Evolución de familias multigénicas 2020
Depuración de datos y alineamiento

Estas notas complementan la demostración práctica hecha en la clase del **25 de mayo**.

Punto de partida:

Partimos de un archivo fasta de secuencias de aminoácidos (en nuestro caso de globinas de deuterostomados, como por ejemplo el archivo ***Deuterostome_Gbs.fasta***, con datos bajados del NCBI y de Ensembl.

Objetivos:

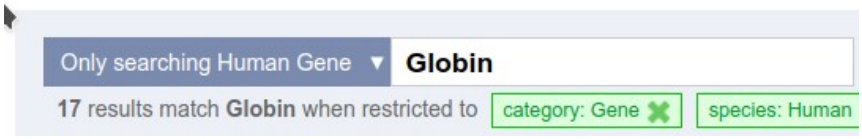
1. Obtener una base de datos reducida y más apropiada para el objetivo (en el ejemplo, reconstrucción del repertorio de globinas de deuterostomados en base a las globinas reportadas en el gusano bellota y al conocimiento de las globinas de vertebrados). Las razones para la eliminación de secuencias son las siguientes:
    a. Secuencias que no son globinas. Pueden haber estado mal anotadas en la base de datos, o podemos haberlas incorporado por error.
    b. Secuencias que no aportan información adicional relevante para la escala de nuestro trabajo. Por ejemplo, variantes menores restringidas a algunas especies, o secuencias de especies muy cercanamente emparentadas en relación a la escala del trabajo.
2. Mejorar el etiquetado de las secuencias:
    a. Retener la ID de la secuencia en la base de datos de la cual fue tomada, pero usar de manera consistente etiquetas, tanto para los genes (“Mb”, por ejemplo) como para los taxones (e.g., “chicken” o “*Gallus gallus*”, pero siempre la misma etiqueta.

Procedimientos:

Partimos de la base que hemos ya identificado una familia multigénica de interés, y que tenemos una noción bastante razonable del grupo de organismos que vamos a estudiar. En este caso, las globinas de deuterostomados. Lo primero es hacer una búsqueda bibliográfica para identificar los globinas presentes el grupo de interés. Como los vertebrados son los más estudiados, podemos empezar allí y buscar todas estas proteínas: *α-* y *β- hemoglobinas*, *Citoglobinas*, *Globinas-E*, *-X* e *-Y*, *Mioglobinas* y *Neuroglobinas*. Aquí conviene pensar que nuestros criterios de búsqueda tienen que ser lo suficientemente laxas como para identificar homologías distantes.

1. Identificar las proteínas a utilizar para sembrar búsquedas. Un buen punto de partida es ir a Ensembl, seleccionar el genoma humano, y buscar todos los genes que tengan el término "Globin".

Esta búsqueda identifica 17 genes, que es claramente demasiado. Entonces,



tenemos que curar esta lista. Inspeccionando con atención, vemos que están la gran mayoría de los genes de interés, pero también tenemos elementos que claramente no pertenecen. Los dos primeros elementos en esa lista regulan a las globinas, pero no pertenecen a esta familia.

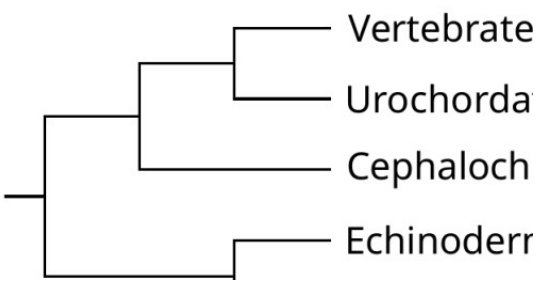
De manera que esta búsqueda sencilla ya nos lleva a una lista bastante razonable de las globinas en humanos, incluyendo la mayoría de las *α-* y *β-hemoglobinas*, *Citoglobina* y *Neuroglobina*, aunque tiene omisiones sorprendentes, como la de la *Mioglobina*. De manera que esta lista es un buen punto de partida.

Aquí utilizamos nuestros conocimientos previos para restringir la lista a una sola de las *α-hemoglobinas*, una sola de las *β-hemoglobinas*, la *Citoglobina*, *Mioglobina* y *Neuroglobina*. Ya sabemos que los humanos no tenemos *Globinas-E*, *-X* e *-Y*, por lo que debemos ir a buscar estos genes en otras especies. Otra vez, gracias a nuestro conocimiento previo sabemos que las tortugas y el celacanto son los 2 grupos que incluyen a la gran mayoría de las globinas de vertebrados.

El archivo ***Globin_search_seeds.fasta*** tiene estas 8 secuencias:
Human_HbA
Human_HbB
Human_Cygb
Latimeria_GbE
Latimeria_GbX1
Latimeria_GbY
Human_Mb
Human_Ngb

De manera que ahora estamos prontos para comenzar a completar nuestra base de datos.

2. El siguiente paso es identificar los genomas a ser muestreados. Aquí buscamos un equilibrio entre muestreo taxonómico denso y selección de genomas de buena calidad. Conviene tener en cuenta que denso se refiere a que haya réplicas, y no a incluir todos los datos posibles. Si nos concentramos en los deuterostomados, estos incluyen 3 filos diferentes: Hemichordata, Echinoderma y Chordata, y este último está dividido en los subfila Urochordata, Cephalochordata y Vertebrata.



Entonces, ahora vamos a buscar las globinas en los genomas de especies representativas. Esto no es tarea sencilla, ya que hay muchos genomas disponibles. Por ejemplo, hay 27 genomas de equinodermos a la fecha (junio 2020), https://www.ncbi.nlm.nih.gov/genome/?term=Echinoderms. Pero vamos a restringirnos a genomas ya curados. En esta caso, vamos a buscar las globinas de las siguientes especies:

Especie	Nombre común	Clasificación
Homo sapiens	Humano	Vertebrado
Pelodiscus sinensis	Tortuga china de caparazón blando	Vertebrado
Xenopus tropicalis	no tiene, es una rana; notar que es un xenopus diploide	Vertebrado
Latimeria chalumnae	celacanto de Comores	Vertebrado
Lepisosteus oculatus	catán pinto;	Vertebrado
Callorhinchus milii	tiburón elefante	Vertebrado
Petromyzon marinus	lamprea marina	Vertebrado
Ciona savignyi	ascidio (tunicado) transparente	Urocordado
Ciona intestinalis	ascidio jarrón	Urocordado
Branchiostoma lanceolatum	pez lanceta	Cefalochordado
Strongylocentrotus purpuratus	erizo de mar rojo	Equiinodermo
Saccoglossus kowalevskii	gusano bellota	Hemicordado

Para eso, utilizamos el algoritmo BLAST y las secuencias ***Globin_search_seeds.fasta***. Con eso empezamos a seleccionar las globinas en los genomas correspondientes, recordando que tenemos que el “query cover” tiene que ser alto, y la identidad es, generalmente mayor a 30%. Verán que se repiten las secuencias, esto es normal, y que muchas veces hay varias isoformas para una misma proteína. Las isoformas son variantes derivados del splicing alternativo, y no las consideramos aqui. Tomamos una proteína por gen, y no hay mayor importancia en cual isoforma escogemos.

Las búsquedas se pueden hacer en Ensembl también.
Aquí conviene hacer una spreadsheet donde guardamos la información de las secuencias con las siguientes columnas: Especie, Nombre común, Clasificación, nombre del gen, número de acceso, largo de la secuencia, localización, “genome

release" y la fuente. Por otro lado, hacemos un archivo multifasta donde guardamos las secuencias. Todo esto se puede guardar en un solo lugar usando paquetes bioinformáticos mas sofisticados como BioPython, pero con esto es suficiente generalmente. En el archivo multifasta, conviene cambiar los nombres para facilitar la visualización de los árboles filogenéticos.

Aquí tienen un ejemplo:

Este es el archivo original

```
>XP_005993267.1 PREDICTED: cytoglobin isoform X1 [Latimeria chalumnae]
MSSEKFLQSFGEDLLLMEKVQGEMEMDRWERSDQLSDTEVESIRQIWSNVYTNCENVGVLVLIRFFVNFP
SAKQYFSQFRHLEDPLDMERSVQLRKHARRVMGAINTVVENVEDQDKIASVLAPVGKAHALKHKVEPVYF
KILSGVILEILAEEYAQHFTPEVQKAWTKLMSIICCHVTATYKEVGWGQLSNSTM
```

Y esta es la misma secuencia con un nombre más informativo:

```
>Celacanto_Cygb_XP_005993267
MSSEKFLQSFGEDLLLMEKVQGEMEMDRWERSDQLSDTEVESIRQIWSNVYTNCENVGVLVLIRFFVNFP
SAKQYFSQFRHLEDPLDMERSVQLRKHARRVMGAINTVVENVEDQDKIASVLAPVGKAHALKHKVEPVYF
KILSGVILEILAEEYAQHFTPEVQKAWTKLMSIICCHVTATYKEVGWGQLSNSTM
```

Al archivo este hay que agregarle dos secuencias de Leghemoglobina de plantas que van a servir de grupos externos.

Están aquí:

```
>Alfalfa_LegHb
MSFTDKQEALVNSSWEAFKQNLPRYSVFFYTVVLEKAPAAKGLFSFLKNSAEVQDSPQLQAHAEKVFGLVRDSAVQL
RATGGVVLGDATLGAIHVRKGVVDPHFVVVKEALLKTIKEAAGDKWSEELNTAWEVAYDALATAIKKAMS

>Yellow_lupin_LegHb
MGALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPELQAHAGKVFKLVYEAAI
QLQVTGVVVTDATLKNLGSVHVSKGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA
```

Ahora estamos prontos para alinear las secuencias. Lo más sencillo es:

1. cargar el archivo fasta en MEGA X (elegir la opción de alinear)
2. alinear con Muscle,
3. salvar el alineamiento,
4. cargar el alineamiento en MEGA X (elegir la opción de analizar),
5. hacer un árbol de NJ (no importa el modelo de sustitución, sin bootstrap),
6. enraizar el árbol con los grupos externos,
7. explorar el árbol buscando ramas muy largas y secuencias que caigan en posiciones extrañas,
8. eliminar las secuencias sospechosas, y anotar por qué,

9. salvar este archivo con un nuevo nombre, y repetir el procedimiento hasta que no haya secuencias extrañas, generalmente es suficiente con una iteración,
10. Alinear el archivo de secuencias curadas usando Muscle, las estrategias L-INS-i, E-INS-i y G-INS-i de Mafft, Kalign,
11. asegurarse de que los archivos resultantes tienen a las secuencias en el mismo orden,
12. comparar los alineamientos usando MUMSA,
13. elegir el que tiene el mejor score, y utilizar este para los análisis filogenéticos,
14. cargar el alineamiento en MEGA X (elegir la opción de analizar),
15. estimar el modelo de sustitución mas adecuado,
16. estimar la filogenia usando máxima verosimilitud bajo el modelo seleccionado, y evaluar apoyo para los nodos usando 100 réplicas de bootstrap,
17. exportar el árbol en formato newick,
18. abrir el árbol en Dendroscope, editar a gusto, los pasos 19 a 21 son opcionales
19. exportar como svg,
20. abrir en Inkscape para dar los toques finales (esto da gráficos de calidad de publicación),
21. reconciliar el árbol de especies con la filogenia de genes obtenida en el paso 16.
22. Escribir el reporte.
23. Hacer esto con otra familia multigénica para el trabajo final del curso.