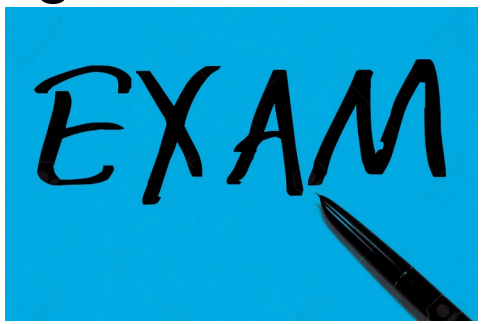
Bioinformatics
CS300
Working With Blast:
manual and “automatic” use

Spring 2021

Oliver BONHAM-CARTER

Exam 1 :: UPDATE!!!
During lab Monday, 12th April

- Multiple choice, Short answer, Short essay,
- Given online on Sakai's "*Tests and Quizzes*" Tab at the beginning of lab.
- You will have **48 hours** to take the exam starting from Monday's lab (2:50pm) [DUE TO SPRING BREAK: 13April]
- The exam will be due on 14th April at 2:50pm and will likely take about two hours to complete during that time.



Blast:
Find your sequence

COVID-19 is an emerging, rapidly evolving situation.

Public health information (CDC) | Research information (NIH)

SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)

GenBank ▾ Send to: ▾

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2

FASTA Graphics

Go to:

```
LOCUS       NC_045512               29903 bp ss-RNA     linear   VRL 18-JUL-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
            complete genome.
ACCESSION   NC_045512
VERSION     NC_045512.2
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus.
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

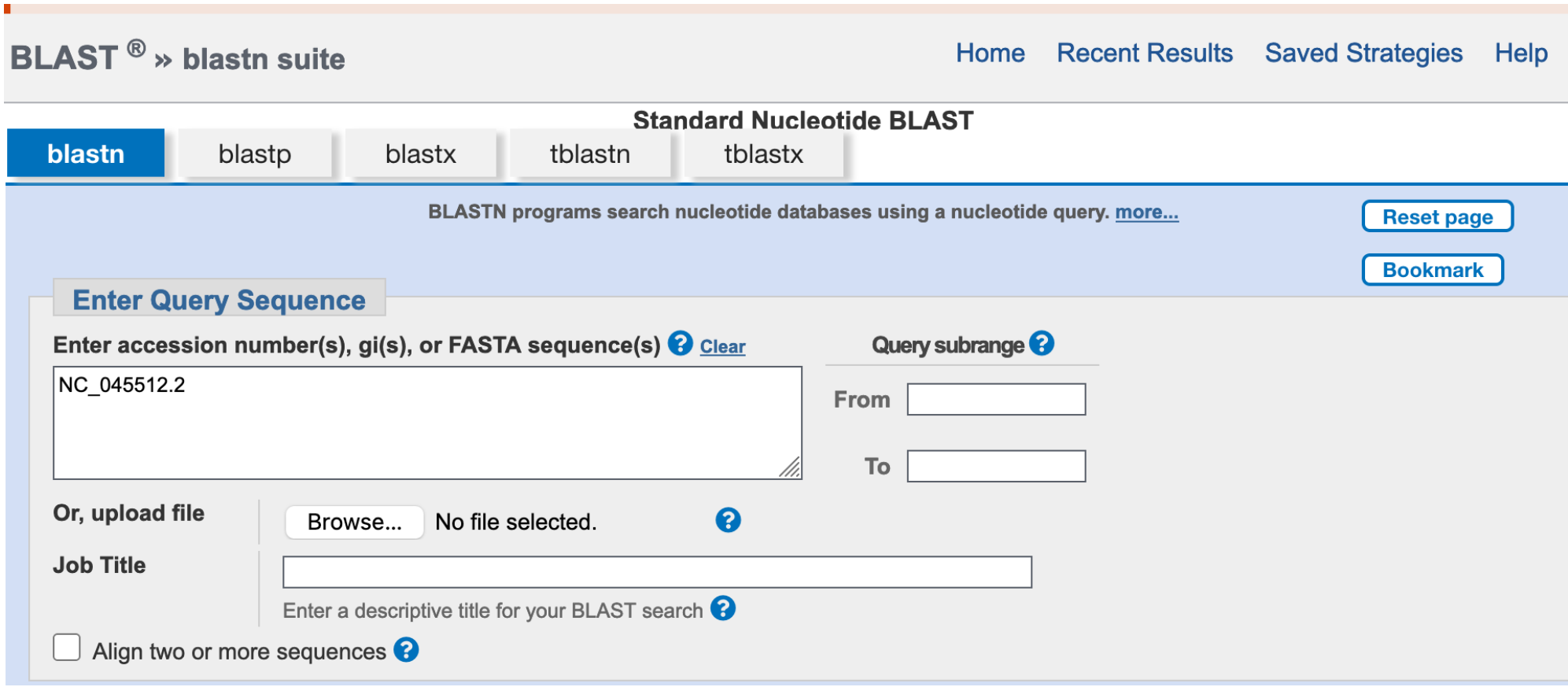
Highlight Sequence Features

Find in this Sequence

NCBI Virus

Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.

Blast:
Enter accession number of sequence



Blast:
Fill in nucleotide parameters

Choose Search Set

Database — Standard databases (nr etc.): rRNA/ITS databases; Genomic + transcript databases; Betacoronavirus

Nucleotide collection (nr/nt)

Organism Optional — Enter organism name or id--completions will be suggested; exclude; Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional — Models (XM/XP); Uncultured/environmental sample sequences

Limit to Optional — Sequences from type material

Entrez Query Optional — YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Blast Initiation:
Running the analysis

BLAST ® » blastn suite » RID-67TST0X3016 Home Recent Results Saved Strategies Help

Format Request Status

[Formatting options]

Job Title: ref|NC_045512.2|

Request ID	**67TST0X3016**
Status	Searching
Submitted at	Tue Mar 30 23:06:01 2021
Current time	Tue Mar 30 23:06:10 2021
Time since submission	00:00:08

This page will be automatically updated in **2** seconds

Blog Support Center

Blast Results:
Resembling sequences

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download New Select columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results New MSA Viewer

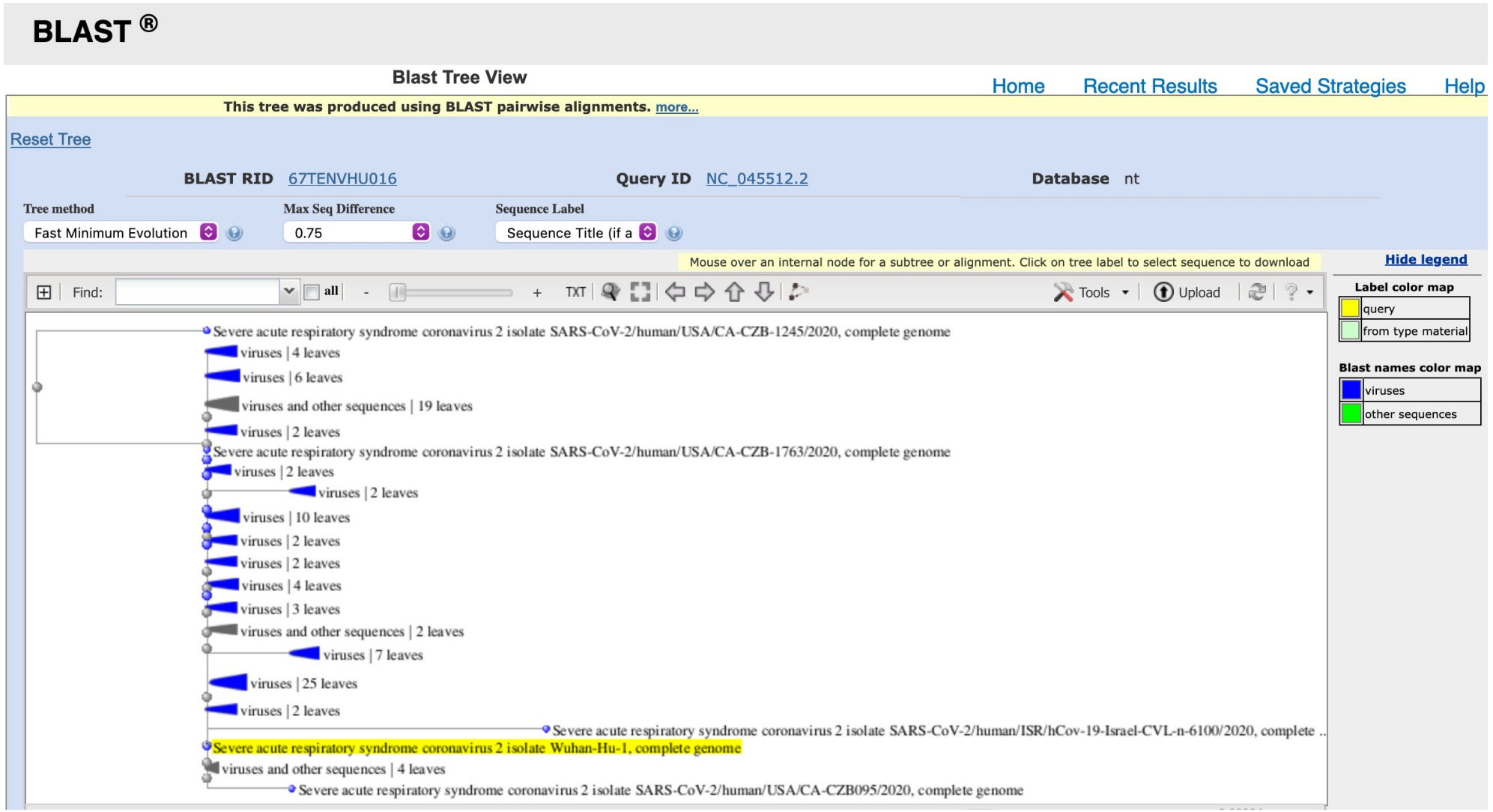
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome	Severe acute r...	55221	55221	100%	0.0	100.00%	29903	NC_045512.2
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/UT-UPHL-2102...	Severe acute r...	55221	55221	100%	0.0	100.00%	29904	MW566244.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/UT-UPHL-2102...	Severe acute r...	55221	55221	100%	0.0	100.00%	29904	MW562722.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CZB-1763/...	Severe acute r...	55217	55217	99%	0.0	100.00%	29901	MT671817.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CDC-0139/...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MT481992.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/France/10009EE/20...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MT470142.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CZB-1040/...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MT438758.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CZB-1004/...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MT438722.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/EGY/EGY_CCHE57...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MW467502.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/EGY/EGY_CCHE57...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MW467494.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/EGY/EGY_CCHE57...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MW467463.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/EGY/EGY_CCHE57...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MW467462.1
Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/EGY/EGY_CCHE57...	Severe acute r...	55215	55215	100%	0.0	100.00%	29903	MW467439.1

Blast:
Making sense of scores

- **Max Score**
  - The score of the best matching segment for local alignment, not global
- **Total Score**
  - The total scores of all matching segments found (same as max score if there is only one matching segment)
- **Query Coverage**
  - The percentage of the query sequence that aligned to some part of the match.
- **E-Value**
  - A statistical measure evaluating how likely it is that a match this good could occur by chance. Lower e-scores indicate that both sequences are truly similar and are not similar by chance alone. Identical sequences have e-scores of zero.
- **Max Indent**
  - The percentage of nucleotides that are identical between the query and the target sequences within the matching regions.

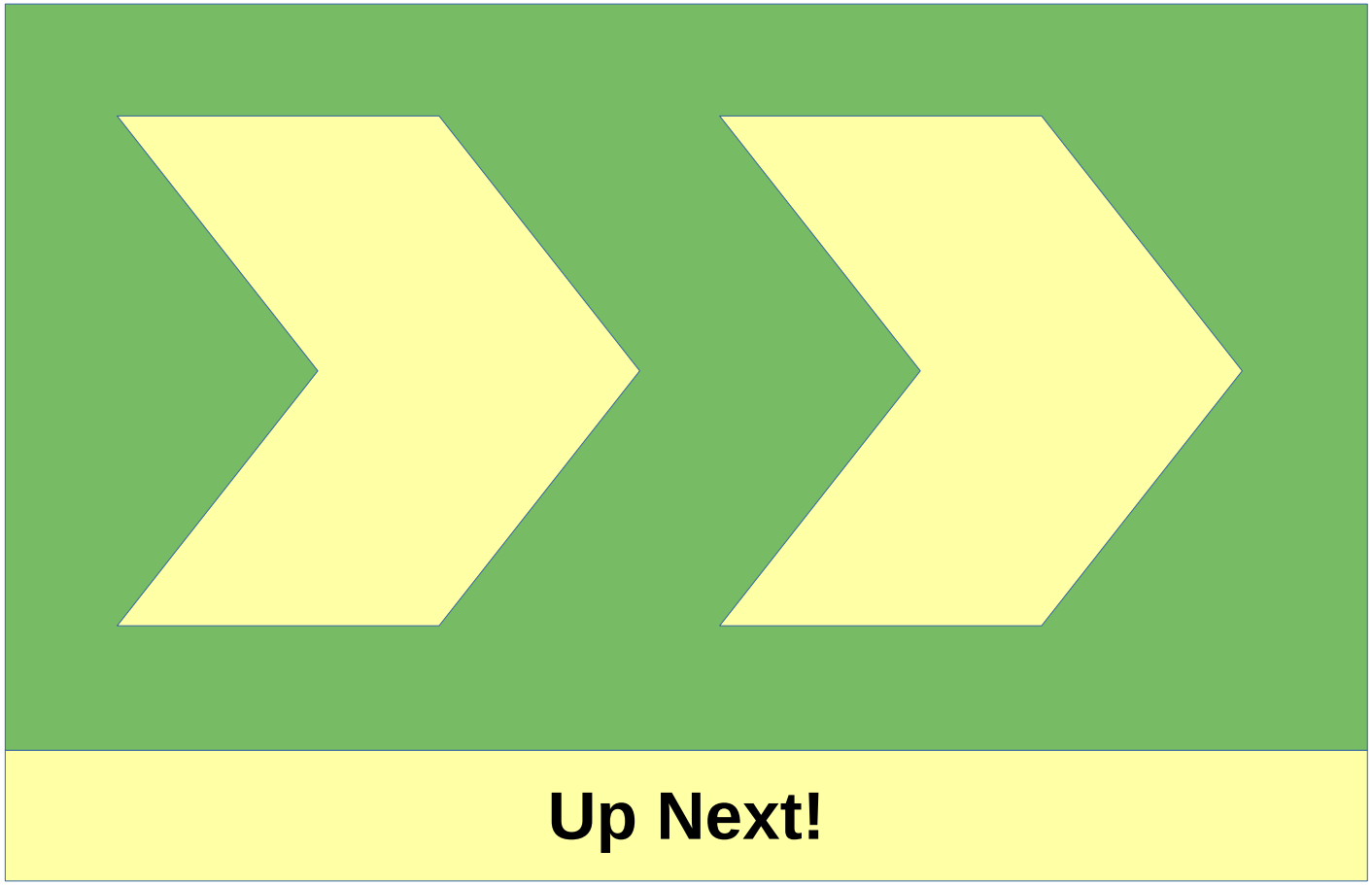
Reference:
https://www.ncbi.nlm.nih.gov/books/NBK62051/

Blast:
Phylogenic trees to show relatedness

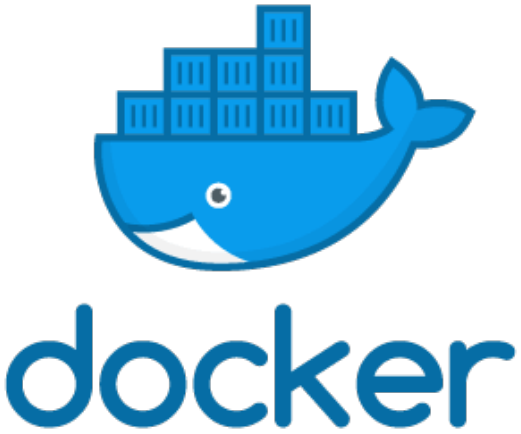


Blast is Cool!

Bring the Code!



Please run your container!!



Dockerfile

(for Docker Desktop)

We will be using BioPython

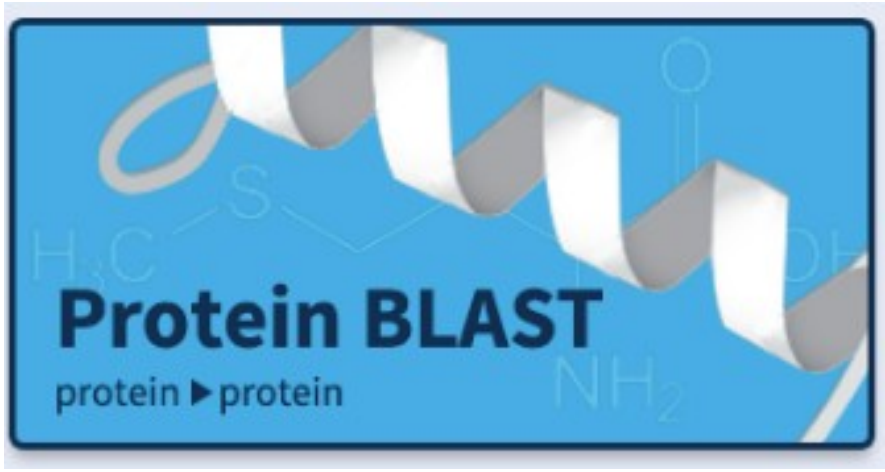
```
FROM ubuntu:20.04

RUN apt-get update && apt-get -y
install git htop vim python3 python3-pip
clustalw


RUN \
pip3 install --upgrade pip \
pip install bioPython

WORKDIR /root

CMD ["bash"]
```

Activity 06

- Activity 06
  - https://forms.gle/1aKbnUTz11P937fm8
- Blast Link:
  - https://blast.ncbi.nlm.nih.gov/Blast.cgi