

## ¿Por qué comparar secuencias...

### ... de proteínas?

- para conocer la función de las proteínas:
  - función general.
  - residuos importantes: p.e. centros activos.
- para predecir la estructura 3D de las proteínas.
- para determinar en qué especies está una proteína.
- ...

### ... de ADN?

- para buscar genes:
  - ESTs.
  - ADN genómico.
- para estudios de genética poblacional (SNPs).
- para comparar secuencias no codificantes.

## ¿Cuál es el objetivo de la comparación?

Durante la evolución las secuencias divergen por:

- mutaciones
- inserciones y deleciones
- barajado de dominios



El objetivo es encontrar el alineamiento que con mayor probabilidad (nunca sabremos si es el real) refleje qué cambios se han producido.

```
RPE_YEAST      6 IAPSIL---ASDFANLGCECHKVINAGADWLHIDVMDGHFVPNITLGQP      51
                ||.:|  ..|...|  .:..:|...:|.||||  |||.|.:...
RPE_MYCPN     10 IAFSLLPLLHQFDRKLL---EQFFADGLRLIHYDVMD-HFVDNTVFQGE      54
```

## -por pares

- alineamiento de dos secuencias
- búsqueda en bases de datos con BLAST.

## -muchas a la vez

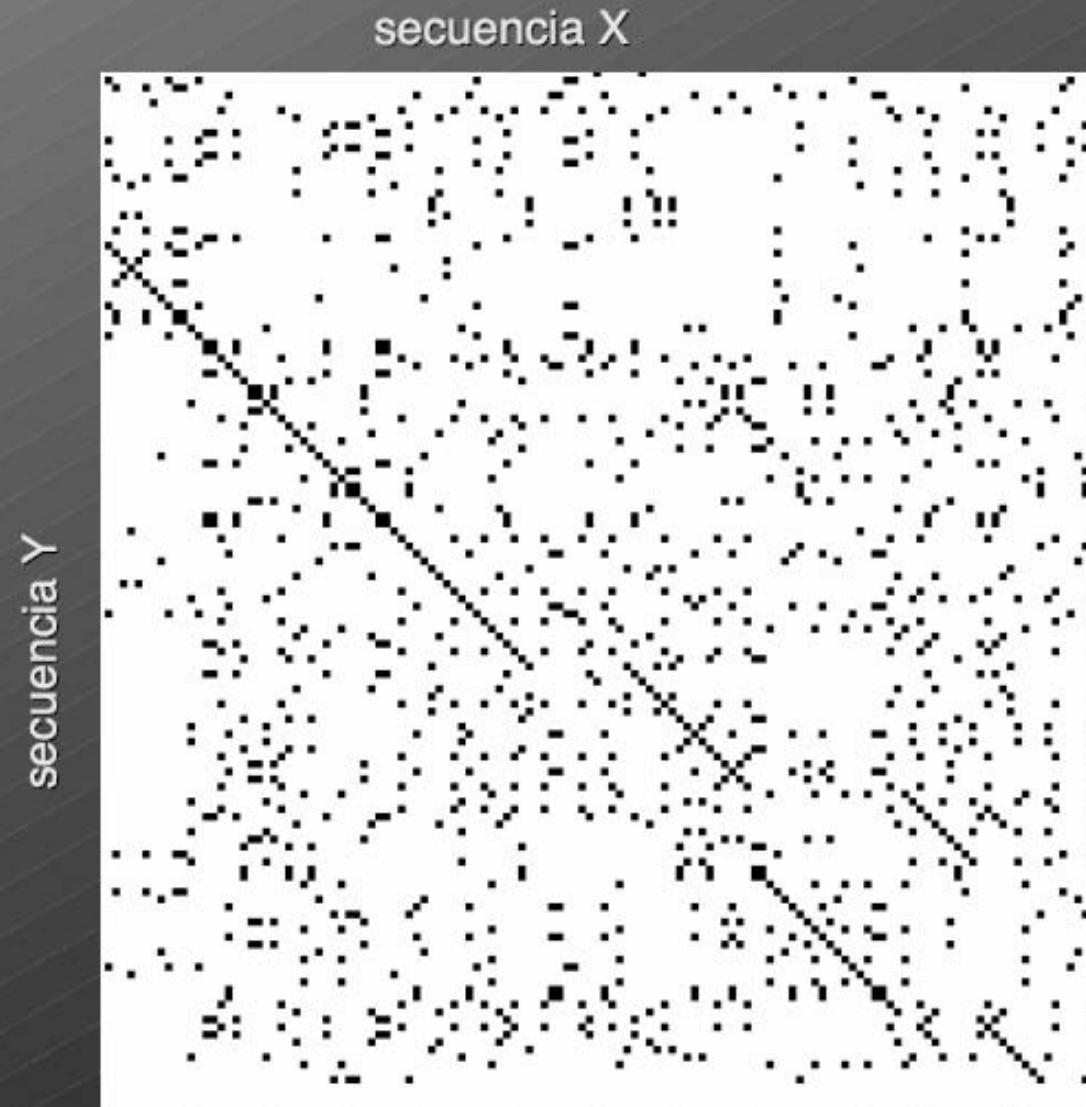
- alineamiento múltiple con Clustalw.

## -con patrones, perfiles y hmm's

- búsqueda en bases de datos con PSI-BLAST.
- bases de datos de interés:
  - PROSITE
  - PFam
  - InterPro

## Alineamiento de pares de secuencias

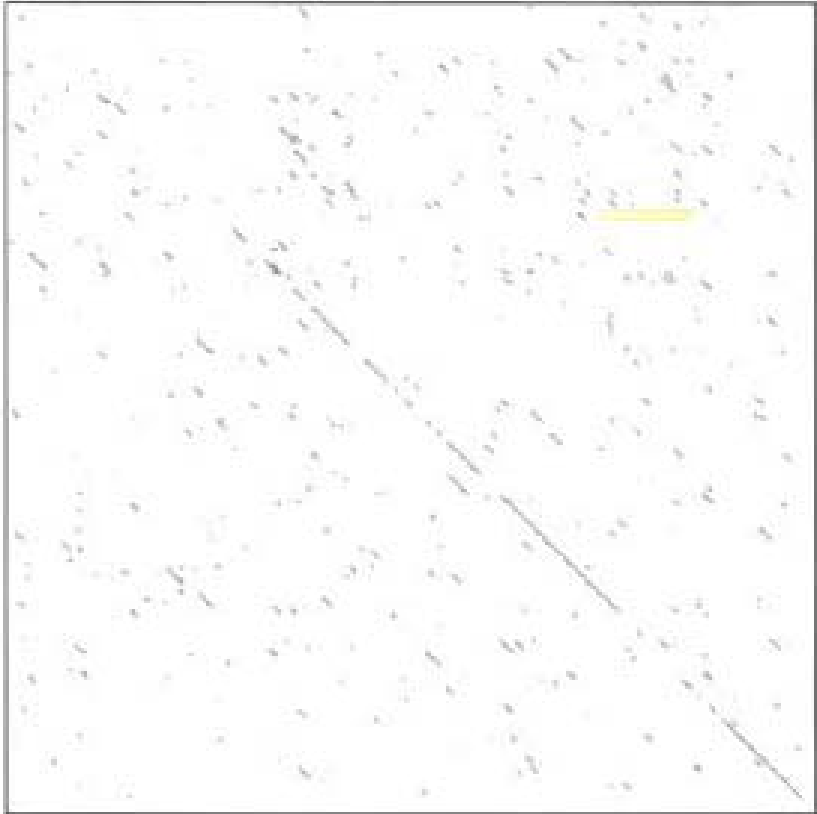
Comparación por identidades: matriz de puntos dot-plot



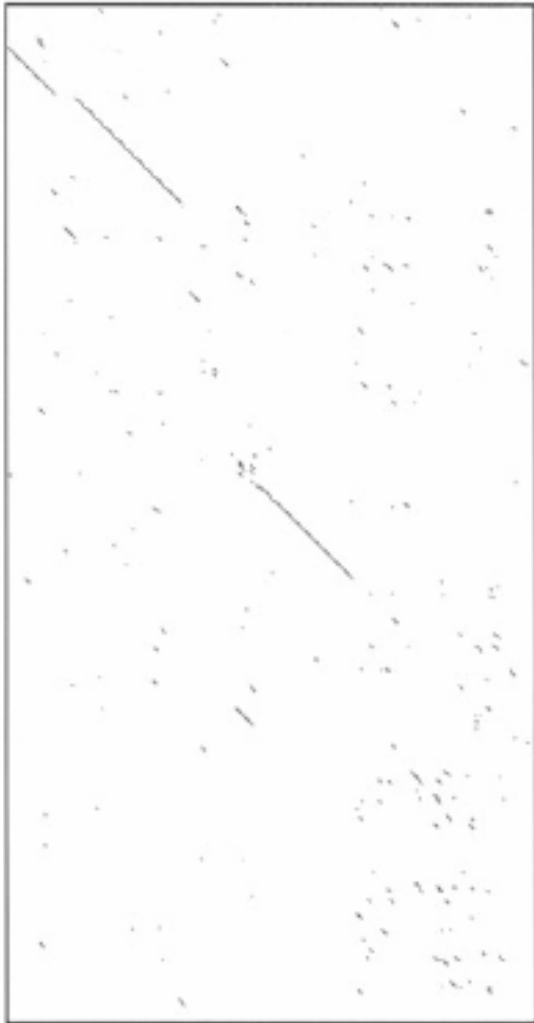


# COMPARACION DE DOS ATPASAS DE PECES

ATPasas lamprey / dogfish



mouse PAX-6 / Drosophila eyeless



PAPA\_CARPA / ACTN\_ACTCH

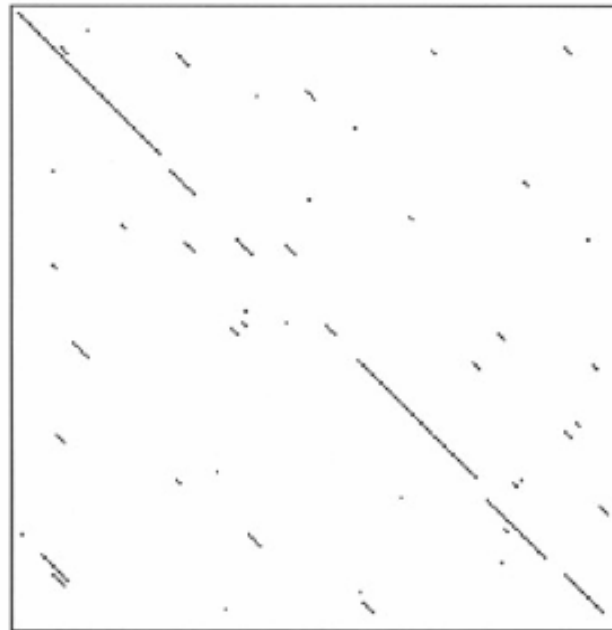


Figure 4.2a: Alignment of papaya papain and kiwi fruit actinidin with the corresponding dotplot.

El DOT PLOT permite una visualización rápida de la similitud entre dos secuencias

Inconvenientes:

No identifica directamente los fragmentos similares

No permite cuantificar el grado de similitud

Ventajas:

Nos muestra inversiones y estructuras repetidas

Computacionalmente eficiente

FIGURE 4.2d

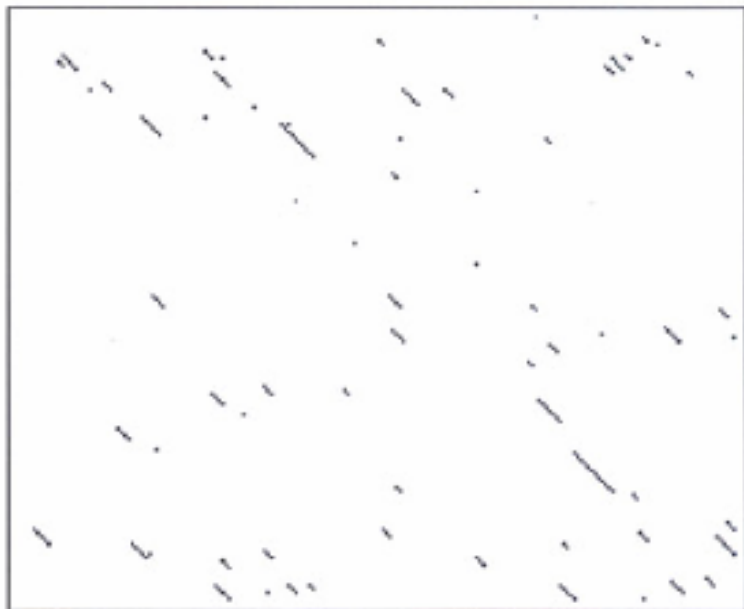


Figure 4.2d: Alignment of papaya papain and *S. aureus* staphopain, with the corresponding dotplot. The alignment is not derivable from this dotplot.

ALINEAMIENTO:

ALINEAR DOS SECUENCIAS CONSISTE EN IDENTIFICAR  
LAS CORRESPONDENCIAS RESIDUO-RESIDUO ENTRE  
AMBAS

EL ALINEAMIENTO ES LA HERRAMIENTA BÁSICA  
EN BIOINFORMATICA

Given two text strings:

- first string  
= a b c  
d e

a reasonable alignment would be

second string

= a c  
d e f

a b c  
d e -  
a - c  
d e f

An uninformative alignment

```
- - - - - - - g c t g a a c g  
c t a t a a t c - - - - - - -
```

An alignment without gaps

```
g c t g a a c g  
c t a t a a t c
```

An alignment with gaps

```
g c t g a - a - - c g  
- - c t - a t a a t c
```

And another

```
g c t g - a a - c g  
- c t a t a a t c -
```

necesitamos criterios para distinguir entre  
Buenos alineamientos y malos alineamiento

Distancia de **Hamming**: dadas dos cadenas de igual longitud, se define como el número de posiciones no coincidentes (mismatches)

Distancia se **Levenshtein**: dadas dos cadenas, no necesariamente de la misma longitud, se define como el número de operaciones de edición (inserción, deleción, alteración) necesarias para convertir una en otra

For example:

agtc  
cgta

Hamming  
distance =  
2

ag-tcc  
cgctca

Levenshtein  
distance =  
3

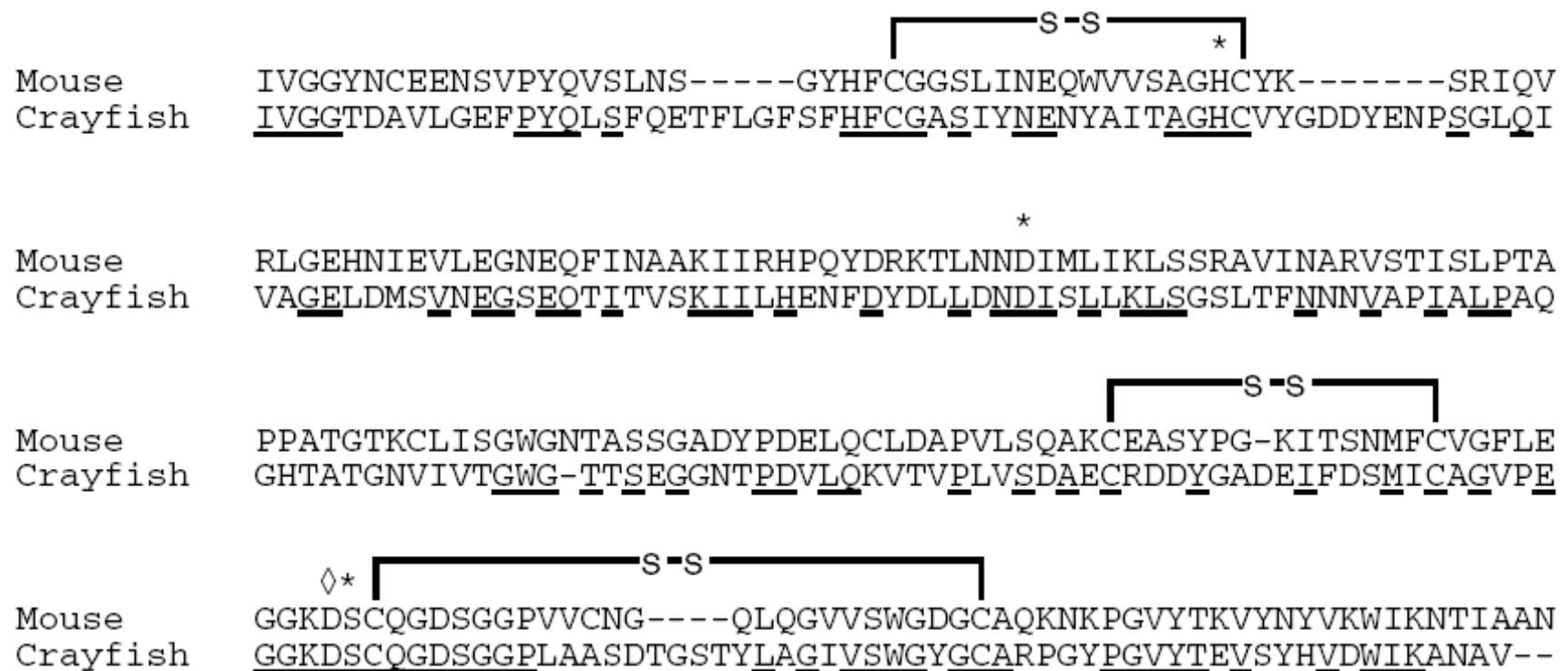
## Scores

Las distancias son medidas de la disimilaridad entre secuencias

El score es una medida de la similaridad entre secuencias

El score tiene introducir una penalización por introducir espacios (gap penalty)

La puntuación asignada a residuos coincidentes no tiene por qué ser igual en todos los casos, y dicha puntuación debería tener un sentido biológico



**Figure 8.1.** Conserved positions are often of functional importance. Alignment of trypsin proteins of mouse (SWISS-PROT P07146) and crayfish (SWISS-PROT P00765). Identical residues are underlined. Indicated above the alignments are three disulfide bonds (–S–S–), with participating cysteine residues conserved, amino acids side chains involved in the charge relay system (asterisk), and active site residue governing substrate specificity (diamond).

# Alineamiento de pares de secuencias

## Comparación por semejanza

**Observación:** hay aa's con propiedades físico-químicas similares:

- aa's ácidos: D, E.
- aa's básicos: K, R, H, ...
- aa's hidrofóbicos: L, I, W, ...
- aa's con estr. similar: Y -P, I -L, D -N, E -Q,...
- etc.

**Objetivo:** utilizar esa información para mejorar el alineamiento.

¿Cómo pasar del conocimiento general qué aa's se parecen a una estimación más precisa, cuantificada?

¿Qué sustituciones se toleran más en la Naturaleza?

**Matrices de sustitución (ejs: PAM, BLOSUM)**

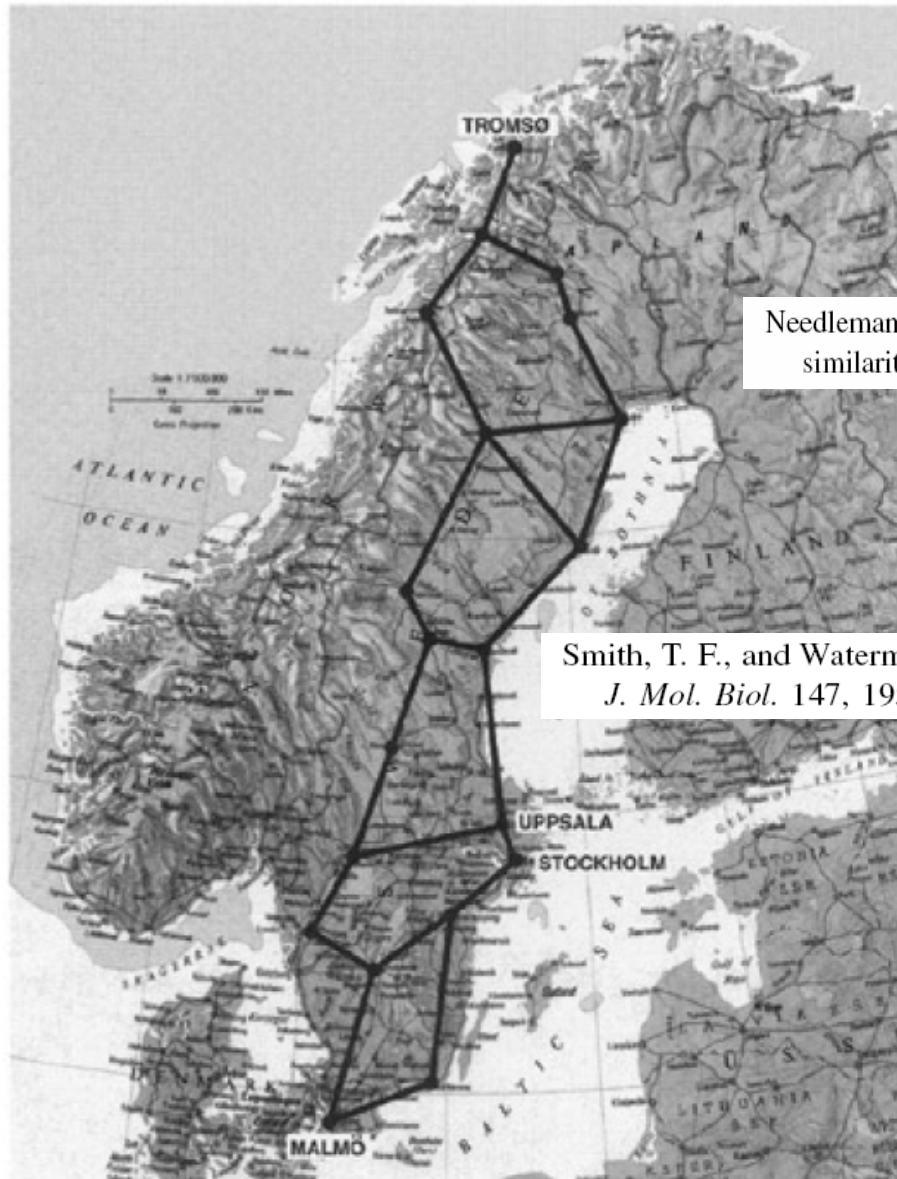
**Matrices de sustitución:** se construyen analizando miles de alineamientos.

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





¿Cuál es la mejor ruta entre Malmo y Tromso?



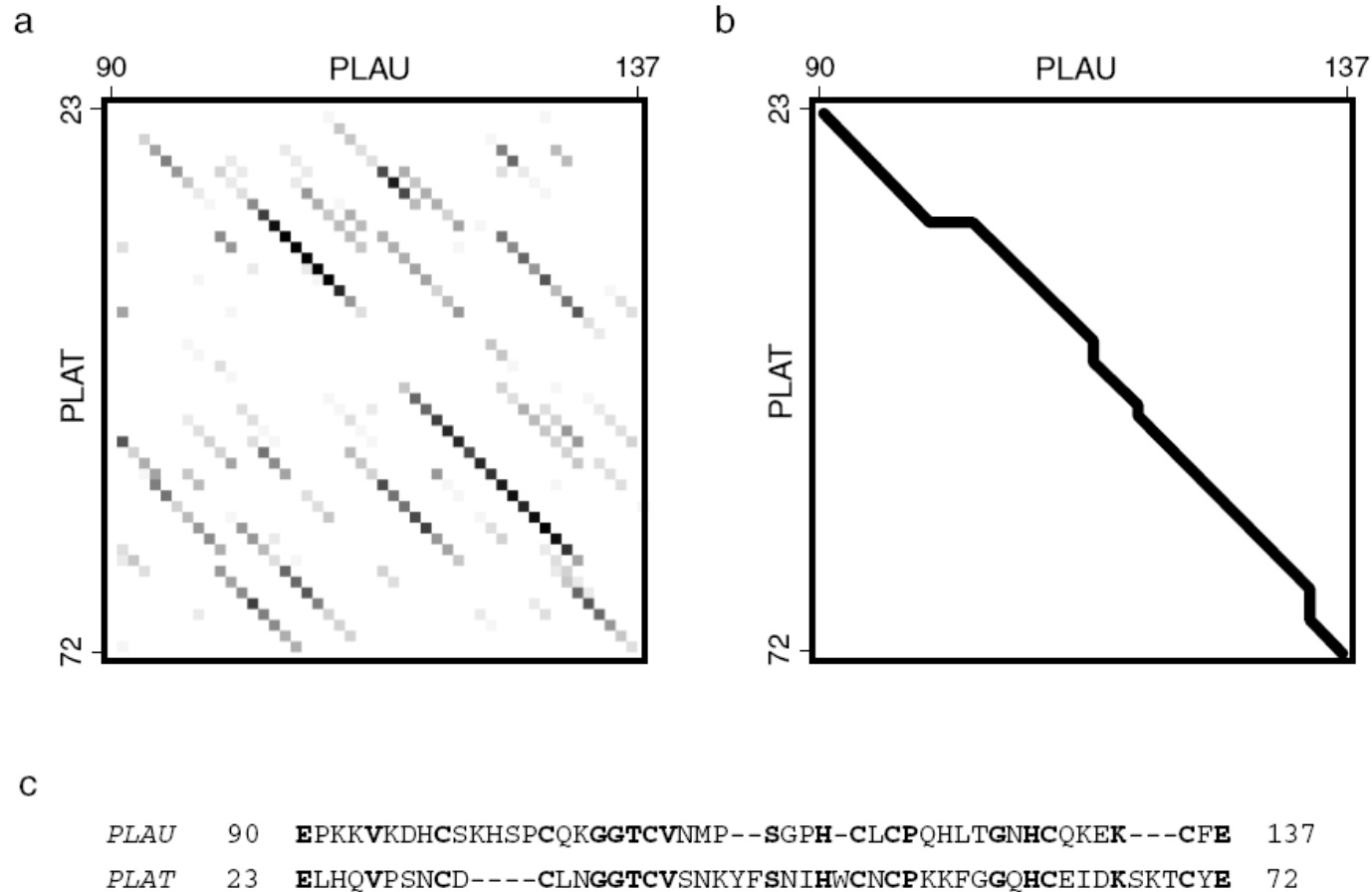
Algoritmos de programación dinámica

Needleman, S. B., and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

global

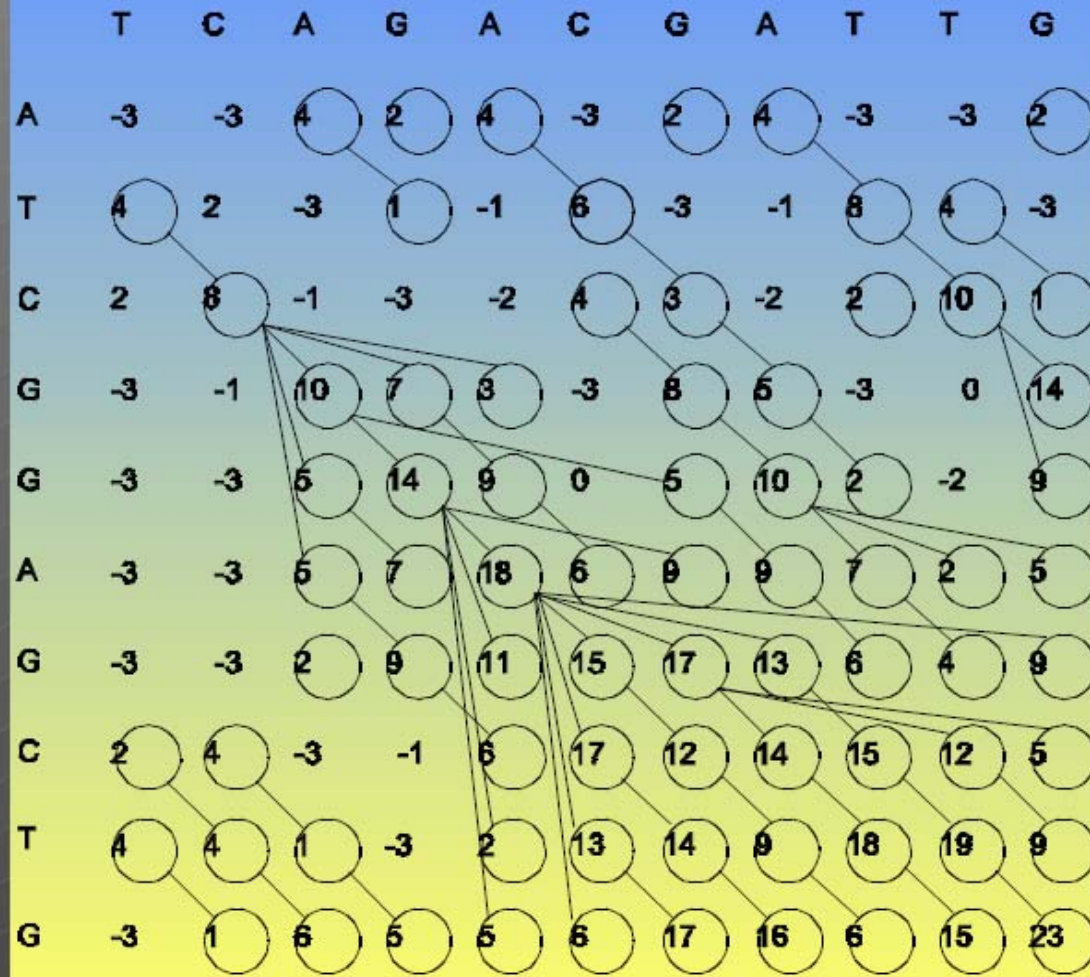
Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.

local



**Figure 8.5.** Dot-matrix, path graph, and alignment. All three views represent the alignment of the EGF similarity domains in the human urokinase plasminogen activator (PLAU; SWISS-PROT P00749) and tissue plasminogen activator (PLAT; SWISS-PROT P00750) proteins. (a) The entire proteins were compared with dotter and an enlargement of the small region corresponding to the EGF domain is shown here. (b) The path graph representation of the alignment found by BLASTP. (c) The BLASTpgp alignment represented in the familiar text form.

### Programacin Dinmica



Esquema de Pesos

[ 4] residuos iguales

[ 2] residuos del mismo tipo

[-3] Resto.

iGap: -5

eGap: -2

Mejor alineamiento:

**TCAGACGATTG**

||.|| . .||

**ATCGGA--GCTG**

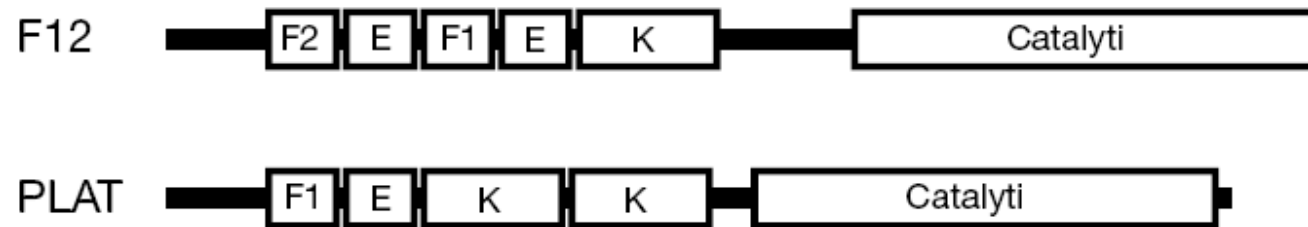
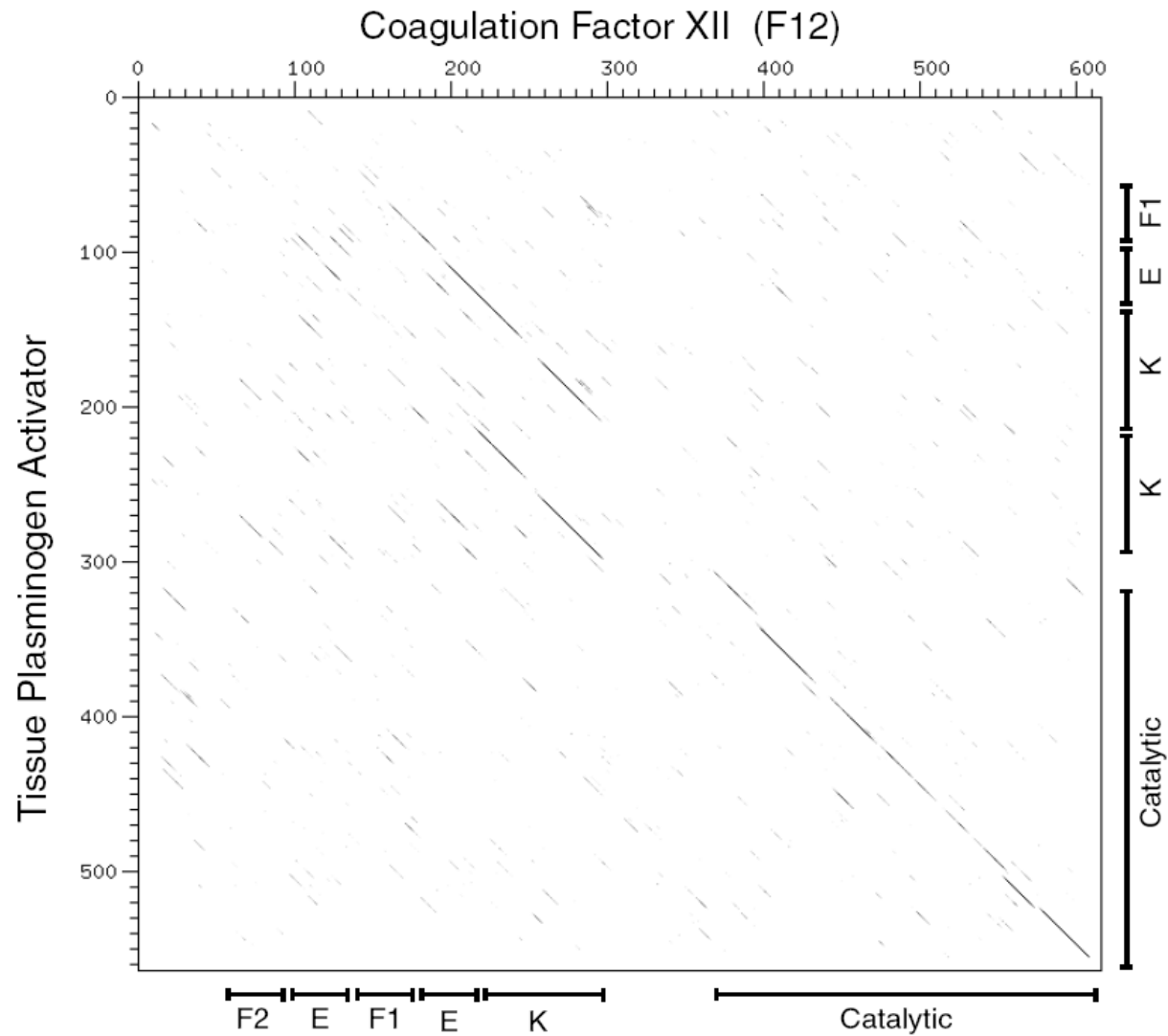


Figure 8.3. Modular structure of two proteins involved in blood clotting. Schematic representation of the modular structure of human tissue plasminogen activator and coagulation factor XII. A module labeled C is shared by several proteins involved in blood clotting. F1 and F2 are frequently repeated units that were first seen in fibronectin. E is a module resembling epidermal growth factor. A module known as a “kringle domain” is denoted K.



**Figure 8.4.** Dot matrix sequence comparison. Dot matrix comparison of the amino acid sequences of human coagulation factor XII (F12; SWISS-PROT P00748) and tissue plasminogen activator (PLAT; SWISS-PROT P00750). The figure was generated using the dotter program (Sonnhammer and Durban, 1996).

## Alineamiento global versus alineamiento local

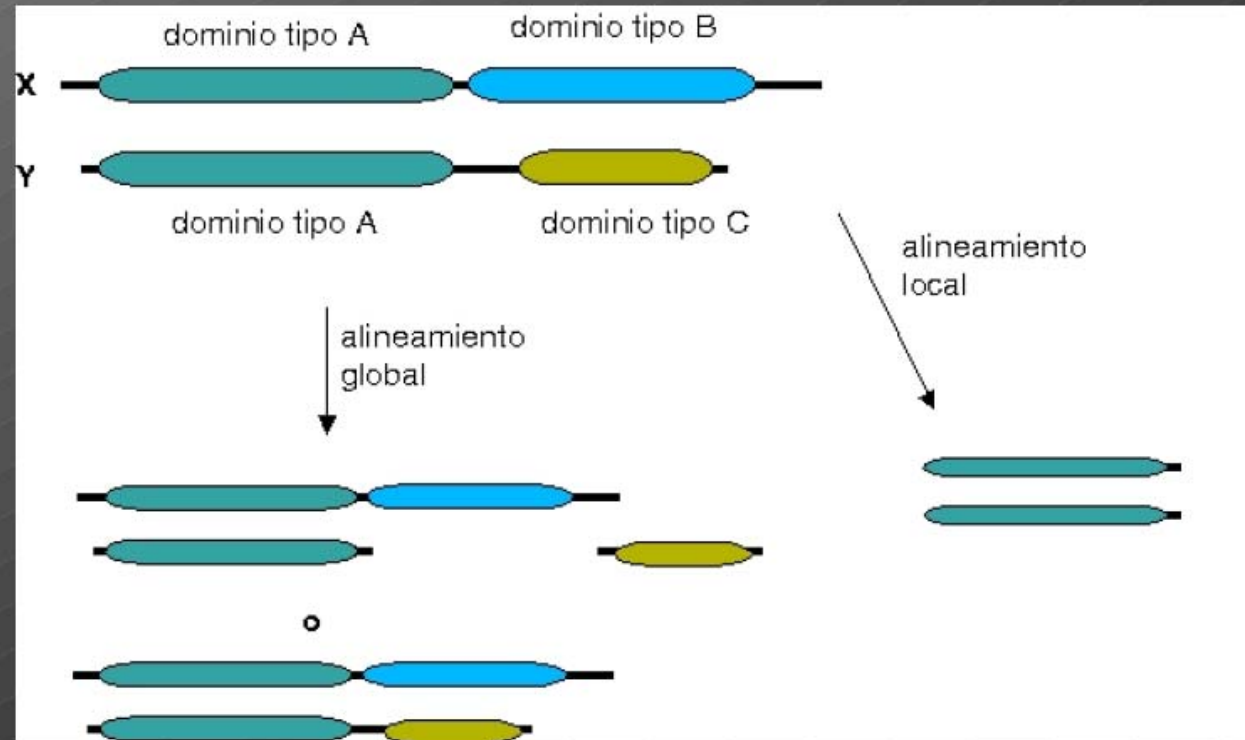


Trata de obtener el mejor alineamiento superponiendo las secuencias completas.

Sólo se debe utilizar cuando las proteínas son homólogas en toda su extensión (tienen los mismos dominios)



Halla aquéllos trozos de las secuencias que superpuestos resultan en una puntuación máxima.



## Comparación incluyendo INDELs (inserciones y deleciones)

**Observación:** además de sustituciones pueden ocurrir inserciones y deleciones.

**Objetivo:** utilizar esa información para mejorar el alineamiento.

**Problemas a resolver:**

- ¿Cómo penalizar los INDELs (los gaps)?

**Apertura y extensión de un gap.**

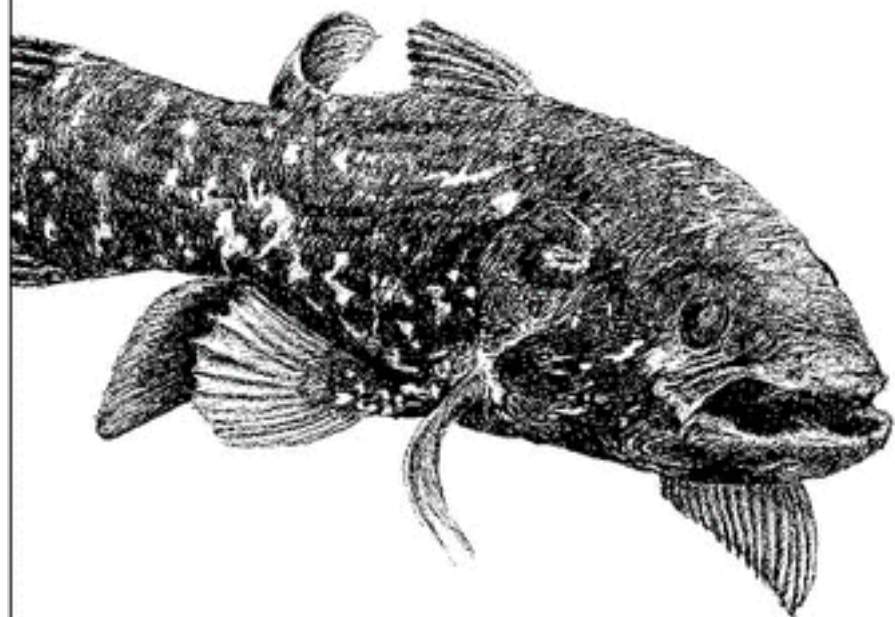
- Las formas de alinear dos secuencias incluyendo gaps son enormes  
=> problema computacional.

**Programación dinámica.**

(Needleman & Wunsch, Smith & Waterman)

*An Essential Guide to the Basic Local Alignment Search Tool*

# BLAST



**O'REILLY®**

*Ian Korf, Mark Yandell & Joseph Bedell*

# Búsqueda en bases de datos con BLAST

## Observaciones:

- Complejidad algorítmica de la programación dinámica:  $N \times M$   
(N y M son las longitudes de las dos secuencias a alinear)
- Conocemos la secuencia de 1,5 millones de proteínas y la de unos 22 millones de ADN (28.000 millones de pdb).

**Problema:** la programación dinámica es demasiado lenta para buscar homólogos en las bases de datos.

**Solución:** aplicar heurísticas (truquillos) para aumentar la velocidad:

- tablas de dispersión.
- k-tuplas.
- búsqueda en las diagonales más probables.


**Heurística:** truquillo que, aunque no garantiza la solución óptima, en la mayoría de los casos funciona.

NCBI BLAST - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Adelante Detener Actualizar Inicio Búsqueda Favoritos Historial Correo Imprimir Modificar Discutir

Dirección <http://www.ncbi.nlm.nih.gov/BLAST/> Ir a Vínculos



**PubMed** **Entrez** **BLAST** **GMM** **Taxonomy** **Structure**

**Info**

- FAQs
- News
- References
- Credits

**Education**

- Program selection guide
- Tutorial
- URL API guide

**Download**

- Executables
- Databases
- Source code

**Support**

- Helpdesk
- Mailing list

# BLAST

**NEW** 10 February 2004 BLAST 2.2.8 has been released. [Read more...](#)

<p><b>Nucleotide</b></p> <ul style="list-style-type: none"> <li>• Discontiguous megablast</li> <li>• Megablast</li> <li>• Nucleotide-nucleotide BLAST (blastn)</li> <li>• Search for short, nearly exact matches</li> <li>• Search trace archives with megablast or discontiguous megablast</li> </ul> <p><b>Translated</b></p> <ul style="list-style-type: none"> <li>• Translated query vs. protein database (blastx)</li> <li>• Protein query vs. translated database (tblastn)</li> <li>• Translated query vs. translated database (tblastx)</li> </ul> <p><b>Special</b></p> <ul style="list-style-type: none"> <li>• Align two sequences (bl2seq)</li> <li>• Screen for vector contamination (VecScreen)</li> <li>• Immunoglobulin BLAST (IgBlast)</li> </ul>	<p><b>Protein</b></p> <ul style="list-style-type: none"> <li>• Protein-protein BLAST (blastp)</li> <li>• PHI- and PSI-BLAST</li> <li>• Search for short, nearly exact matches</li> <li>• Search the conserved domain database (rpsblast)</li> <li>• Search by domain architecture (cdart)</li> </ul> <p><b>Genomes</b></p> <ul style="list-style-type: none"> <li>• Environmental samples <b>NEW</b></li> <li>• Human, mouse, rat</li> <li>• Fugu rubripes, zebrafish</li> <li>• Insects, nematodes, plants, fungi, malaria</li> <li>• Microbial genomes, other eukaryotic genomes</li> </ul> <p><b>Meta</b></p> <ul style="list-style-type: none"> <li>• Retrieve results by RID</li> <li>• Get this page with javascript-free links</li> </ul>
---	--

[Disclaimer](#)  
[Privacy statement](#)

Internet

Inicio C:\USERS\fedex\DOCT... fabascal@gredos.cnb... WS\_FTP LE.urales.cnb... NCBI BLAST - Micr... 16:48

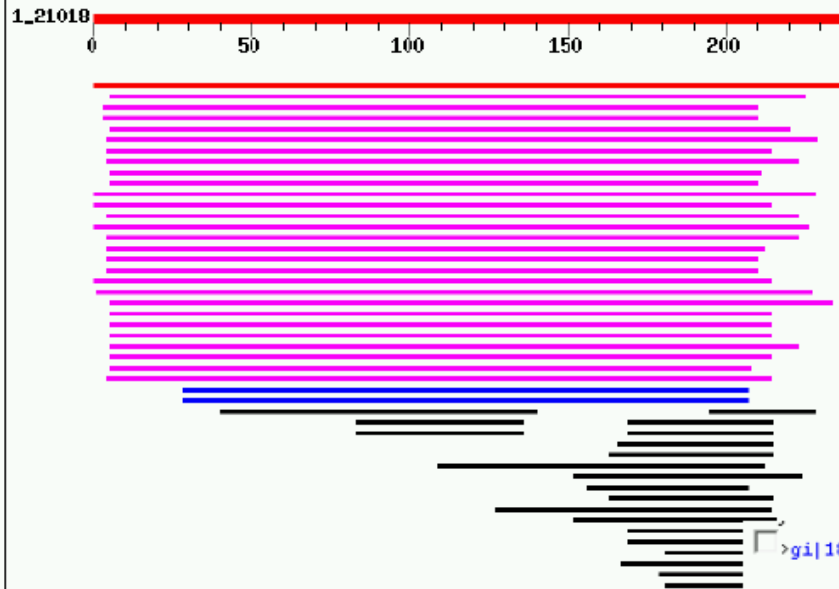
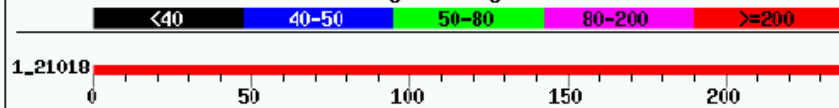
TABLE 8.3. Nucleotide Sequence Databases for use with BLAST

Database	Description
<i>nr</i>	Nonredundant GenBank, excluding the EST, STS, and GSS divisions
<i>month</i>	Subset of <i>nr</i> , which is new or modified within the last 30 days
<i>est</i>	GenBank EST division (expressed sequence tags)
<i>sts</i>	GenBank STS division (sequence tagged sites)
<i>htgs</i>	GenBank HTG division (high-throughput genomic sequences)
<i>gss</i>	GenBank GSS division (genome survey sequences)
<i>ecoli</i>	Complete genomic sequence of <i>E. coli</i>
<i>yeast</i>	Complete genomic sequence of <i>S. cerevisiae</i>
<i>drosoph</i>	Complete genomic sequence of <i>D. melanogaster</i>
<i>mito</i>	Complete genomic sequences of vertebrate mitochondria
<i>alu</i>	Collection of primate Alu repeat sequences
<i>vector</i>	Collection of popular cloning vectors

Distribution of 50 Blast Hits on the Query Sequence

Mouse-over to show define and scores. Click to show alignments

Color Key for Alignment Scores



Sequences producing significant alignments:

Accession	Organism	Score (bits)	E Value
gi 1173139 sp P46969 RPE_YEAS	RIBULOSE-PHOSPHATE 3-EPIMERA...	478	e-135
gi 18203210 sp Q9L0Z5 RPE_STRC	Ribulose-phosphate 3-epimer...	194	1e-49
gi 18269961 sp P71676 RPE_MYCT	Ribulose-phosphate 3-epimer...	184	1e-46
gi 17380276 sp Q9CCP9 RPE_MYCL	Ribulose-phosphate 3-epimer...	180	2e-45
gi 6647758 sp Q34557 RPE_BACSV	Ribulose-phosphate 3-epimera...	170	3e-42
gi 2829613 sp P74061 RPE_SVNY3	Ribulose-phosphate 3-epimera...	162	7e-40
gi 6226024 sp Q67098 RPE_AQUR	Ribulose-phosphate 3-epimera...	158	1e-38
gi 2499728 sp Q43843 RPE_SOLTU	Ribulose-phosphate 3-epimera...	155	8e-38
gi 6647759 sp Q66107 RPE_TREPA	Ribulose-phosphate 3-epimera...	155	9e-38
gi 6094119 sp P51012 RPE_RHOCA	Ribulose-phosphate 3-epimera...	154	2e-37
gi 1169387 sp P32661 RPE_ECOLI	Ribulose-phosphate 3-epimera...	152	5e-37
gi 25091192 sp Q8K940 RPE_BUCAP	Ribulose-phosphate 3-epimer...	151	9e-37

>gi|18203210|sp|Q9L0Z5|RPE\_STRC Ribulose-phosphate 3-epimerase (Pentose-5-phosphate (PPE) (R5P3E))  
Length = 228

Score = 194 bits (492), Expect = 1e-49  
Identities = 101/225 (44%), Positives = 142/225 (63%), Gaps = 19/225 (8%)

Query: 6 IAP SILASDFANLGCCECHKVINAGADWLHIDVMDGHFVFNITLGGPIVTSLRSVPRPGD 65  
I PSIL++DFA L E V GADWLH+DVMD HFVFN+TLG P+V SL R+ P  
Sbjct: 5 INP SILSADFARLADERAKV--EGADWLHVDVMDNHGFVFNITLGGVVFVESLARATDTP-- 60

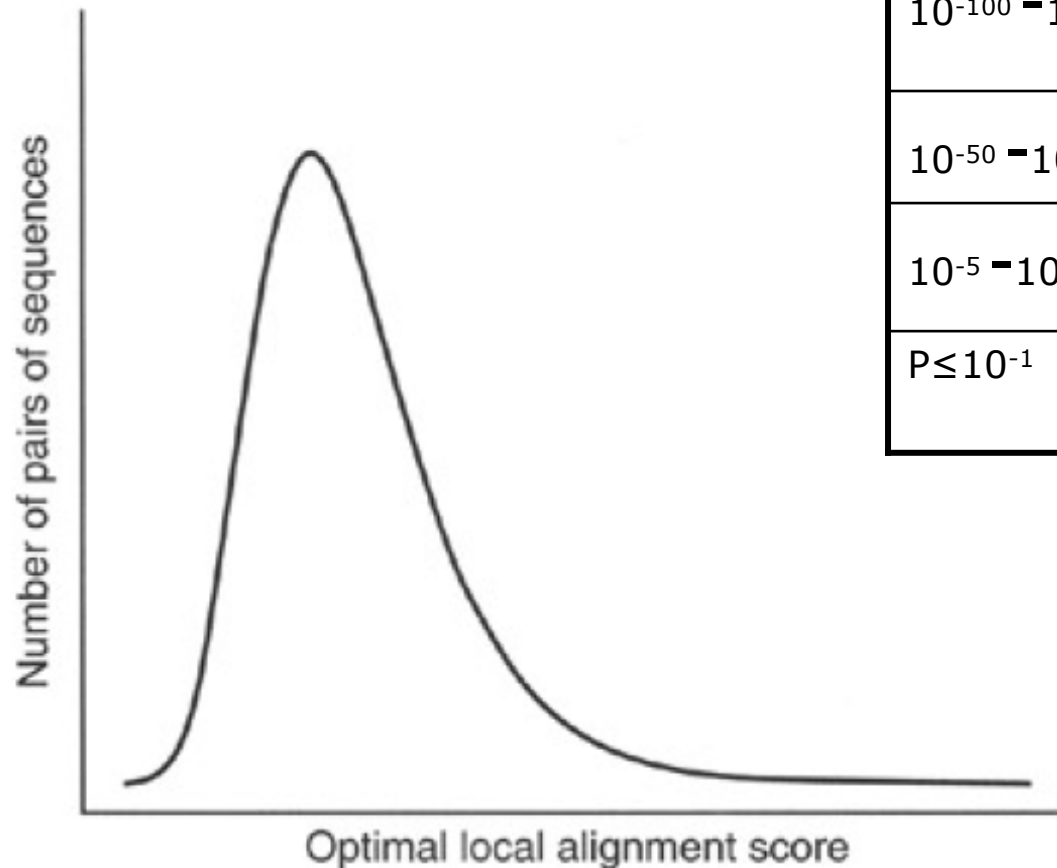
Query: 66 ASNTEKKPTAFFDCMMHVENPEKVVDDFAKCGADQFTFMYEATQDPLHLVKLKIKSGKGIKA 125  
DCM+M+E P++W ++ GA TFM EA P+ L+ I++KG +A  
Sbjct: 61 -----LDCHLMEAPDRWAPQVVERGAGSVTFMREARRAPVRLAREIRAKGARA 109

Query: 126 ACAIKP6T SVDVLFELAPHLDMALVMTVEPGF6GQKFMEDMMPKV----ETLRAKFFPHLN 181  
+ A+KP T V+ +L P LDM L+MTVEPGF6GQ F++ M+PK+ E ++ L  
Sbjct: 110 SMALKPATPVEPYEDLLPELDMLLMTVEPGF6GQRFLLDMLPKIRRTRELKIKKNGLELW 169

Query: 182 IQVDGGGLGKETIPKAKAGANVIVAGT SVFTARDPMDVIFSMKKE 226  
+QVDGG+ TI + A AGA+V VAG++V+ A+DP + + ++ +  
Sbjct: 170 LQVDGGVSAATIERCADGADVFVAGSAVYGRSDPRAEVRALRTQ 214

Para un score  $x$ , la probabilidad de observar un score  $\geq x$  es:

$$P(\text{score} \geq x) = 1 - \exp(-ke^{-\lambda x})$$



$P \leq 10^{-100}$	exacto
$10^{-100} - 10^{-50}$	casi idénticas
$10^{-50} - 10^{-10}$	relacionadas
$10^{-5} - 10^{-1}$	Relación distante
$P \leq 10^{-1}$	Probablemente no relacionadas

Distribución de valor extremo

Z-score = (score-media)/dev estandar

Z=0	Similitud observada equivalente a la aleatoria
Z≥5	Probablemente significativo

E-value de un alineamiento encontrado en una base de datos, es el número esperado de secuencias que por azar dar un score igual o mayor al obtenido. Resulta de multiplicar P por el tamaño de la base de datos

$E \leq 0.02$	Probablemente homólogas
$0.02 < E < 1$	No descartar homología
$E > 1$	Indistinguible del azar

## E-value: algunos consejos prácticos

- Con bases de datos grandes....

Si e-value  $< 1e-05$ : muy-muy fiable

Si  $1e-05 < \text{e-value} < 0.1$ : casi siempre son homólogos

Si e-value  $> 0.1$ : más arriesgado.

- Lo mejor: el propio criterio.
- La prueba definitiva de la homología: el alineamiento múltiple, buscar con métodos más sofisticados (p.e. PSI-BLAST), la estructura de las proteínas, etc.
- En cuanto a los **filtros**, lo mejor es probar con y sin filtrado y determinar si en el caso concreto resultan útiles.

## Interpretación del nivel de similitud entre dos proteínas

<45%	Muy probablemente Idéntica función
45-25%	Alta probabilidad de Estructura y función similar
18-25%	Zona difusa
<18%	Indistinguible del azar