Solutions for Assignment 8: *Metagenomics*

Question 1:
How many total ORFs were predicted identified using the described method?
68 (40 reverse and 28 forward)

Question 2:
Explain what all of the flags for the script 'gmhmmp' are doing.
`-m` specifies the model file (given)
`-o` specifies the output file that you want to name
`-a` outputs the protein sequence of the predicted ORFs
`-d` outputs the nucleotide sequences of the predicted ORFs
`-f G` specifies the output should be in GFF format

Would you have added or dropped any flags for your particular problem or use?
Any answer with a reasonable explanation was acceptable. Common extra flags were -k, and -s.

Question 3:
How many ORFs were predicted using MetaGeneMark?
19

Question 4:
How many ORFs were identified by both MetaGeneMark and your custom ORF caller? How many ORFs were unique to your custom ORF caller? How many ORFs were unique to MetaGeneMark? Please interpret your results and explain why the intersection and set differences are large/small.
There were 8 sequences common to both MetaGeneMark and the custom ORF caller
There were 60 sequences unique to the custom ORF caller
There were 11 sequences unique to MetaGeneMark

NR: Many are unique to the custom ORF caller because it called many more ORFs than MGM (68 vs. 19). The custom ORF is based on a very simple model (as opposed to MGM's HMM-based method) and thus probably contains a lot of false positives. Also, it reports only the longest ORFs, which may actually result in functional proteins. It is encouraging that the the custom ORF caller found ~50% of the same ORFs as MGM. The ORFs that are unique to MGM were not called by the custom ORF caller because they don't start with M or end with a stop codon.

Question 5:
Explain briefly (in a few words) what each one of the parameters mean in:
`qseqid` means Query Seq-id
`sseqid` means Subject Seq-id
`pident` means Percentage of identical matches

`length` means Alignment length
`mismatch` means Number of mismatches
`gapopen` means Number of gap openings
`qstart` means Start of alignment in query
`qend` means End of alignment in query
`sstart` means Start of alignment in subject
`send` means End of alignment in subject
`evalue` means Expect value
`bitscore` means Bit score
`slen` means Subject sequence length
`stitle` means Subject Title
Optional:
`-db` specifies the BLAST database name
`-query` species the input filename
`-out` species the output filename
`-outfmt` species the format of the output file

Question 6:
How many antibiotic resistance genes were identified using BLAST against the CARD database? How many survive the filtering in your Python script?

# of BLAST hits: 493
# of unique AR genes identified using BLAST: 266

# of BLAST hits that survive the filters of doom: 3
# of unique AR genes identified using BLAST that survive the filters of doom: 3

Question 7:
Mention two uses of HMMER, and how you would go about executing it. You could access the documentation here: HMMER userguide.

1. Searching a sequence database with a single profile HMM

One common use of HMMER is to search a sequence database for homologues of a protein family of interest. You need a multiple sequence alignment of the sequence family you're interested in. (Profile HMMs can be trained from unaligned sequences; however, this functionality is temporarily withdrawn from HMMER. We recommend CLUSTALW as an excellent, freely available multiple sequence alignment program.)

2. Searching a query sequence against a profile HMM database

A second use of HMMER is to look for known domains in a query sequence, by searching a single sequence against a library of HMMs. (Contrast the previous section, in which we searched a single HMM against a sequence database.) To do this, you need a library of

profile HMMs. One such library is our PFAM database [Sonnhammer et al., 1997,Sonnhammer et al., 1998], and you can also create your own.

Question 8:
How many antibiotic resistance genes were annotated by Resfams? How does this number compare with the number of antibiotic resistance genes identified using BLAST (question 5)?

There are 7 total genes in Resfams of which **4 are unique** and 266 unique genes based on sseqid from the BLAST output. When looking at genes that survive the filters there are 3 of these. Thus Resfams seems to be finding more resistance genes than BLAST. Further analysis is needed to determine if the two methods predict the same or different antibiotic resistance genes.

Acceptable answers could include filtering or not, so long as the explanation matched the numbers given.