


Algoritmos de Comparación de Secuencias



Evolución Molecular



Evolución Molecular

Existen diferentes **mecanismos** que explican la **biodiversidad** (mutaciones, duplicación de genes, reorganización de genomas e intercambios genéticos como recombinación, reordenamiento y transferencia lateral de genes).

Las **mutaciones** son un **proceso aleatorio**, influenciado por una variedad de factores. **Distintas regiones del genoma presentan patrones y tasas de mutación diferentes.**

Los **alineamientos múltiples** son una **forma de establecer relaciones de homología** (origen evolutivo común) **entre un conjunto de secuencias**. Tienen en cuenta las mutaciones, inserciones y deleciones como principal mecanismo de cambio.

Evolución Molecular

Con el tiempo dos genes acumulan cambios, de modo que puede que los datos de la secuencia en sí mismos no contengan suficiente información sobre la relación entre los dos genes (hayan acumulado demasiada variación).

El término **homología** se usa solo cuando el antepasado común es lo **suficientemente reciente** como para que la información de la secuencia haya retenido suficiente similitud como para hacer inferencias evolutivas.

Los genes son homólogos o no lo son: aunque posean distintos porcentajes de similitud (número de nucleótidos o aminoácidos idénticos, en relación con la longitud de la secuencia).

Alineamiento de secuencias

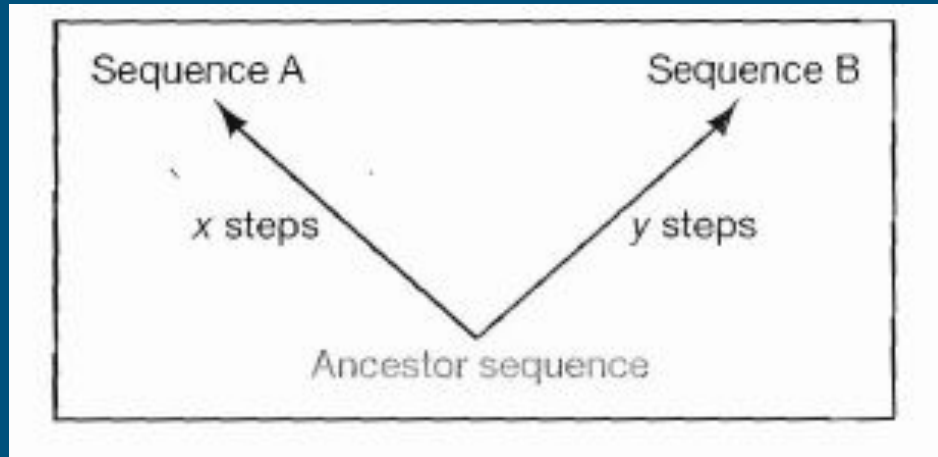
Los alineamientos sirven, entre otras cosas para: cuantificar similitud, encontrar dominios funcionales, buscar posiciones homólogas en las secuencias.

El objetivo del alineamiento es conseguir alinear las posiciones homólogas, con la mejor puntuación, y de la forma más razonable desde un punto de vista biológico.

Cuando hay un cambio de un residuo por otro decimos que hay una sustitución. Cuando falta una base decimos que hay un "gap" (puede corresponder tanto a una deleción como a una inserción).

Alineamiento de secuencias

Un alineamiento implica una hipótesis de evolución



Alineamiento de secuencias

El alineamiento secuencial es un procedimiento por el cual podemos comparar 2 (alineamiento entre pares, **pairwise alignment**) o más secuencias (alineamiento múltiple, **multiple alignment**). Este procedimiento consiste en buscar series de caracteres individuales que se encuentran en el mismo orden en las secuencias a comparar.

Colocación de dos secuencias para que se maximice su similitud. Los caracteres idénticos se ubican en la misma columna (**match**), mientras que los caracteres no-idénticos se pueden ubicar en la misma columna (**mismatch**) o bien alineados con lo que llamamos “gap” (**indel**).

Métricas

Métricas para puntuar el emparejamiento de residuos en secuencias distintas:

- **Puntuaciones que consideran el % de identidad de secuencia o coincidencias:** no todas las sustituciones deben tratadas de la misma manera (no es ideal ya que sabemos que las purinas (A,G) y pirimidinas (C,T), o los residuos aromáticos si hablamos de proteínas, no se intercambian de la misma manera en la evolución).
- **Puntuaciones que tienen en cuenta los gaps** (penalizaciones para los gaps, una para abrir el gap y otra para extenderlo, este último suele ser menos costoso)

ACCTGATCCG

| | | | | | |

AC-TGATCAG

$S=8-4-3=1$

ACCTGATCCG

| | | | |

ACTGA-TCAG

$S=5-4-12=-11$

Identidad != Similitud

TGAAGTA-ACT
TCATGTACACT

Identidad: 1+0+1+0+1+1+1+0+1+1+1 = **8**

Similitud: 1-2+1-1+2+1+1-4+1+2+1 = **3**

C→C, G→G +2

A→A, T→T +1

A→T, T→A -1

C→G, G→C -2

OTROS -2

GAP -4

Las tablas de puntuaciones de sustitución de un residuo por otro se denominan
Matrices de sustitución.

[illegible]
$$R_{ij} = \frac{M_{ij}}{f_j}$$

Matriz PAM70 para 23 aminoácidos. Matrices PAM (mutación puntual aceptada). 'Atlas of Protein Sequence and Structure' Margaret Oakley Dayhoff.

Matrices de sustitución

</

Matrices BLOSUM (BLOCKS of Amino Acid SUBstitution Matrix) con los bloques de secuencias alineadas se calcula una tabla de frecuencias de cada pareja de aminoácidos alineados, obteniendo 210 parejas posibles con sus respectivas frecuencias de aparición que permitirán calcular los odd-ratios (R_{ij}) entre las frecuencias observadas (q_{ij}) y las frecuencias esperadas por casualidad (e_{ij})

$$R_{ij} = \frac{q_{ij}}{e_{ij}}$$

Creadas a partir de alineamientos de secuencias, para mejorar los alineamientos de secuencias divergentes donde las matrices PAM fallan.

Técnicas y programas de alineamiento

Dos tipos posibles de Alineamientos:

- **Global:** alineamiento de la **secuencia completa**. Es **útil cuando se comparan** secuencias muy similares en tamaño y composición, por ejemplo de **dos genes muy conservados**.
- **Local:** cuando sólo nos interesa alinear **regiones similares** entre secuencias. Se utiliza **cuando las secuencias a comparar son diferentes** en tamaño o poseen regiones no conservadas

Técnicas y programas de alineamiento

Local	Global
Mejor alineamiento a lo largo de 2 secuencias	Mejor alineamiento del segmento más largo entre 2 secuencias
Smith Waterman	Needleman Wunsch
Búsquedas de sub-secuencias o regiones	Tienen la misma organización secuencial?
Permite localizar segmentos o dominios comunes	Sirve para proteínas globalmente similares

Ejemplo: Needleman-Wunsch

Este algoritmo que permite encontrar el alineamiento global de puntuación máxima.

Es un ejemplo de algoritmo de programación dinámica: subdivisión de problemas, asegura encontrar la solución óptima o el mejor alineamiento global para 2 secuencias.

Utiliza una matriz cuadrada para asignar puntuación para los distintos alineamientos posibles, dada una puntuación para matches, mismatches y gaps

Ejemplo: Needleman-Wunsch

Este funciona en base a un sistema de puntuaciones de cuán parecidas son dos secuencias. Por ejemplo:

- +1 por cada elemento igual (match)
- 1 por cada elemento desigual (mismatch)
- 1 por cada hueco introducido (gap)

Esta es una puntuación muy simple que se usará sólo para ilustrar el funcionamiento del algoritmo, en la realidad se utilizan las matrices de sustitución que ya hemos introducido (PAM , BLOSUM, etc)

Ejemplo: Needleman-Wunsch

1° Armamos una matriz de comparación de las dos secuencias a alinear.

2° Comenzando desde 0, de arriba hacia abajo y de izquierda a derecha, se calculan tres valores, que son la suma de una celda adyacente más el match/mismatch (MM) de la celda actual

- MM + celda superior
- MM + celda izquierda
- MM + celda superior izquierda

A = A entonces
MM = +1

+1 - 1 = 0
+1 - 1 = 0
+1 + 0 = +1

NOS QUEDAMOS CON EL MAYOR !!

		A	H	C	N	I	R	V	S
	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	+1							
I									
C									
I									
N									
R									
C									
K									

Ejemplo: Needleman-Wunsch

3° Retrocedemos de abajo hacia arriba y de derecha a izquierda, buscando el camino de mayor score:

- Comparamos el primer casillero de abajo a la derecha, con los adyacentes.
- Nos movemos hacia al casillero de mayor score.
- Cuando existen múltiples caminos posibles se exploran todos y se comparan los scores globales.

		A	H	C	N	I	R	V	S
	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	+1	0	-1	-2	-3	-4	-5	-6
I	-2	0	0	-1	-2	-1	-2	-3	-4
C	-3	-1	-1	+1	0	-1	-2	-3	-4
I	-4	-2	-2	0	0	1	0	-1	-2
N	-5	-3	-3	-1	1	0	0	-1	-2
R	-6	-4	-4	-2	0	0	1	0	-1
C	-7	-5	-5	-1	-1	-1	0	0	-1
K	-8	-6	-6	-2	-2	-2	-1	-1	-1

Ejemplo: Needleman-Wunsch

4° Se construye el alineamiento teniendo en cuenta que:

- Cuando nos movemos en diagonal se alinean los dos caracteres.
- Cuando me muevo hacia abajo se introduce un gap en la secuencia horizontal.
- Cuando me muevo hacia la derecha se introduce un gap en la secuencia vertical.

Alineamos de arriba hacia abajo y de izquierda derecha.

```
AHC-NIRVS
| | | | |
A I C I N-RCK
```

		A	H	C	N	I	R	V	S
	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	+1	0	-1	-2	-3	-4	-5	-6
I	-2	0	0	-1	-2	-1	-2	-3	-4
C	-3	-1	-1	+1	0	-1	-2	-3	-4
I	-4	-2	-2	0	0	1	0	-1	-2
N	-5	-3	-3	-1	1	0	0	-1	-2
R	-6	-4	-4	-2	0	0	1	0	-1
C	-7	-5	-5	-1	-1	-1	0	0	-1
K	-8	-6	-6	-2	-2	-2	-1	-1	-1

Ejemplo: Clustal

Alineamiento progresivo:

- Calcula alineamientos de pares entre las secuencias consideradas
- Elige el mejor alineamiento de entre ellos
- Añade progresivamente más secuencias al alineamiento

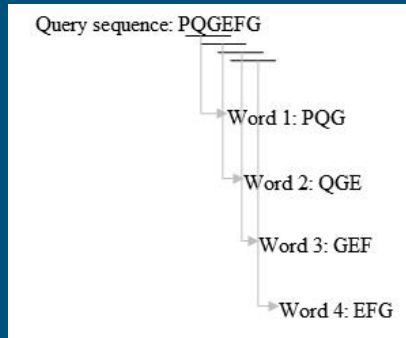
Clustal implementa un algoritmo que consta de 3 fases:

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. Haciendo alineamientos de pares pero según las distancias

Búsqueda de similitud secuencial: BLAST

- Prioriza la velocidad a la sensibilidad (es 50 veces más rápido que SW)
- Se basa en la comparación de palabras o k-tuples
- K -t uples = 3 para proteínas y 11 para ácidos nucleicos

El BLAST realiza un alineamiento y deriva un score de similitud entre las secuencias “query” y cada una de las secuencias contenidas en la base de datos (~250 millones de secuencias).



1° Divide la secuencia query en K cantidad de palabras cortas y asume que un alineamiento es significativo si tiene estas palabras

2° Una vez que encuentra los alineamientos posibles, los extiende y calculando un puntaje, extiende hasta que éste desciende por debajo de un punto de corte., se queda con los de mejor puntaje.

3° Evaluación estadística

Resultados del BLAST

Results (output) of BLAST

Bit-score	E-value	Identity (%)	Similarity (%) Positive score in the substitution matrix	Gaps (%)
Score = 83.6 bits (205), Expect = 3e-14, Method: Compositional matrix adjust.				
Identities = 61/136 (44%), Positives = 73/136 (53%), Gaps = 18/136 (13%)				
Query	184	KPKPKQYPKPVILPSNSTRRISPVTAKTSSSAEGVVVASESPVIAHPGSSSHSRSLSKRRSS	243	
		KP P P+ ILPSN+ +R P S V+ AS+SPVI P+ + RS		
Sbjct	269	KPAPG-LPRFILPSNNPQRQLPPPPSDS-----VIHASQSPVIKPNYAGKPPGFVSARSV	322	
Query	244	GALVDDD-----KRESKHAEQARRNRLAVALHELASLIPA EWKQQNVSAAPSKATT	295	
		L D K+E HK AEQ RRNL AL EL L+P E K+ + PSKATT		
Sbjct	323	RTLSGGDANTGDEFIKVEVHKVAEQGRNRNLNNALAE LNDLLPPELKES--AQVPSKATT	380	
Query	296	VEAACRYIRHL--QQN	309	
		VE AC+YIR L QQN		
Sbjct	381	VELACKYIRQLTGQQN	396	

Resultados del BLAST

E-Value y P-value:

- P-value: dado un score x , el P-value es la probabilidad de encontrar un score S mayor a x .
- E-value: para un alineamiento con un score x , el E-value es el número de alineamientos encontrados al azar en la base de datos de tamaño N , con un score mayor a x .

$$\text{E-value} = \text{P-value} * N$$

Tipos de BLAST

