Practical Questions

1) BLAST based

i) The downloaded file is in the document “Q1” of the submitted folder.

ii) The result of performing BLASTN:

Descriptions | Graphic Summary | Alignments | Taxonomy

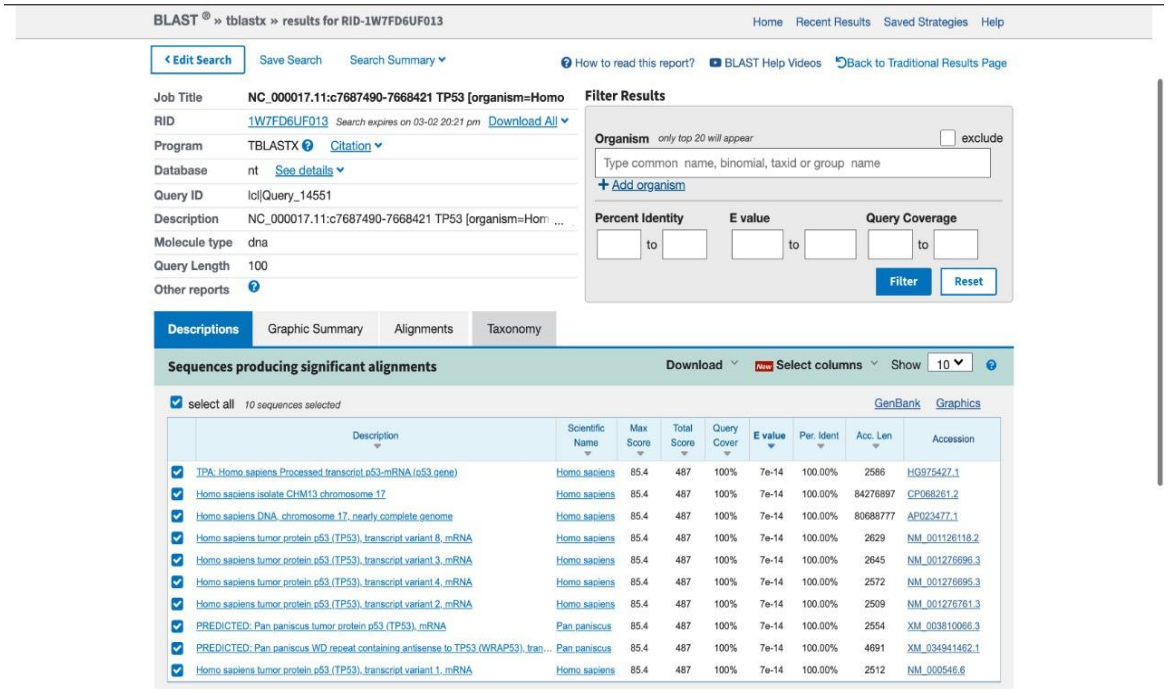
Sequences producing significant alignments Download New Select columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA	Homo sapiens	4639	4639	100%	0.0	100.00%	2512	NM_000546.6
TPA: Homo sapiens Processed transcript p53-mRNA (p53 gene)	Homo sapiens	4639	4639	100%	0.0	100.00%	2586	HG975427.1
Homo sapiens tumor protein p53 (TP53), transcript variant 2, mRNA	Homo sapiens	4619	4619	100%	0.0	99.88%	2509	NM_001276761.3
Homo sapiens p53 protein mRNA, complete cds	Homo sapiens	4599	4599	99%	0.0	99.88%	2521	AF307851.1
Homo sapiens tumor protein p53, mRNA (cDNA clone MGC:646 IMAGE:3544714), complete cds	Homo sapiens	4580	4580	99%	0.0	99.88%	2508	BC003596.1
PREDICTED: Pan troglodytes tumor protein p53 (TP53), transcript variant X2, mRNA	Pan troglodytes	4573	4573	100%	0.0	99.52%	2573	XM_001172077.5
PREDICTED: Gorilla gorilla gorilla tumor protein p53 (TP53), transcript variant X3, mRNA	Gorilla gorilla g…	4556	4556	100%	0.0	99.44%	2562	XM_004058511.3
PREDICTED: Pan paniscus tumor protein p53 (TP53), mRNA	Pan paniscus	4532	4532	100%	0.0	99.28%	2554	XM_003810066.3
Homo sapiens p53 protein (TP53) mRNA, complete cds, alternatively spliced	Homo sapiens	4518	4518	100%	0.0	99.09%	2584	DQ191317.1
Homo sapiens mutant p53 mRNA, complete cds	Homo sapiens	4506	4506	99%	0.0	99.09%	2513	FJ207420.1
Homo sapiens p53 protein (TP53) mRNA, complete cds, alternatively spliced	Homo sapiens	4505	4505	99%	0.0	99.09%	2529	DQ286964.1
Homo sapiens mRNA for P53, complete cds	Homo sapiens	4457	4457	96%	0.0	99.71%	2451	AB082923.1
PREDICTED: Gorilla gorilla gorilla tumor protein p53 (TP53), transcript variant X2, mRNA	Gorilla gorilla g…	4348	4348	95%	0.0	99.42%	2619	XM_019013137.2
PREDICTED: Gorilla gorilla gorilla tumor protein p53 (TP53), transcript variant X1, mRNA	Gorilla gorilla g…	4348	4348	95%	0.0	99.42%	2692	XM_019013136.2
PREDICTED: Pongo abelii tumor protein p53 (TP53), mRNA	Pongo abelii	4322	4322	100%	0.0	97.73%	2570	XM_002826974.4
Homo sapiens tumor suppressor TP53 (TP53) mRNA, complete cds, alternatively spliced	Homo sapiens	4241	4641	100%	0.0	100.00%	2697	JN900492.1
Homo sapiens tumor protein p53 (TP53), transcript variant 8, mRNA	Homo sapiens	4241	4641	100%	0.0	100.00%	2629	NM_001126118.2
Homo sapiens cell line HCT116p53-/- nonfunctional delta40-P53 transcript variant (TP53) mRNA, co…	Homo sapiens	4217	4406	95%	0.0	99.83%	2418	JN214348.1
PREDICTED: Pan troglodytes tumor protein p53 (TP53), transcript variant X1, mRNA	Pan troglodytes	4174	4574	100%	0.0	99.48%	2692	XM_016931470.2
PREDICTED: Nomascus leucogenys tumor protein p53 (TP53), mRNA	Nomascus leuc…	4135	4135	100%	0.0	96.43%	2567	XM_030800485.1
PREDICTED: Hylobates moloch tumor protein p53 (TP53), mRNA	Hylobates mol…	4117	4117	100%	0.0	96.31%	2561	XM_032139337.2
PREDICTED: Piliocolobus tephrosceles tumor protein p53 (TP53), transcript variant X1, mRNA	Piliocolobus te…	3982	3982	100%	0.0	95.35%	2562	XM_023193314.2
PREDICTED: Cercocebus atys tumor protein p53 (TP53), transcript variant X1, mRNA	Cercocebus atys	3971	3971	100%	0.0	95.27%	2556	XM_012052423.1
PREDICTED: Papio anubis tumor protein p53 (TP53), mRNA	Papio anubis	3965	3965	100%	0.0	95.23%	2560	XM_003912272.5

Feedback

The result of performing TBLASTX:

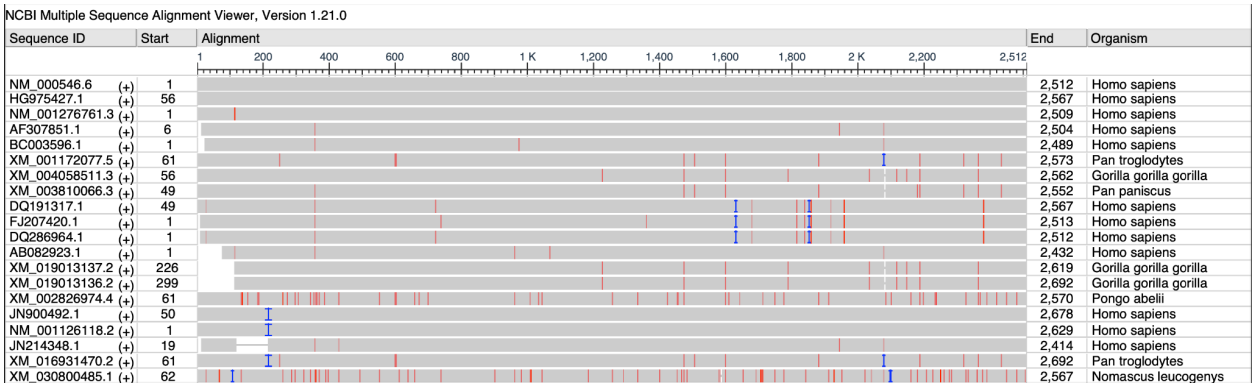


iii) **E-value:** The E value is a number which describes how many times one would expect a match by chance in a database of that size. The lower the E value is, the more significant the match and vice versa. The mathematical equation for E-value is $E = Kmn(e^{-\lambda S})$, where the parameters K and λ represent natural scales for the search space and the scoring system respectively and the rest of the equation represents the size of the query (m), the size of the database (n), and the S score (S).

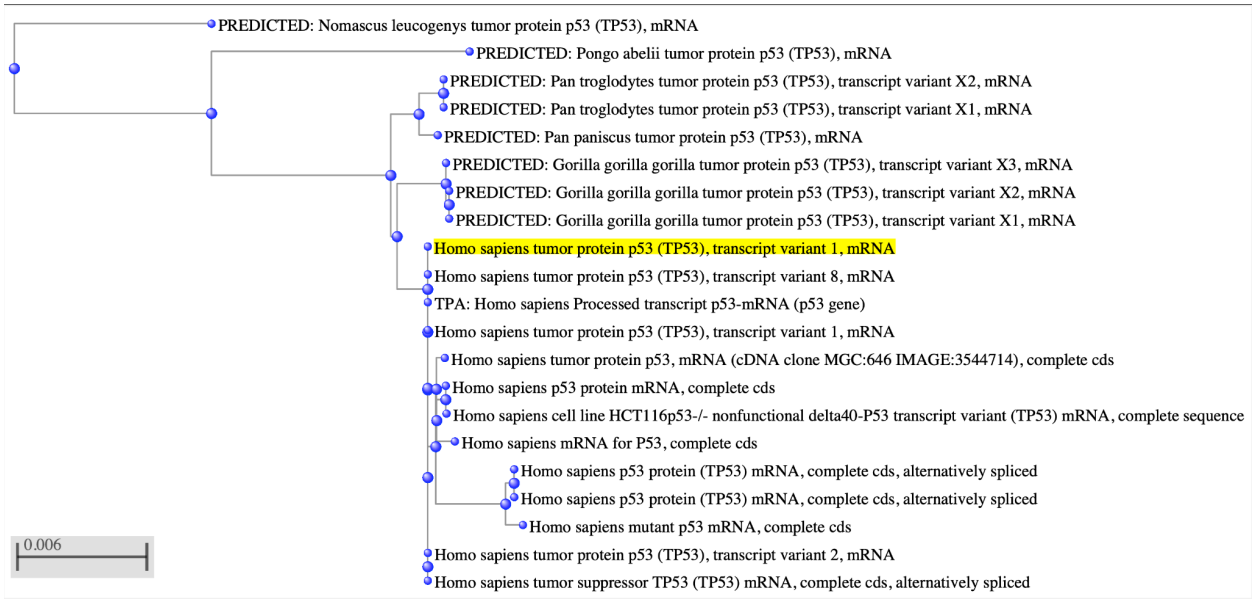
Query Coverage: Query coverage is a number that describes how much of the query sequence is covered by the target sequence. If the target sequence in the database spans the whole query sequence, then the query cover is 100%. This tells us how long the sequences are, relative to each other.

Percent Identity: The percent identity is a number that describes how similar the query sequence is to the target sequence as in how many characters in each sequence are identical. The higher the percent identity is, the more significant the match and vice versa.

iv) MSA of the top 20 resulting sequences based on E-value by performing BLASTN.



v) Phylogenetic Tree of the MSA result obtained in previous part:



2) Central DOGMA based

i) The downloaded file is in the document “Q2” of the submitted folder.

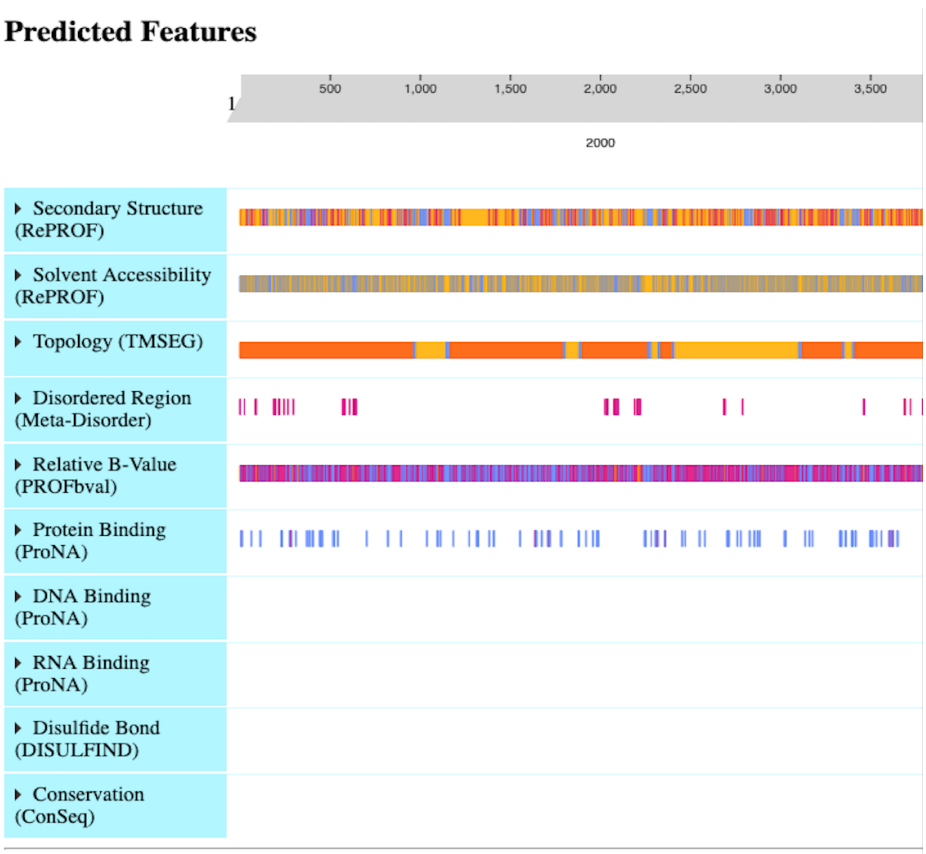
ii) The python file is in the document “Q2” of the submitted folder as ‘ii’.

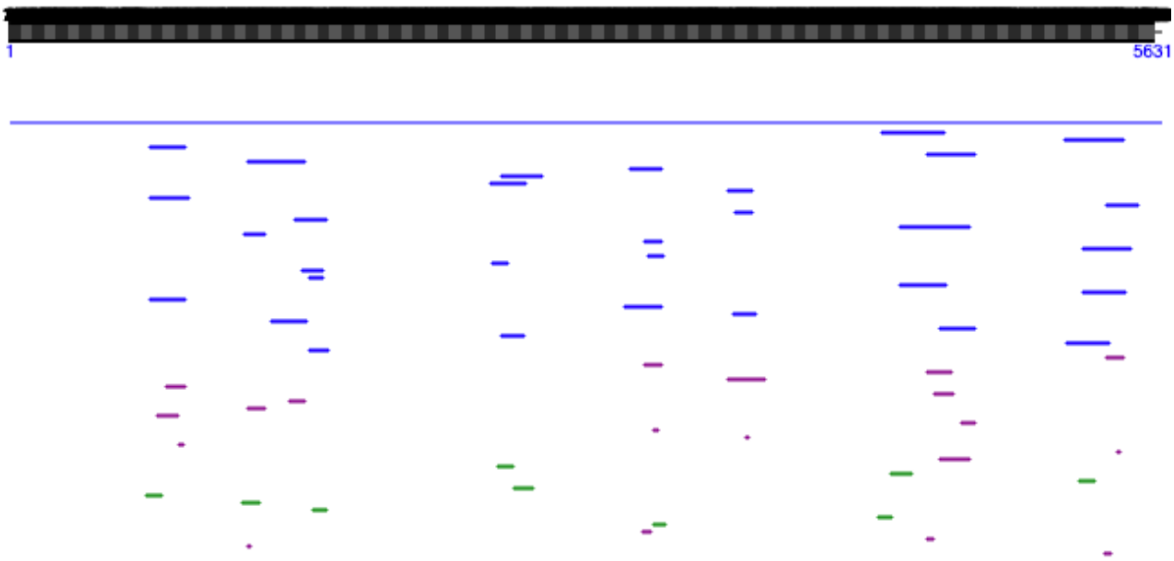
iii) 1220. The python file is in the document “Q2” of the submitted folder as ‘iii’.

iv) a) ${}^{1220}P_5$
b) ${}^{1220}P_3$
c) ${}^{1220}P_{10}$
d) ${}^{1220}P_6$

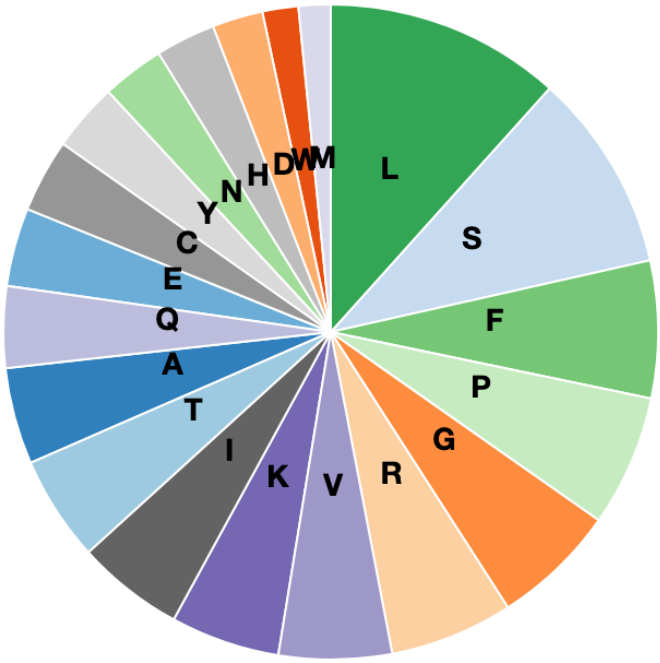
v) The R file is in the document “Q2” of the submitted folder.

vii)





Amino Acid composition



3) BIOMART and ALTERNATIVE SPLICING based

i) Chromosomes are made of DNA, and genes are special units of chromosomal DNA. There are a total of 23 pairs of chromosomes in humans,The total number of genes on each chromosome can be found out using the bm function.
Chromosome 1: 3000
Chromosome 2:2500
Chromosome 3:1900
Chromosome 4:1600
Chromosome 5:1700
Chromosome 6:1900
Chromosome 7:1800
Chromosome 8:1400
Chromosome 9:1400
Chromosome 10:1400
Chromosome 11:2000
Chromosome 12:1600
Chromosome 13:800
Chromosome 14:1200
Chromosome 15:1200
Chromosome 16:1300
Chromosome 17:1600
Chromosome 18:600
Chromosome 19:1700
Chromosome 20:900
Chromosome 21:400
Chromosome 22:800
Chromosome X:1400
Chromosome Y: 200

ii) The sum of the total number of genes is approximately 34300.

iii) a) The chromosome having the maximum number of genes is Chromosome 1. There are about 3543 transcripts.

b) There are on average 3.42 transcripts per gene. For ORF 127 of Chromosome 1, there are 3 transcripts.

C1orf127 Transcripts BETA

C1orf127 — chromosome 1 open reading frame 127 (*Homo sapiens*)

3 transcripts

← Switch to gene view

Edit Download Select columns

Gene ID	Symbol	Transcript	Length (nt)	Protein	Length (aa)	Protein name	Is
148345	C1orf127	XM_017000373.1	2679	XP_016855862.1	787	uncharacterized protein C1orf127	X
148345	C1orf127	NM_001170754.2	2798	NP_001164225.1	823	uncharacterized protein C1orf127	1
148345	C1orf127	NM_001366227.2	2461	NP_001353156.1	684	uncharacterized protein C1orf127	2

c) Transcript length is defined by Ensembl as the total length of the exons in a gene plus the lengths of its untranslated (UTR) regions. The start and end positions for all exons+UTR regions of all transcripts are to be extracted and the gene length based on the combined length between those regions are to be calculated. Using R, the transcripts with the highest transcript length per gene are selected. The maximum length is of the Transcript NM_001170754.2 for the ORF 127 in Chromosome 1.

d) The FASTA nucleotide sequence of the transcript has been downloaded and given in this folder by the name ‘q3_d’.

e) This has been done using Python and the folder has been attached by the name “q3_e”.

g) The top 5 organisms sharing homologous similarity with this transcript:

BLAST ® » blastn suite » results for RID-24WZ8EA6013

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	Nucleotide Sequence
RID	24WZ8EA6013 *Search expires on 03-06 03:17 am* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_15995
Description	None
Molecule type	dna
Query Length	2798
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 to E value to Query Coverage to

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download New Select columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens chromosome 1 open reading frame 127 (C1orf127), transcript variant 1,...	Homo sapiens	5168	5168	100%	0.0	100.00%	2798	NM_001170754.2
PREDICTED: Pongo abelii chromosome 1 C1orf127 homolog (C1H1orf127), transcript...	Pongo abelii	4554	4554	98%	0.0	96.45%	2764	XM_002811509.4
PREDICTED: Pan troglodytes chromosome 1 C1orf127 homolog (C1H1orf127), mRNA	Pan troglody...	4324	4324	88%	0.0	98.34%	2512	XM_016953903.1
PREDICTED: Gorilla gorilla gorilla chromosome 1 C1orf127 homolog (C1H1orf127), m...	Gorilla gorill...	4300	4300	88%	0.0	98.06%	2472	XM_031001361.1
Homo sapiens chromosome 1 open reading frame 127 (C1orf127), transcript variant 2,...	Homo sapiens	4154	4386	84%	0.0	99.96%	2461	NM_001366227.2

The MSA and the tree of the top 50 organisms have been attached in the folder with names ‘q3_g_MSA.pdf’ and ‘q3_g_Tree.pdf’ respectively.

Theoretical Questions

1)

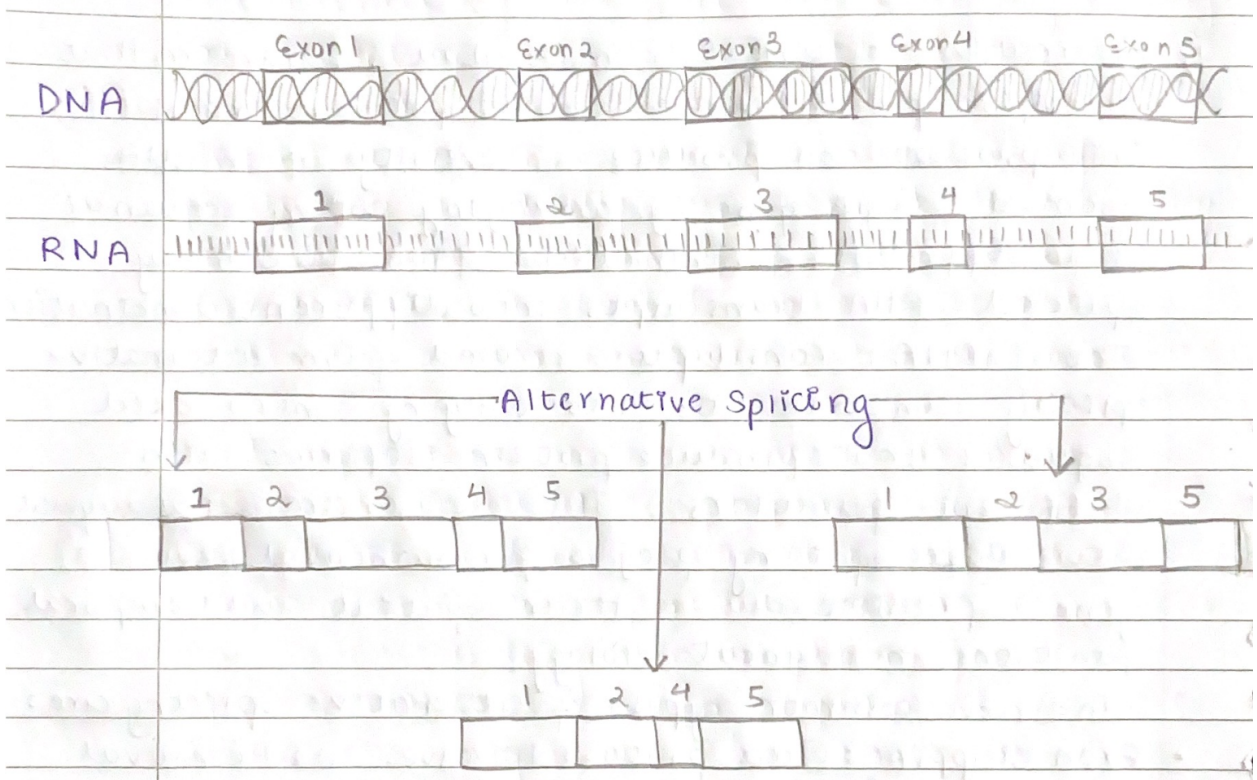
Theoretical Questions

1) Alternative splicing is a molecular mechanism that modifies pre-mRNA constructs prior to translation. This process can produce a diversity of mRNAs from a single gene by arranging coding sequences into different combinations from recently spliced RNA transcripts into different combinations. The mRNA transcripts created from alternative splicing can translate into varying amino acid sequences that produce protein isoforms with different functions. The mechanisms that regulate alternative splicing play a fundamental role in gene expression due to their spatial and temporal functions throughout biology.

The most common types of alternative splicing are:

- Exon skipping: This process involves the removal of certain exons and their adjacent introns from mRNA constructs prior to translation.
- Alternate 5' to 3' splicing: Alternative splicing can also be mediated by joining of exons at alternative 5' or 3' splice sites
- Intron retention: This type happens when non-coding portions of a gene are retained in the final mRNA transcript.

Alternative Splicing



2)

2) BLAST (basic local alignment tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences. A BLAST search enables a researcher to compare a subject protein or nucleotide sequence with a library or database of sequences, and identify database sequences that resemble the query sequence above a certain threshold.

Types of Blast & their uses:

(1) BLASTN: compares a DNA query to a DNA database. searches both strands automatically. It is optimized for speed, rather than sensitivity.

(2) BLASTP: compares a protein query to a protein database

(3) BLASTX: Compares a DNA query to a protein database by translating the query sequence in the 6 possible frames, and comparing each against the database (3 reading frames from each strand of DNA) searching.

(4) TBLASTN: compares a protein query to a DNA database, in the 6 possible frames of the database.

(5) TBLASTX: compares the protein encoded in a DNA query to the protein encoded in a DNA database, in the 6*6 possible frames of both query and database sequences.

(6) BLAST2: It is called advanced BLAST and can perform gapped alignments.

(7) PSI-BLAST - (Position Specific Iterated) Performs iterative database searches.

3)

3) Phylogenetics is the systematic study of reconstructing the past evolutionary history of extant species or taxa, based on present day data, such as morphologies or molecular information. This evolutionary history or phylogeny is ideally represented as a binary tree. In the method of phylogenetic invariants, a pivotal role is played by certain master functions. The aim is to use the observed data as input for their evaluation and dependent on the answers, to infer essential features of phylogenetic history without the need for parameter estimation.

Phylogenetics is important because it enriches our understanding of how genes, genomes, species evolve. Through phylogenetics, we learn not only how the sequences come to be the way they are today, but also general principles that enable us to predict how they will change in the future.

The Machine Learning algorithms which are used for forming phylogenetic trees are Maximum Likelihood and Maximum Parsimony

4)

4) Ways in which quantity of protein can be regulated in a cell:

- Regulation by binding of an effector and by covalent modification:

Protein activity can also be controlled by binding of effector molecules, which often work by inducing conformational changes that produce inactive or active forms of the protein. Effectors that regulate activity by binding to the active site usually take form of inhibitors that compete with the substrate for binding. Often, the product of an enzyme reaction can act as such a competitive inhibitor, allowing the enzyme to regulate itself when too much product might be made.

- Regulation by protein quantity and lifetime:

The activity of a protein can also be regulated by controlling it's amount and lifetime in the cell. This control may be exercised at several places in the flow of information from gene to protein. At it's simplest, the amount of protein can be set by the level of transcription. At the level of protein, quantities are controlled by the lifetime of the molecule, which is determined by it's rate of degradation.

- Regulation by small molecules:

Most enzymes are controlled by changes in their conformation which in turn alter catalytic activity. In many cases such conformational changes result from the binding of small molecules, such as amino acids or nucleotides, that regulate enzyme activity. This type of regulation commonly is responsible for controlling metabolic pathways by feedback inhibition.

- Regulation by Protein-Protein Interactions :
Many proteins consist of multiple subunits, each of which is an independent polypeptide chain. In some proteins the subunits are identical; other proteins are composed of two or more distinct polypeptides. In either case, interactions between the polypeptide chains are important in regulation of protein activity.

- Regulation by Protein Phosphorylation :
This regulation activates enzymes by proteolytic cleavage of inactive precursors. Phosphorylation is readily reversible within the cell and function, to reversibly activate or inhibit a wide variety of cellular proteins in response to environmental changes or signals.

5)

5) Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestlye for each person. This approach will allow doctors and researchers to predict more accurately which treatment and prevention strategies are developed for the average person, with less consideration for the differences between individuals.

I feel it has immense potential to bring a revolution in the field of medicine. It will change the way of treating diseases and will cure many patients through specific medicines. Considering how each and every person is different in the context of their lifestyll and genes, developing such case-specific medicines on a large-scale would be an intensive and costly affair, but it would enable patients to recieve better treatment according to their individual body conditions. The Researchers at National cancer institute hope to we an increased knowledge of the genetics and biology of cancer to find new, more effective treatments for various forms of the disease. This method can truly be revolutionary.