



---

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

GRADO DE BIOLOGÍA



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

Mohammed Bakkali  
Francisco Javier Barrionuevo Jiménez  
Miguel Burgos Poyatos  
Josefa Cabrero Hurtado  
Roberto de la Herrán Moreno  
Manuel Ángel Garrido Ramos  
Michael Hackenberg  
Rafael Jiménez Medina  
María Dolores López León  
Inmaculada López Flores  
Ángel Martín Alganza  
Rafael Navajas Pérez  
Francisco Perfectti Álvarez  
Francisca Robles Rodríguez  
José Carmelo Ruiz Rejón  
Esther Viseras Alarcón  
Federico Zurita Martínez

**Departamento de Genética, Universidad de Granada**

© Mohammed Bakkali  
Francisco Javier Barrionuevo Jiménez  
Miguel Burgos Poyatos  
Josefa Cabrero Hurtado  
Roberto de la Herrán Moreno  
Manuel Ángel Garrido Ramos  
Michael Hackenberg  
Rafael Jiménez Medina  
María Dolores López León  
Inmaculada López Flores  
Ángel Martín Alganza  
Rafael Navajas Pérez  
Francisco Perfectti Álvarez  
Francisca Robles Rodríguez  
José Carmelo Ruiz Rejón  
Esther Viseras Alarcón  
Federico Zurita Martínez

Portada: Carlos Garrido

I.S.B.N.: 978-84-15261-49-0  
Depósito legal: GR:-3572/2011

## ÍNDICE

<i>Problemas</i>	<i>Página 7</i>
<i>Prácticas Laboratorio y simulación</i>	<i>Página 27</i>
<i>1. Aplicación de la PCR al diagnóstico genético: detección de parásitos que infectan a moluscos</i>	
<i>    Página 29</i>	
<i>2. Clonación de un producto de PCR</i>	
<i>    Página 37</i>	
<i>3. Bases de datos de secuencias de ADN y proteínas</i>	<i>Página 45</i>
<i>4. Predicción computacional de genes</i>	<i>Página 67</i>
<i>5. Alineamiento múltiple de secuencias de ADN y proteínas</i>	
<i>    Página 81</i>	
<i>6. Análisis filogenético</i>	<i>Página 99</i>
<i>7. Análisis computacional de datos de expresión génica diferencial obtenidos mediante chips de ADN</i>	<i>Página 115</i>
<i>8. Expresión de genes implicados en el desarrollo testicular de mamíferos</i>	<i>Página 133</i>
<i>9. Estudio de expresión génica mediante RT-PCR</i>	
<i>    Página 141</i>	



# 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.

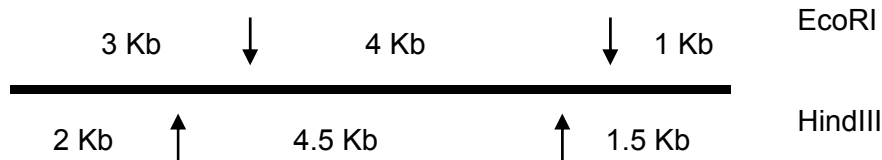
Hay que tener en cuenta la distancia entre dianas, que nos dará el tamaño de las bandas observables pero, además, hay que prestar especial atención a la región con la que hibrida la sonda, pues aquellos fragmentos con los que no hibride, no podrán ser detectados tras el revelado.

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

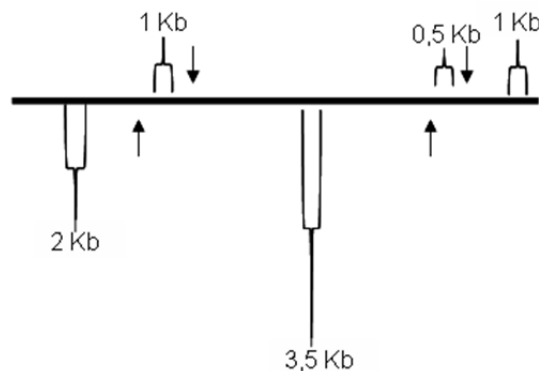
**Problema 1.** Un gen clonado muestra el siguiente mapa de restricción para las enzimas EcoRI y HindIII:



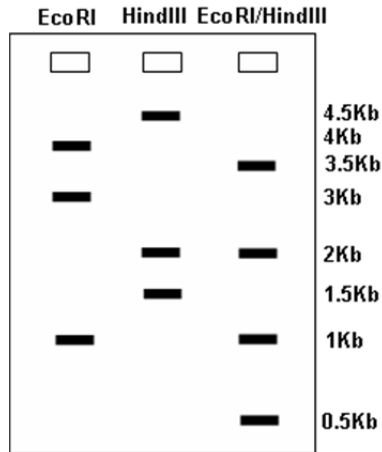
- Dibujar los patrones de los fragmentos de ADN esperados con cada enzima al separar los fragmentos mediante electroforesis en gel de agarosa. Hacer lo mismo para el caso de la digestión doble.
- Dibujar el patrón esperado para una copia mutante del gen que ha perdido el primero de los cortes de EcoRI
- 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*

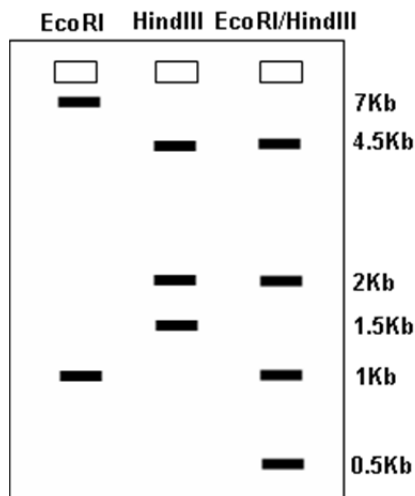
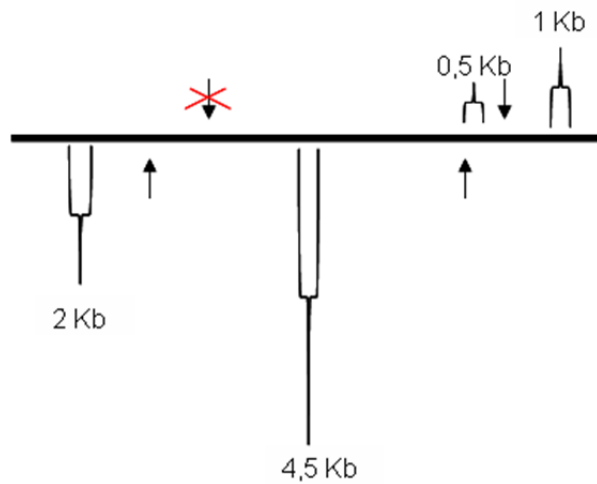
a) Para EcoRI el gen tiene dos dianas, por lo que será cortado en tres fragmentos de tamaños 4Kb+3Kb+1Kb. Para HindIII también tiene dos dianas, pero en diferentes posiciones, por lo que generará tres fragmentos pero de tamaños 4.5Kb+2Kb+1.5Kb. Cuando utilizamos las dos enzimas para digerir el gen, obtendremos 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).



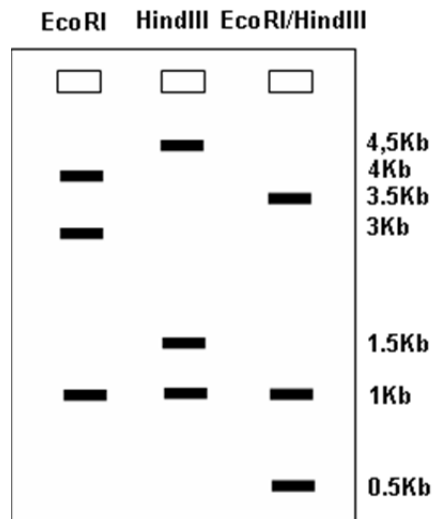
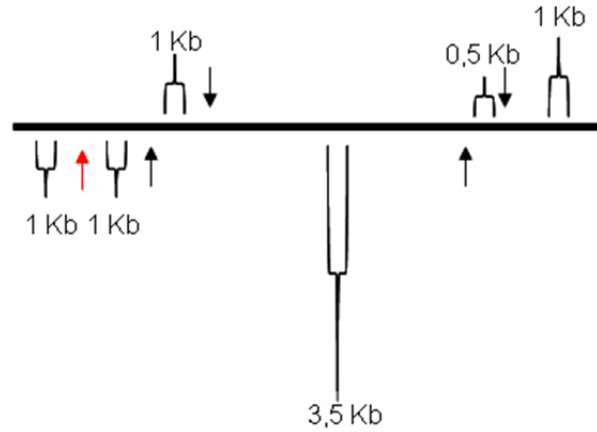
Así, en un gel de agarosa, observaremos los siguientes patrones de bandas:



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).



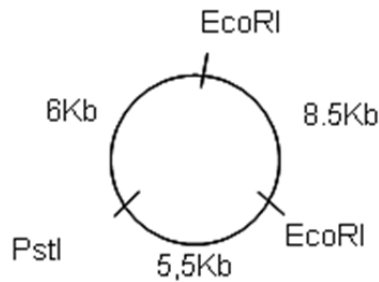
**Problema 2.** Se ha cortado con PstI un plásmido bacteriano circular que contiene un gen de resistencia a la ampicilina. Tras la electroforesis se observa una banda de 20 Kb. ¿Qué deducirías de los resultados que se plantean a continuación?

- Con EcoRI, el plásmido se corta en dos fragmentos: uno de 11,5 Kb y otro de 8,5 Kb
- La digestión PstI+EcoRI genera tres fragmentos de: 6 Kb, 5,5 Kb y 8,5 Kb
- 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.
- Tras cortar uno de los clones recombinantes con PstI se obtienen dos fragmentos: 20 Kb y 6 Kb.
- El clon anterior se corta con EcoRI y se obtienen 10 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 6 Kb y 5.5 Kb por la enzima PstI:

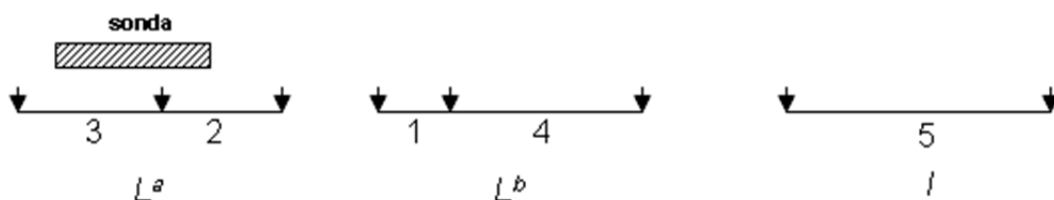


c) Al cortar con PstI, el plásmido se queda en forma lineal con extremos cohesivos para esa enzima. Al poner en contacto fragmentos de ADN cortados también con PstI junto con una enzima ligasa, los fragmentos, que tienen extremos complementarios, se ligan al plásmido y la molécula recirculariza con un 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 esta 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 11,5 Kb.

**Problema 3.** Se conoce un gen autosómico con tres alelos  $L^a$ ,  $L^b$  y  $I$  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

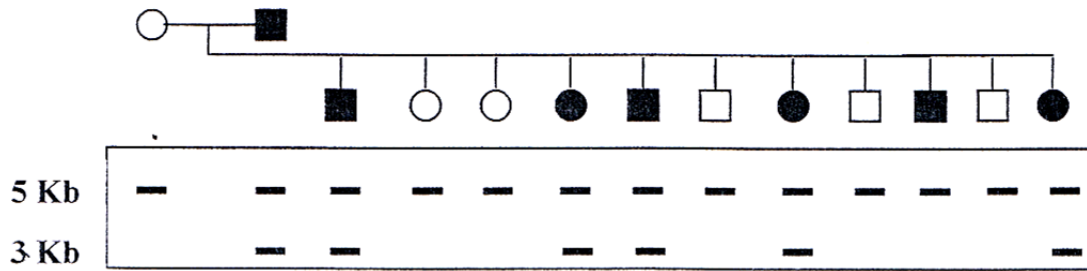
$L^aI$ : bandas de 3Kb y 2Kb para el primer alelo y de 5Kb para el segundo

$L^bL^b$ : bandas de 1Kb y 4Kb

$L^bI$ : bandas de 1Kb y 4Kb para el primer alelo y de 5Kb para el segundo

$I$ : 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.

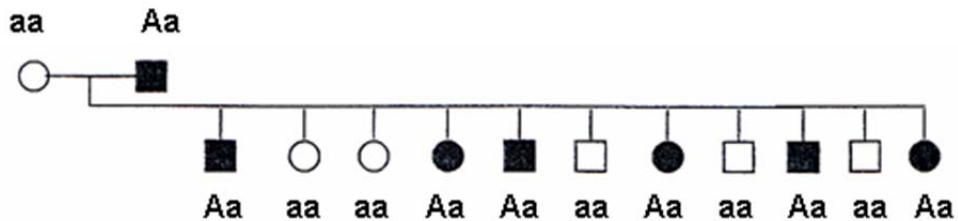


- a) Explica el protocolo seguido y los resultados obtenidos en los distintos individuos  
 b) ¿Podemos usar la sonda con fines diagnósticos para esta enfermedad?

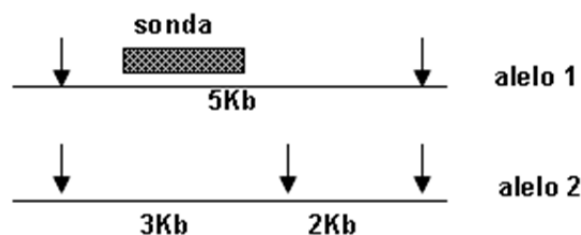
*Respuesta*

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:

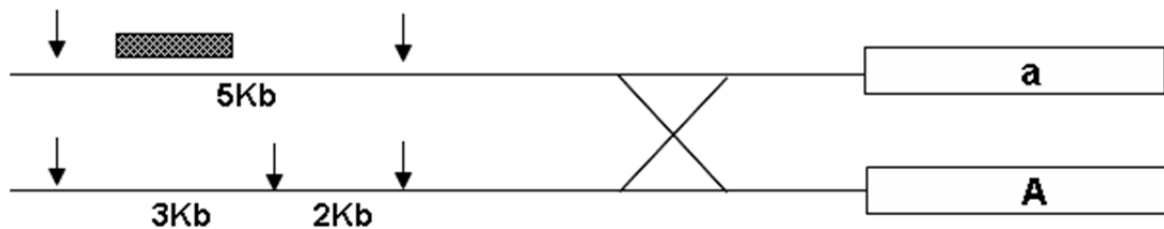


Al comparar los resultados del marcador molecular con los fenotipos (afectados/no afectados por la enfermedad) podemos ver cómo existe coincidencia entre el número y los tamaños de bandas del RFLP y el desarrollo o no de la enfermedad. Así, a excepción de un individuo (II-9), todos los afectados presentan 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 un alelo *a* (5Kb) de la madre y un alelo *A* (recombinante de 5Kb) 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
<b>Locus 1</b>	130	134	134	134	136	138	128	134	128	136
<b>Locus 2</b>	250	256	256	260	258	260	252	260	250	258
<b>Locus 3</b>	140	140	140	144	146	148	138	144	140	150
<b>Locus 4</b>	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 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

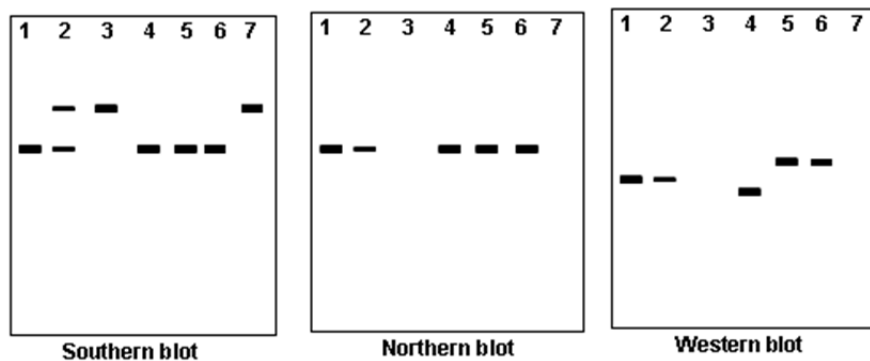
**Problema 3.** Un fragmento lineal de ADN de 11Kb se corta por separado con las enzimas de restricción EcoRI y HaeIII y con una mezcla de ambas obteniéndose los fragmentos indicados a continuación. EcoRI: 6Kb, 3Kb y 2Kb; HaeIII: 7Kb y 4Kb; EcoRI+HaeIII: 5Kb, 3Kb, 2Kb y 1Kb. Dibujar el mapa de restricción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

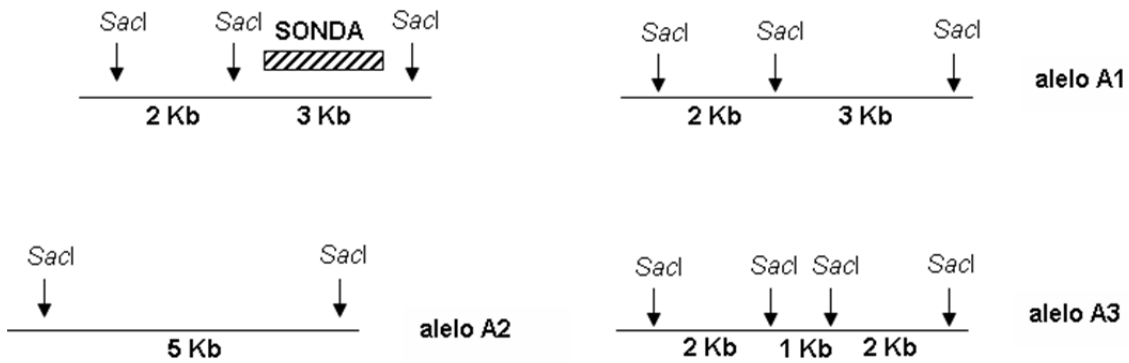
**Problema 5.** Una proteína está codificada por un gen que no tiene intrones. El fragmento de restricción SacI que contiene el gen completo puede ser identificado por hibridación tipo Southern-blot con el ADNc del gen marcado radiactivamente. Para determinar la causa de una enfermedad desconocida, se 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 hibridó con el ADNc marcado como sonda. Igualmente, se extrajo ARN, se sometió a electroforesis, se transfirió a una membrana (Northern-blotting) y se hibridó con el ADNc. Además se realizó la técnica de Western-blotting 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?

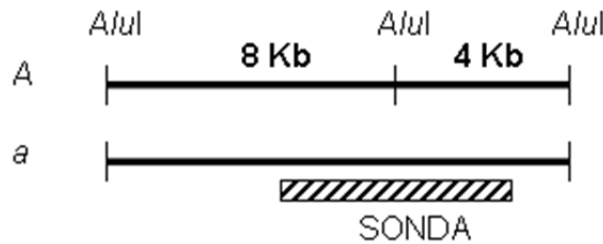


**Problema 6.** Tras una búsqueda de marcadores moleculares para una especie, se diseña una sonda complementaria al ADN genómico de esta especie en la región indicada en el esquema, y que permite diferenciar tres alelos distintos (A1, A2 y A3). El ADN extraído de diferentes individuos se corta con SacI, se realiza una electroforesis y se tranfiere el ADN posteriormente a una membrana de nylon. Esta membrana se hibrida con la sonda (marcada radiactivamente) y se realiza una autorradiografía.



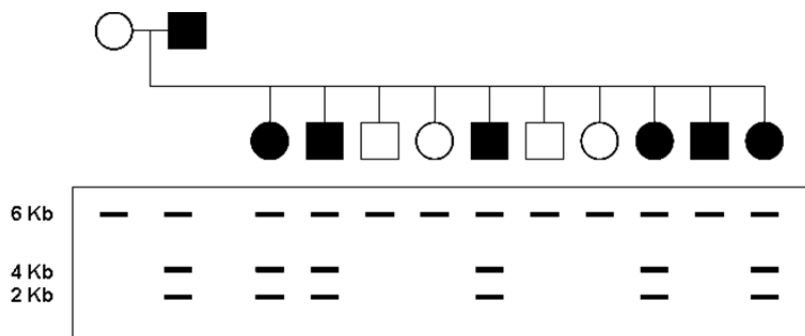
Dibujar esquemáticamente el resultado esperado para: Un homocigoto para A1, un heterocigoto A1A2, un homocigoto A2 y un heterocigoto A1A3.

**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.

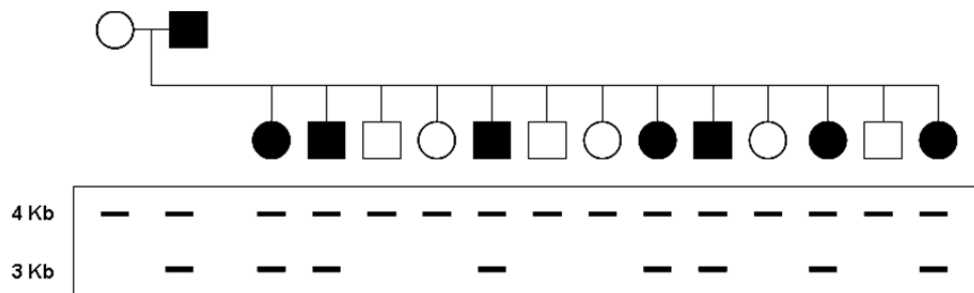
- Explicar los resultados.
- Analizar si existe ligamiento entre el RFLP y el gen causante de la enfermedad.



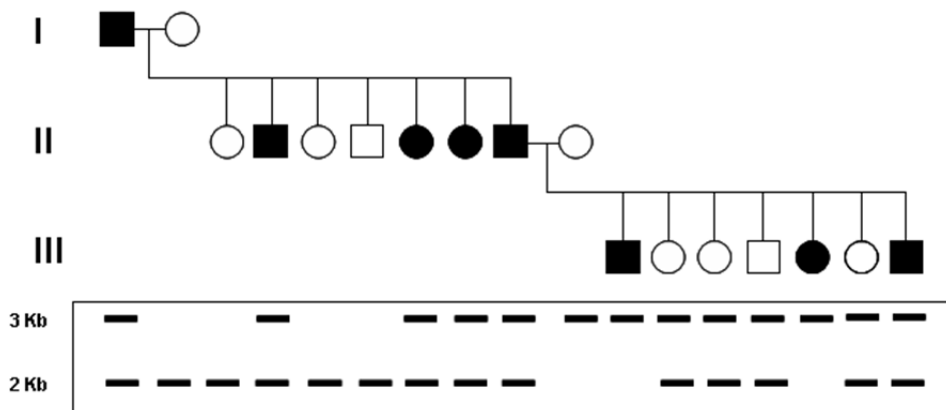
**Problema 9.** Se probaron distintas sondas de ADN hibridando con el ADN genómico de los individuos de una familia numerosa en la que algunos miembros están afectados por una enfermedad autosómica dominante leve. Sobre el Southern-blot obtenido con EcoRI, una de las sondas detecta un polimorfismo en 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.

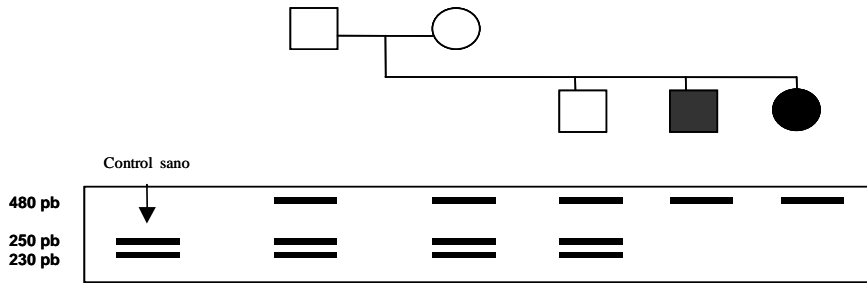
b) Analizar si existe ligamiento entre el RFLP y el gen causante de la enfermedad.



**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:

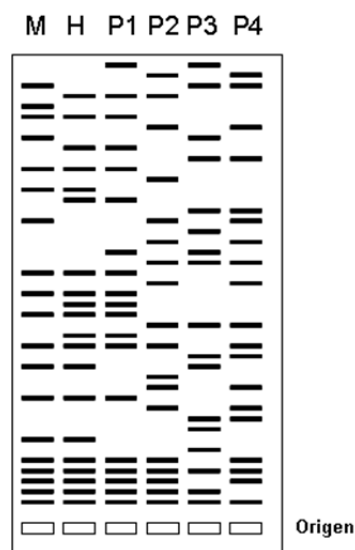


¿Se puede utilizar este marcador como método de diagnóstico? ¿Qué tipo de enfermedad se describe en el pedigrí? Explica, mediante un esquema, los resultados obtenidos a nivel molecular.

**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.

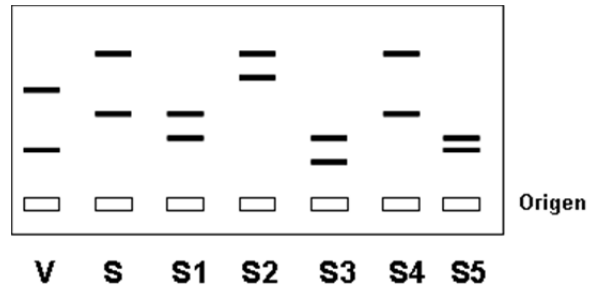
**Problema 13.** Cuatro hombres se disputan la paternidad de un niño. Los forenses deciden utilizar el método de la huella genética para resolver el caso, analizando el ADN de la madre (M), del hijo (H) y de los cuatro posibles padres (P1 a P4). Los resultados obtenidos se muestran en la figura.

- 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,

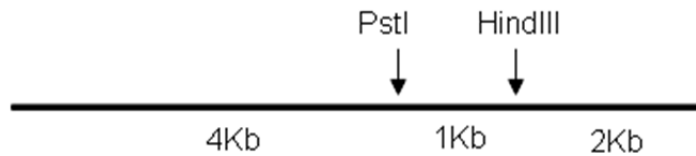
S3, S4 Y S5). Se lleva a cabo un estudio de microsatélites empleando una pareja de cebadores específicos de un locus. Una vez realizada la amplificación con la pareja de cebadores, se obtienen los siguientes resultados:



- Explicar los patrones de amplificación obtenidos.
- ¿Existe algún sospechoso que parezca culpable?

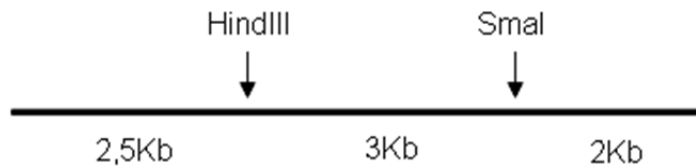
#### 4. SOLUCIONES A LOS PROBLEMAS

##### Problema 1.

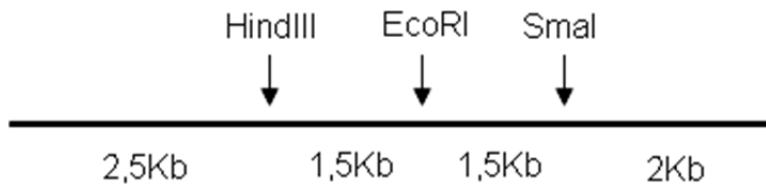


##### Problema 2

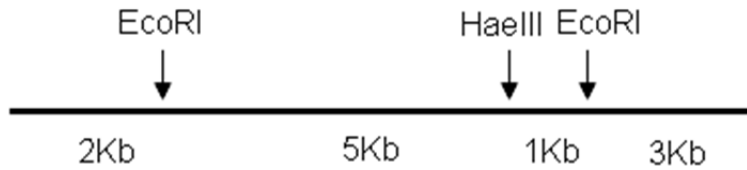
a.



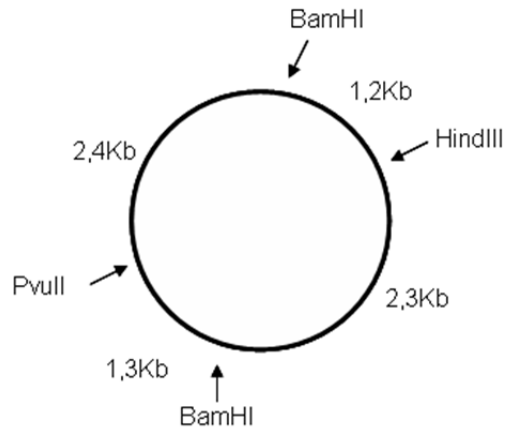
b.



**Problema 3.**



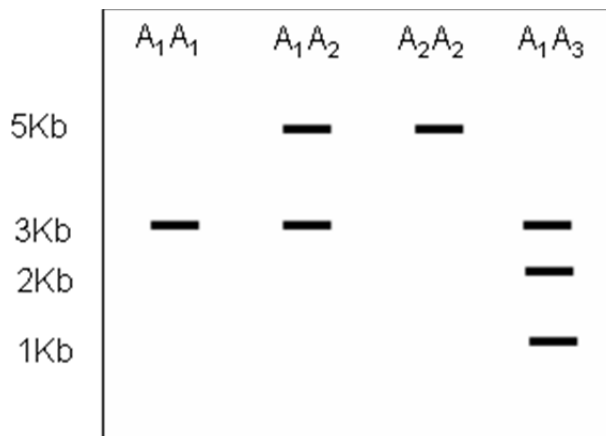
**Problema 4.**



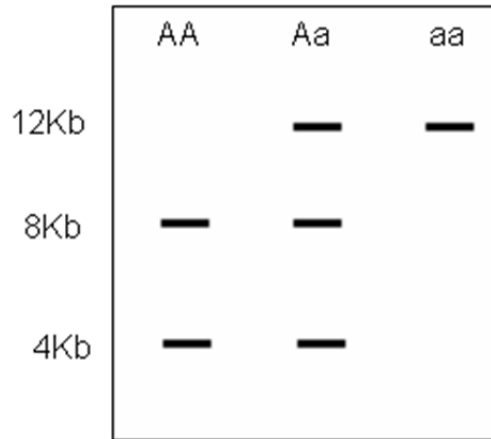
**Problema 5.**

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.**

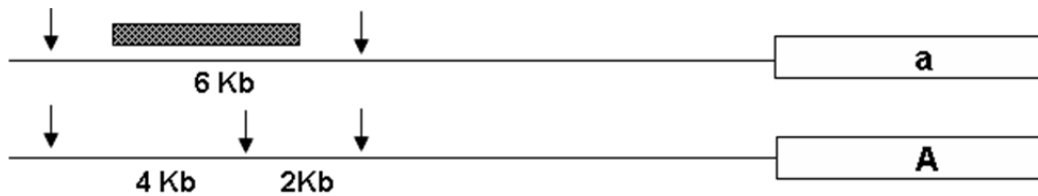


**Problema 7.** Se digiere el ADN genómico de los individuos con la enzima AluI, se somete el ADN cortado a una electroforesis y mediante la técnica de Southern-blot se transfieren los fragmentos a una membrana de nylon. Se realiza una hibridación utilizando la sonda para detectar las regiones homólogas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



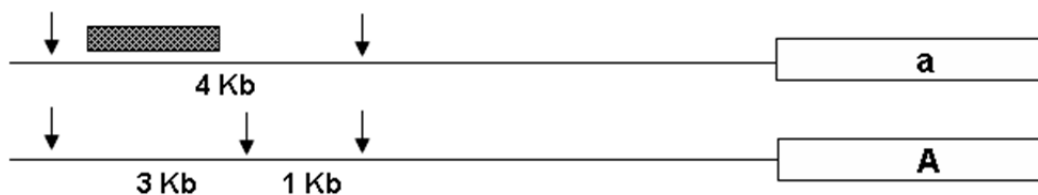
**Problema 8.**

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



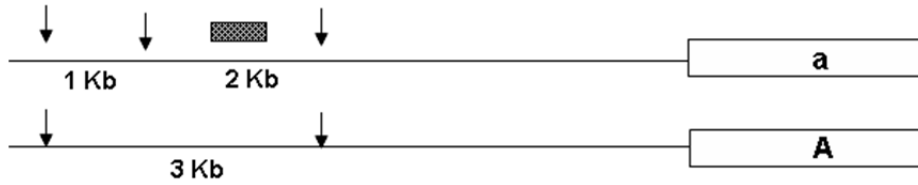
Afectados, Aa (6Kb/4Kb/2Kb); Sanos, aa (6Kb). El genotipo del individuo II-9 es resultado de un entrecruzamiento entre la región que contiene al RFLP y el gen causante de la enfermedad en una de las meiosis del padre.

**Problema 9.** Están ligados, y una posible interpretación de los resultados se muestra en el siguiente esquema:





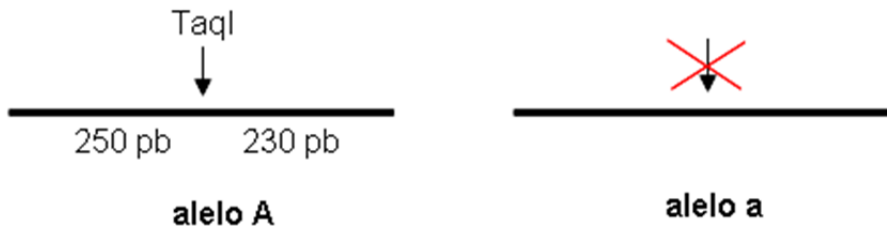
**Problema 10.** Se encuentran ligados. La interpretación podría ser:



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).

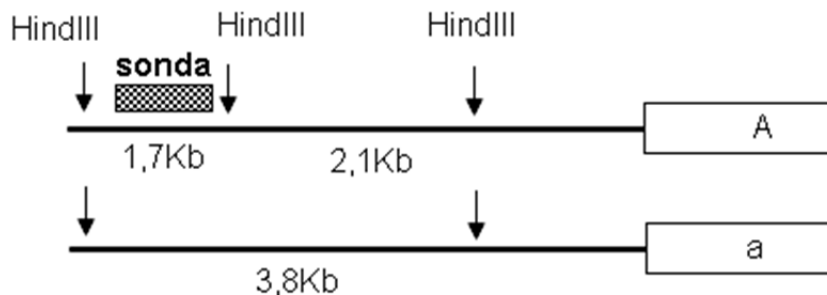
El individuo III-7, es el resultado de un entrecruzamiento entre el RFLP y el gen causante de la enfermedad en el padre heterocigoto (II-7)

**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



APLICACIÓN DE LA PCR AL  
DIAGNÓSTICO GENÉTICO:  
DETECCIÓN DE PARÁSITOS QUE  
INFECTAN A MOLUSCOS



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

## 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.

## 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 de 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 94°C. Posteriormente, el alineamiento de los cebadores se consigue bajando la temperatura hasta un nivel ( $T_m = \text{Temperature of melting}$ ) que permite a éstos unirse

específicamente a su secuencia inversa y complementaria. Dicha temperatura ( $T_m$ ) debe ser aproximadamente similar para los dos cebadores (no más de 5°C de diferencia) y depende tanto de la composición como del tamaño del cebador. Hay una variedad de algoritmos que permiten calcular la  $T_m$ ; una fórmula básica para estimarla es:  $4 \times GC + 2 \times AT$ , 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 (94°C) y para el alineamiento de los cebadores (40-65°C 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 72°C, temperatura similar a la de donde vive este microorganismo. La fase de extensión con la Taq polimerasa se hace, por tanto, a 72°C (otras ADN polimerasas tendrán otras temperaturas óptimas 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  $2^n \times C$  moléculas de ADN amplificado donde n es el número de ciclos de PCR y C la cantidad de moléculas molde iniciales. 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 madreameja (*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 de 760 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.

### 3. METODOLOGÍA

#### **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 µl
- 10% Tampón de PCR 10x 2.5 µl
- 2mM de cada dNTPs 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:	94°C	30 seg.
Alineamiento:	58°C	30 seg.
Extensión:	72°C	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 TAE (0.04 M Tris-acetato, 0.001 M EDTA) y 0,4 g de agarosa.

Calentar utilizando el microondas hasta que se funda la agarosa.

Dejar enfriar hasta aproximadamente 50°C 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 (TAE 1x o TBE 0.5x).

## 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 5  $\mu$ l de tampón de carga a los tubos en los que se desarrolló la PCR y, 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  $\mu$ l 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.

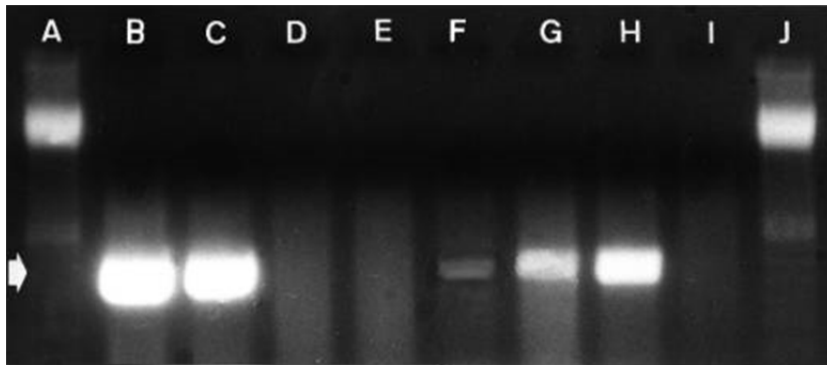
Conectar la fuente de alimentación al gel durante 30 minutos a 50 volts/cm.

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 (ADN clonado a partir del que se diseñaron los oligonucleótidos). 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 (780 pb). Se demuestra la eficacia del método de diagnóstico para *Perkinsus olseni* en cultivos de almejas dado que los oligonucleótidos de que disponemos 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. Así, en los productos de amplificación que se observan en las calles F a H del gel, se puede observar una gradación de menor a mayor en cuanto a la cantidad de producto amplificado, gradación que se corresponde con los niveles de infección de cada uno de los tres cultivos de donde procedían las tres muestras.

## 4. CUESTIONES

- 1.- ¿Qué criterios se debe seguir a la hora de diseñar cebadores para este tipo de análisis?
- 2.- ¿Qué harías si observa amplificación en muestras que claramente sabemos que no están infectadas?
- 3.- ¿Qué es lo que hace de la PCR una técnica idónea para diagnóstico?
- 4.- ¿Podríamos descartar completamente una infección si para una muestra observamos ausencia de amplificación tras la PCR?

# CLONACIÓN DE UN PRODUCTO DE PCR



# CLONACIÓN DE UN PRODUCTO DE PCR

## 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 procariontas (*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. INTRODUCCIÓN

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, 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  $\beta$ -galactosidasa (gen *lacZ*, del operón *lac* de *E. coli*) interrumpido por la región de clonación múltiple (*polylinker*). Para que tenga lugar la expresión del gen de la  $\beta$ -galactosidasa, es necesaria la presencia de *IPTG*, molécula que actúa como un inductor continuo del gen. La proteína  $\beta$ -galactosidasa, en presencia de uno de sus sustratos, *X-gal* (5-Bromo-4-Cloro-3-Indol- $\beta$ -D-galactósido), produce un precipitado azul, ya que *X-gal* es hidrolizado por el 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  $\beta$ -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  $\beta$ -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).



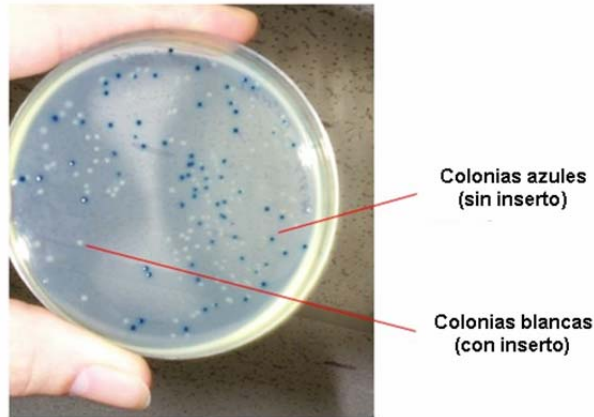


Figura 1: Resultado de un experimento de clonación

Existen otras estrategias similares que se pueden usar con el mismo propósito. Por ejemplo, usando 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.

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-4Z (utiliza el gen de la  $\beta$ -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  $\text{Ca}^{2+}$  (en forma de  $\text{CaCl}_2$ ), 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,  $42^\circ\text{C}$  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  $\mu\text{s}$ ). 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 (entre estas destaca la cepa de *E. coli* DH5 $\alpha$ ).

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 NTS del ADN ribosómico (ADNr) del parásito *Perkinsus olseni*. Como vector de clonación vamos a utilizar el plásmido pGEM-4Z, un vector que presenta las características descritas anteriormente.

### 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-4Z (*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 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. En este proceso interviene una enzima llamada ligasa, que establece un enlace fosfodiéster entre la última base del 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 (Figura 2). Para llevar a cabo esta reacción, se realizan los siguientes pasos:

1. En un microtubo Eppendorf añadir:
  - 6  $\mu$ l del producto de PCR (100-200 ng)
  - 2  $\mu$ l de tampón
  - 1  $\mu$ l del vector pGEM-4Z
  - 1  $\mu$ l de la enzima ligasa
2. Incubar durante 30 minutos a temperatura ambiente.

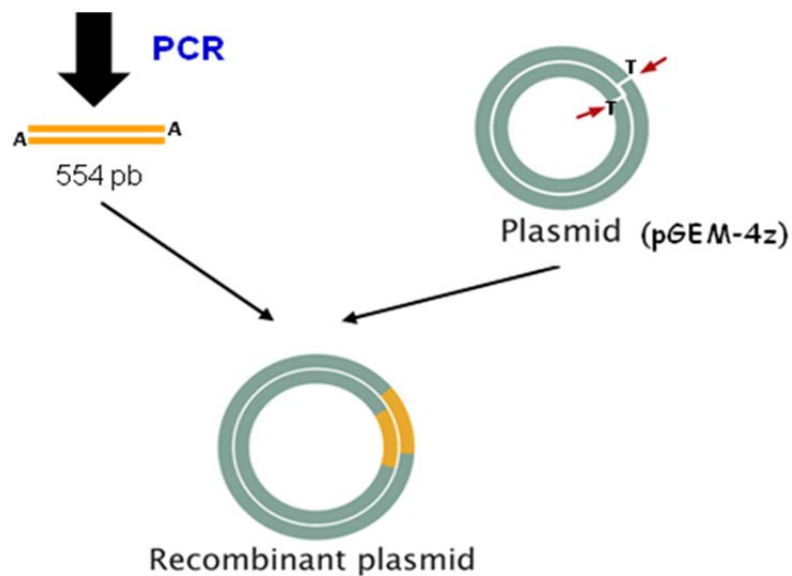


Figura 2: Ligado de un amplificado de PCR al vector de clonación pGEM-4Z

### **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 JM109 de *E. coli*. Se procederá de la siguiente forma:

- Depositar en hielo un tubo Eppendorf durante unos minutos conteniendo 50  $\mu$ l de bacterias competentes.
- Añadir 10  $\mu$ l de la solución de ligación.
- Dejar 20 minutos en hielo.
- Choque térmico introduciendo el tubo Eppendorf con la mezcla en un baño a 42°C durante 45 segundos.
- Inmediatamente, pasar el microtubo a hielo, durante 2 minutos.
- Añadir 250  $\mu$ l de medio de cultivo LB líquido.

- Incubar durante 30-40 minutos a 37°C 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 37°C.

### **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).

## **4. 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-4Z para clonar un fragmento de ADN cortado por enzimas de restricción? ¿Y uno amplificado por una ADN polimerasa de alta fidelidad?
- 4.- ¿Cuáles son las posibles causas para la ausencia de colonias blancas en la placa?
- 5.- ¿Cuáles son las posibles causas para la ausencia de colonias en la placa tras una adecuada incubación a 37°C?

**BASES DE DATOS DE  
SECUENCIAS DE ADN y  
PROTEÍNAS**



# BASES DE DATOS DE SECUENCIAS DE ADN y PROTEÍNAS

## 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.

## 2. FUNDAMENTO TEÓRICO

### 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.

### 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/>).

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, <http://www.embl.de/>) 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/>).

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 famosos programas de alineamiento y comparación de secuencias *Clustal* y *Blast*).

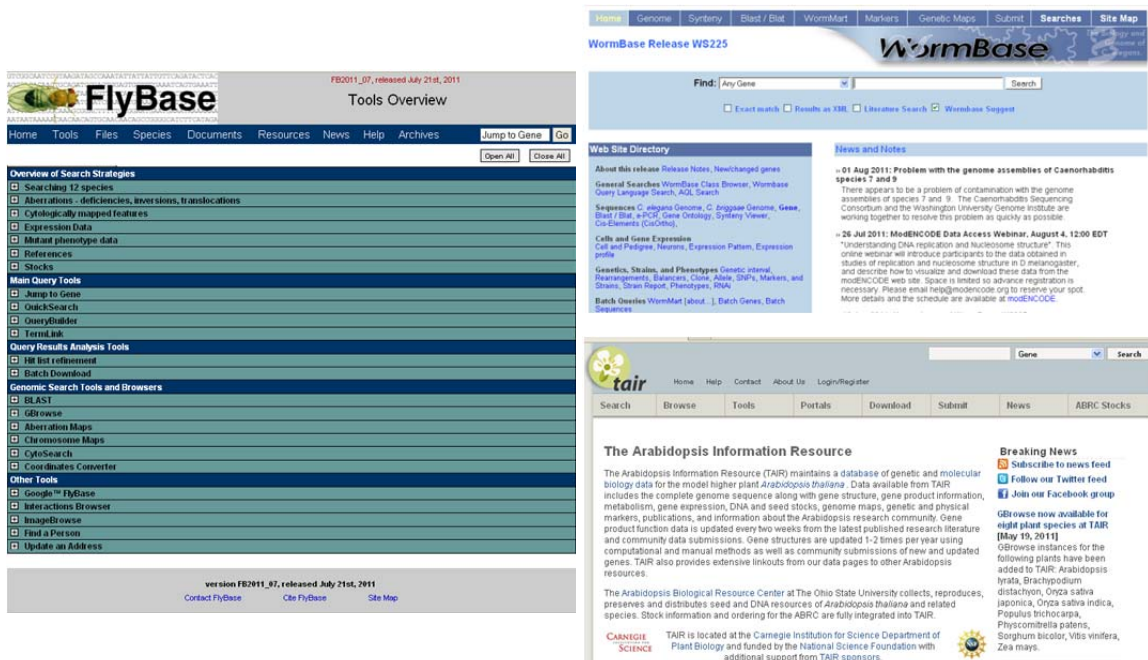


Figura 1: Ejemplos de bases de datos de organismos específicos

The screenshot shows the DDBJ (DNA Data Bank of Japan) website. It features a navigation menu with options like 'HOME', 'Submission', 'How to use', 'Search Analysis', 'FTP and WebAPI', 'Reports/Statistics', and 'Contact us'. The main content is divided into several sections: 'Database Search' (including GenEntry, ARSA, TXSearch, BLAST, and DDBJ Vector Screening System), 'Phylogenetics' (ChastalW), 'Submission of Gene Expression Data' (CIBEX), 'Genome Analyses' (GIB, GIB.V, GTPS, GTOP), 'Next Generation Sequence Analysis' (DDBJ Read Annotation Pipeline), 'Protein Database and Structure' (PMO), and 'Software developed at CIB-DDBJ' (WINA, DendroMaker for Macintosh).

The screenshot shows the ENA website. It includes a search bar and a 'Text search' section with a search query input field and a 'Search' button. There is also a 'Sequence Search' section for entering a nucleotide sequence or accession number.

The screenshot shows the NCBI Entrez website. It features a search bar and a 'Search across databases' section. Below this, there are several database categories listed, such as 'PubMed', 'Books', 'Nucleotides', 'ESTs', 'Genomes', 'Proteins', 'Structures', 'Taxonomy', 'dbGAP', 'UniGene', 'CCDD', 'UniSTS', 'PopSet', 'GEO Profiles', and 'GEO Datasets'.

Figura 2: Bases de datos generales más relevantes

### 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](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*).

The screenshot shows the NCBI Gene database search results for the query 'collagen and human'. The search results are sorted by relevance and show 1 to 20 of 1546 results. The first three results are listed:

- COL5A1**  
 Official Symbol: COL5A1 and Name: collagen, type V, alpha 1 [*Homo sapiens*]  
 Other Aliases: RP11-263F14.1  
 Other Designations: OTTHUMP0000022513; alpha 1 type V collagen; collagen alpha-1(V) chain  
 Chromosome: 9, Location: 9q34.2-q34.3  
 Annotation: Chromosome 9, NC\_000009.11 (137533652..137736689)  
 MIM: 120215  
 ID: 1289
- HMGGA2**  
 Official Symbol: HMGGA2 and Name: high mobility group AT-hook 2 [*Homo sapiens*]  
 Other Aliases: BABL, HMG1-C, HMG1C, LIPO, STQTL9  
 Other Designations: High-mobility group protein HMG1-C; OTTHUMP00000239770; OTTHUMP00000239772; OTTHUM  
 Chromosome: 12, Location: 12q15  
 Annotation: Chromosome 12, NC\_000012.11 (66218240..66360071)  
 MIM: 600698  
 ID: 8091  
[Order cDNA clone](#)
- COL27A1**  
 Official Symbol: COL27A1 and Name: collagen, type XXVII, alpha 1 [*Homo sapiens*]  
 Other Aliases: RP11-8211.1, FLJ11895, KIAA1870, MGC11337

On the right side of the search results, there are filters for 'Filter your results:' (All (1546), Current Only (1334), Genes Genomes (1314), SNP GeneView (1185), In Variation Viewer (194)), a 'Top Organisms' tree (Homo sapiens (741), Mus musculus (416), Rattus norvegicus (244), Caenorhabditis elegans (14), Gallus gallus (72), All other taxa (719)), and a 'Find related data' section with a 'Database' dropdown and a 'Find items' button.

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, número 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 (alineadas, hacer árboles 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

cagaactattctccccagtatgattcatatgatgtcaagtcgggaggtagcagtaggaggactcgaggct
atcctggaccagctggccccccaggcccccccgccccctgggtacatctgggtcatcctggttcccctggatc
tccaggataccaaggaccctgggtgaacctgggcaagctgggtccttcaggccctccaggacctcctgggtgct
ataggtccatctggctcctgctggaaaagatggagaatcaggtagaccgggacgacctggagaccgaggattgc
ctggacctccaggtatcaaaggtccagctgggatacctggattccctgggtatgaaaggacacagaggcttcga
tggacgaaatggagaaaaggggtgaaacaggtgctcctg...ccctggctcctgctgtggtggtggtggagccc
ctgccattgctgggattggagctgaaaaagctggcggttttgcccttattatggagatgaaccaatg
```

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)
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="QNYSPQYDSYDVKSGGVAVGGLAGYPPGAGPPGPPGPPGPGTSGHPG
FT   SPGSPGYQGPPGEPGQAGPSGPPGPPGAIGPSGPAGKDGESGRPRPGDRGLPGPPGIK
FT   GPAGIPGFPMKGRHGFDRNGEKGETGAPGLKGENGLPGENGAPGPMGPRGAPGERGR
FT   PGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPSPGAKGEVGPAGSPGSNGAPGQRG
FT   EPGPQGHAGAQQPPGPPGINGSPPGKGMGPAGIPGAPGLMGARGPPGAPANGAPGLR
FT   GGAGEPGKNGAKGEPGRGERGEAGIPGVPGAKGEDGKDGSPGDPGANGLPGAAGERGA
FT   .....CCGGVGAPAIAGIGAEEKAGGFAPYYGDEPM"
XX
SQ   Sequence 3234 BP; 664 A; 861 C; 1106 G; 603 T; 0 other;
cagaactatt ctccccagta tgattcatat gatgtcaagt cgggscgagat agcagtagga 60
ggactcgcag gctatcctgg accagctggc cccccaggcc cccccggccc ccctggtaca 120
tctggtcacc ctggttcccc tggatctcca ggataccaag gacccccctgg tgaacctggg 180
caagctggtc cttcaggccc tccaggacct cctggtgcta taggtccatc tggctcctgct 240
ggaagaagatg gagaatcagg tagaccgccga cgacctggag accgaggatt gcctggacct 300
ccaggtatca aaggtccagc tgggataacct ggattccctg gtatgaaagg acacagaggg 360
ttcgatggac gaattggaga aaaggtgaaa acaggtgctc ctggattaaa gggtgaaaat 420
..... attg gagctgaaaa agctggcggg tttgccctt attatggaga tgaaccaatg 3234
//

```

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

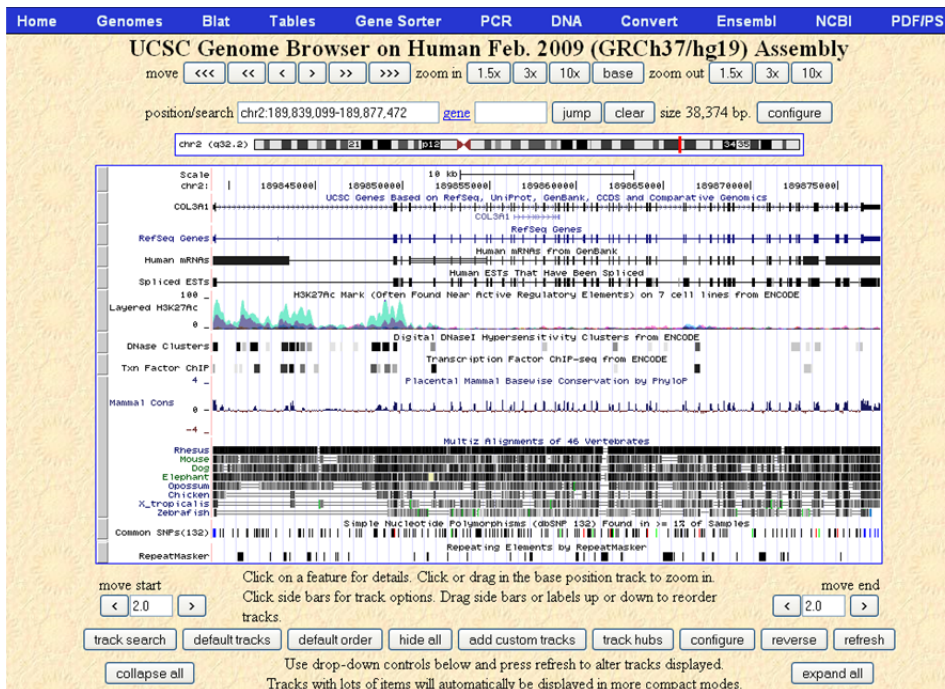
```

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
            ..5490
            /organism="Homo sapiens"
            /mol_type="mRNA"
            /db_xref="taxon:9606"
            /chromosome="2"
            /map="2q31"
            gene
            ..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
            ..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; alphas(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="MMSFVQKGSWLLALLHPTIILAQQEAVEGGCSHLGQSYADRDVWVPEP
            CQICVCDSGSVLCDDICDDQELDCPNPEIPFGECCAVCPQPPTAPTRPPNGQKGDGPPGPGIP
            GRNGDPIFGQPGSPGSPGPPGICESCPTGPNYSPQVDSYDVKSGVAVGLAGYGPAGPPG
            PPGPPTSGHPGSPGSPGYQGPPEPGQAGPSPGPPGPAIGPSPGAGKDGESGRPRPGERG
            LPPGPIKGPAGIPGPGMKHRRGDRNGEKGETGAPLKGENGLPENGAPGPMGPRGAPG
            ERGRPLGAAGARGNDGARGSDGQPGP...VRLPIVDIAPYDIGPQDFGVDVGPVCF"
            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 agggcagggg acaacttgat ggtgctactt
            61 tgaactgctt ttcttttctc ctttttgcac aaagagtctc atgtctgata ttagacatg
            121 atgagctttg tgcaaaaggg gagctggcta cttctcgtcc tgcttcaccc cactattatt
            181 ttggcacaac aggaagctgt tgaaggagga tgttcccatc ttggtcagtc ctatgcggat
            241 agagatgtct ggaagccaga accatgccaa atatgtgtct gtgactcagg atccgtttcc
            5461 ..... caccataaat aaaatatcat attaaaaattc
            //
    
```



## 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 bandeado 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. METODOLOGÍA

### 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.

### 3.2. 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.

#### 3.2.1. GenBank

GenBankR (<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). En agosto de 2009 había aproximadamente 106 533 156 756 bases en 108 431 692 registros de secuencias almacenadas (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.



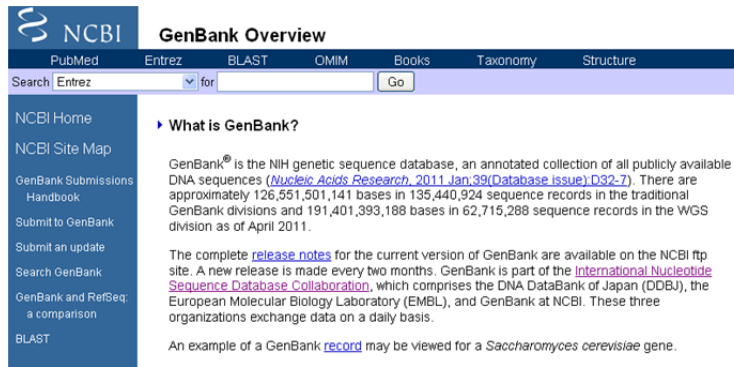


Figura 5: Página web de acceso a GenBank

Debajo del menú principal horizontal, hay un menú desplegable llamado *Search*, en el que se puede escoger la base de datos a utilizar (Figura 6), junto con una caja de texto seguida de un botón *Go*. 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 el botón *Go*.

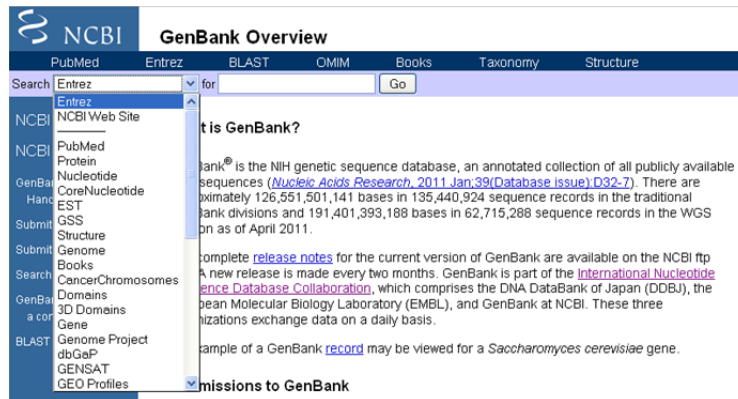


Figura 6: Realizando una búsqueda en GenBank

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.

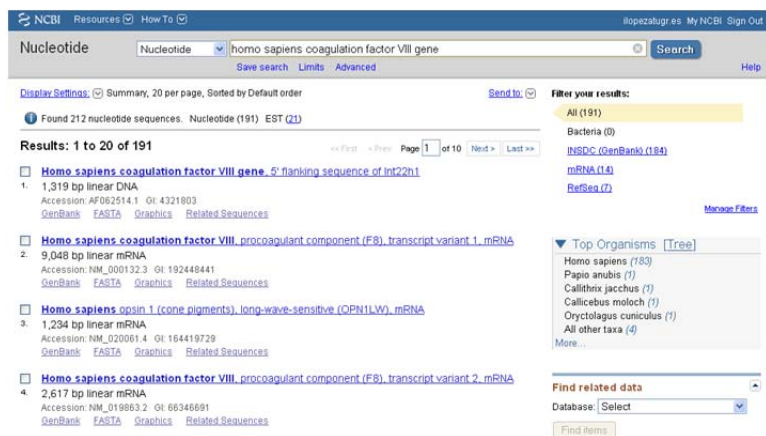


Figura 7. Resultado de una búsqueda en GenBank

Como se observa, se han obtenido 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).

NCBI Resources How To ilopezatugr.es My NCBI Sign Out

Nucleotide Nucleotide Search Limits Advanced Help

Display Settings: GenBank Send

Change region shown

Customize view

### Homo sapiens coagulation factor VIII gene, 5' flanking sequence of Int22h1

GenBank: AF062514.1

FASTA Graphics

LOCUS AF062514 1319 bp DNA linear PRI 28-FEB-1999

DEFINITION Homo sapiens coagulation factor VIII gene, 5' flanking sequence of Int22h1.

ACCESSION AF062514

VERSION AF062514.1 GI:4321803

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 1319)  
AUTHORS Liu, Q. and Sommer, S.S.  
TITLE Subcycling-PCR for multiplex long-distance amplification of regions with high and low GC content: application to the inversion hotspot in the factor VIII gene  
JOURNAL BioTechniques 25 (6), 1022-1028 (1998)  
PUBMED [9863056](#)

REFERENCE 2 (bases 1 to 1319)  
AUTHORS Liu, Q., Norzari, G. and Sommer, S.S.  
TITLE Direct Submission  
JOURNAL Submitted (01-MAY-1998) Molecular Genetics, City of Hope National Medical Center and Beckman Research Institute, 1500 East Duarte Rd, Duarte, CA 91010, USA

FEATURES  
source 1..1319  
/organism="Homo sapiens"  
/mol\_type="genomic DNA"  
/db\_xref="taxon:9606"  
/chromosome="X"  
/map="Xq telomere"  
[gene](#) <1..>1319  
/gene="coagulation factor VIII"  
[intron](#) <1..>1319  
/gene="coagulation factor VIII"  
/number=22  
[misc feature](#) 1..1319  
/gene="coagulation factor VIII"  
/note="5' flanking sequence of Int22h1 after genomic inversion"  
/phenotype="severe hemophilia A"

ORIGIN  
1 ccatacatta gtaaaatcag aatacatttg aatttaccaa gtaggaacaa gagtattaag  
61 ttctactgcc atgcatcagg gcaatgtag ctctcttgtt ttctatcata atatagactc  
121 aagggacctc aaacatcttt acatcccaca agcacaatgc ctgtccatta cactgatgac  
181 atattgctga caggatcttg taaacacata acaaaatagc ctgtcttctt cttgacttta  
241 atgagaatgt aacctgtgtt tcaactgttta atataatggt cactgttagt tcaaaact  
301 tttatgatg tttaaaaagt tttctcctat ccctatttta ttgcaatgg caactgaatt  
361 ttatcaaatg cttttccagc atctttgaca tggtcacatt tctctttgt gttgtcaaat  
421 tatacttaac attcaatagt gtgctgacaa gaaattaaca accccagagg gagcaaatg  
481 agaagaggtt aagaagtaaa ggccttgatt tatagcattg gcagatttcc ctgacataaa  
541 tactactccc atcatgctg gccccagggg tgggaagaga tgcttactta caattggctc  
601 tcacagacaa gtgcaagagc actactgtgc tccatttctg ggaaaaactt cgtcagctc  
661 agtgtggttag ttatttaaaa cttagctgga tccaatttgc caacatttca ttataaatt  
721 ctatactcat attcatgaat gaaatgggtt tagcttttcc agtagcttca cttaccaggt  
781 ttggcatga gggttatatt aaacttgaaa ataaagtggg aaagcttcca ttttttcca  
841 tgaagactat tgctctagaa tagcttattt aatgtaggaa tccagattt aatgagagta  
901 aatgaaaag ccatatgagc catgtgcttt atttagttag agatactagg ctatatttcc  
961 caatgcttat atgattattg ctattctcat ttgtgttccc ttgagttaat gctgcaaat  
1021 gtgcacatta gagtcatata tcccttccctg tgccatacat ctatatctgt atcacacat  
1081 atgtcttttc tcttttttcc ctttctgtat caagagttgg caaggtttg tctatttat  
1141 taatttttca aagaatcagc tcagttttaa acccaactg ggtttttaa gagctgtttc  
1201 tagttcatca atctctgttc aaaactttaa aaattctatt ttcctttctt ttggtttgt  
1261 tctacaatc tggaggtgaa tgcttagtcc acttatttct caactcttgt attttaacg  
//

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.

El formato FASTA (se puede ver picando en el enlace correspondiente en la parte de arriba de la página, picando en *Display settings*) es mucho más sencillo. Consiste en una única línea de anotación, precedida por el símbolo “>”, seguida de la secuencia en la línea siguiente (Figura 9).

The screenshot shows the NCBI GenBank interface. At the top, there's a search bar and a 'Format' dropdown menu. The 'FASTA' option is selected. Below the menu, the sequence is displayed in FASTA format. The header line is: >gi|4321803|gb|AF062514.1| Homo sapiens coagulation factor VIII gene, 5' flanking sequence of Int22h1. The sequence itself is a long string of nucleotide characters (A, T, C, G) on the following line. The page also includes various navigation and utility buttons like 'Send', 'Change region shown', and 'Analyze this sequence'.

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.

NCBI Resources How To ilopezalugr.es My NCBI Sign Out

Protein Protein homo sapiens coagulation factor VIII gene Search Limits Advanced Help

Display Settings: GenPept Send to: Change region shown

**coagulation factor VIII [Homo sapiens]**

GenBank: AAC32196.1  
[FASTA](#) [Graphics](#)

[Go to](#)

LOCUS AAC32196 76 aa linear PRI 18-AUG-1998

DEFINITION coagulation factor VIII [Homo sapiens].

ACCESSION AAC32196

VERSION AAC32196.1 GI:3421393

DBSOURCE accession [AF081784.1](#)

KEYWORDS .

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (residues 1 to 76)  
 AUTHORS Vidal,F., Farssac,E. and Gallardo,D.  
 TITLE Homo sapiens factor VIII (F8c) gene, exon 17  
 JOURNAL Unpublished

REFERENCE 2 (residues 1 to 76)  
 AUTHORS Vidal,F., Farssac,E. and Gallardo,D.  
 TITLE Direct Submission  
 JOURNAL Submitted (31-JUL-1998) Unitat de Recerca, Centre de Transfusio i Banc de Teixits, Pg Vall d'Hebron, 119-129, Barcelona 08035, Spain

COMMENT Method: conceptual translation.

FEATURES  
 Location/Qualifiers  
 source 1..76  
 /organism="Homo sapiens"  
 /db\_xref="taxon:9606"  
 /chromosome="X"  
 /map="Xq28"  
 Protein <1..>76  
 /product="coagulation factor VIII"  
 CDS 1..76  
 /gene="F8c"  
 /coded\_by="AF081784.1:<20..>248"

ORIGIN  
 1 ekdvhsigli pllvchtntl npahgrqvvtv qefalfftif detkswyfte nmerncrapc  
 61 niquecptfk enyrfh  
 //

**Analyze this sequence**  
 Run BLAST  
 Identify Conserved Domains  
 Find in this Sequence

**Articles about the F8 gene**  
 Mutation screening of the F VIII gene in 10 hem [Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2011]  
 Reversible activation of cellular factor XIII by calcium. [J Biol Chem. 2011]  
 Factor VIII A3 domain substitution N1922S results in hemophilia A due to domain-speci [Blood. 2011]  
 See all...

**Pathways for the F8 gene**  
 Blood Clotting Cascade  
 Complement and coagulation cascades  
 Complement and Coagulation Cascades  
 See all...

**Reference sequence information**  
 RefSeq genomic sequence  
 See the genomic reference sequence for the F8 gene (NG\_011403.1).  
 RefSeq protein isoforms  
 See 2 reference sequence protein isoforms for the F8 gene.

**More about the F8 gene**  
 This gene encodes coagulation factor VIII, which participates in the intrinsic pathway of blood coagulation, factor VIII is a cofactor for f...  
 Also Known As: RP11-115M6.7, AHF, DXS1...

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

### 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).

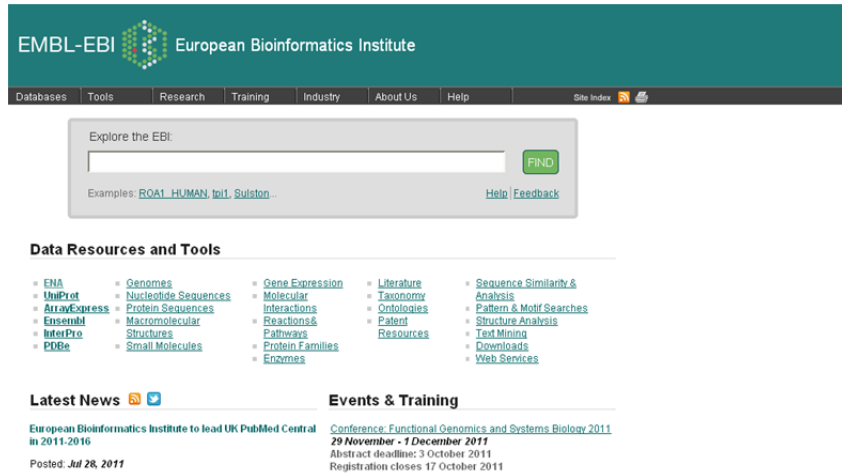


Figura 11: Página de acceso a EMBL

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.L.;
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="G00-119-124"
FT /experiment="experimental evidence, no additional details
FT recorded"
XX
SQ Sequence 1493 BP; 439 A; 265 C; 379 G; 410 T; 0 other:
gagctcacc a tggctacatt ctggtgtaa gagatatac ctatacctgg gccaaatgta 60
aacagcctgg aaagtgtaa ggttaaaac aaacaaaat aataaatga ataatgcc 120
ggtggttatg agtgcattg agaaaatga agccaagag gatatacag atgcaggtyg 180
ggtaaaagag cttacaacat aaatgtgtg ttcctattt aaacctcat caacaggtaa 240
gattggagct gaatgtgaa gtagttgtg gagtggaaact acgtgggaaa ttgggggaaa 300
aggtgttttg gtaaaagaa atagcaagtg ttgaggtcca ggggcatgag tgtgttgat 360
atttaggga agatgaagga gaccagata accagatga gatgagacta cagaggtcag 420
gagaagggc atgcagacca tgtggatgc tctagacct aggccatggt aaagatgag 480
ggttttacc tgatggaggt cagaagcat tggaggatc tgagaagagg atgacagga 540
ctcgtttat agtttaaat tataactata aattatagt ttaaaacaa tagtgccta 600
acctatgtt atatgaaaa ctacagttt aaaaactata aattcctcat actggcaga 660
gtgtgaggg caagggcaaa agcagagaga ctaacagtt gctggttact ctgtcagt 720
caagtgaat ctagaatctt cgacaacat cagaactct ctgtgctg caactcaga 780
agaggggttg agtagctag gaatagagc acaaatataa gctcctgtc actttgact 840
ctccatcct ctcctcctt ccttaaggt ttgattaaa gcagacttat gccctactg 900
ctctcagaag tgaatgggtt aagtttagca gctcctctt tgcacttca gttctctct 960
tggctgctc ccatgataa aaaggaagca atcctatcgg tctactgta gctgtgaga 1020
atccagtggt taagttcct taaatgctc tgcaagaaa ttgggactt tcaataatc 1080
agaaattta ctttttccc ctctgggag ctaaatgat tttagagaag aataaacct 1140
ttgctctcc agtgaacat ttgactaat aagtcatgca aatagagctc tccactgct 1200
ttcttctgt cctttgaga ttctgctta gtgccaccag aagatactac ctgggtgag 1260
tggaaactgc atgggacta atgcaagtg atctcgtgta gctgctgtg gacgcaagt 1320
aaagcactg cctgaggtt ctgctcggg ccagattgt ggggatgaa gctgcttgy 1380
aggaaggtgc agcatcggg ttaggatgt tgtgatgta cctgggccc aaagaacat 1440
ttctgggtaa ggtgtgaca cactctgtt attagcaga atgtaactg ccc 1493
//
    
```

Figura 12: Secuencia de ADN en formato EMBL



### 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

### 3.2.4. 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).

The first completed genomes from [viruses](#), [phages](#) and [organelles](#) were deposited into the EMBL Database in the early 1980's. Since then, molecular biology's shift to obtain the complete sequences of as many genomes as possible combined with major developments in sequencing technology resulted in hundreds of complete genome sequences being added to the database, including [Archaea](#), [Bacteria](#) and [Eukaryota](#). These web pages give access to a large number of complete genomes, [here](#) is available to describe the layout.

**Whole Genome Shotgun Sequences (WGS)**

Methods using whole genome shotgun data are used to gain a large amount of genome coverage for an organism. WGS data for a growing number of organisms are being submitted to DDBJ/EMBL/GenBank.

[More information about WGS projects...](#)

**Last 40 Genome Entries**

Date	Accession	Description
05-AUG-2011	<a href="#">AP012267.1</a>	Equus przewalskii mitochondrial DNA, isolate: Belina
05-AUG-2011	<a href="#">AP012268.1</a>	Equus przewalskii mitochondrial DNA, isolate: Anushka
05-AUG-2011	<a href="#">AP012269.1</a>	Equus przewalskii mitochondrial DNA, isolate: Bonnette
05-AUG-2011	<a href="#">AP012270.1</a>	Equus przewalskii mitochondrial DNA, isolate: Bars
05-AUG-2011	<a href="#">AP012271.1</a>	Equus asinus somalicus mitochondrial DNA
05-AUG-2011	<a href="#">HE577054.1</a>	Paenibacillus polymyxa M1 main chromosome
05-AUG-2011	<a href="#">HE577055.1</a>	Paenibacillus polymyxa M1 plasmid pPPM1a, complete replicon
04-AUG-2011	<a href="#">CP002955.1</a>	Cyclobacterium marinum DSM 745

Figura 14: Página de acceso a *Genomes Pages*

### 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. 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.

```
MAVMAPRTLIV LLLSGALALT QTWAGSHSMR YFSTSVSRPG RGEPRFIAVG YVDDTQFVRF
DSDAASQRME PRAPWIEQEG PEYWRNTRN VKAHSQTDRV DLGTLRGYYN QSEDGSHTIQ
RMYGCDVGS DGRFLRGYQQD AYDGKDYIAL NEDLRSWTAA DMAAEITKRK WEAHFHAEQL
RAYLEGTCTVE WLRRLHLENGK ETLQRTDAPK THMTHHAVSD HEAILRCWAL SFYPAEITLT
WQRDGEDQTO DTELVETRPA GDGTFQKWAA VVVPSCQEQR YTCHVQHEGL PEPLTLRWEP
SSOPTIPIVG ILAGLVLFGA VIAGAVVAAV RWRRKSSDRK GGSYSQAASS DSAQGSVSL
TACKV
```

Figura 15: Secuencia de una proteína anónima

Los algoritmos de rastreo de bases de datos más conocidos son FASTA y BLAST, implementados por los programas FASTA, BLASTn (ADN) y BLASTp (proteínas).

Vamos a rastrear las bases de datos de proteínas con la secuencia de ejemplo de la Figura 15 utilizando el programa BLASTp (Figura 16). En el interfaz gráfico del programa, encontramos una caja de texto donde podemos pegar la secuencia problema, así como un botón *Choose File* 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.

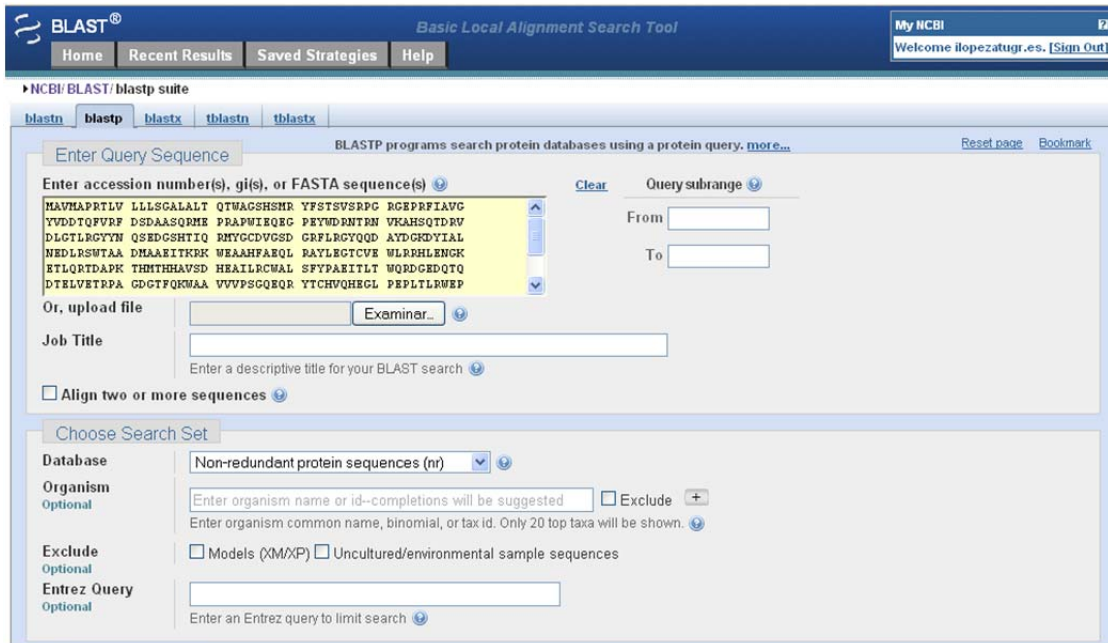


Figura 16: Página de acceso a BLASTp

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.

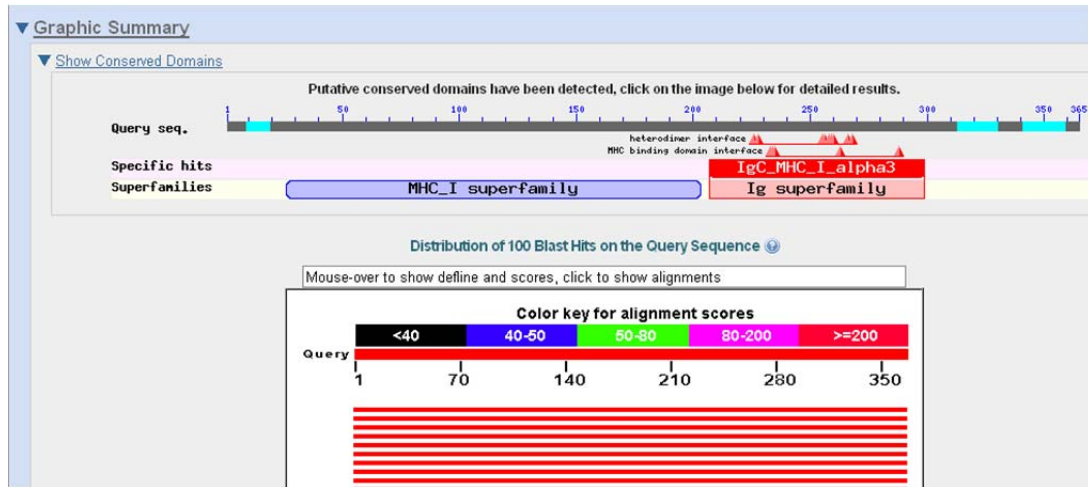


Figura 17: Resumen gráfico del resultado de un rastreo con BLASTp

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.

Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer [P](#) PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
<a href="#">P30375.1</a>	RecName: Full=Class I histocompatibility antigen, Gogo-A*0101 alpha	757	757	100%	0.0	
<a href="#">CAA11708.1</a>	human leucocyte antigen A [Homo sapiens]	723	723	100%	0.0	<a href="#">G</a>
<a href="#">AAB41292.1</a>	HLA class I A locus antigen A*68new [Homo sapiens]	723	723	100%	0.0	<a href="#">G</a> <a href="#">M</a>
<a href="#">P01891.4</a>	RecName: Full=HLA class I histocompatibility antigen, A-68 alpha cha	720	720	100%	0.0	<a href="#">G</a>
<a href="#">CBX51211.1</a>	MHC class I antigen [Homo sapiens]	720	720	100%	0.0	
<a href="#">P10316.2</a>	RecName: Full=HLA class I histocompatibility antigen, A-69 alpha cha	719	719	100%	0.0	<a href="#">G</a> <a href="#">M</a>
<a href="#">CAD23134.1</a>	MHC class I antigen [Homo sapiens]	718	718	100%	0.0	<a href="#">G</a>
<a href="#">CAB59722.1</a>	human leucocyte antigen A [Homo sapiens]	717	717	100%	0.0	<a href="#">G</a>
<a href="#">CRW44121.1</a>	MHC class I antigen [Homo sapiens]	716	716	100%	0.0	
<a href="#">AAA03602.1</a>	HLA-A-6802 [Homo sapiens] >emb CAJ84549.1  MHC class I antigen	716	716	100%	0.0	<a href="#">G</a>
<a href="#">CAD97419.1</a>	MHC class I antigen precursor [Homo sapiens]	715	715	100%	0.0	<a href="#">G</a>
<a href="#">ABP02054.1</a>	MHC class I antigen [Homo sapiens]	714	714	99%	0.0	
<a href="#">CAA80612.1</a>	HLA-A*0210 [Homo sapiens] >gb ACR55713.1  MHC class I antigen [	713	713	100%	0.0	<a href="#">G</a> <a href="#">M</a>
<a href="#">AAD02067.1</a>	MHC class I antigen [Homo sapiens]	713	713	100%	0.0	<a href="#">G</a> <a href="#">M</a>
<a href="#">CAA65501.1</a>	human leucocyte antigen [Homo sapiens] >gb ACR55716.1  MHC clas	713	713	100%	0.0	<a href="#">G</a> <a href="#">M</a>
<a href="#">AAV51797.1</a>	MHC class I antigen [Homo sapiens]	712	712	100%	0.0	<a href="#">G</a> <a href="#">M</a>

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.

```

▼ Alignments
 Select All  Get selected sequences  Distance tree of results  Multiple alignment

>  sp|P30375.1|IA01 GORGO  RecName: Full=Class I histocompatibility antigen, Gogo-A*0101
alpha chain: Flags: Precursor
emb|CAA42810.1| Mhc class I heavy chain [Gorilla gorilla]
Length=365

Score = 757 bits (1954), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 365/365 (100%), Positives = 365/365 (100%), Gaps = 0/365 (0%)

Query 1  MAVHAFRTLWLLLSGALALTQTWAGSHSHRYFSTSVSRPGRGEPRFLAVGYVDDTQFVRF 60
Sbjct 1  MAVHAFRTLWLLLSGALALTQTWAGSHSHRYFSTSVSRPGRGEPRFLAVGYVDDTQFVRF 60

Query 61  DSDAASQRMEPRAPWIEQEPEYMDRNRNPKAHSQTRDRLGLTLRGYYNQSEDSHTIQ 120
Sbjct 61  DSDAASQRMEPRAPWIEQEPEYMDRNRNPKAHSQTRDRLGLTLRGYYNQSEDSHTIQ 120

Query 121  RMYGCDVGSDCGRFLRGYQDAYDQKDYIALNEDLRSTAAADMAAEITKPKWAAHFAEQ 180
Sbjct 121  RMYGCDVGSDCGRFLRGYQDAYDQKDYIALNEDLRSTAAADMAAEITKPKWAAHFAEQ 180

Query 181  RAYLEGTVCVEWLRHLENGKETLQRDAPKTHMTHHAVSDHEAILRCWALSFPYAEITLT 240
Sbjct 181  RAYLEGTVCVEWLRHLENGKETLQRDAPKTHMTHHAVSDHEAILRCWALSFPYAEITLT 240

Query 241  WQRDGEDQTQDTELVETRPAGDGTQKMAAVVPSGQEQRYTCHVQHEGLPELTLRMEP 300
Sbjct 241  WQRDGEDQTQDTELVETRPAGDGTQKMAAVVPSGQEQRYTCHVQHEGLPELTLRMEP 300

Query 301  SSQPTIPIVGIILAGLVLFCAVIAGAVVAAVRWRKSSDRKGGSSYQAASDQAQSDVSL 360
Sbjct 301  SSQPTIPIVGIILAGLVLFCAVIAGAVVAAVRWRKSSDRKGGSSYQAASDQAQSDVSL 360

Query 361  TACKV 365
Sbjct 361  TACKV 365

```

Figura 19: Alineamiento en un rastreo con BLASTp

### 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. Esta 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](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.

## 4. CUESTIONES

### Trabajos de autoaprendizaje

- Búsqueda de secuencias de ADN y proteínas propuestas por el profesor.
- Búsqueda de secuencias de ADN y proteínas de interés para el estudiante.
- Rastreo de bases de datos con proteínas problema propuestas por el profesor (enlazadas en la página web).

### Trabajos de evaluación

- Dada una secuencia de ADN anónima, determinar el tipo de secuencia de que se trata mediante rastreo de una base de datos.
- Dada una secuencia problema de aminoácidos, buscar al menos tres proteínas diferentes que presenten similitud (homología) con ella.
- Localizar un gen concreto (problema) en el genoma utilizando un navegador genómico y resumir la información que se extraiga del mismo.

# PREDICCIÓN COMPUTACIONAL DE GENES



# PREDICCIÓN COMPUTACIONAL DE GENES

## 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.

## 2. FUNDAMENTO TEÓRICO

### 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.

## 3. METODOLOGÍA

### 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 2.653.211 y 2.657.767. Obtendremos esta secuencia en la base de datos *Ensembl*:

<http://www.ensembl.org>

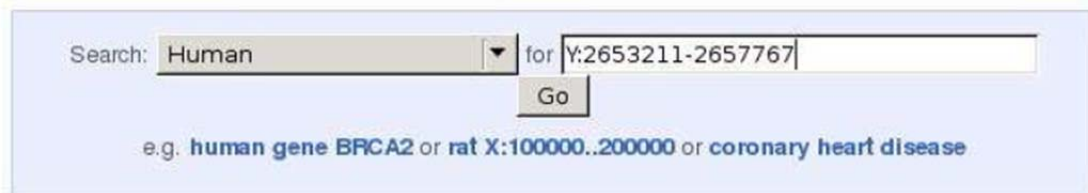
*Ensembl* es una base de datos donde se recogen los genomas de multitud de organismos que se anotan mediante una serie de programas que localizan en las secuencias distintos tipos de características y, entre ellas, la localización de genes, exones e intrones. La base de datos se puede consultar “on line” 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:2653211-2657767

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.

### 3.2. Predicción de ORFs

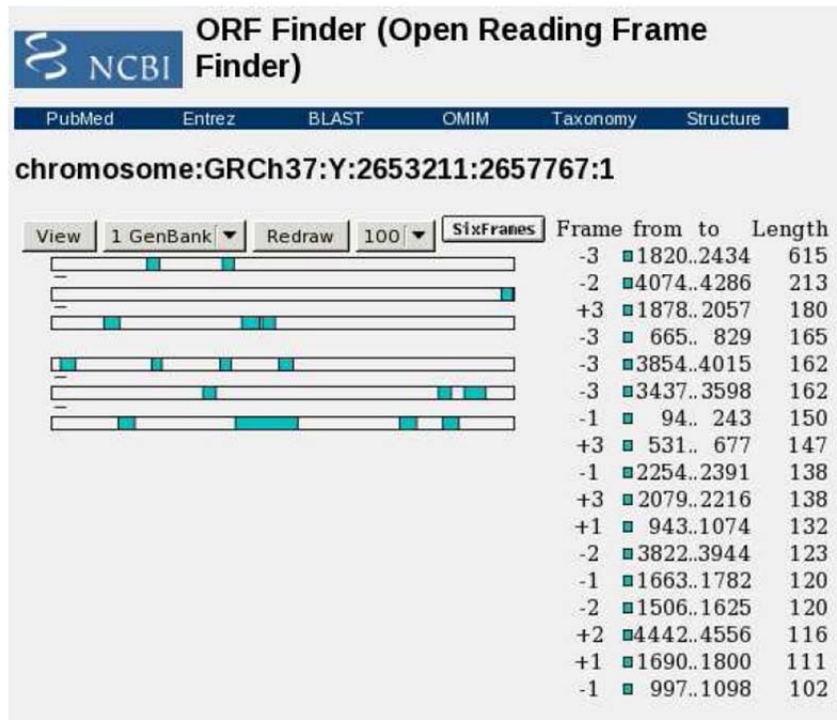
Las *ORF*, del inglés *Open Reading Frame*, o *Marco de Lectura Abierta*, consisten en un fragmento de secuencia que comienza en un codon de inicio y termina en un codon

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:

<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>

pegando la secuencia en el recuadro de texto:

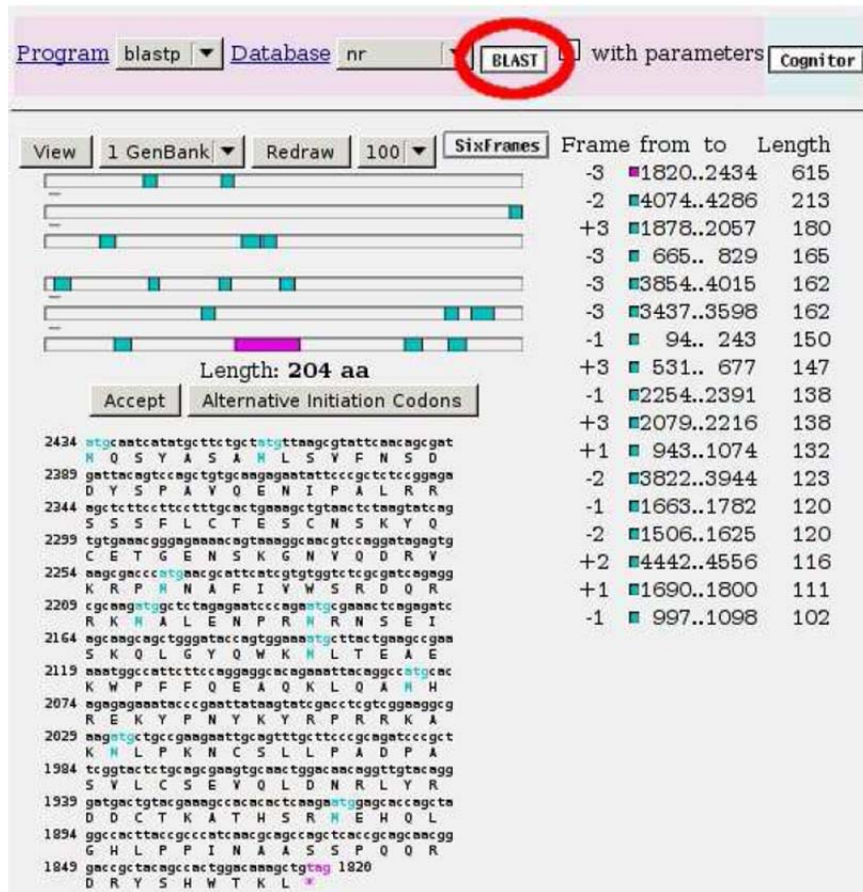
Tras pulsar sobre el botón “ORFFind” el programa buscará las *ORF* y mostrará el resultado como se observa a continuación:





A la izquierda vemos una representación gráfica de las 6 pautas de lectura con las ORF señaladas en color (tres por cada hebra del ADN). A la derecha se indica los nucleótidos donde comienzan y terminan y su longitud total. En la última pauta de lectura se observa una ORF de mayor longitud que las demás, que podría ser indicativa de la presencia de una región codificante.

Picando sobre esa ORF veremos como cambia de color. En la lista de ORF de la derecha también se señalan en color diferente los datos que corresponden a esa ORF y aparece la secuencia de la posible traducción de ese fragmento, indicando con colores los codones de inicio y stop:



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 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.



Query Protein Sequence (204 letters)  
 Database nr  
 Job title Protein Sequence (204 letters)  
 Request ID JUZYJWH9014 View report  Show results in a new window

Format

Show  as   Advanced View  Use old BLAST report format [Reset form to default](#)

Alignment View

Display  Graphical Overview  Linkout  Sequence Retrieval  NCBI-gi

Masking Character:  Color:

Limit results Descriptions:  Graphical overview:  Alignments:

Organism  Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.  
 Enter organism name or id--completions will be suggested  Exclude

Entrez query:

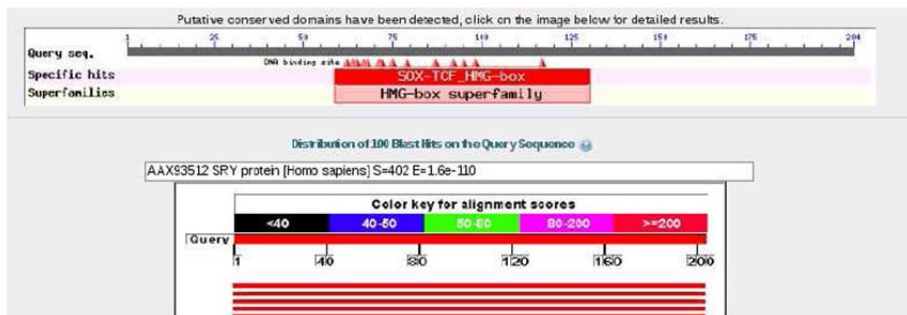
Expect Min:  Expect Max:

Percent Identity Min:  Percent Identity Max:

Format for  PSI-BLAST with inclusion threshold:

Pulsamos entonces sobre el botón “View report” y es posible que tengamos que esperar a que se actualice la página una vez obtenidos los resultados.

Parte de los resultados de BLAST se muestran a continuación:

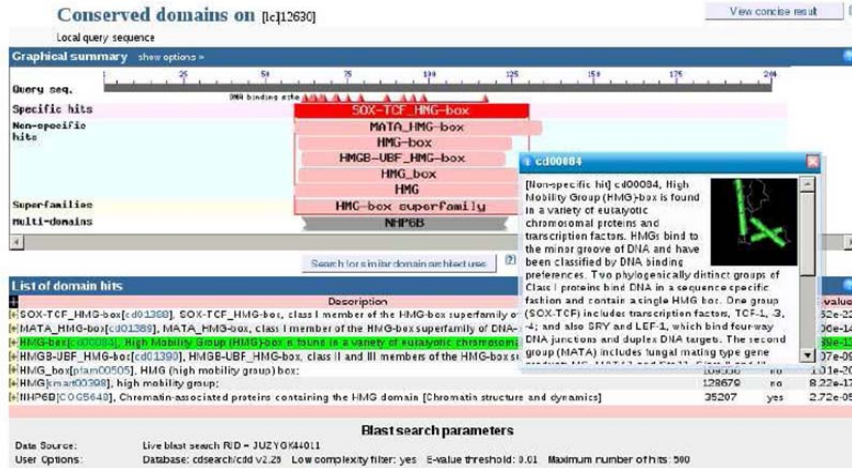


En la parte superior podemos ver que se ha localizado un dominio conservado de tipo HMG. Las líneas inferiores representan las secuencias encontradas con homología con la secuencia de búsqueda. Colocando el cursor sobre la primera línea roja, en el recuadro de texto sobre las líneas aparece información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

AAX93512 SRY (Sex-Determining Region Y) protein [Homo sapiens] S=402 E=1.6e<sup>-110</sup>

donde **S** 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.

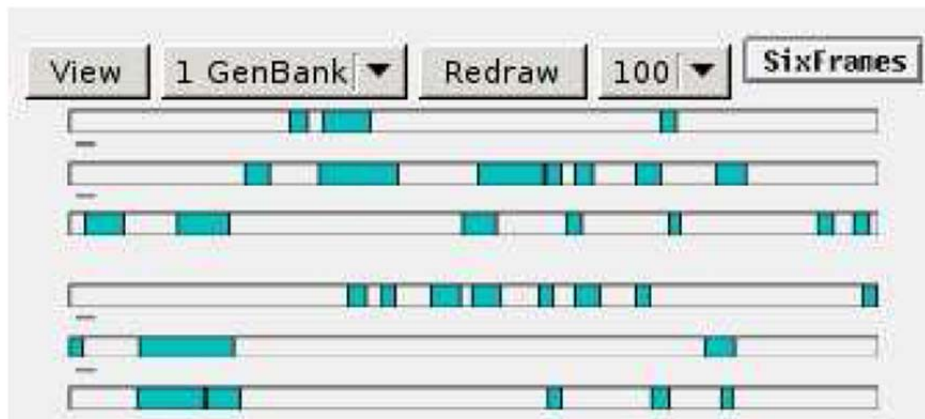
Pulsando sobre la caja HMG de la parte superior accedemos a una página con información adicional sobre este tipo de dominio:



### 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 70116646 a 70123266 del cromosoma 17.

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 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:

```

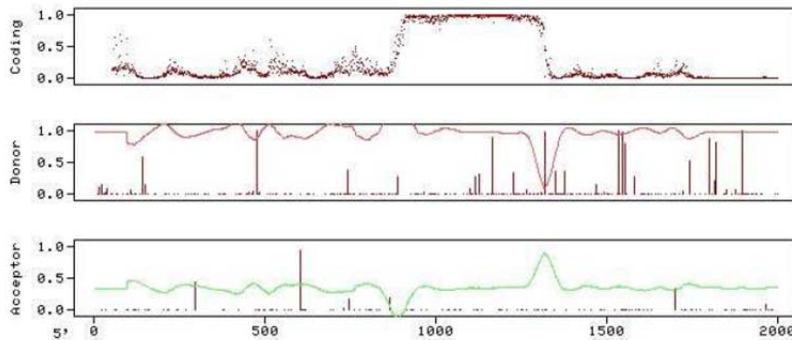
***** NetGene2 v. 2.4 *****

The sequence: 17 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     AGTGAAGAG^GTGAGGGAGG
    
```

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:

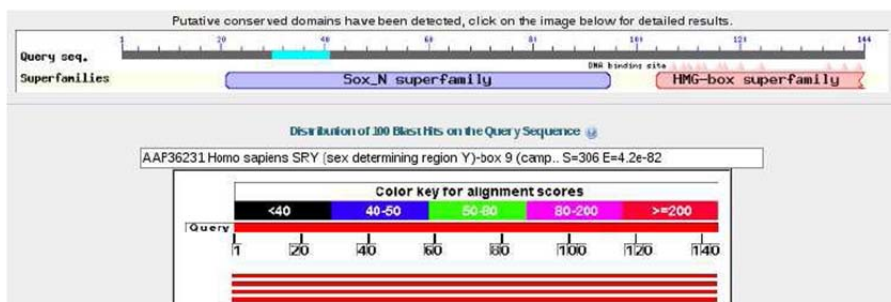


Los tres gráficos corresponden, de arriba a abajo, al potencial codificante, a la localización de sitios donadores y, por último, a la de sitios aceptores. Las líneas verticales corresponden a los posibles puntos donadores y aceptores a los que referían los datos anteriores. La longitud de las líneas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

### 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:

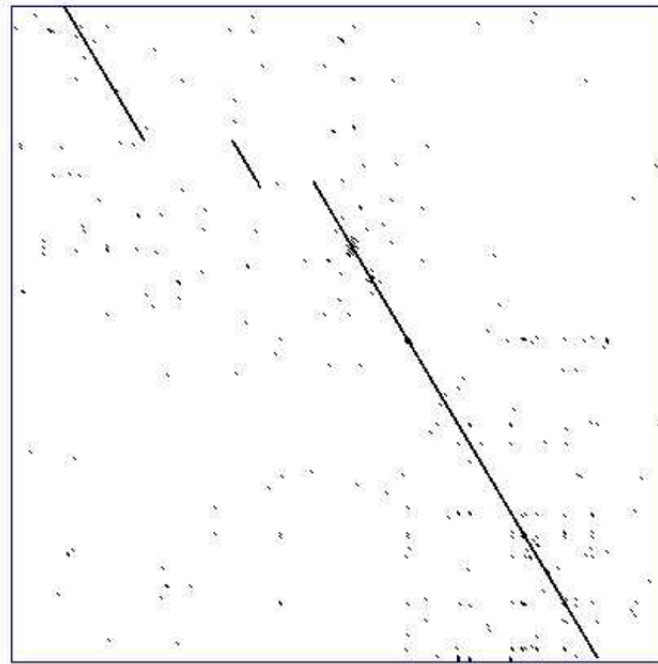
- [Homo sapiens SRY \(sex determining region Y\)-box 9 \(SOX9\), mRNA](#)
- 7. 3,963 bp linear mRNA
- NM\_000346.3 GI:182765453
- [GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

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

<http://www.vivo.colostate.edu/molkit/dnadot/>

nos mostrará la posición de los intrones y exones:





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

#### 4. CUESTIONES

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

- ¿Que indican las 6 barras que aparecen en la pantalla?
- Dentro de estas barras, ¿que indican los fragmentos coloreados?
- ¿Que indican cada una de las columnas numéricas que aparecen a la derecha (“frame”, “from to”, y “length”)?
- De todos los posibles marcos de lecturas abiertos, ¿Cual es el que tiene más probabilidad de ser una región codificante?
- ¿Por qué todos los marcos de lectura abierta comienzan con el triplete ATG?  
¿Con que triplete/s terminan?
- ¿Qué podría ocurrir si un marco de lectura abierto se encuentra interrumpido por un intrón?
- ¿Qué condiciones deberían darse para que un intrón quedase englobado en un marco abierto de lectura?
- ¿Cómo se puede comprobar que un marco de lectura abierta codifica una proteína conocida?
- ¿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:

- ¿Que indican las tablas numéricas “donor splice sites” y “acceptor splice sites”?  
¿Qué indica la última columna en dichas tablas?
- En la columna “confidence”, ¿que indican los valores altos?
- ¿Que se representa en la gráfica superior?

- En las dos gráficas inferiores, ¿Qué representan las líneas verticales?, ¿Y la línea horizontal (roja o verde)?
- Atendiendo a las gráficas, ¿Cómo identificarías lugares con una alta probabilidad de ser donadores/aceptores funcionales?
- ¿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?
- ¿Por qué dos exones del mismo gen pueden aparecer en pautas de lectura diferentes?
- ¿Por qué el potencial codificante decae antes de llegar al final del último exón?





ALINEAMIENTO MÚLTIPLE DE  
SECUENCIAS DE ADN y  
PROTEÍNAS



# ALINEAMIENTO MÚLTIPLE DE SECUENCIAS DE ADN y PROTEÍNAS

## 1. 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.

## 2. 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'-AATGTCATGCGCTGAATCCCCC-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 sustitución 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 cuantos 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 sustituciones). 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 sustituciones 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. Nosotros utilizaremos el programa Clustal X 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.

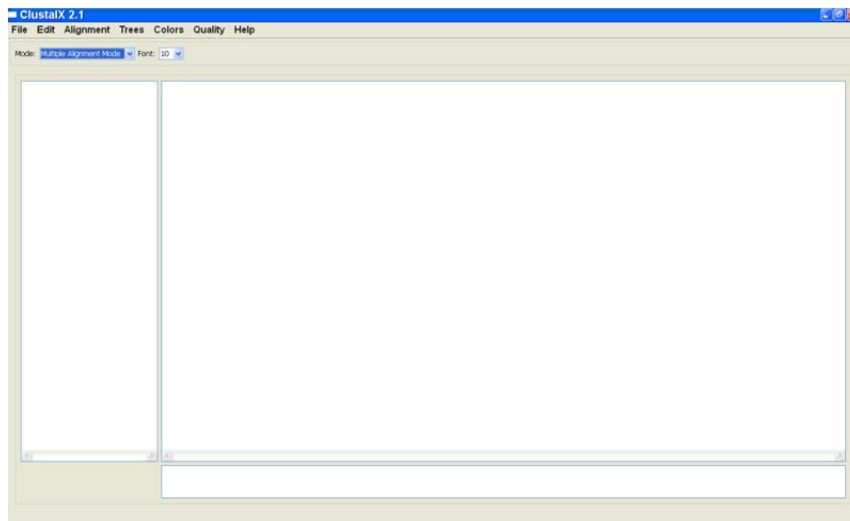
### 3. METODOLOGÍA

El programa ClustalX tiene una interfaz gráfica disponible para diversos sistemas operativos, como Windows o Linux, del programa de alineamiento múltiple de secuencias Clustal W (ver referencias al final de este guión). El diseño del programa permite también la construcción de árboles filogenéticos a partir de las secuencias alineadas.

Este programa reconoce siete formatos diferentes de ficheros de entrada (*input*) con secuencias tanto de ADN como de proteínas: NBRF/PIR, EMBL/SWISSPROT, Pearson (Fasta), Clustal (\*.aln), GCG/MSF (Pileup), GCG9/RSF y GDE. Todos los caracteres (espacios, dígitos, signos de puntuación) se ignoran a excepción de los guiones (“-”) que indican *gaps* (en MSF/RSF los *gaps*, no obstante, son reconocidos como “.”). Nosotros usaremos el formato FASTA, para el que se dan dos ejemplos en el primer Apéndice.

Para obtener el alineamiento procederemos de la siguiente manera:

**Paso 1.** Al abrir el programa aparecerá una ventana como esta:



El programa Clustal X tiene dos modos de actuación diferentes que se pueden seleccionar alternativamente en la ventana ‘MODE’ que está bajo la lista de menús: ‘MULTIPLE ALIGNMENT MODE’ y ‘PROFILE ALIGNMENT MODE’. El primer modo es el que vamos a usar siempre en esta práctica. Nos permite hacer un alineamiento a partir de un conjunto de secuencias. El segundo modo permite trabajar simultáneamente con dos conjuntos de secuencias (cada conjunto se denomina, en este caso, perfil) en dos ventanas diferenciadas e intercambiar secuencias de un perfil a otro así como alinear las secuencias de un perfil con las del otro.

**Paso 2.** A continuación, seleccionamos el fichero que contiene nuestras secuencias en formato Fasta. Para ello desplegamos el menú ‘FILE’, seleccionamos ‘LOAD SEQUENCES’ y navegando por los directorios, descargamos el fichero de entrada cuyas secuencias aparecerán en la ventana del programa de esta manera:



Se pueden añadir nuevas secuencias usando la opción 'APPEND SEQUENCES'.

De la misma manera, desplegando el menú 'EDIT', se pueden eliminar secuencias de manera temporal ('CLEAR SEQUENCE SELECTION'), así como cambiarlas de orden mediante las opciones 'CUT SEQUENCES' y 'PASTE SEQUENCES'.

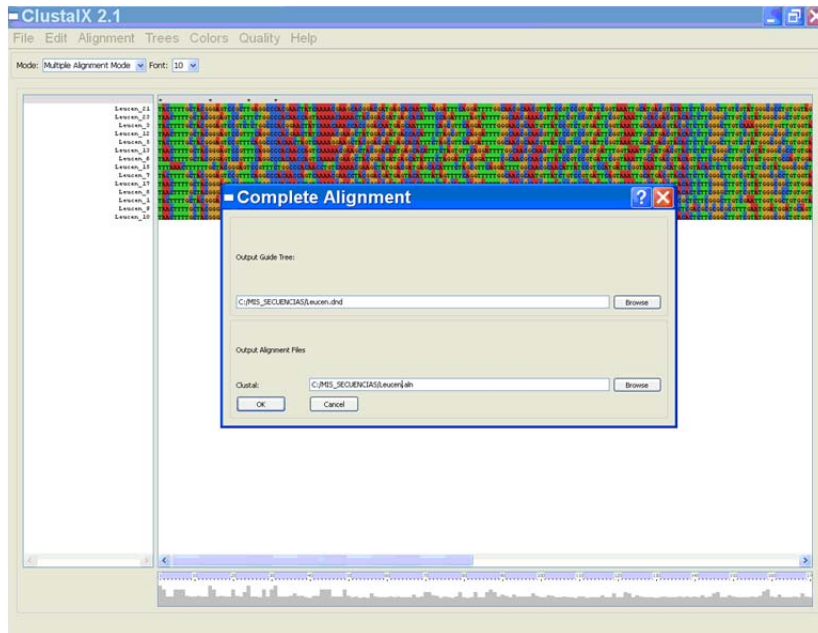
Otras posibilidades de edición (Menú 'EDIT') son 'SEARCH FOR STRING' (buscar en las secuencias una sucesión determinada de residuos, nucleótidos o aminoácidos), 'REMOVE ALL GAPS' (elimina todos los *gaps* existentes en las secuencias, tanto los existentes en el fichero original como los introducidos por Clustal X en el alineamiento) y 'REMOVE GAP-ONLY COLUMNS' (elimina las posiciones en las que en todas las secuencias ya alineadas existe un *gap*, consecuencia de haber eliminado alguna secuencia divergente o tras un re-alineamiento).

**Paso 3.** Procederemos a continuación a realizar el alineamiento múltiple. 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 alineando secuencias de manera progresiva de acuerdo al orden de ramificación del árbol.

Las tres etapas se suceden de manera automática si seleccionamos la opción 'DO COMPLETE ALIGNMENT'. Se puede saltar alguno de los pasos seleccionando 'DO ALIGNMENT FROM GUIDE TREE' (si se dispone de un dendrograma previo que actúa como guía para el alineamiento múltiple) o 'PRODUCE GUIDE TREE ONLY' (si sólo se quiere obtener el árbol de referencia obviando la construcción del alineamiento múltiple).

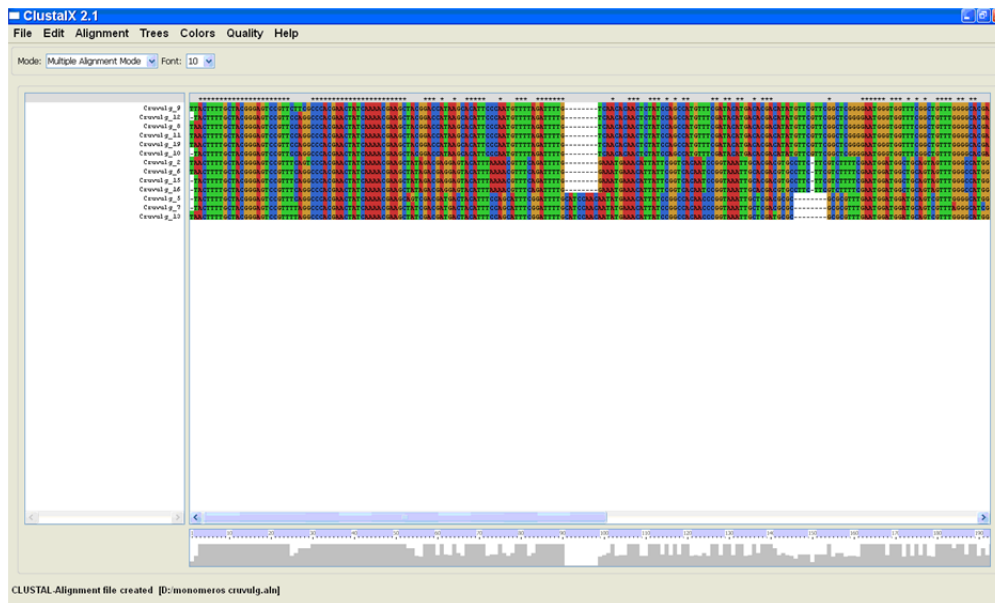
En cualquiera de los tres casos, siempre nos aparecerá una ventana emergente sobre la ventana principal en la que debemos indicar el directorio donde queremos guardar los ficheros de salida (*output*) y el nombre que le daremos a dichos ficheros. Las extensiones son \*.dnd (para el fichero del dendrograma) y \*.aln (para el del alineamiento):





En el Apéndice 2 se muestran dos ficheros de salida típicos: uno con el alineamiento múltiple de un conjunto de secuencias de ADN y otro en el que se describe el dendrograma obtenido en el proceso.

Además, el alineamiento se puede visualizar en la ventana principal del Programa:



Los nucleótidos (o los aminoácidos, según la secuencia utilizada) aparecen coloreados para facilitar el análisis visual del alineamiento. El histograma que aparece debajo de dicho alineamiento indica el grado de similitud entre las secuencias alineadas en cada posición. Los picos altos indican la mayor similitud mientras que los valles indican baja similitud.

En el apéndice 3 se dispone de una descripción de diferentes opciones y parámetros a tener en cuenta a la hora de hacer un alineamiento. En nuestro caso, además de tener en cuenta el resto de opciones, con respecto a las siguientes dos, procederemos así:

- '*PAIRWISE ALIGNMENT PARAMETERS*': Utilizaremos siempre la opción 'Slow-Accurate'. Los parámetros que aparecen por defecto se utilizarán en primera opción y probaremos después varias posibilidades de acuerdo al tipo de secuencias que tengamos.

- '*MULTIPLE ALIGNMENT PARAMETERS*': Utilizaremos los parámetros establecidos por defecto y probaremos después varias posibilidades de acuerdo al tipo de secuencias que tengamos.

**Paso 4.** A continuación, obtendremos un árbol filogenético de las secuencias utilizando en el menú 'TREES' la opción 'DRAW TREE' (extensión de los ficheros de salida: \*.ph).

El programa Clustal X sólo permite la construcción de árboles filogenéticos de secuencias mediante métodos basados en distancias genéticas. Incluye dos algoritmos: UPGMA y Neighbor-Joining (N-J). La opción por defecto es N-J pero se puede cambiar en la opción 'CLUSTERING ALGORITHM' del menú 'TREES'.

El método N-J genera un árbol sin raíz. La raíz del árbol sólo podrá inferirse si se usa una secuencia de referencia externa (*outgroup*: una secuencia que estamos seguros que está en la base del árbol de acuerdo a conocimientos que tengamos de su procedencia biológica). Alternativamente, en el caso de que se carezca de *outgroup*, si asumimos constancia en las tasas de cambio en todas las ramas del árbol ("reloj molecular"), podemos situar la raíz del árbol en el centro del árbol (aproximadamente equidistante a todos los extremos de las ramas). El algoritmo UPGMA genera árboles con raíz.

Podemos evaluar la significación estadística de la topología del árbol obtenido si, para construir el árbol filogenético, utilizamos la opción 'BOOTSTRAP N-J TREE' que nos dará los valores de *bootstrap* para cada nodo del árbol (extensión de los ficheros de salida: \*.php).

En el apéndice 4 se indican algunas directrices a seguir en cuanto a parámetros tener en cuenta a la hora de reconstruir los árboles filogenéticos.

## 4. CUESTIONES

1. Utilizar los ficheros de secuencias de ADN dispuestas en formato FASTA disponibles en la plataforma virtual de enseñanza de las asignaturas para obtener los correspondientes alineamientos múltiples y un árbol filogenético a partir de dichos alineamientos con el programa ClustalX.

2. Realizar búsquedas de 10 secuencias homólogas en bases de datos y realizar un alineamiento múltiple.

## REFERENCIAS

### *Version 2 de ClustalW y ClustalX:*

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947-2948.

### *ClustalX:*

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 25:4876-4882.

### *ClustalW:*

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673-4680.

### *Versiones anteriores:*

#### *Clustal V:*

Higgins, D.G., Bleasby, A.J. and Fuchs, R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *CABIOS* 8, 189-191.

#### *Clustal original:*

Higgins, D.G. and Sharp, P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *CABIOS* 5, 151-153.

Higgins, D.G. and Sharp, P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 73, 237-244.

### *Programa NJPlot:*

Perrière, G. and Gouy, M. (1996) WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, 78, 364-369.

**La última versión de ClustalX se puede conseguir en las siguientes direcciones web:**

<http://www.clustal.org>

<http://www.ebi.ac.uk/Tools/clustalw2/>

**Tutorial on-line:**

<http://www.ebi.ac.uk/2can/tutorials/nucleotide/clustalw.html>

## APÉNDICE 1: FORMATO FASTA

### Ejemplo para secuencias de ADN:

```
>Nombre de la secuencia 1
GTTATCGTAAATTAAATCCAATGGTATTCTGTGGTTTATATCCGATTGATTCTAGAAAAATTAGAATT-
AAACGATTCTGCACTTGAGTTTGAACCAGAAACATTGGATTCTTAGGACTTCTTCATATGGAAATCCTTC
AAGAACGTATTGAACGT
>Nombre de la secuencia 2
GTTATCCTAACTTAAATCCAATGGTTTTCTCTGGTTTATATCCGATTGATTCTTAGAAAAATTAGAATT-
AAACGTTTCTGCAATTGAGTTTGAACCAGAAACATTGGATTCTTAGGACTTCTTCATATGGAAATCCTTC
GGGTACGTATTGAACGT
>Nombre de la secuencia 3
CTTTTCTATCTTATATCCAATCCTTATGGCTCGTATATATCCGATTGATTCTTAGAAAAATTAGAATTA
AAAGGTTTCTGCATTTGAGTTTGAACCAGAAACATTGGATTCTTTGGGCTTCTTCATATGGTTTTTCTTG
GGGTACGTATTGAACGT
```

### Ejemplo para secuencias de proteínas:

```
>Nombre de la secuencia 1
KHPIILEKLEFPDPVISMAIEPKTKKDQEKLSQVLNKFMK-EDPTFRATTDPETGQILIHGMGELHLEIM
VDRMKREYGIENVNGKPQVAYKETIRKKAIGEGKFIKQTGGRGQYGHAIIEIEPLPRGAGFEFIDDIHGG
VIPKEFIPSV
>Nombre de la secuencia 2
KHPLILEKLFFPDPVMSMAIEPKKKKDQEKLSQVLNKFMKKEEDPTFAATTDPETGQILIHGMGELHLEIM
VDRMKREYGIENVNGKPQVAYKETIRKKAIGEGKFIKQTGGRGQYGHAIIEEPLPRGAGFEFIDDIHGG
VIPFEFIPSV
>Nombre de la secuencia 3
PHPLILEKLEFPDPVMAMAIEPTKKKDQEKLQQVLNKFMKKEEPTFA-TTDPETGQILIHGMGELHLEIM
DDRMKREYGIENVNGKPQVAYKETIRKKAIGEGKFIKQTGGRGQYGHAIIEEPLPRGAGFEFIDDIHGG
VIPFEFPP-V
```

## APÉNDICE 2: FICHEROS DE SALIDA TRAS APLICAR LA OPCIÓN 'DO COMPLETE ALIGNMENT'

### Alineamiento múltiple (archivo \*.aln): ADN.

CLUSTAL 2.0.9 multiple sequence alignment

```

Cruvulg_9      TTACTTTTGTACGGGAGTCCGTTCTTCGCCACGAACTATCAAAACGAAGCTACGGACC
Cruvulg_12    T-ACTTTTGTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
Cruvulg_8      TAACTTTTGTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
Cruvulg_11    TAACTTTTGTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
Cruvulg_19    TAACTTTTGTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
Cruvulg_10    T-ACTTTTGTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
Cruvulg_2      TAACTTTTGTACGGGAGTCCGTTTCAGTCCCACGAACTATCAAAACGAAGCTATAGACG
Cruvulg_6      TAACTTTTGTACGGGAGTCCGTTTCAGTCCCACGAACTATCAAAACGAAGCTATAGACG
Cruvulg_15    T-ACTTTTGTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
Cruvulg_16    T-ACTTTTGTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
Cruvulg_5      T-ACTTTTGTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
Cruvulg_7      T-ACTTTTGTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
Cruvulg_13    TAACTTTTGTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
                *****
                *****

Cruvulg_9      ATAAGCACATTCCCAATGTTTTAGATTTG-----TCAACACAACCTCTATCCAGCCA
Cruvulg_12    ATAAGCACATTCCCAATGTTTTAGATTTG-----TCAACACAACCTCTATCCAGCCA
Cruvulg_8      ATAAGCACATTCCCAATGTTTTAGATTTG-----TCAACACAACCTCTATCCAGCCA
Cruvulg_11    ATAAGCACATTCCCAATGTTTTAGATTTG-----TCAACACAACCTCTATCCAGCCA
Cruvulg_19    ATAAGCACATTCCCAATGTTTTAGATTTG-----TCAACACAACCTCTATCCAGCCA
Cruvulg_10    ATAAGCACATTCCCAATGTTTTAGATTTG-----TCAACACAACCTCTATCCAGCCA
Cruvulg_2      AGGAGTACATTTAAAACGTTTCAGATTTG-----GAAATGAAACATTATTCGGTCA
Cruvulg_6      AGGAGTACATTTAAAACGTTTCAGATTTG-----GAAATGAAACATTATTCGGTCA
Cruvulg_15    AGGAGTACATTTAAAACGTTTCAGATTTG-----GAAATGAAACATTATTCGGTCA
Cruvulg_16    AGGAGTACATTTAAAACGTTTCAGATTTG-----GAAATGAAACATTATTCGGTCA
Cruvulg_5      ATGACTACATTTCCAGCATTTCCGATTTTGCATCCAACAATATGAAACATTATCCGGCCA
Cruvulg_7      ATGACTACATTTCCAGCATTTCCGATTTTGCATCCAACAATATGAAACATTATCCGGCCA
Cruvulg_13    ATGACTACATTTCCAGCATTTCCGATTTTGCATCCAACAATATGAAACATTATCCGGCCA
                * * * * *
                * * * * *
    
```

[Los asteriscos indican posiciones del alineamiento en las que en todas las secuencias existe el mismo nucleótido]

### Alineamiento múltiple (archivo \*.aln): Proteínas.

CLUSTAL 2.0.9 multiple sequence alignment

```

Secuencia_1    KHPIILEKLEFPDPVISM AIEPKTKKDKQEKLSQVLNKFMK-EDPTFRATTD PETGQILIH
Secuencia_2    KHPLILEKLEFPDPVMS AIEPKKKKDKQEKLSQVLNKFMK EEDPTFAATTD PETGQILIH
Secuencia_3    PHPLILEKLEFPDPVMAM AIEPTKKKDKQEKLQV LNKFMKEEPTFA -TTDPETGQILIH
                *:*****:*****.*****.***** * * * *
                *****

Secuencia_1    GMGELHLEIMVDRMKREY GIEVNVGKPVAYKETIRKKAIGEGKFIKQTGGRGQYGHAI I
Secuencia_2    GMGELHLEIMVDRMKREY GIEVNVGKPVAYKETIRKKAIGEGKFIKQTGGRGQYGHAI I
Secuencia_3    GMGELHLEIMDDRMRKREY GIEVNVGKPVAYKETIRKKAIGEGKFIKQTGGRGQYGHAI I
                *****

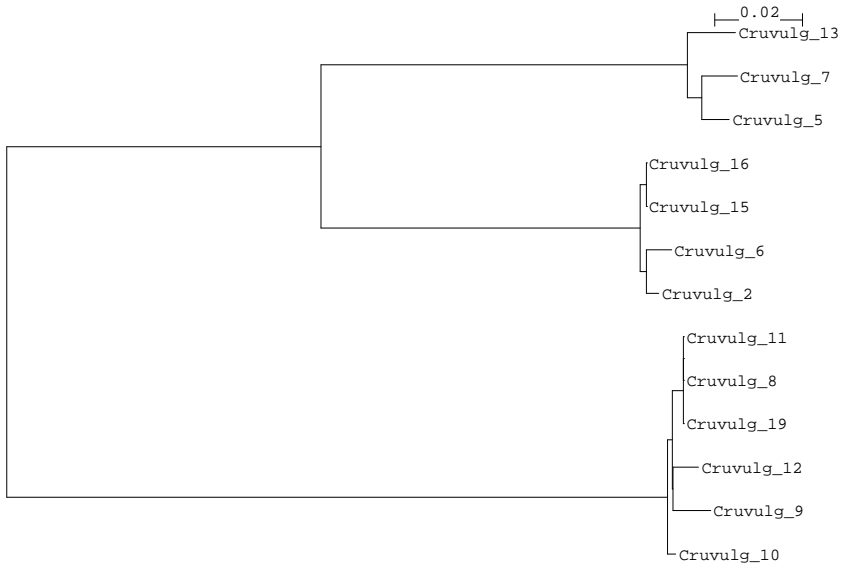
Secuencia_1    EIEPLPRGAGFEFIDDIH GGVIPKEFIPSV
Secuencia_2    EEEPLPRGAGFEFIDDIH GGVIPPEFIPSV
Secuencia_3    EEEPLPRGAGFEFIDDIH GGVIPPEFPP-V
                * * * * *
                * * *
    
```

[Los asteriscos indican posiciones del alineamiento en las que en todas las secuencias existe el mismo aminoácido; ":" indica que uno de los siguientes grupos "fuertes" de aminoácidos con propiedades químicas similares está muy conservado: STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, FYW; "." indica que uno de los siguientes grupos "débiles" está muy conservado: CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, HFY] [Estos son todos los grupos puntuados positivamente de acuerdo a la matriz Gonnet Pam25. Los grupos definidos como "fuertes" tienen una puntuación > 0.5 mientras que los definidos como "débiles" tienen una puntuación ≤ 0.5]

Dendrograma (fichero \*.dnd):

```
(
Cruvulg_8:0.00000,
(
(
(
Cruvulg_9:0.00856,
Cruvulg_12:0.00580)
:0.00026,
(
Cruvulg_10:0.00193,
(
(
(
Cruvulg_2:0.00287,
Cruvulg_6:0.00575)
:0.00147,
(
Cruvulg_15:0.00000,
Cruvulg_16:0.00000)
:0.00143)
:0.07307,
(
(
Cruvulg_5:0.00629,
Cruvulg_7:0.00812)
:0.00335,
Cruvulg_13:0.01105)
:0.08386)
:0.22318)
:0.00100)
:0.00260,
Cruvulg_19:0.00000)
:0.00000,
Cruvulg_11:0.00000);
```

Dendrograma visualizado a partir del fichero anterior con el Programa NJPlot:



### APÉNDICE 3: INSTRUCCIONES BÁSICAS SOBRE PARÁMETROS A TENER EN CUENTA PARA OBTENER EL ALINEAMIENTO MÚLTIPLE DE LAS SECUENCIAS A ANALIZAR

Dentro del menú '**ALIGNMENT**' hay una serie de opciones a tener en cuenta:

**1. 'REALIGN SELECTED SEQUENCES'**: se usa para re-alinear secuencias que, por sus características, se observa que se han alineado de manera incorrecta en el alineamiento global. Se seleccionan las secuencias a re-alinear con el ratón. Las secuencias no seleccionadas quedan "fijadas" y se crea un perfil para ellas siendo re-alineadas cada una de las secuencias seleccionadas al perfil "fijado".

**2. 'REALIGN SELECTED SEQUENCE RANGE'**: se usa para re-alinear una pequeña región del alineamiento. Se marcan con el ratón las columnas a re-alinear.

**3. 'ALIGNMENT PARAMETERS'**: aquí se pueden determinar los parámetros que deseamos aplicar a la hora de hacer el alineamiento si queremos hacer cambios sobre los parámetros utilizados por defecto. Dispone de un submenú:

**3.1. 'RESET NEW GAPS BEFORE ALIGNMENT' y 'RESET ALL GAPS BEFORE ALIGNMENT'**: para eliminar los *gaps* generados en un alineamiento o todos los *gaps* (los generados en el alineamiento más los que existían en las secuencias del fichero de entrada) en el caso de que se quiera hacer un nuevo alineamiento cambiando los parámetros.

**3.2. 'PAIRWISE ALIGNMENT PARAMETERS'**: controla la velocidad y la sensibilidad del proceso de construcción de alineamientos dos a dos (*pairwise*) iniciales.

Entre cada par de secuencias se calcula una distancia (basada en las diferencias entre ellas) y esta distancia se utiliza para construir el árbol filogenético (dendrograma) que guiará en la construcción posterior del alineamiento múltiple. Los valores de dichas distancias se calculan por separado en cada alineamiento por pares y se puede hacer de dos maneras: siguiendo el método de programación dinámica (proceso lento pero preciso) o el método de Wilbur y Lipman (proceso muy rápido pero con un resultado menos preciso). El método lento emplea poco tiempo en el alineamiento de pocas secuencias cortas pero es extremadamente lento para muchas secuencias (por ejemplo, más de 100) y largas (por ejemplo, más de 1000 nucleótidos/aminoácidos). Se puede escoger entre ambos métodos en esta opción, aunque nosotros, en esta práctica, utilizaremos el método preciso:

A. *SLOW-ACCURATE alignment parameters* (proceso lento y preciso):

Estos parámetros no tienen efecto en la velocidad del proceso. Se usan para obtener alineamientos de secuencias dos a dos iniciales que luego son re-valorados para dar puntuaciones de similitud. Estas puntuaciones se convierten luego en distancias entre las secuencias para construir el árbol que servirá de guía para el alineamiento múltiple final. Los parámetros son:

- 'Gap Open Penalty': la penalización por incluir un *gap* en el alineamiento.
- 'Gap Extension Penalty': la penalización por la longitud de un *gap*.
- 'Protein Weight Matrix': la tabla de puntuaciones que describe la similitud de un aminoácido con otro (ver más abajo).
- 'DNA Weight Matrix': la tabla de puntuaciones que se asignan a coincidencias y ausencia de coincidencias (*matches* y *mismatches*) entre dos secuencias (incluidos los símbolos del 'IUB ambiguity codes'). Ver más abajo.

B. *FAST-APPROXIMATE alignment parameters* (proceso rápido y menos exacto): Las puntuaciones de similitud se calculan a partir de alineamientos rápidos, aproximados y globales de secuencias dos a dos y que son controlados por 4 parámetros.

3.3. '*MULTIPLE ALIGNMENT PARAMETERS*': controla la localización y longitud de los *gaps* en el alineamiento múltiple final.

Cada paso hasta llegar al alineamiento final consiste en alinear secuencias dos a dos. Esto se hace de manera progresiva, siguiendo el orden de similitud establecido por el árbol guía. Los parámetros básicos que controlan este procedimiento son dos: la penalización por *gaps* y la puntuación que se da a cada posición nucleotídica de acuerdo a que exista identidad o no entre las dos secuencias que se comparan. Así, tenemos las opciones:

A. '*GAP OPENING*' y '*EXTENSION PENALTIES*': controlan el coste de la expansión cada nuevo *gap* y el coste de cada posición en un *gap*. Incrementando la penalización por cada *gap* ('*GAP OPENING*'), los *gaps* aparecerán con menor frecuencia en el alineamiento final. Incrementando la penalización por la extensión de los *gaps* ('*EXTENSION PENALTIES*'), los *gaps* serán más cortos. Los *gaps* terminales no se penalizan.

B. '*DELAY DIVERGENT SEQUENCES*': esta opción retrasa el alineamiento de las secuencias más divergentes (porcentaje de divergencia que se establece en esta opción) hasta que las más parecidas se han alineado primero.

C. '*TRANSITION WEIGHT*': esta opción da a las transiciones (cambios A↔G o C↔T) un peso entre 0 y 1. Un valor de 0 significa que las transiciones entre las dos secuencias en la posición nucleotídica evaluada se puntúan como posiciones nucleotídicas no coincidentes (*mismatch*), mientras que un peso de 1 da a las transiciones una puntuación igual a la de una coincidencia entre las dos secuencias en la posición nucleotídica evaluada (*match*). Para secuencias muy divergentes, este valor debe ponerse próximo a 0, mientras que es útil asignar un peso próximo a 1 cuando las secuencias son muy parecidas. Por defecto, el valor que aparece es 0.5.

D. '*PROTEIN WEIGHT MATRIX*': Indica las puntuaciones asignadas a cada posición aminoacídica dependiendo de que exista coincidencia (*match*) o no coincidencia (*mismatch*). Esta opción permite escoger una serie de matrices de peso para determinar la similitud de aminoácidos no idénticos. Por ejemplo, el aminoácido tirosina (Tyr) alineado con Fenilalanina (Phe) se valora "mejor" que el alineamiento de Tyr con prolina (Pro).

El programa ofrece tres series de matrices. Cada serie consiste en varias matrices que trabajan de forma diferenciada a diferentes distancias evolutivas. En resumen, el programa almacena varias matrices en su memoria que en conjunto representan todo el abanico de posibilidades de distancias aminoacídicas (desde secuencias casi idénticas a secuencias altamente divergentes). Para secuencias muy similares, es mejor usar matrices estrictas que sólo den puntuaciones altas a posiciones idénticas en todas las secuencias y a las sustituciones (cambios purina-purina o pirimidina-pirimidina) más comunes. Para la comparación de secuencias muy divergentes es mejor usar matrices más laxas en sus requerimientos que den puntuaciones altas también a sustituciones menos frecuentes. Clustal X ofrece la posibilidad de utilizar diferentes series de matrices desarrolladas por diversos autores pero recomienda por defecto el uso de la serie de matrices Gonnet Pam25.



E. 'DNA WEIGHT MATRIX': Indica las puntuaciones asignadas a cada posición nucleotídica dependiendo de que exista coincidencia (*match*) o no coincidencia (*mismatch*) (incluyendo los códigos de ambigüedad IUB). Esta opción permite escoger una matriz simple (no una serie de matrices) para determinar la similitud de nucleótidos no idénticos. Hay disponibles dos matrices:

- IUB. Esta es la matriz que se usa por defecto. Cada coincidencia (*match*) tiene una puntuación de 1.9, mientras que la ausencia de coincidencia (*mismatch*) recibe una valoración de 0. Cualquier X o N en la secuencia es tratada como una coincidencia (*match*) con cualquier símbolo de ambigüedad (de acuerdo al código de ambigüedad IUB). La falta de coincidencia entre secuencias para símbolos IUB se valora como 0.

- CLUSTALW. Cada coincidencia (*match*) tiene una puntuación de 1, mientras que la ausencia de coincidencia (*mismatch*) recibe una valoración de 0. En este caso, las posiciones con X o N, reciben una puntuación de 0.

3.4. 'PROTEIN GAP PARAMETERS': despliega una ventana temporal que permite controlar diferentes parámetros que solo se usan en el alineamiento de secuencias de proteínas.

3.5. La opción 'SECONDARY STRUCTURE PARAMETERS' sólo se puede utilizar en el Modo 'Profile Alignment Mode' y permite controlar parámetros relacionados con la estructura secundaria de las proteínas.

4. 'ITERATION': El proceso de construcción del alineamiento múltiple final puede ser iterado (repetir toda la serie de pasos un cierto número de veces) con el fin de mejorar el resultado final. La iteración se puede hacer en cada uno de los pasos que se vayan dando ('Iterate each alignment step') o tras la obtención final de un primer alineamiento múltiple ('Iterate final alignment'). En cualquiera de los dos casos, el número de iteraciones por defecto es de 3.

5. 'OUTPUT FORMAT OPTIONS': Permite elegir el formato del alineamiento en el fichero de salida (CLUSTAL, GCG, NBRF/PIR, PHYLIP, GDE, NEXUS, FASTA) de acuerdo al programa con el que se vaya a trabajar a continuación con este alineamiento. Se puede escoger más de un formato (incluso los 7 disponibles).

## APÉNDICE 4: OPCIONES DEL MENÚ 'TREES'

1. **'EXCLUDE POSITIONS WITH GAPS'**: Con esta opción, se ignora cualquier posición en la que aparece un *gap*.

2. **'CORRECT FOR MULTIPLE SUBSTITUTIONS'**: Cuando las diferencias entre las secuencias comparadas es pequeña (< 10%) el uso de esta opción no ofrece diferencias con respecto a no usarla. Pero ante grandes divergencias esta opción corrige el hecho de que las distancias observadas sean una subestimación de la distancia evolutiva real. Esto es porque, a mayor divergencia, más probable es que se haya producido más de una sustitución en más de una posición de la secuencia a lo largo de la evolución. Sin embargo, nosotros sólo detectamos una diferencia en dichas posiciones cuando comparamos las secuencias actuales.

3. **'OUTPUT FORMAT OPTIONS'**: Permite que el fichero de salida tenga diversos formatos que pueden ser leídos por diversos programas de inferencia filogenética: ClustalX, Phylip y Nexus.

En cualquier caso, ninguno de los ficheros generados permite la visualización del árbol. Para visualizarlo hay que abrir el fichero generado con alguno de los programas con los que es compatible.

Para visualizarlo nosotros, si generamos un fichero de salida con formato Clustal o Phylip, podremos utilizar el **Programa NJPLOT** (ver apéndice 2) que se distribuye junto con Clustal X y está diseñado por los Drs. Perrière y Gouy de la Universidad de Lyon (Perrière, G. and Gouy, M. (1996) WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, 78, 364-369).

# ANÁLISIS FILOGENÉTICO



# ANÁLISIS FILOGENÉTICO

## 1. 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.

## 2. 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 táxon 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 táxones (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 resolutive 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 sustituciones múltiples y reversiones de tal forma que  $p$  nos estará dando un valor subestimado del número de sustituciones nucleotídicas ocurridas realmente. Por lo tanto, se han desarrollado 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 sustitución 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 sustitución nucleotídica determinado, es decir,  $L = P(\text{datos}|\text{árbol}+\text{modelo})$ . 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 sustitución 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 ( $L = L_1 \times L_2 \times L_3 \times \dots \times L_n$ ). 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 cuál, 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  $\ln L = \ln L_1 + \ln L_2 + \ln L_3 + \dots + \ln L_n$ .

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 sustituciones 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.

### 3. METODOLOGÍA

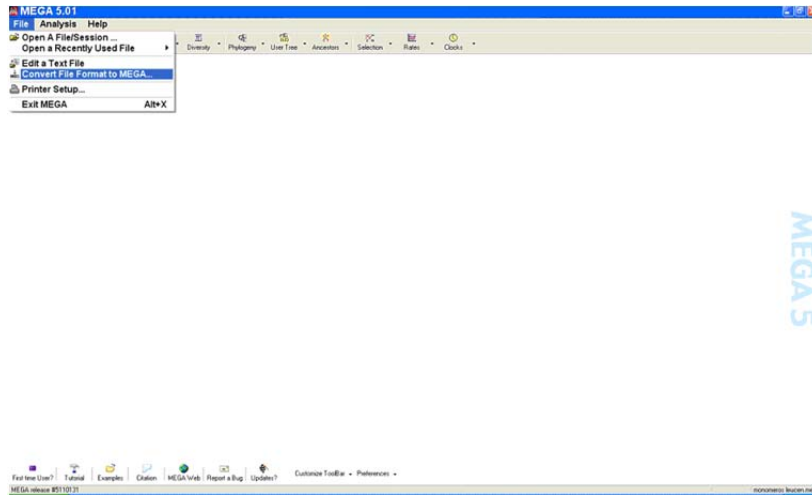
En esta práctica utilizaremos el programa MEGA (*Molecular Evolutionary Genetic Analysis*) desarrollado por el grupo de los Drs. Sudhir Kumar and Koichiro Tamura (ver referencias al final del texto). Concretamente, utilizaremos la versión 5 recientemente actualizada.

El objetivo de este programa ha sido siempre, desde su primera versión, la de proveer de diferentes herramientas para explorar y analizar secuencias de ADN y proteínas desde una perspectiva evolutiva. Ofrece una amplia gama de posibilidades en cuanto a análisis evolutivos. Nosotros, no obstante, nos centraremos sólo en el uso del programa para construir árboles filogenéticos de secuencias de ADN a partir de alineamientos múltiples obtenidos previamente con el Programa Clustal X.

Los ficheros de entrada (*input*) deben ser ficheros básicos del tipo ASCII-tex y pueden contener tanto secuencias de ADN o secuencias de proteínas como una matriz de distancias genéticas o información sobre un árbol filogenético. En esta práctica nos vamos a centrar en filogenias basadas en secuencias de ADN y las explicaciones que se darán en este guión corresponderán, por tanto, al análisis de este tipo de datos. La mayor parte de programas procesadores de textos permiten editar y guardar ficheros ASCII. Estos, normalmente, tienen una extensión del tipo \*.txt. Una vez creado un fichero de este tipo, es conveniente cambiar la extensión al tipo \*.meg de tal manera que se pueda distinguir entre ficheros para usar con MEGA y otros ficheros de texto. En el Apéndice 1 se muestra un alineamiento de secuencias en formato MEGA.

Los ficheros deben contener varias secuencias de igual longitud y, sobre todo, deben estar previamente alineadas. El programa MEGA permite alinear *de novo* las secuencias que le sean suministradas en un fichero en formato fasta (entre otros formatos). También puede usarse un fichero que contenga secuencias alineadas con otros programas. Nosotros utilizaremos los ficheros obtenidos con el programa Clustal X en la práctica realizada anteriormente. Para ello, habrá de convertirse primero los ficheros \*.aln obtenidos con dicho programa a ficheros \*.meg. Existe un convertidor en el propio programa MEGA que genera ficheros con el formato requerido por este programa: una vez abierto el programa MEGA, desplegando el Menú 'File', utilizaremos la opción 'Convert file format to MEGA'. Aparecerá una ventana emergente que nos pedirá que indiquemos el nombre del fichero a convertir y su extensión (\*.aln en este caso). Una vez obtenida la conversión a formato MEGA,

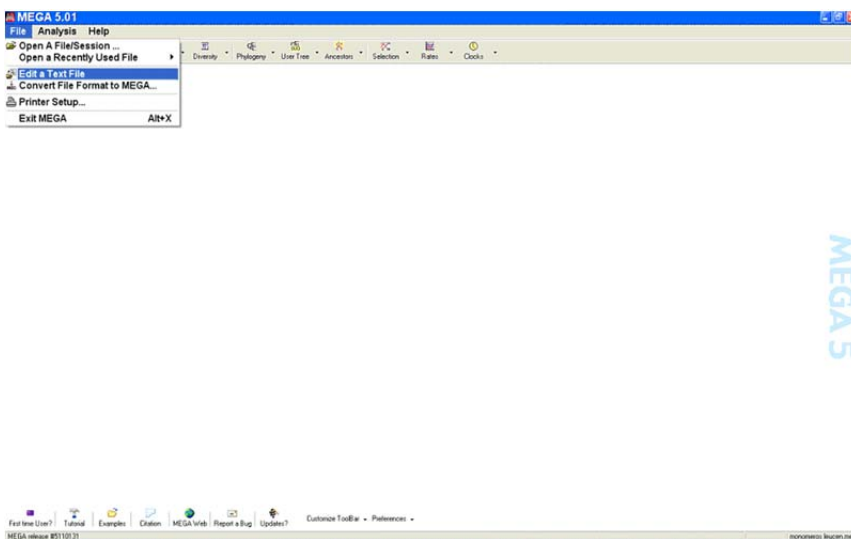
guardaremos el fichero resultante con extensión .meg y lo utilizaremos después para nuestro análisis filogenético.



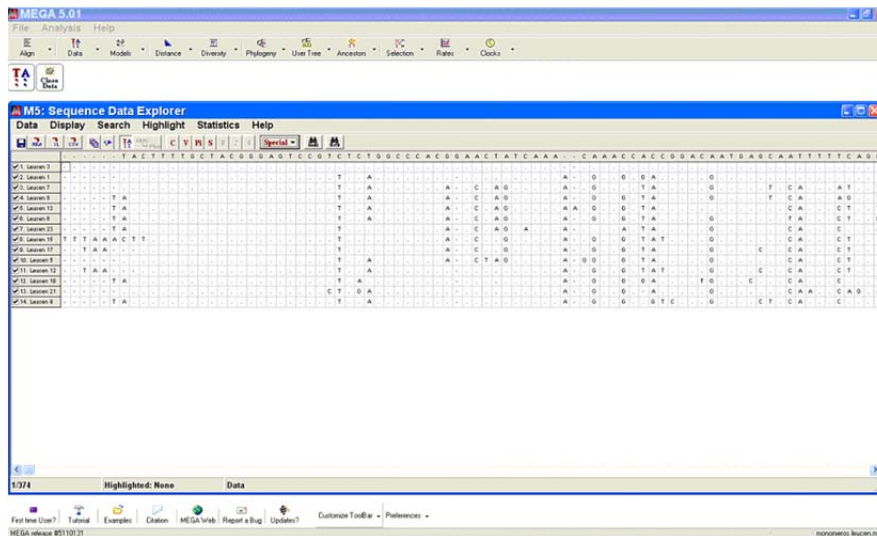
Tanto las secuencias de nucleótidos como las de aminoácidos han de seguir las reglas de código IUPAC. Además, se pueden utilizar los siguientes símbolos especiales:

- Espacios en blanco ( ): son ignorados por MEGA.
- Punto (.): representa identidad en esa posición de la secuencia con el nucleótido/aminoácido de la primera secuencia.
- Interrogación (?): dato desconocido (*missing data*)
- Guión (-): *gap*.

El programa MEGA dispone de un editor de texto (*Text File Editor* invocado con la opción *Edit a Text File* dentro del menú *File*) que es muy útil para crear y editar ficheros ASCII. Se invoca su aparición tanto a requerimiento nuestro como de manera automática por parte del programa en cuanto en el fichero de entrada (*input*) se detectan errores de formato. Una vez hechos los cambios estos se pueden guardar permanentemente.



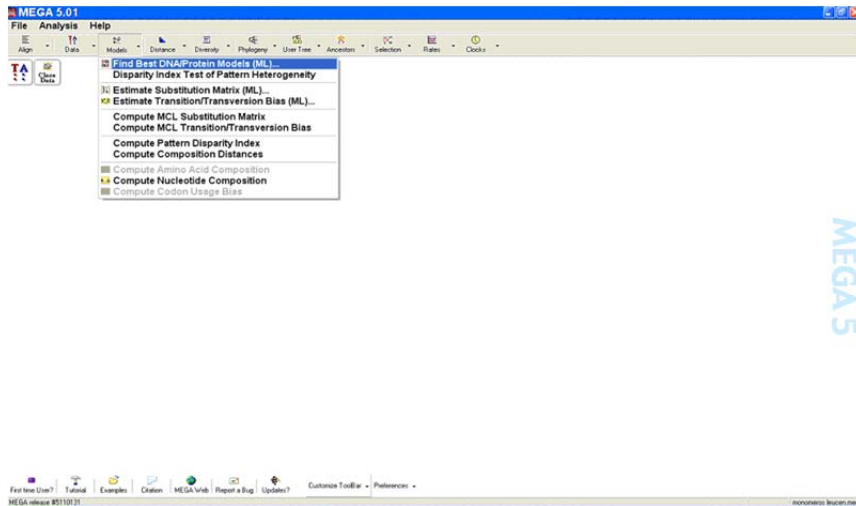
Una vez abierto con el programa MEGA un fichero con secuencias de ADN alineadas en formato MEGA, lo primero que aparecerá en la pantalla será el alineamiento dispuesto en una cuadrícula (*Sequence Data Explorer*):



Este explorador ofrece diversas funciones como por ejemplo: a) la posibilidad de seleccionar diferentes subconjuntos de datos o de secuencias (*Setup/Select Genes and Domains* y *Setup/Select Taxa and Groups*); b) diferentes tipos de análisis estadísticos como la composición nucleotídica; c) identificación de posiciones nucleotídicas según que estén conservadas, sean variables o sean filogenéticamente informativas. En esta práctica no utilizaremos las opciones que permiten definir regiones en las secuencias o seleccionar grupos de secuencias. Sin embargo, sí que podremos visionar las secuencias y determinar algunos datos interesantes como el número de posiciones nucleotídicas variables o la composición nucleotídica de dichas secuencias.

Una vez analizadas visualmente las secuencias, procederemos a hacer los análisis filogenéticos. A continuación, se detallan los pasos a seguir para hacer una reconstrucción filogenética de las secuencias alineadas mediante diferentes métodos de inferencia:

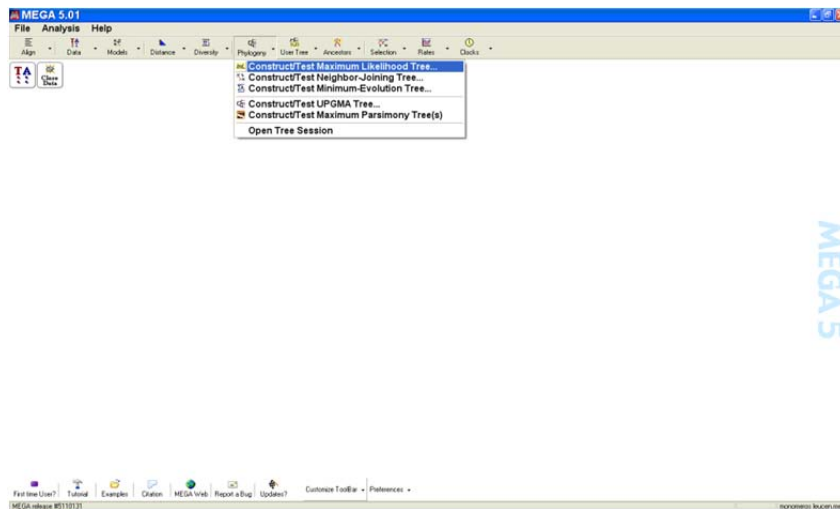
**Paso 1.** Desplegar el menú *Models* y seleccionar la opción *Find Best DNA/Protein Models*. Este análisis inicial es muy útil para conocer el modelo evolutivo al que se ajustan los cambios producidos en las secuencias a analizar. El resultado del análisis es un listado con los diferentes valores obtenidos mediante diferentes criterios (BIC o criterio de inferencia bayesiano, AICc o Criterio de Akaike,  $lnL$  o criterio de máxima verosimilitud). Por lo general, utilizaremos el criterio bayesiano. El modelo que presente la puntuación más baja según este criterio se considera que describe mejor el patrón de sustitución de las secuencias analizadas.



MEGA 5

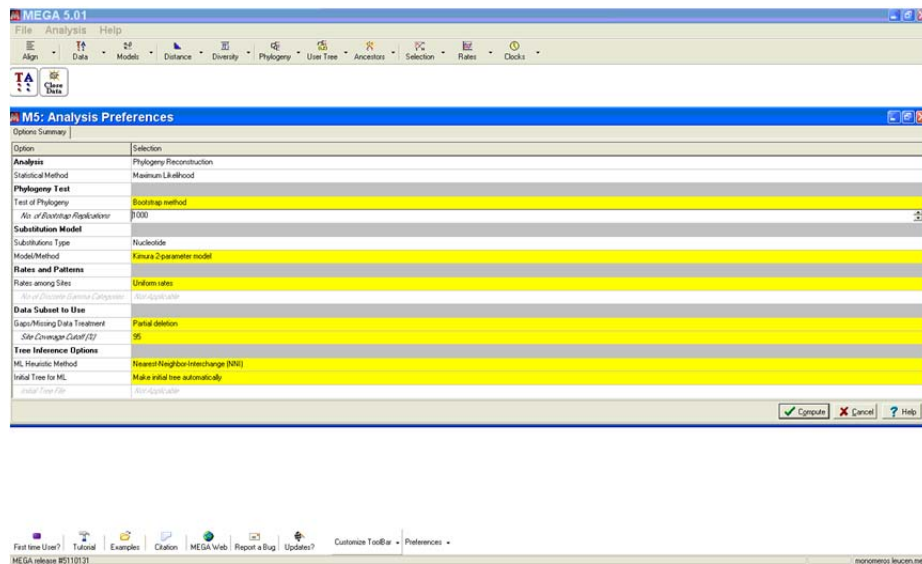
**Paso 2.** A continuación desplegaremos el Menú *Phylogeny* con el fin de realizar el análisis filogenético. La reconstrucción filogenética se va a realizar con tres de los métodos propuestos por el programa: Máxima parsimonia, Máxima verosimilitud y Neighbor-joining (uno de los tres métodos de los que se dispone en el programa para calcular árboles basados en distancias genéticas).

A cualquiera de los métodos se accede desde el menú *Phylogeny*:



MEGA 5

Realizada la elección correspondiente, aparecerá una ventana emergente en la que incluir los parámetros adecuados:



En la opción 'Test of phylogeny' escogeremos la opción 'Bootstrap method' que nos permite evaluar la fiabilidad de las topologías obtenidas mediante el método *bootstrap* de remuestreo para lo que hay que indicar en la ventana emergente que aparece el número de réplicas a realizar en el análisis (1000 puede ser suficiente).

Seguidamente, en la misma ventana, seleccionaremos los parámetros a tener en cuenta según que método estemos utilizando para construir el árbol filogenético:

**2. A. Métodos basados en distancias genéticas.** De los tres disponibles, utilizaremos el método *Neighbor-joining* (N-J). Esta opción requiere el cálculo previo de las distancias genéticas entre las secuencias alineadas, cálculo que está basado en el número de diferencias entre cada par de secuencias. Para ello, en la misma ventana emergente en la que hemos seleccionado el test estadístico a utilizar, seleccionaremos el modelo de sustitución ('*Substitution model*') para calcular las distancias genéticas, para lo que tendremos en cuenta el resultado obtenido en el paso 1.

Si el análisis del paso 1 determinó que los patrones de sustitución no son homogéneos entre diferentes posiciones nucleotídicas de nuestro alineamiento, se seleccionara en el desplegable '*Rates among Sites*' la opción '*Gamma distributed*' y se indicará un valor para el parámetro *gamma* que se puede obtener en el menú *Rates*. Si el análisis del paso 1 determinó que los patrones de sustitución no son homogéneos entre linajes diferentes de la filogenia, en el desplegable '*Pattern among lineages*' se indicará que el patrón es heterogéneo.

La opción '*Data subset to use*' permite manejar los *gaps* y datos desconocidos (*missing data*), incluir o excluir codones, y restringir el análisis a posiciones nucleotídicas marcadas:

### *Gaps and Missing Data*

Se puede escoger eliminar todas las posiciones nucleotídicas en las que en alguna secuencia existe un *gap* o datos desconocidos (*missing data*) antes de iniciar los cálculos utilizando la opción '*Complete-deletion*'. Alternativamente, se puede optar por retener todas esas posiciones inicialmente y excluirlas sólo en las comparaciones entre secuencias dos a dos durante el proceso de cálculo de distancias (opción '*Pairwise-deletion*').

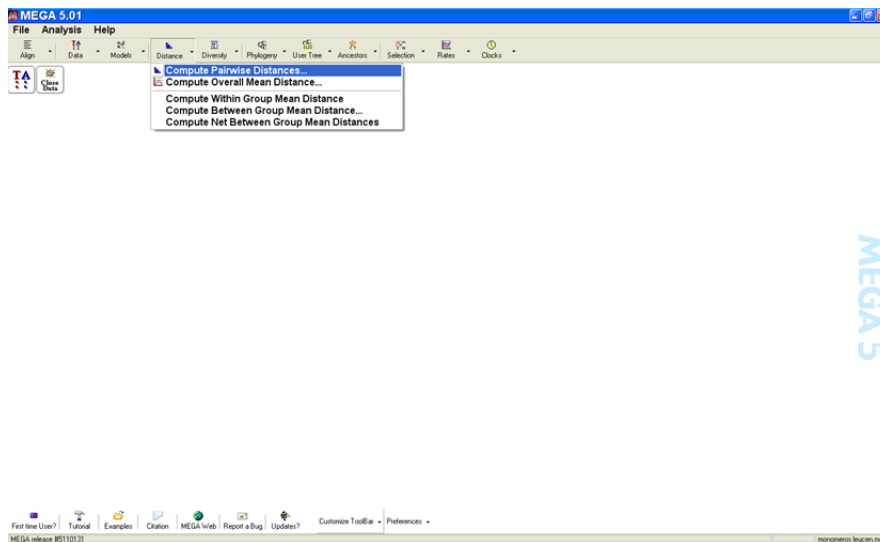
*Include Sites: Codon Positions*

Se puede seleccionar cualquier combinación de codones y de posiciones nucleotídicas no codificadoras con el fin de tratarlas diferencialmente en el análisis.

*Include Sites: Labeled Sites*

Esta opción está disponible sólo si hay marcas asociadas con algunas de las posiciones nucleotídicas del alineamiento. Se puede optar por incluir en el análisis sólo esas posiciones nucleotídicas marcadas.

Si queremos conocer los valores de distancias genéticas calculados entre cada par de secuencias podemos visualizar una tabla de distancias mediante la opción '*Compute pairwise distance*' del menú '*Distance*':



Habrá que introducir el método requerido para el cálculo de distancias y las diferentes opciones referidas anteriormente con respecto al tratamiento de las posiciones nucleotídicas.

Como se puede ver, este menú ofrece muchas otras posibilidades, que no exploraremos, como el cálculo del valor medio de distancia entre todas las secuencias o las distancias medias dentro de grupos particulares de secuencias del alineamiento así como entre grupos de secuencias.

**2. B. Método de máxima parsimonia (MP).** Para construir un árbol MP, sólo se usan las posiciones de las secuencias en las que hay al menos dos clases diferentes de nucleótidos, cada uno representado al menos dos veces (sitios informativos). El resto de sitios variables no se usan en MP aunque sí se usan en métodos de distancias o de ML.

Para la búsqueda del mejor árbol, MEGA ofrece tres tipos diferentes de búsquedas: *max-mini branch-and-bound search*, *min-mini heuristic search*, y *close-neighbor-*

*interchange heuristic search*. Sólo el primero de ellos garantiza la posibilidad de encontrar todos los árboles MP posibles pero el tiempo que se requiere para ello es enorme, sobre todo si en el alineamiento hay más de 15 secuencias. Así que utilizaremos alguno de los métodos heurísticos.

El método MP puede generar diferentes árboles igualmente parsimoniosos. Seleccionando la opción '*Consensus tree*' obtendremos un árbol "consenso" entre todos los árboles posibles. Contamos con dos tipos de árboles consenso: 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.

La opción '*Tree Inference Options*' nos permite definir que método de MP se usará para obtener los árboles más parsimoniosos. Nosotros probaremos el método heurístico (*Close-neighbor interchange*).

La opción '*Data subset to use*' nos ofrece las mismas posibilidades que en el caso anterior.

**2. C. Método de máxima verosimilitud (ML).** Se dispone de las mismas opciones vistas hasta ahora: Test estadístico a aplicar, Modelo de sustitución o Tratamiento que se le dará a los sitios desconocidos y a los *gaps*. A la hora de elegir el modelo evolutivo aplicable a la evolución de las secuencias analizadas tendremos en cuenta el resultado obtenido en el paso 1.

Finalmente se nos da a elegir entre dos métodos heurísticos de inferencia ('*Tree Inference Options*'): *Close-Neighbor-Interchange* (CNI) y *Nearest-Neighbor-Interchange* (NNI). Usaremos este segundo método.

**Paso 3.** Cuando varios nodos internos de un árbol filogenético tienen significación estadística baja, es útil a menudo producir un árbol multifurcado asumiendo que todas las ramas que parten de dichos nodos tienen una longitud igual a 0. Este tipo de árbol se denomina árbol condensado (*Condensed tree*). En MEGA, este tipo de árboles se pueden obtener para cualquier valor de *bootstrap*. Debemos incluir un porcentaje de *bootstrap* para el cual todas las ramas de nodos con soporte menor que dicho porcentaje tendrán reducida su longitud a 0. Dado que las ramas poco significativas son eliminadas, este tipo de árboles enfatizan el resto de ramificaciones. En este tipo de árboles, no obstante, debemos considerar sólo la topología y no prestar excesiva atención a la longitud de las ramas ya que todas quedan modificadas al reducir a 0 la longitud de las poco significativas.

Este tipo de árboles pueden parecer similares a los árboles consenso obtenidos por MP, pero son diferentes. Un árbol consenso es el resultado de escoger un árbol representativo del total de todos los árboles parsimoniosos obtenidos, mientras que un árbol condensado es simplemente una versión simplificada de un árbol.

La opción de producir árboles condensados está disponible para cualquiera de los métodos de inferencia filogenética que vamos a utilizar.

**Paso 4.** Tras la aplicación de cada uno de los tres métodos probados, guardaremos los árboles obtenidos mediante la opción '*Save current session*' del menú *File* que



aparece sobre la reconstrucción filogenética. Este menú también nos permite imprimir el árbol. A partir del menú *Image* se puede seleccionar guardar el árbol como fichero \*.tiff o como fichero \*.pdf.

## 4. CUESTIONES

1. Utilizando los alineamientos de secuencias de ADN obtenidos en la práctica anterior con el programa ClustalX obtener árboles filogenéticos mediante N-J, MP y ML.

## REFERENCIAS

Tamura K, Dudley J, Nei M & Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24: 1596-1599.

Kumar S, Dudley J, Nei M & Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics* 9: 299-306.

**Las diferentes versiones de MEGA, así como un completo manual tanto en formato pdf como on-line se pueden conseguir en la siguiente dirección web:**

<http://www.megasoftware.net/>

**APÉNDICE 1: FORMATO MEGA**

Ejemplo para secuencias de ADN:

```
#MEGA
Title: denominación del fichero (tipo de secuencias, grupo biológico, etc...)

#Cruvulg_9      TTACTTTTGGCTACGGGAGTCCGTTCTTCGCCCACGAACTATCAAAACGAAGCTACGGACC
#Cruvulg_12     T-AC TTTTGGCTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
#Cruvulg_8      TAACTTTTGGCTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
#Cruvulg_11     TAACTTTTGGCTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
#Cruvulg_19     TAACTTTTGGCTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
#Cruvulg_10     T-AC TTTTGGCTACGGGAGTCCGTTCCAGGCCACGAACTATCAAAACGAAGCTACGGACC
#Cruvulg_2      TAACTTTTGGCTACGGGAGTCCGTTTCAGTCCCACGAACTATCAAAACGAAGCTATAGACG
#Cruvulg_6      TAACTTTTGGCTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
#Cruvulg_15     T-AC TTTTGGCTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
#Cruvulg_16     T-AC TTTTGGCTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
#Cruvulg_5      T-AC TTTTGGCTACGGGAGTCCGTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
#Cruvulg_7      T-AC TTTTGGCTACGGGAGTCCGTTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG
#Cruvulg_13     TAACTTTTGGCTACGGGAGTCCGTTTTCAGGCCACGAACTATCAAAACGAAGCTATAGACG

#Cruvulg_9      ATAAGCACATTCCCAATGTTTTAGATTTTG-----TCAACACAACCTCTATCCAGCCA
#Cruvulg_12     ATAAGCACATTCCCAATGTTTTAGATTTTG-----TCAACACAACCTCTATCCAGCCA
#Cruvulg_8      ATAAGCACATTCCCAATGTTTTAGATTTTG-----TCAACACAACCTCTATCCAGCCA
#Cruvulg_11     ATAAGCACATTCCCAATGTTTTAGATTTTG-----TCAACACAACCTCTATCCAGCCA
#Cruvulg_19     ATAAGCACATTCCCAATGTTTTAGATTTTG-----TCAACACAACCTCTATCCAGCCA
#Cruvulg_10     ATAAGCACATTCCCAATGTTTTAGATTTTG-----TCAACACAACCTCTATCCAGCCA
#Cruvulg_2      AGGAGTACATTTAAAACGTTTCAGATTTTG-----GAAATGAAACATTATTCGGTCA
#Cruvulg_6      AGGAGTACATTTAAAACGTTTCAGATTTTG-----GAAATGAAACATTATTCGGTCA
#Cruvulg_15     AGGAGTACATTTAAAACGTTTCAGATTTTG-----GAAATGAAACATTATTCGGTCA
#Cruvulg_16     AGGAGTACATTTAAAACGTTTCAGATTTTG-----GAAATGAAACATTATTCGGTCA
#Cruvulg_5      ATGACTACATTTCCAGCATTTTCGGATTTTGCATCCAACAATATGAAACATTATCCGGCCA
#Cruvulg_7      ATGACTACATTTCCAGCATTTTCGGATTTTGCATCCAACAATATGAAACATTATCCGGCCA
#Cruvulg_13     ATGACTACATTTCCAGCATTTTCGGATTTTGCATCCAACAATATGAAACATTATCCGGCCA

#Cruvulg_9      TGTTTCGATACATGACACGACATATGTTTCGTTTCGGCTCGGGGAATGGGTGGTTTCGGCTG
#Cruvulg_12     TGTTTCGATACATGACACGACATATGTTTCGTTTCGGCTCGGGGAATGGGTGGTTTCGGCTG
#Cruvulg_8      TGTTTCGATACATGACACGACATATGTTTCGTTTCGGCTCGGGGAATGGGTGGTTTCGGCTG
#Cruvulg_11     TGTTTCGATACATGACACGACATATGTTTCGTTTCGGCTCGGGGAATGGGTGGTTTCGGCTG
#Cruvulg_19     TGTTTCGATACATGACACGACATATGTTTCGTTTCGGCTCGGGGAATGGGTGGTTTCGGCTG
#Cruvulg_10     TGTTTCGATACATGACACGACATATGTTTCGTTTCGGCTCGGGGAATGGGTGGTTTCGGCTG
#Cruvulg_2      CAATCCGGTAAATTGCACGACGTGCCTTC-TTCGTCTTTTCGAATGGATGGCTGCAGTAG
#Cruvulg_6      CAATCCGGTAAATTGCACGACGTGCCTTC-TTCGTCTTTTCGAATGGATGGCTGCAGTAG
#Cruvulg_15     CAATCCGGTAAATTGCACGACGTGCCTTC-TTCGTCTTTTCGAATGGATGGCTGCAGTAG
#Cruvulg_16     CAATCCGGTAAATTGCACGACGTGCCTTC-TTCGTCTTTTCGAATGGATGGCTGCAGTAG
#Cruvulg_5      CAATCCGGTAAATTGCATCGACGCGC-----GCGCGTTTGAATGGATGGATGCAGTCCG
#Cruvulg_7      CAATCCGGTAAATTGCATCGACGCGC-----GCGCGTTTGAATGGATGGATGCAGTCCG
#Cruvulg_13     CAATCCGGTAAATTGCATCGACGCGC-----GCGCGTTTGAATGGATGGATGCAGTCCG
```

ANÁLISIS COMPUTACIONAL DE  
DATOS DE EXPRESIÓN GÉNICA  
DIFERENCIAL OBTENIDOS  
MEDIANTE CHIPS DE ADN



# ANÁLISIS COMPUTACIONAL DE DATOS DE EXPRESIÓN GÉNICA DIFERENCIAL OBTENIDOS MEDIANTE CHIPS DE ADN

## 1. OBJETIVOS

La práctica se centra en el análisis de datos de expresión génica obtenidos mediante la técnica de chip de ADN. La práctica se llevará a cabo mediante el programa GEPAS/BABELOMICS. Este programa funciona como servidor-web y se ubica en el Centro de Investigación Príncipe Felipe de Valencia. Para su uso únicamente se necesita un navegador. Este servidor-web implementa todos los métodos necesarios para el análisis permitiendo además el almacenamiento de los datos de los usuarios. Mediante el desarrollo de esta práctica se pretenden alcanzar los siguientes objetivos:

- Entender los principios básicos en los que se basa la tecnología de chip de ADN
- Conocer las fuentes de artefactos técnicos y obtener nociones básicas del preprocesamiento de datos (filtrar muestras que no tienen una calidad suficiente, normalización, sustracción del ruido)
- Conocer los formatos de ficheros con los que trabajamos en esta práctica (datos crudos en CEL & matriz de expresión)
- Saber lo que es un log<sub>2</sub>-ratio y un valor-P y usar estas medidas para obtener una lista de genes expresados diferencialmente
- Obtener nociones básicas de los métodos estadísticos que se emplean incluyendo la corrección para controlar la acumulación de errores de tipo II
- Saber interpretar un heat map

## 2. FUNDAMENTO TEÓRICO

### 2.1. Introducción

Los chips de ADN permiten medir simultáneamente miles de propiedades genómicas como la expresión génica o la existencia de polimorfismos (tanto SNPs como CNV – variaciones en el número de copias). La exactitud y reproducibilidad de esta tecnología han sido altamente probadas durante la última década. Existen artefactos que pueden enmascarar la señal biológica o generar una señal falsa. Estos artefactos incluyen la preparación de la muestra, el proceso de hibridación, sesgo de fluorocromos, hibridación no-específica o diferencias entre los escáneres. Antes de analizar los datos, estos artefactos tienen que ser analizados y corregidos (si es posible). Este paso se llama preproceso e incluye la inspección de los datos crudos, la filtración de muestras con calidad insuficiente, la sustracción del ruido de fondo y la normalización. El resultado de un experimento para detectar expresión diferencial es siempre una lista de genes que caracteriza el experimento.

## 2.2. Expresión diferencial

La expresión diferencial es el cambio de los niveles de expresión de uno o más genes entre dos o varias condiciones. Algunos análisis típicos donde se emplean estas técnicas son:

- Detectar posibles causas de una enfermedad (muestras sanas frente a muestras patológicas o muestras que corresponden a distintas fases de la enfermedad) Caracterizar las diferencias de expresión entre tipos celulares (hígado frente a cerebro, etc.)
- Determinar los genes implicados en el desarrollo (no-diferenciado frente a diferenciado o las diferentes fases de diferenciación)
- Medir los efectos de estímulos externos: medicamentos, luz, alimentación, sueño, etc.
- Comparar células u organismos mutantes frente al tipo común.

Por lo tanto, muchas veces se comparan dos condiciones entre sí. Es común referirse a una de estas condiciones como "casos" (enfermos, con tratamiento, mutantes, etc.) y a la otra como "controles".

## 2.3. Chip de ADN

El principio en que se basa la técnica de chip de ADN es la hibridación entre dos hebras complementarias (Figura 1). Para hacer uso de la hibridación se fijan miles de oligonucleótidos (sondas) en una superficie sólida. Las sondas pueden tener longitudes muy diferentes de un fabricante a otro (60 nt en Agilent y 25 nt en Affymetrix). De estos oligonucleótidos se conoce tanto la posición en la superficie del chip como el gen al que representan. Las muestras analizadas suelen ser ADNc y ARNc etiquetados con fluorocromos. Después de la hibridación entre la muestra y las sondas se eliminan (mediante lavado) las hibridaciones no-específicas (hibridación cruzada). Como último paso se excitan los fluorocromos con un láser leyendo después con un escáner la fluorescencia de cada sonda, siendo la intensidad proporcional al nivel de expresión del gen. La Figura 2 muestra los pasos que engloba un experimento de chip de ADN, desde la extracción del ARN hasta el análisis bioinformático.

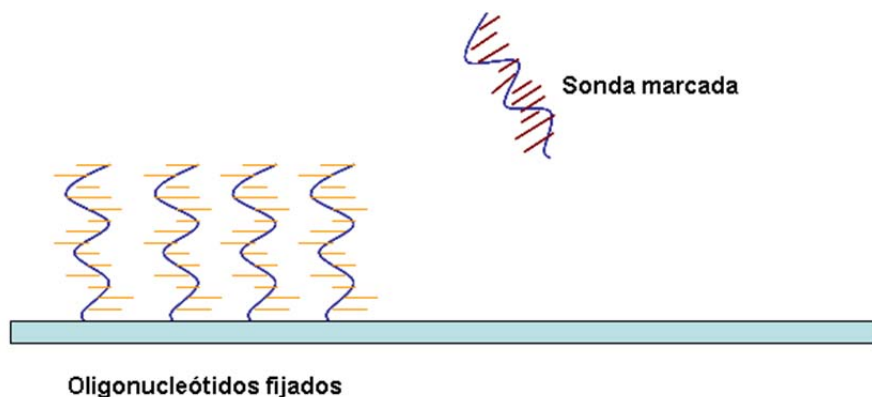


Figura 1: El principio básico de la hibridación entre las sondas y las muestras marcadas con fluorocromos.

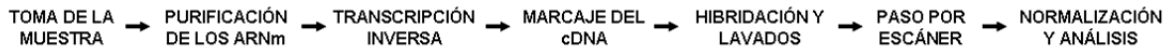


Figura 2: Los diferentes pasos en un análisis de chip de ADN.

### 2.4. Formatos

Cada fabricante maneja formatos diferentes para analizar los resultados crudos del experimento. En el caso de Affymetrix, los ficheros que contienen los datos crudos tienen extensión CEL. Se pueden obtener más datos acerca del formato en la siguiente dirección:

<http://www.affymetrix.com/support/developer/power/tools/changelog/gcos-agcc/cel.html>

Entre otros valores, este fichero contiene las intensidades y las desviaciones típicas de todas las sondas del chip. Por lo tanto, el formato CEL será el punto de partida de nuestro análisis.

Otro fichero importante es el de la matriz de expresión (Figura 3). Este fichero reúne los valores de expresión de las distintas muestras para cada sonda y se genera como resultado de la normalización.

#NUMBER_FEATURES	56675																		
#NUMBER_SAMPLES	10																		
#VARIABLE	knead-down	CATEGORICAL{1,562,CREB}																	
#VALUES	GSM304493.CEL	GSM304494.CEL	GSM304495.CEL	GSM304496.CEL	GSM304497.CEL	GSM304498.CEL	GSM304499.CEL	GSM304500.CEL	GSM304501.CEL	GSM304502.CEL	GSM304503.CEL	GSM304504.CEL	GSM304505.CEL	GSM304506.CEL	GSM304507.CEL	GSM304508.CEL	GSM304509.CEL	GSM304510.CEL	GSM304511.CEL
DESCRIPTION{}																			
AFFX-BI06-5_at	7.25625	7.40737	7.23749	7.40637	7.34597	7.34592	7.47261	7.06995	7.27691	7.46973									
AFFX-BI06-M_at	7.52081	7.00379	7.51106	7.07074	7.6424	7.70306	7.96048	7.49351	7.591	7.97457									
AFFX-BI06-3_at	7.25474	7.42093	7.15845	7.56731	7.21631	7.38323	7.56039	7.09309	7.20534	7.50469									
AFFX-BI06-5_at	0.60199	0.00627	0.62663	0.06627	0.72925	0.75524	0.05612	0.39207	0.63667	0.99369									
AFFX-BI06-3_at	3.14017	3.18143	0.99186	3.24093	0.99073	3.12330	3.31531	0.70477	3.01035	3.37066									
AFFX-BI06n-5_at	10.23937	10.36695	10.10372	10.41274	10.25119	10.26293	10.50732	9.7635	10.26777	10.49168									
AFFX-BI06n-3_at	11.49244	11.58818	11.39346	11.56569	11.35699	11.49299	11.71195	11.27256	11.47972	11.7597									
AFFX-CREB-5_at	12.71793	12.79188	12.68792	12.77999	12.76011	12.76249	12.89519	12.58286	12.79189	12.80281									
AFFX-CREB-3_at	13.14953	13.26549	13.06552	13.19299	13.17099	13.20926	13.30655	13.05695	13.18078	13.24299									
AFFX-DAPK-5_at	3.42588	3.4918	3.37627	3.54084	3.39399	3.26988	3.44672	3.49425	3.47675	3.47596									
AFFX-DAPK-M_at	3.50779	3.63947	3.56408	3.62674	3.69741	3.57905	3.53099	3.50782	3.67961	3.66767									
AFFX-DAPK-3_at	3.67936	3.67034	3.60089	3.60637	3.67202	3.78939	3.76886	3.6996	3.65919	3.67997									
AFFX-LYPK-5_at	3.36709	3.43961	3.3686	3.46966	3.40145	3.38979	3.32958	3.41897	3.44792	3.38999									
AFFX-LYPK-M_at	3.75270	3.67460	3.90616	3.01620	3.75171	3.01194	3.76701	3.00145	3.03940	3.05637									
AFFX-LYPK-3_at	3.40567	3.39325	3.58926	3.47032	3.43001	3.39199	3.3088	3.45031	3.37179	3.56331									
AFFX-PH01-5_at	3.63987	3.61111	3.55439	3.57925	3.47064	3.593	3.48084	3.52134	3.47623	3.46661									
AFFX-PH01-M_at	3.50714	3.62792	3.50831	3.75011	3.65634	3.5892	3.67612	3.58979	3.5493	3.59596									
AFFX-PH01-3_at	5.71979	5.65996	5.75195	5.64905	5.64599	5.46134	5.67611	5.53426	5.53995	6.09561									
AFFX-TH01-5_at	4.2617	4.28934	4.20208	4.36983	4.31919	4.0295	4.27004	4.15127	4.1895	4.33903									
AFFX-TH01-M_at	3.74652	3.79672	3.74989	3.79753	3.84462	3.82842	3.62697	3.7289	3.63958	3.70985									
AFFX-TH01-3_at	4.21699	4.09616	4.25117	4.29149	4.09742	4.09956	4.17697	4.04944	4.04488	4.29755									
AFFX-TYR01-5_at	3.66999	3.84016	3.69948	4.16542	3.67407	3.79769	3.70991	3.85486	3.79588	3.89664									
AFFX-TYR01-M_at	3.75832	3.6989	3.65695	3.74982	3.81179	3.55191	3.79715	3.66367	3.61719	3.83683									
AFFX-TYR01-3_at	3.51530	3.42599	3.38037	3.50065	3.51170	3.38417	3.40667	3.45999	3.3512	3.57526									
AFFX-r2-E0-BI06-5_at	7.60247	7.01990	7.61547	7.96556	7.64080	7.02321	0.04090	7.03904	7.62777	0.15014									

Figura 3: Se muestra la matriz de expresión. El fichero contiene 11 columnas. La primera indica la sonda y las 10 siguientes los valores de expresión para cada una de las 10 muestras.

### 2.5. Preprocesamiento de los datos

El preprocesamiento es necesario para detectar artefactos técnicos y corregirlos en la manera de lo posible. Estos artefactos incluyen diferencias en la preparación de las muestras, diferencias en el proceso de hibridación, hibridación no-específica, sesgo de fluorocromos (más importante en los chips de dos canales) y diferencias entre los escáneres. Para los datos del tipo de chip con que trabajamos aquí (chips de un canal) podemos distinguir 2 pasos importantes en el preprocesamiento:

**Sustracción del ruido de fondo:** Una fuente de ruido de fondo proviene de la hibridación no-específica, es decir de dos secuencias no completamente complementarias que sin embargo pueden formar un híbrido (Figura 1). Para estimar el impacto de este efecto, Affymetrix incluye para cada sonda (PM - perfect match) otra con una base, la decimotercera, cambiada (MM - mismatch). La intensidad de la señal en las sondas MM permitirá estimar el impacto de la hibridación no específica.

**Normalización entre muestras:** Cada muestra esta caracterizada por la distribución de las intensidades de las sondas. Frecuentemente se observa que las medias de estas distribuciones difieren entre las diferentes muestras. Es obvio que este efecto puede llevar a una expresión diferencial artificial ya que las diferencias son debidas a artefactos técnicos y no de origen biológico. El objetivo de este paso es por lo tanto ajustar tanto las medias como la forma de la distribución entre todas las muestras en el análisis (Figuras 9 y 11 que muestran las distribuciones antes y después de la normalización).

Como se ha mencionado antes, Babelomics guarda los resultados del preprocesamiento en una matriz de expresión:

[http://bioinfo.cipf.es/babelomicstutorial/babelomics\\_expression\\_data](http://bioinfo.cipf.es/babelomicstutorial/babelomics_expression_data)

Esta matriz es el fichero con el que se lleva a cabo la detección de genes que se expresan diferencialmente empleando métodos estadísticos.

## 2.6. Estadística básica

Para poder interpretar el resultado de un análisis de expresión diferencial, necesitamos tener algunos conocimientos básicos de las medidas que se emplean para detectarla.

### 2.6.1. $\log_2$ - ratio

Nos permite cuantificar "cuando más se expresa un gen en una condición comparada con otra". Se calcula simplemente como:

$$ratio = \log_2\left(\frac{I_{grupo1}^i}{I_{grupo2}^i}\right)$$

Es decir cómo relación de la intensidad de la señal de la sonda  $i$  entre el grupo 1 y el grupo 2.

El resultado es fácil de interpretar:

- ratio = 0 : el gen se expresa igual en las dos condiciones
- ratio > 0 : el gen se expresa más en el grupo1
- ratio < 0 : el gen se expresa más en el grupo2
- ratio = 1 : el gen se expresa el doble en el grupo1 que en el grupo2
- ratio = -1 : el gen se expresa el doble en el grupo2 que en el grupo1



### 3.6.2. Significación estadística

La ratio definida en la sección anterior define la diferencia de expresión génica entre dos condiciones. Es decir, podemos interpretar ésta medida como la fuerza de la señal. En el próximo paso tenemos que buscar aquellas ratio o diferencias que son estadísticamente significativos. Es decir, aquellas que no se deben al azar. Para ello se pueden emplear muchos tests estadísticos diferentes, algunos de ellos desarrollados específicamente para los chips de ADN. Aquí queremos mencionar solo uno de los más usados en todos los campos de la biológica, la prueba t de Student (t-test). En estos test, el estadístico utilizado tiene una distribución t de Student si la hipótesis nula es cierta. Por lo general se asume que las distribuciones son gaussianas y que las varianzas son del mismo orden. Se calcula en primer lugar el estadístico t como la diferencia entre las medias dividida por la desviación standard del conjunto de datos. El estadístico t se puede convertir directamente en una significación estadística mediante una tabla.

### 3.6.3. Corrección por acumulación de errores de tipo II

Se produce un error de tipo II cuando se rechaza erróneamente una hipótesis nula correcta. En el caso de la expresión génica, la hipótesis nula es que no hay expresión diferencial, es decir que las medias son iguales en controles y casos. Se suele fijar una significación estadística de antemano (por ejemplo 0.05 ó 0.01). Si el valor P de un test es menor se rechaza la hipótesis nula. Si fijamos el nivel en 0.05, eso significa que tenemos una probabilidad de hasta 0.05 de rechazarla erróneamente. Si comprobamos no solo una hipótesis nula sino varias simultáneamente, como es nuestro caso (suele haber decenas de miles de sondas en un chip), se acumulan los errores de tipo II. Para evitarlo y mantener la significación global para el experimento tenemos que corregir los valores P. Existen muchas maneras para lograrlo. La más simple y más conservadora (eleva el número de errores de tipo II) es la corrección de Bonferoni. El valor P corregido se calcula simplemente como el valor P original multiplicado por el número de tests. Otros tests más sofisticados y mejor adaptados al problema de expresión diferencial son el de Benjamini y Hochberg, que se van a emplear en esta práctica. Este test es especialmente indicado cuando se rechazan pocas hipótesis como es el caso en los experimentos de expresión diferencial (suele haber pocos genes que se expresan diferencialmente comparado con el número total de genes en el análisis).

## 3. METODOLOGÍA

La práctica se va a llevar a cabo mediante las herramientas que pone a disposición Babelomics. El servidor web se encuentra en la siguiente dirección:

<http://babelomics.bioinfo.cipf.es/>

El análisis que vamos a llevar a cabo embarca los siguientes pasos:

(i) subir los datos al servidor, (ii) visualizar los datos crudos, (iii) normalizar los datos (eliminar posibles artefactos) y (iv) llevar a cabo la comparación de la expresión génica entre dos condiciones. En un principio no es necesario registrarse. Trabajar como usuario registrado tiene la principal ventaja de que los datos se almacenan en el servidor con lo que se puede reanudar los análisis. Trabajando como usuario anónimo, los datos se pierden en cuando se cierra el navegador. Para estas prácticas no nos

vamos a registrar sino que vamos a trabajar como usuario anónimo. Para ello, pinchamos en el enlace “or start as anonymous user” que se encuentra en la parte derecha de la página principal (Figura 4).



Figura 4: La página principal de Babelomics. A la derecha podemos acceder como usuario, dar de alta a un usuario nuevo o acceder como usuario anónimo.

### 3.1. Estructura general de Babelomics

Después de acceder al servidor como usuario anónimo, podemos observar la estructura de la aplicación (Figura 5).

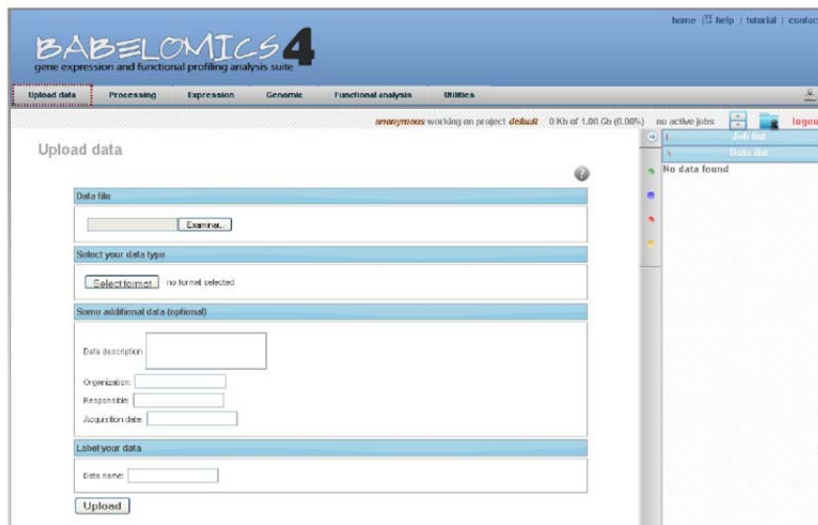


Figura 5: La interfaz para subir los datos al servidor

En el panel superior tenemos 6 opciones:

- [Upload data]: subir datos al servidor
- [Processing]: editar los datos y/o preprocesarlos incluyendo la normalización
- [Expression]: analizar los cambios de expresión para diferentes tipos de experimentos
- [Genomic]: análisis relacionados con polimorfismos de una base (SNPs)

- [Functional analysis]: análisis funcional de una lista de genes (no lo vamos a usar)
- [Utilities]: algunas herramientas para el análisis de datos incluyendo el visor de los datos crudos que vamos a usar en esta práctica.

A la derecha vemos un menú plegable que nos muestra los datos que hemos subido y los procesos que hemos lanzado indicando también su estado actual (pendiente, ejecutándose o terminado)

### 3.2. Subir los datos al servidor

Para poder analizar datos, primero tenemos que subirlos al servidor. Para ello descargamos primero el correspondiente fichero del servidor (<http://mendel.ugr.es/genetica>) y lo guardamos en el disco local. Este fichero comprimido en formato 'zip' contiene los datos de las 10 muestras (10 ficheros en formato 'CEL') que vamos a analizar. Luego pinchamos en "Upload data" en la página de Babelomics (En el menú, arriba). Hay cuatro campos obligatorios que tenemos que rellenar (Figura 5).

- [Data file]: pinchamos en 'Examinar' y navegamos hasta el fichero que queremos analizar
- [Select your data type]: tenemos que especificar el tipo de fichero: Microarray → Expression → One-channel → Affymetrix
- [Some additional data]: aquí podemos anotar los datos
- [Label your data]: aquí podemos asignar un nombre a nuestros datos

Una vez que hemos rellenado los cuatro campos obligatorios, pinchamos en "Upload" y se abre una nueva página donde podemos observar el progreso de la subida de datos. Una vez terminado (Figura 7), tenemos los datos disponibles "Data list" a la derecha.



Figura 6: Las distintas herramientas de Babelomics trabajan con un número alto de ficheros diferentes. En función del tipo de datos, la aplicación solo nos permite hacer los análisis correspondientes. Por eso tenemos que especificar el tipo de datos que estamos subiendo al servidor.

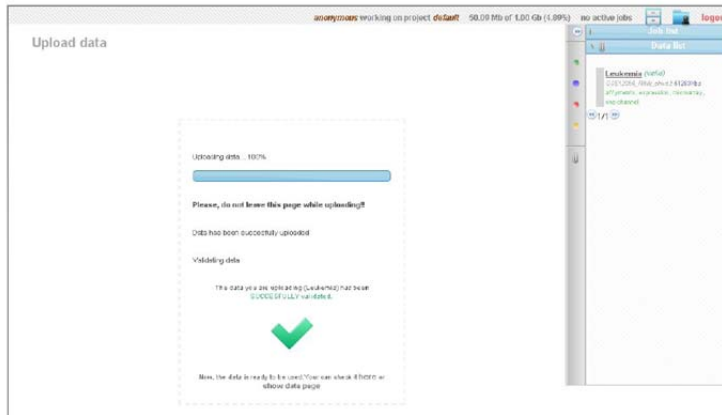


Figura 7: La pagina después de subir con éxito los datos al servidor.

### 3.3. Analizar los datos crudos

Primero vamos a inspeccionar los datos. Para ello, pinchamos en "Utilities" → "Microarray raw-data plots". Se abre otra interfaz donde tenemos que especificar los datos que queremos analizar (Figura 8). Ahí, marcamos los datos que hemos subido antes (Leukemia en el ejemplo).

Después de elegir los datos pinchamos en "Accept". Ahora podemos asignar un nombre a este proceso (por ejemplo: Leukemia raw) y lanzar el proceso.

El análisis tardará algún tiempo dependiendo de la carga de servidor. En la Figura 9 se puede observar la salida. Mientras se esta ejecutando este procesos, vamos a seguir preprocesando. Cuando tengamos los resultados de este paso, volveremos para comparar los dos RMA-box-plots.

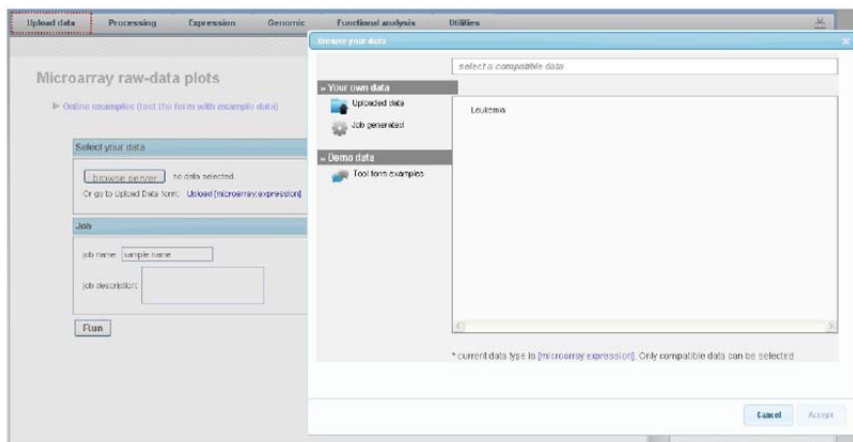


Figura 8: Seleccionar los datos crudos que queremos visualizar.

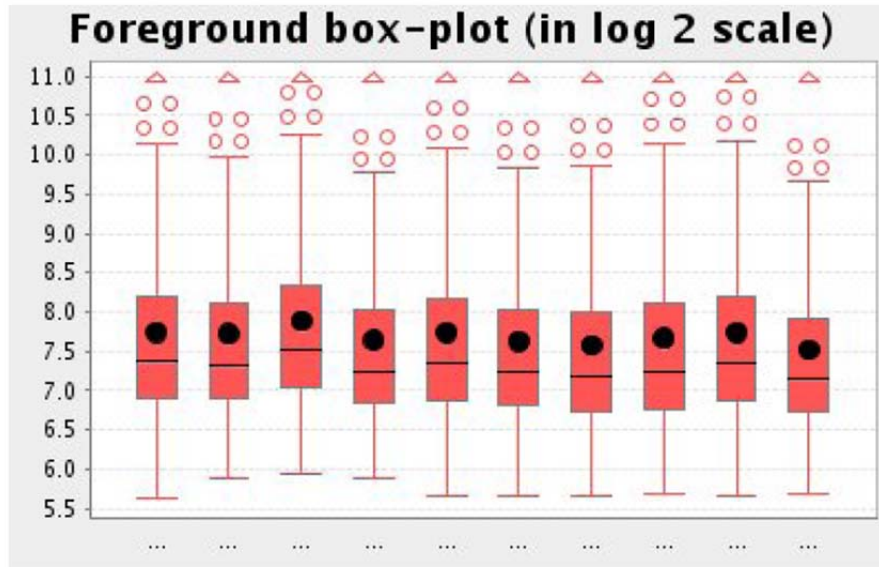


Figura 9: La distribución de las intensidades antes de normalizar los datos

### 3.4. Preprocesar los datos

Este paso elimina posibles artefactos y calcula la matriz de expresión que contiene los valores de expresión de todas las muestras. Para normalizar los datos, elegimos del menú el punto "Processing". En la página que se abre tenemos varias opciones: elegimos Normalize → Expression → One-channel → Affymetrix. Se abre otro formulario (Figura 10) que tenemos que rellenar primero para lanzar el proceso de normalización. Primero en "Select data" pinchamos en "browse server" para elegir los datos que queremos normalizar. Mantenemos el método por defecto (RMA) y asignamos un nombre al proceso (Leukemia norm en el ejemplo).

En cuando haya terminado el proceso, podemos comparar los dos RMA-box-plots, antes (Figura 9) y después (Figura 11) de la normalización. Se estima claramente el efecto de la normalización. Tanto las medias como las distribuciones se asemejan mucho más después de la normalización.



Figura 10: La página para lanzar el proceso de normalización.

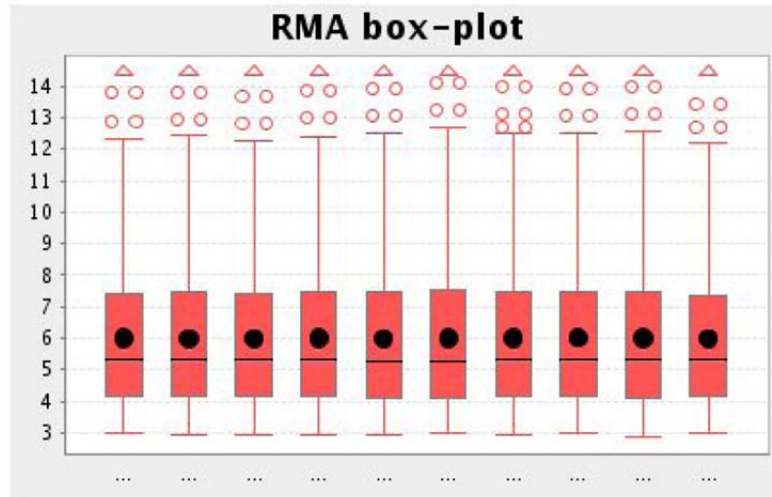


Figura 11: Las distribuciones de intensidad después de la normalización.

### 3.5. Detectar los genes que se expresan diferencialmente

Los datos con los que estamos trabajando se componen de 10 muestras, 5 provienen de la línea celular K562 sin manipular (controles) y 5 con el gen CREB inactivado (knocked-out) (casos). Para poder determinar los genes que se expresan diferencialmente, el programa tiene que asignar todas las muestras a una de las dos condiciones, comparando después la varianza dentro del grupo con la varianza entre los dos grupos. Hasta ahora, el programa no tiene conocimiento de qué muestras pertenecen a cada grupo (condición). Así que el primer paso debe ser asignar una etiqueta a cada muestra que indique su pertenencia a uno de los dos grupos (condiciones). Para ello, pinchamos otra vez en la opción "Processing" y después en el enlace "Edit". Primero tenemos que elegir el conjunto de datos que queremos editar. En "Select your data" pinchamos en "Browse data" y se abre una ventana que nos muestra todos los datos que tenemos a disposición, tanto los que hemos subido (datos crudos) como los que hemos generado mediante la normalización. Así que, en el menú "Your own data" elegimos "Job generated". Pinchamos sobre el nombre que hemos dado a los datos de normalización (leukemia norm en el caso del ejemplo) y aparece el fichero que tenemos que editar, "rma.summary". Lo marcamos y damos a "Aceptar" abajo en la ventana (Figura 12).

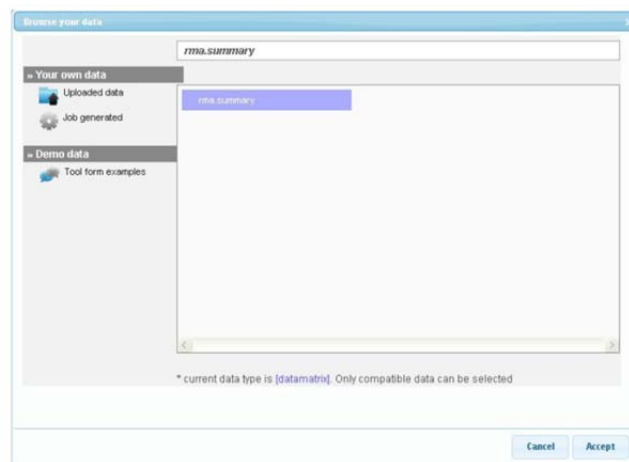


Figura 12: Seleccionar los datos para asignar las muestras a las condiciones/grupos.



El siguiente paso consiste en generar una variable con dos valores. Los valores son los nombres de los dos grupos o condiciones. La elección de los nombres no incluye en los resultados pero deberían ser mnemotécnicos, por ejemplo caso/control o K562/CREB Knock podrían ser nombres posibles.

La tabla muestra los nombres de las muestras que tenemos en el análisis y una breve descripción de la que podemos deducir a qué grupo pertenece cada muestra.

Nombre de la muestra	Descripción
GSM304303	K562 cells control replicate 1
GSM304304	K562 cells control replicate 2
GSM304479	K562 cells control replicate 3
GSM304498	K562 cells control replicate 4
GSM304480	K562 cells control replicate 5
GSM304487	K562 cells CREB Knock-out replicate 1
GSM304488	K562 cells CREB Knock-out replicate 2
GSM304489	K562 cells CREB Knock-out replicate 3
GSM304490	K562 cells CREB Knock-out replicate 4
GSM304491	K562 cells CREB Knock-out replicate 5

Para asignar la variable y sus valores, pinchamos en "Create new variable". Le damos un nombre (por ejemplo CREBknockout) y elegimos "categorical" como tipo de variable. Después podemos pinchar en la "varita" que hay debajo del campo "Values". Se abre una ventana donde podemos especificar el nombre de la etiqueta. Esto tenemos que hacer dos veces, para cada condición una vez. En el ejemplo se ha escogido los nombres "K562" y "CREBKnockout" para caracterizar las dos condiciones. Una vez definidos los nombres de los grupos, podemos asignar las muestras a uno de los grupos. Para ellos pinchamos en "click to edit samples" (Figura 13). Se abre una tabla con los nombres de las muestras y un campo vacío donde tenemos que poner el grupo al que pertenece. Lo podemos rellenar mediante la información que hay en la tabla arriba. Una vez terminado pinchamos en "Submit" y así se guarda la asignación en el servidor. Ahora tenemos los datos preparados para comparar los niveles de expresión entre nuestros dos grupos. Para ello nos vamos al menú "Expression" y allí elegimos "Class comparison". En esta página tenemos que facilitar una serie de datos y parámetros (Figura 14):

- Select your data: mediante esta opción podemos seleccionar los datos que queremos analizar. Para ello pinchamos en "browse server" y se abre la misma ventana que antes y elegimos otra vez los datos de "rma.summary".
- Select the class to analyse: podemos ahora elegir la variable que hemos definido antes y las clases que queremos analizar.
- Select test: según, el número de clases que haya, el programa nos pone a disposición los tests de estadística adecuados. En nuestro caso[ que tenemos dos clases/condiciones podemos elegir entre 3 tests:] T-test, Limma y Fold-change.
- Select multiple-test correction: tenemos 5 modelos diferentes a disposición para corregir por multiple testing (controlar la acumulación de errores de tipo I)

- Select adjusted p-value: tenemos que poner el umbral para el valor P (0.05 por defecto)
- Job: podemos asignar un nombre al proceso

Después de haber rellenado todos los campos, pinchamos en run y vemos como el nuevo proceso aparece a la derecha en "Job list".

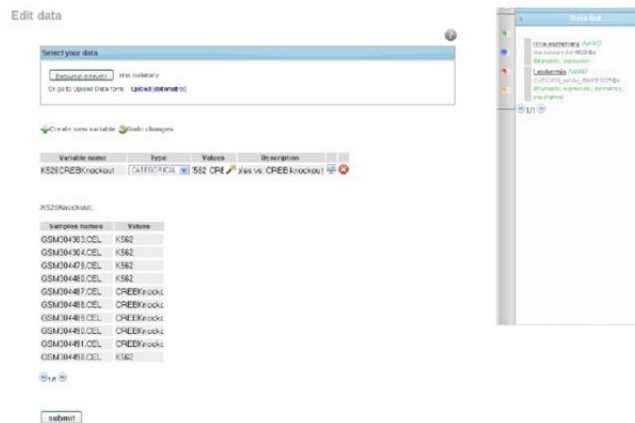


Figura 13: La pantalla donde tenemos que asignar la etiqueta que indica a que condición pertenece cada una de las 10 muestras.

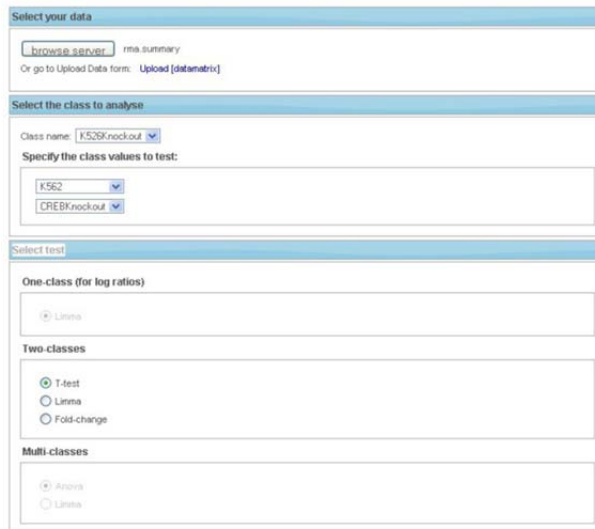


Figura 14: La interfaz para lanzar el proceso que detecta los genes que se expresan diferencialmente.



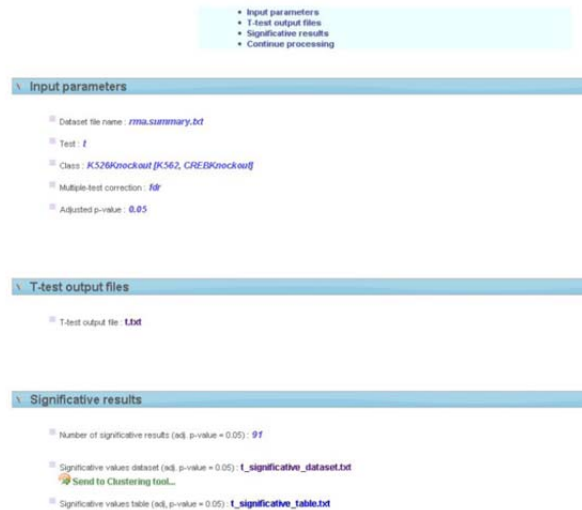


Figura 15: Resumen de los parámetros y los resultados en formato texto para descargar.

### 3.6. Analizar e interpretar los resultados

La página de resultados consta de varios apartados.

La primera parte contiene: (i) resumen de los parámetros, (ii) la salida completa del test estadístico, (iii) la matriz de expresión de los genes que se expresan diferencialmente de forma estadísticamente significativa `t_significative_dataset.txt` y (iv) un fichero que solo contiene los genes significativos `t_significative_table.txt`

La segunda parte muestra los resultados significativos de forma navegable (Figura 16). Podemos observar que en la columna “statistic” (el estadístico) hay tanto valores positivos como negativos. El signo nos permite distinguir entre sobre y infra-expresión. Si el signo es positivo, el gen se expresa más en la primera condición según el orden que hemos puesto en la matriz de expresión (en nuestro caso son los controles: CATEGORICAL {k562,CREB})

La tercera parte de los resultados consiste de una representación gráfica de los resultados en forma de un heat map (Figura 17). Los genes están ordenados por el estadístico. Arriba se muestran los genes sobre-expresados en el primer grupo (controles=K562 en nuestro caso) y abajo los genes reprimidos en los controles o sobre-expresados en los casos (que es equivalente).

name	statistic	p-value	adj. p-value
218055_s_at	15.95	4.69e-7	0.005242
205352_at	14.91	0.0000519	0.04784
241950_at	13.96	0.00002298	0.03695
226751_at	12.43	0.00001591	0.03261
203680_at	10.94	0.000055	0.04784
220520_s_at	10.7	0.000005911	0.02955
222550_at	10.56	0.000006402	0.02955
209362_at	10.54	0.000009515	0.02955
208946_s_at	10.39	0.00001606	0.03261
210758_at	10.11	0.000007856	0.02955

page 1 of 10

Figura 16: Las sondas/genes que se expresan de forma diferencial

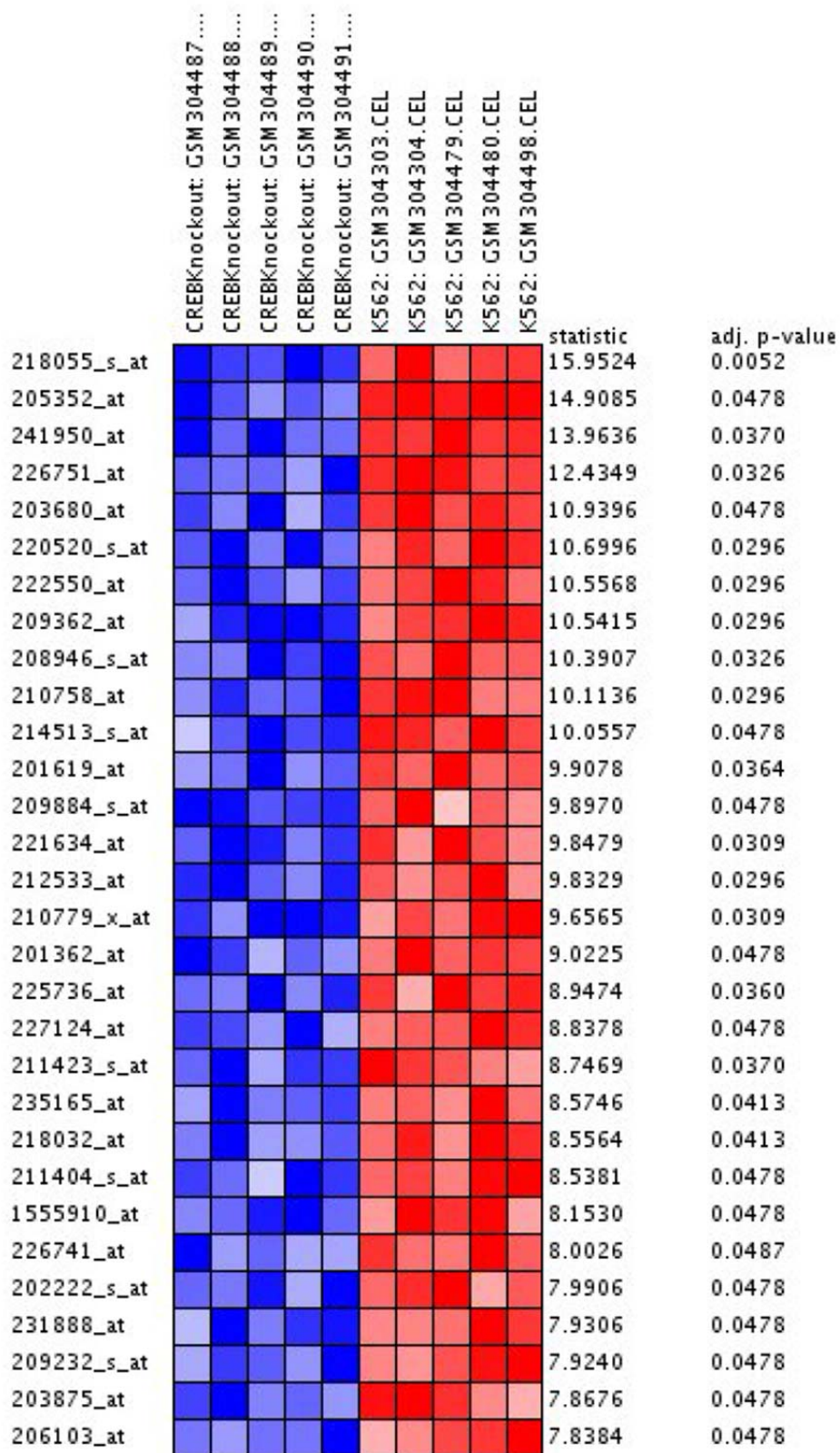


Figura 17: Una representación gráfica de los resultados. El color rojo indica sobre-expresión mientras azul indica represión.



Figura 18: Enlaces directos para llevar a cabo análisis funcionales con la lista genes que se expresan diferencialmente.

Finalmente, la última parte de la página nos permite descargar todos los datos, o llevar a cabo análisis funcionales (Figura 18).

#### 4. CUESTIONES

El alumno puede descargarse más conjuntos de datos para practicar desde la web del Departamento de Genética (<http://mendel.ugr.es>). También se facilitarán los resultados para la comprobación. Preguntas a las que el alumno deberá poder responder después de la práctica:

- Cómo se puede distinguir entre genes que se sobre-expresan y los que se reprimen mediante el estadístico elegido.
- Cuál es el principio básico de un chip de ADN. Qué artefactos técnicos existen y como se pueden eliminar.
- Cuáles son los pasos del análisis bioinformático en un experimento de chip de ADN para detectar expresión diferencial: de los datos crudos en formato "CEL" hasta obtener una lista de genes que se expresan diferencialmente.
- Qué es la expresión diferencial.
- En qué experimentos es de interés y por qué.



EXPRESIÓN DE GENES  
IMPLICADOS EN EL  
DESARROLLO TESTICULAR DE  
MAMÍFEROS



# EXPRESIÓN DE GENES IMPLICADOS EN EL DESARROLLO TESTICULAR DE MAMÍFEROS

## 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.

## 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 subpoblación de células somáticas a diferenciarse como células de Sertoli, encargadas de orquestrar 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 cómo 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*<sup>-/-</sup> (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*<sup>-/-</sup> 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.

### 3. METODOLOGÍA

#### 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'- GCAAAC AGC TTT GTG GTC AA 3'  
Sry-R 5'- GGAAA 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'

-Componentes de la reacción:

- 10X Solución tampón Taq polimerasa 2.5 µl
- 25 mM MgCl<sub>2</sub> 1.5 µl
- 25 mM dNTP 0.2 µl
- DMSO 1.25 µl
- Sry-F (500 ng/µl) 0.5 µl
- Sry-R (500 ng/µl) 0.5 µl
- Mio-F (500 ng/µl) 0.5 µl
- Mio-R (500 ng/µl) 0.5 µl
- H<sub>2</sub>O 16.45 µl
- ADN 100 ng
- Taq polimerasa 0.1 µl
- Volumen total 25 µl

-Pasos de la PCR en el termociclador:

- 94°C, 3 minutos
- 35 ciclos:
  - 91°C, 45 segundos
  - 60°C, 60 segundos
  - 72°C, 45 segundos
- 72°C, 5 minutos
- 4°C, indefinidamente

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).

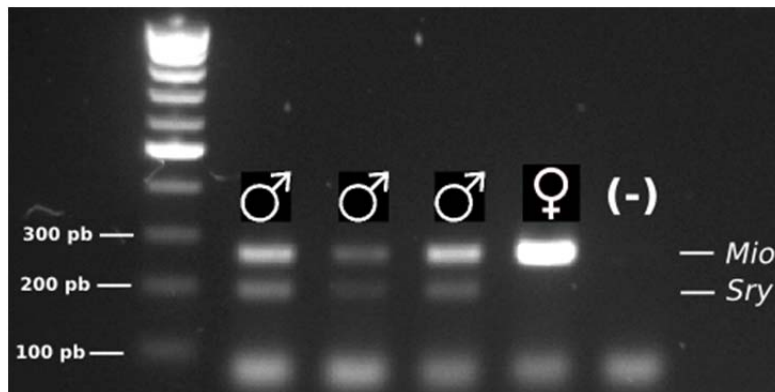


Figura 1: Electroforesis de los productos de una PCR realizada para el sexado de embriones de ratón. La presencia en el gel de una banda correspondiente al gen *Sry* denota la presencia de un macho, mientras que su ausencia indica que ese embrión es hembra. La banda de la miogenina sirve como control 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.

### 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 inmunohistoquímica, 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  $2H_2O_2 \rightarrow 2H_2O + O_2$ . 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).

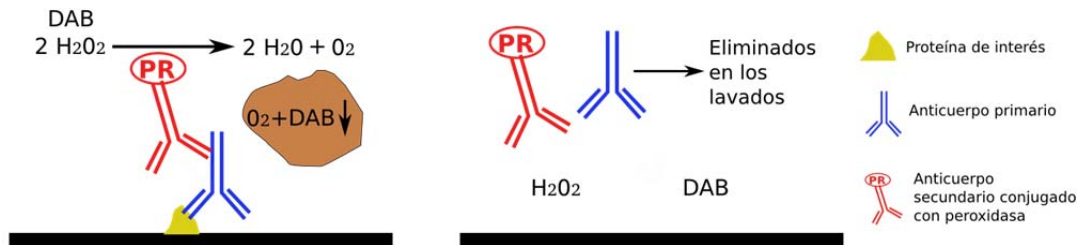


Figura 2: Fundamento de la técnica de inmunohistoquímica. La presencia en la muestra de la proteína de interés (esquema de la izquierda) permite el anclaje a la preparación del complejo compuesto por el anticuerpo primario, el anticuerpo secundario y la peroxidasa, permitiendo la reacción coloreada con 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)

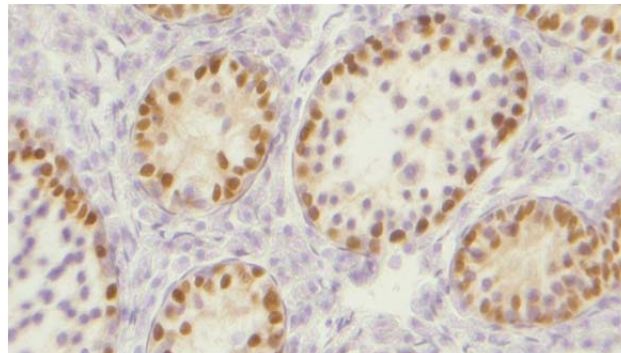


Figura 3: Marcaje inmunohistoquímico de tejido testicular de ratón, usando un anticuerpo primario anti-Sox9. Sólo las células de Sertoli aparecen marcadas con el color marrón. El resto de las células se muestran azul claro por la contra-tinción realizada con hematoxilina.

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.

#### 4. 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).

# ESTUDIO DE EXPRESIÓN GÉNICA MEDIANTE RT-PCR



# ESTUDIO DE EXPRESIÓN GÉNICA MEDIANTE RT-PCR

## 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.

## 2. FUNDAMENTO TEÓRICO

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

Los conductos de Müller (o conductos paramesonéfricos) y los conductos de Wolff (o conductos mesonefricos) 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).

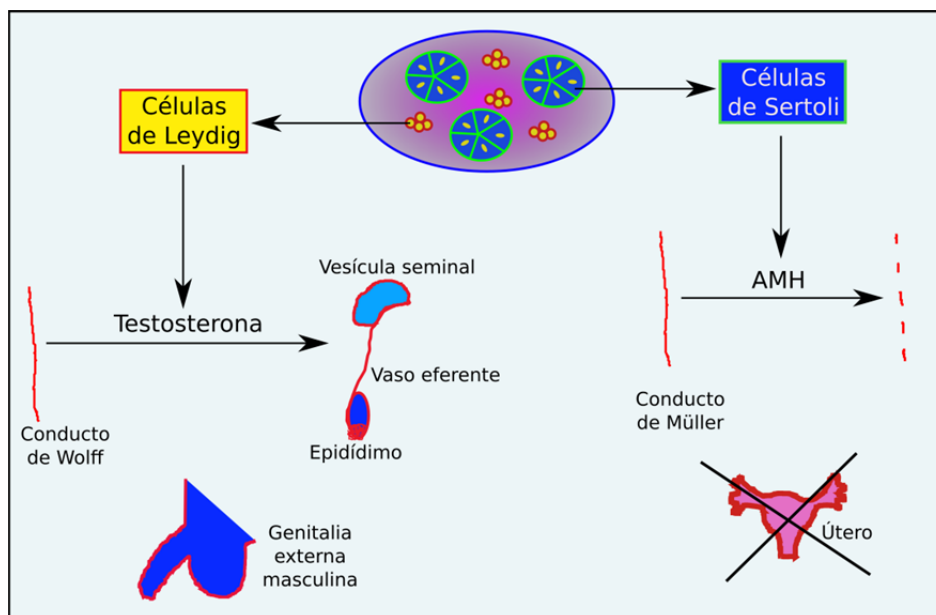


Figura 1: Esquema del desarrollo testicular. La producción hormonal de los testículos incluye testosterona y AMH



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 TGF $\beta$ . 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 transcrito 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)

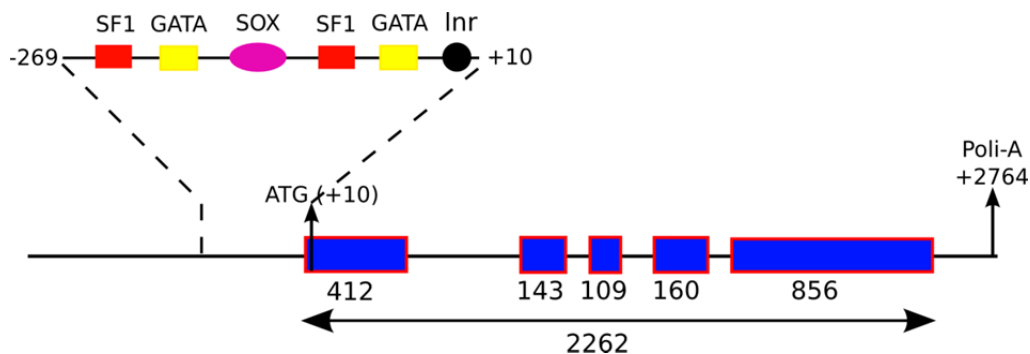


Figura 2: Estructura del gen de la AMH humana.

### **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 anomalías 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 excretorios 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.

## **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 -80°C. 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 desnaturante 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  $\mu$ l buffer de lisis (10  $\mu$ l  $\beta$ -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  $\mu$ l EtOH 70% y mezclar por inversión.
5. Transferir a una columna.
6. Centrifugar 1 minuto a máxima velocidad.
7. Digestión del ADN: Añadir 80  $\mu$ l de solución de Dnasa I, 15 minutos, temperatura ambiente. (preparar gel agarosa para comprobar calidad del ARN)
8. Añadir 700  $\mu$ l de buffer RW1, centrifugar 1 minuto a máxima velocidad, desechar sobrenadante.
9. Añadir 500  $\mu$ l de buffer RPE, centrifugar 1 minuto a máxima velocidad, desechar sobrenadante.
10. De nuevo, añadir 500  $\mu$ l buffer RPE, centrifugar 2 minutos a máxima velocidad, desechar sobrenadante.
11. Colocar la columna en un Eppendorf limpio. Añadir 30  $\mu$ l 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.

- Volumen final de la reacción                    25 µl
  
- ARN total    1 µl
- 5 x Buffer    2.5 µl
- dNTPs (10 mM)                                    1 µl
- Cebador AMH-F (10 mM)                        1 µl
- Cebador AMH-R (10 mM)                        1 µl
- *One step RT-PCR mix*                            1 µl
- Inhibidor de ARNasa                            1 µl
- H<sub>2</sub>O libre de ARNasa                            17.5 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.

#### 4. 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?