Tarea3 Relación secuencia-estructura terciaria de proteínas, y su efecto sobre los alineamientos.

1) Selecciona una superfamilia de proteínas de SCOP(http://scop.berkeley.edu)y extrae la secuencia de aminoácidos (ATOM records) y las coordenadas PDB de variosdominios de la misma.

- **Superfamilia: Fold a.3: Cytochrome c**

Detalles de las 3 estructuras elegidas:

###########
Details for **d2paca**_
PDB Entry: 2pac
PDB Description: solution structure of fe(ii) cytochrome c551 from pseudomonas aeruginosa as determined by two-dimensional 1h nmr
PDB Compounds: (A:) cytochrome c551
###########
Details for **d1j3sa**_
PDB Entry: 1j3s
PDB Description: Solution Structure of Reduced Recombinant Human Cytochrome c
PDB Compounds: (A:) cytochrome c
###########
Details for **d1e8ea**_
PDB Entry: 1e8e
PDB Description: solution structure of methylophilusmethylotrophus cytochrome c". insights into the structural basis of haem-ligand detachment
PDB Compounds: (A:) cytochrome c"
###########

2) Comprueba que las secuencias descargadas coinciden con las coordenadas.

Para comprobar que las secuencias descargadas coincidan con las coordenadas se empleó el siguiente script en bash:

```
#Las siguienteslíneas son para obtener la estructuraprimaria de las
proteínas a partirdelarchivodescargado .pdb (loscualesfueroneditados,
borrándoles las primeraslíneasmanualmente para sólodejar lascolumnas),
el resultadoestaráencódigo de aminoacidosen 3 letras.

awk '{print $4,$6}'  human_coo_cut_pre_comprobation | sed '$d' | uniq |
awk '{print $1}' | tr '\n' ' ' | sed 's/ //g'>human_colapsed
awk '{print $4,$6}'  Pseudomonas_aeruginosa_pre_comprobation | sed '$d'
| uniq | awk '{print $1}' | tr '\n' ' ' | sed 's/ //g'
>Pseudomona_colapsed
awk '{print $4,$6}'  methylophilus_methylotrophus_pre_comprobation | sed
'$d' | uniq | awk '{print $1}' | tr '\n' ' ' | sed 's/ //g'
>methy_colapsed

#Se usó el siguiente script de Python:
http://pldserver1.biochem.queensu.ca/~rlc/work/scripts/seq_convert.py
para convertir las secuencias de aminoácidos de 3 letras a 1.

#Acontinuación se extraesólo la secuencia de aminoácidos de losarchivos
.fastadescargados de SCOP.

sed 's/.*/\U&/' d2paca_\(pseudomona_aureuginosa\).fa  >UPPERdpseudo.fa
sed 's/.*/\U&/' d1j3sa_\(human\).fa > UPPERd1j3sa_\(human\).fa
sed 's/.*/\U&/' d1e8ea_\(methylophilus\ methylotrophus\).fa  >UPPERmethy
#Porútimo se comparan las secuencias de aminoácidosobtenidos de los .pdb
y de los .fastaobservando que son idénticasentodosloscasos.
diff -s UPPERdpseudo.fa Pseudomona_colapsed_3to1
diff -s human_colapsed_3to1 UPPERd1j3sa_\(human\).fa
```

```
diff -s methy_colapsed_3to1 UPPERmethy
```

Las secuencias sí coinciden

3) Calcula al menos dos alineamientos pareados entre secuencias de aminoácidos de las extraídas en 1 y calcula su %identidad como el total de parejas de residuos idénticas / total parejas alineadas.

Los alineamientos pareados globales se realizaron usando el programa NEEDLE, que usa algoritmo Needleman-Wunsch. Los parámetros empleados fueron los predeterminados. El programa se encuentra disponible como parte del servidor web del European Molecular BiologyLaboratory - EuropeanBioinformaticsInstitute (EMBL-EBI) (http://www.ebi.ac.uk/Tools/psa/).

a) Alineamiento pareado global entre *Homo sapiens* y *Pseudomonas aeruginosa*.

Identidad: 19/121 (15.7%)
Similitud: 33/121 (27.3%)
Gaps: 56/121 (46.3%)

```
########################################
# Program: needle
# Rundate: Wed 17 Feb 2016 19:55:28
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20160217-195527-0586-29300432-oy.asequence
#    -bsequence emboss_needle-I20160217-195527-0586-29300432-oy.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: d1j3sa_
# 2: d2paca_Pseudomonas
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 121
# Identity:      19/121 (15.7%)
# Similarity:    33/121 (27.3%)
# Gaps:          56/121 (46.3%)
# Score: 39.0
#
#
#=======================================

d1j3sa_            1 gdvekgkkifimk-csqchtvekggkhktgpnlhglfgrktgqapgysyt     49
                       |..:.:|..| |..||.::   ....||....:.:..|||...:..
d2paca_Pseudo      1 ---edpevlfknkgcvachaid---tkmvgpaykdvaakfagqagaeael     44

d1j3sa_           50 a---anknkgiiwg-------------edtlmeylenpkkyipgtkmifv     83
                     |   .|.::| :||              ..||.:::.:.|
d2paca_Pseudo     45 aqrikngsqg-vwgpipmppnavsddeaqtlakwvlsqk-----------     82

d1j3sa_           84 gikkkeeradliaylkkatne    104

d2paca_Pseudo     83 ----------------------     82


#---------------------------------------
#---------------------------------------
```

b) Alineamiento pareado global entre *Homo sapiens* y *Methylophilus methylotrophus*

Identidad: 30/165 (18.2%)

Similitud: 35/165 (21.2%)

Gaps: 102/165 (61.8%)

```
########################################
# Program: needle
# Rundate: Wed 17 Feb 2016 19:11:13
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20160217-191112-0562-51691871-es.asequence
#    -bsequence emboss_needle-I20160217-191112-0562-51691871-es.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: d1j3sa_
# 2: d1e8ea_
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 165
# Identity:      30/165 (18.2%)
# Similarity:    35/165 (21.2%)
# Gaps:         102/165 (61.8%)
# Score: 43.0
#
#
#=======================================

d1j3sa_            1 -------------------------gdvekgkkifimk----------cs     15
                                               ..:..||..|..|          |:
d1e8ea_            1 dvtnaeklvykytniahsanpmyeapsitdgkiffnrkfktpsgkeaaca     50

d1j3sa_           16 qchtvekggkhktgpnlhglfgrktgqapgysytaanknkgiiwgedtlm     65
                     .|||        ..|                   ||..|.|:.|
d1e8ea_           51 scht--------nnp--------------------anvgknivtg-----     67

d1j3sa_           66 eylenpkkyip------gtkmifvgikkke------------------er     91
                           |.||      .||. |..|.|.|                   |:
d1e8ea_           68 -------keipplaprvntkr-ftdidkvedeftkhcndilgadcspsek    109

d1j3sa_           92 adliaylkkatne--    104
                     |:.||||...|..
d1e8ea_          110 anfiaylltetkptk    124


#---------------------------------------
#---------------------------------------
```

c) Alineamiento pareado global entre *Pseudomonas aeruginosa* y *Methylophilus methylotrophus*

Identidad: 18/132 (13.6%)

Similitud: 31/132 (23.5%)

Gaps: 58/132 (43.9%)

```
########################################
# Program: needle
# Rundate: Wed 17 Feb 2016 19:11:24
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20160217-191123-0859-51914911-oy.asequence
#    -bsequence emboss_needle-I20160217-191123-0859-51914911-oy.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: d2paca_
# 2: d1e8ea_
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 132
# Identity:      18/132 (13.6%)
# Similarity:    31/132 (23.5%)
# Gaps:          58/132 (43.9%)
# Score: 25.5
#
#
#=======================================

d2paca_            1 ---edpevlfknkgcvachaidtkmvgpaykd----vaakfagqagaeae     43
                       ...::::|...  ..|:.:.....|:..|    ...||...:|.||.
d1e8ea_            1 dvtnaeklvykytn--iahsanpmyeapsitdgkiffnrkfktpsgkeaa     48

d2paca_           44 laqrikngsqgvwgpipmppnavsddeaqtlakwvlsqk-----------     82
                     .|....|      .|..:..|.|:..|...||..|.:::
d1e8ea_           49 caschtn------npanvgknivtgkeipplaprvntkrftdidkvedef     92

d2paca_           83 --------------------------------     82

d1e8ea_           93 tkhcndilgadcspsekanfiaylltetkptk    124


#---------------------------------------
#---------------------------------------
```

Es apreciable que, a pesar de que se trata de dominios pertenecientes a la misma superfamilia, la identidad a nivel de estructura primaria es baja (18.2% en el mejor caso).

4) Calcula con mammoth los alineamientos estructurales de los dominios que ya alineaste en 3 en base a su secuencia. Visualízalos con Rasmol como se explica en http://eead-csic-compbio.github.io/bioinformatica_estructural/node32.html. El software está en /home/compu2/algoritmos3D/soft/mammoth-1.0-src para que lo copien y compilen con gfortram como se explica en README, cambiando g77 por gfortran.

Alineamientos obtenidos usando MAMMOTH:

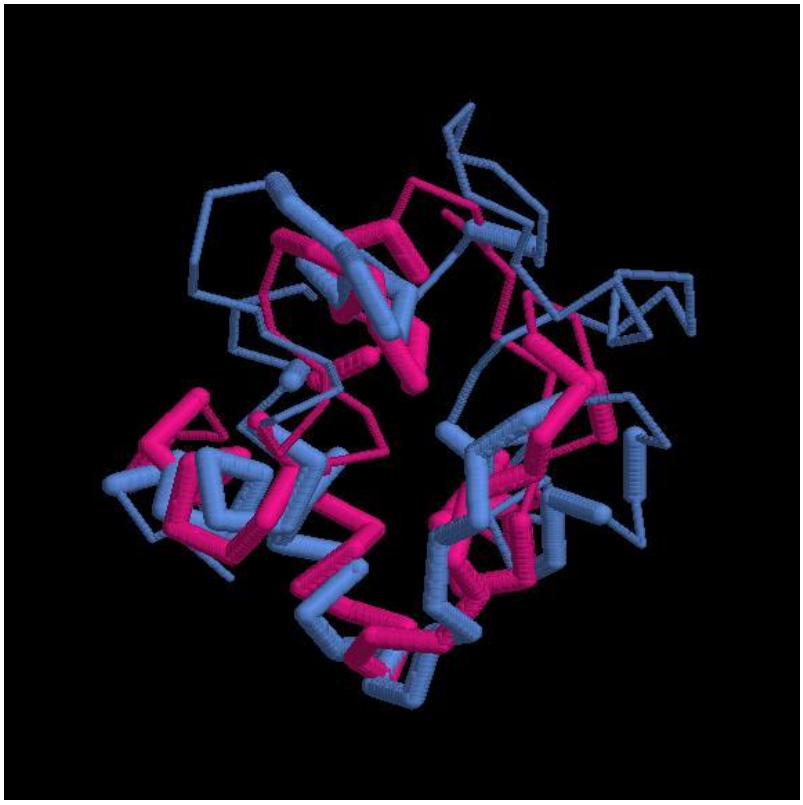
a) Alineamiento estructural entre *Homo sapiens* y *Pseudomonas aeruginosa*.

```
------------------
  Input information
 ------------------
==> PREDICTION:
     Filename: Human
     Number of residues:   104

==> EXPERIMENT:
     Filename: Pseudomonas
     Number of residues:    82

 ----------------------------
  Structural Alignment Scores
 ----------------------------
 PSI(ini)=   98.78  NALI=  81  NORM=  82  RMS=    7.42  NSS=  53
 PSI(end)=   68.29  NALI=  56  NORM=  82  RMS=    3.99
 Sstr(LG)=  821.00  NALI=  56  NORM=  82  RMS=    3.99

E-value=     0.51356241E-03

Z-score=      7.6393014        -ln(E)=      7.5741390
```

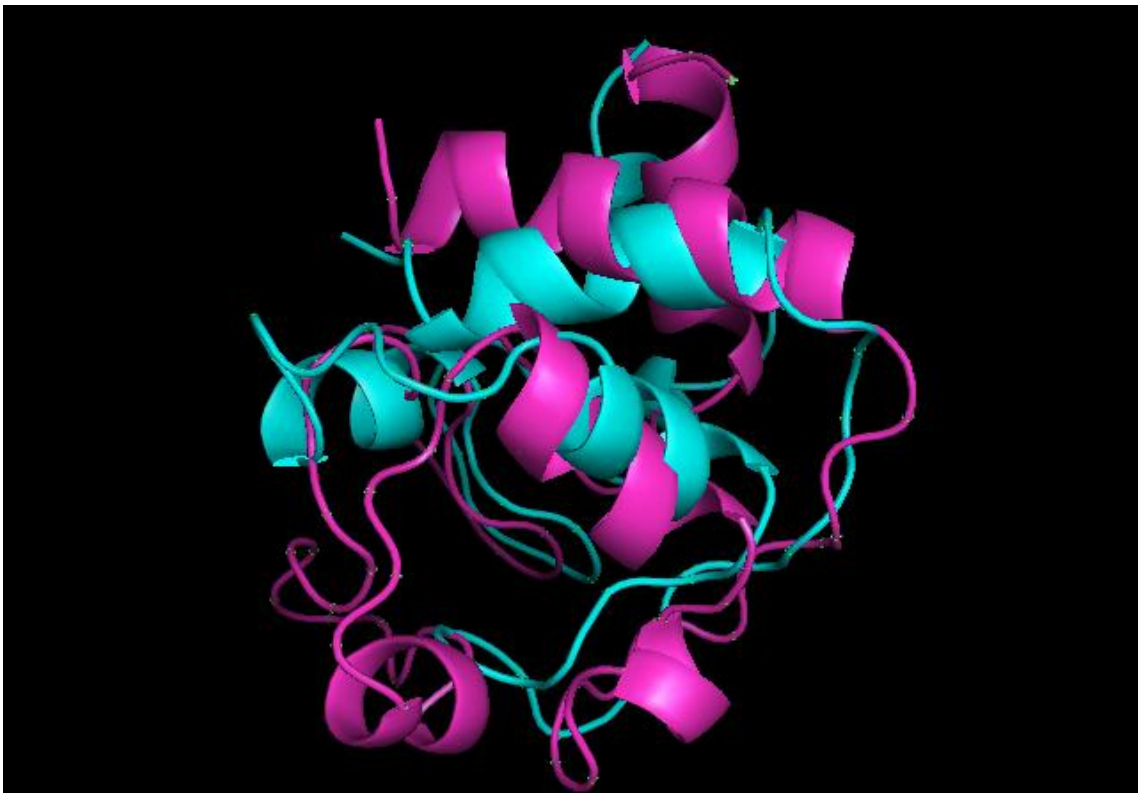
```
 ----------------------------
  Final Structural Alignment
 ----------------------------

                ** ***** *****       * ******* ***    **   *
Prediction GDVEKGKKIF IMKCSQCHTV EKGGKHKTGP NLHGLFGRKT GQAPGYSYTA
Prediction HHHHHHHHHH HHHHH--SSS ----SS---S SSSS-----S SS----SSSS
             ||||    |||||   || ||||||||||  ||         |||||||||
Experiment --HHHHHHHH H-HHH---HH HH-------H HHH----HHH HHHHHHHHHH
Experiment ..EDPEVLFK NKGCV...AC HAIDTKMVGP AYK....DVA AKFAGQAGAE
                ** ***** *****       * ******* ***    **   *

                        ** *    ** ** **  *** ********** *****  **
Prediction ANKNKGIIWG EDTLMEYLEN PKKYIPGTKM IFVGIKKKEE RADLIAYLKK
Prediction SSS---SSS- --HHHHHHHH HHH------- SSSSSS--HH HHHHHHHHHH
                  || |||           |||||   ||     ||| ||||||
Experiment --------HH HHHH------ -SSS-----S SSSSSSS--H HHHHHHHHHH
Experiment ........AE LAQR....IK NGSQGVWGPI PMPPNAVSDD EAQTLAKWVL
                        ** *    ** ** **  *** ********** *****  **

              *
Prediction ATN
Prediction HHH

Experiment HH-
Experiment SQ.
              *
```

Visualización en rasmol (I) y pymol (II)

I)



II)



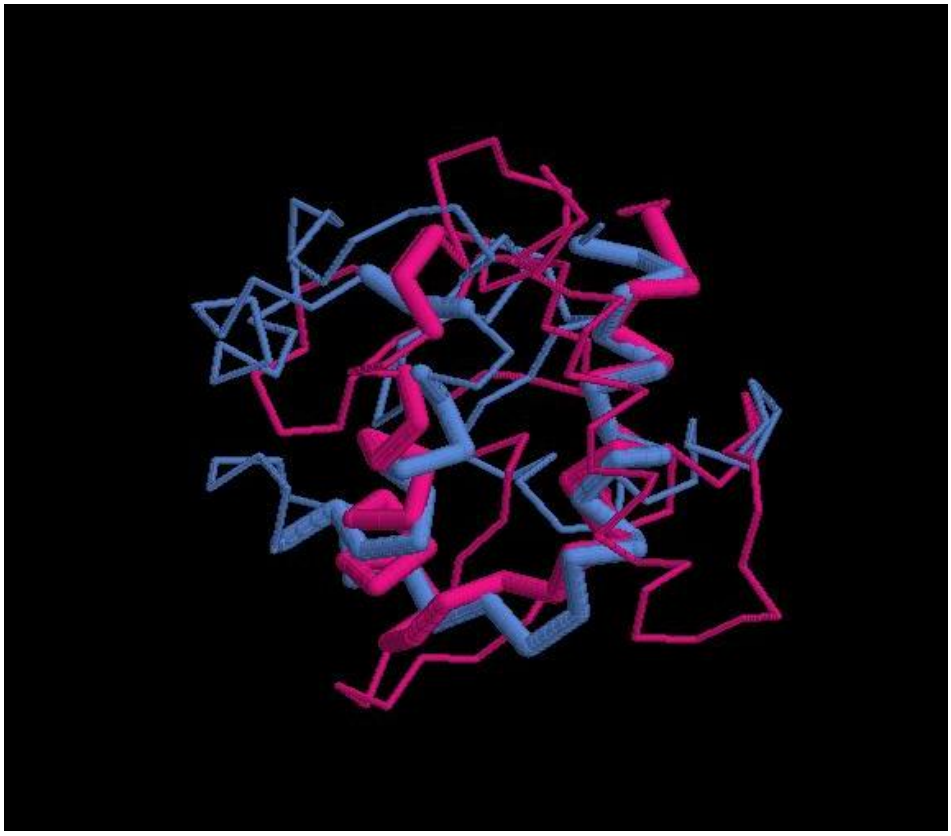
b) Alineamiento estructural entre *Homo sapiens* y *Methylophilus methylotrophus.*

```
------------------
  Input information
 ------------------
==> PREDICTION:
     Filename: Human
     Number of residues:   104

==> EXPERIMENT:
     Filename: Methylophilus
     Number of residues:   124

 ----------------------------
  Structural Alignment Scores
 ----------------------------
 PSI(ini)=   92.31  NALI=  96  NORM= 104  RMS=   10.32  NSS=  61
 PSI(end)=   35.58  NALI=  37  NORM= 104  RMS=    3.56
 Sstr(LG)=  793.09  NALI=  37  NORM= 104  RMS=    3.56

E-value=     0.28058940E-01

Z-score=      3.3514496        -ln(E)=      3.5734480
```

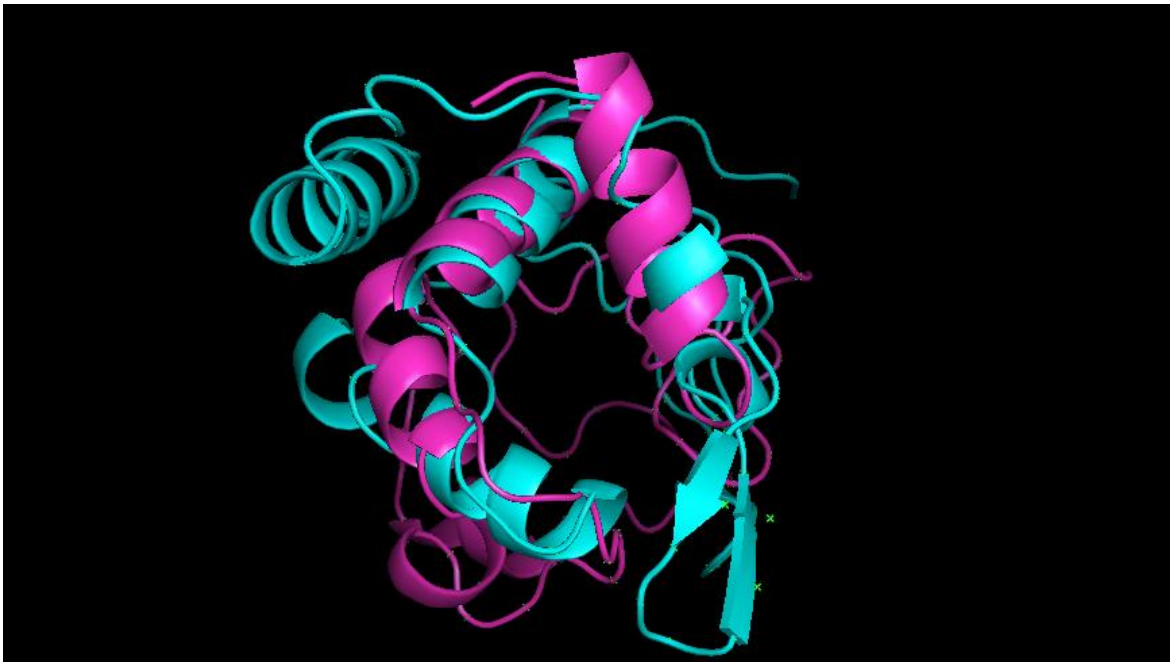
```
        ----------------------------
         Final Structural Alignment
        ----------------------------


       Prediction ..GDVEKGKK IFIMKC.... ..SQCH.... ...TVEKGGK HKTGPNLHGL
       Prediction --HHHHHHHH HHHHHH---- --H--S---- ---SS----S S---SSSSS-
                    |||||||| ||||||       ||          ||             |||
       Experiment HHHHHHHHHH HHHHHHHH-- --HHHHHHHH HHHHHHHHHH -----HHHHH
       Experiment DVTNAEKLVY KYTNIAHSAN PMYEAPSITD GKIFFNRKFK TPSGKEAACA


                                  *                  ***  *** *******
       Prediction FG.....RKT GQAPGYSY.. .TAANKN..K GIIWGEDTLM EYLENPKKYI
       Prediction ---------S SS----SS-- -SSSSS---- -SSS---HHH HHHHHHHH--
                  ||      ||   |||  |       ||||     |||||||||   |||||
       Experiment HHHHH----- SSS-----SS SSS------- SSSS--HHHH HHHHHHH---
       Experiment SCHTNNPANV GKNIVTGKEI PPLAPRVNTK RFTDIDKVED EFTKHCN...
                                  *                  ***  *** *******


                     ****** ********** *******
       Prediction PGTKMIFVGI KKKEERADLI AYLKKATN..
       Prediction -----SSSSS S--HHHHHHH HHHHHHHH--
                        |||| ||||||||||
       Experiment --------HH HHHHHHHHHH HHHHHHHHHH
       Experiment ....DILGAD CSPSEKANFI AYLLTETKPT
                     ****** ********** *******
```

Visualización en rasmol (I) y pymol (II).

I)



II.



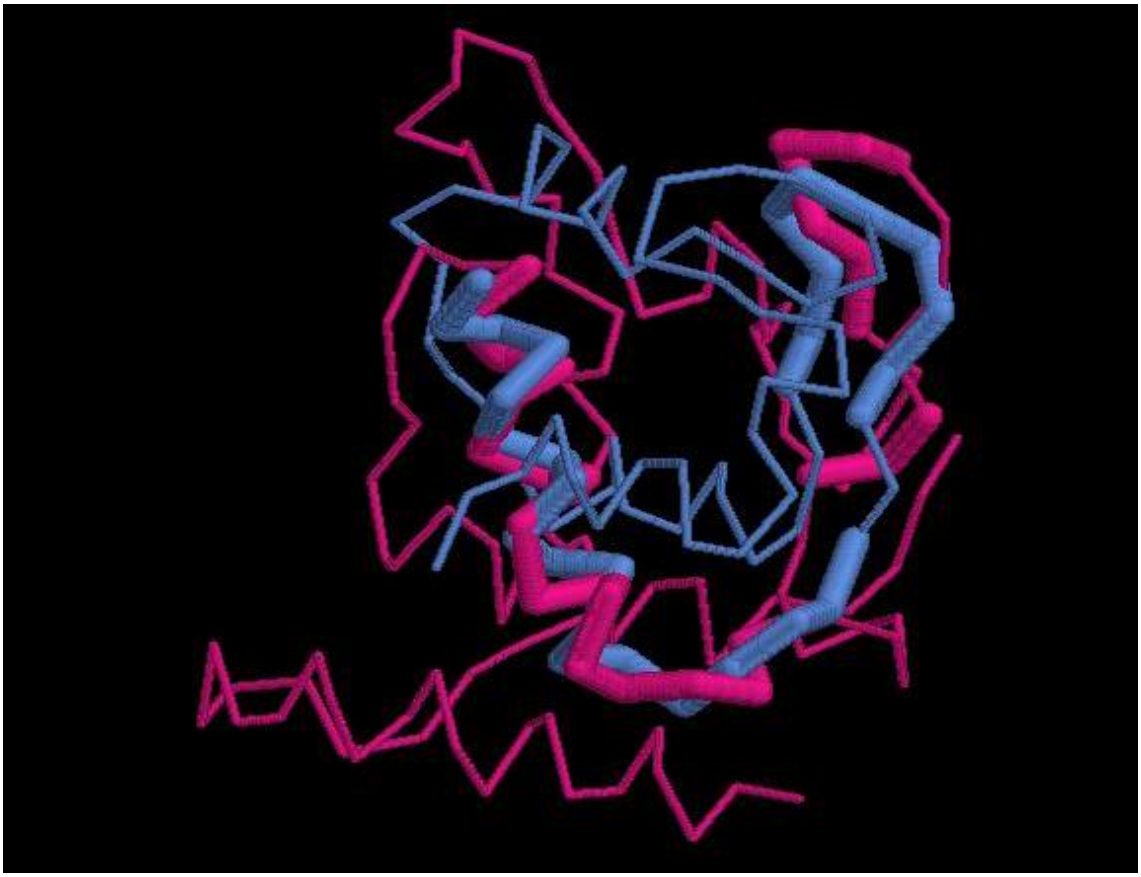
c) Alineamiento estructural entre *Pseudomonas aeruginosa* y *Methylophilus methylotrophus.*

```
------------------
  Input information
 ------------------
==> PREDICTION:
     Filename: Pseudomonas
     Number of residues:    82

==> EXPERIMENT:
     Filename: Methylophilus
     Number of residues:   124

 ----------------------------
  Structural Alignment Scores
 ----------------------------
 PSI(ini)=   98.78  NALI=  81  NORM=  82  RMS=   12.94  NSS=  45
 PSI(end)=   35.37  NALI=  29  NORM=  82  RMS=    3.89
 Sstr(LG)=  509.18  NALI=  29  NORM=  82  RMS=    3.89

E-value=     0.82272800E-01

Z-score=      2.1721472        -ln(E)=      2.4977147
```

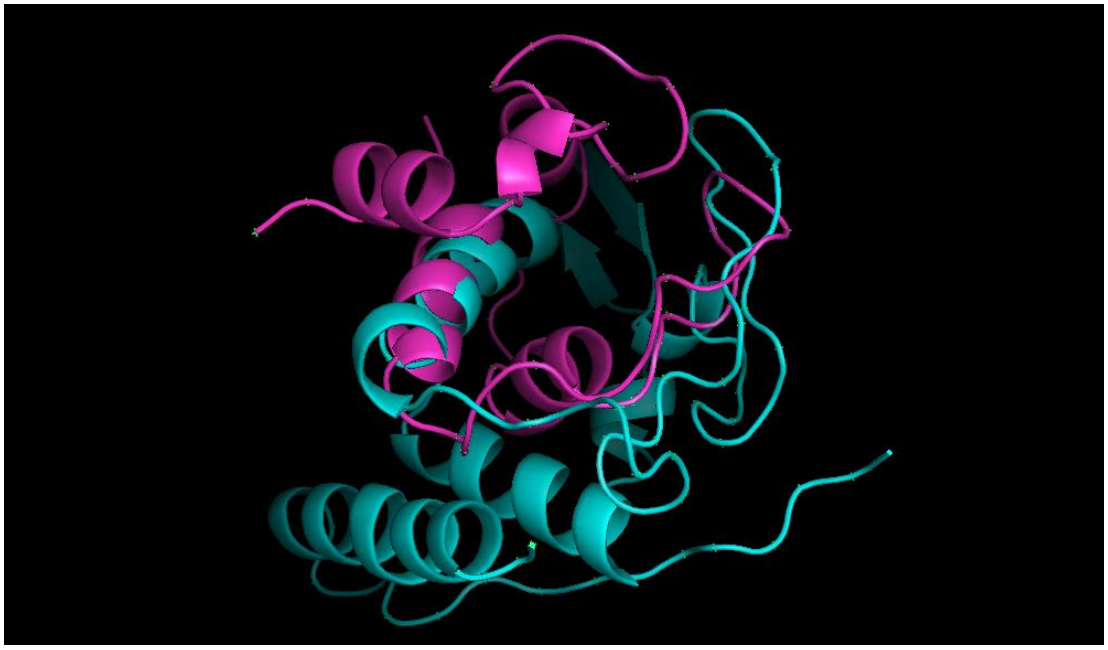
```
      ----------------------------
       Final Structural Alignment
      ----------------------------


     Prediction .EDPEVLFKN KGCVACHAID TKMVGPAYKD VAAKFAG... ...QAGAEAE
     Prediction -HHHHHHHHH -HHHHHHH-- -----HHHHH HHHHHHH--- ---HHHHHHH
                 |||||  || |           |||||||||      |         |||
     Experiment HHHHHHHHHH HHHHHHHH-- --HHHHHHHH HHHHHHHHHH -----HHHHH
     Experiment DVTNAEKLVY KYTNIAHSAN PMYEAPSITD GKIFFNRKFK TPSGKEAACA


                           *******    *** *      ********** ********
     Prediction L..AQRIKNG SQGVWGP... IPMPP..... NAVSDDEAQT LAKWVLSQ..
     Prediction H--HHH---S SS-------- SSSSS----- SSS--HHHHH HHHHHHHH--
                     ||||| ||||||       ||        |||||||| ||
     Experiment HHHHH----- SSS-----SS SSS------- SSSS--HHHH HHHHHHH---
     Experiment SCHTNNPANV GKNIVTGKEI PPLAPRVNTK RFTDIDKVED EFTKHCNDIL
                           *******    *** *      ********** ********


     Prediction .......... .......... ...
     Prediction ---------- ---------- ---

     Experiment -HHHHHHHHH HHHHHHHHHH HHH
     Experiment GADCSPSEKA NFIAYLLTET KPT
```

Visualización en rasmol (I) y pymol (II)

I)



II)



5) Compara los alineamientos obtenidos en 3 y 4. Comenta en qué elementos de estructura secundaria se observan diferencias.

A pesar de que las tres proteínas analizadas pertenecen a la misma superfamilia dentro de SCOP (pregunta 1), es apreciable que las estructuras tridimensionales de dichas proteínas no son idénticas. Sin embargo, en los alineamientos realizados con MAMMOTH se puede ver, en los tres casos, que la coincidencias de estructura secundaria son mucho mayores que las coincidencias a nivel de estructura primaria. Esto deja en evidencia cómo la estructura tridimensional de las proteínas se conserva mucho más que la secuencia durante el tiempo evolutivo.

En las estructuras de las proteínas empleadas en este análisis casi no hay láminas-β, por esta razón las mayores zonas de empalme en las superposiciones estructurales corresponden a hélices. Vale la pena notar que porcentaje elevado de las estructuras corresponde a hazas, giros y conectores; esto podría estar confiriendo mayor movilidad a las proteínas.

6) Utiliza el prog3.1 (http://eead-csic-compbio.github.io/bioinformatica_estructural/node31.html) para calcular el error (RMSD) de los alineamientos obtenidos en 3 y 4 y comenta los resultados. Son mejores o peores los alineamientos basados en secuencia desde el punto de vista del RMSD?

A continuación se muestran los resultados arrojados por el programa prog3.1 En cada caso se indica el nombre de la variación al código (dichos códigos están disponibles en el repositorio de github al igual que todos los archivos de salida).

Alineamientos con secuencia:

1) Human vs Pseudomonas

Programa: human_pseudo.py

total residuos: pdb1 = 104 pdb2 = 82

total residuos alineados = 65

coordenadas originales = original.pdb

superposicion optima:

archivo PDB = align_fit.pdb

 # RMSD = 10.36 Angstrom

2) Human vs Methylophilus

Programa: human_methy.py

total residuos: pdb1 = 104 pdb2 = 124

total residuos alineados = 63

coordenadas originales = original.pdb

superposicion optima:

archivo PDB = align_fit.pdb

 # RMSD = 10.35 Angstrom

3) Pseudomonas vs Methylophilus

Programa: pseudo_methy.py

total residuos: pdb1 = 82 pdb2 = 124

total residuos alineados = 74

coordenadas originales = original.pdb

superposicion optima:

archivo PDB = align_fit.pdb

 # RMSD = 12.41 Angstrom

Alineamientos con estructura:

1) Human vs Pseudomonas

Programa: human_pseudo_str.py

total residuos: pdb1 = 104 pdb2 = 82

total residuos alineados = 79

coordenadas originales = original.pdb

superposicion optima:

archivo PDB = align_fit.pdb

 # RMSD = 7.45 Angstrom

2) Human vs Methylophilus

Programa: human_methy_str.py

total residuos: pdb1 = 104 pdb2 = 124

total residuos alineados = 96

coordenadas originales = original.pdb

superposicion optima:

archivo PDB = align_fit.pdb

 # RMSD = 10.32 Angstrom

3) Pseudomonas vs Methylophilus

Programa: pseudo_methy_str.py

total residuos: pdb1 = 82 pdb2 = 124

total residuos alineados = 81

coordenadas originales = original.pdb

superposicion optima:

archivo PDB = align_fit.pdb

 # RMSD = 10.87 Angstrom

Tal y como los resultados lo muestran, los RMSD de cada alineamiento mejoran (disminuyen) cuando se calcula a parir del alineamiento estructural. Lo anterior tiene sentido, pues el programa MAMMOTH hace un alineamiento basado en el esqueleto de carbonos-alfa, es decir, en una superposición espacial de las dos proteínas. Un algoritmo de alineamiento pareado global únicamente se basa en la secuencia de las proteínas, por tal razón, calcular un RMSD a partir de este tipo de alineamientos ni siquiera tiene sentido.

Sin embargo, para el caso del último alineamiento (Pseudomonas vs Methylophilus), ambos RMSD's son malos, estando por arriba de los 10 A en los dos casos; la posible explicación radica en que la superposición estructural de estas dos proteínas es muy mala, a pesar de tratarse de miembros de la misma supefamilia. Como se aprecia en la sección de la pregunta 4c, solo una región corta de la proteína, correspondiente a

una hélice, es la que se suerpone; el resto está muy separado uno de otro, elevando el valor del RMSD. A pesar de este resultado, el cálculo del RMSD basándose en un alineamiento estructural siempre va a ser más confiable.

Nota: **Todos los archivos de salida generados en el desarrollo de este trabajo se encuentran en el repositorio de GitHub**