Práctica 3

Anastasia Hernández y Alán Muñoz

February 15, 2016

# Práctica 3

1. Selecciona una superfamilia de proteínas de SCOP (<http://scop.berkeley.edu> [ <http://scop.berkeley.edu> ]) y extrae la secuencia de aminoácidos (ATOM records) y las coordenadas PDB de varios dominios de la misma. Podéis ver un ejemplo de dominio en <http://scop.berkeley.edu/sunid=29763> [ <http://scop.berkeley.edu/sunid=29763> ] , y abajo están tanto la secuencia como una liga para descargar las coordenadas.

Guardadas como seqs/fasta/name.fa Guardadas como seqs/pdb/name.pdb

Para visualizar directamente las secuencias: <https://github.com/afermg/bioinfo_lcg/blob/master/Practice> 3/seqs/pdb

1. Comprueba que las secuencias descargadas coinciden con las coordenadas. scripts/filter.sh perl scripts/traduct.pl seqs/aa/name.seqaa > seqs/tocompare/name.tocompare Linearizar Fasta (Manualmente)

* diff name.fa name.tocompare -i

1. Calcula al menos dos alineamiento 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.

* cat seqs/fasta/\*.fa > seqs/concat/seqs\_cat.fa <http://www.ch.embnet.org/software/ClustalW.html>

d1llua2\_d1qora2 Identidad: 46/159 (28.9%)

d1kola2\_d1llub2.al Identidad: 36/232 (15.5%)

Para visualizar directamente los alineamientos: <https://github.com/afermg/bioinfo_lcg/tree/master/Practice%203/seqs/als>

1. 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.

~/algoritmos3D/soft/mammoth-1.0-src/mammoth -p seqs/pdb/name1.pdb -e seqs/pdb/name2.pdb -o seqs/als/name1\_name2.al

rasmol -script seqs/als/rasmol.tcl

Para visualizar directamente los alineamientos: <https://github.com/afermg/bioinfo_lcg/tree/master/Practice%203/seqs/als>

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

Las secuencias son bastante similares al inicio de los alineamientos, pero conforme se avanza en el mismo, ambas divergen; también, puede observarse que en el análisis por secuencias hay una cantidad mucho mayor de gaps debida a la cantidad menor de alineamientos que se encuentran.No se observan secuencias repetidas de forma clara en ninguno de los distintos alineamiento y, dentro del análisis de secuencia en las regiones iniciales y finales del alineamiento se observan grandes regiones de repeticiones H (de aproximadamente 20 posiciones de longitud), ninguna repetición similar a ésta se observa en los alineamientos por secuencia.

1. Utiliza el prog3.1 (en <http://eead-csic-compbio.github.io/bioinformatica_estructural/node31.html> [ <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?

Para d1kola2\_d1llub2 (por secuencia):

total residuos: pdb1 = 193 pdb2 = 166 total residuos alineados = 129

RMSD = 8.23 Angstrom

Para d1kola2\_d1llub2 (por estructura):

total residuos: pdb1 = 193 pdb2 = 166 total residuos alineados = 123

RMSD = 7.82 Angstrom

Para d1qora2\_d1llua2 (por secuencia):

total residuos: pdb1 = 179 pdb2 = 166 total residuos alineados = 152

RMSD = 6.14 Angstrom

Para d1qora2\_d1llua2 (por estructura):

total residuos: pdb1 = 179 pdb2 = 166 total residuos alineados = 123

RMSD = 12.34 Angstrom

= Puede observarse que para ambos casos, en el análisis por estructura, se recupera un total de residuos alineados menor al de secuencia. También, en el primer par de secuencias no se ve una gran diferencia entre los valores de RMSD; pero en el segundo par, el valor de RMSD es más del doble en estructura, demostrando que la diferencia entre los datos del modelo es más grande con esta métrica por lo que la divergencia de las secuencias en el alineamiento es mayor, pero debe hacerse énfasis en que si la secuencia tiene una gran cantidad de dominios múltiples, el valor aumentará por la amplia cantidad de entradas.Por todo esto, parecería que el alineamiento por secuencias es mejor, al menos en términos de divergencia de los datos; pero es incorrecto asegurar esto, sin saber más acerca de la conformación de las secuencias, ya que podemos tener mucho más ruido al obtener el mismo valor para la estructura.