**CS31420 Computational Bioinformatics tob31 assignment Part 2**

# Abstract

The task at hand was to take 4 genomes provided in FASTA file format and find meaningful information about the genomes provided. Then Summarise the information obtained. This meaningful information would include the species identification and similarity analysis of the genomes.

# Introduction

To begin with I used pubmlst an online genomic species identifier to try and identify the species from these genomes.

Next, I took the 4 genomes and ran a kmer analyse using jellyfish. The kmer size was initially 7. I did the same thing for ecoli and bacsillis.

Next, I ran prodigal on the genomes so that I could obtain the ORF’s from the fasta files, from is I was able to obtain the CDS’s,

Using the ORF’s I then proceeded to run blastx on the orf’s to try and find regions of local similarity within the swisprot database. Blastx was too slow, so I then moved on to using diamond which produced the results I needed very quickly.

I then used the orfs I obtained from prodigal to retrieve some orthology results using eggnog.

Looking at the results I had collected using various bioinformatic tools I was able to make some assumptions about were the genomes may have been collected from.

# Materials and methods & results

## Kmer analyse,

I started by running jellyfish on the genomes fasta files. (Explain what jellyfish is). Using the following commands with kmer7. For example

$ ./jellyfish-linux count -m 7 -s 10M -C rug213.fa

$./jellyfish-linux dump mer\_counts.jf > 213kmer7.fa -c

I soon realised that kmer7 was too small to do any real anyalsis so I did it again with kmer27 then kmer61. I choose to use kmer61 to continue with since this is the larger number and would most likely show actual comparisons when comparing files. The files needed to be sorted and compared using the following commands for all files. For example.

$Sort -k 2 -n 61kmer213.fa > 213sort61.txt

$comm -3 213sort61 384sort61 | wc -l > comm213-384.txt

Comm here compares the overlap between the files and wc counts the lines so we can see the total similarity. I had problems comparing the kmer files to find similarities. So I scanned through the files to look for any obvious information I could see.

Just from looking at the kmer7 results. We can see that RUG413 is quite clearly the smallest genome, when compare the others. While quite bacillus is potentially the largest genome, similar to rug545. Because Rug413 is the smallest genome, smaller than E. coli this could indicate that it is a Harmful bacteria / pathogen.

## ORFS

The first item I looked at after reading the orfs from prodigal was the CDS. This is the coding sequence, CDS is the region of DNA or RNA whose sequence determines the sequence of amino acids in a protein. You can obtain the orfs using these command lines.

$ prodigal -i Rug384.fa -o Rug384.gff3 -f gff

$ bedtools getfasta -fi Rug384.fa -bed Rug384.gff3 -s -fo Rug384.orfs.fa

We can get the meta genomics like so

$./prodigal -i RUG384.fa -o RUG384meta.gff -f gff -p meta

$./bedtools getfasta -fi RUG384.fa -bed RUG384.gff -s -fo RUG384.orfs.fa

Looking at the RUG384 metagenomic sequence for the scaffold 88 example, there’s a lot of data we can see, just looking at the header alone we can see,

|  |  |  |
| --- | --- | --- |
| name | description | Rug384 example |
| Seqnum | An ordinal ID for this sequence, beginning at 1. | 1 |
| Seqlen | Number of bases in the sequence | 72170 |
| Seqhdr | The entire FASTA header line | Scaffold\_88 |
| Version | Version of Prodigal used to analyze this sequence | Prodigal.v2.6.3 |
| Run\_type | The type of mode used | Metagenomic |
| Model | Info about the preset training file, (anonymous only) | “33 |
| Gc\_cont | % GC content of the sequence. | 53.00 |
| Transl\_table | The genetic code used to analyze the sequence. | 11 |
| Uses\_sd: | 1 if it used its default RBS finder, 0 if scanned for other motifs | 1 |
| Gene name |  | Pelotomaculum\_thermopropionicum\_SI |

Looking in the first field we can see the start position of the protein and the end position of the protein for example in RUG384 metagenomic sequence for scaffold88 on the first line we can see that the start is on 1 and the stop is on 672 (the + means forward strand and the – means reverse strand).

The fields in the semicolon-delimited string are as follows with the examples from 384:

|  |  |  |
| --- | --- | --- |
| name | description | Rug384 example |
| ID | A unique identifier for each gene, consisting of the ordinal ID of the sequence | 1 |
| partial | An indicator of if a gene runs off the edge of a sequence or into a gap. A "0" indicates the | 10 |
| Start\_type | The sequence of the start codon (usually ATG, GTG, or TTG). If the gene has no start codon, this field will be labeled "Edge". | ATG |
| Stop\_type | The sequence of the stop codon (usually TAA, TGA, or TAG). If the gene has no stop codon, this field will be labeled "Edge". |  |
| rbs\_motif | RBS motif found (e.g. "AGGA" or "GGA") | AGxAGG/AGGxGG |
| rbs\_spacer | The number of bases between the start codon and the observed motif. | 5-10bp |
| gc\_cont | The GC content of the gene sequence. | 0.507 |
| gc\_skew | The GC skew of the gene sequence. |  |
| conf | A confidence score for this gene, representing the probability that this gene is real, i.e. 78.3% means Prodigal believes that gene is real 78.3% of the time and a false positive 21.7% of the time. | 100.00 |
| score | The total score for this gene. | 68.44 |
| cscore | The hexamer coding portion of the score, i.e. how much this gene looks like a true protein. | 59.73 |
| cscore | A score for the translation initiation site for this gene; it is the sum of the following three fields. | 8.71 |
| rscore | A score for the RBS motif of this gene. | 8.35 |
| uscore | A score for the sequence surrounding the start codon | -2.34 |
| tscore | A score for the start codon type (ATG vs. GTG vs. TTG vs. Nonstandard). | 2.70 |
| mscore | A score for the remaining signals (stop codon type and leading/lagging strand information). |  |

You can use a grep command to count how many CDS coding regions have been predicted for the file. For example Rug213 has 1645, Rug384 has 2756, Rug413 has 2038, Rug545 has 2114.

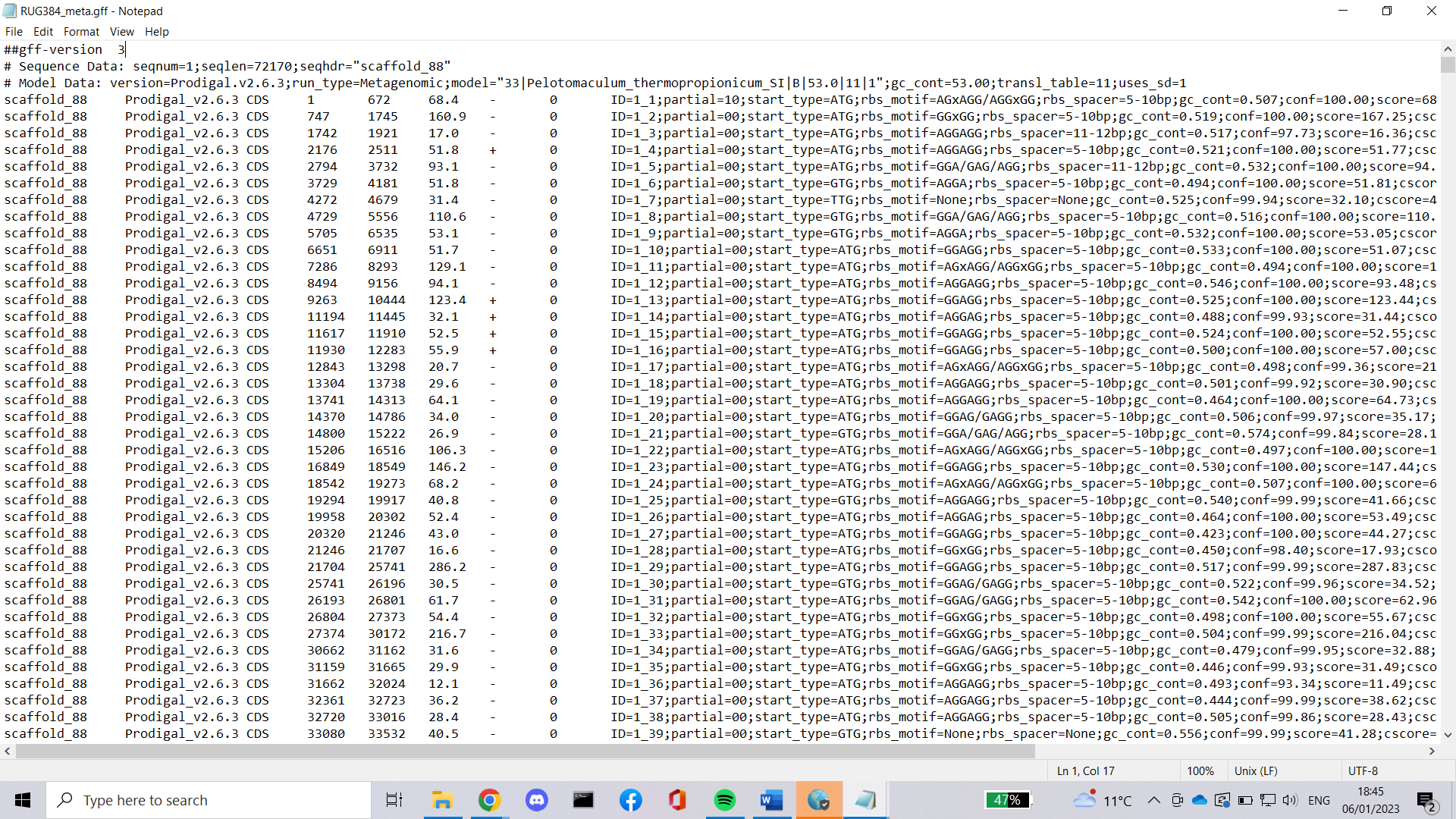
$grep -c “CDS RUG\*\*\*.gff3

|  |
| --- |
| Common gene predictions from the Rug384 metagenomic |
| Pelotomaculum\_thermopropionicum\_SI |
| Desulfotomaculum\_acetoxidans\_DSM\_771 |
| Chlorobium\_phaeobacteroides\_BS1 |
| Escherichia\_coli\_UMN026 |
| Bacteroides\_fragilis\_NCTC\_9343 |
| Marinobacter\_aquaeolei\_VT8 |
| Xenorhabdus\_nematophila\_ATCC\_19061 |
| Akkermansia\_muciniphila\_ATCC\_BAA\_835 |

|  |
| --- |
| Common gene predictions from the Rug213 metagenomic |
| Thermoplasma\_volcanium\_GSS1 |
| Desulfotomaculum\_acetoxidans\_DSM\_771 |
| Ignisphaera\_aggregans\_DSM\_17230 |
| Rickettsia\_conorii\_Malish\_7 |

|  |
| --- |
| Common gene predictions from the Rug413 metagenomic |
| Chlorobium\_tepidum\_TLS |
| Pelotomaculum\_thermopropionicum\_SI |
| Bacteroides\_fragilis\_NCTC\_9343 |
| Rothia\_dentocariosa\_ATCC\_17931 |
| Akkermansia\_muciniphila\_ATCC\_BAA\_835 |

|  |
| --- |
| Common gene predictions from the Rug545 metagenomic |
| Rickettsia\_conorii\_Malish\_7 |
| Thermoplasma\_volcanium\_GSS1 |
| Mycoplasma\_bovis\_PG45 |
| Desulfotomaculum\_acetoxidans\_DSM\_771 |
| Candidatus\_Amoebophilus\_asiaticus\_5a2 |



## PubMLST

PubMLST species identification. My first attempt at finding out what the species the genomes are, was to use PubMLST, Public databases for molecular typing and microbial genome diversity. PubMLST can be used to identify species by inputting fasta files. All of the genomes appear to be bacteria however only one was able to be identified. Compared to other species identifying tools pubmlst was most likely the worst however the results it provided were quick and interesting to see.

**Rug213**



**Rug384**

Rug384 was identified as Succiniclasticum ruminis. Succiniclasticum ruminis is an anaerobe, mesophilic, Gram-negative bacterium that forms circular colonies and was isolated from rumen of cow.

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generated

**Rug413 Rug545**

Graphical user interface, application, table

Description automatically generatedGraphical user interface, application, Word

Description automatically generated

## Sequence similarity

Using blastx and diamond, in order to use blast and diamond to find sequence similarity, I had to install a SwissProt database from Uniprot. Uniprot is a database of protein sequences and can be used to blast the genomes for the sequence alignment.

Blast first.

$ blastx -query orf.fa -db uniprot\_sprot –out outfile.txt

This command can be used to count how many hits occurred. Although because blastx would take too long to run, the number acquired don’t so how many hits there was total but how many I was able to get from the limited time using blastx.

$ wc -l outfile.csv

Rug213 got 1261915, Rug384 got 128823, Rug413 got 426463, Rug545 got 787348

I tried to blast the genomes however they were too large, and it would have taken too long so I didn’t get a fully complete blast report, this made me resort to using diamond instead since it is 20,000 times faster.

$./diamond blastx -d uniprotdia -q RUG384.fa -o diamond384.txt

Looking at the output file we can see a lot of data in this format shown with rug384 metagenomic data from the first line on the first scaffold, scaffold 88.

|  |  |  |
| --- | --- | --- |
| name | Description | Rug384 example |
| Query accession | accession of the sequence that was the search query against the database | scaffold\_88 |
| Target accession | accession of the target database sequence, that the query was aligned against | sp|Q97LC5|SYY1\_CLOAB |
| Sequence identity | The percentage of identical amino acid residues that were aligned against each other | 64.0 |
| Length | The total length of the local alignment | 342 |
| Mismatches | number of non-identical amino acid residues aligned against each other | 123 |
| Gap openings | Number of gap openings | 0 |
| Query start | starting of the local alignment in the query | 1778 |
| Query end | ending of the local alignment in the query | 753 |
| Target start | starting of the local alignment in the target | 65 |
| Target end | ending of the local alignment in the target | 406 |
| E-value | expected value of the hit quantifies the number of alignments | 8.98e-149 |
| Bit score | a scoring matrix independent measure of the (local) similarity of the two aligned sequences | 475 |

Graphical user interface, application, table, Excel

Description automatically generated

The reason why this data is important is because we can see information lie the **sequence identity** what fraction of the amino acids where identical between the query sequence rug384 and the database. The **bit score** is a scoring matrix, higher numbers mean more similar, so a higher score is better in the case of rug384 first line of scaffold 88 475 is strong meaning a high similarity. **E-value** indicated how likely that it is that the sequence would appear in random if it was run this way, these means that a very low e-value is better/more meaningful because it means that the sequence isn’t random and that the sequence has a high homology, in the case of rug384 first line of scaffold 88 8.98e-149 is a strong homology.

## Eggnog

EggNog (evolutionary genealogy of genes: Non-supervised Orthologous Groups) is database of biological information it allows thousands of genomes to be analysed at once to establish orthology relationships between their genes. I used eggnog on the metagenomic orfs I made using prodigal.

Eggnog outputs a wide variety of important information for example the file “out.emapper.decorated” appears to be in similar format to the prodigal output showing cds information And the “out.emapper HITS file” looks to be very similar to the diamond blastx format here we can see e-value and bitscores aswell as sequence identity. However the out.emapper orthologs file is the largest and most interesting file from the eggnog output since it contains a list of the species orthologs from the genome. The screen shot below shows the first few lines from the ortholog file from rug384.

Graphical user interface, text, application

Description automatically generated

# Results

All the results from this assignment are in folders which will be located in the zip file, exepct the kmer results for 21 and 61 since these files are too large.

# Discussion

Comparisons of bioinformatics tools use,

BlastX and Diamond. After annotating my genomes with Prodigal, I moved on to using the results for blast with the SwissProt database from Uniprot, blastx was however too slow to slow to calculate the statistical significance of the whole genome, However I then switched to using diamond, diamond was able to fully align the genome against the swissprot database extremely quickly. Blastx would have taken weeks. This is because Diamond is about 20,000 times faster than blastx however diamond will only report about 80-90% of all matches that blastx finds.

Prodigal, using prodigal to annotate my genomes was ever fast and given the size of the prodigal binary being small, makes it very lightweight. Like other bioinformatic tools prodigal has different modes that can be used. Normal, Anonymous/prediction and Training. For the purpose of this assignment, I used normal mode to learn properties and predict genes. Anonymous/prediction applies pre-calculated training files to the provided input sequence and predict genes based on the best results. Training mode is the same as normal but saves a training file for future use. When compared with other gene annotation tools appears to be one of the fastest however isn’t as sensitive when compared to the likes of RNAmmer and MetageneAnnotator. However, prodigal has greater sensitivity when identifying existing genes accurately.

Bedtools, I used bedtools straight after using prodigal, I used bedtools so that I could extract the sequences of the coding regions as a FASTA file from the GFF3 files. Bedtools is a command line tool that allows you to run a wide range of genomics analysis tasks on genomic files.

Jellyfish, Jellyfish is a tool used to count mer’s(kmers). It’s a fast and memory efficient counting tool. It was the fasted tool that I used for this assignment. Like all of the tools used in this asigment it’s used in the command line and takes dna sequences from fasta files and produces its kmer output. When compared to other kmer counting tools jellyfish is memory efficient however uses the most ram than any other tool, jellyfish is also perhaps one of the slowest when given smaller genomes, only faster when compared to Tallymer, however when given larger genomes jellyfish becomes quite a lot faster in comparison.

# Conclusions

Its interesting to see that e. coli was found in rug384 which is also the same genome that pubmlst identified Succiniclasticum ruminis which is a bacterium that forms circular colonies and was isolated from rumen of cow. Bacteroides fragilis was found and high levels of this may result from reduced digestive capacity or constipation. Akkermansia Muciniphila was found, this is commonly found in the human gut from faecal matter however I believe in this case is has come from the gut of a cow. Desulfotomaculum acetoxidans was found which is a bacteria found in soil this means it could have come from the soil/grass the cow is eating or from the soil where the faecal matter was found, likewise Xenorhabdus nematophila is a pathogen of insects and most likely came from the grass/soil. This is interesting because e. coli is found in cows rumens as well, mostly in the colon but also in the rumen, this suggests that the rug384 may have come from a cow, most likely the rumen from the cows’ faecal matter found in a field.

It was also interesting the see the sizes of the genomes Rug413 is the smallest genome, smaller than E. coli this could indicate that it is a Harmful bacteria / pathogen. Similarly, to Rug384 Bacteroides fragilis was found and the potentially harmful/negative effect of Bacteroides fragilis further supports the idea that rug413 is a harmful genome. Akkermansia Muciniphila was found, this is commonly found in the human gut from faecal matter, unlike rug384 I believe that is has come from human faecal matter, together with the Bacteroides fragilis supporting that this genome is harmful and could potentially have cause diarrhea.

In rug213 Desulfotomaculumacetoxidans, Thermoplasma volcanium and Ignisphaera aggregans were found these together might suggest that rug213 was found from the soil in a volcanic area since Ignisphaera aggregans is a strictly anaerobic, moderately acidophilic, heterotrophic hyperthermophilic and fermentative archaeon isolated from a near neutral, boiling spring, Thermoplasma volcanium functions as a facultative anaerobic chemoorganoheterotroph that is also capable of lithotrophic metabolism through anaerobic sulfur respiration. And finally Desulfotomaculum acetoxidans was one of the first sulfate-reducing bacteria known to grow with acetate as sole energy and carbon source. These all suggest that it came from a volcanic area with high amounts of sulfur, carbon.

In rug545, there are similar genes found suggesting that it comes from a similar volcanic location, however Mycoplasma Bovis was found and this is a species known to infect cattle similarly Candidatus Amoebophilus asiaticus was found which plays a role in symbiose host cell interaction. From this we could suggest that rug545 came from a cow living in a potentially volcanic area that might have once been dangerous but is now safe for farmland.

# Acknowledgements

For this assignment I used various bioinformatic tools like prodigal, bedtools, blastx, diamond, eggnog, jellyfish. These were detrimental to this assignment; I also use the National Center for Biotechnology Information to gather information since this was a very information rich site.

# References

Hyattpd. Understanding the prodigal output · hyattpd/prodigal wiki. Retrieved January 8, 2023, from https://github.com/hyattpd/prodigal/wiki/understanding-the-prodigal-output#gene-coordinates

Buchfink. 1. tutorial · bbuchfink/diamond wiki. Retrieved January 8, 2023, from https://github.com/bbuchfink/diamond/wiki/1.-Tutorial

(Vejandla, *COMPARATIVE ANALYSIS OF KMER COUNTING AND ESTIMATION TOOLS* 2021 from <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1243&context=computerscidiss>)

Software Radius. (2022, December 30). 10 best gene/genome annotation tools & software in 2023. Retrieved January 8, 2023, from https://www.softwareradius.com/best-gene-and-genome-annotation-tools-and-software/

BacDive. Succiniclasticum ruminis SE10: Type strain: DSM 9236: Bacdiveid:103. Retrieved January 8, 2023, from https://bacdive.dsmz.de/strain/103

National Center for Biotechnology Information, Retrieved January 8, 2023, from https://www.ncbi.nlm.nih.gov/