

Análisis de Secuencias I y II

Curso de Biofísicoquímica Macromolecular

Ricardo Cabrera

ricabrer@uchile.cl

23 de Noviembre de 2011

¿qué es un alineamiento de secuencias?

Es una matriz

Posición

Residuos homólogos:
equivalentes o comparables en diferentes proteínas

“Carácter “ que posee 21 estados posibles

1 2 3 4

Secuencias

Proteínas
homólogas

Baboon	V	S	K	P	L	V	P	A	S	F	M
Bushbaby	A	V	K	P	L	V	P	A	S	L	N
Sheep	A	S	K	P	L	V	P	A	S	V	N
Cow	V	S	K	P	L	V	P	A	S	F	M
Pig	V	S	K	P	L	V	P	A	S	F	M
Marsupial	G	D	S	P	K	M	P	V	S	N	



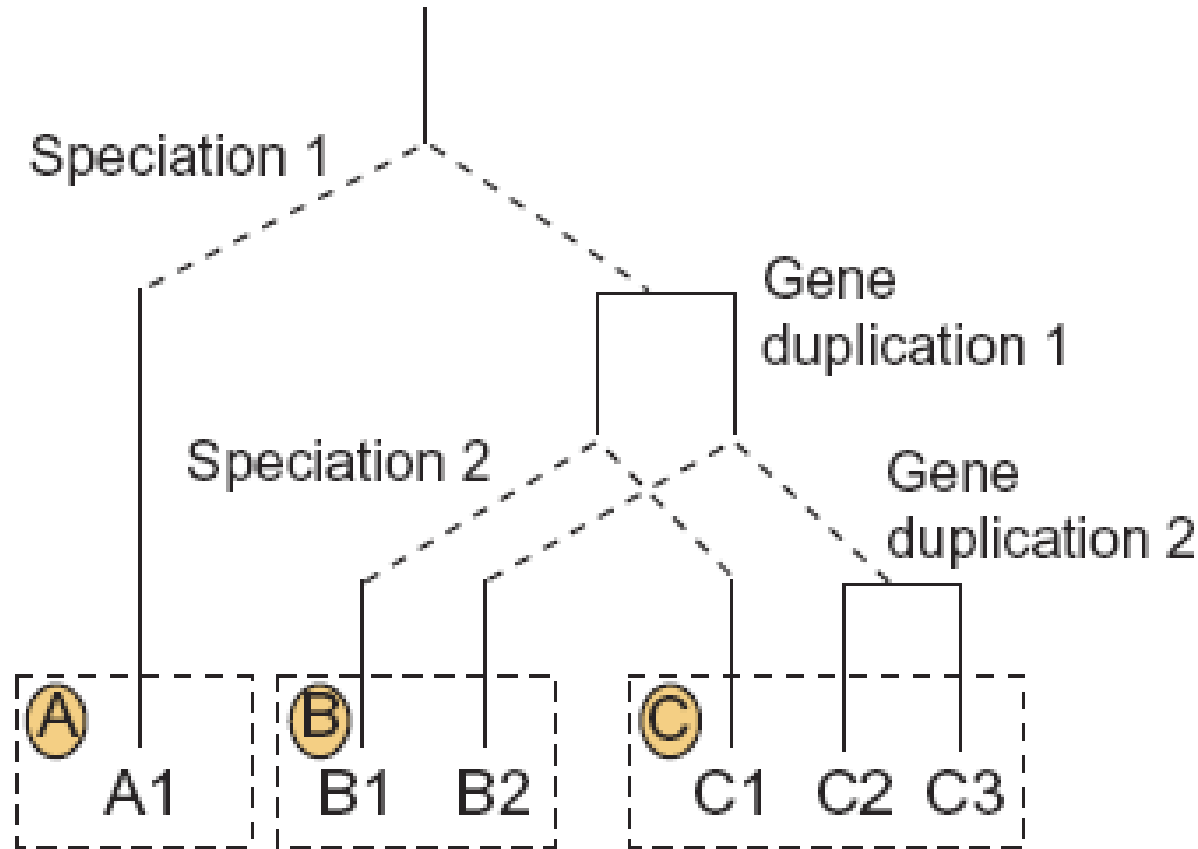
Es una manera de representar los residuos equivalentes en
proteínas homólogas

¿cómo sé que dos proteínas son homólogas?
(¿cómo sé que dos proteínas son alineables?)

¿cómo sé que dos residuos en proteínas homólogas son equivalentes?

Es una manera de representar los ***residuos equivalentes*** en
proteínas homólogas

Duplicación y especiación. Tipos de homología



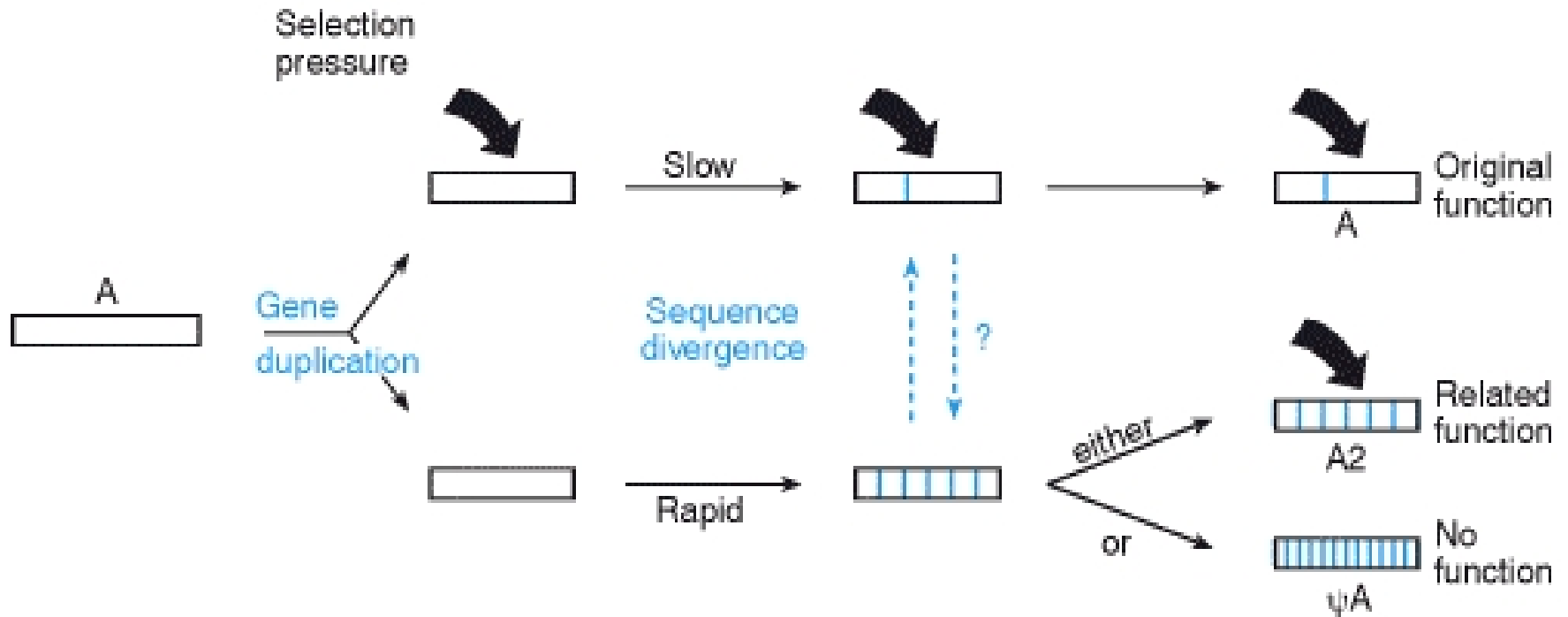
R. Owen fue el primero en definir homología como la propiedad del “mismo órgano que se encuentra en diferentes animales bajo cada variedad de forma y función”

4 eventos de divergencia: dos de especiación y dos de duplicación, resultando en seis genes contemporáneos en tres organismos A, B y C.

A1 tiene 3 ortólogos en la especie C, pero sólo C1 es ortólogo de B1. B2 tiene 2 ortólogos en la especie C (C2 y C3), en tanto que B2 y C1 son parálogos. Los tres genes en C son parálogos.

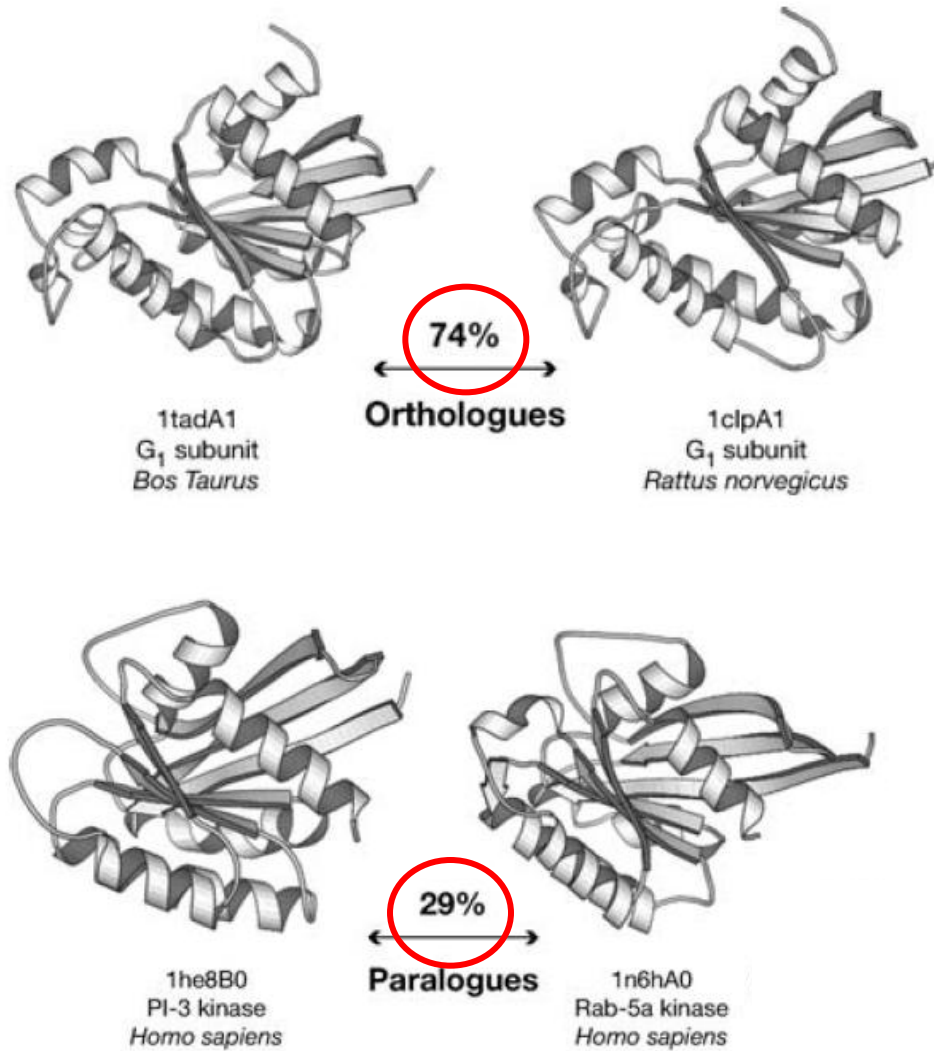
Un determinado gen puede tener más de un ortólogo en una especie y dos genes parálogos no necesariamente están restringidos a pertenecer a la misma especie.

Qué les sucede a las secuencias de los genes después de una duplicación?



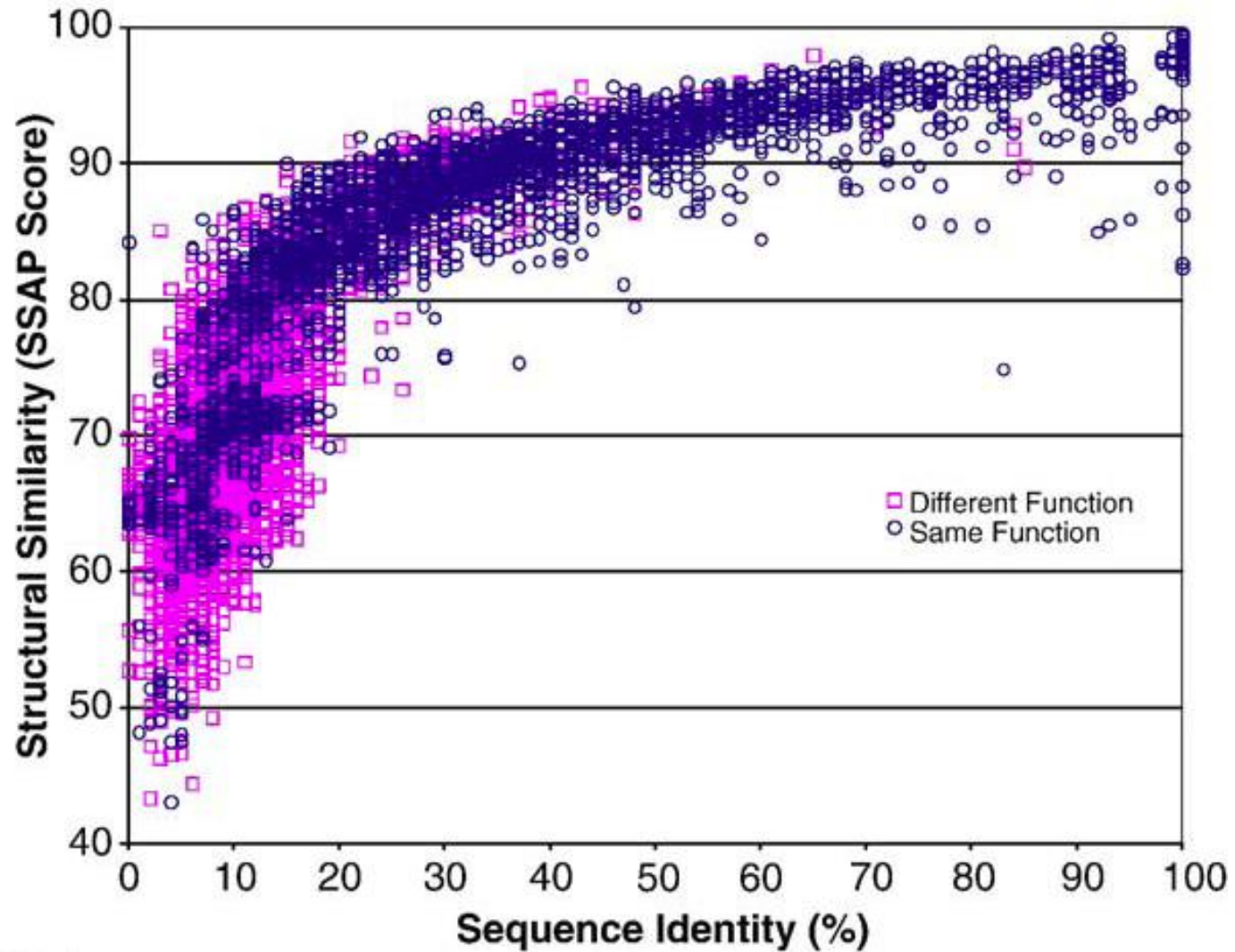
El potencial de adquirir una nueva función puede depender de la capacidad de un plegamiento de aceptar cambios sin “desarmarse”

Divergencia estructural de Ortólogos y Parálogos



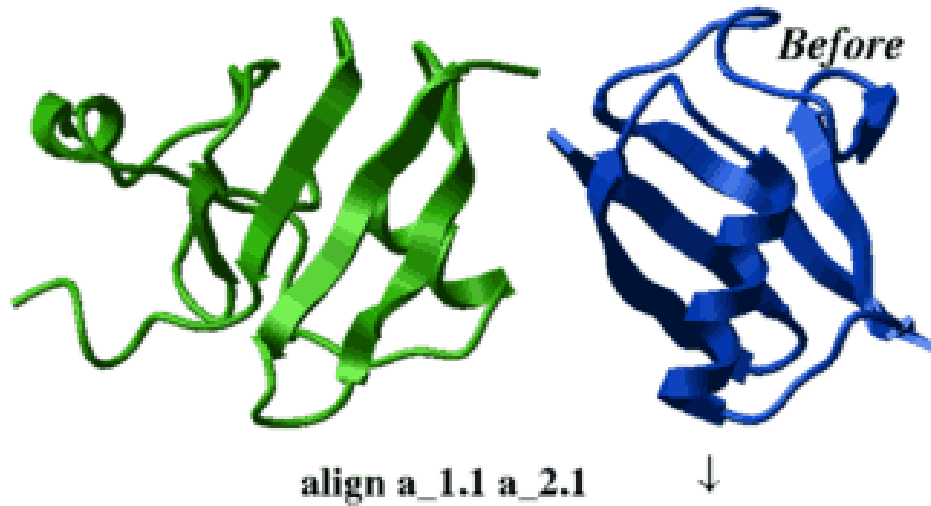
% identidad de secuencia

Divergencia de secuencia versus divergencia de estructura



Superposición estructural

Métodos para establecer equivalencias entre 2 o más estructuras de proteínas sobre la base de su conformación tridimensional

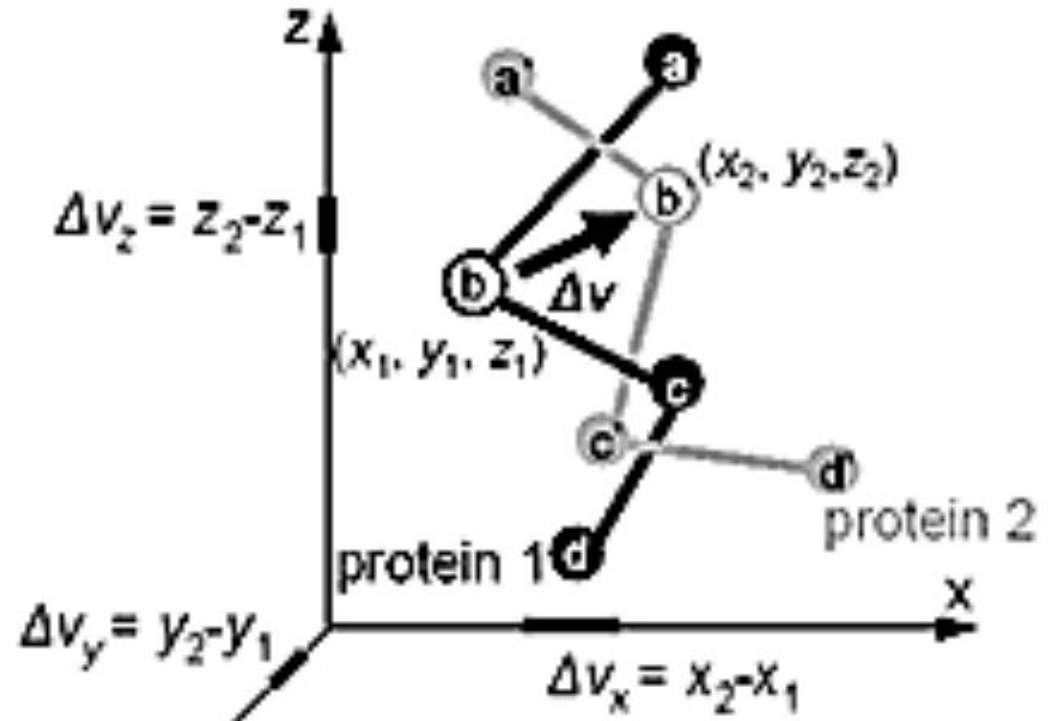


Superposición estructural

Un índice de la semejanza entre dos estructuras luego de una superposición.

$$\text{RMSD} = \left(\frac{1}{n} \sum_n d_n^2 \right)^{1/2}$$

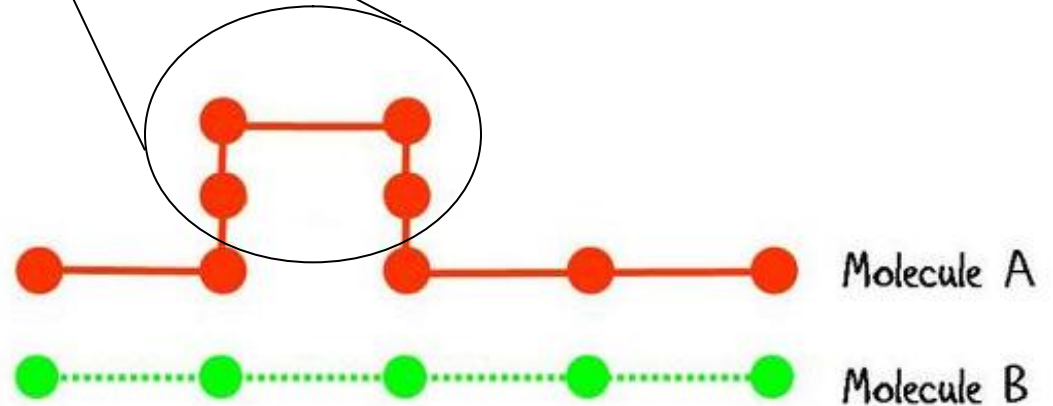
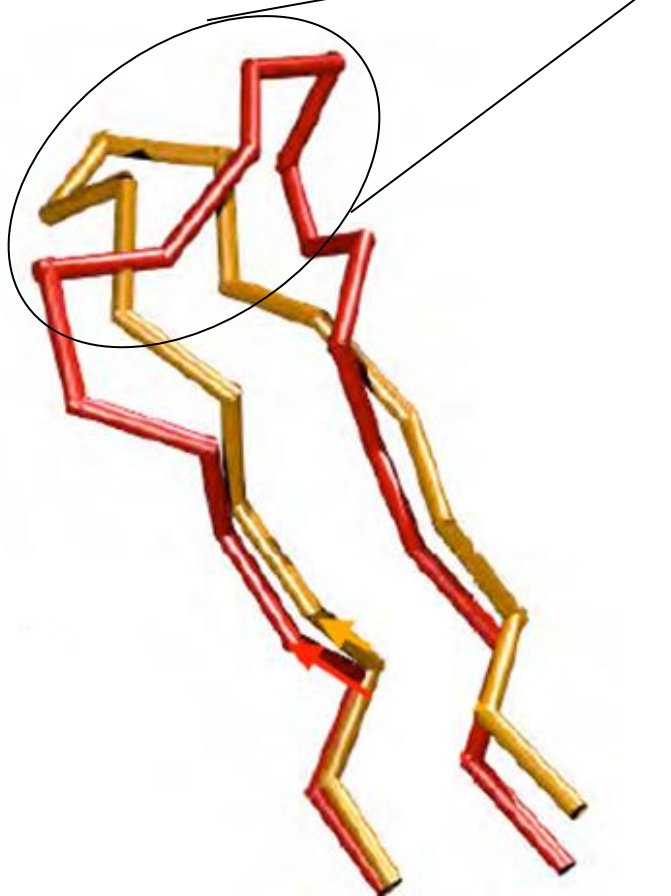
d , representa la distancia entre dos átomos superpuestos y n , el número de residuos superpuestos.



Generalmente comparamos los átomos de la cadena principal o, por simplicidad y eficiencia, sólo los carbonos- α

Alineamientos provenientes de una superposición estructural

Ausencia de correspondencia estructural entre ciertos residuos



Alineamiento de secuencias

Superposición estructural

Tipos de mutaciones:

A) Mutaciones silentes que modifican la secuencia de DNA codificante en una base que no cambia el aminoácido codificado por ese codón; o mutaciones en regiones regulatorias que no alteran la expresión de ningún gen. B) Mutaciones sin sentido que alteran un triplete generando un codón de término dejando una proteína incompleta. C) Mutaciones de cambio de sentido que cambian el aminoácido codificado generando una proteína no funcional con su función modificada. D) Mutaciones de cambio de fase producidas por la inserción o delección de cualquier número de bases que no sea múltiplo de tres que generan una modificación de la secuencia aminoacídica a partir de la inserción (o delección) resultante.

sustituciones sinónimas.- cuando el aminoácido tiene propiedades similares que no afectan la función de la proteína

sustituciones no sinónimas.- el aminoácido modificado tiene propiedades fisicoquímicas muy distintas

Alineamiento de secuencias

- Alineamiento 1 (sin indel):

```
GCGCATGGATTGAGCGAGGAAG
TGCGCCATTGATGACCATGACA
```

- Alineamiento 2 (con indel):

```
-GCGC-ATGGATTGAGCGAGGAAG
TGCGCCATTGAT-GACC-ATGACA
```

conservación (match)

cambio (mismatch)

Inserción-delección (Gap)

Criterio para elegir entre los varios posibles alineamientos:

FUNCIÓN DE *SCORING*

$$S = N_m - N_{mm} - N_g$$

[m = match; mm = mismatch; g = gap]

$$S = N_m V_m + N_{mm} V_{mm} + N_g V_g$$

V_g

toma en cuenta que los tres tipos de equivalencias no pueden sopesarse de la misma manera, por ejemplo, tener un *mismatch* a un *gap*. Esto ya que un proceso de mutación puede ser más probable que un proceso de inserción-delección

Criterio para elegir entre los varios posibles alineamientos:

FUNCIÓN DE *SCORING*

	A	G	T	C
A	2	-1	-1	-1
G	-1	2	-1	-1
T	-1	-1	2	-1
C	-1	-1	-1	2

	A	G	T	C
A	2	-1	-3	-3
G	-1	2	-3	-3
T	-3	-3	2	-1
C	-3	-3	-1	2

Los valores V se extraen de una ***matriz de sustitución***.

Esta es una matriz simétrica, de tamaño $A \times A$ (donde A es el tamaño del alfabeto del cual están hechas las secuencias, en el caso de DNA, $A = 4$), donde el coeficiente MS_{ij} es el score para alinear un carácter i con un carácter j .

Sus valores pueden ser arbitrarios, pero en general dan cuenta de la probabilidad de cambio de un tipo de carácter a otro. En el caso de proteínas, la matriz de sustitución es de 20×20

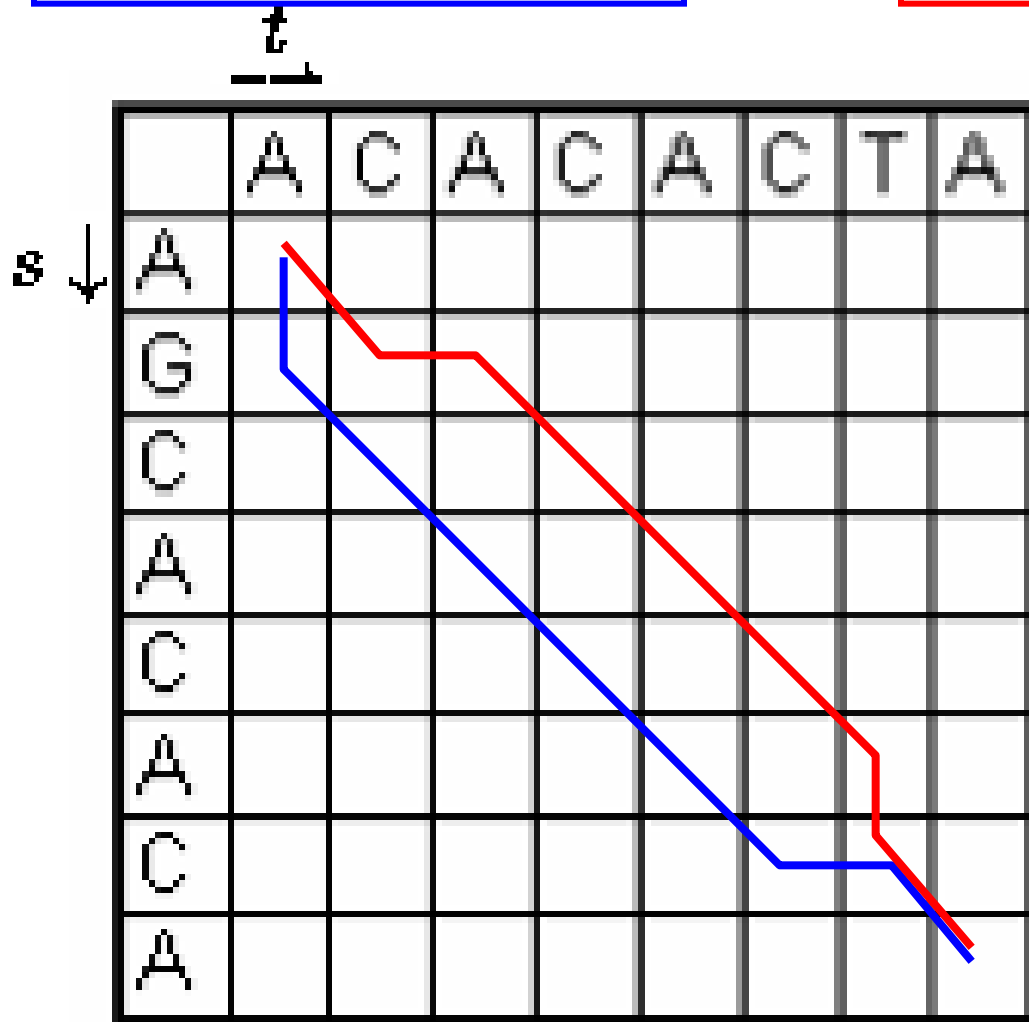
Esta matriz refleja que la transversión es más probable que la transición

Alineamiento de 2 secuencias

s : *A G C A C A C - A*
 t : *A - C A C A C T A*

or

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



las líneas horizontales y verticales nos indican en cuál secuencia hemos introducido *gaps*

Alineamiento de 2 secuencias

Programación dinámica:

matriz de tamaño $(M + 1) \times (N + 1)$
donde M y N son los largos de las
secuencias a alinear

regla de llenado de casilleros de la matriz:

$$S[i, j] = \text{Max} \left[\begin{array}{l} S[i - 1, j - 1] + MS_{ij}, \\ S[i - 1, j] + \text{Gap}, \\ S[i, j - 1] + \text{Gap} \end{array} \right]$$

	gap	a1	a2	a3
gap	0	1 gap	2 gaps	3 gaps
b1	1 gap	$\text{MAX}(X, Y, Z)$		
b2	2 gaps			
b3	3 gaps			

$$\begin{aligned} X &= 0 + \text{match}(a1, b1) \\ Y &= (1 \text{ gap}) + (1 \text{ gap}) \\ Z &= (1 \text{ gap}) + (1 \text{ gap}) \end{aligned}$$

Es decir, elegiremos el
valor máximo entre:

la suma del valor en la
diagonal y el score de
alineamiento para los
residuos i y j (**MS_{ij}**);

la suma de la casilla
superior y el valor de
Gap; y

la suma de la casilla
izquierda con el valor de
gap

Alineamiento de 2 secuencias

Secuencias:
A: G A T C G
B: G C A T C C G

1) Inicializar la matriz de score:

Valores de matriz de sustitución:
Match = 2
Mismatch = -1
Gap = -2

		SeqB							
		G	C	A	T	C	C	G	
SeqA		0	-2	-4	-6	-8	-10	-12	-14
	G	-2							
	A	-4							
	T	-6							
	C	-8							
	G	-10							

2) Llenar la matriz de score y la matriz de traceback:

Matriz de alineamiento

		G	C	A	T	C	C	G
	0	-2	-4	-6	-8	-10	-12	-14
G	-2	2						
A	-4							
T	-6							
C	-8							
G	-10							

Matriz de traceback

		G	C	A	T	C	C	G
G		d						
A								
T								
C								
G								

$$\begin{aligned}
 S[1,1] &= \text{Max} \left[\begin{array}{l} S[0,0] + MS_{ij} = 0 + 2 = 2 \\ S[0,1] + \text{Gap} = -2 + (-2) = -4 \\ S[1,0] + \text{Gap} = -2 + (-2) = -4 \end{array} \right] \\
 S[1,1] &= 2
 \end{aligned}$$

Matriz de alineamiento

		G	C	A	T	C	C	G
	0	-2	-4	-6	-8	-10	-12	-14
G	-2	2	0	-2	-4	-6	-8	-10
A	-4	0	1	2	0	-2	-4	-6
T	-6	-2	-1	0	4	2	0	-2
C	-8	-4	0	-2	2	6	4	2
G	-10	-6	-2	-1	0	4	5	6

Matriz de traceback

	G	C	A	T	C	C	G
G	d	i	i	i	i	i	di
A	a	d	d	i	i	i	i
T	a	da	da	d	i	i	i
C	a	d	dai	a	d	di	i
G	da	a	d	a	a	d	d

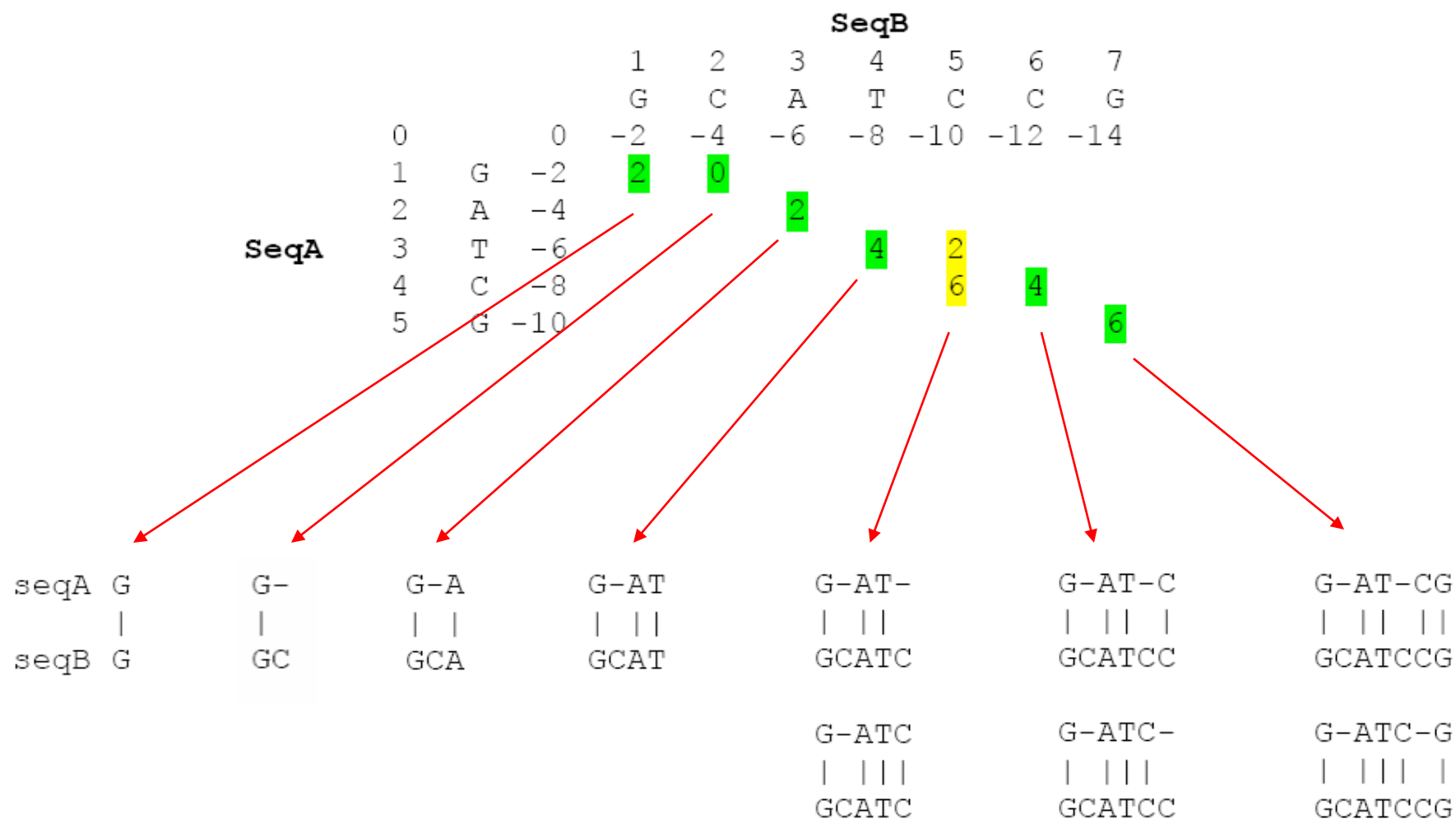
El valor en cada casilla $S[i,j]$ es el del alineamiento de mayor valor entre las subsecuencias de A y B que empiezan en el carácter 1 y terminan en el carácter i y j. En el caso de nuestro alineamiento la casilla $S[m,n]$ tiene como valor 6, lo cual nos dice que el alineamientos global de mayor valor entre ambas secuencias tiene como valor 6.

Si tomamos una submatriz, el valor en la casilla inferior derecha de la submatriz nos indica el puntaje óptimo del “sub”alineamiento

3) Llevar a cabo el *traceback*:

		SeqB							
		G	C	A	T	C	C	G	
SeqA		0	-2	-4	-6	-8	-10	-12	-14
	G	-2	2	0	-2	-4	-6	-8	-10
	A	-4	0	1	2	0	-2	-4	-6
	T	-6	-2	-1	0	4	2	0	-2
	C	-8	-4	0	-2	2	6	4	2
	G	-10	-6	-2	-1	0	4	5	6

		G	C	A	T	C	C	G
G		d	i	i	i	i	i	di
A		a	d	d	i	i	i	i
T		a	da	da	d	i	i	i
C		a	d	dai	a	d	di	i
G		da	a	d	a	a	d	d



para ambos alineamientos el score está dado por:

$$S = 5 * \text{Match} + 2 * \text{Gap} = 5 * 2 + 2 * (-2) = 6$$

Dynamic programming matrix:

		j → (sequence y)								
		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
i ↓ (sequence x)	0	0	-6	-12	-18	-24	-30	-36	-42	-48
	1 T	-6	5	-1	-7	-13	-19	-25	-31	-37
	2 T	-12	-1	3	-3	-2	-8	-14	-20	-26
	3 C	-18	-7	-3	8	2	3	-3	-9	-15
	4 A	-24	-13	-9	2	6	0	1	-5	-4
	5 T	-30	-19	-15	-4	7	4	-2	6	0
	M = 6 A	-36	-25	-21	-10	1	5	2	0	11

Optimum alignment scores 11:

T	-	-	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	-6	-6	+5	+5	-2	+5	+5

Penalidad de Gap

- Introducción de *gaps*:
 - Abrir un gap es más "costoso" que extenderlo
 - Modelos de modificación de costos de gap
 - $P = c + gd$
 - P penalidad total de gaps
 - c penalidad de apertura
 - d penalidad de extensión
 - g longitud del gap - 1

Alineamientos Globales

- El alineamiento global resulta en una comparación de dos secuencias sobre su longitud total
- Los alineamientos globales no son muy efectivos cuando se trata de secuencias muy divergentes y no reflejan el hecho de que dos secuencias pueden compartir sólo una región limitada de secuencia conservada. Por ejemplo cuando sólo un dominio es compartido entre ellas

Algoritmo Needleman-Wunsch

Alineamientos Globales vs Locales

- Los alineamientos locales encuentran la región o regiones de mayor similitud entre dos secuencias

ACTTAGCAGACTAACGTAAC

CCATGACTAACGGGACCTAC

Algoritmo Smith-Waterman

Alineamiento de Múltiples Secuencias

- In theory, making an optimal alignment between **two** sequences is computationally straightforward, but **aligning a large number of sequences** using the same method is **almost impossible**.
- The cost **increases exponentially** with the number of sequences involved (the product of the sequence lengths)

heurístico, ca.

(Del gr. εὕρισκειν, hallar, inventar, y *-ístico*).

1. **adj.** Perteneiente o relativo a la **heurística**.
2. **f.** Técnica de la indagación y del descubrimiento.
3. **f.** Busca o investigación de documentos o fuentes históricas.
4. **f.** En algunas ciencias, manera de buscar la solución de un problema mediante métodos no rigurosos, como por tanteo, reglas empíricas, etc.

Real Academia Española © Todos los derechos reservados

<http://es.wikipedia.org/wiki/Heur%C3%ADstica>

Alineamiento de Múltiples Secuencias

Alineamiento de pares para todas las secuencias

		ACGTACGTCC	ACCTACGTCC	ACCACCGTCC	ACCCCCCTCC	CCCCCCCCCCC
		human	chimp	gorilla	orangutan	maquaque
ACGTACGTCC	human	-				
ACCTACGTCC	chimp	ACGTACGTCC	-			
		ACCTACGTCC				
ACCACCGTCC	gorilla	ACGTACGTCC	ACCTACGTCC	-		
		ACCACCGTCC	ACCACCGTCC			
ACCCCCCTCC	orangutan	ACGTACGTCC	ACCTACGTCC	ACCACCGTCC	-	
		ACCCCCCTCC	ACCCCCCTCC	ACCCCCCTCC		
CCCCCCCCCCC	maquaque	ACGTACGTCC	ACCTACGTCC	ACCACCGTCC	ACCCCCCTCC	-
		CCCCCCCCCCC	CCCCCCCCCCC	CCCCCCCCCCC	CCCCCCCCCCC	

Matriz de distancias con los score obtenidos para cada alineamiento de pares

	human	chimp	gorilla	orangutan	maqaque
human	-				
chimp	1	-			
gorilla	3	2	-		
orangutan	4	3	2	-	
maqaque	6	5	4	2	-

No. de nucleótidos diferentes.

	human	chimp	gorilla	orangutan	maqaque
human	-				
chimp	0,1	-			
gorilla	0,3	0,2	-		
orangutan	0,4	0,3	0,2	-	
maqaque	0,6	0,5	0,4	0,2	-

se pueden usar otras métricas

D, fracción de sitios diferentes

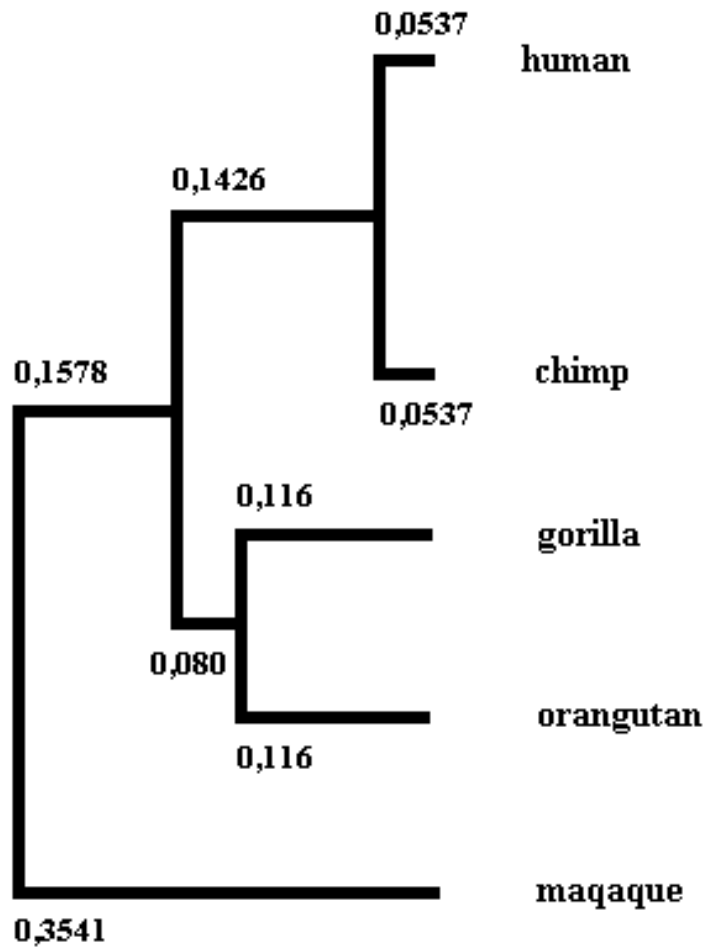
	human	chimp	gorilla	orangutan	maqaque
human	-				
chimp	0,107	-			
gorilla	0,383	0,232	-		
orangutan	0,571	0,383	0,232	-	
maqaque	1,207	0,823	0,571	0,232	-

número de sustituciones por sitio

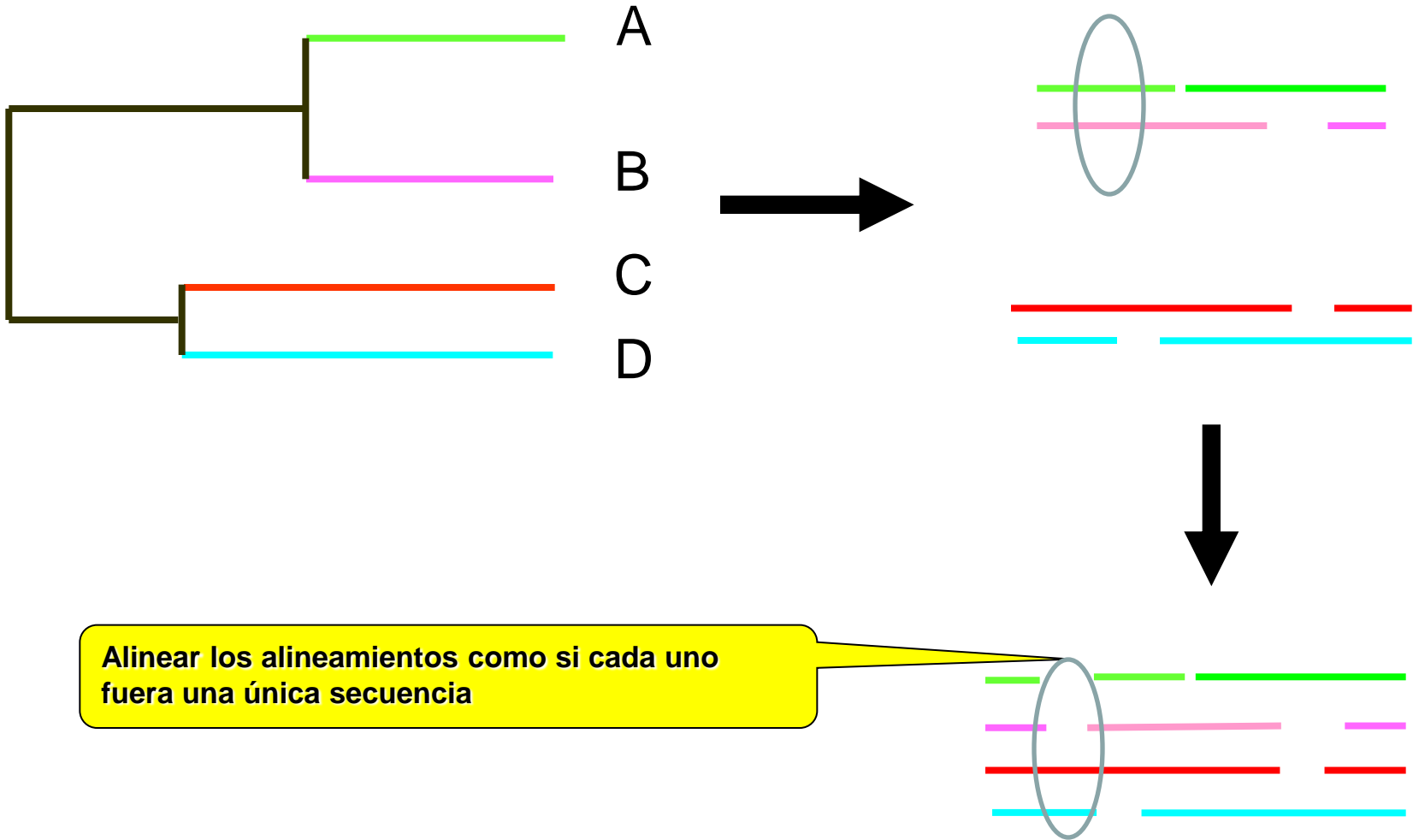
$$d = \frac{-3}{4} \ln\left(1 - \frac{4}{3} D\right)$$

árbol UPGMA

- Unweighted Pair Group Method with Arithmetic mean



Third step:



Puntaje de alineamientos múltiples

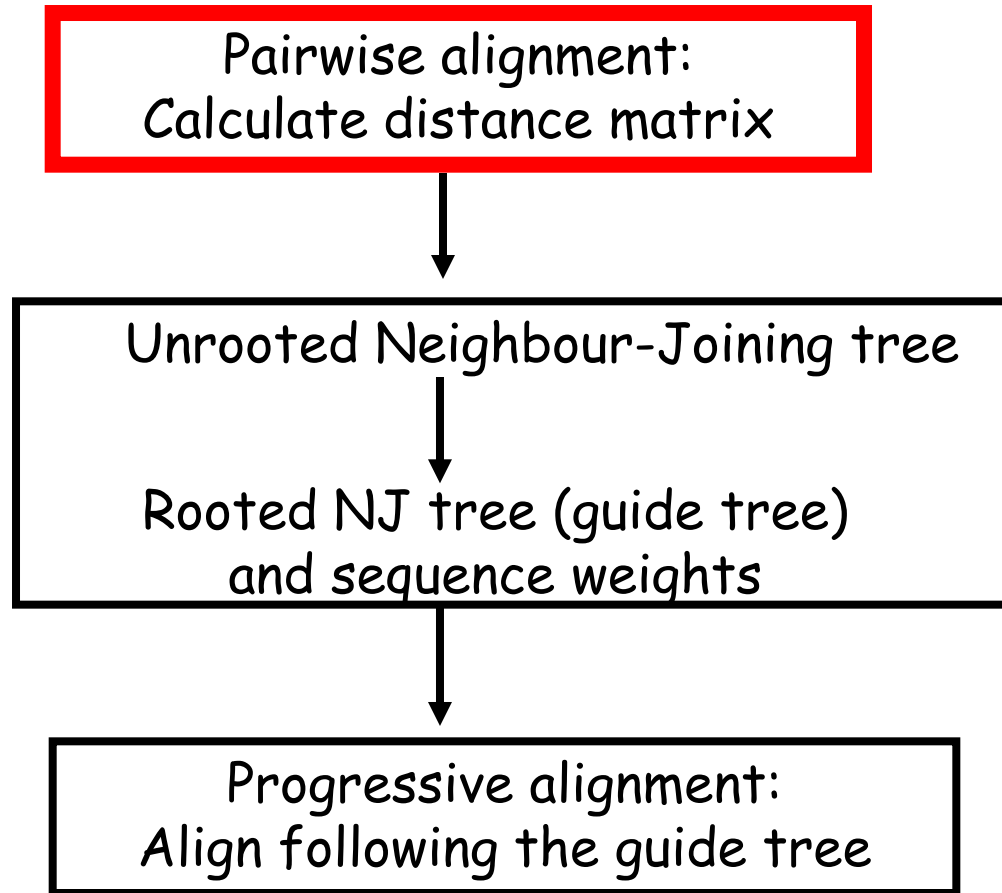
- 1234
 - ACGT
 - ACGA
 - AGGA
- match=1
mismatch=0
- 1: A-A + A-A + A-A = 1+1+1 = 3
 - 2: C-C + C-G + C-G = 1+0+0 = 1
 - 3: G-G + G-G + G-G = 1+1+1 = 3
 - 4: T-A + T-A + A-A = 0+0+1 = 1
- $S(\text{alineamiento}) = S(1) + S(2) + S(3) + S(4) = 3+1+3+1 = 8$

Alineamiento progresivo - pros y cons

- Pros
 - Rápido
 - Suficientemente certero
- Cons
 - La apertura de gaps no se puede corregir
 - Errores en el alineamiento de las primeras secuencias puede tener un gran efecto en el alineamiento total

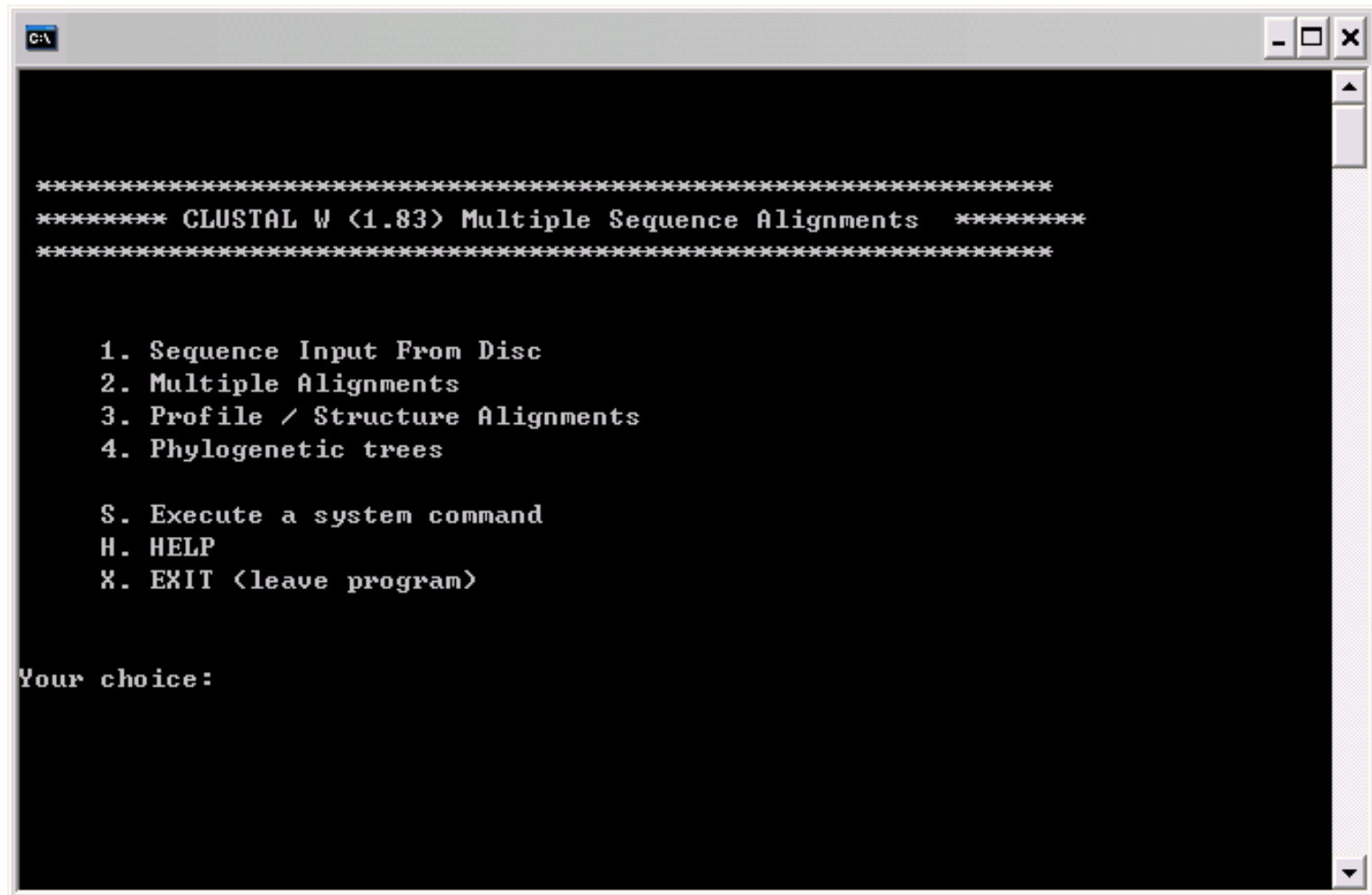
- Un alineamiento múltiple *no* refleja el nivel de similaridad que puede ser obtenido en un alineamiento de a pares (mediante el cual pudieran haber sido alineadas de manera completamente diferente)
- Por debajo de un cierto nivel de identidad de secuencia, los alinamientos (entre dos o más secuencias) son menos confiables
 - *“twilight zone” ~ 20-30% para secuencias de aminoácidos*
- Por debajo de ese umbral es conveniente usar información adicional (ej. incluir secuencias alineadas por superposición de sus estructuras 3D)

ClustalW, método en tres etapas



- **ClustalW** gasta aproximadamente el 96% de su tiempo de corrida en la primera etapa

ClustalW



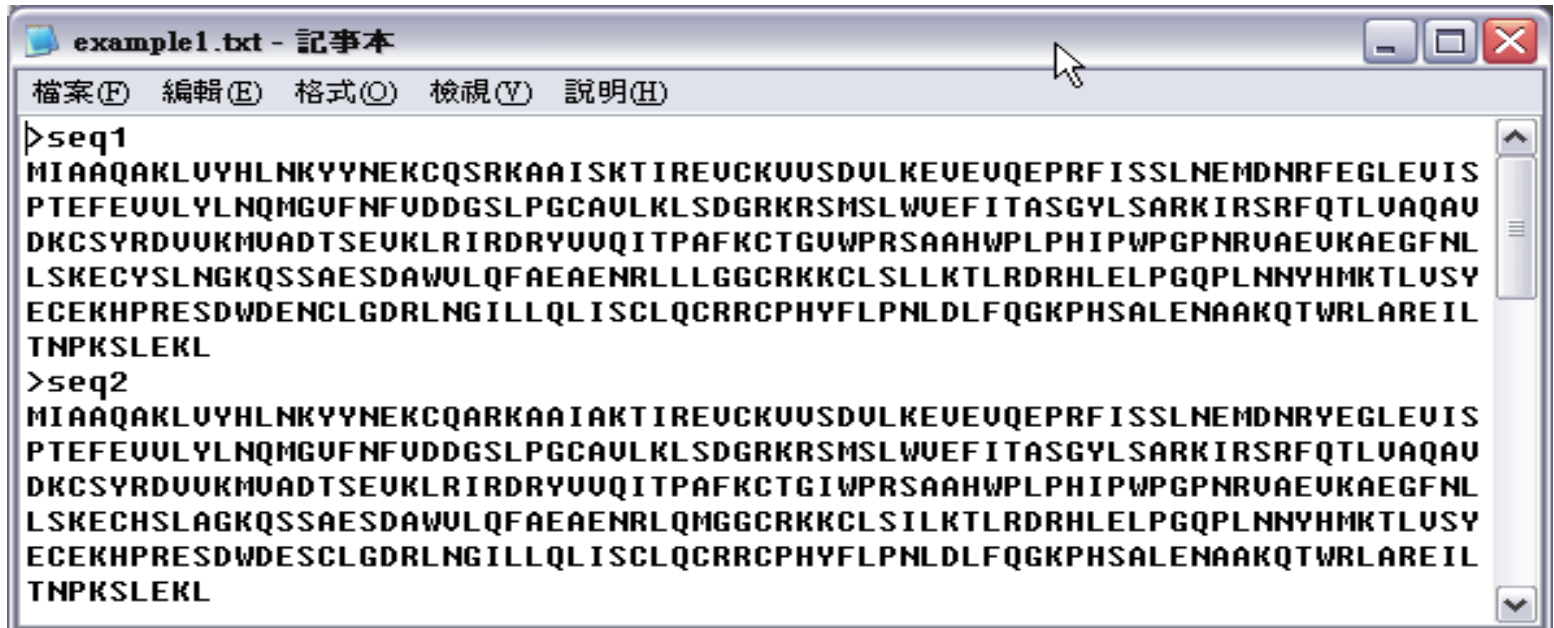
A screenshot of the ClustalW command-line interface running in a Windows-style window. The window has a title bar with a small icon on the left and standard minimize, maximize, and close buttons on the right. The main area is black with white text. The text is as follows:

```
C:\n
```

```
*****  
***** CLUSTAL W (1.83) Multiple Sequence Alignments *****  
*****  
  
1. Sequence Input From Disc  
2. Multiple Alignments  
3. Profile / Structure Alignments  
4. Phylogenetic trees  
  
S. Execute a system command  
H. HELP  
X. EXIT (leave program)  
  
Your choice:
```

Input File

- todas las secuencias en un solo archivo
- 7 formatos pueden ser aceptados:
 - NBRF/PIR, EMBL/Swissport, **Fasta**, GDE, Clustal, GCG/MSF, RSF ... pero **Fasta** es el más común
 - es posible editar en Notepad



```
>seq1
MIAAQAKLUYHLNKYYNEKCQSRKAAISKTIREVCKUUSDULKEVEUQEPRFISSLNEMDNRFEGLVISP
PTEFEVULYLNQMGUFNFUDDGSLPGCAULKLS DGRKRSM LWEIFITASGYLSARKIRS RFQTLVAQAU
DKCSYRDVUKMVADTSEVKLRIRDRYUUQITPAFKCTGUWPRSAAHWPLPHIPWPGPNRVAEUKAEGFNL
LSKECYS L NGKQSSAESDAWULQFAEAENRLLGGCRKKCLSLLKTLRDRHLELPGQPLNNYHMKTLUSY
ECEKHPRES DWDENCLG DRLNGILLQLISCLQCRRCPHYFLPNLDLFQGKPHSALENAAKQTWRLAREIL
TNPKSLEKL

>seq2
MIAAQAKLUYHLNKYYNEKCQARKAAIAKTIREVCKUUSDULKEVEUQEPRFISSLNEMDNRFEGLVISP
PTEFEVULYLNQMGUFNFUDDGSLPGCAULKLS DGRKRSM LWEIFITASGYLSARKIRS RFQTLVAQAU
DKCSYRDVUKMVADTSEVKLRIRDRYUUQITPAFKCTGIWPRSAAHWPLPHIPWPGPNRVAEUKAEGFNL
LSKECHSLAGKQSSAESDAWULQFAEAENRLQMGCCRKKCLSILKTLRDRHLELPGQPLNNYHMKTLUSY
ECEKHPRES DWDENCLG DRLNGILLQLISCLQCRRCPHYFLPNLDLFQGKPHSALENAAKQTWRLAREIL
TNPKSLEKL
```

Main Menu

1. Sequence input from disk
2. Multiple alignment
3. Profile / structure alignment
4. Phylogenetic tree

Multiple alignment menu

1. Do complete multiple alignment now (slow/fast)
2. Produce guide tree only
3. Do alignment using old guide tree file
4. Slow / fast pairwise alignment
5. Pairwise alignment parameter
6. Multiple alignment parameter
7. Reset gaps before alignment
8. Screen display
9. Output format option

Profile / Structure alignment

1. Input 1st. profile
2. Input 2nd. profile / sequence
3. Align 2nd. profile to 1st. profile
4. Align sequences to 1st. profile

Phylogenetic tree

1. Input alignment
2. Exclude position with gaps
3. Correct for multiple substitutions
4. Draw tree now
5. Bootstrap tree

Output File

● CLUSTAL output : [filename].aln

```
CLUSTAL W (1.83) multiple sequence alignment

seq2      MASTUWGGAPWWG-----ARPLTDIDFCGAQLQELTQL-----IQESWSEGPKPGAD-
seq4      -----LTQL-----IQELG-----VQESWSEGEPEPGADL
seq1      MASAUWG-----DFCSGAQLQELTQLIQELGVQESWSDAPKPGPDL
seq3      MASAUWGSAPWWGPPPPAPARPLTDIDFCGAQLQELTQLIQELGVQESWSDGPKPGADL
              :***          :*****:.*:***.*

seq2      LRAKDFUFALLGLUHRQDPRFPPQAELLLLRRGGIREGSLDLGHAPLGPYSRGPHYDAGFT
seq4      LRAKEFUFSLGLUHRQDPRFPPQAELLLLRRGGIREGSVDLGHAPLGPYSRGPHYDAGFT
seq1      LRAKDFUFSLGLUHRDPRFPPQAELLLLRRGGIREGSLDLGLAPLGPYIRGPHYDAGFT
seq3      LRAKDFUFSLGLUHRDPRFPPQAELLLLRRGGIREGSLDLGHAPLGPYARGPHYDAGFT
              ****:*:*:*:*****:*****:*****:*****:*****:*****:*****

seq2      LLUPVFSLDGTGPELLLDLESCSAWLRLELMRGILVUREAWQDCLGPPUPEESDMTHQTH
seq4      ULUPVFSLDGTGPELLLDLESCSAWLRLELMRGILVUREAWLDCLGPPUPEESDMTPQTQ
seq1      LLUPVFSLDGTGQELQLDAESCFARLRLPEQIRGTSUREAWQDCLGPPAPGGRDSVHRTQ
seq3      LLUPMFSLDGT--ELQLDLESCYAQUCLPEMUCGTPIREHWQDCLGPPUPGARDSIHRTE
              :***:*****  ** ** ** * : *** : * :** * *****_*  *  :*.

seq2      SKESPTDRENSVDPSHDYUPEPEPHMSLQKSSSDLSESQSSYKDITNPETPEPLETLSSD
seq4      GKESPTDRGNSVDQSHDCUPEPEPHMSLQKFS-DLG-SQSPYNDIANLEAPELETSPSE
seq1      SEESPKDRQSPUDQPHDG-TEPEPTUSLDQS----SGPEG--QDUIDLELSTPLKLTNGD
seq3      SEESSKDWQSSVDQPHSYUTEHEAPUSLEKSPSDUSASESPQHUVUDLGSTAPLKTMSDD
              .:***.*  ..** .* . * *. :***:  . :. : **: :  . **:* ..:

seq2      ALDAD-ESQVPKPSEAPFUKLAEVAESLIPVPGAPRLUHAARHAGU-
seq4      ALETD-ESPUPRPSEA-----AKUWPTLCPT-----
seq1      LRKAUVSPPIWPSE-----WEAWPTLCPAQUAAWFFASL-----
seq3      UTKAAUESPVPKPSEA-----REAWPTLCSAQUAAWFFATLAUAES
              .:  * :* ***  :.  :* ..
```

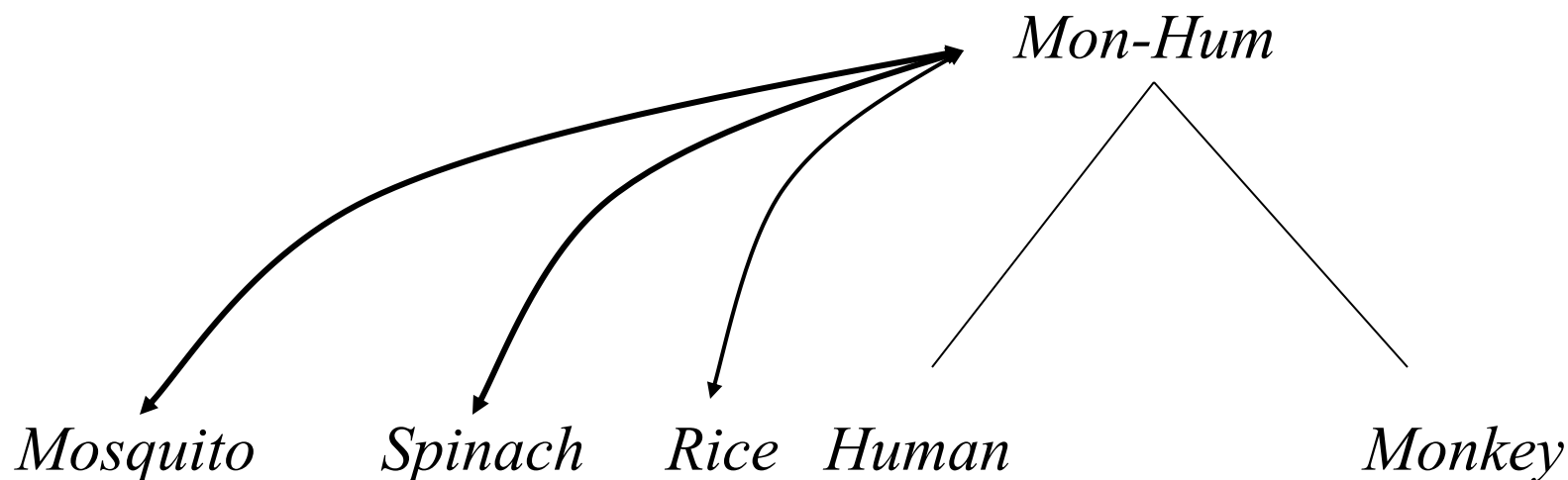
● GUIDE TREE : [filename].dnd

Matriz de distancias

PAM	Spina ch	Rice	Mosquito	Monkey	Human
Spina ch	0.0	84.9	105.6	90.8	86.3
Rice	84.9	0.0	117.8	122.4	122.6
Mosquito	105.6	117.8	0.0	84.7	80.8
Monkey	90.8	122.4	84.7	0.0	3.3
Human	86.3	122.6	80.8	3.3	0.0

Primer paso

La mínima distancia PAM es 3.3 (Human - Monkey) . Unimos Human y Monkey como "MonHum" y recalculamos las distancias de cada secuencia restante a este nuevo taxa



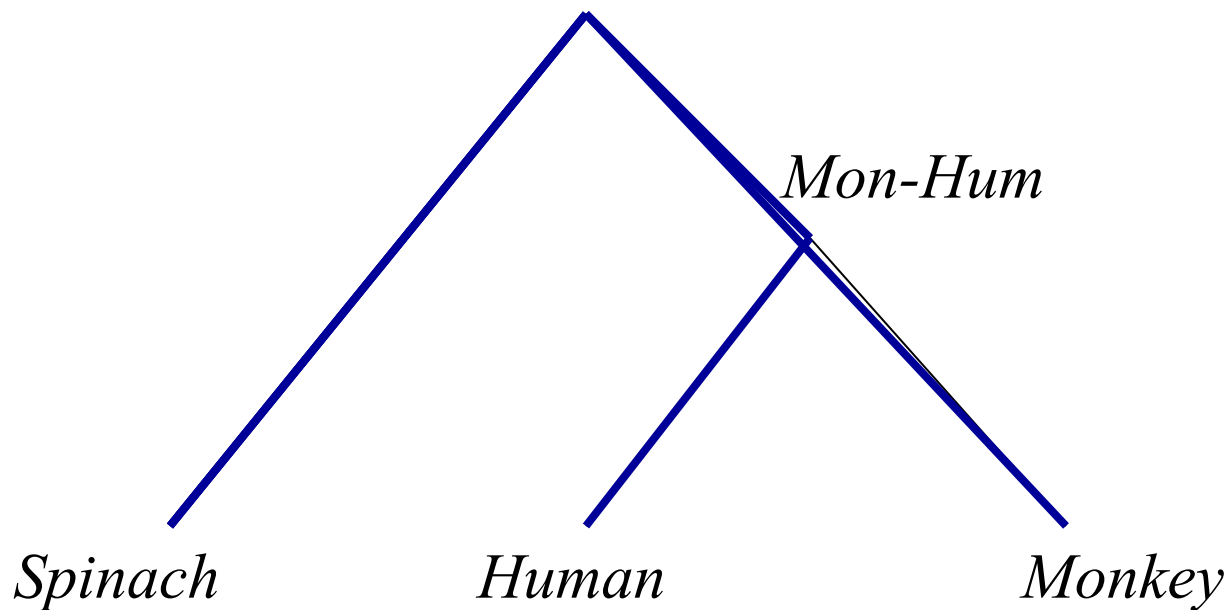
Nuevas distancias

Podemos usar el simple promedio de las distancias:

$Dist[Spinach, MonHum]$

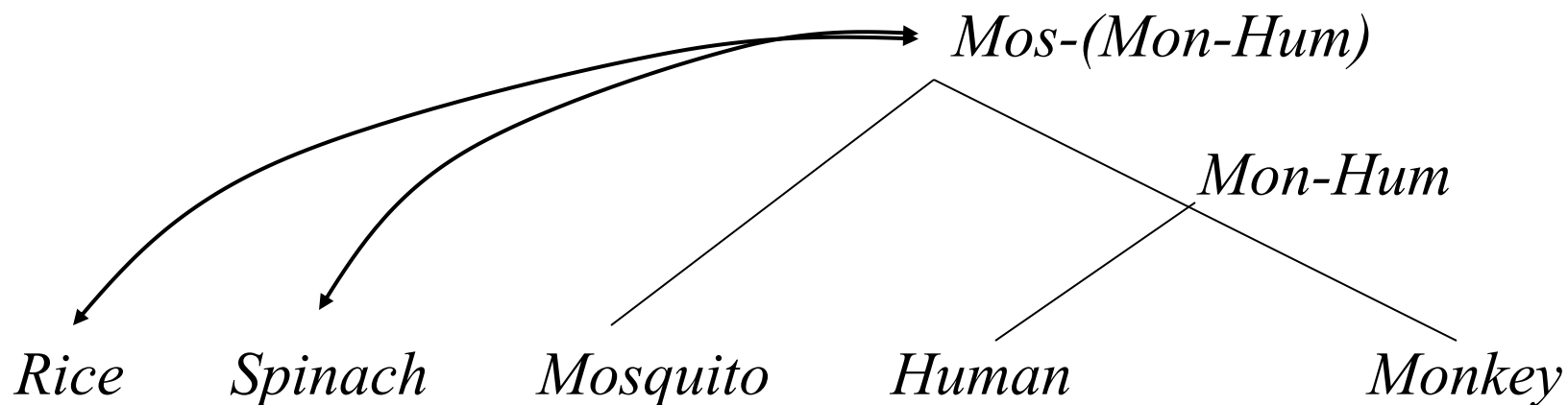
$$= (Dist[Spinach, Monkey] + Dist[Spinach, Human])/2$$

$$= (90.8 + 86.3)/2 = 88.55$$



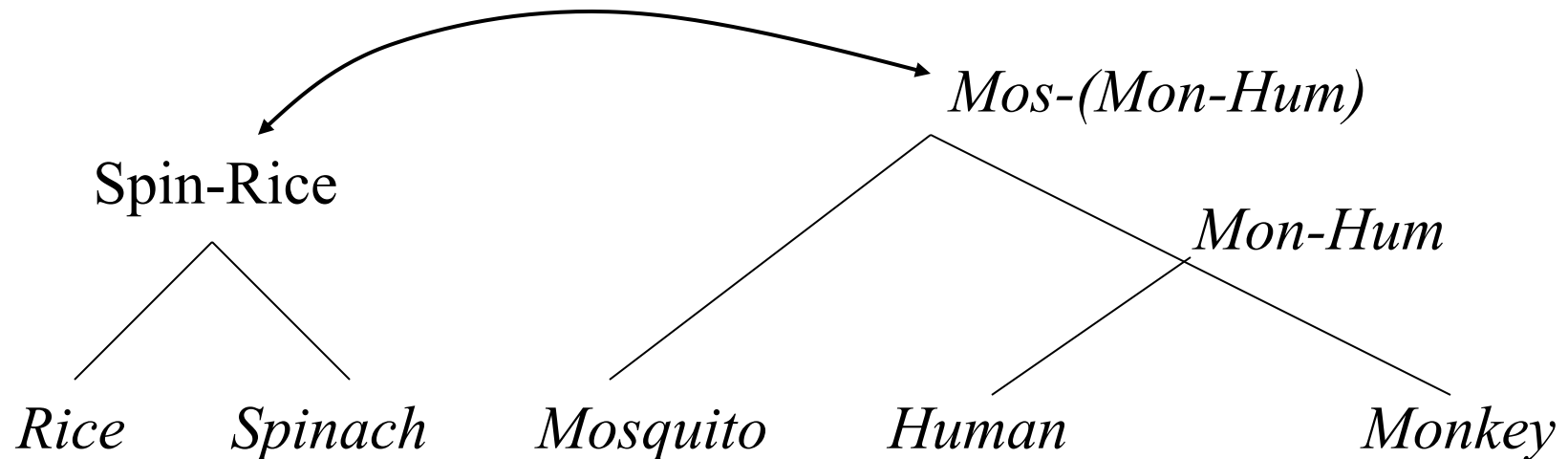
Siguiente ciclo

PAM	Spina ch	Rice	Mosquito	MonHu m
Spina ch	0.0	84.9	105.6	88.6
Rice	84.9	0.0	117.8	122.5
Mosquito	105.6	117.8	0.0	82.8
MonHu m	88.6	122.5	82.8	0.0



Siguiente ciclo

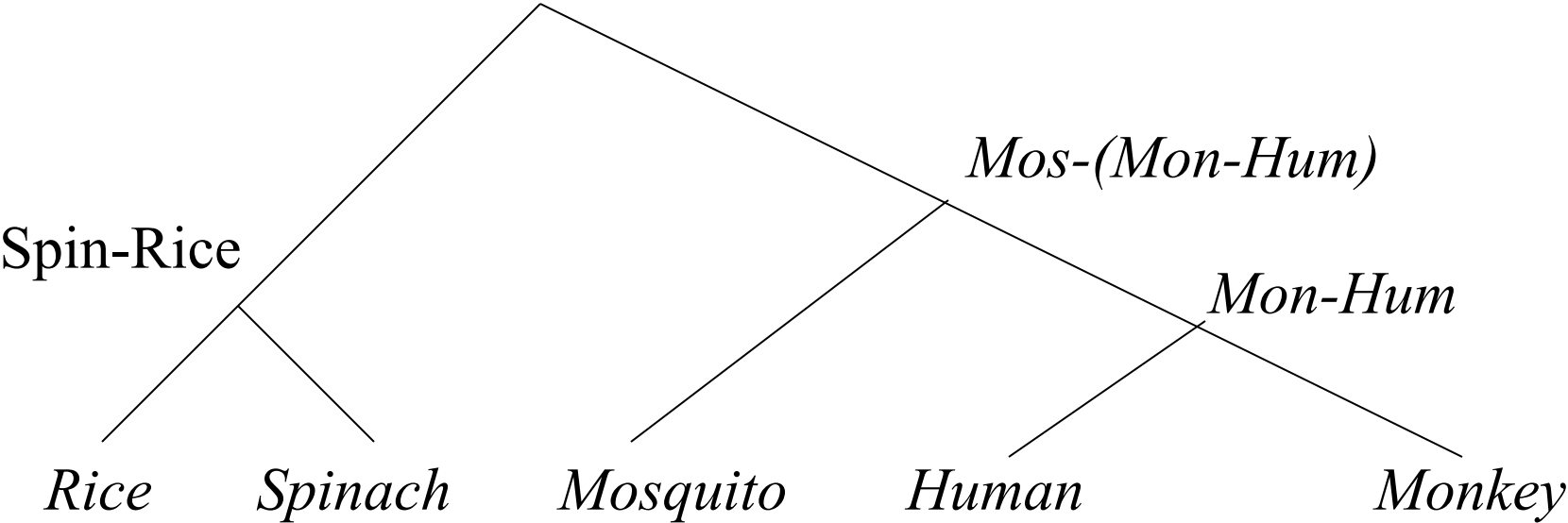
PAM	Spina ch	Rice	MosMonHum
Spina ch	0.0	84.9	97.1
Rice	84.9	0.0	120.2
MosMonHum	97.1	120.2	0.0



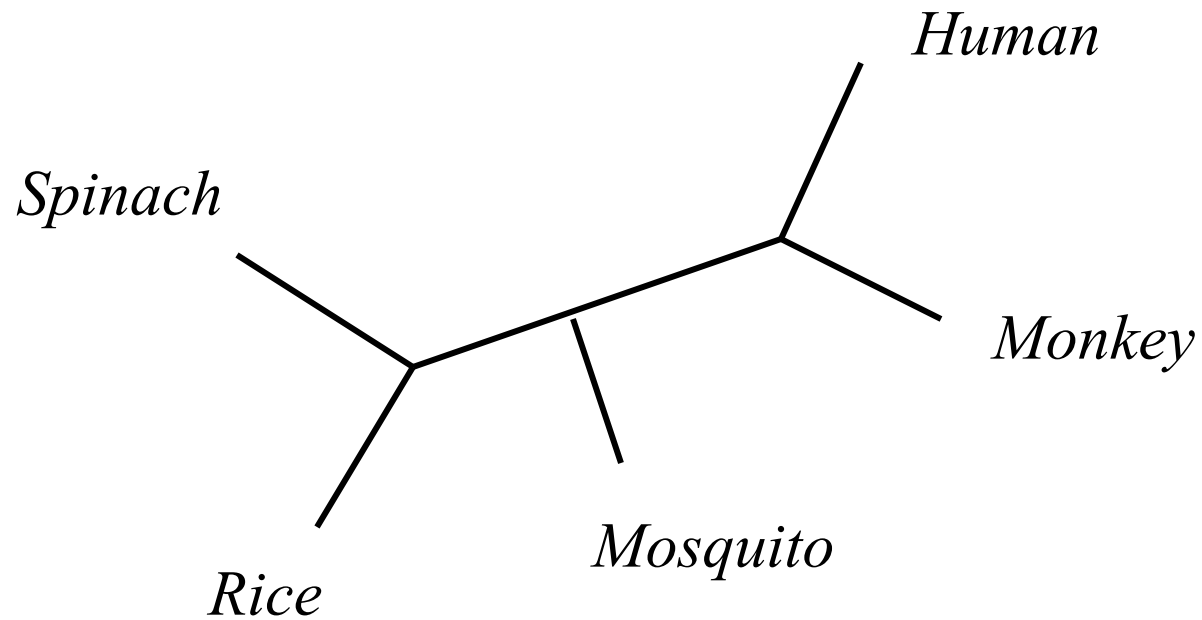
Al final tenemos

PAM	SpinRice	MosMonHum
Spinach	0.0	108.7
MosMonHum	108.7	0.0

(Spin-Rice)-(Mos-(Mon-Hum))



Árbol Neighbor-Joining no enraizado



Bootstrapping:

- Crear un set de datos para analizar
 - Usar el alineamiento original, muestrear con reemplazo para crear “*pseudo-réplicas*” del alineamiento original
- Construir árboles.
 - uno para cada pseudo-réplica
- Calcular valores de bootstrap
 - para cada nuevo árbol, se cuenta el número de puntos de ramificación equivalentes con el árbol original... si todos los árboles muestran las mismas ramificaciones, el valor de bootstrap es 100%.

Bootstrapping:

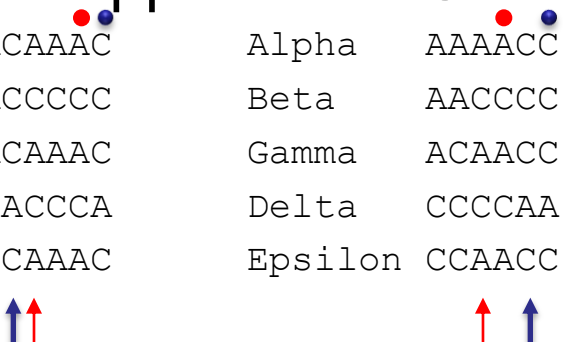
- Este procedimiento nos indica la estabilidad de la topología del árbol
 - Nos ayuda a saber si nuestros datos de secuencia son adecuados para validar una determinada topología
 - El valor mismo es una cuenta (o porcentaje) del número de veces que cada rama de nuestro árbol está presente en el conjunto total de árboles provenientes de *pseudoréplicas*. Por lo tanto, nos indica si nuestro árbol está sujeto a cambios debido a leves variaciones de las secuencias alineadas

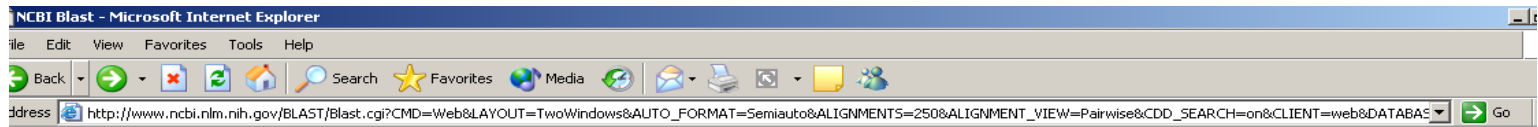
- Test Data Set

Alpha	AACAAC
Beta	AACCCC
Gamma	ACCAAC
Delta	CCACCA
Epsilon	CCAAAC

- Bootstrapped Data Sets

Alpha	ACAAAC	Alpha	AAAACC
Beta	ACCCCC	Beta	AACCCC
Gamma	ACAAAC	Gamma	ACAACC
Delta	CACCCA	Delta	CCCCAA
Epsilon	CCAAAC	Epsilon	CCAACC





[Search](#)

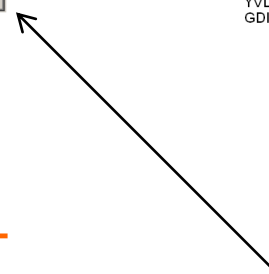
[Set subsequence](#) From: To:

[Choose database](#)

[Do CD-Search](#) ☒

Now: **BLAST!** or [Reset query](#) [Reset all](#)

>1IGR:A INSULIN-LIKE GROWTH FACTOR RECEPTOR
EICGPGDIRNDYQOLKRLNCTVIEGYLHLLISKAEDYRSYRFPKLTVITEYLLLF
VAGLESGLDFPNLTVIRGWKLFYNYALVIFEMTNLKDGLYNLRNITRGAIKNA
DLCYLSTVDVSLILDVSNINYVGNKPPKECGDLCPGTMECKPMCEKTTINNEY
YRCWTTNRCQKMCPCSTCGKRACTENNECHPECLGSCSAPNDTACVACRHY
YYAGVCVPACPPNTYRFEGWRCVDRDFCANILSAESSDSEGFVIHDGECMOEC
PSGFIRNGSQSMYCIPEGPCPKVCEEEKTKTIDSVTSAQMLQGCTIFKGNLLIN
IRRGNNIASELENFMGLIEVVTGYVKIRHSHALVSLSLKNLRLLGEEQLEGNYSF
YVLDNQNLQQLWDWDHRLNLTAKGMYFAFPNPKLCVSEIYRMEEVGTGKGROSK
GDINTRNNGERASCESDVDDDDKEQKLISEEDLN



Aquí se pega la secuencia
aminoacídica

Options for advanced blasting

[Limit by entrez query](#) or select from:

[Composition-based statistics](#) ☐

[Choose filter](#) ☒ Low complexity ☐ Mask for lookup table only ☐ Mask lower case

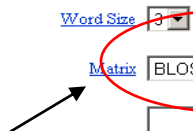
[Expect](#)

[Word Size](#)

[Matrix](#) Gap Cost:

[PSSM](#)

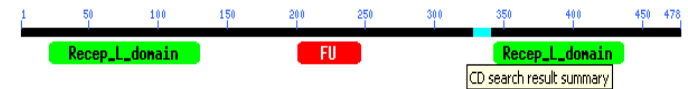
[Other advanced](#)



Matriz de
sustitución
usada

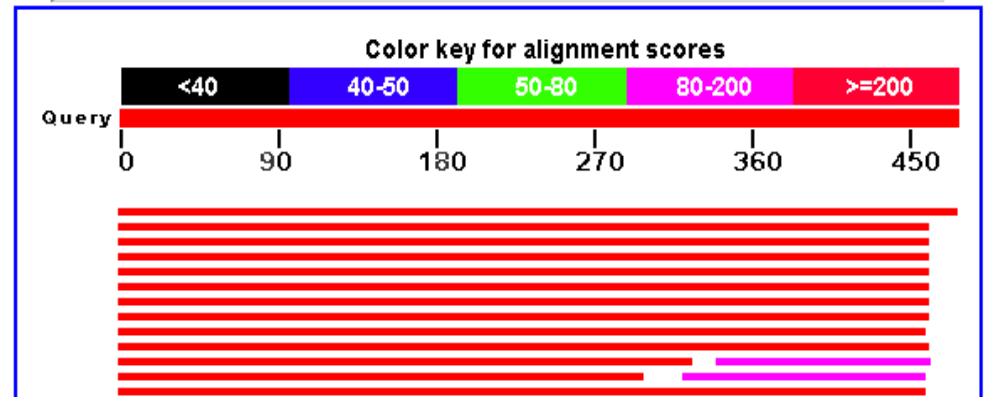
Dominios
conservados

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 1133 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments



Vista gráfica de los
"hits" – coloreado
por similitud

"hits"

Sequences producing significant alignments:		Score	E	
		(bits)	Value	
gi 6435822 pdb 1IGR A	Chain A, Type 1 Insulin-Like Growth F...	979	0.0	S
gi 249616 gb AAB22215.1 	insulin-like growth factor I recep...	947	0.0	
gi 4557665 ref NP_000866.1 	insulin-like growth factor 1 re...	947	0.0	G
gi 47523430 ref NP_999337.1 	IGF-1 receptor [Sus scrofa] >g...	935	0.0	G

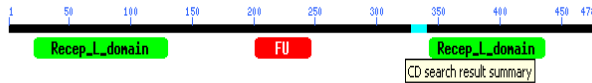
Alineamiento de
cada uno de los hits

> ☐ [gi|6435822|pdb|1IGR|A](#) **S** Chain A, Type 1 Insulin-Like Growth Factor Receptor
1-3)
Length=478

Score = 979 bits (2530), Expect = 0.0
Identities = 478/478 (100%), Positives = 478/478 (100%), Gaps = 0/478 (0%)

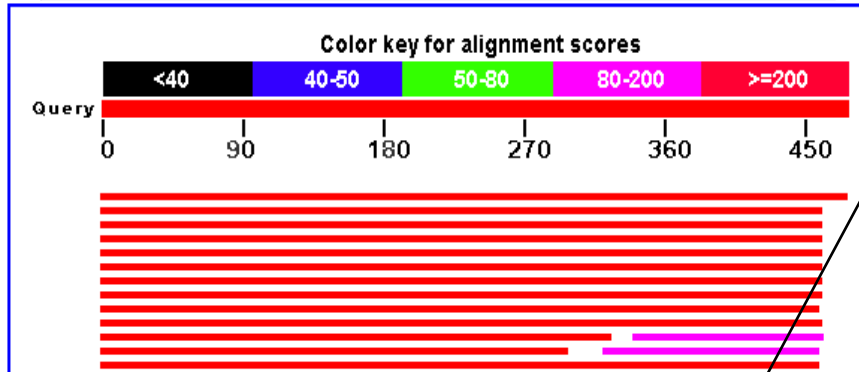
Query	1	EICGPGIDIRNDYQQLKRENECTVIEGYLHILLISKAEDYRSYRFPKLTVITEYLLLFVRV	60
		EICGPGIDIRNDYQQLKRENECTVIEGYLHILLISKAEDYRSYRFPKLTVITEYLLLFVRV	
Sbjct	1	EICGPGIDIRNDYQQLKRENECTVIEGYLHILLISKAEDYRSYRFPKLTVITEYLLLFVRV	60
Query	61	AGLESGLDLPNLTVIRGWKLFYNYALVIFEMTNLKDGLYNLRNITRGAIRIEKNADLC	120
		AGLESGLDLPNLTVIRGWKLFYNYALVIFEMTNLKDGLYNLRNITRGAIRIEKNADLC	

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 1133 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments



Sequences producing significant alignments:

gi 6435822 pdb 1IGR A	Chain A, Type 1 Insulin-Like Growth F...	979	0.0	S
gi 249616 gb AAB22215.1	insulin-like growth factor I recep...	947	0.0	
gi 4557665 ref NP_000866.1	insulin-like growth factor 1 re...	947	0.0	G
gi 47523430 ref NP_999337.1	IGF-1 receptor [Sus scrofa] >g...	935	0.0	G

> [gi|6435822|pdb|1IGR|A](#) S Chain A, Type 1 Insulin-Like Growth Factor Receptor
1-3)
Length=478

Score = 979 bits (2530), Expect = 0.0
Identities = 478/478 (100%), Positives = 478/478 (100%), Gaps = 0/478 (0%)

Query	1	EICGPGIDIRNDYQQLKRENC	TVIEGYLHILLISKAEDYRSYRFPKLT	VITEYLLFRV	60
		EICGPGIDIRNDYQQLKRENC	TVIEGYLHILLISKAEDYRSYRFPKLT	VITEYLLFRV	
Shjct	1	EICGPGIDIRNDYQQLKRENC	TVIEGYLHILLISKAEDYRSYRFPKLT	VITEYLLFRV	60

Query	61	AGLESGLDLPNLT	TVIRGWKLFYNYALVIFEMTNLKD	IGLYNLNITRG	AIKRNADLC	120
		AGLESGLDLPNLT	TVIRGWKLFYNYALVIFEMTNLKD	IGLYNLNITRG	AIKRNADLC	

Bit score: S'

The value S' is derived from the raw alignment score S , but statistical properties of the scoring system have been taken into account. Because bit scores are normalised with respect to the scoring system, they can be used to compare alignment scores from different searches.

E value: Expectation value.

Expected # of alignments with scores equivalent to or better than S to occur by chance. The lower the E value, the more significant the score.

NCBI Blast output help:

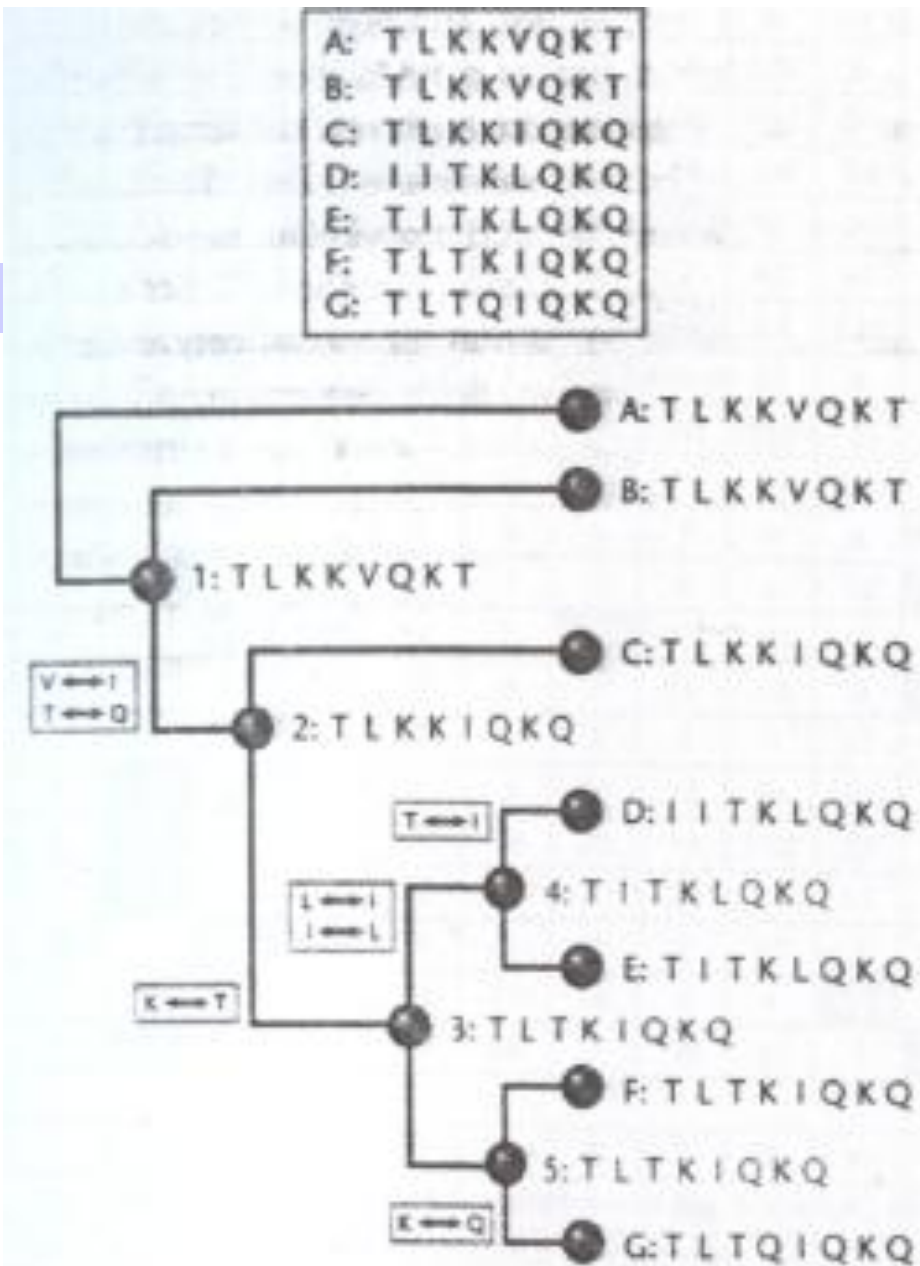
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Blast_output.html

Matrices de sustitución

The PAM family

- PAM matrices are based on global alignments of closely related proteins.
- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.
- Other PAM matrices are extrapolated from PAM1.

	I	K	L	Q	T	V
I	-	-	2	-	1	1
K	-	-	-	1	1	-
L	2	-	-	-	-	-
Q	-	1	-	-	1	-
T	1	1	-	1	-	-
V	1	-	-	-	-	-



Matrices de sustitución

The BLOSUM family

- BLOSUM matrices are based on local alignments.
- BLOSUM 62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.
- All BLOSUM matrices are based on observed alignments; they are **not extrapolated** from comparisons of closely related proteins.
- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. **A search for distant relatives may be more sensitive with a different matrix.**

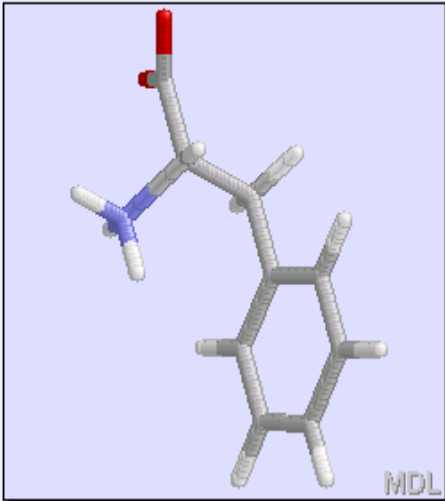
BLOSUM62

A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X

BLOSUM62



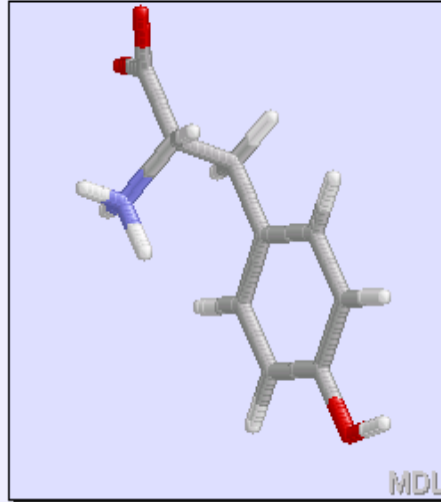
L-phenylalanine [F](#)



Comm

5	
-2	6
0	-2
-3	-4
-3	
1	

L-tyrosine [Y](#)

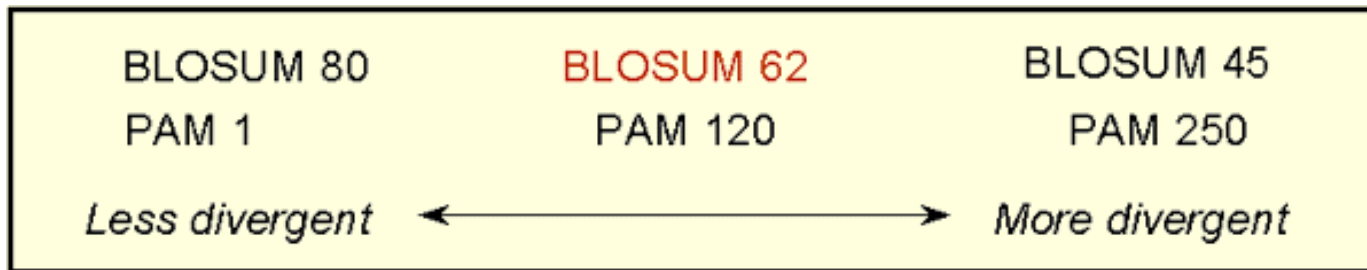


low weights

high weights

[illegible]

The relationship between BLOSUM and PAM substitution matrices. BLOSUM matrices with higher numbers and PAM matrices with low numbers are both designed for comparisons of closely related sequences. BLOSUM matrices with low numbers and PAM matrices with high numbers are designed for comparisons of distantly related proteins. If distant relatives of the query sequence are specifically being sought, the matrix can be tailored to that type of search.



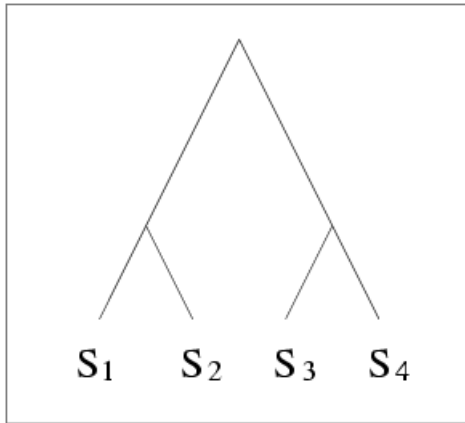
Matríz específica de posición (position-specific scoring matrix)

posición del alineamiento: **1** **2** **3** **4** **5** **6** **7** **8**

- Describe la frecuencia de cada aminoácido en cada posición de un alineamiento múltiple
- De esta manera, podemos contrastar si una nueva secuencia recibe un alto puntaje al ser evaluada por la PSSM. En este caso se asigna a la secuencia NMFWAFGH un puntaje de $0 + -2 + -3 + -2 + -1 + 6 + 6 + 8 = 12$
- Secuencias que reciban un alto puntaje serán muy compatibles con las frecuencias descritas por la PSSM y probablemente serán similares a las secuencias que originalmente se usaron para construir la PSSM.

A	-1	-2	-1	0	-1	-2	0	-2
R	5	0	5	-2	1	-3	-2	0
N	0	6	0	0	0	-3	0	1
D	-2	1	-2	-1	0	-3	-1	-1
C	-3	-3	-3	-3	-3	-2	-3	-3
Q	1	0	1	-2	5	-3	-2	0
E	0	0	0	-2	2	-3	-2	0
G	-2	0	-2	6	-2	-3	6	-2
H	0	1	0	-2	0	-1	-2	8
I	-3	-3	-3	-4	-3	0	-4	-3
L	-2	-3	-2	-4	-2	0	-4	-3
K	2	0	2	-2	1	-3	-2	-1
M	-1	-2	-1	-3	0	0	-3	-2
F	-3	-3	-3	-3	-3	6	-3	-1
P	-2	-2	-2	-2	-1	-4	-2	-2
S	-1	1	-1	0	0	-2	0	-1
T	-1	0	-1	-2	-1	-2	-2	-2
W	-3	-4	-3	-2	-2	1	-2	-2
Y	-2	-2	-2	-3	-1	3	-3	2
V	-3	-3	-3	-3	-2	-1	-3	-3

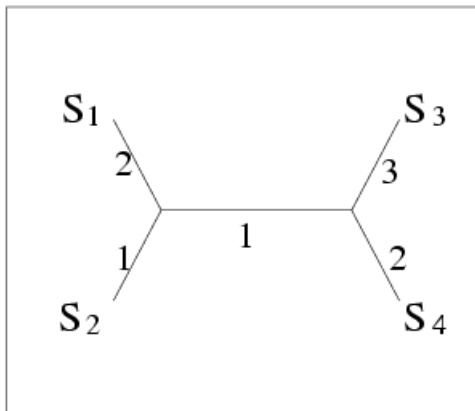
Métodos filogenéticos basados en distancias



TRUE TREE
(desconocido)

S₁ ACAATTAGAAC
S₂ ACCCTTAGAAC
S₃ ACCATTCCAAC
S₄ ACCAGACCAAC

DNA SEQUENCES



INFERRED TREE

METHODS
SUCH AS
NEIGHBOR
JOINING



	S ₁	S ₂	S ₃	S ₄
S ₁	0	3	6	5
S ₂		0	5	4
S ₃			0	5
S ₄				0

DISTANCE MATRIX

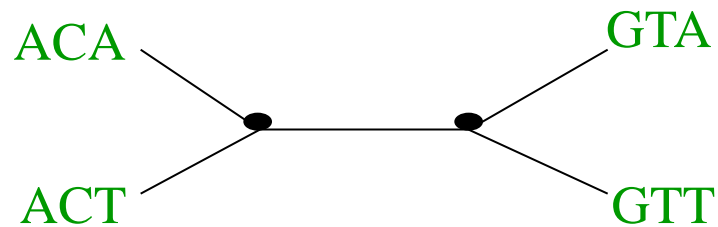
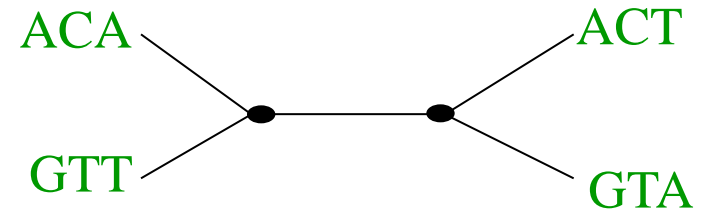
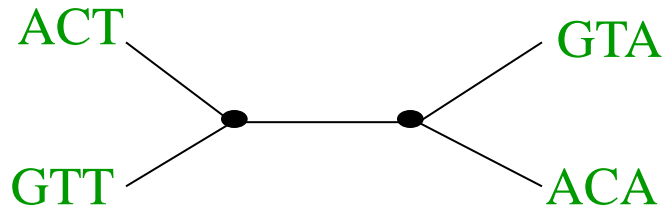
STATISTICAL
ESTIMATION
OF PAIRWISE
DISTANCES

Métodos basados en el análisis de cada posición:

Ej. Maxima parsimonia

- **Input:** Cuatro secuencias
 - ACT
 - ACA
 - GTT
 - GTA

Árboles posibles



Pregunta: cuál de los tres árboles posibles tiene el mejor puntaje de acuerdo con el criterio de máxima parsimonia

el menor número de cambios!

