



Tutorial Modelado por Homología

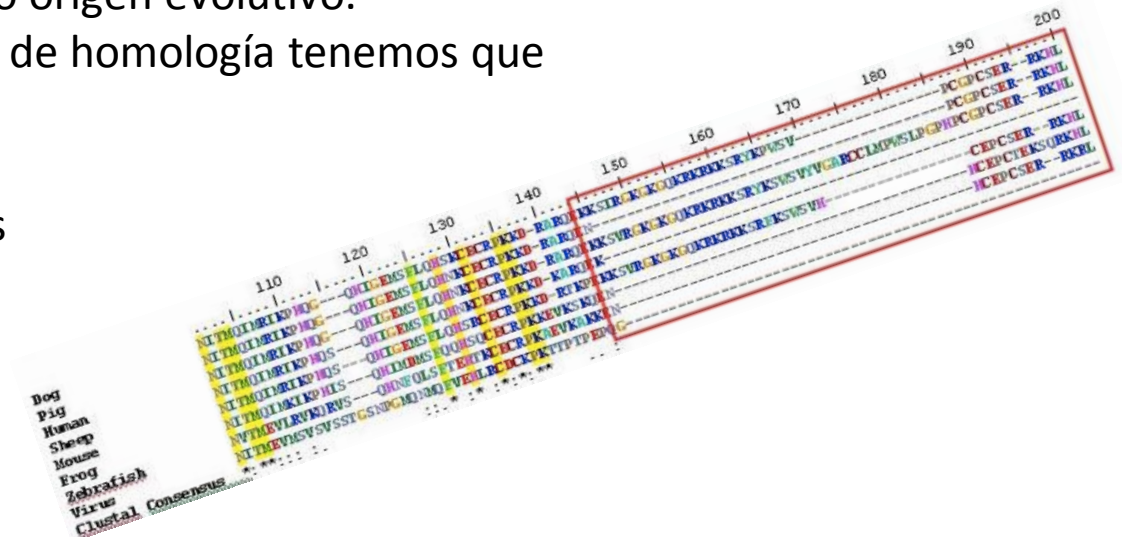
Alineamiento de Secuencias

¿Por qué alinear dos secuencias?

- Similitud : secuencias de cualquier origen que se parecen
- Homología: Secuencias misma especie o especies diferentes con misma función y mismo origen evolutivo.

Por lo tanto cuando hablamos de homología tenemos que considerar:

- Evolución
- Similar función o propiedades



Alineamiento de Secuencias

Alineamiento Global

-Se pretende alinear la secuencia entera empleando tantos caracteres como sea posible de los extremos de las secuencias. Es un alineamiento que se extiende a lo largo de toda la longitud de las secuencias utilizadas.

Global FTFTALILLAVAV
F--TAL-LLA-AV

-Una estrategia general de alineamiento global es el algoritmo de Needleman-Wuncsch basado en programación dinámica.

-Secuencias tamaño parecido

Alineamiento Local

-Se buscan las porciones de las secuencias que presentan mayor cantidad de concordancias.

-El algoritmo de Smith-Waterman es un método general de alineamiento local basado en programación dinámica

-Secuencias tamaño diferente pero se espera que tengan regiones parecidas.

Local FTFTALILL-AVAV
--FTAL-LLAAV--

Alineamiento Local: BLAST

The Basic Local Alignment Search Tool es un programa proporcionado por NCBI que encuentra regiones locales de similitud entre secuencias, este compara secuencias de nucleótidos o proteínas con la base de datos de secuencias y calcula un estadístico significativo.



Algunos tipos...

Blastn: Compara una secuencia de nucleótidos contra una base de datos que contenga también secuencias nucleotídicas.

Blastp: Es un BLAST "con huecos" (o *gaps*) que compara una secuencia de aminoácidos contra una base de datos del mismo tipo. (Usualmente usa la matriz BLOSUM o PAM para realizar los alineamientos, aunque puede usar una matriz definida por el usuario).

Blastx: Consulta la base de datos de nucleótidos traducida utilizando la base de datos de nucleótidos traducida.

Parte Práctica: Alineamiento de secuencias

Utilizando BLAST Blastn

NCBI/ BLAST/ blastn suite Standard Nucleotide B

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

TCGAAATAACGCGTGTCTCAACGCGGTGCGGCAGATGCCTTTGCTCATCAGATGCGACCGC
AACCACGTCCGCGCGCTTGTTCGCGGTCCCGTGCCTCAACCACCACCACGGTGTCGTCTTC
CCCGAACGCGTCCCGGTACGCCAGCCTCCACGCGCGCGCGCGGAGTGCCCATTCGGGC
CGCAGCTGCGACGGTGCCGCTCAGATTCTGTGTGGCAGGCGCGTGTGGAGCTAAA

Query subrange [?](#)

From

To

Or, upload file ningún ...cionado [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database ☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):

☒ Nucleotide collection (nr/nt) [?](#)

Organism [Optional](#) ☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude [Optional](#) ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query [Optional](#) [?](#)

Program Selection

Optimize for ☐ Highly similar sequences (megablast) ☐ More dissimilar sequences (discontiguous megablast) ☒ Somewhat similar sequences (blastn) [?](#)

Choose a BLAST algorithm [?](#)

1- Pegar la secuencia de nucleótidos.

2- Definir la Base de Datos donde se buscará.

3- Selección del programa

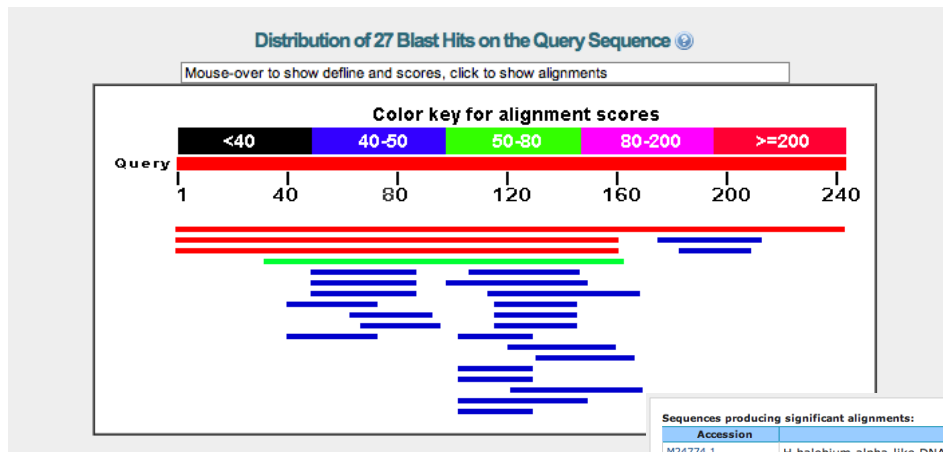
Saldrá una ventana con el título del trabajo y se espera unos segundos para que re-direccione

Job Title: Nucleotide Sequence (242 letters)

Request ID	V84RSESY01N
Status	Searching
Submitted at	Wed May 16 16:50:00 2012
Current time	Wed May 16 16:50:19 2012
Time since submission	00:00:19

This page will be automatically updated in 17 seconds

Obtenemos los resultados en forma gráfica y también nos muestra una lista ordenada con la información



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
M24774.1	H.halobium alpha-like DNA polymerase gene, partial cds	437	437	100%	1e-119	100%	
AM774415.1	Halobacterium salinarum complete genome, strain R1	251	251	66%	1e-63	96%	
AE004437.1	Halobacterium sp. NRC-1, complete genome	251	251	66%	1e-63	96%	
CP002062.1	Halalkalicoccus jeotgali B3, complete genome	60.8	60.8	53%	4e-06	73%	
CP001722.1	Zymomonas mobilis subsp. mobilis NCIMB 11163, complete genome	48.2	48.2	15%	0.027	89%	
CP002019.1	Ketogulonigenium vulgare WSH-001 plasmid 1, complete sequence	42.8	42.8	15%	1.1	87%	
AP012204.1	Microlunatus phosphovorus NM-1 DNA, complete genome	42.8	42.8	15%	1.1	84%	
CP002225.1	Ketogulonigenium vulgare Y25 plasmid pYP1, complete sequence	42.8	42.8	15%	1.1	87%	
CP001349.1	Methylobacterium nodulans ORS 2060, complete genome	42.8	42.8	10%	1.1	96%	
CP000667.1	Salinispora tropica CNB-440, complete genome	42.8	42.8	16%	1.1	86%	

Si continuamos mirando la información que entrega BLAST, vamos a llegar a la sección alineamiento, en la que se muestran las coincidencias encontradas con la base de datos:

Alignments

☐ Select All [Get selected sequences](#) [Distance tree of results](#)

> ☐ [gb|M24774.1|HALADNAPA](#) H.halobium alpha-like DNA polymerase gene, partial cds
Length=242

Score = 437 bits (484), Expect = 1e-119
Identities = 242/242 (100%), Gaps = 0/242 (0%)
Strand=Plus/Plus

Query 1 TCGAAATAACGCGTGTCTCAACGCGGTCGCGCAGATGCCTTTGCTCATCAGATGCGACC 60
Sbjct 1 TCGAAATAACGCGTGTCTCAACGCGGTCGCGCAGATGCCTTTGCTCATCAGATGCGACC 60

Query 61 GCAACCACGTCCGCCGCTTGTTCGCCGTCCCGTGCCTCAACCACCACCACGGTGTCTGT 120
Sbjct 61 GCAACCACGTCCGCCGCTTGTTCGCCGTCCCGTGCCTCAACCACCACCACGGTGTCTGT 120

Query 121 CTTCCCCGAACGCGTCCCGGTTCAGCCAGCCTCCACgagcgcgcgcgcgcgcgGAGTGCCCAT 180
Sbjct 121 CTTCCCCGAACGCGTCCCGGTTCAGCCAGCCTCCACGCGCGCGCGCGGAGTGCCCAT 180

Query 181 CGGGCCGACGCTGCGACGGTGCCGCTCAGATTCTGTGTGGCAGGCGCGTGTGGAGTCTA 240
Sbjct 181 CGGGCCGACGCTGCGACGGTGCCGCTCAGATTCTGTGTGGCAGGCGCGTGTGGAGTCTA 240

Query 241 AA 242
Sbjct 241 AA 242

100% de coincidencia
esa secuencia.

Para obtener mayor
información pinchar en
el número de acceso.

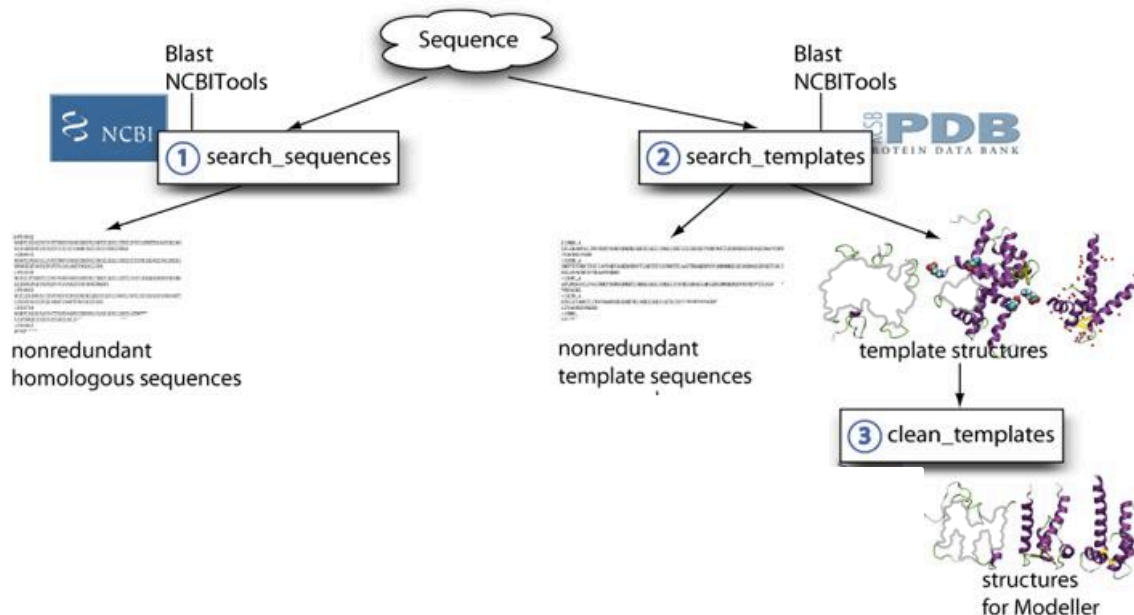
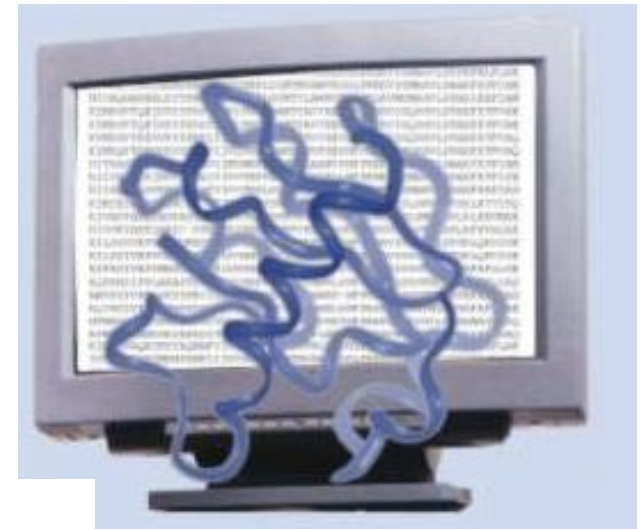
De esta búsqueda obtener la siguiente
información:

- Largo de la secuencia
- La identidad más probable
- Organismo a que pertenece
- Número de acceso
- valor E (e-value)

¿Qué otra información podemos utilizar de BLAST?

Supongamos tenemos una secuencia con estructura desconocida y deseamos crear un modelo estructural para esta secuencia.

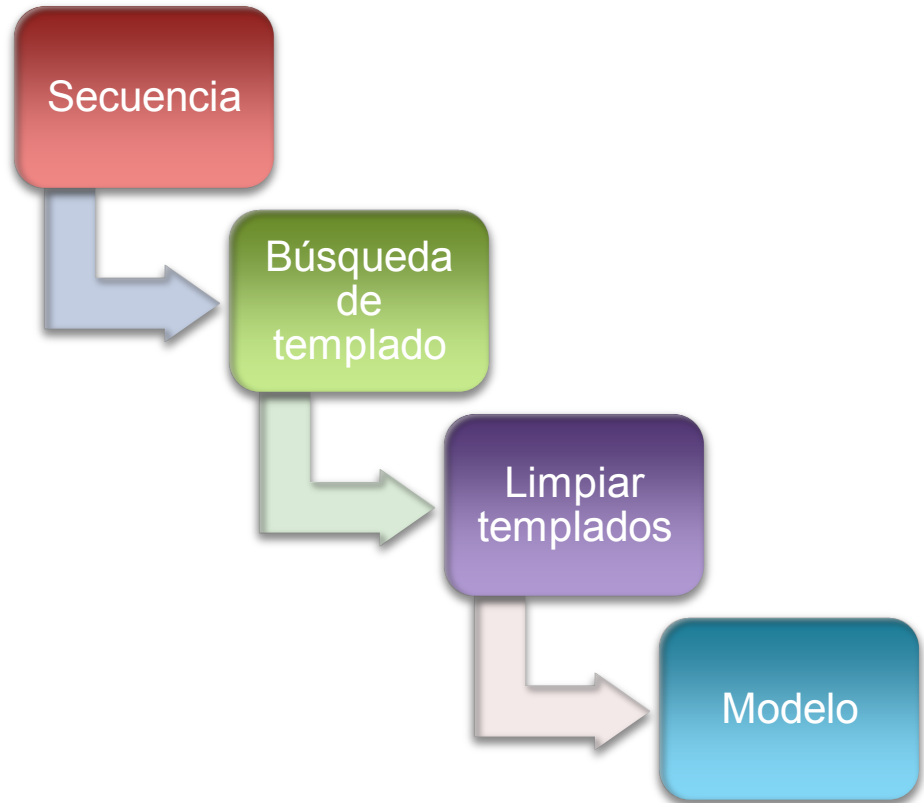
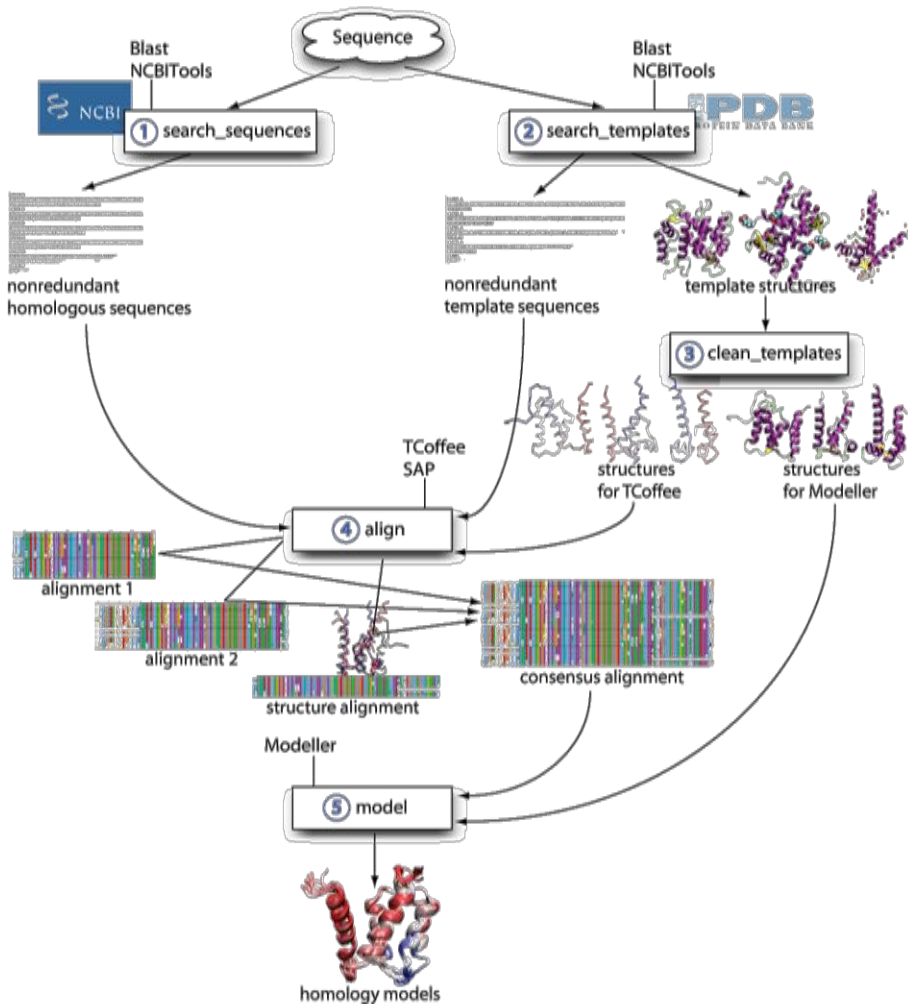
Con ayuda de BLAST podemos encontrar un cristal y realizar un MODELO POR HOMOLOGÍA



Modelo por homología (simple).*

- Usaremos [Modeller](#)
- Libre para uso académico.
- <http://salilab.org/modeller/9v6/modeller9v6.exe>
- Clave de Licencia: MODELIRANJE
- Modeller es una herramienta donde puedes controlar todos los aspectos del proceso de modelado por homología.
- Modeller no tiene interfaz gráfica. Para usarlo tenemos que escribir scripts en python.

+ Proceso:



Secuencia

ENLYFQSMINSFYAFEVKDAKGRTVSLEKYKGK
VSLVVNVASDCQLTDRNYLGLKELHKEFGPSHF
SVLAFPCNQFGESEPRPSKEVESFARKNYGVTF
PIFHKIKILGSEGEPAFRFLVDSSKKEPRWNFWK
YLVNPEGQVVKFWRPEEPIEVIRPDIAALVRQVII
KKKEDL

T0388 LOC493869A, Homo sapiens



CASP target ID

(Critical assessment of techniques
for protein structure prediction)

Búsqueda de templado:

BLAST contra PDB

BLASTP programs search protein databases using a protein query. [more...](#)

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [?](#)

[Clear](#)

Query subrange [?](#)

From

To

ENLYFQSMINSFYAFEVVKDAKGRTVSLEKYKGKVSLLVNVASDCQLTDRNYLGLKELHKE
FGPSHFVLAFFPCNQFGESEPRPSKEVESFARKNYGVTFPIFHKIKILGSEGEPAFRFLV
DSSKKEPRWNFWKYLVNPEGQVVKFWRPEEPIEVIRPDIAALVRQVIKKKEDL

Or, upload file [Browse...](#) [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)





☐ Align two or more sequences [?](#)

Choose Search Set

Database [?](#)




Seleccionando templado

- Existe el cristal de esta secuencia.
- Ignoraremos esta información, y usaremos la cadena A de 2p31.pdb

```
>  pdb|3CYN|A  Chain A, The Structure Of Human Gpx8  
  pdb|3CYN|B  Chain B, The Structure Of Human Gpx8  
  pdb|3CYN|C  Chain C, The Structure Of Human Gpx8  
Length=189
```

```
Score = 357 bits (917), Expect = 9e-100, Method: Compositional matrix adjust.  
Identities = 174/174 (100%), Positives = 174/174 (100%), Gaps = 0/174 (0%)
```

```
Query 1 ENLYFQSMINSFYAFEVKDAKGRTVSLEKYKGKVSLLVNVASDCQLTDRNYLGLKELHKE 60  
      ENLYFQSMINSFYAFEVKDAKGRTVSLEKYKGKVSLLVNVASDCQLTDRNYLGLKELHKE  
Sbjct 16 ENLYFQSMINSFYAFEVKDAKGRTVSLEKYKGKVSLLVNVASDCQLTDRNYLGLKELHKE 75  
  
Query 61 FGPSHFSVLAFFPCNQFGESEPRPSKEVESFARKNYGVTFPIFHKIKILGSEGEPAFRFLV 120  
      FGPSHFSVLAFFPCNQFGESEPRPSKEVESFARKNYGVTFPIFHKIKILGSEGEPAFRFLV  
Sbjct 76 FGPSHFSVLAFFPCNQFGESEPRPSKEVESFARKNYGVTFPIFHKIKILGSEGEPAFRFLV 135  
  
Query 121 DSSKKEPRWNFWKYLVNPEGQVVKFWRPEEPIEVIRPDIAALVRQVIKKKEDL 174  
      DSSKKEPRWNFWKYLVNPEGQVVKFWRPEEPIEVIRPDIAALVRQVIKKKEDL  
Sbjct 136 DSSKKEPRWNFWKYLVNPEGQVVKFWRPEEPIEVIRPDIAALVRQVIKKKEDL 189
```

```
>  pdb|2P31|A  Chain A, Crystal Structure Of Human Glutathione Peroxidase 7  
  pdb|2P31|B  Chain B, Crystal Structure Of Human Glutathione Peroxidase 7  
Length=181
```

```
Score = 210 bits (534), Expect = 3e-55, Method: Compositional matrix adjust.  
Identities = 95/166 (57%), Positives = 123/166 (74%), Gaps = 2/166 (1%)
```

```
Query 1 ENLYFQSMIN--SFYAFEVKDAKGRTVSLEKYKGKVSLLVNVASDCQLTDRNYLGLKELH 58  
      ENLYFQSM      FY F+ + +G+ VSLEKY+G VSLVNVAS+C TD++Y L++L  
Sbjct 16 ENLYFQSMQQEQDFYDFKAVNIRGKLVSEKYRGSVSLVNVASECGFTDQHYRALQQLQ 75  
  
Query 59 KEFGPSHFSVLAFFPCNQFGESEPRPSKEVESFARKNYGVTFPIFHKIKILGSEGEPAFRF 118  
      ++ GP HF+VLAFFPCNQFG+ EP +KE+ESFAR+ Y V+FP+F KI + G+ PAF++  
Sbjct 76 RDLGPHHFNVLAFFPCNQFGQQEPDSNKEIESFARRTYSVSFPMFSKIAVTGTGAHPAFKY 135  
  
Query 119 LVDSSKKEPRWNFWKYLVNPEGQVVKFWRPEEPIEVIRPDIAALVR 164  
      L +S KEP WNFWKYLV P+G+VV W P +E +RP I ALVR  
Sbjct 136 LAQTSKGKEPTWNFWKYLVA PDGKVVGAWDPTVSVEEVRPQITALVR 181
```

Limpiando templados: Creando alineamiento

- Modeller tiene una sofisticada herramienta de alineamiento.
 - Usa la información estructural del template
 - Usa programación dinámica en vez del método de búsqueda de blast.
- Para crear el alineamiento necesitas:
 1. Bajar el archivo PDB de template.
 2. Poner tu secuencia en formato PIR.
 3. Editar el script de alineamiento ([alignment script](#)) en función del template y la cadena.
 4. Ejecutar en modeller: `mod9v6.exe align.py`

Archivo PIR

- Reemplaza la secuencia por la tuya.
- La última línea en debe terminar en *
- No toques nada mas del archivo, o sino no funcionará.
- Nombre del archivo: target.ali

```
>P1;target
```

```
sequence:target:0.00: 0.00
```

```
ENLYFQSMINSFYAFEVKDAKGRTVSLEKYKGKVSLVNVASDCQLTDRNYLGLKELHKE  
FGPSHFSLAFPCNQFGESEPRPSKEVESFARKNYGVTFPIFHKIKILGSEGEPAFRFLV  
DSSKKEPRWNFWKYLVNPEGQVVKFWRPEEPIEVIRPDIAALVRQVIKKKEDL*
```

Align.py

```
from modeller import *  
from modeller.automodel import *
```

```
env = environ()  
aln = alignment(env)
```

```
template='2p31'  
chain='A'
```

← Just change the value of these 2 lines
with your template

```
tc=template+chain
```

```
mdl = model(env, file=template, model_segment=('FIRST:'+chain,'LAST:'+chain))  
aln.append_model(mdl, align_codes=tc, atom_files=template+'.pdb')  
aln.append(file='target.ali', align_codes='target')  
aln.align2d()  
aln.write(file='target-'+tc+'.ali', alignment_format='PIR')  
aln.write(file='target-'+tc+'.pap', alignment_format='PAP')
```


- El alineamiento es diferente al producido por BLAST.
- Modeller ignora los residuos con poca información estructural.

```

Query    1      ENLYFQSMIN--SFYAFEVKDAKGRTVSLEKYKGKVSLVVNVASDCQLTDRNYLGLKELH   58
          ENLYFQSM      FY F+  + +G+ VSLEKY+G VSLVVNVAS+C  TD++Y  L++L
Sbjct   16      ENLYFQSMQQEQDFYDFKAVNIRGKLVSLEKYRGSVSLVVNVASECGFTDQHYRALQQLQ   75

Query    59      KEFGPSHFVLAFFPCNQFGGESEPRPSKEVESFARKNYGVTFPIFHKIKILGSEGEPAFRF   118
          ++ GP HF+VLAFFPCNQFG+ EP  +KE+ESFAR+ Y V+FP+F KI + G+   PAF++
Sbjct   76      RDLGPHHFNVLAFPCNQFGQQEPDSNKEIESFARRTYSVSFPMFSKIAVTGTGAHPAFKY   135

Query   119      LVDSSKKEPRWNFWKYLVNPEGQVVKFWRPEEPIEVIRPDIAALVR   164
          L  +S KEP WNFWKYLV P+G+VV  W P    +E +RP I ALVR
Sbjct  136      LAQTSGKEPTWNFWKYLVA PDGKVVGA WDPTVSVEEVRPQITALVR   181

```

```

_aln.pos      10      20      30      40      50      60
2p31A  -----Q----DFYDFKAVNIRGKLVSLEKYRGSVSLVVNVASECGFTDQHYRALQQLQRDLGPHHFN
target
ENLYFQSMINSFYAFEVKDAKGRTVSLEKYKGKVSLVVNVASDCQLTDRNYLGLKELHKEFGPSHFV
_consrvd      *      * * *      *  * * * * * * * * * * * * * * * * * * *

```

```

_aln.p  70      80      90     100     110     120     130
2p31A
LAFPCNQFGQQEPDSNKEIESFARRTYSVSFPMFSKIAVTGTGAHPAFKYLAQTSGKEPTWNFWKYL
target
LAFPCNQFGGESEPRPSKEVESFARKNYGVTFPIFHKIKILGSEGEPAFRFLVDSSKKEPRWNFWKYL
_consrvd * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

```

```

_aln.pos 140     150     160     170

```

Modelo

```
from modeller import *  
from modeller.automodel import *
```

```
log.verbose()  
env = environ()
```

```
template='2p31'  
chain='A'
```

```
tc=template+chain
```

```
class MyModel(automodel):  
    def get_model_filename(self, sequence, id1, id2,  
file_ext):  
        return sequence+'_'+id2`+file_ext
```

```
    def special_restraints(self, aln):  
        rsr = self.restraints
```

```
a = MyModel(env, alnfile='target-'+tc+'.ali',  
            knowns=tc, sequence='target',  
            assess_methods=(assess.DOPE, assess.GA341))  
a.starting_model = 1  
a.ending_model = 5  
a.make()
```

- 5 modelos son creados
- Cada uno de ellos es ligeramente diferente.

Resultados de modelado

>> Summary of successfully produced models:

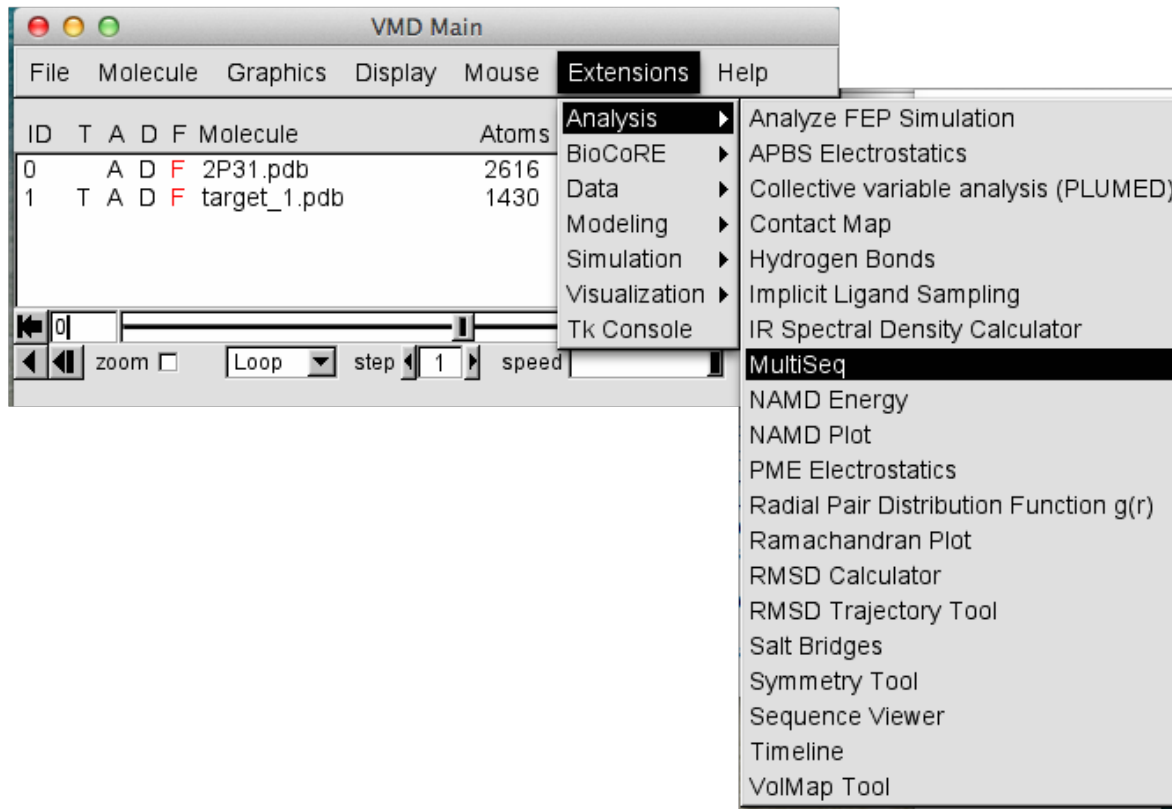
Filename	molpdf	DOPE score	GA341 score

target_1.pdb	1280.53101	-19077.32812	1.00000
target_2.pdb	1570.33606	-18480.83008	1.00000
target_3.pdb	960.32550	-19365.79102	1.00000
target_4.pdb	1415.41724	-18980.71094	1.00000
target_5.pdb	1463.82593	-19077.91016	1.00000

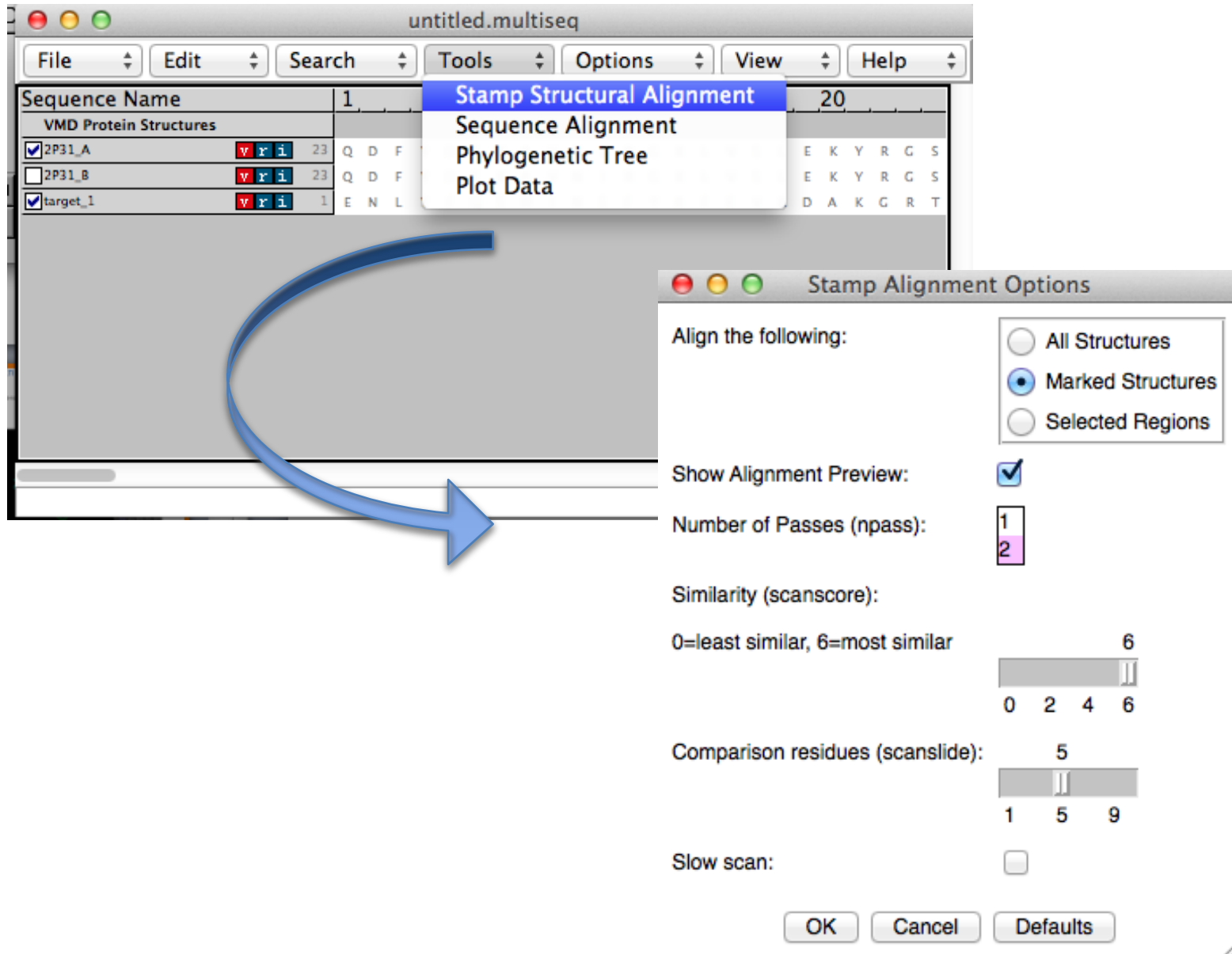
- De acuerdo al puntaje DOPE, el 3 es el mejor modelo y el 2 es el peor.
- El puntaje DOPE mas bajo, es el mejor.
- Veamos que tan diferentes son los modelos.

Viendo los dos PDBs en VMD.

1. Abrir los dos archivos como lo suelen hacer.
2. Ir a extensiones → Multiseq.

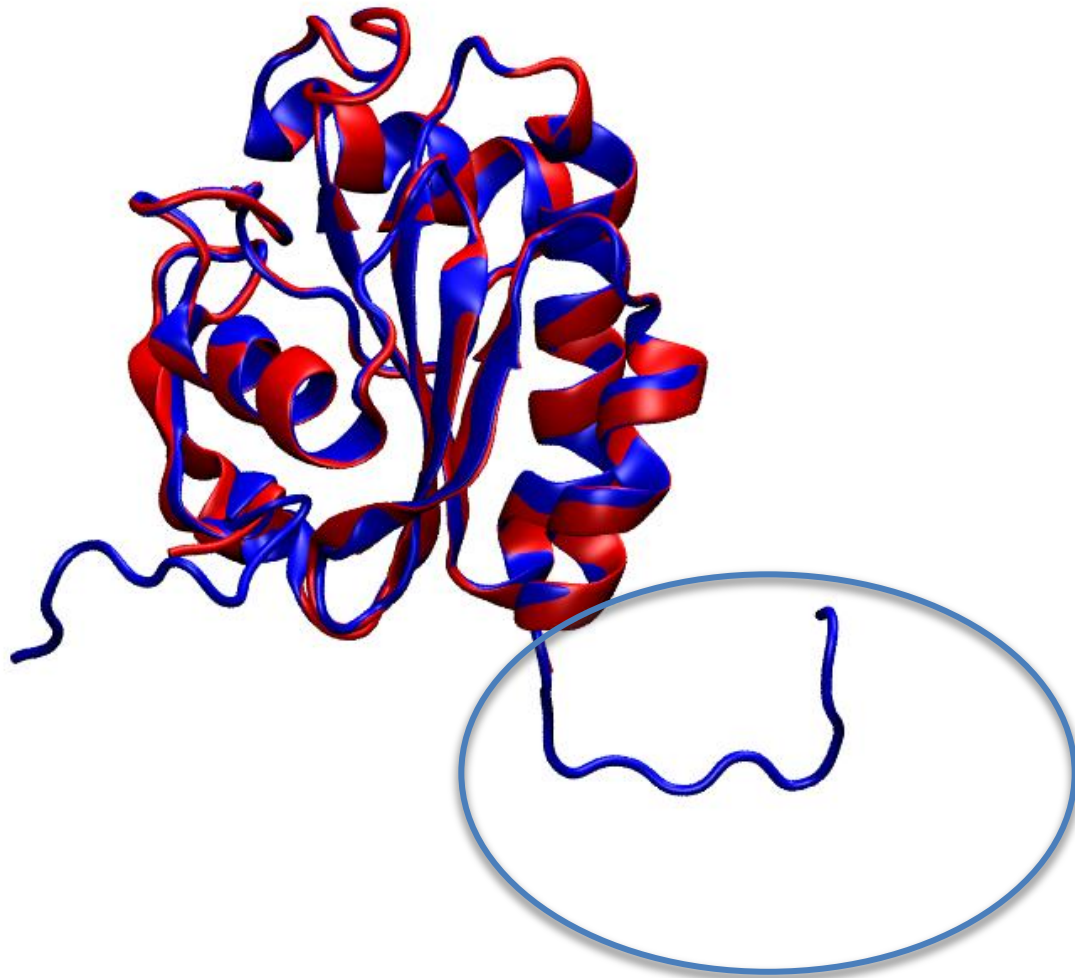


Alinear ambas estructuras.



Proteínas alineadas:

2P31: ROJO
MODELO: AZUL



Actividad próxima clase

Hacer un modelo con la siguiente secuencia

MKEGDSLLRNVAGPLGTPVPMKFFHKILAIGAYTG
IVEVYPIAKAWQEIGNDVTTLHVTFEPMVILKEELEK
AVTRADEEHIVEPVPLNPNADFLANMKNVSQRLKE
KVRELLESEMAAARDWDLVFMVRPVGDKQVFEV
VWKEYQVPMKVLDLHPIMVDGLE

Analiza cual fue el mejor modelo obtenido.
Compara el mejor modelo obtenido con el templado utilizado.
Adjunta una captura de las estructuras alineadas.