Uncovering Small Transmembrane Peptides in Bacteria

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 Experiences (SURE) 2021

Introduction

In the search to identify genes the main focus has been on genes encoding at least 100 amino acids. Genes for much smaller proteins, i.e., those encoding fewer than 50 amino acids, have been mostly overlooked. Among the potential functions of small proteins, the present study searches for those that are *transmembrane*. We have implemented a software pipeline that identifies potential small transmembrane peptides by searching an organism's genome for open reading frames that have a transmembrane signature in their sequence. Since most, if not all, small transmembrane peptides found so far have an alpha helical transmembrane motif, we search for small alpha helical transmembrane peptides. In particular, the side chains of the alpha helix must be hydrophobic, and this characteristic can be searched by existing algorithms. We focus our search on bacteria because bacterial gene structure is simpler (without introns).

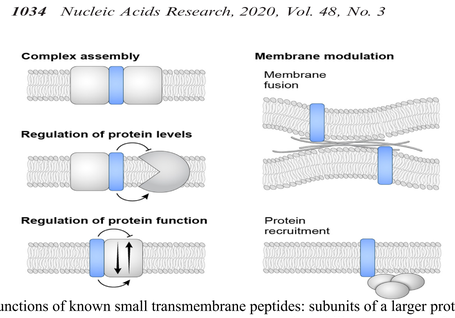

Fig. 1: Functions of known small transmembrane peptides: subunits of a larger protein complex (complex assembly), having an impact on protein stability (regulation of protein levels), regulating transporter activities (regulation of protein function), mediating membrane reorganization (membrane modulation - fusion), and recruiting complexes to cell membranes (membrane modulation – protein recruitment).

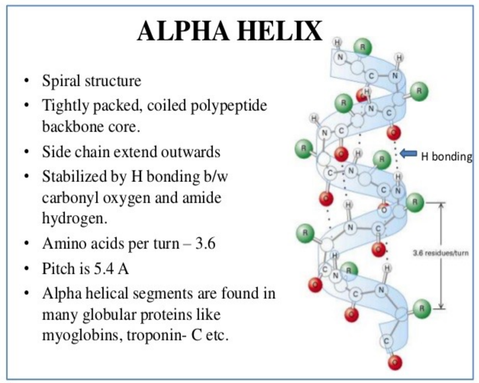

Fig. 2: The side chains of the alpha helix should be hydrophobic so that it can penetrate through the membrane. One full turn of the alpha helix is 5.4 A. Therefore, in order to penetrate the membrane, it would take 7.4 turns which is 27 Amino Acids. Alpha helices are smaller than beta sheets which is why we look for these rather than beta sheets.

Pipeline

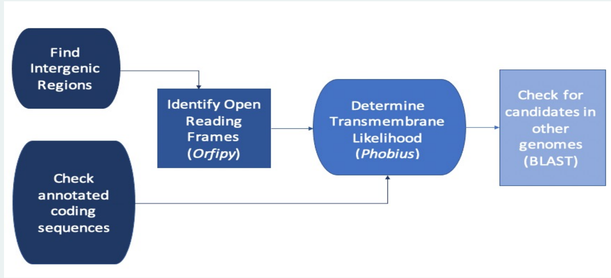

Fig. 3: Our software pipeline is Python code that uses existing bioinformatics tools to take a genome from the National Center for Biotechnology Information (NCBI) and produce a list of putative small transmembrane peptides. The pipeline searches both intergenic regions and annotated coding sequences.

- Identifying Open Reading Frames (*Orfipy*): Restrict to those between 60-180 Nucleotides; Start Codons: ATG, GTG, TTG; Stop Codons: TAA, TGA, TAG
- Determining Transmembrane Likelihood (*Phobius*): Restrict to those between 15-30 Amino Acids
- BLAST Parameters: Exclude other close relatives; Use the Prokaryote genome database; restrict to those with 30 or more hits

Results from *S. pyogenes* HSC5

Number of Blast Hits	Query ID	Query Description	Peptide Sequence
108	lcl|CP006366.1_prot_AGQ28506.1_822	[**protein=citrate lyase**] [location=872034..872168]	MTINMDHLISSFELMALGMAGVFIVLGILYLVAEILIKLFPVSK
304	lcl|CP006366.1_prot_AGQ27766.1_992	[**protein=D-alanyl-lipoteichoic acid biosynthesis protein DltX**] [location=complement(1028105..1028248)]	MIKNKSRMRGIAVFIGKTILFYLILMLLVYFFGYLGHGQSNFIYNEF
120	lcl|CP006366.1_prot_AGQ27781.1_1007	[**protein=hypothetical protein**] [location=complement(1043707..1043835)]	MAFGENGPRKKTTFEKVTMFVVILMVLVTVGGLIASALSVLM
40	lcl|CP006366.1_prot_AGQ28599.1_1141	[**protein=Putative Type I toxin-antitoxin, prophage gene**] [location=1163621..1163716]	MCEAIFTTIIAPLLVGIILLLIQKWLDDSAD
47	lcl|CP006366.1_prot_AGQ28662.1_1455	[**protein=Putative Type I toxin-antitoxin Prophage gene**] [location=1492966..1493061]	MCETIFTTIIAPLLVGIILLLIQKWLDDSAD
121	lcl|CP006366.1_prot_AGQ28282.1_1638	[**protein=secE**] [location=complement(1695067..1695243)]	MGFISGTFKVLKDTTWPNRKQRWKDFISVLEYTAFFTVIIYIFDQLLAKSVLALINLF

Fig. 4: Results of running our pipeline on the HSC5 strain of *S.* pyogenes. Query ID indicates the sequence: those beginning with lcl are annotated coding sequences that our software pipeline has identified as potentially being transmembrane, while query IDs containing "intergenic" (see Fig. 5) are open reading frames found in an intergenic region. Query Description contains brief notes on the putative peptide, including any protein that we have determined the putative peptide is linked to (i.e., is near in the genome).

Figures 4 and 5 contain the results from running our pipeline on two different strains of *Streptococcus* pyogenes. Figure 5 notes instances where the same peptide sequence is found in both strains.

Results from *S. pyogenes* M1 GAS

Number of Blast Hits	Query ID	Query Description	What it is listed as in *S. pyogenes* HSC5	Peptide Sequence
108	NC_002737.2:**intergenic**:973407-973593_ORF.1	[17-149](+) type: complete length:132 frame:3 start:ATG stop:TAA	**Protein: citrate lyase**	MTINMDHLISSFELMALGMAGVFIVLGILYLVAEILIKLFPVSK
304	lcl|NC_002737.2_prot_WP_002984200.1_1049	[**protein=D-alanyl-lipoteichoic acid biosynthesis protein DltX**] [location=complement(1090809..1090952)]	**Protein=D-alanyl-lipoteichoic acid biosynthesis protein DltX**	MIKNKSRMRGIAVFIGKTILFYLILMLLVYFFGYLGHGQSNFIYNEF
120	lcl|NC_002737.2_prot_WP_002989532.1_1064	[**protein=hypothetical protein**] [location=complement(1106412..1106540)]	**Protein=hypothetical protein**	MAFGENGPRKKTTFEKVTMFVVILMVLVTVGGLIASALSVLM
121	lcl|NC_002737.2_prot_WP_002982348.1_1629	[**protein=secE**] [location=complement(1713231..1713407)]	**Protein=secE**	MGFISGTFKVLKDTTWPNRKQRWKDFISVLEYTAFFTVIIYIFDQLLAKSVLALINLF
343	lcl|NC_002737.2_prot_WP_011185066.1_1680	[protein=putative holin-like toxin] [location=1774849..1774992]	**Protein=putative holin-like toxin, prophage gene**	MSGGGAYVCKSQKPKERRRQGLSVYETLTLMIAFGTLIVAIMNNKNK

Fig. 5: Results of running our pipeline on the M1 GAS strain of *S.* pyogenes. The fourth column represents whether-or-not the same peptide sequence was also found in the HSC5 strain (see Figure 4).

Conclusions

Our software pipeline searches bacterial genomes for small transmembrane proteins, and we have analyzed the results from running the pipeline on strains of the pathogen *Streptococcus pyogenes*. A number of annotated coding sequences were identified as putative small transmembrane proteins, but none were identified within intergenic regions.

Further Directions

- Evaluate other tools for identifying novel genes and compare results with our software pipeline.
- Run our pipeline and analyze results on other bacterial genomes, including those which contain well-documented small transmembrane proteins (to verify our pipeline finds these known positives).
- Use RNAseq data to confirm that putative small transmembrane protein genes are being transcribed.

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.