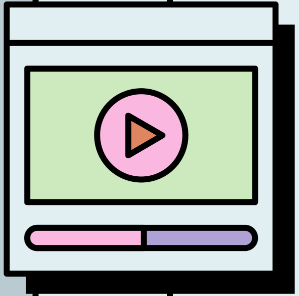
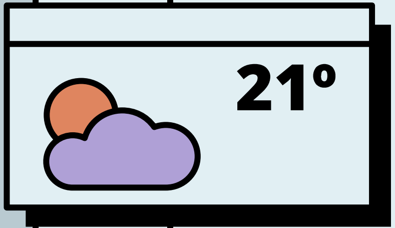
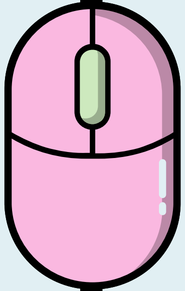




LIFT Bioinformatics Course



Gunica Sharma, Sovie Prasad Shekhar and Brenda Cruz

Table of contents

01 Module Design

02 Module Highlights

03 Testing & Feedback

04 Looking Forward

Course Methodology

Engaging

Wanted a course where students feel engaged with the material that's why we choose to do project based hands on modules

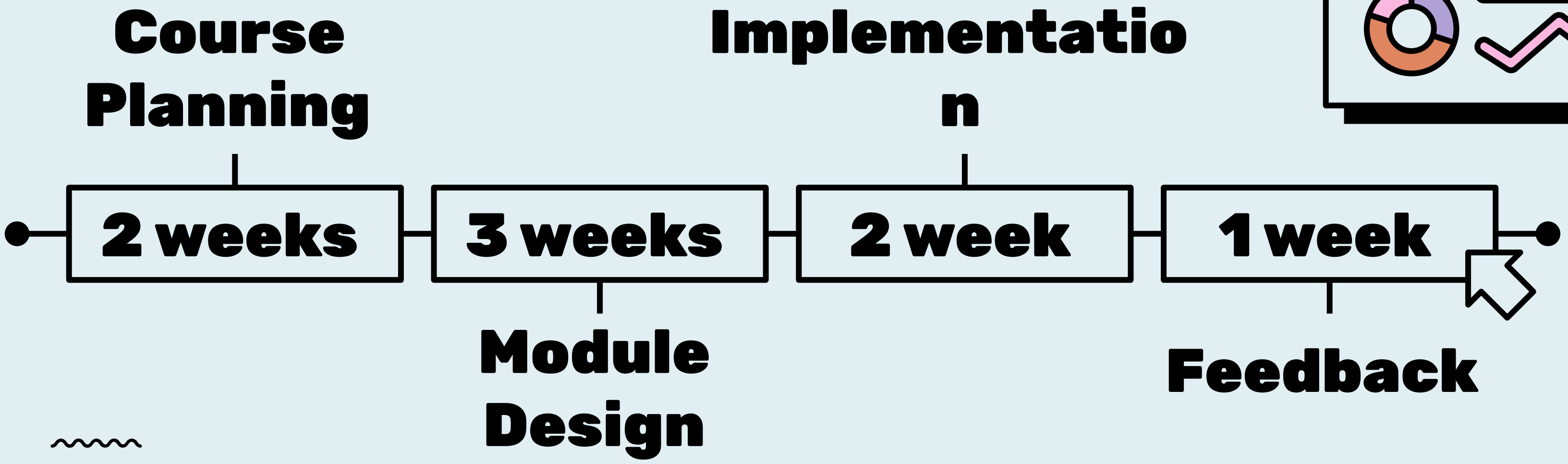
Short

Needed the course to be short enough to finish in a few weeks time we wanted it to be easily integrated in with their busy lives

Meaningful

Want students to walk away with a knowledge of bioinformatics and seek out other opportunities in the field want this course to be their stepping stone

Our Timeline



Course Planning

2 weeks

3 weeks

Module Design

Implementatio
n

2 week

1 week

Feedback

Module Framework

BLAST

Genomic Analysis

Protein Structures

UNIX Commands

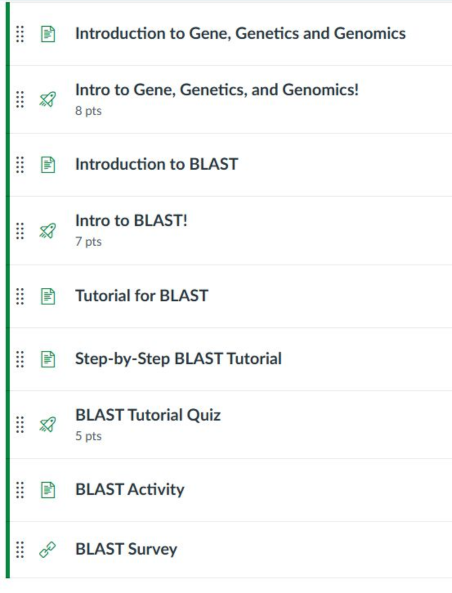
01

BLAST

Exploring and Understanding BLAST

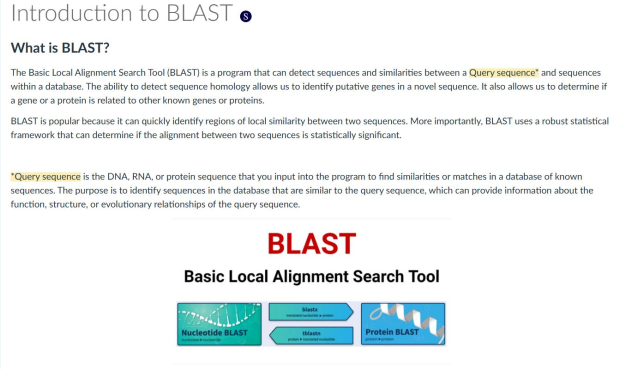
BLAST Module

Format



- Introduction to Gene, Genetics and Genomics
- Intro to Gene, Genetics, and Genomics! 8 pts
- Introduction to BLAST
- Intro to BLAST! 7 pts
- Tutorial for BLAST
- Step-by-Step BLAST Tutorial
- BLAST Tutorial Quiz 5 pts
- BLAST Activity
- BLAST Survey

Introduction & Applications



Introduction to BLAST

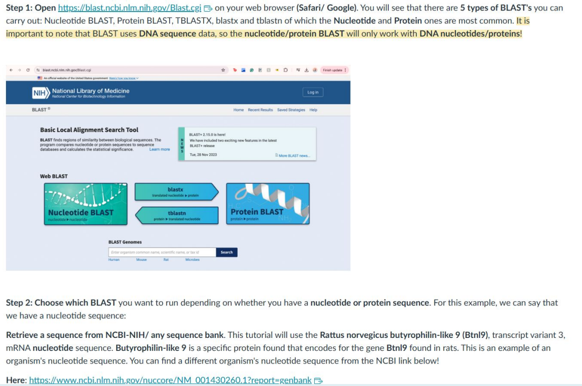
What is BLAST?

The Basic Local Alignment Search Tool (BLAST) is a program that can detect sequences and similarities between a Query sequence* and sequences within a database. The ability to detect sequence homology allows us to identify putative genes in a novel sequence. It also allows us to determine if a gene or a protein is related to other known genes or proteins.

BLAST is popular because it can quickly identify regions of local similarity between two sequences. More importantly, BLAST uses a robust statistical framework that can determine if the alignment between two sequences is statistically significant.

*Query sequence is the DNA, RNA, or protein sequence that you input into the program to find similarities or matches in a database of known sequences. The purpose is to identify sequences in the database that are similar to the query sequence, which can provide information about the function, structure, or evolutionary relationships of the query sequence.

Step-by-step tutorial



Step 1: Open https://blast.ncbi.nlm.nih.gov/Blast.cgi on your web browser **(Safari/ Google)**. You will see that there are **5 types of BLAST's** you can carry out: Nucleotide BLAST, Protein BLAST, TBLASTX, blastx and tblastn of which the **Nucleotide** and **Protein** ones are most common. It is important to note that BLAST uses **DNA sequence** data, so the **nucleotide/protein BLAST** will only work with **DNA nucleotides/proteins**!

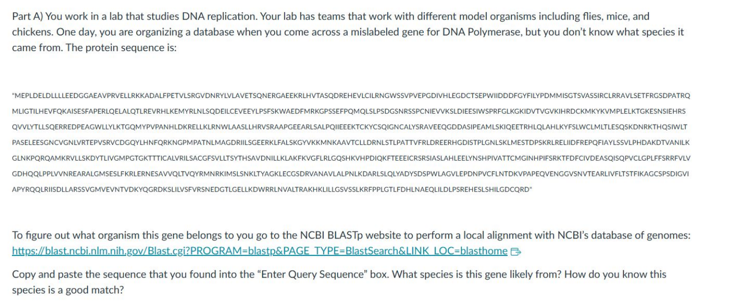
blast.ncbi.nlm.nih.gov/Blast.cgi
Finish update
An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST®
Home
Recent Results
Saved Strategies
Help
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
Learn more
NEWS
BLAST+ 2.15.0 is here!
We have included two exciting new features in the latest BLAST+ release
Tue, 28 Nov 2023
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▸ nucleotide
blastx
translated nucleotide ▸ protein
tblastn
protein ▸ translated nucleotide
Protein BLAST
protein ▸ protein
BLAST Genomes
Enter organism common name, scientific name, or tax id
Search
Human
Mouse
Rat
Microbes

Step 2: Choose which BLAST you want to run depending on whether you have a **nucleotide or protein sequence**. For this example, we can say that we have a nucleotide sequence:

Retrieve a sequence from NCBI-NIH/ any sequence bank. This tutorial will use the **Rattus norvegicus butyrophilin-like 9 (Btnl9)**, transcript variant 3, mRNA **nucleotide** sequence. **Butyrophilin-like 9** is a specific protein found that encodes for the gene **Btnl9** found in rats. This is an example of an organism's nucleotide sequence. You can find a different organism's nucleotide sequence from the NCBI link below!

Here: https://www.ncbi.nlm.nih.gov/nuccore/NM_001430260.1?report=genbank

Do-it-yourself Activity



Part A) You work in a lab that studies DNA replication. Your lab has teams that work with different model organisms including flies, mice, and chickens. One day, you are organizing a database when you come across a mislabeled gene for DNA Polymerase, but you don't know what species it came from. The protein sequence is:

```
"MEPLDELDLLLLEEDGGAEAVPRVELLRKKADALFPETVLSRGVDNRYLVLAVETSQNERGAEEKRLHVTASQDREHEVLCILRNGWSSVPVEPGDIVHLEGDCTSEPWIIDDDFGYFILYPDMMISGTSVASSIRCLRRAVLSETFRGSDPATRQ
MLIGTILHEVFQKAISESFAPERLQELALQTLREVRHLKEMYRLNLSQDEILCEVEEYLPSFSKWAEDFMRKGPSSEFPQMQLSLPSDGSNRSSPCNIEVVKSLDIEESIWSPRFGLKGKIDVTVGVKIHRDCKMKYKVMPLELKTGKESNSIEHRS
QVVLYTLLSQERREDPEAGWLLYLKTGQMYPVPANHLDKRELLKLRNWLAASLLHRVSRAAPGEEARLSALPQIIEEEKTCKYCSQIGNCALYSRAVEEQGDDASIPEAMLSKIQEETRHLQLAHLKYFSLWCLMLTLESQSKDNRKTHQSIWLT
PASELEESGNCVGNLVRTEPVSRVCDGQYLHNFQRKNGPMPATNLMAGDRIILSGEERKLFALSKGYVKKMNKAAVTCLLDRNLSTLPATTVFRLDREERHGDISTPLGNLSKLMESTDPSKRLRELIIDFREPQFIAYLSSVLPHDAKDTVANILK
GLNKPQRQAMKRVLLSKDYTLIVGMPGTGKTTTICALVRILSACGFSVLLTSYTHSAVDNILLKLAKFKVGFLRLGQSHKVHPDIQKFTEEEICRSRSIASLAHLEELYNSHPIVATTCMGINHPIFSRKTFDFCIVDEASQISQPVCLGPLFFSRRFVLV
GDHQQLPPLVVNREARALGMSESLFKRLERNESAVVQLTVQYRMNRKIMSLSNKLTYAGKLECGSDRVANAVLALPNLKDARLSLQLYADYSDSPWLAGVLEPDNPVCFLNTDKVPAPEQVENGGVSNVTEARLIVFLTSTFIKAGCSPSDIGVI
APYRQQLRIISDLLARSSVGMVEVNTVDKYQGRDKSLILVSFVRSNEDGTLGELLKDWRRLNVALTRAKHKLILLGSVSSLKRFPPLGTLFDHLNAEQLILDLPSREHESLSHILGDCQRD"
```

To figure out what organism this gene belongs to you go to the NCBI BLASTp website to perform a local alignment with NCBI's database of genomes: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Copy and paste the sequence that you found into the "Enter Query Sequence" box. What species is this gene likely from? How do you know this species is a good match?

02

Genomic Data Analysis

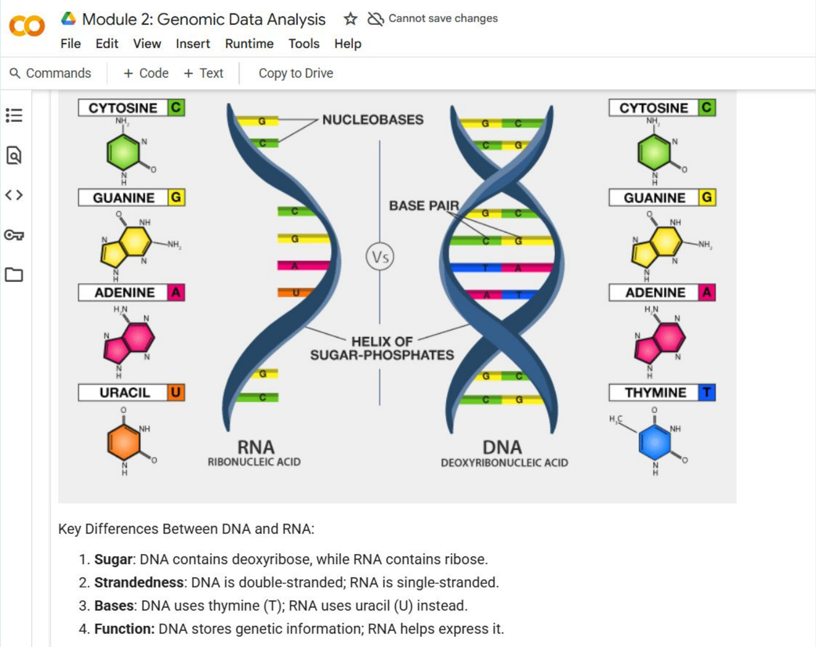
Working with DNA & RNA sequences to analyze them in different ways

Learning Objectives

1. Understanding DNA & RNA sequences
2. Basic sequence operations: transcription (DNA → RNA) and translation (RNA -> protein)
3. What a Single Nucleotide Polymorphism is and how to find one by comparing reference and mutated sequences
4. Visualize GC content to compare sequences

Module Format

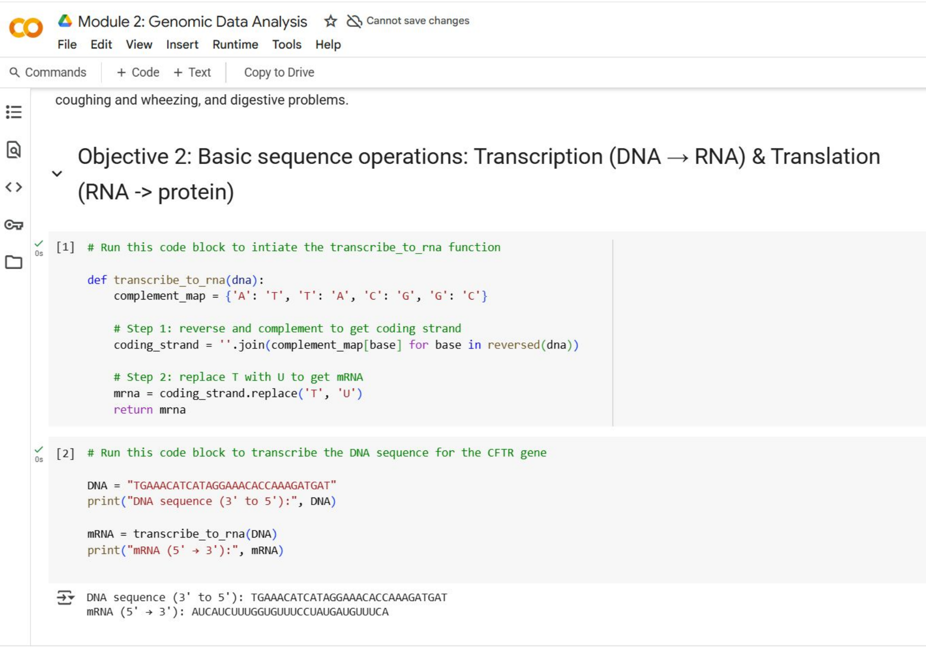
Written explanations & images



Key Differences Between DNA and RNA:

1. **Sugar**: DNA contains deoxyribose, while RNA contains ribose.
2. **Strandedness**: DNA is double-stranded; RNA is single-stranded.
3. **Bases**: DNA uses thymine (T); RNA uses uracil (U) instead.
4. **Function:** DNA stores genetic information; RNA helps express it.

Guided coding & testing out



coughing and wheezing, and digestive problems.

Objective 2: Basic sequence operations: Transcription (DNA → RNA) & Translation (RNA -> protein)

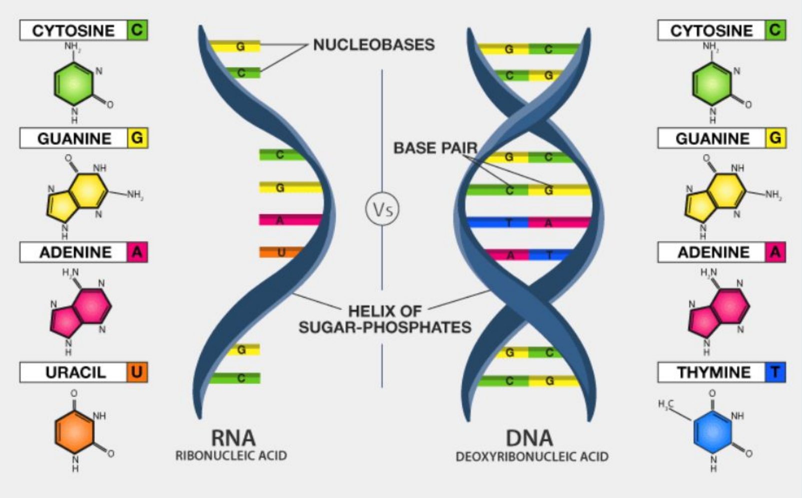
```
[1] # Run this code block to intiate the transcribe_to_rna function

    def transcribe_to_rna(dna):
        complement_map = {'A': 'T', 'T': 'A', 'C': 'G', 'G': 'C'}

        # Step 1: reverse and complement to get coding strand
        coding_strand = ''.join(complement_map[base] for base in reversed(dna))

        # Step 2: replace T with U to get mRNA
        mrna = coding_strand.replace('T', 'U')
        return mrna
```

```
[2] # Run this code block to transcribe the DNA sequence for the CFTR gene

    DNA = "TGAAACATCATAGGAAACACCAAAGATGAT"
    print("DNA sequence (3' to 5'):", DNA)

    mRNA = transcribe_to_rna(DNA)
    print("mRNA (5' → 3'):", mRNA)
```

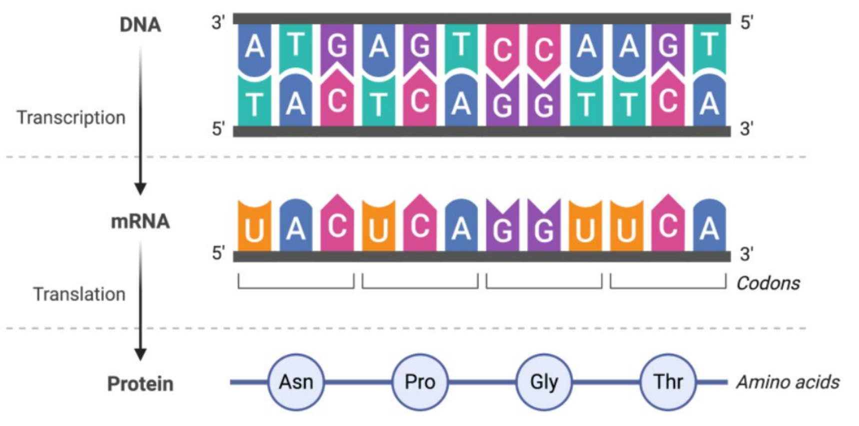
```
DNA sequence (3' to 5'): TGAAACATCATAGGAAACACCAAAGATGAT
mRNA (5' → 3'): AUCAUCUUUGGUGUUUCCUAUGAUGUUUCA
```

RNA & DNA Background

Difference between RNA and DNA



Central Dogma (DNA→protein)



Case Study

Walk-through

CFTR gene (Cystic Fibrosis Transmembrane Conductance Regulator)

- Chloride ions can't move across cell membranes normally.
- Mucus builds up, especially in the lungs and pancreas.
- Chronic lung infections, coughing and wheezing, and digestive problems

Mutated sequence: missense mutation

Mini-project Deliverable

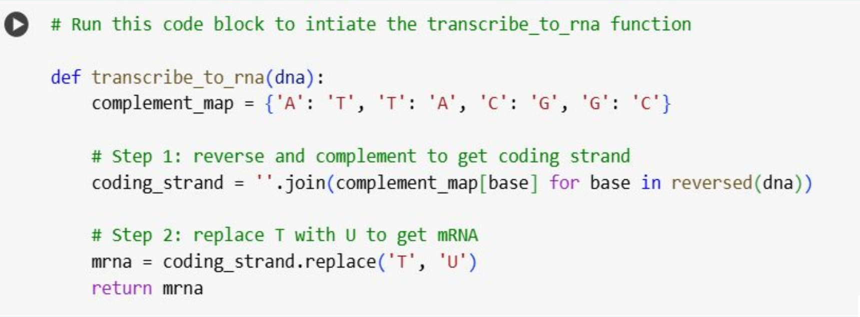
HBA1 gene encodes the alpha-globin subunit of hemoglobin, which is part of the oxygen-carrying molecule in red blood cells.

- Ineffective red blood cell production
- Anemia, fatigue
- Enlarged spleen

Mutated sequence: silent mutation

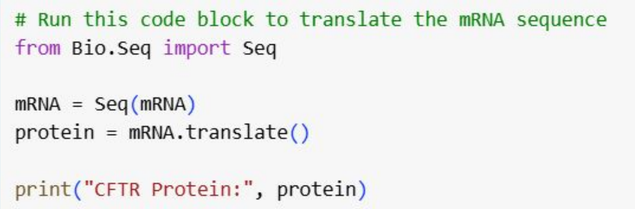
Basic Sequence Operations & SNP

Transcription - function



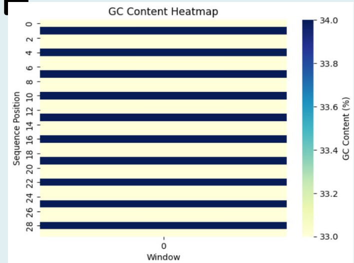
```
# Run this code block to intiate the transcribe_to_rna function

def transcribe_to_rna(dna):
    complement_map = {'A': 'T', 'T': 'A', 'C': 'G', 'G': 'C'}

    # Step 1: reverse and complement to get coding strand
    coding_strand = ''.join(complement_map[base] for base in reversed(dna))

    # Step 2: replace T with U to get mRNA
    mrna = coding_strand.replace('T', 'U')
    return mrna
```

Translation - package

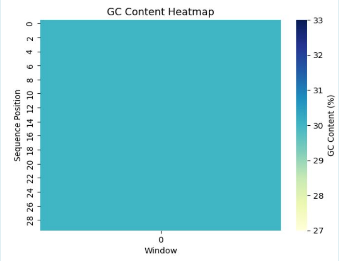


```
# Run this code block to translate the mRNA sequence
from Bio.Seq import Seq

mRNA = Seq(mRNA)
protein = mRNA.translate()

print("CFTR Protein:", protein)
```

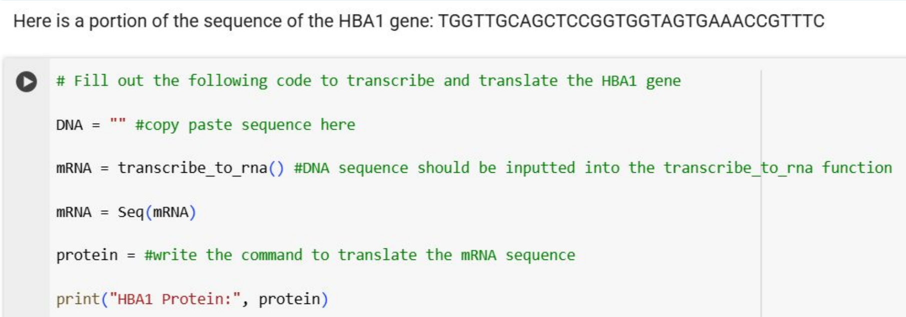
CFTR Gene Wild Type



CFTR Gene Mutated



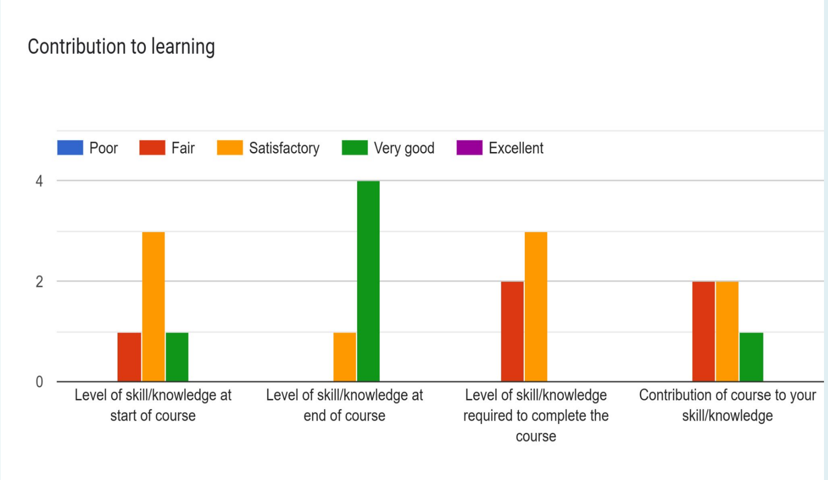
Mini-project

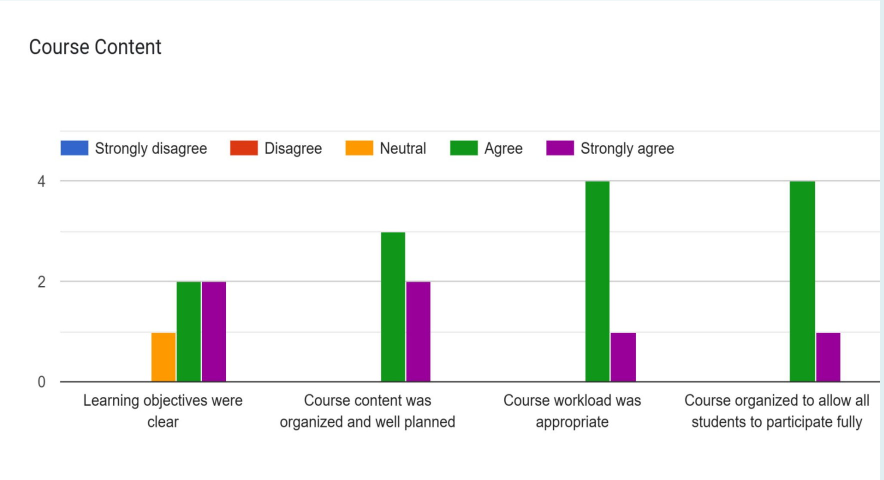


Here is a portion of the sequence of the HBA1 gene: TGGTTGCAGCTCCGGTGGTAGTGAAACCGTTTC

```
# Fill out the following code to transcribe and translate the HBA1 gene

DNA = "" #copy paste sequence here

mRNA = transcribe_to_rna() #DNA sequence should be inputted into the transcribe_to_rna function

mRNA = Seq(mRNA)

protein = #write the command to translate the mRNA sequence

print("HBA1 Protein:", protein)
```

Feedback





Most students’ knowledge went from fair/satisfactory to very good after taking course.

Learning objectives, course content, workload, and organization were well planned and clear.

Feedback

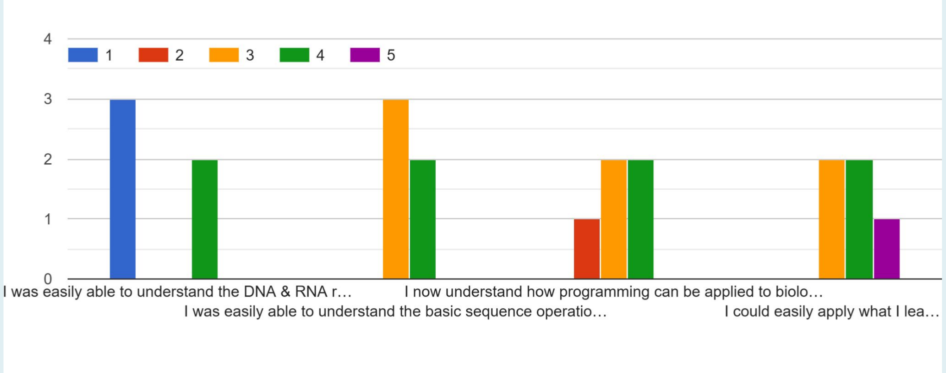
What aspects of this course were most useful?

- Explanation of connection between code and biological principles
- Practicing Python myself
- Explanation of each code block
- Mini Project

What could be improved?

- Directions for data entry
- Troubleshooting for coding parts
- More involved

4 out of 5 students took **30 minutes** and all took less than an hour to complete the module.



Most students were confident that they could apply what they learned from the course and had a fairly easy time understanding the coding format.

03

Exploring Protein Structures

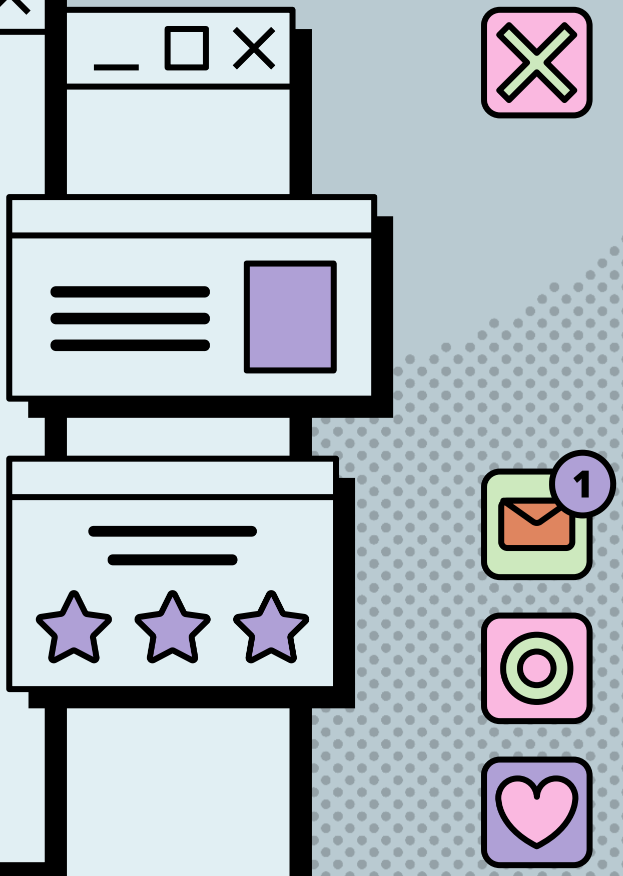
Understanding Protein Structures and Analyzing Differences between them

Learning Objectives

- Understand how to obtain protein sequences from NCBI database
- Learn how to use alpha fold to predict protein structure
- Compare protein structures to experimentally generated ones
- Identify structural features and relate them to protein function

Key Concepts Introduced

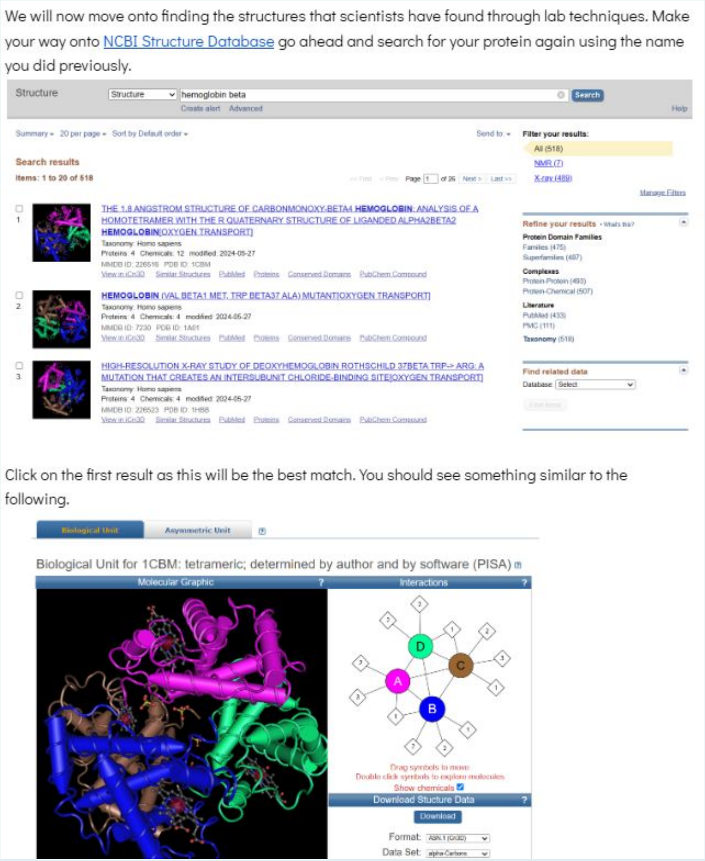
- Protein function depends on structure
- Introduction to AlphaFold: AI-based structure prediction
- Real-world protein structure data from NCBI
- Gene → Protein → Structure → Function → Disease



Module Format

Project Workflow

1. Choose a gene from a curated list (e.g., INS, TP53, GFP, HBB)
2. Get DNA/protein sequence from NCBI Gene
3. Find experimental structure using NCBI Structure Database
4. Find predicted structure using AlphaFold DB
5. Compare structures: real vs. predicted
6. Analyze: How do differences impact function?



We will now move onto finding the structures that scientists have found through lab techniques. Make your way onto NCBI Structure Database go ahead and search for your protein again using the name you did previously.

Structure: hemoglobin beta

Search results

Items: 1 to 20 of 518

1. THE 1.8 ANGSTROM STRUCTURE OF CARBONMONOXY-BETA4 **HEMOGLOBIN**: ANALYSIS OF A HOMOTETRAMER WITH THE R QUATERNARY STRUCTURE OF LIGANDED ALPHA2BETA2 **HEMOGLOBIN**[OXYGEN TRANSPORT]
   Taxonomy: Homo sapiens
   Proteins: 4 Chemicals: 12 modified: 2024-05-27
   MMDB ID: 226516 PDB ID: 1CBM
2. **HEMOGLOBIN** (VAL BETA1 MET, TRP BETA37 ALA) MUTANT[OXYGEN TRANSPORT]
   Taxonomy: Homo sapiens
   Proteins: 4 Chemicals: 4 modified: 2024-05-27
   MMDB ID: 7230 PDB ID: 1A01
3. HIGH-RESOLUTION X-RAY STUDY OF DEOXYHEMOGLOBIN ROTHSCHILD 37BETA TRP-> ARG: A MUTATION THAT CREATES AN INTERSUBUNIT CHLORIDE-BINDING SITE[OXYGEN TRANSPORT]
   Taxonomy: Homo sapiens
   Proteins: 4 Chemicals: 4 modified: 2024-05-27
   MMDB ID: 226523 PDB ID: 1HBB

Filter your results:
All (518)
NMR (7)
X-ray (489)

Refine your results
Protein Domain Families: Families (475), Superfamilies (497)
Complexes: Protein-Protein (493), Protein-Chemical (507)
Literature: PubMed (433), PMC (111)
Taxonomy (518)

Click on the first result as this will be the best match. You should see something similar to the following.

Biological Unit for 1CBM: tetrameric; determined by author and by software (PISA)

Format: ASN.1 (Cn3D)
Data Set: alpha-Carbons

A look into the module

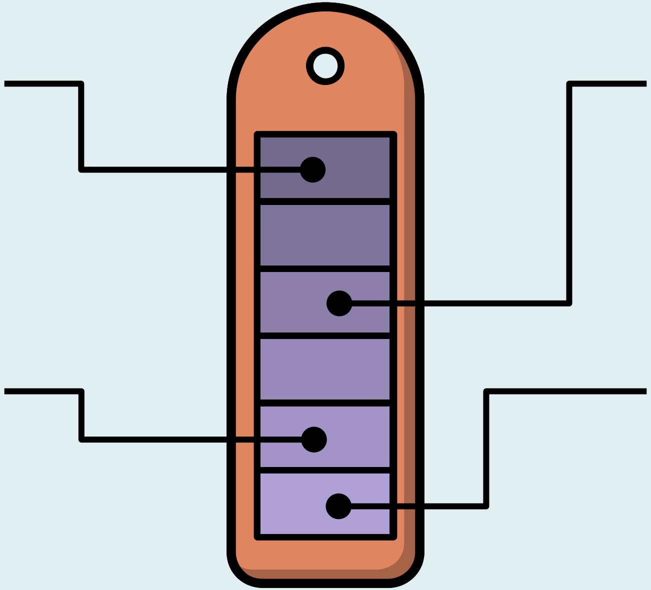
Link

Meet us

Feedback

Participant	What Worked Well	What Was Confusing or Challenging	Suggestions for Improvement	Overall Rating (1–5)
A	Loved using AlphaFold; visuals were engaging	Confused about finding the right FASTA sequence	Add short video/gif showing how to navigate NCBI	⭐⭐⭐⭐
B	Found comparing structures interesting	Unsure how to interpret differences in structures	Include example comparison (before vs. after screenshot)	⭐⭐⭐⭐
C	Clear instructions and well-paced	Didn't know what to look for in the 3D structure	Provide checklist of things to look at in structures	⭐⭐⭐⭐⭐
D	Fun to explore protein shape changes	Didn't understand the purpose of NCBI vs. AlphaFold	Add a slide or box explaining the difference between databases	⭐⭐⭐⭐
E	Enjoyed learning real-world tools	Needed more context on how mutations affect function	Include real-world example like sickle cell earlier in activity	⭐⭐⭐

How We Improved



Improved Guide

Added a more clear step by step guide with screenshots

Summary

Defined terminology and databases better for understanding and context

Added Example

Students did not understand how to compare structures added a case study example

Visual Checklist

Added a table of different proteins for students to pick from

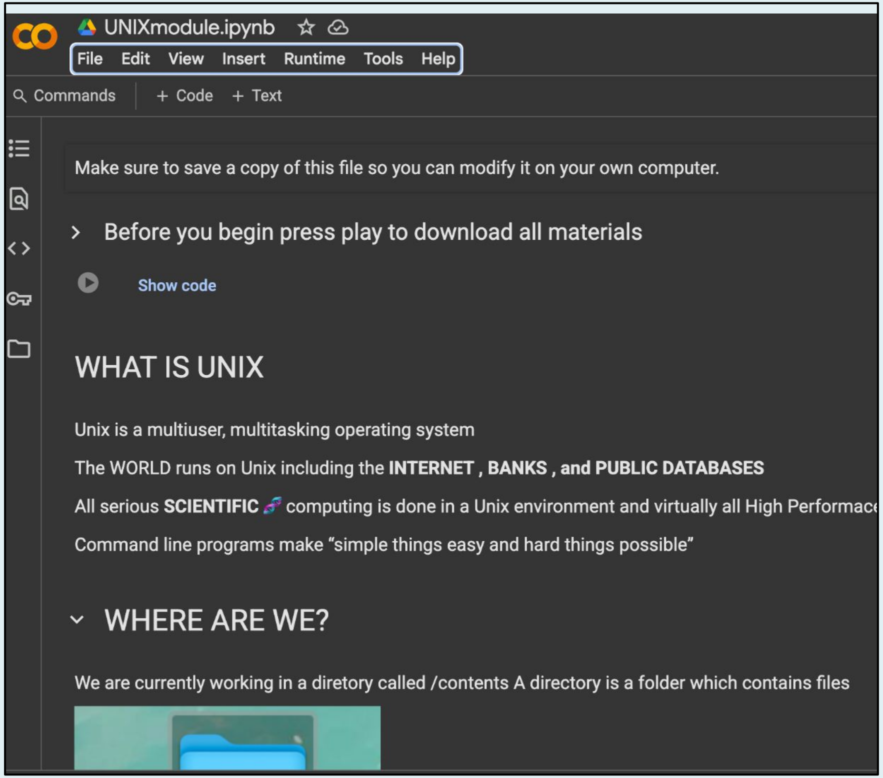
04

UNIX Commands in Bioinformatics

Understanding and Working with UNIX commands
and applying it to BLAST

Unix Module format

- Chose to do it in google colab because you can use unix/python/blast without having to download it on your computer
- Start with WHY Unix is important
- Intro to a few commands
- Learn the commands useful for a small bioinformatics test case
- Be able to apply a command by themselves on a “real world” example
- Redirect Students to more in depth Workshops and Resources



Make sure to save a copy of this file so you can modify it on your own computer.

Before you begin press play to download all materials

WHAT IS UNIX

Unix is a multiuser, multitasking operating system

The WORLD runs on Unix including the **INTERNET , BANKS , and PUBLIC DATABASES**

All serious **SCIENTIFIC** 🧬 computing is done in a Unix environment and virtually all High Performac

Command line programs make “simple things easy and hard things possible”

WHERE ARE WE?

We are currently working in a diretory called /contents A directory is a folder which contains files

Case Study:

HPV identification

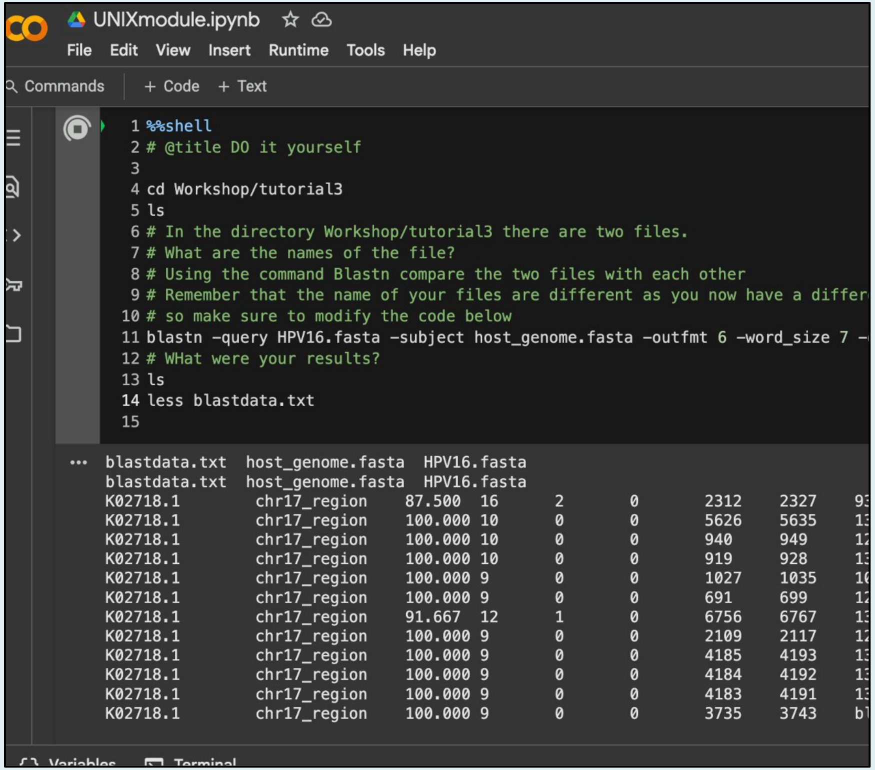
Students dealt with 2 Fasta files.

1. HPV16.fasta: The known viral sequence (HPV16 DNA).

2. host_genome.fasta: A small segment of a host genome

HPV16 (Human Papillomavirus type 16) is a high-risk strain of HPV that is strongly associated with several types of cancer

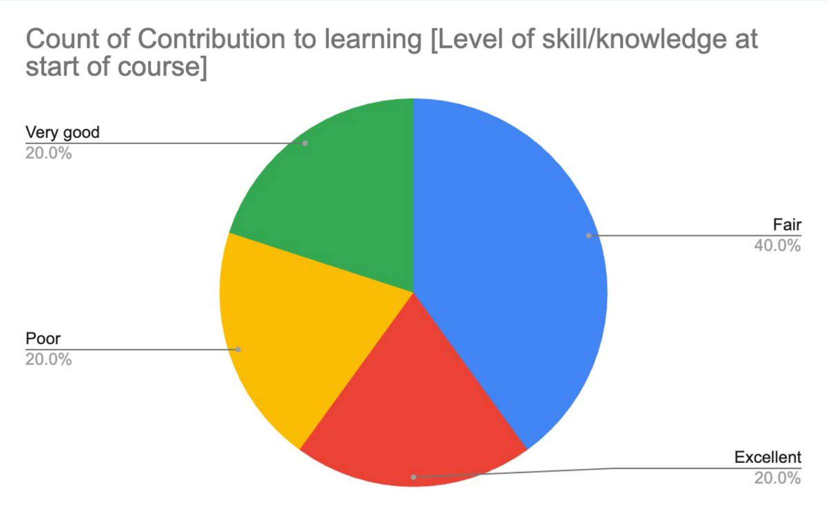
I explained what a FASTA file was and how to identify it after receiving feedback
Then the students run BLAST to see if there's a match between the 2 files

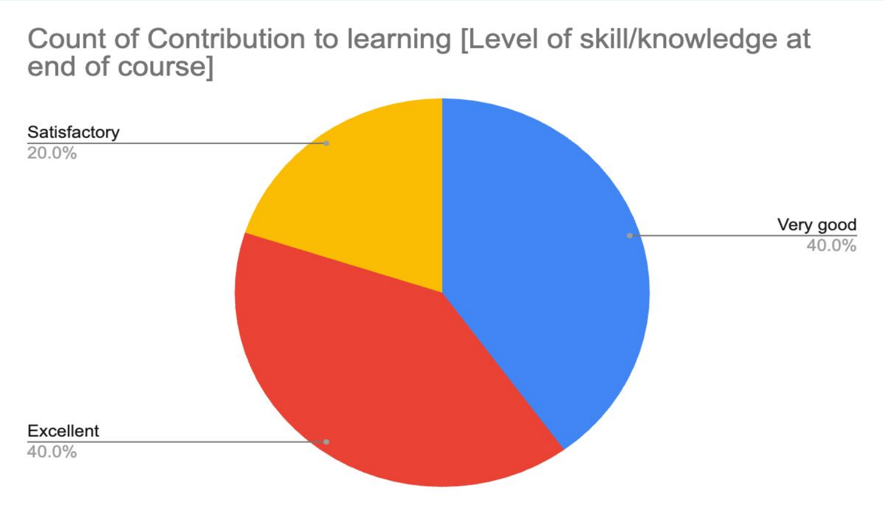


```
%%shell
# @title DO it yourself

cd Workshop/tutorial3
ls
# In the directory Workshop/tutorial3 there are two files.
# What are the names of the file?
# Using the command Blastn compare the two files with each other
# Remember that the name of your files are different as you now have a differ
# so make sure to modify the code below
blastn -query HPV16.fasta -subject host_genome.fasta -outfmt 6 -word_size 7 -
# WHat were your results?
ls
less blastdata.txt

```

```
blastdata.txt  host_genome.fasta  HPV16.fasta
blastdata.txt  host_genome.fasta  HPV16.fasta
K02718.1        chr17_region    87.500  16      2       0       2312    2327    93
K02718.1        chr17_region    100.000 10      0       0       5626    5635    13
K02718.1        chr17_region    100.000 10      0       0       940     949     12
K02718.1        chr17_region    100.000 10      0       0       919     928     13
K02718.1        chr17_region    100.000 9       0       0       1027    1035    10
K02718.1        chr17_region    100.000 9       0       0       691     699     12
K02718.1        chr17_region    91.667  12      1       0       6756    6767    13
K02718.1        chr17_region    100.000 9       0       0       2109    2117    12
K02718.1        chr17_region    100.000 9       0       0       4185    4193    13
K02718.1        chr17_region    100.000 9       0       0       4184    4192    13
K02718.1        chr17_region    100.000 9       0       0       4183    4191    13
K02718.1        chr17_region    100.000 9       0       0       3735    3743    bl
```

Level of skill before and after the course


Count of Contribution to learning [Level of skill/knowledge at start of course]


Count of Contribution to learning [Level of skill/knowledge at end of course]

60%

Strongly Agree that the learning objectives were clear

80%

The level of difficulty for the Unix Section was a 4 on a scale of 5. 5 being easy to understand

40%

The level of difficulty for the BLASTn Section was a 4 on a scale of 5. 5 being easy to understand

Verbal Feedback

Pros

“I enjoyed the user friendly and step by step process for getting started with the course.”

“I liked the guided practice”

“Though I have minimal knowledge of this field it was helpful with the instructions”

Why did you want to take this course?

“I wanted to learn more about CS and this seemed the least intimidating way to start”

Cons

A few people were confused and were wondering WHERE the Workshop was going to be held.

We had to remind them that it was online, and all they needed to do was click on the link.

The Blastn Section was the most confusing for students. They didn’t know that it was a program that could be run with a few “words” (commands).

Recap

01 Built 3 Separate Modules

02 Conducted user testing with a group of participants.

03 Modified Modules based on feedback

04 Looking Forward

Looking forward

01

Promote the Workshop to Rising Sophomores

02

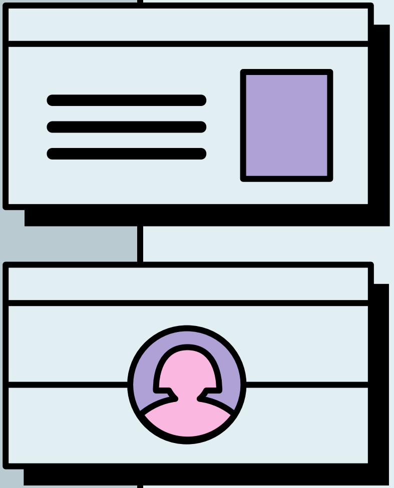
Work on Promotional Flyers/Art to grab people's attention

03

Create Canvas Course and run the Course for 4 Consecutive Weeks

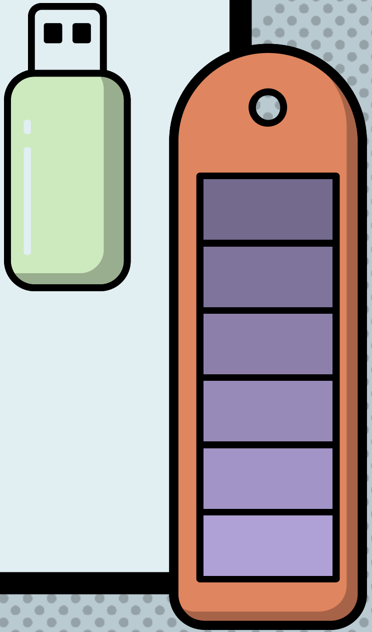
04

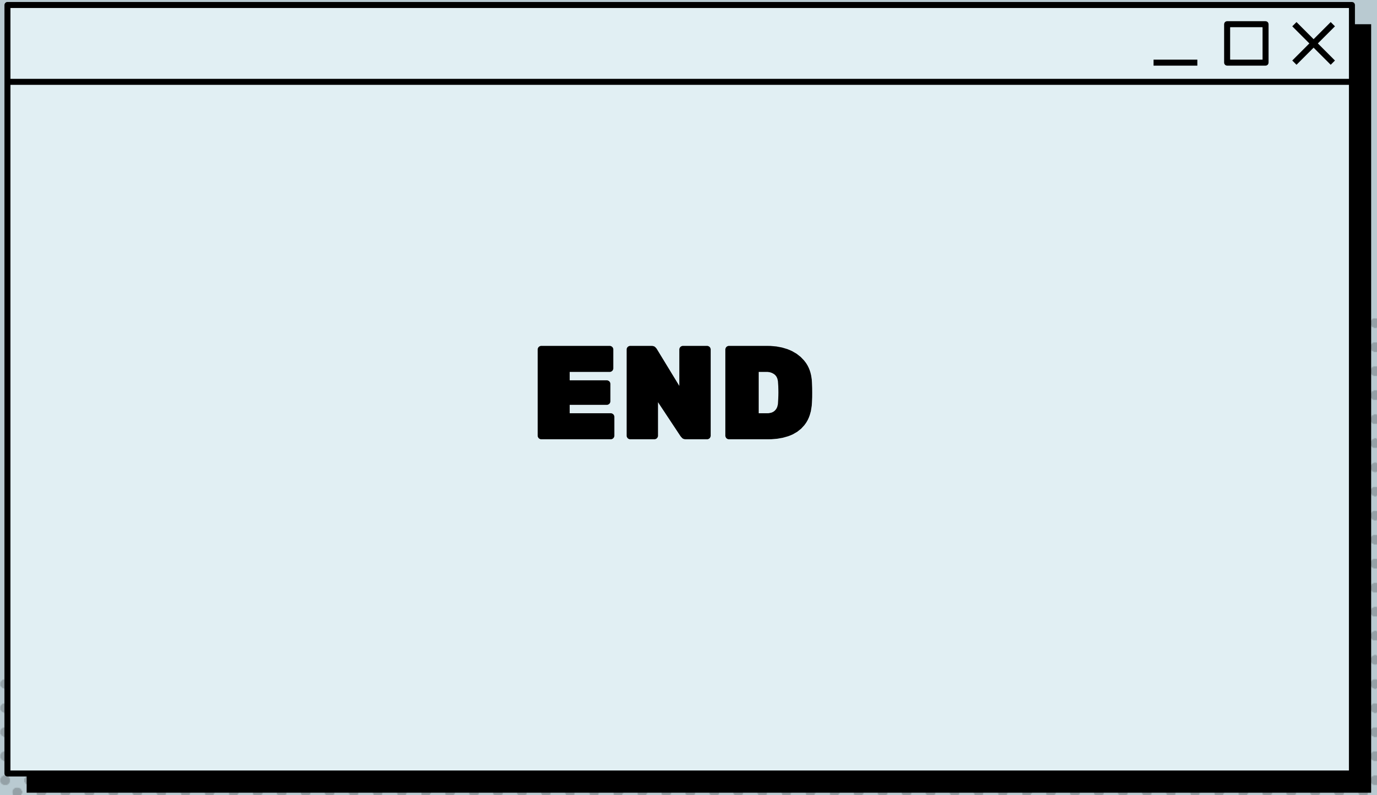
Try to set it up online discussion, so that people can ask questions as they work



Questions

Thank You!





END