# Introduction

CS482/682

Computational Techniques in Biological Sequence Analysis

# Course Logistics

- Instructor: Bin Ma (DC 3345, http://www.cs.uwaterloo.ca/~binma)
- TA: Ruisheng (Benson) Guo. https://cs.uwaterloo.ca/~r9guo/
- Course webpage: https://cs.uwaterloo.ca/~binma/cs482
- No required textbooks.
- Lecture notes will be provided.
- Class attendance is required.

# Grades

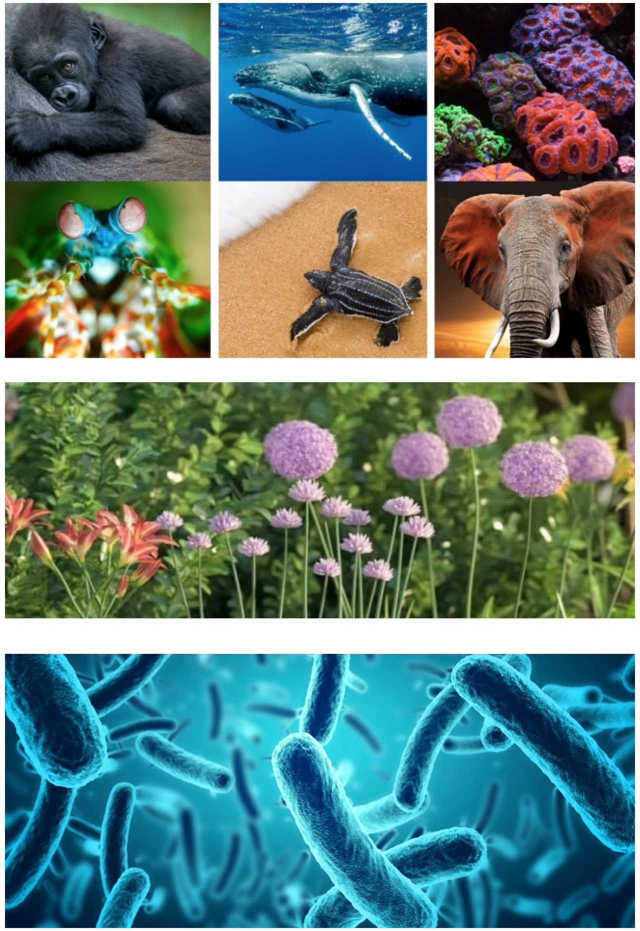
- 4 assignments:  60% = 15+15+15+15
- Final exam: 40%

# Bioinformatics

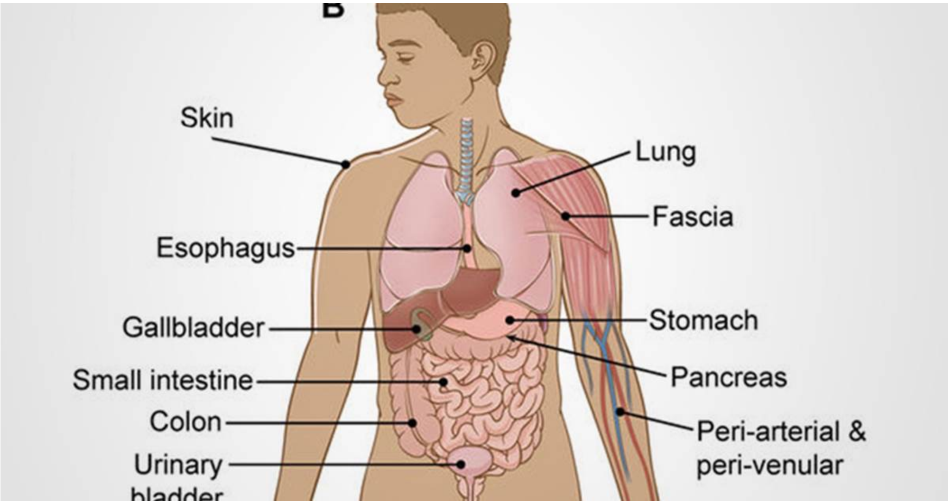
Biology: the reason, goal, purpose.

Informatics: the method.

# Biology Can be Studied at Different Scales

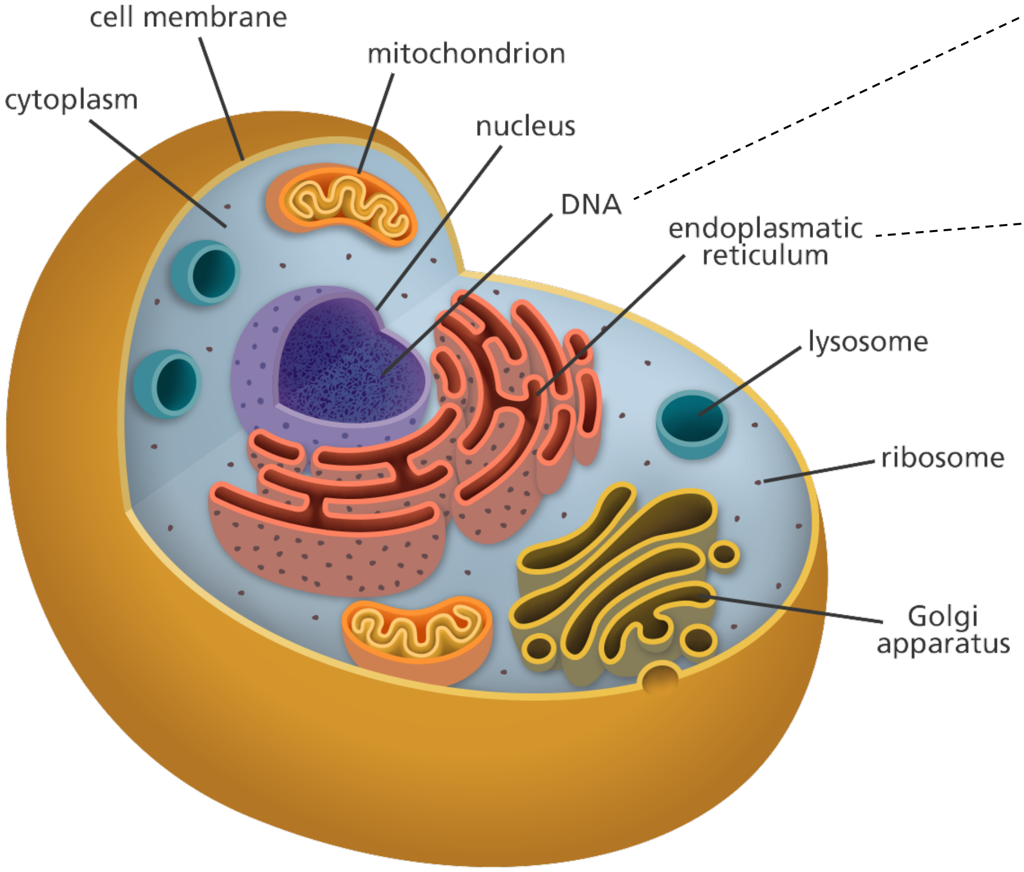


Organisms: living things



Organs and tissues

# Cell Level

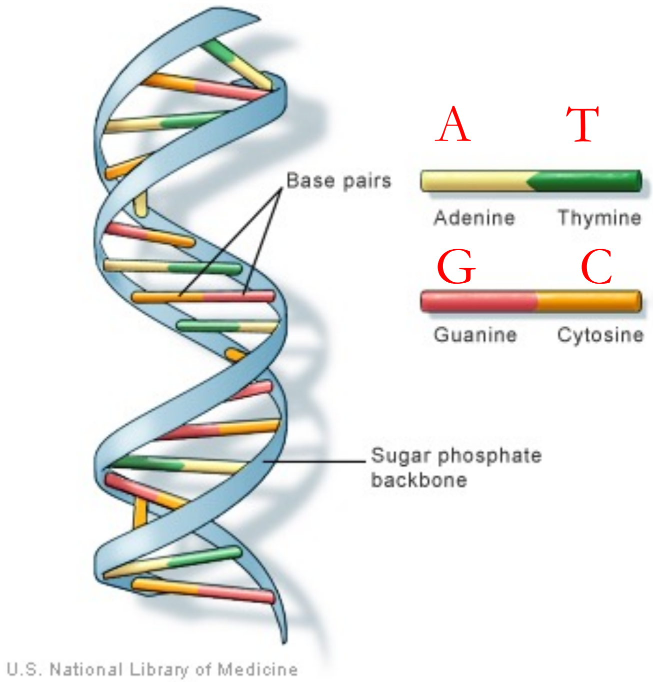


The Animal Cell

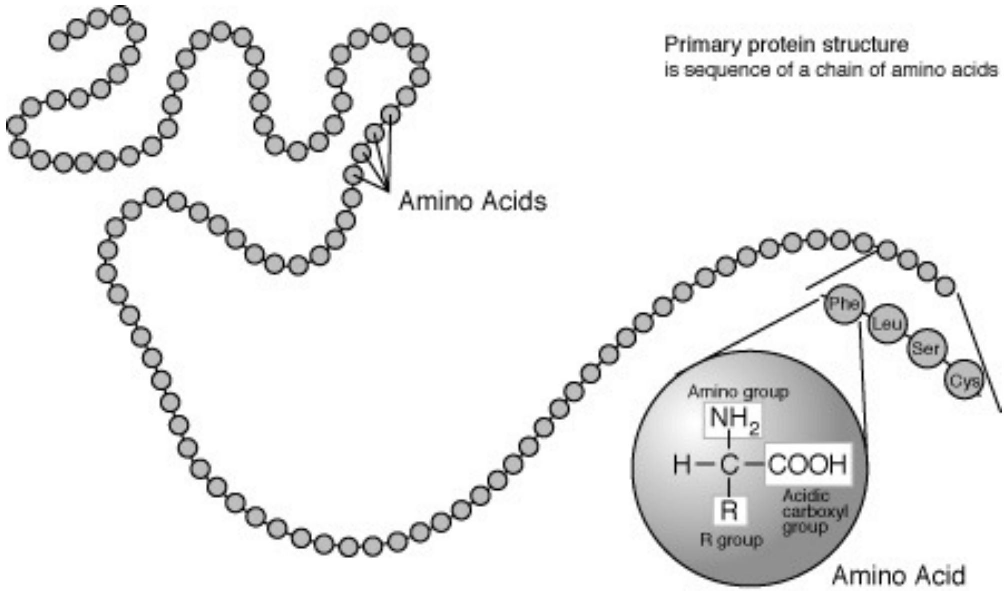
Genetic information storage

Production of functional proteins from genetic info.

# Molecular Level



RNA



DNA: chain of nucleotide bases

Protein: chain of amino acids.

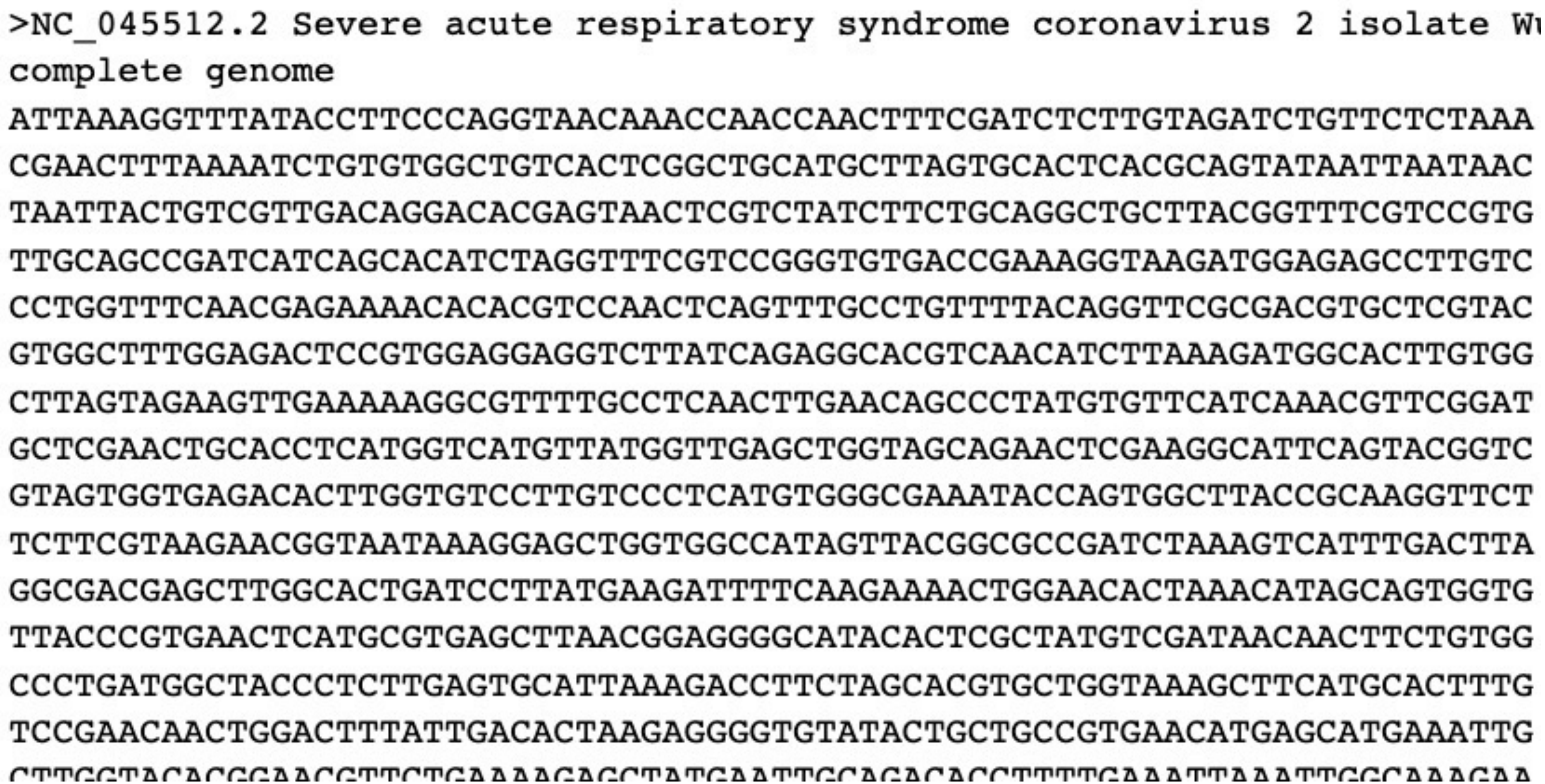
**…GCTTACACGTCACCAT…**

**…LVQSGAEVKKP…**

# Public Molecular Data

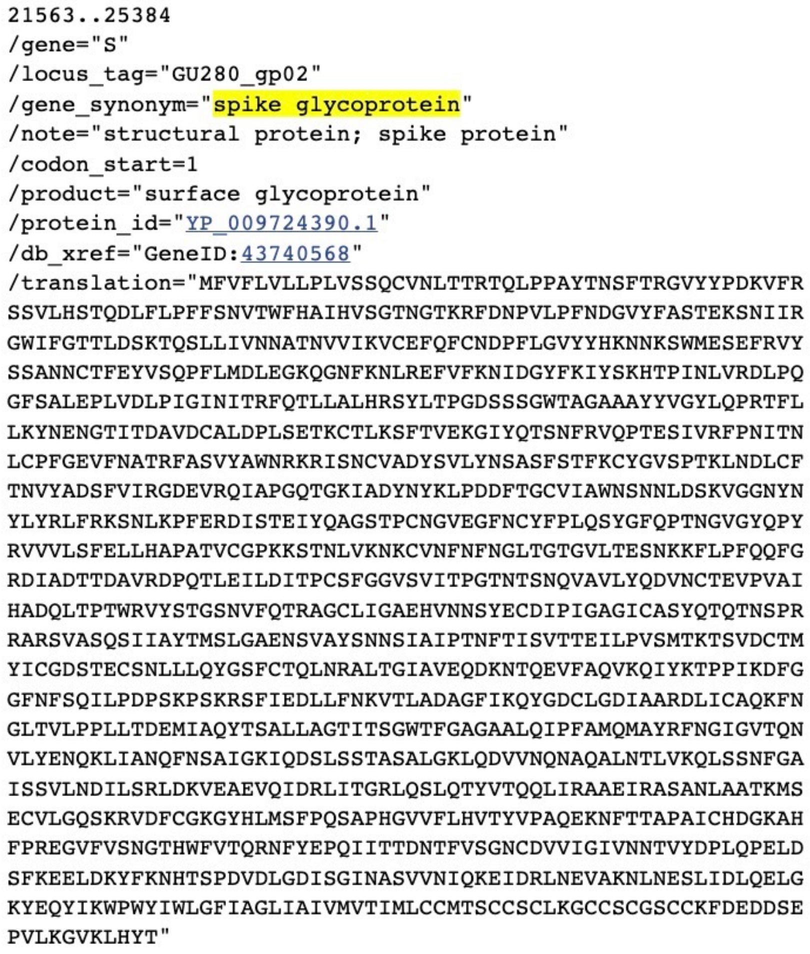
- There are tremendous amount of public biomolecule data and free software.
- NCBI’s sequence data bank:
  - E.g. https://www.ncbi.nlm.nih.gov/nuccore/NC_045512
- PDB protein structure database
  - E.g. https://www.rcsb.org/structure/6vxx

# DNA

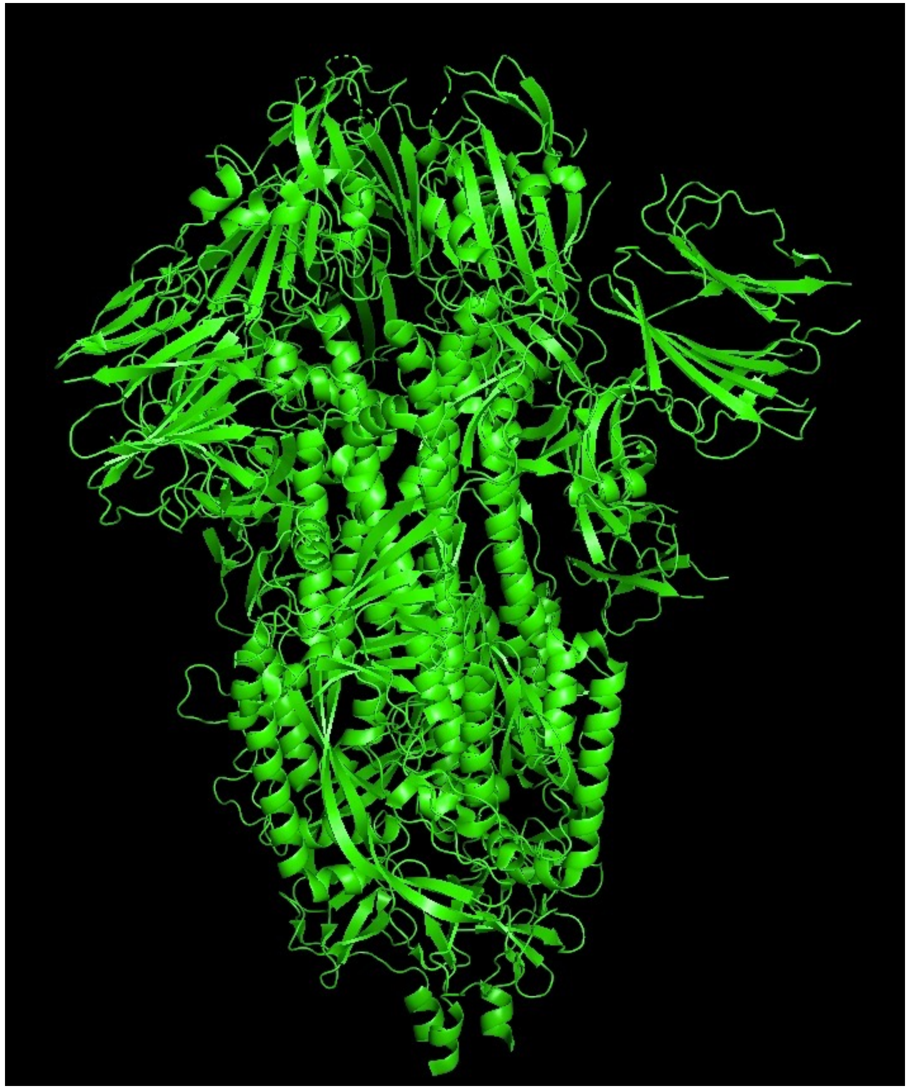


```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wu
complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
```

# Protein Sequence and Structure



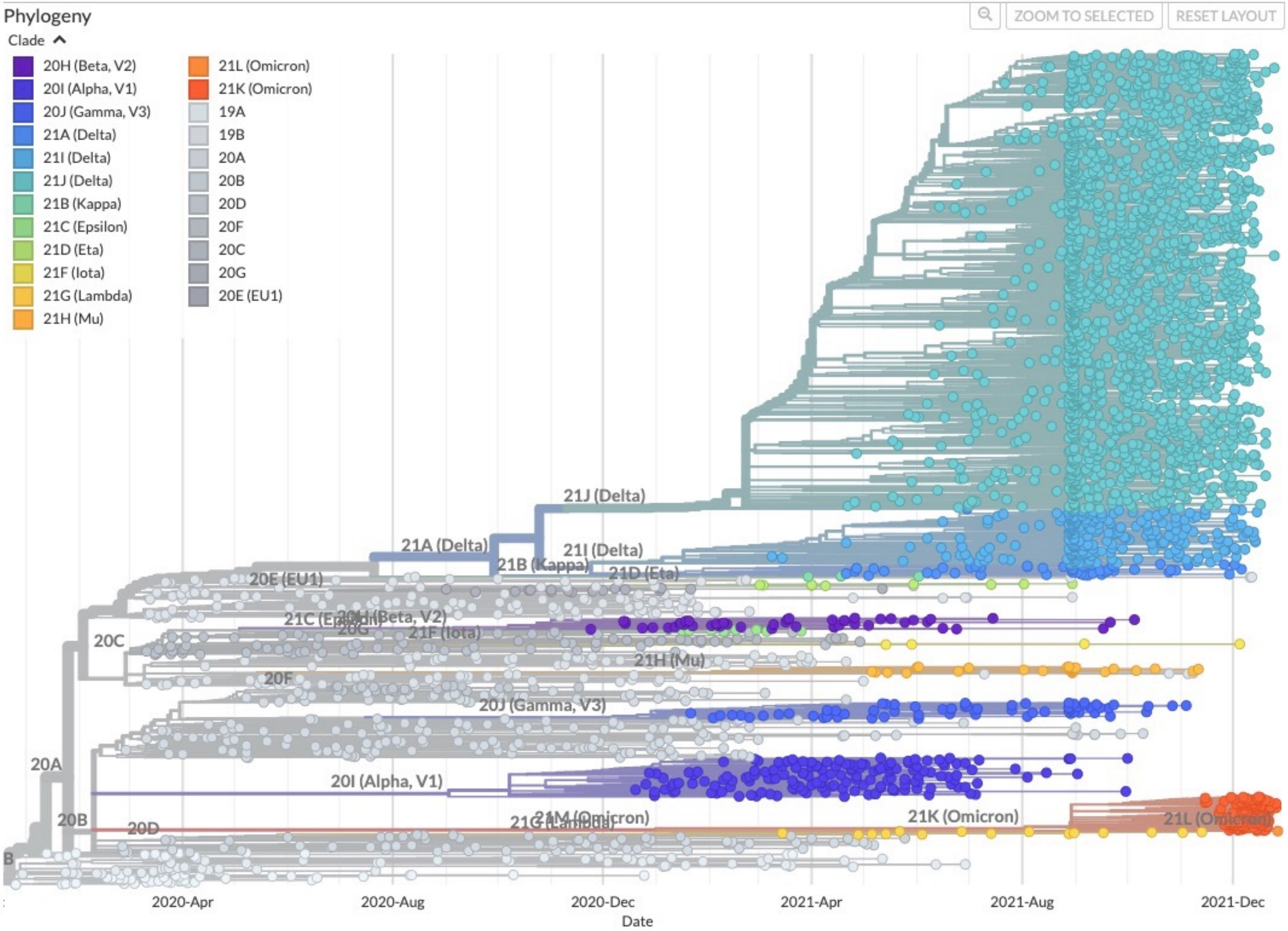
```
21563..25384
/gene="S"
/locus_tag="GU280_gp02"
/gene_synonym="spike glycoprotein"
/note="structural protein; spike protein"
/codon_start=1
/product="surface glycoprotein"
/protein_id="YP_009724390.1"
/db_xref="GeneID:43740568"
/translation="MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFR
SSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIR
GWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVY
SSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQ
GFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFL
LKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITN
LCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY
RVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG
RDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAI
HADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPR
RARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTM
YICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG
GFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFN
GLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQN
VLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGA
ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMS
ECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELD
SFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELG
KYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSE
PVLKGVKLHYT"
```



https://www.rcsb.org/structure/6vxx

# Why Do People Do Bioinformatics?

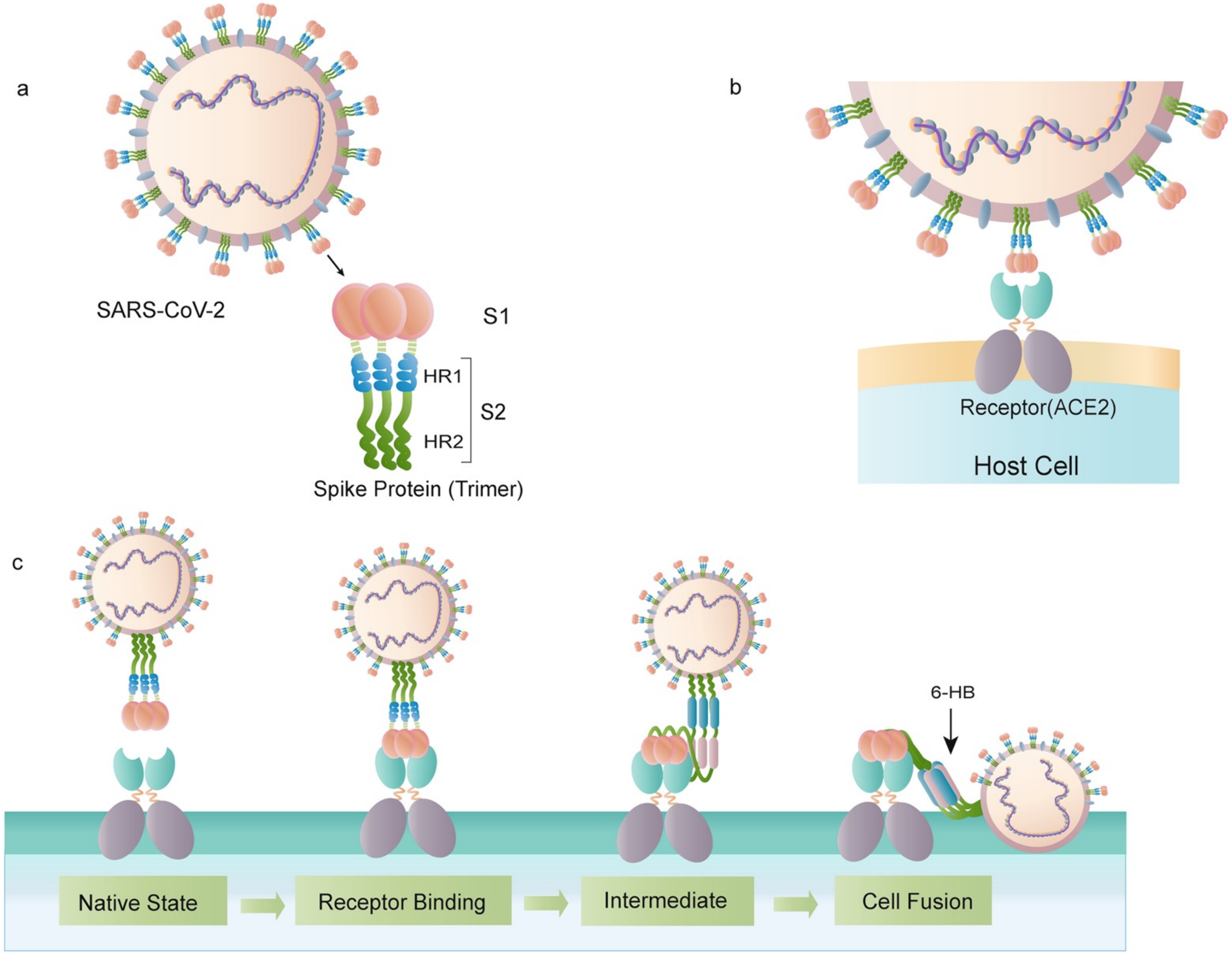
- Understand life at molecular level
- Human health.
  - E.g. Sequencing SARS-Cov2 genomes allowed people study the evolution of this virus.
  - E.g. Study the structure of the spike protein and its iteraction with the host cells
  - E.g. A lot of human diseases are related to genetics.



Phylogeny tree of 3475 SARS-Cov2 genomes sampled between Dec 2019 and Dec 2021. Image credit: https://nextstrain.org/

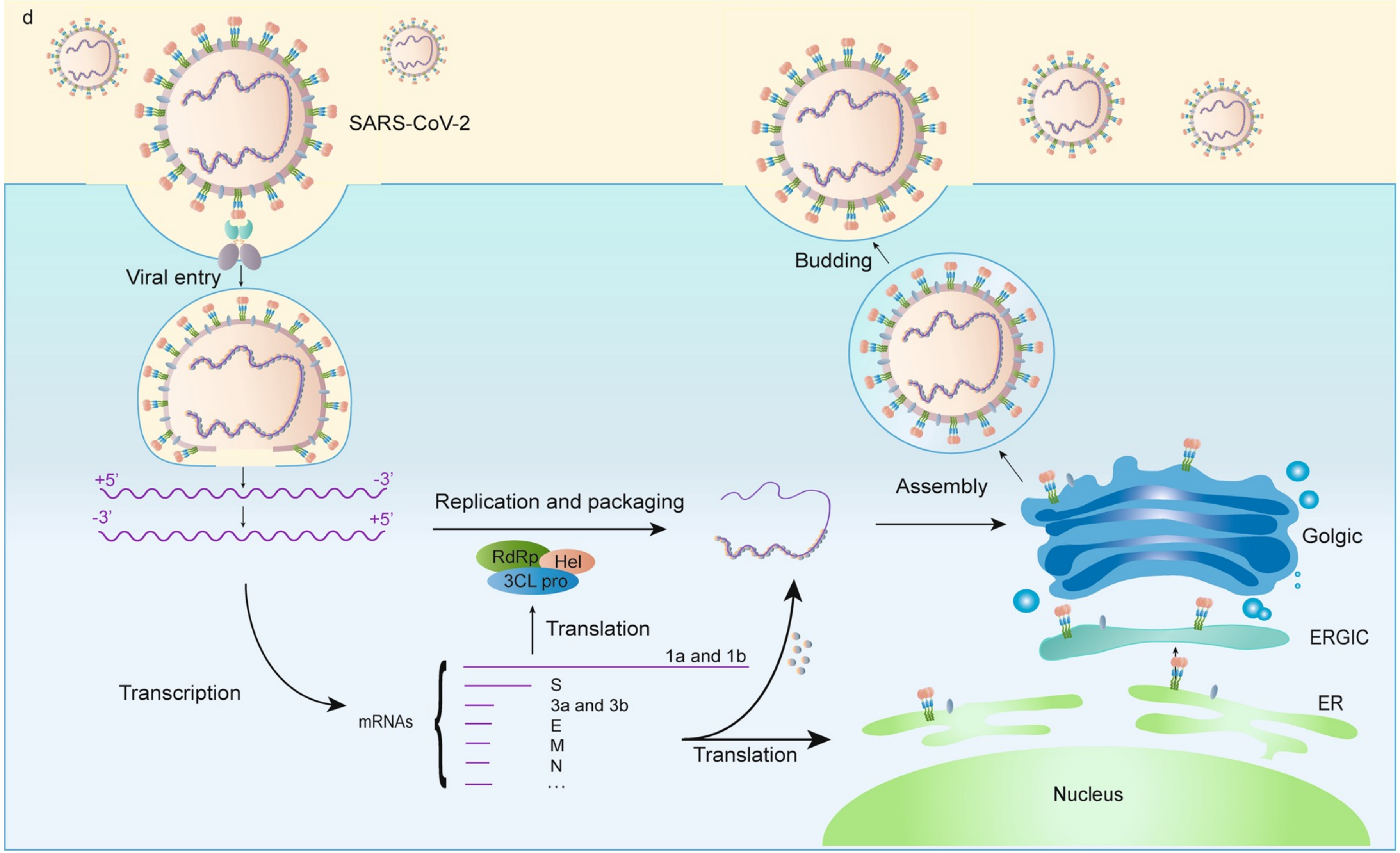
- Sequencing
- Comparison
- Phylogeny

# How SARS-Cov2 Invades Human Cells



- Gene prediction
- Protein identification
- Structure prediction
- Protein-protein interaction
- Structure determination

# How SARS-Cov2 Invades Human Cells



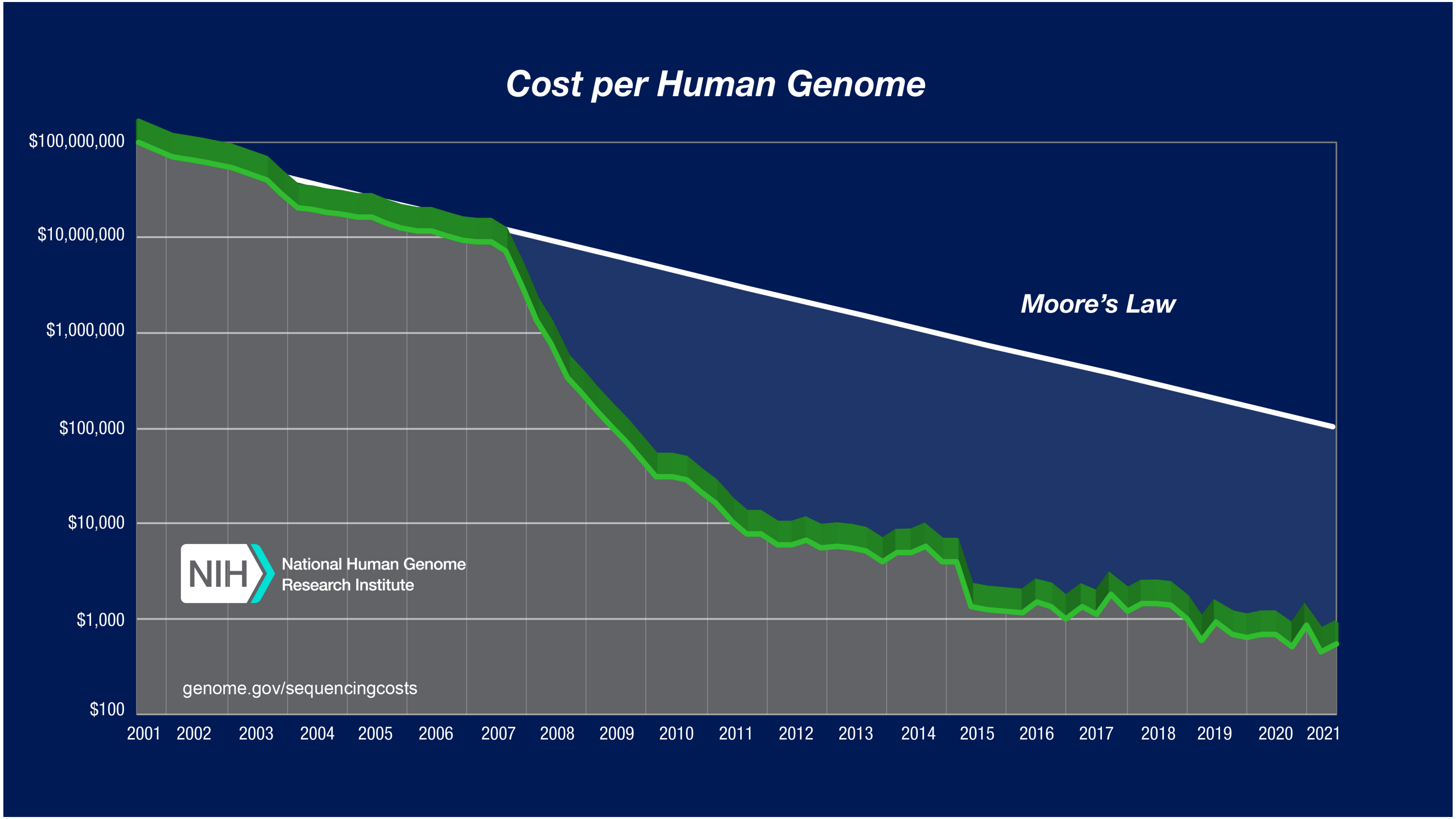
- Immune system
- Antigen &Antibody
- MHC & T-cells

# Bioinformatics

- Determine the molecule information
  - Through analyzing the data produced by measuring instruments
  - Usually in large scale and high throughput
- Use the molecular data to make inference

# Difference Made by Bioinformatics

- Example: genome sequencing.
- Human Genome Project
  - 3B$ from 1990-2003 to study human genome.
  - identify all the approximately 20,000-25,000 genes in human DNA,
  - determine the sequences of the 3 billion chemical base pairs that make up human DNA.
  - Bioinformatics played an essential role in analyzing the data and assemble the genome.
- Today one can sequence a human's genome with <1000$ in a couple of weeks. Bioinformatics is the key to utilize the NGS (next generation sequencing) data for genome sequencing.
- As such, today's cancer treatment starts to become *personalized*. And many new drugs now require gene sequencing as companion diagnostic.


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts

# Objectives of This Course

- Know bioinformatics
  - Purpose and method
  - General topics
- Learn classic problems and **algorithms** in bioinformatics
- Learn wide-applicable computational techniques
  - String algorithms
  - Hidden Markov Model
  - Log likelihood ratio score
  - Statistical validation
  - A bit of machine learnine

# A Typical Problem

- Human genome has ~ 3G base pairs (letters).

```
>A substring of the genome
GCTTACACGTCACCATCTGTGCCACCACCCATGTCTCTAGTGAT
CCCTCATAAGTTCCAACAAAGTTTGCGAGTACTCAACACCAACA
TTGATGGGCAATGGAAAATAGCCTTCGCCATCACACCATTAAGG
GTGATGTTGAGGAAAGCAGACATTGACCTCACCGAGAGGGCAGG
CGAGCTCAGGTAGGATGAGGTGGAGCATATGATCACCATCATAC
AGAACTCACCAAGATTCCAGACTGGTTC
```

- Only 1-2% encode proteins (genes).
  - where are they?

# An Analog

- Find the English words.

```
fbjpsikocxltfestkvdvjiixjsasisxmhbqpvpwb
ulfddurluvwrritrbsbhcpeyhbekydaibmwyfntj
nwvporabwuahvsdgknpkzihjqagrpspixtzqphhk
tvwbioinformaticsisusefulcrsqibcadosyvoz
vhuzdxabqrjzfagiysqfcmyrkqdytjtjbusysqga
etyllzbinma@uwaterloo.cawpxuprgokixkoiyv
```

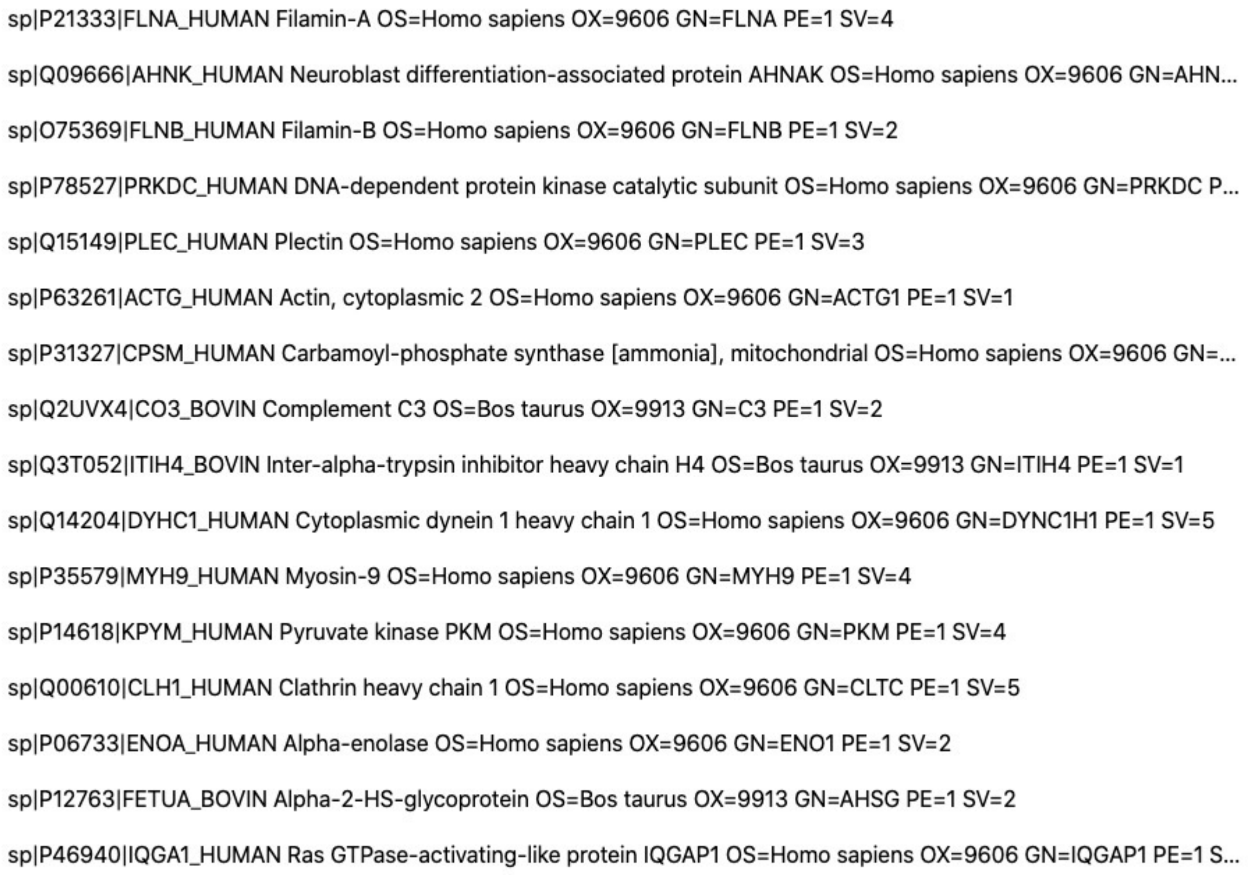
# Another Typical Problem

- Find the longest shared substring between human and mouse genomes
  - Each has $3\times10^9$ base pairs.
  - Cannot afford $3\times10^9\times3\times10^9$ comparisons.

# Longest Common Substring

- Longest Common Substring can be done in linear time!
- What if all similarities (instead of exact matches) are to be found?
  - This leads to the homology search problem.
  - Some of the key techniques are developed by profs in this school.

# A Third Typical Problem



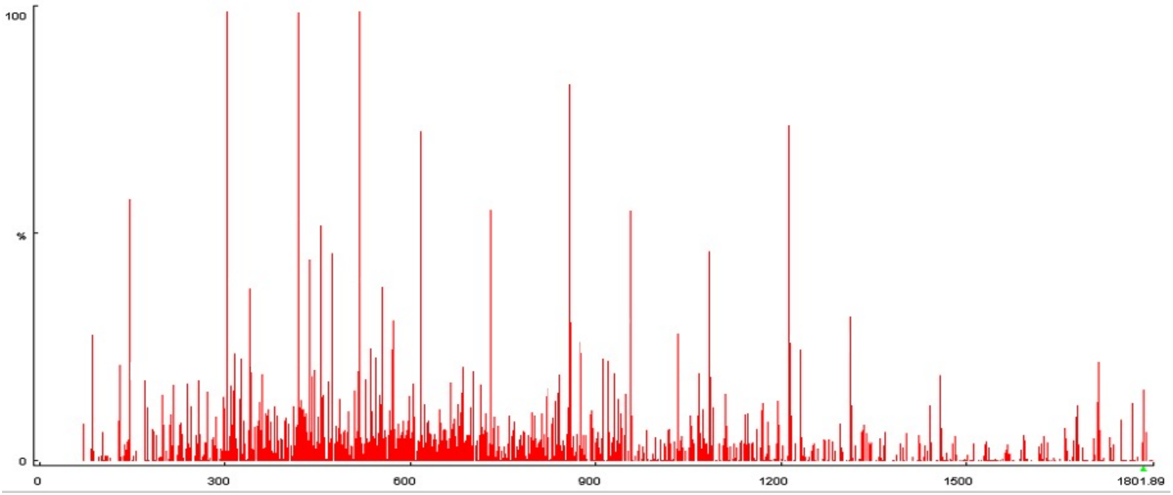
```
sp|P21333|FLNA_HUMAN Filamin-A OS=Homo sapiens OX=9606 GN=FLNA PE=1 SV=4
sp|Q09666|AHNK_HUMAN Neuroblast differentiation-associated protein AHNAK OS=Homo sapiens OX=9606 GN=AHN...
sp|O75369|FLNB_HUMAN Filamin-B OS=Homo sapiens OX=9606 GN=FLNB PE=1 SV=2
sp|P78527|PRKDC_HUMAN DNA-dependent protein kinase catalytic subunit OS=Homo sapiens OX=9606 GN=PRKDC P...
sp|Q15149|PLEC_HUMAN Plectin OS=Homo sapiens OX=9606 GN=PLEC PE=1 SV=3
sp|P63261|ACTG_HUMAN Actin, cytoplasmic 2 OS=Homo sapiens OX=9606 GN=ACTG1 PE=1 SV=1
sp|P31327|CPSM_HUMAN Carbamoyl-phosphate synthase [ammonia], mitochondrial OS=Homo sapiens OX=9606 GN=...
sp|Q2UVX4|CO3_BOVIN Complement C3 OS=Bos taurus OX=9913 GN=C3 PE=1 SV=2
sp|Q3T052|ITIH4_BOVIN Inter-alpha-trypsin inhibitor heavy chain H4 OS=Bos taurus OX=9913 GN=ITIH4 PE=1 SV=1
sp|Q14204|DYHC1_HUMAN Cytoplasmic dynein 1 heavy chain 1 OS=Homo sapiens OX=9606 GN=DYNC1H1 PE=1 SV=5
sp|P35579|MYH9_HUMAN Myosin-9 OS=Homo sapiens OX=9606 GN=MYH9 PE=1 SV=4
sp|P14618|KPYM_HUMAN Pyruvate kinase PKM OS=Homo sapiens OX=9606 GN=PKM PE=1 SV=4
sp|Q00610|CLH1_HUMAN Clathrin heavy chain 1 OS=Homo sapiens OX=9606 GN=CLTC PE=1 SV=5
sp|P06733|ENOA_HUMAN Alpha-enolase OS=Homo sapiens OX=9606 GN=ENO1 PE=1 SV=2
sp|P12763|FETUA_BOVIN Alpha-2-HS-glycoprotein OS=Bos taurus OX=9913 GN=AHSG PE=1 SV=2
sp|P46940|IQGA1_HUMAN Ras GTPase-activating-like protein IQGAP1 OS=Homo sapiens OX=9606 GN=IQGAP1 PE=1 S...
```

# Peptide Identification

*peptide sequence:* LGSSEVEQVQLVVDGVK

*tandem mass spectrometry:*

*MS/MS spectrum*



# Assignments

- All are programming assignments. You submit source code and a half-page document.
  - Based on your own work. Use of library needs to be documented.
- Evaluation is mostly based on correctness and the performance of the program (speed, accuracy, etc.).

# Assignments

- 1. Pairwise sequence alignment and COVID variants assignment
  - Every Bioinformatics course does this.
- 2. Is it a natural/real peptide?
  - A taste of scoring and machine learning.
- 3. Peptide identification from mass spec.
- 4. Predict structure of SARS-Cov2 spike protein.

# Read More

- Modern molecular biology studies a few types of biologically important molecules: DNA, RNA, protein, lipid, glycan
- Bioinformatics has mostly studied DNA, then RNA and protein.
  - Because they are “easier”
  - their primary structures are sequences.
  - Also because the measuring technology has been developed.
- If you don’t have much biology background, read the following articles (and other related articles) from Wikipedia: Protein, DNA, RNA, gene, genome, genetic code.
- We will also <u>briefly</u> review the necessary biology knowledge when needed.

# Summary

- We talked about:
- course logistics
- basic biology (wikipedia good resource)
- course topics

- Next time: sequence alignment