

## La Estadística detrás del Análisis Filogético

Hugo Flores Arguedas

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

Escuela de Otono, Biología Matemática, Oct 7-11, 2024



# ¿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. Mahendrarajah](#), [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

## 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?

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

8 matches

2 mismatches

## 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  
ASFTPPRGRKRSWTIG

## 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

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

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](#)

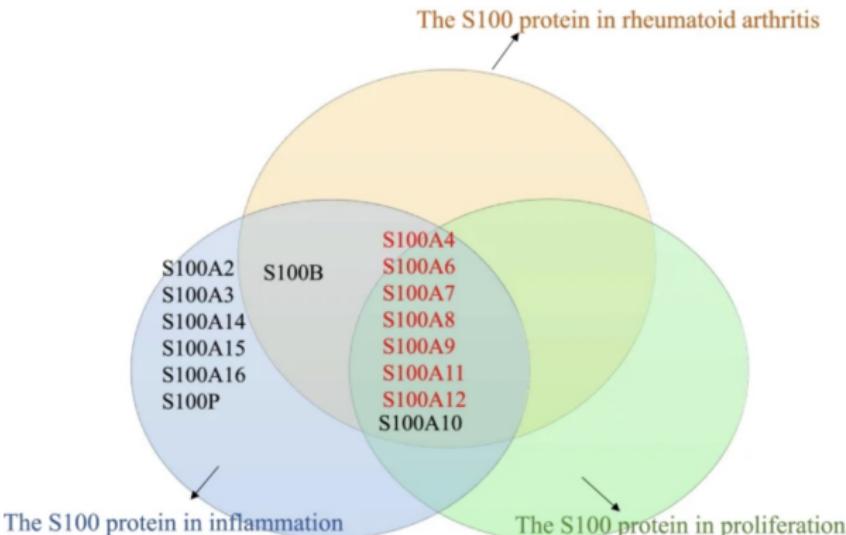
4418 Accesses | 23 Citations | [Metrics](#)

### 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

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



Develop

Use NCBI APIs and code libraries to build applications



**Download**

Transfer NCBI data to your computer



Analyze

Identify an NCBI tool for your data analysis task



**Learn**

Find help documents, attend a class or watch a tutorial



Research

Explore NCBI research and collaborative projects



---

### 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

A new interface for viewing ligand-protein interactions in iCn3D.

NCBI

# Base de datos NCBI

National Library of Medicine  
National Center for Biotechnology Information

Protein      Protein: S100A4      Search      Log in

Species: Summary 20 per page Sort by Default order      Send to: Filters: Manage Filters

Source databases: PDB (89) RefSeq (0) UniProtKB / Swiss-Prot (0) Customize...      clear

Results by taxon: Top Organisms [Tree]  
Homo sapiens (77)  
Danio rerio (6)  
unidentified (6)

Find related data: Database: Select      Find items

Search details: S100A4 [All Fields] AND pdb[filter]

RefSeq Sequences

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

FASTA ▾ Send to: ▾

**Chain B, S100A4 Metastasis Factor**

PDB: 2Q91\_B

[GenPept](#) [Identical Proteins](#) [Graphics](#)

> pdb|2Q91|B Chain B, S100A4 Metastasis Factor  
MACPLEKALDWIVMVFHFKYSGKEGDKFKLNKSELKELLTRELPSFLGKRTDEAAFQKLMNSNLDNRDNEV  
DFQEYCVFLSCIAMMNEFFEGFPDKOPRKK

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.

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

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

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

## 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



PLOS BIOLOGY

The logo for PLOS BIOLOGY features the word "PLOS" in a black sans-serif font and "BIOLOGY" in a larger blue sans-serif font. Above the "PLOS" part, there is a small circular icon containing a stylized DNA helix. To the right of the main title, there are three smaller links: "PUBLISH", "SUBMIT", and "ABOUT".

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)
SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVTTQTTG
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFLVNLQEGRS
VRRHFDHQPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQWSNI
GDHVLYATVQAGQVLQPNGITAIRHRKSGGQY
```

Fragmentation into words:

SWVSQASFTPPGIM → SWV WVS VSQ SQA QAS ASF SFT ...  


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

	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

SWV (4+11+4 = 19)	Synonyms above threshold 11... (others not shown)
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)	Synonyms below threshold 11... (others not shown)
SFV (4+1+4 = 9)	
SRV (4-3+4 = 5)	

## 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	

→ 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**

Word match from Step 1      Extension until score drops

RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKR  
 .| | | | | | | | :| :| | | | |  
 ::TAMLVSWVSQASFNPPGLTIALAKE.RAEGLDHSGD

## 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      SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGHVTTQTTGRH----- 52  
       +G++FA+ L+ K+QR RQ+ E Q+DRT+QAVGRIIGS+ V+T + H  
Sbjct 393    AGSDFAQVLKKAKKQRSPRQSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTIPHPEVEEP 452  
  
Query 53     -----QGILTSWVQSASFTPPGIMLAIPGEFDAYGLAGQNKAFLVNLLQEGRSVRRHFDH 107  
       +L SWVSQASF PPG+ +A+ E A GL AFVLN+L+EG ++RRHF  
Sbjct 453    QLEVPTAMLVSWVSQASFNFNPPGLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK 511  
  
Query 108    QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVLO 167  
       P G++ F+ L +NGC +L + LAYLEC VQS GDH L+YATV G+VLO  
Sbjct 512    SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVNNNGKVLQ 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  $T$**  aumento 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  $T$**  aumento 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:

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

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:

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

donde  $M$  es el puntaje proveniente de la matrix de similitud,  $c$  es el números 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:

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

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-values por secuencia (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

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-values por secuencia (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.

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

# 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 $\times 10^5$	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	-0.401	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	0.1102	-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	-0.408	28	23	19	16	13	11	0
	18	9	0.0290	-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 National Library of Medicine BLAST homepage. At the top, there's a header with the NIH logo, "National Library of Medicine", and "National Center for Biotechnology Information". On the right side of the header are links for "Log in", "Home", "Recent Results", "Saved Strategies", and "Help". Below the header, the word "BLAST®" is displayed. In the center, there's a section titled "Basic Local Alignment Search Tool" with a brief description of what BLAST does. To the right of this, a box contains news about BLAST+ 2.15.0, mentioning two new features and the release date (Tue, 28 Nov 2023). Below this, there's a link to "More BLAST news...". At the bottom, there are three main search tool options: "Nucleotide BLAST" (nucleotide ➤ nucleotide), "blastx" (translated nucleotide ➤ protein), and "tblastn" (protein ➤ translated nucleotide). To the right of these is "Protein BLAST" (protein ➤ protein).

# BLASTN

select all 18 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"/>	<i>Piscinibacter gummiphilus</i> strain NBRC 109400 chromosome, complete genome	<i>Piscinibacter gummiphilus</i>	1624	1624	100%	0.0	100.0%	6398100	CP024645.1
<input checked="" type="checkbox"/>	<i>Piscinibacter gummiphilus</i> strain NS21, complete genome	<i>Piscinibacter gummiphilus</i>	1624	1624	100%	0.0	100.0%	6398096	CP151118.1
<input checked="" type="checkbox"/>	<i>Piscinibacter gummiphilus</i> strain SBD 7-3 chromosome, complete genome	<i>Piscinibacter gummiphilus</i>	643	1008	84%	2e-179	82.46%	5594180	CP136336.1
<input checked="" type="checkbox"/>	<i>Piscinibacter gummiphilus</i> strain SBD 7-3 plasmid unnamed1, complete sequence	<i>Piscinibacter gummiphilus</i>	593	593	84%	2e-164	81.28%	200976	CP136337.1
<input checked="" type="checkbox"/>	<i>Acidovorax delafieldii</i> phbA gene for PBS(A) depolymerase, complete cds	<i>Acidovorax delafieldii</i>	529	529	84%	6e-145	79.74%	915	AB06349.1
<input checked="" type="checkbox"/>	Synthetic construct fast-polyethylene terephthalate hydrolase gene, partial cds	synthetic construct	398	398	82%	2e-105	76.93%	795	OR020855.1
<input checked="" type="checkbox"/>	<i>Streptomyces ferrugineus</i> strain CCTCC AA2014009 chromosome, complete genome	<i>Streptomyces ferrugineus</i>	71.3	71.3	9%	5e-07	82.50%	9859088	CP063373.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. 11x1 chromosome, complete genome	<i>Streptomyces</i> sp. 11x1	67.6	67.6	6%	6e-09	88.89%	10541094	CP122458.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. NBC_01262 chromosome, complete genome	<i>Streptomyces</i> sp. NBC_01262	62.1	62.1	7%	3e-04	83.58%	9736065	CP108462.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. NBC_00285 chromosome, complete genome	<i>Streptomyces</i> sp. NBC_00285	60.2	60.2	4%	0.001	92.86%	10263655	CP108055.1
<input checked="" type="checkbox"/>	<i>Streptomyces canus</i> strain NBC_00868 chromosome, complete genome	<i>Streptomyces canus</i>	60.2	60.2	4%	0.001	92.86%	10632801	CP108818.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. NBC_00882 chromosome, complete genome	<i>Streptomyces</i> sp. NBC_00882	60.2	60.2	4%	0.001	92.86%	10957798	CP108797.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. NBC_01478 chromosome, complete genome	<i>Streptomyces</i> sp. NBC_01478	58.4	58.4	4%	0.004	92.68%	12124816	CP109444.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. NBC_01261 chromosome, complete genome	<i>Streptomyces</i> sp. NBC_01261	58.4	58.4	4%	0.004	92.68%	11585837	CP108463.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. NBC_01622 chromosome, complete genome	<i>Streptomyces</i> sp. NBC_01622	58.4	58.4	4%	0.004	94.59%	11969068	CP109293.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. NBC_00989 chromosome, complete genome	<i>Streptomyces</i> sp. NBC_00989	58.4	58.4	4%	0.004	92.68%	11950319	CP108728.1
<input checked="" type="checkbox"/>	<i>Streptomyces</i> sp. NBC_00988 chromosome, complete genome	<i>Streptomyces</i> sp. NBC_00988	58.4	58.4	4%	0.004	92.68%	12375414	CP108730.1
<input checked="" type="checkbox"/>	<i>Streptomyces prunicolor</i> strain NBC_01021 chromosome, complete genome	<i>Streptomyces prunicolor</i>	58.4	58.4	4%	0.004	92.68%	10948847	CP108678.1

# BLASTP

Safari File Edit View History Bookmarks Window Help

blast.ncbi.nlm.nih.gov Tue 27 Feb 2:53

Descriptions Graphic Summary Alignments Taxonomy

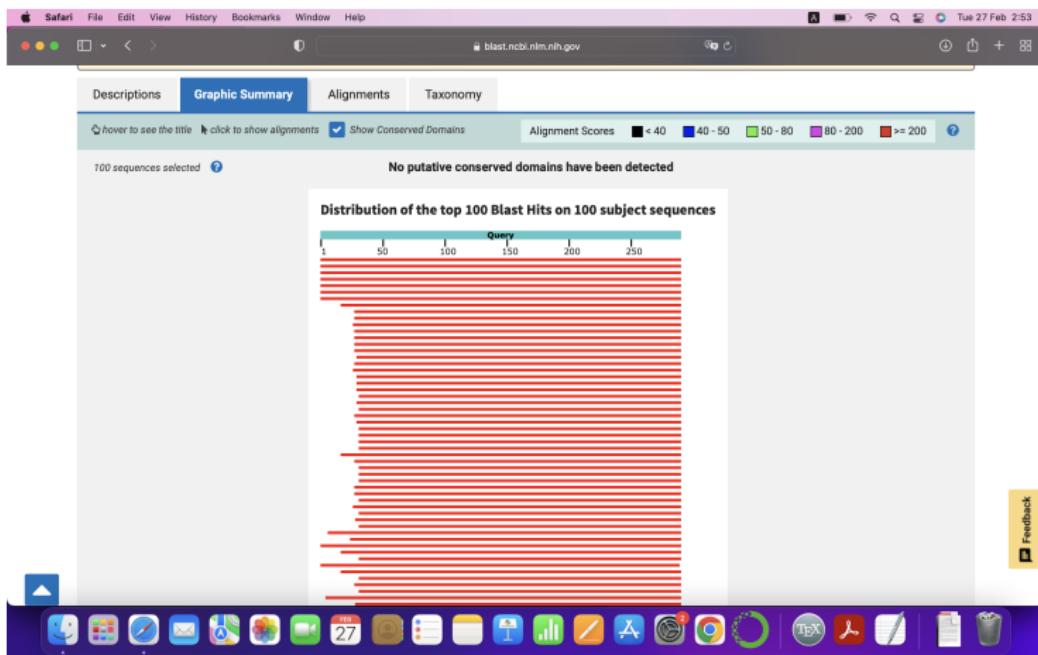
Sequences producing significant alignments Download Select columns Show 100 ↴

select all 100 sequences selected

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident.	Acc. Len	Accession
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	590	590	100%	0.0	100.0%	298	EEQD_A
deneclactone hydrolase family protein [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	589	589	100%	0.0	100.0%	290	WP_054022242_1
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	584	584	100%	0.0	99.31%	298	7QGB_A
deneclactone hydrolase family protein [Rhizobacter sp.]	Rhizobacter sp.	577	577	100%	0.0	97.83%	290	MBX3625691_1
deneclactone hydrolase family protein [Piscinibacter gummichilus]	Piscinibacter gummichilus	575	575	100%	0.0	97.24%	290	WP_316704339_1
deneclactone hydrolase family protein [Piscinibacter gummichilus]	Piscinibacter gummichilus	567	567	100%	0.0	95.52%	290	WP_316702395_1
Chain A. PET hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	566	566	100%	0.0	96.55%	298	6KY5_A
IsPETase-Catcher [synthetic construct]	synthetic construct	548	548	94%	0.0	96.35%	525	WCH76318_1
IsPETase-Cat [synthetic construct]	synthetic construct	546	545	90%	0.0	100.0%	467	WCH76319_1
IsPETase-Soy [synthetic construct]	synthetic construct	541	541	90%	0.0	100.0%	405	WCH76316_1
IsPETase-Tag [synthetic construct]	synthetic construct	540	540	91%	0.0	99.62%	313	WCH76317_1
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	539	539	90%	0.0	100.0%	272	6ANE_A
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	538	538	90%	0.0	100.0%	270	6LWJ_A
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	538	538	90%	0.0	100.0%	282	8GU4_A
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	535	535	90%	0.0	99.62%	270	6LXJ_A
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	533	533	90%	0.0	100.0%	268	8XG0_A
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	533	533	90%	0.0	99.24%	272	8J17_A
Chain A. Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]	Piscinibacter sakaiensis	533	533	91%	0.0	98.86%	272	8YFE_A

Feedback

# BLASTP



# BLASTP

The screenshot shows a BLASTP search results page from NCBI. The search query is "Chain A, Poly(ethylene terephthalate) hydrolase [Piscinibacter sakaiensis]". The results table displays 100 sequences selected, with the first few entries shown below:

Score	Expect	Method	Identites	Positives	Gaps
590 bits(1521)	0.0	Compositional matrix adjust.	290/290(100%)	290/290(100%)	0/290(0%)
Query 1	MNIFPRASRLM	AAVLLGGLMAVSAAAATQDTPYARGNPITAASLESAGPFTVRSGFTVSRP	68		
Sbjct 1	MNIFPRASRLM	AAVLLGGLMAVSAAAATQDTPYARGNPITAASLESAGPFTVRSGFTVSRP	68		
Query 61	SGYAGTVYYP	TNAAGTVGAIAIVPGYTAQDOSKIKWMPRLASHGFWVITIDTNSTLDP	128		
Sbjct 61	SGYAGTVYYP	TNAAGTVGAIAIVPGYTAQDOSKIKWMPRLASHGFWVITIDTNSTLDP	128		
Query 123	SSRSQQMALRQDV	ASLNGTTSSPPIYGKVDTARMGVGMWSGGGGSLISAAANPSLKAA	188		
Sbjct 123	SSRSQQMALRQDV	ASLNGTTSSPPIYGKVDTARMGVGMWSGGGGSLISAAANPSLKAA	188		
Query 181	PQAPMDSTNSF	SVTVPTLIFACENDSIAPIVNSALPIYDMSMRNAKQFLLEINGGHS	248		
Sbjct 181	PQAPMDSTNSF	SVTVPTLIFACENDSIAPIVNSALPIYDMSMRNAKQFLLEINGGHS	248		
Query 241	NSGNNSQALIG	KKGKVAVMKRPMNDTRYSTFACEPNSTRVSDFRTANC	290		
Sbjct 241	NSGNNSQALIG	KKGKVAVMKRPMNDTRYSTFACEPNSTRVSDFRTANC	290		

Below the table, there are sections for "Related Information", "Structure - 3D structure displays", and "Identical Proteins - Identical proteins to 6EQD\_A". The bottom of the screen shows the Mac OS X dock with various application icons.

## Scoring Matrices

## Determining Probabilities

What is the probability of being born in January?

## Determining Probabilities

What is the probability of being born in January?

What is the probability of being born in February?

## Determining Probabilities

What is the probability of being born in January?

What is the probability of being born in February?

What is the probability of being born in July 4th?

## Determining Probabilities

What is the probability of being born in January?

What is the probability of being born in February?

What is the probability of being born in July 4th?

What is the probability of being born in February 29th?

## Determining Probabilities

What is the probability of being born in January?

What is the probability of being born in February?

What is the probability of being born in July 4th?

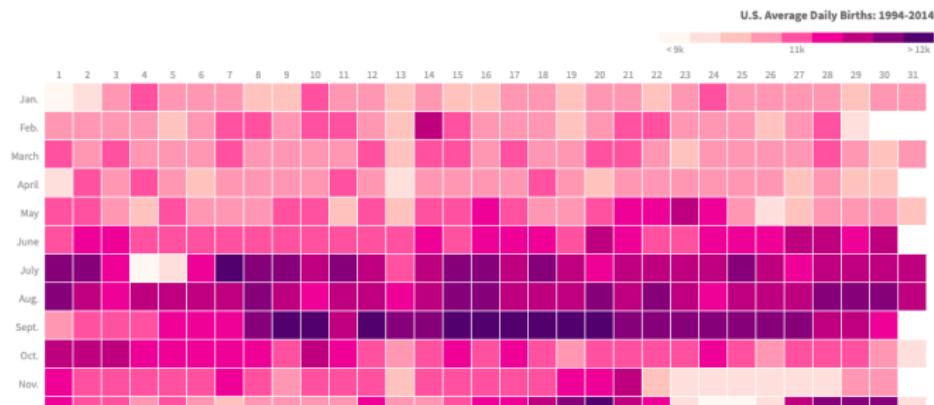
What is the probability of being born in February 29th?

What is the probability of being born in April 31st?

# Determining Probabilities

## HOW POPULAR IS YOUR BIRTHDAY?

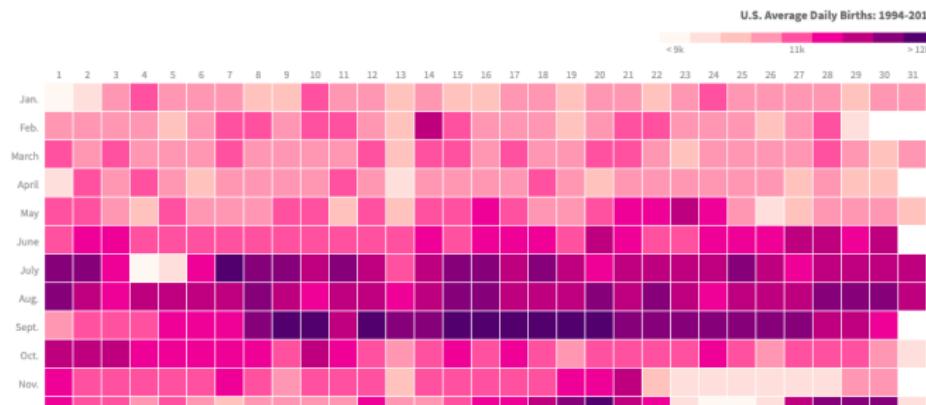
Two decades of American birthdays, averaged by month and day.



## Determining Probabilities

### HOW POPULAR IS YOUR BIRTHDAY?

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



Sometimes, it's better to propose a probability based on empirical and trustful results than in the classical method.

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) for finding common motifs (Motif is a region of a protein or DNA sequence that has a specific structure).

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) for finding common motifs (Motif is a region of a protein or DNA sequence that has a specific structure). To calculate a BLOSUM matrix, the following equation is used:

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

where

$p_{ij}$   $\equiv$  probability of two amino acids  $i$  and  $j$  replacing each other in a homologous sequence.

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) for finding common motifs (Motif is a region of a protein or DNA sequence that has a specific structure). To calculate a BLOSUM matrix, the following equation is used:

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

where

$p_{ij}$   $\equiv$  probability of two amino acids  $i$  and  $j$  replacing each other in a homologous sequence.

$q_i$   $\equiv$  background probability of finding amino acid  $i$  in any protein sequence.

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) for finding common motifs (Motif is a region of a protein or DNA sequence that has a specific structure). To calculate a BLOSUM matrix, the following equation is used:

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

where

$p_{ij}$   $\equiv$  probability of two amino acids  $i$  and  $j$  replacing each other in a homologous sequence.

$q_i$   $\equiv$  background probability of finding amino acid  $i$  in any protein sequence.

$q_j$   $\equiv$  background probability of finding amino acid  $j$  in any protein sequence.

## BLOSUM

Blocks amino acid Substitution Matrix (BLOSUM) for finding common motifs (Motif is a region of a protein or DNA sequence that has a specific structure). To calculate a BLOSUM matrix, the following equation is used:

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

where

$p_{ij}$   $\equiv$  probability of two amino acids  $i$  and  $j$  replacing each other in a homologous sequence.

$q_i$   $\equiv$  background probability of finding amino acid  $i$  in any protein sequence.

$q_j$   $\equiv$  background probability of finding amino acid  $j$  in any protein sequence.

$\lambda$   $\equiv$  scaling factor, set such that the matrix contains easily computable integer values.