Identificação GH158

SEXTA-FEIRA, 28/02/2020

GH158 em metagenomas

O primeiro passo foi a busca por GH158 dentro do banco de dados de metagenomas:

```
#run-dbcan v 2.0.6
#CAZyDB.07312019.fa
#dbCAN-HMMdb-V8.txt



run_dbcan.py ~/CNPEM/Metagenomes_LNBR/dbcan/uniprot/${x} protein --out_dir
~/CNPEM/Metagenomes_LNBR/dbcan/${x} --dia_cpu 8 --hmm_cpu 8 --hotpep_cpu 8 --tf_cpu 8 --stp_cpu 8 --db_dir
db/
#Pra cada arquivo de metagenoma foi rodado separadamente, os resultos obtidos estão em
/home/ABTLUS/lucas.maciel/CNPEM/Metagenomes_LNBR/dbcan
```

Os resultados foram concatenados com grep e armazenados no arquivo
`/home/ABTLUS/lucas.maciel/CNPEM/Metagenomes_LNBR/dbcan/GH158_results.txt`

GH158_results.txt

O arquivo fasta destas proteínas está salvo em
`~/bin/seqtk subseq ../all_files.fasta ID_GH158.txt > GH158_meta.fasta`
`#/home/ABTLUS/lucas.maciel/CNPEM/Metagenomes_LNBR/dbcan/GH158_meta.fasta`

GH158 no cazy

As proteínas presentes no cazy já classificadas como GH158 foram obtidas a partir da lista presente no próprio site:
http://www.cazy.org/GH158_all.html
Com o código de acesso busquei no Entrez-batch : https://www.ncbi.nlm.nih.gov/sites/batchentrez

```
#Arquivo salvo em /home/ABTLUS/lucas.maciel/CNPEM/Metagenomes_LNBR/cazy/GH_158.fasta
#A partir deste arquivo removi a redundância dele utilizando CD-HIT
~/bin/cd-hit-v4.8.1-2019-0228/cd-hit -i GH_158.fasta -o GH158_cdhit.out -M 3000 -T 8 -c 1 -t 1 -g 1
#Junto com as proteínas de GH157 eles formam o arquivo non-redundant.fasta
cat GH158_cdhit.out GH_157.fasta > non-redundant.fasta
```

GH158 no Uniprot

Procurei então GH158 no uniprot. Isso foi feito no servidor. Pra fazer isso eu retirei do do arquivo dbCAN-HMMdb-V8.txt o perfil de GH157 e GH158 em um arquivo chamado profile, este arquivo então foi preparado para busca

```
#/home/ABTLUS/lucas.maciel/GH158/dbcan/profiles.txt

hmmpress profiles.txt
```

Uma vez com o perfil preparado, busquei este perfil contra o uniprot

```
#!/bin/bash

#$ -q highmem.q
#$ -cwd
#$ -pe smp 24
#$ -N HMMER
#$ -l h_vmem=20G

hmmsearch --domtblout GH157-8.out -E 1e-15 --cpu 24 profiles.txt
/bioinf/databases/Uniprot/uniprot_trembl.07102018.fasta

#Resultados salvos em #/home/ABTLUS/lucas.maciel/GH158/uniprot_trembl
#Arquivos que apresentaram hits destes dois perfis foram salvos em hits.*
```

Estes arquivos foram então transferidos para o PC local e divididos entre GH157 e GH158 e processados para análise com dbcan e remoção de redundância

```
#/home/ABTLUS/lucas.maciel/CNPEM/Metagenomes_LNBR/dbcan/uniprot/hits.fasta
#Com estes arquivos rodei novamente o dbcan
run_dbcan.py ~/CNPEM/Metagenomes_LNBR/dbcan/uniprot/hits.fasta protein --out_dir dbcan --dia_cpu 8 --
hmm_cpu 8 --hotpep_cpu 8 --tf_cpu 8 --stp_cpu 8 --db_dir db/
#Os que tiveram hits positivos contra GH158 foram recuperados por grep e salvos em
/home/ABTLUS/lucas.maciel/CNPEM/Metagenomes_LNBR/dbcan/uniprot/GH158_ID.txt

#A redundância foi removida com CD-HIT
~/bin/cd-hit-v4.8.1-2019-0228/cd-hit -i GH_158.fasta -o GH158_cdhit.out -M 3000 -T 8 -c 1 -t 1 -g 1
```

Unindo os datasets

Para unir os datasets foi importante retirar a redundância do uniprot das proteínas que já estavam presentes no cazy

```
~/bin/cd-hit-v4.8.1-2019-0228/cd-hit-2d -i ../cazy/non-redundant.fasta -i2 uniprot/uniprot_cdhit.out -o
uniprot-unique.fasta -c 1 -g 1 -M 3000

#Obtive então o fasta deles
~/bin/seqtk subseq uniprot-unique.fasta GH158_ID.txt > GH158.fasta

#Após isso todas foram concatenados
cat GH158_meta.fasta uniprot-unique.fasta cazy/GH158_cdhit.out > uniprot/merge/all_GH158.fasta
#/home/ABTLUS/lucas.maciel/CNPEM/Metagenomes_LNBR/dbcan/uniprot/merge
```