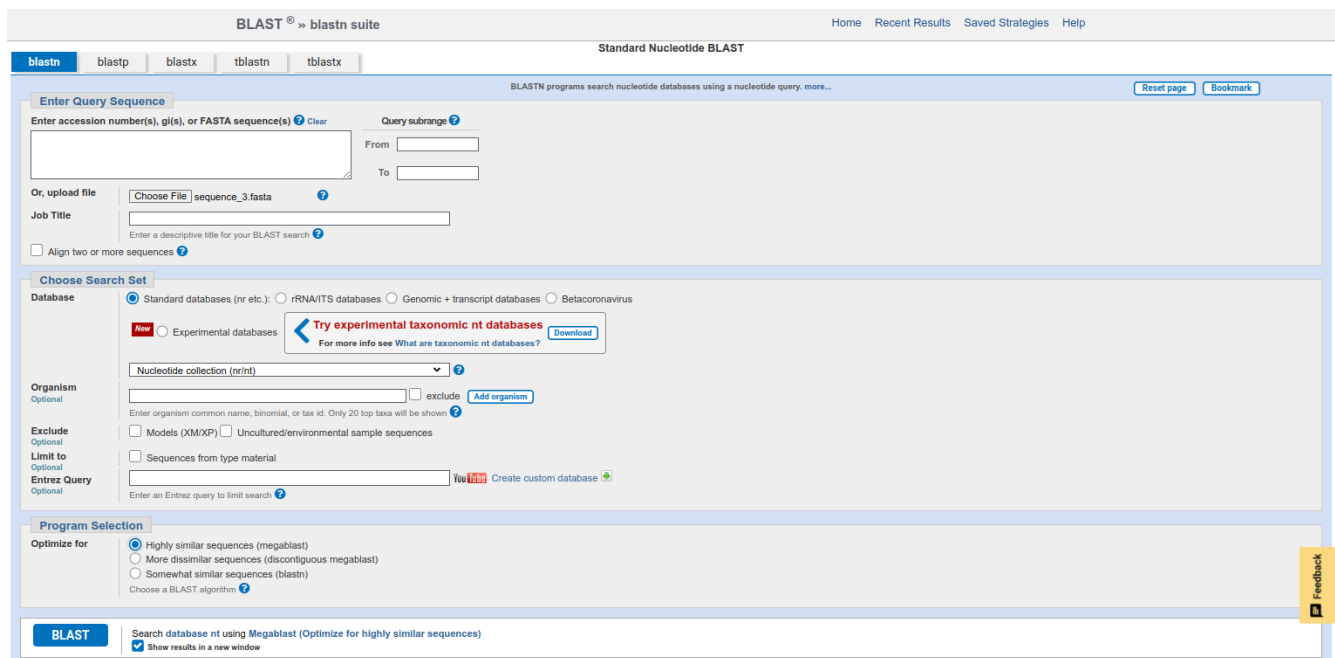


The task is to analyse a specific nucleotide sequence using blast. phylogenetic tree analysis using MEGA, multiple sequence alignment, swiss model repository and open reading frame, and also homologs. we will need to explain and explain the results obtained from the tests above, especially the phylogenetic tree relationships.

BLAST Analysis

BLAST url: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

As the sequence is DNA, we will use Nucleotide BLAST (blastn).



The image shows the NCBI Nucleotide BLAST Search Page. The page has a header with the BLAST logo and navigation links: Home, Recent Results, Saved Strategies, and Help. Below the header, there are tabs for different BLAST programs: blastn, blastp, blastx, tblastn, and tblastx. The 'blastn' tab is selected. The main section is titled 'Standard Nucleotide BLAST'. It contains several input fields and options for configuring the search. The 'Enter Query Sequence' section has a text box for the accession number(s), GI(s), or FASTA sequence(s), and a 'Query subrange' section with 'From' and 'To' fields. There is also an 'Or, upload file' section with a 'Choose File' button and a 'Job Title' field. The 'Choose Search Set' section includes a 'Database' dropdown menu (set to 'Standard databases (nr etc.)'), an 'Organism' dropdown menu (set to 'Nucleotide collection (nr/nt)'), and checkboxes for 'Exclude' and 'Limit to'. The 'Program Selection' section has radio buttons for 'Optimize for' (set to 'Highly similar sequences (megablast)') and a 'Choose a BLAST algorithm' dropdown menu. At the bottom, there is a 'BLAST' button and a checkbox for 'Show results in a new window'.

Figure 1: NCBI Nucleotide Blast Search Page

National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST® » blastn suite » results for RID-PSG2R74X01N

HomeRecent ResultsSaved StrategiesHelp

Edit Search
Save Search
Search Summary

How to read this report?
BLAST Help Videos
Back to Traditional Results Page

Job Titlesequence_3
RIDPSG2R74X01N Search expires on 12-04 15:46 pm Download All
ProgramBLASTN Citation
Database nt See details
Query ID lcl|Query_326195
Description sequence_3
Molecule type dna
Query Length 1858
Other reports Distance tree of results MSA viewer

Filter Results
Organism only top 20 will appear exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity E value Query Coverage
Filter Reset

Descriptions
Graphic Summary
Alignments
Taxonomy

Sequences producing significant alignments
Download Select columns Show 100

select all 8 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"/> Synechococcus elongatus PCC 6301 DNA complete genome	Synechococcus elongatus PCC 6301	3199	3199	93%	0.0	100.00%	2696255	AP008231.1
<input checked="" type="checkbox"/> Synechococcus elongatus PCC 7942 = FACHB-805 strain UTEX 2434 chromosome co...	Synechococcus elongatus PCC 7942...	3193	3193	93%	0.0	99.88%	2695906	CP130602.1
<input checked="" type="checkbox"/> Synechococcus sp. UTEX 2973 complete genome	Synechococcus elongatus UTEX 2973	3193	3193	93%	0.0	99.94%	2690418	CP006471.1
<input checked="" type="checkbox"/> Synechococcus elongatus PCC 6301 chromosome complete sequence	Synechococcus elongatus PCC 6301	3193	3193	93%	0.0	99.88%	2695389	CP085785.1
<input checked="" type="checkbox"/> Synechococcus elongatus PCC 6311 chromosome	Synechococcus elongatus PCC 6311	3193	3193	93%	0.0	99.88%	2689791	CP088958.1
<input checked="" type="checkbox"/> Synechococcus elongatus PCC 7943 chromosome	Synechococcus elongatus PCC 7943	3193	3193	93%	0.0	99.88%	2689559	CP088961.1
<input checked="" type="checkbox"/> Synechococcus elongatus PCC 7942 = FACHB-805 strain PCC 7942 chromosome com...	Synechococcus elongatus PCC 7942...	3193	3193	93%	0.0	99.88%	2695903	CP000100.1
<input checked="" type="checkbox"/> Synechococcus elongatus UTEX 3055 chromosome complete genome	Synechococcus elongatus UTEX 3055	2473	2473	93%	0.0	92.44%	2767524	CP033061.1

Figure 2: NCBI Nucleotide Blast Result Page

Here we get sequence_3 has 100% identity with AP008231.1 (Synechococcus elongatus PCC 6301). Sequence_3 matches position 293963-295693 in AP008231.1 sequence which is a coding sequence having protein_id BAD78445.1. We can download that protein sequence as fasta file to use in SWISS-MODEL Repository.

```

CDS
complement(293963..295693)
/locus_tag="syc0255_c"
/codon_start=1
/transl_table=11
/product="hypothetical protein"
/protein_id="BAD78445.1"
/translation="MQNAFSTADSSDQSLLEIEVGLPNPKPRLKVKGWSWVYGGAG
LLLGVFGLGSLGLFAIQEARRQVDRSLEAIEVAQIIDYYQIRLLFRFNSYLESRQET
DNKLYQLGQEQLLDSIDKLKSFYVQPPNPEQLKEIAALSALIQSKLEEQQNLMNAKQE
ATPSLDASYNVHSQINQIVQNERRILDRVINVDSYRLLTNVLLILGSLGLTLAIA
LYQEQRREREQMSVIDHDYQEKEDTLNHKLKVLQLEQKLSSLLTCSRSTEEIKKILED
FFQRWFPQAQGAVLEISASRDTLVEIARFGELELPSLAMPSPDCWAMRRGECYHSSQAE
FTYPCGLCHHLHGEIIPDNIICIPLOAHEQLIGILHLTNVDPKSQKIVESFGQQLALP
LAVMHLQEQLKQLSYRDSNTALYNRRFLDEILERTLLTAIRRNESRSLGDAPYSVGLI
FLDVKFKDFNTRFGHAVGDQVLQTLGQTMLESRRGEDLACRYGGEEFVLILPGMDE
DMTYQRAEQIRLAVSQKAVSNCRITISLGVAAFPSAGQTPSELLKAANMAMLKAKLNG
RNQTVRISQL"

```

Figure 3: Position of Sequence 3 in AP008231.1

Homolog sequences are:

1. AP008231.1
2. CP130602.1
3. CP006471.1
4. CP085785.1
5. CP088958.1
6. CP088961.1
7. CP000100.1
8. CP033061.1

Multiple Sequence Alignment (MEGA)

Open MEGA and load fasta file having homolog sequences:

ALIGN → Edit/Build Alignment → Retrieve a sequence from a file → Open (homolog sequences.fasta)

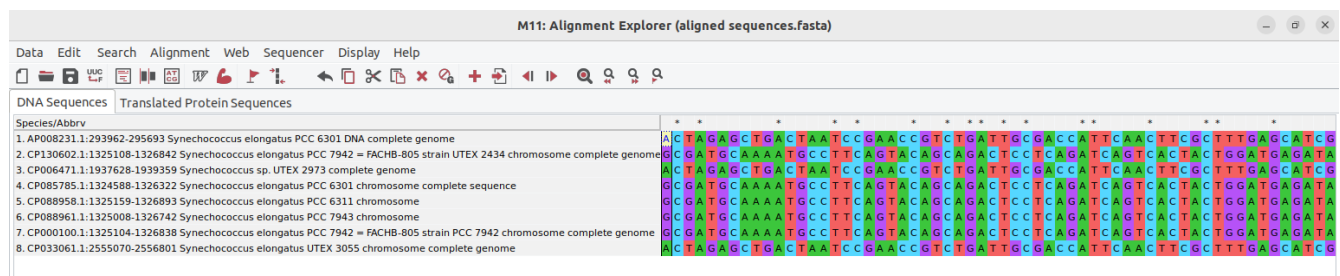


Figure 4: Loaded Homolog Fasta File in MEGA

MUSCLE Alignment Options	
Option	Setting
GAP PENALTIES	
Gap Open	<input checked="" type="checkbox"/> -400.00
Gap Extend	<input checked="" type="checkbox"/> 0.00
MEMORY/ITERATIONS	
Max Memory in MB	<input checked="" type="checkbox"/> 2048
Max Iterations	<input checked="" type="checkbox"/> 16
ADVANCED OPTIONS	
Cluster Method (Iterations 1,2)	<input checked="" type="checkbox"/> Neighbor Joining
Cluster Method (Other Iterations)	<input checked="" type="checkbox"/> Neighbor Joining
Min Diag Length (Lambda)	<input checked="" type="checkbox"/> 24
<input type="button" value="Help"/> <input type="button" value="Reset"/> <input type="button" value="Cancel"/> <input type="button" value="OK"/>	

Figure 5: Parameters of Muscle Alignment for MSA

DNA Sequences		Translated Protein Sequences
Species/Abbrv		
1. AP008231.1:293962-295693 Synechococcus elongatus PCC 6301 DNA complete genome	CTAGAGCTGACTAAATCCGAACCGT.....CTGATTGC.....
2. CP130602.1:1325108-1326842 Synechococcus elongatus PCC 7942 = FACHB-805 strain UTEX 2434 chromosome complete genome		GC GATG C AAAATGCC TTC AGTACAGCAGACTCCCTCAGATCAGTCACTACTGGATGAGATA
3. CP006471.1:1937628-1939359 Synechococcus sp. UTEX 2973 complete genome		ACTAGAGCTGACTAAATCCGAACCGT.....CTGATTGC.....
4. CP085785.1:1324588-1326322 Synechococcus elongatus PCC 6301 chromosome complete sequence		GC GATG C AAAATGCC TTC AGTACAGCAGACTCCCTCAGATCAGTCACTACTGGATGAGATA
5. CP088958.1:1325159-1326893 Synechococcus elongatus PCC 6311 chromosome		GC GATG C AAAATGCC TTC AGTACAGCAGACTCCCTCAGATCAGTCACTACTGGATGAGATA
6. CP088961.1:1325008-1326742 Synechococcus elongatus PCC 7943 chromosome		GC GATG C AAAATGCC TTC AGTACAGCAGACTCCCTCAGATCAGTCACTACTGGATGAGATA
7. CP000100.1:1325104-1326838 Synechococcus elongatus PCC 7942 = FACHB-805 strain PCC 7942 chromosome complete genome		GC GATG C AAAATGCC TTC AGTACAGCAGACTCCCTCAGATCAGTCACTACTGGATGAGATA
8. CP033061.1:2555070-2556801 Synechococcus elongatus UTEX 3055 chromosome complete genome	CTAGAGCTGACTAAATCCGAACCGT.....CTGATTGC.....

Figure 6: First Few Bases of Aligned sequences using MUSCLE

Phylogenetic Tree Analysis (MEGA)

M11: Analysis Preferences

Phylogeny Reconstruction

Option	Setting
ANALYSIS	
Scope	→ All Selected Taxa
Statistical Method	→ Neighbor-joining
PHYLOGENY TEST	
Test of Phylogeny	→ Bootstrap method
No. of Bootstrap Replications	→ 500
SUBSTITUTION MODEL	
Substitutions Type	→ Nucleotide
Model/Method	→ Maximum Composite Likelihood
Substitutions to Include	→ d: Transitions + Transversions
RATES AND PATTERNS	
Rates among Sites	→ Uniform Rates
Gamma Parameter	→ Not Applicable
Pattern among Lineages	→ Same (Homogeneous)
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ Complete deletion
Site Coverage Cutoff (%)	→ Not Applicable
SYSTEM RESOURCE USAGE	
Number of Threads	→ 3

? Help

✕ Cancel

✓ OK

Figure 7: Parameters for Phylogenetic Analysis in MEGA

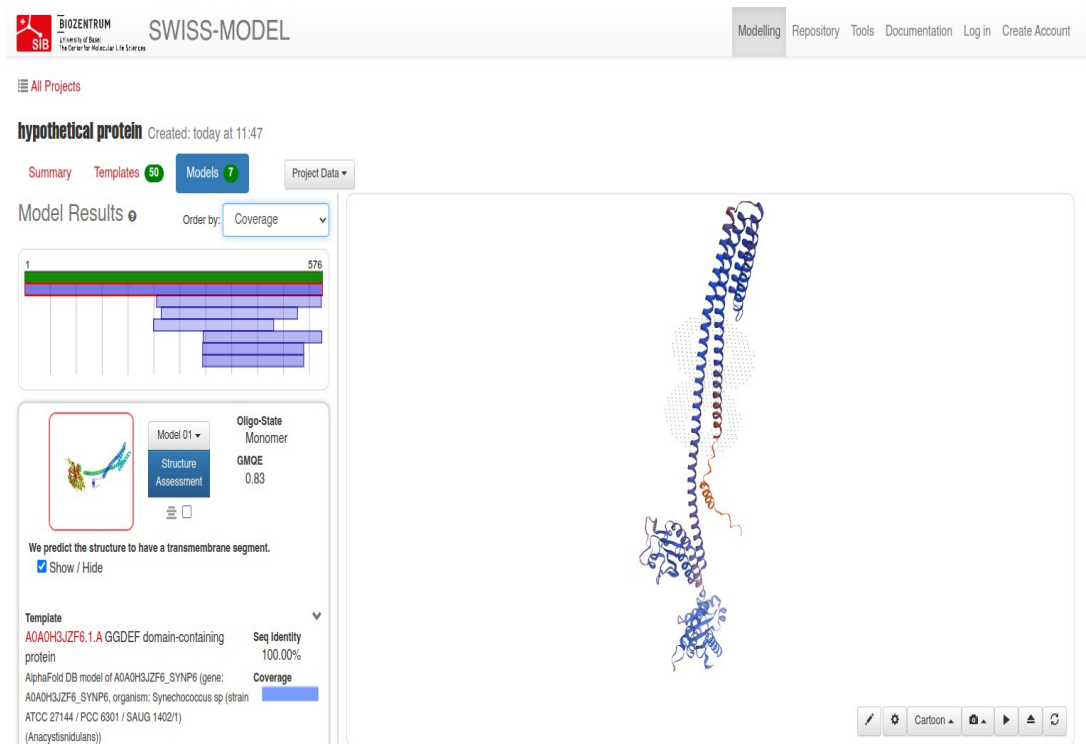


Figure 10: 3D Structure for Hypothetical Protein

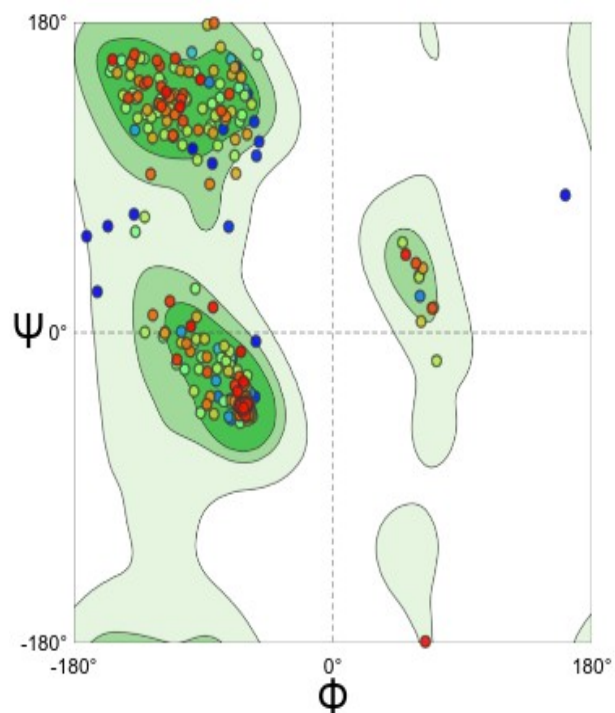


Figure 11: Ramachandran Plots for Assessment of Protein Structure

MolProbity Results	
MolProbity Score	1.08
Clash Score	0.76
Ramachandran Favoured	96.52%
<input type="checkbox"/> Ramachandran Outliers	0.52%
A10 SER, A7 THR, A3 ASN	
<input type="checkbox"/> Rotamer Outliers	1.39%
A389 LEU, A267 THR, A295 SER, A9 ASP, A336 THR, A30 LEU, A7 THR	
C-Beta Deviations	0
<input type="checkbox"/> Bad Bonds	1 / 4668
A329 HIS	
<input type="checkbox"/> Bad Angles	28 / 6307
A489 ASP, A375 ASP, (A391 LEU-A392 PRO), A7 THR, A173 HIS, A329 HIS, (A31 LYS-A32 VAL), A246 HIS, (A23 LEU-A24 PRO), A182 ASN, A435 ASN, (A350 ILE-A351 PRO), A344 HIS, A370 HIS, A235 HIS, A343 HIS, A346 HIS, A133 GLN, A397 HIS, (A443 ALA-A444 PRO), (A541 PHE-A542 PRO), A3 ASN, A529 ASN, A362 HIS, A466 HIS	
<input type="checkbox"/> Cis Prolines	2 / 19
(A127 GLN-A128 PRO), (A541 PHE-A542 PRO)	
Results obtained using MolProbity version 4.4	

Figure 12: MolProbity Results as Assessment of Protein Structure

Ramachandran Favoured score is 96.52%.

6 Domains and 1 Alpha Fold.

Domains are:

- ✓ GGDEF_dom - IPR000160
- ✓ GGDEF - TIGR00254
- ✓ GGDEF - SM00267
- ✓ GGDEF - PS50887
- ✓ GGDEF - PF00990
- ✓ GGDEF - cd01949

Additional Info:

- ✓ https://www.uniprot.org/uniprotkb/A0A0H3JZF6/entry#subcellular_location
- ✓ <https://www.ebi.ac.uk/interpro/protein/UniProt/A0A0H3JZF6/>

Open Reading Frame Analysis:

NCBI orffinder url: <https://www.ncbi.nlm.nih.gov/orffinder/>

Start Codon: ATG

Stop Codons: TAA, TAG, TGA

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x84.

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt



Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

AP008231.1

From: 293962

To: 295693

Choose Search Parameters

Minimal ORF length (nt): 75

Genetic code: 1, Standard

ORF start codon to use:
☒ "ATG" only
☐ "ATG" and alternative initiation codons
☐ Any sense codon

☒ Ignore nested ORFs

Start Search / Clear

Submit

Clear

Figure 13: NCBI OrfFinder Search Page

Synechococcus elongatus PCC 6301 DNA, complete genome

ORFs found: 5 Genetic code: 1 Start codon: 'ATG' only
ORFs were calculated on the interval from 293962 to 295693 nt



ORF3

SmartBLAST

BLAST

Marked set (0)

SmartBLAST best hit titles...

BLAST

BLAST Database:

UniProtKB/Swiss-Prot (swissprot) ▼

[Mark subset...](#)

Marked: 0

Download marked set

as Protein FASTA ▾

Six-frame translation...

Figure 14: NCBI OrfFinder Result Page