Comparación de dos secuencias
El alineamiento

Biocomputación
Grado en Bioquímica

¿Para que comparar dos secuencias?

La comparación de dos secuencias mediante su alineamiento es uno de los pilares de la bioinformática y es la base de muchos métodos diferentes

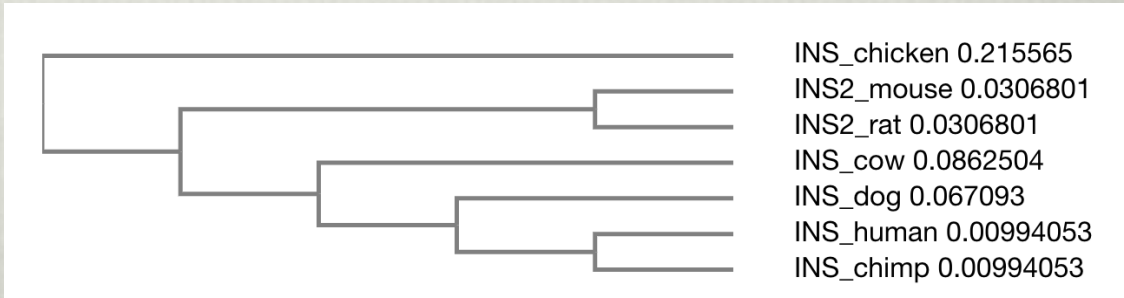
1) Comparación de secuencias homólogas para hacer un análisis evolutivo

- Detectar regiones conservadas
- Determinar el grado de parentesco entre varias secuencias (y de las especies que las portan)

```
INS_chicken      MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ
INS2_mouse       MALWMRFLPLLALLFLWESHPTQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVED
INS2_rat         MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVED
INS_cow          MALWTRLAPLLALLALWAPAPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG
INS_dog          MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVED
INS_human        MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
INS_chimp        MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
                 **** *  ***.** :     : * .:************************:* :**:.*

INS_chicken      PLVSSPL-RGEAGV--LPFQQEEYEKVKRGIVEQCCHNTCSLYQLENYCN
INS2_mouse       PQVAQLELGGGPGAGDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN
INS2_rat         PQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN
INS_cow          PQVGALELAGGPGAGGL-----EGPPQKRGIVEQCCASVCSLYQLENYCN
INS_dog          LQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCTSICSLYQLENYCN
INS_human        LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
INS_chimp        LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
                   *      *  *   *     *    *****:*** . ***********
```

Regiones conservadas

Árbol filogenético



¿Para que comparar dos secuencias?

2) Comparar una secuencia de entrada frente a todas en una base de datos (BLAST)

3) Encontrar una secuencia dentro del genoma (BLAT)

```
000000135 acatgtcatgctgcccaaggacatagccaagctggtccctaaaacccatc 000000184
>>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>>
154950463 acatgtcatgctgcccaaggacatagccaagctggtccctaaaacccatc 154950512

000000185 tgatgtctgaatctgaatggaggaatcttggcgttcagcagagtcaggga 000000234
>>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>>
154950513 tgatgtctgaatctgaatggaggaatcttggcgttcagcagagtcaggga 154950562

000000235 tgggtccattatatgatccatgaaccag 000000262
>>>>>>>>>> ||||||||||||||||||||||||||||| >>>>>>>>>>
154950563 tgggtccattatatgatccatgaaccag 154950590
```

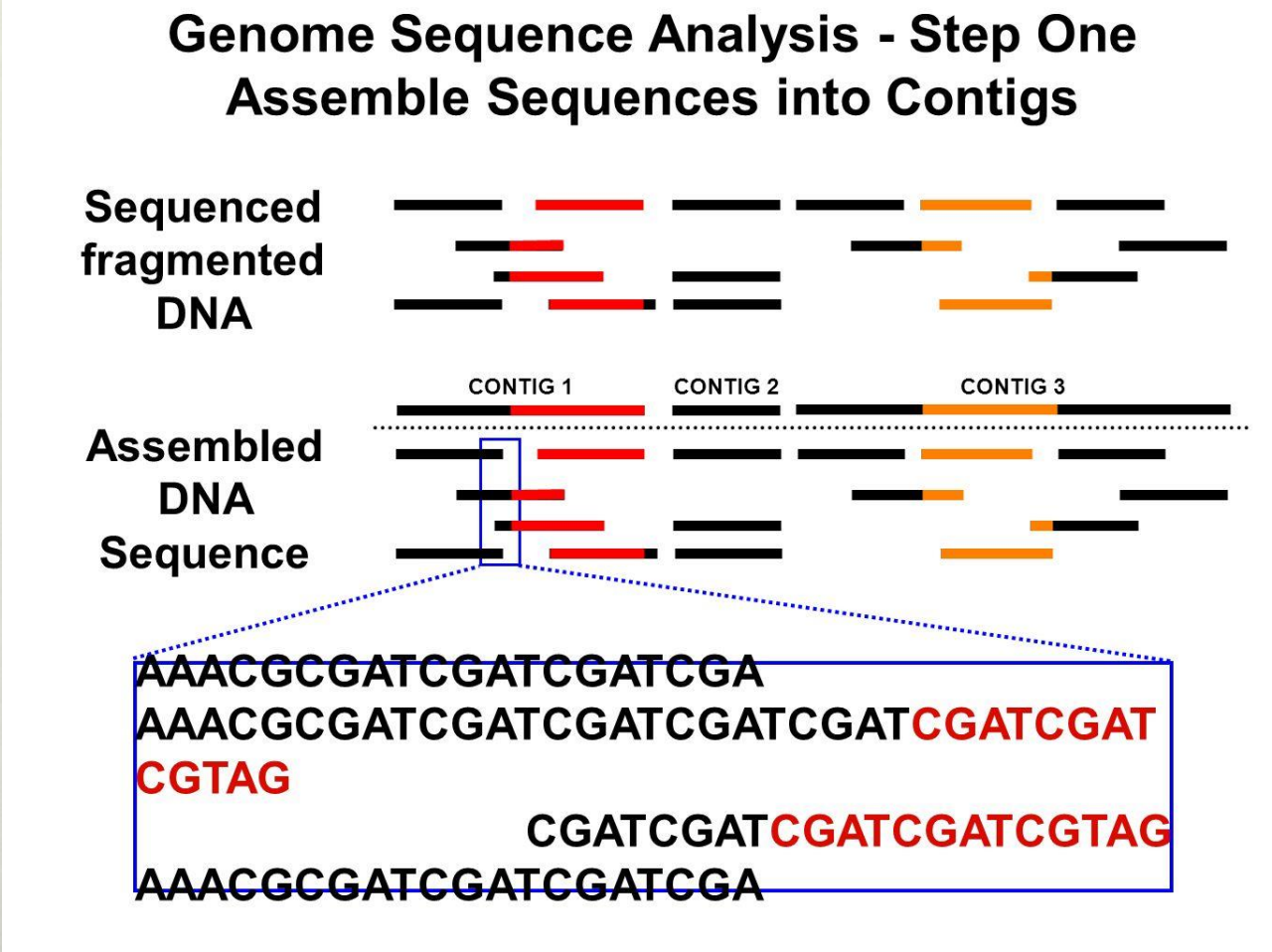
```
gtgctccgtg gggctcccat tatgttgtag ataatcaact agaagactat  154950196
aaggtccatg aaggcggaga aaatgtgcct tgaacttaag tgctcattca  154950246
ctggaattat tcactaattt gtatttttaa tatactctcg gaatttatat  154950296
aacatagcct gtatttctag ctgctgctac cccacaataa aattatataa  154950346
actctgacat caaaaaccag tttttctttc ctttcttggc cttgtaaaca  154950396
ggcatttgtc taagagactg tatctggtac taacataaat tccccacttc  154950446
cccgtttctg ttacagACAT GTCATGCTGC CCAAGGACAT AGCCAAGCTG  154950496
GTCCCTAAAA CCCATCTGAT GTCTGAATCT GAATGGAGGA ATCTTGGCGT  154950546
TCAGCAGAGT CAGGGATGGG TCCATTATAT GATCCATGAA CCAGgtcagt  154950596
gcactggcta aaaacaacca tatagaactg ctacactgag agaatgaaag  154950646
aataagattg tataacccaa atagggagat aggaaatggt ttactggttc  154950696
cttccccctc cagtcgtggg ggatttttt aaaaaaaaac tagtgaccaa  154950746
aaataagact aaaatatctg ggaagttcag agacaacctg tcactgaaaa  154950796
acctcctgta atctttcatt caatcagagg gtattctttt taaggccaca  154950846
tatagcctga tcatagcccc tgcctcattc tccatcgaaa acattcttgg  154950896
atgtgttcta aataagcaaa ggaaagtata tttattgata agacaccaga  154950946
cacccagctg ccaggcaaaa ctaataaggg acaccctggg gctgtataaa  154950996
catagcaaaa gaactgatat taacaattct gtacttggca gacagtccag  154951046
acttctgggt ctgcttctaa ggccatatgc ttaagtcttt atttagttat  154951096
aaagatctga gtgggtgata gctggggagg tggtagtgga atacactata  154951146
ggttgtacat agaatggtgt gagcttgtct cttagatttc cctcactctt  154951196
tcagAACCTC ACATCTTGCT GTTCCGGCGC CCACTACCCA AGAAACCAAA  154951246
GAAATGAAGC TGGCAAGCTA CTTTTCAGCC TCAAGCTTTA CACAGCTGTC  154951296
CTTACTTCCT AACATCTTTC TGATAACATT ATTATGTTGC CTTCTTGTTT  154951346
CTCACTTTGA TATTTAAAAG ATGTTCAATA CACTGTTTGA ATGTGCTGGT  154951396
AACTGCTTTG CTTCTTGAGT AGAGCCACCA CCACCATAGC CCAGCCAGAT  154951446
GAGTGCTCTG TGGACCCACA GCCTaAGCTG AGTGTGACCC CAGaagccac  154951496
gatgtgctct gtatccagaa cacacttggc agatggagga agcatctgag  154951546
tttgagacca tggctgttac agggatcatg taaacttgct gtt
```

Dos exones de NM_001252546 (chimpancé) alineados en el genoma humano (bases en azul)

```
000000263 aacctcacatcttgctgttccggcgcccactacccaagaaaccaaagaaa 000000312
>>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>>
154951201 aacctcacatcttgctgttccggcgcccactacccaagaaaccaaagaaa 154951250

000000313 tgaagctggcaagctacttttcagcctcaagctttacacagctgtcctta 000000362
>>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>>
154951251 tgaagctggcaagctacttttcagcctcaagctttacacagctgtcctta 154951300

000000363 cttcctaacatctttctgataacattattatgttgccttcttgtttctca 000000412
>>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>>
154951301 cttcctaacatctttctgataacattattatgttgccttcttgtttctca 154951350

000000413 ctttgatatttaaaagatgttcaatacactgtttgaatgtgctggtaact 000000462
>>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>>
154951351 ctttgatatttaaaagatgttcaatacactgtttgaatgtgctggtaact 154951400

000000463 gctttgcttcttgagtagagccaccaccaccatagcccagccagatgagt 000000512
>>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>>
154951401 gctttgcttcttgagtagagccaccaccaccatagcccagccagatgagt 154951450

000000513 gctctgtggacccacagcctcagctgagtgtgaccccag 000000551
>>>>>>>>>> ||||||||||||||||||||| ||||||||||||||||||| >>>>>>>>>>
154951451 gctctgtggacccacagcctaagctgagtgtgaccccag 154951489
```

¿Para que comparar dos secuencias?

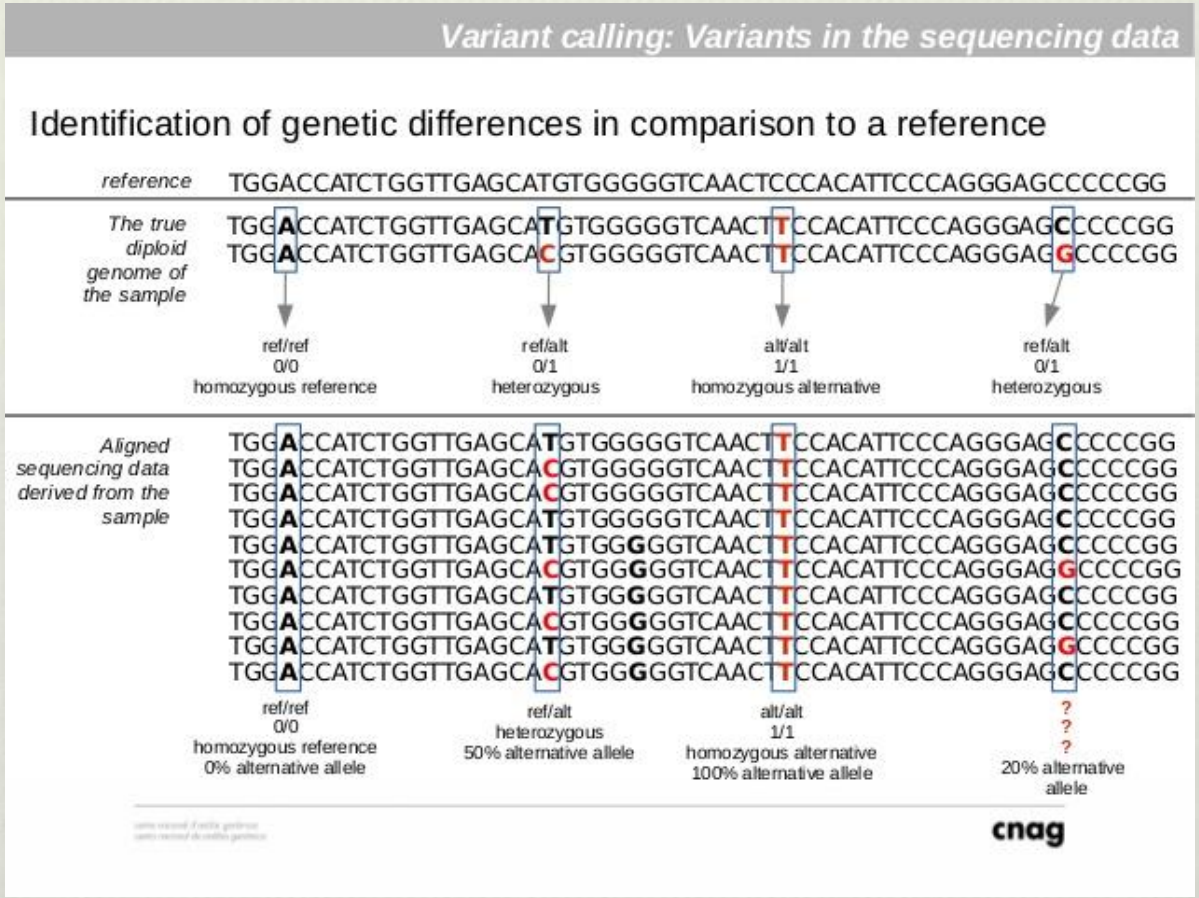
4) Ensamblar lecturas cortas (ensamblaje *de novo* de un genoma)



http://slideplayer.com/slide/9940475/

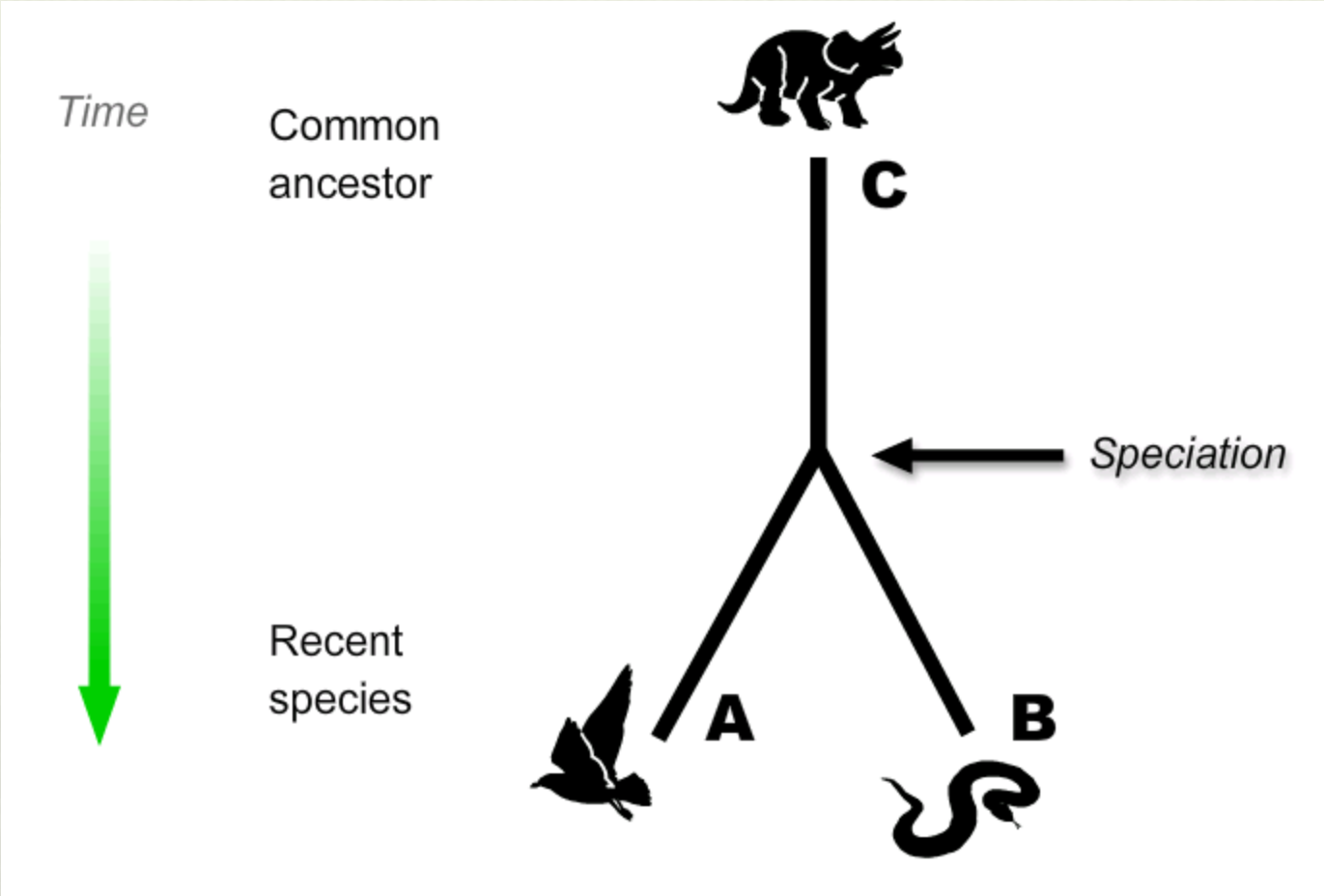
¿Para que comparar dos secuencias?

5) Resecuenciar un genoma para encontrar variantes de secuencia y cambios estructurales



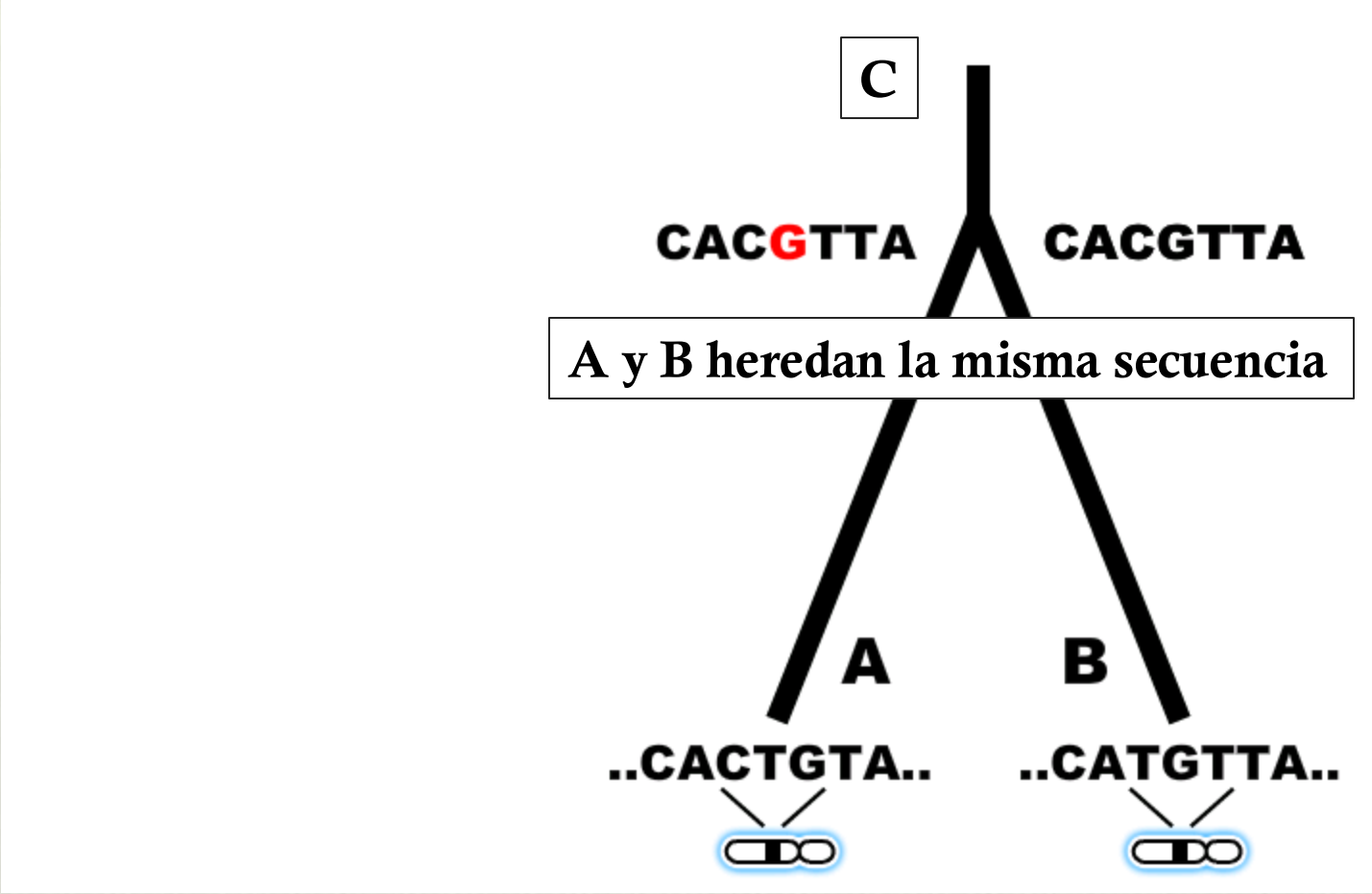
https://www.slideshare.net/VHIR/variant-calling-and-how-to-prioritize-somatic-mutations-and-inheritated-variants-in-ngs-data-an-advanced-overview

El alineamiento en un contexto evolutivo

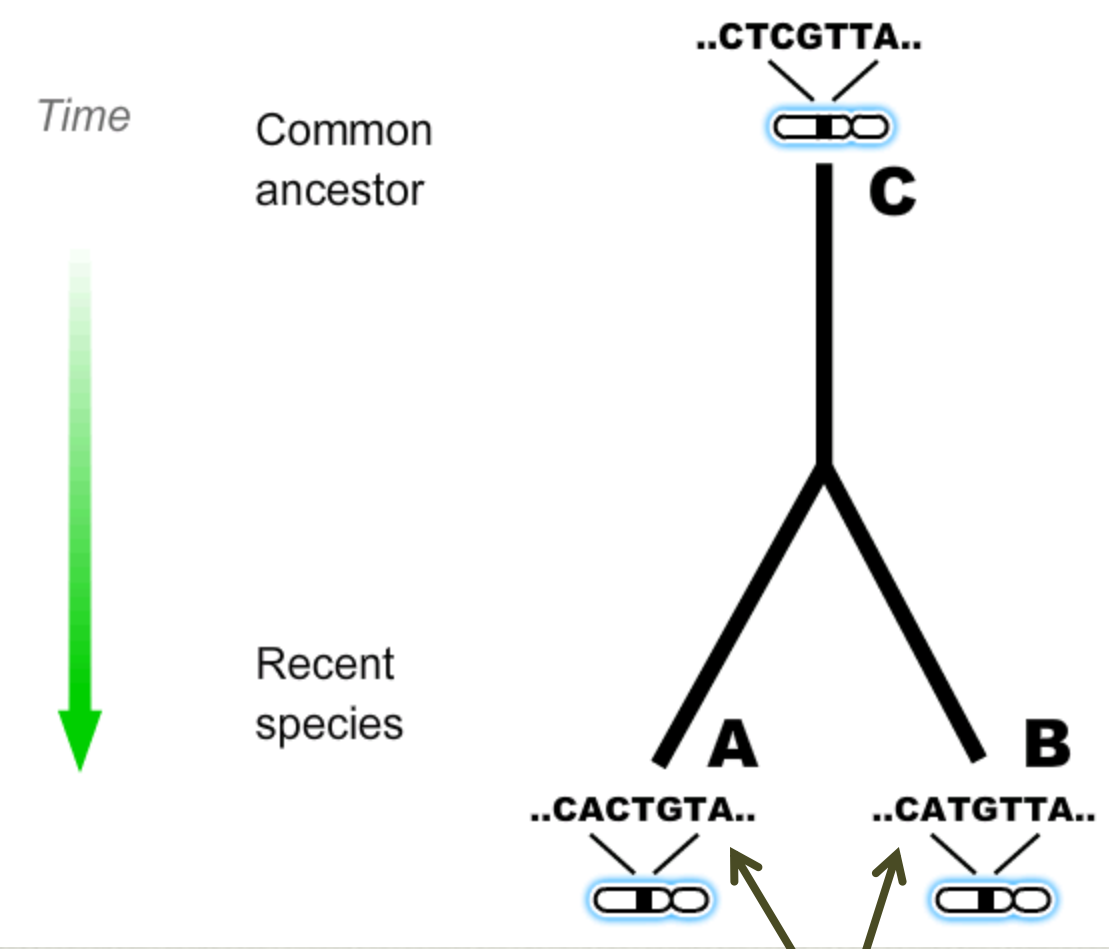


http://statgen.ncsu.edu/slse/animations/module1.html

Cambios evolutivos y el alineamiento



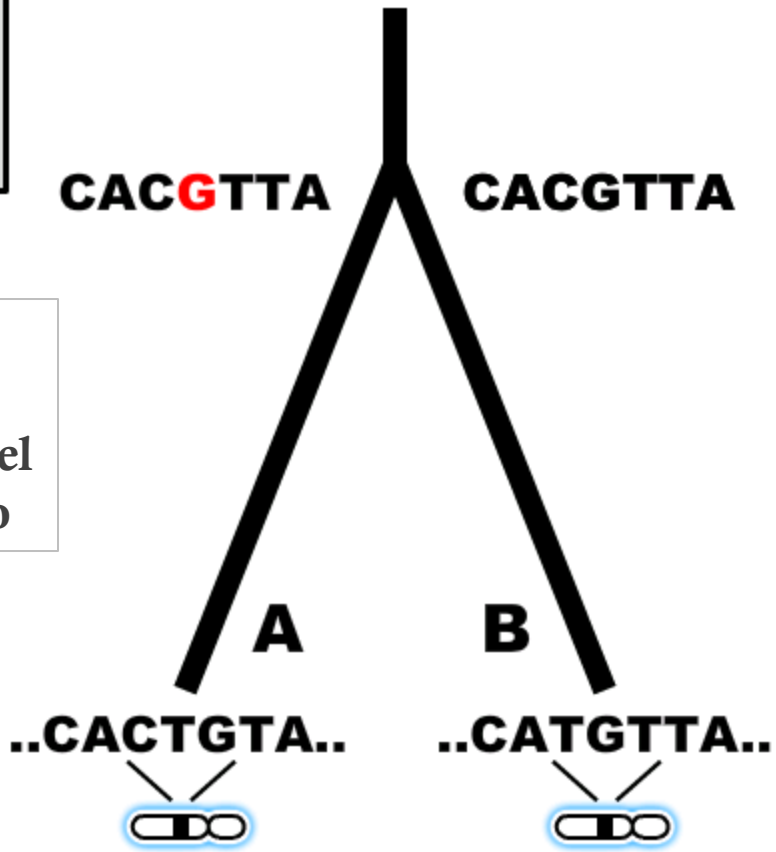
El alineamiento en un contexto evolutivo



Estas dos secuencian son similares debido a que derivan de un ancestro común y no debido al azar

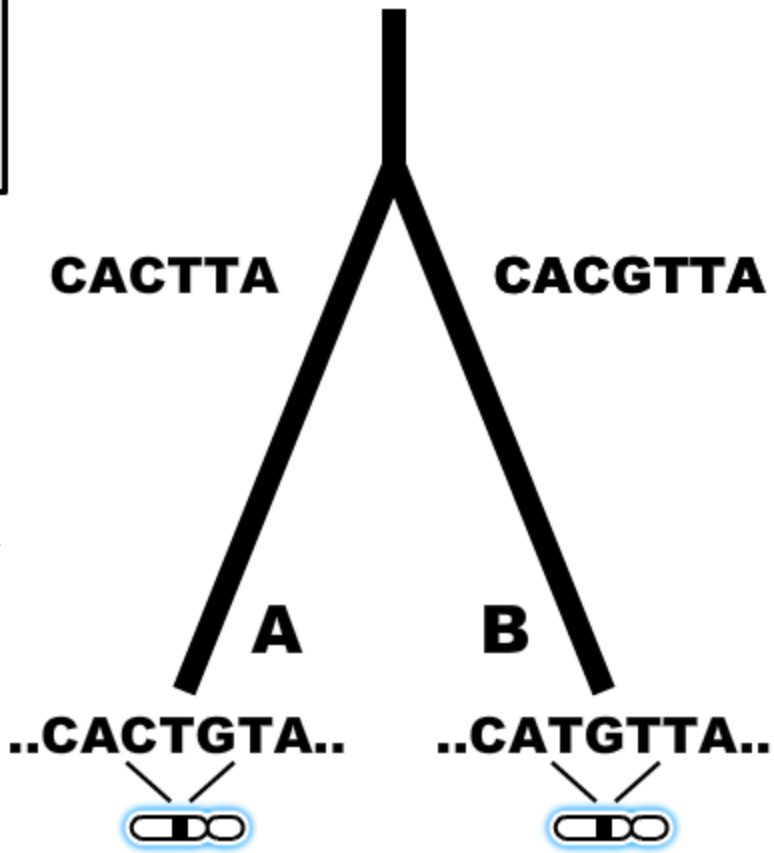
Cambios evolutivos y el alineamiento

En el momento de la especiación, las dos secuencias son idénticas y el 'alineamiento' es perfecto

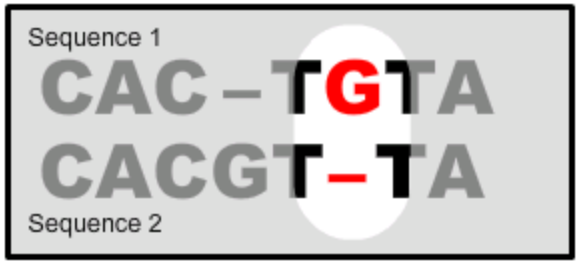


Cambios evolutivos y el alineamiento

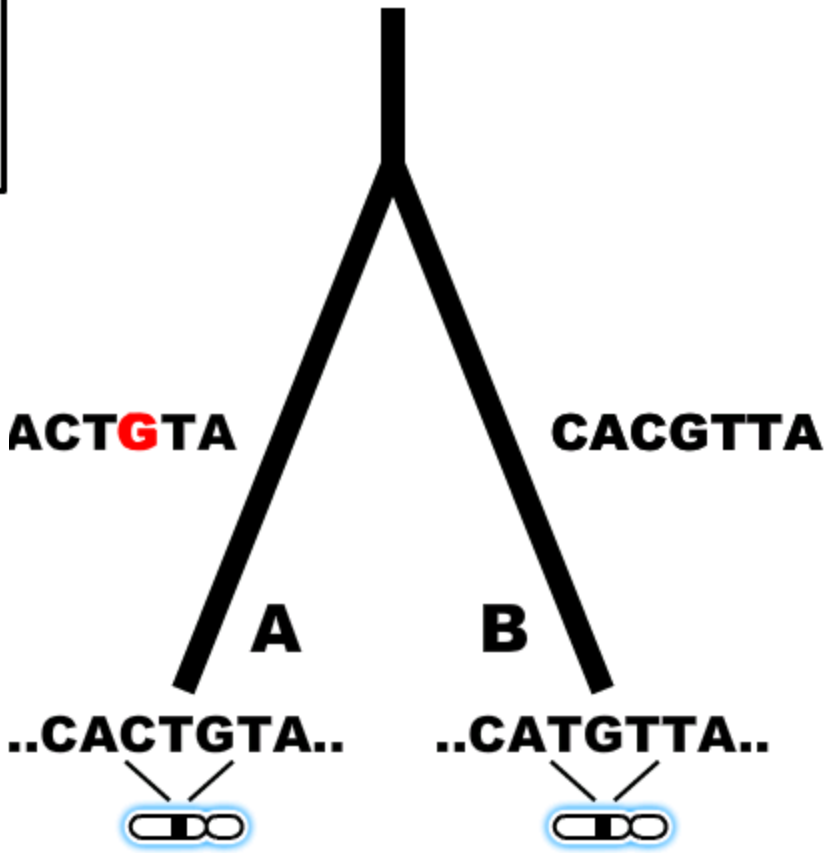
La deleción de un nucleótido de la secuencia 1 en la especie A se refleja en el alineamiento



Cambios evolutivos y el alineamiento

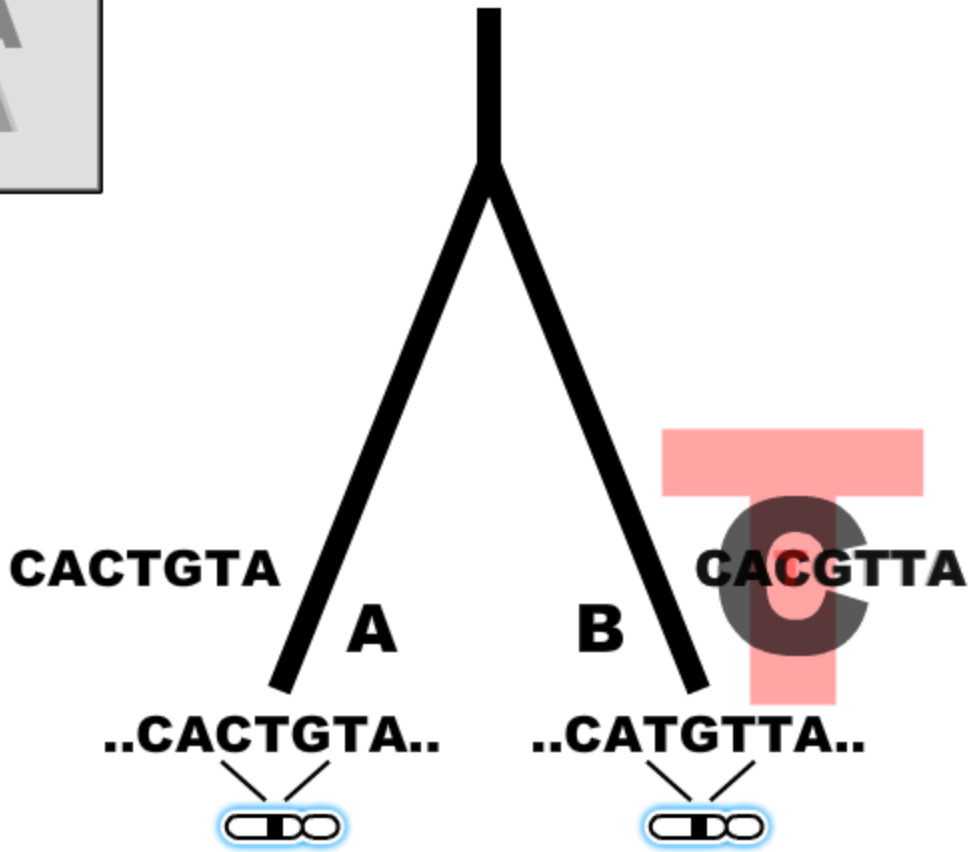


La inserción de un nucleótido en la secuencia 2 de la especie b se refleja en el alineamiento



Cambios evolutivos y el alineamiento

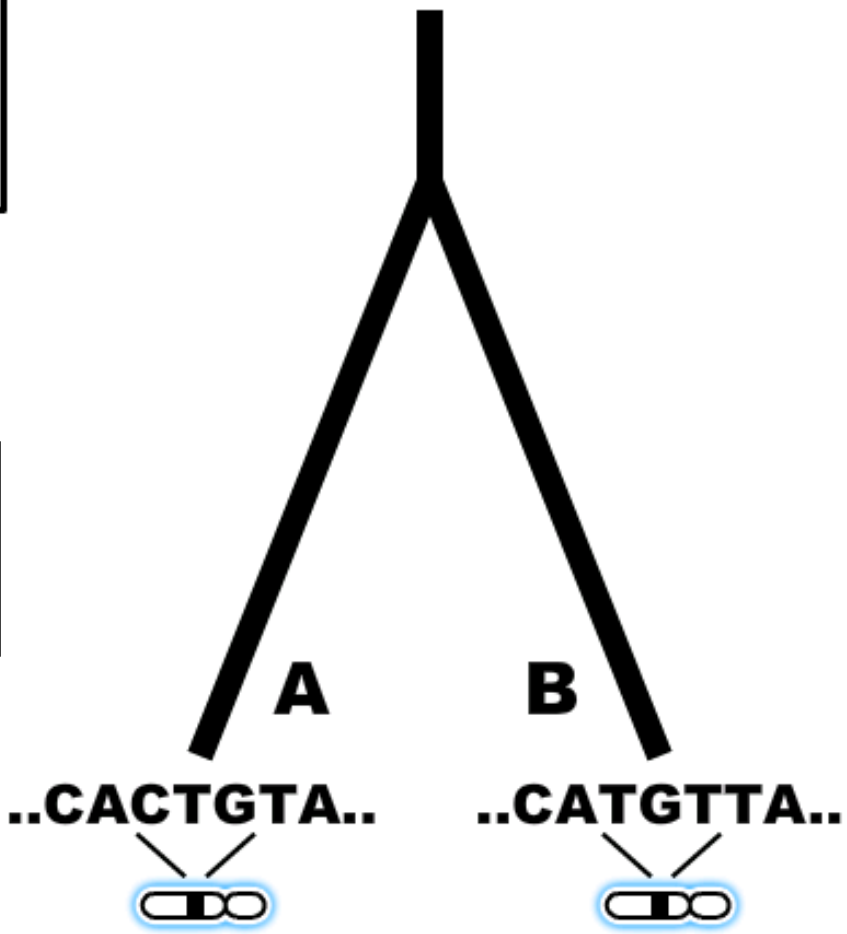
La sustitución de una C por una T en la especie B se refleja como desemparejamiento (mismatch) en el alineamiento



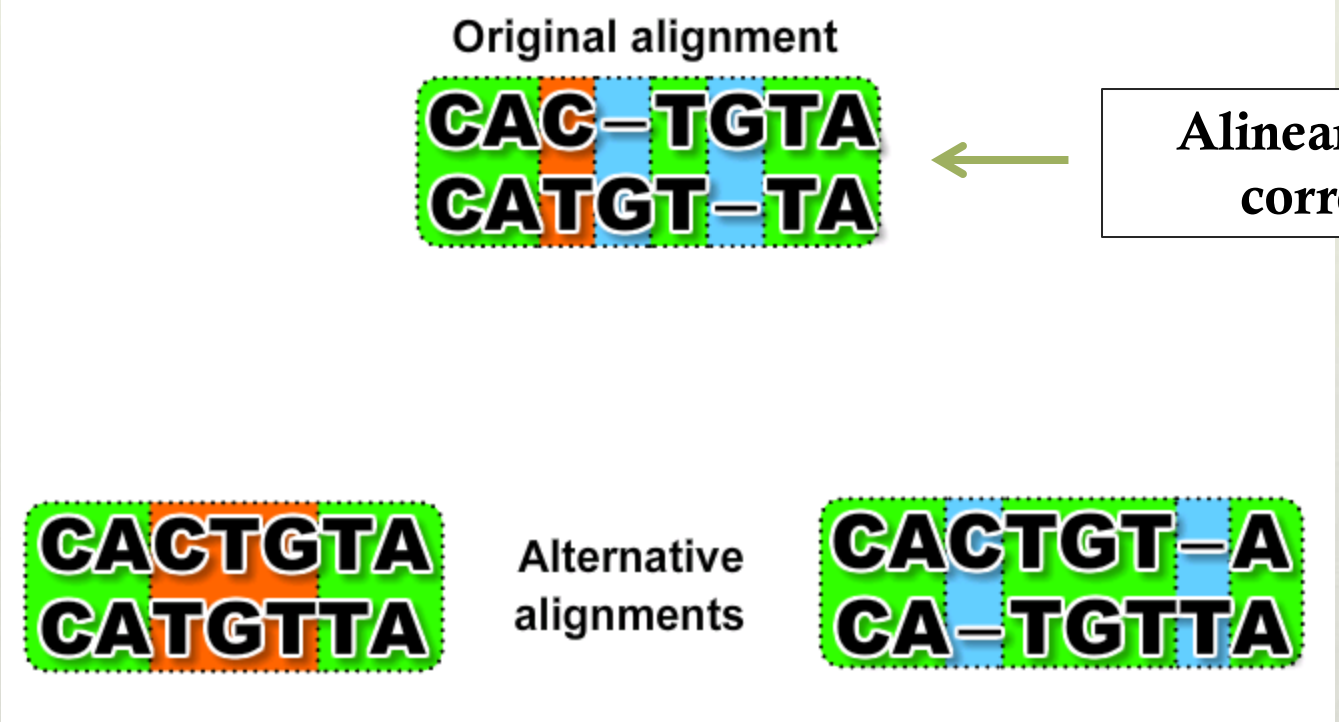
Cambios evolutivos y el alineamiento

Sequence 1
CAC–TGTA
CATGT–TA
Sequence 2

Este alineamiento refleja la 'historia' evolutiva de las dos secuencias



Cambios evolutivos y el alineamiento



Alineamiento correcto

OBJETIVO:
Encontrar el alineamiento que refleja la 'verdadera' historia evolutiva de las secuencias

→

Puntuar los alineamientos con el esquema de puntuación adecuado

Puntuar el alineamiento

Punto de partida:
Dos o mas secuencias

```
Sec1   ACGTATAGCG
Sec2   ACGTAGCG
```

Alinear las secuencias

```
ACGTATAGCG
|||||   |||
ACGTA--GCG
```

Dos huecos/ gaps

```
ACGTATAGCG
|||||   |
ACGTAGCG
```

Dos desemparejamientos

¿Cual de los dos alineamientos es mas probable?

Puntuar el alineamiento

```
ACGTATAGCG
|||||   |||
ACGTA--GCG
```

```
ACGTATAGCG
|||||   |
ACGTAGCG
```

Puntuar los alineamientos:
Emparejamiento: +1
Desemparejamiento: -1
Hueco: -5

Puntuación /Score: 3

Puntuación /Score: 4

El alineamiento mas probable dado el sistema de puntuación

Puntuar el alineamiento

```
ACGTATAGCG
|||||   |||
ACGTA--GCG
```

```
ACGTATAGCG
|||||   |
ACGTAGCG
```

Puntuar los alineamientos:
Emparejamiento: +1
Desemparejamiento: -2
Hueco: -5

Puntuación /Score:

Puntuación /Score:

Puntuar el alineamiento

```
ACGTATAGCG
|||||   |||
ACGTA--GCG
```

```
ACGTATAGCG
|||||   |
ACGTAGCG
```

Puntuar los alineamientos:
Emparejamiento: +1
Desemparejamiento: -2
Hueco: -5

Puntuación /Score: 3

Puntuación /Score: 2

↓

El alineamiento mas probable dado el sistema de puntuación

Puntuar el alineamiento

ACGTA**TA**GCG

ACGTA**TA**GCG

¡¡El alineamiento mas probable depende del sistema de puntación utilizado!!

El esquema de puntuación elegido depende del conocimiento previo que tengamos de nuestras secuencias y/o la finalidad concreta del alineamiento

Similitud

Similitud de secuencia = numero de coincidencias / longitud del alineamiento

```
ACGTATAGCG
|||||   |||
ACGTA--GCG
```

Ungapped: 100%

Gapped: 80%

```
ACGTATAGCG
|||||   |
ACGTAGCG
```

75%

Similitud y homologia

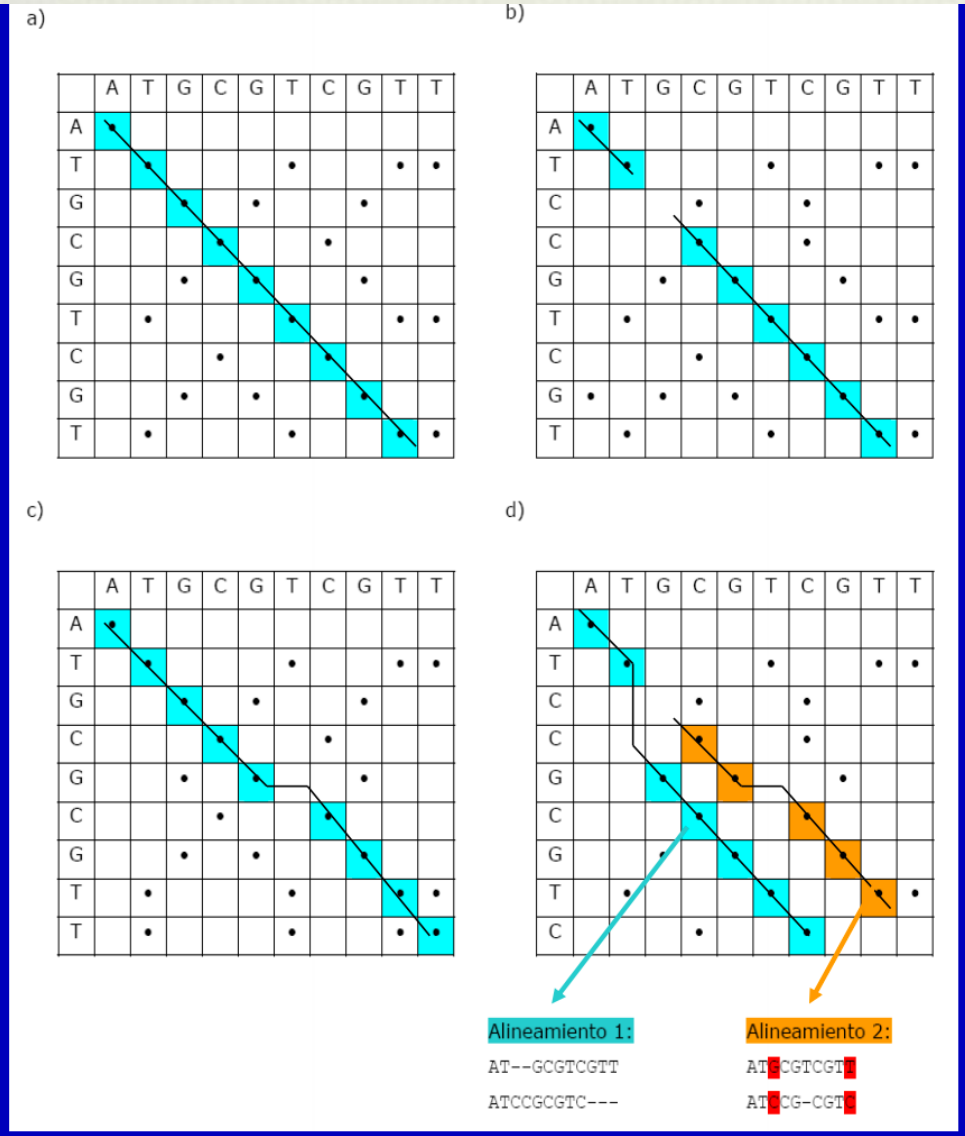
La similitud de secuencia NO ES LO MISMO que homología

Similitud de secuencia:
grado de similitud que hay entre dos secuencias

Homología:
las dos secuencias derivan de la misma secuencia ancestral presente en el ancestro común

Si observamos un alto grado de similitud de secuencia podemos deducir que las dos secuencias son homólogas
(son similares debido a la procedencia común y no debido al azar)

La matriz de puntos



a) **Las dos secuencias son idénticas en la parte alineada.**

b) **Las dos secuencias muestran un desemparejamiento debido a una sustitución; la posición (3,3) se queda en blanco.**

c) **Las dos secuencias difieren por una inserción/deleción (indel), dando lugar a un hueco o gap; nótese el quiebro o zig-zag de la diagonal principal.**

d) **Dos posibles alineamientos mostrando desemparejamientos y huecos. El alineamiento 1 supondría en total cinco huecos (o un hueco de dos nucleótidos y otro hueco terminal de tres nucleótidos) y ningún desemparejamiento, mientras que el alineamiento 2 supondría un hueco y dos desemparejamientos.**

Matriz de puntuación (BLOSUM62)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Matriz de puntuación (PAM250)

PAM = Point Accepted Mutations

PAM 250	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1

Nomenclatura

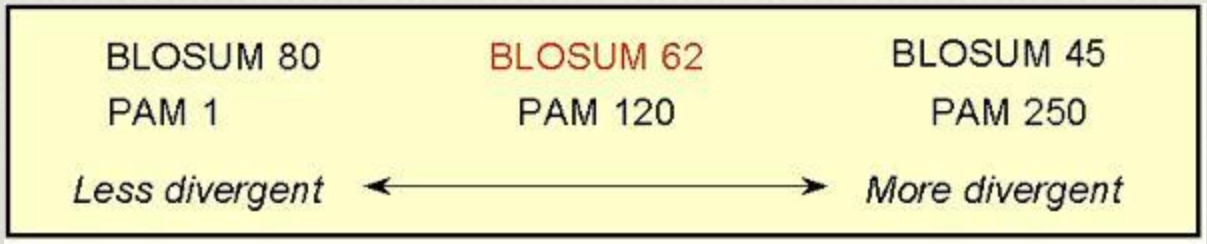
One letter code	Three letter code	Amino acid	Possible codons
A	Ala	Alanine	GCA, GCC, GCG, GCT
B	Asx	Asparagine or Aspartic acid	AAC, AAT, GAC, GAT
C	Cys	Cysteine	TGC, TGT
D	Asp	Aspartic acid	GAC, GAT
E	Glu	Glutamic acid	GAA, GAG
F	Phe	Phenylalanine	TTC, TTT
G	Gly	Glycine	GGA, GGC, GGG, GGT
H	His	Histidine	CAC, CAT
I	Ile	Isoleucine	ATA, ATC, ATT
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	CTA, CTC, CTG, CTT, TTA, TTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAC, AAT
P	Pro	Proline	CCA, CCC, CCG, CCT
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	AGA, AGG, CGA, CGC, CGG, CGT
S	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
T	Thr	Threonine	ACA, ACC, ACG, ACT
V	Val	Valine	GTA, GTC, GTG, GTT
W	Trp	Tryptophan	TGG
X	X	any codon	NNN
Y	Tyr	Tyrosine	TAC, TAT
Z	Glx	Glutamine or Glutamic acid	CAA, CAG, GAA, GAG
*	*	stop codon	TAA, TAG, TGA

Propiedades de los aminoácidos

Property		Amino acids
small		Ala, Gly
acidic / amide		Asp, Glu, Asn, Gln
charged	negative	Asp, Glu
	positive	Lys, Arg
polar		Ala, Gly, Ser, Thr, Pro
hydrophobic		Val, Leu, Ile, Met
size	big	Glu, Gln, His, Ile, Lys, Leu, Met, Phe, Trp, Tyr
	small	Ala, Asn, Asp, Cys, Gly, Pro, Ser, Thr, Val
aliphatic		Ile, Leu, Val
aromatic		His, Phe, Tyr, Trp

Comparación de las matrices

PAM1: 1% de los residuos han cambiado (secuencias altamente relacionadas)
PAM250: Sustituciones múltiples (2.5 por residuo) para secuencias remotamente relacionadas



Percent Accepted Mutation PAM	Blocks Substitution Matrix
Based on explicit evolutionary model	Based on empirical frequencies
Represents a specific evolutionary distance	Always a blend of distances as seen in the database and PROSITE
Ranges from identical to completely random	Narrower range than PAM matrix

Tipos de alineamientos

Global: se fuerza un alineamiento de las dos secuencias en su longitud completa

Local: búsqueda de subsecuencias con alto grado de similitud entre dos secuencias mas largas que pueden estar muy divergidas

```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

Illustration of global and local alignments demonstrating the 'gappy' quality of global alignments that can occur if sequences are insufficiently similar

Exacto: Un método exacto nos garantiza encontrar el mejor alineamiento entre dos secuencias (Smith-Waterman, Needleman- Wünsch): **LENTO**

Heurístico: Solución aproximada; no garantiza encontrar el mejor alineamiento (BLAST; BLAT): **MAS RAPIDO**

```
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

HBA_HUMAN          1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                     || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN          1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48
```

Secuencias de DNA: alfabeto y formato

A = adenina
C = citosina
G = guanina
T = timina

R = G A (pu**r**ina)
Y = T C (pirimidina - *pyrimidine*)
K = G T (ceto, en inglés ***keto***)
M = A C (a**m**ino)
S = G C (enlaces fuertes, en inglés ***strong***)
W = A T (enlaces débiles, en inglés ***weak***)
B = G T C (cualquiera excepto A)
D = G A T (cualquiera excepto C)
H = A C T (cualquiera excepto G)
V = G C A (cualquiera excepto T)
N = A G C T (cualquiera)

ID / nombre **Descripción**

```
>gi|4558520|gb|AF033819.3|   HIV-1, complete genome
GGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCC
TCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAACTAGAGA
TCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGCGCCCGAACAGGGACCTGAAAGCGAA
AGGGAAACCAGAGGAGCTCTCTCGACGCAGGACTCGGCTTGCTGAAGCGCGCACGGCAAGAGGCGAGGGG
CGGCGACTGGTGAGTACGCCAAAAATTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGCG
TCAGTATTAAGCGGGGGAGAATTAGATCGATGGGAAAAAATTCGGTTAAGGCCAGGGGGAAAGAAAAAAT
ATAAATTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAACGATTCGCAGTTAATCCTGGCCTGTTAGA
AACATCAGAAGGCTGTAGACAAATACTGGGACAGCTACAACCATCCCTTCAGACAGGATCAGAAGAACTT
AGATCATTATATAATACAGTAGCAACCCTCTATTGTGTGCATCAAAGGATAGAGATAAAAGACACCAAGG
AAGCTTTAGACAAGATAGAGGAAGAGCAAAACAAAAGTAAGAAAAAAGCACAGCAAGCAGCAGCTGACAC
AGGACACAGCAATCAGGTCAGCCAAAATTACCCTATAGTGCAGAACATCCAGGGGCAAATGGTACATCAG
GCCATATCACCTAGAACTTTAAATGCATGGGTAAAAGTAGTAGAAGAGAAGGCTTTCAGCCCAGAAGTGA
TACCCATGTTTTCAGCATTATCAGAAGGAGCCACCCCACAAGATTTAAACACCATGCTAAACACAGTGGG
GGGACATCAAGCAGCCATGCAAATGTTAAAAGAGACCATCAATGAGGAAGCTGCAGAATGGGATAGAGTG
CATCCAGTGCATGCAGGGCCTATTGCACCAGGCCAGATGAGAGAACCAAGGGGAAGTGACATAGCAGGAA
CTACTAGTACCCTTCAGGAACAAATAGGATGGATGACAAATAATCCACCTATCCCAGTAGGAGAAATTTA
TAAAAGATGGATAATCCTGGGATTAAATAAAATAGTAAGAATGTATAGCCCTACCAGCATTCTGGACATA
AGACAAGGACCAAAGGAACCCTTTAGAGACTATGTAGACCGGTTCTATAAAACTCTAAGAGCCGAGCAAG
CTTCACAGGAGGTAAAAAATTGGATGACAGAAACCTTGTTGGTCCAAAATGCGAACCCAGATTGTAAGAC
TATTTTAAAAGCATTGGGACCAGCGGCTACACTAGAAGAAATGATGACAGCATGTCAGGGAGTAGGAGGA
CCCGGCCATAAGGCAAGAGTTTTGGCTGAAGCAATGAGCCAAGTAACAAATTCAGCTACCATAATGATGC
AGAGAGGCAATTTTAGGAACCAAAGAAAGATTGTTAAGTGTTTCAATTGTGGCAAAGAAGGGCACACAGC
CAGAAATTGCAGGGCCCCTAGGAAAAAGGGCTGTTGGAAATGTGGAAAGGAAGGACACCAAATGAAAGAT
TGTACTGAGAGACAGGCTAATTTTTTAGGGAAGATCTGGCCTTCCTACAAGGGAAGGCCAGGGAATTTTC
TTCAGAGCAGACCAGAGCCAACAGCCCCACCAGAAGAGAGCTTCAGGTCTGGGGTAGAGACAACAACTCC
CCCTCAGAAGCAGGAGCCGATAGACAAGGAACTGTATCCTTTAACTTCCCTCAGGTCACTCTTTGGCAAC
GACCCCTCGTCACAATAAAGATAGGGGGGCAACTAAAGGAAGCTCTATTAGATACAGGAGCAGATGATAC
AGTATTAGAAGAAATGAGTTTGCCAGGAAGATGGAAACCAAAAATGATAGGGGGAATTGGAGGTTTTATC
AAAGTAAGACAGTATGATCAGATACTCATAGAAATCTGTGGACATAAAGCTATAGGTACAGTATTAGTAG
GACCTACACCTGTCAACATAATTGGAAGAAATCTGTTGACTCAGATTGGTTGCACTTTAAATTTTCCCAT
```

Secuencias de proteínas: alfabeto y formato

One letter code	Three letter code	Amino acid
A	Ala	Alanine
B	Asx	Asparagine or Aspartic acid
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
X	X	any codon
Y	Tyr	Tyrosine
Z	Glx	Glutamine or Glutamic acid
*	*	stop codon

```
>gi|28872819|ref|NP_057849.4| Gag-Pol [Human immunodeficiency virus 1]
MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGLLETSEGCRQILGQLQPSLQT
GSEELRSLYNTVATLYCVHQRIEIKDTKEALDKIEEEQNKSKKKAQQAAADTGHSNQVSQNYPIVQNIQG
QMVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLKETINEEAA
EWDRVHPVHAGPIAPGQMREPRGSDIAGTTSTLQEQIGWMTNNPPIPVGEIYKRWIILGLNKIVRMYSPT
SILDIRQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTAC
QGVGGPGHKARVLAEAMSQVTNSATIMMQRGNFRNQRKIVKCFNCGKEGHTARNCRAPRKKGCWKCGKEG
HQMKDCTERQANFLREDLAFLQGKAREFSSEQTRANSPTRRELQVWGRDNNSPSEAGADRQGTVSFNFPQ
VTLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAI
GTVLVGPTPVNIIGRNLLTQIGCTLNFPISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEK
EGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGD
AYFSVPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQY
MDDLYVGSDLEIGQHRTKIEELRQHLLRWGLTTPDKKHQKEPPFLWMGYELHPDKWTVQPIVLPEKDSWT
VNDIQKLVGKLNWASQIYPGIKVRQLCKLLRGTKALTEVIPLTEEAELELAENREILKEPVHGVYYDPSK
DLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARMRGAHTNDVKQLTEAVQKITTESIVIWGKTPKFKLPI
QKETWETWWTEYWQATWIPEWEFVNTPPLVKLWYQLEKEPIVGAETFYVDGAANRETKLGKAGYVTNRGR
QKVVTLTDTTNQKTELQAIYLALQDSGLEVNIVTDSQYALGIIQAQPDQSESELVNQIIEQLIKKEKVYL
AWVPAHKGIGGNEQVDKLVSAGIRKVLFLDGIDKAQDEHEKYHSNWRAMASDFNLPPVVAKEIVASCDKC
QLKGEAMHGQVDCSPGIWQLDCTHLEGKVILVAVHVASGYIEAEVIPAETGQETAYFLLKLAGRWPVKTI
HTDNGSNFTGATVRAACWWAGIKQEFGIPYNPQSQGVVESMNKELKKIIGQVRDQAEHLKTAVQMAVFIH
NFKRKGGIGGYSAGERIVDIIATDIQTKELQKQITKIQNFRVYYRDSRNPLWKGPAKLLWKGEGAVVIQD
NSDIKVVPRRKAKIIRDYGKQMAGDDCVASRQDED
```

Coordenadas del alineamiento relativas a la secuencia de entrada

Coordenadas cromosómicas del alineamiento

Longitud de la región en el cromosoma

Score/puntuación

Similitud de secuencia

ACTIONS		QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser	details	NM_000346.4	3929	1	3931	3931	100.0%	chr17	+	72121020	72126416	5397
browser	details	NM_000346.4	244	586	912	3931	91.0%	chr16	+	982121	983836	1716
browser	details	NM_000346.4	211	571	803	3931	95.3%	chr22	-	37983357	37983589	233
browser	details	NM_000346.4	127	690	902	3931	82.3%	chr20	-	64048851	64049259	409
browser	details	NM_000346.4	108	685	902	3931	74.8%	chr13	+	112067809	112068026	218
browser	details	NM_000346.4	100	684	887	3931	74.6%	chr3	+	181712480	181712683	204
browser	details	NM_000346.4	82	686	857	3931	73.9%	chr20	+	326043	326214	172
browser	details	NM_000346.4	81	681	858	3931	90.2%	chr6	+	21594705	21594882	178
browser	details	NM_000346.4	73	820	906	3931	92.0%	chr8	+	54459087	54459173	87
browser	details	NM_000346.4	62	690	791	3931	80.4%	chr8	-	10730195	10730296	102
browser	details	NM_000346.4	55	686	830	3931	69.0%	chr3	+	137764807	137764951	145
browser	details	NM_000346.4	42	675	717	3931	100.0%	chr13	-	94711996	94712044	49
browser	details	NM_000346.4	35	819	859	3931	92.7%	chr17	-	7589358	7589398	41
browser	details	NM_000346.4	32	686	719	3931	97.1%	chr1	+	204123700	204123733	34
browser	details	NM_000346.4	31	3414	3503	3931	94.3%	chr19	+	48279421	48279551	131
browser	details	NM_000346.4	30	3411	3442	3931	96.9%	chr3	-	116220181	116220212	32
browser	details	NM_000346.4	29	2527	2563	3931	84.4%	chr16	-	71518534	71518568	35