Annotating the functions in the bins using OrfM and the SEED

Another way to annotate the metagenomes is to identify all the open reading frames (i.e. stretches of DNA that start at a translational start site (typically: ATG), and end at a translational stop site (typically: TAA, TGA, TAG). In complete genomics, there are heuristics used to filter those open reading frames to ensure that we have the open reading frames are most likely to be the real protein encoding genes. For example, if you have two overlapping open reading frames it is much more likely that the longer sequence is a real protein encoding gene and the shorter sequence is less likely to be a real protein encoding gene.

For metagenomics, however, we can identify all the open reading frames, and then compare them to the databases and suggest that the open reading frames that have real matches are real, while the others are potentially not real open reading frames, and we can ignore them.

A quick way to identify all the open reading frames is to use orfM (64), a very fast open reading frame extractor. This just takes a fasta file as its argument, and generates a fasta file with amino acids in it:

```
orfm contigs.fna  > orfs.faa
```

Notice that here our DNA file has the standard extension `.fna` for "fasta nucleic acids" (i.e. DNA) and our output file has the standard extension `.faa` for "fasta amino acids" (i.e. proteins). These may or may not be recognized by your operating system though!

Note also, that we are redirecting the output to the `orfs.faa` file as we discussed with Linux

There are lots of ways that you can compare these sequences to databases. One way is to copy and paste a few of the sequences into the BLAST web interface. However, you can only do a few of the amino acid sequences at a time.

Another way to annotate the sequences is using the SEED server framework (65). We have included that on the virtual box image. You can rapidly assign functions to proteins using the command:

```
svr_assign_using_figfams < orfs.faa > orfs.fn
```

Here, we are redirecting the input from `orfs.faa` and redirecting the output to `orfs.fn`.

This will give you a function for those sequences that we can find a "reliable" match to. By default, a "reliable" match means that there are three 7-mer amino acid matches between your protein sequence and our database, however you can change that reliability to get more hits. For more information, run `svr_assign_using_figfams` with a `-h` flag to see more options.

It is worth taking a few of those sequences, to which you have a match from the SEED servers and running blast using the NCBI website above and exploring those hits.

- Does the SEED annotation make sense?
- Does the NCBI annotation make sense?
- Which organisms are matched?