

## La Estadística detrás del Análisis Filogético

Hugo Flores Arguedas

Departamento Ciencias y Matemáticas  
Arkansas State University Campus Queretaro  
[hfloresarguedas@astate.edu.mx](mailto:hfloresarguedas@astate.edu.mx)

Red Mexicana de Biología y Matemáticas, Nov 3, 2025



Comparando Secuencias: Valorando la Similitud

¿Cómo obtengo las secuencias?

(Scoring) Matrices de Puntaje

## Motivaciones

# ¿Por qué Matemáticas?



Ideas worth spreading

WATCH DISCOVER ATTEND PAR



TED2016 • February 2016 | 2.1M views

Like (64K)

Share

Add

## What's so sexy about math?

Cédric Villani

[Read transcript](#)

Villani Talk

# ¿Por qué Análisis Filogético?

## nature ecology & evolution

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [nature ecology & evolution](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 12 July 2024

### The nature of the last universal common ancestor and its impact on the early Earth system

[Edmund R. R. Moody](#) , [Sandra Álvarez-Carretero](#), [Tara A. Mahendarajah](#), [James W. Clark](#), [Holly C. Betts](#), [Nina Dombrowski](#), [Lénárd L. Szánthó](#), [Richard A. Boyle](#), [Stuart Daines](#), [Xi Chen](#), [Nick Lane](#), [Ziheng Yang](#), [Graham A. Shields](#), [Gergely J. Szöllősi](#), [Anja Spang](#), [Davide Pisani](#) , [Tom A. Williams](#) , [Timothy M. Lenton](#)  & [Philip C. J. Donoghue](#) 

[Nature Ecology & Evolution](#) 8, 1654–1666 (2024) | [Cite this article](#)

102k Accesses | 7 Citations | 1381 Altmetric | [Metrics](#)

Luca Nature

# Inteligencia Artificial Generativa



ChatGPT ▾

Log in



¿Por qué es importante el Análisis Filogenético?

El **Análisis Filogenético** es una herramienta fundamental en biología evolutiva y en muchas otras ramas de las ciencias naturales. Su importancia radica en que permite **reconstruir las relaciones evolutivas entre organismos**, es decir, entender **cómo están emparentadas las distintas especies** (o genes, o proteínas) y **cómo han evolucionado a lo largo del tiempo**.

# Inteligencia Artificial Generativa

Pero, ¿en serio se necesita estadística para entenderlo?

¡Excelente pregunta! 🧠

Sí, **la estadística es fundamental para entender y realizar un análisis filogenético**, aunque el grado de complejidad depende del nivel al que trabajes. Te explico por qué:

# Inteligencia Artificial Generativa

Pero, ¿en serio se necesita estadística para entenderlo?

¡Excelente pregunta! 🧠

Sí, la estadística es fundamental para entender y realizar un análisis filogenético, aunque el grado de complejidad depende del nivel al que trabajes. Te explico por qué:

## 💡 En resumen:

No necesitas ser un experto en estadística para entender los conceptos básicos del análisis filogenético, pero sí es indispensable para hacerlo rigurosamente y evaluar los resultados con confianza.

Piensa en la estadística como las "lentes" que te permiten ver la evolución con claridad en medio del ruido biológico.

## Introducción

## Homología

El concepto de homología, **ascendencia evolutiva común**, es central para el análisis computacional de secuencias de proteínas y de ADN.

## Homología

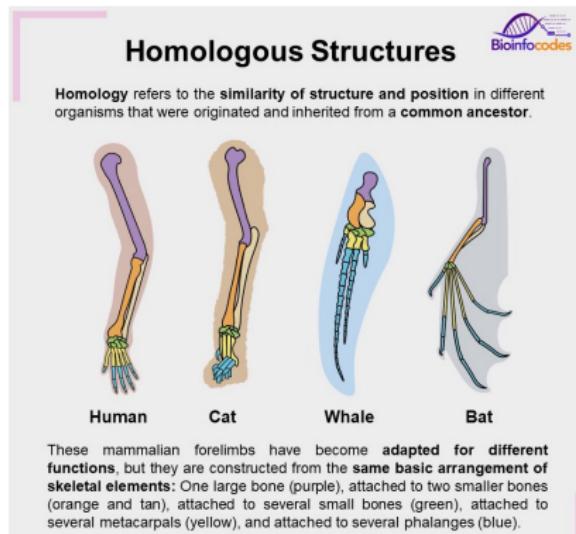
El concepto de homología, **ascendencia evolutiva común**, es central para el análisis computacional de secuencias de proteínas y de ADN.

**Inferimos** homología cuando dos secuencias o estructuras comparten mayor similitud que lo que se espera por azar.

## Homología

El concepto de homología, **ascendencia evolutiva común**, es central para el análisis computacional de secuencias de proteínas y de ADN.

**Inferimos** homología cuando dos secuencias o estructuras comparten mayor similitud que lo que se espera por azar.



## Homología

Para el análisis de secuencias, cuando un **exceso de similitud** es observado, la explicación más sencilla para este exceso es que las dos secuencias no *surgieron* independientemente, sino que surgen de un ancestro en común.

## Homología

Para el análisis de secuencias, cuando un **exceso de similitud** es observado, la explicación más sencilla para este exceso es que las dos secuencias no *surgieron* independientemente, sino que surgen de un ancestro en común.

Esta es **la explicación más sencilla**, pero no necesariamente la indicada. Por lo tanto, buscaremos determinar la significancia estadística de esta similitud.

## Homología

Para el análisis de secuencias, cuando un **exceso de similitud** es observado, la explicación más sencilla para este exceso es que las dos secuencias no *surgieron* independientemente, sino que surgen de un ancestro en común.

Esta es **la explicación más sencilla**, pero no necesariamente la indicada. Por lo tanto, buscaremos determinar la significancia estadística de esta similitud.

Cuando una búsqueda de similitud encuentra un **coincidencia (match) estadísticamente significativa**, podemos inferir que las dos secuencias son homólogas.

## Homología

Para el análisis de secuencias, cuando un **exceso de similitud** es observado, la explicación más sencilla para este exceso es que las dos secuencias no *surgieron* independientemente, sino que surgen de un ancestro en común.

Esta es **la explicación más sencilla**, pero no necesariamente la indicada. Por lo tanto, buscaremos determinar la significancia estadística de esta similitud.

Cuando una búsqueda de similitud encuentra un **coincidencia (match) estadísticamente significativa**, podemos inferir que las dos secuencias son homólogas.

**¿Qué pasa si no encontramos coincidencias estadísticamente significativas en una base de datos?**

## Homología

Para el análisis de secuencias, cuando un **exceso de similitud** es observado, la explicación más sencilla para este exceso es que las dos secuencias no *surgieron* independientemente, sino que surgen de un ancestro en común.

Esta es **la explicación más sencilla**, pero no necesariamente la indicada. Por lo tanto, buscaremos determinar la significancia estadística de esta similitud.

Cuando una búsqueda de similitud encuentra un **coincidencia (match) estadísticamente significativa**, podemos inferir que las dos secuencias son homólogas.

**¿Qué pasa si no encontramos coincidencias estadísticamente significativas en una base de datos?**

No podemos estar seguros que no hayan homólogos presentes.

## Homología

**¿Cómo determinamos la significancia estadística?**

## Homología

**¿Cómo determinamos la significancia estadística?**

A través de una Prueba de Hipótesis!

## Homología

**¿Cómo determinamos la significancia estadística?**

A través de una Prueba de Hipótesis!

**Objetivo:** Minimizar los Errores

## Homología

¿Cómo determinamos la significancia estadística?

A través de una Prueba de Hipótesis!

**Objetivo:** Minimizar los Errores



## Errores en las Pruebas de Hipótesis

Type I Error



Type II Error



	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

## Homología

*¿Cómo determinamos la significancia estadística?*

## Homología

**¿Cómo determinamos la significancia estadística?**

A través de una Prueba de Hipótesis!

## Homología

**¿Cómo determinamos la significancia estadística?**

A través de una Prueba de Hipótesis!

**Objetivo:** Minimizar los Errores

## Homología

### ¿Cómo determinamos la significancia estadística?

A través de una Prueba de Hipótesis!

**Objetivo:** Minimizar los Errores

- **Falso Positivo:** No-homólogos con puntuaciones significativas (Error Tipo I)

# Homología

## ¿Cómo determinamos la significancia estadística?

A través de una Prueba de Hipótesis!

**Objetivo:** Minimizar los Errores

- **Falso Positivo:** No-homólogos con puntuaciones significativas (Error Tipo I)
- **Falso Negativo:** Homólogos con puntajes no significativos (Error Tipo II)

## Homología

### ¿Cómo determinamos la significancia estadística?

A través de una Prueba de Hipótesis!

**Objetivo:** Minimizar los Errores

- **Falso Positivo:** No-homólogos con puntuaciones significativas (Error Tipo I)
- **Falso Negativo:** Homólogos con puntajes no significativos (Error Tipo II)

¿Podemos minimizar ambos simultáneamente?

Ejemplo: Pruebas de Hipótesis



## Ejemplo: Pruebas de Hipótesis



Hugo quisiera determinar si un estudiante de su clase de Estadística sabe lo suficiente para exentar .

## Ejemplo: Pruebas de Hipótesis



Hugo quisiera determinar si un estudiante de su clase de Estadística sabe lo suficiente para exentar .

**Criterio de Exención:** Premiar el alcance temprano de los objetivos del curso.

## Ejemplo: Pruebas de Hipótesis



## Ejemplo: Pruebas de Hipótesis



¿Qué se espera?

## Ejemplo: Pruebas de Hipótesis



### ¿Qué se espera?

- Que un estudiante que haya alcanzado los objetivos sea exento.

## Ejemplo: Pruebas de Hipótesis



### ¿Qué se espera?

- Que un estudiante que haya alcanzado los objetivos sea exento.
- Que un estudiante que no haya alcanzado los objetivos no sea exento.

## Ejemplo: Pruebas de Hipótesis



## Ejemplo: Pruebas de Hipótesis



**Riesgos al determinar el criterio:**

## Ejemplo: Pruebas de Hipótesis



### Riesgos al determinar el criterio:

- Algún estudiante que ya alcanzó los objetivos no sea exento (Si se escoge un 9 como valor de corte).

## Ejemplo: Pruebas de Hipótesis



### Riesgos al determinar el criterio:

- Algún estudiante que ya alcanzó los objetivos no sea exento (Si se escoge un 9 como valor de corte).
- Algún estudiante que no ha alcanzado los objetivos sea exento (Si se escoge un 7 como valor de corte).

## Ejemplo: Pruebas de Hipótesis



### Riesgos al determinar el criterio:

- Algún estudiante que ya alcanzó los objetivos no sea exento (Si se escoge un 9 como valor de corte).
- Algún estudiante que no ha alcanzado los objetivos sea exento (Si se escoge un 7 como valor de corte).

¿Podemos minimizar ambos riesgos simultáneamente?

## Homología

### ¿Cómo determinamos la significancia estadística?

A través de una Prueba de Hipótesis!

**Objetivo:** Minimizar los Errores

- **Falso Positivo:** No-homólogos con puntuaciones significativas (Error Tipo I)
- **Falso Negativo:** Homólogos con puntajes no significativos (Error Tipo II)

## Homología

### ¿Cómo determinamos la significancia estadística?

A través de una Prueba de Hipótesis!

**Objetivo:** Minimizar los Errores

- **Falso Positivo:** No-homólogos con puntuaciones significativas (Error Tipo I)
- **Falso Negativo:** Homólogos con puntajes no significativos (Error Tipo II)

Herramientas de búsqueda para similitud de secuencias como BLAST y HMMER minimizan los falsos positivos, sin embargo, no realizan ninguna afirmación acerca de los falsos negativos.

## Comparando Secuencias: Valorando la Similitud

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

Por ejemplo, si comparamos las siguientes dos secuencias:

- CAGTCCTATT
- CAGTGGTATT

# Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

Por ejemplo, si comparamos las siguientes dos secuencias:

- CAGTCCTATT
- CAGTGGTATT

**Son similares?**

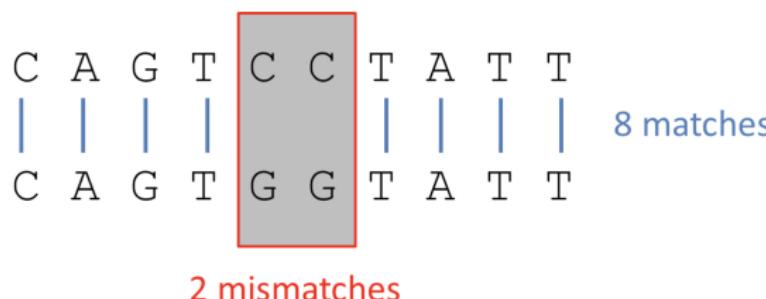
# Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

Por ejemplo, si comparamos las siguientes dos secuencias:

- CAGTCCTATT
- CAGTGGTATT

Son similares?



## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

|APRGKRSTWTIG  
ASFTPPRGKRSTWTIG

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

|APRGKRSTWTIG  
ASFTPPRGKRSTWTIG

Son similares?

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

APRGKRSTWTIG  
ASFTPPRGKRSTWTIG

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

APRGKRSTWTIG  
ASFTPPRGKRSTWTIG

**Son similares?**

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

**Para cuantificar la similitud**, es necesario alinear las dos secuencias, y hasta entonces, podemos calcular un puntaje (score) basado en el alineamiento.

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

**Para cuantificar la similitud**, es necesario alinear las dos secuencias, y hasta entonces, podemos calcular un puntaje (score) basado en el alineamiento.

A- - - -PRGKRSTWTIG  
ASFTP PRGKRSTWTIG

## Similitud

Si estamos estudiando un par de genes o proteínas en particular, una pregunta importante es hasta qué punto son las dos secuencias similares.

**Para cuantificar la similitud**, es necesario alinear las dos secuencias, y hasta entonces, podemos calcular un puntaje (score) basado en el alineamiento.

A - - - PRGKRSTWTIG  
ASFTP PRGKRSTWTIG

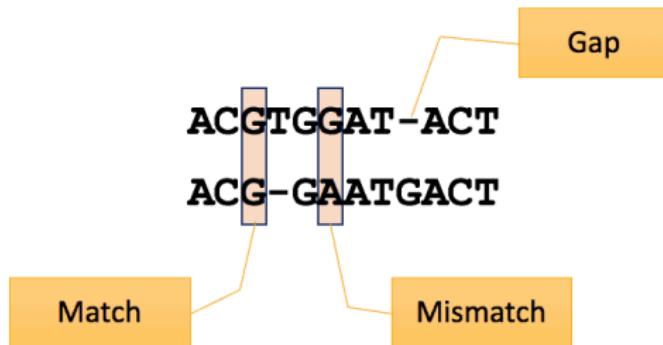
Existen dos tipos de alineamientos, **global y local**.

## Similitud

Al contar con un alineamiento, podemos asignar un puntaje:

# Similitud

Al contar con un alineamiento, podemos asignar un puntaje:



$$\text{Similarity}(\%) = \frac{\text{Match}}{\text{Match} + \text{Mismatch}} \times 100$$

## Similitud

Un **alineamiento global** es un alineamiento de la longitud completa de dos secuencias, de principio a fin, por ejemplo, de dos secuencias de proteínas o dos secuencias de ADN. La optimización puede incluir grandes extensiones de baja similitud.

## Similitud

Un **alineamiento global** es un alineamiento de la longitud completa de dos secuencias, de principio a fin, por ejemplo, de dos secuencias de proteínas o dos secuencias de ADN. La optimización puede incluir grandes extensiones de baja similitud.

Un **alineamiento local** es un alineamiento de una parte de una secuencia con una parte de otra secuencia; las partes que terminan siendo alineadas son las más similares, y son determinadas por el algoritmo de alineamiento.

# Similitud

Un **alineamiento global** es un alineamiento de la longitud completa de dos secuencias, de principio a fin, por ejemplo, de dos secuencias de proteínas o dos secuencias de ADN. La optimización puede incluir grandes extensiones de baja similitud.

Un **alineamiento local** es un alineamiento de una parte de una secuencia con una parte de otra secuencia; las partes que terminan siendo alineadas son las más similares, y son determinadas por el algoritmo de alineamiento.

## A. Examples of Sequences

Sequence1 : ATACCGGATATT

Sequence2 : AACGGACCCCT

## B. Global Alignment

ATACCGGATATT	
- - AACGGACCCCT	
	match

## C. Local Alignment

ATACCGGATATT	
- - AACGGA-----	
	match

# Similarity

## A. Examples of Sequences

Sequence1 : ATACCGGATATT

Sequence2 : AACGGACCCCT

## B. Global Alignment

ATACCGGATATT  
| . | | | | . . . |  
- - AACGGACCCCT  
          brace     brace  
                match

## C. Local Alignment

ATACCGGATATT  
| . | | | |  
- - AACGGA - - -  
          brace  
                match

**FIGURE 2.** Example of global and local alignment in two sequence. In the figure, A is an example of two base sequences: Sequence 1 and Sequence 2. B shows the method for global alignment and C shows the method for local alignment. There are dots, long lines, and short lines. Non-matching pairs are indicated with dots '.', matching pairs are indicated with long vertical lines '|', and insertions or deletions, which show blank spaces, are indicated with short horizontal dashes '-'.

**Figure:** Kim, Ji, & Yi, *A Review on Sequence Alignment Algorithms for Short Reads Based on Next-Generation Sequencing*

## Similitud

Muchas medidas de similitud empiezan con una **matriz de similitud**, donde se asigna un puntaje a todas los posibles pares de residuos. Identidades y reemplazamientos conservadores tienen puntajes positivos mientras que reemplazamientos poco probables tienen puntajes negativos.

# Similitud

Muchas medidas de similitud empiezan con una **matriz de similitud**, donde se asigna un puntaje a todas los posibles pares de residuos. Identidades y reemplazamientos conservadores tienen puntajes positivos mientras que reemplazamientos poco probables tienen puntajes negativos.

Substitution Matrix\*

	R	G	I	K	F	S	T	W	V
R	5	0	-1	-1	-2	1	0	-3	0
G		6	-4	-2	-3	0	-2	-2	-3
I			4	-3	0	-2	-1	-3	3
K				5	-3	0	-1	-3	-2
F					6	-2	-2	1	-1
S						4	1	-3	-2
T							5	-2	0
W								11	-3
V									4

\*A portion of the BLOSUM 62 matrix

¿Cómo obtengo las secuencias?

# Pregunta de interés

Review | [Open access](#) | Published: 31 January 2022

## Role of the S100 protein family in rheumatoid arthritis

[Yuan-yuan Wu](#), [Xiao-feng Li](#), [Sha Wu](#), [Xue-ni Niu](#), [Su-qin Yin](#), [Cheng Huang](#)✉ & [Jun Li](#)✉

*Arthritis Research & Therapy* **24**, Article number: 35 (2022) | [Cite this article](#)

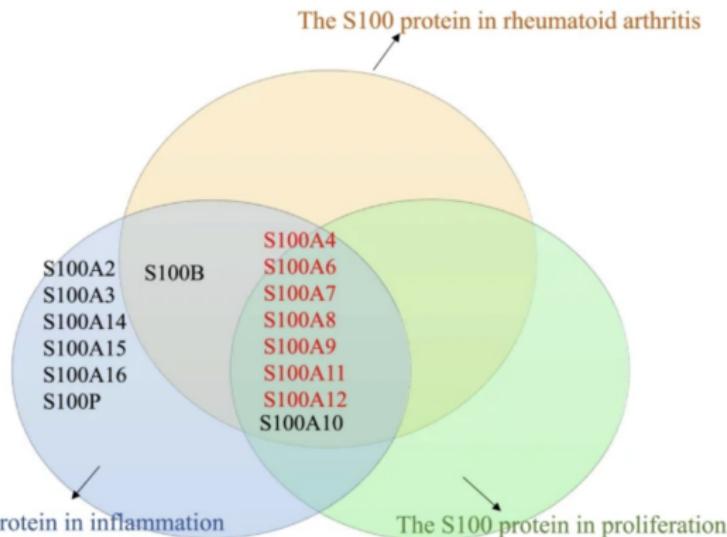
**4418** Accesses | **23** Citations | [Metrics](#)

### Abstract

Rheumatoid arthritis is a chronic systemic autoimmune disease characterized by synovial hyperplasia, inflammatory cell infiltration, and proliferation of inflammatory tissue (angiogranuloma). The destruction of joints and surrounding tissues eventually causes joint deformities and dysfunction or even loss. The S100 protein family is one of the biggest subtribes in the calcium-binding protein family and has more than 20 members. The overexpression of most S100 proteins in rheumatoid arthritis is closely related to its pathogenesis. This paper reviews the relationship between S100 proteins and the occurrence and development of rheumatoid arthritis. It will provide insights into the development of new clinical diagnostic markers and therapeutic targets for rheumatoid arthritis.

## Familia de proteínas S100

Fig. 1



S100 protein was associated with proliferation and inflammation of cells and RA

# Base de datos NCBI

An official website of the United States government [Here's how you know](#)

**National Library of Medicine**  
National Center for Biotechnology Information

[Log in](#)

All Databases [Search](#)

---

**NCBI Home**

[Resource List \(A-Z\)](#)

[All Resources](#)

[Chemicals & Bioassays](#)

[Data & Software](#)

[DNA & RNA](#)

[Domains & Structures](#)

[Genes & Expression](#)

[Genetics & Medicine](#)

[Genomes & Maps](#)

[Homology](#)

[Literature](#)

[Proteins](#)

[Sequence Analysis](#)

[Taxonomy](#)

[Training & Tutorials](#)

[Variation](#)

---

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

---

**Submit**

Deposit data or manuscripts into NCBI databases



NCBI

**Download**

Transfer NCBI data to your computer



NCBI

**Learn**

Find help documents, attend a class or watch a tutorial



NCBI

---

**Develop**

Use NCBI APIs and code libraries to build applications



NCBI

**Analyze**

Identify an NCBI tool for your data analysis task



NCBI

**Research**

Explore NCBI research and collaborative projects



NCBI

---

### Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

---

### NCBI News & Blog

[NCBI Taxonomy Updates to Prokaryotes](#)  
02 Oct 2024

As previously announced, NCBI is continuing to improve our Taxonomy resource. The International Code of

[Viewing Ligand-Protein Interactions In iCn3D](#)  
30 Sep 2024

Access this content at [https://www.ncbi.nlm.nih.gov](#)

# Base de datos NCBI

The screenshot shows the NCBI Protein search interface. The search term "S100A4" is entered in the search bar. The results page displays information for the protein S100A4, also known as S100 calcium binding protein A4. Key details include:

- Summary:** 20 per page, Sort by Default order.
- Species:** Animals (83)
- Source databases:** PDB (89), RefSeq (0), UniProtKB / Swiss-Prot (0)
- Sequence length:** Custom range...
- Molecular weight:** Custom range...
- Release date:** Custom range...
- Revision date:** Custom range...

The main content area shows the protein's gene information, including its aliases (18A2, 42A, CAPL, FSP1, MTS1, P9KA, PEL98), Gene ID (6275), and links to RefSeq transcripts (2), RefSeq proteins (2), RefSeqGene (1), and PubMed (386). Buttons for Orthologs, Genome Data Viewer, and BLAST are also present.

On the right side, there are filters, results by taxon (Top Organisms: Homo sapiens (77), Danio rerio (6), unidentified (6)), find related data (Database: Select), and search details (S100A4[All Fields] AND pdb[filter]).

At the bottom, it says "Items: 1 to 20 of 89".

# Base de datos NCBI: Archivo Fasta

An official website of the United States government [Here's how you know](#) ✓

 National Library of Medicine  
National Center for Biotechnology Information

Log in

Protein   Help

Advanced

FASTA ▾ Send to: ▾ Change region shown

**Chain B, S100A4 Metastasis Factor**

PDB: 2Q91\_B  
GenPept, Identical Proteins, Graphics

> pdb|2Q91|B Chain B, S100A4 Metastasis Factor  
MACPLEKALDVMVSTFHKTSGKEGDKFKLNKSELKLLTRELPSFLGKRTDEAAFQKLMSNLDSNRDNEV  
DFQEYCVFLSCIAMMNEFFEGFPDKQPRKK

Analyze this sequence  
Run BLAST  
Identify Conserved Domains

Protein 3D Structure  
  
Structure of the Ca<sup>2+</sup>-Bound Activated Form of the S100A4 Metastasis Factor  
PDB: 2Q91  
Source: Homo sapiens  
Method: X-ray Diffraction  
Resolution: 1.63 Å

See all 15 structures...

## Basic Local Alignment Search Tool

Stephen F. Altschul<sup>1</sup>, Warren Gish<sup>1</sup>, Webb Miller<sup>2</sup>  
Eugene W. Myers<sup>3</sup> and David J. Lipman<sup>1</sup>

<sup>1</sup>*National Center for Biotechnology Information  
National Library of Medicine, National Institutes of Health  
Bethesda, MD 20894, U.S.A.*

<sup>2</sup>*Department of Computer Science  
The Pennsylvania State University, University Park, PA 16802, U.S.A.*

<sup>3</sup>*Department of Computer Science  
University of Arizona, Tucson, AZ 85721, U.S.A.*

*(Received 26 February 1990; accepted 15 May 1990)*

## Par de segmentos Maximal

Dadas las reglas de similitud, para secuencias de amino ácidos o secuencias de ADN, los autores definen un **par de segmentos Maximal (MSP)** como el par de longitudes idénticas de mayor puntaje escogidos entre dos secuencias.

## Par de segmentos Maximal

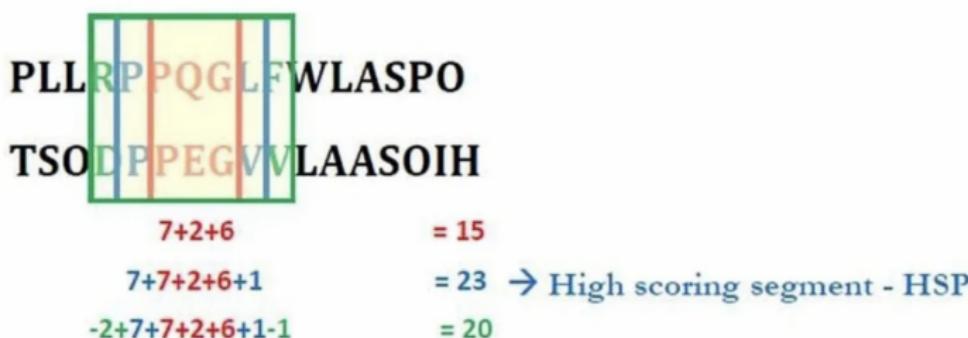
Dadas las reglas de similitud, para secuencias de amino ácidos o secuencias de ADN, los autores definen un **par de segmentos Maximal (MSP)** como el par de longitudes idénticas de mayor puntaje escogidos entre dos secuencias.

Los límites del MSP se escogen para maximizar su puntaje, por lo que el MSP puede ser de cualquier longitud.

## Par de segmentos Maximal

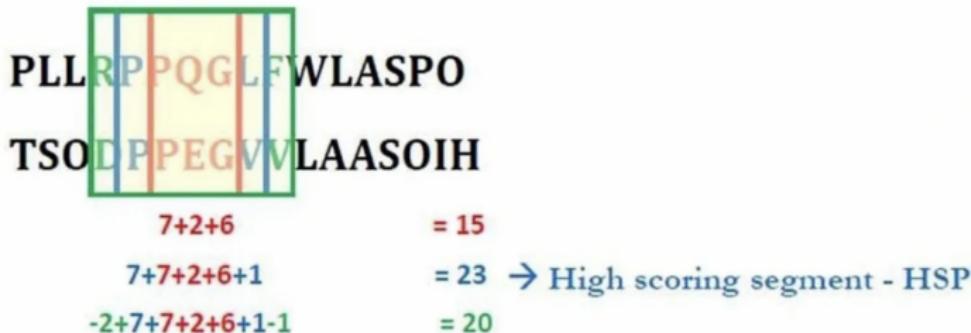
Dadas las reglas de similitud, para secuencias de amino ácidos o secuencias de ADN, los autores definen un **par de segmentos Maximal (MSP)** como el par de longitudes idénticas de mayor puntaje escogidos entre dos secuencias.

Los límites del MSP se escogen para maximizar su puntaje, por lo que el MSP puede ser de cualquier longitud.



extending hits example

## Par de segmentos Maximal



extending hits example

## Par de segmentos Maximal

PLL RPPQGLFWLASPO

TSOD D PPEGVV LAASOIH

$$7+2+6$$

$$= 15$$

$$7+7+2+6+1$$

= 23 → High scoring segment - HSP

$$-2+7+7+2+6+1-1$$

$$= 20$$

extending hits example

El puntaje MSP provee una medida de similitud local para cualquier par de secuencias.

## Par de segmentos Maximal

PLL RPPQGLFWLASPO

TSOD D PPEGVV LAASOIH

$$7+2+6$$

$$= 15$$

$$7+7+2+6+1$$

= 23 → High scoring segment - HSP

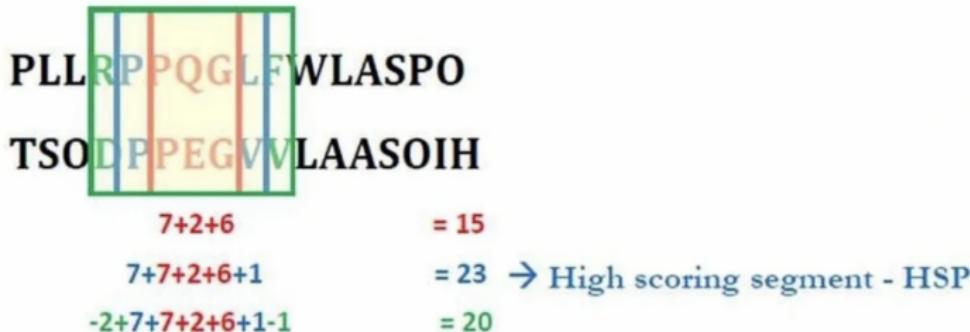
$$-2+7+7+2+6+1-1$$

$$= 20$$

extending hits example

El puntaje MSP provee una medida de similitud local para cualquier par de secuencias.

## Par de segmentos Maximal



extending hits example

**BLAST** puede buscar todos los pares de segmentos localmente maximales con puntajes sobre cierto valor de corte.

## Par de segmentos Maximal

**BLAST puede buscar** todos los pares de segmentos localmente maximales con **puntajes sobre cierto valor de corte**.

## Par de segmentos Maximal

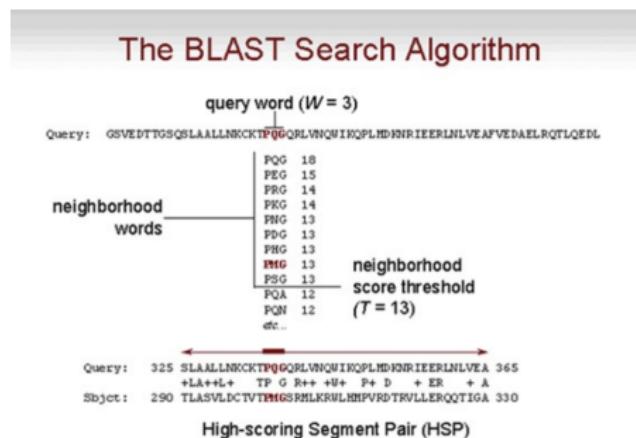
**BLAST puede buscar** todos los pares de segmentos localmente maximales con **puntajes sobre cierto valor de corte**.

BLAST Glossary

## Par de segmentos Maximal

**BLAST puede buscar** todos los pares de segmentos localmente maximales con **puntajes sobre cierto valor de corte**.

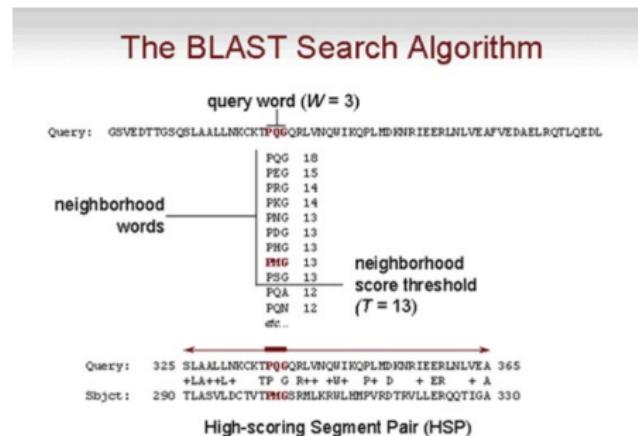
### BLAST Glossary



## Par de segmentos Maximal

**BLAST puede buscar** todos los pares de segmentos localmente maximales con **puntajes sobre cierto valor de corte**.

## BLAST Glossary



Una ventaja importante de los MSP es que **podemos estimar su significancia estadística** bajo un modelo aleatorio apropiado.

## Aproximación rápida de los puntajes MSP

En una base de datos de miles de secuencias, ¿cuántas esperas encontrar que sean **homólogas** a la secuencia de consulta (query)?

## Aproximación rápida de los puntajes MSP

En una base de datos de miles de secuencias, ¿cuántas esperas encontrar que sean **homólogas** a la secuencia de consulta (query)?

Secuencias homólogas: secuencias que tiene un origen evolutivo común.

## Aproximación rápida de los puntajes MSP

En una base de datos de miles de secuencias, ¿cuántas esperas encontrar que sean **homólogas** a la secuencia de consulta (query)?

Secuencias homólogas: secuencias que tiene un origen evolutivo común.

Nos interesa hacer una búsqueda rápida, cometiendo poco errores.

## Aproximación rápida de los puntajes MSP

En una base de datos de miles de secuencias, ¿cuántas esperas encontrar que sean **homólogas** a la secuencia de consulta (query)?

Secuencias homólogas: secuencias que tiene un origen evolutivo común.

Nos interesa hacer una búsqueda rápida, cometiendo poco errores.

**Interés:** identificar solo aquellas entradas de secuencias con puntajes MSP sobre cierto valor de corte  $S$ .

## Referencia

BROWSE · PUBLISH · ABOUT

# PLOS BIOLOGY

 OPEN ACCESS

EDUCATION

## Using BLAST to Teach “E-value-tionary” Concepts

Cheryl A. Kerfeld  Kathleen M. Scott

Published: February 1, 2011 • <https://doi.org/10.1371/journal.pbio.1001014>

## Aproximación rápida de los puntajes MSP

Considera una palabra-par (par de segmentos de longitud fija  $w$ ).

## Aproximación rápida de los puntajes MSP

Considera una palabra-par (par de segmentos de longitud fija  $w$ ).

La **principal estrategia del BLAST** es buscar solo pares de segmentos que contengan una palabra-par con un puntaje de al menos  $T$ .

# Aproximación rápida de los puntajes MSP

Considera una palabra-par (par de segmentos de longitud fija  $w$ ).

La **principal estrategia del BLAST** es buscar solo pares de segmentos que contengan una palabra-par con un puntaje de al menos  $T$ .

## 1. Generate words from sequence above threshold (e.g. T=11)

Query Sequence:

```
>gi|16329320 (residues 412 to 594)
SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTQTG
RHQGILTSWVSQASFTPPGIMLAIPEGFDAYGLAGQNKAFLNLLQEGRS
VRRHFDHQPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNI
GDHVLVYATVQAGQVLQPNGITAIRHRKSGGQY
```

Fragmentation into words:

Selection of words scoring above threshold (for word SWV):

Substitution Matrix\*

	R	G	I	K	F	S	T	W	V
R	5	0	-1	-1	-2	1	0	-3	0
G	6	-4	-2	-3	0	-2	-2	-3	
I		4	-3	0	-2	-1	-3	3	
K			5	-3	0	-1	-3	-2	
F				6	-2	-2	1	-1	
S					4	1	-3	-2	
T						5	-2	0	
W							11	-3	
V								4	

SWV (4+11+4 = 19)

SWI (4+11+3 = 18)

TWV (1+11+4 = 16)

GWV (0+11+4 = 15)

KWV (0+11+4 = 15)

SWS (4+11-2 = 13)

SFV (4+1+4 = 9)

SRV (4-3+4 = 5)

Synonyms above  
threshold 11...  
(others not shown)

Synonyms below  
threshold 11...  
(others not shown)

\*A portion of the BLOSUM 62 matrix

## Aproximación rápida de los puntajes MSP

Escaneando una secuencia, **se puede determinar rápidamente** si contiene una palabra de longitud  $w$  que puede emparejar con la secuencia de consulta para producir una palabra-par con un puntaje mayor o igual al corte  $T$ .

## Aproximación rápida de los puntajes MSP

Escaneando una secuencia, **se puede determinar rápidamente** si contiene una palabra de longitud  $w$  que puede emparejar con la secuencia de consulta para producir una palabra-par con un puntaje mayor o igual al corte  $T$ .

**Cualquiera de esos aciertos es extendido** para determinar si está contenido dentro de un par de segmentos cuyo puntaje es mayor o igual a  $S$ .

## Aproximación rápida de los puntajes MSP

Escaneando una secuencia, **se puede determinar rápidamente si** contiene una palabra de longitud  $w$  que puede emparejar con la secuencia de consulta para producir una palabra-par con un puntaje mayor o igual al corte  $T$ .

**Cualquiera de esos aciertos es extendido** para determinar si está contenido dentro de un par de segmentos cuyo puntaje es mayor o igual a  $S$ .

Selection of words scoring above threshold (for word SWV):

Substitution Matrix\*

	R	G	I	K	F	S	T	W	V
R	5	0	-1	-1	-2	1	0	-3	0
G	6	-4	-2	-3	0	-2	-2	-3	
I		4	-3	0	-2	-1	-3	3	
K			5	-3	0	-1	-3	-2	
F				6	-2	-2	1	-1	
S					4	1	-3	-2	
T						5	-2	0	
W							11	-3	
V								4	

→

SWV (4+11+4 = 19)  
 SWI (4+11+3 = 18)  
 TWV (1+11+4 = 16)  
 GWV (0+11+4 = 15)  
 K WV (0+11+4 = 15)  
 SWS (4+11-2 = 13)  
 SFV (4+1+4 = 9)  
 SRV (4-3+4 = 5)

Synonyms above threshold 11...  
(others not shown)

Synonyms below threshold 11...  
(others not shown)

\*A portion of the BLOSUM 62 matrix

### 2. Search the database for words matching those generated

### 3. Extend matching hits in both directions

RHQGILTSWVSQASFTPPGIMLAIPGEFDAYLAGQNKR...  
 ...TAMLVSWVSQASFNPPGLTIALAKE.RAEGLDHSGD

Word match from Step 1      Extension until score drops

# Aproximación rápida de los puntajes MSP

## 4. Generate alignment and calculate statistics

```
>ref|YP_002482587.1| flavin reductase domain protein FMN-binding [Cyanothece sp.  
PCC 7425]  
gb|ACL44226.1| flavin reductase domain protein FMN-binding [Cyanothece sp. PCC  
7425]
```

Length=585

Score = 176 bits (446), Expect = 1e-42, Method: Compositional matrix adjust.  
Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)

Query 1	SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTQTTGRH-----	52
	+G++FA+ L+ K+QR RQ+ E Q+DRT+QAVGRIIGS+ V+T + H	
Sbjct 393	AGSDFAQVLKKAKKQRSPRSQNSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP	452
Query 53	-----QGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFLVNLQEGRSVRRFDH	107
	+L SWVSQASF PPG+ +A+ E A GL AFVLN+L+EG ++RRHF	
Sbjct 453	QLEVPTAMLVSWVSQASFNPPLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK	511
Query 108	QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVHQ	167
	P G++ F+ L +NGC +L + LAYLEC VQS GDH L+YATV G+VHQ	
Sbjct 512	SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLHYATVNNGKVHQ	569
Query 168	PNGITAIRHRKSGGQY 183	
	P G TA++HRKSG QY	
Sbjct 570	PTGTTAVQHRKSGNQY 585	

## Aproximación rápida de los puntajes MSP

Entre más bajo sea  $T$ , mayor es la probabilidad de que un par de segmentos con puntaje al menos de  $S$  contenga una palabra-par con un puntaje de al menos  $T$ .

## Aproximación rápida de los puntajes MSP

Entre más bajo sea  $T$ , mayor es la probabilidad de que un par de segmentos con puntaje al menos de  $S$  contenga una palabra-par con un puntaje de al menos  $T$ .

**Un valor bajo de  $T$**  aumenta el número de aciertos y por lo tanto, el tiempo de ejecución del algoritmo.

## Aproximación rápida de los puntajes MSP

Entre más bajo sea  $T$ , mayor es la probabilidad de que un par de segmentos con puntaje al menos de  $S$  contenga una palabra-par con un puntaje de al menos  $T$ .

**Un valor bajo de  $T$**  aumenta el número de aciertos y por lo tanto, el tiempo de ejecución del algoritmo.

**Simulaciones aleatorias** permiten seleccionar el valor de  $T$  que balancea estas consideraciones.

## Significancia Estadística

Para calcular el puntaje en bruto al comparar dos secuencias, se utiliza la fórmula:

$$S = \left( \sum M_{ij} \right) - cO - dG$$

donde  $M$  es el puntaje proveniente de la matrix de similitud,  $c$  es el número de *huecos o espacios*,  $O$  es la penalización por la existencia de un espacio,  $d$  es la longitud total de los espacios,  $G$  es la penalización por residuo por extender el espacio.

## Significancia Estadística

Para calcular el puntaje en bruto al comparar dos secuencias, se utiliza la fórmula:

$$S = \left( \sum M_{ij} \right) - cO - dG$$

donde  $M$  es el puntaje proveniente de la matrix de similitud,  $c$  es el número de *huecos o espacios*,  $O$  es la penalización por la existencia de un espacio,  $d$  es la longitud total de los espacios,  $G$  es la penalización por residuo por extender el espacio.

En general, este puntaje puede cambiar según las matrices y las penalizaciones utilizadas, por lo que para calcular probabilidades, se utiliza la expresión:

$$S' = (\lambda S - \ln K) / \ln 2$$

donde  $\lambda$  y  $K$  cuantifica dicha variación. ( $S'$  se conoce como el puntaje de bits)

## Significancia Estadística: el E-valor

El **E-valor** es el número esperado de secuencias por encontrar en la base de datos con un puntaje de bits igual o mayor que el calculado por el alineamiento entre la secuencia de consulta y la secuencia a comparar (subject), basado en puro azar.

## Significancia Estadística: el E-valor

El **E-valor** es el número esperado de secuencias por encontrar en la base de datos con un puntaje de bits igual o mayor que el calculado por el alineamiento entre la secuencia de consulta y la secuencia a comparar (subject), basado en puro azar.

$$E = Knme^{-\lambda S}$$

## Significancia Estadística: el E-valor

El **E-valor** es el número esperado de secuencias por encontrar en la base de datos con un puntaje de bits igual o mayor que el calculado por el alineamiento entre la secuencia de consulta y la secuencia a comparar (subject), basado en puro azar.

$$E = Knme^{-\lambda S}$$

Los E-valores de secuencias (subject) que son muy similares a la secuencia de consulta serán muy pequeños, y se utilizan para valorar la confianza con la que se afirma que la secuencia (subject) y la secuencia de consulta son homólogas.

## Significancia Estadística: el E-valor

**A medida que el e-valor decrece, la significancia biológica aumenta.**

## Significancia Estadística: el E-valor

**A medida que el e-valor decrece, la significancia biológica aumenta.**

En la práctica, es usual consider el caso  $E < 0.00001$  como convincente de que dos secuencias son homólogas.

# BLAST

*¿Cómo escoger w y T?*

## BLAST

¿Cómo escoger w y T?

**Table 1**  
*The probability of a hit at various settings of the parameters w and T, and the proportion of random MSPs missed by BLAST*

w	T	Probability of a hit $\times 10^3$	Linear regression $-\ln(q) = aS + b$		Implied % of MSPs missed by BLAST when S equals						
			a	b	45	50	55	60	65	70	75
3	11	253	0.1236	-1.005	1	1	0	0	0	0	0
	12	147	0.0875	-0.746	4	3	2	1	1	0	0
	13	83	0.0625	-0.570	11	8	6	4	3	2	2
	14	48	0.0463	-0.461	20	16	12	10	8	6	5
	15	26	0.0328	-0.353	33	28	23	20	17	14	12
	16	14	0.0232	-0.263	46	41	36	32	29	26	23
	17	7	0.0158	-0.191	59	55	51	47	43	40	37
	18	4	0.0109	-0.137	70	67	63	60	57	54	51
4	13	127	0.1192	-1.278	2	1	1	0	0	0	0
	14	78	0.0904	-1.012	5	3	2	1	1	0	0
	15	47	0.0686	-0.802	10	7	5	4	3	2	1
	16	28	0.0519	-0.634	18	14	11	8	6	5	4
	17	16	0.0390	-0.498	28	23	19	16	13	11	9
	18	9	0.0200	-0.387	40	35	30	26	22	19	17
	19	5	0.0215	-0.298	51	46	41	37	33	30	27
	20	3	0.0159	-0.234	62	57	53	49	45	41	38
5	15	64	0.1137	-1.525	3	2	1	1	0	0	0
	16	40	0.0882	-1.207	6	4	3	2	1	1	0
	17	25	0.0679	-0.939	12	9	6	4	3	2	2
	18	15	0.0529	-0.754	20	15	12	9	7	5	4
	19	9	0.0413	-0.608	29	23	19	15	13	10	8
	20	5	0.0327	-0.506	38	32	28	23	20	17	14
	21	3	0.0257	-0.420	48	42	37	32	29	25	22

## Opciones en el BLAST

- blastn: nucleótido vs nucleótido
- blastp: proteína vs proteína
- mas tipos

The screenshot shows the BLAST homepage with a dark blue header. The NIH logo and "National Library of Medicine" are on the left, and "National Center for Biotechnology Information" is below it. On the right are "Log in", "Home", "Recent Results", "Saved Strategies", and "Help". A green sidebar on the left has "NEWS" at the top. The main content area features a "Basic Local Alignment Search Tool" section with a brief description and a "Learn more" link. To the right is a news box for "BLAST+ 2.15.0 is here!" with a "More BLAST news..." link. Below these are three large buttons for "Web BLAST": "Nucleotide BLAST" (nucleotide ➔ nucleotide), "blastx" (translated nucleotide ➔ protein), and "tblastn" (protein ➔ translated nucleotide). To the right is a "Protein BLAST" button (protein ➔ protein).

Comparando Secuencias: Valorando la Similitud  
¿Cómo obtengo las secuencias?  
(Scoring) Matrices de Puntaje

## BLASTN

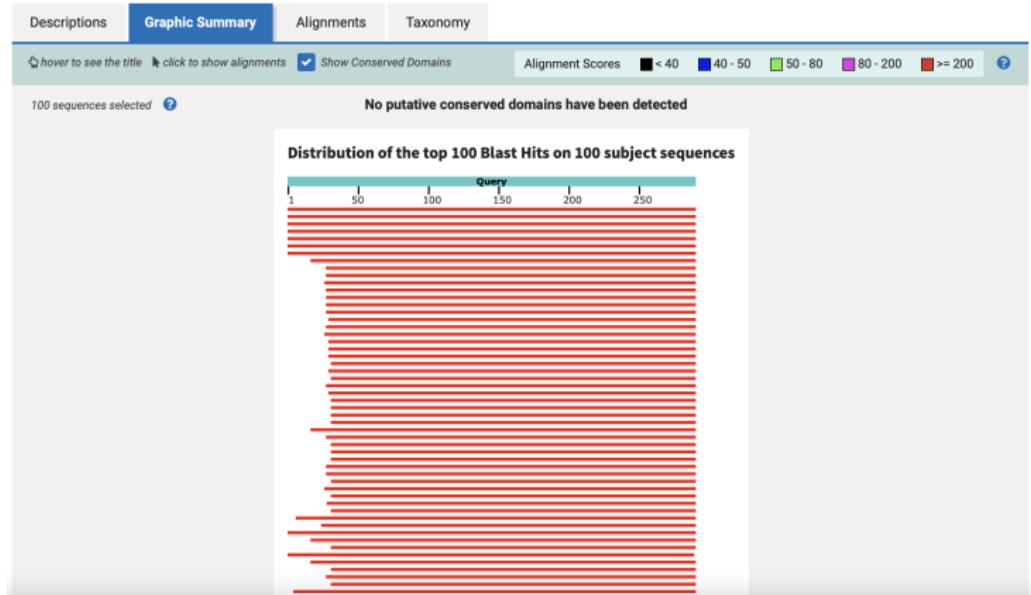
select all 78 sequences selected

			GenBank	Graphics	Distance tree of results	MSA Viewer			
Description		Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	<a href="#">Piscinibacter gummiphilus strain NBRC 109400 chromosome, complete genome</a>	Piscinibacter gummiphilus	1624	1624	100%	0.0	100.0%	6398100	CP024645.1
<input checked="" type="checkbox"/>	<a href="#">Piscinibacter gummiphilus strain NS21, complete genome</a>	Piscinibacter gummiphilus	1624	1624	100%	0.0	100.0%	6398096	CP015118.1
<input checked="" type="checkbox"/>	<a href="#">Piscinibacter gummiphilus strain SBD 7-3 chromosome, complete genome</a>	Piscinibacter gummiphilus	643	1008	84%	2e-179	82.46%	5594180	CP136336.1
<input checked="" type="checkbox"/>	<a href="#">Piscinibacter gummiphilus strain SBD 7-3 plasmid unnamed1, complete sequence</a>	Piscinibacter gummiphilus	593	593	84%	2e-164	81.28%	200976	CP136337.1
<input checked="" type="checkbox"/>	<a href="#">Acidovorax delafieldii pbsA gene for PBS(A) depolymerase, complete cds</a>	Acidovorax delafieldii	529	529	84%	6e-145	79.74%	915	AB068349.1
<input checked="" type="checkbox"/>	<a href="#">Synthetic construct fast-polyethylene terephthalate hydrolyase gene, partial cds</a>	synthetic construct	398	398	82%	2e-105	76.93%	795	DR020855.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces ferrugineus strain CCTCC AA2014009 chromosome, complete genome</a>	Streptomyces ferrugineus	71.3	71.3	9%	5e-07	82.50%	9959088	CP063373.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. 11x1 chromosome, complete genome</a>	Streptomyces sp. 11x1	67.6	67.6	6%	6e-06	88.89%	10541094	CP122458.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. NBC_01262 chromosome, complete genome</a>	Streptomyces sp. NBC_01262	62.1	62.1	7%	3e-04	83.58%	9736085	CP108462.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. NBC_00285 chromosome, complete genome</a>	Streptomyces sp. NBC_00285	60.2	60.2	4%	0.001	92.86%	10263655	CP108055.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces canus strain NBC_00866 chromosome, complete genome</a>	Streptomyces canus	60.2	60.2	4%	0.001	92.86%	10632801	CP108818.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. NBC_00882 chromosome, complete genome</a>	Streptomyces sp. NBC_00882	60.2	60.2	4%	0.001	92.86%	10657798	CP108797.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. NBC_01478 chromosome, complete genome</a>	Streptomyces sp. NBC_01478	58.4	58.4	4%	0.004	92.68%	12124816	CP109444.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. NBC_01261 chromosome, complete genome</a>	Streptomyces sp. NBC_01261	58.4	58.4	4%	0.004	92.68%	11585837	CP108463.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. NBC_01622 chromosome, complete genome</a>	Streptomyces sp. NBC_01622	58.4	58.4	4%	0.004	94.59%	11696066	CP109293.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. NBC_00989 chromosome, complete genome</a>	Streptomyces sp. NBC_00989	58.4	58.4	4%	0.004	92.68%	11950319	CP108728.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces sp. NBC_00988 chromosome, complete genome</a>	Streptomyces sp. NBC_00988	58.4	58.4	4%	0.004	92.68%	12375414	CP108730.1
<input checked="" type="checkbox"/>	<a href="#">Streptomyces prunicolor strain NBC_01021 chromosome, complete genome</a>	Streptomyces prunicolor	58.4	58.4	4%	0.004	92.68%	10948847	CP108678.1

## BLASTP

Descriptions		Graphic Summary	Alignments	Taxonomy										
Sequences producing significant alignments										Download	Select columns	Show	100	?
<input checked="" type="checkbox"/> select all 100 sequences selected			GenPept	Graphics	Distance tree of results			Multiple alignment		MSA Viewer				
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	590	590	100%	0.0	100.00%	298	EEQD_A					
<input checked="" type="checkbox"/>	dienelactone hydrolase family protein [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	589	589	100%	0.0	100.00%	290	WP_054022242.1					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	584	584	100%	0.0	99.31%	298	ZOSB_A					
<input checked="" type="checkbox"/>	dienelactone hydrolase family protein [Rhizobacter sp.]	Rhizobacter sp.	577	577	100%	0.0	97.93%	290	MBX36255601.1					
<input checked="" type="checkbox"/>	dienelactone hydrolase family protein [Piscinibacter gummiphilus]	Piscinibacter gummiphilus	575	575	100%	0.0	97.24%	290	WP_316704339.1					
<input checked="" type="checkbox"/>	dienelactone hydrolase family protein [Piscinibacter gummiphilus]	Piscinibacter gummiphilus	567	567	100%	0.0	95.52%	290	WP_316702295.1					
<input checked="" type="checkbox"/>	Chain A, PET hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	566	566	100%	0.0	96.55%	298	OKYS_A					
<input checked="" type="checkbox"/>	IspETase-Catcher [synthetic construct]	synthetic construct	548	548	94%	0.0	96.35%	525	WCH76318.1					
<input checked="" type="checkbox"/>	IspETase-Cat [synthetic construct]	synthetic construct	545	545	90%	0.0	100.00%	467	WCH76319.1					
<input checked="" type="checkbox"/>	IspETase-Spy [synthetic construct]	synthetic construct	541	541	90%	0.0	100.00%	405	WCH76316.1					
<input checked="" type="checkbox"/>	IspETase-Tag [synthetic construct]	synthetic construct	540	540	91%	0.0	99.62%	313	WCH76317.1					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	539	539	90%	0.0	100.00%	272	6ANE_A					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	538	538	90%	0.0	100.00%	270	6ILW_A					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	538	538	90%	0.0	100.00%	282	8GU4_A					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	535	535	90%	0.0	99.62%	270	6ILX_A					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	533	533	90%	0.0	100.00%	268	5XG0_A					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	533	533	90%	0.0	99.24%	272	8J17_A					
<input checked="" type="checkbox"/>	Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	533	533	91%	0.0	98.86%	272	5YFE_A					

# BLASTP



## BLASTP

Descriptions   Graphic Summary   **Alignments**   Taxonomy

Alignment view   Pairwise   ?   Restore defaults

100 sequences selected   ?

Download ▾   GenPept Graphics

**Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]**

Sequence ID: [6EQD\\_A](#) Length: 298 Number of Matches: 1

[See 10 more title\(s\)](#) ▾   [See all Identical Proteins\(IPG\)](#)

Range 1: 1 to 290   [GenPept](#)   [Graphics](#)   [▼ Next Match](#)   [▲ Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
590 bits(1521)	0.0	Compositional matrix adjust.	290/290(100%)	290/290(100%)	0/290(0%)
<hr/>					
Query 1	MNFPRASRLMQAAVLGGMLMAVSAAAATQTNPYARGNPNTAASLEASAGPFTVRSFTVSRP		60		
Sbjct 1	MNFPRASRLMQAAVLGGMLMAVSAAAATQTNPYARGNPNTAASLEASAGPFTVRSFTVSRP		60		
Query 61	SGYAGTVYYPTNAAGGTGVAIAIVPGYTAQQSIKWWGPRLASHGFVVITIDNSTLDQP		120		
Sbjct 61	SGYAGTVYYPTNAAGGTGVAIAIVPGYTAQQSIKWWGPRLASHGFVVITIDNSTLDQP		120		
Query 121	SSRSSQ0MAALRQVASLNGTSSSPIYGKVDTARMGVGMWSMGGGSLISAANPSLKAAA		180		
Sbjct 121	SSRSSQ0MAALRQVASLNGTSSSPIYGKVDTARMGVGMWSMGGGSLISAANPSLKAAA		180		
Query 181	PQAPWDSSTNFSSVTPTLIFACENDSIAPIVNNSALPIYDSMSRNNAQFLEINGSHSCA		240		
Sbjct 181	PQAPWDSSTNFSSVTPTLIFACENDSIAPIVNNSALPIYDSMSRNNAQFLEINGSHSCA		240		
Query 241	NSGNNSNQALIGKKGVAVMKRFDMDNTRYSTFACENPNSTRVSDFRANC斯 290				
Sbjct 241	NSGNNSNQALIGKKGVAVMKRFDMDNTRYSTFACENPNSTRVSDFRANC斯 290				

(Scoring) Matrices de Puntaje

## Calculando Probabilidades

¿Cuál es la probabilidad de haber nacido en Enero?

## Calculando Probabilidades

¿Cuál es la probabilidad de haber nacido en Enero?

¿Cuál es la probabilidad de haber nacido en Febrero?

## Calculando Probabilidades

¿Cuál es la probabilidad de haber nacido en Enero?

¿Cuál es la probabilidad de haber nacido en Febrero?

¿Cuál es la probabilidad de haber nacido el 4 de julio?

## Calculando Probabilidades

¿Cuál es la probabilidad de haber nacido en Enero?

¿Cuál es la probabilidad de haber nacido en Febrero?

¿Cuál es la probabilidad de haber nacido el 4 de julio?

¿Cuál es la probabilidad de haber nacido el 29 de Febrero?

## Calculando Probabilidades

¿Cuál es la probabilidad de haber nacido en Enero?

¿Cuál es la probabilidad de haber nacido en Febrero?

¿Cuál es la probabilidad de haber nacido el 4 de julio?

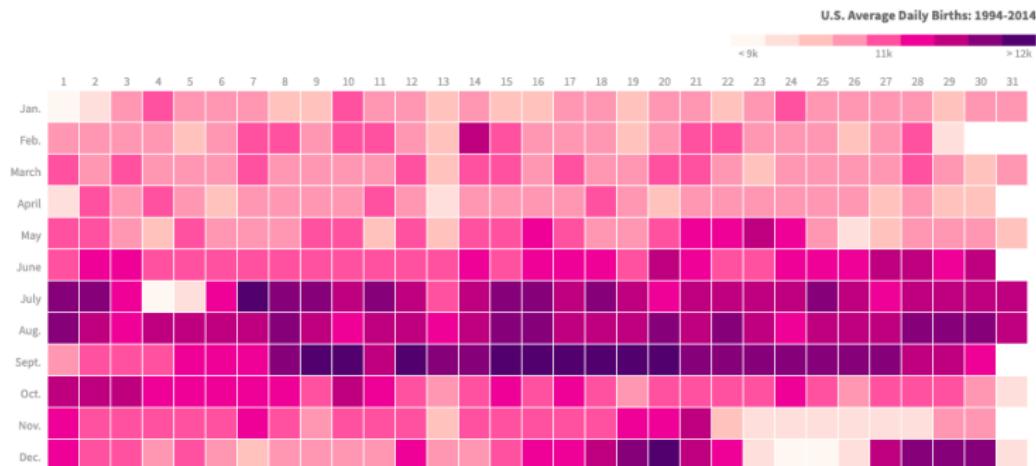
¿Cuál es la probabilidad de haber nacido el 29 de Febrero?

¿Cuál es la probabilidad de haber nacido el 31 de Abril?

# Calculando Probabilidades

## HOW POPULAR IS YOUR BIRTHDAY?

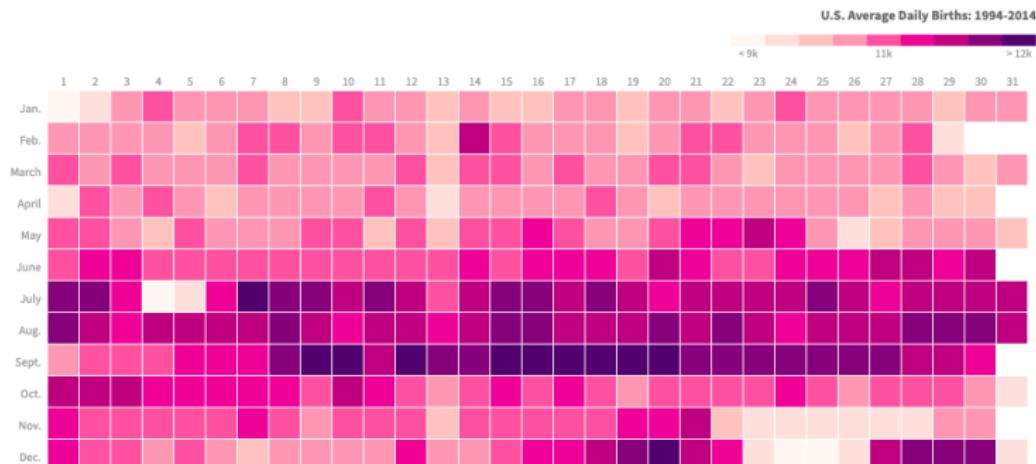
Two decades of American birthdays, averaged by month and day.



# Calculando Probabilidades

## HOW POPULAR IS YOUR BIRTHDAY?

Two decades of American birthdays, averaged by month and day.



Para responder a estas preguntas, es más confiable utilizar un enfoque empírico para calcular probabilidades

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) para encontrar un motivo de secuencia en común. (El motivo es una region de una proteína que posee una estrutura específica).

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) para encontrar un motivo de secuencia en común. (El motivo es una region de una proteína que posee una estrutura específica). Para calcular la matriz BLOSUM, se utiliza la siguiente ecuación:

$$S_{ij} = \frac{1}{\lambda} \log \left( \frac{p_{ij}}{q_i \cdot q_j} \right)$$

donde

$p_{ij}$  ≡ probabilidad de que dos amino ácidos  $i$  y  $j$  se reemplacen uno al otro en una secuencia homologa.

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) para encontrar un motivo de secuencia en común. (El motivo es una region de una proteína que posee una estrutura específica). Para calcular la matriz BLOSUM, se utiliza la siguiente ecuación:

$$S_{ij} = \frac{1}{\lambda} \log \left( \frac{p_{ij}}{q_i \cdot q_j} \right)$$

donde

$p_{ij}$  ≡ probabilidad de que dos amino ácidos  $i$  y  $j$  se reemplacen uno al otro en una secuencia homologa.

$q_i$  ≡ prob. de encontrar el amino ácido  $i$  en alguna secuencia de proteínas.

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) para encontrar un motivo de secuencia en común. (El motivo es una region de una proteína que posee una estrutura específica). Para calcular la matriz BLOSUM, se utiliza la siguiente ecuación:

$$S_{ij} = \frac{1}{\lambda} \log \left( \frac{p_{ij}}{q_i \cdot q_j} \right)$$

donde

$p_{ij}$  ≡ probabilidad de que dos amino ácidos  $i$  y  $j$  se reemplacen uno al otro en una secuencia homologa.

$q_i$  ≡ prob. de encontrar el amino ácido  $i$  en alguna secuencia de proteínas.

$q_j$  ≡ prob. de encontrar el amino ácido  $j$  en alguna secuencia de proteínas.

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) para encontrar un motivo de secuencia en común. (El motivo es una region de una proteína que posee una estrutura específica). Para calcular la matriz BLOSUM, se utiliza la siguiente ecuación:

$$S_{ij} = \frac{1}{\lambda} \log \left( \frac{p_{ij}}{q_i \cdot q_j} \right)$$

donde

$p_{ij}$  ≡ probabilidad de que dos amino ácidos  $i$  y  $j$  se reemplacen uno al otro en una secuencia homologa.

$q_i$  ≡ prob. de encontrar el amino ácido  $i$  en alguna secuencia de proteínas.

$q_j$  ≡ prob. de encontrar el amino ácido  $j$  en alguna secuencia de proteínas.

$\lambda$  ≡ factor de escala, determinado de manera que la matriz contenga valores enteros que faciliten los cálculos.

## Calculando Probabilidades

## Substitution Matrix\*

	R	G	I	K	F	S	T	W	V
R	5	0	-1	-1	-2	1	0	-3	0
G		6	-4	-2	-3	0	-2	-2	-3
I			4	-3	0	-2	-1	-3	3
K				5	-3	0	-1	-3	-2
F					6	-2	-2	1	-1
S						4	1	-3	-2
T							5	-2	0
W								11	-3
V									4

\*A portion of the BLOSUM 62 matrix