

Alineamiento: Análisis computacional de secuencias

Andrés M. Pinzón
Centro de Bioinformática
Instituto de Biotecnología
Universidad Nacional de Colombia

Por qué y para qué...

Tengo una secuencia de DNA/Proteína...

¿... es esta secuencia **homóloga** a otra(s) secuencia(s)?

¿... o tal vez tiene una **función** idéntica o similar a otra(s)?

¿... quizás tiene una **estructura común** con otras secuencias, como **dominios** o **motivos** ?

El **Alineamiento de Secuencias** me permite responder estas y otras preguntas.

Es el procedimiento más común en bioinformática

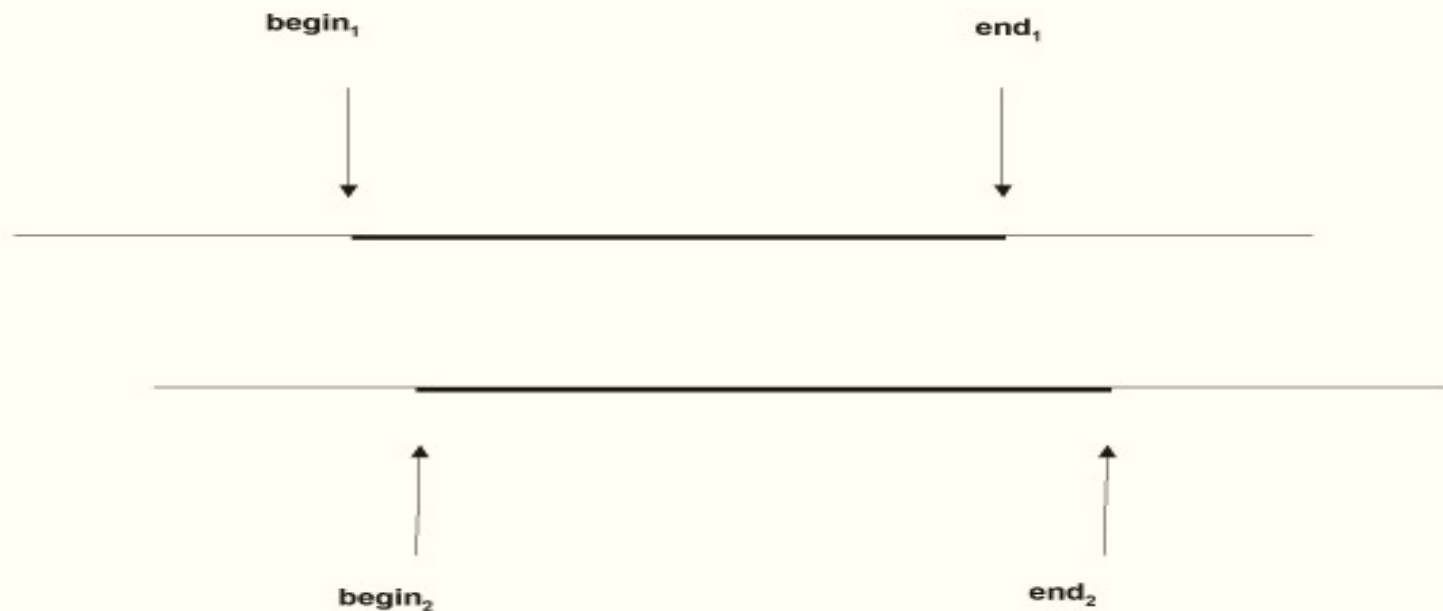
¿En qué consiste un alineamiento de secuencias?

Sec. 1

MALVEDNN~~AVAVS~~FLRNSFSEEQILVKYGV~~DG~~

Sec. 2

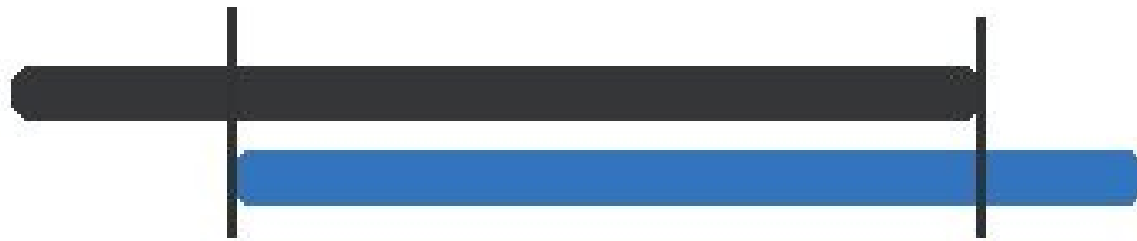
HKSF~~EDNN~~AVAVSSEFALLEQILVK~~DG~~



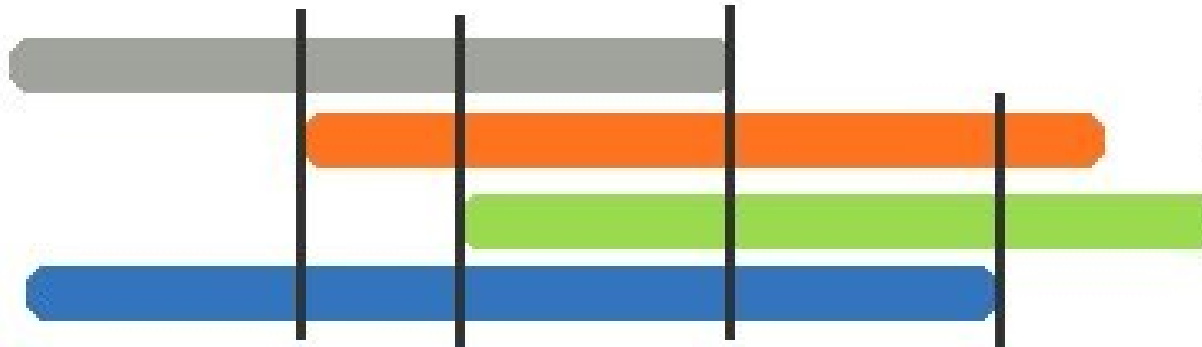
Existen diferentes tipos de alineamientos...

De acuerdo al número de secuencias a alinear:

- Pareado ($=2$)
- Múltiple (>2)



Pareado



Múltiple

De acuerdo a la región a alinear:

- Local (sub región de la secuencia)
- Global (secuencia completa)



Local



Global

¿Cómo luce un alineamiento de secuencias?

Human Chr.21 37218151-37247016,+
Chimp Chr.22 37250843-37279887,+

```
37218151 GGAGCAATGCACAAGCTCTCTAGTTC----- 37218176
          ||||||||||||||||||||||||||||
37250843 GGAGCAATGCACAAGCTCTCTAGTTGCAAGGTAAAGGTGCGACCTTTCT 37250892

37218177 -----CCCAAGGTAAAGGTGCGACCTTCTGTTGGG 37218206
          | ||||||||||||||||||||||||||||
37250893 GGGGGCAGCAGTCCAGTTCTCAAGGTAAAGGTGCGACCTTCTGTTGGG 37250942

37218207 GCAGCACGTCCAGTTCCCAAGGTAAAGGTGCGACCTTCTAGGGGTGG 37218256
          |||||||||||||||| ||||||||||||||||||||||||||||
37250943 GCAGCACGTCCAGTTACCAAGGTAAAGGTGCGACCTTCTAGGGGTGG 37250992

37218257 CACATCCACCTTCTTCTGCTTCTGACTCACTGGGGGGACCAAAG 37218306
          |||||||||||||||||||||||||||| ||||||||||||||||||
37250993 CACATCCACCTTCTTCTGCTTCTGGCTCACTGGGGGGACCAAAG 37251042

37218307 GGGTCCACTGGCAGGCAACTCTGCAGGGGTGGAGGGTGGGGGGCTGGCTC 37218356
          |||||||||||||||||||||||||||| ||||||||||||||||
37251043 GGGTCCACTGGCAGGCAACTCTGCAGGGGTGGAGGGTGGGGGGCTGGCTC 37251092

37218357 TGCCCAACACACAGATTGAGGTCCAGCGGTCTAAGCCGACAGCTGC 37218406
```

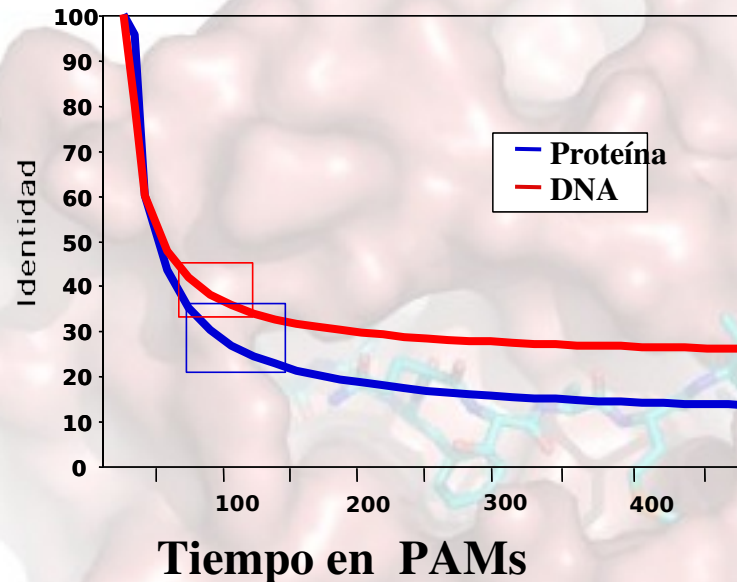
Similaridad, identidad y homología

Similaridad es una cantidad observable que puede ser expresada como, por ejemplo, un porcentaje de identidad.

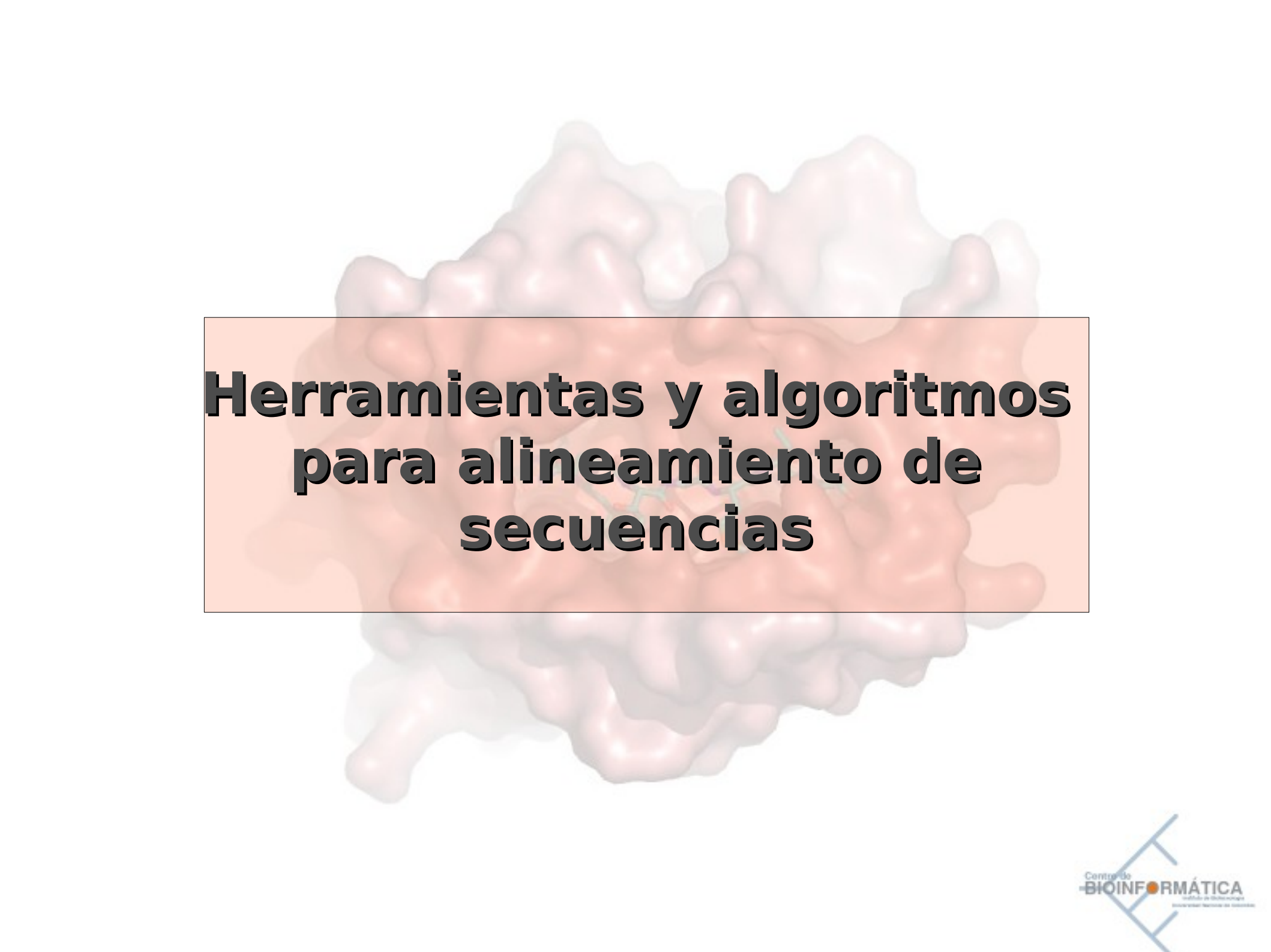
Homología es una conclusión basada, en parte, en la similaridad y de la que se infiere una relación evolutiva entre nuestras secuencias.

Un gen es o no es homólogo NO existe un “porcentaje de homología”.

La divergencia ocasiona que el parecido entre secuencias se aproxime asintóticamente al azar



- Al aumentar la divergencia se dificulta distinguir homólogos de no homólogos.
- Las comparaciones de proteína conservan mas información que las de ADN y pueden detectar homología mas remota.



Herramientas y algoritmos para alineamiento de secuencias

Dos algoritmos fundamentales...

- Smith and waterman (local)
- Needleman Wunsch (global), 1970: genera un alineamiento óptimo aún introduciendo gaps.

Aproximaciones heurísticas de Smith and waterman:

- FASTA (Pearson and Lipman, 1998)
- BLAST (altschul et al., 1990)

Elementos característicos de un alineamiento

Human Chr.21 37218151-37247016,+
Chimp Chr.22 37250843-37279887,+

Gap (in/del)

```
37218151 GGAGCAATGCACAAGCTCTCTAGTTC 37218176
          |||||
37250843 GGAGCAATGCACAAGCTCTCTAGTTGCAAGGTAAAGGTGCGACCTTTCT 37250892
          Match

37218177 -----CCCAAGGTAAAGGTGCGACCTTCTGTTGGG 37218206
          |||||
37250893 GGGGGCAGCAAGTCCAGTTCTCAAGGTAAAGGTGCGACCTTCTGTTGGG 37250942
          Mismatch

37218207 GCAGCAAGTCCAGTTCCCAAGGTAAAGGTGCGACCTTCTAGGGGTGG 37218256
          |||||
37250943 GCAGCAAGTCCAGTTACCAAGGTAAAGGTGCGACCTTCTAGGGGTGG 37250992
          |||||

37218257 CACATCCACCTTCTTCTGCTTCTGCTCACTGGGGGGACCAAAG 37218306
          |||||
37250993 CACATCCACCTTCTTCTGCTTCTGCTCACTGGGGGGACCAAAG 37251042
          |||||

37218307 GGTCCACTGGCAGGCAACTCTGCAGGGGTGGAGGGTGGGGGGCTGGCTC 37218356
          |||||
37251043 GGTCCACTGGCAGGCAACTCTGCAGGGGTGGAGGGTGGGGGGCTGGCTC 37251092
          |||||

37218357 TGGGGACACACAGATTCCAGGTCCAGGGGTCTAACGGGACAGCTGGC 37218406
```

En definitiva...

Lo que buscamos con un alineamiento óptimo es reducir al mínimo los “gaps” y los “mismatches” y maximizar los “matches”

Para lograrlo se debe implementar una manera de “calificar” el alineamiento con base en esas 3 variables!

SISTEMAS DE PUNTUACIÓN

- Deben definirse para poder cuantificar la *similitud* entre dos cadenas S y T .
- Así para cada alineamiento se calcula un número que, a mayor valor, represente la mayor significancia (biológica)
- Una vez definido el sistema a usar, la *puntuación* (score) de una pareja de caracteres s, t alineados se define como $p(s, t)$



SISTEMAS DE PUNTUACIÓN

Puede ser un esquema sencillo como (para DNA):

A

Match = 1

Mismatch = -2

B

Match = 1

Transición = -1.5

Transversión = -2

C

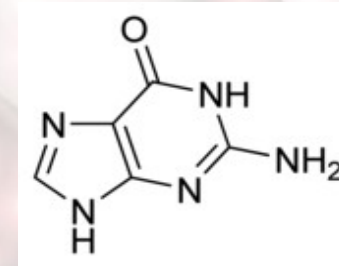
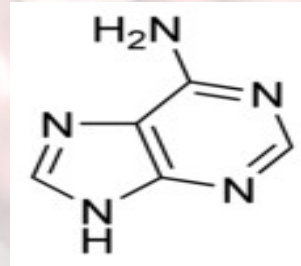
Match = 1

Transición = -1.5

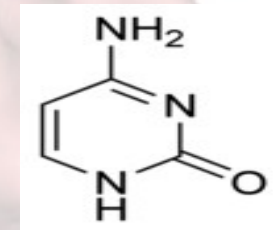
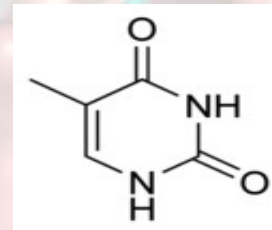
Transversión = -2

Gap Opening = -5

Gap extension = -2



Purinas: **A**denina - **G**uanina



Pirimidinas: **T**imina - **C**itosina

- El sistema para puntuar los apareamientos entre AA debe reflejar su relación química y biológica

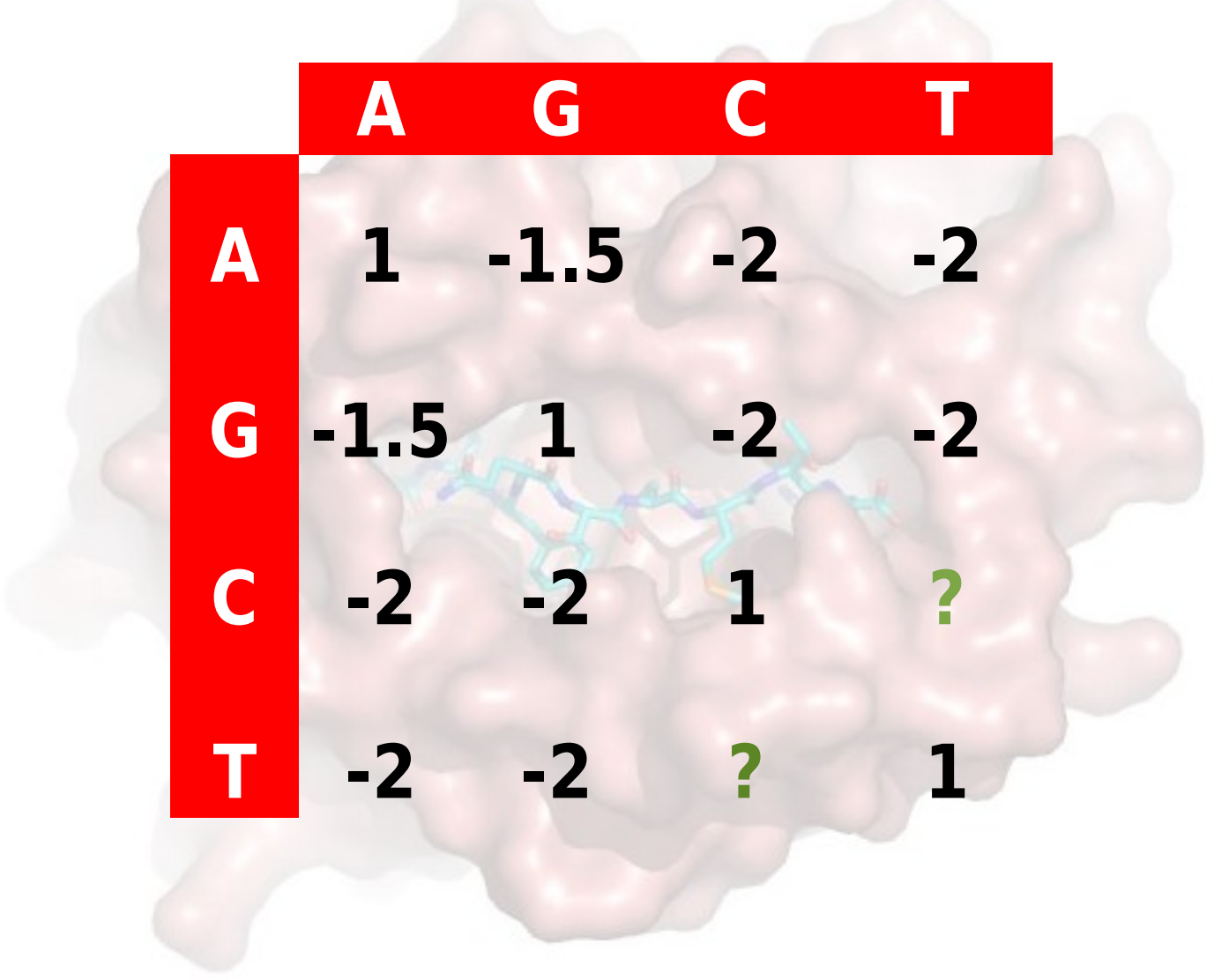
MATRICES DE SUSTITUCIÓN

Contienen valores proporcionales a la probabilidad de que un aminoácido 'i' sea substituído por un aminoácido 'j', calculados para todos los pares de aminoácidos posibles.

[illegible]

El término “**sustitución**” indica que el objetivo de puntear un apareamiento es valorar el coste evolutivo y funcional de cambiar un residuo por otro

Una matriz sencilla...



	A	G	C	T
A	1	-1.5	-2	-2
G	-1.5	1	-2	-2
C	-2	-2	1	?
T	-2	-2	?	1

Tipos de matrices...

- No hay una matriz única que se pueda usar siempre.
- Se utilizan según la familia de proteínas y el grado de similitud esperado.
- Las más utilizadas.
 - **PAM:** *Percent Accepted Mutation Matrix*
 - Derivadas de alineamientos globales de secuencias cercanamente relacionadas.
 - PAM40 PAM250. A mayor N° mayor distancia evolutiva
 - **BLOSUM**
 - Derivadas de alineamientos locales de secuencias distantes
 - BLOSUM90 BLOSUM45 El N° representa porcentaje de identidad



PAM

Porcentaje de mutaciones aceptadas

- La cuestión era observar qué sustituciones ocurrían entre proteínas homólogas.
- Accepted Point Mutation (PAM): cambio, en una proteína, de un aa por otro que ha sido "aceptado" por la selección natural (SN).
- los cambios más aceptados por la SN son los de tipo conservativo, ej. isoleucina por valina (hidrofóbicos y de tamaño similar).
- Permite relacionar familias de proteínas por su distancia evolutiva.



PAM1/PAM250

- Con base en las probabilidades de ocurrencia de cada aa se generó una matriz de probabilidad de mutación.
- Probabilidad de que un aa sea mutado en otro, durante un periodo evolutivo definido.
- Un intervalo de estos es por ejemplo de 1 PAM: "La unidad de divergencia evolutiva en la que 1% de los aa han sido cambiados entre 2 secuencias dadas".
- Notar que la unidad NO está dada en años, sino en % de divergencia.
- PAM 250: distancia evolutiva en la que las proteínas comparten cerca del 20% de identidad. Es decir en la que el 80% de los aa se han cambiado.

Pam250 Matrix

C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
11.5	0.1	-0.5	-3.1	0.5	-2.0	-1.8	-3.2	-3.0	-2.4	-1.3	-2.2	-2.8	-0.9	-1.1	-1.5	0.0	-0.8	-0.5	-1.0	C
	2.2	1.5	0.4	1.1	0.4	0.9	0.5	0.2	0.2	-0.2	-0.2	0.1	-1.4	-1.8	-2.1	-1.0	-2.8	-1.9	-3.3	S
		2.5	0.1	0.6	-1.1	0.5	0.0	-0.1	0.0	-0.3	-0.2	0.1	-0.6	-0.6	-1.3	0.0	-2.2	-1.9	-3.5	T
			7.6	0.3	-1.6	-0.9	-0.7	-0.5	-0.2	-1.1	-0.9	-0.6	-2.4	-2.6	-2.3	-1.8	-3.8	-3.1	-5.0	P
				2.4	0.5	-0.3	-0.3	0.0	-0.2	-0.8	-0.6	-0.4	-0.7	-0.8	-1.2	0.1	-2.3	-2.2	-3.6	A
					6.6	0.4	0.1	-0.8	-1.0	-1.4	-1.0	-1.1	-3.5	-4.5	-4.4	-3.3	-5.2	-4.0	-4.0	G
						3.8	2.2	0.9	0.7	1.2	0.3	0.8	-2.2	-2.8	-3.0	-2.2	-3.1	-1.4	-3.6	N
							4.7	2.7	0.9	0.4	-0.3	0.5	-3.0	-3.8	-4.0	-2.9	-4.5	-2.8	-5.2	D
								3.6	1.7	0.4	0.4	1.2	-2.0	-2.7	-2.8	-1.9	-3.9	-2.7	-4.3	E
									2.7	1.2	1.5	1.5	-1.0	-1.9	-1.6	-1.5	-2.6	-1.7	-2.7	Q
										6.0	0.6	0.6	-1.3	-2.2	-1.9	-2.0	-0.1	2.2	-0.8	H
											4.7	2.7	-1.7	-2.4	-2.2	-2.0	-3.2	-1.8	-1.6	R
												3.2	-1.4	-2.1	-2.1	-1.7	-3.3	-2.1	-3.5	K
													4.3	2.5	2.8	1.6	1.6	-0.2	-1.0	M
														4.0	2.8	3.1	1.0	-0.7	-1.8	I
															4.0	1.8	2.0	0.0	-0.7	L
																3.4	0.1	-1.1	-2.6	V
																	7.0	5.1	3.6	F
																		7.8	4.1	Y
																			14.2	W

- Residue Identity
- Hydrophobic Similarity
- Hydrophilic Similarity

BLOSUM

(matriz de sustitución de bloques)

- Steve Henikoff (1992)
- Los valores están basados en la observación de frecuencias de sustitución en bloques de ALINEAMIENTOS LOCALES de proteínas relacionadas
- De esta manera se enfoca en las regiones conservadas.
- BLOSUM62, une todas las proteínas en un alineamiento que $\geq 62\%$ de identidad en sus aminoácidos.
- se penalizan más para aquellas secuencias que comparten menos del 62% de identidad.



BLOSUM 62

	A	C	D	E	F	G	H	
A	4	0	-2	-1	-2	0	-2	
C	0	9	-3	-4	-2	-3	-3	
D	-2	-3	6	2	-3	-1	-1	
E	-1	-4	2	5	-3	-2	0	
F	-2	-2	-3	-3	6	-3	-1	
G	0	-3	-1	-2	-3	4	0	
H	-2	-3	-1	0	-1	0	3	

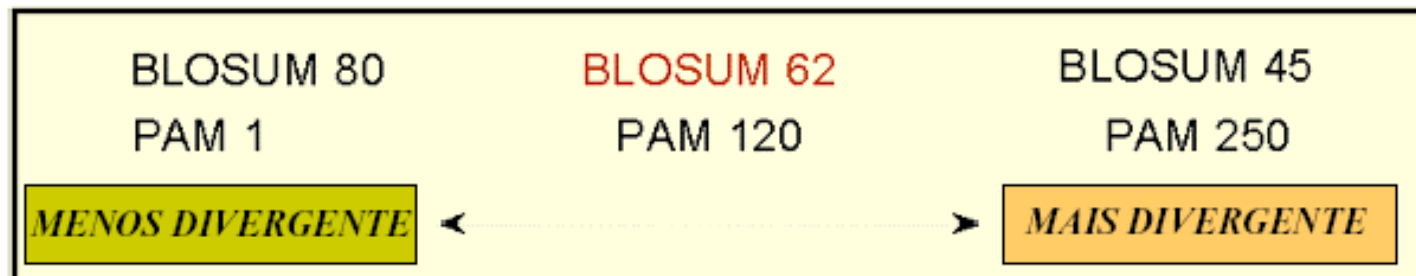
BLOSUM 62

El número hace referencia al mínimo porcentaje de homología de los bloques usados para construir la Matriz.

BLOSUM 62 fue calculada a partir de la comparación de secuencias con una divergencia mínima de 62%

BLOSUM62 Substitution Scoring Matrix. The BLOSUM 62 matrix shown here is a 20 x 20 matrix of which a section is shown here in which every possible identity and substitution is assigned a score based on the observed frequencies of such occurrences in alignments of related proteins. Identities are assigned the most positive scores. Frequently observed substitutions also receive positive scores and seldom observed substitutions are given negative scores.

PAM	BLOSUM
Calculada a partir de alineamientos globales	Calculada a partir de alineamientos locales
Secuencias de proteínas usadas en alineamiento tienen >99% de identidad	Puede seleccionar un nivel de similaridad entre las secuencias usadas en el análisis
Las matrices son extrapolaciones matemáticas de la matriz PAM1	Cada matriz es el resultado del análisis de un alineamiento de bloques conservados (análisis real)
Es posible montar un modelo evolutivo y así generar nuevas matrices a partir de la primera	No permite generar un modelo evolutivo
	Permiten detectar las mejores secuencias con relación biológica



PENALIZACIÓN POR “GAPS”

- En un sistema de puntuación es importante definir el coste de insertar o eliminar un residuo, lo que en el alineamiento aparece como un hueco (“gap”)
- Suele penalizarse diferente
 - el primer hueco (“gap opening”)
 - que los restantes (“gap extension”) que parten de él
- La variación de estos parámetros puede tener efectos importantes en el alineamiento final



EFFECTO DEL VALOR DE LA PENALIZACIÓN

Costo de apertura de gap	Costo de extensión del gap	Comentario
Grande	Grande	<ul style="list-style-type: none">• Pocas inserciones o eliminaciones• Adecuado para proteínas muy relacionadas
Grande	Pequeño	<ul style="list-style-type: none">• Algunas inserciones grandes• Adecuado si hay la posibilidad que se hayan insertado dominios completos
Pequeño	Grande	<ul style="list-style-type: none">• Muchas inserciones pequeñas• Adecuado si se trata de proteínas distantes



```
xterm <3>

1 #####
2 # Program: needle
3 # Rundate: Wed May 17 2006 14:13:25
4 # Align_format: srspair
5 # Report_file: needle.needle
6 #####
7
8 #=====
9 #
10 # Aligned_sequences: 2
11 # 1: seq1
12 # 2: seq2
13 # Matrix: EDNAFULL
14 # Gap_penalty: 10.0
15 # Extend_penalty: 0.5
16 #
17 # Length: 33
18 # Identity:      19/33 (57.6%)
19 # Similarity:    19/33 (57.6%)
20 # Gaps:          12/33 (36.4%)
21 # Score: 56.0
22 #
23 #
24 #=====
25
26 seq1      1 -----ag-tgctgtga-ctacgatcgataca      24
27              || |||| | ||||| |||||.|.|
28 seq2      1 gctcgatagatgct---agctacgatcgaaaaa      30
29
30
31 #-----
32 #-----

needle.needle      1,1      Todo
"water.water" 32L, 735C escritos

1 #####
2 # Program: water
3 # Rundate: Wed May 17 2006 14:13:06
4 # Align_format: srspair
5 # Report_file: water.water
6 #####
7
8 #=====
9 #
10 # Aligned_sequences: 2
11 # 1: seq1
12 # 2: seq2
13 # Matrix: EDNAFULL
14 # Gap_penalty: 10.0
15 # Extend_penalty: 0.5
16 #
17 # Length: 23
18 # Identity:      17/23 (73.9%)
19 # Similarity:    17/23 (73.9%)
20 # Gaps:          4/23 (17.4%)
21 # Score: 56.0
22 #
23 #
24 #=====
25
26 seq1      3 tgctgtga-ctacgatcgataca      24
27              |||| | ||||| |||||.|.|
28 seq2     11 tgct---agctacgatcgaaaaa      30
29
30
31 #-----
32 #-----

water.water      1,1      Todo
```



Herramientas para alineamiento

Alineamiento local

- water
- NCBI-BLAST, WU-BLAST
- FASTA

Alineamiento global:


- needle (EMBOSS)

Alineamiento múltiple:

- clustal-w
- MUSCLE
- T-COFFEE

<http://bioinf.ibun.unal.edu.co/dotlet/>

EMBnet Colombia

Inicio  :: Centro de Bioinformática del Instituto de Biotecnología :: Mapa del Sitio

Inicio | **Quiénes somos** | Documentos | Cursos | Contacto | 27 de marzo de 2006



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Herramientas

| [EMBOSS](#) | [BLAST](#) | [BLEE](#) | [SMS SEWER](#) | [BACTNAME](#) | [SINCO](#) | [DOTLET](#)

Puede consultar nuestra página de ayuda para mayor [información](#) [acerca de estas herramientas](#).

Curso Microarreglos

Del 28 de Noviembre al 2 de Diciembre de 2005 tendrá lugar el curso Internacional: "**Introducción al Análisis de Microarreglos**", organizado por el CBIB en colaboración con varios grupos e institutos a nivel mundial. [Mayor información](#)

Curso Bioinformática

Curso: "**Bioinformática: Introducción al Análisis de secuencias**", organizado por el CBIB en colaboración con el Laboratorio de Investigación en Sistemas Inteligentes. [Mayor información](#)

Usuari@s de Mac: Al parecer existe un problema en varias versiones del navegador Netscape, el cual hace que Dotlet falle con el error **SecurityViolation**. Existe también un error en Microsoft Internet Explorer el cual no permite las operaciones de cortar y pegar. Lo lamentamos, pero la solución de estos problemas está fuera de nuestro alcance.

[Acerca de dotlet](#) [Ayuda](#) [Aprende por medio de ejemplos](#)

Input one or more sequence(s)

seq_1

seq_1


Blosum62

15

1:1

compute

horizontal: seq_1
vertical: seq_1
matrix: Blosum62
sliding window: 15
zoom: 1:1
score range: -60 to 165
gray scale: 0% - 100%



seq_1|8
RPPPLAVBMKRGICLAALLCLFNLYLGAGHGENAEDIRNSEGKVNFFSLDSNLKKNKSKHNVRKR
RPPPLAVBMKRGICLAALLCLFNLYLGAGHGENAEDIRNSEGKVNFFSLDSNLKKNKSKHNVRKR
seq_1|8

Entrada

Matriz

Zoom

input seq_1 seq_1 Blosum62 15 1:1 compute

Tamaño de ventana

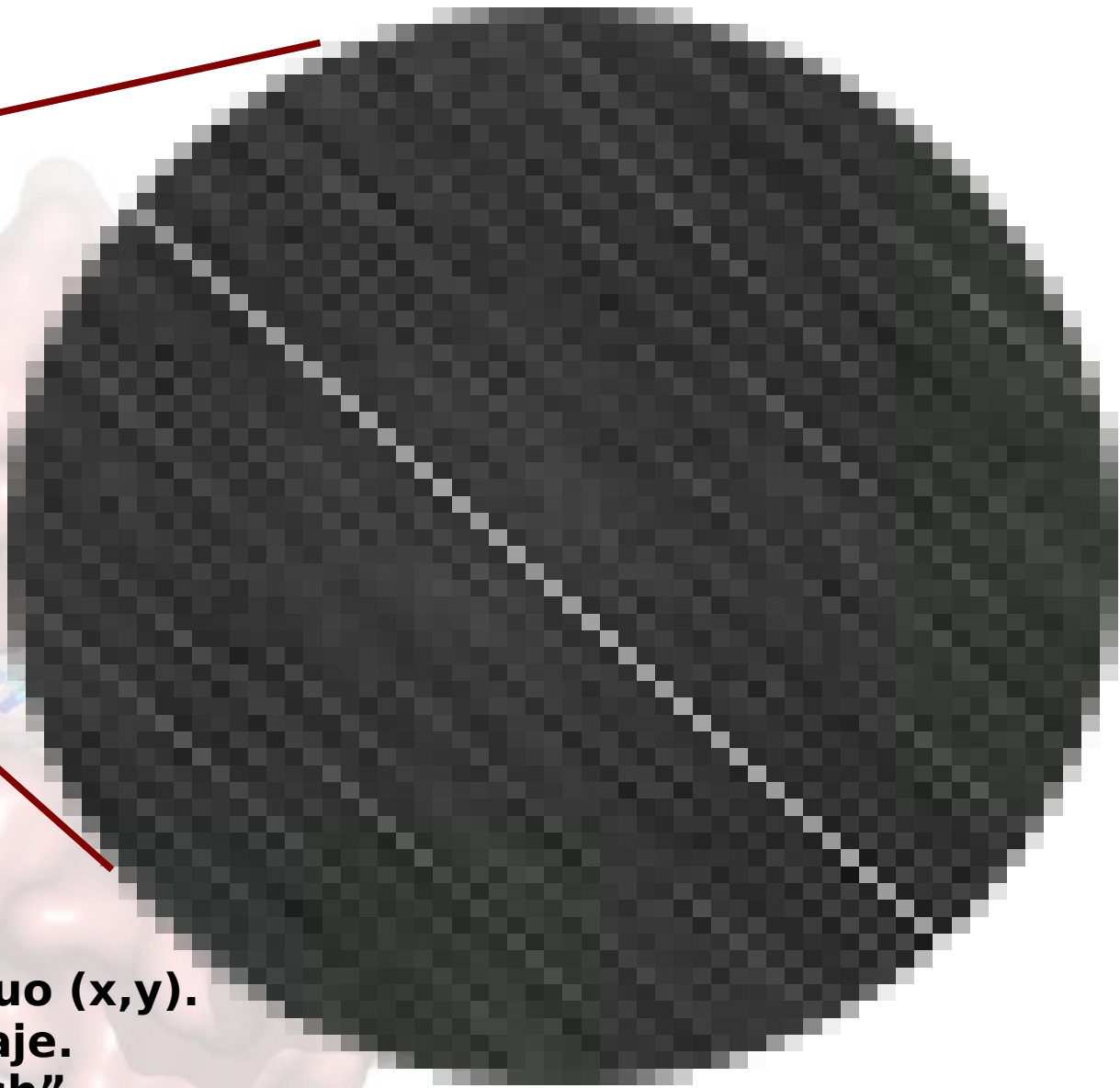
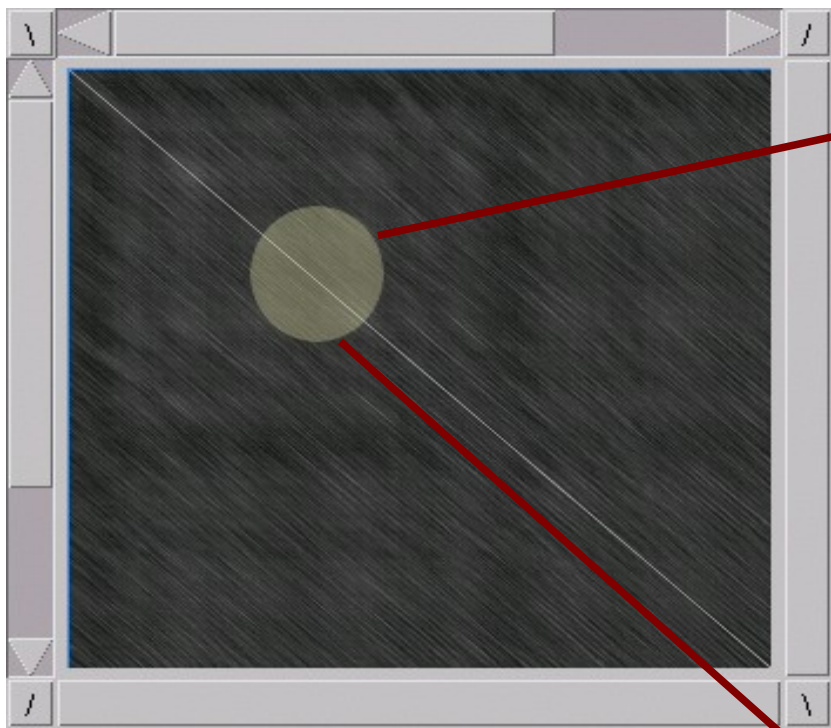
horizontal: seq_1
vertical: seq_1
matrix: Blosum62
sliding window: 15
zoom: 1:1
score range: -60 to 165
gray scale: 0% - 100%

Ventana de puntos

Histograma

```
seq_1|8  
RBPPLAVBMKRGICLAALLCLFNLYLGAGHGENAEDIRNSEGKVNFFSLDSNLKKNKSKHNRVKRR  
RBPPLAVBMKRGICLAALLCLFNLYLGAGHGENAEDIRNSEGKVNFFSLDSNLKKNKSKHNRVKRR  
seq_1|8
```

Ventana de alineamiento



Cada pixel representa un residuo (x,y).
Cada pixel representa un puntaje.
Un alto puntaje == buen “match”.



Bajo

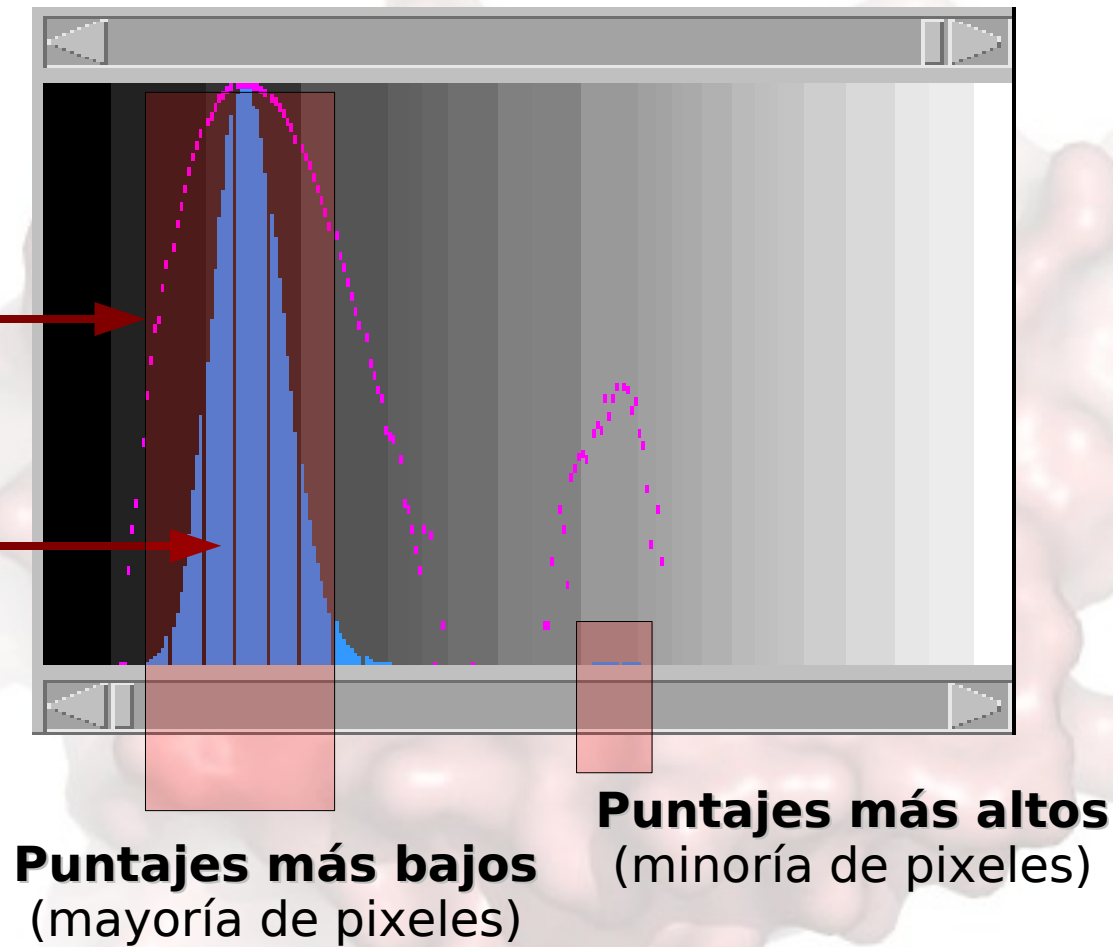
Alto

Histograma

Representa la frecuencia de cada score

Escala
logaritmica

Escala
lineal



BLAST

Basic Local Alignment Search Tool

NCBI →
BLAST

Latest news: 6 December 2005 : BLAST 2.2.13 released

About

- Getting started
- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobulin BLAST (IgBlast)
- SNP BLAST

Meta

- Retrieve results



WU-BLAST

Welcome to the Washington University BLAST Archives

Serving the world community since 1995, now celebrating more than a decade of fast gapped alignments

Faster at any sensitivity, more sensitive at any speed, the *original* gapped BLAST with statistics, providing the performance, features and reliability demanded by technical professionals:

- WU BLAST 2.0 ... setting a higher standard

The latest release is dated [22-Mar-2006] and is free for academic and nonprofit use.

Licensing information is available [here](#).

If you're not using WU BLAST, you don't know what you're missing!

Announcements

[18-Mar-2006] If system panics are experienced when using 64-bit binaries under Mac OS X 10.4 "Tiger", restrict these binaries to using a single CPU (cpus=1).

[02-Nov-2005] **New! EXALIN** — an improved spliced alignment program

[09-Jun-2005] "Universal" binaries are now available for Mac OS X 10.4 on Intel and PowerPC processors

[21-May-2005] 64-bit Solaris™ X64 is now supported

[22-Mar-2005] **Tabular** and XML output are now supported; See the **mformat** option.

[10-Dec-2003] **Licensing** of BLAST 2.0 for academic and nonprofit use can now be performed on-line.

General and specialized search services running WU BLAST 2.0

- European Bioinformatics Institute BLAST server at Hinxton
- Institut Pasteur
- Biowulf at the NIH

<http://blast.wustl.edu/>

<http://ncbi.nlm.nih.gov/BLAST/>



Nucleotide

- ◆ Quickly search for highly similar sequences (megablast)
- ◆ Quickly search for divergent sequences (discontiguous megablast)
- ◆ Nucleotide-nucleotide BLAST (blastn)
- ◆ Search for short, nearly exact matches
- ◆ Search trace archives with megablast or discontiguous megablast

Protein

- ◆ Protein-protein BLAST (blastp)
- ◆ Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- ◆ Search for short, nearly exact matches
- ◆ Search the conserved domain database (rpsblast)
- ◆ Protein homology by domain architecture (cdart)

Translated

- ◆ Translated query vs. protein database (blastx)
- ◆ Protein query vs. translated database (tblastn)
- ◆ Translated query vs. translated database (tblastx)

Genomes

- ◆ Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- ◆ Chicken, puffer fish, zebrafish
- ◆ Fly, honey bee, other insects
- ◆ Microbes, environmental samples
- ◆ Plants, nematodes
- ◆ Fungi, protozoa, other eukaryotes

Special

- ◆ Search for gene expression data (GEO BLAST)
- ◆ Align two sequences (bl2seq)
- ◆ Screen for vector contamination (VecScreen)
- ◆ Immunoglobulin BLAST (IgBlast)
- ◆ SNP BLAST

Meta

- ◆ Retrieve results


[Search](#)

[Set subsequence](#)

From: To:

[Choose database](#)

Now:

BLAST! or  **Reset all**

Options for advanced blasting

[Limit by
entrez query](#)

or select from:

[Choose filter](#)

☒ Low complexity ☐ Human repeats ☐ Mask for lookup table only ☐ Mask lower case

[Expect](#)

[Word Size](#)

[Other](#)

AGCGTGCTCGTGATAGATAGACAGATAGACAGTCGCTCGCTGAGATAGATAGATGAACAA
TGCCTGATGATAGATAGATAGACAGATAGTGC TAGCTACACACAAAAA

[Search](#)

[Set subsequence](#)

From: To:

[Choose database](#)

Now: **BLAST!** or

Options for advanced blasting

[Limit by](#) or select from:

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

[Expect](#)

[Word Size](#)

[Other](#)

Your request has been successfully submitted and put into the Blast Queue.

Query = (110 letters)

The request ID is

Format! or **Reset all**

The results are estimated to be ready in 10 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the results of a different search by entering any other valid request ID to see other recent jobs.

Format

Show ☒ [Graphical Overview](#) ☒ [Linkout](#) ☒ [Sequence Retrieval](#) ☒ [NCBI-genbank](#) Alignment in [format](#)

☐ [CDS feature](#)

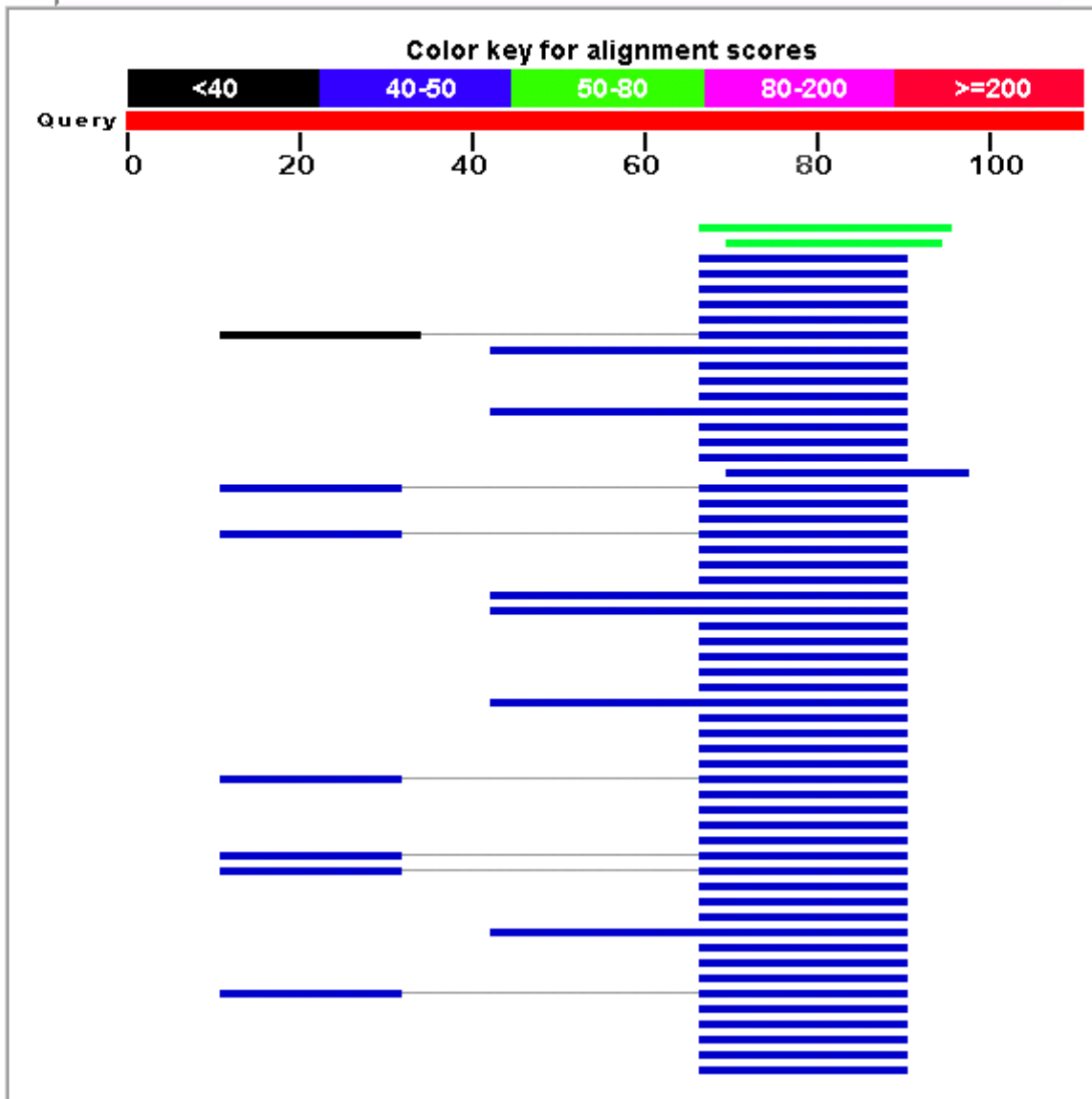
[Masking Character](#) [Masking Color](#)

Number of: [Descriptions](#) [Alignments](#) [Graphic overview](#)

[Alignment view](#)

Distribution of 239 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



Get selected sequences Select all Deselect all Tree View

> [gi|23462928|gb|AC122034.3](#) **D** Mus musculus BAC clone RP24-403M18 from chromosome 15, complete sequence
Length=182891

Score = 52.0 bits (26), Expect = 3e-04
Identities = 26/26 (100%), Gaps = 0/26 (0%)
Strand=Plus/Minus

Query 70 GATAGATAGATAGACAGATAGTGCTA 95
|||||
Sbjct 74017 GATAGATAGATAGACAGATAGTGCTA 73992

Score = 40.1 bits (20), Expect = 1.3
Identities = 23/24 (95%), Gaps = 0/24 (0%)
Strand=Plus/Minus

Query 67 GATGATAGATAGATAGACAGATAG 90
|||||
Sbjct 74052 GATGATAGATAGATAGATAGATAG 74029

Score = 38.2 bits (19), Expect = 5.0
Identities = 22/23 (95%), Gaps = 0/23 (0%)
Strand=Plus/Minus

Query 70 GATAGATAGATAGACAGATAGTG 92
|||||
Sbjct 75409 GATAGATAGATAGATAGATAGTG 75387

> [gi|25896278|emb|AL672240.13](#) **D** Mouse DNA sequence from clone RP23-27I16 on chromosome 11, complete sequence
Length=224603

Score = 50.1 bits (25), Expect = 0.001
Identities = 25/25 (100%), Gaps = 0/25 (0%)
Strand=Plus/Minus

Query 70 GATAGATAGATAGACAGATAGTGCT 94
|||||
Sbjct 69410 GATAGATAGATAGACAGATAGTGCT 69386

Tree view **NEW**

Sequences producing significant alignments:

	Score (Bits)	E Value
gi 23462928 gb AC122034.3 Mus musculus BAC clone RP24-403M18...	52.0	3e-04
gi 25896278 emb AL672240.13 Mouse DNA sequence from clone RP...	50.1	0.001
gi 56382171 gb AC101277.16 Mus musculus chromosome 1, clone RP2	48.1	0.005
gi 2911721 gb AC004225.1 Homo sapiens chromosome 5, Pac clon...	48.1	0.005
gi 77020723 gb AC163686.3 Mus musculus BAC clone RP23-3C10 f...	48.1	0.005
gi 50657543 gb AC144384.4 Pan troglodytes BAC clone CH251-45...	48.1	0.005
gi 51036414 gb AC138290.9 Mus musculus chromosome 10, clone ...	48.1	0.005
gi 50726841 gb AC113548.22 Mus musculus chromosome 9, clone RP2	48.1	0.005
gi 50058027 gb AC110744.6 Mus musculus chromosome 8, clone RP23	48.1	0.005
gi 76362957 gb AC158933.3 Mus musculus chromosome 1, clone RP23	48.1	0.005
gi 46359924 gb AC147883.2 Xenopus tropicalis clone CH216-117C13	48.1	0.005
gi 41386868 gb AC121097.6 Mus musculus chromosome 13, clone ...	48.1	0.005
gi 40254085 gb AC114998.13 Mus musculus chromosome 8, clone RP2	48.1	0.005
gi 76150334 emb AM085148.1 Poicephalus robustus robustus micros	48.1	0.005
gi 38304354 gb AC146275.2 Pan troglodytes BAC clone RP43-30A5 f	48.1	0.005
gi 28416265 gb AC126038.4 Mus musculus BAC clone RP24-230C18...	48.1	0.005
gi 25046672 gb AC124418.4 Mus musculus BAC clone RP24-318G17...	48.1	0.005
gi 28316667 gb AC122490.3 Mus musculus BAC clone RP24-388D8 ...	48.1	0.005
gi 24270642 gb AC090107.16 Homo sapiens 12 BAC RP11-643D8 (R...	48.1	0.005
gi 22857827 gb AC115632.4 Mus musculus BAC clone RP24-136O10...	48.1	0.005
gi 23499680 gb AC122850.4 Mus musculus BAC clone RP23-167J19 fr	48.1	0.005
gi 27452965 gb AC122840.3 Mus musculus BAC clone RP23-193M8 fro	48.1	0.005
gi 93140414 gb DP000044.2 Callithrix jacchus target 106 genomic	48.1	0.005
gi 37202222 gb AC108774.10 Mus musculus chromosome 1, clone RP2	48.1	0.005
gi 73745987 gb AC164704.4 Mus musculus BAC clone RP23-238D20...	48.1	0.005

TAREA!!!!

http://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.htm