Bitacora para el manejo de secuencias ensambladas y búsqueda con *Blastn*

Edith

Para el siguiente ejercicio es necesario tener el Blast+ instalado en la computadora

https://www.ncbi.nlm.nih.gov/guide/data-software/ (https://www.ncbi.nlm.nih.gov/guide/data-software/)

Se utilizarán los contigs formados por el ensamblaje obtenido en la bitacora anterior SOAP

``

In [2]:

```
import os
from pandas import Series, DataFrame
import pandas as pd
from Bio import SeqIO, AlignIO, SeqRecord
from Bio.SeqRecord import SeqRecord
from Bio.Seq import Seq
import matplotlib.pyplot as plt
```

In [3]:

```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/
```

```
/LUSTRE/bioinformatica_data/lga/edith/data/microalgas
```

In [9]:

```
os.makedirs('img',exist_ok=True)
```

In [4]:

```
ls
```

```
CA2/
CA2_blastn.tsv
CA2_S27_L001_R1_001.fastq
CA2_S27_L001_R1_001_fastqc.html
CA2_S27_L001_R1_001_fastqc.zip
```

```
CA2_S27_L001_R1_001.fastq_trimming_report.txt
CA2_S27_L001_R1_001_unpaired_1.fq
CA2_S27_L001_R1_001_val_1_fastqc.html
CA2_S27_L001_R1_001_val_1_fastqc.zip
CA2_S27_L001_R1_001_val_1.fq
CA2_S27_L001_R2_001.fastq
CA2_S27_L001_R2_001_fastqc.html
CA2_S27_L001_R2_001_fastqc.zip
CA2_S27_L001_R2_001.fastq_trimming_report.txt
CA2_S27_L001_R2_001_unpaired_2.fq
CA2_S27_L001_R2_001_val_2_fastqc.html
CA2_S27_L001_R2_001_val_2_fastqc.zip
CA2_S27_L001_R2_001_val_2.fq
CN2/
CN2_S28_L001_R1_001.fastq
CN2_S28_L001_R1_001_fastqc.html
CN2_S28_L001_R1_001_fastqc.zip
CN2_S28_L001_R1_001.fastq_trimming_report.txt
CN2_S28_L001_R1_001_unpaired_1.fq
CN2_S28_L001_R1_001_val_1_fastqc.html
CN2_S28_L001_R1_001_val_1_fastqc.zip
CN2_S28_L001_R1_001_val_1.fq
CN2_S28_L001_R2_001.fastq
CN2_S28_L001_R2_001_fastqc.html
CN2_S28_L001_R2_001_fastqc.zip
CN2_S28_L001_R2_001.fastq_trimming_report.txt
CN2_S28_L001_R2_001_unpaired_2.fq
CN2_S28_L001_R2_001_val_2_fastqc.html
CN2_S28_L001_R2_001_val_2_fastqc.zip
CN2_S28_L001_R2_001_val_2.fq
config_CcalcitransCA.txt
img/
lecturas_fq_gz.txt
lecturas.txt
MA2/
MA2_S25_L001_R1_001.fastq
MA2_S25_L001_R1_001_fastqc.html
MA2_S25_L001_R1_001_fastqc.zip
MA2_S25_L001_R1_001.fastq_trimming_report.txt
MA2_S25_L001_R1_001_unpaired_1.fq
MA2_S25_L001_R1_001_val_1_fastqc.html
MA2_S25_L001_R1_001_val_1_fastqc.zip
MA2_S25_L001_R1_001_val_1.fq
MA2_S25_L001_R2_001.fastq
MA2_S25_L001_R2_001_fastqc.html
MA2_S25_L001_R2_001_fastqc.zip
MA2_S25_L001_R2_001.fastq_trimming_report.txt
MA2_S25_L001_R2_001_unpaired_2.fq
MA2_S25_L001_R2_001_val_2_fastqc.html
MA2_S25_L001_R2_001_val_2_fastqc.zip
```

```
MA2_S25_L001_R2_001_val_2.fq
MN2/
MN2_S26_L001_R1_001.fastq
MN2_S26_L001_R1_001_fastqc.html
MN2_S26_L001_R1_001_fastqc.zip
MN2_S26_L001_R1_001.fastq_trimming_report.txt
MN2_S26_L001_R1_001_unpaired_1.fq
MN2_S26_L001_R1_001_val_1_fastqc.html
MN2_S26_L001_R1_001_val_1_fastqc.zip
MN2_S26_L001_R1_001_val_1.fq
MN2_S26_L001_R2_001.fastq
MN2_S26_L001_R2_001_fastqc.html
MN2_S26_L001_R2_001_fastqc.zip
MN2_S26_L001_R2_001.fastq_trimming_report.txt
MN2_S26_L001_R2_001_unpaired_2.fq
MN2_S26_L001_R2_001_val_2_fastqc.html
MN2_S26_L001_R2_001_val_2_fastqc.zip
MN2_S26_L001_R2_001_val_2.fq
numero_lecturas_arc.csv
numero_lecturas.csv
numero_lecturas_fq.csv
numero_lecturas_fq_gz.csv
soap_config/
tiempoCA2_blastn.txt
XA2/
XA2_S29_L001_R1_001.fastq
XA2_S29_L001_R1_001_fastqc.html
XA2_S29_L001_R1_001_fastqc.zip
XA2_S29_L001_R1_001.fastq_trimming_report.txt
XA2_S29_L001_R1_001_unpaired_1.fq
XA2_S29_L001_R1_001_val_1_fastqc.html
XA2_S29_L001_R1_001_val_1_fastqc.zip
XA2_S29_L001_R1_001_val_1.fq
XA2_S29_L001_R2_001.fastq
XA2_S29_L001_R2_001_fastqc.html
XA2_S29_L001_R2_001_fastqc.zip
XA2_S29_L001_R2_001.fastq_trimming_report.txt
XA2_S29_L001_R2_001_unpaired_2.fq
XA2_S29_L001_R2_001_val_2_fastqc.html
XA2_S29_L001_R2_001_val_2_fastqc.zip
XA2_S29_L001_R2_001_val_2.fq
XN2/
XN2_S30_L001_R1_001.fastq
XN2_S30_L001_R1_001_fastqc.html
XN2_S30_L001_R1_001_fastqc.zip
XN2_S30_L001_R1_001.fastq_trimming_report.txt
XN2_S30_L001_R1_001_unpaired_1.fq
XN2_S30_L001_R1_001_val_1_fastqc.html
XN2_S30_L001_R1_001_val_1_fastqc.zip
XN2_S30_L001_R1_001_val_1.fq
```

```
XN2_S30_L001_R2_001.fastq
XN2_S30_L001_R2_001_fastqc.html
XN2_S30_L001_R2_001_fastqc.zip
XN2_S30_L001_R2_001.fastq_trimming_report.txt
XN2_S30_L001_R2_001_unpaired_2.fq
XN2_S30_L001_R2_001_val_2_fastqc.html
XN2_S30_L001_R2_001_val_2_fastqc.zip
XN2_S30_L001_R2_001_val_2.fq
```

Se analizarán con blastn los contigs obtenidos a la base de datos *nt*

Verifique la localización de la base de datos, en este caso se encuentra en `~/DATA/nt/` o corrija si es necesario

In [14]:

```
!grep ">" CA2/CA2_out.contig |wc -l
!grep ">" CN2/CN2_out.contig |wc -l
!grep ">" MA2/MA2_out.contig |wc -l
!grep ">" MN2/MN2_out.contig |wc -l
!grep ">" XA2/XA2_out.contig |wc -l
!grep ">" XN2/XN2_out.contig |wc -l
```

```
1790273
1746104
1387216
1529856
2921160
2725480
```

In [15]:

```
!grep -c ">" CA2/CA2.fasta
!grep -c ">" CN2/CN2.fasta
!grep -c ">" MA2/MA2.fasta
!grep -c ">" MN2/MN2.fasta
!grep -c ">" XA2/XA2.fasta
!grep -c ">" XN2/XN2.fasta
```

```
122626
62186
73443
80891
138890
112965
```

```
In [5]:  pwd
```

```
Out[5]:  '/LUSTRE/bioinformatica_data/lga/edith/data/microalgas'
```

```
In [17]:  !find CA2/CA2_out.contig
```

```
CA2/CA2_out.contig
```

```
In [6]:  ls /LUSTRE/bioinformatica_data/BD/blast/db/NT/
```

```
archivos_tar/  nt.04.nhi  nt.08.nin  nt.12.nni  nt.16.nsq  nt.21.nhi
nt.25.nin
nt.00.nhd      nt.04.nhr  nt.08.nnd  nt.12.nog  nt.17.nhd  nt.21.nhr
nt.25.nnd
nt.00.nhi      nt.04.nin  nt.08.nni  nt.12.nsq  nt.17.nhi  nt.21.nin
nt.25.nni
nt.00.nhr      nt.04.nnd  nt.08.nog  nt.13.nhd  nt.17.nhr  nt.21.nnd
nt.25.nog
nt.00.nin      nt.04.nni  nt.08.nsq  nt.13.nhi  nt.17.nin  nt.21.nni
nt.25.nsq
nt.00.nnd      nt.04.nog  nt.09.nhd  nt.13.nhr  nt.17.nnd  nt.21.nog
nt.26.nhd
nt.00.nni      nt.04.nsq  nt.09.nhi  nt.13.nin  nt.17.nni  nt.21.nsq
nt.26.nhi
nt.00.nog      nt.05.nhd  nt.09.nhr  nt.13.nnd  nt.17.nog  nt.22.nhd
nt.26.nhr
nt.00.nsq      nt.05.nhi  nt.09.nin  nt.13.nni  nt.17.nsq  nt.22.nhi
nt.26.nin
nt.01.nhd      nt.05.nhr  nt.09.nnd  nt.13.nog  nt.18.nhd  nt.22.nhr
nt.26.nnd
nt.01.nhi      nt.05.nin  nt.09.nni  nt.13.nsq  nt.18.nhi  nt.22.nin
nt.26.nni
nt.01.nhr      nt.05.nnd  nt.09.nog  nt.14.nhd  nt.18.nhr  nt.22.nnd
nt.26.nog
nt.01.nin      nt.05.nni  nt.09.nsq  nt.14.nhi  nt.18.nin  nt.22.nni
nt.26.nsq
nt.01.nnd      nt.05.nog  nt.10.nhd  nt.14.nhr  nt.18.nnd  nt.22.nog
nt.27.nhd
nt.01.nni      nt.05.nsq  nt.10.nhi  nt.14.nin  nt.18.nni  nt.22.nsq
nt.27.nhi
nt.01.nog      nt.06.nhd  nt.10.nhr  nt.14.nnd  nt.18.nog  nt.23.nhd
nt.27.nhr
nt.01.nsq      nt.06.nhi  nt.10.nin  nt.14.nni  nt.18.nsq  nt.23.nhi
nt.27.nin
nt.02.nhd      nt.06.nhr  nt.10.nnd  nt.14.nog  nt.19.nhd  nt.23.nhr
nt.27.nnd
nt.02.nhi      nt.06.nin  nt.10.nni  nt.14.nsq  nt.19.nhi  nt.23.nin
nt.27.nni
nt.02.nhr      nt.06.nnd  nt.10.nog  nt.15.nhd  nt.19.nhr  nt.23.nnd
nt.27.nog
nt.02.nin      nt.06.nni  nt.10.nsq  nt.15.nhi  nt.19.nin  nt.23.nni
nt.27.nsq
nt.02.nnd      nt.06.nog  nt.11.nhd  nt.15.nhr  nt.19.nnd  nt.23.nog
nt.nal
nt.02.nni      nt.06.nsq  nt.11.nhi  nt.15.nin  nt.19.nni  nt.23.nsq
nt.ndb
nt.02.nog      nt.07.nhd  nt.11.nhr  nt.15.nnd  nt.19.nog  nt.24.nhd
nt.nos
nt.02.nsq      nt.07.nhi  nt.11.nin  nt.15.nni  nt.19.nsq  nt.24.nhi
nt.not
nt.03.nhd      nt.07.nhr  nt.11.nnd  nt.15.nog  nt.20.nhd  nt.24.nhr
nt.ntf
nt.03.nhi      nt.07.nin  nt.11.nni  nt.15.nsq  nt.20.nhi  nt.24.nin
nt.nto
nt.03.nhr      nt.07.nnd  nt.11.nog  nt.16.nhd  nt.20.nhr  nt.24.nnd
taxdb.btd
nt.03.nin      nt.07.nni  nt.11.nsq  nt.16.nhi  nt.20.nin  nt.24.nni
taxdb.bti
nt.03.nnd      nt.07.nog  nt.12.nhd  nt.16.nhr  nt.20.nnd  nt.24.nog
nt.03.nni      nt.07.nsq  nt.12.nhi  nt.16.nin  nt.20.nni  nt.24.nsq
nt.03.nog      nt.08.nhd  nt.12.nhr  nt.16.nnd  nt.20.nog  nt.25.nhd
nt.03.nsq      nt.08.nhi  nt.12.nin  nt.16.nni  nt.20.nsq  nt.25.nhi
nt.04.nhd      nt.08.nhr  nt.12.nnd  nt.16.nog  nt.21.nhd  nt.25.nhr
```

Ejecuta el blastn en slurum

In [9]:

```
fout = open("blastn_CN2.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
```

```
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/l
ib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/NT

#
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CN2
date > tiempoCN2_blastn.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin/blastn \\
 -query CN2.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out CN2_blastn.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames scomnames sblastna
mes strand"
date >> tiempoCN2_blastn.txt

head CN2_blastn.tsv
echo ""
grep -c CN2_blastn.tsv

"""
fout.write(linea)
fout.close()
```

In [30]:

```
!head -100 blastn_CN2.sh
```

```
#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0
/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/NT

#
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CN2
date > tiempoCN2_blastn.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin/blastn \
 -query CN2.fasta \
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \
 -out CN2_blastn.tsv \
 -evalue 1E-6 \
 -max_target_seqs 1 \
 -num_threads 24 \
 -outfmt "6 std sskingdoms stitle staxids sscinames scomnames sblast
names strand"
date >> tiempoCN2_blastn.txt

head CN2_blastn.tsv
echo ""
grep -c CN2_blastn.tsv
```

Mandar cola de trabajo en el servidor

```
In [11]: !sbatch blastn_CN2.sh
```

```
Submitted batch job 166788
```

```
In [33]: !head tiempoCN2_blastn.txt
```

```
head: cannot open 'tiempoCN2_blastn.txt' for reading: No such file o
r directory
```

Comando que verifica el nodulo donde esta llevandose a cabo el proceso en slurum

```
In [36]: !squeue
```

```
             JOBID PARTITION     NAME     USER ST       TIME  NODES
NODELIST(REASON)
            166792    cicese   blastn elizondo PD       0:00      1
(AssocMaxJobsLimit)
            166793    cicese   blastn elizondo PD       0:00      1
(AssocMaxJobsLimit)
            166813       d30   copTHR  gvkaren PD       0:00      1
(AssocMaxJobsLimit)
            166814    cicese  copLTHR  gvkaren PD       0:00      1
(AssocMaxJobsLimit)
            166818    cicese   blastn elizondo PD       0:00      1
(AssocMaxJobsLimit)
            166782    cicese      wTO   rgomez  R 1-02:50:37      1
nodo7
            166783    cicese      wTO   rgomez  R 1-02:50:37      1
nodo8
            166784    cicese      wTO   rgomez  R 1-02:50:37      1
nodo9
            166791    cicese   blastn elizondo  R   10:23:14      1
nodo3
            166812       d30  nCopTHR  gvkaren  R    1:45:37      1
nodo10
            166807    cicese  copLGPB  gvkaren  R    2:38:38      1
nodo6
            166805    cicese  copACHE  gvkaren  R    2:41:16      1
nodo5
            166802       d30  nCopGPB  gvkaren  R    2:46:01      1
nodo4
            166789    cicese   blastn elizondo  R 1-00:08:46      1
nodo18
            166817    cicese     bash   sylvia  R       8:43      2
nodo[1,21]
            166786       d30   spr4_4  sduenas  R 1-00:44:02      4
nodo[11,13-15]
```

In [13]:

```
fout = open("blastn_CA2.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
```

```
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/NT

#
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CA2
date > tiempoCA2_blastn.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin/blastn \\
 -query CA2.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out CA2_blastn.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames scomnames sblastnames strand"
date >> tiempoCA2_blastn.txt

head CA2_blastn.tsv
echo ""
grep -c CA2_blastn.tsv

"""
fout.write(linea)
fout.close()
```

Mandar a la cola de trabajo del servidor

In [14]:

```
!sbatch blastn_CA2.sh
```

```
Submitted batch job 166789
```

In [34]:

```
fout = open("blastn_MA2.sh", "w")
```

```
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/NT

#
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/MA2
date > tiempoMA2_blastn.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin/blastn \\
 -query MA2.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out MA2_blastn.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames scomnames sblastnames strand"
date >> tiempoMA2_blastn.txt

head MA2_blastn.tsv
echo ""
grep -c MA2_blastn.tsv

"""
```

```
fout.write(linea)
fout.close()
```

In [35]:

```
!sbatch blastn_MA2.sh
```

```
Submitted batch job 166818
```

In [17]:

```
fout = open("blastn_MN2.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/l
ib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/NT

#
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/MN2
date > tiempoMN2_blastn.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin/blastn \\
 -query MN2.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out MN2_blastn.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames scomnames sblastnames strand"
date >> tiempoMN2_blastn.txt

echo ""  >> tiempoMN2_blastn.txt
head MN2_blastn.tsv >> tiempoMN2_blastn.txt
echo ""  >> tiempoMN2_blastn.txt
grep -c MN2_blastn.tsv >> tiempoMN2_blastn.txt

"""
fout.write(linea)
fout.close()
```

In [18]:

```
!sbatch blastn_MN2.sh
```

```
Submitted batch job 166791
```

In [19]:

```
fout = open("blastn_XA2.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/NT
```

```
#
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XA2
date > tiempoXA2_blastn.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin/blastn \\
 -query XA2.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out XA2_blastn.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames scomnames sblastna
mes strand"
date >> tiempoXA2_blastn.txt

echo ""  >> tiempoMN2_blastn.txt
head MN2_blastn.tsv >> tiempoMN2_blastn.txt
echo ""  >> tiempoMN2_blastn.txt
grep -c MN2_blastn.tsv >> tiempoMN2_blastn.txt

"""
fout.write(linea)
fout.close()
```

In [20]:

```
!sbatch blastn_XA2.sh
```

```
Submitted batch job 166792
```

In [21]:

```
fout = open("blastn_XN2.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
```

```
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/NT

#
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XN2
date > tiempoXN2_blastn.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin/blastn \\
 -query XN2.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out XN2_blastn.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames scomnames sblastnames strand"
date >> tiempoXN2_blastn.txt

echo ""  >> tiempoMN2_blastn.txt
head MN2_blastn.tsv >> tiempoMN2_blastn.txt
echo ""  >> tiempoMN2_blastn.txt
grep -c MN2_blastn.tsv >> tiempoMN2_blastn.txt
echo ""  >> tiempoMN2_blastn.txt
"""
fout.write(linea)
fout.close()
```

In [22]:

```
!sbatch blastn_XN2.sh
```

```
Submitted batch job 166793
```

In [2]:

```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/
```

```
/LUSTRE/bioinformatica_data/lga/edith/data/microalgas
```

Comando para verificar el contenido de los archivos *.err que se generan como resultado de las corridas

In [3]:

```
!for f in blastn.*.err; do echo $f; ls -lh $f; head $f; echo "--------------"; done
```

```
blastn.nodo16.166790.err
blastn.nodo16.166790.err
-rw-r--r-- 1 elizondo gen_acuicola 188 Mar  5 05:32 blastn.nodo16.16
6790.err
ModuleCmd_Load.c(213):ERROR:105: Unable to locate a modulefile for '
gcc-7.2'
Warning: [blastn] Examining 5 or more matches is recommended

real    167m55.961s
user    204m46.201s
sys     70m10.681s
--------------
blastn.nodo17.166788.err
-rw-r--r-- 1 elizondo gen_acuicola 189 Mar  5 02:44 blastn.nodo17.16
6788.err
ModuleCmd_Load.c(213):ERROR:105: Unable to locate a modulefile for '
gcc-7.2'
Warning: [blastn] Examining 5 or more matches is recommended

real    657m54.144s
user    197m53.851s
sys     309m28.164s
--------------
blastn.nodo17.166792.err
-rw-r--r-- 1 elizondo gen_acuicola 189 Mar  5 23:00 blastn.nodo17.16
6792.err
ModuleCmd_Load.c(213):ERROR:105: Unable to locate a modulefile for '
gcc-7.2'
Warning: [blastn] Examining 5 or more matches is recommended

real    420m44.620s
user    383m23.501s
sys     145m56.196s
--------------
blastn.nodo18.166789.err
-rw-r--r-- 1 elizondo gen_acuicola 190 Mar  6 03:21 blastn.nodo18.16
6789.err
ModuleCmd_Load.c(213):ERROR:105: Unable to locate a modulefile for '
gcc-7.2'
Warning: [blastn] Examining 5 or more matches is recommended

real    2134m35.757s
user    535m34.637s
sys     995m59.113s
--------------
blastn.nodo20.166818.err
-rw-r--r-- 1 elizondo gen_acuicola 111 Mar  5 20:18 blastn.nodo20.16
6818.err
Warning: [blastn] Examining 5 or more matches is recommended

real    209m49.312s
user    201m25.977s
```

```
sys     58m54.002s
--------------
blastn.nodo22.166793.err
-rw-r--r-- 1 elizondo gen_acuicola 110 Mar  5 21:45 blastn.nodo22.16
6793.err
Warning: [blastn] Examining 5 or more matches is recommended

real    296m4.717s
user    280m14.132s
sys     81m26.629s
--------------
blastn.nodo3.166791.err
-rw-r--r-- 1 elizondo gen_acuicola 113 Mar  5 23:03 blastn.nodo3.166
791.err
Warning: [blastn] Examining 5 or more matches is recommended

real    1050m37.566s
user    281m12.837s
sys     474m59.335s
--------------
```

Verificar el archivo de blast XA2 que muestra error en blastn.nodo17.166792.err

In [5]:

```
!head blastn.nodo17.166792.err
```

```
ModuleCmd_Load.c(213):ERROR:105: Unable to locate a modulefile for '
gcc-7.2'
Warning: [blastn] Examining 5 or more matches is recommended

real    420m44.620s
user    383m23.501s
sys     145m56.196s
```

In [6]:

```
!head blastn.nodo17.166792.log
```

```
5564481 CP019388.1      97.500  200     4       1       3       201
1993520 1993719 1.30e-89        340     Bacteria
Winogradskyella sp. J14-2, complete genome      1936080 Winogradskye
lla sp. J14-2   Winogradskyella sp. J14-2       CFB group bacteria
5564531 CP045367.1      97.030  202     5       1       1       201
3752429 3752630 4.66e-89        339     Bacteria        Marinobacter
sp. THAF39 chromosome, complete genome  2587857 Marinobacter sp. THA
F39     Marinobacter sp. THAF39 g-proteobacteria
5564567 CP019388.1      86.500  200     27      0       1       200
158033  158232  1.76e-53        220     Bacteria
Winogradskyella sp. J14-2, complete genome      1936080 Winogradskye
lla sp. J14-2   Winogradskyella sp. J14-2       CFB group bacteria
5564595 AB266130.2      93.035  201     14      0       1       201
14053   13853   1.02e-75        294     Bacteria        Uncultured b
acterium DNA, fosmid clone, clone: 04E12        77133   uncultured b
acterium        uncultured bacterium    bacteria
5564603 LT629752.1      79.048  105     20      2       54      157
1124637 1124534 1.88e-08        71.3    Bacteria        Polaribacter
sp. KT25b genome assembly, chromosome: I        1855336 Polaribacter
sp. KT25b       Polaribacter sp. KT25b  CFB group bacteria
5564619 XR_004637948.1  91.489  47      4       0       91      137
740     786     8.74e-07        65.8    Eukaryota       PREDICTED: S
etaria viridis kinesin-like protein KIN-12E (LOC117844868), transcri
pt variant X4, misc_RNA 4556    Setaria viridis Setaria viridis mono
cots
5564657 CP002825.1      91.026  156     11      3       48      201
671962  671808  1.37e-49        207     Bacteria        Lacinutrix s
p. 5H-3-7-4, complete genome    983544  Lacinutrix sp. 5H-3-7-4 Laci
nutrix sp. 5H-3-7-4     CFB group bacteria
5564685 CP019388.1      85.000  160     24      0       1       160
2163516 2163357 3.01e-36        163     Bacteria
Winogradskyella sp. J14-2, complete genome      1936080 Winogradskye
lla sp. J14-2   Winogradskyella sp. J14-2       CFB group bacteria
5564705 CP019288.1      86.567  201     27      0       1       201
4788863 4789063 4.90e-54        222     Bacteria        Kordia antar
ctica strain IMCC3317 chromosome, complete genome       1218801 Kord
ia antarctica   Kordia antarctica       CFB group bacteria
5564729 CP013195.1      88.764  89      8       2       113     201
206084  205998  1.43e-19        108     Bacteria        Prevotella e
noeca strain F0113, complete genome     76123   Prevotella enoeca
Prevotella enoeca       CFB group bacteria
```

Este error corresponde al archivo blastn_XA2.sh por lo que hay que visualizarlo

```
In [7]: !head -100 blastn_XA2.sh
```

```
#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0
/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/NT

#
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XA2
date > tiempoXA2_blastn.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin/blastn \
 -query XA2.fasta \
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \
 -out XA2_blastn.tsv \
 -evalue 1E-6 \
 -max_target_seqs 1 \
 -num_threads 24 \
 -outfmt "6 std sskingdoms stitle staxids sscinames scomnames sblast
names strand"
date >> tiempoXA2_blastn.txt

head XA2_blastn.tsv
echo ""
grep -c XA2_blastn.tsv
```

se visualiza el contenido del archivo de salida XA2_blastn.tsv

In [11]:

```
%%bash
head XA2/XA2_blastn.tsv
echo "numero de resultados es:"
wc -l XA2/XA2_blastn.tsv
```

```
5564481 CP019388.1      97.500  200     4       1       3       201
1993520 1993719 1.30e-89        340     Bacteria
Winogradskyella sp. J14-2, complete genome      1936080 Winogradskye
lla sp. J14-2   Winogradskyella sp. J14-2       CFB group bacteria
5564531 CP045367.1      97.030  202     5       1       1       201
3752429 3752630 4.66e-89        339     Bacteria        Marinobacter
sp. THAF39 chromosome, complete genome  2587857 Marinobacter sp. THA
F39     Marinobacter sp. THAF39 g-proteobacteria
5564567 CP019388.1      86.500  200     27      0       1       200
158033  158232  1.76e-53        220     Bacteria
Winogradskyella sp. J14-2, complete genome      1936080 Winogradskye
lla sp. J14-2   Winogradskyella sp. J14-2       CFB group bacteria
5564595 AB266130.2      93.035  201     14      0       1       201
14053   13853   1.02e-75        294     Bacteria        Uncultured b
acterium DNA, fosmid clone, clone: 04E12        77133   uncultured b
acterium        uncultured bacterium    bacteria
5564603 LT629752.1      79.048  105     20      2       54      157
1124637 1124534 1.88e-08        71.3    Bacteria        Polaribacter
sp. KT25b genome assembly, chromosome: I        1855336 Polaribacter
sp. KT25b       Polaribacter sp. KT25b  CFB group bacteria
5564619 XR_004637948.1  91.489  47      4       0       91      137
740     786     8.74e-07        65.8    Eukaryota       PREDICTED: S
etaria viridis kinesin-like protein KIN-12E (LOC117844868), transcri
pt variant X4, misc_RNA 4556    Setaria viridis Setaria viridis mono
cots
5564657 CP002825.1      91.026  156     11      3       48      201
671962  671808  1.37e-49        207     Bacteria        Lacinutrix s
p. 5H-3-7-4, complete genome    983544  Lacinutrix sp. 5H-3-7-4 Laci
nutrix sp. 5H-3-7-4     CFB group bacteria
5564685 CP019388.1      85.000  160     24      0       1       160
2163516 2163357 3.01e-36        163     Bacteria
Winogradskyella sp. J14-2, complete genome      1936080 Winogradskye
lla sp. J14-2   Winogradskyella sp. J14-2       CFB group bacteria
5564705 CP019288.1      86.567  201     27      0       1       201
4788863 4789063 4.90e-54        222     Bacteria        Kordia antar
ctica strain IMCC3317 chromosome, complete genome       1218801 Kord
ia antarctica   Kordia antarctica       CFB group bacteria
5564729 CP013195.1      88.764  89      8       2       113     201
206084  205998  1.43e-19        108     Bacteria        Prevotella e
noeca strain F0113, complete genome     76123   Prevotella enoeca
Prevotella enoeca       CFB group bacteria
numero de resultados es:
16757 XA2/XA2_blastn.tsv
```

In [12]:

```
ls -lh *.tsv
```

```
-rw-r--r-- 1 elizondo gen_acuicola 0 Feb 22 16:00 CA2_blastn.tsv
```

In [18]:

```
ls -ld *.tsv
```

```
-rw-r--r-- 1 elizondo gen_acuicola 0 Feb 22 16:00 CA2_blastn.tsv
```

se visualizan los archivos .tsv que son los que tienen la informacion del blastn

In [23]:

```
ls */*.tsv
```

```
CA2/CA2_blastn.tsv  MA2/MA2_blastn.tsv  XA2/XA2_blastn.tsv
CN2/CN2_blastn.tsv  MN2/MN2_blastn.tsv  XN2/XN2_blastn.tsv
```

se copian los archivos .tsv desde Lustre hasta mi caprteta tsv en mi direccion de omica

In [25]:

```
%%bash
for f in ls */*.tsv
do
echo $f
cp $f ~/data/microalgas/tsv/
done
```

```
ls
CA2/CA2_blastn.tsv
CN2/CN2_blastn.tsv
MA2/MA2_blastn.tsv
MN2/MN2_blastn.tsv
XA2/XA2_blastn.tsv
XN2/XN2_blastn.tsv
```

```
cp: cannot stat 'ls': No such file or directory
```

In [29]:

```
cd ~/data/microalgas/
```

```
/home/elizondo/data/microalgas
```

El comando tar es usado para comprimir los archivos de interés que posteriormente serán descargados del directorio y así ser analizados en excel cada resultado de blastn. Y para empaquetar una compresión de alguna carpeta se debe usar el algo así como comprimir archivo/s o carpeta/s, se debe realizar de la siguiente manera:

!tar -czv

```
In [27]: !tar -czvf tsv.tar.gz ./tsv
```

```
./tsv/
./tsv/XA2_blastn.tsv
./tsv/MA2_blastn.tsv
./tsv/CA2_blastn.tsv
./tsv/XN2_blastn.tsv
./tsv/CN2_blastn.tsv
./tsv/MN2_blastn.tsv
```

```
In [28]: cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas
```

```
/LUSTRE/bioinformatica_data/lga/edith/data/microalgas
```

```
In [32]: !grep ">" ~/data/microalgas
```

```
grep: /home/elizondo/data/microalgas: Is a directory
```

```
In [6]: !grep ">"
~/Documents/secuenciacion_masiva/sec2019/microalgas/CN2/CN2_clc/Contig
_CN2_T.fasta |wc -l
```

```
    6542
```

```
In [7]: !grep ">"
~/Documents/secuenciacion_masiva/sec2019/microalgas/MA2/MA2_clc//Conti
g_MA2_T.fasta |wc -l
```

```
   21282
```

```
In [8]: !grep ">"
~/Documents/secuenciacion_masiva/sec2019/microalgas/MN2/MN2_clc/Contig
_MN2_T.fasta |wc -l
```

```
   20167
```

In [9]:

```
!grep ">"
~/Documents/secuenciacion_masiva/sec2019/microalgas/XA2/XA2_clc/Contig
_XA2_T.fasta |wc -l
```

```
40958
```

In [10]:

```
!grep ">"
~/Documents/secuenciacion_masiva/sec2019/microalgas/XN2/XN2_clc/Contig
_XN2_T.fasta |wc -l
```

```
44350
```

In [ ]:

```
pwd
```

In [ ]:

```
ls
```

In [ ]:

```
!head -2 CA2_blastn.tsv
```

In [ ]:

```
encabezado =("qseqid", "sseqid", "pident", "length", "mismatch", "gapo
pen","qstart",
             "qend", "sstart","send", "evalue", "bitscore", "sskingdom
s", "stitle",
             "staxids", "sscinames", "scomnames", "sblastnames")
```

In [ ]:

```
ftab=pd.read_csv("CA2_blastn.tsv", header=None, sep = "\t" , names= en
cabezado)
ftab.head()
```

In [ ]:

```
len(ftab)
```

In [ ]:

```
ftab1= ftab.groupby("sskingdoms")["qseqid"].count()
ftab1 = DataFrame(ftab1)
ftab1
```

Cuántos contigs no son eucariotas?

Guadardo la base de datos en formato csv

In [ ]:

```
ftab.to_csv("CA2_blastn.csv", header=True, index= None)
```

en caso de haber el hecho el análisis previo y querer recuperar el archivo anterior

```
In [ ]: ftab= pd.read_csv("contigs_blastn.csv")
        ftab.head(2)
```

Hay algún contig con más de una asignación (duplicados)?

```
In [ ]: ftab1= ftab.groupby("qseqid")["qseqid"].count()
        ftab1 = DataFrame(ftab1)
        ftab1
```

Es necesario eliminar duplicados

```
In [ ]: ftab1 =ftab.drop_duplicates(subset = 'qseqid', keep='first', inplace =
        False)
        ftab1
```

Cuántos grupos hay a parte del que pertenece la especie analizada?

```
In [ ]: ftab2= ftab1.groupby(["sskingdoms", "sblastnames"])["qseqid"].count()
        ftab2 = DataFrame(ftab2)
        ftab2
```

```
In [ ]: ftab2= ftab1.groupby(["sskingdoms", "sscinames"])["qseqid"].count()
        ftab2.sort_values(axis = 0, ascending=False, inplace=True)
        ftab2
```

Procedimiento para simplificar tabla y graficar las 10 primeras categorías y el resto ponerlas en "otras"

In [ ]:

```
linea10=ftab2[:10]
linea11=ftab2[10:]
#linea10
otro=sum (linea11)
#otro
otros = pd.DataFrame({0:otro}, index=["Other"])
otros
linea10=linea10.append(otros)
linea10
```

In [ ]:

```
linea10.plot(kind='barh', figsize= (8,6))
plt.axis([-1, max(linea10[0])+5, -1, 10.8], label=None)
plt.legend().set_visible(False)
plt.xlabel("Frecuencia")
plt.ylabel("Especies")
plt.title("Especies con resultado de blastn")
plt.show()
```