

**FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS  
DEPARTAMENTO DE BIOTECNOLOGÍA Y TECNOLOGÍA ALIMENTARIA  
UNIVERSIDAD ARGENTINA DE LA EMPRESA**

# **Bioinformática**

**ANÁLISIS COMPUTACIONAL DE SECUENCIAS**

**Dr. Lucas L. Maldonado (PhD)**

**Lic. Biotechnologist and Molecular Biologist**

**Bioinformatics and genomics specialist**

**CONICET**

**Fac. de Medicina - UBA**

**Fac. de Ciencias Exactas y Naturales – UBA**

[lucamaldonado@uade.edu.ar](mailto:lucamaldonado@uade.edu.ar)

[lmaldonado@fmed.uba.ar](mailto:lmaldonado@fmed.uba.ar)

[luscas.l.maldonado@gmail.com](mailto:luscas.l.maldonado@gmail.com)

# Alineamiento de secuencias

**Principios: que es? Para que?**

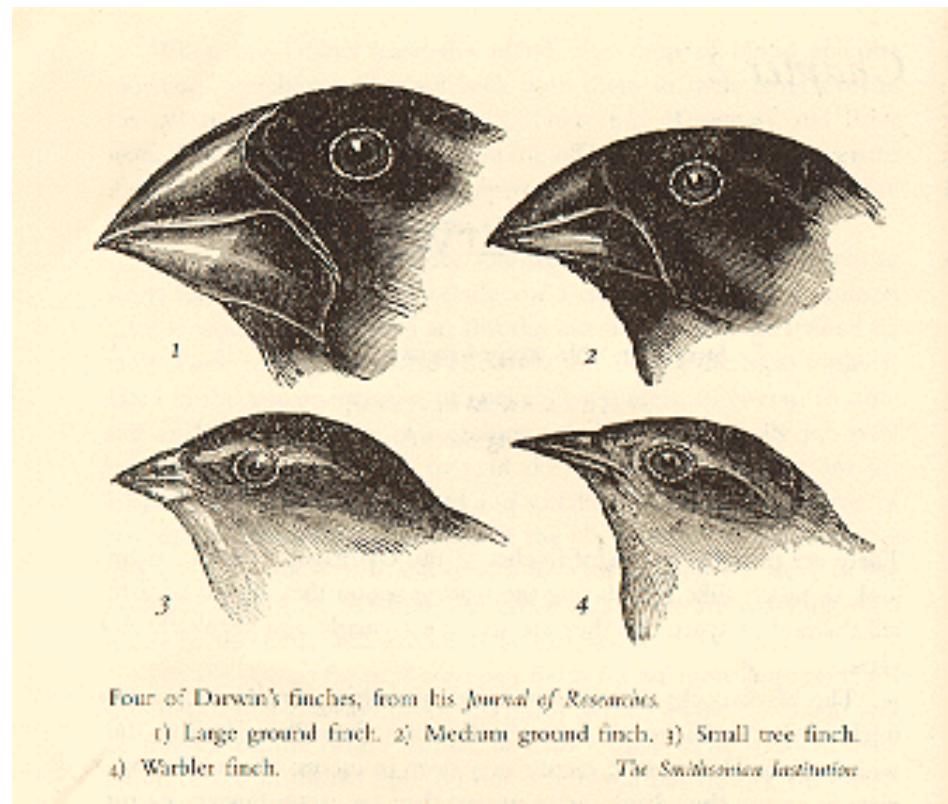
- **Algoritmos**
- **Dot-Plots**
- **NW Puntuación: Simple y Matrices**
- **Programas: BLAST, Fasta**
- **Alineamiento Múltiple**

# Análisis comparativo

## ¿Qué comparamos en biología?

El alineamiento de secuencias es similar a otros tipos de análisis comparativo.

En ambos es necesario cuantificar las similitudes y diferencias (scoring) entre un grupo relacionado de entidades.



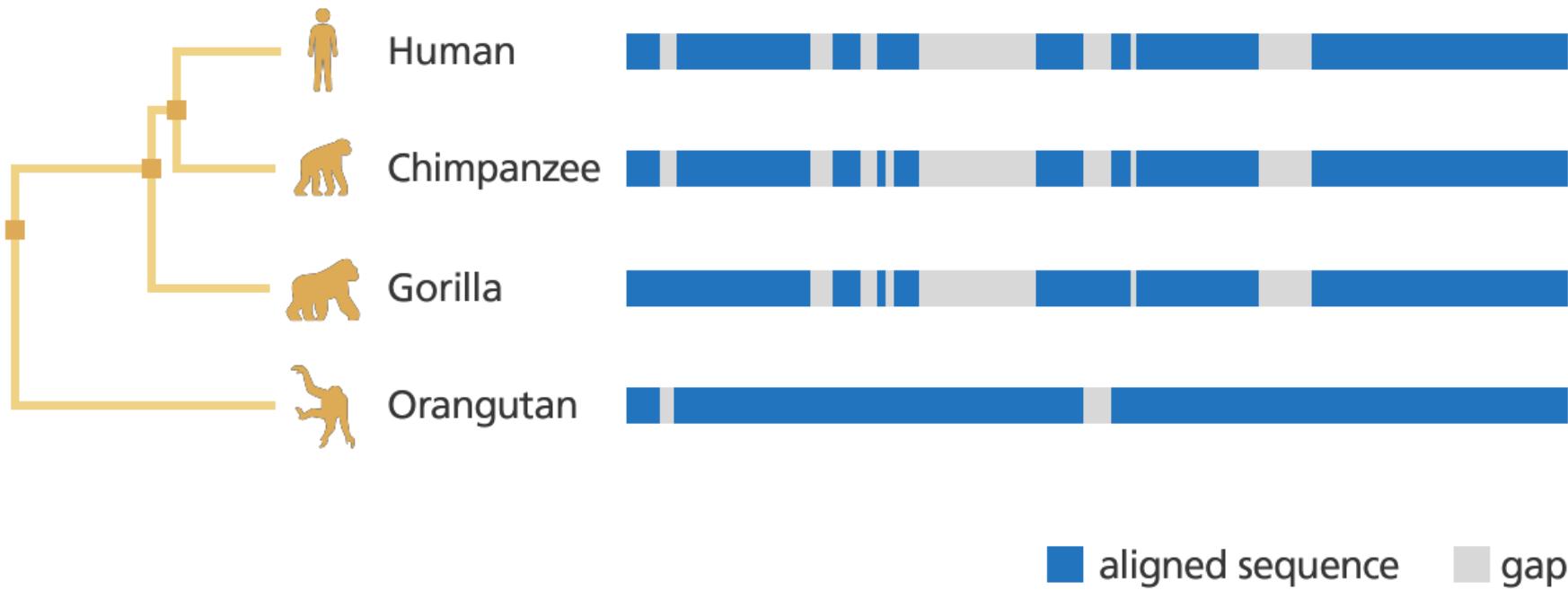
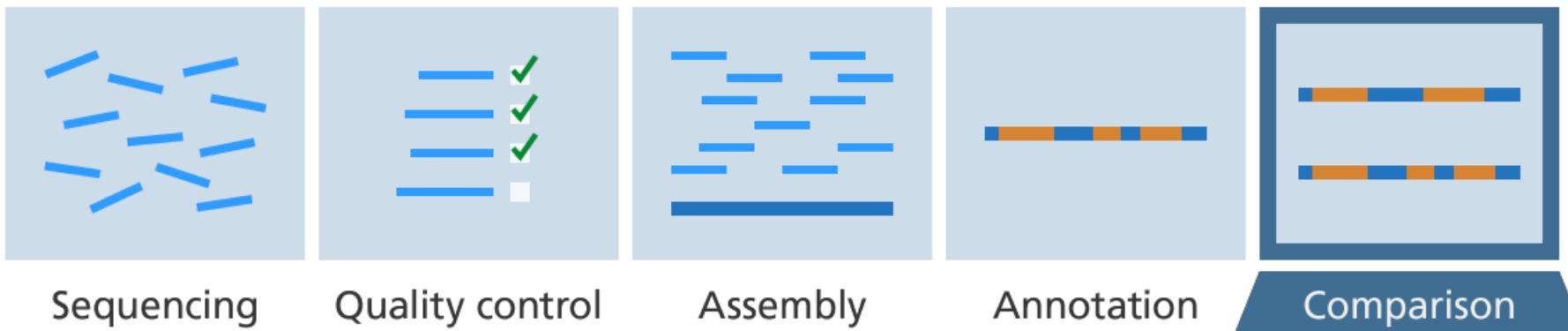
Four of Darwin's finches, from his *Journal of Researches*.

- 1) Large ground finch.
- 2) Medium ground finch.
- 3) Small tree finch.
- 4) Warbler finch.

The Smithsonian Institution

Finches of the Galápagos Islands observed by Charles Darwin on the voyage of HMS *Beagle*

# Comparación de secuencias



# Comparación de secuencias

31	V P Q H R G H V C Y	L G V C R T H R L A	E I I Y W I R C L H
30	V S I L K S G R L C S	L G T C Q T H R L P	E I I Y W I L R S A S

## Introduction

Over the past 20 years, sequence comparison has evolved from an obscure pursuit of few evolutionary biologists to a routine event that is performed 100,000's times a day by more than 10,000 different scientists in 100 different countries. This is because sequence comparison is the simplest, quickest and most inexpensive way of determining whether a newly sequenced gene or protein is in fact "new" and whether this new gene might do something interesting. By comparing a sequence to others that have already been painstakingly characterized, it is possible to infer not only functional and structural similarity, but also detailed phylogenetic relationships -- simply on the basis of sequence similarity alone. In many respects, sequence searching and the assessment of sequence similarity lies at the heart of bioinformatics.



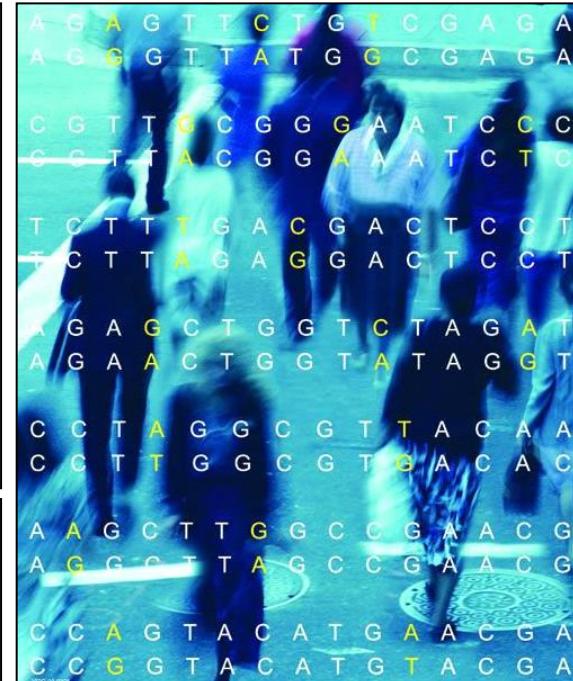
La comparación de secuencias es uno de los pilares de la Bioinformática

# Comparación de secuencias

## Why we need to compare sequences?

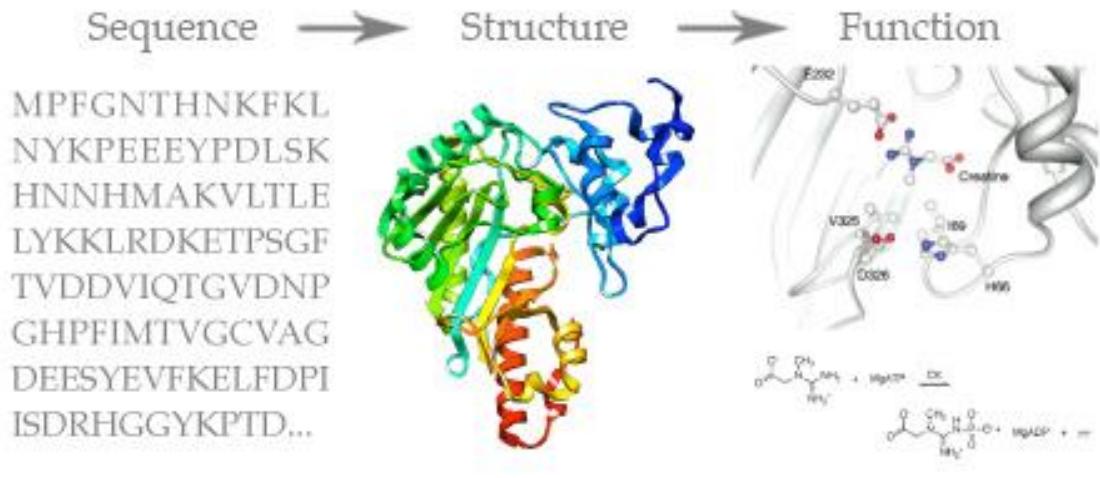
- ✓ Genome is already sequenced
- ✓ There are methods that predict DNA coding regions (genes)
- ✓ We can find out what protein (sequence) a gene encodes
- ✓ But we still do not know what this protein does...
- ✓ However we can search for known proteins with **similar** sequences and known functions

- Use sequence similarity to infer homology and/or structural similarity between 2 or more genes/proteins
- Identify more conserved regions of a protein, potentially identifying regions of most functional importance
- Compare and contrast homologs (perhaps into groups) based on shared positions or regions
- Infer evolutionary distance from sequence dissimilarity



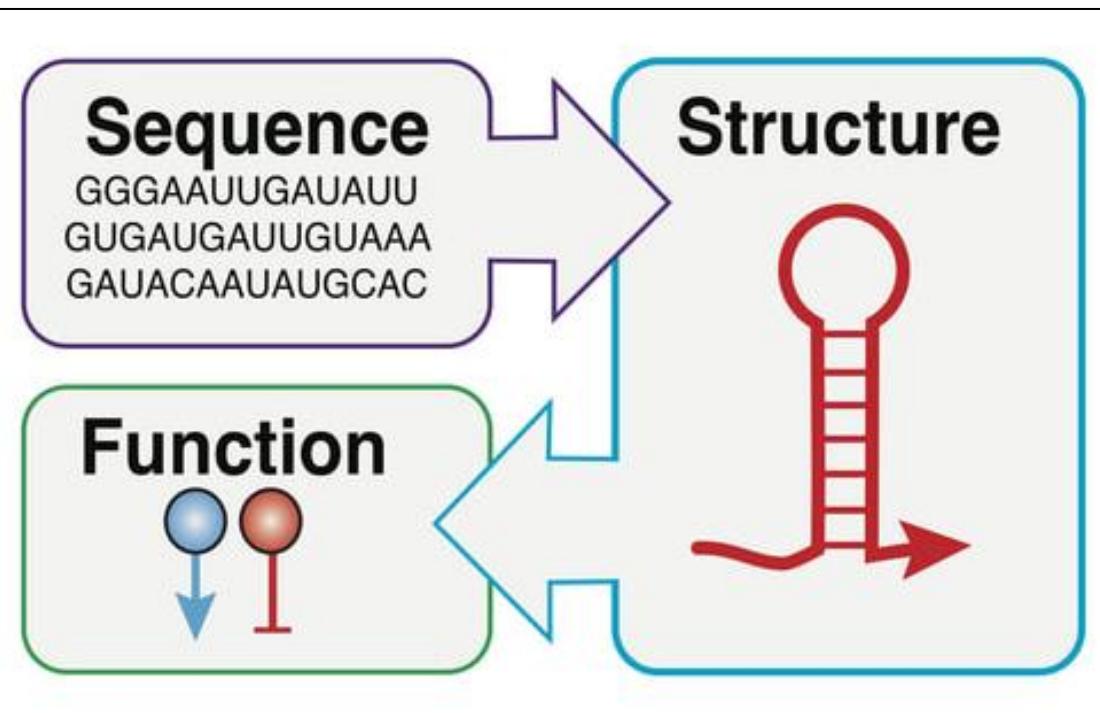
¿Para qué se comparan las secuencias?

# Comparación de secuencias



Las secuencia del ADN determina la secuencia de una proteína.

La secuencia de una proteína determina su estructura 3D.



La estructura 3D de una proteína determina su función biológica.

Por tanto, es muy probable que secuencias similares den lugar a proteínas con estructura y función parecidas.

Secuencia → Estructura → Función

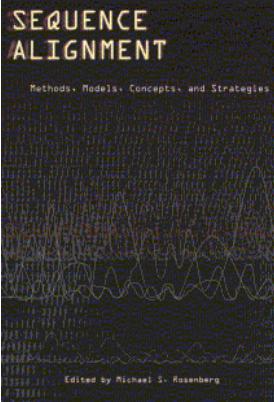
# Comparación de secuencias

¿Que es un alineamiento?

¿Para que lo hago?

- El alineamiento es una manera/metodo de comparación de dos o más secuencias (de DNA o proteínas). El procedimiento de comparación de dos (o más) secuencias que **busca una serie de caracteres individuales** o patrones de caracteres **que se encuentren en el mismo orden en ambas secuencias**
- Es probablemente la herramienta más utilizada en bioinformática
- Su objetivo es utilizar el concepto de culpa por asociacion y por ende “asignar” funciones (y otras caracteristicas) tentativas a “secuencias” desconocidas

# Comparación de secuencias



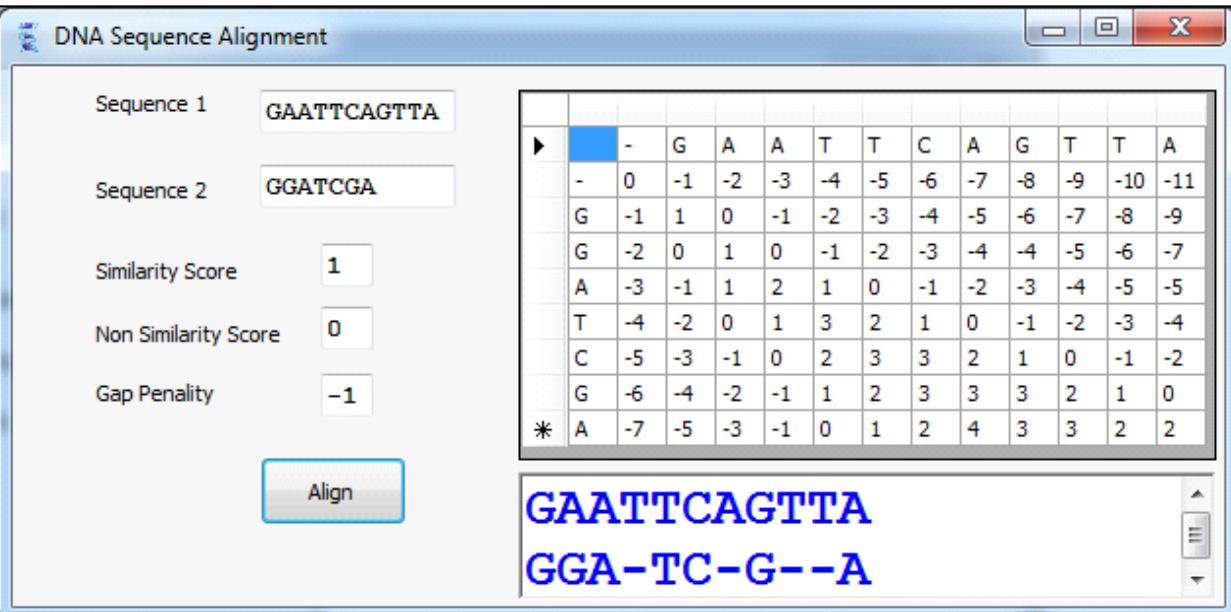
Sequence alignment is a fundamental procedure (implicitly or explicitly) conducted in any biological study that compares two or more biological sequences (whether DNA, RNA, or protein). It is the procedure by which one attempts to infer which positions (sites) within sequences are homologous, that is, which sites share a common evolutionary history.

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

Similarity between sequences:

- ✓ may indicate their common ancestral origin
- ✓ may indicate similarity of biological functions
- ✓ may indicate similarity of biological structures

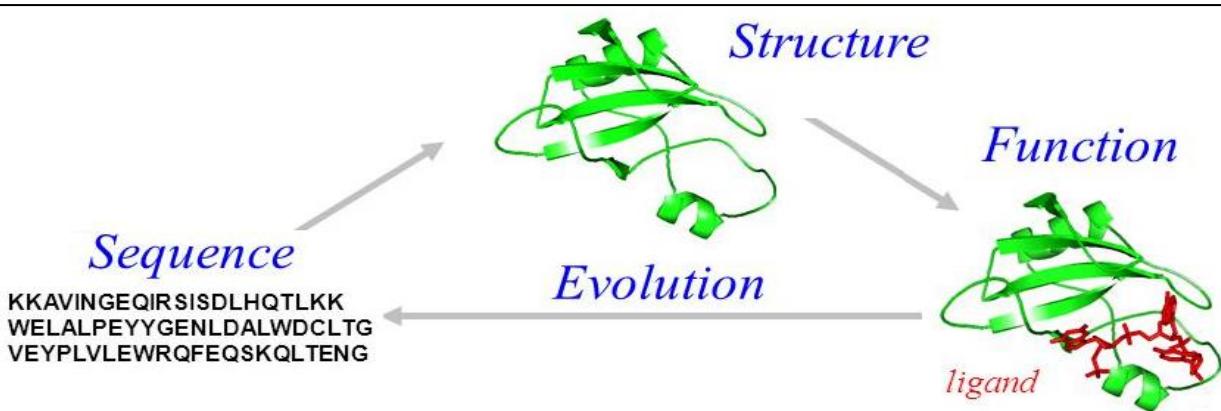
# Comparación de secuencias



El alineamiento de secuencias es la técnica que permite establecer el grado de similitud que hay entre ellas.

Cuando el grado de similitud entre dos o más secuencias es elevado, existe una probabilidad muy alta de que se trate de secuencias homólogas.

El alineamiento de secuencias es una herramienta básica de la bioinformática porque permite obtener información funcional, estructural y evolutiva



- Similar sequence leads to **similar structure**
- Similar structure leads to **similar function**

## Alineamiento de secuencias (2)

# Comparación de secuencias

## Definiciones básicas:

- **Secuencia:** conjunto de letras ordenadas seleccionadas de un alfabeto
- **Complejidad del alfabeto:** es el numero de caracteres posibles, ADN(4), Proteínas(20), otros
- **Identidad:** % de igualdad de caracteres en una comparacion de secuencias
- **Similitud:** medida del parecido entre dos caracteres de un alfabeto
- **Homología:** caracteristica biologica que implica un antecesor comun
- **Algoritmo:** es un conjunto de pasos sucesivos de instrucciones o reglas bien definidas, ordenadas y finitas que permite realizar una actividad
- **Programa:** es la implementación computacional de un algoritmo

# Comparación de secuencias

## Mas definiciones...

- **Alineamiento:** es el procedimiento consistente en comparar dos (“*pairwise*”) o más (“*multiple*”) secuencias buscando los caracteres o patrones que aparezcan en el mismo orden en las secuencias
- **Match:** coincidencia de caracteres (en un alineamiento)
- **Mismatch:** no coincidencia de caracteres
- **Gap:** carácter alineado contra “nada”
- Podemos distinguir entre **alineamientos**
  - **Globales:** Alineamiento de secuencias completas
  - **Locales :** Alineamiento de subsecuencias

# Comparación de secuencias

## 1.- Alineamiento global

Un alineamiento global se extiende por toda la longitud de la secuencia

```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC  
| | | | | | | | | | | | | | | | | | | | | | | | | |  
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
```

Homología

## 2.- Alineamiento local

Un alineamiento local se limita a una región concreta de la secuencia

```
gagcatgcagagactcccAGTTATGTCAGggacacgagcatgca  
| | | | | | | | | | | | | | | | | | | | | | | | | |  
gccgccgtcggtttcagCAGTTATGTCAGatcgccgccgtcggtt
```

Motivos conservados

## 3.- Alineamiento semiglobal

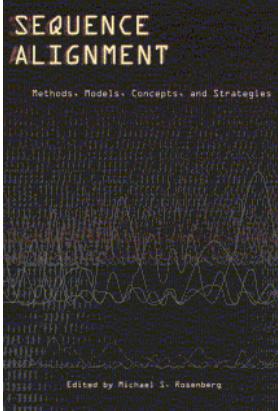
Un alineamiento semiglobal se produce entre el final de una secuencia y el inicio de otra

```
CAGTTATGTCAGggacacgagcatgca  
| | | | | | | | | | | | | | | | | | | | | | | | | |  
gccgccgtcggtttcagCAGTTATGTCAG
```

Ensamblaje de secuencias

## Tipos de alineamiento de dos secuencias

# Comparación de secuencias



The biological goal of alignment is the inference of site *homology*. Homology is similarity in a character or trait due to inheritance from a common ancestor. With respect to comparing biological sequences, homology can have three different interpretations: (1) The sequences can be homologous; (2) the sites within homologous sequences can be homologous; (3) the observed characters at a homologous site can be homologous. Sequence alignment (as discussed in this book) is mostly concerned with the second of these. The general purpose of alignment is to identify positions in homologous sequences that are descended from a common ancestral sequence, that is, to identify which sites in a pair (or more) of sequences are themselves homologous. A pair of sites is “homologous” if the position in both sequences corresponds to the identical position in the common ancestral sequence. A pair of sites is “identical” if both sequences contain the same nucleotide (or amino acid for protein sequences); identity could be due to homology (i.e., the specific nucleotide was inherited by both sequences from the common ancestral sequence with no substitutions) but may often be due to convergent or parallel substitutions or by misalignment

## El objetivo de un alineamiento de secuencias

# Comparación de secuencias

However, before going any further, it is important to develop two important definitions concerning sequence relatedness:



1) **Similarity** - in sequence analysis, this refers to the likeness or percent identity between any two sequences. Sequences are similar if they share a statistically significant number of amino acids in approximately the same positions. Similarity does not infer homology, it is only a descriptive term that carries no suggestion of shared ancestry or origin.



2) **Homology** - in sequence analysis, this refers to a shared ancestry. Two sequences are homologous if they are derived from a common ancestral sequence or if one of the two sequences has diverged (through evolution) to be different in its amino acid sequence from its parent. Homology usually implies similarity.

It is common for many biologists, biochemists and drug researchers to confuse similarity and homology when talking about protein or DNA sequences.

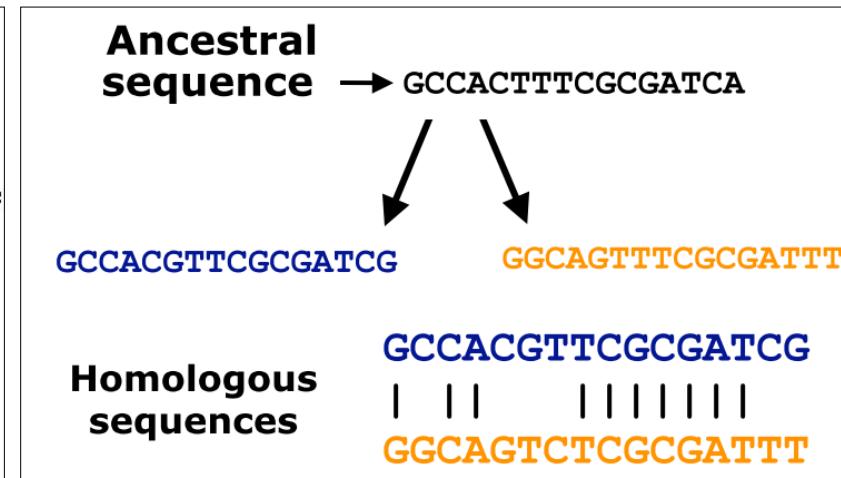
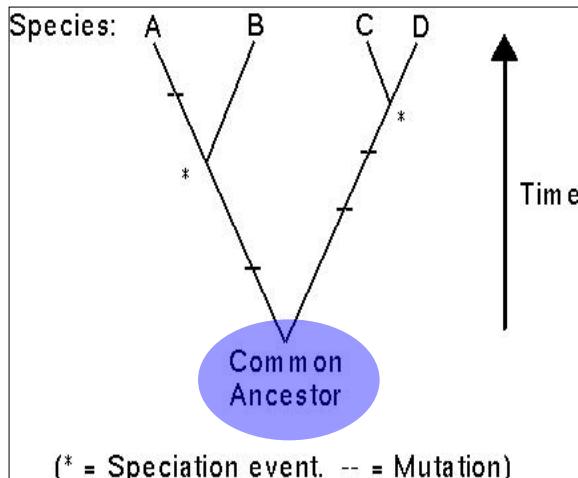
## Homología y similitud

# Comparación de secuencias

## Similarity implies homology

*The probability of two independent randomly evolving sequences converging over any but very small lengths is infinitesimally small.*

*Sequences more similar than expected from random are therefore inferred to have evolved from a common ancestor.*



**Significantly similar sequences** (such as from a BLAST search) are inferred to have come from a common ancestor

La similitud implica homología ...

# Comparación de secuencias

## Homología vs similitud

- Homología entre dos entes biológicos implica una herencia compartida
- Homología es un término cualitativo
- Se es homólogo o no se es
- Similitud implica una apreciación cuantitativa o una cuantificación directa de algún carácter
- Podemos usar una medida de similitud para **inferir** homología

# Comparación de secuencias

AAGCAGCT

| | | | | |

AGGCACCT

what can i do with sequence similarity

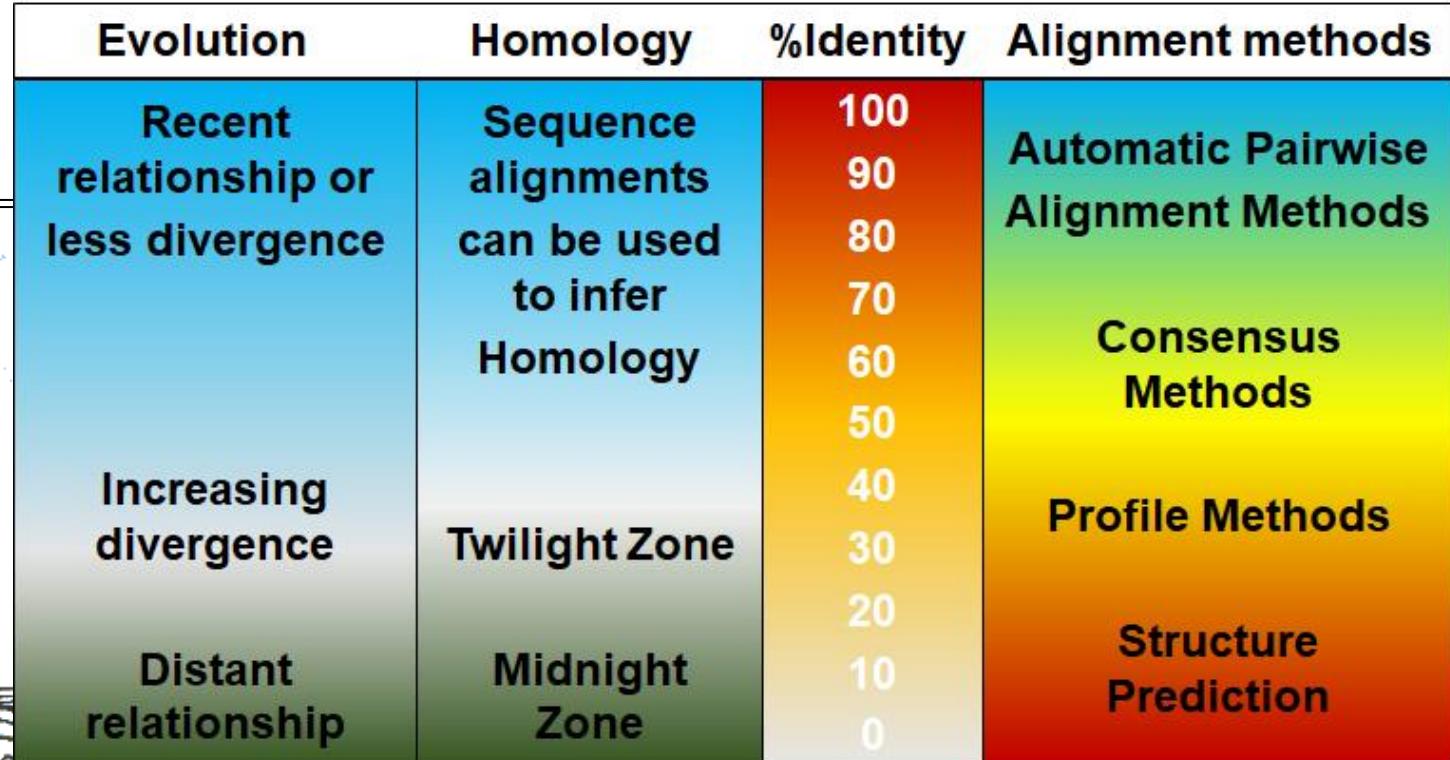
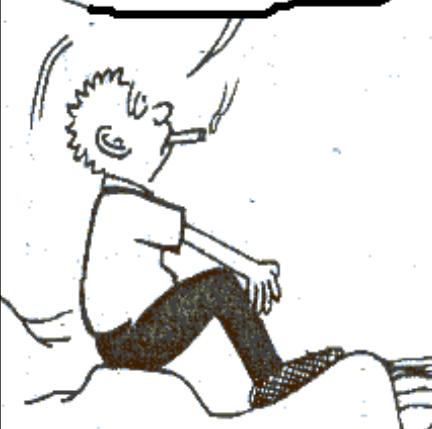
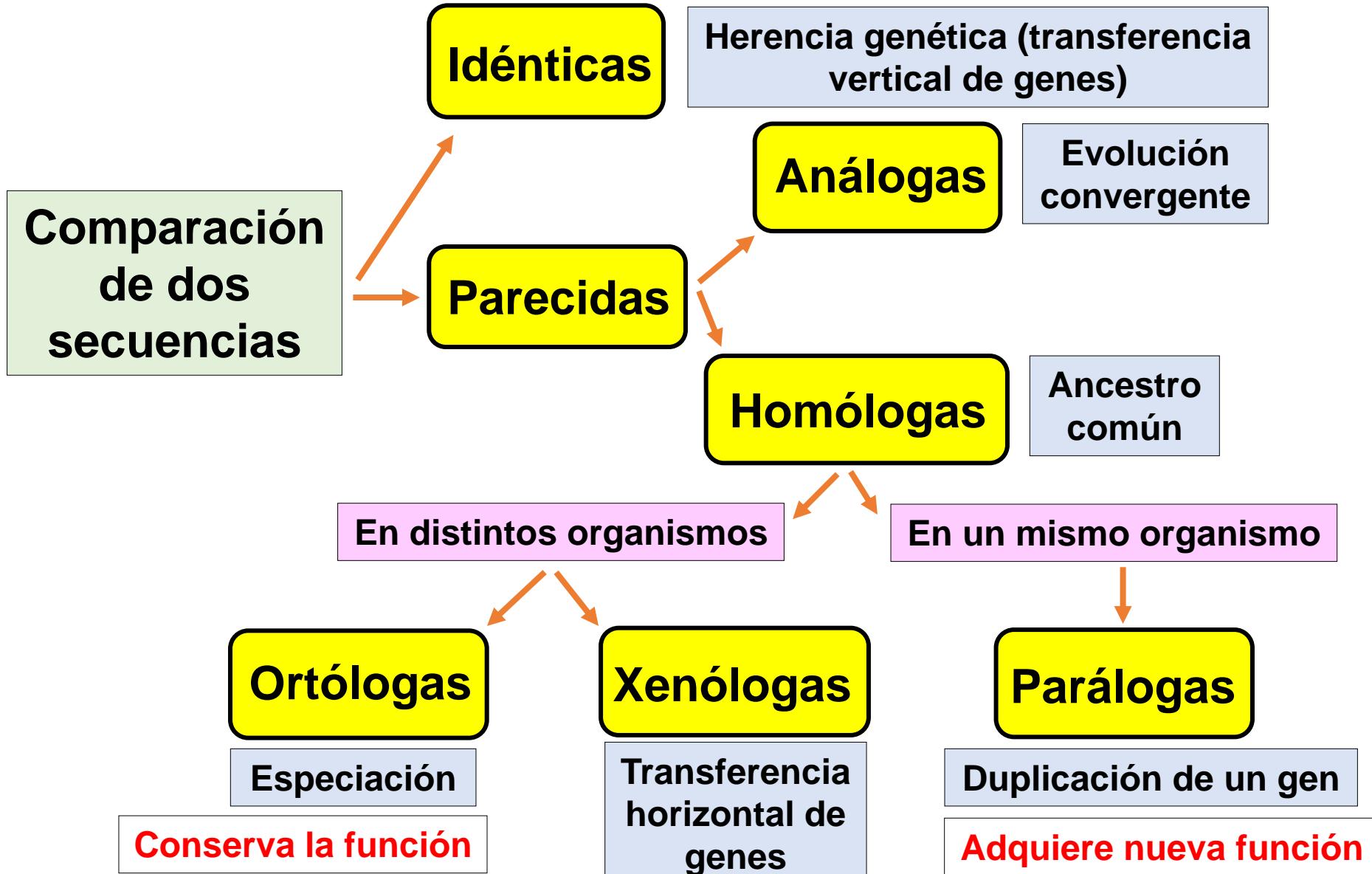


Fig 4.1 Percentage identity is an important indicator of the level of evolutionary divergence and functional/structural similarity between compared sequences. Different alignment methods have different areas of optimum application. Pairwise alignment algorithms, for example, perform well at high levels of identity, but below ~50%, the use of consensus information (from multiple alignments) may be necessary. Below ~30%, profile methods are generally used, because they allow insertions, deletions and substitutions to be modelled. Finally, at the lowest levels of identity, where alignments are no longer statistically significant, structure prediction algorithms tend to be used.

Que puedo comparar? ... pero todo tiene un límite

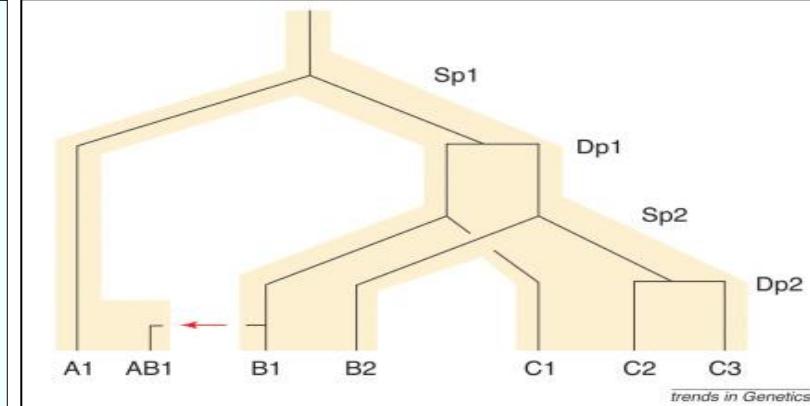
# Comparación de secuencias



Posibles causas del parecido entre dos secuencias

# Comparación de secuencias

Homólogas: secuencias que proceden de una misma secuencia ancestral y que, por tanto, presentan cierto grado de similitud.

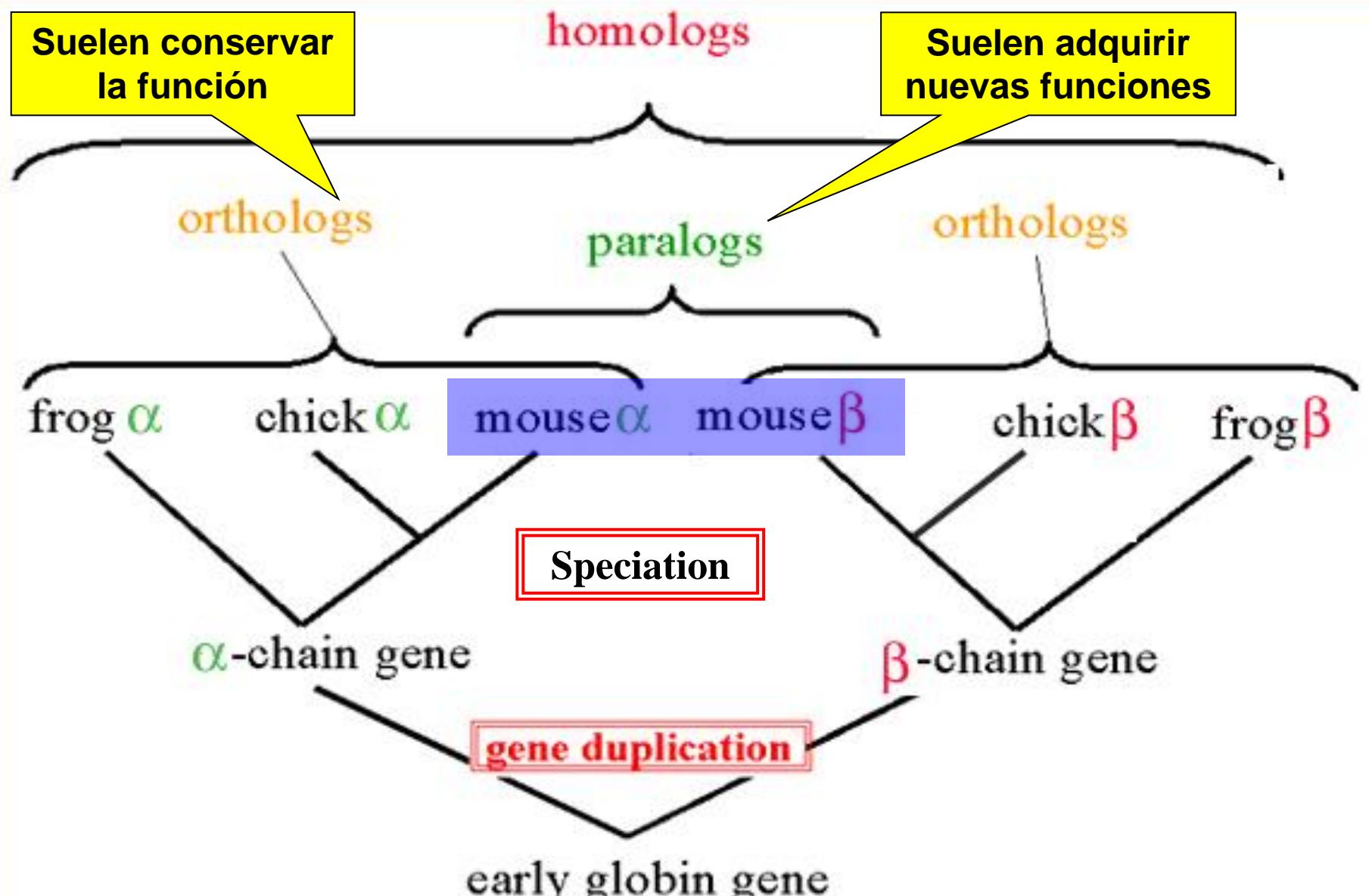


Parálogas: secuencias de un mismo organismo, que han aparecido tras un proceso de duplicación génica. Pueden adquirir distinta función.

Ortólogas: secuencias de organismos distintos, que han aparecido durante el proceso de especiación. Conservan la misma función.

Xenólogas: secuencias de organismos distintos, que han aparecido tras un proceso de transferencia horizontal de genes (virus, simbiosis, etc.)

# Comparación de secuencias



**Por ejemplo:** Ortólogos y parálogos

# Comparación de secuencias

Vamos a comparar secuencias!!!

# Comparación de secuencias

2 differences

ACCTCTG**T**ATCTATTGG**C**ATCATCAT  
ACCCCTGAATCTATTGGGATCATCAT

ACCTCTGTATCTATTGGGATCATCAT  
ACCTCTG**A**ATCTAT**C**GGGATCAT**G**AT

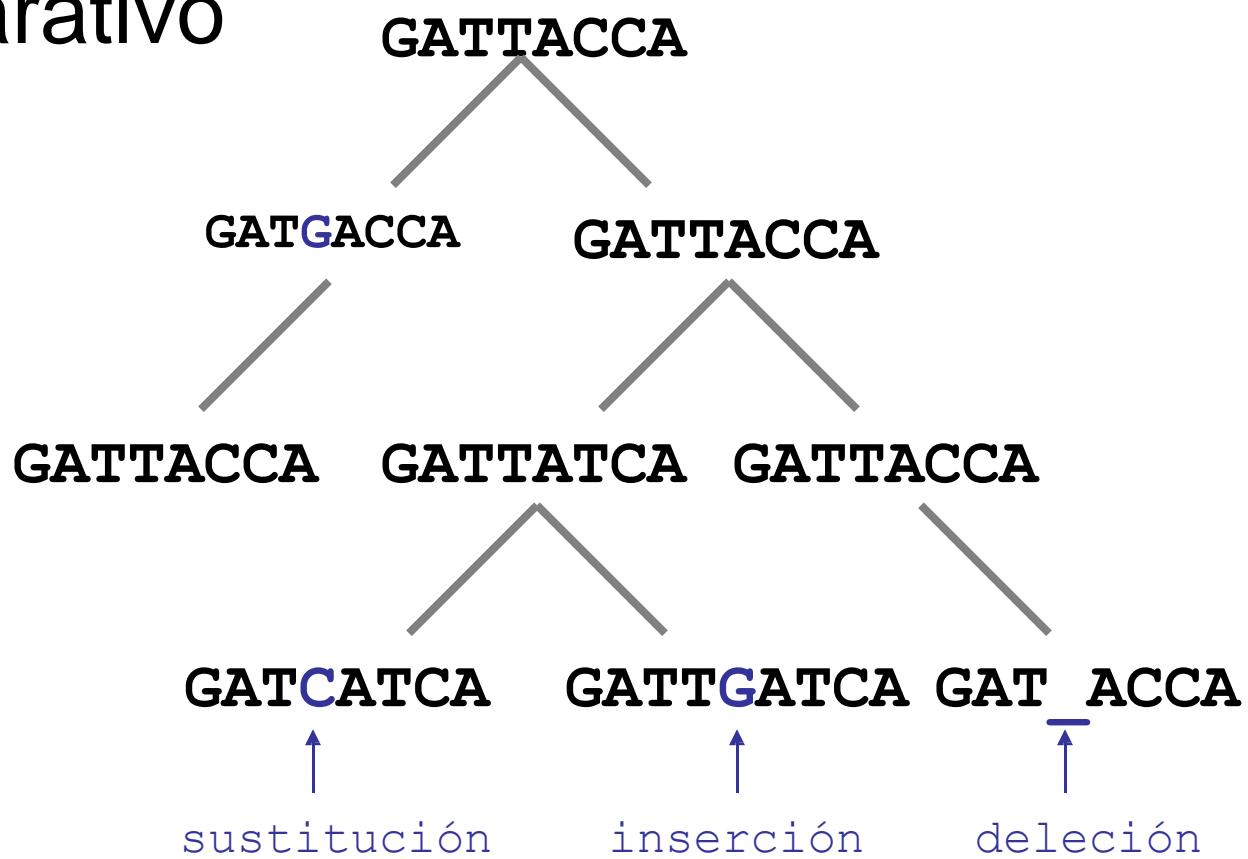
3 differences

¿Cuales secuencias son mas parecidas?

# Comparación de secuencias

Los algoritmos que alinean secuencias modelan procesos evolutivos

## Análisis comparativo



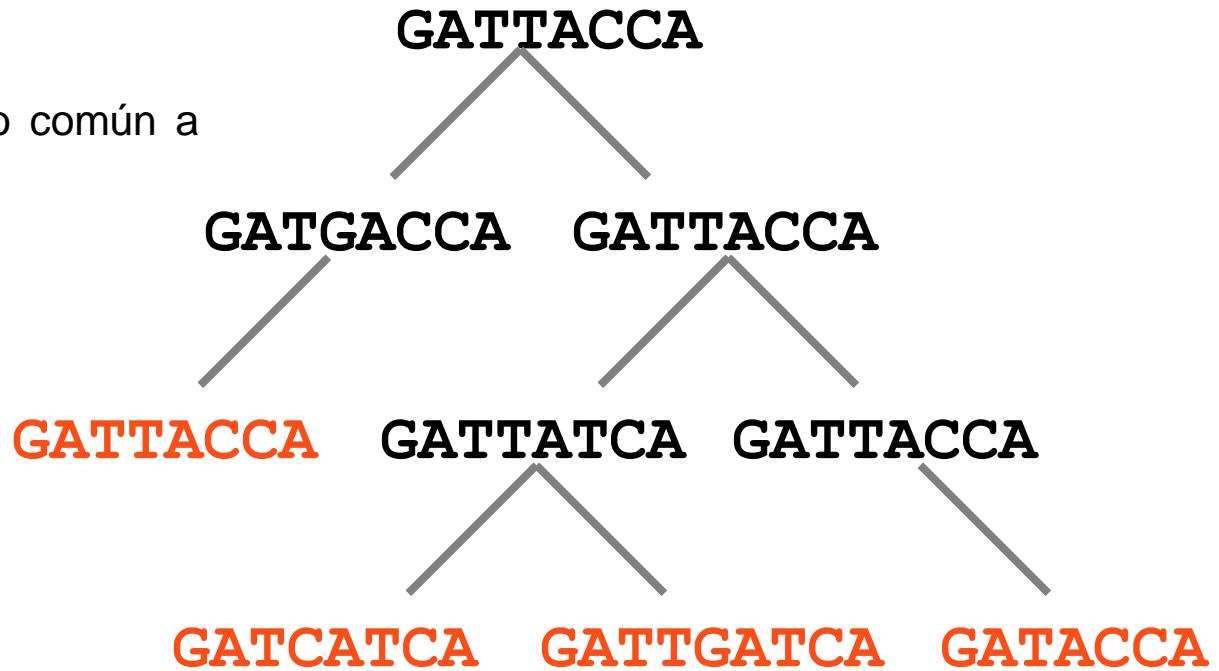
**Deriva de un ancestro común** a través de cambios incrementales debido a errores en la replicación del DNA, mutaciones, daño o crossing-over desigual.

# Comparación de secuencias

Los algoritmos que alinean secuencias modelan procesos evolutivos

## Análisis comparativo

Deriva a partir de un ancestro común a través de cambio incremental.



Sólo las secuencias actuales son conocidas, las secuencias ancestrales se postulan.

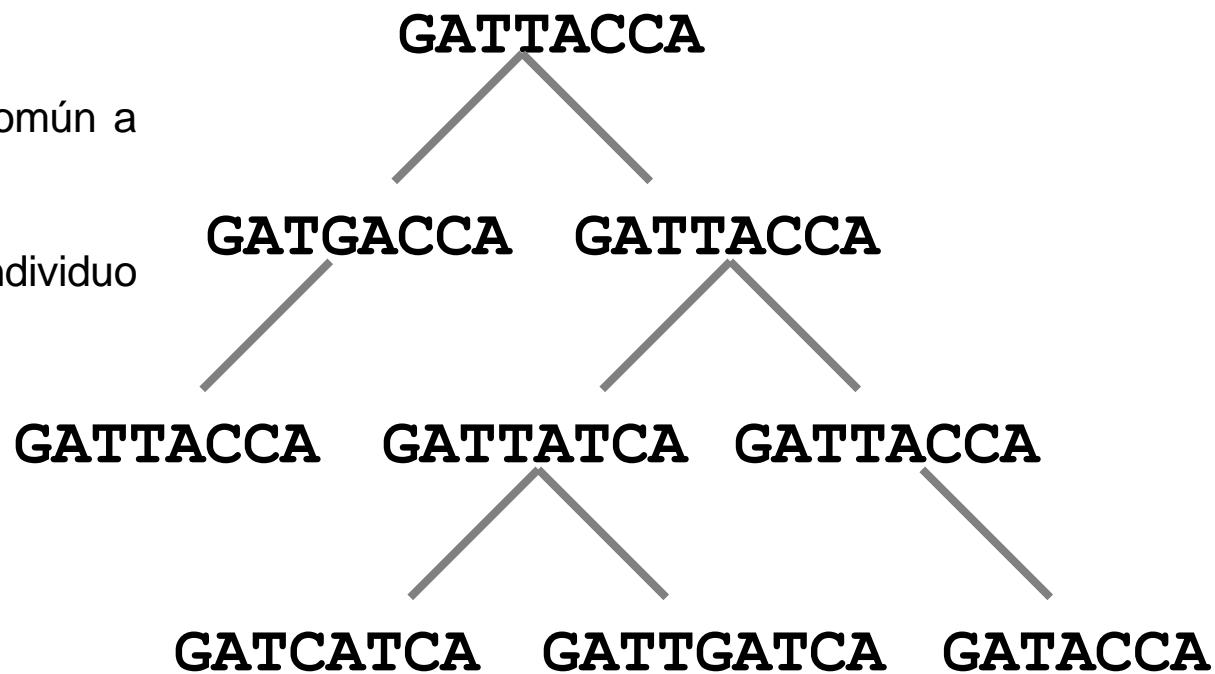
# Comparación de secuencias

Los algoritmos que alinean secuencias modelan procesos evolutivos

## Análisis comparativo

Deriva a partir de un ancestro común a través de cambio incremental.

Mutaciones que no matan al individuo pueden pasar a la población.

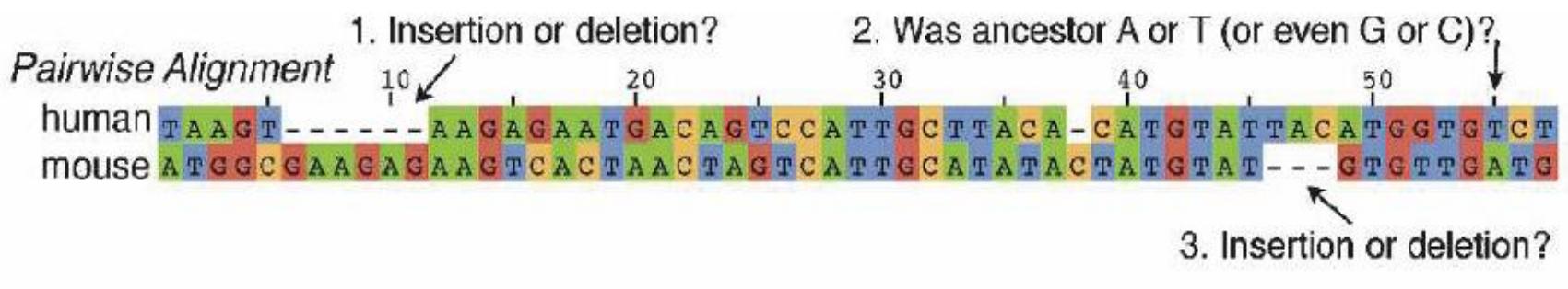


La palabra **homología** implica una herencia común (un ancestro común), el cual puede ser inferido a partir de observaciones de **similitud** de secuencia.

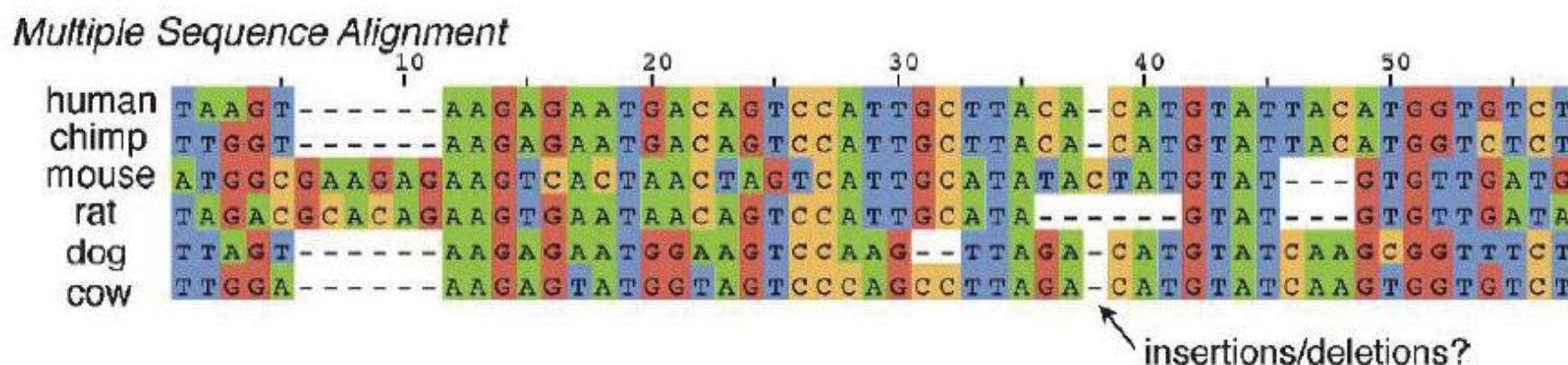
# Comparación de secuencias

Según el número de secuencias que se comparan podemos distinguir:

## 1.- Alineamiento de dos secuencias

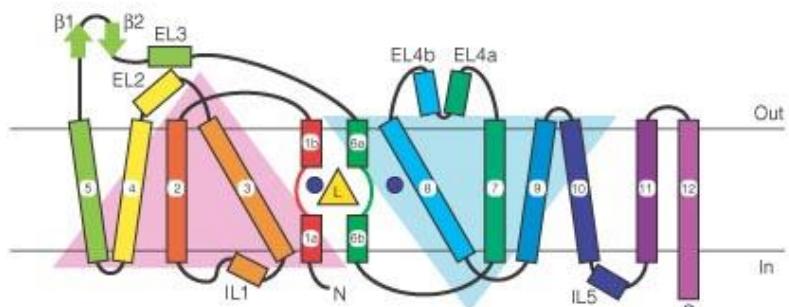
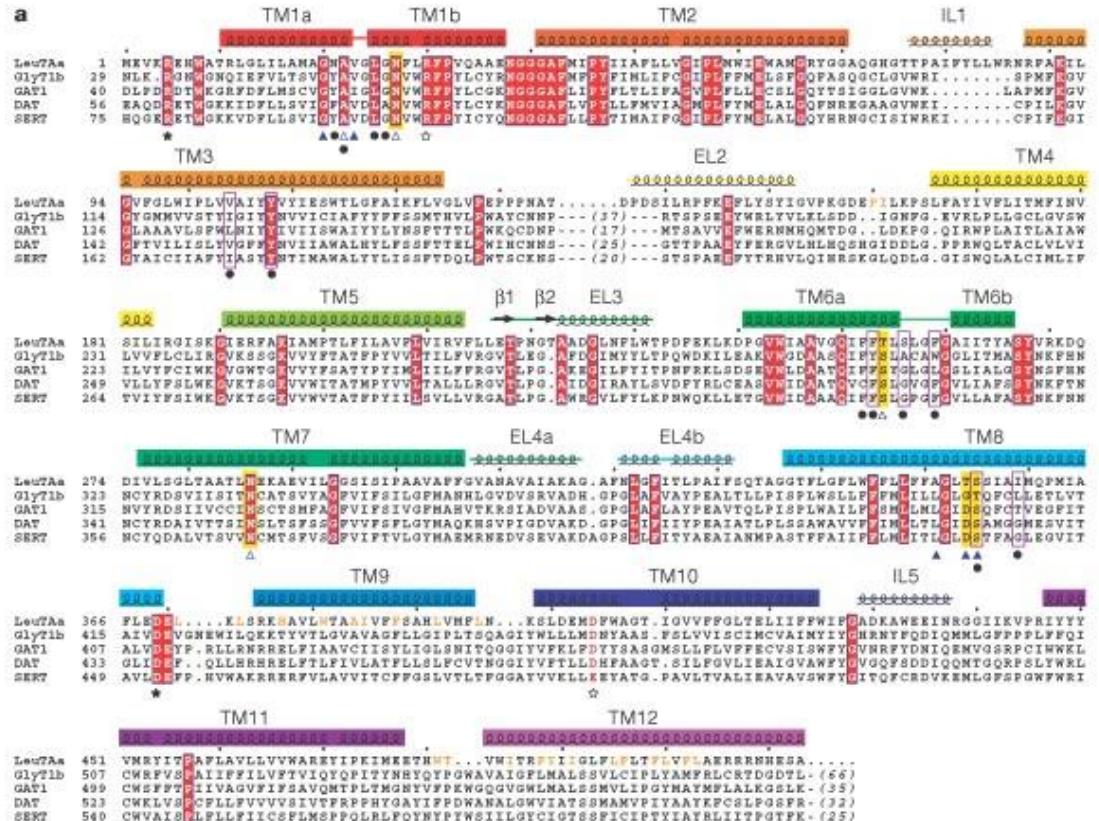


## 2.- Alineamiento múltiple de secuencias (MSA)



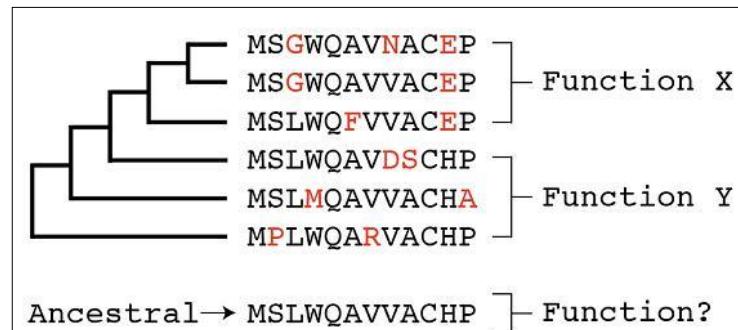
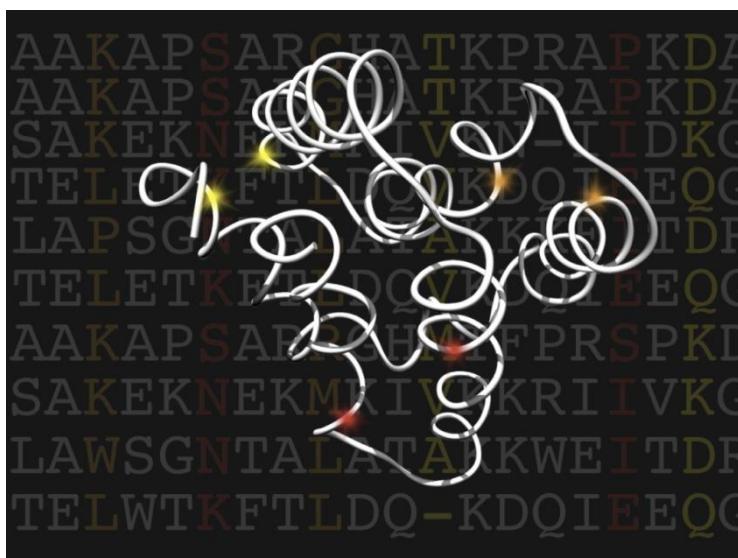
## Tipos de alineamiento

# Comparación de secuencias



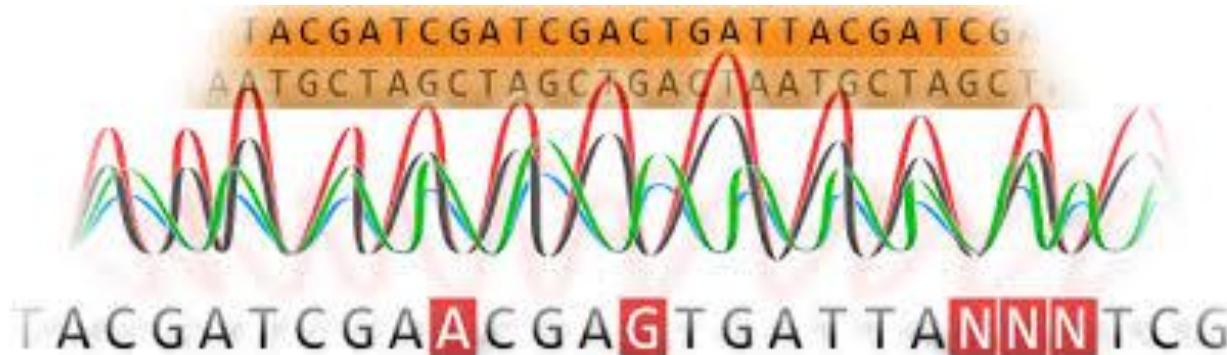
# Alineamiento múltiple de secuencias (MSA)

**Se conservan las regiones que son importantes para mantener la estructura y/o función**



# Comparación de secuencias

¿Qué creen que es mas fácil alinear o comparar?  
¿Secuencias de nucleótidos o de aminoácidos?

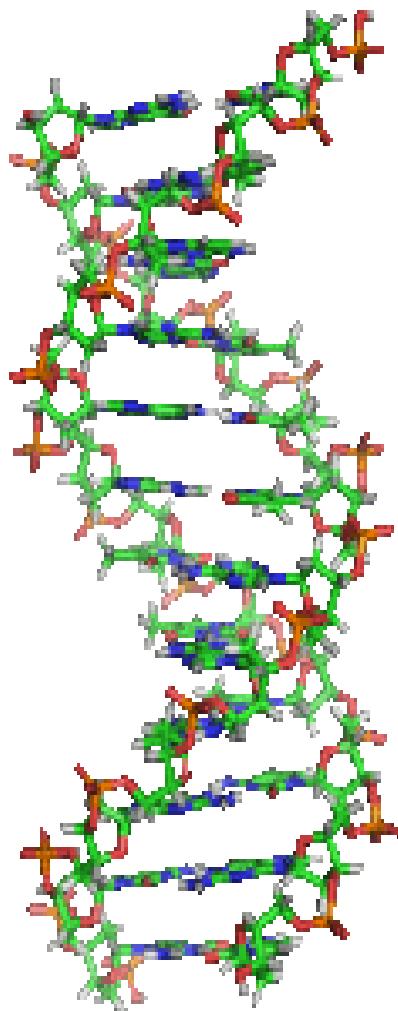


SOFL/1-63 MESANIFI GS --DEDCRS CESGWITMYLASQSHDRDD-CY-----Y-----GDDDEE EEDSDGGDSMDSDASS G P ME  
AtSOFL4/1-71 MDK -----EECSSSESGWT TYLSSP I KVDEDE-VV-DEDYYYEGYNLY-----NY SS KVEHEE ERNKDS DDSMASDASS G P NY  
AtSOFL3/1-70 MER -----EECSSSESGWT TYIISRMEEEEEE-VI -DE -VYYEGH I IEKD-----RR KYANEYE I NKDS DDSMASDASS GPS Y  
AtSOFL5/1-72 MLG -----SSSGCESGWITLYLDQS VSSSPSPS-CFRDS-NGFDSRRRSKD-----SWDQNYVHQEEEEE DDLSM I SDASS G P R N  
AtSOFL1/1-86 MESPRNHGGSEE EYSSCESGWITMYIEDA F HGNDQSSVVVDD-DDDTQVKEARO---GYE NDDGDT S DDGGDEES S DDSMASDASS G S N  
AtSOFL2/1-87 MESPRIHGG--AEEKSSCESGWITMYIEDT F HGNHHSVEYYEE- EDDGF SVKEVDDDGDE D DDDDDDDDS SNNES DDSMTSDASSWPS T  
AtSOFL6/1-58 MDFSDL -----DYS DAGD SGWITMLGHS S SVSLH--HF-D-----YHNGETKQEHD EDS S MVS DASS G P Y



# Comparación de secuencias

Son menos sensibles que los alineamientos de proteínas porque:



En las bases de datos, los **4 nucleótidos** aparecen con la misma frecuencia. Cada nucleótido aporta 2 bits de información

Todos los cambios posibles tienen una probabilidad similar.

Se basa fundamentalmente en la coincidencia directa entre los textos

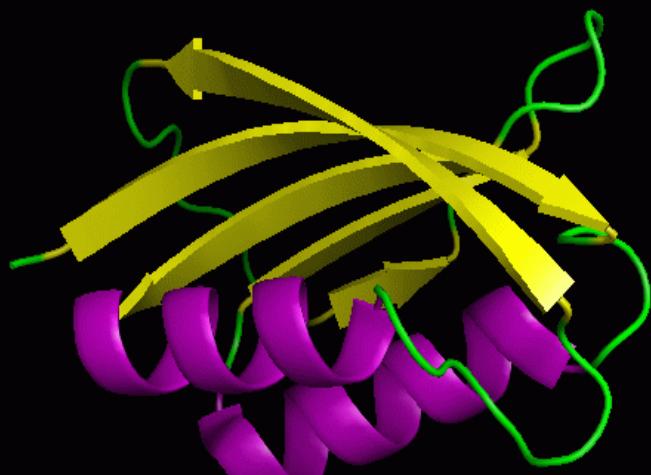
Método lento, porque las bases de datos de ácidos nucleicos contienen un número muy elevado de caracteres

Es preferible “traducir” una secuencia de ADN a 6 proteínas (los 6 ORF) y alinear las secuencias de proteínas

Si se trata de secuencias no codificantes, no queda más remedio que hacerlo

Alineamientos de secuencias de **ácidos nucleicos**

# Comparación de secuencias



1.- Aportan más información (más de 4 bits por aa).

2.- Se obtienen resultados estadísticamente significativos con alineamientos más cortos

3.- El código genético es redundante, casi 1/3 de las bases no están sometidas a presión selectiva y generan ruido, lo que afecta a la sensibilidad de la búsqueda

4.- Las búsquedas en bases de datos de ácidos nucleicos son más lentas porque son mucho más grandes a causa de los proyectos genómicos y, además, contienen muchas secuencias no codificantes.

5.- A diferencia de los nucleótidos, las probabilidades de sustituir un aa por otro son muy distintas y, con ello, la eficacia de la búsqueda aumenta notablemente.

## Alineamientos de secuencias de proteínas

# Comparación de secuencias

Consideremos estas dos secuencias:

ATGGAGCTGATCTCATCAGCGATCTCAGCGCTGATCGTCGAGTGA  
ATGGAATTAAATTAGTAGTGCTATTAGTGCTTTAATTGTTGAATAA

Hagamos un alineamiento sin huecos:

ATGGAGCTGATCTCATCAGCGATCTCAGCGCTGATCGTCGAGTGA  
||| | | | | | | | | | | | | | | | | | | | | | | | | |  
ATGGAATTAAATTAGTAGTGCTATTAGTGCTTTAATTGTTGAATAA

Hay 23 nucleótidos idénticos de un total de 45 (Un 51% de identidad)

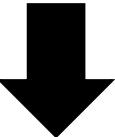
Alineamientos de 2 secuencias de ADN

# Comparación de secuencias

Secuencias de nucleótidos:

ATGGAGCTGATCTCATCAGCGATCTCAGCGCTGATCGTCGAGTGA  
ATGGAATTAAATTAGTAGTGCTATTAGTGCTTTAATTGTTGAATAA

Traducción a proteínas:



Secuencias de aminoácidos:

MELISSAISALIVE

MELISSAISALIVE

A nivel de aminoácidos, las dos secuencias son idénticas

Alineamiento de las proteínas codificadas

# Comparación de secuencias

sequence comparison is best done when the sequences (or character strings) exhibit "**complexity**", meaning that they are composed of a large number of different characters. In this regard, protein sequences (unless they are highly repetitive or only have a small number of amino acids) with their 20 letter alphabet are far more complex or informatically "richer" than DNA sequences which only have a 4 letter alphabet. Consequently, string comparisons between two DNA sequences are more likely to yield ambiguous results or potentially false alignments than string comparisons conducted against protein sequences. It is for this reason that most experienced bioinformaticians insist that DNA sequences be translated to protein sequences before they are submitted for pairwise alignment or database comparisons (Doolittle, 1986). Indeed, the only reason why one would not want to translate a DNA sequence into the corresponding protein sequence is if the DNA does not code for any protein product or if it corresponds to a tRNA or rRNA gene.

So **ALWAYS TRANSLATE THOSE DNA SEQUENCES!** 

**La complejidad no es mala**

# Métodos de alineamientos y algoritmos

**Existen diversos métodos para el alineamiento de dos secuencias:**

**1.- El algoritmo de la fuerza bruta**

**2.- Matrices de puntos (*dot plots*)**

**3.- El algoritmo de programación dinámica**

**4.- Métodos heurísticos (FASTA, BLAST)**

Usan “palabras” para anclar dos secuencias

**Estrategias para alinear dos secuencias**

# Métodos de alineamientos

## A mano:

se mueven las secuencias a mano y se observan... (no muy practico!)

## Dot plot (Método gráfico)

se grafica en una matriz de a “ventanas”

## Dynamic programming

Smith-Waterman  
Needle-Wunsch



(lento, óptimo, garantiza el mejor posible alineamiento)

## Heuristic methods

(rápido, se aproxima a un buen resultado. No lo garantiza)

BLAST and FASTA: Usan “palabras” para anclar dos secuencias!

# Métodos de alineamientos y algoritmos

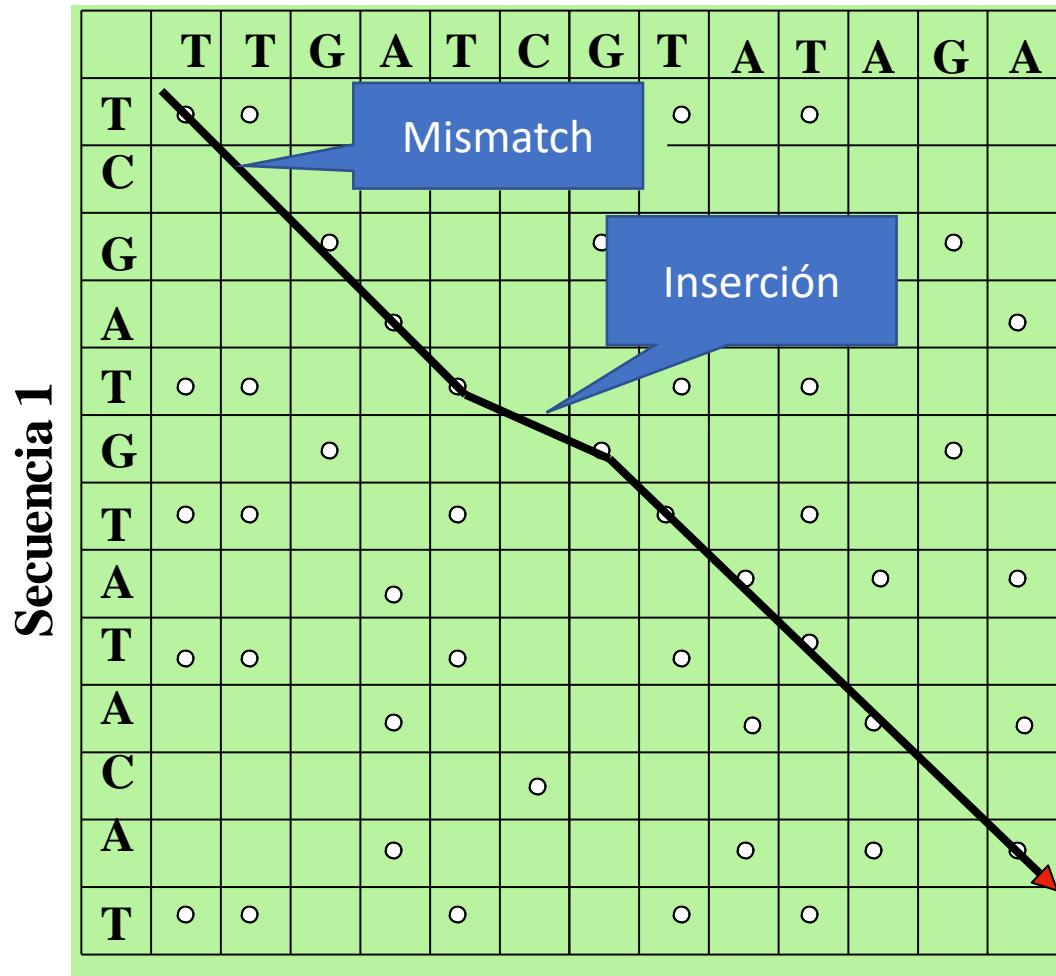
## Métodos gráficos de alineamiento

### Dot plot

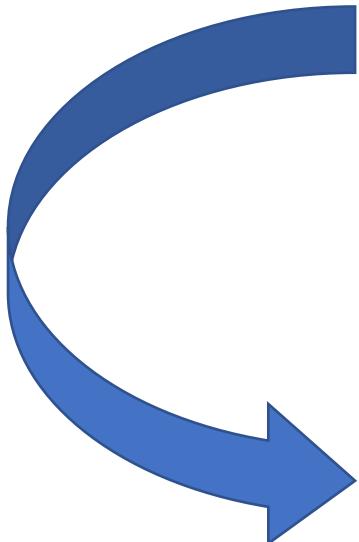
- Dot plot es una representación gráfica de similaridad entre dos secuencias
- Está compuesta por una matriz, cuyos ejes se forman con las dos secuencias que se desean alinear.
- Luego se llena con un punto la interjección donde coinciden los caracteres.
- Por última se traza una línea uniendo las diagonales formadas por las regiones similares.

# Métodos de alineamientos y algoritmos

## **Secuencia 2**



# Dot plot



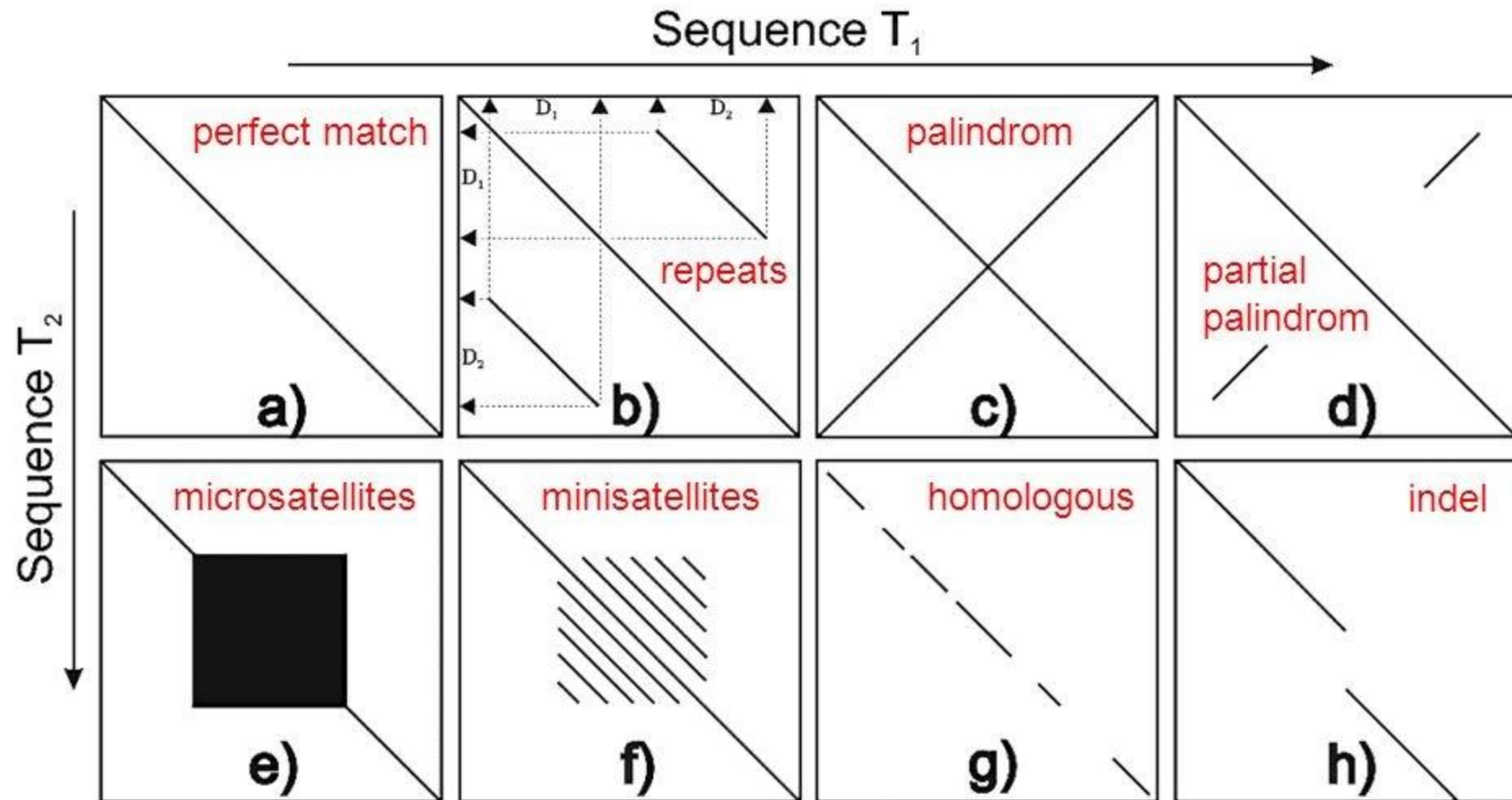
## Secuencia 1 TCGAT-GTATAACAT

— — — — — — — — — — — —

## Secuencia 2 TTGATCGTATAGA-

# Métodos de alineamientos y algoritmos

## Interpretation of dot plot – summary



# Métodos de alineamientos y algoritmos

## Analysis of dot plot matrix

- Region of similarity appears as diagonal run of dots.
- Principal diagonal shows identical sequence.
- Displacements in the main diagonal shows indels
- Global and local alignment are shown.
- Multiple diagonal indicate repetition
- Reverse diagonal (perpendicular to diagonal) indicate INVERSION.
- Reverse diagonal crossing diagonal (X) indicate PALINDROMES.
- Formation of box indicate the low complexity region.

# Métodos de alineamientos y algoritmos

## Umbral de severidad (“Stringency threshold”)

- Para facilitar la visualización, se opta a menudo por mostrar únicamente las diagonales formadas por un número mínimo de puntos (**umbral de severidad**)
- Si el umbral de severidad es alto:
  - Eliminamos el ruido de fondo (“filtrado alto”)
  - Solo detecta similitudes muy altas
- Si es bajo:
  - Hay ruido de fondo
  - Detecta relaciones distantes

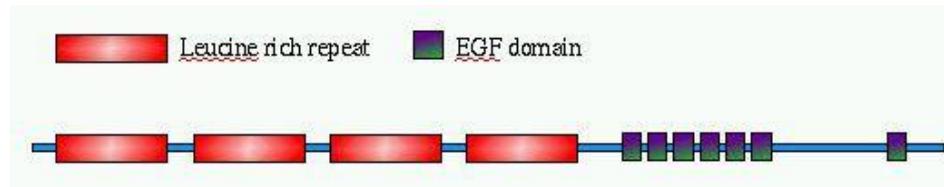
# Métodos de alineamientos y algoritmos

## Parámetros del filtro/umbral:

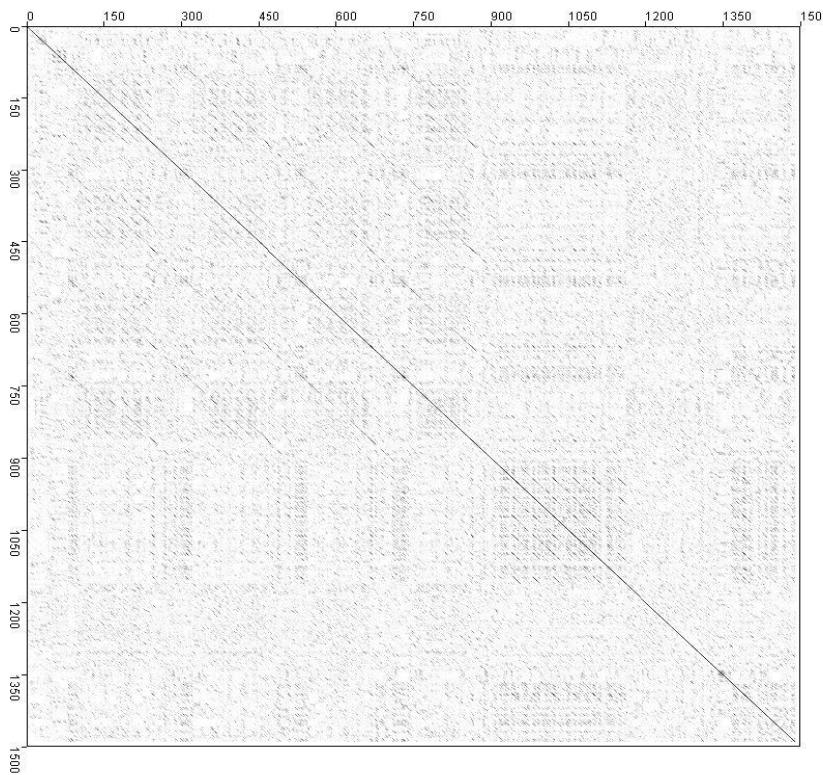
- **Window Size:** Es el número de bases que generan un punto en la comparación.
  - Por ejemplo puedo decidir que 9 bases me generan un punto.
- **Mismatch Limit:** Determina que tan similares dos secuencias en una ventana (window size) tienen que ser para considerarlas un match.
  - Ventana 9, mismatch limit 2, entonces si hay hasta dos letras que no son idénticas igual lo considero un match.

# Métodos de alineamientos y algoritmos

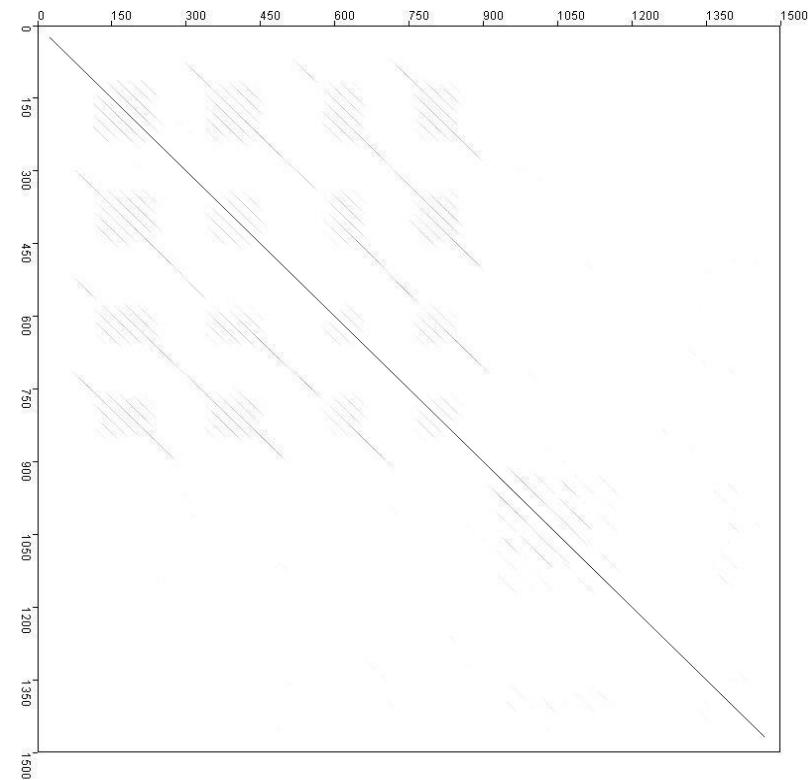
## Ejemplo:



Alin consigo misma



Umbral bajo



Umbral alto

# Métodos de alineamientos y algoritmos

## Dot-Plot

### Ventajas

- Es una forma rápida y gráfica para encontrar regiones de apareamiento entre dos secuencias.
- Es útil para encontrar regiones repetidas e invertidas. Es útil como primer paso antes de aplicar algoritmos de programación dinámica.

### Desventajas

- A veces no es fácil encontrar el mejor apareamiento de forma objetiva.
- No es cuantitativo es cualitativo.
- Cuando se analiza una secuencia con una base de datos de secuencias, ¿de que forma encontrar las secuencias que mayor similitud tengan? y en tal caso ¿como poder inferir una homología?

# Métodos de alineamientos y algoritmos

## Exhaustive alignment: brute force

- Having the scoring scheme we can proceed to generate and evaluate alignments:
- Brute force:** Generate the list of all possible alignments between two sequences, score them and select the alignment with the best score.
- The number of possible global alignments between two sequences of length  $L$  is  $2^{2L} / (\pi L)^{1/2}$ . For two sequences of 250 bases this is  $\sim 10^{149}$ .
- Practically useless...


$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

**A lo bestia: el algoritmo de fuerza bruta**

# Métodos de alineamientos

Trata de encontrar la secuencia común de mayor tamaño entre dos secuencias X e Y de longitudes  $m$  y  $n$ , respectivamente.

Se determinan todas las subsecuencias posibles de X ( $2^m$ ) y se comparan con todas las subsecuencias posibles de Y ( $2^n$ )

**En total, hay que hacer  $4^{(m+n)}$  comparaciones**

Con gaps, hay que repetir los cálculos  $2N$  veces para examinar la presencia de gaps en todas las posiciones posibles de las dos secuencias

Según Waterman (1989) comparar dos secuencias de 300 aminoácidos requiere examinar  $10^{88}$  posibilidades, casi el mismo número de partículas elementales que hay en el Universo.

En la práctica, resulta imposible, tanto por el tiempo que se necesita como por los recursos de memoria que le harían falta al ordenador

Con alineamientos locales es aún peor

# Métodos de alineamientos y algoritmos

groan	vermiform-----
:	
grown	-----formation

elephant	vermiform
:	::   :::::
eleg-ant	formation

colo-r	disestablishment
	:
colour	dis-----sent

disestablishment	theatre
	::
dis-----s--ent	theater

Cualquier pareja de secuencias se puede alinear. Basta con introducir un elevado número de huecos en el alineamiento.

El objetivo de un alineamiento consiste en colocar una secuencia encima de la otra de modo que el número de caracteres idénticos o similares sea máximo, introduciendo el mínimo número de huecos.

Los huecos también se denominan indels o gaps, ya que pueden ser inserciones en una secuencia o delecciones en la otra. Dos indels no pueden estar alineados.

	Match		indels	
Sequence 1	A	C	T	- G
Sequence 2	A	T	T	C -

Mismatch      Gap      Gap

No todos los alineamientos son iguales

# Métodos de alineamientos

- Ahora, alineamos por ejemplo:

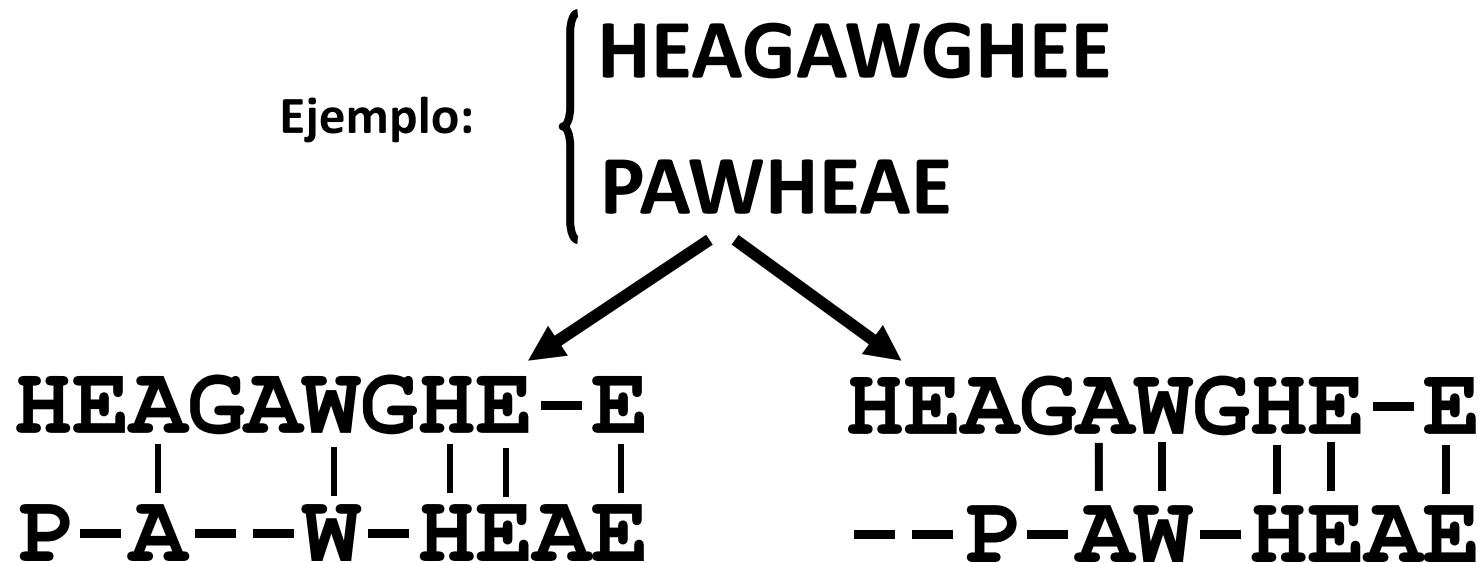
- VELIIQDIVLGGA
- VDIVLQDLGGA

- Possible Resultado:
  1. VELIIQ**DIVL**---GGA
  2. V-----**DIVL**QDLGGA
- Alternativamente:
  1. **VELIIQDIVL**GGA
  2. **VDIVLQDL**--GGA

Cual elijo? Por que?

# Métodos de alineamientos y algoritmos

Dos secuencias se pueden alinear de muchas formas distintas



¿Cuál es mejor?

Para determinar cuál es el mejor alineamiento se necesita un sistema de puntuación.

El alineamiento que obtenga la puntuación más elevada se denomina alineamiento óptimo

La necesidad de disponer de un sistema de puntuación

# Métodos de alineamientos y algoritmos

What are we comparing?

- **DNA or RNA**
  - Four nucleic acids (basic set)
- **Protein**
  - Twenty amino acids (basic set)



# Métodos de alineamientos y algoritmos

- Differences in the sequence can be caused by deletions or insertions in the DNA, or by point mutations
- These changes can be seen at the protein level as well (changes in the translation of the protein)

# Métodos de alineamientos y algoritmos

This scheme works fine as long as **you assume that all possible mutations** occur at the same frequency. However, nature doesn't work this way. It has been found that in DNA, transitions occur more often than transversions.

# Métodos de alineamientos y algoritmos

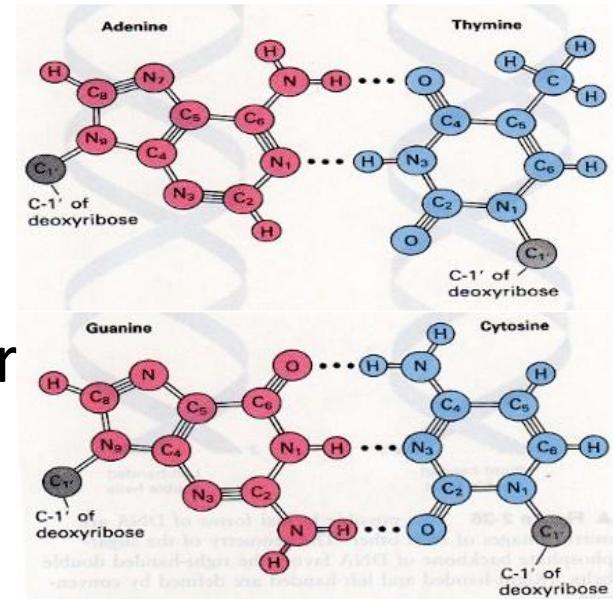
For DNA:

Purines (A,G) are 2-ring bases

Pyrimidines (C,T) are 1-ring bases

**Transition:** purine to purine or  
pyrimidine to pyrimidine

**Transversion:** purine to pyrimidine or  
pyrimidine to purine



Transitions conserve ring number Transversions change  
ring number



# Métodos de alineamientos y algoritmos

For proteins, the situation is far more complex

- Amino acids can be grouped by a number of classifications:
  - **Chemical:** aromatic, aliphatic, sulphuric
  - **Functional:** hydrophobic, hydrophilic, acidic, basic
  - **Charge:** positive, negative, neutral
  - **Structural:** internal, external

# Métodos de alineamientos y algoritmos

## SEQUENCE ALIGNMENT

Methods, Models, Concepts, and Strategies

Edited by Michael S. Rosenberg



To be able to compare potential sequence alignments, one needs to be able to determine a value (or score) that estimates the quality of each alignment. The formulas behind an alignment score are generally known as *objective functions*; they range from simple cost-benefit sums to complex maximum-likelihood values. This introduction will use a simple cost-benefit approach, but much more advanced scoring algorithms and mechanisms are available, many of which are discussed in detail throughout this book.

When using a cost-benefit approach to evaluating a pairwise alignment, one must specify scores for the various ways in which a pair of sites can be compared. In the simplest case, three scores are specified: (1) the benefit of aligning a pair of sites that contain the same character (state) in both sequences; (2) the cost of aligning a pair of sites that contain different characters in the sequences; and (3) the cost of aligning a character in one sequence with a gap in the other sequence. Depending on how one defines the scores, the eventual goal could be to find the alignment that maximizes the benefit or to find the alignment that minimizes the cost. This is essentially an arbitrary choice based on how one chooses to define the costs and benefits. Many of the popular alignment programs in use today (e.g., ClustalW) find the maximum score.



## Sistema de puntuación = función objetiva

# Métodos de alineamientos y algoritmos

## Consideraciones para el sistema de puntuación

Para saber cuál es el mejor alineamiento entre dos secuencias es necesario establecer un sistema de puntuación.

El sistema de puntuación consta de dos componentes: (1) una matriz de puntuación que asigna un valor a cada una de las posibles coincidencias y sustituciones y (2) una penalización por la introducción de indels.

La puntuación del alineamiento resulta de sumar la puntuación de cada posición, en función de que los residuos coincidan (match), sean distintos (mismatch) o haya un hueco (indel).

Cada uno de los posibles alineamientos recibe una puntuación. Se considera alineamiento óptimo aquél que consigue la puntuación más elevada.

# Métodos de alineamientos y algoritmos

El sistema más sencillo consiste en otorgar una puntuación discreta a las coincidencias (*match*), otra a las diferencias (*mismatch*) y otra a los huecos (*gaps*).

- Match score: +1
- Mismatch score: 0
- Gap penalty: -1

Sistema de puntuación

ACGTCTGATA**A**CGCCGTAT**A**GTCTATCT  
||| | | | | | | | | | | | | | | | | |  
----CTGATT**T**CGC---AT**C**GTCTATCT

- Matches:  $18 \times (+1)$
- Mismatches:  $2 \times 0$
- Gaps:  $7 \times (-1)$

Score = +11

El sistema de puntuación más sencillo

# Métodos de alineamientos y algoritmos

Si hay más de un alineamiento con igual puntuación, será criterio del investigador decidir cuál es el más probable.

**GAATTTCAG**

| | | | |

**GGA-TC-G**

**GAATTTC-A**

| | | | |

**GGA-TCGA**

**GAATTTCAG**

| | | | |

**GCAT-C-G**

**GAATTTC-A**

| | | | |

**GCAT-CGA**

¿Cuál elijo?

# Métodos de alineamientos y algoritmos

Un sistema de puntuación simple puede conducir a obtener muchas soluciones

**Solucion:**

**Utilizar matrices de puntuación (scoring) o de substitución (substitution)**

- Una forma usual de definir el sistema de puntuación es utilizando una matriz de substitución
  - Es una tabla que contiene las puntuaciones que asignamos a cada pareja posible de caracteres.
    - sirve para las coincidencias y las no-coincidencias
  - El término “substitución” refleja que lo que se pretende al puntuar un emparejamiento es valorar el costo evolutivo de cambiar un residuo por otro.

# Métodos de alineamientos y algoritmos

## Puntuación (score) de los alineamientos

### Scoring matrices

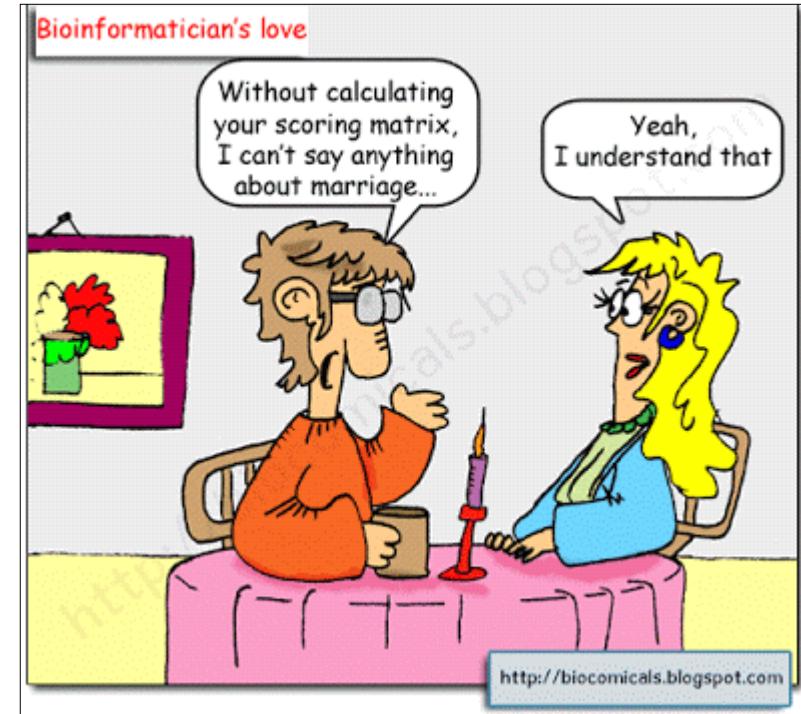
- ❖ Sequence alignment and database searching programs compare sequences to each other as a series of characters.
- ❖ All algorithms (programs) for comparison rely on some scoring scheme for that.
- ❖ Scoring matrices are used to assign a score to each comparison of a pair of characters.

# Métodos de alineamientos y algoritmos

## Puntuación (score) de los alineamientos - Para Ácidos nucleicos

En muchos casos se utiliza una matriz de puntuación (*scoring matrix*) donde se tienen en cuenta todas las sustituciones posibles.

A cada sustitución se le asigna una puntuación distinta porque:

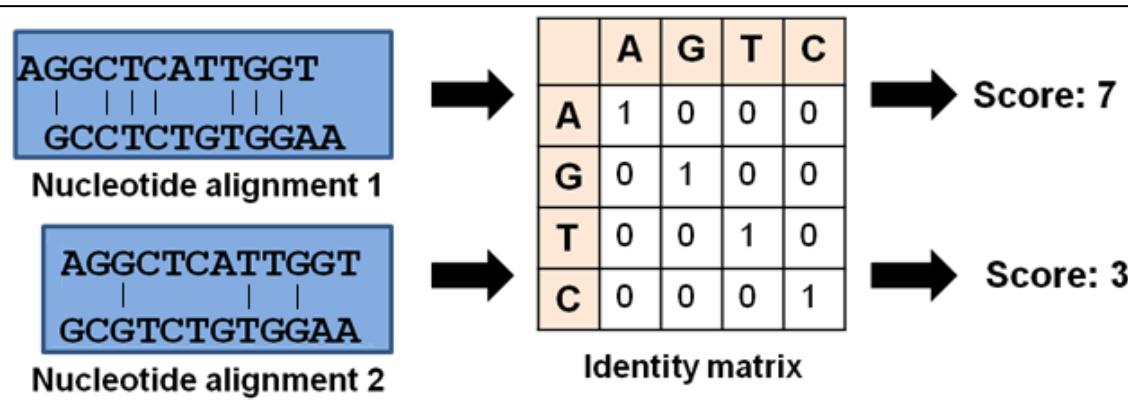


1.- No todos los nucleótidos sustituyen a otros con la misma probabilidad (las transiciones son más probables que las transversiones).

2.- No todos los aminoácidos sustituyen a otros con la misma probabilidad (muchas de las sustituciones observadas son conservativas).

# Métodos de alineamientos y algoritmos

## Puntuación (score) de los alineamientos



Matriz de identidad que otorga una puntuación de 1 en caso de coincidencia. En caso contrario, la puntuación es 0.

## Matriz de identidad



También se pueden incluir valores distintos de 0 y una penalización por introducir huecos

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

value of matching G with A

value of matching C with gap

value of matching G with G

# Métodos de alineamientos y algoritmos

Puntuación (score) de los alineamientos- Modelo de Jukes-Cantor (uniforme)

## Mutation probability matrix (PAM-1)

**Table 3.4.** Nucleotide mutation matrix for an evolutionary distance of 1 PAM, which corresponds to a probability of a change at each nucleotide position of 1%

A. Model of uniform mutation rates among nucleotides

	A	G	T	C
A	0.99			
G	0.00333	0.99		
T	0.00333	0.00333	0.99	
C	0.00333	0.00333	0.00333	0.99

Values are frequency of change at each site, or of no change for all base combinations.

**Table 3.5.** Nucleotide substitution matrix at 1 PAM of evolutionary distance

A. Model of uniform mutation rates among nucleotides

	A	G	T	C
A	2			
G	-6	2		
T	-6	-6	2	
C	-6	-6	-6	2

Units are log odds scores obtained as described in the text.

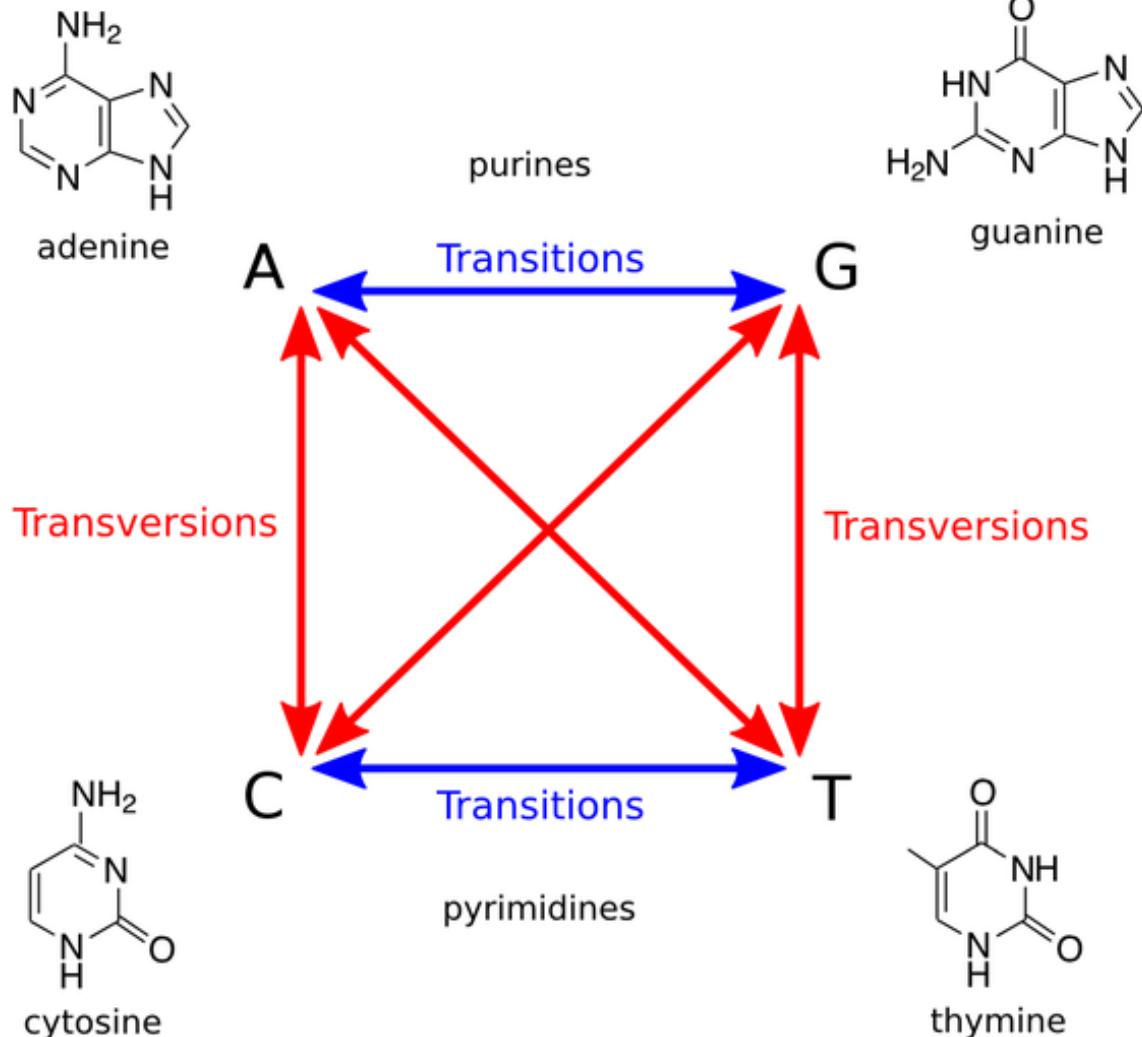
Se contempla un periodo evolutivo en el que ha habido una mutación puntual aceptada por cada 100 bases (PAM-1).

Se considera un modelo mutacional de Markov en el que las mutaciones son aleatorias e independientes.

Todas las mutaciones son igual de probables y todas las bases aparecen con la misma frecuencia.

$$s_{ij} = \log_2 (p_i M_{ij} / p_i p_j)$$

# Métodos de alineamientos y algoritmos



Transiciones y transversiones

Transición ( $A \leftrightarrow G$ ) ( $C \leftrightarrow T$ )

(purina $\leftrightarrow$ purina)  
(pirimidina $\leftrightarrow$ pirimidina)

Transversión  
( $A \leftrightarrow T$ ) ( $A \leftrightarrow C$ )  
( $G \leftrightarrow T$ ) ( $G \leftrightarrow C$ )

(purina $\leftrightarrow$ pirimidina)  
(pirimidina $\leftrightarrow$ purina)

Las transiciones son tres veces más probables que las transversiones.

Modelo de Kimura (sesgado)

# Métodos de alineamientos y algoritmos

## Mutation probability matrix (PAM-1)

**Table 3.4.** Nucleotide mutation matrix for an evolutionary distance of 1 PAM, which corresponds to a probability of a change at each nucleotide position of 1%

	A	G	T	C
A	0.99			
G	0.006	0.99		
T	0.002	0.002	0.99	
C	0.002	0.002	0.006	0.99

Values are frequency of change at each site, or of no change for all base combinations.

**Table 3.5.** Nucleotide substitution matrix at 1 PAM of evolutionary distance

### B. Model of threefold higher transitions than transversions

	A	G	T	C
A	2			
G	-5	2		
T	-7	-7	2	
C	-7	-7	-5	2

Units are log odds scores obtained as described in the text.

Se contempla un periodo evolutivo en el que ha habido una mutación puntual aceptada por cada 100 bases (PAM-1).

Se considera un modelo mutacional de Markov: las mutaciones son aleatorias e independientes.

Las transiciones son 3 veces más probables que las transversiones. Todas las bases aparecen con la misma frecuencia.

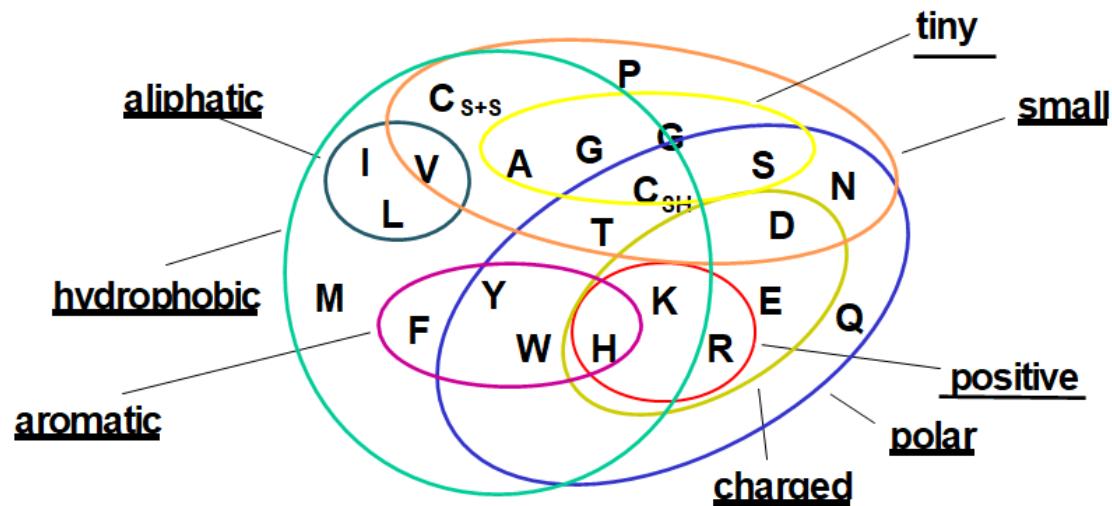
$$s_{ij} = \log_2(p_i M_{ij} / p_i p_j)$$

## Modelo de Kimura (sesgado)

# Métodos de alineamientos y algoritmos

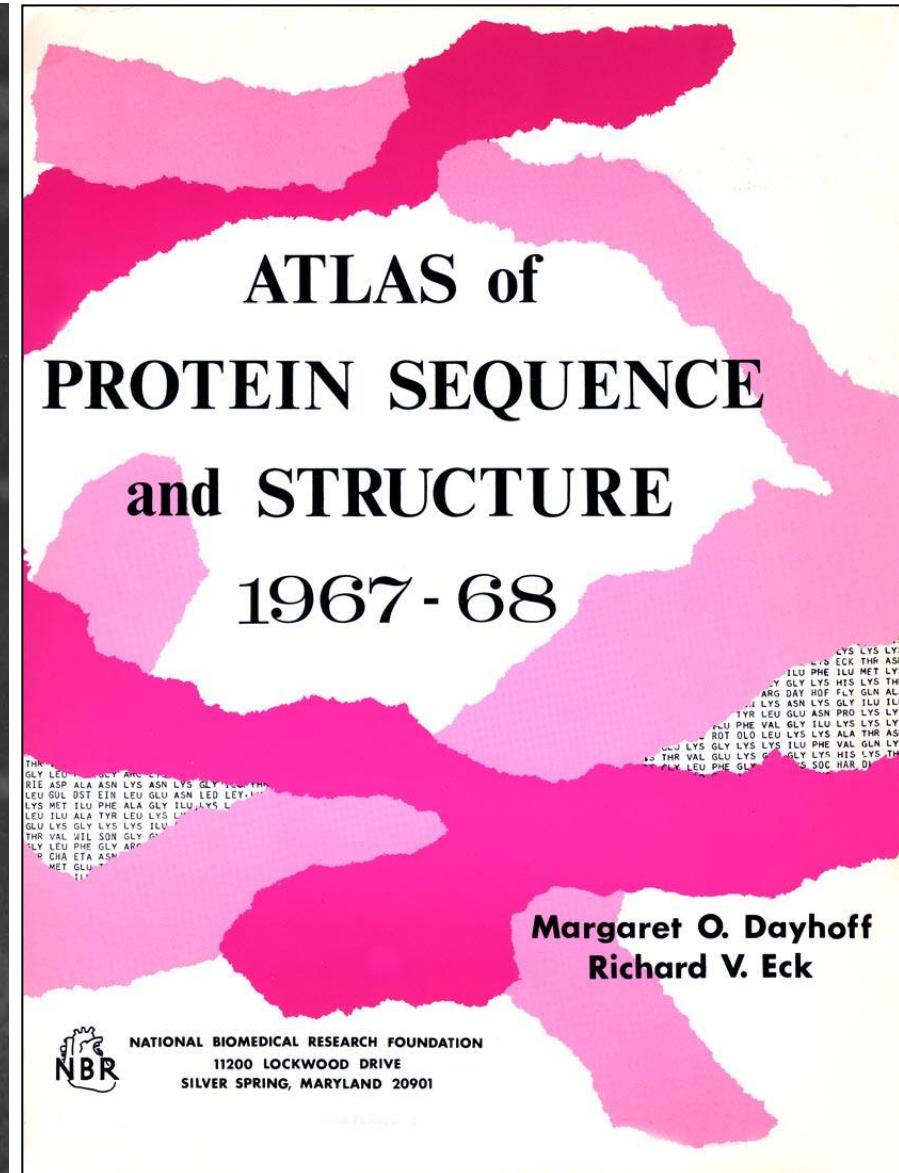
## Puntuación (score) de los alineamientos - Para proteínas

- A substitution matrix contains values proportional to the probability that amino acid i mutates into amino acid j for all pairs of amino acids.
- Substitution matrices are constructed by assembling a large and diverse sample of verified pairwise alignments (or multiple sequence alignments) of amino acids.
- Substitution matrices should reflect the true probabilities of mutations occurring through a period of evolution.
- The two major types of substitution **matrices** are **PAM** and **BLOSUM**.



Los aminoácidos tienen distintas propiedades posibilidades distintas de ser sustituidos unos por otros en la evolución

# Métodos de alineamientos y algoritmos



## Matrices PAM para aminoácidos

# Métodos de alineamientos y algoritmos

## 22 A Model of Evolutionary Change in Proteins

1978

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt

### Accepted Point Mutations

An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein; the second is the acceptance of the mutation by the species as the new predominant form. To be accepted, the new amino acid usually must function in a way similar to the old one: chemical and physical similarities are found between the amino acids that are observed to interchange frequently.

Any complete discussion of the observed behavior of amino acids in the evolutionary process must consider the frequency of change of each amino acid to each other one and the propensity of each to remain unchanged. There are  $20 \times 20 = 400$  possible comparisons. To collect a useful amount of information on these, a great many observations are necessary. The body of data used in this study includes 1,572 changes in 71 groups of closely related proteins appearing in the *Atlas* volumes through Supplement 2.

Una mutación puntual aceptada (PAM) es la sustitución de un aminoácido por otro que ha sido aceptada por la selección natural.

We have assumed that the likelihood of amino acid X replacing Y is the same as that of Y replacing X,

By comparing observed sequences with inferred ancestral sequences, rather than with each other, a sharper picture of the acceptable point mutations is obtained.

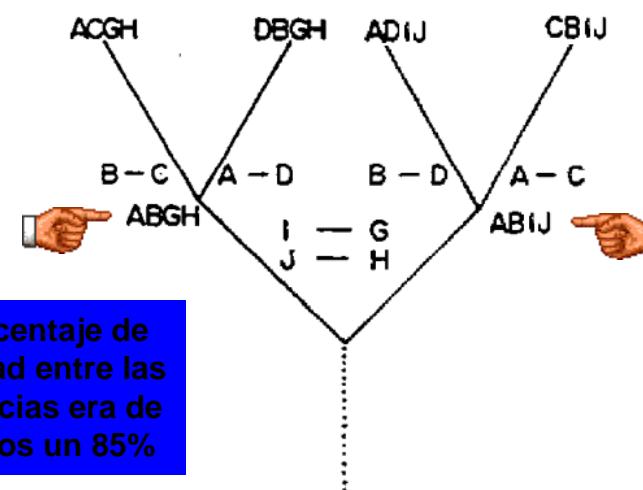


Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

El artículo original de Dayhoff y col.

# Métodos de alineamientos y algoritmos

La matriz de *log odds* es simétrica ( $S_{i,j} = S_{j,i}$ ). A partir de ella se puede deducir que:

— Si  $S_{i,j} > 0$ , el aa *i* sustituye al aa *j* con más frecuencia de lo que se podría esperar por simple azar

— Si  $S_{i,j} < 0$ , el aa *i* sustituye al aa *j* con menos frecuencia de lo que se podría esperar por simple azar

— Si  $S_{i,j} = 0$ , el aa *i* sustituye al aa *j* con la frecuencia que se podría esperar por simple azar

PAM250 MATRIX

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

Table 1 - The log odds matrix for 250 PAMs (multiplied by 10)

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

$$S_{i,j} = 10 \times \log_{10}(R_{i,j})$$



Matriz de probabilidades relativas ( $R_{i,j}$ ) para PAM = 250

# Métodos de alineamientos y algoritmos

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

Figure 3.14. The log odds form (the mutation data matrix or MDM) of the PAM250 scoring matrix.

**PAM 250 (*log odds matrix*)**

# Métodos de alineamientos y algoritmos

## What do the numbers mean in a log odds matrix?

---

A score of +2 indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance.

A score of 0 is neutral.

A score of -10 indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one tenth as frequent as the chance alignment of these amino acids.

¿Cómo se interpretan los valores de la matriz?

# Métodos de alineamientos y algoritmos

$$S_{i,j} = 10 \times \log_{10}(R_{i,j})$$

$$Q \leftrightarrow E = 2 \rightarrow$$

En realidad es 0,2 porque se multiplicó por 10

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

$$0,2 = \log_{10} (\text{probabilidad relativa})$$

Un valor  $> 0$  indica que el cambio es más frecuente de lo esperable al azar

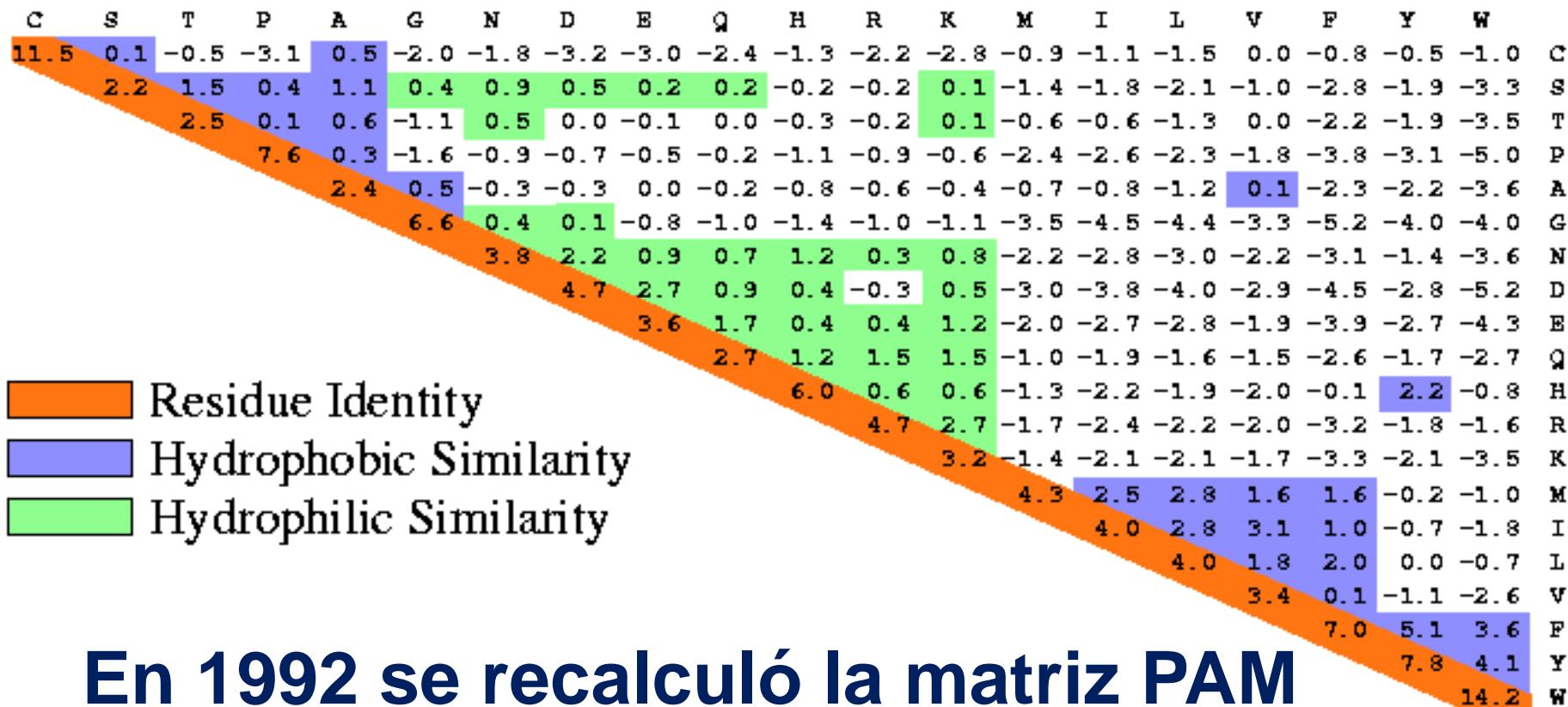
$$\text{Prob. rel.} = 10^{0,2}$$
$$\text{Prob. rel.} = 1,6$$

el cambio  $Q \leftrightarrow E$  es 1,6 veces más frecuente de lo esperable al azar

## Un ejemplo

# Métodos de alineamientos y algoritmos

Gonnet, Cohen, Benner (1992). Exhaustive matching of the entire protein sequence database. *Science*. 256:1443-1445.



En 1992 se recalcularó la matriz PAM

Gonnet PAM250

# Métodos de alineamientos y algoritmos

The alignments used by Dayhoff had ~85% identity

However, frequencies of substitutions are expected to depend on the rate of divergence between sequences: the number of substitutions increases with time.

In order to take into account the divergence rate, Margaret Dayhoff calculated a **series of scoring matrices**, each reflecting a certain level of divergence

## PAM001

- rates of substitutions between amino-acid pairs expected for proteins with an average of 1% substitution per position

## PAM050

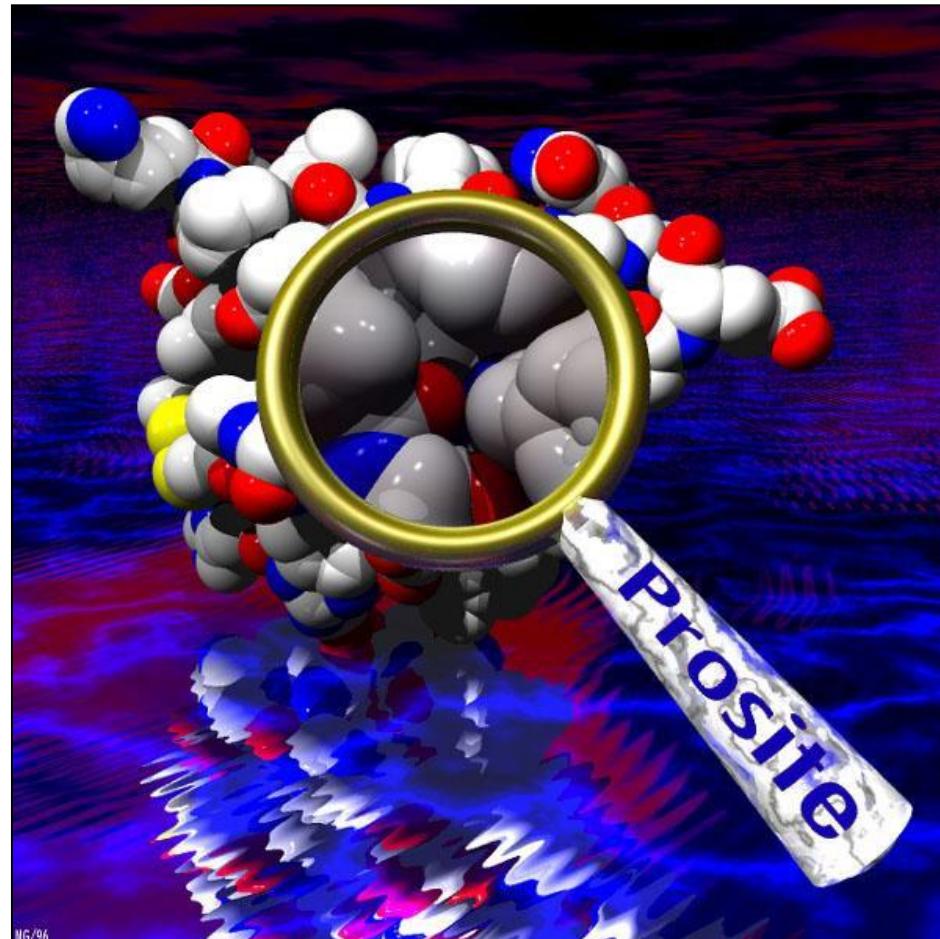
- rates of substitutions between amino-acid pairs expected for proteins with an average of 50% substitution per position

## PAM250

- 250% mutations/position (**note:** a position could mutate several times)

The substitution matrix must this be chosen according to the relatedness of the sequences to be aligned

# Métodos de alineamientos y algoritmos (BLOcks SUbstitution Matrix)



**PROSITE**  
Database of protein families and domains

**Steven y Jorja Henikoff**

# Métodos de alineamientos y algoritmos

Proc. Natl. Acad. Sci. USA  
Vol. 89, pp. 10915–10919, November 1992  
Biochemistry

## Amino acid substitution matrices from protein blocks

(amino acid sequence/alignment algorithms/data base searching)

STEVEN HENIKOFF\* AND JORJA G. HENIKOFF

Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98104

Communicated by Walter Gilbert, August 28, 1992 (received for review July 13, 1992)

**ABSTRACT** Methods for alignment of protein sequences typically measure similarity by using a substitution matrix with scores for all possible exchanges of one amino acid with another. The most widely used matrices are based on the Dayhoff model of evolutionary rates. Using a different approach, we have derived substitution matrices from about 2000 blocks of aligned sequence segments characterizing more than 500 groups of related proteins. This led to marked improvements in alignments and in searches using queries from each of the groups.

new sequence and every other sequence. For example, if the residue of the new sequence at position 1 is A and the residue of the first column of the first block is 1 A residue and 1 S residue, then there is a 1 AS mismatch. This procedure is repeated for all residues in the new sequence and all blocks with the summed results stored in a table. Then the new sequence is added to the group. For each residue, the same procedure is followed, summing the counts for each residue in the new sequence to the group leads to a table of counts of all possible amino acid

Se parte de un conjunto de secuencias de proteínas relacionadas extraídas de la BD PROSITE 9.0

Las secuencias corresponden a 559 familias de proteínas (muchos más datos de partida que en el caso de las matrices PAM)

Los programas MOTIF y PROTOMAT detectaron 2106 bloques en esas secuencias

AABCD	---	BBCDA
DABCD	-A-	BBCBB
BBBCDBA	-	BCCAA
AAACDC	-DC	BCDDB
CCBADB	-DB	BBDCC
AAACA	---	BBCCC

El artículo original de los Henikoff

# Métodos de alineamientos y algoritmos

Un bloque es un alineamiento local y sin huecos de una región conservada en una familia de proteínas

Los bloques constituyen una característica distintiva de la familia, ya que suelen contener los aa responsables de la función bioquímica común a todos sus miembros

En cada bloque, cada línea corresponde a una proteína, y cualquiera de ellas puede ser ancestro de la otra (un modelo evolutivo que se denomina starburst)

bloque 1

WWYIR  
WFYVR  
WYYVR  
WYFIR

bloque 2

CASILRKIYIYGPV  
CASICLRLHYHRSPA  
AAAavarHIYLRLKTV  
AASICRHLIYIRSPA

bloque 3

GVSRLRTAYGGRKNRG  
GVGSITKIYGGRKNG  
GVGRLRKVHGSTKNRG  
GIGSFEKIYGGRRRRG

Características de los bloques

# Métodos de alineamientos y algoritmos

## Histone Block Segment

```
KKA S KPKKAASKAP T KKPATPVKKAKKK L AATPKK A KKPK T VK
KKA A KPKKAASKAP S KKPATPVKKAKKK P AATPKK A KKPK V VK
KKA A KPKKAASKAP S KKPATPVKKAKKK P AATPKK A KKPK I VK
KKA A KPKKAASKAP S KKPATPVKKAKKK P AATPKK T KKPK T VK
KKA S KPKKAASKAP T KKPATPVKKAKKK L AATPKK A KKPK T VK
```

Matches = 39 columns x 6 rows = 234

BLOSUM	Identity	Evolutionary Distance
BLOSUM45	Up to 45%	Largest
BLOSUM50	Up to 50%	Large
BLOSUM52	Up to 52%	Medium
BLOSUM60	Up to 60%	Short
BLOSUM80	Up to 80%	Shorter
BLOSUM90	Up to 90%	Shortest

BLOSUM K

Para la construcción de la matrix BLOSUM se analizaron 2000 segmentos /bloques

Para 500 grupos de proteínas relacionadas que fueron agrupadas de acuerdo a su porcentaje de identidad

# Métodos de alineamientos y algoritmos

	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				basic						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		aromatic	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	

BLOSUM 62

# Métodos de alineamientos y algoritmos

## Blosum 62

High Penalty for very different aminoacids

Small positive score for changes in similar aminoacids

Small positive score for common aminoacids

Infrequent aminoacids have high score

	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	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	-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

$$S_{ij} = 2 \cdot \log_2 \frac{p_{ij}}{p_i p_j}$$

¿Cómo se interpretan los valores de la matriz?

# Métodos de alineamientos y algoritmos

Si  $S_{i,j} > 0$ , el aa  $i$  sustituye al aa  $j$  con más frecuencia de lo que se podría esperar por simple azar

Si  $S_{i,j} < 0$ , el aa  $i$  sustituye al aa  $j$  con menos frecuencia de lo que se podría esperar por simple azar

Si  $S_{i,j} = 0$ , el aa  $i$  sustituye al aa  $j$  con la frecuencia que se podría esperar por simple azar

**La puntuación del alineamiento es la suma de los logaritmos de las probabilidades relativas de cada pareja de aa alineada**

```

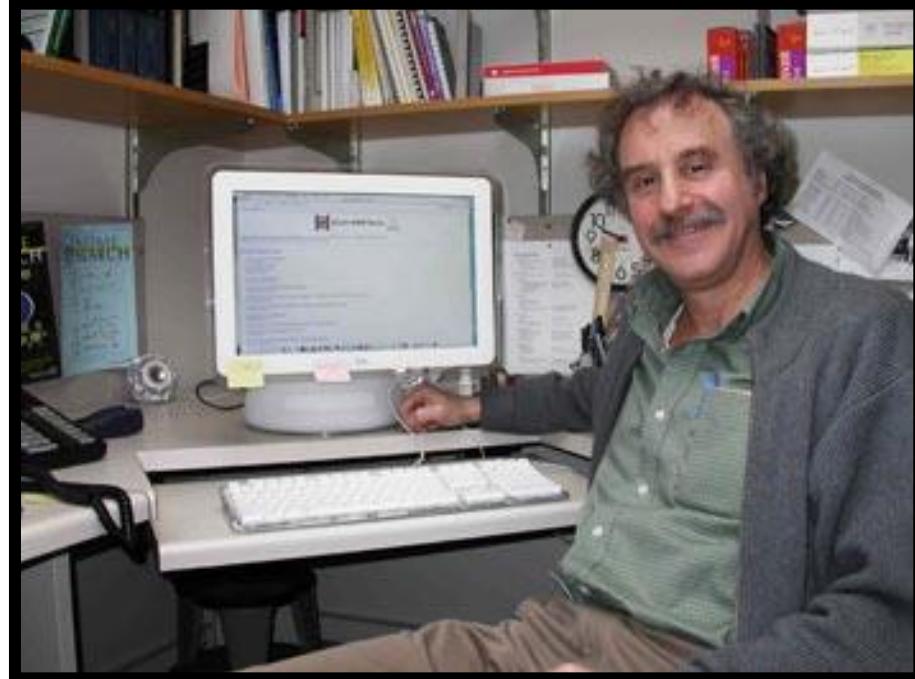
=====
# Aligned_sequences: 2 Length: 125 #
# 1: RPC1_LAMBD Identity: 60/125 (48.0%) #
# 2: NP_059606.1 Similarity: 82/125 (65.6%) #
# Matrix: EBLOSUM62 Gaps: 3/125 ( 2.4%) #
# Gap_penalty: 15.0 Score: 297.0  #
# Extend_penalty: 1.0 #
=====#

```

LAMBD	104	YPVFSHVQAGMFSPELRTFTKGDAERWVSTTKASDSAFWLEVEGNSMTA	153	
		.: . . . .: .: .: ... ... . .. . .. .: .. .: .. .: ..		
NP_059606.1	86	YPLISWVSAGQWMEAEPYHKRAIENWHDTTVDCSEDSFWLDVQGDSMTA	135	
LAMBD	154	PTGSKPSFPDGMLILVDPEQAVEPGDFCIARLGGD-EFTFKKLIRDGQV	202	
		. .  . .: .. :     ... ... .: .. .: .. .. .. .: ..		
NP_059606.1	136	PAGL--SIPEGMIIILVDPEVEPRNGKLVVAKLEGNEATFKKLVMDAGRK	183	
LAMBD	203	FLQPLNPQYPMIPCNECSVVGKVI	227	
		:     .. .: .. .: ..		
NP_059606.1	184	FLKPLNPQYPMIEINGNCKIIGVVV	208	

## Puntuación de un alineamiento con la matriz BLOSUM 62

# Métodos de alineamientos y algoritmos



Se construye a partir de  
alineamientos globales

Pocos datos de partida

Se basa en un modelo  
evolutivo mutacional  
(proceso de Markov)

Los errores en PAM-1 se  
amplifican 250 veces en PAM250

Las secuencias de  
partida son muy  
similares (> 85%)

Cómputo de cambios  
basado en el método  
de máxima parsimonia

Se construye a partir de  
alineamientos locales

Se basa en un modelo  
evolutivo del tipo *starburst*

Los errores en BLOSUM  
se deben a alineamientos  
incorrectos

Se computan los cambios agrupando las  
secuencias que superan un umbral de similitud

**PAM versus BLOSUM (1)**

# Métodos de alineamientos y algoritmos

## PAM

Para detectar homología en secuencias alejadas se utilizan matrices PAM con un número elevado

PAM con números elevados indican más divergencia

Diseñadas para desvelar el parentesco evolutivo de las proteínas

BLOSUM 80

PAM 1

*Less divergent*

## BLOSUM

Para detectar homología en secuencias alejadas se utilizan matrices BLOSUM con un número bajo

BLOSUM con números elevados indican menos divergencia

Diseñadas para encontrar dominios conservados en las proteínas

BLOSUM 62

PAM 120

BLOSUM 45

PAM 250

*More divergent*

## **PAM versus BLOSUM (2)**

# Métodos de alineamientos y algoritmos

## En resumen:

- No hay una matriz única que se pueda usar siempre
- Según la familia de proteínas y el grado de similitud esperado se usará una u otra
- Las más utilizadas PAM y BLOSUM:

- **PAM: Percent Accepted Mutation Matrix:**

Derivadas de alineamientos globales de secuencias próximas  
PAM40 → PAM250. A mayor nº mayor distancia evolutiva

- **BLOSUM:**

Derivadas de alineamientos locales de secuencias distantes  
BLOSUM90 → BLOSUM45 El nº representa porcentaje de identidad

# Métodos de alineamientos y algoritmos

## En resumen:

Different substitution scoring matrices have been established PAM  
(Dayhoff, 1979)

BLOSUM (Henikoff & Henikoff, 1992)

Residue categories (Phylip)

Substitution matrices allow to detect similarities between more distant proteins than what would be detected with the simple identity of residues.

The matrix must be chosen carefully, depending on the expected rate of conservation between the sequences to be aligned.

With **PAM** matrices

the score indicates the percentage of substitution per position  
→ **higher numbers** are appropriate for **more distant** proteins

With **BLOSUM** matrices

the score indicates the percentage of conservation  
→**higher numbers** are appropriate for **more conserved** proteins

# Métodos de alineamientos y algoritmos

## Otros tipos de matrices

**Matriz de identidad**

**Matriz de sustitución de codones**

**Matriz de hidrofobicidad**

**Cadenas laterales de los aminoácidos**

# Métodos de alineamientos y algoritmos

## y los “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 distinto
  - 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.

**Existen regiones muy conservadas y con funciones muy delicadas que no tolerarían ningún cambio**

# Métodos de alineamientos y algoritmos

## Penalización afín - y los “gaps”

Desde un punto de vista evolutivo, es más realista suponer que la naturaleza ha insertado/eliminado fragmentos en la secuencia de una sola vez. Por eso se introduce una penalización ( $g_o$ ) para la inclusión de un indel (gap open penalty) y otra penalización ( $g_e$ ), menos costosa, que depende de la longitud del indel (gap extension penalty).

La inserción/eliminación es mucho menos probable que cualquier sustitución de aa, por radical que ésta sea. Por tanto, la  $g_o$  debe estar muy penalizada para que se introduzcan indels donde sea preciso, y no por toda la secuencia

Una vez que se ha introducido un indel en un punto de la secuencia, su extensión ( $g_e$ ) es mucho más probable y debe estar menos penalizada.

# Métodos de alineamientos y algoritmos

En la penalización afín hay dos maneras distintas de penalizar la extensión del indel :

Modelo lineal: Para todo  $n > 1$ ,  $p(n+1) - p(n) = p(n) - p(n-1)$

(La penalización es proporcional a la longitud del indel)

$$G = g_o + ng_e$$

$$G = g_o + (n-1)g_e$$

Modelo convexo: Para todo  $n > 1$ ,  $p(n+1) - p(n) < p(n) - p(n-1)$

(Cada tramo adicional del indel penaliza menos que el anterior. Es el modelo que más se ajusta a la realidad, pero desde el punto de vista computacional es muy difícil incluirlo en el algoritmo )

$$G = g_o + k \log (n)$$

Dos modelos de penalización afín (1)

# Métodos de alineamientos y algoritmos

Es importante seleccionar una penalización apropiada en función de la matriz de puntuación elegida para que no se excluyan los indels, pero que tampoco se propaguen por todo el alineamiento.

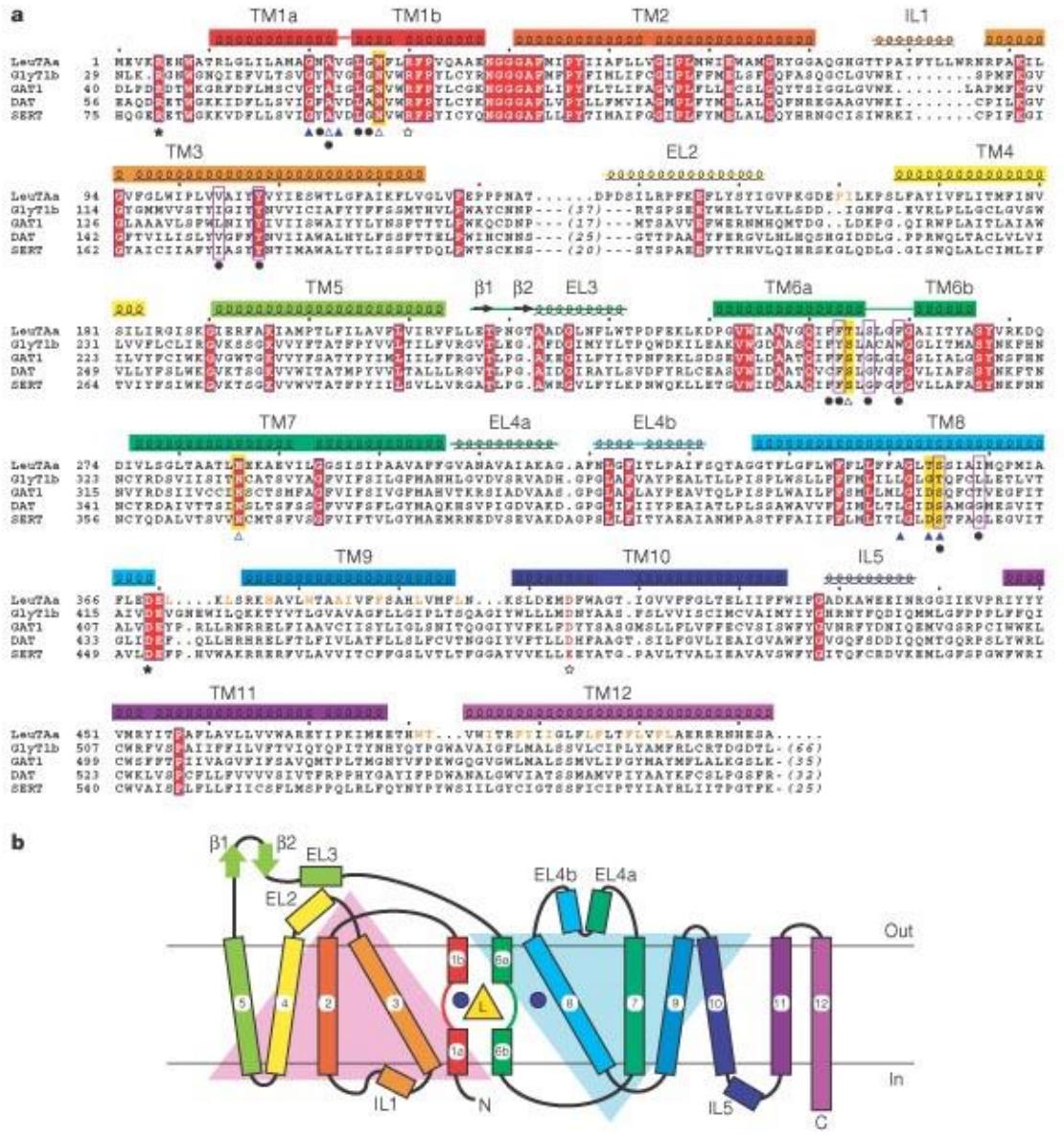
No hay una mecanismo formal para calcular el valor de la penalización. La mayor parte de los programas hacen sus propias recomendaciones, que están basadas en métodos de ensayo y error, pero que no garantizan que para tu caso concreto sean las más adecuadas. Deberás hacer varias pruebas.

Algunos valores típicos:

Matriz	gap opening	gap extension
BLOSUM 62	-12	- 3
BLOSUM 50	-15	- 8
PAM 250	-15	- 5

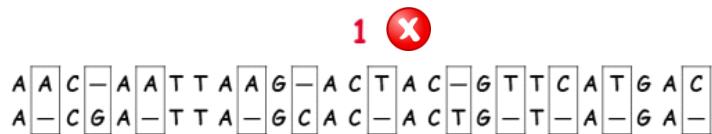
Algunas recomendaciones

# Métodos de alineamientos y algoritmos



Los huecos suelen incluirse en los bucles que conectan los elementos de estructura secundaria

Se considera más lógico introducir un hueco de longitud  $n$  que  $n$  huecos de longitud 1.



Dónde y cómo introducir huecos

# Métodos de alineamientos y algoritmos

## ¿Cómo los uso?

Coste de apertura de <i>gap</i>	Coste de extensión del <i>gap</i>	Comentario
Grande	Grande	Pocas inserciones o eliminaciones Bueno para proteínas muy relacionadas
Grande	Pequeño	Algunas inserciones grandes Bueno si puede que se hayan insertado dominios completos
Pequeño	Grande	Muchas inserciones pequeñas Bueno si se trata de proteínas distantes

# Métodos de alineamientos y algoritmos

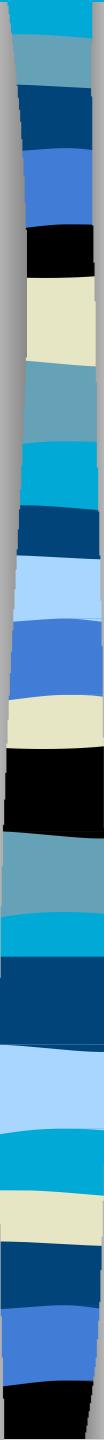
## En resumen...

Una vez fijado un sistema de puntuación

- Matriz de substitución (Identidad, PAMxx, BLOSUM...)
  - Coste de la apertura y de la extensión de “gaps”
- 
- El problema se reduce a encontrar el alineamiento óptimo o el mejor
  - Se define el alineamiento óptimo entre dos secuencias como aquel cuya puntuación es máxima entre todos los posibles alineamientos.

**¿Como encuentro el alineamiento de mayor puntuación u óptimo?**

**Lo veremos la próxima clase!!!**



# Métodos de alineamientos y algoritmos

## Programación dinámica

### Alineamiento óptimo

The optimal alignment of two sequences is one that finds the longest segment of high sequence similarity.