

CURSO: CC471 - 2018 -1

Examen Parcial.

Apellidos:_____ Nombres: _____Codigo:_____

Tiempo: 2 Horas

1. Cuales son las diferencias entre el algoritmo Needleman Wunch y el de Smith Waterman

R:

a). Needleman Wunch se usa para alineamientos globales y Smith Waterman para alineamientos locales

b). La inicializacion de la matriz de scoring. En Smith Waterman la primera columna y la primera fila se inicializan en 0, mientras que en Needleman Wunch se inicializan con factores del valor de gap, empezando en cero.

c). En Smith Waterman los valores de scoring no pueden ser menores que cero.

d). En Smith Waterman la ruta de alineamiento en la fase de backtrace empieza con el mayor valor de la matriz y termina en el primer 0 encontrado

2.Utilice el algoritmo Needleman Wunch para alinear a mano las siguientes secuencias:

a) GATAGC

b) GTAC

Cuales son score y alineamiento optimos si la penalidad por gap es -1, por mismatch es 0 y el puntaje por match es 1?. Presente la Matriz de scoring y el camino de backtrack

R:

Match=1
Mismatch=0
Gap=-1

Needleman Wunch

Gap=-1
Mismatch=0
Match=1

$F(i,j)$

$F(i-1,j)+g$
 $F(i-1,j-1)+s(i,j)$
 $F(i,j-1)+g$

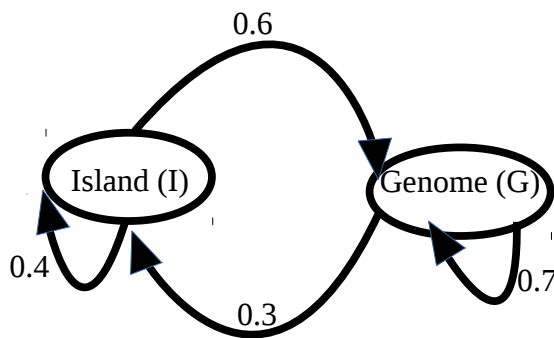
Alineamiento:

		G	T	A	C
	0	-1	-2	-3	-4
G	-1	1	0	0	0
A	-2	0	0	1	0
T	-3	0	1	0	0
A	-4	0	0	1	0
G	-5	1	0	0	0
C	-6	0	0	0	1

G -- T A -- C
G A T A G C

Scoring: +1-1+1+1-1+1=2

3. En la siguiente HMM (Hidden Markov Model) identifique cual es la secuencia mas probable de estados internos para la siguiente secuencia de simbolos emitida : CAT



$$P_1(S) = \begin{bmatrix} (I) & (G) \\ 0.5 & 0.5 \end{bmatrix}$$

$$P(X|S=(I)) = \begin{bmatrix} C & G & A & T \\ 0.4 & 0.1 & 0.1 & 0.4 \end{bmatrix}$$

$$P(X|S=(G)) = \begin{bmatrix} 0.2 & 0.3 & 0.3 & 0.2 \end{bmatrix}$$

El estado inicial mas probable para el primer simbolo de la secuencia “C”:

$$P(X_1=C | S_1=(G)enome) = P_1 * P(C|S=G) = (0.5) * (0.2) = 0.10$$

$$P(X_1=C | S_1=(I)sland) = P_1 * P(C|S=I) = (0.5) * (0.4) = 0.2$$

Para el segundo simbolo de la secuencia “A”:

$$P(X_2=A | S_1=(G)enome) = 0.10 * 0.7 * 0.3 = 0.021$$

$$P(X_2=A | S_1=(I)sland) = 0.2 * 0.4 * 0.1 = 0.008$$

$$P(X_2=A | S_1=(G)enome) = 0.10 * 0.3 * 0.1 = 0.003$$

$$P(X_2=A | S_1=(I)sland) = 0.2 * 0.6 * 0.3 = 0.036$$

Para el tercer simbolo “T”:

$$P(X_3=T | S_2=(G)enome) = 0.036 * 0.7 * 0.2 = 0.00504$$

$$P(X_3=T | S_2=(I)sland) = 0.008 * 0.4 * 0.4 = 0.00128$$

$$P(X_3=T | S_2=(G)enome) = 0.036 * 0.3 * 0.4 = 0.00432$$

$$P(X_3=T | S_2=(I)sland) = 0.008 * 0.6 * 0.2 = 0.00096$$

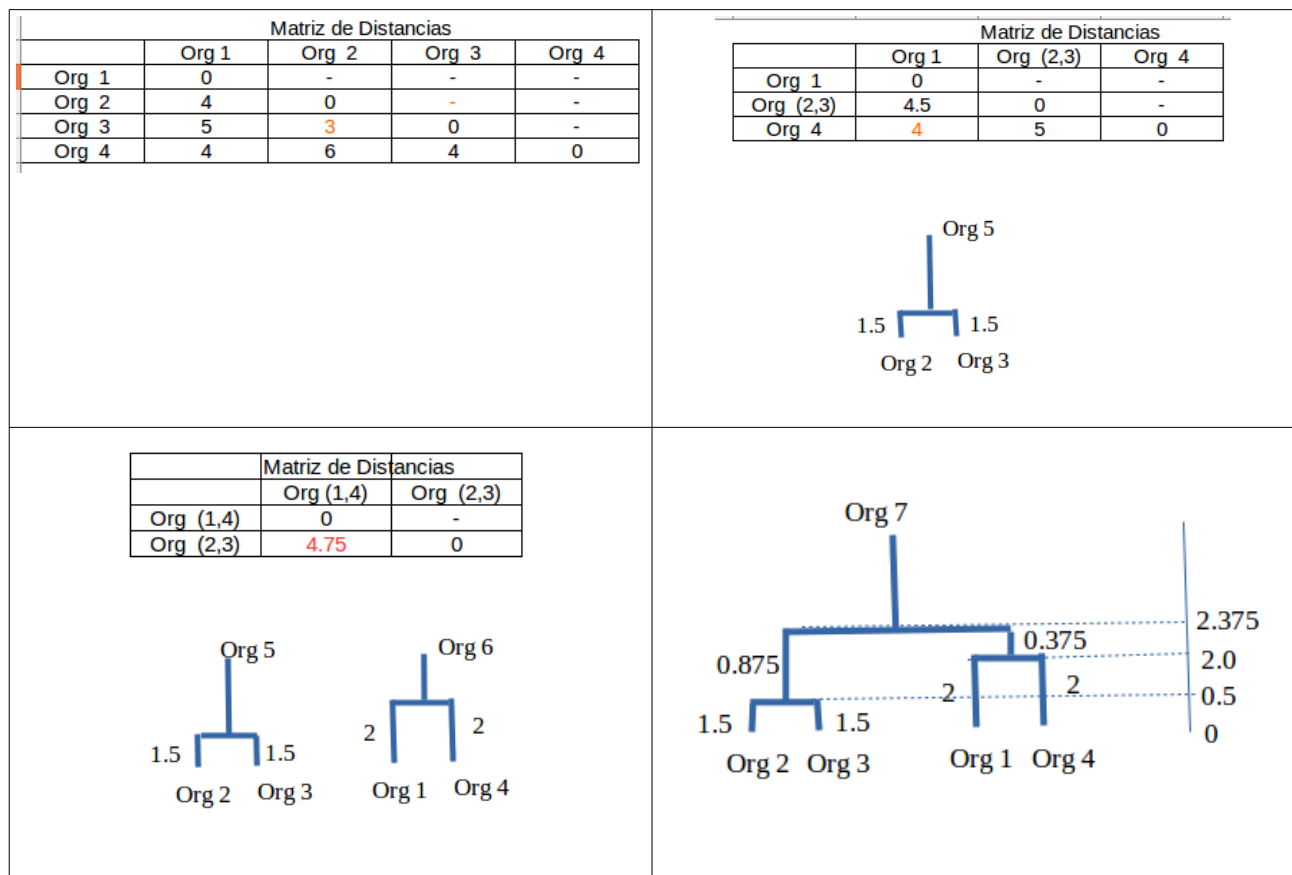
(Emitidos)	C	A	T
(Genome)	0.1	0.036	0.00504
(Island)	0.2	0.008	0.00432

La secuencia más probable de estados internos: (Island, Genome, Genome)

4. Dado el siguiente alineamiento de secuencias de 4 organismos, genere la matriz de distancias entre cada uno de ellos – Distancia de Hamming-. Una vez obtenida la matriz de distancias utilice UPGMA para generar un árbol que las represente. Presente el Árbol incluyendo las longitudes de sus ramas.

Org1.ACCTCACGTTATCC
 Org2.ACCTCGCGTGAGCG
 Org3.ACGTCGCCCTTAGCG
 Org4.AGGTCACATTATCG

R:



5. Explique detalladamente que significa y para que sirve

- PAM 250
- BLOSUM62

R:

a)

- Las Matrices PAM (Point Accepted Mutation) se utilizan en algoritmos de alineamiento de Proteínas.
- Se derivan de un grupo de secuencias que son > del 85% milares.
- fue diseñado para detectar los orígenes evolutivos de las proteínas.
- Point Accepted Mutation – Una mutación en un aminoácido que es aceptada por que no cambia la funcionalidad.
- 1 PAM es la matriz de probabilidades de que 1 aminoácido sustituido por cada 100.
- PAM250 se obtiene multiplicando PAM1 por si misma 250 veces. Generalmente PAM250 se utiliza para analizar secuencias muy distintas.

b)

-Las matrices BLOSUM (Block Substitution Matrix) se utilizan en algoritmos de Alineamiento de proteínas.

-Están basadas en comparaciones de secuencias de una base de datos denominada Blocks que contiene segmentos sin gaps en alineamientos múltiples de proteínas (Las regiones más conservadas de las proteínas)

BLOSUM62 se deriva de BD Blocks que contienen secuencias que son más del 62% idénticas en un alineamiento sin gaps.

-BLOSUM62 es la matriz por defecto para el programa estándar BLAST de proteínas.