BioE Final Project

Quentin Delepine, Raymond Guo, Aditya Srivastava, Roshan Rao

Introduction

In this project we examine the use of pathogenic protein prediction software on various isolates of *Salmonella* genomes. In particular, we examine the Metagenomic Pathogenic Protein Prediction (MP3) tool, a commonly used software available as a command line tool and as a webserver. We seek to determine whether MP3 is able to identify which of these genomes is pathogenic, and if so to mark those genomes as pathogenic. We also want to compare the different genomes by their pathogenicity.

Since MP3 takes in protein sequences as input, we first assemble our genome. The assembled genome is then annotated to identify coding regions, from which protein sequences are extracted. Finally, we run MP3 on the resulting sequences to determine which proteins are classified as pathogenic.

We perform this analysis on all available data from 13 groups (ours + 12 other groups which made their assemblies available). Using this information, we construct a database of proteins from each assembly, including pathogenicity information.

By querying this database, we find that each genome has more than 30% pathogenic proteins, well over the threshold for pathogenicity identified in the original MP3 paper. This suggests that either all genomes are pathogenic or that MP3 is not able to differentiate pathogenic from non-pathogenic *Salmonella* isolates. We make our database available at Github, as well as include several starter scripts and query examples for further exploration.

Background

Salmonella

Salmonella are an enteric bacteria, meaning that they're Gram-negative, facultatively anaerobic rod-shaped bacilli, and members of the Enterobacteriacae family. Today, *Salmonella* is considered to be one of the major causes of foodborne gastrointestinal diseases seen world-wide, due to the consumption of contaminated food items.

Salmonella is composed of an extremely diverse group of bacteria broken down into two main species: *Salmonella bongori and Salmonella enterica*, which is further divided into 6 subspecies and over 2500 serovars. Each subspecies and serovar of *Salmonella* is variable in its pathogenicity, which is also dependent on the host species. For example,

McWhorter and Chousalkar (2015) tested 10 different strains of *Salmonella*, including *Salmonella enterica* strains *S*. Adelaide, *S*. Bredeney, *S*. Cerro, *S*. Orion, *S*. Senftenberg, *S*. Virchow, and *S*. Typhimurium definitive types 44 (DT 44), 170=108 (DT170=108), 135 (DT135), and 193 (DT193), and found that both invasive capacity and pathogenicity varied across strains. In particular, *S*. Typhimurium DT 170=108 was significantly more invasive than other strains, with *S*. Adelaide and *S*. Cerro showing negligible invasion, while the DT135, DT170=108, and DT193 strains were significantly more pathogenic than other strains based on mortality rate of contaminated mice.

Other studies have found similar variations between the pathogenicity of different *Salmonella* strains. For example, Swearingen et al. (2012) studies 32 different *Salmonella* strains and found that "the strains clearly differed in their ability to colonize the mice and to be shed in feces."

The virulent nature of *Salmonella* have been mapped to a group of pathogenicity islands (SPIs). Even though there are at least 21 SPIs in *Salmonella*, the two most student SPIs are SPI-1 and SPI-2.

SPI-1 is a 40-kb DNA region, most responsible for the invasion of nonphagocytic cells for replication. The island encodes a type III secretion system, which is responsible for transporting bacterial proteins into the cytosol of host cells, while the SP-2 is a 40-kb DNA region encoding another type III secretion system that is responsible for modifying the intracellular environment. Outside of SPI-1 and SPI-2, there are other genes that are heavily associated with virulence:

- *mgtC*, in SPI-3, which is takes part in macrophage invasion
- *sopB*, in SPI-5, which encodes a phosphate phosphatase which affects the inositol phosphate signaling pathways
- *spvR, spvA, spvB, spvC, spvD*, a group of genes found in the *spv* locus (*Salmonella* plasmid virulence)
- *sopE*, which encodes an accessory epithelial cell invasion factor
- *sodCI* and *sodCII*, encoding periplasmic superoxide dismutase enzymes
- *shdA*, often associated with fecal shedding and adaptation to the intestine environment of warm-blooded animals
- There are other genes that are considered virulent as well.

Overall, the differing virulent nature of *Salmonella* can be attributed partly to the genetic variation between different strains. Between the majority of strains, there are a set of virulent genes that are shared; in particular Campioni et al. (2012) investigated 128 strains and found that almost all of the strains harbored a set of 13 virulence genes

(*InvA, sipA, sipD, sopB, sopD, sopE2, ssaR, sifA, spvB, Prot6E, flgK, fljB, flgL, sdfl*). One strain were negative for the presence of SPI-1 gene *sipA,* while 2 strains were negative for the presence of *prot6E*, a plasmidial gene.

Nevertheless, there is still significant variation between the strains. A comparative genomic analysis of *Salmonella enterica* conducted by Jacobsen et al. (2011) found that the *S.* Enteriditis genome contained regions of difference not present in all serovars, including *S.* Typhimurium. The regions of difference included two coding sequences within SPI-19, genes within ROD21, as well as the *peg* fimbrial operon. Similarly, genes such as *sopE* and *shdA* are encoded only in some isolates of *S.* Typimurium and other subspecies I serovars. *Salmonella* strains also differ in the presence of *sodCI*.

MP3

MP3 is a combined Support Vector Machine (SVM) and Hidden Markov Model (HMM) approach to predicting pathogenicity. First, an HMM is used to identify Pfam domains within proteins in the training dataset. Domains are then classified as exclusively pathogenic (only found in pathogenic proteins), exclusively non-pathogenic (only found in non-pathogenic proteins), or shared (can be found in either). Then, given a query protein, the HMM classifies it as pathogenic if it contains at least one exclusively pathogenic domain, non-pathogenic if it contains at least one non-pathogenic domain and no pathogenic domains, and unknown if there are no known domains or it consists only of shared domains.

If the HMM fails to classify the protein, the task is then passed on to the SVM. The SVM takes in a bag-of-words representation of amino acids and dipeptides. For the amino acid count, each protein is collapsed into a length 20 vector, with each entry containing the percentage of the protein consisting of that amino acid. Similarly for dipeptides, the protein is collapsed to a length 400 vector, with each entry corresponding to the percentage of protein consisting of that dipeptide. This feature vector is then passed to an SVM with a polynomial kernel (d=3). This SVM then makes the prediction for all unclassified proteins.

Methods

Assembly

We assembled our genome using the open source program SPAdes, which was designed specifically to assemble small genomes such as ours. The command:

spades.py -o .
-1 /bigdata/FinalProject_data/190724_SARA_Genomes/SARA_5_S28_L004_R1_001.fastq.gz
-2 /bigdata/FinalProject_data/190724_SARA_Genomes/SARA_5_S28_L004_R2_001.fastq.gz -t 1

Once the genome had been assembled by SPAdes, we ran the generated contigs through assembly-stats to get an overview of the genome:
assembly-stats ./contigs.fasta ./scaffolds.fasta

Annotations

Using rna_hmm3 we generated a HMM over the assembly, and using bedtools, we extracted 16S ribosomal RNA sequences from the contigs using the generated HMM. We then fed the 16S sequences into SeqMatch to find that we had a *Salmonella* genome. Finally, we uploaded our contigs to RAST for annotation.
rna_hmm3.py -i /bigdata/FinalProject_groups/Group_5/assembly/contigs.fasta
-o ./rna_hmm3_o -L /bigdata/FinalProject_groups/Group_5/rna_hmm3/HMM3

bedtools getfasta -fi ./contigs.fasta -bed ./rna_hmm3_16 -fo ./nucleic_acids

MP3

We downloaded the annotated protein sequences from the RAST job (genbank) and ran these through MP3 to find pathogenic proteins. This generated a table of results which we parsed into a pandas dataframe to function as a database of pathogenicity results. To extend our database, we further annotated all the other groups' genomes using RAST, downloaded the respective annotated protein sequences from the associated RAST jobs, and ran these through MP3 and our parsing script as well. We also added the amino acid sequence of each protein to our database using Genbank data and added a group # identifier to distinguish between all the groups (since every group had a *Salmonella* genome).

We then demonstrated some sample queries using our constructed database and also derived statistics about the database.
mp3 Group5.faa 1 30 -0.2

Results & Analysis

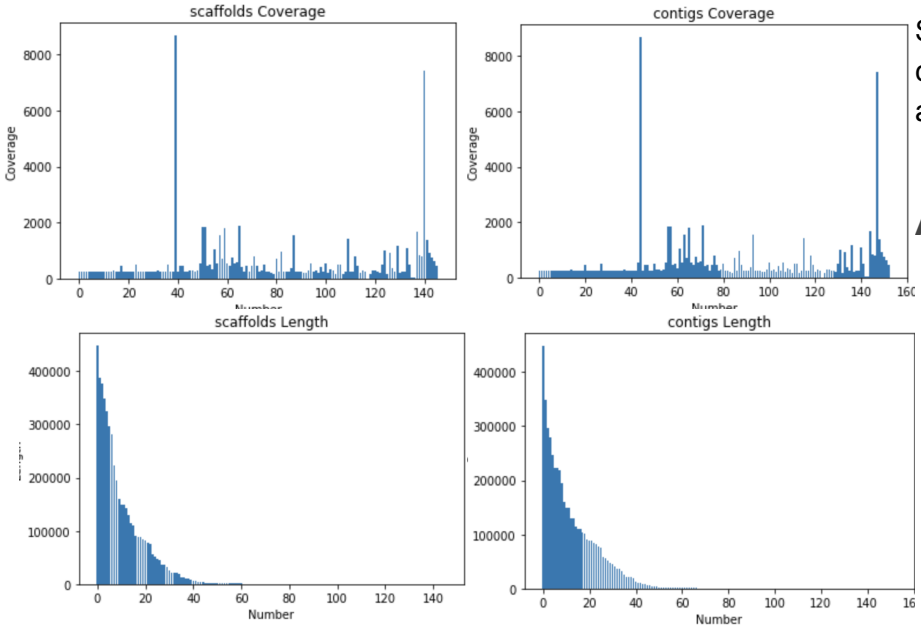
Assembly Output

stats for ./contigs.fasta
sum = 4960322, n = 153, ave = 32420.41, largest = 449208
N50 = 194186, n = 9
N60 = 130292, n = 13
N70 = 109439, n = 17
N80 = 87947, n = 22
N90 = 51460, n = 29
N100 = 56, n = 153

stats for ./scaffolds.fasta
sum = 4960972, n = 146, ave = 33979.26, largest = 449208
N50 = 223794, n = 8
N60 = 159283, n = 10
N70 = 143384, n = 13
N80 = 89581, n = 18
N90 = 56692, n = 24
N100 = 56, n = 146

N_count = 0
Gaps = 0

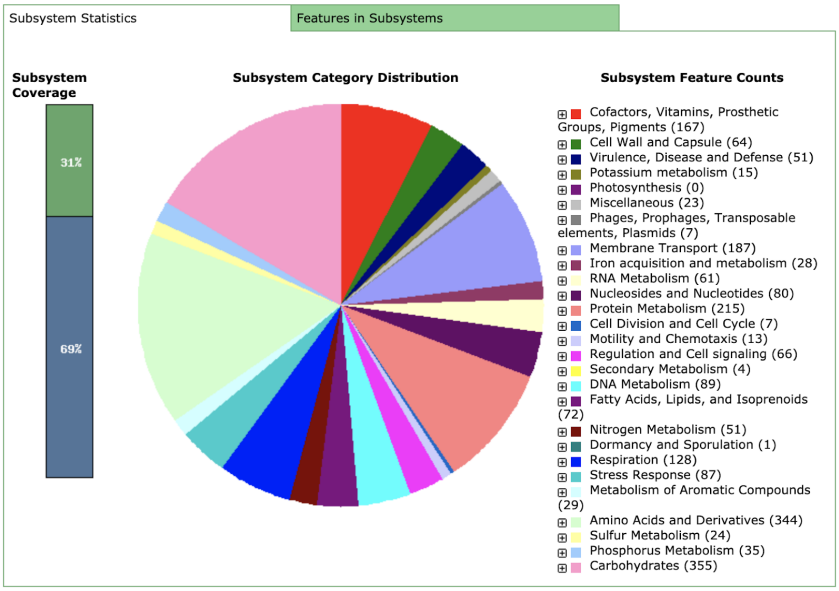
N_count = 700
Gaps = 7



Summary graphs from contigs/scaffolds coverage and length distributions.

Annotations (RAST)

Subsystem Information



Database & Analysis

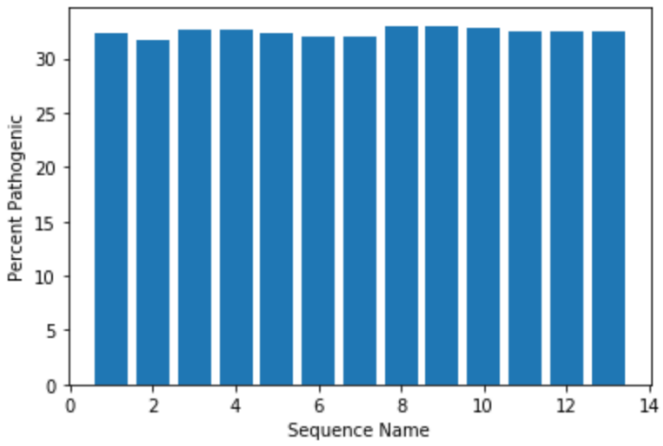
We developed a database across 13 Groups' data (all data which was on the server). We then ran genomes through RAST and downloaded genbank outputs. Those outputs were then fed through the mp3 algorithm. Two files, a genbank file and a table of results, were integrated to form a searchable database of proteins in the genome which is searchable by: group, sequence name, type of protein, pathogenic assignment, sequence, protein product, and organism

	Group	Sr._No.	Sequence_Name	Type_of_Pfam_domains	HMM_Prediction	SVM_Score	SVM_prediction	Hybrid_Prediction	Assignment
0	1	1	fig|6666666.498498.peg.1337	Uncalssified protein	----------	-0.090801	Pathogenic	Pathogenic	S
1	1	2	fig|6666666.498498.peg.1338	Excl. Non-pathogenic	Non-Pathogenic	0.495358	Pathogenic	Non-Pathogenic	H
2	1	3	fig|6666666.498498.peg.1339	Excl. Non-pathogenic	Non-Pathogenic	0.029591	Pathogenic	Non-Pathogenic	H
3	1	4	fig|6666666.498498.peg.1340	Uncalssified protein	----------	-0.623353	Non-Pathogenic	Non-Pathogenic	S
4	1	5	fig|6666666.498498.peg.1341	Uncalssified protein	----------	1.387766	Pathogenic	Pathogenic	S

Sequence	Product	Organism
RQAWRTIALFCVTECFPEDVITDKVEPLTPVYLMTTLMPDVPLTDA...	hypothetical protein	salmonella sp. Bacteria.
MQAIASELSARLNTPVEVGGVEANMAVAGALTTPGCDAPLAILDLG...	Glycerol dehydratase reactivation factor large...	salmonella sp. Bacteria.
MEFFREPLSPSVFAKVVYLKEGELIPVDNQTSLEKIRLVRRQAKEK...	Glycerol dehydratase reactivation factor large...	salmonella sp. Bacteria.
MPTAIEKALDFIGGMNTSASVPHSMDESTAKGILKYLHDLGVPVSP...	Uncharacterized protein YoaC	salmonella sp. Bacteria.
MKNNIEETIGKYLPILMILPLAGLAELASLYSIQALLPKLSEVYNI...	Uncharacterized MFS-type transporter STM0328.s	salmonella sp. Bacteria.

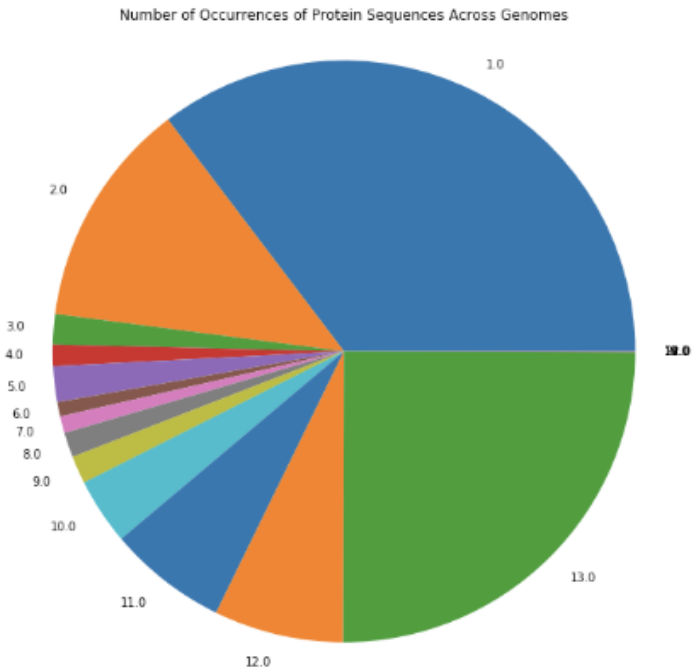
Pathogenicity of Strains

We found that the genomes we scanned had a minimum of 31.7% pathogenic proteins, a maximum of 32.5%, and an average of 33.0%. The original MP3 paper found that pathogenic genomes had between 20-30% of proteins classified as pathogenic, while non-pathogenic genomes had ~10% of proteins classified as pathogenic. Results suggest that all strains are pathogenic, or that MP3 does a poor job of classifying Salmonella isolates as pathogenic/non-pathogenic.



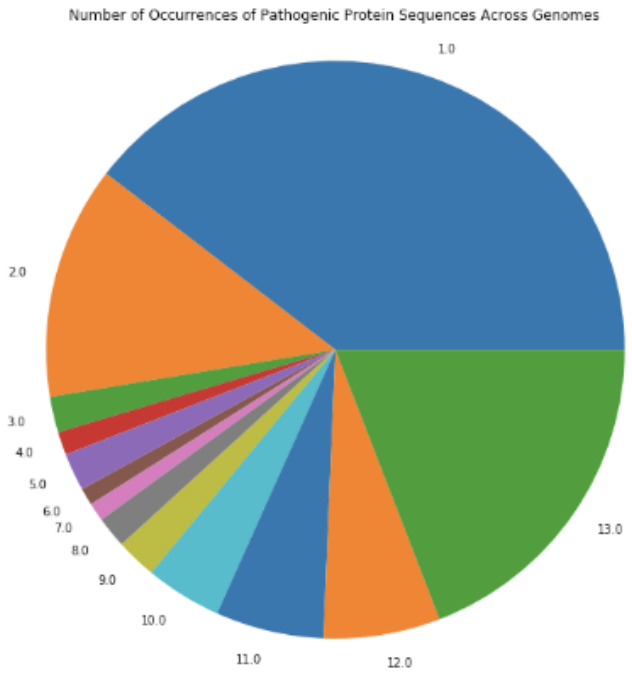
Number of Protein Sequence Occurrences Across Strains

We found that 35.3% of proteins were unique to only one genome and 64.7% of proteins occurred in two more more genomes. The most common protein sequence was MSKEKFERTKPHVNVGTIGHVDH, corresponding to Translation elongation factor Tu. It occurred in 11 genomes, twice each and is non-pathogenic.



Number of Pathogenic Protein Sequence Occurrences Across Strains

We found that 39.6% of pathogenic proteins were unique to only one genome and 60.4% of pathogenic proteins occurred in two more more genomes. The most common protein sequences occurred 13 times across all the genomes and comprised 19.1% of all proteins.



Future Work

Improve Database

We have a few ideas for how to improve the database, including increasing the number of keys to search it, link similar proteins to each other, possibly using BLAST, and improve the sqlite3 relational database we made by perhaps adding more useful indexing.

Increase automation of pathogenic prediction pipeline

We envision a small CLI to run parts (or all) of the analysis automatically - e.g. given a set of reads, our tool would do the assembly, annotation, database creation, and some preliminary analysis automatically by essentially functioning as a wrapper/shim around the CLI tools we used for this project (bedtools, MP3 etc.)

Improve MP3 Classification

We are not sure that the MP3 software is still state-of-the-art at accurately classifying pathogenic vs. non-pathogenic strains. We would like to explore further models for classification, perhaps something involving a neural network trained on a large corpus of genomic data.

References & Resources

- Github Repo with code and examples: https://github.com/rmrao/cmpbio-final-project

- Presentation of this Report: https://docs.google.com/presentation/d/1RbJZauYtNAMGt-8NderWXZFYrPkHX5TcpzVry1Rf-8M/edit#slide=id.g7a6dc42288_1_35
- MP3 Software: (n.d.). Retrieved from http://metagenomics.iiserb.ac.in/mp3/index.php.
- Chaudhary, J. H., Nayak, J. B., Brahmbhatt, M. N., & Makwana, P. P. (2015). Virulence genes detection of Salmonella serovars isolated from pork and slaughterhouse environment in Ahmedabad, Gujarat. *Veterinary world*, *8*(1), 121–124. doi:10.14202/vetworld.2015.121-124
- Campioni, Fábio & Bergamini, Alzira & Falcão, Juliana. (2012). Genetic diversity, virulence genes and antimicrobial resistance of Salmonella Enteritidis isolated from food and humans over a 24-year period in Brazil. Food microbiology. 32. 254-64. 10.1016/j.fm.2012.06.008.
- Dodd, Christine & Aldsworth, T. & Stein, R.A. & Cliver, D.O. & Riemann, H.P.. (2017). Foodborne Diseases: Third Edition.
- Gupta, A., Kapil, R., Dhakan, D. B., & Sharma, V. K. (2014). MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PloS one*, *9*(4), e93907. doi:10.1371/journal.pone.0093907
- Jacobsen, A., Hendriksen, R. S., Aaresturp, F. M., Ussery, D. W., & Friis, C. (2011). The Salmonella enterica pan-genome. *Microbial ecology*, *62*(3), 487–504. doi:10.1007/s00248-011-9880-1
- McWhorter, A. R., & Chousalkar, K. K. (2015). Comparative phenotypic and genotypic virulence of Salmonella strains isolated from Australian layer farms. *Frontiers in microbiology*, *6*, 12. doi:10.3389/fmicb.2015.00012
- Suez, J., Porwollik, S., Dagan, A., Marzel, A., Schorr, Y. I., Desai, P. T., … Gal-Mor, O. (2013). Virulence gene profiling and pathogenicity characterization of non-typhoidal Salmonella accounted for invasive disease in humans. *PloS one*, *8*(3), e58449. doi:10.1371/journal.pone.0058449
- Swearingen, M. C., Porwollik, S., Desai, P. T., McClelland, M., & Ahmer, B. M. (2012). Virulence of 32 Salmonella strains in mice. *PloS one*, *7*(4), e36043. doi:10.1371/journal.pone.0036043