

## Tarea 4: Modelado de proteínas por homología

Lucía Graña

Programa de doctorado en ciencias Biomédicas

El objetivo de esta tarea es conocer la estructura tridimensional de una proteína a partir de su secuencia aminoacídica. Esto se realiza estimando las coordenadas de los átomos de la proteína a partir de la estructura de proteínas homólogas.

Los pasos a seguir son los siguientes:

1. Búsqueda de estructuras proteicas relacionadas
2. Selección de uno o más *templates*
3. Alinear *target-template*
4. Construir un modelo
5. Evaluar el modelo

Para realizar la tarea se escogió uno de los dominios trabajado en la Tarea 3, el dominio 1auz A de la bacteria *Bacillus subtilis*. La secuencia de este dominio va a ser el *query* en este trabajo, se intentará predecir su estructura tridimensional a partir de la secuencia de aminoácidos.

El primer paso fue buscar un dominio homólogo cuya estructura sea conocida. Para esto se utilizó HHpred. Este *software* es uno de los más eficientes en encontrar *templates* y alinear el *target* con el *template*. Esta búsqueda detectó 100 dominios con diferentes porcentajes de identidad con el *target*, de los cuales se escogió como *template* el dominio 1sbo A de *Mycobacterium avium*.

El alineamiento *target-template* se realizó y se utilizó el programa MODELLER (Sali and Blundell, 1993) para generar dos modelos. MODELLER es un programa muy popular para modelar proteínas, funciona intentando satisfacer las restricciones espaciales impuestas dada la supuesta homología con el *template* (Sali and Blundell, 1993). Este programa requiere como *input* el alineamiento *target-template* en formato .pir y los archivos .pdb de cada uno. El output son los dos modelos generados. Ambos modelos se muestran en las figuras 1 y 2.



Figura 1. Modelo 1 predicho por MODELLER

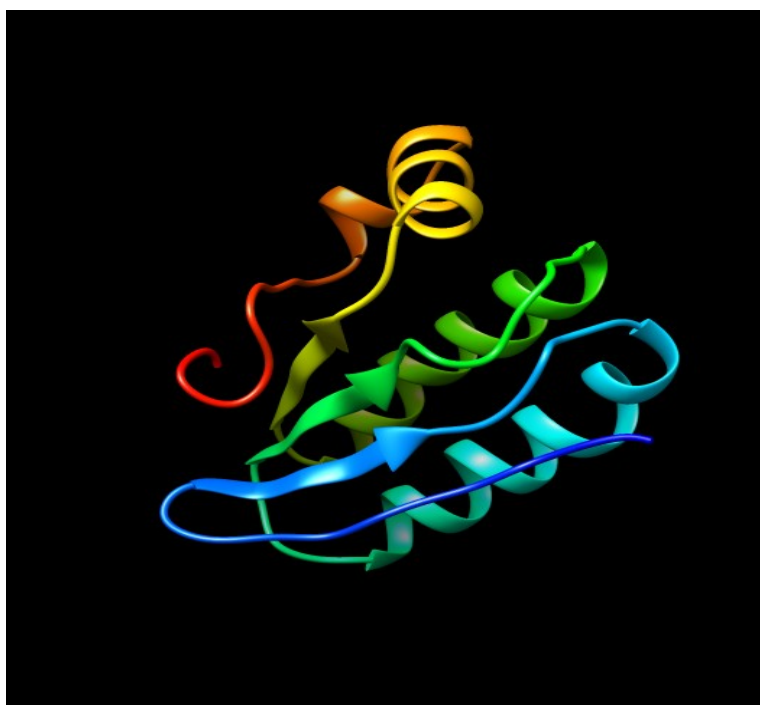


Figura 2. Modelo 2 predicho por MODELLER.

Los modelos generados realmente no muestran variación apreciable a simple vista entre ellos.

Estos dos modelos fueron comparados con MAMMOTH (Ortiz et al., 2002), este programa permite evaluar los modelos usando DOPE (**D**iscrete **O**ptimized **P**rotein **E**nergy), que es una función de energía derivada de el análisis de estructuras proteicas conocidas del Protein Data Bank, por lo que también se lo conoce como potencial estadístico (Shen and Sali, 2006). DOPE permite valorizar la predicción de la estructura proteica basada en homología.

DOPE se presenta como un Z-score (*normalized DOPE method*); los scores positivos indican modelos pobres, mientras que si el score es menor a -1, el modelo es posible que esté prediciendo la estructura nativa (<<https://salilab.org/modeller/tutorial/basic.html>>).

En este caso ambos modelos presentan un Z-score ampliamente positivo: para el modelo 1 el Z-score es 16.106750 y para el modelo 2 es el mismo. Realmente es sorprendente encontrar este resultado, incluso Modeller se volvió a correr y esta vez para generar tres modelos, nuevamente no se detectaron diferencias entre los modelos generados.

El programa prog3.1.py (<[https://github.com/eead-csic-compbio/bioinformatica\\_estructural/](https://github.com/eead-csic-compbio/bioinformatica_estructural/)>) calcula la superposicion en 3D equivalente a un alineamiento de secuencia de dos proteínas del PDB. Genera un fichero PDB con la superposicion obtenida. Este archivo se visualizó con chimera.

Se utilizó MAMMOTH (Ortiz et al., 2002) para comparar los modelos generados con Modeller, con la estructura terciaria conocida del dominio. Estos son los resultados:

## Modelo 1 vs Estructura Experimental

- Alineamiento:

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

*****
Prediction SLGIDMNVKE SVLCIRLTGE LDHHTAETLK QKVTQSLEKD DIRHIVLNLE
Prediction SSSSSSS---SSSSSS--S S---HHHHHH HHHHHHHHH- ---SSSSS-
      ||||| ||||| ||||| ||||| |||||
Experiment SSSSSSSS--SSSSSS--HHHHHHH HHHHHHHHH- ---SSSSS
Experiment SLGIDMNVKE SVLCIRLTGE LDHHTAETLK QKVTQSLEKD DIRHIVLNLE
*****

*****
Prediction DLSFMDSSGL GVILGRYKQI KQIGGEMVVC AISPAVKRLF DMSGFLFKIIR
Prediction ---HHHHHHH HHHHHHHHHH H---SSSS SS---HHHHH HH-SSSS--S
      ||||| ||||| ||||| ||||| |||||
Experiment ---SSS-HHH HHHHHHHHHH H---SSSS SS---HHHHH HH-SSSS--S
Experiment DLSFMDSSGL GVILGRYKQI KQIGGEMVVC AISPAVKRLF DMSGFLFKIIR
*****

**** **
Prediction FEQSEQQALL TLGVA
Prediction SS---HHHHH H---

Experiment SS-----
Experiment FEQSEQQ... ..
**** **
```

- RMSD = 2.37 Angstrom

Este es un valor de similitud moderada.

- Superposición de estructuras:



## Modelo 2 vs Estructura Experimental

- Alineamiento:

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

```

*****
Prediction SLGIDMNVKE SVLCIRLTGE LDHHTAETLK QKVTQSLEKD DIRHIVLNLE
Prediction SSSSSSS-- --SSSSS--S S---HHHHHH HHHHHHHHH- ---SSSSS-
||||||| ||||||| ||||||| ||| |||||||
Experiment SSSSSSSS-- --SSSSS-- ---HHHHHH HHHHHHHHH- ----SSSSS
Experiment SLGIDMNVKE SVLCIRLTGE LDHHTAETLK QKVTQSLEKD DIRHIVLNLE
*****

*****
Prediction DLSFMDSSGL GVILGRYKQI KQIGGEMVVC AISPAVKRLF DMSGFLFKIIR
Prediction ---HHHHHH HHHHHHHHH H---SSSSS SS---HHHH HH-SSSS--S
||||||| ||||||| ||||||| ||||||| |||||||
Experiment ---SSS-HHH HHHHHHHHH H---SSSSS SS---HHHH HH-SSSS--S
Experiment DLSFMDSSGL GVILGRYKQI KQIGGEMVVC AISPAVKRLF DMSGFLFKIIR
*****

**** **
Prediction FEQSEQQALL TLGVA
Prediction SS---HHHHH H---

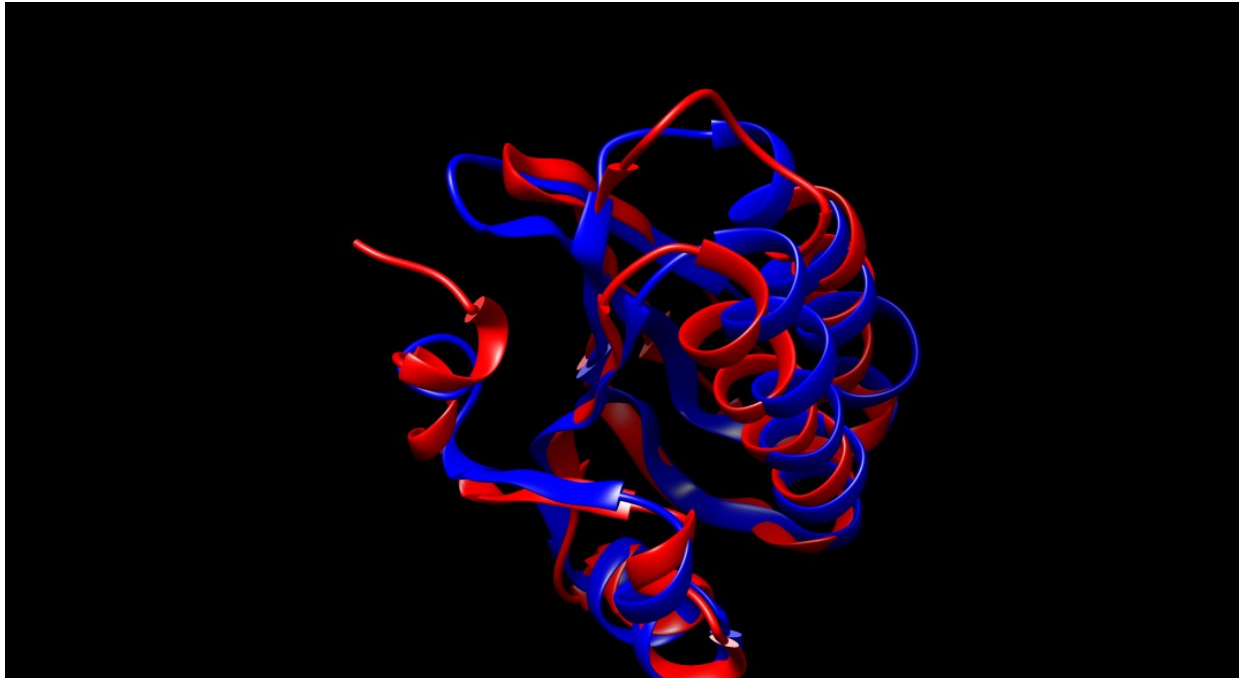
Experiment SS-----
Experiment FEQSEQQ... ..
**** **

```

- RMSD = 2.35 Angstrom

Este valor indica una similitud moderada entre el modelo propuesto y la estructura real de la proteína.

- Superposición de estructuras:



Los dos modelos hipotéticos generados son realmente muy similares, no se los puede distinguir con DOPE, o sea, no se puede decir que uno es mejor que otro. En la visualización de la superposición de sus estructuras, con la estructura experimental del *target*, que fue descargada del Protein Data Bank, parece que el modelo no difiere mucho de la estructura original. Aún así, los valores de RMSD (que mide las desviaciones en las posiciones de los residuos a partir de las secuencias y los archivos de coordenadas), indican una similitud moderada.

## Bibliografía

Ortiz, A.R., Strauss, C.E.M., and Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci. Publ. Protein Soc.* 11, 2606–2621.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815.

Shen, M., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15, 2507–2524.