

# Introducción a la Bioinformática

## BIO 267

Jessica Liliana Campo Giraldo

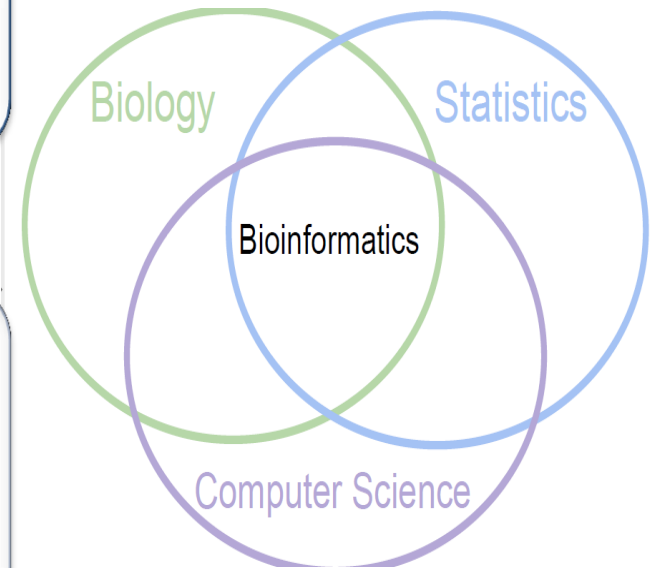
# Bioinformática

## Definición

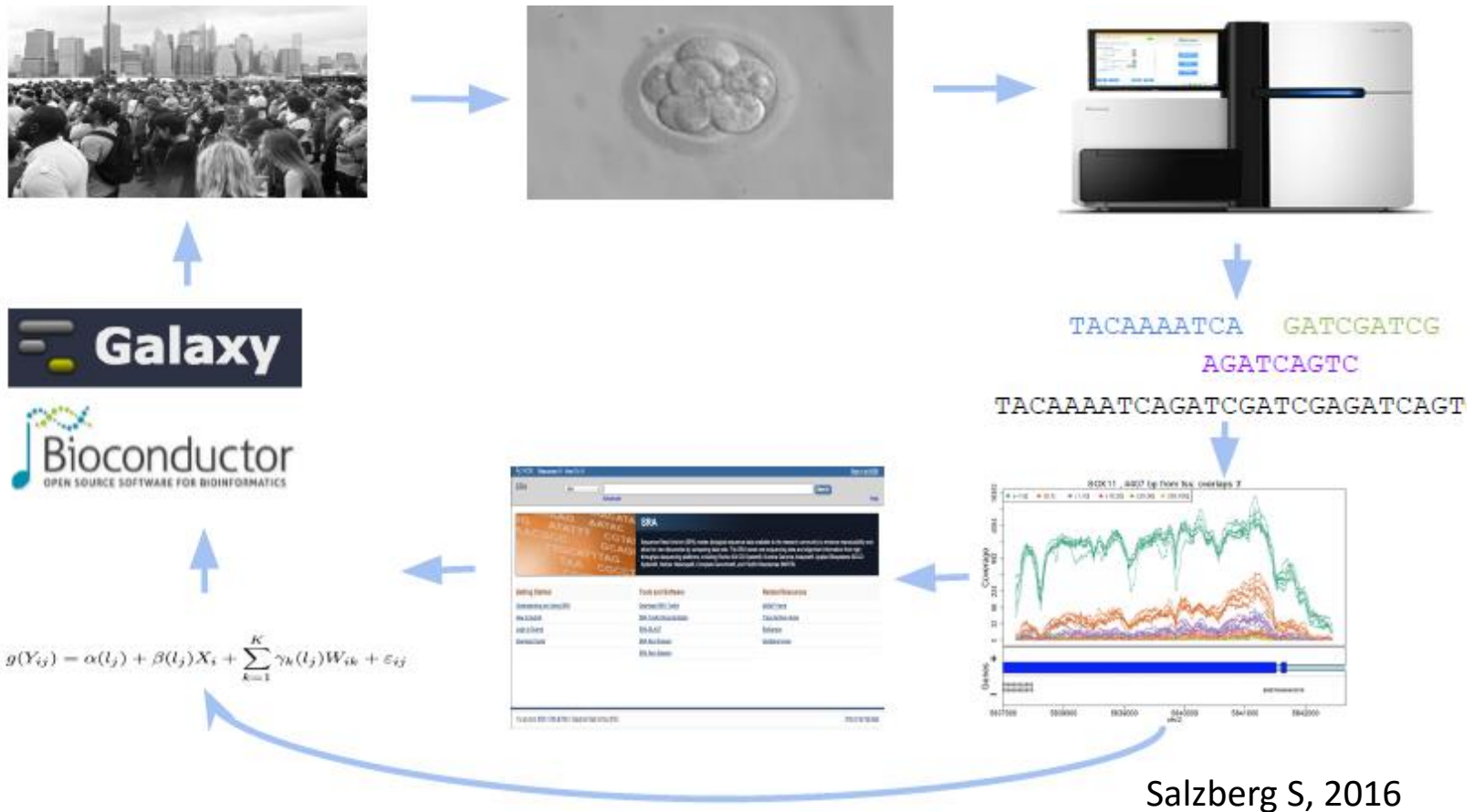
- Campo de la ciencia que se ocupa de la información y flujo de información en los sistemas biológicos, apoyado en el uso de métodos computacionales en genética y genómica(OED, 2015).

## Objetivo

- Facilitar el descubrimiento de nuevas ideas biológicas así como crear perspectivas globales a partir de las cuales se puedan discernir principios unificadores en biología (NCBI, 2001).



# ¿Cuál es el potencial de esta área?



# ¿Qué hemos aprendido a la fecha?



TACAAAATCA GATCGATCG  
AGATCAGTC  
TACAAAATCAGATCGATCGAGATCAG!

Desarrollo de  
hipótesis

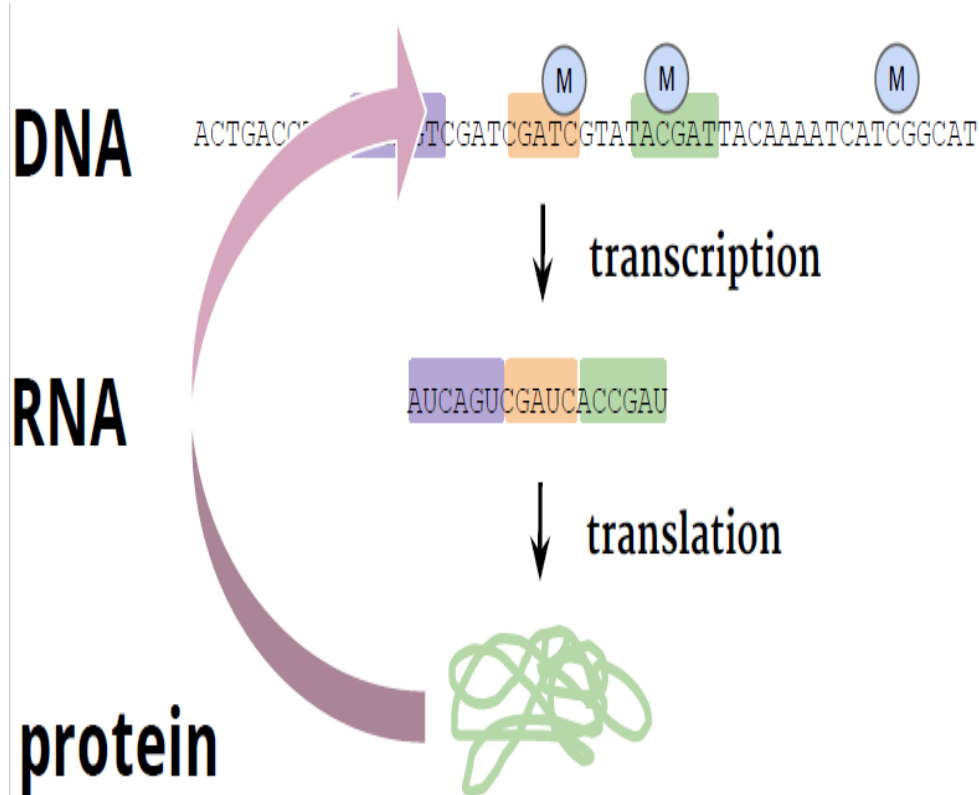


Diseño  
experimental



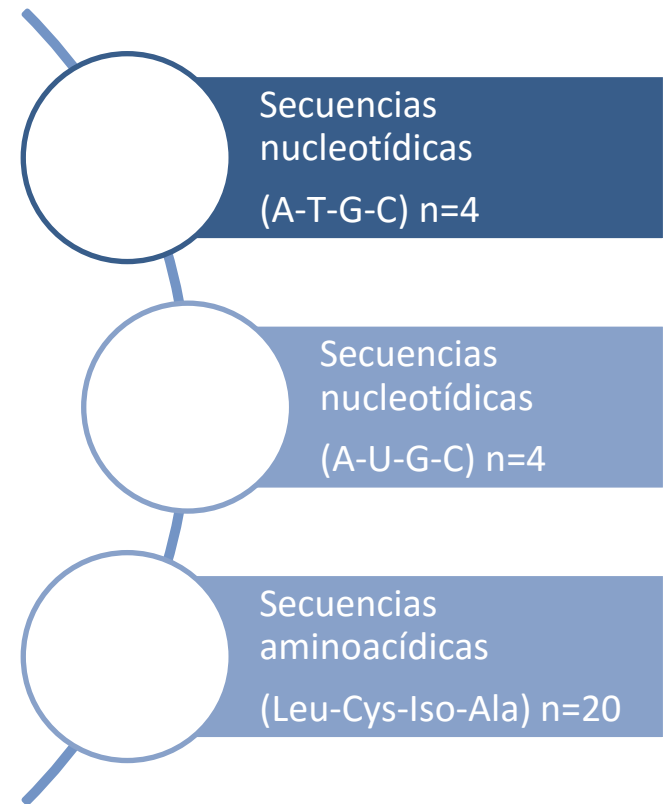
Análisis de  
resultados

# Dogma central de la biología

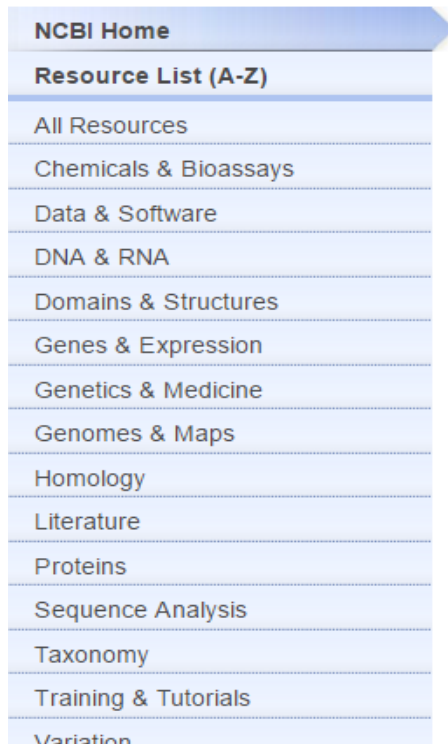
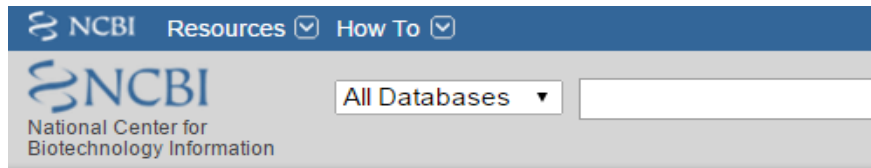


slide adapted from Alyssa Frazee

# Flujo de la información



# ¿Qué nos falta aprender?



## Welcome to NCBI

The National Center for Biotechnology  
biomedical and genomic information

[About the NCBI](#) | [Mission](#) | [Organ](#)

### Submit

Deposit data or manuscripts  
into NCBI databases



### Develop

Use NCBI APIs and code  
libraries to build applications

Uso de bases  
de datos



Herramientas  
de comparación  
de secuencias



Búsqueda de  
propiedades en  
una secuencia

# **Bases de datos**

# Bases de datos

ScienceDirect



EBSCO Information Services

WEB OF SCIENCE™




## PubMed


PubMed comprises more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.



# Ingresar a:

## <http://www.ncbi.nlm.nih.gov/>

Resources ☒ How To ☒Sign in to NCBI

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




#### Download

Transfer NCBI data to your computer



#### Learn

Find help documents, attend a class or watch a tutorial



#### 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

### Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

### NCBI Announcements

HTTPS at NCBI: Guidance for NCBI web API users

27 Jul 2016

As originally announced on June 10, NCBI will be moving all web services to

9

# Obteniendo secuencia nucleotídica

<http://www.ncbi.nlm.nih.gov/nuccore>

NCBI Resources ▾ How To ▾

Nucleotide Nucleotide ▾

[Create alert](#) [Advanced](#)

NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! [Read more...](#)

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

Animals (2,002)  
Plants (1,609)  
Fungi (2,342)  
Protists (2,846)  
Bacteria (160,923)  
Archaea (1,558)  
Viruses (42)  
Customize ...

Molecule types  
genomic DNA/RNA (168,981)  
mRNA (2,256)  
Customize ...

Source databases  
INSDC (GenBank) (85,520)  
RefSeq (85,814)  
Customize ...

Genetic compartments

**Items: 1 to 20 of 171348**

<< First < Prev Page 1 of 8568 Next > Last >>

Found 173215 nucleotide sequences. Nucleotide (171348) EST (1834) GSS (33)

☐ [Trypanosoma grayi glutamate dehydrogenase partial mRNA](#)

1. 3,000 bp linear mRNA  
Accession: XM\_009315222.1 GI: 686642102  
[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Human glutamate dehydrogenase \(GDH\) mRNA, complete cds](#)

2. 2,950 bp linear mRNA  
Accession: M37454.1 GI: 483857  
[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Escherichia coli glutamate dehydrogenase \(gdhA\) gene, complete cds](#)

3. 1,779 bp linear DNA  
Accession: U00096.1 GI: 1146100

# Secuencia nucleotídica

## Escherichia coli glutamate dehydrogenase (gdhA) gene, complete cds

GenBank: J01615.1

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

**Formato**

Go to: ☒

---

LOCUS	ECOGDHA	1779 bp	DNA	linear	BCT 10-FEB-2004
DEFINITION	Escherichia coli glutamate dehydrogenase (gdhA) gene, complete cds.				
ACCESSION	J01615 K00565 M23171 X00988				
VERSION	J01615.1 GI:146123				
KEYWORDS	NADP-specific glutamate dehydrogenase; dehydrogenase; gdhA gene; glutamate dehydrogenase.				
SOURCE	Escherichia coli				
ORGANISM	<a href="#">Escherichia coli</a> Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.				
REFERENCE	1 (bases 361 to 714)				
AUTHORS	Mattaj,I.W., McPherson,M.J. and Wootton,J.C.				
TITLE	Localisation of a strongly conserved section of coding sequence in glutamate dehydrogenase genes				
JOURNAL	FEBS Lett. 147 (1), 21-25 (1982)				
PUBMED	<a href="#">6754449</a>				
REFERENCE	2 (bases 1 to 498)				
AUTHORS	Valle,F., Sanvicente,E., Seeburg,P., Covarrubias,A., Rodriguez,R.L. and Bolivar,F.				
TITLE	Nucleotide sequence of the promoter and amino-terminal coding region of the glutamate dehydrogenase structural gene of Escherichia coli				
JOURNAL	Gene 23 (2), 199-209 (1983)				
PUBMED	<a href="#">6225701</a>				
REFERENCE	3 (bases 121 to 1779)				

---

# Formato FASTA

## Escherichia coli glutamate dehydrogenase (gdhA) gene, complete cds

GenBank: J01615.1

[GenBank](#) [Graphics](#)

>gi|146123|gb|J01615.1|ECOGDHA Escherichia coli glutamate dehydrogenase (gdhA) gene, complete cds

```
CCGGGTGGCAAACTTTAGCGTCTGAGGTTATCGCATTGGTTATGAGATTACTCTCGTTATTAATTTGC
TTTCCTGGGTCATTTTTTTCTTGCTTACCGTCACATTCTTGATGGTATAGTCGAAAAGCACA
TGACATAAACACATAAGCACAATCGTATTAATATATAAGGGTTTTATATCTATGGATCAGACATATTCT
CTGGAGTCATTCTCAACCATGTCCAAAAGCGCGACCCGAATCAAACCGAGTTCGCGCAAGCCGTTCTG
AAGTAATGACCACACTCTGGCCTTTCTTGAACAAAATCCAAAATATCGCCAGATGTCATTACTGGAGCG
TCTGGTTGAACCGAGCGCGTGATCCAGTTTCGCGTGGTATGGGTTGATGATCGCAACCAGATACAGGTC
AACCGTGCATGGCGTGTGCAGTTCAGCTCTGCCATCGGCCCGTACAAAGGCGGTATGCGCTTCCATCCGT
CAGTTAACCTTTCCATTCTCAAATTCCTCGGCTTTGAACAAACCTTCAAAAATGCCCTGACTACTCTGCC
GATGGGCGGTGGTAAAGGCGGCAGCGATTTTCGATCCGAAAGGAAAAAGCGAAGGTGAAGTGATGCGTTTT
TGCCAGGCGCTGATGACTGAACTGTATCGCCACCTGGGCGCGGATACCGACGTTCCGGCAGGTGATATCG
GGGTTGGTGGTCGTGAAGTCGGCTTTATGGCGGGGATGATGAAAAAGCTCTCCAACAATACCGCCTGCGT
CTTACCGGTAAGGGCCTTTCAATTTGGCGGCAGTCTTATTCGCCCAGGCTACCGGCTACGGTCTGGTT
TATTTACAGAAGCAATGCTAAAACGCCACGGTATGGGTTTGAAGGGATGCGCGTTTCCGTTTCTGGCT
CCGGCAACGTCGCCAGTACGCTATCGAAAAGCGATGGAATTTGGTGCTCGTGTGATCACTGCGTCAGA
CTCCAGCGGCACTGTAGTTGATGAAAGCGGATTCACGAAAGAGAACTGGCACGTCTTATCGAAATCAA
GCCAGCCGCGATGGTCGAGTGGCAGATTACGCCAAAGAATTTGGTCTGGTCTATCTCGAAGGCCAACAGC
CGTGGTCTCTACCGGTTGATATCGCCCTGCCTTGCGCCACCCAGAATGAACTGGATGTTGACGCCGCGCA
TCAGCTTATCGCTAATGGCGTTAAAGCCGTGCGCGAAGGGGCAAAATATGCCGACCACCATCGAAGCGACT
GAACTGTTCCAGCAGGCAAGGCGTACTATTTGCACCGGGTAAAGCGGCTAATGCTGGTGGCGTCGCTACAT
CGGGCCTGAAAATGGCACAAAACGCTGCGCGCTGGGCTGGAAAGCCGAGAAAGTTGACGCACGTTTGCA
TCACATCATGCTGGATATCCACCATGCCTGTGTTGAGCATGGTGGTGAAGGTGAGCAAACCAACTACGTG
CAGGGCGCGAACATTGCCGGTTTTGTGAAGGTTGCCGATGCGATGCTGGCGCAGGGTGTGATTTAAGTTG
TAAATGCCTGATGGCGCTACGCTTATCAGGCCTACAAATGGGCACAATTCATTGCAGTTACGCTCTAATG
TAGGCCGGGCAAGCGCAGCGCCCCGGCAAAATTTAGGCGTTTATGAGTATTTAACGGATGATGCTCCC
CACGGAACATTTCTTATGGGCCAACGGCATTCTTACTGTAGTGCTCCCAAACTGCTTGTGTAACGAT
AACACGCTTCAAGTTCAGCATCCGTTAAC
```

# **Herramientas de comparación:**

## **BLAST n y BLAST p**

# BLAST n

## Caso de estudio

Tengo una secuencia aminoacídica de un gen XX

¿Cómo puedo determinar la identidad del gen o similitud con otros ?

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information

**BLAST** » blastn suite

blastn blastp blastx tblastn tblastx

### Enter Query Sequence

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

U01615.1

**Ingreso de secuencia**

From

To

Or, upload file  Ningún archivo seleccionado [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.):  
Nucleotide collection (nr/nt) [Exclude](#) [+](#)

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

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

Limit to Optional ☐ Sequences from type material

Entrez Query Optional  [YouTube](#) [Create custom database](#)

### Program Selection

Optimize for ☐ Highly similar sequences (megablast) ☐ More distant sequences (discontiguous megablast) ☒ Somewhat similar sequences (blastn) [Choose a BLAST algorithm](#)

**BLAST** Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences) ☒ Show results in a new window

# Resultados



U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

**BLAST**® » blastn suite » RID-U4RW0AC3014

## BLAST Results

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

gb|J01615.1| (1779 letters)

RID [U4RW0AC3014](#) (Expires on 08-05 05:17 am)

Query ID [gi|146123|gb|J01615.1|ECOGDHA](#)

Description *Escherichia coli* glutamate dehydrogenase (gdhA) gene, complete cds

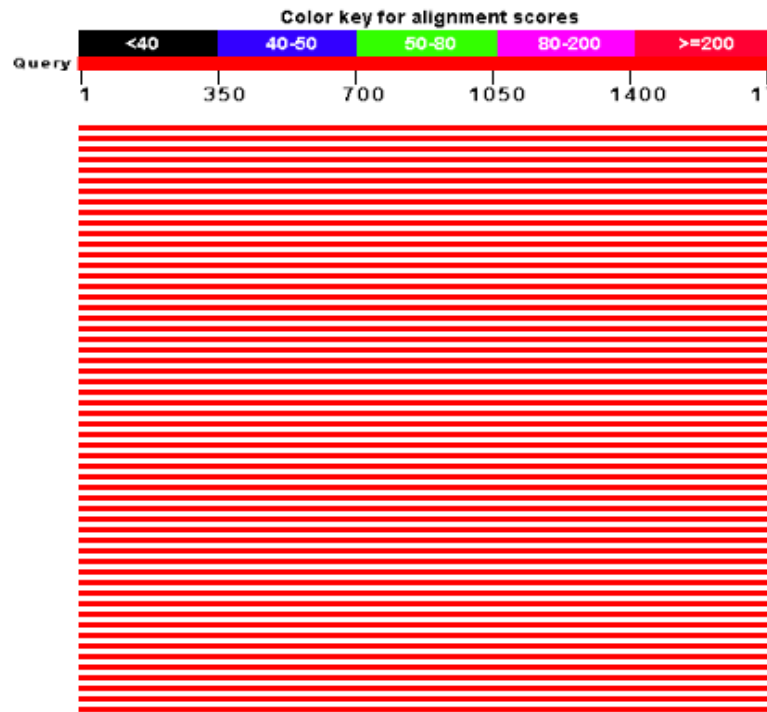
Molecule type nucleic acid

Query Length 1779

Database Name nr  
Description Nucleotide collection (nt)  
Program BLASTN 2.5.0+ [Citation](#)

Distribution of 200 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



Sequences producing significant alignments:

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

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of re](#)

	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> <a href="#">Escherichia coli glutamate dehydrogenase (gdhA) gene, complete cds</a>	3209	3209	100%	0.0	100%	<a href="#">J01615.1</a>
<input type="checkbox"/> <a href="#">Escherichia coli isolate NCTC86EC genome assembly chromosome:1</a>	3200	6558	100%	0.0	99%	<a href="#">LT801384.1</a>
<input type="checkbox"/> <a href="#">Escherichia coli strain EC690, complete genome</a>	3200	6547	100%	0.0	99%	<a href="#">CP016182.1</a>
<input type="checkbox"/> <a href="#">Escherichia coli strain ER1821R, complete genome</a>	3200	6603	100%	0.0	99%	<a href="#">CP016018.1</a>
<input type="checkbox"/> <a href="#">Escherichia coli str. K-12 substr. MG1655 strain JW5437-1, complete genome</a>	3200	6603	100%	0.0	99%	<a href="#">CP014348.1</a>

- Colores
- E-value
- %identidad
- %Cobertura

# BLAST p

## Caso de estudio

Tengo una secuencia aminoacídica de una proteína

¿Cómo puedo determinar la identidad de la proteína?

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information

**BLAST** » blastp suite

blastn **blastp** blastx tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

LL EEFDEDEEMLDLVDRSPAELTVSLDDIVLTDQEAEMSKVAKAIEHRDYLATIGLNEVGKLLFVG  
PPGTGKTSARGLAHQDLDPFVEVKLSMITSQYLGETAKNVEKVFEVAKRLSPICILFMDEFDFVATTRTG  
DEHNAIKRAVNTLLKSIDISLVTDQVLLIGATNHPDELDAAANRRFDEILSFPRPDEGMRAIIISLVTS  
EVDIADFPAALAAETSGLTGSOLRLVLRRAVLDAVEDRTELTDOLMAAISEFEDRDHLRNLDLTLEDA  
LDOTHSOHTDDHPAPNPE

Clear Query subrange

From

To

Or, upload file  Ningún archivo seleccionado

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism Optional

Exclude Optional

Entrez Query Optional

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

Enter an Entrez query to limit search

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

**BLAST**

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

☐ Show results in a new window

Ingreso de secuencia



# **Propiedades de las proteínas**

# Base de datos con información sobre proteínas

<http://www.uniprot.org/>




UniProtKB  Advanced


[BLAST](#) [Align](#) [Retrieve/ID mapping](#) [Help](#) [Contact](#)

## UniProtKB results

[About UniProtKB](#) [Basket](#)

**Filter by<sup>i</sup>**






 Reviewed (17)  
Swiss-Prot

 Unreviewed (5,004)  
TrEMBL

**Popular organisms**  
E. coli K12 (2)

[BLAST](#) [Align](#) [Download](#) [Add to basket](#) [Columns](#) [Share](#)

◀ 1 to 25 of 5,021 ▶ Show 25 ▼

<input type="checkbox"/>	Entry ▾	Entry name ▾		Protein names ▾ 	Gene names ▾	Organism ▾	Length ▾ 
<input type="checkbox"/>	P0A915	OMPW_ECOLI		<b>Outer membrane protein W</b>	ompW yciD, b1256, JW1248	Escherichia coli (strain K12)	212
<input type="checkbox"/>	A0A069B979	A0A069B979_BURPE		<b>Membrane protein</b>	ompW_3 DP46_1899, DP49_123, ERS012314_05325,	Burkholderia pseudomallei (Pseudomonas pseudomallei)	243

# Copiar código UniProt

UniProtKB - P0A915 (OMPW\_ECOLI)

## Display

[BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#)

Entry

Feature viewer

Feature table

None

☒ Function

☒ Names & Taxonomy

☒ Subcellular location

☐ Pathology & Biotech

☒ PTM / Processing

☐ Expression

☒ Interaction


Protein | **Outer membrane protein W**

Gene | **ompW**

Organism | *Escherichia coli* (strain K12)

Status |  Reviewed - Annotation score:  - Experimental evidence at protein level<sup>i</sup>

## Function<sup>i</sup>

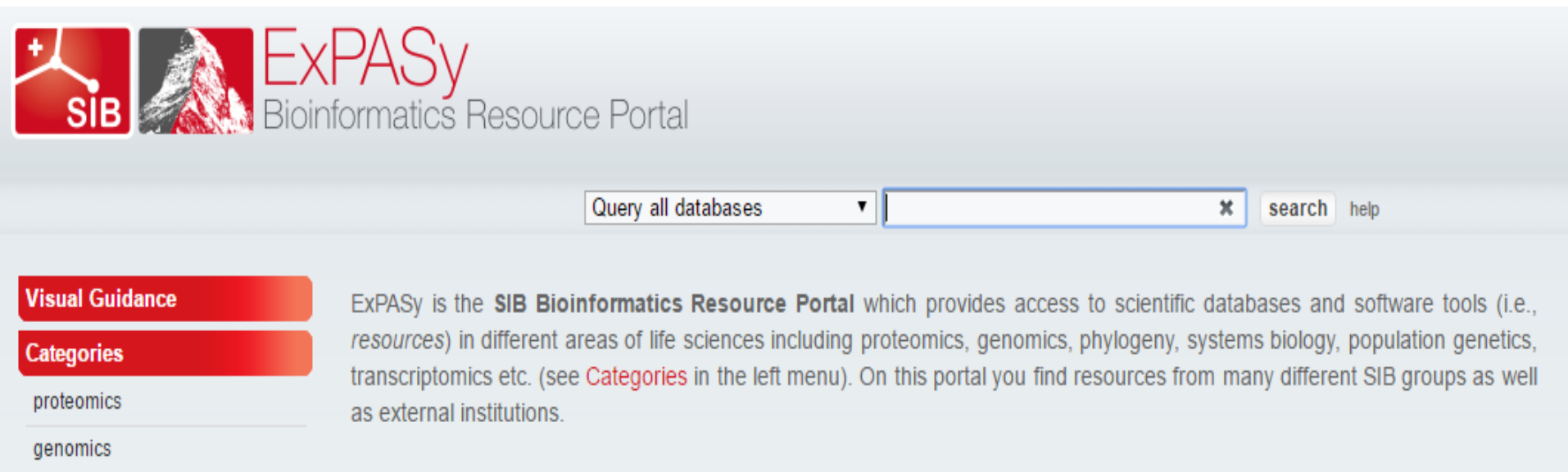
Acts as a receptor for colicin S4.  1 Publication ▼

## Enzyme and pathway databases

BioCyc <sup>i</sup>	EcoCyc:EG11124-MONOMER. ECOL316407:JW1248-MONOMER.
---------------------	---

# Base de datos con información sobre proteínas

<http://www.expasy.org/>



The image shows the top section of the ExPASy Bioinformatics Resource Portal. On the left, there are two logos: a red square with a white plus sign and 'SIB' (Swiss Institute of Bioinformatics), and a red square with a white mountain-like shape. To the right of these logos is the text 'ExPASy' in a large, red, serif font, followed by 'Bioinformatics Resource Portal' in a smaller, grey, sans-serif font. Below the logos and text is a search bar with a dropdown menu set to 'Query all databases', a search input field with a blue border and a red 'x' icon, and buttons for 'search' and 'help'. On the left side of the page, there are two red buttons: 'Visual Guidance' and 'Categories'. Below the 'Categories' button, there are two links: 'proteomics' and 'genomics'. On the right side of the page, there is a paragraph of text describing the portal.

**Visual Guidance**

**Categories**

proteomics

genomics

ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see **Categories** in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

# Peso molecular y propiedades fisicoquímicas



## ProtParam tool

**ProtParam** ([References](#) / [Documentation](#)) is a tool which allows the user to enter a protein sequence. The computed parameters include molecular weight, isoelectric point, instability index, aliphatic index and grand average of hydropathicity.

Please note that you may only fill out **one** of the following fields at a time:

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example F00000)

Or you can paste your own amino acid sequence (in one-letter code)

**Ingreso código Uniprot**

RESET

Compute parameters

Ruta:

➤ Página inicio ExPASy

➤ Resources A..Z

Compute ProtParam tool

# Punto isoeléctrico

Ruta:

- Página inicio  
ExPASy
- Resources A..Z  
Compute pI/Mw

 **ExPASy**  
Bioinformatics Resource Portal

Compute pI/Mw

---

### Compute pI/Mw tool

---

**Compute pI/Mw** is a tool which allows the computation of the theoretical pI (isoelectric point) and Mw (molecular weight) for UniProtKB/Swiss-Prot protein identifiers (ID) (e.g. *ALBU\_HUMAN*) or UniProt KB entries or for user entered sequences [[reference](#)].

---

[Documentation](#) is available.

---

#### Compute pI/Mw for Swiss-Prot/TrEMBL entries or a user-entered sequence

Please enter one or more UniProtKB/Swiss-Prot protein identifiers (ID) (e.g. *ALBU\_HUMAN*) or UniProt KB entries or for user entered sequences in single letter code. The theoretical pI and Mw (n) will be computed.

P0A915

**Ingreso código Uniprot**

Or upload a file from your computer, containing one Swiss-Prot/TrEMBL ID/AC or one sequence per line:

Resolution: ☒ Average or ☐ Monoisotopic

[Click here to compute pI/Mw](#) [Reset](#)

# Estructura secundaria

Ruta:

- Página inicio ExPASy
- Resources A..Z

SOPMA

The screenshot shows the PRABI-GERLAND website header with the logo and navigation links (Home, Services, Teaching, Publications). The main heading is 'SOPMA SECONDARY STRUCTURE PREDICTION METHOD'. Below this are links for '[Abstract]', '[NPS@ help]', and '[Original server]'. There is a text input field for 'Sequence name (optional)'. A large text area is provided for 'Paste a protein sequence below : help'. At the bottom, there is an 'Output width' dropdown set to '70', and 'SUBMIT' and 'CLEAR' buttons.

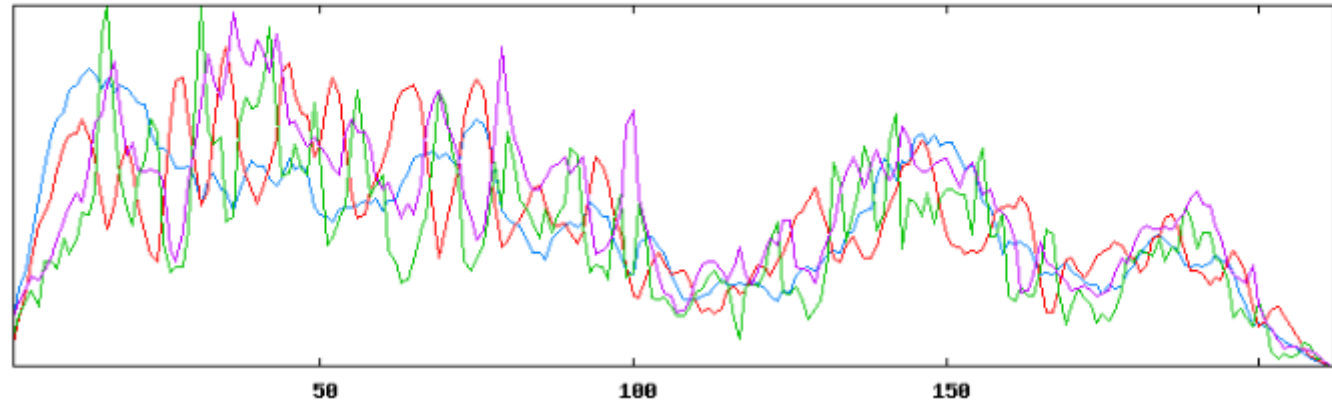
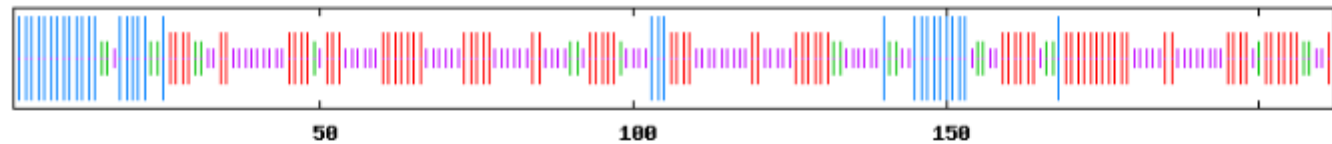
**Ingresa código Uniprot o  
secuencia aminoacídica  
en formato FASTA**

# Resultados

Sequence length : 212

SOPMA :

Alpha helix	(Hh)	:	35	is	16.51%
3 <sub>10</sub> helix	(Gg)	:	0	is	0.00%
Pi helix	(Ii)	:	0	is	0.00%
Beta bridge	(Bb)	:	0	is	0.00%
Extended strand	(Ee)	:	74	is	34.91%
Beta turn	(Tt)	:	21	is	9.91%
Bend region	(Ss)	:	0	is	0.00%
Random coil	(Cc)	:	82	is	38.68%
Ambiguous states (?)		:	0	is	0.00%
Other states		:	0	is	0.00%





# Dominios transmembrana



## Phobius

A combined transmembrane topology and signal peptide predictor

Ruta:

- Página inicio
- ExPASy
- Resources A..Z
- Phobius

[Normal prediction](#) [Constrained prediction](#) [PolyPhobius](#) [Instructions](#) [Download](#) [Mirror site at KU](#)

### Normal prediction

Paste your protein sequence here in Fasta format:

**Ingreso código Uniprot o  
secuencia aminoacídica  
en formato FASTA**

Or: Select the sequence file you wish to use  Ningún archivo seleccionado

Select output format:

- ☐ Short
- ☐ Long without Graphics
- ☒ Long with Graphics

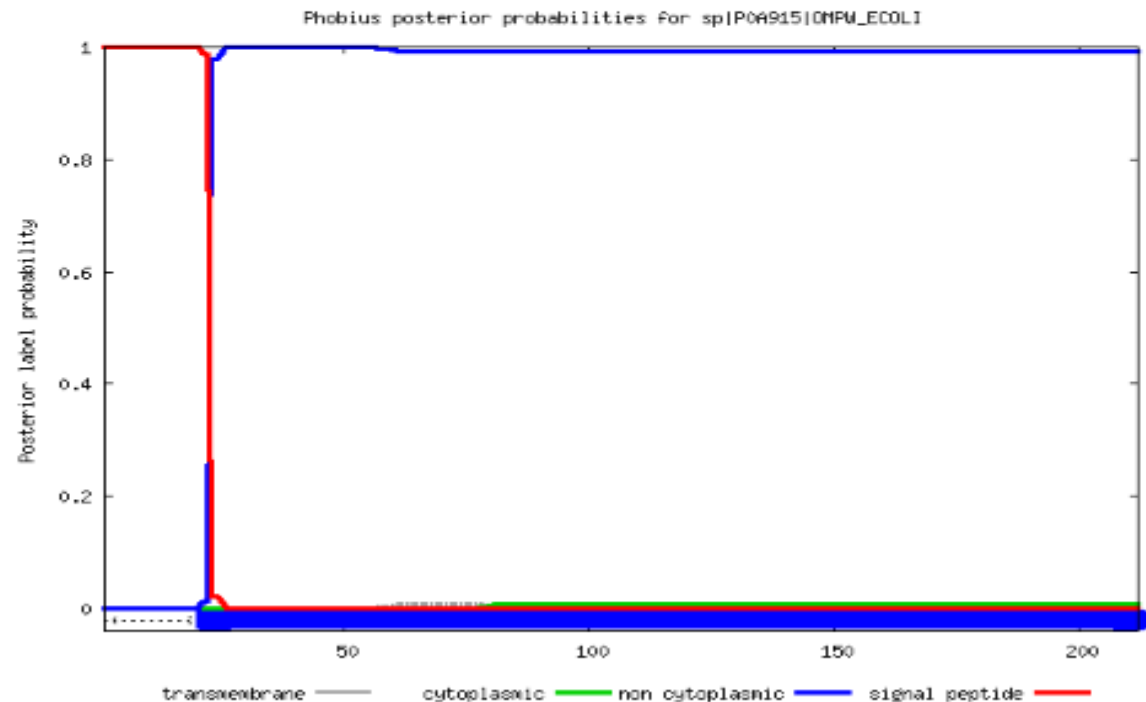
# Resultados

## Phobius prediction

### Prediction of sp|P0A915|OMPW\_ECOLI

ID	sp P0A915 OMPW_ECOLI			
FT	SIGNAL	1	21	
FT	REGION	1	3	N-REGION.
FT	REGION	4	15	H-REGION.
FT	REGION	16	21	C-REGION.
FT	TOPO_DOM	22	212	NON CYTOPLASMIC.

¿Posee péptido  
señal o dominios  
transmembrana?



The probability data used in the plot is found [here](#), and the gnuplot script is [here](#).

¿Preguntas, sugerencias,  
comentarios?