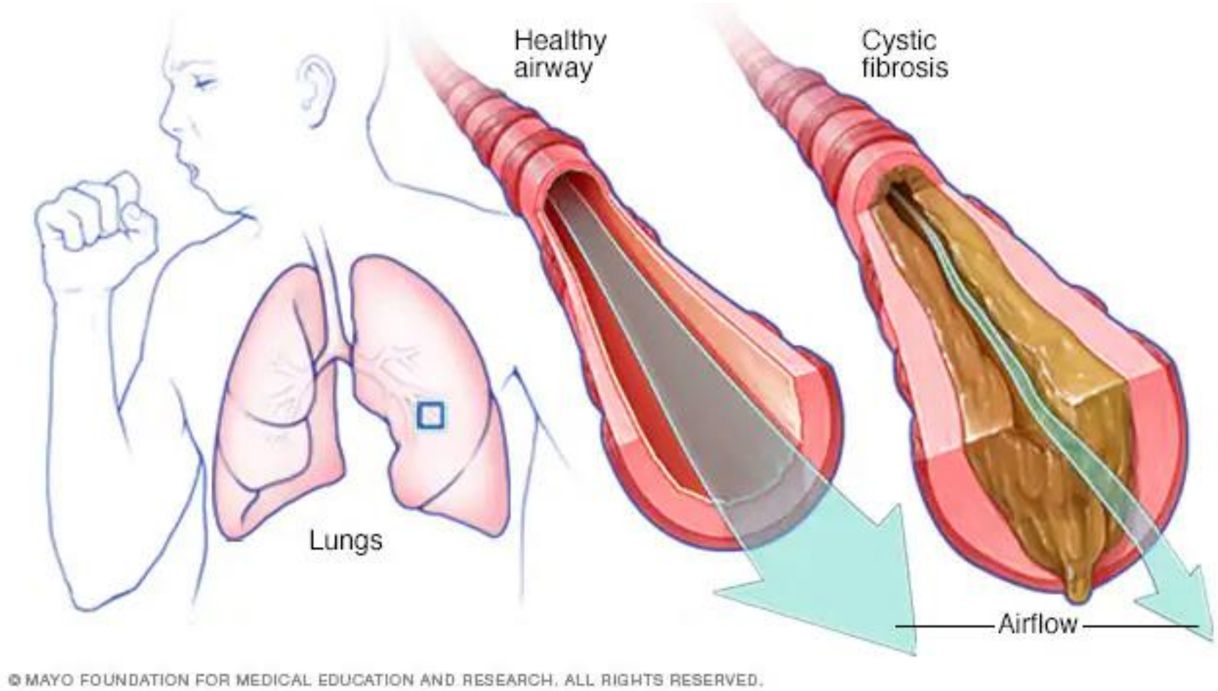
Project – Main Assignment

Practical information

- **Time**: Every afternoon, from ~15:00 to 17:00
- **Duration**: Throughout the entire week
- Apply your knowledge to tackle a real-world problem at a larger scale than the exercises
- Work on your own or in groups
- TAs available for questions
- **Not mandatory** but highly recommended
- Solutions will be published on **Friday** after the lectures

Background


Healthy airway
Cystic fibrosis
Lungs
Airflow
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

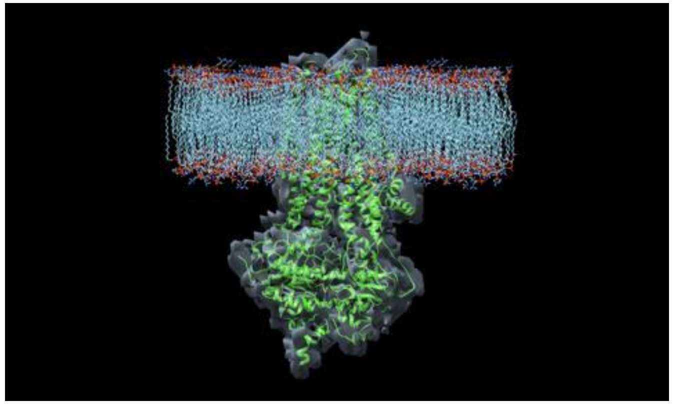
source: mayoclinic.org

Cystic fibrosis (CF)

- Genetic inherited disease
- Produces thick and sticky mucus in organs, including lungs and the pancreas
- Clogs the airways of patients and makes them difficult to breathe
- No cure available but only symptom management, such as airway clearance

Genomic facts of Cystic Fibrosis

- CF is caused by mutations in Cystic Fibrosis Transmembrane Conductance Regulator (CFTR).
- The CFTR protein is an ion channel protein, acting like gates in a cell membrane that control the traffic of molecules through the membrane.
- For regular people, CFTR acts as a gate for chloride ions. When chloride leaves the cell, it carries water with it, which makes mucus less thick.
- For patients with CF, gene mutations in CFTR prevent this functionality, causing the mucus stays sticky and thick.



source: cff.org

More about the CFTR gene

- CFTR gene is located on chromosome 7 of the human genome
- Over 1,500 mutations known to cause CF
- One type of mutations
  - Non-synonymous (with amino acid changing) mutations that generate a premature termination codon (PTC), that further leads to a truncated CFTR protein (shortened length).

Goal of the project

Write a python program that:

- Extract the correct CFTR transcript from the human genome
- Translate it into its corresponding amino acid sequence
- Determine if one or more patients have a premature stop codon

You will be guided step by step towards the final goal

Data

- Human reference genome
  - Chromosome 7 in fasta format
  - Gene annotations in GTF (Gene Transfer Format) format
- Genome sequencing data from five patients
  - Chromosome 7 in fasta format

Fasta format

```
>MT dna:chromosome chromosome:GRCh38:MT:1:16569:1 REF
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
```

GTF format

- GTF stands for Gene transfer format
- Holds information about gene structure
- Tab-delimited

Columns of GTF file

<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes]

1. **<seqname>**: The name of the sequence (typically a chromosome).
2. **<source>**: The source of the annotation (e.g., ENSEMBL).
3. **<feature>**: The type of feature (e.g., gene, transcript, exon).
4. **<start>**: The starting position of the feature in the sequence.
5. **<end>**: The ending position of the feature in the sequence.
6. **<score>**: A score between 0 and 1000, or . if not applicable (indicating the reliability of the annotation).
7. **<strand>**: The strand on which the feature is located (`+` for the forward strand, `-` for the reverse strand).
8. **<frame>**: The reading frame, one of '0', '1' or '2', or `.` if not applicable.
9. **[attribute]**: A list of key-value pairs providing additional information about the feature.

Attribute of GTF

- A semicolon (;)-separated list of key-value pairs
- For each key-value pair, key is one word, and the value is quoted by double quotes, which may contain multiple words
- A key can be repeated multiple times.

Some attributes (always semi-colon separated key-value pairs):

- gene_id: The stable identifier for the gene
- gene_version: The stable identifier version for the gene
- gene_name: The official symbol of this gene
- gene_source: The annotation source for this gene
- transcript_id: The stable identifier for this transcript
- transcript_name: The symbold for this transcript derived from the gene name
- exon_id: The stable identifier for this exon

GTF example

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame>
[attributes]


1 havana gene 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "5";
gene_name "DDX11L1"; gene_source "havana"; gene_biotype
"transcribed_unprocessed_pseudogene";

1 havana transcript 11869 14409 . + . gene_id "ENSG00000223972"; gene_version
"5"; transcript_id "ENST00000456328"; transcript_version "2"; gene_name
"DDX11L1"; gene_source "havana"; gene_biotype
"transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-202";
transcript_source "havana"; transcript_biotype "processed_transcript"; tag
"basic"; transcript_support_level "1";

1 havana exon 11869 12227 . + . gene_id "ENSG00000223972"; gene_version "5";
transcript_id "ENST00000456328"; transcript_version "2"; exon_number "1";
gene_name "DDX11L1"; gene_source "havana"; gene_biotype
"transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-202";
transcript_source "havana"; transcript_biotype "processed_transcript";
exon_id "ENSE00002234944"; exon_version "1"; tag "basic";
transcript_support_level "1";
```

Getting started

- Create a folder called project
- Download and extract the project files in this folder
- Work with Jupyter or any text editor (e.g., Spyder, Sublime)
- Ask questions if something is unclear
- Speak to a TA or discuss with your neighbour
- Or use the discussion section in Canvas
- Find out more info at Canvas -> Modules -> Project