TEMA 3.2. Obtención de matrices: búsqueda de homologías, alineamiento

Contacto: Isabel Draper (Isabel.draper@uam.es)

INTRODUCCIÓN

Una vez obtenidas las secuencias, es necesario agruparlas para construir la matriz de datos. Para ello se colocan las secuencias superpuestas, de tal manera que cada fila corresponde a una muestra y cada columna a un carácter (que en el caso de las secuencias será molecular –un nucleótido o un conjunto de nucleótidos-, pero que también puede ser morfológico). Este es el proceso que se denomina alineamiento, y busca enfrentar aquellas partes de las secuencias que son homólogas (idénticas o con el menor número de cambios posible), para así poder identificar mutaciones, inserciones, deleciones, inversiones, etc. producidas entre las muestras estudiadas (Fig. 1).

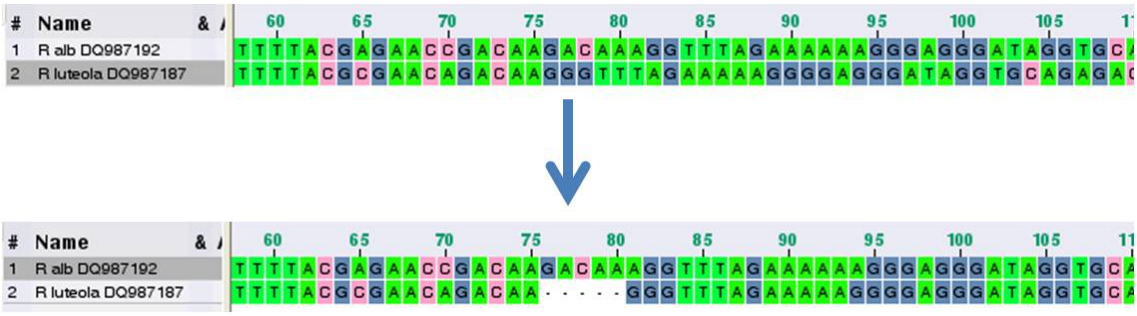


Figura 1. Ejemplo de búsqueda de homología para realizar un alineamiento.

La correcta elaboración del alineamiento constituye un paso fundamental para la obtención de una buena filogenia, dado que la topología de los árboles filogenéticos que se obtengan al final del proceso está totalmente condicionada por las asunciones de homología reflejadas en la matriz de datos (Simmons et al, 2001; Simmons y Freudenstein, 2003). Sin embargo, en algunas ocasiones, el alineamiento de las secuencias es difícil, dado que las secuencias pueden haber acumulado multitud de cambios que pueden ser interpretados de diferentes maneras (Fig. 2), o que implican unas secuencias tan diferentes entre sí que se complica la búsqueda de zonas homólogas. Además, dos secuencias de ADN pueden presentar hasta un 25% de identidad sólo por azar (Simmons y Freudenstein, 2003).

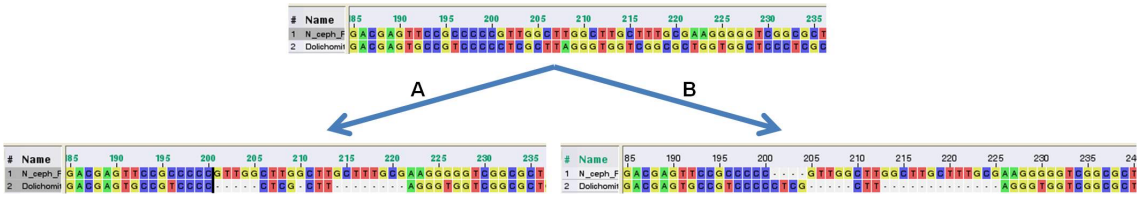


Figura 2. Distintas alternativas (A y B) para el alineamiento de las dos mismas secuencias.

Las mutaciones del ADN pueden ser: sustituciones (cambio de un nucleótido por otro en una misma posición), inserciones (adición de nuevos nucleótidos a la secuencia) o deleciones (pérdidas de nucleótidos en la secuencia). Las mutaciones más frecuentes son las sustituciones, sin embargo no es raro encontrar inserciones y deleciones de diferente tamaño. La presencia de inserciones o deleciones en determinadas secuencias supone un reto en el alineamiento al obligar a la inclusión de un nuevo carácter (denominado *gap* o *indel y* representado por el símbolo "-") en la(s) secuencia(s) que carezcan de esa posición. En el ejemplo de la figura 2 se puede observar como la posibilidad A implica la creación de tres huecos –*gaps*- (por inserción en la primera secuencia o deleción en la segunda) y una mutación entre las posiciones 200 y 222, mientras la posibilidad B implica tres gaps y ninguna mutación en ese mismo conjunto de bases.

En general, se considera que el mejor alineamiento es aquel que tiene más sentido biológico. Existen numerosos programas informáticos que realizan alineamientos de manera automática, estableciendo cuál de los posibles alineamientos tiene más sentido biológico mediante sistemas de evaluación (*scoring*). La manera más simple de evaluar los alineamientos consiste en puntuar de manera diferente las mutaciones y la apertura de *gaps* en el alineamiento (por ejemplo sumando un punto por cada base coincidente, dado un valor de cero a cada mutación y restando un punto por cada base que falte). En cualquier caso, los alineamientos obtenidos siempre deben ser revisados manualmente para valorar el sentido biológico de los cambios propuestos (ej. Doyle y Gaut, 2000; Kelchner, 2000). Finalmente es muy importante decidir cómo tratar los *gaps* del alineamiento a la hora de elaborar la reconstrucción filogenética (Simmons y Ochoterena, 2000). En la mayor parte de los programas la opción establecida por defecto considera los *gaps* como información desconocida (*missing data*, representado el matriz por el símbolo "?"). Sin embargo, muchos autores sostienen que los *gaps* pueden tener sentido filogenético. Existen dos maneras de considerar los *gaps* como informativos para los análisis: como un quinto estado (A, T, C, G, -), con el inconveniente de que cada base ausente es considerada como un carácter independiente (cuando puede no serlo si ha habido una inserción o deleción múltiple), o codificando cada *indel* (inserción o deleción; Simmons y Ochotorena, 2000).

Una vez obtenido el alineamiento, este puede ser guardado en diferentes formatos de texto para su posterior análisis. Los formatos más frecuentemente utilizados son fasta, nexus y phylip.

El formato **fasta** (Fig. 3) incluye el nombre de la secuencia, que se distingue por empezar por el símbolo >, y los datos de la secuencia, nucleótidos representados por un código de letras que normalmente es:

Código	Significado	Código	Significado
A	Adenosina	S	G / C
C	Citosina	W	A / T
G	Guanina	B	G / T / C
T	Timidina	D	G / A / T
U	Uracilo	H	A / C / T
R	G / A	V	G / C / A
Y	T / C	N	A / G / C / T
K	G / T	X	máscara
M	A / C	-	*gap*

```
>R_alb_DQ987192¶
CTTTCAAATTCAGAGAAACCCTGGAATTAACAATGGGCAACCCTGAGCCAAATCCTGTTTTACGAGAACCGAC
AAGACAAAGGTTTAGAAAAAAGGGAGGGATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACAAATGGAGTTC
ACTGCCTTGTGTTCCTTATGTAATACTATTTTGAAATTTAATATTCAGTAACAAAAAATCACTCCATAGTCTG
ATTAATAACTGATTAATCGGACGAGAATAAAGATAGAGTCCTATTCTACATGTCAATACTGACAACAATGAAA
TTTATCGTAAGATGAAAATCCGTCGACTTTTAAAATCGTGAGGGTTCAAGTCCCTCTATCCCCAACTTTATCC
CTCCAAAAGAGTCTGTTTGAGGCTCTACCTAATTTTAATTTTAGTTATTCAAAAATTCATTATCATTTGCGTT
GATCCTACTCTTTTACAAACGTATCTGAGCAGAATTTTGTATCTTATTACAAGTCTTGTAGAATATATGAGAC
TCATACAAATGAGAAAAAAATACCGATTTGACTGATTTACAATCTATAGCATTATTCATTTTAAAACTTATAA
AGTATTCCTTTTGAAAATCTAAGAAATTCCCGGTCCAAGACTTAATACTTTTAGTTTTCTTTTCATTGACATA
GACCTAAGTCATCGCGTAAAATGAAGATGATGCTTCGGTAA??????????????????????????????????¶
>R_luteola_DQ987187¶
CTTTCAAATTCAGAGAAACCCTGGAATTAACAATGGGCAACCCTGAGCCAAATCCTGTTTTACGCGAACAGAC
AAGGGTTTAGAAAAAGGGGAGGGATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACAAATGGAGTTCACTAC
CTTGTGTTCCTAAAGAGAAAAACTTATATTGCATAAGTATATCTAATGTAATACTATTTTAAAATTGAATATT
GATTTTAAAAAAATGCATTCCATAGTCTGATAAATAACTGATTAATCGGACGAGAATAAAGATAGAGTCCTAT
TCTACATGTCAATACTGACAACAATGAAATTTATCGTAAGATGAAAATCCGTCGACTTTTAAAATCGTGAGGG
TTCAAGTCCCTCTATCCCCAACTTTATCCCTCCAAAAGAGTCTGTTTGATGCTCTACCAAATTTTATTTTTTA
GTTATTCAACAATTCATTATCATTTGCGTTGATCCTACTTTTTTACAAACGTATCTGAGCAGAATTTTATATC
TTATTACAAGTCTTGTAGGATATATAAGACTCATACAAATGAGCAAAAAATACCGATTTGACTGATTTACAAT
CTATAGCATTATTCATTTTAAAACTTATAAAGTATTCCTTTTGAAAATCTAAGAAATTCCCGGTCCAAGACTT
AATACTTTGATTTTTCTTTTCATTGACATAGACCTAAGTCATCGCGTAAAATGAAGATGATGCTTCGGTAA??¶
```

Figura 3. Representación de dos secuencias de ADN de *Reseda luteola* y *R. alba* en formato fasta.

El formato **nexus** (Fig. 4) incluye un bloque inicial, previo a la matriz de datos, mediante el cual se especifican:

- el formato (#Nexus)
- las dimensiones de la matriz (ntax=número de secuencias y nchar=número de caracteres)
- el tipo de datos incluidos (datatype=dna o restriction –si se trata de gaps codificados- o standard –si son caracteres morfológicos codificados-)
- la manera de alternar las secuencias (interleave=yes –si en la matriz se alternan la primera línea de la muestra 1, la primera línea de la muestra 2, la segunda línea de la muestra 1, la segunda línea de la muestra 2, etc.- o interleave=no –si la matriz está constituida por la primera secuencia completa, seguida por la segunda secuencia completa, etc.-)
- el código de símbolos utilizado

A continuación de este bloque aparece la matriz de datos, que en el caso de secuencias moleculares suele estar constituida por el nombre de la secuencia (en este caso no se acepta el comienzo con >) y sus nucleótidos. La matriz siempre termina con ; END;

```
#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=2 NCHAR=730;
FORMAT DATATYPE=DNA INTERLEAVE=NO SYMBOLS="ACTG" GAP=- MISSING=? ;

MATRIX
R_alb_DQ987192
CTTTCAAATTCAGAGAAACCCTGGAATTAACAATGGGCAACCCTGAGCCAAATCCTGTTTTACGAGAACCGACA
AGACAAAGGTTTAGAAAAAAGGGAGGGATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACAAATGGAGTTCA
CTGCCTTGTGTTCCTTATGTAATACTATTTTGAAATTTAATATTCAGTAACAAAAAATCACTCCATAGTCTGATTAA
TAACTGATTAATCGGACGAGAATAAAGATAGAGTCCTATTCTACATGTCAATACTGACAACAATGAAATTTATCG
TAAGATGAAAATCCGTCGACTTTTAAAATCGTGAGGGTTCAAGTCCCTCTATCCCCAACTTTATCCCTCCAAAAG
AGTCTGTTTGAGGCTCTACCTAATTTTAATTTTAGTTATTCAAAAATTCATTATCATTTGCGTTGATCCTACTCTTT
TACAAACGTATCTGAGCAGAATTTTGTATCTTATTACAAGTCTTGTAGAATATATGAGACTCATACAAATGAGAAA
AAAATACCGATTTGACTGATTTACAATCTATAGCATTATTCATTTTAAAACTTATAAAGTATTCCTTTTGAAAATCT
AAGAAATTCCCGGTCCAAGACTTAATACTTTTAGTTTTCTTTTCATTGACATAGACCTAAGTCATCGCGTAAAATG
AAGATGATGCTTCGGTAA??????????????????????????????????
R_luteola_DQ987187
CTTTCAAATTCAGAGAAACCCTGGAATTAACAATGGGCAACCCTGAGCCAAATCCTGTTTTACGCGAACAGACA
AGGGTTTAGAAAAAGGGGAGGGATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACAAATGGAGTTCACTACC
TTGTGTTCCTAAAGAGAAAAACTTATATTGCATAAGTATATCTAATGTAATACTATTTTAAAATTGAATATTGATTT
TAAAAAAATGCATTCCATAGTCTGATAAATAACTGATTAATCGGACGAGAATAAAGATAGAGTCCTATTCTACAT
GTCAATACTGACAACAATGAAATTTATCGTAAGATGAAAATCCGTCGACTTTTAAAATCGTGAGGGTTCAAGTCC
CTCTATCCCCAACTTTATCCCTCCAAAAGAGTCTGTTTGATGCTCTACCAAATTTTATTTTTTAGTTATTCAACAA
TTCATTATCATTTGCGTTGATCCTACTTTTTACAAACGTATCTGAGCAGAATTTTATATCTTATTACAAGTCTTGT
AGGATATATAAGACTCATACAAATGAGCAAAAAATACCGATTTGACTGATTTACAATCTATAGCATTATTCATTTT
AAAACTTATAAAGTATTCCTTTTGAAAATCTAAGAAATTCCCGGTCCAAGACTTAATACTTTGATTTTTCTTTTCAT
TGACATAGACCTAAGTCATCGCGTAAAATGAAGATGATGCTTCGGTAA??
;
END;
```

Figura 4. Representación de las mismas dos secuencias de ADN de *Reseda luteola* y *R. alba* en formato nexus.

El formato **Phylip** también incluye una primera línea en la que se indican las características de la matriz (Fig. 5). En este caso esa primera línea está constituida por un primer número que se refiere al número de muestras, un segundo número que se refiere al número de caracteres, y una letra (i si los datos aparecen interleaved, s si aparecen secuenciales). A continuación de esta línea aparece la matriz, constituida por el nombre de la secuencia, seguida por los nucleótidos que la constituyen.

```
2··730··s¶
¶
R_alb_DQ98¶
CTTTCAAATTCAGAGAAACCCTGGAATTAACAATGGGCAACCCTGAGCCAAATCCTGTTTTACGAGAACCGAC
AAGACAAAGGTTTAGAAAAAAGGGAGGGATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACAAATGGAGTTC
ACTGCCTTGTGTTCCTTATGTAATACTATTTTGAAATTTAATATTCAGTAACAAAAAATCACTCCATAGTCTG
ATTAATAACTGATTAATCGGACGAGAATAAAGATAGAGTCCTATTCTACATGTCAATACTGACAACAATGAAA
TTTATCGTAAGATGAAAATCCGTCGACTTTTAAAATCGTGAGGGTTCAAGTCCCTCTATCCCCAACTTTATCC
CTCCAAAAGAGTCTGTTTGAGGCTCTACCTAATTTTAATTTTAGTTATTCAAAAATTCATTATCATTTGCGTT
GATCCTACTCTTTTACAAACGTATCTGAGCAGAATTTTGTATCTTATTACAAGTCTTGTAGAATATATGAGAC
TCATACAAATGAGAAAAAAATACCGATTTGACTGATTTACAATCTATAGCATTATTCATTTTAAAACTTATAA
AGTATTCCTTTTGAAAATCTAAGAAATTCCCGGTCCAAGACTTAATACTTTTAGTTTTCTTTTCATTGACATA
GACCTAAGTCATCGCGTAAAATGAAGATGATGCTTCGGTAA??????????????????????????????????¶
R_luteola_¶
CTTTCAAATTCAGAGAAACCCTGGAATTAACAATGGGCAACCCTGAGCCAAATCCTGTTTTACGCGAACAGAC
AAGGGTTTAGAAAAAGGGGAGGGATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACAAATGGAGTTCACTAC
CTTGTGTTCCTAAAGAGAAAAACTTATATTGCATAAGTATATCTAATGTAATACTATTTTAAAATTGAATATT
GATTTTAAAAAAATGCATTCCATAGTCTGATAAATAACTGATTAATCGGACGAGAATAAAGATAGAGTCCTAT
TCTACATGTCAATACTGACAACAATGAAATTTATCGTAAGATGAAAATCCGTCGACTTTTAAAATCGTGAGGG
TTCAAGTCCCTCTATCCCCAACTTTATCCCTCCAAAAGAGTCTGTTTGATGCTCTACCAAATTTTATTTTTTA
GTTATTCAACAATTCATTATCATTTGCGTTGATCCTACTTTTTTACAAACGTATCTGAGCAGAATTTTATATC
TTATTACAAGTCTTGTAGGATATATAAGACTCATACAAATGAGCAAAAAATACCGATTTGACTGATTTACAAT
CTATAGCATTATTCATTTTAAAACTTATAAAGTATTCCTTTTGAAAATCTAAGAAATTCCCGGTCCAAGACTT
AATACTTTGATTTTTCTTTTCATTGACATAGACCTAAGTCATCGCGTAAAATGAAGATGATGCTTCGGTAA??¶
¶
```

Figura 5. Representación de las mismas dos secuencias de ADN de *Reseda luteola* y *R. alba* en formato phylip.

PROGRAMAS NECESARIOS

Existen numerosos programas informáticos para el alineamiento de secuencias, tanto para hacer un alineamiento manual secuencia a secuencia, como para obtener un alineamiento automático. Algunos de estos programas pueden ser utilizados directamente en aplicaciones de Internet, como por ejemplo en:

http://www.ebi.ac.uk/Tools/msa/

En esta sección te proponemos la utilización de ClustalW2 (Larkin et al, 2007; Thompson et al, 1994) y MUSCLE (Edgar, 2004) para el alineamiento automático (disponibles en la dirección de Internet indicada), y PhyDE (Müller et al: http://www.phyde.de/) para el alineamiento manual.

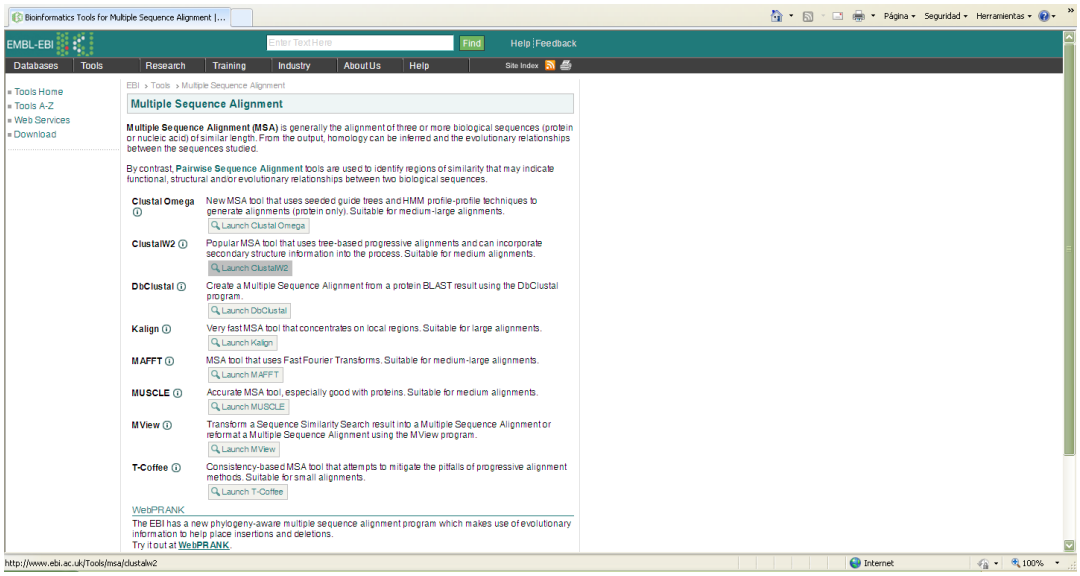
PhyDE es un editor de alineamientos con el que puedes realizar manualmente un alineamiento desde cero, o modificar un alineamiento previamente obtenido, como vamos a hacer a continuación.

Este programa puede ser descargado desde http://www.phyde.de/download.html. Al ser instalado, automáticamente se creará una carpeta llamada PhyDE-Data. Para poder utilizar los plugins desde PhyDE debes buscar dónde se localiza esta carpeta en tu ordenador, crear dentro de ella dos subcarpetas: tmp y Plugins, e instalar los Plugins en la carpeta Plugins que acabas de crear. El manual de utilización de PhyDE puedes consultarlo en: http://www.phyde.de/docu/docu.html.

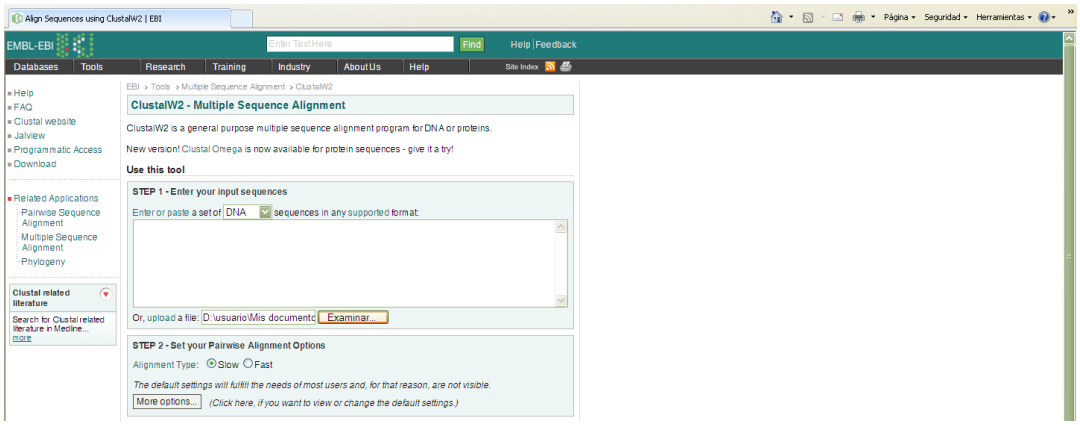
METODOLOGÍA Y PRÁCTICA

I. Alineamiento mediante ClustalW2

Paso 1. Desde la página http://www.ebi.ac.uk/Tools/msa/ abre ClustalW2:



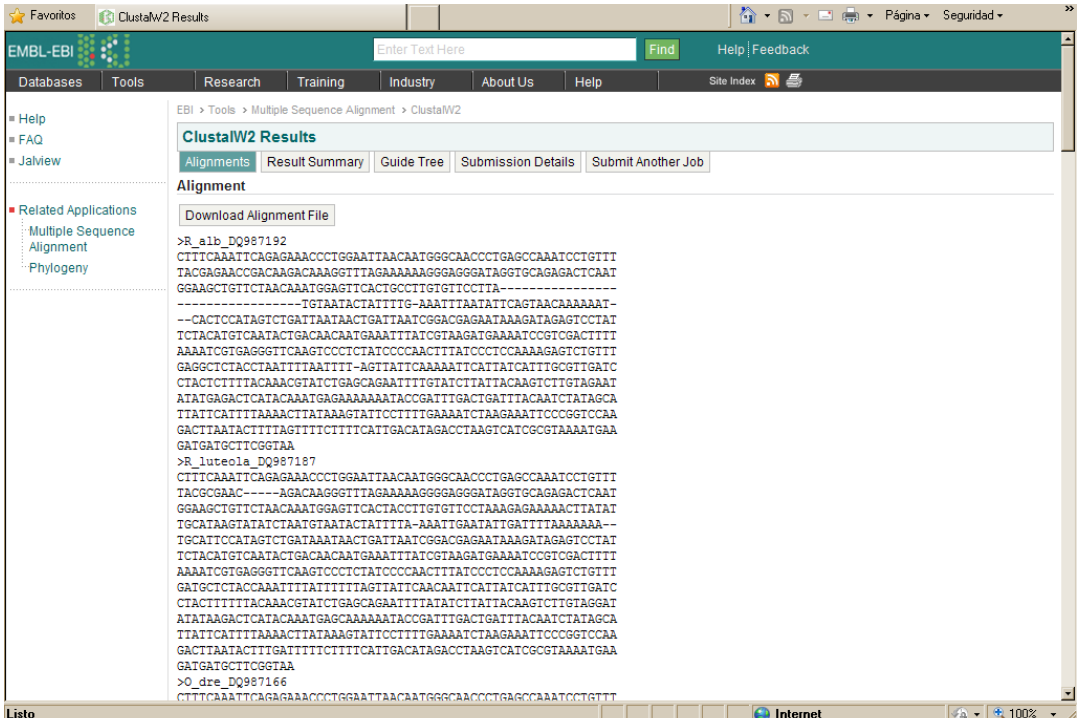
Paso 2. El primer paso (*step* 1) consiste en seleccionar el tipo de secuencias (DNA en este caso) y cargar el archivo para alinear. Carga Secuencias_LF.fas generado desde *GenBank* en el tema 3.1:



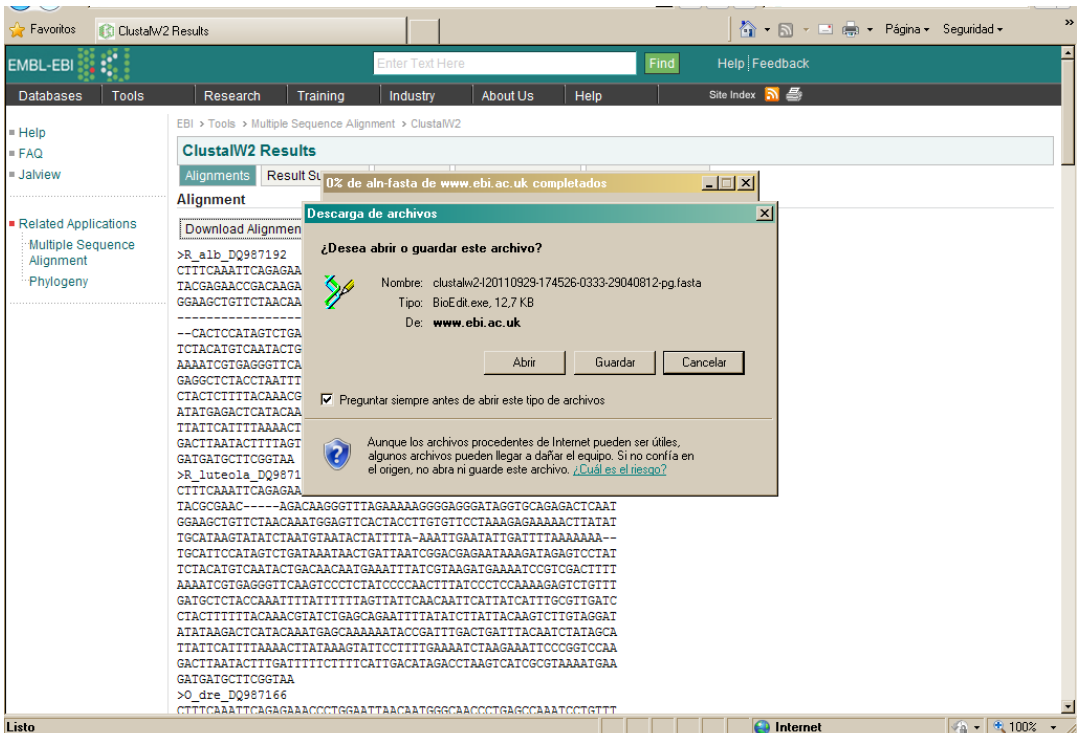
Paso 3. De momento vamos a dejar los parámetros que vienen seleccionados por defecto para los *steps* 2 y 3. Ten en cuenta que en estos pasos es en los que se pueden modificar las penalizaciones que se dan a la apertura y extensión de *gaps*.

Paso 4. En el *step* 3 se pueden seleccionar además opciones de formato del *output* (alineamiento de salida): el tipo de formato en el que se guardará y el orden en el que aparecerán las secuencias. Selecciona fasta y orden como en el *input* (como en el archivo de entrada). Pincha en *submit* para obtener el alineamiento.

Paso 5. El alineamiento automático obtenido aparece en una ventana, desde la cual puede ser descargado:

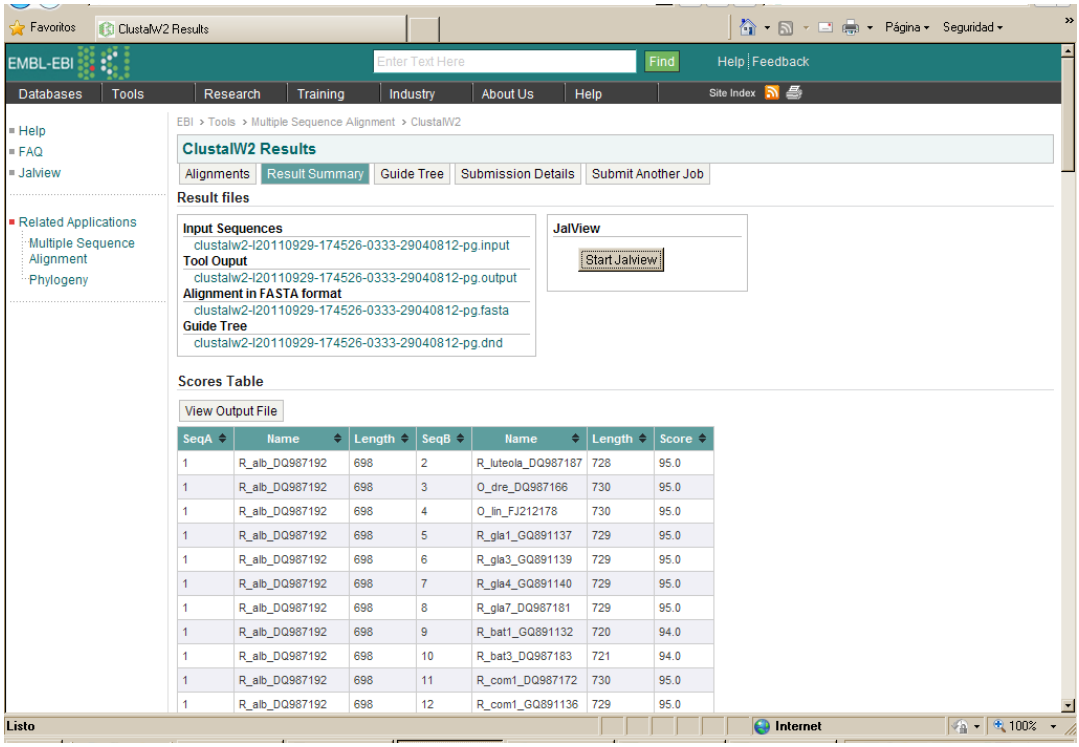


Paso 6. Guarda el alineamiento obtenido pinchando en *Download Alignment File*:

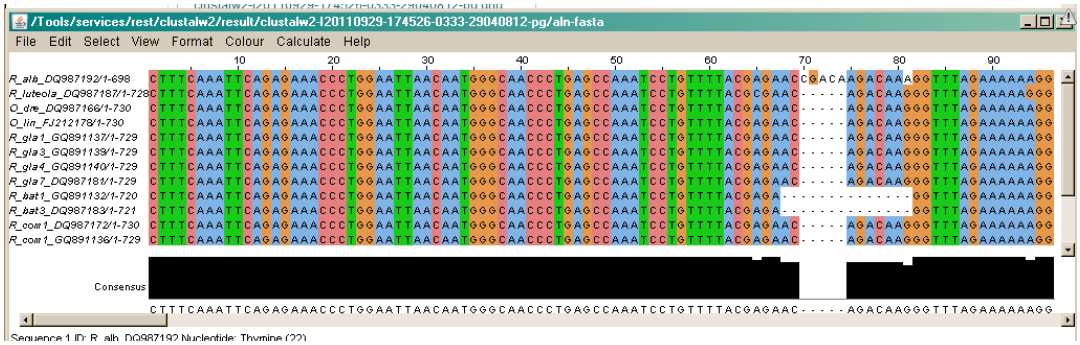


Guarda el archivo con el nombre “Alineamiento_LF_clustal”, para indicar que el alineamiento ha sido obtenido con Clustal y así poder compararlo posteriormente con los alineamientos resultantes de otros programas que utilizaremos.

Paso 7. En la pestaña de *Result Summary* puedes ver una tabla con información sobre la longitud de las secuencias (una vez introducidos los *gaps*), etc.:



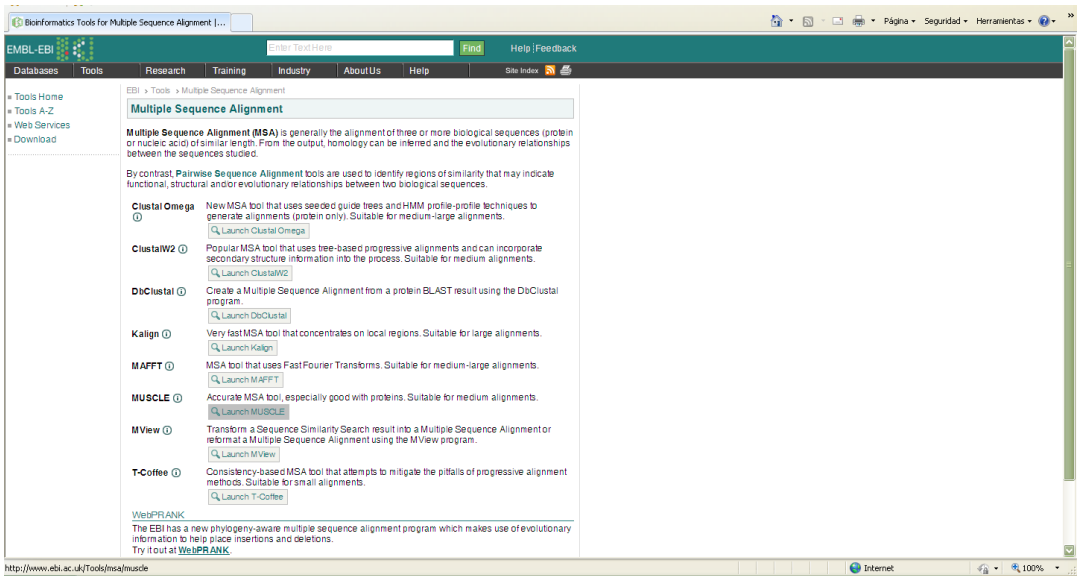
En esa misma pestaña, pinchando en *Start Jalview* puedes visualizar el alineamiento. Esta visualización te permite identificar fácilmente las zonas que están mejor conservadas, las que han sufrido mutaciones, y los puntos en los que es necesario introducir un *gap* porque se ha producido una inserción o una deleción:



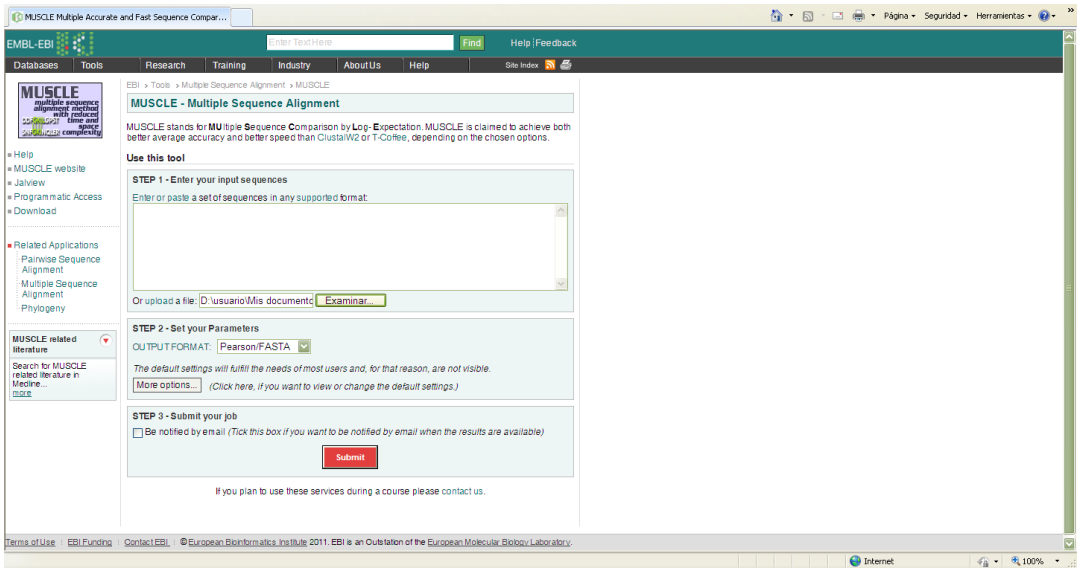
Ejercicio 3.2.1. Repite el proceso para la región del ADN ribosómico nuclear ITS y guarda el alineamiento bajo el nombre “Alineamiento_ITS_clustal”.

II. Alineamiento mediante MUSLCE

Paso 1. Desde la página http://www.ebi.ac.uk/Tools/msa/ abre ClustalW2:

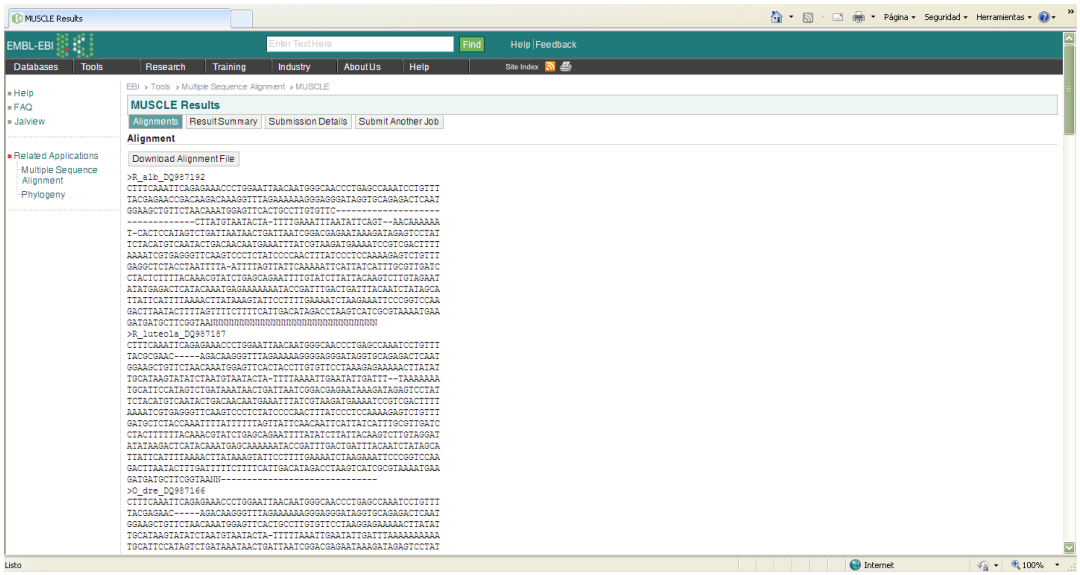


Paso 2. Como en el caso anterior, el primer paso (*step* 1) consiste en cargar el archivo para alinear. Carga de nuevo Secuencias_LF.fas generado desde *GenBank*:

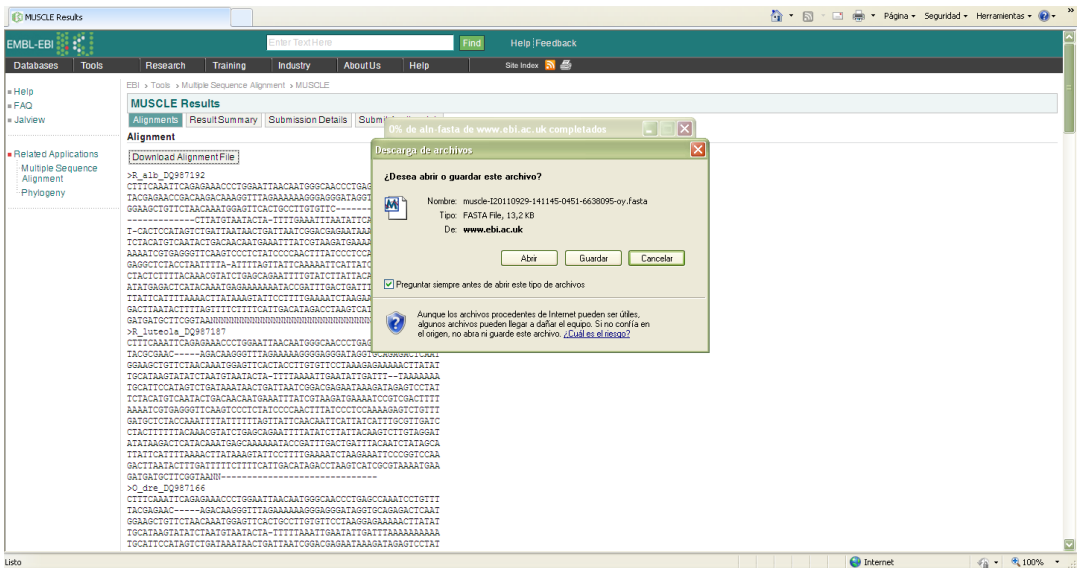


Paso 3. En el *step* 2 se puede seleccionar el tipo de formato en el que se guardará el alineamiento obtenido. Otra vez selecciona fasta y pincha en *submit* para obtener el alineamiento.

Paso 4. Como en el caso anterior, el alineamiento automático obtenido aparece en una ventana, desde la cual puede ser descargado:

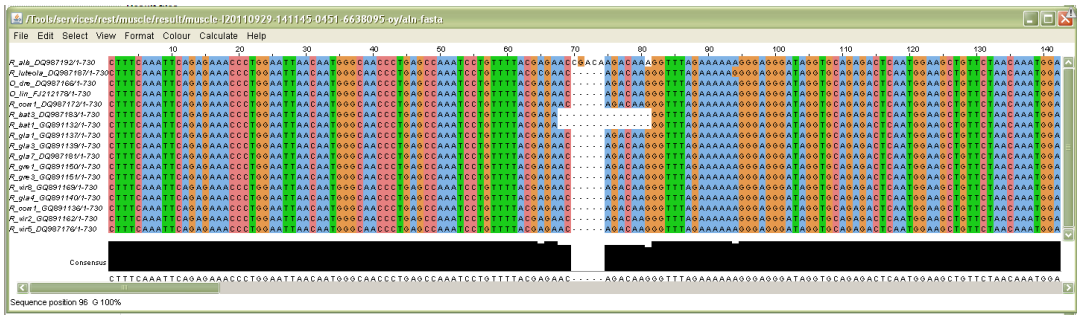


Paso 5. Guarda el alineamiento obtenido pinchando en *Download Alignment File*:



Guarda el alineamiento con el nombre “Alineamiento_LF_muscle”.

Paso 6. En la pestaña de *Result Summary* puedes descargar tres tipos de archivos: el input, que corresponde a las secuencias que introdujimos para analizar y por tanto ya tenemos, el output, donde se pueden ver los datos sobre el alineamiento (penalización que se da a los gaps, etc.), y el alineamiento propiamente dicho, que descarga el mismo archivo fasta que hemos obtenido en el paso anterior. En esta misma pestaña, podemos visualizar el alineamiento pulsando en *Start Jalview*:

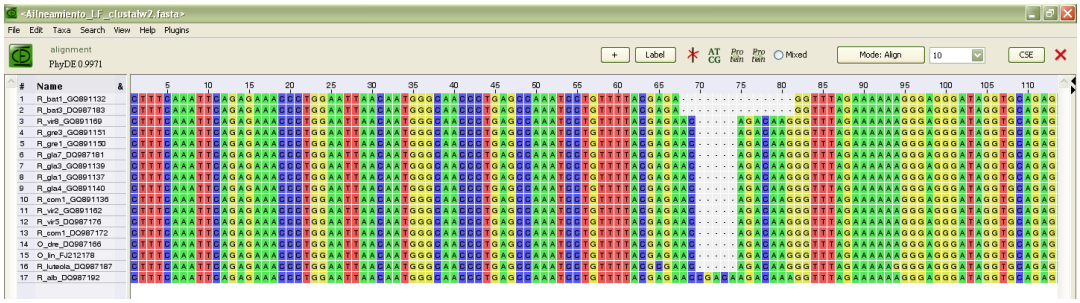


Ejercicio 3.2.2. Repite el proceso para la región del ADN ribosómico nuclear ITS y guarda el alineamiento bajo el nombre “Alineamiento_ITS_muscle”.

III. Revisión de los alineamientos obtenidos mediante alineamiento manual con PhyDE

Paso 1. Abre el programa PhyDE:

Paso 2. Abre uno de los alineamientos obtenidos, por ejemplo el de Alineamiento_LF_Clustal.fas, haciendo click en *File* y *Open*:



PhyDE tiene tres modos de trabajo: *Locked*, *Align* y *Edit*. En el modo locked no se puede modificar nada del alineamiento, pero en el modo align se pueden crear gaps (colocando el cursor donde se quiera y dando al espacio), borrarlos (seleccionándolos y dando a suprimir) y moverlos (seleccionando la parte de la secuencia que se quiera mover y manteniendo pulsado el botón izquierdo del ratón). Además, en el modo edit se puede modificar la secuencia, por ejemplo si se quiere cambiar algún nucleótido tras revisar el cromatograma. Para cambiar de un modo a otro, pulsar en el botón de modo que está a la derecha, en la barra superior de herramientas.

Paso 3. Abre el mismo alineamiento obtenido mediante otro programa, por ejemplo Alineamiento_LF_muscle.fas, haciendo de nuevo click en *File* y *Open*. Puedes visualizar las dos ventanas a la vez seleccionando *View* y *Arrange windows*.

Ejercicio 3.2.3. Compara los dos alineamientos obtenidos y responde a las siguientes preguntas: (a) ¿Qué diferencias observas entre ellos? (b) ¿Cuál de las opciones te parece que tiene mayor sentido biológico en cada caso? (c) ¿Existe algún punto en el alineamiento en el que manualmente hubieras propuesto una solución diferente a la obtenida automáticamente mediante Clustal y MUSCLE? (d) ¿Por qué?

Paso 4. Observarás que MUSCLE, al crear los *gaps*, prolonga las secuencias con Ns al final del alineamiento. Ten en cuenta que el alineamiento definitivo no debe llevar estas Ns y *gaps* al final. Puedes eliminarlos en PhyDE, si seleccionas el modo *edit* y suprimes estas extensiones.

Paso 5. Confirma tus respuestas al ejercicio 3.2.2. con las soluciones proporcionadas en el archivo “ResEj_3.2”. Selecciona el mejor alineamiento y prepáralo para usarlo en las reconstrucciones filogenéticas cortando, si es necesario, los extremos de las secuencias. Guarda este archivo en el que tienes el alineamiento revisado bajo el nombre “Alineamiento_LF_revisado”.

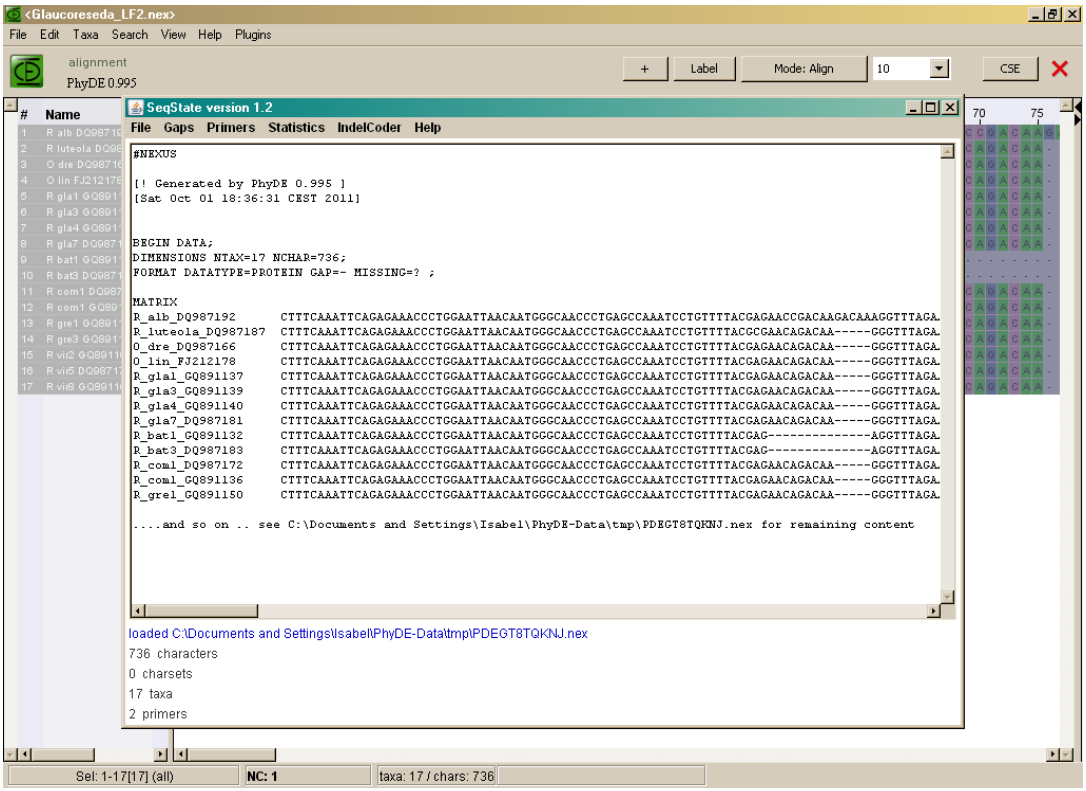
IV. Tratamiento de los gaps mediante SeqState (implementado como Plugin de PhyDE)

Como se ha mencionado en la introducción, los *gaps* producidos como resultado del alineamiento pueden ser utilizados de diferentes maneras a la hora de hacer la reconstrucción filogenética. En la mayor parte de los programas tendremos la opción de indicar si queremos que se ignoren (como *missing data*) o que se traten como un quinto estado. Si lo que optamos es por codificarlos para que cada indel sea considerado como un único evento o cambio, podemos hacerlo de manera manual (incluyendo unas columnas al final de la matriz en las que indicaremos la presencia o ausencia de los gaps), o automáticamente mediante SeqState, uno de los Plugins disponibles para PhyDE. A continuación te proponemos que codifiques los gaps de los alineamientos que has seleccionado como definitivos utilizando este programa:

Paso 1. Abre el alineamiento definitivo de la región *trn*L-F (Alineamiento_LF_revisado) con PhyDE.

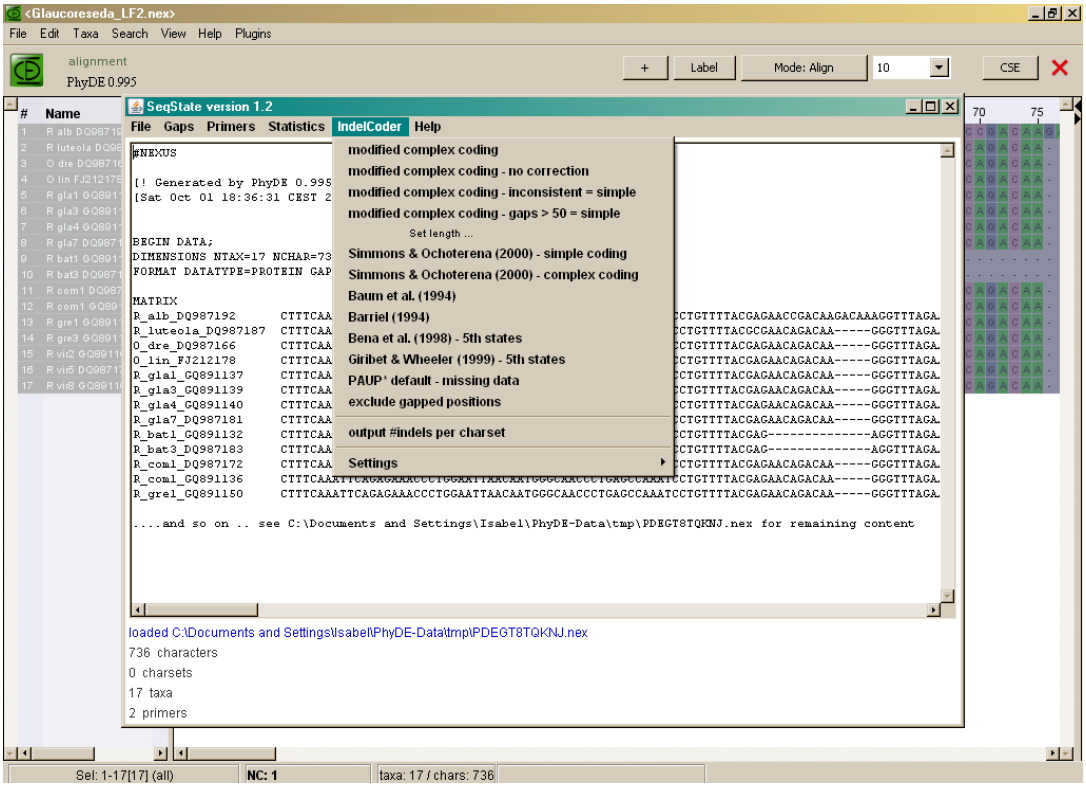
Paso 2. Selecciona todas las secuencias utilizando el ratón o pulsando *Taxa* y *Select all*.

Paso 3. Abre SeqState pulsando *Plugins* y *SeqState*. Se tiene que abrir SeqState en una ventana nueva:



La parte superior de la ventana da información sobre la matriz de datos cargada. En la parte inferior de la ventana iremos viendo información sobre lo que hagamos en SeqState.

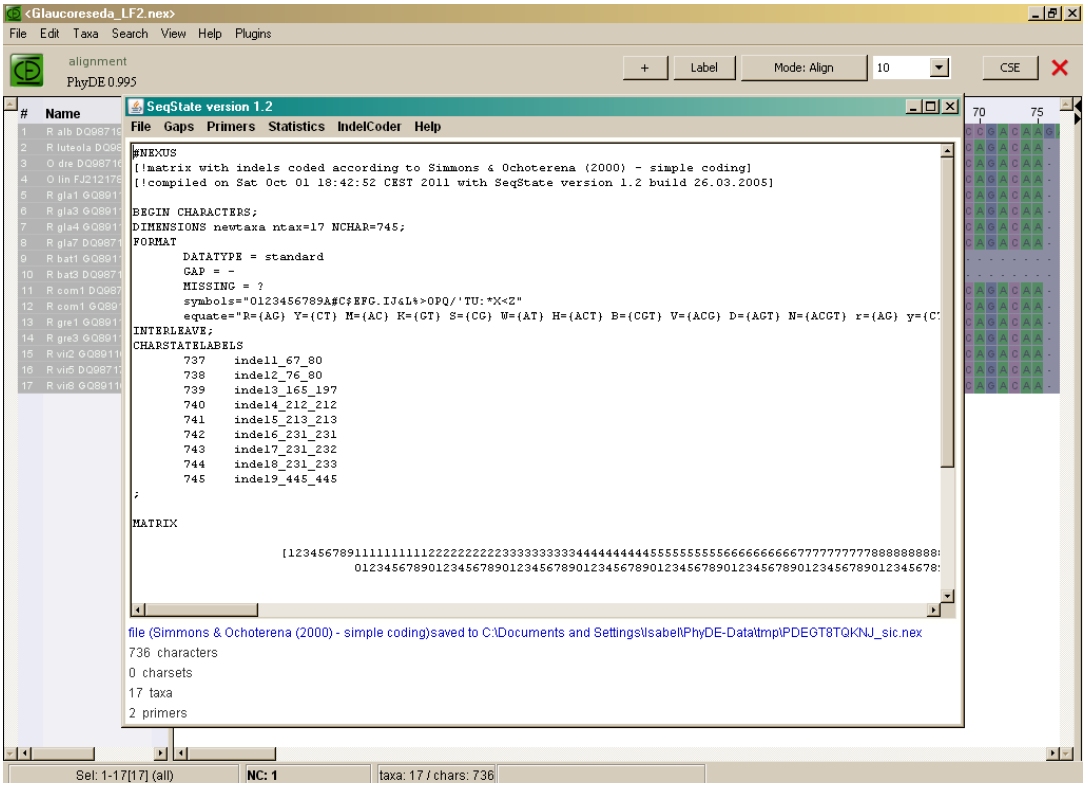
Paso 4. Para codificar los gaps tenemos que utilizar la aplicación *IndelCoder* de la barra superior de herramientas:



Al pulsar en esta opción se despliegan las diferentes opciones para codificar gaps. Recomendamos utilizar la codificación simple de Simmons y Ochoterena (2000).

Paso 5. Cuando pulsamos *IndelCoder* y *Simmons & Ochoterena (2000) - simple coding* el programa crea un archivo nexus con la nueva matriz de datos (que corresponde a la matriz original con unas columnas finales, una por cada gap, presentes -1- o ausentes -0- en cada muestra).

En la ventana superior de SeqState aparecerá una lista en la que se indica el número de carácter que corresponde a cada gap (en el ejemplo inferior los gaps están codificados en las columnas 737 a 745; el primer gap, en la columna 737, corresponde al gap que aparece en la matriz en las posiciones 67 a 80, etc.). En la ventana inferior de SeqState se indica dónde se ha guardado esta nueva matriz (normalmente en la carpeta tmp de PhyDE-Data), y con qué nombre (en el ejemplo inferior el archivo generado se llama PDEGT8TQNJ_sic.nex):



Paso 6. Busca el archivo nexus generado con los gaps codificados en tu ordenador, cámbiale el nombre por uno que te permita reconocer el contenido del archivo (por ejemplo “Glaucoreseda_LF_gaps” y guárdalo en la carpeta en la que tengas el resto de las matrices generadas. Luego puedes abrirlo con un visor de alineamientos (por ejemplo con PhyDE) para ver cómo ha quedado la nueva matriz.

V. Combinación de matrices (con PhyDE)

Cuando las diferentes regiones de ADN estudiadas son congruentes, puede ser de utilidad combinarlas para crear una única matriz con más información para la reconstrucción filogenética. Existen diferentes maneras de comprobar si las regiones del ADN analizadas son congruentes entre sí y pueden ser combinadas. En general, lo más seguro es estudiar primero las regiones por separado. Si se observa alguna incongruencia, se recomienda utilizar algún test de congruencia para evaluar si la información filogenética proporcionada por ambas regiones es significativamente diferente, en cuyo caso de desaconseja la combinación de ambas regiones. La manera de proceder sería, por tanto:

1) Elaborar una aproximación filogenética para cada región de ADN (en nuestro caso, una para ITS y otra para *trn*LF)

2) Inspeccionar visualmente si las reconstrucciones obtenidas son congruentes (puede ocurrir que una región resuelva una parte que el otro no resuelve, pero no podríamos combinar las regiones si cada región propone una solución diferente y con buen soporte para un mismo clado)

3) En el caso de que las reconstrucciones filogenéticas obtenidas fueran congruentes, elaboraríamos una matriz combinada que utilizaríamos para la reconstrucción filogenética definitiva

A continuación te explicamos cómo podrías hacer esta matriz combinada, una vez comprobada la compatibilidad de las regiones objeto de estudio.

Una posibilidad, que desaconsejamos, es modificar a mano los archivos utilizando un editor de texto: abriríamos los dos archivos, seleccionaríamos los bloques de letras que corresponden a cada secuencia y elaboraríamos un nuevo archivo en el que cada muestra estuviera caracterizada por un bloque de letras constituido por ambas regiones, copiando y pegando las secuencias una a continuación de la otra. Ojo, porque si utilizas esta opción, en los archivos nexus y phylip, además de modificar las secuencias que caracterizan cada muestra, tienes que modificar la línea de comandos en la que se indica el número de caracteres de cada secuencia.

La segunda posibilidad, que te recomendamos utilices, es combinar las matrices utilizando un editor de secuencias como PhyDE. Te proponemos que combines las matrices de ITS y *trn*L-F de *Glaucoreseda*:

Paso 1. Abre uno de los archivos de alineamiento definitivo, por ejemplo, “Alineamiento_LF_revisado”.

Paso 2. Abre el otro archivo de alineamiento definitivo (en este ejemplo “Alineamiento_ITS_revisado”).

Paso 3. Pulsa *View* y *Arrange windows* para poder ver las dos matrices a la vez.

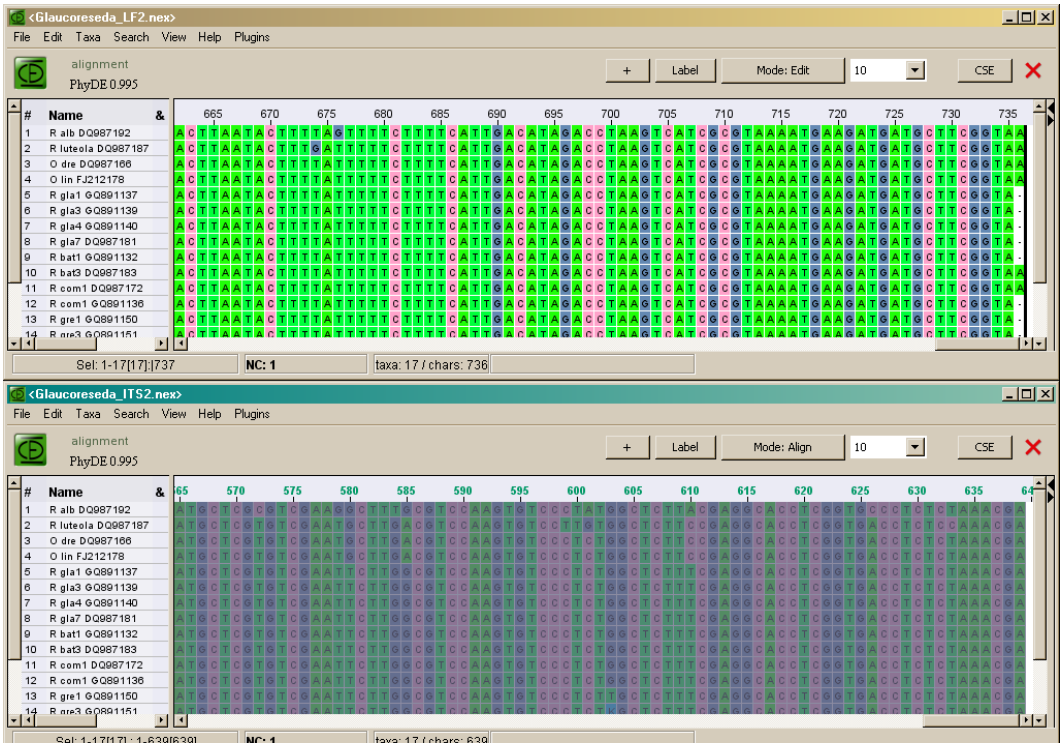
Paso 4. Comprueba que todas las secuencias aparecen en el mismo orden en las dos matrices.

Paso 5. En una de las ventanas, por ejemplo en Alineamiento_LF_revisado, pulsa *Mode: edit*.

Paso 6. En la otra ventana (en este caso Alineamiento_ITS_revisado), selecciona todas las secuencias (no sus nombres) utilizando el ratón sin soltar el botón izquierdo o mediante el uso combinado de la tecla de mayúsculas y las flechas.

Paso 7. Copia las secuencias seleccionadas pulsando *Edit* y *Copy* (o con ctrl+c).

Paso 8. En la primera ventana (Alineamiento_LF_revisado) coloca el cursor al final de la matriz (de todas las secuencias) pinchando con el ratón sin soltar el botón izquierdo o pulsando a la vez la tecla de mayúsculas y la flecha hacia abajo:



Paso 9. Pega las secuencias de ITS a continuación de las de *trn*L-F, pulsando *Edit* y *Paste* (o con ctrl+v). Como resultado, obtienes una matriz con los 1375 caracteres que resultan de unir los 639 de la matriz final de ITS y los 737 de la matriz final de *trn*L-F. Guarda esta nueva matriz con un nombre que la identifique como por ejemplo “Glaucoreseda_ITS_LF” en formato nexus.

VI Conversión de formatos

Como se ha mencionado, existen diferentes tipos de formatos para los archivos de alineamientos, y según el programa de reconstrucción filogenética que vayamos a utilizar necesitaremos tener el alineamiento en un tipo u otro de formato. Como en el caso de la combinación de matrices, la conversión de los archivos de un formato a otro se puede hacer de manera manual o utilizando algún programa informático.

De manera manual, se debe abrir el archivo que se quiera convertir en un editor de texto (WordPad para PC, o TextWrangler para MAC; OJO es muy importante que no lo abráis con Word), y modificar las líneas de comando iniciales según se indicó en la introducción (por ejemplo, si se quiere pasar de fasta a nexus habrá que eliminar el símbolo > del principio de los nombres de las secuencias y habrá que añadir el bloque inicial previo para indicar el tipo de formato, las dimensiones de la matriz, etc.).

Para convertir los formatos automáticamente también se puede utilizar PhyDE (para fasta o nexus):

1. Abre el archivo que quieras convertir en PhyDE.
2. Pulsa *File* y *Export as*
3. Selecciona el tipo de formato al que quieres convertir la matriz.

4. Indica el nombre con el quieres guardar la matriz y el lugar en el ordenador y pulsa *Export*.

Además existen otros conversores en Internet, con los que también puedes manejar otros formatos como Phylip. Algunas direcciones útiles son:

http://hcv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html

http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html