**Lab sheet1: Information retrieval and sequence analysis**

**Q1**

**COX-2 (prostaglandin H2 synthase-2 (PTGS2)) gene**

COX-2 has been thoroughly studied because of its role in prostaglandin synthesis. Prostaglandins have a wide range of roles in our body from aiding in digestion to propagating pain and inflammation.

Aspirin is a general inhibitor of prostaglandin synthesis and therefore, helps reduce pain.

However, aspirin also inhibits the synthesis of prostaglandins that aid in digestion. Therefore, aspirin is a poor choice for pain and inflammation management for those with ulcers or other digestion problems.

Recent advances in targeting specific prostaglandin-synthesizing enzymes have lead to the development of Celebrex, which is marketed as an arthritis therapy. Celebrex is a potent and specific inhibitor of COX-2. Celebrex is considered specific because it doesn’t inhibit COX-1, which is involved in synthesizing prostaglandins that aid in digestion.

This is a remarkable accomplishment given the great similarity between COX-1 and COX-2.

This achievement has paved the way for developing new therapies that bind more specifically to their target and therefore have fewer side effects. Understanding the enzyme structures of COX-1 and COX-2 helped researchers develop a drug that would only bind and inhibit COX-2. Many of the types of information and tools used by researchers for these types of studies are freely available on the web .

GenBank, SwissProt, Sequence Manipulation suite are some of the websites.

1. Access the entries for Human PTGS1 and PTGS2 in the “Gene” database at the NCBI (<https://www.ncbi.nlm.nih.gov/>) Website.
   1. PTGS1 and PTGS2 are isozymes. Isozymes catalyze the same reaction but are separate genes. What types of reactions do PTGS enzymes catalyze? Also, what pathway are these enzymes a part of?
   2. How is the expression of PTGS1 and PTGS2 different?
   3. Which isozyme ( PTGS1 or PTGS2 ) is required to inhibit inflammation?
   4. The drug Celebrex selectively inhibits PTGS2 while aspirin and other NSAID’s inhibit both PTGS1 and PTGS2 in the same way. Why do you think researchers wanted to discover a selective inhibitor to PTGS2?
   5. Describe how studying 3-D structures of PTGS1 and PTGS2 could help researchers design a drug that binds to PTGS1, but not to PTGS2.
2. Considering the Homo sapiens PTGS2 gene entry in NCBI gene <https://www.ncbi.nlm.nih.gov/gene/> database,
   1. What is the gene name?
   2. What is the GeneID number?
   3. Where in the human genome is this gene located?
   4. What is the RefSeq accession number for the mRNA sequence of H o m o sa pie n s prostaglandin-endoperoxide synthase 2?
   5. Download the prostaglandin-endoperoxide synthase 2 Reference mRNA sequence in “FASTA” format.
   6. What is the RefSeq accession number for the H o m o sa pie n s PTGS2 protein sequence? Download the sequence in “FASTA” format.
3. Search for the UniProt entry for PTGS2 in Expasy <https://www.expasy.org/>website.
   1. What are the alternate names for this protein.
   2. What types of drugs target this protein?
   3. What amino acid is acetylated by aspirin (amino acid type)?
4. Translate the mRNA sequence of PTGS2 into Protein. Use “Translate “ tool in ExPASy. Explain the output.

Readings:

<http://www.aspree.org/AUS/aspree-content/aspirin/how-aspirin-works.aspx>

**Q2. Python Exercises**

1. Below shows some files with embedded sample names:

lane1\_NewCode\_L001\_R1.fastq.gz

lane1\_NoIndex\_L001\_R1.fastq.gz

lane1\_NoIndex\_L001\_R2.fastq.gz

pipeline\_processing\_output.log

lane7027\_ACTGAT\_JH25\_L001\_R1.fastq.gz

lane7027\_ACTTGA\_E30\_1\_2\_Hap4\_24h\_L001\_R1.fastq.gz

lane7027\_AGTTCC\_JH14\_L001\_R1.fastq.gz

lane7027\_CGGAAT\_JH37\_L001\_R1.fastq.gz

lane7027\_GCCAAT\_E30\_1\_2l\_Hap4\_log\_L001\_R1.fastq.gz

lane7127\_GGCTAC\_E30\_1\_4\_Hap4\_48h\_L001\_R1.fastq.gz

Write a Python code to extract the sample name from these files ignoring any files which do not match the format given below.

The format is:

1. Written lane number
2. Barcode
3. Sample name
4. Numeric lane number (starting with L)
5. Read number (R1/2/3/4)
6. File extension

Eg. Lane8127\_GCCAAT\_S30\_1\_2l\_Hap4\_log\_L001\_R1.fastq.gz the sample name would be,

S30\_1\_2l\_Hap4\_log

1. Create a FASTA file by obtaining 10 Dengue 1- Envelop gene DNA sequences from NCBI. Write a Python-program that reads the FASTA file, cleans up the header line to have only Accession number & gene-name and print headers and sequences to standard output as multi-FASTA-file again.
2. A chart with different colored text

   Description automatically generated with medium confidenceWrite a Python program to search the DNA Sequence for the presence of one of the following Transcription Factor Binding Sites(TFBS) with ambiguity codes. Search for all the positions in the sequence where TFBS is located.

|  |  |
| --- | --- |
| **Transcription Factor** | **Consensus Sequence** |
| RUNX1 | BHTGTGGTYW |
| TGIF1 | WGACAGB |
| IKZF1 | BTGGGARD |

The sequence is shown below.

>search\_seq

GACACCTCAGTACTAGGATGNNNNNNTATCAGCCTGAACTAGCAGGCCTGGTTCCAAATT

TTTTTATCAACACTCGTAGGGGGATTATCCTAGAGGGGGTCTGGGATTTCTTTGACATCA

GAGTATTTTTGCCTTGCTCCTTCACAATTTGGGAACAAATAATTTAGTGGTTATTAACCC

TGGCTACGCACTGGAAACTTTAAAAATAATGCTGGTATGAAATTTACACAGAGTATCGTG

AAAATTTTCACTGAGTACCATGTGGTTATACATTGGATAAGGCTCCAGGAAGCAGCTACT

GGAAGACAGCCATGCCAAGAGTGGTTAGTGGTTGGAATTTTGGCAAGTCAGTTTTAGTCT

GCCTTATCAAATACATGGGCATACAGATAAATCCTTAGATGGCTCTCCTACTTACTGAAA

CATTTTCTATCTATCTATCTATCTATCTATCTATTTGGGAAGCTATCTATCTATCTATCA

TTTATTTAAGGTAGTCTCTATCTGCCTCTGTCTCTGTCTGTCTCTGTGTCTCTGTGTCTG

TCTGCTCTCTCTCTCTCTCTGTGGGAATCTCTCTCTGTGTGTGTGTGTGTATGTGTGTGT

GTGTGTGTGTGGTGTGCATGAACATGAGTAAAATCCATAAGGAAACTTTCAGAGTTGGTC

CTCTCCTTATATCAAATGGATCCAGGAATTAAACTCAGGTTCAATTCTTGGTGCCTTTAC

TAGTTGAGCCATCTCACTGGCTCTTCATCATCTTTAGAATAAACTCACTTTATTACACAC

ACACACACACACACAACCTGGGAGTACACACACACACACAACCAAAGCCCCAACGGAAAA

CTACAATATTATAATGAATACACAGGTTCTCAACATAGTCTCTGCCACGCTTGCAGACAA

AGATGAGTAGAAGTAGAAAGAACCAGGGAAACGTGGAGCAAGTCAGAAGGAATAACAGTC

AGAAGGAATAACAGTCAGAAGGAATAACAGTCAGAAGGAGTAACAGTCAGAAGGAATAGC

AGTCAGAAGGAATAACAGTCAGAAGACAGCACAGTCAGAAGGAATAACAGTCAGAAGGAA

TAACAGTCAGAAGGAATAACAGTCAGAAGGAATAACAGTCAGAAGGAATAGCAGTCAGAA

GGAATAACAGTCAGAAGGAATAACAGTCAGAAGGAATAACAGTCAAAGAAATAGCAGTCA

GAAGGAATAGCAGTCAGAAGGAATAACAGTCAAAGGAGCAGTCAGAAGGAGTAACAGTCA

GAAGGAATAACAGTCAGAAGGAATAACAGTCAAAGGAATAGCAGTCAGAAGGAGTAACAG

TCAGAGCAAACACAGAGATGACAAAGGCAATGGGGTCAGAGACTTCACCACTCTCCAAGA

**Q3 – Biopython**

**Biopython Tutorial and Cookbook** [**https://biopython.org/DIST/docs/tutorial/Tutorial.html#sec2**](https://biopython.org/DIST/docs/tutorial/Tutorial.html#sec2)

1. Write a Biopython program that asks the user to input a DNA-sequence and then translates the sequence to protein sequence.
2. Write a Biopython program that will find all articles related to Alzheimer’s in PubMed. Print the total number of articles available and the authors.
3. Write a Biopython-program that finds CpG-islands from a given DNA-sequence.