

Introduction to protein annotation with Hidden Markov Models

Frank Aylward

Abstract

Here is a general tutorial on how to begin annotating proteins with Hidden Markov Models. A small test set of HMMs is provided in the Git repo downloaded in the first step.

Citation: Frank Aylward Introduction to protein annotation with Hidden Markov Models. **protocols.io**

dx.doi.org/10.17504/protocols.io.pijdkcn

Published: 17 Apr 2018

Protocol

Get the data

Step 1.

Let's clone some data files from my Github repo to get started.

```
git clone https://github.com/faylward/hmm_tutorial
```

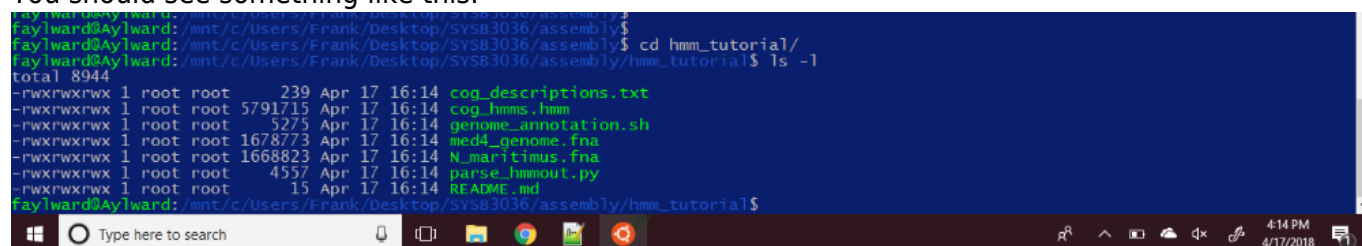
Now let's go inside the new folder:

```
cd hmm_tutorial
```

And let's see what's inside:

```
ls -l
```

You should see something like this:



```
faylward@Aylward: /mnt/c/Users/Frank/Desktop/SYSB3036/assembly$  
faylward@Aylward: /mnt/c/Users/Frank/Desktop/SYSB3036/assembly$ cd hmm_tutorial/  
faylward@Aylward: /mnt/c/Users/Frank/Desktop/SYSB3036/assembly/hmm_tutorial$ ls -l  
total 8944  
-rw-rw-rw- 1 root root 239 Apr 17 16:14 cog_descriptions.txt  
-rw-rw-rw- 1 root root 5791715 Apr 17 16:14 cog_hmms.hmm  
-rw-rw-rw- 1 root root 5275 Apr 17 16:14 genome_annotation.sh  
-rw-rw-rw- 1 root root 1678773 Apr 17 16:14 med4_genome.fna  
-rw-rw-rw- 1 root root 1668823 Apr 17 16:14 N_maritimus.fna  
-rw-rw-rw- 1 root root 4557 Apr 17 16:14 parse_hmmout.py  
-rw-rw-rw- 1 root root 15 Apr 17 16:14 README.md  
faylward@Aylward: /mnt/c/Users/Frank/Desktop/SYSB3036/assembly/hmm_tutorial$
```

Predict genes

Step 2.

Now we have some genome files (.fna) but we need to get some proteins to begin annotating. To do this let's use Prodigal.

Prodigal will predict genes from chromosomes (or contigs), translate those genes into amino acids, and produce annotation summary files (such as "gene feature format", or gff, files), depending on what options you use.

```
prodigal -i med4_genome.fna -a med4.proteins.faa -d med4.genes.fna -f gff -o med4.prodigal.gff
```

Protein prediction QC

Step 3.

Always good to check the files we just created to make sure we know what's inside. Let's use seqtk to take a quick look at what's inside the .faa files.

```
seqtk comp med4.proteins.faa | head
```

And how many proteins total were predicted?

```
seqtk comp med4.proteins.faa | wc
```

Now query the proteins against the HMMs that we have in the folder

Step 4.

Now run the HMMER command. Note that the last two arguments are "positional arguments" since they do not have flags in front of them. The .hmm file and the query protein file must always be provided at the end, in that order.

```
hmmsearch --tblout med4.hmmout -o med4.output cog_hmms.hmm med4.proteins.faa
```

Parse the output to get the best matches for the query proteins

Step 5.

The main tabulated output we want is in med4.hmmout. Unfortunately the authors of HMMER made the output space-delimited, so it's a bit hard to look at or put in an Excel spreadsheet.

I made a small Python script that will parse through this output, pull out the best hit for each query protein, and put it in a tab-delimited output.

```
python parse_hmmout.py med4.hmmout > med4.hmmout.parsed
```

Play around with the parameters and employ cutoffs to ensure good matches

Step 6.

Now you may notice some hits that have very low bit scores. This is because we did not use any quality cutoffs when we ran HMMER. Just like with BLAST, there is an e-value cutoff option that we can use.

For that we can use the following command:

```
hmmsearch -E 1e-10 --tblout med4.hmmout -o med4.output cog_hmms.hmm med4.proteins.faa
```

Now try again with another genome

Step 7.

Now let's practice again with another genome and see what we get. A new genome is in the file N_maritimus.fna. This is an Archaea called Nitrosopumilis maritimus, an abundant ammonia-oxidizing microbe in the ocean.

Here is the overall workflow:

```
prodigal -i N_maritimus.fna -a N_maritimus.faa -f gff -o N_maritimus.gff
```

```
hmmsearch -E 1e-10 --tblout N_maritimus.hmmout -o N_maritimus.output cog_hmms.hmm
```

```
N_maritimus.faa
```

```
python parse_hmmout.py N_maritimus.hmmout > N_maritimus.hmmout.parsed
```

what functional genes are present here that are not present in Prochlorococcus? Which genes are present in both?