Alineamiento múltiple de secuencias

Guillermo Barturen
gbarturen@ugr.es

Alineamiento múltiple

- Los alineamientos globales y locales son métodos válidos a la hora de buscar información en secuencias desconocidas por medio de la comparación individual con secuencias similares y/o con bases de datos.

- Sin embargo, el **estudio filogenético** de una secuencia o el estudio de la **funcionalidad de la estructura de una proteína** a partir de su secuencia lineal requiere de la comparación de múltiples secuencias de manera simultánea.

- De hecho, en algunos casos dos secuencias alejadas evolutivamente pueden presentar un mal alineamiento entre ellas, pero revelar regiones conservadas si se incluyen secuencias más cercanas en el mismo alineamiento.

Diferencias entre alineamiento por pares y múltiple

Pairwise Alignment	Multiple Sequence Alignment (MSA)
An alignment procedure comparing two biological sequences of either protein, DNA or RNA	An alignment procedure comparing three or more biological sequences of either protein, DNA or RNA
Pairwise alignments can be generally categorized as global or local alignment methods.	MSA is generally a global multiple sequence alignment
Comparatively simple algorithm is used	Complex sophisticated algorithm is used
A general global alignment technique is the Needleman–Wunsch algorithm. A general local alignment method is Smith–Waterman algorithm.	A technique called progressive alignment method is employed. In this approach, a pairwise alignment algorithm is used iteratively, first to align the most closely related pair of sequences, then the next most similar one to that pair, and so on.
Applications: a) Primarily to find out conserved regions between the two sequences. b)Similarity searches in a database	**Applications:** a) To detect regions of variability or conservation in a family of proteins b) Phylogenetic analysis (inferring a tree, estimating rates of substitution, etc.) c) Detection of homology between a newly sequenced gene and an existing gene family prediction of protein structure d) Demonstration of homology in multigene families
Examples of pairwise alignment tools: • LALIGN • BLAST • EMBOSS Needle • EMBOSS Water	**Examples of Multiple Sequence Alignment tools:** • MUSCLE • T-Coffee • MAFFT • CLUSTALW

Como abordar un alineamiento múltiple

1. El primer paso, y de suma importancia para poder resolver la pregunta que nos estemos haciendo, es seleccionar las secuencias que vamos a utilizar:

    - Buscar **secuencias homólogas** conocidas en las bases de datos
    - **Evitar redundancias** de manera que aumentemos la variabilidad pero sin introducir mucho ruido
    - En el caso de las **filogenias** es importante incluir una secuencia alejada del grupo a comparar (***outgroup***), para enraizarla y para determinar la dirección de los cambios en el grupo a comparar

2. Seleccionar el algoritmo que utilice una función de puntuación óptima para el problema a resolver

3. Ajustar los parámetros del algoritmo

4. Revisar visualmente los alineamientos, un alineamiento múltiple puede revelar mucha información incluso aunque el alineamiento no sea perfecto. Ya que las secuencias pueden cambiar pero la funcionalidad permanecer.

Tipos de alineamientos múltiples

Existen tres aproximaciones principales al alineamiento múltiple:

1. **Alineamiento exacto**

El alineamiento exacto se basa en programación dinámica, similar al algoritmo de Needleman visto en el alineamiento global. Es extremadamente lento aunque asegura un alineamiento óptimo.

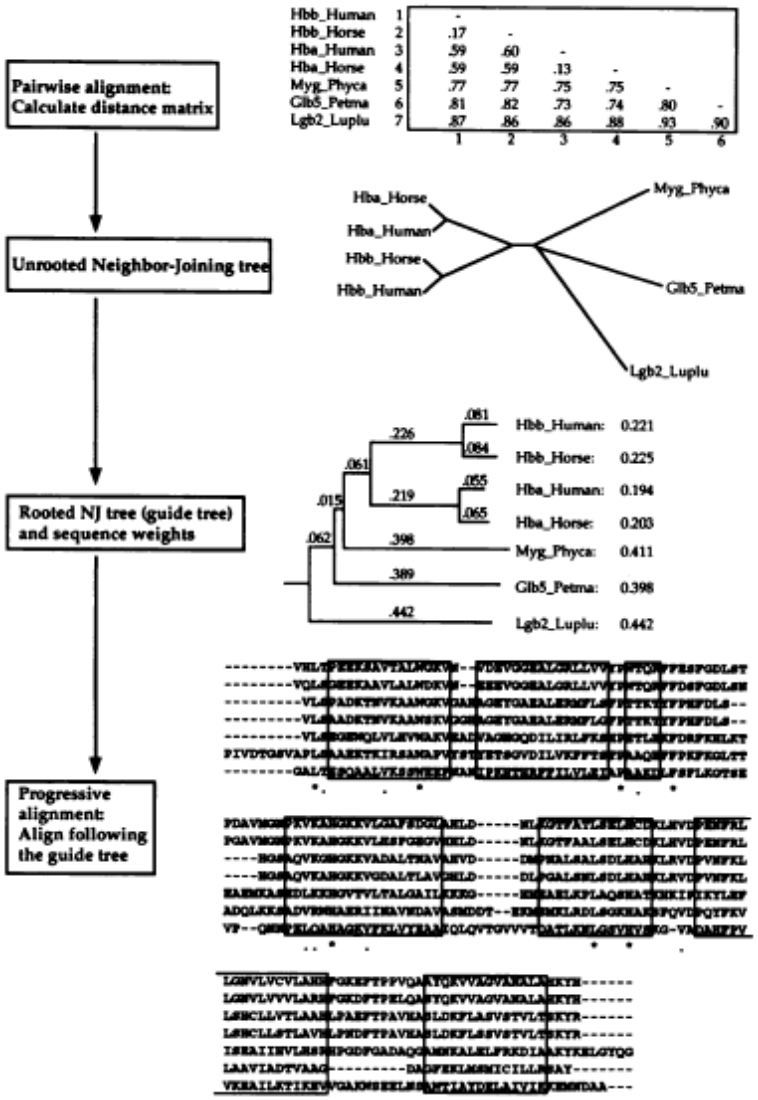
2. **Alineamiento progresivo**

El alineamiento progresivo es más rápido y puede usarse con múltiples secuencias, pero son métodos incompletos donde su solución no necesariamente es el alineamiento óptimo.

3. **Alineamiento iterativo**

Los alineamientos iterativos son alineamientos progresivos que revisan cada paso del alineamiento progresivo modificándolos en función de información adicional. En general, este tipo de alineamiento suele encontrar mejores soluciones que los alineamientos progresivos sin corrección.

Alineamiento progresivo (Clustal)



1. Alineamiento global por pares

2. Se transforma la similitud de secuencias a distancia mediante el algoritmo *Feng & Doolittle.*

$S(s_i, s_j)$ = Similitud entre secuencias

$S_{max}(s_i, s_j) = (S(s_i, s_i) + S(s_j, s_j)) / 2$

$S_{rand}(s_i, s_j)$ = Media similitud secuencias randomizadas

$$S_{eff}(s_i, s_j) = [S(s_i, s_j) - S_{rand}(s_i, s_j)] / [S_{max}(s_i, s_j) + S_{rand}(s_i, s_j)]$$

Distancia = $-\ln(S_{eff})$

3. Se genera un árbol guía uniendo las ramas de las secuencias con distancias más cortas, método "*neighbor-joining*" (Saitou & Nei). En cada paso, se calcula la suma de todas las ramas y se elige el par de OTUs que origina la suma mínima.

4. Se alinean las secuencias más cercanas en el árbol y se van añadiendo las más próximas progresivamente hasta llegar a la raíz del árbol (alineamiento progresivo).

Alineamiento iterativo (MUSCLE)

- El alineamiento mediante Clustal va heredando en cada sucesivo alineamiento progresivo los huecos introducidos en los pasos previos y no son corregidos, dando de esta manera al alineamiento una estructura de bloques.

- Los alineamientos iterativos cuentan con una gran precisión y rapidez, funcionan de manera **similar a los progresivos pero re-evaluan en cada paso los alineamientos anteriores**. Calculan una solución subóptima mediante un alineamiento progresivo y luego modifican el alineamiento mediante programación dinámica hasta que la solución converge.

- En el caso de MUSCLE, se **re-estiman las distancias entre las secuencias mediante el conteo de k-meros**.

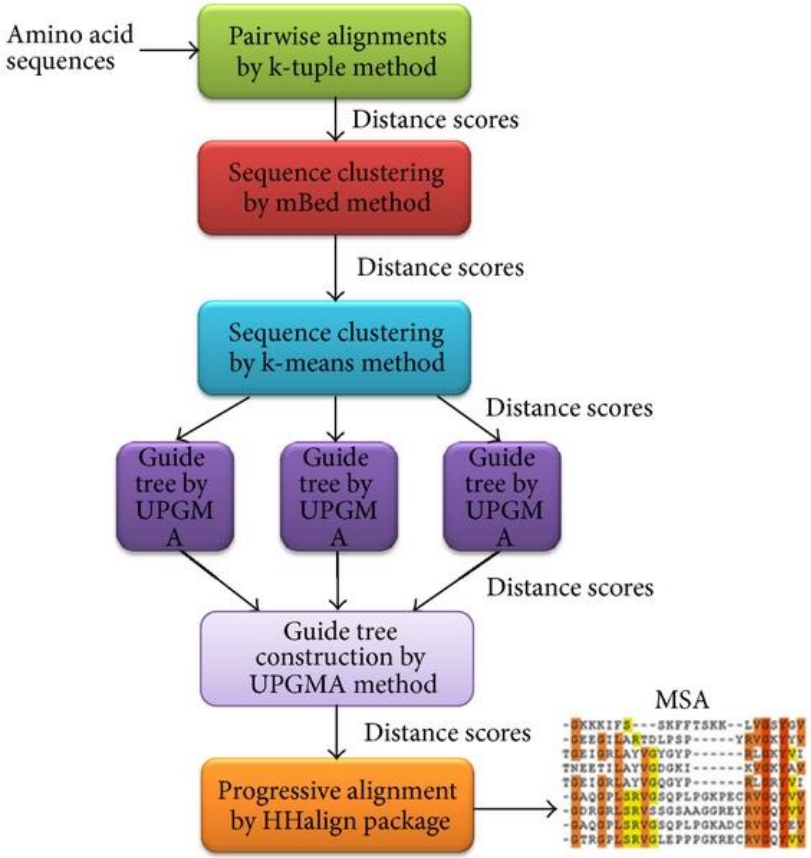
Otros alineamientos iterativos

- **Métodos basados en la consistencia (ProbCons / T-Coffee)**:
  - Incorporan información de las distintas secuencias en la creación de cada alineamiento de pares.
  - **T-Coffee** por ejemplo, **calcula todos los alineamientos de pares** globales entre secuencias, utilizando el algoritmo de Needleman. Después calcula los 10 alineamientos de pares locales con puntuación más alta. Con estas puntuaciones asigna puntuaciones a cada par de nucleótidos/aminoácidos alineados.
  - Realiza un **alineamiento progresivo utilizando estas puntuaciones en la fase de refinamiento iterativo**.

- **Métodos basados en la estructura (PRALINE / PipeAlign / Expresso):**
  - Estos métodos se basan en que las **estructuras terciarias de proteínas** evolucionan más lentamente que la estructura primaria.

Clustal Omega

Clustal Omega es la nueva versión del algoritmo clustal que realiza el **alineamiento múltiple en varios pasos seriados**, es extraordinariamente rápido mejorando la precisión de los alineamientos de su predecesor.



1. Genera todos los alineamientos por pares **basado en k-meros**.

2. Se agrupan las secuencias en base a dos métodos seriados mBed y k-means.

3. Utiliza el método UPGMA para construir los árboles filogenéticos, este es un método basado en **agrupaciones jerárquicas**.

4. Realiza un **alineamiento progresivo basado en modelos ocultos de markov (HMM)**.

Visualización de alineamiento múltiples

- Salida estándar

```
-------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
-----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
             :   :   :   : .. .    .      ::   *    *.

QFSSPEDCLSSPEFLDHIRKVMLVI---DAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS
NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHLDNLKGTFATLSELHCDKLH--VDPE
HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAEIKPLAQSHATKHK--IPVK
          . ..  *  .::        :                    :    :

SFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDGE----
HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE---
QFVVVKEALLKTIKAAVGDK-WSDELSRAWEVAYDELAAAIKKA--------
NFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH------
YLEFISECIIQVLQSKHPGD-FGADAQGAMNKALELFRKDMASNYKELGFQG
 :  :   ::  :        :      * .     .   :
```

- ∗ Conserved sequence (identical)
- : Conservative mutation
- · Semi-conservative mutation
- ( ) Non-conservative mutation
- \- Gap

- JalView