Alineamientos de múltiples secuencias

Rodrigo Santamaría

Alineamientos de múltiples secuencias

- Introducción
  - Motivación
  - Definición
  - Usos
- Algoritmos
- Benchmarking
- Visualización
- Bases de Datos

Introducción

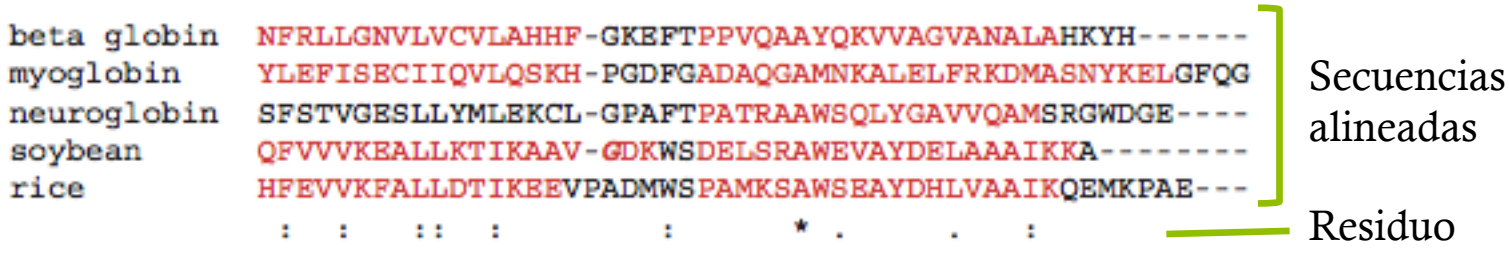
- Multiple Sequence Alignment (MSA)
- Hemos visto cómo comparar una secuencia con otra (alineamiento de pares)
- Hemos visto cómo comparar una secuencia con muchas otras en una BD (muchos alineamientos de pares - BLAST)
- Ahora veremos cómo comparar múltiples secuencias simultáneamente, no de dos en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

Definición

- Un alineamiento múltiple es una colección de tres o más secuencias de aminoácidos o nucleótidos parcial o completamente alineados
  - Residuo: secciones homólogas de las secuencias, en un sentido
    - Evolutivo: presumiblemente provenientes de un ancestro común
    - Estructural: suelen ocupar lugares relevantes en la estructura 3D



[NOTA: A veces se llama residuo a cada columna del alineamiento]

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)

Usos típicos

- 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
- Clasificación y generación de BBDD de proteínas una vez secuenciado el genoma completo de un organismo
- Primer paso (y el más importante) en la generación de árboles filogenéticos

Alineamientos de múltiples secuencias

Introducción

Algoritmos

- Métodos exactos
- Progresivos – Clustal
- Iterativos – MUSCLE
- Consistencia – T-Coffee
- Estructura

Benchmarking

Visualización

Bases de Datos

Algoritmos

- Existen cinco aproximaciones algorítmicas distintas al MSA
  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

- Se basan en programación dinámica
  - Similar a un NW para alineamiento global de pares
- Aseguran un alineamiento óptimo, pero son lentos
  - No son factibles ni en espacio ni en tiempo si tenemos más de unas pocas secuencias
  - Para N secuencias de longitud media L, el coste en tiempo es $O(2^N L^N)$
- Se prefieren los métodos inexactos, mucho más rápidos
  - ClustalW: $O(N^4+L^2)$
  - MUSCLE: $O(N^4+ NL^2)$

Alineamiento progresivo

- "Progresivo":
  - Calcula alineamientos de pares entre las secuencias consideradas
  - Elige el mejor alineamiento de entre ellos
  - Añade *progresivamente* más secuencias al alineamiento
- El programa de alineamiento progresivo más usado es ClustalW
  - http://www.ebi.ac.uk/clustsalw

Clustal

- Clustal implementa el algoritmo de Feng y Doolittle, que consta de 3 fases
  1. Alineamiento global 2 a 2 mediante el algoritmo de NW
     - Las puntuaciones de similitud se traducen a una matriz de distancias
  2. Se crea un árbol guía a partir de la matriz de distancias
  3. Se crea el alineamiento múltiple paso a paso
     1. Haciendo alineamientos de pares pero según las distancias

- Dos versiones:
  - ClustalW (línea de comandos)
  - ClustalX (interfaz gráfica)

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	Len(aa)	SeqB	Name	Len(aa)	Score
1	beta_globin	147	2	myoglobin	154	25
1	beta_globin	147	3	neuroglobin	151	15
1	beta_globin	147	4	soybean	144	13
1	beta_globin	147	5	rice	166	21
2	myoglobin	154	3	neuroglobin	151	16
2	myoglobin	154	4	soybean	144	8
2	myoglobin	154	5	rice	166	12
3	neuroglobin	151	4	soybean	144	17
3	neuroglobin	151	5	rice	166	18
4	soybean	144	5	rice	166	43

Las puntuaciones se traducirán a distancias para que puedan usarse para generar el árbol

1 ← Mejor alineamiento

Fase 1. alineamiento global de pares

- Conversión de similitud a distancia (Feng y Doolittle)
  - Sea $S_{real(ij)}$ la similitud entre las secuencias *i* y *j*
    - Sea $S_{rand(ij)}$ la media de las similitudes calculadas para las 2 secuencias aleatorizadas muchas veces (p. ej. 1000)
    - Sea $S_{iden(ij)}$ la media de las similitudes identidad:

$$S_{iden(ij)} = \frac{S_{real(ii)} + S_{real(jj)}}{2}$$

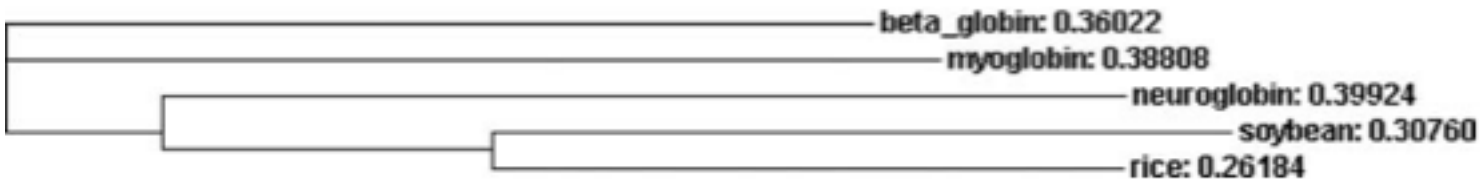
    - Sea $S_{eff(ij)} = \frac{S_{real(ij)} + S_{rand(ij)}}{S_{iden(ij)} + S_{rand(ij)}} \times 100$
  - La distancia entre las secuencias *i* y *j* es $D_{ij} = -\ln S_{eff(ij)}$

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

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

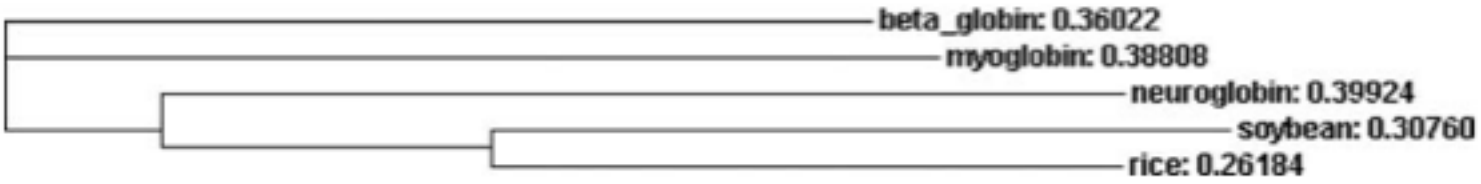
Formato Newick (.nwk)



Fase 3. Creación del alineamiento múltiple

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

Fase 3. Creación del alineamiento múltiple




```
beta globin    ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG-  47
myoglobin      ----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK-  48
neuroglobin    ------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR  47
soybean        ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA-  49
rice           MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR-  59
                          :   :   :   : .. .    .      ::   *     *.

beta globin    DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin      HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin    QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean        --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice           --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
                 .       . ..  *  .::          :           :              :

beta globin    NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin      YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin    SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean        QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA-------- 144
rice           HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
                 :   :   ::   :         :          .      .    :
```

. coincidencia
: coincidencia alta
* coincidencia exacta

Interpretación del alineamiento

- Hay una fenilanina muy conservada (flecha roja)
- Hay una histidina muy conservada (flecha hueca)
  - Regula el enlace hemo
- Hay otra histidina que a pesar de saberse que está muy conservada no se ha alineado bien (flecha negra)
- **Ejercicio**: realizad este alineamiento múltiple mediante el ClustalW del EBI: http://www.ebi.ac.uk/clustalW

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"
- Loytynoja y Goldman (2005) demostraron que alineamientos con más huecos (menos compactos) coinciden mejor con la filogenia y la estructura de proteínas conocidas como la globina

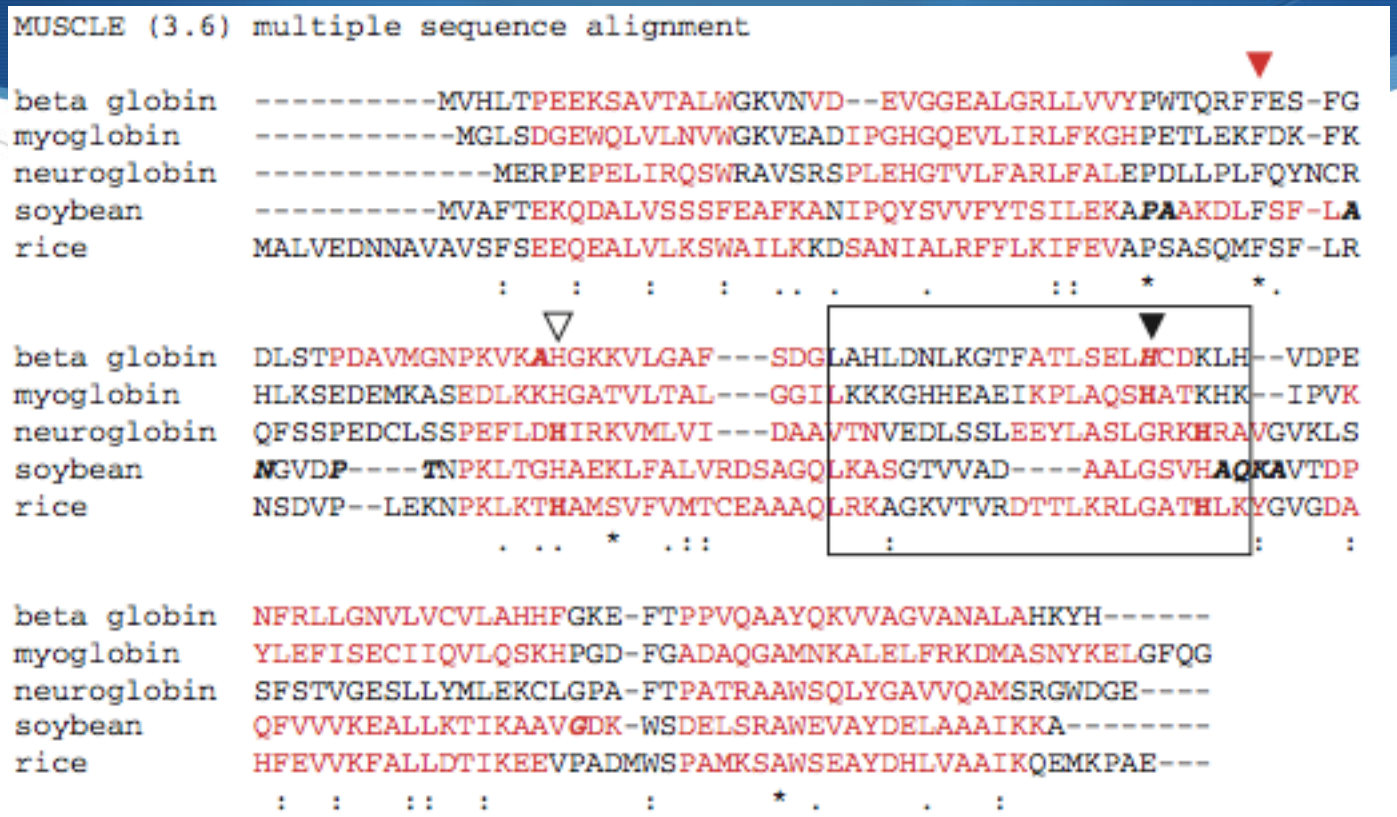
Aproximaciones iterativas

- Calculan una solución subóptima mediante un alineamiento progresivo
- Luego modifican el alineamiento mediante programación dinámica hasta que la solución converge
- En un alineamiento progresivo normal, una vez que cometemos un error, no lo podemos corregir
  - La aproximación iterativa soluciona esto

MUSCLE

- Multiple Sequence Comparison by Log-Expectation
- Es un programa muy popular por su precisión y rapidez
  - Alinea 1000 proteínas de tamaño ~300 en 21s
- http://www.ebi.ac.uk/muscle

MUSCLE



- MUSCLE realiza el alineamiento de las globinas usando menos huecos, pero sigue teniendo problemas con la histidina (flecha negra)
- **Ejercicio**: calculad el alineamiento de las globinas mediante MUSCLE

Aproximaciones basadas en la consistencia

- Esta aproximación incorpora la información de las distintas secuencias en la creación de cada alineamiento de pares
  - En la primera fase de un alineamiento progresivo clásico, se utiliza sólo la información de dos secuencias para cada alineamiento de pares
- Esta estrategia suele generar alineamientos de secuencia mucho más precisos, según los estudios de benchmarking
- ProbCons y T-Coffee son los dos algoritmos más conocidos

T-Coffee

- Tree-based **Consistency** Objective Function For alignmEnt Evaluation
  - Disponible en http://www.ebi.ac.uk/tcoffee
- Algoritmo
  - Calculamos todos los alineamientos de pares globales entre secuencias, utilizando el algoritmo NW, *Y* calculamos también los 10 alineamientos de pares locales con puntuación más alta
    - Con estas puntuaciones damos pesos a cada par de nucleótidos/aminoácidos alineados
  - Realizamos un alineamiento progresivo iterativo, utilizando dichos pesos en la fase de refinamiento iterativo

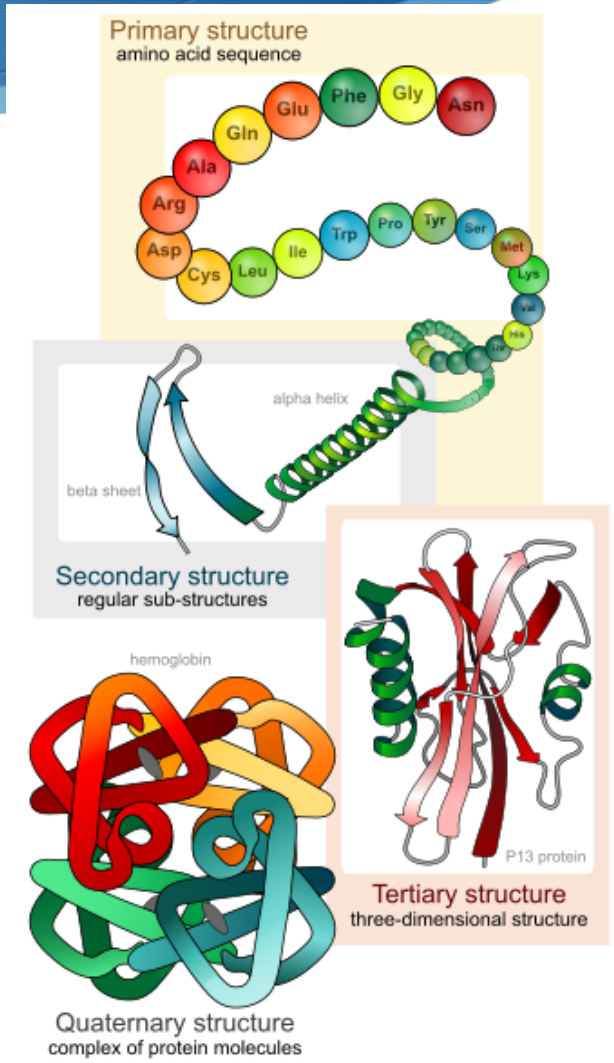
T-Coffee

```
CLUSTAL FORMAT for T-COFFEE Version_5.13

beta globin  ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
myoglobin    ----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin  ----------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-FLA
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
                       :   :   :   : .. .   .      ::   *    *.

beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVDP
myoglobin    HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIPV
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLASLGRKH-RAVGVKL
soybean      NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
             .         . ..  *  .::          :                *.  *

beta globin  ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------
myoglobin    KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin  SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean      Q-FVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA--------
rice         H-FEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
               :   :   ::  :        :      * .     .   :
```

- T-Coffee consigue alinear más correctamente las cinco globinas
- **Ejercicio**: realizad el alineamiento múltiple de las globinas con T-Coffee

Aproximaciones basadas en la estructura

- Las estructuras terciarias evolucionan más lentamente que la estructura primaria
  - Por ejemplo, la beta-globina y la mioglobina humanas tienen poca similitud (están en la "dimensión desconocida") pero sus estructuras están claramente relacionadas
- Estas aproximaciones utilizan información sobre la estructura 3D de una o más de las secuencias
- Algunos implementaciones son PRALINE, PipeAlign y Expresso (módulo de T-Coffee)



Alineamientos de múltiples secuencias

Introducción
Algoritmos
Benchmarking
Visualización
Bases de Datos

Benchmarking

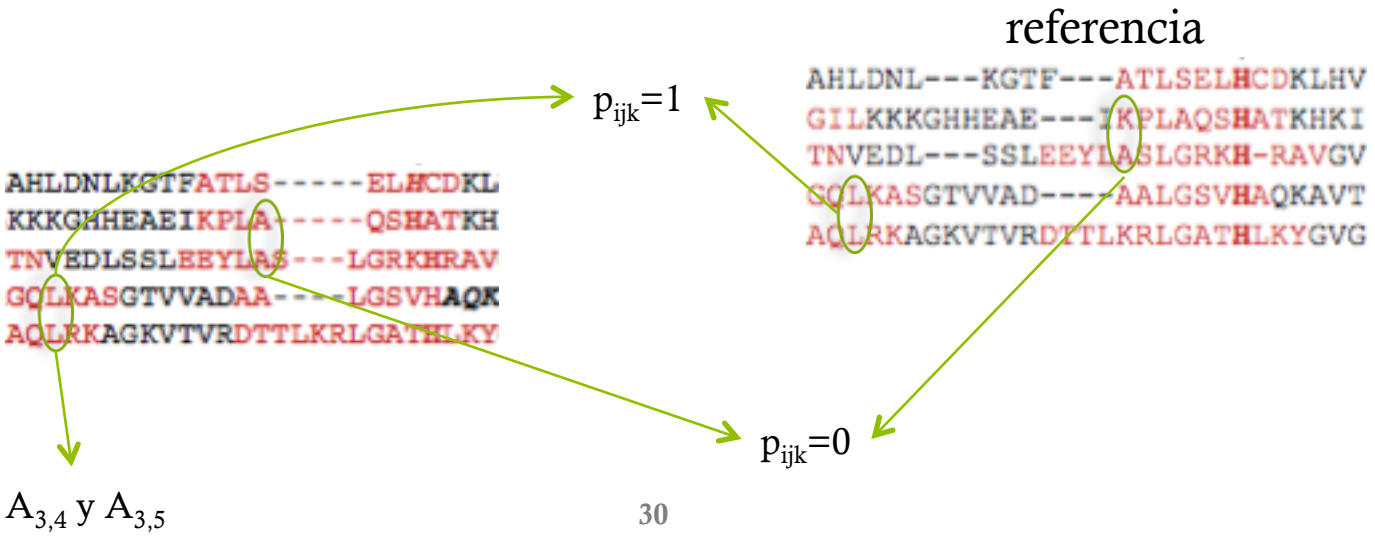
- Hay 5 aproximaciones distintas al MSA
  - Y de cada aproximación, múltiples implementaciones
  - ¿Cómo determinamos su precisión y rendimiento?
- Solución: Comparar el MSA de nuestro método con el MSA canónico de secuencias con estructuras 3D conocidas
  - **Benchmark**: alineamiento "perfecto" con el que comparar otros
  - Existen varios conjuntos de alineamientos de benchmark
- La "bondad" de un MSA se mide como un valor relativo al benchmark elegido, calculando una función de puntuación

Puntuación de suma de pares (SPS)

- Métrica más usada para comparación con benchmarks
- Sea un alineamiento de N secuencias en M columnas
  - A_{i1}… A_{iN} son los residuos para la columna i

SPS

- Sea un par de residuos A_{ij} y A_{ik}
- p_{ijk} es
  - 1 si A_{ij} y A_{ik} están alineados en nuestro alineamiento y en el de referencia
  - 0 en cualquier otro caso



SPS

- Para la columna i, la puntuación S_i es:

$$S_i = \sum_{j=1, j \neq k}^{N} \sum_{k=1}^{N} p_{ijk}$$

- Y para el alineamiento múltiple completo:
  - S_{ri} es la puntuación del propio alineamiento de referencia

$$SPS = \frac{\sum_{i=1}^{M} S_i}{\sum_{i=1}^{M_r} S_{ri}}$$

Conjuntos de datos de Benchmark

Database	Reference	URL
BAliBASE	Thompson et al. (2005)	http://www-bio3d-igbmc.u-strasbg.fr/balibase/
HOMSTRAD	Mizuguchi et al. (1998)	http://www-cryst.bioc.cam.ac.uk/~homstrad/
IRMBASE	Subramanian et al. (2005)	http://dialign-t.gobics.de/main
OxBench	Raghava et al. (2003)	http://www.compbio.dundee.ac.uk/Software/Oxbench/oxbench.htm
Prefab	Edgar (2004b)	http://www.drive5.com/muscle/prefab.htm
SABmark	Van Walle et al. (2005)	http://bioinformatics.vub.ac.be/databases/content.html

Conclusiones de los estudios de Benchmarking

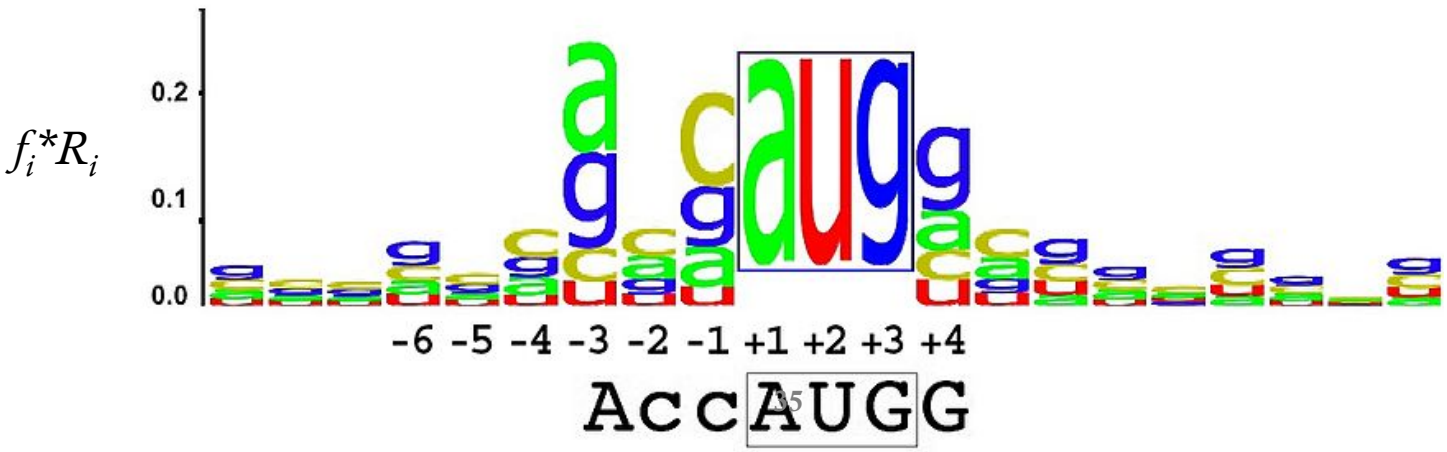
- Añadir más homólogos a un MSA mejora su precisión
- Para grupos de secuencias con baja identidad la precisión se reduce, siendo especialmente grave con <25% de similitud
- Una secuencia huérfana es una proteína divergente respecto al resto de su familia. Contra todo pronóstico, no estropean el MSA, sobre todo si usamos alineamientos globales
- Generalmente, el alineamiento global es mejor que el local para MSA, excepto
  - En proteínas con muchos grupos carboxilo o amino
  - En secuencias muy divergentes

Alineamientos de múltiples secuencias

Introducción
Algoritmos
Benchmarking
Visualización
Bases de Datos

Logo de secuencia

- Representación gráfica de la conservación de residuos
  - También llamado logo de consenso
- Parte de un alineamiento, y representa los residuos más grandes cuanto más conservados estén



Logo de secuencia

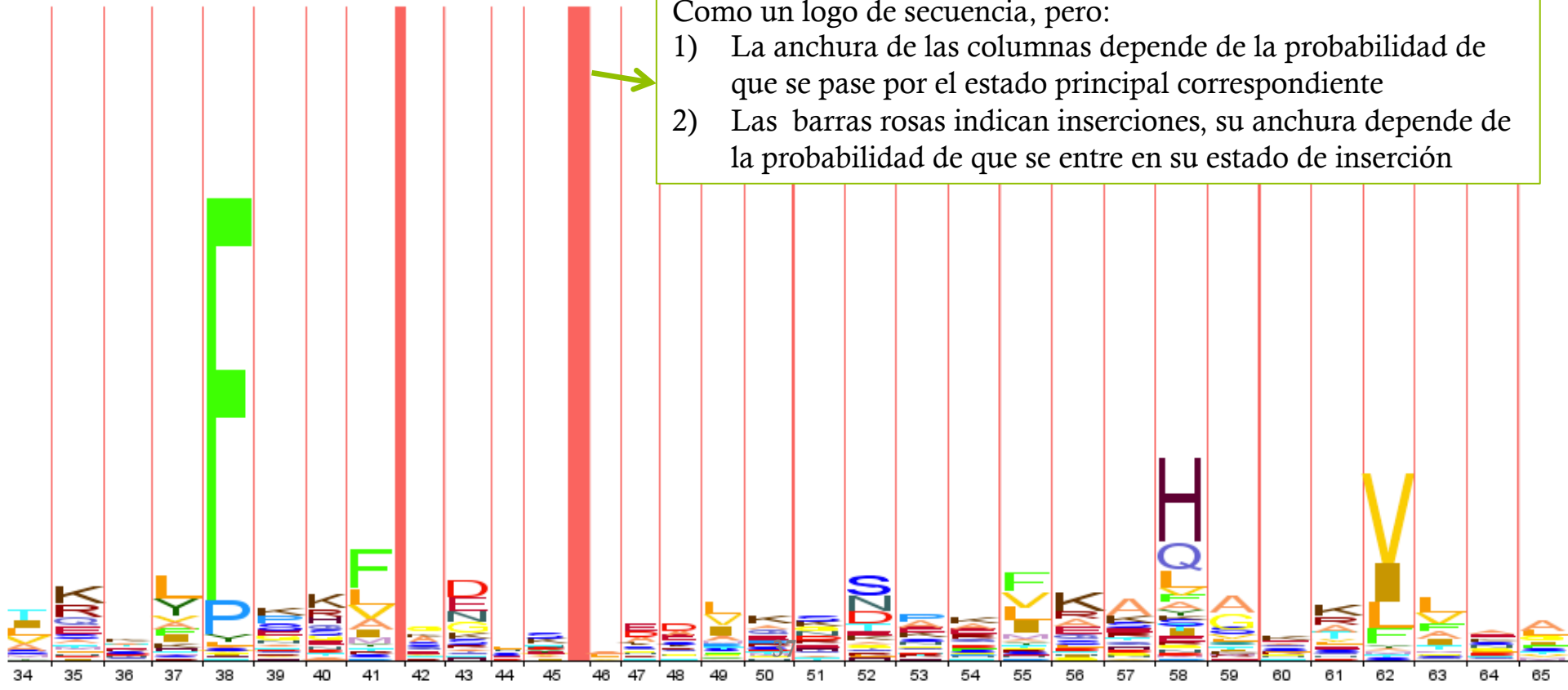
- $R_i = \log_2(s) - (H_i - e_n)$
  - s es el nº de elementos (4 para nucleótidos, 20 para aminoácidos)
  - n es el nº de secuencias en el alineamiento
- H_i es la incertidumbre (o entropía) de Shannon de la posición i

$$H_i = -\sum f_i * \log_2 f_i$$

  - f_i es la frecuencia relativa del residuo en la posición i
- e_n es una corrección al tamaño

$$e_n = \frac{s-1}{2 * \ln(2) * n}$$

Logo HMM



Como un logo de secuencia, pero:

1) La anchura de las columnas depende de la probabilidad de que se pase por el estado principal correspondiente
2) Las barras rosas indican inserciones, su anchura depende de la probabilidad de que se entre en su estado de inserción

Alineamientos de múltiples secuencias

Introducción
Algoritmos
Benchmarking
Visualización
Bases de Datos

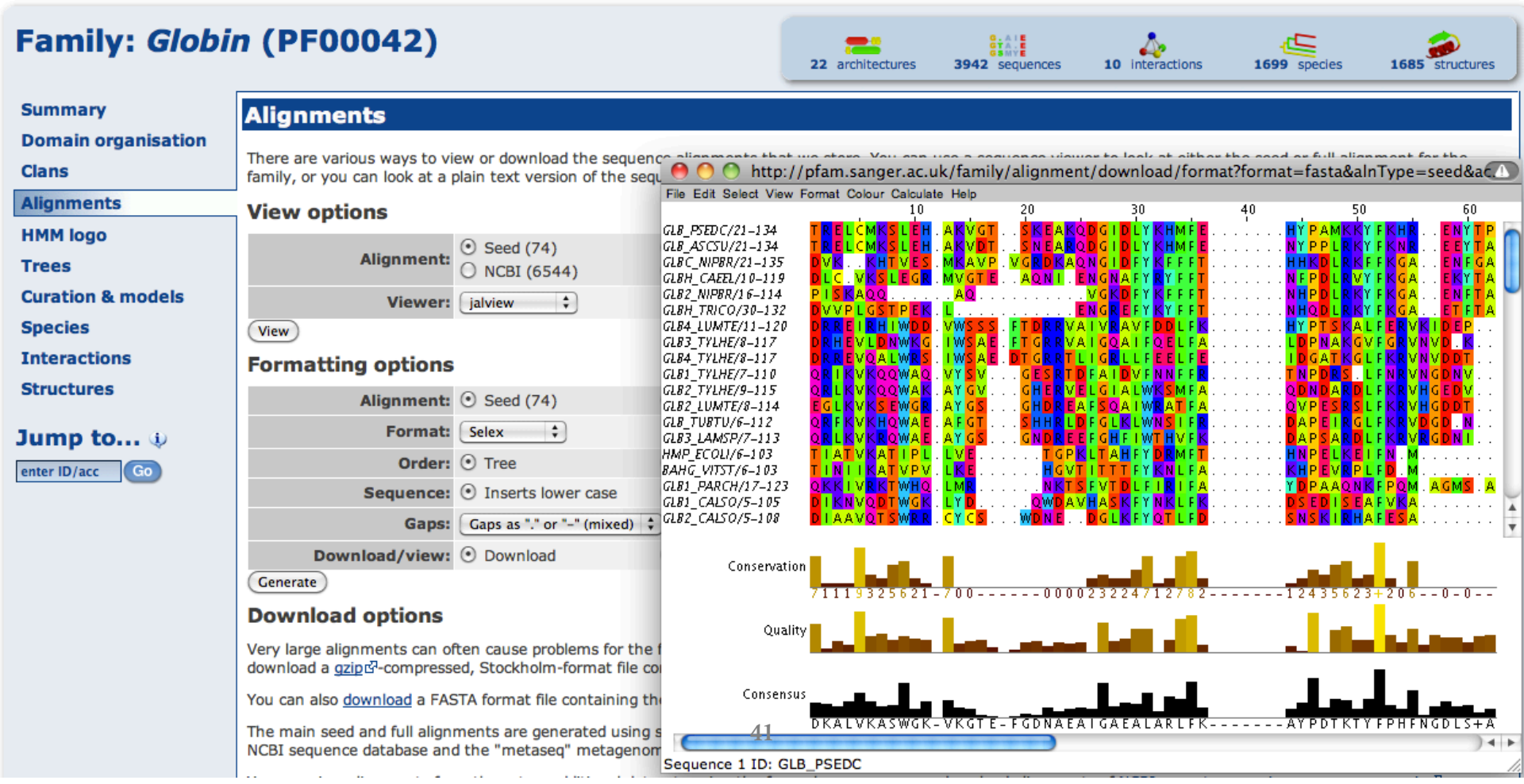
Bases de Datos

- Veremos algunas BBDD que almacenan información de familias de proteínas, juntos con sus MSA correspondientes
- Suelen ser BBDD consultables por secuencias o por texto (nombres de proteínas, de familias, dominios, etc.)
- Algunos ejemplos
  - Pfam
  - SMART
  - CDD

Pfam

- Protein Family Database
  - Colección de familias de proteínas, cada una representada por uno o más MSAs y perfiles HMM creados mediante HMMER
  - Se basa en los datos de Swiss-Prot y SP-TrEMBL
    - Aproximadamente cubre el 75% de sus entradas (2007)
- Desarrollado por el Sanger Institute (UK)
  - http://pfam.sanger.ac.uk/

Pfam: globinas



HMMER

- Herramienta de búsqueda de homólogos de proteínas en BBDD
  1. Toma como entrada un alineamiento múltiple de proteínas
  2. Construye su perfil HMM
  3. Busca en BBDD por homólogos que coincidan con el perfil HMM, y los alinea.

- En esencia es una herramienta similar a BLAST, pero con un fundamento estadístico más potente
  - A partir de HMMER3, presumiblemente tan rápido como BLAST

HMMER

biosequence analysis using profile hidden Markov models

SMART

- Simple Modular Architecture Research Tool
- BD de familias de proteínas implicadas en señales celulares, dominios extracelulares y la función de la cromatina
- Como Pfam, usa perfiles HMM creados mediante HMMER
- Mantenida por EMBL
  - http://smart.embl.heidelberg.de

CDD

- Conserved Domain Database
- Herramienta para búsquedas por secuencia o texto en Pfam o SMART
  - Con el propósito principal de identificar dominios conservados
    - Para ello utiliza RPS-BLAST, un método más sensible que BLAST (ver tema de BLAST)
- Mantenida por el NCBI
  - http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

InterPRO

- Recurso que integra la mayoría de las BBDD de alineamiento
  - Cada BD usa un algoritmo y unos métodos de búsqueda distintos
    - Unos usan HMMs, otros se centran en dominios, otros en motivos…
  - Esta BBDD integrada facilita la exploración de las características de una proteína desde múltiples puntos de vista
- El proyecto es un esfuerzo coordinado de ocho centros de investigación, siendo los principales el EBI y Sanger
  - http://www.ebi.ac.uk/interpro/
- Contiene alineamientos de Pfam, PROSITE, PRINTS, ProDom, SMART y TIGRFAMs

Alineamientos de múltiples secuencias

Introducción
Algoritmos
Benchmarking
Visualización
Bases de Datos
MSA de secuencias genómicas

MSA de secuencias genómicas

- Cada vez hay más genomas secuenciados
  - Compararlos puede ser útil para encontrar regiones que han cambiado dentro de un linaje (selección positiva) o que se conservan (selección negativa).
- Se utilizan modificaciones del alineamiento progresivo
  - Tratando de adaptarse a las particularidades del alineamiento de secuencias genómicas

Diferencias con el MSA convencional

- En el MSA convencional, comparamos muchas secuencias (100-1000) cortas (<1000 residuos)
  - En MSA genómico, comparamos pocas secuencias (varias decenas) con longitudes de millones de pares.
- Al comparar genomas de organismos muy distintos, encontramos islas bastante conservadas separadas por regiones muy poco conservadas.
  - Esto va a llevar al concepto de "anclajes" que veremos luego.

Diferencias con el MSA convencional

- Los genomas eukariotas están llenos de elementos repetitivos (por ejemplo, transposones) que ocupan regiones sustanciales del genoma.
- También existen recolocaciones (deleciones, duplicaciones, inversiones, translocaciones) que atañen a millones de pares.
- El MSA para genomas debe adaptarse a estas dos peculiaridades
- Además, todavía no existen datos de Benchmark para controlar la calidad de los alineamientos

Algoritmos

- Los algoritmos más usados para MSA genómico son:
  - TBA divide el genoma en bloques de secuencias que alinea mediante programación dinámica con MULTIZ
  - MLAGAN usa un alineamiento progresivo tipo ClustalW. Es una evolución del algoritmo LAGAN (alineamiento de pares)
  - MAVID usa también alineamiento progresivo al que se han añadido varias optimizaciones para secuencias genómicas largas
- El UCSC Genome Browser provee MSAs de ADN genómico para muchas especies, mediante estos tres programas.

Ejercicio

- Continuamos con el gen propuesto como "nuevo" en la sesión anterior
  - Buscar mediante BLAST secuencias homólogas en su especie
  - Realizar un estudio con distintos alineamientos múltiples de secuencia (distintos programas y parámetros), discutiendo los resultados
    - En el proceso, puede decidirse argumentadamente eliminar o añadir secuencias homólogas al alineamiento.

Resumen

- El alineamiento múltiple de secuencias (MSA) es el proceso por el que todos los miembros de una familia de proteínas o ADN se agrupan juntos
- Las filas corresponden a secuencias, y las columnas a residuos, los residuos alineados en la misma columna implican un ancestro común y/o una posición compartida en su estructura 3D
- Existen una gran cantidad de herramientas y aproximaciones al problema. La mayoría de ellas funcionan muy bien con secuencias similares (>40%) pero para secuencias distantes los resultados pueden variar mucho, sobre todo respecto a los huecos
- Para un usuario normal, se recomienda realizar MSAs con distintos programas, variando los parámetros de búsqueda (sobre todo con la penalización de huecos)
- Los algoritmos de MSA están cambiando con la tecnología, enfocándose ahora en el análisis de secuencias de ADN genómico, donde todavía no hay benchmarks para decidir cuál es la mejor opción
- Es una tendencia extendida el uso de bases de datos de MSAs (Pfam, InterPro), acompañadas de anotaciones de expertos, y con un enfoque en la integración de recursos

Preguntas para debate

- Feng y Doolittle introdujeron la regla de “una vez que hay un hueco, siempre hay un hueco”, indicando que las dos secuencias más parecidas que se alinean inicialmente deben tener más peso en la asignación de huecos. ¿Por qué es necesario introducir esta regla?

- ¿Cuáles son algunos de los problemas asociados con adaptar los programas de MSA a grandes regiones de ADN genómico?

Lecturas adicionales

- Pevsner, 2009: Ch 6 *Multiple Sequence Alignment*
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. *CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic Acids Res. 22, 4673–4680 (1994)
  - PMCID: PMC308517
- Larkin, M. A. et al. *Clustal W and Clustal X version 2.0.* Bioinformatics 23(21) 2947-2948 (2007)