ALINEAMIENTOS Y FILOGENIA
1ª PRÁCTICA OBLIGATORIA

1.- Objetivo global

El objetivo final es encontrar un gen "nuevo" (una secuencia de ADN que se encuentra en una base de datos y no está anotada) y caracterizarlo todo lo que podamos. En particular debemos

- Realizar una justificación razonada de la "novedad" del gen mediante el uso de BLAST.
- Analizar su familia: identificación de homólogos, alineamiento de múltiples secuencias.
- Analizar su filogenia: construcción del árbol filogenético de la familia.

Un par de ejemplos o motivaciones de este ejercicio:

- Quieres estudiar una globina que nadie ha caracterizado antes, quizás en un organismo de tu interés tal como una planta.
- Estás interesado en las lipocalinas, y has visto en un artículo que hay una que se encuentra en las lágrimas de los hamsters. ¿Podría haber un gen no descubierto que codifica una lipocalina que se encuentre en las lágrimas humanas?

Esta práctica es básicamente la integración de los distintos ejercicios propuestos en clase sobre la búsqueda de un gen "nuevo" (sección "Descubrimiento de Genes" del tema 4 y secciones de ejercicios dentro de las presentaciones de los temas 5 y 6). Dichos ejercicios se pueden integrar o ampliar para formar el cuerpo de esta práctica.

2.- Selección del gen/proteína

Primero, seleccionad un gen que os resulte interesante en base a algún estudio (más o menos científico) que hayáis leído o a algún tema biológico o médico. Luego seleccionad su producto genético o proteína (tened en cuenta que un gen puede tener varias isoformas y por tanto, varios productos distintos).

Por ejemplo, podéis buscar genes relacionados con enfermedades o su cura/tolerancia, relacionados con alergias, con fenotipos físicos (color de ojos, pelo, sexo), etc. Cualquiera es válido siempre y cuando el tema os interese. Pero no te encariñes demasiado con tu gen, igual no encuentras nada novedoso y debes buscar en otro.

Veréis que aparecen muchísimos genes/productos genéticos relacionados con vuestra búsqueda. Algunos criterios para filtrarlos:

- Limitad la búsqueda inicial al máximo, mediante el uso de distintos filtros (especialmente nombre de gen y organismo). Por ejemplo, si buscamos "BRCA1" en Entrez Protein nos devuelve decenas de miles de resultados. Si buscamos "BRCA1[Gene]" sólo unas mil, y si buscamos "BRCA1[Gene] AND Homo sapiens[Organism] se reduce a unas cien.
- Utilizad bases de datos no redundantes, como UniGene, RefSeq, etc.
- Si dudáis, realizad una búsqueda en PubMed en estudios sobre el tema para buscar el identificador exacto de la proteína. Si es relativamente popular, seguramente Wikipedia también os pueda ayudar.
- Elegid una proteína con un funcionamiento comprensible (por ejemplo, "inducir la muerte celular") en vez de una muy complicada que quizás no seamos capaces de abarcar desde un punto biológico, por falta de conocimientos especializados.

3.- Búsqueda de genes nuevos a partir de nuestra elección

Comenzando con la secuencia de la proteína elegida, una búsqueda *tblastn* sobre una base de datos de secuencias genómicas o (mejor) de secuencias expresadas (EST) nos va a dar secuencias similares. Vuestro cometido es determinar la base de datos, parámetros de búsqueda, organismos seleccionados y, sobre todo, inspeccionar los resultados en busca de secuencias similares relevantes. No dudéis en repetir la búsqueda *blast* tantas veces como sea necesario para encontrar algo relevante, o en cambiar de proteína de inicio si no encontráis nada claro.

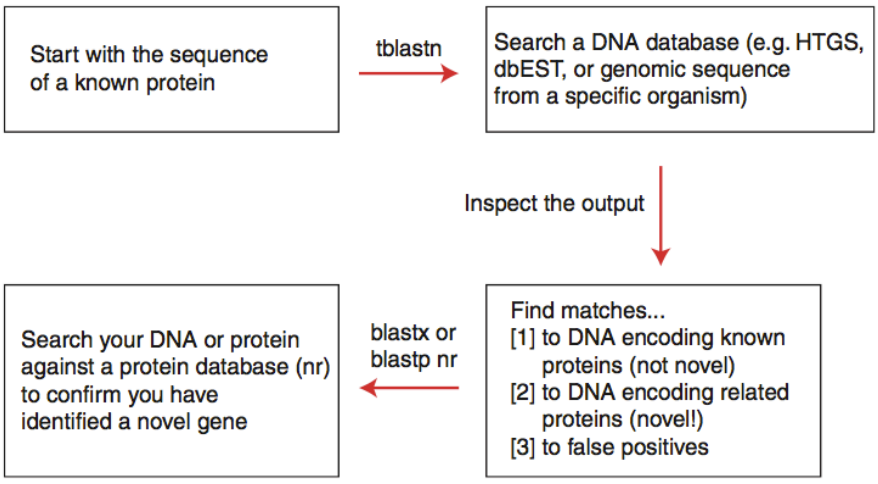
En general, tened en cuenta los valores de coincidencia y E-valor, además de los alineamientos:

- Descartad coincidencias con E-valores muy altos (>0.6)
- Si la coincidencia es exacta (100%), será una proteína ya descubierta
- Si la coincidencia es alta (>20-30%) pero no exacta, probablemente sea una proteína no descubierta aún.
- Si la coincidencia es muy baja, probablemente no sea un homólogo si no un falso positivo (estamos en la "dimensión desconocida").

Una vez seleccionado el gen "nuevo" candidato, caracterízalo: ponle un nombre y recoge el identificador de la secuencia y el nombre de la especie.

A continuación realiza una búsqueda *blastx* o *blastp* en la base de datos *nr* del NCBI, con la intención de buscar posibles proteínas ya identificadas con este gen:

- Si hay una coincidencia del 100% con alguna proteína de la BD, de la misma especie que el gen "nuevo", es que el gen no es nuevo (ya está en la BD de proteínas, aunque esté con nombre "unknown")
- Si la mejor coincidencia es <100%, es probable que sea una proteína nueva y hayamos tenido éxito.
- Si la coincidencia es del 100%, pero en una especie distinta, hemos encontrado un gen nuevo.
- Ojo: si en la búsqueda *blast* no está la proteína original seleccionada en el apartado 2, habremos encontrado un ADN/proteína que no es homólogo con la consulta original. Debemos empezar de nuevo, seguramente habremos hecho algo mal.



Esquema de proceso, tomado de Pevsner 2009 (p 170)

4.- Análisis de su familia genética

Ahora, partiendo de la proteína inicial, la proteína "nueva" y otras proteínas que seleccionemos de su familia (por ejemplo, si la proteína nueva y la inicial son de distintas especies, sus equivalentes de otras especies; u otras proteínas de la misma familia que la proteína inicial). Escoge como mínimo entre 5-10 y como máximo 30.

Con esta familia, podemos utilizar las distintas aproximaciones de alineamiento múltiple para compararlas. A continuación, utilizaremos el alineamiento múltiple para construir su árbol filogenético.

Discute las relaciones entre los miembros de la familia y la fiabilidad del alineamiento y la filogenia.

5.- Conclusiones

Describe lo que has aprendido sobre este gen/proteína, y sobre su familia. Describe también tus impresiones sobre el proceso de análisis, los puntos más complicados, errores que hayas podido cometer, utilidad biológica, etc.

Evaluación

Se tendrán en cuenta los siguientes aspectos

- *Comprensión*: de los métodos utilizados y de la función (en términos generales) del gen/proteína inicial seleccionado. Explicación de por qué se han elegido dichos métodos (y sus parámetros).
- *Análisis crítico*: por qué se ha seleccionado un determinado gen/proteína como nuevo. Valoración de la calidad de las búsquedas, alineamientos y filogenias realizadas.
- *Concreción*: capacidad de síntesis en cuanto a la explicación del problema y su análisis, no repetir lo dicho por otros.
- *Claridad*: claridad en la redacción, corrección ortográfica, organización en apartados. La claridad del informe se considera un indicativo del grado de comprensión.
- *Compleción*: variedad de tipos de análisis y configuraciones de parámetros probados, número de herramientas utilizadas, número de especies/genes/productos génicos relevantes que han sido explorados o al menos contemplados.
- *Ética*: rigor en las citas (publicaciones o páginas web), evitad copiar textos tal cual (entenderlos, sintetizarlos y explicarlos si es necesario, si no lo es simplemente citarlos). La copia, total o en gran medida, del trabajo, es motivo suficiente para suspender de manera automática la práctica. Sé un buen profesional, no copies.

Entrega

Entregad un informe redactando los principales hallazgos, dificultades, etc. que habéis encontrado en cada paso.

Especialmente, documentad:

- Elección de Genes/proteínas: identificadores, organismo, secuencia, descripción, razones para su elección
- Búsqueda de ADN codificante "nuevo": descripción de la búsqueda (base de datos, blast, parámetros), razones para la elección de dicha búsqueda, calidad de los filtros utilizados/resultados obtenidos (E-valores, falsos positivos, etc.)
- Confirmación de la novedad: discusión de la búsqueda inversa en blast
- Caracterización del gen "nuevo": identificadores, secuencia, especie, descripción.
- Alineamiento y filogenia: búsqueda de secuencias homólogas en su especie o inter-especies, estudio de alineamientos múltiples de las secuencias, discusión de resultados. Construcción de árboles filogenéticos a partir del alineamiento, discusión de métodos y resultados.

La entrega se hará a través de moodle, el nombre del fichero debe tener OBLIGATORIAMENTE el siguiente formato (sin tildes):

`NombreApellido1Apellido2.pdf`

Si se desean incluir más ficheros distintos a los dos anteriores (con nombres arbitrarios), hacedlo en un fichero comprimido `NombreApellido1Apellido2.zip`

Los ficheros con nombres distintos a éstos (fuera del .zip) serán ignorados