

Introducción a la Filoinformática:
IBBM-UNLP/CONYCET, Argentina. 2-6 Julio 2018.

Pablo Vinuesa (vinuesa@ccg.unam.mx)

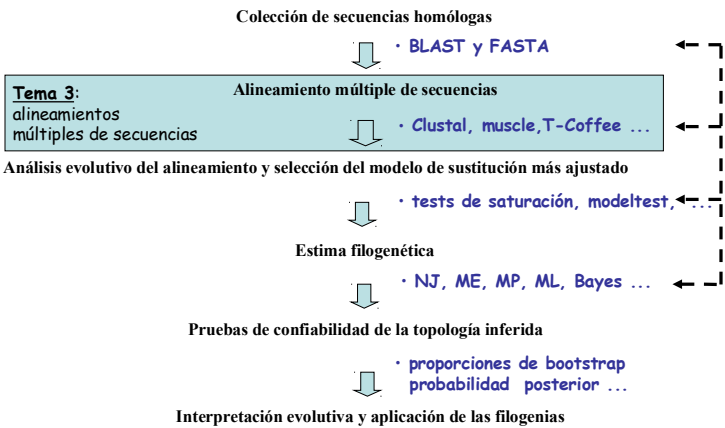
Programa de Ingeniería Genómica, CCG-UNAM, México
<http://www.ccg.unam.mx/~vinuesa/>

Todo el material del curso (presentaciones, tutoriales y datos de secuencias) lo encontrarás en:
<https://github.com/vinuesa/intro2phyloinfo>

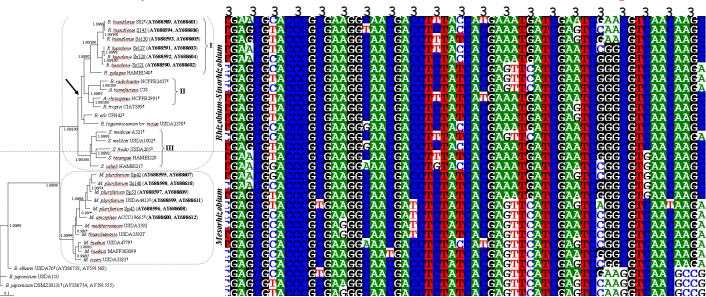
Tema 3: Alineamientos múltiples

- 1. Alineamientos múltiples y el problema de las repeticiones, sustituciones e indeles
- 2. Alineamientos múltiples progresivos usando programas de la familia Clustal
- 3. Scripts de Perl para automatizar procesos: alinear muchos archivos y hacer interconversiones de formatos de secuencia sobre múltiples archivos.
- 4. Formatos de secuencia
- 5. Alineamiento de secuencias codificadoras de proteínas usando RevTrans
- 6. Alineamiento de genes ribosomales usando RDP-II y GreenGenes
- 7. Aln. múltiples usando muscle

Protocolo básico para un análisis filogenético de secuencias moleculares



- Cualquier estudio de filogenético o de evolución molecular basado en secuencias necesita de un **alineamiento múltiple para determinar las correspondencias de homología a nivel de los residuos individuales o caracteres.**
- **El alineamiento representa una hipótesis sobre la homología de los caracteres**
- La mejor manera de representar estas homología entre caract. es escribiendo los residuos homólogos en columnas, generándose una matriz de m x n (secs. x posic) residuos, en la que **cada columna contiene a residuos o caracteres homólogos**

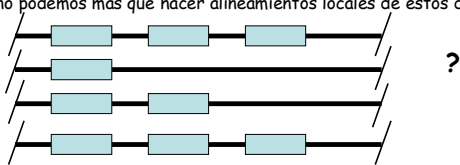


- Comparar los aln. múltiples en el contexto de una filogenia nos puede revelar mucho acerca de los patrones y tasas de sustitución, haciendo posible la identificación de estados de carácter ancestrales y derivados, etc.

Generación de alineamientos múltiples - consideraciones generales

El problema de las repeticiones

Muchas proteínas multidominio pueden presentar diverso grado de **repetición de dominios** particulares. Puede llegar a ser muy complejo o imposible hacer el alineamiento global de las proteínas si difieren en el número y orientación de estas regiones repetidas. A veces no podemos más que hacer alineamientos locales de estos dominios.

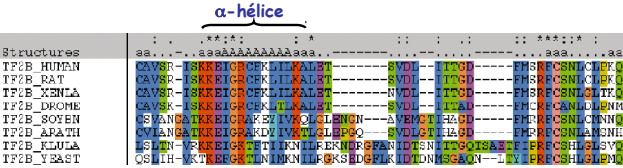


A nivel de DNA se dan también regiones repetidas, muchas veces involucrando a unos pocos nts, como es el caso de los **microsatélites y otras regiones repetidas**. Con frecuencia estas regiones son imposibles de alinear objetivamente. Suelen acumularse en regiones no codificantes del genoma, incluyendo intrones, o en regiones codificantes hipervariables como espaciadores intergénicos transcritos o regiones reguladoras (UTRs).

Este tipo de "repeats" cortos son poco frecuentes a nivel de aminoácidos, si bien a este nivel es común encontrar regiones o dominios **"de gran escala" repetidos**. Un ejemplo clásico de este fenómeno son las calmodulinas.

Generación de alineamientos múltiples - consideraciones generales

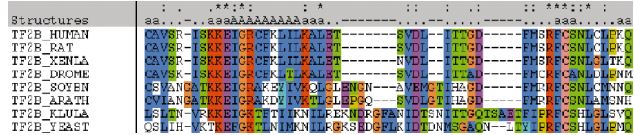
- El problema de las sustituciones
- Al examinar alns. múltiples de proteínas se observan dos patrones de sustitución:
 - Existen bloques de 5 a 20 residuos con alto nivel de identidad y similitud dispersos entre regiones de menor similitud. Estos bloques corresponden típicamente a elementos estructurales como **α-hélices** y **pliegues beta** que evolucionan más lentamente que los loops o bucles que los interconectan



- Las **columnas alineadas** con múltiples estados de carácter **tienden a presentar residuos de características bioquímicas similares** (I, A, V, L; S, T; R, K; etc.). Esta conservación de residuos similares es particularmente patente en los bloques correspondientes a elementos de estructura secundaria, sitios activos o de unión a ligandos. La propiedad bioquímica más conservada es la de polaridad/hidrofobicidad.

Generación de alineamientos múltiples - consideraciones generales

- El problema de los indeles (inserciones/deleciones)
- Cuando por eventos de inserción o deleción (**indeles**) las secuencias homólogas presentan distintas longitudes, es necesario introducir **"gaps"** en el alineamiento para mantener la correspondencia entre sitios homólogos situados antes y después de las regiones afectadas por indeles. Estas regiones se identifican mediante guiones (-).

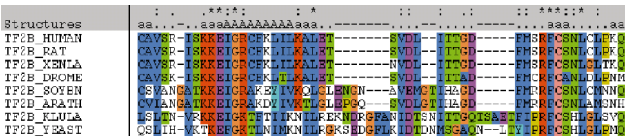


Los indeles no se distribuyen aleatoriamente en las secuencias codificadoras. Casi siempre aparecen **ubicados entre dominios funcionales o estructurales**, preferentemente en bucles (loops) que conectan a dichos dominios. Esto vale tanto para RNAs estructurales (tRNAs y rRNAs) como para proteínas. No suelen interrumpir el marco de lectura.

- Generalmente se usan **sistemas de penalización de gaps afines**: $GP = gap + e(L - 1)$ en el cálculo del score de un alineamiento múltiple [gap = costo apertura de gap; e = costo extensión del indel; L longitud del indel]

Generación de alineamientos múltiples - consideraciones generales

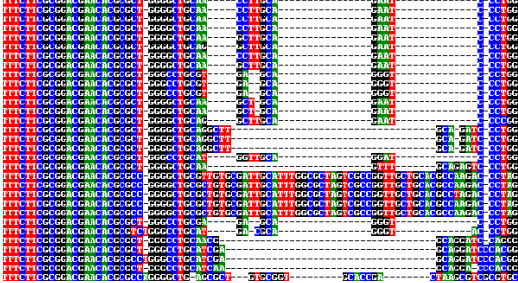
- El problema de las sustituciones
- Es importante recordar que por debajo del **20% de identidad** a nivel de sec. de AA es ya imposible que se pueda obtener un alineamiento múltiple (o pareado) confiable si nos basamos para obtenerlo sólo en la secuencia primaria, ya que entramos en la **zona de penumbra** (saturación mutacional)



- Un par de secuencias de nts al azar presentarán en promedio un 25 % de dentidad.
- Por tanto, siempre que sea posible, hay que realizar los alineamientos múltiples en base a las secuencias traducidas, es decir, sobre AAs (igual que al hacer búsquedas en bases de datos de secuencia), que son mucho más informativos (20 caracteres vs. 4) y con tasa evolutiva mucho menor.

Generación de alineamientos múltiples - consideraciones generales

- A mayor **distancia genética** (evolutiva) entre un par de secuencias, mayor será el número de mutaciones acumuladas. Dependiendo del tiempo de separación de los linajes y la tasa evolutiva del locus, puede llegar a ser imposible alinear ciertas regiones debido a fenómenos de **saturación mutacional** o **acúmulo de indeles**. En loci de evolución muy rápida como intrones o espaciadores intergénicos, los fenómenos de saturación mutacional se observan incluso cuando se comparan secuencias de organismos evolutivamente próximos (mismo género o familia).



Las regiones de homología dudosa deben de ser excluidas de un análisis filogenético! Debemos de maximizar a toda costa la relación entre señal/ruído

Alineamientos múltiples - algoritmos

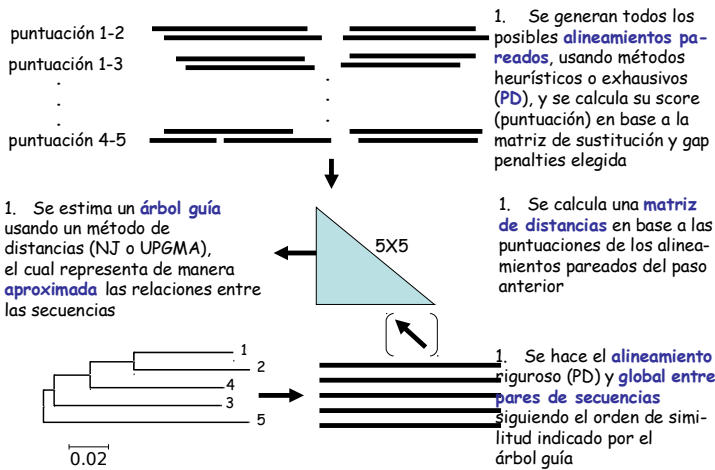
Existen diversas estrategias computacionales para obtener alineamientos múltiples de manera (semi)automática para conjuntos grandes (cientos - miles) de secuencias.

1.- Implementación de algoritmos de alineamiento progresivo.

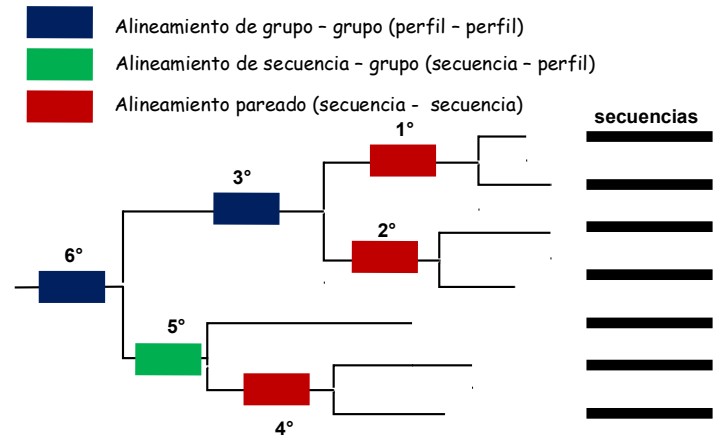
Así como los alns. múltiples son indispensables para reconstruir filogenias a partir de secs, un árbol de relaciones filogenéticas representa información muy valiosa para guiar la generación de un aln. múltiple.

La mayor parte de los alineadores automáticos modernos se basan en este tipo de algoritmos. Construyen un árbol guía aproximado a partir de distancias calculadas entre todos los pares posibles de secuencias. De la matriz de distancias resultantes se construye un árbol usando un método algorítmico (NJ o UPGMA). El árbol guía resultante se emplea para construir el alineamiento de manera progresiva. Las dos secuencias más similares se alinean primero usando PD y una matriz o esquema de ponderación particular. Una vez alineado el primer par, los gaps generados ya no se mueven. Este par es tratado como una sola secuencia y es alineada contra la siguiente secuencia o grupo de secuencias más próximas en el árbol. Se repite el proceso hasta que todas las secs. están alineadas. El proceso es suficientemente rápido como para alinear varios cientos de secuencias. Son menos precisos que los métodos basados en la WSPs, pero muchísimo más rápidos.

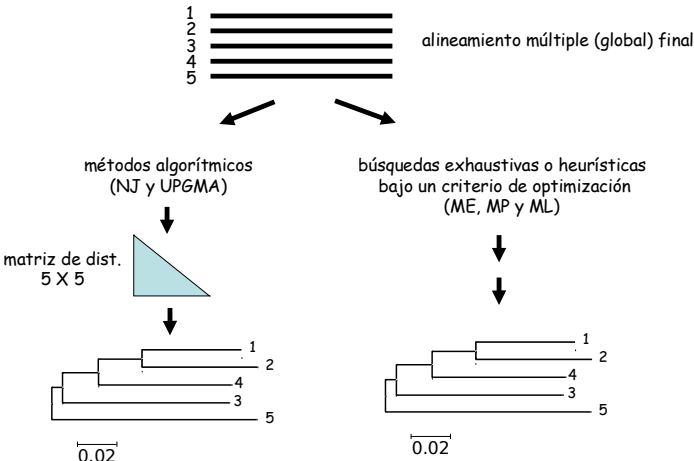
Pasos en la generación de un alineamiento múltiple siguiendo la estrategia de alineamiento progresivo



Pasos en la generación de un alineamiento múltiple siguiendo la estrategia de alineamiento progresivo con un árbol guía

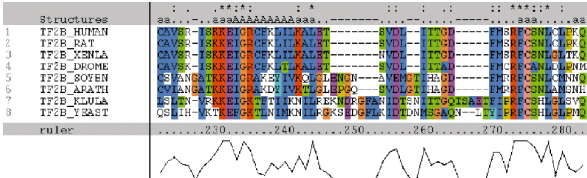


Pasos en la generación de un alineamiento múltiple siguiendo la estrategia de alineamiento progresivo - y su uso para estimar una filogenia

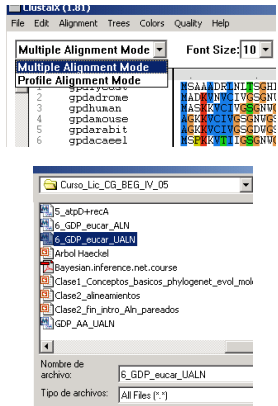


Alineamientos múltiples progresivos usando Clustal

- La familia Clustal es posiblemente la más popular para hacer AMs de nt y aa
- Existen versiones para todas las plataformas y en red (<http://www.ebi.ac.uk.clustalw>)
- La primera versión (Clustal) salió en 1988, la última, **ClustalX**, en 2007 (última Vers. = 2.0)
- **ClustalX (X-windows Clustal)** lee secuencias en diversos formatos, calcula un **árbol guía NJ**, usando algoritmos heurísticos o exhaustivos sobre aln. locales basado en **distintas matrices de pesado y de penalización de gaps afines y sitio-específicos**. Puede hacer **alineamientos de perfiles** y existen diversas **herramientas de control de calidad del AM**. Permite incluir criterios estructurales para guiar el AM, usando **máscaras estructurales**. Partes del alineamiento o secuencias particulares pueden ser **realineadas** para ir obteniendo un aln global cada vez mejor. Es decir, ClustalX no sólo genera alineamientos (como ClustalW), sino que éstos pueden ser editados y mejorados interactivamente por el usuario. Además, ClustalX (y ClustalW) permite la **reconstrucción y visualización de árboles NJ** y hacer **análisis de bootstrap** sobre los alineamientos. Finalmente, los AMs pueden ser escritos en **diversos formatos de salida** (CLUSTAL, FASTA, NEXUS, PHYLIP ...)



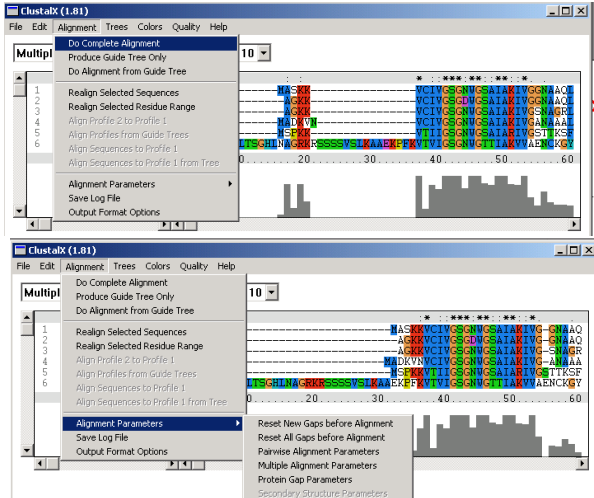
Alineamientos múltiples progresivos usando Clustal
-un ejemplo: alineamiento de GDPs dependientes de NAD



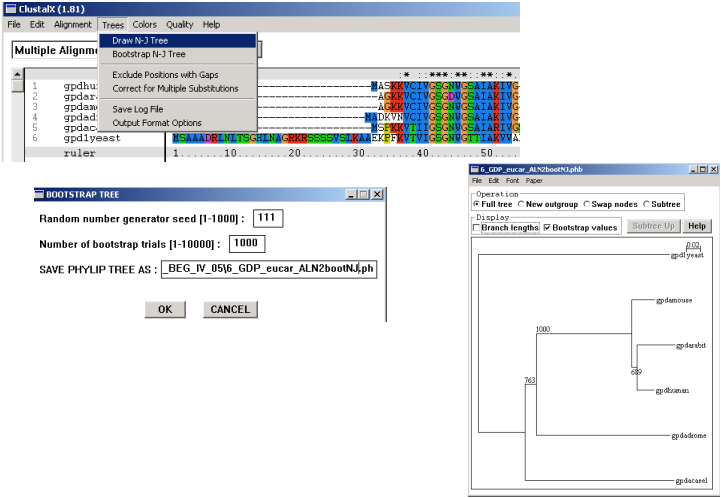
- 1.- Selecciona modo de aln y fichero a alinear (en este caso las secs. están escritas en formato FASTA)

```
>gpdlyeast
MSAAADRLNLTSGHLNAGKRKSSSSVSLKAAEKPFVKVTVIGSGNWGTTIA
KVVAENCKGYPEVFAPVQMWWFEEENGINEKLTETINTRHQNVKYLPGIT
LFDNLVANFDLSDVKVDVITVFNIHQFELFRICQSLKQGHVDSHVRASC
LKGFEVGAQGVQLLSYITTEELGQCGALSGANIEVAQEHSETTVAY
HIPKDFRSGKVDHVKLALHHPVHFVSVTEVAGISICSAKRWVAL
GCGFVGLGWNNASAAIQRVGLGEIIRFGQMFPESERETTYQESAGVA
DLITTCAGGRNVKVARLMATSGKDANECEKLLNGQSAQGLITCKEVHEW
LTCGSDVEDFLFVAVYQIVYNNYPMKNLPMIEELDLHED
>gpdadrome
MADKVVIVIGSGNWGSAIAKIVGANAAALPEFEERVTMTFVEEELIDGKK
LTEIINETHENVKYLKGHLKLPNAVDPDLVEAKNADILIFVVPHQFIPN
FCQQLLKGKIFNAITSLIKGFDKAEQGGIDLIHSHITRIPCAVLMGANL
ANEVAGHFCETTIGCTOKKYGKVLRLDLFQAHFRVYVVDADAVEGCGA
LNMIVCGAGFVDSKLDGNTKAVATRLGLMEIRLVQVYQSGKLSFFF
ESCGVADLITTVRVSEAFVTGKTEBLEKMLNVQKLQGPPTAEVNY
```

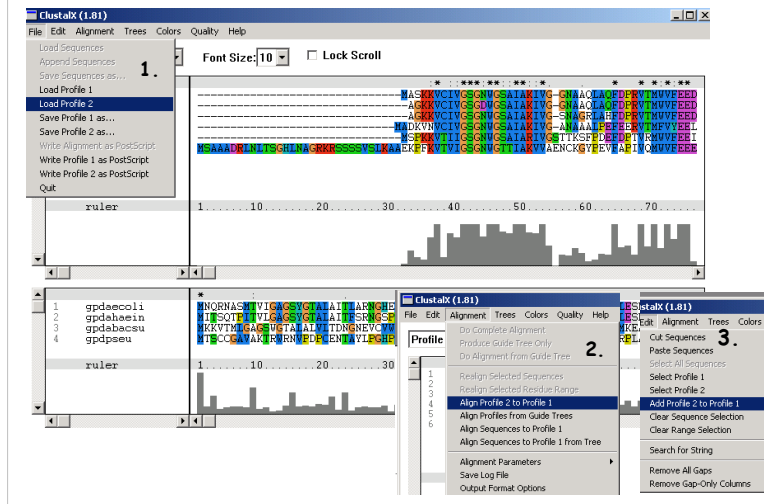
Alineamientos múltiples progresivos usando Clustal
-un ejemplo: alineamiento de GDPs dependientes de NAD



Alineamientos múltiples progresivos usando Clustal
-un ejemplo: alineamiento de GDPs dependientes de NAD

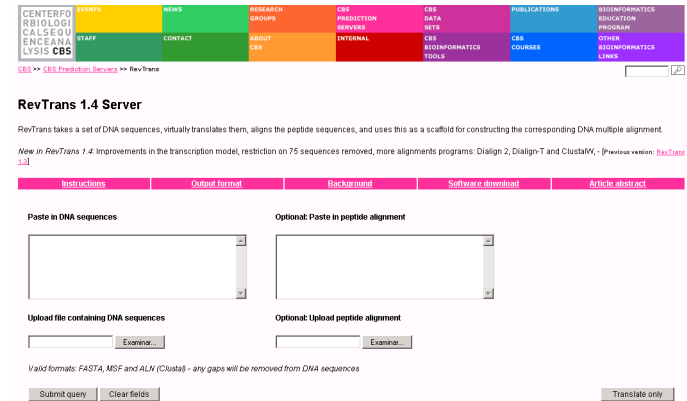


Alineamientos múltiples progresivos usando Clustal
-un ejemplo: alineamiento de GDPs dependientes de NAD



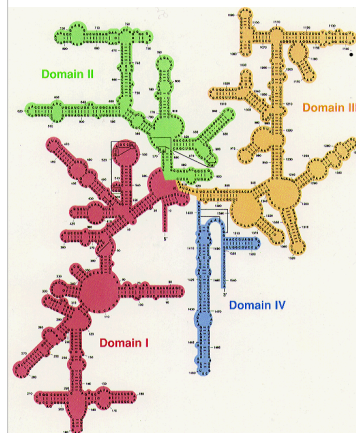
Servidores para alinear nts. en base a un alineamiento de proteínas

<http://www.cbs.dtu.dk/services/RevTrans/>



Servidores para alinear secuencias de rRNAs o rDNAs

- Los genes ribosomales representan un problema muy particular en el contexto de alineamientos múltiples. Deben de guiarse usando máscaras de información estructural.



Servidores como **GreenGenes** y **RDP-II** proveen herramientas muy útiles en este contexto.

Si quieres ver unos tutoriales sobre el uso de estos servidores, visita mi sitio web y busca bajo phylogeny tutorials:

<http://www.ccq.unam.mx/~vinuesa/>

Using_the_GreenGenes_and_RDPII_servers.html

Formatos de secuencias

I) FASTA

- Existen una gran cantidad de estilos o formatos de presentación de secuencias. Muchos programas de análisis filogenético usan su propio formato (Phylip, Nexus, Mega ...)
- El formato más sencillo es el **FASTA**, en el que cada secuencia se identifica mediante un renglón descriptor que comienza con > en el siguiente renglón comienza la secuencia

>|c|1 R_galegae
CCGCTGGTCACTCCGGCAAGCGGCCATCCACCAGGAAGCGCCTTCTTA
CGTCGATCAGTCACCGAAGGCCAGATCTGGTCAACGGCATCAAGGTCG

> lcl|2 M._plurifarium
CCGGTCGACGCCGTGAGCTGCGTGCCATCCACCAGCCGGCTCCGGCCTA
TGTCGACCAGTCGACGGAAGCGCAGATCCTGGTTACCGGCATCAAGGTTTC

> lcl|3 B._japonicum
CCGGTCAAGTCGGAAGGCCCTGCGGCCATCCACCAGGAAGCGCCGACCTA
CACCGACCAGTCCACCGAAGCTGAAATTCTCGTACCGGCATCAAGGTCG

Formatos de secuencias
II) PHYLIP

- **Phylip (interleaved):** no. seqs, no. caracteres
nombre secuencias (máx 10 caracteres) espacio, secuencia ...

```
3      100
R._galegae CCGCUGGUCA CCUCCGGCAA GCGCGCCAUC CACCAGGAAG CGCCUCCUA
M._plurifa ...G.C.A.G ..GU..AGCU ...U..... CCG. .U..GG....
B._japonic ...G.CAAGU .GGAA...CU ..... GA....

CGUCGAUCAG UCGACCGAAG GCCAGAUCU GGUCACCGGC AUCAAGGUCC
U.....C... ..G.... CG..... ..U..... UC
.AC...C... ..C..... CUG.A..U.. C.....
```

- **Phylip (sequential or non-interleaved)**

```
3 100
R._galegae CCGCTGGTCA CCTCCGGCAA GCGCGCCATC CACCAGGAAG CGCCTTCCTA
CGTCGATCAG TCGACCGAAG GCCAGATCCT GGTCACCGGC ATCAAGGTCG
M._plurifa CCGGTCGACG CCGTCGAGCT GCGTGCCATC CACCAGCCGG CTCGGGCTTA
TGTCGACCAG TCGACGGAAG CGCAGATCCT GGTACCGGC ATCAAGGTTT
B._japonic CCGGTCAAAGT CGGAAGGCCT GCGCGCCATC CACCAGGAAG CGCCGACCTA
CACCGACCAG TCCACCGAAG CTGAAATTCT CGTCACCGGC ATCAAGGTCG
```

Formatos de secuencias
III) NEXUS

```
#NEXUS
[OJO!!!, no usar guiones- (reservado para gaps!), sólo guiones bajos_]

BEGIN TAXA;          [taxa block]
DIMENSIONS NTAX=3;
TAXLABELS
R._galegae;
M._plurifarium;
B._japonicum;
END;

BEGIN CHARACTERS;    [character block]
DIMENSIONS NCHAR=100;
FORMAT DATATYPE=DNA MISSING=? GAP=- MATCHCHAR=. INTERLEAVE=yes ;
MATRIX
[
      10      20      30      40      50]
[
      *      *      *      *      *]
R._galegae CCGCTGGTCACCTCCGGCAAGCGGCCATCCACCAGGAAGCGCCTTCCTA
M._plurifarium ...G.C.A.G..GT..AGCT...T.....CCG..T..GG....
B._japonicum ...G.CAAGT.GGAA...CT.....GA....

[
      60      70      80      90      100]
[
      *      *      *      *      *]
R._galegae CGTCGATCAGTCGACCGAAGGCCAGATCCTGGTCACCGGCATCAAGGTCG
M._plurifarium T....C.....G....CG.....T.....TC
B._japonicum .AC...C.....C.....CTG.A..T..C.....
;
END;
```

Formatos de secuencias:
su interconversión

- Cuando preparamos un fichero con nuestras propias secuencias generalmente lo más adecuado es hacerlo en formato FASTA
- Si necesitamos pasarlo a otro formato, una buena posibilidad es hacerlo con **ReadSeq**

<http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi>

ReadSeq reconoce automáticamente el formato de entrada y si se trata de aas o nts

- Otra alternativa es escribir un sencillo script de **Perl** que haga uso de los objetos y métodos del módulo Bio::AlignIO de **BioPerl** (<http://www.bioperl.org>) para interconvertir Formatos ... veremos un ejemplo más adelante.

Alineamientos múltiples progresivos usando Clustalw

Desde el directorio en el que están las secuencias llamar a ClustalW

```
vinuesa@teide:~/cd Cursos/BGEIV-06/
vinuesa@teide:~/Cursos/BGEIV-06> ls
ClustalW_cmmds.txt myoglobins.fasta
vinuesa@teide:~/Cursos/BGEIV-06> clustalw
```

```
*****
***** CLUSTAL W (1.83) Multiple Sequence Alignments *****
*****
```

1. Sequence Input From Disc
2. Multiple Alignments
3. Profile / Structure Alignments
4. Phylogenetic trees

5. Execute a system command
- H. HELP
- X. EXIT (leave program)

Your choice:1

Alineamientos múltiples progresivos usando Clustalw

¿No se ve nada divertido, verdad?, ¿pero qué tal esto?:

```
clustalw -infile=myoglobins.fasta -align -pwmatrix=blosum -pwgapopen=12  
-pwgapext=0.2 -matrix=blosum -gapopen=12 -gapext=0.2 -outorder=aligned  
-convert -outfile=myoglobins_aln1.phy -output=phylip
```

- Y si tenemos muchos archivos fasta de proteína para alinear (extensión .faa), podemos escribir una simple línea de shell como la siguiente para alinearlos todos consecutivamente usando los parámetros por defecto de clustalw:

```
for file in *.faa; do clustalw -infile=$file -align -convert -output=fasta; done
```

Las opciones de clustal en línea de comando se ven así:

```
clustalw -options
```

y la ayuda así:

```
clustalw -help # puede no funcionar si el documento de ayuda no está en el path
```