Bioinformática estructural

Tarea 3-Algoritmos 3D

Carina Paola Cornejo Páramo

Claudia Saraí Reyes Ávila

Marzo-2016

Introducción

Las estructuras terciarias sirven para capturar las relaciones espaciales entre diferentes partes o elementos de estructuras secundarias de una misma molécula. La estructura terciaria de una proteína está determinada por su secuencia. En esta práctica seleccionaremos proteínas relacionadas, que pertenecen a una misma superfamilia, y haremos alineamientos estructurales para saber que tanto se parecen. La predicción de la estructura terciaria a partir de la secuencia es todavía un gran reto, y este tipo de comparaciones nos pueden mostrar que tanto cambia la estructura de una proteína cuando se dan, por ejemplo, mutaciones puntuales y esto tiene gran relevancia desde el punto de vista evolutivo y del estudio de patologías.

Desarrollo:

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 varios dominios de la misma. Podéis ver un ejemplo de dominio en http://scop.berkeley.edu/sunid=29763 , y abajo están tanto la secuencia como una liga para descargar las coordenadas.

Las proteínas que seleccionamos pertenecen a la clase de proteínas *All alpha*, y a la superfamilia de Citocromos p450.

Seleccionamos diferentes 3 diferentes citocromos P450:

d2j0da1:citocromo P450 3a4 de Homo sapiens

d2v0ma1: citocromo P450 3a4 de Homo sapiens

d1dt6a1: citocromo p450 2c5 de Oryctolagus cuniculus (conejo).

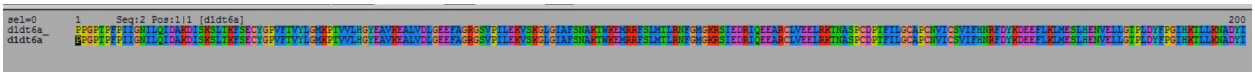
2) Comprueba que las secuencias descargadas coinciden con las coordenadas. Comprobar que coincidan las coordenadas

Usamos un programa de perl para extraer las secuencias a partir de los archivos pbd que descargamos. Posteriormente usamos el visualizador SeaView para comparar las secuencias. A continuación se muestran las imágenes de las secuencias en SeaView:

d2v0ma1:

d2j0da1

d1dt6a



Las secuencias comparadas son exactamente iguales como podemos observar por el código de colores en las imágenes.

3) 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. me falta una secuencia que me equivoque

Reaizamos los alineamientos con BLAST y los descargamos. En las siguientes imágenes se muestran los alineamientos así como los porcentajes de identidad.

Alineamiento de d2j0da1 contra d1dt6a:

```
Query_216117  d2j0da1 a.104.1.1 (A:30-496) Mammalian cytochrom...  113     7e-32

ALIGNMENTS
>Query_216117 d2j0da1 a.104.1.1 (A:30-496) Mammalian cytochrome P450 3a4 {Human
(Homo sapiens) [TaxId: 9606]}
Length=445

 Score = 113 bits (282),  Expect = 7e-32, Method: Compositional matrix adjust.
 Identities = 115/441 (26%), Positives = 185/441 (42%), Gaps = 53/441 (12%)

Query  2    PGPTPFPIIGNILQIDAKDISKSLTKFS-EC---YGPVFTVYLGMKPTVVLHGYEAVKEA  57
            PGPTP P +GNIL        K    F  EC   YG V+  Y G +P + +   + +K
Sbjct  10   PGPTPLPFLGNIL-----SYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAITDPDMIKTV  64

Query  58   LV-DLGEEFAGR---GSVPILEKVSKGLGIAFSNAKTWKEMRRFSLMTLRNFGMGK-RSI  112
            LV +    F  R   G V  ++       I+ +  + WK +R  SL++   F  GK + +
Sbjct  65   LVKECYSVFTNRRPFGPVGFMKS-----AISIAEDEEWKRLR--SLLS-PTFTSGKLKEM  116

Query  113  EDRIQEEARCLVEELRKT--NASPCDPTFILGCAPCNVICSVIFHNRFDYKDEEFLKLME  170
               I +    LV  LR+      P     + G    +VI S  F    D   ++    +E
Sbjct  117  VPIIAQYGDVLVRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDSNPQD--PFVE  174

Query  171  SLHENVELLGTPLDYFPGIHKTL--LKNADY---IKNFIMEKVKEH------QKLLDVNN  219
            +  + +      +  FP +   L  L    +   + NF+ + VK        Q ++D  N
Sbjct  175  NTKKLLRFFFLSITVFPFLIPILEVLNICVFPREVTNFLRKSVKRMKESRFLQLMIDSQN  234

Query  220  PRDFIDCFLIKMEQENNLEFTLESLVIAVSDLFGAGTETTSTTLRYSLLLLLKHPEVAAR  279
               +          ++LE   +S++        AG ETTS+ L + +  L  HP+V  +
Sbjct  235  SHKAL----------SDLELVAQSIIFIF-----AGYETTSSVLSFIMYELATHPDVQQK  279

Query  280  VQEEIERVIGRHRSPCMQDRSRMPYTDAVIHEIQRFIDLLPTNLPHAVTRDVRFRNYFIP  339
            +QEEI+ V+     P      +M Y D V++E  R   +    L     +DV     FIP
Sbjct  280  LQEEIDAVLPNKAPPTYDTVLQMEYLDMVVNETLRLFPI-AMRLERVCKKDVEINGMFIP  338

Query  340  KGTDIITSLTSVLHDEKAFPNPKVFDPGHFLDESGNFKKSDYFMPFSAGKRMCVGEGLAR  399
            KG  ++    ++  D K +  P+ F P  F  ++ +      + PF +G R C+G   A
Sbjct  339  KGVVVMIPSYALHRDPKYWTEPEKFLPERFSKKNKDNIDPYIYTPFGSGPRNCIGMRFAL  398

Query  400  MELFLFLTSILQNFKLQSLVE  420
            M + L L  +LQNF  +   E
Sbjct  399  MNMKLALIRVLQNFSFKPCKE  419
```

Alineamiento de d2v0ma1 contra d2j0da1:

```
Query_8931  d2j0da1 a.104.1.1 (A:30-496) Mammalian cytochrome ...  881     0.0

ALIGNMENTS
>Query_8931 d2j0da1 a.104.1.1 (A:30-496) Mammalian cytochrome P450 3a4 {Human
(Homo sapiens) [TaxId: 9606]}
Length=445

 Score = 881 bits (2277),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 441/457 (96%), Positives = 441/457 (96%), Gaps = 16/457 (4%)

Query  1    HGLFKKLGIPGPTPLPFLGNILSYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAITDPDM  60
            HGLFKKLGIPGPTPLPFLGNILSYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAITDPDM
Sbjct  1    HGLFKKLGIPGPTPLPFLGNILSYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAITDPDM  60

Query  61   IKTVLVKECYSVFTNRRPFGPVGFMKSAISIAEDEEWKRLRSLLSPTFTSGKLKEMVPII  120
            IKTVLVKECYSVFTNRRPFGPVGFMKSAISIAEDEEWKRLRSLLSPTFTSGKLKEMVPII
Sbjct  61   IKTVLVKECYSVFTNRRPFGPVGFMKSAISIAEDEEWKRLRSLLSPTFTSGKLKEMVPII  120

Query  121  AQYGDVLVRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDSLNNPQDPFVENTKK  180
            AQYGDVLVRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDS  NPQDPFVENTKK
Sbjct  121  AQYGDVLVRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDS--NPQDPFVENTKK  178

Query  181  LLRFDFLDPFFLSITVFPFLIPILEVLNICVFPREVTNFLRKSVKRMKESRLEDVDFLQL  240
            LLRF     FFLSITVFPFLIPILEVLNICVFPREVTNFLRKSVKRMKESR     FLQL
Sbjct  179  LLRF-----FFLSITVFPFLIPILEVLNICVFPREVTNFLRKSVKRMKESR-----FLQL  228

Query  241  MIDSQ----ALSDLELVAQSIIFIFAGYETTSSVLSFIMYELATHPDVQQKLQEEIDAVL  296
            MIDSQ    ALSDLELVAQSIIFIFAGYETTSSVLSFIMYELATHPDVQQKLQEEIDAVL
Sbjct  229  MIDSQNSHKALSDLELVAQSIIFIFAGYETTSSVLSFIMYELATHPDVQQKLQEEIDAVL  288

Query  297  PNKAPPTYDTVLQMEYLDMVVNETLRLFPIAMRLERVCKKDVEINGMFIPKGVVVMIPSY  356
            PNKAPPTYDTVLQMEYLDMVVNETLRLFPIAMRLERVCKKDVEINGMFIPKGVVVMIPSY
Sbjct  289  PNKAPPTYDTVLQMEYLDMVVNETLRLFPIAMRLERVCKKDVEINGMFIPKGVVVMIPSY  348

Query  357  ALHRDPKYWTEPEKFLPERFSKKNKDNIDPYIYTPFGSGPRNCIGMRFALMNMKLALIRV  416
            ALHRDPKYWTEPEKFLPERFSKKNKDNIDPYIYTPFGSGPRNCIGMRFALMNMKLALIRV
Sbjct  349  ALHRDPKYWTEPEKFLPERFSKKNKDNIDPYIYTPFGSGPRNCIGMRFALMNMKLALIRV  408

Query  417  LQNFSFKPCKETQIPLKLSLGGLLQPEKPVVLKVESR  453
            LQNFSFKPCKETQIPLKLSLGGLLQPEKPVVLKVESR
Sbjct  409  LQNFSFKPCKETQIPLKLSLGGLLQPEKPVVLKVESR  445
```

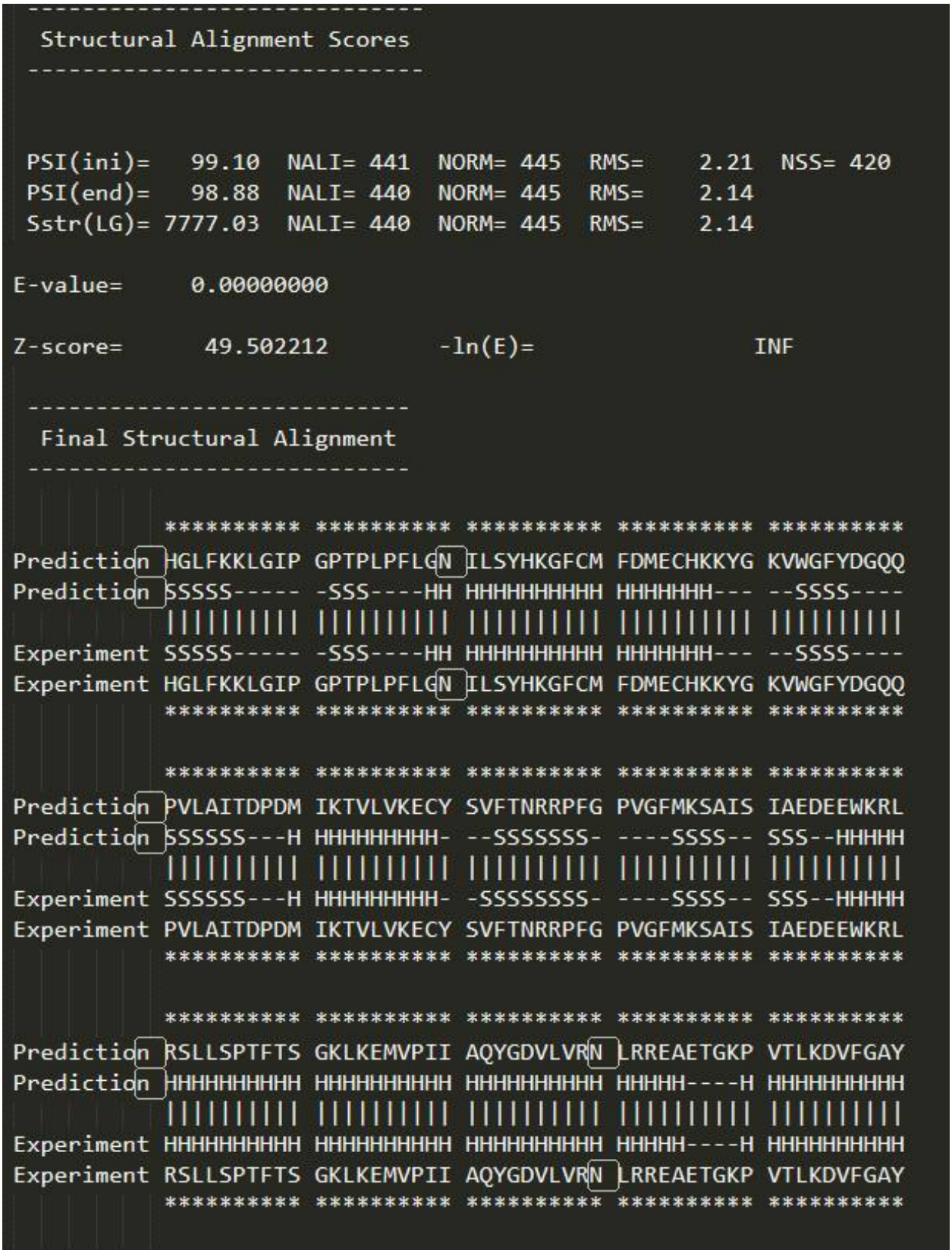
Podemos observar que la identidad cambia mucho, pues entre d2v0ma1 y d2j0da1 es del 96% ya que ambas son proteínas humanas, mientras que la identidad entre d2j0da1 y d1dt6a la identidad es de 26% pues una proteína es de humano mientras que la otra es de conejo.

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_estructlsural/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.

Alineamiento estructural de d2j0da1 contra d1dt6a:

```
 ----------------------------
  Structural Alignment Scores
 ----------------------------


 PSI(ini)=   95.28  NALI= 424  NORM= 445  RMS=    5.00  NSS= 360
 PSI(end)=   85.39  NALI= 380  NORM= 445  RMS=    3.37
 Sstr(LG)= 5580.54  NALI= 380  NORM= 445  RMS=    3.37

E-value=     0.40115480E-17

Z-score=      42.332242        -ln(E)=      40.057354

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

                * ** ***  ***** * ******** ********** **********
Prediction HGLFKKLGIP GPTPLPFLGN ILSYHKGFCM FDMECHKKYG KVWGFYDGQQ
Prediction SSSSS----- -SSS----HH HHHHHHHHHH HHHHHHH--- --SSSS----
                |||| |||  ||||      |||||| |||||||||| ||||||||||
Experiment ------SSSS -SS-S----S SSSSS---HH HHHHHHH--- --SSSS----
Experiment ......PPGP TPF.PIIGNI LQIDAKDISK SLTKFSECYG PVFTVYLGMK
                * ** ***  ***** * ******** ********** **********

           ********** ******* ** ********** *  ******* **********
Prediction PVLAITDPDM IKTVLVKECY SVFTNRRPFG P..VGFMKSA ISIAEDEEWK
Prediction SSSSSS---H HHHHHHHHH- -SSSSSSS- ------SSSS --SSS--HHH
           |||||||||| ||||||| || ||||    |       |||| ||||||||||
Experiment -SSSSS---H HHHHHHH-HH HHHHHHHHHH HH----SSS- -------HHH
Experiment PTVVLHGYEA VKEALVD.LG EEFAGRGSVP ILEKVSKGLG IAFSNAKTWK
           ********** ******* ** ********** *  ******* **********

           ********** ****  **** ********** ********** **********
Prediction RLRSLLSPTF TSGKLKEMVP IIAQYGDVLV RNLRREAETG KPVTLKDVFG
Prediction HHHHHHHHHH HHHHHHHHHH HHHHHHHHHH HHHHHHH--- -HHHHHHHHH
           |||||||||         ||| |||||||||| ||     ||| ||||||||||
Experiment HHHHHHHHHH HHHH---HHH HHHHHHHHHH HHHHHHHH-- --HHHHHHHH
Experiment EMRRFSLMTL RNFGMGKRSI EDRIQEEARC LVEELRKTNA SPCDPTFILG
           ********** ****  **** ********** ********** **********
```

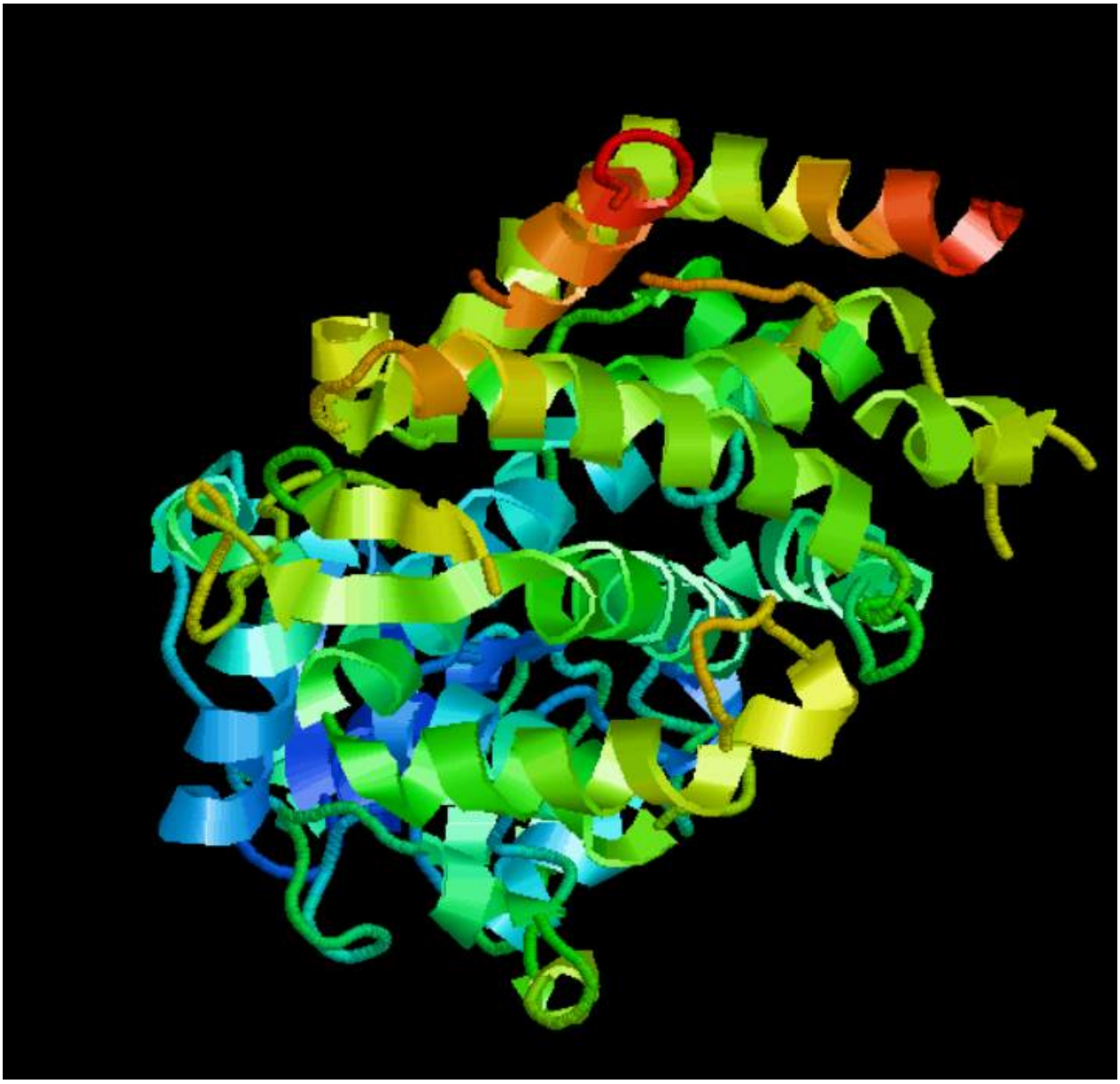
Alineamiento estructural de d2v0ma1 contra d2j0da1:



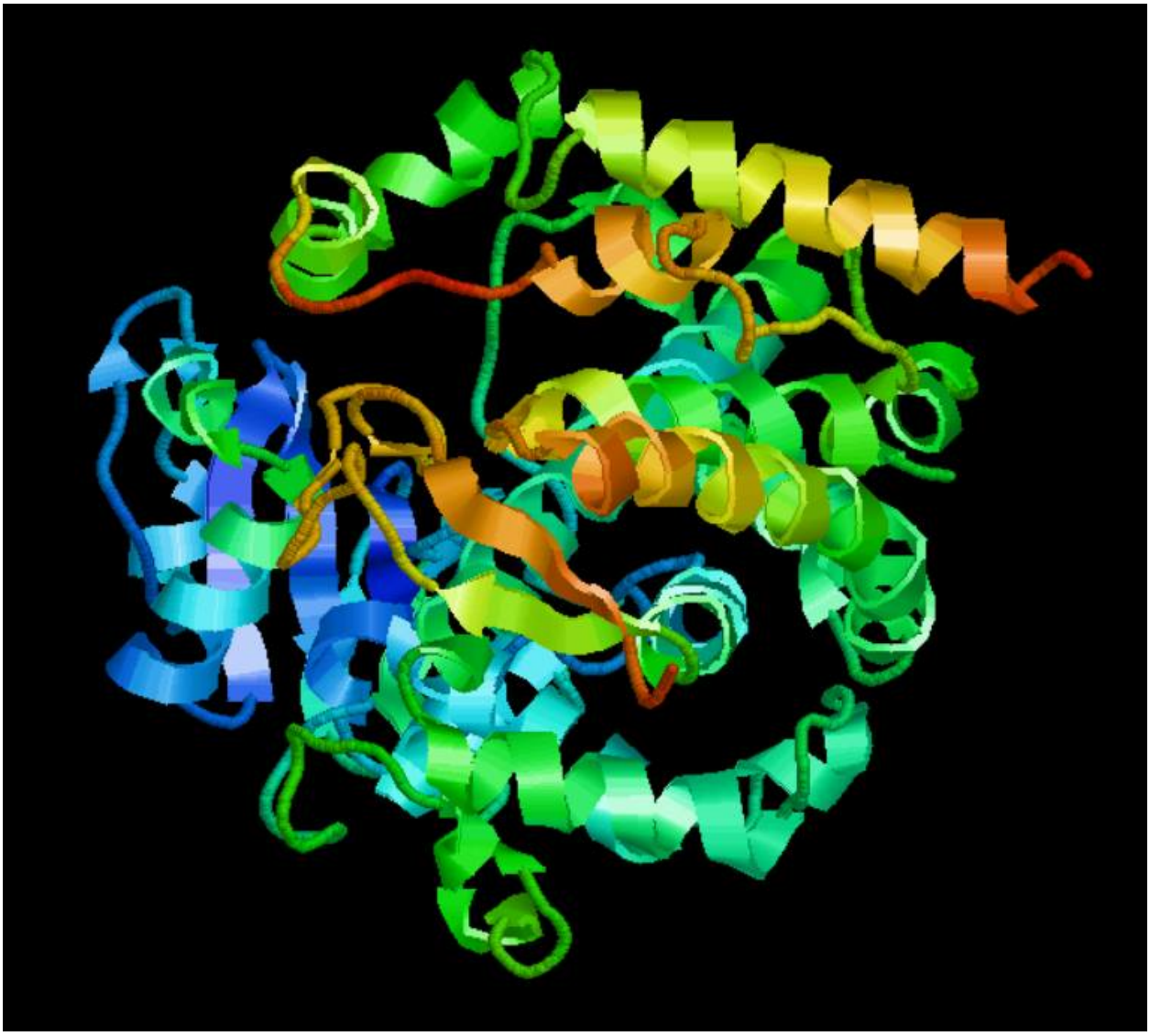
5) Compara los alineamientos obtenidos en 3 y 4. Comenta en qué elementos de estructura secundaria se

observan diferencias.

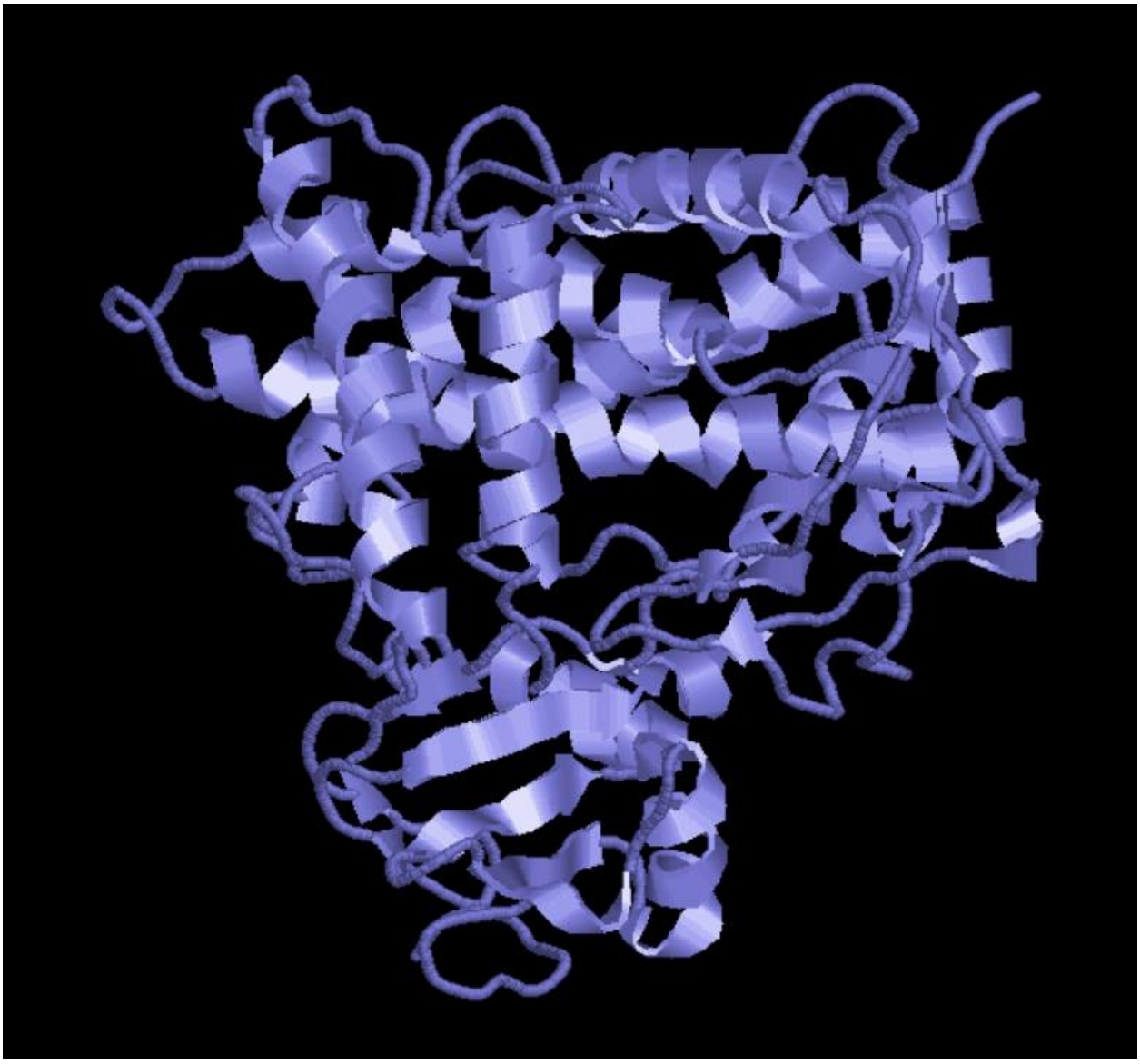
d2j0da1:



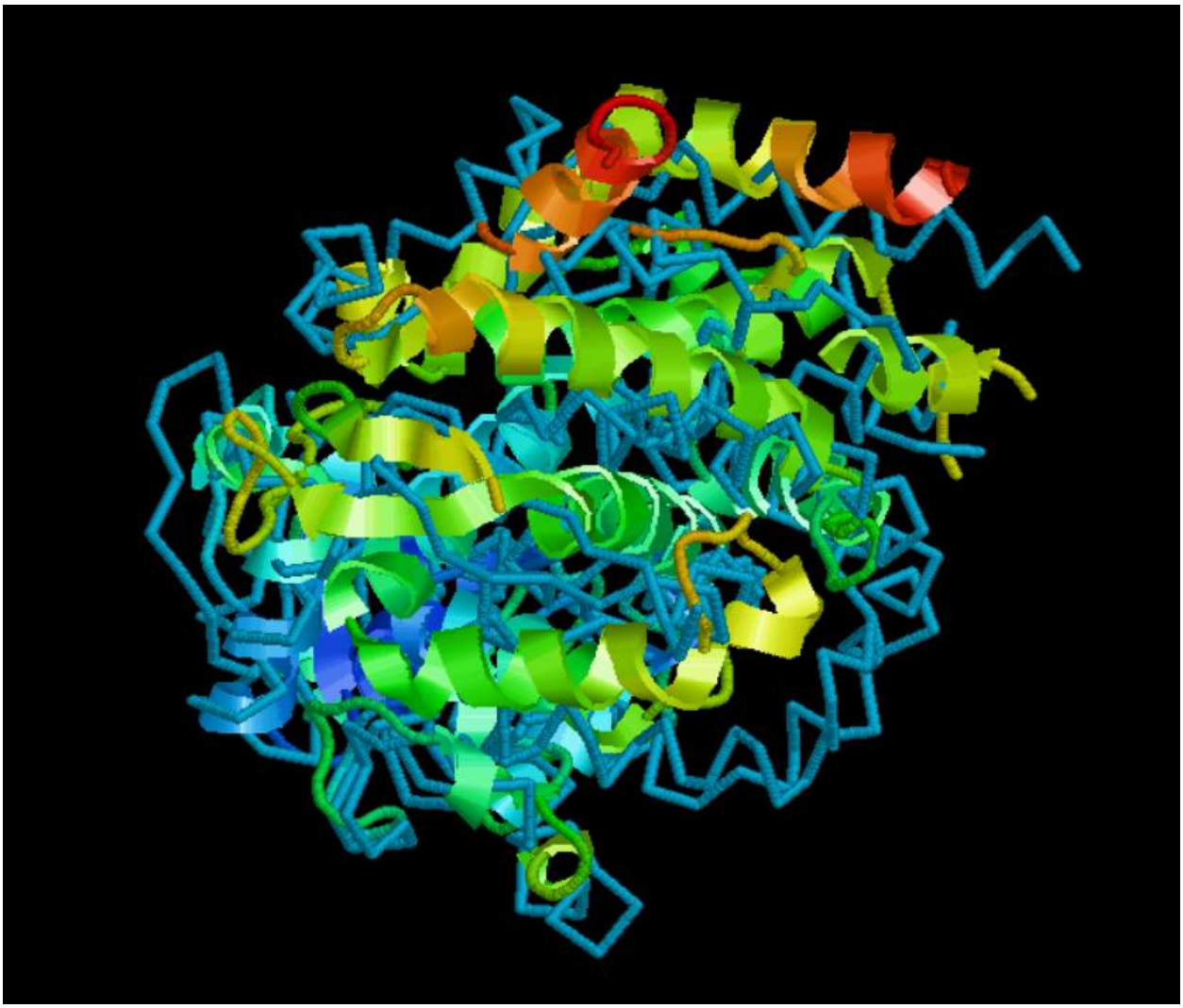
d2v0ma1:



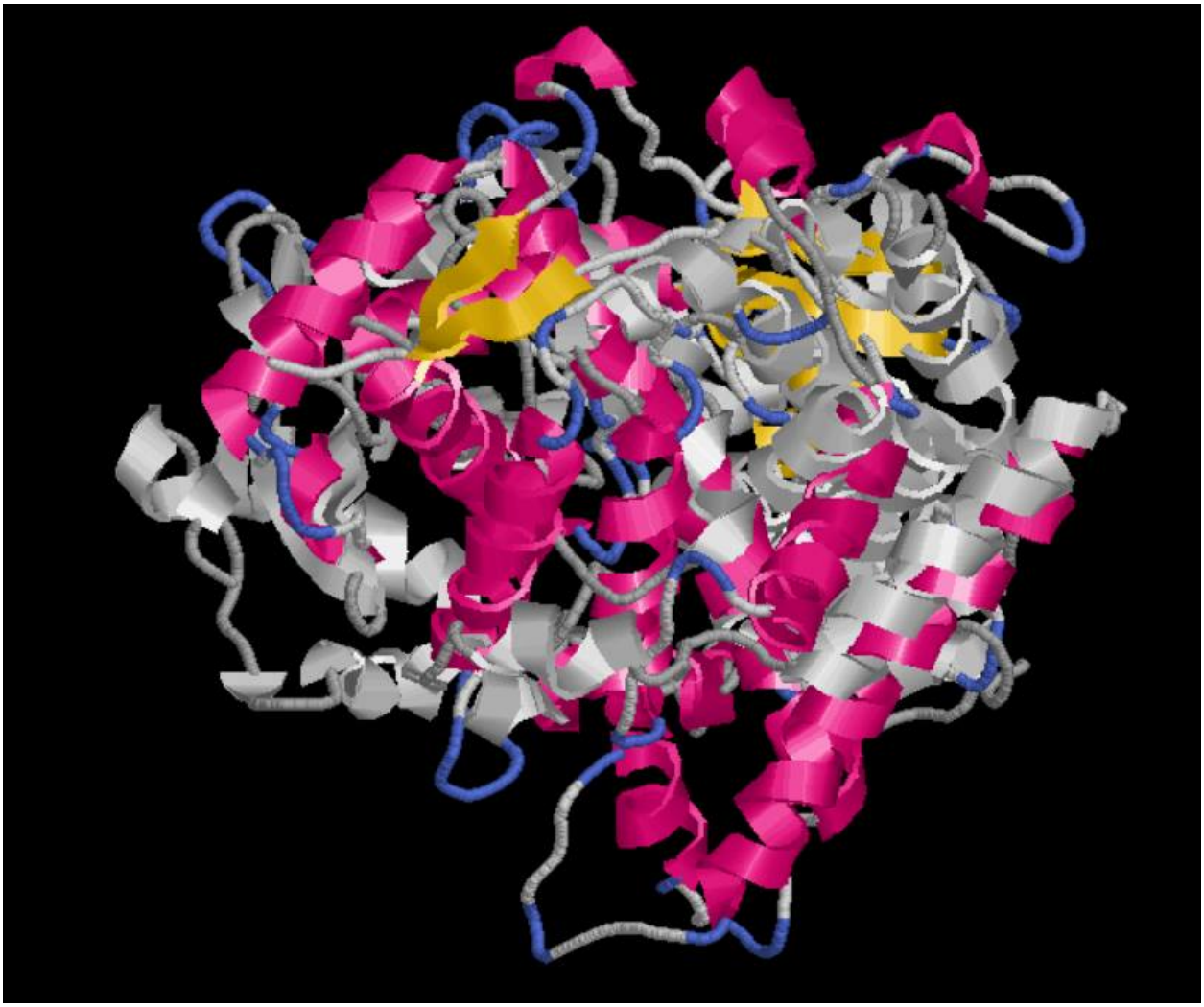
d1dt6a:



d2v0ma1 vs d2j0da1:



d2j0da1 vs d1dt6a:



En las tres proteínas podemos observar que hay pequeñas regiones en las que hay ojas plegadas, entre las proteínas que pertenecen a humano esta región es muy similar, sin embargo está región parece un poco diferente en la proteína de conejo; sin embargo las proteínas son muy similares, y en base a los z scores concluimos que aunque la proteína de conejo es mucho más diferente en secuencia que las proteínas de humano, parece que las diferencias estructurales entre las proteínas de humano y de conejo no son mucho más diferentes que las dos proteínas de humano entre sí, por lo tanto, la estructra se conserva mucho más que la secuencia.

6) Utiliza el prog3.1 (en 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?

d2v0ma1 vs d2j0da1

```
# total residuos: pdb1 = 445 pdb2 = 455
# total residuos alineados = 441

# coordenadas originales = original.pdb
# superposicion optima:
# archivo PDB = align_fit.pdb
# RMSD = 1.74 Angstrom
```

d2j0da1 vs d1dt6a

```
# total residuos: pdb1 = 449 pdb2 = 445
# total residuos alineados = 388

# coordenadas originales = original.pdb
# superposicion optima:
# archivo PDB = align_fit.pdb
# RMSD = 13.57 Angstrom
```

Liga al programa: https://github.com/carinapaola/Bioinformatica_estructural/blob/master/RMSD.ipynb

Los valores de RMSD representan la diferencia entre dos secuencias, podemos observar que la diferencia entre las proteínas humanas (1.74 Angstrom)es mucho menor que cuando se comparan una de las proteínas humanas con la de conejo (13.57 Angstrom).