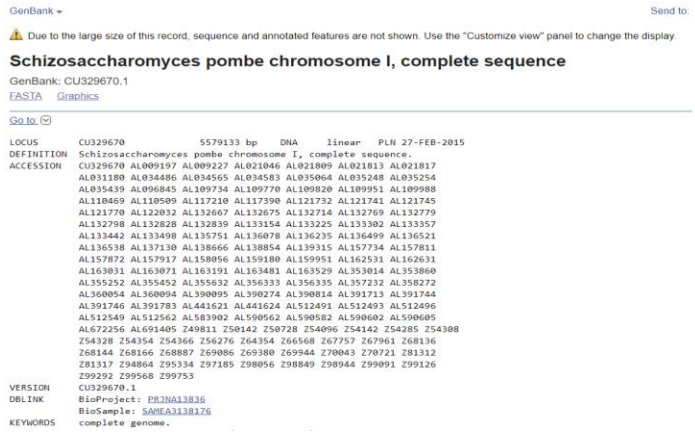
BMEN 6387/BIOL 5376: Applied Bioinformatics
Problem Set 1 (based on material from weeks 1-3)

Due 10 AM 9/8/20

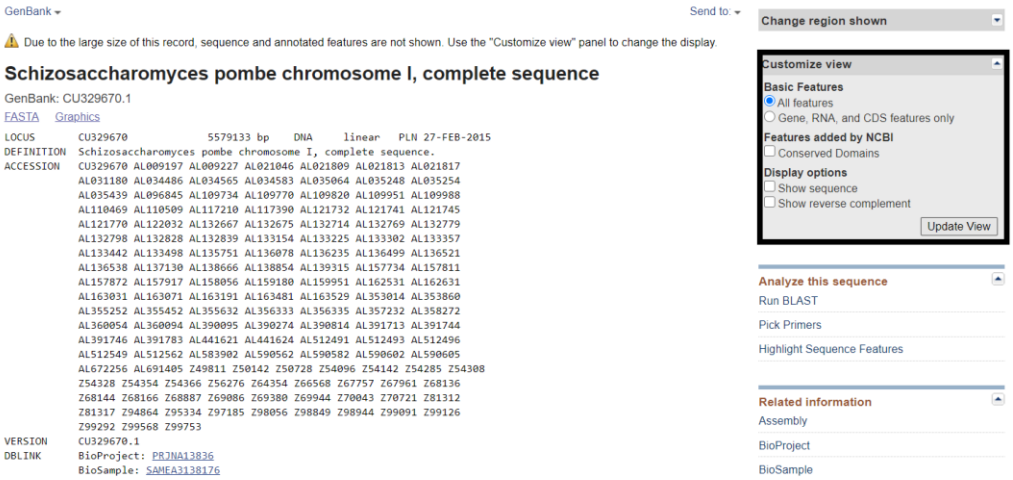
Name: LIPI VIKRAM THAKKER
NET-ID: LXT190004

1. For this exercise you will need to access GenBank by going to the NCBI website and use the dropdown window to search "nucleotide". Note that the definition of the coding strand is the strand of DNA within the gene that is identical to the genetic code. Conversely, the template strand is the strand that is complementary to the coding strand.

   A. Use the following accession number to access the nucleotide sequence in GenBank: CU329670



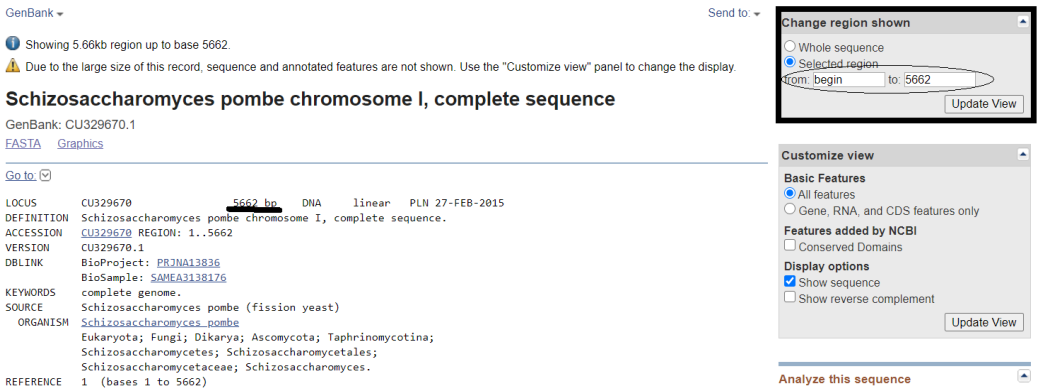
   B. Go to the FEATURES section of the record.



   C. Link to the CDS to gain access to the first 5662 nucleotides of the sequence.

BMEN 6387/BIOL 5376: Applied Bioinformatics
Problem Set 1 (based on material from weeks 1-3)

Due 10 AM 9/8/20



D. Name the protein product of the CDS.

- The protein product of the CDS is “RecQ type DNA helicase”.

E. Write the first four amino acids (starting from the N-terminus).

- The first four amino acid of the sequence is M, V,V and A.
- M -> Methionine
- V -> Valine
- A -> Alanine

BMEN 6387/BIOL 5376: Applied Bioinformatics
Problem Set 1 (based on material from weeks 1-3)

Due 10 AM 9/8/20

```
/db_xref="InterPro:IPR001650"
/db_xref="InterPro:IPR001878"
/db_xref="InterPro:IPR011545"
/db_xref="InterPro:IPR014001"
/db_xref="InterPro:IPR027417"
/db_xref="PomBase:SPAC212.11"
/db_xref="PomBase:SPAC212.11.1"
/translation="MVVASEIAKVASKTARDIAGCFTCQCGTQFDNVERIVQHFKECR
YRDETCKDDDIVVYEPSSFVQDEKKDKPIIVEAASEATSEEACNSSKERQLPALSALS
ALSTLTTSANDDLWTARLIWQSTNDTKLDNSPSSNYTDLNHKLANYGLSILSIHALMC
VECECLLNVIHTAQHMQIVHKLELNEDLLWFQELRTLKLKSPTNVLQTHSSQTHVYPY
IRGLPVLLNGYECVPCTKNGTGFVHAIMDTFRHHVRRTHGKVIKLENCIRRTALQTVK
NKYAQRCQFFKVDYVPLNGGEEEEEEEGEEKEDAQNIKERMVDFCFSKFMEKNQQRRE
QQDKGENKKRQDDVDQATDNNTNTILEDDEKDNDEEEEEEIVNAREKNLLNQQFNWTA
IVKKLGENWDQLVRFEYTNGIVTLDTIVNQLIRYYYRGFRHLSGMTMGMRRMFTQGGS
YSAQERGLCRLEQKDTVVRYAQSAALYLIFLLRRPSADSGIRRHLEAMCGATVERKEG
GSNSSSNISNVANFDSAEDDNDNDNDNDRDSNNNNNNNNTNTDDDDKLAYLELHEALK
LAFLQQYDFSKNVQDLEIMEFLACMSLHKDGTSKYAYEISACFAPLIYTCRLVAACEL
QRLIDEKQIDLLSIPSFQTAGSIAYAHVFCFITLGQRNLYDVLYETQKVVRDIIRTEG
YANTLQGLSPSTVLFQPRSNSMYPCIGDAFNNMVRLDLSELTALYEGMFAKVQDLLKE
LCFDMNVEKLLPISLLRSIGDDINNSKLGYSFFKESIEIRSSHSVLLRTILKNSELCH
RFFPSMSKKDLTKLFGGVSDQQRNECDNYSNHYNDNSNDNDNDVFLKLHWSKSAIKKY
ETKASIFNELLFCLVYISAGQPARAQEMVYWTLRNGKYKTRELYLMFGRLMIYSRYDK
TRNMKFAEKPIPRFLSEPLSILALRYYVLVRPLEALMKYVTTADRSKVAVYLDFMFVI
AGERLQRDLPYRIFPKATYQCIQKPLGFRNYRHIAHYFKEKNIEEEMTRESYFDLQAG
HTRNTALYIYGRTMDNLHYLPSDYFANFFRASYKWQELLQIRDNPTHGLLVETKHPFI
KRVDQLEEALNEKLARLVGEQMVEGDKEKDKTNEEKNKDEVKAEMTQPVVNQDSHDLQ
DQLATTPTAPTAFHYRPGLLQPSQTSVQHCCWALSQYYGLEAKFRSLKQFQSVYFSLL
NRMNLITVLPTGGGKSLSFLIPALIEKKRQTPGKVMNMVTLVLVPMMSLRQDMMLRVN
EKGLLVCSGNWTAFKDVRLTLETQLPDLFILTYESALTNSGLRFFESLATLGRLARVV
IDEAHLLLTSGAWRTALSRASRLSGLYAPLHLLSATFPRQLEMVARQTFCTNFYVLRE
```

F. Write the nucleotide sequence of the coding strand that corresponds to these amino acids.

- The nucleotide sequence of the coding strand that corresponds to these amino acids are TAC, CAG, CAG and CGA.

G. Write the nucleotide sequence of the template strand that corresponds to these amino acids.

- The nucleotide sequence of the template strand that corresponds to these amino acids are ATG GTC GTC GCT.

H. Using the sequence shown in the record, give the nucleotide number range that corresponds to these amino acids.

- the nucleotide number range that corresponds to these amino acids is 5550-5562.

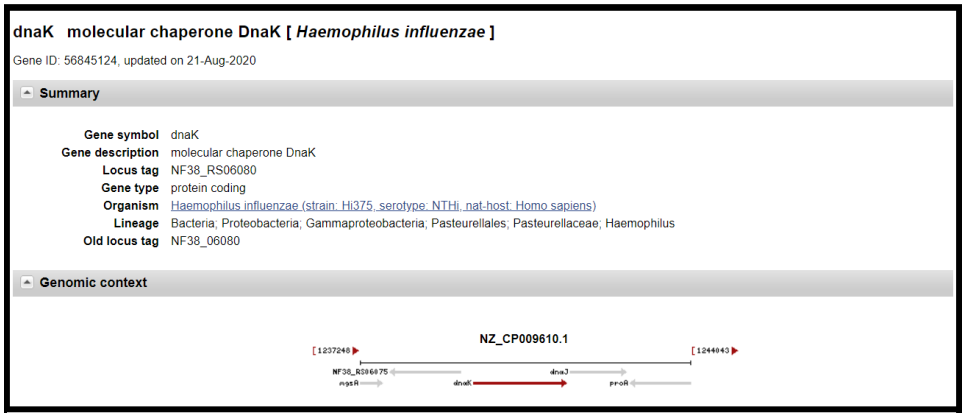
BMEN 6387/BIOL 5376: Applied Bioinformatics
Problem Set 1 (based on material from weeks 1-3)

Due 10 AM 9/8/20

2. Retrieve the sequence of the gene *dnaK* from the organism *Haemophilus influenzae.*

A. Give as much information as possible about this genome (*e.g.*, prokaryotic vs. eukaryotic, linear vs. circular, etc.)

- The genomic sequence is NZ_CP009610.1.
- Gene ID: 56845124.
- strain="Hi375"
- size: 1908 bp.
- genome structure: Linear.
- H*aemophilus influenzae* is a prokaryotic organism (without well-defined nucleus).
- Proteins length: 635 aa
- Protein position: 139
- Also called molecular chaperone DnaK.



B. What are the coordinates of this gene?

- The co-ordinates of the gene dnaK is 1,239,572- 1,241,479

C. What is the biological function of the gene product?
The biological function of the chaperone is :

- cellular response to unfolded protein.
- chaperone cofactor-dependent protein refolding
- protein refolding
- response to unfolded protein

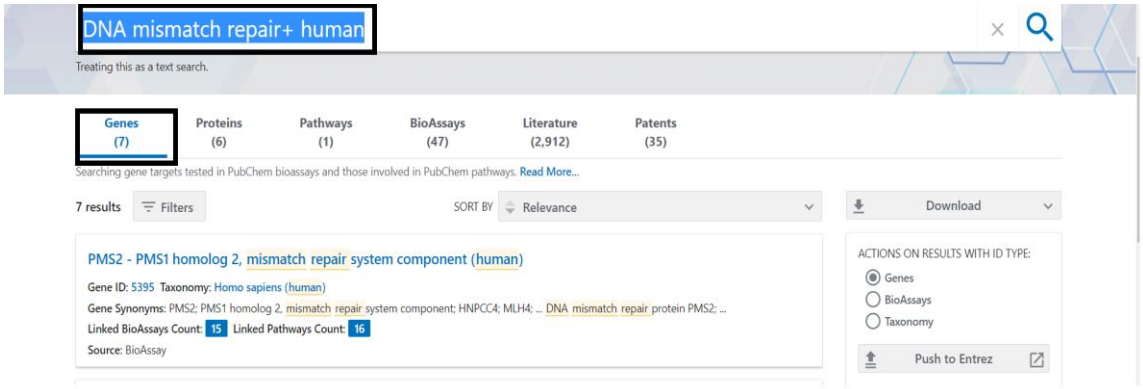
BMEN 6387/BIOL 5376: Applied Bioinformatics
Problem Set 1 (based on material from weeks 1-3)

Due 10 AM 9/8/20

3. Retrieve a list of all known human genomic loci that are involved in DNA mismatch repair. Describe in reasonable detail the steps you took to obtain this list.

- The reference of the below answer is from PubChem database.
- The result obtained from the database is thoroughly based on literature.
- I obtained the below list by writing [DNA mismatch repair+ human] as search query.
- The list for the human genomic locu involved in DNA mismatch repair is:

PMS1
PMS2
MSH2
MSH3
MSH6
MLH1
MLH3



4. Locate the RefSeq record for SARS-Cov-2

A. Locate the sequence of the spike protein "S" in the record. What are the values of the CDS coordinates, /product keyword, and the protein_id?

- **21563..25384 // RANGE OF S PROTEIN IN PROTEIN STRAND.**
- **/gene="S"**
- /locus_tag="GU280_gp02"
- /gene_synonym="spike glycoprotein"
- /note="structural protein; spike protein"
- /codon_start=1

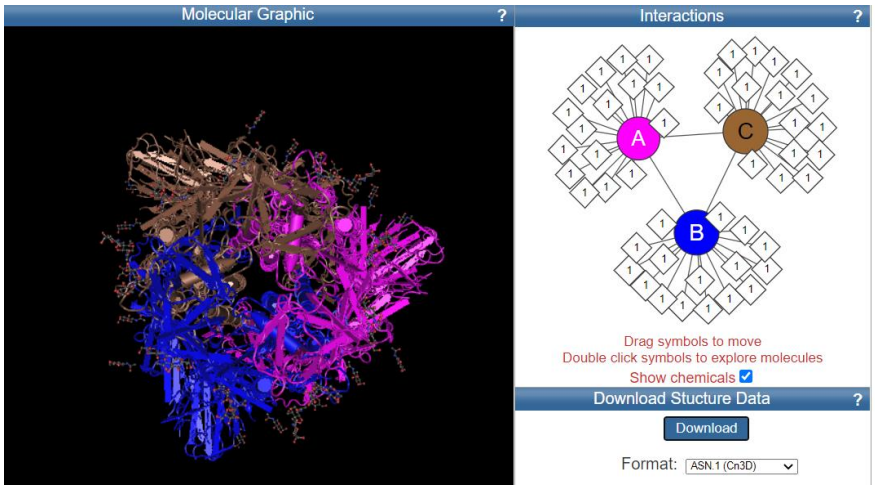
BMEN 6387/BIOL 5376: Applied Bioinformatics
Problem Set 1 (based on material from weeks 1-3)

Due 10 AM 9/8/20

- /product="surface glycoprotein"
- **/protein_id="YP_009724390.1"**
- /db_xref="GeneID:43740568"
-

B. Find the structure of the S protein in the PDB (MMDB).

- 6X79: Prefusion SARS-CoV-2 S ectodomain trimer covalently stabilized in the closed conformation
- MMDB ID: 191449
- PDB DEPOSITION DATE: 2020/05/29.



C. How many x-ray structures of this protein are available in the database?

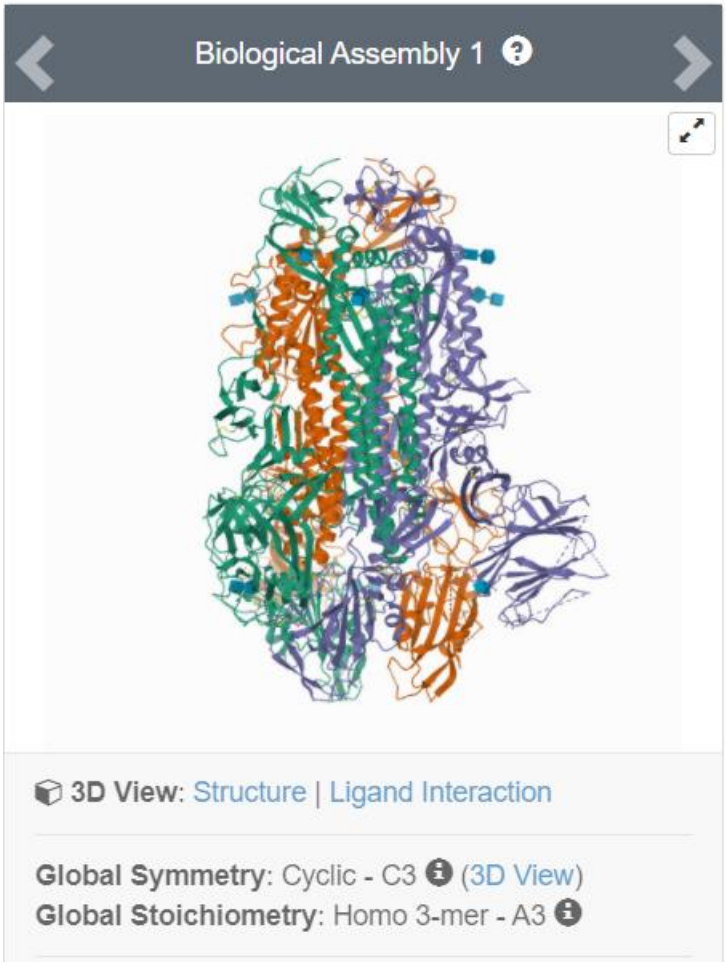
- Till date, there is only one x-ray structure of this protein available on MMDB.

D. What is the stoichiometry of the S protein in the biological assembly?

- The symmetry of the S protein in the biological assembly is cyclic-c3 globally.
- The stoichiometry of S Protein in biological assembly is homo 3-mer-A3.

BMEN 6387/BIOL 5376: Applied Bioinformatics
Problem Set 1 (based on material from weeks 1-3)

Due 10 AM 9/8/20



E. (*Extra credit*) Include a screenshot of the biological assembly such that all polypeptides in the complex are shown in distinct colors.