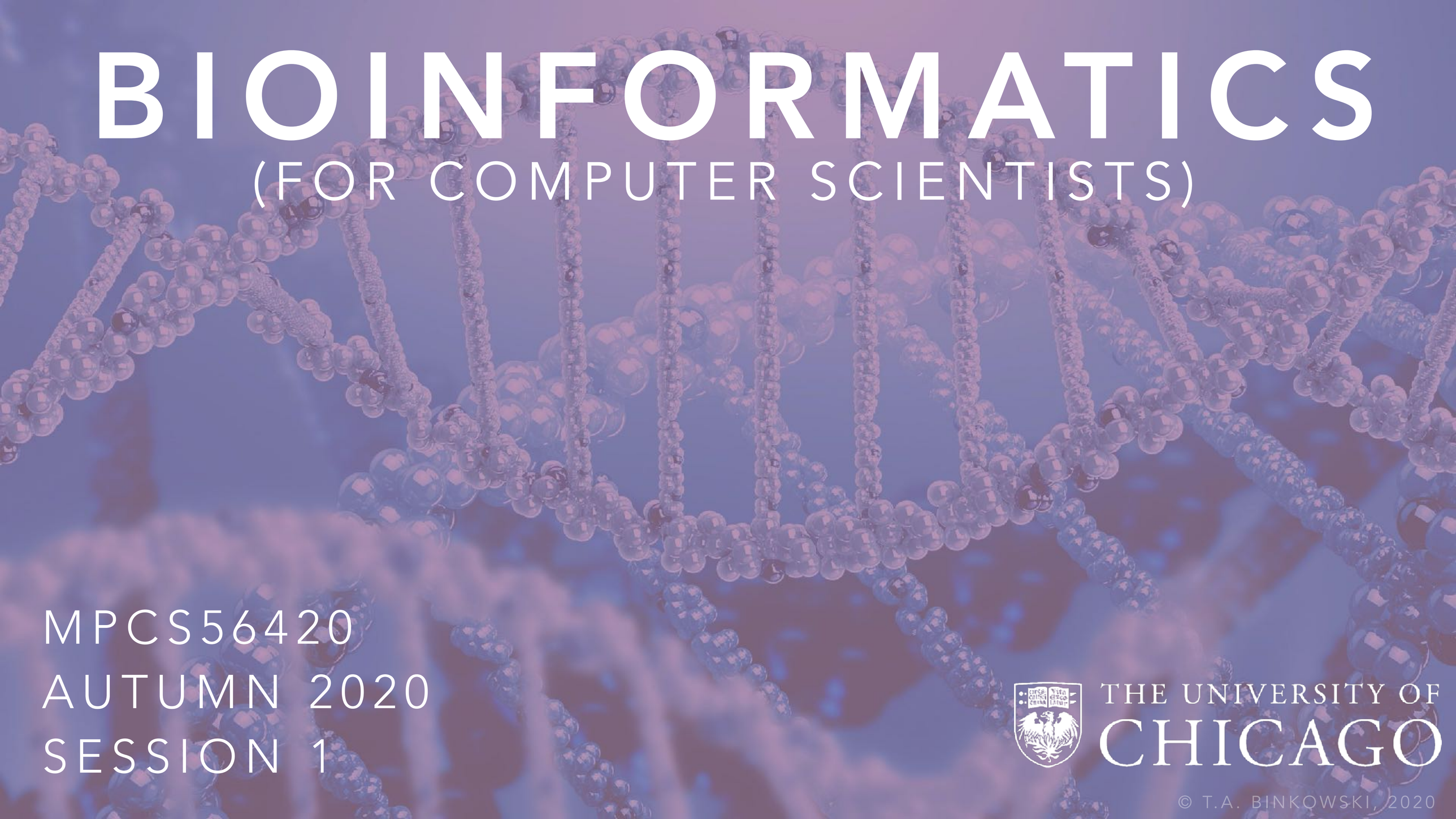
BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)



MPCS56420
AUTUMN 2020
SESSION 1

THE UNIVERSITY OF CHICAGO


WELCOME TO
BIOINFORMATICS

COURSE LOGISTICS

- Prerequisites
  - Algorithms
  - Core Programming
- Consent of instructor
  - Requires departmental approval

COURSE LOGISTICS

- Flipped Classroom
  - Videos of lectures
  - Meet for discussions and office hours

COURSE LOGISTICS

- Regular Class Meeting
  - Thursday 5:30-6:30pm
- Lecture Video Recap/Demos
  - Monday 1:00 (See slack poll)
- Office hours
  - Tuesday 5:30 (Neal Conrad)
  - Monday 1:30pm (Andrew) (See slack poll)

Introduce upcoming weeks material
Introduce/update assignment
Open Q and A

Student presentations

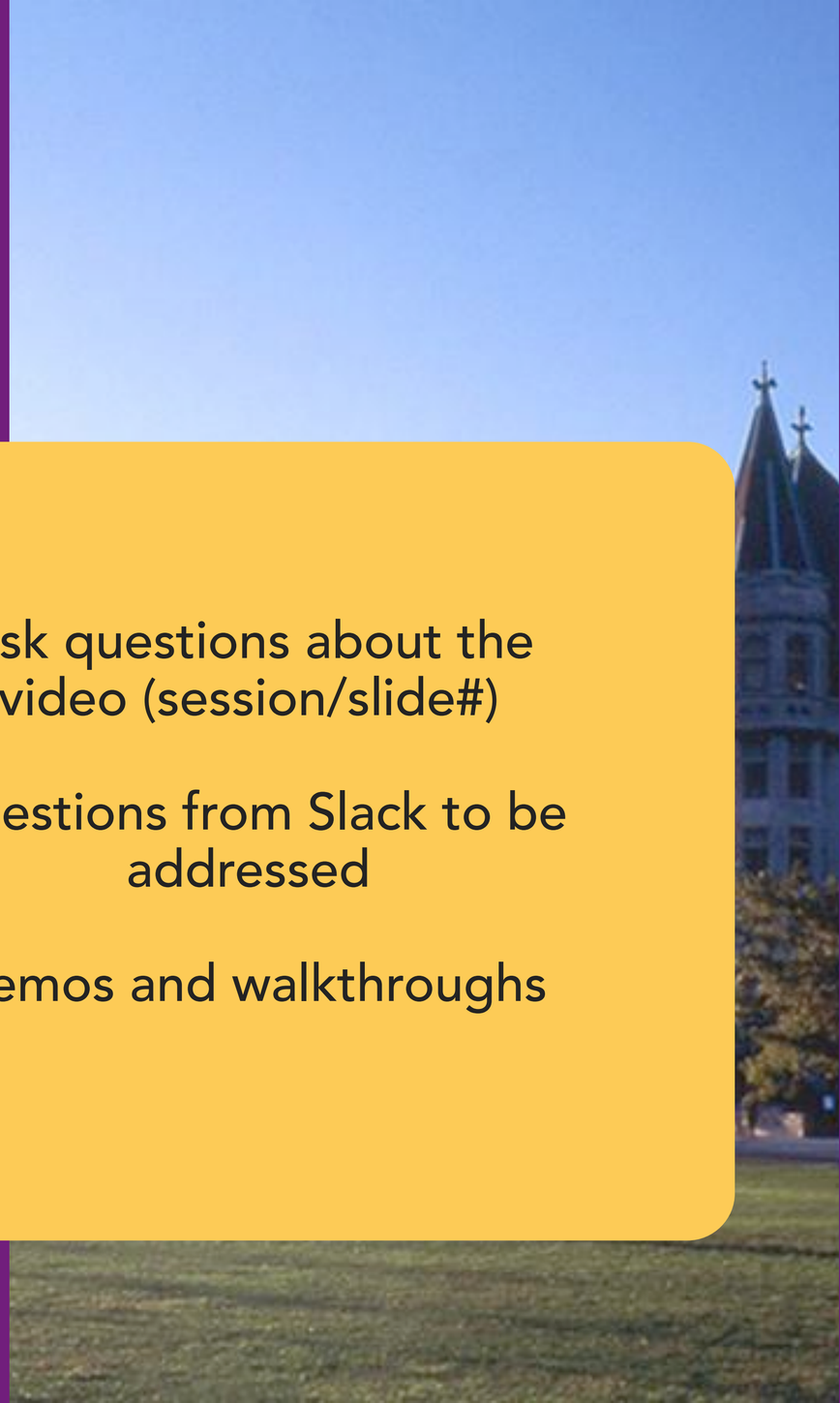
COURSE LOGISTICS

- Regular Class Meeting
  - Thursday 5:30-6:30pm
- Lecture Video Recap/Demos
  - Monday 1:00 (See slack poll)
- Office hours
  - Tuesday 5:30 (Neal Conrad)
  - Monday 1:30pm (Andrew) (See slack poll)

Ask questions about the video (session/slide#)

Questions from Slack to be addressed

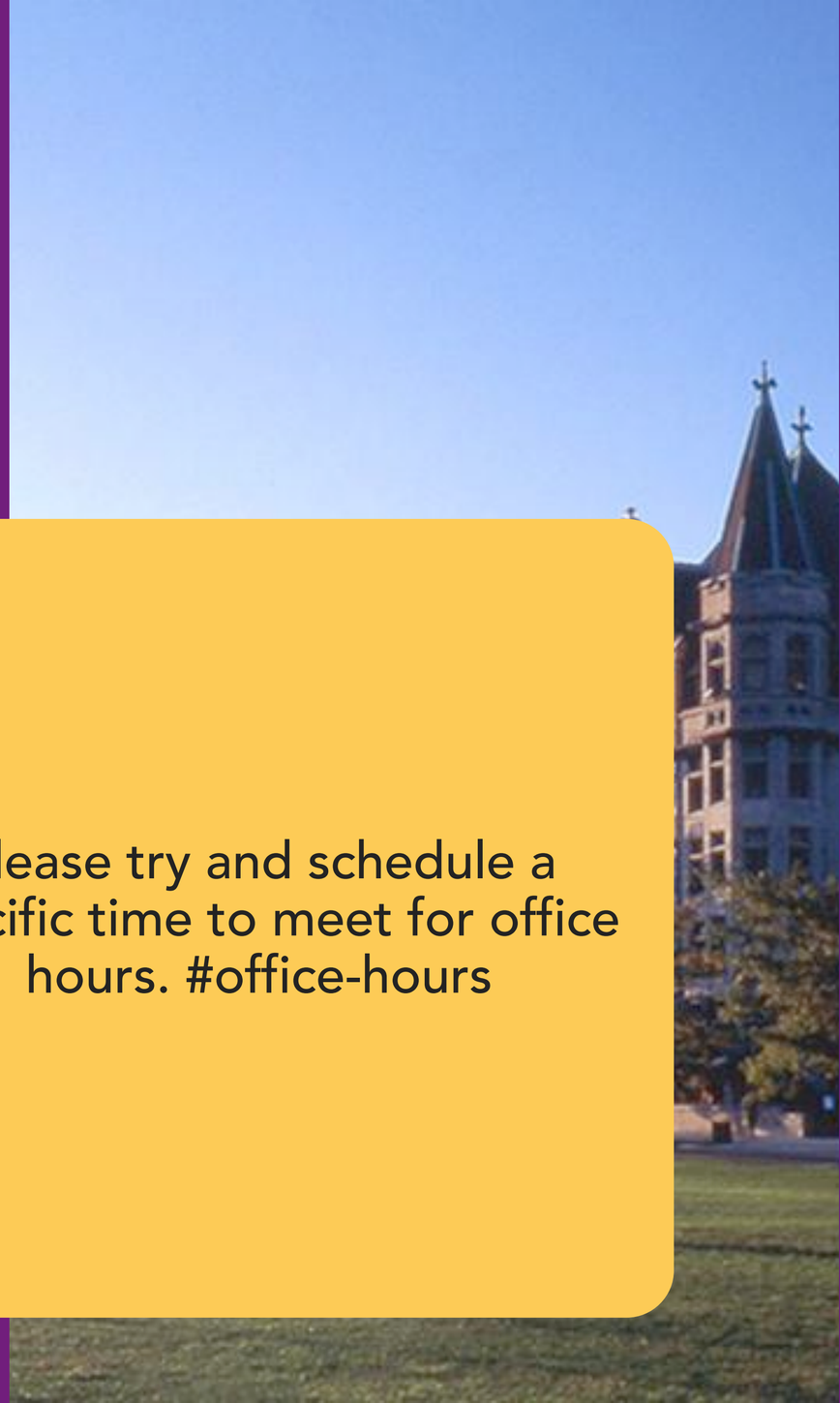
Demos and walkthroughs



COURSE LOGISTICS

- Regular Class Meeting
  - Thursday 5:30-6:30pm
- Lecture Video Recap/Demos
  - Monday 1:00 (See slack poll)
- Office hours
  - Tuesday 5:30 (Neal Conrad)
  - Monday 1:30pm (Andrew) (See slack poll)

Please try and schedule a specific time to meet for office hours. #office-hours

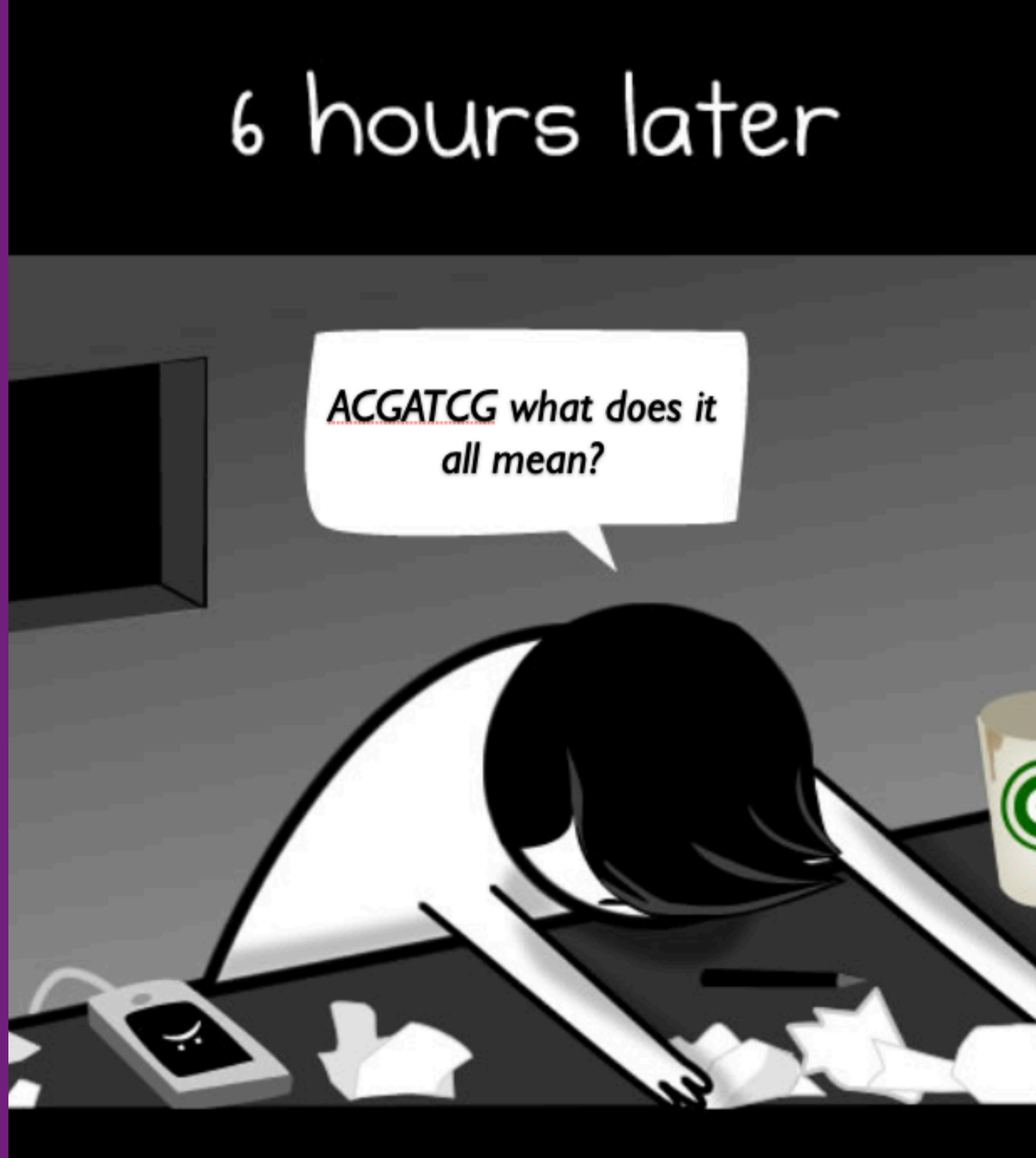


COURSE LOGISTICS

- No biology required
  - Weekly “need to know” basis
  - Class will adjust for student backgrounds
  - Online resources for learning biology
    - Khan Academy, EdX, Coursera, Udacity

COURSE LOGISTICS

- Biology is a learn-as-you-go
- Researchers study one gene for a lifetime



COURSE DESIGN

INSTRUCTORS

- Andrew
  - The University of Chicago
    - Center for Structural Genomics of Infectious Diseases
    - MPCS Program: Python, iOS, Advanced iOS, Bioinformatics, Practicum
  - ANL
    - Midwest Center for Structural Genomics, Leadership Computing Facility
    - Computation Institute

COURSE DESIGN

INSTRUCTORS

- Research programs
  - Structural Genomics, molecular modeling for drug discovery
  - HPC for bioinformatics pipelines; INCITE
  - STEM Education (K-12)

COURSE DESIGN

INSTRUCTORS

A BIOLOGIST TURNED COMPUTER SCIENTIST

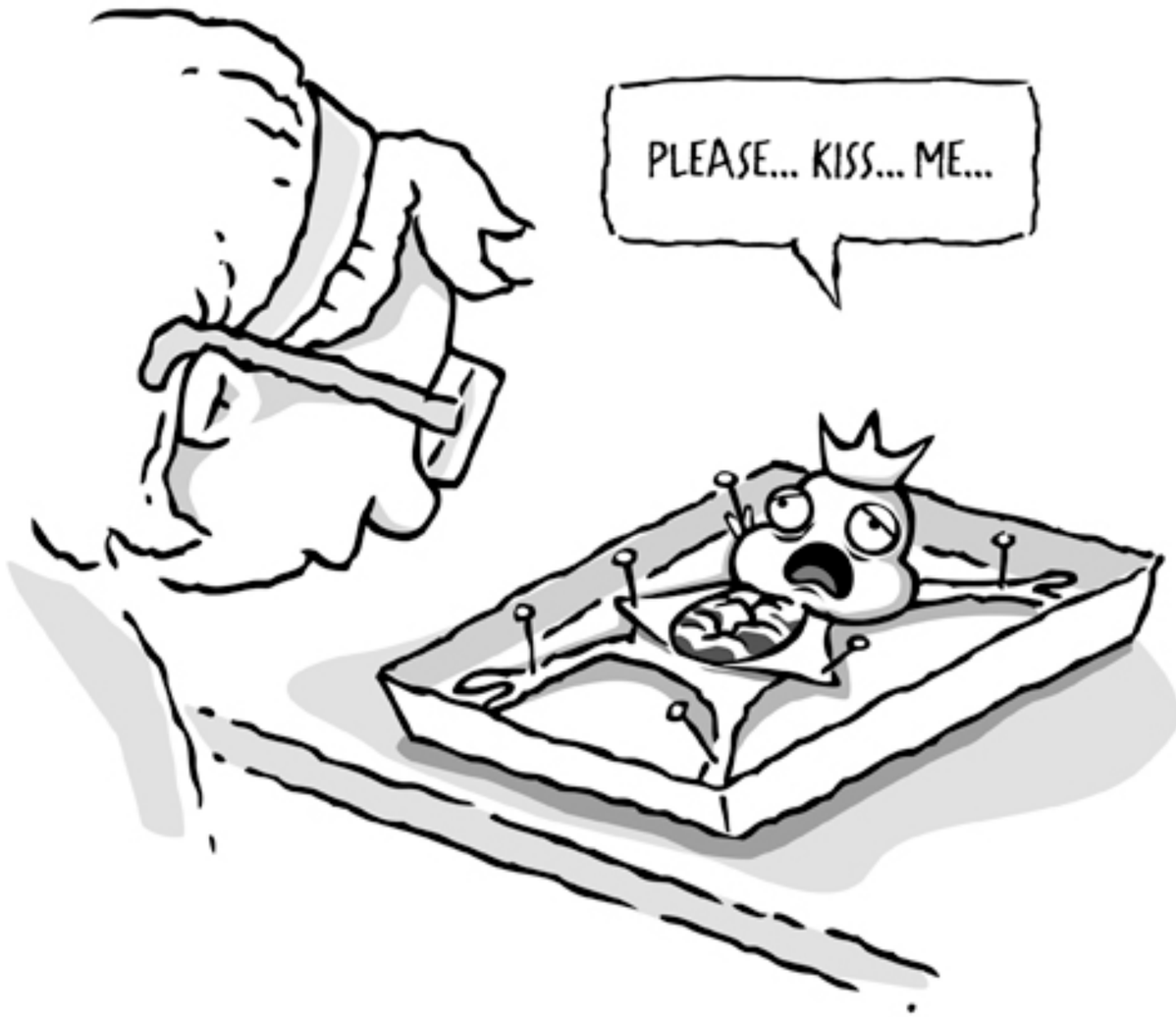
- Neal Conrad
  - Graduated UChicago in 2013 with a degree in Biology after working on a project related to the cytoskeleton and cancer
  - Started the masters program in 2014 to transition from wet lab research to data analysis
  - Graduated MCPS program in 2017
  - Semi-professional gambler

Course Design

COURSE DESIGN

CHALLENGES

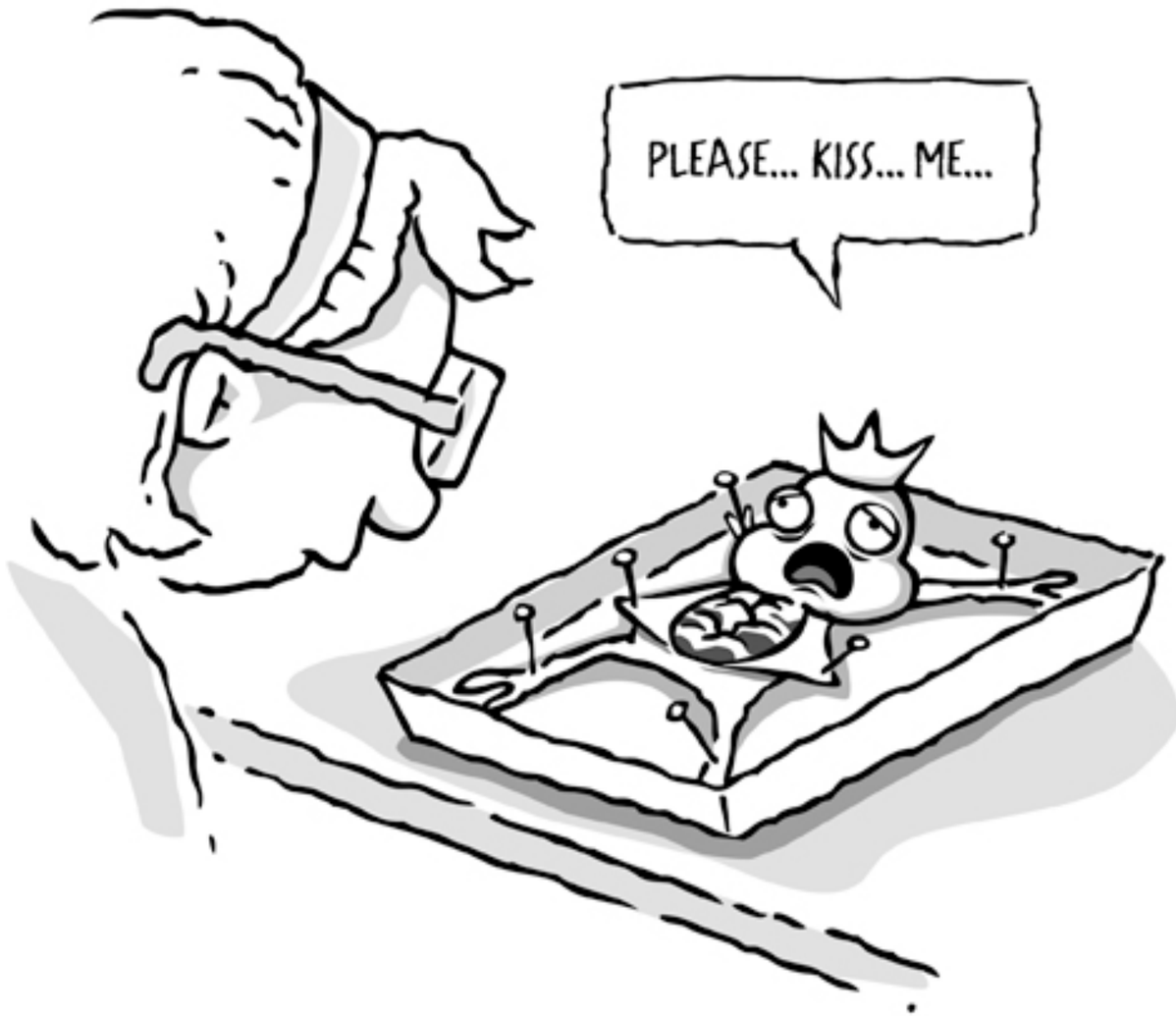
- People from different backgrounds
    - No biology since high school



COURSE DESIGN

CHALLENGES

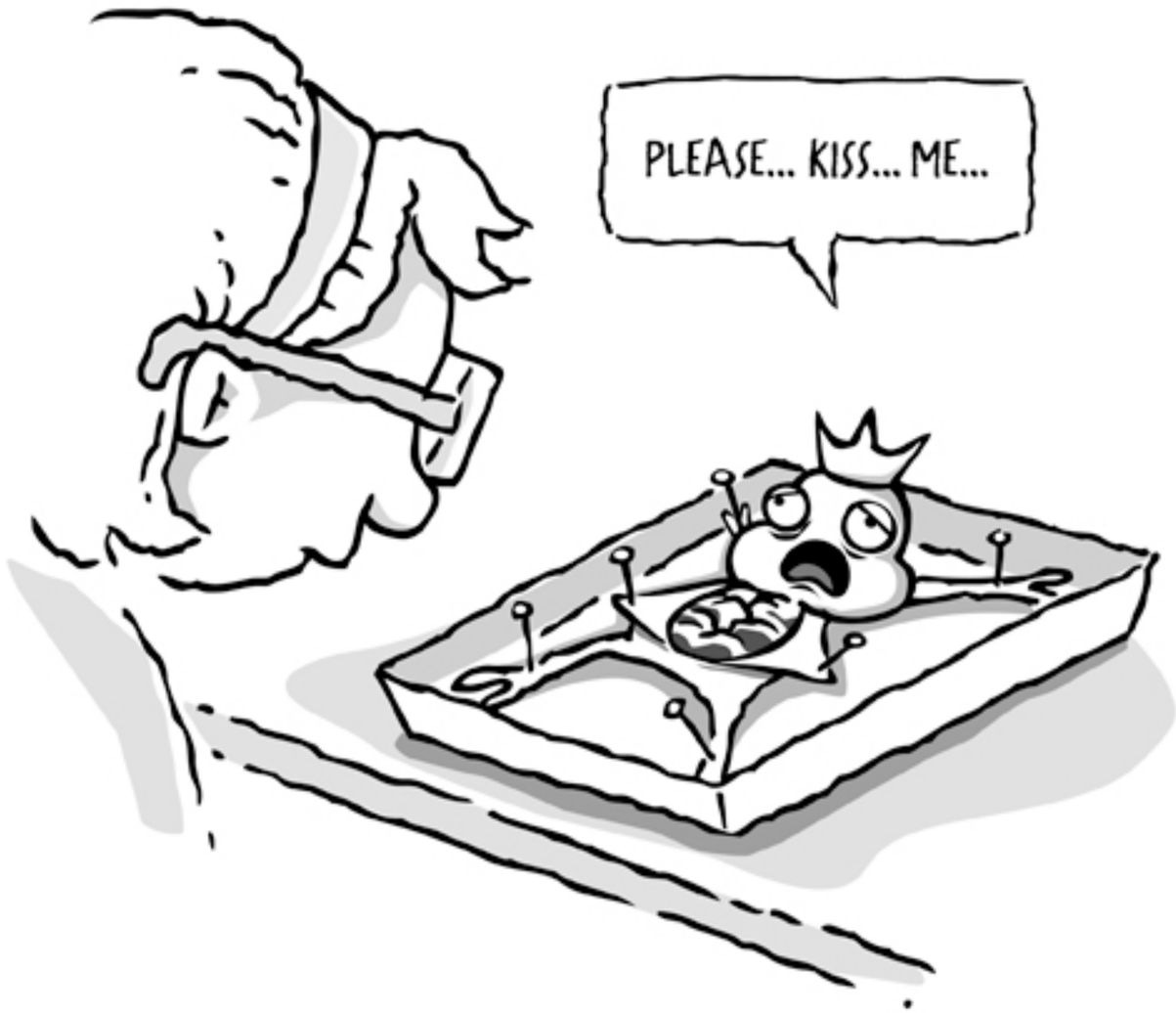
- Wide topic that touches many fields and sub-fields
  - Molecular biology, structural biology, genomics, medicine



COURSE DESIGN

CHALLENGES

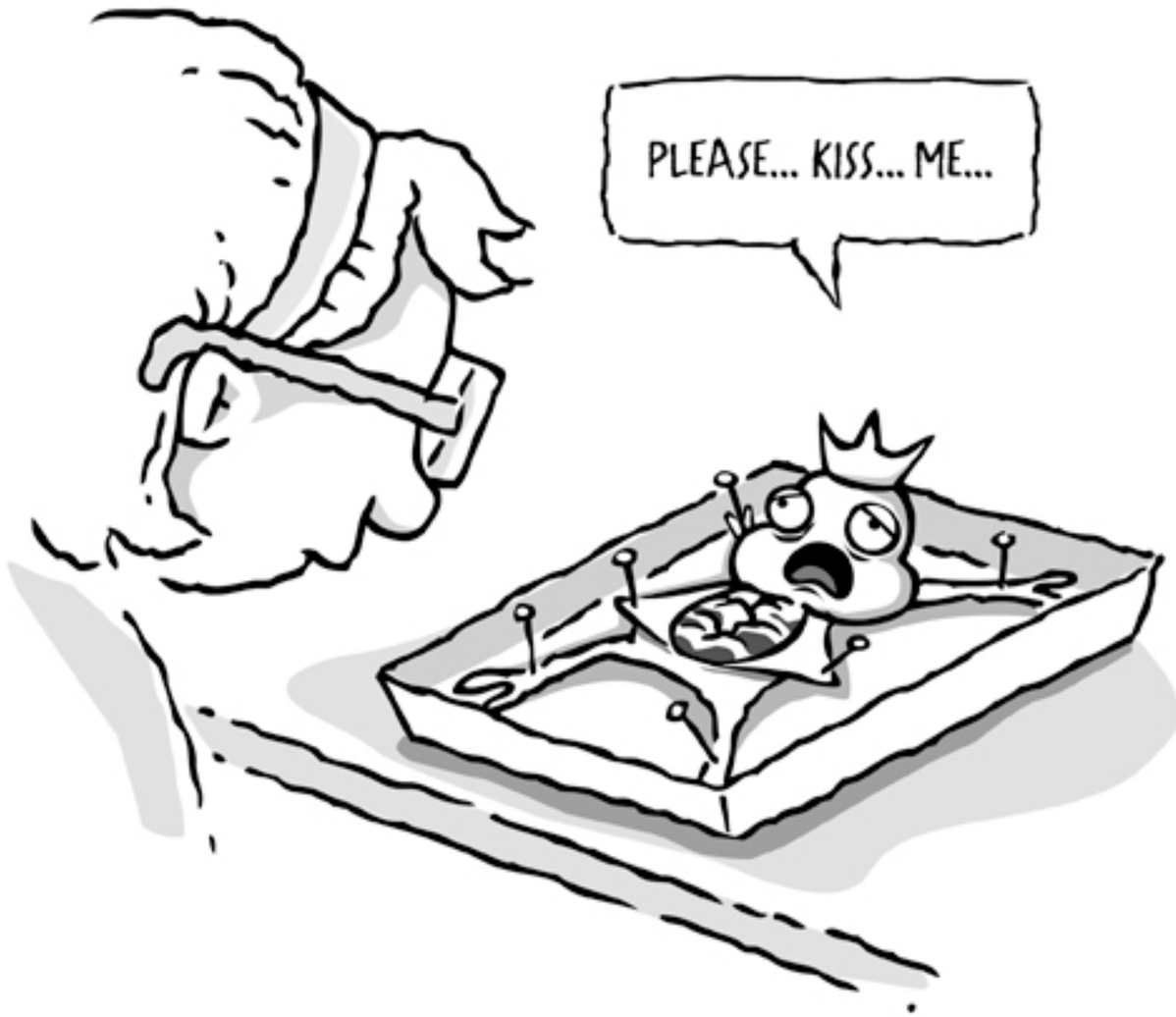
- Different programming backgrounds and skill sets
  - HPC, scripting, compiled languages, unix, web, etc.



COURSE DESIGN

CHALLENGES

- Different interests and goals for the course
    - Casual interest in biology
    - Elective credit



COURSE DESIGN

PROVIDE
KNOWLEDGE AND TOOLS TO
UNDERSTAND AND SOLVE A
RESEARCH PROBLEM
(IN BIOLOGICAL SPACE) WITH THE
AID OF COMPUTATION

COURSE DESIGN

COURSE GOALS

- Introduce major techniques and bioinformatics algorithms
  - Implement some of them
- Develop applications for bioinformatic analysis
  - Make them accessible to other researchers
- Become proficient in most important bioinformatics databases
  - NCBI, EBI, PDB, etc.
- Provide practical experience in research

COURSE DESIGN

SECONDARY GOAL: BECOME "COCKTAIL PARTY" CONVERSANT IN BIOLOGY

Syllabus

COURSE DESIGN

- Week 1,2 - Biology and Biological Databases
- Week 3,4 - Sequence Alignment
- Week 5,6 - Protein Bioinformatics
- Week 7,8,9 - Discovery, Genomes and Human Variation

COURSE DESIGN

- Week 1: Genomics, Bioinformatics and Molecular Biology
  - A high-level view of increasingly important role of computing in the biological sciences will be presented.

COURSE DESIGN

- Week 2: Genomes, Sequences and Databases
  - A survey of the current state of the art in storing, organizing and analyzing large data sets will be discussed
  - The advantages and disadvantages of these methods will be explored in the context of academic and commercial research initiatives.

COURSE DESIGN

- Week 3: Sequence Alignment
  - Fast, reliable alignment of text strings started the bioinformatics revolution. This lecture will show how these seemingly simple strings form the basis of almost all bioinformatics research.

COURSE DESIGN

- Week 4: Analyzing DNA and Protein Sequences
  - Techniques for sequence analysis will be discussed

COURSE DESIGN

- Week 5: Protein Structure and Function
  - Proteins are central building blocks of all organisms
  - Take bioinformatics to the third-dimension, showcasing how the spatial assembly and interactions of proteins support life and cause of disease

COURSE DESIGN

- Week 6: Molecular Modeling
  - Understanding protein function holds the promise developing therapeutics and curing diseases, but the computational complexity of analyzing three-dimensional models presents obstacles that have been difficult to overcome. This lecture will discuss approaches to shape analysis and comparison that can be scaled to large data sets.

COURSE DESIGN

- Week 7: In-Silico Drug Discovery
  - Approaches to using computer models to develop new drugs will be presented. We will discuss how years of playing Tetris might be more useful than you thought in combating antibiotic resistant pathogens.

COURSE DESIGN

- Week 8: Student Presentations / The Human Genome and Disease
  - The cause of diseases can be as simple as a single misplaced letter in a DNA sequence. From gene to disease, we will trace the genetic origin of disease. We will explore different approaches to cataloging and analyzing these changes.
- Week 9: Personal Genomics and Drug Discovery
  - Personalized genomic analysis is being used by consumers to better understand their health and their ancestry. The technologies used to power these services will be introduced as well as the different approaches used to provide web services to analyze the data.

COURSE DESIGN

- Final Exam Week 10: Final Project Presentations
  - Students will present a research topic in bioinformatics of their own choosing

Course Work

COURSE WORK

- 6 assignments (10% each)
  - Split between practical problems and implementation problems
- Gene Study Presentations (15%)
- Final project (25%):
  - Implement a bioinformatic method of your choosing

WE WILL DISCUSS THIS IN DETAIL LATER ON

COURSE WORK

SUBJECT TO CHANGE

Week	Material	Homework	Final Project
1	Bioinformatics, Genomics, and Molecular Biology	Assignment 1	
2	Biological Databases	Assignment 2	
3	Sequence Alignment	Assignment 3	
4	Analyzing Sequences	Assignment 4	
5	Protein Structure and Function	Assignment 5	
6	Molecular Modelling	Assignment 6	
7	in-silico drug discovery	Gene Study	Proposals Due before class
8	Student Gene Presentations		
9	The Human Genome and Disease		
Final	Student Video Presentations		Projects Due

COURSE WORK

GENE REPORT

- Identify a gene of interest
  - Related to disease, biofuels, manufacturing, etc.
  - Family of genes, what pathway, what does it do?
  - Find a gene of unknown function, predict its function…
- Become an expert on that gene
  - Publications
  - Bioinformatic analysis on it
  - Sequence, structure and function
- Present to class 10 minutes (5 minutes for questions)
  - Required elements of presentation

ASSIGNMENTS & PROJECTS

FINAL PROJECT

- Project of your choosing
- Examples
  - Implement a published bioinformatic method
  - Invent your own
  - Research a topic
  - Anything related to anything we've talked about...
- Present it to the class

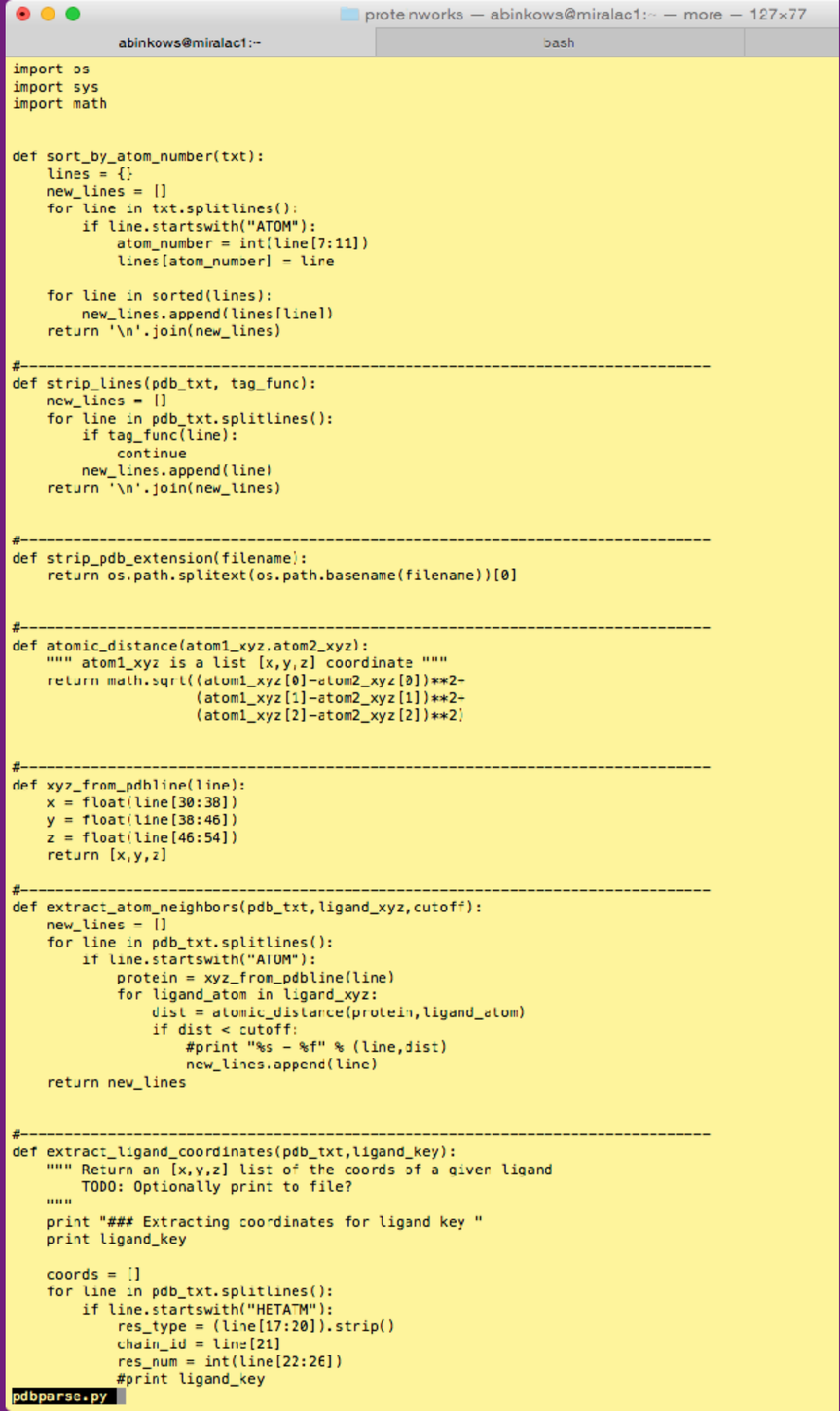
ASSIGNMENTS & PROJECTS

- Honor Code
  - All the assignments should be your own work
  - Department policies are strictly enforced
- Citing Resources
  - Cite any resources you use on homework, presentation and projects
  - Includes online resources
    - StackOverfow, BioStars, blogs, GitHub, etc.
- Third-party libraries and software
  - There are many great libraries for bioinformatics
  - Unless specifically stated, you should not use them

Course Technologies

COURSE TECHNOLOGIES

- Bioinformatics is focused on developing solutions to biological problems
  - Not on mastery of any particular language
  - Many bioinformatics resources violate all CS good design and implementation rules
  - Utility trumps all
    - Flat files are the lifeblood of many bioinformatic pipelines
    - The rise and fall of DoubleTwist



COURSE TECHNOLOGIES

- Bioinformatics requires aptitude in a variety of programming languages
  - Scripting (Python, Perl)
  - Command line (Bash, wget, curl, etc.)
  - Compiled languages (Fortran, C, C++)
  - Specialized languages (R, Matlab)
  - Web programming (Javascript, PHP, HTML)

MORE IMPORTANT THAN YOU MIGHT THINK

COURSE TECHNOLOGIES

- Command line tools
  - “Dirty little secret” of bioinformatics
  - You have to do something with you data before/after you big cluster runs

Command-line tools can be 235x faster than your Hadoop cluster

Sat 25 January 2014 by Adam Drake

Introduction

As I was browsing the web and catching up on some sites I visit periodically, I found a cool article from Tom Hayden about using Amazon Elastic Map Reduce (EMR) and mrjob in order to compute some statistics on win/loss ratios for chess games he downloaded from the millionbase archive, and generally have fun with EMR. Since the data volume was only about 1.75GB containing around 2 million chess games, I was skeptical of using Hadoop for the task, but I can understand his goal of learning and having fun with mrjob and EMR. Since the problem is basically just to look at the result lines of each file and aggregate the different results, it seems ideally suited to stream processing with shell commands. I tried this out, and for the same amount of data I was able to use my laptop to get the results in about 12 seconds (processing speed of about 270MB/sec), while the Hadoop processing took about 26 minutes (processing speed of about 1.14MB/sec).

After reporting that the time required to process the data with 7 c1.medium machine in the cluster took 26 minutes, Tom remarks

> "This is probably better than it would take to run serially on my machine but probably not as good as if I did some kind of clever multi-threaded application locally."

This is absolutely correct, although even serial processing may beat 26 minutes. Although Tom was doing the project for fun, often people use Hadoop and other so-called *Big Data (tm)* tools for real-world processing and analysis

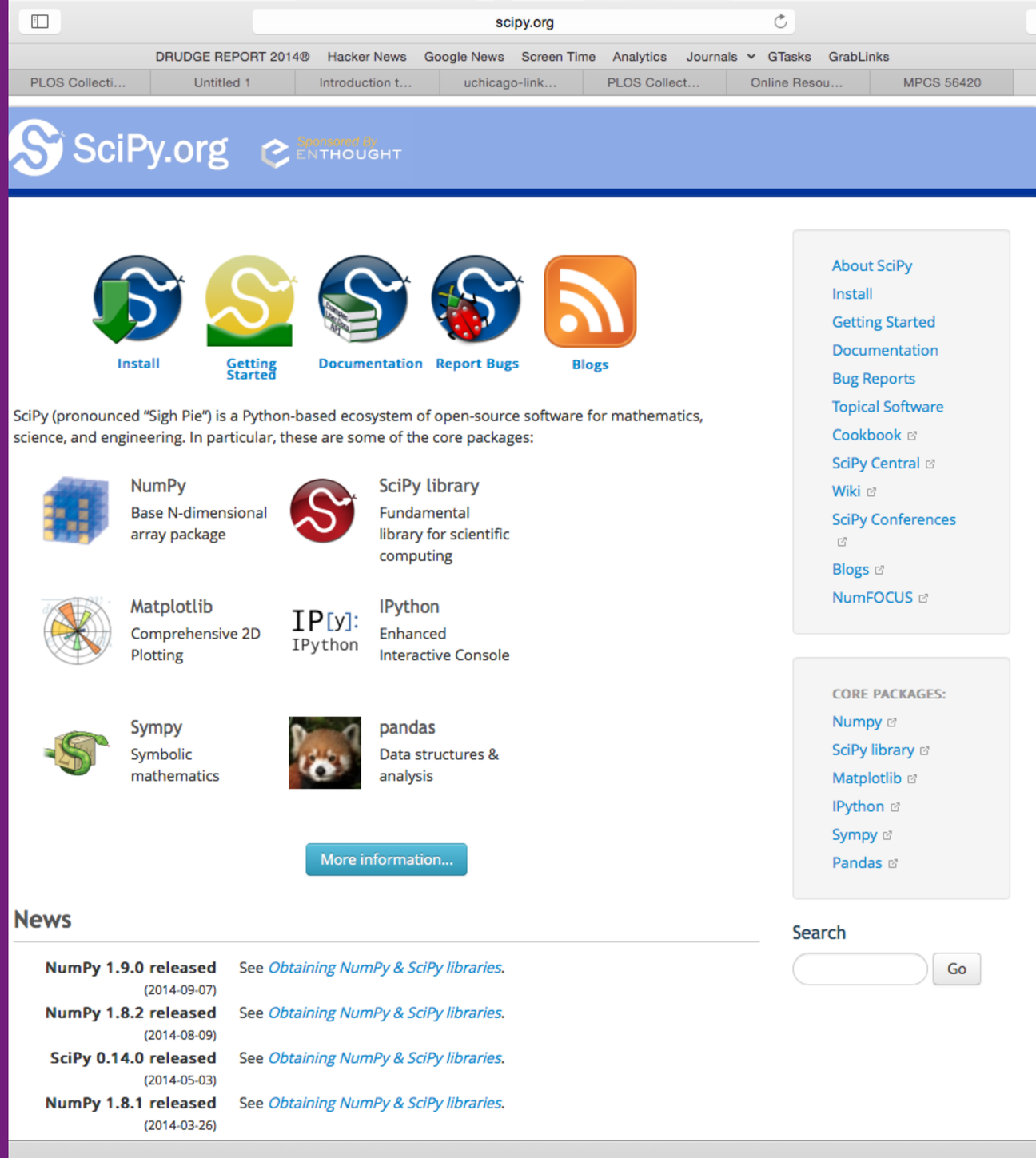
COURSE TECHNOLOGIES

CONSIDERATIONS FOR BIOINFORMATICS PROGRAMMING

- Portability - Will it run on my desktop and a legacy SGI machine?
- Scalability - Will it run on a cluster?
- Development speed - How long will it take to write
- Longevity - Is this a one time script or a full fledged application?
- Deliverability - Will this eventually be a web app or in the App Store?
- Target Audience - Who (besides me) may be running this? What is their background?

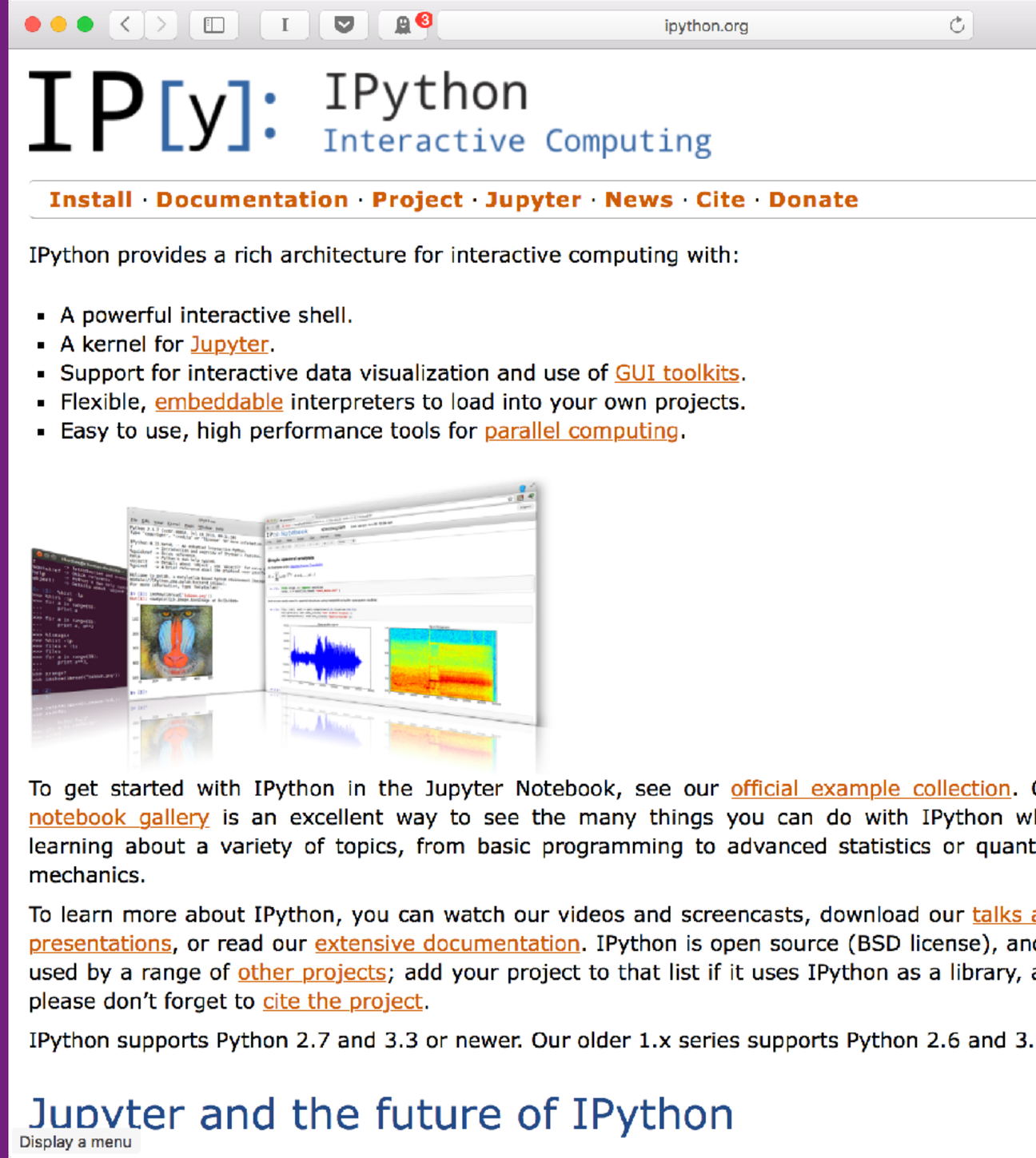
COURSE TECHNOLOGIES

- Programming languages for this course
  - Lectures and demonstrations will be conducted mostly in Python
  - Historical language of bioinformatics is Perl
- Why Python?
  - Great online resources and support
  - Support for SciPy scientific programming
  - Extensibility for C modules
    - Optimize when you need it
  - Native web application language



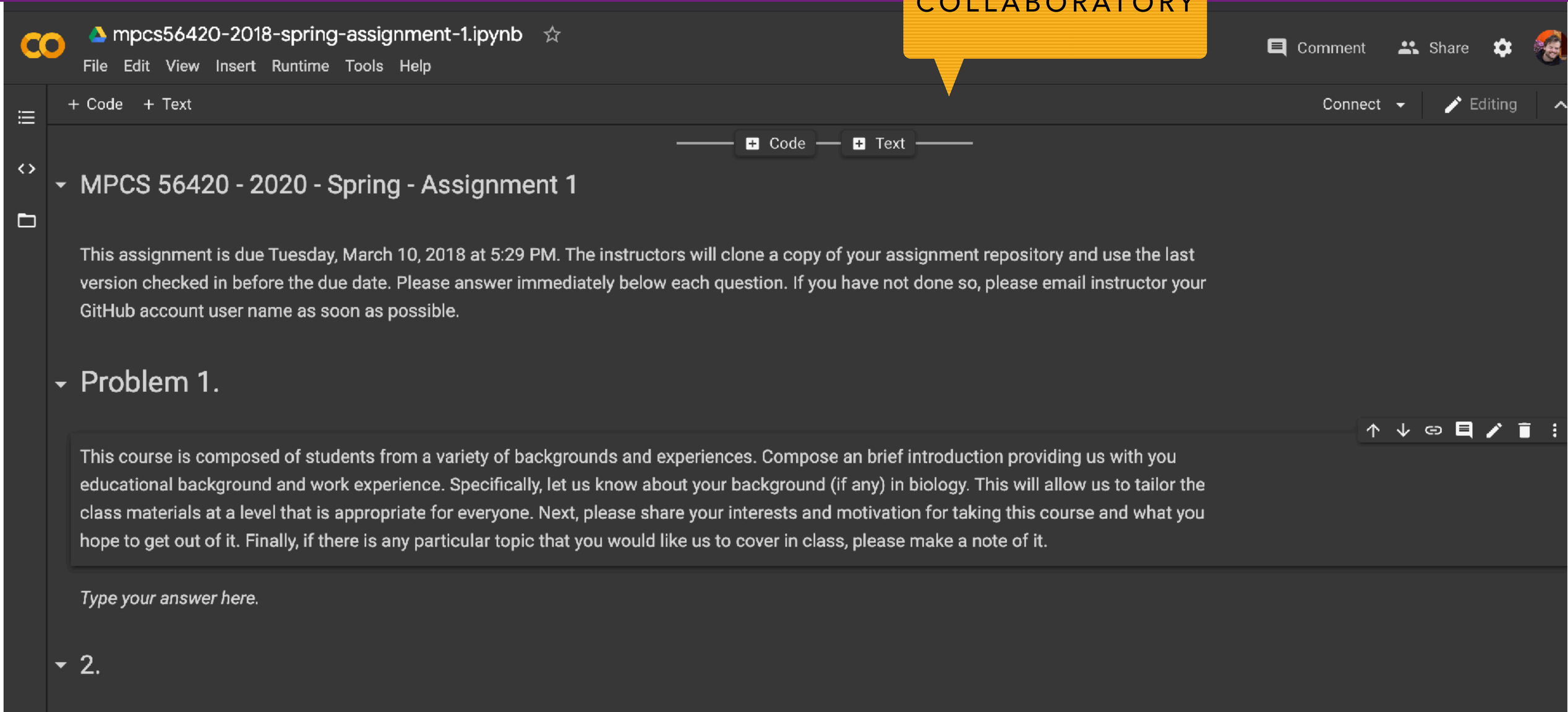
COURSE TECHNOLOGIES

- Jupyter notebooks
  - Support for reproducible workflows
  - Excellent support for SciPy
  - Consistent environment across platforms
  - Quickly becoming a “industry” standard
    - Publishing reproducible results



COURSE TECHNOLOGIES

GOOGLE COLLABORATORY

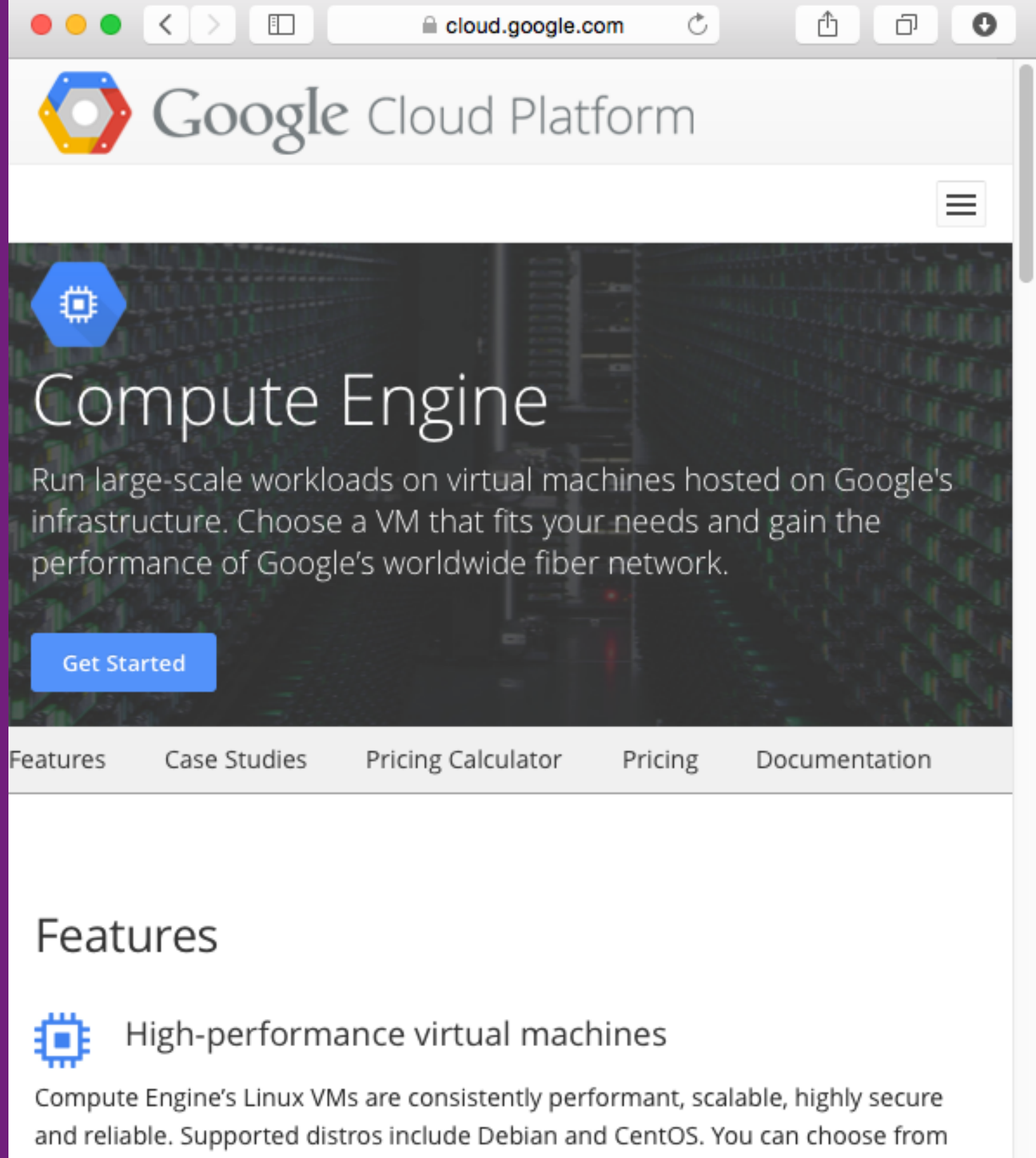


COURSE TECHNOLOGIES

- Biopython
  - Open sources framework for bioinformatics
  - Great resources and place to get started for algorithms and data structures
  - Unless specified, you should not use it

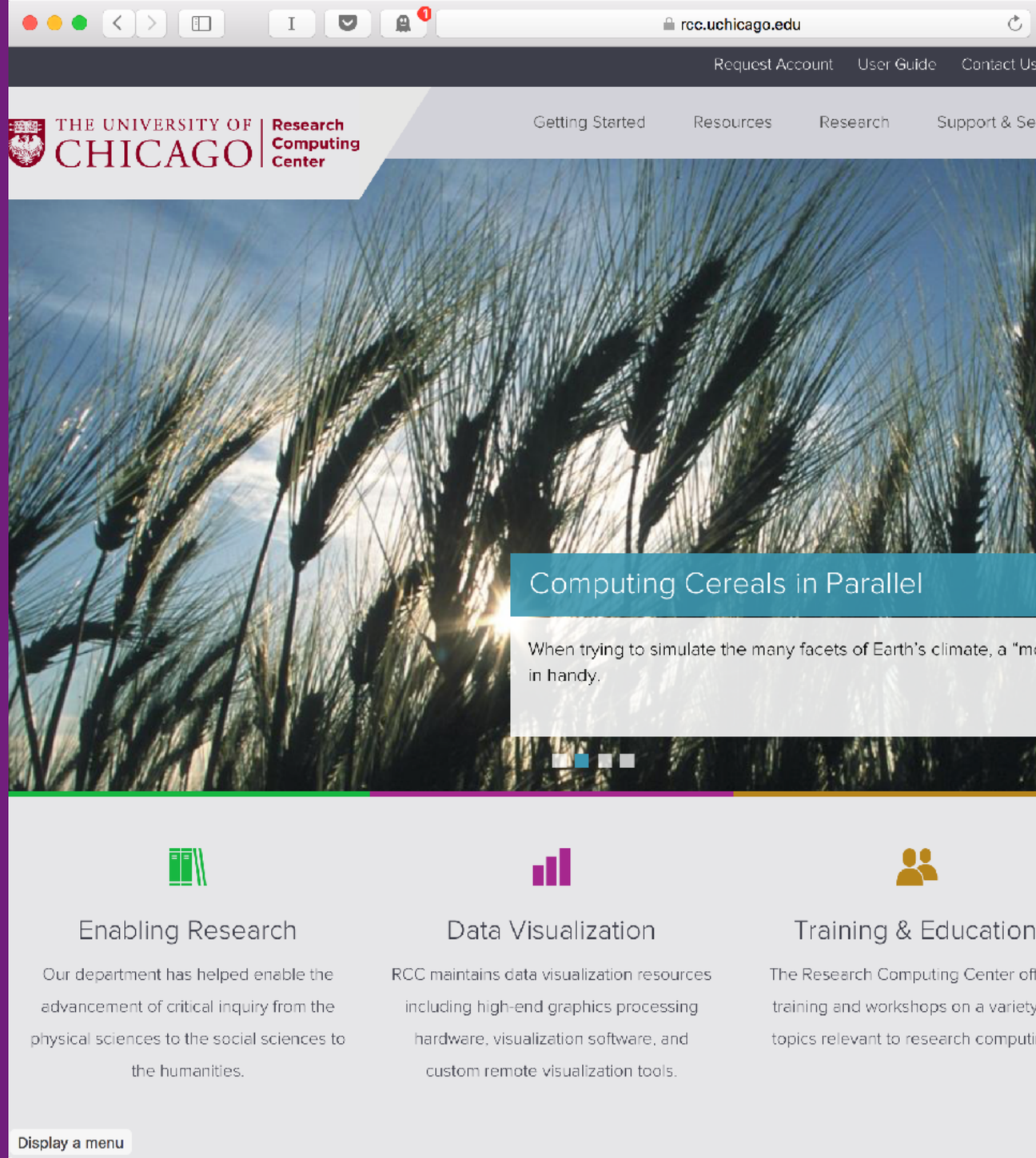
COURSE TECHNOLOGIES

- Google Could Platform
  - App Engine, Compute Engine, Storage, Genomics
  - Python, Java, PHP and Go
  - NumPy and Matplotlib are available on Google App Engine
  - Support for “big data” workflows



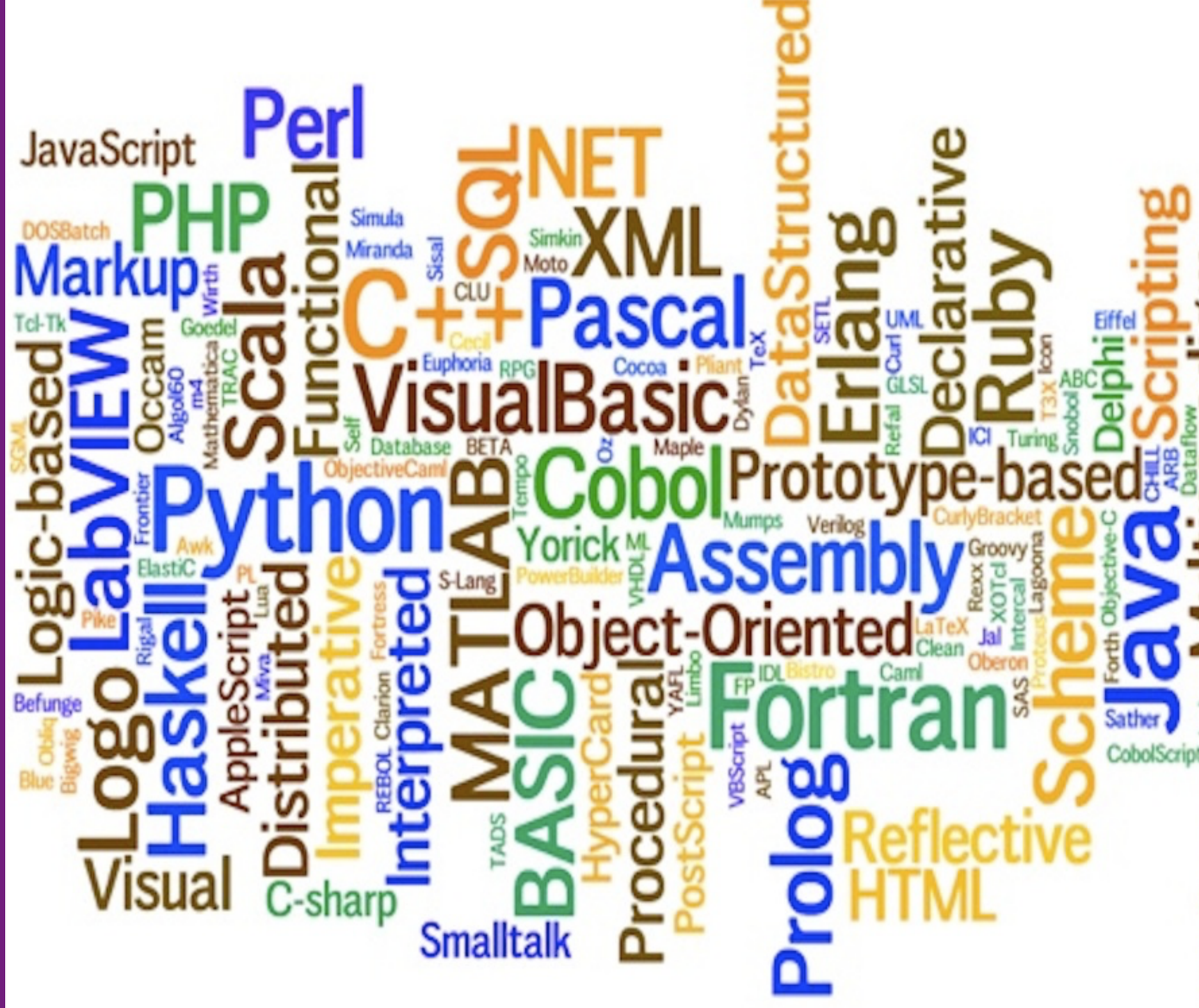
COURSE TECHNOLOGIES

- Research Computing Center
  - High-performance computing and visualization resources
  - High-capacity storage and backup
  - Software
  - High-speed networking
  - Hosted data sets



COURSE TECHNOLOGIES

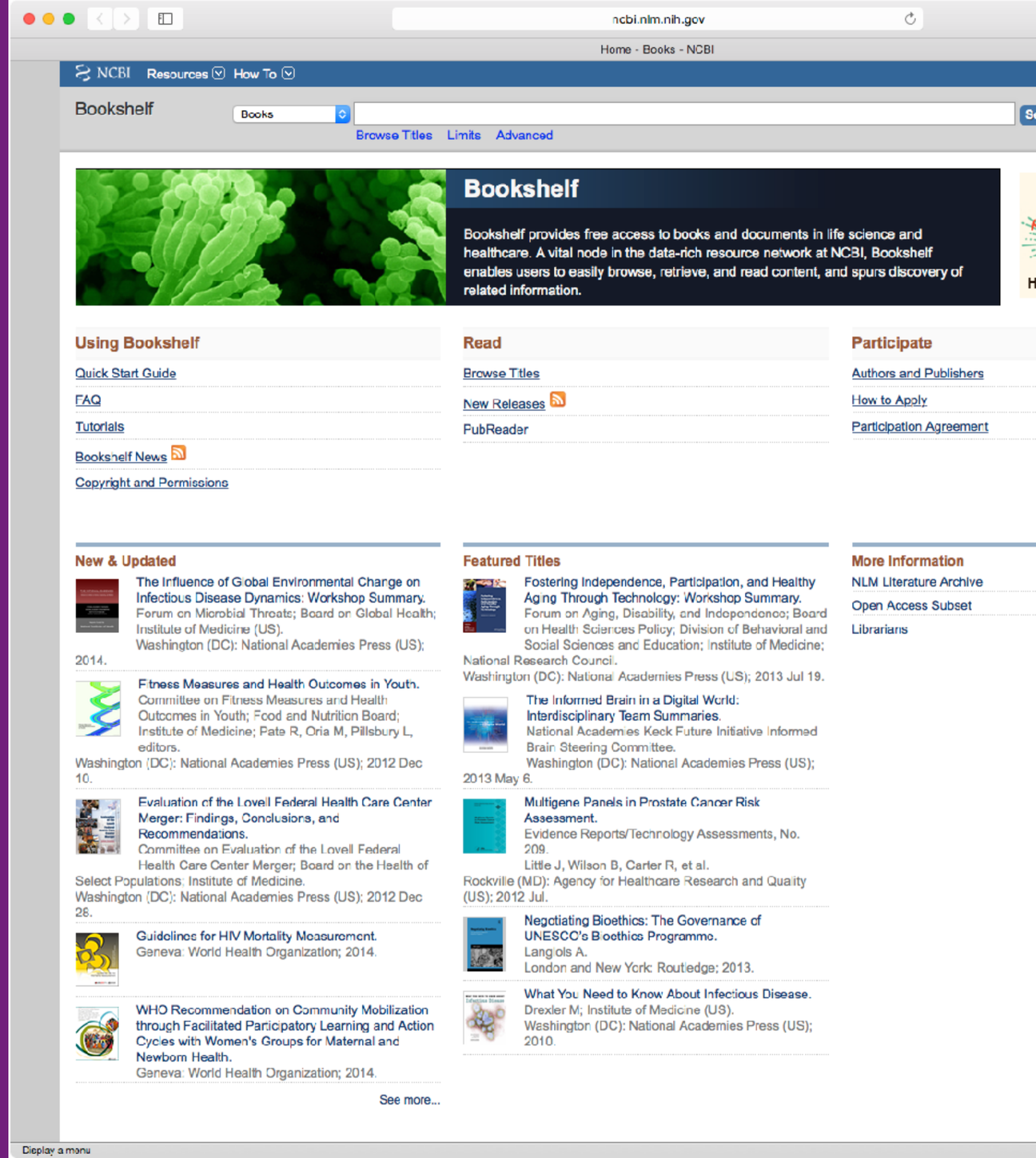
- You can use any other languages and technologies for your final projects
- Limited support from instructors



Course Resources

COURSE RESOURCES

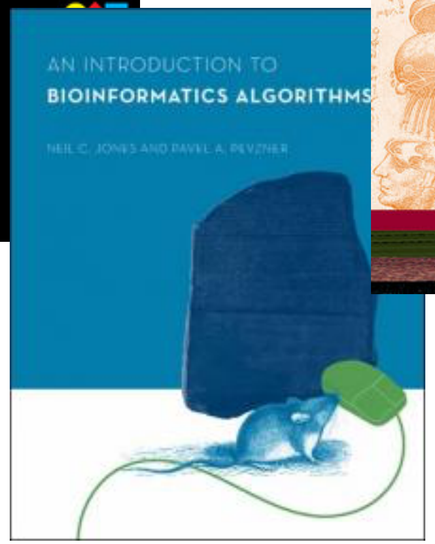
- Text Books
  - No required text books
  - Books become outdated quickly
  - Publications (scientific journals) have latest information
  - Quality online references from reputable sources
    - NCBI Handbook (http://www.ncbi.nlm.nih.gov/books/NBK143764/?term=handbook)

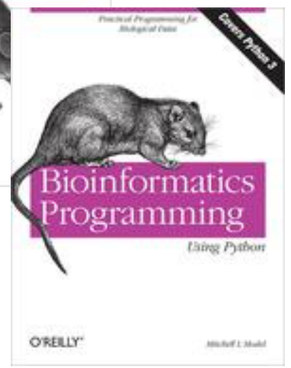


COURSE RESOURCES

TEXT BOOKS

FREE DIGITAL COPIES AVAILABLE THROUGH UCHICAGO LIBRARY

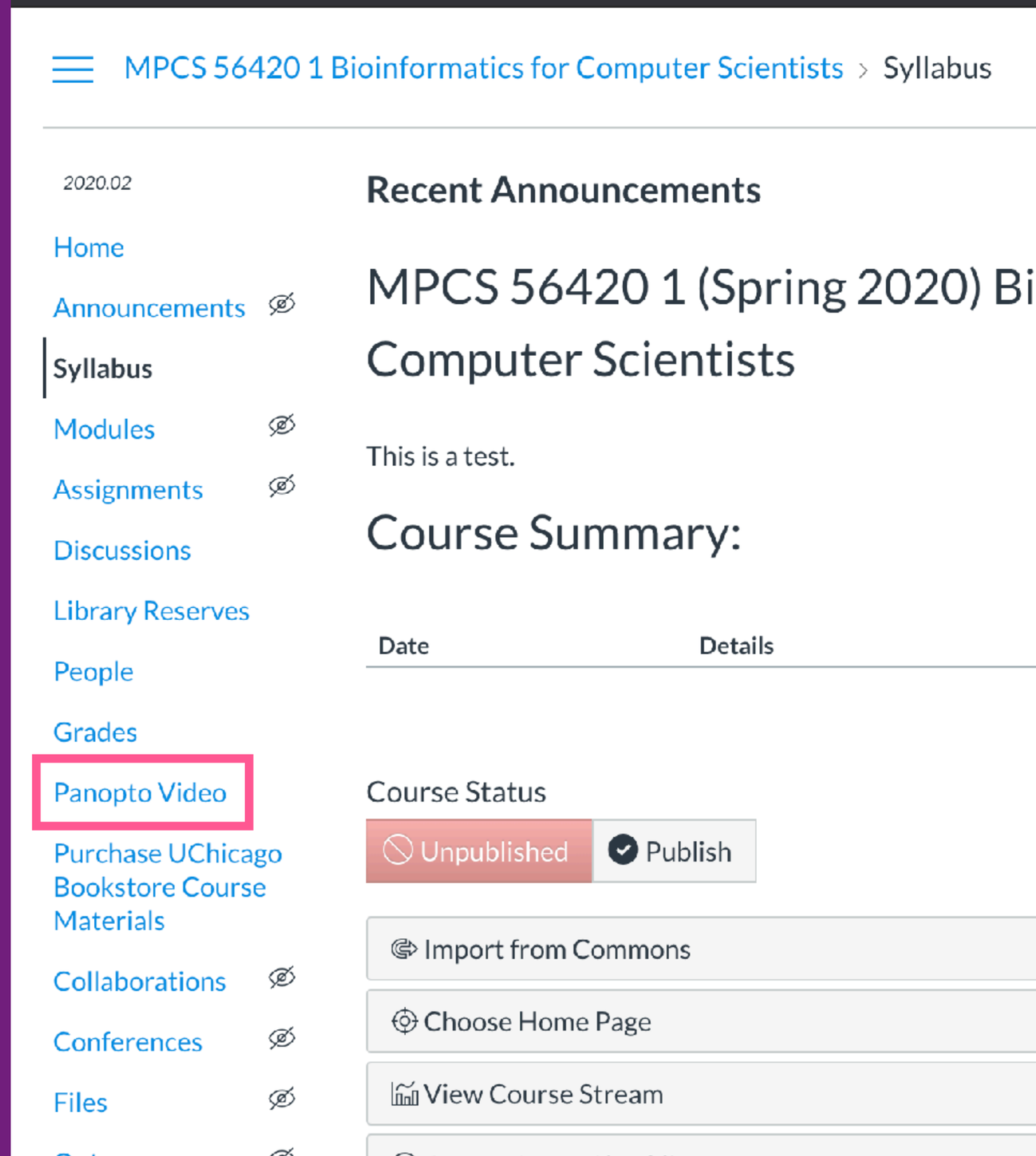




RECOMMENDED

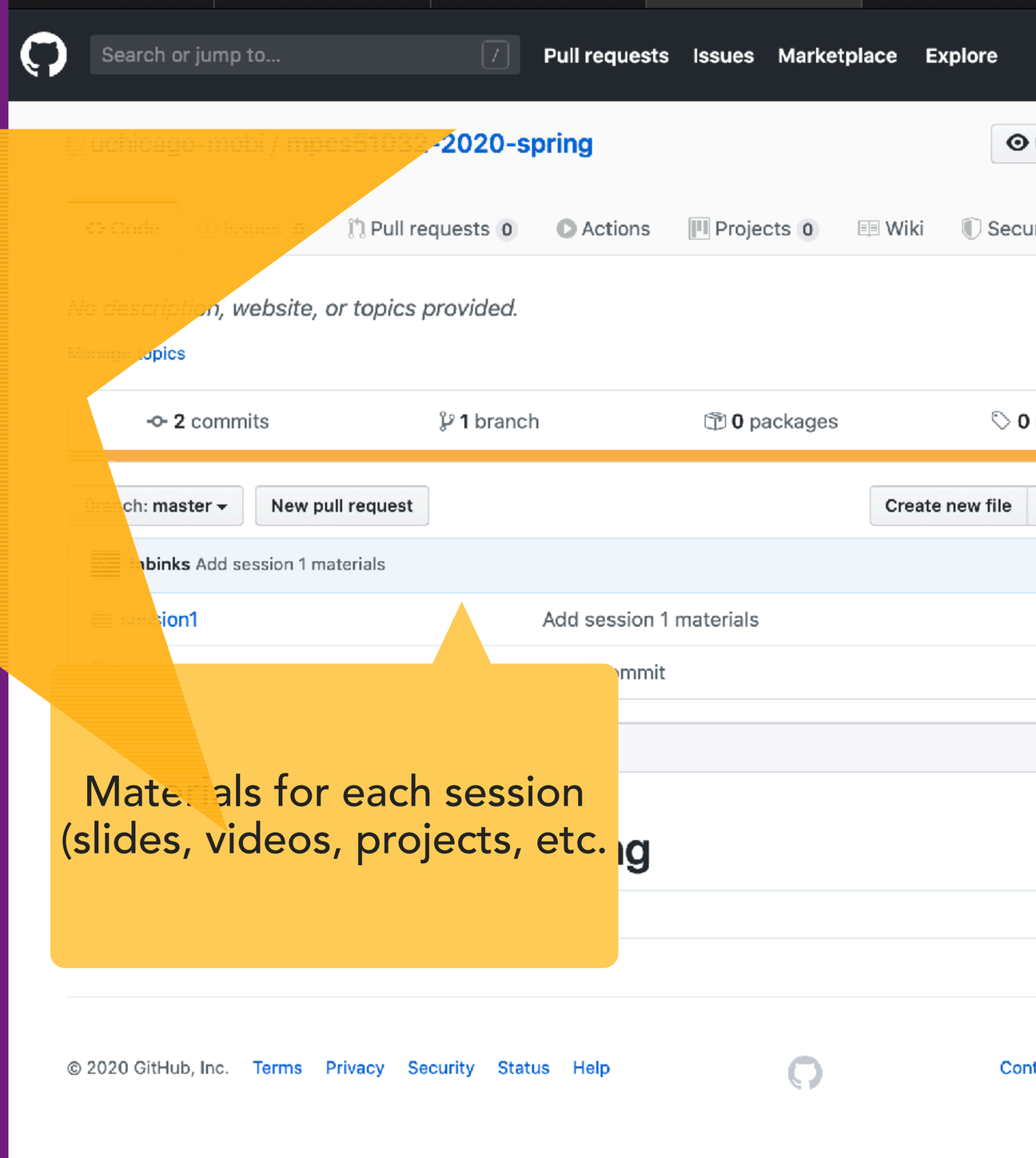
COURSE RESOURCES

- Canvas "Lite"
- Links to materials
  - Github
  - Panopto



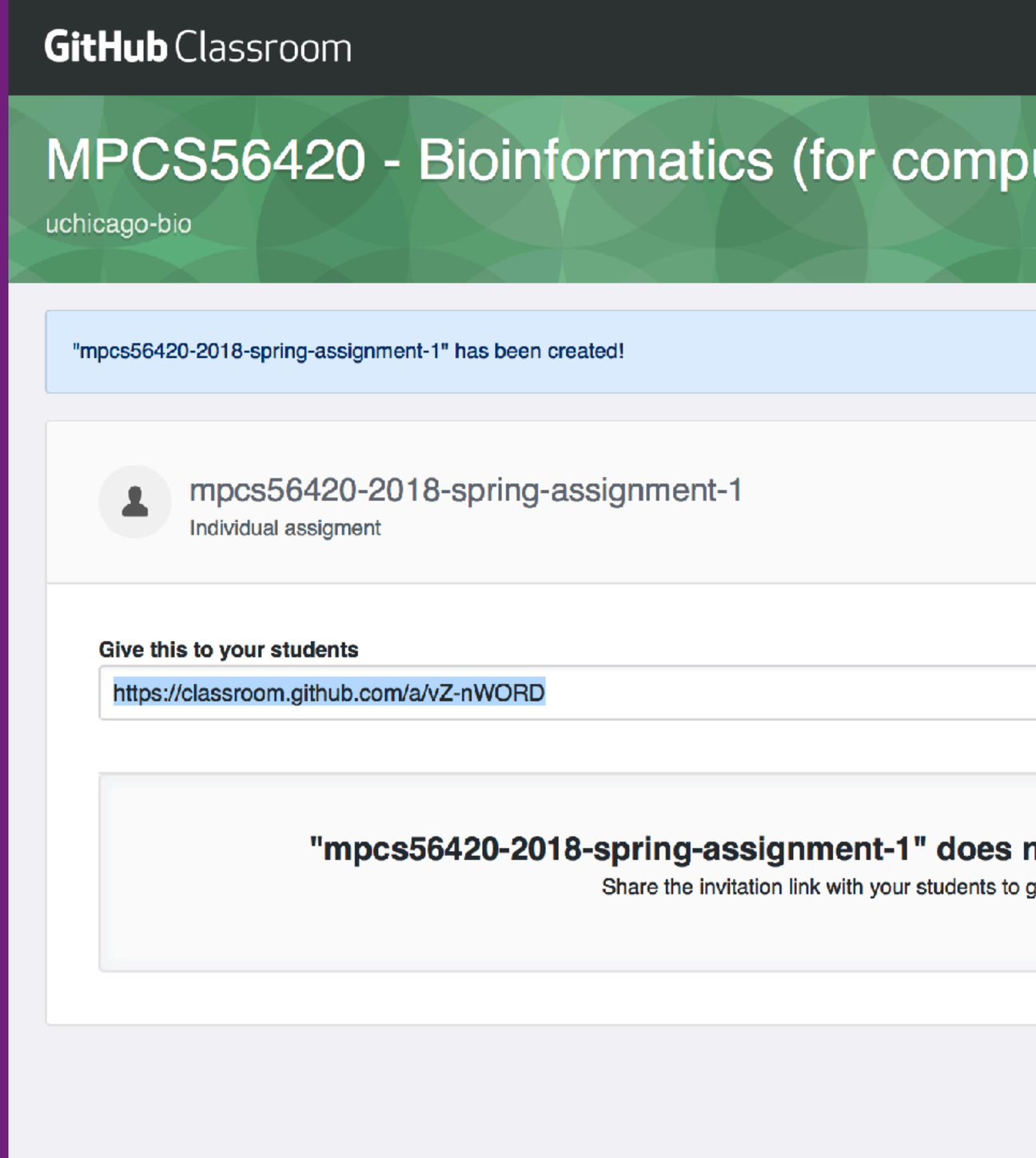
COURSE RESOURCES

- Github Repository for all course materials (no website 😢)
- https://github.com/uchicago-mobi/mpcs51032-2020-spring



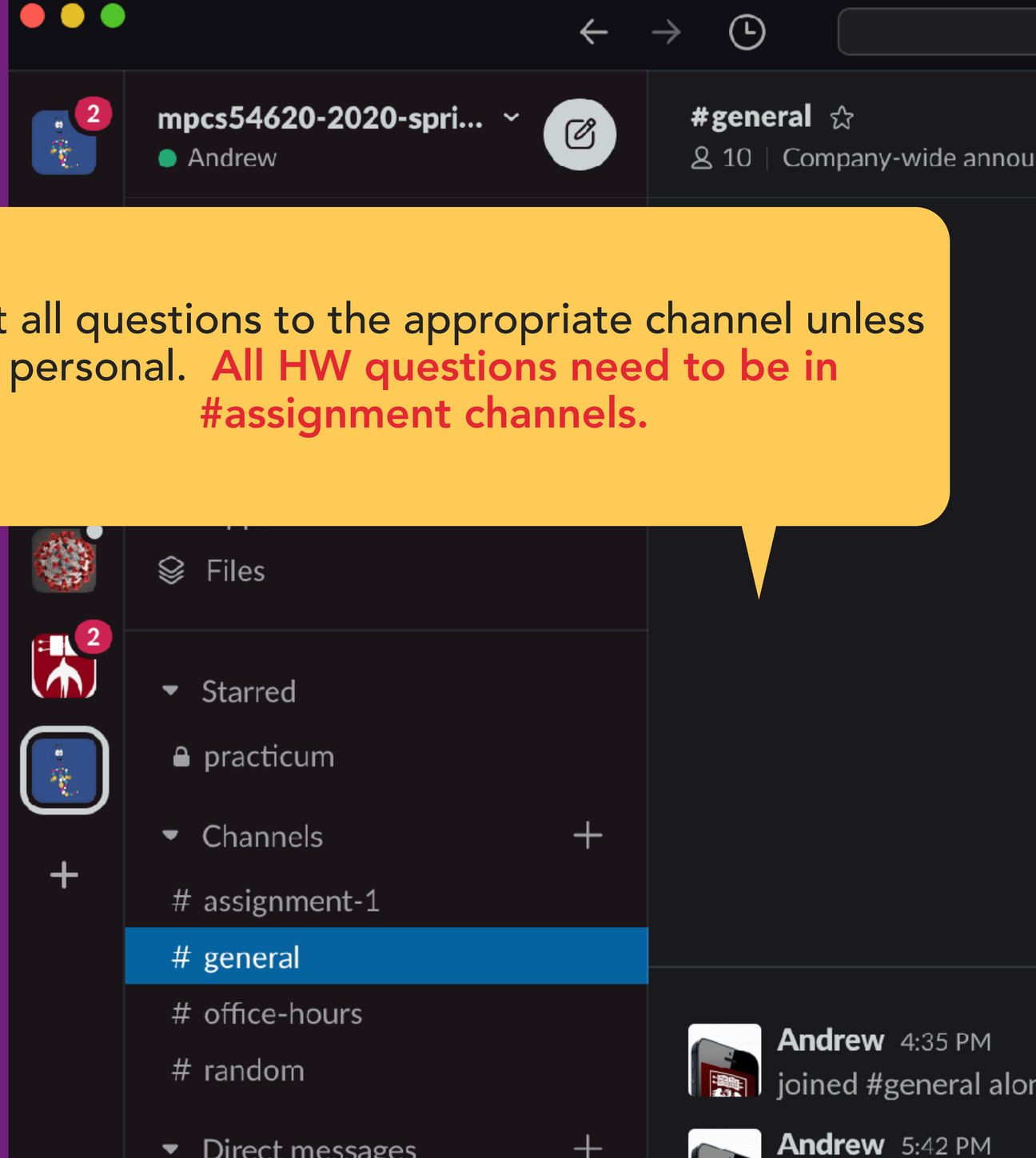
COURSE RESOURCES

- Github for coursework
  - Commit your repo to “turn in” assignments
  - Github has provided private repositories
    - I will create and share with you for course
- Slack for forum, announcements
- Email for personal matters



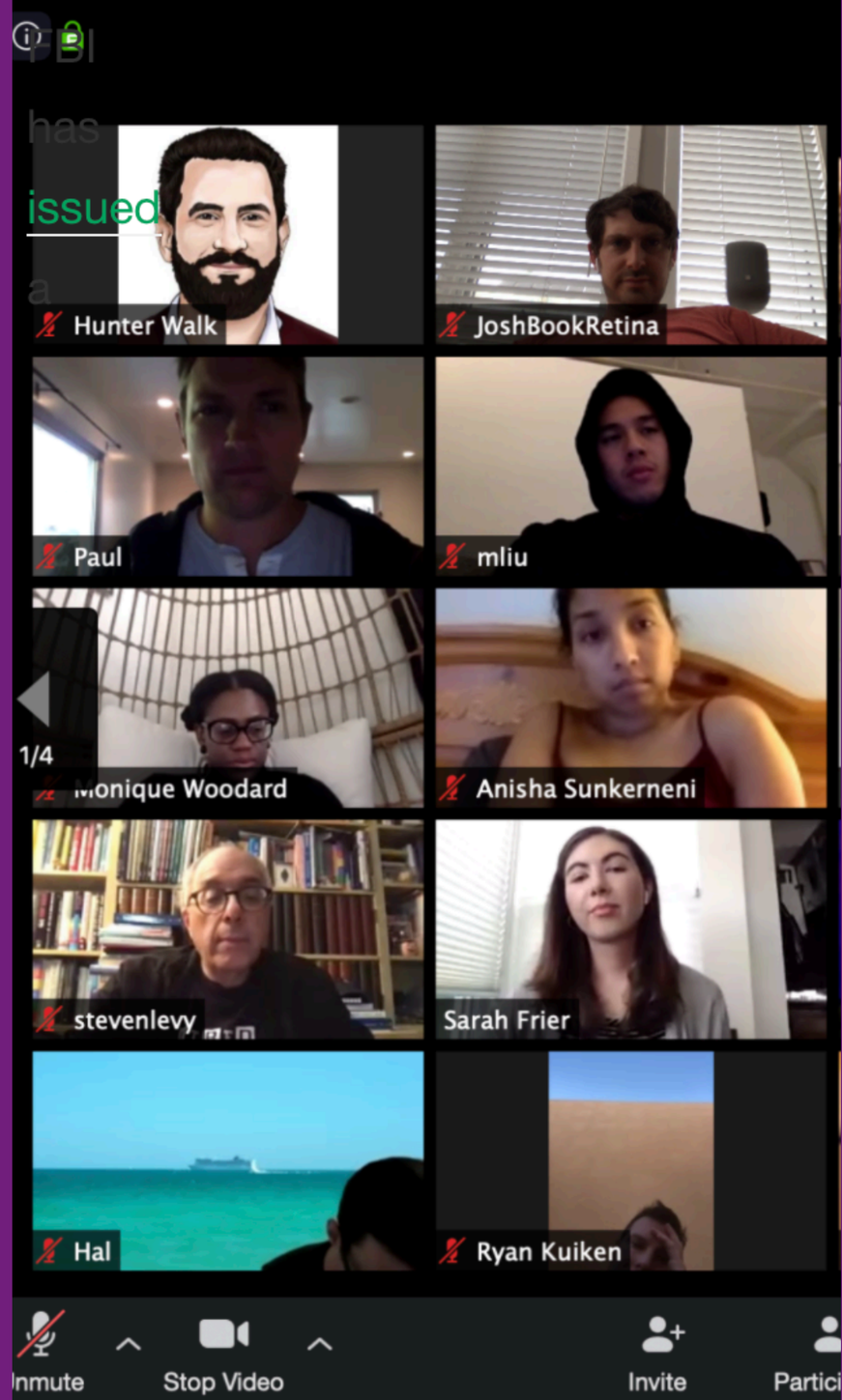
COURSE LOGISTICS

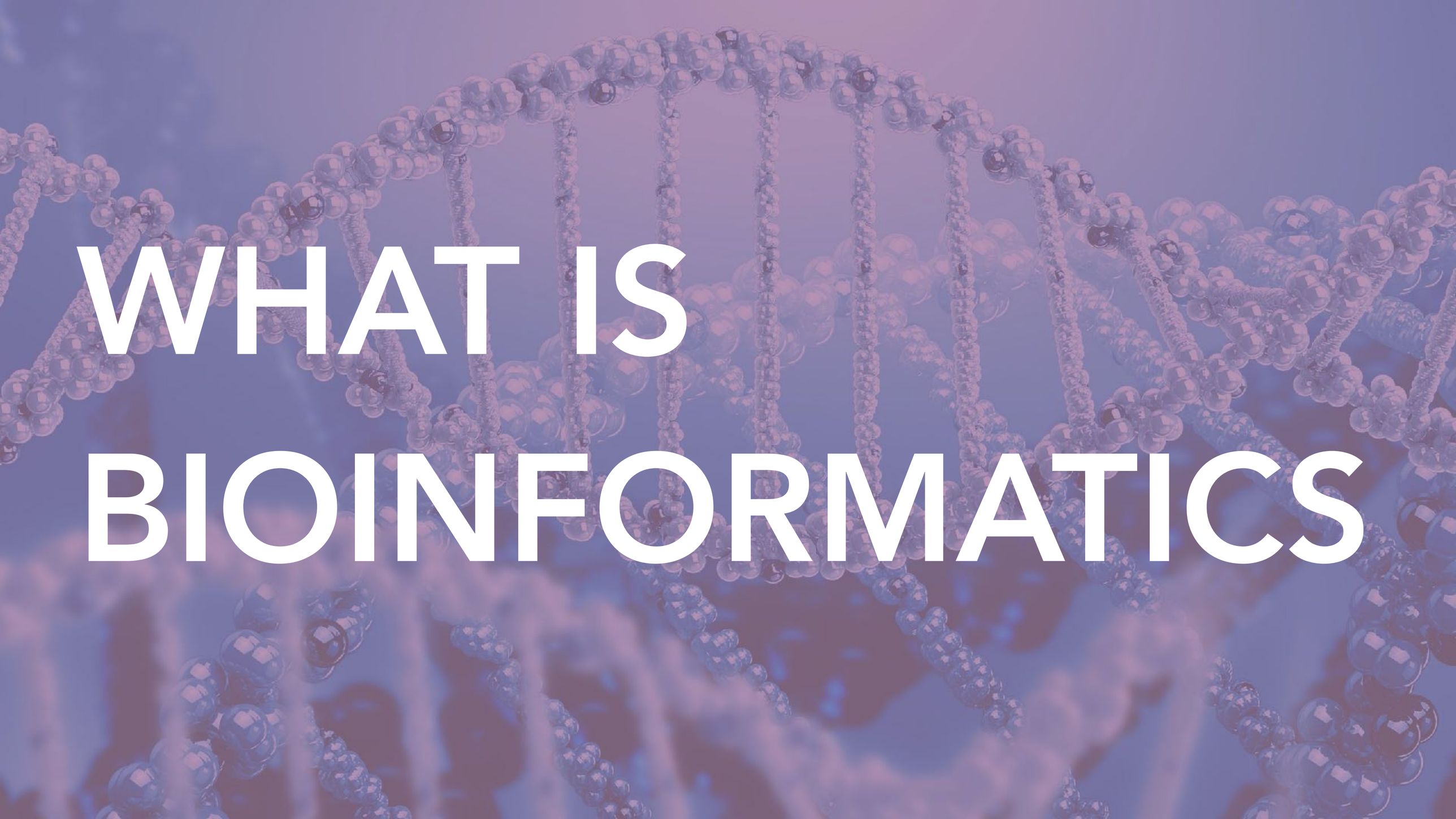
- Discussion forum rules
  - Anyone can answer a question
  - Post sources (if available) when answering questions



COURSE RESOURCES

- Please make sure you are muted if you are not talking
- Questions
  - "Raise Hand" if you have a question
  - Type it in the chat
  - Post in slack
- All sessions will be recorded
  - You can opt-out by turning camera and microphone off
  - Will not be distributed outside of class




WHAT IS
BIOINFORMATICS

WHAT IS BIOINFORMATICS

- "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline”
  - Official definition from the National Center for Biotechnology Information (NCBI)

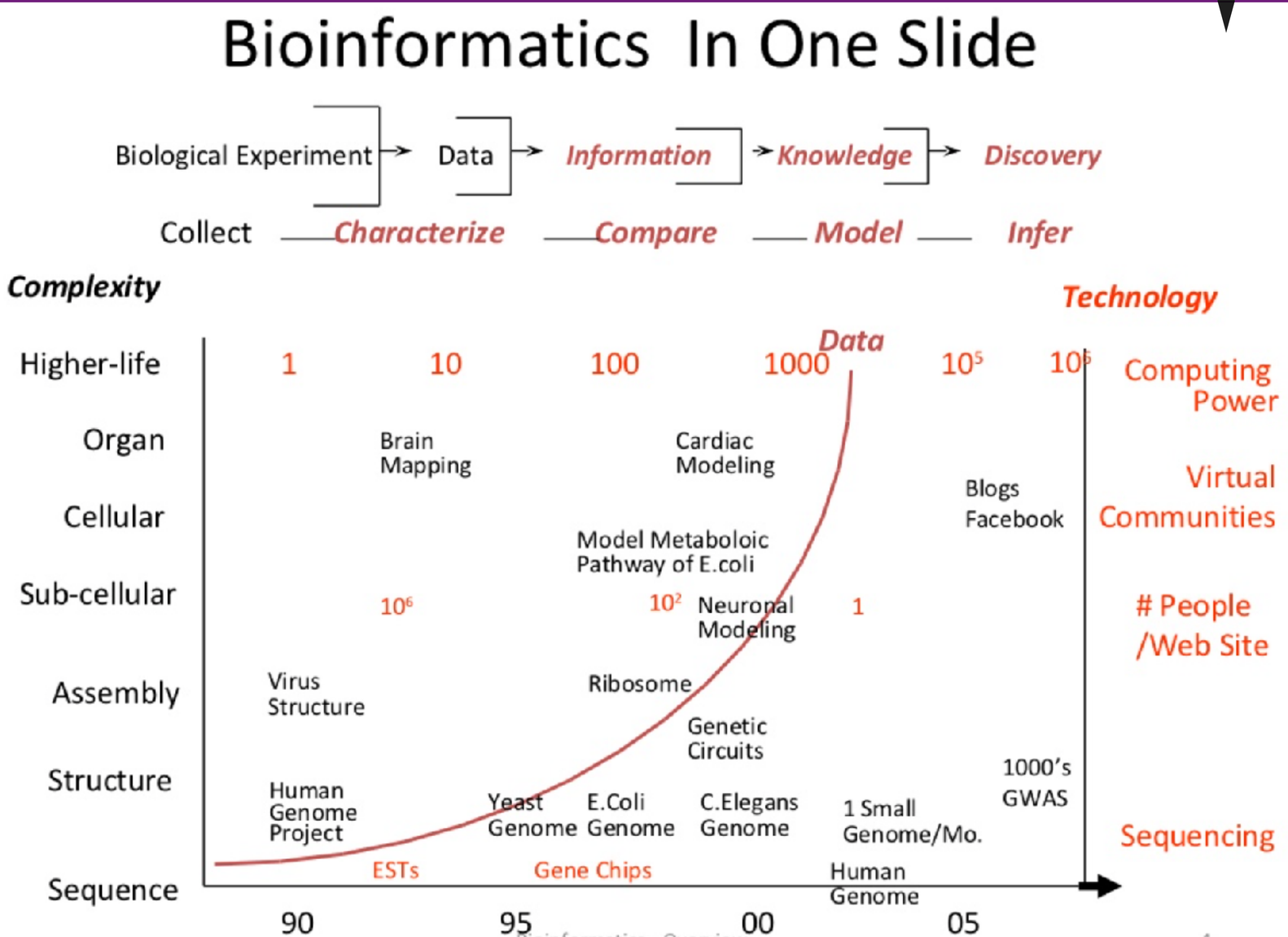
WHAT IS BIOINFORMATICS

THREE IMPORTANT SUB-DISCIPLINES

- Development of new algorithms and statistics with which to assess relationships among members of large data sets
- Analysis and interpretation of various types of data
  - Nucleotide and amino acid sequences, proteins
- Development and implementation of tools that enable efficient access and management of different types of information

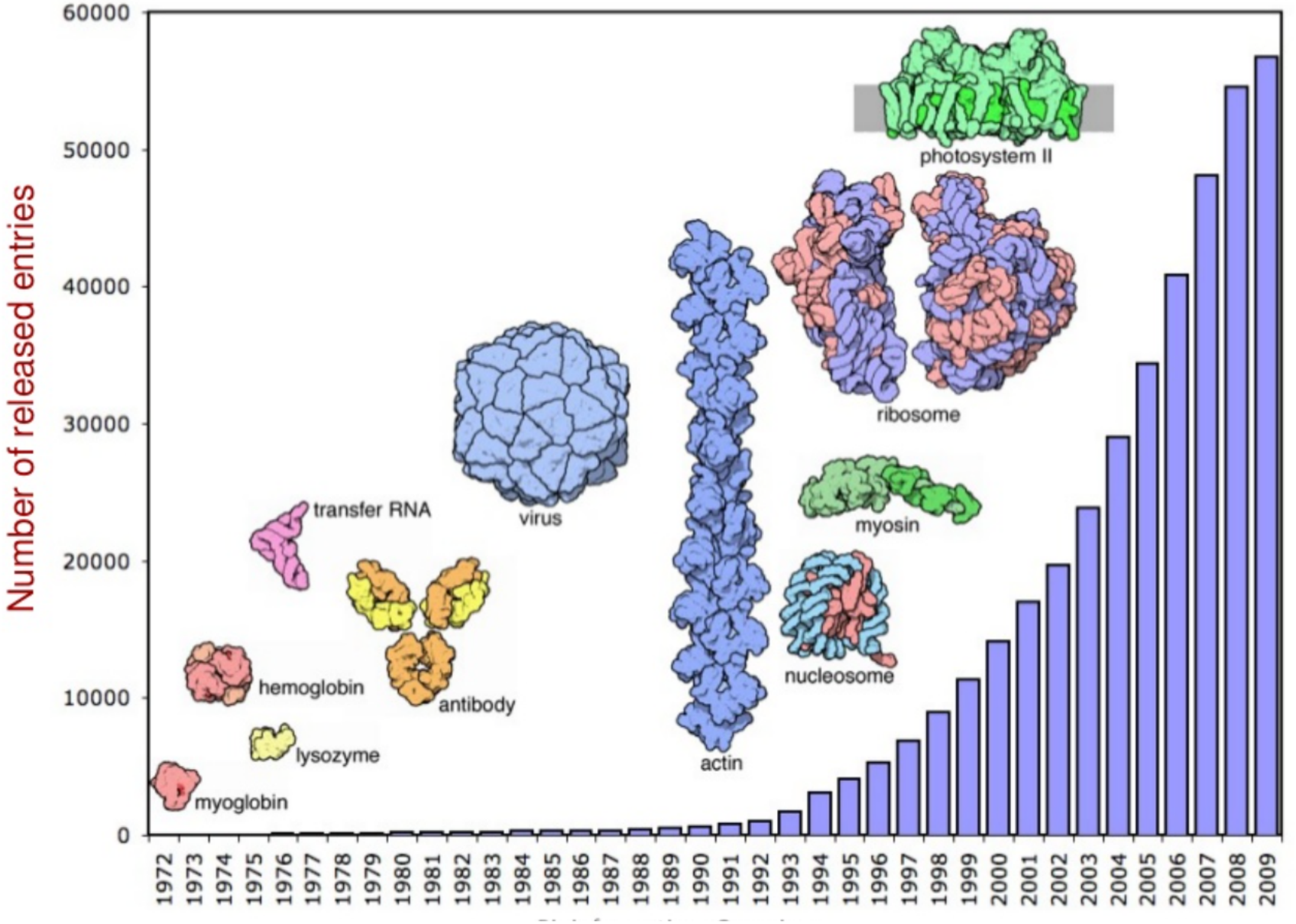
WHAT IS BIOINFORMATICS?

PHILIP BOURNE, UCSD



WHAT IS BIOINFORMATICS

- Growth of data
  - Protein Data Bank (PDB)



WHAT IS BIOINFORMATICS

- Using computers to answer biological questions
  - Including management and use of biological information
- Not LIMS (laboratory information management systems)
  - Collect data from experiments
  - Organize lab notebooks

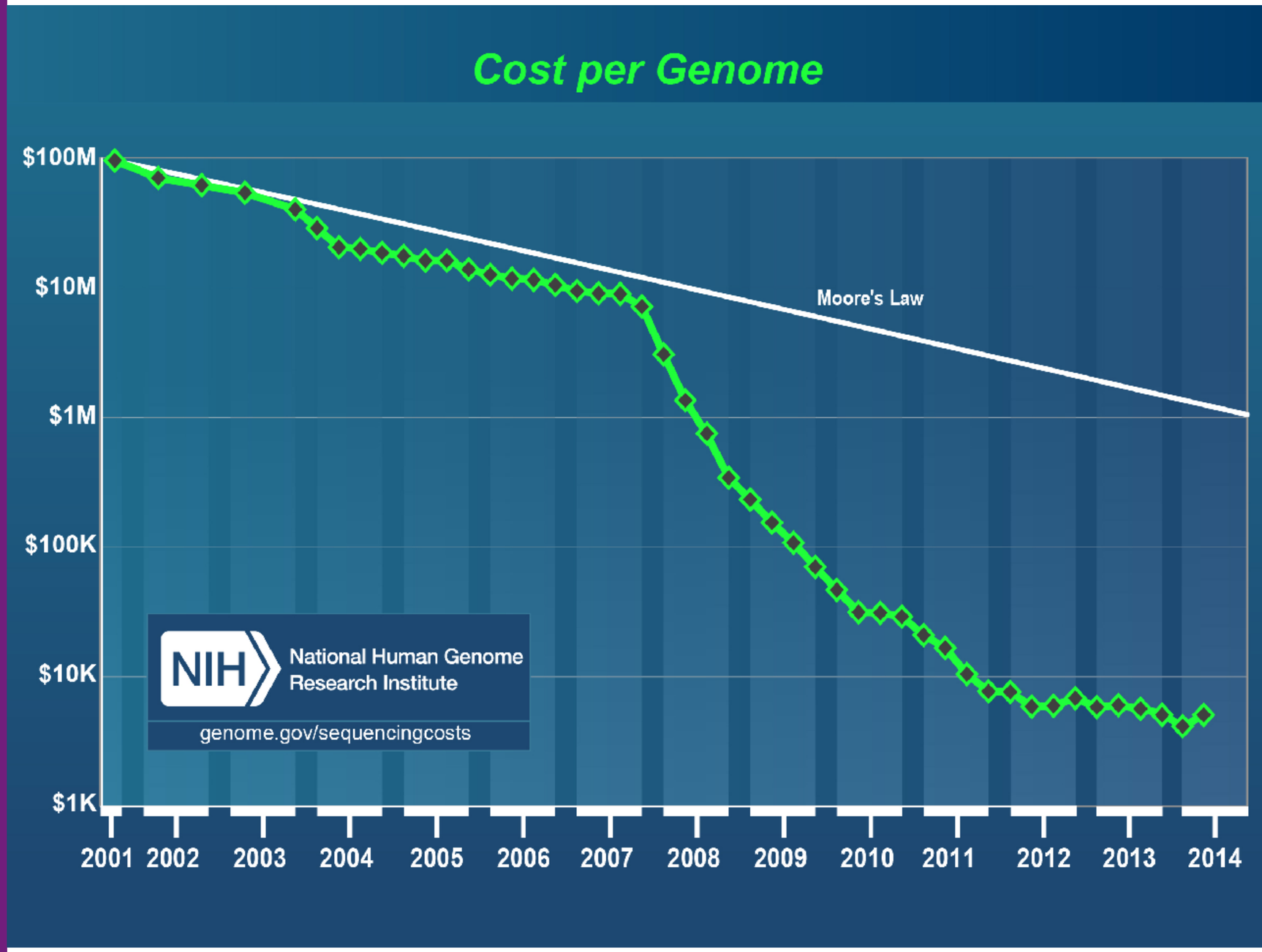
WHAT IS BIOINFORMATICS

BANKING USES COMPUTERS BUT ITS NOT CALLED BANKFORMATICS

- My answer: Its the next step in biology and biological research

WHAT IS BIOINFORMATICS

- The drive for the next-new-new revolution in bioinformatics
  - Genome sequencing costs
  - Do things that were literally not possible years ago



WHAT IS BIOINFORMATICS

- What can you do with bioinformatics?
  - Really important scientific research that contributes to a better humanity

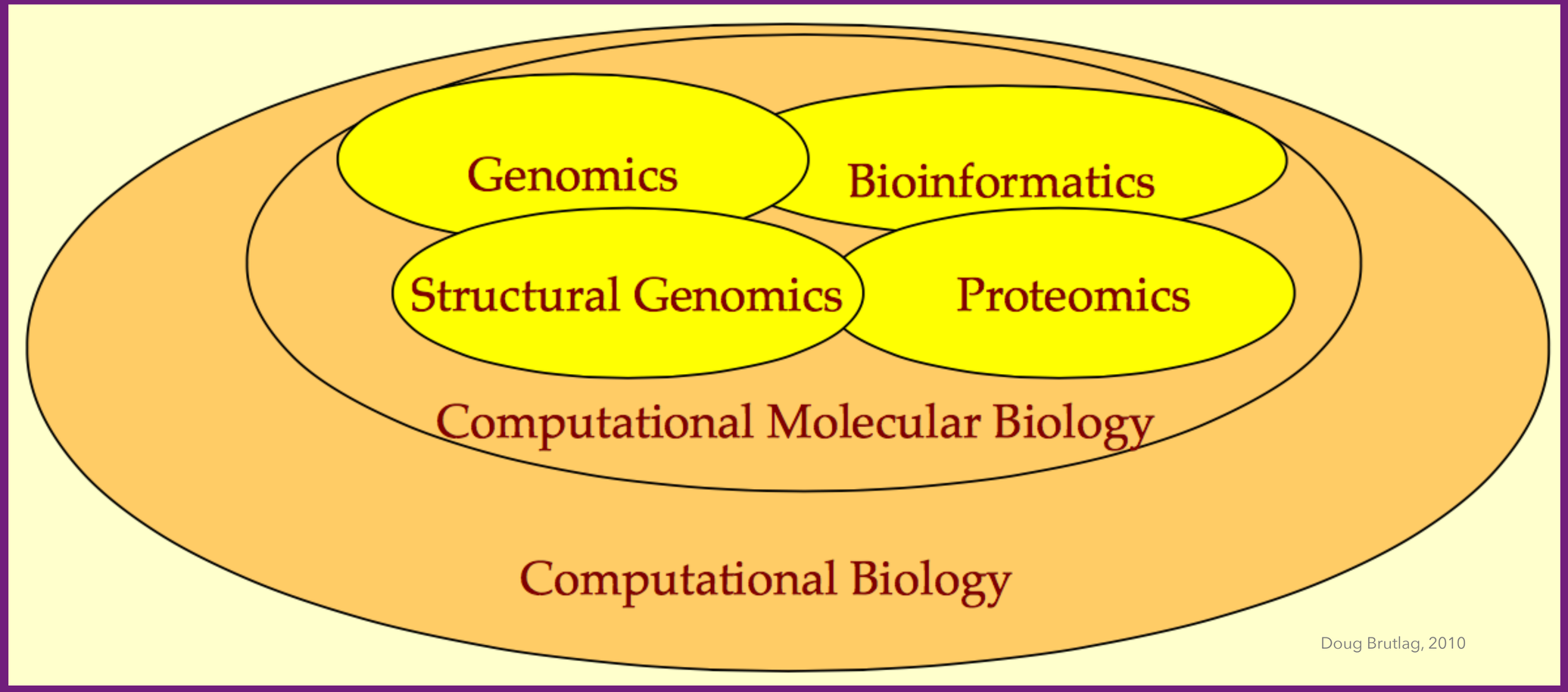
WHAT IS BIOINFORMATICS

WHAT ABOUT "COMPUTATIONAL BIOLOGY"?

INCLUDES BIOINFORMATICS

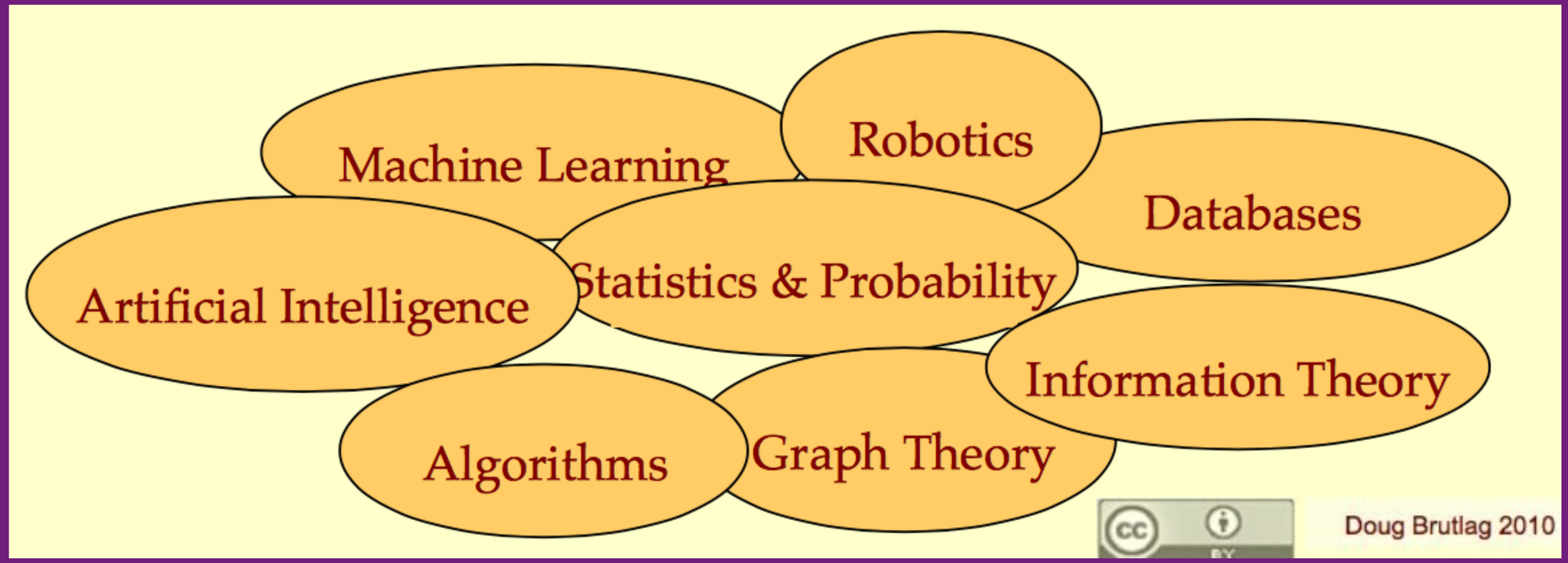
- Computational biology is an interdisciplinary field that applies the techniques to address biological problems
  - Computer science
  - Applied mathematics
  - Statistics
- A broader term (Can be used interchangeably, in my opinion)

WHAT IS BIOINFORMATICS



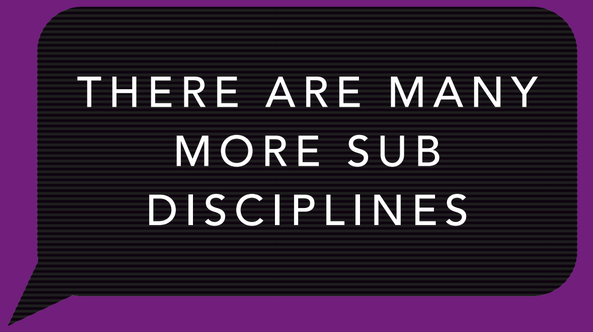
- Using computers to answer biological question

WHAT IS BIOINFORMATICS?



- Computational biology pulls from many subjects for ideas
  - Unconventional ways (mosquito robot, docking game)

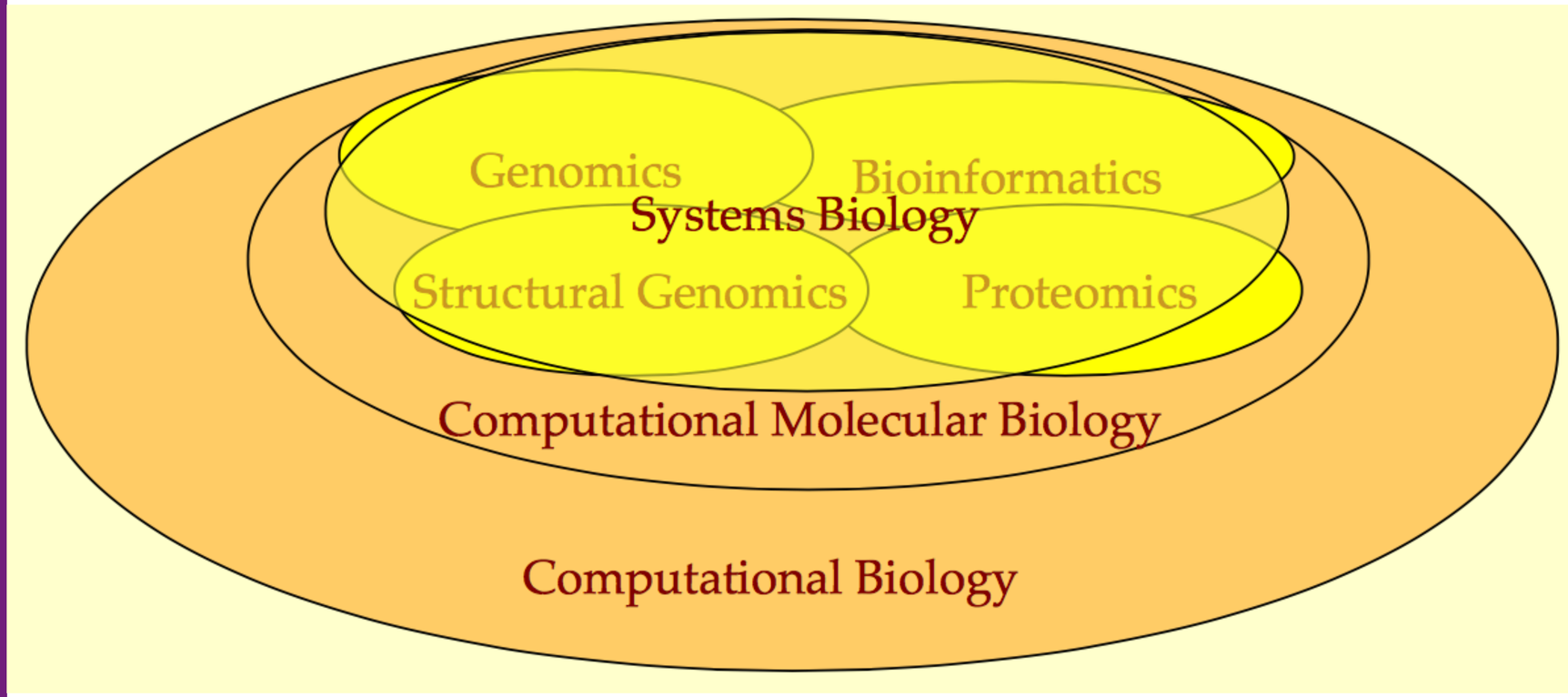
WHAT IS BIOINFORMATICS?



- What about all the “-omics”?
  - Genomics - Discovery, arrangement, expression of genes
  - Proteomics - Study the proteins in a system; mass spec analysis; markers
  - Structural Genomics - Study structural representatives of all proteins
  - Bioinformatics - Studying biological information

WHAT IS BIOINFORMATICS

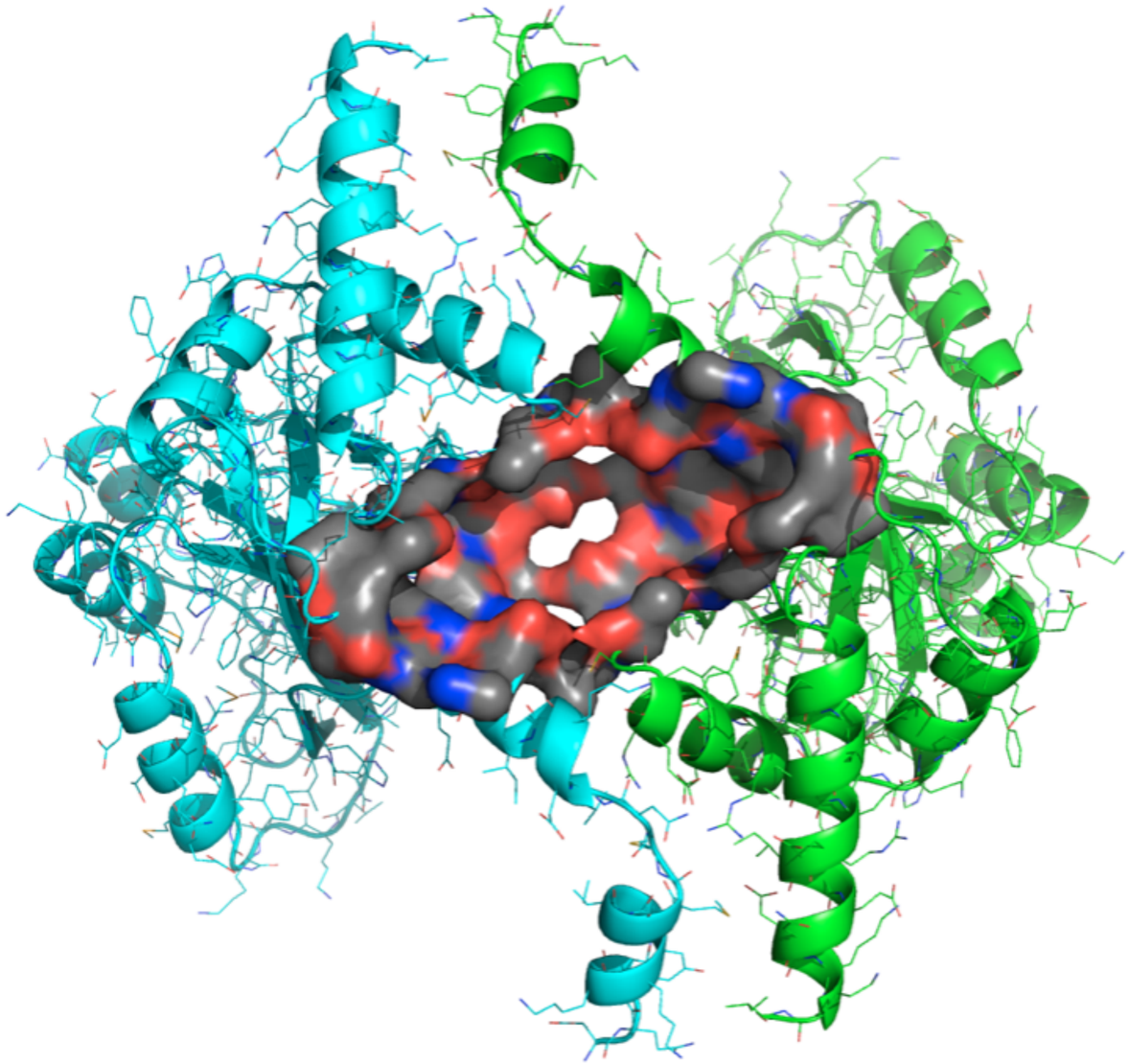
- Systems biology
  - New(er) field uses information to understand how systems work together
  - Enabled by the other fields



Doug Brutlag, 2010

WHAT IS BIOINFORMATICS

- Means different things to different people
  - Population modeling, disease modeling, molecular modeling, etc.
- Spend a career in one small subset
  - Protein surface analysis
- We will try to cover a little bit of everything



Goals of Bioinformatics

GOALS OF BIOINFORMATICS

- Discover
- Predict
- Infer
- Organize
- Integrate
- Simulate
- Engineer

GOALS OF BIOINFORMATICS

- Analysis of genes and proteins
  - Homolog detection
  - Alignment (the residual-level mapping among homologous genes/proteins)
  - Application of the alignments
    - Detect the conserved residues, Functional sites, Prediction of protein structures, Motif finding (cis-elements)
  - Phylogeny
  - Engineer

03.30.18

CRISPR-altered plants are not going to be regulated (for now)

[Photo: dimitrisvetsikas1969/Pixabay]

Good news for people who like genetically altered tomatoes and other plants. The U.S. Department of Agriculture announced it will no longer regulate them.

The USDA not only rolled back Obama-era rules regulating genetically edited plants, but now it claims that plants whose genomes have been altered using gene-editing technology

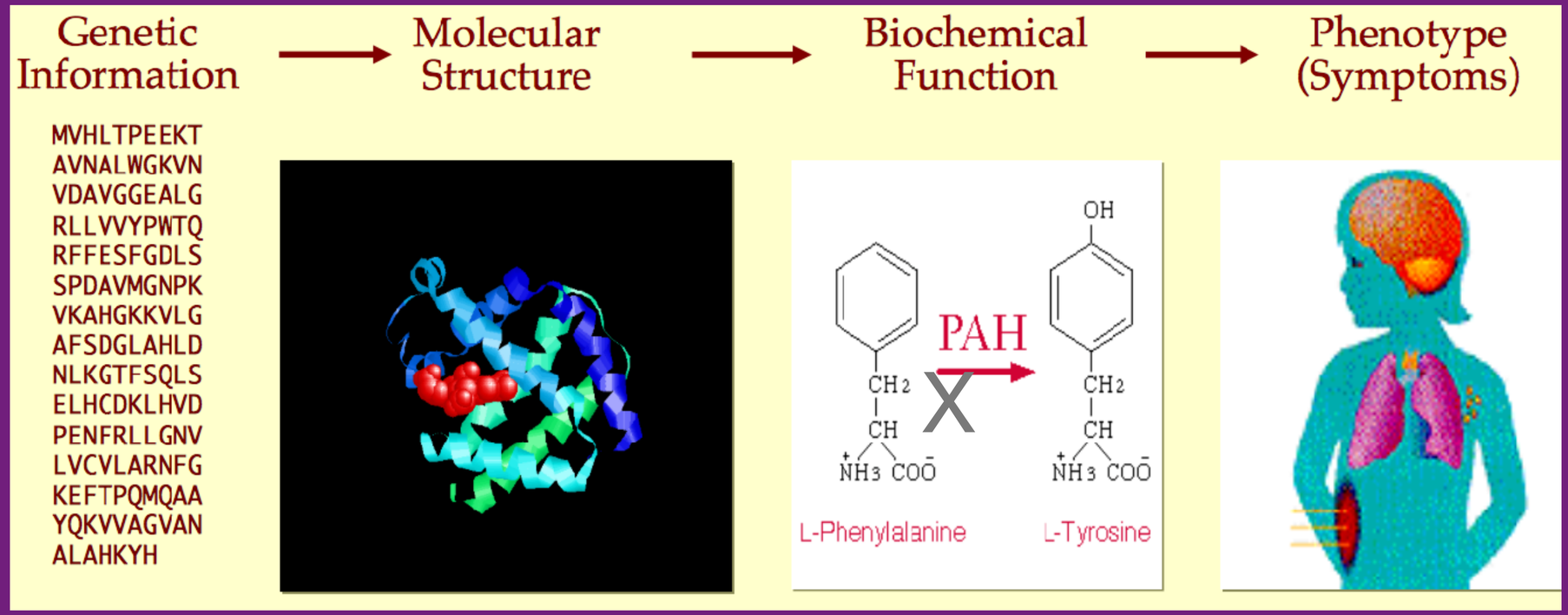
GOALS OF BIOINFORMATICS

LEARN BY COMPARISON AND INFERENCE

- Is Protein A similar to Protein B?
  - Sequence similarity (alignment!)
  - Structure similarity (structural comparison)
  - Co-expression (Microarray data analysis)
  - Any types of correlation (operon-structure, etc)

GOALS OF BIOINFORMATICS

CENTRAL PARADIGM OF BIOINFORMATICS: PREDICT/INFER



- Phenyalaline hydroxylase (PAH) gene adds a hydroxyl group to create tyrosine
  - Tyrosine import to formation of hormones and neurotransmitters
- Mutation in PAH allows toxic buildup of Phenynlalaine
- PKU (Phenylketonuria) - intellectual disability and seizures; low birth weight in babies

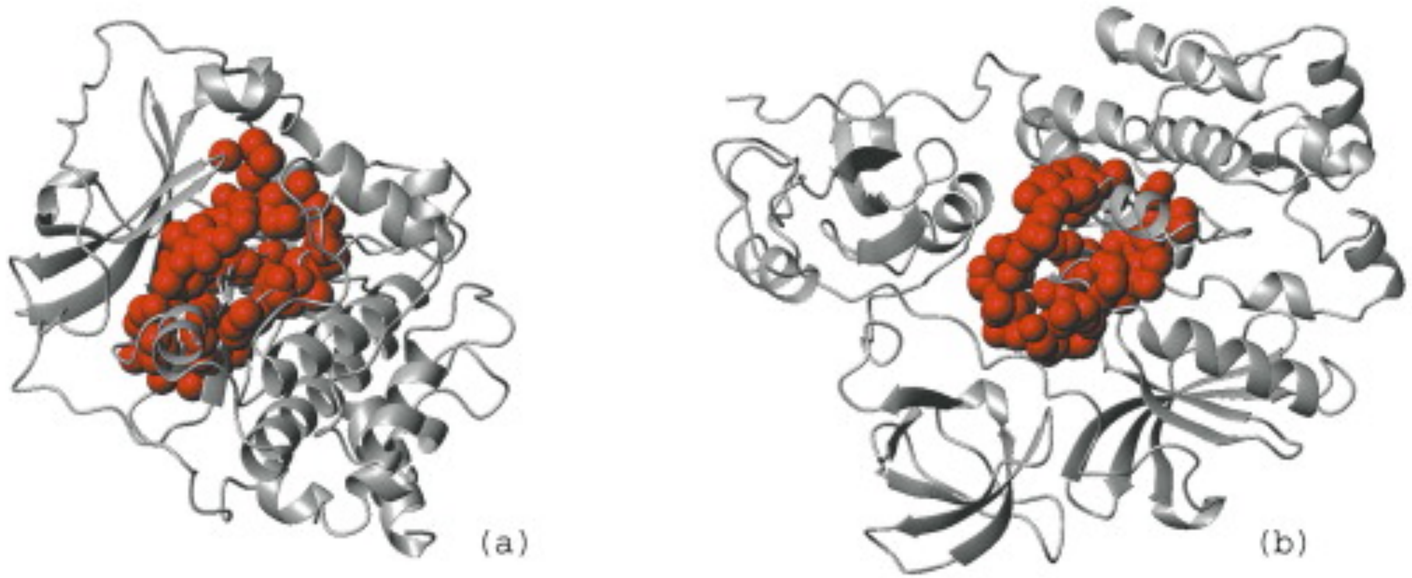
GOALS OF BIOINFORMATICS

- Challenges in bioinformatics (molecular biology)
  - Genetic information is redundant
  - Structural information is redundant
  - Genes and proteins aren’t stable
    - Flexibility important for function
  - Genes have multiple functions
  - Translate 1 dimensional to 3 dimensional data

GOALS OF BIOINFORMATICS

- Challenges: 1D to 3D

IMPORTANT RESIDUES ARE NEIGHBORS IN 3D SPACE, NOT IN 2D



(c)

>1cdk_A

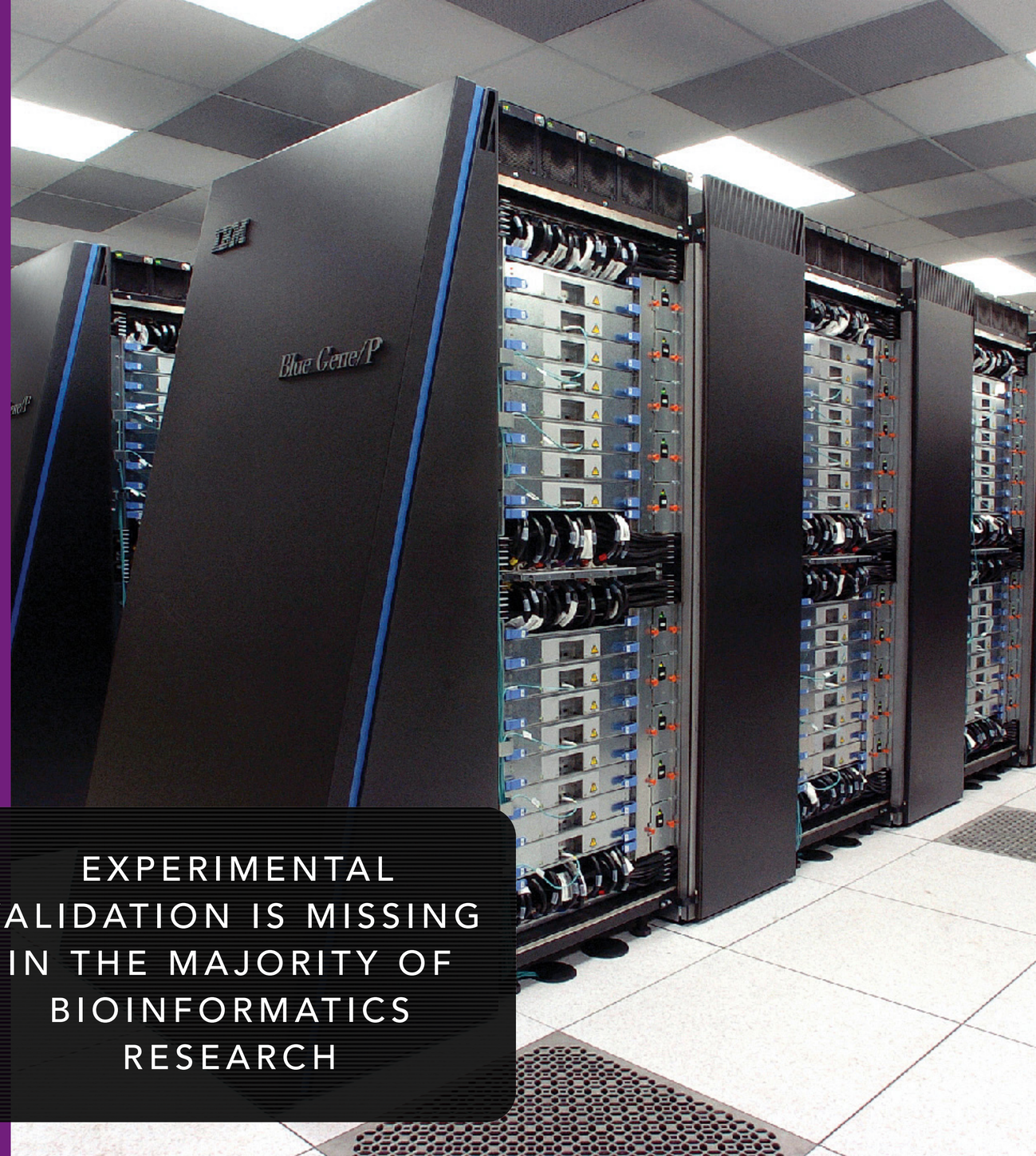
GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHKETGNHFAMKILD
KQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQI
VLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALG
VLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKDGVNDIKNHKWFATT
DWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF

>2src_

MVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRR
ESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYS
KHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPGTMSPEA
FLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMN
YVHRDLRAANILVGENLVCKVADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELT
TKGRVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQXQPGE
NL

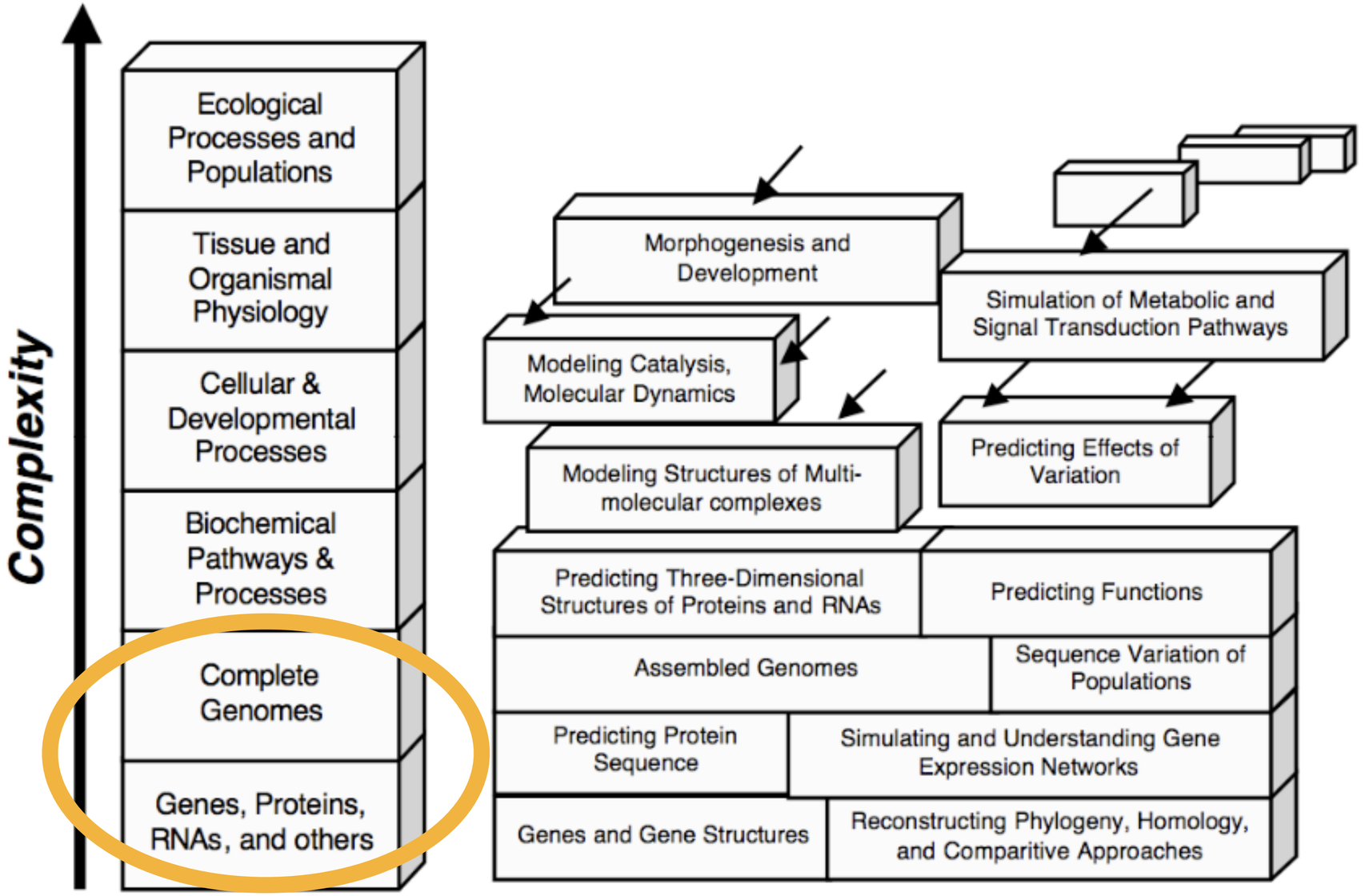
GOALS OF BIOINFORMATICS

- Challenges in bioinformatics (computing)
  - Growth, complexity and volume of biological data
  - Propagation of errors
    - Problem of inference
  - Economics of research
    - Computation is cheap (and doesn't go on vacation)
    - People are expensive
    - Wet lab work is expensive



EXPERIMENTAL VALIDATION IS MISSING IN THE MAJORITY OF BIOINFORMATICS RESEARCH

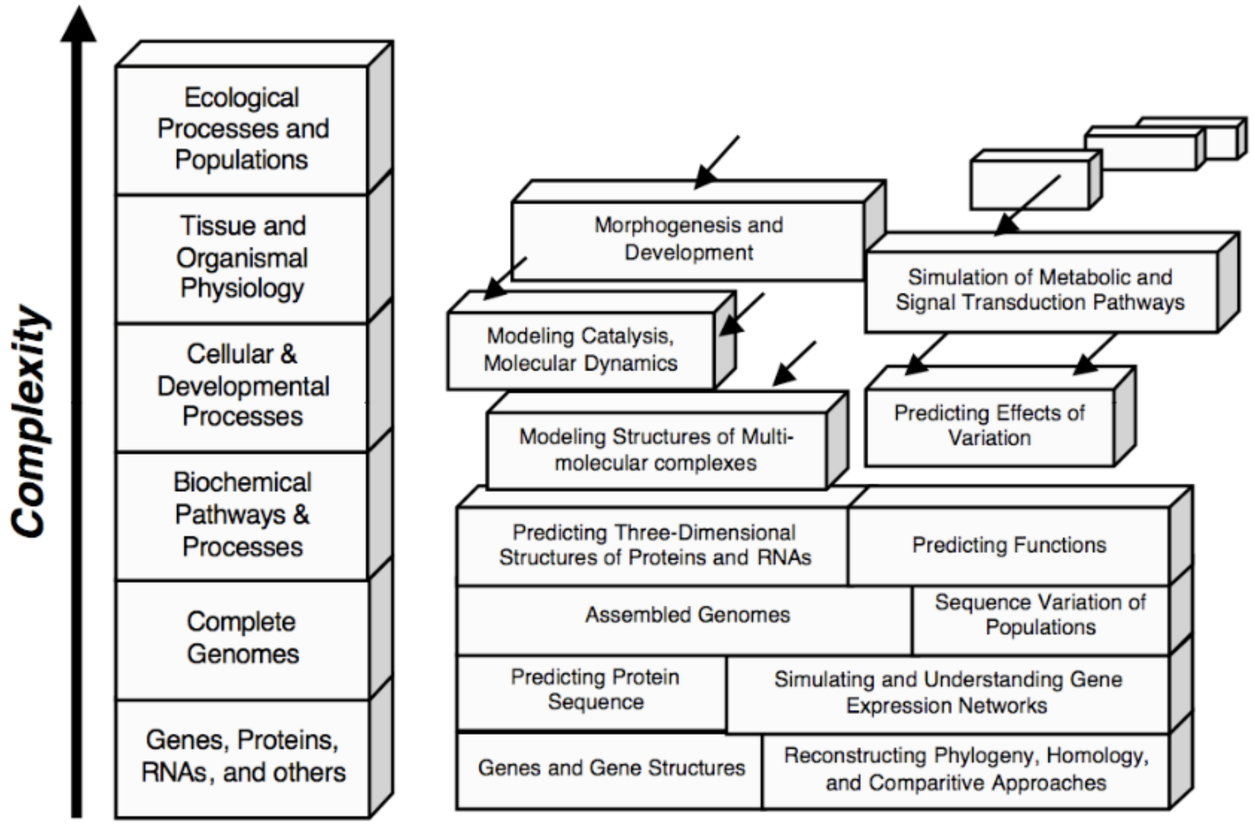
GOALS OF BIOINFORMATICS



Impact of Bioinformatics

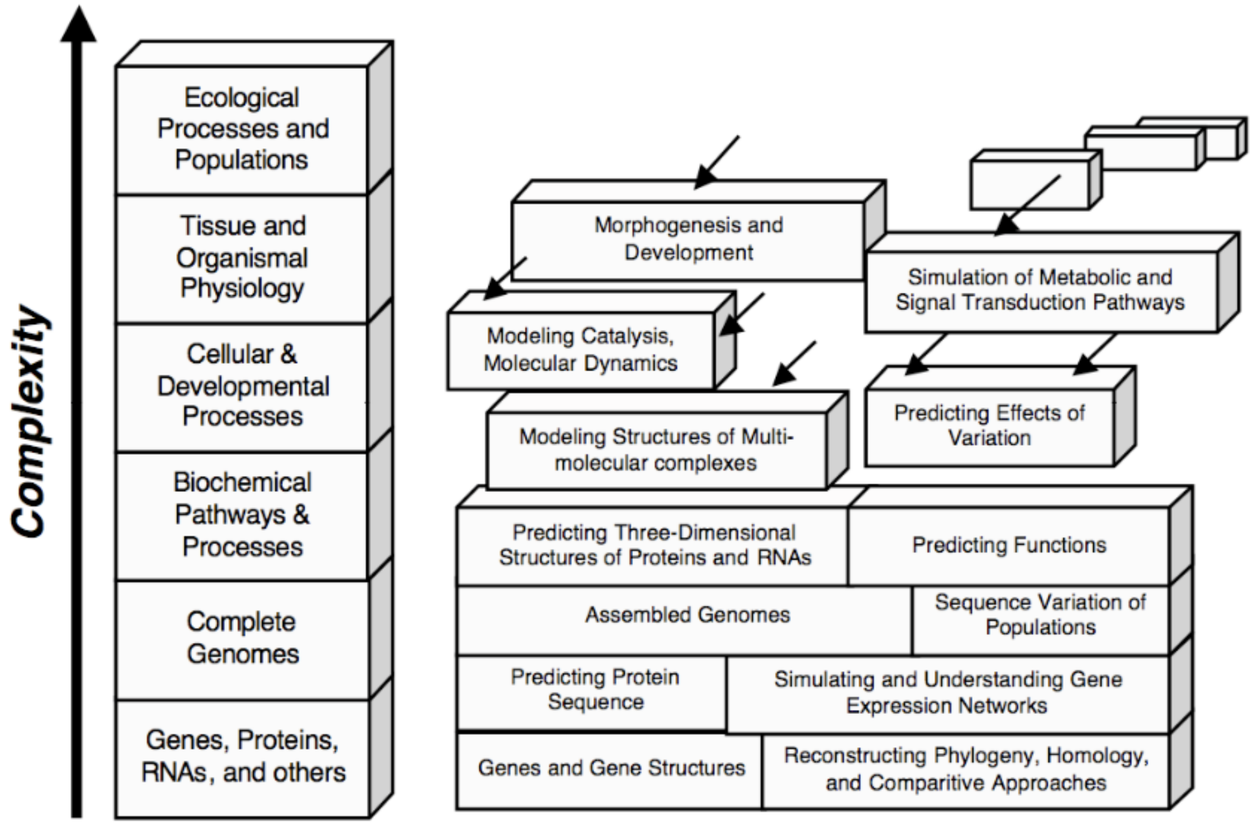
IMPACT OF BIOINFORMATICS

- On Biological sciences
  - Large scale experimental techniques
    - Analyze an entire genome
  - Information growth
  - Integration of data between fields



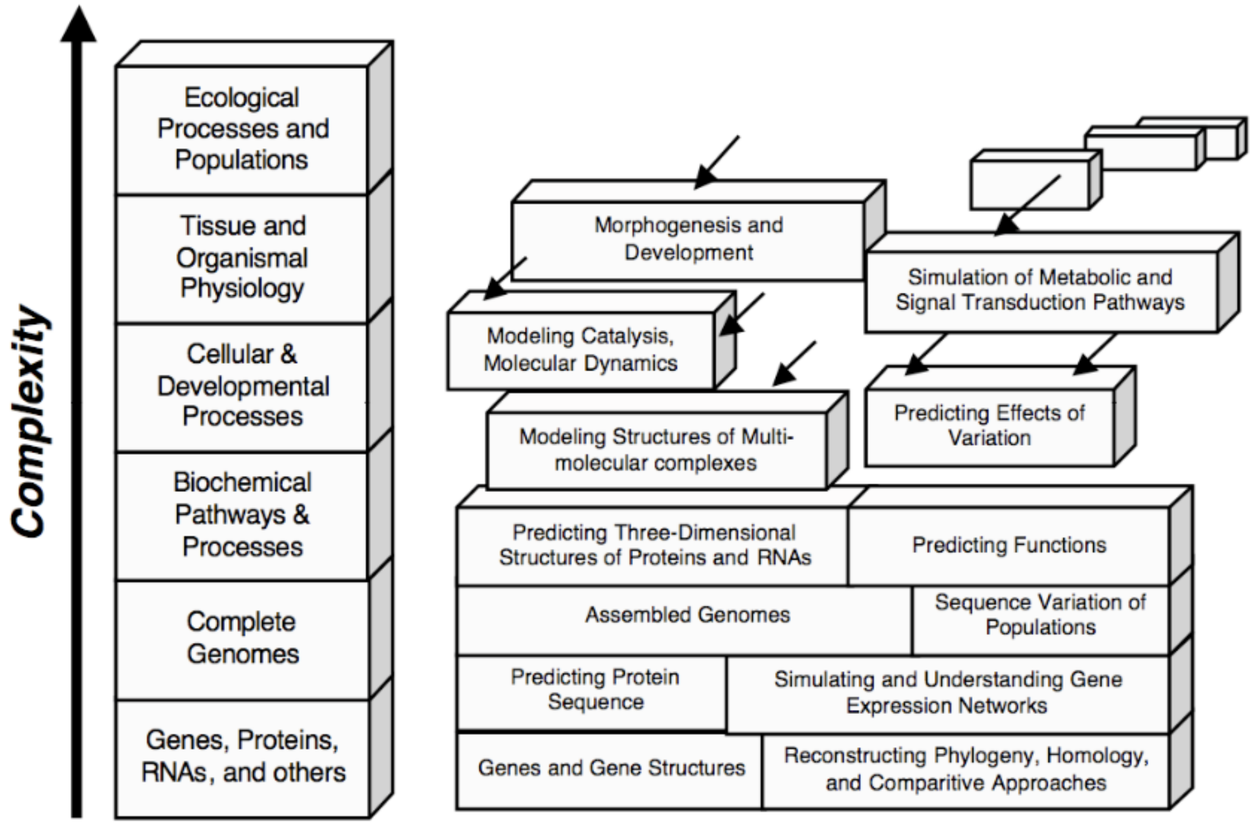
IMPACT OF BIOINFORMATICS

- On computational sciences
  - New algorithmic and statistical problems
  - Big data

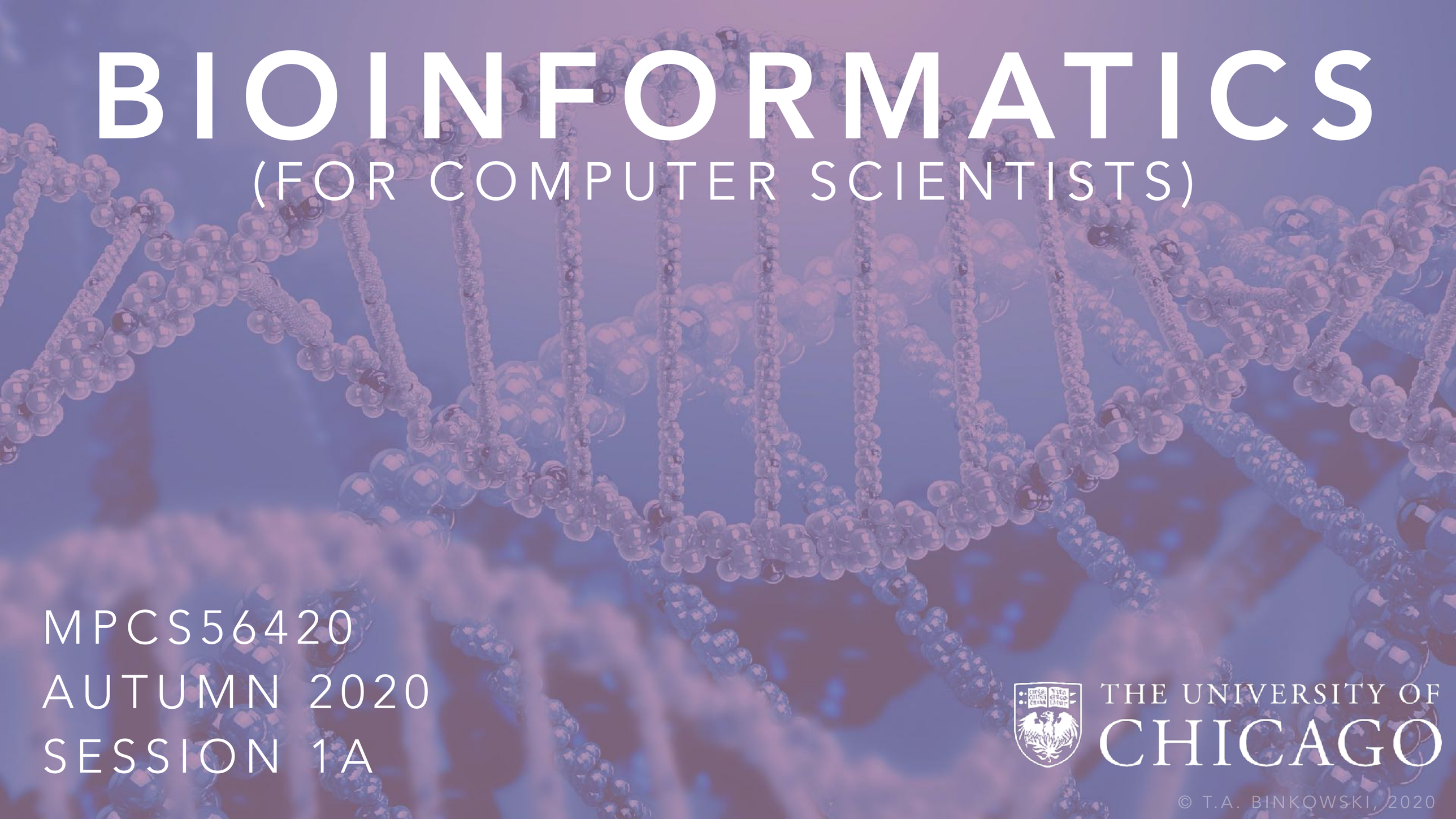


IMPACT OF BIOINFORMATICS

- On Medical sciences
  - Translational research
  - Personalized treatment



SESSION 1A



BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
AUTUMN 2020
SESSION 1A

THE UNIVERSITY OF CHICAGO


BIOLOGY 101

A Timeline of Molecular Biology

TIMELINE OF MOLECULAR BIOLOGY

- 1665
  - Robert Hooke discovered organisms are made up of cells
- 1859
  - Charles Darwin published the "On the Origin of Species"
- 1865
  - Gregor Mendel investigated "traits" passed from parents to prodigy and coined the terms dominant and recessive traits

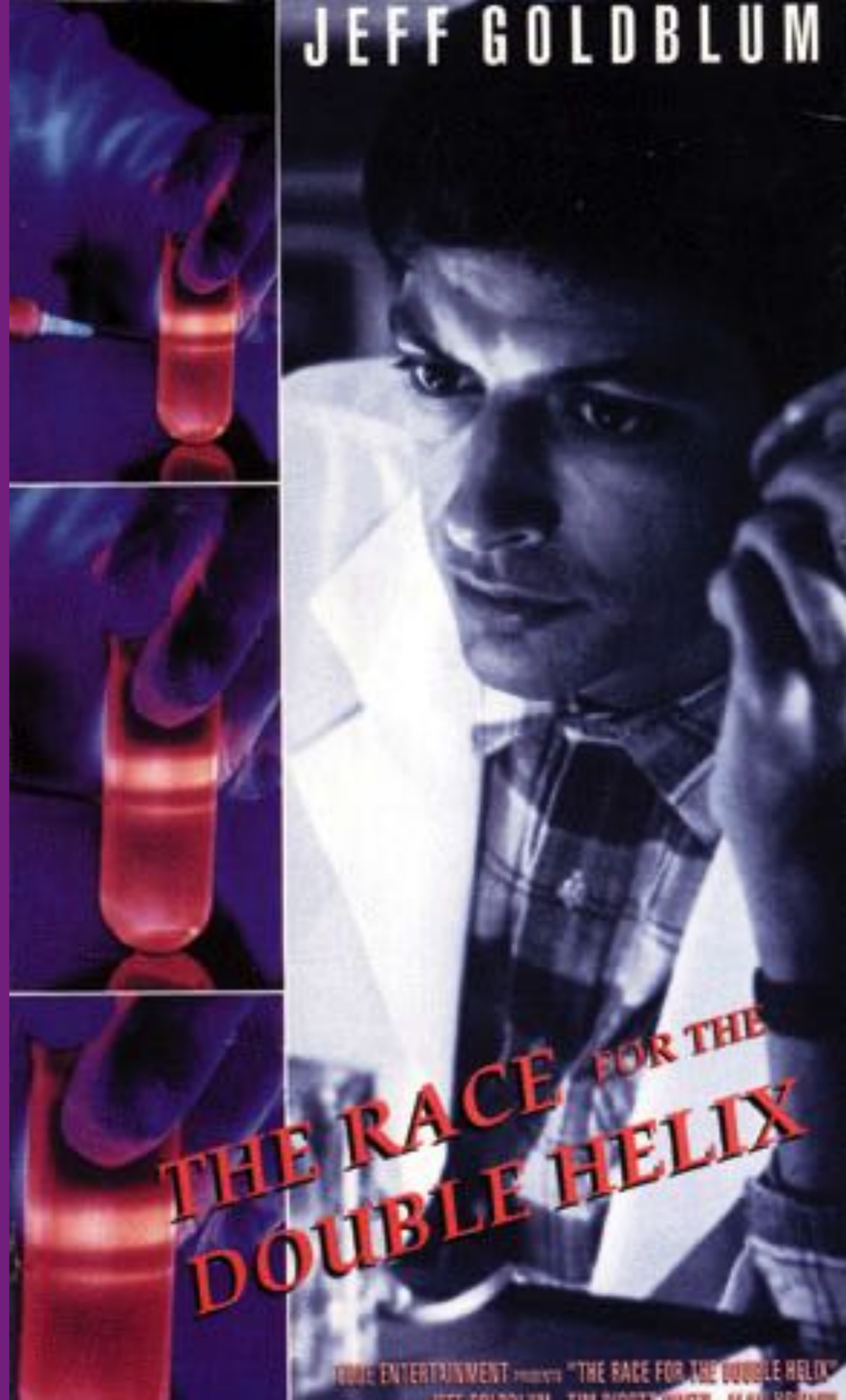
TIMELINE OF MOLECULAR BIOLOGY

- 1900
  - Chemical structures of all 20 amino acids had been identified
- 1902
  - Emil Hermann Fischer wins Nobel prize: showed amino acids are linked and form proteins
- 1941
  - George Beadle and Edward Tatum identify that genes make proteins


Emil
Fischer

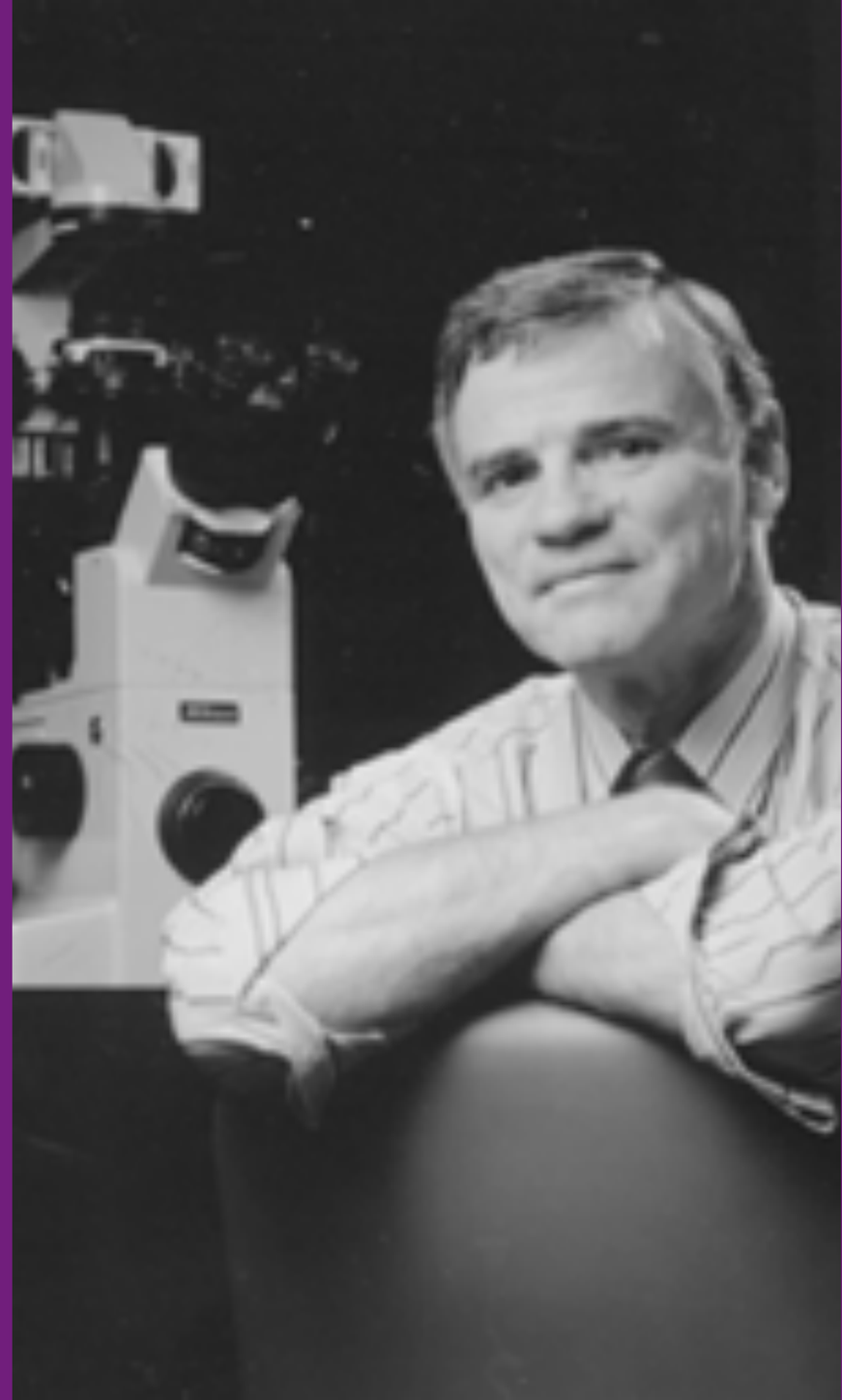
TIMELINE OF MOLECULAR BIOLOGY

- 1952-1953
  - Rosalind Franklin produce X-ray diffraction of DNA
  - James D. Watson and Francis H. C. Crick deduced the double helical structure of DNA (Race for the Double Helix movie)
- 1956
  - George Emil Palade showed the site of enzymes manufacturing in the cytoplasm is made on RNA organelles called ribosomes
- 1977
  - Phillip Sharp and Richard Roberts demonstrated that pre-mRNA is processed by the excision of introns and exons are spliced together



TIMELINE OF MOLECULAR BIOLOGY

- 1977
  - Phillip Sharp and Richard Roberts demonstrated that pre-mRNA is processed by the excision of introns and exons are spliced together
- 1986
  - Leroy Hood developed automated sequencing mechanism
- 1986
  - Human Genome Initiative announced



TIMELINE OF MOLECULAR BIOLOGY

- 1990
  - The 15 year Human Genome project is launched
- 1995
  - John Craig Venter: First bacterial genomes sequenced
  - Automated fluorescent sequencing instruments and robotic operations

TIMELINE OF MOLECULAR BIOLOGY

- 1996
  - First eukaryotic genome-yeast-sequenced
- 1997
  - E. Coli sequenced
- 1999
  - First human chromosome (22) sequenced

Nature. 1999 Dec 2;402(6761):489-95.

The DNA sequence of human chromosome 22.

Dunham I[1], Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM
Smink LJ, Ainscough R, Almeida JP, Babbage A, Bagguley C, Bailey J, Barlow K, Bates K
Bird CP, Blakey S, Bridgeman AM, Buck D, Burgess J, Burrill WD, O'Brien KP, et al.

Author information

Erratum in
Nature 2000 Apr 20;404(6780):904.

Abstract

Knowledge of the complete genomic DNA sequence of an organism allows a syster
to defining its genetic components. The genomic sequence provides access to the
structures of all genes, including those without known function, their control elemen
inference, the proteins they encode, as well as all other biologically important seque
Furthermore, the sequence is a rich and permanent source of information for the de
biological studies of the organism and for the study of evolution through cross-speci
comparison. The power of this approach has been amply demonstrated by the dete
the sequences of a number of microbial and model organisms. The next step is to c
complete sequence of the entire human genome. Here we report the sequence of th
part of human chromosome 22. The sequence obtained consists of 12 contiguous s
spanning 33.4 megabases, contains at least 545 genes and 134 pseudogenes, and
first view of the complex chromosomal landscapes that will be found in the rest of th

Comment in
Do we need a huge new centre to annotate the human genome? [Nature. 2000]
Tiny chromosome is rich in genes and medical promise. [Nature. 1999]
'Finishing' success marks major genome sequencing milestone...as researchers po
data. [Nature. 1999]
The book of genes. [Nature. 1999]

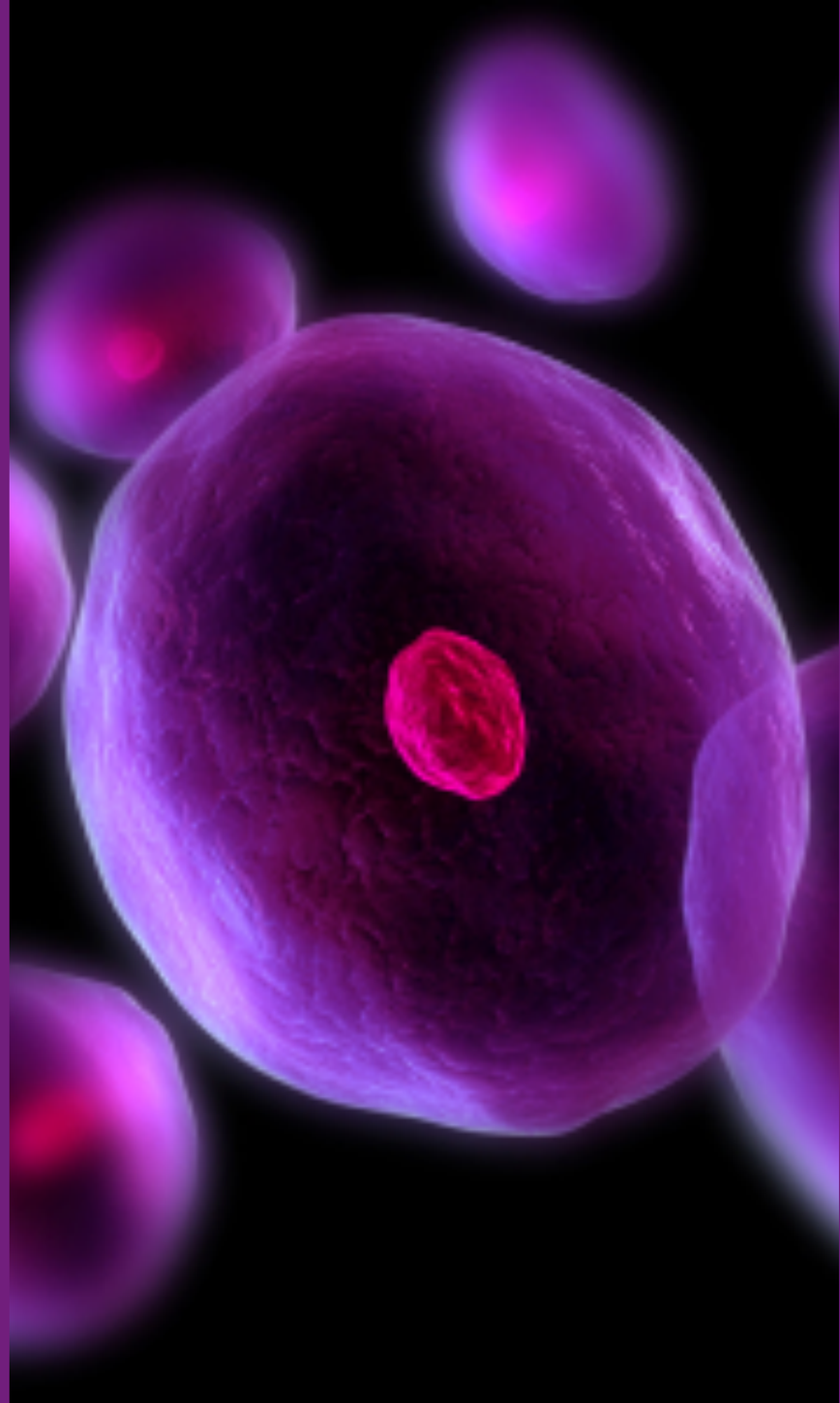
TIMELINE OF MOLECULAR BIOLOGY

- 2000
  - Complete sequence of the Drosophila melanogaster genome
- 2001
  - International Human Genome Sequencing publishes first draft of the sequence of the human genome
- 2003
  - Human Genome Project “completed”

Cells

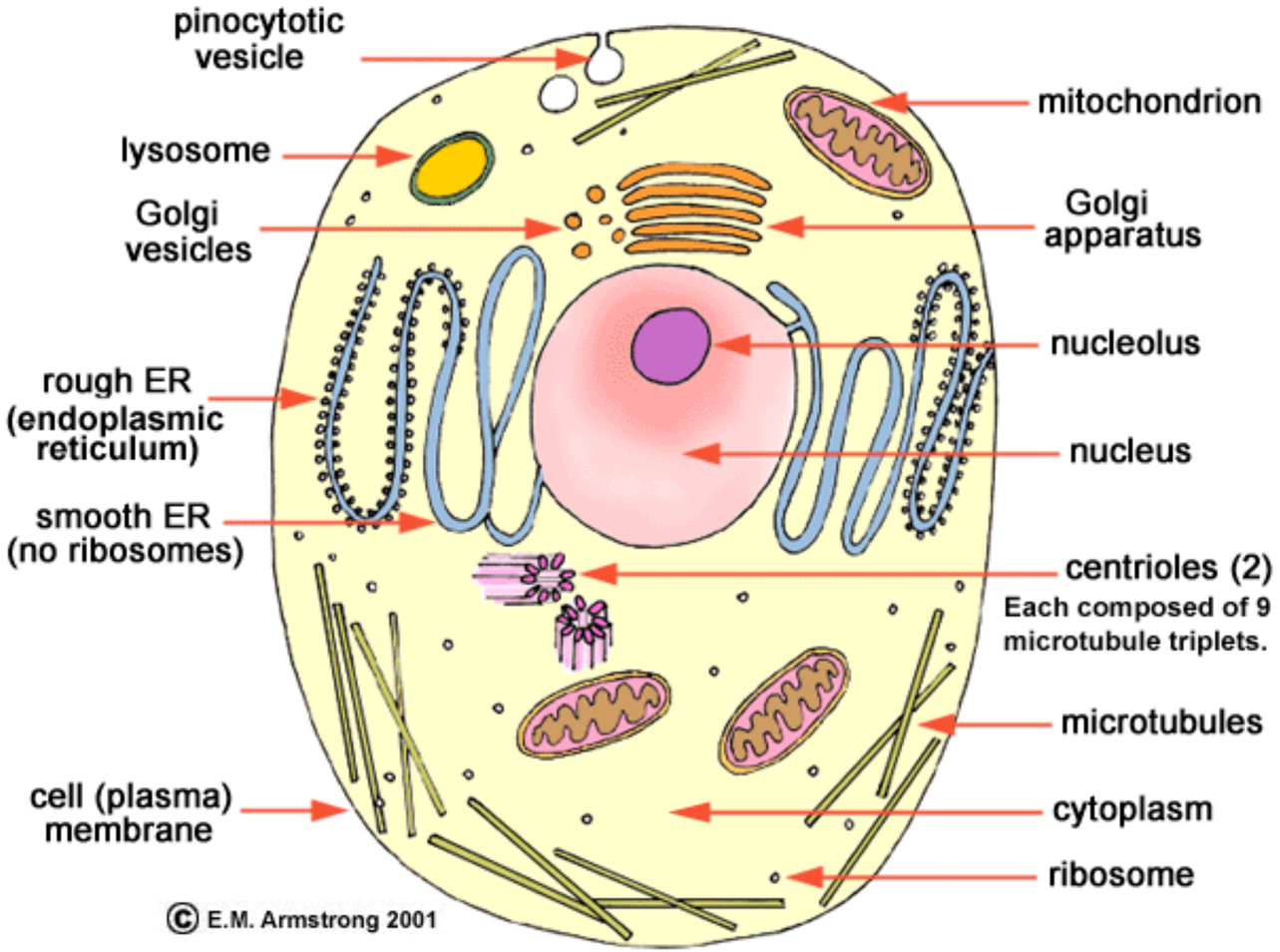
CELLS

- Cell Theory
  - All living organisms are composed of one or more cells
  - The cell is the most basic unit of life
  - All cells arise from pre-existing, living cells
  - Cells contain heredity information that is passed during cell division

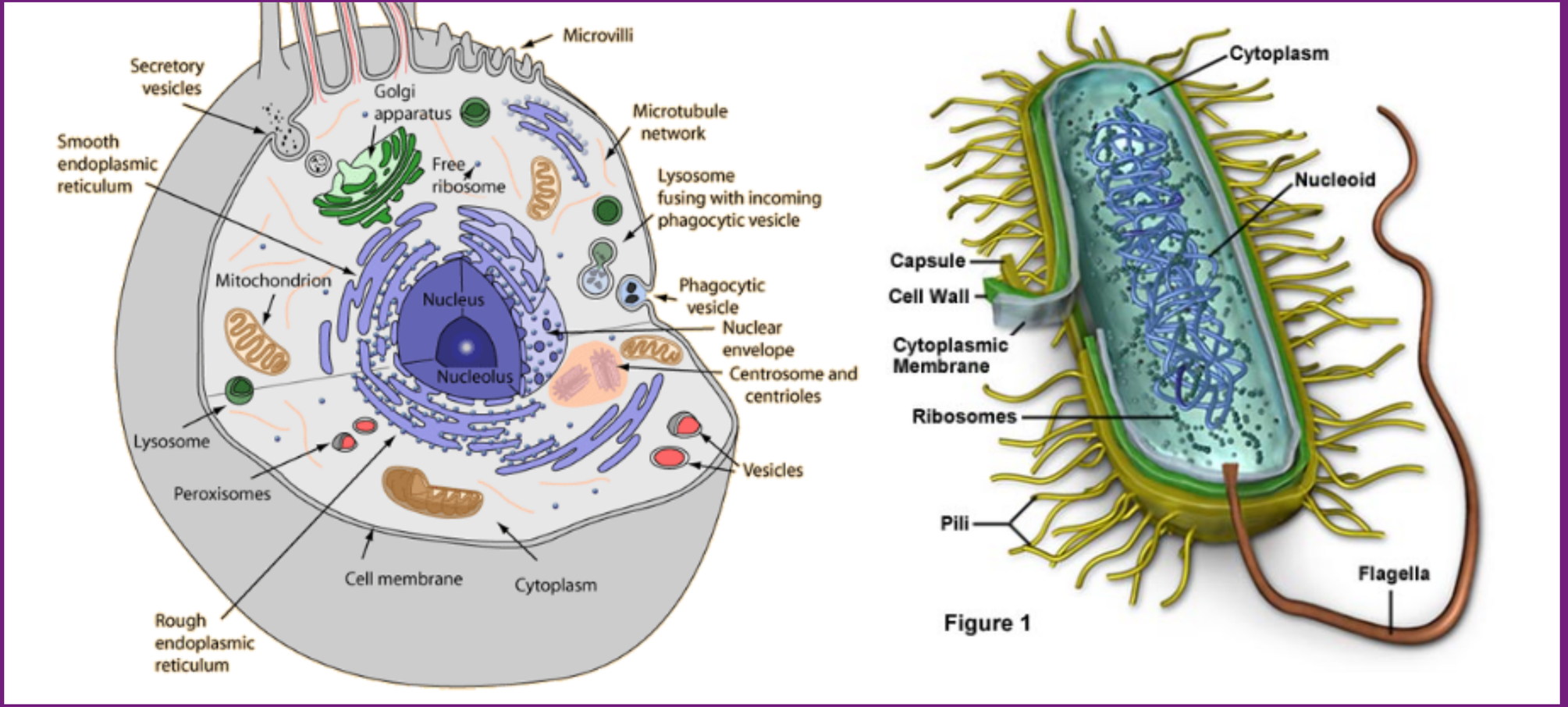


CELLS

- Organisms can be of single cells or multiple cells
  - Most living organisms are single cells (e.g. E. coli, yeast)
  - Multicellular organisms have trillions of cells
- Cells have many parts
  - Organelles - specialized structures that perform certain tasks within the cell



CELLS

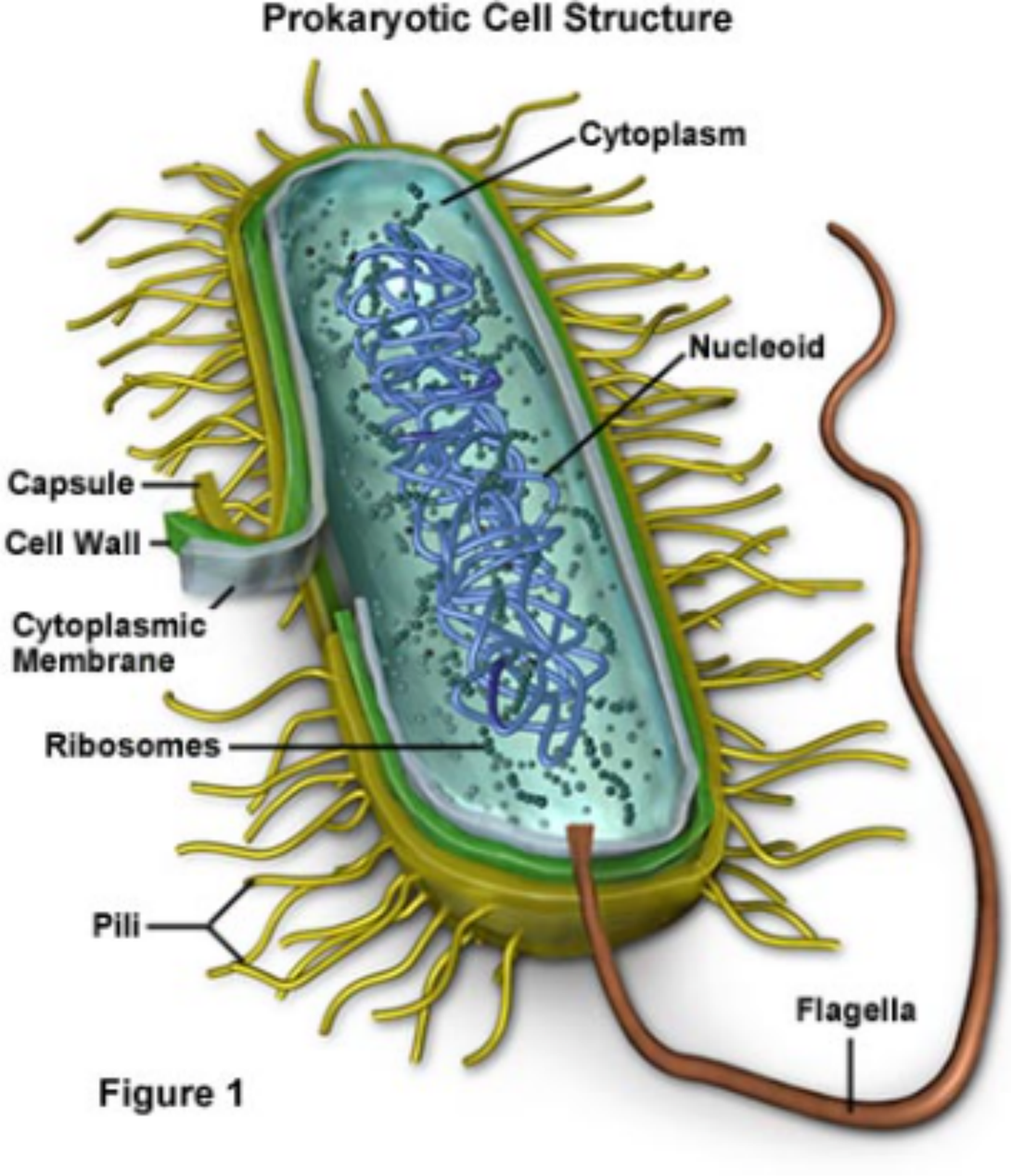


- Organisms classified by type of cells: prokaryotic, eukaryotic

CELLS

PROKARYOTES

- Lack a cell nucleus
- Lack membrane bound organelles
- DNA in a single loop
- Mostly single-cellular (some multi-cellular)
- Two groups
  - Bacteria
  - Archaea (extremophiles)

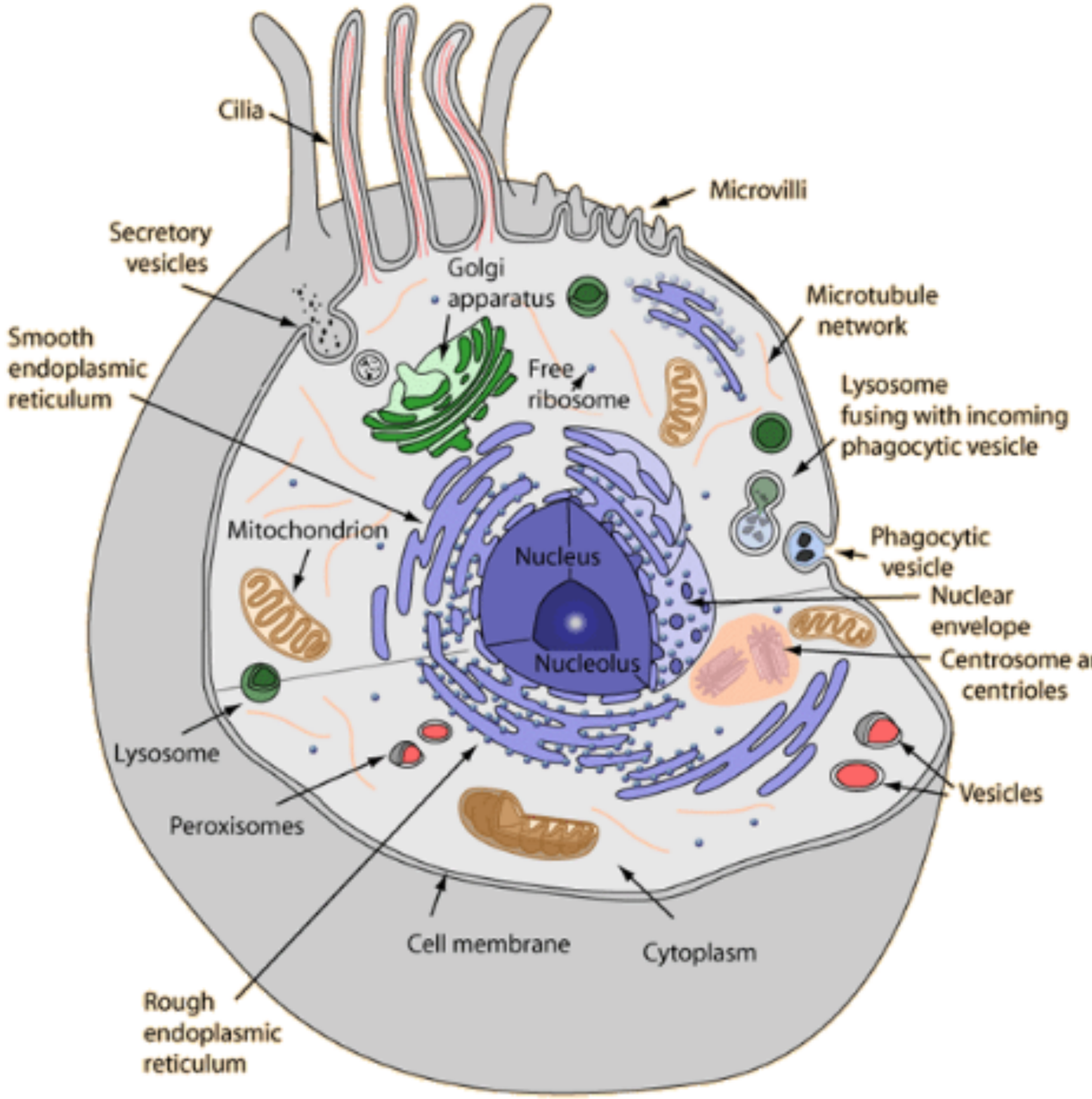


Prokaryotic cell structure

CELLS

EUKARYOTES

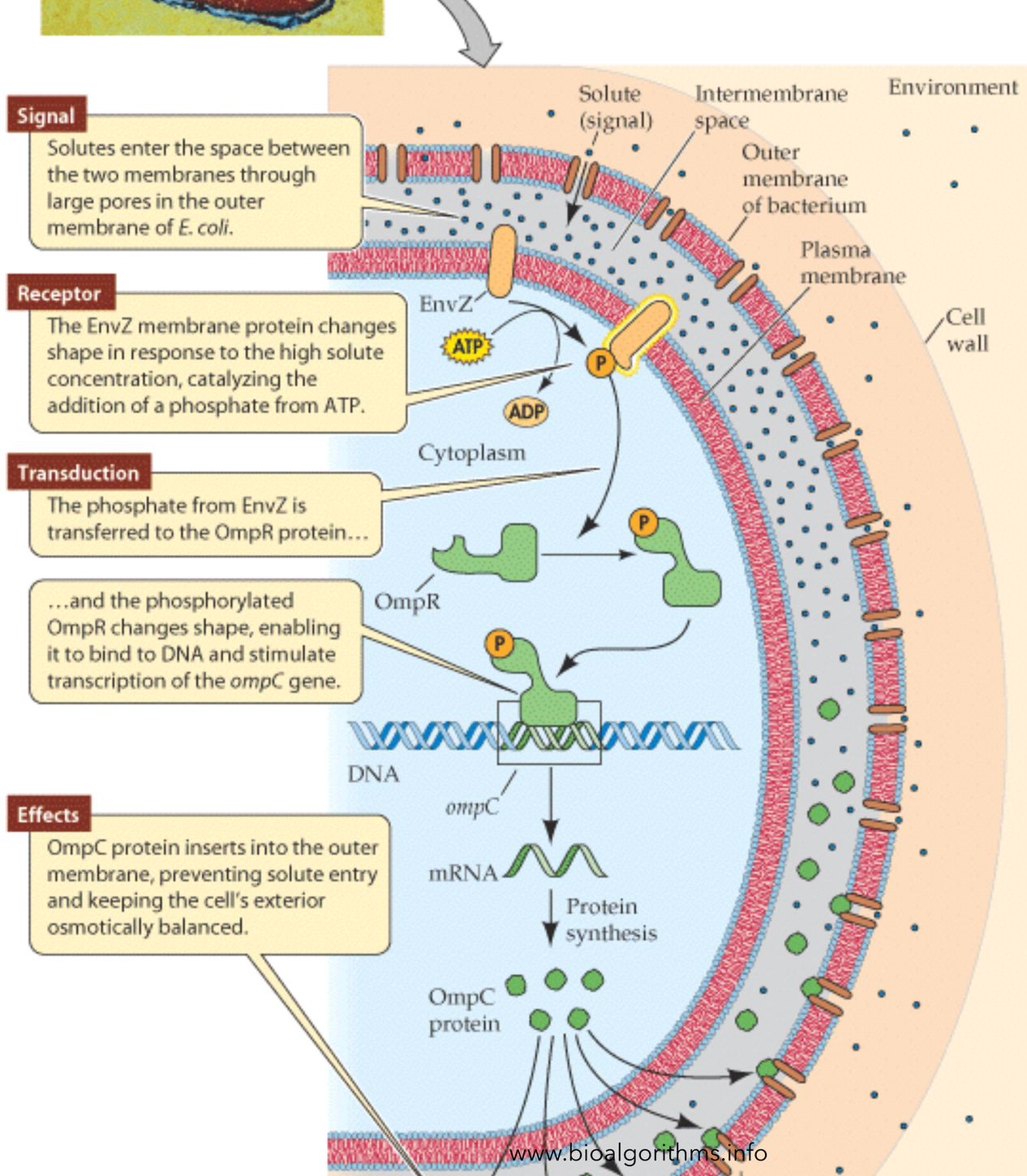
- Cells posses membrane-bound organelles
- Genetic material contained in nucleus
- DNA organized into chromosomes
- Can be single or multi-cellular
- Examples:
  - Animals, plants, fungi



http://hyperphysics.phy-astr.gsu.edu/hbase/biology/imgbio/cellhlabel.gif

CELLS

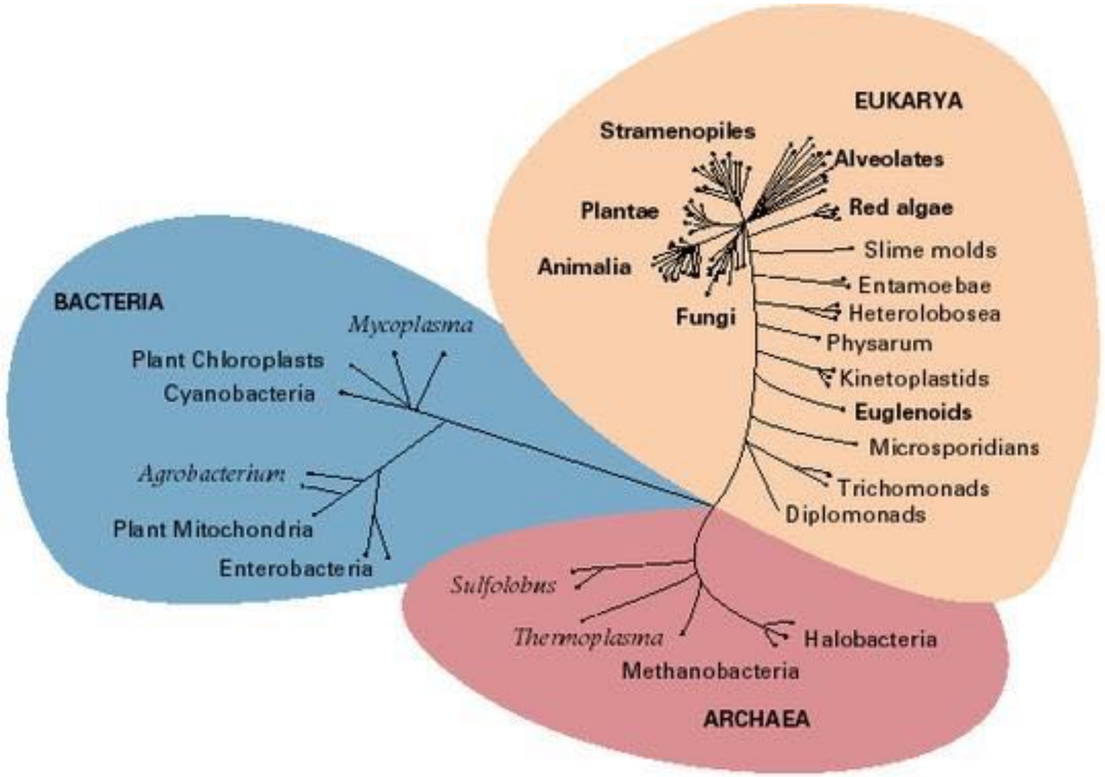
- Cells make decision through complex networks of chemical reactions, called pathways
  - Synthesize new materials
  - Break other materials down for spare parts
  - Signal to eat or die



CELLS

- According to the most recent evidence, there are three main branches to the tree of life

EVOLVING MODEL AS WE LEARN MORE (NEW ORGANISMS, SEQUENCES)



CELLS

- Global oceanic sampling expedition

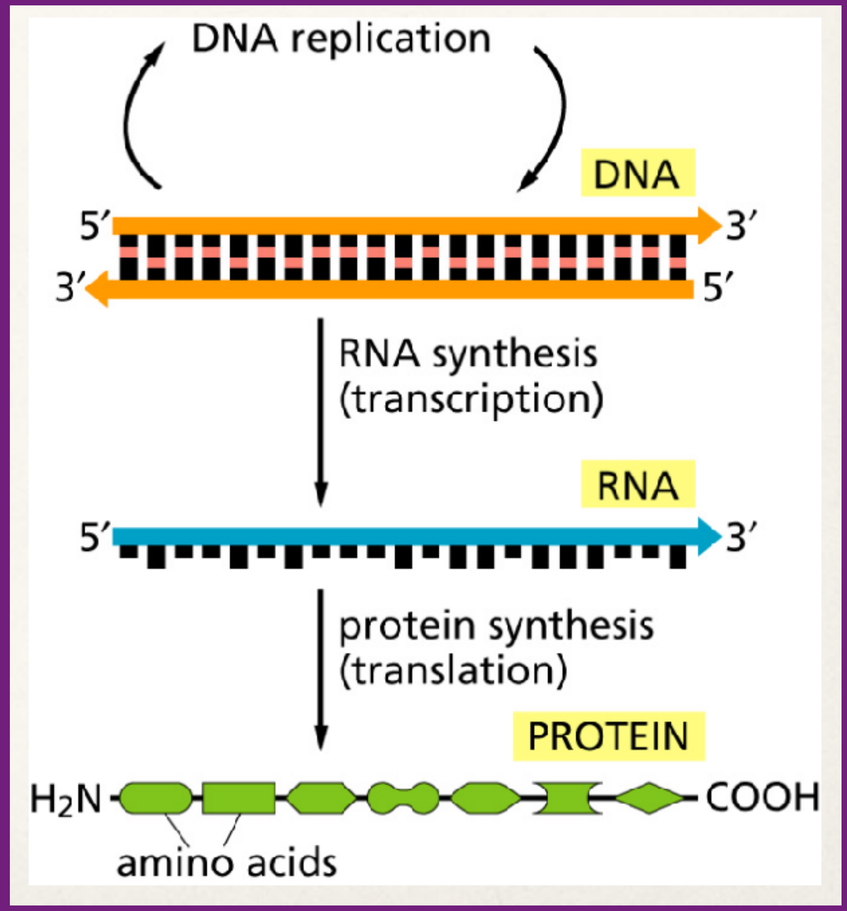
Central Dogma of Molecular Biology

CENTRAL DOGMA OF MOLECULAR BIOLOGY

DNA CAN REPLICATE

INFORMATION IN DNA IS PASSED TO RNA

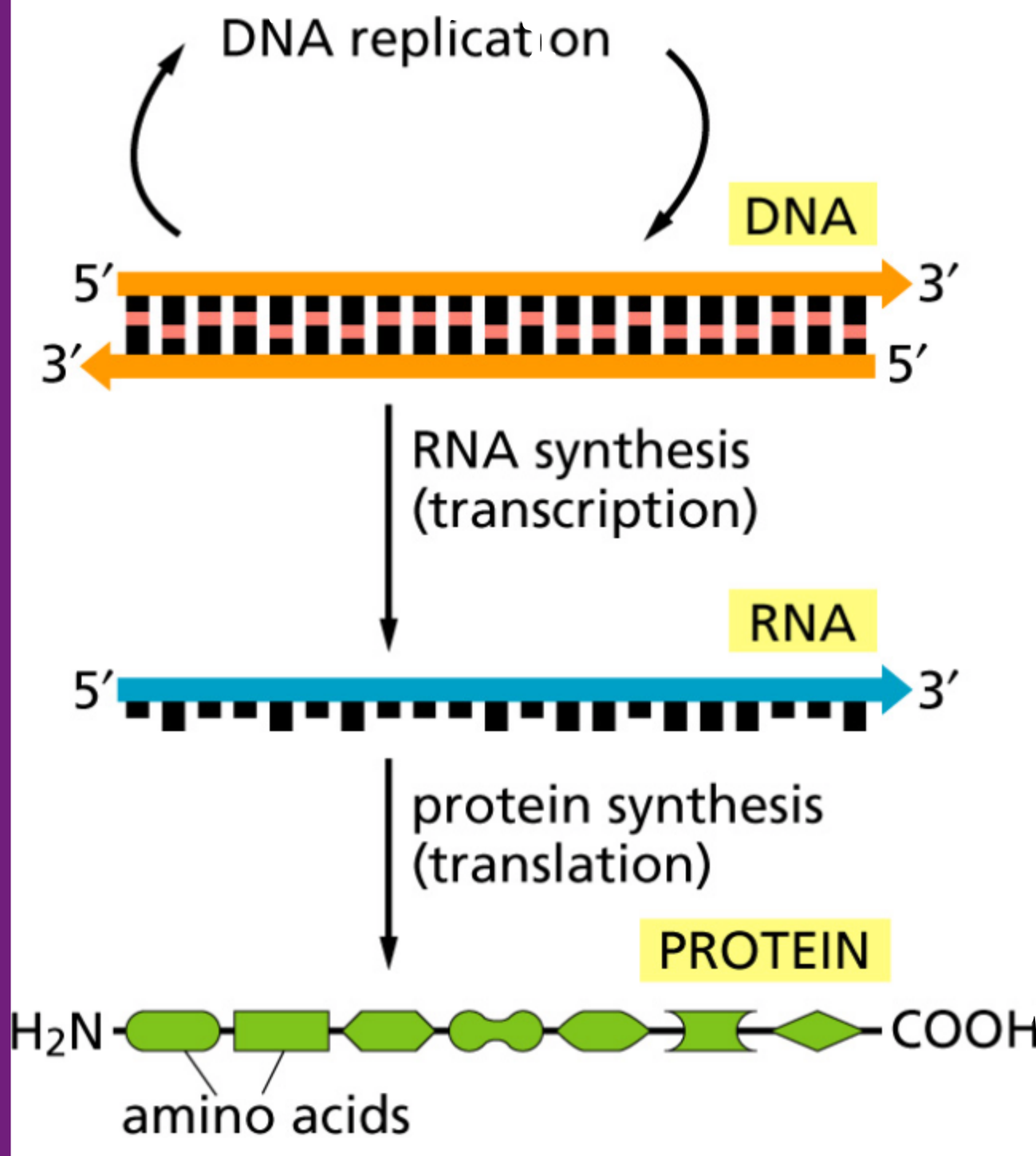
INFORMATION IN RNA IS PASSED TO PROTEINS



- Describes the flow of information in cells

CENTRAL DOGMA OF MOLECULAR BIOLOGY

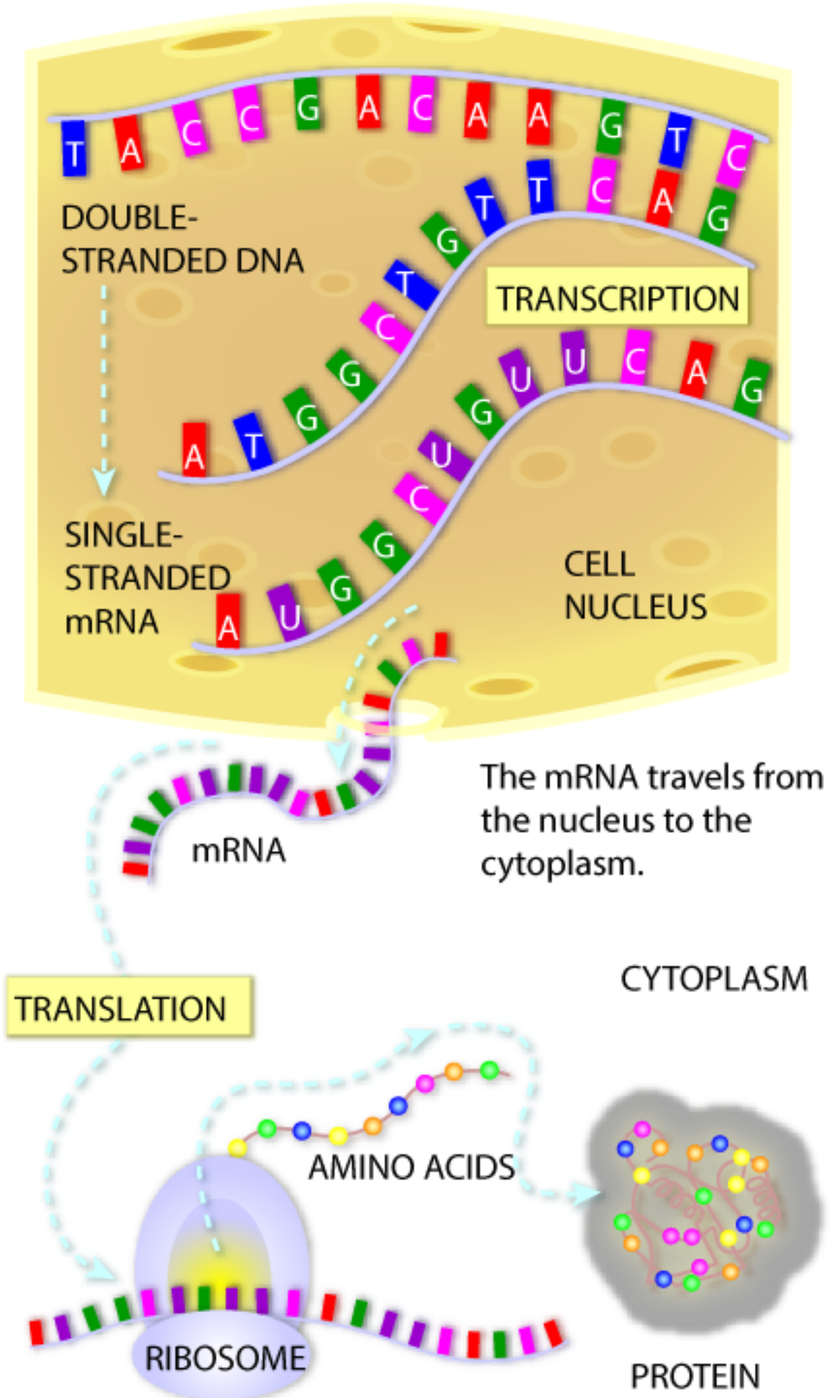
- All forms of life follow this scheme
  - Slight variations between organisms
- Genomic DNA encodes all the molecules necessary for life of an organism



CENTRAL DOGMA OF MOLECULAR BIOLOGY

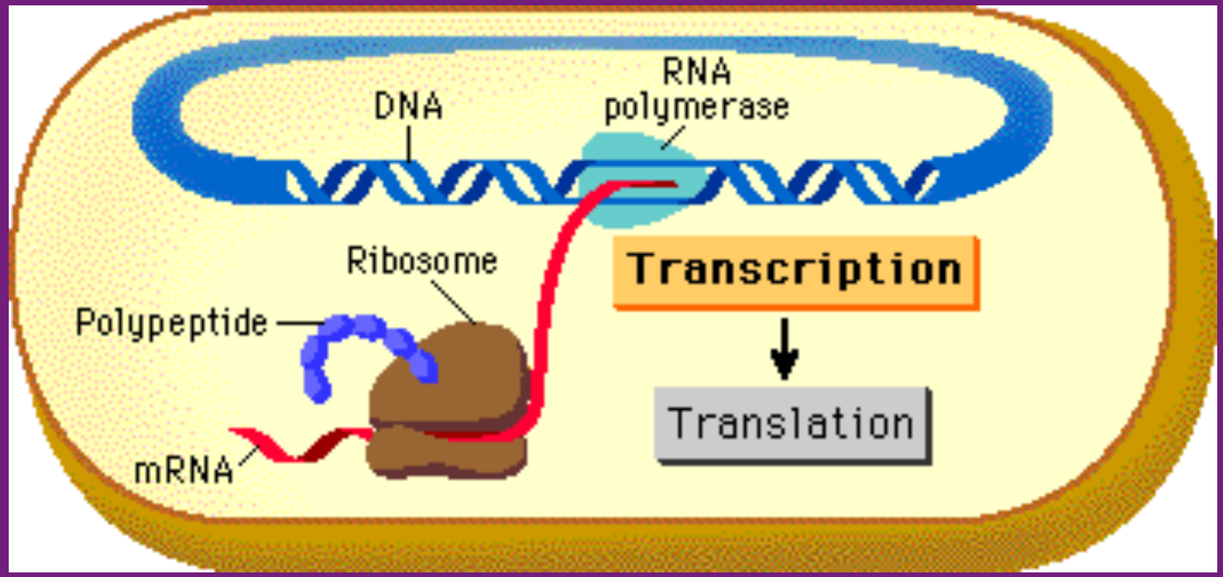
CENTRAL DOGMA OF MOLECULAR BIOLOGY

- Transcription
  - Synthesis of an RNA copy of a segment of DNA
  - RNA is synthesized by the enzyme RNA polymerase
- Translation
  - Ribosome reads the mRNA sequence
  - Translates to amino acid sequence of the protein



http://learn.genetics.utah.edu/content/molecules/transcribe/

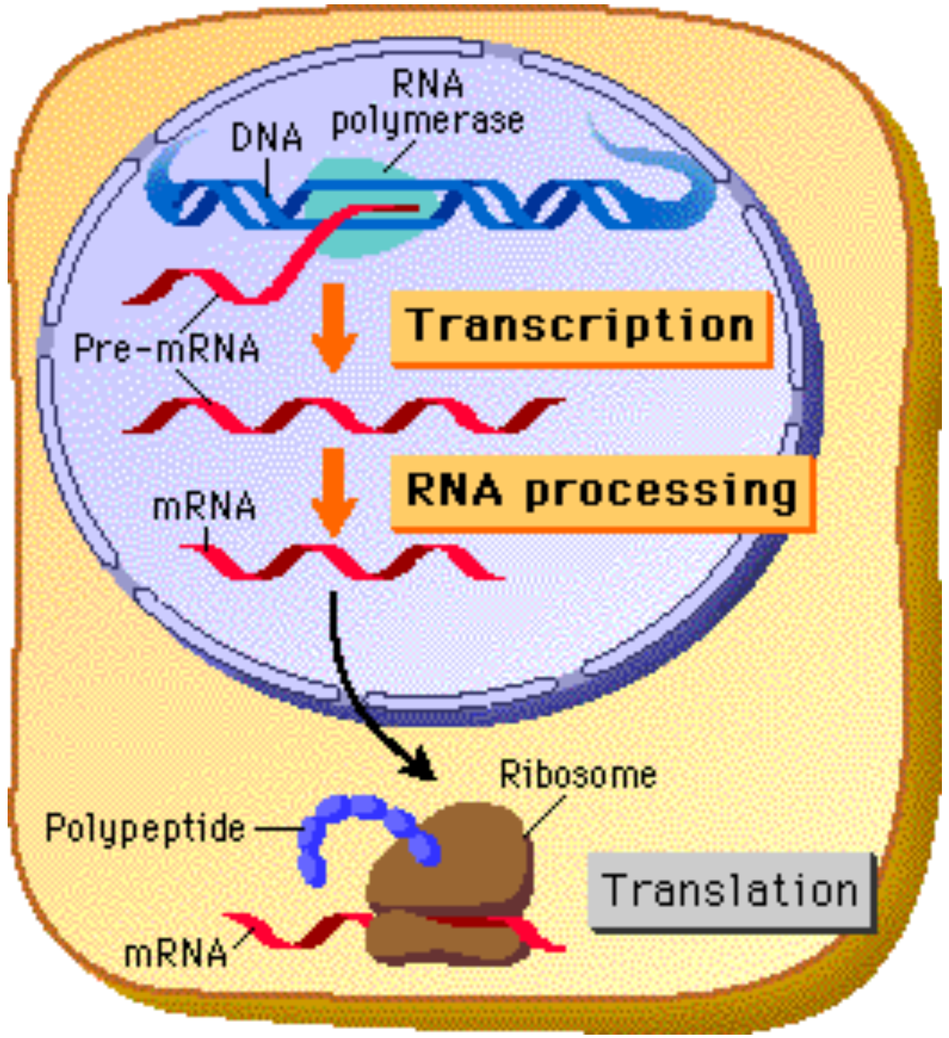
CENTRAL DOGMA OF MOLECULAR BIOLOGY



- Prokaryotic cell
  - Transcripts are immediately translated (no nucleus to cross)

CENTRAL DOGMA OF MOLECULAR BIOLOGY

- Eukaryotic cell
  - Transcription occurs in nucleus
  - Pre-mRNA produced
  - RNA processing
  - Mature mRNA exits
  - Translated in the cytoplasm



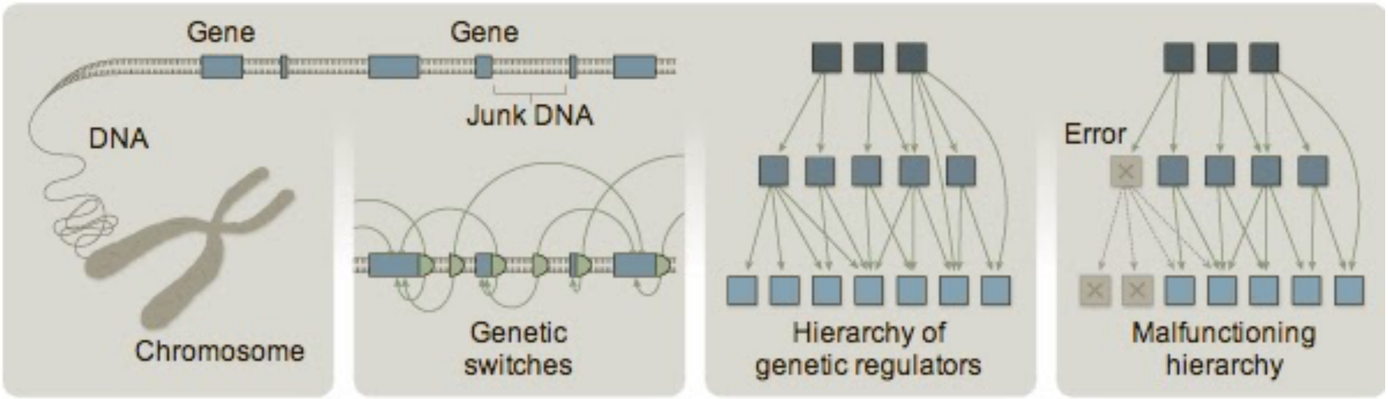
CENTRAL DOGMA OF MOLECULAR BIOLOGY

- Not all genetic information in DNA encodes proteins
- Non-coding “junk” DNA
  - 98% is non-coding in human genome
  - Play other roles
    - Regulation
    - Transcription factor site
    - Operators, promoters
    - Undiscovered functionality

CENTRAL DOGMA OF MOLECULAR BIOLOGY

Rethinking 'Junk' DNA

A large group of scientists has found that so-called junk DNA, which makes up most of the human genome, does much more than previously thought. Related Article »

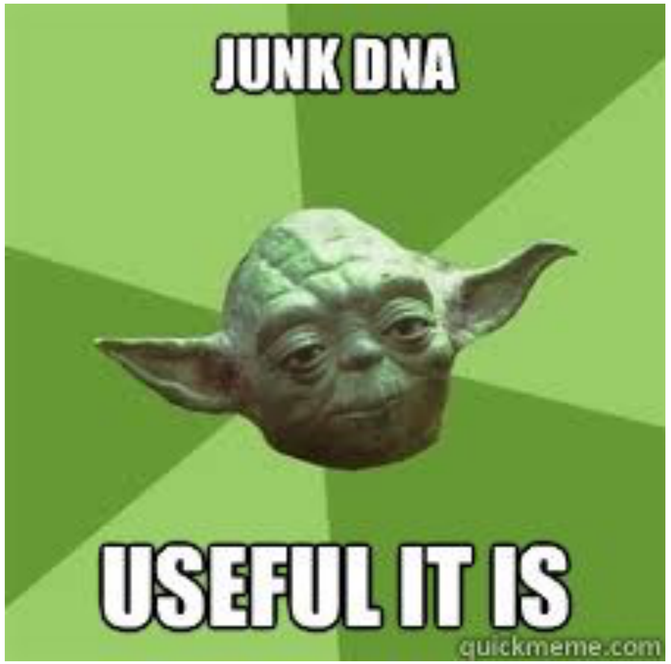


GENES
Each human cell contains about 10 feet of DNA, coiled into a dense tangle. But only a very small percentage of DNA encodes genes, which control inherited traits like eye color, blood type and so on.

JUNK DNA
Stretches of DNA around and between genes seemed to do nothing, and were called junk DNA. But now researchers think that the junk DNA contains a large number of tiny genetic switches, controlling how genes function within the cell.

REGULATION
The many genetic regulators seem to be arranged in a complex and redundant hierarchy. Scientists are only beginning to map and understand this network, which regulates how cells, organs and tissues behave.

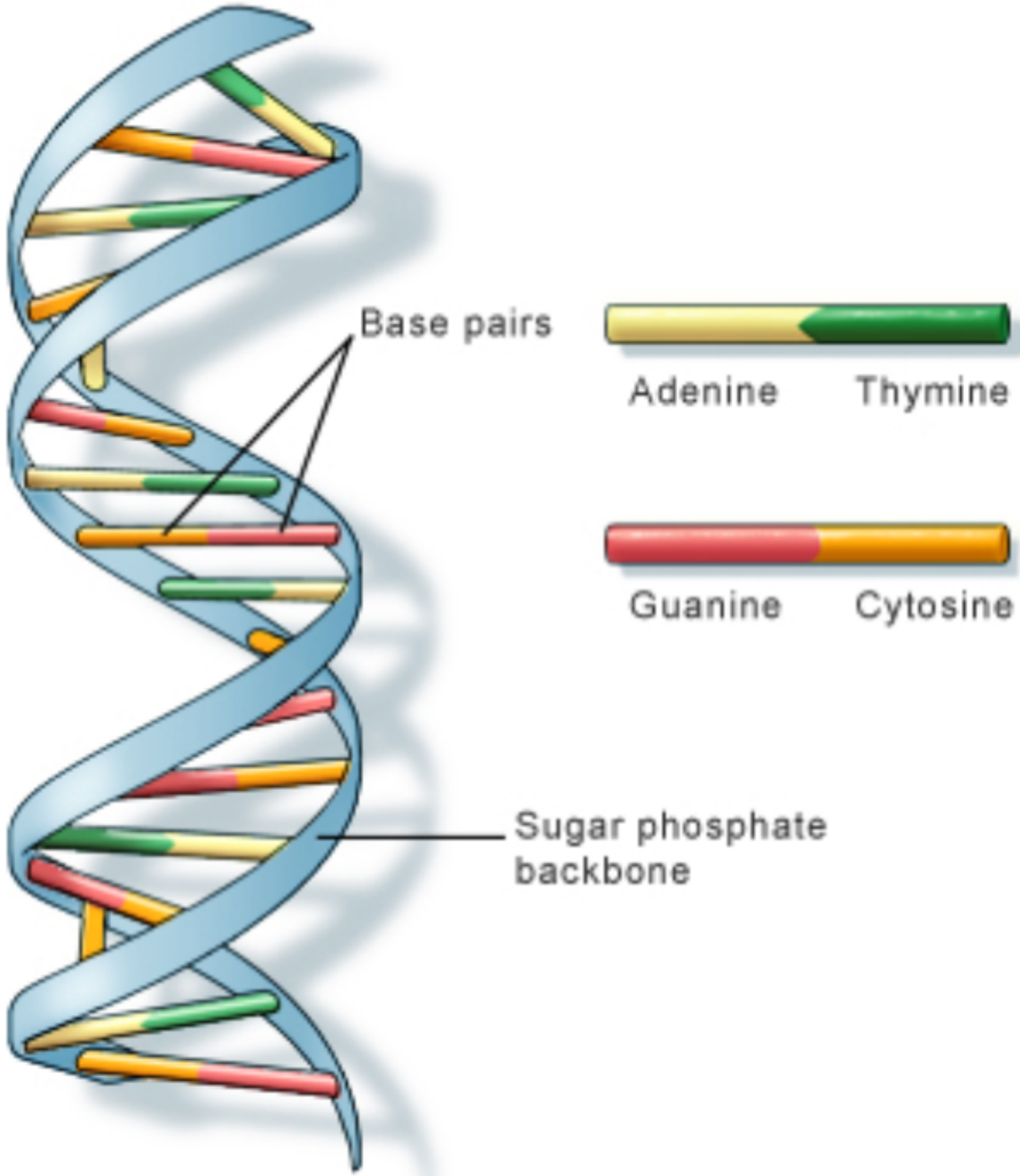
DISEASE
Errors or mutations in genetic switches can disrupt the network and lead to a range of diseases. The new findings will spur further research and may lead to new drugs and treatments.



DNA

DNA

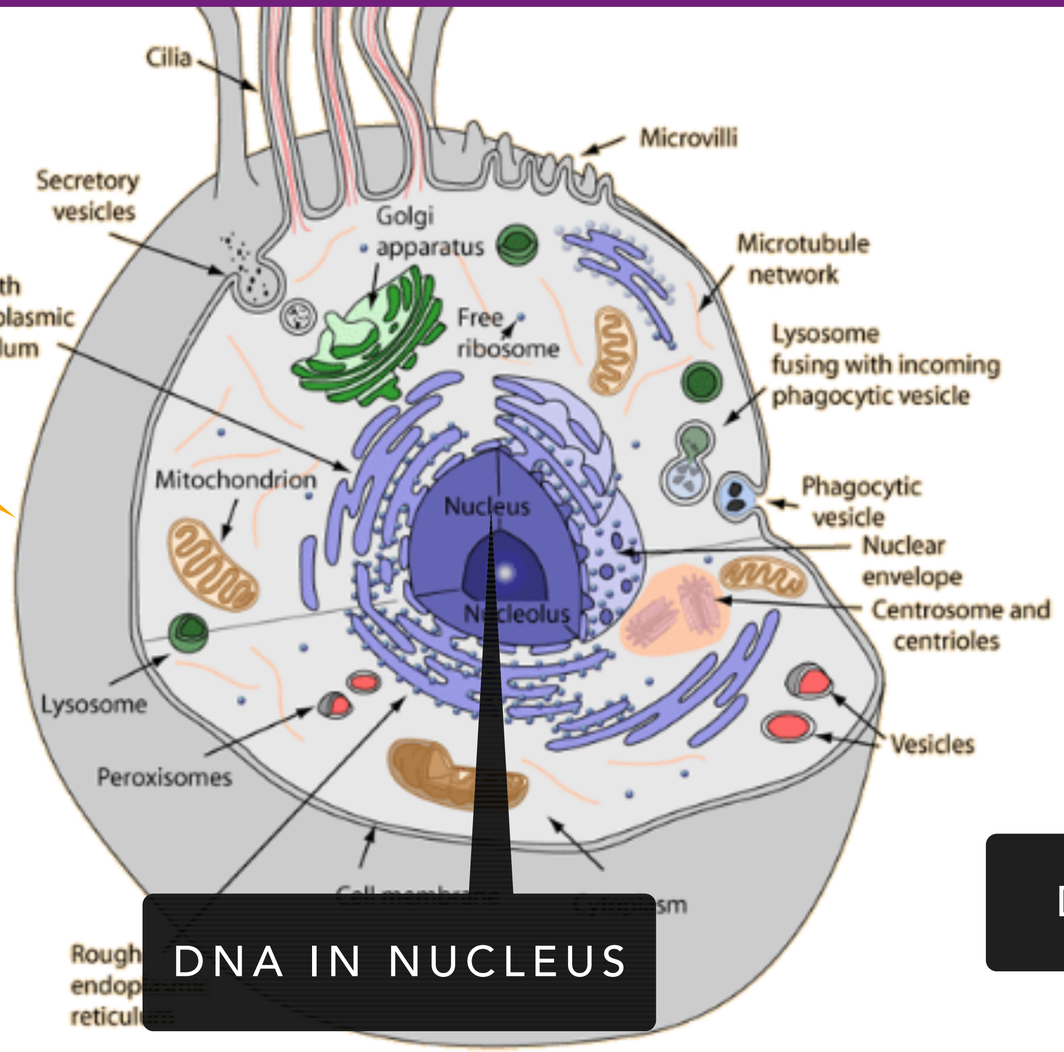
- DNA (**d**eoxyribo**n**ucleic **a**cid)
  - Hereditary material in organisms
- Nearly every cell in a person's body has the same DNA



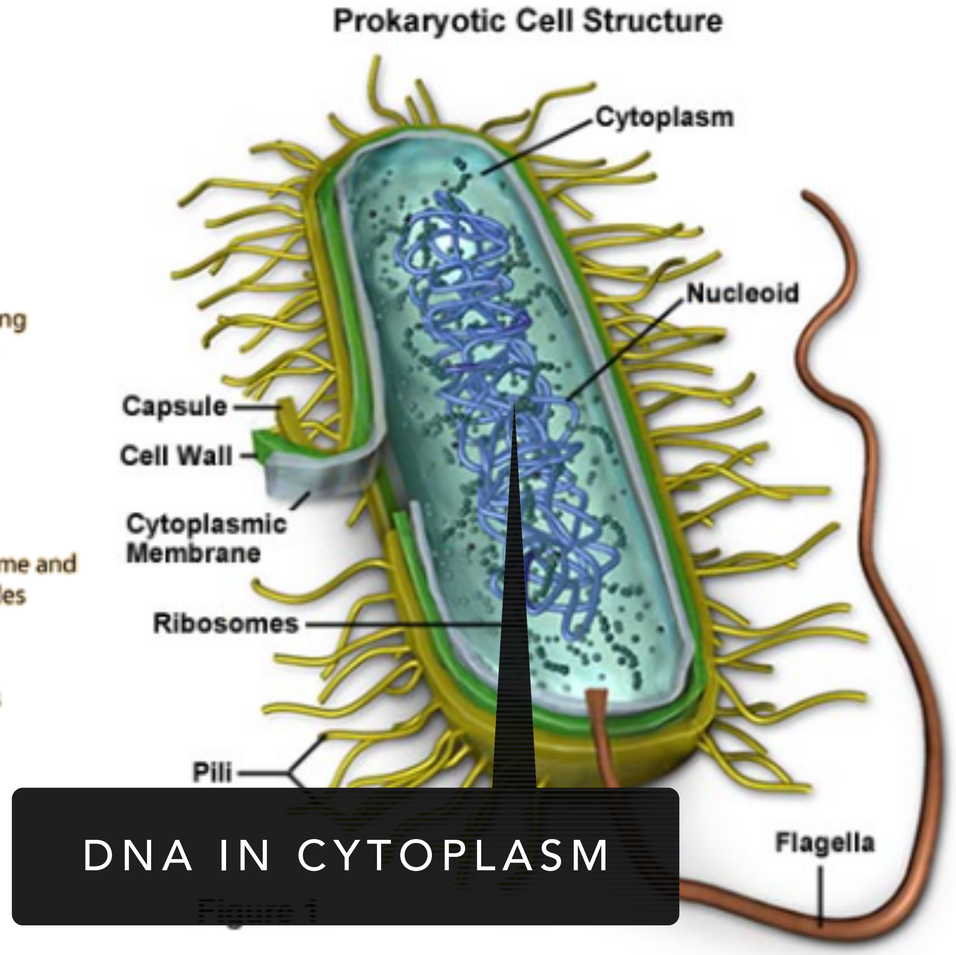
U.S. National Library of Medicine

DNA

DNA IN MITOCHONDRION



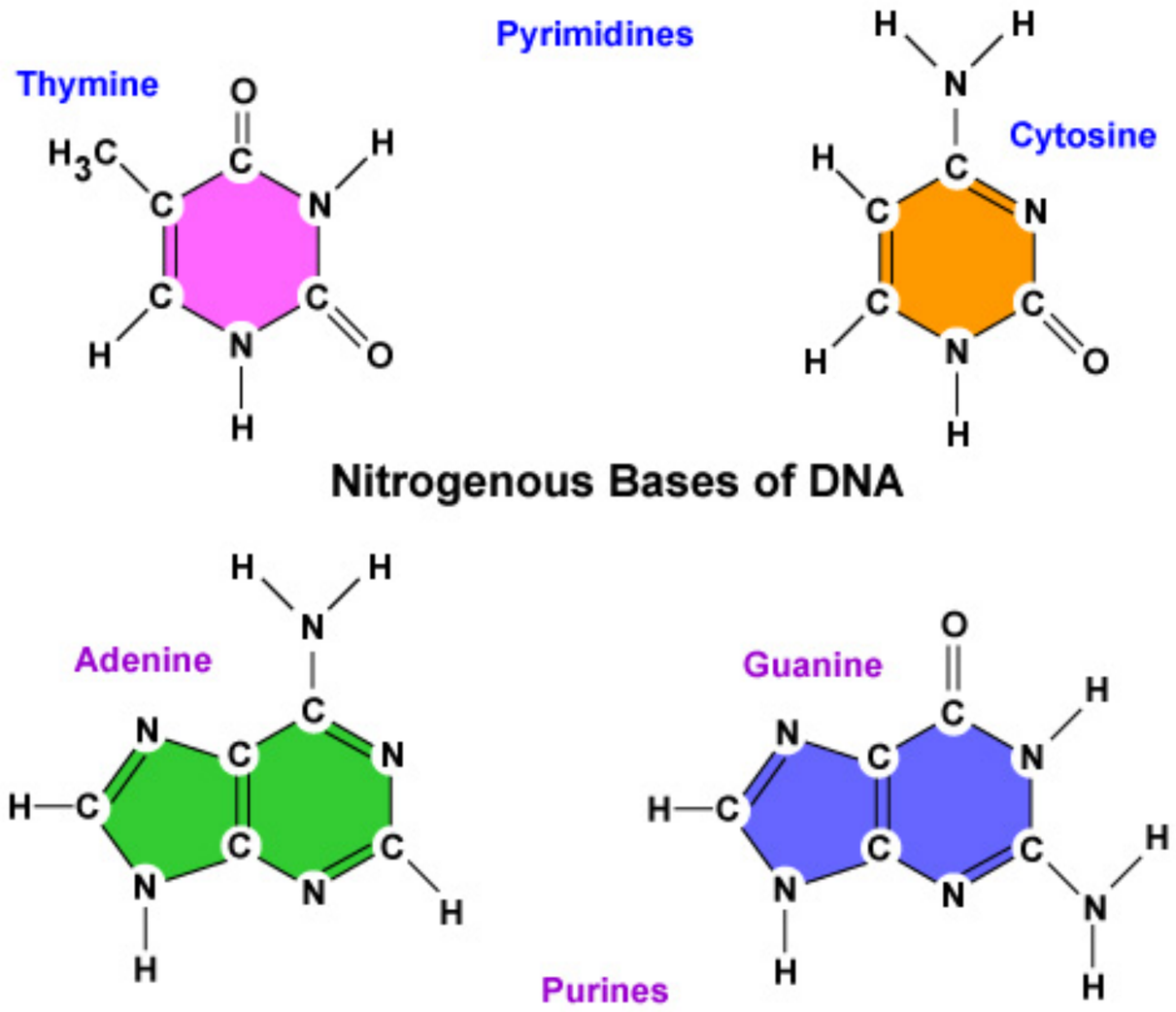
Animal cell structure



Prokaryotic cell structure

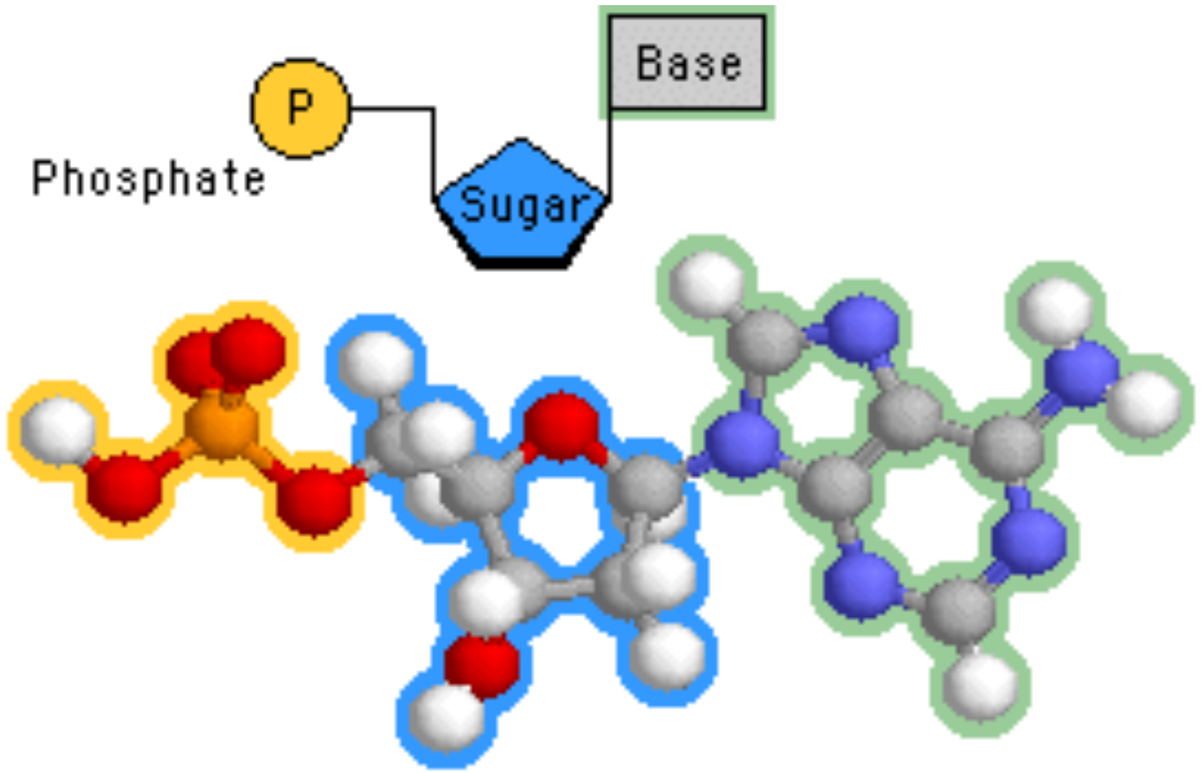
DNA

- The information in DNA is stored as a code made up of four chemical bases
  - adenine (A)
  - guanine (G)
  - cytosine (C)
  - thymine (T)

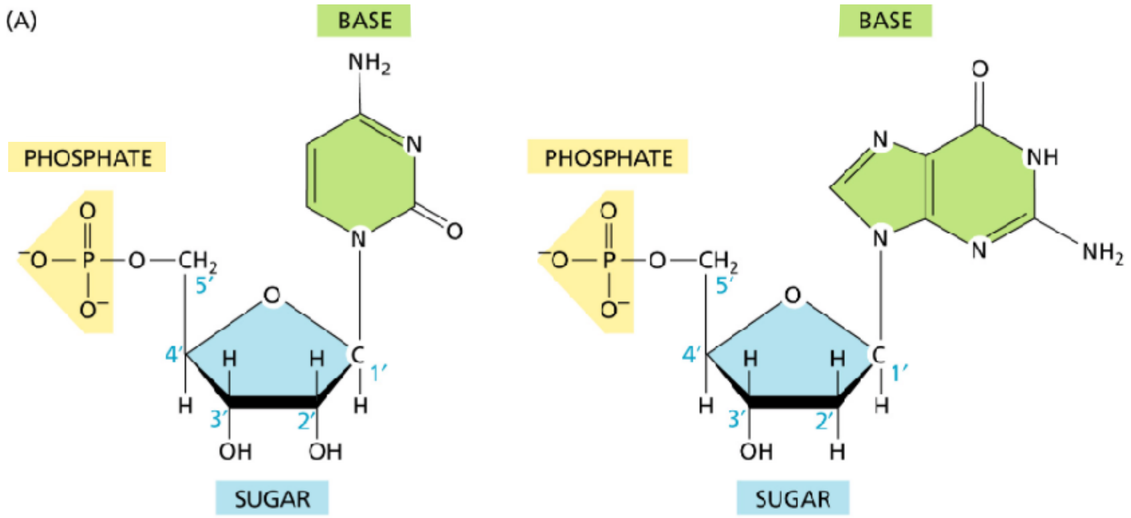


DNA

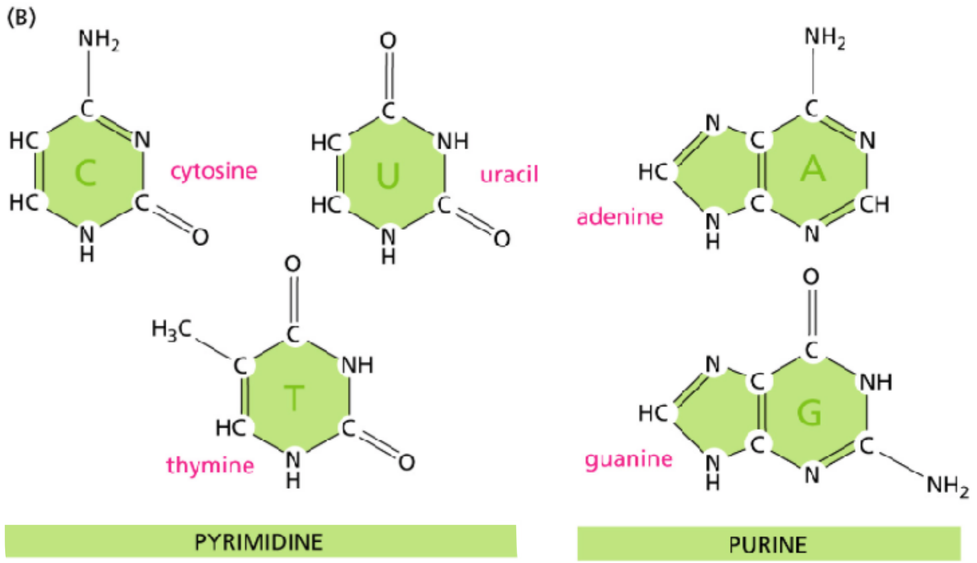
- Each nucleotide has three components
  - 1 or more phosphate groups
  - 5-carbon, or pentose, sugar
  - Nucleotide

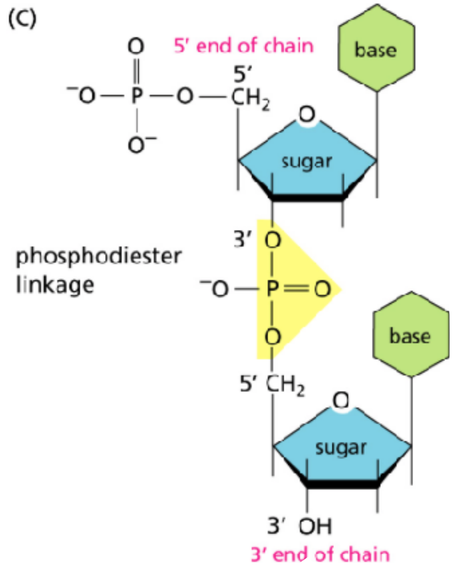


DNA



- Purine or pyrimidine bases
- 5-carbon (pentose) sugar
- 1 or more phosphate groups

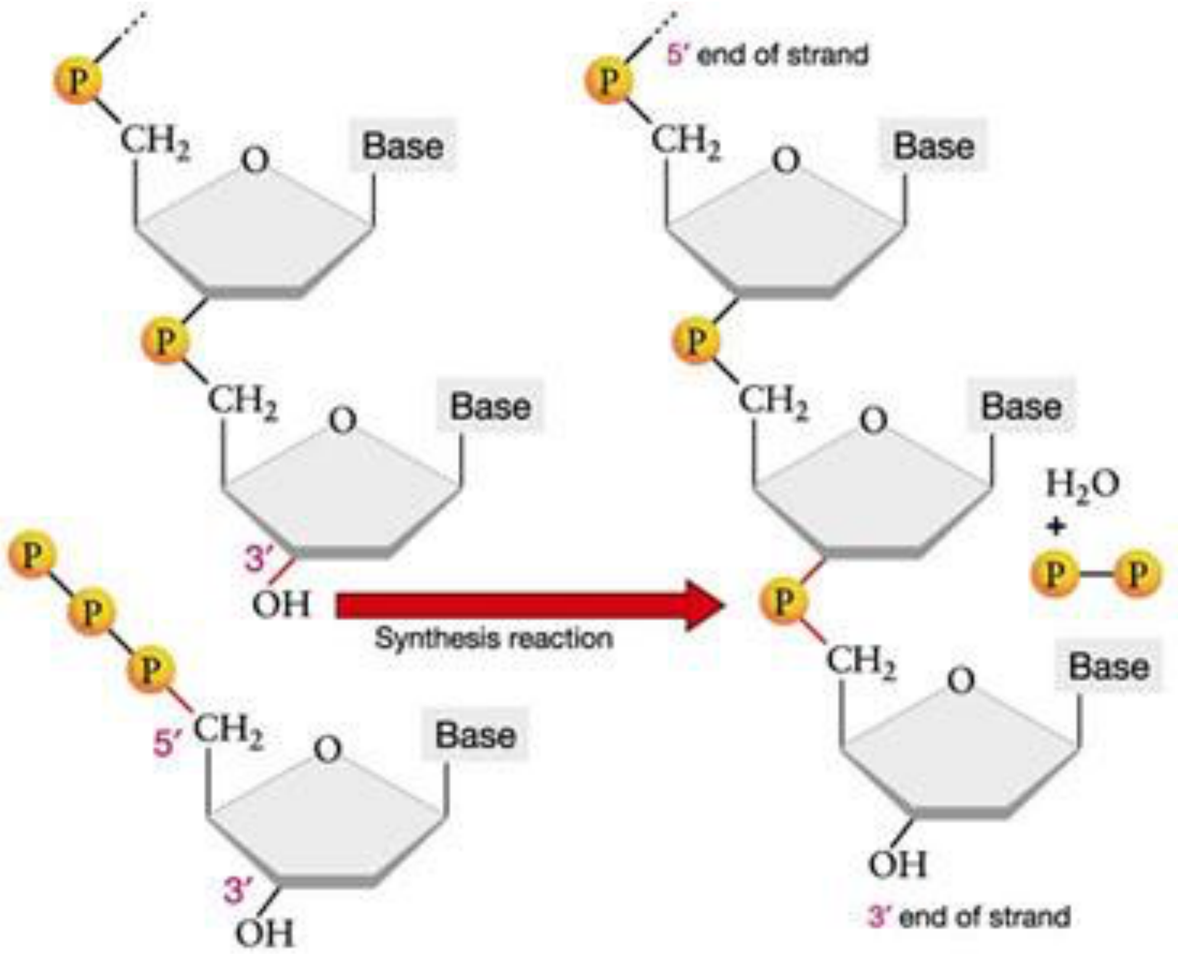




POLYMERIZATION INTO CHAIN

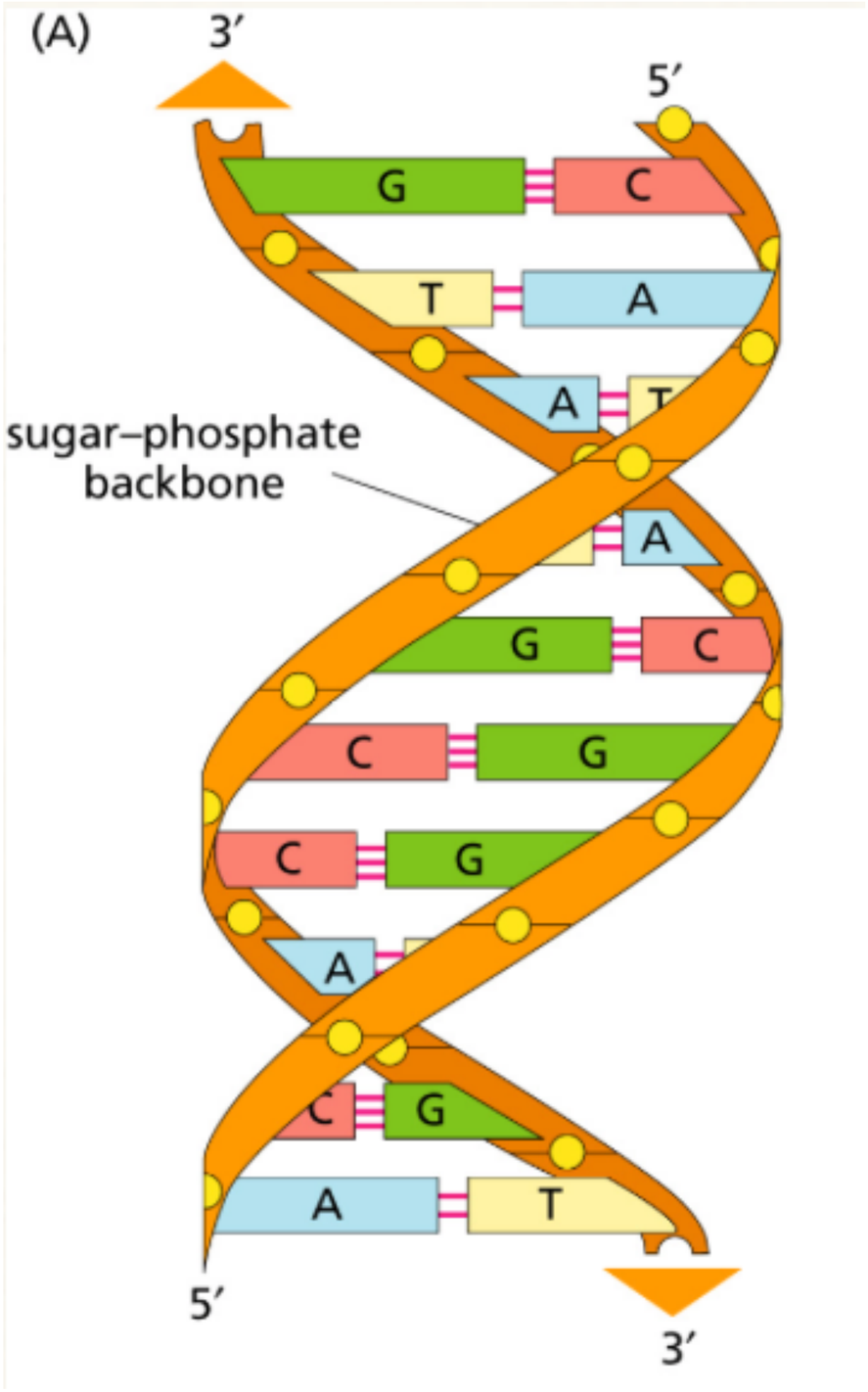
DNA

- Nucleic acids are built by polymerizing nucleotides
  - Millions of bases
  - Sequence encodes information



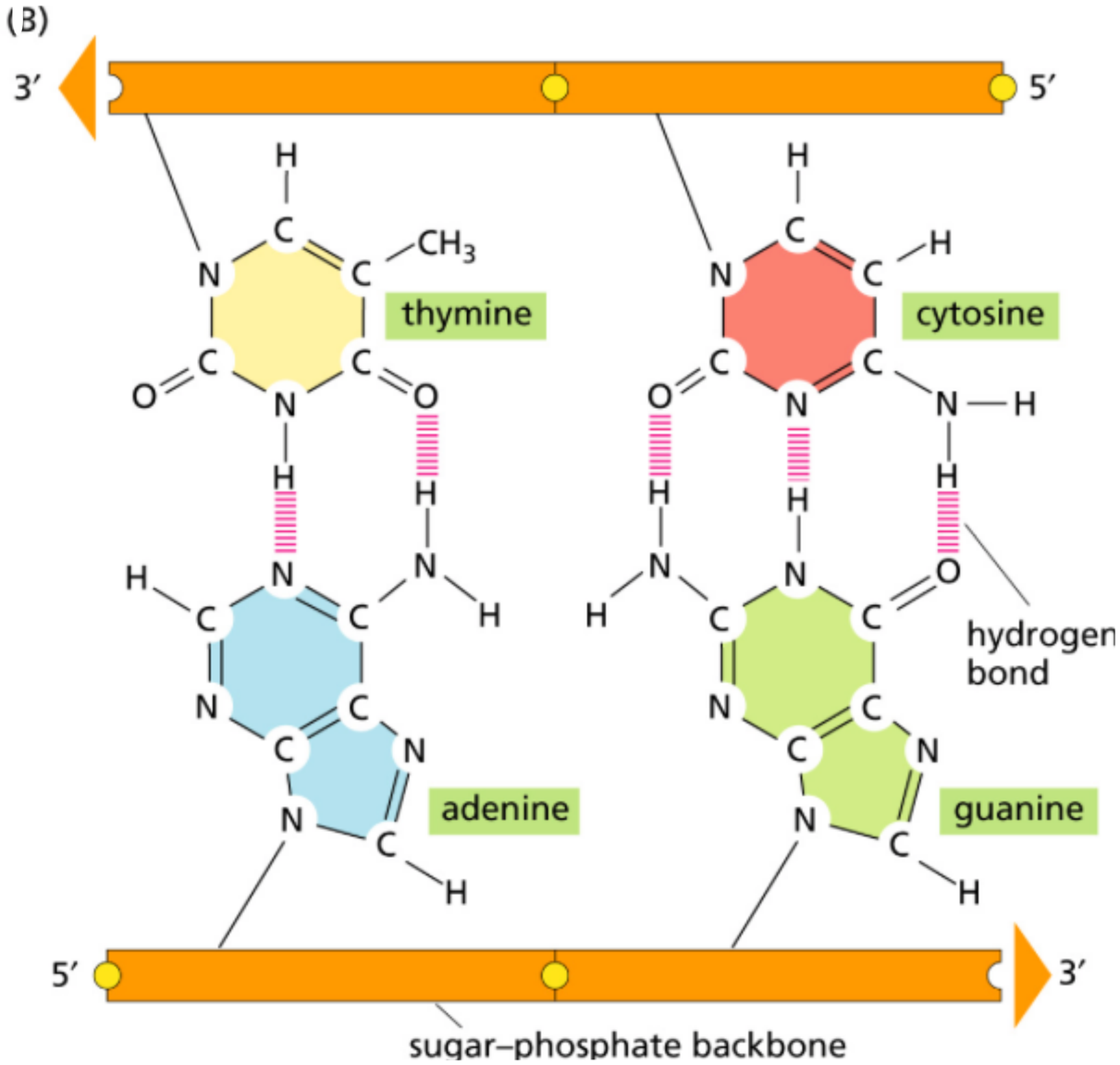
DNA

- DNA has a double helix structure
  - The backbone consists of alternating deoxyribose and phosphate groups
  - Each strand has a base sequence that is complementary to its partner strand



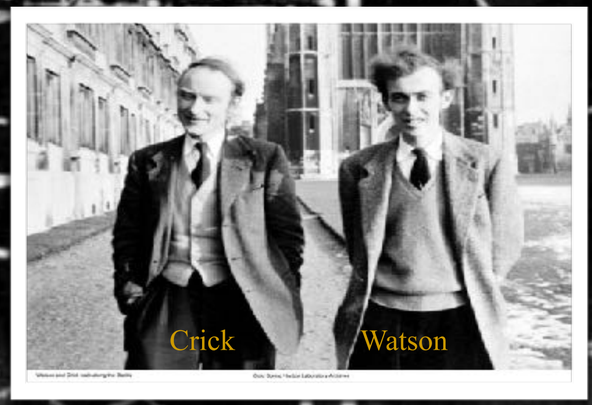
DNA

- Watson-Crick base-pairing
  - A will only base-pair with T
  - C will only base-pair with G
- Base-pairs
    - A and T contain two H-bonds
    - G and C contain three H-bonds (more stable that AT)



DISCOVERY OF DNA

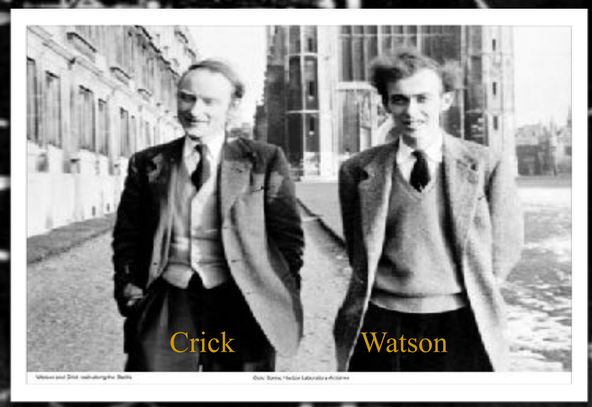
- DNA Sequences
  - Chargaff and Vischer, 1949
    - DNA consisting of A, T, G, C
  - Chargaff Rule (#A≈#T and #G≈#C)
    - A “strange but possibly meaningless” phenomenon.



RESOURCES ON COURSE WEBSITE UNDER SESSION 1

DISCOVERY OF DNA

- DNA Structure
  - Watson and Crick, Nature, April 25, 1953
  - Rich, 1973l Structural biologist at MIT
    - DNA's structure in atomic resolution

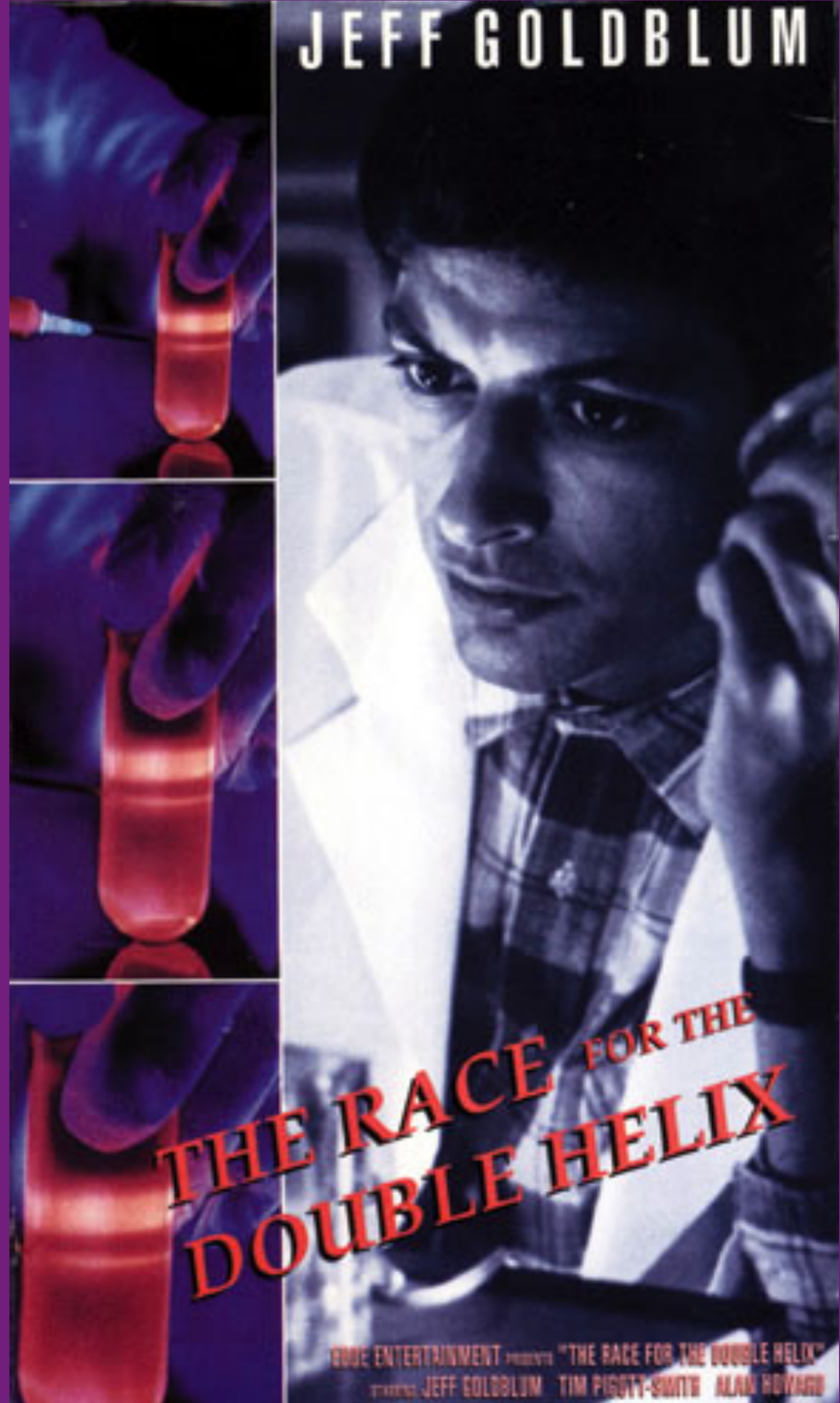


DNA

- The players
  - Watson: a zoologist
  - Crick: a physicist “In 1947 Crick knew no biology and practically no organic chemistry or crystallography”
  - Rosalind Franklin - Xray crystallographer
  - Watkins - Showed an X-ray to Watson
- The Discovery
  - Applying Chagraff’s rules and the X-ray image from Rosalind Franklin, they constructed a “tinkertoy” model showing the double helix
- 1962 Nobel Prize for Watson, Crick, Wilkins
  - Rosalind Franklin died in 1958

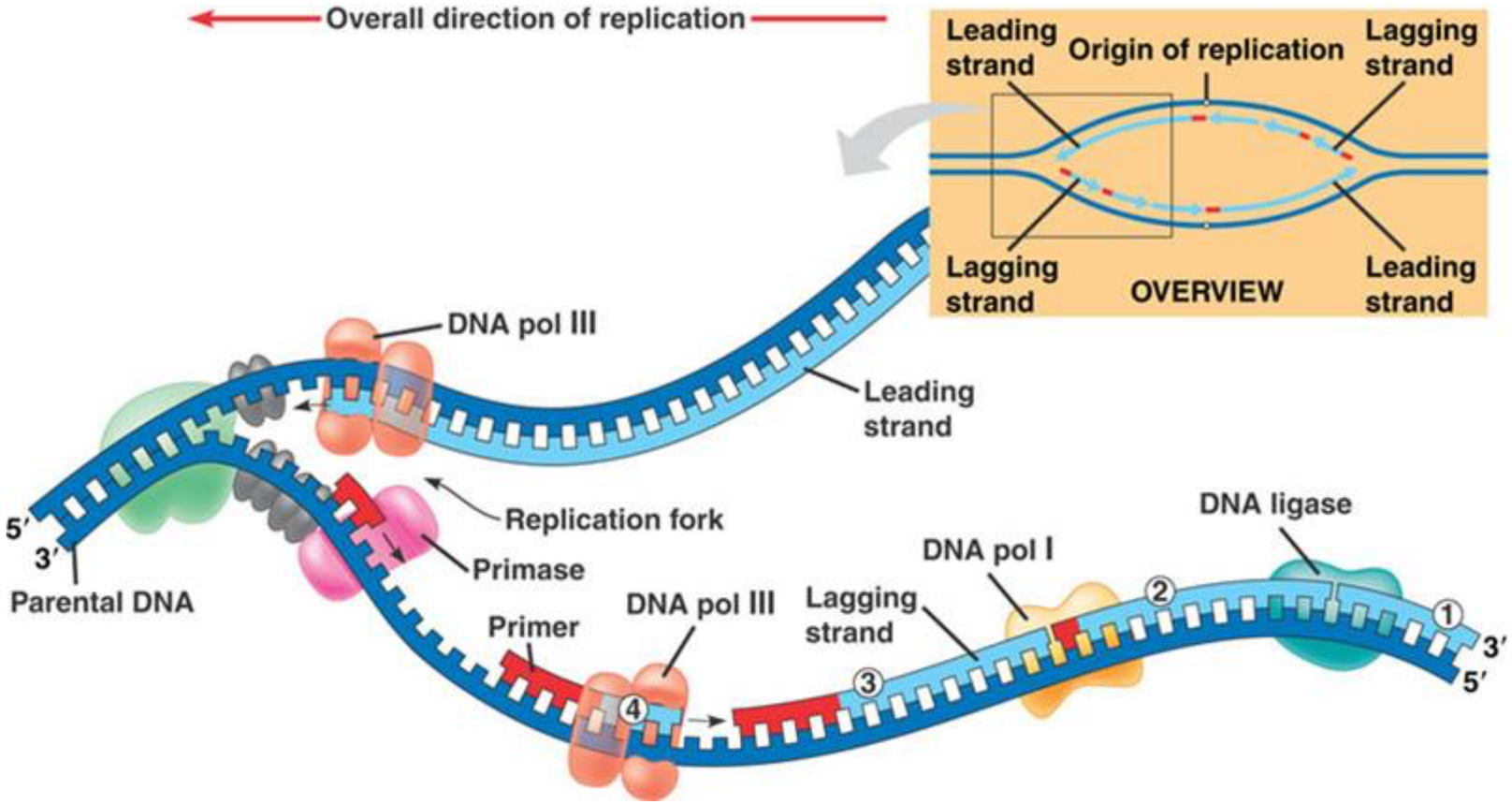
DNA

- Life Story: The Race for the Double Helix



DNA

- DNA Replication
  - Process of producing two identical replicas from one original DNA molecule

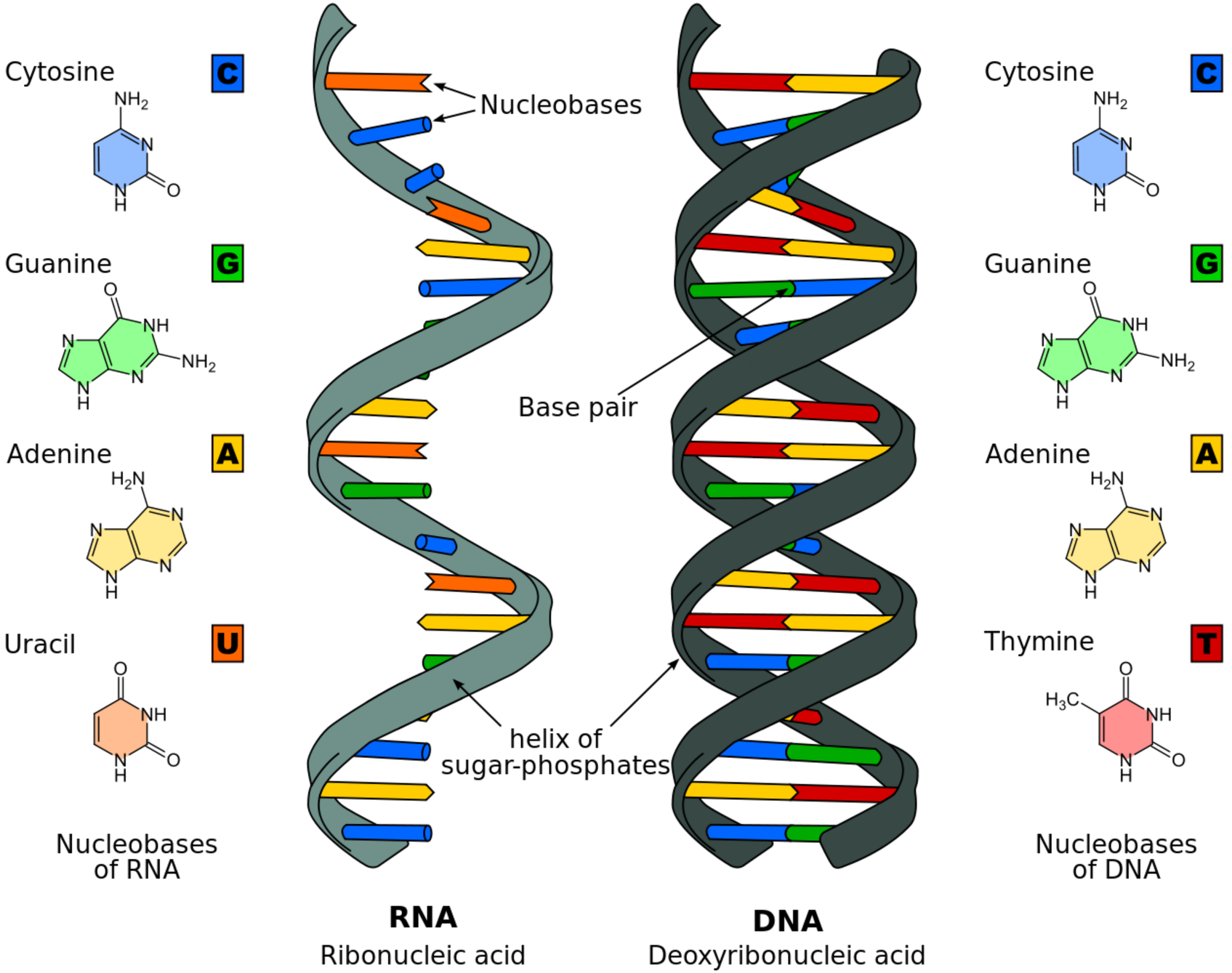


RESOURCES ON COURSE WEBSITE UNDER SESSION 1

RNA

RNA

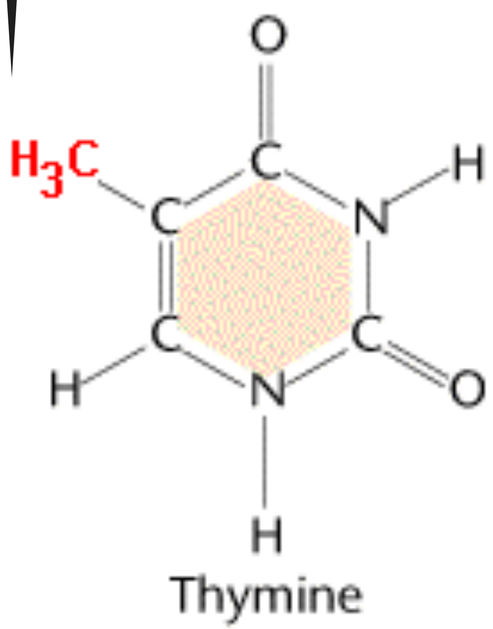
- Ribonucleic acid
- RNA is chemically similar to DNA chemically
  - Different base
  - Different sugar



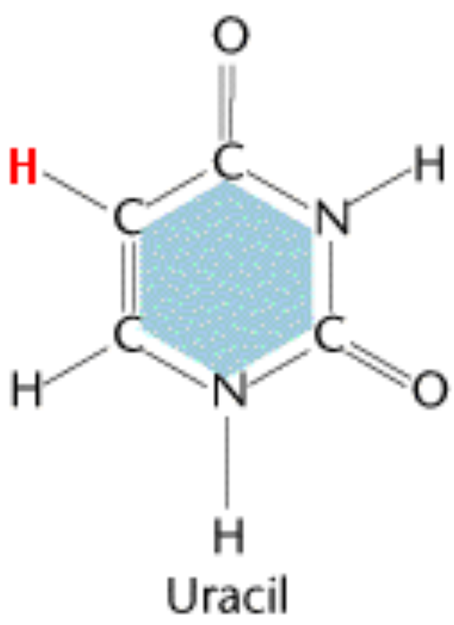
RNA

- T(hyamine) is replaced by U(racil)
- Methylation of DNA
  - More structural stability
  - More efficient for replication

METHYL GROUP



Thymine

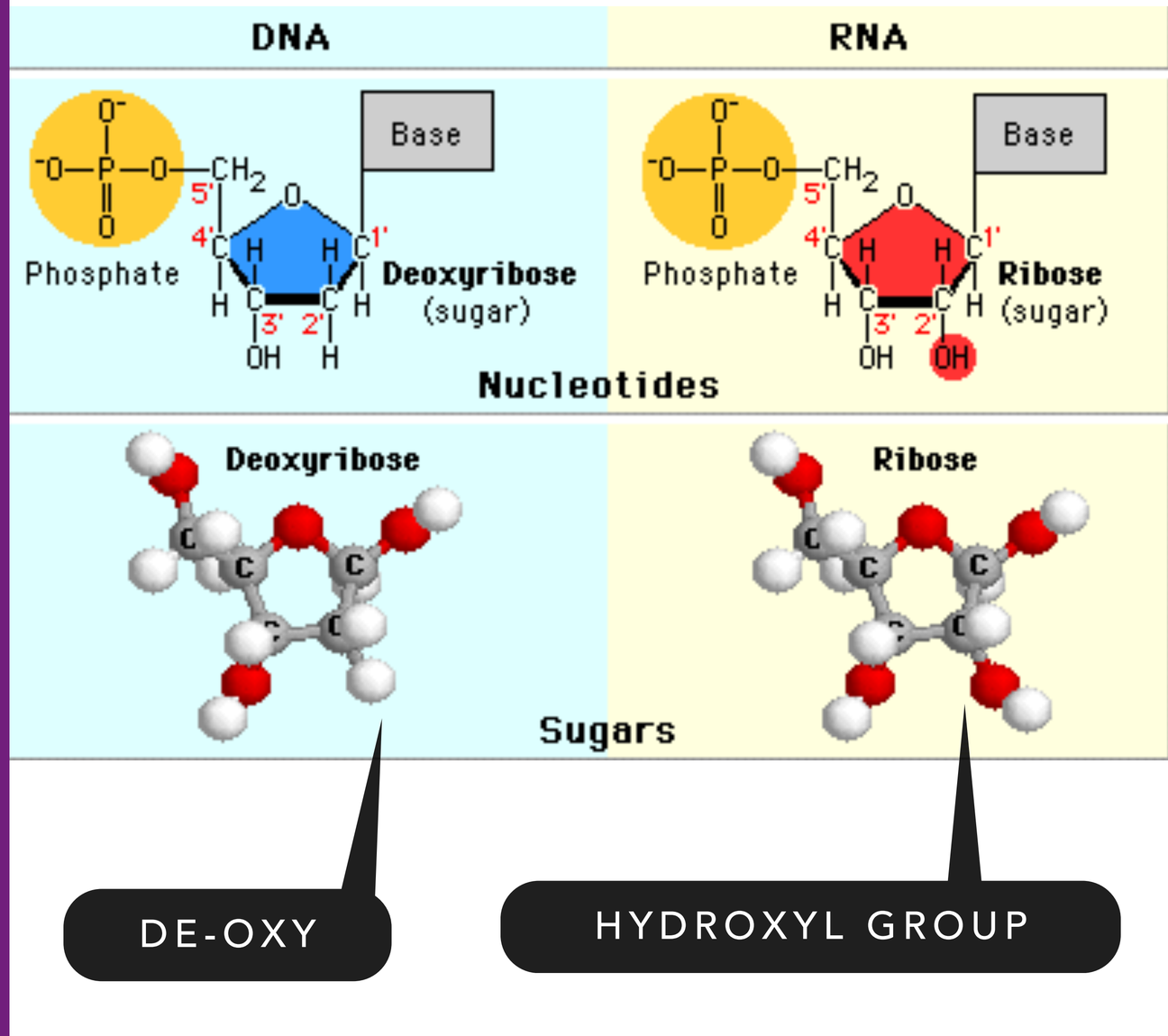


Uracil

(Klug & Cummings 1997)

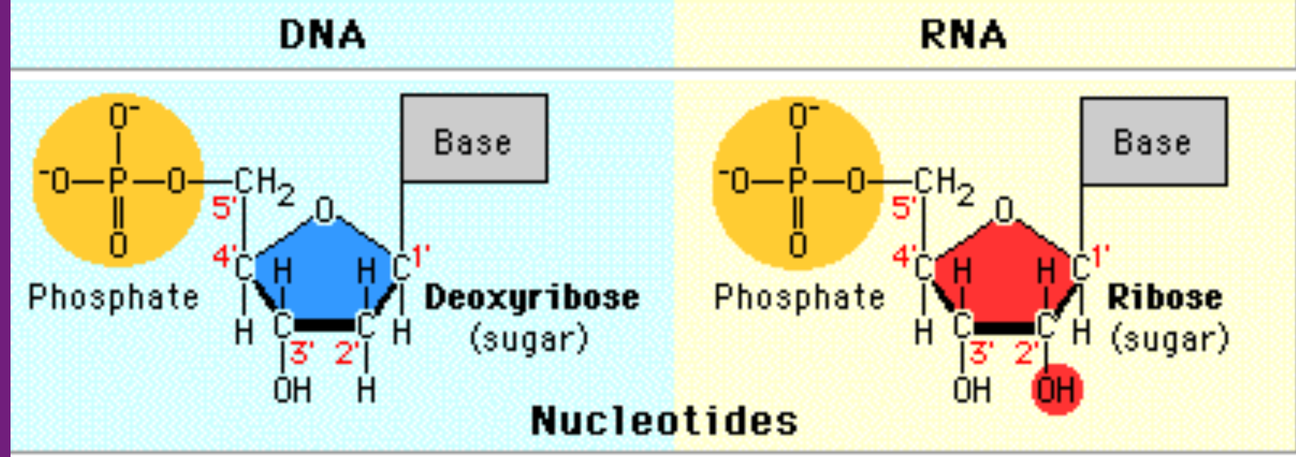
RNA

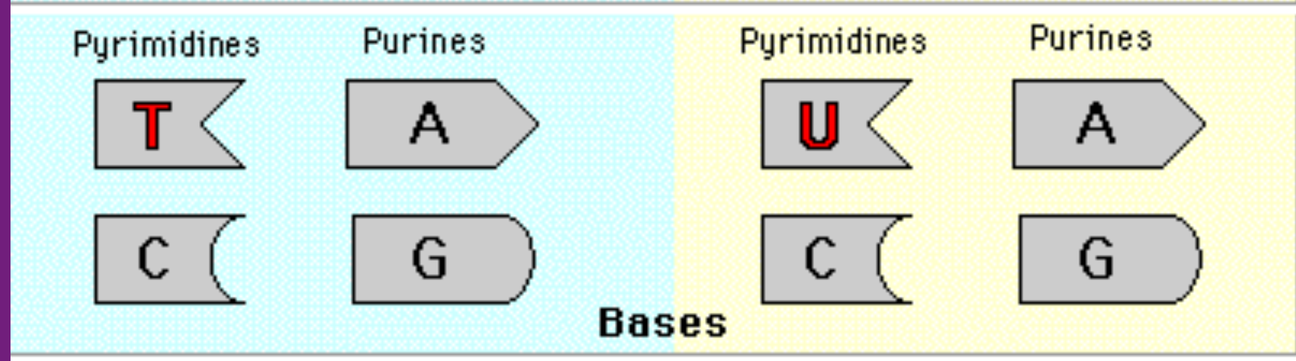
- RNA sugar is ribose
- DNA sugar is Deoxyribose (minus one oxygen atom)
  - Important for the enzymes that recognize DNA and RNA inside organisms

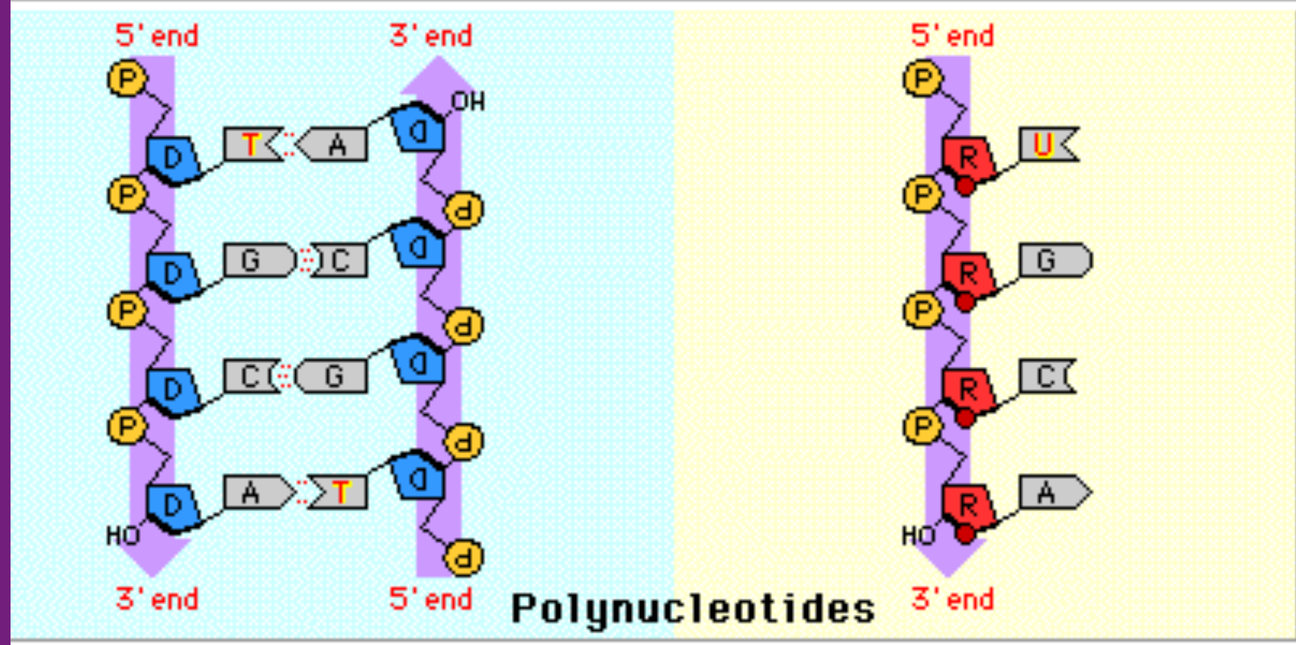


RNA

- Properties of RNA
  - More structurally flexible than DNA
  - Less stable than DNA
  - More reactive (OH group)
  - Almost all RNA is single stranded

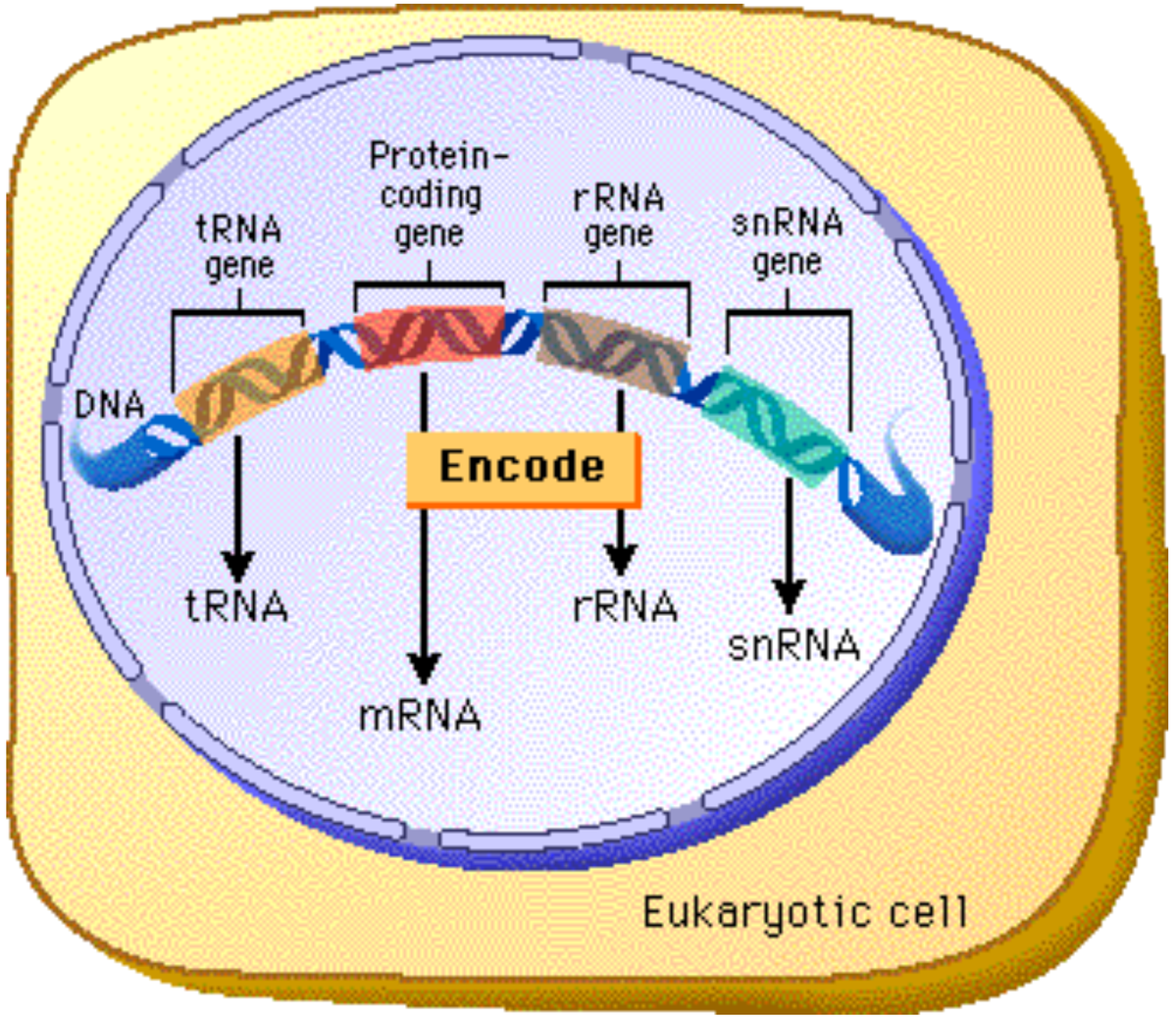






RNA

- 4 types of RNA, each encoded by a different gene
  - **mRNA** - Messenger RNA: Encodes amino acid sequence of a polypeptide
  - **tRNA** - Transfer RNA: Brings amino acids to ribosomes during translation
  - **rRNA** - Ribosomal RNA: With ribosomal proteins, makes up the ribosomes
  - **snRNA** - Small nuclear RNA; combine with proteins for post-transcriptional processing (in eukaryotes only)

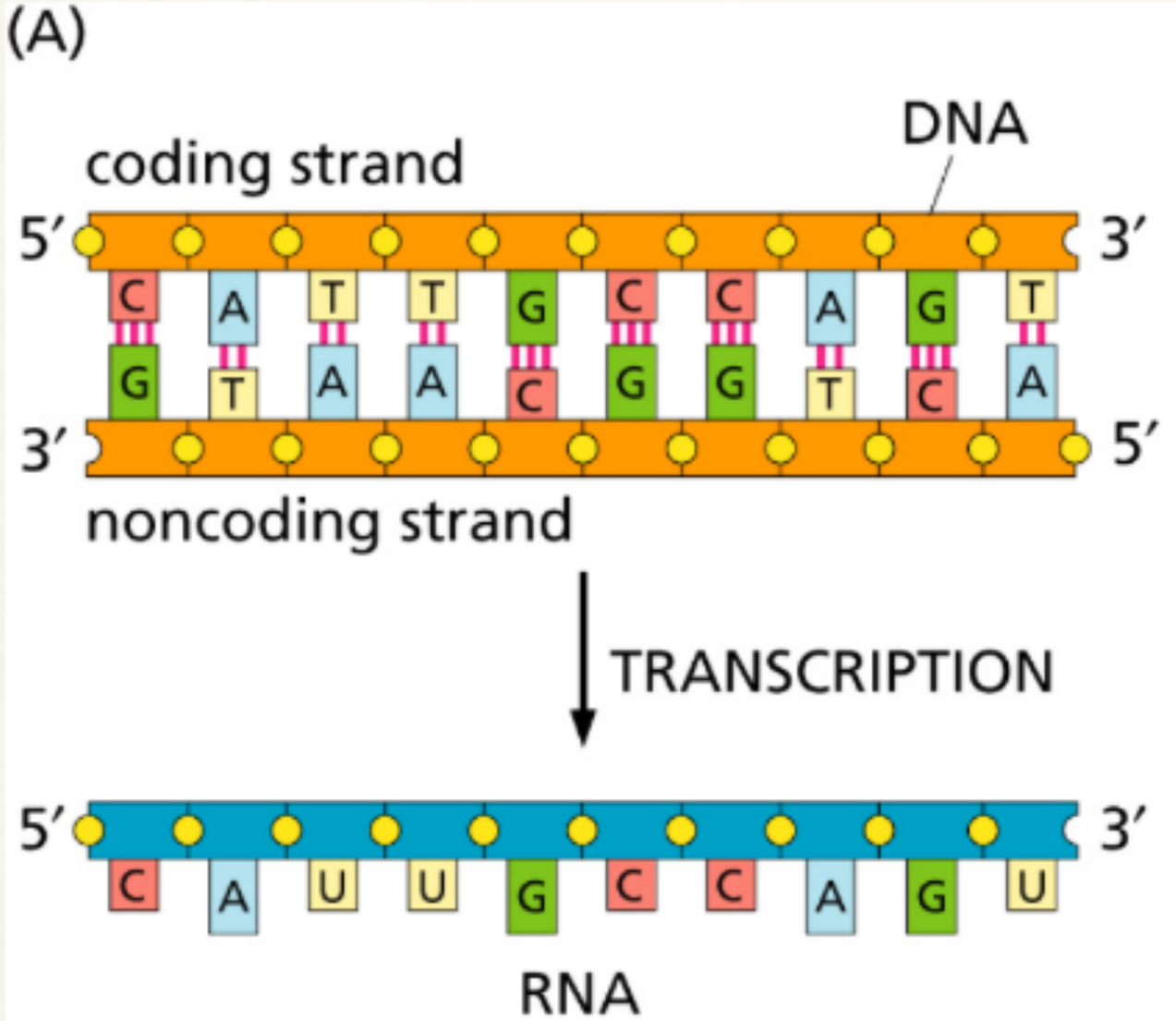


Transcription

TRANSCRIPTION

- The synthesis of an RNA copy of a segment of DNA
  - When a gene is being transcribed into RNA, the gene is said to be <u>expressed</u>

GENE IS TRANSCRIBED FROM DNA

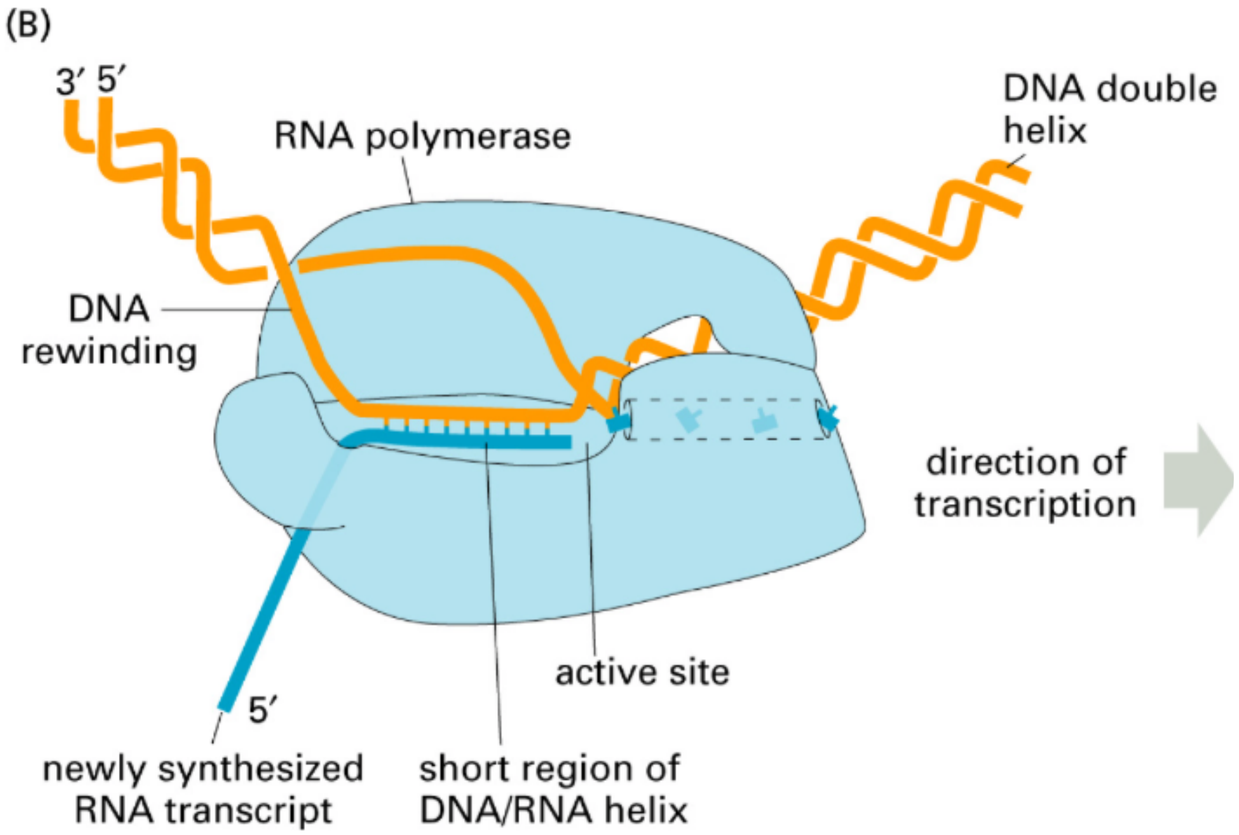


TRANSCRIPTION

- Basic: https://www.youtube.com/watch?v=5MfSYnItYvg
- Advanced: https://www.youtube.com/watch?v=SMtWvDbfHLo

TRANSCRIPTION

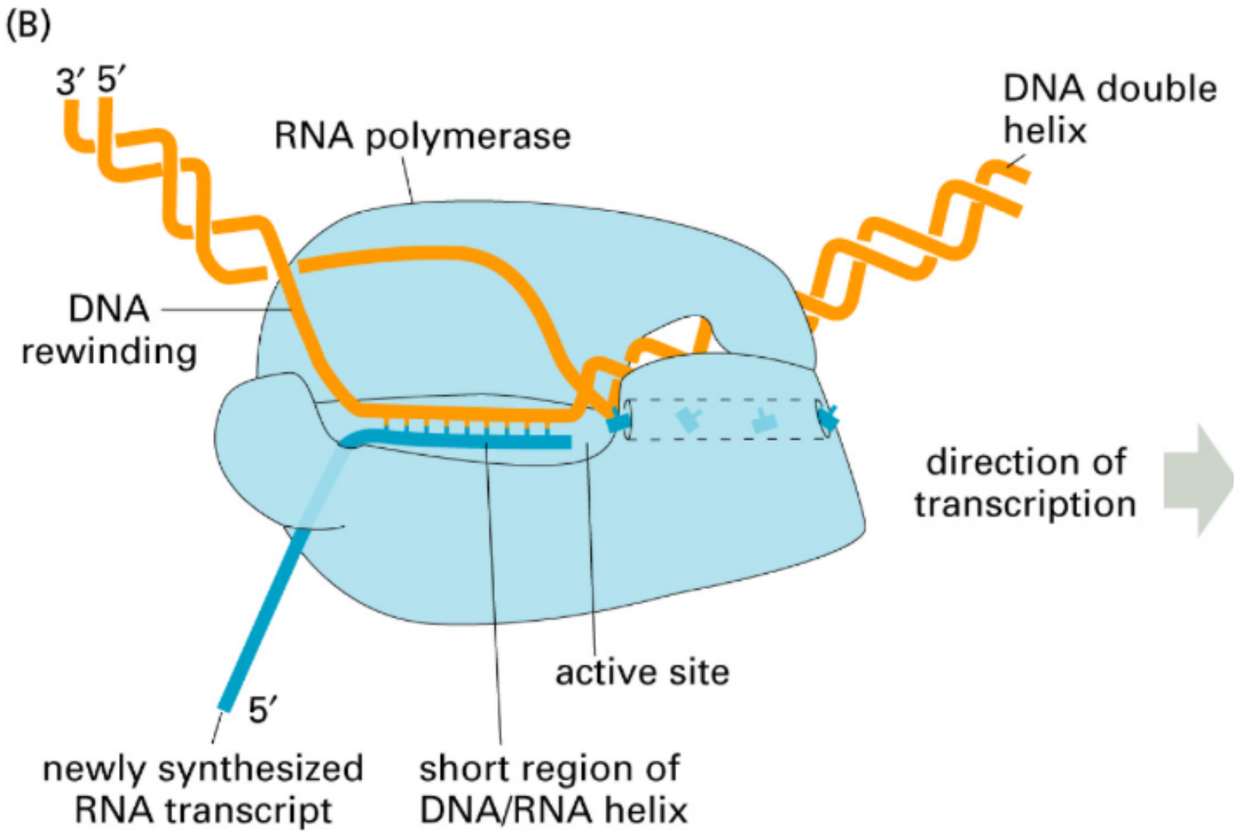
- RNA synthesis
  - Separation of the DNA strands
  - Use one DNA strand as a template
  - Synthesis of RNA in the 5' to 3' direction by **RNA polymerase**



http://www.cs.rice.edu/~nakhleh/COMP572/Slides/MolCellBiolReview.pdf

TRANSCRIPTION

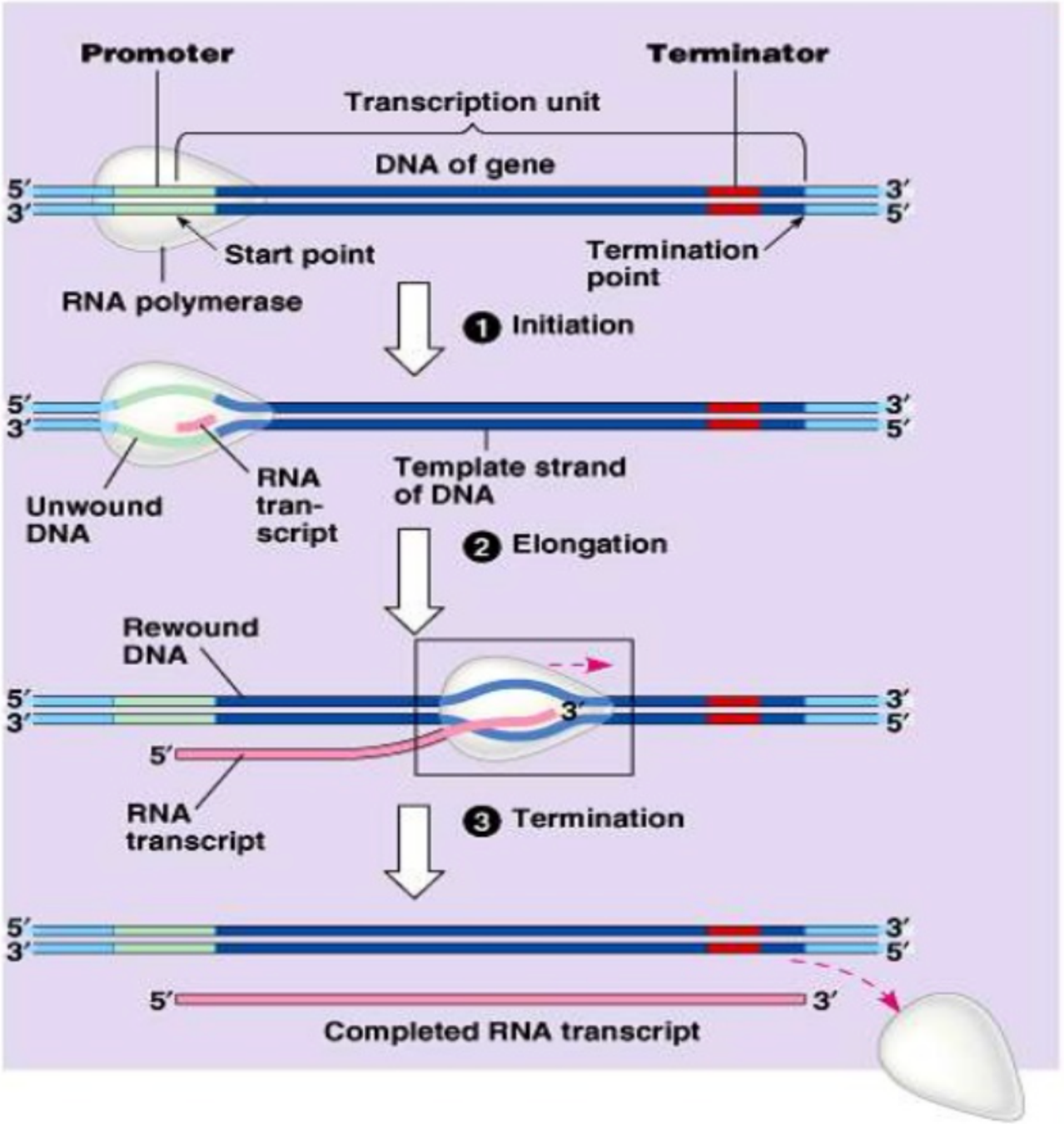
- RNA Polymerase
  - Multi-subunit enzyme
  - Reads the DNA and recruits the correct building blocks of RNA to string them together based on the DNA code



http://www.cs.rice.edu/~nakhleh/COMP572/Slides/MolCellBiolReview.pdf

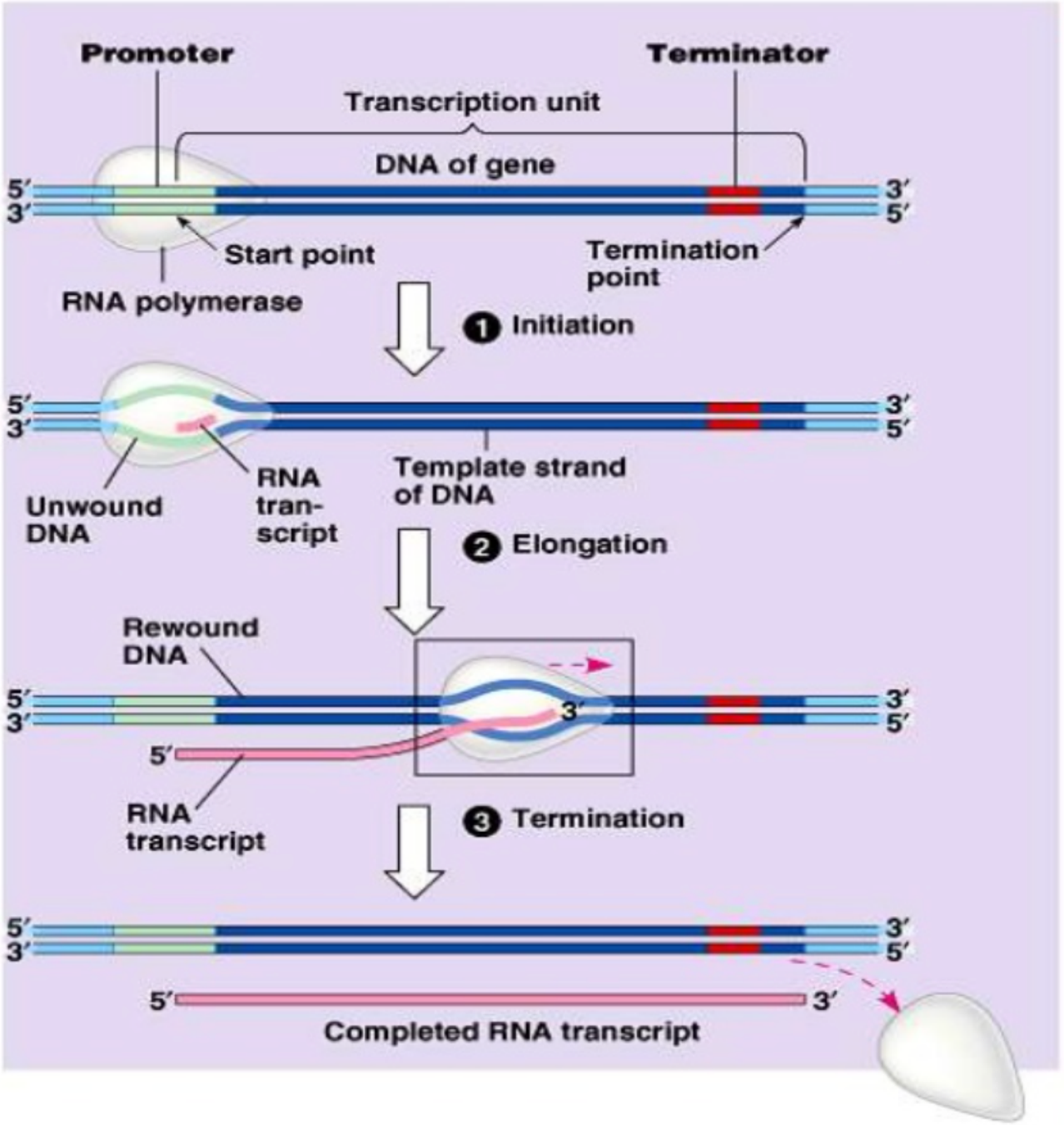
TRANSCRIPTION

- Initiation Steps
  - Transcription factors
    - Proteins bind together on the promoter region to form a transcription initiation complex
  - Promoter
    - Special sequence of nucleotides indicating the starting point for RNA synthesis
  - RNA polymerase
    - Binds the transcription initiation complex

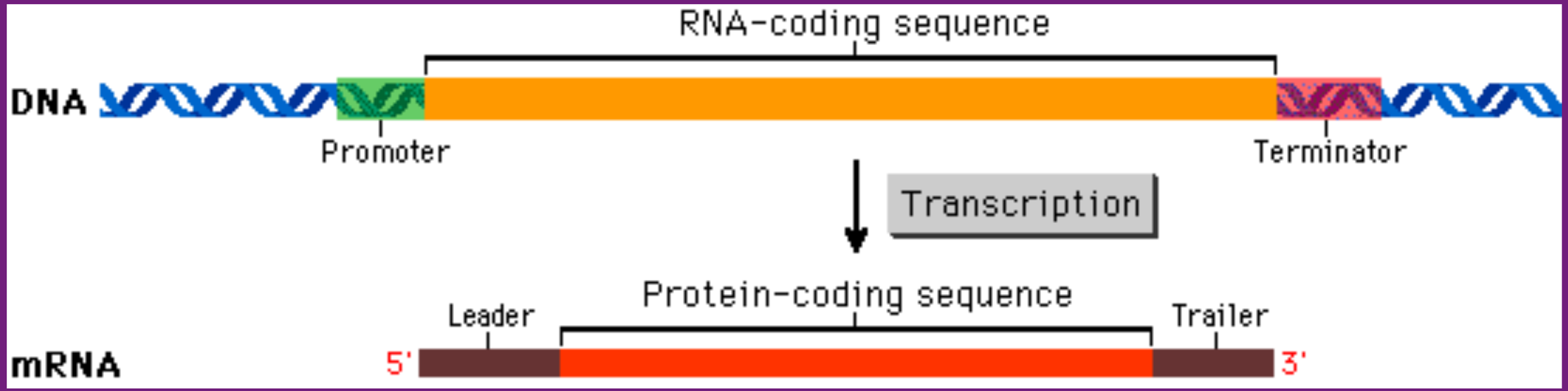


TRANSCRIPTION

- Elongation
  - RNA polymerase starts moving along the template strand from the 5′ to 3′ adding free RNA to make mRNA
- Termination
  - RNA polymerase stops at a termination codon
  - Terminator
    - Signal in DNA that halts transcription

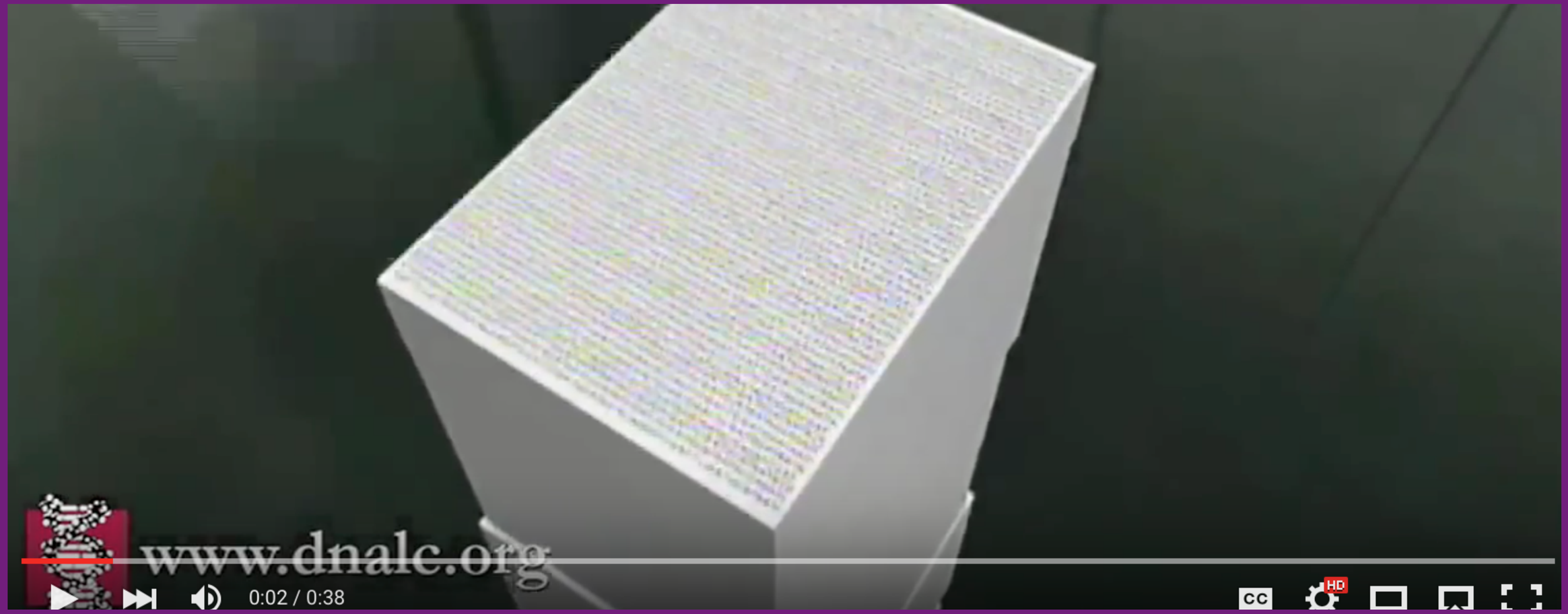


TRANSCRIPTION



- mRNA in prokaryotes
  - Colinear with the translated mRNA
  - The transcript of the gene is the molecule that is translated directly to polypeptide

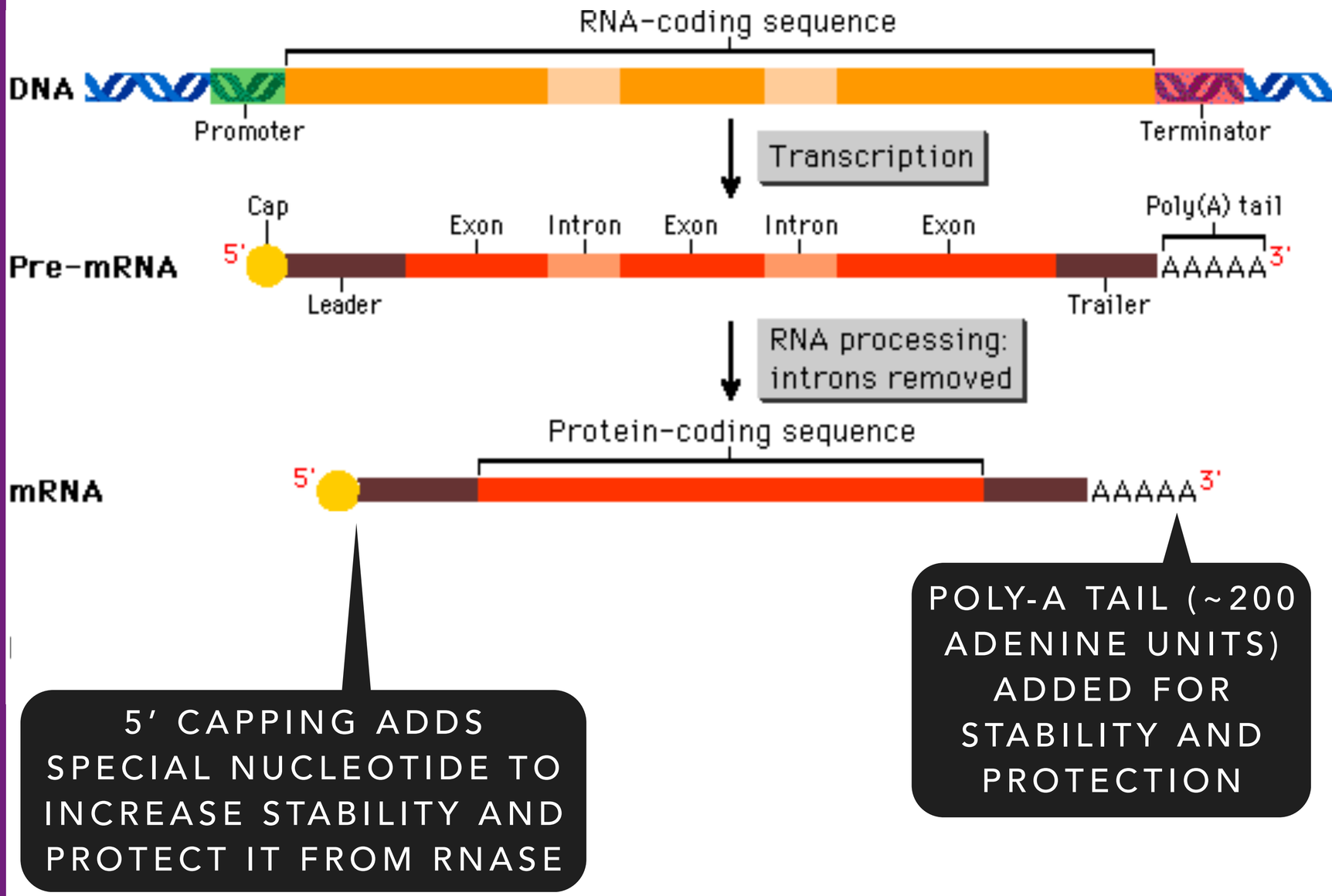
TRANSCRIPTION



- https://www.youtube.com/watch?v=hV6NSHjTR1s

TRANSCRIPTION

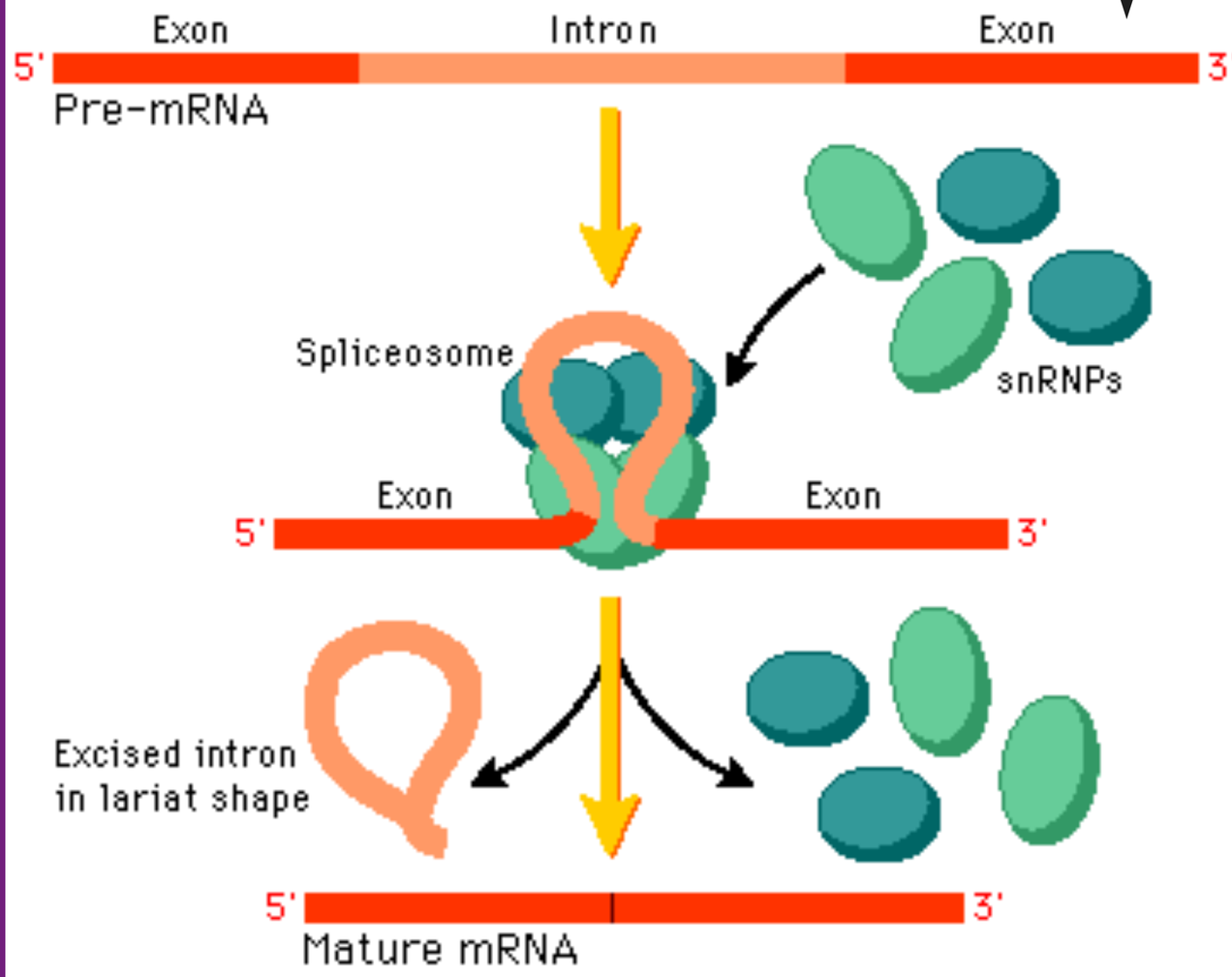
- mRNA in eukaryotes
  - Post-transcriptional modification
  - Spliceosome processes gene to remove extra sequences (introns)
  - Exits nucleus to cytoplasm



TRANSCRIPTION

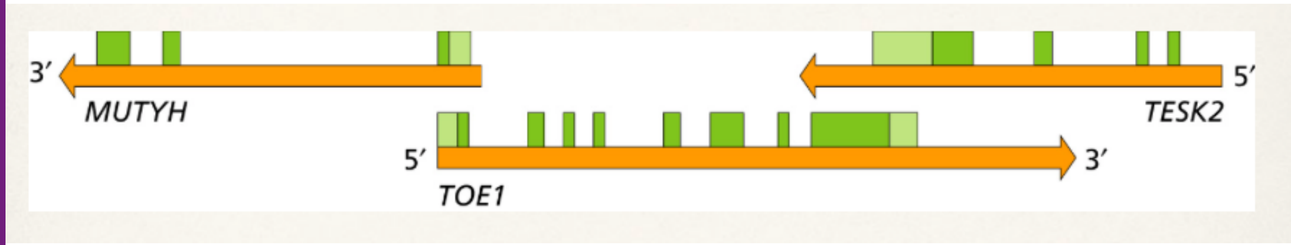
- Pre-mRNA Processing (splicing)
  - The intron loops out as snRNPs bind to form the spliceosome
  - Intron is excised
  - Exons are spliced together
  - Resulting mature mRNA exits the nucleus

SNRNP - SMALL NUCLEAR RIBONUCLEOPROTEIN PARTICLES; COMPLEXES OF SNRNAS AND PROTEINS; FORM THE SPLICEOSOME



TRANSCRIPTION

- Only one segment of DNA is transcribed for any given gene
  - Genes can overlap so that one (or both) strands encode different parts of proteins
  - Efficient for small genomes
- At any given time a cell only expresses a fraction of the genes in its genome
  - Regulated gene expression



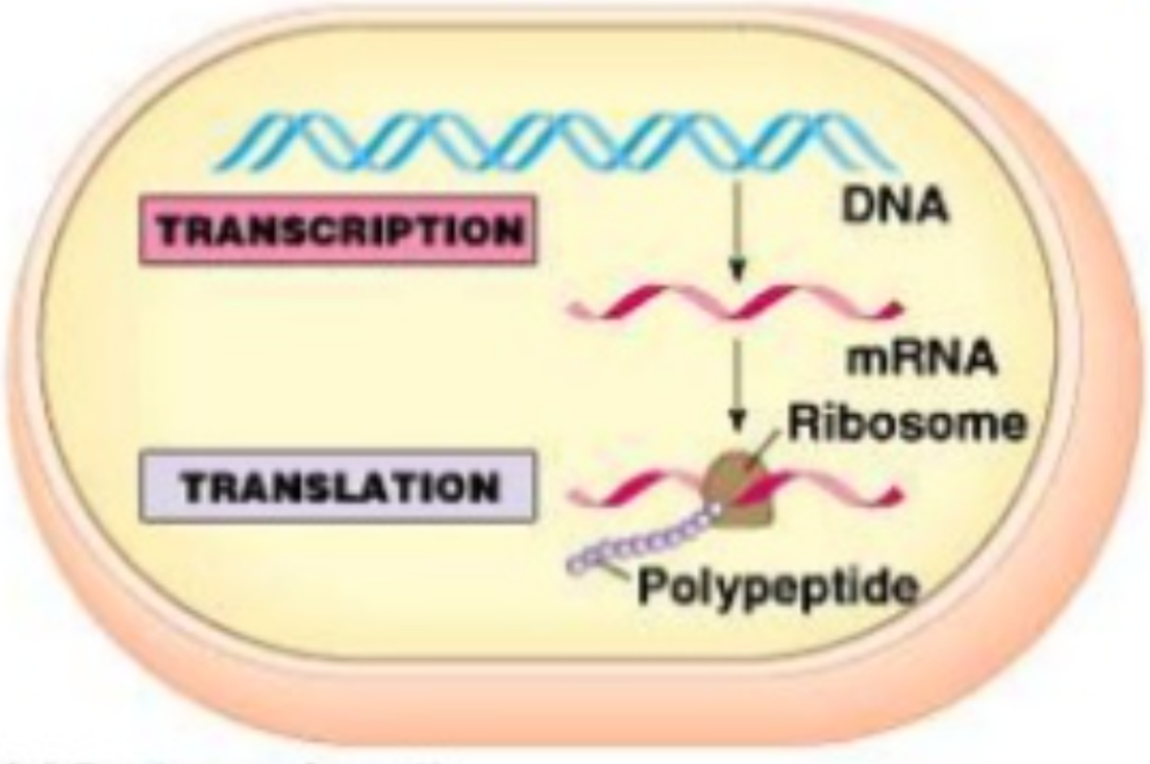
Translation

TRANSLATION

- https://www.youtube.com/watch?v=8dsTvBaUMvw Basic
- https://www.youtube.com/watch?v=TfYf_rPWUdY Advanced

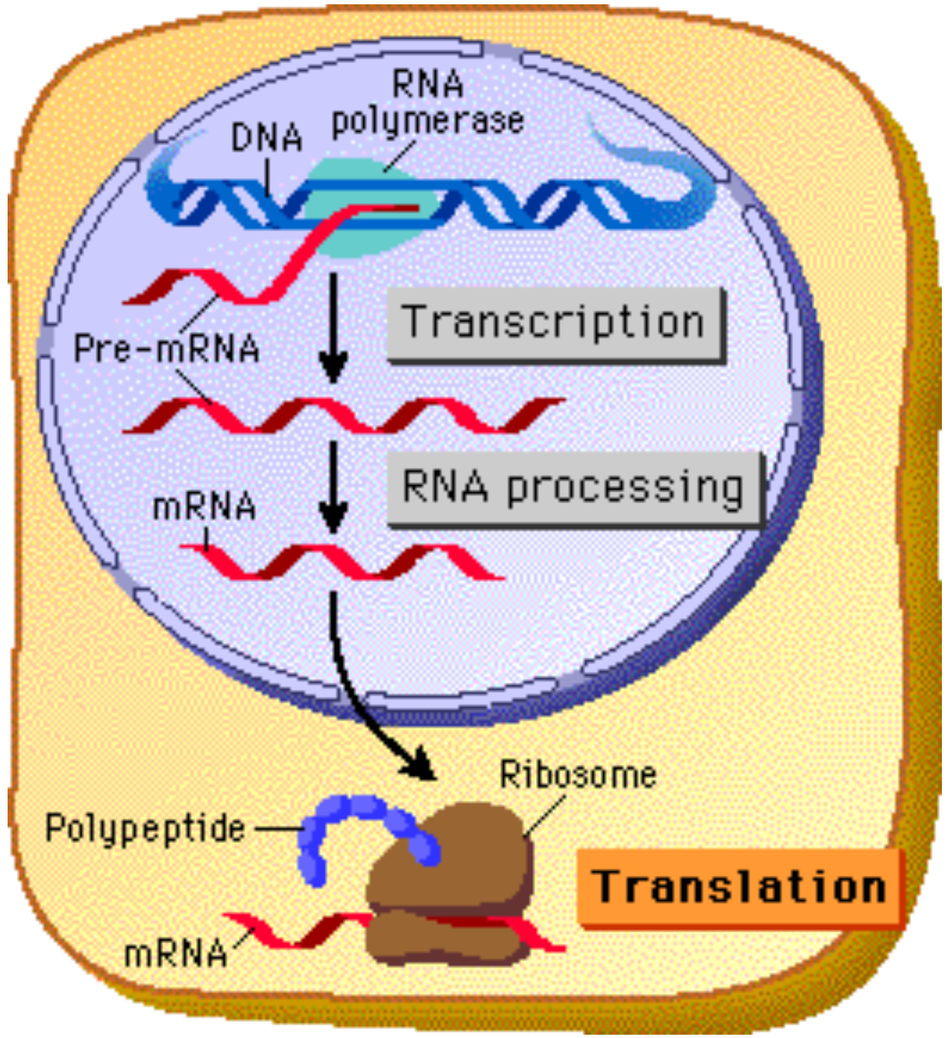
TRANSLATION

- Translation in prokaryotic cell
  - mRNA is translated by ribosomes to produce polypeptide chain

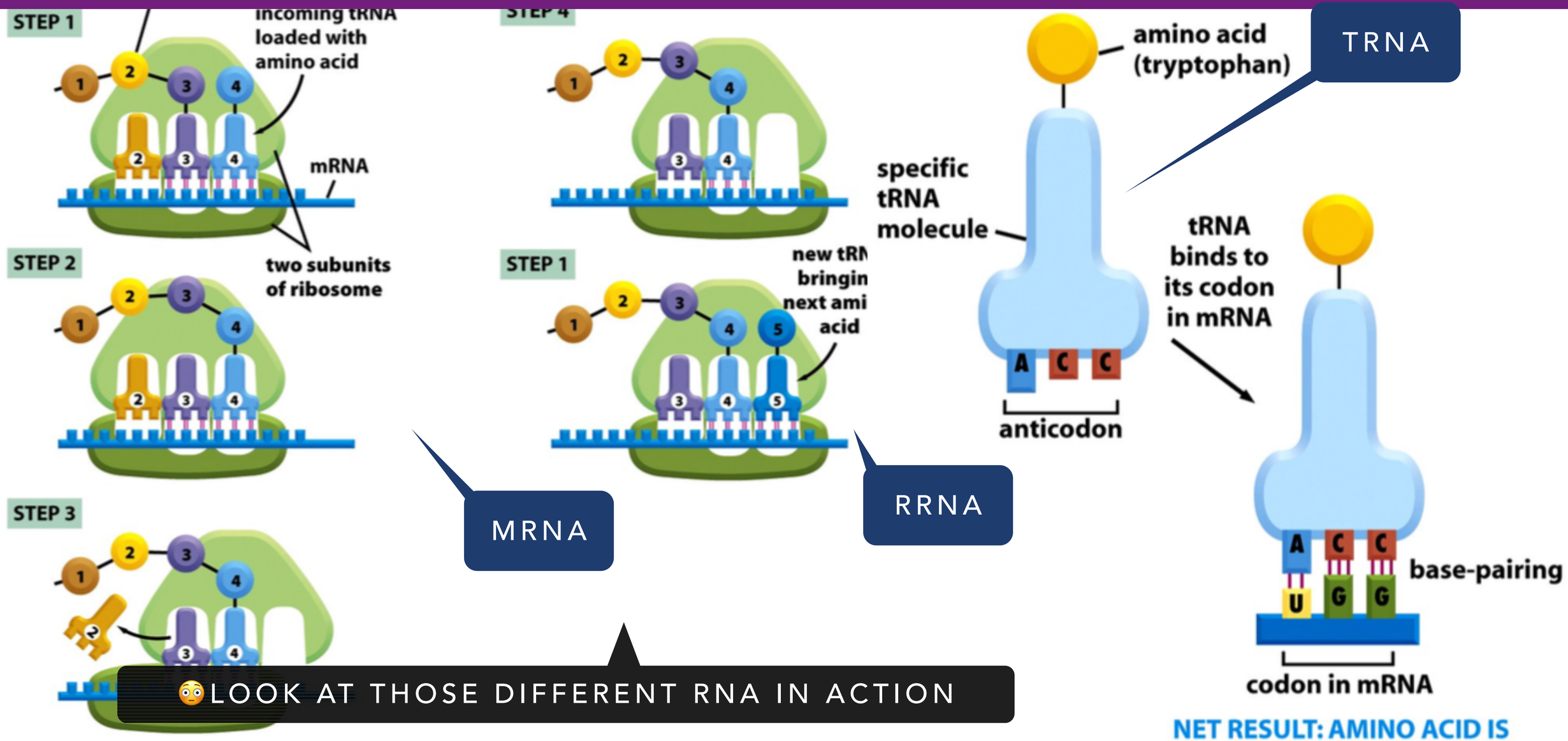


TRANSLATION

- Translation in a Eukaryotic Cell
  - A protein-coding gene is transcribed into a pre-mRNA
  - Pre-mRNA is processed into a mature mRNA
  - mRNA exits the nucleus
  - mRNA is translated on ribosomes to produce the polypeptide chain

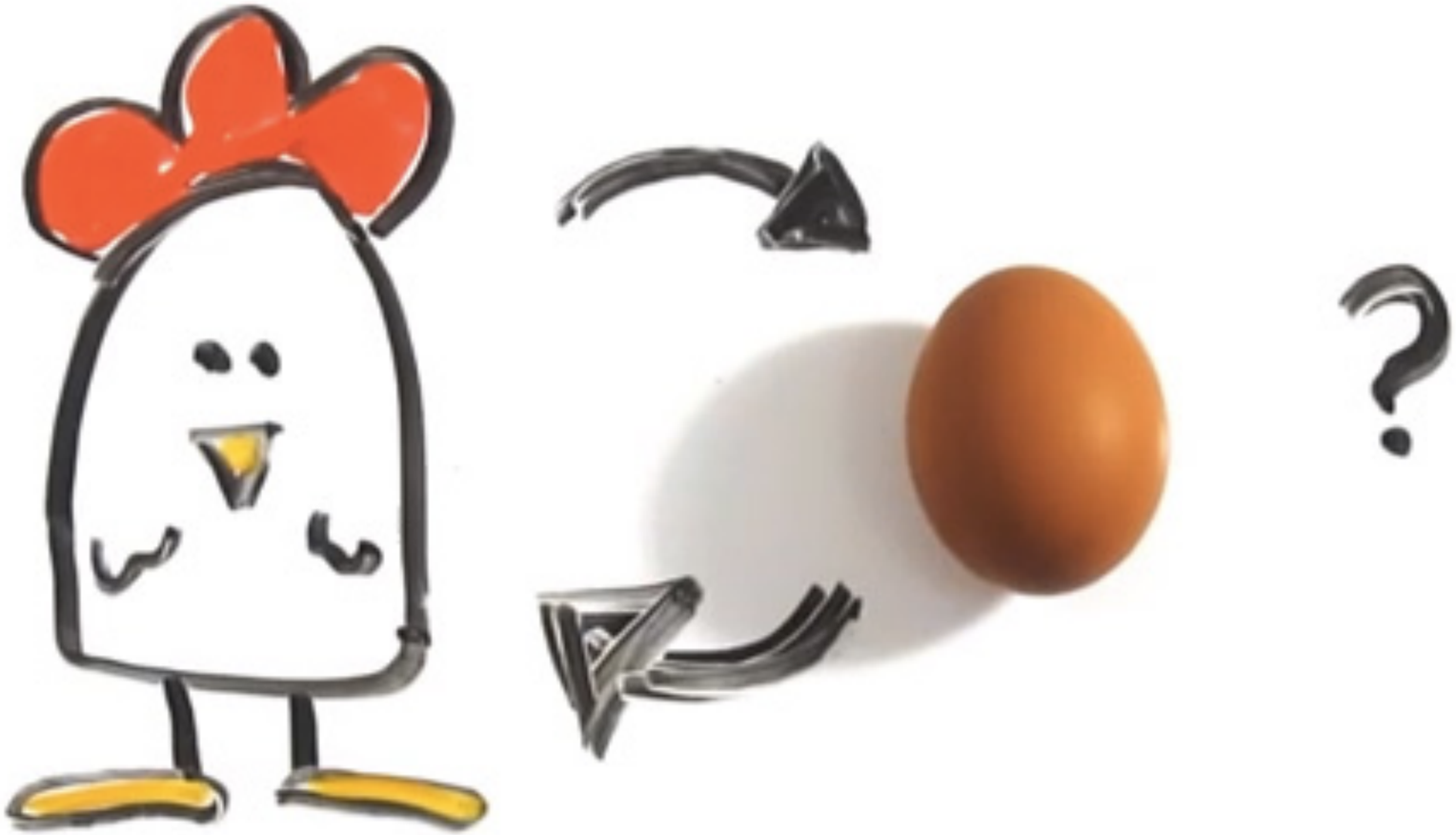


TRANSLATION



😳LOOK AT THOSE DIFFERENT RNA IN ACTION

TRANSLATION



TRANSLATION

- If ribosomes are needed to make proteins but they are also made of proteins, which came first?

From left, Venkatraman Ramakrishnan of the MRC Laboratory of Molecular Biology in Cambridge, England; Thomas A. Steitz of Yale University; and Ada E. Yonath of the Weizmann Institute of Science in Rehovot, Israel, will share the 2009 Nobel Prize in Chemistry.

THE ANSWER, DR. BERG SAID, IS THAT THE ACTIVE CORE OF THE RIBOSOME IS MADE OF RNA. THE PROTEIN SEEMS TO HAVE BEEN ADDED LATER, WHICH MEANS THE RIBOSOME IS “AN RNA-BASED MACHINE THAT EVOLVED THE ABILITY TO MAKE PROTEINS.”

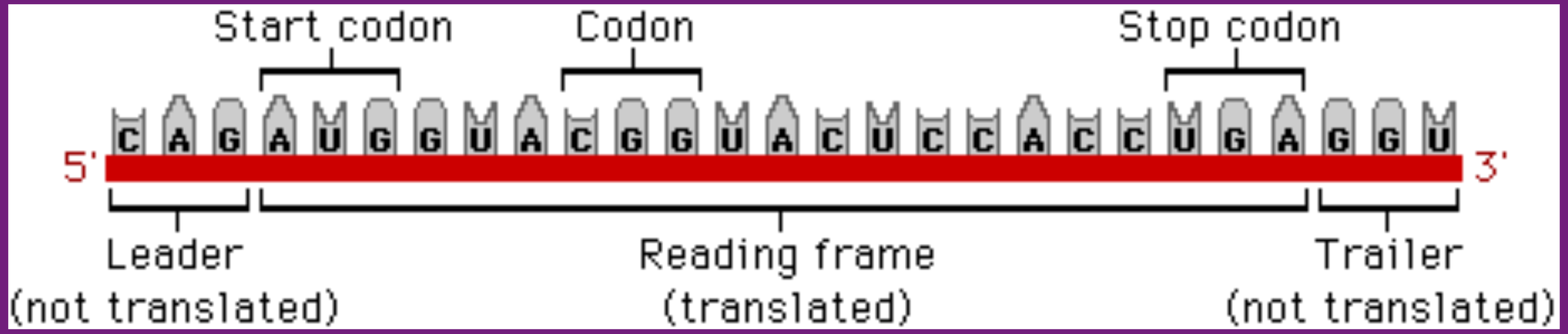
TRANSLATION

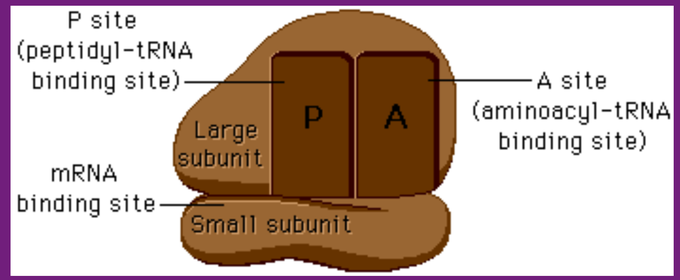
- mRNA is translated into a protein according the genetic code
- Codon
  - 3 nucleotides read at a time
  - Start codon (AUG)
  - Stop codon (UAA,UAG,UGA) tell the ribosome the protein is complete

STANDARD GENETIC CODE

First letter	Second letter U	Second letter C	Second letter A	Second letter G
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly

TRANSLATION

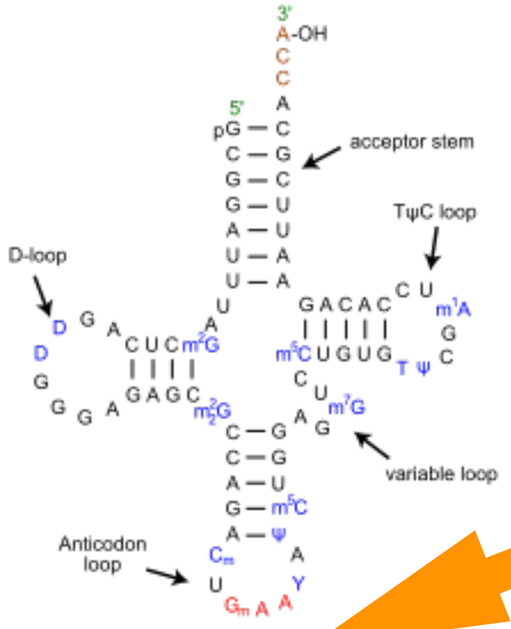


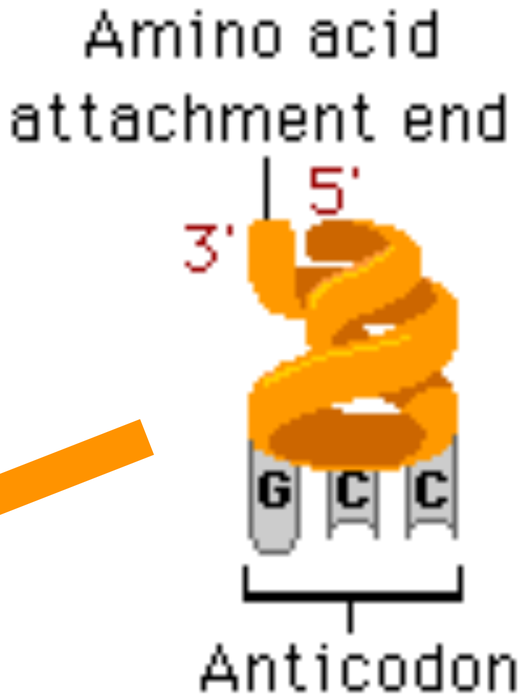


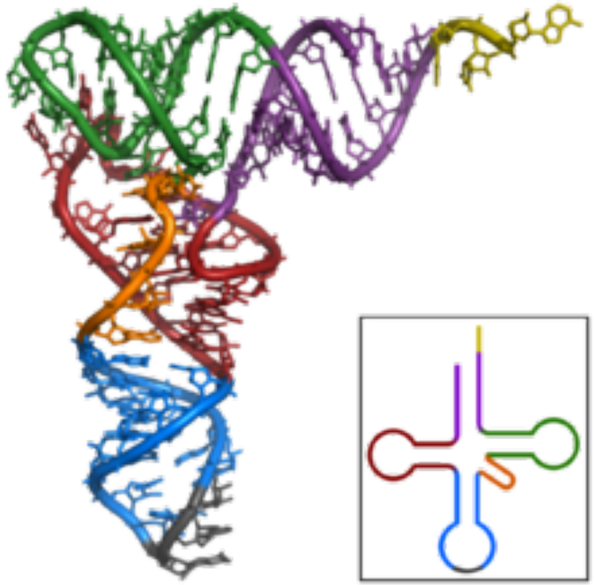
- mRNA passes through the ribosome
- Codons are recognized by tRNAs carrying the specified amino acids
- Each ribosomal subunit consists of rRNA (ribosomal RNA, encoded by rRNA genes) and ribosomal proteins

TRANSLATION

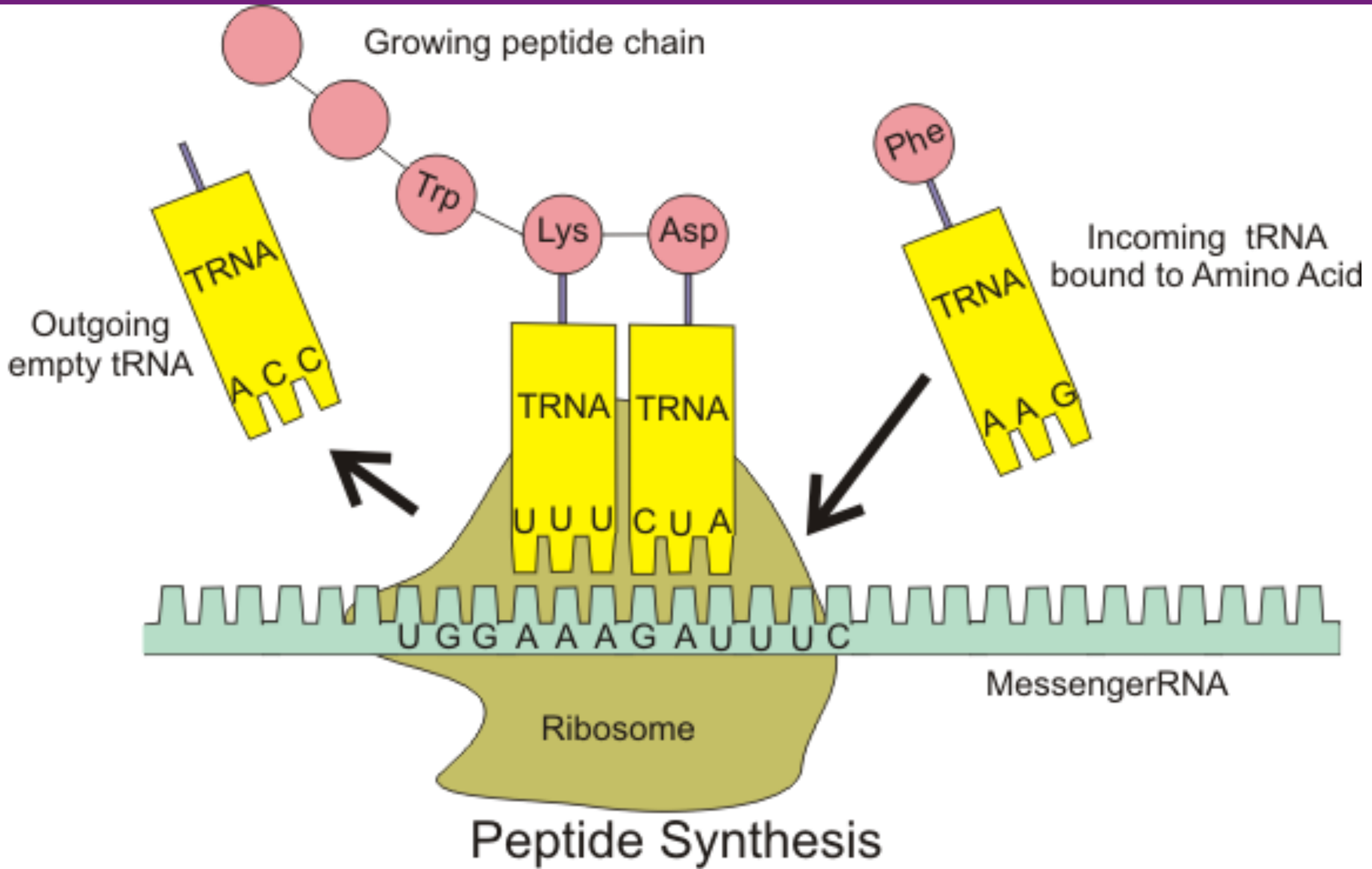
- Transfer RNA
  - Encoded by tRNA genes
  - All tRNA molecules are similar in size and shape (cloverleaf)
  - All tRNAs have CCA at the 3' end to which the amino acid attaches
  - Anticodon “reads" the matching codon on the mRNA





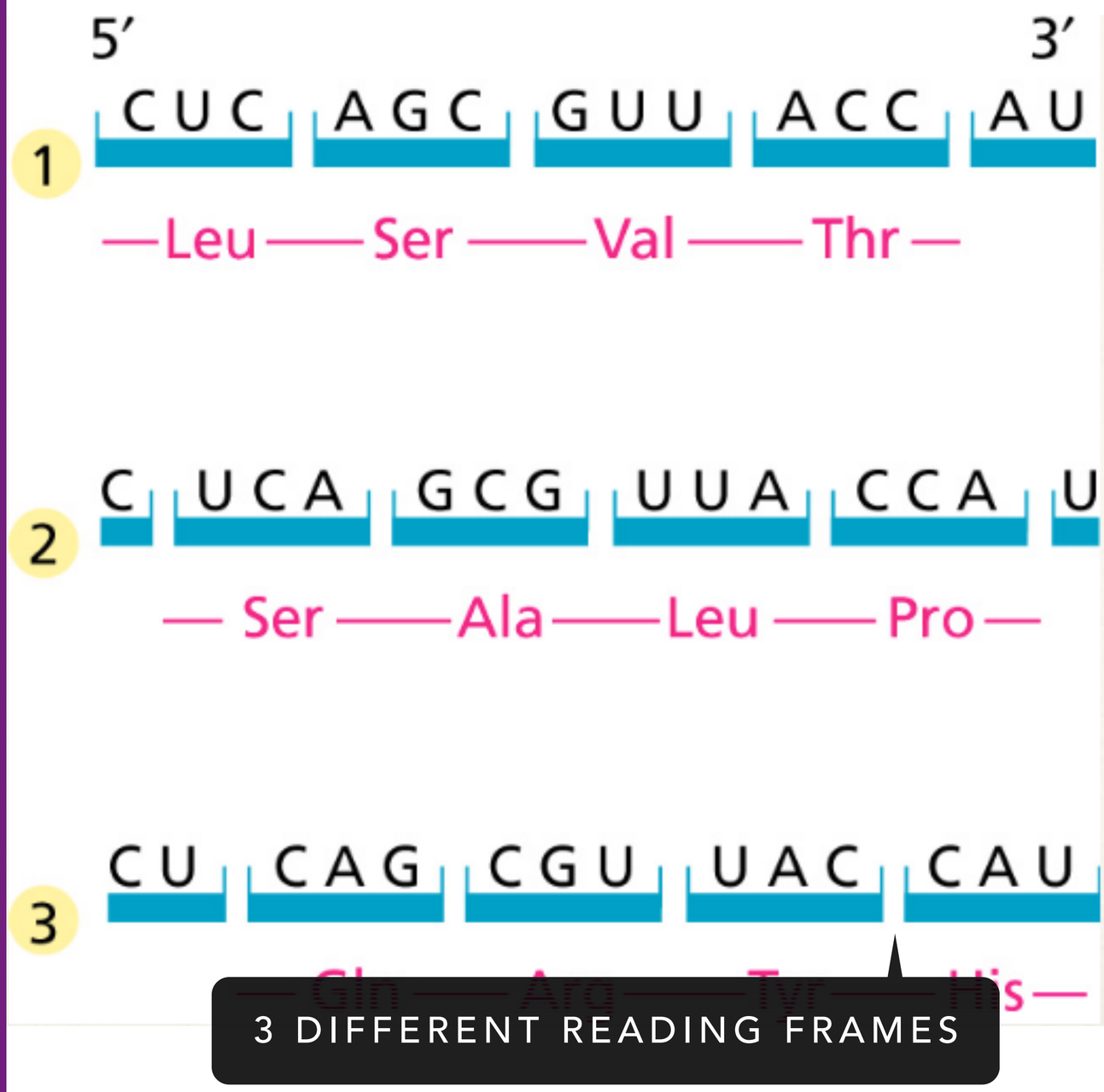


TRANSLATION

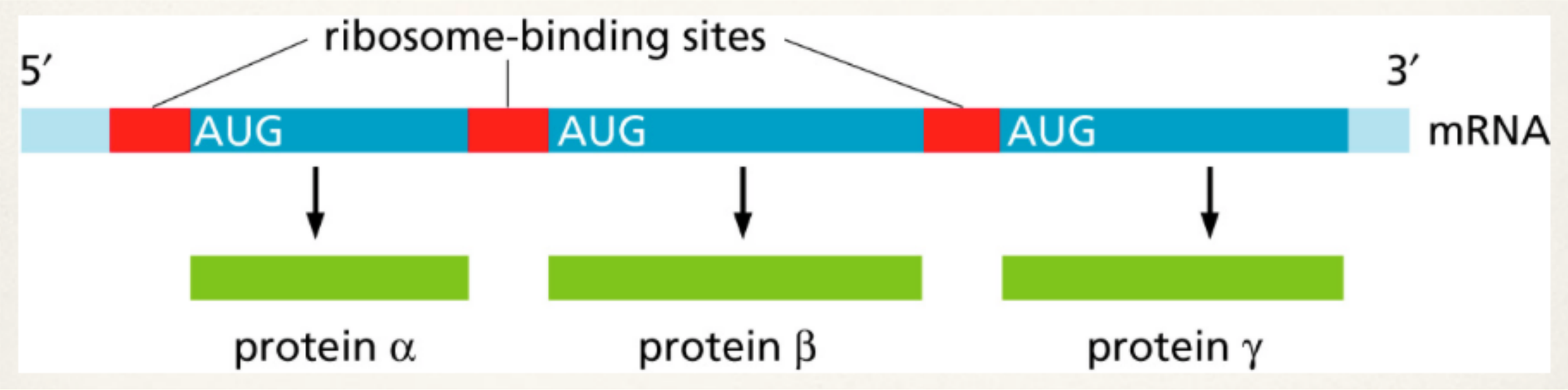


TRANSLATION

- Reading frames
  - Translation occurs in nonoverlapping sets of three bases
  - There are thus three possible ways to translate any nucleotide sequence
  - These three reading frames give three different protein sequences
- Detailed control signals ensure that only the appropriate reading frame is translated into protein



TRANSLATION



- Functionally related protein-coding sequences are often clustered together into **operons**
- Each operon is transcribed as a single mRNA transcript; proteins are separately translated from this one long molecule
  - A single **operator** activate the simultaneous expression of all genes in the operon
- Allows efficient, coordinated protein synthesis

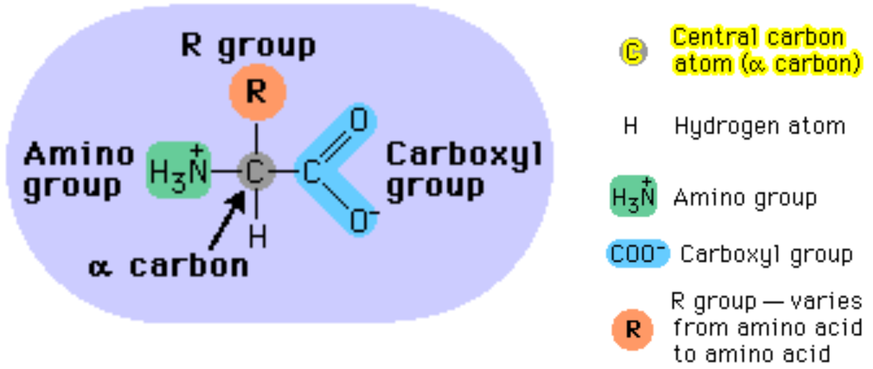
Proteins

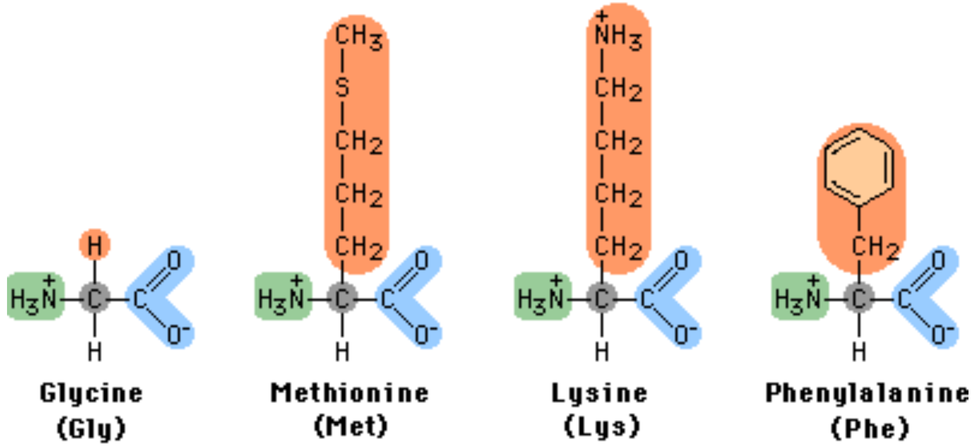
PROTEINS

- Proteins are large, complex molecules made of amino acid residues
- Workhorse of cells
- Required for the structure, function, and regulation of the body’s tissues and organs

PROTEINS

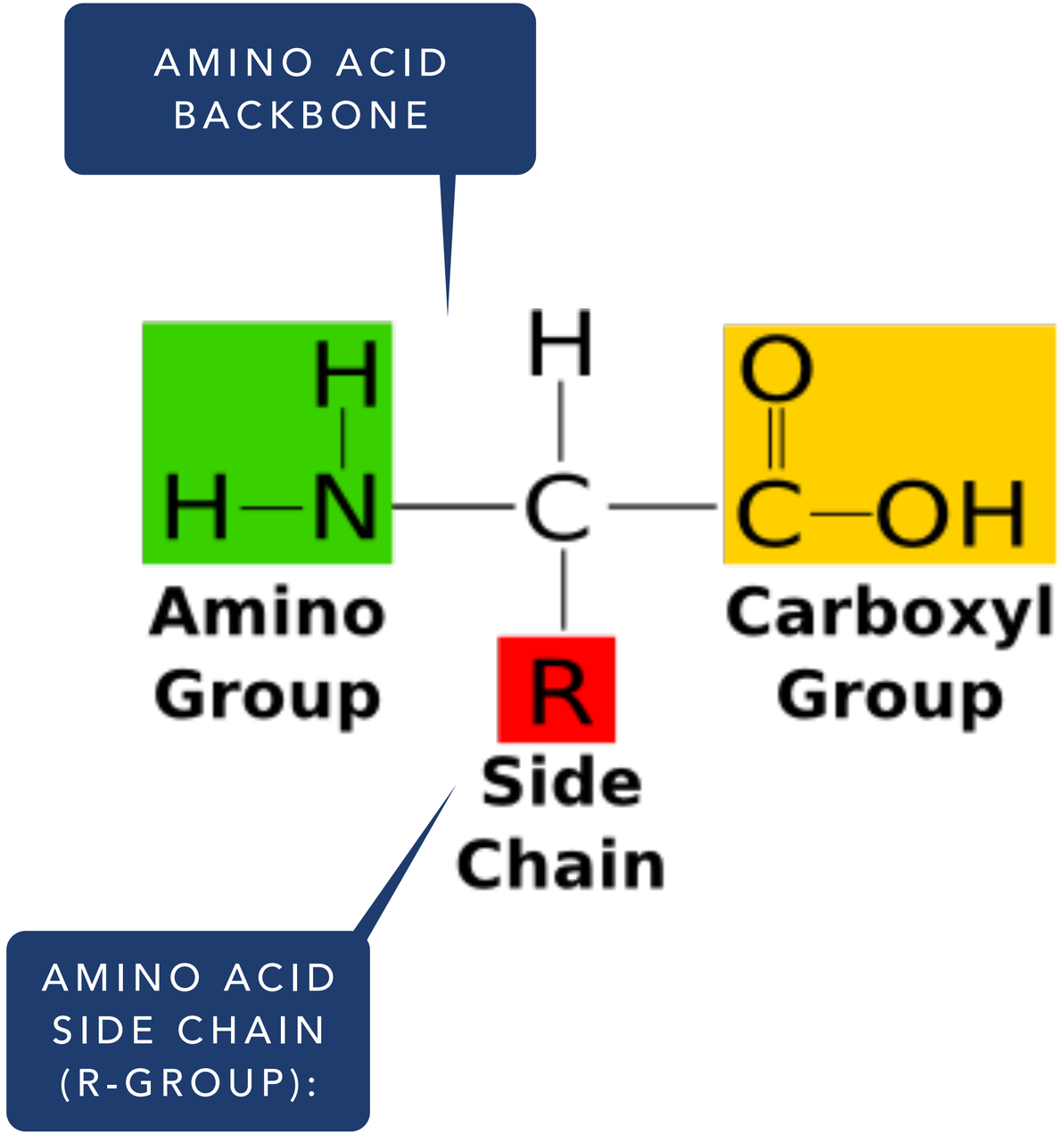
- Proteins structure
  - Building blocks are amino acids
  - There are 20 standard amino acids
    - Differ only by R group



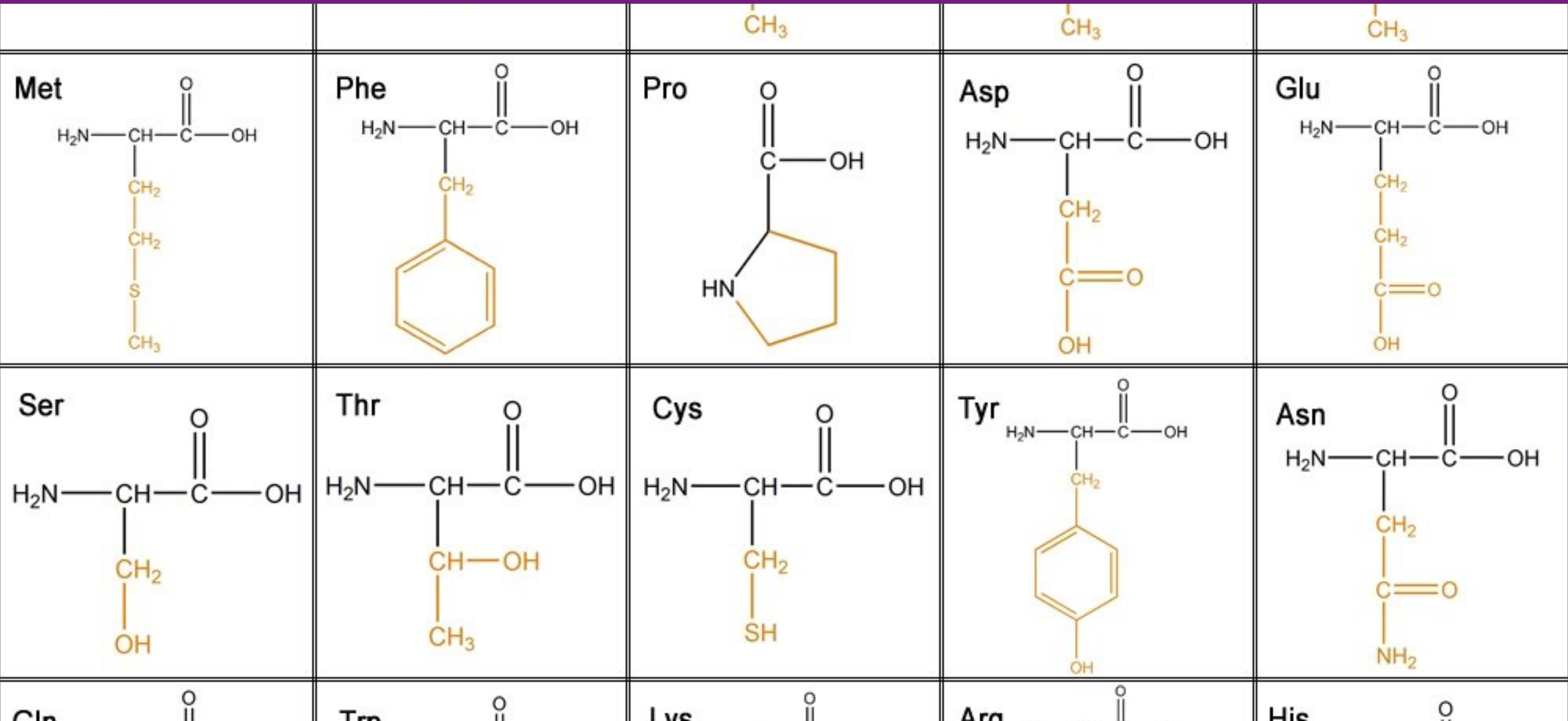


PROTEINS

- Amino acid side chains
- Differ by their side chains (R-groups)
- Determine chemical properties

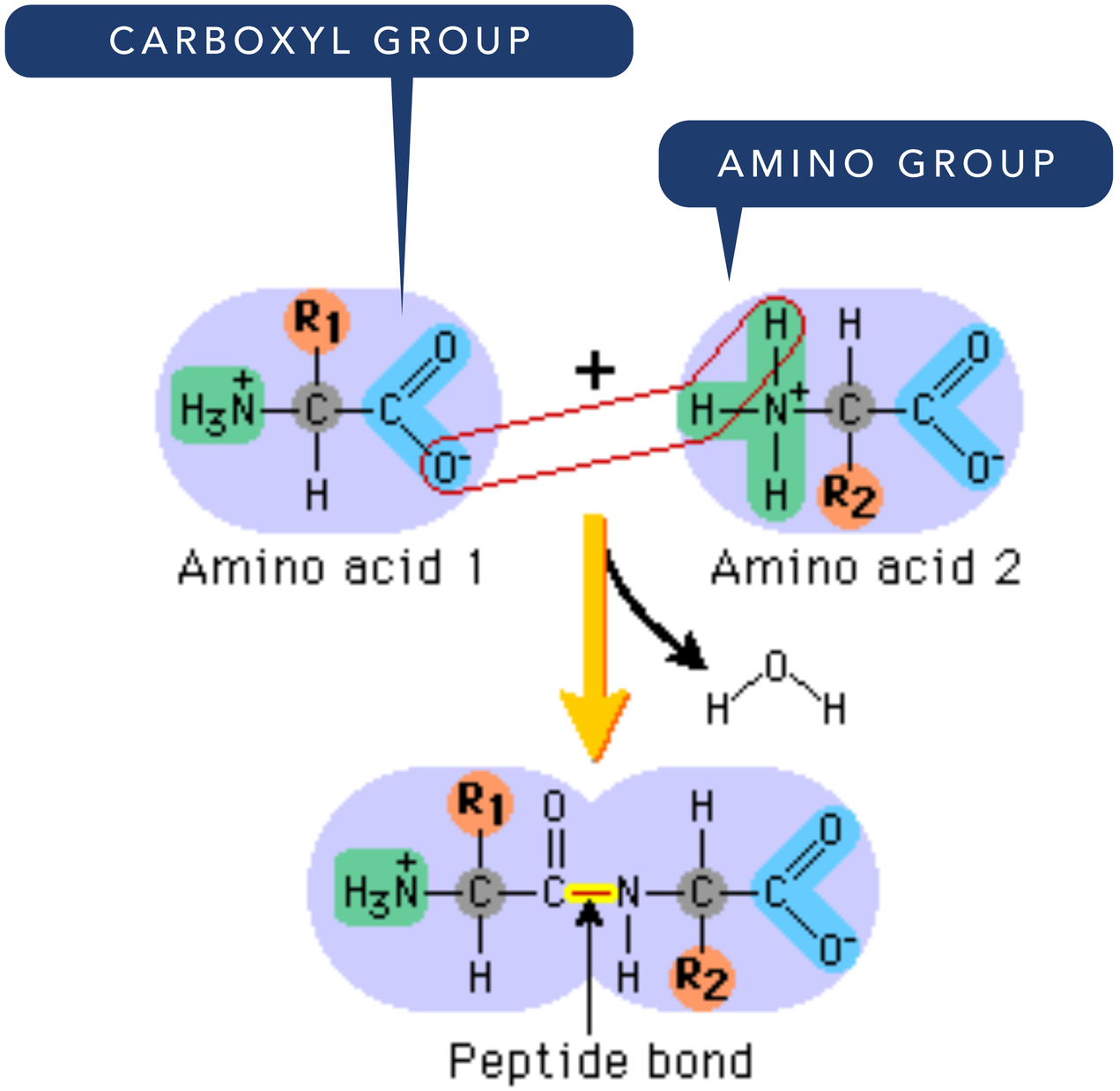


PROTEINS



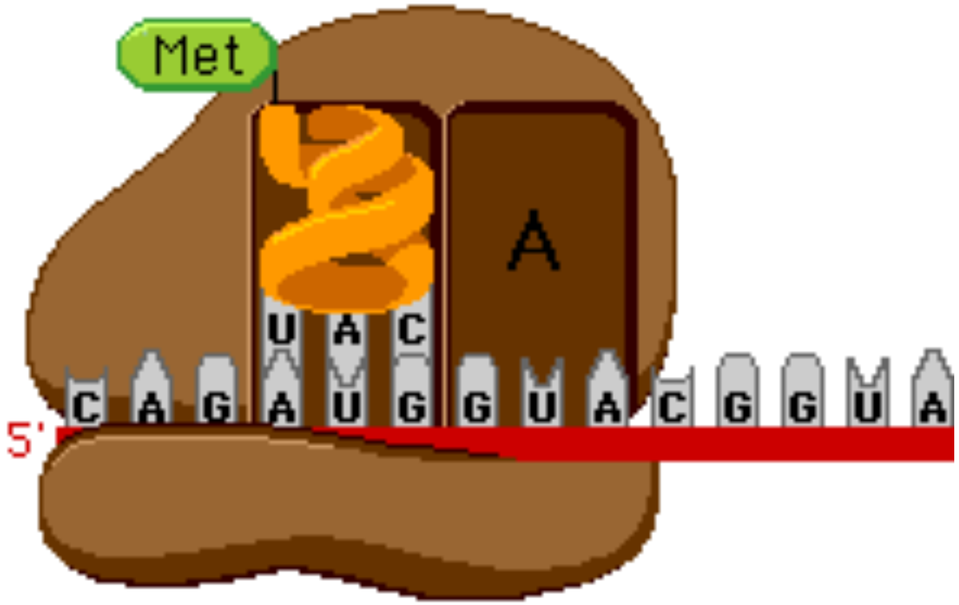
PROTEINS

- The Peptide Bond
  - Joins amino acids in proteins
  - Forms between the carboxyl group of one amino acid and the amino group of the adjacent amino acid



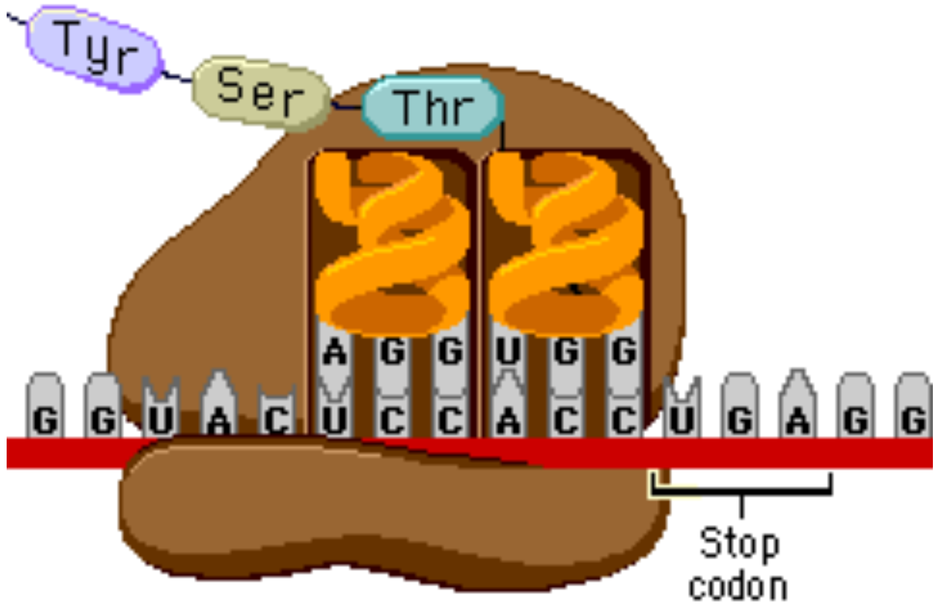
PROTEINS

- Elongation of the peptide chain
  - tRNA anti-codon recognized the complementary mRNA codon
  - Amino acids are joined by peptide bond



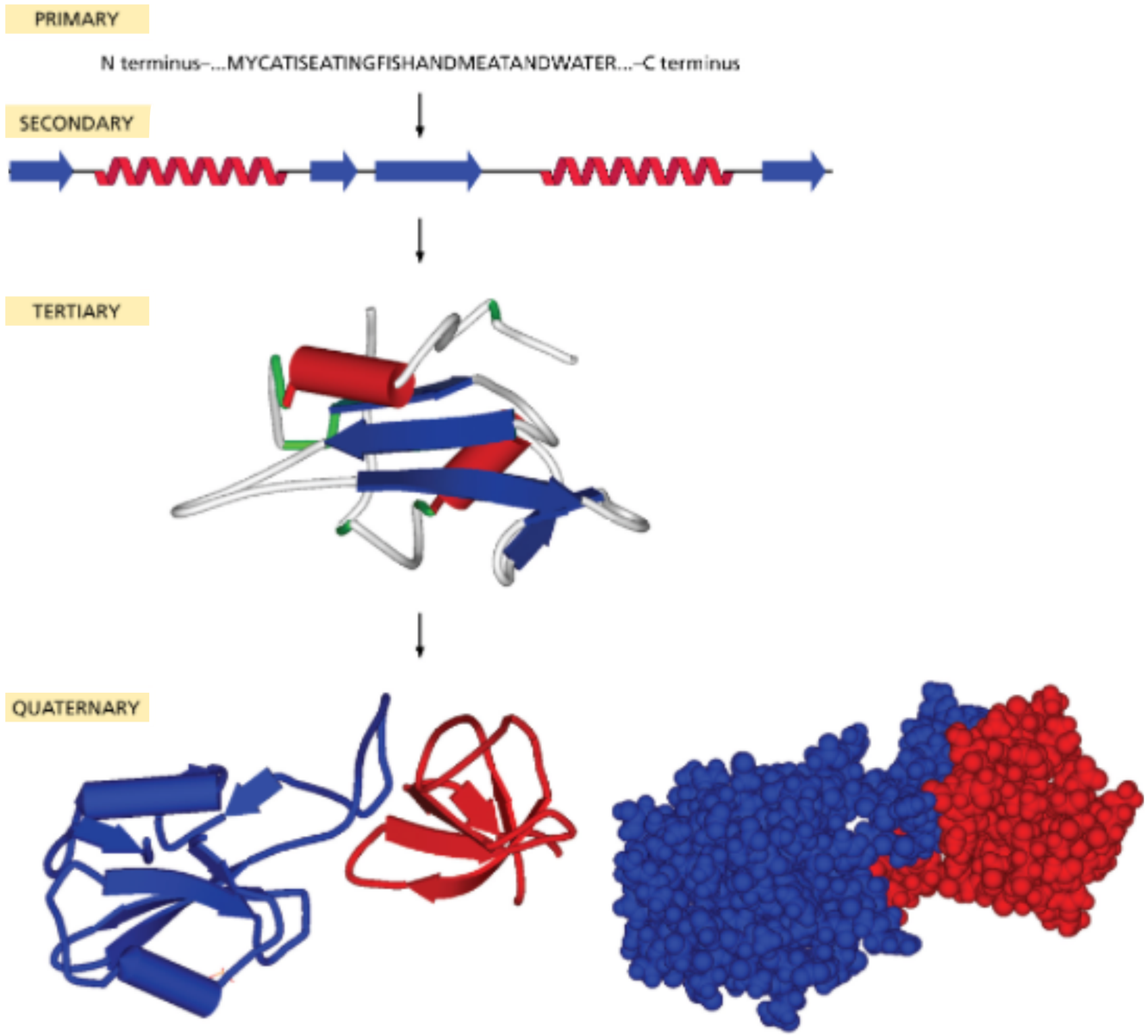
PROTEINS

- Termination of the peptide chain
  - Release factor - protein that recognizes the stop code in an mRNA sequence



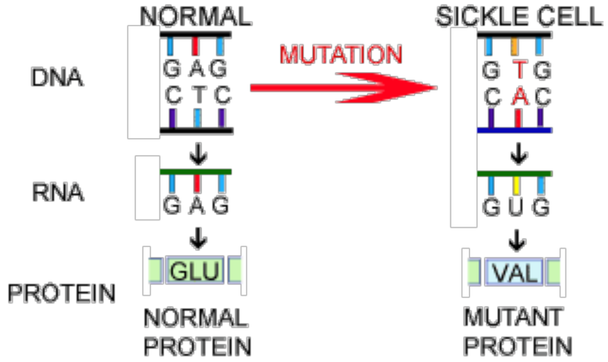
PROTEINS

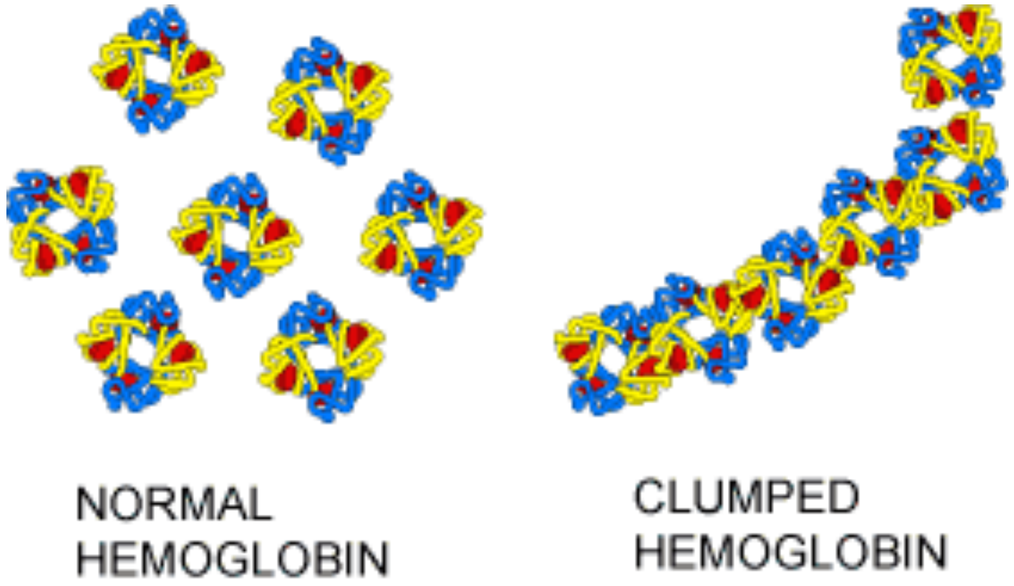
- Primary sequence
- Secondary structure
- Tertiary structure
  - 3 dimensional
  - X-ray crystallography
- Quarternary
  - Assembled molecular units



PROTEINS

- Sickle cell anemia
  - Mutation in hemoglobin
    - Carries oxygen in red blood cells
  - Deprived of oxygen cells become sickle-shaped
  - Carrier experiences pain and fatigue
  - Carrier are resistant to malaria
    - Parasites are killed in blood cells





Mutations in DNA

MUTATIONS

- Change of nucleotide or amino acid at a given position
- Driven by evolution or environmental factors

TRANSCRIPTIONS & TRANSLATION

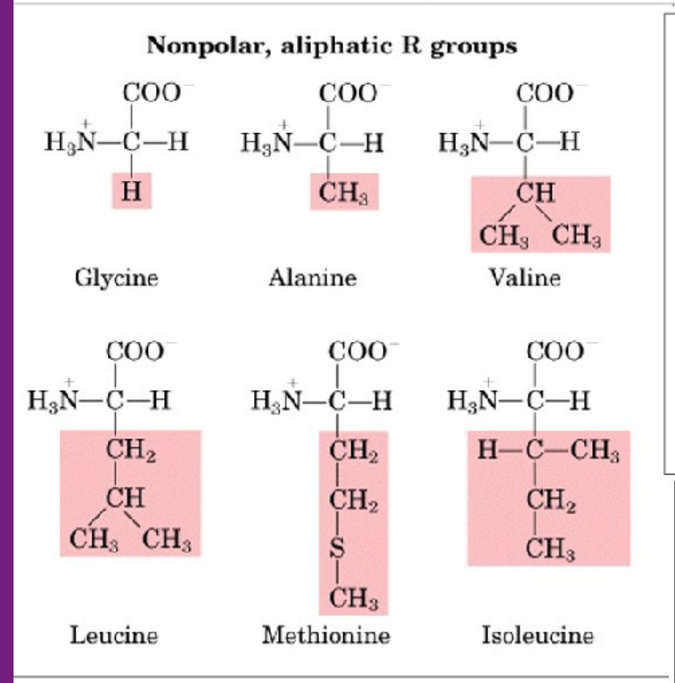
- Mutations in DNA
  - Substitution
    - Non-synonymous mutation - amino acid sequence is changed (“sense”)
    - Synonymous mutation - do not change an amino acid sequence (“silent mutation”)
    - Nonsense mutation - base changes to stop codon, premature termination of translation
  - Insertion
  - Deletion
  - Frameshift

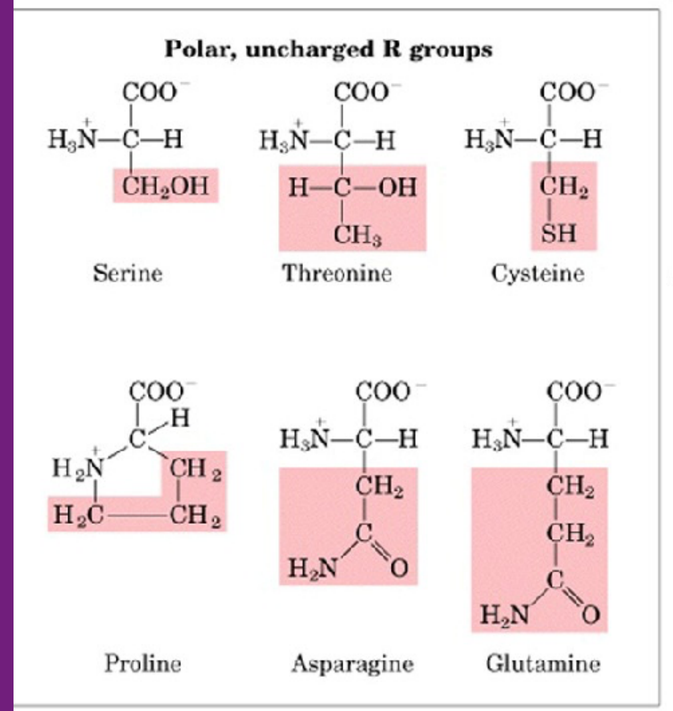
Second Position			
U	C	A	G
UUU F	UCU S	UAU Y	UGU C
UUC F	UCC S	UAC Y	UGC C
UUA L	UCA S	UAA stop	UGA stop
UUG L	UCG S	UAG stop	UGG W
CUU L	CCU P	CAU H	CGU R
CUC L	CCC P	CAC H	CGC R
CUA L	CCA P	CAA Q	CGA R
CUG L	CCG P	CAG Q	CGG R
AUU I	ACU T	AAU N	AGU S
AUC I	ACC T	AAC N	AGC S
AUA I	ACA T	AAA K	AGA R
AUG M	ACG T	AAG K	AGG R
GUU V	GCU A	GAU D	GGU G
GUC V	GCC A	GAC D	GGC G
GUA V	GCA A	GAA E	GGA G
GUG V	GCG A	GAG E	GGG G

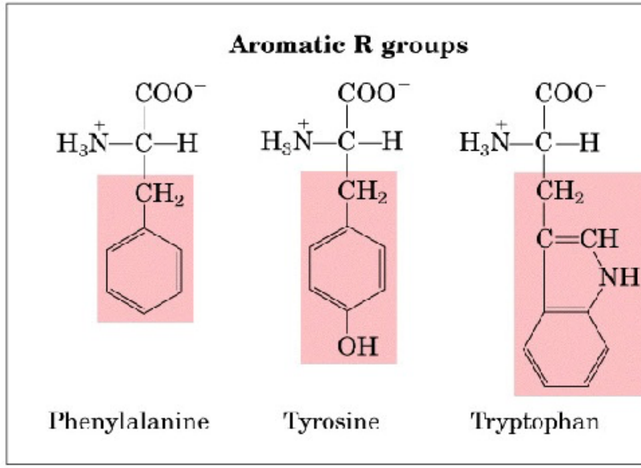
http://www.nature.com/scitable/topicpage/genetic-mutation-441

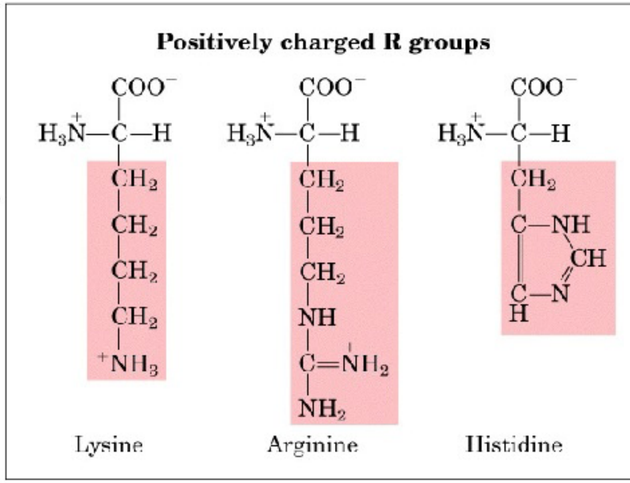
MUTATIONS

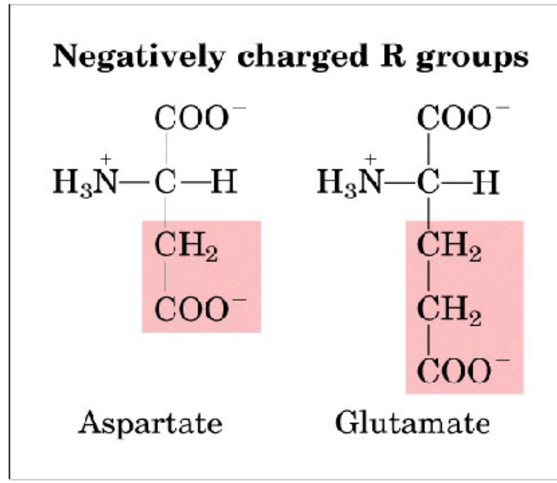
- Amino acids differ by their side chains (R-groups)
- Grouped by properties
  - Hydrophobic or non polar
    - which are amino acids with side chains that repel water
  - Hydrophilic or polar
    - which are amino acids with side chains that are attracted to water
  - Acidic or negatively-charged
    - Amino acids that contain carboxyl groups (-COO-) as side chains
  - Basic or positively-charged side chains
    - Amino acids that contain amine groups (-NH3+) as side chains





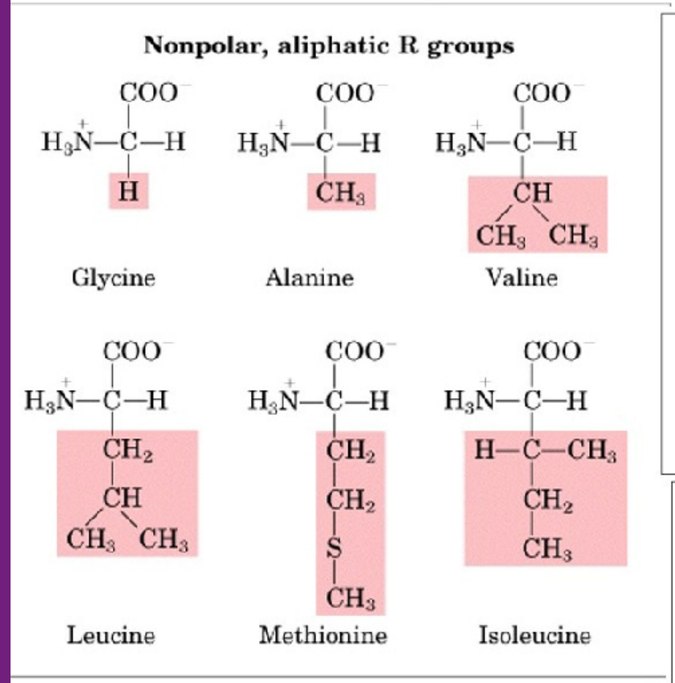


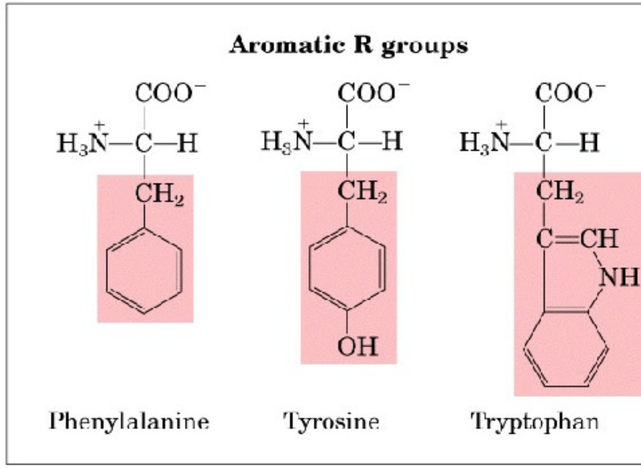


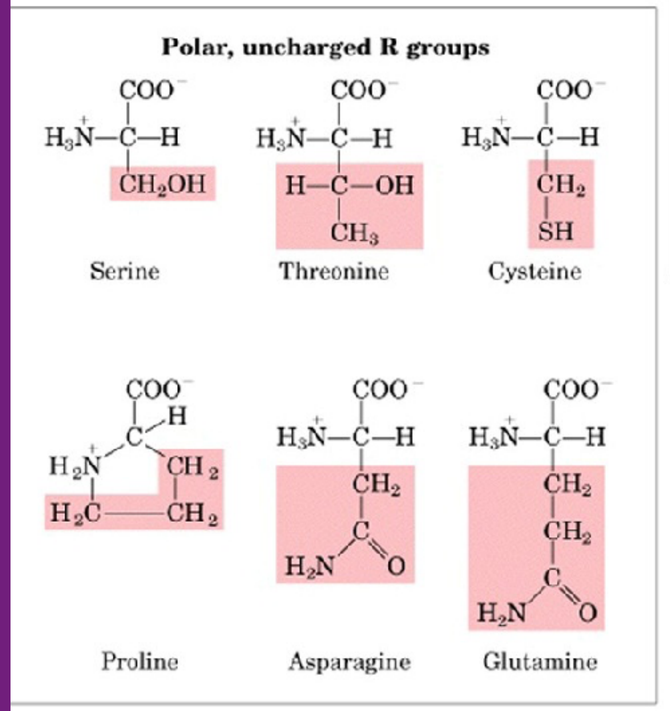


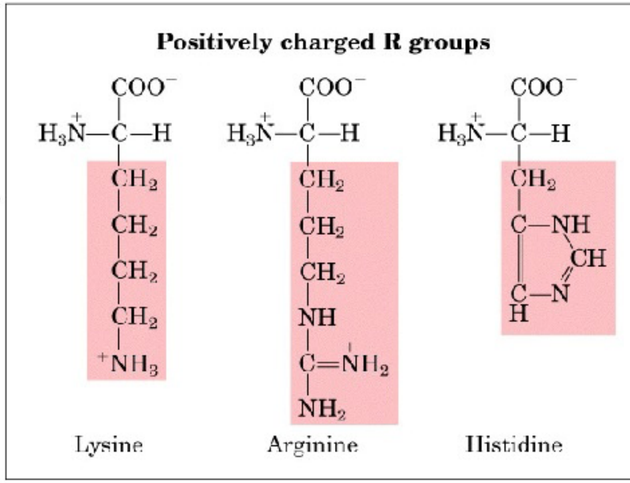
MUTATIONS

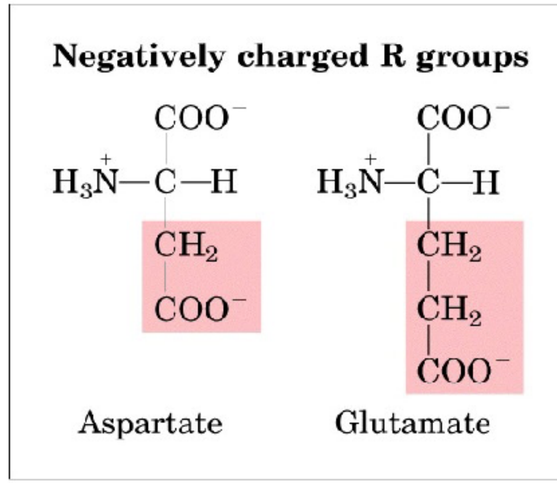
- Mutations in protein sequence
  - Non-conserved
    - results in different side chain
  - Conserved
    - results in same side chain





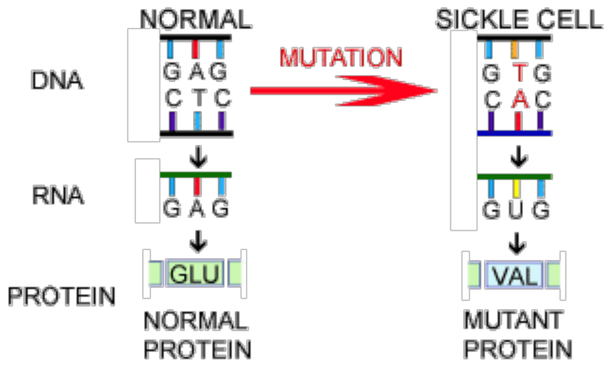


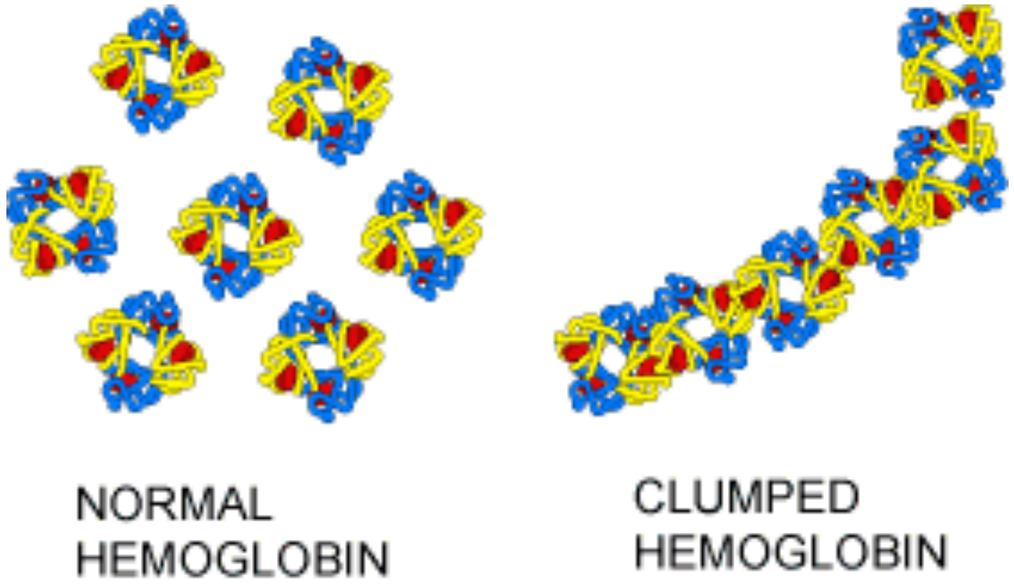


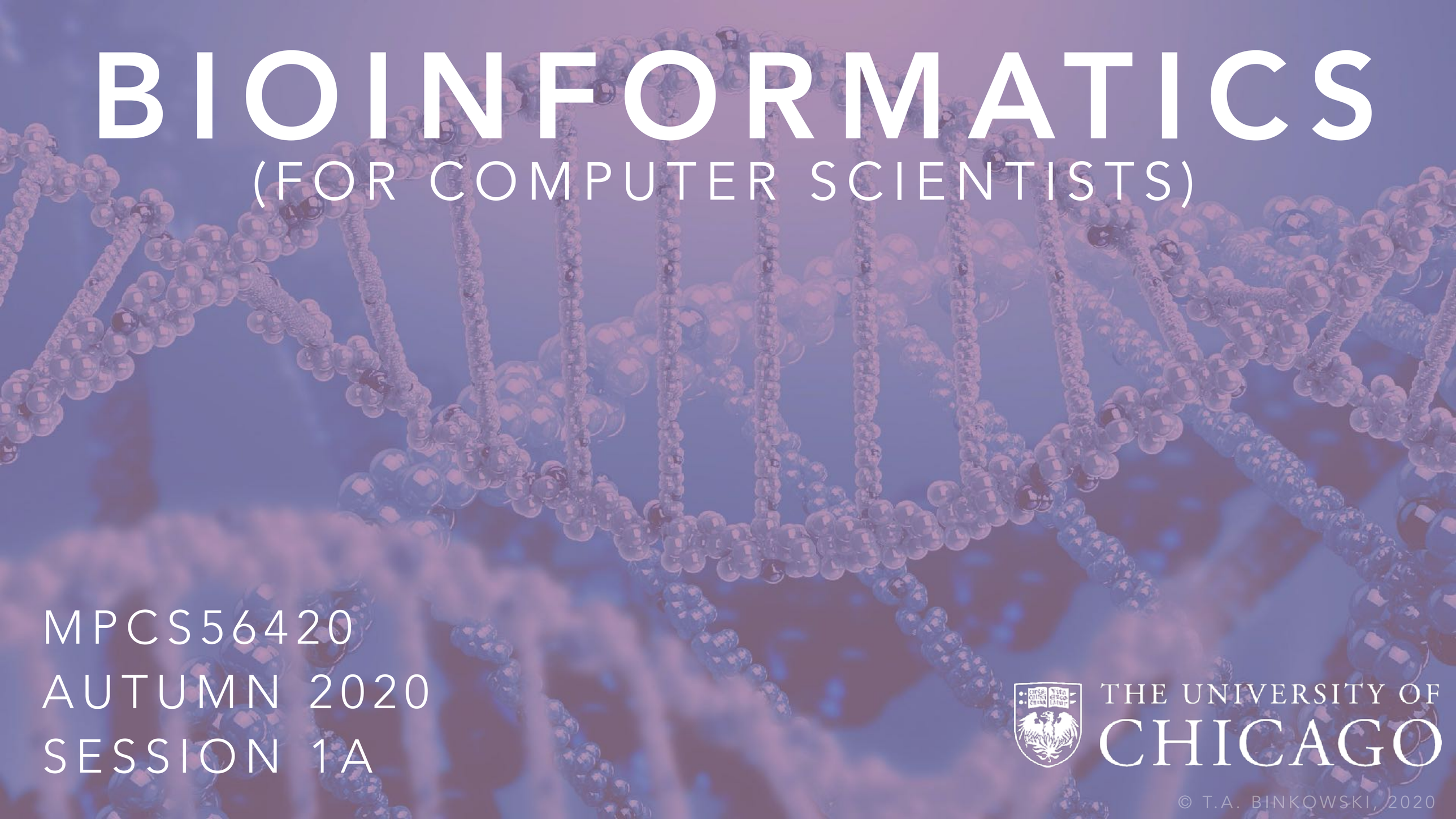


TRANSCRIPTIONS & TRANSLATION

- Sickle cell anemia
  - Mutation in hemoglobin
    - Carries oxygen in red blood cells
  - Deprived of oxygen cells become sickle-shaped
  - Carrier experiences pain and fatigue
  - Carrier are resistant to malaria
    - Parasites are killed in blood cells







BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

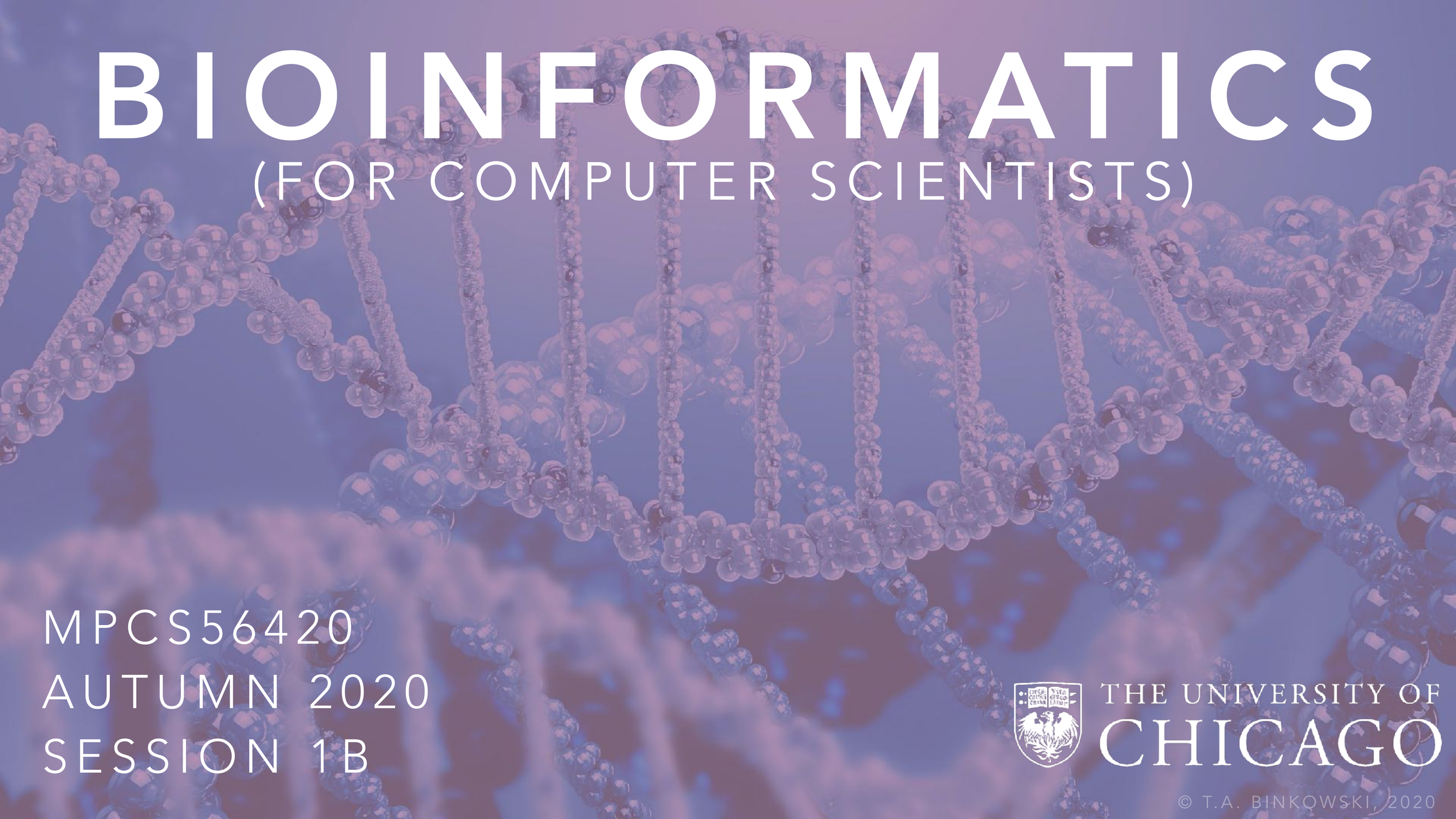
MPCS56420
AUTUMN 2020
SESSION 1A

THE UNIVERSITY OF CHICAGO

SESSION 1B

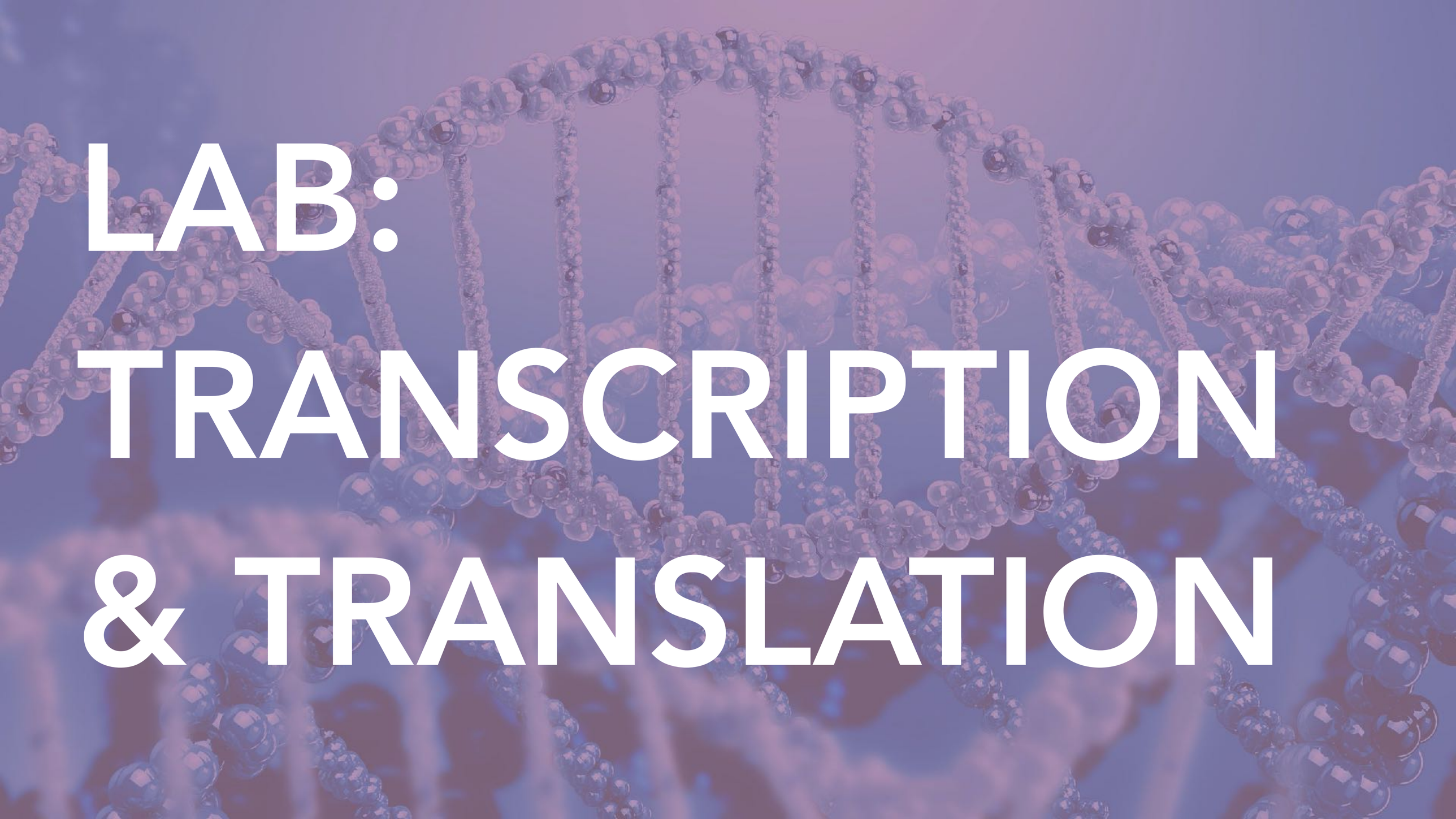
BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)



MPCS56420
AUTUMN 2020
SESSION 1B

THE UNIVERSITY OF CHICAGO



LAB: TRANSCRIPTION & TRANSLATION

TRANSCRIPTION & TRANSLATION

WHAT YOU SHOULD HAVE GOT OUT OF THE LAST HOUR

- DNA contains the genetic information that encodes traits
- DNA is double stranded, complementary and anti-parallel
- The beginning of a DNA strand is called the 5’ (“five prime”) region and the end of a DNA strand is called the 3’ (“three prime”) region
- Proteins are produced through the processes of transcription and translation
- Amino acids are encoded by nucleotide triplets called codons
- mRNA transcripts contain “start” and “stop” codons that initiate and terminate protein translation

TRANSCRIPTION & TRANSLATION

- DNA is complementary and anti-parallel

Gene or
coding *or*
sense strand

5′– CCGATGTCATAAGAC – 3′

DNA IS READ 5′ TO 3′

GENES ARE TRANSCRIBED FROM 5′ TO 3′

TRANSCRIPTION & TRANSLATION

- DNA is complementary and anti-parallel

Gene or
coding *or*
sense strand

Template or
non-coding or
anti-sense strand

5′– CCGATGTCATAAGAC – 3′

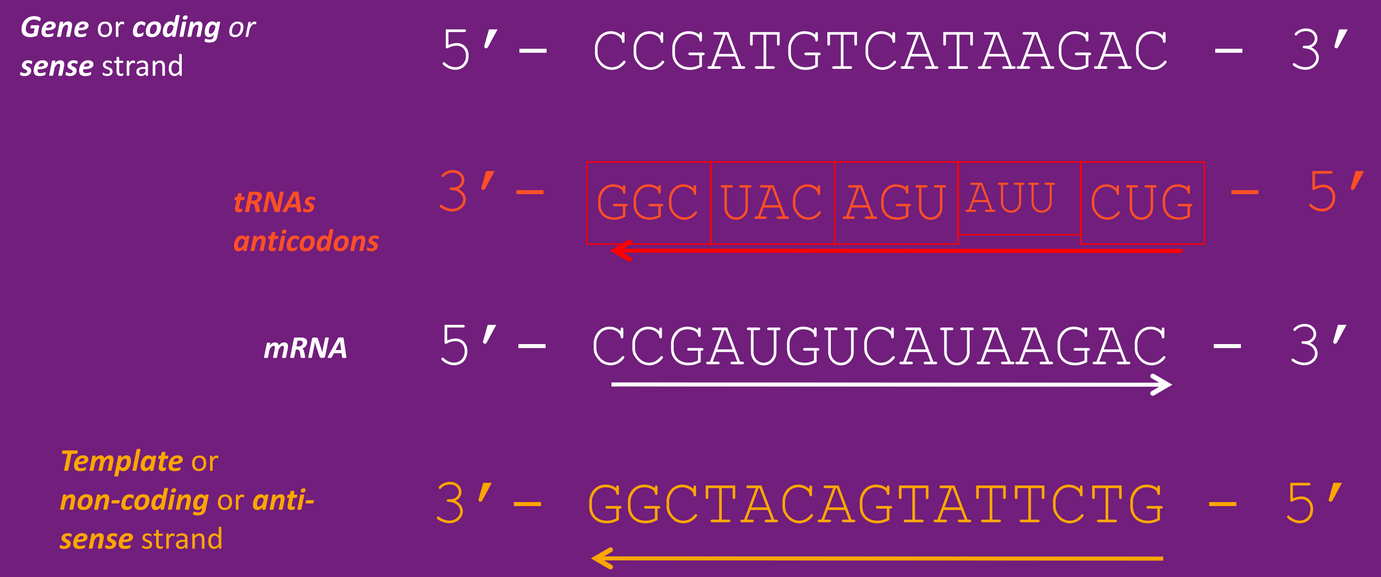
3′– GGCTACAGTATTCTG – 5′

DNA STRANDS ARE ANTI-PARALLEL, PARALLEL BUT OPPOSITE

DNA STRANDS ARE COMPLEMENTARY; ADENINE (A) PAIRS WITH THYMINE (T), CYTOSINE (C) PAIRS WITH GUANINE (G)

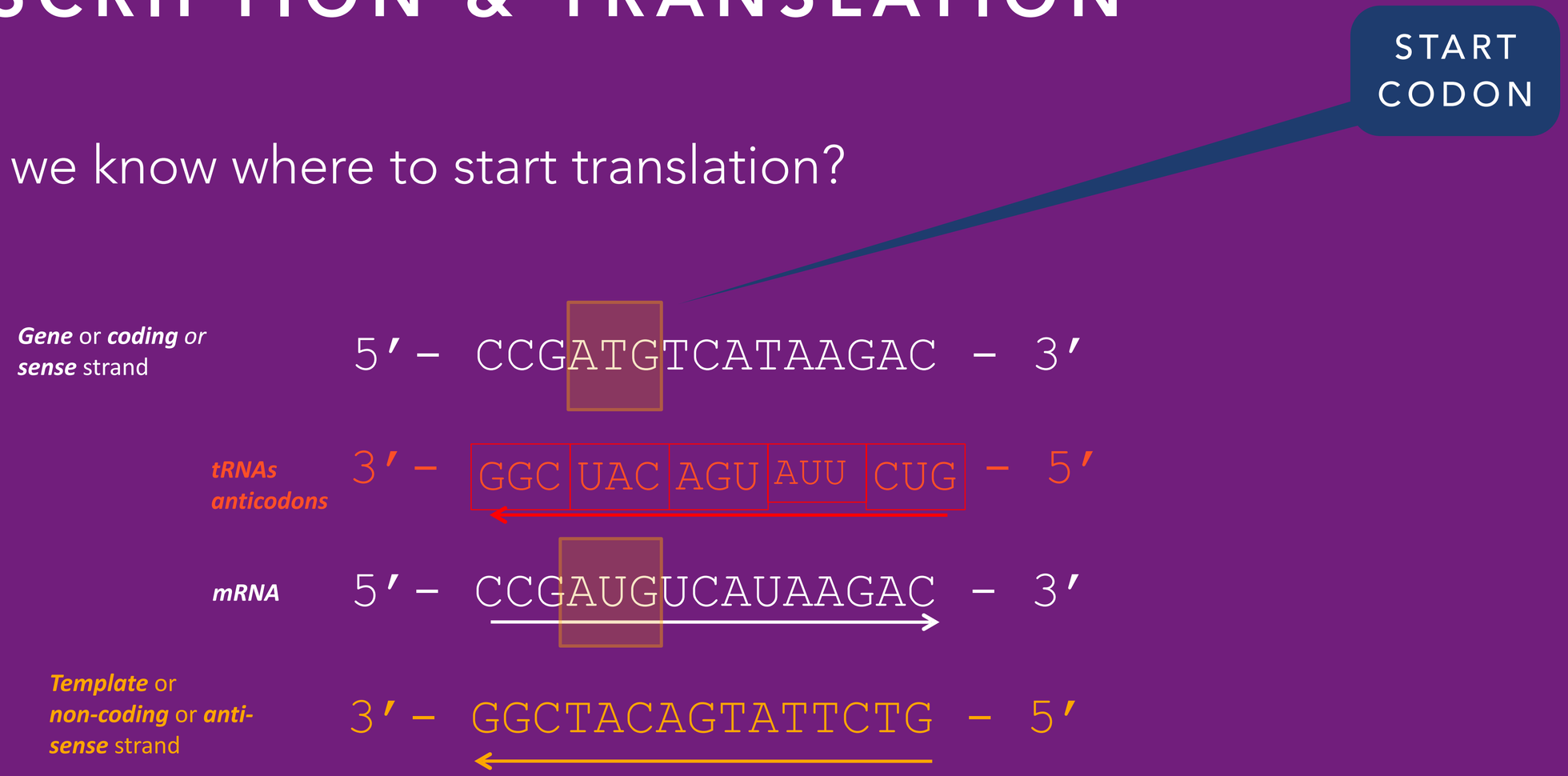
TRANSCRIPTION & TRANSLATION

- Translating DNA into proteins



TRANSCRIPTION & TRANSLATION

- How do we know where to start translation?



TRANSCRIPTION & TRANSLATION

- Codon
  - Series of 3 nucleotides in a row
  - Specifies the genetic code information for a particular amino acid (e.g. AAU = I)
  - Also called “nucleotide triplet”
  - Codon table

First Position 5′	Second Position				Third Position 3′
	U	C	A	G	
U	UUU F UUC F UUA L UUG L	UCU S UCC S UCA S UCG S	UAU Y UAC Y UAA stop UAG stop	UGU C UGC C UGA stop UGG W	U C A G
C	CUU L CUC L CUA L CUG L	CCU P CCC P CCA P CCG P	CAU H CAC H CAA Q CAG Q	CGU R CGC R CGA R CGG R	U C A G
A	AUU I AUC I AUA I AUG M	ACU T ACC T ACA T ACG T	AAU N AAC N AAA K AAG K	AGU S AGC S AGA R AGG R	U C A G
G	GUU V GUC V GUA V GUG V	GCU A GCC A GCA A GCG A	GAU D GAC D GAA E GAG E	GGU G GGC G GGA G GGG G	U C A G

STOP CODONS

START CODON

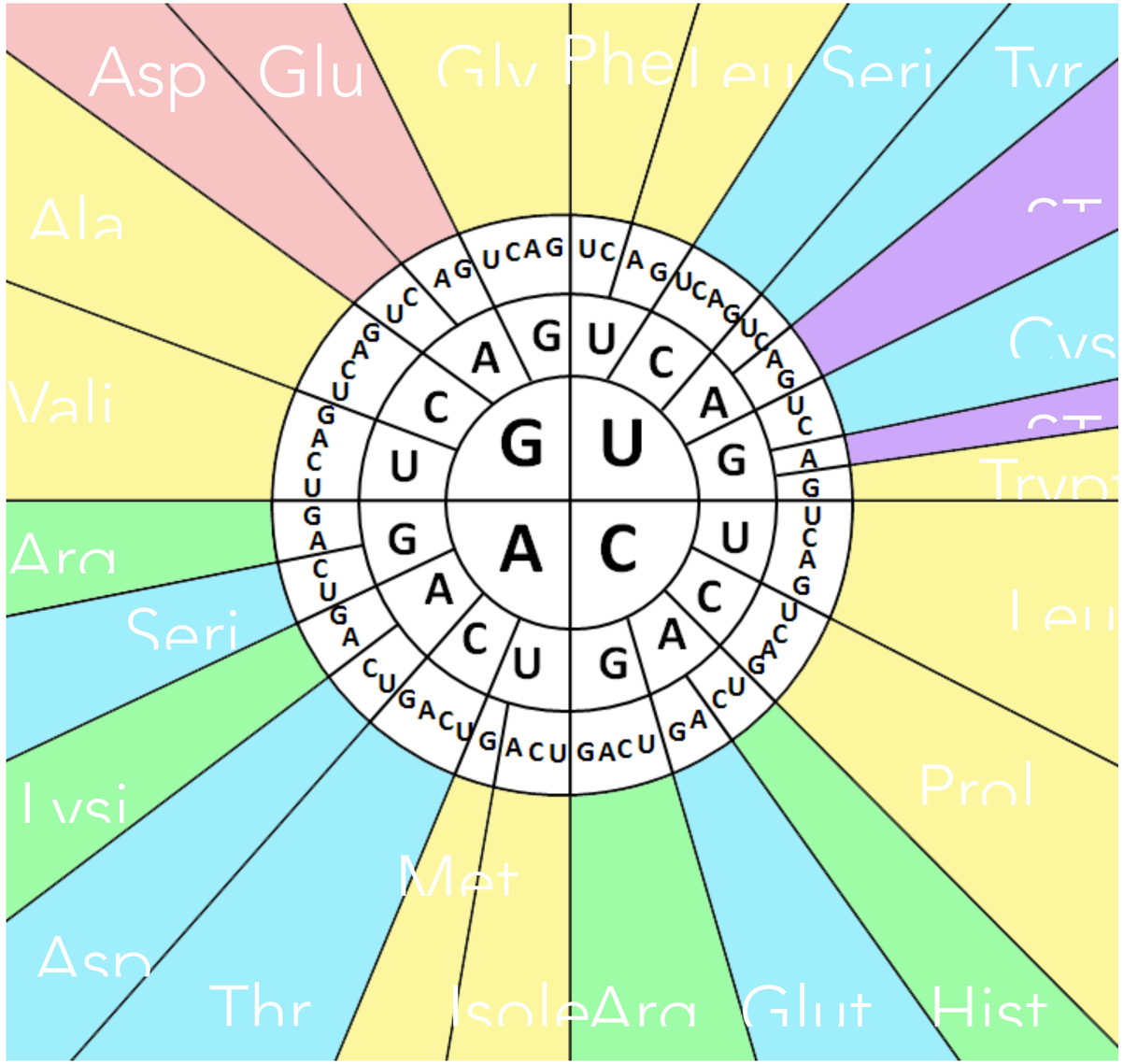
AMINO ACID SINGLE-LETTER

TRANSCRIPTION & TRANSLATION

- Built-in redundancy

Side Chain (R-Group) Chemistry:

- Hydrophobic / Nonpolar
- Hydrophilic/ Polar
- Acidic / Negative
- Basic / Positive



TRANSCRIPTIONS & TRANSLATION

- Reading frames
  - Non-overlapping sequence of three-nucleotide codons
  - There are 3 possible reading frames in an mRNA strand
  - There are 6 in a double-stranded DNA molecule (three reading frames from each of the two DNA strands)
  - Nomenclature
    - 1,2,3 coding strand
    - -1,-2,-3 for template strand

TRANSCRIPTIONS & TRANSLATION

"Gene" Sequence: thecatatetherat.

PERIOD IS "STOP CODON"

Reading Frame +1 starts at the **first** letter:

the cat ate the rat.

Reading Frame +2 starts at the **second** letter:

t hec ata tet her at.

Reading Frame +3 starts at the **third** letter:

th eca tat eth era t.

Reading Frames -1, -2 & -3 would be like reading the sentence "backwards."

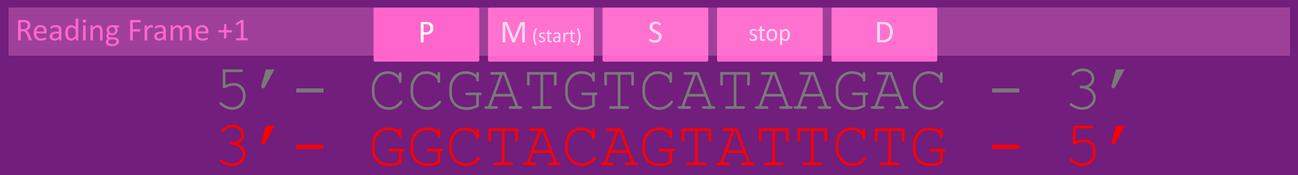
TRANSCRIPTIONS & TRANSLATION

- Open reading frame (ORF)
  - A reading frame that contains a start codon and a stop codon, with multiple three-nucleotide codons in between
  - Hypothesis for correct reading frame from which to translate the DNA into protein
  - May contain introns (non-coding regions) in eukaryotes

- Coding Sequence (CDS)
  - The actual region of DNA that is translated to protein

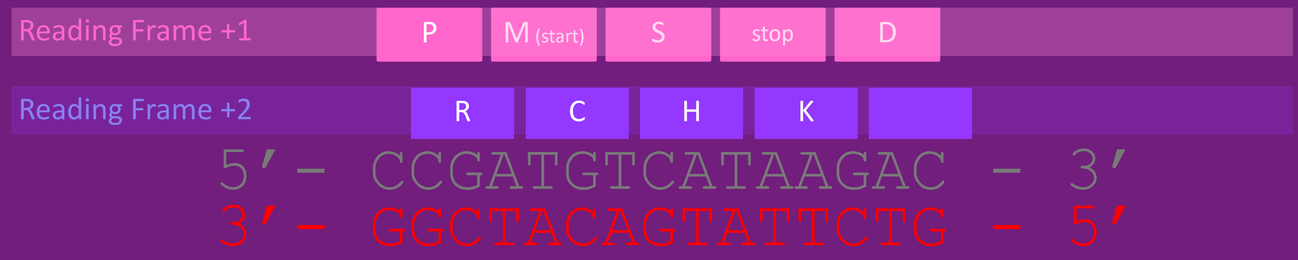
TRANSCRIPTIONS & TRANSLATION

- Reading frames



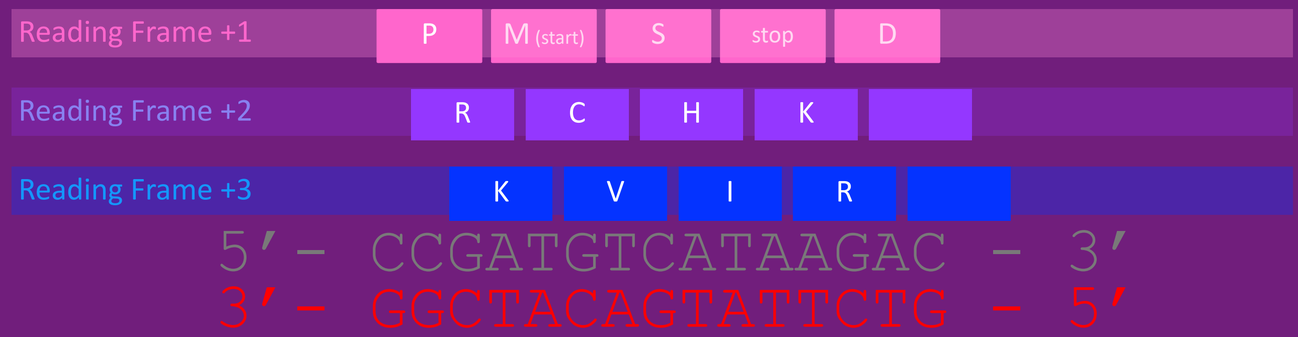
TRANSCRIPTIONS & TRANSLATION

- Reading frames



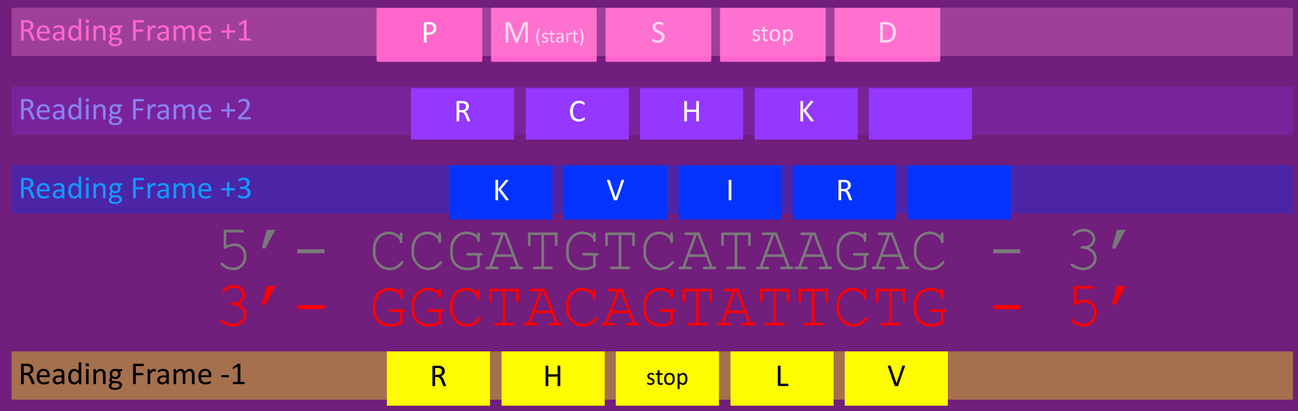
TRANSCRIPTIONS & TRANSLATION

- Reading frames



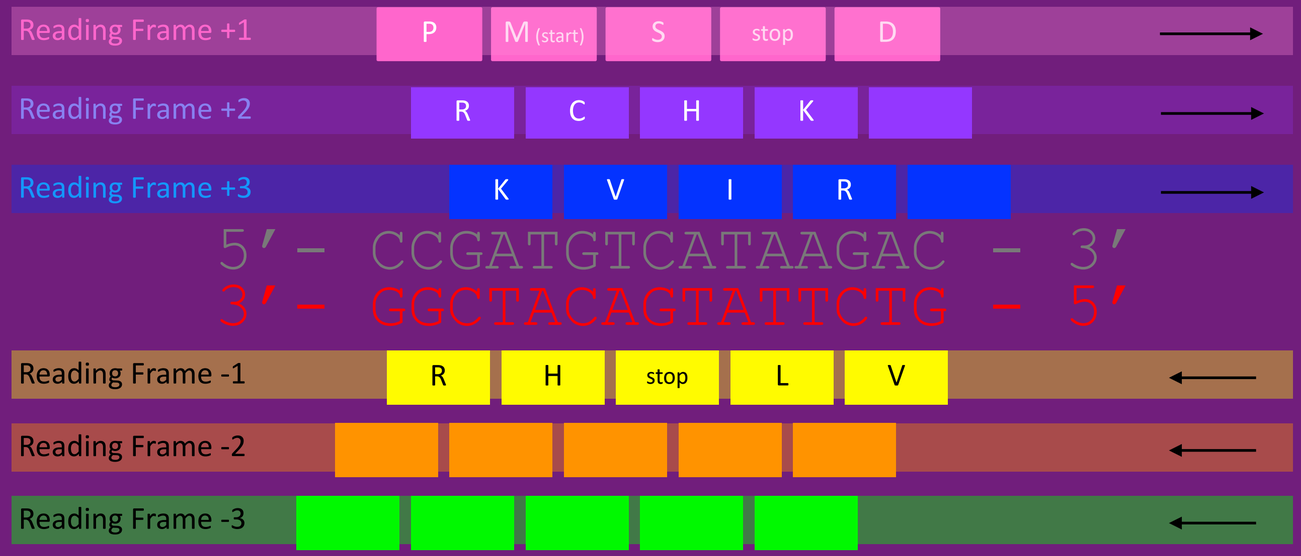
TRANSCRIPTIONS & TRANSLATION

- Reading frames



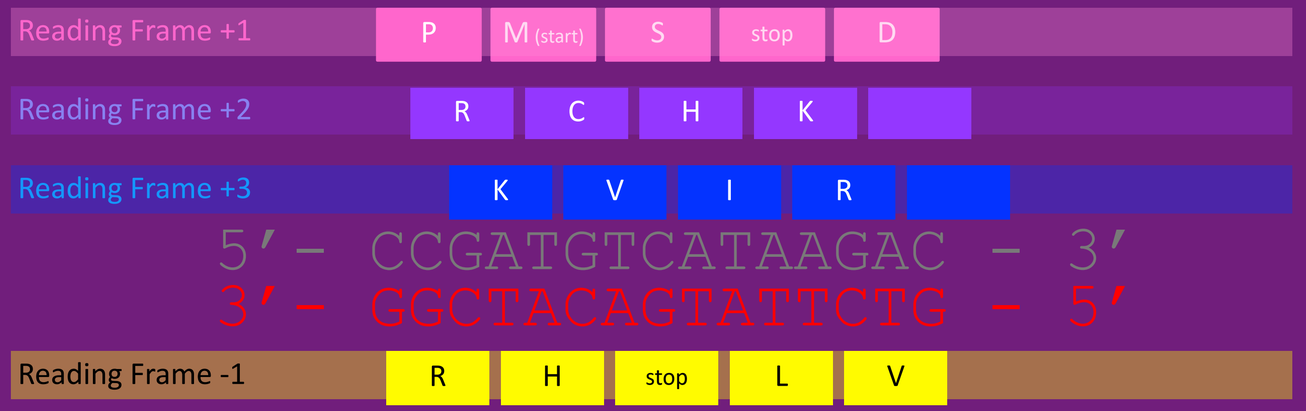
TRANSCRIPTIONS & TRANSLATION

- Reading frames



TRANSCRIPTIONS & TRANSLATION

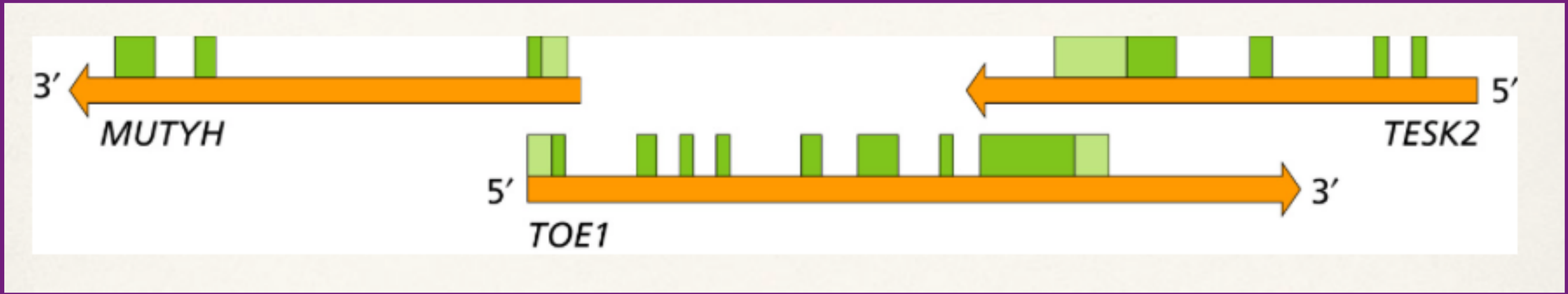
- Possible Sequences
  - PMS-stop-D
  - RCHK
  - KVIR
  - VL-stop-HR



TRANSCRIPTIONS & TRANSLATION

- Possible Sequences
  - PMS-stop-D
  - RCHK
  - KVIR
  - VL-stop-HR
- Which one is an actual coding sequence?
  - PMS-stop-D
  - Coding sequence has to have a start (M) and stop codon with at least one amino acid in between

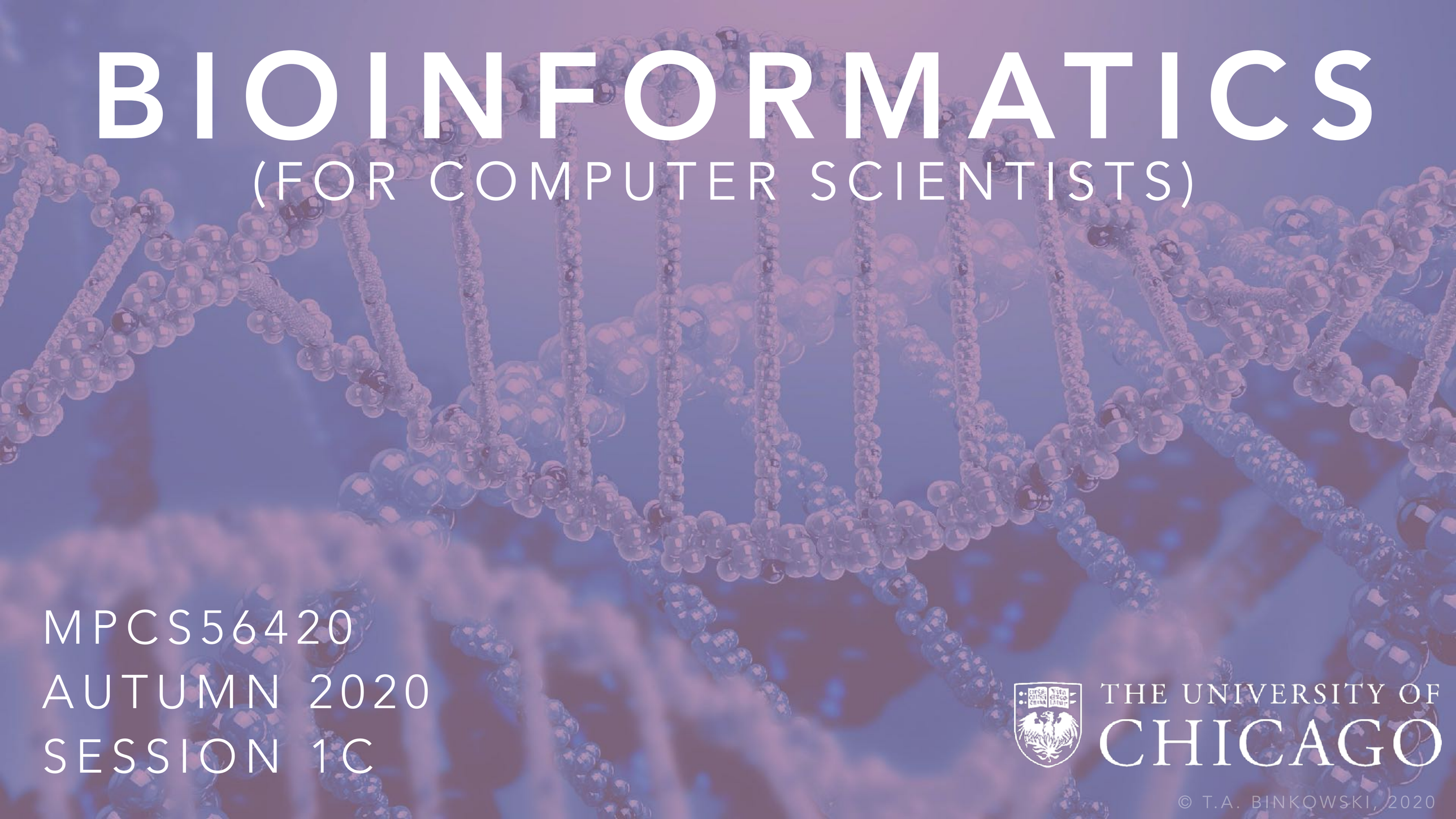
TRANSCRIPTIONS & TRANSLATION



- Common misconceptions
  - Translation always starts with the first letter of a DNA sequence
  - Translation begins at the first start codon (AUG/ATG)
  - All DNA codes for proteins
  - Genes are found only on one of the strands of DNA

BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)



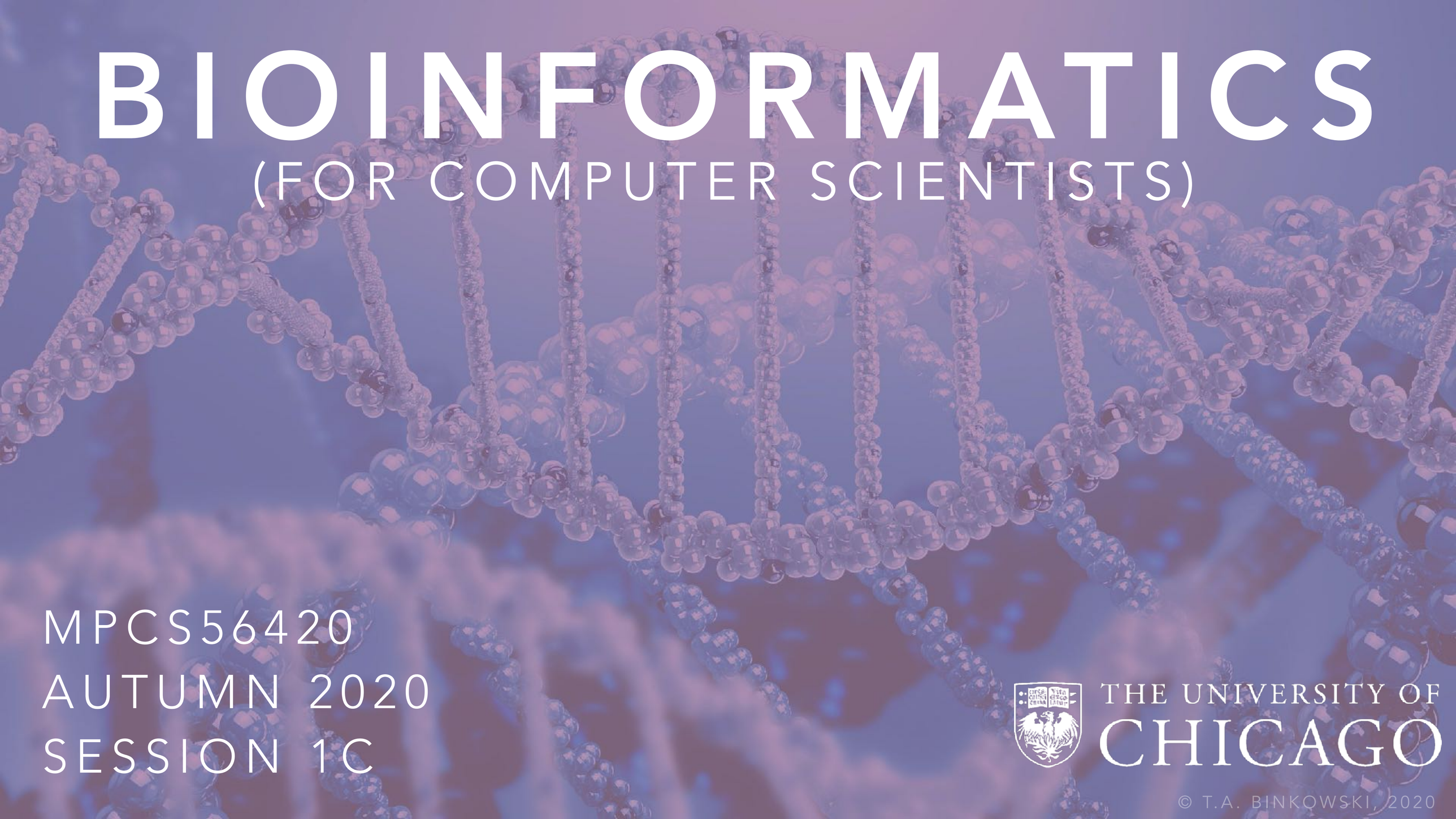
MPCS56420
AUTUMN 2020
SESSION 1C

THE UNIVERSITY OF CHICAGO

SESSION 1C

BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)



MPCS56420
AUTUMN 2020
SESSION 1C

THE UNIVERSITY OF CHICAGO


ASSIGNMENT

IPYTHON QUICK START

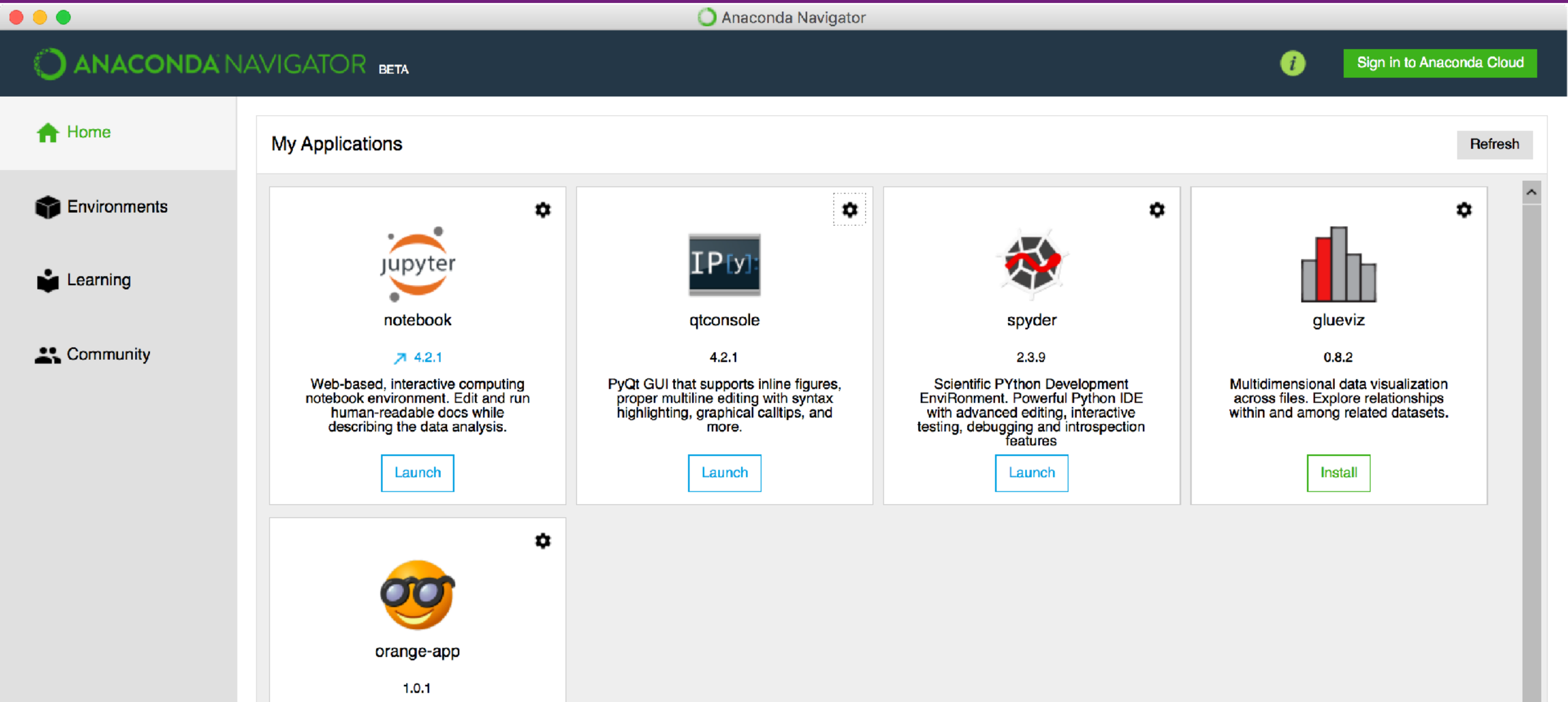
jupyter

Installing Jupyter Notebook

Contents

This information explains how to install the Jupyter Notebook and the IPython kernel.

IPYTHON QUICK START

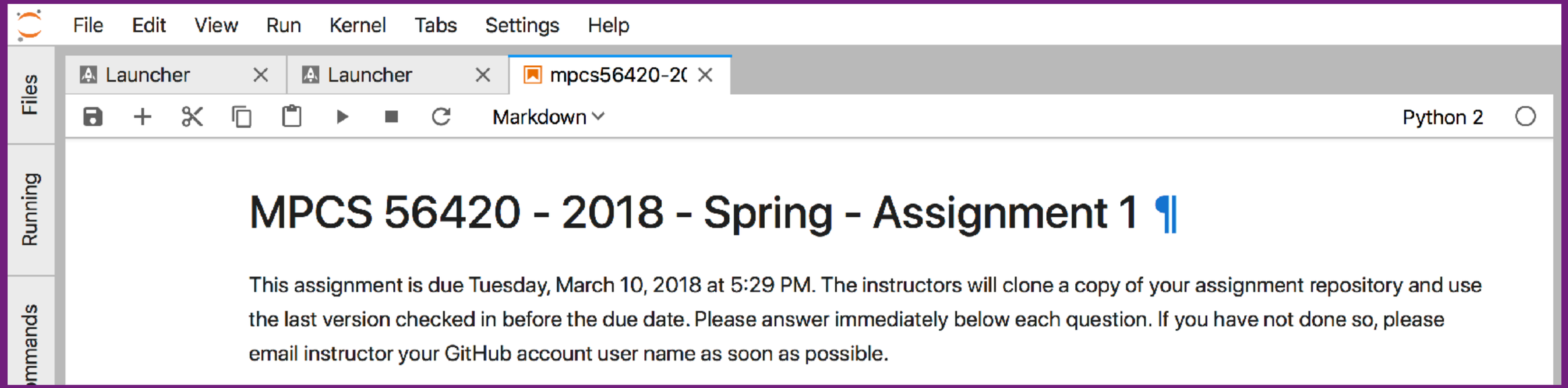


IPYTHON QUICK START

tbinkowski — jupyter_mac.command — python ◂ -bash — 127×22

```
ski@TABinkosMBP2013 ~
Users/tbinkowski/anaconda/bin/jupyter_mac.command ; exit;
8:00.089 NotebookApp] Unrecognized JSON config file version, assuming version 1
8:00.479 NotebookApp] [nb_conda_kernels] enabled, 1 kernels found
8:00.895 NotebookApp] ✓ nbpresent HTML export ENABLED
8:00.895 NotebookApp] ✗ nbpresent PDF export DISABLED: No module named nbbrowserpdf.exporters.pdf
8:00.900 NotebookApp] [nb_conda] enabled
8:00.987 NotebookApp] [nb_anacondacloud] enabled
8:00.999 NotebookApp] Serving notebooks from local directory: /Users/tbinkowski
8:00.999 NotebookApp] 0 active kernels
8:00.999 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/
8:00.999 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
pt: OpenScripting.framework - scripting addition "/Library/ScriptingAdditions/Adobe Unit Types.osax" cannot be us
rrent OS because it has no OSAXHandlers entry in its Info.plist.
8:01.901 NotebookApp] 404 GET /apple-touch-icon-precomposed.png (::1) 9.12ms referer=None
8:01.924 NotebookApp] 404 GET /apple-touch-icon.png (::1) 1.48ms referer=None
```

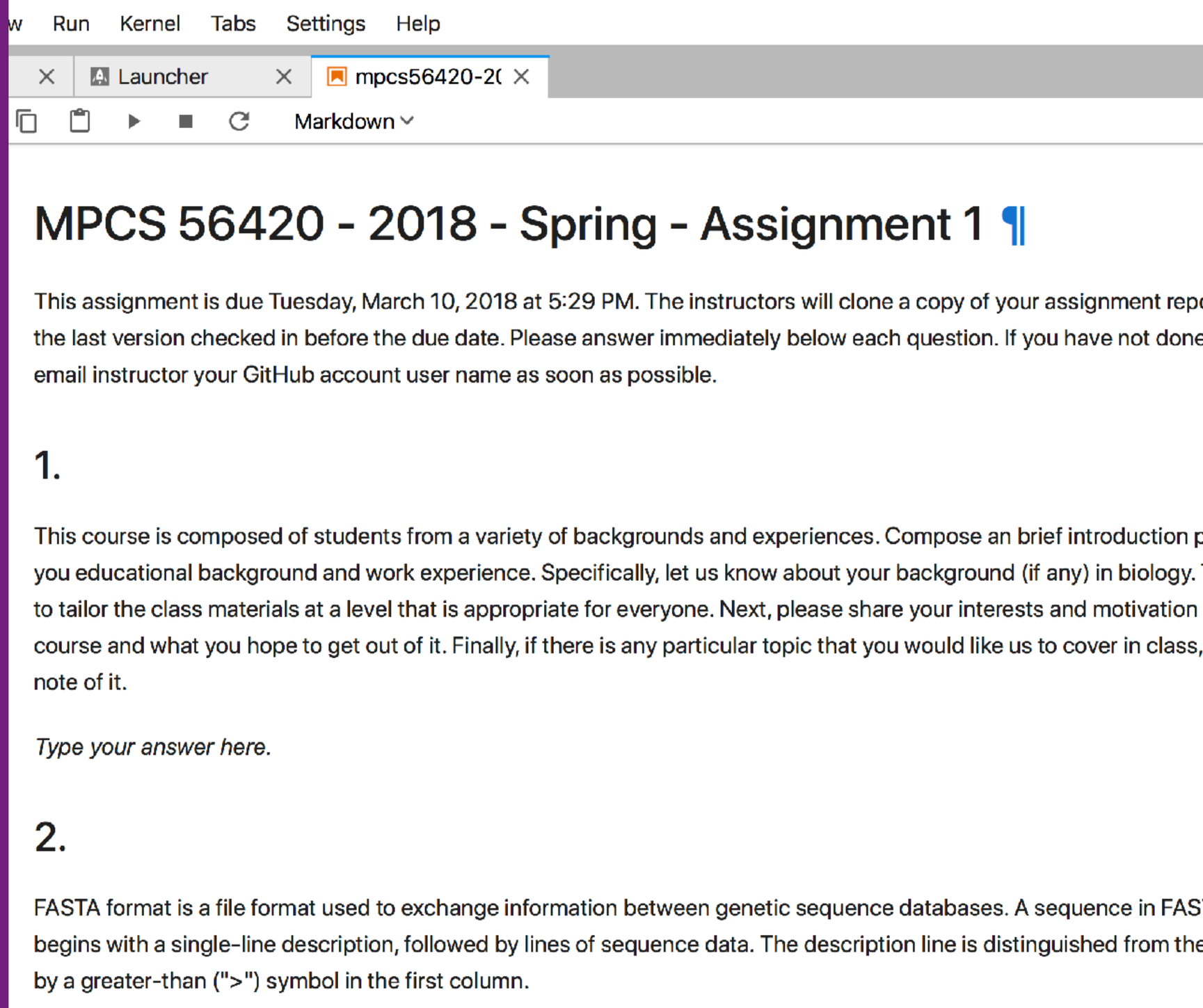
IPYTHON QUICK START



- Launches a web server that will show all notebooks in that directory
- Click on on to launch

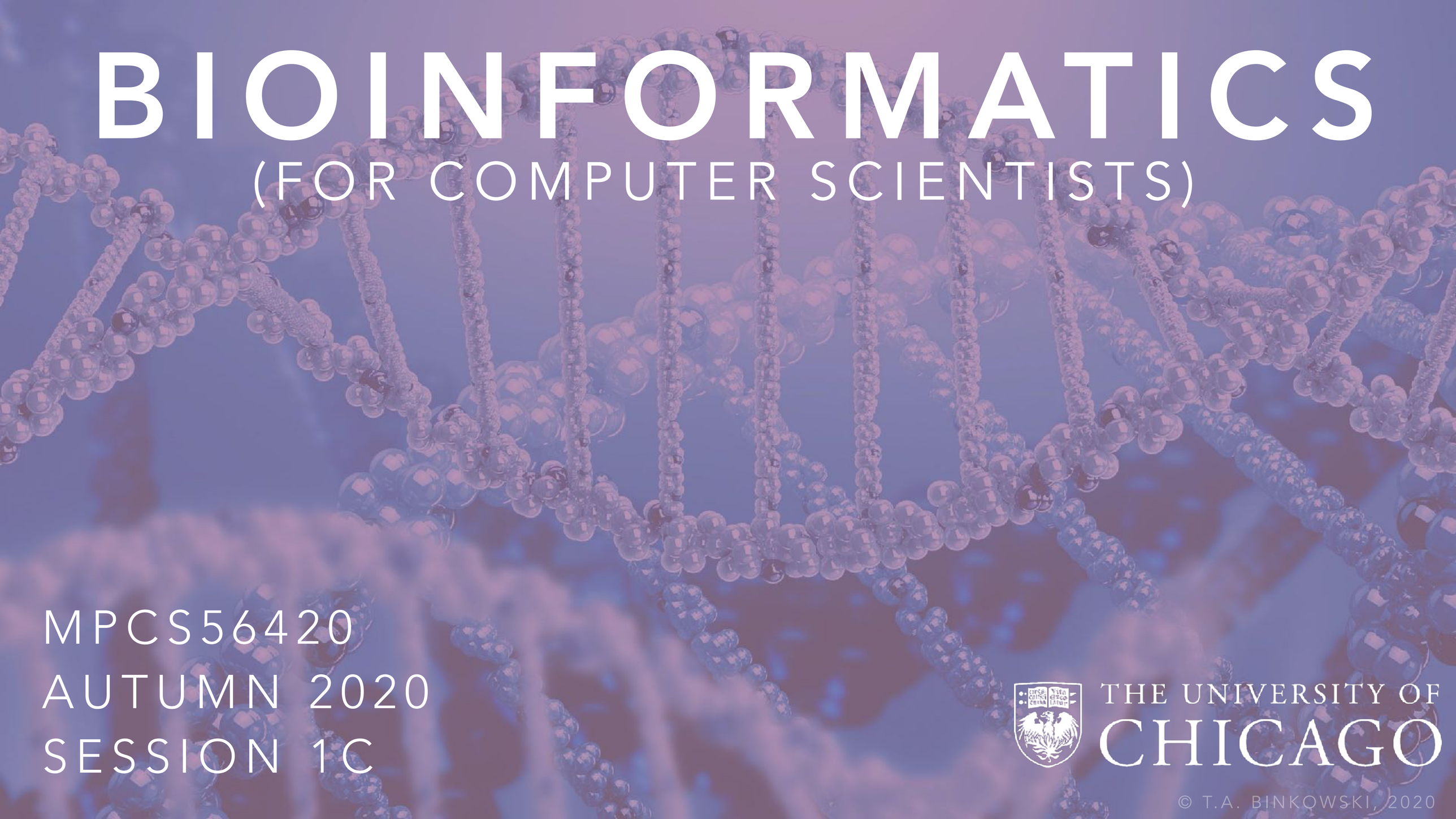
IPYTHON QUICK START

- Double click on cell to edit
- Shift+Return will execute cell



BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)



MPCS56420
AUTUMN 2020
SESSION 1C

THE UNIVERSITY OF CHICAGO