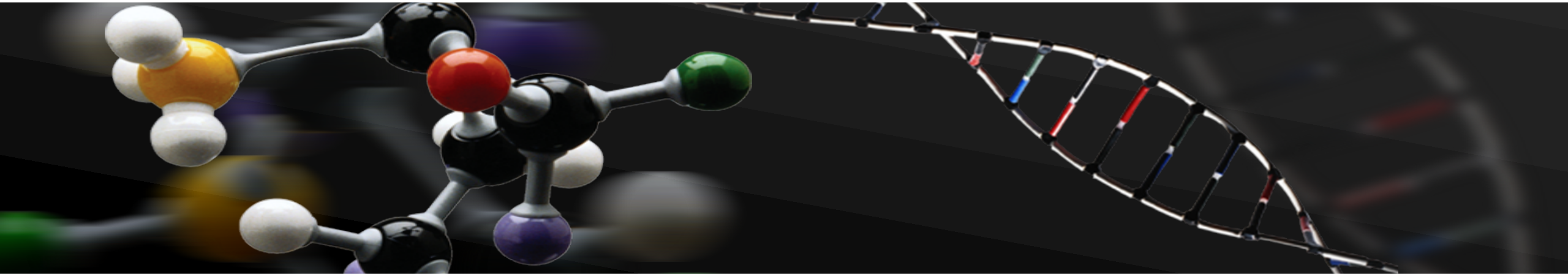
BIOINFORMATICS 2

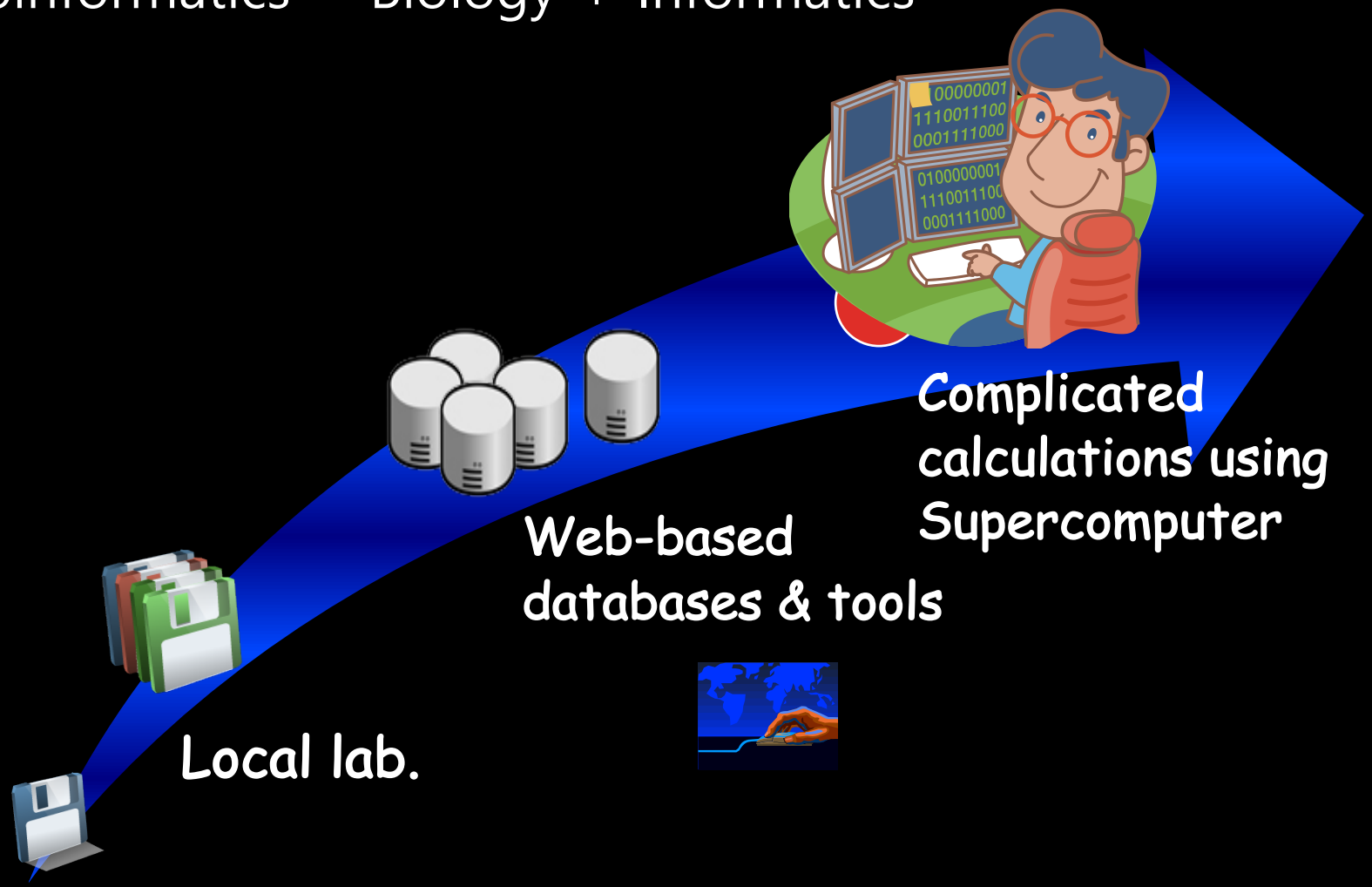


16th November

UST 과학기술연합대학원대학교
UNIVERSITY OF SCIENCE & TECHNOLOGY

WHAT IS BIOINFORMATICS..?

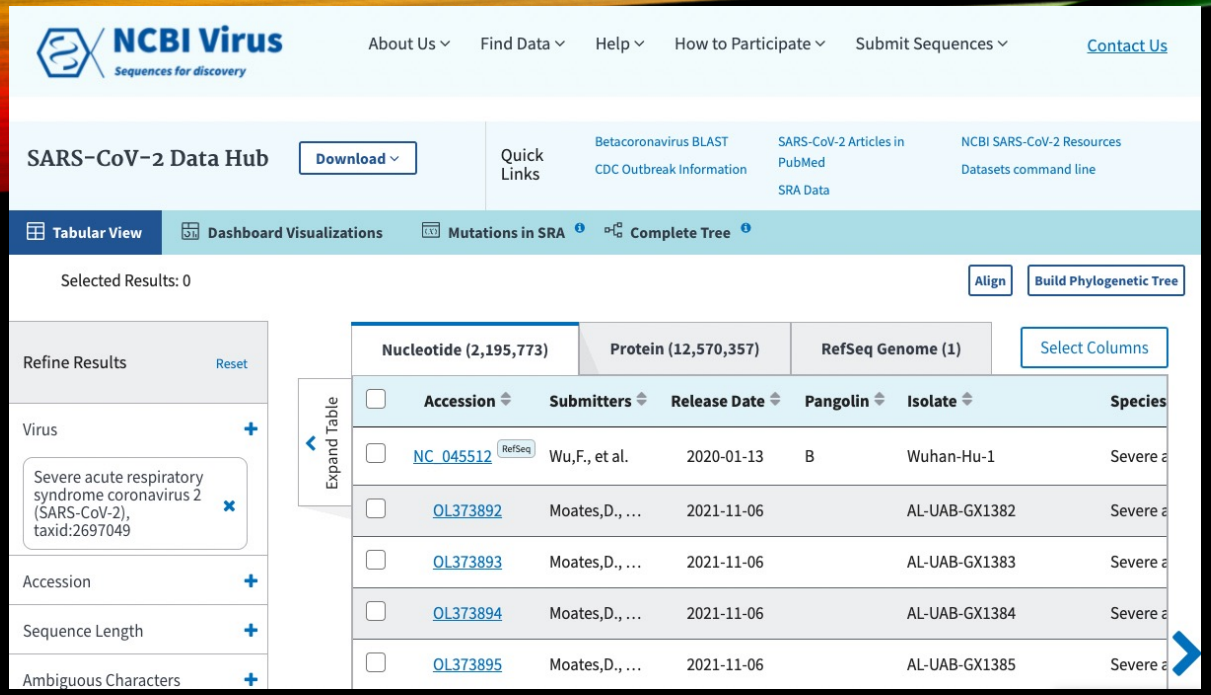
- Bioinformatics = Biology + Informatics

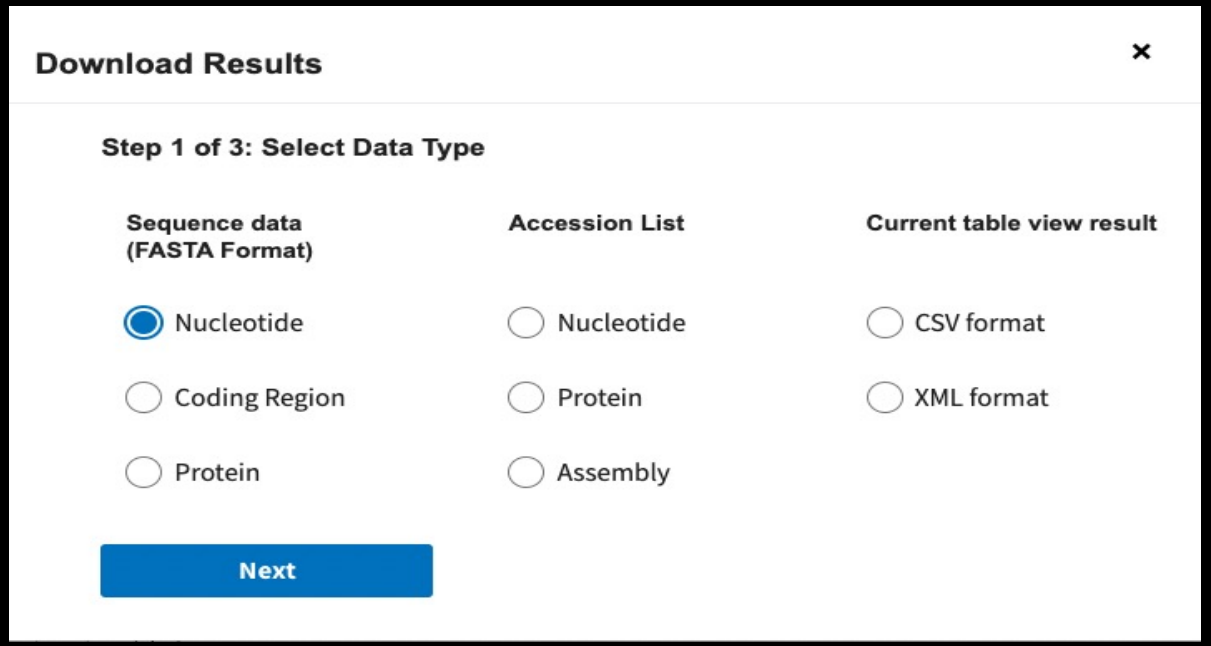


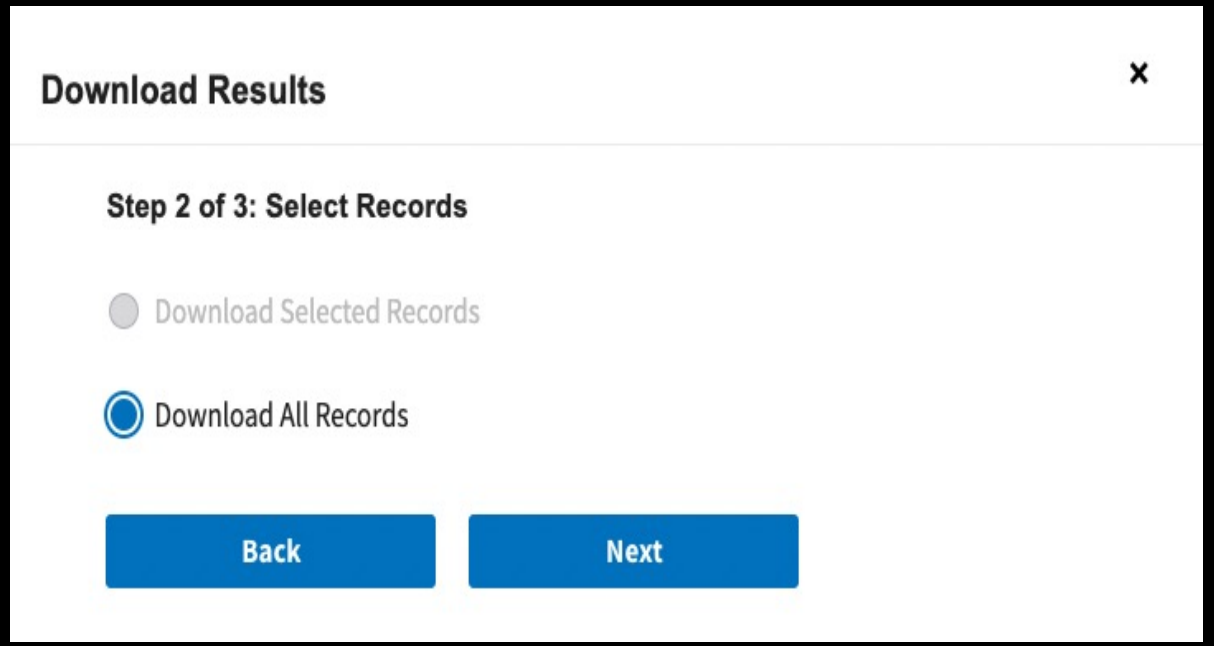
ACCESS TO DATA IN NCBI

- Entrez Nucleotide (https://www.ncbi.nlm.nih.gov/nucleotide)
  - ✓Search GenBank for sequence identifiers and annotations
- BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)
  - ✓Search and align GenBank sequences to a query sequence
- NCBI e-utils (https://www.ncbi.nlm.nih.gov/books/NBK25501)
  - ✓Search, link, and download sequences programatically
- FTP server (https://ftp.ncbi.nlm.nih.gov/ncbi-asn1)
  - ✓The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server

ENTREZ NUCLEOTIDE

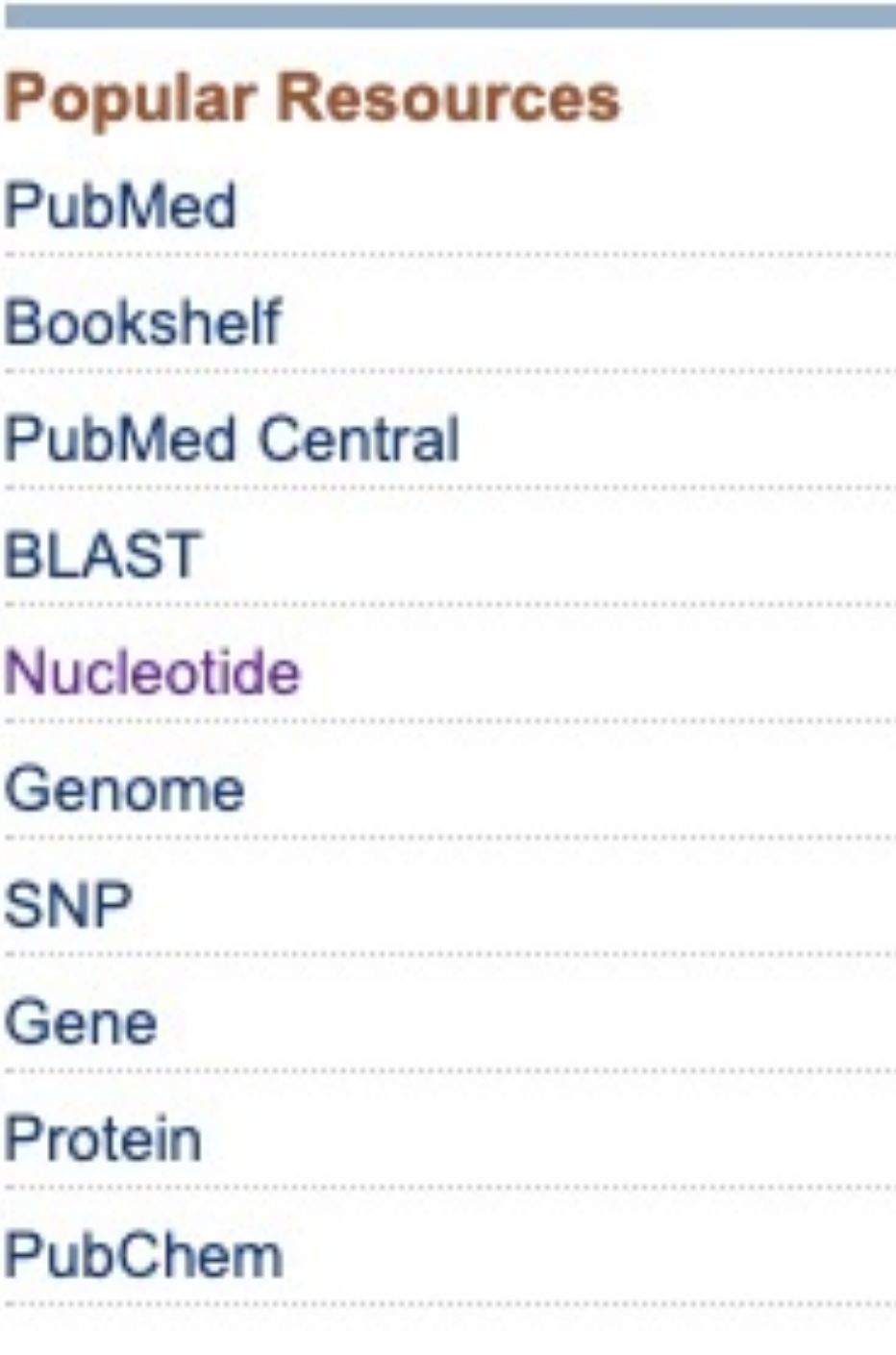






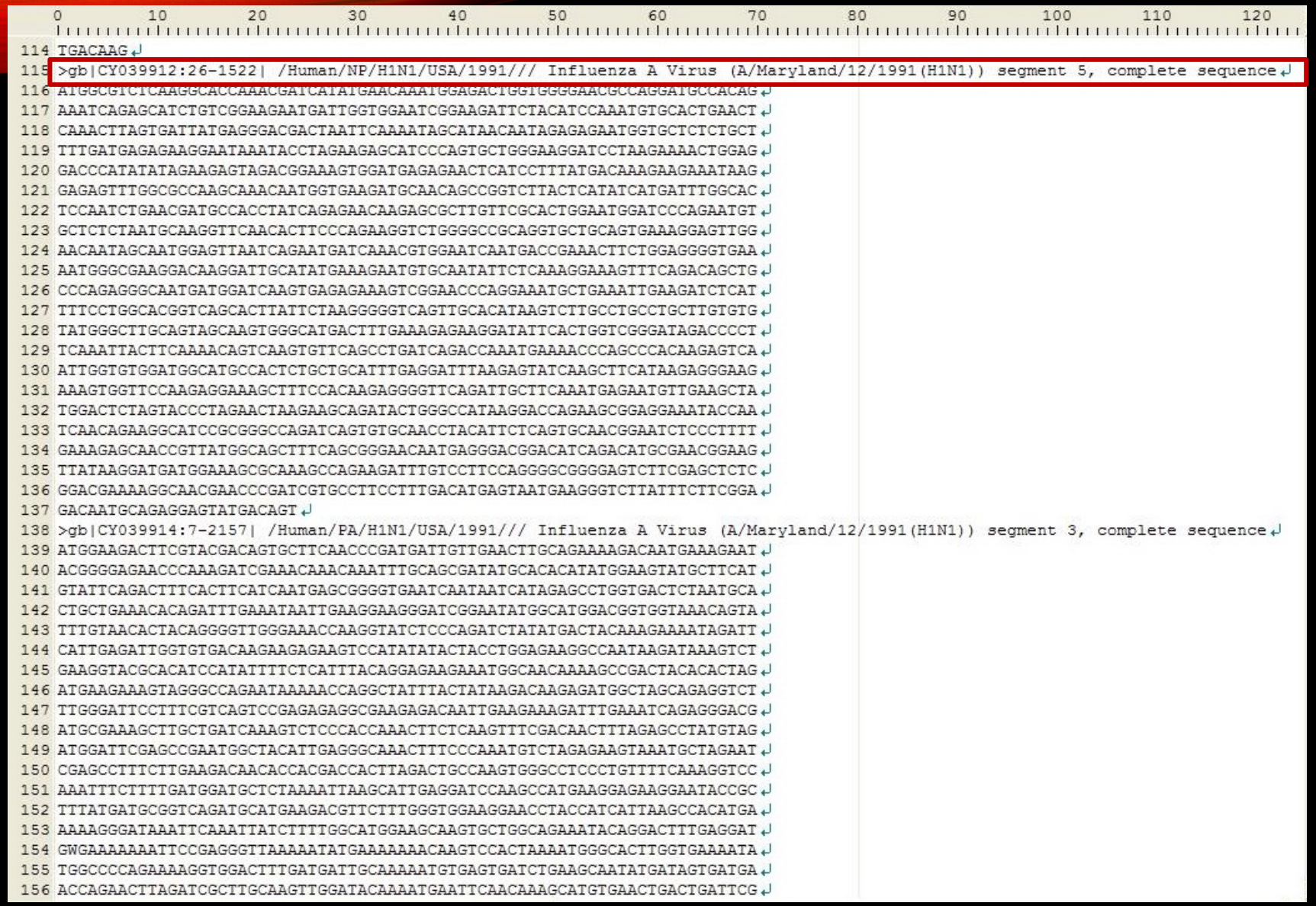
E-UTILS

- Searching a Database
  - Input: Entrez database (&db); Any Entrez text query (&term)
  - Output: List of UIDs matching the Entrez query
  - Example: *Get the PubMed IDs (PMIDs) for articles about SARS-CoV-2 published in Nature in 2021*
    - https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=nature[journal]+SARS-CoV-2+2021[pdat]



STEP 2: PARSING ANNOTATION DATA

BEFORE



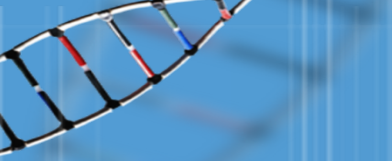
```
0        10        20        30        40        50        60        70        80        90        100       110       120
TGACAAG
>gb|CY039912:26-1522| /Human/NP/H1N1/USA/1991/// Influenza A Virus (A/Maryland/12/1991(H1N1)) segment 5, complete sequence
ATGGCGTCTCAAGGCACCAAACGATCATATGAACAAATGGAGACTGGTGGGGAACGCCAGGATGCCACAG
AAATCAGAGCATCTGTCGGAAGAATGATTGGTGGAATCGGAAGATTCTACATCCAAATGTGCACTGAACT
CAAACTTAGTGATTATGAGGGACGACTAATTCAAAATAGCATAACAATAGAGAGAATGGTGCTCTCTGCT
TTTGATGAGAGAAGGAATAAATACCTAGAAGAGCATCCCAGTGCTGGGAAGGATCCTAAGAAAACTGGAG
GACCCATATATAGAAGAGTAGACGGAAAGTGGATGAGAGAACTCATCCTTTATGACAAAGAAGAAATAAG
GAGAGTTTGGCGCCAAGCAAACAATGGTGAAGATGCAACAGCCGGTCTTACTCATATCATGATTTGGCAC
TCCAATCTGAACGATGCCACCTATCAGAGAACAAGAGCGCTTGTTCGCACTGGAATGGATCCCAGAATGT
GCTCTCTAATGCAAGGTTCAACACTTCCCAGAAGGTCTGGGGCCGCAGGTGCTGCAGTGAAAGGAGTTGG
AACAATAGCAATGGAGTTAATCAGAATGATCAAACGTGGAATCAATGACCGAAACTTCTGGAGGGGTGAA
AATGGGCGAAGGACAAGGATTGCATATGAAAGAATGTGCAATATTCTCAAAGGAAAGTTTCAGACAGCTG
CCCAGAGGGCAATGATGGATCAAGTGAGAGAAAGTCGGAACCCAGGAAATGCTGAAATTGAAGATCTCAT
TTTCCTGGCACGGTCAGCACTTATTCTAAGGGGGTCAGTTGCACATAAGTCTTGCCTGCCTGCTTGTGTG
TATGGGCTTGCAGTAGCAAGTGGGCATGACTTTGAAAGAGAAGGATATTCACTGGTCGGGATAGACCCCT
TCAAATTACTTCAAAACAGTCAAGTGTTCAGCCTGATCAGACCAAATGAAAACCCAGCCCACAAGAGTCA
ATTGGTGTGGATGGCATGCCACTCTGCTGCATTTGAGGATTTAAGAGTATCAAGCTTCATAAGAGGGAAG
AAAGTGGTTCCAAGAGGAAAGCTTTCCACAAGAGGGGTTCAGATTGCTTCAAATGAGAATGTTGAAGCTA
TGGACTCTAGTACCCTAGAACTAAGAAGCAGATACTGGGCCATAAGGACCAGAAGCGGAGGAAATACCAA
TCAACAGAAGGCATCCGCGGGCCAGATCAGTGTGCAACCTACATTCTCAGTGCAACGGAATCTCCCTTTT
GAAAGAGCAACCGTTATGGCAGCTTTCAGCGGGAACAATGAGGGACGGACATCAGACATGCGAACGGAAG
TTATAAGGATGATGGAAAGCGCAAAGCCAGAAGATTTGTCCTTCCAGGGGCGGGGAGTCTTCGAGCTCTC
GGACGAAAAGGCAACGAACCCGATCGTGCCTTCCTTTGACATGAGTAATGAAGGGTCTTATTTCTTCGGA
GACAATGCAGAGGAGTATGACAGT
>gb|CY039914:7-2157| /Human/PA/H1N1/USA/1991/// Influenza A Virus (A/Maryland/12/1991(H1N1)) segment 3, complete sequence
ATGGAAGACTTCGTACGACAGTGCTTCAACCCGATGATTGTTGAACTTGCAGAAAAGACAATGAAAGAAT
ACGGGGAGAACCCAAAGATCGAAACAAACAAATTTGCAGCGATATGCACACATATGGAAGTATGCTTCAT
GTATTCAGACTTTCACTTCATCAATGAGCGGGGTGAATCAATAATCATAGAGCCTGGTGACTCTAATGCA
CTGCTGAAACACAGATTTGAAATAATTGAAGGAAGGGATCGGAATATGGCATGGACGGTGGTAAACAGTA
TTTGTAACACTACAGGGGTTGGGAAACCAAGGTATCTCCCAGATCTATATGACTACAAAGAAAATAGATT
CATTGAGATTGGTGTGACAAGAAGAGAAGTCCATATATACTACCTGGAGAAGGCCAATAAGATAAAGTCT
GAAGGTACGCACATCCATATTTTCTCATTTACAGGAGAAGAAATGGCAACAAAAGCCGACTACACACTAG
ATGAAGAAAGTAGGGCCAGAATAAAAACCAGGCTATTTACTATAAGACAAGAGATGGCTAGCAGAGGTCT
TTGGGATTCCTTTCGTCAGTCCGAGAGAGGCGAAGAGACAATTGAAGAAAGATTTGAAATCAGAGGGACG
ATGCGAAAGCTTGCTGATCAAAGTCTCCCACCAAACTTCTCAAGTTTCGACAACTTTAGAGCCTATGTAG
ATGGATTCGAGCCGAATGGCTACATTGAGGGCAAACTTTCCCAAATGTCTAGAGAAGTAAATGCTAGAAT
CGAGCCTTTCTTGAAGACAACACCACGACCACTTAGACTGCCAAGTGGGCCTCCCTGTTTTCAAAGGTCC
AAATTTCTTTTGATGGATGCTCTAAAATTAAGCATTGAGGATCCAAGCCATGAAGGAGAAGGAATACCGC
TTTATGATGCGGTCAGATGCATGAAGACGTTCTTTGGGTGGAAGGAACCTACCATCATTAAGCCACATGA
AAAAGGGATAAATTCAAATTATCTTTTGGCATGGAAGCAAGTGCTGGCAGAAATACAGGACTTTGAGGAT
GWGAAAAAAATTCCGAGGGTTAAAAATATGAAAAAAACAAGTCCACTAAAATGGGCACTTGGTGAAAATA
TGGCCCCAGAAAAGGTGGACTTTGATGATTGCAAAAATGTGAGTGATCTGAAGCAATATGATAGTGATGA
ACCAGAACTTAGATCGCTTGCAAGTTGGATACAAAATGAATTCAACAAAGCATGTGAACTGACTGATTCG
```

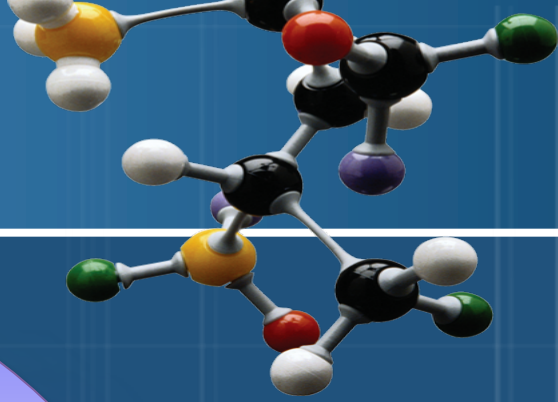
AFTER


0 10 20 30 40 50 60
1 >HA/S67220_HA/Swine/H1N1/USA/1992/A
2 AUGAAGGCAAUACCAUUAGUCUUGCUAUAUACAUUUACAGCCGCAAAUGCAGACACACUAUGUAUAGGU
3 >PB2/AJ564805_PB2/Human/H1N1/Fiji/1983/A
4 AUGGAAAGAAUAAAAGAGCUAAGGAAUCUGAUGUCGCAGUCCCGCACUCGCGAGAUACUUACAAAAACC
5 >NA/CY039911_NA/Human/H1N1/USA/1991/A
6 AUGAAUACAAACCAAAGAAUAAUAACCAUCGGGACAGUCUGUCUGAUAGUUGGAAUAGUUAGUCUAUUA
7 >NP/CY039912_NP/Human/H1N1/USA/1991/A
8 AUGGCGUCUCAAGGCACCAAACGAUCAUAUGAACAAAUGGAGACUGGUGGGGAACGCCAGGAUGCCACA
9 >PB1/CY039915_PB1/Human/H1N1/USA/1991/A
10 AUGGAUGUCAAUCCGACUUUACUUUUCCUGAAAGUGCCAGCACAAAAUGCUAUAAGCACAACGUUUCCU
11 >M1/GQ404564_M1/Avian/H1N1/Czech Republic/2008/A
12 AUGAGUCUUCUAACCGAGGUCGAAACGUACGUUCUCUCCAUCAUCCCGUCGGGCCCCCUCAAAGCCGAC
13 >M2/GQ404564_M2/Avian/H1N1/Czech Republic/2008/A
14 AUGAGUCUUCUAACCGAGGUCGAAACGCCUACCAGAAACGGAUGGGAGUGCAAAUGCAGCGAUUCAAGU
15 >M2/GQ404616_M2/Swine/H1N1/Czech Republic/1992/A
16 AUGAGUCUUCUAACCGAGGUCGAAACGCCUACCAGAAACGAAUGGGAGUGCAGAUGCAGCGAUUCAAGU
17 >M2/GQ404581_M2/Swine/H1N1/Czech Republic/1957/A
18 AUGAGCCUUCUAACCGAGGUCGAAACGCCUACCAGAAACGAAUGGGGGUGCAGAUGCAACGAUUCAAGU
19 >M2/GQ404585_M2/Swine/H1N1/Belarus/1965/A
20 AUGAGUCUUCUAACCGAGGUCGAAACGCCUAUCAGAAACGAAUGGGGGUGCAGAUGCAACGGUUCAAGU
21 >M2/GQ404587_M2/Swine/H1N1/USA/1968/A
22 AUGAGCCUUCUAACCGAGGUCGAAACGCCUAUCAAAAGCGAAUGGGGGUGCAGAUGCAACGAUUCAAGU

MAKE MYSQL INPUT FILES

```
CREATE TABLE influ_A_nt_20091207 (
accession varchar (50) NOT NULL default '',
host varchar(20) default NULL,
gene varchar(20) default NULL,
subtype varchar(20) default NULL,
country varchar(20) default NULL,
year int(11) NOT NULL default '0',
len int(11) NOT NULL default '0',
sequence text,
species varchar(5) default NULL,
PRIMARY KEY (accession)
) TYPE = MyISAM;
INSERT INTO influ_A_nt_20091207 VALUES ('S67220_HA', 'swine', 'HA', 'H1N1', 'USA', 1992, 1728, 'AUGAAGGCAAUACCAUUAGUC
INSERT INTO influ_A_nt_20091207 VALUES ('AJ564805_PB2', 'human', 'PB2', 'H1N1', 'Fiji', 1983, 2295, 'AUGGAAAGAAUAAAAC
INSERT INTO influ_A_nt_20091207 VALUES ('CY039911_NA', 'human', 'NA', 'H1N1', 'USA', 1991, 1407, 'AUGAAUACAAACCAAAGAA
INSERT INTO influ_A_nt_20091207 VALUES ('CY039912_NP', 'human', 'NP', 'H1N1', 'USA', 1991, 1494, 'AUGGCGUCUCAAGGCACCA
INSERT INTO influ_A_nt_20091207 VALUES ('CY039915_PB1', 'human', 'PB1', 'H1N1', 'USA', 1991, 2271, 'AUGGAUGUCAAUCCGAC
INSERT INTO influ_A_nt_20091207 VALUES ('GQ404564_M1', 'avian', 'M1', 'H1N1', 'Czech Republic', 2008, 756, 'AUGAGUCUU
INSERT INTO influ_A_nt_20091207 VALUES ('GQ404564_M2', 'avian', 'M2', 'H1N1', 'Czech Republic', 2008, 291, 'AUGAGUCUU
INSERT INTO influ_A_nt_20091207 VALUES ('GQ404616_M2', 'swine', 'M2', 'H1N1', 'Czech Republic', 1992, 291, 'AUGAGUCUU
INSERT INTO influ_A_nt_20091207 VALUES ('GQ404581_M2', 'swine', 'M2', 'H1N1', 'Czech Republic', 1957, 288, 'AUGAGCCUU
INSERT INTO influ_A_nt_20091207 VALUES ('GQ404585_M2', 'swine', 'M2', 'H1N1', 'Belarus', 1965, 291, 'AUGAGUCUUCUAACCC
INSERT INTO influ_A_nt_20091207 VALUES ('GQ404587_M2', 'swine', 'M2', 'H1N1', 'USA', 1968, 291, 'AUGAGCCUUCUAACCGAGGU
INSERT INTO influ_A_nt_20091207 VALUES ('GQ404620_M1', 'swine', 'M1', 'H1N1', 'Germany', 1993, 756, 'AUGAGUCUUCUAACCC
INSERT INTO influ_A_nt_20091207 VALUES ('CY021709_HA', 'human', 'HA', 'H1N1', 'USA', 1945, 1701, 'AUGAAGGCAAGACUACUGC
INSERT INTO influ_A_nt_20091207 VALUES ('CY021710_M1', 'human', 'M1', 'H1N1', 'USA', 1945, 759, 'AUGAGUCUUCUAACCGAGGU
INSERT INTO influ_A_nt_20091207 VALUES ('CY021710_M2', 'human', 'M2', 'H1N1', 'USA', 1945, 294, 'AUGAGUCUUCUAACCGAGGU
INSERT INTO influ_A_nt_20091207 VALUES ('CY021711_NA', 'human', 'NA', 'H1N1', 'USA', 1945, 1410, 'AUGAAUCCAAAUCAGAAAA
INSERT INTO influ_A_nt_20091207 VALUES ('CY021712_NP', 'human', 'NP', 'H1N1', 'USA', 1945, 1497, 'AUGGCGUCCCAAGGCACCA
```





Practice #1

Extract sequence
using Entrez from Biopython

What is Biopython..?

- ✓ Biopython is a python library to make the bioinformatics works easier.
- ✓ Using this, you can import necessary information from bioinformatics file formats (FASTA, FASTAQ, BAM, VCF..) and utilize the tools used in bioinformatics.

Practice #1: Call Entrez in Python

1. Step 1: import Entrez
   - from Bio import Entrez

2. Step 2: enter your e-mail
   - Entrez.email = ”your_email_address"

Practice #1: Call ID list using esearch

3. Step 3: Call esearch to find IDs
   - keyWord = "avian influenza"
   - minDate='2021/11/01'
   - maxDate='2021/11/01'
   - search_results = Entrez.esearch(db = "nucleotide", term = keyWord, mindate=minDate , maxdate=maxDate)

Practice #1: Get ID list using esearch

3. Step 4: get a list of IDs out of esearch
   - records = Entrez.read(search_results)
   - print(records.keys())

   - identifiers = records['IdList']
   - count = int(records["Count"])
   - print("* Found %i results" % count)

Count	IdList
RetMax	TranslationSet
RetStart	TranslationStack
QueryKey	QueryTranslation
WebEnv	

Practice #1: Get Sequence using efetch

4. Step 5: use efetch to retrieve entries
   - fetch_handle = Entrez.efetch(db="nucleotide",id = identifiers, rettype="fasta", retmode="text", retmax=str(count))
   - result_text = fetch_handle.read()

```
>OK205887.1 Influenza A virus (A/chicken/Veracruz/28159-398/1995(H5N2)) segment 8
nuclear export protein (NEP) and nonstructural protein 1 (NS1) genes, complete cds
AGCAAAAGCAGGGTGACAAATACATAATGGACTCCAACACGATAACCTCGTTTCAGGTAGATTGCTATCT
ATGGCACATACGAAAGCTACTCAGCATGAGAGACATGTGTGATGCTCCCTTTGATGACAGACTCAGAAGA
GACCAAAAGGCATTGAAGGGAAGAGGCAGCACACTTGGACTCGACCTGCGAGCGGCCACAATAGAAGGCA
AAAAGATTGTTGAAGACATCCTAAAGAGTGAGACGGATGAAATTCTCAAAATTGCAATTGCATCCAGCCC
TGCTCCTCGGTATATTACCGATATGAGCATAGAGGAAATAAGCAGGGAATGGTACATGCTCATGCCAAGG
CAGAAAATAACAGGAGGCCTGATAGTAAAAATGGATCAGGCCATCATGGACAAGAGAATAATACTCAAGG
CAAACTTCTCTGTCCTATTTGATCAACTGGAAACATTAGTCTCACTGAGGGCTTTCACAGACGATGGCGC
CATTGTAGCTGAAATATCTCCCTTTCCTTCTATGCCAGGACATTCTACAGAGGATGTCAAAATTGCAATT
GGAATCCTCATCGGTGGACTTGAATGGAATGATAACTCAATTCGAGCGTCTGAAAATATACAGAGATTTG
CTTGGGGAGTCCGTGATGAGAATAGGGGACCTTCACTCCCTCCAAAGCAGAAACGCTACATGGCGAGAAG
AATTGAGTCAAAAGTTTGAAGAGATCAGATGGCTAATTGCAGAGTGTAGAAACATATTAACCAAAACTGA
GAACAGCTTCGAGCAGATAACGTTCTTGCAGGCATTGCAACTCTTACTTGAAGTTGAGAGTGAGATAAGG
ACATTTTCTTTTCAGCTTATTTAGTACTAAAAAACACCCTTGTTTCTACT

>OK205886.1 Influenza A virus (A/chicken/Veracruz/28159-398/1995(H5N2)) segment 7
matrix protein 2 (M2) and matrix protein 1 (M1) genes, complete cds
AGCAAAAGCAGGTAGATATTGAAAGATGAGTCTTCTAACCGAGGTCGAAACGTACGTTCTCTCTATCGTC
CCGTCAGGCCCCCTCAAAGCCGAGATCGCGCAGAGACTTGAAGATGTCTTTGCAGGGAAGAACACCGATC
TTGAGGCACTCATGGAATGGCTAAAGACAAGACCAATCCTGTCACCTCTGACTAGGGGGATTTTAGGATT
TGTGTTCACGCTCACCGTGCCCAGTGAGCGAGGACTGCAGCGTAGACGCTTTGTCCAAAATGCCCTTAAT
GGGAATGGGGATCCAAACAACATGGACAGAGCGGTCAAACTGTACAGGAAGCTAAAAAGGGAAATAACAT
TCCATGGGGCAAAAGAAGTGGCACTCAGTTACTCAGCTGGTGCACTTGCCAGTTGCATGGGCCTCATATA
CAACAGAATGGGGACTGTGACCACCGAAGTGGCATTTGGCCTGGTGTGCGCCACATGTGAGCAAATTGCT
GATTCCCAGCATCGGTCTCACAGACAAATGGTGACAACGACCAACCCACTGATTAGACATGAAAACAGAA
TGGTACTGGCCAGTACTACGGCAAAAGCCATGGAGCAAATGGCAGGGTCAAGTGAACAGGCAGCAGAGGC
TATGGAGGTTGCTAGTCAGGCTAGACAGATGGTGCAGGCAATGAGGACCATTGGAACCCATCCTAGCTCC
AGTGCTGGTCTAAAAGATGATCTTCTTGAAAATTTGCAGGCCTACCAGAAACGGATGGGAGTGCAAATAC
AGCGATTCAAGTGATACTCTCGTTATTGCCGCAAGCATCATTGGGATCCTGCACTTGATATTGTGGATTC
TTGATCGTCTTTTCTTCAAATGCATTTATCGTCGCTTTAAATACGGTTTGAAAAGAGGGCCTTCTACGGA
AGGAGTGCCTGAATCTATGAGGGAAGAATATCGGCAGGAACAGCAGAGTGCTGTGGATGTTGACGATGGT
CATTTTGTCAACATAGAGCTGGAGTAAAAAACTACCTTGTTTCTACT

>OK205885.1 Influenza A virus (A/chicken/Veracruz/28159-398/1995(H5N2)) segment 6
neuraminidase (NA) gene, complete cds
AGCAAAAGCAGGAGTGAAGATGAATCCAAATCAGAAGATAATAACAATTGGCTCCGTCTCTCTAACCATT
GCAACAGTATGTTTCCTCATGCAGATTGCCATCTTAACAACGACTGTGACACTGCATTTCAAGCAAAACG
AATGCAGCATCCCCGCTAACAACCAAGTAGTGCCATGTGAACCAATCATAATAGAGAGGAACATAGACTA
```

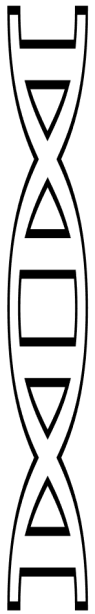
Practice #2

Pilot study using
influenza sequences

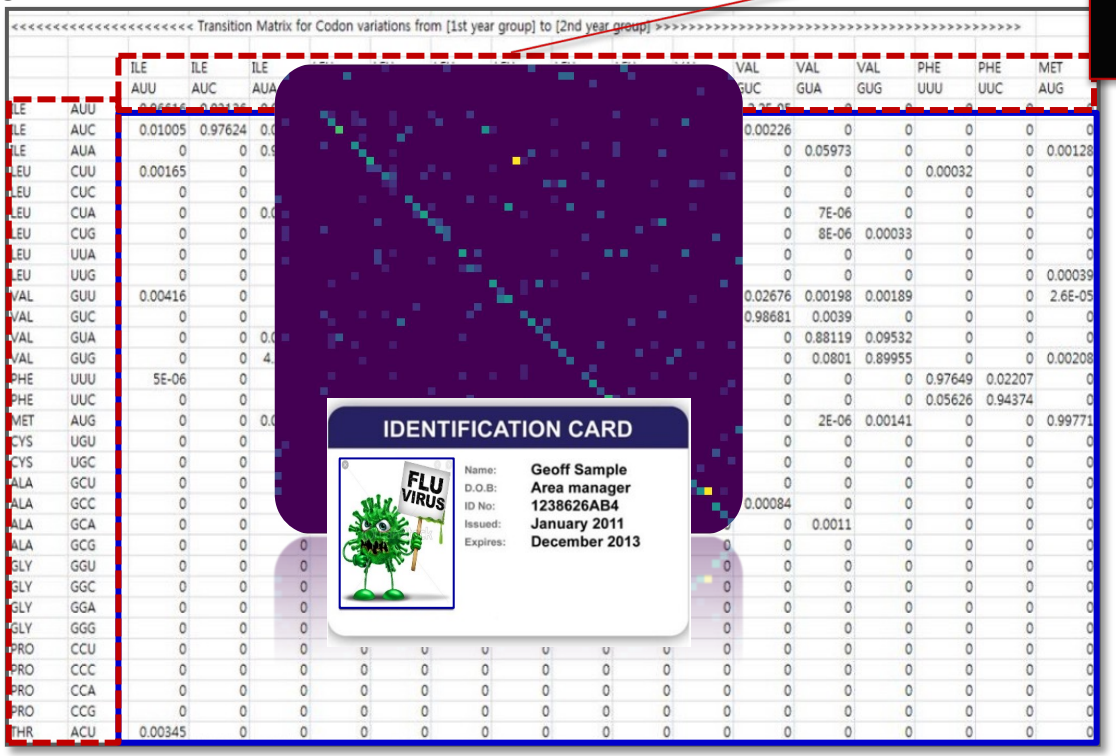
Create Variant vector

Reference sequence

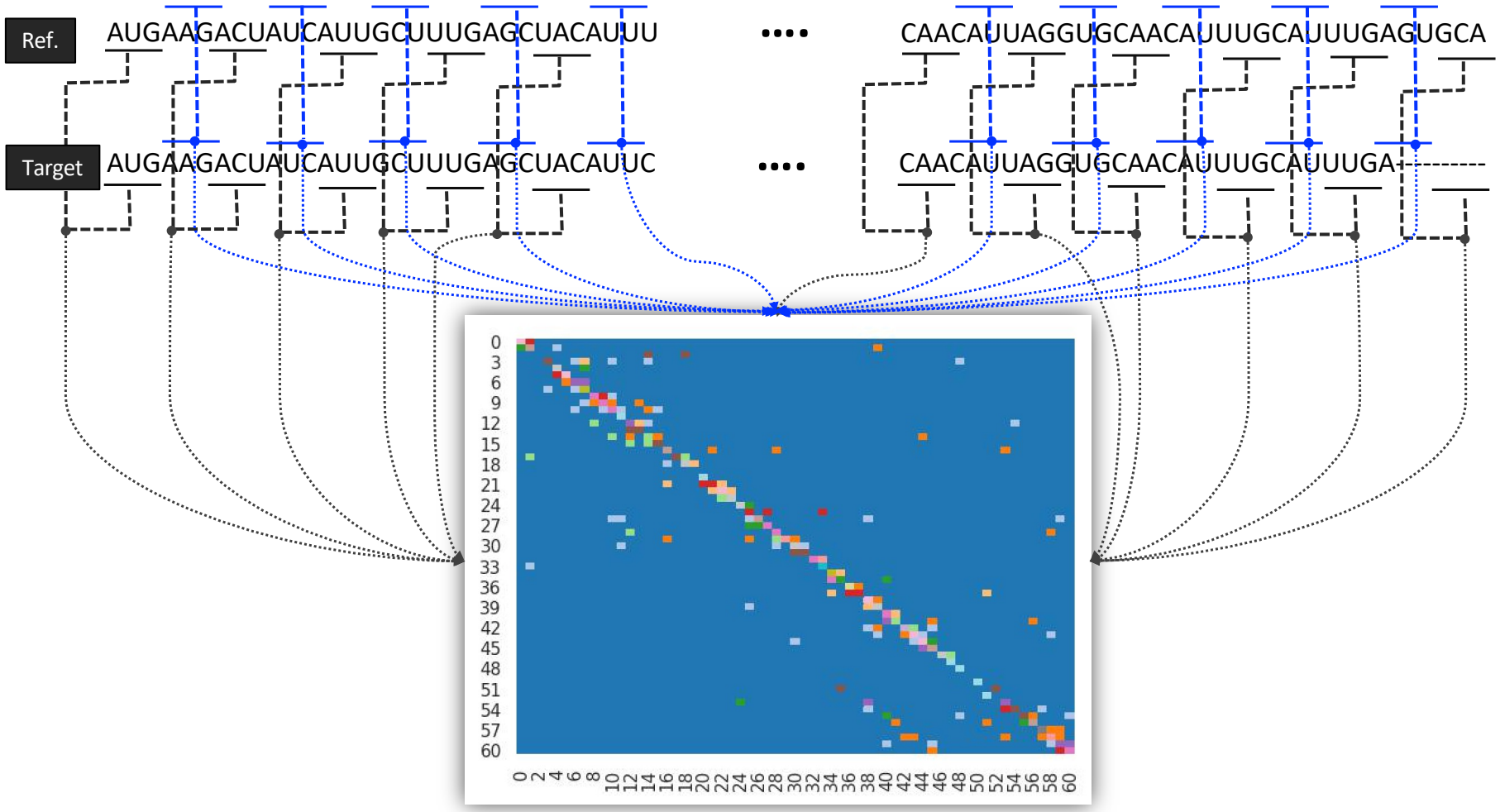
Target sequence



Comparison of variants at the codon (3 bases) level



Create Variant vector



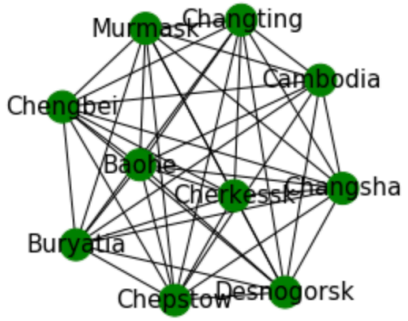
What is synonymous codon..?

First base	Second base: U	Second base: C	Second base: A	Second base: G	Third base
U	UUU, UUC: Phenylalanine (Phe)	UCU, UCC, UCA, UCG: Serine (Ser)	UAU, UAC: Tyrosine (Tyr)	UGU, UGC: Cysteine (Cys)	U
	UUA, UUG: Leucine (Leu)		UAA, UAG: Stop Codon	UGA: Stop Codon; UGG: Tryptophan (Trp)	C, A, G
C	CUU, CUC, CUA, CUG: Leucine (Leu)	CCU, CCC, CCA, CCG: Proline (Pro)	CAU, CAC: Histidine (His)	CGU, CGC, CGA, CGG: Arginine (Arg)	U, C, A, G
			CAA, CAG: Glutamine (Gln)		
A	AUU, AUC, AUA: Isoleucine (Ile)	ACU, ACC, ACA, ACG: Threonine (Thr)	AAU, AAC: Asparagine (Asn)	AGU, AGC: Serine (Ser)	U, C
	AUG: Methionine (Met) Start codon		AAA, AAG: Lysine (Lys)	AGA, AGG: Arginine (Arg)	A, G
G	GUU, GUC, GUA, GUG: Valine (Val)	GCU, GCC, GCA, GCG: Alanine (Ala)	GAU, GAC: Aspartic acid (Asp)	GGU, GGC, GGA, GGG: Glycine (Gly)	U, C, A, G
			GAA, GAG: Glutamic acid (Glu)		

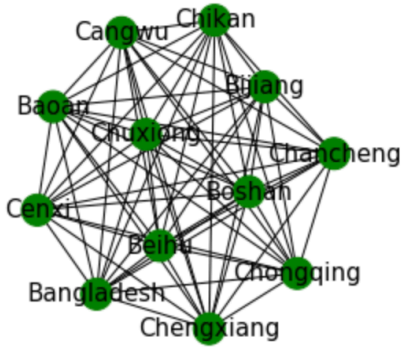
https://www.researchgate.net/figure/Synonymous-Codons-of-20-Amino-Acids_fig2_324469014

Tracking the spread of influenza between countries through cluster analysis

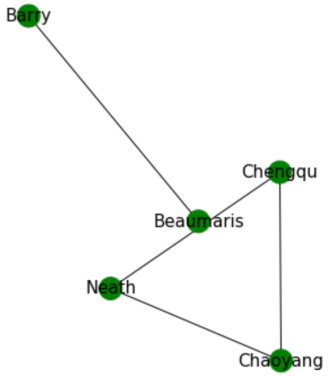
H3N2 2019 (Correlation Coefficient > 0.9)



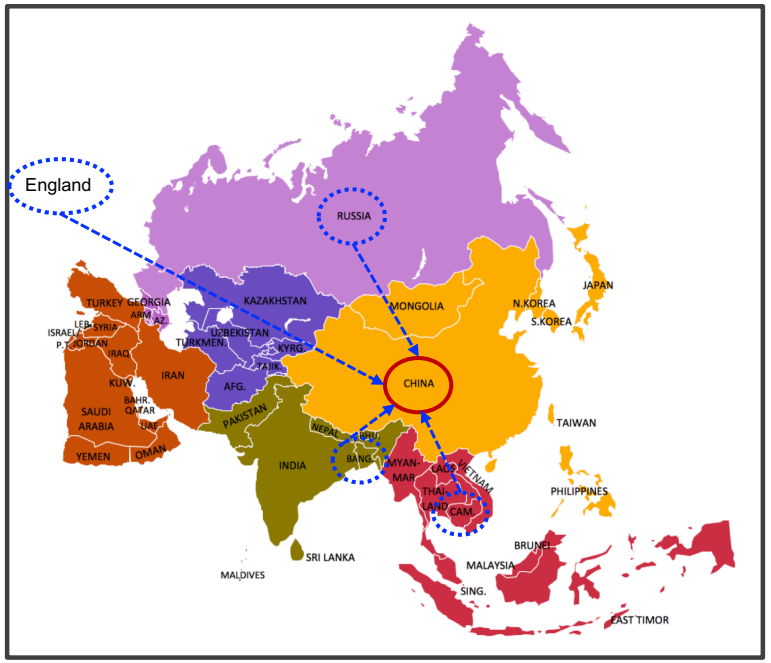
China & Russia & Cambodia



China & Bangladesh

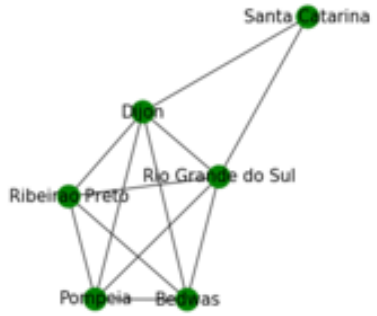


China & England

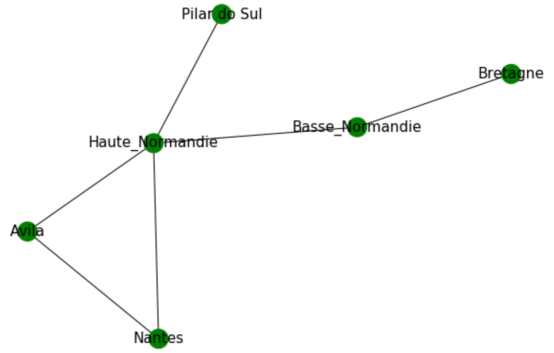


Tracking the spread of influenza between countries through cluster analysis

H3N2 2019 (Correlation Coefficient > 0.9)

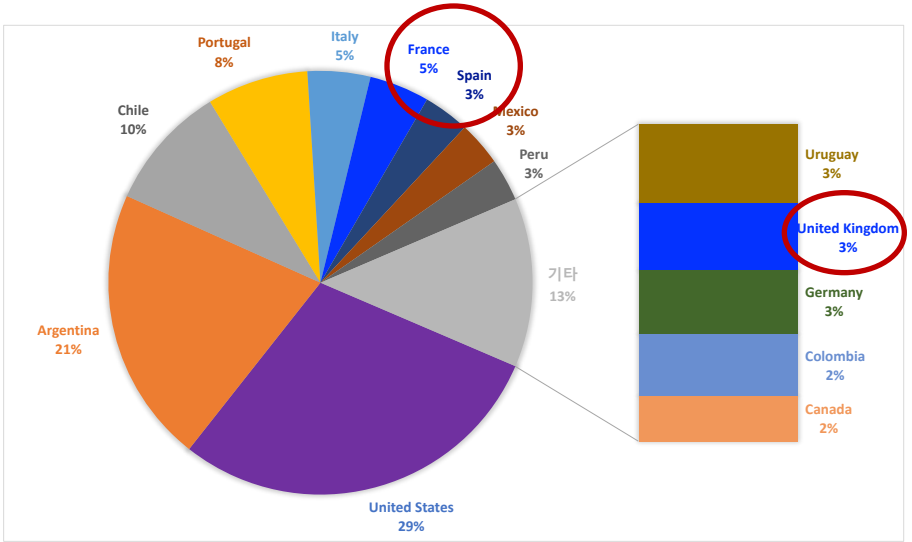


Brazil & France & England



Brazil & Spain & France

Southern America & Western Europe



Assignments

Assignments

1. **Install** biopython library in python and try crawling various data including sequence data from NCBI.
2. **Design** a pilot study related to COVID-19 to be conducted using the big data and technologies learned during the remainder of the semester.
3. **Armanc** from KIST, **Chen Jingyu** from KRICT, please prepare a 5-minute presentation next week.

THANK YOU