Q1. R Script

```
> # BIOL 365 Assign2 Gene Annotation using DECIPHER
> # Michelle Trinh
> # This assignment uses DECIPHER and dependent R packages
> R.version.string
[1] "R version 4.4.2 (2024-10-31 ucrt)"
> date()
[1] "Mon Feb  3 14:52:42 2025"
>
> # Task 1
> #Loading A2contig FASTA file from local copy to R
> A2contig <- readDNAStringSet("C:/Users/miche/OneDrive/Documents/University of Waterloo/Fifth Year 2024-2025/BIOL 365/A2contig.fasta", format="fasta")
>
> #Confirming data has loaded
> A2contig
DNAStringSet object of length 1:
    width seq                                                                                names
[1] 27162 GTTGAAAAATTAATCCTCTGTCCGTTTCTGCTTCATTGA...ACATTGAGTTAAATTCCGAAGGAATTGGAGCATTATGT JQGU01000058.1 Le...
>
> #Retrieving name of data: "JQGU01000058.1 Leptospira alexanderi strain 56643 Contig058, whole genome shotgun sequence"
> names(A2contig)
[1] "JQGU01000058.1 Leptospira alexanderi strain 56643 Contig058, whole genome shotgun sequence"
```

Q2. How many protein-coding genes were identified? 31 protein-coding genes were identified.

```
> #Finding genes using FindGenes
> orfs <- FindGenes(A2contig, showPlot=TRUE, allScores=TRUE)
Iter  Models Start Motif  Init  Fold UpsNt  Term   RBS  Auto  Stop Genes
7          3  8.39  1.97  5.31  0.52  3.60  2.55  2.07  1.84  0.17    31

Time difference of 6.99 secs

>
> #Viewing orfs results
> orfs
Genes object of size 2,980 specifying:
31 protein coding genes from 168 to 2,550 nucleotides.
2,949 open reading frames from 60 to 2,562 nucleotides.

  Index Strand Begin End TotalScore ... Gene
1     1      1     2  73      -4.65 ...    0
2     1      1     2  76       5.20 ...    0
3     1      1     2  88      -1.86 ...    0
4     1      1     2  91      -2.25 ...    0
5     1      1     2 100     -11.81 ...    0
6     1      1     2 109       6.45 ...    0
... with 2,974 more rows.
```

Q3. How long are the shortest and longest genes? Shortest: 168 bp Longest: 2550 bp

```
> #Subset ORFs that are predicted as genes
> genes <- orfs[orfs[, "Gene"]==1,]
>
> #Extract genes using ExtractGenes
> dna <-  ExtractGenes(genes, A2contig)
> dna
DNAStringSet object of length 31:
     width seq
 [1]  1164 ATGAAAAAGAAAATTCTCCTTTTA...GCTTCTAAAATTAAAGTCGATTAA
 [2]  2550 GTGGCACTCGATAAAGTCACAAGC...ATGTGTACGGACGAATCGTCTTAA
 [3]   792 GTGACTTTATCGAGTGCCACAAAA...TTAGAAGACGATACTGCAGATTAA
 [4]   420 ATGGAAGTCGTACTTTGTGGATTG...AAAAAGATGGAATATAAACTCTAA
 [5]   969 ATGAATGATAAAGGGAACTTTATA...AGTGTTCTAAACAACACGATTTAA
 ...   ... ...
[27]   516 ATTGGAAACCGTAGAACTCTTGTT...AGTTATGCTATTCAAGAATTTTGA
[28]   387 ATGACAGGAAAACAAAAGGCAATT...GCGATTCTATCAAAGGCCGACTAA
[29]   270 TTGGCGCTTTATTACCTGTGTAAT...TTCTATTACGGAACTAAACGTTAG
[30]  1605 GTGGTTAGGGTACTTGTCAATCTA...CTGGAATCGACCGATCATGGATGA
[31]   273 ATGAGGCTTTTTACTTGCAAAAAG...TCGGAAGCGTTTCTTCGCTCCTAA
>
> #Determine shortest and longest genes
> w <- which(width(dna) < 200)
> dna[w]
DNAStringSet object of length 2:
    width seq
[1]   192 ATGGCTCAATTTTTCTCTCAGTTGT...GGCAAAAGGTTTCAGGAATTGTAA
[2]   168 GTGAAAATTGGAGTTAAAGCTCCGA...AAAGATTCCGGAAAAAAAGGCTGA
> w <- which(width(dna) > 2000)
> dna[w]
DNAStringSet object of length 1:
    width seq
[1]  2550 GTGGCACTCGATAAAGTCACAAGCC...ATGTGTACGGACGAATCGTCTTAA
```

Q4. What are the start site locations and orientations for the first three genes?

```
> #Determine start site locations and orientations for first three genes
> w <- which(width(dna) == 1164)
> genes[w]
Genes object of size 1 specifying:
1 protein coding gene of 1,164 nucleotides.

   Index Strand Begin  End TotalScore ... Gene
44     1      0   710 1873     159.69 ...    1
> w <- which(width(dna) == 2550)
> genes[w]
Genes object of size 1 specifying:
1 protein coding gene of 2,550 nucleotides.

    Index Strand Begin  End TotalScore ... Gene
296     1      1  1870 4419     286.89 ...    1
> w <- which(width(dna) == 792)
> genes[w]
Genes object of size 1 specifying:
1 protein coding gene of 792 nucleotides.

    Index Strand Begin  End TotalScore ... Gene
488     1      0  4400 5191      91.07 ...    1
```

Gene #1: start site location at 710, orientation is on top strand (Strand = 0), otherwise known as sense strand that codes for proteins in 5’ to 3’ direction

Gene #2: start site location at 1870, orientation is on bottom strand (Strand = 1), otherwise known as antisense or template strand in 3’ to 5’ direction

Gene #3: start site location at 4400, orientation is on top strand (Strand = 0), otherwise known as sense strand that codes for proteins in 5’ to 3’ direction

Q5. What are the corresponding protein sequences for the first three genes?

```
> #Translate genes to obtain predicted protein sequences
> aa <- ExtractGenes(genes, A2contig, type="AAStringSet")
> aa
AAStringSet object of length 31:
     width seq
 [1]   388 MKKKILLLGSGELGKEFVIAAQRLGQYVIAVDSYDGAPAMQVAHEKEIINMLDGNLL...LDIASEMPESDFRIFGKPITRPYRRMGVALSYSAKGEGTSSLRKRAILLASKIKVD*
 [2]   850 MALDKVTSPLSCDLSVRQPGKRKRKIRFLIVSRILSVSIAIRMEEIKSSAIISANIA...LTKKGNYLIRDSDEILKQGLKILKQKKIVFLENENVKILDHNLIQYYANMCTDESS*
 [3]   264 MTLSSATKKYNGFFLFFLFLSYLLISSINCATYFQNRKNDLKDIFTLGVETPGYGAG...QAFNKRKDGYPKSFLYQIEVYLGIYVGLRIGFNPAELFDFLVGLAGLDLLEDDTAD*
 [4]   140 MEVVLCGLEFRLRSITVKKILSTFSILVFILILFFCSSENKKQPPRQEFQPNSDIRM...NTQDGAILEWYLVSTDYQKNSYKTLAEKPETISEDTKFIRLTIDKKGVIKKMEYKL*
 [5]   323 MNDKGNFIKTVLVTGGSGSLGLILLPELVKNYRVVCIGRNLSSFPDSIRFHSNFVFY...FNLANSLLANVFKKKVNLEYIQESSIVISDRIQKELDFQFKMDFEEGIQSVLNNTI*
 ...   ... ...
[27]   172 MGNRRTLVKRNPMPYPTDAILSPENRVSIAGTYFKDPSIHYGNASTRLRLFGFDGFI...FVFGKTSFYDELKQEMKKNKIKALFDSRLDIEQTAYLFWLIQKKCIRFKSYAIQEF*
[28]   129 MTGKQKAILILSSLSGFLGVALGAFGAHALKPILTPELFAIYETGNRYHLIHSIPPL...FSSILFLTGILVFSGSLYILAIAGIKILGAITPVGGIAFLIAWGLLSFYAILSKAD*
[29]    90 MALYYLCNEAPPQTLALAFLIGDSKVCGSMLGVLEGKRREYVHQLMAEQKEAEIAKKESAVQGLLIIAEGLITRRLIEKKGKFYYGTKR*
[30]   535 MVRVLVNLKEEFLLSTEMKEYENLFTEQIRRELKSFTEYRIAYQEIWTLRNVLSVLQ...DFSSLLNWLKRNVHWKGKFYSAKDLIRSATGSDPDSSYLIQYLEKKLIELESTDHG*
[31]    91 MRLFTCKKYNFLIEKNLPSFPASSPHHPKSGWGRACFTEDLSEFQQIYLRIQVLVGKVVDSQSRIPARLCSEASHFGLGTSCSSEAFLRS*
> write.csv(aa)
"","x"
"1","MKKKILLLGSGELGKEFVIAAQRLGQYVIAVDSYDGAPAMQVAHEKEIINMLDGNLLDQVVEKYKPDLIVPEIEAIRTERFYEYEKQGYQIVPSAKAANFTMNRKSIRDLAAKDLKLLTAKYLYASS
EEELIKATEILGFPCVVKPLMSSSGKGQSVIKSPKDISKAWETSQTKGRTSTTEIIVEEFISFKSEITLLTVTQKNGKTLFCPPIGHRQERGDYQESWQPAETSEVQLKEAQRMAGAVTKELTGFGIWGVEF
FLTEDQVYFSELSPRPHDTGMVTLAGTQNFNEFELHIRAILGIPISEITQERKGASAVILASTDGKIPEIKGLDIASEMPESDFRIFGKPITRPYRRMGVALSYSAKGEGTSSLRKRAILLASKIKVD*"
"2","MALDKVTSPLSCDLSVRQPGKRKRKIRFLIVSRILSVSIAIRMEEIKSSAIISANIAVLGGGPMGIYLSAYLSPKAEEIFLWYDDRKKAAKIEKERIATVLEDSIALPENVRVKSEFDFLKSGSWIL
VIAVPSRLMEGILDELIKILDKNSSHYVFAFTKGLLSISTRKKTNCITYSEYIYKLSVTGELKNVEYTAVNGPNILGELKRGHHSFYCLASSGTQSIEIFETLFGDSRSHTKTYEDLMGLEIFGVMKNPIAI
ACGIASGIPECGSNFEGELISLGYSEIIALLETLGIPTKPVQEYGLADLIASCTSRYSRNKAYGHRFIHKLISGEDRPNLIERIELFFNPAEFIQKEVSQSESHVEGAFALASIISLAEEKNVDIPLYSILF
QILTRRVSPTELIRFVSKSTSDEVRHISKTTTKRSGLGMASGKKFQEALSKNVLRHINSQPGMTDRIVKQSSLLVKSLEKRYKEAEESKDITDLLQIPKEIQLWNEVENKFQKTGNKDLTRILDFYVSEIAD
DYKPFLRDTLINLIIPIRYILNGFKSGTGLPKIGGCVKEVKALASRYDILYTPTHRSHLDSIEVAFGLKWLGLPVPRYAADKKVMATPGLANVLKSLGAYMVDRKRNRNILYLECLTQYSTMMLEAGIPTLV
YPEGTRSRTGGILPIKTGILSTSVEAYKHTGSEVIVVPIVLSYENVPEDEEFCGKDKKSGFKDFFYKRKEVYMDLCEPIPVSRYIREEDPVGMIGFEITQGWRKYRRILPNQLVARLIVETGTEVKMGELRN
LIKETILTKKGNYLIRDSDEILKQGLKILKQKKIVFLENENVKILDHNLIQYYANMCTDESS*"
"3","MTLSSATKKYNGFFLFFLFLSYLLISSINCATYFQNRKNDLKDIFTLGVETPGYGAGLRIGPLAAGFVFQGGESTPGKRDLGKGYGLRSGYFGLYRSQQLIFGILGSDTFFPLEAETQTFETEETSE
TISSETAEEFKSLENSKTPGDLVPEFLNERYNIKSQKLRYLSFYNIPTAERRKRKKEEFYKRFIEGQNFDRNDPAVQNALQAFNKRKDGYPKSFLYQIEVYLGIYVGLRIGFNPAELFDFLVGLAGLDLLED
DTAD*"
```

Gene #1:

MKKKILLLGSGELGKEFVIAAQRLGQYVIAVDSYDGAPAMQVAHEKEIINMLDGNLLDQVVEKYKPDLIVPEIEAIRTERFY
EYEKQGYQIVPSAKAANFTMNRKSIRDLAAKDLKLLTAKYLYASSEEELIKATEILGFPCVVKPLMSSSGKGQSVIKSPKDISK
AWETSQTKGRTSTTEIIVEEFISFKSEITLLTVTQKNGKTLFCPPIGHRQERGDYQESWQPAETSEVQLKEAQRMAGAVTK
ELTGFGIWGVEFFLTEDQVYFSELSPRPHDTGMVTLAGTQNFNEFELHIRAILGIPISEITQERKGASAVILASTDGKIPEIKG
LDIASEMPESDFRIFGKPITRPYRRMGVALSYSAKGEGTSSLRKRAILLASKIKVD*

Gene #2:

MALDKVTSPLSCDLSVRQPGKRKRKIRFLIVSRILSVSIAIRMEEIKSSAIISANIAVLGGGPMGIYLSAYLSPKAEEIFLWYDD
RKKAAKIEKERIATVLEDSIALPENVRVKSEFDFLKSGSWILVIAVPSRLMEGILDELIKILDKNSSHYVFAFTKGLLSISTRKKT
NCITYSEYIYKLSVTGELKNVEYTAVNGPNILGELKRGHHSFYCLASSGTQSIEIFETLFGDSRSHTKTYEDLMGLEIFGVMK
NPIAIACGIASGIPECGSNFEGELISLGYSEIIALLETLGIPTKPVQEYGLADLIASCTSRYSRNKAYGHRFIHKLISGEDRPNLIE
RIELFFNPAEFIQKEVSQSESHVEGAFALASIISLAEEKNVDIPLYSILFQILTRRVSPTELIRFVSKSTSDEVRHISKTTTKRSGL
GMASGKKFQEALSKNVLRHINSQPGMTDRIVKQSSLLVKSLEKRYKEAEESKDITDLLQIPKEIQLWNEVENKFQKTGNK
DLTRILDFYVSEIADDYKPFLRDTLINLIIPIRYILNGFKSGTGLPKIGGCVKEVKALASRYDILYTPTHRSHLDSIEVAFGLKWL
GLPVPRYAADKKVMATPGLANVLKSLGAYMVDRKRNRNILYLECLTQYSTMMLEAGIPTLVYPEGTRSRTGGILPIKTGILS
TSVEAYKHTGSEVIVVPIVLSYENVPEDEEFCGKDKKSGFKDFFYKRKEVYMDLCEPIPVSRYIREEDPVGMIGFEITQGWR
KYRRILPNQLVARLIVETGTEVKMGELRNLIKETILTKKGNYLIRDSDEILKQGLKILKQKKIVFLENENVKILDHNLIQYYAN
MCTDESS*

Gene #3:

MTLSSATKKYNGFFLFFLFLSYLLISSINCATYFQNRKNDLKDIFTLGVETPGYGAGLRIGPLAAGFVFQGGESTPGKRDLGK
GYGLRSGYFGLYRSQQLIFGILGSDTFFPLEAETQTFETEETSETISSETAEEFKSLENSKTPGDLVPEFLNERYNIKSQKLRYLS
FYNIPTAERRKRKKEEFYKRFIEGQNFDRNDPAVQNALQAFNKRKDGYPKSFLYQIEVYLGIYVGLRIGFNPAELFDFLVGL
AGLDLLEDDTAD*

Q6. Proposed gene names and functions for the first three genes/proteins.

Gene #1: formate-dependent phosphoribosylglycinamide formyltransferase [Leptospira alexanderi] or PurT is an enzyme involved in the de novo purine biosynthesis pathway by catalyzing the transfer of a formyl group from formate to GAR, which then forms FGAR.

Gene #2: NAD-dependent glycerol-3-phosphate dehydrogenase C-terminal domain protein [Leptospira borgpetersenii serovar Pomona str. 200901868] or G3PDH is an enzyme involved in the glycerophospholipid metabolism pathway by catalyzing the reversible redox conversion between DHAP and G3P for lipid biosynthesis and energy metabolism.

Gene #3: LIC13411 family adhesin [Leptospira alexanderi] is a protein that enables a bacterium to adhere to host cells, particularly to human endothelial cells.

Q7. Non-coding gene sequence, location, length, and type.

```
> #Obtain pre-built models of non-coding RNAs commonly found in bacteria
> data(NonCodingRNA_Bacteria)
> x <- NonCodingRNA_Bacteria
>
> #Search for these models in A2contig
> rnas <- FindNonCoding(x, A2contig)
  |================================================| 100%

Time difference of 3.05 secs
> rnas
Genes object of size 1 specifying:
1 non-coding RNA of 73 nucleotides.

  Index Strand Begin   End TotalScore Gene
1     1      0 26511 26583      52.11   -8
>
> #Match Gene value with corresponding model
> annotations <- attr(rnas, "annotations")
> m <- match(rnas[, "Gene"], annotations)
> sort(table(names(annotations)[m]))
tRNA-Gly
       1
>
> #Determine sequence of non-coding gene
> gene_sequence <- subseq(A2contig, start = 26511, end = 26583)
> write.csv(gene_sequence)
"","x"
"JQGU01000058.1 Leptospira alexanderi strain 56643 Contig058, whole genome shotgun sequence","GCGGGCATGGTGTAATGGCTAGCACTGTAGCCTTCCAA
GCTTCCAGTGAGGGTTCGAGTCCCTCTGCCCGCAA"
```

Sequence:
GCGGGCATGGTGTAATGGCTAGCACTGTAGCCTTCCAAGCTTCCAGTGAGGGTTCGAGTCCCTCTGCCCGCAA

Location: 26511-26583 Length: 73 bp Type: tRNA-Gly

Q8. Whether there is overlap between coding and non-coding genes. No

Q9. If there is an overlap, can both models be correct?

Yes, since it is possible due to either biological or technical reasons. Regarding biological, the protein-coding genes would lie in regions that would code for functioning proteins, while non-coding could lie nearby that act as regulation factors, which may have overlapping but alternative reading frames. Other reasons may include that non-coding RNAs are transcribed in the antisense direction of a coding gene.

Q10. A brief explanation of how you would resolve the question of gene overlap.

There are various solutions, whether that be from additional analyses or laboratory methods. Solutions may include considering alternative splicing or presence of overlapping transcription start sites, performing an experiment that silences the non-coding RNA and determine whether it affects gene regulation (i.e., confirms that it is a regulatory element and thus has overlapping ORFs), etc.