Task 1 DNA/Protein Sequence Analysis

Sequence Similarity Analysis between Streptococcus pyogenes M-Protein and Human Cardiac Myosin

Introduction

Protein sequence comparison is one of the most important bioinformatics techniques used to study evolutionary relationships, structural similarity, and potential functional overlap between proteins.

The **M protein** of *Streptococcus pyogenes* is a major virulence factor. It helps the bacteria evade the host immune system by resisting phagocytosis. However, certain regions of this bacterial protein show **molecular mimicry** with human heart proteins, especially **cardiac myosin**. This similarity is believed to play a key role in **autoimmune cross-reactivity**.

To explore this concept at the molecular level, a **protein sequence comparison** was performed between the M protein of *S. pyogenes* and human cardiac myosin using NCBI databases and BLAST tools.

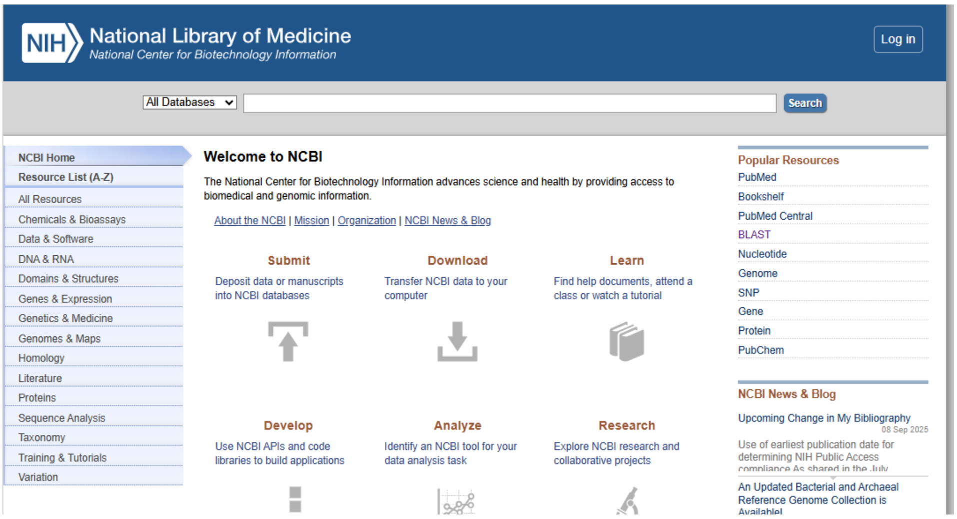
Background Information

1. **M Protein (Streptococcus pyogenes)**
   - A surface-exposed protein responsible for virulence.
   - Highly variable in sequence, with over 200 known serotypes.
   - Some conserved regions show cross-reactivity with human tissues.
2. **Cardiac Myosin (Human)**
   - A major contractile protein in cardiac muscle.
   - Plays a key role in muscle contraction and heart function.
   - Has both conserved functional domains and variable regions.

Procedure:

Steps to Retrieve FASTA sequences from NCBI:

1. Open the **NCBI** website



Welcome to NCBI

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

Deposit data or manuscripts into NCBI databases

Transfer NCBI data to your computer

Find help documents, attend a class or watch a tutorial

Use NCBI APIs and code libraries to build applications

Identify an NCBI tool for your data analysis task

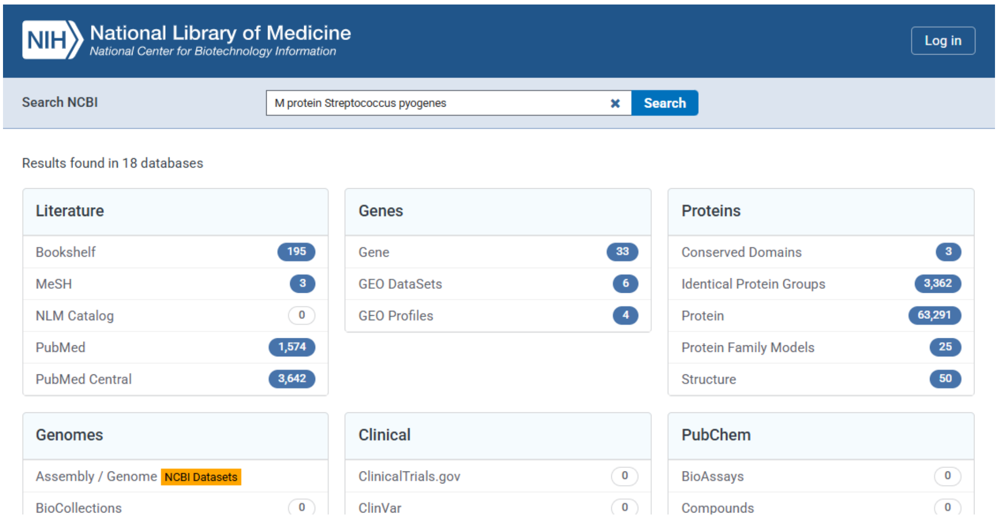
Explore NCBI research and collaborative projects

Upcoming Change in My Bibliography
08 Sep 2025

Use of earliest publication date for determining NIH Public Access compliance As shared in the July

An Updated Bacterial and Archaeal Reference Genome Collection is Available!

2. In the search bar, type the **protein name**


NIH National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
Search

M protein Streptococcus pyogenes

Results found in 18 databases

Literature
Bookshelf 195
MeSH 3
NLM Catalog 0
PubMed 1,574
PubMed Central 3,642

Genes
Gene 33
GEO DataSets 6
GEO Profiles 4

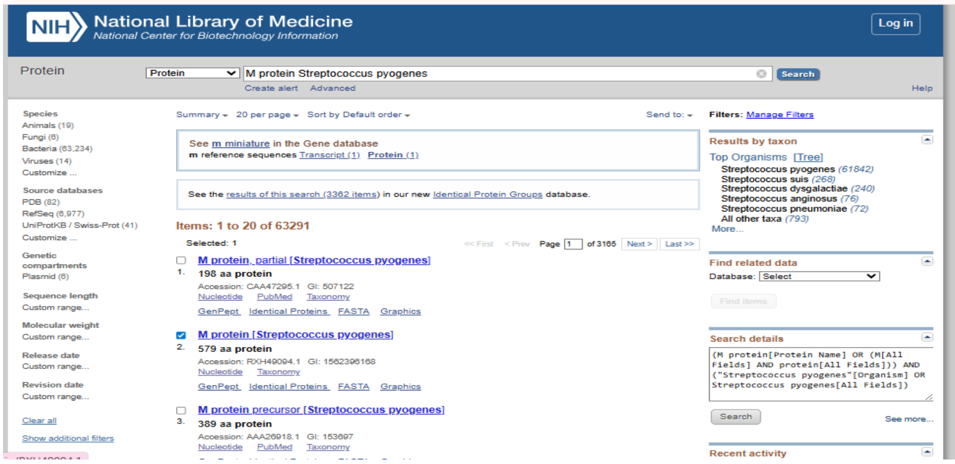
Proteins
Conserved Domains 3
Identical Protein Groups 3,362
Protein 63,291
Protein Family Models 25
Structure 50

Genomes
Assembly / Genome NCBI Datasets
BioCollections 0

Clinical
ClinicalTrials.gov 0
ClinVar 0

PubChem
BioAssays 0
Compounds 0

3. From the results, open the **Protein** record of the selected protein.



M protein Streptococcus pyogenes

Species
Animals (19)
Fungi (6)
Bacteria (63,234)
Viruses (14)
Customize ...

Source databases
PDB (82)
RefSeq (6,977)
UniProtKB / Swiss-Prot (41)
Customize ...

Genetic compartments
Plasmid (6)

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

See m miniature in the Gene database
m reference sequences Transcript (1) Protein (1)

See the results of this search (3362 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 63291

Selected: 1

Page 1 of 3165

☐ 1. M protein, partial [Streptococcus pyogenes]
198 aa protein
Accession: CAA47295.1 GI: 507122
Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

☑ 2. M protein [Streptococcus pyogenes]
579 aa protein
Accession: RXH49094.1 GI: 1562396168
Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

☐ 3. M protein precursor [Streptococcus pyogenes]
389 aa protein
Accession: AAA26918.1 GI: 153697
Nucleotide PubMed Taxonomy

Filters: Manage Filters

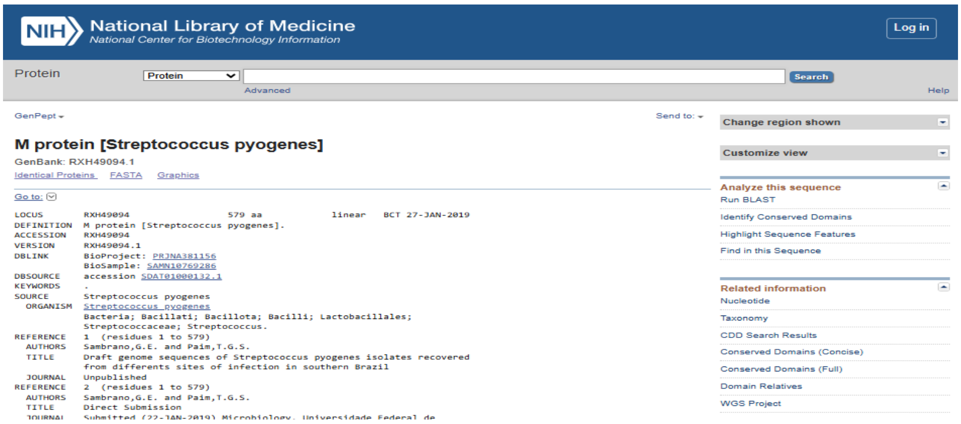
Results by taxon
Top Organisms [Tree]
Streptococcus pyogenes (61842)
Streptococcus suis (268)
Streptococcus dysgalactiae (240)
Streptococcus anginosus (76)
Streptococcus pneumoniae (72)
All other taxa (793)
More...

Find related data
Database: Select

Search details
(M protein[Protein Name] OR (M[All Fields] AND protein[All Fields])) AND ("Streptococcus pyogenes"[Organism] OR Streptococcus pyogenes[All Fields])

Recent activity

4. Scroll to the top-right section and click on the **"FASTA"** link.


NIH National Library of Medicine
National Center for Biotechnology Information
Log in
Protein
Protein
Search
Advanced
Help
GenPept
Send to:
Go to:

M protein [Streptococcus pyogenes]

GenBank: RXH49094.1

Identical Proteins FASTA Graphics

```
LOCUS       RXH49094                 579 aa            linear   BCT 27-JAN-2019
DEFINITION  M protein [Streptococcus pyogenes].
ACCESSION   RXH49094
VERSION     RXH49094.1
DBLINK      BioProject: PRJNA381156
            BioSample: SAMN10769286
DBSOURCE    accession SDAT01000132.1
KEYWORDS    .
SOURCE      Streptococcus pyogenes
  ORGANISM  Streptococcus pyogenes
            Bacteria; Bacillati; Bacillota; Bacilli; Lactobacillales;
            Streptococcaceae; Streptococcus.
REFERENCE   1  (residues 1 to 579)
  AUTHORS   Sambrano,G.E. and Paim,T.G.S.
  TITLE     Draft genome sequences of Streptococcus pyogenes isolates recovered
            from differents sites of infection in southern Brazil
  JOURNAL   Unpublished
REFERENCE   2  (residues 1 to 579)
  AUTHORS   Sambrano,G.E. and Paim,T.G.S.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-JAN-2019) Microbiology, Universidade Federal de
```

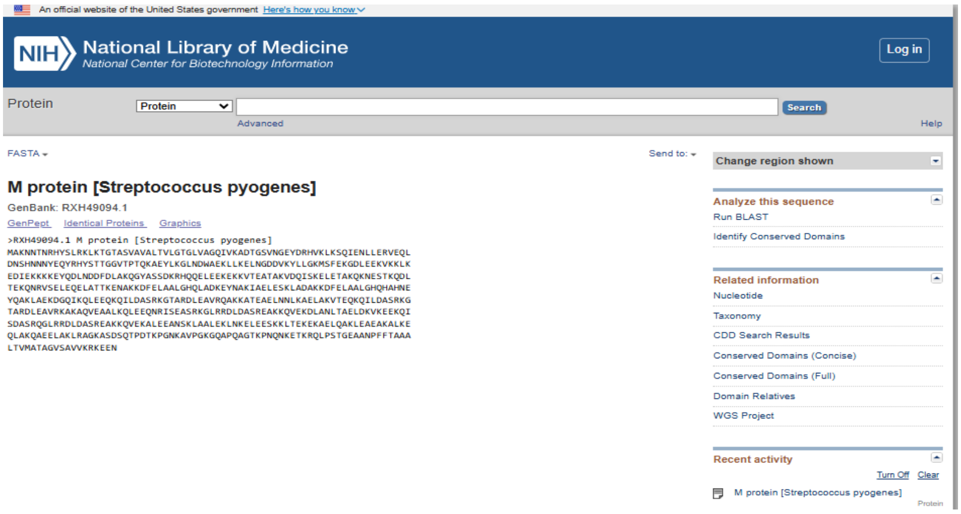
Change region shown

Customize view

Analyze this sequence
Run BLAST
Identify Conserved Domains
Highlight Sequence Features
Find in this Sequence

Related information
Nucleotide
Taxonomy
CDD Search Results
Conserved Domains (Concise)
Conserved Domains (Full)
Domain Relatives
WGS Project

5. The FASTA format sequence will appear; you may **copy the sequence directly** for use or to download the sequence, click on **"Graphics"** at the top of the record.



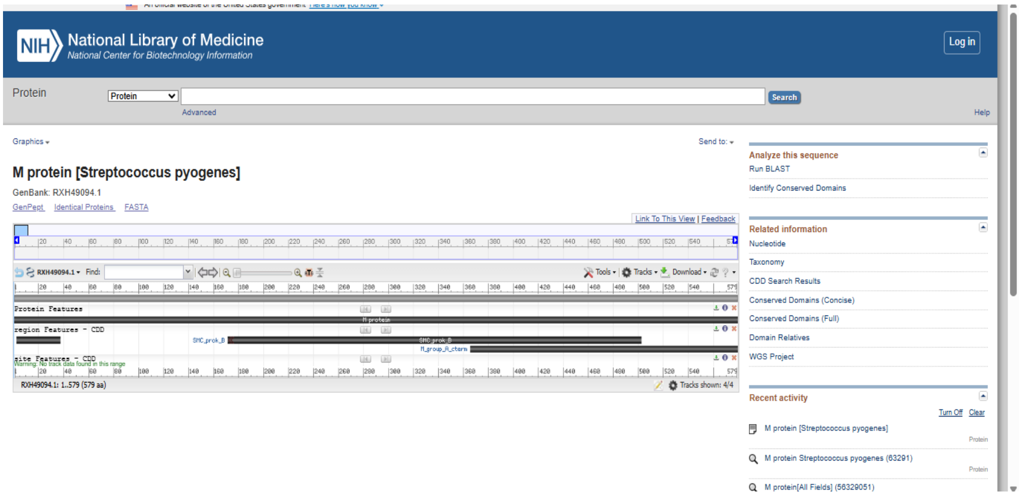
M protein [Streptococcus pyogenes]

GenBank: RXH49094.1

GenPept Identical Proteins Graphics

```
>RXH49094.1 M protein [Streptococcus pyogenes]
MAKNNTNRHYSLRKLKTGTASVAVALTVLGTGLVAGQIVKADTGSVNGEYDRHVKLKSQIENLLERVEQL
DNSHNNNYEQYRHYSTTGGVTPTQKAEYLKGLNDWAEKLLKELNGDDVKYLLGKMSFEKGDLEEKVKKLK
EDIEKKKKEYQDLNDDFDLAKQGYASSDKRHQQELEEKEKKVTEATAKVDQISKELETAKQKNESTKQDL
TEKQNRVSELEQELATTKENAKKDFELAALGHQLADKEYNAKIAELESKLADAKKDFELAALGHQHAHNE
YQAKLAEKDGQIKQLEEQKQILDASRKGTARDLEAVRQAKKATEAELNNLKAELAKVTEQKQILDASRKG
TARDLEAVRKAKAQVEAALKQLEEQNRISEASRKGLRRDLDASREAKKQVEKDLANLTAELDKVKEEKQI
SDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKLNKELEESKKLTEKEKAELQAKLEAEAKALKE
QLAKQAEELAKLRAGKASDSQTPDTKPGNKAVPGKGQAPQAGTKPNQNKETKRQLPSTGEAANPFFTAAA
LTVMATAGVSAVVKRKEEN
```

6. Select **"Download"** and choose **FASTA format** to save the file.

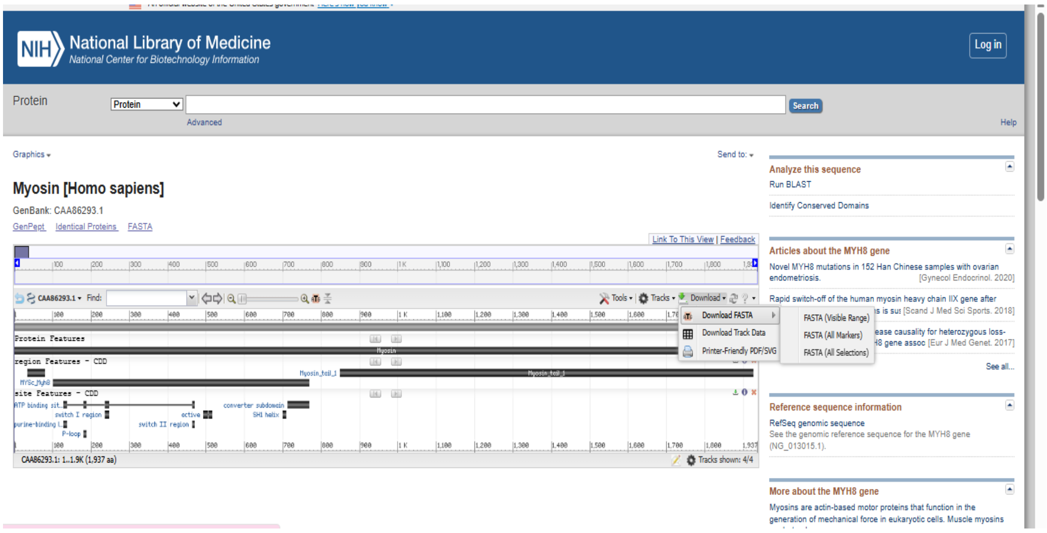


M protein [Streptococcus pyogenes]

GenBank: RXH49094.1

GenPept Identical Proteins FASTA

7. Repeat the same steps for the **second protein** (e.g., Myosin).



Myosin [Homo sapiens]

GenBank: CAA86293.1

GenPept Identical Proteins FASTA

Novel MYH8 mutations in 152 Han Chinese samples with ovarian endometriosis. [Gynecol Endocrinol. 2020]

Rapid switch-off of the human myosin heavy chain IIX gene after ...s is su: [Scand J Med Sci Sports. 2018]

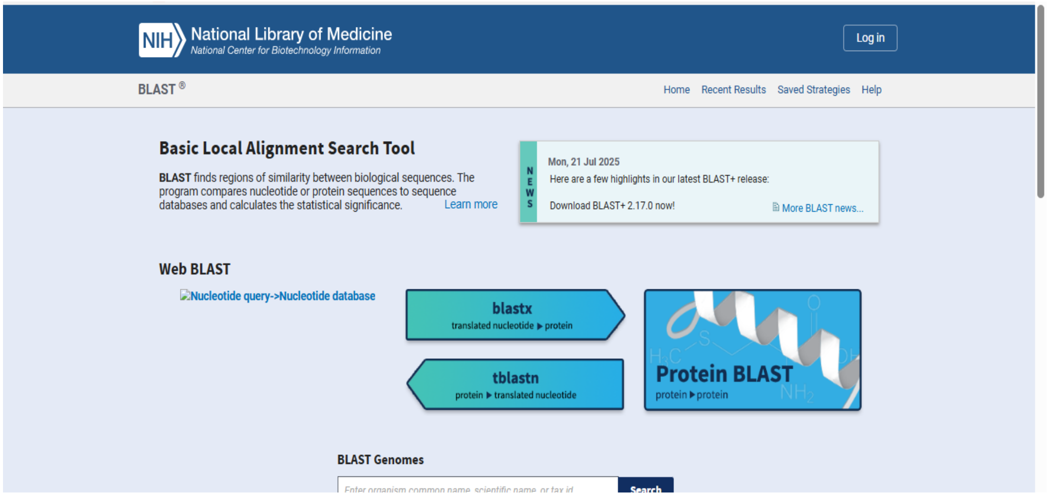
...ease causality for heterozygous loss- ...H8 gene assoc [Eur J Med Genet. 2017]

See the genomic reference sequence for the MYH8 gene (NG_013015.1).

Myosins are actin-based motor proteins that function in the generation of mechanical force in eukaryotic cells. Muscle myosins

Steps to Perform Multiple Sequence Alignment Using NCBI BLAST

1. Open the **BLAST** page on NCBI



NIH National Library of Medicine
National Center for Biotechnology Information

Log in

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

NEWS

Mon, 21 Jul 2025

Here are a few highlights in our latest BLAST+ release:

Download BLAST+ 2.17.0 now!

More BLAST news...

Web BLAST

Nucleotide query->Nucleotide database

blastx
translated nucleotide ▸ protein

tblastn
protein ▸ translated nucleotide

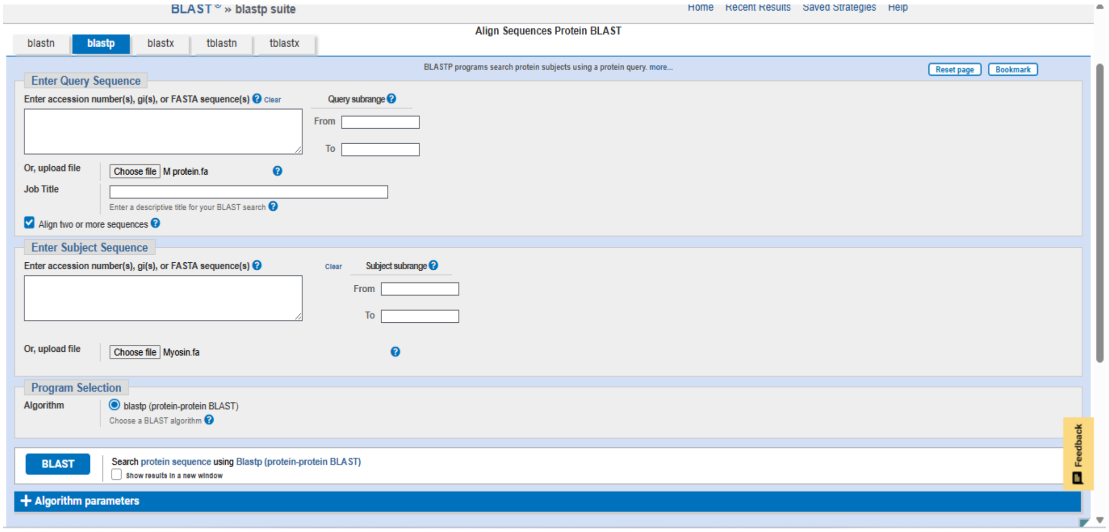
Protein BLAST
protein ▸ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

2. Select **Protein BLAST (blastp)**.
3. Under the query box, click **“Align two or more sequences”**.
4. Upload or paste the **FASTA sequences** of both proteins (Actin and Myosin).
5. Scroll down and click **“BLAST”** to run the alignment.
6. Wait for the results to load; the alignment output will be displayed below.



BLAST ® » blastp suite

Home Recent Results Saved Strategies Help

Align Sequences Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein subjects using a protein query. more...

Reset page | Bookmark

Enter Query Sequence

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

Query subrange

From

To

Or, upload file: Choose file M protein.fa

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

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

Clear

Subject subrange

From

To

Or, upload file: Choose file Myosin.fa

Program Selection

Algorithm: ● blastp (protein-protein BLAST)

Choose a BLAST algorithm

BLAST

Search protein sequence using Blastp (protein-protein BLAST)

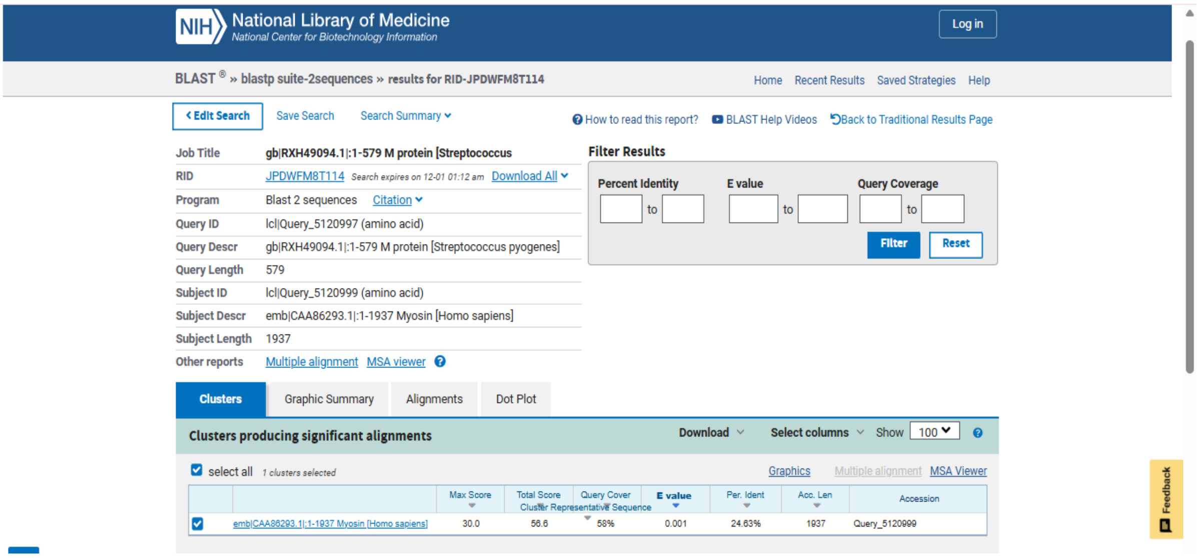
□ Show results in a new window

+ Algorithm parameters

Feedback

Understanding the BLAST Alignment Output

- **Matching letters** → identical amino acids at that position.
- **Dots (.)** → similar amino acids (conserved substitution).
- **Spaces ( )** → mismatch; no similarity at that position.
- **Query** → first protein sequence you input.
- **Subject** → second protein sequence.
- **Score / Bit score** → strength of alignment (higher = better).
- **E-value** → significance of the match (lower = more significant).



NIH National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ® » blastp suite-2sequences » results for RID-JPDWFM8T114
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	gb|RXH49094.1|:1-579 M protein [Streptococcus
RID	JPDWFM8T114 *Search expires on 12-01 01:12 am* Download All
Program	Blast 2 sequences Citation
Query ID	lcl|Query_5120997 (amino acid)
Query Descr	gb|RXH49094.1|:1-579 M protein [Streptococcus pyogenes]
Query Length	579
Subject ID	lcl|Query_5120999 (amino acid)
Subject Descr	emb|CAA86293.1|:1-1937 Myosin [Homo sapiens]
Subject Length	1937
Other reports	Multiple alignment MSA viewer

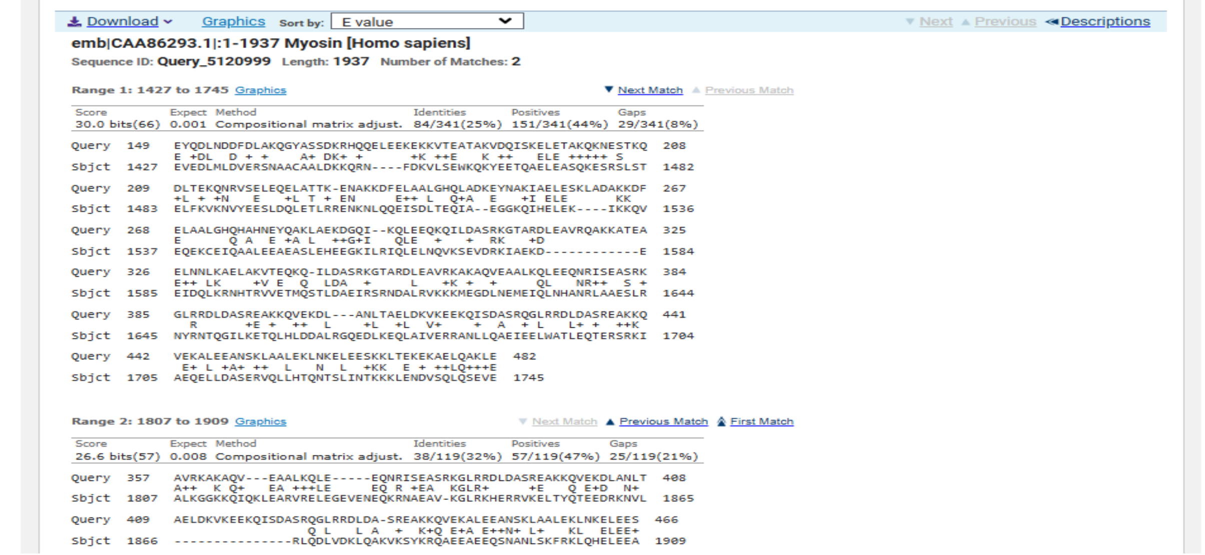
Filter Results

Percent Identity [ ] to [ ] E value [ ] to [ ] Query Coverage [ ] to [ ]

Clusters producing significant alignments

☑ select all *1 clusters selected*

	Cluster Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	emb|CAA86293.1|:1-1937 Myosin [Homo sapiens]	30.0	56.6	58%	0.001	24.63%	1937	Query_5120999



emb|CAA86293.1|:1-1937 Myosin [Homo sapiens]

Sequence ID: **Query_5120999** Length: **1937** Number of Matches: **2**

Range 1: 1427 to 1745 Graphics ▼ Next Match ▲ Previous Match

```
 Score          Expect  Method                         Identities     Positives       Gaps
 30.0 bits(66)  0.001   Compositional matrix adjust.   84/341(25%)    151/341(44%)    29/341(8%)

Query  149    EYQDLNDDFDLAKQGYASSDKRHQQELEEKEKKVTEATAKVDQISKELETAKQKNESTKQ  208
              E +DL  D + +    A+ DK+ +      +K ++E   K ++   ELE +++++ S
Sbjct  1427   EVEDLMLDVERSNAACAALDKKQRN----FDKVLSEWKQKYEETQAELEASQKESRSLST  1482

Query  209    DLTEKQNRVSELEQELATTK-ENAKKDFELAALGHQLADKEYNAKIAELESKLADAKKDF  267
              +L + +N   E   +L T + EN     E++ L  Q+A  E   +I ELE      KK
Sbjct  1483   ELFKVKNVYEESLDQLETLRRENKNLQQEISDLTEQIA--EGGKQIHELEK----IKKQV  1536

Query  268    ELAALGHQHAHNEYQAKLAEKDGQI--KQLEEQKQILDASRKGTARDLEAVRQAKKATEA  325
              E      Q A  E +A L  ++G+I   QLE  +   +  RK   +D
Sbjct  1537   EQEKCEIQAALEEAEASLEHEEGKILRIQLELNQVKSEVDRKIAEKD-----------E  1584

Query  326    ELNNLKAELAKVTEQKQ-ILDASRKGTARDLEAVRKAKAQVEAALKQLEEQNRISEASRK  384
              E++ LK    +V E  Q  LDA  +     L   +K +  +     QL   NR++  S +
Sbjct  1585   EIDQLKRNHTRVVETMQSTLDAEIRSRNDALRVKKKMEGDLNEMEIQLNHANRLAAESLR  1644

Query  385    GLRRDLDASREAKKQVEKDL---ANLTAELDKVKEEKQISDASRQGLRRDLDASREAKKQ  441
                R      +E +  ++  L    +L  +L  V+    +  A  + L   L+ +  ++K
Sbjct  1645   NYRNTQGILKETQLHLDDALRGQEDLKEQLAIVERRANLLQAEIEELWATLEQTERSRKI  1704

Query  442    VEKALEEANSKLAALEKLNKELEESKKLTEKEKAELQAKLE  482
               E+ L +A+ ++  L   N  L  +KK  E + ++LQ+++E
Sbjct  1705   AEQELLDASERVQLLHTQNTSLINTKKKLENDVSQLQSEVE  1745
```

Range 2: 1807 to 1909 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

```
 Score          Expect  Method                         Identities     Positives      Gaps
 26.6 bits(57)  0.008   Compositional matrix adjust.   38/119(32%)    57/119(47%)    25/119(21%)

Query  357    AVRKAKAQV---EAALKQLE-----EQNRISEASRKGLRRDLDASREAKKQVEKDLANLT  408
              A++  K Q+   EA +++LE     EQ R +EA  KGLR+     +E   Q E+D  N+
Sbjct  1807   ALKGGKKQIQKLEARVRELEGEVENEQKRNAEAV-KGLRKHERRVKELTYQTEEDRKNVL  1865

Query  409    AELDKVKEEKQISDASRQGLRRDLDA-SREAKKQVEKALEEANSKLAALEKLNKELEES  466
                               Q L   L A  +  K+Q E+A E++N+ L+   KL  ELEE+
Sbjct  1866   ---------------RLQDLVDKLQAKVKSYKRQAEEAEEQSNANLSKFRKLQHELEEA  1909
```