

BLAST: Búsquedas en Bases de Datos de Secuencias

Luis E. Garreta U
luis.garreta@javerianacali.edu.co

Curso de Bioinformática
Pontificia Universidad Javeriana – Cali
Facultad de Ingeniería - Carrera de Biología

8 de septiembre de 2018

Objetivos

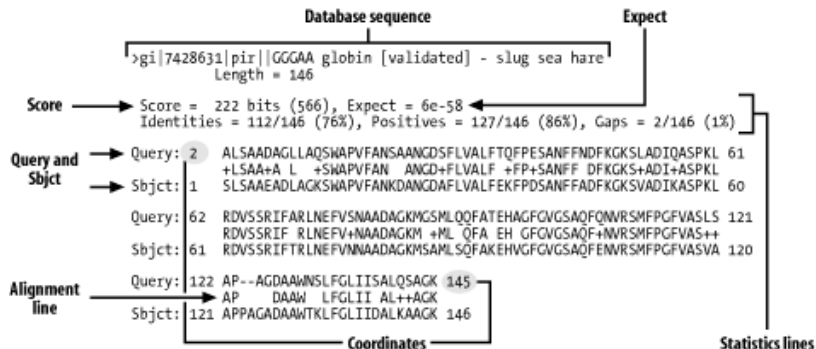
- ▶ Descubrir porque las búsquedas de similaridades son tan importantes
- ▶ Entender la relación entre homología, similaridad, e "identidad"
- ▶ Ejecutar BLAST e interpretar las salidas
- ▶ Entender el concepto de *e-values*
- ▶ Conocer cómo hacer preguntas biológicas con BLAST

- ▶ Significado biológico de **similitud** entre secuencias
- ▶ Homología, identidad, y similitud
- ▶ Ejecución de BLAST
- ▶ Interpretación de la salida de BLAST
- ▶ Análisis biológicos con BLAST
- ▶ Ejecución de PSI-BLAST

Similaridad entre Secuencias

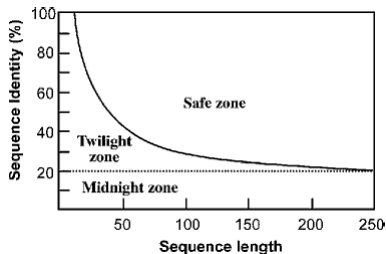
- ▶ Dos secuencias de proteínas con más del 25 % de aminoácidos idénticos (sobre 100 AA) son homologas.
- ▶ Dos secuencias de ADN con más del 70 % de nucleótidos idénticos (sobre 100 NN) son homologas.
- ▶ Secuencias homologas tienen:
 - ▶ Un ancestro común (proteínas y ADN)
 - ▶ Una estructura 3D similar (proteínas)
 - ▶ A menudo una función similar (proteínas)

Alineamientos



Homología

- ▶ Cuando dos proteínas tienen menos que el 25 % de identidad
 - ▶ Pueden ser o no homólogos
 - ▶ Es imposible decir que es verdad
- ▶ Este rango de identidad es llamado:
 - ▶ "la zona de penumbra"
 - o
 - ▶ "*Twilight Zone*"



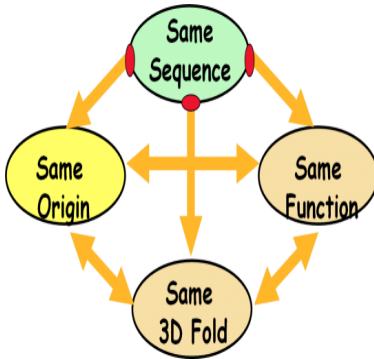
Homología, Similitud, e Identidad

- ▶ Identidad es una medida realizada sobre el alineamiento
 - ▶ Secuencia A puede ser "32 % idéntica" a la secuencia B
- ▶ Similitud es una medida de que tan cerca están dos aminoácidos:
 - ▶ Por ejemplo, isoleucina y leucina son similares
- ▶ Homología es una propiedad que existe o no
 - ▶ Secuencia A **ES** o **NO ES** homóloga a la secuencia B
 - ▶ Secuencia A **no puede** ser 40 % homóloga a B
- ▶ Homología se establece con base en la similitud e identidad medidas

Cómo Establecer Homología

- ▶ Comparar proteína A con cada proteína en una base de datos (e.g. Swiss-Prot)
- ▶ Identificar una Proteína B que es 40 % idéntica a su proteína
 - ▶ Es mejor usar el *e-value* pero la idea es la misma (...)
- ▶ Se puede concluir que A y B son probablemente homólogos si ambas son muy similares
 - ▶ Es como decir, "Juan y María son probablemente hermanos ya que ellos son muy similares"
- ▶ Entonces, si se conoce la estructura o función de B, entonces A y B probablemente tienen la misma estructura.

- ▶ Cuando ya se logra establecer que dos proteínas (A y B) son homologas, entonces se puede extrapolar todo lo que conoce de la proteína A hacia la proteína B:



- ▶ Es como realizar un "experimento virtual"
- ▶ Esto es biología *in-silico*

- ▶ BLAST: Basic Local Alignment Search Tool
- ▶ BLAST es una herramienta para comparar una secuencia con TODAS las otras secuencias dentro de una base de datos
- ▶ BLAST puede comparar:
 - ▶ Secuencias de ADN
 - ▶ Secuencias de Proteínas
- ▶ BLAST es más exacto al comparar secuencias de proteínas que al comparar secuencias de ADN

BLAST (continuación)

- ▶ BLAST realiza **alineamientos locales**:
 - ▶ Solamente alinea lo que puede ser alineado
 - ▶ e ignora el resto
- ▶ BLAST es muy rápido
 - ▶ Sólo unos pocos segundos para explorar la BD Swiss-Prot en un PC estándar

Existen varios *sabores* de BLAST para realizar diferentes tareas:

The screenshot shows the NCBI BLAST website. At the top, there's a navigation bar with the NIH logo, "U.S. National Library of Medicine", and "NCBI". On the right, it says "Sign in to NCBI". Below this, the "BLAST" logo is prominent, with links for "Home", "Recent Results", "Saved Strategies", and "Help". The main heading is "Basic Local Alignment Search Tool". A text block explains that BLAST finds regions of similarity between biological sequences. To the right, a "NEWS" sidebar announces "BLAST+ 2.4.0 released". Below the main heading, the "Web BLAST" section features three large buttons: "blastx" (translated nucleotide to protein), "tblastn" (protein to translated nucleotide), and "Protein BLAST" (protein to protein). Each button has a corresponding icon: a DNA helix for blastx, a protein ribbon for tblastn, and a protein ribbon for Protein BLAST.

BLASTing de una Secuencia de Proteínas

- ▶ Seleccione el tipo de BLAST correcto para proteínas

Qué es lo que quiero hacer?

- ▶ Quiero encontrar algo acerca de la función de mi proteína **blastp**, para compara su proteína con otras contenidas en BDs de proteínas

- ▶ Quiero descubrir nuevos genes que codifican proteínas: **tblastn**, para comparar su proteína con secuencias de ADN trasladadas en los 6 posibles marcos de lectura (3 en cada hebra)

Ejecución de blastp

1 Ingrese la secuencia a buscar:

Nombre: Ingresar el código o número de acceso (gen o proteína)

Secuencia: Cortar y pegar la secuencia cruda (ADN o Aminoácidos)

2 Seleccione uno de los servidores públicos:

NCBI: www.ncbi.nlm.nih.gov/blast

EBI: www.ebi.ac.uk/blast

EMBLnet: www.expasy.ch/blast

3 Seleccione la base de datos donde buscar:

NR para encontrar cualquier secuencia de proteína

Swiss-Prot para encontrar proteínas con funciones conocidas

PDB para encontrar proteínas con estructuras 3D conocidas

4 Ejecute BLAST dando click en el botón **BLAST**

Ejercicio 1: BLASTING Blast para una secuencia de nucleótidos desconocida

- ▶ Descargue la secuencia desconocida del github
- ▶ Ejecute BLAST sobre esta secuencia

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



Download

Transfer NCBI data
to your computer



Learn

Find help
documents, attend
a class or watch a
tutorial



Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[PubMed Health](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

Develop

Use NCBI APIs and
code libraries to
build applications



Analyze

Identify an NCBI
tool for your data
analysis task



Research

Explore NCBI
research and
collaborative
projects



BLAST®

[Home](#)[Recent Results](#)[Saved Strategies](#)[Help](#)

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

NEW

BLAST+ 2.4.0 released

A new version (2.4.0) of the BLAST+ executables is now available.

Thu, 02 Jun 2016 14:00:00 EST

[More BLAST news...](#)


Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

```
>gil9266|emb|X16893.1| Tarantula mRNA for hemocyanin subunit a
GAATCGGAGAGTGTGGTCACTTAGCGCGGGGAACATCGAGCAATCCAAGATGACCATTTTGC
ACGACAAGCAGGTTCAAGGCTGAAGTTGTTGAGAGAAGCTCAGCGTAGCCGCCACTGGTGAGC
CAGTTCCTGCAGACCAGATCGACGAAAGGCTTAGAAACATCACAACTTAGGTCCCAATGAATC
TTCTCTTGCTTTTATCCAGACCACCTTGGAACAAGCCAAGAGAGTCTACGAAGTTTCTGCCATC
```

From

To

Or, upload file

no file selected [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database

☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):

[?](#)

Organism

Optional

☐ Exclude [+](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude

Optional

☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to

Optional

Entrez Query

Optional

☐ Sequences from type material

[YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for

☐ Highly similar sequences (megablast)
☐ More dissimilar sequences (discontiguous megablast)
☒ Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)



BLAST

Search **database Nucleotide collection (nr/nt)** u

☐ Show results in a new window

Use megablast to find the best matching sequence or change to blastn to find related sequences from other organisms

Lectura de los resultados de BLAST:

- Graphic Display:
 - Vista Gráficos de los alineamientos
- Hit List:
 - Descripción de los alineamientos
- Alignments:
 - Detalle de los alineamientos

The screenshot shows a BLAST search result page. The search query is '16893:Tarantula mRNA for hemocyanin subunit...'. The query ID is 'q19266|emb|X161'. The query description is 'Tarantula mRNA for hemocyanin subunit...'. The molecule type is 'nucleic acid'. The query length is '2110'. The search results are displayed in a table with columns for 'Hit', 'Score', 'E-value', 'Identity', and 'Accession'. The first hit is '16893:Tarantula mRNA for hemocyanin subunit...' with a score of 100.0 and an E-value of 0.0. The graphic display shows a sequence alignment between the query and the hit. The hit list shows the top hits with their scores and E-values. The alignments section shows the detailed sequence alignment between the query and the hit.

Search Summary

The length of your query sequence

Graphic Display

Hit List

Alignments

Resultados: Search Summary

NIH U.S. National Library of Medicine NCBI

BLAST® » **blastn suite** » **RID-SS58EJRE014** Home Recent Results Saved Strategies

BLAST Results

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#) [YouTube](#) [How to read this page](#)

X16893:Tarantula mRNA for hemocyanin subunit...

RID [SS58EJRE014](#) (Expires on 07-19 07:30 am)

Query ID	gi 9266 emb X16893.1
Description	Tarantula mRNA for hemocyanin subunit...
Molecule type	nucleic acid
Query Length	2110

The length of your query sequence

Database Name	nr
Description	Nucleotide collection (nt)
Program	BLASTN 2.4.0+ Citation

Database name and description

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#)

☒ **Graphic Summary**

The search settings

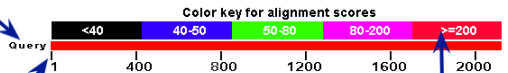
Copyright © 2016 [Digital World Biology LLC](#) All rights reserved.

Resultados: Vista Gráfica

the sequence that was used to search the database. This is called the query sequence.

Distribution of 146 Blast Hits on the Query Sequence

Refine, click to show alignments



Colors show BLAST scores for matching sequences. BLAST scores > 200 are red.

The scale shows the residue position. The first nucleotide is at position 1 and the last nucleotide is at position 2110.



Scroll down

Each bar shows where part of a sequence from the database matches part of the query sequence.

Lines show regions where the query and subject sequences are different.

Resultados: Lista de Hits

Sequences producing significant alignments:

Select the number of the sequence to view details

A description of the sequence

The Total score is the sum of the blast scores from each region where the query sequence and subject sequence align. If the two sequences align in multiple places, the total score is larger than the Max score.

Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Tarantula mRNA for hemocyanin subunit a	3806	3806	100%	0.0	100%	X16893.1
<input type="checkbox"/>	Nephila inaurata madagascariensis mRNA for hemocyanin subunit A (hc-a gene)	1045	1045	90%	0.0	72%	AJ547807.1
<input type="checkbox"/>	PREDICTED: Parasteatoda tepidariorum hemocyanin A chain-like (LOC107451659), mRNA	1034	1034	90%	0.0	72%	XM_016067823.1
<input type="checkbox"/>	PREDICTED: Parasteatoda tepidariorum hemocyanin A chain-like (LOC107440514), mRNA	996	996	83%	0.0	73%	XM_016053461.1
<input type="checkbox"/>	Euphrynichus bacillifer mRNA for hemocyanin subunit a (HcA gene)	713	713	76%	0.0	70%	FR865913.1
<input type="checkbox"/>	Mastigoproctus giganteus mRNA for hemocyanin subunit a (HcA gene)	527	527	72%	3e-145	68%	FR865920.1
<input type="checkbox"/>	Nephila inaurata madagascariensis mRNA for hemocyanin subunit F (hc-f gene)	502	502	71%	1e-137	68%	AJ547811.1

The blast score from the part of the subject sequence that aligns best to the query.

The fraction of the query sequence that aligns to the subject sequence

The percent of bases in the best aligned region that are identical.

The accession number serves as an identity number for a sequence.

See the next page to

Results: *e-values* o valor esperado

The E (Expect) value is equal to the number of matching sequences you would expect to find if you searched a database of random sequences.

Two important parameters that influence the E value are:

The number sequences in the database

The length of the query sequence

The E value increases when the database is larger and / or the query sequence is shorter. Both of these changes increase the probability of finding a matching sequence.

E
Value

0.0
1e-21
7e-20
4e-18
2e-14
1e-09
4e-09
1e-08
1e-08
1e-08
0.014
0.014
0.014
0.014
0.014
0.014
0.22
0.22
0.86
0.86
0.86
0.86
3.4
3.4

If the E value is close to zero, the program rounds the value off to zero. The letter “e” in the number means exponent.

In this example, the E value equals

$$1 \times 10^{-21}$$

An E value this low corresponds to a very low chance of finding a random sequence that matches this well.

This sequence has an E value of 3.4. That means you would expect to find 3.4 sequences in a random database that match the query as well as this one.

Results: Descripciones de las Secuencias *Hits*

The sequence descriptions are linked to alignments between the query sequence and the subject sequence

The accession number is linked to the database entry for that sequence.

Results							
	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Tarantula mRNA for hemocyanin subunit a	3806	3806	100%	0.0	100%	X16893.1
<input type="checkbox"/>	Nephila inaurata madagascariensis mRNA for hemocyanin subunit A (hc-a gene)	1045	1045	90%	0.0	72%	AJ547807.1
<input type="checkbox"/>	PREDICTED: Parasteatoda tepidariorum hemocyanin A chain-like (LOC107451659), mRNA	1034	1034	90%	0.0	72%	XM_016067823.1
<input type="checkbox"/>	PREDICTED: Parasteatoda tepidariorum hemocyanin A chain-like (LOC107440514), mRNA	996	996	83%	0.0	73%	XM_016053461.1
<input type="checkbox"/>	Euphrynichus bacillifer mRNA for hemocyanin subunit a (HcA gene)	713	713	76%	0.0	70%	FR865913.1
<input type="checkbox"/>	Mastigoproctus giganteus mRNA for hemocyanin subunit a (HcA gene)	527	527	72%	3e-145	68%	FR865920.1
<input type="checkbox"/>	Nephila inaurata madagascariensis mRNA for hemocyanin subunit F (hc-f gene)	502	502	71%	1e-137	68%	AJ547811.1

Results: Detalle de los Alineamientos

Tarantula mRNA for hemocyanin subunit a

Sequence ID: [emb|X16893.1](#) Length: 2110 Number of Matches: 1

Range 1: 1 to 2110 [GenBank](#) [Graphics](#)

This is the title, accession number, and length of the matching region from the subject sequence

Score	Expect	Identities	Gaps	Plus/Plus
3806 bits(4220)	0.0	2110/2110(100%)	0/2110(0%)	
Query 1	GAATCGGAGAGTGTGGTCACTTAGCGCGGGGAACATCGAGCAATTCCAAGATGACCATT	60		
Sbjct 1	GAATCGGAGAGTGTGGTCACTTAGCGCGGGGAACATCGAGCAATTCCAAGATGACCATT	60		
Query 61	TTGCACGACAAGCAGGTTCAAGCACTGAAGTTGTTTCGAGAAGCTCAGCGTAGCCGCCACT	120		
Sbjct 61	TTGCACGACAAGCAGGTTCAAGCACTGAAGTTGTTTCGAGAAGCTCAGCGTAGCCGCCACT	120		
Query 121	GGTGAGCCAGTTCTTCGAGACCAGATCGACGAAAGGCTTAGAAACATCACAACTTAGGT	180		
Sbjct 121	GGTGAGCCAGTTCTTCGAGACCAGATCGACGAAAGGCTTAGAAACATCACAACTTAGGT	180		
Query 181	CCCAATGAATCTTCTTCTTGCTTTTACCCAGACCACTTGGAAACAAGCCAAGAGAGTCTAC	240		
Sbjct 181	CCCAATGAATCTTCTTCTTGCTTTTACCCAGACCACTTGGAAACAAGCCAAGAGAGTCTAC	240		
Query 241	GAAAGTTTCTGCCATGCTGCTAACTTCGATGACTTCGTCAGCTTGGCAAAGCAAGCGCGA	300		
Sbjct 241	GAAAGTTTCTGCCATGCTGCTAACTTCGATGACTTCGTCAGCTTGGCAAAGCAAGCGCGA	300		

Notice, the end of each line is at 60 nucleotides, but alignments continue on the next line

In this alignment the raw blast score is 4420 (about twice the number of nucleotides), the E value is 0.0. All 2110 aligned bases are identical and there aren't any gaps.

Results: Otro alineamiento con dos matches

Download ▾ GenBank Graphics Sort by: E value ▾

Euphrynichus bacillifer mRNA for hemocyanin subunit e (HcE gene)
Sequence ID: [emb|FR865917.1](#) Length: 1994 Number of Matches: 2

Range 1: 469 to 1193 GenBank Graphics ▾ Next Match ▴ Previous Match

Score	Expect	Identities	Gaps	Strand
342 bits(378)	3e-89	521/731(71%)	12/731(1%)	Plus/Plus
Query 500	AAGACACTGGTAATATCTGGATCCAGAAATACAAACTCGCTACTTCCGAGAGGATATG	559		
Sbjct 469	AAGAAACAGGAAACATCTTGGATGAAGAGTACAAAGCTGGCTTACTTCGGTGAAGATGCG	528		
Query 560	GAGTGAACGCTCATCACTGGCATTTGGCATGTTGTTTACCCCTTACCTACGATCCCGCTT	619		
Sbjct 528	GGTGAACGCTCATCACTGGCATTTGGCATGTTGTTTACCCCTTACCTACGATCCCGCTT	587		
Query 619	AAGGACAGGAAGGGAGAGCTCTTCTATTACATGCATCAGCAGATG	678		
Sbjct 587	AAGGACAGGAAGGGAGAACTTTTCTACTACATGCATCAGCAGATG	647		
Query 678	TGTGAGCGATTGTCTAATGGCCTGAACAGGATGATTCCTTCCAC	738		
Sbjct 647	TGCGAGCGCTCTGTCCATCGGATTGCAGAGAATGATCCCTTCCAC	707		
Query 739	AACCTTCACGAAACCCCTTGGTGGTTATGCCGCTCATCTGACCCATGTTGCCAGTGGTGGC	796		
Sbjct 708	AACCTTTCAGAGCCCTTGGAAAGCTACGCACTCATCTTAAGCTCAGTTCAGTGGAC	767		
Query 796	CATTATTCAGCAGAGGCCAGATGG-TCTGGC-CATGACATATGTCGTGAAGTACAGCTG	855		
Sbjct 768	AACCTACGCTCTCGTCTGAGGGATTCACTCTCAGG-TATCTTAAACAGCTCGATGTC	824		
Query 856	CAAGACATGGGAGGTGGACTGAACGTATCATGGAAAGCCATCG-TTTACGTAGGGTCAT	914		
Sbjct 825	CAGGAGATGATCCGTTGGAGGGAACGTATCTTGGAAAGCATCCATCTTGGCTACG-TCAT	883		
Query 914	CG-ACTCAACAGCACTCACATTCCTCTTGACAAGGACCATGGCGCAGACATCCTTGGAG	942		
Query 974	CACCTCATCGAATTCAGCTACGAATCCAAAGAACGAGGCTATTACGGAAGCCTTCAACAAT	1033		
Sbjct 943	CACCTATTAGAGTCAGCTACGAGTCCAAAGAACGAGGATTATACGGCAGCTTCCACAAT	1002		
Query 1034	GGGGACATGTTATATGCTTACATTCATGATCCTGATGGCAGATTCAAGGAAACACCAG	1093		
Sbjct 1003	GGGGGACGTCATATATGGCCAGAGTACACGACCCGACGCGAGATTCCAGAAATCCAG	1062		
Query 1094	GTGTCACTGACTGACAGCGCCACAGTCTTAGGGATCCAACTCTTACAGATACACAGAT	1153		
Sbjct 1063	GTGTATGAGCGACACTTCTACCTCCATCCGTGACCCAACTCTTACAGATACACAGAT	1122		
Query 1154	TCACTGCACAAGCTTTTCCAAAGATACAAGAAACTCTGCCAGTGACAGCAAGAACATC	1213		
Sbjct 1123	TTGTGTACAACATTTCCAGAGCTTCAAGCTGTCCCTGAATCTCTTACACCAAGAGAGC	1182		
Query 1214	TGGACTTCCTCT 1224			
Sbjct 1183	TGGACTTCCTCT 1193			

These sequences align in two regions.

Mismatched bases

The numbers show the nucleotide positions where the alignments begin and end.

Gaps

Range 2: 1323 to 1492 GenBank Graphics ▾ Next Match ▴ Previous Match

Score	Expect	Identities	Gaps	Strand
55.4 bits(60)	0.005	114/170(67%)	0/170(0%)	Plus/Plus
Query 1354	TACCATCATTTGGATCAGAAATCTTTCTCTCATCATTCAGCGCCAGAACACAGCAAT	1523		
Sbjct 1323	TACAAACACTTGGATCAGCAACCTTTTCGCTACAACTCAGCGTTGAAACAGACTGGA			
Query 1414	GCTGACAAAGCAGGCAACTGTTGGAATCTTCTTGGCCCCACATACGATGAACTCGGAAT			
Sbjct 1383	GGATCTCAAAACAGCTACCGTCCGTATCTTCTTGGACCCAGAAACAGCAGGATTGGCAAT			
Query 1474	GACATCTCACTAGACGAAACAGGAGAGCTGACATTTGAAATGGACAATTT 1523			
Sbjct 1443	CTCTTGGAACTTGACGACAGCGCGCGCTGCAATCGAGTTGGACAATTT 1492			

BLAST: Búsquedas en Bases de Datos de Secuencias

Luis E. Garreta U luis.garreta@javerianacali.edu.co

Results: Registro de la Base de Datos

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

Alignments								Download	GenBank	Graphics	Distance tree of results	
	Description	Max score	Total score	Query cover	E value	Ident	Accession					
<input type="checkbox"/>	Tarantula mRNA for hemocyanin subunit a	3806	3806	100%	0.0	100%	X16893.1					
<input type="checkbox"/>	Nephila inaurata madagascariensis mRNA for hemocyanin subunit A (hc-a gene)	1045	1045	90%	0.0	72%	AF1607.1					
<input type="checkbox"/>	PREDICTED: Parasteatoda tepidariorum hemocyanin A chain-like (LOC107451659), mRNA	1034	1034	90%	0.0	72%	XM_016067823.1					
<input type="checkbox"/>	PREDICTED: Parasteatoda tepidariorum hemocyanin A chain-like (LOC107440514), mRNA	996	996	83%	0.0	73%	XM_016053461.1					
<input type="checkbox"/>	Euphrynichus bacillifer mRNA for hemocyanin subunit a (HcA gene)	713					AF1607.1					
<input type="checkbox"/>	Mastigoproctus giganteus mRNA for hemocyanin subunit a (HcA gene)	527	527	72%	0.0	68%	AF1607.1					
<input type="checkbox"/>	Nephila inaurata madagascariensis mRNA for hemocyanin subunit F (hc-f gene)	502	502	71%	1e-137	68%	AJ547811.1					

View the database record

Results: Encabezado Registro de la Base de Datos

Tarantula mRNA for hemocyanin subunit a

GenBank: X16893.1

[FASTA](#) [Graphics](#)

[Go to:](#) ☐

Sequence length & kind of molecule

2110 bp

mRNA

linear

INV 18-APR-2005

LOCUS X16893
DEFINITION Tarantula mRNA for hemocyanin subunit a.

ACCESSION X16893

VERSION X16893.1 GI:9266

KEYWORDS hemocyanin; hemocyanin subunit a.

SOURCE Aphonopelma sp.

ORGANISM [Aphonopelma sp.](#)

A link to the taxonomy database and classification

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Chelicerata; Arachnida;
Araneae; Mygalomorphae; Theraphosidae; Aphonopelma.

REFERENCE 1

AUTHORS Voit,R. and Feldmaier-Fuchs,G.

TITLE Arthropod hemocyanins. Molecular cloning and sequencing of cDNAs
encoding the tarantula hemocyanin subunits a and e

J. Biol. Chem. 265 (32), 19447-19452 (1990)

PUBMED [2246235](#)

Publication

REFERENCE 2 (bases 1 to 2110)

AUTHORS Voit,R.

TITLE Direct Submission

JOURNAL Submitted (12-OCT-1989) Voit R., Zoologisches Institut,

Universitaet Muenchen, Luisenstrasse 14, D-8000 Muenchen 2, FRG

COMMENT Data kindly reviewed (26-MAR-1990) by Voit R.

FEATURES Location/Qualifiers

Results: *Features* Registro de la Base de Datos

FEATURES	Location/Qualifiers
source	1..2110 /organism="Aphonopelma sp." /mol_type="mRNA" /db_xref="taxon:29932" /clone="lambda-K1" /clone_lib="lambda-gt10"
CDS	52..1947 /note="unnamed protein product; h 1-631)" /codon_start=1 /protein_id="CAA34771.1" /db_xref="GI:9267" /db_xref="GOA:P14750" /db_xref="InterPro:IPR000896" /db_xref="InterPro:IPR002227" /db_xref="InterPro:IPR005203" /db_xref="InterPro:IPR005204" /db_xref="PR008922" /db_xref="PR013788" /db_xref="PR014756" /db_xref="UniProtKB/Swiss-Prot:P14750" /translation="MTILHDKQVQALKLFEKLSVAATGEPVPADQIDERLNRITTLGP NEFFSCFYPDHLEQAARVYEVFCHAANFDDFVSLAKQARSFMNSTLFAFSAEVALLHR EDCRGVIVPPVQEVFADRFIPADSLIKAFTLATTTQPGDESIDIIVDKDTGNILDPEY KLAYFREDIGVNAHHWHVVPSTYDPAFFGKVKDRKGELFYMHQMCARYDCERL SNGLNRMPFHNFNPEPLGGYAAHLTHVASGRHYAQRPDGLAMHDVREVVDQDMRWTE RIMEAIDLRRVISPTGEYIPLDEEHGADILGALIESSYESKNRGYGSLSHNWGHVMMMA YIHDPDGRFRETPGVMTDTATSLRDPIFYRYHRFIDNVFQYKKTLPVYSKDNLDFPQ VTITDVVKVAKIPNVVHTFIREDELELSHCLHFAKPGCSVRARYHHLDHESFSYIIISAQ NNSNADKQATVRIFLAPTYDELGNDISLDEQRRLYIEMDKFYHTLRPGKNTIVRSSTD SSVTLSVVHTFKELLRGEDLVEGQTEFCSCGWPQHLLVPKGNEKGMQFDLFVMLTDAS VDRVQSGDGTPLCADALSYCGVLDQKYPDKRAMGYFDRKITADTHEEFLTGMNINSH VTVRFQD" misc feature 2089..2094 /note="polyA signal"

Where did the source material come from?

CDS = coding sequence.
The mRNA region between
nucleotides 52 and 1947
codes for protein.


The predicted amino acid sequence


MTILHDKQVQALKLFEKLSVAATGEPVPADQIDERLNRITTLGP
NEFFSCFYPDHLEQAARVYEVFCHAANFDDFVSLAKQARSFMNSTLFAFSAEVALLHR
EDCRGVIVPPVQEVFADRFIPADSLIKAFTLATTTQPGDESIDIIVDKDTGNILDPEY
KLAYFREDIGVNAHHWHVVPSTYDPAFFGKVKDRKGELFYMHQMCARYDCERL
SNGLNRMPFHNFNPEPLGGYAAHLTHVASGRHYAQRPDGLAMHDVREVVDQDMRWTE
RIMEAIDLRRVISPTGEYIPLDEEHGADILGALIESSYESKNRGYGSLSHNWGHVMMMA
YIHDPDGRFRETPGVMTDTATSLRDPIFYRYHRFIDNVFQYKKTLPVYSKDNLDFPQ
VTITDVVKVAKIPNVVHTFIREDELELSHCLHFAKPGCSVRARYHHLDHESFSYIIISAQ
NNSNADKQATVRIFLAPTYDELGNDISLDEQRRLYIEMDKFYHTLRPGKNTIVRSSTD
SSVTLSVVHTFKELLRGEDLVEGQTEFCSCGWPQHLLVPKGNEKGMQFDLFVMLTDAS
VDRVQSGDGTPLCADALSYCGVLDQKYPDKRAMGYFDRKITADTHEEFLTGMNINSH
VTVRFQD"

misc feature

2089..2094
/note="polyA signal"

Results: Publicación

 **NCBI** Resources ▾ How To ▾

 **PubMed**
US National Library of Medicine
National Institutes of Health


Advanced

Format: Abstract ▾ Send to ▾

[J Biol Chem.](#) 1990 Nov 15;265(32):19447-52.

Arthropod hemocyanins. Molecular cloning and sequencing of cDNAs encoding the tarantula hemocyanin subunits a and e.

[Voit R¹](#), [Feldmaier-Fuchs G.](#)

 **Author information**

Abstract

cDNA clones comprising the entire coding region of two out of the seven heterogeneous subunits of hemocyanin from the tarantula, *Eurypelma californicum*, were isolated from four cDNA libraries constructed from total RNA from the heart tissue of single spiders. Hybridization was first carried out using a tarantula hemocyanin subunit e partial cDNA, and several positive clones were isolated, including one containing a 2.2-kilobase full-length cDNA (lambda M1). The cDNA comprises an open reading frame for 623 amino acids, 34 nucleotides of the 5'-noncoding region, and 286 nucleotides of the 3'-noncoding region. To select for other hemocyanin subunits, two 17-mer oligonucleotide mixtures, corresponding to the conserved regions in the copper A and copper B oxygen-binding site of chelicerate hemocyanins, were used as probes. Among the positive clones obtained, full-length cDNAs coding for subunit a were identified. The cDNA sequence determined from clone lambda K1 provides an open reading frame coding for 630 amino acids and includes the 5'- and 3'-noncoding regions. Northern blot analysis revealed single transcripts for subunits a and e, each 2.3 kilobases long. The cDNAs for subunits a and e were both found to lack any leader peptide sequence. This supports the idea that the mature protein accumulates in the cytoplasm and is released by cell rupture.

- ▶ E-value significa valor esperado (*expecta*)
- ▶ E-value es la medida más usada para estimar similaridad entre secuencias
- ▶ Cuantas veces es puedo encontrar al azar un alineamiento igual
- ▶ Si un alineamiento es altamente inesperado, este probablemente no se dará solo por azar
 - ▶ Origin común es la explicación más probable
 - ▶ Es así como se infiere homología

Qué valores son aceptables para los *E-values*

- ▶ Bajo e-value **implica** un buen hit
 - ▶ 1 = bad e-Value
 - ▶ $10e-3$ = borderline E-value
 - ▶ $10e-4$ = good E-value
 - ▶ $10e-10$ = very good E-value
- ▶ E-values menores que $10e-4$ indican posible homología
- ▶ E-values más altos que $10e-4$ requieren evidencia extra evidence para soportar la homología


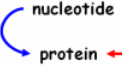
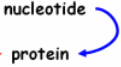
Porqué usar *e-values*

- ▶ e-values hacen posible comparar alineamientos de diferentes longitudes
- ▶ e-values son usados por la mayoría de programas de comparación de secuencias
 - ▶ BLAST
 - ▶ FASTA
 - ▶ PSI-BLAST
 - ▶ ...

BLASTing Secuencias de ADN

- ▶ Secuencias de ADN:
 - ▶ ADN codificante
 - ▶ ADN no-codificante
- ▶ BLASTing de secuencias de ADN es menos exacto que realizar BLAST con secuencias de proteínas

BLASTing Secuencias de ADN

Program	Query	Database
blastn	nucleotide	nucleotide
blastx		
tblastx		

Ejercicio2: Realizar el BLAST de la proteína P09405