SYSB 3036 W05 Introduction to Hidden Markov Models

Frank Aylward[1]

[1]Virginia Tech

Feb 28, 2019

Working

dx.doi.org/10.17504/protocols.io.yhqft5w

Frank Aylward
Virginia Tech

PROTOCOL STATUS

Working
We use this protocol in our group and it is working

1 First let's download some data from GitHub:

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

and then

```
cd hmm_introduction
```

2 There should be two files there, nifh.faa.gz and nosz.faa.gz.
We will use the nifh.faa.gz file today, and the nosz.faa.gz file will be for your homework.

The nifh.faa.gz file contains proteins that belong to the NifH family. NifH is a core component of the nitrogenase enzyme, which fixes atmospheric nitrogen into ammonia. It is a key component of the global nitrogen cycle.

Let's unzip the nifh.faa.gz file

```
gunzip nifh.faa.gz
```

And now let's get some stats on how many sequences are in this file, and how long they are:

```
seqkit fx2tab -inl nifh.faa
```

3 For starters we are going to create a global alignment of these NifH proteins using Clustal Omega, a popular alignment program. Because it is nice to visualize these things we will go to the main webpage and use the web interface:

https://www.ebi.ac.uk/Tools/msa/clustalo/

In the window you can paste the contents of the nifh.faa file (you can open the file in a text editor to copy the contents, or you can use the "more" command, scroll through all of the contents, and then copy the sequences directly from the command line if you prefer).

Click the "submit" button and wait for the results to appear. You should eventually see a multi-sequence alignment.

4 Click on the "send to mview" button to export the alignment to the MView tool. This will take you to a new submission website, where all you should need to do is click "submit" again. You should eventually see a multi-sequence alignment with statistics for base conservation on the bottom.

5 Now we want to do something similar in the command line.

For alignment we can still use Clustal Omega:

```
clustalo -i nifh.faa -o nifh.aln
```

And now let's take a look at the file:

```
more nifh.aln
```

Note that the alignment file is still in FASTA format, though now there are "-" characters to signify gaps. FASTA format is generally used for alignments, though for visualization tools like MView are nicer since you can see all of the aligned regions more clearly.

Let's see how long these sequences are:

```
seqkit fx2tab -iln nifh.aln
```

Note that all of the sequences are the same length now. This should always be the case for an alignment, since gap characters should effectively lengthen shorter sequences.

6 Now let's create an HMM from the multi-sequence alignment of NifH proteins. For this we can use the hmmbuild command in HMMER.

```
hmmbuild nifh.hmm nifh.aln
```

You should see the nifh.hmm file now. Take a look with "more".

7 Now we want to test this new nifh.hmm out to see how well it predicts NifH proteins.
For this we will download the genome of Methanococcus vinelandii

```
wget -O methanococcus.faa.gz ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/017/165/GCF_000017165.1_ASM1716v1/GCF_000017165.1_ASM1716v1_protein.faa.gz
```

and unzip it:

```
gunzip methanococcus.faa.gz
```

8 To compare a protein file to a HMM we can use the hmmsearch command in HMMER.

```
hmmsearch nifh.hmm methanococcus.faa > hmmout.txt
```

Or, if we want a tabulated output and introduce an E-value threshold,

```
hmmsearch --tblout hmm_table.txt  -E 1e-10  nifh.hmm  methanococcus.faa > hmmout.txt
```

We can browse the results with "more". There are two proteins with very good matches to our HMM.

9 Now let's compare this annotation to one that we would do with BLASTP.
First we need to make a BLASTP database from the nifh.faa file.

```
makeblastdb -in nifh.faa -dbtype prot
```

And then we can compare all of the Methanococcus proteins to this BLAST database.

```
blastp -query methanococcus.faa -db nifh.faa -evalue 1e-5 -outfmt 6 > blastpout.txt
```

Here the annotations are not as clear as with hmmsearch. Note that some of the hits that we are getting for our query proteins have very low % identity, so it is difficult to assess with BLASTP alone whether or not the proteins belong to the same family.

10