INTRODUCTION TO COMPUTATIONAL MOLECULAR BIOLOGY

Bioinformatics studies the flow of information in molecular biology

- Information flow from genotype to phenotype

  DNA ⇝Protein ⇝Function ⇝Organism ⇝Population ⇝DNA

- Experimental information flow for creating and testing models

  Hypothesis ⇝Experiment ⇝Data ⇝Conflict ⇝Hypothesis

Learning Objectives for this Course

1. Understand main computational data structures and algorithms in molecular biology

2. Implement and test some key algorithms that establishes context for rest of fields

Textbook

- David W. Mount, *Bioinformatics: Sequence and Genome Analysis*, Second edition (July, 2004), Cold Spring Harbor Laboratory Press,
  ISBN: 0879696877

- There are many others books in bioinformatics

Course Instructor

Name and Title Dr. Alioune Ngom, Associate Professor of Computer Science.

Office School of Computer Science, 5107 Lambton Towe
University of Windsor, 401 Sunset Avenue.

Contact Information

Phone: 519-253-3000 extension 3789

Fax: 519-973-7093

E-mail: angom@cs.uwindsor.ca

Url: http://www.cs.uwindsor.ca/~angom/cs558.html

Lectures Tuesday, 08:30–11:20 AM, DH 263

Office Hours Friday, 08:30–11:30 AM, or by appointment. Walk-ins with very short questions are encouraged anytime.

Course Prerequisites

- Basic probability and statistics
- Basic mathematics
- Ability to analyze and design algorithms
- Interest in biology (may have to do some reading)
- Ability to program at least in C, C++, Java, Perl or MatLab
- Unix, Linux, Windows
- Access to WWW, Acrobat Viewer

Course Grading

Class Participation: 5%

Presentation or Survey: 15%

Assignments: 25%

Take-home exam (midterm or final): 25%

Project: 30%. An in-depth effort on a particular aspect of bioinformatics. A relatively extensive literature search in the area is expected with a subsequent bibliography. Good projects are typically as follows

- Best: Some of your own original thinking and proposal of a method or approach to a given problem, typically well benefited by some computer simulation to bear out potential.
- Very Good: Starting from an in-depth study of some current techniques, strive to extend it through some new mechanism.
- Good: A study of a current problem/method with an in-depth analysis of its strength
- Bad: A description of a current model

The earlier you start the better. You should use your own initiative and the resources available to peruse and find any topic of interest to you, regardless of whether it will be discussed in class.

Letter Grading

The final letter grade, L, will be given from the numeric grade based on the following conversion rule:

Letter Grade	Numeric Grade Range
A+	$93 \leq G < 100$
A	$86 \leq G < 93$
A-	$80 \leq G < 86$
B+	$77 \leq G < 80$
B	$73 \leq G < 77$
B-	$70 \leq G < 73$
C+	$67 \leq G < 70$
C	$63 \leq G < 67$
C-	$60 \leq G < 63$
D+	$57 \leq G < 60$
D	$53 \leq G < 57$
D-	$50 \leq G < 53$
F	$35 \leq G < 50$
F-	$0 \leq G < 35$

Course Policies

Attendance and preparation Lecture attendance is mandatory and students are expected to come well-prepared for every class. Notetaking is encouraged to help understand ideas more deeply.

Assignment submission All assignments must be handed in to me in classroom at the beginning of the lecture on the due dates and in envelopes with the School of Computer Science and University of Windsor logo on them. **Late submission will not be accepted (tolerated)**. Students are responsible for making sure that I receive their assignments by or on the due dates. All assignments as well as envelopes must be clearly marked with the student name, student number, course name and number, section number and the instructor's name.

Academic honesty *You are expected to do all of your work on assignments and examinations individually. That is, collaboration on the assignments and/or plagiarism is not accepted; what you turn in should be your own work.* **Anyone found cheating on any graded assignment or examination will get no points at all for that homework assignment or question in exam**. *The instructor reserves the right to assign anyone involved in cheating a failing grade (F-) and will initiate the proceedings for disciplinary actions by the department and the university. This will be irrespective of who cheated from whom. In other words, you are responsible to protect your work from others.* **Please read the University of Windsor regulations on cheating**.

Makeup/Incomplete *Makeup work or incomplete grade are only given in unusual circomstances, and only when work has been completed satisfactorily up to the point when the incomplete was requested. If you suspect that you will be unable to attend an examination because of a* **valid and verifiable reason**, *you* **must** *give me a prior notice,* **at least** *one full day before the examination. Even if you are sick or face unavoidable circumstances, you* **must** *notify me or the department through phone, email, fax, etc. along with a valid documentary evidence. I* **must** *receive a* **proper documentary evidence within a week** *of the examination.* **In the absence of such notice and a proper documented proof, makeup examination(s) will not be allowed**. *Unless mentioned otherwise, all examinations will be closed book, closed notes and closed neighbors. Date and place for makeup examination will be announced at an appropriate time. It will be your responsibility to get the necessary information about the makeup examination.* **Please read the University of Windsor regulations**.

Appeal Students who wish to appeal an assignment or exam mark should do it within two weeks of the reception of the mark. I will be glad to remark your work and explain my marking scheme to you. Numerical errors in adding marks will be corrected when identified. In case of a total disagreement on a mark, you must then submit a formal appeal. **Please read the University of Windsor regulations on appealing**

BioInformatics

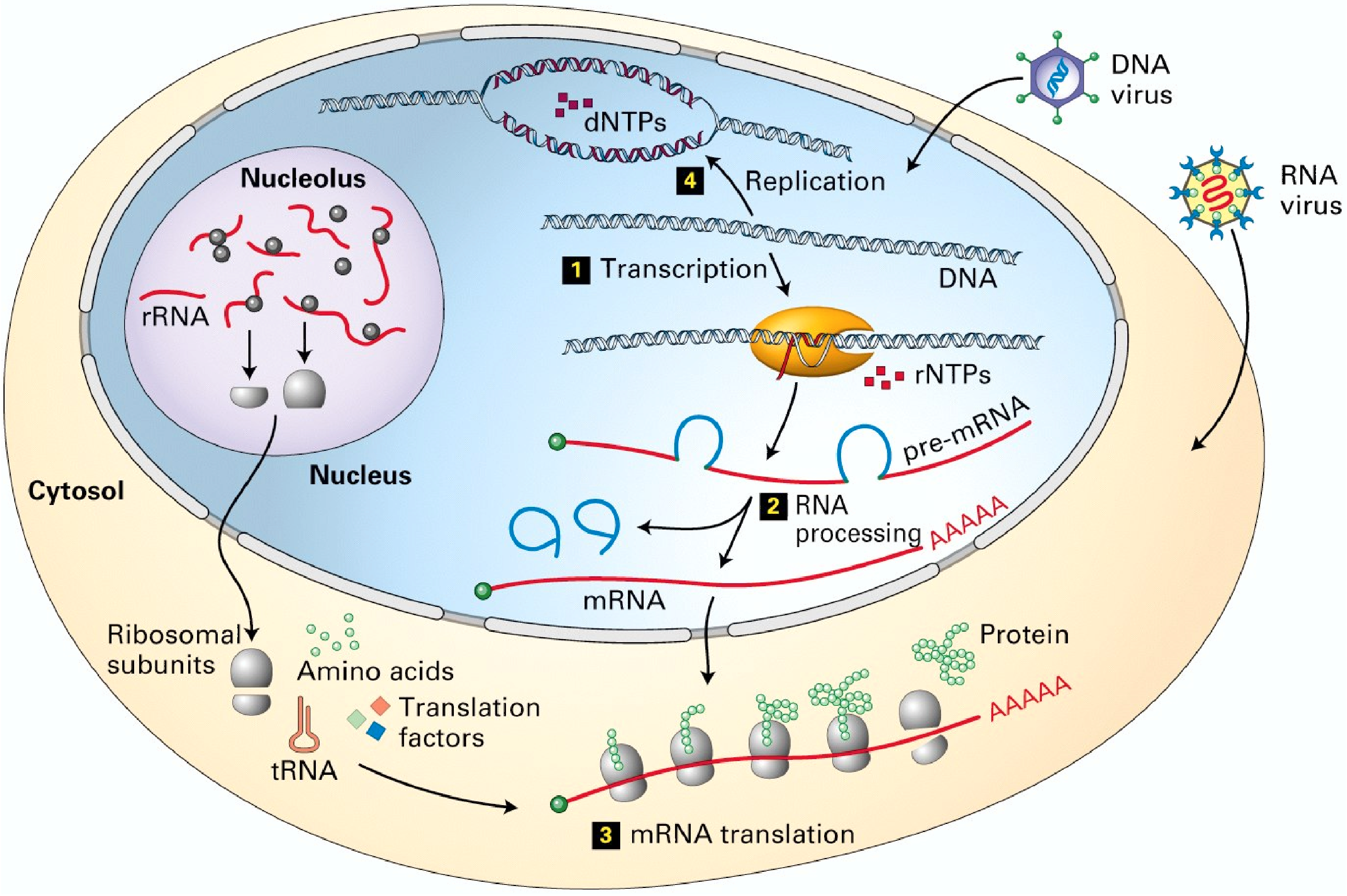
Application of computer and information technologies to problems in molecular biology

- Explosion of experimental data
- Difficulty in interpreting data
- Need for new paradigms for computing with data and extracting new knowledge from it

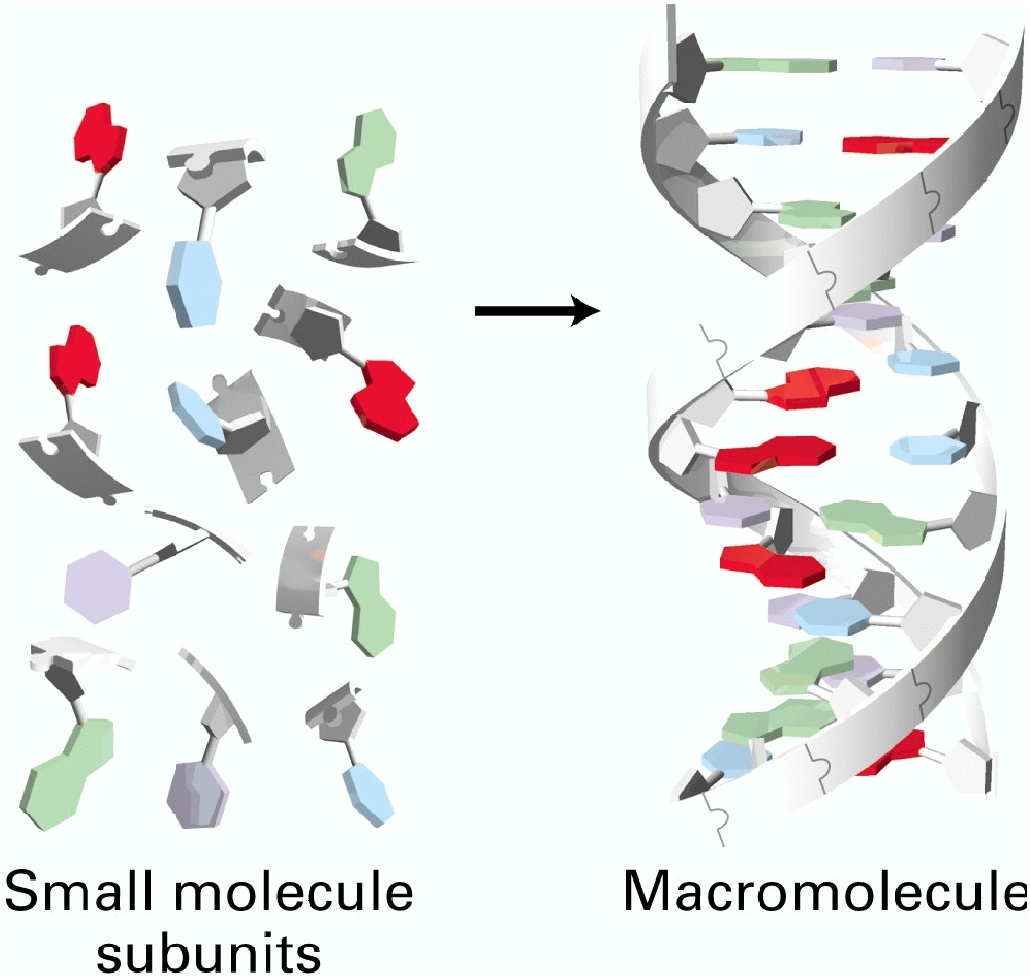
What This Course Will Do

1. Give you a feeling for main issues in molecular biological computing: sequence, structure and function.
2. Give you an opportunity to implement some widely used algorithms.
3. Give you exposure to classic computational problems, as manifested in biology
4. Give you exposure to classic biological problems, as represented computationally

BioInformatics Schematic of a Cell



Molecules of Life



Macromolecule (Polymer)	**Monomer**
DNA	Deoxyribonucleotides (dNTP)
RNA	Ribonucleotides (NTP)
Protein or Polypeptide	Amino Acid

Nucleic Acids: DNA and RNA

Form the genetic material of all organisms.

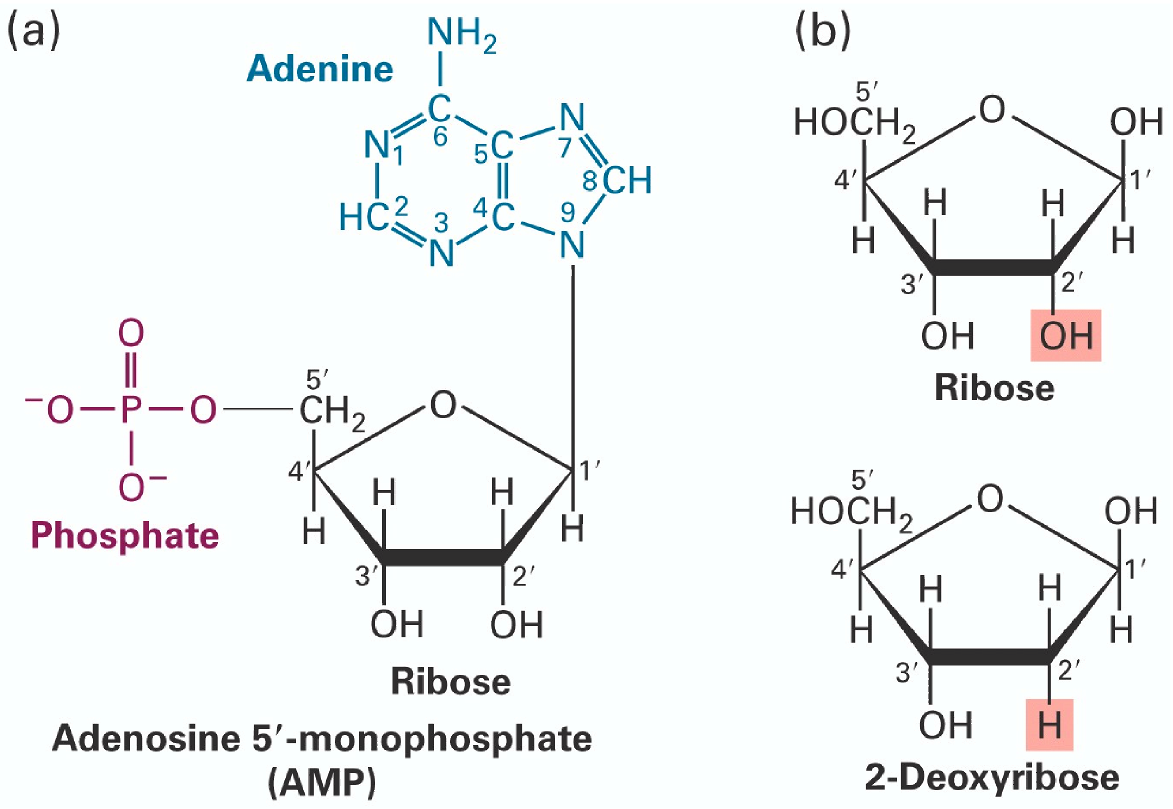
Found mainly in the *nucleus* of a cell (hence "nucleic")

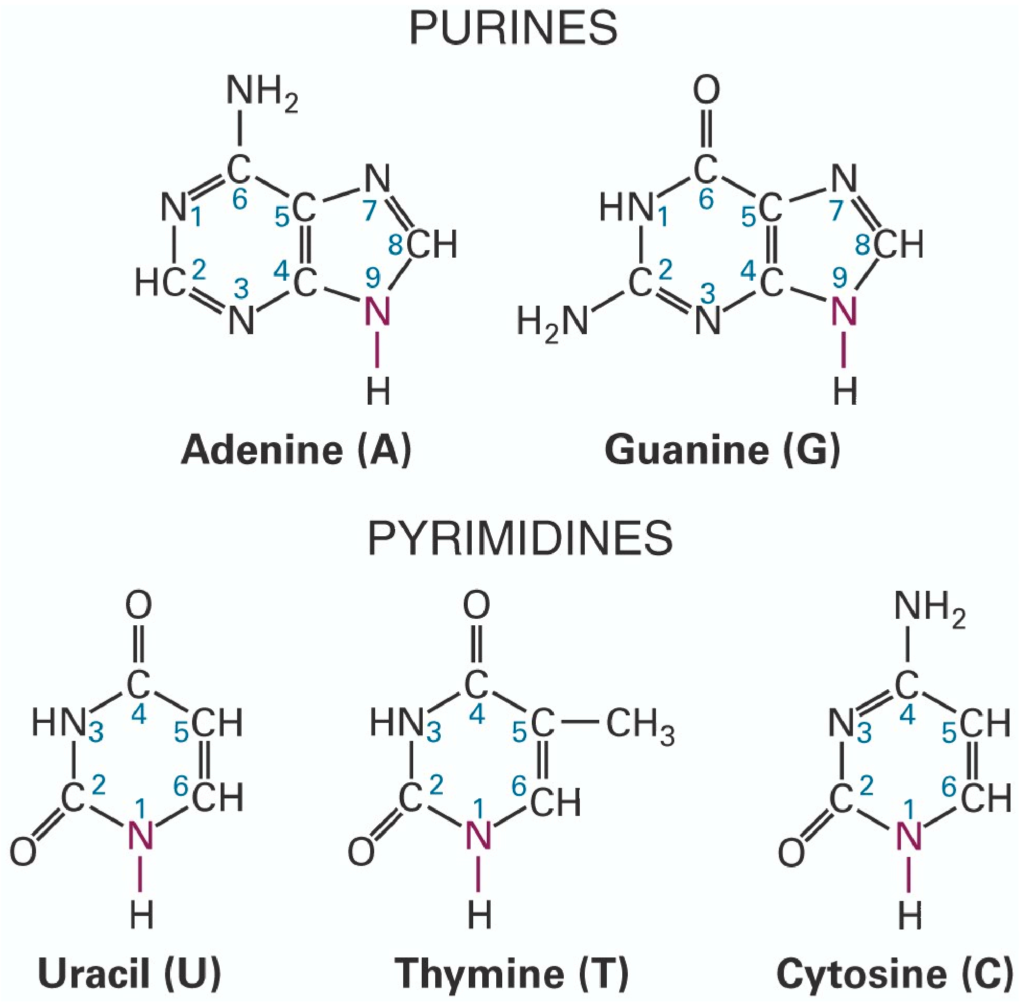
Contain phosphoric *acid* as a component (hence "acid")

They are made up of *nucleotides*. A nucleotide has 3 components

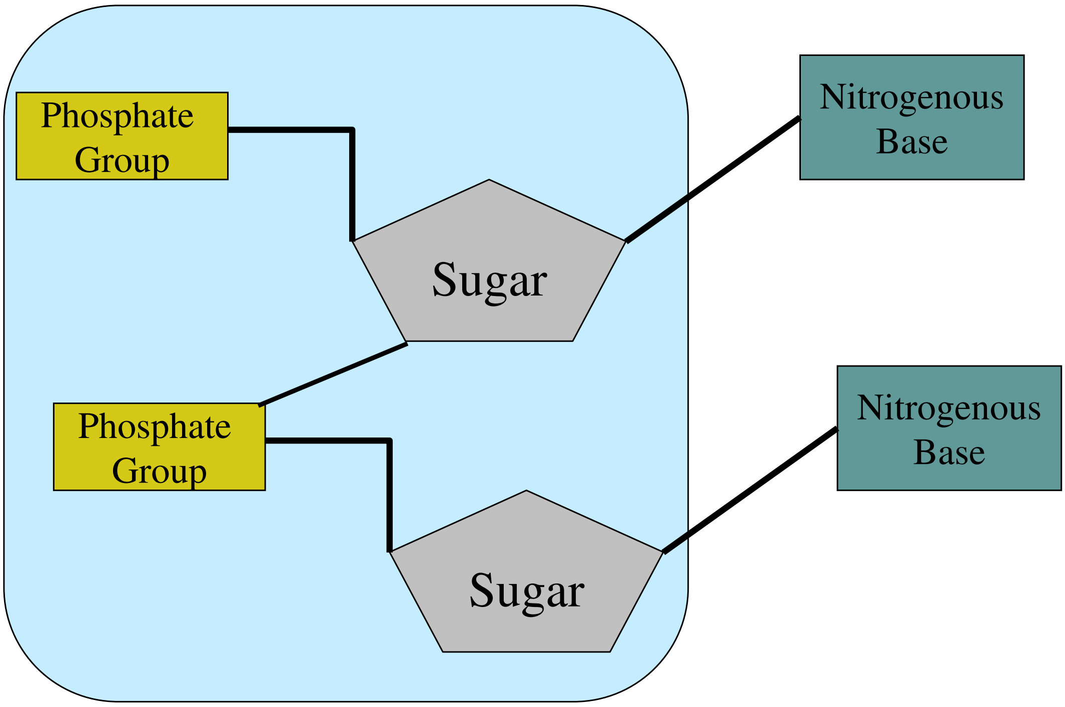
1. Sugar: ribose in RNA, deoxyribose in DNA
2. Phosphoric acid
3. Nitrogen base
   - Adenine (A)
   - Guanine (G)
   - Cytosine (C)
   - Thymine (T) or Uracil (U)

Nucleotide and Nitrogen Bases

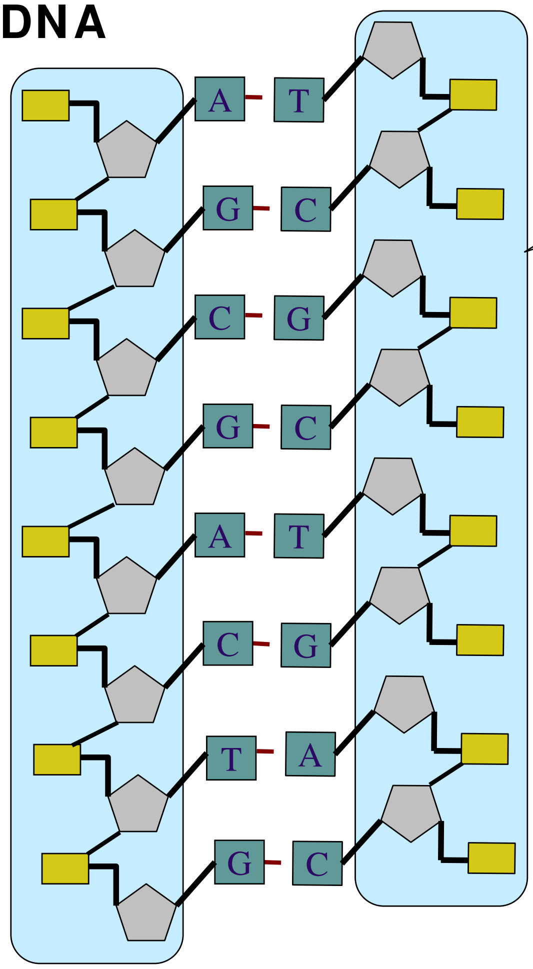



PURINES
NH2
N 1, 6 C, 5 C, 7 N, 8 CH, HC 2, 3 N, 4 C, 9 N
H
Adenine (A)
O
HN 1, 6 C, 5 C, 7 N, 8 CH, C 2, 3 N, 4 C, 9 N
H2N
H
Guanine (G)
PYRIMIDINES
O
HN 3, 4 C, 5 CH, 6 CH, C 2, 1 N
O
H
Uracil (U)
O
HN 3, 4 C, 5 C—CH3, 6 CH, C 2, 1 N
O
H
Thymine (T)
NH2
N 3, 4 C, 5 CH, 6 CH, C 2, 1 N
O
H
Cytosine (C)

DNA and RNA Strands

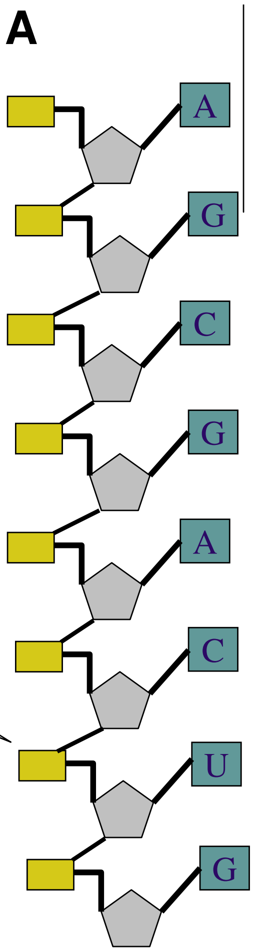


DNA



A = T
G = C

RNA



T → U

Protein

Protein molecule is a chain of amino acids

- There are 20 different amino acids with distinct physico-chemical properties

  1. Hydrophilic amino acids:

     THR, SER, PRO, TYR, ASN, GLN

  2. Charged amino acids:

     ASP, GLU, LYS, ARG, HIS

  3. Hydrophobic amino acids:

     VAL, LEU, ILE, PHE, ALA, GLY, TRP

  4. Sulfur containing amino acids:

     MET, CYS

- The interaction of these properties allows a chain to fold into a unique reproducible 3D shape

Protein Chain

Common repeating backbone (yellow)

Unique sidechains (blue)

Since the backbone is always there, we can specify the protein by specifying the sequence of amino acids

- Example:

Alanine–Tyrosine–Valine

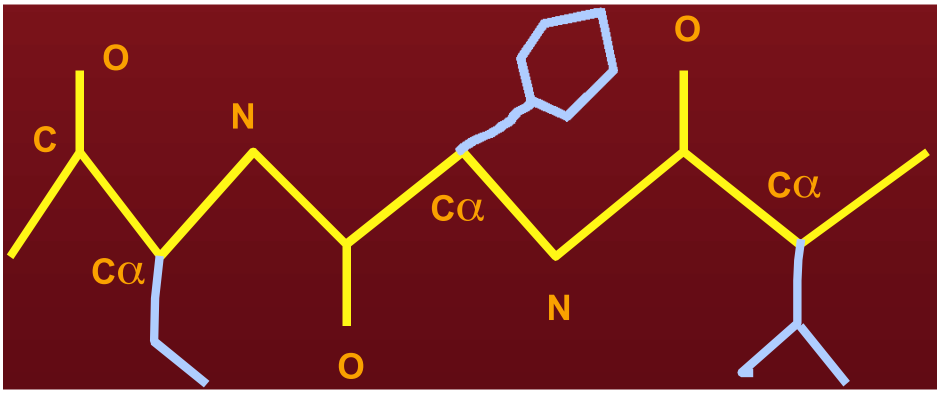
↓

ALA-TYR-VAL

↓

A-Y-V

↓



Protein Structure and Function

A linear sequence of amino acids folds to form a complex 3D structure

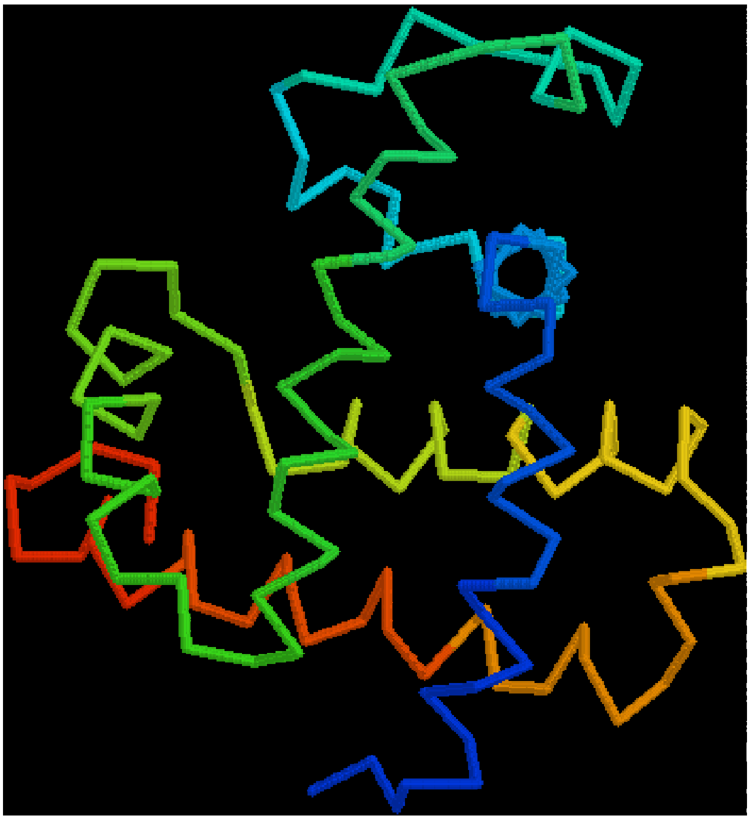
The structure of a protein is intimately connected to its function

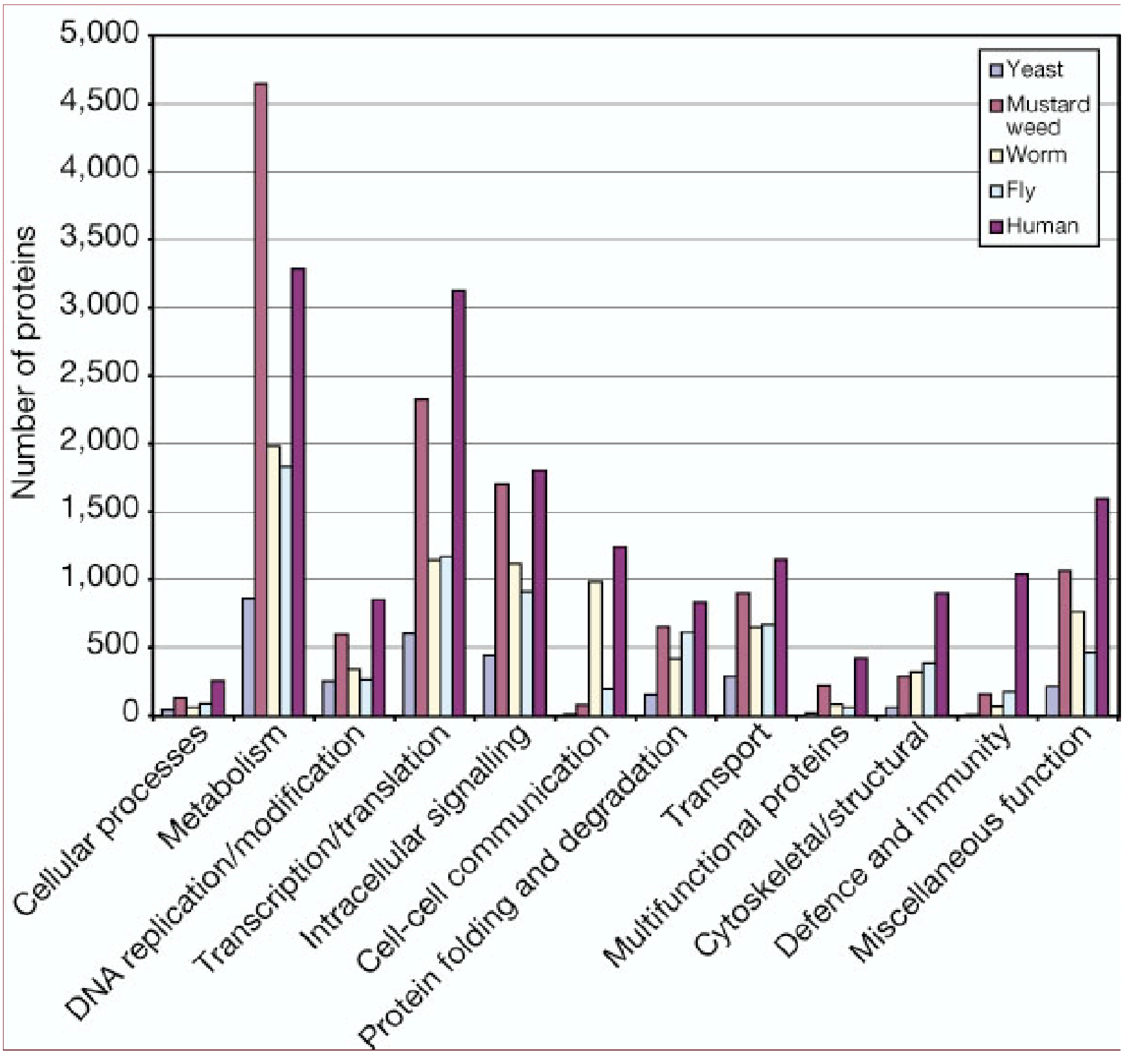
It is the 3D shape of proteins that gives them their working ability — generally speaking, the ability to bind with other molecules in very specific ways

Protein Functions

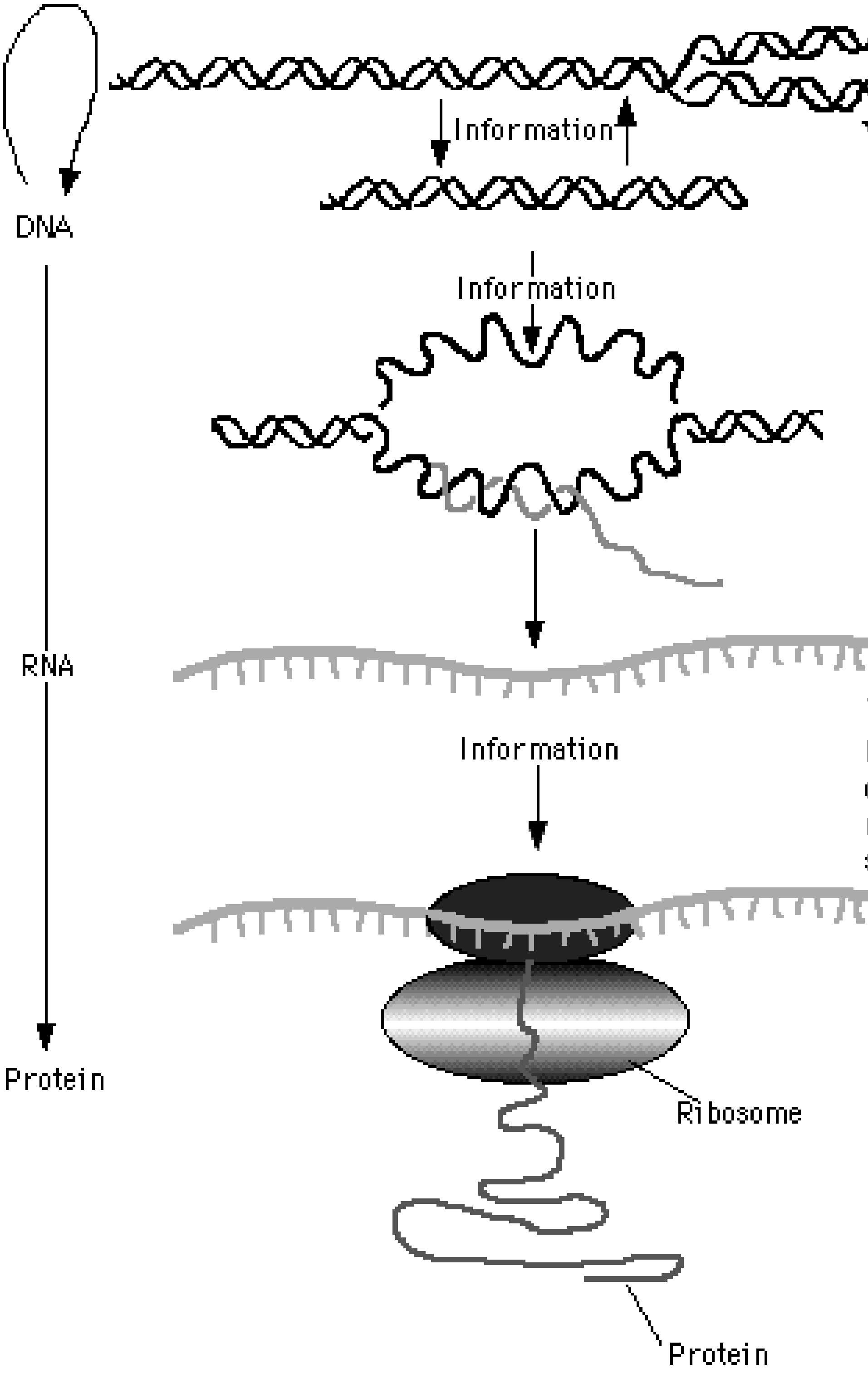
- Tissue building blocks
- Enzymes: catalyst of chemical reactions
- Transport or storage of molecules
- Antibody defense
- Regulate or control cell processes

Protein Structure and Function





Central Dogma of Molecular Biology



Replication

(DNA passes coded information with replication)

Transcription

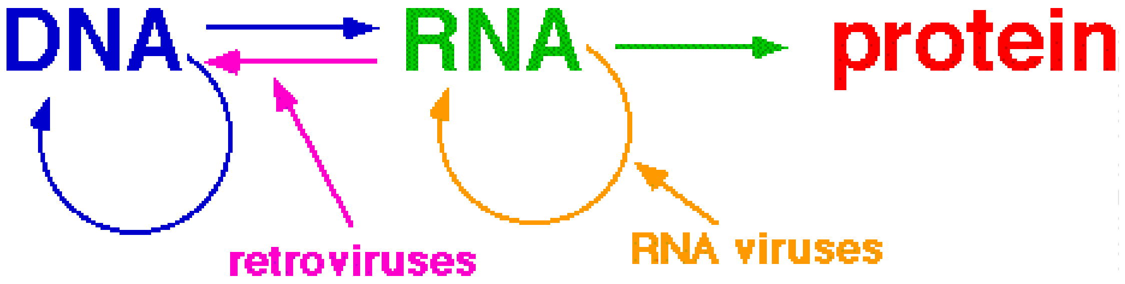
RNA synthesis: coded information passed into RNA during transcription

Translation

Messenger RNA carries coded information to ribosome during protein synthesis

The buck stops here. Proteins refuse to give away any information.

Proteins provide structure and help carry out almost all biological activity.



Central Dogma of Molecular Biology

Replication See figure below

Transcription A gene is copied into a complementary strand of a *messenger RNA* (mRNA)

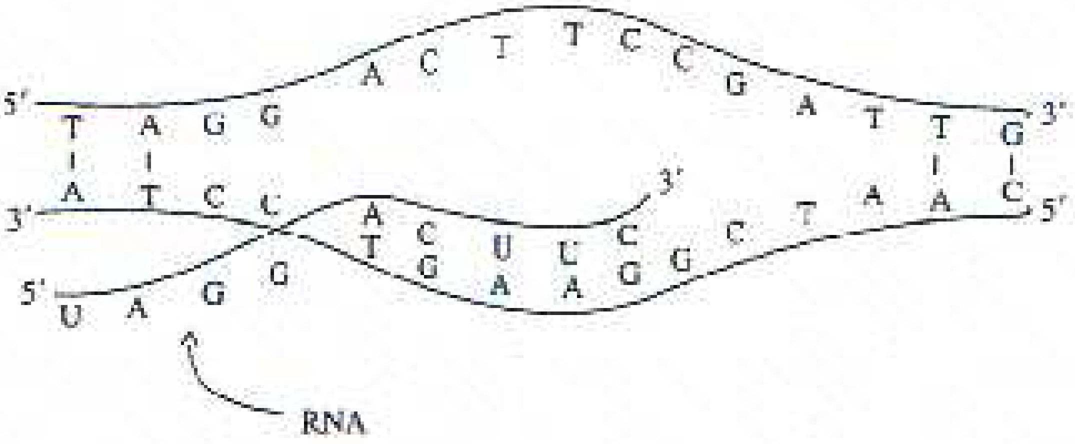
1. RNA polymerase splits open a double-stranded DNA
2. Finds transcription start site
3. Free RNA bases bind to complementary DNA bases of the template single-stranded DNA
4. Transcription stop site signals end of transcription

Translation mRNA sequence is *read* and *interpreted* to synthesize a protein

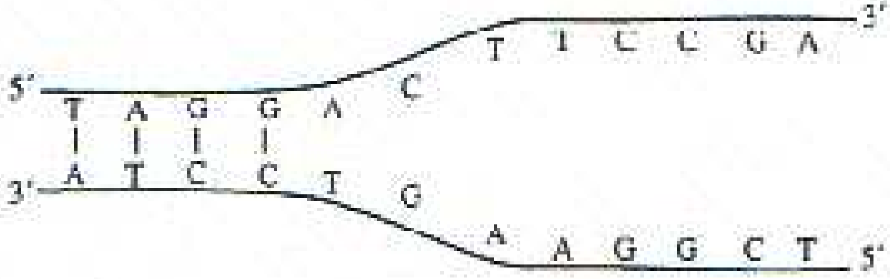
1. A codon (3 nucleotides) at a time enters the ribosome
2. Anticodon of free *transfer RNA* (tRNA) binds to the current codon
3. The amino acid attached to the tRNA is added to the chain of protein being made

Information is stored in DNA as *nucleotide sequences* and used in protein synthesis

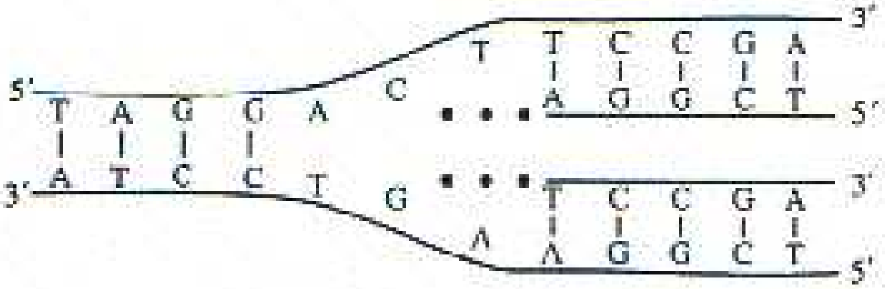
Replication



Making a new DNA from one already existing is called DNA *replication*. We begin with a double helix that has been separated into two single strands.

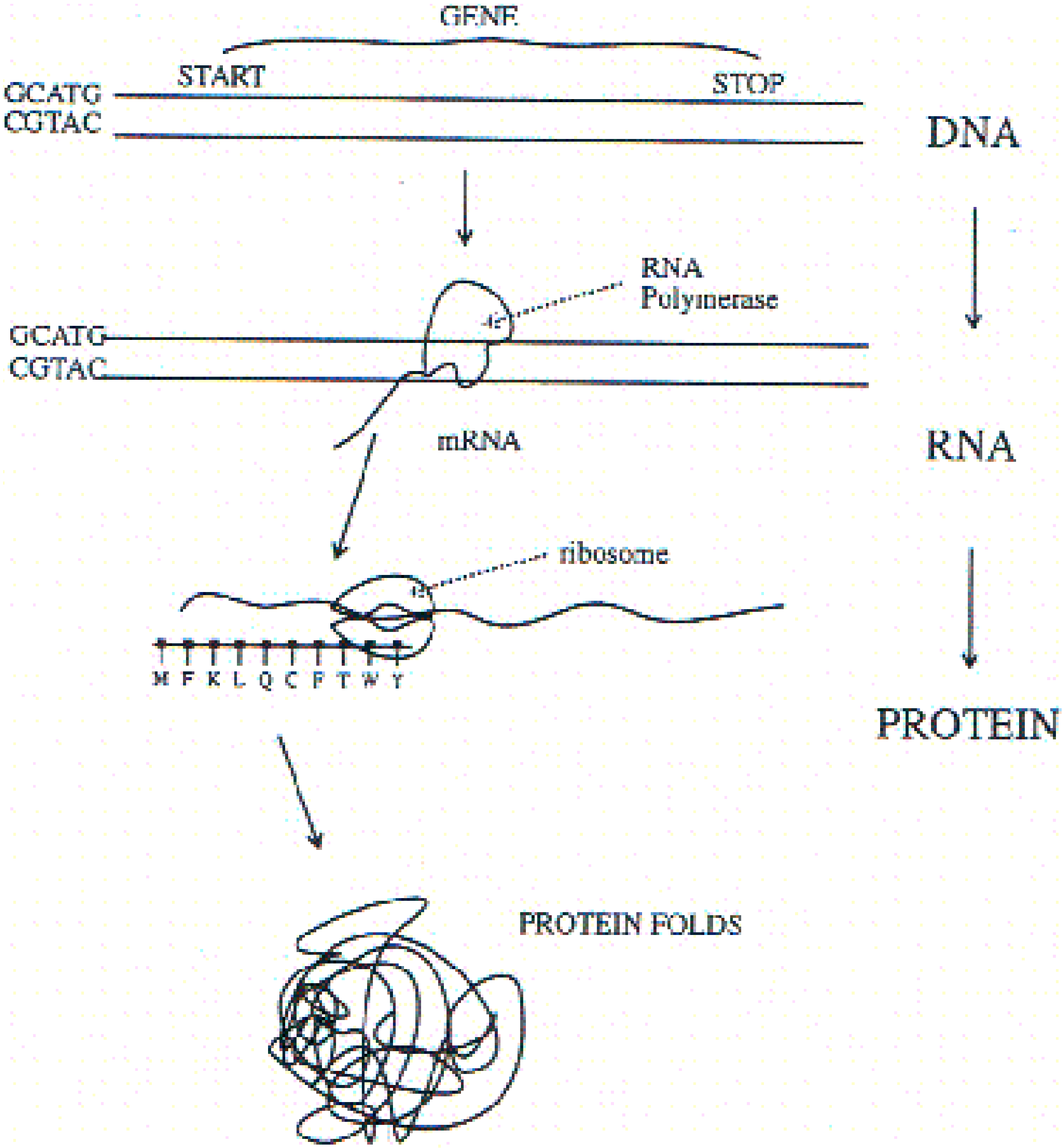


Then the single strands are used to template new double strands.



In this way two identical DNA molecules are made, each having one strand of the original molecule. Replication in this picture proceeds from right to left.

Transcription and Translation



More on Transcription

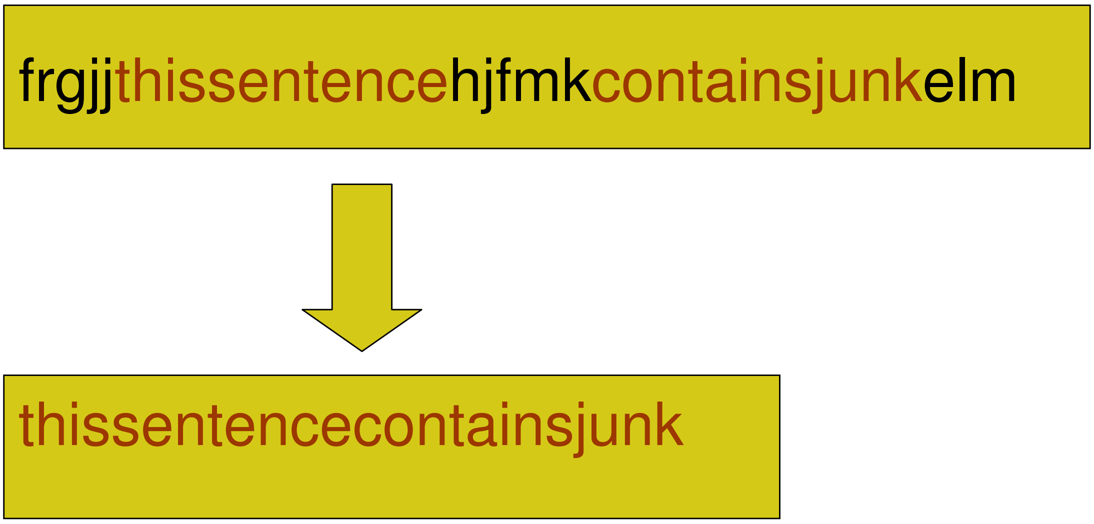
The RNA message is sometimes *edited*

Exons are nucleotide segments whose codons will be **ex**pressed

Introns are **in**tervening segments (genetic gibberish) that are snipped out

Transcription details

1. Gene is first transcribed into a *pre-mRNA*
2. Exons are spliced out of pre-mRNA together to form the final mRNA used in translation

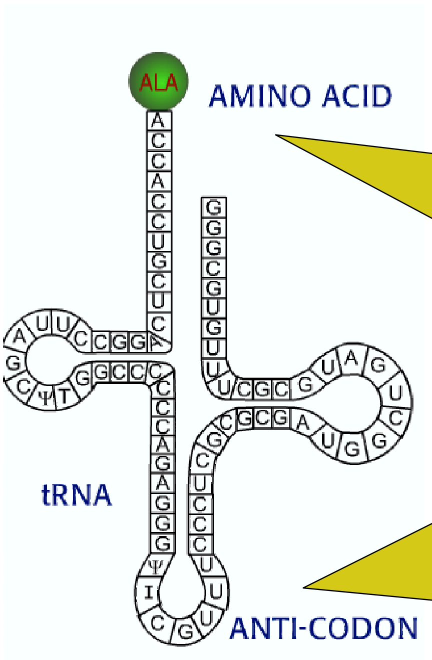


frgjjthissentencehjfmkcontainsjunkelm

thissentencecontainsjunk

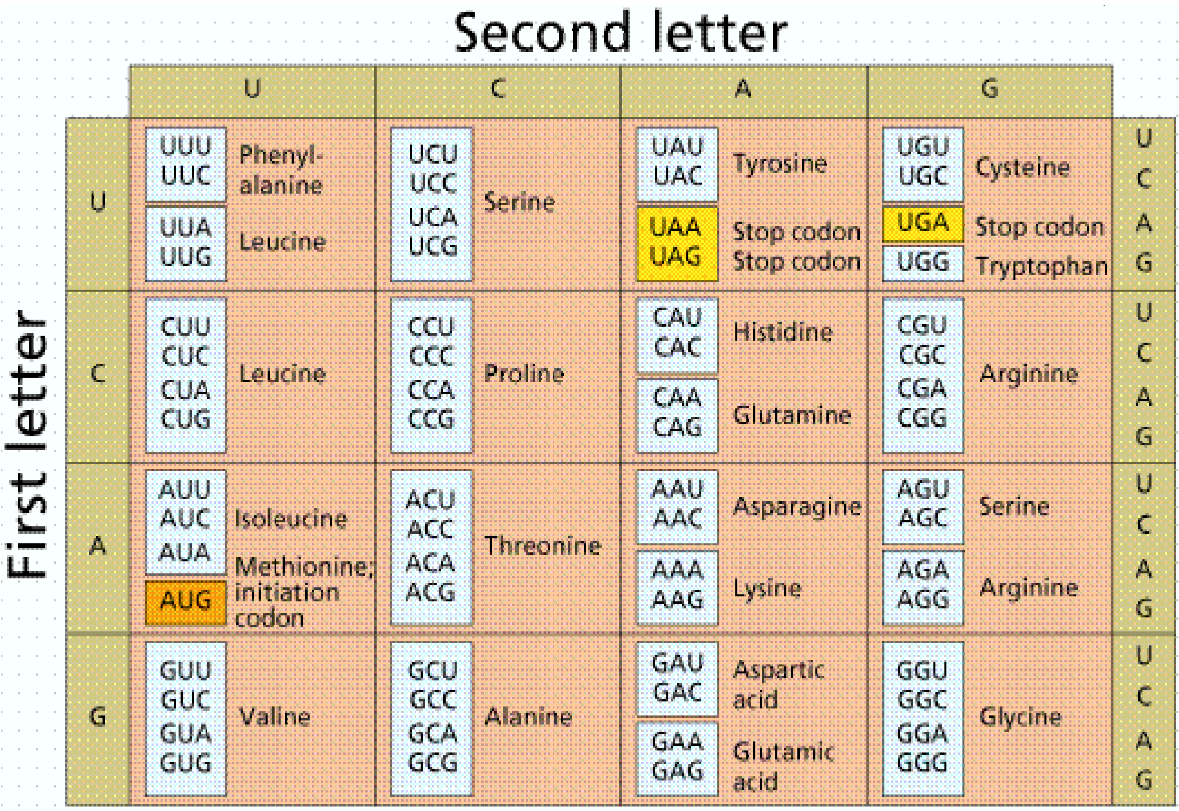
The Genetic Code

tRNA



One end of the tRNA links with a specific amino acid, which it finds floating free in the cytoplasm.

It employs its opposite end to form base pairs with nucleic acids – with a codon on the mRNA tape that is being read inside the ribosome.



First letter	Second letter U	Second letter C	Second letter A	Second letter G	
U	UUU UUC Phenyl-alanine UUA UUG Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine UAA Stop codon UAG Stop codon	UGU UGC Cysteine UGA Stop codon UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine CAA CAG Glutamine	CGU CGC CGA CGG Arginine	U C A G
A	AUU AUC AUA Isoleucine AUG Methionine; initiation codon	ACU ACC ACA ACG Threonine	AAU AAC Asparagine AAA AAG Lysine	AGU AGC Serine AGA AGG Arginine	U C A G
G	GUU GUC GUA GUG Valine	GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid GAA GAG Glutamic acid	GGU GGC GGA GGG Glycine	U C A G

The code is *universal* and redundant / *degenerate*

Gene and Genome

- Genome = The entire DNA sequence within the nucleus

  Human genome has 3,000,000,000 bases divided into 23 chromosomes

- Gene = segment of DNA that codes for a protein

  A gene has on average 1340 DNA bases, thus specifying a protein of about 447 amino acids

  Humans have about 30,000 genes = 40,000,000 DNA bases = 3% of total DNA in genome

- Human have another 2,960,000,000 bases for control information

  Control information = when, where, how long, . . .

- Junk DNA: ≈90% of genome is non-coding

- Genotype = Genetic sequences associated with an individual organism

- Phenotype = Observable non-sequence features of an individual organism

Example: Myoglobin

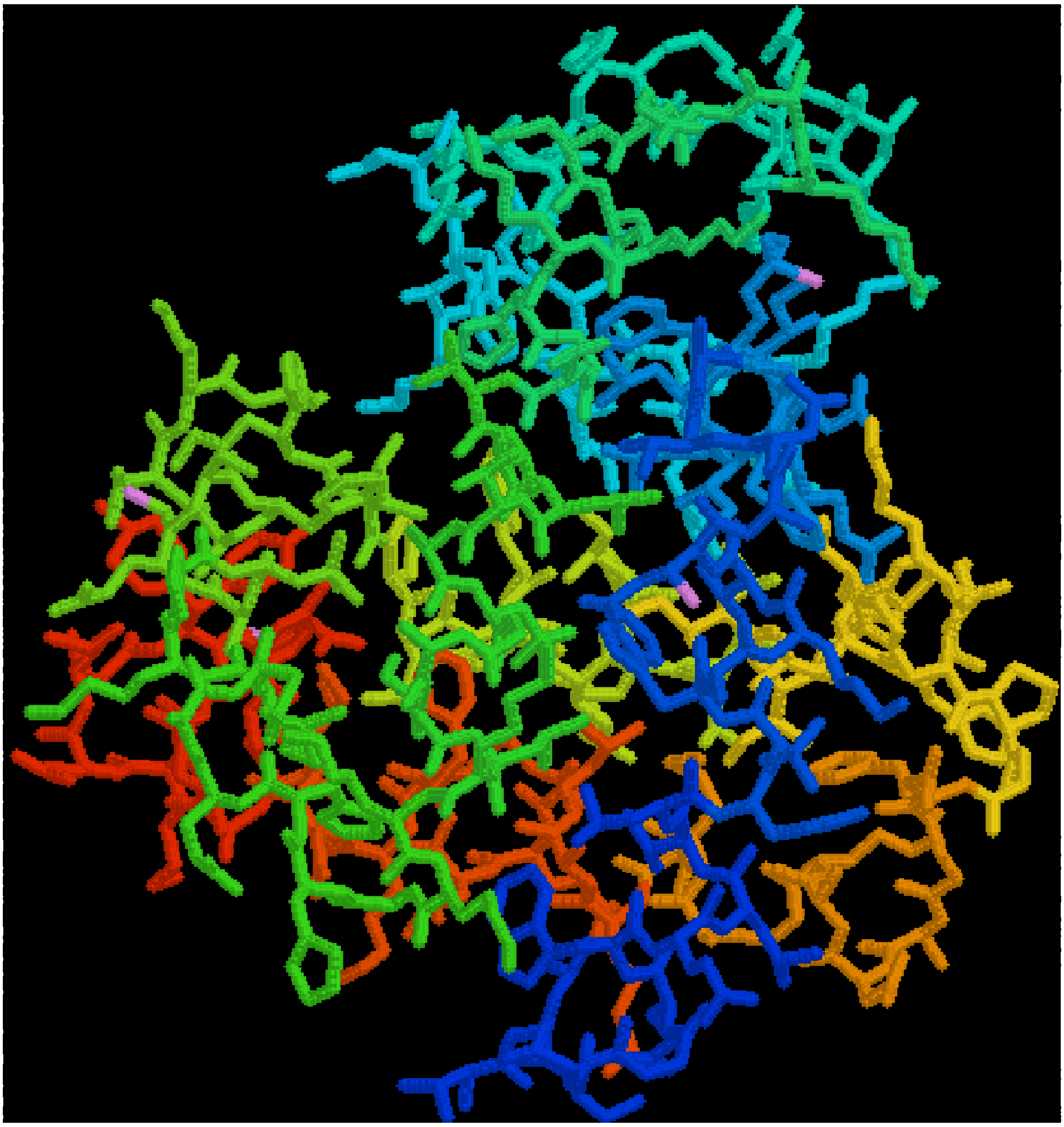
The GENE for the protein myoglobin:

```
ctgcagataa ctaactaaag gagaacaaca acaatggttc tgtctgaagg
tgaatggcag ctggttctgc atgtttgggc taaagttgaa gctgacgtcg
ctggtcatgg tcaggacatc ttgattcgac tgttcaaatc tcatccggaa
actctggaaa aattcgatcg tttcaaacat ctgaaaactg aagctgaaat
gaaagcttct gaagatctga aaaaacatgg tgttaccgtg ttaactgccc
taggtgctat ccttaagaaa aaagggcatc atgaagctga gctcaaaccg
cttgcgcaat cgcatgctac taaacataag atcccgatca aatacctgga
attcatctct gaagcgatca tccatgttct gcattctaga catccaggta
acttcggtgc tgacgctcag ggtgctatga acaaagctct cgagctgttc
cgtaaagata tcgctgctaa ctgggttacc agggttaatg aggtacc

BASE COUNT       155 a    108 c    115 g    129 t
```

The protein sequence for myoglobin:

```
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPG
NFGADAQGAMNKALELFRKDIAAKYKELGYQG
```



Genes

The behavior, development, state and structure of an organism are determined by its genes and their control — the genotype

To understand biology, we must understand how the gene products interact and produce observed features and traits — the phenotype

Many genes are shared among organisms, but each has made changes in their detailed sequence of DNA

Functions of genes

- Signal transduction
- Structural support
- Enzymatic catalysis
- Transport
- Movement
- Transcription control
- Trafficking

Computational Molecular Biology

In order to gather insight into the ways in which genes and proteins function, we DO

Sequence Analysis Analyze DNA and protein sequences, searching for clues about structure, function and control

Structure Analysis Analyze biological structures, searching for clues about sequence, function and control

Function Analysis Understand how sequence and structure lead to functions or phenotype

Goal of Human Genome Project:

1. Identify and characterize all proteins

2. Identify and characterize all interactions

3. Characterize all phenotypes

Challenges Understanding Genetic Information

Genetic information is redundant

- Double-stranded DNA
- Genetic code
- Base or amino acid replacements
- Mutation
- Alternative splicing
- Sequencing errors

Single genes have multiple functions

Single function/phenotype from multiple genes

Function depends on structure depend on sequence

Junk DNA

Huge and growing databases

Annotations problems

BioInformatics Depends on Data

DNA and RNA sequence data

Protein sequence data

Gene or protein microarray data

Three-dimensional structural data

Proteomics, genomics or metabolomics data

Literature and annotation data

Where is the Information?

GENEBANK:
http://www.ncbi.nih.gov

Swiss-prot:
http://us.expasy.org/sprot/relnotes

Protein Information Resource:
http://www.pir.org

Protein Data Bank (PDB):
http://www.rcsb.org/pdb/holdings.html

Stanford Microarray DB:
http://smd.stanford.edu

MedLine or PubMed

http://genome.ucsc.edu or
http://www.ebi.ac.uk/ensembl