

Tutorial: 16S rRNA identification/MSA (21/Mayo/2020 y 28/Mayo/2020).

Autor: Roberto Durán

1. Identificación de bacterias mediante 16S rRNA

16S rRNA es uno de los componentes de la subunidad menor (30S) del ribosoma de bacterias y arqueas y es un gen que ha sido ampliamente utilizado en la identificación taxonómica de estos grupos desde la década del 80 debido a la descripción del SSU rRNA (small subunit rRNA) como el “último cronómetro molecular” (Woese, C.R. 1987, *Bacterial evolution. Microbiol. Rev.*). Aunque actualmente se considera que no es el mejor marcador filogenético (múltiples copias del operón rRNA, poca resolución a nivel de especie en algunas taxas), aún se sigue usando en identificación de bacterias como un primer paso a un posicionamiento filogenético de la cepa en estudio.

Actualmente se manejan algunos valores de similitud de 16S rRNA entre la cepa en estudio y las cepas tipo, siendo la cepa tipo la primera cepa que se describe para crear un nuevo taxón y que en general corresponden a una especie “**sp. nov.**” o género “**gen. nov.**” nuevo y se designan con una *t* mayúscula en superíndice en papers de taxonomía (e.g. *Hydrocarboniphaga daqingensis* sp. nov.; *Hydrocarboniphaga daqingensis* B2-9^T).

Los valores utilizados como parámetros generales son de 98.7% para especie, 94.5% para género, 86.5% para familia, 82% para orden, 78.5% para clase y 75% para filo (Yarza et al., 2014). De todas formas estos parámetros indican que si obtienen una similitud dentro de estos rangos **con seguridad** podrían decir que se trata de una taxa nueva. Esto solo es una parte del análisis taxonómico polifásico y es necesario realizar otros análisis (fenotípicos y genotípicos) para asegurarse de su identificación.

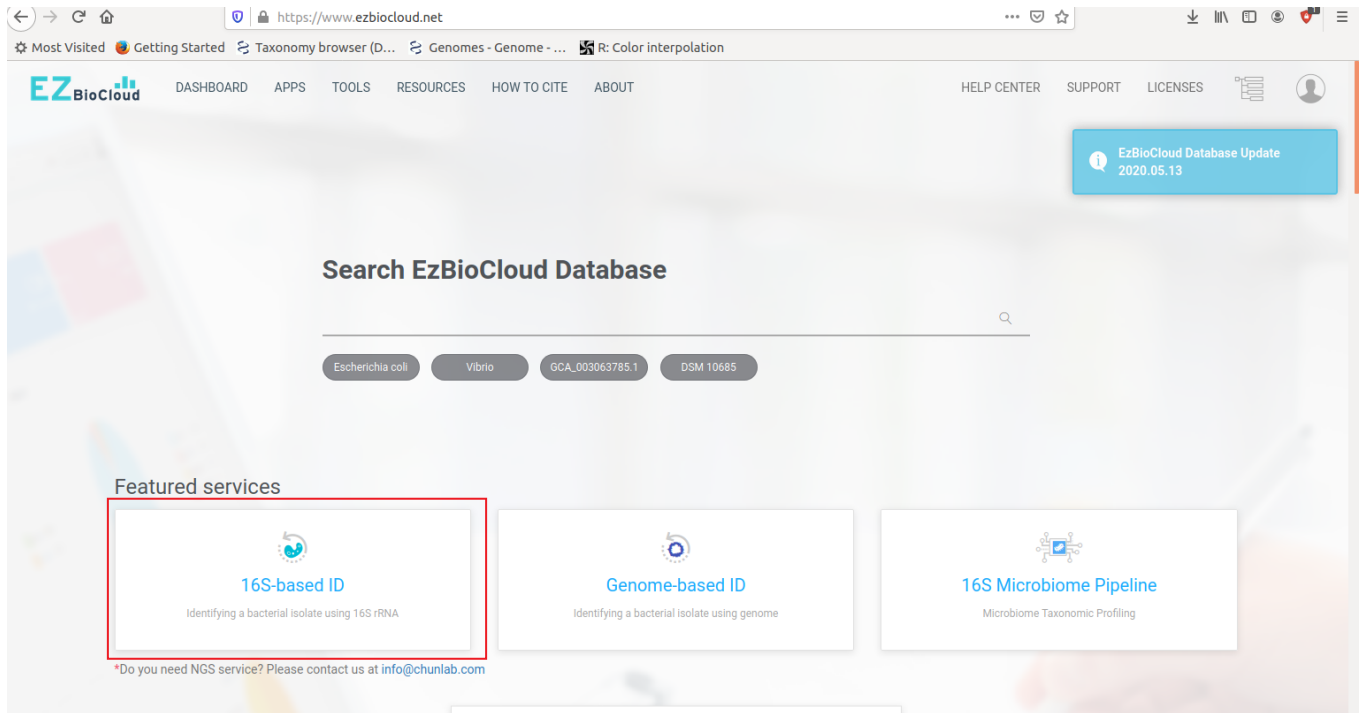
Table 1 | **Taxonomic thresholds of bacteria and archaea***

	Genus	Family	Order	Class	Phylum
Number of taxa	568	201	85	39	23
Median sequence identity	96.4% (96.2, 96.55)	92.25% (91.65, 92.9)	89.2% (88.25, 90.1)	86.35% (84.7, 87.95)	83.68% (81.6, 85.93)
Minimum sequence identity	94.8% (94.55, 95.05)	87.65% (86.8, 88.4)	83.55% (82.25, 84.8)	80.38% (78.55, 82.5)	77.43% (74.95, 79.9)
Threshold sequence identity	94.5%	86.5%	82.0%	78.5%	75.0%

*Results based on the Living Tree Project (LTP) 102 data set. Values given are the Hodges–Lehmann estimator (also known as the ‘pseudo-median’) and its 95% confidence interval (in parentheses) of all of the taxa median and minimum sequence identities for the 16S ribosomal RNA genes. Values were calculated using the Wilcoxon signed rank test (*wilcox.test*), which was implemented in the R package ‘stats’⁵³.

Como vimos en clases anteriores los alineamientos locales y globales dan diferentes resultados en cuanto a similitud (identidad) de las secuencias por lo que la obtención de similitud de 16S rRNA se requiere seguir un protocolo fijo.

Para esto se recomienda realizar un pairwise alignment utilizando un algoritmo de alineamiento global (e.g. Needleman-Wunsch o Myers-Miller) o se pueden usar bases de datos con información curada que realizan este análisis de forma automática (EzBioCloud database: <https://www.ezbiocloud.net/>).



2. Multiple Sequence Alignment (MSA)

Vamos a comparar algunos alineadores múltiples y los resultados que entregan. Para esto vamos a utilizar dos set de datos que estaban en el correo enviado la semana pasada:

Secuencias nucleotídicas del regulador OxyR: "oxyR_class.faa"

Secuencias aminoacídicas del regulador OxyR: "oxyR_class.fna"

ClustalO (<https://www.ebi.ac.uk/Tools/msa/clustalo/>):

Ingresar cada set de datos por separado, setear el output format como "NEXUS" y guardar los alineamientos en el pc como "clustal_aa_oxyR.nexus" y "clustal_nt_oxyR.nexus"

Clustal Omega

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

DNA

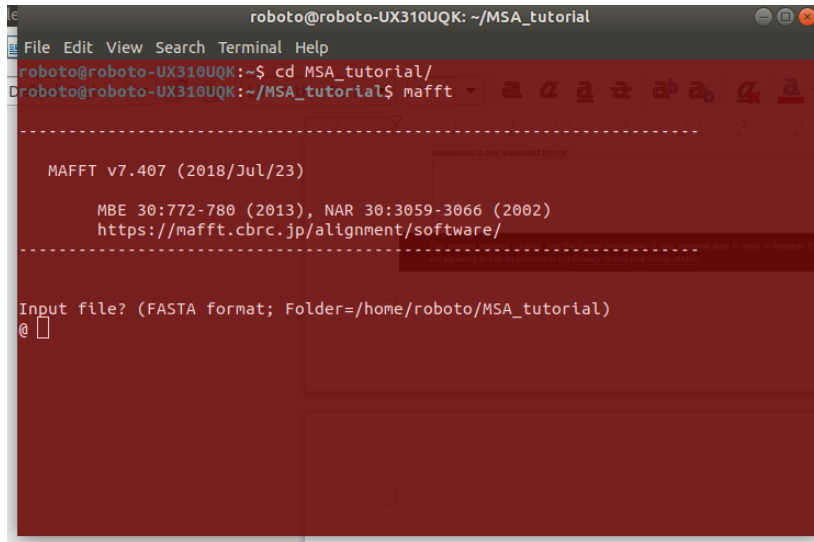
sequences in any supported format:

This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our [Privacy Notice](#) and [Terms of Use](#).

[I agree, dismiss this banner](#)

MAFFT (mafft):

Ingresa mediante la terminal a la carpeta donde contengan los set de datos e ingresa a mafft:



```
roboto@roboto-UX310UQK: ~/MSA_tutorial
File Edit View Search Terminal Help
roboto@roboto-UX310UQK:~$ cd MSA_tutorial/
roboto@roboto-UX310UQK:~/MSA_tutorial$ mafft

-----
MAFFT v7.407 (2018/Jul/23)
-----
MBE 30:772-780 (2013), NAR 30:3059-3066 (2002)
https://mafft.cbrc.jp/alignment/software/
-----

Input file? (FASTA format; Folder=/home/roboto/MSA_tutorial)
@
```

Aquí entra a la interfaz de mafft y te pide lo necesario para realizar el MSA.

1. Input file?

Aquí deben ingresar el nombre del archivo que quieran ingresar “oxyR_class.fna” o “oxyR_class.faa”.

2. Output file? Nombre del archivo de salida

mafft_nt_oxyR.phylip o mafft_aa_oxyR.phylip

3. Output format?

Se puede elegir entre clustal, nexus y philip, usaremos philip en este caso.

Opción 5

4. Strategy? Estrategia de alineamiento, cuál es el mejor algoritmo para su set de datos. Esto depende de lo que requieran y la cantidad de secuencias. Para este caso necesitamos un alineamiento global de las secuencias y más preciso (accurate) ya que son pocas secuencias.

Opción 4.

y luego se pueden ingresar otros argumentos (para otra ocasión).

Enter enter y listo!

TranslatorX (<http://translatorx.co.uk/>):

TranslatorX es un alineador de secuencias nucleotídicas que alinea las secuencias en base a la secuencia traducida en los 6 marcos de lectura abierto (ORF).

T
N
T

-
S
O

R
L
R

A
A
X

TranslatorX server
Nucleotide sequence alignment and alignment cleaning based on amino acid information.
Powered by ReadSeq, GBlocks, Jalview, Muscle, ClustalW, MAFFT, T-Coffee and PRank.
[Download local version of TranslatorX](#)

Help

Nucleotide sequences
Paste your nt-sequences below (most formats are accepted by ReadSeq)
[Example \(Mollusca_CyB genes\)](#)

or upload a file: No file selected.

Protein alignment Two alternatives are available:
☒ A. Let TranslatorX automatically compute the protein alignment.
Which method do you prefer?
☒ Muscle
☐ MAFFT
☐ ClustalW
☐ T-coffee
☐ Prank
☐ B. Provide your own alignment.

Genetic code
☒ A. Same for all taxa
Translate according to the genetic code
☐ B.1. Variable. Define it interactively.
☐ B.2. Variable. A priori definition.
☐ Guess most likely reading frame?

Alignment cleaning
☐ Remove poorly aligned sites from the protein alignment before back-translate to nucleotides?

☐ Subscribe to news and updates.

Citation: Abascal F, Zardoya R, Telford MJ (2010)

Se agrega el set de datos de nucleotidos, luego el codigo genético y luego agregar el alineamiento de proteínas de las secuencias a analizar. Aquí hay dos opciones, TranslatorX calcula el alineamiento de proteínas en base a las secuencias nucleotídicas (usando MUSCLE, MAFFT, ClustalW, T-coffee o Prank) o uno puede agregar su propio alineamiento.