Bitácora de Blastn de ensamblajes de microalgas

Edith Elizondo

Se cargan las funciones que se utilizarán en este proceso

In [ ]:
```
from Bio import SeqIO, pairwise2, AlignIO, Phylo, Entrez, SeqRecord, S
eq, SearchIO
from Bio.Align.Applications import ClustalwCommandline
from Bio.Blast import NCBIWWW, NCBIXML
from Bio.Seq import Seq
from Bio.SeqUtils import GC
from Bio.SeqRecord import SeqRecord

from matplotlib import *
import matplotlib.pyplot as plt
from matplotlib_venn import venn3_unweighted, venn2_unweighted

import os, pylab

from pandas import DataFrame
import pandas as pd

import pylab as pl
from pylab import *
```

Se definen funciones a utilizar en la bitácora

In [ ]:
```
def cpg(secuencia):
    g= secuencia.count("G")
    c= secuencia.count("C")
    cg= secuencia.count("CG")
    lar= len(secuencia)
    cpG=0
    try:
        g*c==0
    except:
        cpG=0
    else:
        if g == 0 or c== 0:
            cpG =0
        else:
            cpG=(round(cg/(g*c)*(lar**2/(lar-1)) ,8))
    return (cpG)

def generoespecie(genesp):
    genero=genesp[:genesp.find(" ")]
    #print(genero)
    especie = genesp[genesp.find(" ")+1:]
    especie = especie[:especie.find(" ")]
    #print(especie)
    genesp1 = genero+" "+especie
    return(genesp1)

def gespecie(genesp):
    genero=genesp[:1]+"._"
    #print(genero)
    especie = genesp[genesp.find(" ")+1:]
    especie = especie[:especie.find(" ")]
    #print(especie)
    genesp1 = genero+especie
    return(genesp1)
```

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

In [ ]:
```
ls /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CA2
```

In [ ]:
```
ls ~/data/microalgas/tsv
```

Número de secuencias en el archivo tsv que contiene todas las secuencias generadas por el ensamblaje y ya analizadas en blastn

In [ ]:
```
!grep /tsv/CA2_blastn.tsv
```

In [ ]:
```
!grep ">" /tsv/CA2_blastn.tsv  |wc -l
```

In [ ]:
```
!grep -c ">" tsv/CA2_blastn.tsv
```

In [ ]:
```
pwd
```

Se inicia el análisis de los datos de blastn

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "mismatch", "gapo
pen", "qstart",
             "qend", "sstart", "send", "evalue", "bitscore", "sskingdo
ms", "stitle",
             "staxids", "sscinames", "scomnames", "sblastnames")
```

In [ ]:
```
ftsv=pd.read_csv("CA2_blastn.tsv", sep = "\t", header=None , names= en
cabezado, engine="python")
ftsv.head()
```

Guardando los datos en formato csv

In [ ]:
```
ftsv.to_csv("CA2_blastn.csv", header=True, index= None)
```

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

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "mismatch", "gapo
pen", "qstart",
             "qend", "sstart", "send", "evalue", "bitscore", "sskingdo
ms", "stitle",
             "staxids", "sscinames", "scomnames", "sblastnames")
```

In [ ]:
```
ftsv=pd.read_csv("CN2_blastn.tsv", sep = "\t", header=None , names= en
cabezado, engine="python")
ftsv.head()
```

Guardando los datos en formato csv

In [ ]:
```
ftsv.to_csv("CN2_blastn.csv", header=True, index= None)
```

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

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "mismatch", "gapo
pen", "qstart",
             "qend", "sstart", "send", "evalue", "bitscore", "sskingdo
ms", "stitle",
             "staxids", "sscinames", "scomnames", "sblastnames")
```

In [ ]:
```
ftsv=pd.read_csv("MA2_blastn.tsv", sep = "\t", header=None , names= en
cabezado, engine="python")
ftsv.head()
```

Guardando los datos en formato csv

In [ ]:
```
ftsv.to_csv("MA2_blastn.csv", header=True, index= None)
```

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

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "mismatch", "gapo
pen", "qstart",
             "qend", "sstart", "send", "evalue", "bitscore", "sskingdo
ms", "stitle",
             "staxids", "sscinames", "scomnames", "sblastnames")
```

In [ ]:
```
ftsv=pd.read_csv("MN2_blastn.tsv", sep = "\t", header=None , names= en
cabezado, engine="python")
ftsv.head()
```

Guardando los datos en formato csv

In [ ]:
```
ftsv.to_csv("MN2_blastn.csv", header=True, index= None)
```

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

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "mismatch", "gapo
pen", "qstart",
             "qend", "sstart", "send", "evalue", "bitscore", "sskingdo
ms", "stitle",
             "staxids", "sscinames", "scomnames", "sblastnames")
```

In [ ]:
```
ftsv=pd.read_csv("XA2_blastn.tsv", sep = "\t", header=None , names= en
cabezado, engine="python")
ftsv.head()
```

Guardando los datos en formato csv

In [ ]:
```
ftsv.to_csv("XA2_blastn.csv", header=True, index= None)
```

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

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "mismatch", "gapo
pen", "qstart",
             "qend", "sstart", "send", "evalue", "bitscore", "sskingdo
ms", "stitle",
             "staxids", "sscinames", "scomnames", "sblastnames")
```

In [ ]:
```
ftsv=pd.read_csv("XN2_blastn.tsv", sep = "\t", header=None , names= en
cabezado, engine="python")
ftsv.head()
```

Guardando los datos en formato csv

In [ ]:
```
ftsv.to_csv("XN2_blastn.csv", header=True, index= None)
```

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

Estadísticas de GC de cada ensamblaje con Soap

In [ ]:
```
cd  /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XN2
```

In [ ]:
```
ls /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XN2
```

In [ ]:
```
!head -100 XN2_out.scafStatistics
```

Cantidad de lecturas ensambladas con el SOAP

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XN2
```

In [ ]:
```
!grep -c "^>" XN2_out.contig
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XA2
```

In [ ]:
```
!grep -c "^>" XA2_out.contig
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/MN2
```

In [ ]:
```
!grep -c "^>" MN2_out.contig
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/MA2
```

In [ ]:
```
!grep -c "^>" MA2_out.contig
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CA2
```

In [ ]:
```
!grep -c "^>" CA2_out.contig
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CN2
```

In [ ]:
```
!grep -c "^>" CN2_out.contig
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CA2
```

In [ ]:
```
ls /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CA2
```

In [ ]:
```
!head -100 CA2_out.scafStatistics
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/MA2
```

In [ ]:
```
!head -100 MA2_out.scafStatistics
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XA2
```

In [ ]:
```
!head -100 XA2_out.scafStatistics
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/CN2
```

In [ ]:
```
!head -100 CN2_out.scafStatistics
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/MN2
```

In [ ]:
```
!head -100 MN2_out.scafStatistics
```

In [ ]:
```
cd /LUSTRE/bioinformatica_data/lga/edith/data/microalgas/XN2
```

In [ ]:
```
!head -100 XN2_out.scafStatistics
```