ALINEAMIENTO MÚLTIPLE DE SECUENCIAS
EJERCICIOS

Ejercicio 1

Hazte familiar con los tres modos de encontrar secuencias de interés para alineamientos múltiples: HomoloGene, CDD y BLAST.

- Busca una determinada palabra clave (por ejemplo cytochrome, ferritin, S100 o trypsin) en NCBI Entrez HomoloGene. Identifica un grupo de proteínas homólogas y haz click en su enlace. Puedes descargarte sus secuencias FASTA o ver su alineamiento múltiple desde la página de su enlace.
- Haz la misma búsqueda pero en Entrez Conserved Domains, que contiene dominios conservados de las bases de datos Pfam, CDD, Smart y COG. Selecciona una con identificador CDD (por ejemplo cd07484 para trypsin). Puedes ver el alineamiento múltiple y árbol filogenético asociado. También puedes cambiar el formato del alineamiento a mFASTA y copiar el texto resultante a un fichero local.
- Por último, puedes hacer un BLASTP, por ejemplo sobre la cadena ligera de la ferritina (NP_000137) e inspeccionar los alineamientos de pares resultantes. Selecciona los que quieras, y al final de la página tienes opciones para obtener sus secuencias o hacer su alineamiento múltiple, y posteriormente descargarlos.

Ejercicio 2

Usando HomoloGene, busca familias de homólogos para S100, y descarga en fasta la familia con id 55916. Ahora ve a los programas para alineamiento múltiple del EBI (ClustalW, MAFFT, Muscle y T-Coffee). Realiza los alineamientos con los 4 y compara los resultados. ¿Cómo varían? ¿Podrías decidir cuál es probablemente el más preciso? En los casos en los que se pueda, trata de ajustar los parámetros (matrices de puntuación, penalizaciones a los huecos, número de iteraciones) para ver los efectos que producen en los alineamientos.

Ejercicio 3

En teoría veíamos que un problema de las iteraciones en ClustalW es que puede dar mucha importancia a una proteína divergente respecto al resto, por su política de huecos ("una vez ocurre un hueco, siempre hay un hueco")

Probemos con estas cinco beta globinas (muy relacionadas):

```
>human_NP_000509
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>Pan_troglodytes_XP_508242
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>Canis_familiaris_XP_537902
MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAKVKAHGKKVLN
SFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVAN
ALAHKYH
>Mus_musculus_NP_058652
MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPKVKAHGKKVIT
AFNEGLKNLDNLKGTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVAT
ALAHKYH
>Gallus_gallus_XP_444648
MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLT
SFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAH
ALARKYH
```

Primero hacemos el alineamiento múltiple normal, y luego lo repetimos, pero añadiendo 5 veces más la secuencia del pollo (Gallus gallus) para ver si eso distorsiona el alineamiento.

Ahora probemos con estas otras 5 globinas (menos relacionadas), y replicando 5 veces la segunda (mioglobina NP_005359.1):

```
>beta_globin 2hhbB NP_000509.1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVL
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFR
KDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVML
VIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAV
VQAMSRGWDGE
>soybean_globin 1FSL leghemoglobin P02238 LGBA_SOYBN [Glycine max]
MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLANGVDPTNPKLTGHAEKLFALV
RDSAGQLKASGTVVADAALGSVHAQKAVTDPQFVVVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAAA
IKKA
>rice_globin 1D8U rice Non-Symbiotic Plant Hemoglobin NP_001049476.1 [Oryza
sativa (japonica cultivar-group)]
MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLRNSDVPLEKNPK
LKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDAHFEVVKFALLDTIKEEVPADMWS
PAMKSAWSEAYDHLVAAIKQEMKPAE
```

¿Cuál es el efecto de dar más peso a una de las secuencias en ambos casos?¿Puedes explicar por qué sucede el efecto?

Ejercicio 4

Vamos a probar ahora el uso de información estructural para validar o mejorar nuestros alineamientos. Usaremos las siguientes 8 lipocalinas:

```
   >human_RBP4 gi|55743122|ref|NP_006735.2| retinol-binding protein 4, plasma precursor
[Homo sapiens]
   MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
   MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
   LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
   >rat_OBP gi|20302101|ref|NP_620258.1| odorant binding protein I f [Rattus norvegicus]
   MVKFLLIVLALGVSCAHHENLDISPSEVNGDWRTLYIVADNVEKVAEGGSLRAYFQHMECGDECQELKII
   FNVKLDSECQTHTVVGQKHEDGRYTTDYSGRNYFHVLKKTDDIIFFHNVNVDESGRRQCDLVAGKREDLN
   KAQKQELRKLAEEYNIPNENTQHLVPTDTCNQ
   >1qwd NP_006735 retinol-binding protein 4 [Homo sapiens]
   MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
   MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
   LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
   >1qwdA Bacterial Lipocalin Blc E. Coli
   MSYYHHHHHHLESTSLYKKSSSTPPRGVTVVNNFDAKRYLGTWYEIARFDHRFERGLEKVTATYSLRDDG
   GLNVINKGYNPDRGMWQQSEGKAYFTGAPTRAALKVSFFGPFYGGYNVIALDREYRHALVCGPDRDYLWI
   LSRTPTISDEVKQEMLAVATREGFDVSKFIWVQQPGS
   >1z24A Chain A, The Molecular Structure Of Insecticyanin From The Tobacco Hornworm
Manduca Sexta L. At 2.6 A Resolution.
   GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLPLENENQGKCTIAEYKYDGKKASVYNSFVSNGVKEYM
   EGDLEIAPDAKYTKQGKYVMTFKFGQRVVNLVPWVLATDYKNYAINYNCDYHPDKKAHSIHAWILSKSKV
   LEGNTKEVVDNVLKTFSHLIDASKFISNDFSEAACQYSTTYSLTGPDRH
   >2blg  Bovine Beta-Lactoglobulin
   LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENDECAQKK
   IIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSLVCQCLVRTPEVDDEALEKFDKAL
   KALPMHIRLSFNPTQLEEQCHI
   >1pboA Bovine Odorant Binding Protein (Obp)
   AQEEEAEQNLSELSGPWRTVYIGSTNPEKIQENGPFRTYFRELVFDDEKGTVDFYFSVKRDGKWKNVHVK
   ATKQDDGTYVADYEGQNVFKIVSLSRTHLVAHNINVDKHGQKTELTGLFVKLNVEDEDLEKFWKLTEDKG
   IDKKNVVNFLENEDHPHPE
   >1e5pA Aphrodisin Female Hamster
   QDFAELQGKWYTIVIAADNLEKIEEGGPLRFYFRHIDCYKNCSEXEITFYVITNNQCSKTTVIGYLKGNG
   TYETQFEGNNIFQPLYITSDKIFFTNKNXDRAGQETNXIVVAGKGNALTPEENEILVQFAHEKKIPVENI
   LNILATDTCPE
```

Nota que algunas de ellas llevan un identificador extraño (1QWD, 1Z24). Estos son identificadores de PBD, una BD de estructuras 3D de proteínas,. Puedes encontrar estos identificadores por Entrez Structure, o en la página web de PBD.

Con estas secuencias:

- Alinéalas con T-Coffee con la versión del EBI
- Alinéalas con T-Coffee con la versión oficial (http://www.tcoffee.org/)
- En esa misma página, evalúa el alineamiento con el programa iRMSD. Automáticamente incluirá la información estructural de las proteínas que estén especificadas con identificadores de PBD. Cuidado con la notación pues a veces da errores (mayúsculas y minúsculas, etc.)
- Alinea las secuencias de nuevo con Expresso, también en la misma página, para incorporar ahora la información estructural de las lipocalinas con ids de PBD. A continuación vuelve a ejecutar el programa iRMSD. Anota a la puntuación obtenida y compárala con la anterior. ¿Mejora el alineamiento?¿En qué se diferencia de los anteriores?¿En qué medida podemos confiar en la puntuación de iRMSD sobre la bondad del alineamiento?