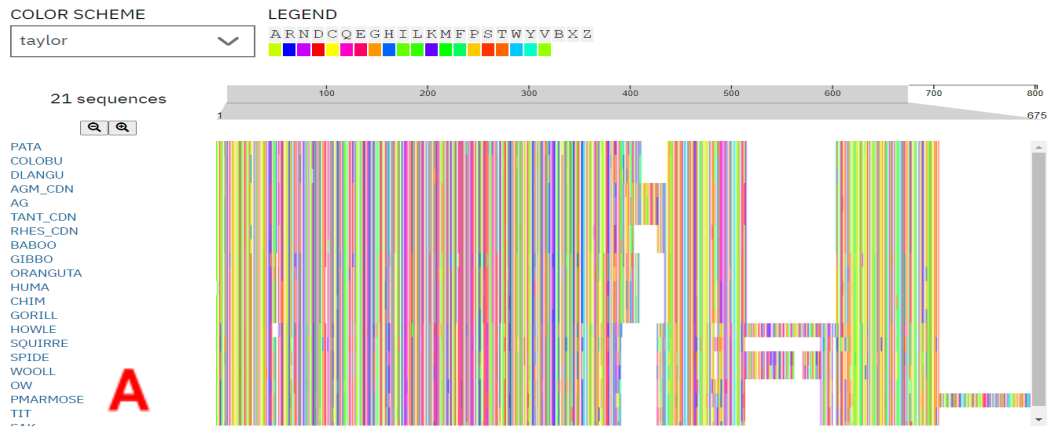
Taller 9 - Alineamientos múltiples y HMMs

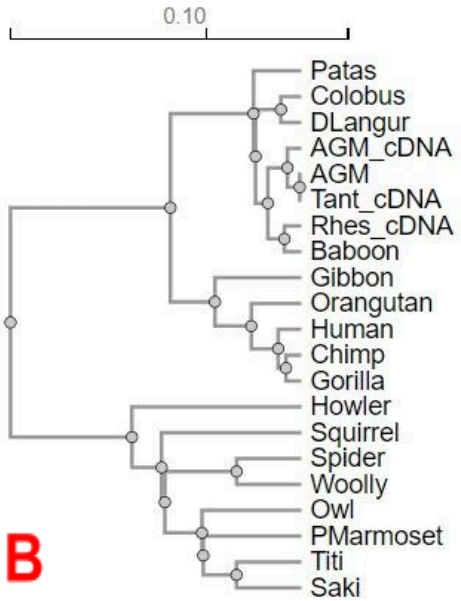
Grupo 04 - sección 02

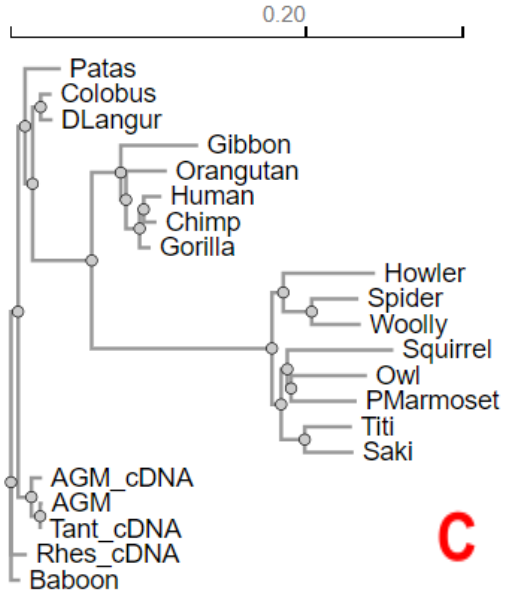
Integrantes:
- Nicolas Montoya Leon - 202310678
- John Anderson Acosta - 202212004
- Raquel Bautista Escobar – 202310296

1. Corriendo diferentes alineadores múltiples.

- Realice el alineamiento de ClustalOmega. Guarde el alineamiento y, seleccione FastA o ClustalW como formato de salida, asegúrese de cambiar el nombre del archivo de salida para evitar sobrescribir el archivo original.





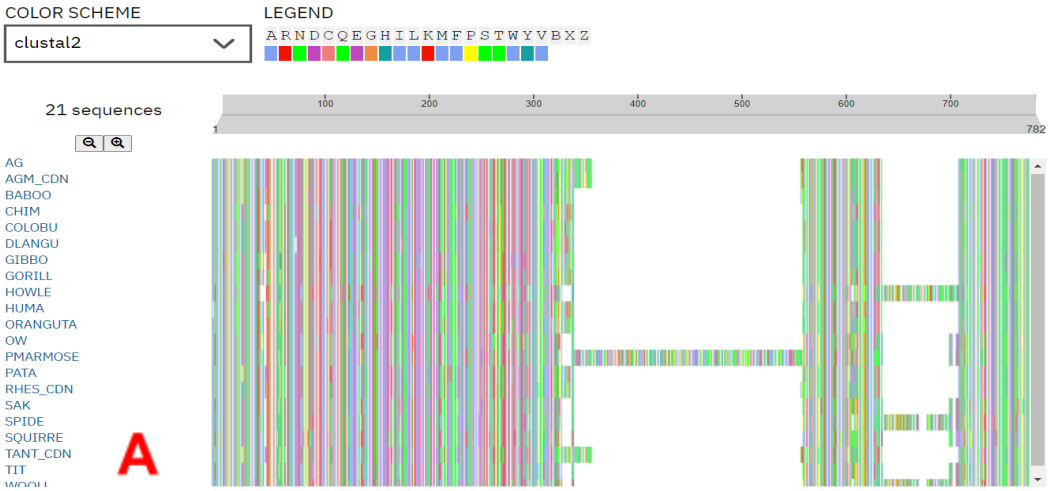


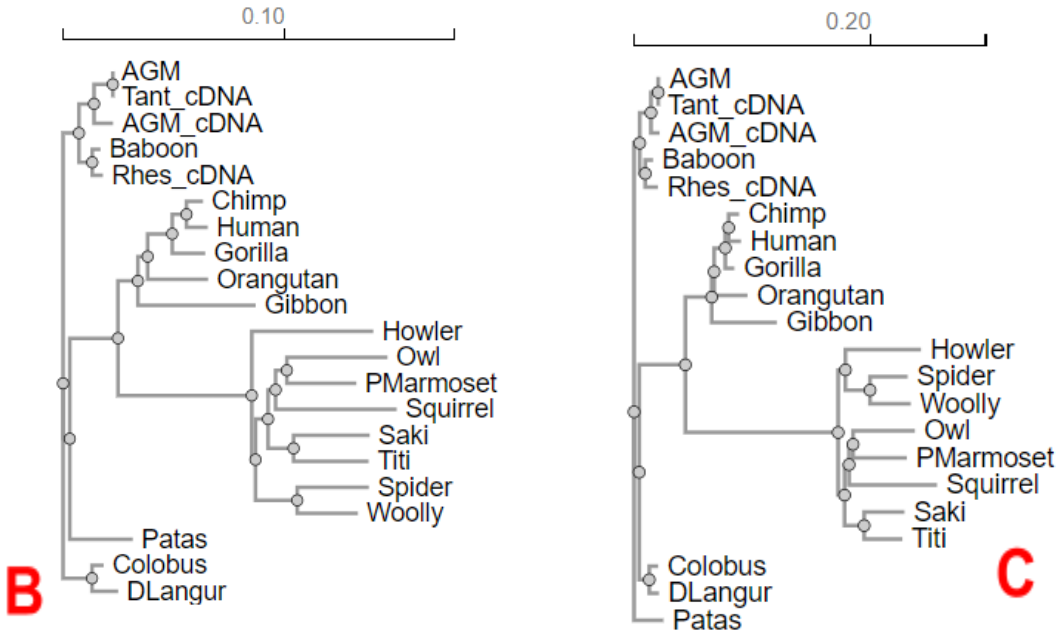
Se observan en las ***Figura A, Figura B y Figura C*** los resultados generados por EMBL-EBI para el alineamiento progresivo (ClustalW o ClustalOmega). La ***Figura A*** indica los 21 alineamientos de las secuencias aminoácidos de las distintas especies. Asimismo, se percibe que la longitud de alineamiento, con gaps o espacios incluidos, es de 802 aminoácidos. Sumado a esto, la ***Figura B*** representa el Guide Tree del alineamiento, en el que se observa que hay 3 grandes grupos o clados que se formaron. Del mismo modo, la ***Figura C*** plasma el árbol filogenético del alineamiento, hay que aclarar que este no es un árbol filogenético de las relaciones evolutivas, sino una reconstrucción de las relaciones en base de las secuencias usadas. Finalmente, estos resultados son complementarios a los del alineamiento de salida del Clustal, y pueden ser encontrados y revisados con otros resultados que analizan otros apartados en Results for Job ID: clustalo-I20240404-145644-0951-20423802-p1m en la página de EMBL-EBI.

```
CLUSTAL O(1.2.4) multiple sequence alignment


Patas          MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKKSMLYKEEERSCPVCR     60
Colobus        MASGILVNIKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYKEGERSCPVCR     60
DLangur        MASGILVNIKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYKEGERSCPVCR     60
AGM_cDNA       MASGILVNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYKEEERSCPVCR     60
AGM            MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYKEEERSCPVCR     60
Tant_cDNA      MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYKEEERSCPVCR     60
Rhes_cDNA      MASGILLNVKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYKEGERSCPVCR     60
Baboon         MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHRKSMLYKEGERSCPVCR     60
Gibbon         MASGILVNVKEKVTCPICLELLTQPLSLDCGHSFCQACLTANHKTSMPD-EGERSCPVCR     59
Orangutan      MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSTLD-KGERSCPVCR     59
Human          -MSGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLD-KGESSCPVCR     58
Chimp          MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLD-KGESSCPVCR     59
Gorilla        MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLD-KGESSCPVCR     59
Howler         MASKILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKES-----RERSCPLCR     55
Squirrel       MASRILGSIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESMLH-QGERSCPLCR     59
Spider         MASEILLNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLH-QGERSCPLCR     59
Woolly         MASEILVNIKEEVTCPICLDLLTEPLSLDCGHSFCQACITADHKESTLH-QGERSCPLCR     59
Owl            MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKKSMPH-QGERSCPLCR     59
PMarmoset      MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLH-QGERSCPLCR     59
Titi           MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLH-QGERSCPLCR     59
Saki           MASRILMNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKKSMLH-QGERSCPLCR     59
                 * ** .:**:*******:***:**** *********:**:*: *      * ***:**

Patas          ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDRKVICWLCE     120
Colobus        ISYQPENIRPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDRKVICWLCE     120
DLangur        ISYQPENIRPNRHVANIV-KLREVKLSPEEGQKVDHCARHGEKLLLFCQEDRKVICWLCE     119
AGM_cDNA       ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE     120
AGM            ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE     120
Tant_cDNA      ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE     120
Rhes_cDNA      ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE     120
Baboon         ISYQPENIQPNRHVANIVEKLREVKLSPEEGLKVDHCARHGEKLLLFCQEDSKVICWLCE     120
Gibbon         ISYQHKNIRPNRHVANIVEKLREVKLSPEEGQKVDHCARHGKKLLLFCQEDRKVICWLCE     119
Orangutan      VSYQPKNIRPNRHVANIVEKLREVKLSPE-GQKVDHCARHGEKLLLFCKEDGKVICWLCE     118
Human          ISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCARHGEKLLLFCQEDGKVICWLCE     117
Chimp          ISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCAHHGEKLLLFCQEDGKVICWLCE     118
Gorilla        ISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCARHGEKLLLFCQEDGKVICWLCE     118
Howler         VSYHSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARHGEKLLLFCQQHGNVICWLCE     115
Squirrel       LPYQSENLRPNRHLASIVERLREVMLRPEERQNVDHCARHGEKLLLFCEQDGNIICWLCE     119
Spider         VSYQSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARHGEKLLLFCQQHGNVICWLCE     119
Woolly         VGYQSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARHGEKLLLFCQQHGNVICWLCE     119
Owl            ISYSSENLRPNRHLVNIVERLREVMLSPEEGQKVDHCAHHGEKLVLFCQQDGNVICWLCE     119
PMarmoset      MSYPSENLRPNRHLANIVERLKEVMLSPEEGQKVDHCARHGEKLLLFCQQDGNVICWLCE     119
Titi           ISYPSENLRPNRHLANIVERLREVVLSPEEGQKVDLCARHGEKLLLFCQQDGNVICWLCE     119
Saki           ISYPSENLRPNRHLANIVERLREVMLSPEEGQKVDHCARHGEKLLLFCQQDGNVICWLCE     119
               : *  :*::****:..*. :*:** * **   :** **:**:**:***::. ::******

Patas          RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK     180
Colobus        RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK     180
DLangur        RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDCDK     179
AGM_cDNA       RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK     180
AGM            RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK     180
Tant_cDNA      RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK     180
Rhes_cDNA      RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK     180
Baboon         RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK     180
Gibbon         RSQEHRGHHTFLTEEVAQEYQMKLQAALQMLRQKQQEAEELEADIREEKASWKTQIQYDK     179
Orangutan      RSQEHRGHHTFLTEEVAQKYQVKLQAALEMLRQKQQEAEELEADIREEKASWKTQIQYDK     178
```

Como se observa en la imagen de arriba, este es el resultado de salida para el alineamiento de Clustal Omega, se encuentra en formato ClustalW y, como se mencionó en los otros resultados proporcionados por EMBL-EBI, género un alineamiento de 802 aminoácidos y se puede determinar que existen regiones, dominios y motivos que se conservan entre las especies para el gen TRIM5. No obstante, se observa que existen regiones en las que pocas secuencias fueron alineadas capaces de alinearse, lo que puede indicar que son regiones que solo se presentan en pocas especies o en ese organismo.

- Ahora, realice el alineamiento con T-Coffee. Cuando el procedimiento de alineamiento haya terminado será direccionado a una nueva página con enlaces a los archivos de salida. Guarde el alineamiento en formato Fasta o ClustalW en su computador.



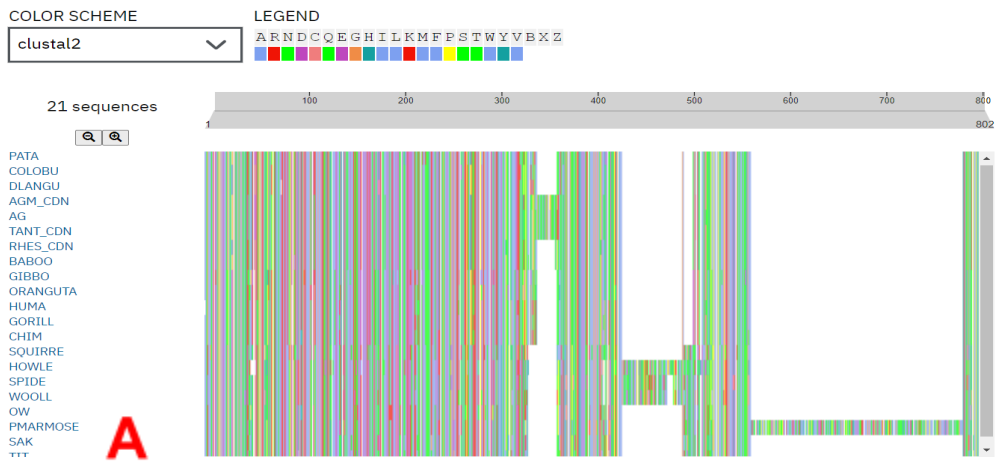


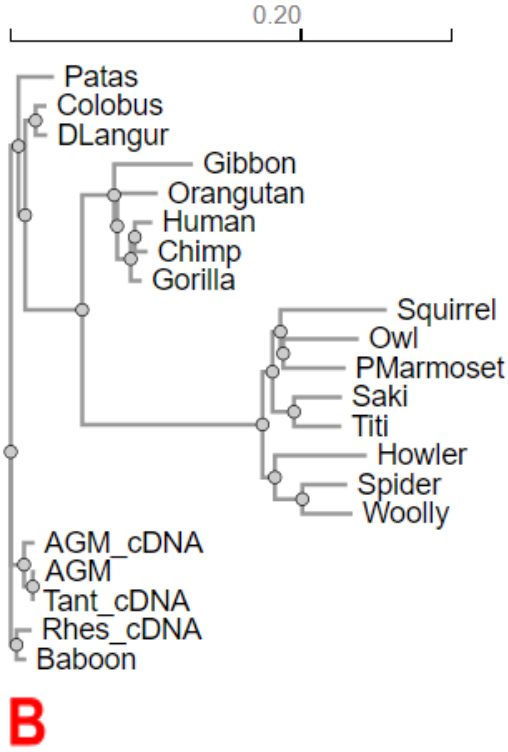
De la misma manera que el resultado de Clustal Omega, en el caso de T-Coffee las ***Figura A, Figura B y Figura C*** son resultados generados por EMBL-EBI para el alineamiento basado en consistencias. La ***Figura A*** indica los 21 alineamientos de las secuencias aminoácidos de las distintas especies, en el que contrasta con Clustal Omega, al tener una longitud de 782. La ***Figura B*** representa el Guide Tree y la ***Figura C*** el árbol filogenético del alineamiento, cabe señalar que ambos árboles generaron una relación similar entre las especies, agrupando en categorías la mayoría de especies en su mayor parte de la misma forma. Finalmente, estos resultados pueden ser encontrados y revisados con otros resultados en Results for Job ID: tcoffee-I20240404-150221-0227-7836564-p1m en la página de EMBL-EBI.

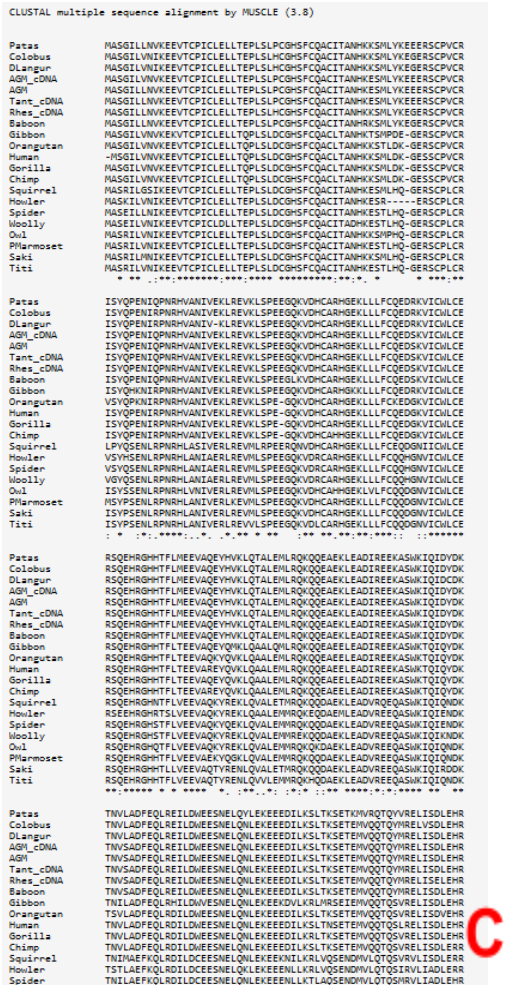
```
CLUSTAL W (1.83) multiple sequence alignment

AGM             MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYK
AGM_cDNA        MASGILVNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYK
Baboon          MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHRKSMLYK
Chimp           MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLDK
Colobus         MASGILVNIKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYK
DLangur         MASGILVNIKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYK
Gibbon          MASGILVNVKEKVTCPICLELLTQPLSLDCGHSFCQACLTANHKTSMPDE
Gorilla         MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLDK
Howler          MASKILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKES----
Human           M-SGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLDK
Orangutan       MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSTLDK
Owl             MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKKSMPHQ
PMarmoset       MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLHQ
Patas           MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKKSMLYK
Rhes_cDNA       MASGILLNVKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYK
Saki            MASRILMNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKKSMLHQ
Spider          MASEILLNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLHQ
Squirrel        MASRILGSIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESMLHQ
Tant_cDNA       MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYK
Titi            MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLHQ
Woolly          MASEILVNIKEEVTCPICLDLLTEPLSLDCGHSFCQACITADHKESTLHQ
                * * ** .:**:*******:***:**** *********:**:*: *

AGM             EEERSCPVCRISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARH
AGM_cDNA        EEERSCPVCRISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARH
Baboon          EGERSCPVCRISYQPENIQPNRHVANIVEKLREVKLSPEEGLKVDHCARH
Chimp           -GESSCPVCRISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCAHH
Colobus         EGERSCPVCRISYQPENIRPNRHVANIVEKLREVKLSPEEGQKVDHCARH
DLangur         EGERSCPVCRISYQPENIRPNRHVANIV-KLREVKLSPEEGQKVDHCARH
Gibbon          -GERSCPVCRISYQHKNIRPNRHVANIVEKLREVKLSPEEGQKVDHCARH
Gorilla         -GESSCPVCRISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCARH
Howler          -RERSCPLCRVSYHSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARH
Human           -GESSCPVCRISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCARH
Orangutan       -GERSCPVCRVSYQPKNIRPNRHVANIVEKLREVKLSPE-GQKVDHCARH
Owl             -GERSCPLCRISYSSENLRPNRHLVNIVERLREVMLSPEEGQKVDHCAHH
PMarmoset       -GERSCPLCRMSYPSENLRPNRHLANIVERLKEVMLSPEEGQKVDHCARH
Patas           EEERSCPVCRISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARH
Rhes_cDNA       EGERSCPVCRISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARH
Saki            -GERSCPLCRISYPSENLRPNRHLANIVERLREVMLSPEEGQKVDHCARH
Spider          -GERSCPLCRVSYQSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARH
Squirrel        -GERSCPLCRLPYQSENLRPNRHLASIVERLREVMLRPEERQNVDHCARH
Tant_cDNA       EEERSCPVCRISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARH
Titi            -GERSCPLCRISYPSENLRPNRHLANIVERLREVVLSPEEGQKVDLCARH
Woolly          -GERSCPLCRVGYQSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARH
                  * ***:**: *  :*::****:..*. :*:** * **   :** **:*

AGM             GEKLLLFCQEDSKVICWLCERSQEHRGHHTFLMEEVAQEYHVKLQTALEM
AGM_cDNA        GEKLLLFCQEDSKVICWLCERSQEHRGHHTFLMEEVAQEYHVKLQTALEM
Baboon          GEKLLLFCQEDSKVICWLCERSQEHRGHHTFLMEEVAQEYHVKLQTALEM
Chimp           GEKLLLFCQEDGKVICWLCERSQEHRGHHTFLTEEVAREYQVKLQAALEM
Colobus         GEKLLLFCQEDRKVICWLCERSQEHRGHHTFLMEEVAQEYHVKLQTALEM
DLangur         GEKLLLFCQEDRKVICWLCERSQEHRGHHTFLMEEVAQEYHVKLQTALEM
Gibbon          GKKLLLFCQEDRKVICWLCERSQEHRGHHTFLTEEVAQEYQMKLQAALQM
Gorilla         GEKLLLFCQEDGKVICWLCERSQEHRGHHTFLTEEVAQEYQVKLQAALEM
Howler          GEKLLLFCQQHGNVICWLCERSEEHRGHRTSLVEEVAQKYREKLQAALEM
Human           GEKLLLFCQEDGKVICWLCERSQEHRGHHTFLTEEVAREYQVKLQAALEM
Orangutan       GEKLLLFCKEDGKVICWLCERSQEHRGHHTFLTEEVAQKYQVKLQAALEM
Owl             GEKLVLFCQQDGNVICWLCERSQEHRGHQTFLVEEVAQKYREKLQVALEM
PMarmoset       GEKLLLFCQQDGNVICWLCERSQEHRGHHTFLVEEVAEKYQGKLQVALEM
Patas           GEKLLLFCQEDRKVICWLCERSQEHRGHHTFLMEEVAQEYHVKLQTALEM
Rhes_cDNA       GEKLLLFCQEDSKVICWLCERSQEHRGHHTFLMEEVAQEYHVKLQTALEM
Saki            GEKLLLFCQQDGNVICWLCERSQEHRGHHTLLVEEVAQTYRENLQVALET
Spider          GEKLLLFCQQHGNVICWLCERSQEHRGHSTFLVEEVAQKYQEKLQVALEM
Squirrel        GEKLLLFCEQDGNIICWLCERSQEHRGHNTFLVEEVAQKYREKLQVALET
Tant_cDNA       GEKLLLFCQEDSKVICWLCERSQEHRGHHTFLMEEVAQEYHVKLQTALEM
Titi            GEKLLLFCQQDGNVICWLCERSQEHRGHHTFLVEEVAQTYRENLQVVLEM
Woolly          GEKLLLFCQQHGNVICWLCERSQEHRGHSTFLVEEVAQKYREKLQVALEM
                *:**:***::. ::********:***** * * ****. *: :**..*:

AGM             LRQKQQEAEKLEADIREEKASWKIQIDYDKTNVSADFEQLREILDWEESN
AGM_cDNA        LRQKQQEAEKLEADIREEKASWKIQIDYDKTNVSADFEQLREILDWEESN
```

Por último, la última imagen representa el archivo de salida para el alineamiento de T-Coffee, se encuentra en formato ClustalW, posee un alineamiento de 782 aminoácidos y se puede concluir que existen 3 regiones o dominios que se presentan entre las especies para el gen TRIM5.

- Finalmente, realice el alineamiento en MUSCLE. Cargue su archivo de secuencias usando la opción Upload a file. Asegúrese que la salida esté en formato FastA o ClustalW.







Para el alineamiento en MUSCLE se obtuvieron los siguientes resultados ***Figura A, Figura B y Figura C*** generados por EMBL-EBI. La ***Figura A*** representa los 21 alineamientos múltiples de las secuencias aminoácidos de las distintas especies para un alineamiento iterativo. Igualmente, se percibe que el tamaño del alineamiento es de 802 aminoácidos. La ***Figura B*** muestra el árbol filogenético del alineamiento en MUSCLE, siendo muy similar al generado por Clustal Omega. Por último, la ***Figura C*** es el archivo de salida para el alineamiento de MUSCLE, se encuentra, al igual que los otros alineamientos, en formato ClustalW y de que hay 4 regiones o dominios para la secuencia en MUSCLE entre las especies para el gen TRIM5.

Bono [0.25 puntos] Realice el alineamiento desde el cluster con la línea de comandos (debe llamar el módulo muscle y el comando muscle, revise el manual para saber qué argumentos usar)

```
[biol2205@hypatia Grupo-04]$ module load muscle/5.1
[biol2205@hypatia Grupo-04]$ module list

Currently Loaded Modules:
  1) anaconda/conda4.12.0   2) muscle/5.1

[biol2205@hypatia Grupo-04]$
```

A

```
Basic usage

    muscle -in <inputfile> -out <outputfile>

Common options (for a complete list please see the User Guide):

    -in <inputfile>    Input file in FASTA format (default stdin)
    -out <outputfile>  Output alignment in FASTA format (default stdout)
    -diags             Find diagonals (faster for similar sequences)
    -maxiters <n>      Maximum number of iterations (integer, default 16)
    -maxhours <h>      Maximum time to iterate in hours (default no limit)
    -html              Write output in HTML format (default FASTA)
    -msf               Write output in GCG MSF format (default FASTA)
    -clw               Write output in CLUSTALW format (default FASTA)
    -clwstrict         As -clw, with 'CLUSTAL W (1.81)' header
    -log[a] <logfile>  Log to file (append if -loga, overwrite if -log)
    -quiet             Do not write progress messages to stderr
    -version           Display version information and exit
```

B

```
[biol2205@hypatia Taller-09]$ muscle -in TRIM5.fasta -clwout TRIM5.clw

MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

TRIM5 21 seqs, lengths min 492, max 709, avg 515
00:00:00    20 MB(-1%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00    20 MB(-1%)  Iter   1  100.00%  K-mer dist pass 2
00:00:00    29 MB(-2%)  Iter   1  100.00%  Align node
00:00:00    29 MB(-2%)  Iter   1  100.00%  Root alignment
00:00:00    29 MB(-2%)  Iter   2  100.00%  Refine tree
00:00:00    29 MB(-2%)  Iter   2  100.00%  Root alignment
00:00:00    29 MB(-2%)  Iter   2  100.00%  Root alignment
00:00:00    29 MB(-2%)  Iter   3  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter   4  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter   5  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter   6  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter   7  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter   8  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter   9  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  10  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  11  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  12  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  13  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  14  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  15  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  16  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  17  100.00%  Refine biparts
00:00:00    29 MB(-2%)  Iter  18  100.00%  Refine biparts
00:00:01    29 MB(-2%)  Iter  19  100.00%  Refine biparts
00:00:01    29 MB(-2%)  Iter  20  100.00%  Refine biparts
[biol2205@hypatia Taller-09]$
```

C

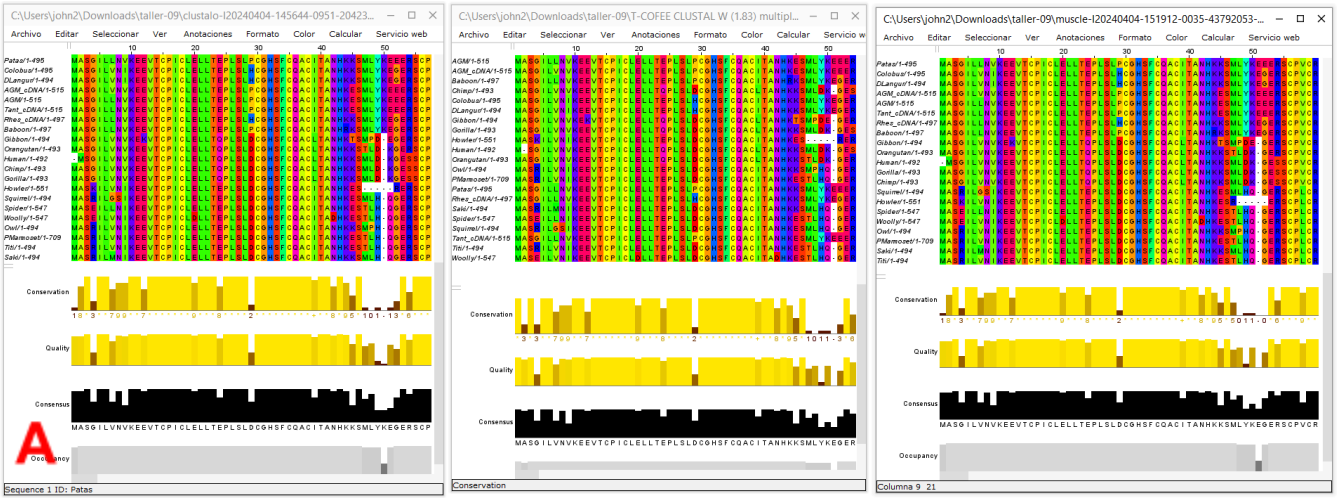
```
[biol2205@hypatia Taller-09]$ cat TRIM5.clw
MUSCLE (3.8) multiple sequence alignment


Patas           MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKKSMLYKEEERSCPVCR
Colobus         MASGILVNIKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYKEGERSCPVCR
DLangur         MASGILVNIKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYKEGERSCPVCR
AGM_cDNA        MASGILVNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYKEEERSCPVCR
AGM             MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYKEEERSCPVCR
Tant_cDNA       MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHKESMLYKEEERSCPVCR
Rhes_cDNA       MASGILLNVKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYKEGERSCPVCR
Baboon          MASGILLNVKEEVTCPICLELLTEPLSLPCGHSFCQACITANHRKSMLYKEGERSCPVCR
Gibbon          MASGILVNVKEKVTCPICLELLTQPLSLDCGHSFCQACLTANHKTSMPDE-GERSCPVCR
Orangutan       MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSTLDK-GERSCPVCR
Human           -MSGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLDK-GESSCPVCR
Gorilla         MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLDK-GESSCPVCR
Chimp           MASGILVNVKEEVTCPICLELLTQPLSLDCGHSFCQACLTANHKKSMLDK-GESSCPVCR
Squirrel        MASRILGSIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESMLHQ-GERSCPLCR
Howler          MASKILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESR-----ERSCPLCR
Spider          MASEILLNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLHQ-GERSCPLCR
Woolly          MASEILVNIKEEVTCPICLDLLTEPLSLDCGHSFCQACITADHKESTLHQ-GERSCPLCR
Owl             MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKKSMPHQ-GERSCPLCR
PMarmoset       MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLHQ-GERSCPLCR
Saki            MASRILMNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKKSMLHQ-GERSCPLCR
Titi            MASRILVNIKEEVTCPICLELLTEPLSLDCGHSFCQACITANHKESTLHQ-GERSCPLCR
                  * ** .:**:*******:***:**** *********:**:*. *      * ***:**

Patas           ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDRKVICWLCE
Colobus         ISYQPENIRPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDRKVICWLCE
DLangur         ISYQPENIRPNRHVANIV-KLREVKLSPEEGQKVDHCARHGEKLLLFCQEDRKVICWLCE
AGM_cDNA        ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE
AGM             ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE
Tant_cDNA       ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE
Rhes_cDNA       ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE
Baboon          ISYQPENIQPNRHVANIVEKLREVKLSPEEGLKVDHCARHGEKLLLFCQEDSKVICWLCE
Gibbon          ISYQHKNIRPNRHVANIVEKLREVKLSPEEGQKVDHCARHGKKLLLFCQEDRKVICWLCE
Orangutan       VSYQPKNIRPNRHVANIVEKLREVKLSPE-GQKVDHCARHGEKLLLFCKEDGKVICWLCE
Human           ISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCARHGEKLLLFCQEDGKVICWLCE
Gorilla         ISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCARHGEKLLLFCQEDGKVICWLCE
Chimp           ISYQPENIRPNRHVANIVEKLREVKLSPE-GQKVDHCAHHGEKLLLFCQEDGKVICWLCE
Squirrel        LPYQSENLRPNRHLASIVERLREVMLRPEERQNVDHCARHGEKLLLFCEQDGNIICWLCE
Howler          VSYHSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARHGEKLLLFCQQHGNVICWLCE
Spider          VSYQSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARHGEKLLLFCQQHGNVICWLCE
Woolly          VGYQSENLRPNRHLANIAERLREVMLSPEEGQKVDRCARHGEKLLLFCQQHGNVICWLCE
Owl             ISYSSENLRPNRHLVNIVERLREVMLSPEEGQKVDHCAHHGEKLVLFCQQDGNVICWLCE
PMarmoset       MSYPSENLRPNRHLANIVERLKEVMLSPEEGQKVDHCARHGEKLLLFCQQDGNVICWLCE
Saki            ISYPSENLRPNRHLANIVERLREVMLSPEEGQKVDHCARHGEKLLLFCQQDGNVICWLCE
Titi            ISYPSENLRPNRHLANIVERLREVVLSPEEGQKVDLCARHGEKLLLFCQQDGNVICWLCE
                : *  :*:.****:..*. .*.** * **   :** **.**:**:***::  ::***:**

Patas           RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK
Colobus         RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDYDK
```

D

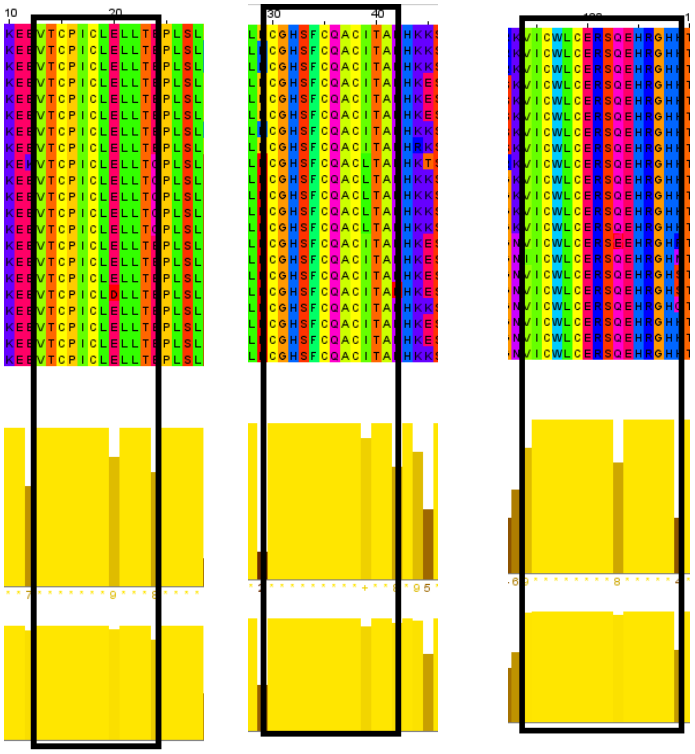
Para poder llevar a cabo el alineamiento desde el cluster fue necesario primero cargar el módulo muscle/5.1 ***Figura A.*** Posteriormente, ver los parámetros necesarios del comando muscle, para esto se apoyó usando el comando -help como se observa en la ***Figura B.*** Después, se seleccionó el parámetro -in para introducir el archivo de entrada, que en este caso fue TRIM5.fasta, sumado a esto se usó el parámetro -clwout, para señalar que el archivo de salido es un Clustalw y no fasta. De igual forma, se agregó al archivo de salida el nombre TRIM5.clw como se observa en la ***Figura C***. Finalmente se observa el archivo de salida, en el que primeramente resalta su gran similitud con el archivo de salida del alineamiento de MUSCLE en el punto anterior ***Figura D***.

2. Comparando alineamientos múltiples.

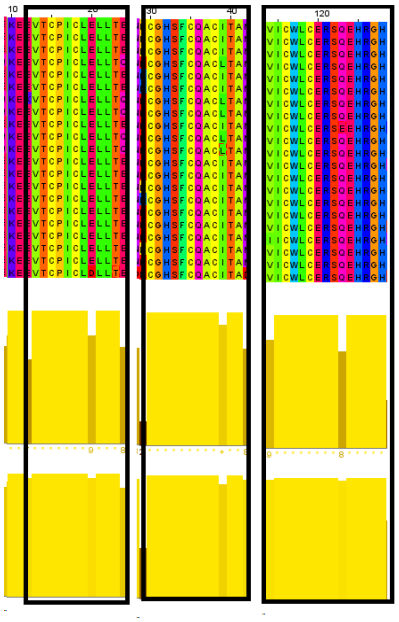
- Usando JalView para comparar los resultados obtenidos en la parte A de ClustalOmega, T-Coffee y MUSCLE, mencione dónde comienzan y acaban las primeras tres regiones conservadas para cada algoritmo; indique las posiciones y qué tan largos son. Asuma que las regiones conservadas son de mínimo 10 aminoácidos de longitud.



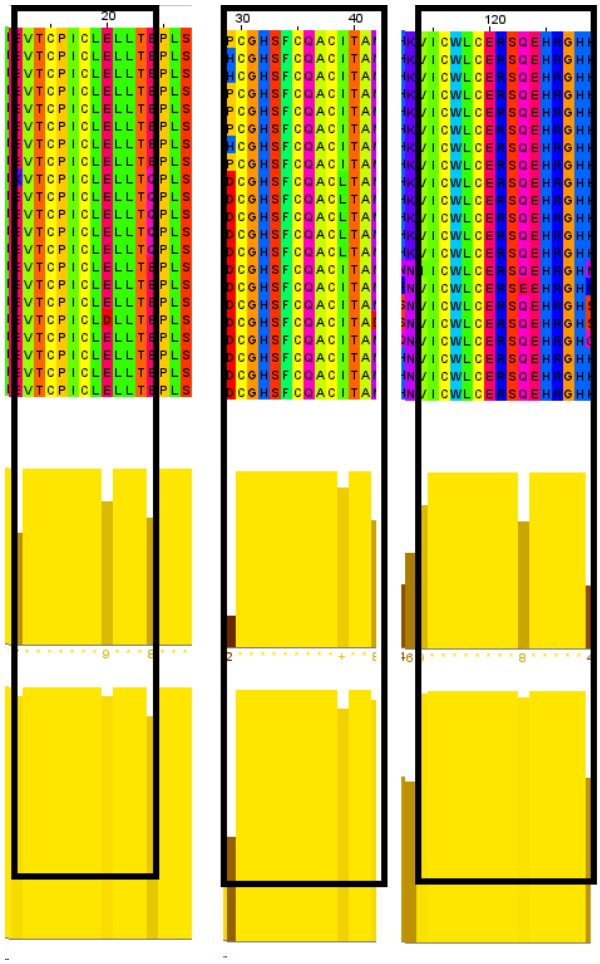
En primer lugar se analizó las secuencias de forma separada, cada una en única ventana, es decir, en vez de comparar cada alineamiento por conjunto**,** se optó por comparar los tres algoritmos en su propio alineamiento, como se señala en la ***Figura A***. En segundo lugar el método de comparación fue observar la calidad del alineamiento para decidir si es una región conservada, debido a que hay aminoácidos pudieron deberse a mutaciones pero que pertenecen aún al mismo grupo, por ejemplo básicos, ácidos, polares, apolares, etc.



La primera región conservada en Clustal Omega va de la posición 13 a la posición 23, teniendo una longitud de 11 aminoácidos seguidos, siendo una única secuencia en la especie Woolly monkey la que no corresponde en el alineamiento, en este caso la especie tiene el ácido aspártico (D) en vez del ácido glutámico (E). La segunda región va de la posición 30 a la posición 41 y posee 12 aminoácidos de longitud, sobre este caso se tiene que cuatro especies difieren del aminoácido isoleucina (I) por una Leucina (L). Por último, la última región conservada va de la posición 115 a la posición 128 y contiene 14 aminoácidos, al igual que los anteriores regiones, hay un cambio en la especie Howler del aminoácido Glutamina(Q) por el Ácido glutámico(E). Sin el caso de que fuera estrictamente una región seguida para todas las especies, sólo habría una región conservada en la posición 378 hasta la posición 388.



La primera región conservada en T-Coffee va de la posición 13 a la posición 23, teniendo una longitud de 11 aminoácidos seguidos, siendo una única secuencia en la especie Woolly monkey la que no corresponde en el alineamiento, al igual que en Cluster Omega. La segunda región va de la posición 30 a la posición 41 y posee 12 aminoácidos de longitud, sobre este caso se tiene en este caso cinco especies difieren del aminoácido isoleucina (I) por una Leucina (L). Por último, la última región conservada, al igual que las otras dos, va de la posición 115 a la posición 128 y contiene 14 aminoácidos, una gran similitudes que comparte con Cluster Omega. No obstante, en algo que contrasta es que en el caso de que fuera estrictamente una región seguida para todas las especies, sólo habría una región conservada en la posición 573 hasta la posición 582, siendo de una región de longitud de 10 aminoácidos.



Al igual que en las otras tres secuencias anteriores, el alineamiento al usar MUSCLE tambien posee 3 regiones conservadas en las posiciones 13 a la posición 23, en la posición la posición 30 a la posición 41 y en la posición posición 115 a la posición 128, teniendo las mismas causas de los mismatches presentes en las especies iguales. Algo que tiene relación con Cluster Omega y que contrasta con T-Coffee, es que posee en el caso de que fuera estrictamente una región

seguida de aminoácidos para todas las especies, sólo habría una región conservada en la posición 378 hasta la posición 388.

- ¿Qué modificaciones se podrían hacer a los alineamientos? ¿Hay alguna(s) secuencia(s) que se podría(n) eliminar?

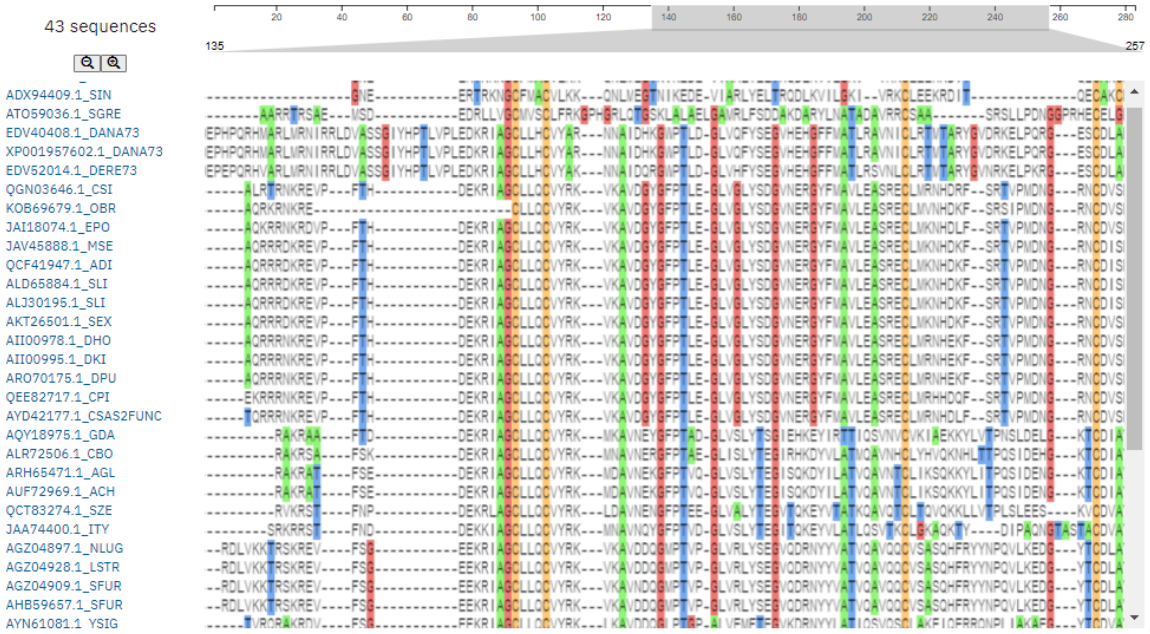
Como se ha mencionado anteriormente, la secuencias Woolly monkey y Howler generan mismatches en el momento de alinear las 21 secuencias de las especies, algo que si se modificara o se eliminará produciría un efecto en que las regiones conservadas fueron para todas las especies sin ninguna excepción y teniendo una conservación del 100% para todo el alineamiento múltiples para cada uno de los tres algoritmos.

- ¿Qué métodos de alineamientos presentan las mayores diferencias entre sí? ¿Qué métodos presentan las menores? Explique basado en la metodología de cada algoritmo sus resultados.

Los alineamientos que poseen las mayores diferencias serían Cluster Omega con T-Coffee y MUSCLE con T-coffee, y del mismo modo, los alineamientos que presentan las menores diferencias es MUSCLE y Cluster Omega, esto se podria explicar debido a que tanto Cluster Omega como MUSCLE son alineamientos progresivos o iterados, un alineamiento muy similar que no tienen muchas diferencias en reads muy cortos, al contrario de T-Coffee basado en consistencias.

3. Perfiles de HMMs y homologías.

- Busque en la carpeta ~/Talleres/Taller-04 el archivo de nombre Group_11.fasta. Realice un alineamiento múltiple con MUSCLE, guarde el archivo en formato fasta. El alineamiento se puede hacer en la herramienta web del EBI o en el módulo en el clúster.



- Construya el HMM para nuestras secuencias problema Group_11.fasta, estas serán usadas como señuelo para atrapar los receptores olfativos de interés en el grupo generalista de receptores. Esto debe hacerlo desde el clúster llamando el módulo hmmer y usando el comando con los siguientes argumentos.

hmmbuild perfil.hmm alineamiento.fasta

```
[biol2205@hypatia Grupo-04]$ hmmbuild Group11.hmm Group_11_aligned.fasta
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.4 (Aug 2023); http://hmmer.org/
# Copyright (C) 2023 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             Group_11_aligned.fasta
# output HMM file:                  Group11.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                 nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     Group_11_aligned        43   283   185     1.33  0.591

# CPU time: 0.17u 0.00s 00:00:00.17 Elapsed: 00:00:00.19
```

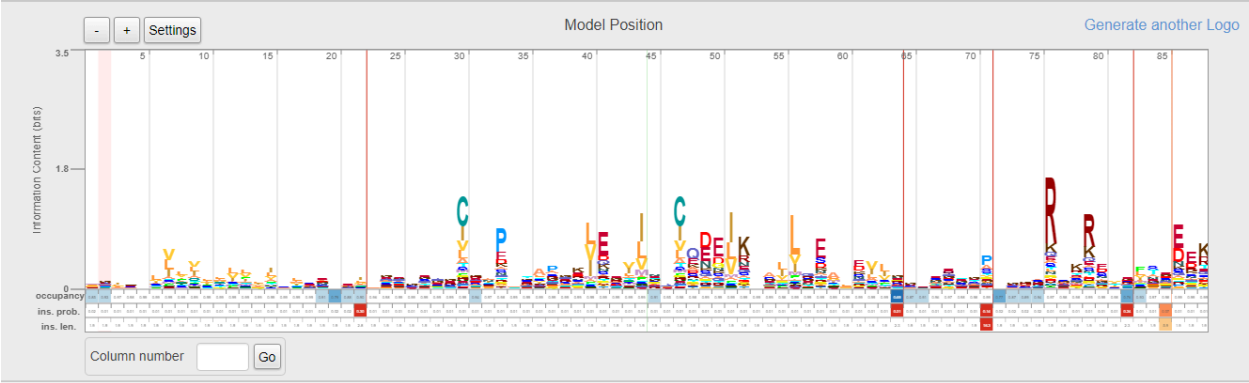
- Ahora realizaremos la búsqueda en la base de datos general (OBP_ALL_1.fasta). Para eso use el siguiente comando:

hmmsearch perfil.hmm base_datos.fasta > salida search.out

Los resultados restantes tendrán tres partes, donde la primera parte contiene aquellas secuencias que fueron rescatadas y sus respectivos mejores dominios. De acuerdo con lo que conoce sobre el e-value, ¿piensa que los resultados son confiables?

```
Internal pipeline statistics summary:
-------------------------------------
Query model(s):                            1  (185 nodes)
Target sequences:                       2736  (431750 residues searched)
Passed MSV filter:                      2019  (0.737939); expected 54.7 (0.02)
Passed bias filter:                     1526  (0.557749); expected 54.7 (0.02)
Passed Vit filter:                       904  (0.330409); expected 2.7 (0.001)
Passed Fwd filter:                       451  (0.164839); expected 0.0 (1e-05)
Initial search space (Z):               2736  [actual number of targets]
Domain search space  (domZ):             451  [number of targets reported over threshold]
# CPU time: 1.11u 0.02s 00:00:01.13 Elapsed: 00:00:00.70
```

- Ahora para la visualización de los HMMs se hará en forma de logos. Esto puede realizarse en el enlace Skylight. Use la aplicación y genere el logo correspondiente para perfil generado con HMMer (perfil.hmm). Explique ¿Qué significa un logo? ¿Qué información puedo extraer de él?



Los logos son representaciones gráficas de la información contenida en perfiles HMM; estos muestran la probabilidad de ocurrencia de cada aminoácido o nucleótido en cada posición del alineamiento múltiple. Las posiciones del alineamiento se representan en columnas, donde la altura de las letras en cada columna indica la frecuencia relativa de cada aminoácido o nucleótido en esa posición.

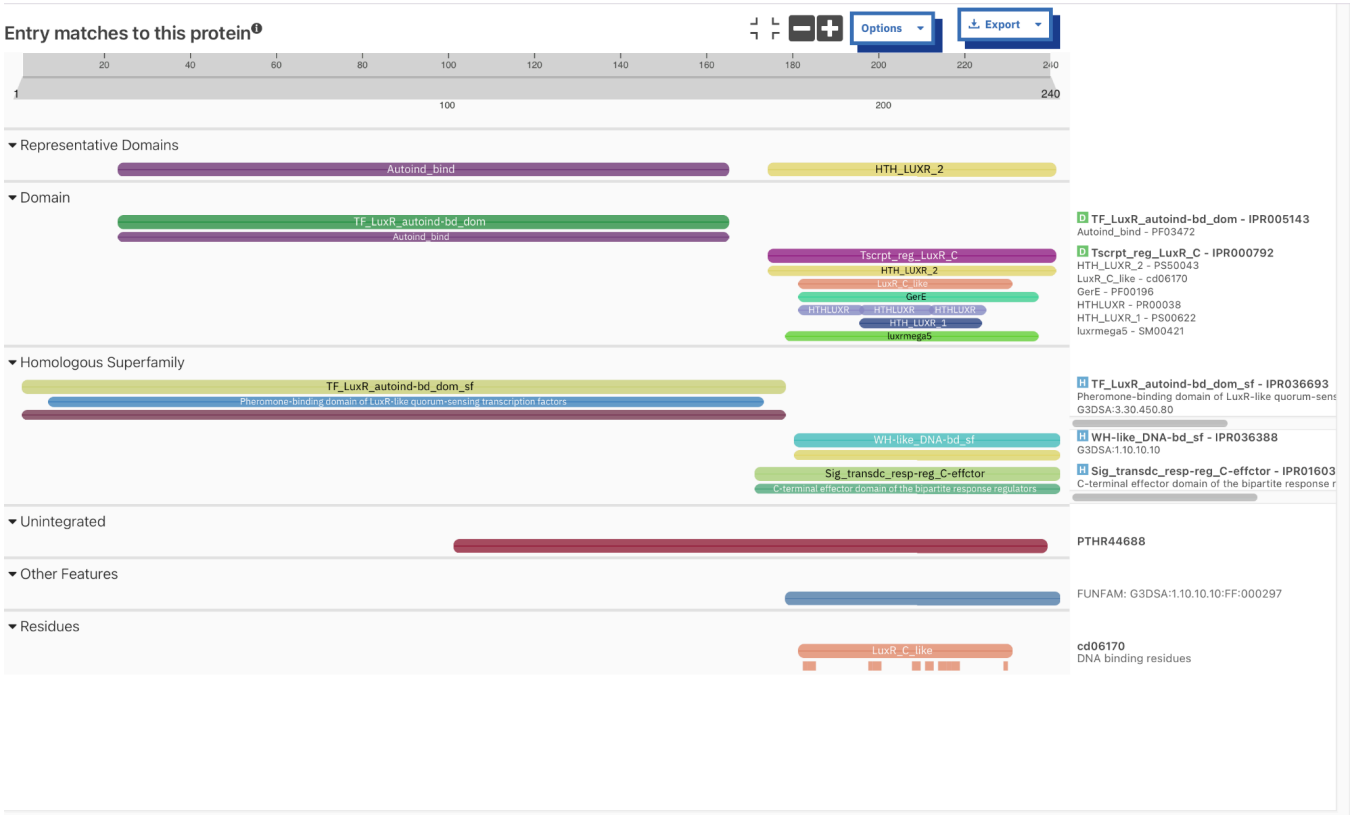
Mediante los logos es posible obtener información sobre la conservación de las diferentes regiones de las secuencias alineadas, siendo aquellas que tienen letras más altas, las más conservadas dado que dicho nucleótido se repite más.

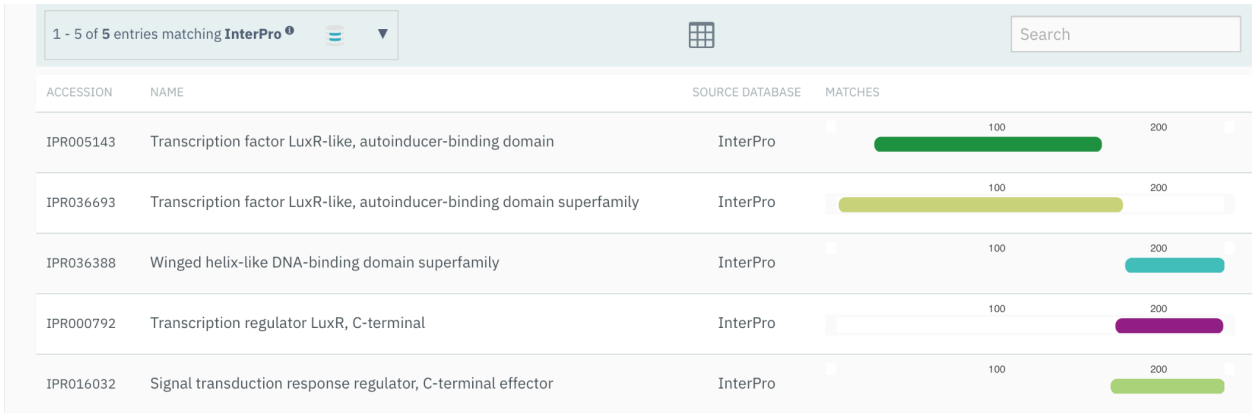
- Describa biológicamente, qué significa este resultado con respecto a la pregunta biológica sobre los receptores olfativos de los insectos.

A partir del análisis tanto del alineamiento múltiple, como del logo obtenido del archivo .hmm, se puede evidenciar la presencia de múltiples regiones homólogas; lo cual es un indicativo positivo para existencia de regiones olfativas específicas dentro de grupo de insectos, que les permitirían detectar el olor a flores.

4. Analizando dominios y familias de proteínas.

Descargue el archivo RHLR.fasta del directorio ~/Talleres/Taller-09 del cluster y utilice InterPro para identificar todos los dominios o motivos que tenga esta proteína. ¿Cuántos dominios y superfamilias homólogas encontró? ¿En qué posiciones? ¿Qué funciones reportadas poseen estas regiones?





Se encontraron 11 dominios, de los cuales los más representativos son Autoind_bind y HTH_LUXR_2 en las posiciones 25-164 y 174-239. Por otra parte, se presentan 7 superfamilias homólogas, en las posiciones 3-177 y 171-240-

Funciones reportadas:

- Procesos biológicos: regulación de la transcripción con plantilla de ADN (GO:0006355)
- Función molecular: vinculante de ADN (GO:0003677)

https://www.ebi.ac.uk:443/interpro//result/InterProScan/iprscan5-R20240410-153327-0112-94670255-p1m/