

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/slides#)

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

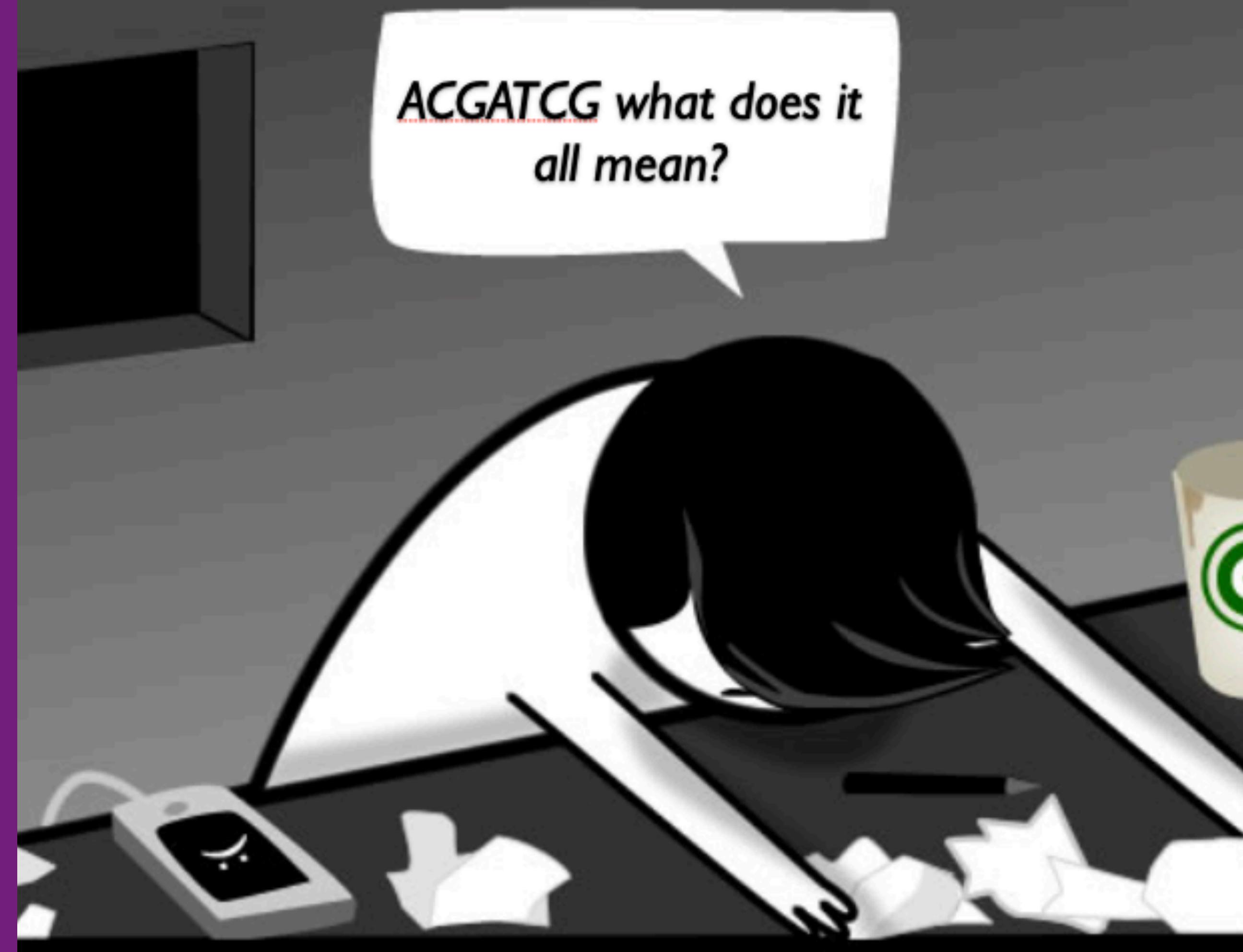
COURSE LOGISTICS

- 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

6 hours later



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

- 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

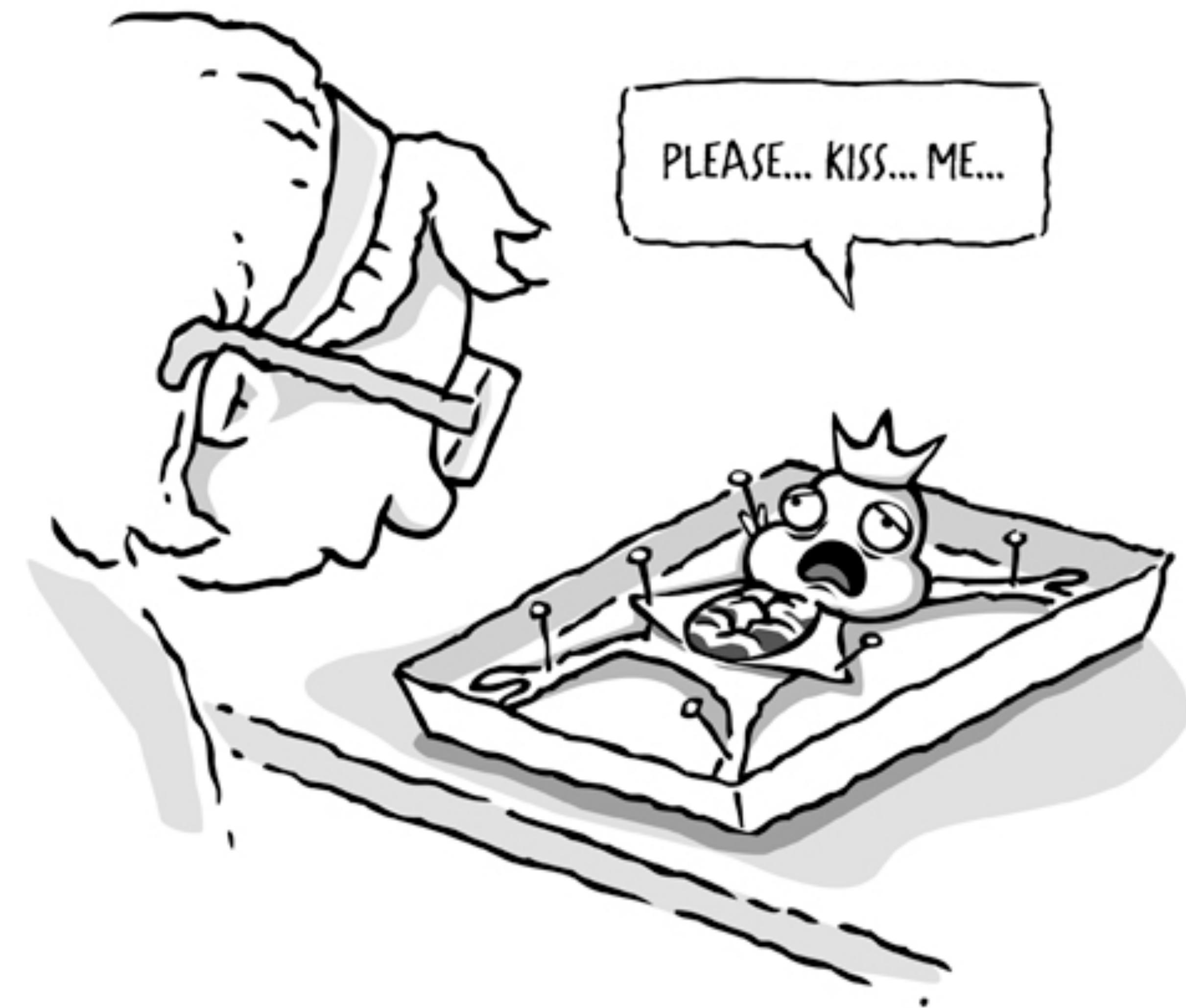
A BIOLOGIST
TURNED
COMPUTER
SCIENTIST

COURSE DESIGN

COURSE DESIGN

CHALLENGES

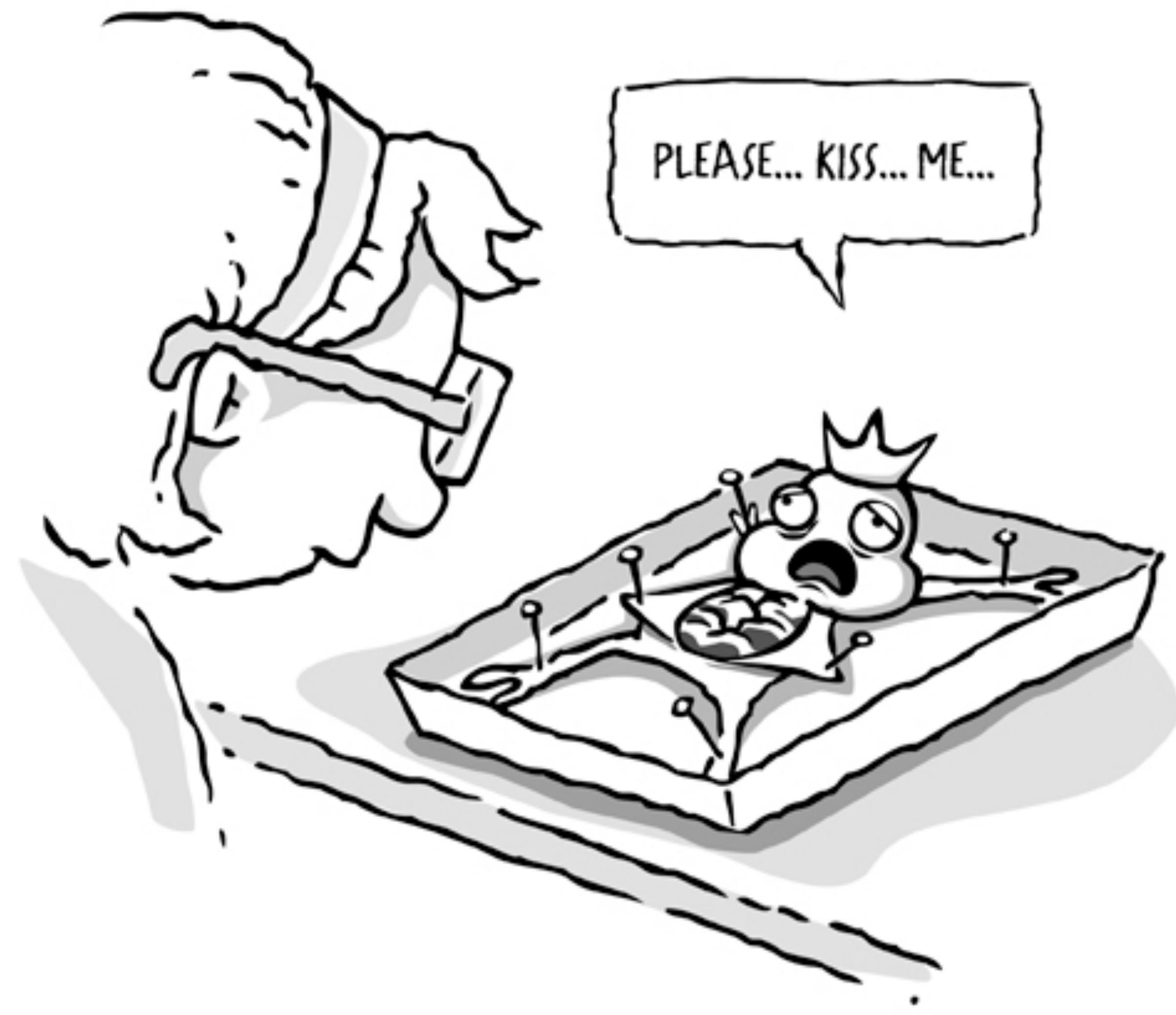
- People from different backgrounds
 - No biology since high school



COURSE DESIGN

CHALLENGES

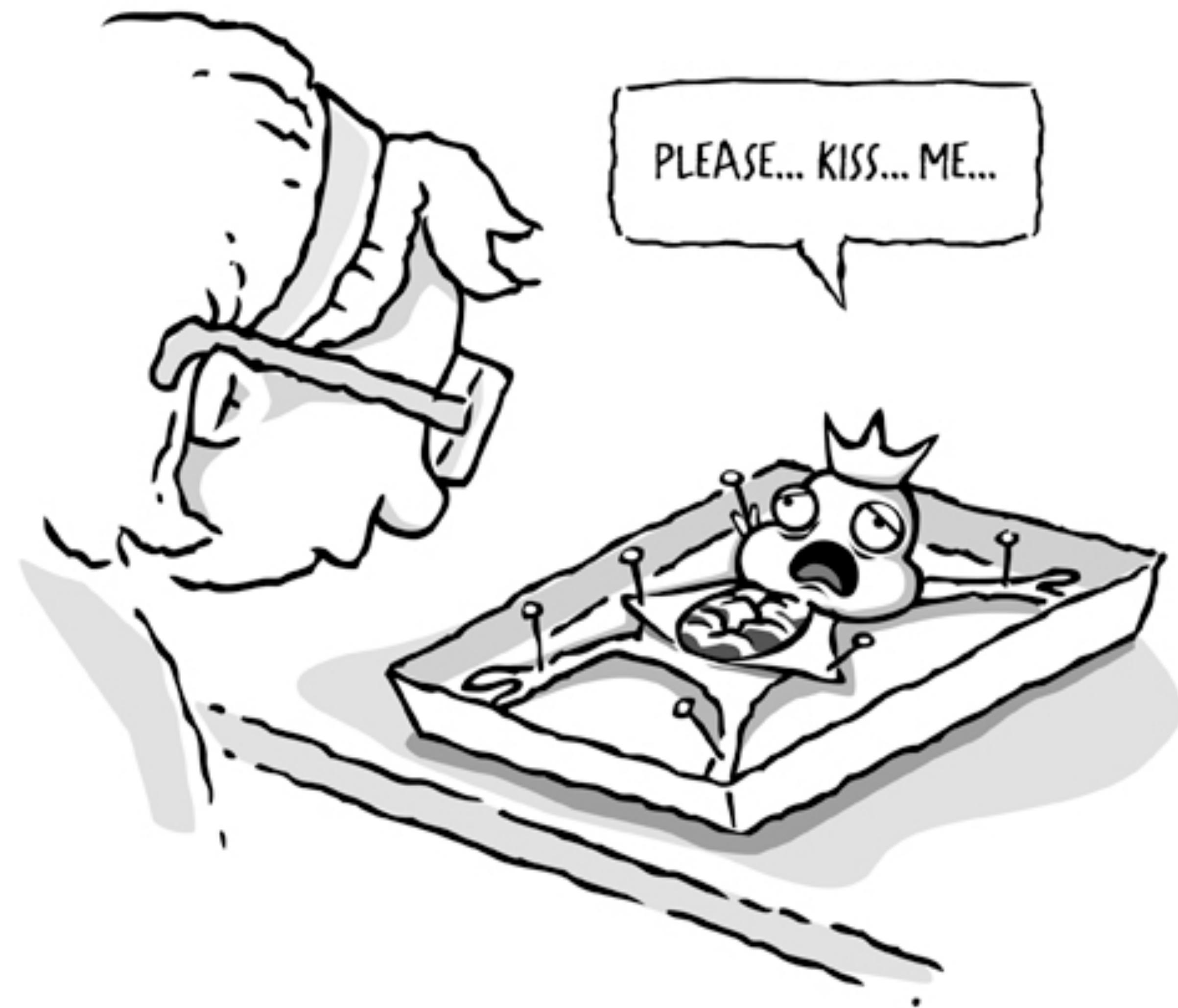
- Wide topic that touches many fields and sub-fields
 - Molecular biology, structural biology, genomics, medicine



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
 - StackOverflow, 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

```
proteinworks — abinkows@miraclac1:~ more — 127x77
abinkows@miraclac1:~ bash

import os
import sys
import math

def sort_by_atom_number(txt):
    lines = {}
    new_lines = []
    for line in txt.splitlines():
        if line.startswith("ATOM"):
            atom_number = int(line[7:11])
            lines[atom_number] = line

    for line in sorted(lines):
        new_lines.append(lines[line])
    return '\n'.join(new_lines)

#-
def strip_lines(pdb_txt, tag_func):
    new_lines = []
    for line in pdb_txt.splitlines():
        if tag_func(line):
            continue
        new_lines.append(line)
    return '\n'.join(new_lines)

#-
def strip_pdb_extension(filename):
    return os.path.splitext(os.path.basename(filename))[0]

#-
def atomic_distance(atom1_xyz, atom2_xyz):
    """ atom1_xyz is a list [x,y,z] coordinate """
    return math.sqrt((atom1_xyz[0]-atom2_xyz[0])**2-
                    (atom1_xyz[1]-atom2_xyz[1])**2-
                    (atom1_xyz[2]-atom2_xyz[2])**2)

#-
def xyz_from_pdbleline(line):
    x = float(line[30:38])
    y = float(line[38:46])
    z = float(line[46:54])
    return [x,y,z]

#-
def extract_atom_neighbors(pdb_txt, ligand_xyz, cutoff):
    new_lines = []
    for line in pdb_txt.splitlines():
        if line.startswith("ATOM"):
            protein = xyz_from_pdbleline(line)
            for ligand_atom in ligand_xyz:
                dist = atomic_distance(protein, ligand_atom)
                if dist < cutoff:
                    #print "%s - %f" % (line, dist)
                    new_lines.append(line)
    return new_lines

#-
def extract_ligand_coordinates(pdb_txt, ligand_key):
    """ Return an [x,y,z] list of the coords of a given ligand
    TODO: Optionally print to file?
    """
    print "## Extracting coordinates for ligand key "
    print ligand_key

    coords = []
    for line in pdb_txt.splitlines():
        if line.startswith("HETATM"):
            res_type = (line[17:20]).strip()
            chain_id = line[21]
            res_num = int(line[22:26])
            #print ligand_key
            coords.append([res_type, chain_id, res_num])

    return coords
```

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)

```
import os
import sys
import math

def sort_by_atom_number(txt):
    lines = {}
    new_lines = []
    for line in txt.splitlines():
        if line.startswith("ATOM"):
            atom_number = int(line[7:11])
            lines[atom_number] = line

    for line in sorted(lines):
        new_lines.append(lines[line])
    return '\n'.join(new_lines)

#-----
def strip_lines(pdb_txt, tag_func):
    new_lines = []
    for line in pdb_txt.splitlines():
        if tag_func(line):
            continue
        new_lines.append(line)
    return '\n'.join(new_lines)

#-----
def strip_pdb_extension(filename):
    return os.path.splitext(os.path.basename(filename))[0]

#-----
def atomic_distance(atom1_xyz, atom2_xyz):
    """ atom1_xyz is a list [x,y,z] coordinate """
    return math.sqrt((atom1_xyz[0]-atom2_xyz[0])**2+
                     (atom1_xyz[1]-atom2_xyz[1])**2+
                     (atom1_xyz[2]-atom2_xyz[2])**2)

#-----
def xyz_from_pdbleline(line):
    x = float(line[30:38])
    y = float(line[38:46])
    z = float(line[46:54])
    return [x,y,z]

#-----
def extract_atom_neighbors(pdb_txt,ligand_xyz,cutoff):
    new_lines = []
    for line in pdb_txt.splitlines():
        if line.startswith("ATOM"):
            protein_xyz = xyz_from_pdbleline(line)
            for ligand_atom in ligand_xyz:
                dist = atomic_distance(protein_xyz,ligand_atom)
                if dist <= cutoff:
                    new_lines.append(line)
    return new_lines

#-----
def extract_ligand_coordinates(pdb_txt,ligand_key):
    """ Return an [(x,y,z)] list of the coords of a given ligand
        TODO: Optionally print to file?
    """
    print "### Extracting coordinates for ligand key "
    print ligand_key

    coords = []
    for line in pdb_txt.splitlines():
        if line.startswith("HETATM"):
            res_type = (line[17:20]).strip()
            chain_id = line[21]
```

MORE IMPORTANT
THAN YOU MIGHT
THINK

COURSE TECHNOLOGIES

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

DRUDGE REPORT 2014® Hacker News Google News Screen Time Analytics Journals GTasks GrabLinks

PLOS Collect... Untitled 1 Introduction t... uchicago-link... PLOS Collect... Online Resou... MPCS 56420

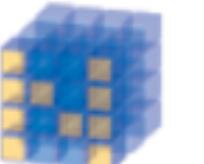
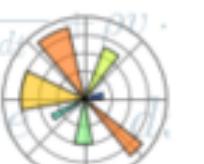
SciPy.org

Sponsored By ENTHOUGHT

Install Getting Started Documentation Report Bugs Blogs

SciPy (pronounced "Sigh Pie") is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:

| | |
|---|---|
|  NumPy Base N-dimensional array package |  SciPy library Fundamental library for scientific computing |
|  Matplotlib Comprehensive 2D Plotting |  IP[y]: IPython Enhanced Interactive Console |
|  Sympy Symbolic mathematics |  pandas Data structures & analysis |

[More information...](#)

CORE PACKAGES:

Numpy ▾ SciPy library ▾ Matplotlib ▾ IPython ▾ Sympy ▾ Pandas ▾

News

NumPy 1.9.0 released See [Obtaining NumPy & SciPy libraries.](#)
(2014-09-07)

NumPy 1.8.2 released See [Obtaining NumPy & SciPy libraries.](#)
(2014-08-09)

SciPy 0.14.0 released See [Obtaining NumPy & SciPy libraries.](#)
(2014-05-03)

NumPy 1.8.1 released See [Obtaining NumPy & SciPy libraries.](#)
(2014-03-26)

Search Go

COURSE TECHNOLOGIES

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

ipython.org

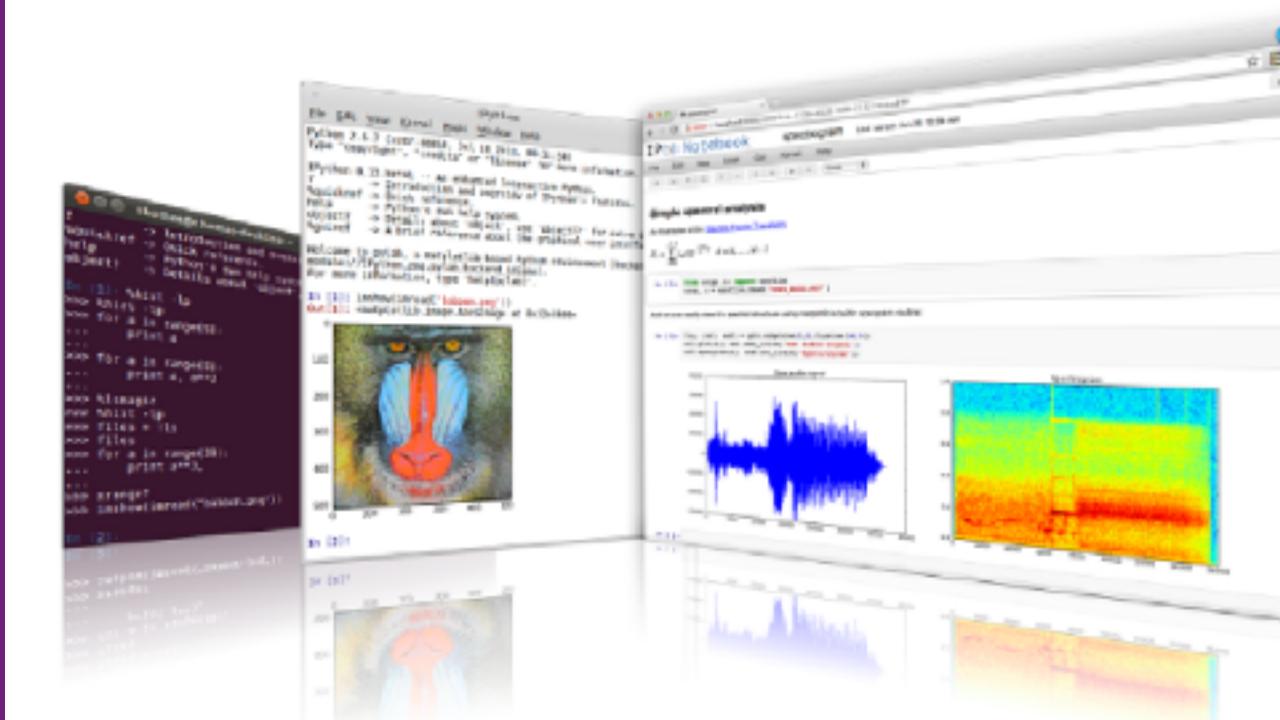
IP[y]: IPython

Interactive Computing

[Install](#) · [Documentation](#) · [Project](#) · [Jupyter](#) · [News](#) · [Cite](#) · [Donate](#)

IPython provides a rich architecture for interactive computing with:

- A powerful interactive shell.
- A kernel for [Jupyter](#).
- Support for interactive data visualization and use of [GUI toolkits](#).
- Flexible, [embeddable](#) interpreters to load into your own projects.
- Easy to use, high performance tools for [parallel computing](#).



To get started with IPython in the Jupyter Notebook, see our [official example collection](#). Our [notebook gallery](#) is an excellent way to see the many things you can do with IPython while learning about a variety of topics, from basic programming to advanced statistics or quantum mechanics.

To learn more about IPython, you can watch our videos and screencasts, download our [talks and presentations](#), or read our [extensive documentation](#). IPython is open source (BSD license), and used by a range of [other projects](#); add your project to that list if it uses IPython as a library, and please don't forget to [cite the project](#).

IPython supports Python 2.7 and 3.3 or newer. Our older 1.x series supports Python 2.6 and 3.1.

Jupyter and the future of IPython

Display a menu

COURSE TECHNOLOGIES

GOOGLE
COLLABORATORY

The screenshot shows a Google Collaboratory interface. At the top, there's a purple header with the title "COURSE TECHNOLOGIES". Below it is a yellow speech bubble containing the text "GOOGLE COLLABORATORY". The main area is a notebook titled "mpcs56420-2018-spring-assignment-1.ipynb". The notebook menu bar includes File, Edit, View, Insert, Runtime, Tools, Help, Comment, Share, and a profile icon. The notebook content pane shows a sidebar with "+ Code" and "+ Text" buttons, and a main area with a section titled "MPCS 56420 - 2020 - Spring - Assignment 1". This section contains a text block about assignment due dates and GitHub account user names. Below this is a section titled "Problem 1." which contains a text block about introducing oneself and sharing interests. A text input field is present for "Type your answer here." At the bottom, there's another section labeled "2.".

mpcs56420-2018-spring-assignment-1.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text

Comment Share

MPCS 56420 - 2020 - Spring - Assignment 1

This assignment is due Tuesday, March 10, 2018 at 5:29 PM. The instructors will clone a copy of your assignment repository and use the last version checked in before the due date. Please answer immediately below each question. If you have not done so, please email instructor your GitHub account user name as soon as possible.

Problem 1.

This course is composed of students from a variety of backgrounds and experiences. Compose a brief introduction providing us with your educational background and work experience. Specifically, let us know about your background (if any) in biology. This will allow us to tailor the class materials at a level that is appropriate for everyone. Next, please share your interests and motivation for taking this course and what you hope to get out of it. Finally, if there is any particular topic that you would like us to cover in class, please make a note of it.

Type your answer here.

2.

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



Page Discussion

Biopython

(Redirected from [Main Page](#))

Navigation

[Main Page](#)
[Downloads](#)
[Mailing lists](#)
[Documentation](#)
[Cookbook](#)
[News](#)
[Source Code](#)
[New issue tracker](#)
[Old issue tracker](#)
[Buildbot Tests](#)
[Participants](#)
[Script Central](#)
[Recent changes](#)
[Random page](#)

Toolbox

[What links here](#)
[Related changes](#)
[Special pages](#)
[Printable version](#)
[Permanent link](#)

Introduction

Biopython is a set of freely available tools for biological computation written in [Python](#) by an international team of developers.

It is a distributed collaborative effort to develop Python libraries and applications which address the needs of bioinformatics. The source code is made available under the [Biopython License](#), which is extremely permissive and has been adopted by many other projects around the world. We work along with the [Open Bioinformatics Foundation](#), who generously host this wiki.

This wiki will help you download and install Biopython, and start using the libraries and tools.

Get Started

- [Download Biopython](#)
- [Installation help](#) (PDF)

Get help

- [Tutorial](#) (PDF)
- [Documentation on this wiki](#)
- [Cookbook \(working examples\)](#)
- [Discuss and ask questions](#)

- [W](#)
- [D](#)
- [G](#)
- [R](#)

The latest release is [Biopython 1.65](#), released on 17 December 2014.

This page was last modified on 17 December 2014, at 21:09.

This page has been accessed 1,875,589 times.

Content is available under [GNU Free Documentation License 1.2](#).

[Privacy policy](#) [About Biopython](#) [Disclaimers](#)

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

The screenshot shows the Google Cloud Platform Compute Engine landing page. At the top, there's a navigation bar with icons for back, forward, search, and account. The URL 'cloud.google.com' is visible. Below the header, there's a large image of server racks with a blue hexagonal icon containing a white circuit board symbol overlaid. The text 'Compute Engine' is prominently displayed in large white letters. A subtitle explains: 'Run large-scale workloads on virtual machines hosted on Google's infrastructure. Choose a VM that fits your needs and gain the performance of Google's worldwide fiber network.' A 'Get Started' button is located in the bottom-left corner of the main image area. At the very bottom, there's a navigation bar with links for 'Features', 'Case Studies', 'Pricing Calculator', 'Pricing', and 'Documentation'.

Features



High-performance virtual machines

Compute Engine's Linux VMs are consistently performant, scalable, highly secure and reliable. Supported distros include Debian and CentOS. You can choose from

COURSE TECHNOLOGIES

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

The screenshot shows the homepage of the Research Computing Center (RCC) at the University of Chicago. The header features the university's crest and the text "THE UNIVERSITY OF CHICAGO" next to "Research Computing Center". The main visual is a close-up photograph of wheat stalks against a blue sky. A blue banner across the image contains the text "Computing Cereals in Parallel". Below the banner, a caption reads: "When trying to simulate the many facets of Earth's climate, a 'mugshot' of wheat can be very handy." At the bottom of the page, there are three columns with icons and text: "Enabling Research" (book icon), "Data Visualization" (bar chart icon), and "Training & Education" (two people icon). Each column also has a brief description.

THE UNIVERSITY OF
CHICAGO | Research Computing Center

Computing Cereals in Parallel

When trying to simulate the many facets of Earth's climate, a "mugshot" of wheat can be very handy.

Enabling Research

Our department has helped enable the advancement of critical inquiry from the physical sciences to the social sciences to the humanities.

Data Visualization

RCC maintains data visualization resources including high-end graphics processing hardware, visualization software, and custom remote visualization tools.

Training & Education

The Research Computing Center offers training and workshops on a variety of topics relevant to research computing.

Display a menu

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>)

ncbi.nlm.nih.gov

Home - Books - NCBI

NCBI Resources How To

Bookshelf Books Browse Titles Limits Advanced

 Bookshelf

Bookshelf provides free access to books and documents in life science and healthcare. A vital node in the data-rich resource network at NCBI, Bookshelf enables users to easily browse, retrieve, and read content, and spurs discovery of related information.

Using Bookshelf

Quick Start Guide

FAQ

Tutorials

Bookshelf News 

Copyright and Permissions

New & Updated

The Influence of Global Environmental Change on Infectious Disease Dynamics: Workshop Summary. Forum on Microbial Threats; Board on Global Health; Institute of Medicine (US). Washington (DC): National Academies Press (US); 2014.

Fitness Measures and Health Outcomes in Youth. Committee on Fitness Measures and Health Outcomes in Youth; Food and Nutrition Board; Institute of Medicine; Pate R, Oria M, Pillsbury L, editors. Washington (DC): National Academies Press (US); 2012 Dec 10.

Evaluation of the Lovell Federal Health Care Center Merger: Findings, Conclusions, and Recommendations. Committee on Evaluation of the Lovell Federal Health Care Center Merger; Board on the Health of Select Populations; Institute of Medicine. Washington (DC): National Academies Press (US); 2012 Dec 28.

Guidelines for HIV Mortality Measurement. Geneva: World Health Organization; 2014.

WHO Recommendation on Community Mobilization through Facilitated Participatory Learning and Action Cycles with Women's Groups for Maternal and Newborn Health. Geneva: World Health Organization; 2014.

See more...

Read

Browse Titles

New Releases 

PubReader

Participate

Authors and Publishers

How to Apply

Participation Agreement

Featured Titles

Fostering Independence, Participation, and Healthy Aging Through Technology: Workshop Summary. Forum on Aging, Disability, and Independence; Board on Health Sciences Policy; Division of Behavioral and Social Sciences and Education; Institute of Medicine; National Research Council. Washington (DC): National Academies Press (US); 2013 Jul 19.

The Informed Brain in a Digital World: Interdisciplinary Team Summaries. National Academies Keck Future Initiative Informed Brain Steering Committee. Washington (DC): National Academies Press (US); 2013 May 6.

Multigene Panels in Prostate Cancer Risk Assessment. Evidence Reports/Technology Assessments, No. 209. Little J, Wilson B, Carter R, et al. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 Jul.

Negotiating Bioethics: The Governance of UNESCO's Bioethics Programme. Langlois A. London and New York: Routledge; 2013.

What You Need to Know About Infectious Disease. Drexler M; Institute of Medicine (US). Washington (DC): National Academies Press (US); 2010.

More Information

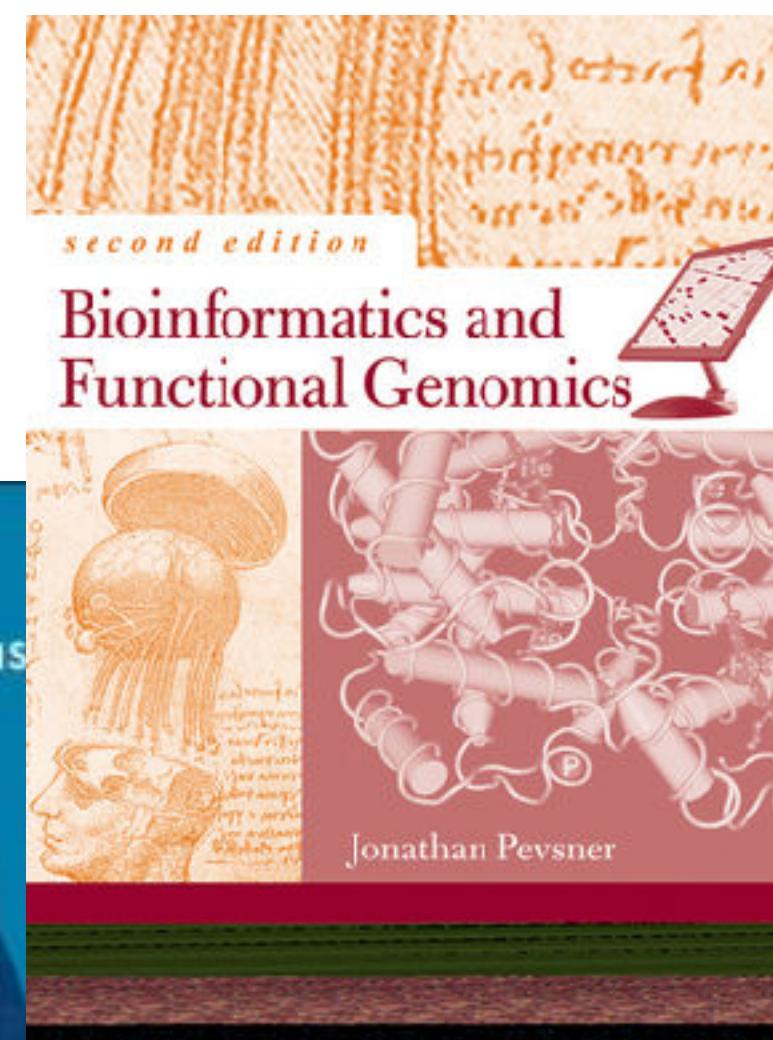
NLM Literature Archive

Open Access Subset

Librarians

COURSE RESOURCES

TEXT BOOKS



RECOMMENDED

FREE DIGITAL COPIES
AVAILABLE THROUGH
UCHICAGO LIBRARY



COURSE RESOURCES

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

2020.02

[Home](#)

[Announcements](#) 

[Syllabus](#)

[Modules](#) 

[Assignments](#) 

[Discussions](#)

[Library Reserves](#)

[People](#)

[Grades](#)

[Panopto Video](#)

[Purchase UChicago Bookstore Course Materials](#)

[Collaborations](#) 

[Conferences](#) 

[Files](#) 

[Syllabus](#) 

Recent Announcements

MPCS 56420 1 (Spring 2020) Bioinformatics for Computer Scientists

This is a test.

Course Summary:

Date

Details

Course Status

 Unpublished

 Publish

 Import from Commons

 Choose Home Page

 View Course Stream

COURSE RESOURCES

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

The screenshot shows a GitHub repository page. At the top, there's a search bar with the placeholder "Search or jump to...". Below the search bar are navigation links: Pull requests, Issues, Marketplace, and Explore. The repository name is "uchicago-mobi/mpcs51032-2020-spring". Underneath the repository name, there are links for Code, Issues (0), Pull requests (0), Actions, Projects (0), Wiki, and Security. A note says "No description, website, or topics provided." Below this, there's a "Topics" section with a link to "Edit". Statistics show 2 commits, 1 branch, 0 packages, and 0 issues. A "Create new file" button is visible. The main area shows a commit titled "Add session 1 materials" from "mbinks" with a timestamp of "1 day ago". A yellow callout box points to this commit with the text "Materials for each session (slides, videos, projects, etc.)". At the bottom of the page, there's a footer with links to GitHub, Inc., Terms, Privacy, Security, Status, Help, and a GitHub logo.

No description, website, or topics provided.

Topics

2 commits 1 branch 0 packages 0 issues

Branch: master New pull request

Add session 1 materials

Add session 1 materials

Materials for each session (slides, videos, projects, etc.)

© 2020 GitHub, Inc. Terms Privacy Security Status Help

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

GitHub Classroom

MPCS56420 - Bioinformatics (for compu

uchicago-bio

"mpcs56420-2018-spring-assignment-1" has been created!



mpcs56420-2018-spring-assignment-1

Individual assignment

Give this to your students

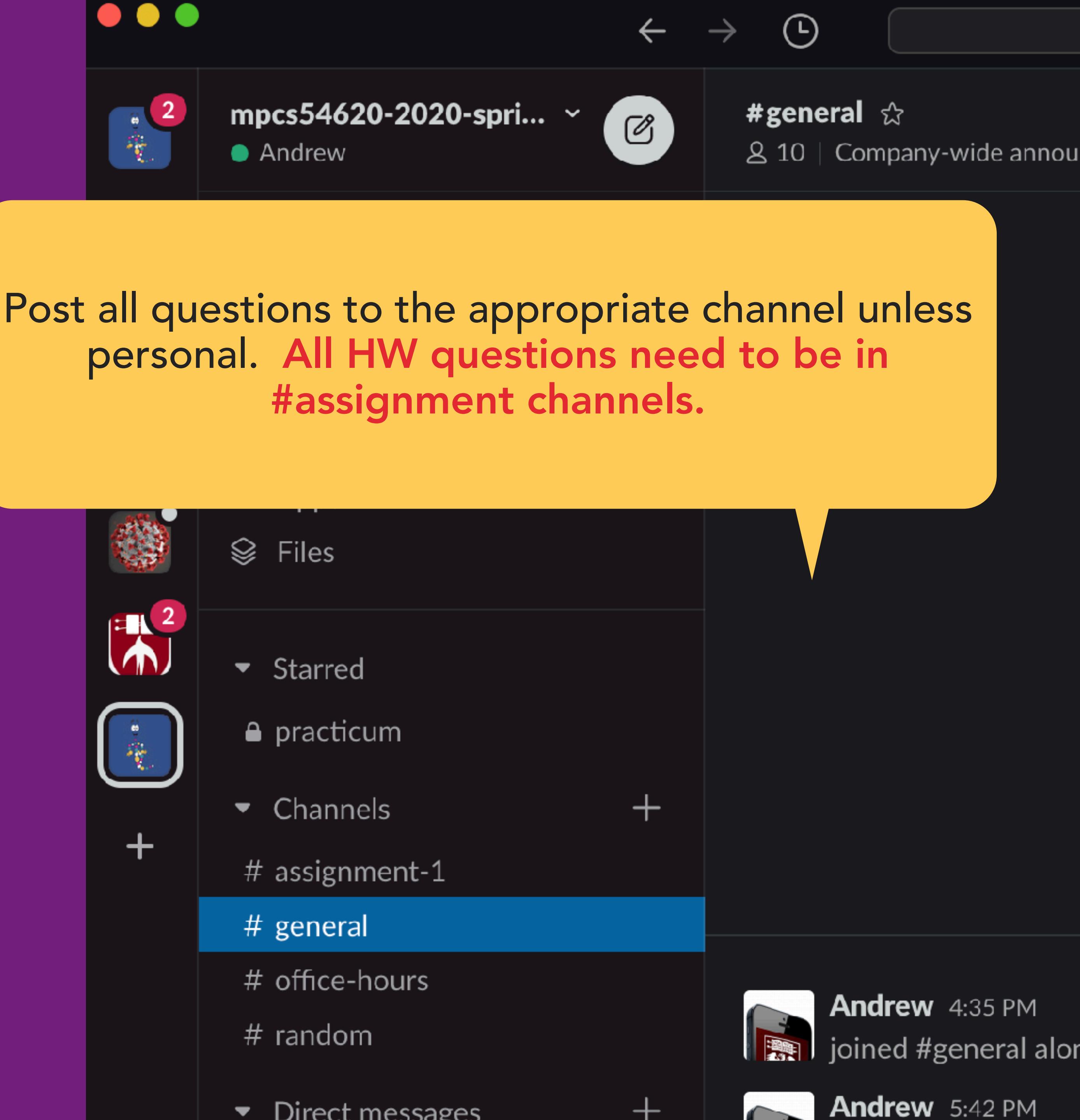
<https://classroom.github.com/a/vZ-nWORD>

"mpcs56420-2018-spring-assignment-1" does not exist

Share the invitation link with your students to get started

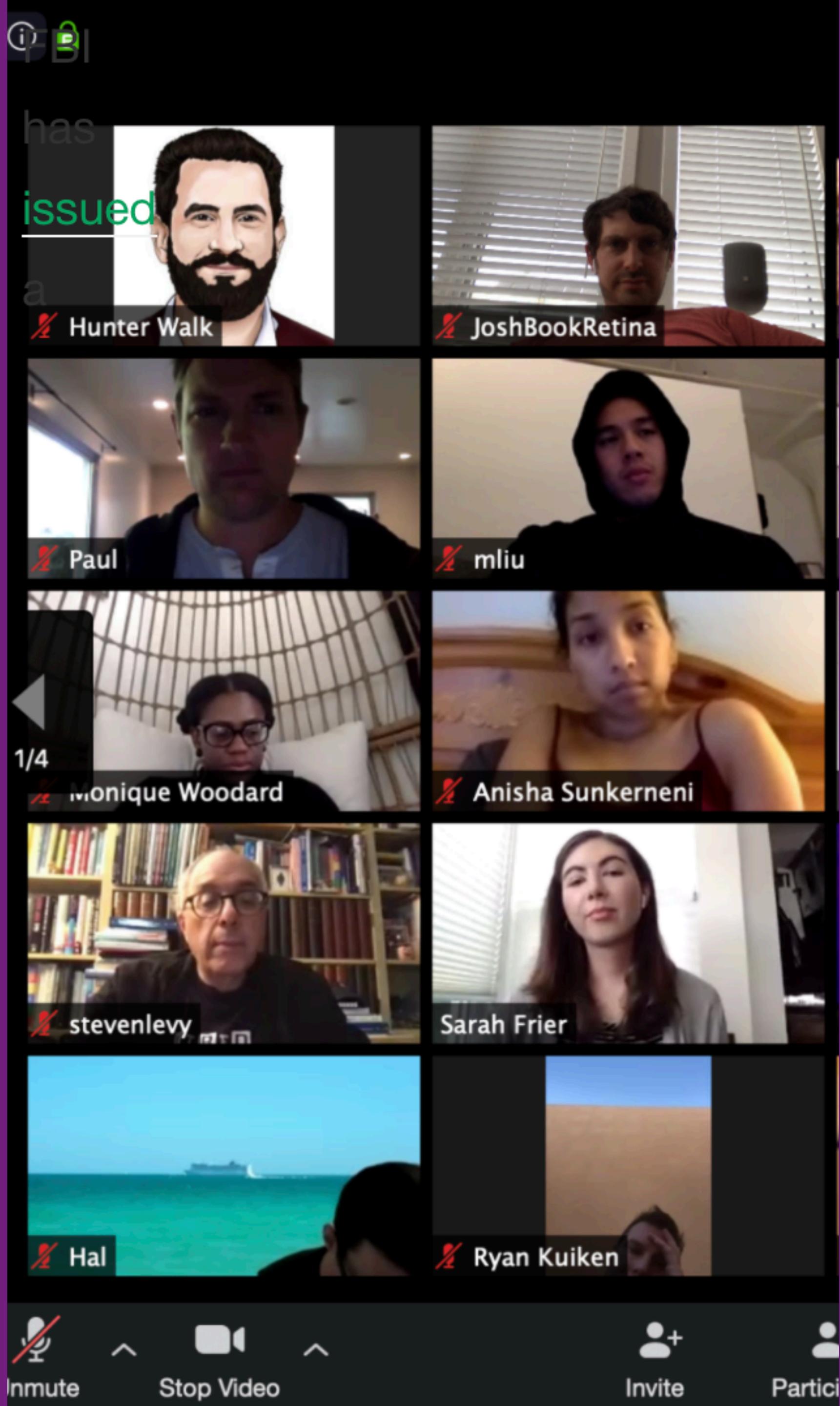
COURSE LOGISTICS

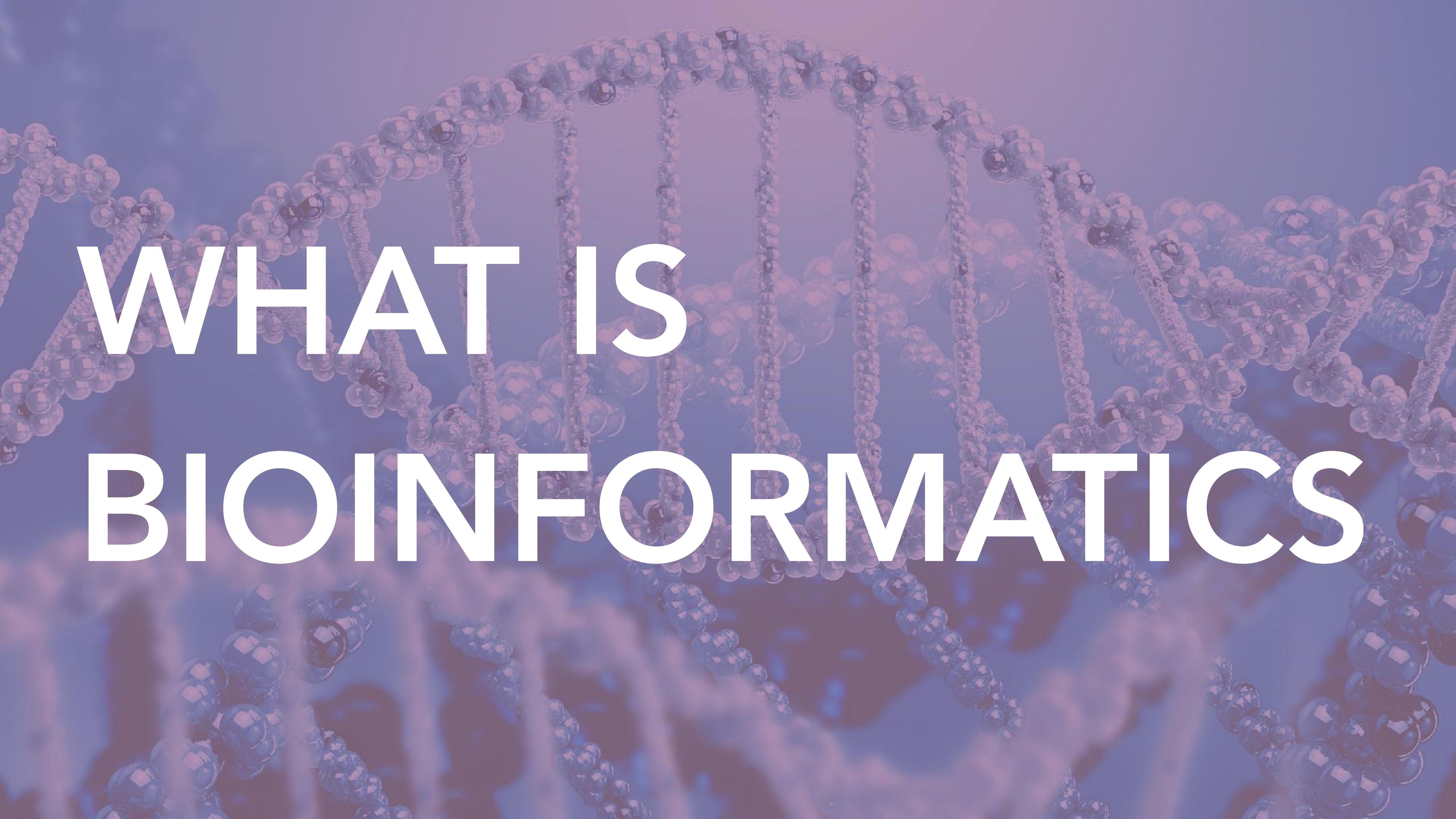
- Discussion forum rules
 - Anyone can answer a question
 - Post sources (if available) when answering questions



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

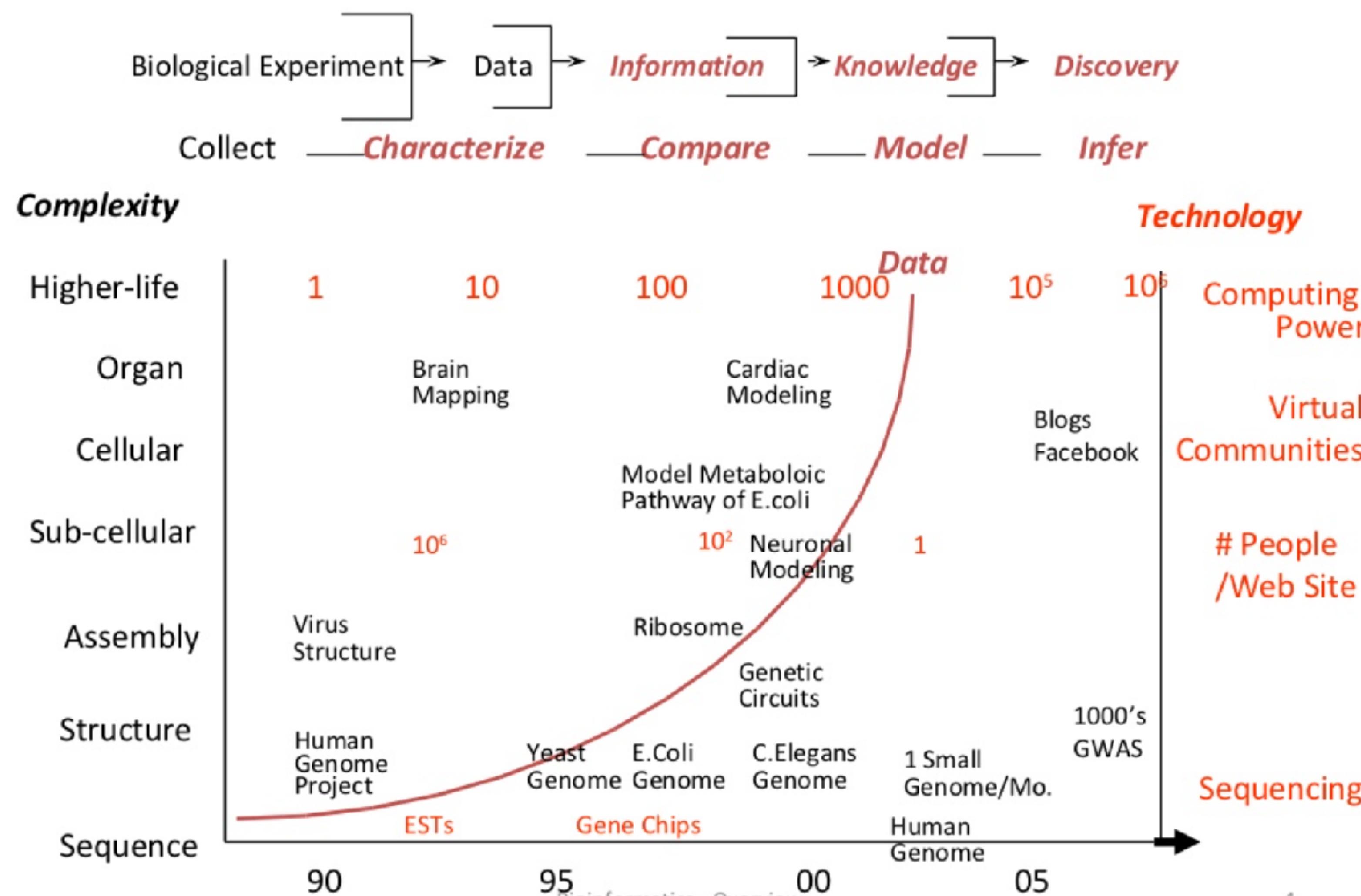
THREE IMPORTANT SUB-DISCIPLINES

- 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?

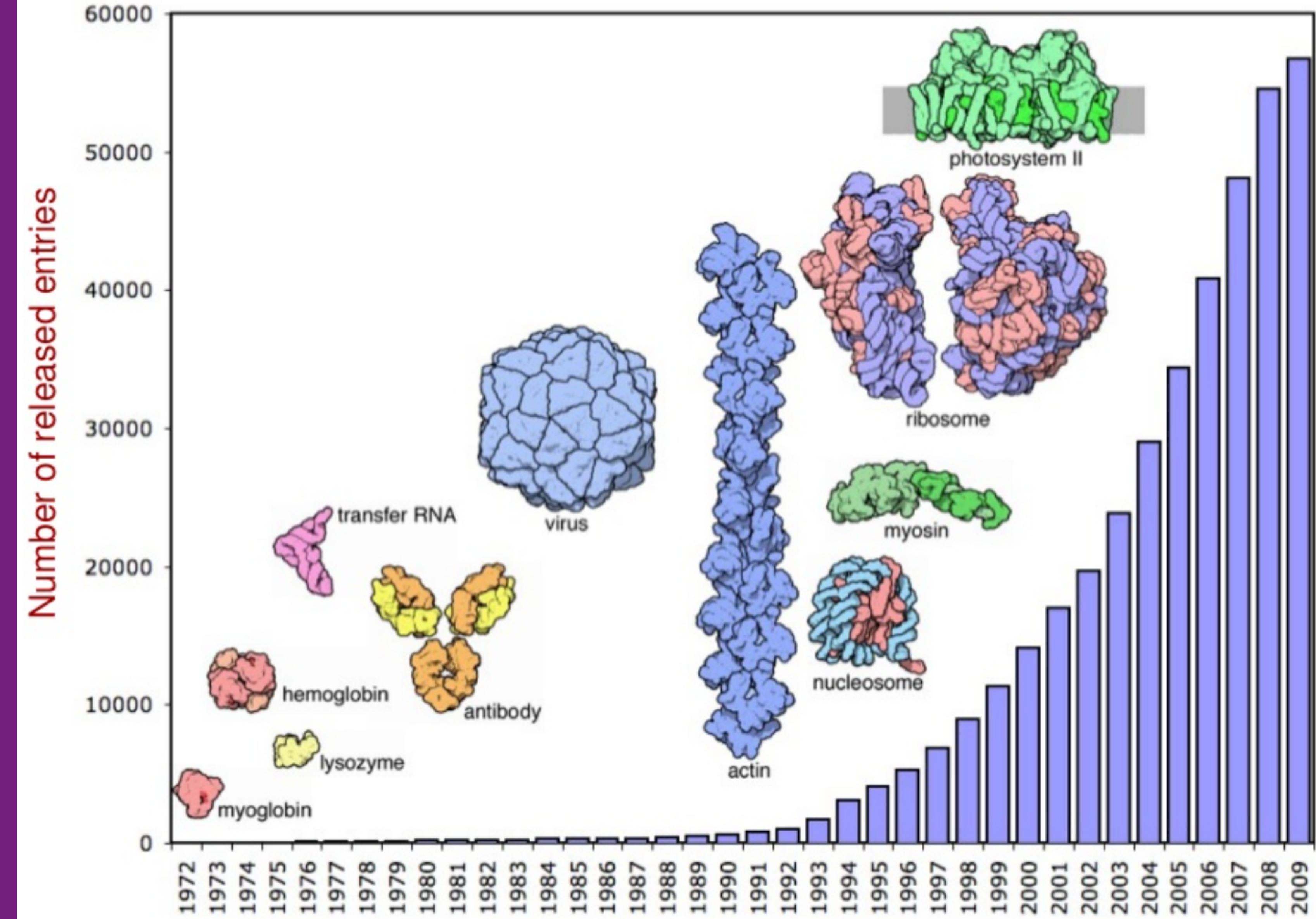
PHILIP BOURNE,
UCSD

Bioinformatics In One Slide



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

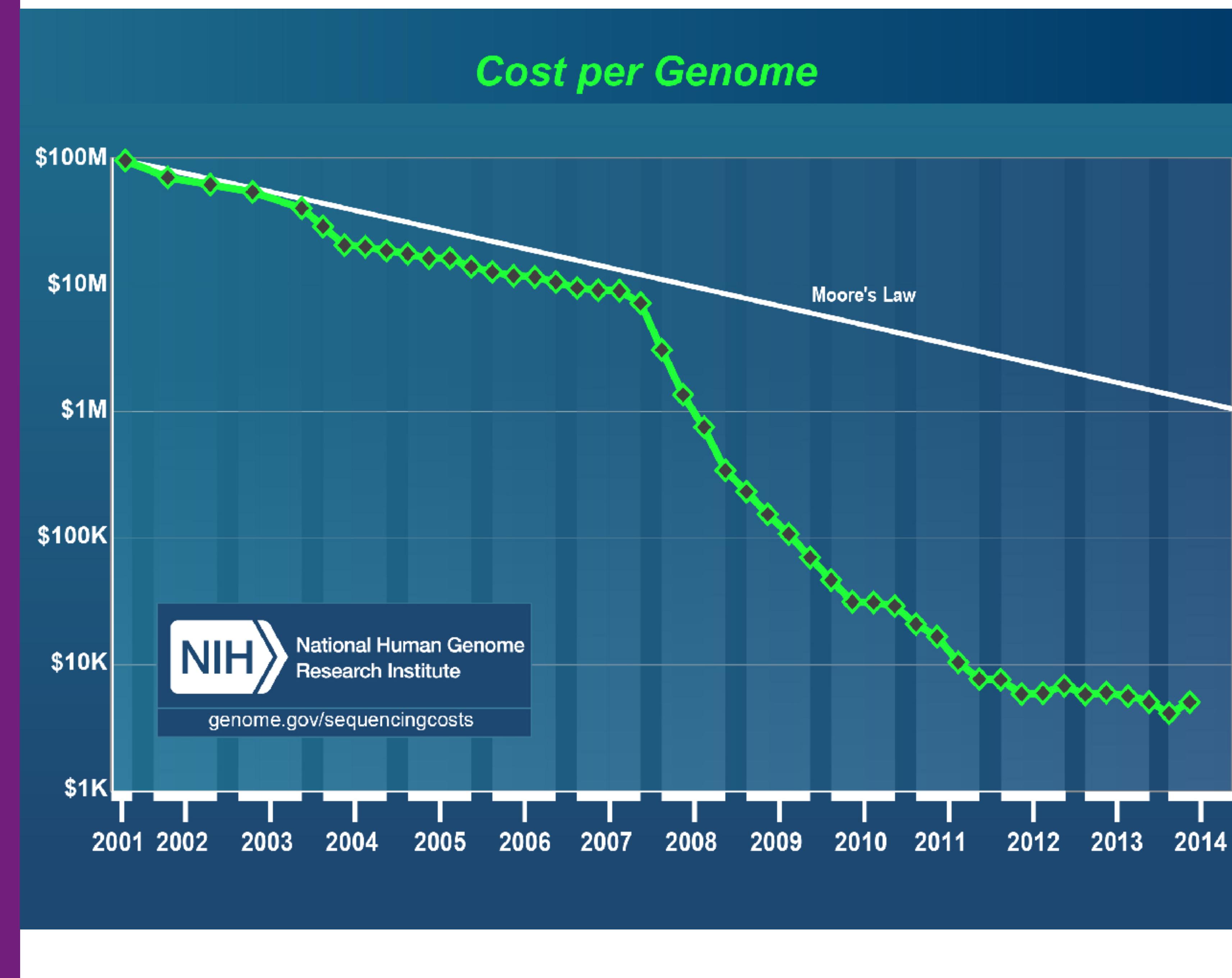
BANKING USES
COMPUTERS BUT ITS
NOT CALLED
BANKFORMATICS



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

WHAT IS BIOINFORMATI CS

- 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



Police confirm waiter spat in customer's soda using DNA testing

by Reuters Videos 0:47 mins

A disgruntled customer who found spit in his drink after visiting a Chili's restaurant in Clay, New York, is suing the company that owns Chili's, the franchise owner and a former waiter.

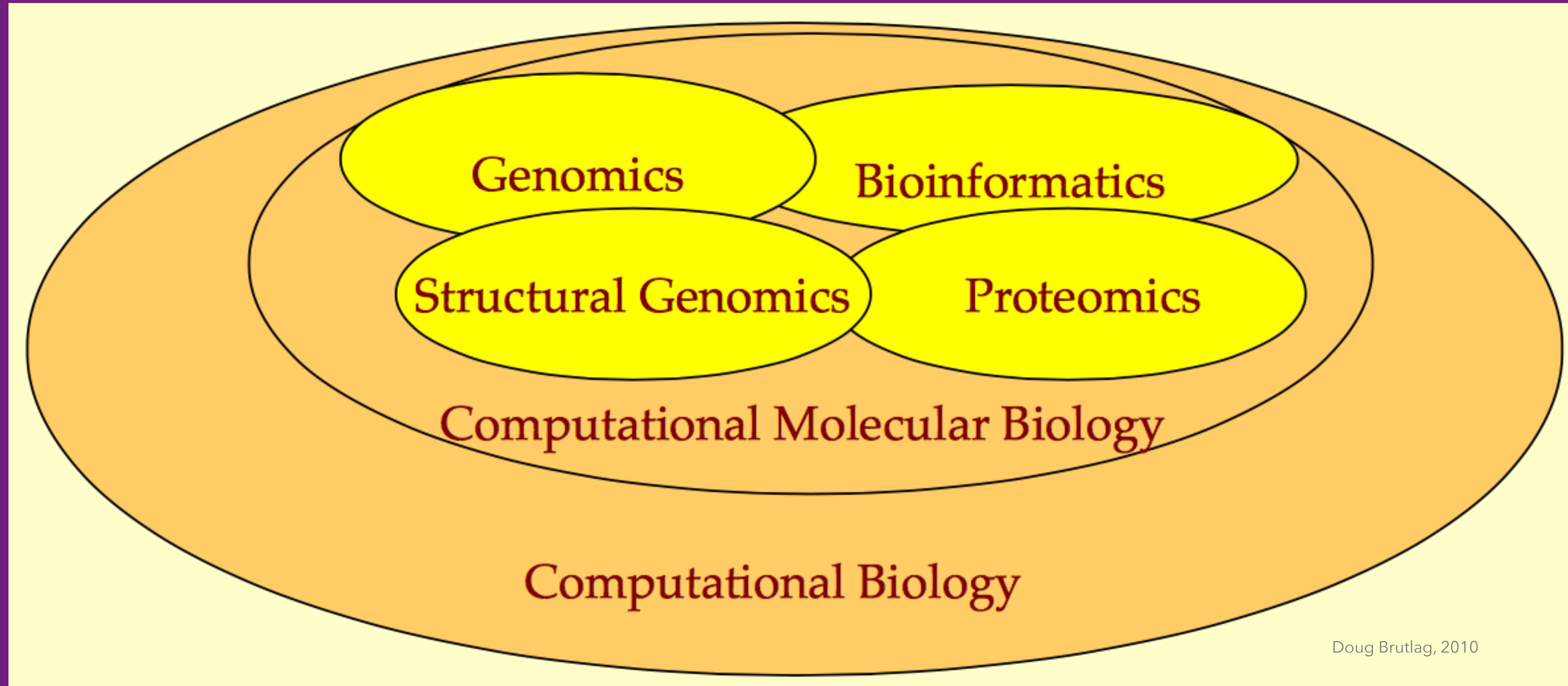
WHAT IS BIOINFORMATICS

WHAT ABOUT "COMPUTATIONAL BIOLOGY"?

INCLUDES
BIOINFORMATICS

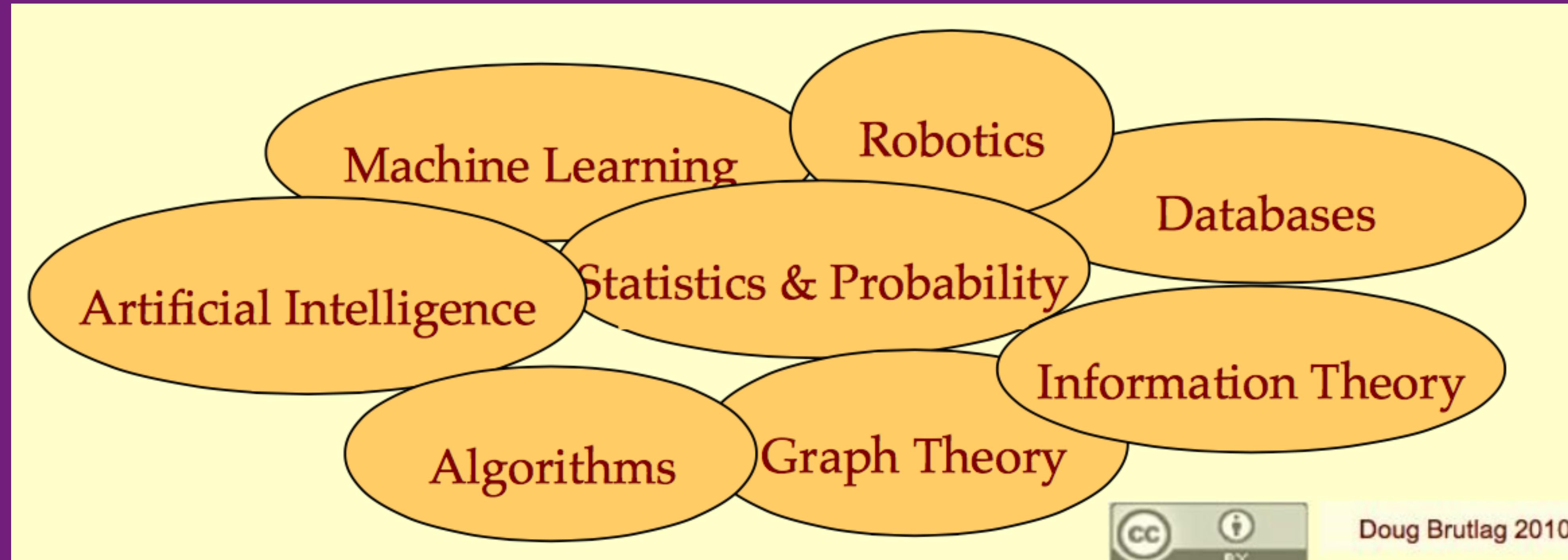
- Computational biology is an interdisciplinary field that applies the techniques to address biological problems
 - Computer science
 - Applied mathematics
 - Statistics
- A broader term (Can be used interchangeably, in my opinion)

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)

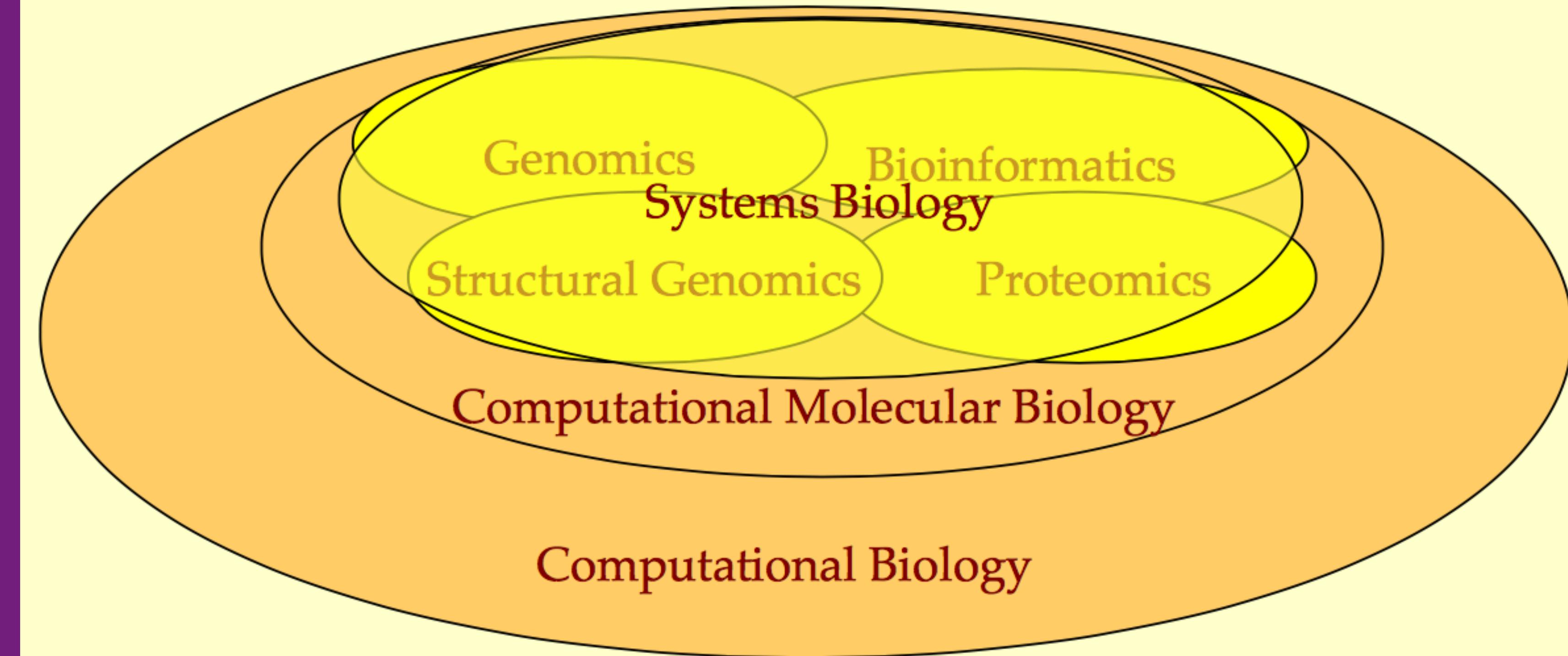
WHAT IS BIOINFORMATICS?

THERE ARE MANY
MORE SUB
DISCIPLINES

- 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

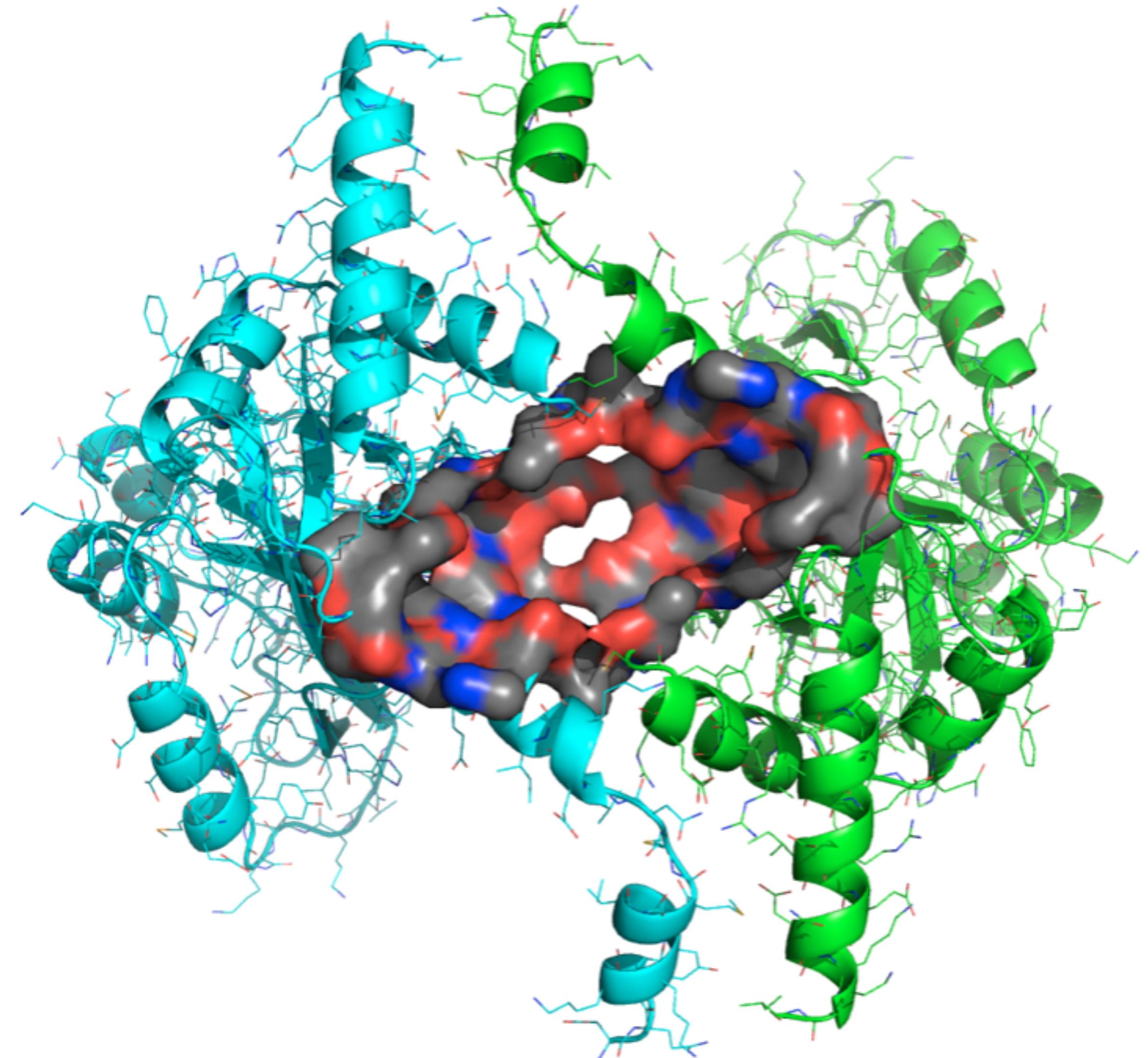
- Systems biology
 - New(er) field uses information to understand how systems work together
 - Enabled by the other fields



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

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

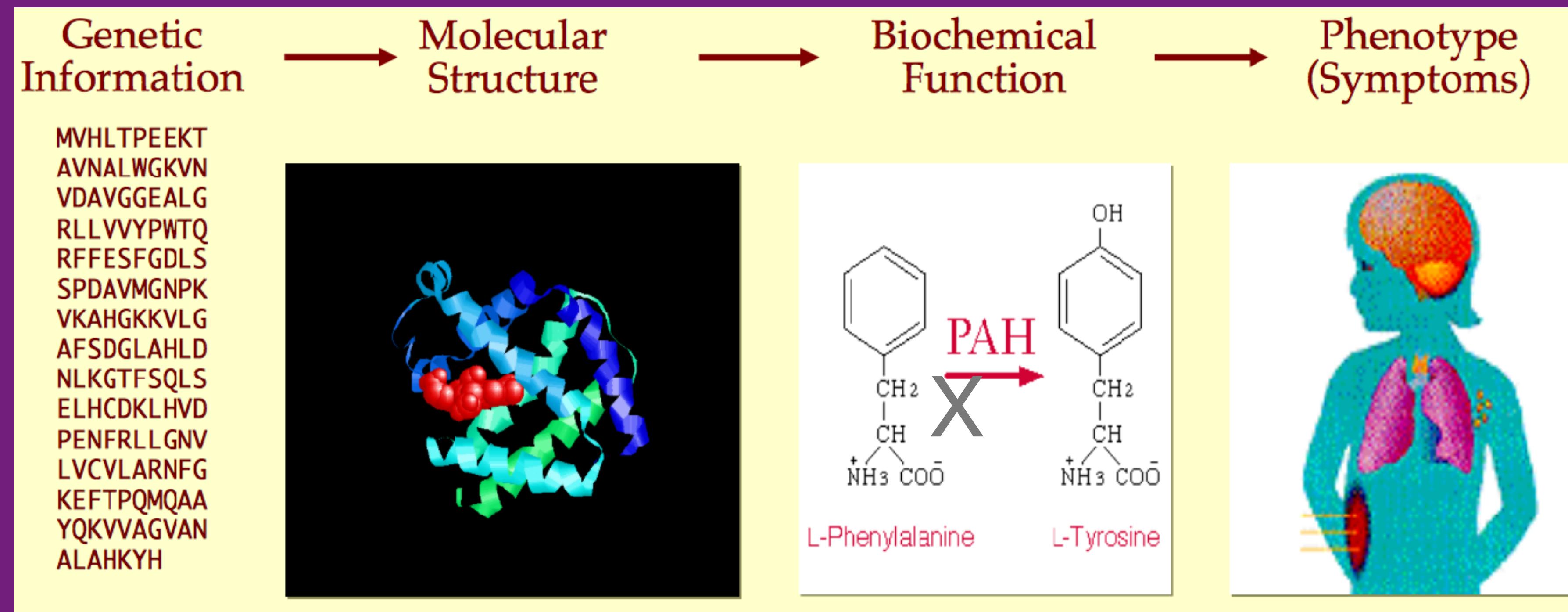
GOALS OF BIOINFORMATICS

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/INFERENCE



- Phenylalanine 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 Phenylalanine
- 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

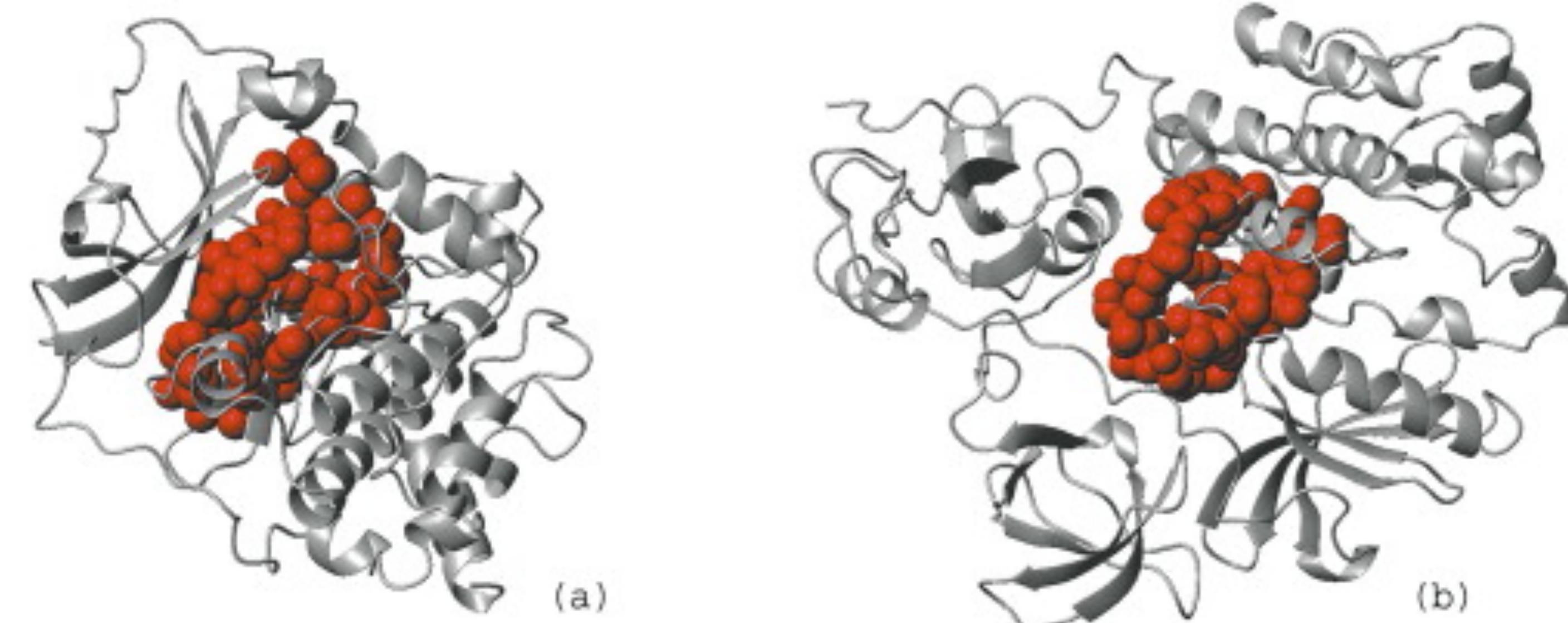
Genetic Information

MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHLD
NLKGTFSQLS
ELHCDKLHVD
PENFRLLGKV
LVCVLARNFG
KEFTPQMQAA
YQKVVAGVAN
ALAHKYH

GOALS OF BIOINFORMATICS

- Challenges: 1D to 3D

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



(c)

>1cdk_A

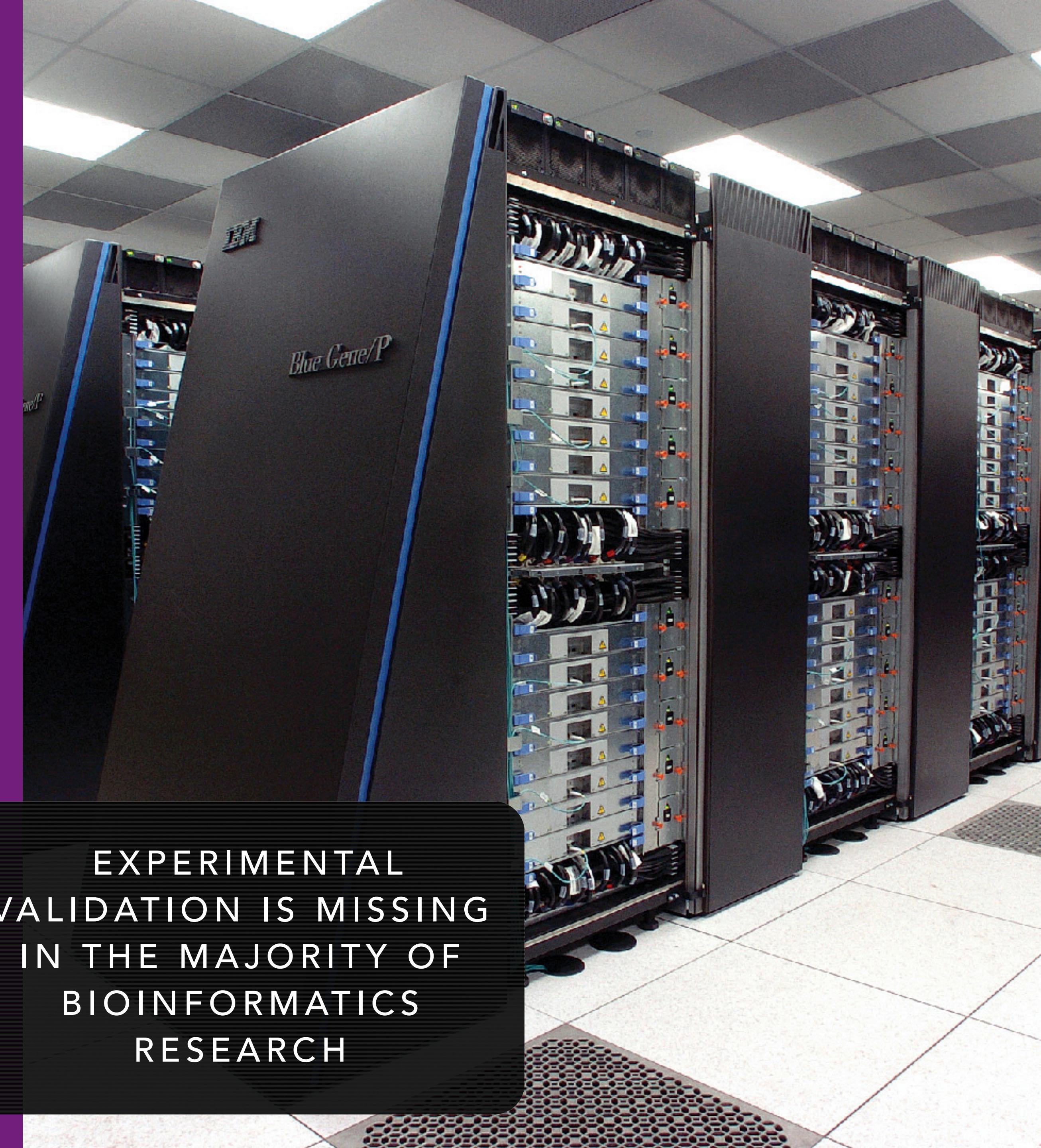
GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHDQFERIKT**LGTGSFGRV**MLVKHKETGNHF**AMKILD**
KQKVVKLQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQI
VLTFEYLHSLDLIYRDLKPENLLIDQQGYIQV**TDFGF**AKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWALG
VLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHSSDLKDLLRNLLQVDLTKRGPNLKDGVNIDKNHKWFATT
DWIAIYQRKVEAPFIPFKFGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF

>2src_

MVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRR
ESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVHYKIRKLDSGGFYITSRTQFNSLQQLVAYYS
KHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEV**KLGQGCFGEV**WMGTWNGTTRV**AIKTL**KPGT**MSPEA**
FLQEAQVMKKLRHEKLVQLYAVVSEEPYIVT**EYMSKGSLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMN**
YVHRDL**RAANIL**VGENLVCKV**ADFGLARLIEDNEYTARQGAKFPIKWT**PEAALYGRFTIKSDVWSFGILLTEL
TKGRVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLDYFTSTEPQXQPGE
NL

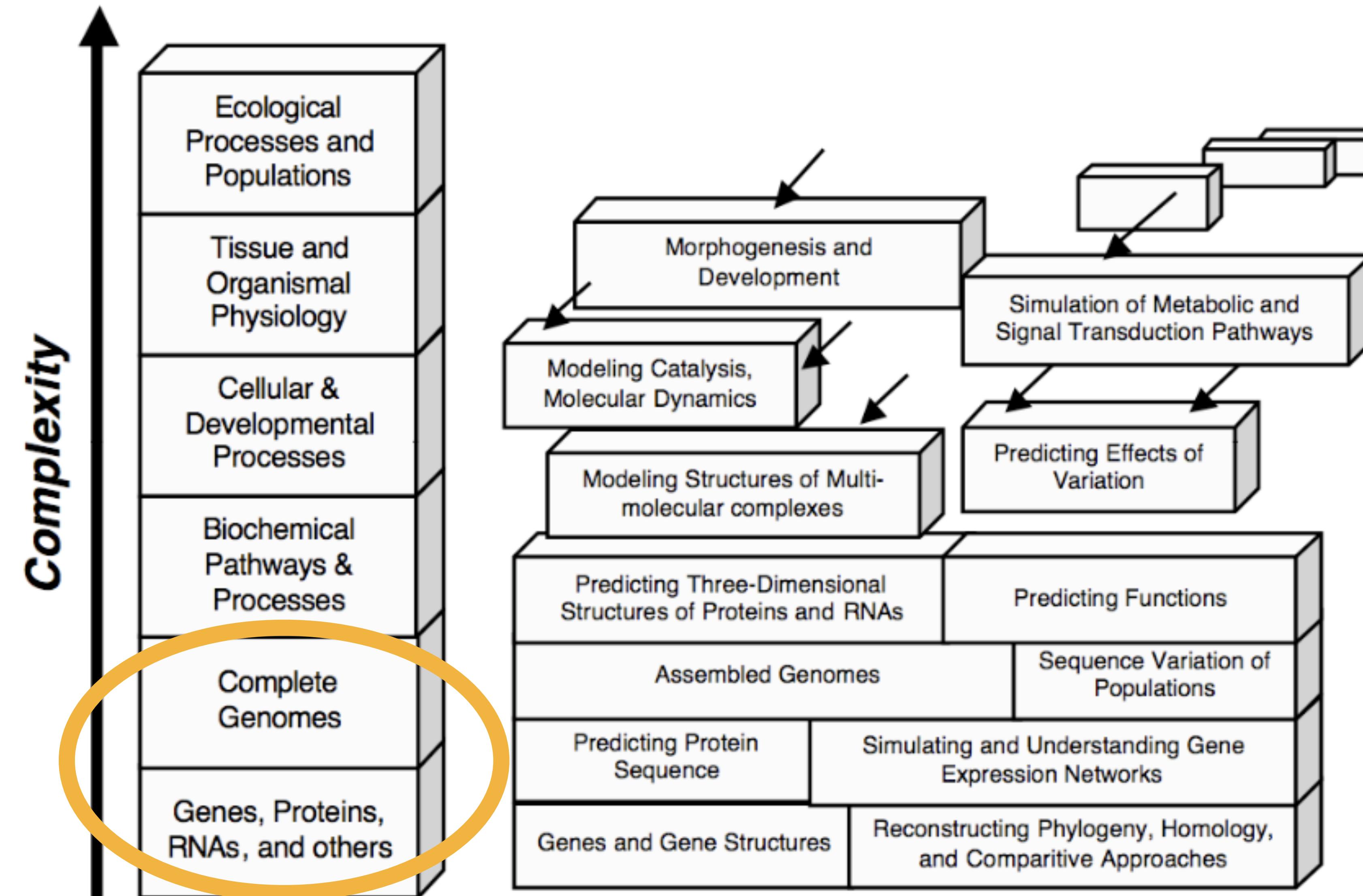
GOALS OF BIOINFORMATICS

- 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

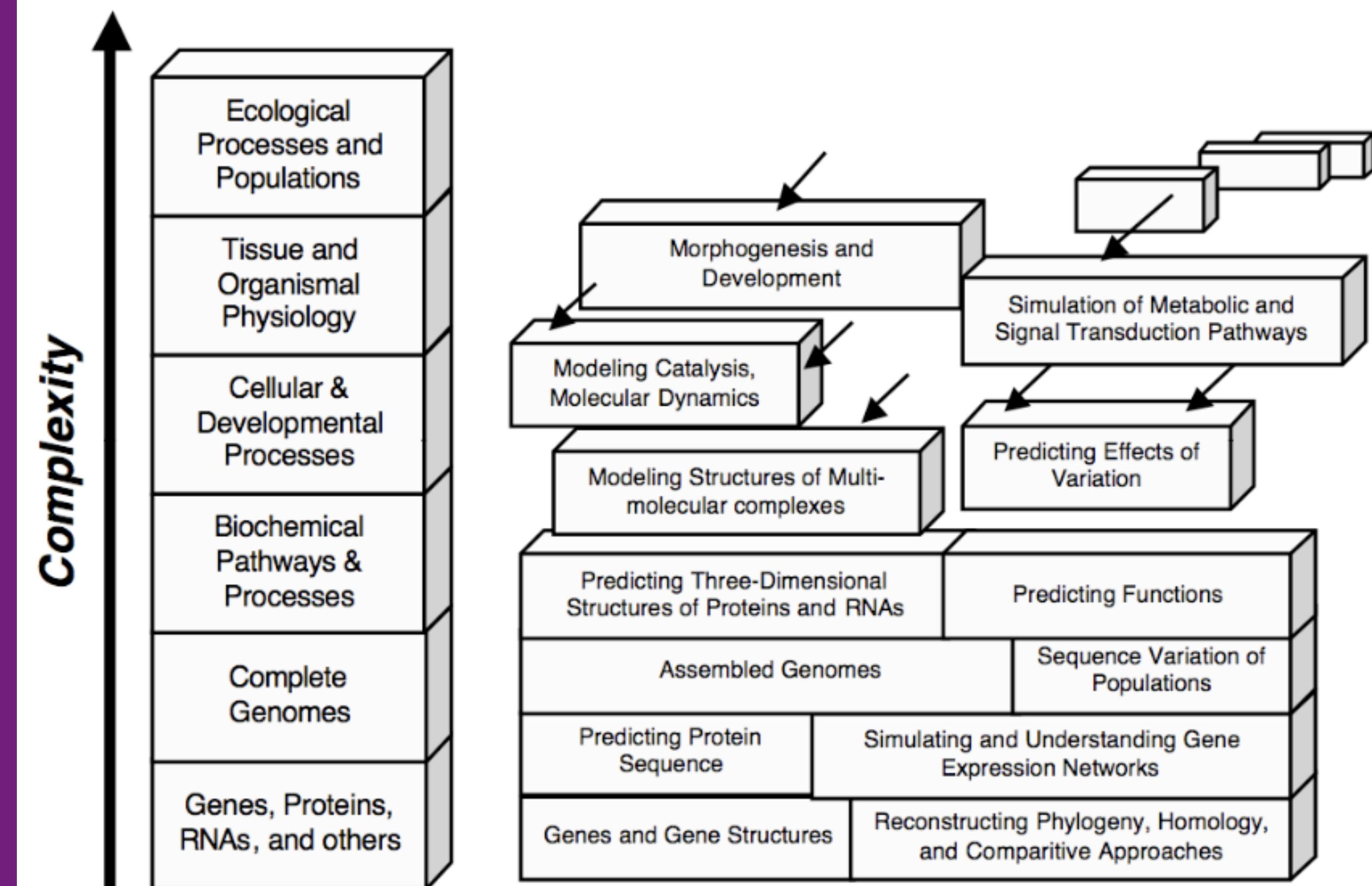
GOALS OF BIOINFORMATICS



IMPACT OF BIOINFORMATICS

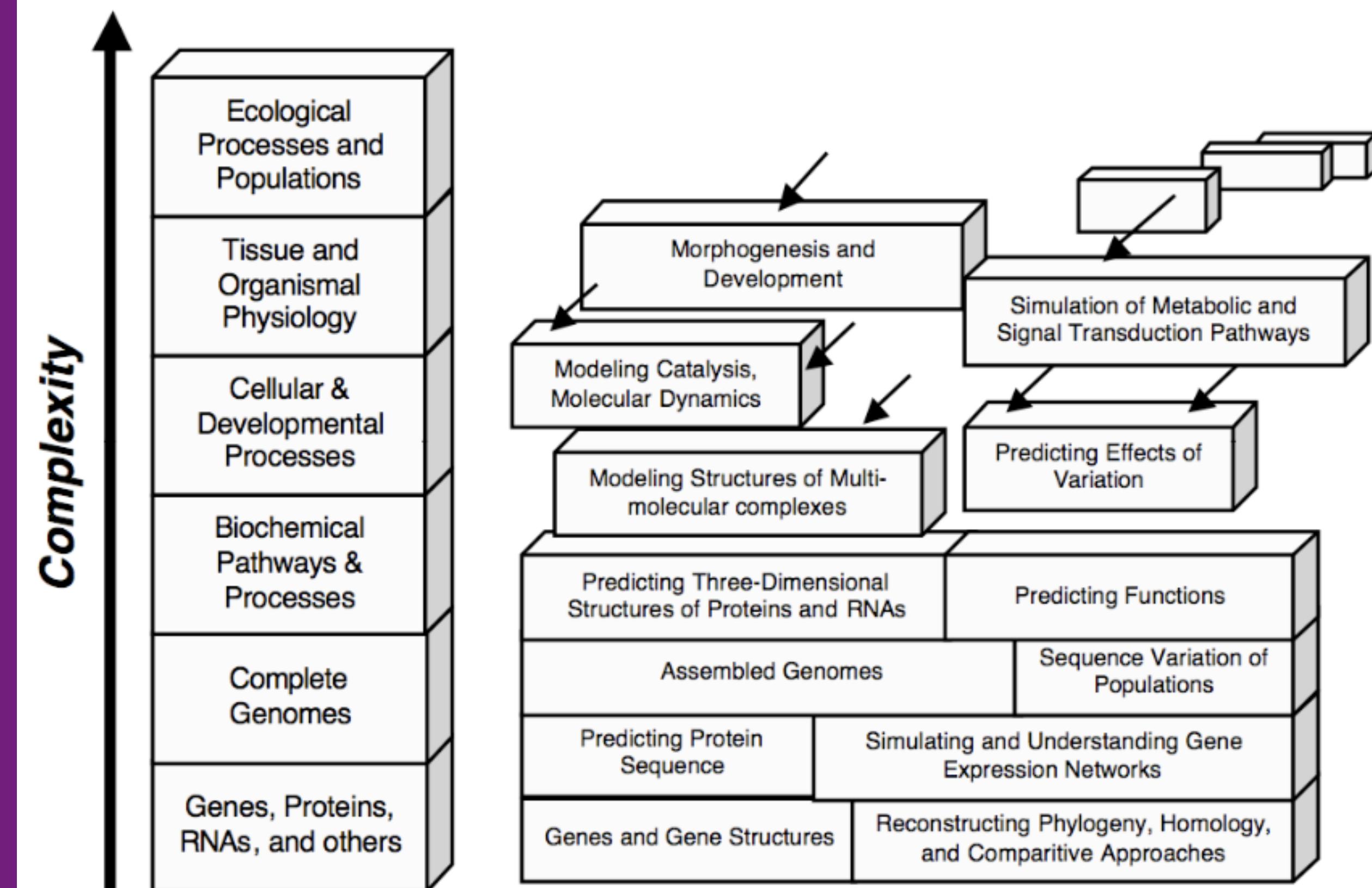
IMPACT OF BIOINFORMATICS

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



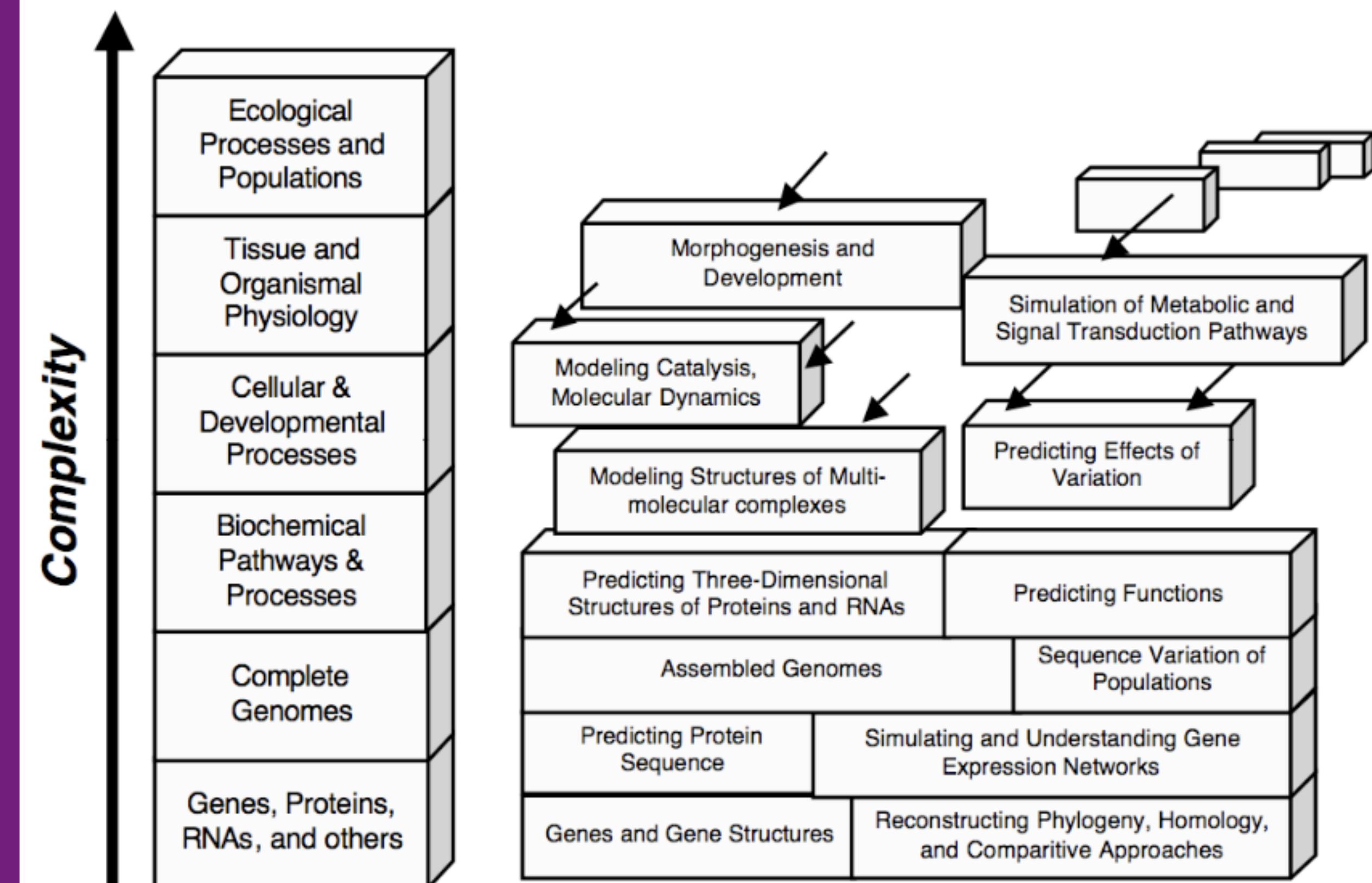
IMPACT OF BIOINFORMATICS

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



IMPACT OF BIOINFORMATICS

- On Medical sciences
 - Translational research
 - Personalized treatment



SESSION 1A

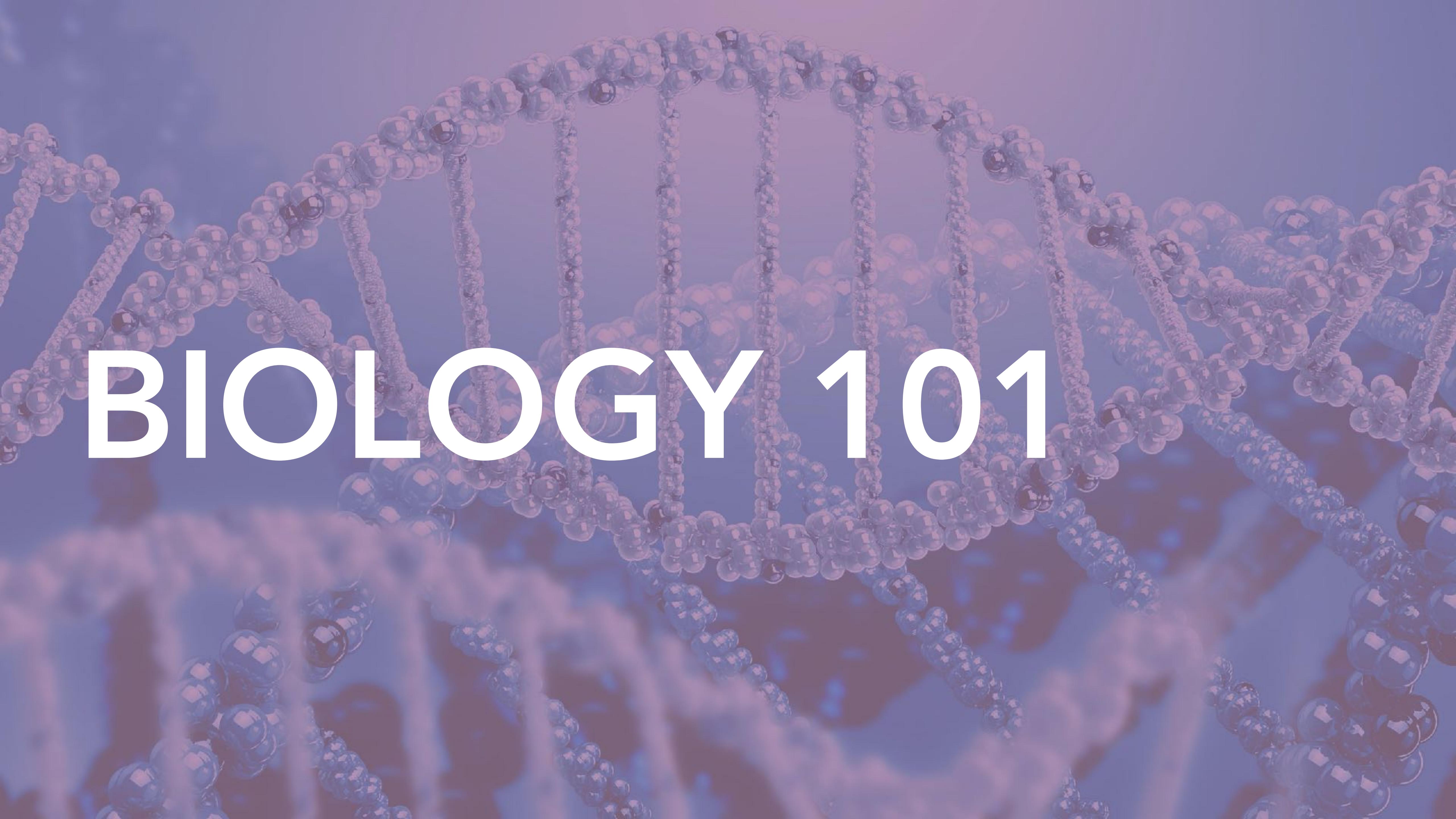
BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
AUTUMN 2020
SESSION 1A



THE UNIVERSITY OF
CHICAGO

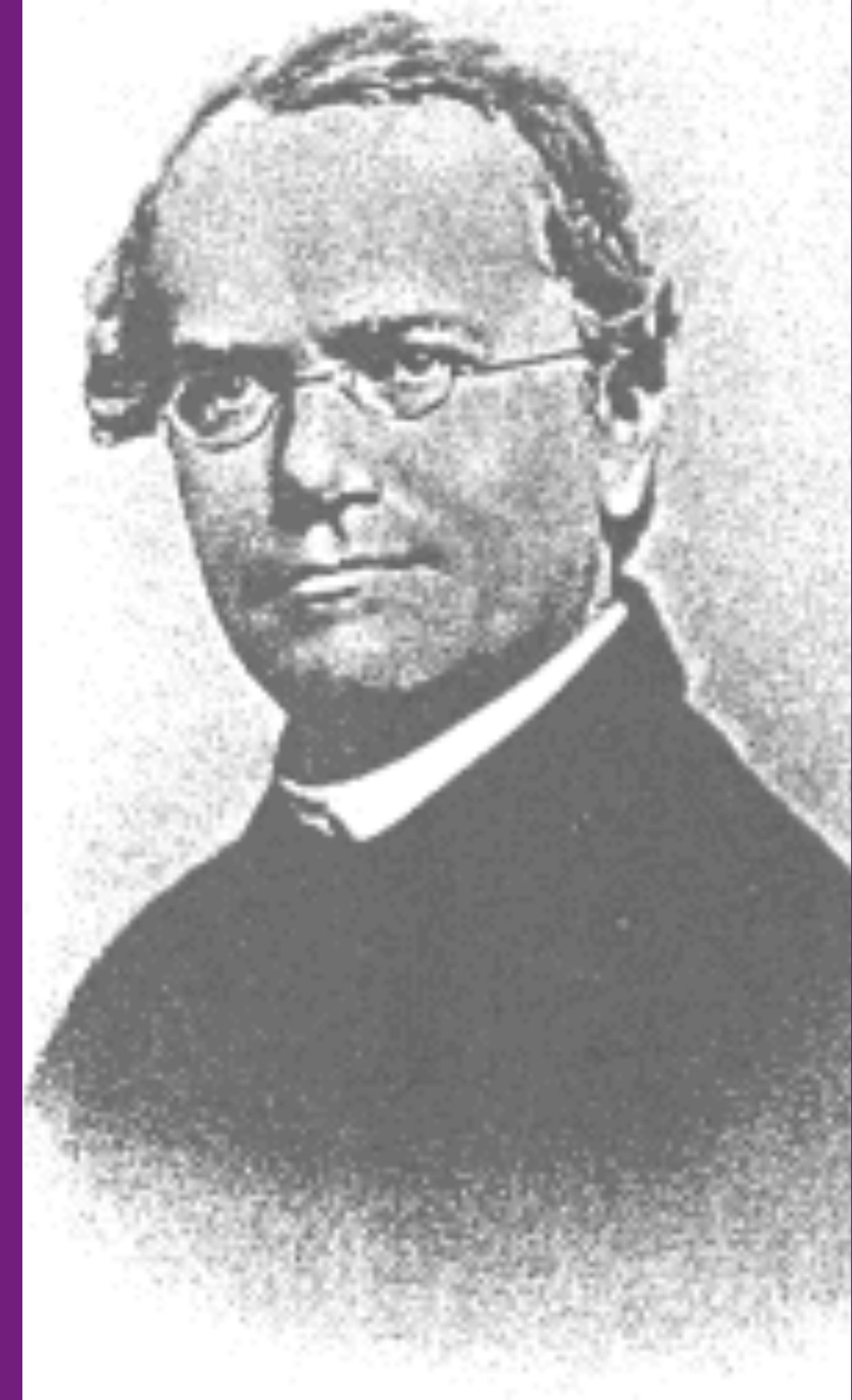


BIOLOGY 101

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 progeny and coined the terms dominant and recessive traits



TIMELINE OF MOLECULAR BIOLOGY

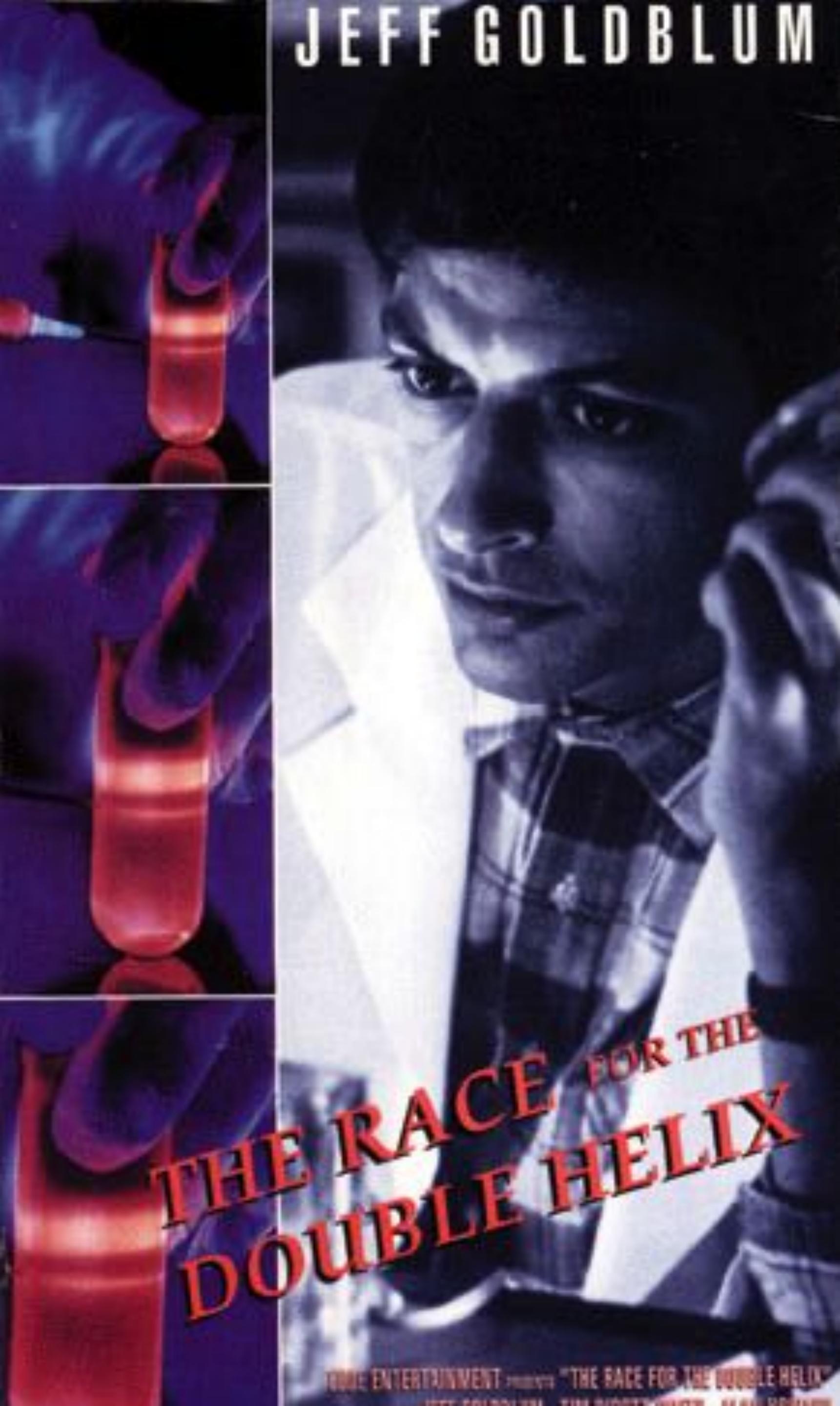
- 1900
 - Chemical structures of all 20 amino acids had been identified
- 1902
 - Emil Hermann Fischer wins Nobel prize: showed amino acids are linked and form proteins
- 1941
 - George Beadle and Edward Tatum identify that genes make proteins



Emil
Fischer

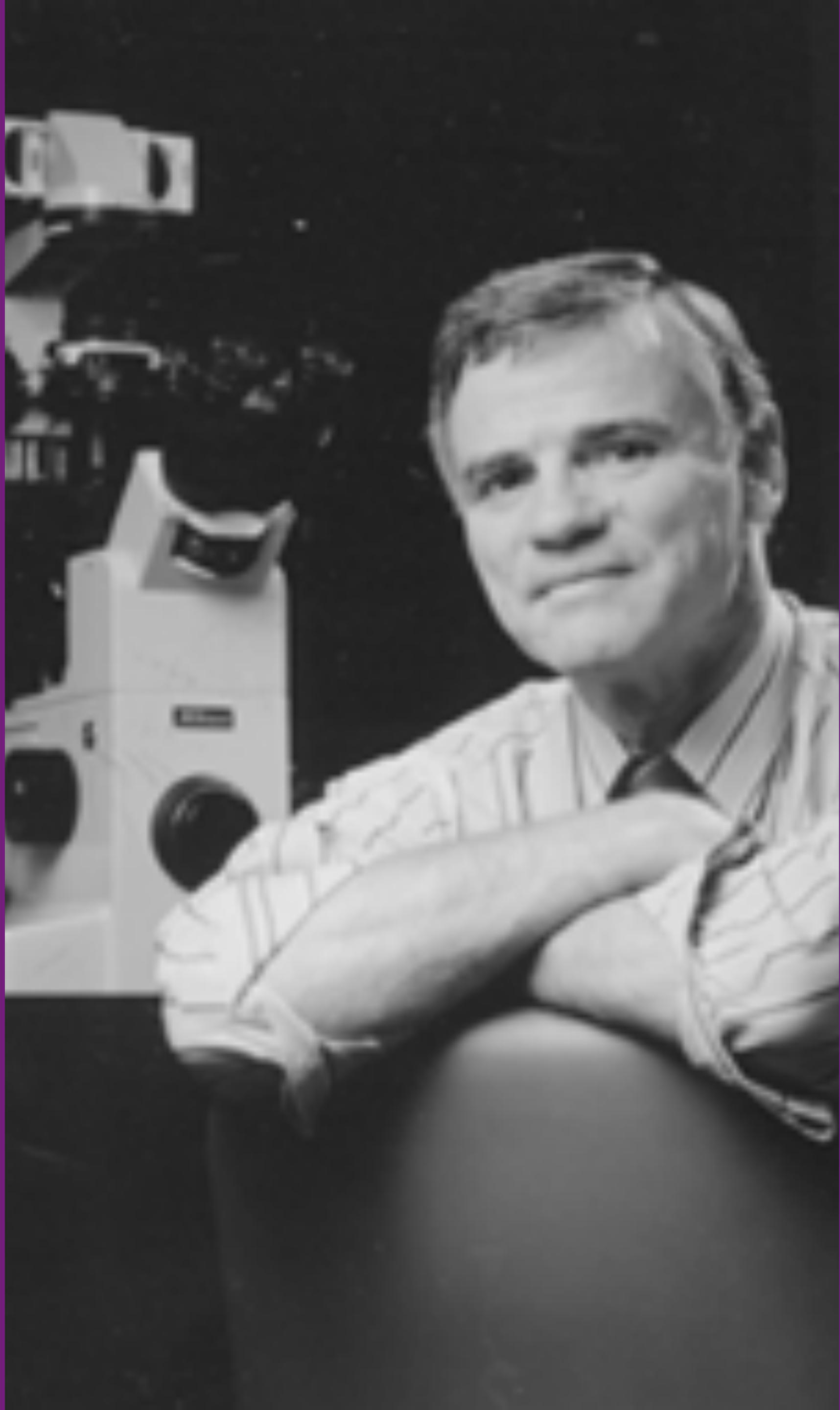
TIMELINE OF MOLECULAR BIOLOGY

- 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 I¹, 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

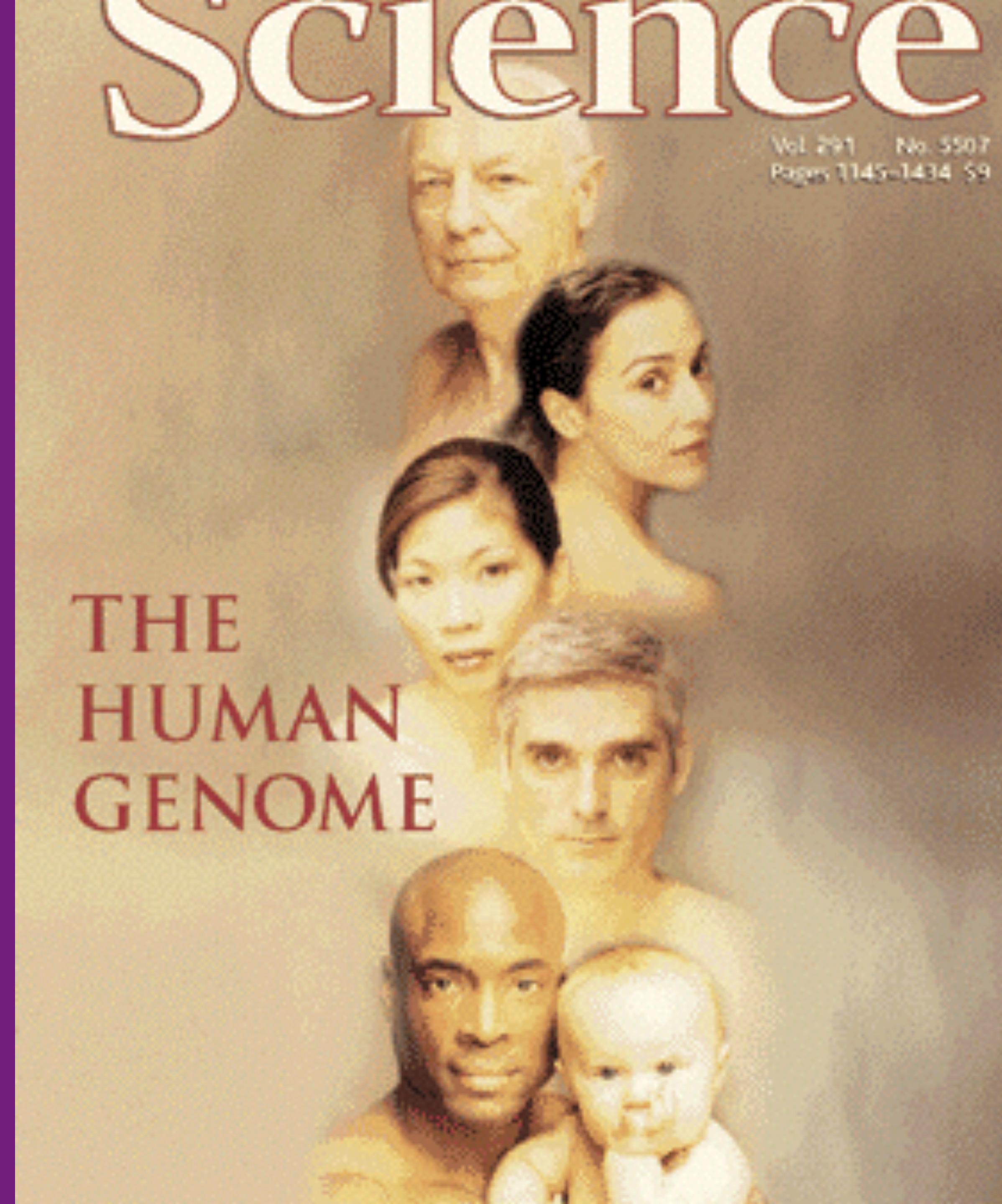
Knowledge of the complete genomic DNA sequence of an organism allows a systematic approach to defining its genetic components. The genomic sequence provides access to the structures of all genes, including those without known function, their control elements, their inferred function by homology inference,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

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 pore over data. [Nature. 1999]
The book of genes. [Nature. 1999]

TIMELINE OF MOLECULAR BIOLOGY

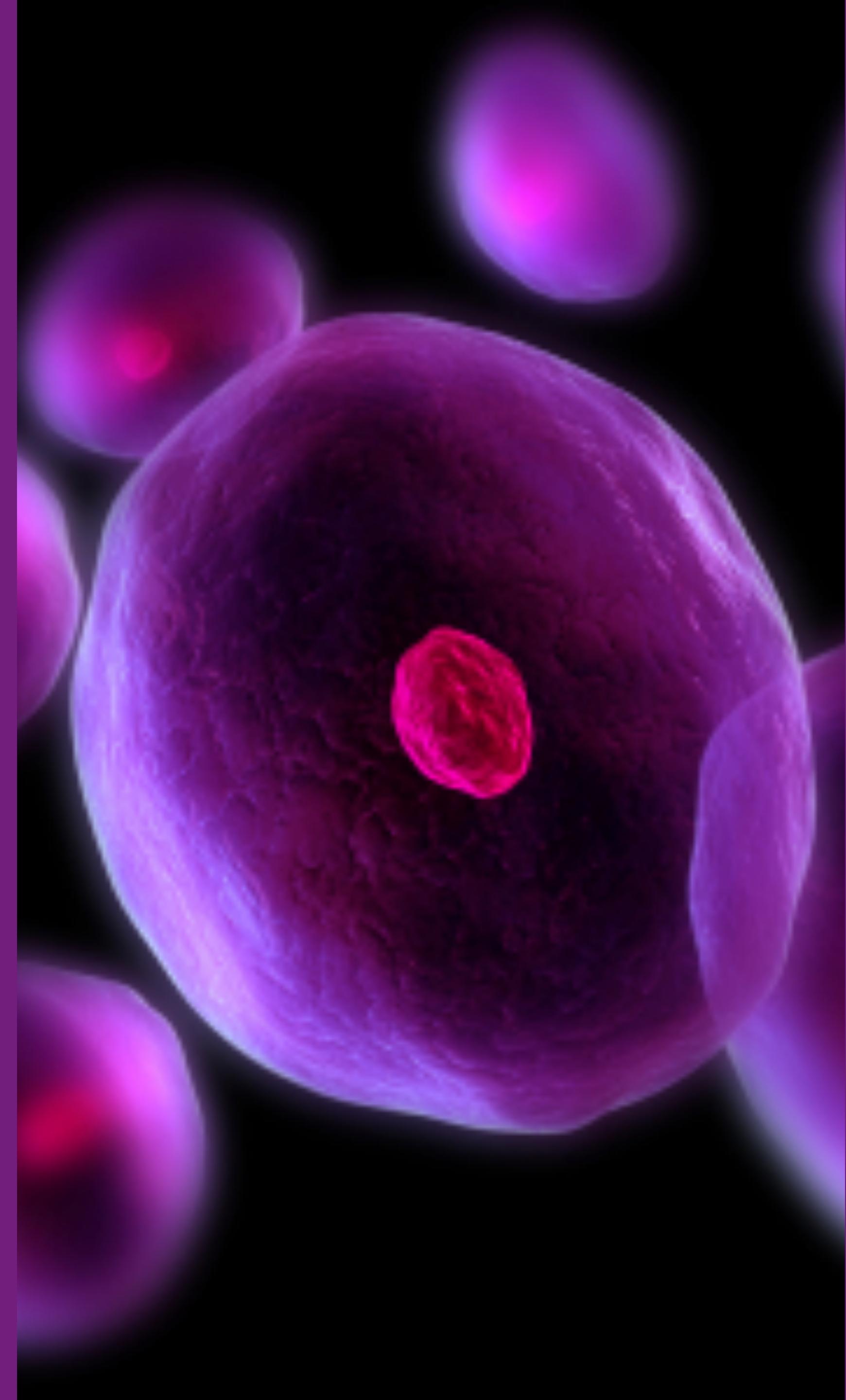
- 2000
 - Complete sequence of the *Drosophila melanogaster* genome
- 2001
 - International Human Genome Sequencing publishes first draft of the sequence of the human genome
- 2003
 - Human Genome Project "completed"



CELLS

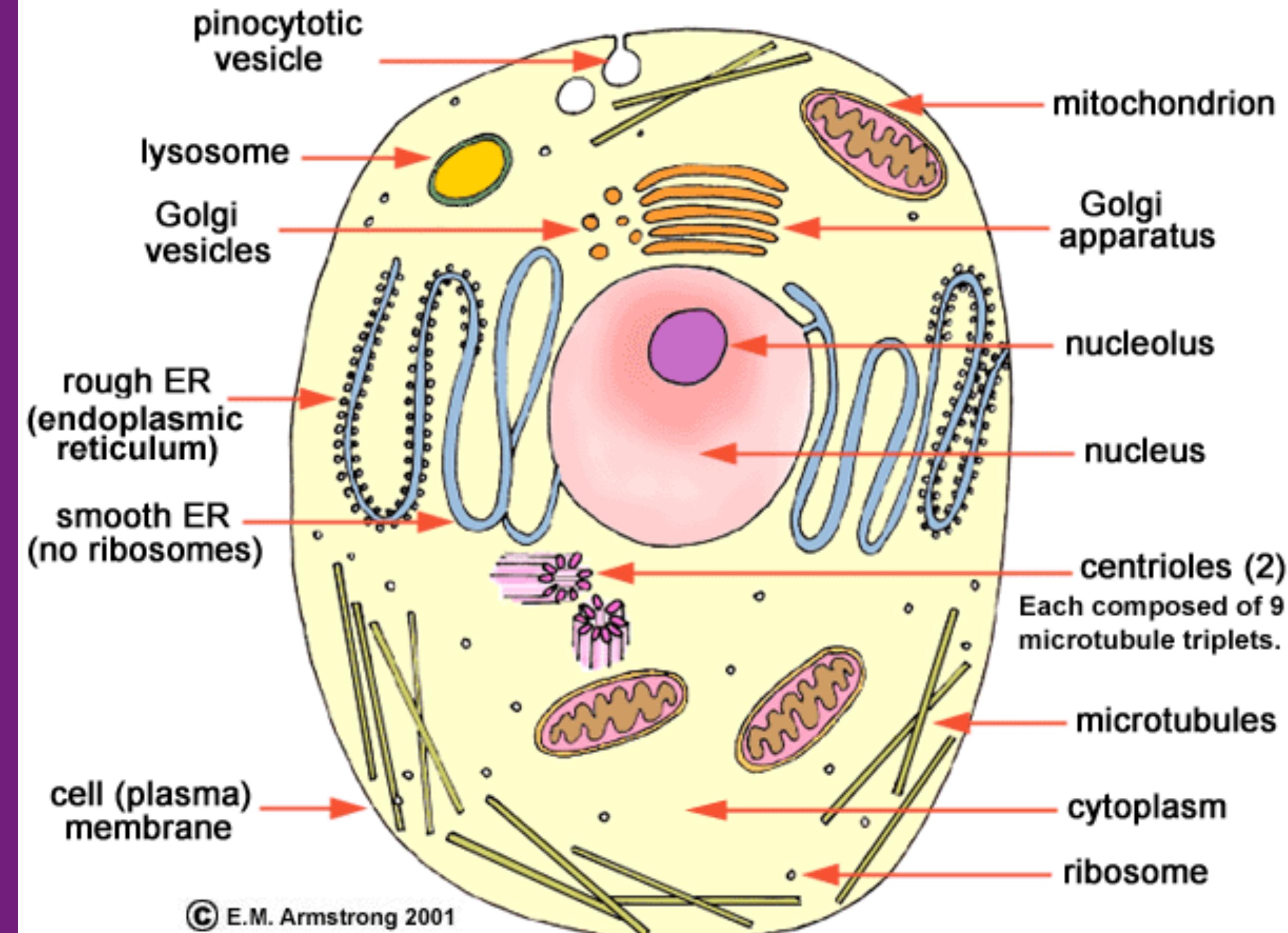
CELLS

- 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

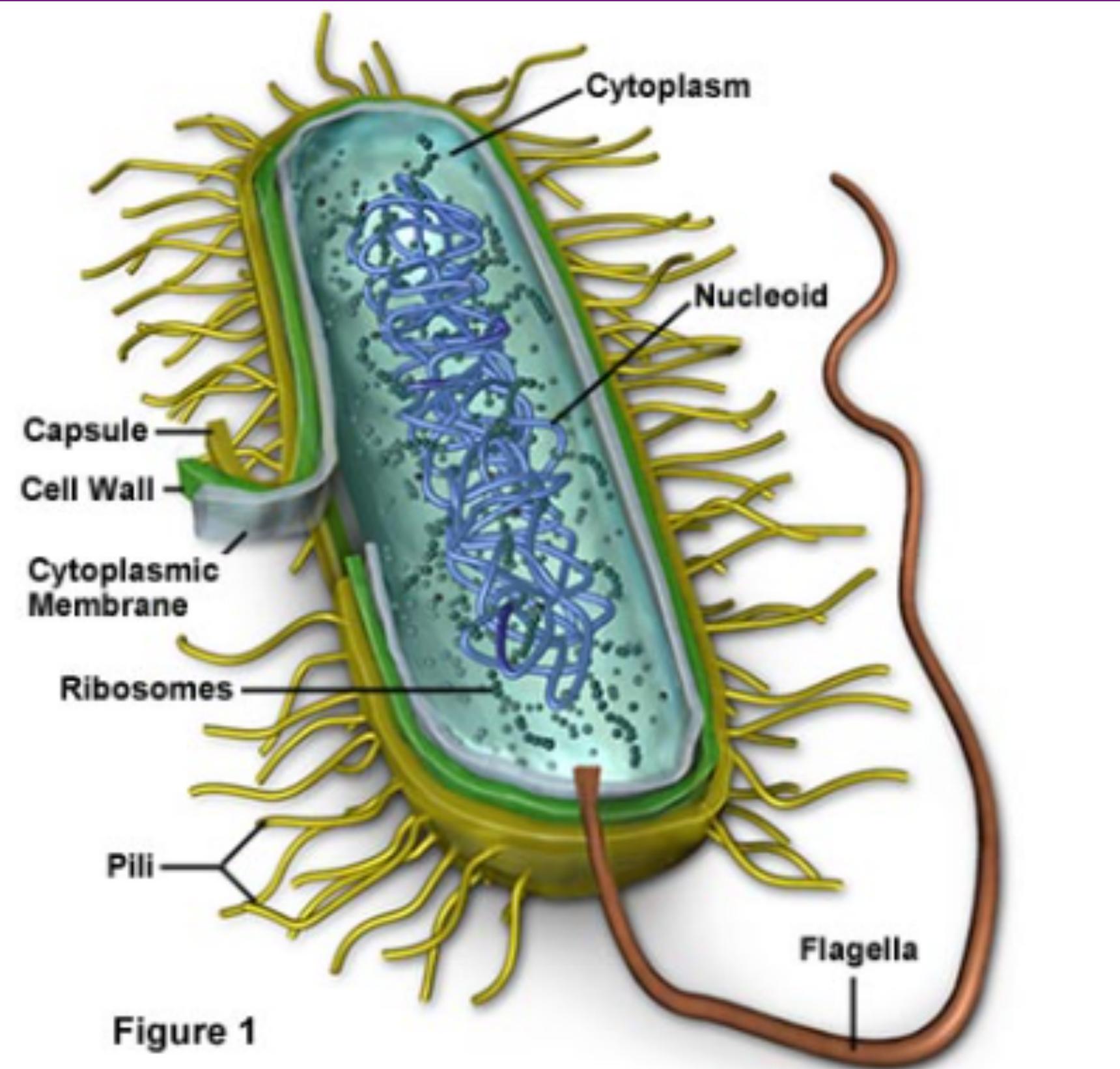
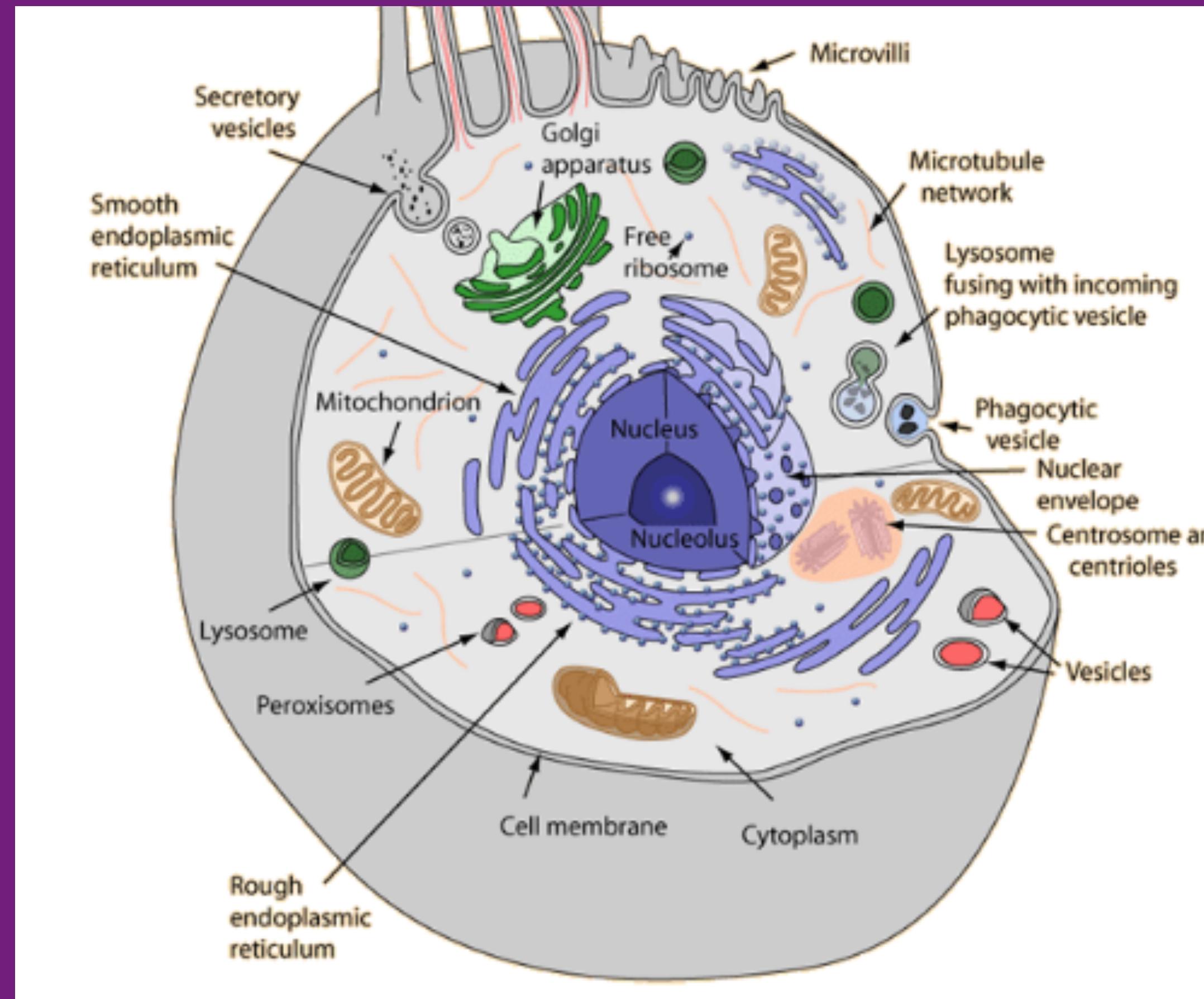


Figure 1

- 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

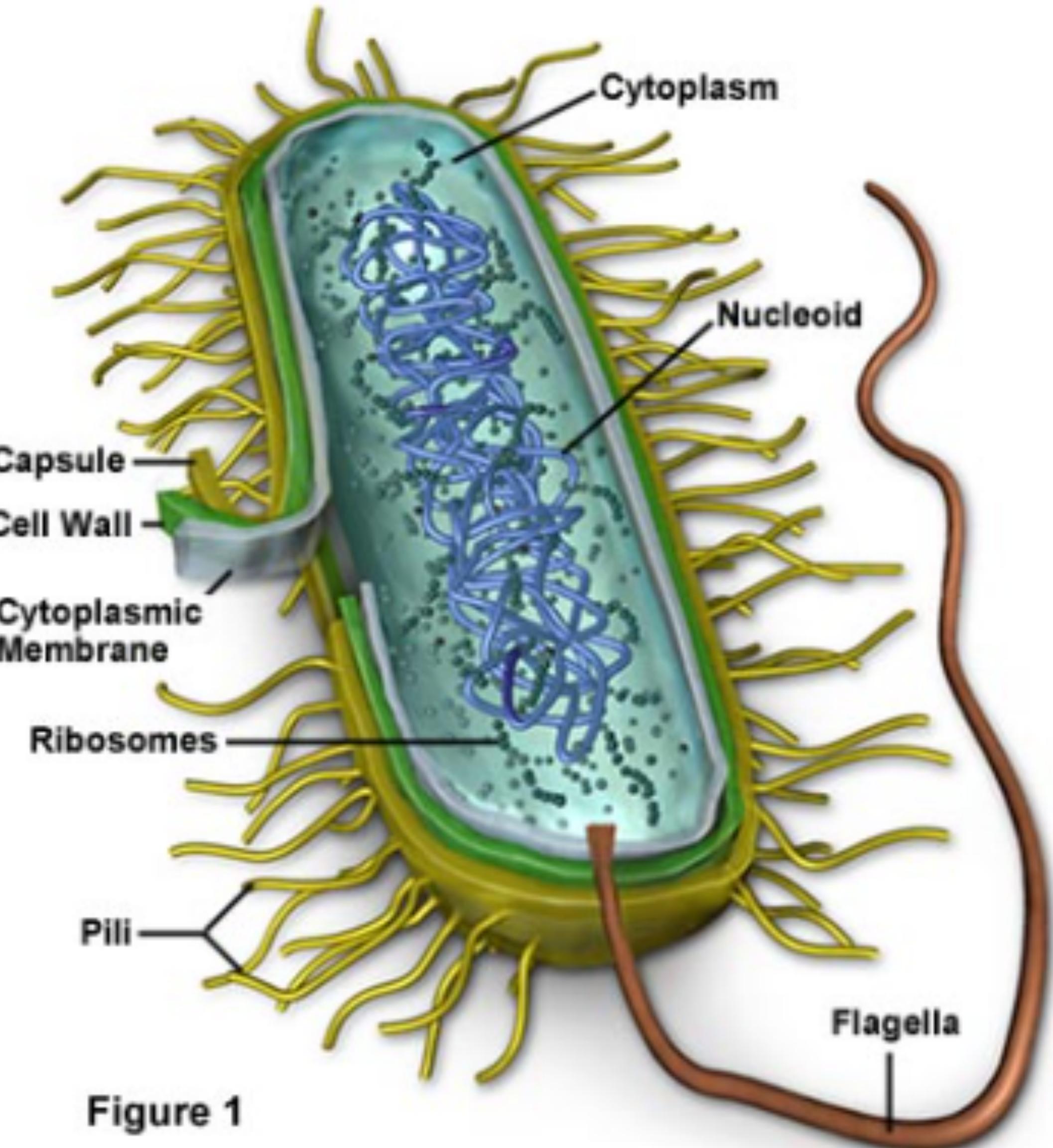


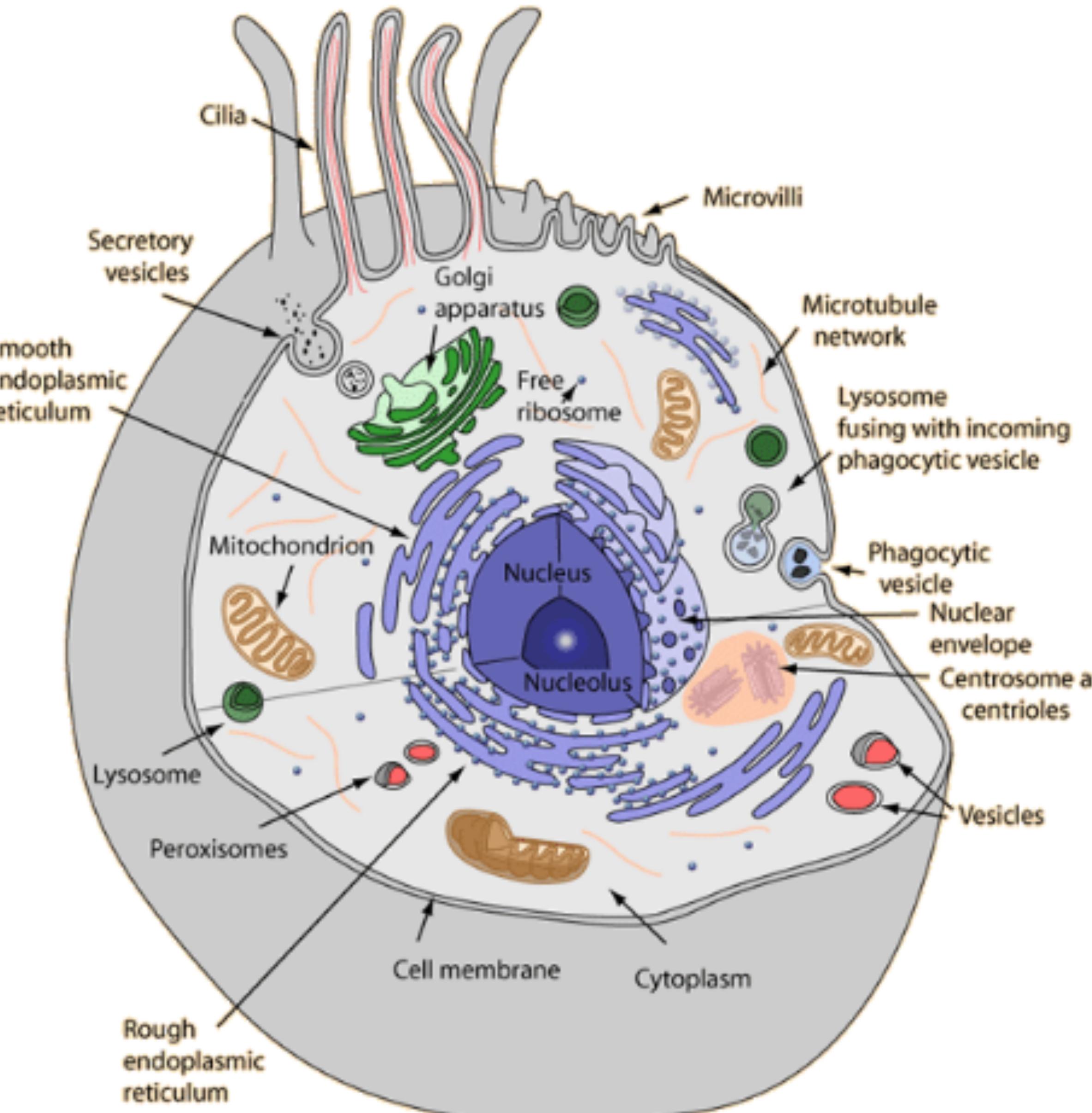
Figure 1

Prokaryotic cell structure

CELLS

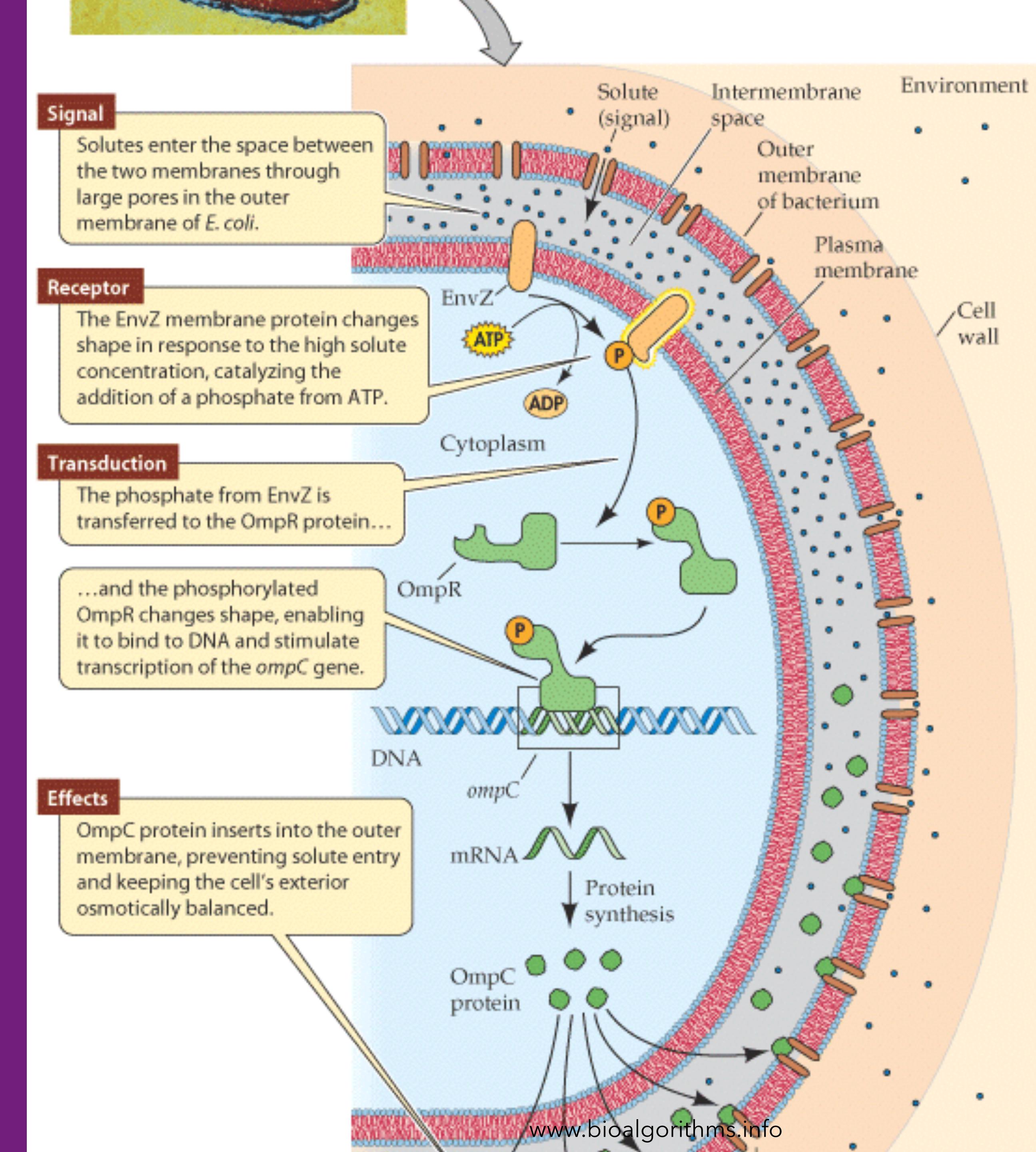
EUKARYOTES

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



CELLS

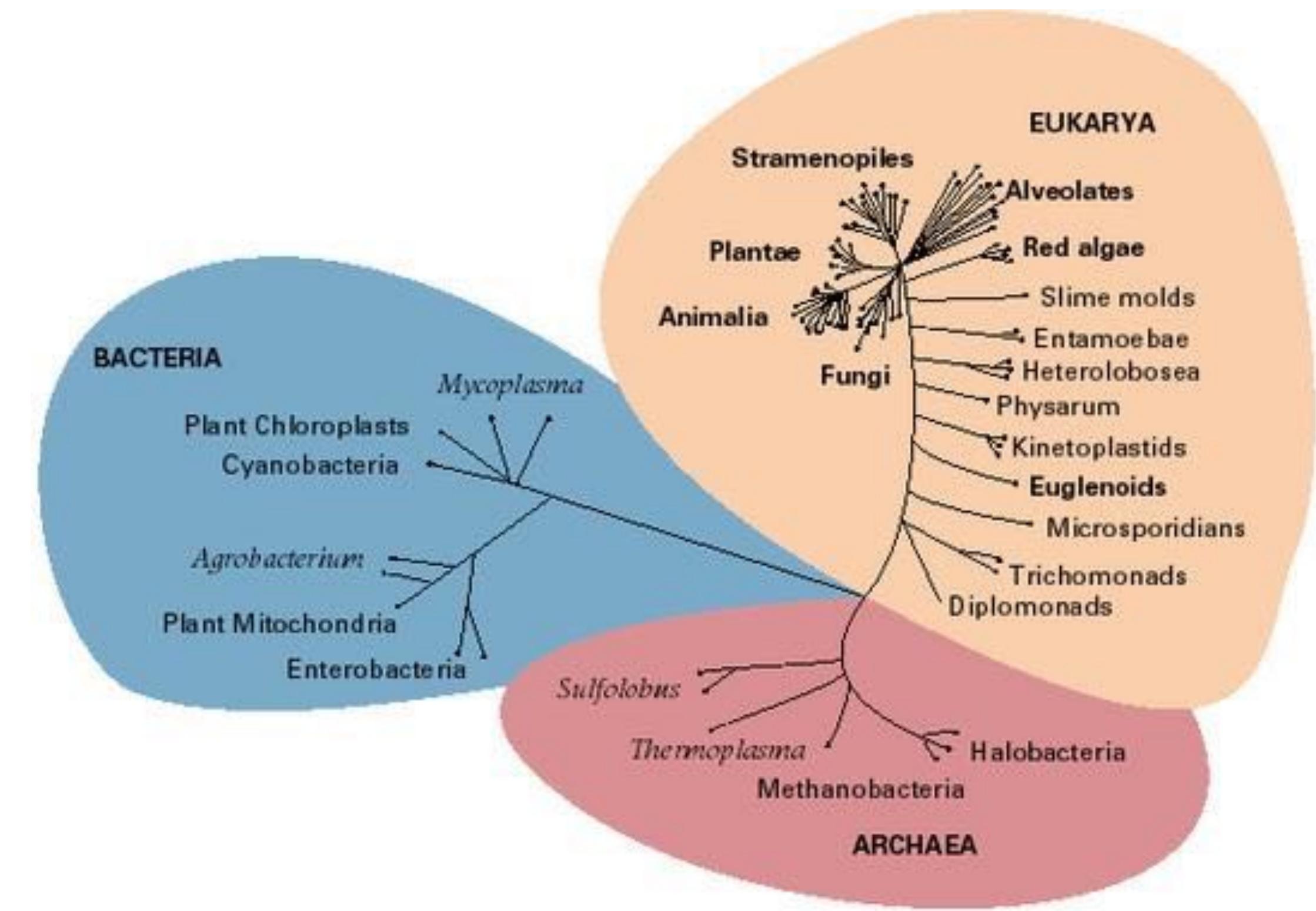
- Cells make decision through complex networks of chemical reactions, called pathways
- Synthesize new materials
- Break other materials down for spare parts
- Signal to eat or die



CELLS

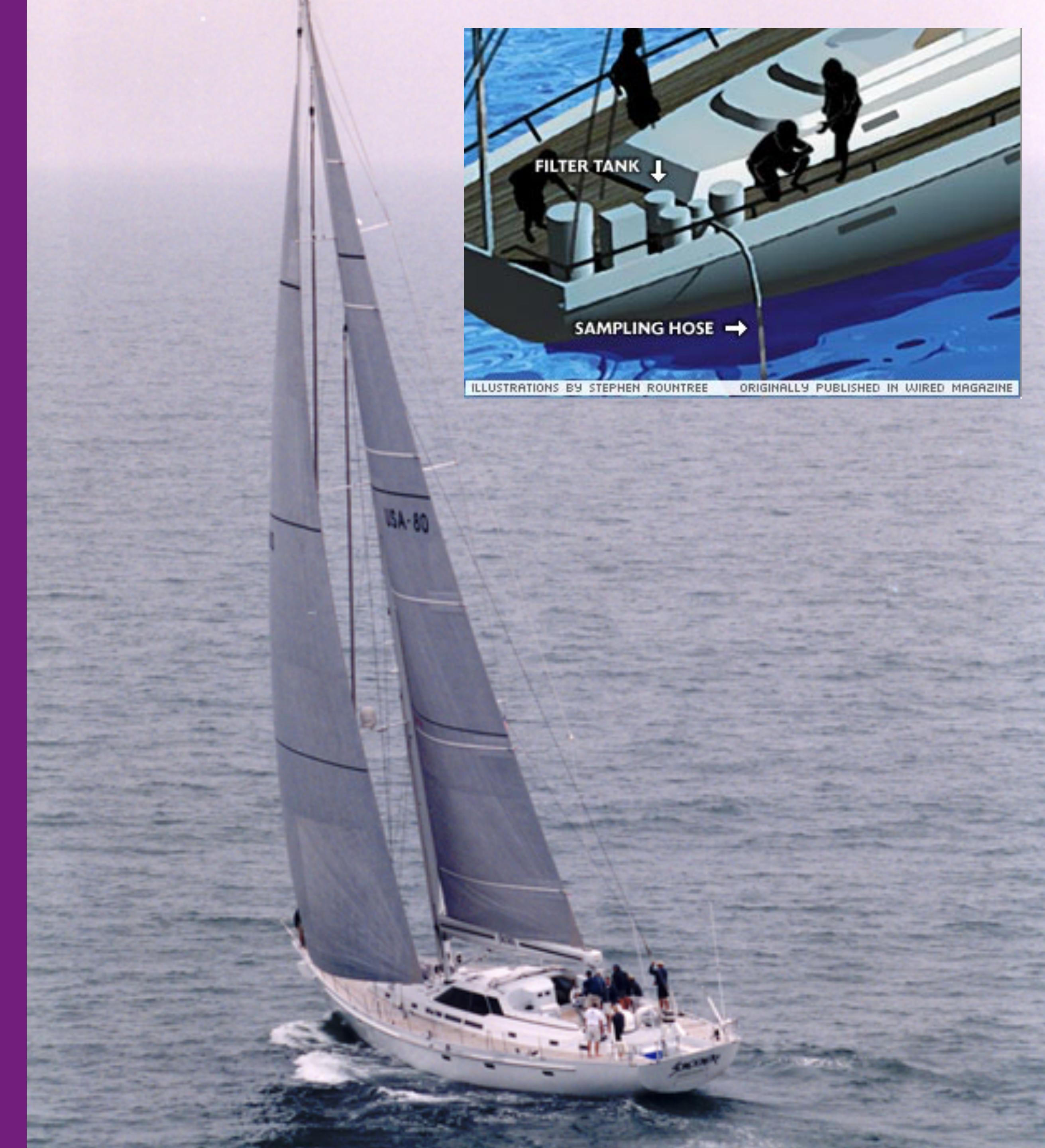
- According to the most recent evidence, there are three main branches to the tree of life

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



CELLS

- Global oceanic sampling expedition



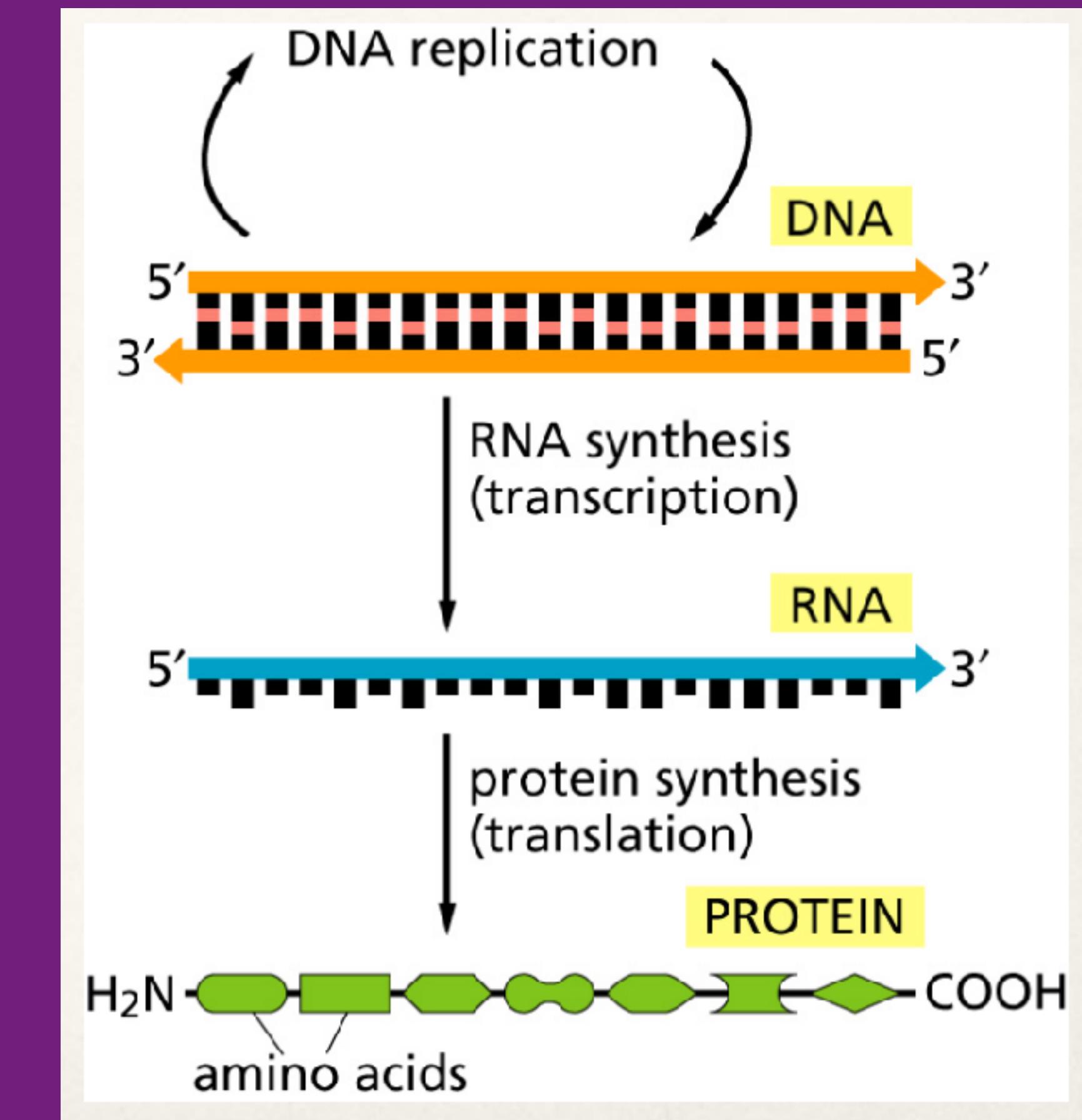
CENTRAL DOGMA OF MOLECULAR BIOLOGY

CENTRAL DOGMA OF MOLECULAR BIOLOGY

DNA CAN
REPLICATE

INFORMATION IN
DNA IS PASSED TO
RNA

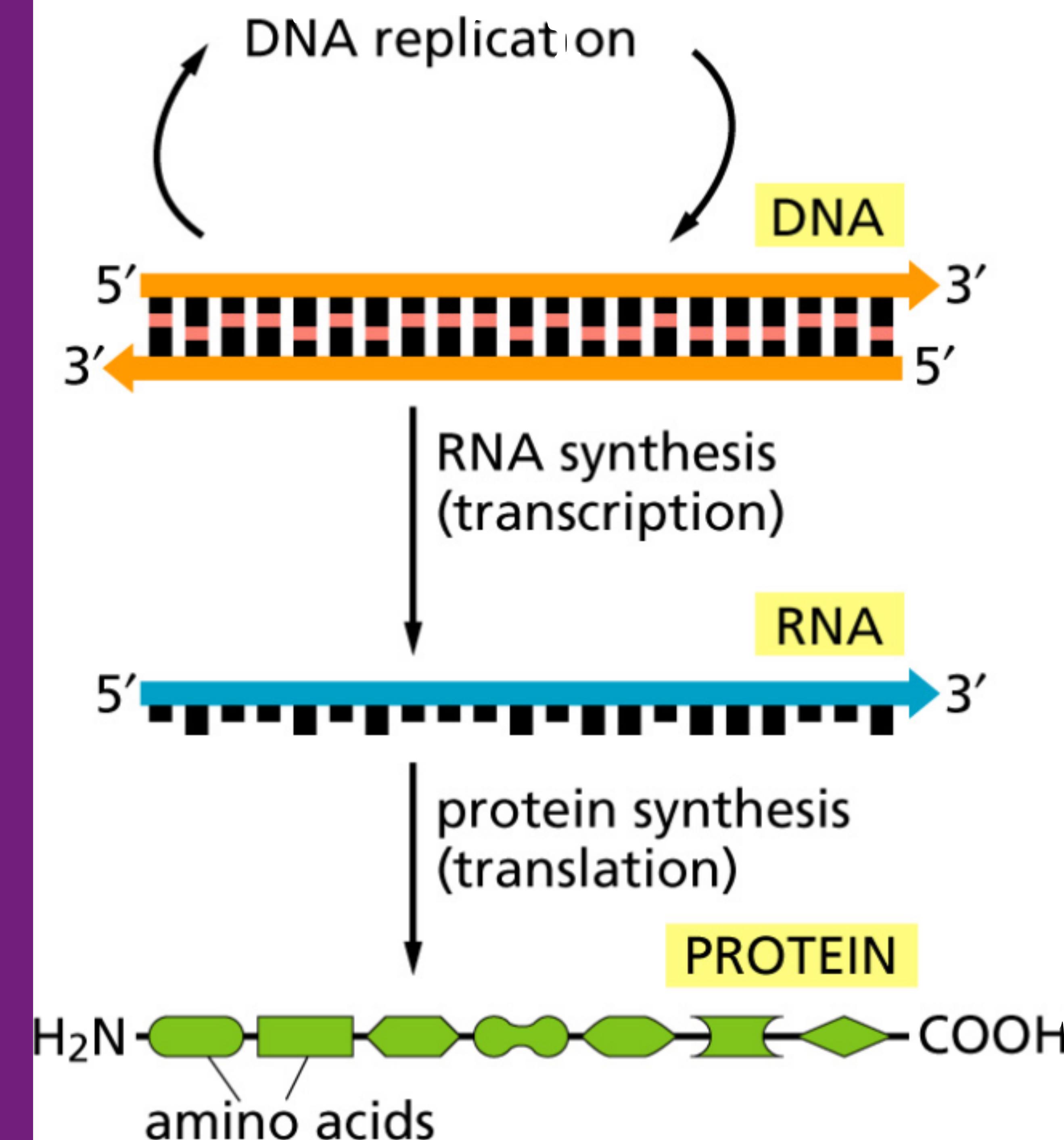
INFORMATION IN
RNA IS PASSED TO
PROTEINS



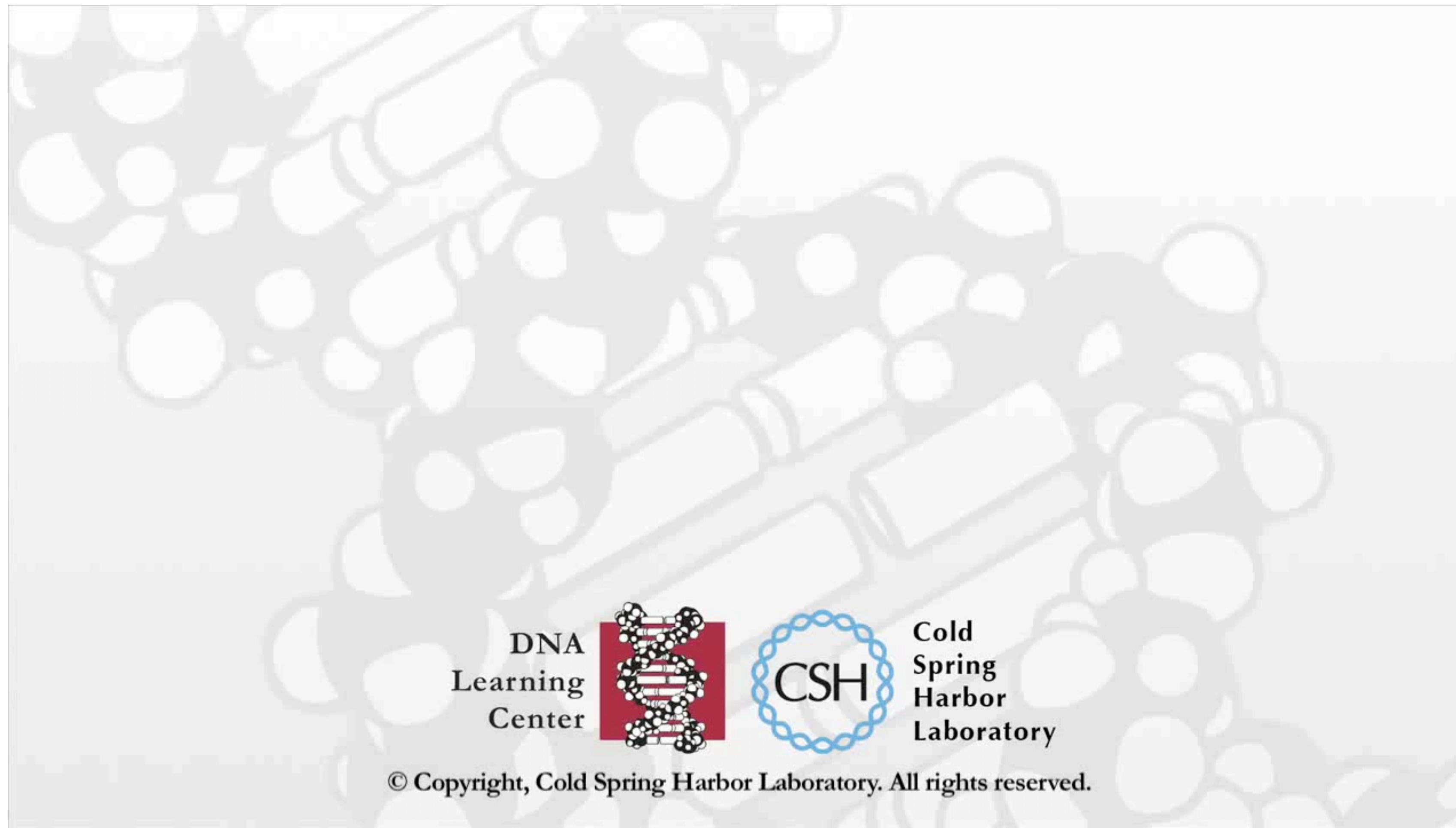
- Describes the flow of information in cells

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



DNA
Learning
Center

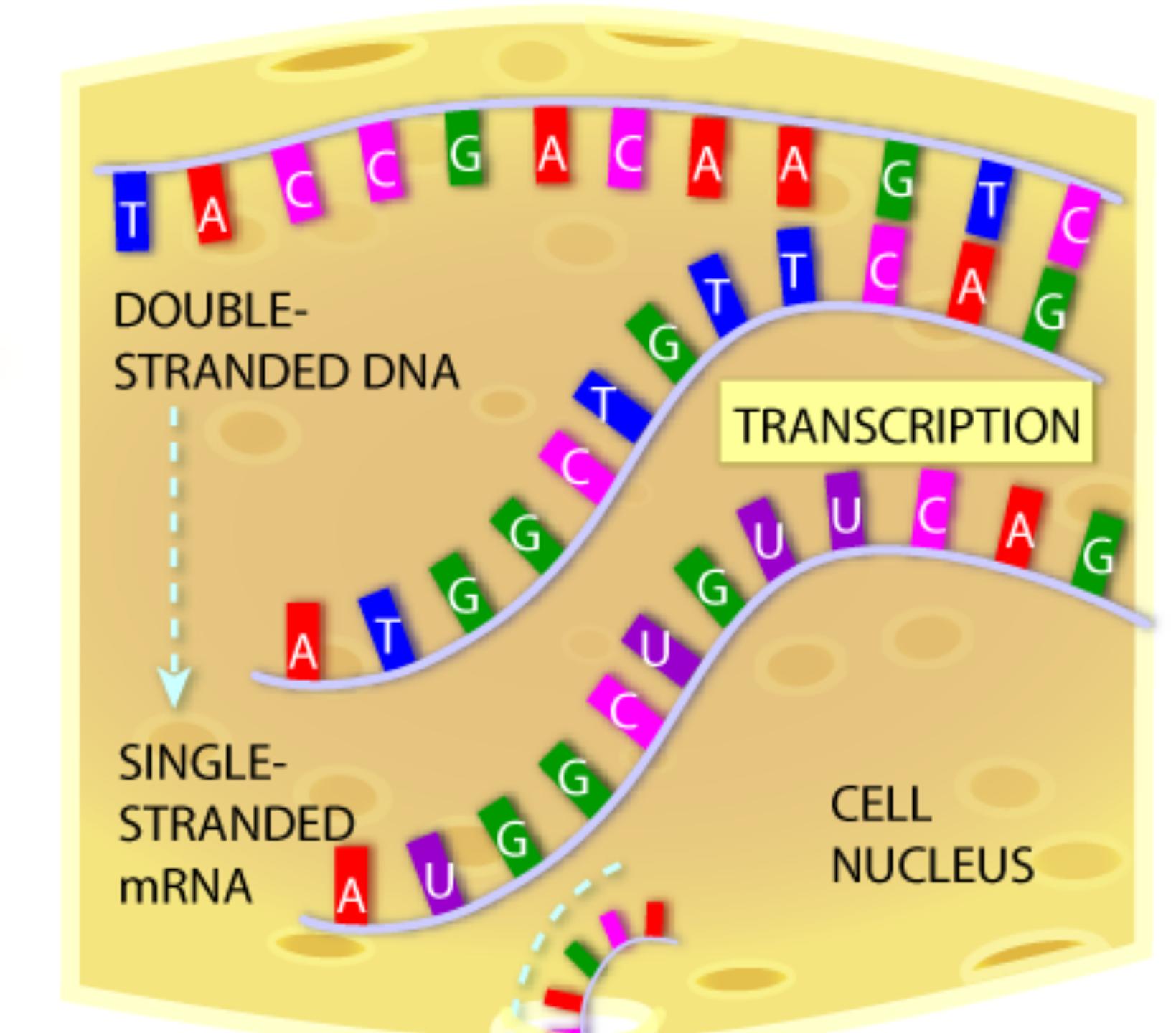


Cold
Spring
Harbor
Laboratory

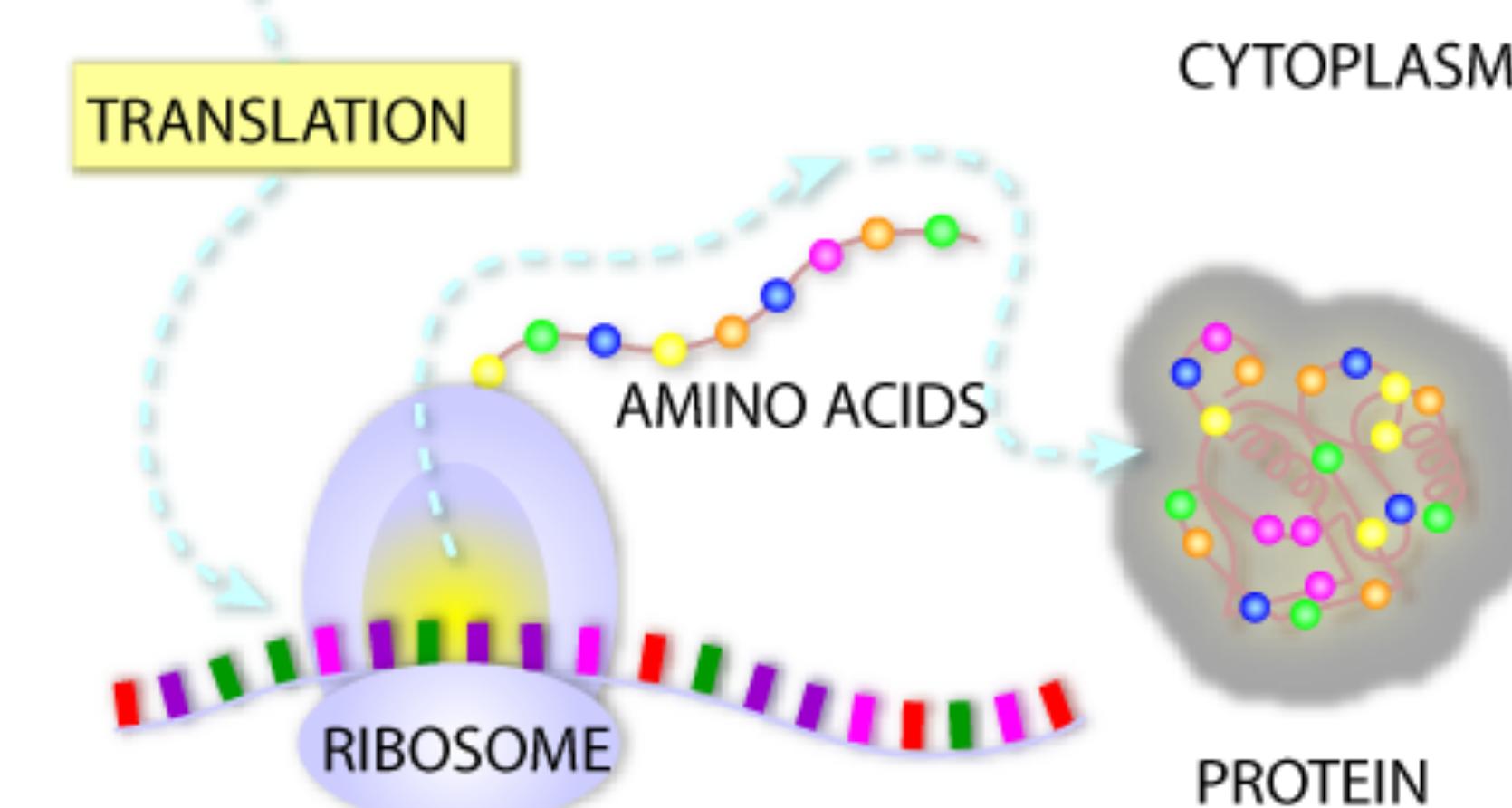
© Copyright, Cold Spring Harbor Laboratory. All rights reserved.

CENTRAL DOGMA OF MOLECULAR BIOLOGY

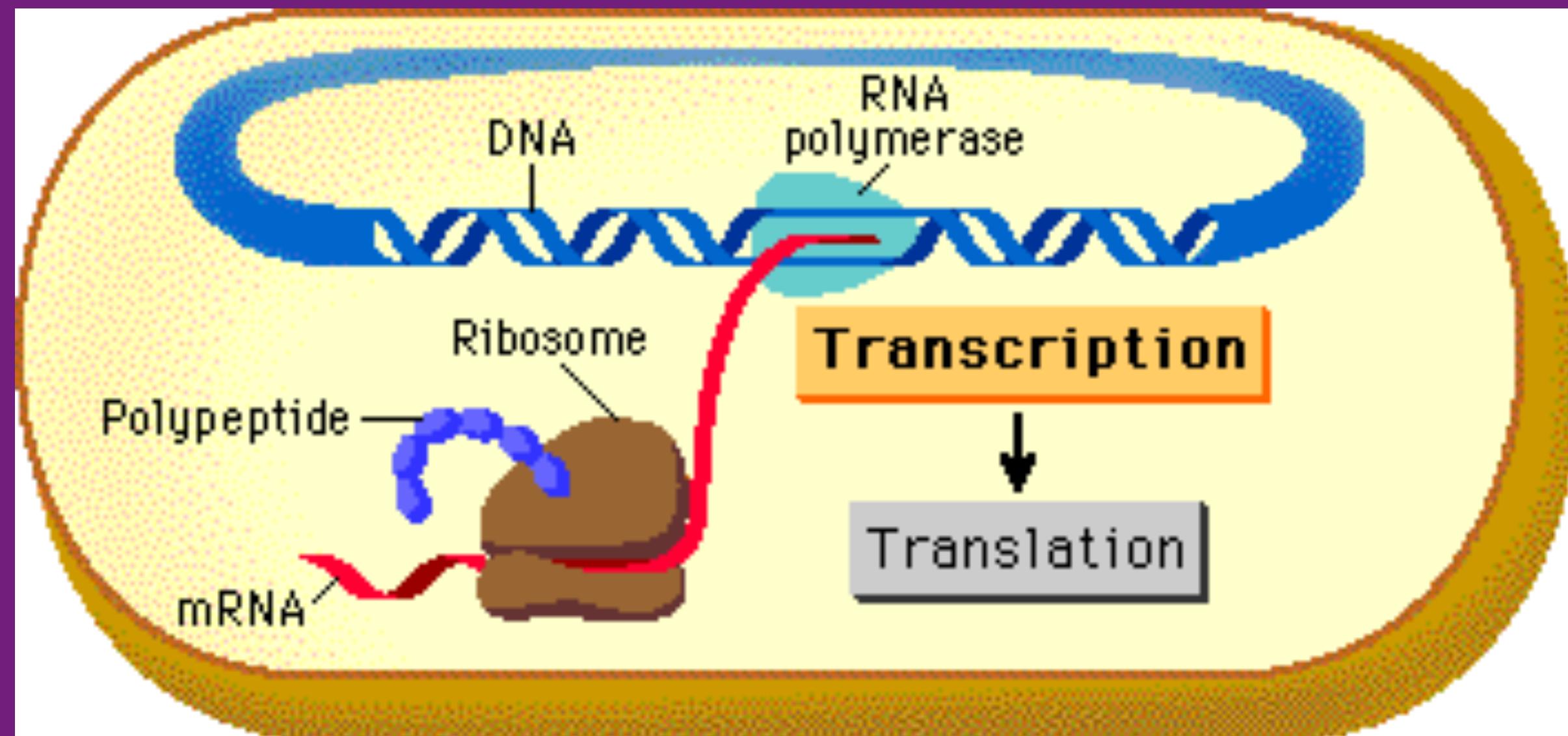
- Transcription
 - Synthesis of an RNA copy of a segment of DNA
 - RNA is synthesized by the enzyme RNA polymerase
- Translation
 - Ribosome reads the mRNA sequence
 - Translates to amino acid sequence of the protein



The mRNA travels from the nucleus to the cytoplasm.



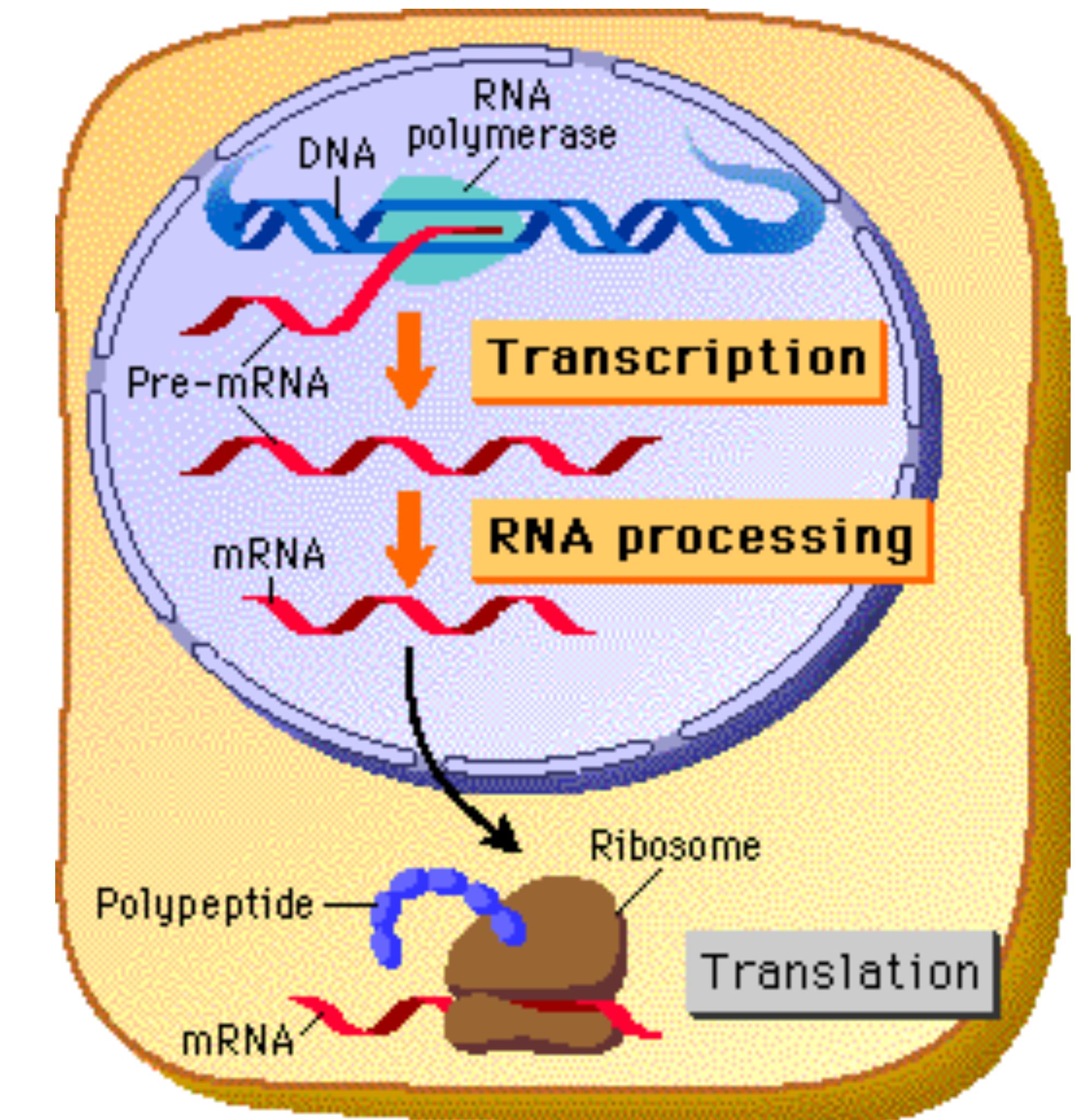
CENTRAL DOGMA OF MOLECULAR BIOLOGY



- 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

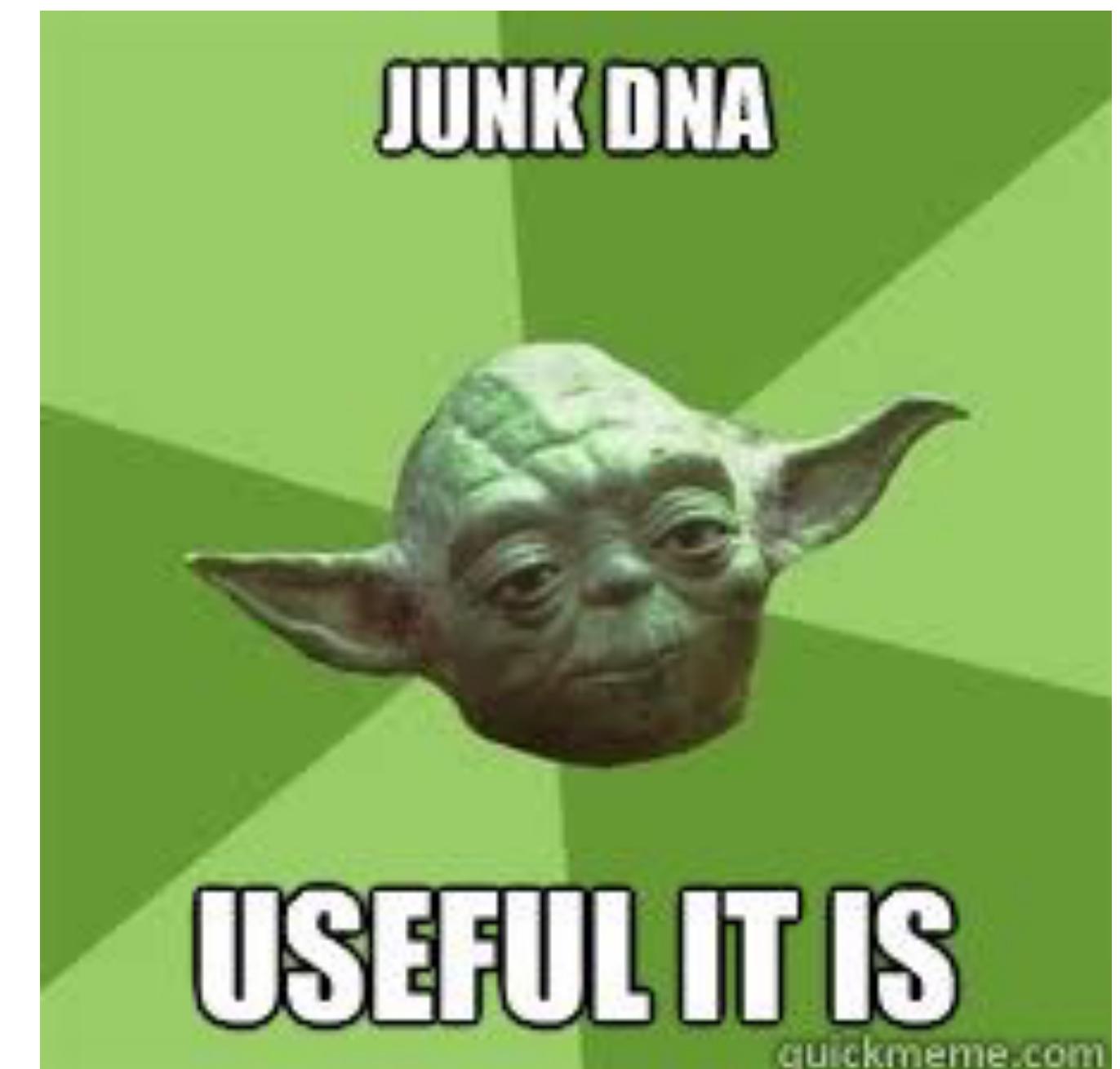
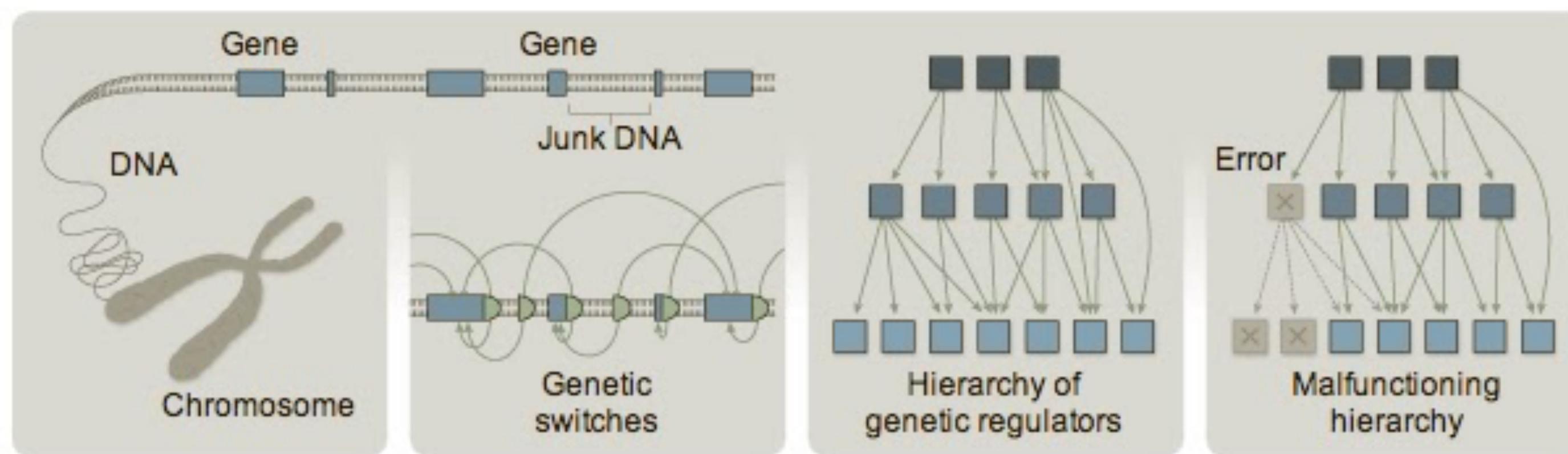
Just because of you don't understand,
you can't call us “Junk”!



CENTRAL DOGMA OF MOLECULAR BIOLOGY

Rethinking 'Junk' DNA

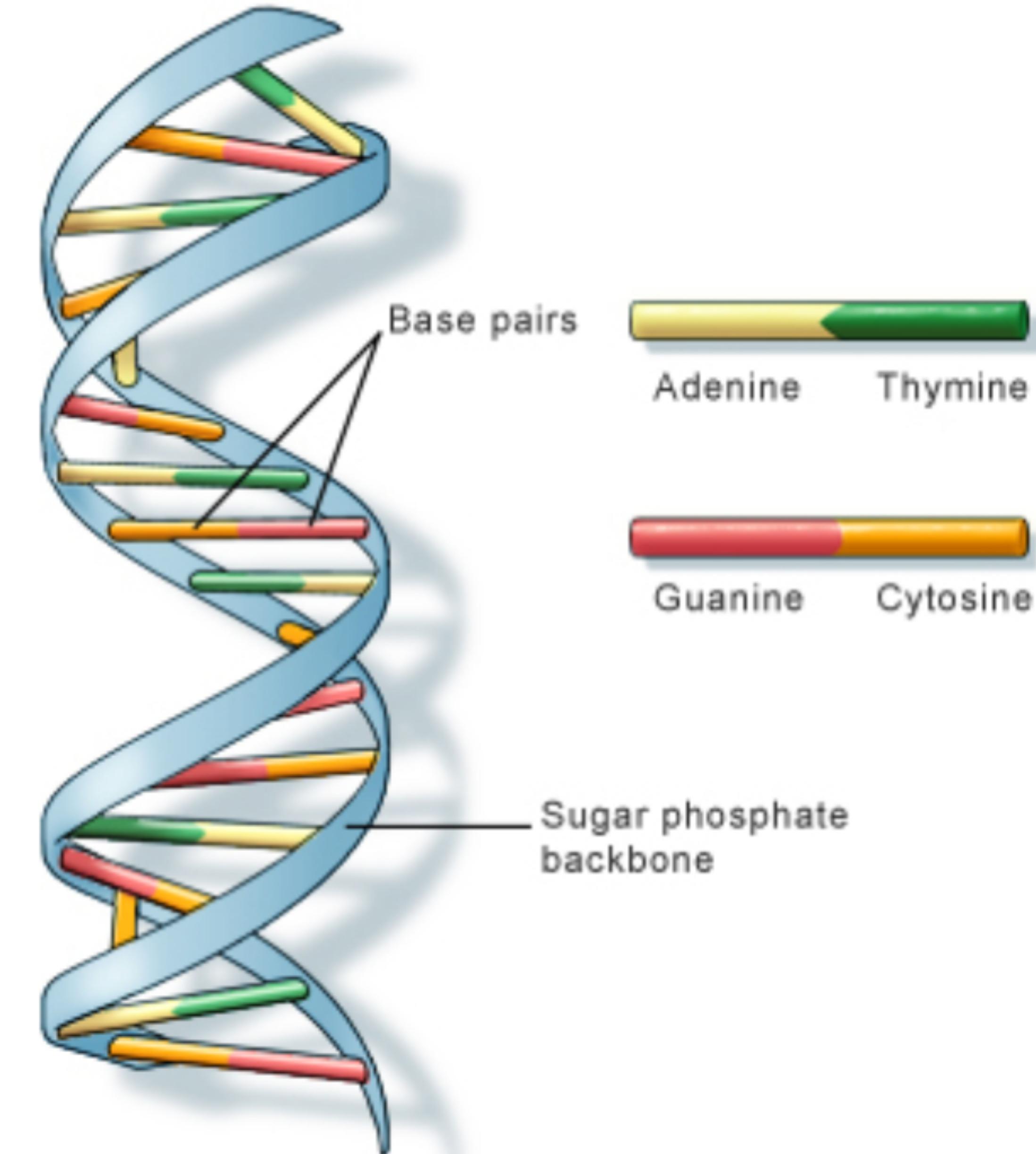
A large group of scientists has found that so-called junk DNA, which makes up most of the human genome, does much more than previously thought. [Related Article »](#)



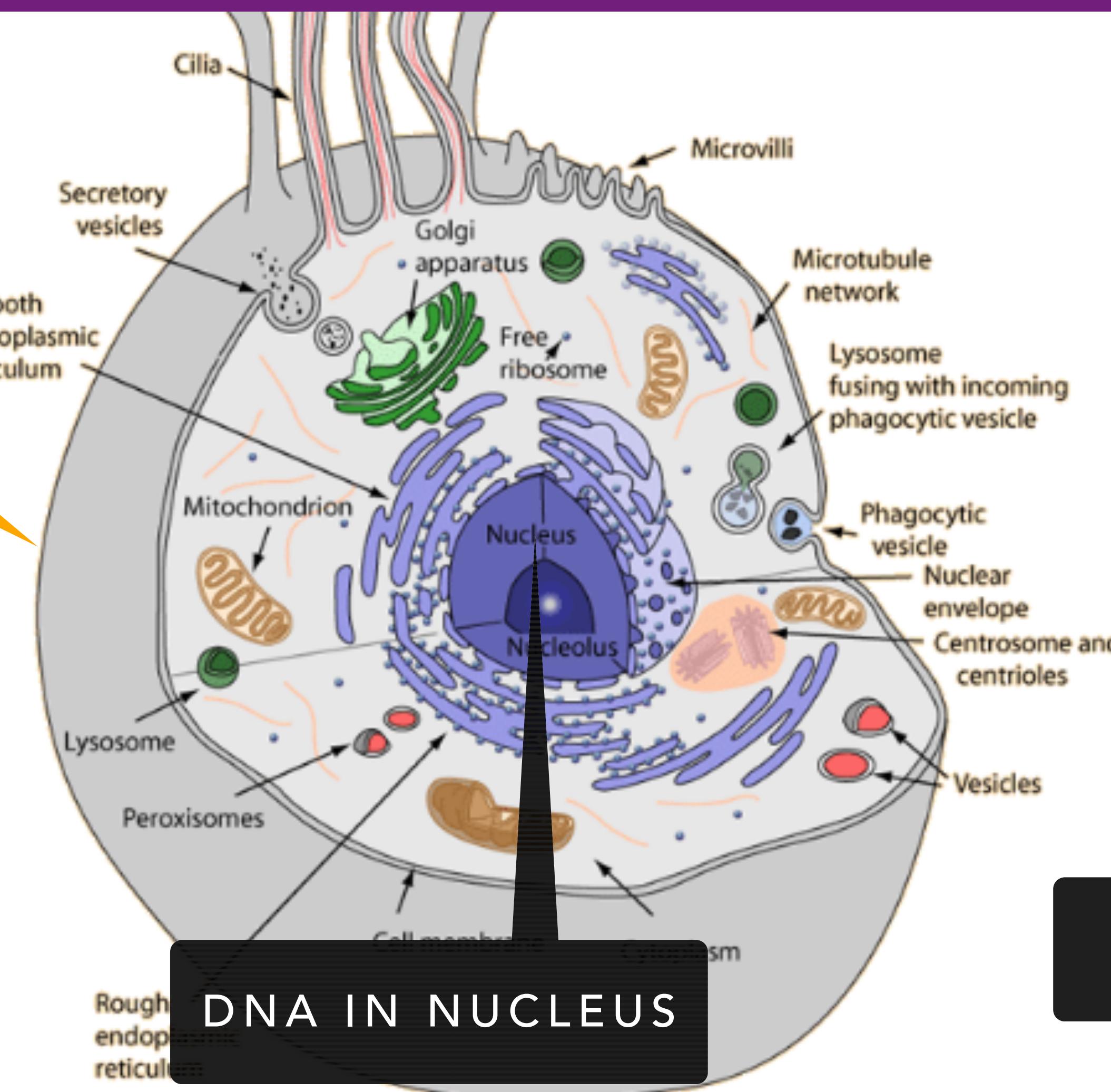
DNA

DNA

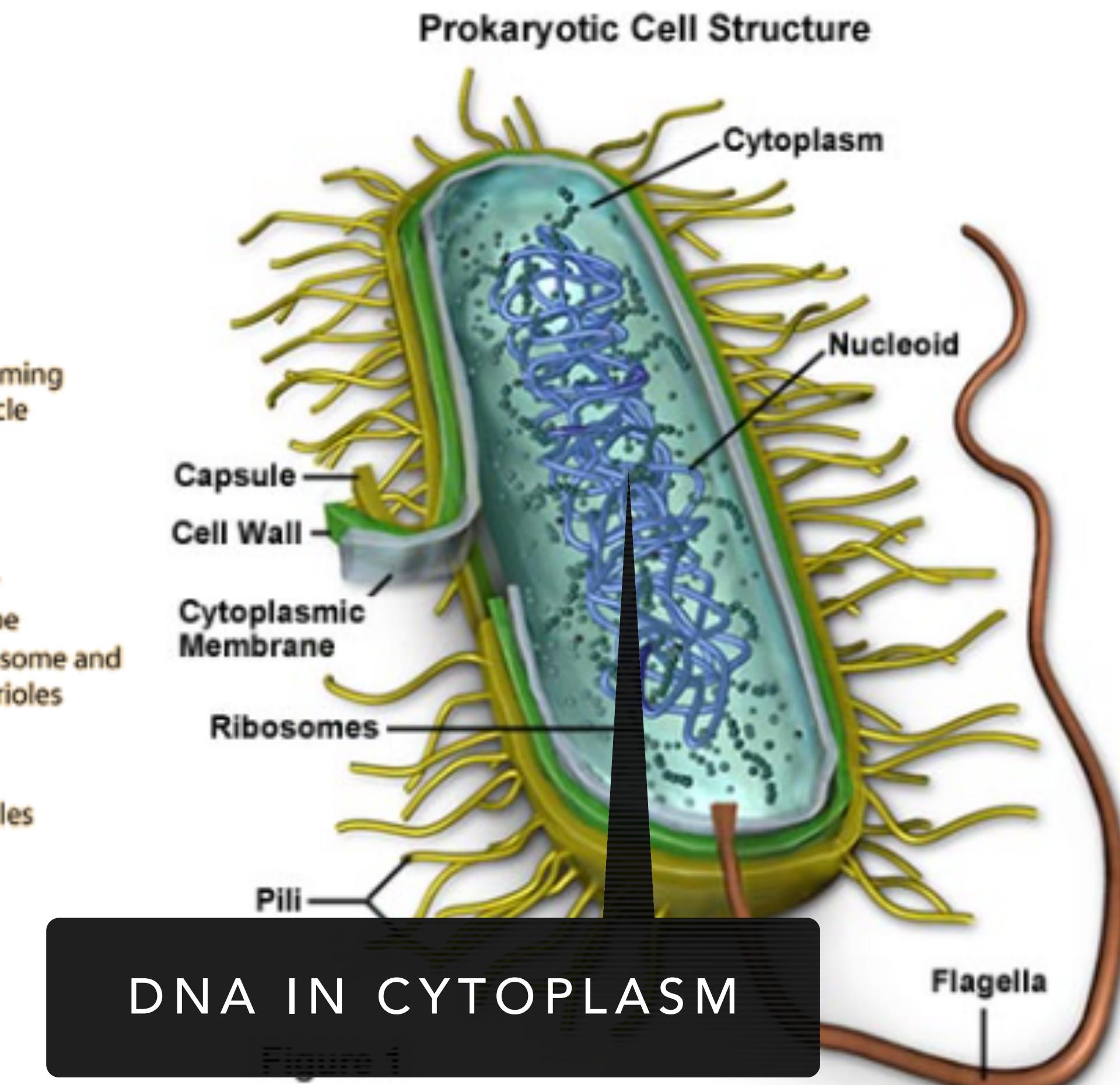
- DNA (**d**eoxyribonucleic **a**cid)
 - Hereditary material in organisms
 - Nearly every cell in a person's body has the same DNA



DNA



Animal cell structure

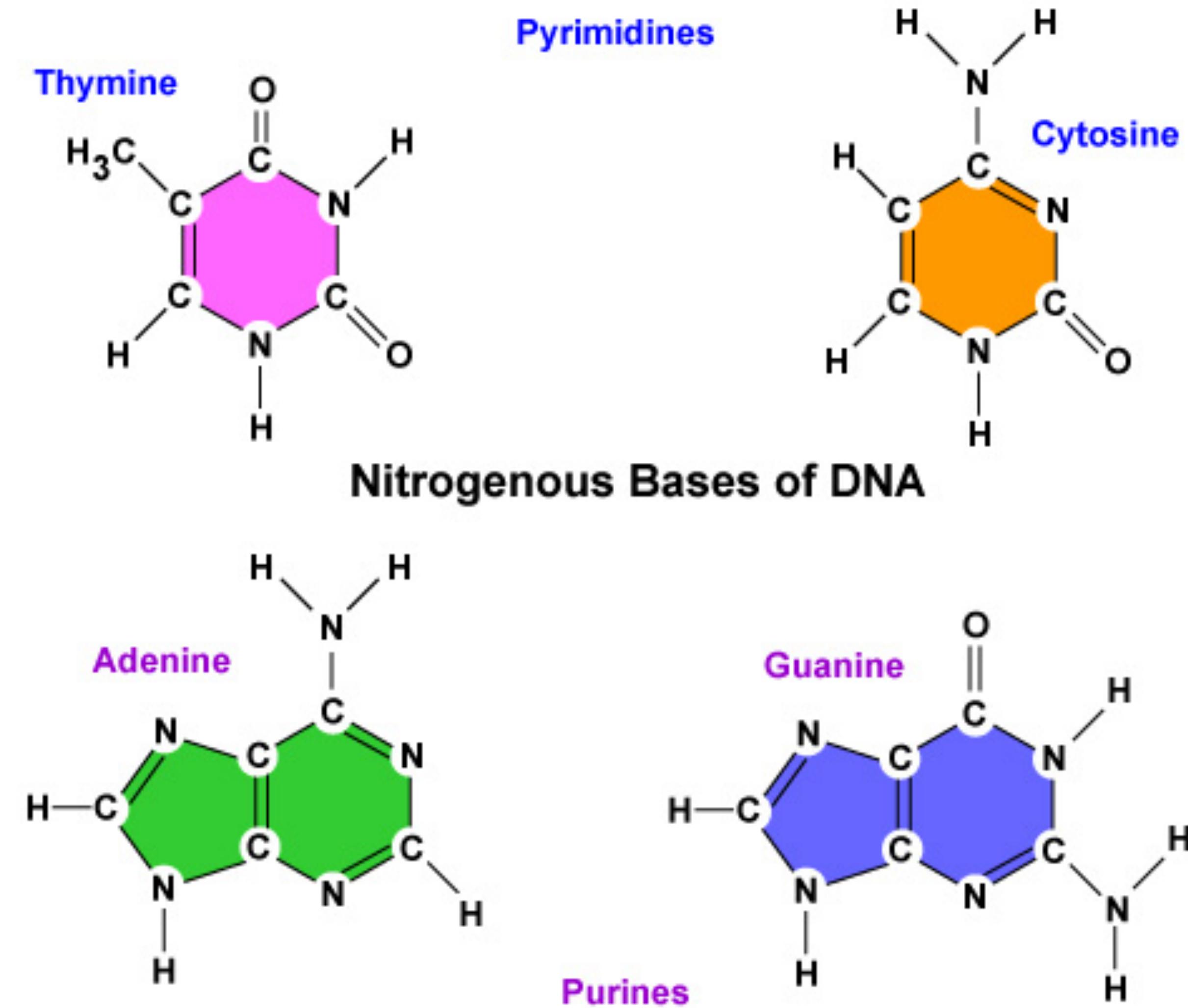


Prokaryotic cell structure

Figure 1

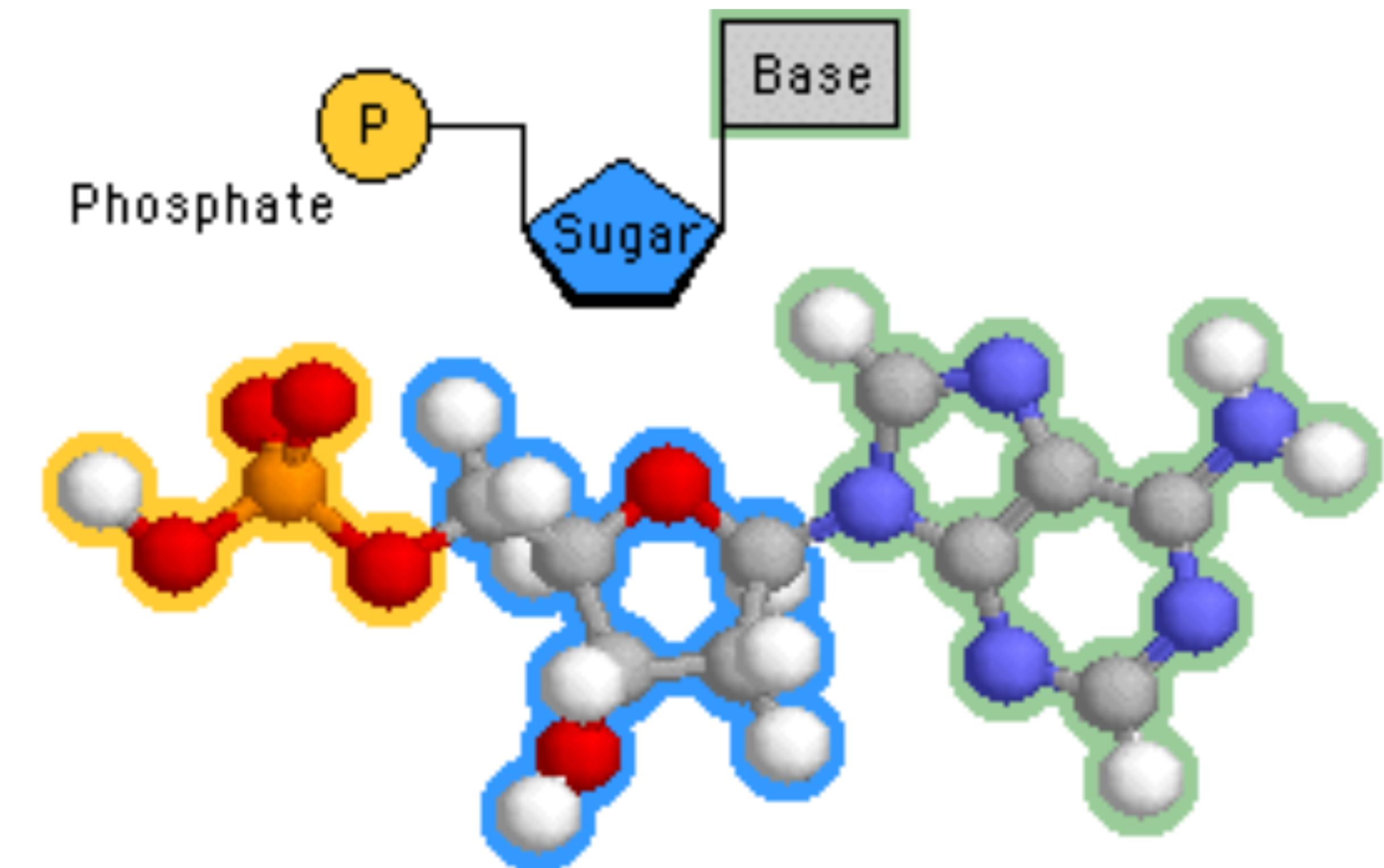
DNA

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

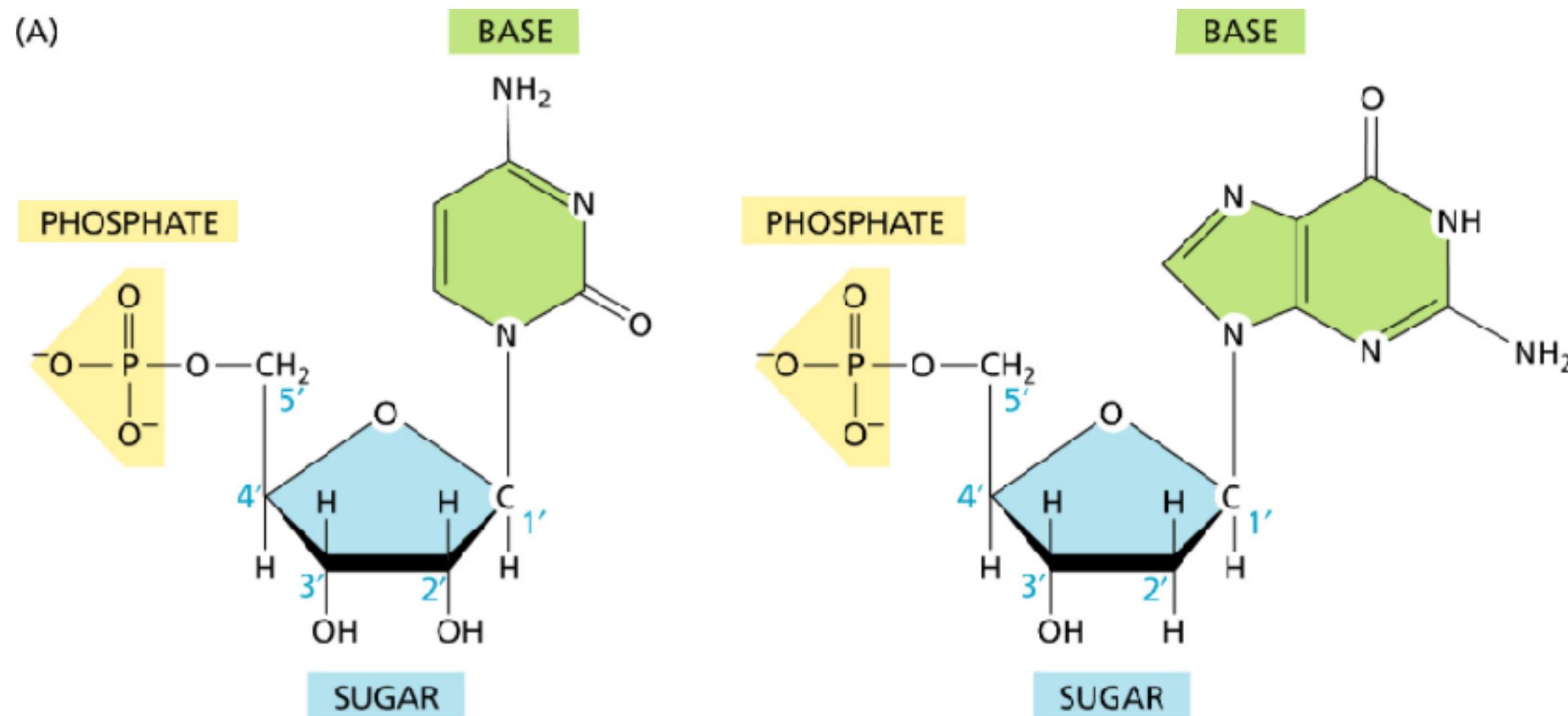


DNA

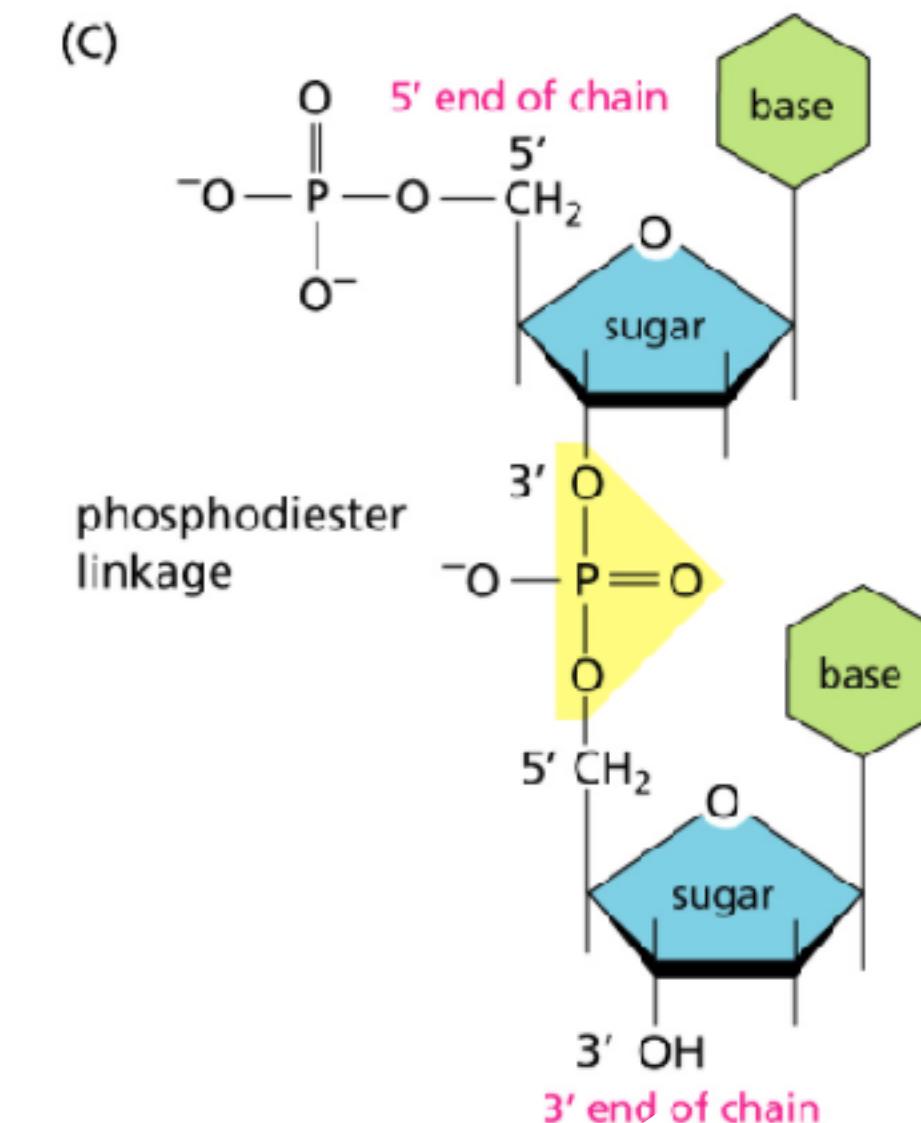
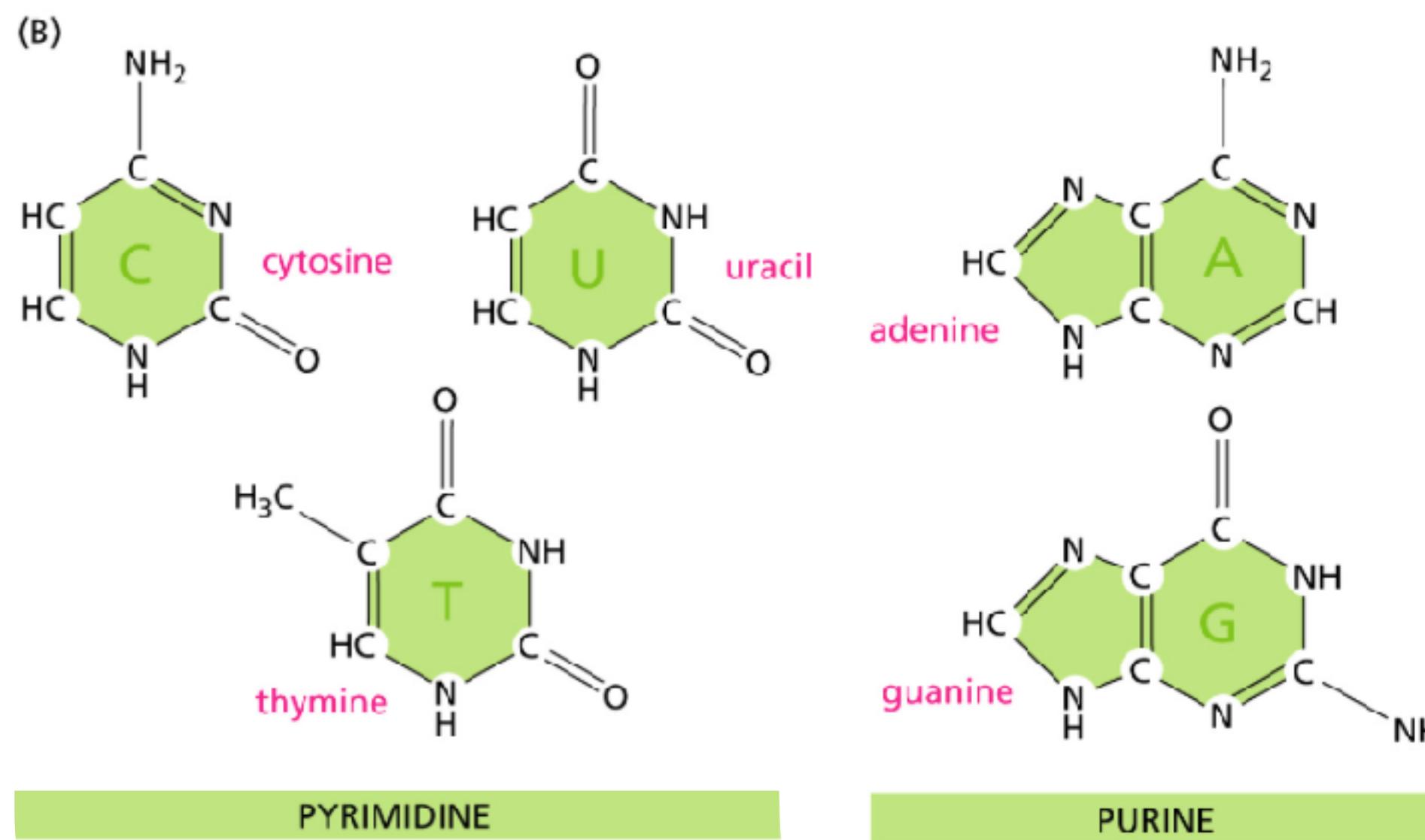
- Each nucleotide has three components
 - 1 or more phosphate groups
 - 5-carbon, or pentose, sugar
 - Nucleotide



DNA



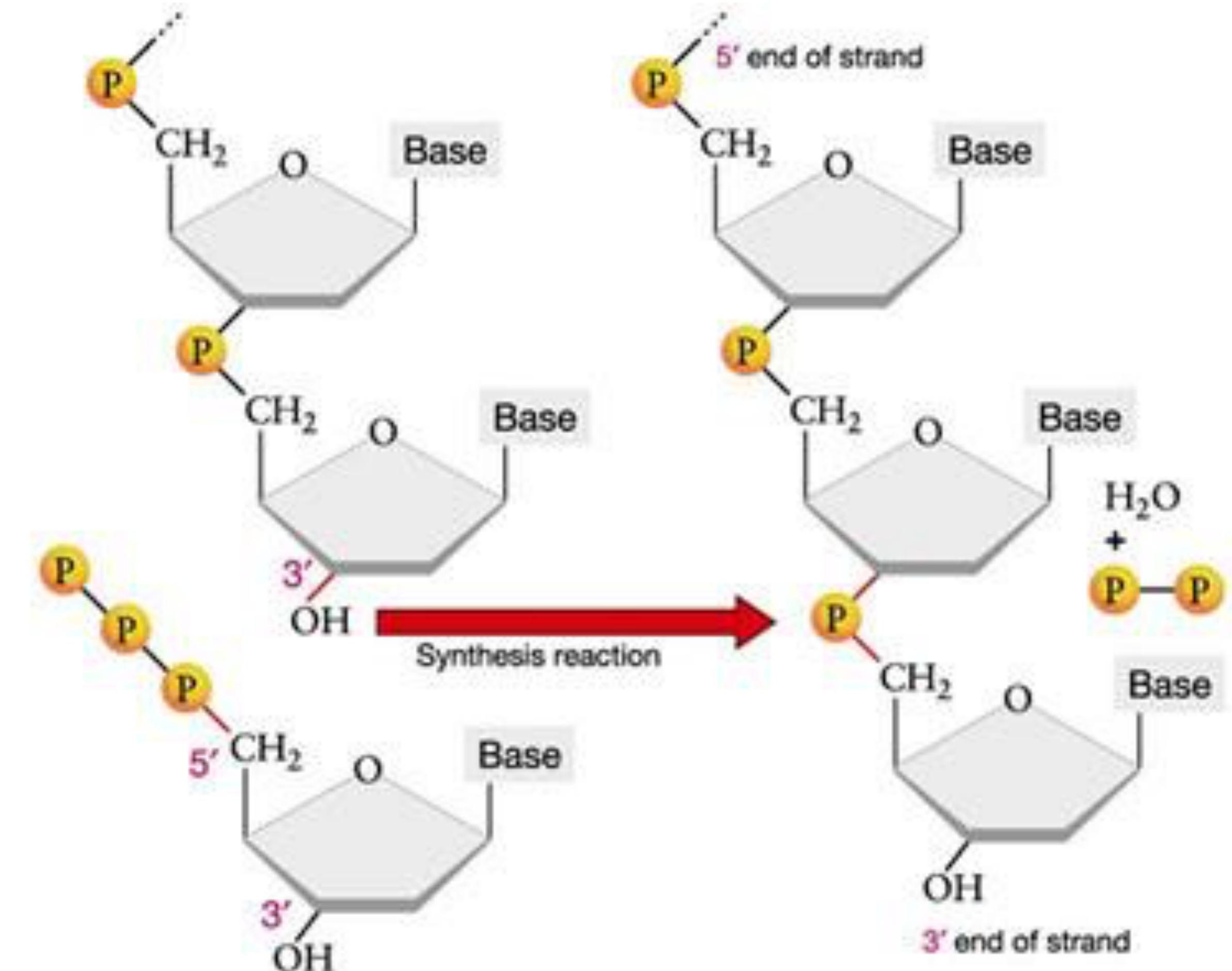
- Purine or pyrimidine bases
- 5-carbon (pentose) sugar
- 1 or more phosphate groups



POLYMERIZATION
INTO CHAIN

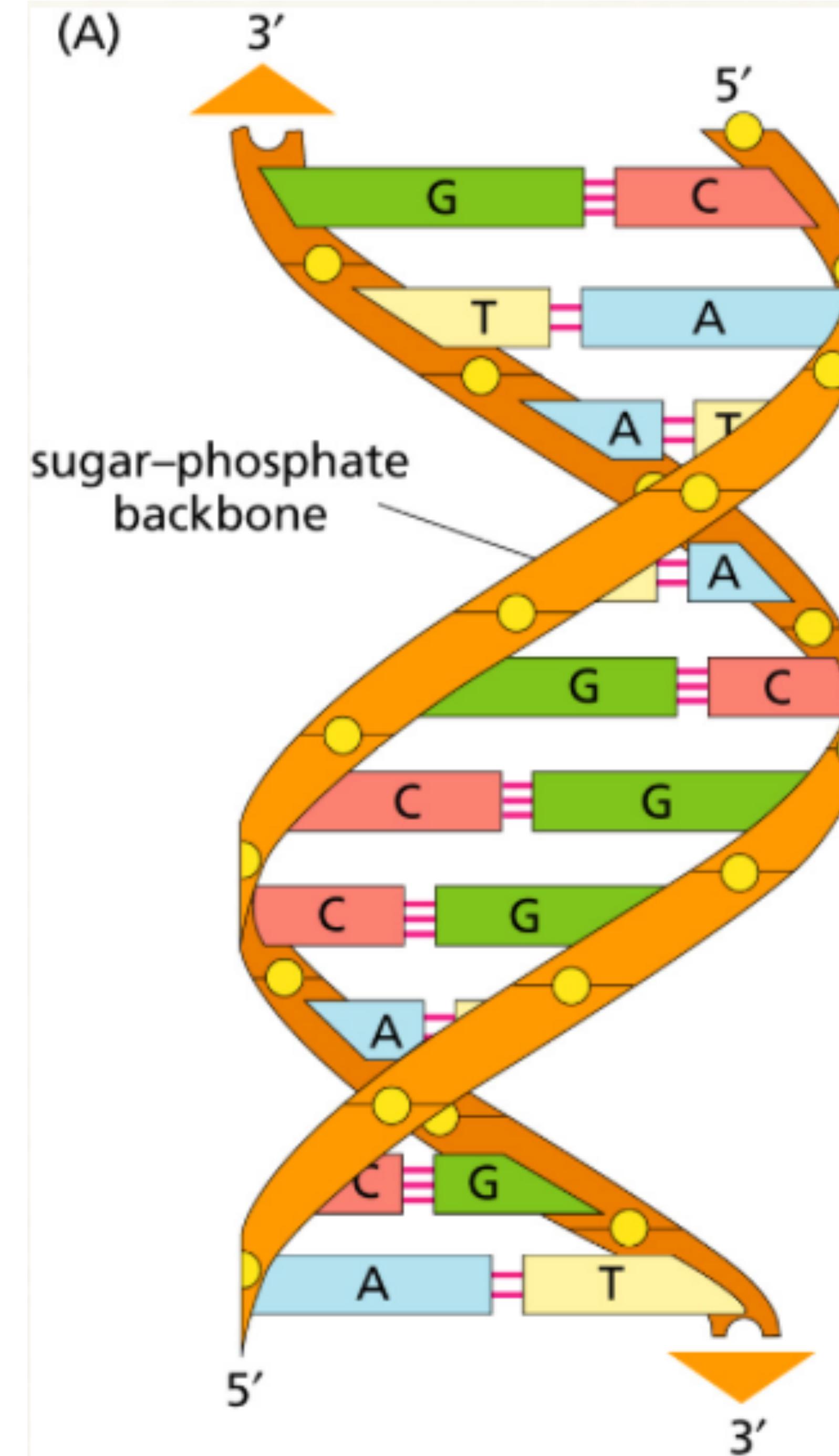
DNA

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



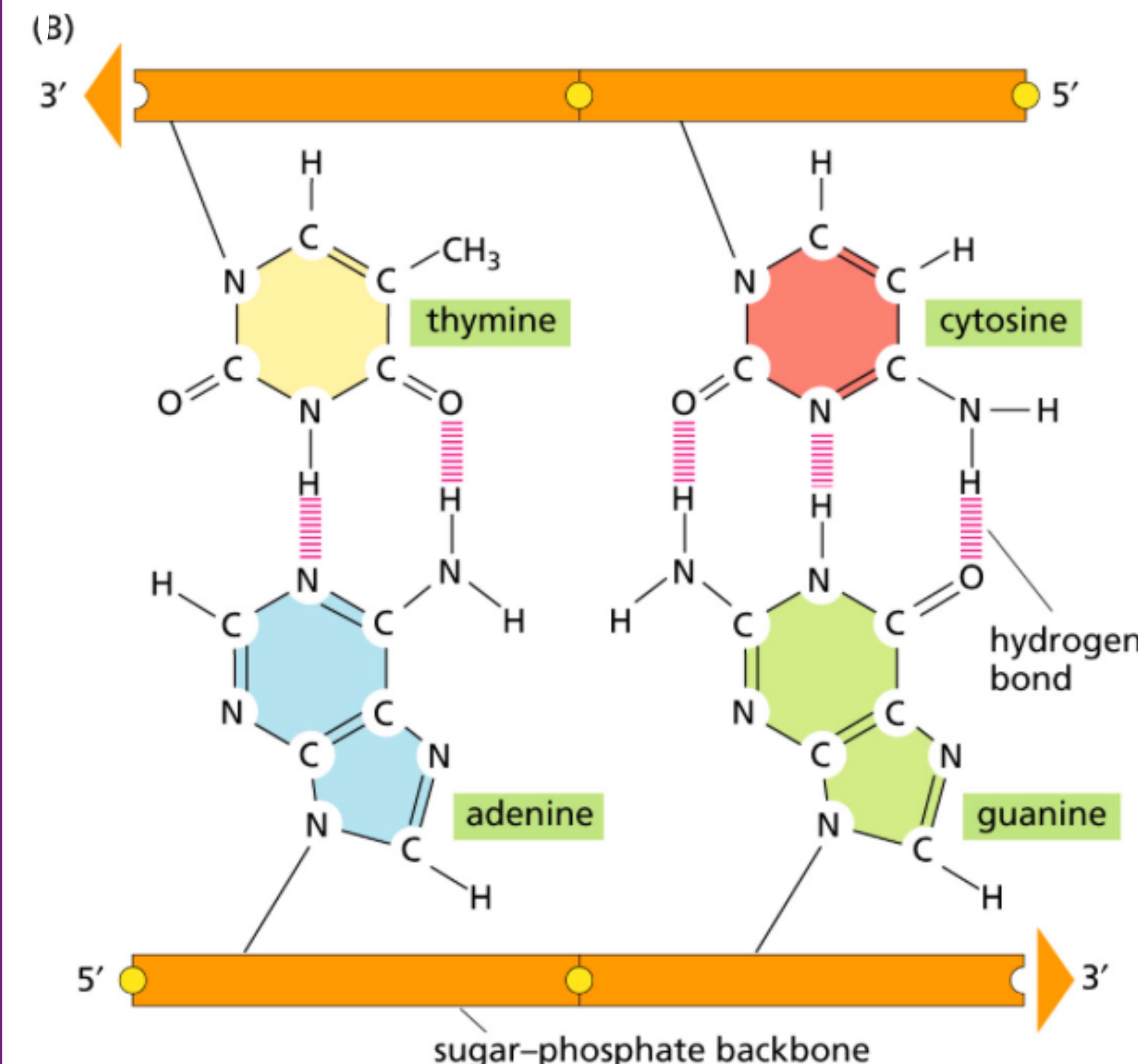
DNA

- 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



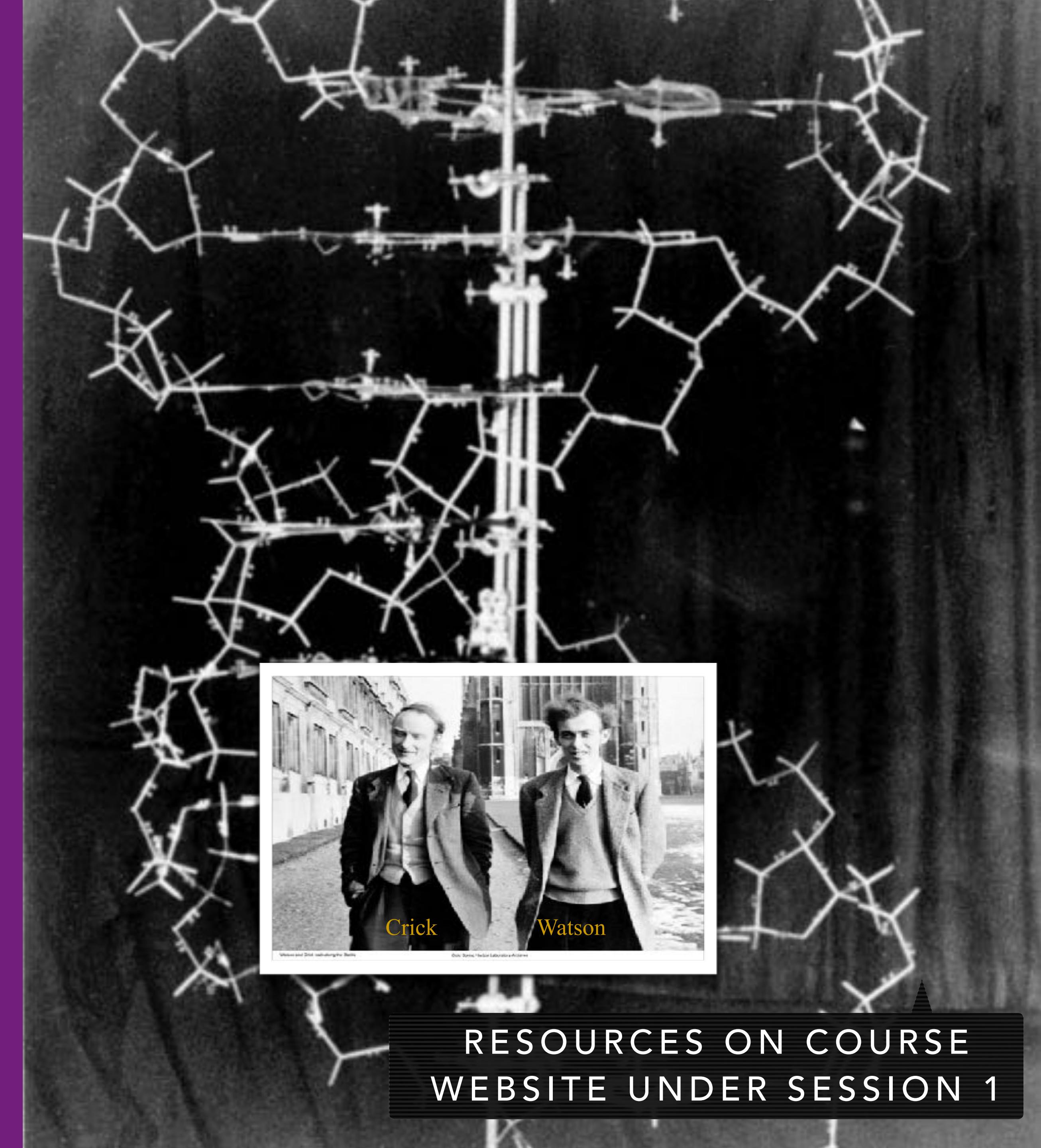
DNA

- 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 than AT)



DISCOVERY OF DNA

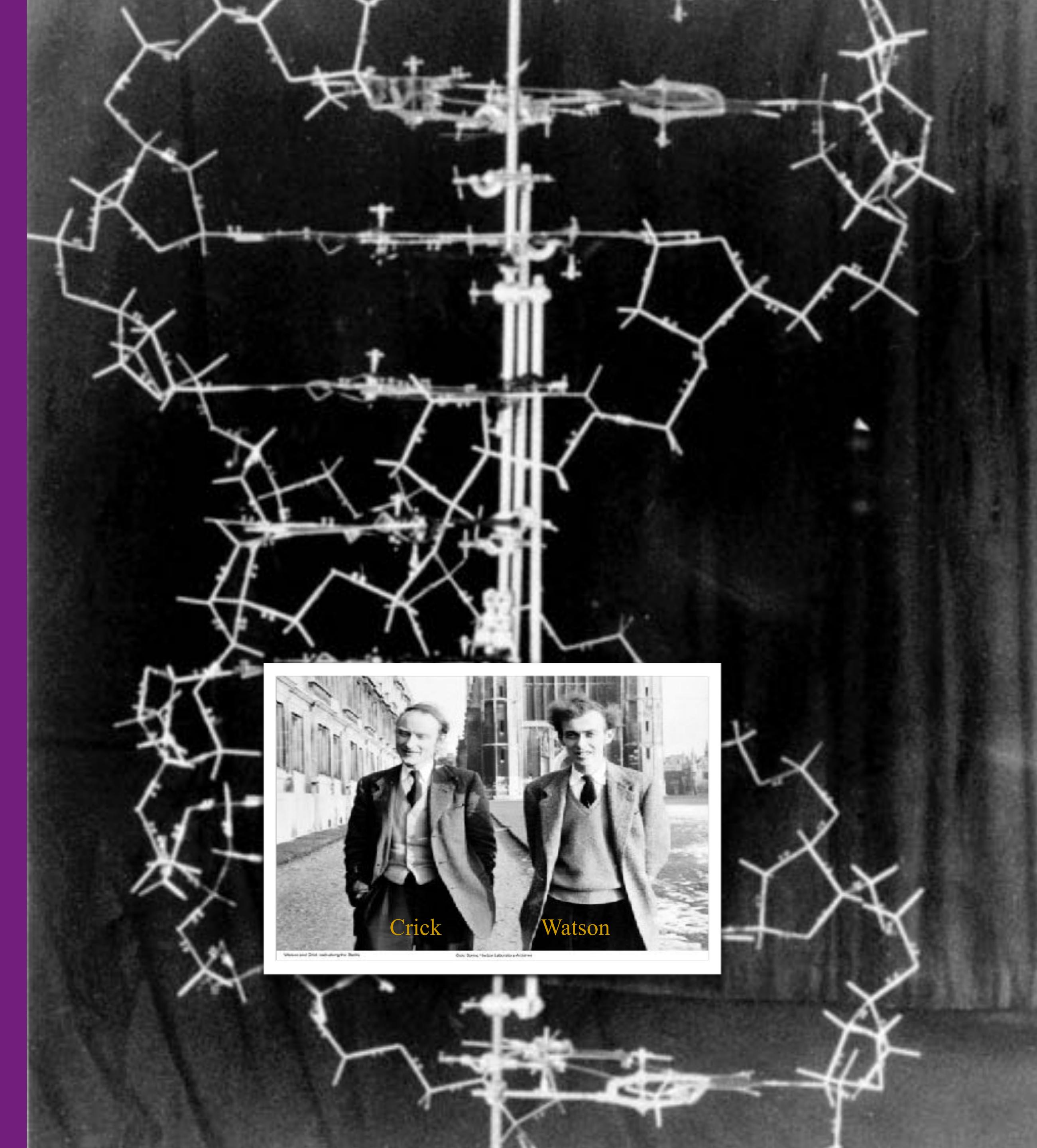
- DNA Sequences
 - Chargaff and Vischer, 1949
 - DNA consisting of A, T, G, C
 - Chargaff Rule ($\#A \approx \#T$ and $\#G \approx \#C$)
 - A “strange but possibly meaningless” phenomenon.



RESOURCES ON COURSE
WEBSITE UNDER SESSION 1

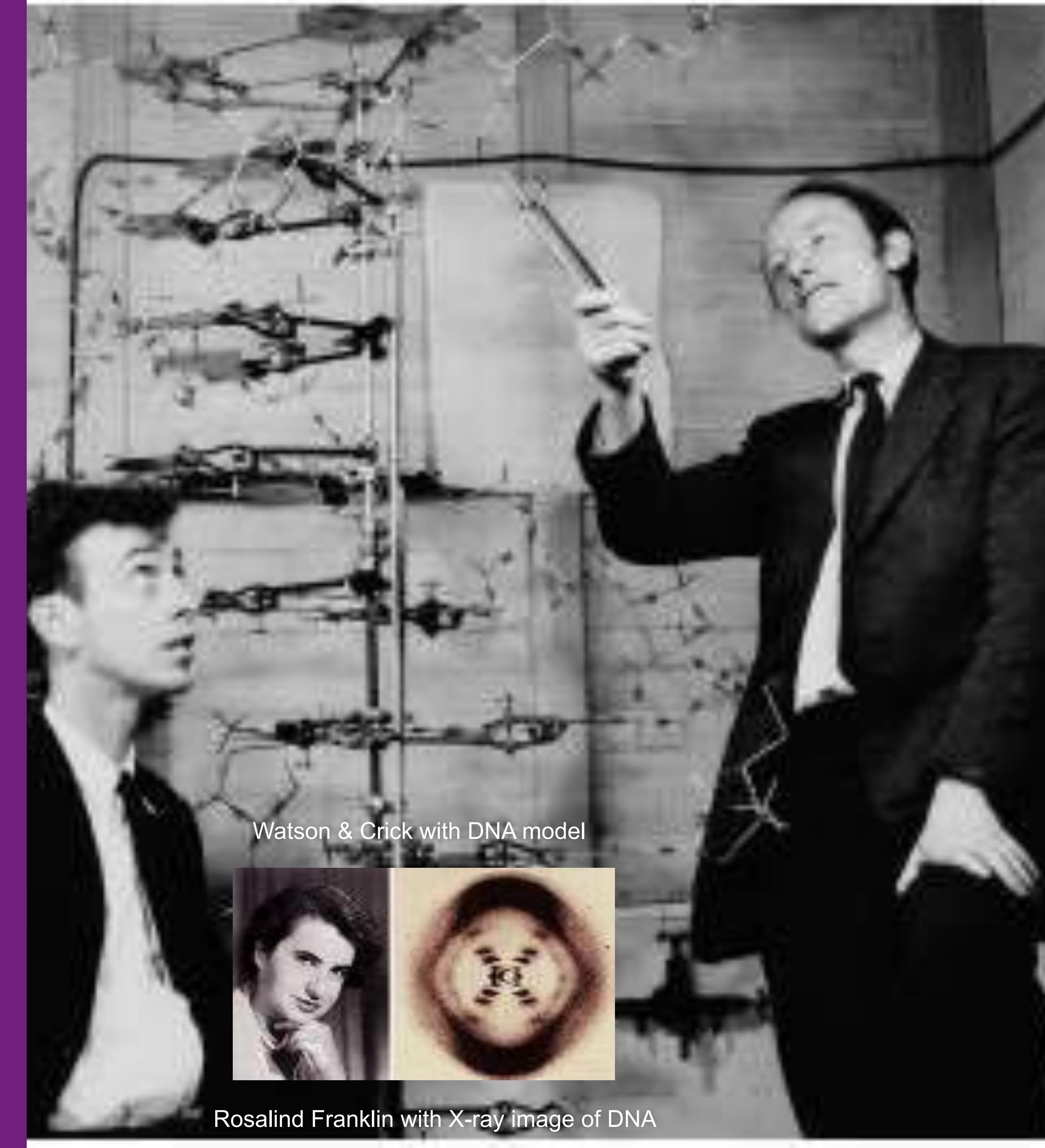
DISCOVERY OF DNA

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



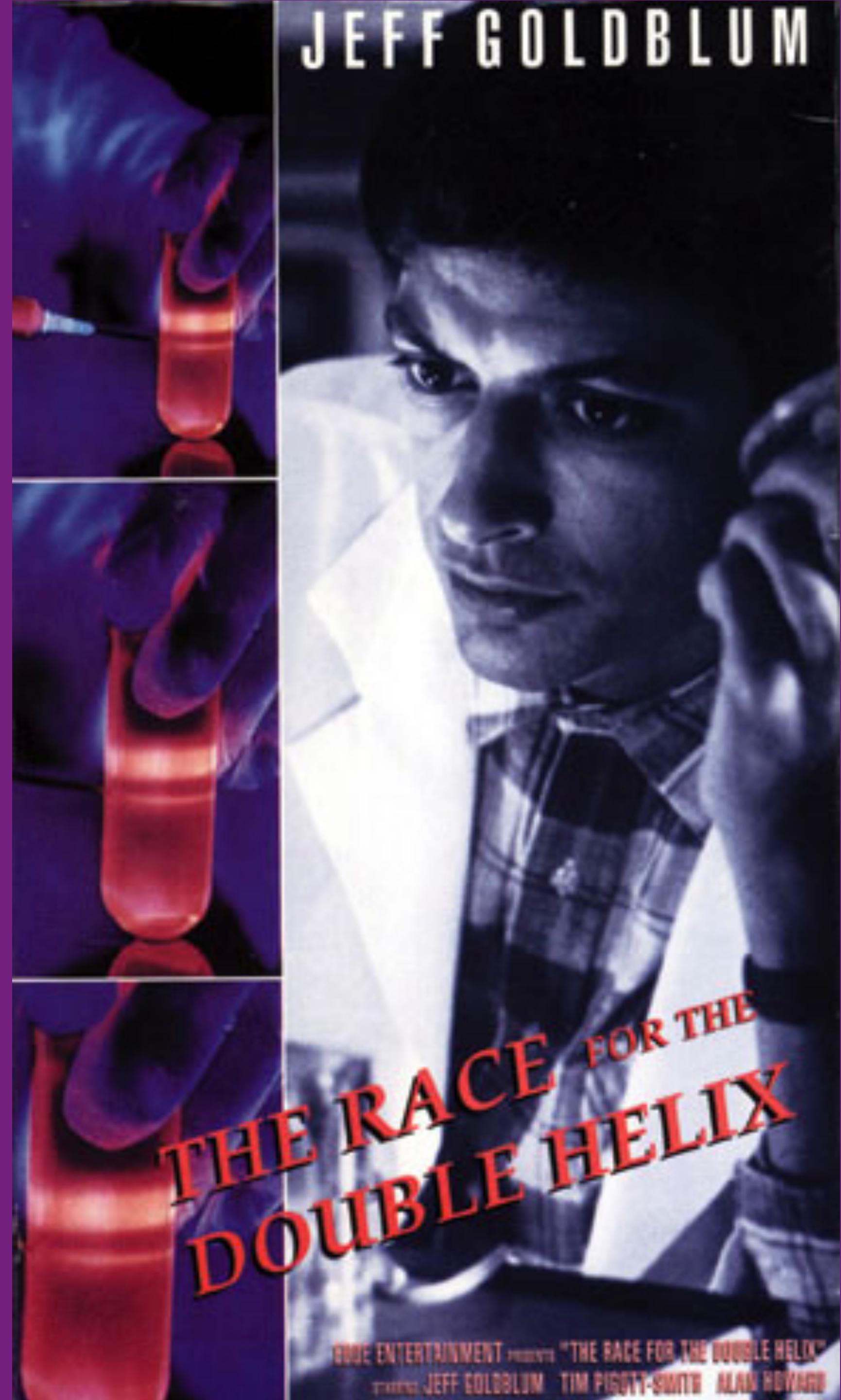
DNA

- 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
 - Wilkins - Showed an X-ray to Watson
- The Discovery
 - Applying Chargaff'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



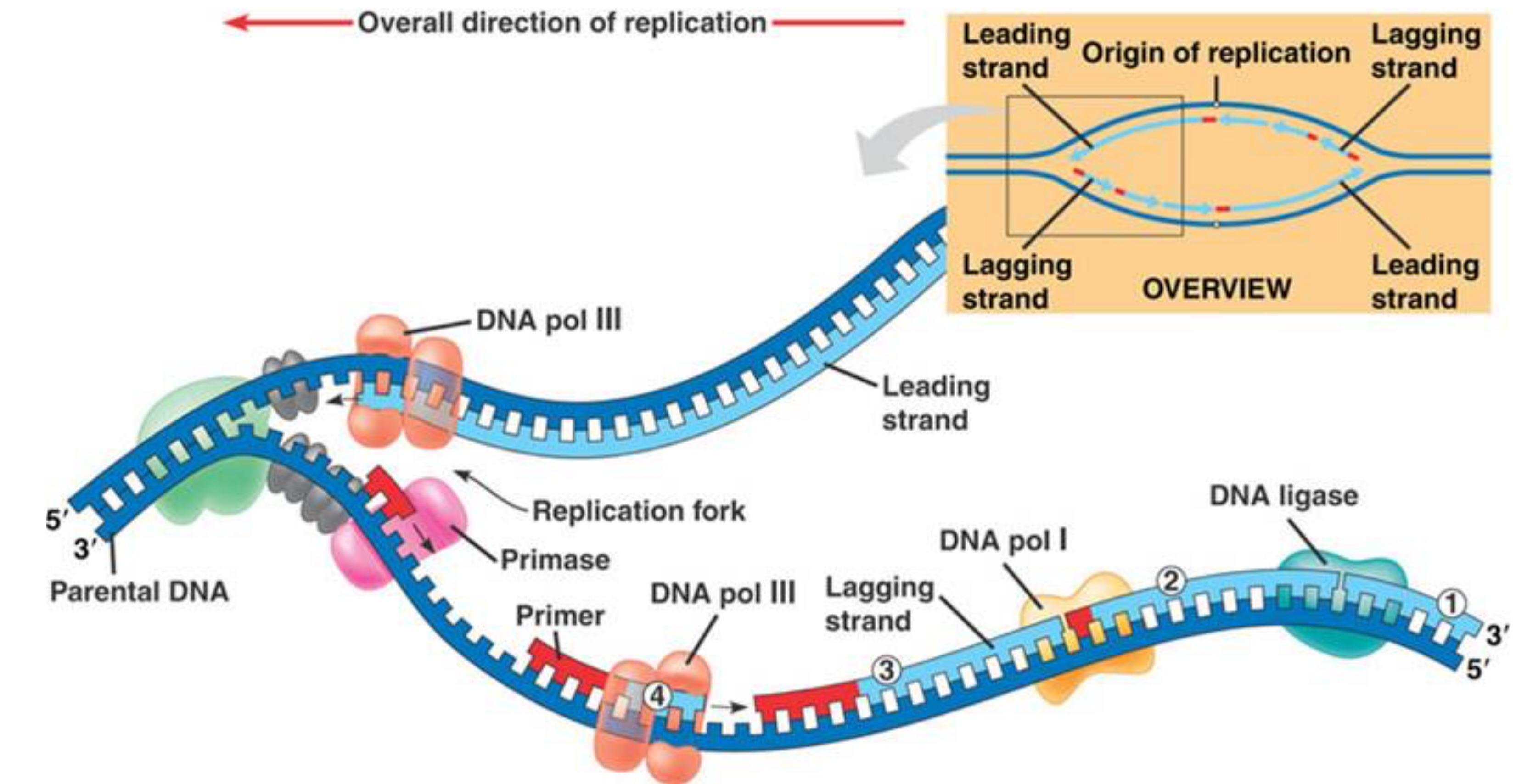
DNA

- Life Story: The Race for the Double Helix



DNA

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

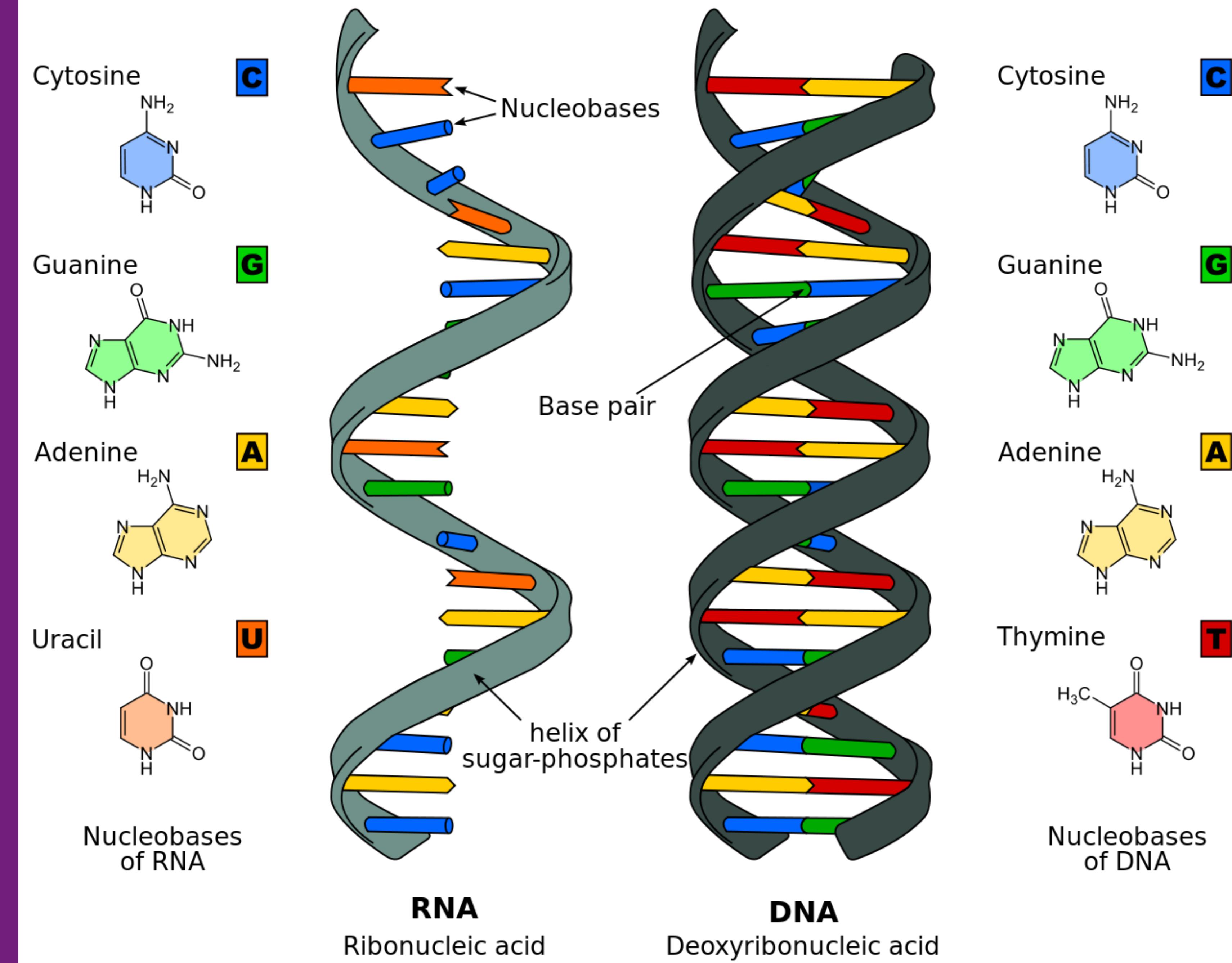


RESOURCES ON COURSE
WEBSITE UNDER SESSION 1

RNA

RNA

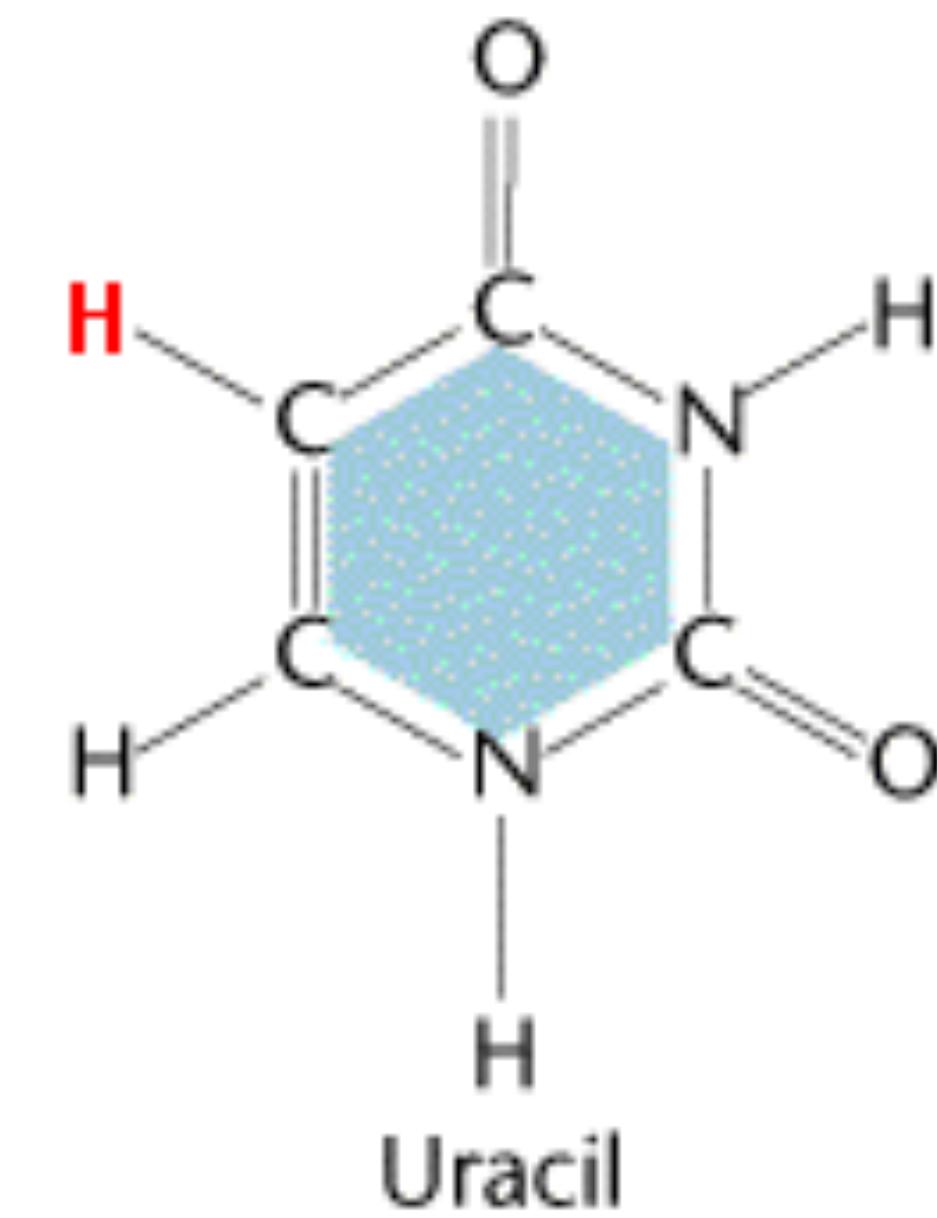
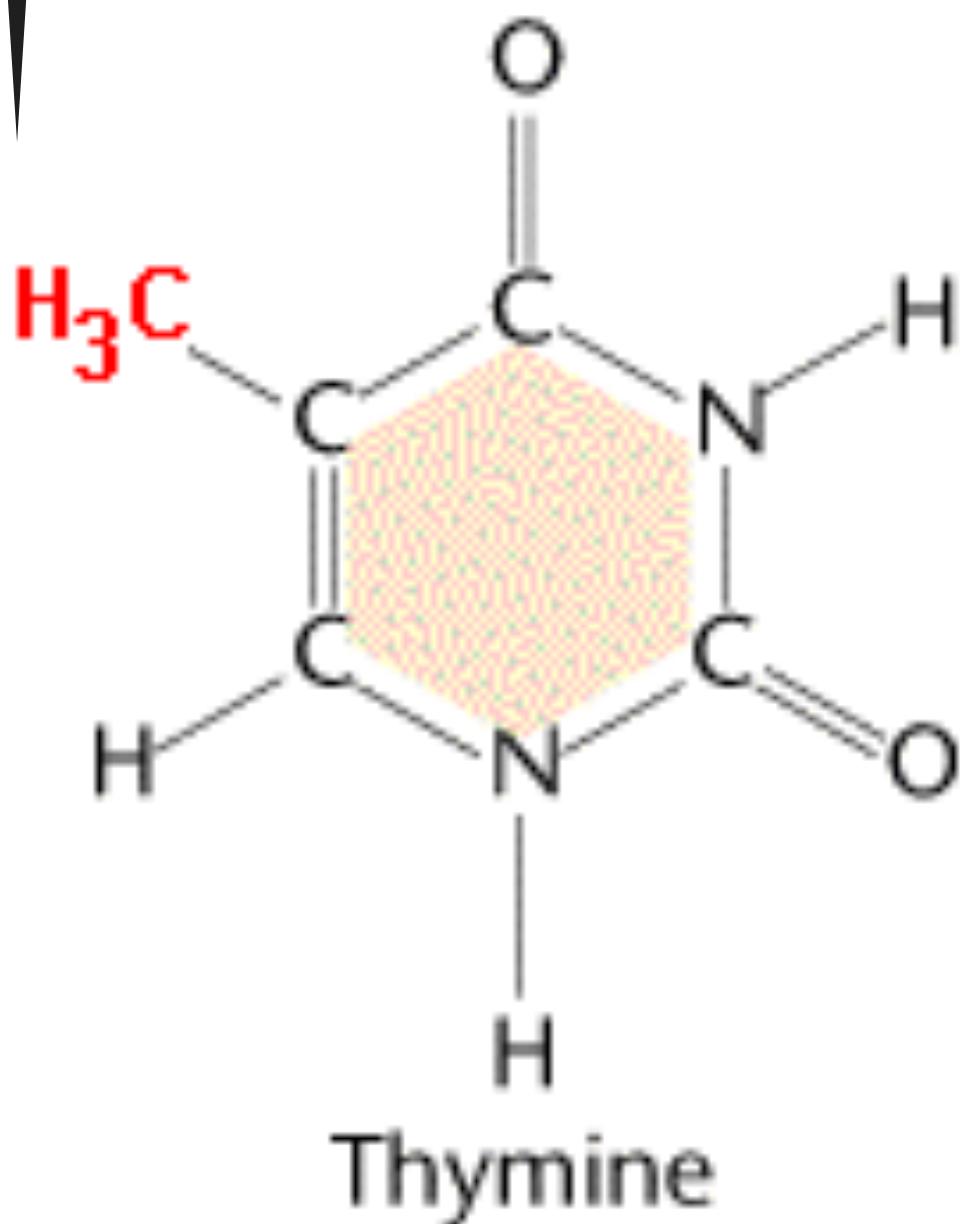
- Ribonucleic acid
- RNA is chemically similar to DNA chemically
 - Different base
 - Different sugar



RNA

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

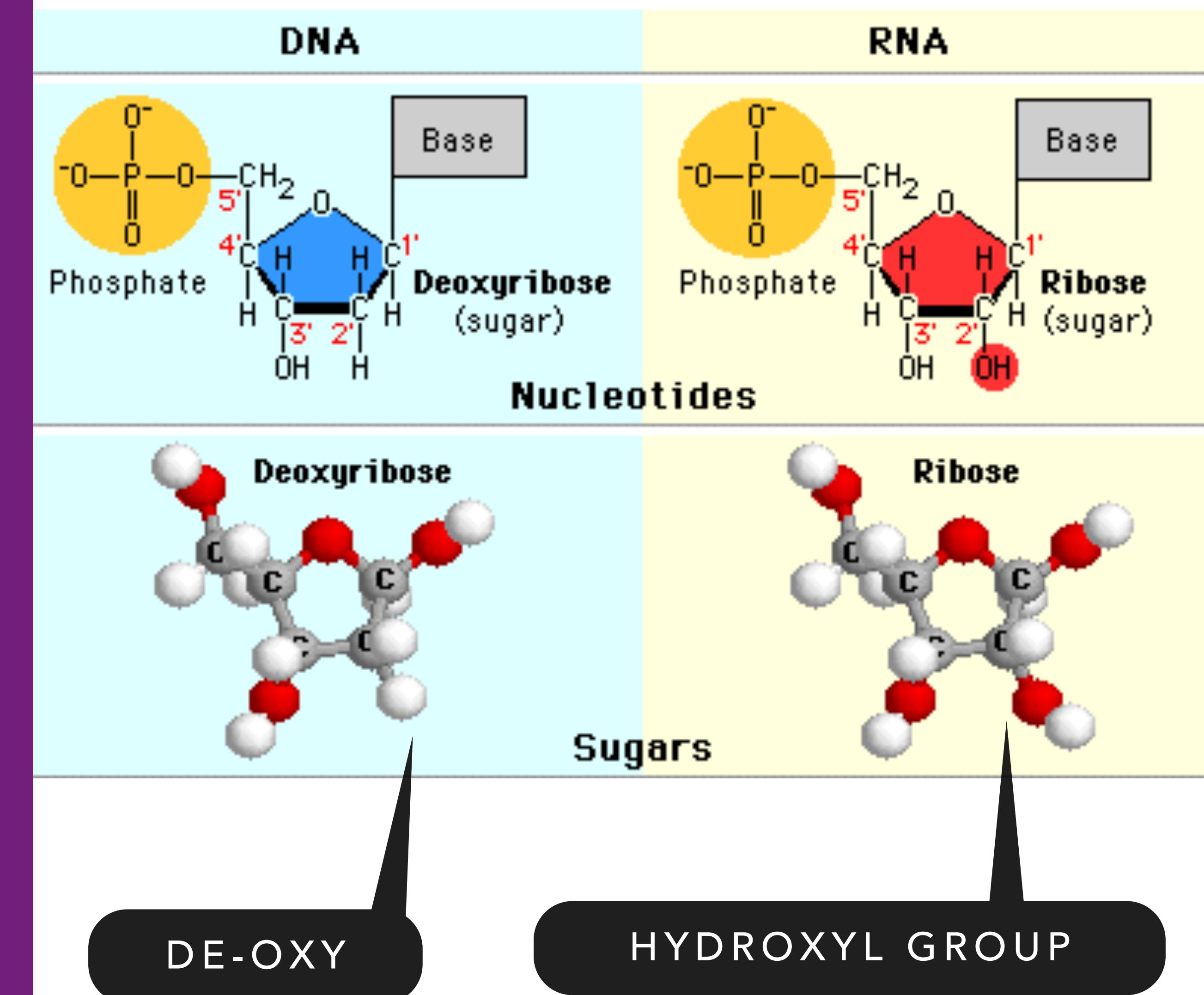
METHYL GROUP



(Klug & Cummings 1997)

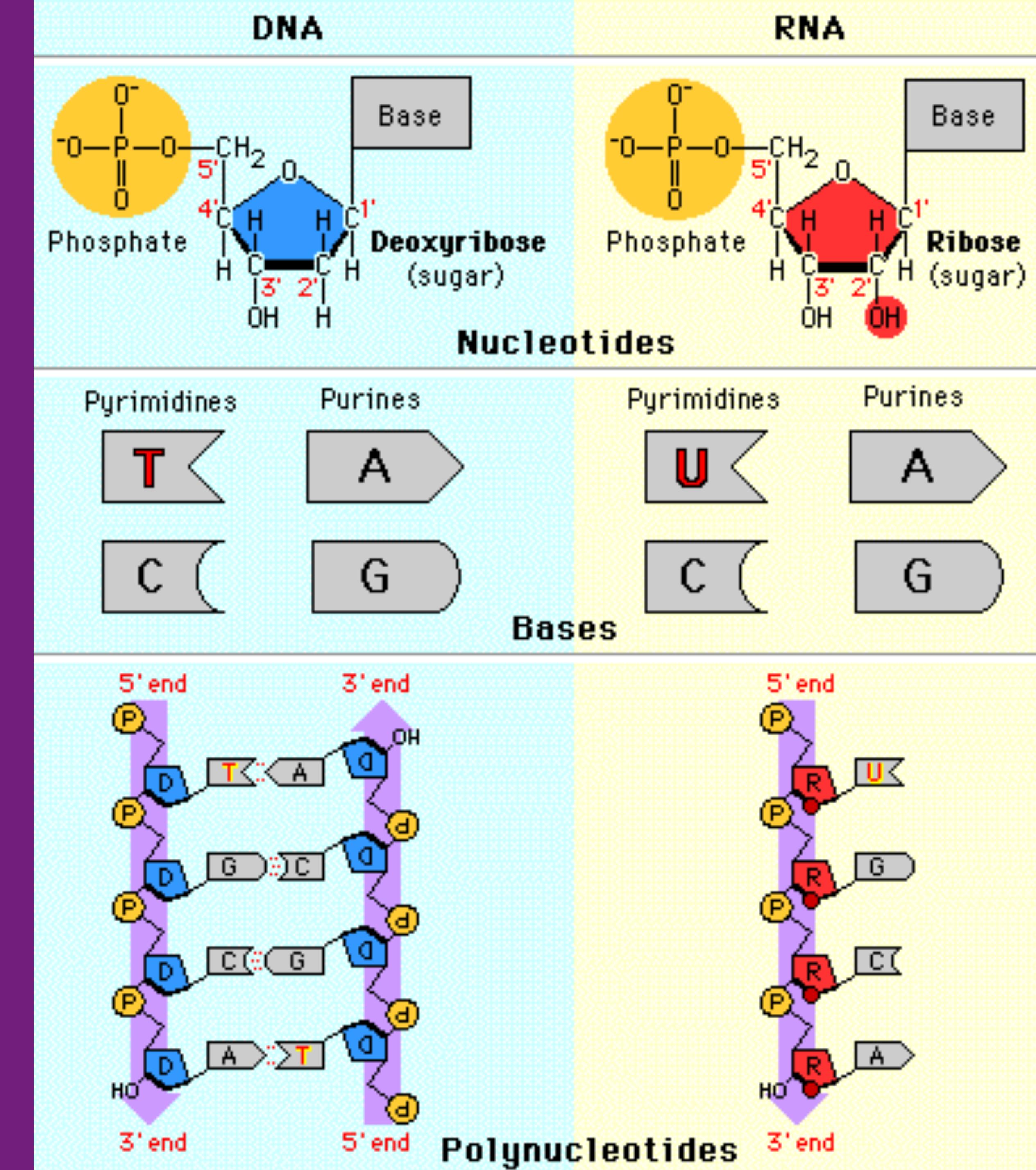
RNA

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



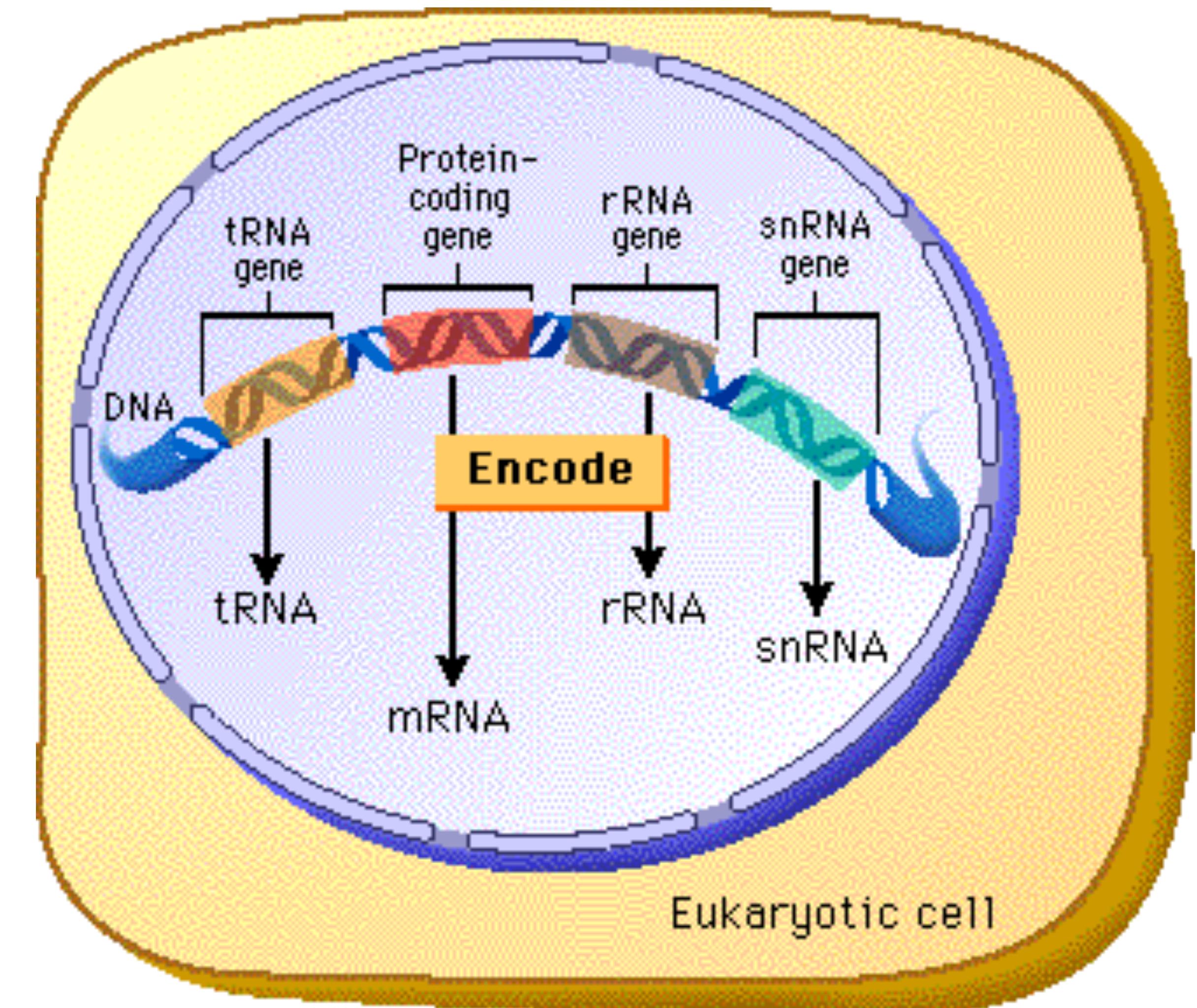
RNA

- 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

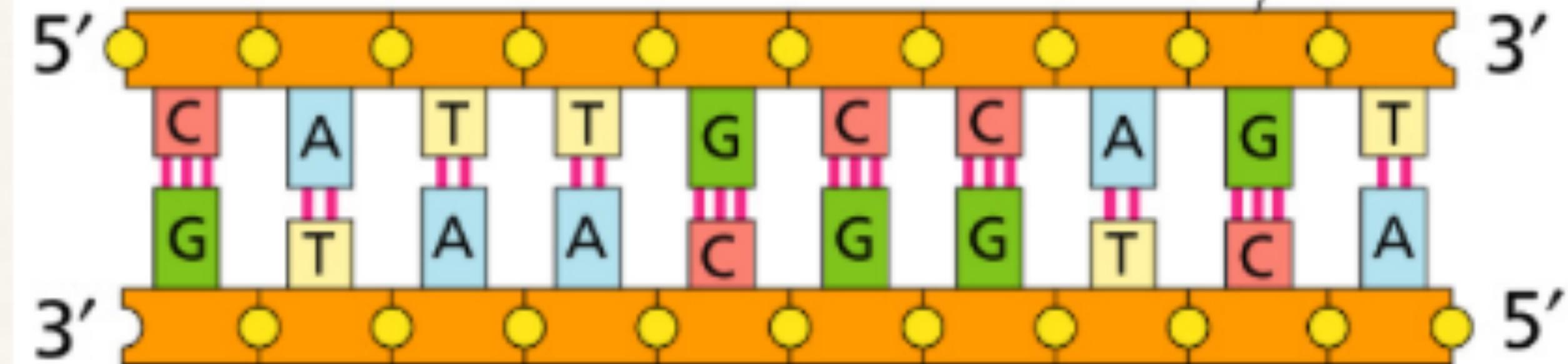
TRANSCRIPTION

- 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 expressed

GENE IS TRANSCRIBED
FROM DNA

(A)

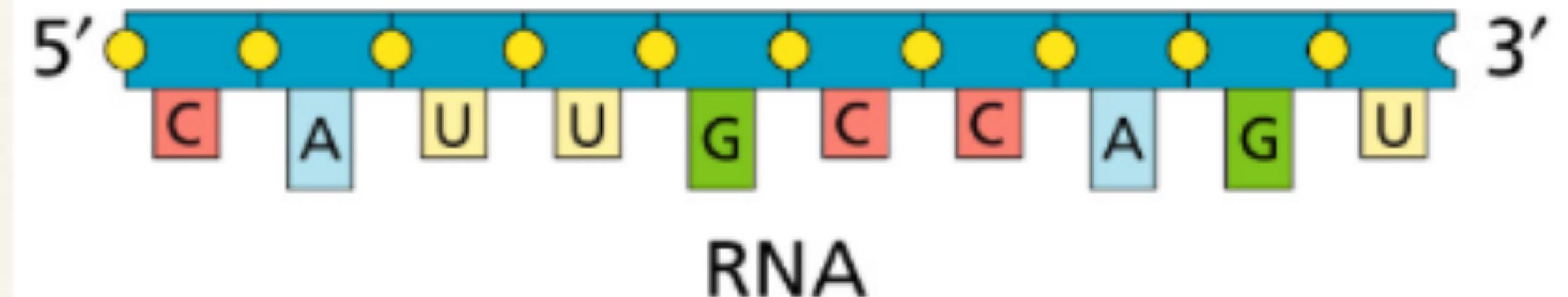
coding strand



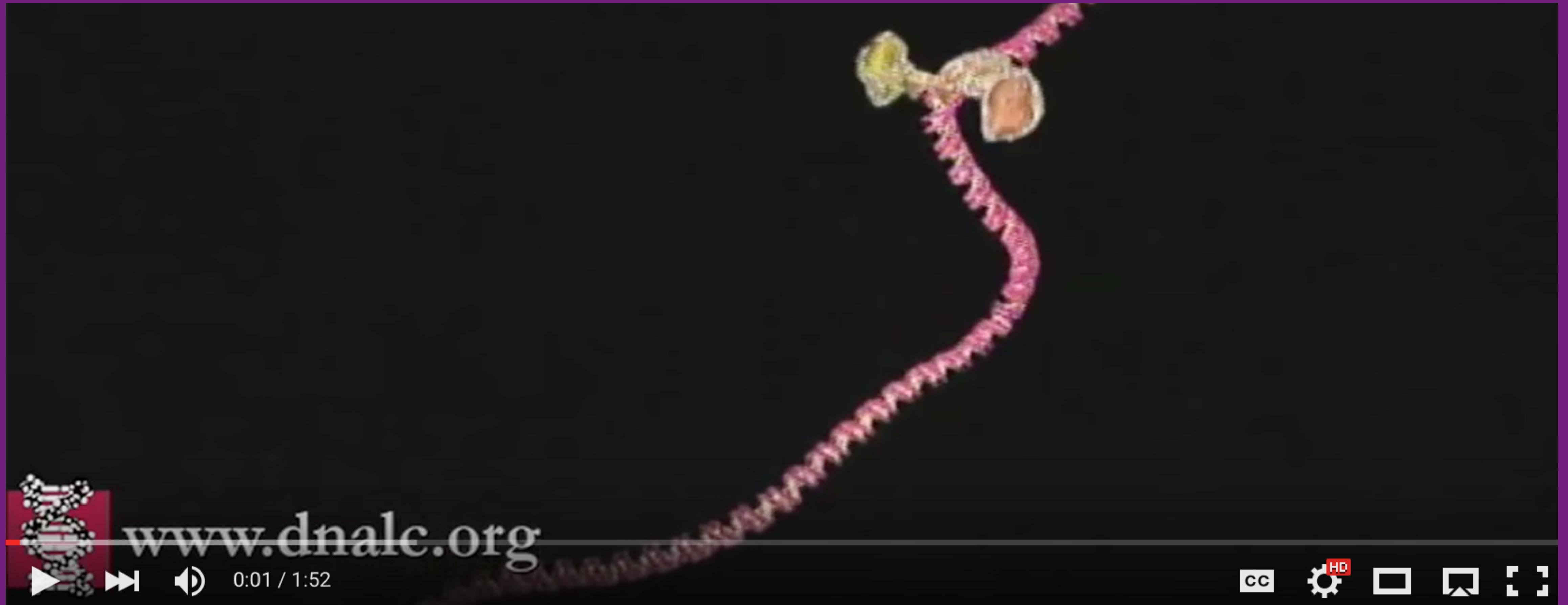
noncoding strand



TRANSCRIPTION



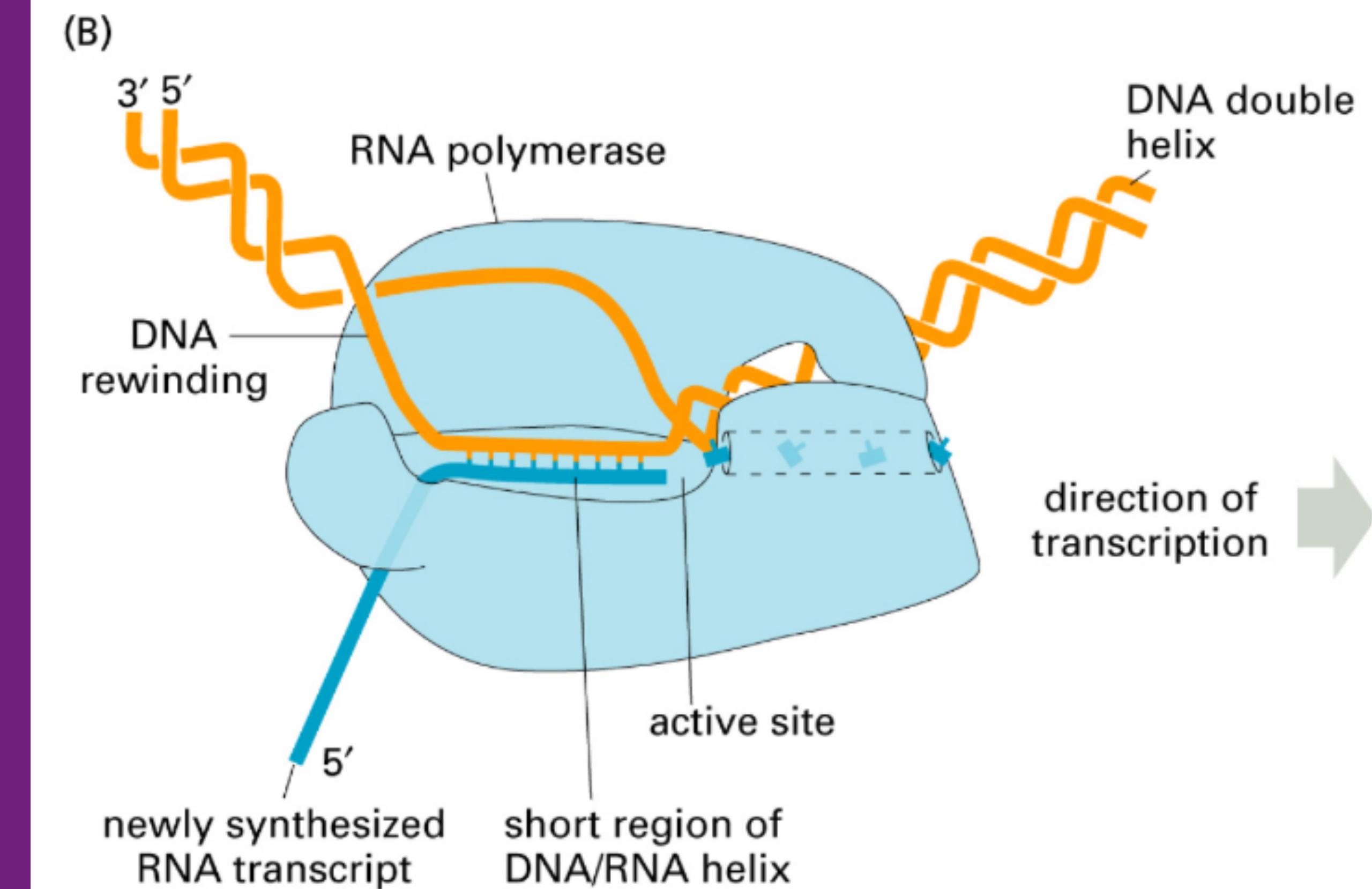
TRANSCRIPTION



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

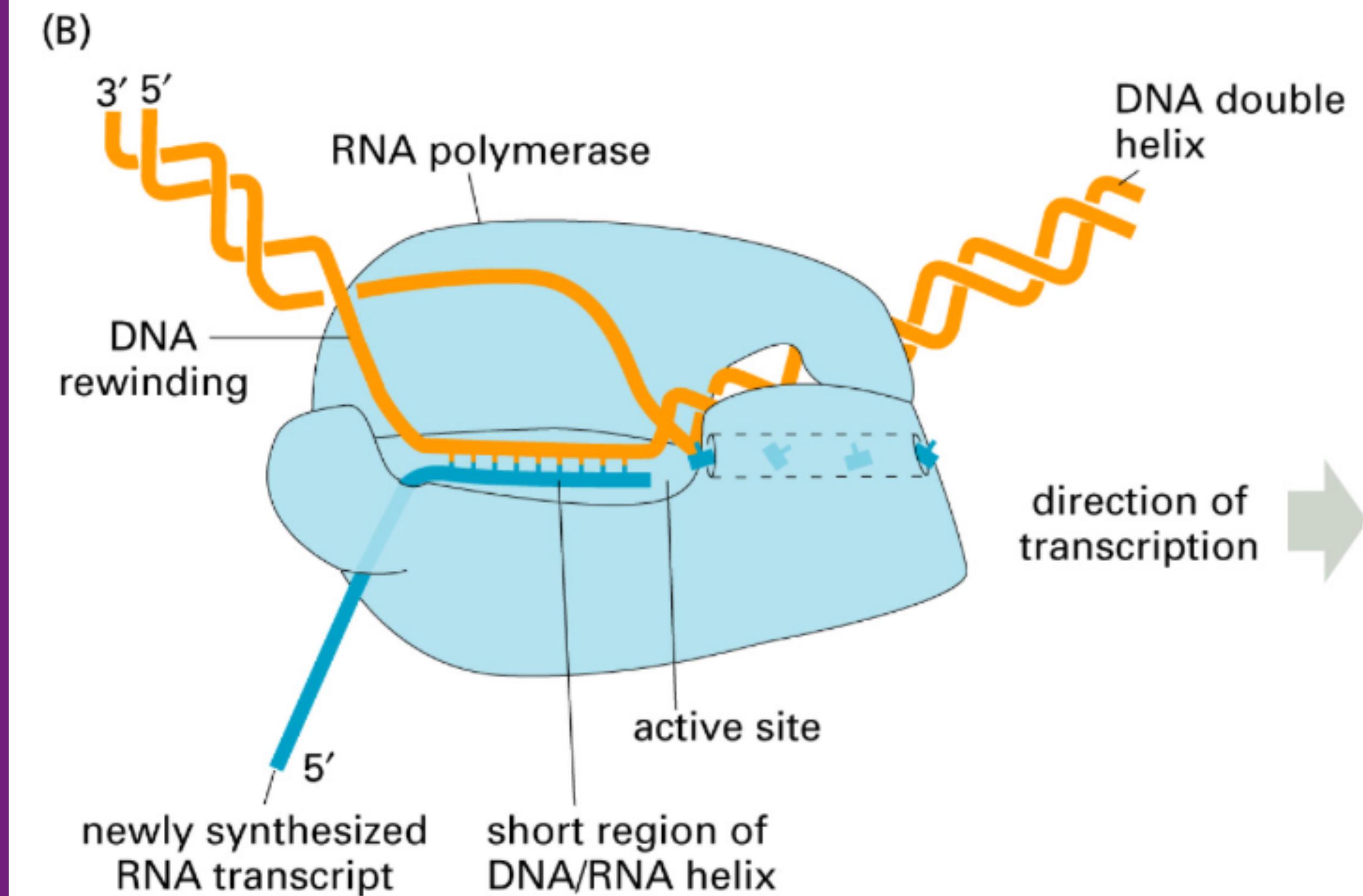
TRANSCRIPTION

- 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**



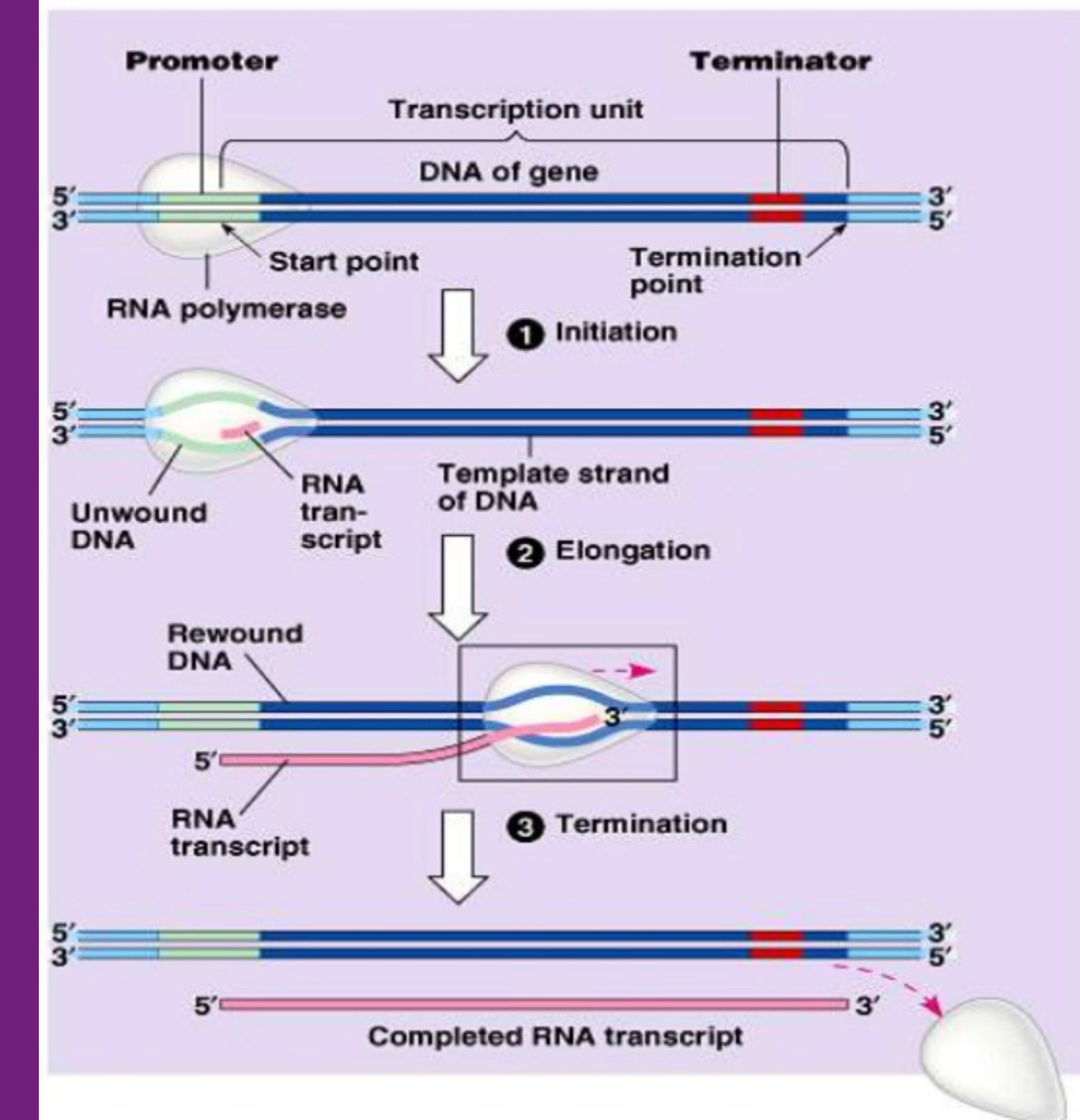
TRANSCRIPTION

- 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



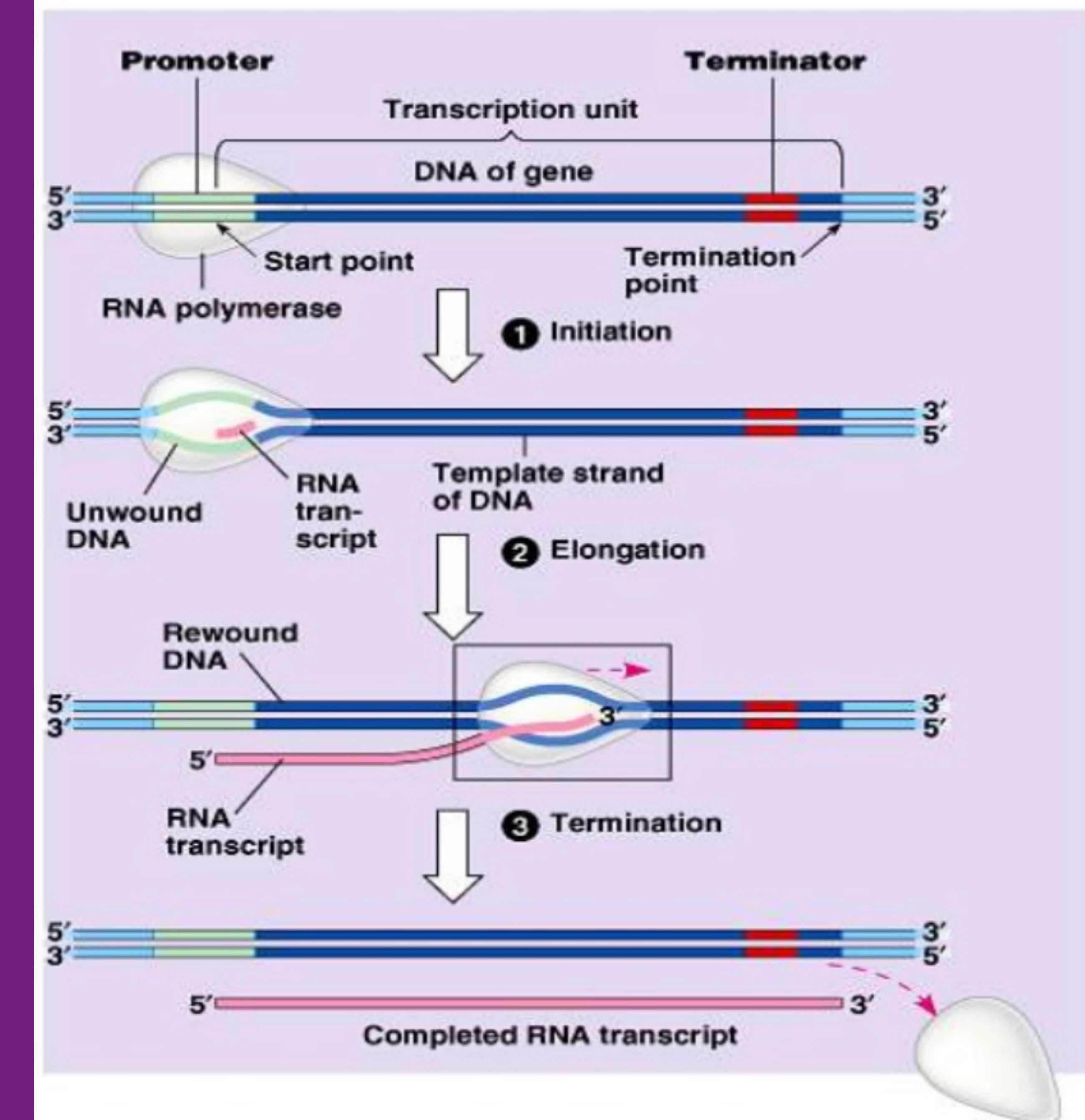
TRANSCRIPTION

- 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

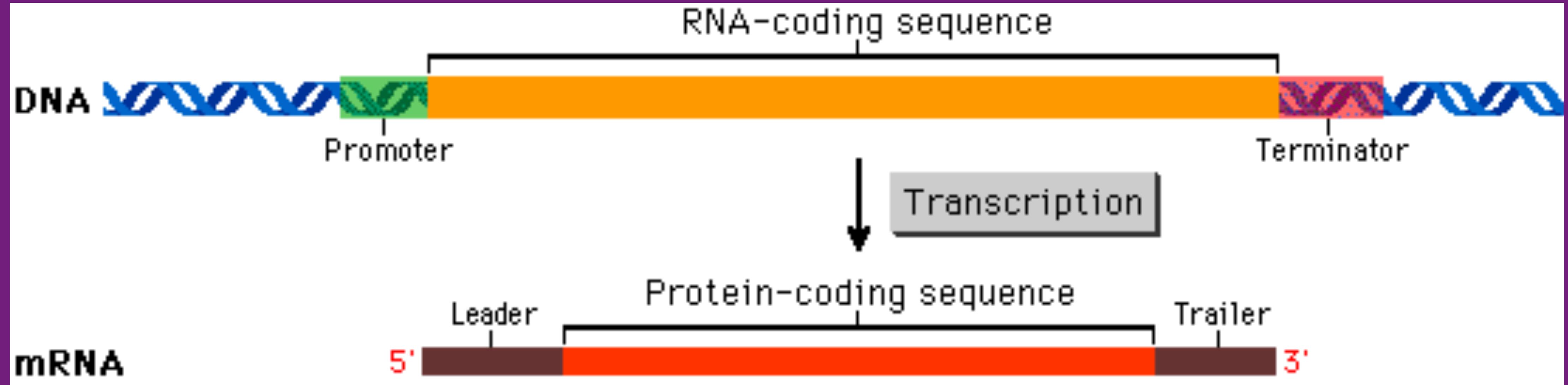


TRANSCRIPTION

- 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

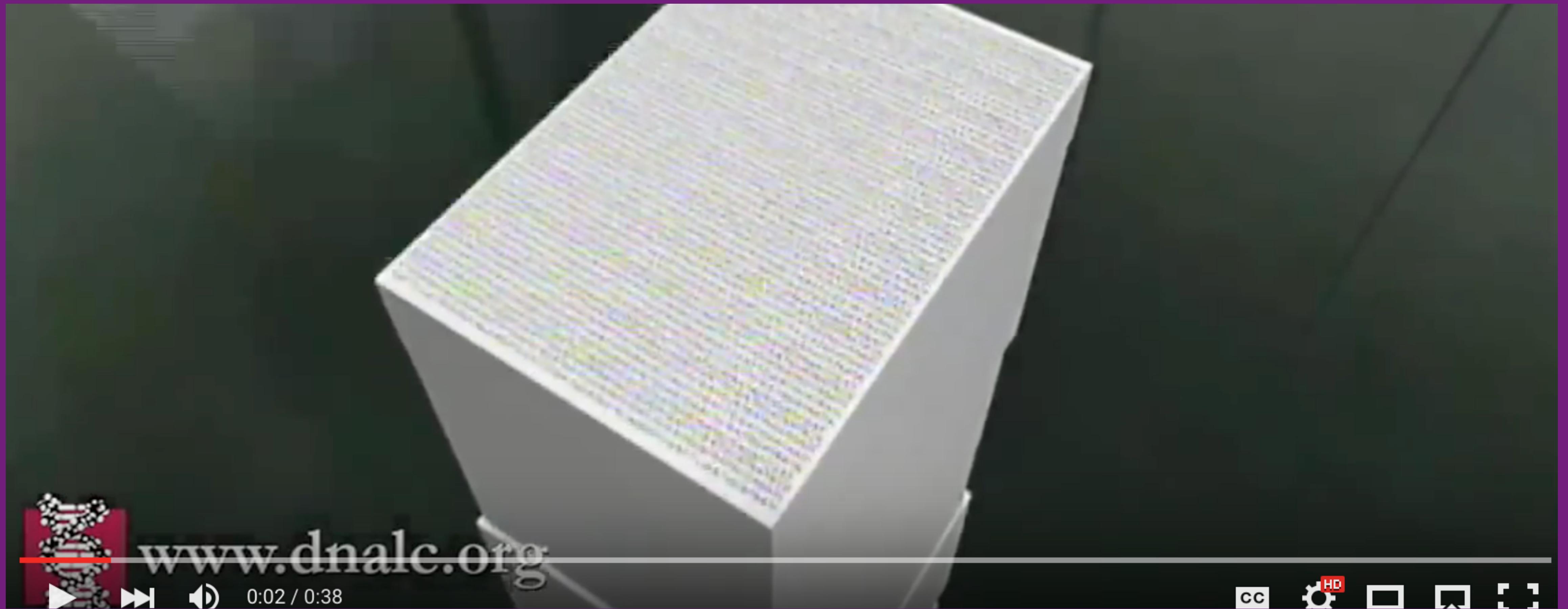


TRANSCRIPTION



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

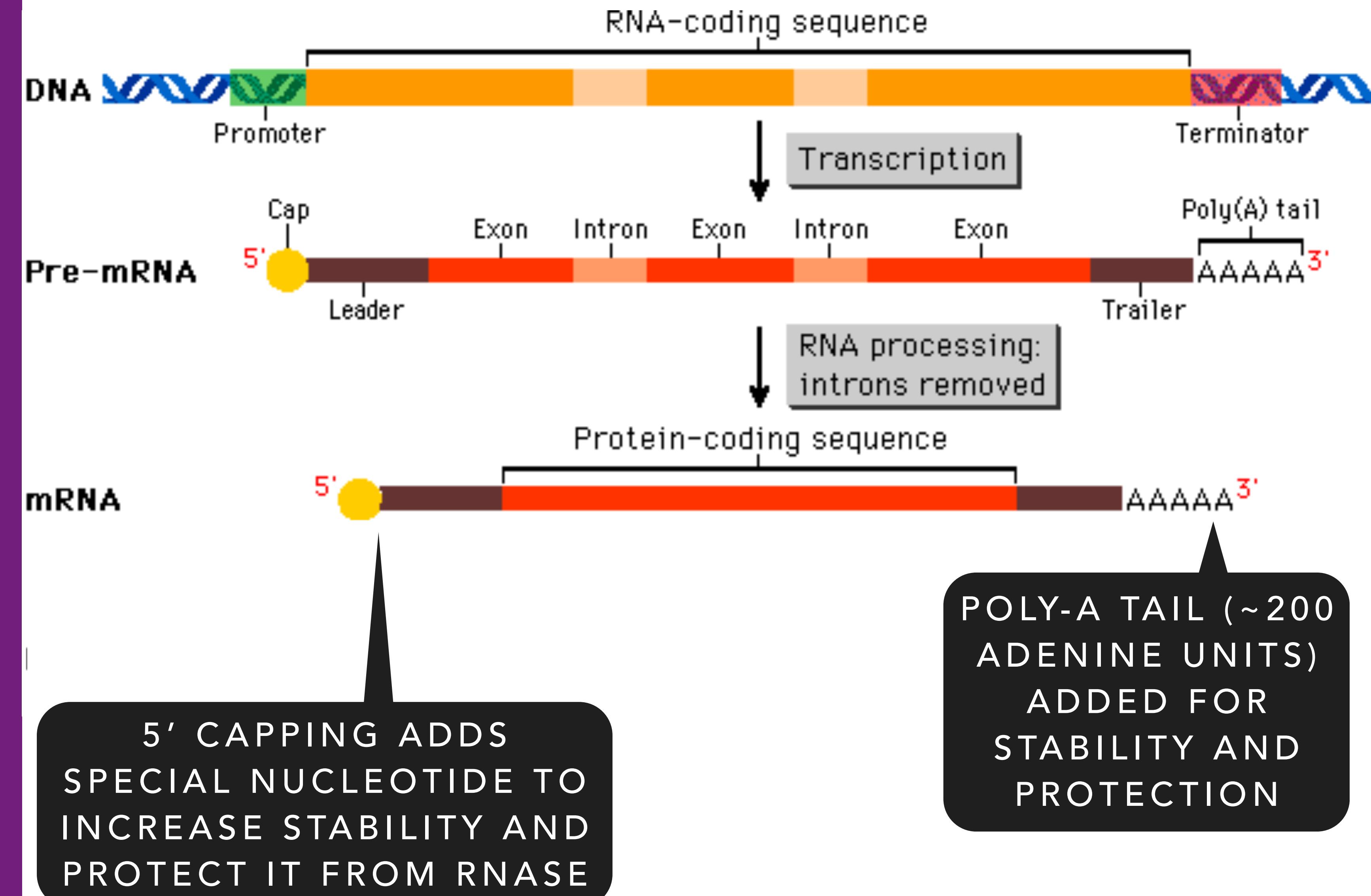
TRANSCRIPTION



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

TRANSCRIPTION

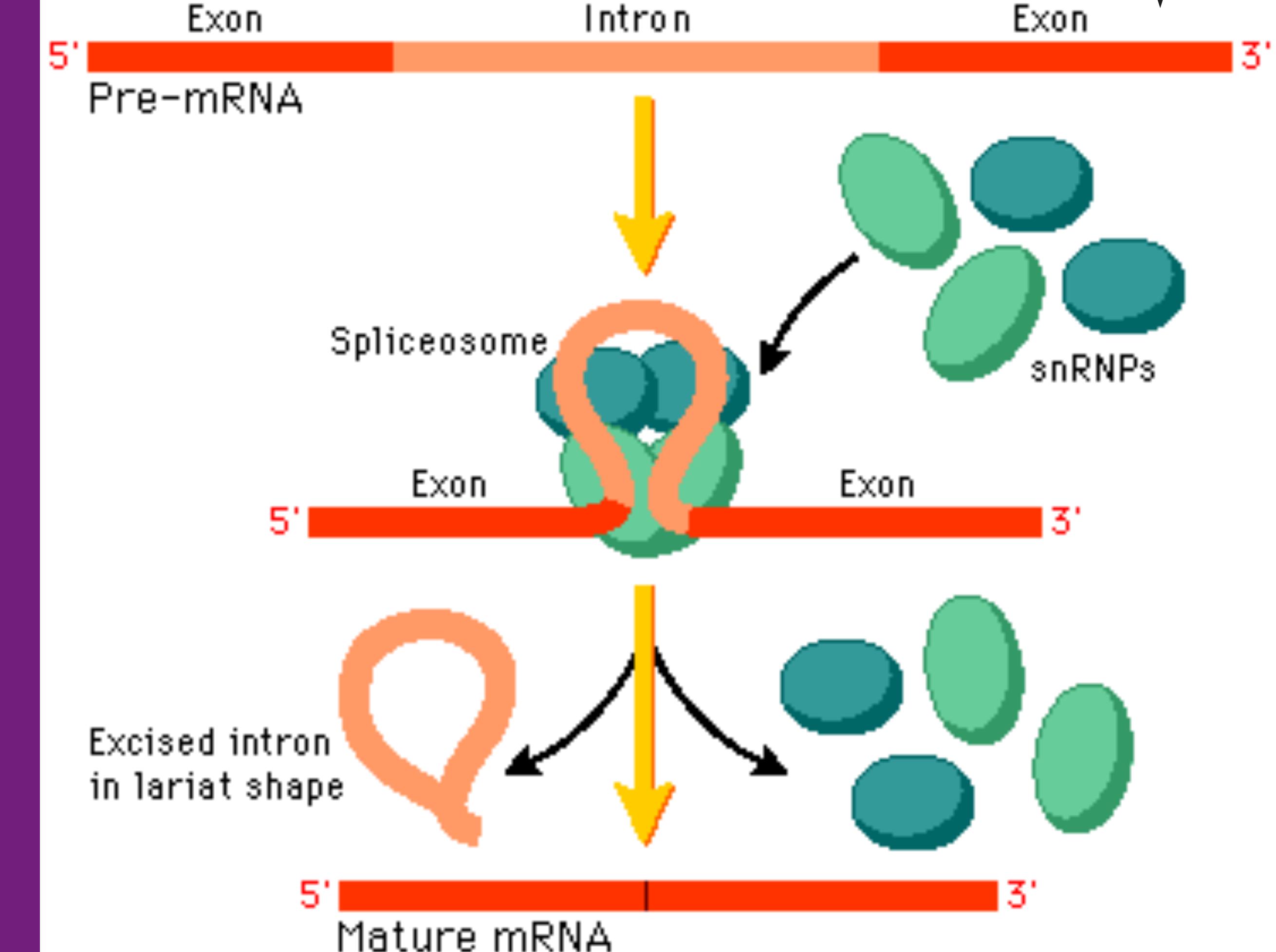
- mRNA in eukaryotes
 - Post-transcriptional modification
 - Spliceosome processes gene to remove extra sequences (introns)
 - Exits nucleus to cytoplasm



TRANSCRIPTION

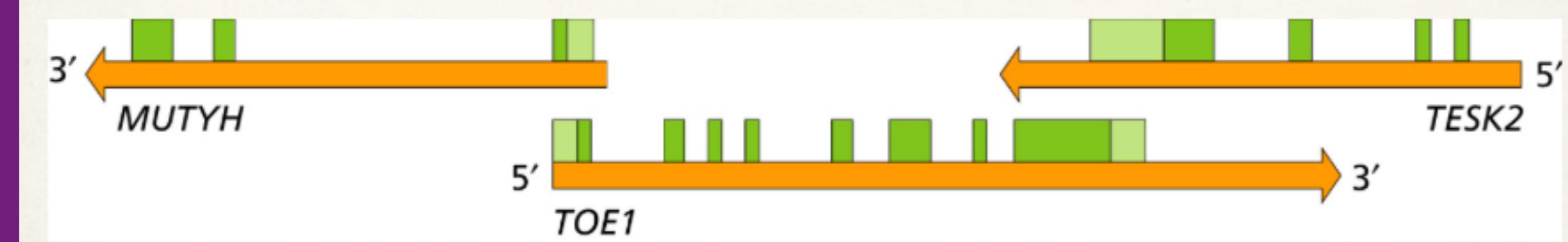
SNRNP - SMALL NUCLEAR RIBONUCLEOPROTEIN PARTICLES; COMPLEXES OF SNRNAs AND PROTEINS; FORM THE SPLICEOsome

- 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



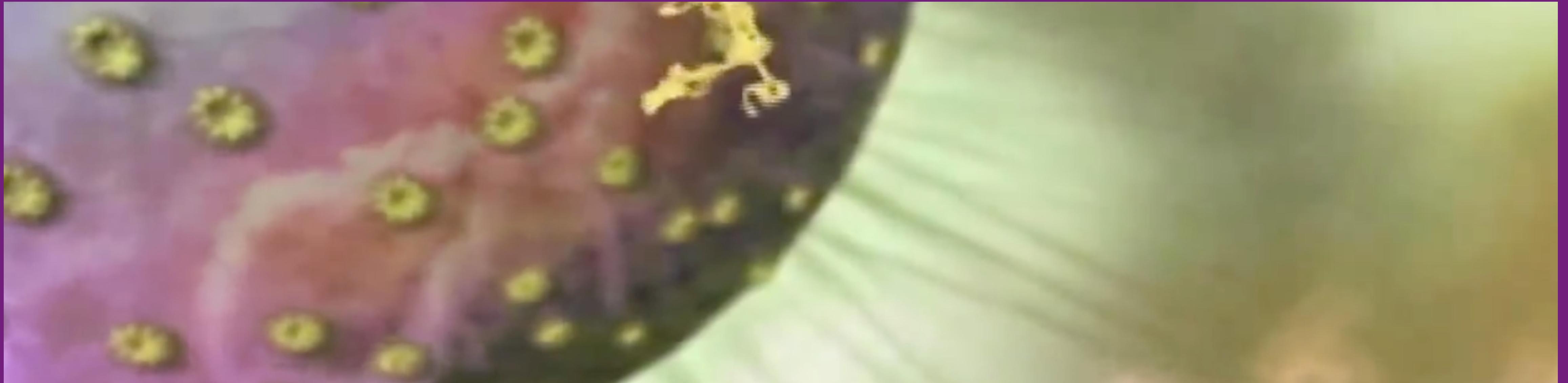
TRANSCRIPTION

- 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

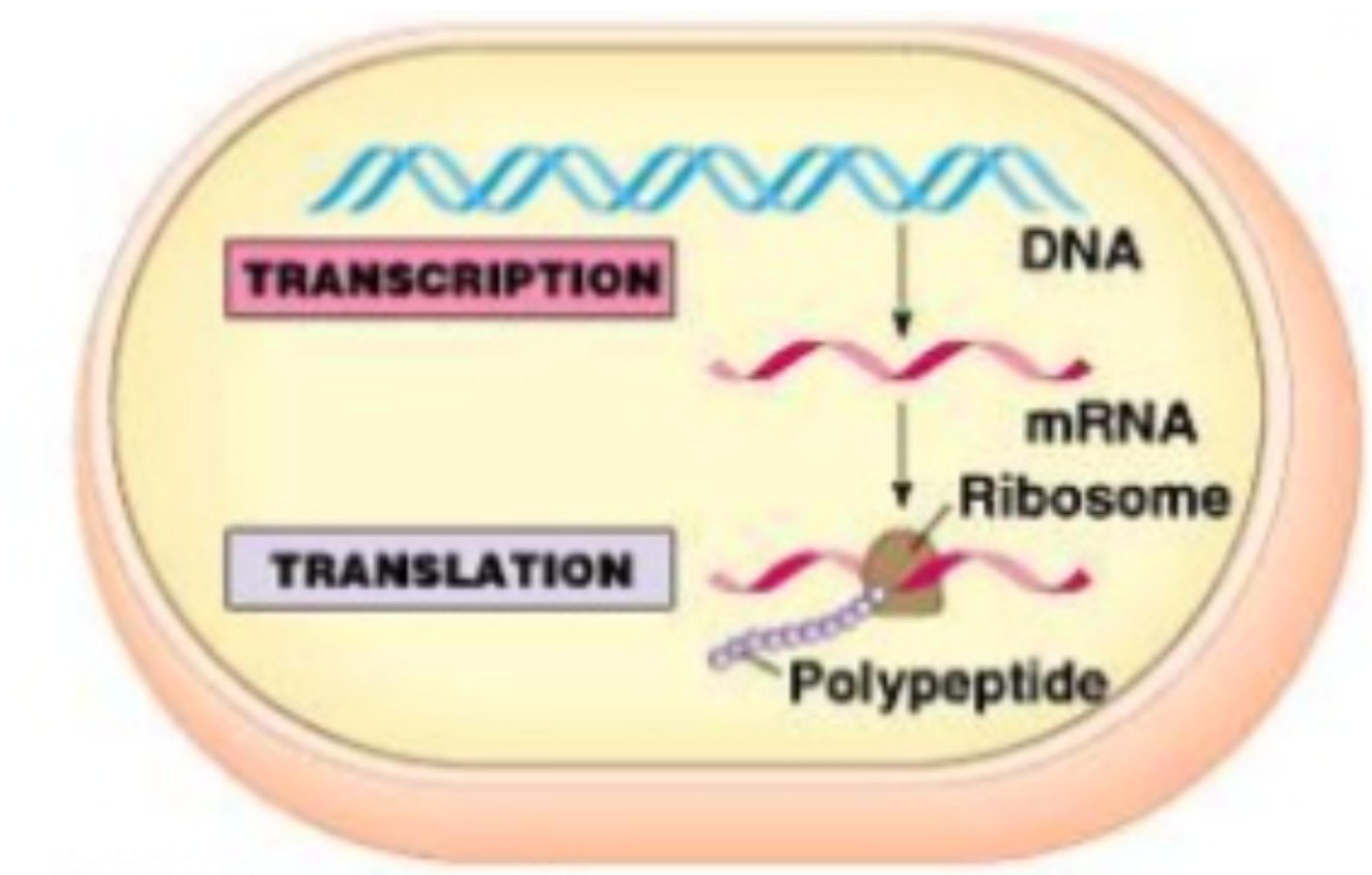
TRANSLATION



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

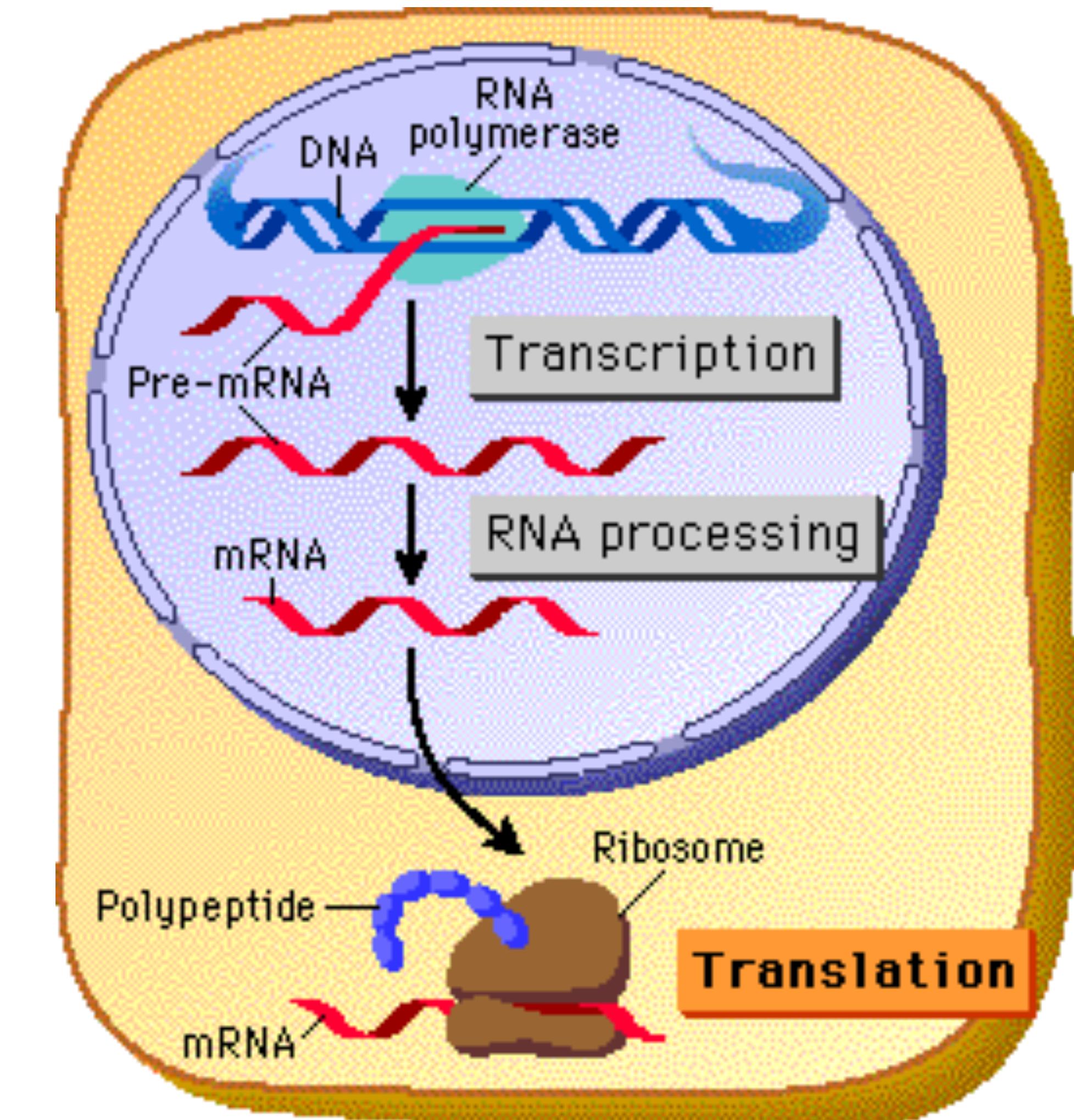
TRANSLATION

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

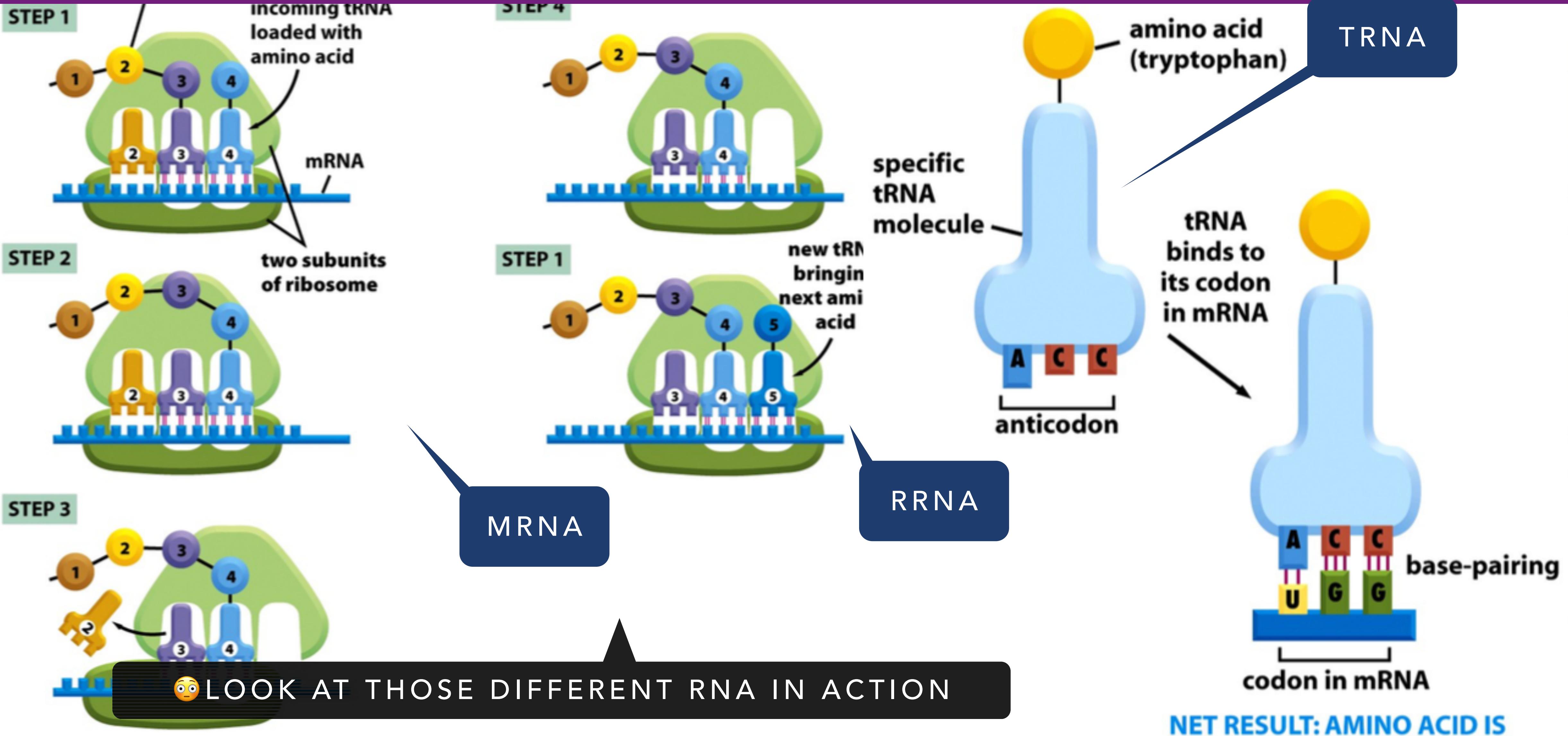


TRANSLATION

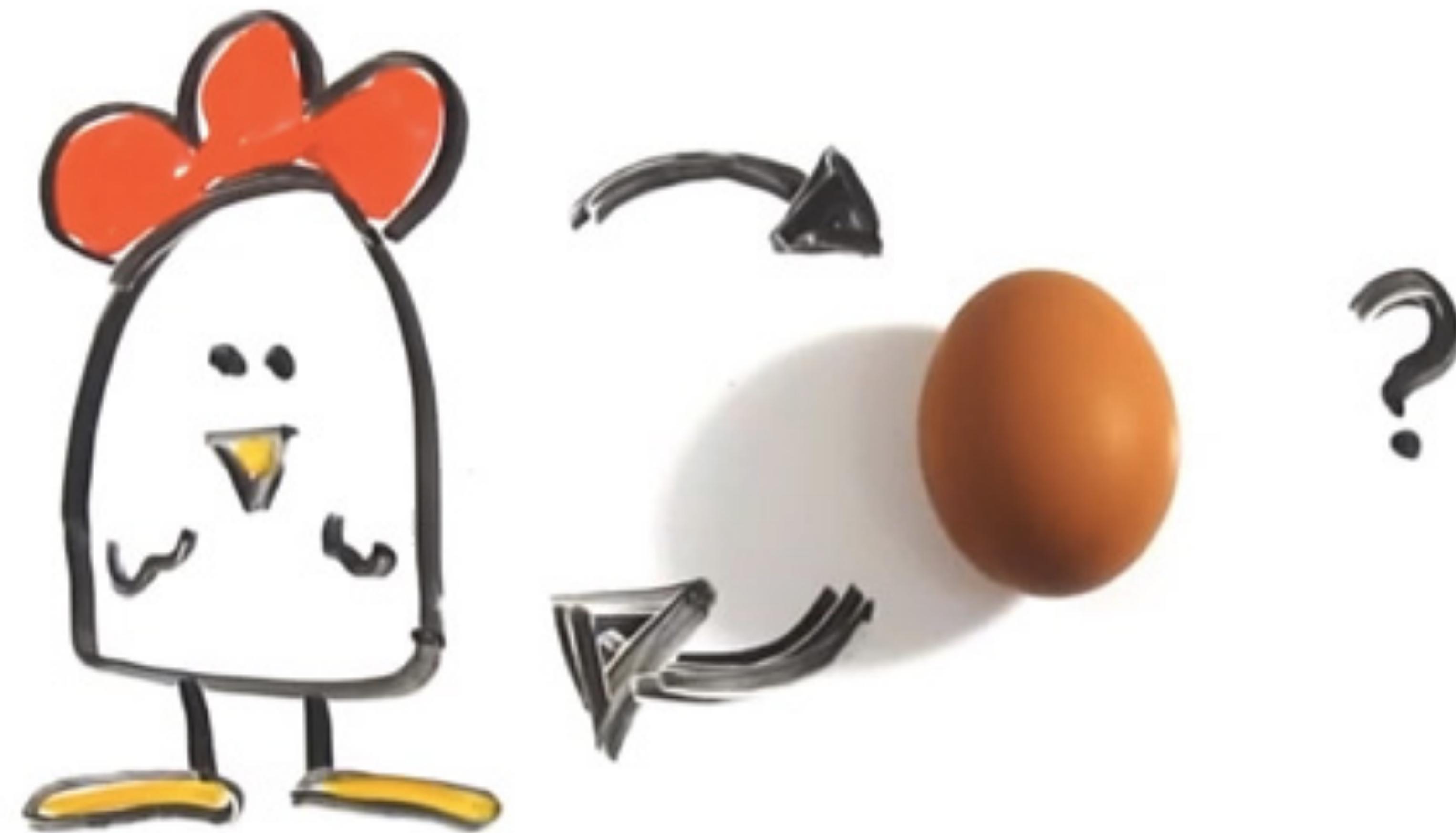
- Translation in a Eukaryotic Cell
 - A protein-coding gene is transcribed into a pre-mRNA
 - Pre-mRNA is processed into a mature mRNA
 - mRNA exits the nucleus
 - mRNA is translated on ribosomes to produce the polypeptide chain



TRANSLATION



TRANSLATION



TRANSLATION

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

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."



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.

TRANSLATION

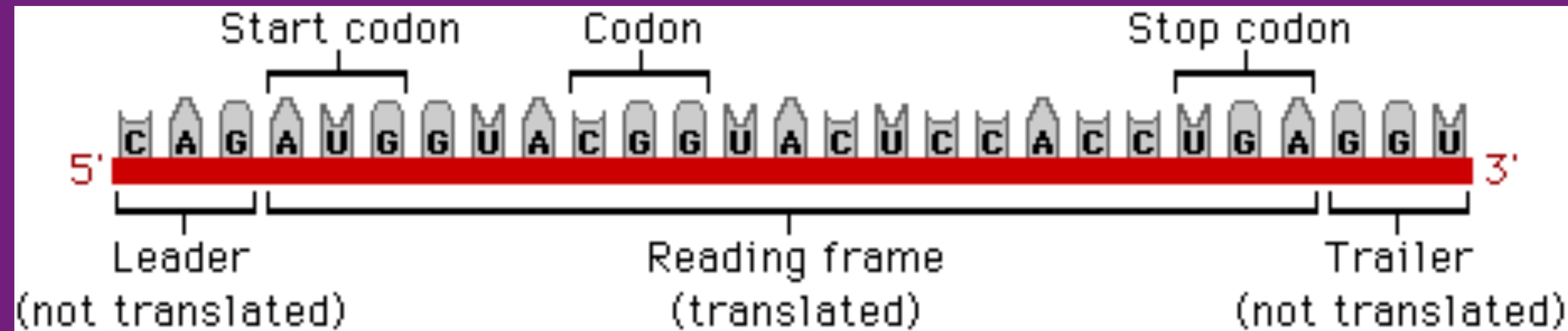
- mRNA is translated into a protein according the genetic code
- Codon
 - 3 nucleotides read at a time
 - Start codon (AUG)
 - Stop codon (UAA,UAG,UGA) tell the ribosome the protein is complete

STANDARD GENETIC CODE

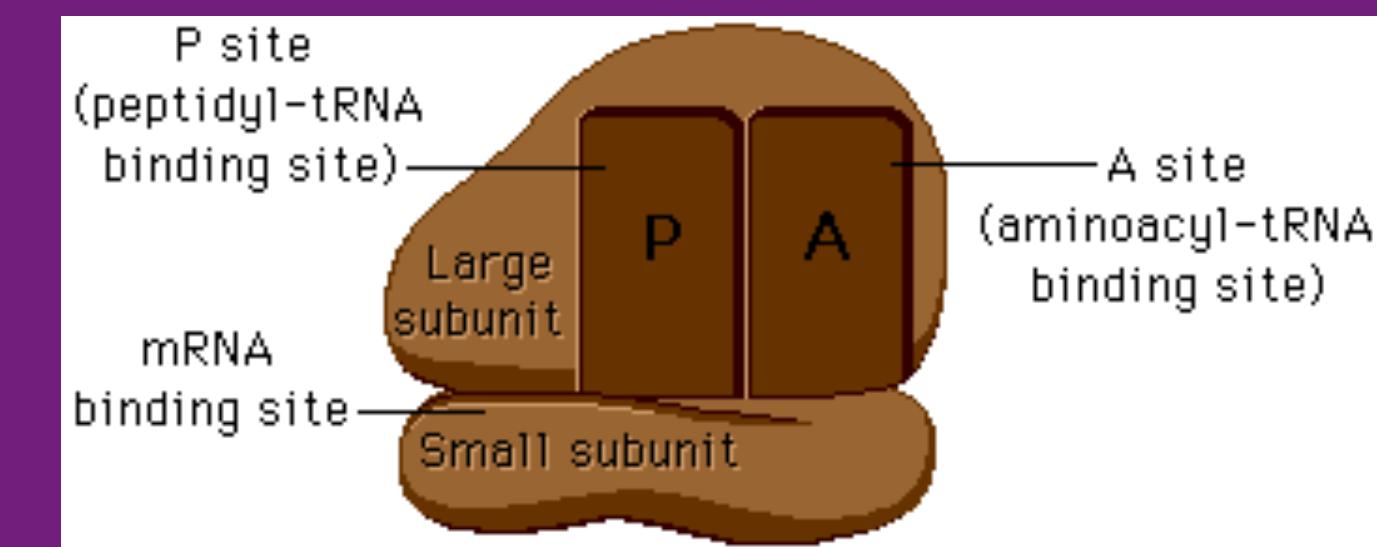
Second letter

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

TRANSLATION

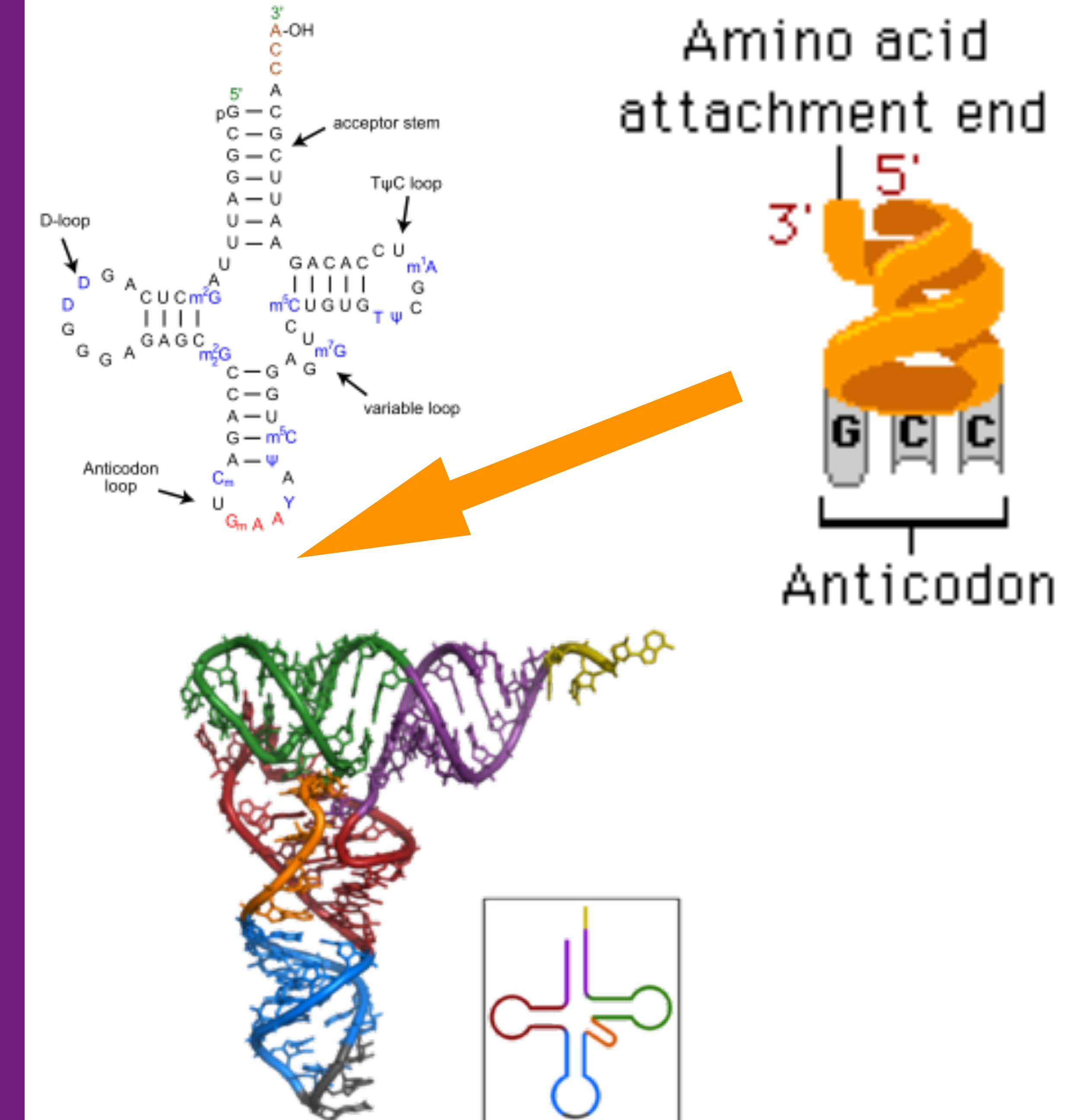


- mRNA passes through the ribosome
- Codons are recognized by tRNAs carrying the specified amino acids
- Each ribosomal subunit consists of rRNA (ribosomal RNA, encoded by rRNA genes) and ribosomal proteins

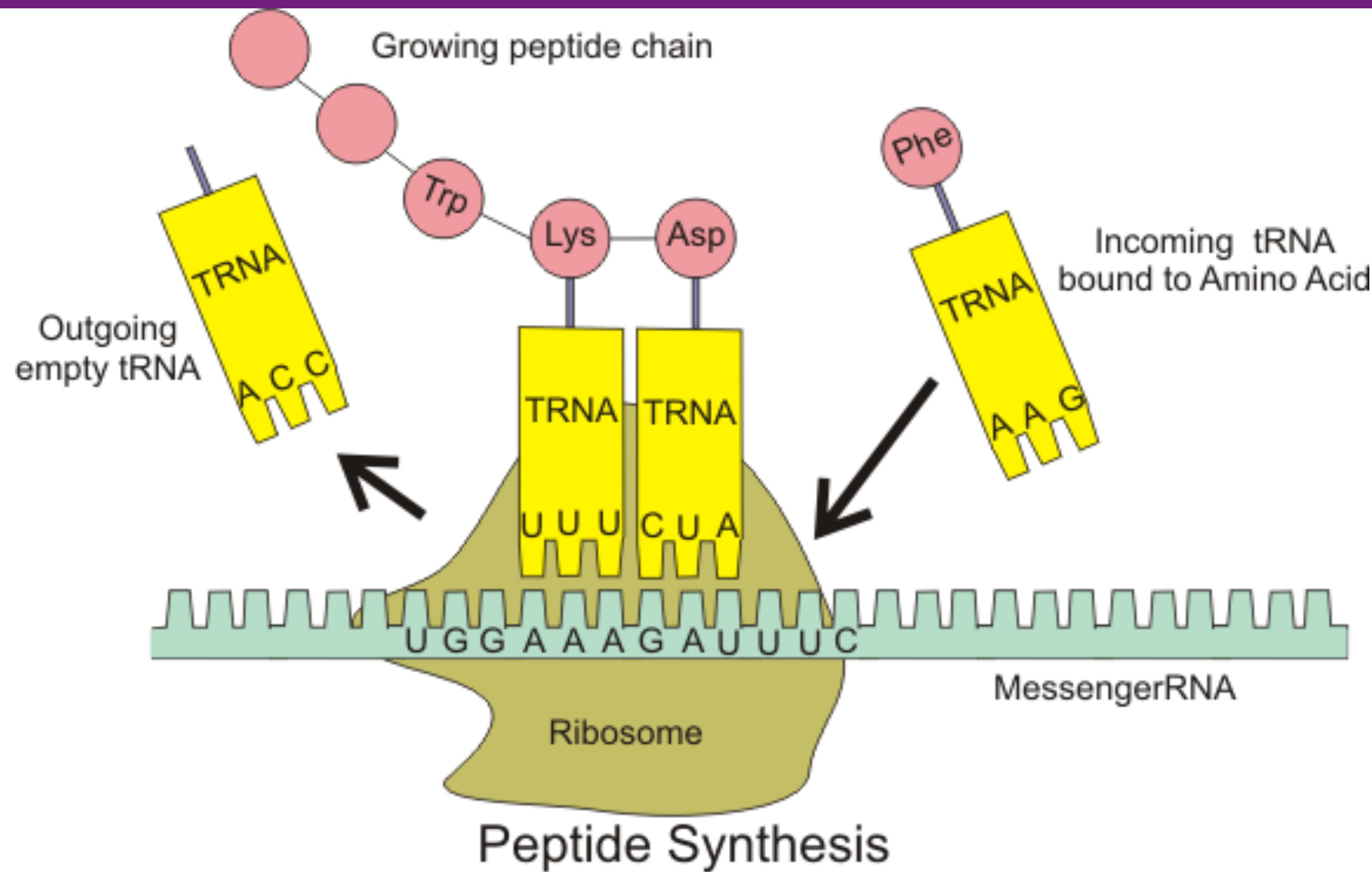


TRANSLATION

- 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

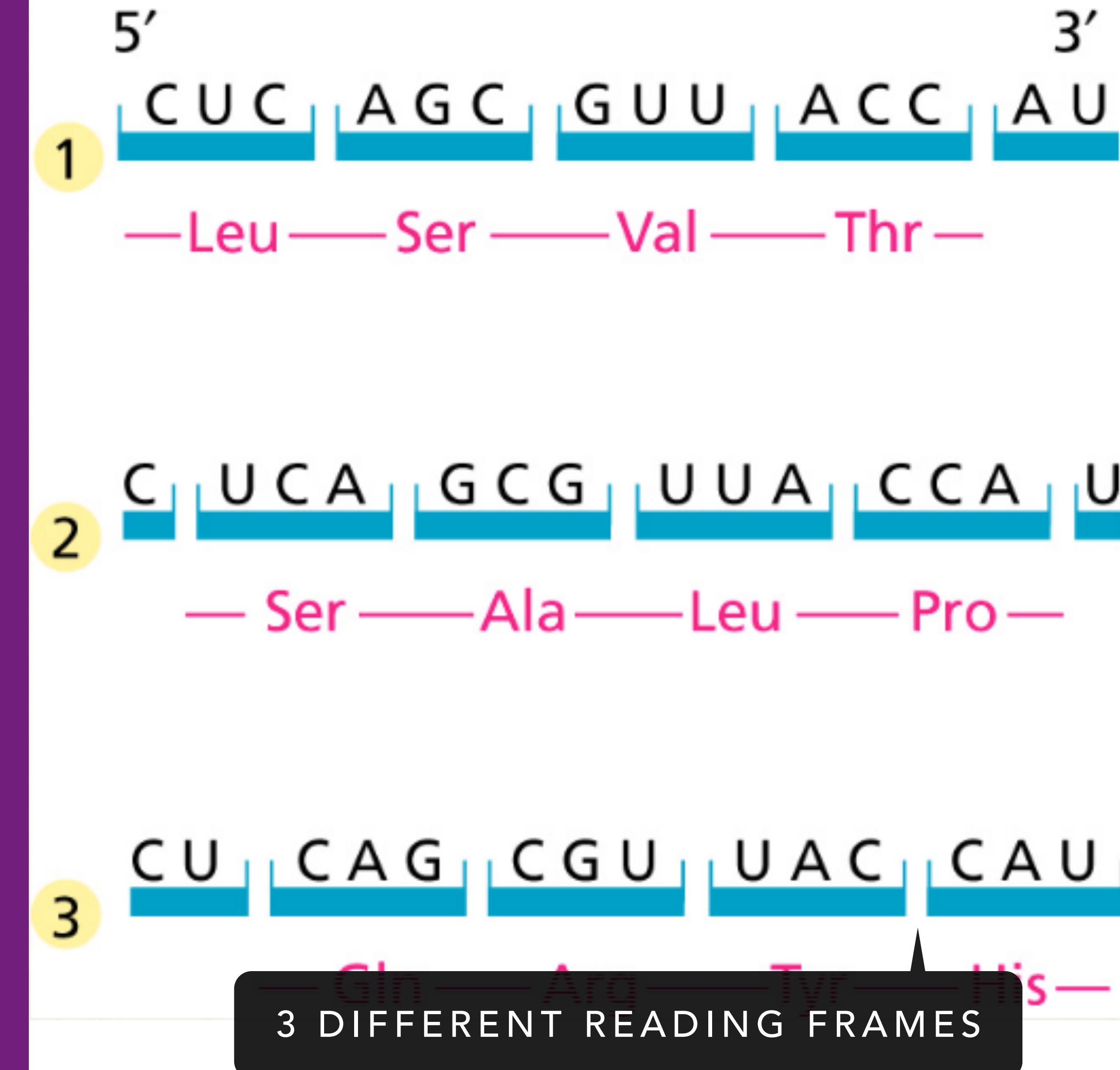


TRANSLATION

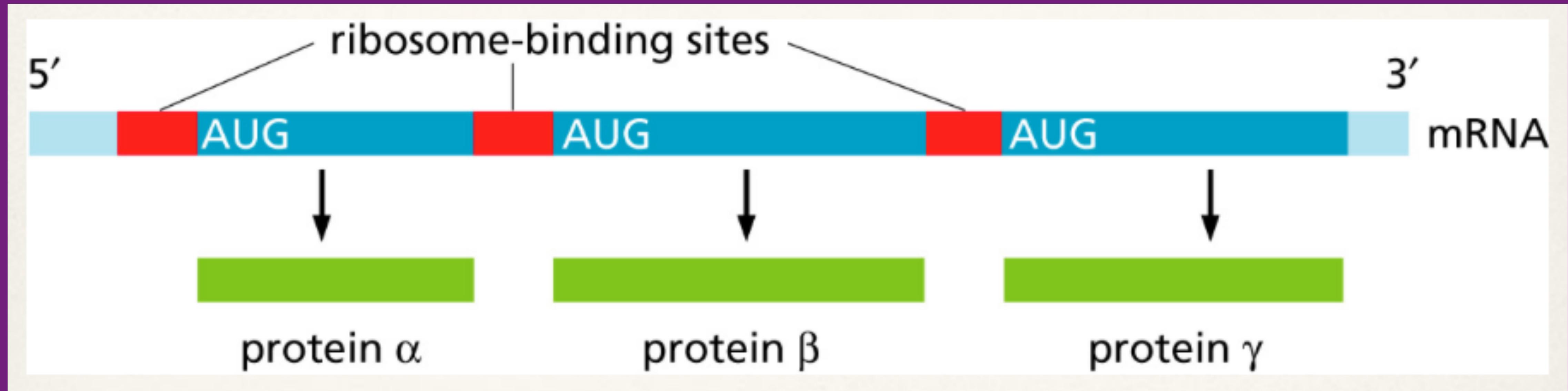


TRANSLATION

- 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** activates the simultaneous expression of all genes in the operon
- Allows efficient, coordinated protein synthesis

PROTEINS

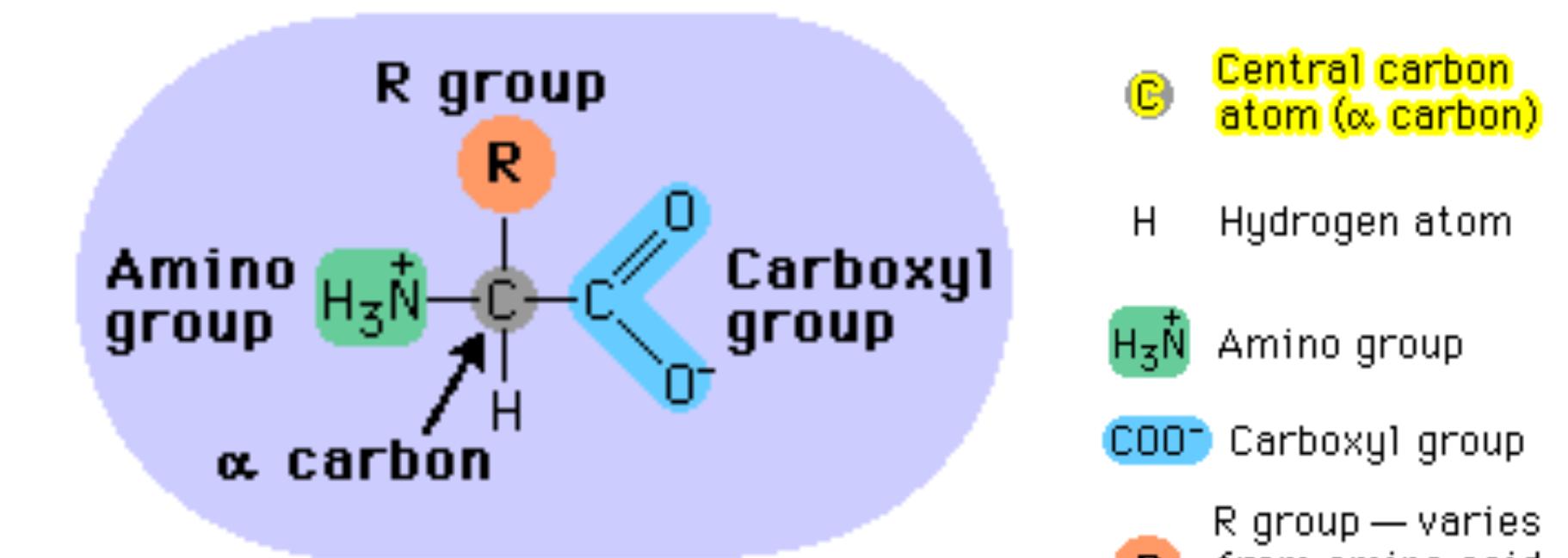
PROTEINS

- 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

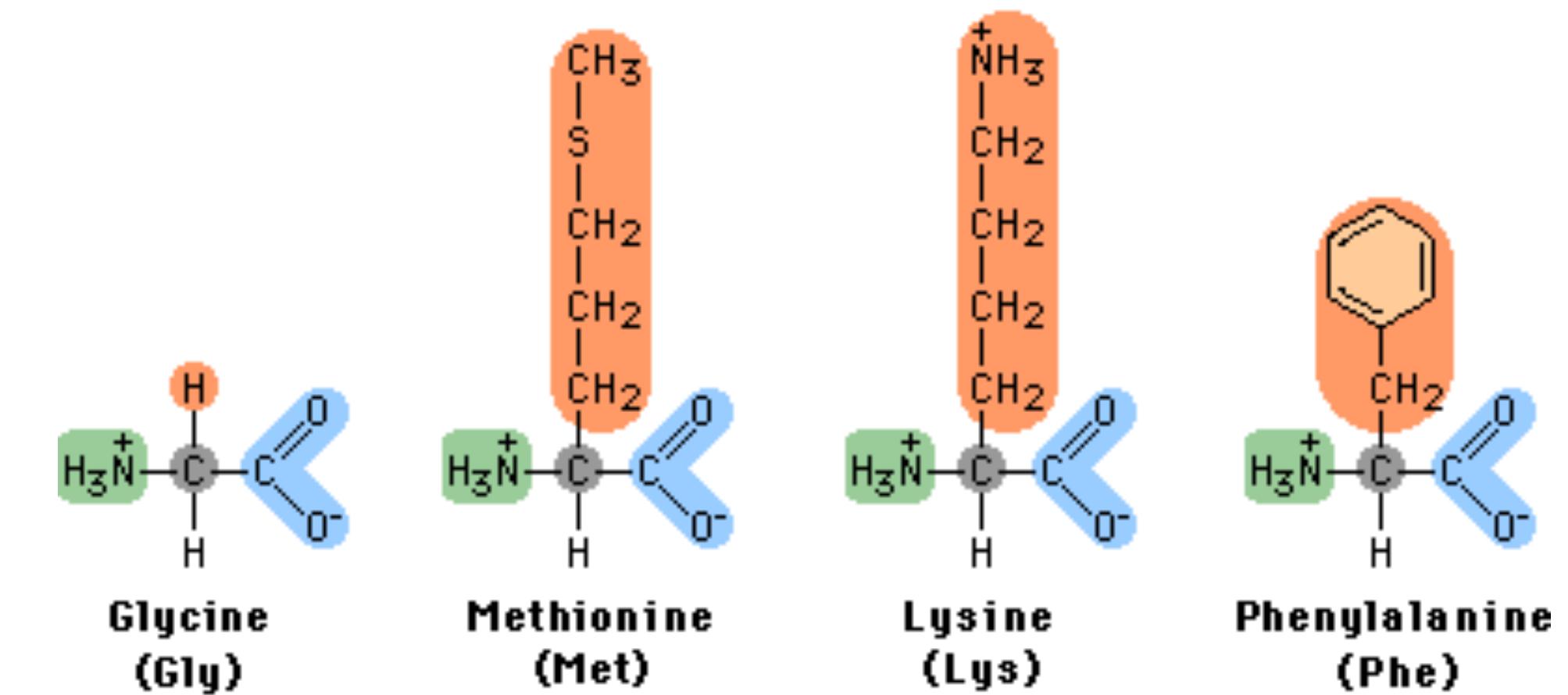


PROTEINS

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



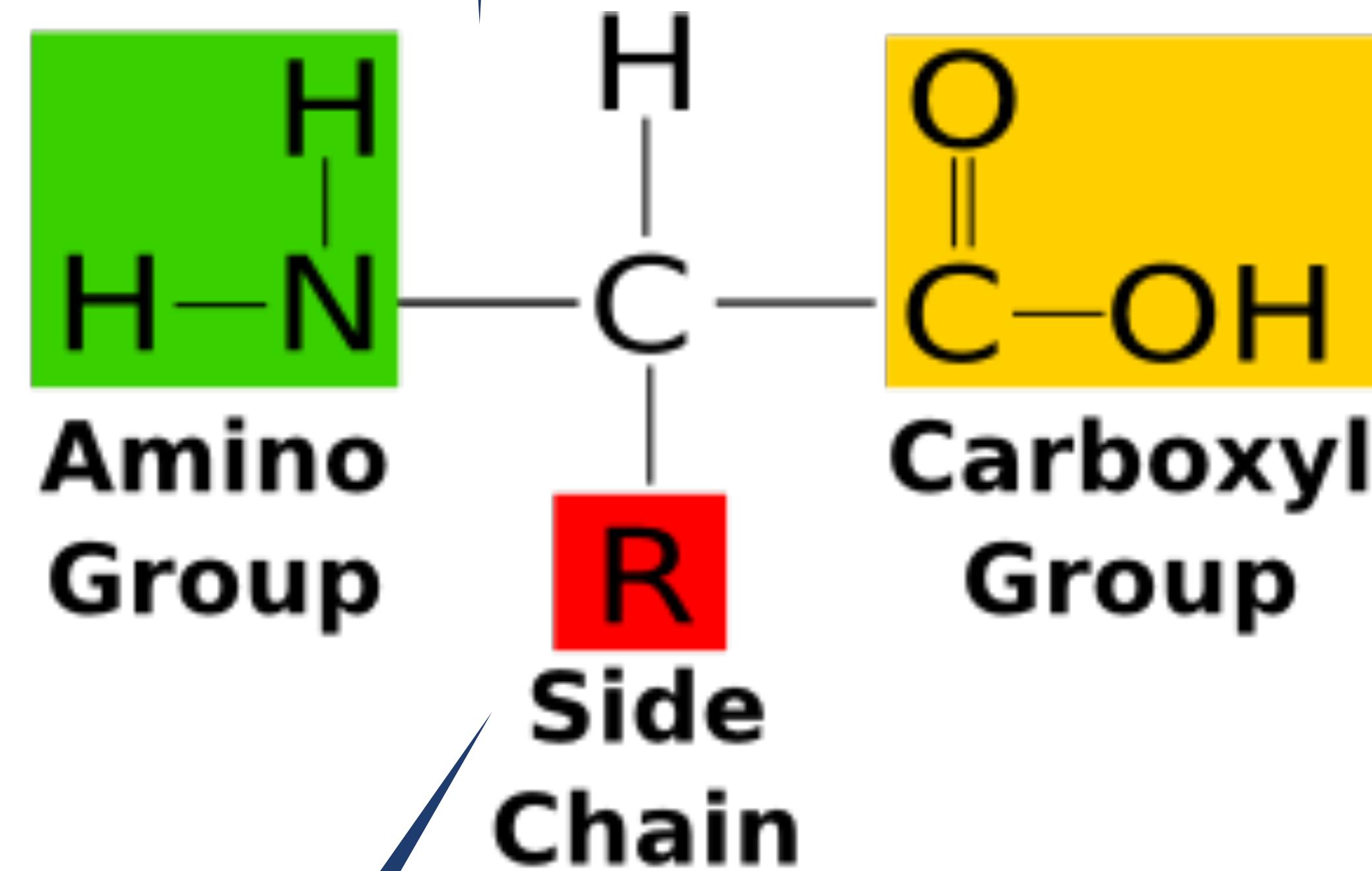
Legend:
C Central carbon atom (α carbon)
H Hydrogen atom
 H_3N^+ Amino group
 COO^- Carboxyl group
R R group — varies from amino acid to amino acid



PROTEINS

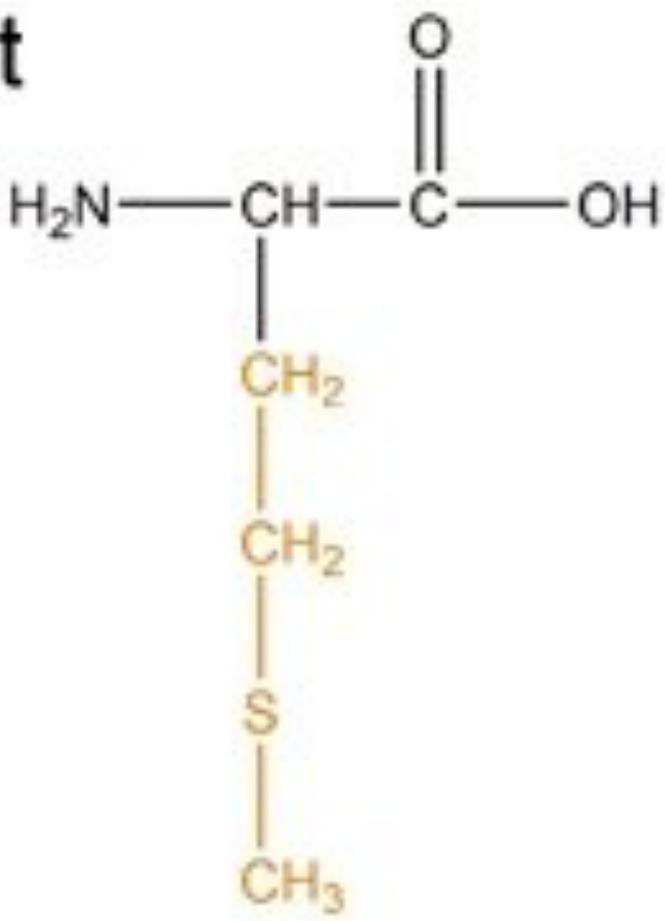
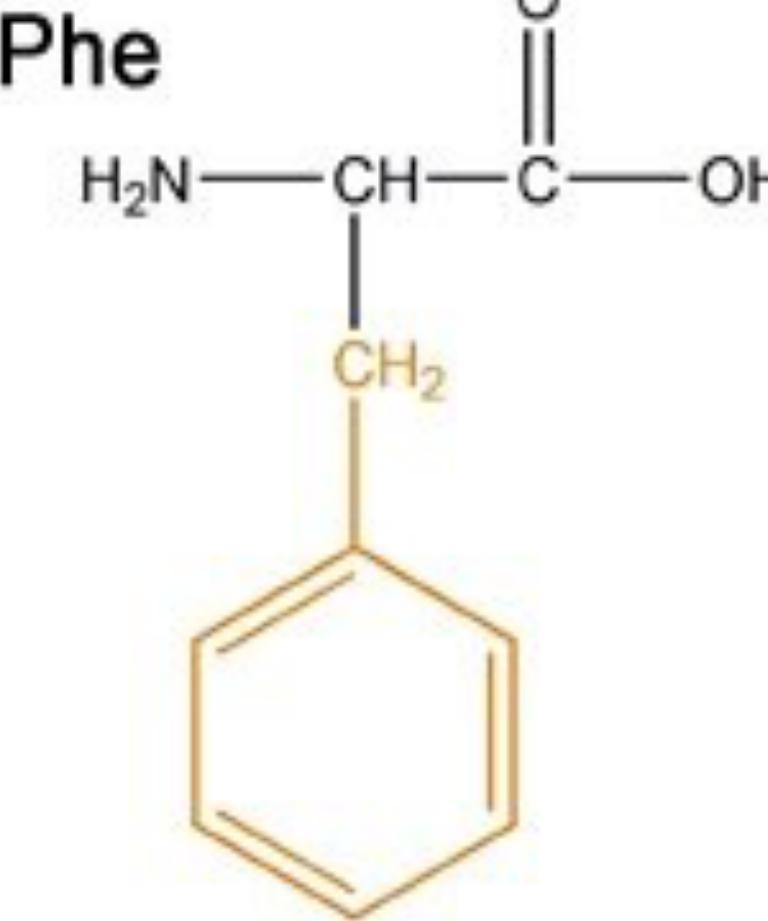
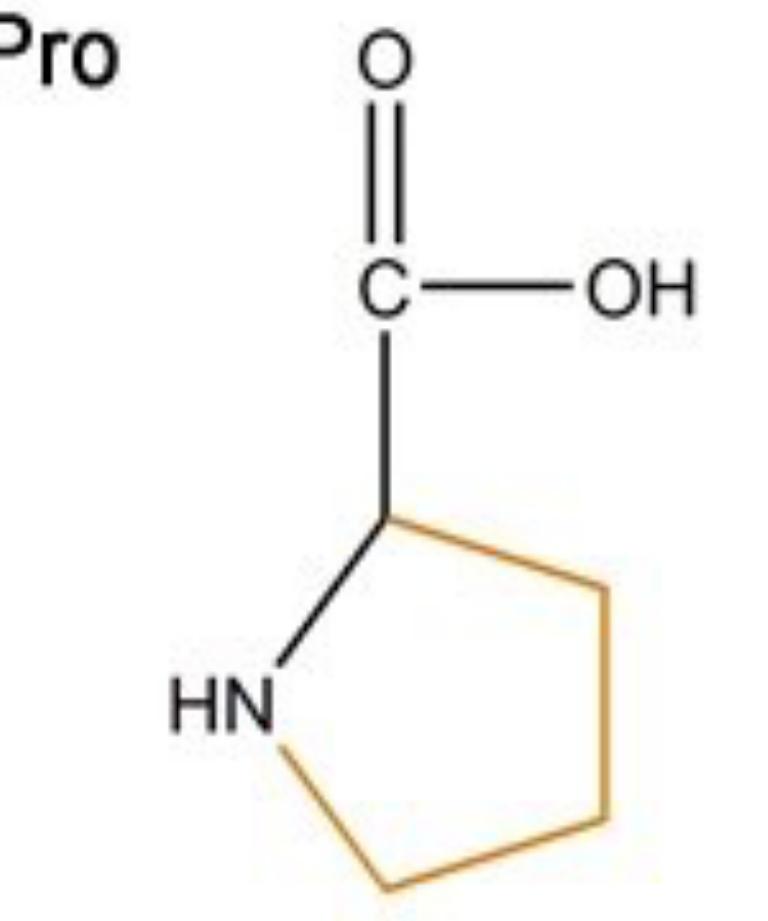
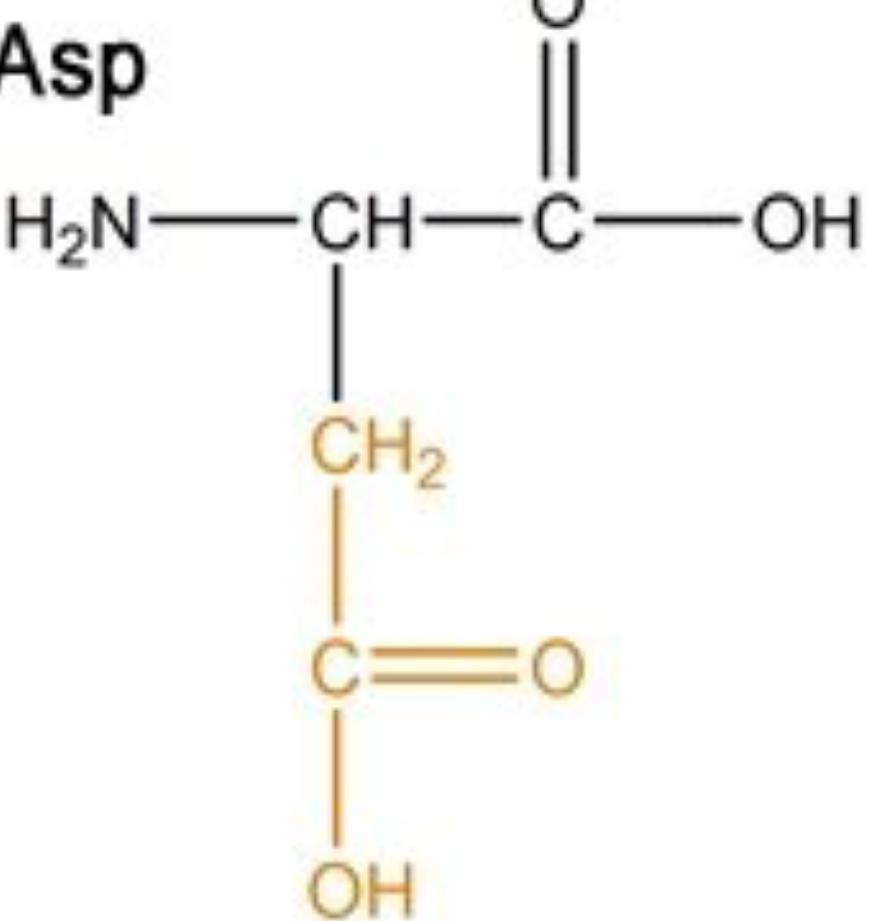
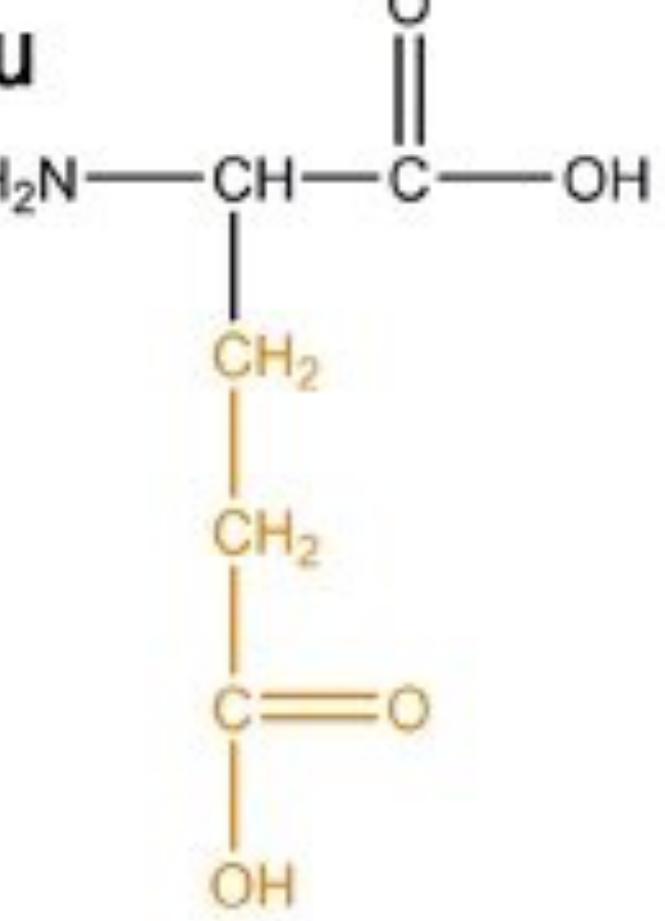
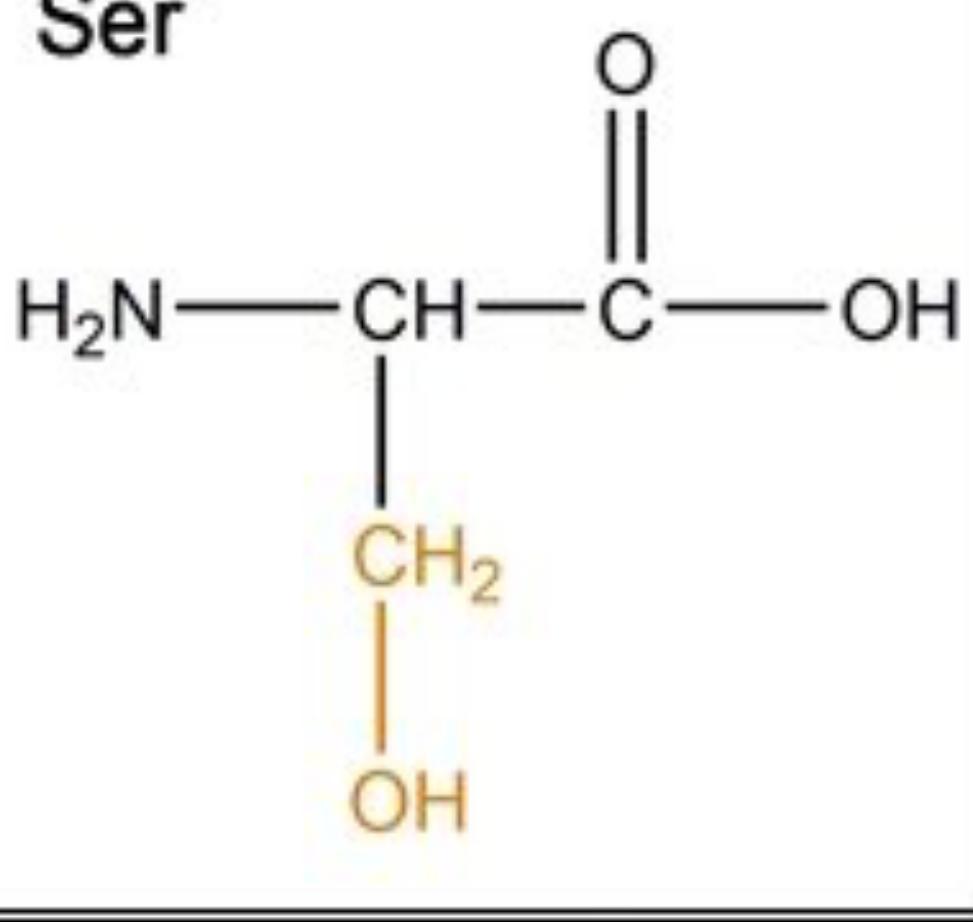
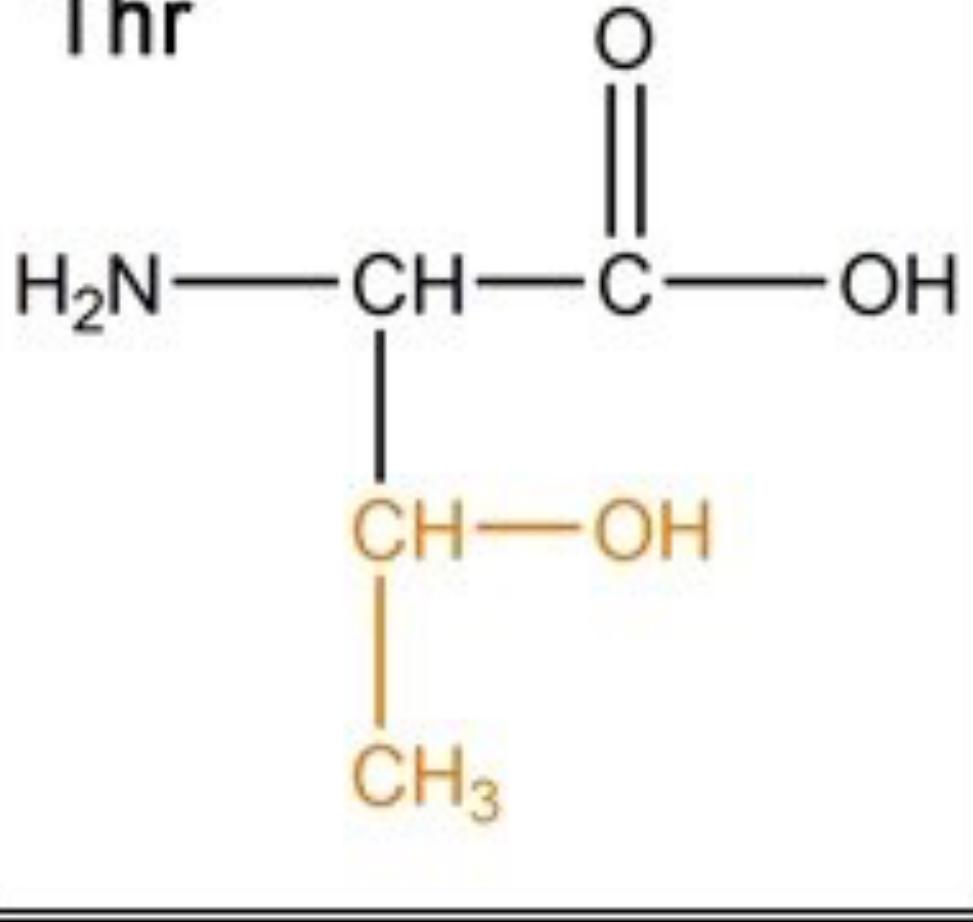
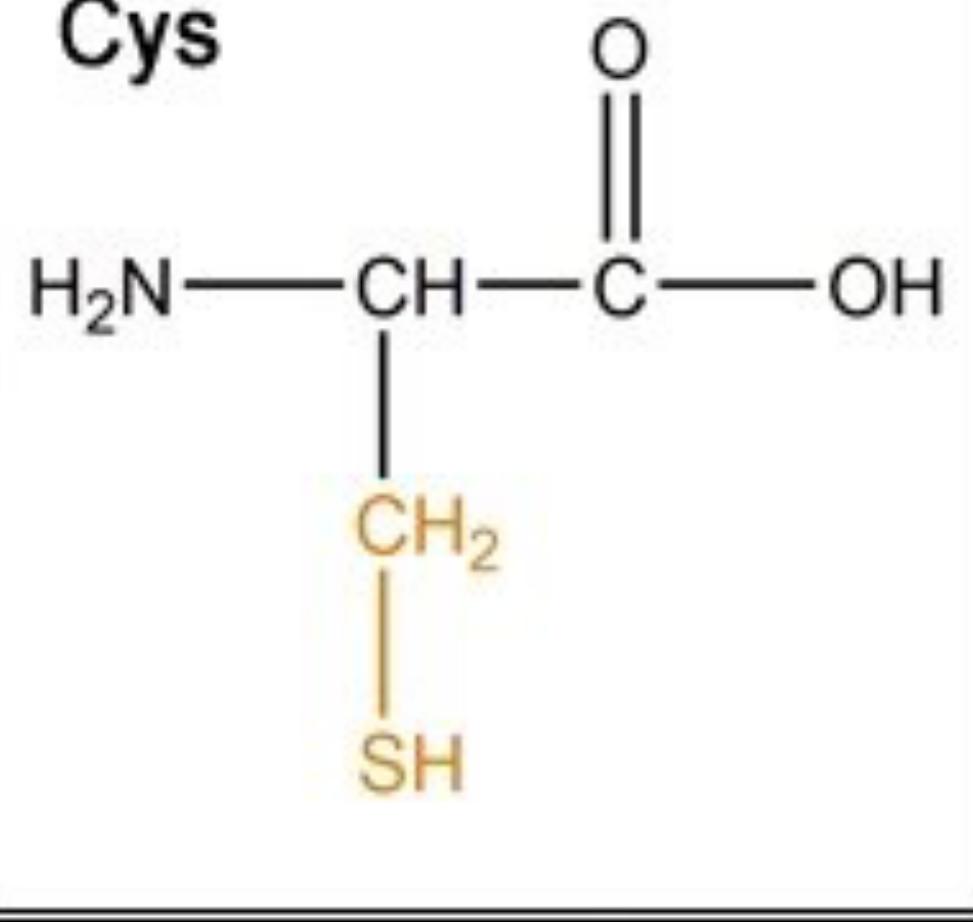
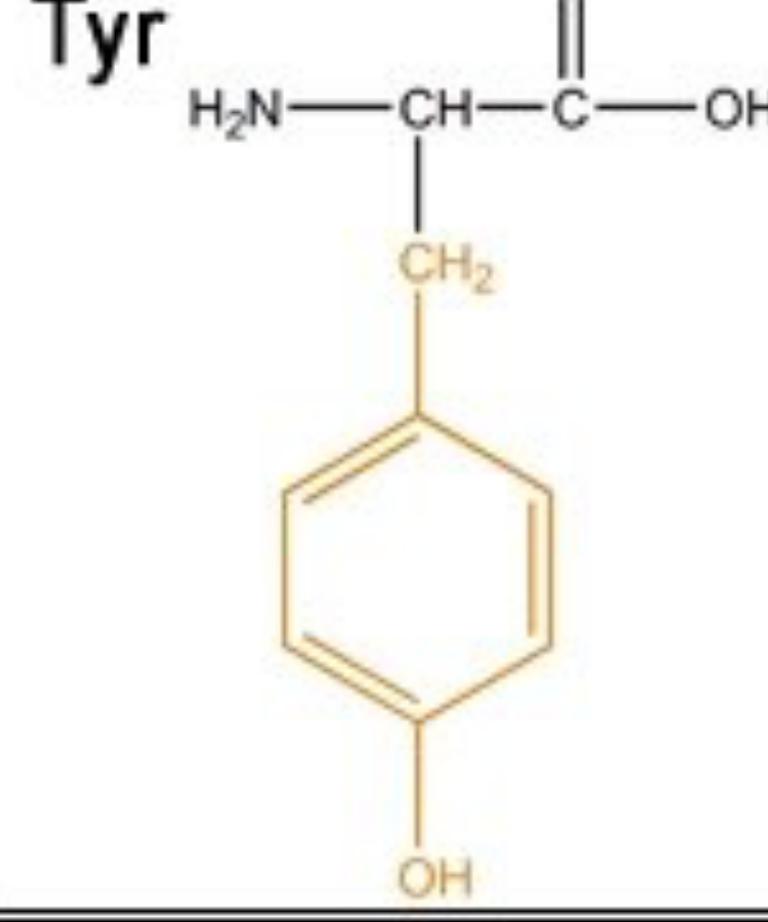
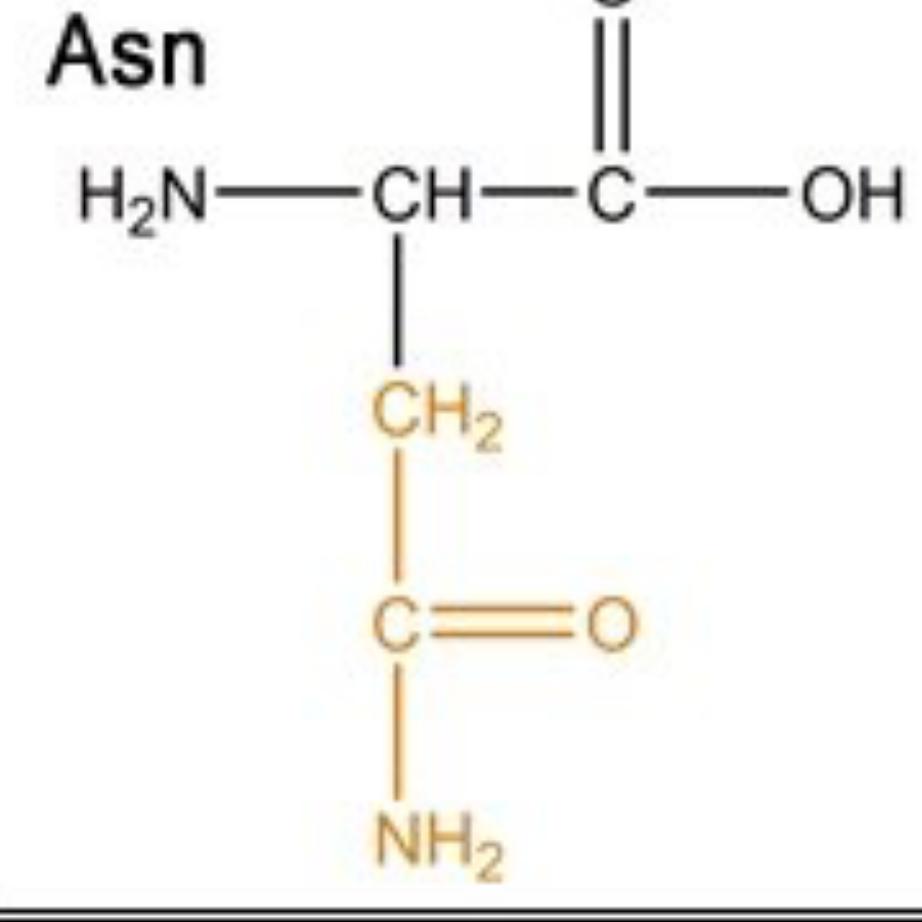
- Amino acid side chains
- Differ by their side chains (R-groups)
- Determine chemical properties

AMINO ACID BACKBONE



AMINO ACID SIDE CHAIN (R-GROUP):

PROTEINS

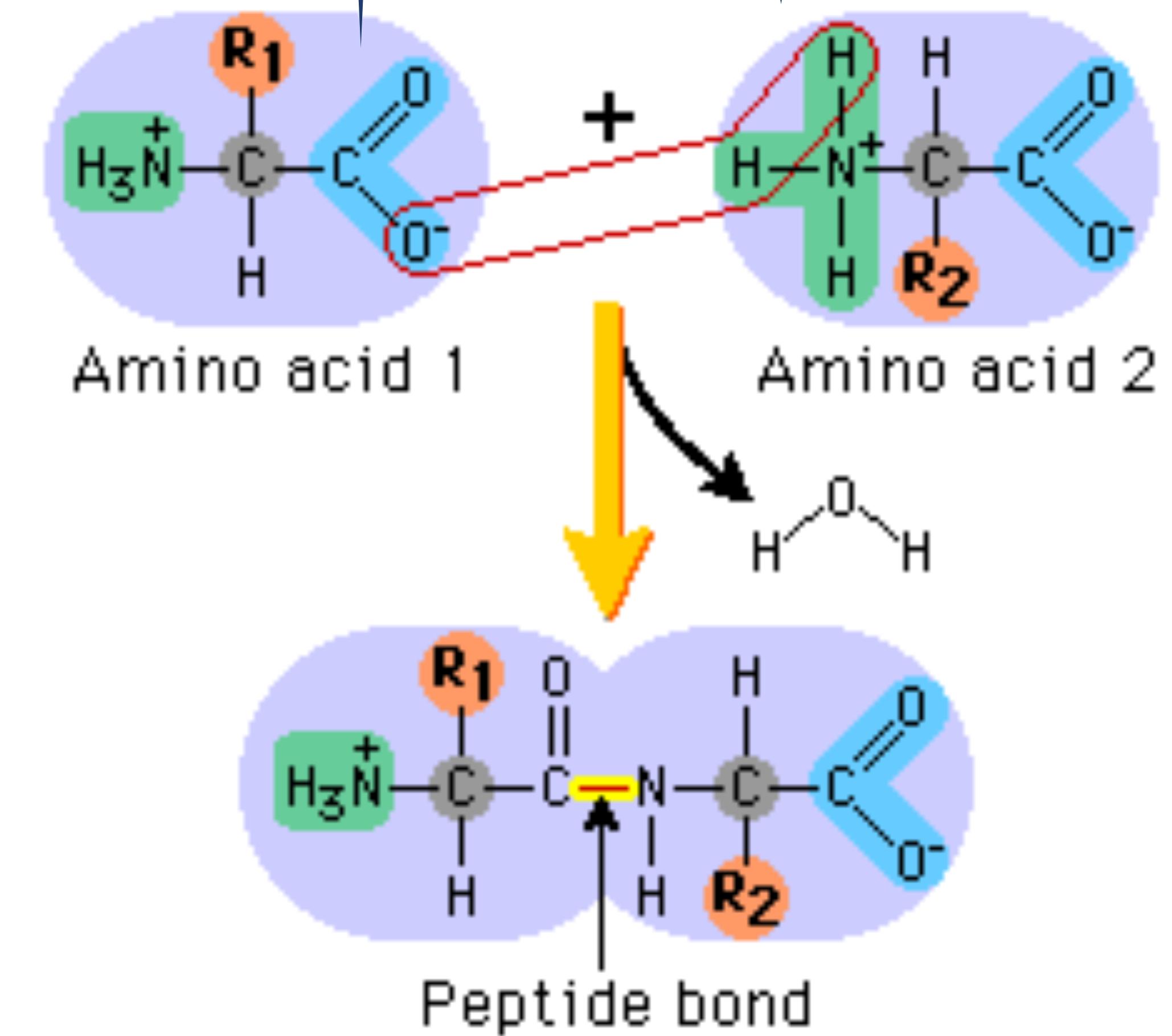
| | | | | |
|--|---|---|---|---|
| Met | Phe | Pro | Asp | Glu |
|  |  |  |  |  |
| Ser | Thr | Cys | Tyr | Asn |
|  |  |  |  |  |
| Cys | Trp | Lys | Arg | His |

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

CARBOXYL GROUP

AMINO GROUP



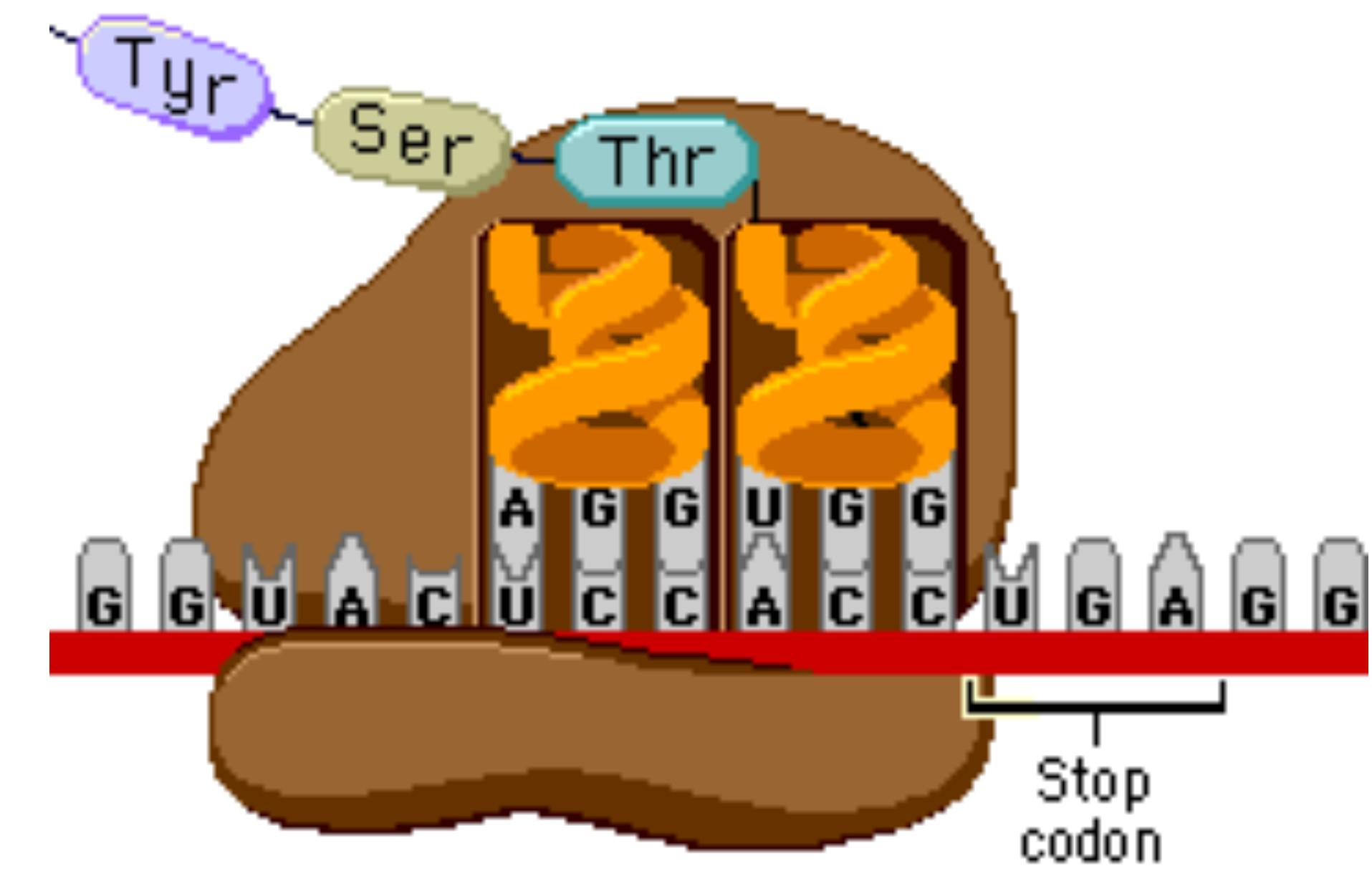
PROTEINS

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



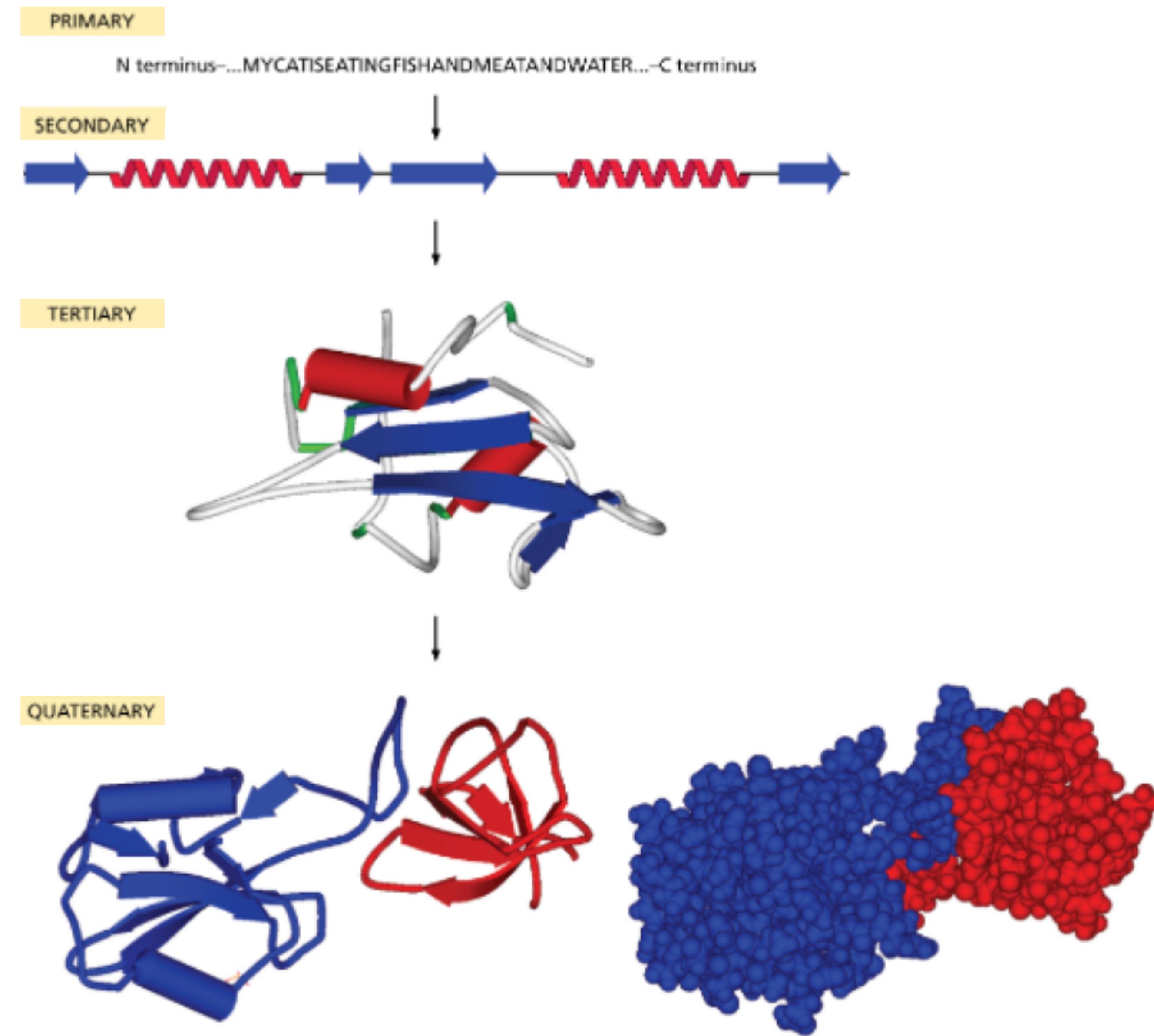
PROTEINS

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



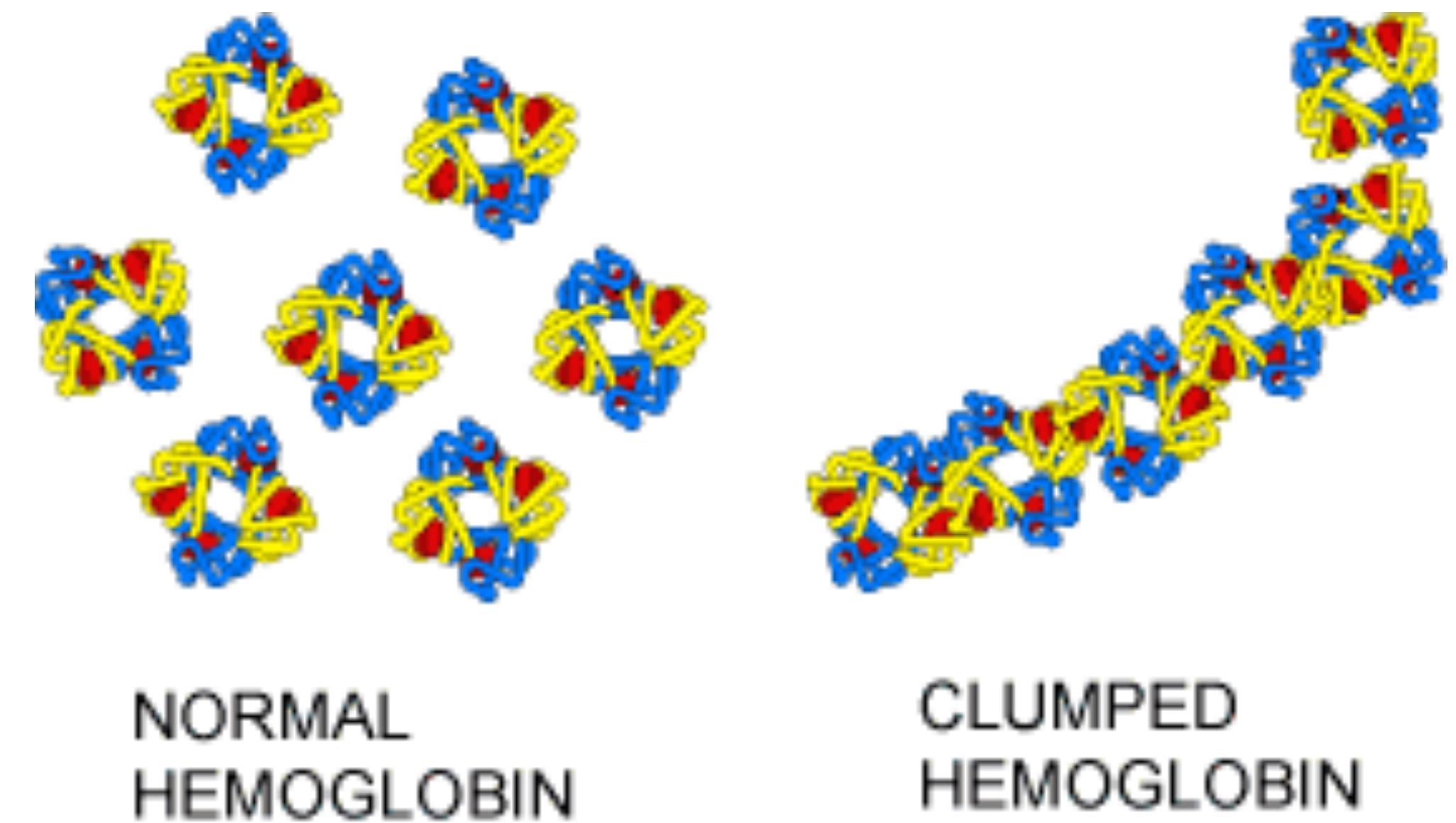
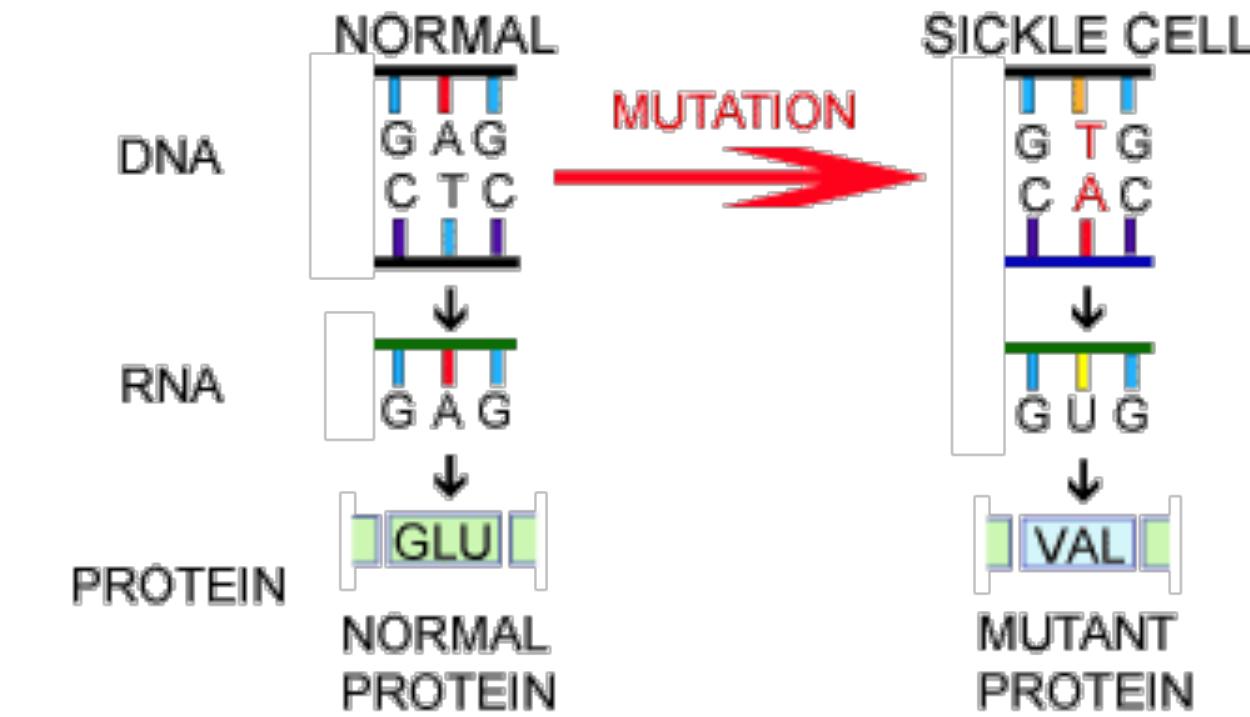
PROTEINS

- 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



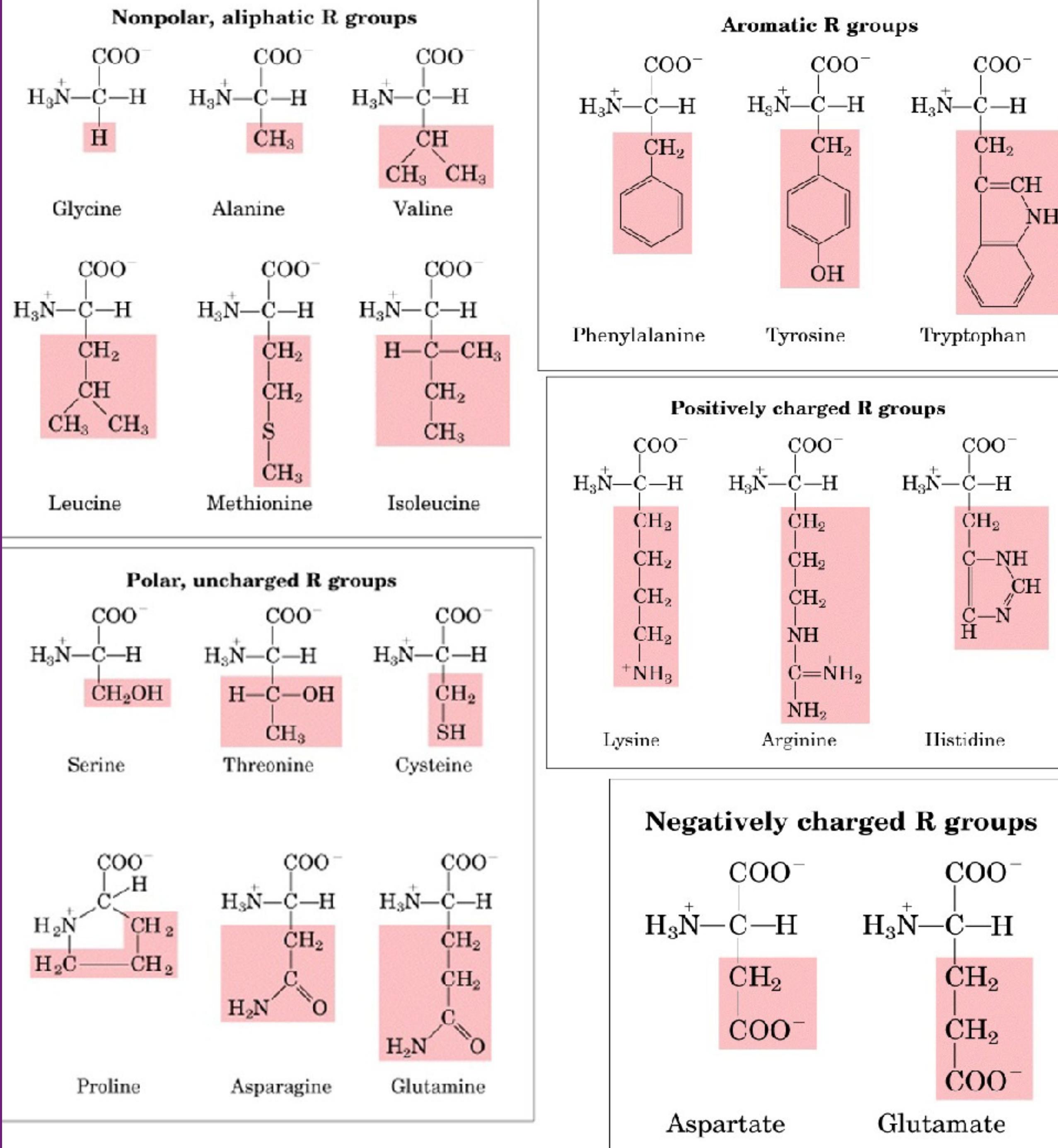
TRANSCRIPTIONS & TRANSLATION

- 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 |

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

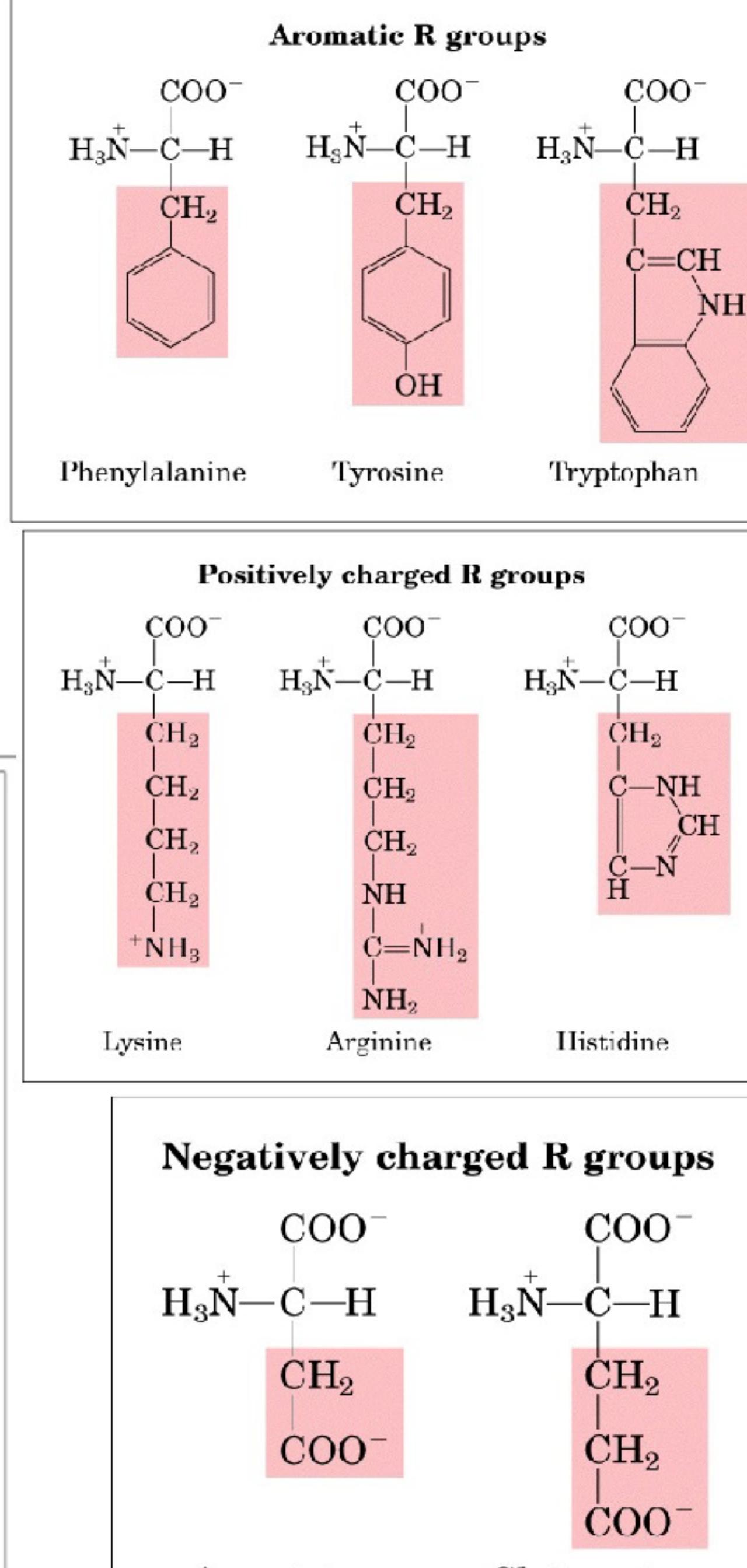
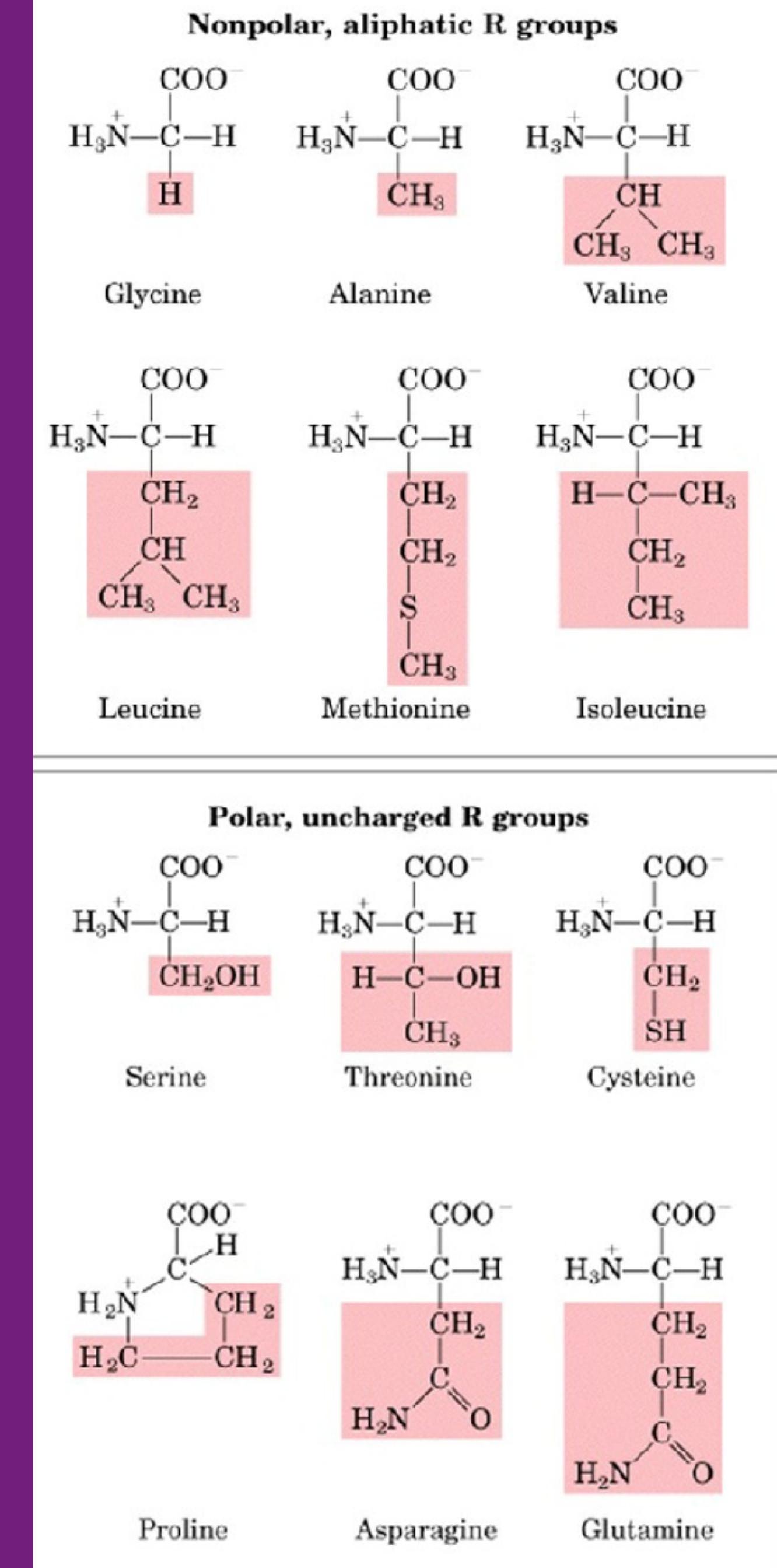
- Mutations in protein sequence

- Non-conserved

- results in different side chain

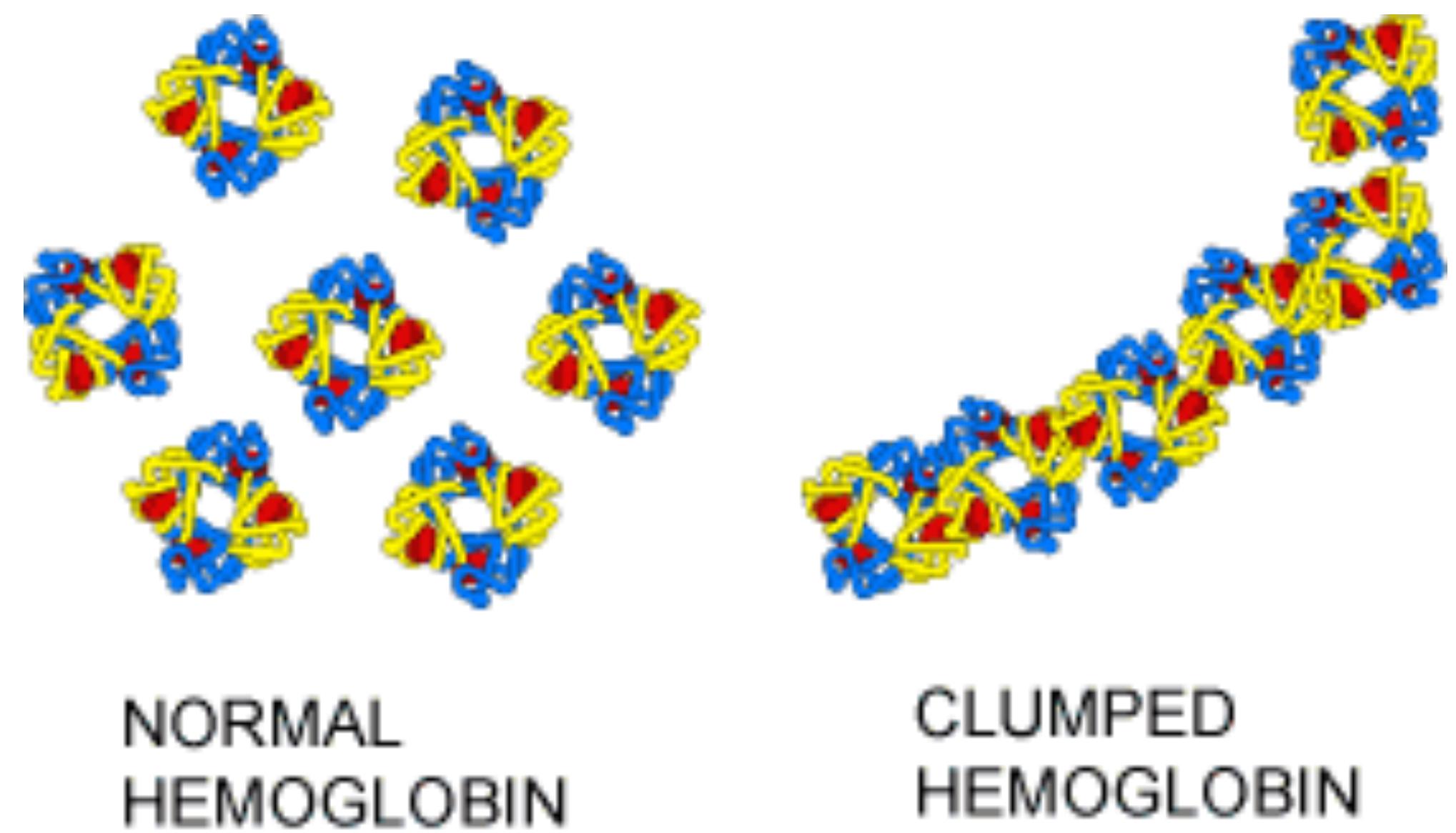
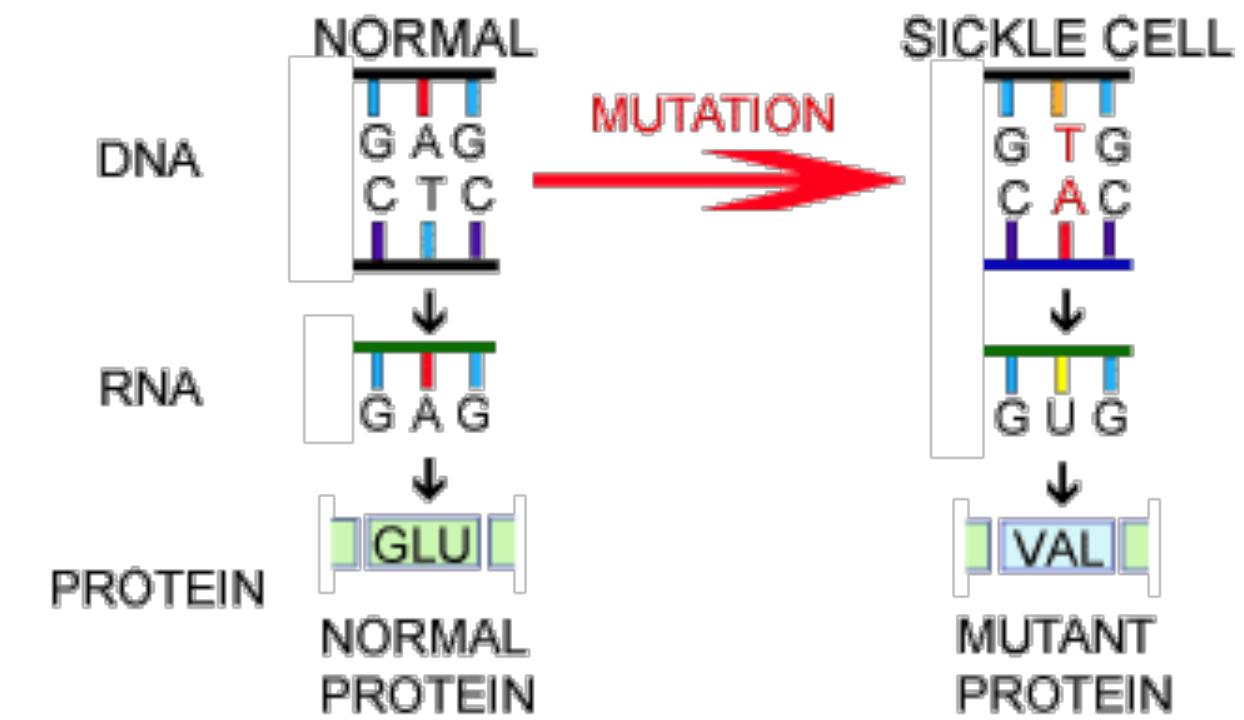
- Conserved

- 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

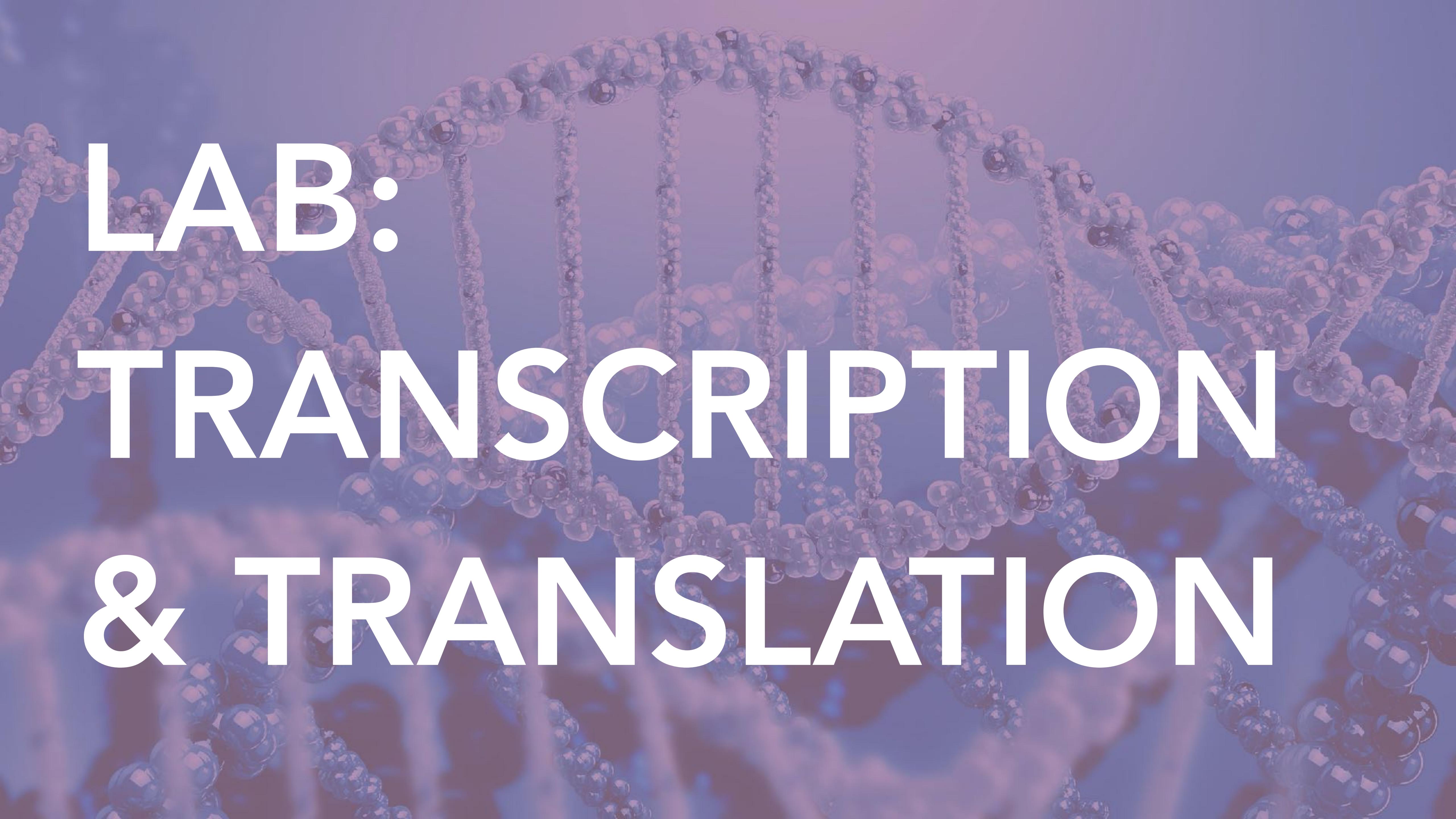
BIOINFORMATICS

(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' - CCGATGTATAAGAC - 3'

DNA IS READ 5' TO 3'

GENES ARE TRANSCRIBED FROM 5' TO 3'

TRANSCRIPTION & TRANSLATION

- DNA is complementary and anti-parallel

DNA STRANDS ARE
ANTI-PARALLEL,
PARALLEL BUT
OPPOSITE

*Gene or
coding or
sense* strand

5' - CCGATGTCA**TAAAGAC** - 3'

*Template or
non-coding or
anti-sense* strand

3' - GGCTACAGTATTCTG - 5'

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

TRANSCRIPTION & TRANSLATION

- Translating DNA into proteins

*Gene or coding or
sense strand*

5' - CCGATGTCATAAGAC - 3'

*tRNAs
anticodons*

3' - GGC UAC AGU AUU CUG - 5'


mRNA

5' - CCGAUGUCAUAAGAC - 3'


*Template or
non-coding or anti-
sense strand*

3' - GGCTACAGTATTCTG - 5'


TRANSCRIPTION & TRANSLATION

START
CODON

- How do we know where to start translation?

*Gene or coding or
sense strand*

5' - CCGATGTCATAAGAC - 3'

*tRNAs
anticodons*

3' - GGC UAC AGU AUU CUG - 5'

mRNA

5' - CCGAUGUCAUAAGAC - 3'

*Template or
non-coding or anti-
sense strand*

3' - GGCTACAGTATTCTG - 5'

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

STOP CODONS

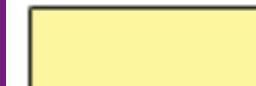
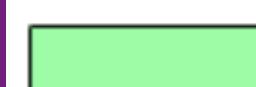
| First Position 5' | Second Position | | | | Third Position 3' |
|----------------------|----------------------------------|----------------------------------|--|-------------------------------------|----------------------|
| | U | C | A | G | |
| U | UUU F UUC F UUA L UUG L | UCU S UCC S UCA S UCG 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 Y GUC Y GUA Y GUG Y | 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 |

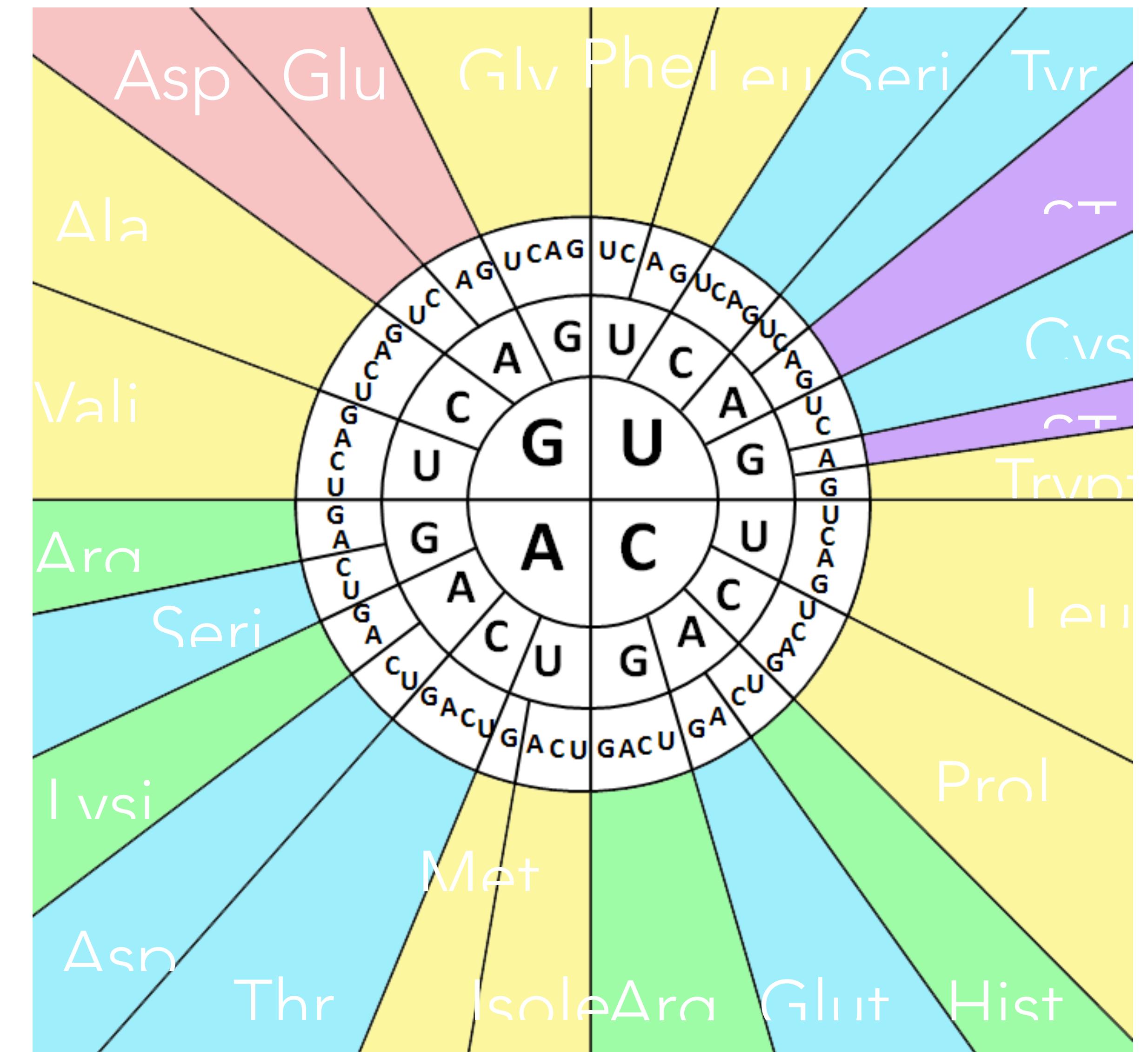
START CODON

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.

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

the cat ate the rat.

PERIOD IS "STOP CODON"

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

t **he**c ata tet her at.

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

th **e**ca tat **et**h era t.

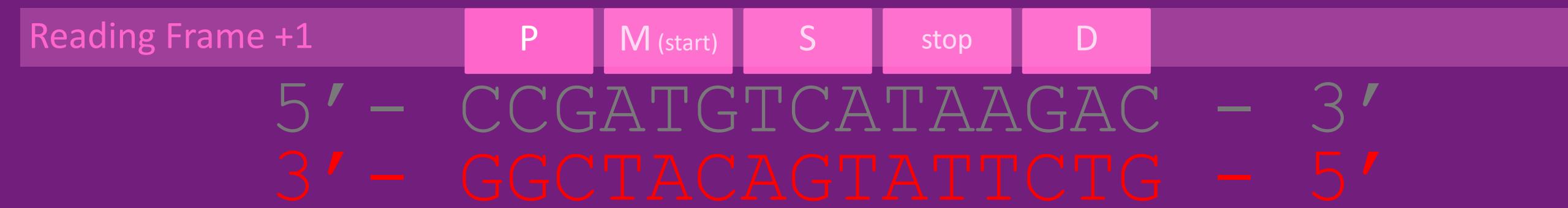
Reading Frames -1, -2 & -3 would be like reading the sentence "backwards."

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

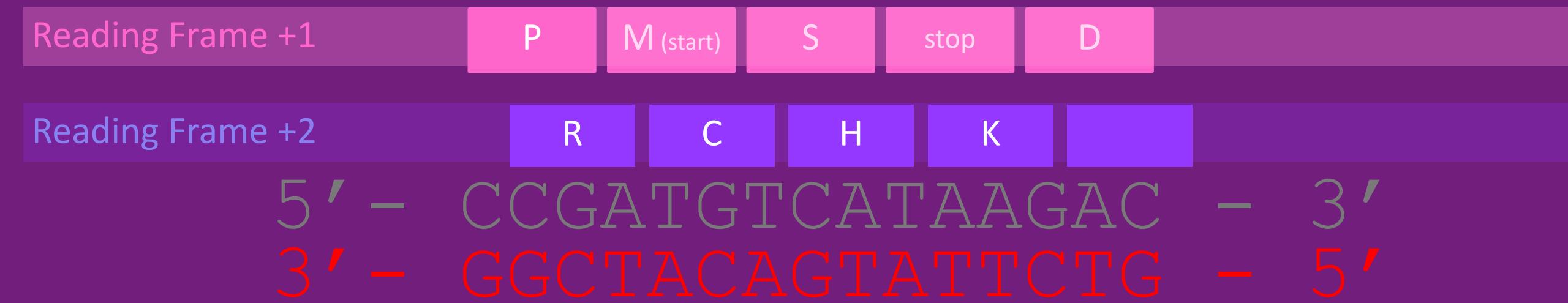
TRANSCRIPTIONS & TRANSLATION

- Reading frames



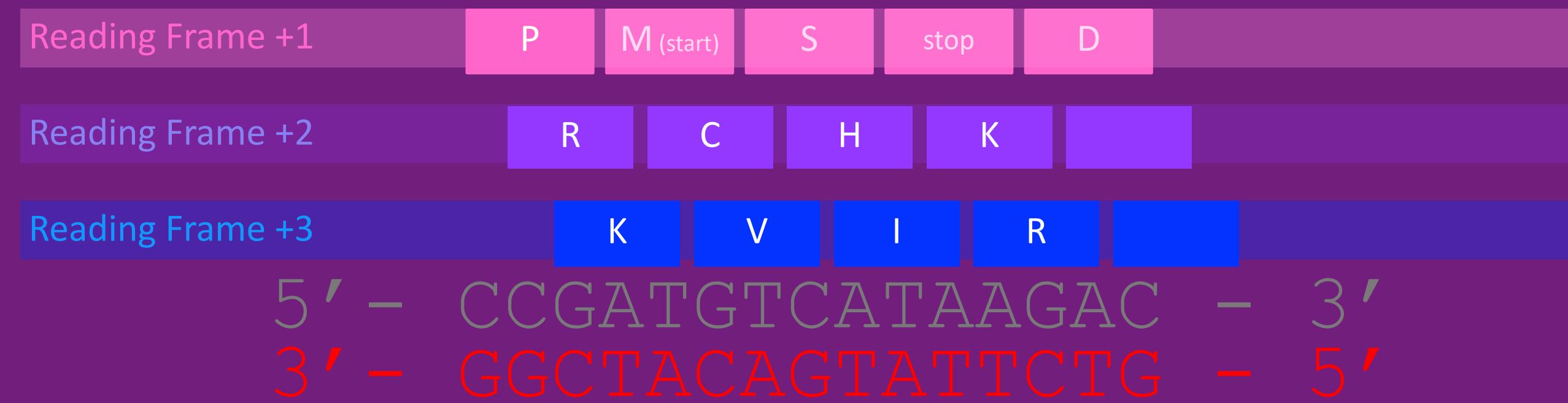
TRANSCRIPTIONS & TRANSLATION

- Reading frames



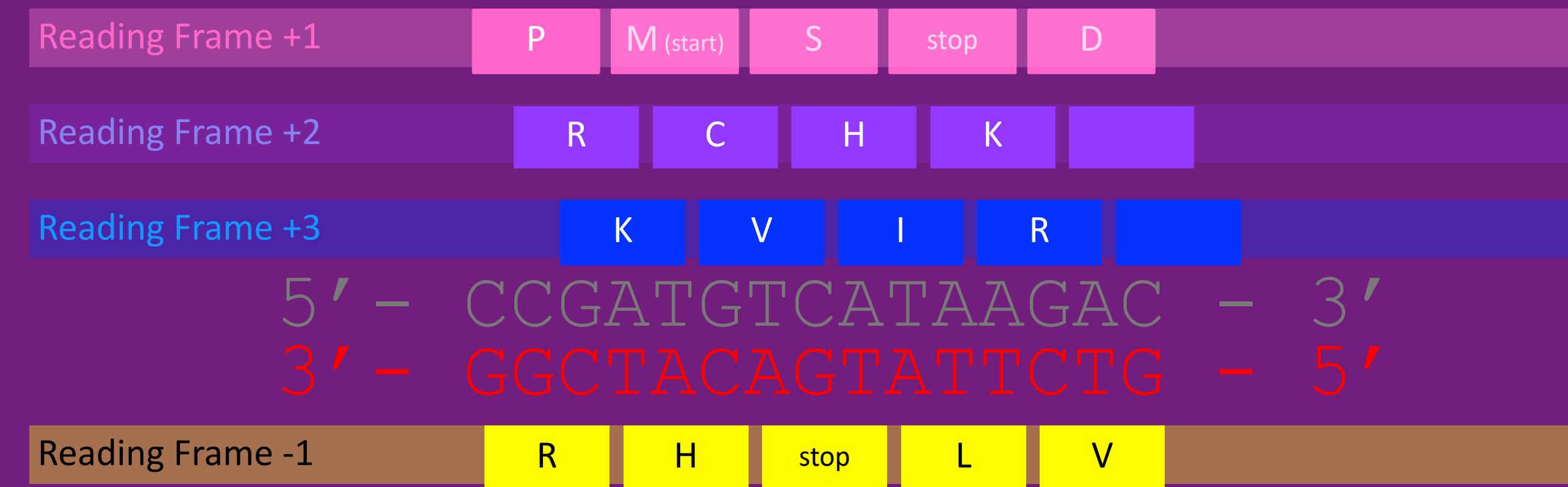
TRANSCRIPTIONS & TRANSLATION

- Reading frames



TRANSCRIPTIONS & TRANSLATION

- Reading frames



TRANSCRIPTIONS & TRANSLATION

- Reading frames



TRANSCRIPTIONS & TRANSLATION

- Possible Sequences

- PMS-stop-D



- RCHK



- KVIR



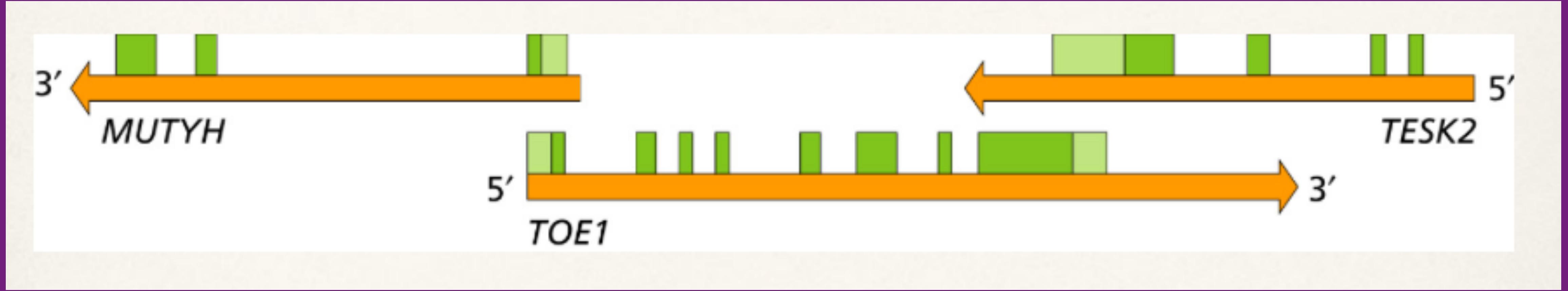
- VL-stop-HR



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

Jupyter Notebook
Quickstart

Architecture Guides

Narratives and Use
Cases

IPython

Installation,

Installing Jupyter Notebook

Contents

- [Prerequisite: Python](#)
- [Installing Jupyter using Anaconda and conda](#)
- [*Alternative for experienced Python users: Installing Jupyter with pip*](#)

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

IPYTHON QUICK START

■ ■ ■

Anaconda Navigator

ANACONDA NAVIGATOR BETA

i Sign in to Anaconda Cloud

Home

Environments

Learning

Community

My Applications

Refresh

jupyter notebook 4.2.1

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Launch

IP[y]: qtconsole 4.2.1

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Launch

spyder 2.3.9

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Launch

glueviz 0.8.2

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Install

orange-app 1.0.1

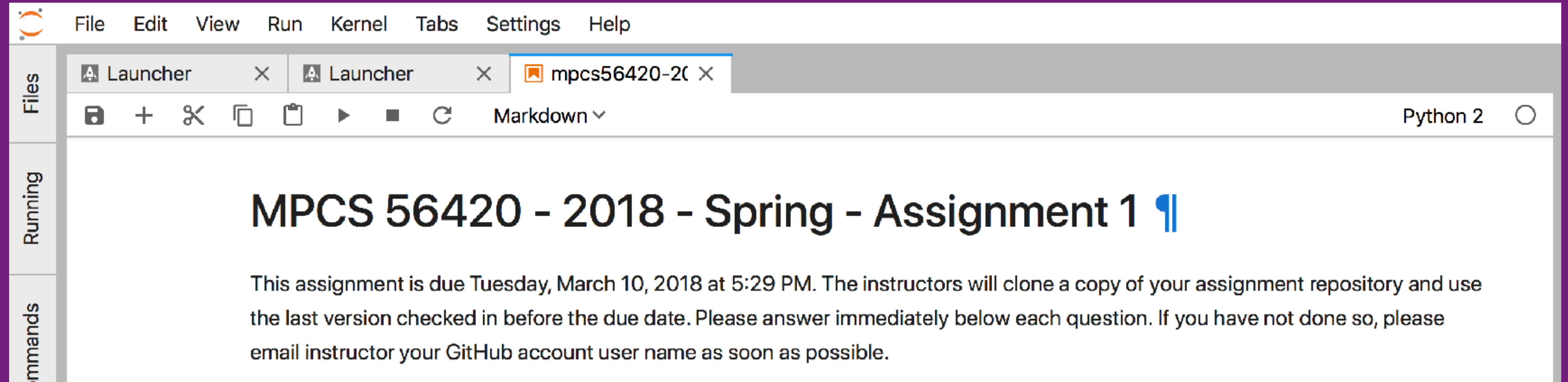
The screenshot shows the Anaconda Navigator interface. On the left, there's a sidebar with icons for Home, Environments, Learning, and Community. The main area is titled 'My Applications' and lists several data science tools: Jupyter notebook (version 4.2.1), IP[y]: qtconsole (version 4.2.1), spyder (version 2.3.9), and glueviz (version 0.8.2). Each application has a small icon, a version number, a brief description, and a 'Launch' or 'Install' button. A 'Refresh' button is located at the top right of the application list.

IPYTHON QUICK START

```
tbinkowski — jupyter_mac.command — python -bash — 127x22
```

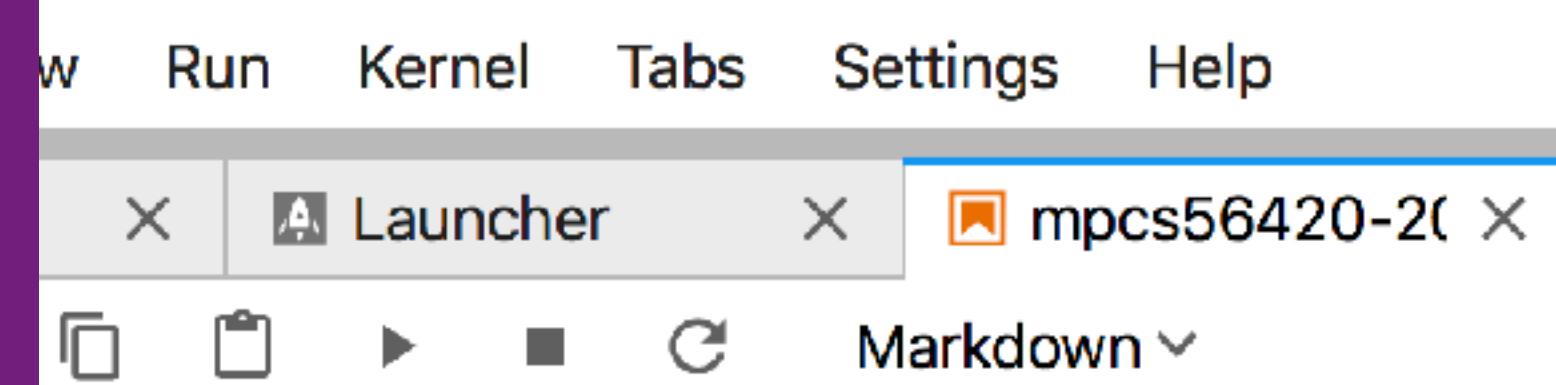
```
ski@TABinkosMBP2013 ~
Users/tbinkowski/anaconda/bin/jupyter_mac.command ; exit;
8:00.089 NotebookApp] Unrecognized JSON config file version, assuming version 1
8:00.479 NotebookApp] [nb_conda_kernels] enabled, 1 kernels found
8:00.895 NotebookApp] ✓ nbpresent HTML export ENABLED
8:00.895 NotebookApp] ✗ 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

MPCS 56420 - 2018 - Spring - Assignment 1

This assignment is due Tuesday, March 10, 2018 at 5:29 PM. The instructors will clone a copy of your assignment repository to their local machine. You must commit and push the last version checked in before the due date. Please answer immediately below each question. If you have not done so already, please let your instructor know via email and provide your GitHub account user name as soon as possible.

1.

This course is composed of students from a variety of backgrounds and experiences. Compose a brief introduction about yourself, including your educational background and work experience. Specifically, let us know about your background (if any) in biology. Next, please share your interests and motivation for taking this course and what you hope to get out of it. Finally, if there is any particular topic that you would like us to cover in class, please let us know and note of it.

Type your answer here.

2.

FASTA format is a file format used to exchange information between genetic sequence databases. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column.

BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
AUTUMN 2020
SESSION 1C



THE UNIVERSITY OF
CHICAGO