Bioinformática Estructural

Montserrat Justo, Diana García

2019-02-03

Ejercicio 3

El tercer ejercicio consiste en construir y evaluar modelos por homología de la secuencia P1.faa, la proteína que eligieron ayer. Deberán hacerlo por etapas:

1. Búsqueda de estructuras de proteínas homólogas: moldes o templates. Para ello son útiles herramientas de fold recognition como HHpred porque tienen una mayor sensibilidad y permiten encontrar similitudes de secuencia significativas muy sutiles, más allá de herramientas como PSI-BLAST.

```
head -n 25 "hhpred_5833734.hhr"
```

```
## Query         Q_5833734
## Match_columns 86
## No_of_seqs    143 out of 646
## Neff          6.83276
## Searched_HMMs 46778
## Date          Thu Jan 31 17:10:58 2019
## Command       hhsearch -cpu 8 -i ../results/full.a3m -d /cluster/toolkit/production/databases/hh-suite/
##
##  No Hit                             Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
##   1 2WPT_A COLICIN-E2 IMMUNITY PRO 100.0 7.6E-36 1.6E-40  183.4  10.4   84    1-86      1-85  (86)
##   2 1GXG_A COLICIN E8 IMMUNITY PRO 100.0 7.9E-35 1.7E-39  178.4  10.4   84    1-86      1-84  (85)
##   3 1UNK_B COLICIN E7; IMMUNITY PR 100.0   1E-34 2.1E-39  178.7  10.8   86    1-86      1-86  (87)
##   4 2VLQ_A COLICIN-E9 IMMUNITY PRO 100.0   9E-35 1.9E-39  178.6  10.5   84    1-86      1-85  (86)
##   5 4UHP_D LARGE COMPONENT OF PYOC 100.0 1.1E-34 2.4E-39  182.3   9.5   83    1-86      1-89  (98)
##   6 4QKO_C Pyocin-S2 immunity prot 100.0 2.6E-33 5.6E-38  175.1   9.5   82    3-86      1-86  (95)
##   7 4F37_B Colicin-E7 immunity pro 100.0 9.8E-33 2.1E-37  179.9  10.5   86    1-86     17-102 (124)
##   8 5WNW_C Spheroplast protein Y,   98.3 1.6E-08 3.5E-13   53.9   5.0   38    6-45      1-40  (40)
##   9 3THG_A Ribulose bisphosphate c  89.0    0.75 1.6E-05   29.3   3.9   34    7-41      2-35  (107)
##  10 6B58_B Fumarate reductase flav  86.4     1.7 3.7E-05   25.3   3.9   65   14-80     14-79  (79)
##  11 2ELC_B Anthranilate phosphorib  82.5      13 0.00027   26.8   7.5   52    8-78     11-62  (329)
##  12 2DSJ_B Pyrimidine-nucleoside (  79.4      23  0.0005   26.7   8.2   60    8-86     14-73  (423)
##  13 1KHD_A Anthranilate phosphorib  79.3      18 0.00039   26.3   7.3   51    8-77     24-74  (345)
##  14 6E4J_A Uncharacterized protein  78.9      11 0.00023   23.0   5.0   39   13-53     31-69  (80)
##  15 1VQU_B Anthranilate phosphorib  78.8      21 0.00044   26.5   7.6   52    8-78     38-89  (374)
##  16 1X6I_B Hypothetical protein yg  78.4     7.3 0.00016   23.3   4.3   68   14-83     22-90  (91)
```

1.1. Elige 1 ó 2 moldes no redundantes (T1,T2).

```
cat "P1_P2_P3.faa"
```

```
## >P1;UKNP
## sequence:UKNP:1    :A:90  :A::::
## --LKNSISDYTEAEFVQLLKEIEKENVAA----TDDVLDVLLEHFVKITEHPDGTDLIYYPSDNRDDSPEGIVKEIKEWRAANGKPGFKQ*
## >P2;5WNW
## structure:5WNW:1   :C:40  :C::Escherichia coli:1.79:
## -----SISDY-EAE-VQLLKEIEKEN--V--AATDDVLDVLLEHFVK-T---------------------------------------*
## >P3;4QKO
## structure:4QKO:1   :C:86  :C::Pseudomonas aeruginosa:1.8:
## --MKSKISEYTEKEFLEFVKDIYTNN--KKKFPTEESHIQAVLEFKKLTEHPSGSDLLYYPNENREDSPAGVVKEVKEWRASKGLPGFKA*
```

```
head "5wnw.pdb"
```

```
## HEADER    CHAPERONE                               01-AUG-17   5WNW
## TITLE     CHAPERONE SPY BOUND TO IM7 6-45 ENSEMBLE
## CAVEAT     5WNW    THERE ARE SIGNIFICANT ATOMIC CLASHES IN THE STRUCTURE. CHAIN
## CAVEAT   2 5WNW    C DOES NOT SHOW CLEAR ELECTRON DENSITY. THERE ARE ATOMS
## CAVEAT   3 5WNW    WITH OCCUPANCY LARGER THAN 1. THERE ARE GAPS BETWEEN A MET
## CAVEAT   4 5WNW    53 AND A ARG 55, BETWEEN B MET 53 AND B ARG 55, BETWEEN C
## CAVEAT   5 5WNW    TYR 10 AND C GLU 12, BETWEEN C GLU 14 AND C VAL 16, BETWEEN
## CAVEAT   6 5WNW    C LYS 43 AND C THR 45.
## COMPND    MOL_ID: 1;
## COMPND   2 MOLECULE: PERIPLASMIC CHAPERONE SPY;
```

```
head "4qko.pdb"
```

```
## HEADER    ANTIMICROBIAL PROTEIN                   07-JUN-14   4QKO
## TITLE     THE CRYSTAL STRUCTURE OF THE PYOCIN S2 NUCLEASE DOMAIN, IMMUNITY
## TITLE    2 PROTEIN COMPLEX AT 1.8 ANGSTROMS
## COMPND    MOL_ID: 1;
## COMPND   2 MOLECULE: PYOCIN-S2 IMMUNITY PROTEIN;
## COMPND   3 CHAIN: A, C, E, G;
## COMPND   4 ENGINEERED: YES;
## COMPND   5 MOL_ID: 2;
## COMPND   6 MOLECULE: PYOCIN-S2;
## COMPND   7 CHAIN: B, D, F, H;
```

1.2. En base a T1-T2 modela la estructura terciaria de P1 desde HHpred, con ayuda de MODELLER. Guarda los ficheros PDB de salida, así como sus anotaciones.

Figure 1: Resultados de Modeller

```
head "Modeller_6097201.pdb"
```

```
## EXPDTA    THEORETICAL MODEL, MODELLER 9.21 2019/01/31 17:26:39
## REMARK   6 MODELLER OBJECTIVE FUNCTION:      1729.7787
## REMARK   6 MODELLER BEST TEMPLATE % SEQ ID:  91.892
## REMARK   6 SEQUENCE: 6097201
## REMARK   6 ALIGNMENT: alignment.pir
## REMARK   6 SCRIPT: modeller_script.py
## REMARK   6 TEMPLATE: 5WNW 1:C - 40:C MODELS 6:X - 45:X AT 91.9%
## REMARK   6 TEMPLATE: 4QKO 1:C - 86:C MODELS 3:X - 86:X AT 51.2%
## ATOM      1  N   LEU X   3     -16.246  -5.532   7.173  1.00103.44           N
## ATOM      2  CA  LEU X   3     -15.269  -4.426   7.136  1.00103.44           C
```

```
head "HHPRED_summary.txt"
```

```
## Check Hits for overlap and repeat domains:
## 6 4QKO_C  3 86  1 86
## 8 5WNW_C  6 45  1 40
##
##
##
##
## Selected templates 8, 6.
## 8: 5WNW_C Spheroplast protein Y, Colicin-E7 immunity; Chaperone; HET: IMD, ZN; 1.79
## 6: 4QKO_C Pyocin-S2 immunity protein; HNH Nuclease Domain, Colicin Nuclease; HET: B
```

2. Trata de modelar P1 por predicción de contactos con EVfold (ver sección "Modelado de proteínas por predicción de contactos"). Guarda los ficheros PDB de salida. [OJO: EVfold puede tardar muchas horas].

No pudimos obtener los resultados de EVfold porque el job falló en múltiples ocasiones

3. Evalúa la calidad de los modelos obtenidos:

3.1. Graficando su diagrama de Ramachandran.

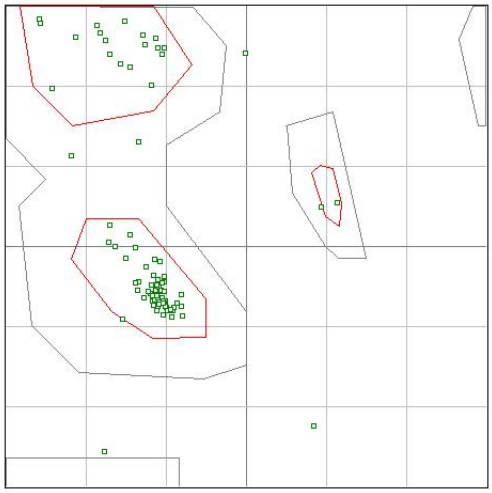

Figure 2: Ramachandran plot de 1AYI

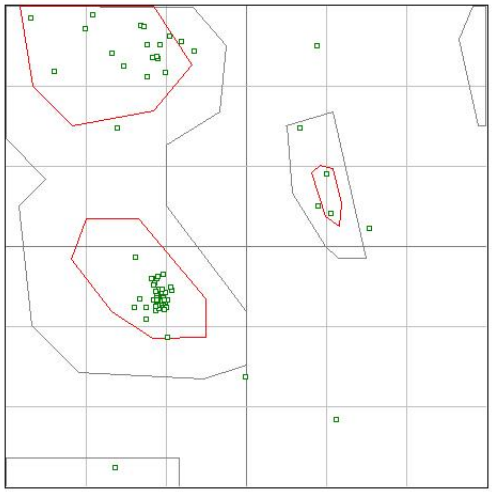

Figure 3: Ramachandran plot del modelo obtenido por HHPred

3.2. Comparando las estructuras modeladas con la solución experimental P1.pdb. Para ello deben calcular sus superposiciones con MAMMOTH (RMSD y E-valor, instalado en tepeu.lcg.unam.mx) y con TMalign (TMscore, lo tienen que instalar de http://zhanglab.ccmb.med.umich.edu/TM-align/TM-align-C/TMalignc.tar.gz).

```
cat "mammoth.txt"
```

```
##  Predicted path:
##  Experimental path:
##  ________________________________________________________________________________
##
##                                 M  A  M  M  O  T  H
##
##
##                MAtching Molecular Models Obtained from THeory
##  ________________________________________________________________________________
##
##
##  -------------------
##   Input information
##  -------------------
##
##
## ==> PREDICTION:
##
##      Filename: Modeller_6097201.pdb
##      Number of residues:    84
##
##
## ==> EXPERIMENT:
##
##      Filename: P1.pdb
##      Number of residues:    86
##
##
##  ----------------------------
##   Structural Alignment Scores
##  ----------------------------
##
##
##  PSI(ini)=   98.81  NALI=  83  NORM=  84  RMS=    1.89  NSS=  69
##  PSI(end)=   98.81  NALI=  83  NORM=  84  RMS=    1.89
##  Sstr(LG)= 1461.77  NALI=  83  NORM=  84  RMS=    1.89
##
```

```
## E-value=     0.34265275E-05
##
## Z-score=      12.990349        -ln(E)=      12.583963
##
##  ----------------------------
##   Final Structural Alignment
##  ----------------------------
##
##              ******** ********** ********** ********** **********
## Prediction ..LKNSISDY TEAEFVQLLK EIEKENVAAT DDVLDVLLEH FVKITEHPDG
## Prediction --SSS--SSS S-HHHHHHHH HHHHHH---- -HHHHHHHHH HHHH----SS
##              |||||||| |||||||||         ||| |||||||||| ||||||||||
## Experiment SSSS----SS S-HHHHHHHH HHHHHHH--- HHHHHHHHHH HHHH----SS
## Experiment MELKNSISDY TEAEFVQLLK EIEKENVAAT DDVLDVLLEH FVKITEHPDG
##              ******** ********** ********** ********** **********
##
##            ********** ********** ********** *****
## Prediction TDLIYYPSDN RDDSPEGIVK EIKEWRAANG KPGFK
## Prediction SSS--SSS-- -SS---HHHH HHHHHHH--- -SSSS
##            |||||||||| |||||||||| |||||||||
## Experiment SSS--SSS-- -SS---HHHH HHHHHHH--- -SSSS
## Experiment TDLIYYPSDN RDDSPEGIVK EIKEWRAANG KPGFK
##            ********** ********** ********** *****
##
##
##  ---------
##   Timings
##  ---------
##
##    < Initialization:                     0.000 sec >
##    < Secondary Structure assignment      0.000 sec >
##    < Structure alignment:                0.010 sec >
##    < Tertiary structure matching:        0.000 sec >
##    < Text Output                         0.000 sec >
##
## <MAMMOTH> NORMAL_EXIT
```

```
cat "TMAlign_result.txt"
```

```
##
##  ************************************************************************
##  * TM-align (Version 20160521): A protein structural alignment algorithm     *
##  * Reference: Y Zhang and J Skolnick, Nucl Acids Res 33, 2302-9 (2005)       *
##  * Please email your comments and suggestions to Yang Zhang (zhng@umich.edu) *
```

```
##  ****************************************************************************
##
## Name of Chain_1: /home/diana/Dropbox/Curso_Bioinformatica/Tareas/Tarea_2/P1.pdb (to be superimposed ont
## Name of Chain_2: /home/diana/Dropbox/Curso_Bioinformatica/Tareas/Tarea_3/Modeller_6097201.pdb
## Length of Chain_1: 86 residues
## Length of Chain_2: 84 residues
##
## Aligned length= 82, RMSD=   1.51, Seq_ID=n_identical/n_aligned= 0.915
## TM-score= 0.85642 (if normalized by length of Chain_1, i.e., LN=86, d0=3.33)
## TM-score= 0.87490 (if normalized by length of Chain_2, i.e., LN=84, d0=3.29)
## (You should use TM-score normalized by length of the reference protein)
##
## (":" denotes residue pairs of d <  5.0 Angstrom, "." denotes other aligned residues)
## MELKNSISDYTEAEFVQLLKEI-EKE-NVAATDDVLDVLLEHFVKITEHPDGTDLIYYPSDNRDDSPEGIVKEIKEWRAANGKPGFKQ
##   :::::::::::::::::::: ::: .:: :: ::::::::::::::::::::::::::::::::::::::::::::::::::::::
## --LKNSISDYTEAEFVQLLKEIEKENVAAT-DD-VLDVLLEHFVKITEHPDGTDLIYYPSDNRDDSPEGIVKEIKEWRAANGKPGFKQ
##
##
## Total running time is  0.01 seconds
```

El alineamiento estructural generado con MAMMOTH muestra que la similitud estructural entre ambos modelos están muy relacionadas (E value < 0.0001). Igualmente el alineamiento mediante TMalign muestra la alta similitud estre ambos modelos (TM-score= 0.87490). Se utilizó la página http://tomcat.cs.rhul.ac.uk/home/mxba001/

Observando las gráficas de Ramachandran podemos concluir que el modelo obtenido de la predicción es muy parecido al modelo de nuestra proteína original. La mayoría de los residuos de la proteína se distribuyen de manera similar entre ambos modelos.