HMM analyses

LJM

2019-11-15

Contents

Chapter 1

Find annotated viruses in the `.emapper.annotations` files

Search all viruses from eggnogg mapper annotations and store them in `viruses.emapper.annotations`-file. The goal is to use the identifiers in this file to fetch out protein sequences of all these viruses.

One possible way to go might be to create profile HMMs of some clear groups of viruses, e.g. all *Myoviridae* are used to create a *Myoviridae* profile, all *Siphoviridae* another profile, etc. and scan the fOTUs with these HMM profiles to classify them.

1.1 Gather all eggnogg mapper annotated viruses in one file

```
ANNOT="../data/annotations/"
ANALYSES="../analyses/HMM_scan/"

grep -E 'Viruses' "$ANNOT"*.emapper.annotations > "$ANALYSES"viruses.emapper.annot
head -n 2 "$ANALYSES"viruses.emapper.annot
```

1.2 Survey what categories of viruses are there

Now that we have all the viruses found, it would be good to try to survey what type of categories of viruses are there so that they can be used to create the profiles. Maybe one way to see the categories is the following:

```
ANNOT="../data/annotations/"
ANALYSES="../analyses/HMM_scan/"

awk '{print $5}' "$ANALYSES"viruses.emapper.annot | sort | uniq
```

We thus see that there are 9 different categories of viruses with varying specificity.

1.3 Gather and group all viral protein sequences

Next task is to gather groupwise all viral protein sequences. Maybe one simple way to go about doing that is by pattern matching the 5th field in the previously produced `viruses.emapper.annot` file with the found categories and saving each category in own file.

```
INPUT="../analyses/HMM_scan/viruses.emapper.annot"
SCRIPT="../scripts/virusSeparator.awk"
RESULTS_DIR="../analyses/HMM_scan/virus_annotations_categorised/"
cat "$INPUT" | awk -v output_dir="$RESULTS_DIR" -f "$SCRIPT"
```

This seems to be possible by searching the correct file indexing it with exonerate v. 2.4.0 (Slater and Birney 2005) command `fastaindex` and then fetching it with another tool in the exonerate utility suite called `fastafetch`. Lastly each of the fetched files are appended to a new file with hopefully most of the annotation data not lost (but contained in the fasta header).

```
#ml bioinfo-tools exonerate/2.4.0
VIRUS_ANNOT="../analyses/HMM_scan/virus_annotations_categorised/"
PROTEOMS="../data/proteoms/"
SCRIPT="../scripts/virusProteomGatherer.awk"
PROTEOM_INDEXS="../analyses/HMM_scan/proteom_indexes/"
VIRUS_PROTEOMS="../analyses/HMM_scan/virus_proteoms/"

read -r -a ANNOTs <<< $( find $VIRUS_ANNOT -name "*.annot" -and -type f -print0 | xargs -0 e
for ANNOT in "${ANNOTs[@]}"; do
  FILE=$(echo $(basename "$ANNOT"))
  VIR_CATEGORY=$(echo $(basename "$ANNOT") | awk -F "." '{print $1}')
  cat "$VIRUS_ANNOT""$FILE" | awk -v proteom_dir="$PROTEOMS" -v proteom_index_dir="$PROTEOM_
done
```

Chapter 2

Prepare for finding viral sequences in fOTUs

The tool to be used in the HMM analyses is called HMMer v. 3.2.1 ("HMMER v. 3.2.1 (June 2018)" n.d.). It requires multiple sequence alignments (MSA) as a source for building a profile HMM. Let's now use MAFFT v7.407 (2018/Jul/23) (Katoh and Standley 2013) to build MSAs from each of the viral multifasta files.

2.1 Build MSAs of same category of viral proteins

```
VIRUS_PROTEOMS="../analyses/HMM_scan/virus_proteoms/"
PROTEOM_MSAs="../analyses/HMM_scan/virus_MSAs/"

read -r -a mFASTAs <<< $( find $VIRUS_PROTEOMS -name "*.faa" -and -type f -print0 | xargs -0
for mFASTA in "${mFASTAs[@]}"; do
  FILE=$(echo $(basename "$mFASTA"))
  VIR_CATEGORY=$(echo $(basename "$mFASTA") | awk -F "." '{print $1}')
  mafft --auto --thread 4 "$VIRUS_PROTEOMS""$FILE" > "$PROTEOM_MSAs""$VIR_CATEGORY"".aln.faa
done
```

2.2 Build profile HMMs from the viral MSAs

```
PROTEOM_MSAs="../analyses/HMM_scan/virus_MSAs/"
VIRUS_HMMs="../analyses/HMM_scan/virus_HMMs/"

read -r -a MSAs <<< $( find $PROTEOM_MSAs -name "*.faa" -and -type f -print0 | xargs -0 echo
for MSA in "${MSAs[@]}"; do
  FILE=$(echo $(basename "$MSA"))
  VIR_CATEGORY=$(echo $(basename "$MSA") | awk -F "." '{print $1}')
  #echo "$FILE $VIR_CATEGORY"
  hmmbuild --informat afa --amino --cpu 4 -o "$VIRUS_HMMs""$VIR_CATEGORY"".out" -O "$VIRUS_F
done
#cut -f 2- "$ANALYSES"viruses.emapper.annotations | head -n 3
```

2.3 Create target protein multifasta files

The target protein multifasta files for searching with the built HMM profiles should contain all the protein sequences belonging to each fOTU. The file defining which bins belong to which fOTU are defined in file `../data/fOTUs.csv`.

Let's now create ~ 2800 multifasta files corresponding to each fOTU.

```
fOTUs="../data/fOTUs.csv"
PROTEOMS="../data/proteoms/"
SCRIPT="../scripts/fOTUproteomGatherer.awk"
fOTUproteoms_DIR="../analyses/fOTU_proteomes/"
cat "$fOTUs" | awk -F, -v proteom_dir="$PROTEOMS" -v fOTUproteoms="$fOTUproteoms_DIR" -f "$S
```

Chapter 3

Identify viral content in fOTUs with a HMM profile

The identification with the 9 HMM profiles can be executed in the following way:

```
#ml bioinfo-tools hmmer/3.2.1
fOTUproteoms_DIR="../analyses/fOTU_proteomes/"
VIRUS_HMMs_DIR="../analyses/HMM_scan/virus_HMMs/"
HMMer_OUTPUT_DIR="../analyses/HMM_scan/virus_hmm_scan_out/"

read -r -a fOTUproteoms <<< $( find $fOTUproteoms_DIR -name "*.faa" -and -type f -print0 | x
read -r -a VIRUS_HMMs <<< $( find $VIRUS_HMMs_DIR -name "*.hmm" -and -type f -print0 | xargs
for fOTUproteom in "${fOTUproteoms[@]}"; do
  fOTU_FILE=$(echo $(basename "$fOTUproteom"))
  fOTU=$(echo $(basename "$fOTUproteom") | awk -F "." '{print $1}')
  for VIRUS_HMM in "${VIRUS_HMMs[@]}"; do
    VIR_HMM_FILE=$(echo $(basename "$VIRUS_HMM"))
    VIR_CATEGORY=$(echo $(basename "$VIRUS_HMM") | awk -F "." '{print $1}')
    hmmsearch --cpu 4 -o "$HMMer_OUTPUT_DIR""out/""$fOTU""_""$VIR_CATEGORY"".out" -A "$HMMer
  done
done
```

References

"HMMER v. 3.2.1 (June 2018)." n.d. Accessed November 13, 2019. http://hmmer.org/.

Katoh, Kazutaka, and Daron M Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Slater, Guy St C., and Ewan Birney. 2005. "Automated generation of heuristics for biological sequence comparison." *BMC Bioinformatics* 6 (February): 31. https://doi.org/10.1186/1471-2105-6-31.