1). Complete Biological Central Dogma

a).

i.

Frame used: 5'3' Frame 3

R V G A A A N P A L R S R A L P A P A R R P A P G C P C C E G R R A R R L P Q P S C L A A G P R R A R A A G G G A A A R R R R P E P Met A Y S Q G G G K K K V C Y Y Y D G D I G N Y Y Y G Q G H P Met K P H R I R Met T H N L L L N Y G L Y R K Met E I Y R P H K A T A E E Met T K Y H S D E Y I K F L R S I R P D N Met S E Y S K Q Met Q R F N V G E D C P V F D G L F E F C Q L S T G G S V A G A V K L N R Q Q T D Met A V N W A G G L H H A K K S E A S G F C Y V N D I V L A V L E L L K Y H Q R V L Y I D I D I H H G D G V E E A F Y T T D R V Met T V S F H K Y G E Y F P G T G D L R D I G A G K G K Y Y A V N F P Met R D G I D D E S Y G Q I F K P I I S K V Met E Met Y Q P S A V V L Q C G A D S L S G D R L G C F N L T V K G H A K C V E V V K T F N L P L L Met L G G G G Y T I R N V A R C W T Y E T A V A L D C E I P N E L P Y N D Y F E Y F G P D F K L H I S P S N Met T N Q N T P E Y Met E K I K Q R L F E N L R Met L P H A P G V Q Met Q A I P E D A V H E D S G D E D G E D P D K R I S I R A S D K R I A C D E E F S D S E D E G E G G R R N V A D H K K G A K K A R I E E D K K E A E D K R T D V K E E D K S K D N S G E K T D T K G A K S E Q L N N P Stop I Stop L P N L R N L E K Stop D D S G I R N L R C L R K F G F I L Y C F G Met D C I Y F Q N G L F L V F L G K F Y C E F F Stop L Stop S K F F S T Met L Y V I V F K L Met C Y Y V K K K K P D L L K K Q L A F L S Stop F F H L L Stop L S L L K N C T W Met V F C L F I Met K A C F Q A R L Stop D W S I P Met V I S V A E T L S L S L Y V L T G S N R Y F L F Q Stop K Stop H H V T L Y S Stop K K K K K

ii.

In gene expression there are 6 possible reading frames. This is because in a double strand nucleic acid there are two strands. Each strand is a made of a series of 4 nucleobases A, C, G, T. For Transcription, the strand is grouped into triplets which are called codons. Now there can be three ways to read a strand. For example:

DNA Strand 1:

AGGTGACACCGCAAGCCTTATATTAGC

Frame 1: AGG·TGA·CAC·CGC·AAG·CCT·TAT·ATT·AGC

Frame 2: A·GGT·GAC·ACC·GCA·AGC·CTT·ATA·TTA·GC

Frame 3: AG·GTG·ACA·CCG·CAA·GCC·TTA·TAT·TAG·C

So, codons depend on where the reading starts in a strand.

Similarly, for the second strand (the complementary strand) it will be:

DNA Strand 2:

TCCACT…….

Frame 4: TCC·ACT………

Frame 5: T·CCA·CT……..

Frame 6: TC·CAC.T………

iii.

The code is attached in the files.

Nucleotide sequence: CGCGCGTCGGCGCCGCTGCCAACCCCGCTCTGCGATCTCGGGCCCTCCCG

Protein Sequence I got: RASAPLPTPLCDLGPS

Amino Acids: ['Arginine', 'Alanine', 'Serine', 'Alanine', 'Proline', 'Leucine', 'Proline', 'Threonine', 'Proline'

'Leucine', 'Cysteine', 'Aspartic Acid', 'Leucine', 'Glycine', 'Proline']

b).

i.

We can use Blastx which is a part of Basic Local Alignment Search Tool (BLAST) on NCBI. It finds similarities between biological sequences. It uses the submitted nucleotide sequence to find similar protein and organisms.

ii. The sequence closely matches with [Rattus norvegicus](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=10116&lvl=3&lin=f&keep=1&srchmode=1&unlock) (Norway rat). The description in Blastx shows that it has a query cover of 70 %.

iii. The name of the protein which matches to the sequence is Hdac2 protein [Rattus norvegicus].

The function of this protein is to positively regulate the cell differentiation, positively regulate the macromolecule metabolic process and respond to alkaloid. It localizes to chromatin, cytoplasm, and nucleus. This protein is used to study Peyronie's disease; alcohol use disorder; and anxiety disorder. It exhibits other functions: including deacetylase activity; heat shock protein binding activity; and promoter-specific chromatin binding activity. It is also used as a biomarker of bronchitis; chronic obstructive pulmonary disease; endometriosis; pulmonary emphysema; and visual epilepsy.

I have researched online and found this information from this url: <https://rgd.mcw.edu/rgdweb/report/gene/main.html?id=619976>

2). Biological Sequence alignment.

a).

ACCTAGCTAG

ACCCACCGGG

Use Match: +1, Mismatch: -1 and Indel: -2

i.

Alignment 1:

ACCTAGCTAG

ACCCACCGGG

+1 +1 +1 -1 +1 -1 + 1 -1 -1 + 1 = 6 – 4 = 2

Alignment 2:

Indel **C** at position 4 sequence 1

ACC**C**TAGCTAG

ACCCACCGGG

-2 +1 +1 +1 + 1 - 1 -1 - 1 -1 -1 -1 = 4 -8 = -4

Alignment 3:

Indel **C** at position 6 sequence 1

ACCTA**C**GCTAG

ACCCACCGGG

-2 +1 +1 +1 - 1 + 1 + 1 -1 -1 -1 -1 = 5 -7 = -2

Alignment 4:

Indel **G** at position 8 sequence 1

ACCTAGC**G**TAG

ACCCACCGGG

-2 +1 +1 +1 -1 +1 -1 + 1 +1 -1 – 1 = 6 – 6 = 0

Alignment 5:

Indel **G** at position 9 sequence 1

ACCTAGCT**G**AG

ACCCACCGGG

-2 +1 +1 +1 -1 +1 -1 + 1 -1 + 1 – 1 = 6 – 6 = 0

Alignment 6:

Indel **T** at position 4 in sequence 2

ACCTAGCTAG

ACC**T**CACCGGG

-2 +1 +1 +1 + 1 -1 -1 + 1 -1 -1 + 1 = 6 - 6 = 0

Alignment 7:

Indel **G** at position 6 in sequence 2

ACCTAGCTAG

ACCCA**G**CCGGG

-2 +1 +1 +1 -1 +1 +1 + 1 -1 -1 + 1= 7 – 5 = 2

Alignment 8:

Indel **T** at position 8 in sequence 2

ACCTAGCTAG

ACCCACC**T**GGG

-2 +1 +1 +1 -1 +1 -1 + 1 +1 -1 + 1 = 7 – 5 = 2

Alignment 9:

Indel **A** at position 9 in sequence 2

ACCTAGCTAG

ACCCACCG**A**GG

-2 +1 +1 +1 -1 +1 -1 + 1 -1 +1 + 1 = 7 – 5 = 2

ii. I have tried only 10 alignments with single indel in both upper and lower sequences. We can see that the best score I got was in Alignments: 1, 7, 8, 9 that is 2. There can be many possible alignments by adding more nucleotide bases, but It would reduce the total score of alignment due to negative scoring of Indel. So, I think these alignments are the best ones so far.

b).

i*.*Homo sapiens (Human)

>sp|P55073|IOD3\_HUMAN Thyroxine 5-deiodinase OS=Homo sapiens OX=9606 GN=DIO3 PE=1 SV=4

MPRQATSRLVVGEGEGSQGASGPAATMLRSLLLHSLRLCAQTASCLVLFPRFLGTAFMLW

LLDFLCIRKHFLGRRRRGQPEPEVELNSEGEEVPPDDPPICVSDDNRLCTLASLKAVWHG

QKLDFFKQAHEGGPAPNSEVVLPDGFQSQHILDYAQGNRPLVLNFGSCTUPPFMARMSAF

QRLVTKYQRDVDFLIIYIEEAHPSDGWVTTDSPYIIPQHRSLEDRVSAARVLQQGAPGCA

LVLDTMANSSSSAYGAYFERLYVIQSGTIMYQGGRGPDGYQVSELRTWLERYDEQLHGAR

PRRV

Rattus norvegicus (Rat)

>sp|P49897|IOD3\_RAT Thyroxine 5-deiodinase OS=Rattus norvegicus OX=10116 GN=Dio3 PE=1 SV=3

MPRQAASRLVVGEGEGPPGASGPAATMLRSLLLHSLRLCAQTASCLVLFPRFLGTAFMLW

LLDFLCIRKHFLRRRHPDHPEPEVELNSEGEEMPPDDPPICVSDDNRLCTLASLKAVWHG

QKLDFFKQAHEGGPAPNSEVVRPDGFQSQRILDYAQGTRPLVLNFGSCTUPPFMARMSAF

QRLVTKYQRDVDFLIIYIEEAHPSDGWVTTDSPYVIPQHRSLEDRVSAARVLQQGAPGCA

LVLDTMANSSSSAYGAYFERLYVIQSGTIMYQGGRGPDGYQVSELRTWLERYDEQLHGTR

PRRL

Bos taurus (Bovine)

>sp|Q5I3B1|IOD3\_BOVIN Thyroxine 5-deiodinase OS=Bos taurus OX=9913 GN=DIO3 PE=2 SV=3

MSRQAAPRWVVGEGRGTLGGAATMLRSLLLHSLRLCSQTASCLVLFPRFLGTAFMLWLLD

FLCIRKHLLGRRRRGQPEIEVELNSDGEEVPPDDPPVCVSDDNRLCTLASLRAVWHGQKL

DFFKQAHEGGPAPNSEVVLPDGFQNQHILDYARGNRPLVLNFGSCTUPPFMARMSAFQRL

VTKYQRDVDFLIIYIEEAHPSDGWVTTDSPYSIPQHRSLEDRVSAARVLQQGAPECALVL

DTMTNSSSSAYGAYFERLYIIQSGTIMYQGGRGPDGYQVSEVRTWLERYDEQLHGPQPRR

V

Ovis aries (Sheep)

>sp|Q6DN07|IOD3\_SHEEP Thyroxine 5-deiodinase (Fragment) OS=Ovis aries OX=9940 GN=DIO3 PE=2 SV=3

VVGEGRGALGGAATMLRSLLLHSLRLCAQTASCLVLFPRFLGTAFMLWLLDFLCIRKHLL

GRRRRGQPEIEVELNSDGEEVPPDDPPVCVSDDNRLCTLASLRAVWHGQKLDFFKQAHEG

GPAPNSEVVLPDGFQNQHILDYARGNRPLVLNFGSCTUPPFMARMSAFQRLVTKYQRDVD

FLIIYIEEAHPSDGWVTTDSPYSIPQHRSLEDRVSAARVLQQGAPECALVLDTMTNSSSS

AYGAYFERLYIIQSGTIMYQGGRGPDGYQVSELRTWLERYDEQLHGPQPRRV

Xenopus laevis (African clawed frog)

>sp|P49899|IOD3\_XENLA Thyroxine 5-deiodinase OS=Xenopus laevis OX=8355 GN=dio3 PE=1 SV=2

MLHCAGPHTGKLVKQVAACCLLLPRFLLTGLMLWLLDFQCIRRRVLLTAREESTAEHEDP

PLCVSDSNRMCTVESLRAVWHGQKLDYFKSAHLGCSAPNTEVVMLEGRRLCKILDFSQGK

RPLVVNFGSCTUPPFMARLQAYRRLAAQHVGIADFLLVYIEEAHPSDGWLSTDASYQIPQ

HQCLQDRLAAAQLMLQGAPGCRVVVDTMDNSSNAAYGAYFERLYIVLEGKVVYQGGRGPE

GYKISELRMWLEQYQQGLMGTKGSGQVVIQV

ii.

A picture containing text

Description automatically generated

iii.

The alignment results show the protein sequences for all 5 organisms share a lot of similarities, but some organisms share more similarities than other. The Xenopus laevis (African clawed frog) has very different sequence in the beginning as compared to other sequences. We can see that some amino acids are missing in Xenopus laevis (African clawed frog). Also Bovine, Sheep, Rat, and Human share a lot of similarities except Bovine and Sheep having a different sequence in the beginning. The ‘\*’ symbol shows positions which have a single, fully conserved residue. Looking at the figure above we can say that all 5 sequences have many fully conserved residues, and the sequence has been maintained by natural selection. A ‘:’ (colon) indicates conservation between groups of strongly similar properties and a ‘.’ (period) indicates conservation between groups of weakly similar properties.

iv.

Graphical user interface, text

Description automatically generated

The Phylogenetic tree shows the relationship of the protein sequence among 5 different organisms. In the phylogenetic tree we can see that Human and Rat protein sequences are closely related (similar) to one another as they are on the same level in the tree. They are on the top level of the tree displayed in two separate branches, illustrating two separate entities (but closely related as compared to others). Next branch in the tree splits into two parts which shows that Bovine and sheep are under one branch and African clawed frog is under another branch. This shows that the protein sequence of Bovine and sheep share similarities whereas the protein sequence from African clawed frog is different hence displayed in a separate branch.

3). Hidden Markov Model (HMM)

i.

GATTAG

= 1\* 1 \* 0.5 \* 0.25 \* 0.2 \* 0.5 \* 0.4 \* 0.15 \* 0.6 \* 0.25 \* 1 \* 0.5 \* 1

=0.00005625

ii.

GTAAG

B -> S1-> S2 -> S4 -> S5 -> S7-> E

=>1 \* 1 \* 0.5 \* 0.5 \* 0.4 \* 0.4 \* 0.6 \* 0.25 \* 1 \* 0.5 \* 1 = .003

B -> S1-> S2 -> S4 -> S6 -> S7-> E

=> 1\* 1 \* 0.5 \* 0.5 \* 0.4 \* 04 \* 0.4 \* 0 \* 0.7 \* 0.5 \* 1 = 0

B -> S1-> S3-> S4 -> S6 -> S7-> E

=> 0

B -> S1-> S3-> S4 -> S5 -> S7-> E

=> 1 \* 1 \* 0.5 \* 0.3 \* 0.4 \* 0.4 \* 0.6 \* 0.25 \* 1 \* 0.5 \* 1 = 0.0018

The probability of seeing GTAAG is sum of all possible probabilities to get the given sequence. That is:

= 0.003 + 0 + 0 + 0.0018 = 0.0048

iii.

GTACGG

B -> S1-> S2-> S3-> S4 -> S6 -> S7-> E

= 1 \* 1 \* 0.5 \* 0.5 \* 0.6 \* 0.25 \* 0.4 \* 0.15 \* 0.4 \* 0.30 \* 0.7 \* 0.5 \* 1=0.0000945

B -> S1-> S2-> S3-> S4 -> S5 -> S7-> E

= 1 \* 1 \* 0.5 \* 0.5 \* 0.6 \* 0.25 \* 0.4 \* 0.15 \* 0.6 \* 0.05 \* 1 \* 0.5 \* 1 = 0.00003375

B -> S1-> S3 -> S2-> S4 -> S6 -> S7-> E

= 1 \* 1 \* 0.5 \* 0.30 \* 0.2 \* 0.5 \* 0.4 \* 0.15 \* 0.4 \* 0.3 \* 0.7 \* 0.5 \* 1 = 0.0000378

B -> S1-> S3 -> S2-> S4 -> S5 -> S7-> E

= 1 \* 1 \* 0.5 \* 0.30 \* 0.2 \* 0.5 \* 0.4 \* 0.15 \* 0.6 \* 0.05 \* 1 \* 0.50 \* 1= 0.0000135

B -> S1-> S3 -> S3-> S4 -> S6 -> S7-> E

= 1 \* 1 \* 0.5 \* 0.30 \* 0.4 \* 0.25 \* 0.4 \* 0.25 \* 0.4 \* 0.30 \* 0.7 \* 0.5 \* 1 = 0.000063

B -> S1-> S3 -> S3-> S4 -> S5 -> S7-> E

= 1 \* 1 \* 0.5 \* 0.30 \* 0.4 \* 0.25 \* 0.4 \* 0.25 \* 0.6 \* 0.05 \* 1 \* 0.5 \* 1= 0.0000225

B -> S1-> S3 -> S4 -> S6 -> S6 -> S7-> E

= 1 \* 1 \* 0.5 \* 0.30 \* 0.4 \* 0.4 \* 0.4 \* 0.4 \* 0.3 \* 0.3 \* 0.7 \* 0.5 \* 1 = 0.00012096

B -> S1-> S2 -> S4 -> S6 -> S6 -> S7-> E

= 1 \* 1 \* 0.5 \* 0.5 \* 0.4 \* 0.4 \* 0.4 \* 0.4 \* 0.3 \* 0.3 \* 0.7 \* 0.5 \* 1 = 0.0002016

The probability of seeing GTACGG is sum of all possible probabilities to get the given sequence. That is:

= 0.0000945 + 0.00003375 + 0.0000378 + 0.0000135 + 0.000063 + 0.0000225 + 0.00012096 + 0.0002016 = 0.00058761

4). Next Generation sequencing

i.

The highest number of SNP per exons is 56. It is given in the figure in step 5 below.

Step 1: Extracting EXONs

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, application, Word

Description automatically generated

Step 2: Extracting SNPs

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generated

Step 3: Join both SNPs and Exons.

Graphical user interface, text, application

Description automatically generated

Step 4: Grouping data

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

Step 5: Sort by Descending order

Graphical user interface, text, application, email

Description automatically generated

ii.

Step 6: Select Top 5 exons with highest number of SNPsGraphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generatediii.

Screen shots are provided above.