

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

- 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

GenBank Send to

Due to the large size of this record, sequence and annotated features are not shown. Use the "Customize view" panel to change the display.

Schizosaccharomyces pombe chromosome I, complete sequence

GenBank: CU329670.1

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

[Go to](#)

LOCUS CU329670 5579133 bp DNA linear PLN 27-FEB-2015

DEFINITION Schizosaccharomyces pombe chromosome I, complete sequence.

ACCESSION CU329670 AL009197 AL009227 AL021046 AL021809 AL021813 AL021817

AL031180 AL034486 AL034565 AL034583 AL035064 AL035248 AL035254

AL035439 AL096845 AL109734 AL109770 AL109820 AL109951 AL109988

AL110469 AL110509 AL117210 AL117390 AL121732 AL121741 AL121745

AL121770 AL122032 AL132667 AL132675 AL132714 AL132769 AL132779

AL132798 AL132828 AL132839 AL133154 AL133225 AL133302 AL133357

AL133442 AL133498 AL135751 AL136078 AL136235 AL136499 AL136521

AL136538 AL137130 AL138666 AL138854 AL139315 AL157734 AL157811

AL157872 AL157917 AL158056 AL159180 AL159951 AL162531 AL162631

AL163031 AL163071 AL163191 AL163481 AL163529 AL163804 AL163860

AL355252 AL355452 AL355632 AL356333 AL356335 AL357232 AL358272

AL360054 AL360094 AL390095 AL390274 AL390814 AL391713 AL391744

AL391746 AL391783 AL441621 AL441624 AL512491 AL512493 AL512496

AL512549 AL512562 AL583902 AL590562 AL590582 AL590602 AL590605

AL672256 AL691405 AL98111 Z50142 Z50728 Z54096 Z54142 Z54285 Z54308

Z54328 Z54354 Z54366 Z56276 Z64354 Z66568 Z67757 Z67961 Z68136

Z68144 Z68166 Z68887 Z69086 Z69380 Z69944 Z70043 Z70721 Z81312

Z81317 Z94864 Z95334 Z97185 Z98056 Z98849 Z98944 Z99091 Z99126

Z99292 Z99568 Z99753

VERSION CU329670.1

DBLINK BioProject: [PRJNA13836](#)

KEYWORDS BioSample: [SAMEA3138176](#)
complete genome.

- B. Go to the FEATURES section of the record.

GenBank Send to

Due to the large size of this record, sequence and annotated features are not shown. Use the "Customize view" panel to change the display.

Schizosaccharomyces pombe chromosome I, complete sequence

GenBank: CU329670.1

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

LOCUS CU329670 5579133 bp DNA linear PLN 27-FEB-2015

DEFINITION Schizosaccharomyces pombe chromosome I, complete sequence.

ACCESSION CU329670 AL009197 AL009227 AL021046 AL021809 AL021813 AL021817

AL031180 AL034486 AL034565 AL034583 AL035064 AL035248 AL035254

AL035439 AL096845 AL109734 AL109770 AL109820 AL109951 AL109988

AL110469 AL110509 AL117210 AL117390 AL121732 AL121741 AL121745

AL121770 AL122032 AL132667 AL132675 AL132714 AL132769 AL132779

AL132798 AL132828 AL132839 AL133154 AL133225 AL133302 AL133357

AL133442 AL133498 AL135751 AL136078 AL136235 AL136499 AL136521

AL136538 AL137130 AL138666 AL138854 AL139315 AL157734 AL157811

AL157872 AL157917 AL158056 AL159180 AL159951 AL162531 AL162631

AL163031 AL163071 AL163191 AL163481 AL163529 AL163804 AL163860

AL355252 AL355452 AL355632 AL356333 AL356335 AL357232 AL358272

AL360054 AL360094 AL390095 AL390274 AL390814 AL391713 AL391744

AL391746 AL391783 AL441621 AL441624 AL512491 AL512493 AL512496

AL512549 AL512562 AL583902 AL590562 AL590582 AL590602 AL590605

AL672256 AL691405 AL98111 Z50142 Z50728 Z54096 Z54142 Z54285 Z54308

Z54328 Z54354 Z54366 Z56276 Z64354 Z66568 Z67757 Z67961 Z68136

Z68144 Z68166 Z68887 Z69086 Z69380 Z69944 Z70043 Z70721 Z81312

Z81317 Z94864 Z95334 Z97185 Z98056 Z98849 Z98944 Z99091 Z99126

Z99292 Z99568 Z99753

VERSION CU329670.1

DBLINK BioProject: [PRJNA13836](#)
BioSample: [SAMEA3138176](#)

Customize view

Basic Features

☒ All features

☐ Gene, RNA, and CDS features only

Features added by NCBI

☐ Conserved Domains

Display options

☐ Show sequence

☐ Show reverse complement

[Update View](#)

Analyze this sequence

[Run BLAST](#)

[Pick Primers](#)

[Highlight Sequence Features](#)

Related information

[Assembly](#)

[BioProject](#)

[BioSample](#)

- 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

GenBank

Showing 5.66kb region up to base 5662

Due to the large size of this record, sequence and annotated features are not shown. Use the "Customize view" panel to change the display.

Schizosaccharomyces pombe chromosome I, complete sequence

GenBank: CU329670.1

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

[Go to](#)

LOCUS CU329670 5662 bp DNA linear PLN 27-FEB-2015

DEFINITION Schizosaccharomyces pombe chromosome I, complete sequence.

ACCESSION [CU329670](#) REGION: 1..5662

VERSION CU329670.1

DBLINK BioProject: [PRJNA13836](#)
BioSample: [SAMFA3138176](#)

KEYWORDS complete genome.

SOURCE Schizosaccharomyces pombe (fission yeast)

ORGANISM Schizosaccharomyces pombe

Eukaryota; Fungi; Dikarya; Ascomycota; Taphrinomycotina;
Schizosaccharomycetes; Schizosaccharomycetales;
Schizosaccharomycetaceae; Schizosaccharomyces.

REFERENCE 1 (bases 1 to 5662)

Change region shown

☐ Whole sequence

☒ Selected region

from begin to 5662

Update View

Customize view

Basic Features

☒ All features

☐ Gene, RNA, and CDS features only

Features added by NCBI

☐ Conserved Domains

Display options

☒ Show sequence

☐ Show reverse complement

Update View

Analyze this sequence

D. Name the protein product of the CDS.

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

```
CDS
complement(<1..5662)
/gene="tlh1"
/locus_tag="SPAC212.11"
/codon_start=1
/product="RecQ type DNA helicase"
/protein_id="CAC05745.1"
/db_xref="EnsemblGenomes-Gn:SPAC212.11"
/db_xref="EnsemblGenomes-Tr:SPAC212.11.1"
/db_xref="GOA:P0CT33"
/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"
```

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="MVAASEIAKVASKTARDIAGCFTCQCGTQFDNVERIVQHFKECR  
YRDETKDDDIVVYEPSSFVQDEKKDKPIIVEAASEATSEEACNSSKERQLPALSA  
ALSTLTSSANDDLWARTLIWQSTNDTKLDNSPSSNYTDLNHLKLANYGSLSIHALMC  
VECECLLNVIHTAQHMQIVHKLELNEDLLWFQELRTLKLSPTNVLQTHSSQTHVYPY  
IRGLPVLLNGYECVPCTKNGTGFVHAIMDTFRHHVRRTHGKVIKLENCIRRTALQTVK  
NKYAQRCCQFKVDYVPLNGGEEEEEEGEEKEDAQNIKERMVDFCFSKFMEKNQQRRE  
QQDKGENKKRQDDVDQATDNNTNTILEDEKDNDEEEEEIYNAREKNLLNQQFNWTA  
IVKKLGENWDQLVRFEYTINGIVTLDITVNLIRYYRGFRHLSGMTMGMRMFTQGGG  
YSAQERGLCRLEQKDTVVRYAQAALYLIFLLRRPSADSGIRRHLEAMCGATVERKEG  
GSNSSNISNVANFDSAEDDNDNDNDNRDSNNNNNNNTNTDDDDKLAYELHEALK  
LAFLLQYDFSKNVQDLEIMEFLACMSLHKDGTSKYAYEISACFAPLIYTCRLVAACEL  
QRLIDEKQIDLLSIPSFQTAGSIAYAHVFCITLGQRNLYDVLVETQKVVRDIIRTEG  
YANTLQGLSPSTVLFPQRSNSMYPICIGDAFNMMVRLDSELTALYEGMFAKVQDLLKE  
LCFDMNVEKLLPISLLRSIGDDINNSKLGYSFFKESIEIRSSSHSVLLRTILKNSELCH  
RFFPSMSKKDLTKLFGGVSQQRNECDNYSNHYNDNSNDNDNDVFLKLHWSKSAIKKY  
ETKASIFNELLFCLVYISAGQPARAQEMVYVTLRNGKYKTRELYLMFGRLLMIYSRYDK  
TRNMKFAEKPIPRFLSEPLSILALRYVYVLRPLEALMKYVTTADRSKVAVYLDPMFVI  
AGERLQRDLPRYIFPKATYQCIQKPLGFRNYRHIAHYFKEKNIEEEMTRESYFDLQAG  
HTRNTALYIYGRITMDNLHYLPSDYFANFRASYKQWELLQIRDNPTHGLLVETKHPFI  
KRVDQLEALNEKLARLVGEQMVGEKDKEKDTNEEKNKDEVKAEMTQPVVNQDSHDLQ  
DQLATPTAPTAFHYRPGLLQPSQTSVQHCCWALSQYYGLEAKFRSLKQFQSVYFSL  
NRMNLITVLP TGGGKSLSFLIPALIEKKRQTGKVMNMVTLVLPMMSLRQDMMLRVN  
EKGLLVCSGNWTAFAKDVRLTLETQLPDLFILTYESALTNSGLRFFESLATLGRLARVV  
IDEAHL L TSGAWRTALSRASRLSGLYAPLHLLSATFPRQLEMVARQTECTNFYVLR
```

- 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

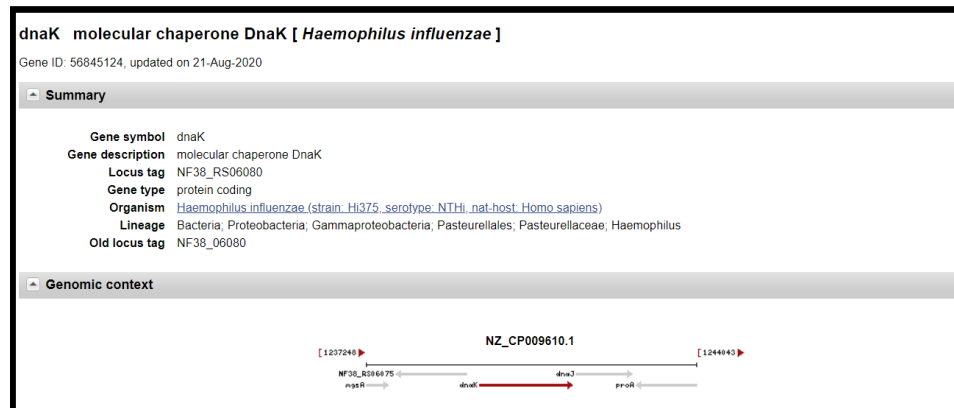
Problem Set 1 (based on material from weeks 1-3)

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.
- *Haemophilus 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 loci involved in DNA mismatch repair is:

PMS1
PMS2
MSH2
MSH3
MSH6
MLH1
MLH3

The screenshot shows a PubChem search interface. At the top, the search bar contains the text "DNA mismatch repair+ human". Below the search bar, there are tabs for different result types: Genes (7), Proteins (6), Pathways (1), BioAssays (47), Literature (2,912), and Patents (35). The "Genes" tab is selected. The search results show 7 results. The first result is highlighted: "PMS2 - PMS1 homolog 2, mismatch repair system component (human)". Below this title, it shows the Gene ID: 5395, Taxonomy: Homo sapiens (human), and Gene Synonyms: PMS2; PMS1 homolog 2; mismatch repair system component; HNPCC4; MLH4; ... DNA mismatch repair protein PMS2; ... It also shows Linked BioAssays Count: 15 and Linked Pathways Count: 16. On the right side, there are options to download the results and actions on results with ID type: Genes, BioAssays, and Taxonomy. There is also a "Push to Entrez" button.

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

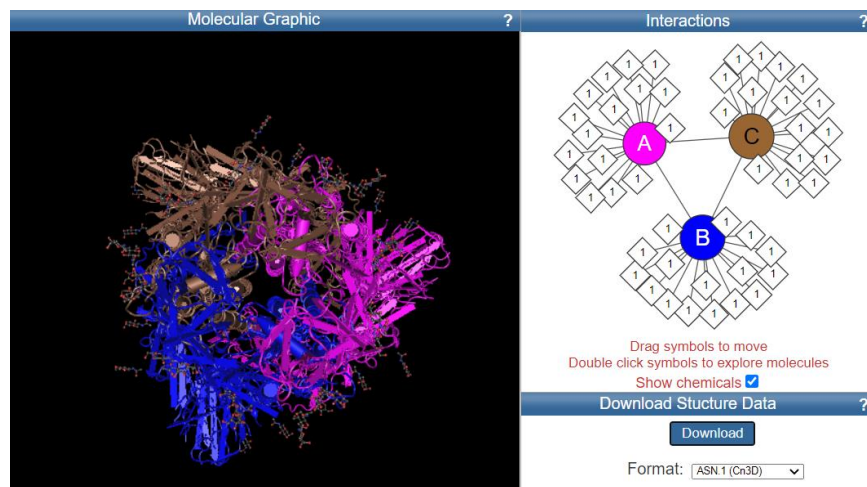
Problem Set 1 (based on material from weeks 1-3)

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.

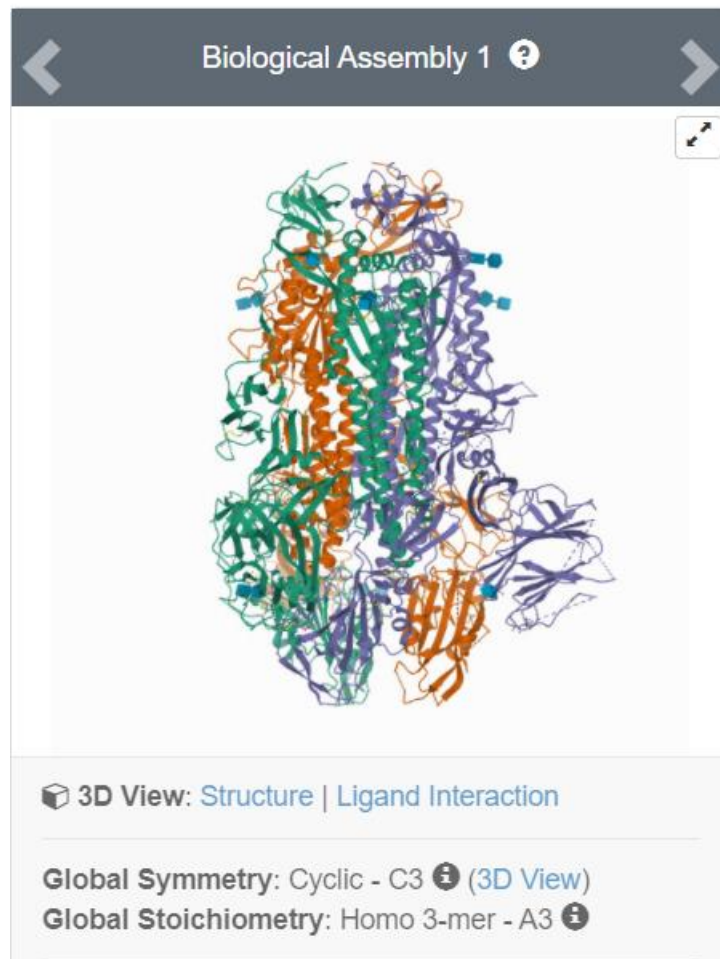
D. What is the stoichiometry of the S protein in the biological assembly?

- 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.

