Alineamientos Múltiples de Secuencias

Luis E. Garreta U

Curso de Bioinformática
Pontificia Universidad Javeriana – Cali
Facultad de Ingeniería - Carrera de Biología

17 de septiembre de 2018

Introducción

- Hemos visto cómo comparar una secuencia con otra (alineamiento de pares)
- Hemos visto cómo comparar una secuencia con muchas otras en una BD (muchos alineamientos de pares - BLAST)
- Ahora veremos cómo comparar múltiples secuencias simultáneamente, no de dos en dos.

Motivación

- Las secuencias biológicas a menudo se agrupan en familias
  - ✓ Genes relacionados de un organismo (parálogos)
  - ✓ Genes relacionados de distintas especies (ortólogos)
  - ✓ Secuencias dentro de una población (variantes polimórficas)
- Dos secuencias pueden tener un alineamiento no muy bueno entre ellas, pero pueden alinearse vía una tercera
  - ✓ Identificación de familias y regiones conservadas

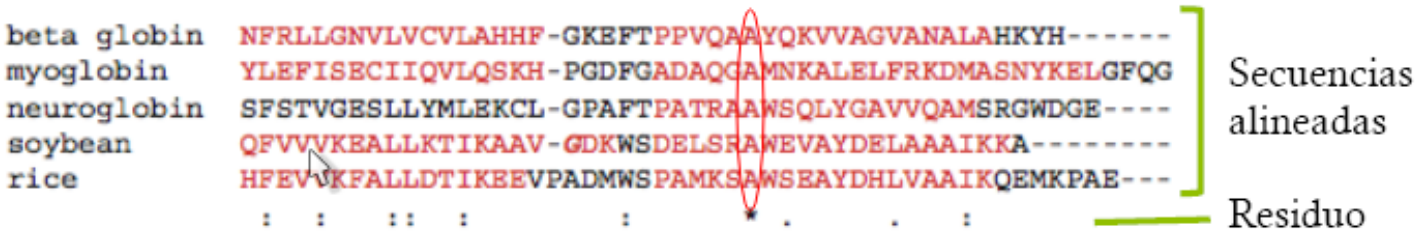
Alineamientos Múltiples de Secuencias

- Un alineamiento múltiple de secuencias es un alineamiento de más de dos secuencias.
- Estas secuencias, como en el caso de los alinenamientos por parejas pueden ser ADN, ARN o proteína.

```
Hsa_TMEM66      ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI 103
Ptr_TMEM66      ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI 103
Ppy_TMEM66      ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI 103
Mml_TMEM66      ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI 103
Mfa_TMEM66      ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI 103
Mne_TMEM66      ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI 103
Ssc_TMEM66      ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDV 103
Bta_TMEM66      ALTLYYDRYTTSRRLEPIPQLKCVGGTAGCDSYTPKVIQCQNRGWDGYDVQWECKTDLDV 103
Cfa_TMEM66      ALTLHHDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI 103
Mmu_TMEM66      ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCEAYTPRVIQCQNKGWDGYDVQWECKTDLDI 104
Rno_TMEM66      ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCDAYTPKVVQCQNKGWDGYDVQWECKTDLDI 104
Ocu_TMEM66      ALTLHYDRYTTSRRLEPIPQLKCVGGTAGCDAYTPKVIQCQNKGWDGYDVQWECKTDLDV 97
Laf_TMEM66      ALTLHYNRYTTSRRLDPVPQLKCIGGTAGCNSYTPKVIQCQNKGWDGYDVQWECKTDLDI 89
Mdo_TMEM66      ALTLHRDRFTTARRTAPIPQLQCLGGSAGCPAHIPEIVQCRNKGWDGFDVQWECKAELDT 119
Gga_TMEM66      VLTLHRGRYTTARRTAAVPQLQCIGGSAGCS-DIPEVVQCYNRGWDGYDVQWQCKADLEN 94
Xla_TMEM66      TITLYADRYTNARRSAPVPQLKCIGGNAGCHAMVPQVVQCHNRGWDGLDVQWECRVDMDN 93
Xtr_TMEM66      AITLYADRYTNARRSAPVPQLKCIGGSAGCHTMVPQVVQCHNRGWDGFDVQWECKVDMDN 93
Dre_TMEM66      VLTLYRGRYTTARRSSPVPQLQCIGGSAGCGSFTPEVVQCYNRGSDGIDAQWECKADMDN 93
Ssa_TMEM66      VLTLYKGKYTTARRSSAVPQLQCVGGSAGCGSFIPEVVQCKNKGWDGVDAQWECKTDMDN 93
Tru_TMEM66      VLTLYRGLYTTARRSSPVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGMDIQWECRTDMDN 99
Tni_TMEM66      TLTLYRGRYTTARRSSPVPQLRCVGGSAGCQAFVPEVVQCQNRGWDGVDVQWECKTDMDN 89
Gac_TMEM66      ALTLYKNRYTTARRASPVPQLQCVGGSAGCQAFVPEVVQCQNKGWDGVDVQWECRTDMDN 92
Ppr_TMEM66      VLTLYKGRYTTARRSSPVLQLQCAGGTAGCGSFVPEVVQCYNRGSDGIDTQWECKADMDN 93
Cel_TMEM66      AITLHKGKMTTGRRVSPTFQLKCVGG-SAKGAFTPKVVQCANQGFDGSDVQWRCDADLPH 96
Cre_TMEM66      AITLNKGKMTTGRRVAPTLQLKCVGG-SAKGAFTPKVVQCSNQGFDGSDVQWRCDADLPH 96
Cbr_TMEM66      AITLHKGKMTTGRRVAPALQLKCVGG-SAKGQFSPKVVQCANQGFDGSDVQWRCDADLPH 96
                .:**  .  *..**  .  **:* ** :.     *.:::** *:* ** * **.* .::
```

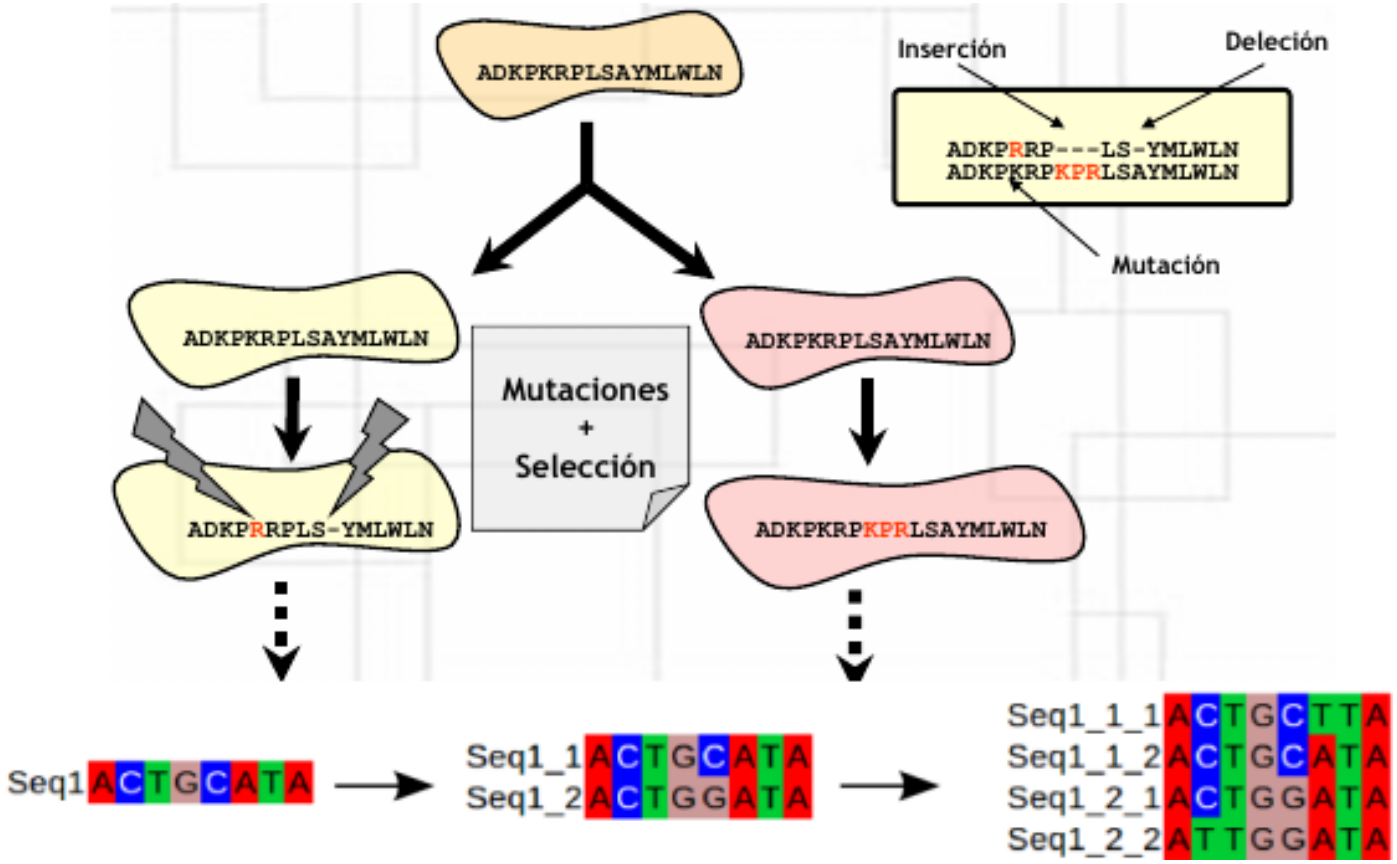
Definiciones

- **Residuo**:
  - ✓ Se refiere a secciones homólogas de las secuencias
  - ✓ A veces se llama residuo a cada columna del alineamiento
- Los residuos se pueden originarese ya sea
  - ✓ **Evolutivamente**: presumiblemente provenientes de un ancestro común
  - ✓ **Estructuralmente**: suelen ocupar lugares relevantes en la estructura 3D



Un Alineamiento nos cuenta una historia

En todos los casos los algoritmos de alineamiento múltiple asumen que las secuencias que se están alineando descienden de un antepasado común y lo que se intenta hacer es alinear las posiciones homólogas.



Pasos básicos

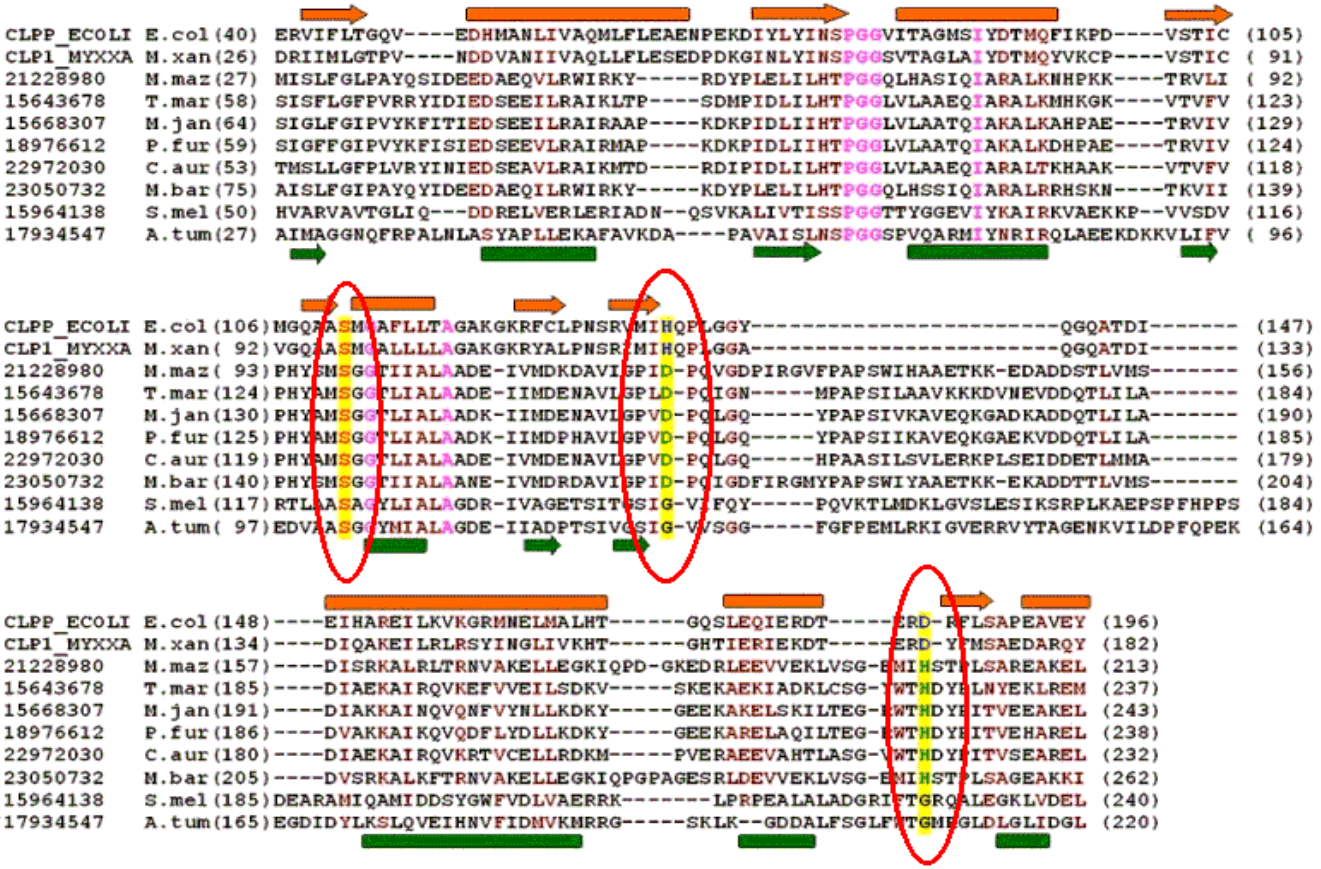
- Para hacer un alineamiento, generalmente necesitamos seleccionar:
  - ✓ 1. Las secuencias homólogas a alinear
  - ✓ 2. El software que utilice una función de puntuación óptima
  - ✓ 3. Los parámetros adecuados (fundamentalmente huecos)
- No hay un alineamiento perfecto
  - ✓ Las secuencias evolucionan más rápido que las estructuras o funcionalidades
  - ✓ La secuencia puede variar y la estructura o función seguir invariante

Aplicaciones de los Alineamientos Múltiples

- Dar información acerca de la función, estructura y evolución de una secuencia
  - ✓ Al conocer cómo se alinea respecto a un grupo de secuencias
  - ✓ Válido para análisis de genes, proteínas o poblaciones
- Encontrar miembros distantes de una familia de proteínas
  - ✓ Es muy frecuente que estén distantes, y el alineamiento de pares no suele ser lo suficientemente preciso para encontrarlos
- Primer paso (y el más importante) en la generación de árboles filogenéticos

Aplicaciones:

Determinar elemento común que corresponde al Sitio Activo de la Sesrina



Aplicaciones:

Extrapolación: determinar la función de una proteína desconocida con base en proteína conocidas

```
chite     ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat     --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr     KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
unknown   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
                ***. ::: .: ..  .    :  . .      *  .  *: *

chite     AATAKQNYIRALQEYERNGG-
wheat     ANKLKGEYNKAIAAYNKGESA
trybr     AEKDKERYKREM---------
unknown   AKDDRIRYDNEMKSWEEQMAE
          *   : .* . :
```

Aplicaciones:

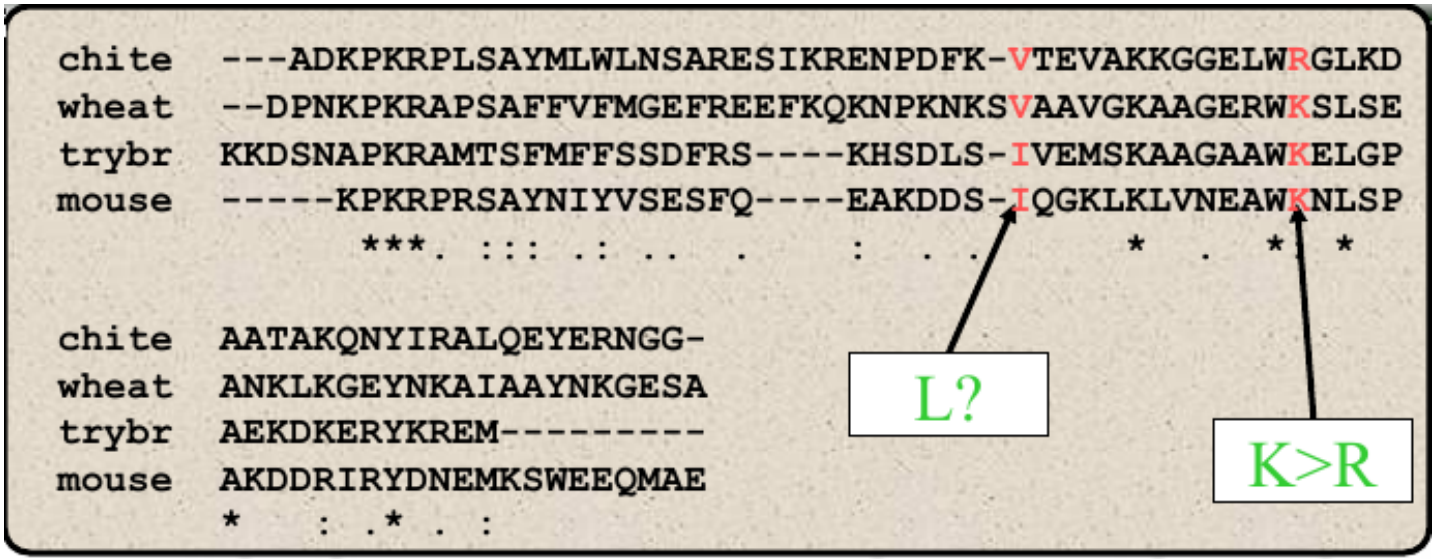
Bloques o patrones comúnes que corresponden a Motivos o Dominios de Proteínas

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Aplicaciones

Elementos representativos comunes que definen el perfil de una familia de proteínas

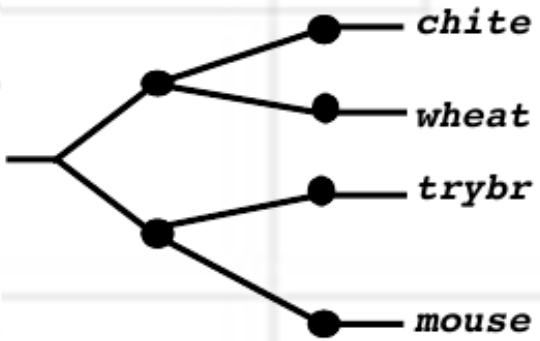


Aplicaciones

El alineamiento me guía para la construcción árboles filogenéticos

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: .. .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

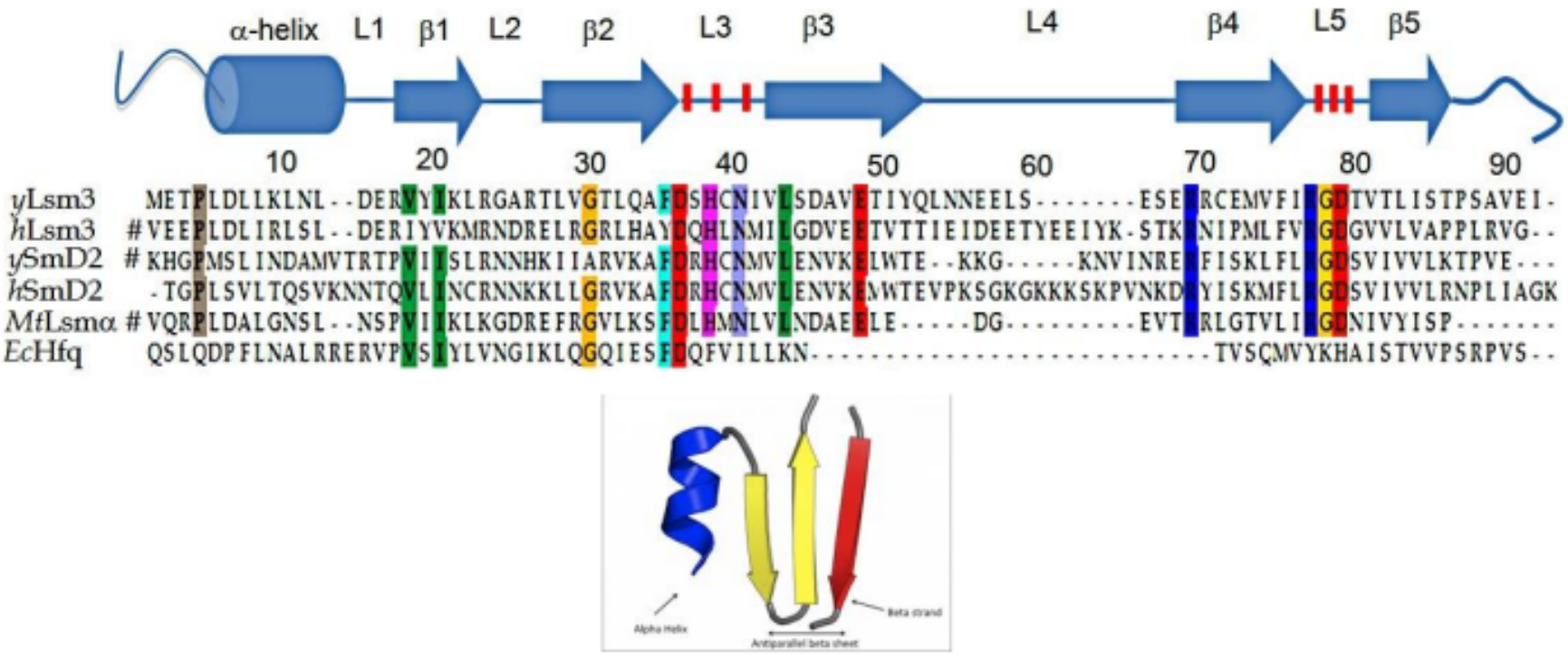


-Evolución

-Paralogía/Ortología

Aplicaciones de los AMS:

Determinación de la estructura 3D de una proteína



Selección de Secuencias para AMS

- Es su trabajo el de seleccionar las secuencias
- Dos factores a tener en cuenta:
  - ✓ El número de secuencias
  - ✓ La naturaleza de las secuencias
- Un número razonable de secuencias: 20 a 50
  - ✓ Inicia con 10 o 15 secuencias
  - ✓ Alineamientos pequeños son fáciles de mostrar y analizar
- Tipos de secuencias:
  - ✓ ADN, ARN, Proteínas (AA)
- Evite secuencias muy similares o muy distantes

Qué tipo de secuencias seleccionar: ADN o Proteínas

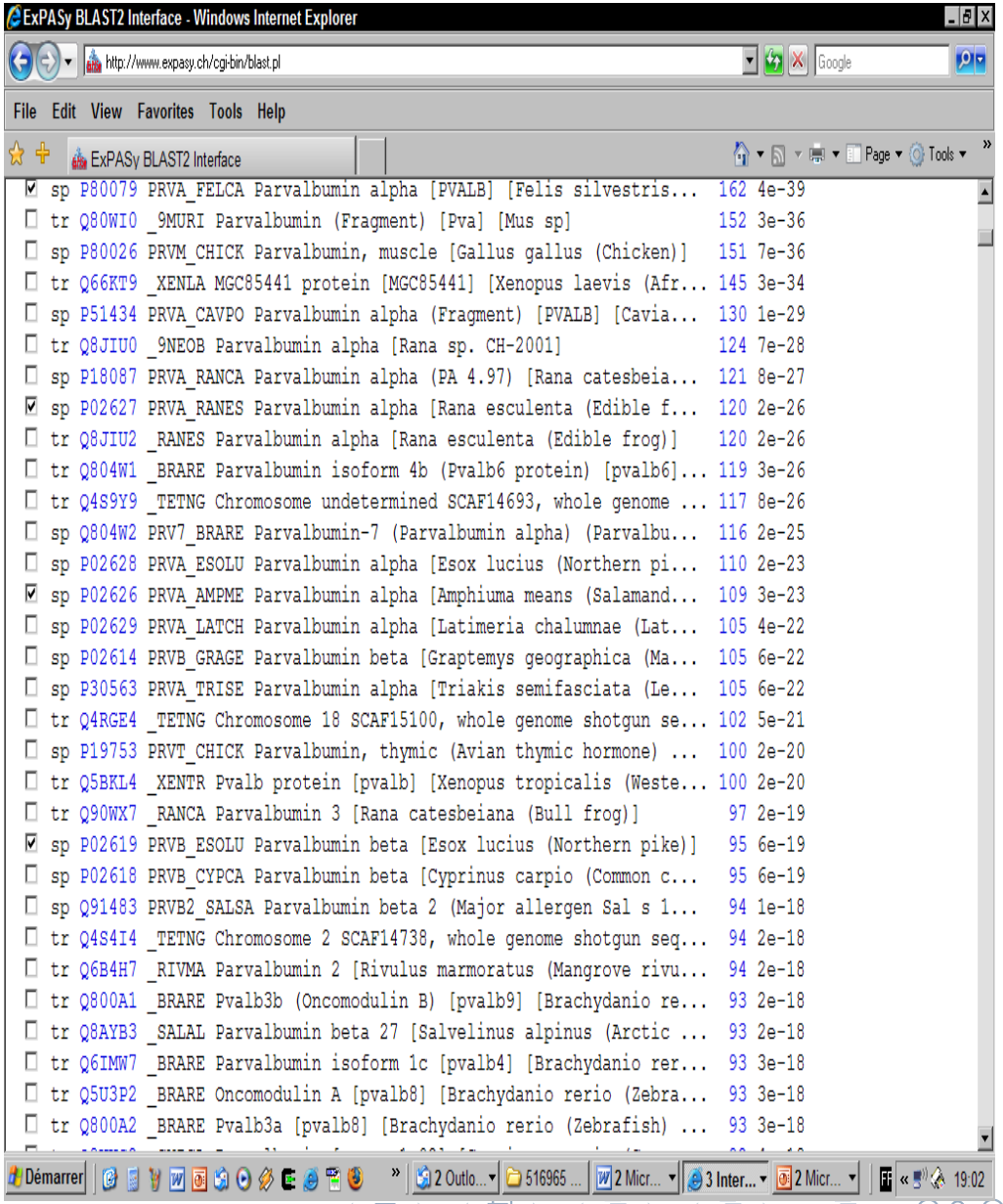
- Secuencias de ADN son más complicadas de alinear que proteínas ya que se nucleótidos se repiten demasiado
- Mayoría métodos trabajan con proteínas
- Si su ADN es codificante, trasládelo a proteínas
- Si las secuencias son homologas:
  - ✓ **En toda su extención**: Use alineamientos progresivos
  - ✓ **Solo en partes:** Use descubrimiento de motivos.

Obtención de Secuencias con BLAST

- El método más conveniente para seleccionar sus secuencias es a través de un servidor BLAST
- Servidores BLAST tienen integrados múltiples métodos de alineamiento:

Obtención de Secuencias con BLAST

- Seleccione algunas del tope
- Seleccione uniformemente algunas del medio que no sean tan idénticas
- La idea es tener secuencias:
  - ✓ No tan similares
  - ✓ No tan distantes



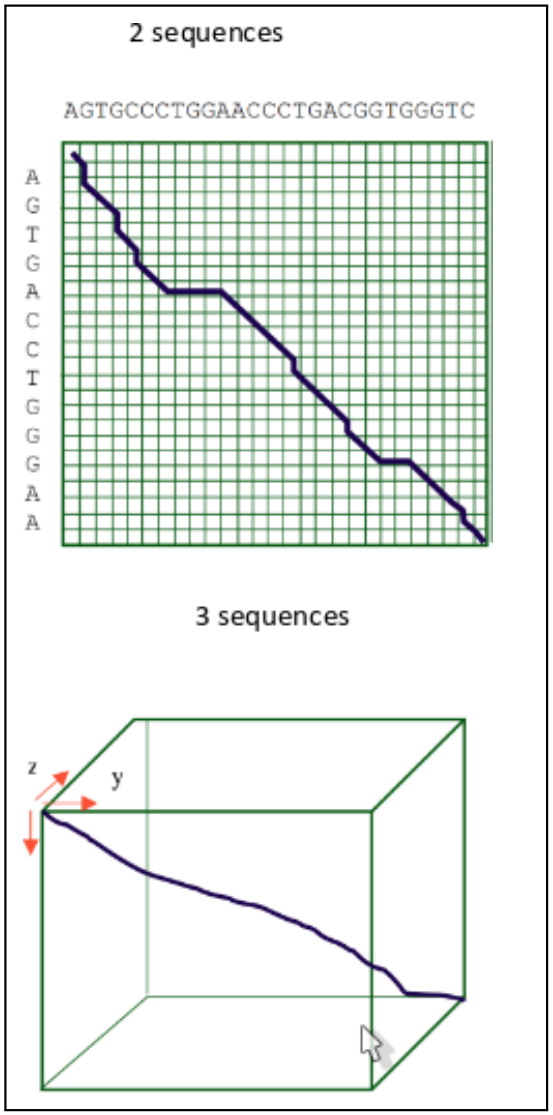
Muchos Algoritmos y Herramientas

- Existen muchos algoritmos
- La mayoría son métodos aproximativos donde no se garantiza el alineamiento perfecto:
- Todos los métodos tienen pros y contras:
  - ✓ CLUSTALW: alineamiento progresivo
  - ✓ MUSCLE: alineamiento iterativo
  - ✓ T-COFFEE: alineamiento basado en consistencia
  - ✓ EXPRESSO: alineamiento basada en la estructuras 3D

Algoritmos para Alineamiento Múltiple

- Existen cinco tipos de algoritmos:
  - ✓ 1. Métodos exactos
  - ✓ 2. Alineamiento progresivo
  - ✓ 3. Aproximaciones iterativas
  - ✓ 4. Métodos basados en la consistencia
  - ✓ 5. Métodos basados en la estructura
- Las aproximaciones no son excluyentes
  - ✓ Las tres últimas, por ejemplo, utilizan alineamiento progresivo

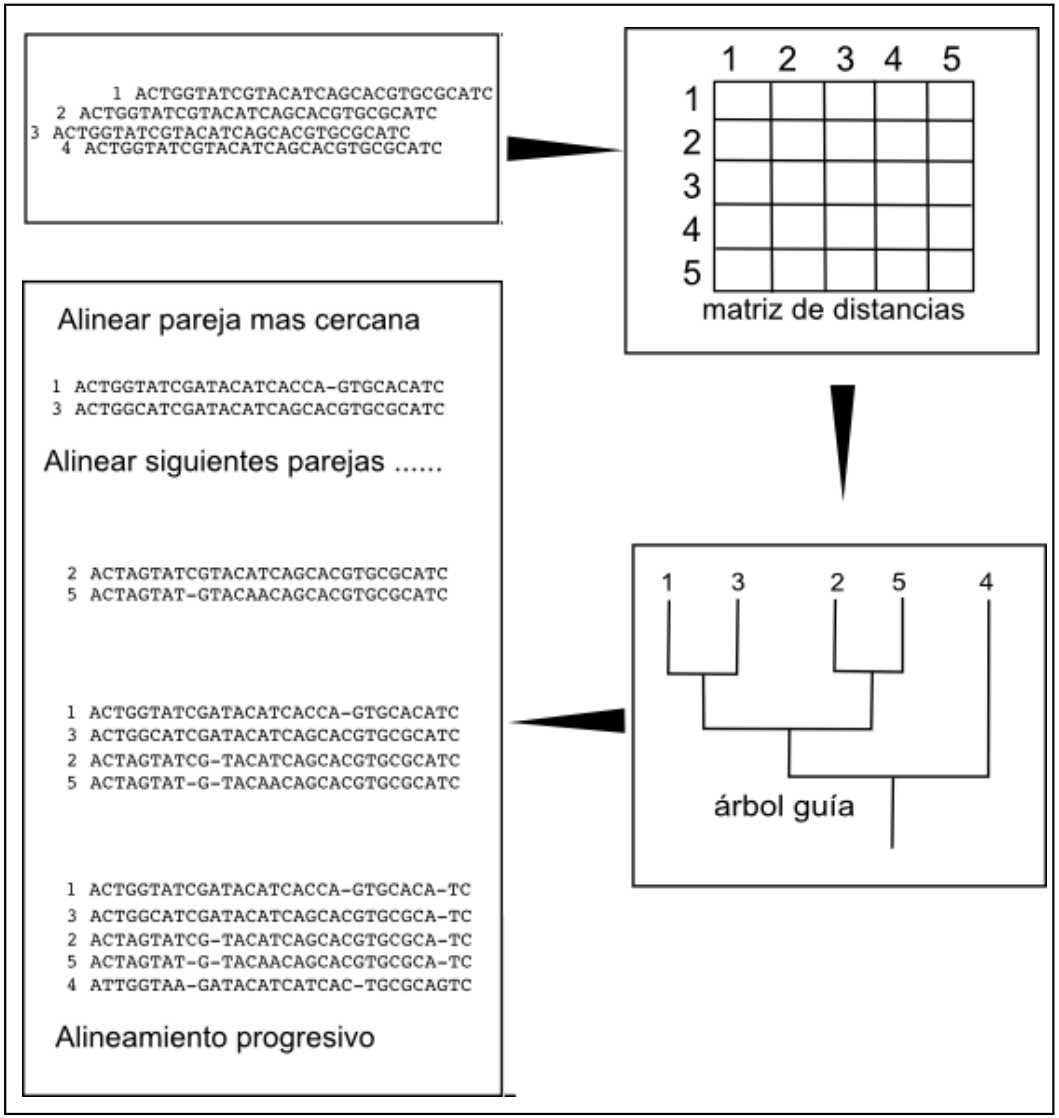
Métodos exactos



- Similar a alineamiento global de pares pero ahora con más de 3 secuencias:
  - ✓ Se basan en programación dinámica
- Aseguran un alineamiento óptimo (perfecto), pero son lentos
  - ✓ No son factibles ni en espacio ni en tiempo si tenemos más de unas pocas secuencias
- Se prefieren los métodos inexactos, mucho más rápidos
  - ✓ ClustalW
  - ✓ MUSCLE
  - ✓ T-COFFEE

Alineamiento progresivo

- "Progresivo"
  - ✓ Calcula alineamientos de pares entre las secuencias consideradas
  - ✓ Elige el mejor alineamiento de entre ellos
  - ✓ Añade progresivamente más secuencias al alineamiento escojido
- El programa de alineamiento progresivo más usado, pero ya no el mejor es ClustalW



Algoritmo de ClustalW

Clustal implementa el algoritmo de Feng y Doolittle, que consta de 3 fases:

1. Alineamiento global 2 a 2 mediante el algoritmo de NW
2. Se crea un árbol guía
3. Se crea el alineamiento múltiple paso a paso

Fase 1: Alineamiento global de pares

- Ejemplo: cinco globinas muy conocidas, bastante distantes:
  - ✓ NP_000509, NP_005359, NP_067080, 1FSL, 1D8U
- Para 5 secuencias tendremos 10 alineamientos
- Para n secuencias tendremos n!/(2(n-2)!) alineamientos

SeqA	Name	Length	SeqB	Name	Length	Score
1	beta_globin	147	2	myoglobin	154	25.17
1	beta_globin	147	3	neuroglobin	151	15.65
1	beta_globin	147	4	soybean_globin	144	13.19
1	beta_globin	147	5	rice_globin	166	21.09
2	myoglobin	154	3	neuroglobin	151	16.56
2	myoglobin	154	4	soybean_globin	144	8.33
2	myoglobin	154	5	rice_globin	166	12.99
3	neuroglobin	151	4	soybean_globin	144	17.36
3	neuroglobin	151	5	rice_globin	166	18.54
4	soybean_globin	144	5	rice_globin	166	43.06

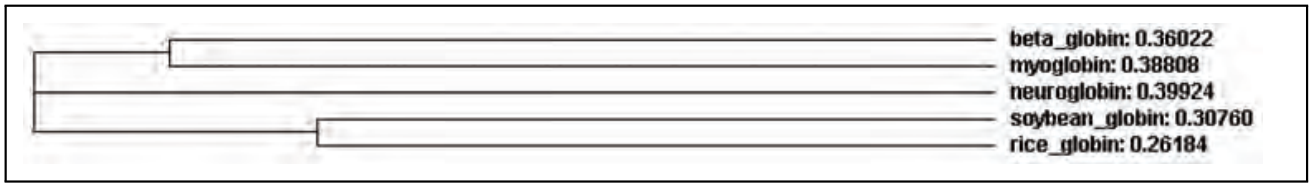
Fase 2: Creación del árbol guía

- La longitud de las ramas depende de las distancias
- Se unen las ramas de las secuencias con distancias más cortas
- Árbol:

Formato de Newid (.nwk)

```
(
(
beta_globin:0.36022,
myoglobin:0.38808)
:0.06560,
neuroglobin:0.39924,
(
soybean_globin:0.30760,
rice_globin:0.26184)
:0.13652);
```

Representación Gráfica:



Fase 3: Creación del alineamiento múltiple

- Se seleccionan las dos secuencias más cercanas según el árbol guía
- Se seleccionan las dos secuencias más cercanas siguientes
  - ✓ Si alguna coincide, se añade al alineamiento de pares, dando lugar a un alineamiento de 3+ secuencias, o perfil
  - ✓ Si ninguna coincide con las anteriores, se realiza su alineamiento de pares, para después realizar alineamiento de perfiles.
- El alineamiento continúa hasta llegar a la raíz del árbol

Fase 3: Creación del alineamiento múltiple

Notación:

- Punto (.): Coincidencia
- Dos puntos (:): Coincidencia Alta.
- Asterisco (*): Coincidencia Exacta

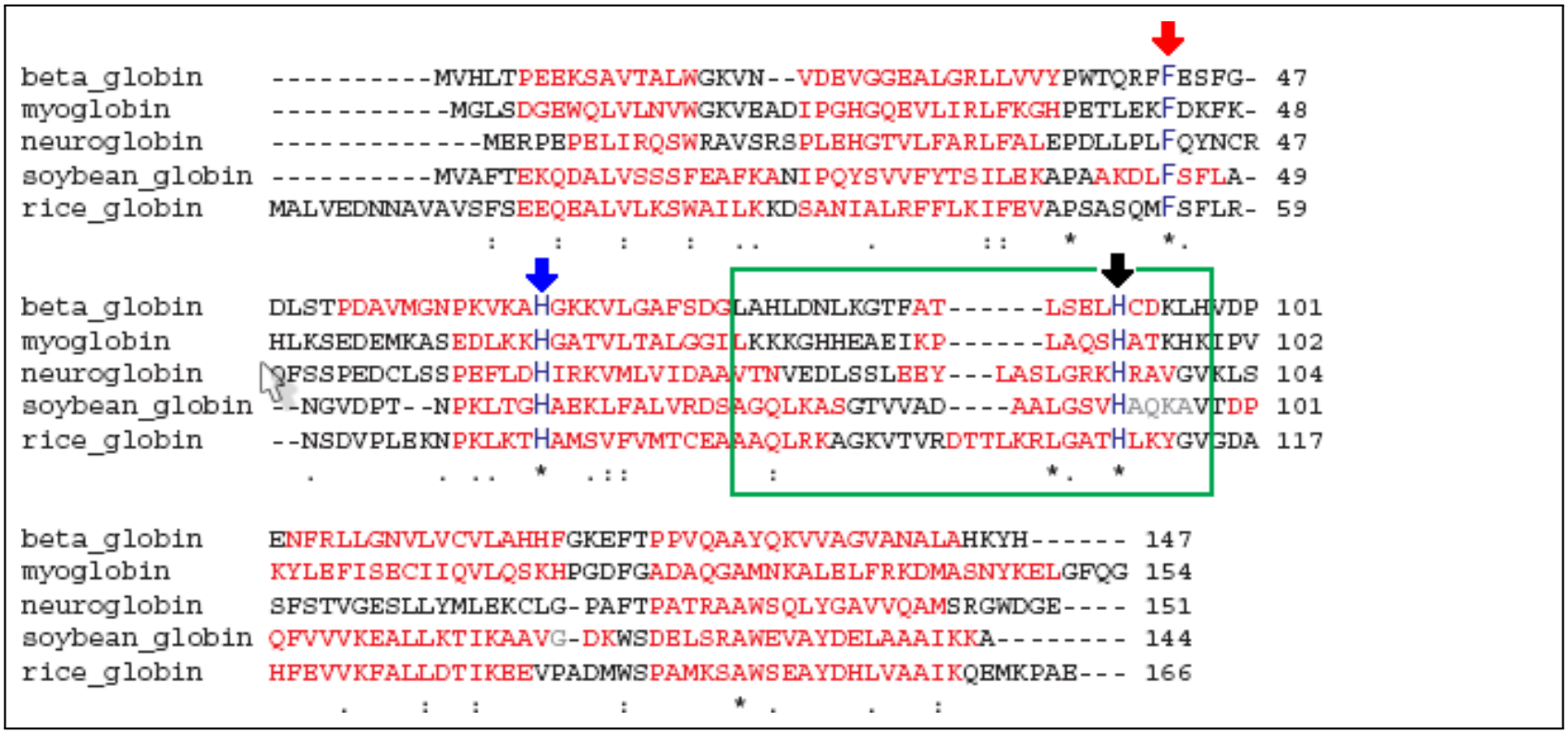
```
beta_globin     ----------MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin       ----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin     ------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean_globin  ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice_globin     MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
                          :   :   :   :  ..      .      ::   *     *.

beta_globin     DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP 101
myoglobin       HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPV 102
neuroglobin     QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS 104
soybean_globin  --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADLAALGSVHAQKAVTDP 101
rice_globin     --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
                  .       . ..  *  .::       :                *.  *

beta_globin     ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin       KYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin     SFSTVGESLLYMLEKCLG-PAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean_globin  QFVVVKEALLKTIKAAVG-DKWSDELSRAWEVAYDELAAAIKKA-------- 144
rice_globin     HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
                     .    :  :       :      * .     .   :
```

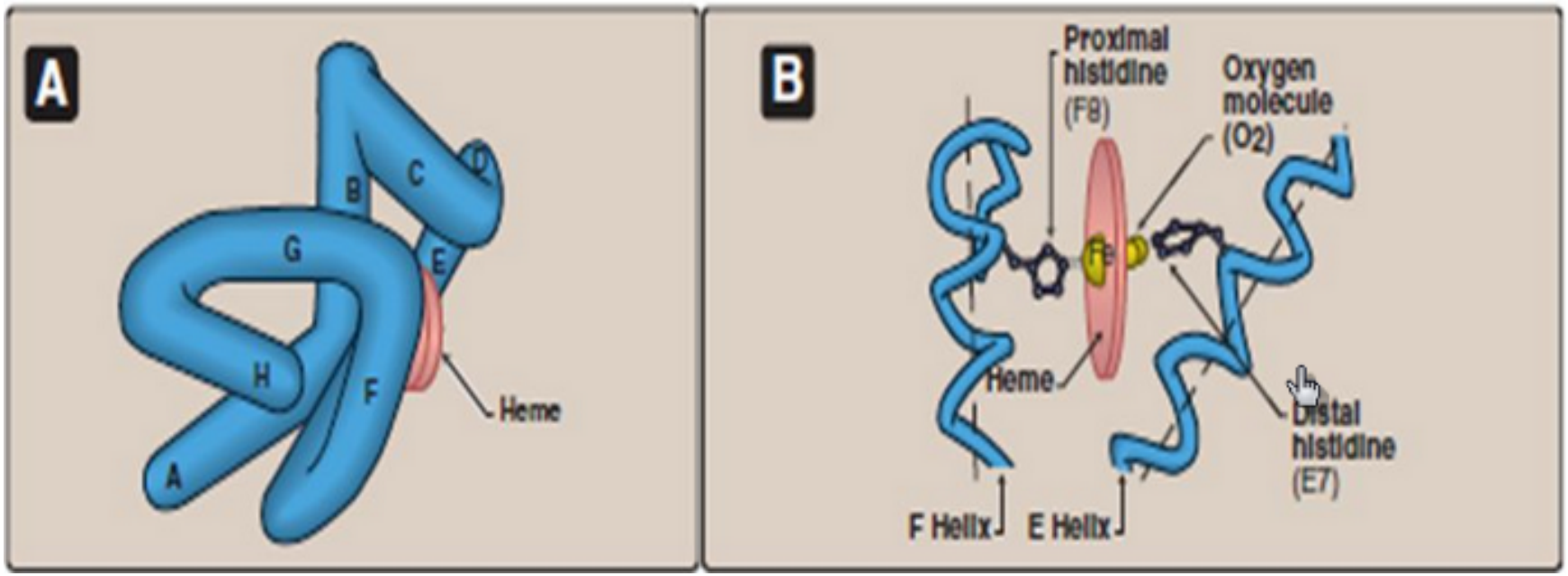
Interpretación del Alineamiento

- Hay una fenilanina muy conservada (flecha roja)
- Hay dos histidinas muy conservada (flecha azul y flecha negra)
  - ✓ Estas histidinas son importantes en la coordinación del enlace de la proteína al grupo hemo



Estructura y Función de la Mioglobina

- Presente en el corazón y en los músculos esqueléticos
- Funciona como repositorio de oxígeno dentro de las células del músculo
- Consiste de una único polipéptido conectado al grupo hemo



- Fig A: Hélices A hasta H
- Fig B: Sitio del Ligamento del oxígeno de la mioglobina

Ejercicio 1

- Realizar el anterior alineamiento múltiple mediante el ClustalW del EBI: http://www.ebi.ac.uk/clustalW y compare con el alineamiento anterior:

- Cinco beta globinas bastante distantes:

>beta_globin 2hhbB NP_000509.1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVL
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFR
KDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVML
VIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAV
VQAMSRGWDGE
>soybean_globin 1FSL leghemoglobin P02238 LGBA_SOYBN [Glycine max]
MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLANGVDPTNPKLTGHAEKLFALV
RDSAGQLKASGTVVADAALGSVHAQKAVTDPQFVVVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAAA
IKKA
>rice_globin 1D8U rice Non−Symbiotic Plant Hemoglobin NP_001049476.1
MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLRNSDVPLEKNPK
LKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDAHFEVVKFALLDTIKEEVPADMWS
PAMKSAWSEAYDHLVAAIKQEMKPAE

Ejercicio 2

- Realizar el siguiente alineamiento múltiple mediante el ClustalW del EBI: http://www.ebi.ac.uk/clustalW

- Cinco globinas bastante cercanas:

>human_NP_000509
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>Pan_troglodytes_XP_508242
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>Canis_familiaris_XP_537902
MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAKVKAHGKKVLN
SFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVAN
ALAHKYH
>Mus_musculus_NP_058652
MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPKVKAHGKKVIT
AFNEGLKNLDNLKGTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVAT
ALAHKYH
>Gallus_gallus_XP_444648
MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLT
SFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAH
ALARKYH

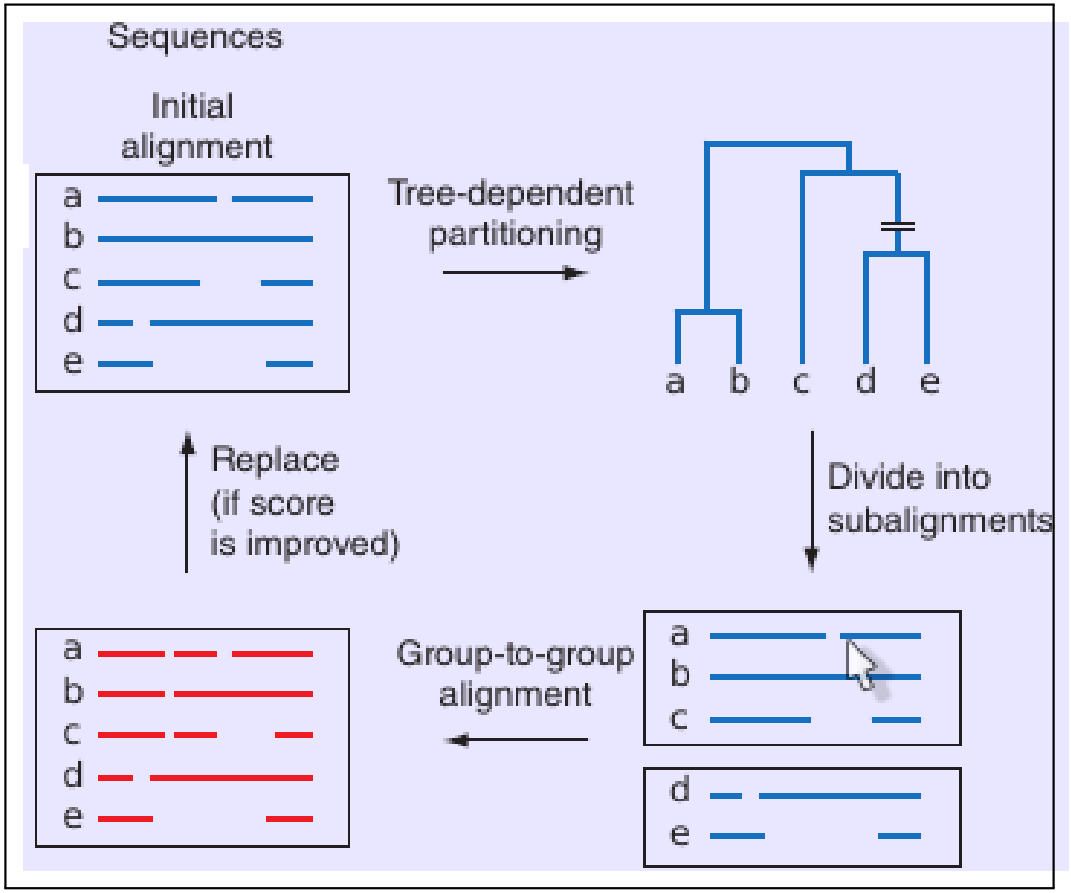
Alineamiento Beta Globinas Cercanas

- The coloring scheme (ClustalW program) includes groups such as:
  - ✓ acidic amino acids (blue),
  - ✓ basic amino acids (magenta), and
  - ✓ hydrophobic res- idues (red).

ClustalW y huecos

- ClustalW sigue la política: "una vez se encuentra un hueco, siempre hay un hueco"
  - ✓ Cuando hay un hueco en un alineamiento, se fomenta que se conserve en alineamientos posteriores
  - ✓ Da al alineamiento múltiple una estructura de "bloques
- Generalmente, alineamientos con más huecos (menos compactos) coinciden mejor con la filogenia y la estructura de proteínas conocidas como la globina

Algoritmos Iterativas



- Calculan una solución subóptima mediante un alineamiento progresivo
- Modifican el alineamiento y repiten el proceso el alineamiento como inicio hasta que la solución converge
- Se corrige el error de los alineamiento progresivo, donde una vez que cometemos un error, no lo podemos corregir
- Algoritmos conocidos: MAFFT, MUSCLE

Aproximaciones basadas en la consistencia

- Este enfoque combina alineamientos progresivos e iterativos y métodos probabilisticos
- Esta estrategia suele generar alineamientos de secuencia mucho más precisos
- ProbCons y T-Coffee son los dos algoritmos más conocidos
- M-COFFEE: usa muchos algoritmos (T-COFFEE, ClustalW, MAFFT, MUSCLE, and ProbCons)

Taller1: Comparaciones de diferentes Algoritmos de MSA

- Realizar el alineamiento múltiple de las siguientes 9 globinas usando: (a) CLUSTALW, (b) MAFFT, (b) MUSCLE, y (d) T-COFFEE.
  - ✓ (1) hbb_human (human NP_000509.1, 1HBB);
  - ✓ (2) hbb_chimp (Pan_troglodytes XP_508242.1, no structure);
  - ✓ (3) hbb_dog (Canis lupus familiaris NP_001257813.1, 2QLS|B);
  - ✓ (4) hbb_mouse (Mus_musculus NP_058652.1, 3HRW|B);
  - ✓ (5) hbb_chicken (Gallus_gallus NP_990820.1, 1HBR|B);
  - ✓ (6) myoglobin (human NP_005359.1, 3RGK);
  - ✓ (7) neuroglobin (human NP_067080.1,1OJ6|A);
  - ✓ (8) globin_soybean (Glycine max leghemoglobin A, NP_001235928.1, 1FSL); and
  - ✓ (9) globin_rice (Oryza sativa (japonica cultivar-group) NonSymbiotic Plant Hemoglobin NP_001049476.1, 1D8U).

Taller1: Secuencias

Las secuencias anteriores deben parecerse a las siguientes:

>hbb_human NP_000509.1
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLA
HLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>hbb_chimp Pan_troglodytes XP_508242
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLA
HLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>hbb_dog Canis_familiaris XP_537902
MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAKVKAHGKKVLNSFSDGLK
NLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH
>hbb_mouse Mus_musculus NP_058652
MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPKVKAHGKKVITAFNEGLK
NLDNLKGTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH
>hbb_chicken Gallus_gallus XP_444648
MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAVK
NLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH
>myoglobin_human 2MM1 Homo sapiens NP_005359.1
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGIL
KKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
>neuroglobin_human 1OJ6A Homo sapiens NP_067080.1
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT
NVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWDGE
>globin_soybean 1FSL leghemoglobin P02238 LGBA_SOYBN [Glycine max]
MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLANGVDPTNPKLTGHAEKLFALVRDSAGQL
KASGTVVADAALGSVHAQKAVTDPQFVVVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAAAIKKA
>globin_rice 1D8U rice Non-Symbiotic Plant Hemoglobin NP_001049476.1
MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLRNSDVPLEKNPKLKTHAMS
VFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDAHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHL
VAAIKQEMKPAE

Taller1: Trabajo Casa

- Realizar y comparar los alineamientos con los cuatro algoritmos ClustalW, MAFF, Muscle, T-Coffee
- Investigar alineamientos con estructuras 3D y realizar el anterior alineamiento utilizando la herramienta EXPRESSO
- Respondes:
  - ✓ Para usted, cuál algoritmo presenta el mejor alineamiento? Porqué?
  - ✓ Qué elementos comunes describe este grupo de globinas?
  - ✓ Existe algún bloque común?
  - ✓ Qué tipo de algoritmo realiza el mejor alineamiento y en general como lo hace?