

Basic Local Alignment Search Tool (BLAST)

Actualizado en: 27/02/2023

Objetivos

- Comprender el funcionamiento de BLAST.
- Conocer los diferentes tipos de búsquedas posibles.
- Capacidad de utilizar la interfaz web de NCBI BLAST.

BLAST, el algoritmo

Contexto histórico

- Lipman & Pearson. 1985. Rapid and sensitive protein similarity searches. *Science* 227(4693):1435-1441.

“One of the most rigorous programs for comparing amino acid sequences, SEQHP (5), requires more than 8 hours to compare a 200-residue protein to the 500,000-residue NBRF (National Biomedical Research Foundation) protein library on the VAX 11/750 computer.”

- Altschul et al. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215:403-410.

Para qué sirve

- Comparar dos secuencias y encontrar similitudes locales (como Smith-Waterman).
- Buscar secuencias parecidas a una *query* en una base de datos (*target*).
- Múltiples aplicaciones:
 - Identificación de bacterias usando 16S rRNA.
 - Diseño de cebadores para amplificar un gen.
 - Anotación de regiones codificantes o proteínas.
 - Identificación de dominios en una proteína.
 - Recopilar secuencias homólogas para crear una filogenia.

Para qué sirve

Partiendo de una secuencia nucleotídica o de proteína:

- ¿Con qué otras está relacionada? ¿Cuál es su función? (Homología, dominios conservados).
- ¿Está ya presente en la base de datos? (Identificación).
- ¿Dónde se encuentra o cómo está organizada? (Anotación, ensamblaje).

Qué meritos tiene

- Rapidez.
- Sensibilidad.
- Estadística.

Cómo funciona

1. Detecta e ignora regiones repetitivas o de *baja complejidad* de la *query*.
2. Hace una lista de palabras de k letras de la *query* ($k = 11$ para DNA):

PQGEFG
PQG
QGE
GEF
EFG

3. Añade a la lista palabras *vecinas* que alinearan con puntuación de al menos T .

Cómo funciona

4. Busca las palabras de la lista entre las secuencias de la base de datos (indexadas).
5. Alarga la *semilla* de los alineamientos encontrados (*High-scoring Segment Pair*, HSP).

Query sequence: R P P Q G L F
Database sequence: D P P E G V V

└─>Exact match is scanned.

Score: -2 7 7 2 6 1 -1

└─>HSP

Optimal accumulated score = $7+7+2+6+1 = 23$

Cómo funciona

6. Enumera HSPs con puntuación mayor de la que se produciría por azar.
7. Evalúa la significación de los HSPs.
8. Combina dos o más HSP en uno.
9. Muestra el alineamiento local Smith-Waterman de cada resultado.
10. Enumera los resultados con valor E menor o igual a un cierto umbral.

Evaluación estadística de los resultados

La distribución de puntuaciones de HSPs entre dos secuencias de longitudes m y n está descrita por los parámetros K y λ . El número esperado de HSPs con una puntuación de al menos S (**valor E**) es:

$$E = K m n e^{-\lambda S}$$

En una búsqueda en una base de datos, n es la longitud total de la base de datos entera. Los parámetros K y λ deben ser estimados mediante permutaciones. La probabilidad de observar al menos un HSP con una puntuación de al menos S por casualidad, es (distribución de Poisson):

$$P = 1 - e^{-E}$$

Este sería el valor p .

Test

- El mismo HSP, en bases de datos de tamaños diferentes, ¿dónde tendrá un valor E mayor?
- ¿Qué es mejor, una base de datos grande o una pequeña?
- ¿Cómo afectará el tamaño de palabra, k , a la sensibilidad? ¿Y al tiempo de ejecución?
- ¿Para qué sirve conocer la distribución teórica de puntuaciones de HSPs?

Ejemplos

NCBI BLAST

blastn

blastp

blastx

tblastn

tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

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

Query subrange [?](#)

From

To

Or, upload file

Browse...

No file selected. [?](#)

Job Title

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

☐ Align two or more sequences [?](#)

Choose Search Set

Database

☐ Standard databases (nr etc.): ☒ rRNA/ITS databases ☐ Genomic + transcript databases ☐ Betacoronavirus

16S ribosomal RNA sequences (Bacteria and Archaea)

[Targeted Loci Project Information](#)

Organism

Optional

☐ exclude [Add organism](#)

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

☐ Sequences from type material

Entrez Query

Optional

[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 [?](#)

NCBI BLAST

Algorithm parameters

General Parameters

Max target sequences

100

Select the maximum number of aligned sequences to display ?

Short queries

☒ Automatically adjust parameters for short input sequences ?

Expect threshold

0.05

?

Word size

28

?

Max matches in a query range

0

?

Scoring Parameters

Match/Mismatch Scores

1,-2

?

Gap Costs

Linear

?

Filters and Masking

Filter

☒ Low complexity regions ?

☐ Species-specific repeats for: Homo sapiens (Human) ?

Mask

☒ Mask for lookup table only ?

☐ Mask lower case letters ?

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

Identificación de especies

Secuencias parciales de genes 16S rRNA de bacterias no cultivadas:

AM179943.1
AM179942.1
AM179941.1
AM179940.1
AM179939.1
AM179938.1
AM179937.1
AM179936.1
AM179935.1
AM179934.1
AM179933.1
AM179932.1
AM179931.1

Identificación de especies

BLAST® » blastn suite » results for RID-ZT2S1WU801N

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

[Edit Search](#) [Save Search](#) [Search Summary](#)

[How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title

AM179943:Uncultured bacterium partial 16S...

RID

ZT2S1WU801N Search expires on 02-28 17:50 pm [Download All](#)

Results for

13:embjAM179931.1 Uncultured bacterium partial 16S rRNA gene, c

Program

BLASTN [Citation](#)

Database

rRNA_typestrains/16S_ribosomal_RNA [See details](#)

Query ID

AM179931.1

Description

Uncultured bacterium partial 16S rRNA gene, clone M55

Molecule type

dna

Query Length

493

Other reports

[Distance tree of results](#) [MSA viewer](#)

Filter Results

Organism only top 20 will appear ☐ exclude

[Add organism](#)

Percent Identity

to

E value

to

Query Coverage

to

Filter

Reset

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

[Download](#) [Select columns](#) [Show](#)

☒ select all 585 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"/>	Malacoplasma muris strain RIII-4 16S ribosomal RNA, partial sequence	Malacoplasma m...	527	527	92%	3e-149	87.10%	1462	NR_044664.2
<input checked="" type="checkbox"/>	Malacoplasma microti strain IL371 16S ribosomal RNA, partial sequence	Malacoplasma m...	520	520	92%	5e-147	87.28%	1435	NR_025055.1
<input checked="" type="checkbox"/>	Malacoplasma iowae 695 16S ribosomal RNA, partial sequence	Malacoplasma io...	486	486	91%	5e-137	85.34%	1461	NR_044669.2
<input checked="" type="checkbox"/>	Ureaplasma cati strain F2 16S ribosomal RNA, partial sequence	Ureaplasma cati	470	470	91%	5e-132	85.53%	1476	NR_115604.1
<input checked="" type="checkbox"/>	Ureaplasma urealyticum serovar 8 str. ATCC 27618 16S ribosomal RNA, partial sequence	Ureaplasma urea...	464	464	91%	2e-130	85.34%	1435	NR_041710.1
<input checked="" type="checkbox"/>	Ureaplasma felinum strain FT2-B 16S ribosomal RNA, partial sequence	Ureaplasma felin...	464	464	91%	2e-130	85.28%	1473	NR_025879.1
<input checked="" type="checkbox"/>	Ureaplasma diversum strain A417 16S ribosomal RNA, partial sequence	Ureaplasma dive...	459	459	91%	1e-128	85.13%	1478	NR_025878.1
<input checked="" type="checkbox"/>	Ureaplasma gallorale strain D6-1 16S ribosomal RNA, partial sequence	Ureaplasma gall...	453	453	91%	5e-127	84.91%	1455	NR_026027.1
<input checked="" type="checkbox"/>	Ureaplasma canigenitalium strain D6P-C 16S ribosomal RNA, partial sequence	Ureaplasma cani...	449	449	91%	7e-126	84.70%	1480	NR_025877.1
<input checked="" type="checkbox"/>	Ureaplasma parvum strain ATCC 27815 16S ribosomal RNA, complete sequence	Ureaplasma parv...	435	435	91%	2e-121	84.19%	1517	NR_074762.2
<input checked="" type="checkbox"/>	Ureaplasma parvum serovar 3 strain ATCC 27815 16S ribosomal RNA, partial sequence	Ureaplasma parv...	435	435	91%	2e-121	84.19%	1439	NR_027532.1
<input checked="" type="checkbox"/>	Mycoplasma fastidiosum strain 4822 16S ribosomal RNA, partial sequence	Mycoplasmaoides...	390	390	87%	4e-108	83.14%	1455	NR_024987.1
<input checked="" type="checkbox"/>	Mycoplasma testudinis strain 01008 16S ribosomal RNA, partial sequence	Mycoplasma test...	383	383	91%	7e-106	82.11%	1416	NR_029175.1

17/24

Diseño de cebadores para el gen MPO

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Nucleotide Search

Advanced Help

GenBank

Send to:

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Show in Genome Data Viewer

Articles about the MPO gene

Paediatric-onset generalized pustular psoriasis secondary to myeloper [Clin Exp Dermatol. 2023]

Associations between myeloperoxidase and paraoxonase-1 a [BMC Cardiovasc Disord. 2022]

iPS cells from Chediak-Higashi syndrome patients recapitulate the giant ϵ [Pediatr Int. 2022]

See all...

Go to: ☐

LOCUS NM_000250 3216 bp mRNA linear PRI 12-FEB-2023

DEFINITION Homo sapiens myeloperoxidase (MPO), mRNA.

ACCESSION NM_000250

VERSION NM_000250.2

KEYWORDS RefSeq; MANE Select.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 3216)

AUTHORS Perez-Feal P, Moreiras-Arias N, Vega A, Suarez-Penaranda JM, Lopez-Franco MM, Rodriguez-Blanco I, Sanchez-Aguilar D and Vazquez-Osorio I.


TITLE Paediatric-onset generalized pustular psoriasis secondary to myeloperoxidase mutations

JOURNAL Clin Exp Dermatol 48 (2), 130-132 (2023)

PUBMED [36730508](#)

REMARK GeneRIF: Paediatric-onset generalized pustular psoriasis secondary

Diseño de cebadores para el gen MPO

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

Primer-BLAST A tool for finding specific primers

Finding primers specific to your PCR template (using Primer3 and BLAST).

Primers for target on one template Primers common for a group of sequences

PCR Template Retrieve recent results Publication Tips for finding specific primers

Enter accession, gi, or FASTA sequence (A refseq record is preferred) ? Clear
NM_000250.2

Or, upload FASTA file Browse... No file selected.

Range ? Clear
Forward primer From To
178
Reverse primer 2415

Primer Parameters

Use my own forward primer (5'→3' on plus strand) ? Clear
Use my own reverse primer (5'→3' on minus strand) ? Clear

PCR product size Min Max
70 1000

of primers to return 10

Primer melting temperatures (T_m) Min Opt Max Max T_m difference
57.0 60.0 63.0 3 ?

Exon/intron selection A refseq mRNA sequence as PCR template input is required for options in the section ?

Exon junction span No preference ?

Exon junction match Min 5' match Min 3' match Max 3' match
7 4 8

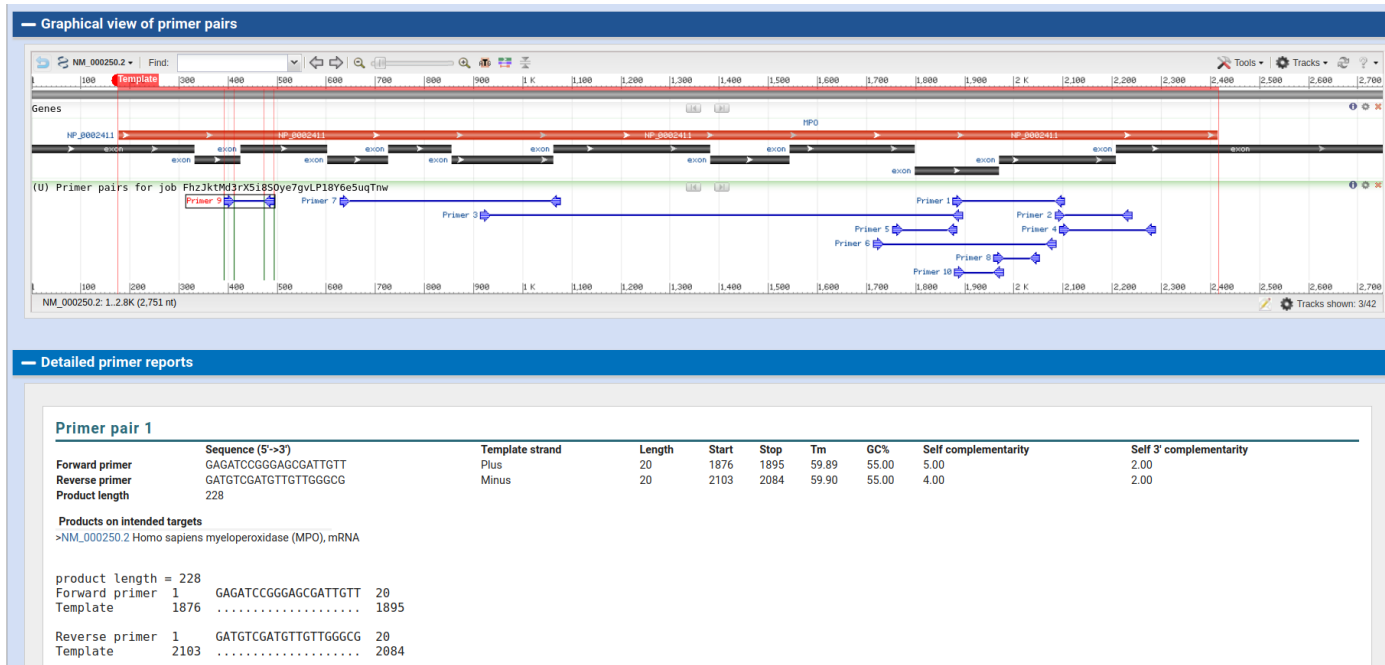
Minimal and maximal number of bases that must anneal to exons at the 5' or 3' side of the junction ?

☐ Primer pair must be separated by at least one intron on the corresponding genomic DNA ?

Intron inclusion

Intron length range Min Max
1000 10000 ?

Diseño de cebadores para el gen MPO



Filogenia mitocondrial de hominidos

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Nucleotide Nucleotide Search

Advanced Help

GenBank

Send to: Change region shown

Customize view

Lemur catta mitochondrion complete mitochondrial genome

GenBank: AJ421451.1

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

Go to: ☑

LOCUS AJ421451 17036 bp DNA circular PRI 26-JUL-2016

DEFINITION Lemur catta mitochondrion complete mitochondrial genome.

ACCESSION AJ421451

VERSION AJ421451.1

KEYWORDS 12S ribosomal RNA; 12S rRNA gene; atpase6 gene; ATPase6 protein; atpase8 gene; ATPase8 protein; COI gene; COII gene; COIII gene; complete genome; control region; cytb gene; cytochrome b; D-loop; mitochondrion; NADH dehydrogenase subunit 1; NADH dehydrogenase subunit 2; NADH dehydrogenase subunit 3; NADH dehydrogenase subunit 4; NADH dehydrogenase subunit 4L; NADH dehydrogenase subunit 5;

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

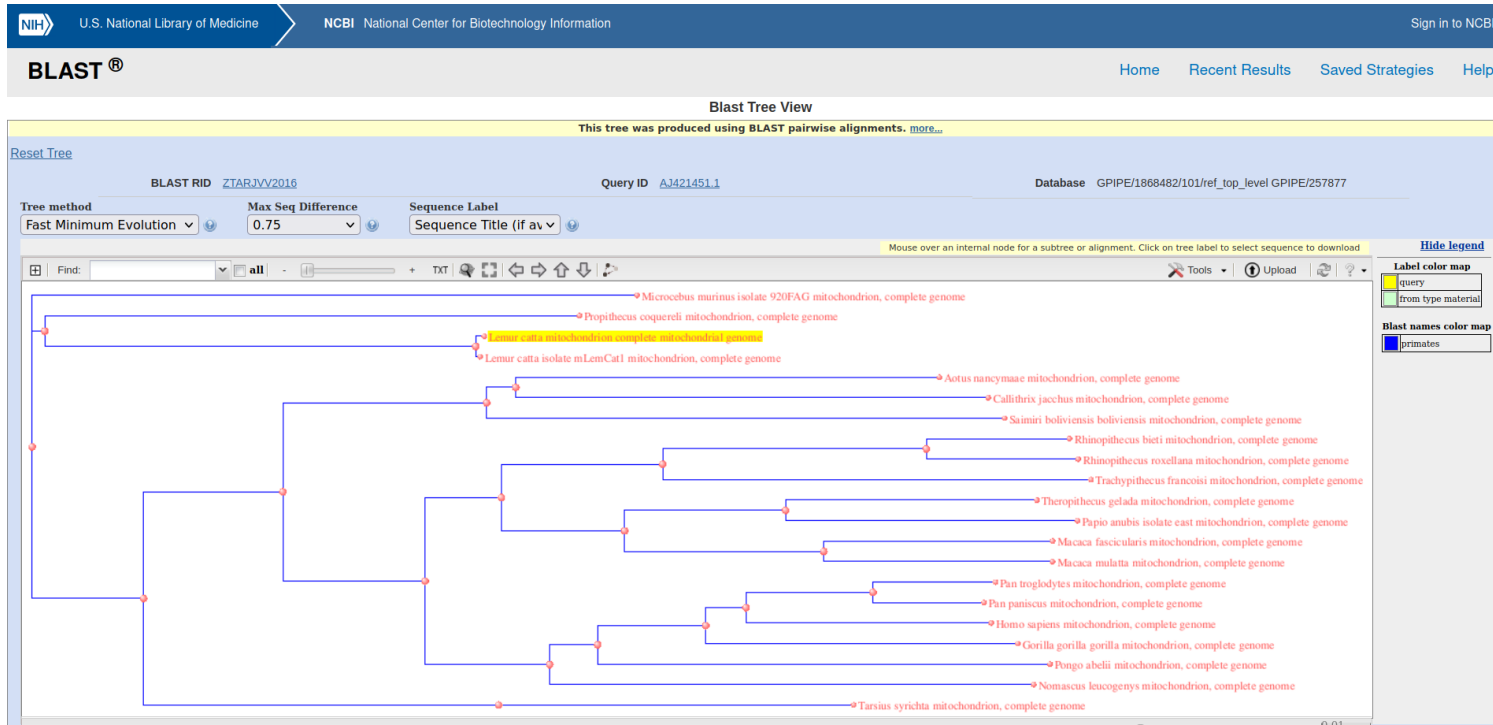
Find in this Sequence

Related information


Protein

Usamos el mtDNA de *Lemur catta* como outgroup

Filogenia mitocondrial de hominidos



Anotación de un contig de DNA

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

Log in

Nucleotide Nucleotide Search Advanced Help

GenBank Send to Change region shown Customize view

Marine metagenome ERR594294-C1343, whole genome shotgun sequence
GenBank: MIZB01000007.1
[FASTA](#) [Graphics](#)

Go to: Find regions of similarity between this sequence and other sequences using BLAST.

LOCUS MIZB01000007 13728 bp DNA linear

DEFINITION Marine metagenome ERR594294-C1343, whole genome shotgun sequence

ACCESSION MIZB01000007 [MIZB01000000](#)

VERSION MIZB01000007.1

DBLINK BioProject: [PRJNA335308](#)
BioSample: [SAMN05440186](#)

KEYWORDS WGS; ENV.

SOURCE marine metagenome

ORGANISM [marine metagenome](#)
unclassified sequences; metagenomes; ecological metagenomes.

REFERENCE 1 (bases 1 to 13728)

AUTHORS Haro-Moreno,J.M., Rodriguez-Valera,F., Lopez-Garcia,P., Moreira,D. and Martin-Cuadrado,A.-B.

TITLE New Insights into Marine Group III Euryarchaeota, from dark to light

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 13728)

AUTHORS Haro-Moreno,J.M., Rodriguez-Valera,F., Lopez-Garcia,P., Moreira,D. and Martin-Cuadrado,A.-B.

TITLE Direct Submission

JOURNAL Submitted (05-SEP-2016) Evolutionary Genomics Group, Universidad

Analyze this sequence
[Run BLAST](#)
[Pick Primers](#)
[Find in this Sequence](#)

Recent activity Turn Off Clear

Marine metagenome ERR594294-C1343, whole genome shotgun sequence Nucleotide

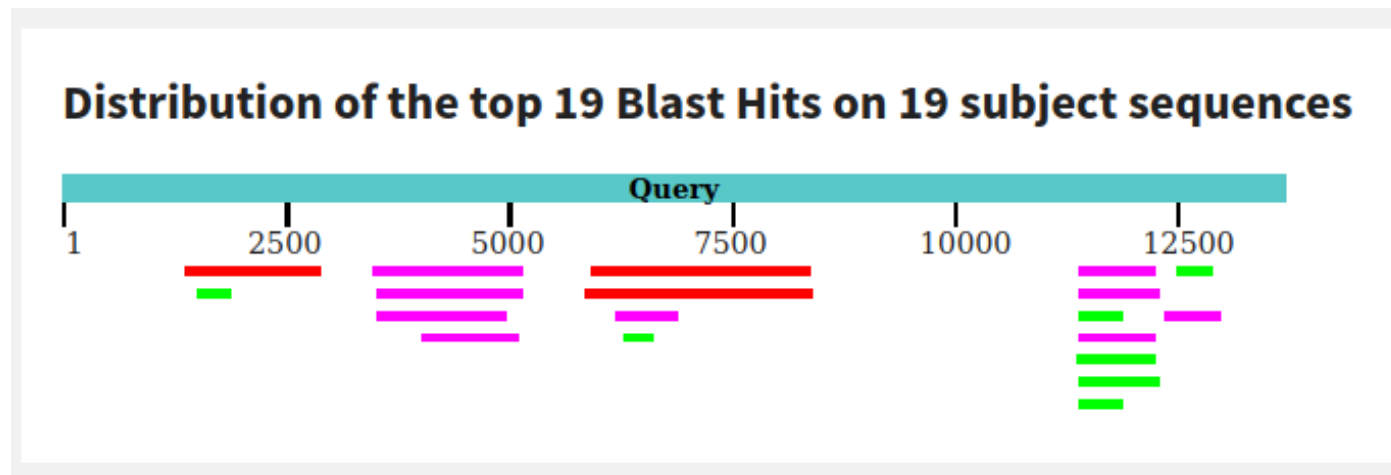
freshwater metagenome genome assembly, contig: UFOP-RE-15AUG16-175-C1 Nucleotide

unannotated metagenome shotgun (66939) Nucleotide

ERR594294 (7) SRA

marine metagenome JCVI_SCAF_1096627126648 genor Nucleotide

Anotación de un contig de DNA



Los resultados del BLASTX sugieren que hay cuatro o cinco genes codificantes de proteínas en el contig.