

# Comparación de secuencias de nucleótidos y proteínas.

Dr. Alexis Salas Burgos

V 1.1

# Objetivos

- ¿Por qué tiene sentido comparar secuencias?
- ¿Cómo puedo comparar dos secuencias?
- ¿Cómo puedo alinear dos secuencias?
- ¿Cómo yo puedo buscar en una base de datos secuencias similares?

¿Por qué tiene sentido comparar secuencias?

EVOLUCIÓN

# ¿Por qué tiene sentido comparar secuencias?

```
trigo  --DPNKPGRAMTSFVFFMSEFRSEFKQKHSKLKSIVEMVKAAGER
      | | | | | | | | | | | | | | | | | | | | | |
????? KKDSNAPKRAMTSFMFFSSDFRS----KHSDL-SIVEMSKAAGAA
```

# Extrapolar

# Homología?

# SwissProt

??????

[illegible]

# ¿Por qué tiene sentido comparar secuencias?

[ExPASy Home page](#) [Site Map](#) [Search ExPASy](#) [Contact us](#) [SWISS-PROT](#)

Mirror sites: [Australia](#) [Canada](#) [China](#) [Taiwan](#)

## NiceProt View of SWISS-PROT: [P40623](#)

[\[General\]](#) [\[Name and origin\]](#) [\[References\]](#) [\[Comments\]](#) [\[Cross-references\]](#) [\[Keywords\]](#) [\[Features\]](#) [\[Sequence\]](#) [\[Tools\]](#)

**General information about the entry**

Entry name	HMGB1_CHITE
Primary accession number	<b>P40623</b>
Secondary accession number(s)	None
Entered in SWISS-PROT in	Release 31, February 1995
Sequence was last modified in	Release 31, February 1995
Annotations were last modified in	Release 32, November 1995

**Name and origin of the protein**

Protein name	MOBILITY GROUP PROTEIN 1B
Synonym(s)	None
Gene name(s)	HMGB1
From	<a href="#">Chironomus tentans</a> (Midge)
Taxonomy	Eukaryota, Metazoa, Arthropoda, Tracheata, Hexapoda, Insecta, Pterygota, Neoptera, Endopterygota, Diptera, Nematulimorpha, Chironomidae, Chironominae, Chironomus

**References**

[1]  
SEQUENCE FROM N. A.  
TISSUE=EMBRYONIC EPITHELIUM,  
MEDLINE: 9238101 [NCBI] [ExPASy](#), [Israel](#), [Japan](#)  
[Watanabe J, R. Scholze E:](#)  
"Insect proteins homologous to mammalian high mobility group protein 1. Characterization and DNA-binding properties.",  
[J. Biol. Chem.](#) 267:17170-17177(1992).

**Comments**

- **FUNCTION:** FOUND IN CONDENSED CHROMOMERES. BINDS PREFERENTIALLY TO AT-RICH DNA.
- **SUBCELLULAR LOCATION:** NUCLEAR.
- **SIMILARITY:** BELONGS TO THE HMGB/HMGB2 PROTEIN FAMILY.
- **SIMILARITY:** CONTAINS 1 HMGB BOX.

**Copyright**

This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.ebi.ac.uk/venues/> or send an email to [license@ebi.ac.uk](mailto:license@ebi.ac.uk)).

**Cross-references**


EMBL	M93254; AAA21713.1; -. [ <a href="#">EMBL</a> ] / [ <a href="#">GenBank</a> ] / [ <a href="#">DDBJ</a> ] [ <a href="#">GoDnaSequence</a> ]
HSFP	<a href="#">Q85783</a> , 1HMA. [ <a href="#">HSFP ENTRY</a> ] / [ <a href="#">SWISS-3DIMAGE</a> ] / [ <a href="#">EDB</a> ]
PFAM	<a href="#">PF00305</a> , HMGB_box, 1.
PRODOM	[ <a href="#">Domain structure</a> ] / [ <a href="#">List of seq. sharing at least 1 domain</a> ]
BLOCKS	<a href="#">P40623</a> .
DCMO	<a href="#">P40623</a> .
PROTOMAP	<a href="#">P40623</a> .
PREPAGE	<a href="#">P40623</a> .
DIP	<a href="#">P40623</a> .
SWISS-2DPAGE	<a href="#">GET REGION ON 2D PAGE</a> .

**Keywords**

Nuclear protein, Chromosomal protein, DNA-binding.

**Features**

DNA_BIND	5 - 71	HMGB_BOX	
DOMAIN	104 - 110	ASP/GLU-RICH (ACIDIC) .	

 [VIEWER logo](#) [FT table viewer](#)

**Sequence information**

Length: 110 AA	Molecular weight: 12150 Da	CRC64: B3491735713333C4 [This is a checksum on the sequence]
----------------	----------------------------	--

10	20	30	40	50	60
MAKRPKPLS	AYNLVLSAR	ESIRKRNDF	KUTTVAKKG	ELVRLKDS	EWEAKAATK
70	80	90	100	110	
QNVIRALQEV	ERNGGGDDK	GEKRGGAAPK	EGAGKSKKG	AKSDDGDSK	

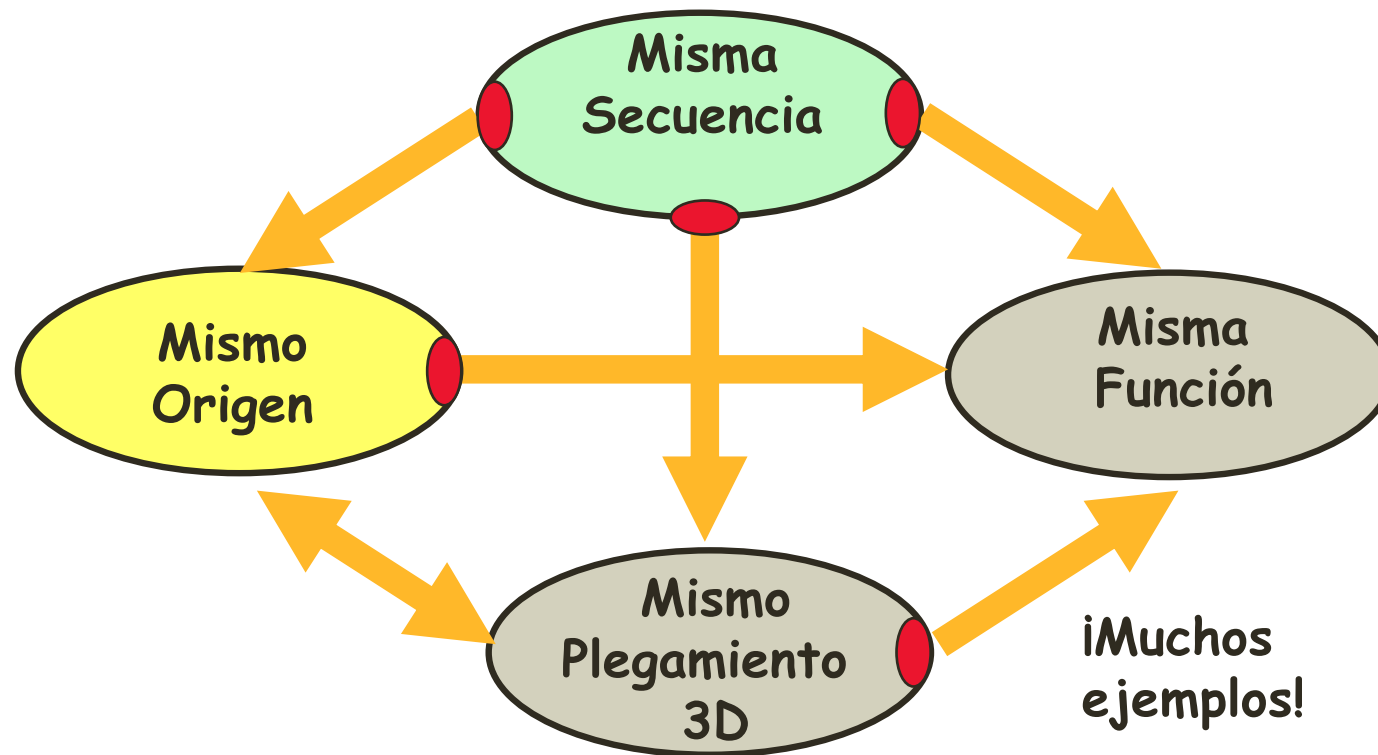
P40623 in [FASTA format](#)

# ¿Por qué tiene sentido comparar secuencias?

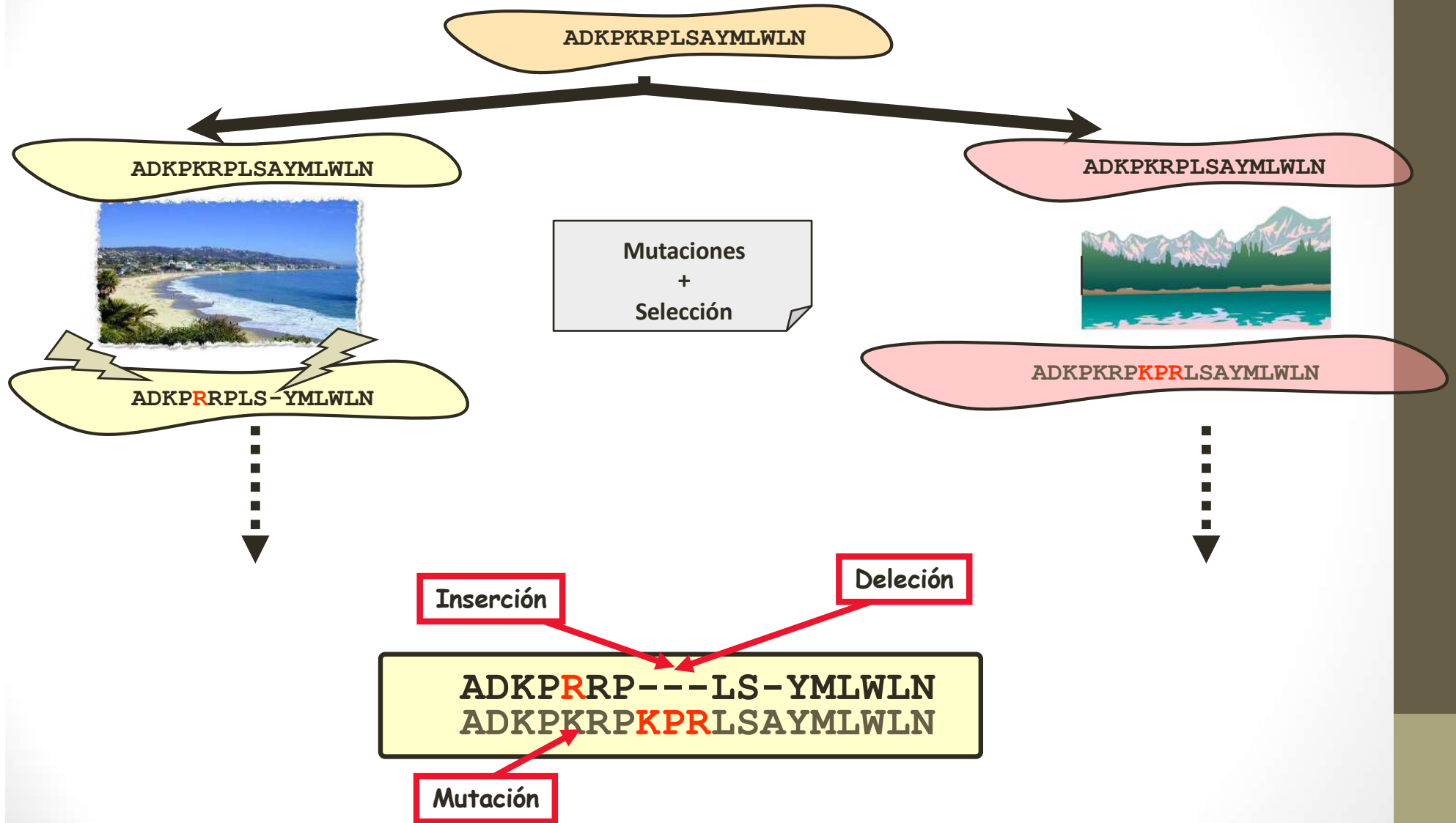
- Evolución es una herramienta real.
- La naturaleza va reutilizando secuencias.
- La mayoría de las veces es divergente.



# ¿Por qué tiene sentido comparar secuencias?



# Un alineamiento es una historia





# Evolución no siempre es divergente...

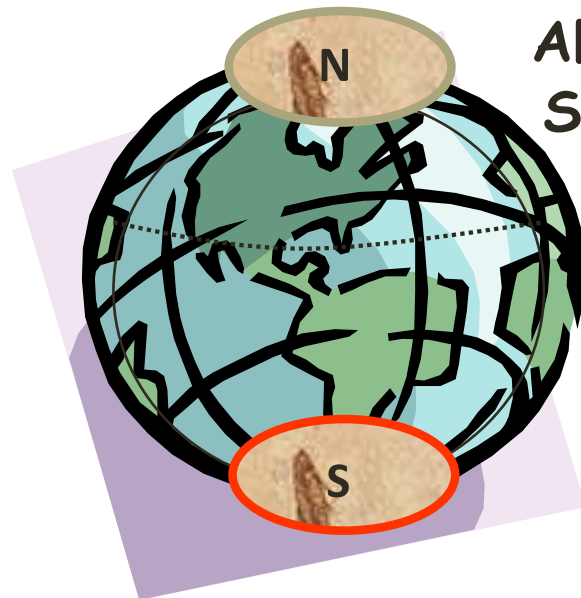
*Proc. Natl. Acad. Sci. USA*  
Vol. 94, pp. 3811–3816, April 1997  
Evolution

## Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish

(repetitive sequences/gene duplication/environmental selection/*de novo* amplification)

LIANGBIAO CHEN, ARTHUR L. DeVRIES, AND CHI-HING C. CHENG\*

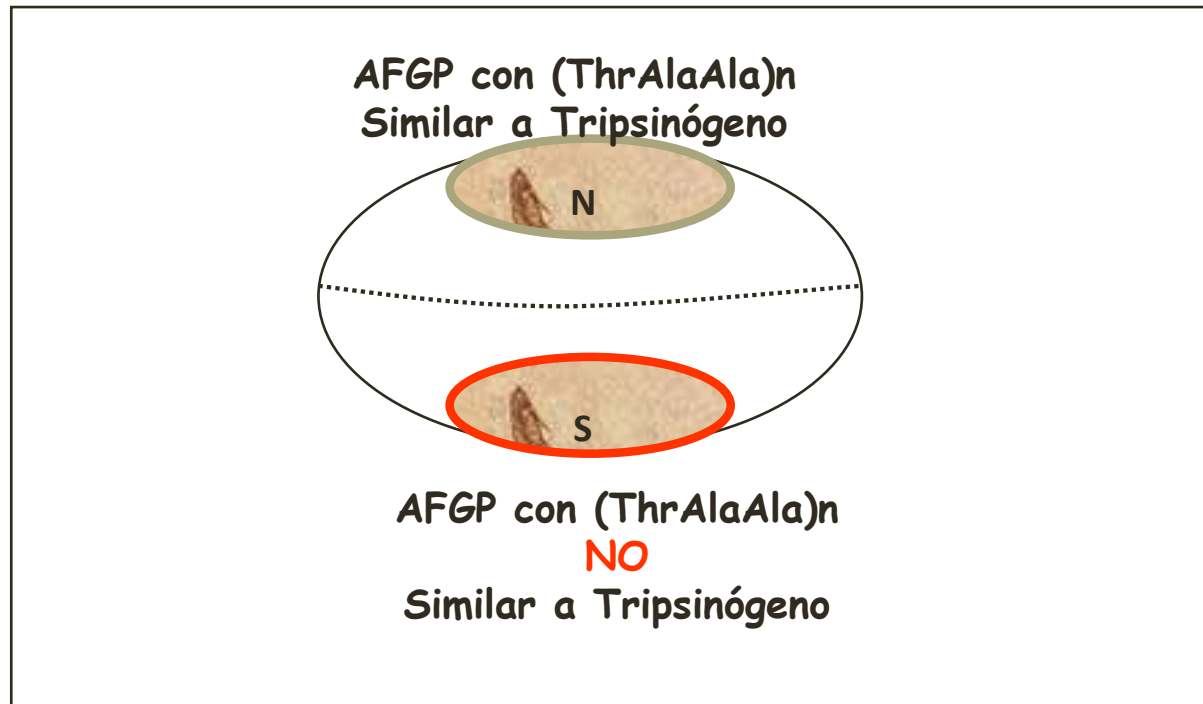
Department of Molecular and Integrative Physiology, University of Illinois, Urbana, IL 61801



AFGP con (ThrAlaAla)<sub>n</sub>  
Similar a Tripsinógeno

AFGP con (ThrAlaAla)<sub>n</sub>  
**NO** Similar a  
Tripsinógeno

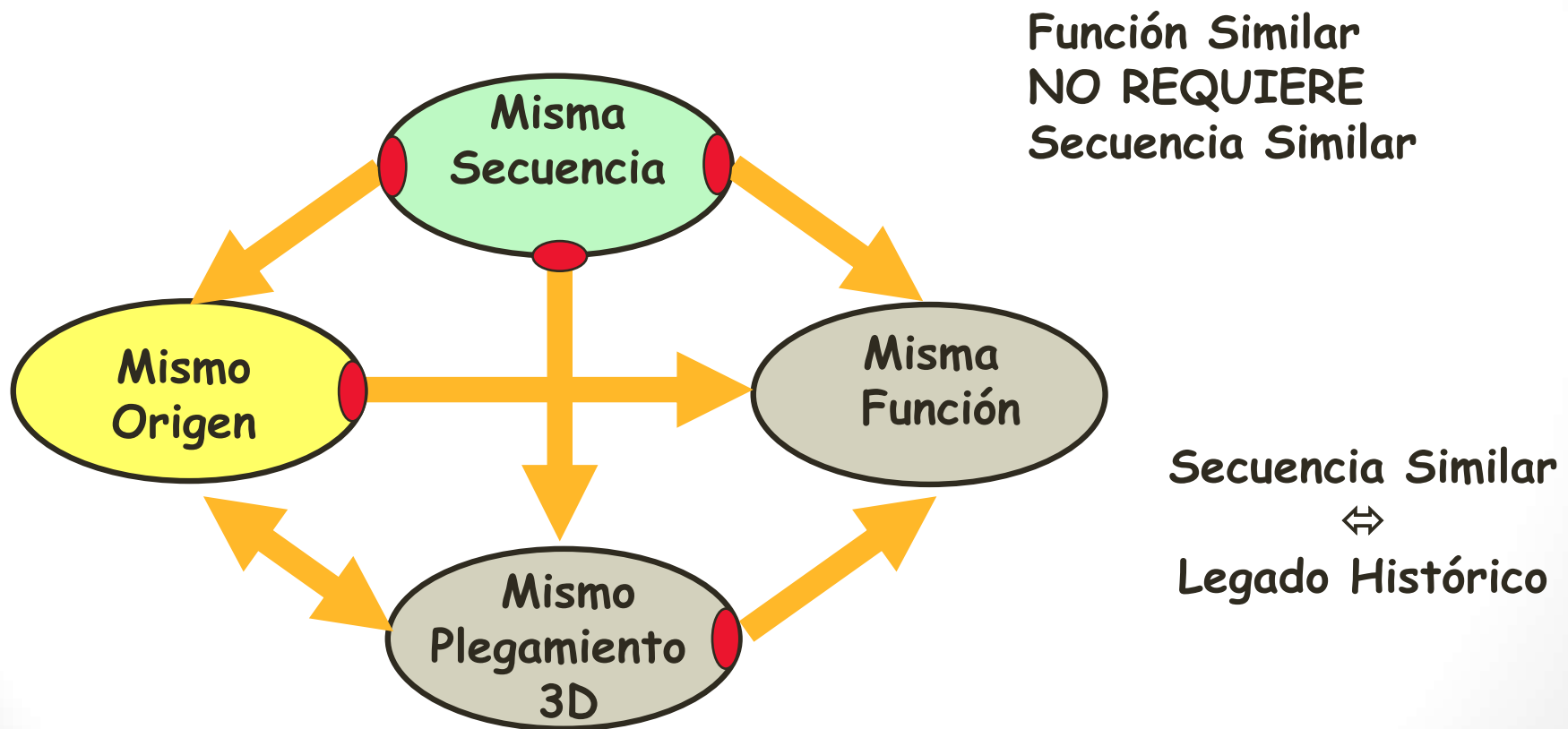
# Evolución no siempre es divergente...



Secuencias **SIMILARES**  
**PERO**  
Orígenes **DIFERENTES** origin

# Evolución no siempre es divergente...

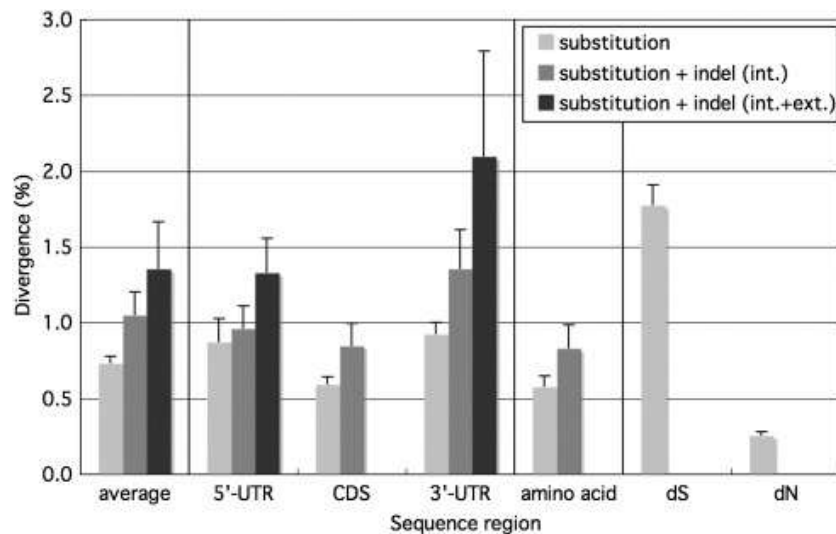
Pero en la **Mayoría** de los casos, tú puedes asumir esto:



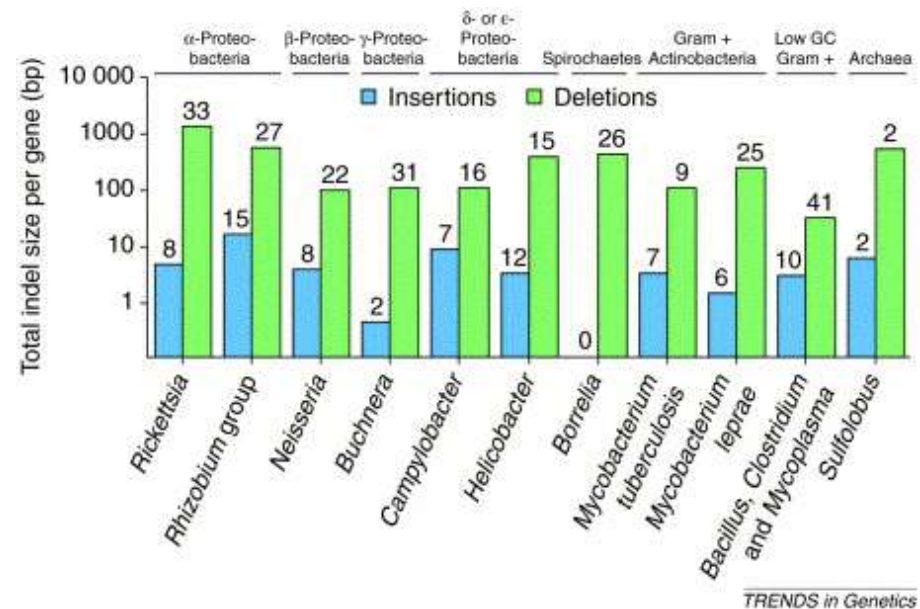
# ¿Cómo evolucionan las secuencias?

Cada porción del genoma posee su propia Agenda.  
La presión del medio ambiente influye.

## Región del Gen



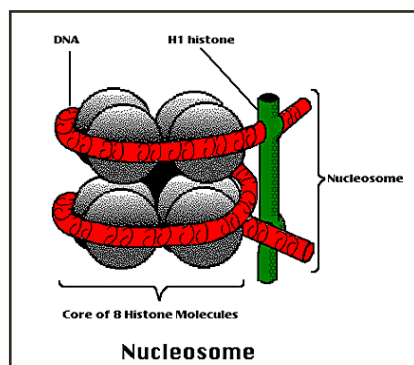
## Especies Bacteriales



# ¿Cómo evolucionan las secuencias?

**Restringido** Posiciones del Genoma Evolucionan Lentamente

**CADA** Familia de Proteínas Posee su Propio nivel de Restricción.

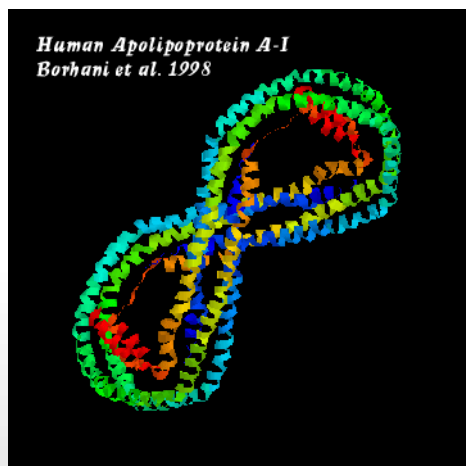


Familia	$K_S$	$K_A$
Histona3	6.4	0
Insulina	4.0	0.1
Interleuquina I	4.6	1.4
$\alpha$ -Globina	5.1	0.6
Apolipoprot. AI	4.5	1.6
Interferón G	8.6	2.8

Velocidades en Sustituciones/sitio/Billón de Años  
Referencia Mouse Vs Human (80 Millones Años)

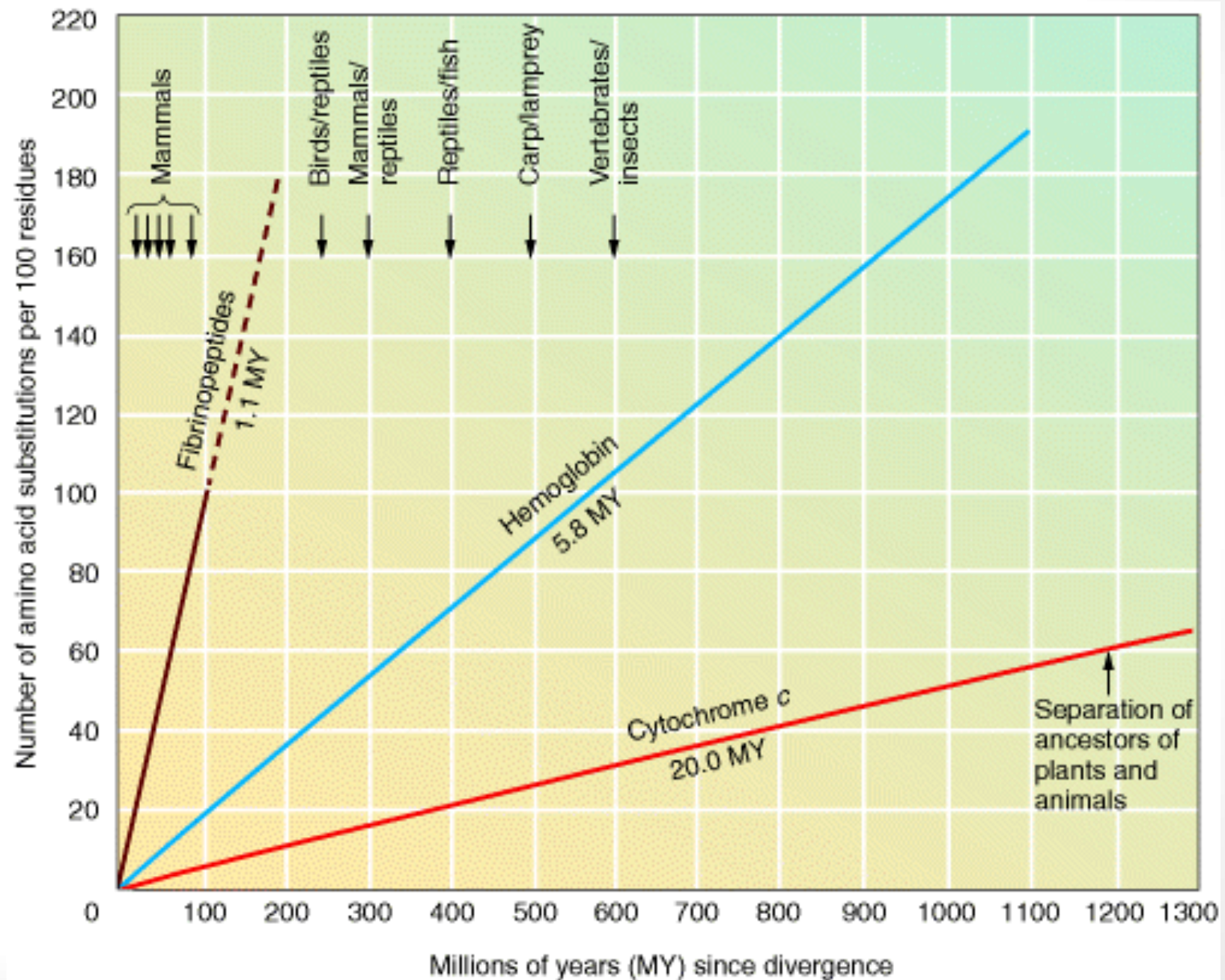
$K_S$ : Mutaciones Sinónimas

$K_A$ : Mutaciones No neutrales.





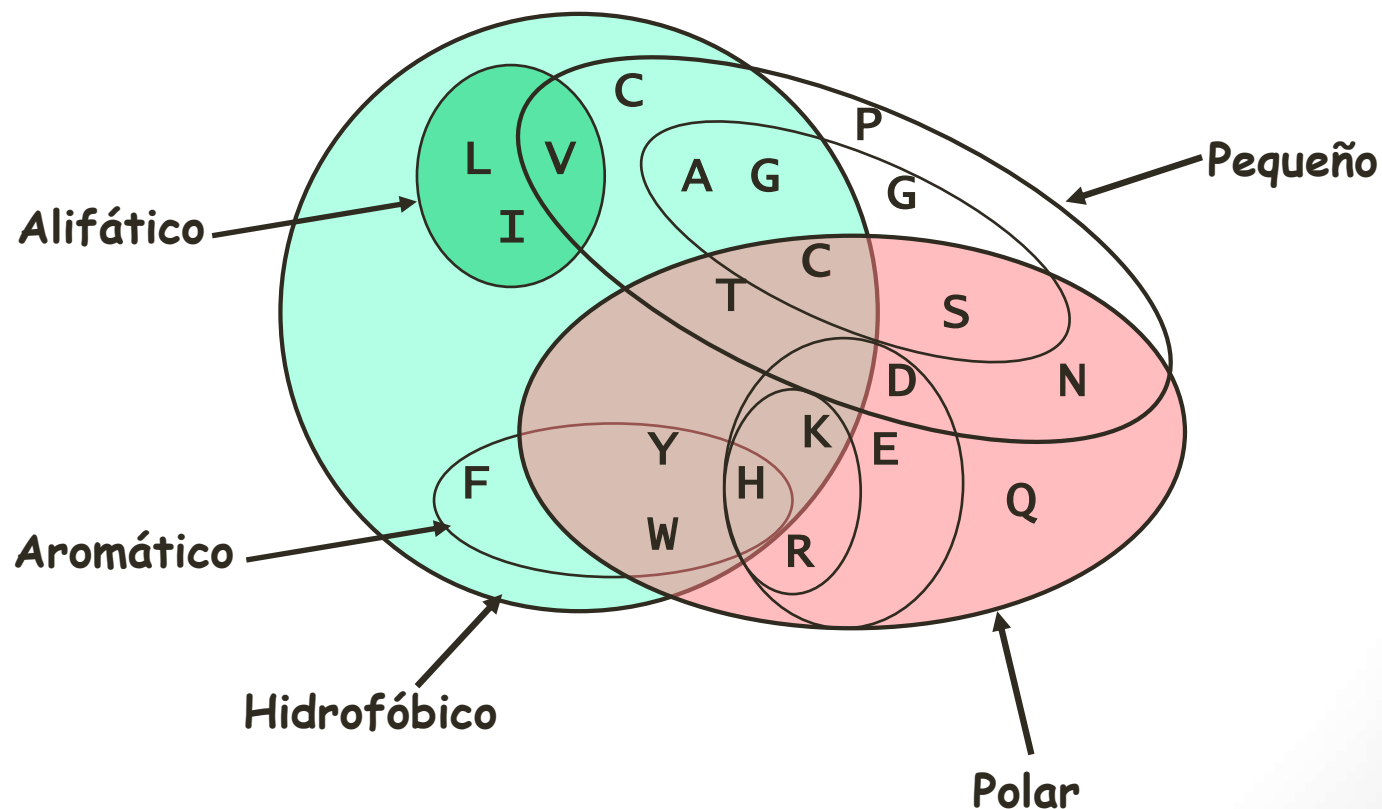
## Diferentes relojes moleculares para diferentes proteínas.



# ¿Cómo evolucionan las secuencias?

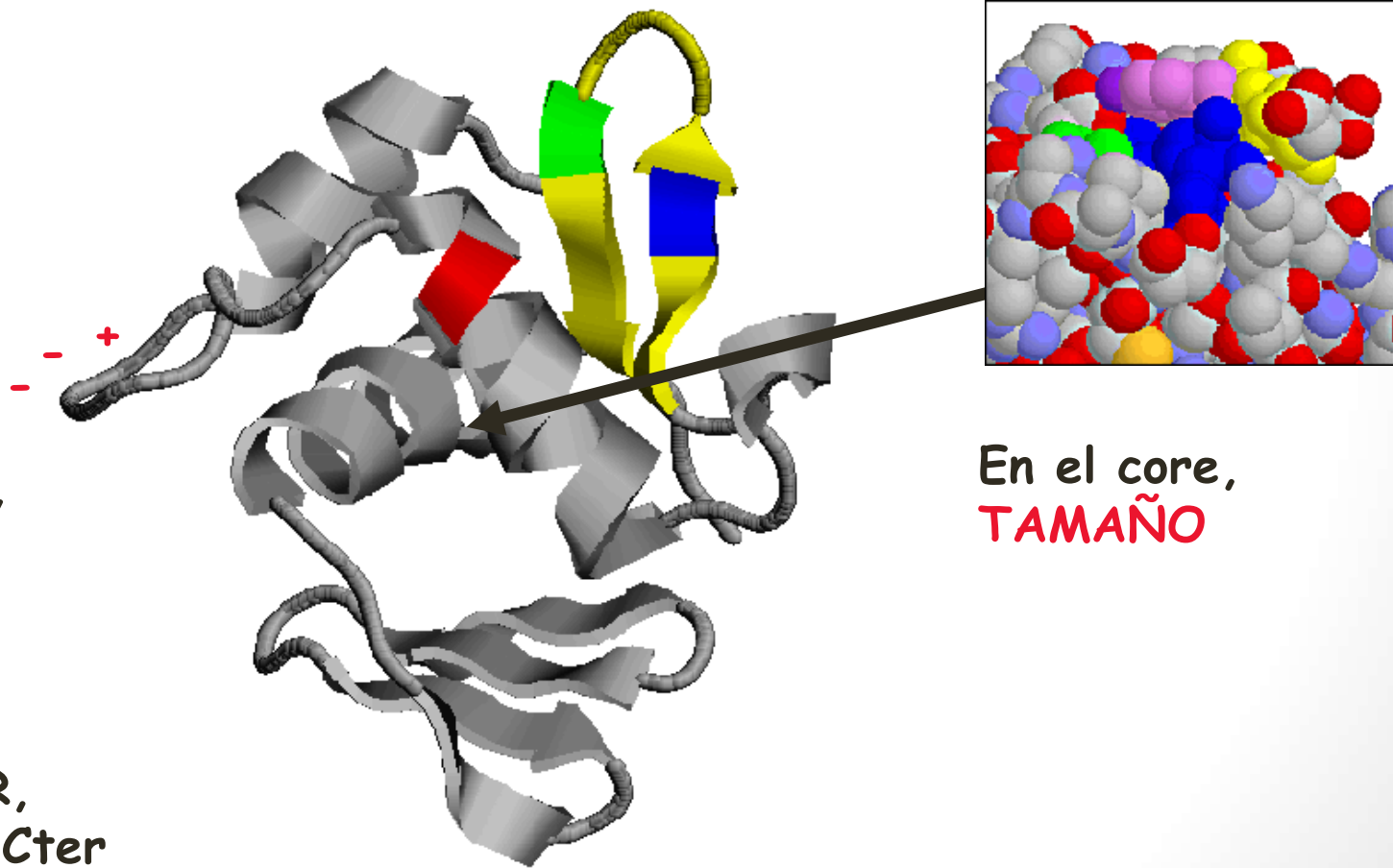
## Diagrama de Venn de Aminoácidos

Cada Aminoácido Posee su "Personalidad".



# ¿Cómo evolucionan las secuencias?

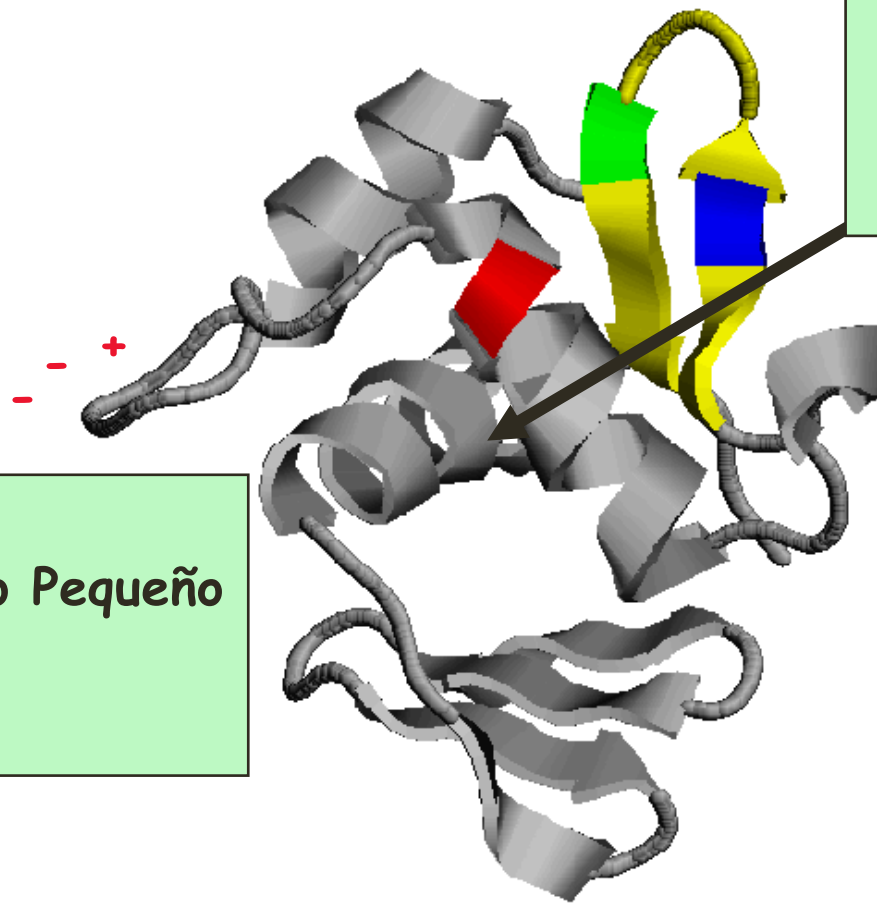
En una estructura, cada Aminoácido juega un papel fundamental.





# ¿Cómo evolucionan las secuencias?

Mutaciones Aceptadas Dependen de la Estructura.



Cargado -> Cargado  
Pequeño <-> Grande o Pequeño  
**DELECIONES**

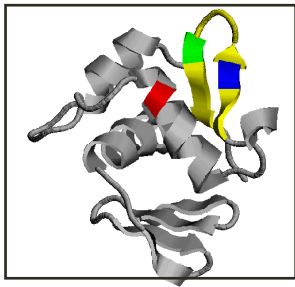
Grande -> Grande  
Pequeño ->Pequeño  
**NO DELECIÓN**

¿Cómo puedo comparar dos secuencias?

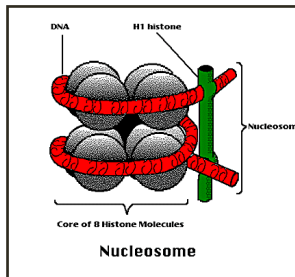
# MATRICES DE SUSTITUCIÓN

# ¿Cómo puedo comparar dos secuencias?

Para Comparar Dos Secuencias, Necesitamos:



Su Estructura

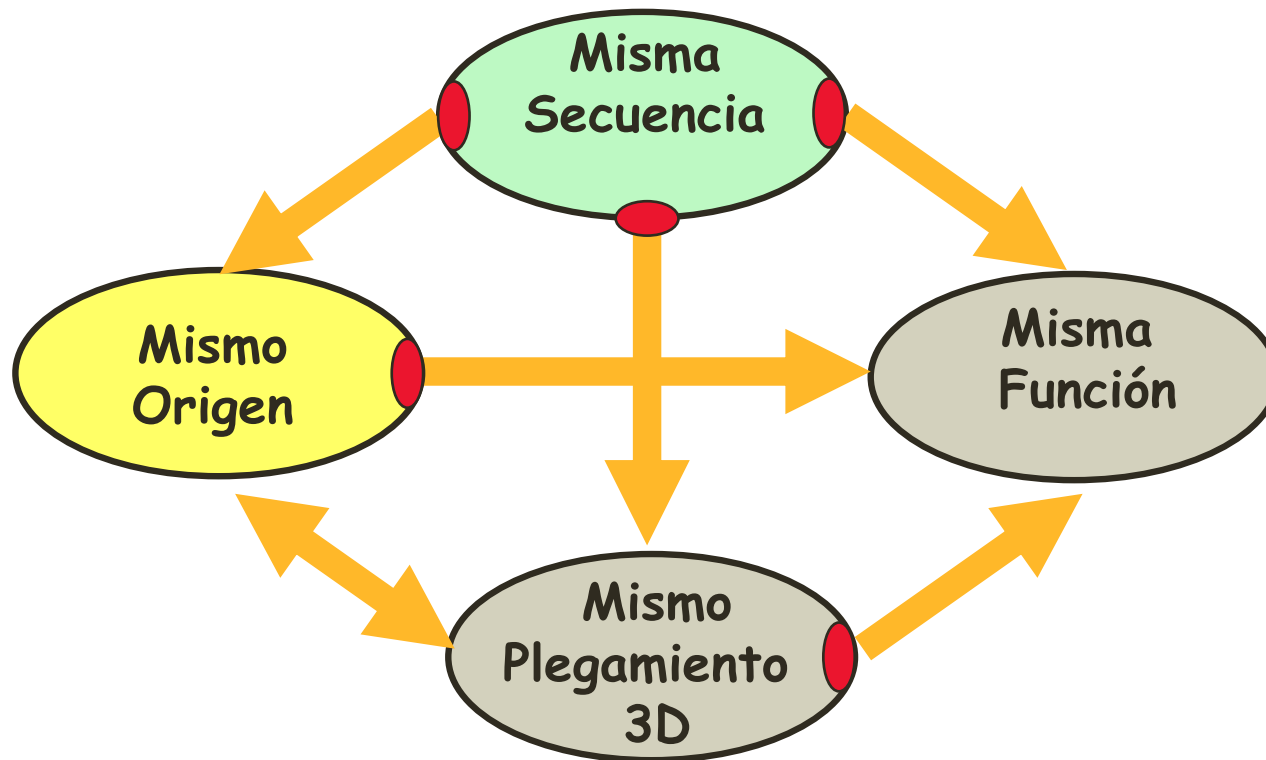


Su Función

Generalmente no  
contamos con las dos  
informaciones !!!

# ¿Cómo puedo comparar dos secuencias?

Nosotros podríamos necesitar reemplazar la información estructural con la información de secuencia.



**No Podemos Trabajar con todos al mismo tiempo!!!**

# ¿Cómo puedo comparar dos secuencias?

Para comparar secuencias, nosotros necesitamos comparar aminoácidos

Nosotros necesitamos conocer cual es el **COSTO** al **SUSTITUIR**

una Alanina por una Isoleucina  
un Triptófano por una Glicina

...

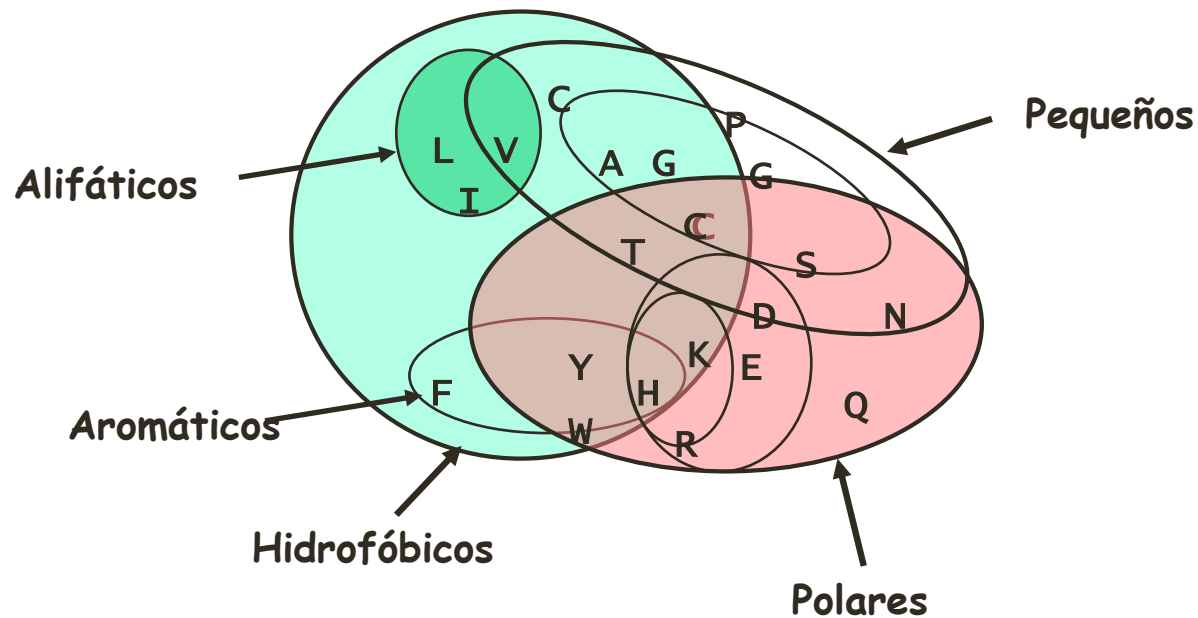
La tabla que contiene el costo para todas las posibles sustituciones es llamada la **MÁTRIZ de SUSTITUCIÓN**.

Cómo determinar está matriz?



# ¿Cómo puedo comparar dos secuencias?

Utilizando Conocimiento Podría Trabajar...



Pero, nosotros no conocemos su Evolución y Estructura 3D.

El uso de Datos trabaja mejor.

# ¿Cómo puedo comparar dos secuencias?

## “Cocinando” una matriz de sustitución.



- Toma 100 pares de secuencias de proteínas, fáciles de alinear (80% de identidad).

- Alineelas...

- Cuenta cada mutación en el alineamiento.

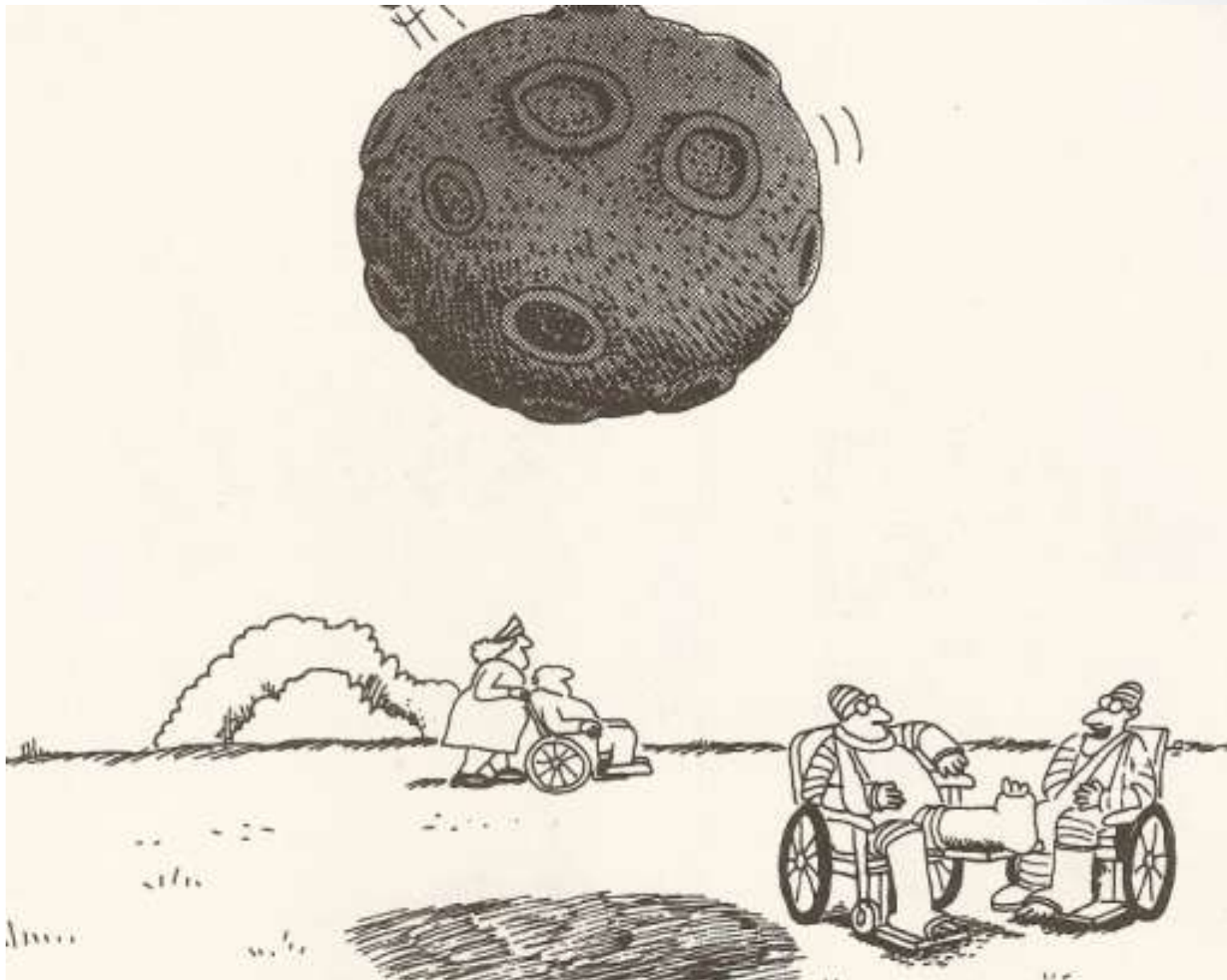
- 25 Trp a Phe

- 30 Iso a Leu

- ...

- Para cada mutación, calcula el puntaje de sustitución el radio (**log odd**):

$$\text{Log} \left( \frac{\text{Observed}}{\text{Expected by chance}} \right)$$



**You' re kidding! ... I was struck by a lightning twice too!!**

Garry Larson, The Far Side



# ¿Cómo puedo comparar dos secuencias?

## “Cocinando” una matriz de sustitución.



-Toma 100 pares de secuencias de proteínas, fáciles de alinear (80% de identidad).

-Alineelas...

-Cuenta cada mutación en el alineamiento.

-25 Trp a Phe

-30 Iso a Leu

...

- Para cada mutación, calcula el puntaje de sustitución el radio (**log odd**):

$$\text{Log} \left( \frac{\text{Observed}}{\text{Expected by chance}} \right)$$

$$\log \left( \frac{p_{ij}}{q_i * q_j} \right)$$

# ¿Cómo puedo comparar dos secuencias?

## “Cocinando” una matriz de sustitución.

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

*La Diagonal indica cuánto cuesta que evolucione ese aminoácido.  
W es MUY Conservado*

*Algunos residuos son fáciles de evolucionar en uno parecido.*

*Cisteínas que se encuentran en puentes disulfuro no han sido incluidas*

¿Cómo puedo comparar dos secuencias?  
 “Cocinando” una matriz de sustitución.

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

# ¿Cómo puedo comparar dos secuencias?

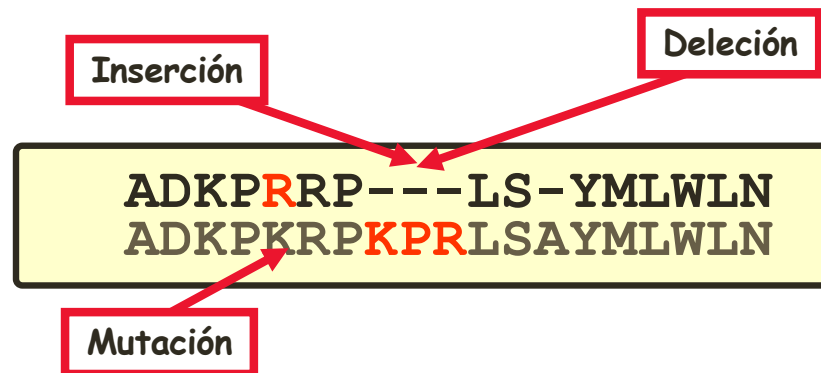
## Unidades PAM

- PAM (mutación puntual aceptada) es una unidad de distancia evolutiva entre dos secuencias de aminoácidos.
- 1 PAM = 1 mutación puntual aceptada ( no inserciones o deleciones) por cada 100 aminoácidos.
- 200 PAM = 200 mutaciones puntuales/ 100 aa (asume que las mutaciones pueden ocurrir múltiples veces en la misma posición).
- 2 secuencias que divergen por 200 PAM  $\cong$  25% identidad.

PAM a veces es también definida como el “porcentaje de mutaciones aceptada”

# ¿Cómo puedo comparar dos secuencias? Utilizando la matriz de sustitución

*Dadas dos secuencias y una matriz de sustitución,  
Nosotros deberíamos calcular el alineamiento más **BARATO***



# Función de Puntaje

Las matrices de sustitución más populares son:

- PAM250
- Blosum62 (Más utilizada)

## Puntaje crudo

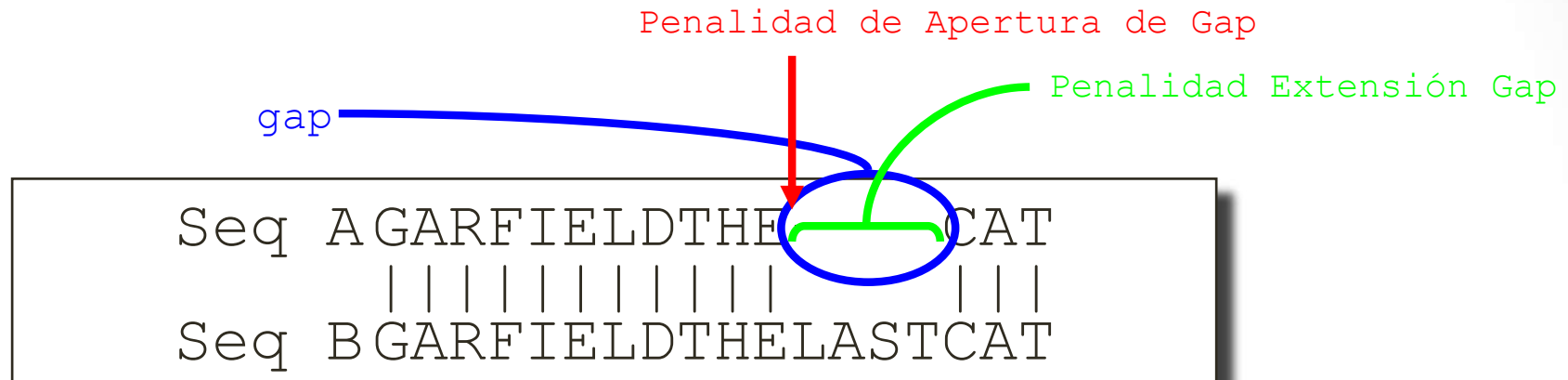
Puntaje = 1 + 6 + 0 + 2 = 9

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

- Pregunta: ¿Es posible tener un buen alineamiento sólo por azar?

# Inserciones y Deleciones

Penalidad de «Gap»

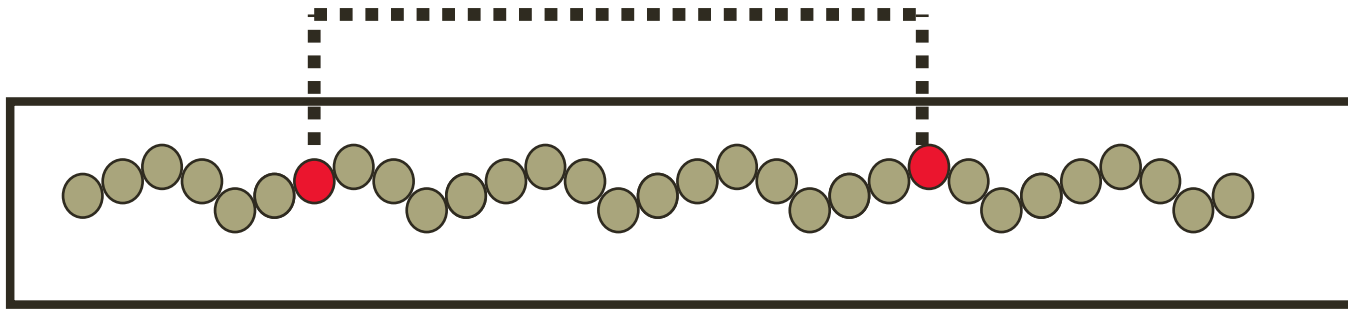


- Apertura de Gap es más costosa que la extensión.

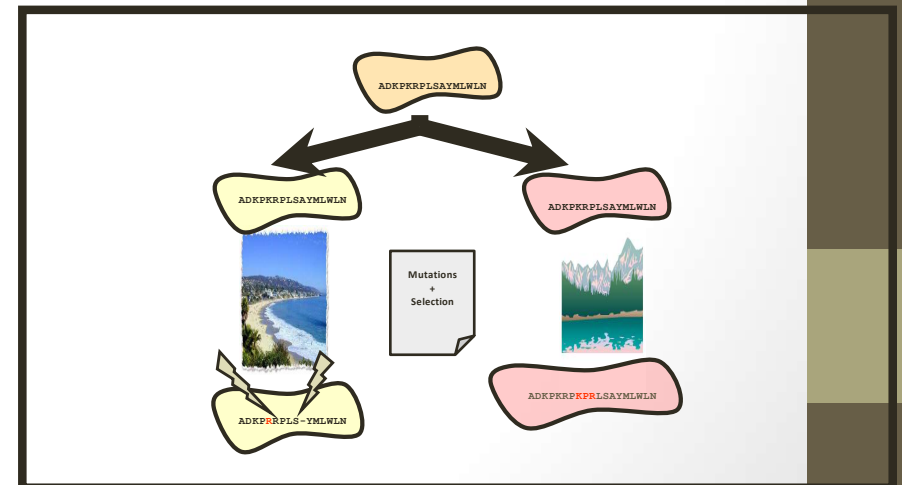
# ¿Cómo puedo comparar dos secuencias?

## Límites de la matriz de sustitución

Ellas ignoran interacciones no locales y asumen que residuo idénticos se comportan de igual manera



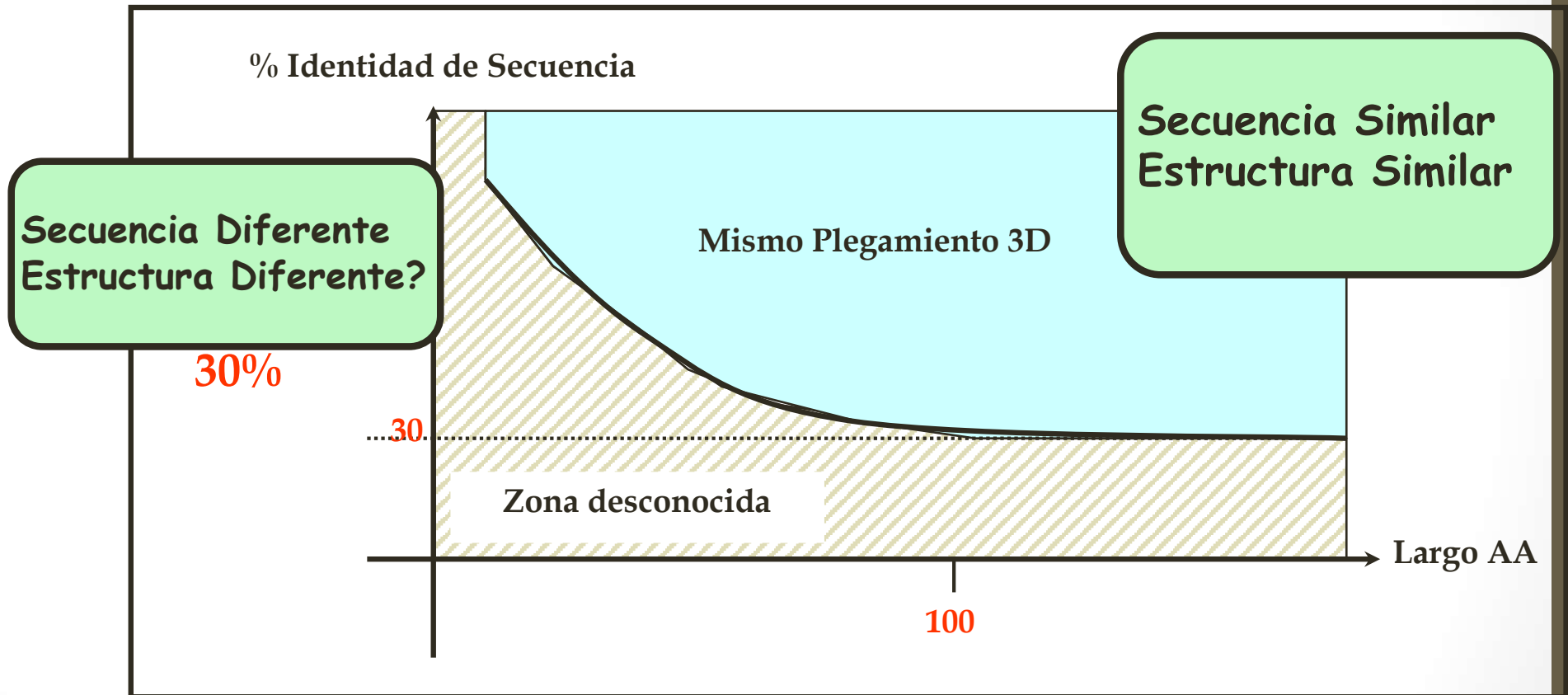
Ellas asumen que la velocidad de evolución es constante.





# ¿Cómo puedo comparar dos secuencias?

## La zona desconocida



Matrices de sustitución trabajan razonablemente bien en secuencias con más de un 30% de identidad y con sobre 100 aminoácidos.

# Matrices PAM

- El conjunto de matrices **PAM**, o **Point Accepted Mutation** (del inglés, mutación puntual aceptada).
- Introducida por Margaret Dayhoff a finales de los años 70.
- PAM1 estima el ritmo de sustitución esperado entre dos aminoácidos si el 1% de los aminoácidos cambian.
- Otras matrices PAM se derivan de la multiplicación de la PAM1 por sí misma ( $\text{PAM250} = \text{PAM1}^{250}$ ), para secuencias más remotas.
- Actualmente trabajan muy bien, PAM-250 es utilizada rutinariamente para buscar homólogos distantes.
- Pero existen algunos problemas con el modelo:
  - Este asume que todos los aminoácidos son igualmente mutables.
  - El modelo es obtenido utilizando las regiones más mutables y no las más conservadas, que reflejan regiones con importantes propiedades químicas y estructurales.
  - Derivada desde un pequeño grupo de secuencias de proteínas globulares disponible en la base de datos de 1979.

# Matrices BLOSUM

- BLOSUM (BLOcks of Amino Acid SUBstitution Matrix, o matriz de sustitución de bloques de aminoácidos. Se introdujo en 1992 por primera vez en un artículo de Henikoff y Henikoff.
- BLOSUM se usa para puntuar alineamientos entre secuencias de proteínas evolutivamente divergentes.
- Se basa en alineamientos locales, derivada desde la base de datos de BLOCKS, que es derivada desde PROSITE.
- BLOCKS generada desde secuencias de alineamiento múltiple sin “gaps” agrupada a varios umbrales de similaridad y corregida para impedir muestreo dirigido.
- Derivada desde datos que representan segmentos de secuencias altamente conservados desde proteínas.

# Matrices BLOSUM

- Muchas secuencias desde familias alineadas son utilizadas para generar las matrices.
- Las secuencias son idénticas a  $>X\%$  son eliminadas para impedir sobrerrepresentación de proteínas en la base de datos.
- Las matrices específicas se refieren a estos “cut-off”. Por ejemplo, BLOSUM62 refleja las sustituciones observada entre segmentos con menos de 62% de identidad.
- En analogía las matrices PAM, un logaritmo es calculado desde las frecuencias  $A_{ij}$  de residuos observado  $i$  en un clúster alineado contra el residuo  $j$  en otro clúster.

# Matrices PAM v/s BLOSUM

- Las matrices BLOSUM han reemplazado a las PAM como matrices por defecto en varios sitios de búsqueda de bases de datos. Como BLAST <http://blast.ncbi.nlm.nih.gov> y FASTA <http://www.ebi.ac.uk/Tools/sss/fasta>.
- Tanto PAM120 y BLOSUM62 trabajan mejor para proteínas moderadamente divergentes y podrían quedar similitudes fuera de su ventana óptima.
- PAM provee la alternativa para secuencias cortas fácilmente accesible (no existe una apropiada versión de BLOSUM disponible).
- La mejor solución es proveer un rango de sistemas de puntaje, que es actualmente la práctica de servidores principales.
- La configuración apropiada de penalidades de “gap” posee un gran efecto en la sensibilidad de la matriz.

# ¿Cómo puedo comparar dos secuencias?

## ¿Qué matriz utilizar?

La inicial matriz PAM fue calculada en proteínas con 80% de similitud.

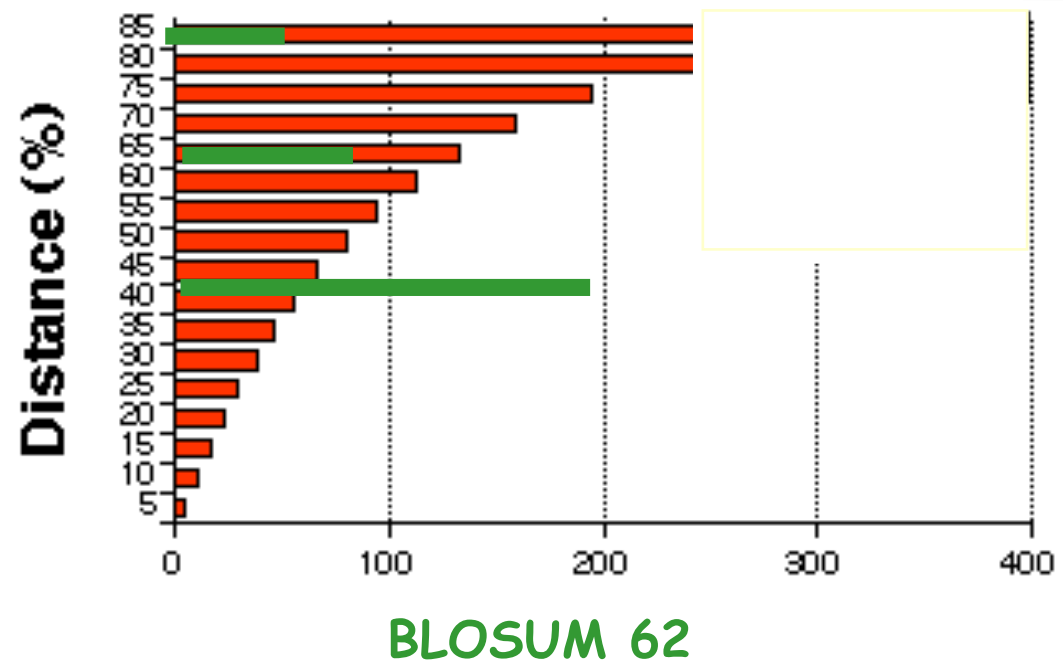
Esto ha sido extrapolado a secuencias de proteínas más distantes.

PAM 250

PAM 350

BLOSUM 42

BLOSUM 62



# ¿Cómo puedo comparar dos secuencias?

## ¿Qué matriz utilizar?

**PAM: Proteínas Distantes ⇔ Alto índice (PAM 350)**

**BLOSUM: Proteínas Distantes ⇔ Bajo Índice (BLOSUM30)**

**La elección de la matriz correcta puedes ser engañoso**

- **GONNET 250 > BLOSUM62 > PAM 250.**
- **Esto depende de:**
  - **La familia.**
  - **El programa usado y su configuración.**
- **Inserciones, Deleciones.**

¿Cómo puedo alinear dos secuencias?

MATRIZ DE PUNTO  
ALINEAMIENTO GLOBAL  
ALINEAMIENTO LOCAL



# Comparación métodos de alineamiento de pares

Método	Situación	Referencia
Dot-plot	<b>Exploración general de secuencias</b> Descubrir repeticiones Buscar inserciones y deleciones Extraer porciones de la secuencia	
Alineamiento local	<b>Comparar secuencias con homología parcial</b> Realizar alineamientos de alta calidad Realizar análisis de AA por AA	
Alineamiento global	<b>Comparar dos secuencias en su largo entero</b> Identificar largas inserciones/deleciones Verificar la calidad de tu data Identificar mutaciones en tu secuencia.	

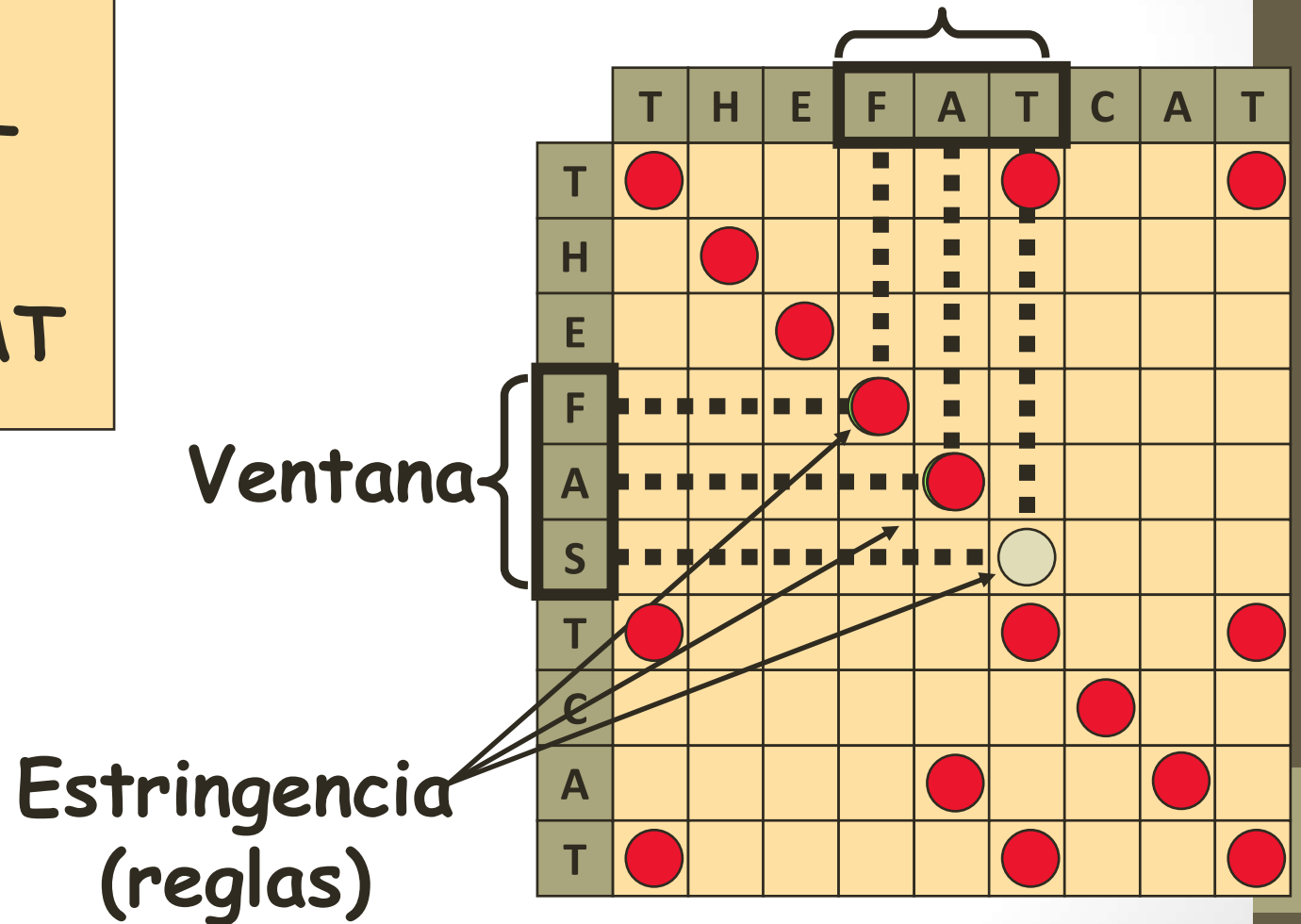
# Matrices de Punto

## Pregunta

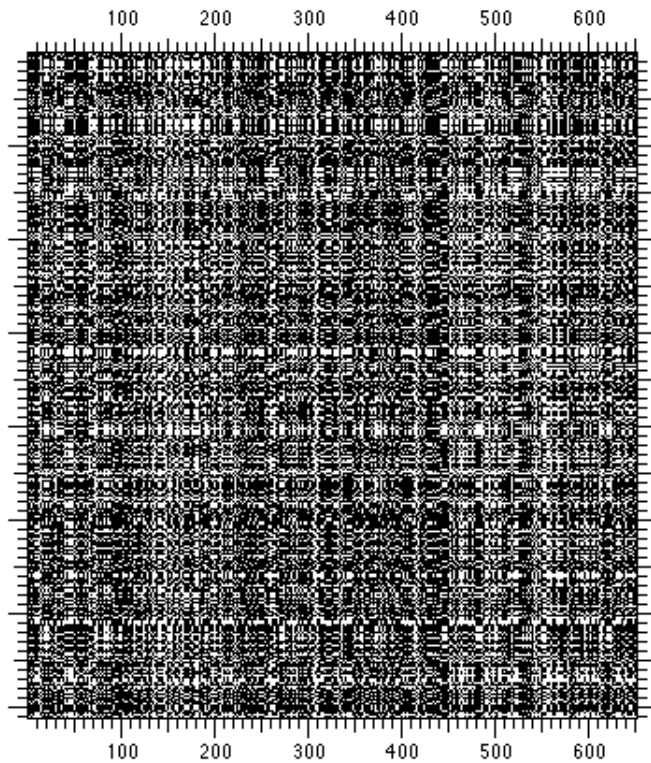
¿Cuáles son los elementos compartidos por las secuencias?

# Matrices de Punto

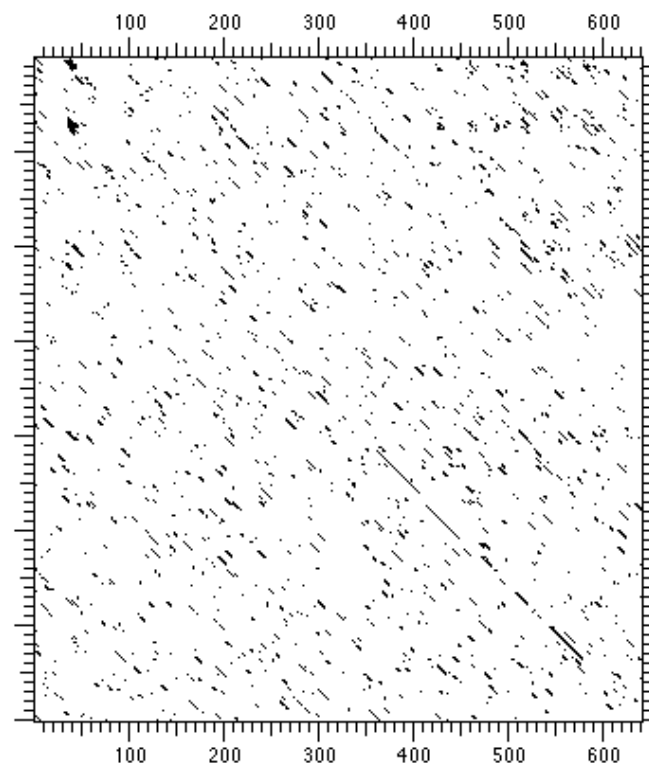
>Seq1  
THEFATCAT  
>Seq2  
THELASTCAT



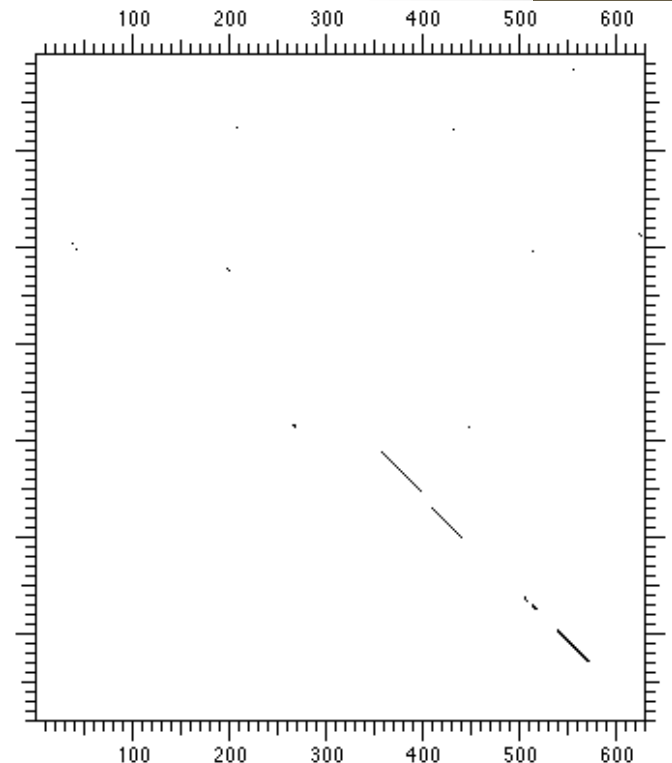
# Matrices de Punto Estringencia



*ventanas=1*  
*estringencia=1*



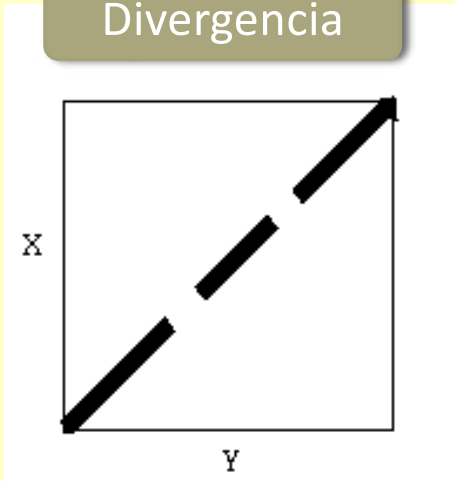
*ventanas=11*  
*estringencia=7*



*ventanas=25*  
*estringencia=15*

# Matrices de Punto Evaluación

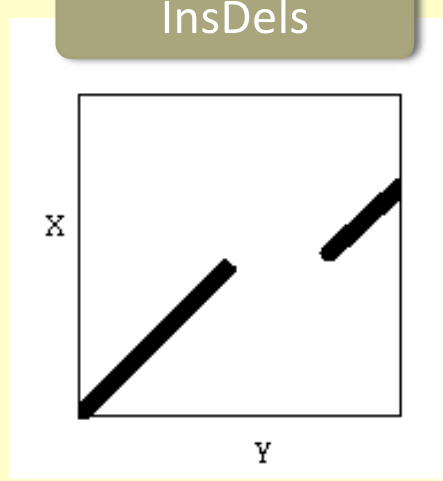
Divergencia



X 

Y 

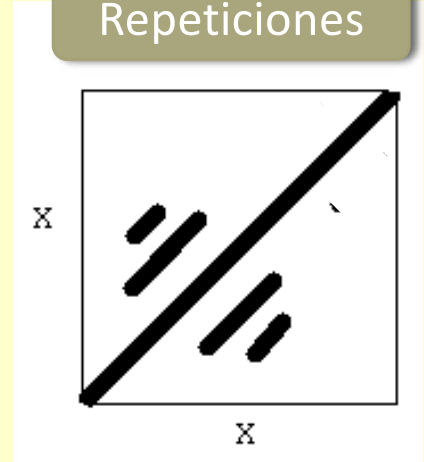
InsDels



X 

Y 

Repeticiones



X 

# Matrices de Punto

## Limitaciones

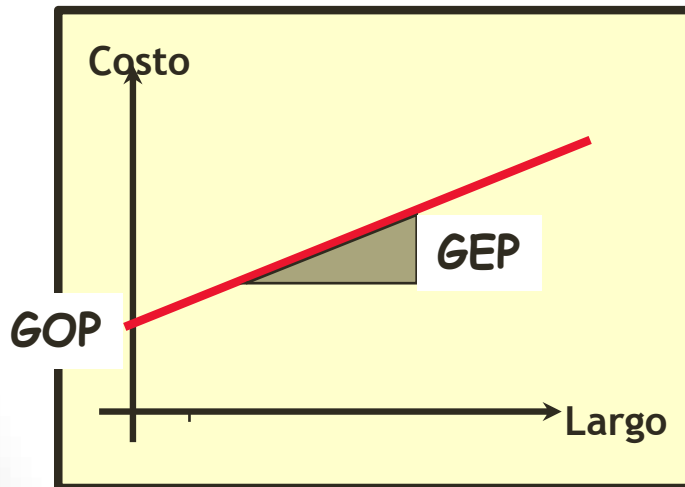
- Ayuda Visual
- Es la mejor vía para **EXPLORAR** la organización de una secuencia.
- **NO** nos provee con un **ALINEAMIENTO**

```
trigo  --DPNKPKRAMTSFVFFMSEFRSEFKQKHSKLKSIVEMVKAAGER
      | | ||||| | | | | | | | | | | | | | | | | |
????? KKDSNAPKRAMTSFMFFSSDFRS----KHSDL-SIVEMSKAAGAA
```

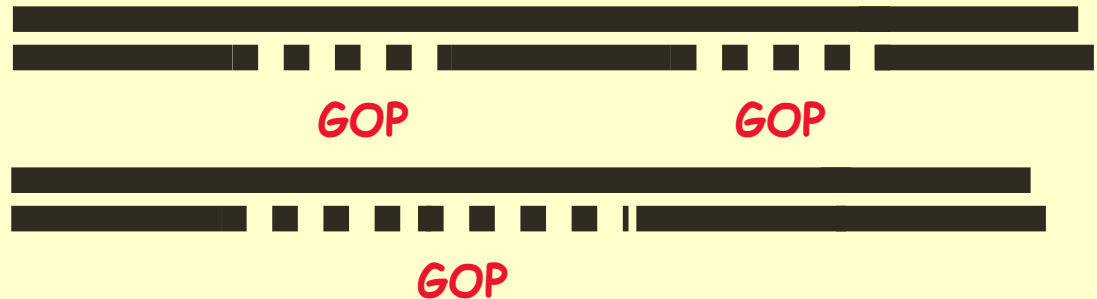
# Alineamiento Global



- Tomar dos secuencias de proteínas.
- Una buena matriz de sustitución (blosum)
- Una penalidad de apertura de Gap (GOP)
- Una penalidad de extensión de Gap (GEP)



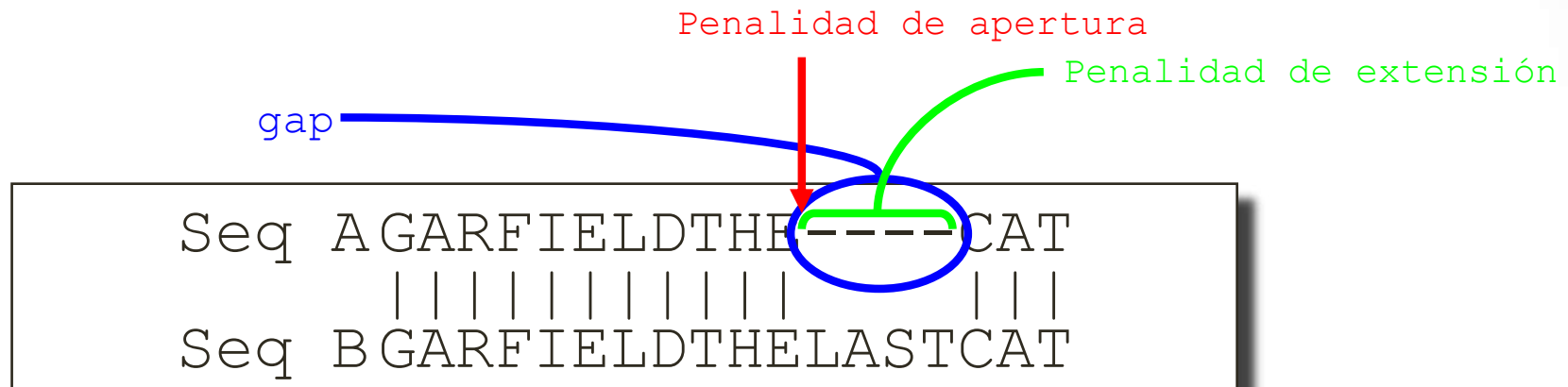
Penalidad de Gap



Parsimonia:  
Evolución toma la vía más simple  
(Eso es lo que pensamos...)

# Inserciones y Delecciones

## Penalidades de Gap



- Apertura de Gap es más costosa que la extensión.

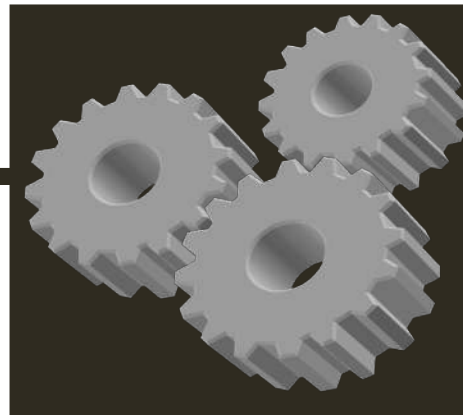


# Alineamiento Global



- Tomar dos secuencias de proteínas.
- Una buena matriz de sustitución (blosum).
- Una penalidad de apertura de Gap (GOP).
- Una penalidad de extensión de Gap (GEP).
- PROGRAMACIÓN DINÁMICA.**

>Seq1  
THEFATCAT  
>Seq2  
THEFASTCAT



THEFA-TCAT  
THEFASTCAT

**PROGRAMACIÓN DINÁMICA**

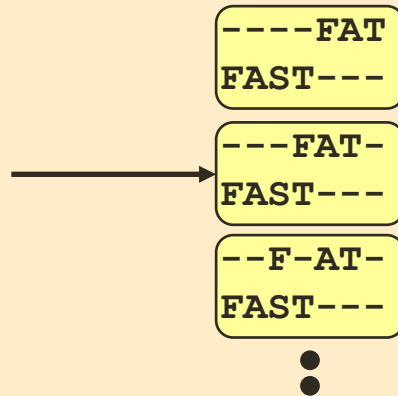
# Alineamiento Global

## Programación Dinámica

### Enumeración de Fuerza Bruta

F A S T

F A T



$$\left( \frac{(L1+L2)!}{(L1)!*(L2)!} \right)^2$$

# Alineamiento Global

## Programación Dinámica

Programación Dinámica (Needlman and Wunsch)

Match=1 MisMatch=-1 Gap=-1

	F	A	S	T	
F	0	-1	-2	-3	-4
A	-1	1	0		
T	-2	0	2		
	-3				

	F	A	S	T	
F	0	-1	-2	-3	-4
A	-1	1	0	-1	0
T	-2	0	2	1	0
T	-3	-1	-1	1	2

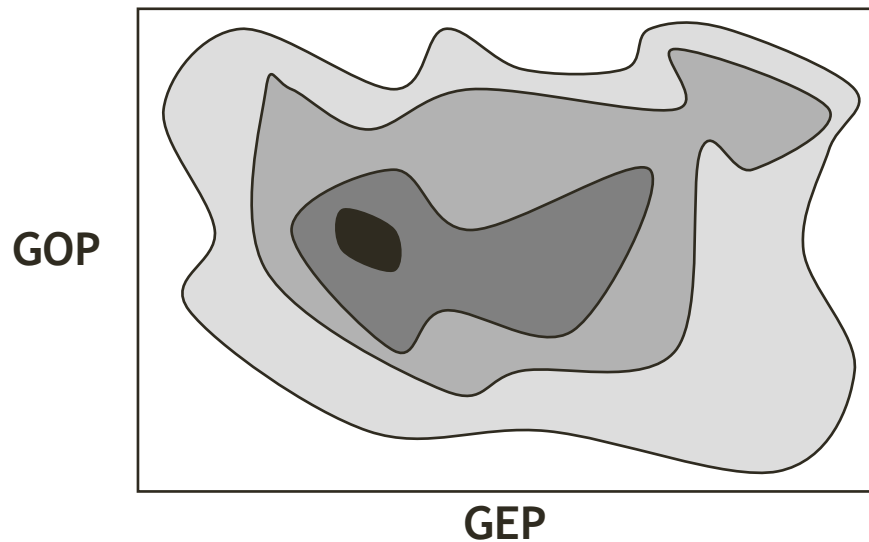
	F	A	S	T	
F	0	-1	-2	-3	-4
A		1			
T			2	1	
					2

F A S T  
F A - T

# Alineamiento Global

## Programación Dinámica

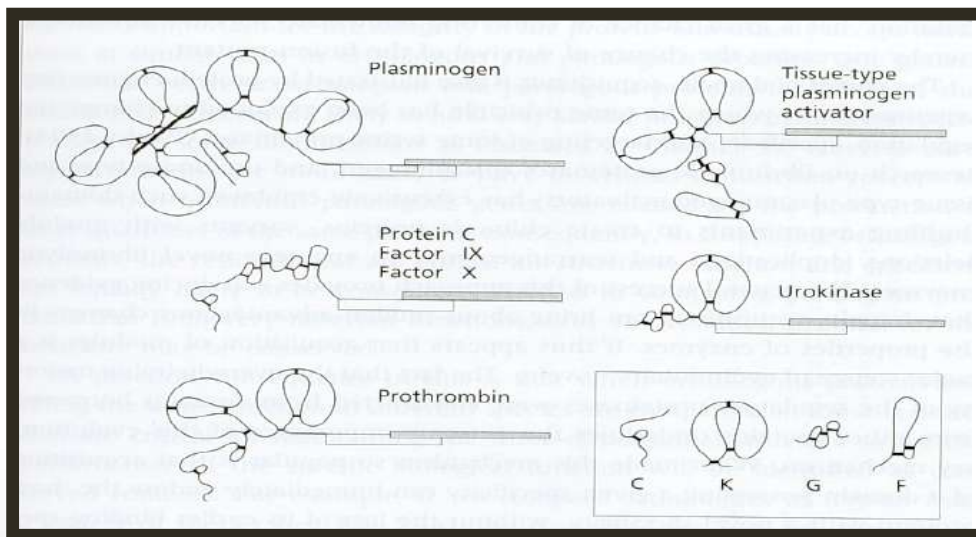
**Alineamiento Globales son muy sensibles a penalidades de "gap".**



# Alineamiento Global Programación Dinámica

Alineamientos globales son muy sensible a las penalidades de "gap".

Alineamientos globales no consideran la naturaleza **MODULAR** de las proteínas.



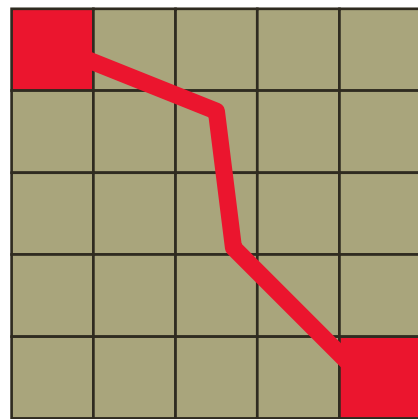
C: vitamina K dep. Ca

K: Dominio Kring

G: Dominio Factor Crecimiento

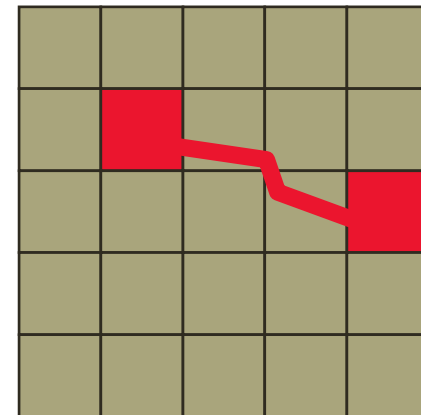
F: Modulo dedos

# Alineamiento Local



Global FTFTALILLAVAV  
F--TAL-LLA-AV

Local FTFTALILL-AVAV  
--FTAL-LLAAV--



Alineamiento  
**GLOBAL**

Alineamiento  
**LOCAL**

# Alineamiento Local

- El algoritmo SW fue propuesto por Temple Smith y Michael Waterman en 1981.
- El algoritmo de Smith-Waterman es una reconocida estrategia para realizar alineamiento local de secuencias biológicas (ADN, ARN o proteínas); es decir que determina regiones similares entre un par de secuencias.
- Este garantiza la búsqueda de un alineamiento local con respecto a un sistema de puntaje (la matriz de sustitución y el esquema de puntaje dependiente de gaps).
- La principal diferencia con el algoritmo de Needleman–Wunsch es la matriz de puntaje negativa configurada a ceros, que es poblada con el alineamiento local (puntuajes positivos).

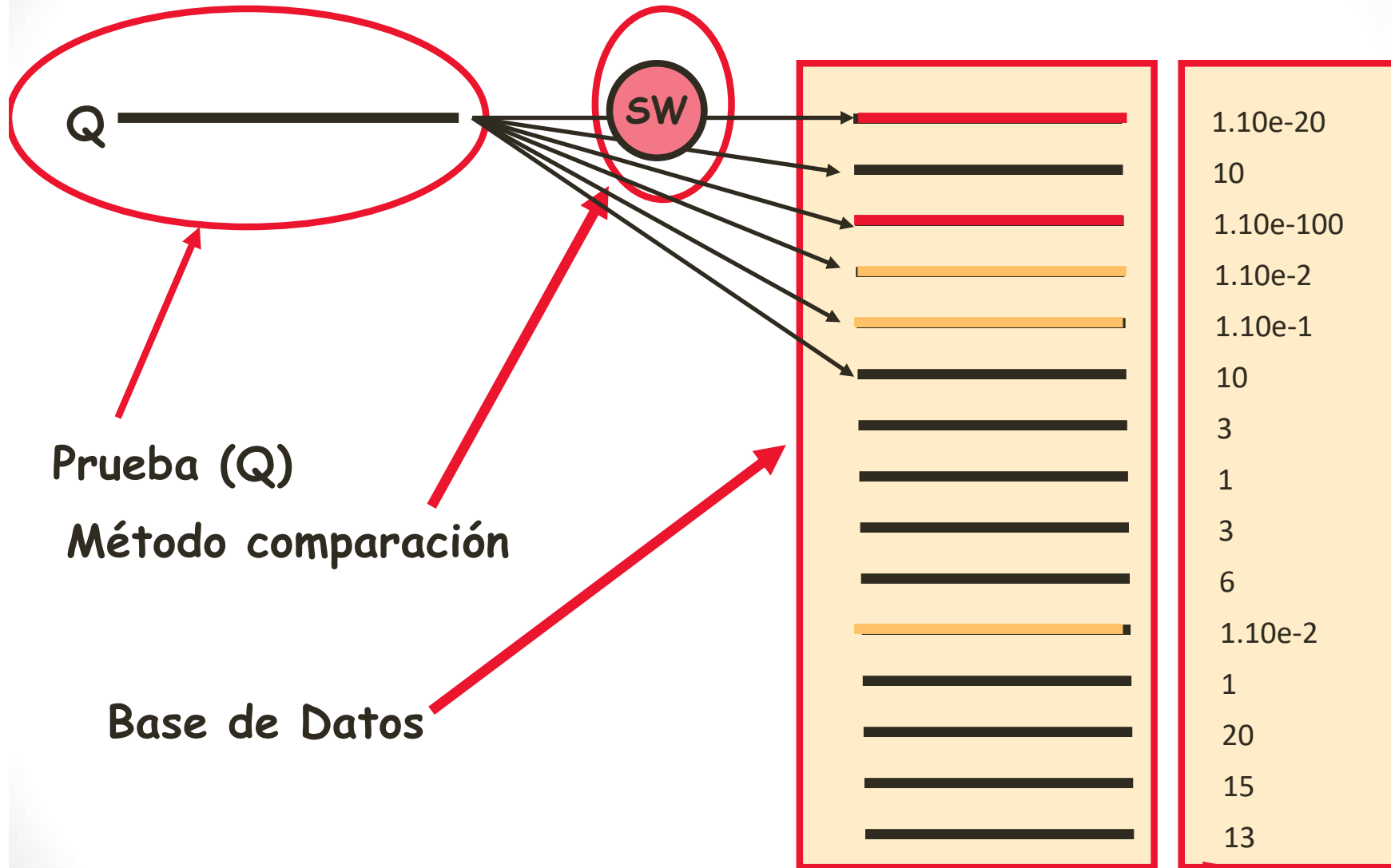
# Alineamiento Local seudocódigo

```
for i=0 to length(A)
  F(i,0) ← d*i
for j=0 to length(B)
  F(0,j) ← d*j
for i=1 to length(A)
  for j=1 to length(B)
  {
    Match ← F(i-1,j-1) + S(Ai, Bj)
    Delete ← F(i-1, j) + d
    Insert ← F(i, j-1) + d
    F(i,j) ← max(Match, Insert, Delete)
  }
```

```
AlignmentA ← ""
AlignmentB ← ""
i ← length(A)
j ← length(B)
while (i > 0 or j > 0)
{
  if (i > 0 and j > 0 and F(i,j) == F(i-1,j-1) + S(Ai, Bj))
  {
    AlignmentA ← Ai + AlignmentA
    AlignmentB ← Bj + AlignmentB
    i ← i - 1
    j ← j - 1
  }
  else if (i > 0 and F(i,j) == F(i-1,j) + d)
  {
    AlignmentA ← Ai + AlignmentA
    AlignmentB ← "-" + AlignmentB
    i ← i - 1
  }
  else (j > 0 and F(i,j) == F(i,j-1) + d)
  {
    AlignmentA ← "-" + AlignmentA
    AlignmentB ← Bj + AlignmentB
    j ← j - 1
  }
}
```



# Búsqueda en base de datos



Prueba (Q)

Método comparación

Base de Datos

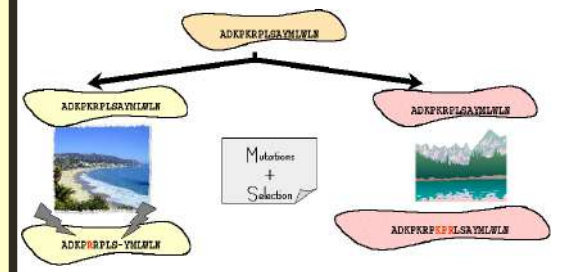
Valores E-

¿Cuántas veces esperas tener este alineamiento al azar?

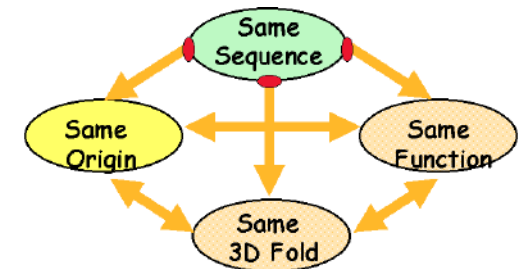
# RESUMEN

# Comparación de Secuencias

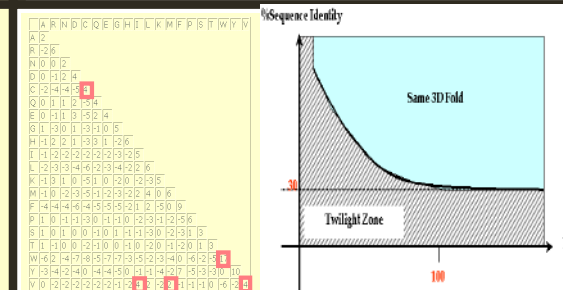
- Gracias a la Evolución, nosotros podemos comparar secuencias.



- Esta es una relación entre secuencia y estructura.



- Las matrices de sustitución sólo trabajan bien con secuencias similares (Más que 30% id).

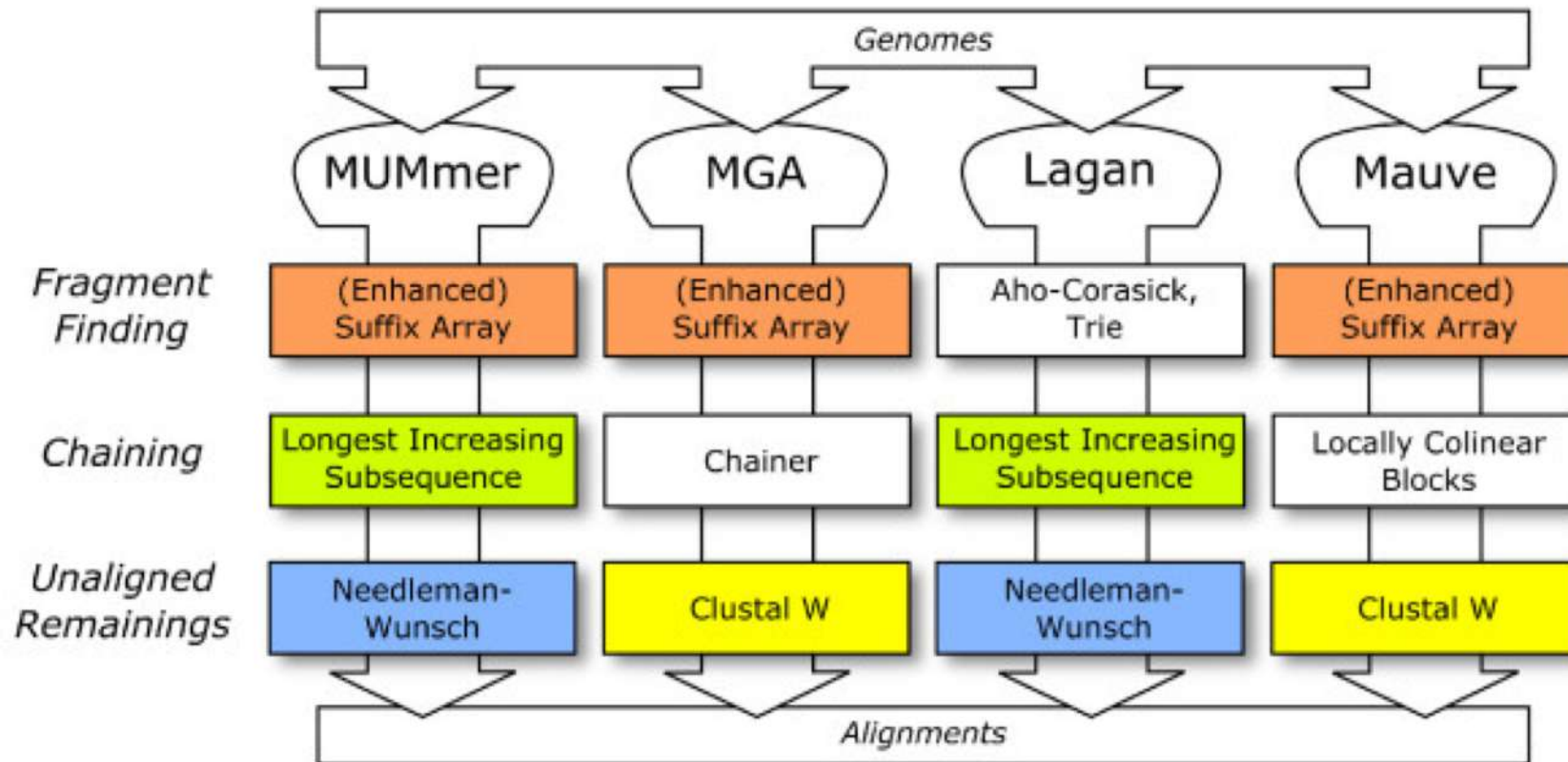


La vía más fácil de comparar secuencias es un gráfico de puntos.



RECURSOS ONLINE

# Recursos online



## Online Pairwise Alignment Programs

<i>Name</i>	<i>Address</i>	<i>Alignment type</i>
<b>lalign</b>	<a href="http://www.ch.embnet.org/software/LALIGN_form.html">www.ch.embnet.org/software/LALIGN_form.html</a>	Global/Local
<b>lalign</b>	<a href="http://fasta.bioch.virginia.edu/fasta_www/plalign.htm">http://fasta.bioch.virginia.edu/fasta_www/plalign.htm</a>	Global/Local
<b>USC</b>	<a href="http://www-hto.usc.edu/software/seqaln/seqaln-query.html">www-hto.usc.edu/software/seqaln/seqaln-query.html</a>	Global/Local/Exotic(!)
<b>alion</b>	<a href="http://fold.stanford.edu/alion/">fold.stanford.edu/alion/</a>	Global/Local
<b>align</b>	<a href="http://genome.cs.mtu.edu/align.html">genome.cs.mtu.edu/align.html</a>	Global/Local
<b>align</b>	<a href="http://www.ebi.ac.uk/emboss/align/">www.ebi.ac.uk/emboss/align/</a>	Global/Local
<b>Blast2seqs</b>	<a href="http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html">www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html</a>	Local BLAST
<b>Blast2seqs</b>	<a href="http://web.umassmed.edu/cgi-bin/BLAST/blast2seqs">web.umassmed.edu/cgi-bin/BLAST/blast2seqs</a>	Local BLAST

---

## Online Pairwise Alignments Analyses

---

<i>Name</i>	<i>Address</i>	<i>Function</i>
lalnview	<a href="http://www.expasy.ch/tools/sim-prot.html">www.expasy.ch/tools/sim-prot.html</a>	Visualization
prss	<a href="http://www.ch.embnet.org/software/PRSS_form.html">www.ch.embnet.org/software/PRSS_form.html</a>	Evaluation
prss	<a href="http://fasta.bioch.virginia.edu/fasta/prss.htm">fasta.bioch.virginia.edu/fasta/prss.htm</a>	Evaluation
graph-align	<a href="http://darwin.nmsu.edu/cgi-bin/graph_align.cgi">darwin.nmsu.edu/cgi-bin/graph_align.cgi</a>	Evaluation

---

---

## Online Pairwise Alignments Analyses

---

<i>Name</i>	<i>Address</i>	<i>Function</i>
lalnview	<a href="http://www.expasy.ch/tools/sim-prot.html">www.expasy.ch/tools/sim-prot.html</a>	Visualization
prss	<a href="http://www.ch.embnet.org/software/PRSS_form.html">www.ch.embnet.org/software/PRSS_form.html</a>	Evaluation
prss	<a href="http://fasta.bioch.virginia.edu/fasta/prss.htm">fasta.bioch.virginia.edu/fasta/prss.htm</a>	Evaluation
graph-align	<a href="http://darwin.nmsu.edu/cgi-bin/graph_align.cgi">darwin.nmsu.edu/cgi-bin/graph_align.cgi</a>	Evaluation

---



---

## Various flavors of dot-plot programs

---

<i>Name</i>	<i>For</i>	<i>Range</i>	<i>URL</i>	<i>Platforms</i>
<b>Dotlet</b>	Proteins DNA	10.000	<a href="http://www.ch.embnet.org">www.ch.embnet.org</a>	All
<b>Dotter</b>	Proteins DNA	100.000	<a href="http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html">www.cgr.ki.se/cgr/ groups/sonnhammer/ Dotter.html</a>	Unix Linux Windows
<b>Dottup</b>	DNA	Complete Genomes	<a href="http://www.emboss.org">www.emboss.org</a>	Unix Linux

---

# Guidelines for using PAM matrices

The relative entropy H of PAM matrices (from Table 1)		
PAM distance	H (bits)	Min. signif length (30 bits)
40	2-26	14
120	0-98	31
250	0-36	83

Ranges of local alignment lengths for which  
various PAM matrices are appropriate  
(from Table 3)

PAM matrix	93% efficiency range for database searching (30 bits)
40	9-21
120	19-50
240	47-123

from Altschul, "Amino acid substitution matrices from an information theoretic perspective"