12/09/20016

- Baixei os genomes et proteomas do site <http://www.ncbi.nlm.nih.gov/genome/genomes/2411>

Animal: Sheep

País: Austrália, Argentina, scotland

Strains: C231, PAT10, 3/99-5

Animal: Goat

Pais: Brasil, US/california

Strains: 1002, 226

-formatar os proteomas

maquina : MacMini

/Users/julianab/projects/actives/Monika/data/genome/formatDb.sh

14/09/20016

-Criar arquivos com variaves de ambientes

-Maquina : clusterLGM

export BASE\_DIR=~/projects/actives/Monika/;

export RESULT\_DIR=$BASE\_DIR/results/;

export DATA\_DIR=$BASE\_DIR/data/;

export GENOME\_DIR=$DATA\_DIR/genome/;

export GENOME\_LIST=$GENOME\_DIR/species.list;

export FILES\_DIR=$BASE\_DIR/files/;

export FILES\_RUN\_DIR=$FILES\_DIR/run/;

export LOG\_DIR=$FILES\_DIR/log/;

export LOG\_OUT\_DIR=$LOG\_DIR/out/;

export PERL\_LIB=~/projects/perlLib/;

export HMM\_RES\_DIR=$RESULT\_DIR/hmmscan;

Salvo em : ~/projects/actives/Monika/files/env\_var/environment\_variables\_LGM.sh

Env Var : export ENV\_VARS\_MONIKA=~/projects/actives/Monika/files/env\_var/environment\_variables\_LGM.sh

-Rodar Hmmscan para ver o % de não anotadas

cd $FILES\_RUN\_DIR

nano hmmscan.sh

export FASTA\_FILE=$1;

export OUTPUT\_FILE=$2;

export HMM\_DB=$3;

hmmscan --cpu 8 --cut\_ga --acc --noali --domtblout $OUTPUT\_FILE $HMM\_DB $FASTA\_FILE

chmod +x  hmmscan.sh

nano hmmscan.mpi.sh

export ENV\_VAR\_FILE=$1;

source $ENV\_VAR\_FILE;

for dir in 1002 226 3\_99-5 PAT10 c231;

do

echo $dir;

mkdir -p $HMM\_RES\_DIR/${dir};

qsub -V -S /bin/bash -N Pfam$dir -pe make 8 -e $LOG\_OUT\_DIR -o $LOG\_OUT\_DIR $FILES\_RUN\_DIR/hmmscan.sh $GENOME\_DIR/$dir/${dir}\_protein.faa.ftt $HMM\_RES\_DIR/$dir/hmmscan.out /LGM/common/db/Pfam/27.0/Pfam-A.hmm

done

Executar : bash hmmscan.mpi.sh /home/julianab/projects/actives/Monika/files/env\_var/environment\_variables\_LGM.sh

26/09/2016 Pfam analyses

-Transfer hmmscan results to Mesu cluster

-format hmmscan results for CLADE format

source $ENV\_VARS\_MONIKA

bash $PERL\_LIB/run\_sh/formatHmmscanClade.sh \

$ENV\_VARS\_MONIKA \

$GENOME\_LIST \

$LOG\_DIR/ \

fthmm \

1 \

$HMM\_RES\_DIR/FAMILY\_FACTOR/hmmscan.out \

$HMM\_RES\_DIR/FAMILY\_FACTOR/hmmscan.out.ftt

bash ~/projects/scriptsCluster/run\_mesu.sh 5 $LOG\_DIR fthmm\_runFamily fthmm 4:00:00 > $LOG\_DIR/fthmm.mesu.run

qsub -o $LOG\_OUT\_DIR -e $LOG\_OUT\_DIR $LOG\_DIR/fthmm.mesu.run

-find domain architectures

bash $PERL\_LIB/run\_sh/pfamArchs.sh \

$ENV\_VARS\_MONIKA \

$GENOME\_LIST \

$LOG\_DIR/ \

archPfam \

1 \

$HMM\_RES\_DIR/FAMILY\_FACTOR/hmmscan.out.ftt \

$HMM\_RES\_DIR/FAMILY\_FACTOR/hmmscan.out.archs

bash ~/projects/scriptsCluster/run\_mesu.sh 5 $LOG\_DIR archPfam\_runFamily archPfam 1:00:00 > $LOG\_DIR/archPfam.mesu.run

qsub -q f32c -o $LOG\_OUT\_DIR -e $LOG\_OUT\_DIR $LOG\_DIR/archPfam.mesu.run

-Count number of proteins without domains

for dir in 1002 226 3\_99-5 PAT10 c231;

do

cut -f 4 $dir/hmmscan.out.archs | sort | uniq | wc –l

done

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Strain | #proteins | #domains | #proteins wo domains |
| Goat  País: Brasil, US/california | 1002 | 1984 | 2402 | 316 |
| 226 | 1935 | 2338 | 309 |
| Sheep  País: Austrália, Argentina, scotland | 3\_99-5 | 1994 | 2415 | 320 |
| c231 | 1964 | 2382 | 309 |
| PAT10 | 1970 | 2380 | 322 |

29/09/2016 Clade analyse

-1002

bash ~/projects/actives/CLADE/27/files/run/CASH/search/cash\_test.sh \

~/projects/actives/CLADE/27/files/run/env\_var/environment\_variables\_LGM.sh \

~/projects/actives/CLADE/27/data/pfamLists/pfam\_27.full \

~/projects/actives/CLADE/27/files/log/ \

cladeMo \

116 \

~/projects/actives/Monika/data/genome/1002/1002\_protein.faa.ftt \

~/projects/actives/CLADE/27/data/models/pssms/FAMILY\_FACTOR.tar.gz \

~/projects/actives/CLADE/27/data/models/hmms/FAMILY\_FACTOR.hmm \

~/projects/actives/Monika/results/clade/1002/domainsPfam/FAMILY\_FACTOR/ \

~/projects/actives/CLADE/27/data/pfamLists/used/FAMILY\_FACTOR/list.txt.final \

1

bash ~/projects/scriptsCluster/qsub.sh cladeMo\_runFamily.sh cladeMo ~/projects/actives/CLADE/27/files/log/out/ > cladeMo.jobs

03/10/2016 Clade analyse

* Concat all proteomes and run CLADE in Mesu (Riccardo)
* Sent to Riccardo

05/10/2016 Clade analyse

- Separate archs -> ~/projects/actives/Monika/results/clade/separateArchs.sh

for dir in 1002 226 3\_99-5 PAT10 c231;

do

echo "cat bacteria.archs.txt | awk '\$4 ~ /$dir/' > $dir.archs"

done

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Strain | #proteins | #domains | #proteins wo domains |
| Goat  País: Brasil, US/california | 1002 | 1984 | 2402 | 316 |
| 226 | 1935 | 2338 | 309 |
| Sheep  País: Austrália, Argentina, scotland | 3\_99-5 | 1994 | 2415 | 320 |
| c231 | 1964 | 2382 | 309 |
| PAT10 | 1970 | 2380 | 322 |