Instructor: Sael Lee

CS549 Spring – Computational Biology

# LECTURE 1:
# INTRODUCTION TO COMPUTATIONAL BIOLOGY

Resources used: Lecture slides from Steven Skiena's Computational Biology class and Daisuke Kihara's Protein Bioinformatics class

# WATCH A VIDEO

*Cellular Vision: The Inner Life of a Cell [8.5 min]*
TED talk: David Bolinsky: Visualizing the wonder of a living cell

"The Inner Life of the Cell is an 8.5-minute 3D computer graphics animation illustrating the molecular mechanisms that occur when a white blood cell in the blood vessels of the human body is activated by inflammation (Leukocyte extravasation). It shows how a white blood cell rolls along the inner surface of the capillary, flattens out, and squeezes through the cells of the capillary wall to the site of inflammation where it contributes to the immune reaction."[ wikipedia, Wired News. March 14, 2007]

# COURSE INFORMATION

- Please look at the course information sheet.
- Many of the class materials will be distributed through e-mail.

# WHY COMPUTATIONAL BIOLOGY?

- Computational biology is particularly exciting today because:
  - the problems are large enough to motivate efficient algorithms,
  - the problems are accessible, fresh and interesting,
  - biology is increasing becoming a computational science
- Computational biology is increasing of interest in both life science and computational science departments.
- Source of complex questions and real-life data.
  - Many problem ideas go from biology to CS: e.g. fragment assembly, sequence analysis, algorithms for phylogenic trees.
  - Many problem ideas go from CS to biology: e.g. sequencing by hybridization, DNA computing.

# COMPUTER SCIENTIST VS BIOLOGIST

- Similarity:
  - There are many different types of life scientists (biologists, ecologists, medical doctors, etc.), just as there are many different types of computational scientists (algorists, software engineers, statisticians, etc.).
- Many cultural differences
  - *Nothing is ever completely true or false in biology, where everything is either true or false in computer science /* mathematics.
  - Biologists are comfortable with the idea that all data has errors; computer scientists are not.
  - Biologists strive to understand the very complicated, very messy natural world; computer scientists seek to build their own clean and organized virtual worlds.

* Information extracted from Steve Skiena's slide. Thanks, Steve. ☺

- Biologists are *data driven;* while computer scientists are *algorithm driven.* Although nowadays CS are becoming more data driven.
- Biotechnology/drug companies are largely **science driven**, while the computer industry is **more engineering/marketing driven.**
- The Platonic ideal of a biologist runs a big laboratories with many people. The Platonic ideal of a computer scientists is a hacker in garage.
- Biologists are much more obsessed with being the **first to discover something**; computer scientists **invent** more than discover.
- Biologists can get/spend **infinitely more research money** than computational scientists.
- Biologists seek to publish in prestigious **journals** like *Science and Nature.* Computer scientists seek to publish in prestigious refereed **conference proceedings**.
  - One consequence is life science journals get refereed faster than computational science journals.

# BIOINFORMATICS JOURNALS / CONFERENCES & MEETINGS

## JOURNALS

- Nature/Science
- Proceeding Nat. Academy of Sci. USA

- J. Computational Biology
- PLOS Computational Biology
- J. Molecular Biology
- Bioinformatics
- BMC Bioinformatics
- Nucleic Acid Research
- Genome Research
- Proteins
- Protein Science

## CONFERENCES

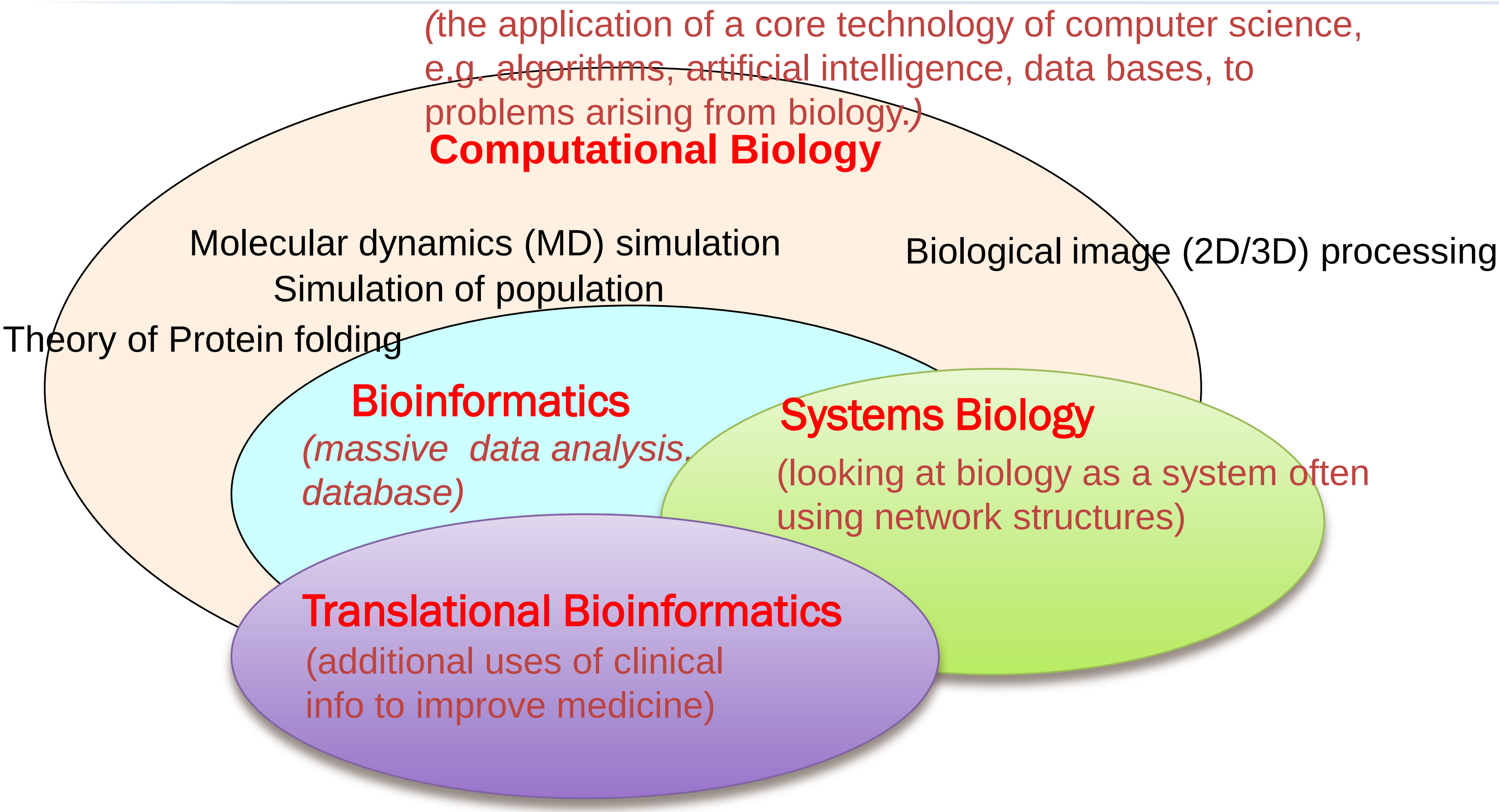
- ISMB (Int. Conf. on Intelligent Systems for Molecular Biology)
- RECOMB (Int. Conf. on Research in Molecular Biology)
- PSB (Pacific Symposium on Biocomputing)

- Genome Informatics (CSHL)
- 2013 Joint Summits TBI-CRI

# TOPICS COVERED

| Problems | | Methods |
|---|---|---|
| DNA-binding site identification | ← | Basic information theory |
| Finding nucleosome positions | ← | Mixture models |
| Gene network construction | ← | Graph structure learning |
| Graph searching | ← | Biomolecule searching |
| Feature selection | ← | Biomarker discovery |
| Feature extraction | ← | Protein surface analysis |

# Bioinformatics/Computational Biology



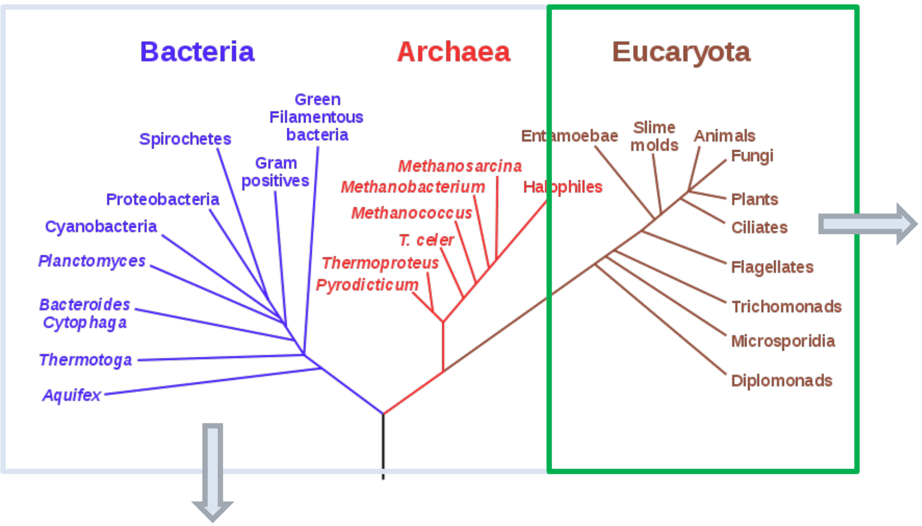

In this course, Bioinformatics ≈ Computational Biology

# YOUTUBE VIDEO ON CELL STRUCTURE

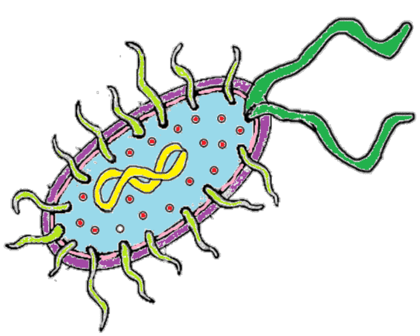
Biology: Cell Structure
Nucleus Medical Media
https://www.youtube.com/watch?v=URUJD5NEXC8

# PROKARYOTIC VS EUKARYOTIC

Phylogenetic Tree of Life




In prokaryotes, DNA is stored in what is essentially a single long loop.



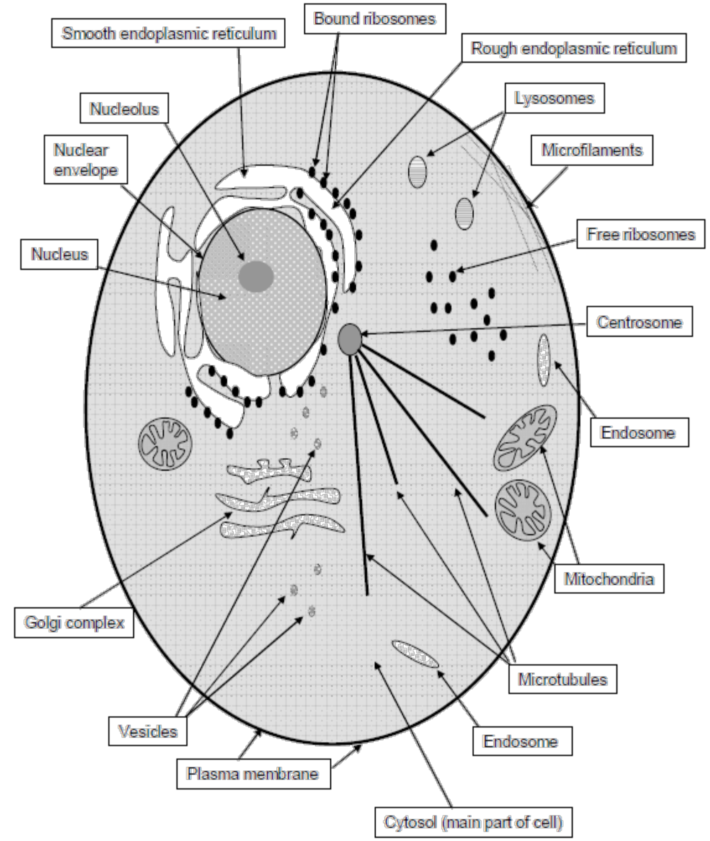



**Figure 3. Internal organization of a eukaryotic animal cell.**

In eukaryotes, DNA is stored in complexes called chromosomes, wrapped around protein complexes called nucleosomes.

W.W. Cohen, “A Computer Scientist’s Guide to Cell Biology”

# PROKARYOTIC VS EUKARYOTIC CHROMOSOMES

## PROKARYOTIC CHROMOSOMES

- Many contain a single circular chromosome.
- Chromosomes are condensed in the nucleoid via DNA supercoiling and the binding of various architectural proteins.
- DNA can interact with cytoplasm, so transcription and translation occur simultaneously.
- Most prokaryotes contain only one copy of each gene (i.e., they are haploid).
- Genomes are efficient and compact, containing little repetitive DNA.

## EUKARYOTIC CHROMOSOMES

- Eukaryotes contain multiple linear chromosomes.
- Chromosomes are condensed in a membrane-bound nucleus via histones.
- Transcription occurs in the nucleus, and translation occurs in the cytoplasm.
- Most contain two copies of each gene (i.e., they are diploid).
- Contain large amounts of noncoding and repetitive DNA.

A. Criswold, "Genome Packaging in Prokaryotes: the Circular Chromosome of *E. coli*", *2008 Nature Education*

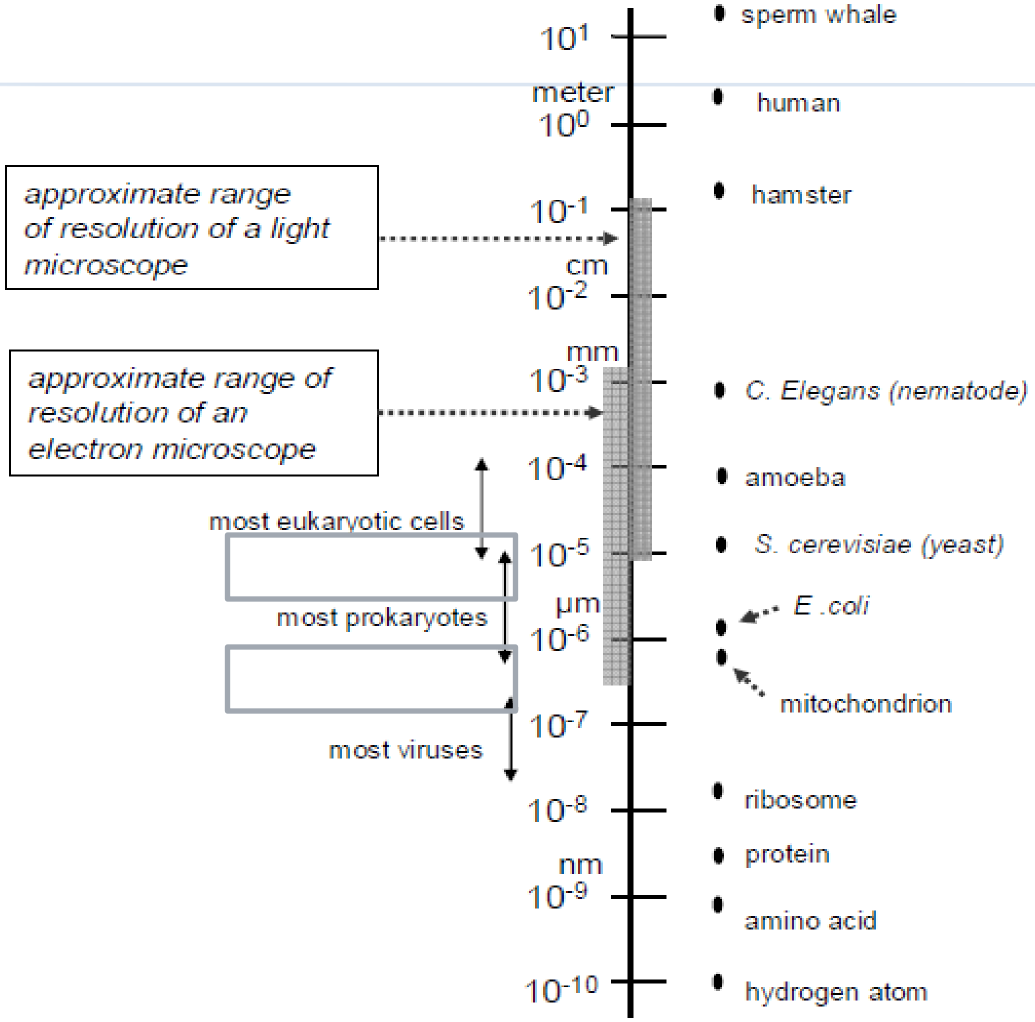



**Figure 2. Relative sizes of various biological objects.**

From W. Cohen, “A Computer Scientist’s Guide to Cell Biology”

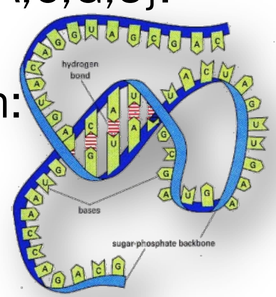
# INFORMATION CONTENT IN BIOLOGY

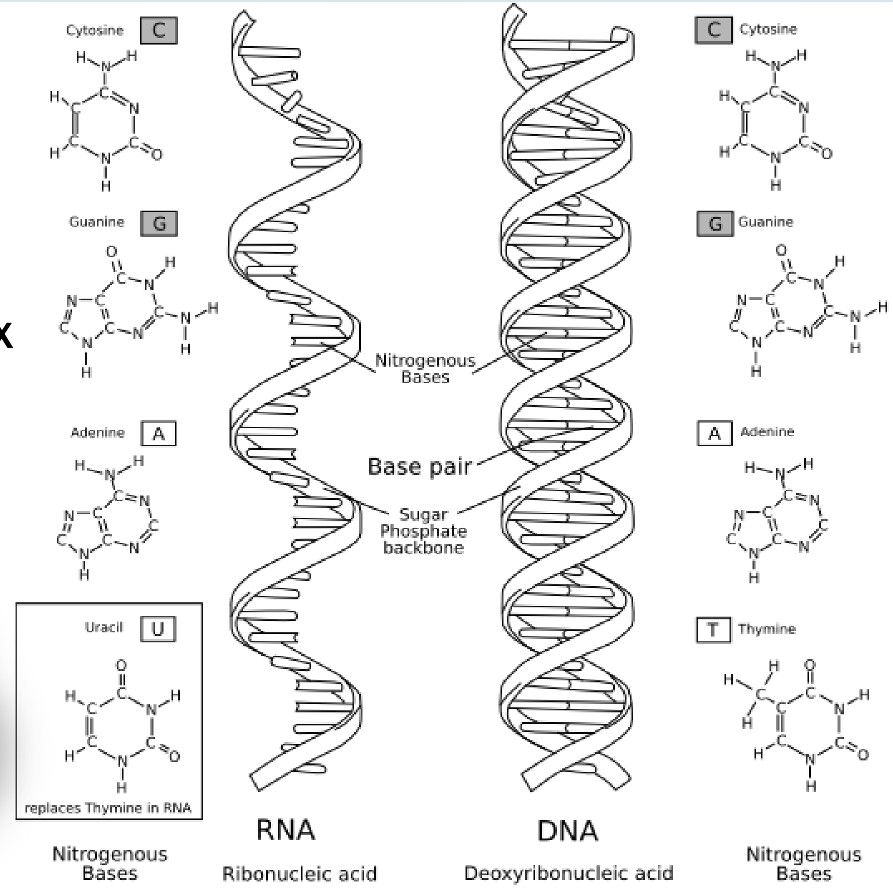
## DNA

- DNA sequences can be thought of as strings of bases on a four-letter alphabet, {A,C,G,T}, called nucleic acids.
- Binding: A=T; C-G
- Stable structural form : **double helix**

## RNA

- RNA sequences can also be thought of as strings of bases on a four-letter alphabet, {A,C,G,U}.
- Binding: A=U; C-G
- Stable structural form:

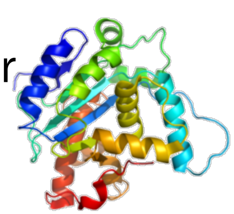



http://en.wikipedia.org/wiki/Nucleic_acids

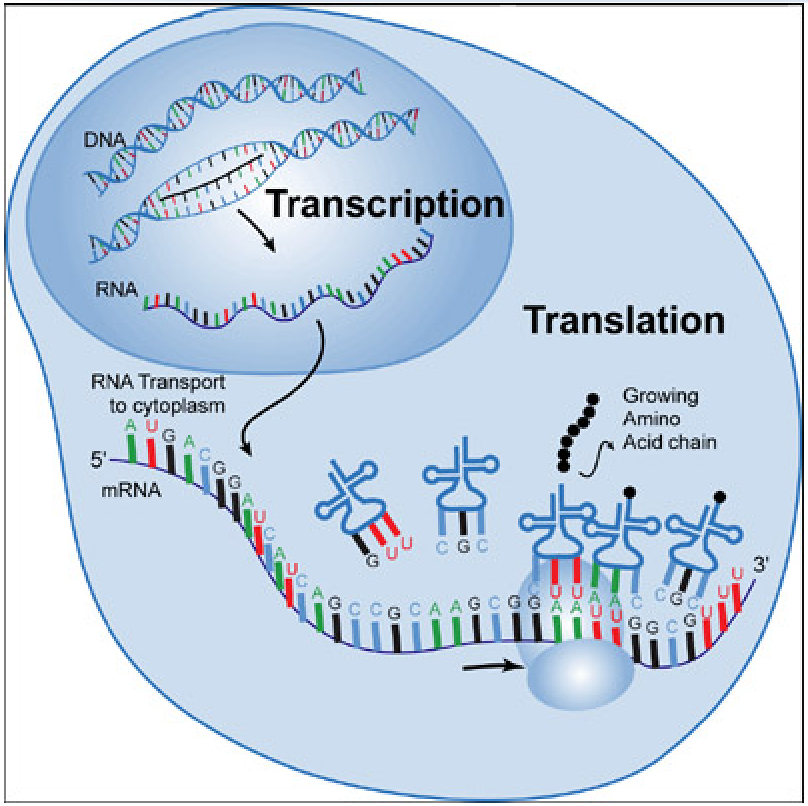
## Proteins

- Proteins sequence can be thought of as string of 21-letter alphabet
- Binding: covalent bonding, van der Waals force, hydrophobicity, etc.
- Stable structural form:

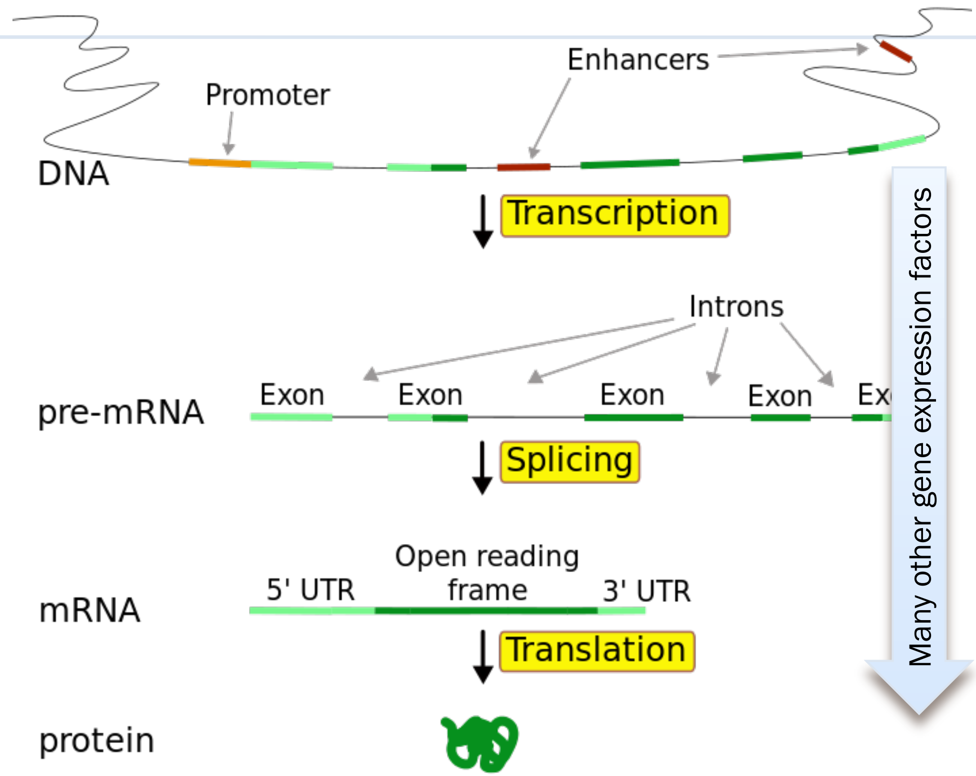


A LQNHTFLHTVYCQDGSPSVGLSEA …
DIFSCIVTHEPDRYTAIAYWVPRNALPS

# CENTRAL DOGMA OF BIOLOGY




http://www.tokresource.org/tok_classes/biobiobio/biomenu/transcription_translation/transcription_2.jpg




Youtube: From DNA to protein - 3D [3min]

**Common Abbreviations**

- DNA: Deoxyribonucleic acid
- RNA: Ribonucleic acid
- mRNA: messenger RNA
- tRNA: transfer RNA
- rRNA: ribosomal RNA
- siRNA: Small interfering RNA

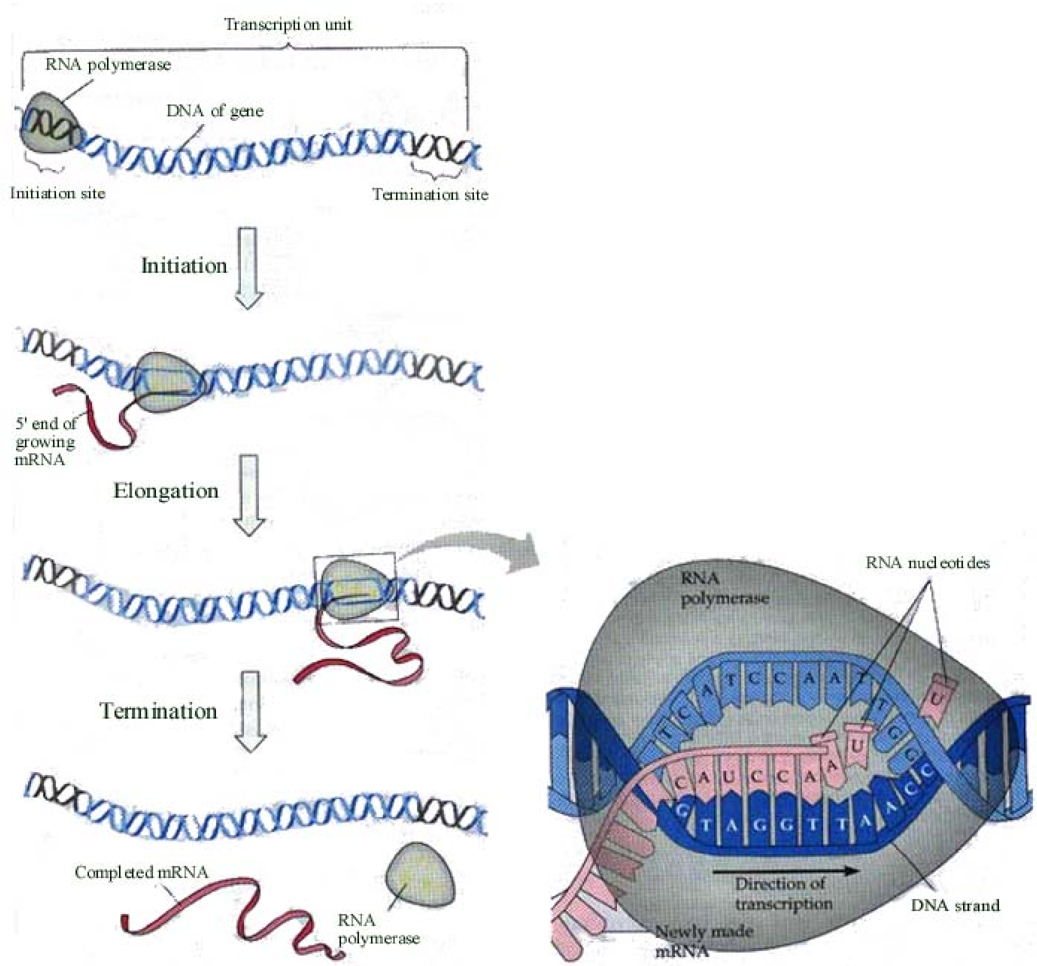
# TRANSCRIPTION PROCESS

**RNA polymerase** ‘unzips’ the DNA from **initiation site**.

**Elongation:** create a RNA strand by coping DNA strand

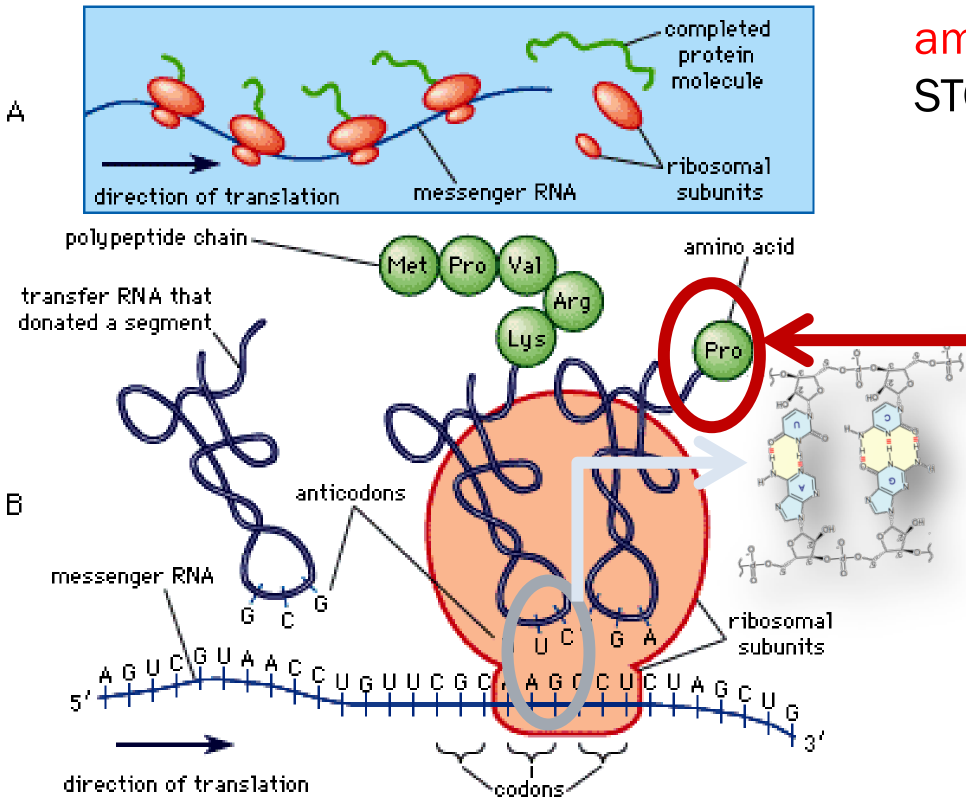
Stops at **termination site**

**Posttranslational modification**




http://biology12-lum.wikispaces.com/file/view/transcription,nbk.jpg/177394793/transcription,nbk.jpg

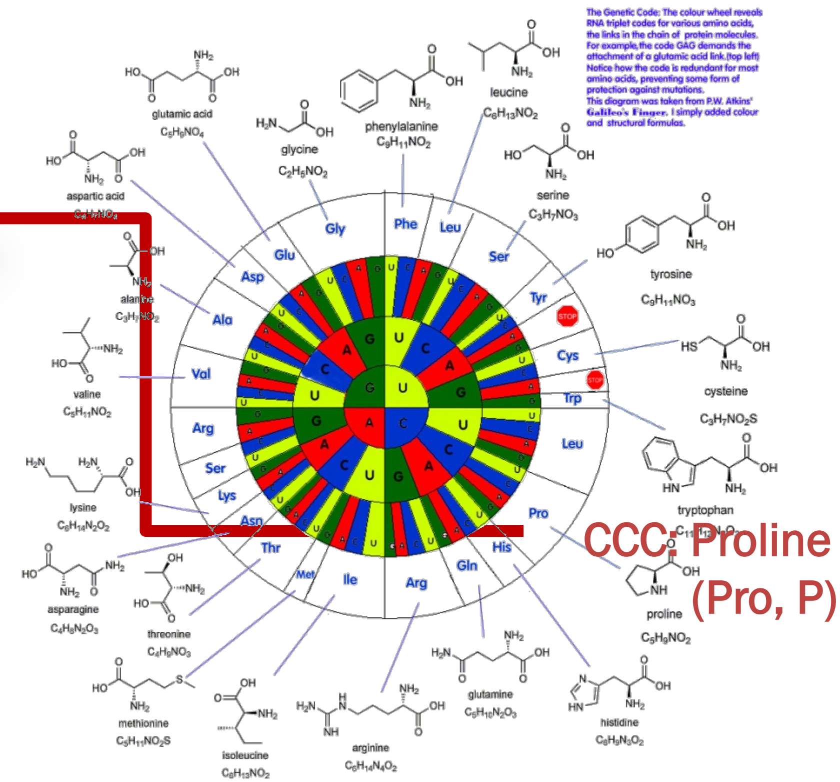
# TRANSLATION PROCESS




© 2006 Encyclopædia Britannica, Inc.

http://content.answcdn.com/main/content/img/BritannicaConcise/images/780.gif

*Codon*: Three nucleic acid coding one of 20 amino acid (alphabet of 20 size) + START & STOP CODEN

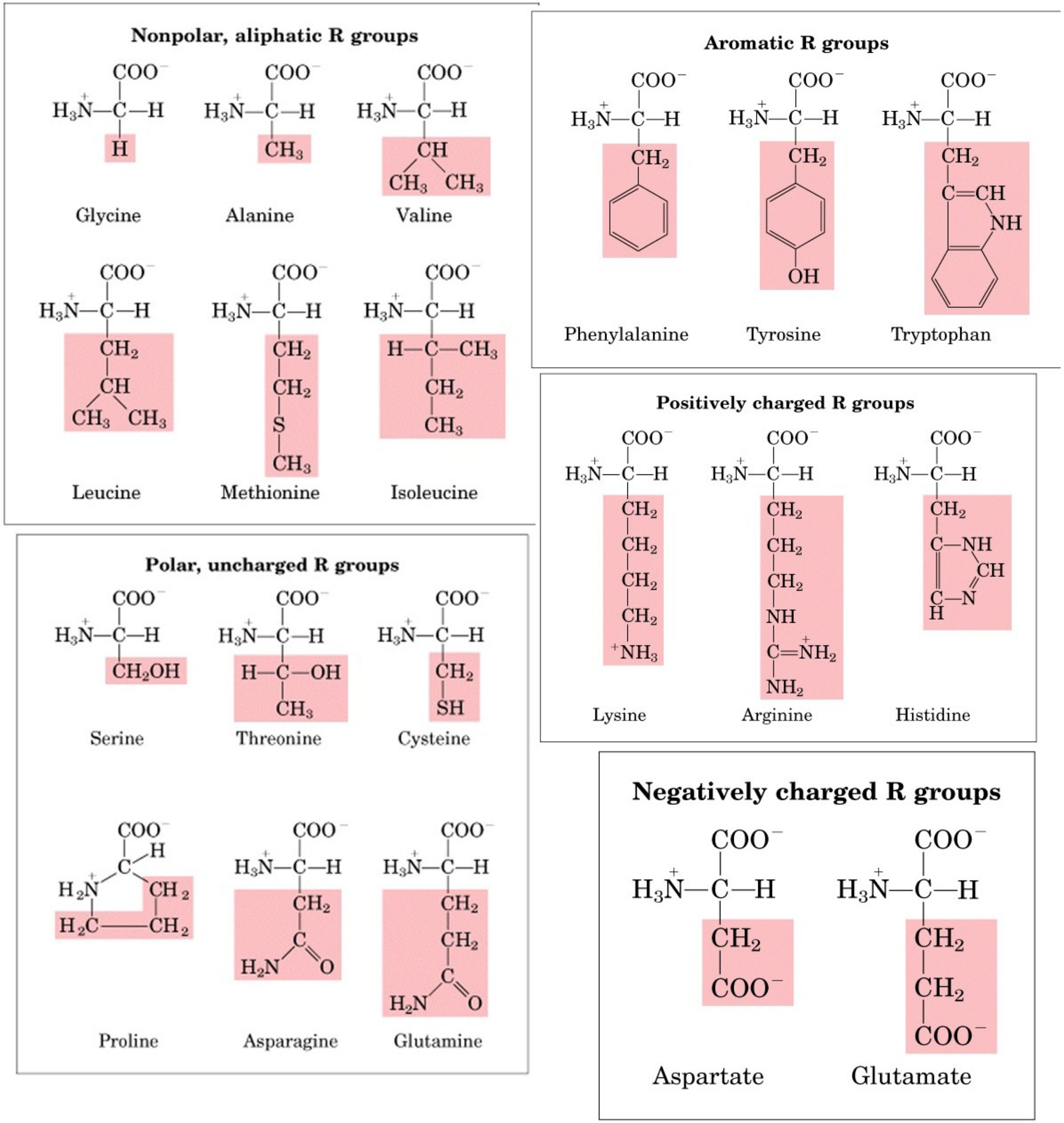



CCC: Proline (Pro, P)

Start codon: AUG ( also Methionine (Met, M))

Stop codon: UAA, UAG, UGA

# 20 AMINO ACIDS FORM ALL NATURAL PROTEINS



# PROTEIN FUNCTIONS

- Signaling
  - Interact to send signal to other proteins
- Involved in molecular recognition
  - Recognize and bind to specific molecules (DNA, RNA, proteins).
  - In the case of DNA they may recognize a specific sequence of nucleotides, or even a specific pattern
- Their function depends on the 3D structure
  - May be turned active and inactive
  - Protein conformation may change after binding to other molecules
- Molecular motors
  - Proteins may act as molecular motors through repeated changes in their 3D structure
  - Used for particle transportation or for cell locomotion
- Self-assembly
  - By binding to another protein, some new binding sites may be unveiled, for other proteins to bind, etc.

# MOLECULES INTERACT WITH OTHER MOLECULES IN THE CELL

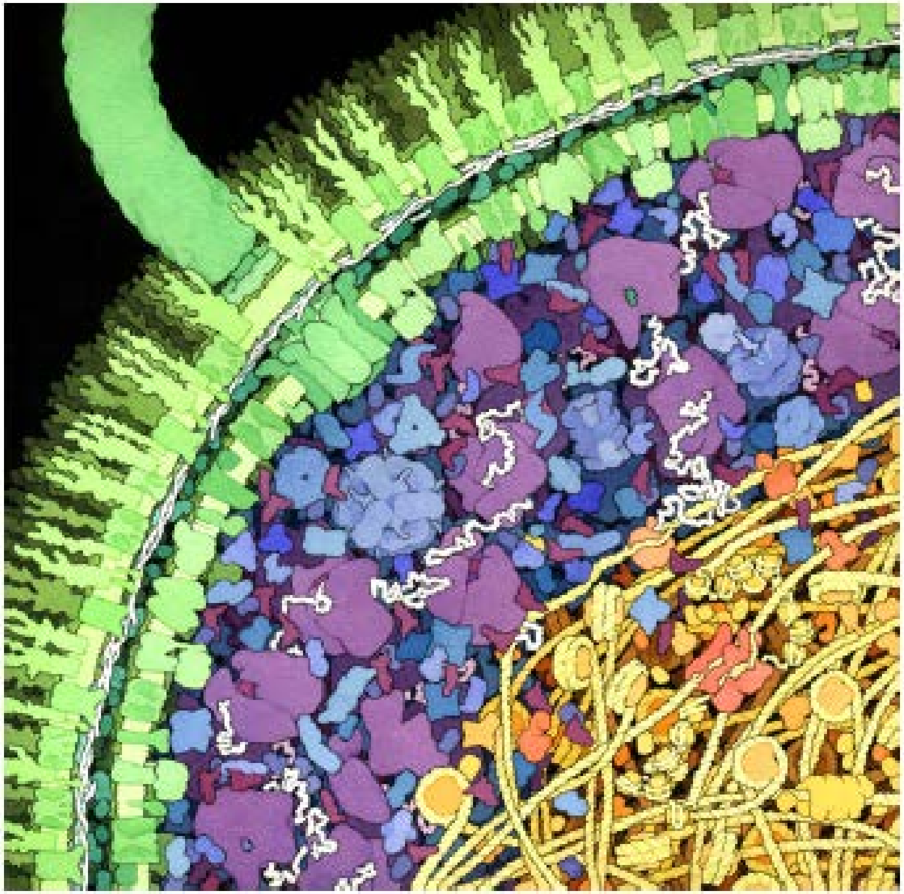


**Figure 1.** Representation of the approximate numbers, shapes and density of packing of macromolecules inside a cell of *Escherichia coli*. (Illustration by David S Goodsell; reprinted with permission.)

- Protein Interactions
- Biological macromolecules occur at concentration of 300-400g/liter
- 20-30% of the total volume

# FUNCTIONAL BASICS

- Function by molecular interaction or binding
- Classify by duration
  - Transient binding
  - Obligate binding
- Classify by components
  - Protein-protein binding
  - Protein-ligand binding
  - Protein-ion binding
  - Protein-DNA/RNA binding
  - Etc.
- Classify by function
  - Signaling
  - Structure forming
  - etc

# ANIMATIONS

TED talk: Drew Berry - Astonishing Molecular Machines
[14.27 min]

*Cellular Vision: The Inner Life of a Cell [8.5 min]*
TED talk: David Bolinsky: Visualizing the wonder of a living cell

Functions viewed: metabolite transportation, protein-protein binding, DNA replication, DNA ligase, microtubule formation/dissipation, protein synthesis, Kinesin protein delivering proteins