
# MANUAL DE PROBLEMAS, PRÁCTICAS DE LABORATORIO Y SIMULACIÓN DE GENÉTICA II

GRADO EN BIOLOGÍA

# MANUAL DE PROBLEMAS, PRÁCTICAS DE LABORATORIO Y SIMULACIÓN DE GENÉTICA II

Este manual para la asignatura Genética II es parte de los manuales de "Problemas y Casos Prácticos de Genética" (ISBN: 978-84-15261-50-6 ) y "Manual de Prácticas de Genética" (ISBN: 978-84-15261-49-0) elaborados por profesores del Departamento de Genética de la Universidad de Granada, en el marco de un Proyecto de Innovación Docente titulado "Nuevos recursos docentes para las prácticas del Departamento de Genética en el marco del EEES" (curso académico 2010/2011) financiado por el Vicerrectorado de Garantía de la Calidad de la Universidad de Granada.

En el curso académico 2017/2018, este manual ha sido revisado en el marco de los Proyectos de Innovación Docente titulados "Actualización de material didáctico para la docencia práctica de la asignatura *Genética II: de la secuencia a la función* del Grado en Biología" y "Desarrollo de medios audiovisuales y virtualización de contenidos en asignaturas del área de Genética" concedidos por el Vicerrectorado de Garantía de la Calidad de la Universidad de Granada para el periodo 2016/2018.

# ÍNDICE

# PROBLEMAS

# GENÉTICA MOLECULAR

## 1. GUÍA DE RESOLUCIÓN DE PROBLEMAS

### *Mapas de restricción*

Un mapa de restricción representa una secuencia lineal de los sitios en los que diferentes enzimas de restricción poseen dianas en una molécula de ADN particular. Consiste en la ordenación de una serie de dianas para enzimas de restricción en una molécula de ADN concreta. En el mapa se representan las distancias entre dichas dianas, distancias que se miden en pares de bases (o en kilobases).

Cuando una molécula de ADN es cortada con una enzima de restricción y los fragmentos generados se separan por electroforesis en un gel de agarosa, se puede determinar el número de sitios de restricción y la distancia entre ellos a partir del número y la posición de las bandas en el gel. Cabe distinguir entre moléculas de ADN lineal y ADN circular:

**ADN lineal**: hay que tener en cuenta que el número de fragmentos que se generan tras una digestión, es el número de dianas presentes en su secuencia para esa enzima más uno. La suma del tamaño de los fragmentos debe de coincidir con el tamaño total del ADN digerido. Pero hay que tener en cuenta que el número de fragmentos no es siempre coincidente con el número de bandas que aparecen en un gel de agarosa, ya que puede haber fragmentos de igual tamaño que migran juntos.

**ADN circular**: el número de fragmentos que se generan tras una digestión, es el mismo que el número de dianas presentes en su secuencia para esa enzima. Cuando una enzima corta una vez sólo, nos revela el tamaño del ADN circular. La suma del tamaño de los fragmentos debe de coincidir con el tamaño total del ADN digerido. Como antes, hay que tener en cuenta que el número de fragmentos no es siempre coincidente con el número de bandas que aparecen en un gel de agarosa, ya que puede haber fragmentos de igual tamaño que migran juntos.

En cualquier caso, la información obtenida mediante electroforesis no nos revela el orden ni la localización de las dianas de restricción. Para poder realizar un mapa, se ha de cortar una muestra del ADN a mapear con una enzima de restricción, una segunda muestra del mismo ADN con otra enzima diferente y una tercera muestra de dicho ADN con las dos enzimas simultáneamente (digestión doble). Esta tercera digestión nos da la clave para determinar el orden de las dianas para ambas enzimas de restricción.

### *Marcadores moleculares*

Es importante asignar los genotipos a los individuos del pedigrí para intentar ver la coincidencia entre sus alelos y los patrones de bandas.

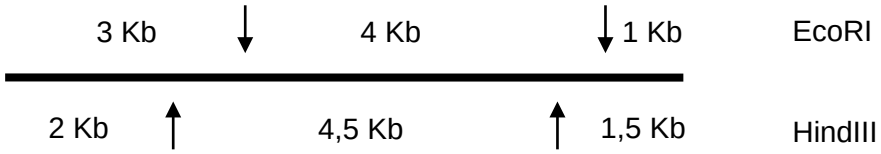
Hay que tener en cuenta la distancia entre dianas, que nos dará el tamaño de las bandas observables pero, además, hay que prestar especial atención a la región con la que hibrida la sonda, pues aquellos fragmentos con los que no hibride, no podrán ser detectados tras el revelado.

En el caso de los microsatélites, los diferentes tamaños amplificados para un locus, pueden considerarse como alelos. Los microsatélites presentan herencia mendeliana simple y son

codominantes. Para un locus microsatélite, cada uno de los alelos presentes en el genotipo de un individuo (tamaño de amplificado) procede uno del padre y otro de la madre.

## 2. PROBLEMAS RESUELTOS

**Problema 1.** Un gen clonado muestra el siguiente mapa de restricción para las enzimas EcoRI y HindIII (↓ sitio de corte):



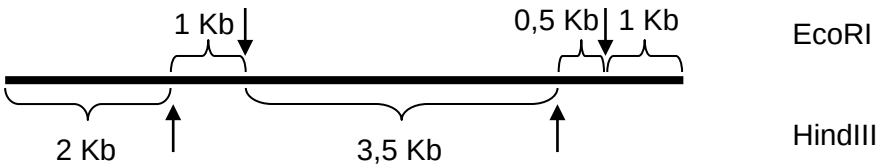

a) Dibujar los patrones de los fragmentos de ADN esperados con cada enzima al separar los fragmentos mediante electroforesis en gel de agarosa. Hacer lo mismo para el caso de la digestión doble.

b) Dibujar el patrón esperado para una copia mutante del gen que ha perdido el primero de los cortes de EcoRI

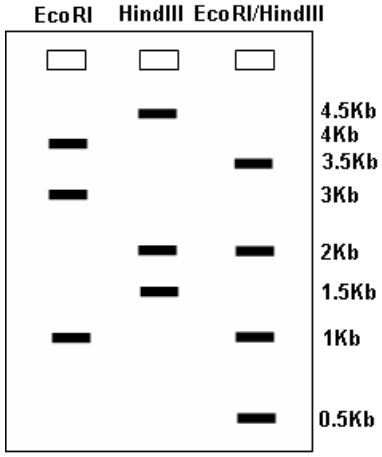
c) Dibujar el patrón esperado para una copia mutante del gen en la que ha aparecido una nueva diana para HindIII en el centro del fragmento de 2Kb.

*Respuesta*

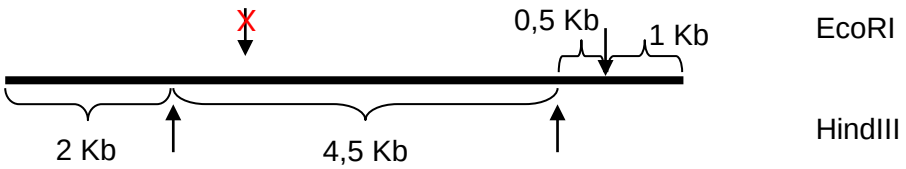
a) Para EcoRI el gen tiene dos dianas, por lo que será cortado en tres fragmentos de tamaños 4Kb+3Kb+1Kb. Para HindIII también tiene dos dianas, pero en diferentes posiciones, por lo que generará tres fragmentos pero de tamaños 4.5Kb+2Kb+1.5Kb. Cuando utilizamos las dos enzimas para digerir el gen, obtendremos 5 fragmentos diferentes (existen 4 puntos de corte, generando fragmentos de diana a diana de ambas enzimas), aunque dos de ellos presentan el mismo tamaño (1Kb), por lo que los observaremos como una única banda en el gel de agarosa. Los tamaños serán, por tanto, de 3.5Kb+2Kb+1(x2)Kb+0.5Kb (ver figura).

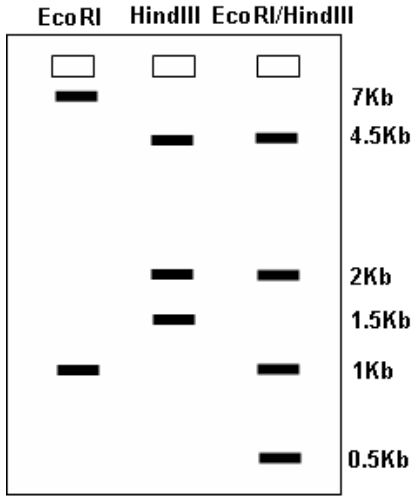


Así, en un gel de agarosa, observaremos los siguientes patrones de bandas:

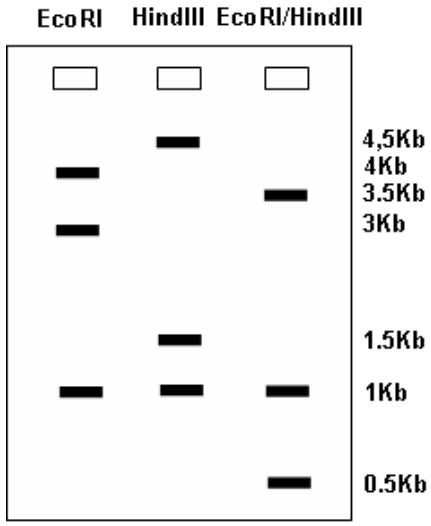



b) Si la copia mutante del gen pierde una diana para EcoRI, al cortar con esta enzima, obtendremos solo dos fragmentos, siendo uno de ellos, la suma de los dos entre los cuales se encontraba la diana perdida para EcoRI, 7Kb+1Kb. Para el corte con HindIII el patrón de bandas no se vería afectado, pero sí nuevamente para la digestión doble, ya que hay un corte menos, 4,5Kb+2Kb+1Kb+0,5Kb (ver figura).



c) En este caso, cuando cortamos el gen con HindIII, al tener una diana más (tres puntos de corte) obtendríamos un fragmento más. Sin embargo, en el gel, no aparecerían 4 bandas, ya que se han generado dos fragmentos de igual tamaño (1Kb), por lo que correrán de igual forma. Los fragmentos para HindIII serían 4,5Kb+1,5Kb+1Kb(x2). Para EcoRI el patrón no se ve afectado y para la digestión doble los fragmentos generados serían 3,5Kb+1Kb(x4)+0,5Kb (ver figura).


**Problema 2.** Se ha cortado con PstI un plásmido bacteriano circular que contiene un gen de resistencia a la ampicilina. Tras la electroforesis se observa una banda de 20 Kb. ¿Qué deducirías de los resultados que se plantean a continuación?

a) Con EcoRI, el plásmido se corta en dos fragmentos: uno de 11.5Kb y otro de 8.5Kb

b) La digestión PstI+EcoRI genera tres fragmentos de: 6Kb, 5.5Kb y 8.5Kb

c) El ADN del plásmido cortado con PstI se ha mezclado y ligado con fragmentos de ADN cortados con PstI. Todos los clones recombinantes son resistentes a la ampicilina.

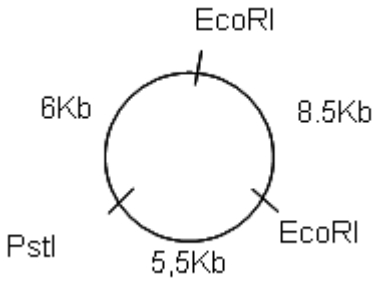
d) Tras cortar uno de los clones recombinantes con PstI se obtienen dos fragmentos: 20 Kb y 6 Kb.

e) El clon anterior se corta con EcoRI y se obtienen10 Kb, 8.5 Kb y 7.5 Kb.

*Respuesta*

a) Al sumar los fragmentos 11.5 Kb + 8.5 Kb nos da un valor de 20 Kb. Este valor es coincidente con el fragmento generado con PstI, lo que significa que el plásmido tiene un tamaño de 20 Kb y que, por tanto, PstI lo corta una sola vez mientras que EcoRI tiene dos dianas dentro de la molécula circular.

b) Con la digestión doble podemos, ahora, obtener un mapa de restricción de esta molécula circular. Podemos deducir que el fragmento de 11.5 Kb generado por EcoRI, es cortado en dos fragmentos menores de 6Kb y 5.5Kb por la enzima PstI:

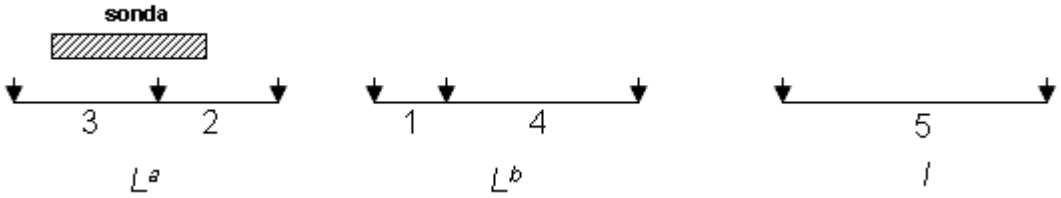



c) Al cortar con PstI, el plásmido se queda en forma lineal con extremos cohesivos para ese enzima. Al poner en contacto fragmentos de ADN cortados también con PstI junto con una enzima ligasa, los fragmentos, que tienen extremos complementarios, se ligan al plásmido y la molécula recircularia con un inserto dentro de ella, obteniendo un plásmido recombinante. Si la diana para PstI estuviera dentro del gen de la ampicilina, el inserto "rompería" este gen, por lo que quedaría inactivo y las bacterias serían sensibles a la ampicilina. Por eso, podemos deducir que la diana para PstI no se encuentra dentro del gen de resistencia a la ampicilina.

d) Al cortar de nuevo con la enzima PstI, lo que estamos haciendo es separar nuevamente el plásmido del inserto, por lo que obtenemos un fragmento de 20 Kb correspondiente al plásmido y otro de 6 Kb que sería el tamaño del fragmento clonado.

e) Al aparecer un nuevo fragmento cuando cortamos con EcoRI el plásmido recombinante, que no aparecía en el plásmido bacteriano inicial, significa que existe una nueva diana para este enzima. La diferencia entre el plásmido bacteriano inicial y el plásmido recombinante, es la presencia del inserto de 6 Kb. Por eso, deducimos que el fragmento clonado tiene una diana para EcoRI y que se encuentra situada entre las dos dianas EcoRI separadas por 11,5Kb.

**Problema 3.** Se conoce un gen autosómico con tres alelos $L^a$, $L^b$ y *l* que se diferencian en una diana para la enzima de restricción PstI (↓ sitio de corte):




Diseñar un experimento para diferenciar los genotipos de los diferentes individuos que pudieren existir en una población si utilizamos como sonda el fragmento homólogo de ADN señalado en el esquema.

*Respuesta*

La diferencia entre los distintos alelos del gen corresponde a diferencias en secuencia nucleotídica. En este caso, estas diferencias de nucleótidos afectan a secuencias dianas para la enzima PstI. Esta información la vamos a utilizar para realizar un experimento que detecte un marcador RFLP (polimorfismo en la longitud de los fragmentos de restricción).

Para ello, debemos seguir los siguientes pasos:

a) Digerir todo el ADN genómico con la enzima PstI, ya que inicialmente, no podemos aislar nuestro gen del resto del genoma.

b) Para separar los fragmentos generados, según su tamaño, debemos ahora realizar una electroforesis en gel de agarosa.

c) Los fragmentos de ADN, tal como se encuentran ordenados en el gel de agarosa, deben de transferirse a una membrana de nylon mediante la técnica de Southern-blot.

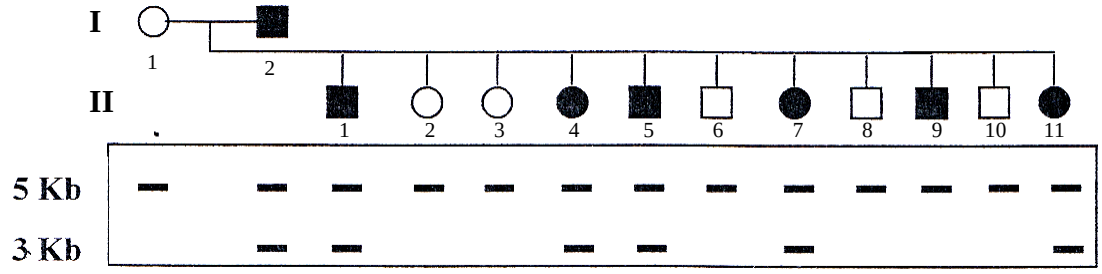
d) Mediante una hibridación tipo Southern-blot, utilizando la sonda señalada en el esquema, podremos localizar específicamente la región correspondiente al gen estudiado, ya que es complementaria a esta región.

e) El revelado de la hibridación pondrá de manifiesto los fragmentos de ADN genómico con los que la sonda ha hibridado

Así, tenemos diferentes genotipos posibles, que coincidirán con patrones de bandas:

$L^aL^a$: bandas de 3Kb y 2Kb
$L^aL^b$: bandas de 3Kb y 2Kb para el primer alelo y de 1Kb y 4Kb para el segundo alelo
$L^al$: bandas de 3Kb y 2Kb para el primer alelo y de 5Kb para el segundo
$L^bL^b$: bandas de 1Kb y 4Kb
$L^bl$: bandas de 1Kb y 4Kb para el primer alelo y de 5Kb para el segundo
*ll*: banda de 5Kb

**Problema 4.** El siguiente pedigrí representa a una familia con alguno de sus miembros afectado por una enfermedad autosómica dominante. El ADN de todos los individuos fue digerido con la enzima PstI y sometido a electroforesis en gel de agarosa. Se analiza este ADN mediante hibridación tipo Southern con una sonda que corresponde a un fragmento de ADN humano clonado en un plásmido bacteriano. Los resultados del revelado de la hibridación se muestran junto al pedigrí.



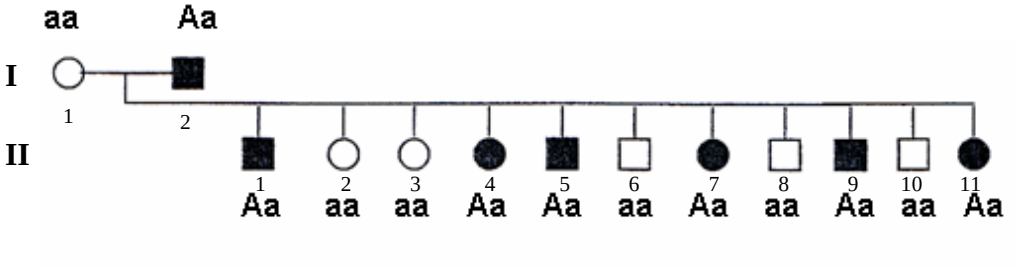

a) Explica el protocolo seguido y los resultados obtenidos en los distintos individuos

b) ¿Podemos usar la sonda con fines diagnósticos para esta enfermedad?

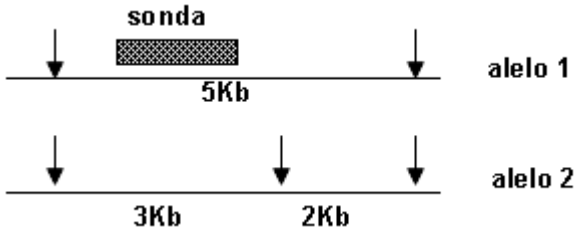
*Respuesta*

a) El marcador molecular utilizado en este análisis corresponde con un RFLP. El ADN genómico se ha digerido con la enzima PstI y los fragmentos generados se han separado mediante electroforesis en gel de agarosa. A continuación, se realiza una transferencia de esos fragmentos (tal y como han migrado en el gel) a una membrana de nylon, a la cual, se fijan. Sobre esta membrana se realiza una hibridación (hibridación tipo Southern) con un fragmento de ADN marcado (sonda). Al revelar esta hibridación, nos aparecen bandas de diferentes pesos moleculares, indicando tamaños de fragmentos de ADN genómico que son homólogos a la sonda.

El primer paso consiste en asignar los genotipos a los individuos del pedigrí, teniendo en cuenta que la enfermedad es autosómica dominante:



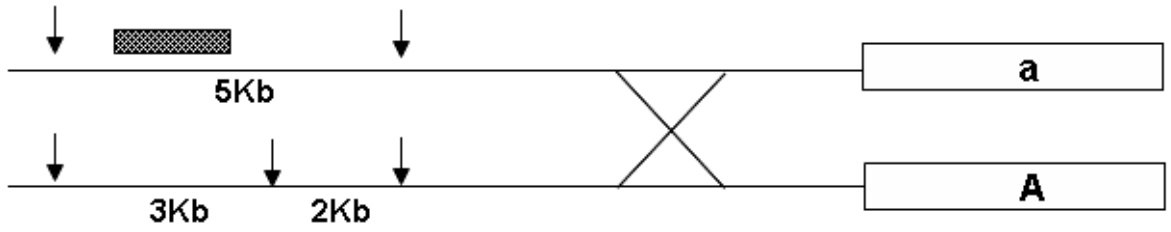

Al comparar los resultados del marcador molecular con los fenotipos (afectados/no afectados por la enfermedad) podemos ver cómo existe coincidencia entre el número y los tamaños de bandas del RFLP y el desarrollo o no de la enfermedad. Así, a excepción de un individuo (II-9), todos los afectados presentan dos bandas de 5Kb y 3Kb y los no afectados una única banda de 5Kb. Teniendo en cuenta que, en especies diploides, existen parejas de regiones homólogas (cromosomas homólogos), debemos de "identificar" dos alelos. Así, la diferencia en secuencia entre estos alelos podría ser detectada si afectara a una diana para la enzima PstI, tal como se muestra en el esquema:




Si la sonda hibrida en la región indicada, y teniendo en cuenta que la banda de 3Kb es exclusiva para los afectados, podemos deducir, que el *alelo 1* del esquema anterior corresponde al alelo *a* del pedigrí, mientras que el *alelo 2*, corresponde con el alelo *A*, que es el causante de la enfermedad. Entonces, los individuos heterocigotos (Aa) presentan dos bandas, 5Kb del alelo *a* y 3Kb del alelo *A* (ya que el fragmento de 2Kb de este alelo *A* no es detectado por la sonda). Los individuos homocigotos sanos (aa) presentan una única banda de 5Kb. Los hipotéticos individuos homocigotos (AA) presentarían una única banda de 3Kb.

b) Al establecer la relación entre los genotipos y el patrón de bandas, observamos que el individuo II-9 no presenta esta correspondencia. Esto se podría explicar porque las diferencias observables en el patrón de bandas no son debidas a cambios en la secuencia del propio gen causante de la enfermedad, sino a regiones cercanas a él. Es decir, el RFLP que detectamos no

se corresponde con diferencias en la secuencia de los alelos del gen, sino que se encuentra en regiones ligadas al mismo, tal como muestra el siguiente esquema:




Este esquema ilustra el caso del padre (individuo I-2; genotipo Aa) que se encuentra afectado. Si durante la formación de los gametos de este individuo existiera un entrecruzamiento entre el RFLP y el gen causante de la enfermedad (como se indica en la figura), se generaría un gameto con genotipo *A* pero con marcador RFLP de 5Kb. Esto es lo que le ocurre al individuo II-9, que tiene una alelo *a* (5Kb) de heredado de la madre y un alelo *A* (recombinante de 5Kb) heredado del padre.

Por tanto, la sonda se podría usar como diagnóstico, pero debemos tener en cuenta que existe un porcentaje de error debido a la posibilidad de recombinación entre el marcador RFLP y el gen causante de la enfermedad.

**Problema 5.** En un análisis con 4 marcadores de microsatélites se obtuvieron los siguientes resultados para 5 individuos (los números indican tamaños de fragmentos amplificados en pb):

| | Individuo 1 | | Individuo 2 | | Individuo 3 | | Individuo 4 | | Individuo 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Alelo 1 | Alelo 2 | Alelo 1 | Alelo 2 | Alelo 1 | Alelo 2 | Alelo 1 | Alelo 2 | Alelo 1 | Alelo 2 |
| Locus 1 | 130 | 134 | 134 | 134 | 136 | 138 | 128 | 134 | 128 | 136 |
| Locus 2 | 250 | 256 | 256 | 260 | 258 | 260 | 252 | 260 | 250 | 258 |
| Locus 3 | 140 | 140 | 140 | 144 | 146 | 148 | 138 | 144 | 140 | 150 |
| Locus 4 | 187 | 193 | 185 | 187 | 183 | 189 | 185 | 191 | 181 | 189 |

a) ¿Qué diferencia existe entre los diferentes alelos de un mismo locus de microsatélite?

b) ¿Cuántos alelos tienen los diferentes loci de microsatélites analizados en este estudio?

c) ¿Se puede saber el número de repeticiones para cada uno de ellos?¿Y el motivo de repetición?

d) Si el individuo 1 es la madre del individuo 2, ¿cuáles de los otros tres individuos pueden descartarse como posibles padres?

*Respuesta*

a) Entre los diferentes alelos de un mismo microsatélite las diferencias existentes corresponden a un número variable de repeticiones de un motivo (generalmente dinucleótido, trinucleótido o tetranucleótido).

b) Con esta muestra no podemos saber el número de alelos totales existentes en la población, ya que pueden existir más alelos que no están representados en estos individuos. En los individuos analizados tenemos en el locus 1, 5 alelos; en el locus 2, 5 alelos; en el locus 3, 6 alelos y en el locus 4, 7 alelos.

c) A la hora de amplificar las repeticiones de los microsatélites se utilizan las regiones flanqueantes para diseñar los primers. La distancia en pares de bases entre el motivo repetido y las regiones donde se diseñan los primers son variables para cada microsatélite, por lo que el fragmento amplificado incluye el motivo repetido y parte de las regiones flanqueantes cuyo tamaño, en este caso, no conocemos. Por eso, no podemos saber el número de repeticiones en cada alelo. Tampoco sabemos el motivo de repetición, pues no tenemos información de la secuencia. Lo que sí sabemos, es que en los cuatro microsatélites, este motivo corresponde a dos nucleótidos, ya que los alelos tienen variaciones de dos pares de bases entre ellos.

d) Podemos descartar a los individuos 3 y 5, pues el individuo 2, debe de tener para cada microsatélite un alelo procedente de la madre y otro del padre.

## 3. PROBLEMAS PARA RESOLVER

**Problema 1.**Un fragmento de ADN se corta con PstI y HindIII por separado. Posteriormente, se utiliza una mezcla de ambas enzimas obteniéndose los fragmentos indicados a continuación:

PstI: 3Kb y 4Kb
HindIII: 2Kb y 5Kb
PstI+HindIII: 1Kb, 2Kb y 4Kb

Dibujar el mapa de restricción de este segmento de ADN

**Problema 2.** Se digiere un fragmento de ADN clonado con las enzimas de restricción HindIII y SmaI y con una mezcla de ambas. Se obtiene:

HindIII: 2,5Kb y 5Kb
SmaI: 2Kb y 5.5Kb
HindIII+SmaI: 2Kb, 2,5Kb y 3Kb.

a) Dibujar el mapa de restricción

b) Cuando la mezcla de fragmentos producida por la actuación de las dos enzimas a la vez se corta además con la enzima EcoRI, se observa la pérdida del fragmento de 3Kb y la aparición de una banda de 1,5Kb en el gel de agarosa. Indicar sobre el mapa anterior el punto de corte de EcoRI

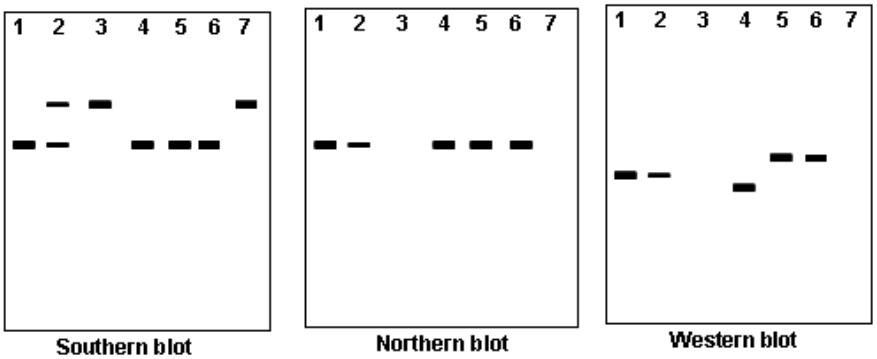
**Problema 3**. Un fragmento lineal de ADN de 11Kb se corta por separado con las enzimas de restricción EcoRI y HaeIII y con una mezcla de ambas obteniéndose los fragmentos indicados a continuación. EcoRI: 6Kb, 3Kb y 2Kb; HaeIII: 7Kb y 4Kb; EcoRI+HaeIII: 5Kb,3Kb, 2Kb y 1Kb. Dibujar el mapa de restricción de este segmento de ADN.

**Problema 4**. El plásmido pAl21 se cortó con diferentes enzimas de restricción y se observaron las siguientes bandas en un gel de agarosa: BamHI (3.7 Kb, 3.5Kb), PvuII (7.2Kb), HindIII (7.2Kb), BamHI+PvuII (3.5Kb, 2.4Kb, 1.3Kb), PvuII+HindIII (3.6Kb). Dibuja el mapa de restricción del plásmido.

**Problema 5.** Una proteína está codificada por un gen que no tiene intrones. El fragmento de restricción SacI que contiene el gen completo puede ser identificado por hibridación tipo Southern-blot con el ADNc del gen marcado radiactivamente. Para determinar la causa de una enfermedad desconocida, se obtuvo sangre de pacientes y de personas sanas como controles. Se extrajo su ADN, se cortó con la enzima SacI, se transfirió a una membrana de nylon y se híbrido con el ADNc marcado como sonda. Igualmente, se extrajo ARN, se sometió a electroforesis, se transfirió a una membrana (Northern-blot) y se hibridó con el ADNc. Además se realizó la técnica de Western-blot y se probó la proteína codificada por el gen mediante el uso de un anticuerpo específico frente a ella.

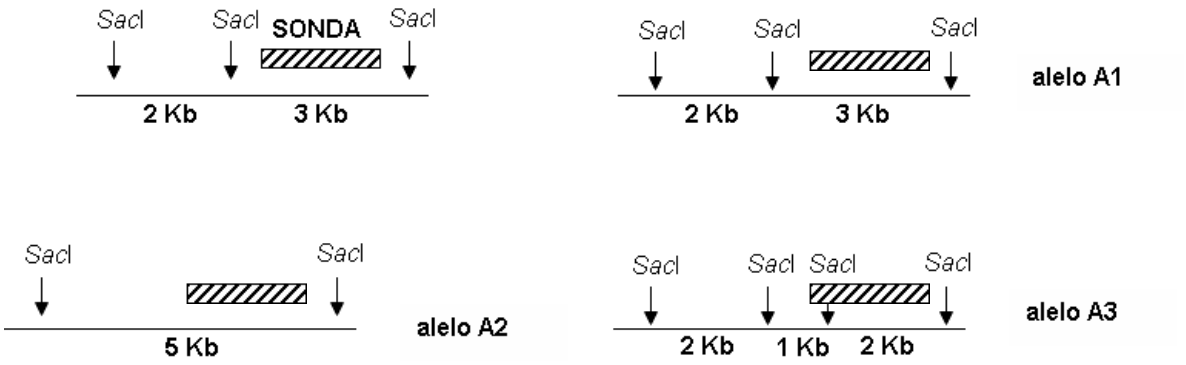
Los resultados se muestran a continuación (las personas 1 y 2 son controles sanos y las personas 3, 4, 5, 6 y 7 son enfermos).

¿Cuál puede ser la causa de la enfermedad en cada uno de los individuos enfermos?



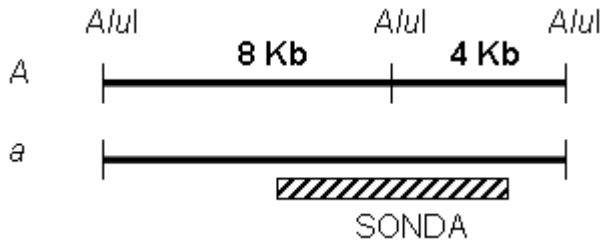

**Problema 6.**Tras una búsqueda de marcadores moleculares para una especie, se diseña una sonda complementaria al ADN genómico de esta especie en la región indicada en el esquema, y que permite diferenciar tres alelos distintos (A1, A2 y A3). El ADN extraído de

diferentes individuos se corta con SacI, se realiza una electroforesis y se transfiere el ADN posteriormente a una membrana de nylon. Esta membrana se hibrida con la sonda (marcada radiactivamente) y se realiza una autorradiografía.




Dibujar esquemáticamente el resultado esperado para: Un homocigoto para A1, un heterocigoto A1A2, un homocigoto A2 y un heterocigoto A1A3.

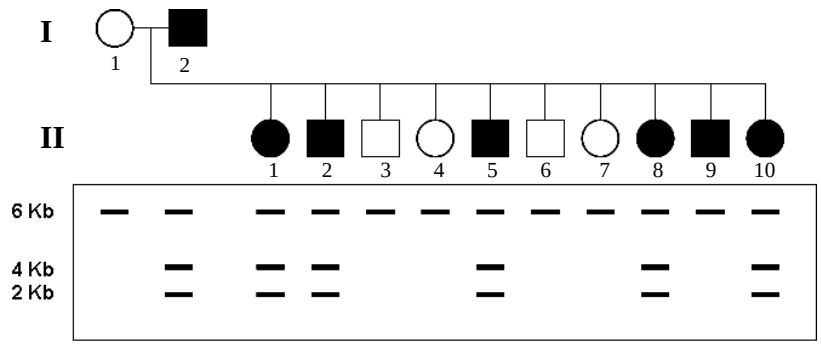
**Problema 7.**Un gen autosómico con dos alelos *A* y *a* se diferencian para la enzima de restricción AluI según indica la figura. Diseña un experimento para diferenciar los distintos genotipos de una población. Dibuja los posibles resultados.




**Problema 8.** Se prueban distintas sondas de ADN hibridando con el ADN genómico de los individuos de una familia numerosa en la que algunos miembros están afectados por una enfermedad autosómica dominante de manifestación tardía (aproximadamente a los 40 años). Sobre el Southern-blot obtenido con TaqI, una de las sondas detecta un polimorfismo en la longitud de los fragmentos de restricción (RFLP). Los patrones del RFLP de cada individuo del pedigrí se muestran en la figura.

a) Explicar los resultados.

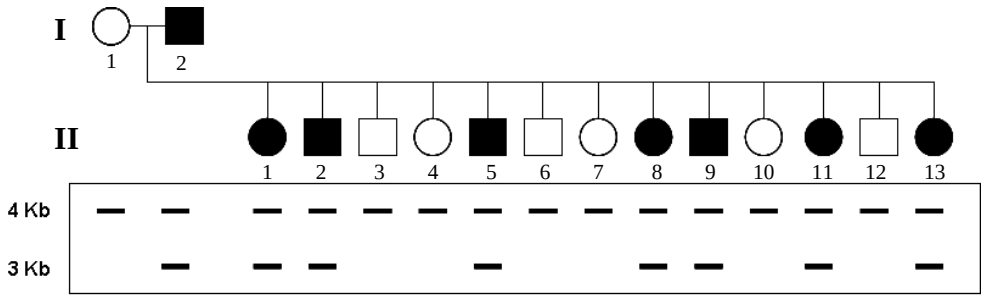
b) Analizar si existe ligamiento entre el RFLP y el gen causante de la enfermedad.



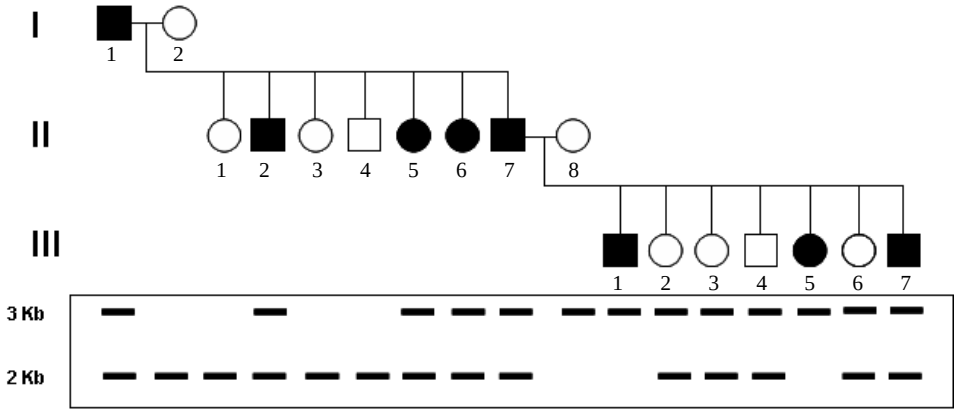
**Problema 9.** Se probaron distintas sondas de ADN hibridando con el ADN genómico de los individuos de una familia numerosa en la que algunos miembros están afectados por una enfermedad autosómica dominante leve. Sobre el Southern-blot obtenido con EcoRI, una de las sondas detecta un polimorfismo en la longitud de los fragmentos de restricción (RFLP). Los patrones del RFLP de cada individuo del pedigrí se muestran en la figura.

a) Explicar los resultados.

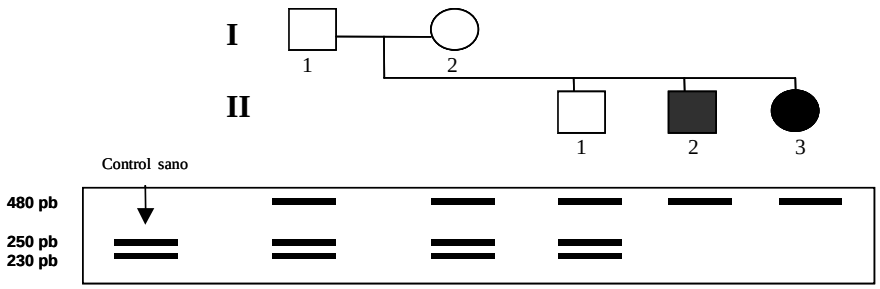
b) Analizar si existe ligamiento entre el RFLP y el gen causante de la enfermedad.




**Problema 10.** Se extrae ADN genómico de los miembros de una familia en la que existen afectados para una enfermedad autosómica dominante. Se digiere con PvuII y los fragmentos se separan en gel. Tras hibridación tipo Southern-blot con una sonda que detecta un RFLP se obtienen los resultados de la figura. ¿Está ligado el RFLP al gen causante de la enfermedad?




**Problema 11.** Una enfermedad está asociada a la ausencia de actividad de una enzima determinada. En cada miembro de la familia que se muestra a continuación, se amplificó el exón 2 del gen que codifica dicha enzima y se digirió con la enzima de restricción EcoRI, obteniéndose los siguientes resultados:

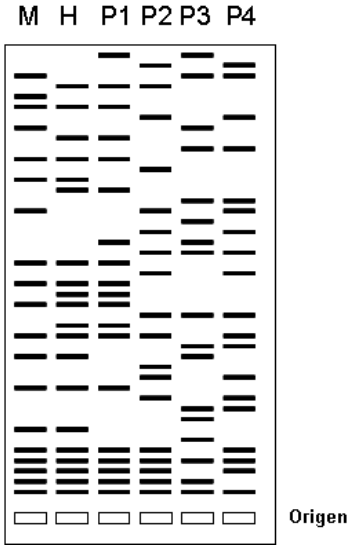


¿Se puede utilizar este marcador como método de diagnóstico? ¿Qué tipo de enfermedad se describe en el pedigrí? Explica, mediante un esquema, los resultados obtenidos a nivel molecular.

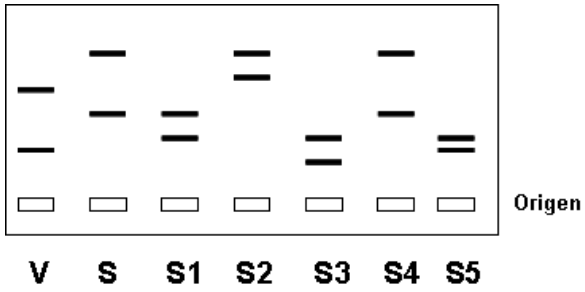
**Problema 12.** Una sonda detecta un RFLP con dos alelos alternativos de 1,7Kb y 3,8 Kb a partir de ADN de ratón digerido con HindIII. Un ratón, heterocigoto para un alelo dominante que determina cola curvada y con alelos para el RFLP de 1,7Kb y 3,8Kb, se cruza con un ratón silvestre que muestra sólo el fragmento de 3,8Kb. La mitad de los descendientes presenta cola curvada. Al analizar estos ratones con cola curvada para el RFLP, encontramos que un 20% de ellos son homocigotos para el alelo de 3,8Kb y el 80% son heterocigotos para los alelos 3,8Kb y 1,7Kb. a) ¿Está ligado el locus que determina la cola curvada al RFLP?; b) ¿Si lo está, a qué distancia se encuentran?; c) Explica estos resultados mediante un esquema.

**Problema 13.** Cuatro hombres se disputan la paternidad de un niño. Los forenses deciden utilizar el método de la huella genética para resolver el caso, analizando el ADN de la madre (M), del hijo (H) y de los cuatro posibles padres (P1 a P4). Los resultados obtenidos se muestran en la figura.
a) ¿Quién es más probable que sea el padre?

b) Atribuir el mayor número posible de bandas al padre y a la madre.




**Problema 14.**Se extrae el ADN de la sangre de una víctima de violación (V), del semen extraído de su cuerpo (S) y de muestras tomadas de 5 sospechosos (S1, S2, S3, S4 Y S5). Se lleva a cabo un estudio de microsatélites empleando una pareja de cebadores específicos de un locus. Una vez realizada la amplificación con la pareja de cebadores, se obtienen los siguientes resultados:

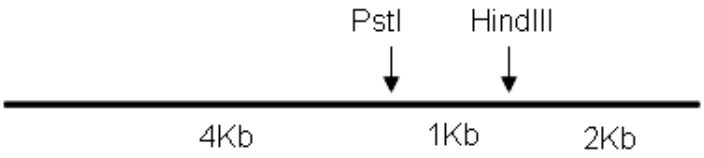


a) Explicar los patrones de amplificación obtenidos.

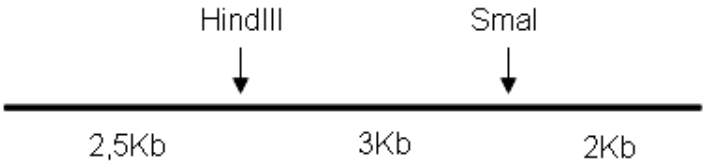
b) ¿Existe algún sospechoso que parezca culpable?

## 4. SOLUCIONES A LOS PROBLEMAS

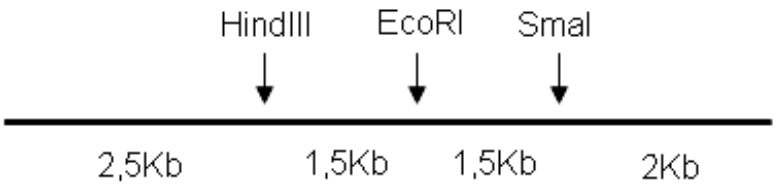
**Problema 1.**



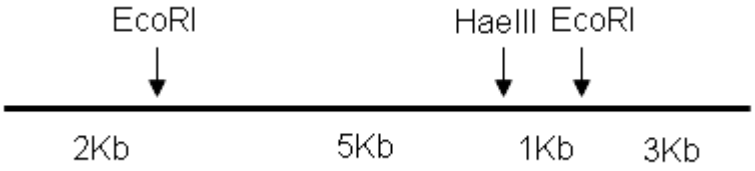

**Problema 2**
**a**)



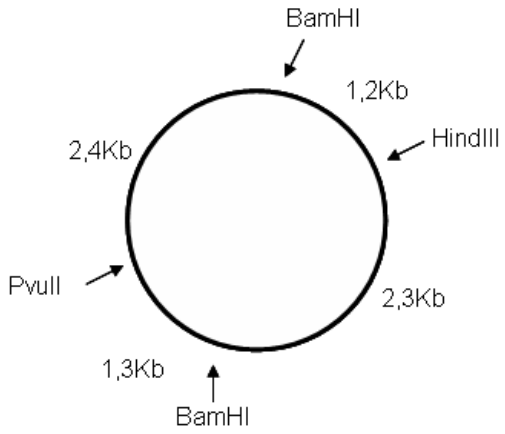

**b)**




**Problema 3.**



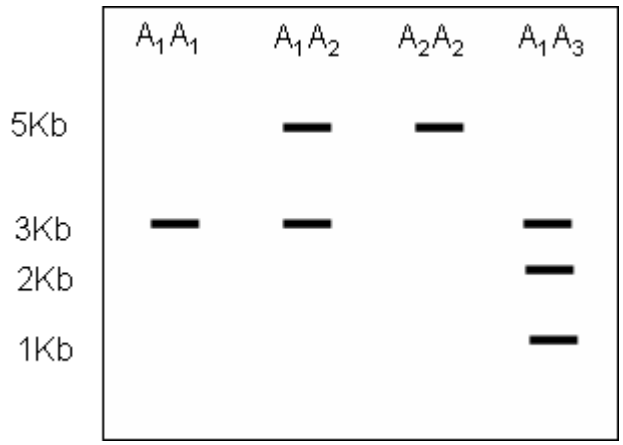
**Problema 4.**



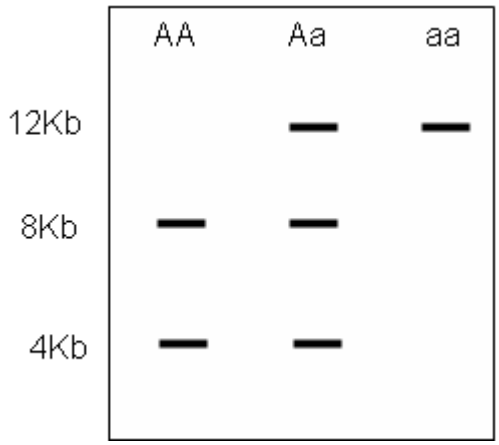

**Problema 5.**

No se transcribe el gen (individuos 3 y 7)o las proteínas producidas son defectuosas (individuos 4, 5 y 6). El individuo 1 es homocigótico para alelo normal y el 2 es heterocigótico portador de alelo que no se transcribe y produce la mitad de ARNm y de proteína que el 1. Los individuos 3 y 7 son homocigóticos para este último alelo y no producen proteína. Los individuos 4, 5 y 6, son homocigóticos para alelo que se transcribe y produce ARNm de igual longitud que alelo normal pero han sufrido algún cambio que cambia la secuencia de la proteína.

**Problema 6.**

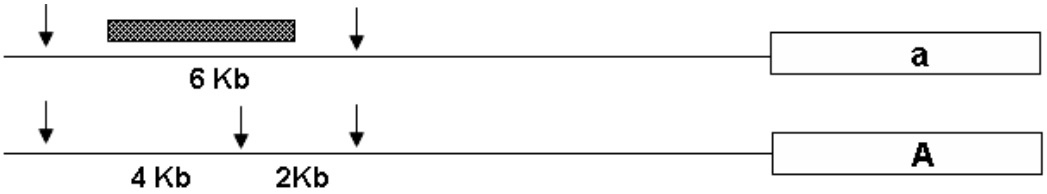



**Problema 7.**Se digiere el ADN genómico de los individuos con la enzima AluI, se somete el ADN cortado a una electroforesis y mediante la técnica de Southern-blot se transfieren los fragmentos a una membrana de nylon. Se realiza una hibridación utilizando la sonda para detectar las regiones homólogas a ellas. Los resultados posibles son:




**Problema 8.**

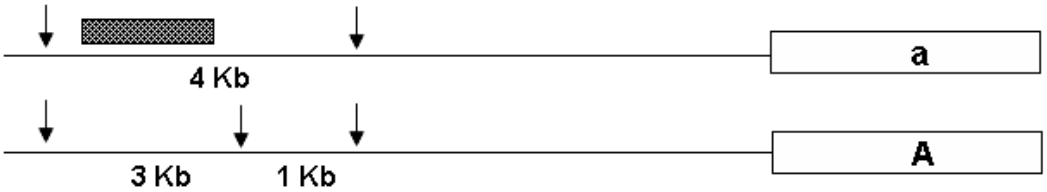
Están ligados, y una posible interpretación de los resultados se muestra en el siguiente esquema:




Afectados, Aa (6Kb/4Kb/2Kb); Sanos, aa (6Kb). El genotipo del individuo II-9 es resultado de un entrecruzamiento entre la región que contiene al RFLP y el gen causante de la enfermedad en una de las meiosis del padre.

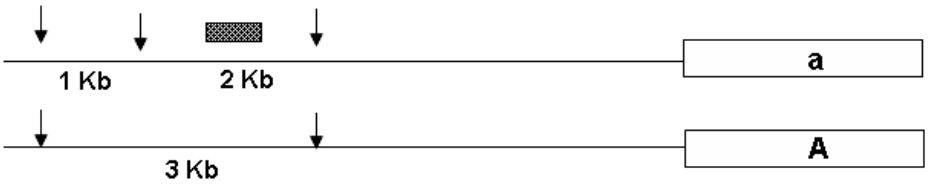
**Problema 9.**

Están ligados, y una posible interpretación de los resultados se muestra en el siguiente esquema:



**Problema 10.**

Se encuentran ligados. La interpretación podría ser:

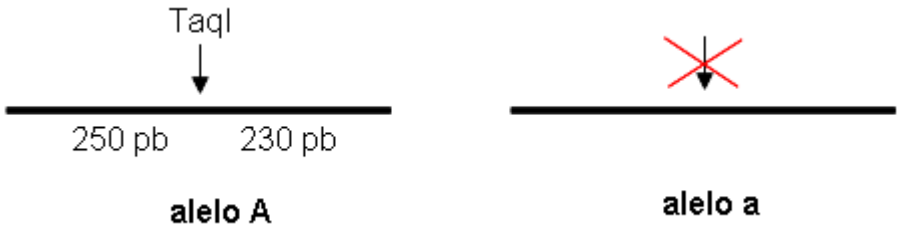



Afectados, Aa (3Kb/2Kb); Sanos aa (2Kb). Esto sería para las generaciones I y II. El individuo II-8 es un individuo procedente de otra familia, en el que, ahora, el alelo *a* está asociado a la banda 3Kb. Esto hace que el patrón de bandas cambie en la generación III. Así, los afectados Aa (A del padre y a de la madre) sean homocigóticos (3Kb/3Kb) y los sanos aa (a del padre y a de la madre) sean heterocigóticos (3Kb/2Kb).

El individuo III-7, es el resultado de un entrecruzamiento entre el RFLP y el gen causante de la enfermedad en el padre heterocigoto (II-7)

**Problema 11.**

Sí. Una enfermedad autosómica recesiva.

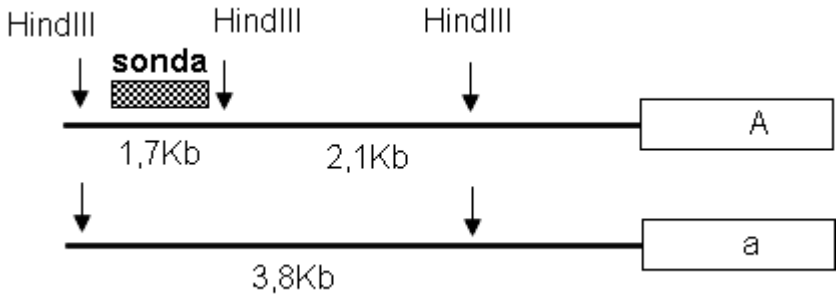



**Problema 12.**

**a)** Sí, se encuentran ligados.

**b)** 20 u.m.

**c)**



**Problema 13.**

**a)** El individuo P1.

**b)** La mitad deben de estar presentes en la madre y la otra mitad en el padre.

**Problema 14.**

**a**) Son debidos a la diferencia en el número de repeticiones en tándem

**b)** El S4

# PRÁCTICAS DE LABORATORIO Y SIMULACIÓN

# 1.- APLICACIÓN DE LA PCR AL DIAGNÓSTICO GENÉTICO: DETECCIÓN DE PARÁSITOS QUE INFECTAN A MOLUSCOS

## 1.1. OBJETIVO

En esta práctica se pretende comprobar la eficacia de la PCR en el diagnóstico genético de enfermedades e infecciones parasitarias en moluscos bivalvos. Se utiliza la especificidad de los *primers* para amplificar una región concreta del genoma del parásito cuando se encuentra presente en una muestra.

## 1.2. FUNDAMENTO TEÓRICO

La PCR es una técnica que permite la amplificación (multiplicación) específica de ADN *in vitro.* Para llevarla a cabo se necesita un ADN molde, un par de cebadores o *primers* que marcan los puntos del inicio de síntesis de la cadena 3' y 5' del ADN a amplificar, una cantidad suficiente de desoxiribonucleótidos tri-fosfato (dATP, dTTP, dCTP y dGTP), una ADN polimerasa, su tampón, y las condiciones para una eficiente reacción. La reacción es cíclica y, tras una etapa inicial de desnaturalización del ADN molde (de 2 a 5 minutos), consta generalmente de unos 25 a 35 ciclos. Cada uno de los ciclos está compuesto por una etapa de **desnaturalización** (unos 30-60 segundos), una de **alineamiento** de los cebadores al ADN molde (unos 30-60 segundos), y una de **extensión** (o polimerización, cuyo tiempo depende del tamaño del ADN a amplificar y de la polimerasa usada, y, por lo general, una aproximación es un minuto por kilobase de ADN a amplificar). Tras los ciclos de desnaturalización, alineamiento y extensión, la reacción termina con una etapa de extensión final que suele ser de cinco minutos.

Aunque la PCR puede detectar desde una única molécula de ADN, para su buen funcionamiento el ADN molde debe ser de buena calidad (no degradado) y libre de inhibidores de actividad enzimática. Por su parte, la región de ADN a amplificar (amplicón) debe tener un tamaño no superior a las 3 ó 4 kilobases y, preferiblemente, sin estructuras secundarias (éstas bloquean la progresión de la ADN polimerasa durante la síntesis). Por su parte, los cebadores son la cadena inversa y complementaria a la secuencia de inicio de síntesis de cada una de las dos hebras del ADN a amplificar. Deben ser específicos, de forma que se alineen exclusivamente con la región complementaria en el fragmento de ADN que se desea amplificar y no se unan a ninguna otra secuencia de ADN cercana. Suelen ser de un tamaño entre 15 y 35 nucleótidos (cuantos más nucleótidos, más especificidad). Los cebadores deben tener una composición equilibrada de CGs (bases Citosina y Guanina) y ATs (bases Adenina y Timina) y, sobre todo, no deben tener estructuras secundarias ni complementariedad interna o con el otro cebador (de lo contrario se plegarían sobre sí o se alinearían entre sí formando dímeros).

La desnaturalización del ADN se consigue mediante su incubación a 94ºC. Posteriormente, el alineamiento de los cebadores se consigue bajando la temperatura hasta un nivel (Tm= *Temperature of melting*) que permite a éstos unirse específicamente a su secuencia inversa y complementaria. Dicha temperatura (Tm) debe ser aproximadamente similar para los dos cebadores (no más de 5ºC de diferencia) y depende tanto de la composición como del

tamaño del cebador. Hay una variedad de algoritmos que permiten calcular la Tm; una formula básica para estimarla es: 4 x GC + 2 x AT, donde GC es el número de Gs y Cs en el cebador y AT el de As y Ts. Dichos algoritmos también permiten chequear el potencial de formación de estructuras secundarias o de complementariedad tanto interna como entre cebadores.

En principio cualquier ADN polimerasa puede servir para sintetizar ADN *in vitro*. Sin embargo, la PCR requiere altas temperaturas para la desnaturalización del ADN molde (94ºC) y para el alineamiento de los cebadores (40-65ºC o más dependiendo de los cebadores). Por eso, la PCR requiere ADN polimerasas termoestables, las cuales se consiguen a partir de microorganismos que viven en lugares con altas temperaturas y cuyas polimerasas están adaptadas a esta situación. La ADN polimerasa comúnmente utilizada para PCR es la polimerasa Taq, obtenida a partir de la bacteria termófila *Thermus aquaticus*, la cual tiene una temperatura óptima de polimerización del ADN de 72ºC, temperatura similar a la de donde vive este microorganismo. La fase de extensión con la Taq polimerasa se hace, por tanto, a 72ºC (otras ADN polimerasas tendrán otras temperaturas optimas de síntesis de ADN).

Como cualquier reacción bioquímica, la PCR necesita una solución tampón que es una mezcla de sales y reactivos (entre los cuales destaca el cloruro de magnesio). Las repeticiones cíclicas de diferentes temperaturas a lo largo de la reacción de PCR se consiguen mediante el uso de **termocicladores**. Estos son aparatos capaces de conseguir temperaturas precisas, mantenerlas durante un tiempo determinado y cambiar entre temperaturas de forma homogénea y rápida.

Así, la PCR consiste en la desnaturalización que abre la doble cadena del ADN molde, el alineamiento que permite el anclaje de los cebadores a sus correspondientes secuencias inversas y complementarias, y la extensión que permite la síntesis de ADN partiendo desde el último nucleótido 3' del cebador anclado a su correspondiente hebra de ADN molde. Un ciclo resulta en la duplicación del número de moléculas correspondientes al ADN a amplificar, tras el segundo ciclo habrá cuatro veces ese número, tras el tercer ciclo habrá ocho veces ese número de moléculas, etc. Al final habrá una cantidad teórica de $2^n$ x C moléculas de ADN amplificado donde n es el número de ciclos de PCR y C la cantidad de moléculas molde inicial. Se recomienda no superar los 35 ciclos de PCR ya que, por un lado, la ADN polimerasa tiene una tasa de error de síntesis (cerca de uno por cada millón de nucleótidos incorporados) y, por otro lado, el agotamiento diferencial de productos en la reacción puede resultar en más errores (por ejemplo si se agotan los dATPs, puede que la Taq inserte un dTTP en una posición correspondiente a un dATP).

Una vez finalizada la reacción de PCR se visualizan los productos de la reacción mediante la técnica de electroforesis en gel agarosa. Al cargar el producto de la PCR en el gel agarosa y someter este último en un campo eléctrico directo, se aprovecha la carga eléctrica negativa del ADN para hacerlo migrar diferencialmente desde el polo negativo al polo positivo del campo eléctrico directo (de polos positivo y negativo estables). La porosidad del gel de agarosa hará que, a medida que migren desde el polo negativo hacia el polo positivo, las moléculas de ADN se separen en base a su tamaño de forma que las moléculas más cortas migren más rápido (y por consiguiente avancen más en el gel). Para tener una referencia, se separan también las moléculas de ADN de una mezcla de fragmentos de tamaños conocidos y cantidades relativas (marcadores de peso molecular). La separación simultánea, pero por separado (en diferentes pocillos), de los productos de la PCR y del marcador de peso molecular en el mismo gel permite al investigador determinar los tamaños moleculares de los productos de la PCR que deben coincidir con los esperados.

La presencia de *Perkinsus spp.* es conocida prácticamente en todas las aguas cálidas del mundo y ha sido históricamente asociada a mortalidades masivas de moluscos bivalvos. La presencia de *Perkinsus olseni* en las almejas del litoral europeo se conoce desde 1987. Este

parásito se ha detectado por ejemplo en la almeja fina (*Ruditapes decussatus*), en la almeja japonesa (*R. philippinarum*), en la madrealmeja (*Venerupis pullastra*), en el pirulo (*V. aurea*) y en la almeja rubia (*V. rhomboides*). *Perkinsus olseni* puede considerarse en la actualidad el principal problema patológico para el desarrollo del cultivo de almejas en el litoral europeo. Hasta ahora, su diagnóstico precisaba de técnicas que requerían de tres a cinco días, y cuyo desarrollo y eficacia oscilaba entre el 60-90%. La puesta en marcha de nuevas técnicas más sensibles y rápidas constituye un avance muy importante en el control, ordenación y protección de las poblaciones y cultivos de moluscos bivalvos. Desde su implantación, la aplicación de la técnica de amplificación de ADN mediante la reacción en cadena de la polimerasa (PCR) ha revolucionado el diagnóstico de enfermedades infecciosas en Acuicultura. La sensibilidad y la rapidez son las cualidades más notables de estas técnicas.

En esta práctica se pretende determinar la presencia de parásitos en distintas muestras de moluscos bivalvos mediante la amplificación por PCR de un fragmento de ADN cuya secuencia es específica del parásito. Se trata de un fragmento del espaciador intergénico de los genes ribosómicos (Figura 1). Los genes que codifican para tres de los cuatro ARNs que forman parte del ribosoma (ARN ribosómicos 18S, 5.8S y 28S) se disponen formando una unidad de transcripción compuesta por la secuencia ETS (espaciador externo que se transcribe por delante del gen 18S), el gen 18S, ITS-1 (espaciador interno entre el gen 18S y el 5.8S), el gen 5.8S, ITS-2 (espaciador interno entre el gen 5.8S y el 28S), 28S y otro ETS (espaciador externo que se transcribe por detrás del gen 28S). En un locus ribosómico, varios cientos de estas unidades de transcripción se repiten en tándem separados por una secuencia NTS (espaciador no transcrito). Juntos, el NTS y los ETSs constituyen el llamado IGS (espaciador intergénico). Los genes ribosómicos se caracterizan por su elevado grado de conservación. No es el caso de los espaciadores entre estos genes, que al no codificar ningún producto génico, no están sujetos a presión selectiva, y por tanto, su secuencia es muy variable entre especies. Esto hace que el fragmento de ADN que nosotros vamos a amplificar (un fragmento de760 pb del NTS de *P. olseni*)tenga una secuencia específica del parásito.


**Figura 1.** Organización de los genes ribosómicos en los genomas eucarióticos. Las flechas indican el lugar de anclaje de los cebadores específicos.

# 1.3. METODOLOGÍA

### *Reacción de Amplificación (PCR)*

En un microtubo de 200µl añadir, siguiendo el orden indicado, los siguientes reactivos para un volumen final de 10µl:

- Agua estéril 2,5 µl
- 10% Tampón de PCR (10x) 1 µl
- 2mM de cada dNTPs (10 mM) 1 µl
- Primer PkI (0,2 µM) 2 µl
- Primer PkII (0,2 µM) 2 µl
- ADN de almeja 1 µl
- Taq polimerasa (2U) 0,5 µl

A continuación se colocan los microtubos en el termociclador y se programa para 35 ciclos según el programa:

| | | |
|---|---|---|
| Desnaturalización: | 94ºC | 30 seg. |
| Alineamiento: | 58ºC | 30 seg. |
| Extensión: | 72ºC | 30 seg. |

Se analizarán muestras de ADN procedentes de distintas almejas para determinar si están infectadas o no por el parásito. Una vez terminada la PCR, se someterán las muestras a una electroforesis en gel de agarosa y se procederá al diagnóstico de los individuos.

### *Preparación del gel de agarosa*

En un matraz de 250 ml de capacidad, añadir 40 ml de tampón TBE (0,04 M Tris-acetato; 0,04 ácido bórico; 0,01 M EDTA) y 0,4 g de agarosa (1% de agarosa).

Calentar utilizando el microondas hasta que se funda la agarosa.

Dejar enfriar hasta aproximadamente 50ºC y añadir 4 µl de solución de colorante para ADN SYBR® Safe (10.000x).

Mientras se enfría la agarosa, colocar en el adaptador el molde en el que se preparará el gel. Dejar el molde en el adaptador en una superficie horizontal y situar el peine que labrará los pocillos a unos centímetros del borde.

Una vez se ha enfriado la agarosa, se añade la solución al molde con cuidado de retirar las burbujas que se formen. Dejar gelificar la agarosa hasta que adquiera una apariencia translúcida.

Retirar el peine y el molde del adaptador. Colocar el gel en la cubeta de electroforesis y cubrirlo con tampón de electroforesis (TBE).

### *Electroforesis*

Con cuidado de no romper los pocillos, cargar en el gel las diferentes muestras correspondientes a cada una de las reacciones de amplificación. Para ello, añadir 2 µl de tampón de carga a los 10 µl de la reacción de PCR, una vez mezclado con el ADN, con la ayuda de una micropipeta, cargar la mezcla en un pocillo del gel (una muestra por pocillo). Cargar en otro pocillo 4 µl de la mezcla ya preparada de marcador de peso molecular, que nos servirá de referencia para determinar el tamaño de los fragmentos que queremos caracterizar.

Conectar la fuente de alimentación al gel durante 30 minutos a 50 volts/cm.

Analizar los resultados mediante la observación en un transiluminador.

### *Diagnóstico de los individuos*

En aquellos individuos donde observemos una amplificación correspondiente a 760 pb estará presente el parásito, y por tanto, los podremos diagnosticar como positivos para esta enfermedad.

Se espera un resultado como el de la figura:

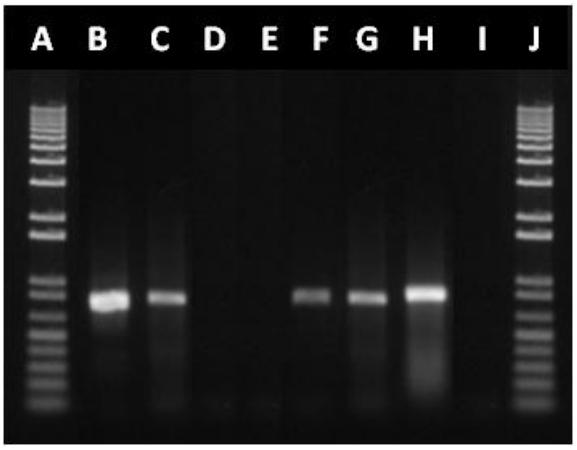



**Figura 2.** Se muestra el resultado del test de diagnóstico para *Perkinsus*. En un gel de electroforesis se cargaron diferentes muestras con el contenido de la reacción de amplificación: A y J, marcadores de ADN para determinación del tamaño del fragmento amplificado. B, amplificación de producto a partir de una muestra utilizada como control positivo (ADN de *Perkinsus*). C, amplificación de producto a partir de una muestra utilizada como control positivo. D y E, ausencia de amplificación del producto en muestras de tejidos de almejas procedentes de cultivos no infectados. F-H, amplificación del producto en muestras de tejidos de almejas procedentes de cultivos infectados. I, ausencia de amplificación del producto en muestras carentes de material biológico (prueba de control negativo). En esta Figura, la flecha señala los fragmentos amplificados de ADN (760 pb). Se demuestra la eficacia del método de diagnóstico para *Perkinsus olseni* en cultivos de almejas con los primers PkI y PkII, que detectan la presencia del parásito en almejas infectadas y no ocasiona problemas de falsos positivos puesto cultivos no infectados no mostraron amplificación. Además, se demuestra la gran sensibilidad del método, puesto que detecta la presencia del parásito en cultivos aun cuando el nivel de infección es mínimo.

## 1.4. RECURSOS WEB

A través del siguiente link de YouTube se puede acceder a diferentes vídeos de interés para las asignaturas de Genética.

https://www.youtube.com/playlist?list=PLBa9sJUx0zXWnO2Wu4H6qmJrEOIFNCzaI

De especial utilidad para esta práctica son los video-tutoriales para la preparación de un gel de agarosa, y el que lleva el nombre de esta práctica "Aplicación de la PCR al diagnóstico genético: detección de parásitos que infectan a moluscos", que muestran todos y cada uno de los pasos llevados a cabo en el laboratorio durante la sesión práctica.

## 1.5. CUESTIONES

**1.** ¿Qué criterios se deben seguir a la hora de diseñar cebadores para este tipo de análisis?

**2.** ¿Qué es un control negativo y un control positivo en la técnica de PCR?

**3.** ¿Qué harías si observa amplificación en muestras que claramente sabemos que no están infectadas?

**4.** ¿Qué es lo que hace de la PCR una técnica idónea para diagnóstico?

**5.** ¿Podríamos descartar completamente una infección si para una muestra observamos ausencia de amplificación tras la PCR?

**6.** ¿Podría detectarse mediante PCR, en un experimento similar al realizado en esta práctica, la presencia de varios parásitos al mismo tiempo?

# 2.- CLONACIÓN DE UN PRODUCTO DE PCR

## 2.1. OBJETIVO

Esta práctica tiene como objetivo conocer el procedimiento para clonar un fragmento de ADN. Para ello, construiremos una molécula de ADN recombinante que, en nuestro caso, estará formada por un vector de clonación y un fragmento de ADN del parásito *Perkinsus olseni.* Utilizaremos como vector de clonación un plásmido que presenta una serie de características especialmente favorables: 1) se facilita la inserción de un fragmento de ADN, 2) se replica de forma autónoma en células procariotas (*E. coli*) y 3) permite distinguir entre las colonias que han incorporado plásmidos recombinantes y las que incorporaron plásmidos sin inserto. De esta forma, se pretende dar a conocer una técnica de gran utilidad y de uso rutinario en los laboratorios de Genética Molecular.

## 2.2. FUNDAMENTO TEÓRICO

En Biología Molecular, el término clonación hace referencia a una técnica mediante la cual se logra introducir un fragmento de ADN de interés en un vector, siendo esta "construcción genética" introducida posteriormente en células bacterianas, de forma que logre mantenerse y multiplicarse (replicarse) dentro de las mismas.

Por lo tanto, los componentes principales de un experimento de clonación son: a) el fragmento de ADN a clonar, que se denomina inserto una vez integrado en el vector, b) el vector de clonación, donde el inserto es introducido permitiendo así su incorporación dentro de la célula, y c) las bacterias donde es introducida la construcción formada por inserto más vector (plásmido recombinante), permitiendo obtener muchas copias del mismo. Este tipo de experimentos se engloban dentro de lo que se conoce hoy día como tecnología del ADN recombinante, dado que se construyen moléculas de ADN compuestas por fragmentos de diferentes orígenes.

El inserto puede ser cualquier fragmento de ADN, sea cual fuere su origen. No obstante, el tamaño máximo a insertar está limitado por la capacidad del vector usado. En el caso de los plásmidos más comunes el tamaño del inserto no suele superar las 10 kilobases (Kb), y un inserto de mayor tamaño suele generar una construcción recombinante cuyo tamaño obstaculiza su eficiente penetración en las células bacterianas. Cuando necesitemos clonar fragmentos de ADN de mayor tamaño se puede recurrir a otros vectores, como el fago lambda, en el que podemos clonar fragmentos de unos 15 Kb, los cósmidos, que pueden aceptar hasta 40 Kb, los BACs (Bacterial Artificial Chromosome), que pueden aceptar hasta 200 Kb, los YACs (Yeast Artificial Chromosome), que pueden aceptar hasta 2 Mb y los MACs (Mammalian Artificial Chromosome), que permiten clonar fragmentos de varias Mb.

Para obtener el inserto de interés hay que recurrir a una fuente de ADN que lo incluya, por ejemplo, el genoma de un animal o de una planta. Después hay que seleccionar una técnica que nos permita aislar el fragmento de interés, como, por ejemplo, mediante digestión del ADN genómico con enzimas de restricción (véase el protocolo más abajo), o bien mediante amplificación del inserto mediante PCR (véase el guión y práctica correspondientes a esta técnica). En ambos casos obtendremos un fragmento de ADN de tamaño conocido, por lo que realizaremos una electroforesis en gel de agarosa, identificaremos el fragmento

adecuado y el ADN será recuperado mediante una técnica de purificación de ADN a partir de geles de agarosa.
En esta práctica usaremos un plásmido como vector de clonación. Un plásmido es una molécula de ADN bacteriano circular que, no siendo imprescindible para la supervivencia y multiplicación de la bacteria, puede coexistir y replicarse en el protoplasma celular como molécula extra-cromosómica y transmitirse a las células hijas. Por lo tanto, para que un plásmido pueda ser usado como vector de clonación, tiene que ser capaz de mantenerse y replicarse dentro de la célula. Esto es posible porque el plásmido contiene una secuencia denominada *origen de replicación*, específica de cada especie bacteriana, donde se une la ADN polimerasa y comienza la replicación del plásmido. En cuanto al número de copias existentes en el interior de una bacteria, los plásmidos se pueden clasificar en dos categorías: 1) *relajados*, si existen múltiples copias, y 2) *restringidos*, si hay una única o muy pocas copias por célula. Los plásmidos relajados suelen ser más ventajosos ya que permite una multiplicación eficiente del plásmido y, por consiguiente, del inserto.

Para facilitar la integración del inserto, los plásmidos poseen una región que contiene varias secuencias diana específicas de diversas enzimas de restricción (sitio de clonación múltiple o *polylinker*). En esta región se pueden insertar fragmentos de ADN obtenidos por digestión con enzimas de restricción cuyas dianas se encuentran en este sitio de clonación múltiple. Para clonar insertos amplificados mediante PCR, se suelen usar vectores abiertos (no circulares) cuyos extremos 3' terminan con un nucleótido de timina protuberante (que no tiene nucleótido complementario en la otra hebra). Estos plásmidos aprovechan el hecho de que la *Taq* polimerasa (la enzima que permite la amplificación de ADN mediante PCR) añade un nucleótido de adenina a cada extremo 3' del ADN amplificado. Las adeninas en los extremos del fragmento amplificado se pueden emparejar con las timinas de los extremos del plásmido, hecho que facilita la inserción del amplicón (se denomina así al producto de la PCR) en el plásmido.

Durante la introducción de los plásmidos en el interior de las bacterias (proceso conocido como transformación) sólo una proporción de las mismas incorporarán el plásmido (la eficiencia de transformación nunca es del 100%). Los plásmidos comúnmente usados para este fin, contienen además uno o varios genes de resistencia a antibióticos, lo que permite seleccionar las células transformadas (que han incorporado el plásmido). Para ello, tras el proceso de transformación, se cultivan todas la bacterias en un medio que contiene el antibiótico ante el cual el plásmido confiere resistencia y, como consecuencia, sólo aquellas que hayan incorporado el plásmido sobrevivirán. Además, el plásmido puede contener algún sistema que le permita discriminar entre las células que llevan el vector con el inserto y las que llevan el vector recircularizado (sin inserto). Un sistema muy usado es el que utiliza la secuencia del gen de la β-galactosidasa (gen *lac*Z, del operón *lac* de *E. coli*) interrumpido por la región de clonación múltiple (*polylinker*). Para que tenga lugar la expresión del gen de la β-galactosidasa, es necesaria la presencia de *IPTG*, molécula que actúa como un inductor continuo del gen. La proteína β-galactosidasa, en presencia de uno de sus sustratos, *X-gal* (5-Bromo-4-Cloro-3-Indol-β-D-galactósido), produce un precipitado de color azul: el *X-gal* es hidrolizado por la enzima, dando lugar a galactosa y 5-bromo-4-cloro-3-hidroxindol, que es oxidado originando 5,5'-dibromo-4,4'-dicloro-índigo, un compuesto azul insoluble. Así, si cultivamos en medio sólido bacterias transformadas con un plásmido que contenga el sistema de la β-galactosidasa en presencia de *IPTG* y *X-gal*, las bacterias que hayan incorporado plásmidos recombinantes (con inserto en el sitio de clonación múltiple) tendrán inactivo el gen de la β-galactosidasa, y no se formará el precipitado azul (darán lugar a colonias blancas), mientras aquellas que se hayan transformado con plásmidos sin inserto podrán producir el enzima, por poseer intacto su gen, y originarán colonias de color azul (Figura 1).

Existen otras estrategias similares que se pueden usar con el mismo propósito. Por ejemplo, usar plásmidos que contienen la secuencia de un gen letal interrumpida por el sitio de clonación múltiple. En este caso, la inclusión del inserto en el sitio de clonación múltiple

interrumpirá al gen letal, siendo las bacterias transformadas con plásmidos con inserto las únicas que sobrevivan.

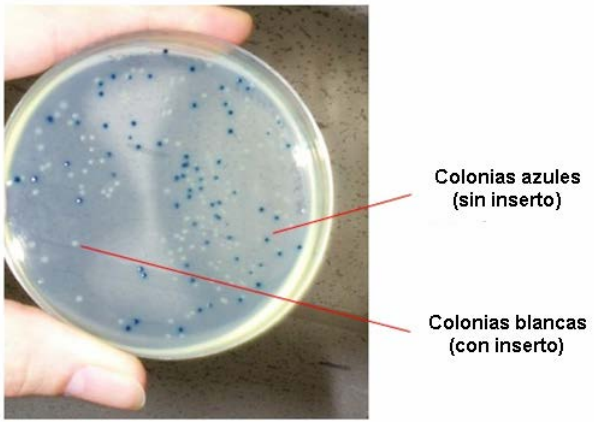



Figura 1: Resultado de un experimento de clonación

Hoy día, para usos convencionales, el investigador no tiene la necesidad de construir sus propios vectores, ya que existe una amplia variedad de vectores diseñados y producidos por empresas de Biotecnología para todo tipo de usos en clonación (por ejemplo los vectores pGEM-T (utiliza el gen de la β-galactosidasa como marcador de selección) y el vector TOPO (utiliza un gen letal).

La especie y cepa bacteriana usada durante el proceso de clonación también debe tener una serie de características especiales. No debe ser patógena (obviamente para evitar riesgos al personal investigador y a la población en general) y debe ser fácil de cultivar (se utilizan cepas no patógenas *de Escherichia coli*). Es preferible que tenga una reproducción (multiplicación) eficiente y que esté modificada de forma que se evite la recombinación entre el plásmido (vector) y su propio cromosoma (de lo contrario, se corre el riesgo de perder el inserto). De forma natural, una bacteria puede adquirir un estado fisiológico que la capacita ("permeabiliza") para sufrir un proceso de transformación. En esta situación se dice que la bacteria es "competente". Sin embargo, esta competencia natural ocurre con una frecuencia muy baja y no es útil con fines de clonación. Por ello, en el laboratorio se recurre a inducir artificialmente este estado con diversos métodos, proceso que se denomina "competencia artificial". Dicha permeabilización se puede inducir por métodos químicos. Para ello, las células se enfrían en presencia de cationes divalentes como $Ca^{2+}$ (en forma de $CaCl_2$), lo que prepara las membranas celulares para ser permeables al ADN plasmídico. Después, las células son incubadas en hielo con el ADN y luego se someten brevemente un choque térmico (por ejemplo, 42°C por 30-120 segundos), lo que facilita que el ADN entre en la célula. La permeabilización también puede conseguirse usando elementos físicos, como la corriente eléctrica. En este caso, las células bacterianas se someten a una corriente eléctrica de alto voltaje (alrededor de 2000V para el caso de las bacterias) y corta duración (varios µs). Como en el caso de los vectores, existe una gran variedad de cepas de bacterias "competentes" proporcionadas por empresas biotecnológicas para todo tipo de usos en clonación (por ejemplo las cepas de *E. coli* DH5α y JM109).

Entre los múltiples usos de la clonación, podemos citar la multiplicación de las copias de un fragmento, ya que, al replicarse el plásmido recombinante dentro de la célula y al multiplicarse esta última, se consiguen muchas moléculas de ADN. La clonación también permite la discriminación entre diferentes secuencias o variantes de un ADN amplificado. Por lo general, cada bacteria transformada adquiere un sólo plásmido recombinante. Al ser cultivadas en medio sólido, cada bacteria originará una colonia de bacterias idénticas a la original y, por lo tanto, con el mismo inserto. La secuenciación de los insertos procedentes de diferentes colonias nos permitirá tener una idea sobre la variabilidad de las secuencias de ADN originales. Otra utilidad de la clonación es la generación de una genoteca o librería

genómica, que consiste en un conjunto de clones bacterianos cada uno de los cuales porta un fragmento de ADN del genoma de la especie objeto de estudio. Cada fragmento está incluido en un clon, y entre todos los clones, componen el genoma entero. Las genotecas también pueden contener fragmentos de ADNc (ADN complementario o copia), obtenidos por retro-transcripción de ARNm. En este caso el número de clones es menor ya que sólo estarán representados los genes que se expresaban en el tejido que se usó para extraer el ARNm. Igualmente los insertos serán, en general, de menor tamaño ya que los genes clonados no contendrán intrones.

La clonación también puede permitir que un gen se exprese dentro de una célula bacteriana. Para ello es necesario clonar el fragmento en fase con la pauta de lectura abierta del gen (normalmente el ADNc obtenido a partir del ARN mensajero) en un vector de expresión. El vector de expresión tiene un promotor especial que permite la inducción controlada de la transcripción de la secuencia insertada. Como resultado, las bacterias pueden sintetizar la proteína codificada por el inserto, permitiendo la producción de enzimas y otras proteínas de interés científico, farmacológico o comercial.

En esta práctica, vamos a clonar fragmentos de ADN que previamente hemos amplificado por PCR. Dichos fragmentos contienen una región del ADN espaciador IGS del ADN ribosómico (ADNr) del parásito *Perkinsus olseni*. Como vector de clonación vamos a utilizar el plásmido pGEM-T, un vector que presenta las características descritas anteriormente.

## 2.3. METODOLOGÍA

Para llevar a cabo la clonación, vamos a seguir los siguientes pasos:

***Obtención del fragmento a clonar y del vector de clonación***

El ADN a clonar será el producto obtenido en la práctica de PCR (véase el guión y práctica correspondientes a esta técnica). El vector de clonación corresponde al plásmido comercial pGEM-T (*Promega*). En una reacción de PCR, la Taq-polimerasa tiene una actividad transferasa-terminal, no dependiente del ADN molde, que añade un nucleótido de adenina en los extremos 3´ de los productos amplificados. El vector pGEM-T se encuentra en forma lineal y presenta en sus extremos 3´ un nucleótido de timina. Esto permite una eficiencia mucho mayor de la unión entre el fragmento amplificado y el vector (Figura 2).

***Ligado***

Se trata, en este caso, de ligar los fragmentos de ADN obtenidos por PCR con el vector pGEM-T. En este proceso interviene una enzima llamada ligasa, que establece un enlace fosfodiéster entre la última base del producto amplificado por PCR (A) y la primera base de los extremos del vector T sin incorporar un nuevo nucleótido. Esto tiene como consecuencia que se produzca la unión entre las cadenas de ADN correspondientes al vector y al inserto por complementariedad entre las bases (Figura 2). Para llevar a cabo esta reacción, se realizan los siguientes pasos:

1. En un microtubo Eppendorf añadir:

    - 7 µl del producto de PCR (100-200 ng)
    - 1 µl de tampón de clonación (10x)
    - 1 µl del vector pGEM-T
    - 1 µl de la enzima ligasa

2. Incubar durante 30 minutos a temperatura ambiente.

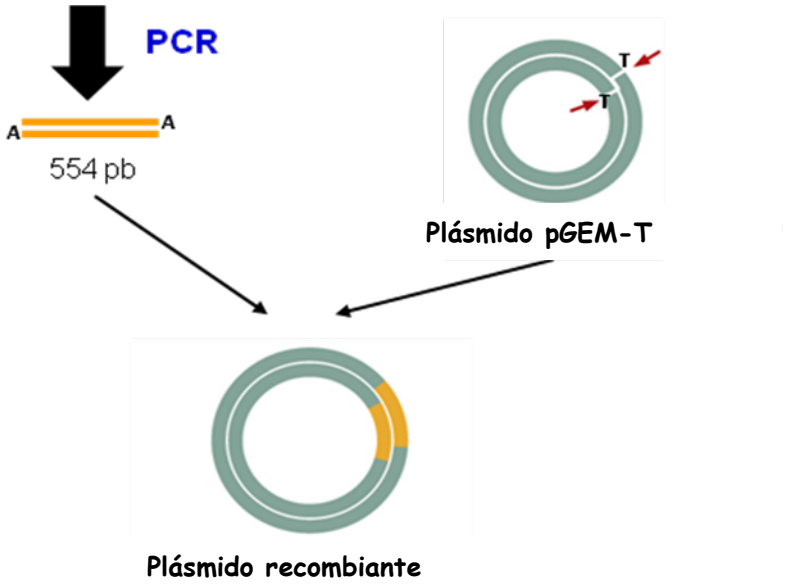



Figura 2: Ligado de un amplificado de PCR al vector de clonación pGEM-4Z

***Transformación***

Como se indicó en la introducción, transformación es el proceso mediante el cual los plásmidos se introducen en células bacterianas. Para ello utilizaremos bacterias competentes de la cepa JM109de *E. coli.* Se procederá de la siguiente forma:

- Depositar en hielo un tubo Eppendorf durante unos minutos conteniendo 50µl de bacterias competentes.
- Añadir 10µl de la solución de ligación.
- Dejar 20 minutos en hielo.
- Choque térmico introduciendo el tubo Eppendorf con la mezcla en un baño a 42ºC durante 45 segundos.
- Inmediatamente, pasar el microtubo a hielo, durante 2 minutos.
- Añadir 1000 µl de medio de cultivo LB líquido.
- Incubar durante 30-40 minutos a 37ºC con agitación.
- Sembrar 50 µl del cultivo líquido en placas de medio LB sólido con Ampicilina, X-gal e IPTG.
- Incubar las placas en posición invertida durante toda la noche en una estufa a 37ºC.

***Observación de los resultados***

Tras un periodo de incubación de entre 16 y 24 horas, observaremos la placa resultante en la cual podremos encontrar colonias de color blanco (con el plásmido recombinante).

## 2.4. RECURSOS WEB

A través del link de YouTube se puede acceder al video-tutorial de la práctica.

https://www.youtube.com/playlist?list=PLBa9sJUx0zXWnO2Wu4H6qmJrEOIFNCzaI

## 2.5. CUESTIONES

**1.** ¿Qué son células competentes? ¿Qué características tienen?

**2.** En el proceso de clonación, ¿en qué paso se introduce el plásmido recombinante en la bacteria?

**3.** ¿Serviría el vector pGEM-T para clonar un fragmento de ADN cortado por enzimas de restricción? ¿Y uno amplificado por una ADN polimerasa de alta fidelidad?

**4.** ¿En qué se parece/diferencia la técnica de PCR a la de clonación?

**5.** ¿Cuál de las dos técnicas se parece más a lo que ocurre en la fase S del ciclo celular?

**6.** ¿Qué marcadores seleccionables hay en el vector p-GEM-T que son de utilidad para discriminar entre colonias portadoras de plásmidos recombinantes de las portadoras de plásmidos no recombinantes? ¿Se te ocurren otros posibles marcadores?

**7.** Tras una eficiente transformación ¿por qué se encuentran en la célula más colonias azules que blancas?

# 3.- BASES DE DATOS DE SECUENCIAS DE ADN y PROTEÍNAS

## 3.1. OBJETIVO

Con esta práctica se pretende introducir al alumno en el conocimiento y manejo de las bases de datos de secuencias de ADN y proteínas.

## 3.2. FUNDAMENTO TEÓRICO

### 3.2.1. La información biológica

Como todas las ciencias, la Biología no cesa de generar cantidades cada vez más extensas de información. A diario, los biólogos están constantemente haciendo descubrimientos y produciendo datos (información) sobre aspectos relacionados con los seres vivos. Esta información abarca desde características básicas (por ejemplo, la estructura molecular y configuración tridimensional de una proteína) hasta aspectos más complejos (por ejemplo, la taxonomía, relación filogenética y ecología de los organismos).

Al mismo tiempo, para conocer el estado de un tema de investigación, los biólogos necesitan acceder continuamente a información y datos obtenidos previamente por otros investigadores. Además de la bibliografía, los genetistas, por ejemplo, necesitan información sobre metodologías, técnicas y reactivos utilizados habitualmente. Pero también necesitan otro tipo de información sobre la especie objeto de su investigación, como pueden ser información sobre secuencias de genes (o ADN en general) o proteínas, sus variantes, su función, las interacciones de esos genes con otros, su relación con secuencias en otros organismos, lo que se sabe sobre su patrón de expresión, efecto de silenciamiento (o mutación), etc.

La información ya disponible sobre un tema en concreto es la base sobre la cual se desarrollan nuevas ideas, y su conocimiento es lo que evita que se investigue repetidamente sobre hechos sobradamente conocidos. Se podría decir que el avance en el conocimiento científico tiene un primer paso, muy importante y necesario, que es la revisión de los trabajos de investigación que han sido desarrollados hasta el momento sobre el tema en cuestión.

### 3.2.1. Almacenamiento de la información biológica

Buena parte de la información que se adquiere se olvida con facilidad a no ser que sea almacenada de una o varias formas. Antiguamente, e incluso en ciertas civilizaciones actuales, la conservación de la información se lleva a cabo a través de la denominada memoria colectiva, de transmisión oral de padres a hijos. Esta forma tradicional de transmisión de la información tiene como desventajas la limitación en la cantidad de información que puede “almacenarse” y el riesgo (casi inevitable) de la deformación de la información. Almacenar la información de forma escrita ofrece una capacidad de almacenaje de la información prácticamente ilimitada y una fiabilidad absoluta. Antes de la existencia de ordenadores, en ciencias, al igual que en otras disciplinas, la única forma de dar a conocer, almacenar u obtener información de los experimentos ya realizados era mediante su

publicación en revistas científicas. La información más relevante publicada acababa formando parte de libros científicos y de texto. En esas condiciones, obtener bibliografía, conocer métodos o información previa podía llegar a ser un obstáculo ya que uno debía tener acceso físico a la revista o revistas que contenían la información buscada. Este hecho implicaba que, además de tener que adquirir todos los libros que se pudiera, había que suscribirse a revistas científicas y guardar todos los ejemplares de un modo que permitiera saber dónde estaba la información y poder recuperarla cuando fuera necesario consultarla.

El desarrollo de los ordenadores personales, limitados al principio por su poca capacidad de almacenaje, supuso un avance significativo ya que permitía poder guardar la información en formato digital. Pero todavía había que depender de material físico (disquetes) para conseguir la información digitalizada o para trasladarla entre ordenadores. Sin embargo, al igual que se hizo ciencia antes del descubrimiento de la electricidad e incluso de la máquina de escribir, hasta los años ochenta, los investigadores no podían contar con la poderosa herramienta que es internet. Desde su aparición, internet supuso un salto tanto cuantitativo como cualitativo para la publicación, almacenaje, búsqueda y obtención de datos. Teniendo acceso a internet, desde cualquier punto en el mundo, un investigador puede conseguir desde bibliografía hasta datos sobre el gen o proteína que le interesa, incluidas las secuencias, sus variantes, secuencias homólogas, datos de expresión, de efecto, de mutación, de función, etc. Además, prácticamente todas las revistas científicas están actualmente disponibles *online* (muchas requieren suscripción pero otras son de acceso libre). Incluso muchos artículos publicados en fechas anteriores a la existencia del ordenador están ahora digitalizados. Internet, junto con la cada vez más potente capacidad de almacenaje de los discos duros de los ordenadores, ofrecieron la posibilidad de centralizar las formas de almacenaje y organización de la información en forma de bases de datos.

En genética las bases de datos son hoy por hoy una herramienta vital para la investigación. Para la genética actual, es imprescindible tener acceso a las secuencias de ADN (incluidos genomas), ARN (incluidos transcriptomas) y proteínas (incluidos proteomas) que ya están identificadas. A menudo se requiere también información sobre vías y redes génicas que ofrecen información sobre las interacciones génicas. Esta, junto a información sobre la expresión, función y evolución del gen de interés están disponibles en bases de datos que son cada vez más completas. No se exagera si se dice que un genetista actual no puede desarrollar su actividad investigadora sin acceso a las bases de datos.

En lo que se refiere al análisis de secuencias de ADN o proteínas, los investigadores disponen actualmente de bases de datos donde se almacenan estas secuencias además de sus variantes, secuencias homólogas, y una gran cantidad de información sobre su localización cromosómica, propiedades, expresión, función, relaciones filogenéticas, etc. Obviamente, la procedencia de estas secuencias es la ciencia misma, ya que cada vez que un grupo de investigación identifica una secuencia, o genoma, las sube a la base de datos y, subir secuencias a las bases de datos es un requisito para la publicación en revistas científicas de los hallazgos relacionados con dicha secuencia.

En ocasiones, la enorme logística requerida para construir y mantener una base de datos depende de proyectos científicos individuales (cuando se trata de bases de datos orientadas hacia un organismo específico) o de un esfuerzo gubernamental o incluso intergubernamental (como es el caso de las bases de datos generales con más uso). Ejemplos del primer caso incluyen las bases de datos sobre los organismos modelo (Figura 1):

La mosca de la fruta, *Drosophila melanogaster* (http://flybase.org/)
El nematodo *Caenorhabditis elegans* (http://www.wormbase.org/)
La planta *Arabidopsis thaliana* (http://www.arabidopsis.org/).

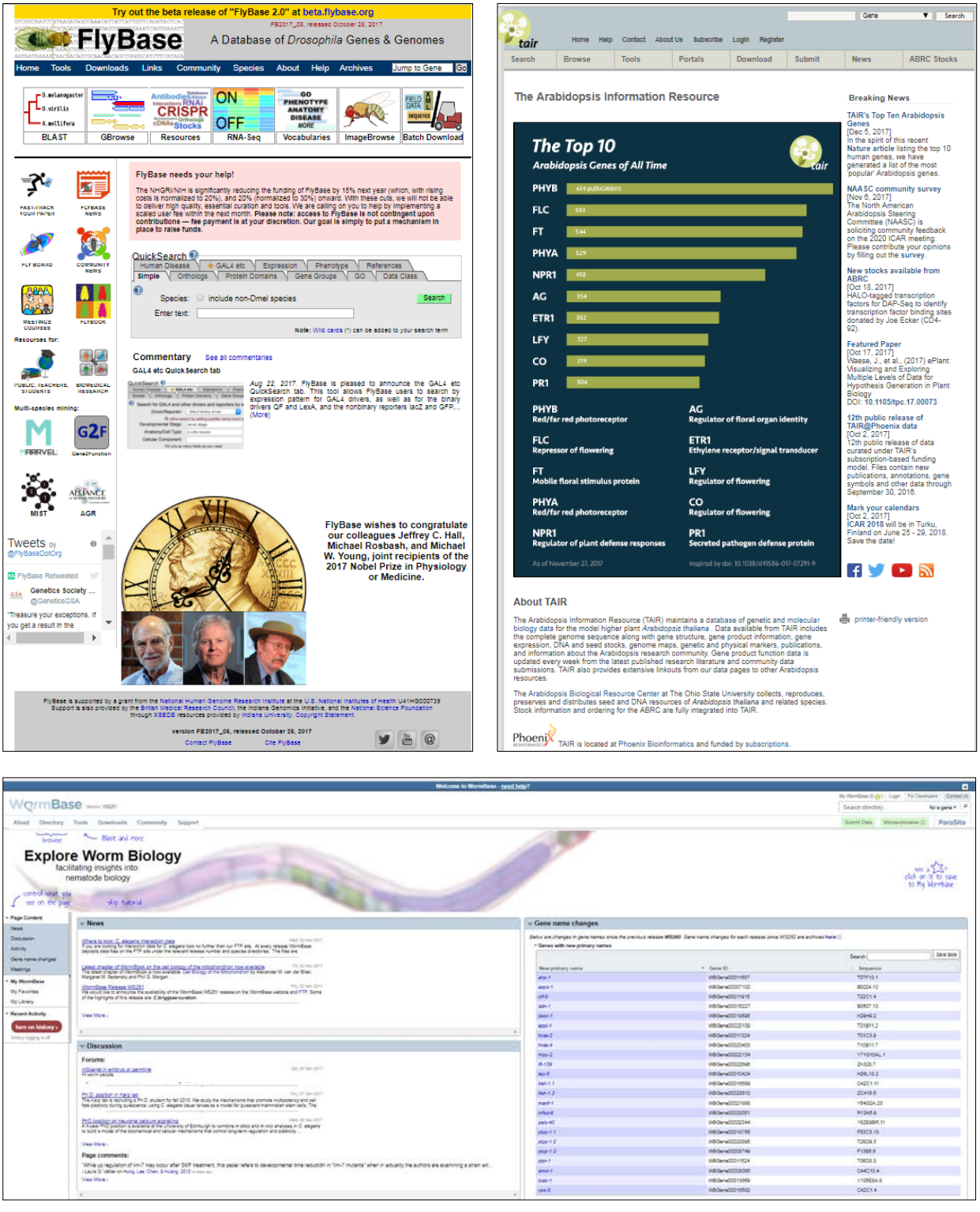






Figura 1: Ejemplos de bases de datos de organismos específicos

Entre las bases de datos generales, las más relevantes son (Figura 2):

- *The ADN DataBank of Japan* (DDBJ, http://www.ddbj.nig.ac.jp/index-e.html)
- *The European Molecular Biology Laboratory* (EMBL,) y su "hermana" *The European Nucleotide Archive* (http://www.ebi.ac.uk/ena/)
- *GenBank*, una base de datos del estadounidense *The National Center for Biotechnology Information* (http://www.ncbi.nlm.nih.gov/genbank/). De ellas surge el proyecto internacional *The International Nucleotide Sequence Database collaboration* (http://www.insdc.org/).

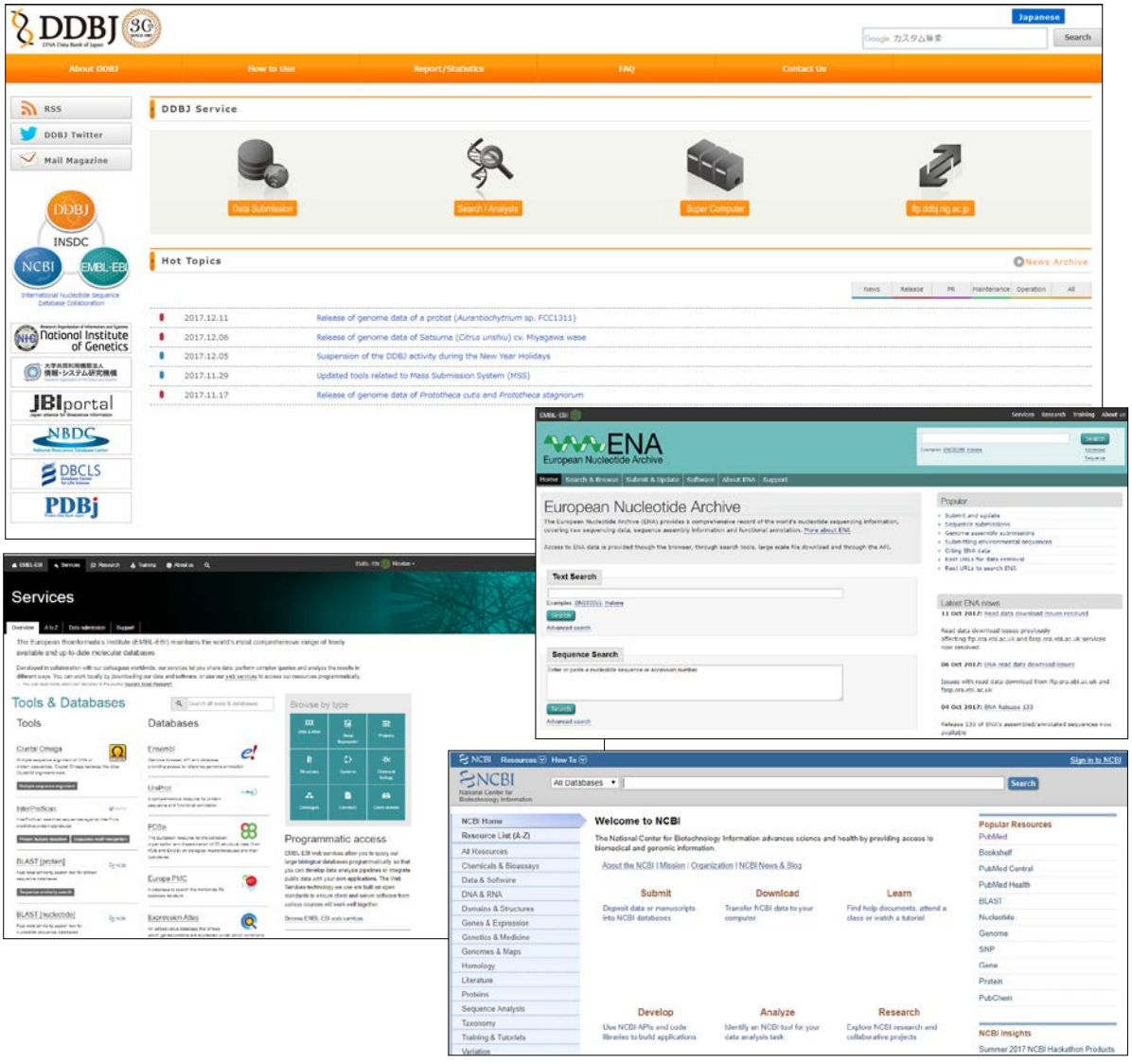



Figura 2: Bases de datos generales más relevantes

Además del personal investigador, las bases de datos también dependen de algoritmos (programas informáticos) que permiten la automatización del proceso de obtención, almacenaje, organización y eficiente presentación y accesibilidad de su contenido. Otros algoritmos, integrados en las bases de datos, permiten el análisis de las secuencias de

interés y su comparación con otras (ejemplos de estos son los conocidos programas de alineamiento y comparación de secuencias *Clustal* y *Blast*).

### 3.2.3. Identificación y formato de las secuencias de nucleótidos y aminoácidos en las bases de datos

Para conseguir una secuencia desde una base de datos se puede recurrir a 4 tipos de búsquedas. La secuencia se puede encontrar utilizando el nombre del gen y organismo correspondiente (o bien con el nombre del gen y luego seleccionando el organismo que nos interesa); la Figura 3 muestra el resultado de la búsqueda en el directorio de genes de la base de datos NCBI de la secuencia de los genes del colágeno en humanos. Sin embargo, las secuencias en las bases de datos están catalogadas y etiquetadas con un número de acceso y un identificador únicos, y unas etiquetas informativas sobre su origen y otras más características (véase formatos de secuencias). Esto ofrece la posibilidad de conseguir directamente la secuencia buscando por su número de acceso o identificador. En el caso del gen de colágeno humano de tipo 3 alpha 1, la secuencia puede obtenerse buscando en el directorio de genes de *GenBank* (la base de datos más completa) por el número de acceso X15332, o por el identificador COL3A1.

Dos formas más indirectas de conseguir las secuencias son mediante búsqueda *Blast* (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome) con secuencias homólogas (secuencias del gen pertenecientes a organismos filogenéticamente cercanos), o bien mediante navegación en el cromosoma correspondiente utilizando navegadores genómicos como *Genome Browser* (http://genome.ucsc.edu/cgi-bin/hgGateway) o *Ensembl* (http://www.ensembl.org) en el caso de que se conozca el genoma del organismo y la localización de la secuencia de interés. En el caso del gen del colágeno habría que navegar alrededor de los nucleótidos 189833342 y 189883227 en la banda 32 del brazo largo del cromosoma 2 (*chromosome: 2; Location: 2q32.2*).

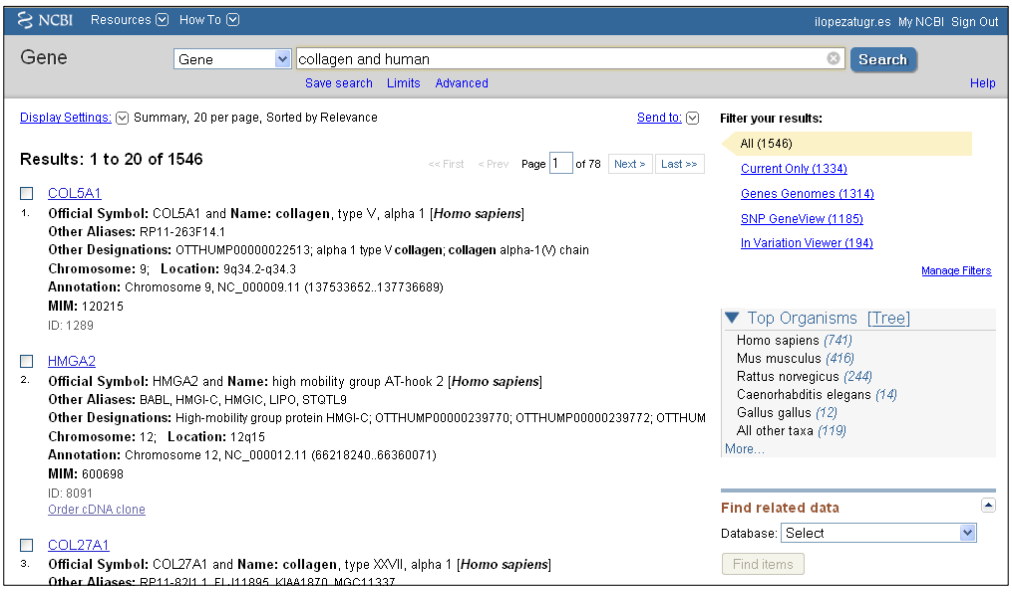



Figura 3: Búsqueda de secuencias de genes del colágeno humanos en NCBI

Una vez conseguida, la secuencia puede estar presentada en un formato u otro dependiendo de la base de datos de la que se obtengan; aquí introduciremos los tres formatos más utilizados. Se trata de los formatos “europeo” EMBL, el “estadounidense” GenBank (ambos incluyen información y varias etiquetas identificadoras de la secuencia y de su procedencia) y

el “sencillo y universal” fasta que puede no incluir más que un encabezamiento con el nombre de la secuencia.
Como hemos mencionado antes, el formato fasta es el más sencillo ya que incluye solo una parte comentario, o título; cuyo inicio está señalado por el símbolo “>”, y que suele ser el nombre de la secuencia, su procedencia, numero de acceso a la base de datos, seguido por un salto de línea y la secuencia de nucleótidos o aminoácidos que suele estar presentada en líneas de 80 o 120 residuos, aunque, aparte del primer salto de línea entre el título y la secuencia, el formato ignora espacios y acepta secuencias en forma de residuos continuos sin espacio o salto de línea. El fin de la secuencia es simplemente el último carácter (residuo) de la misma (véase el ejemplo que sigue). Al ser tan sencillo, el formato fasta es el formato base requerido por la gran mayoría de programas y algoritmos de análisis de secuencias y, por lo tanto, el más usado por los investigadores a la hora de manejar secuencias (alinearlas, hacer arboles filogenéticos, hacer búsquedas blast, etc.). El fichero fasta puede ser un fichero de texto simple o tener una de las extensiones “.fas” o “.fasta”. Un fichero con secuencias fasta puede tener una o varias secuencias cada una con su línea identificativa (que empieza por “>”).

**Ejemplo de formato fasta** (los puntos en negrita dentro de la secuencia indican que hemos quitado residuos para ahorrar espacio, ya que la secuencia completa es de unos 5kb):

```
>embl|X15332|X15332 Human COL3A1 mRNA for pro alpha-1 (III) collagen

cagaactattctccccagtatgattcatatgatgtcaagtcgggcggagtagcagtaggaggactcgcaggct
atcctggaccagctggccccccaggcccccccggcccccctggtacatctggtcatcctggttcccctggatc
tccaggataccaaggacccccctggtgaacctgggcaagctggtccttcaggccctccaggacctcctggtgct
ataggtccatctggtcctgctggaaaagatggagaatcaggtagacccggacgacctggagaccgaggattgc
ctggacctccaggtatcaaaggtccagctgggatacctggattccctggtatgaaaggacacagaggcttcga
tggacgaaatggagaaaagggtgaaacaggtgctcctg...ccctggtccttgctgtggtggtgttggagccc
ctgccattgctgggattggagctgaaaaagctggcggttttgccccttattatggagatgaaccaatg
```

Por su parte, tanto el formato EMBL como GeneBank son más elaborados e incluyen más identificadores e información sobre la secuencia. Ambos comparten la característica de tener, en su parte inicial, anotaciones que indican el número de acceso de la secuencia y, al igual que el fasta, pueden tener una o varias secuencias cada una marcada por su identificador. La columna izquierda del fichero EMBL contiene dos letras (abreviatura del término en inglés) que indican la naturaleza de la anotación del campo correspondiente (por ejemplo ID es el identificador, KW es la palabra clave, etc.). El formato EMBL empieza con un identificador (ID) de la secuencia seguido por anotaciones como pueden ser el número de acceso (AC), fechas de creación y actualización (DT), descripción (DE), palabras clave (KW), organismo o especie de origen (OS), clasificación de la especie (OC), datos sobre la referencia bibliográfica (páginas (RP), Autores (RA), título del trabajo (RT), Revista, volumen, año y páginas de la publicación (RL), o comentarios (CC). Las letras FT marcan otras características de la secuencia como puede ser su traducción, identificador de la proteína, etc. El comienzo de la secuencia está marcado con las letras SQ y su fin con el símbolo “//”. Cada línea de la secuencia contiene sesenta residuos se parados de diez en diez por un espacio. La línea termina con una tabulación y la posición del último residuo de la correspondiente línea. El formato GenBank tiene una estructura similar a la del formato EMBL con las siguientes diferencias: la primera línea del fichero empieza con la palabra “LOCUS” y contiene información sobre la secuencia (número de acceso, nombre, etc.), en lugar de utilizar abreviaturas en la primera columna, como en EMBL, el formato GenBank utiliza una palabra completa descriptiva de la anotación del campo (de esta forma el formato GenBank es más intuitivo que el EMBL). El comienzo de la secuencia está marcado con la palabra “ORIGIN” y, como en EMBL, su fin por el símbolo “//”. Igual que en el formato EMBL, cada línea de la secuencia GenBank contiene sesenta residuos separados de diez en diez por un espacio. En

el caso GenBank, sin embargo, la línea comienza con un número que marca la posición del primer residuo de la correspondiente línea (en EMBL es el último residuo que está marcado).

**Ejemplo de formato EMBL** (los puntos en negrita indican lo explicado anteriormente):

```
ID   X15332; SV 1; linear; mRNA; STD; HUM; 3234 BP.
XX
AC    X15332;
XX
DT    06-JUL-1989 (Rel. 20, Created)
DT    05-AUG-1995 (Rel. 44, Last updated, Version 2)
XX
DE    Human COL3A1 mRNA for pro alpha-1 (III) collagen
XX
KW    COL3A1 gene; collagen.
XX
OS    Homo sapiens (human)
OC    Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
      Mammalia;
OC    Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini;
      Hominidae;
OC    Homo.
XX
RN    [1]
RP    1-3234
RA    Janeczko R., Ramirez F.;
RT    ;
RL    Submitted (19-MAY-1989) to the EMBL/GenBank/DDBJ databases.
RL    Janeczko R., Ramirez F., Suny Health Science Centre, 450 Clarkson
      Avenue-RL Box 44, Brooklyn NY 11203, U S A.
XX
RN    [2]
RX    DOI; 10.1093/nar/17.16.6742
RX    PUBMED; 2780304.
RA    Janeczko R.A., Ramirez F.;
RT    "Nucleotide and amino acid sequences of the entire human alpha 1 (III)
RT    collagen";
RL    Nucleic Acids Res. 17(16):6742-6742(1989).
XX
DR    GDB; 174873.
DR    H-InvDB; HIT000321499.
XX
CC    The sequence overlaps with that reported by Chu et. al. in
CC    J. Biol. Chem. 260:4357-4363(1985), by Toman et. al. in
CC    Nucl. Acids Res. 16:7201-7201(1988) and by Mankoo et. al. in
CC    Nucl. Acids Res. 16:2337-2337(1988).
XX
FH    Key Location/Qualifiers
FH
FT    source 1..3234
FT    /organism="Homo sapiens"
FT    /map="2q31"
FT    /mol_type="mRNA"
FT    /db_xref="taxon:9606"
FT    CDS <1..>3234
FT    /codon_start=1
FT    /product="alpha-1 (III) collagen"
FT    /protein_id="CAA33387.1"
FT    /translation="QNYSPQYDSYDVKSGGVAVGGLAGYPGPAGPPGPPGPPGTSGHPG
FT    SPGSPGYQGPPGEPGQAGPSGPPGPPGAIGPSGPAGKDGESGRPGRPGDRGLPGPPGIK
FT    GPAGIPGFPGMKGHRGFDGRNGEKGETGAPGLKGENGLPGENGAPGPMGPRGAPGERGR
FT    PGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEVGPAGSPGSNGAPGQRG
FT    EPGPQGHAGAQGPPGPPGINGSPGGKGEMGPAGIPGAPGLMGARGPPGPAGANGAPGLR
FT    GGAGEPGKNGAKGEPGPRGERGEAGIPGVPGAKGEDGKDGSPGDPGANGLPGAAGERGA
FT    ..........CCGGVGAPAIAGIGAEKAGGFAPYYGDEPM"
XX
SQ    Sequence 3234 BP; 664 A; 861 C; 1106 G; 603 T; 0 other;
cagaactatt ctccccagta tgattcatat gatgtcaagt cgggcggagt agcagtagga   60
ggactcgcag gctatcctgg accagctggc cccccaggcc cccccggccc ccctggtaca  120
tctggtcatc ctggttcccc tggatctcca ggataccaag gacccctgg tgaacctggg  180
caagctggtc cttcaggccc tccaggacct cctggtgcta taggtccatc tggtcctgct  240
ggaaaagatg gagaatcagg tagacccgga cgacctggag accgaggatt gcctggacct  300
ccaggtatca aaggtccagc tgggatacct ggattccctg gtatgaaagg acacagaggc  360
ttcgatggac gaaatggaga aaagggtgaa acaggtgctc ctggattaaa gggtgaaaat  420
..... attg gagctgaaaa agctggcggt tttgcccctt attatggaga tgaaccaatg 3234
//
```

**Ejemplo de formato GenBank** (los puntos en negrita indican lo explicado anteriormente):

```
LOCUS       NM_000090         5490 bp     mRNA     linear     PRI 29-JAN-2011
DEFINITION  Homo sapiens collagen, type III, alpha 1 (COL3A1), mRNA.
ACCESSION   NM_000090
VERSION     NM_000090.3 GI:110224482
KEYWORDS    .
SOURCE      Homo sapiens (human)
ORGANISM    Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires;
            Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE   1 (bases 1 to 5490)
AUTHORS     Kronenberg D, Bruns BC, Moali C, Vadon-Le Goff S, Sterchi EE, Traupe H, Bohm M, Hulmes DJ, Stocker W, Becker-Pauly C
TITLE       Processing of procollagen III by meprins: new players in extracellular matrix assembly?
JOURNAL     J. Invest. Dermatol. 130 (12), 2727-2735 (2010)
PUBMED      20631730
REMARK       GeneRIF: meprins could be important players in several remodeling processes involving collagen fiber deposition
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence was derived from
            BP374999.1, BC028178.1, X14420.1 and AC066694.7.
FEATURES    Location/Qualifiers
source      1..5490
            /organism="Homo sapiens"
            /mol_type="mRNA"
            /db_xref="taxon:9606"
            /chromosome="2"
            /map="2q31"
gene        1..5490
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /note="collagen, type III, alpha 1"
            /db_xref="GeneID:1281"
            /db_xref="HGNC:2201"
            /db_xref="HPRD:00365"
            /db_xref="MIM:120180"
exon        1..196
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /inference="alignment:Splign"
            /number=1
CDS         118..4518
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /note="Ehlers-Danlos syndrome type IV, autosomal dominant;
            collagen, fetal; collagen alpha-1(III) chain; alpha1 (III)
            collagen"
            /codon_start=1
            /product="collagen alpha-1(III) chain preproprotein"
            /protein_id="NP_000081.1"
            /db_xref="GI:4502951"
            /db_xref="CCDS:CCDS2297.1"
            /db_xref="GeneID:1281"
            /db_xref="HGNC:2201"
            /db_xref="HPRD:00365"
            /db_xref="MIM:120180"
            /translation= "MMSFVQKGSWLLLALLHPTIILAQQEAVEGGCSHLGQSYADRDVWKPEP
             CQICVCDSGSVLCDDIICDDQELDCPNPEIPFGECCAVCPQPPTAPTRPPNGQKGDPGPPGIP
             GRNGDPGIPGQPGSPGSPGPPGICESCPTGPQNYSPQYDSYDVKSGVAVGGLAGYPGPAGPPG
             PPGPPGTSGHPGSPGSPGYQGPPGEPGQAGPSGPPGPPGAIGPSGPAGKDGESGRPGRPGERG
             LPGPPGIKGPAGIPGFPGMKGHRGFDGRNGEKGETGAPGLKGENGLPGENGAPGPMGPRGAPG
             ERGRPGLPGAAGARGNDGARGSDGQPGP....VRLPIVDIAPYDIGGPDQEFGVDVGPVCFL"
sig_peptide  118..186
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
proprotein   187..4515
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /product="collagen alpha-1(III) chain proprotein"
mat_peptide  577..3780
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /product="collagen alpha-1(III) chain"
STS         2303..2528
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /standard_name="GDB:178411"
            /db_xref="UniSTS:155007"
exon        2347..2400
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /inference="alignment:Splign"
            /number=32
            /gene_synonym="EDS4A; FLJ34534"
STS         5334..5460
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /standard_name="WI-16343"
            /db_xref="UniSTS:68589"
STS          5359..5419
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /standard_name="COL3A1"
            /db_xref="UniSTS:480020"
polyA_signal  5468..5473
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
polyA_signal  5481..5486
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
polyA_site  5490
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
ORIGIN
          1 ggctgagttt tatgacgggc ccggtgctga agggcaggga acaacttgat ggtgctactt
         61 tgaactgctt ttcttttctc ctttttgcac aaagagtctc atgtctgata tttagacatg
        121 atgagctttg tgcaaaaggg gagctggcta cttctcgctc tgcttcatcc cactattatt
        181 ttggcacaac aggaagctgt tgaaggagga tgttcccatc ttggtcagtc ctatgcggat
        241 agagatgtct ggaagccaga accatgccaa atatgtgtct gtgactcagg atccgttctc
       5461 .......... cacccataat aaaatatcat attaaaattc
//
```

### *Genome Browser*

Entre los algoritmos y utilidades que una base de datos de secuencias puede ofrecer está un visor que permite tener información sobre la secuencia que nos interesa teniendo en cuenta su localización cromosómica. Por consiguiente solo es posible utilizar en caso de secuencias procedentes de organismos cuyos genomas están parcialmente o completamente secuenciados ensamblados y anotados. Se trata de la herramienta llamada *Genome Browser*. Como su nombre indica, se trata de una herramienta que permite al investigador navegar en el genoma (cada cromosoma aparte). Dicha navegación no solo es posible en dirección horizontal (es decir ver qué secuencias lindan con nuestra secuencia o locus de interés) sino que también lo es en dirección vertical (zoom) permitiendo así el movimiento entre varios niveles de enfoque desde el citogenético (por ejemplo para ver la información de bandeo cromosómico en la región) hasta la secuencia propiamente dicha y sus características (promotor, sitio de unión de factores de transcripción…). Además el *Genome Browser* permite incluir todo tipo de información y anotaciones sobre cada secuencia del cromosoma. De esta forma, si nos dirigimos al *Genome Browser* para el genoma humano y buscamos el gen de colágeno que hemos utilizado de ejemplo antes (COL3A1) veremos que efectivamente (Figura 4) se encuentra en locus comprendido entre los nucleótidos 189833342 y 189883227 del ensamblaje del cromosoma 2 humano; zona que corresponde a la banda citológica (cromosómica) 32 del brazo largo de dicho cromosoma. Veremos que además de la información sobre en qué cromosoma, brazo, y región se encuentra nuestra secuencia, *Genome Browser* también nos ofrece información sobre su naturaleza (gen, promotor, intrón, etc.), datos de expresión, variantes (incluidos SNPs), datos sobre la función, interacciones génicas, datos estructurales de la proteína, los ortólogos de la secuencia, sus relaciones filogenéticas, bibliografía, etc. Todo tipo de anotación (información) disponible sobre esa secuencia. Todo esto hace que *Genome Browser* sea la herramienta más informativa en caso de secuencias de organismos con genomas secuenciados y ensamblados (aunque parcialmente).

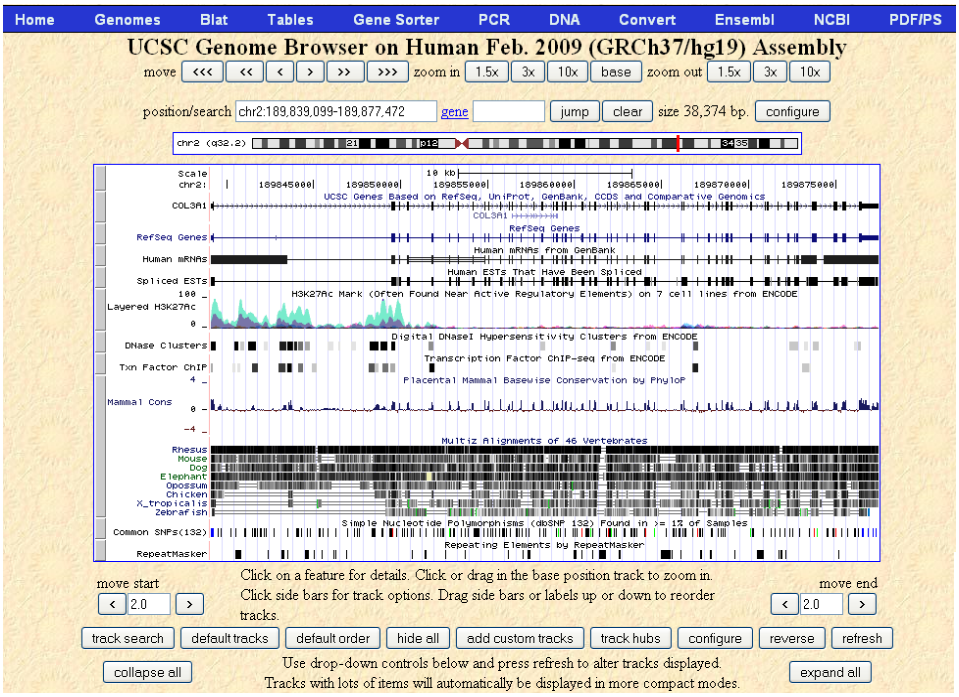



**Figura 4.** Captura de pantalla mostrando una parte del resultado de la búsqueda de la secuencia del gen humano Col3A1 en el *Genome Browser*

# 3.3. METODOLOGÍA

### 3.3.1. Introducción a la Bioinformática

Las técnicas de análisis genético han sufrido una evolución muy rápida en los últimos años, habiendo pasado de ser manuales, lentas, costosas y producir relativamente poca información, a ser automáticas, cada vez más rápidas y baratas y producir cantidades enormes de información. Con las tecnologías de secuenciación masiva, por ejemplo, se pueden obtener secuencias de genomas completos en poco tiempo. El almacenamiento, tratamiento y análisis de toda esa información, requieren la utilización de herramientas computacionales rápidas y potentes. La bioinformática es la disciplina encargada de elaborar las herramientas necesarias para ello (perfil de desarrollo o programación), así como la utilización de esas herramientas para llevar a cabo los análisis que, al final, derivan en conocimiento biológico.

Las herramientas bioinformáticas pueden clasificarse como herramientas de almacenamiento y recuperación de la información (bases de datos) y programas de manipulación y análisis de dicha información. Desde el punto de vista de la Genética, las primeras son fundamentalmente las bases de datos de secuencias de ADN y proteínas, mutaciones, expresión, regulación, metilación, etc., y van a ser el objeto de trabajo de esta sesión práctica, mientras que algunas de las segundas (predicción computacional de genes, alineamiento múltiple y reconstrucciones filogenéticas, análisis computacional de expresión génica diferencial) se trabajarán en las prácticas siguientes.

### 3.3.2. Bases de datos de secuencias de ADN y proteínas

Las bases de datos son sistemas de almacenamiento estructurado de la información que permiten la localización y recuperación de los datos de interés de forma rápida, sencilla y eficiente, entre cantidades enormes de datos, mediante un programa llamado motor de la base de datos. En esta práctica vamos a ver algunas de las bases de datos de secuencias de ADN y proteínas de uso más extendido.

#### 3.3.2.1. GenBank

GenBank (http://www.ncbi.nlm.nih.gov/genbank/) es la base de datos de secuencias genéticas de los *National Institutes of Health* (NIH) de Estados Unidos, una colección anotada de todas las secuencias de ADN disponibles públicamente (Nucleic Acids Research, 2008 Jan; 36 (Database issue): D25-30) (Figura 5). La base de datos está alojada en los servidores del Centro Nacional Para la Información Biotecnológica (*The National Center for Biotechnology Information*) en Estados Unidos.

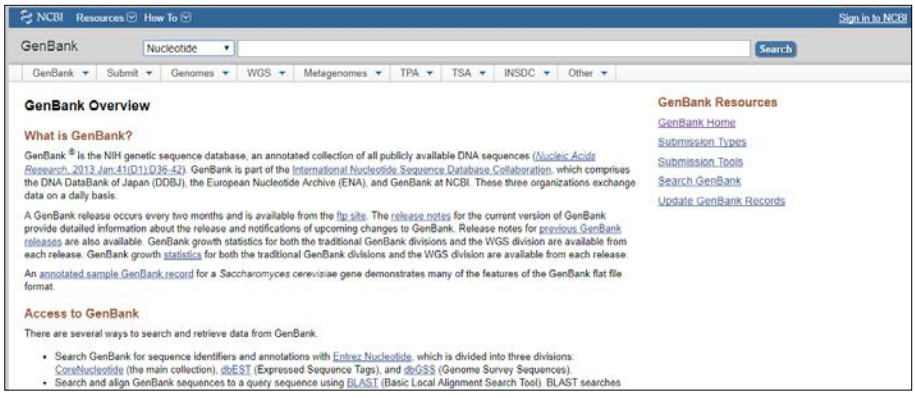



Figura 5: Página web de acceso a GenBank

El menú desplegable permite escoger la base de datos a utilizar (Figura 6), junto con una caja de texto seguida de un botón *Search*.

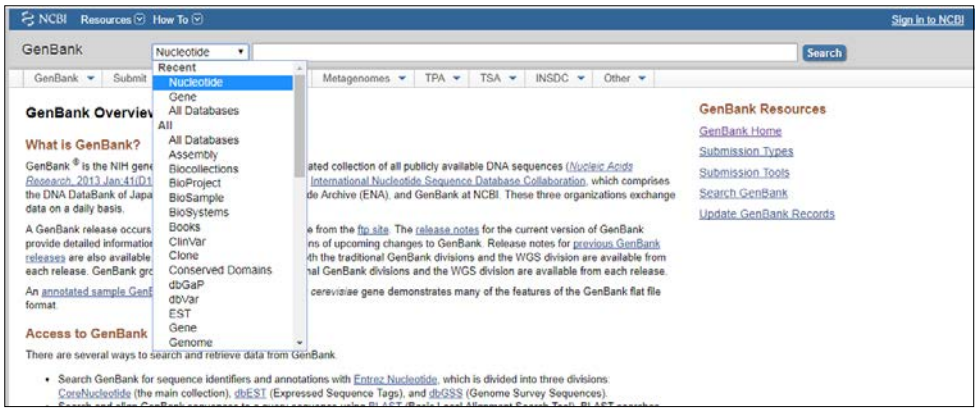



Figura 6: Realizando una búsqueda en GenBank

Para realizar una búsqueda, seleccionamos primero la base de datos a utilizar (*Nucleotide* para ADN, *Protein* para proteínas, *PubMed* para bibliografía, etc.) y después introducimos una cadena de búsqueda en el cuadro de texto; finalmente, picamos en *Search*. Como ejemplo, si quisiéramos buscar la secuencia del gen que codifica para el factor de coagulación VIII humano, escogeríamos la base de datos de nucleótidos y escribiríamos *Homo sapiens coagulation factor VIII gene* en la caja de texto. El resultado de esa búsqueda se muestra en la Figura 7.

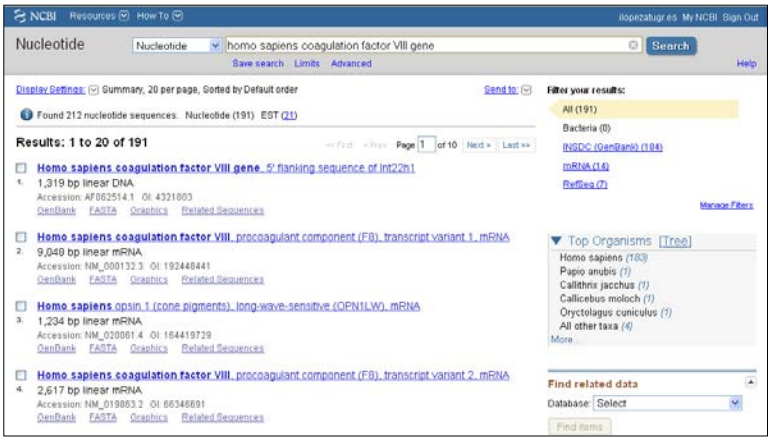



Figura 7. Resultado de una búsqueda en GenBank

Como se observa, en el momento en el momento en el que se realizó, se obtuvieron coincidencias de la cadena de búsqueda con 191 registros de la base de datos. Para cada uno de los resultados, se muestra el nombre de la secuencia enlazado (en azul y subrayado) al registro en la base de datos, el tipo (ADN o ARN) de secuencia y su longitud, el número de acceso (*Accession number*) del registro en la base de datos (que lo identifica de forma única, y enlaces a la secuencia en los formatos GenBank y FASTA, así como a un navegador gráfico de secuencias y un listado de secuencias relacionadas.

Picando en el enlace con el nombre de la secuencia accedemos a la información almacenada en el registro correspondiente, que está estructurada en diferentes campos de información (Figura 8).

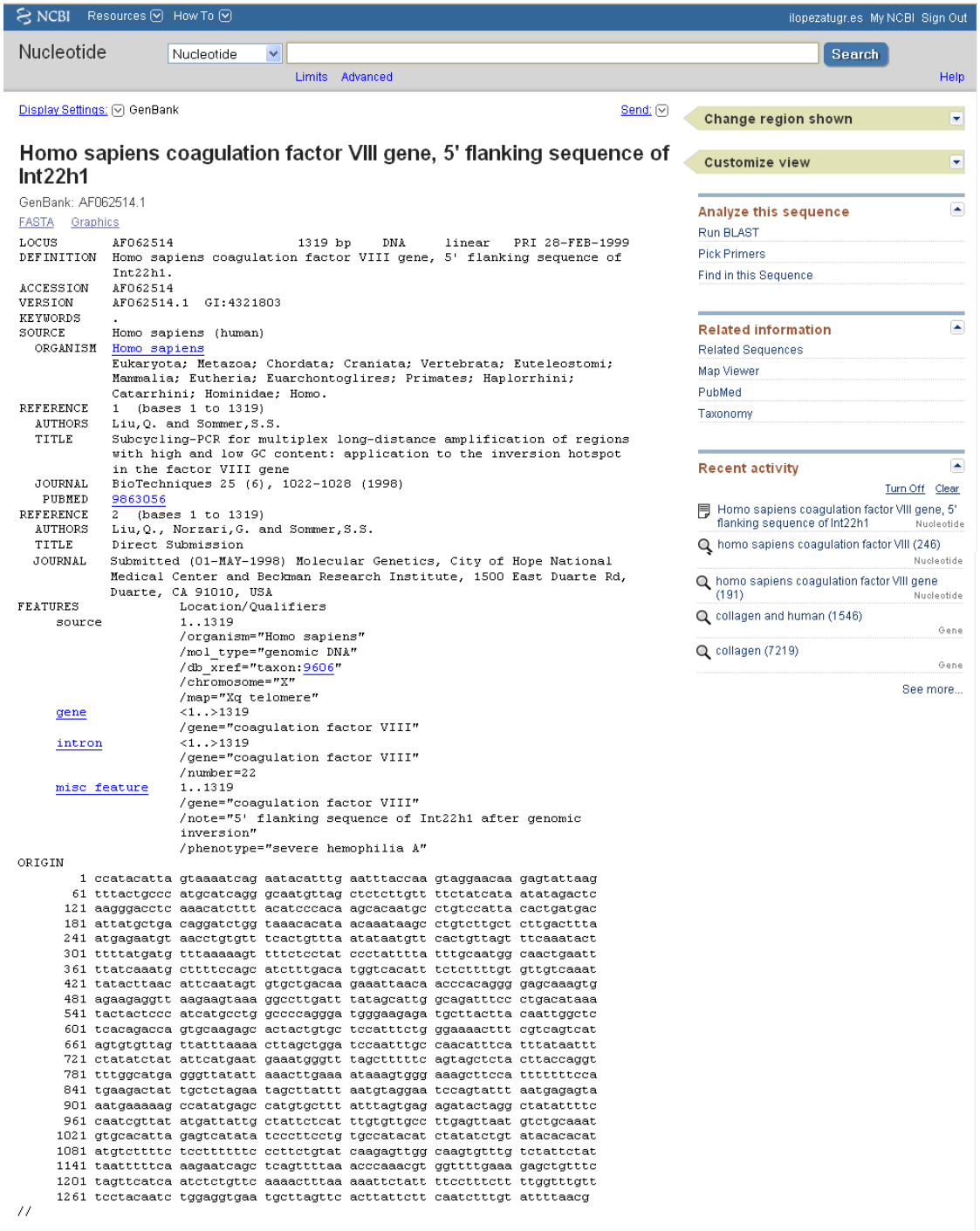



Figura 8: Registro de una secuencia de ADN almacenada en formato GenBank

Algunos campos relevantes del formato GenBank son los siguientes:

***Locus*** Contiene un identificador (no necesariamente único) de la secuencia, así como su longitud (1319 pares de bases en el ejemplo), el tipo de secuencia (ADN lineal) y la fecha de su publicación en la base de datos.

***Definition*** Contiene información más detallada acerca de la secuencia almacenada en ese registro.

***Accession*** Es el número de acceso de la secuencia en la base de datos, que la identifica de forma inequívoca.

***Source*** Son campos que contienen información acerca del origen de la secuencia almacenada, la especie a la que pertenece y su clasificación taxonómica.

***Reference*** Son campos que contienen referencias bibliográficas sobre la secuencia, su publicación en revistas o bases de datos científicas, etc.

***Features*** Contienen la anotación de la secuencia, que describe qué está contenido concretamente en las distintas posiciones de la secuencia. En el caso del ejemplo en la figura 8, el gen que codifica para el factor VIII de coagulación en humanos.

***Origin*** Es el último campo del registro, que almacena la secuencia de nucleótidos. El final de registro viene marcado por los caracteres "//" situados en una línea nueva.

Se puede acceder a la secuencia en formato FASTA (picando en el enlace correspondiente en la parte superior izquierda de la página) (Figura 9).

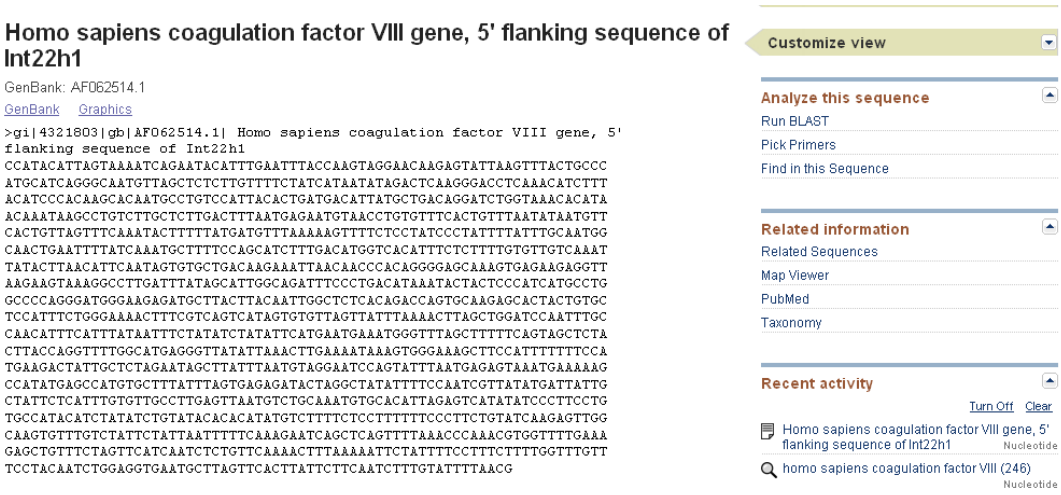



Figura 9: Secuencia del factor de coagulación VIII en formato FASTA

La búsqueda de una secuencia de aminoácidos se realiza en GenBank de forma análoga, escogiendo la base de datos de proteínas en el menú desplegable y tecleando la cadena de búsqueda en la caja de texto. Se puede ver un ejemplo en la Figura 10, que muestra el registro correspondiente a la proteína codificada por el gen del ejemplo anterior, es decir, el factor VIII de coagulación en el hombre.

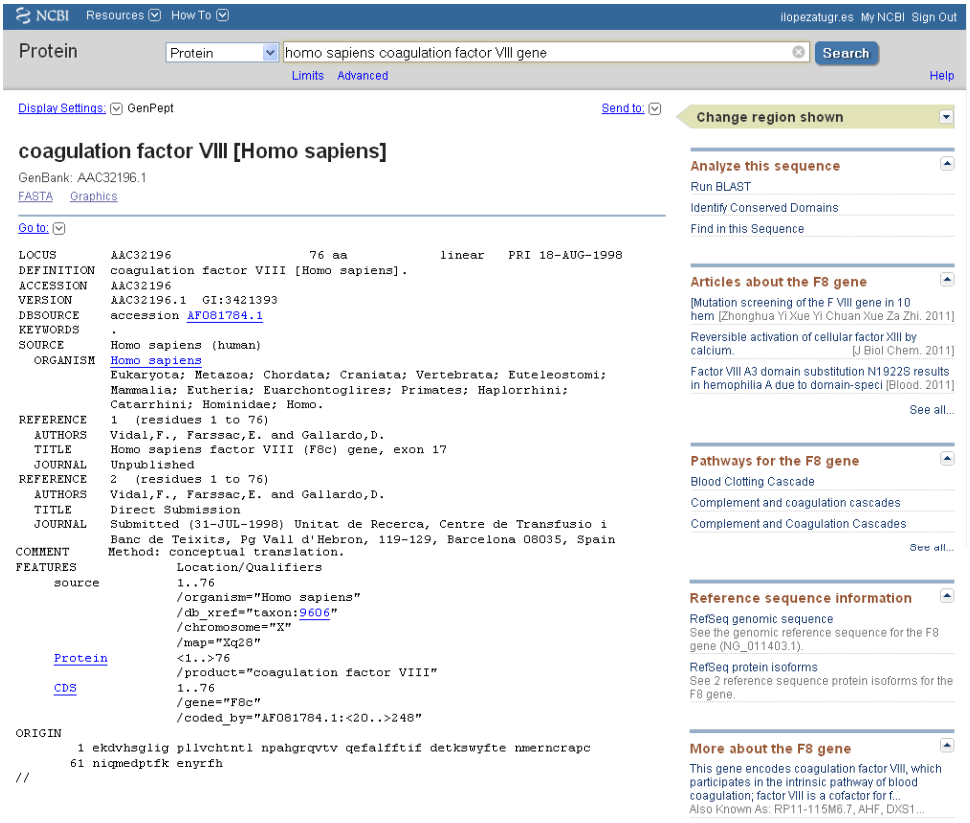



Figura 10: Registro de una secuencia de proteína almacenada en formato GenBank

### 3.3.2.2. EMBL

La base de datos de secuencias EMBL pertenece al Laboratorio Europeo de Biología Molecular (*European Molecular Biology Laboratory*, EMBL), y se encuentra alojada en los servidores del Instituto Europeo de Bioinformática (*European Bioinformatics Institute*, EBI, http://www.ebi.ac.uk/). Igual que GenBank, EMBL contiene bases de datos de secuencias de ADN y proteínas, de estructura, expresión, genomas completos, literatura científica, etc. (Figura 11).

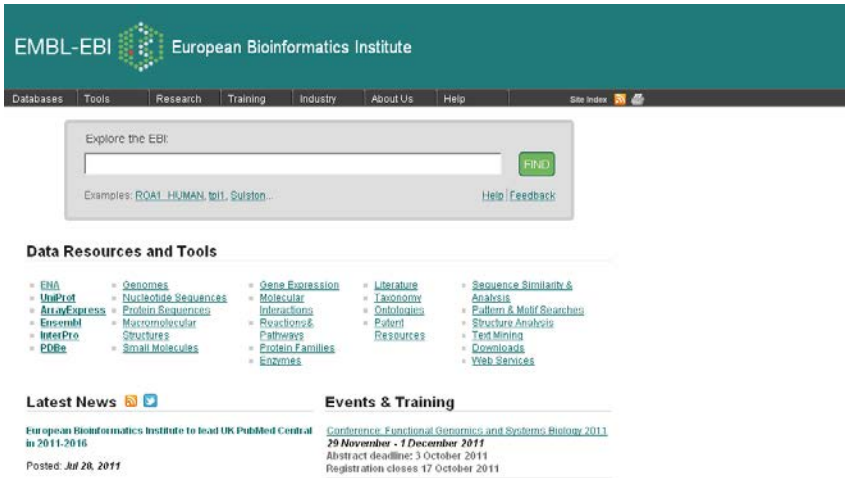



Figura 11: Página de acceso a EMBL

La utilización de las bases de datos de EMBL y el formato de almacenamiento de sus secuencias (Figura 12) son muy similares a los de GenBank.

```
ID   M88628; SV 1; linear; genomic DNA; STD; HUM; 1493 BP.
XX
AC   M88628;
XX
DT   07-AUG-1992 (Rel. 33, Created)
DT   19-OCT-2008 (Rel. 97, Last updated, Version 6)
XX
DE   Homo sapiens coagulation factor VIII (F8C) gene, exon 1.
XX
KW   coagulation factor; factor VIII.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-1493
RX   DOI; 10.1093/hmg/1.3.199.
RX   PUBMED; 1303178.
RA   Gitschier J., Wood W.I.;
RT   "Sequence of the exon-containing regions of the human factor VIII gene";
RL   Hum. Mol. Genet. 1(3):199-200(1992).
XX
DR   EPD; EP14077; HS_F8.
DR   Ensembl-Gn; ENSG00000185010; Homo_sapiens.
DR   Ensembl-Tr; ENST00000360256; Homo_sapiens.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1493
FT                   /organism="Homo sapiens"
FT                   /map="Xq28"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:9606"
FT   exon            1006..1318
FT                   /gene="F8C"
FT                   /number=1
FT                   /note="GOO-119-124"
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1493 BP; 439 A; 265 C; 379 G; 410 T; 0 other;
     gagctcacca tggctacatt ctgatgtaaa gagatatatc ctatacctgg gccaaatgta        60
     aacagcctgg aaaagtgtta ggttaaaaac aaaacaaaat aaataaatga ataaatgcca       120
     ggtggttatg agtgctattg agaaaaatga agccaagagg gatatcagtg atgcaggtgg       180
     gggtaaagag cttacaacat aaatgtggtg ttccatattt aaacctcatt caacagggaa       240
     gattggagct gaaatgtgaa ggagttgtgg gagtggaact acgtgggaaa tctgggggaa       300
     aggtgttttg ggtaaaagaa atagcaagtg ttgaggtcca ggggcatgag tgtgcttgat       360
     attttaggga agagtaagga gaccagtata accagagtga gatgagacta cagaggtcag       420
     gagaaagggc atgcagacca tgtgggatgc tctaggacct aggccatggt aaagatgtag       480
     ggttttaccc tgatggaggt cagaagccat tggaggattc tgagaagagg agtgacagga       540
     ctcgctttat agtttttaaat tataactata aattatagtt tttaaaacaa tagttgccta       600
     acctcatgtt atatgtaaaa ctacagtttt aaaaactata aattcctcat actggcagca       660
     gtgtgagggg caagggcaaa agcagagaga ctaacaggtt gctggttact cttgctagtg       720
     caagtgaatt ctagaatctt cgacaacatc cagaacttct cttgctgctg ccactcagga       780
     agagggttgg agtaggctag gaataggagc acaaattaaa gctcctgttc actttgactt       840
     ctccatccct ctcctccttt ccttaaaggt tctgattaaa gcagacttat gcccctactg       900
     ctctcagaag tgaatgggtt aagtttagca gcctcccttt tgctacttca gttcttcctg       960
     tggctgcttc ccactgataa aaaggaagca atcctatcgg ttactgctta gtgctgagca      1020
     atccagtggg taaagttcct taaaatgctc tgcaaagaaa ttgggacttt tcattaaatc      1080
     agaaatttta ctttttttccc ctcctgggag ctaaagatat tttagagaag aattaacctt      1140
     ttgcttctcc agttgaacat ttgtagcaat aagtcatgca aatagagctc tccacctgct      1200
     tctttctgtg ccttttgcga ttctgcttta gtgccaccag aagatactac ctgggtgcag      1260
     tggaactgtc atgggactat atgcaaagtg atctcggtga gctgcctgtg gacgcaaggt      1320
     aaaggcatgt cctgtagggt ctgatcgggg ccaggattgt ggggatgtaa gtctgcttgg      1380
     aggaaggtgc agacatcggg ttaggatggt tgtgatgcta cctgggcccc aaagaaacat      1440
     ttctgggtaa ggtgtgcaca catctgtgtt attagcagaa atgctaactg ccc              1493
//
```

Figura 12: Secuencia de ADN en formato EMBL

### 3.3.2.3. UniProt

UniProt (http://www.uniprot.org/uniprot/) es una de las bases de datos de proteínas más utilizadas (Figura 13). La consulta a la base de datos es similar a la de las bases de datos anteriores.


Figura 13: Página de acceso a UniProt

### 3.3.2.4. Bases de datos de genomas completos

Ya existen también bases de datos que almacenan genomas completos, como la *Genomes Pages* (http://www.ebi.ac.uk/genomes/) del EBI (Figura 14).

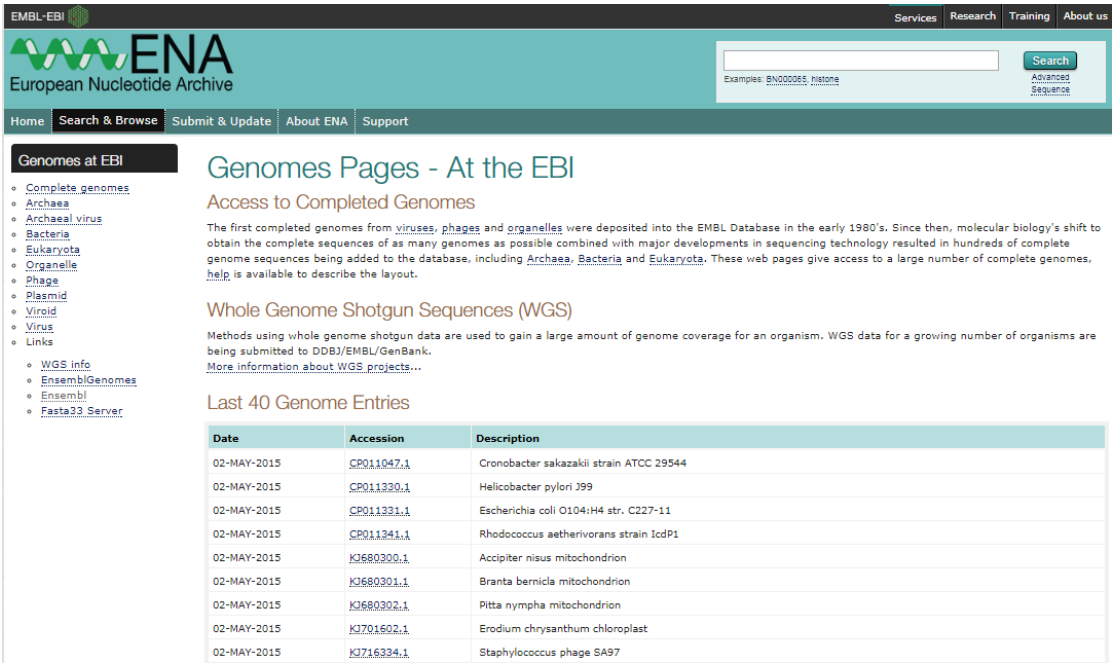



Figura 14: Página de acceso a *Genomes Pages*

### 3.3.2.5. Bases de datos de bibliografía científica

También existen bases de datos de bibliografía científica, alojadas en los servidores de los centros de investigación mencionados anteriormente, como Entrez, EMBL, PubMed, NCBI Bookshelf, etc.

### 3.3.3. Rastreo de bases de datos

Además de buscar secuencias de ADN o proteínas por su nombre, especie, etc., podemos estar interesados en buscar secuencias que presenten similitud (¿homología?) con una secuencia problema dada (ejemplo en la Figura 15), es decir, lo que se conoce como rastrear bases de datos.

```
MAVMAPRTLV LLLSGALALT QTWAGSHSMR YFSTSVSRPG RGEPRFIAVG YVDDTQFVRF
DSDAASQRME PRAPWIEQEG PEYWDRNTRN VKAHSQTDRV DLGTLRGYYN QSEDGSHTIQ
RMYGCDVGSD GRFLRGYQQD AYDGKDYIAL NEDLRSWTAA DMAAEITKRK WEAAHFAEQL
RAYLEGTCVE WLRRHLENGK ETLQRTDAPK THMTHHAVSD HEAILRCWAL SFYPAEITLT
WQRDGEDQTQ DTELVETRPA GDGTFQKWAA VVVPSGQEQR YTCHVQHEGL PEPLTLRWEP
SSQPTIPIVG IIAGLVLFGA VIAGAVVAAV RWRRKSSDRK GGSYSQAASS DSAQGSDVSL
TACKV
```

Figura 15: Secuencia de una proteína anónima

Uno de los algoritmos de rastreo de bases de datos más conocido es BLAST, implementado por los programas BLASTn (ADN) y BLASTp (proteínas) (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Vamos a rastrear las bases de datos de proteínas con la secuencia de ejemplo de la Figura 15 utilizando el programa BLASTp (Figura 16a y b). En el interfaz gráfico del programa, encontramos una caja de texto donde podemos pegar la secuencia problema, así como un botón *Seleccionar archivo* que nos permite escoger un fichero que contenga la secuencia problema en nuestro ordenador. Más abajo encontramos el botón *Blast* para la ejecución del rastreo.

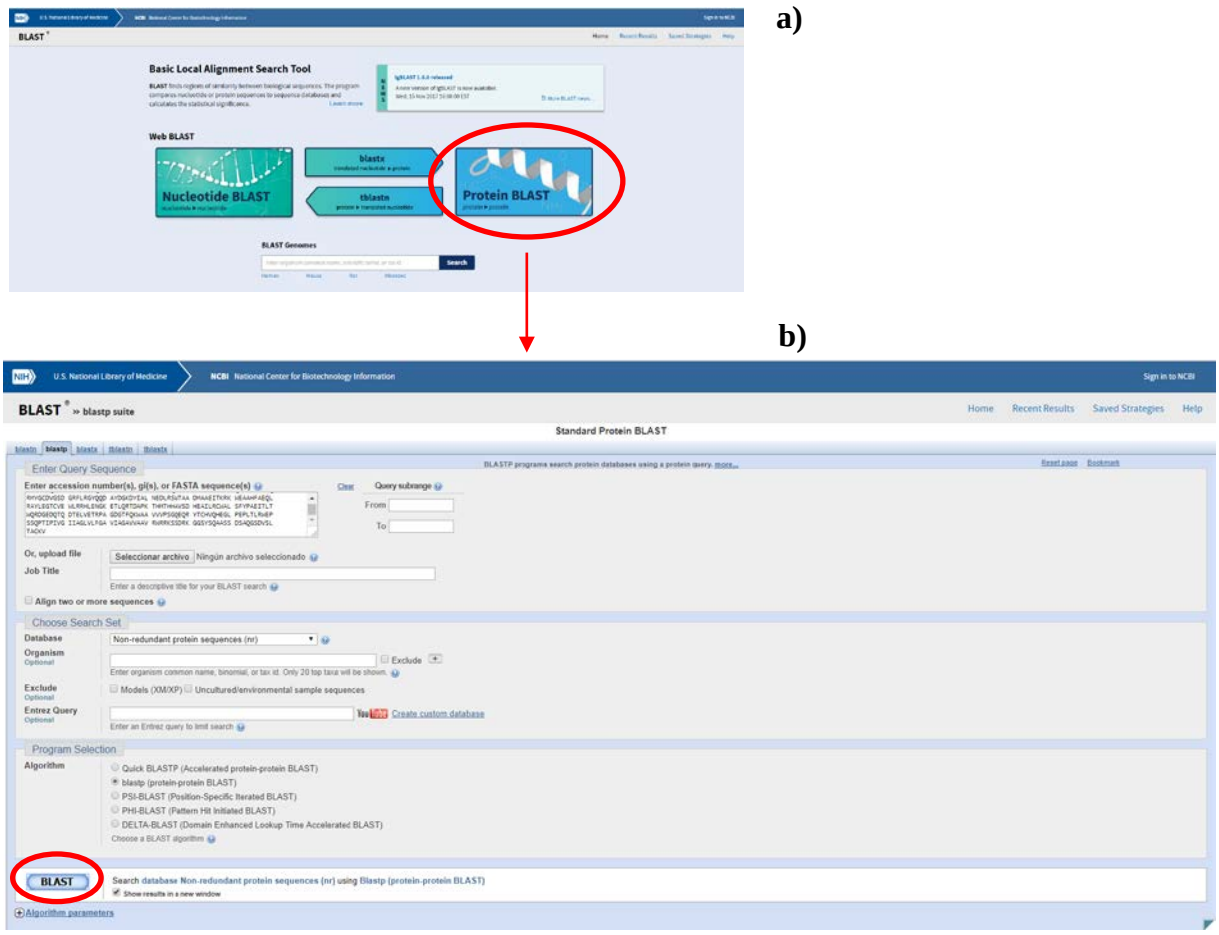



Figura 16: Página de acceso a BLASTp

El resultado de un rastreo con BLASTp tiene tres partes, un resumen gráfico interactivo (Figura 17), un resultado detallado en forma de tabla (Figura 18) y un listado de los alineamientos de las secuencias encontradas (Figura 19). Como puede comprobarse en todos ellos, la proteína problema era el antígeno de histocompatibilidad humano de clase I.

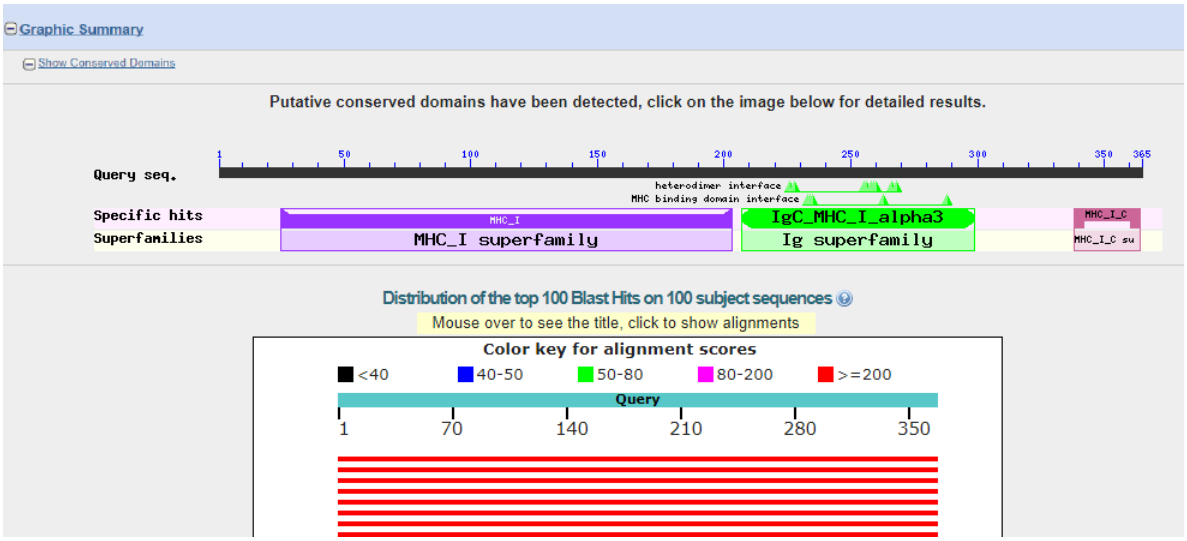
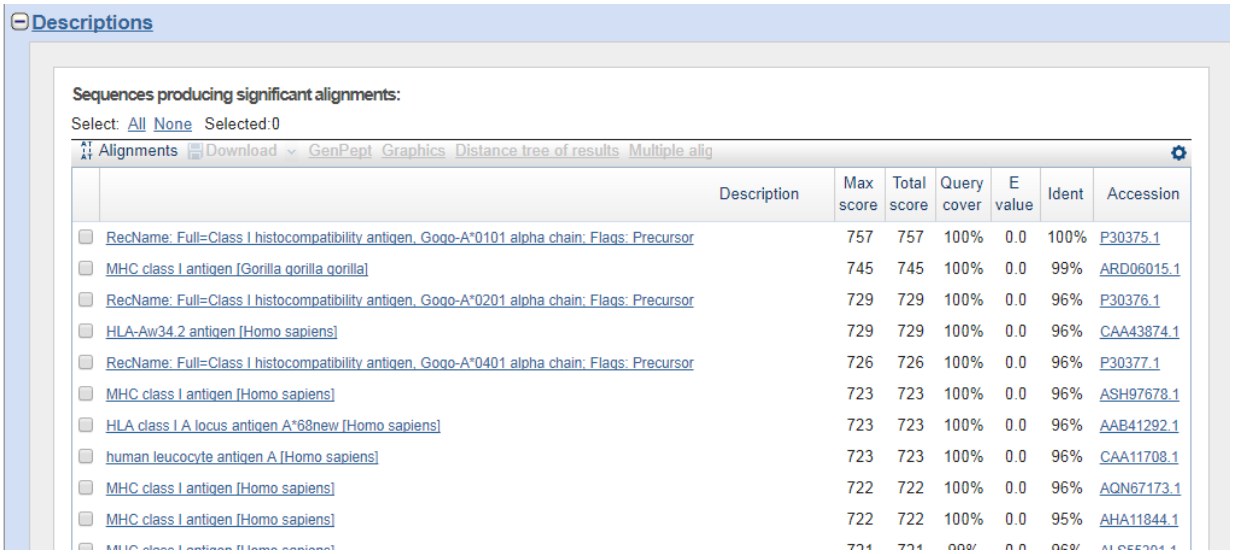



Figura 17: Resumen gráfico del resultado de un rastreo con BLASTp

La tabla que recoge los resultados (Figura 18) presenta en la primera columna el número de acceso de cada una de las secuencias de la base de datos que presentan similitud (encontradas mediante un algoritmo de alineamiento de secuencias) con la secuencia problema. El número de acceso es también un enlace al registro que almacena la secuencia en cada caso. La segunda columna contiene la descripción de la secuencia. Las siguientes presentan la puntuación del alineamiento, el porcentaje de superposición de las secuencias y, por último, el valor E de probabilidad, que representa la probabilidad de que la similitud entre la secuencia anónima problema y la encontrada en la base de datos sea al azar. Valores pequeños indican que el parecido no se debe al azar y, por tanto, las secuencias están relacionadas o, como en el caso de la primera secuencia obtenida (E = 0), son la misma secuencia.




| | Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|---|
| ☐ | RecName: Full=Class I histocompatibility antigen, Gogo-A*0101 alpha chain; Flags: Precursor | 757 | 757 | 100% | 0.0 | 100% | P30375.1 |
| ☐ | MHC class I antigen [Gorilla gorilla gorilla] | 745 | 745 | 100% | 0.0 | 99% | ARD06015.1 |
| ☐ | RecName: Full=Class I histocompatibility antigen, Gogo-A*0201 alpha chain; Flags: Precursor | 729 | 729 | 100% | 0.0 | 96% | P30376.1 |
| ☐ | HLA-Aw34.2 antigen [Homo sapiens] | 729 | 729 | 100% | 0.0 | 96% | CAA43874.1 |
| ☐ | RecName: Full=Class I histocompatibility antigen, Gogo-A*0401 alpha chain; Flags: Precursor | 726 | 726 | 100% | 0.0 | 96% | P30377.1 |
| ☐ | MHC class I antigen [Homo sapiens] | 723 | 723 | 100% | 0.0 | 96% | ASH97678.1 |
| ☐ | HLA class I A locus antigen A*68new [Homo sapiens] | 723 | 723 | 100% | 0.0 | 96% | AAB41292.1 |
| ☐ | human leucocyte antigen A [Homo sapiens] | 723 | 723 | 100% | 0.0 | 96% | CAA11708.1 |
| ☐ | MHC class I antigen [Homo sapiens] | 722 | 722 | 100% | 0.0 | 96% | AQN67173.1 |
| ☐ | MHC class I antigen [Homo sapiens] | 722 | 722 | 100% | 0.0 | 95% | AHA11844.1 |

Figura 18: Resultado detallado de un rastreo con BLASTp

Finalmente, aparecen los alineamientos de la secuencia problema con cada una de las secuencias obtenidas de la base de datos (Figura 19), en los que se pueden observar las secuencias completas y, entre ambas, la secuencia consenso. Es fácil observar las coincidencias y diferencias entre las secuencias alineadas.

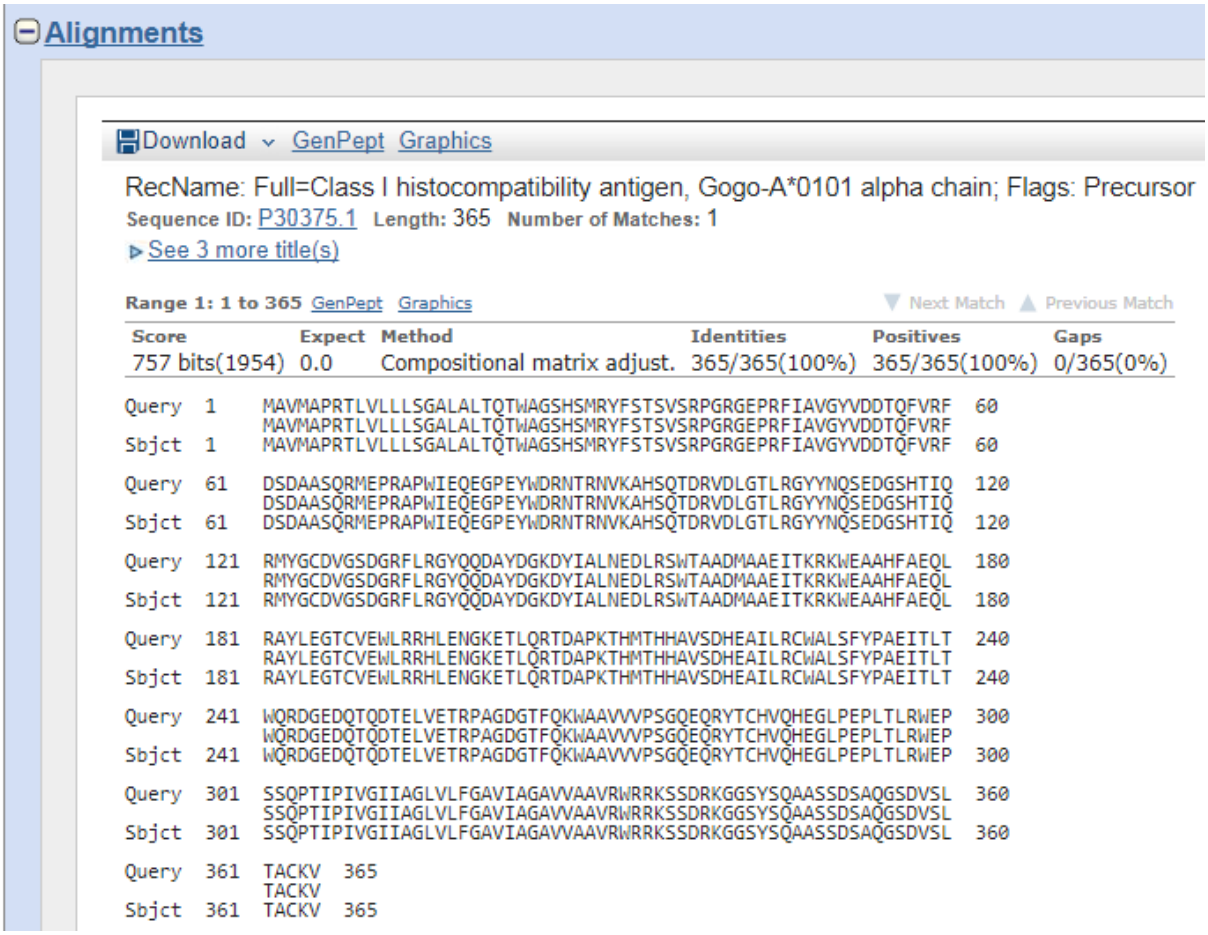



Figura 19: Alineamiento en un rastreo con BLASTp

## BLAT

Este programa funciona similar al BLAST (de hecho significa Blast Like Alignment Tool) y sirve para encontrar una secuencia de entrada dentro de un genoma. BLAT reporta todas las posiciones (cromosoma, inicio y final) del genoma donde el alineamiento de la secuencia de entrada da una puntuación (score) y similitud por encima de los umbrales mínimos. La salida del BLAT y su interpretación se explica en la siguiente figura:

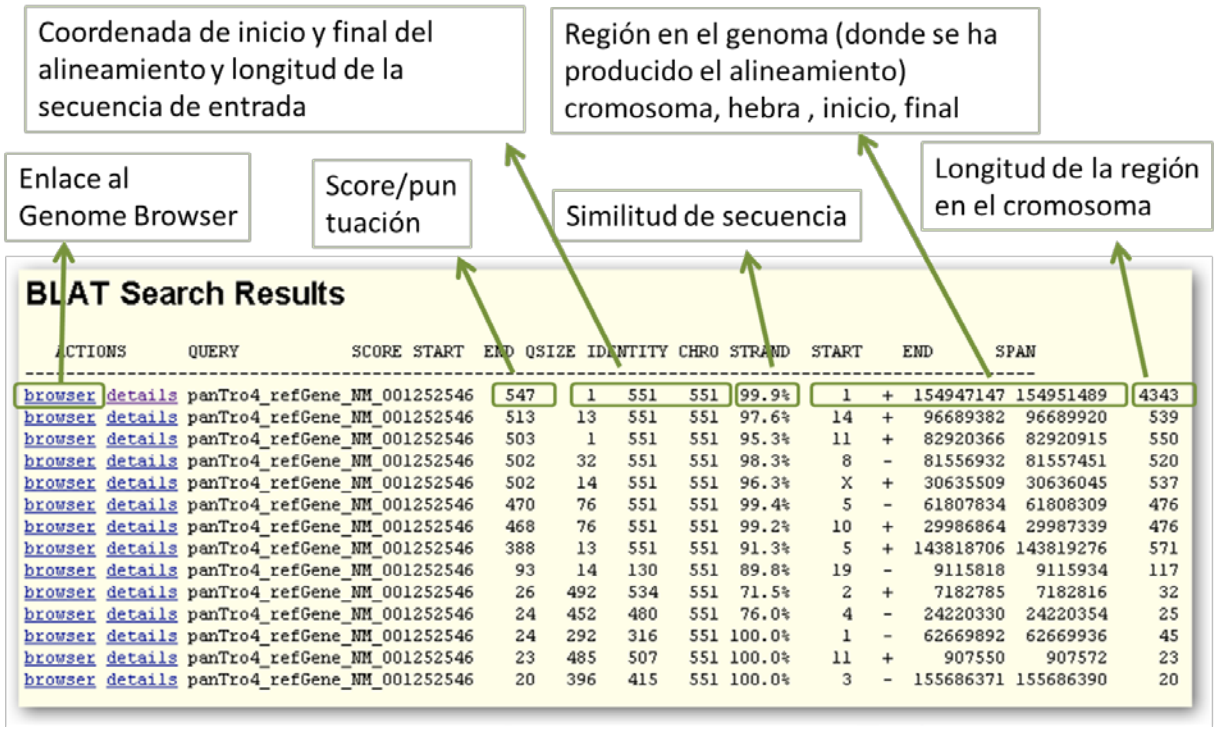



Figura 20: Rastreo de una secuencia en un genoma con BLAT

### 3.3.4. Navegadores genómicos

Un navegador genómico (*genome browser*) es una representación gráfica de un genoma, como puede deducirse de lo comentado anteriormente. Existen navegadores genómicos diferentes, pero todos ellos permiten visualizar las anotaciones y otras características genómicas. En general, los navegadores genómicos son aplicaciones informáticas que pueden ser independientes u operar a través de internet, y que permiten acceder a gran cantidad de información sobre los genomas, como por ejemplo, identificar secuencias de ADN correspondientes a genes concretos dentro de un genoma completo (al cual se accede a través de una base de datos determinada), identificar elementos funcionales, llevar a cabo comparación entre especies, etc.

Algunos de los navegadores genómicos más utilizados son:

*Apollo Genome Annotation Curation Tool* (http://apollo.berkeleybop.org/current/index.html)
Este navegador genómico ofrece muchas posibilidades, incluyendo la capacidad de realizar anotaciones. Está basado en Java, por lo que puede utilizarse en Windows, Mac OS X, o cualquier sistema operativo basado en Unix.

*Generic Genome Browser (GBrowse)* (http://www.gmod.org/wiki/GBrowse)
Desarrollado por GMOD (http://www.gmod.org/wiki/Main_Page), permite a los usuarios configurar rápidamente un navegador según sus necesidades.

UCSC *Genome Browser* (http://genome.ucsc.edu/cgi-bin/hgGateway)
Desarrollado por el Genome Bioinformatics Group of UC Santa Cruz (Universidad de California), proporciona diferentes genomas para analizar.

*Ensembl* (http://www.ensembl.org/index.html)
*Ensembl* es un proyecto conjunto entre el EMBL-EBI y el *Wellcome Trust Sanger Institute*, y facilita el acceso a diferentes genomas eucariotas para analizar.

## 3.4. EJERCICIOS Y CUESTIONES

A continuación, se plantean ejemplos y ejercicios basados en los algoritmos BLAST y BLAT:

- Dada una secuencia problema de aminoácidos, buscar proteínas diferentes que presenten similitud (homología) con ella.
- Dada una secuencia de ADN anónima, determinar el tipo de secuencia de que se trata mediante rastreo de una base de datos.
- Localizar una secuencia concreta en el genoma humano y resumir la información que se extraiga del análisis.

**BLAST:**

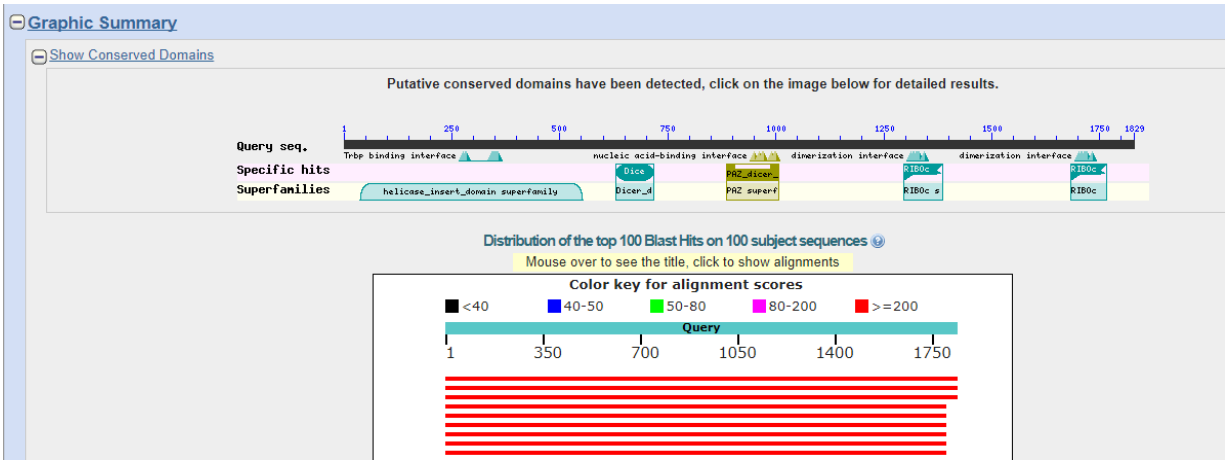
**Ejemplo: Secuencia de la proteína de Dicer(isoforma 2)**

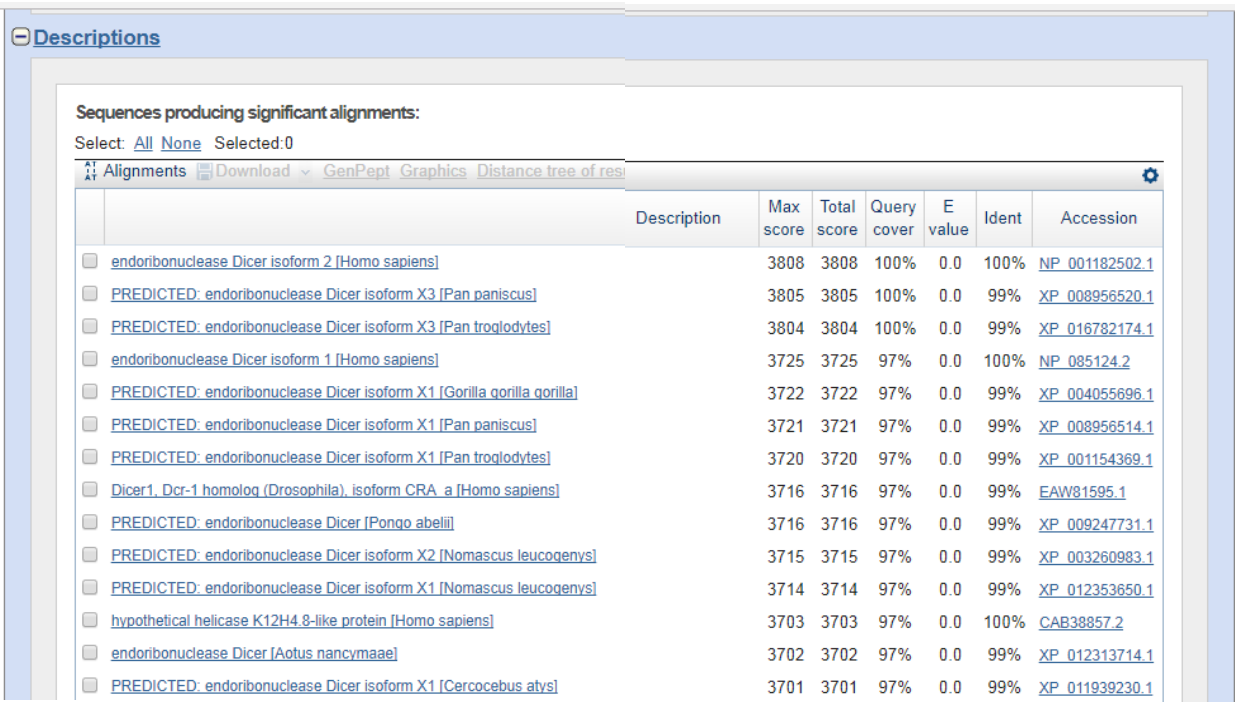
Ir a la página de BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Antes de empezar el análisis:

1. ¿ De qué tipo de secuencia se trata?
2. ¿Qué 'tipo' de BLAST tenemos que usar?
3. ¿Qué diferencias hay entre BLASTn y BLASTp?
4. ¿Qué base de datos usaremos?

Elegir el programa de BLAST a utilizar (BLASTn o BLASTp), pegar la secuencia de Dicer (isoforma2) en la caja de texto y rastrear las bases de datos haciendo click en BLAST al fondo de la página a la izquierda (ver captura de pantalla en la página 57 de este guión). Las siguientes imágenes corresponden a capturas de pantalla de los resultados:





Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of res

| | Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|---|
| endoribonuclease Dicer isoform 2 [Homo sapiens] | | 3808 | 3808 | 100% | 0.0 | 100% | NP_001182502.1 |
| PREDICTED: endoribonuclease Dicer isoform X3 [Pan paniscus] | | 3805 | 3805 | 100% | 0.0 | 99% | XP_008956520.1 |
| PREDICTED: endoribonuclease Dicer isoform X3 [Pan troglodytes] | | 3804 | 3804 | 100% | 0.0 | 99% | XP_016782174.1 |
| endoribonuclease Dicer isoform 1 [Homo sapiens] | | 3725 | 3725 | 97% | 0.0 | 100% | NP_085124.2 |
| PREDICTED: endoribonuclease Dicer isoform X1 [Gorilla gorilla gorilla] | | 3722 | 3722 | 97% | 0.0 | 99% | XP_004055696.1 |
| PREDICTED: endoribonuclease Dicer isoform X1 [Pan paniscus] | | 3721 | 3721 | 97% | 0.0 | 99% | XP_008956514.1 |
| PREDICTED: endoribonuclease Dicer isoform X1 [Pan troglodytes] | | 3720 | 3720 | 97% | 0.0 | 99% | XP_001154369.1 |
| Dicer1, Dcr-1 homolog (Drosophila), isoform CRA_a [Homo sapiens] | | 3716 | 3716 | 97% | 0.0 | 99% | EAW81595.1 |
| PREDICTED: endoribonuclease Dicer [Pongo abelii] | | 3716 | 3716 | 97% | 0.0 | 99% | XP_009247731.1 |
| PREDICTED: endoribonuclease Dicer isoform X2 [Nomascus leucogenys] | | 3715 | 3715 | 97% | 0.0 | 99% | XP_003260983.1 |
| PREDICTED: endoribonuclease Dicer isoform X1 [Nomascus leucogenys] | | 3714 | 3714 | 97% | 0.0 | 99% | XP_012353650.1 |
| hypothetical helicase K12H4.8-like protein [Homo sapiens] | | 3703 | 3703 | 97% | 0.0 | 100% | CAB38857.2 |
| endoribonuclease Dicer [Aotus nancymaae] | | 3702 | 3702 | 97% | 0.0 | 99% | XP_012313714.1 |
| PREDICTED: endoribonuclease Dicer isoform X1 [Cercocebus atys] | | 3701 | 3701 | 97% | 0.0 | 99% | XP_011939230.1 |

El análisis de los datos que arroja BLAST se puede resumir en los siguientes puntos:

- Cada línea representa el alineamiento frente una secuencia en la base de datos (que supera ciertos umbrales)

- La primera línea representa el alineamiento de la secuencia de entrada frente a la misma secuencia en la base de datos (*Query Cover*=100% e *Ident*=100%) y la longitud de NP_001182502 es idéntica a la longitud del alineamiento.

- Observamos alineamientos con un valor de *Ident* alto frente a secuencias de otras especies (corresponden a secuencias homólogas presentes en otras especies).

- Observamos en la cuarta línea un alineamiento frente a una secuencia en *Homo sapiens* con *Ident* = 100% y *Query Cover* = 97%: este alineamiento corresponde a otra isoforma, y por eso no toda la secuencia de entrada ha sido alineada (solo el 97%)
- Si cliqueamos en el *accession number* de las secuencias (última columna en *Descriptions)*, nos lleva a la página del NCBI en la que podemos consultar la información acerca de las secuencias con las que la secuencia analizada presenta identidad. Ver la página de NCBI gene para obtener información acerca de este gen: (http://www.ncbi.nlm.nih.gov/gene/?term=NP_001182502)

**Ejercicio: Identificación de secuencia por comparación**

De un paciente con síntomas de gripe se ha realizado un test molecular (PCR). El tamaño del producto amplificado no era el esperado, por lo que se ha llevado a cabo la secuenciación del fragmento de ADN amplificado.

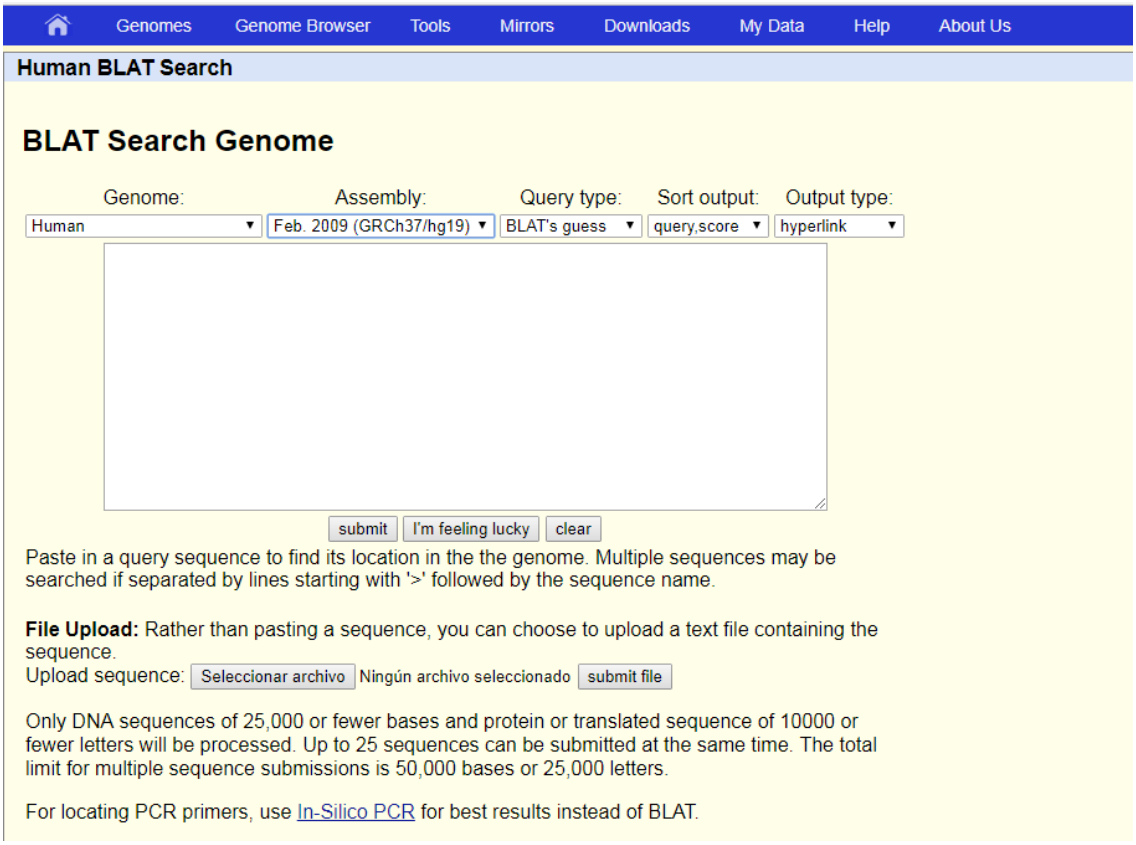
a) De qué organismo proviene esta secuencia?
b) ¿La secuencia obtenida del paciente está contenida en la base de datos?
c) El alineamiento, ¿es una propiedad que tienen dos secuencias?
d) ¿Qué es la similitud entre dos secuencias?
e) ¿La similitud entre dos secuencias es lo mismo que la homología entre dos secuencias?
f) ¿Qué es el score del alineamiento?
g) ¿Qué significado tiene el valor *e* que reporta BLAST?
h) ¿Por qué observamos frecuentemente más de un resultado en una búsqueda con BLAST?

**BLAT:**

**Ejemplo: Localización de una secuencia Alu activa en el genoma humano**

Los retrotransposones Alu representan el elemento transponible más frecuente del genoma humano (y de la mayoría de los primates). Hay más de 1,1 millones de copias en el genoma humano, lo que supone el 11% de todo el genoma .Aunque la mayor tasa de amplificación la tuvieron hace más de 30 millones de años, todavía hay Alus activas en el genoma humano que producen polimorfismos de inserción (relacionados en algunos casos con determinadas enfermedades).

Ir a la página de BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat).



Usaremos el programa BLAT para localizar las posiciones de inserción **de esta Alu activa** en el ensamblado hg19 (*Homo sapiens*). Para ello, pegar la secuencia de la Alu en la caja de texto y rastrear el genoma haciendo click en Submit debajo de la caja con la secuencia. La siguiente imagen corresponde a la captura de pantalla de los resultados:

**BLAT Search Results**

| ACTIONS | QUERY | SCORE | START | END | QSIZE | IDENTITY | CHRO | STRAND | START | END | SPAN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| browser details | YourSeq | 281 | 1 | 281 | 281 | 100.0% | 4 | + | 36469922 | 36470202 | 281 |
| browser details | YourSeq | 279 | 1 | 281 | 281 | 99.7% | 8 | - | 92152900 | 92153180 | 281 |
| browser details | YourSeq | 279 | 1 | 281 | 281 | 99.7% | 6 | - | 50452515 | 50452795 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | X | - | 63618385 | 63618665 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 8 | - | 52648125 | 52648405 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 7 | - | 102476075 | 102476355 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 7 | - | 101334235 | 101334515 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 6 | - | 89424180 | 89424460 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 6 | - | 51517395 | 51517675 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 6 | - | 20936500 | 20936780 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 5 | - | 143922850 | 143923130 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 5 | - | 121963515 | 121963795 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 5 | - | 81333085 | 81333365 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 5 | - | 33200890 | 33201170 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 5 | - | 21338265 | 21338545 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 4 | - | 186198905 | 186199185 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 4 | - | 166172780 | 166173060 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 4 | - | 133908565 | 133908845 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 4 | - | 95499000 | 95499280 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 4 | - | 68986125 | 68986405 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 3 | - | 101653670 | 101653950 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 3 | - | 47458135 | 47458415 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 3 | - | 43031140 | 43031420 | 281 |
| browser details | YourSeq | 277 | 1 | 281 | 281 | 99.3% | 3 | - | 26712525 | 26712805 | 281 |

El análisis de los datos que arroja BLAST se puede resumir en los siguientes puntos:

- Cada línea corresponde a una posición en el genoma donde la secuencia Alu ha alineado con mayor similitud de secuencia que un valor umbral determinado.
- A pesar de alinear en diferentes loci, la secuencia Alu activa se encuentra en un solo locus (sólo la primera línea representa un alineamiento en el que la secuencia Alu completa, 281pb, muestra una similitud del 100% con la secuencia presente en esa posición o locus).
- El resto de posiciones en las que la secuencia Alu ha alineado corresponden a secuencias Alu que con el paso de generaciones han ido acumulando mutaciones y se diferencian por ello a nivel de secuencia.

**Ejercicio: Polimorfismo de inserción por secuencia Alu**

El locus entre las coordenadas 27144072-27144384 en el cromosoma 12 (chr12) corresponde a una AluYb9. La inserción de este elemento Alu es polimórfica en la población humana (polimorfismo de inserción).
Para detectar si un determinado individuo tiene la inserción de esta Alu, se puede amplificar la región mediante PCR. Tenemos las secuencias de dos cebadores o *primers* candidatos:

5'*primer*: GCAGACAGTACCCACTTATTTTTGT
3'*primer*: GAAGAAACAAATGCTTTATAGAACCA

a) ¿Son adecuados estos dos cebadores para amplificar una región que contenga dicha AluYb9?

b) En caso negativo ¿qué pareja de *primers* podríamos utilizar?

c) ¿Qué longitud tendrá el producto de la PCR amplificado por la pareja de cebadores apropiada?

d) ¿Cuál será el tamaño de los productos amplificados a partir del ADN de un individuo heterocigoto para la inserción de la Alu?

e) Dibuja un esquema en el que representes los productos obtenidos por PCR para los genotipos posibles en relación a este polimorfismo de inserción.

f) ¿A qué preguntas biológicas se pueden responder mediante BLAST y BLAT?

Notas

- Las secuencias de los cebadores deben pegarse en la caja de texto de BLAT en formato FASTA:

>5′*primer*
GCAGACAGTACCCACTTATTTTTGT
>3′*primer*:
GAAGAAACAAATGCTTTATAGAACCA

- Programas auxiliares: EMBOSS revseq

# 4.- PREDICCIÓN COMPUTACIONAL DE GENES

## 4.1. OBJETIVO

En la actualidad la cantidad de información genética está aumentando enormemente debido principalmente a que muchos proyectos de secuenciación de genomas completos han finalizado o lo harán próximamente. Una vez que la secuencia de un genoma está disponible es de capital importancia el reconocimiento de regiones codificantes de proteínas. Para ello, hoy día se han desarrollado diversos programas informáticos que, a partir de una secuencia no caracterizada, predicen el número y la localización de genes, incluyendo la localización exacta de exones e intrones (en eucariotas). El objetivo de esta práctica es la aproximación al conocimiento y manejo de este tipo de programas.

## 4.2. FUNDAMENTO TEÓRICO

### 4.2.1. Recursos en la web

La mayoría de los programas de ordenador utilizados en bioinformática se ejecutan en línea de comandos en máquinas con entorno UNIX, sin embargo se han desarrollado para ellos diversos tipos de interfaces que facilitan su uso. Algunos de estos interfaces consisten en páginas web que recogen los datos suministrados por el usuario y devuelven los resultados proporcionados por el programa a través de la propia web o mediante correo electrónico.

Puede consultarse una lista de software con enlaces a las páginas originales de cada aplicación en:

http://www.sanger.ac.uk/resources/software/

Otro software relacionado directamente con la predicción computacional de genes se puede usar y descargar de:

http://opal.biology.gatech.edu/GeneMark/

Contiene varias versiones de la aplicación GeneMark para predicción de genes en procariotas, eucariotas o una versión “autoentrenable” y enlaces para la descarga gratuita de los programas para un uso no comercial con licencia renovable para dos años.

## 4.3. METODOLOGÍA

### 4.3.1. Búsqueda y análisis de una secuencia (A)

En primer lugar obtendremos una secuencia de ADN sobre la que podamos usar programas de predicción computacional de genes. La secuencia elegida es un fragmento del cromosoma Y humano, comprendido entre los nucleótidos 2.784.990 y 2.789.726. (ensamblado GRCh38/hg19). Obtendremos esta secuencia en la base de datos *Ensembl*.

*Ensembl* es una base de dados donde se recogen los genomas de multitud de organismos que se anotan mediante una serie de programas que localizan en las secuencias distintos tipos de características y, entre ellas, la localización de genes, exones e intrones. La base de

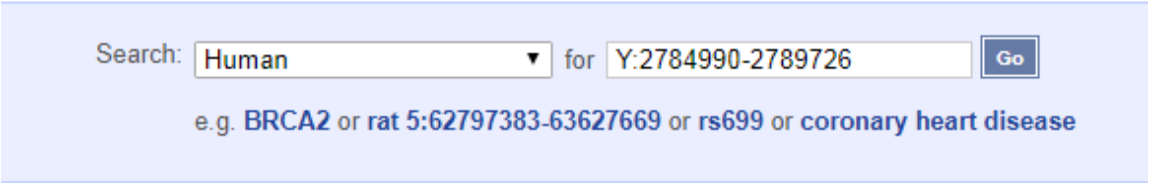
datos se puede consultar “on line” a través de la web o desde programas que pueden acceder a esta base en remoto utilizando librerías escritas en lenguaje *perl*.

Para obtener nuestra secuencia iremos en primer lugar a la página principal de Ensembl: http://www.ensembl.org

En el desplegable de la parte superior de la página elegimos “Human” y en el campo de texto de búsqueda pondremos el cromosoma de interés y los nucleótidos inicial y final de la siguiente forma:

Y:2784990-2789726

Tal y como se ve en la siguiente figura:




Tras pulsar el botón “Go” nos aparecerá una representación de la zona del genoma elegida. Puesto que lo que queremos es la secuencia de nucleótidos de esa región, pulsaremos sobre el botón “Export data” situado a la izquierda.

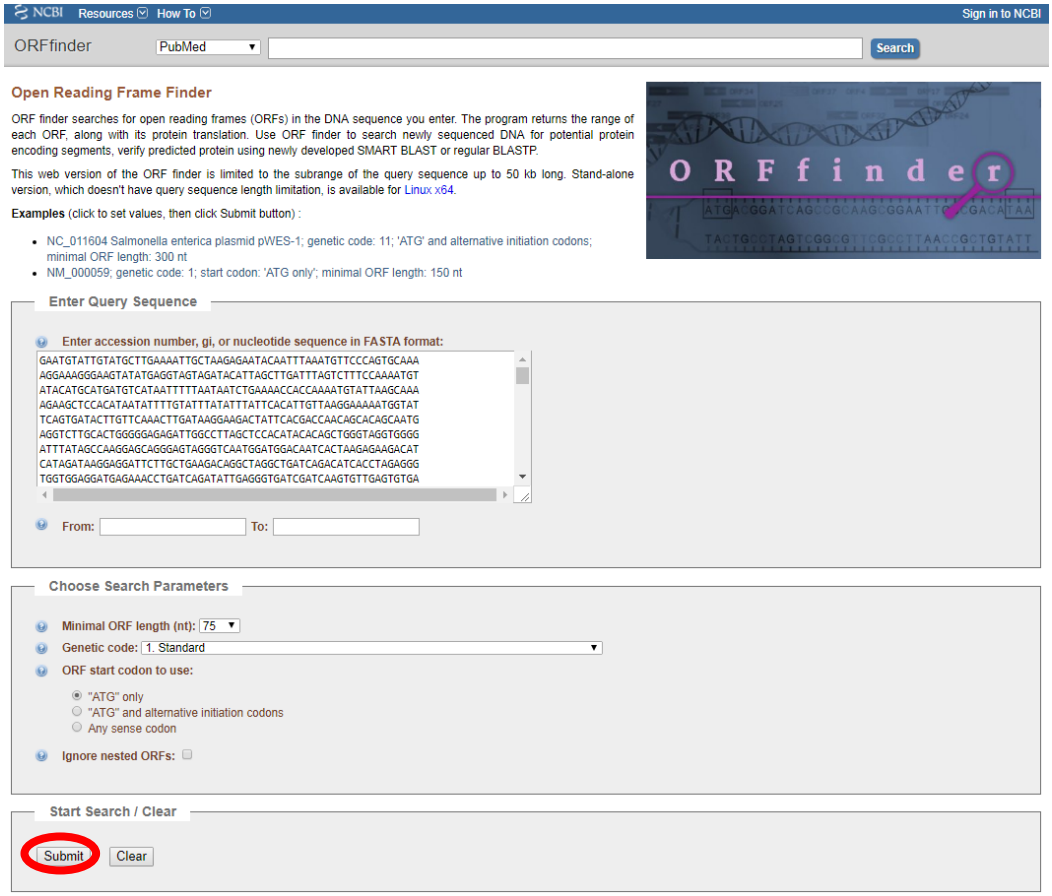

Aparecerá una nueva ventana en la que se podrán elegir diferentes opciones acerca de los datos que pretendemos exportar. La opción por defecto es exportar la secuencia en formato FASTA, que precisamente es lo que pretendemos, así que únicamente deberemos pulsar el botón “Next”. Se nos ofrece entonces la posibilidad de descargar la secuencia en diferentes formatos. Elegiremos “Text” y obtenemos así la secuencia que podremos archivar o copiar y pegar en un editor de texto o en la interfaz web de algún otro programa.

### 4.3.2. Predicción de ORFs

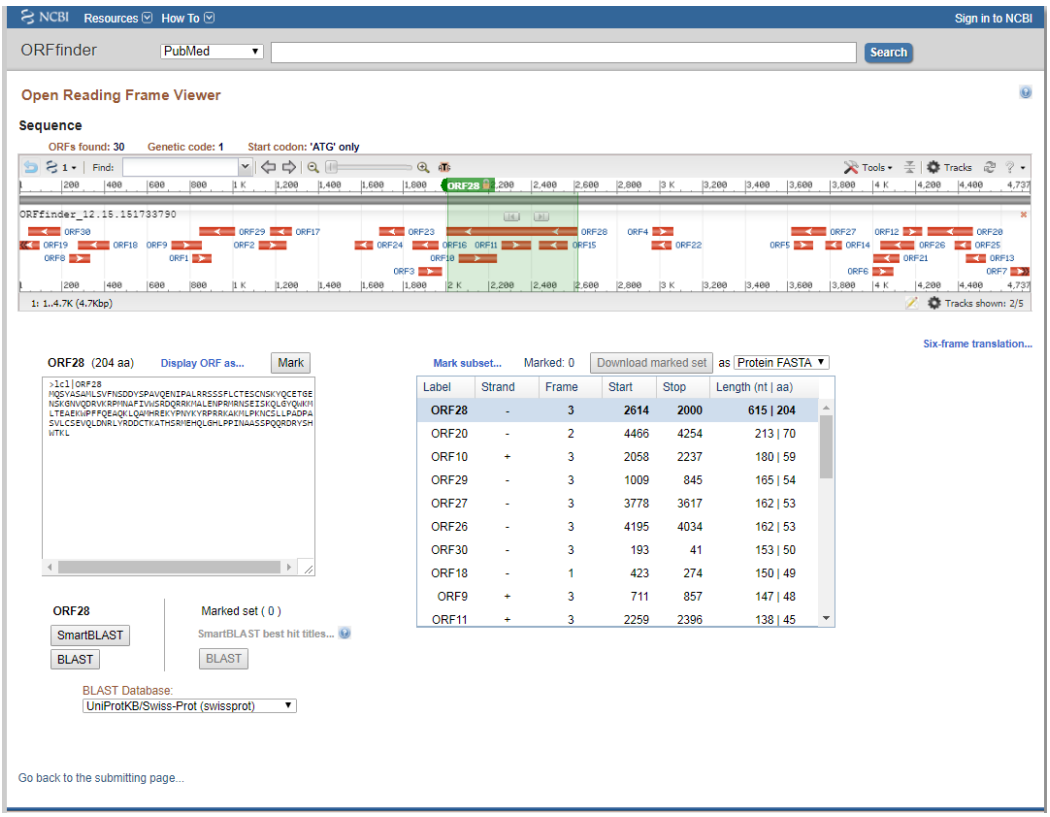
Las *ORF*, del inglés *Open Reading Frame*, o *Marco de Lectura Abierta*, consisten en un fragmento de secuencia que comienza en un codón de inicio y termina en un codón de stop. Si la distancia entre ambos codones es lo suficientemente grande estas *ORF* podrían ser indicativas de la presencia de una región codificante. Buscaremos *ORF* con el programa *ORFfinder* en:

https://www.ncbi.nlm.nih.gov/orffinder/

pegando la secuencia en el recuadro de texto:

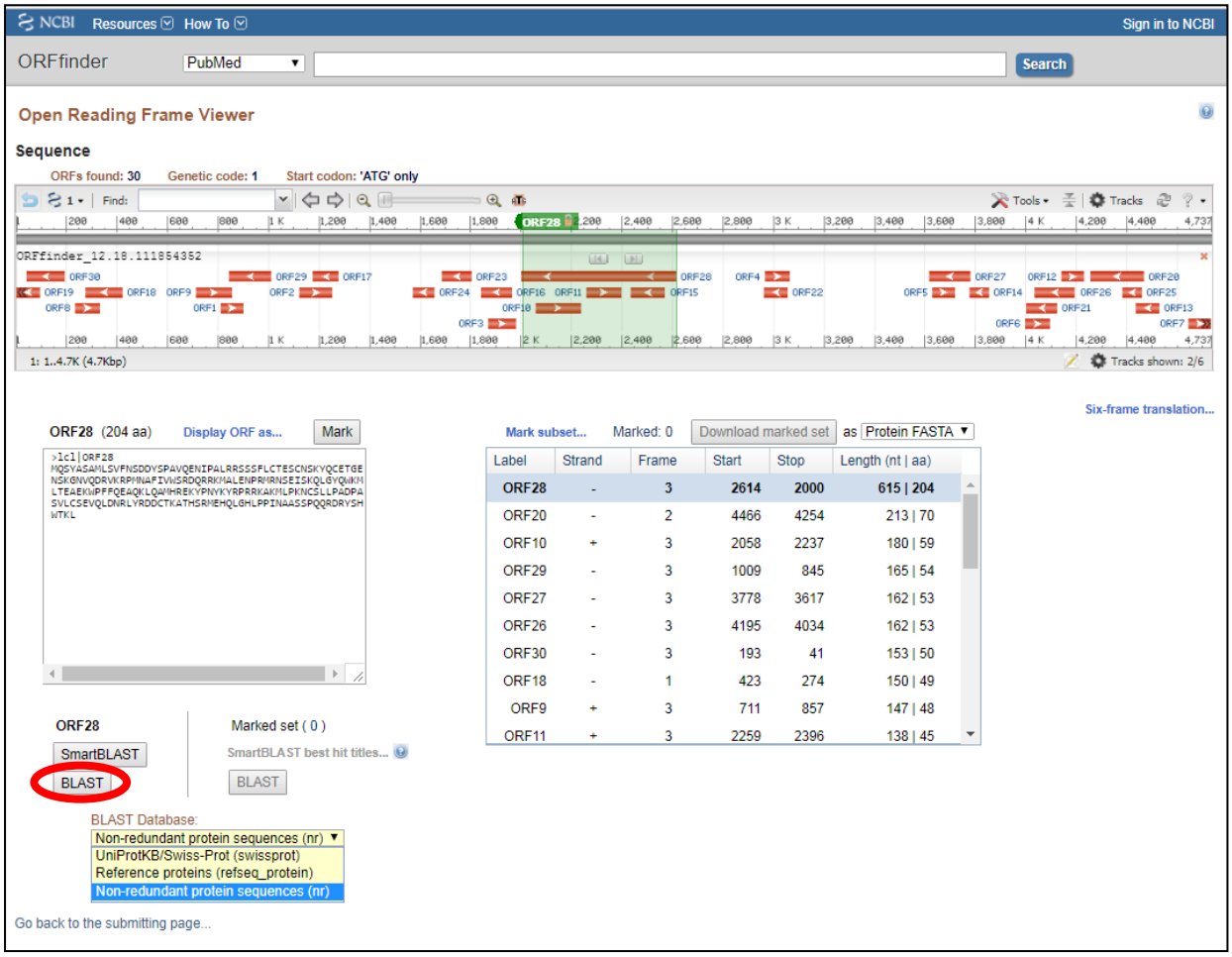


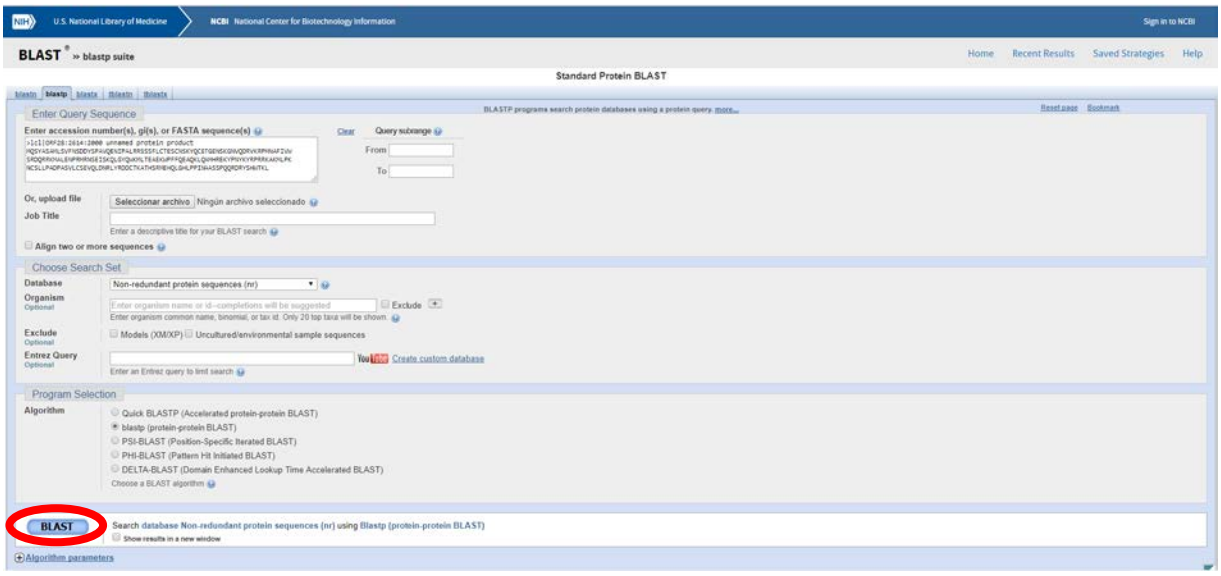

Tras pulsar sobre el botón “Submit” el programa buscará las *ORF* y mostrará el resultado como se observa a continuación:



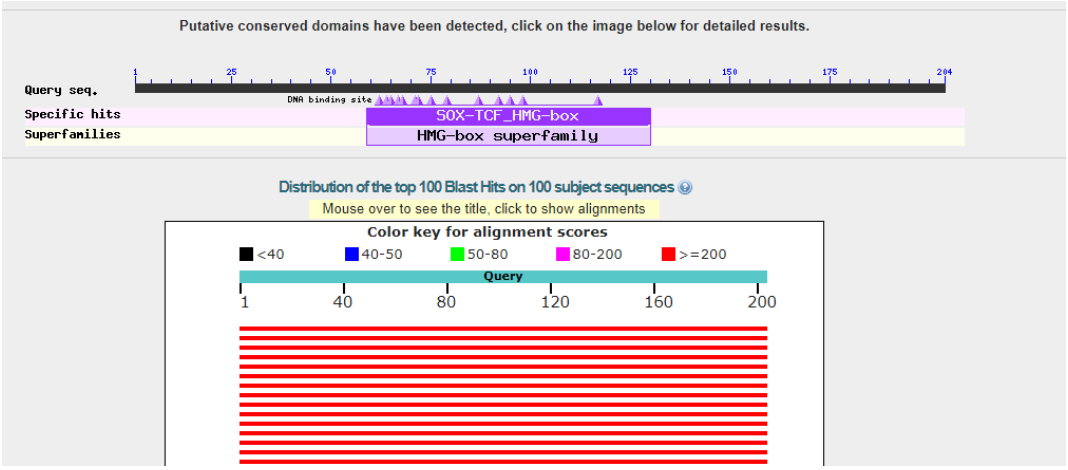
Arriba vemos una representación gráfica de todas las posibles *ORF* señaladas en color rojo y nombre en azul. Debajo del gráfico, en la caja de la derecha, se indica para cada posible ORF, la hebra (+ ó -), la pauta de lectura (1, 2 ó 3), los nucleótidos donde comienzan y terminan y su longitud total en nucleótidos | amino ácidos. Las ORFs aparecen ordenadas por longitud. Aparece marcada, tanto en el gráfico cómo en la caja, la ORF de mayor longitud, que podría ser indicativa de la presencia de una región codificante. Debajo del gráfico, la caja de la izquierda muestra la secuencia de amino ácidos de la potencial región codificante (la posible traducción de ese fragmento de nucleótidos o potencial ORF).

Deberíamos ver a continuación si ese marco de lectura corresponde con alguna proteína conocida. Desde ésta misma página de resultados es posible realizar una búsqueda mediante BLASTp frente a la base de datos “nr” (no redundante) que contiene todas las secuencias conocidas habiendo eliminado los datos redundantes. Para ello elegimos la opción en el desplegable debajo de la caja con la secuencia y pulsamos sobre el botón “BLAST” señalado en la figura. Al pulsar éste botón los datos se envían al NCBI (National Center for Biotechnology Information) mostrándonos una página con los datos enviados y las opciones elegidas para realizar la búsqueda.



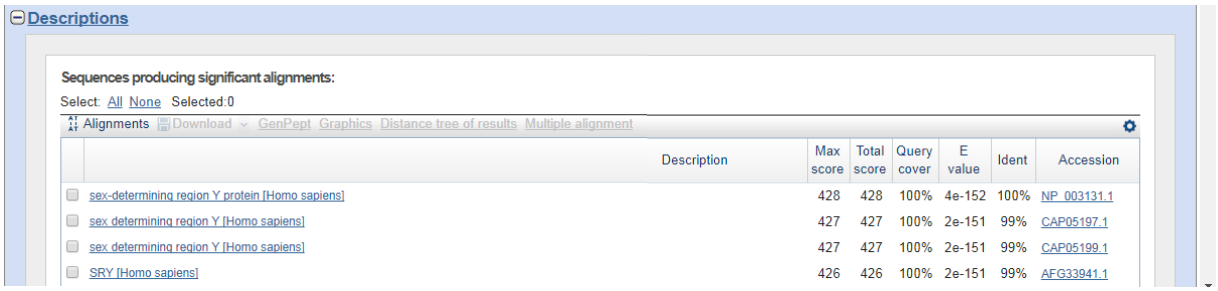



Una vez en la página del NCBI, pulsamos de nuevo sobre el botón “BLAST”. Parte de los resultados de BLAST se muestran a continuación:



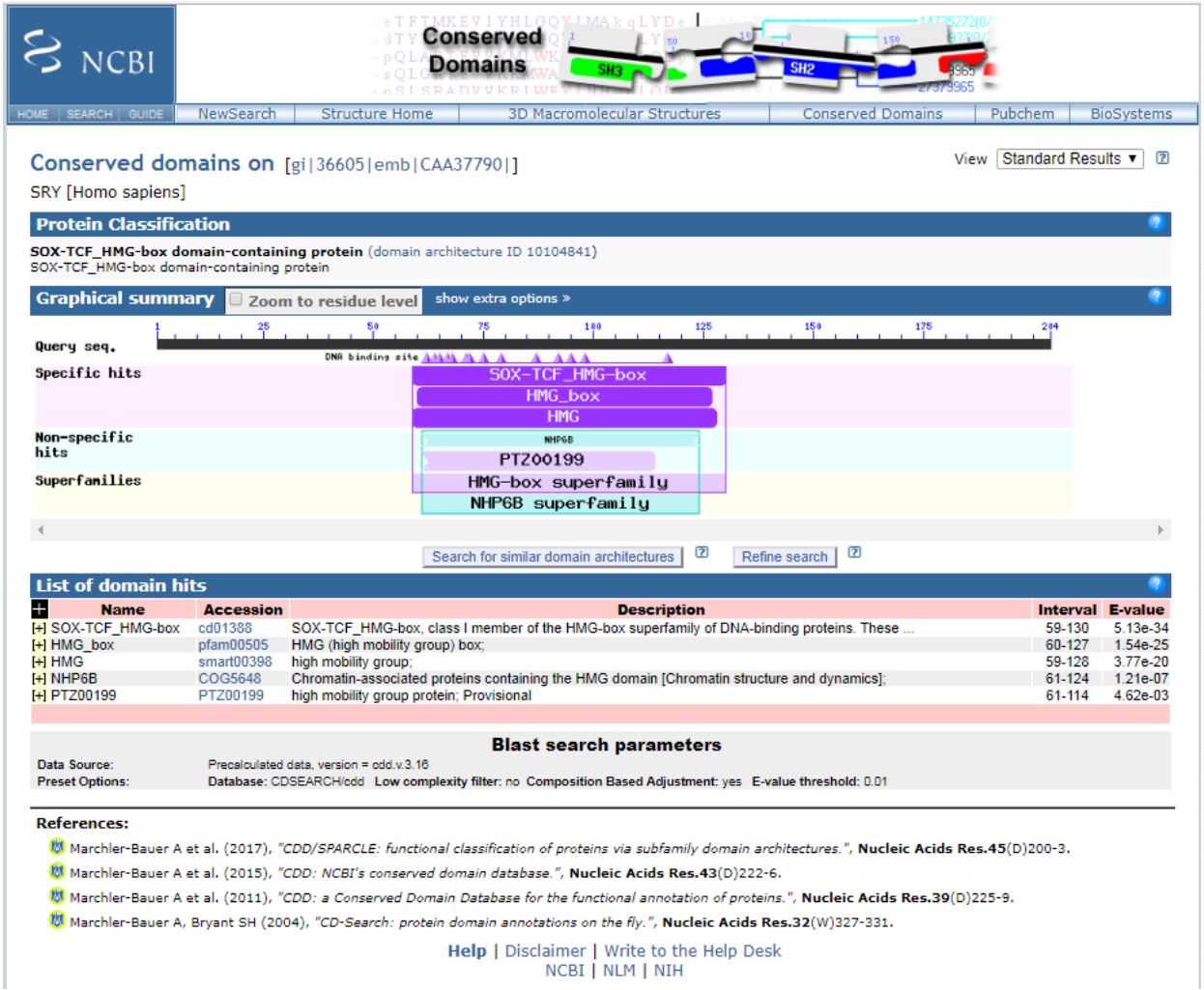

En la parte superior podemos ver que se ha localizado un domino conservado de tipo HMG. Las líneas inferiores representan las secuencias encontradas con homología con la secuencia de búsqueda, tal y como se explicó en la práctica anterior. Colocando el cursor sobre la primera línea roja, en el recuadro de texto sobre las líneas aparece información sobre la secuencia que esa línea representa. En este caso se trata de:

Sex-Determining Region Y protein [Homo sapiens] Score=428, E=$4e^{-152}$

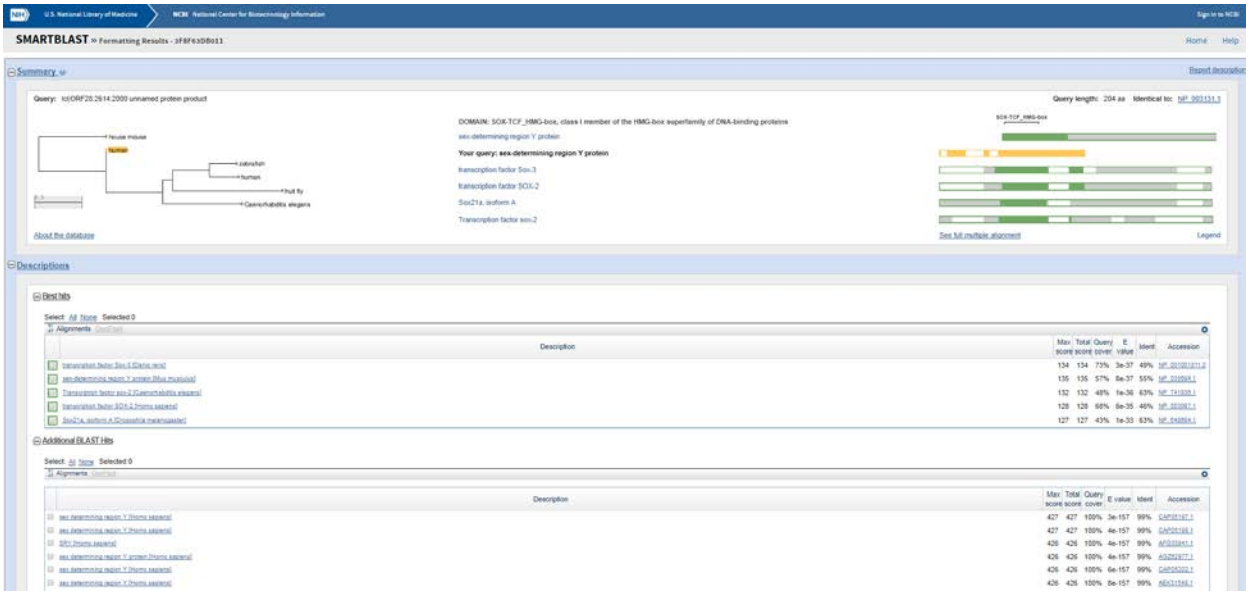


donde, como se explicó también en la práctica anterior, ***Score*** es la puntuación obtenida en el alineamiento y **E** es el valor "Expect", es decir, el número de veces que se esperaría encontrar la secuencia buscada por azar en la base de datos.

Pulsando sobre la caja HMG de la parte superior accedemos a una página con información adicional sobre este tipo de dominio:



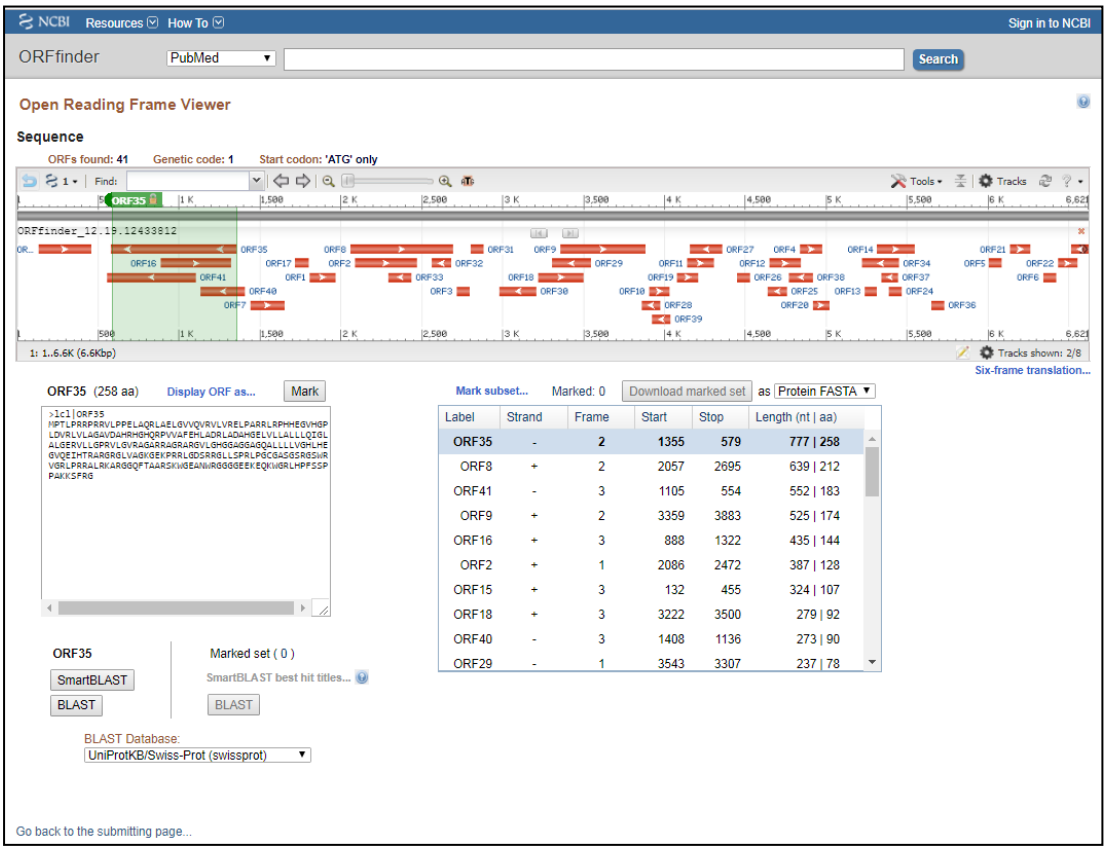

La otra opción disponible en el ORFfinder para BLAST, nombrada SmartBLAST, permite realizar el análisis BLAST de un modo algo diferente. SmartBLAST procesa la consulta de la secuencia presentando un resumen de las cinco proteínas caracterizadas en especies diferentes (siempre que esto sea posible) incluidas en la base de datos de referencia que presentan homología con la secuencia problema. Si SmartBLAST no puede encontrar cinco coincidencias en la base de datos de referencia, usará coincidencias de la base de datos de no redundante (nr). SmartBLAST obtiene estos resultados utilizando una combinación de una búsqueda BLASTp optimizada, una nueva implementación de BLAST destinada a encontrar coincidencias estrechamente relacionadas y a generar un alineamiento múltiple. Además, SmartBLAST presenta las coincidencias encontradas de la secuencia problema en la base de datos de dominios conservados. Las coincidencias adicionales con la base de datos nr se presentan después de las cinco primeras.




### 4.3.3. Búsqueda y análisis de una secuencia (B)

De la misma forma que recuperamos anteriormente una secuencia del cromosoma Y humano, recuperaremos ahora la secuencia correspondiente a los nucleótidos 72.120.505 a 72.127.125 del cromosoma 17 (ensamblado GRCh38/hg19). Para ello, seguir los pasos descritos anteriormente para la secuencia del gen SRY (En http://www.ensembl.org, Human, buscar 17: 72.120.505-72.127.125, Export data, Next, Text)

Analizaremos entonces la secuencia en el *ORFfinder*. El resultado obtenido es un poco confuso:

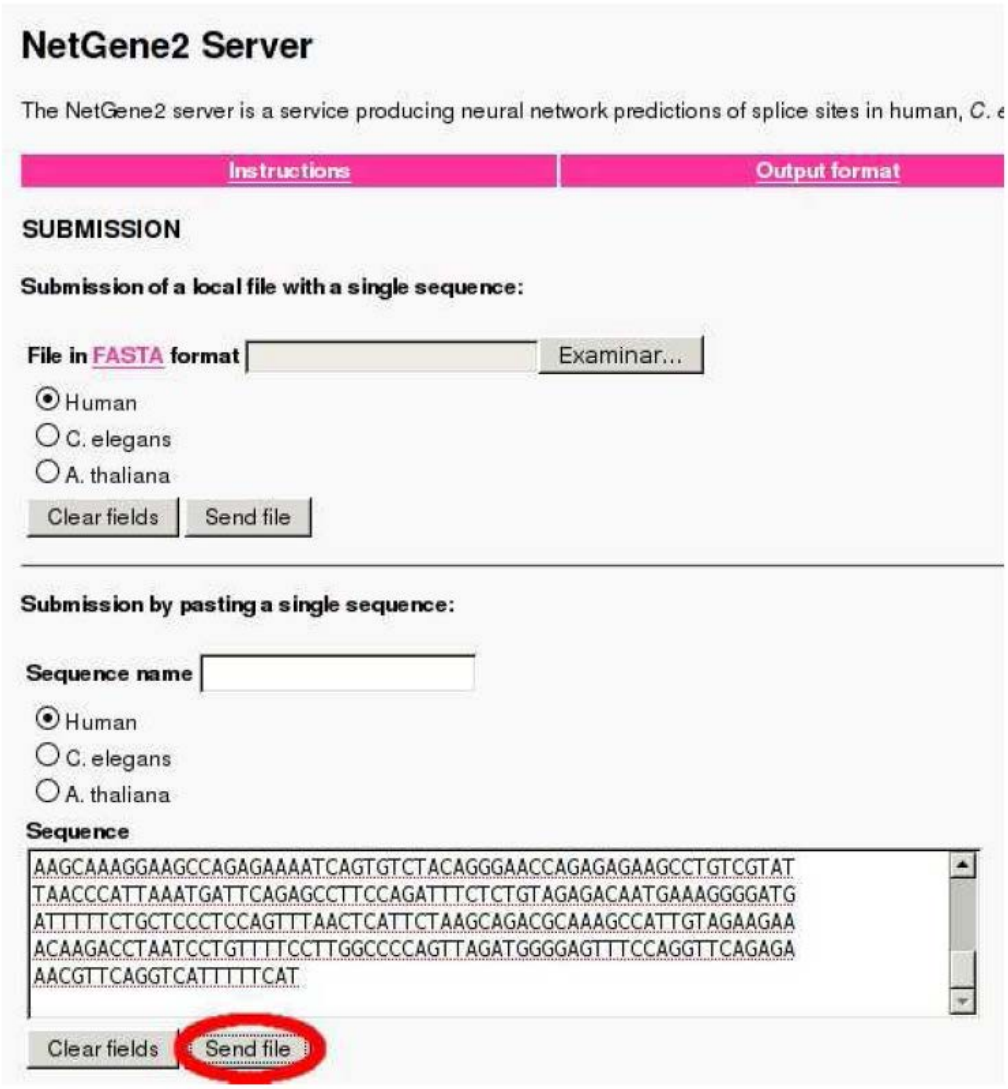


Se observan muchos marcos de lectura abierta dispersos pero no hay ninguna que claramente se diferencie de todos los demás en tamaño. Sabemos de antemano que esta región del cromosoma 17 contiene un gen, luego cabría preguntarse acerca de la eficiencia de este método que se está utilizando para localizar genes. El problema radica en que el gen contenido en esta porción de ADN posee varios intrones que interrumpen el marco de lectura abierta. Teniendo en cuenta que la mayoría de los genes de eucariotas están interrumpidos por intrones, esto supone realmente un problema para estimar correctamente donde se localizan los genes basándose únicamente en la presencia de marcos de lectura abierta.

Por tanto se hace necesario estimar la posición de los posibles principios y finales de intrones presentes en la secuencia, que reciben respectivamente el nombre de sitios "donadores" y de sitios "aceptores". Para ello analizaremos la secuencia con el programa NetGene2:

http://www.cbs.dtu.dk/services/NetGene2/

Una vez en la página de NetGene2 pegamos nuestra secuencia en el recuadro de texto inferior y pulsamos en botón "Send file":




Nos aparecerá una página que se actualizará automáticamente a intervalos regulares hasta que el trabajo esté completado, momento en el que aparecerá la página con los resultados. A continuación se muestran parte de los resultados obtenidos con la secuencia problema:

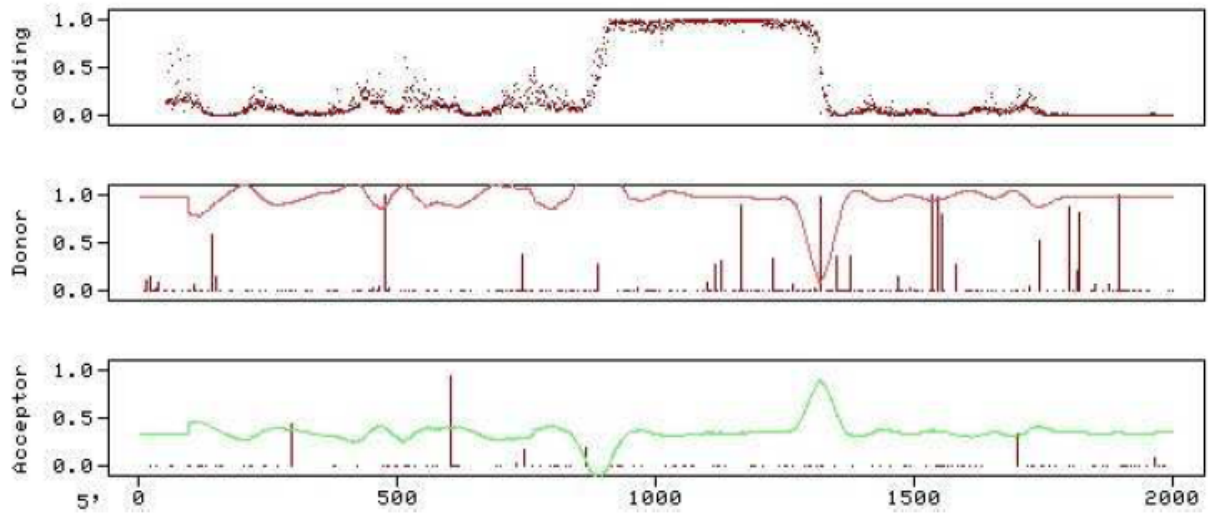
```
********************* NetGene2 v. 2.4 ************************


The sequence: sequence1 has the following composition:

Length: 6621 nucleotides.
23.8% A, 27.3% C, 25.2% G, 23.7% T, 0.0% X, 52.6% G+C

Donor splice sites, direct strand
---------------------------------
               pos 5'->3'  phase strand  confidence  5'      exon intron    3'
                   474       0      +        0.79       TGGCTCTAAG^GTGAGGCGGA
                  1164       0      +        0.34       GCCCATGCCG^GTGCGCGTCA
                  1319       2      +        0.95       AGCTCTGGAG^GTAGGACCCG H
                  1532       0      +        0.63       GAGGGGGGTG^GTAAGTGGAA
                  1545       1      +        0.50       AGTGGAAGAG^GTGAGGGAGG
                  1800       2      +        0.32       CTGGAAATAG^GTGGGAGTGT
                  1894       1      +        0.49       GTTGGGGGCG^GTAAGTCGAG
                  2011       1      +        0.35       GACCGCTCAG^GTCAGACTGC
                  2469       1      +        1.00       GAGCACTCGG^GTGAGTCGCC H
                  4682       0      +        0.37       GAAGCATTTG^GTAAGCTTTA
                  4820       0      +        0.62       TAAGAAAGAG^GTAAAAGGCA
                  5114       2      +        0.35       TCCTCAAAGG^GTATGGTCAT
```

En esta porción de la salida se muestran los posibles sitios donadores (límite exón/intrón). En las columnas se muestra de izquierda a derecha: la posición del punto de corte exón/intrón, la pauta en que se encuentra, la hebra, el nivel de confianza y, por último, la secuencia del sitio. Los niveles de confianza próximos a 1 pueden indicar lugares funcionales. De la misma forma se muestran en la página los posibles lugares aceptores. Finalmente se muestra una representación gráfica de los resultados:

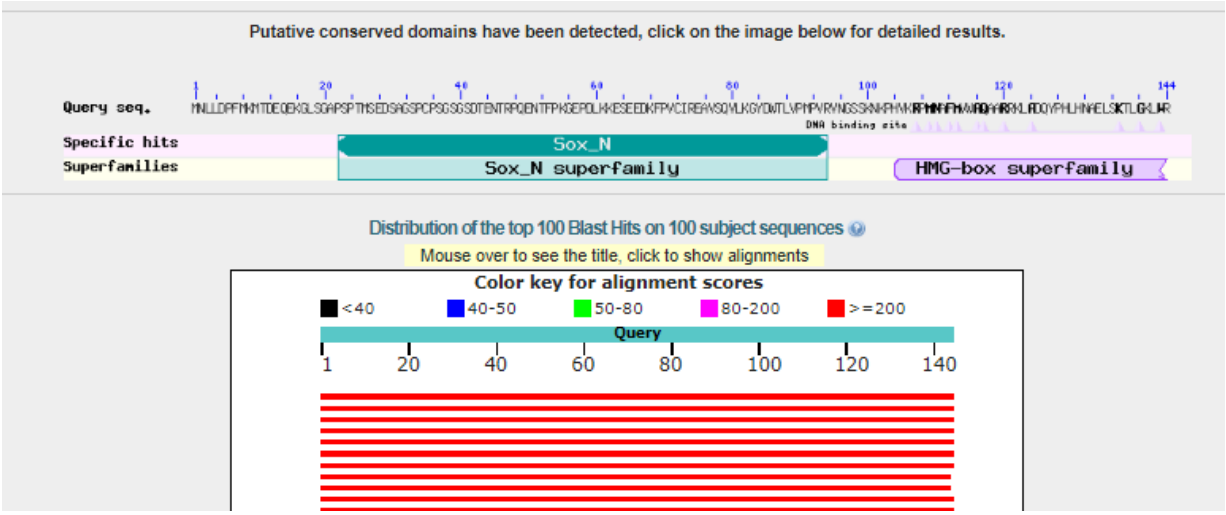



Los tres gráficos corresponden, de arriba a abajo, al potencial codificante, a la localización de sitios donadores y, por último, a la de sitos aceptores. Las líneas verticales corresponden a los posibles puntos donadores y aceptores a los que referían los datos anteriores. La longitud de las líneas se corresponde con los niveles de confianza.

Las curvas que se observan en la segunda y tercera gráfica se derivan de los cambios de pendiente de la curva de potencial codificante. Para identificar los sitios donadores o aceptores con potencial biológico real, deberían coincidir sus posiciones con los límites de las regiones potencialmente codificantes. De esta forma, los límites entre exones/intrones y entre intrones/exones deberían coincidir con líneas verticales de longitud próxima a 1 y bajadas significativas en las curvas respectivas que, a su vez, coinciden con cambios de pendiente en la curva de potencial codificante.

El primer potencial sitio donador (exón/intrón) corresponde a la posición 1319, y la siguiente posición que podría actuar como aceptor (intrón/exón) es la 2214. Estos dos puntos corresponderían a un primer intrón, por tanto la secuencia codificante del posible gen debería comenzar antes de la posición 1319 y terminar en los alrededores de ésta.

Observando los resultados de *ORFfinder* vemos que el segundo marco de lectura abierta en la pauta 3 termina en el nucleótido 1322. Si hacemos un BLAST desde *ORFfinder* con la secuencia de aminoácidos codificada por este marco obtenemos lo siguiente:




Vemos que efectivamente corresponde con dominios conservados y se trata de un fragmento del gen *SOX9* ya que las primeras coincidencias corresponden a este gen en distintas especies. Una de ellas corresponde al gen *SOX9* Humano.

El siguiente exón debería comenzar después de la posición 2214, donde se encuentra el primer sitio donador. La pauta de lectura 1 muestra una *ORF* entre los nucleótidos 2086-2472, y la pauta 2 entre 2057-2695. Un BLAST con la primera de ellas no arroja resultados significativos, pero en la segunda se observa el final de una caja conservada HMG.

Podremos concluir por tanto que la *ORF* de la pauta 3 y la siguiente de la pauta 2 corresponden a dos exones de un mismo gen, interrumpido por un intrón situado en medio de la región que codifica un dominio conservado de tipo HMG. Siguiendo esta estrategia podremos localizar el resto de las secuencias que corresponden a los exones de *SOX9*

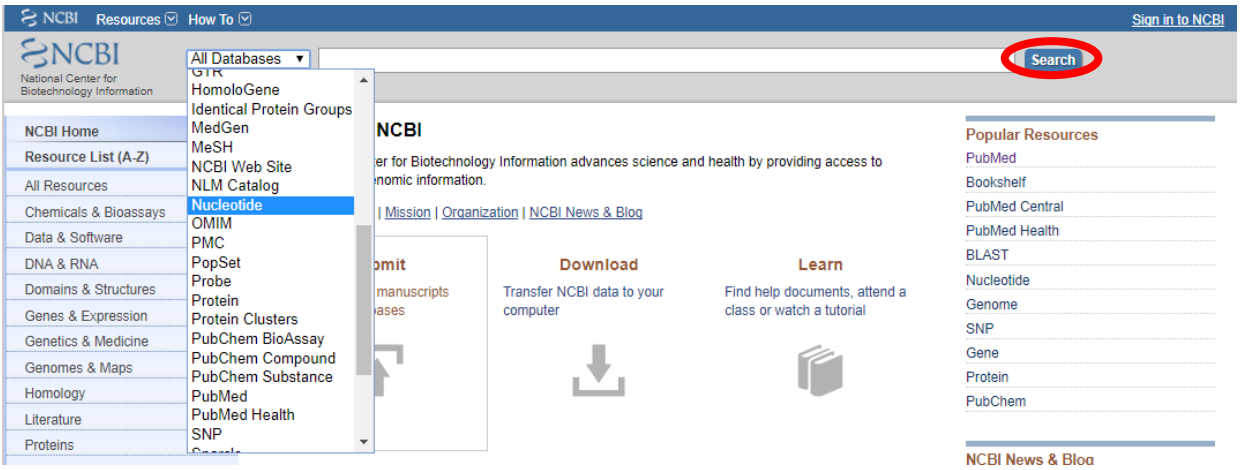
### 4.3.4. Localización de intrones mediante “dot plot”

Simularemos un experimento de laboratorio en el que se aislaría el ARNm del gen de interés una vez conocida parte o toda su secuencia. Posteriormente se obtendría la secuencia de este ARNm y se compararía con la secuencia genómica del mismo gen, poniendo de manifiesto las regiones que corresponden a exones e intrones. En lugar de obtener la

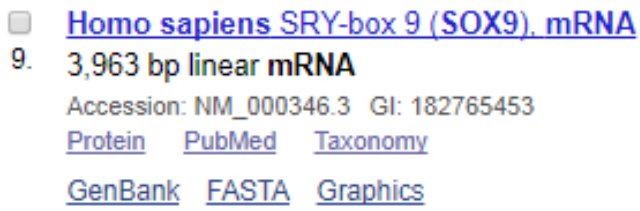
secuencia del mensajero en el laboratorio, la obtendremos en una base de datos, ya que se trata en realidad de un gen conocido. Para ello iremos a la página:

http://www.ncbi.nlm.nih.gov/

En la parte superior seleccionaremos la base de datos de nucleótidos, en la línea de texto escribiremos como palabras clave “sox9 mrna homo sapiens” y pulsaremos el botón “Search”:




Entre los resultados obtenidos veremos:




pulsando sobre la descripción del gen podremos recuperar la secuencia del mensajero. Un análisis de “dot plot” en:

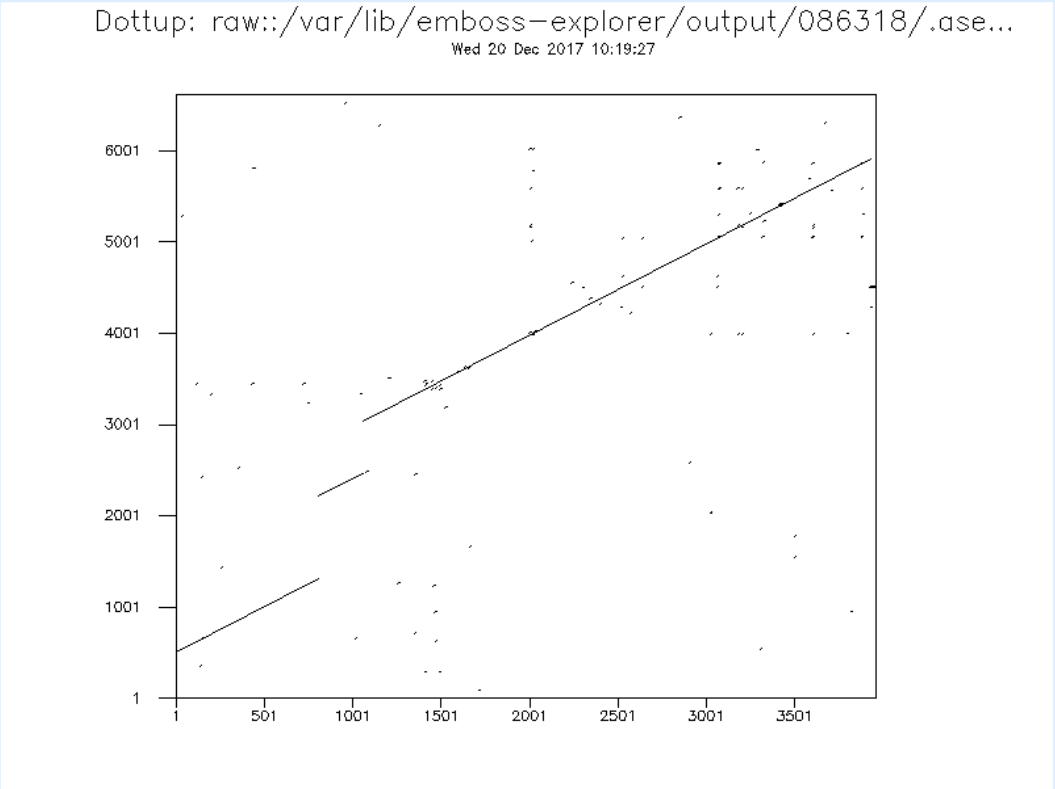
http://emboss.bioinformatics.nl/cgi-bin/emboss/dottup


nos mostrará la posición de los intrones y exones cuando comparemos ambas secuencias (se puede pegar cada secuencia en una caja o seleccionar los archivos en cada caso):

**OUTPUT FILE** .stdout

```
Created dottup.1.png
```

**IMAGE FILE** dottup.1.png




La traducción del ARNm revelará donde se encuentra el codón de stop, y proporciona una explicación de porqué el potencial codificante decae antes del final del tercer exón, como puede apreciarse en la salida de NetGene2.

## 4.4. CUESTIONES

En relación con la consulta realizada con el programa *ORFfinder*:

**1.-** ¿Qué indican las barras de color rojo que aparecen en el gráfico?

**2.-** Dentro de estas barras, ¿que indican las flechas?

**3.-** ¿Qué indican cada una de las columnas que aparecen en la caja debajo a la derecha del gráfico ("Strand", “Frame”, “Start”, "Stop" y “Length”)?

**4.-** De todos los posibles marcos de lecturas abiertos, ¿Cuál es el que tiene más probabilidad de ser una región codificante?

**5.-** ¿Por qué todos los marcos de lectura abierta comienzan con el triplete ATG? ¿Con qué triplete/s terminan?

**6.-** ¿Qué podría ocurrir si un marco de lectura abierto se encuentra interrumpido por un intrón?

**7.-** ¿Qué condiciones deberían darse para que un intrón quedase englobado en un marco abierto de lectura?

**8.-** ¿Cómo se puede comprobar que un marco de lectura abierta codifica una proteína conocida?

**9.-** ¿Qué ocurre con el programa *ORFfinder* si en nuestra secuencia a analizar se encuentra un gen compuesto por varios exones e intrones?

En relación a la consulta realizada en el programa NetGene2:

**1.-** ¿Qué indican las tablas numéricas “donor splice sites” y “aceptor splice sites”? ¿Qué indica la última columna en dichas tablas?

**2.-** En la columna “confidence”, ¿qué indican los valores altos?

**3.-** ¿Qué se representa en la gráfica superior?

**4.-** En las dos gráficas inferiores, ¿Qué representan las líneas verticales?, ¿Y la línea horizontal (roja o verde)?

**5.-** Atendiendo a las gráficas, ¿Cómo identificarías lugares con una alta probabilidad de ser donadores/aceptores funcionales?

**6.-** ¿Por qué en el caso de *SOX9* los comienzos y finales de las *ORF* localizadas con *ORFfinder* no coinciden exactamente con los puntos donadores y aceptores predichos por NetGene2?

**7.-** ¿Por qué dos exones del mismo gen pueden aparecer en pautas de lectura diferentes?

**8.-** ¿Por qué el potencial codificante decae antes de llegar al final del último exón?

# 5. A) ALINEAMIENTO MÚLTIPLE DE SECUENCIAS DE ADN y PROTEÍNAS.

## 5.1A. OBJETIVO

Cuando se quieren comparar secuencias homólogas de nucleótidos (ADN) o de aminoácidos (proteínas) de especies diferentes con el fin de analizar las diferencias existentes entre ellas y sus relaciones evolutivas, un paso previo imprescindible en dicho análisis es el de establecer un alineamiento múltiple de todas las secuencias. El objetivo de esta práctica es adquirir las destrezas necesarias para llevar a cabo alineamientos múltiples de secuencias y familiarizarse con el uso de los programas informáticos que nos permiten hacerlos.

El procedimiento a seguir tiene varios pasos, el primero de los cuáles consiste en alinear todas las secuencias dos a dos. Por ello, en primer lugar, describiremos como se procede a la hora de hacer un alineamiento entre dos secuencias homólogas.

## 5.2A. FUNDAMENTO TEÓRICO

***Alineamiento de dos secuencias homólogas de nucleótidos o de aminoácidos***

Mediante comparación de dos secuencias homólogas de ADN o de proteínas se puede llegar a establecer un alineamiento por emparejamiento, base a base, de las bases de cada una de las dos secuencias. Por ejemplo, para el caso de ADN:

**5'-AATGTCATGCGCTGAATCCCCCC-3'**
**5'-AAGGTCTTGCCCT-AATGCCCCC-3'**

Si las dos secuencias que se comparan tienen diferente longitud es porque alguna de ellas o las dos han incorporado o perdido algún residuo (nucleótido o aminoácido, dependiendo de las secuencias que se comparen). Así, lo primero a identificar es la localización de las inserciones y deleciones que han podido ocurrir en cada especie desde que están divergiendo de una especie ancestral común.

En el emparejamiento base a base del alineamiento, nos podemos encontrar con una de tres posibilidades de sitios o posiciones nucleotídicas/aminoacídicas:

- Coincidencias (*matches*): la misma base/aminoácido en las dos secuencias.
- Ausencia de coincidencias (*mismatches*): una base/aminoácido diferente en cada secuencia.
- Inserciones/deleciones (*gaps*): los *gaps* se representan por guiones (-) y significan que en una de las dos secuencias se produjo una inserción o una deleción en esa posición.

Cuando comparamos una secuencia parcial de un/a gen/proteína obtenida a partir de una especie con la secuencia completa de dicho/a gen/proteína, el alineamiento se realizará proponiendo un enorme *gap* terminal que representaría a la información desconocida (*missing data*). Estas posiciones del alineamiento se suelen representar muchas veces con el signo de interrogación (?) en la secuencia incompleta.

La obtención del alineamiento correcto es fundamental para que todos los análisis evolutivos y filogenéticos posteriores no se vean afectados. Dicho alineamiento se puede hacer manualmente si no hay muchos *gaps* y si las secuencias son cortas y no muy divergentes. Sin embargo, se han desarrollado métodos que facilitan el trabajo y la fidelidad del resultado en cualquier tipo de comparaciones:

1. El método de la **matriz de puntos** (*dot matrix*) sigue el siguiente procedimiento: una de las secuencias se dispone en el eje vertical, y la otra secuencia en el eje horizontal, de una matriz bidimensional. Cada vez que existe un nucleótido/aminoácido idéntico en ambas secuencias, se coloca un punto en el recuadro correspondiente a la posición *x* de la secuencia horizontal y a la posición *y* de la vertical. El alineamiento se obtiene mediante una línea diagonal que une los puntos a través de la matriz comenzando en el recuadro superior izquierdo y tratando de acabar en el inferior derecho. El trazado puede revelar diferentes situaciones tal como podemos ver en las siguientes matrices de puntos para dos secuencias nucleotídicas hipotéticas:

A. Las dos secuencias son idénticas:

| | A | G | C | T | T | G | C | A | G | C |
|---|---|---|---|---|---|---|---|---|---|---|
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| T | | | | ● | ● | | | | | |
| T | | | | ● | ● | | | | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |

B. Las dos secuencias son iguales en tamaño pero difieren en secuencia:

| | A | G | C | T | T | G | C | A | G | C |
|---|---|---|---|---|---|---|---|---|---|---|
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| T | | | | ● | ● | | | | | |
| T | | | | ● | ● | | | | | |
| G | | ● | | | | ● | | | ● | |
| T | | | | ● | ● | | | | | |
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |

C. Las dos secuencias difieren en tamaño (sólo inserciones y/o deleciones explicarían las diferencias entre ellas):

| | A | G | C | T | T | G | C | A | G | C |
|---|---|---|---|---|---|---|---|---|---|---|
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| T | | | | ● | ● | | | | | |
| T | | | | ● | ● | | | | | |
| C | | | ● | | | | ● | | | ● |
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |

D. Las dos secuencias difieren en tamaño (inserciones y/o deleciones explicarían partes de las diferencias entre ellas) y en secuencia (cambios por substitución de un residuo por otro):

| | A | G | C | T | T | G | C | A | G | C |
|---|---|---|---|---|---|---|---|---|---|---|
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| T | | | | ● | ● | | | | | |
| T | | | | ● | ● | | | | | |
| C | | | ● | | | | ● | | | ● |
| A | ● | | | | | | | ● | | |
| C | | | ● | | | | ● | | | ● |
| C | | | ● | | | | ● | | | ● |

En una secuencia más larga y con más cambios de los reflejados aquí se hace mucho más difícil establecer el alineamiento pudiendo existir más rutas alternativas que explicarían las diferencias entre dos secuencias.

De hecho, lo normal es que exista un número muy abundante de puntos en la matriz que, junto con la ausencia de una diagonal perfecta, dificulta el trazado del alineamiento. Se ha ideado un método que permite mejorar la definición del alineamiento. Consiste en comparar las dos secuencias usando "ventanas deslizantes" que van haciendo las comparaciones de *n* en *n* residuos, en lugar de nucleótido a nucleótido. Una coincidencia (*match*) en este caso se determina a partir de un umbral determinado. Así, dos parámetros son fundamentales en este tipo de comparaciones: el **tamaño de la ventana** (*windows size*) y la **astringencia** (*stringency*). Una vez establecido un tamaño de ventana, éste se mantiene constante en todo el análisis. Consiste en determinar cada cuántos residuos se hace una comparación. Así, si el tamaño de la ventana es de cinco residuos, quiere decir que comparamos las dos secuencias progresivamente de 5 en 5 residuos. La astringencia determina el umbral: número de

residuos que deben ser coincidentes dentro de esa ventana. Con esto se eliminan muchos de los puntos de identidad falsos de la matriz.
2. Un segundo método consiste en definir un alineamiento como aquel en el que el número de disimilitudes (*mismatches*) y *gaps* están minimizados de acuerdo a unos criterios determinados. El problema radica en que para aumentar el número de coincidencias suele ser necesario aumentar el número de *gaps.* Por tanto, según este criterio, son posibles varias opciones de alineamiento por lo que se ha diseñado un procedimiento consistente en calcular un **índice de divergencia** o **disimilitud** entre las dos secuencias que se comparan. Este índice tendrá diferentes valores para cada uno de los alineamientos alternativos obtenidos. Aquel alineamiento con menor índice de divergencia será el escogido como mejor de todos.

El cálculo del índice de divergencia depende del **coste o penalización por *gaps*** *(gap penalty*) que suele tener dos componentes: penalización por cada *gap* introducido en el alineamiento (*gap-opening penalty*) y penalización por la extensión de cada *gap* (*gap-extension penalty*). Las penalizaciones por *gaps* son factores por los que se multiplican los valores de los *gaps* (el número y la longitud de los *gaps*) con el fin de establecer una equivalencia entre esos valores y el valor de los des-emparejamientos o *mismatches* (número de substituciones). Así, la penalización se basa en nuestra propia experiencia a través de la comparación entre el cálculo de la frecuencia de inserciones y deleciones que han ocurrido en la evolución desde la separación de las dos especies cuyas secuencias están siendo alineadas y la frecuencia con la que han ocurrido substituciones nucleotídicas (o aminoacídicas).

En el caso de secuencias de proteínas, las disimilitudes en las diferentes posiciones aminoacídicas pueden ser valoradas con diferente peso según que el cambio producido sea a un aminoácido más o menos similar en sus propiedades bioquímicas. Así, se han establecido ciertos grupos de aminoácidos por afinidad bioquímica cuyos emparejamientos en un alineamiento reciben mayor o menor puntuación de acuerdo a diferentes criterios, en lugar de una puntuación de cero que es lo que reciben los sitios en los que hay una disimilitud y los aminoácidos emparejados no guardan ninguna afinidad bioquímica.

### *Alineamientos múltiples*

Los alineamientos múltiples siguen un procedimiento similar al descrito, pero la complejidad de los cálculos se hace mayor al incrementarse el número de secuencias a alinear. Existen diferentes programas informáticos que pueden hacer este tipo de alineamientos, como Clustal X, o la versión de Clustal Omega *online*, que implementa el algoritmo Clustal (Higgins y Sharp, 1988). En este caso, los alineamientos se realizan en un proceso de tres etapas. Primero, se comparan todas las secuencias dos a dos (alineamientos *pairwise*). A continuación se construye un dendrograma (similar a un árbol filogenético) que agrupa las secuencias por similitud. En tercer lugar, el alineamiento múltiple se hace usando el dendrograma como guía y alineando secuencias de manera progresiva de acuerdo al orden de ramificación del árbol. Es decir, primero se alinean las dos secuencias con mayor similitud y se van añadiendo secuencias al alineamiento de manera progresiva por orden de similitud decreciente.

# 5. B) ANÁLISIS FILOGENÉTICO

## 5.1B. OBJETIVO

La filogenia molecular consiste en el estudio de las relaciones evolutivas entre organismos a partir de datos moleculares ordenados en un alineamiento múltiple de secuencias de ADN o de proteínas. El objetivo de esta práctica es introducirnos en la teoría y la metodología utilizadas en el análisis filogenético así como familiarizarnos con el uso de programas informáticos de análisis filogenético.

## 5.2B. FUNDAMENTO TEÓRICO

Para simplificar la redacción de este texto, nos referiremos a partir de ahora siempre a secuencias de ADN, siendo aplicable todo lo que se dice también al análisis de las secuencias de proteínas.

En el análisis filogenético, el objetivo es la construcción de un **árbol filogenético** que ilustre la historia evolutiva de un grupo de especies. Un árbol filogenético es un gráfico compuesto de **nodos** y **ramas** en el que una rama conecta dos nodos adyacentes. Los nodos representan a las especies y las ramas definen las relaciones entre esas especies en términos de descendencia y ascendencia. El patrón de ramificación se denomina **topología** del árbol. Hay que distinguir entre **nodos terminales** y **nodos internos**. Estos últimos representan a especies ancestrales hipotéticas mientras que los nodos terminales representan a especies existentes en la actualidad. Las especies que están conectadas por ramas a un mismo nodo interno, comparte ese nodo ancestral. Las ramas que conectan nodos externos con nodos internos se denominan **ramas externas** o **terminales** mientras que las que conectan nodos internos son **ramas internas**. Un nodo puede ser **bifurcado** si tiene sólo dos descendientes o **multifurcado** si tiene más de dos. Por lo general, la representación más común de las filogenias emplea árboles bifurcados dado que se asume que el proceso de especiación es binario: dos especies descendientes a partir de una especie ancestral común. Una multifurcación o **politomía** en un árbol puede interpretarse de dos maneras: a) representa una realidad, es decir, un ancestral ha dado lugar a más de dos especies descendientes; b) existe una ambigüedad a la hora de determinar el correcto patrón de bifurcación porque los datos disponibles no son resolutivos.

Un **clado natural** o **grupo monofilético** consiste en un grupo de táxones (especies, o grupo de especies como un género, una familia, un orden o una clase) que derivan de un ancestral común que no es compartido con ningún otro taxón fuera del grupo. Se espera que un grupo taxonómico (género, familia, orden o clase) sea monofilético. Sin embargo, algunos grupos taxonómicos establecidos actualmente pueden ser no monofiléticos: la filogenia molecular ha demostrado, en algunos casos, que un grupo taxonómico tiene un ancestral común compartido con otros taxones (grupo **parafilético**); un grupo **polifilético** está formado por dos linajes que han adquirido un mismo carácter por convergencia evolutiva (los organismos clasificados en un mismo grupo polifilético comparten homoplasias fenotípicas).

Un árbol puede ser un **árbol con raíz** cuando existe un nodo, la raíz, que de forma inequívoca es el ancestral común más reciente de todas las especies comparadas. Desde la raíz, una única ruta evolutiva da lugar a cada uno de los nodos. Un **árbol sin raíz** es un árbol que sólo especifica las relaciones de parentesco entre las especies comparadas sin describir los pasos evolutivos que han conducido desde un ancestral común a dichas especies.

Un **árbol escalado** es aquel en el que sus ramas están escaladas, es decir, la longitud de cada rama es proporcional al número de cambios producidos entre las secuencias que se comparan. En un **árbol no escalado** las longitudes de las ramas no son proporcionales a ese número de cambios con lo que los nodos terminales aparecerán alineados.

Para un grupo determinado de especies existen diferentes árboles posibles y el número de estos se incrementa en relación al número de especies comparadas. Sin embargo, sólo uno de esos árboles es el árbol correcto que, dependiendo de la precisión de nuestros datos y de nuestros análisis, puede coincidir o no con el árbol inferido en nuestra reconstrucción filogenética.

En cualquier caso, siempre tenemos que tener presente que en nuestro análisis lo que comparamos son secuencias homólogas de ADN obtenidas de cada una de las especies que estamos estudiando. Por tanto, en principio, lo que obtenemos es un **árbol génico**. Sin embargo, cada gen puede tener diferentes historias evolutivas y los ritmos y los modos de éstas pueden no reflejar coherentemente la historia evolutiva de las especies. Por tanto, para obtener un árbol de especies lo más preciso posible, es más correcto analizar la historia de diferentes genes y secuencias no génicas.

La tasa de cambio de las secuencias comparadas es algo que debemos tener muy en cuenta a la hora de elegir qué tipo de secuencias vamos a utilizar en nuestro análisis filogenético. Así, si el grupo a comparar está formado por especies muy próximas filogenéticamente, se requiere una secuencia que evolucione más rápidamente y haya acumulado suficientes cambios en el proceso de diversificación del grupo comparado. En este caso, es interesante recurrir a secuencias no génicas que cambian más rápidamente. El uso de secuencias de genes conservados con una función importante en el organismo estaría desaconsejado en este caso, dado que es muy probable que se hayan producido muy pocos cambios en las secuencias comparadas y, por tanto, exista poca señal filogenética con capacidad resolutiva para la reconstrucción filogenética. No obstante, suele ser útil el uso de secuencias génicas de ADN mitocondrial que tienen una tasa de evolución más rápida que las secuencias de ADN nuclear. Cuando la comparación es entre especies de grupos taxonómicos alejados, por el contrario, las secuencias no génicas pueden ser muy dispares y ser poco aconsejables para el análisis filogenético. En este caso, es más conveniente el uso de secuencias más conservadas.

### *Métodos de reconstrucción filogenética*

La mayoría de los diferentes métodos de inferencia filogenética propuestos por diversos autores definen un **criterio de optimización** determinado que persigue elegir el mejor árbol de entre todos los posibles que podrían explicar los datos de partida. Este criterio da diferentes valores a cada árbol posible. Este valor es el que se usa para comparar los diferentes árboles. Existen diferentes **algoritmos** que permiten computar dichos valores e identificar el mejor árbol de acuerdo al criterio de optimización.

En la actualidad disponemos de los siguientes métodos de inferencia filogenética: a) métodos basados en matrices de distancias genéticas; b) método de máxima parsimonia; c) método de máxima verosimilitud; d) método bayesiano.

#### *Métodos basados en matrices de distancias*

Existen varios métodos de reconstrucción de árboles filogenéticos basados en matrices de distancias genéticas. En todos ellos, lo primero que se debe hacer es construir dicha matriz de distancias. Para ello se estiman las diferencias entre cada par de secuencias del alineamiento. La forma más simple de calcular la distancia genética es calculando el número

de diferencias (*p*) entre las secuencias. Sin embargo, si *p* tiene un valor alto (las secuencias han divergido considerablemente) puede ocurrir que, en cada sitio del alineamiento se hayan producido substituciones múltiples y reversiones de tal forma que *p* nos estará dando un valor subestimado del número de substituciones nucleotídicas ocurridas realmente. Por lo tanto, se han desarrollo un número amplio de métodos de cálculo de distancias corregidas basados en modelos probabilísticos. Los cálculos de dichas distancias son valores corregidos de *p* según dichos modelos. Cada modelo asume un patrón evolutivo diferente con respecto a composición nucleotídica y tasas de cambio para cada tipo de substitución nucleotídica, para cada posición nucleotídica y para cada linaje. Más adelante, cuando estudiemos los métodos de máxima verosimilitud, volveremos a hablar de estos modelos.

Una matriz de distancias típica tiene esta apariencia:

| | Especie 1 | Especie 2 | Especie 3 | Especie 4 | Especie 5 |
|---|---|---|---|---|---|
| Especie 1 | | 0,012 | 0,018 | 0,022 | 0,035 |
| Especie 2 | | | 0,013 | 0,020 | 0,032 |
| Especie 3 | | | | 0,021 | 0,033 |
| Especie 4 | | | | | 0,020 |

Los valores de distancias de esta matriz son los que se utilizan para reconstruir el árbol, siendo la longitud de las ramas proporcional a dichos valores. Como se decía al principio, existen diferentes métodos de inferencia basados en distancias, pero el más popular es el **método del vecino más próximo**, conocido normalmente con su denominación en inglés (**Neighbor-joining** o método **N-J**). Este método se basa en un algoritmo que trata de buscar el árbol más corto, es decir, aquel que minimiza la longitud total del árbol, entendida ésta como la suma de las longitudes de todas sus ramas. Primero se identifican las dos secuencias que más se parecen (menor distancia genética hay entre ellas). Es decir, de entre todos los pares de secuencias comparados, se identifican aquellas dos secuencias cuya suma de las longitudes de sus ramas es la menor. Ese par de secuencias constituyen el primer par de "vecinos", conectados a través de un nodo interno. El siguiente paso es considerar a este par como una sola secuencia computándose la distancia media aritmética entre ellas y el resto de secuencias y construyendo una nueva matriz de distancias. A continuación se elige de nuevo el par de secuencias cuya suma de las longitudes de sus ramas es la menor, procedimiento que se continúa hasta que se identifican todos los nodos internos del árbol.
Como ejercicio, se podría tratar de construir manualmente un árbol por este método a partir de la matriz de distancias mostrada más arriba.

*Método de máxima parsimonia*

El método de **máxima parsimonia** persigue construir una filogenia con la topología que requiera el menor número de cambios evolutivos para explicar las diferencias observadas entre las secuencias alineadas. A veces, este criterio lo cumplen dos o más árboles que serán

igualmente parsimoniosos. Para aplicar este criterio, cada uno de los sitios nucleotídicos de la secuencia se clasifica de la siguiente manera:

-**Invariable**: todas las secuencias presentan el mismo nucleótido en dicha posición.
-**Informativo**: un sitio es filogenéticamente informativo desde el punto de vista de la máxima parsimonia cuando hay al menos dos clases diferentes de nucleótidos, cada uno representado al menos dos veces en el alineamiento.
-**No informativo**: un sitio que, siendo variable, no cumple el anterior requisito.

Una vez clasificados los sitios del alineamiento e identificados los sitios informativos, para cada árbol posible se calcula el número mínimo de sustituciones necesarias para explicar cada sitio informativo. Sumando el número de cambios para el conjunto de todos los sitios informativos para cada árbol posible, se elegirá aquel árbol que se explique con el menor número de cambios.

Si hay más de un árbol con ese número, se puede obtener un **árbol consenso**, del que podemos distinguir: a) consenso estricto (*strict consensus*), en el que todas las ramas conflictivas se resuelven colapsándolas a un único nodo multifurcado; b) consenso por la regla de la mayoría (*majority-rule consensus*) en el que las ramas en conflicto se resuelven mediante la selección del patrón de ramificación observado en más del 50% de los árboles obtenidos.

*Método de máxima verosimilitud*

La verosimilitud, *L*, de un árbol filogenético es la probabilidad de que los datos observados en un alineamiento se puedan explicar a partir de esa filogenia construida según un modelo evolutivo de substitución nucleotídica determinado, es decir, ***L* =P(datos|árbol+modelo)**. El objetivo del método de máxima verosimilitud es encontrar el árbol con el mayor valor de *L*, de entre todos los árboles posibles que explicarían los datos observados.

La pregunta que hay que plantearse es: ¿Cuál es la probabilidad de que una filogenia determinada haya generado los datos observados en un alineamiento asumiendo un determinado modelo evolutivo de substitución nucleotídica?

Para responder a la pregunta, asumiendo que cada sitio del alineamiento evoluciona independientemente, hay que calcular *L* para cada sitio separadamente ($L_n$) y en conjunto ($L = L_1 \times L_2 \times L_3 \times ... \times L_n$). Para calcular cada $L_n$ se deben considerar todos los posibles escenarios a través de los cuáles se ha llegado al nucleótido actual en cada secuencia a partir de un nucleótido ancestral. Algunos escenarios serán más plausibles que otros pero todos tendrán al menos alguna probabilidad de ser los que han generado la situación actual. Por tanto, cada $L_n$ tiene una probabilidad que es igual a la suma de las probabilidades de cada posible reconstrucción filogenética que explique los datos actuales desde la situación ancestral. Estas probabilidades dependen del modelo evolutivo que asumamos y de la longitud de las ramas la cual, a su vez, depende de la tasa de sustitución y del tiempo evolutivo. Por conveniencia, la verosimilitud se calcula mediante transformación logarítmica ($\ln L$) con lo que tendremos que $\ln L = \ln L_1 + \ln L_2 + \ln L_3 + ... + \ln L_n$.

Un árbol filogenético inferido por este método solo es válido para el modelo evolutivo asumido pero puede no ser válido para otro modelo evolutivo. Por ello, es fundamental una correcta elección del modelo evolutivo aplicable a las secuencias analizadas. Existen diferentes modelos evolutivos que tratan de explicar el patrón de sustitución nucleotídica que siguen las secuencias analizadas. Desde un modelo general en el que se asume que cada tipo de sustitución nucleotídica tiene una tasa diferente y que cada nucleótido aparece en la secuencia en una proporción diferente hasta un modelo más simple en el que asumimos que

todos los nucleótidos aparecen con la misma frecuencia (25% para cada uno de los cuatro nucleótidos) y existe una misma tasa de cambio para todos los tipos de sustitución nucleotídica. Pasando por diferentes modelos en los que se tienen en cuenta las diferencias en la proporción de nucleótidos o no y se consideran de manera diferenciada los diferentes tipos de substituciones nucleotídicas (diferentes tipos de transiciones y de transversiones). Además cada modelo puede asumir que las tasas de cambio difieren entre sitios nucleotídicos del alineamiento diferentes o entre linajes de la filogenia diferentes. Se hace, por tanto, necesario testar qué modelo evolutivo se ajusta mejor a las secuencias analizadas.

*Método de inferencia bayesiana*

La inferencia bayesiana de una filogenia está basada en una cantidad llamada *probabilidad posterior de árboles de distribución*, que es la probabilidad de un árbol condicionado por las observaciones [**P(árbol+modelo|datos)**]. El condicionamiento se logra a través del teorema de Bayes. No es posible calcular analíticamente la probabilidad posterior de árboles de distribución. A cambio, se utiliza una técnica de simulación llamada Monte Carlo de cadena de Markov (MCMC) para aproximar esta probabilidad.

**Fiabilidad de la reconstrucción filogenética**

Para responder a la pregunta que nos podamos hacer con respecto al árbol obtenido por cualquiera de los métodos existentes sobre la fiabilidad del mismo existen métodos que nos permiten estimar el soporte estadístico de la topología obtenida. Uno de los más populares es el método de re-muestreo con re-emplazamiento o ***bootstrap***. Una vez obtenido un árbol filogenético a partir de un alineamiento de secuencias y con un método determinado, esta filogenia se convierte en la hipótesis nula a comprobar mediante *bootstrap*. Para ello, se construyen nuevos alineamientos diferentes (un número apropiado podría ser entre 500 y 1000) mediante re-muestreo con re-emplazamiento. Es decir, se construyen diferentes alineamientos al azar re-emplazando un número determinado de posiciones nucleotídicas con otras posiciones del alineamiento, cada una de las cuales tiene la misma probabilidad de re-emplazar a las demás. Por tanto en el nuevo alineamiento, un sitio puede estar repetido más de una vez a costa de otros sitios. Así, si el alineamiento tiene esta secuencia de posiciones nucleotídicas:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Los diferentes re-muestreos pueden dar lugar a alineamientos como éstos:

1 1 1 4 4 6 6 6 6 10 11 12 12 12 12 12 17 18 19 20 20 20 20 24 24

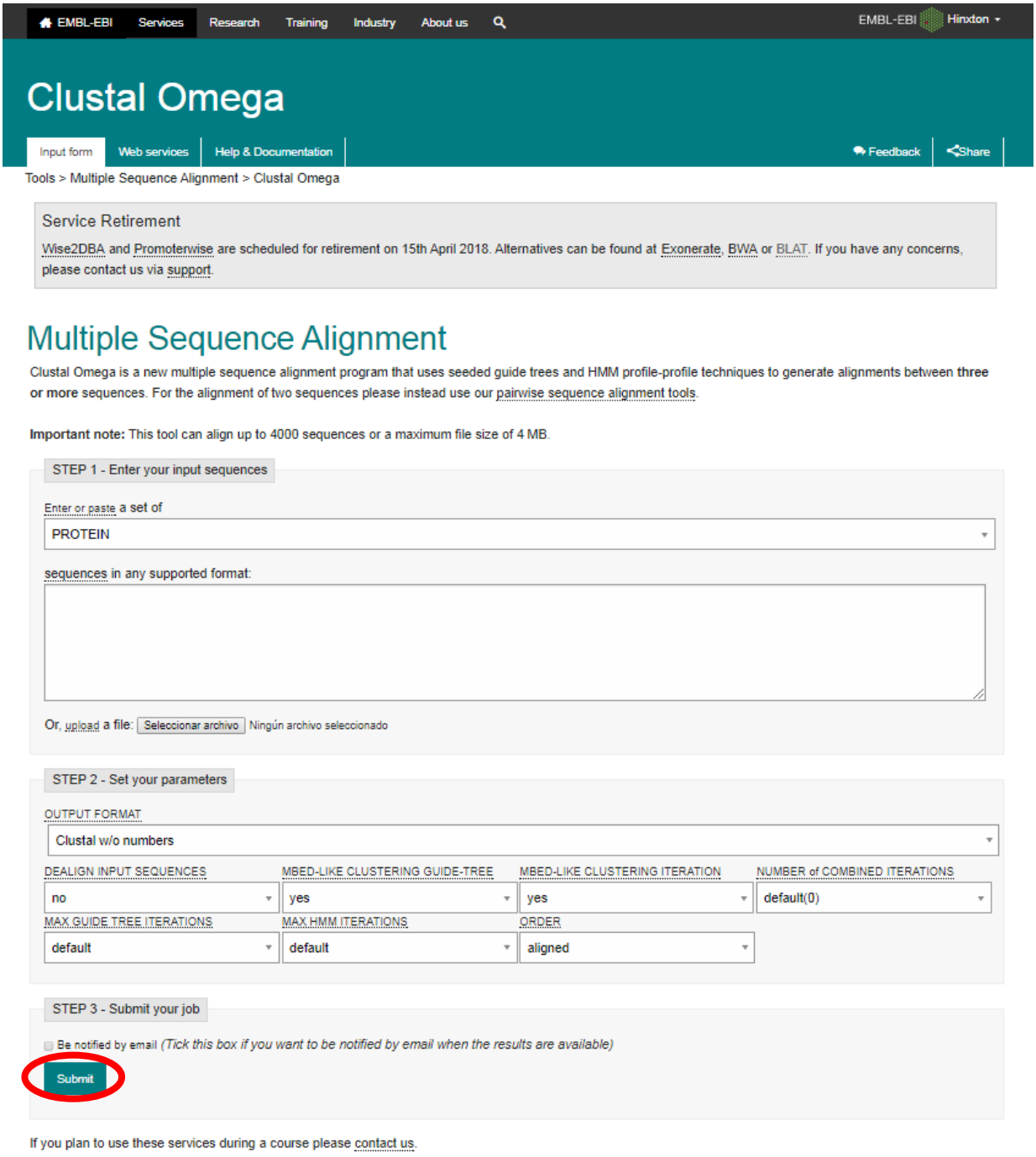
1 1 3 3 3 6 7 8 8 8 8 13 14 15 15 15 15 19 19 19 25 25 25 25

2 2 2 4 5 5 5 8 9 10 10 12 12 12 16 16 17 19 20 21 21 24 24 24 25

A partir de cada uno de los nuevos alineamientos se infiere un nuevo árbol filogenético utilizando el mismo método utilizado con el alineamiento inicial. El porcentaje de veces que cada rama interior del árbol inicial se confirma en el conjunto de los árboles obtenidos por *bootstrap*, constituye el valor de *bootstrap* de cada rama. Como regla general, si el valor de *bootstrap* de una rama interior determinada es superior al 95%, se acepta que la topología de esa rama es correcta.

## 5.3. METODOLOGÍA

La práctica se va a llevar a cabo ejecutando online el programa Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/)




En primer lugar, hay que seleccionar en "STEP 1" el tipo de secuencia que se va a analizar (protein, DNA o RNA). En la caja, cargar las secuencias a analizar (pegar las secuencias o subir el archivo, en formato FASTA en cualquier caso), dejando las opciones de análisis por defecto que aparecen en "STEP2". Pulsar en "Submit" para llevar a cabo el alineamiento múltiple de las secuencias.

Cómo se ha indicado anteriormente, el alineamiento se realiza en un proceso de tres etapas:

1.- Se comparan todas las secuencias dos a dos (alineamientos *pairwise*).

2.- Se construye un dendrograma (similar en aspecto a un árbol filogenético) que agrupa las secuencias por similitud.

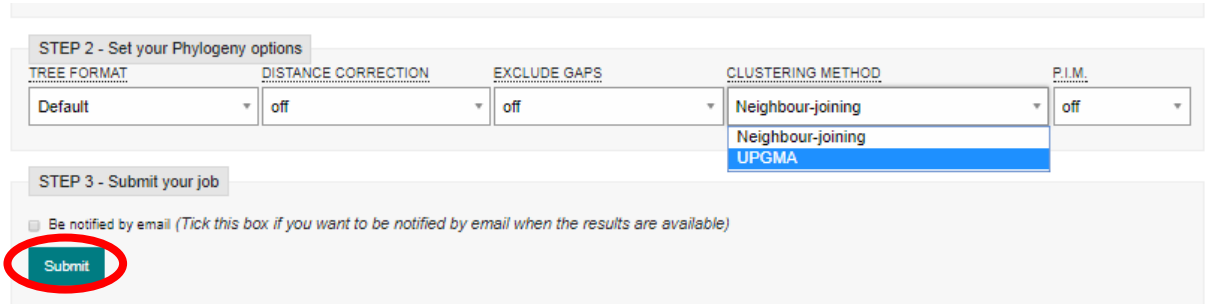
3.- Se utiliza el dendrograma como guía, alineando secuencias de manera progresiva de acuerdo al orden de ramificación del árbol. Se alinean las dos secuencias con mayor similitud y se van añadiendo secuencias al alineamiento de manera progresiva por orden de similitud decreciente.

El programa muestra la mejor combinación de alineamiento de las secuencias, mostrando las similitudes y diferencias.

El símbolo "*" indica posiciones del alineamiento en las que en todas las secuencias existe el mismo residuo (nucleótido o amino ácido).

En el caso de alineamiento de secuencias de proteínas, el símbolo ":" indica que uno de los siguientes grupos "fuertes" de amino ácidos con propiedades químicas similares está muy conservado: STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, FYW, mientras que el símbolo "." indica que uno de los siguientes grupos "débiles" está muy conservado: CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, HFY.

Una vez obtenido el alineamiento múltiple, el programa ofrece diferentes opciones, entre ellas descargar el alineamiento o llevar a cabo un análisis filogenético sencillo (*Send to simple_phylogeny*). En esta opción, se elegirá el análisis mediante UPGMA:




Por último, pulsar en "Submit" para obtener el árbol de secuencias.

## 5.4. EJERCICIOS Y CUESTIONES

A continuación, se plantean ejemplos y ejercicios utilizando secuencias de nucleótidos y proteínas. Se llevará a cabo un alineamiento múltiple (alineamiento progresivo) de las secuencias homólogas. A partir de los datos de comparación de secuencias, se obtendrá una matriz de distancias, y a partir de la matriz de distancias, se obtendrá un dendrograma mediante un método algorítmico como es UPGMA (Unweigtthed Pair Group Method with Aritmetic mean).

Algunos conceptos clave en este tipo de análisis son:

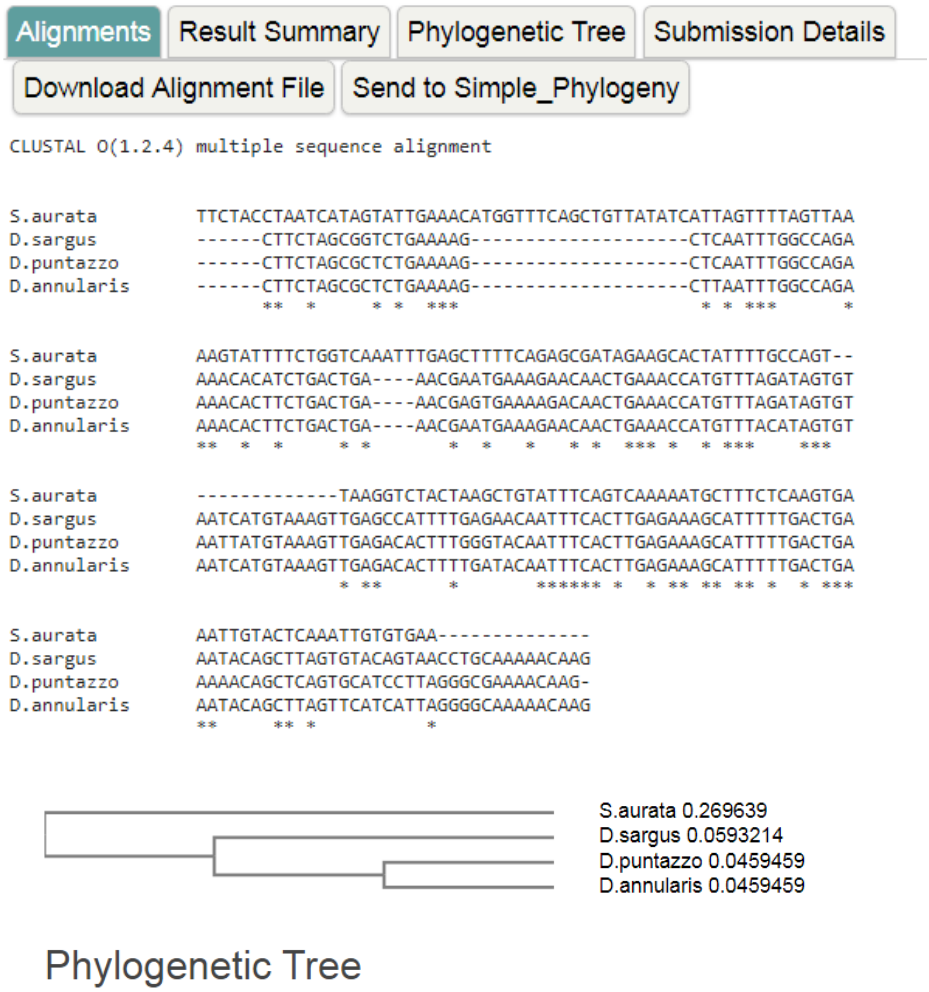
- La reconstrucción filogenética basada en distancias que llevaremos a cabo está fundamentada en que la distancia entre taxones es reflejo de la relación filogenética entre ellos.
- La distancia evolutiva es la media de cambios que se han producido en una posición entre dos pares de secuencias a lo largo de su evolución desde su ancestro común.
- La filogenia molecular es una estimación de las relaciones filogenéticas basada en la comparación de secuencias de ADN o proteínas pertenecientes a dichos taxones.

**<u>Ejemplo alineamiento secuencias nucleótidos:</u> Alineamiento múltiple de cuatro secuencias pertenecientes a la familia de ADN satélite EcoRI en cuatro especies diferentes de peces espáridos.**

Utilizando el fichero de secuencias de esta familia de ADN repetido en formato FASTA disponible en la plataforma PRADO2 de la asignatura, obtener el correspondiente alineamiento múltiple y el árbol filogenético utilizando el programa Clustal Omega *online*.

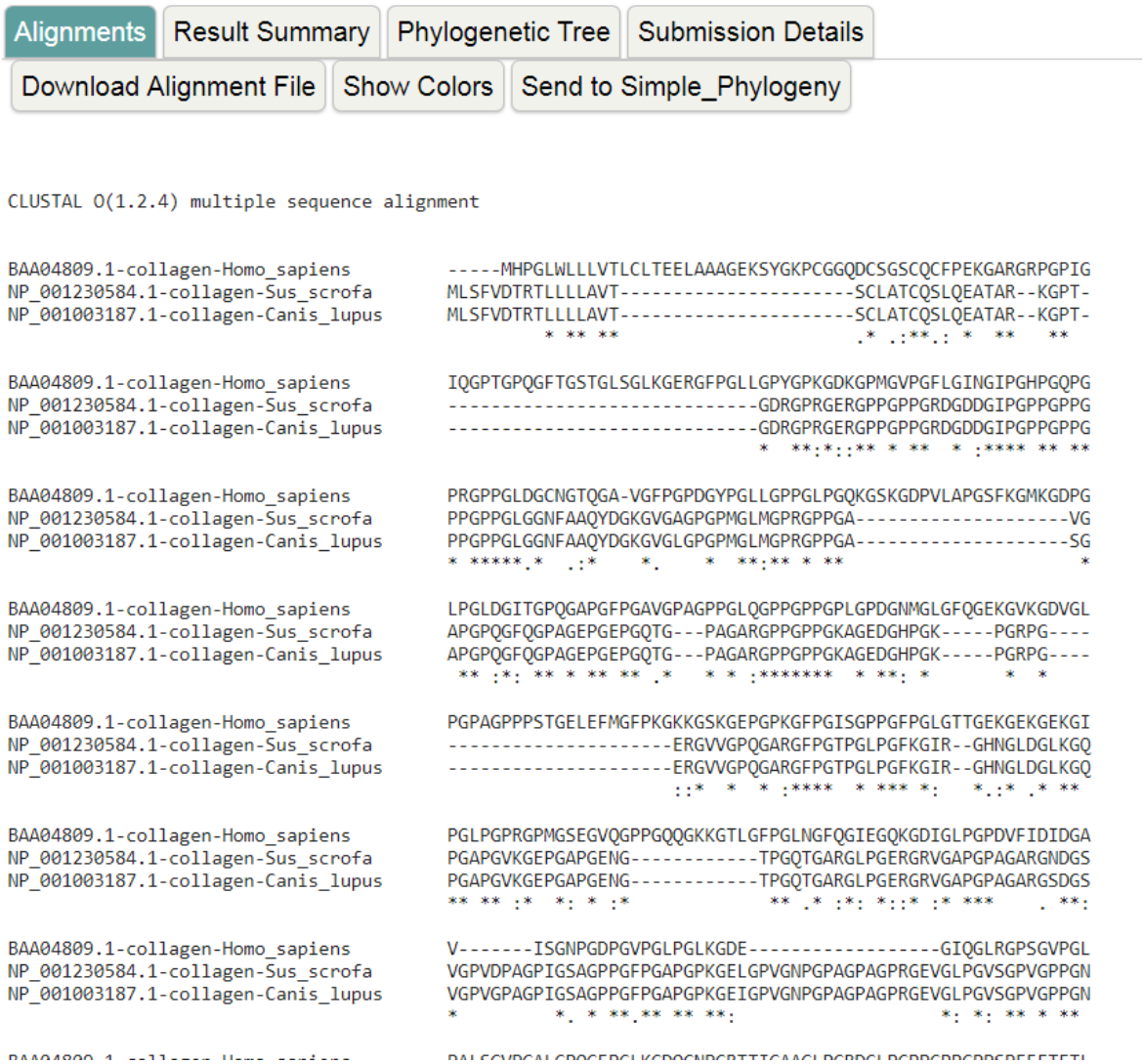
Para ello, sigue la metodología descrita en el apartado anterior para el programa Clustal Omega *online* (https://www.ebi.ac.uk/Tools/msa/clustalo/).

Debes obtener un alineamiento y un árbol de secuencias similares a los de la siguiente figura:



**<u>Ejemplo alineamiento secuencias amino ácidos:</u> Alineamiento múltiple de tres secuencias pertenecientes al colágeno.**

Utilizando ahora el fichero de secuencias del colágeno, debes obtener un alineamiento similar al de la siguiente figura parcial:

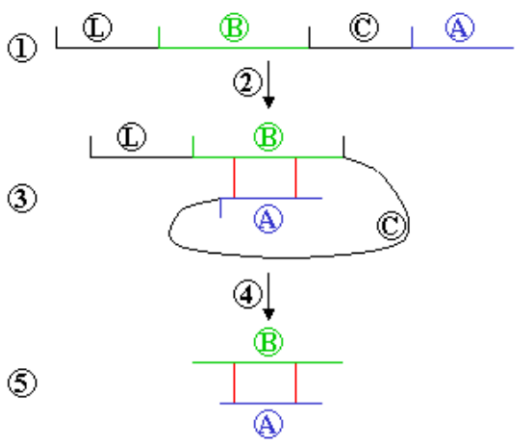



**<u>Ejercicio 1</u>: Búsqueda de secuencias homólogas en la base de datos del NCBI y realización de un alineamiento múltiple perteneciente al gen de la insulina de diferentes especies.**

Busca los siguientes *accession numbers* en la base de datos, tal y como se explicó en la práctica correspondiente a bases de datos de secuencias de ADN y proteínas (http://www.ncbi.nlm.nih.gov/genbank), y realiza un alineamiento múltiple con estas secuencias.

| | | |
|---|---|---|
| AAP36446 | NP_776351 | P01318 |
| NP_001008996 | NP_032413 | P01315 |
| NP_001123565 | NP_062003 | |

A continuación, compara el alineamiento múltiple obtenido con el siguiente esquema:




L: 1-24
B: 25-54
C: 55-89
A: 90-110

1.Preproinsulina (Lguía, B cadena, C cadena, A cadena); proinsulina consiste BCA, sin L
2.plegamiento espontáneo
3.Cadenas A y B unidas por puentes sulfuros
4.Guía y la cadena C son cortadas
5.Restos de insulina
http://es.wikipedia.org/wiki/Insulina

Ahora, incluye en el alineamiento anterior las siguientes secuencias:

AAA37041
XP_006008147
XP_006033708
NP_990553

Y por último, establece las relaciones de parentesco entre las especies realizando un árbol filogenético tal y como se ha explicado en el ejemplo anterior (*Send to simple_phylogeny;* UPGMA; Submit). Explica y discute los resultados obtenidos.

**Ejercicio 2: Búsqueda de secuencias homólogas relacionadas con la primera práctica de laboratorio de la asignatura.**

Busca en la base de datos del NCBI las secuencias correspondientes a los siguientes *accesión numbers*:

AF509333
AY305326
AF497479

a) ¿A qué organismo y a qué tipo de secuencia hace referencia cada *accession number*?

b) Realiza un alineamiento múltiple con ellas.

c) Explica y discute los resultados del alineamiento

**Ejercicio 3: Búsqueda de secuencias homólogas de ADN ribosómico**

Realiza un alineamiento múltiple y un árbol filogenético con los siguientes *accession numbers* correspondientes a secuencias del gen de ADN ribosómico 18S en diferentes especies:

| | | |
|---|---|---|
| L11288 | X00686 | AF173630 |
| AF173605 | NR_033238 | AF173611 |
| AF115860 | AF173614 | AF173612 |

**<u>Ejercicio 4</u>: Identificación de secuencia de ADN por similitud filogenética sobre el árbol del ejercicio anterior**

Utilizando el fichero llamado "Secuencia problema" disponible en la plataforma PRADO2 de la asignatura, que contiene la secuencia del gen de ADN ribosómico 18S de una especie sin identificar, y teniendo como base el alineamiento múltiple y el árbol de secuencias obtenido en el ejercicio anterior, trata de determinar, en la medida de lo posible, a qué organismo pertenece.

# 6. EXPRESIÓN DE GENES IMPLICADOS EN EL DESARROLLO TESTICULAR DE MAMÍFEROS

## 6.1. OBJETIVO

Que el alumno aprenda un método, basado en el diagnóstico molecular, que es usado habitualmente para el sexado de embriones de mamíferos así como a identificar órganos embrionarios en los que se expresa el gen *SOX9.*

## 6.2. FUNDAMENTO TEÓRICO

***Determinación genética del sexo en mamíferos***

En mamíferos, la presencia de un cromosoma Y determina el sexo masculino, mientras que su ausencia implica un desarrollo femenino. Al inicio del desarrollo embrionario, la gónada es indiferenciada y bipotencial, lo que significa que puede seguir dos rutas de desarrollo alternativas y, en condiciones normales, mutuamente excluyentes: testículo u ovario. En la gónada embrionaria XY, el gen *SRY* (localizado en el cromosoma Y; * ver nota sobre la tipografía correcta de los genes de mamíferos al final de este guión) inicia una cascada de activación génica que induce a una sub-población de células somáticas a diferenciarse como células de Sertoli, encargadas de orquestar el desarrollo testicular. Estas células de Sertoli se organizan formando cordones sexuales (precursores de los túbulos seminíferos del testículo adulto) en el interior de los cuales se localizan las células germinales que dejan de proliferar (arresto mitótico). Las células de Sertoli controlan también la diferenciación de células de Leydig, células secretoras de testosterona y dihidrotestosterona que masculinizarán el soma del individuo. En la ruta masculina de desarrollo gonadal de ratón, la proteína SRY se une, junto con el factor esteroidogénico SF1, a una secuencia intensificador del gen *Sox9* y lo activa. Las mutaciones en que el gen *Sox9* se activa en una gónada XX, hacen que ésta siga la ruta testicular, mientras que si este gen permanece inactivo en una gónada XY, ésta seguirá la ruta ovárica. Por tanto, *Sox9,* al igual que *Sry,* son necesarios y suficientes para activar la organogénesis testicular. SOX9 activa el gen *Fgf9* que a su vez estabiliza la expresión de *Sox9,* estableciéndose un bucle de automantenimiento de la expresión de éste último en la gónada masculina. *SOX9* activa también la expresión de otros genes como *Amh* (hormona antimülleriana), *Vnn1* (Vanin-1), y *Pgds* (prostaglandina sintetasa) que se sabe están implicados en la diferenciación testicular. Sobre la base de lo expuesto, se puede decir que *SOX9* es el gen alrededor del cual pivota el desarrollo testicular, y lo hace no sólo en mamíferos sino en todos los vertebrados.

En la gónada XX la ausencia de cromosoma Y, y por tanto del gen *SRY,* implica la inactividad de *SOX9* y la activación de *RSPO1* y *WNT4,* que inician la cascada de activación génica que conduce al desarrollo ovárico. Al no expresarse el gen *Sox9*, las células somáticas bipotenciales de la gónada embrionaria se diferencian como células pre-foliculares (no como células pre-Sertoli), mientras que las células de la línea esteroidogénica se diferenciarán como células de la teca (en vez de como células de Leydig) y las células germinales inician la meiosis, que se detiene poco después en la profase I (arresto meiótico). En resumen, en ausencia de *Sry*, la organogénesis gonadal sigue la ruta ovárica y el fenotipo somático del individuo será femenino.
La visión clásica acuñada por Jost (1953) de que la ruta ovárica es la ruta constitutiva, cambió sobre la base de nuevos datos en los que se describió la reversión sexual parcial o

total de individuos XX de ratón, que presentaron mutaciones de pérdida de función en genes como *Wnt4* y *Rspo1*. Los individuos XX *Wnt4*-/- (homocigotos para el alelo mutado) mostraron gónadas parcialmente masculinizadas y expresión de los genes *Sox9 y Fgf9*, diferenciación de células de Leydig, migración celular desde el mesonefros adyacente hacia el interior de la gónada (evento morfológico específico de la gónada XY) y desarrollo de un patrón vascular específico de testículo. La mutación de pérdida de función en el gen *RSPO1* provoca una reversión sexual completa de hembra a macho, es decir machos XX. Este fue el primer caso descrito de una única mutación en un gen que provoca reversión sexual completa de hembra a macho y esta mutación sitúa a *RSPO1* como el probable determinante ovárico en mamíferos. *RSPO1* activaría los genes implicados en el desarrollo ovárico e inhibiría directa o indirectamente los genes implicados en la ruta testicular de desarrollo gonadal. Otro gen que interviene en la ruta ovárica es *FOXL2,* necesario para el desarrollo y mantenimiento de la estructura ovárica**.** La ausencia de células de la granulosa funcionales conlleva la iniciación prematura de la foliculogénesis y un fallo ovárico prematuro. Sin embargo, la ausencia de reversión sexual de hembra a macho de ratones mutantes *Foxl2*-/- indica que no es un determinante ovárico. En esta práctica vamos a amplificar un fragmento del gen *Sry*, y comprobaremos que está presente en células masculinas (XY), mientras que las células femeninas (XX) carecen de dicho gen.

### *SOX9: Un gen pleiotrópico*

El gen *SOX9* fue inicialmente identificado como el gen responsable del síndrome displasia campomélica (DSCM), una malformación del esqueleto asociado con reversión sexual XY. *SOX9* es un factor de transcripción perteneciente a la familia de proteínas SOX (Sry-like HMG box). En humanos se encuentra localizado en la región cromosómica 17q24.3-q25.1 y está compuesto por tres exones y dos intrones. *SOX9* se expresa en un gran número de tejidos embrionarios entre los que se incluye condrocitos, células de Sertoli, células de la placoda ótica, células pancreáticas, células del epitelio intestinal, células de la cresta neural, células del epitelio pulmonar, células de la notocorda y varios tejidos más. Esto sugiere que *SOX9* tiene múltiples funciones durante el desarrollo embrionario de mamíferos, y para poner de manifiesto el papel que *Sox9* tiene en el desarrollo de los diferentes órganos en que se expresa se han generado ratones mutantes para este gen. En el ratón, este gen está localizado en el cromosoma 11. El primer ratón mutante para *Sox9* fue descrito en 2001. Estos ratones mutantes heterocigóticos para *Sox9* reproducían la mayor parte de las malformaciones del esqueleto mostradas por los pacientes con DSCM, aunque otras anomalías, como la reversión sexual no se ponían de manifiesto. Los ratones mutantes heterocigóticos para *Sox9* morían alrededor del nacimiento, por lo que no era posible generar ratones mutantes homocigóticos. Debido a esto último, se generaron ratones mutantes condicionales para los diversos tejidos donde *Sox9* se expresa, es decir, animales que sólo carecen de la función del gen en tejidos u órganos concretos. Así,*Sox9* ha sido inactivado condicionalmente en homocigosis en condrocitos, lo que provocó la ausencia completa de cartílago y huesos. Los embriones con *Sox9* inactivado en condrocitos exhibían una condrodisplasia generalizada. *Sox9* también ha sido inactivado condicionalmente durante el desarrollo testicular de ratón. En estos ratones se observó que los individuos XY se desarrollaban fenotípicamente como hembras que tenían ovarios en lugar de testículos. A pesar de ello, el gen determinante de testículo, *Sry*, continuaba expresándose indicando que *Sox9* actúa posteriormente en la cascada génica que regula el desarrollo testicular. La inactivación condicional homocigótica de *Sox9* en ratón ha mostrado que también es necesario para la diferenciación de las células gliales de la espina dorsal, la formación de la válvulas y el tabique cardiaco, el desarrollo de la notocorda, el mantenimiento de la células madre pancreáticas, la invaginación de la placoda ótica, el desarrollo de la próstata, la supervivencia de las células de la cresta neural y el mantenimiento de la espermatogénesis. La segunda parte de esta práctica va a consistir en la observación de cortes histológicos a los que se ha realizado una inmunohistoquímica con un anticuerpo anti-SOX9.

# 6.3. METODOLOGÍA

### 6.3.1. Detección del gen *Sry* mediante PCR

Para la detección del gen *Sry* haremos uso de la técnica PCR (*Polimerase Chain Reaction*). Para ello, hemos diseñado cebadores específicos, por un lado del gen *Sry*, que se encuentra en el cromosoma Y, por lo que es específico de machos, y por otro lado del gen de la Miogenina, gen autosómico que nos va a servir como control positivo. Haremos una "PCR duplex", es decir, una PCR en la que en una única reacción los cebadores de ambos genes están presentes, y por lo tanto podemos amplificar simultáneamente los fragmentos correspondientes a los dos genes. Las secuencias de los cebadores son las siguientes:

-*Oligonucleótidos para la amplificación del gen Sry de ratón:*

Sry-F 5'- GCA AAC AGC TTT GTG GTC AA 3'
Sry-R 5'- GGA AAA GGG GAT GAA ATG GT 3'

-*Oligonucleótidos para la amplificación del gen de la Miogenina de ratón:*

Mio-F 5'- TTA CGT CCA TCG TGG ACA GCA T 3'
Mio-R 5' TGG GCT GGG TGT TAG CCT TAT G 3'

***Reacción de Amplificación (PCR)***

En un microtubo de 200µl añadir, siguiendo el orden indicado, los siguientes reactivos para un volumen final de 25µl:

- Agua estéril 16,5 µl
- 10% Tampón de PCR (10x) 2,5 µl
- $MgCl_2$ (25mM) 1,5 µl
- DMSO 1,2 µl
- Primer Sry-F (500 ng/µl) 0,5 µl
- Primer Sry-R (500 ng/µl) 0,5 µl
- Primer Mio-F (500 ng/µl) 0,5 µl
- Primer Mio-R (500 ng/µl) 0,5 µl
- dNTPs (25 mM) 0,2 µl
- Taq polimerasa (2U) 0,1 µl
- ADN (100 ng/µl) 1 µl

A continuación se colocan los microtubos en el termociclador y se programa para 35 ciclos según el programa:

| | | |
|---|---|---|
| Desnaturalización: | 91ºC | 45 seg. |
| Alineamiento: | 60ºC | 60 seg. |
| Extensión: | 72ºC | 45 seg. |

Una vez terminada la PCR, se someterán las muestras a una electroforesis en gel de agarosa y se procederá al sexado de los individuos.

Tras la reacción de PCR, se amplificará, en el caso del gen *Sry,* un fragmento de 179 pb, y en el caso del gen de la Miogenina un fragmento de 246 pb. Ambos amplicones se pueden separar perfectamente mediante una electroforesis en gel de agarosa, que se realizará a continuación. En el caso de un macho se distinguirán ambas bandas, mientras que en el caso de una hembra sólo se apreciará la banda de 246 pb. Si no se observara ninguna banda indicaría que la PCR no ha funcionado (Figura 1).

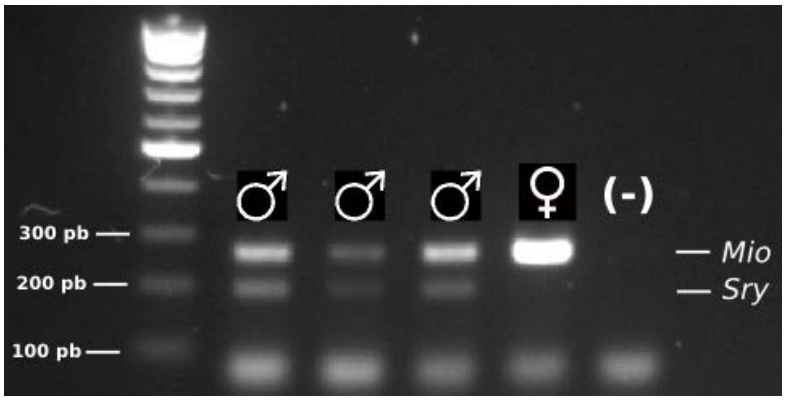



Figura 1: Electroforesis de los productos de una PCR realizada para el sexado de embriones de ratón. La presencia en el gel de una banda correspondiente al gen Sry denota la presencia de un macho, mientras que su ausencia indica que ese embrión es hembra. La banda de la miogenina sirve como control de calidad (control positivo) de la reacción de PCR. (-) es el control negativo (reacción sin molde), que indica la ausencia de contaminación de ADN en la mezcla de reacción de la PCR.

### 6.3.2. Observación de preparaciones de inmunohistoquímica para SOX9

Actualmente existen varias técnicas para detectar la expresión de genes en tejidos. Una de estas técnicas es la inmunohistoquímica, que nos permite identificar el tipo celular donde se localiza una proteína de interés, situación que en la mayoría de los casos implica que el gen que codifica para dicha proteína se está expresando en ese tipo celular. En una técnica immunohistoquímica, la localización de la proteína de interés se pone de manifiesto mediante una reacción enzimática, siendo la catalizada por la peroxidasa de rábano una de las más usadas en la actualidad. Una de las formas de realizar una inmunohistoquímica mediante el método de la peroxidasa consiste en fijar el tejido de interés, deshidratarlo, incluirlo en parafina, y realizar cortes histológicos. Tras desparafinar e hidratar los cortes histológicos, se incuban con una solución que contiene el anticuerpo primario, específico de nuestra proteína de interés. En esta situación, en aquellas células donde la proteína de interés esté presente, se producirá la unión entre la proteína de interés y el anticuerpo primario. Dado que la proteína de interés está fijada en el interior de la célula, el complejo también permanecerá en el interior celular. Posteriormente se lavan intensamente las preparaciones para eliminar el anticuerpo primario que no se ha unido a la proteína de interés, y se vuelve a incubar con una solución que contiene un anticuerpo secundario, que es un anticuerpo específico contra la inmunoglobulina G de la especie donde se generó el anticuerpo primario. El anticuerpo secundario está conjugado con la peroxidasa de rábano (anti-Ig-Peroxidasa). Esto hace que se forme un complejo entre la proteína de interés, el anticuerpo-primario y el anticuerpo secundario conjugado, que permanece en el interior de las células donde esté presente la proteína de interés. Después, se vuelven a lavar las preparaciones para eliminar el anticuerpo secundario libre y se incuba con una solución que contiene $H_2O_2$ y di-amino bencidina (DAB). La peroxidasa cataliza la reacción $2H_20_2$ ---> $2H_20 + O_2$. Esto hace que se libere $O_2$ en el interior de aquellas células donde está retenido el complejo que oxida a la DAB, dando lugar a un precipitado marrón (Figura 2).

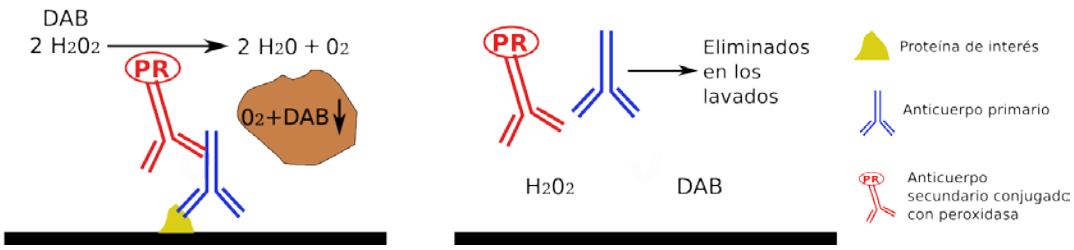



Figura 2: Fundamento de la técnica de inmunohistoquímica. La presencia en la muestra de la proteína de interés (esquema de la izquierda) permite el anclaje a la preparación del complejo compuesto por el anticuerpo primario, el anticuerpo secundario y la peroxidasa, permitiendo la reacción coloreada con DAB. Su ausencia (esquema de la derecha) permite el lavado de todos los componentes, no habiendo reacción alguna.

Finalmente se hace una contra-tinción de las preparaciones con hematoxilina, se deshidratan y se montan con DePeX para su observación al microscopio óptico. Tras este proceso, observaremos las células positivas para la proteína de interés de color marrón, mientras que el núcleo de las células negativas se ve de color azul (hematoxilina; Figura 3)

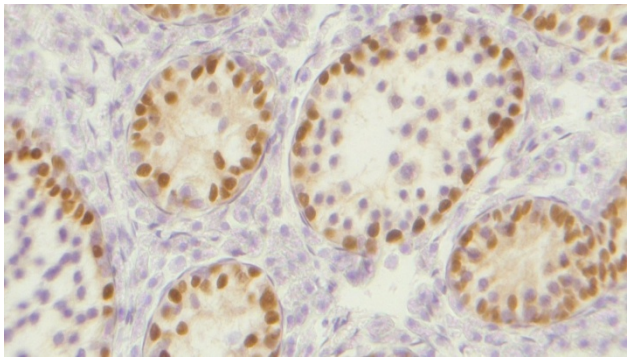


Figura 3: Marcaje inmunohistoquímico de tejido testicular de ratón, usando un anticuerpo primario anti-Sox9. Sólo las células de Sertoli aparecen marcadas con el color marrón. El resto de las células se muestran azul claro por la contra-tinción realizada con hematoxilina.

En esta práctica se suministrarán a los alumnos preparaciones inmunohistoquímicas, realizadas mediante el método de la peroxidasa, para la proteína SOX9 en embriones en el estadio embrionario 12.5 de ratón (E12.5). Dado que *Sox9* es un gen pleiotrópico, su expresión se detectará en diferentes tejidos y órganos embrionarios. El objetivo de esta práctica consistirá en identificar la presencia o ausencia de expresión de este gen en los diferentes órganos y tejidos observados en los cortes de embriones examinados.

## 6.4. RECURSOS WEB

A través del link de YouTube se puede acceder al video-tutorial de la práctica.

https://www.youtube.com/playlist?list=PLBa9sJUx0zXWnO2Wu4H6qmJrEOIFNCzaI

## 6.5. CUESTIONES

**1.** ¿Qué otras técnicas inmunológicas existen en la actualidad para detectar la presencia de una proteína de interés en un tejido?

**2.** ¿Qué ocurre en mamíferos cuando el gen *SRY* está mutado? ¿Y si está translocado al cromosoma X?

**3.** ¿Un gen pleiotrópico tiene la misma función en todos los tejidos donde se expresa? Pon un ejemplo que incluya a *SOX9*.

*NOTA: La nomenclatura correcta de los genes de mamíferos es la siguiente:

Los nombre de los genes se escriben en cursiva con letras mayúsculas (p. ej. *SOX9*), para todas las especies, excepto para el ratón y la rata, en cuyo caso se escriben en cursiva con la primera letra en mayúscula y las demás en minúscula (p. ej. *Sox9*). Los nombres de las correspondientes proteínas siempre se escriben sin cursiva y con mayúsculas (p. ej. SOX9).

# 7. ESTUDIO DE EXPRESIÓN GÉNICA MEDIANTE RT-PCR

## 7.1. OBJETIVO

El objetivo de esta práctica es que el alumno aprenda un método de purificación de ARN y su uso para un estudio de expresión génica mediante la aplicación de la técnica de RT-PCR.

## 7.2. FUNDAMENTO TEÓRICO

***Identificación de la Hormona Anti-Mülleriana***

Los conductos de Müller (o conductos paramesonéfricos) y los conductos de Wolff (o conductos mesonéfricos) son dos estructuras tubulares embrionarias que aparecen lateralmente en el primordio urogenital durante el desarrollo embrionario de mamíferos. En hembras, los conductos de Müller se diferencian en varias estructuras del tracto urogenital femenino: los oviductos (en mujeres se denominan trompas de Falopio), el útero, el cuello del útero y parte superior de la vagina, mientras que los conductos de Wolff degeneran. Por el contrario, en machos, los conductos de Wolff dan lugar a los conductos eferentes, epidídimos y vesículas seminales, degenerando los conductos de Müller.

Las primeras evidencias sobre el mecanismo molecular responsable de la degeneración de los conductos de Müller se obtuvieron en la década de 1940-1950, a partir del trabajo de Alfred José que transplantó tejido testicular en fetos de conejo que previamente habían sido castrados y observó que los conductos de Wolff se diferenciaban en los conductos eferentes, epidídimos y vesículas seminales, mientras que los conductos de Müller degeneraban. Posteriormente, observó que un cristal de propionato de testosterona era capaz de inducir la diferenciación de los conductos de Wolff en los fetos de ratones castrados, pero no afectaban el desarrollo de los conductos de Müller, que formaban los oviductos, el útero, el cuello del útero y parte superior de la vagina. De estos experimentos se dedujo que un factor difusible, producido por el testículo, diferente de la testosterona, era responsable de la regresión de los conductos de Müller en el feto masculino. A este factor lo llamó inicialmente sustancia inhibidora de los conductos de Müller. Sin embargo, la identificación de esta sustancia no resultó ser fácil, y no fue hasta 1984 cuando se pudo purificar y caracterizar. A esta sustancia se la conoce actualmente como Hormona Anti-Mülleriana (AMH), o sustancia inhibidora del conducto de Müller (MIS). Experimentos posteriores confirmaron que la AMH era la responsable de la degeneración de los conductos de Müller. Uno de estos experimentos fue la identificación de esta sustancia como el agente causal del *freemartinismo*, un fenómeno descrito en mamíferos desde principios del Siglo XX. Un *freemartin* es un individuo XX con ovarios no funcionales y con una anatomía reproductiva anormal caracterizada por genitales externos femeninos y genitales internos con un número variable de estructuras fenotípicas masculinas. Los casos de *freemartinismo* siempre se producen cuando un individuo XX tiene un gemelo fraterno XY. Debido a esto se hipotetizó que ciertos factores masculinizantes viajarían desde el feto masculino hasta el feto femenino. De acuerdo con esto último, varios investigadores descubrieron que en los casos de *freemartinismo* el feto femenino en el útero tiene fusionado su corion con el corion de un feto masculino, lo que permite que los vasos sanguíneos estén interconectados. En 1984 se confirmó que la sustancia difusible que viajaba a través de los vasos sanguíneos entre los fetos masculino y femenino en los casos de *freemartinismo* era la AMH**.**

### *AMH y desarrollo testicular*

En mamíferos, la expresión del gen determinante de testículo, *SRY*, en las células pre-Sertoli del primordio gonadal XY, hace que la gónada bipotencial comience a diferenciarse como testículo. Poco después de la expresión del *SRY,* varios genes involucrados en el control de la ruta masculina, como *SOX9* y *SF1* son activados en las células pre-Sertoli, y éstas se diferencian en las células de Sertoli. Las células de Sertoli sufren una transición mesénquima-epitelio y forman los cordones testiculares. Poco después, en el mesénquima que rodea a los cordones testiculares, se diferencian las células de Leydig. Las células de Sertoli son las encargadas de producir la AMH, desde donde es secretada y transportada al mesénquima que rodea al conducto de Müller. En estas células se produce la unión con su receptor, el receptor de tipo II de la hormona Anti-Mülleriana (AMHR2). La unión ligando-receptor (AMH-AMHR2) desencadena una cascada génica conducente a la degeneración del conducto de Müller mediante apoptosis. A su vez, las células de Leydig producen testosterona, que es necesaria para que se produzca el desarrollo de los conductos de Wolff. En la hembra no ocurre la diferenciación de las células de Sertoli, por lo que no se produce AMH y el conducto de Müller no degenera. Tampoco se diferencian las células de Leydig, por lo que no se produce testosterona y la falta de desarrollo del conducto de Wolff impide que se formen los órganos sexuales masculinos (Figura 1).

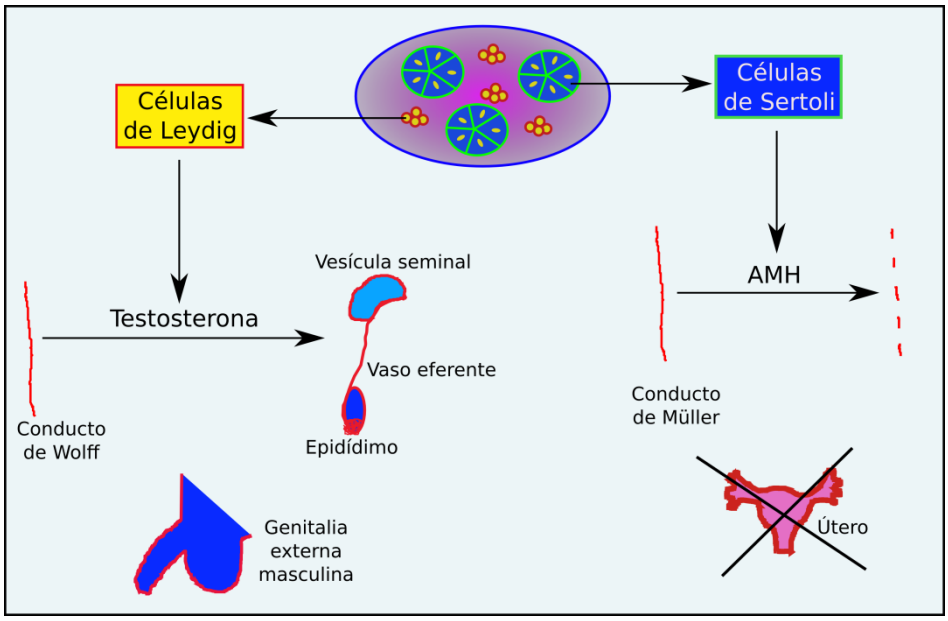



Figura 1: Esquema del desarrollo testicular. La producción hormonal de los testículos incluye testosterona y AMH

Una vez que estos eventos han tenido lugar y transcurre algún tiempo, la función de las células de Sertoli cambia durante la pubertad, cuando sufren una transformación, tanto morfológica como funcional que las prepara para respaldar el ciclo espermatogénico. En este proceso, conocido como maduración de las células de Sertoli, éstas cambian su morfología, pasando a un estado maduro no proliferativo. Sufren una transformación que las prepara para ejercer sus nuevas funciones. Si este proceso de maduración no tiene lugar, la entrada de las células germinales en meiosis y su posterior transformación en espermatozoides no ocurre. La expresión de AMH continúa en el testículo hasta la pubertad, coincidiendo con la maduración de las células de Sertoli, por lo que su inactivación parece estar asociada con el comienzo de la maduración, aunque no se conoce el mecanismo molecular que controla este proceso. La AMH también se expresa en las células de la granulosa del ovario, comenzando en el periodo post-natal y terminado al comienzo de la menopausia, donde tiene un papel en la regulación de la maduración los folículos ováricos.

### *El gen de la AMH*

La AMH está formada por un homodímero de gluco-proteína de unos 140 KD muy conservada entre diferentes especies. La región carboxi-terminal comparte una gran homología con los miembros de la superfamilia de factores transformantes del crecimiento TGFβ. La AMH humana está codificada por un gen de 2.75 Kbp divididos en 5 exones caracterizados por un alto contenido en GC. La región 5' no traducida es de aproximadamente 10 nucleótidos, mientras que la señal de poliadenilación está a 90 nucleótidos corriente abajo del codón de terminación TGA. En rata, se han descrito dos tipos de ARNm que se diferencian en la longitud de la cola de poli-A. Durante el periodo de diferenciación testicular se ha observado la presencia en el testículo de un ARNm de unos 2.0 Kb, cuya abundancia va disminuyendo en los estadios posteriores de gestación, y en los estadios post-natales prácticamente sólo se detecta un trascrito de unos 1.8 Kb. El promotor de la AMH bovina, de ratón y de rata contiene una caja TATA y un único sitio de iniciación de la transcripción, localizado 10 pb corriente arriba del codón de iniciación ATG. Contrariamente, la AMH humana no contiene una caja TATA o CCAAT, sino que posee un elemento iniciador funcional (Inr), que es específicamente reconocido por el factor de transcripción TFII-I. En la región promotora de humano se han encontrado sitios de unión funcionales para los factores de transcripción *SOX9*, *SF1* y *GATA* (Figura 2)

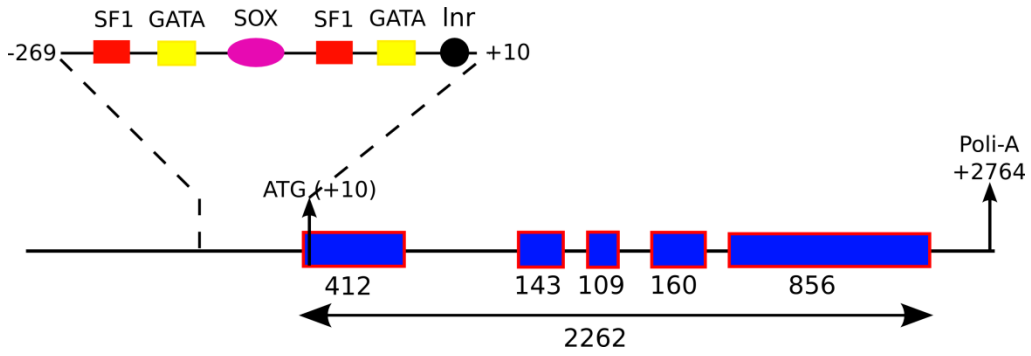



Figura 2: Estructura del gen de la AMH humana.

### *Mutaciones en la AMH*

El síndrome de persistencia del conducto de Müller (PMDS, persistent Müllerian duct syndrom) es una enfermedad rara de origen genético caracterizada por anomalías del tracto reproductor. Los pacientes desarrollan testículos, y en el momento del nacimiento son identificados como varones sin ambigüedad aparente. No obstante, una observación más detallada revela que los pacientes desarrollan anormalidades genitales, entre las que se incluye criptorquidismo, que es un defecto del desarrollo en el que uno (criptorquidismo unilateral) o ambos testículos (criptorquidismo bilateral) no consiguen descender desde el abdomen al escroto. Este descenso testicular al escroto es esencial para la fertilidad masculina, ya que en el escroto los testículos se encuentran a una temperatura menor que la corporal, condición necesaria para el desarrollo normal de la espermatogénesis. Además estos pacientes mantienen estructuras derivadas de los conductos de Müller, como un útero y trompas de Falopio. Dado que estas estructuras son internas, a no ser que un hermano mayor sea identificado con esta condición, para el correcto diagnóstico del síndrome es necesario el uso de la cirugía. Los testículos se diferencian normalmente, y en el caso que no haya tenido lugar un criptorquidismo prolongado suelen contener células germinales. Sin embargo, los conductos excretores no suelen estar conectados correctamente, ya que frecuentemente desarrollan una aplasia del epidídimo y de la parte superior de los conductos eferentes.

Los análisis genéticos realizados en más de 100 familias con PMDS han mostrado que las mutaciones en el gen de la *AMH* son la causa de la enfermedad en el 45% de los casos. En el 40% de los casos se debe a mutaciones en el gen que codifica el receptor de la AMH, AMHR2. En ambos casos la condición se transmite siguiendo un patrón autosómico recesivo, y son sintomáticas sólo en los varones. En un 5% de los casos de PMDS, las causas son desconocidas.

## 7.3. METODOLOGÍA

En esta práctica se comprobará que el gen de la AMH se expresa en tejido testicular. Para ello vamos a extraer ARN total de testículos y de ovarios (como control negativo) de ratones en estadio neonatal. Con el ARN total realizaremos una reacción de retro-transcripción seguida de una reacción en cadena de la polimerasa, RT-PCR, para detectar la presencia de transcritos de *AMH*.

### *Extracción de ARN*

Para la extracción de ARN se proveerá al alumno de un tubo Eppendorf que contiene una pequeña muestra de tejido testicular u ovárico, que previamente ha sido extraído de ratón en el estadio neonatal y congelado a -80ºC. Se utilizarán columnas extracción de ARN que contienen una membrana de sílice. Las muestras biológicas inicialmente serán lisadas y homogeneizadas en presencia de un tampón altamente desnaturalizante que además contiene tiocianato de guanidina, que inactiva inmediatamente las ARNasas, lo que evita la degradación del ARN. Después se añade etanol, lo que proporciona a la solución unas condiciones físico-químicas que favorecen la unión del ARN a la membrana de sílice, mientras que el resto de los componentes celulares permanecen en disolución. Se hace pasar el lisado a través de una columna de extracción mediante centrifugación. Tras este proceso, el ARN permanecerá unido a la columna, y el resto de componentes celulares se eliminarán con el sobrenadante. Después de lavar la columna usando varias soluciones, se pone agua en la columna. En presencia de agua, el ARN se desprende de la membrana de sílice y pasa a solución acuosa, que será recuperada en el sobrenadante tras una centrifugación.

### *Procedimiento*

1. Añadir 350 µl buffer de lisis (10 µl β-ME por 1 ml Buffer RLT).
2. Homogeneizar pasándolo unas 10 veces por una jeringa con una aguja de 0.8 mm de diámetro.
3. Centrifugar y pasar el sobrenadante a un Eppendorf limpio.
4. Añadir 350 µl EtOH 70% y mezclar por inversión.
5. Transferir a una columna.
6. Centrifugar 1 minuto a máxima velocidad.
7. Digestión del ADN: Añadir 80 µl de solución de Dnasa I, 15 minutos, temperatura ambiente. (preparar gel agarosa para comprobar calidad del ARN)
8. Añadir 700 µl de buffer RW1, centrifugar 1 minuto a máxima velocidad, desechar sobrenadante.

9. Añadir 500 µl de buffer RPE, centrifugar 1 minuto a máxima velocidad, desechar sobrenadante.
10. De nuevo, añadir 500 µl buffer RPE, centrifugar 2 minuto a máxima velocidad, desechar sobrenadante.
11. Colocar la columna en un Eppendorf limpio. Añadir 30 µl de agua libre de ARNasa, esperar 1 minuto, centrifugar 1 minuto a máxima velocidad.
12. En el sobrenadante se recupera el ARN total.

***RT-PCR de un paso***

Una reacción de RT-PCR se puede hacer de dos formas diferentes. En la forma conocida como RT-PCR de dos pasos (two-step RT-PCR), primero se hace la retro-transcripción, con lo que se genera ADN complementario (ADNc) y después, partiendo de este material, en un tubo de reacción diferente se lleva a cabo una PCR convencional. En este tipo de RT-PCR, para la retro-transcripción se usan cebadores universales (Poli-T o cebadores degenerados), con lo que el ADNc generado es representativo de todos los ARNm expresados en el tejido objeto de estudio. Otra forma diferente de llevar una RT-PCR es la que se conoce como RT-PCR de un paso (one-step PCR). En este caso, retro-transcripción y PCR tienen lugar en el mismo tubo de reacción, usándose unos cebadores específicos del gen de interés. Esto implica que en la retro-transcripción inicial sólo se generará ADNc específico del gen de interés, que inmediatamente después servirá de molde para la amplificación de un fragmento mediante PCR.

Nosotros vamos a realizar una RT-PCR de un sólo paso para la AMH. Para ello usaremos los dos cebadores siguientes, localizados en dos exones diferentes del gen de la AMH:

AMH-F: 5'-ACC CTT CAA CCA AGC AGA GA-3'

AMH-R: 5'-CCT CAG GCT CCA GGG ACA-3'

También usaremos una mezcla de enzimas, “One step RT-PCR mix”, que contiene la retro-transcriptasa y la ADN polimerasa.

En un microtubo de 200µl añadir, siguiendo el orden indicado, los siguientes reactivos para un volumen final de 25µl:

- $H_20$ libre de ARNasa 14 µl
- Tampón RT-PCR (5x) 5 µl
- Cebador AMH-F (10 mM) 1 µl
- Cebador AMH-R (10 mM) 1 µl
- dNTPs (10 mM) 1 µl
- Inhibidor de ARNasa 1 µl
- ARN total 1 µl
- One step RT-PCR mix 1 µl

Con el producto de la reacción de RT-PCR se realizará una electroforesis en gel de agarosa. En caso de que en el tejido de partida haya expresión de la AMH, se observará un amplicón de aproximadamente 200 pb.

## 7.4. RECURSOS WEB

A través del link de YouTube se puede acceder al video-tutorial de la práctica.

https://www.youtube.com/playlist?list=PLBa9sJUx0zXWnO2Wu4H6qmJrEOIFNCzaI

## 7.5. CUESTIONES

**1.** ¿Cuál es el factor clave en el procedimiento de extracción de ARN usado en esta práctica? Explicar brevemente por qué.

**2.** ¿Qué enzimas contiene el *One step RT-PCR mix*? ¿Cuál es la función de cada una de ellas?

**3.** ¿Por qué utilizamos tejido testicular prepuberal en esta práctica, y no es apropiado el tejido adulto?

**4.** ¿Por qué es necesario hacer una retro-transcripción previa a la PCR en un estudio de expresión génica como éste?