CoRNonCOB App Note

Ethan Holleman

April 24, 2020

1 Abstract

CoRNonCOB (Compare Really Non-Conserved Oligopeptides in Bacteria) is a tool that identifies peptides in the non-coding regions of bacterial genomes that may contribute to a phenotype of interest. CornNonCOB is designed to work at the very specific scale of individual phenotypes of specific strains of bacteria and because of this RNA-seq data that could illuminate the expression profiles of the bacteria of interest may not be available. Therefore, CornNonCOB attempts to leverage genomic data by identifying peptides in non-coding regions that are conserved in the members of a positive phenotype and absent in a control. With this tool, we hope to identify the peptides that are unique to Lactobacillus jensenii that produce a microbial agent to kill E. Coli bacteria. Through identifying these unique peptides, we hope to provide another parameter to characterize phenotypes. Our tool takes x amount of genomes in FASTA format from each phenotype, and generates the peptides unique to each respective phenotype. Our tool is available at the GitHub page here, where additional information on the parameters required and our current analyses can be found. Documentation is available at the CornOnCOB website here.

2 Introduction

Currently, there are no tools for identifying peptides located in the non-coding regions of bacterial genomes. There are tools that identify peptides in other organisms like plants. For example, the Small Peptide Alignment Discovery Application (SPADA), is an application to identify the small peptides in plant genomes. SPADA uses homology-based modeling to predict peptides that are one to two exons in length. Unfortunately, SPADA is not equipped to handle bacterial genomes because their genomes do not consist of introns.1

There is a database that stores the current known antimicrobial peptides called the Antimicrobial Peptide Database (APD). APD has aided in classifying Antimicrobial Peptides (AMP) and identifying peptides with similar sequences.2 AMPs are peptides associated with the ability to kill bacteria, viruses and fungi.3 CoRNonCOB is designed to identify peptides in the noncoding regions of bacterial genomes. For this specific project, we are looking at L. jensenii. L. jensenii are associated with a healthy vaginal and urinary tract microbiome environment. In fact, Lactobacillus probiotics have been found to prevent recurrent Urinary Tract Infections (UTIs) in women.4

There are two phenotypes of L. jensenii that we hope to distinguish by identifying unique peptides. One phenotype characterized by their ability to produce a microbial agent to kill E. Coli, while the other phenotype does not appear to produce any agent to kill E. Coli. With CoRNONCOB we hope to identify if there are any AMPs in the L. jensenii that show characteristics of killing E.Coli.

3 Implementation

3.1 Workflow

CornOnCob is written in python and incorporates Prokka for prokaryotic gene prediction, BioPython for sequence translation and IO, the CD-HIT suite of programs for sequence clustering and the modelamp package for chemical property calculations. The generalized workflow can be seen in figure ??.

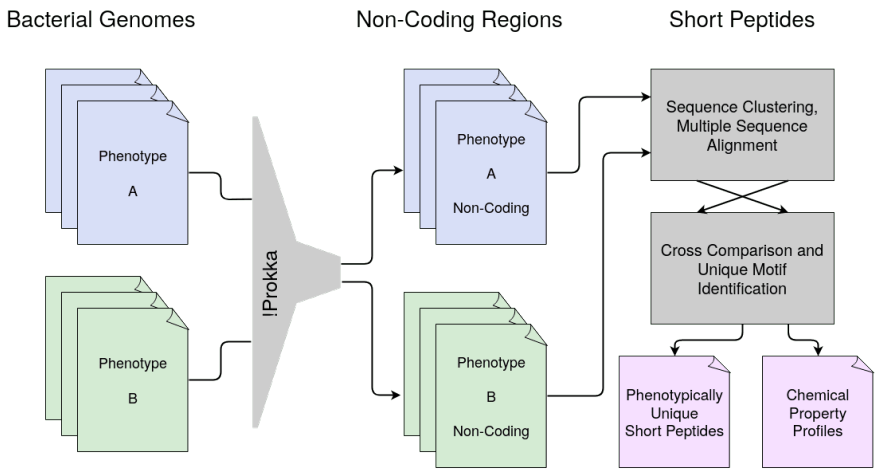

Figure 1: Histogram of candidate peptide properties identified from Putoni Lab *Lactobacillus crispatus* genomes vs. all AMPs available from the AMP3 Database.

The user supplies paths to directories containing the genomes of the members of each phenotype. Running prokka on each genome produces a gff file of predicted coding regions which CornOnCob then reads and inverts to yield predicted non-coding regions for each genome. Non-coding regions are translated into all six reading frames, written as fasta files in the program output. These fasta files representing the predicted non-coding peptides from individual genomes are then concatenated to create a file containing the predicted non-coding peptides of an entire phenotype. CD-HIT is then used to cluster the concatenated non-coding peptide files by sequence similarity. CD-HIT uses a greedy algorithm that sorts sequences by length and then searches for similar sequences via kmer comparison allowing for extremely fast sequence clustering. The longest sequence in each cluster is considered the representative sequence for that cluster. The user can control the minimum overlap of subject to representative sequence and minimum identify thresholds in the CoRNonCOB arguments.

Peptide clusters meeting a certain threshold of genome participation are considered conserved and written to a new fasta file representing the predicted conserved non-coding peptides for an individual phenotype. CD-HIT-2D, which uses the same basic algorithm of CD-HIT but looks for peptides in one library that

are similar to another library, compares the conserved non-coding peptides between the two phenotypes. CoRNonCOB then parses this output and removes non-unique sequences to produce a fasta file of candidate peptides that are uniquely conserved in the positive phenotype. Lastly, using the modelamp python package basic chemical properties of each peptide are predicted and written to a csv file.

3.2 Run Time and Memory Usage Estimates

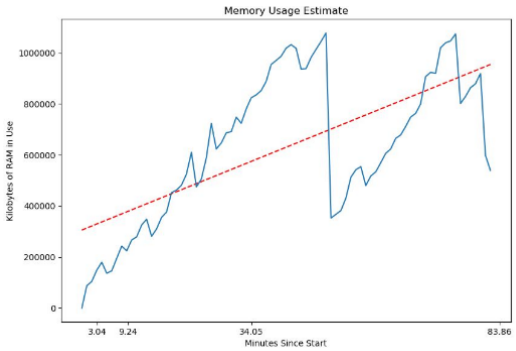

Figure 2: CoRNonCOB Memory sage with increasing numbers of genomes.

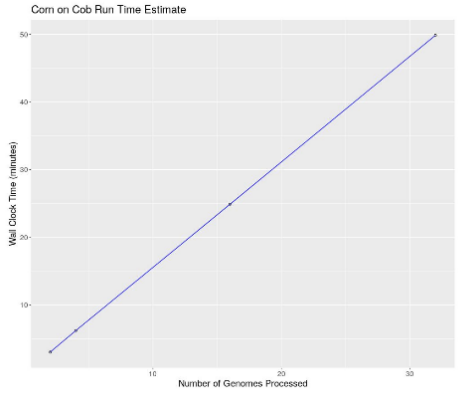

Figure 3: CoRNonCOB allclock runtime with increasing numbers of genomes.

Figure X shows the time for CoRNonCOB to execute given an increasing number of genomes. All of the genomes used were of similar length. WIth this metric CoRNonCOB appears to run in linear time. Operations performed in python by CoRNonCOB such as non-coding region extraction and translation are done in linear time with respect to number of genomes and number of non-coding regions.

However, Prokka is by far the rate limiting step in CoRNonCOB's runtime which is estimated as $O(n^2)$ where n is the length of the genome used for prediction. Therefore if genome size remains relatively constant, runtime will reflect figure x but will increase to $O(n^2)$ as genome size increases.

4 Results

Using CoRNonCOB in test mode and with the bacterial genomes provided from the Putonti Lab and the University of Wisconsin we randomly inserted three experimentally verified AMPs into the genome of the positive phenotype member which we would expect to be present in the final output if CoRNonCOB is functioning as envisioned.

Test Type	Peptides in Test	Peptides Expected in Final Results	Peptides in Final Results	Percent Correctness
Positive	9	9	7	77
Negative	9	0	0	100

Currently, the CoRNonCOB pipeline has been able to recover two of the three test peptides along with a large number of other candidate peptides. We are currently debugging to determine where the third test peptide may have been filtered out and plan on running additional tests using more positive peptides, negative peptides which we would expect to filter out of our results, and additional genomes from NCBI.

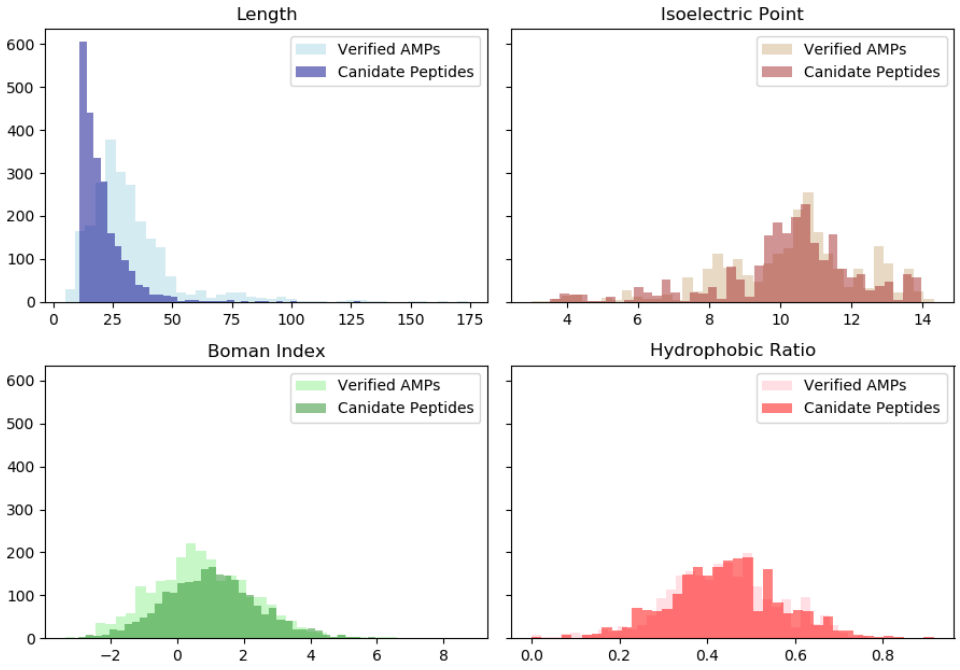

Figure 4: Histogram of candidate peptide properties identified from Putoni Lab *Lactobacillus crispatus* genomes vs. all AMPs available from the AMP3 Database.

We have also reached out to the Sheshu lab at George Mason university to obtain the source code for their AMP classification model described in their 2018 publication Deep learning improves antimicrobial peptide recognition in order to try and narrow down the candidate peptides in the final fasta output.[1]

References

[1] Timo Duchrow, Timur Shtatland, Daniel Guettler, Misha Pivovarov, Stefan Kramer, and Ralph Weissleder. Enhancing navigation in biomedical databases by community voting and database-driven text classification. *BMC Bioinformatics*, 10(1):317, 2009.

[2] T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[3] Boris Vishnepolsky and Malak Pirtskhalava. Prediction of linear cationic antimicrobial peptides based on characteristics responsible for their interaction with the membranes. *Journal of Chemical Information and Modeling*, 54(5):1512–1523, 2014.

[4] Z. Wang. Apd: the antimicrobial peptide database. *Nucleic Acids Research*, 32(90001), Jan 2004.

[5] Peng Zhou, Kevin At Silverstein, Liangliang Gao, Jonathan D Walton, Sumitha Nallu, Joseph Guhlin, and Nevin D Young. Detecting small plant peptides using spada (small peptide alignment discovery application). *BMC Bioinformatics*, 14(1):335, 2013.