Activities Sheet for Afternoon Session Tuesday, November 5, 2019

BIN420 Functional Metagenomics Course

The InterProScan annotation files are located in the following folder on the CIGENE server:

/mnt/project/Courses/BIN420-2019/Day2_GenesAndAnnotations/iprscan/

1. Instead of running everything on the head node, let's initiate an interactive session with the srun command, allocating 6000 Mb of memory and 1 core:

srun --mem=6G -n1 -N1 --pty --preserve-env $SHELL

Let's take a closer look at the output from the interproscan run, the "proteins_nostops.gff3" file:

head proteins_nostops.gff3

tail proteins_nostops.gff3

What is the format of the file at the beginning? How does the file look at the end?

There's a line conisting of "##FASTA" that denotes the transition from the table to the fasta format section. You may see this portion of the file with the following grep search, listing 10 lines before the match and 10 lines after:

grep "##FASTA" proteins_nostops.gff3 -A10 -B10

What are the different columns of the table at the beginning of the gff3 format file? See the wikipedia page for an explanation: https://en.wikipedia.org/wiki/General_feature_format

The dataset we are using for the course is derived from a microbial community enriched by addition of cellulose. So, let's start off by looking for cellulose degradation genes such as "cellulase":

grep "cellulase" proteins_nostops.gff3

There's 1 hit. In the last column of the gff table (the attributes column), there's portion that says "Name=cd04080". Search "cd04080" in the NCBI conserved domains database at:

https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

What is the function of this conserved domain?

Let's find out to what bin this contig belongs. The contig name appears in the first column, or sequence column, of the tabular part of the gff file. It's name is "NODE_32_length_88662_cov_470.213076_68", but the "_68" was added to the contig name later, presumably by Prodigal to denote multiple open reading frames that can occur on a single, long contig. Hence, search the genome bins for the contig ID as follows:

```
grep "NODE_32_length_88662_cov_470.213076"
../../Day1_AssemblyAndBinning/binning/Bins_metaBat2_from_Ben/*.fa
```

The output should indicate which bin contigs file had a match for the contig ID. Which bin contigs file was it?

Upload this bin contigs file to the online Microbial Genomes Atlas (MiGA) as a population genome. First, go to the Microbial Genomes Atlas (MiGA) at http://www.microbial-genomes.org/ and select “NCBI Prokaryotes”. Give your search a name, select the type of dataset as “Popgenome…” and type of input as “Assembly in FastA format”. Upload the bin contigs file from the genome bin and submit it. Be sure to click “Execute MyTaxa Scan analysis” when the new page comes up showing that it is currently running. The analysis may take hours or overnight to complete, so it would be best if you obtain a login for their site first so that you can return to a saved copy of your analysis results.

Check back in a few hours or the next day once the analysis has completed online. What is the nearest relative identified in the public databases for this genome bin? How confident may one be in this match?

Try a literature search for the name of this organism and "cellulose degradation". Is anything known about its ability to degrade cellulose?

Does the organism have other synonymous names? To find out, search it's name in the NCBI taxonomy database:

https://www.ncbi.nlm.nih.gov/taxonomy

You may find details about this organism in the literature under prior names for the organism.

What else is co-located on this contig with the putative cellulase gene? You can explore this with the following command:

```
grep "NODE_32_length_88662_cov_470.213076" proteins_nostops.gff3
```

Let's copy this output to a file and thereafter open it in a spreadsheet so that it's better organized:

grep "NODE_32_length_88662_cov_470.213076" proteins_nostops.gff3 > ~/cellulase_contig_all_domains.tsv

How many open reading frames (ORF) appear to be on this contig?

Do you notice multiple, redundant annotations for a given ORF? Why are they present in the interproscan output? Do the redundant annotations overlap?

Let's look at just the cellulase ORF:

grep "NODE_32_length_88662_cov_470.213076_68" proteins_nostops.gff3

How many annotations are there for this ORF? Which annotations overlap, and which indicate distinct domain regions within the ORF?

Let's try a different approach with this contig. Let's extract the contig sequence from the original assembly contigs file and upload it to an online tool for annotation of carbohydrate active enzymes. To extract the contig sequence, we'll use the samtools command "faidx", which creates an index of the fasta file for rapid searching of individual fasta sequences in a multi-sequence fasta file. First though, let's copy the assembly contigs file to our home directory where the index file may be created alongside it:

cp ../../Day1_AssemblyAndBinning/assembly/contigs.fasta ~

Now let's use samtools to simultaneously create the index and extract the contig, named "NODE_32_length_88662_cov_470.213076 ", and save it to a new output file named "cellulase_contig.fasta" in our home directory:

samtools faidx ~/contigs.fasta NODE_32_length_88662_cov_470.213076 > ~/cellulase_contig.fasta

Copy the "cellulase_contig.fasta" file to your laptop, and then upload the contig to the annotation tool for the dbCAN database, a database of carbohydrate active enzymes:

http://bcb.unl.edu/dbCAN2/blast.php

Explore the output. Can you tell from the output whether some of these genes are co-located close to one another?

The attributes column for the cellulase also contains the annotation "signature_desc=CBM6_cellulase-like". In this context, CBM is an abbreviation for Carbohydrate Binding Module. Let's perform a broader search for "CBM":

grep "CBM" proteins_nostops.gff3

However, there are many lines that match "CBM". To make more sense of the hits, let's save the output to a Tab Separated Values (TSV) file in our home directory on the CIGENE server:

grep "CBM" proteins_nostops.gff3 > ~/CBM_output.tsv

Now, copy the "CBM_output.tsv" file to your laptop and open it with a spreadsheet program with tab as the separator or delimiter. There should be 9 columns.

2. Note that some proteins were annotated with the word hypothetical:

grep "hypothetical" PROKKA_07272017.ffn | wc -l

What does "conserved hypothetical protein" mean in this context?

How many lines of the gff file have "DUF" present in them?

grep "DUF" proteins_nostops.gff3 | wc -l

What does "DUF" refer to in this context?

You can see that there are many DUFs in Pfam:

http://pfam.xfam.org/family/browse?browse=d

3. You may try searching some proteins with domains annotated as “hypothetical” or "DUF" on the Phyre2 Protein Fold Recognition Server website (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). It may take an hour or so for the server to process the sequence and generate results.

The following sequence has a DUF, although it also has a number of well-defined domains:

grep ">NODE_32_length_88662_cov_470.213076_73" /mnt/project/Courses/BIN420-2019/Day2_GenesAndAnnotations/prodigal/proteins_nostops.faa -A 15

Search this sequence on the Phyre2 web server. What does the protein function appear to be? How high is the confidence, and over what extent of the protein sequence?

4. You can download the HMM for McrA from the FunGene pipeline and repository at http://fungene.cme.msu.edu/hmm_download.spr?hmm_id=16

However, we have already placed this on the server for you, and you may copy it to your home directory with the following command:

cp /mnt/project/Courses/BIN420/d2_function/mcrA.hmm ~

Now, use the HMM to search for McrA among the protein sequences:

module load hmmer

hmmsearch ~/mcrA.hmm /mnt/project/Courses/BIN420-2019/Day2_GenesAndAnnotations/prodigal/proteins.faa

How many McrA hits do you find? How long is McrA, and how long are the matches from the HMM search in comparison? Why the discrepancy (hint: look at the read coverages, or "cov" in the contig identifiers). Does it appear that the mcrA gene assembled well?

5. Next, let's use hmm models from Pfam to perform a search. Note that while one can download the entire Pfam set of HMMs and extract a desired HMM using the tool hmmer, the full set of HMMs is several hundred megabytes in size (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz). Thus, if there's only one Pfam you wish to search, one can download it directly from the website. Given that one of the genome bins corresponds to an organism that is known for its ability to produce ethanol, we'll use the Pfam HMM for an alcohol dehydrogenase. The Pfam number is PF00465. Go to the webpage for this Pfam:

http://pfam.xfam.org/family/PF00465

and then click on the tab "Curation and model" on the lefthand side of the webpage. Then, at the bottom of the page, you will see a download link to "download the raw HMM for this family". You may copy this link into a wget command to download the hmm to your home directory:

wget -O ~/PF00465.hmm http://pfam.xfam.org/family/PF00465/hmm

Next, perform the HMM search with the model:

module load hmmer

hmmsearch ~/PF00465.hmm /mnt/project/Courses/BIN420-2019/Day2_GenesAndAnnotations/prodigal/proteins_nostops.faa

How do you interpret the output? Does the length of the matching region vary as one descends through the list of hits?

Check the following hit from the HMMer search against the contigs in the genome bins to see which bin it matches:

grep "NODE_19_length_154378_cov_445.119399" /mnt/project/Courses/BIN420-2019/Day1_AssemblyAndBinning/binning/Bins_metaBat2_from_Ben/*.fa

6. Now try a search of multiple proteins at once using a collection of hidden markov models specific to carbohydrate active enzymes. Begin by copying the dbCAN database and parser script to your home directory:

cp /mnt/project/Courses/BIN420/d2_function/dbCAN-fam-HMMs.txt ~

cp /mnt/project/Courses/BIN420/d2_function/hmmscan-parser.sh ~

Note that these files are originally available at http://bcb.unl.edu/dbCAN2/download/

Next, make the database files:

hmmpress ~/dbCAN-fam-HMMs.txt

You should see the following output:

Working... done.

Pressed and indexed 360 HMMs (360 names and 1 accessions).

Models pressed into binary file: dbCAN-fam-HMMs.txt.h3m

SSI index for binary model file: dbCAN-fam-HMMs.txt.h3i

Profiles (MSV part) pressed into: dbCAN-fam-HMMs.txt.h3f

Profiles (remainder) pressed into: dbCAN-fam-HMMs.txt.h3p

Now perform the search (note that we use the miniature set to save time as this step can take a while):

hmmscan --domtblout ~/dbCAN.out.dm ~/dbCAN-fam-HMMs.txt /mnt/project/Courses/BIN420-2019/Day2_GenesAndAnnotations/prodigal/proteins_nostops_mini.faa > ~/dbCAN_hmmscan.out

Now parse the output into a table with dbCAN's custom script:

sh ~/hmmscan-parser.sh ~/dbCAN.out.dm > ~/dbCAN.out.dm.parsed.tsv

Check first few lines of the file:

head ~/dbCAN.out.dm.parsed.tsv

Which gives the following output (which is significantly reduced compared to the full dataset):

```
GH23.hmm 135      NODE_1_length_571491_cov_39.794040_16      644      2.8e-29  15       123      511      635
         0.8

GT2.hmm  168      NODE_1_length_571491_cov_39.794040_61      795      1.3e-13  3        166      212      392
         0.970238095238095
```

How do you interpret this output? Which HMMs had matches in the miniature protein dataset? What protein functions do these HMMs represent? Do the protein families represented by each HMM encompass a single, or rather multiple, functions?