Using Bioinformatics to Uncover Small Transmembrane Proteins in Bacteria and Archaea

Tara Hoffman[1], Jeff Kinne[1], Kyu Hong Cho[2]

[1]Department of Mathematics and Computer Science, Indiana State University

[2]Department of Biology, Indiana State University

Summer Undergraduate Research Experience (SURE) 2022

Introduction

Biologists have previously focused on the genes that encode more than 100 amino acids, and the smaller genes for proteins of less than 50 amino acids have been overlooked. These smaller genes have been found to produce small peptides that are biologically active. Among small proteins, our study focuses on small transmembrane proteins (STMPs) because they can be searched through bioinformatic tools with relatively high accuracy. There has been little attempt to find STMPs systematically. STMPs have been found experimentally and serendipitously and they are involved in functions such as cell division, signal transduction, regulation of transporters and drug efflux pumps, as well as stress response.

Pipeline

•Identifying Open Reading Frames (*Orfipy*): Between 60– 180 Nucleotides; Start Codons: ATG, GTG, TTG; Stop Codons: TAA, TGA, TAG

• Determining Transmembrane Likelihood (*Phobius*): Between 15-30 Amino Acids

•BLAST Parameters: Exclude the same species; Use the Prokaryote genome database; results have 30 or more hits

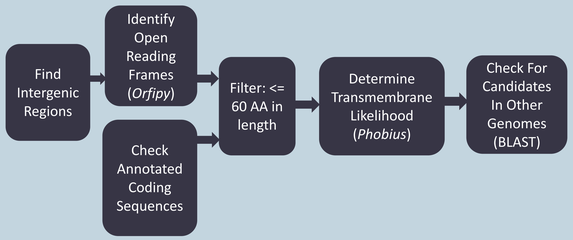


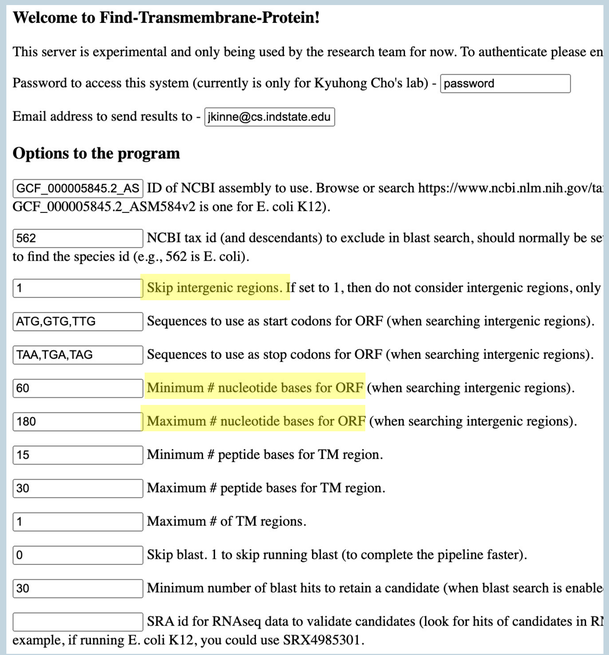
Fig. 1: Our software pipeline coded with Python uses existing bioinformatics tools to take a genome from the National Center for Biotechnology Information (NCBI) and produce a list of putative small transmembrane proteins. The pipeline searches both intergenic regions and annotated coding sequences.

Website

Our program allows the user to input the species genome assembly of their choice. The user is allowed to change if they want to search intergenic regions, the start and stop codons of the ORFs, the minimum and maximum nucleotide bases of the ORF, the size of the transmembrane region, if they want to run BLAST on the candidates, and if you want to include RNAseq data to validate the expression of candidates.

Website link:

cs.indstate.edu/find-transmembrane



Welcome to Find-Transmembrane-Protein!

This server is experimental and only being used by the research team for now. To authenticate please en

Password to access this system (currently is only for Kyuhong Cho's lab) - password

Email address to send results to - jkinne@cs.indstate.edu

Options to the program

GCF_000005845.2_AS ID of NCBI assembly to use. Browse or search https://www.ncbi.nlm.nih.gov/ta
GCF_000005845.2_ASM584v2 is one for E. coli K12).

562 NCBI tax id (and descendants) to exclude in blast search, should normally be se
to find the species id (e.g., 562 is E. coli).

1 Skip intergenic regions. If set to 1, then do not consider intergenic regions, only

ATG,GTG,TTG Sequences to use as start codons for ORF (when searching intergenic regions).

TAA,TGA,TAG Sequences to use as stop codons for ORF (when searching intergenic regions).

60 Minimum # nucleotide bases for ORF (when searching intergenic regions).

180 Maximum # nucleotide bases for ORF (when searching intergenic regions).

15 Minimum # peptide bases for TM region.

30 Maximum # peptide bases for TM region.

1 Maximum # of TM regions.

0 Skip blast. 1 to skip running blast (to complete the pipeline faster).

30 Minimum number of blast hits to retain a candidate (when blast search is enable

SRA id for RNAseq data to validate candidates (look for hits of candidates in RI
example, if running E. coli K12, you could use SRX4985301.

This poster was created using www.PosterPresentations.com

Results from *E. coli* K-12

#hits	query len	query id	query description	query pep sequence
400	49	NC_000913.3_cds-NP_415283.1	[protein=multidrug efflux pump accessory protein AcrZ]	MLELLKSLVFAVIMVPVVMAIILGLIYGLGEVFNIFSGVGKKDQPGQNH
127	46	NC_000913.3_cds-NP_415576.1	[protein=DUF2770 domain-containing protein YceO]	MRPFLQEYLMRRLLHYLINNIREHLMLYLFLWGLLAIMDLIYVFYF
40	34	NC_000913.3_cds-NP_416302.2	[protein=protein YoaI]	MNDQMFVETLIITSSFFAIAVVLVLSVLLIERTG
500	52	NC_000913.3_cds-NP_417152.1	[protein=Pmp3 family protein YqaE]	MGFWRIVITIILPPLGVLLGKGFGWAFIINILLTLLGYIPGLIHAFWVQTRD
139	32	NC_000913.3_cds-NP_418215.1	[protein=ilvXGMEDA operon leader peptide]	MTALLRVISLVVISVVVVIIIPPCGAALGRGKA
64	30	NC_000913.3_cds-YP_001165313.1	[protein=cytochrome bd-II accessory subunit AppX]	MWYLLWFVGILLMCSLSTLVLVWLDPRLKS
84	31	NC_000913.3_cds-YP_001165318.1	[protein=uncharacterized protein YncL]	MNVSSRTVVLINFFAAVGLFTLISMRFGWFI
119	31	NC_000913.3_cds-YP_001165319.1	[protein=small protein MgtS]	MLGNMNVFMAVLGIILFSGFLAAYFSHKWDD
117	27	NC_000913.3_cds-YP_001165320.1	[protein=uncharacterized protein YdgU]	MVGRYRFEFILILILCALITARFYLS
129	29	NC_000913.3_cds-YP_001165321.1	[protein=cytochrome bd-I accessory subunit CydH]	MSTDLKFSLVTTIIVLGLIVAVGLTAALH
99	29	NC_000913.3_cds-YP_001165331.1	[protein=membrane-depolarizing toxin TisB]	MNLVDIAILILKLIVAALQLLDAVLKYLK
158	32	NC_000913.3_cds-YP_002791242.1	[protein=putative membrane protein YoaK]	MRIGIIFPVVIFITAVVFLAWFFIGGYAAPGA
72	24	NC_000913.3_cds-YP_002791243.1	[protein=uncharacterized protein YoaJ]	MKKTTIIMMGVAIIVVLGTELGWW
82	27	NC_000913.3_cds-YP_002791249.1	[protein=putative membrane protein YohP]	MKIILWAVLIIFLIGLLVVTGVFKMIF
243	23	NC_000913.3_cds-YP_002791250.1	[protein=putative membrane protein YpdK]	MKYFFMGISFMVIVWAGTFALMI
31	29	NC_000913.3_cds-YP_004831120.1	[protein=small regulatory membrane protein PmrR]	MKNRVYESLTTVFSVLVVSSFLYIWFATY
86	31	NC_000913.3_cds-YP_009518771.1	[protein=protein YmiC]	MINTNMKYWSWMGAFSLSMLFWAELLWIITH
63	58	NC_000913.3_cds-YP_009518796.1	[protein=protein YecU]	MIKIFIGHYINVFYSTADITLKKQPLLFLAKLMVYSAALTFFTANFHCNMTRKINEYA
51	56	NC_000913.3_cds-YP_009518818.1	[protein=protein YqiD]	MFIAWYWIVLIALVVVGYFLHLKRYCRAFRQDRDALLEARNKYLNSTREETAEKVE
59	35	NC_000913.3_cds-YP_025297.1	[protein=small toxic polypeptide LdrA]	MTLAQFAMIFWHDLAAPILAGIITAAIVSWWRNRK
104	35	NC_000913.3_cds-YP_025298.1	[protein=small toxic polypeptide LdrB]	MTLAQFAMTFWHDLAAPILAGIITAAIVGWWRNRK
59	35	NC_000913.3_cds-YP_025299.1	[protein=small toxic polypeptide LdrC]	MTLAQFAMIFWHDLAAPILAGIITAAIVSWWRNRK
208	49	NC_000913.3_cds-YP_025301.1	[protein=toxin HokB]	MKHNPLVVCLLIICITILTFTLLTRQTLYELRFRDGDKEVAALMACTSR
100	41	NC_000913.3_cds-YP_025303.1	[protein=beta-lactam resistance protein]	MNRLIELTGWIVLVVSVILLGVASHIDNYQPPEQSASVQHK
51	35	NC_000913.3_cds-YP_026227.1	[protein=small toxic polypeptide LdrD]	MTFAELGMAFWHDLAAPVIAGILASMIVNWLNKRK
282	50	NC_000913.3_cds-YP_026229.1	[protein=small toxic polypeptide]	MPQKYRLLSLIVICFTLLFFTWMIRDSLCELHIKQESYELAAFLACKLKE
198	46	NC_000913.3_cds-YP_588449.3	[protein=uncharacterized protein YmiA]	MRLAMPSGNQEPRRDPELKRKAWLAVFLGSALFWVVVALLIWKVWG
221	35	NC_000913.3_cds-YP_588460.1	[protein=UPF0387 family protein YohO]	MRIAKIGVIALFLFMALGGIGGVMLAGYTFILRAG

Fig. 2: Results of running our pipeline on the K-12 strain of *E. coli*. Number of hits indicate how many blast hits each candidate has. Query length is based on the Amino Acid sequence and its length. Query ID indicates the sequence ID from the NCBI annotated coding sequence file that our software pipeline has identified as potentially being transmembrane. Query Description contains brief notes on the putative peptide function from NCBI. Then the final column lists the candidate's peptide function.

Results Summary

	Gram Positive Bacteria	Gram Negative Bacteria	Other Bacteria	Archaea
# of Species Studied	31	40	3	11
# of Candidates	185	215	11	28
Average	5.97	5.38	3.67	2.55

Fig. 3: Results from running our program on 85 prokaryotes. The results were split up based on their biological make-up. The figure lists the total number of species study, candidates, and the average number of candidates for each species. There have tended to be fewer results found for less well studied organisms, with the results for Archaea fitting that pattern.

Conclusion

Our software pipeline searches bacterial genomes for small transmembrane proteins, and we have analyzed the results from running the pipeline on multiple different species. Many small transmembrane proteins were found in the well studied *E. coli* K-12 strain and proves that our program finds annotated coding sequences that are small transmembrane peptides. Our program can be used on lesser-known species to discover new small transmembrane proteins.

Further Directions

- Evaluate other tools for identifying novel genes and compare results with our software pipeline
- Work towards publishing a paper of our findings and our program

Acknowledgments

Funding Sources: BD4ISU, NIH 1R25MD011712-04

References

1. Garai, Preeti, and Anne Blanc-Potard. “Uncovering Small Membrane Proteins in Pathogenic Bacteria: Regulatory Functions and Therapeutic Potential.” *Molecular Microbiology*, vol. 114, no. 5, 2020, pp. 710–720., doi:10.1111/mmi.14564.
2. Kamar, Rita, et al. “DltX of Bacillus Thuringiensis Is Essential for D-Alanylation of Teichoic Acids and Resistance to Antimicrobial Response in Insects.” *Frontiers in Microbiology*, vol. 8, 2017, doi:10.3389/fmicb.2017.01437.
3. Orr, Mona Wu, et al. “Alternative ORFs and Small ORFs: Shedding Light on the Dark Proteome.” *Nucleic Acids Research*, vol. 48, no. 3, 2019, pp. 1029–1042., doi:10.1093/nar/gkz734.
4. Sousa, Maria E., and Michael H. Farkas. “Micropeptide.” *PLOS Genetics*, vol. 14, no. 12, 2018, doi:10.1371/journal.pgen.1007764.