Assembly and annotation of a novel *T. siberians* strain with a focus on glycosyltransferases

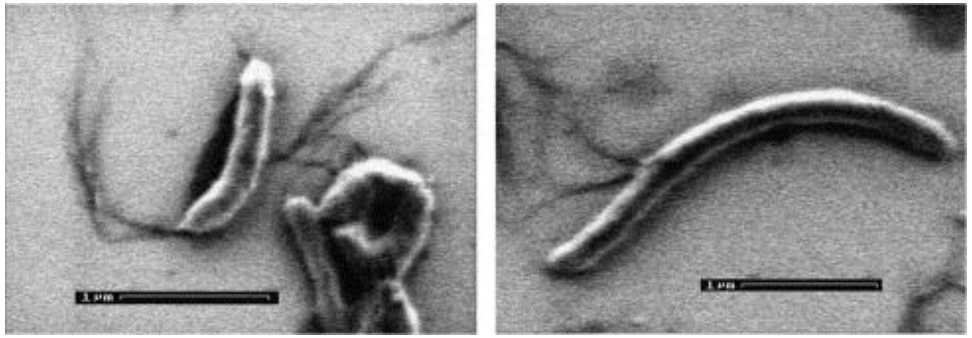
Prepared by Rajesh Chikatla, Madeline Fossitt, Juhee Lim, Elijah Walker

Contents

- Background on the Bacteria
- Genome Assembly- Trimmomatic & QUAST
- Annotation and the Genome Assembly –PROKKA
- Identifying Protein of Interest - Blastp
- Comparison structure – NCBI CDD
- Active Site & binding Site - PyMol
- Superimposable images- PyMol

Background on *Telmatosprillum siberians*

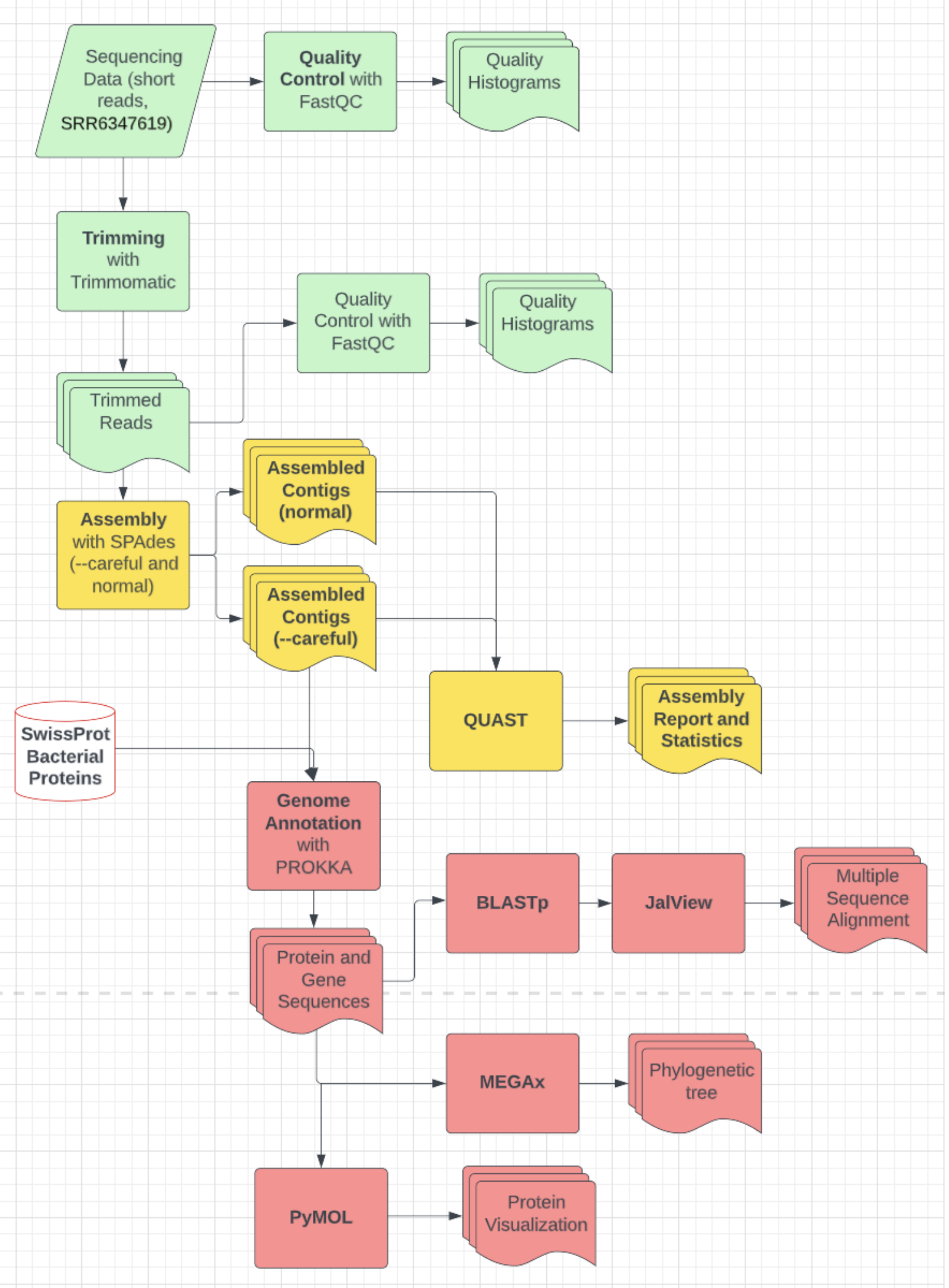
- Facultative, gram-negative anaerobic bacteria
    - Mesophilic; grows well at around 28°C
- Isolated from a Siberian marshy peatland
- Novel strains with designations of 26-4b1, 26-2, and K-1
    - We will be focusing on 26-4b1



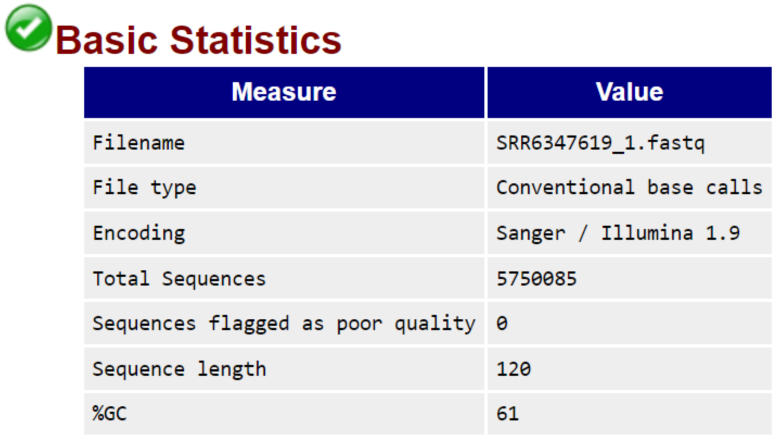
Project Workflow

- Green: Quality control of reads
- Yellow: Assembly of draft *T. siberians* genome using SPAdes
- Red: Annotation and analysis using PROKKA, Blastp, MegaX, PyMOL
- Key:
  - Diamond: start of workflow
  - Rectangle: Bioinformatic program
  - Stacked: resulting files from program
  - Cylinder: web database

Rajesh

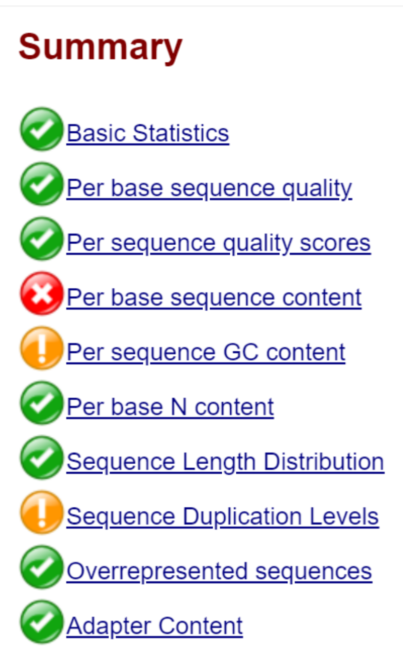

Sequencing Data (short reads, SRR6347619)
Quality Control with FastQC
Quality Histograms
Trimming with Trimmomatic
Quality Control with FastQC
Quality Histograms
Trimmed Reads
Assembled Contigs (normal)
Assembly with SPAdes (--careful and normal)
Assembled Contigs (--careful)
SwissProt Bacterial Proteins
QUAST
Assembly Report and Statistics
Genome Annotation with PROKKA
BLASTp
JalView
Multiple Sequence Alignment
Protein and Gene Sequences
MEGAx
Phylogenetic tree
PyMOL
Protein Visualization

Genome Assembly – Quality Assessment



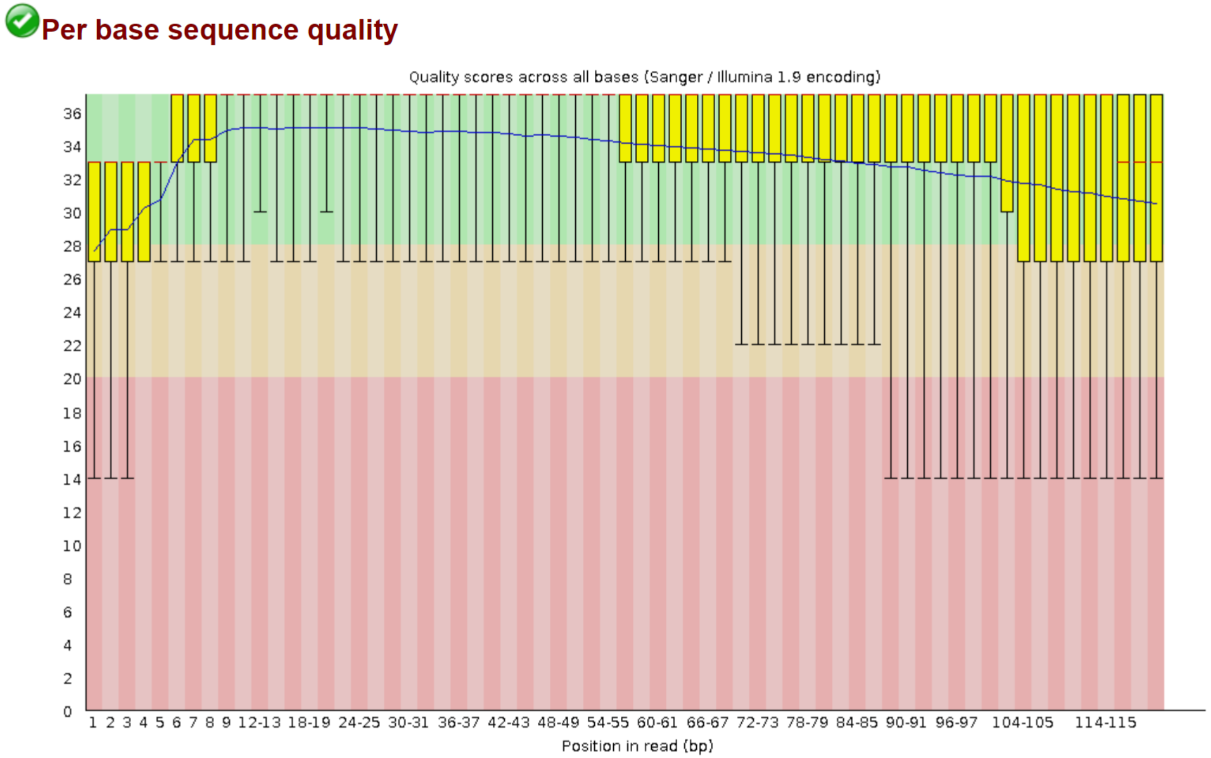
✓ Basic Statistics

Measure	Value
Filename	SRR6347619_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5750085
Sequences flagged as poor quality	0
Sequence length	120
%GC	61



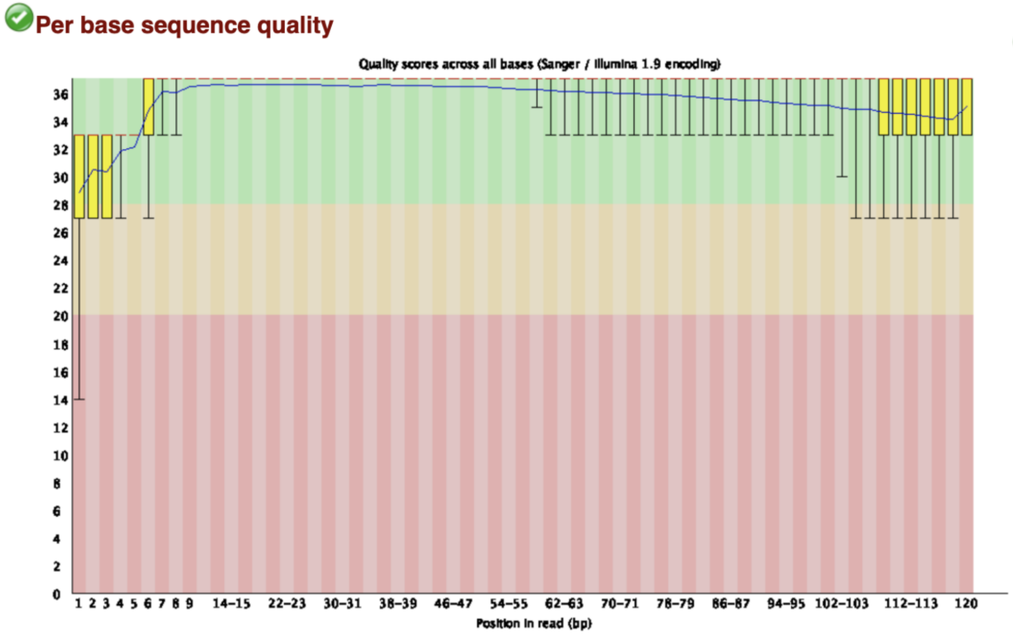
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

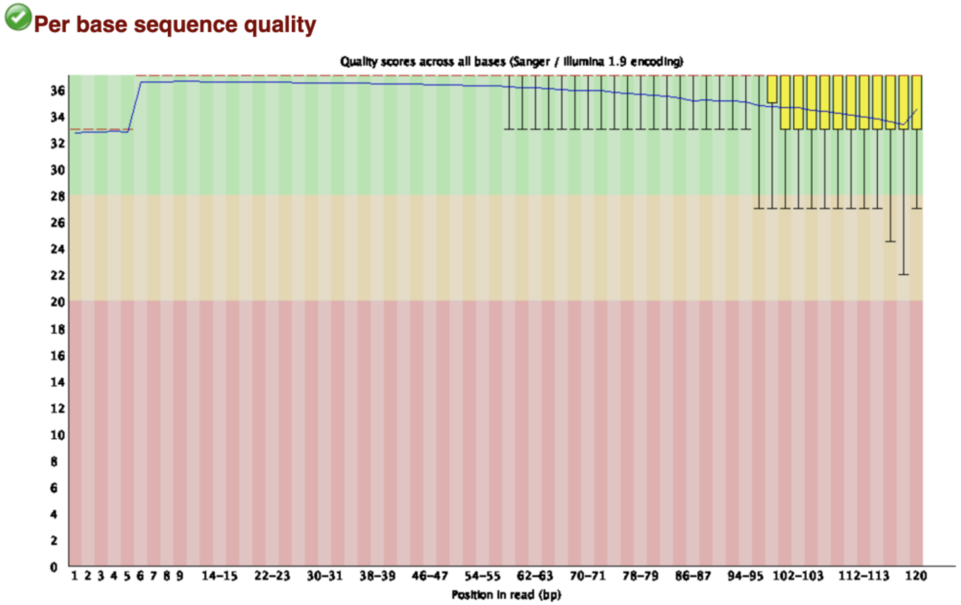


✓ Per base sequence quality

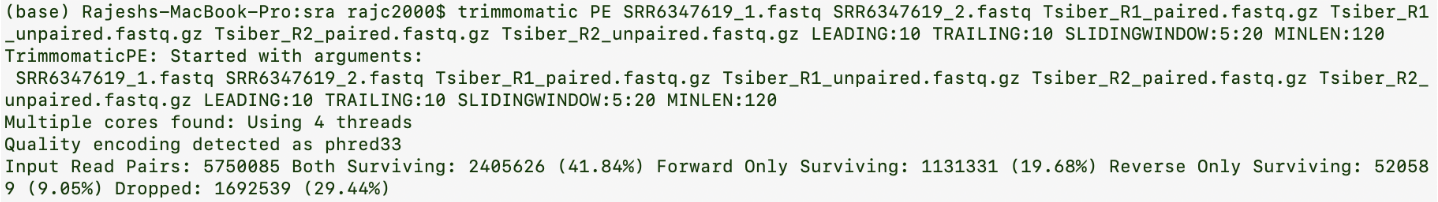
Genome Assembly – After Trimmomatic



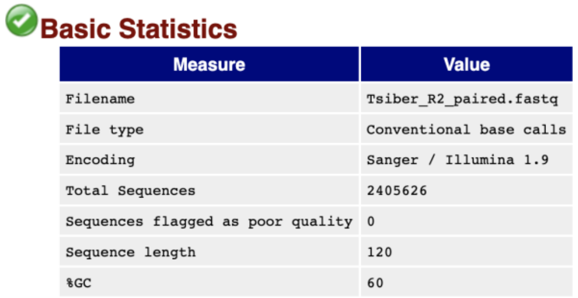
Tsiber_R1_paired_fastqc
(FWD READ)



Tsiber_R2_paired_fastqc
(REV READ)



```
(base) Rajeshs-MacBook-Pro:sra rajc2000$ trimmomatic PE SRR6347619_1.fastq SRR6347619_2.fastq Tsiber_R1_paired.fastq.gz Tsiber_R1
_unpaired.fastq.gz Tsiber_R2_paired.fastq.gz Tsiber_R2_unpaired.fastq.gz LEADING:10 TRAILING:10 SLIDINGWINDOW:5:20 MINLEN:120
TrimmomaticPE: Started with arguments:
 SRR6347619_1.fastq SRR6347619_2.fastq Tsiber_R1_paired.fastq.gz Tsiber_R1_unpaired.fastq.gz Tsiber_R2_paired.fastq.gz Tsiber_R2_
unpaired.fastq.gz LEADING:10 TRAILING:10 SLIDINGWINDOW:5:20 MINLEN:120
Multiple cores found: Using 4 threads
Quality encoding detected as phred33
Input Read Pairs: 5750085 Both Surviving: 2405626 (41.84%) Forward Only Surviving: 1131331 (19.68%) Reverse Only Surviving: 52058
9 (9.05%) Dropped: 1692539 (29.44%)
```

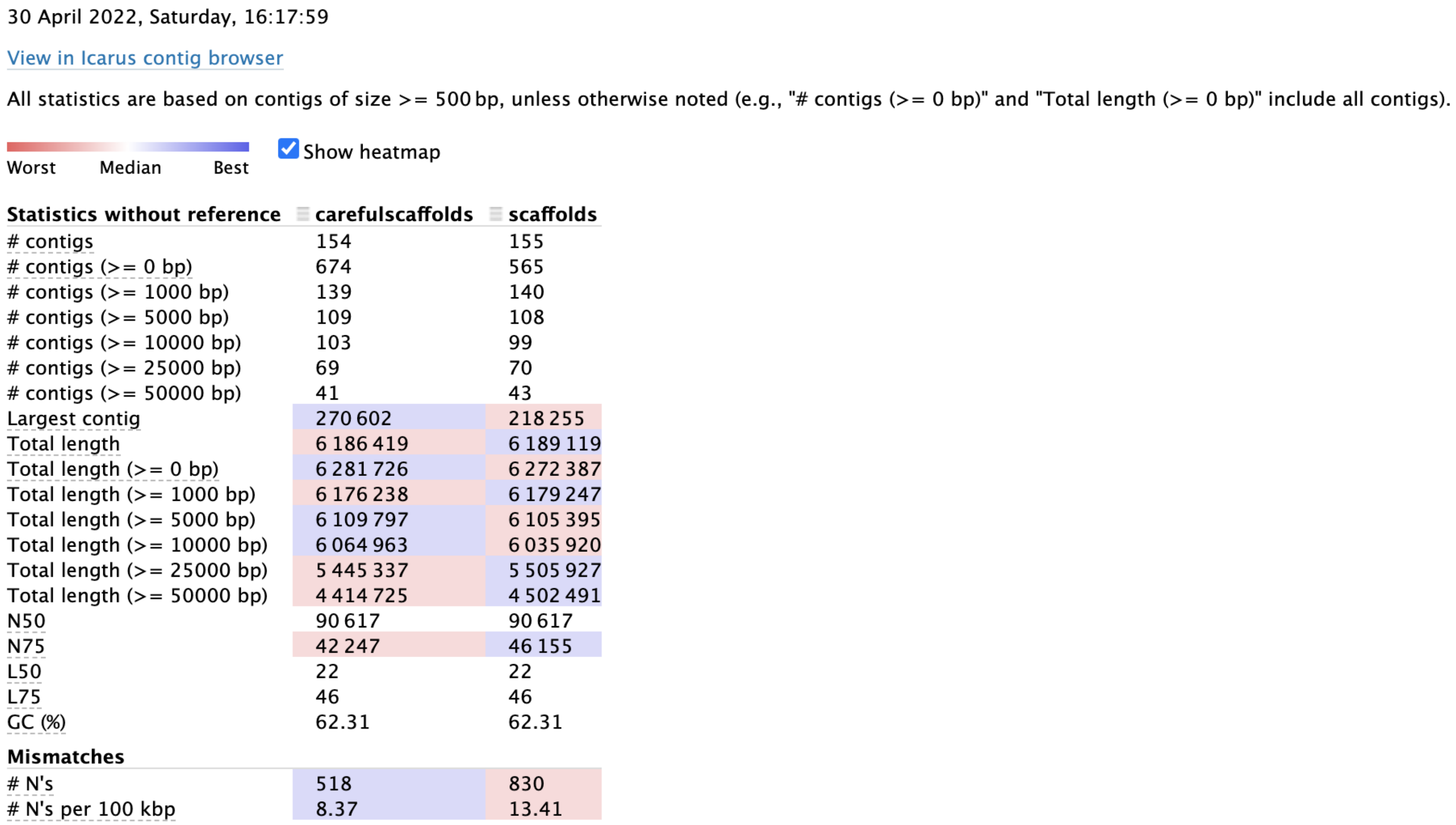


✓ Basic Statistics

Measure	Value
Filename	Tsiber_R2_paired.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2405626
Sequences flagged as poor quality	0
Sequence length	120
%GC	60

Genome Assembly: Quality Assessment

- **QUAST** – Quality Assessment Tool for Genome Assemblies
- Uses various metrics to assess the quality of an assembled genome.
- QUAST was used to compare two SPAdes assemblies run with –careful parameter and default parameters, respectively.
- Annotation was performed with carefulscaffolds.fasta due to less mismatches.



30 April 2022, Saturday, 16:17:59

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

☑ Show heatmap

Statistics without reference	carefulscaffolds	scaffolds
# contigs	154	155
# contigs (>= 0 bp)	674	565
# contigs (>= 1000 bp)	139	140
# contigs (>= 5000 bp)	109	108
# contigs (>= 10000 bp)	103	99
# contigs (>= 25000 bp)	69	70
# contigs (>= 50000 bp)	41	43
Largest contig	270 602	218 255
Total length	6 186 419	6 189 119
Total length (>= 0 bp)	6 281 726	6 272 387
Total length (>= 1000 bp)	6 176 238	6 179 247
Total length (>= 5000 bp)	6 109 797	6 105 395
Total length (>= 10000 bp)	6 064 963	6 035 920
Total length (>= 25000 bp)	5 445 337	5 505 927
Total length (>= 50000 bp)	4 414 725	4 502 491
N50	90 617	90 617
N75	42 247	46 155
L50	22	22
L75	46	46
GC (%)	62.31	62.31
Mismatches		
# N's	518	830
# N's per 100 kbp	8.37	13.41

Genome Summary – PROKKA/QUAST Results

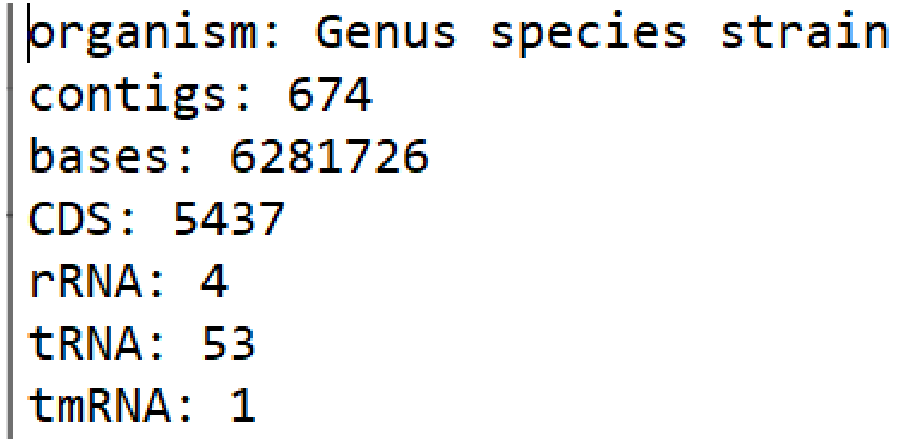
Contigs: 674

Genome size: 6,281,726

# of Coding Sequences: 5,437

Number of RNAs: 58

GC Content: 62.31%



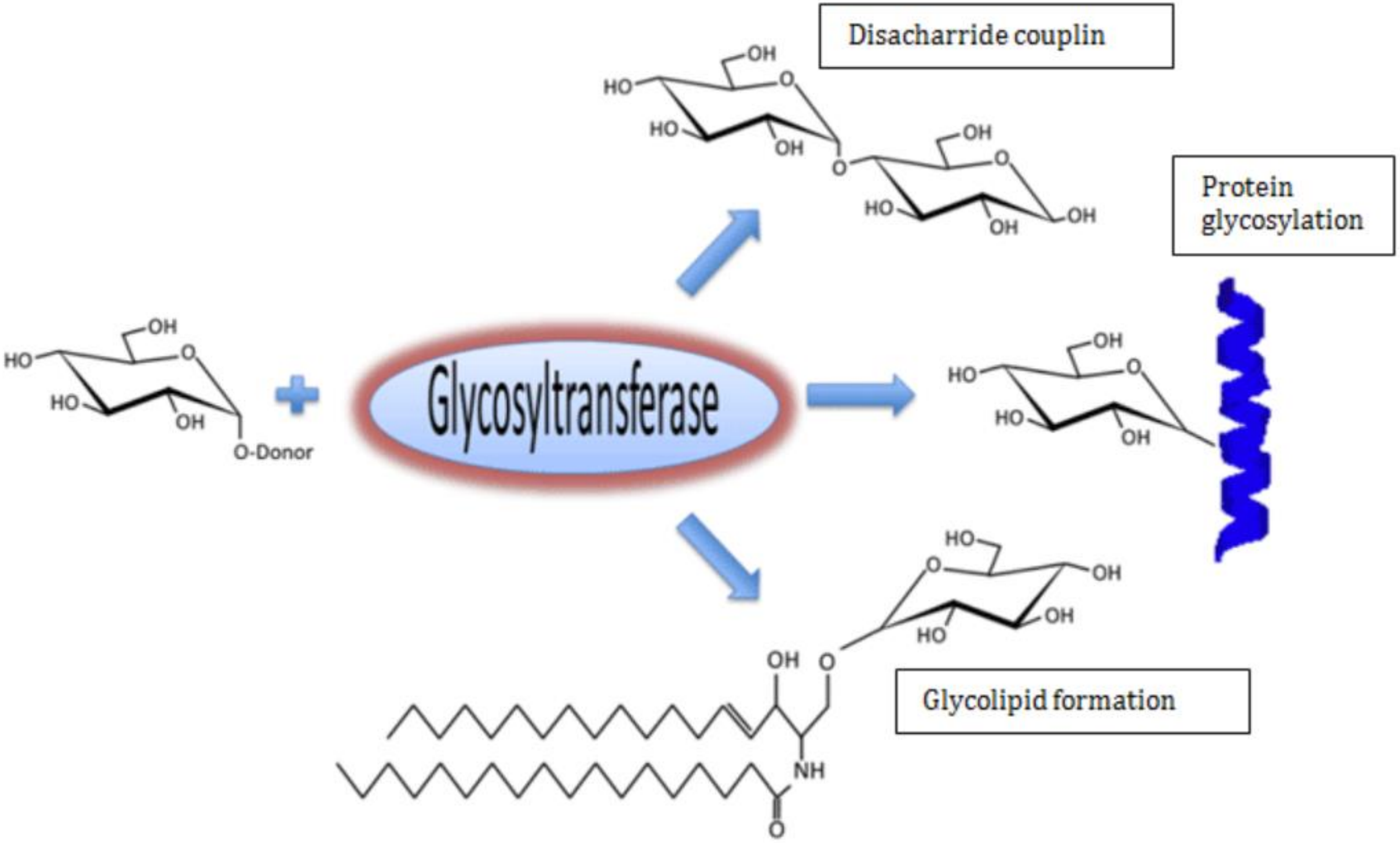
```
organism: Genus species strain
contigs: 674
bases: 6281726
CDS: 5437
rRNA: 4
tRNA: 53
tmRNA: 1
```

QUAST

L75	46
GC (%)	62.31
Mismatches	

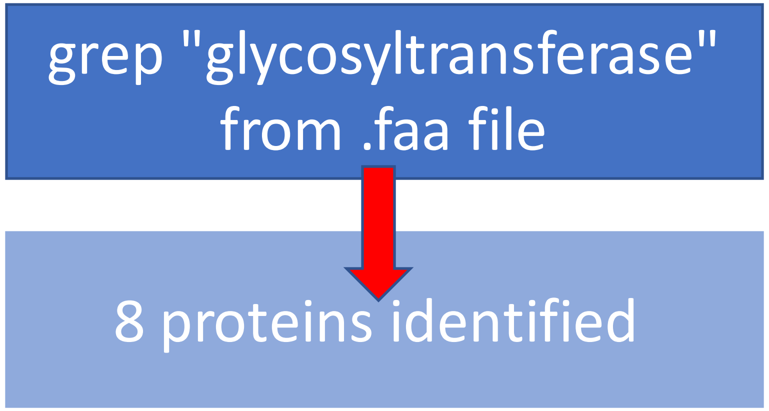
Glycosyltransferases

- Enzyme subclass responsible for the initiation and elongation of glycan chains
- Aids in post-translational modification of proteins



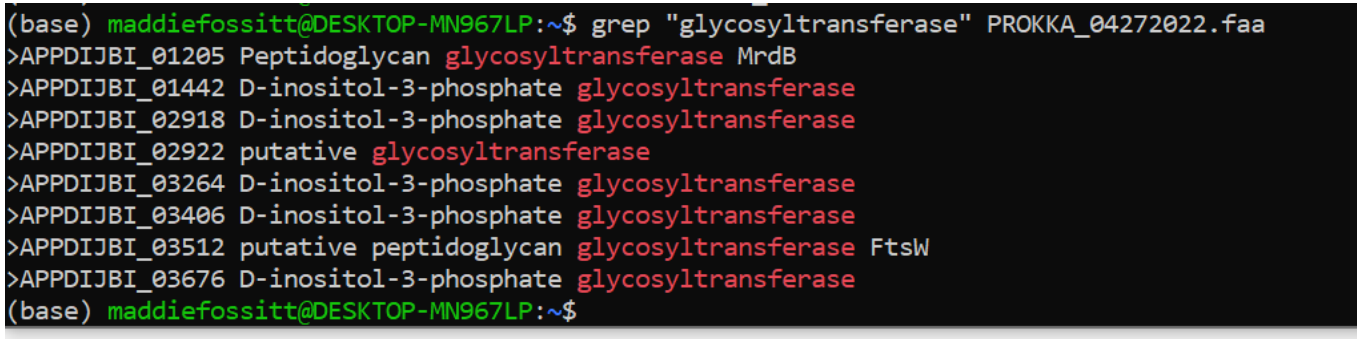
https://www.sbhsciences.com/Glycosyltransferase.asp

Filtering for Glycosyltransferases



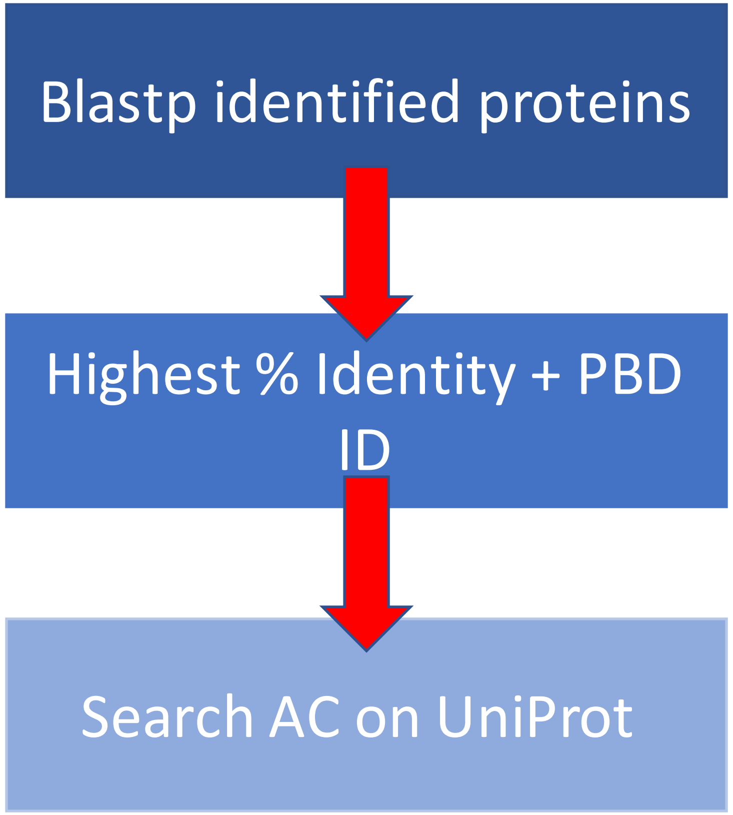
grep "glycosyltransferase" from .faa file

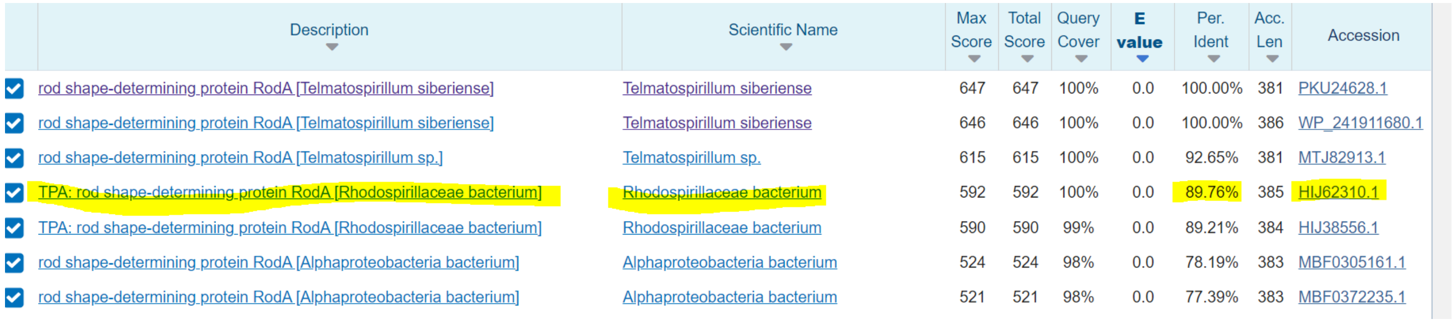
8 proteins identified



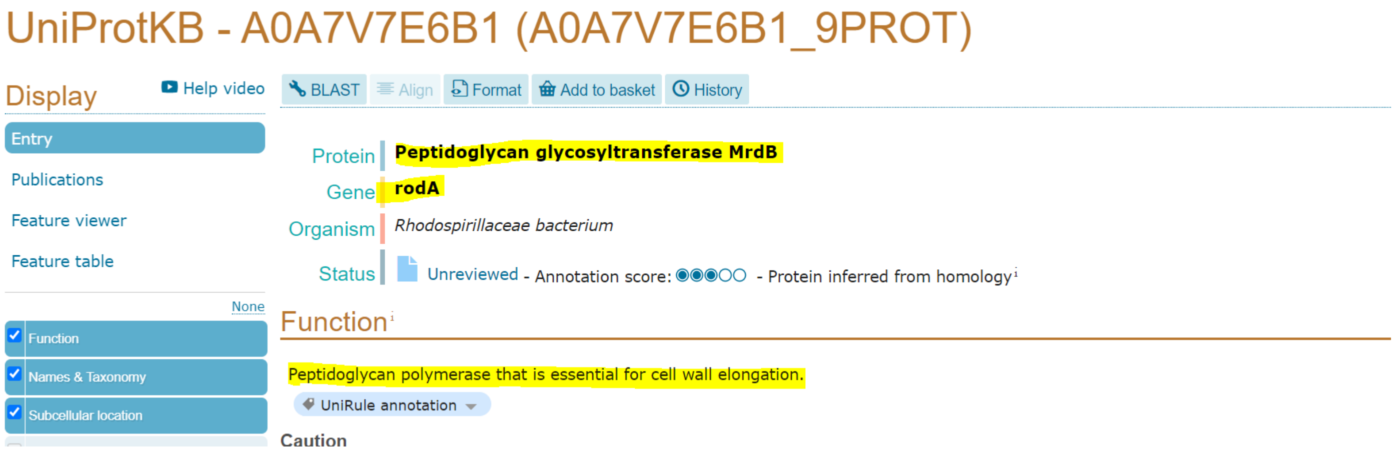
```
(base) maddiefossitt@DESKTOP-MN967LP:~$ grep "glycosyltransferase" PROKKA_04272022.faa
>APPDIJBI_01205 Peptidoglycan glycosyltransferase MrdB
>APPDIJBI_01442 D-inositol-3-phosphate glycosyltransferase
>APPDIJBI_02918 D-inositol-3-phosphate glycosyltransferase
>APPDIJBI_02922 putative glycosyltransferase
>APPDIJBI_03264 D-inositol-3-phosphate glycosyltransferase
>APPDIJBI_03406 D-inositol-3-phosphate glycosyltransferase
>APPDIJBI_03512 putative peptidoglycan glycosyltransferase FtsW
>APPDIJBI_03676 D-inositol-3-phosphate glycosyltransferase
(base) maddiefossitt@DESKTOP-MN967LP:~$
```

Identifying Protein of Interest





	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	rod shape-determining protein RodA [Telmatospirillum siberiense]	Telmatospirillum siberiense	647	647	100%	0.0	100.00%	381	PKU24628.1
☑	rod shape-determining protein RodA [Telmatospirillum siberiense]	Telmatospirillum siberiense	646	646	100%	0.0	100.00%	386	WP_241911680.1
☑	rod shape-determining protein RodA [Telmatospirillum sp.]	Telmatospirillum sp.	615	615	100%	0.0	92.65%	381	MTJ82913.1
☑	TPA: rod shape-determining protein RodA [Rhodospirillaceae bacterium]	Rhodospirillaceae bacterium	592	592	100%	0.0	89.76%	385	HIJ62310.1
☑	TPA: rod shape-determining protein RodA [Rhodospirillaceae bacterium]	Rhodospirillaceae bacterium	590	590	99%	0.0	89.21%	384	HIJ38556.1
☑	rod shape-determining protein RodA [Alphaproteobacteria bacterium]	Alphaproteobacteria bacterium	524	524	98%	0.0	78.19%	383	MBF0305161.1
☑	rod shape-determining protein RodA [Alphaproteobacteria bacterium]	Alphaproteobacteria bacterium	521	521	98%	0.0	77.39%	383	MBF0372235.1



UniProtKB - A0A7V7E6B1 (A0A7V7E6B1_9PROT)

Protein **Peptidoglycan glycosyltransferase MrdB**

Gene **rodA**

Organism *Rhodospirillaceae bacterium*

Status Unreviewed - Annotation score: ●●●○○ - Protein inferred from homology[i]

Function[i]

Peptidoglycan polymerase that is essential for cell wall elongation.

UniRule annotation

Caution

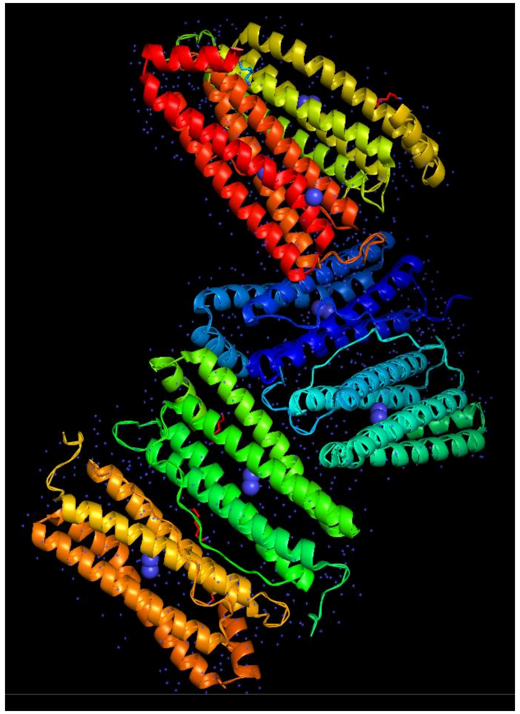
Protein of Interest

Peptidoglycan glycosyltransferase MrdB

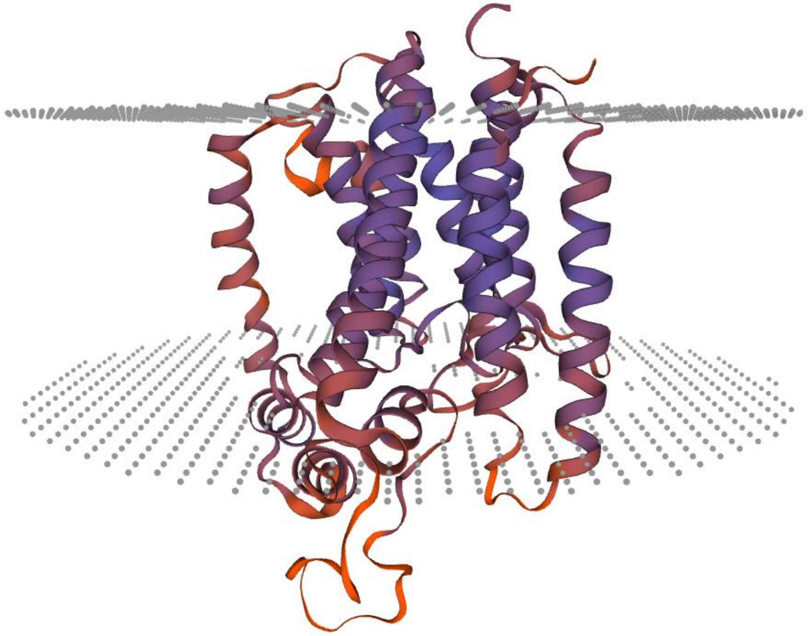
PDB ID: 6LP5

Gene: rodA

Function: Peptidoglycan polymerase that is essential for cell wall elongation

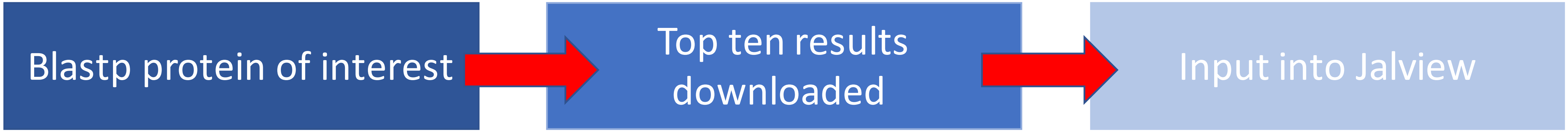


PyMOL 🥰



Swiss Model

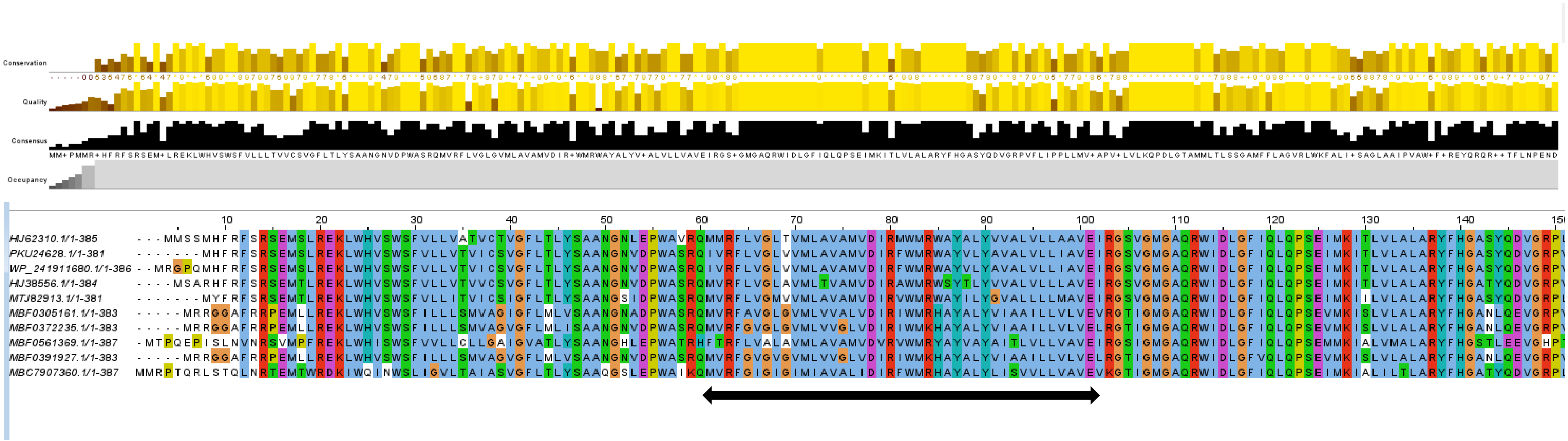
MSA



Blastp protein of interest

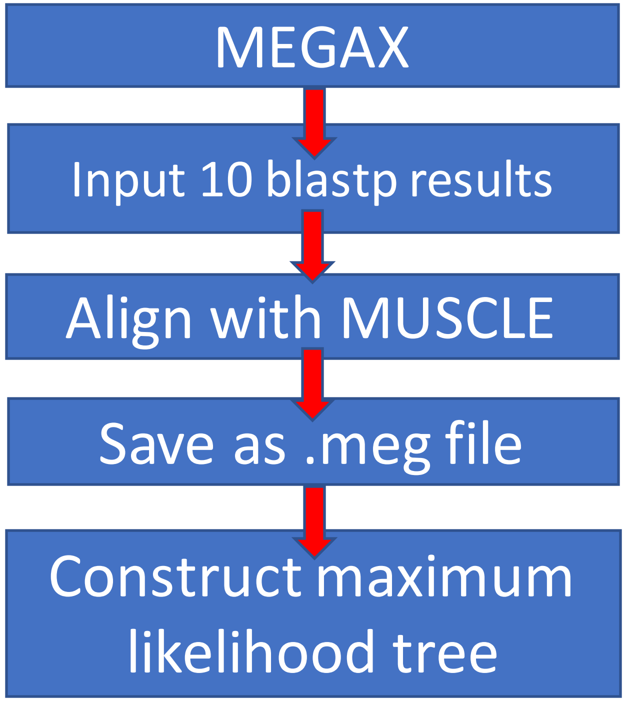
Top ten results downloaded

Input into Jalview


Conservation
Quality
Consensus
MM+PMMR+HFRFSRSEM+LREKLWHVSWSFVLLLTVVCSVGFLTLYSAANGNVDPWASRQMVRFLVGLGVMLAVAMVDIR+WMRWAYALYV+ALVLLAVEIRGS+GMGAQRWIDLGFIQLQPSEIMKITLVLALARYFHGASYQDVGRPVFLIPPLLMV+APV+LVLKQPDLGTAMMLTLSSGAMFFLAGVRLWKFALI+SAGLAAIPVAW+F+REYQRQR++TFLNPEND
Occupancy
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150

```
HJ62310.1/1-385       ---MMSSMHFRFSRSEMSLREKLWHVSWSFVLLVATVCTVGFLTLYSAANGNLEPWAVRQMMRFLVGLTVMLAVAMVDIRMWMRWAYALYVVALVLLAAVEIRGSVGMGAQRWIDLGFIQLQPSEIMKITLVLALARYFHGASYQDVGRPL
PKU24628.1/1-381      -------MHFRFSRSEMSLREKLWHVSWSFVLLVTVICSVGFLTLYSAANGNVDPWASRQIVRFLVGLVVMLAVAMVDIRFWMRWAYVLYAVALVLLIAVEIRGSVGMGAQRWIDLGFIQLQPSEIMKITLVLALARYFHGASYQDVGRPV
WP_241911680.1/1-386  --MRGPQMHFRFSRSEMSLREKLWHVSWSFVLLVTVICSVGFLTLYSAANGNVDPWASRQIVRFLVGLVVMLAVAMVDIRFWMRWAYVLYAVALVLLIAVEIRGSVGMGAQRWIDLGFIQLQPSEIMKITLVLALARYFHGASYQDVGRPV
HJ38556.1/1-384       ----MSARHFRFSRSEMTLREKLWHVSWSFVLLVTVVCSVGFLTLYSAANGNVDPWASRQMVRFLVGLAVMLTVAMVDIRAWMRWSYTLYVVALVLLLAVEIRGSVGMGAQRWIDLGFIQLQPSEIMKITLVLALARYFHGATYQDVGRPV
MTJ82913.1/1-381      -------MYFRFSRSEMTLREKLWHVSWSFVLLITVICSIGFLTLYSAANGSIDPWASRQMVRFLVGMVVMLAVAMVDIRVWMRWAYILYGVALLLLMAVEIRGSVGMGAQRWIDLGFIQLQPSEIMKIILVLALARYFHGASYQDVGRPL
MBF0305161.1/1-383    -----MRRGGAFRRPEMLLREKLWHVSWSFILLLSMVAGIGFLMLVSAANGNADPWASRQMVRFAVGLGVMLVVALVDIRIWMRHAYALYVIAAILLVLVEVRGTIGMGAQRWIDLGFIQLQPSEVMKISLVLALARYFHGANLQEVGRPV
MBF0372235.1/1-383    -----MRRGGAFRRPEMLLREKLWHVSWSFILLLSMVAGVGFLMLISAANGNVDPWASRQMVRFGVGLGVMLVVGLVDIRIWMKHAYALYVIAAILLVLVELRGTIGMGAQRWIDLGFIQLQPSEVMKISLVLALARYFHGANLQEVGRPV
MBF0561369.1/1-387    -MTPQEPISLNVNRSVMPFREKLWHISWSFVVLLCLLGAIGVATLYSAANGHLEPWATRHFTRFLVALAVMLAVAMVDVRVWMRYAYVAYAITLVLLVAVEIRGSIGMGAQRWIDLGFIQLQPSEMMKIALVMALARYFHGSTLEEVGHPT
MBF0391927.1/1-383    -----MRRGGAFRRPEMLLREKLWHVSWSFILLLSMVAGVGFLMLVSAANGNVDPWASRQMVRFGVGVGVMLVVGLVDIRIWMKHAYALYVIAAILLVLVELRGTIGMGAQRWIDLGFVQLQPSEVMKISLVLALARYFHGANLQEVGRPV
MBC7907360.1/1-387    MMRPTQRLSTQLNRTEMTWRDKIWQINWSLIGVLTAIASVGFLTLYSAAQGSLEPWAIKQMVRFGIGIGIMIAVALIDIRFWMRHAYALYLISVVLLVAVEVKGTIGMGAQRWIDLGFIQLQPSEIMKIALILTLARYFHGATYQDVGRPL
```

Phylogenetic Tree



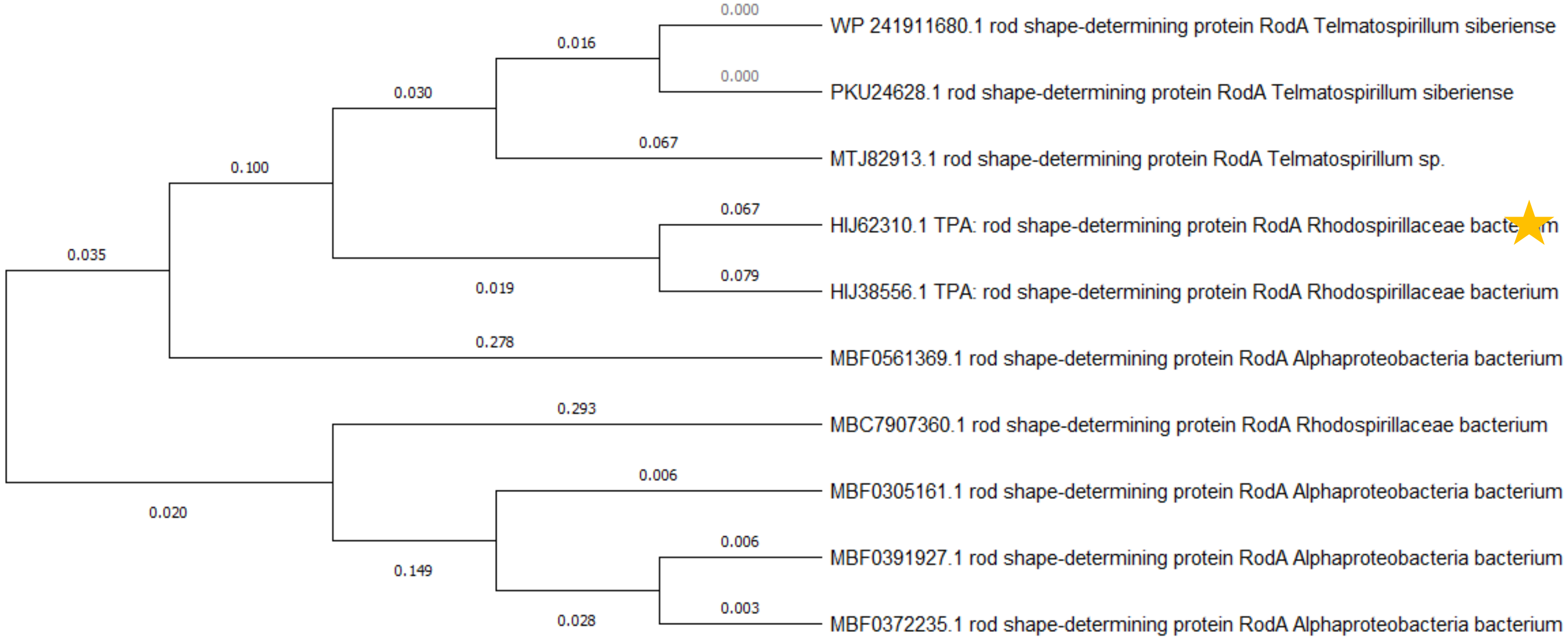
MEGAX

Input 10 blastp results

Align with MUSCLE

Save as .meg file

Construct maximum likelihood tree

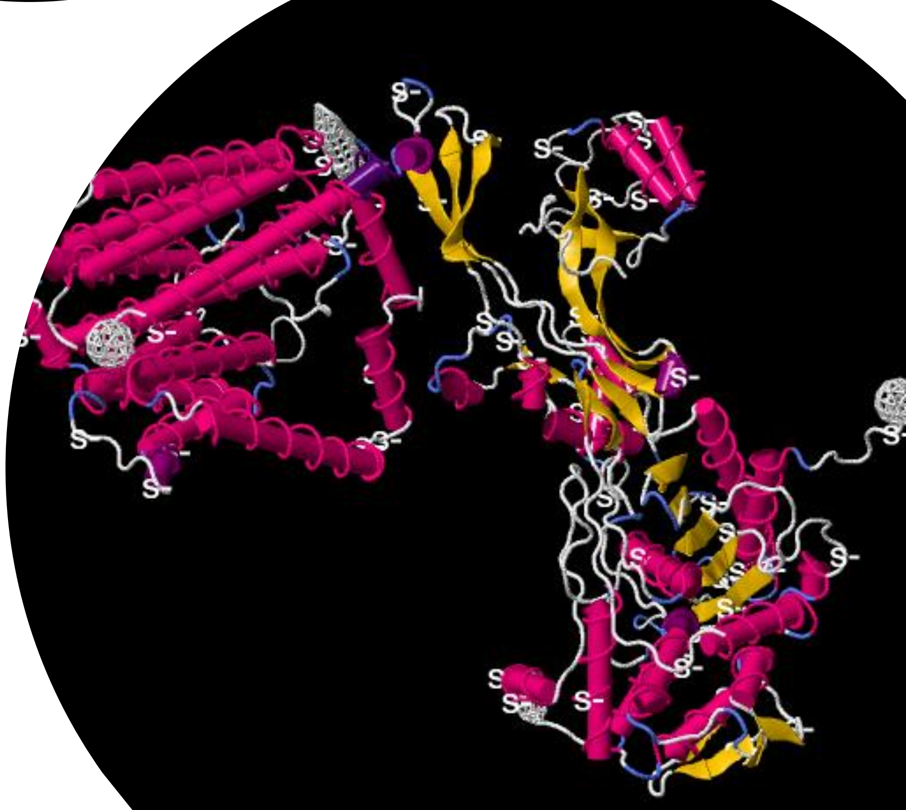


0.000
WP 241911680.1 rod shape-determining protein RodA Telmatospirillum siberiense
0.016
0.000
PKU24628.1 rod shape-determining protein RodA Telmatospirillum siberiense
0.030
0.067
MTJ82913.1 rod shape-determining protein RodA Telmatospirillum sp.
0.100
0.067
HIJ62310.1 TPA: rod shape-determining protein RodA Rhodospirillaceae bacterium
0.035
0.019
0.079
HIJ38556.1 TPA: rod shape-determining protein RodA Rhodospirillaceae bacterium
0.278
MBF0561369.1 rod shape-determining protein RodA Alphaproteobacteria bacterium
0.293
MBC7907360.1 rod shape-determining protein RodA Rhodospirillaceae bacterium
0.006
MBF0305161.1 rod shape-determining protein RodA Alphaproteobacteria bacterium
0.020
0.006
MBF0391927.1 rod shape-determining protein RodA Alphaproteobacteria bacterium
0.149
0.003
0.028
MBF0372235.1 rod shape-determining protein RodA Alphaproteobacteria bacterium

3D Structure (6pl5)

Characteristics of the protein:

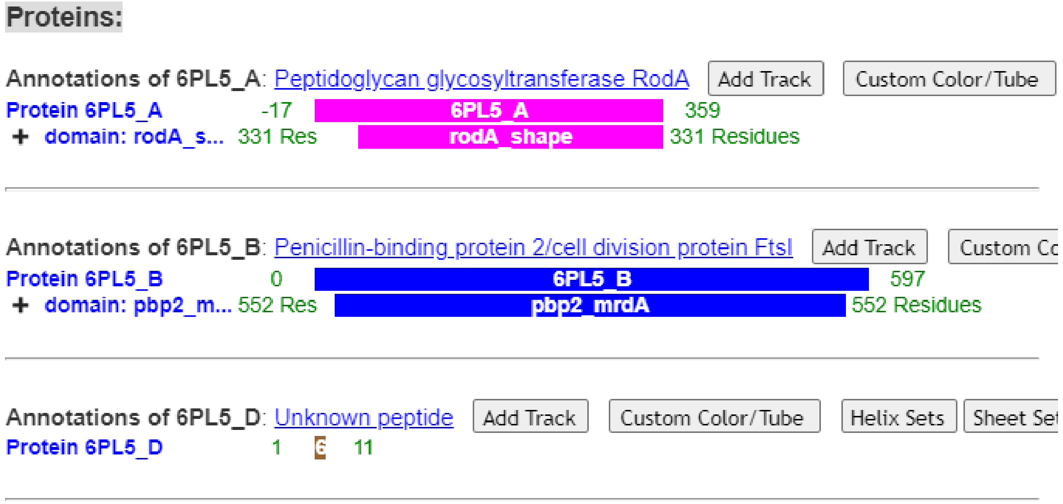
- This protein is involved in the pathway peptidoglycan biosynthesis
- Peptidoglycan polymerase that is essential for cell wall elongation.
- Located within the cellular transmembrane
- Cell wall shape-determining protein



Elijah

Domains, Motifs, or Physicochemical Characteristics

- 3 Domains : 6pl5_A (Rod A), 6pl5_B (penicillin-binding protein 2), 6pl5_D (unknown peptide)
- RodA 331 residues
- Pbp2 552 residues



Proteins:

Annotations of 6PL5_A: Peptidoglycan glycosyltransferase RodA

Protein 6PL5_A -17 6PL5_A 359

+ domain: rodA_s... 331 Res rodA_shape 331 Residues

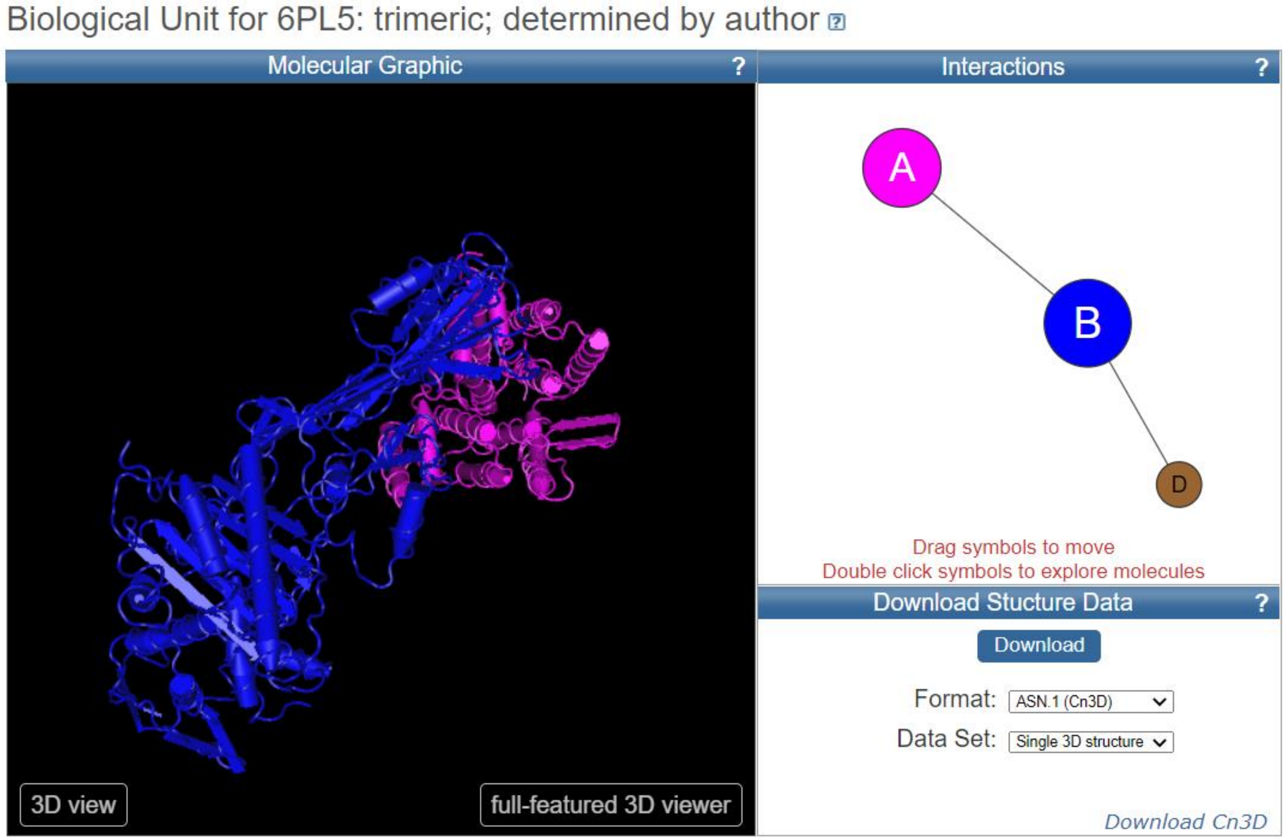
Annotations of 6PL5_B: Penicillin-binding protein 2/cell division protein FtsI

Protein 6PL5_B 0 6PL5_B 597

+ domain: pbp2_m... 552 Res pbp2_mrdA 552 Residues

Annotations of 6PL5_D: Unknown peptide

Protein 6PL5_D 1 6 11



Biological Unit for 6PL5: trimeric; determined by author

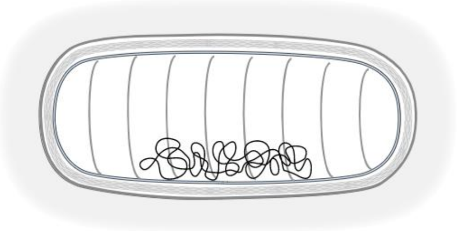
Drag symbols to move

Double click symbols to explore molecules

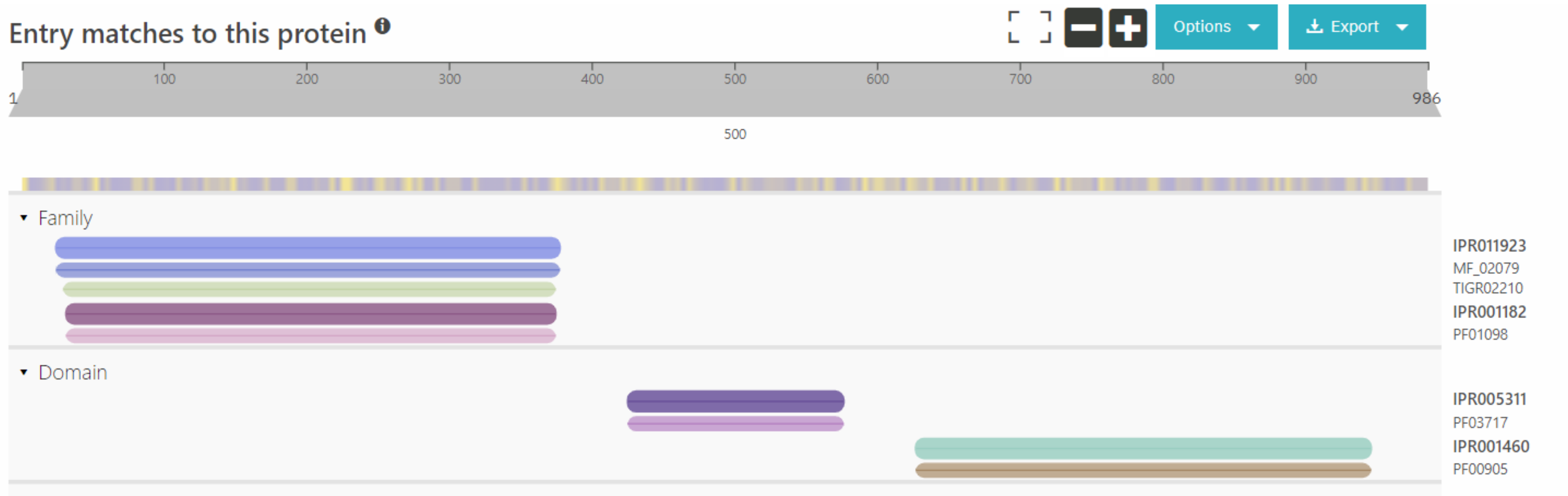
Format: ASN.1 (Cn3D)

Data Set: Single 3D structure

Domains



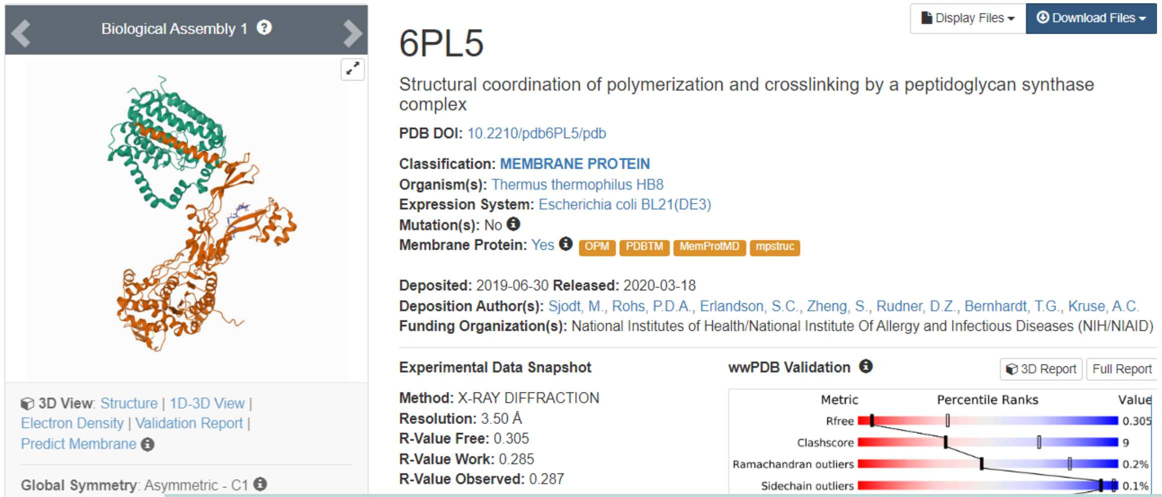
- Main domain : chain A and B
  - Rod shape & strongly implicated in PBP polymerisation.
- Using InterProScan, chain A is Family and Chain B is Domain



Entry matches to this protein

Comparison structures

Thermus thermophilus VS Campylobacter jejuni



6PL5

Structural coordination of polymerization and crosslinking by a peptidoglycan synthase complex

PDB DOI: 10.2210/pdb6PL5/pdb

Classification: MEMBRANE PROTEIN
Organism(s): Thermus thermophilus HB8
Expression System: Escherichia coli BL21(DE3)
Mutation(s): No
Membrane Protein: Yes

Deposited: 2019-06-30 **Released:** 2020-03-18
Deposition Author(s): Sjodt, M., Rohs, P.D.A., Erlandson, S.C., Zheng, S., Rudner, D.Z., Bernhardt, T.G., Kruse, A.C.
Funding Organization(s): National Institutes of Health/National Institute Of Allergy and Infectious Diseases (NIH/NIAID)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 3.50 Å
R-Value Free: 0.305
R-Value Work: 0.285
R-Value Observed: 0.287

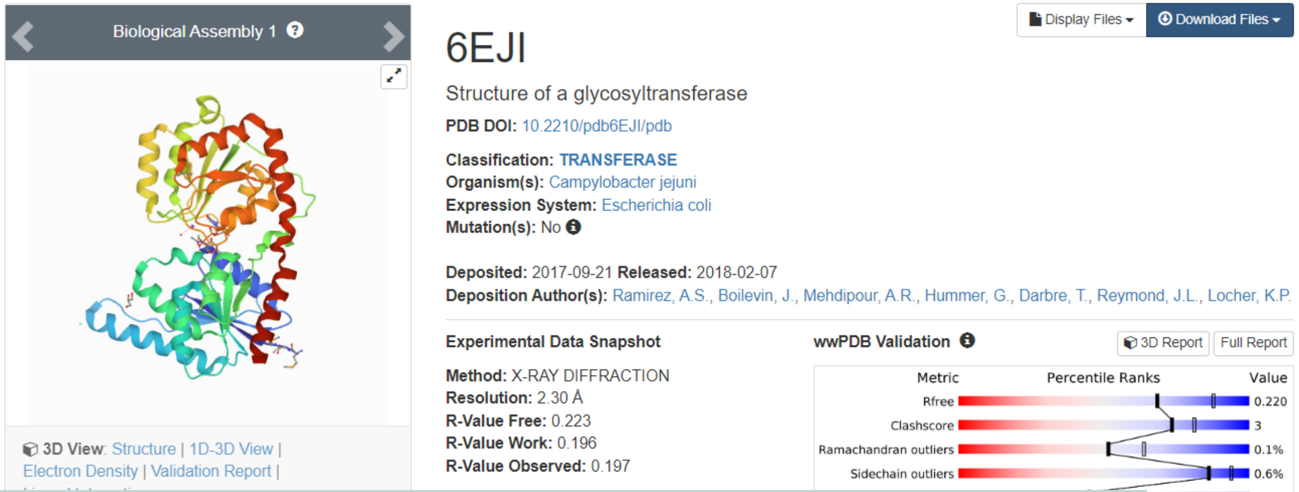
wwPDB Validation

Metric	Percentile Ranks	Value
Rfree		0.305
Clashscore		9
Ramachandran outliers		0.2%
Sidechain outliers		0.1%

3D View: Structure | 1D-3D View |
Electron Density | Validation Report |
Predict Membrane

Global Symmetry: Asymmetric - C1
Global Stoichiometry

Find Similar Assemblie



6EJI

Structure of a glycosyltransferase

PDB DOI: 10.2210/pdb6EJI/pdb

Classification: TRANSFERASE
Organism(s): Campylobacter jejuni
Expression System: Escherichia coli
Mutation(s): No

Deposited: 2017-09-21 **Released:** 2018-02-07
Deposition Author(s): Ramirez, A.S., Boilevin, J., Mehdipour, A.R., Hummer, G., Darbre, T., Reymond, J.L., Locher, K.P.

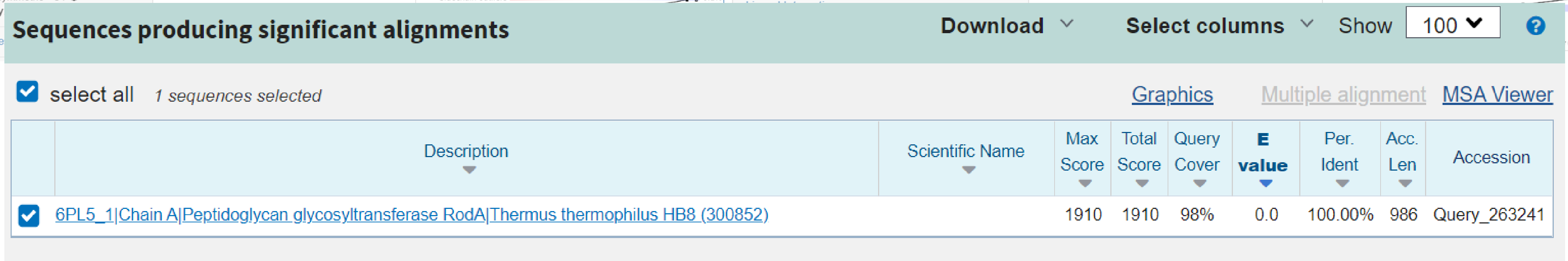
Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.30 Å
R-Value Free: 0.223
R-Value Work: 0.196
R-Value Observed: 0.197

wwPDB Validation

Metric	Percentile Ranks	Value
Rfree		0.220
Clashscore		3
Ramachandran outliers		0.1%
Sidechain outliers		0.6%
		3.4%

3D View: Structure | 1D-3D View |
Electron Density | Validation Report |



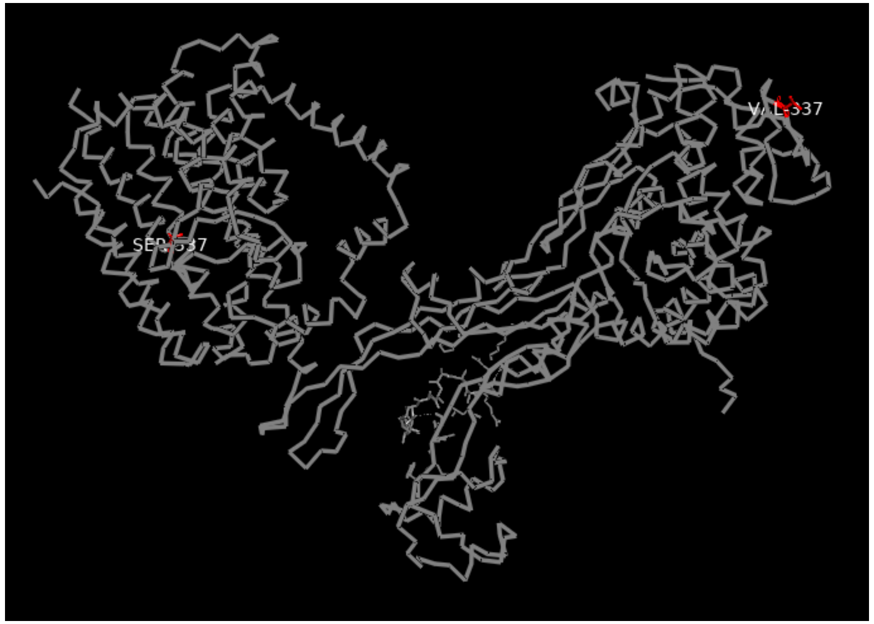
Sequences producing significant alignments

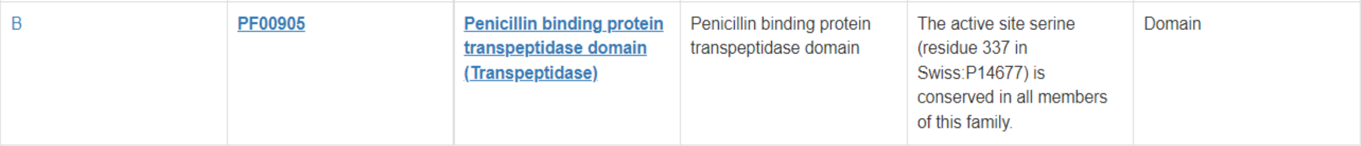
☑ select all *1 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	6PL5_1|Chain A|Peptidoglycan glycosyltransferase RodA|Thermus thermophilus HB8 (300852)		1910	1910	98%	0.0	100.00%	986	Query_263241

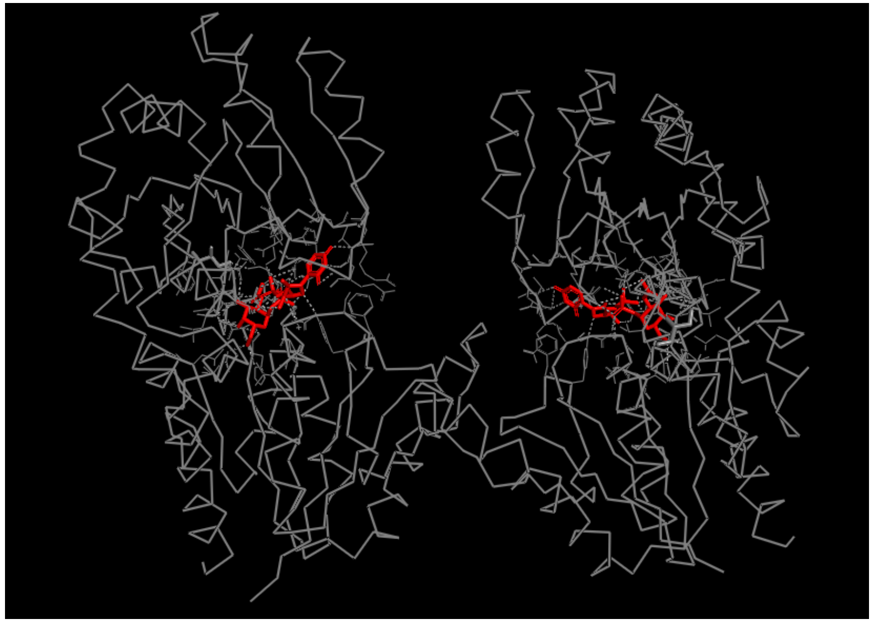
Active site

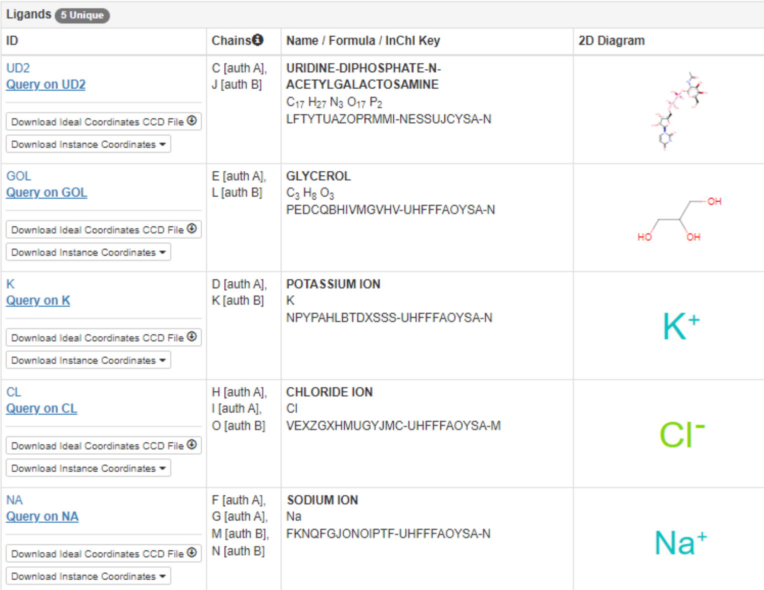
6PL5 VS 6EJI





B	PF00905	Penicillin binding protein transpeptidase domain (Transpeptidase)	Penicillin binding protein transpeptidase domain	The active site serine (residue 337 in Swiss:P14677) is conserved in all members of this family.	Domain





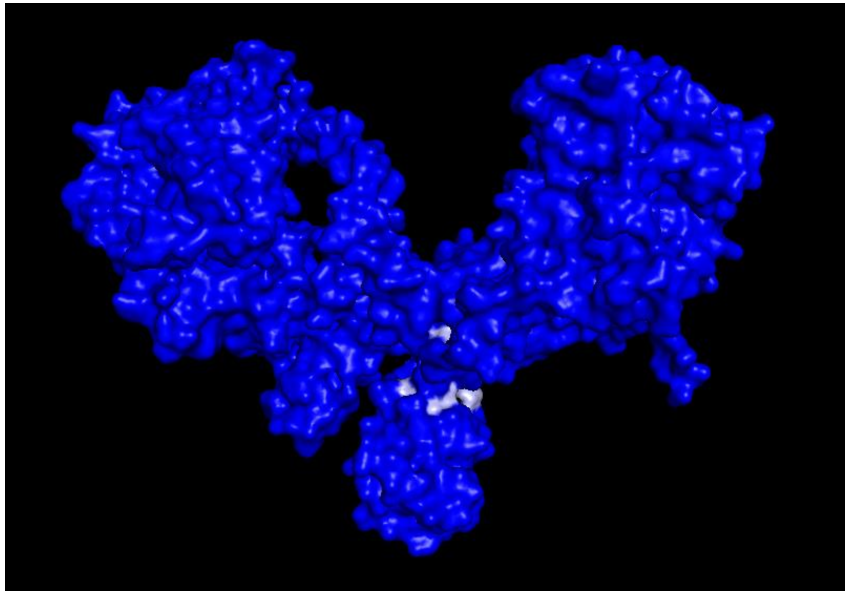
Ligands 5 Unique

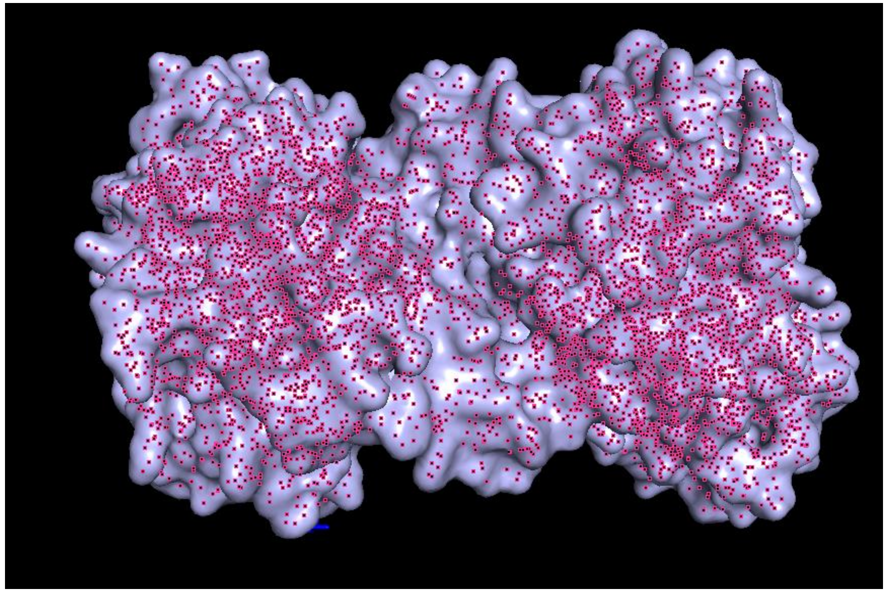
ID	Chains	Name / Formula / InChI Key	2D Diagram
UD2 Query on UD2 Download Ideal Coordinates CCD File Download Instance Coordinates	C [auth A], J [auth B]	URIDINE-DIPHOSPHATE-N-ACETYLGALACTOSAMINE $C_{17} H_{27} N_3 O_{17} P_2$ LFTYTUAZOPRMMI-NESSUJCYSA-N	
GOL Query on GOL Download Ideal Coordinates CCD File Download Instance Coordinates	E [auth A], L [auth B]	GLYCEROL $C_3 H_8 O_3$ PEDCQBHIVMGVHV-UHFFFAOYSA-N	
K Query on K Download Ideal Coordinates CCD File Download Instance Coordinates	D [auth A], K [auth B]	POTASSIUM ION K NPYPAHLBTDXSSS-UHFFFAOYSA-N	K^+
CL Query on CL Download Ideal Coordinates CCD File Download Instance Coordinates	H [auth A], I [auth A], O [auth B]	CHLORIDE ION Cl VEXZGXHMUGYJMC-UHFFFAOYSA-M	Cl^-
NA Query on NA Download Ideal Coordinates CCD File Download Instance Coordinates	F [auth A], G [auth A], M [auth B], N [auth B]	SODIUM ION Na FKNQFGJONOIPTF-UHFFFAOYSA-N	Na^+

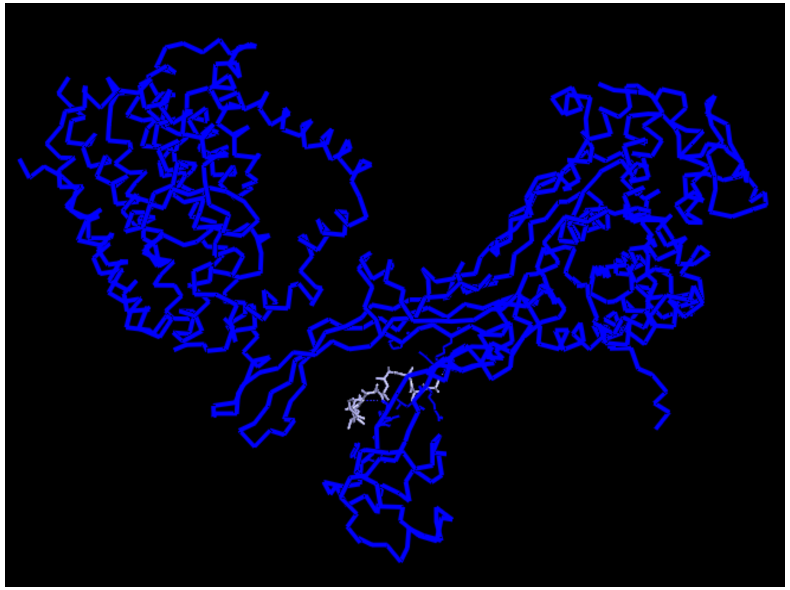
Binding site 6PL5 VS 6EJI

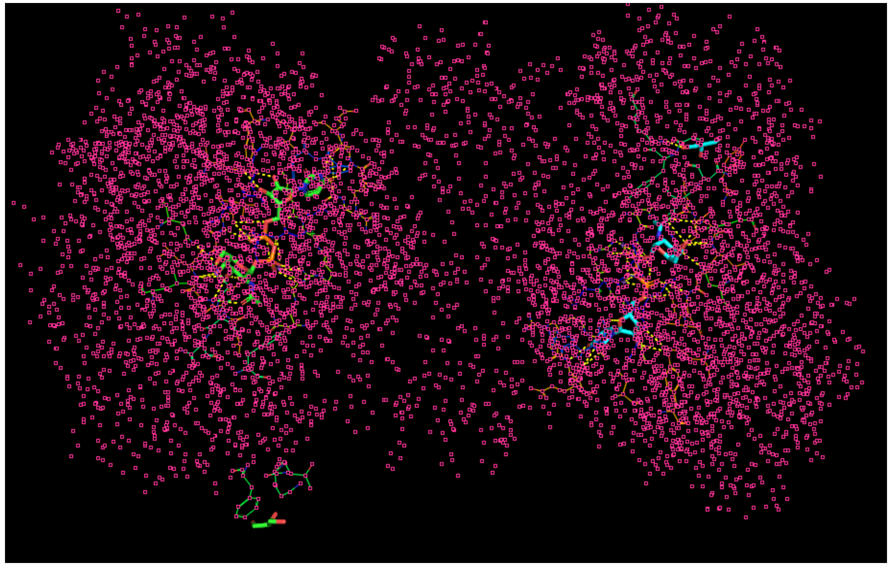
6PL5: white region is where the ligand is bind

6EJI: red dot which is entire protein is where the ligand is bind



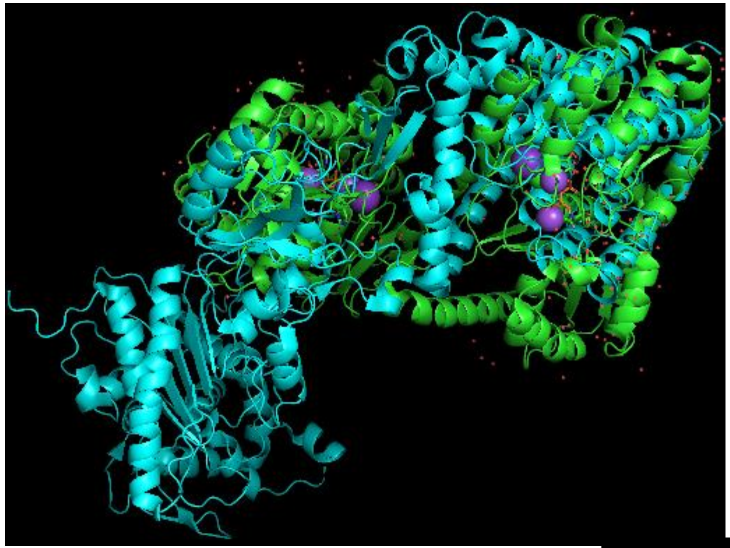




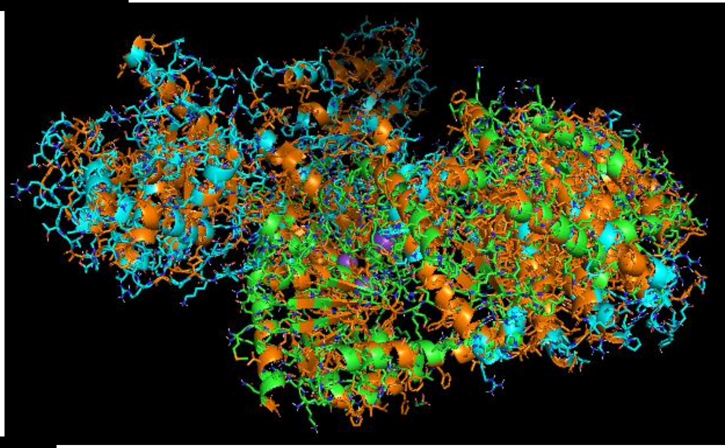


Superimposable Images

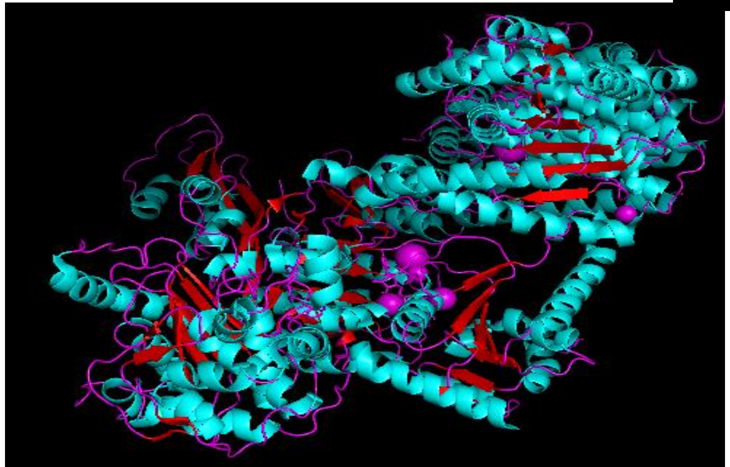
- Using Pymol we were able to construct 3 images related to the 2 proteins
- The orange color is used to represent the hydrophobes
- The pink red color is used to represent the secondary structures present



- Superimposable



- Hydrophobic Interactions



- Secondary Structures

Questions?