Stepe to explore cas proteins in bacterial genome (next step explore them in metagenome)

0) https://www.biostars.org/p/214355/
Download complete bacterial genomes and associated plasmid sequences from NCBI

For this project, it is downloaded like this:
(https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#downloadservice)

to download genomic FASTA sequence for all RefSeq bacterial complete genome assemblies:

Start with an "all[filter]" query on Assembly (https://www.ncbi.nlm.nih.gov/assembly)
Select "Bacteria" from the "Organism group" facet in the left-hand sidebar
Select "Complete genome" from the "Assembly level" facet in the left-hand sidebar
Click on the "Download Assemblies" button to open the download menu
Leave "Source database" set to RefSeq
Select "Genomic FASTA" from the "File type" menu
Wait for the "calculating size..." message to be replaced by an estimated size
Click Download, you may get a pop-up window asking if/where you want to save the genome_assemblies.tar archive file
After the download has finished, expand the tar archive
The resulting folder named "genome_assemblies" will contain:
a report.txt file that provides a summary of what was downloaded
a folder named like "ncbi-genomes-YYYY-MM-DD", where YYYY-MM-DD is the date of the download, containing:
a README.txt file
an md5checksums.txt file
many data files with names like *_genomic.fna.gz, in which the first part of the name is the assembly accession followed by the assembly name

---

1) annotate bacterial genome (make faa files), then make protein database

```
TAGS=$(ls /home/cas_pipeline/all_initial_input/*.fna)

for file in $TAGS; do tag=${file%.fna}; prokka --outdir $all_final_output/prokka/$tag --force --prefix Bacterialgenome "$file"; done

tags=$(ls *.faa)
cat $tags > seqdb
```

---

2) using hmmbuild: build a profile HMM from an alignment (566 cas protein alignment files was used)

```
for alignfile in *.FASTA; do hmmbuild /home/sedreh/Downloads/Supplementary_Dataset_2.profiles/hmm_profiles/"${alignfile%.*}" $alignfile; done
for alignfile in *.sr; do hmmbuild /home/sedreh/Downloads/Supplementary_Dataset_2.profiles/Type_VI_profiles/hmm_profiles/"${alignfile%.*}" $alignfile; done
for alignfile in *.sr; do hmmbuild
```

```
/home/sedreh/Downloads/Supplementary_Dataset_2.profiles/Type_V_profiles/
hmm_profiles/"${alignfile%.*}" $alignfile; done


# using hmmsearch: search a sequence database with a profile HMM

for i in *.hmm; do hmmsearch --tblout
/home/sedreh/Downloads/all_that_I_have/clustering/profiles/hmmsearch_results/$
{i}.tbl ${i} /home/sedreh/Downloads/all_that_I_have/clustering/test/faa_files/seqdb
; done

# now we have all significant hits from hmmseaech! I need to take first column of
each file(I mean Ids) and then we need to prepare bash files to extract sequence of
seach Id from protein sequence database
```

---

solving the problem of esl-sfetch for fetching the significant hit sequences from database

Steps to create a permanent Bash alias:
Open the Terminal app

Edit ~/.bash_aliases or ~/.bashrc file using: nano ~/.bash_aliases
Append your bash alias -->> alias update='sudo yum update'
For example append: alias esl-sfetch='/home/sedreh/hmmer/hmmer-3.1b2-linux-intel-x86_64/binaries/./esl-sfetch'
Save and close the file.
Activate alias by typing: source ~/.bash_aliases

---

3) Extracting significant hits using HMMSEARCH

```
database=$'/home/sedreh/Downloads/all_that_I_have/clustering/test/genomes/prokka/
DATABASE'
INDIR=$'/home/sedreh/Downloads/all_that_I_have/clustering/profiles'
OUTDIR=$'/home/sedreh/Downloads/all_that_I_have/clustering/profiles/
hmmsearch_results'
mkdir $OUTDIR
FILES=$(ls $INDIR/*.hmm)
for i in $FILES; do hmmsearch --tblout $OUTDIR/${i}.tbl ${i} $database ; done
```

---

4) Creating final fastas containg sequences from final hits (extracting first column)

```
INDIR=$'/home/sedreh/Downloads/all_that_I_have/clustering/profiles/
hmmsearch_results'
OUTDIR=$'/home/sedreh/Downloads/all_that_I_have/clustering/profiles/
hmmsearch_results/lists'
cd $INDIR
mkdir $OUTDIR
mkdir $OUTDIR_new
tables=$(cd $INDIR && ls *.hmm.tbl)
for i in $tables; do grep -v "^#" ${i} | awk '{print $1}' >>
$OUTDIR/$i.cleaned_fasta; done

INDIR=$'/home/sedreh/Downloads/all_that_I_have/clustering/profiles/
hmmsearch_results/lists'
```

```
OUTDIR=$'/home/sedreh/Downloads/all_that_I_have/clustering/profiles/
hmmsearch_results/final'
mkdir $OUTDIR
tables=$(cd $INDIR && ls *.hmm.tbl.cleaned_fasta)
for i in $tables; do esl-sfetch -f $database $INDIR/${i} > $OUTDIR/${i%.fasta};
done
```

---

5) cluster the sequences

```
INDIR=$'/home/sedreh/Downloads/all_that_I_have/clustering/profiles/
hmmsearch_results/final'
OUTDIR=$'/home/sedreh/Downloads/all_that_I_have/clustering/profiles/clustering'
mkdir OUTDIR

links=$(ls $INDIR/*.cleaned_fasta)


ln -s $links $cd_hit_input

# this is an important path, it must needs be modified for containerisation
path_to_cdhit=$'/home/sedreh/Downloads/cdhit-master/psi-cd-hit'

cd_hit_input=$"$path_to_cdhit"

TAGS=$(ls $cd_hit_input/*.cleaned_fasta | xargs -n 1 basename)


for i in $TAGS; do cd $path_to_cdhit/; ./psi-cd-hit.pl -i ${i} -o ${i
%.hmm.tbl.cleaned_fasta} -c 0.95; done
```

---

6) Then do multiple sequence alignment

```
for i in $(ls *.clean.fasta); do muscle -in $i -out
/home/sedreh/Downloads/all_that_I_have/clustering/hmmsearch/alignments/fastas/
clean/$i.fasta; done
muscle -in /home/sedreh/Downloads/cdhit-master/psi-cd-hit/Cas10_0_III -out
/home/sedreh/Downloads/cdhit-master/psi-cd-hit/Cas10_0_III.alignment
```

---

7) Then make tree

```
for f in *.fasta; do iqtree -s $f -bb 1000 -alrt 1000 -nt 6
iqtree -s /home/sedreh/Downloads/cdhit-master/psi-cd-hit/Cas10_0_III.alignment -bb
1000 -alrt 1000 -nt 6
```

---

next task

1) cluster the results of hmmsearch (clean fastas) using cd-hit
inputfiles should be in the pATH (/home/sedreh/cdhit-4.8.1/psi-cd-hit)

```
for i in *.clean.fasta; do ./psi-cd-hit.pl -i $i -o /home/sedreh/cdhit-4.8.1/psi-
cd-hit/cas_output/${i%.*}; -c 0.9; done
```

```
./psi-cd-hit.pl -i Cas12c_0_VC.hmm.tbl.cleaned_fasta -o Cas12c_0_VC -c 0.95
```

move all representative sequence files to the folder:

```
for f in *.clean; do
    mv -vn "$f" "/home/sedreh/Downloads/all_that_I_have/clustering/cd_hit_results";
done
```

2) blast representative sequence of each cluster

```
blastp -query cow.small.faa -db human.1.protein.faa -out
cow_vs_human_blast_results.fasta
```

3) select blast hits with 50% similarity

4) make tree

```
muscle -in /home/sedreh/Downloads/cdhit-master/psi-cd-hit/complete/complete.1 -
out /home/sedreh/complete.1.alignment
iqtree -s /home/sedreh/Desktop/example/CLEANED_ALIGN -bb 1000 -alrt 1000 -nt 6
```

5) Compare them with previous trees

```
export PATH="$PATH:/path/to/dir"
```