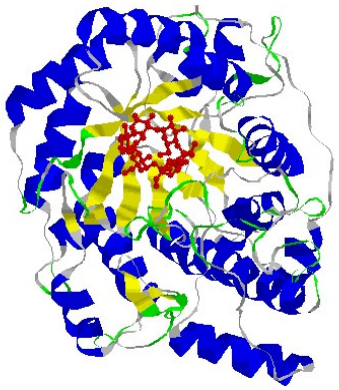
Introduction to Bioinformatics

Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
khuri@cs.sjsu.edu
www.cs.sjsu.edu/faculty/khuri

Biology Review



- DNA
- RNA
- Proteins
- Central Dogma
- Transcription
- Translation

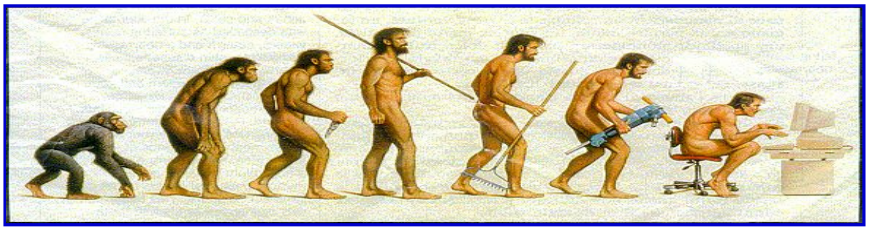
Last revised: July 18, 2010

Understanding Biology I

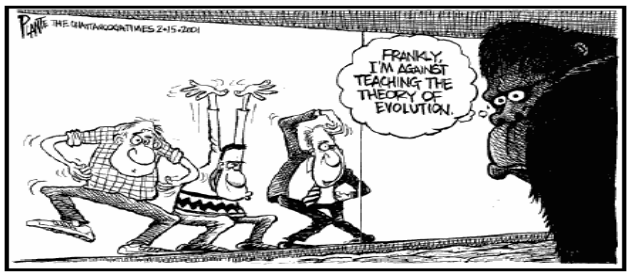
Nothing in biology makes sense, except in the light of evolution.

Dobzhansky, Russian geneticist (1900-75)

"I've only just bought this bronze stuff and you're telling me I ought to upgrade to iron?"



Understanding Biology II



- All organisms are (probably) **evolutionarily** related to each other; i.e., descended from a single common ancestor.
- **Living organisms** are "imperfect replication machines".
- Biology is not an exact science.

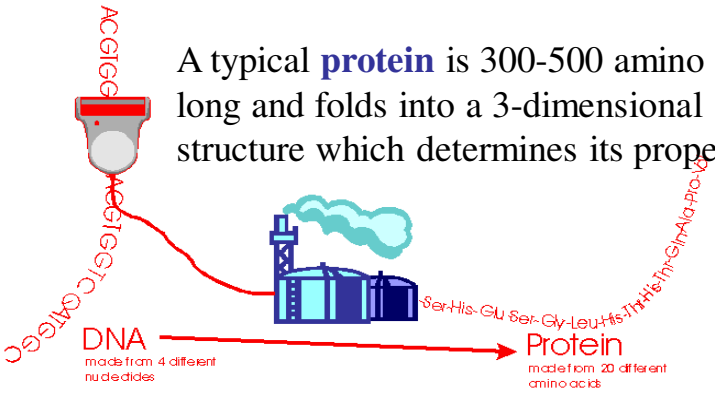
"We *are* our Proteins" Doolittle

Limitless Diversity From Combinatorial Assemblies of Limited Building Blocks

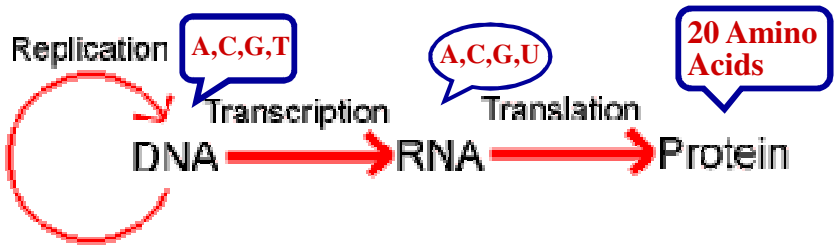
Protein Factory

Proteins: basis of how biology gets things done.

A typical **protein** is 300-500 amino acids long and folds into a 3-dimensional structure which determines its properties.

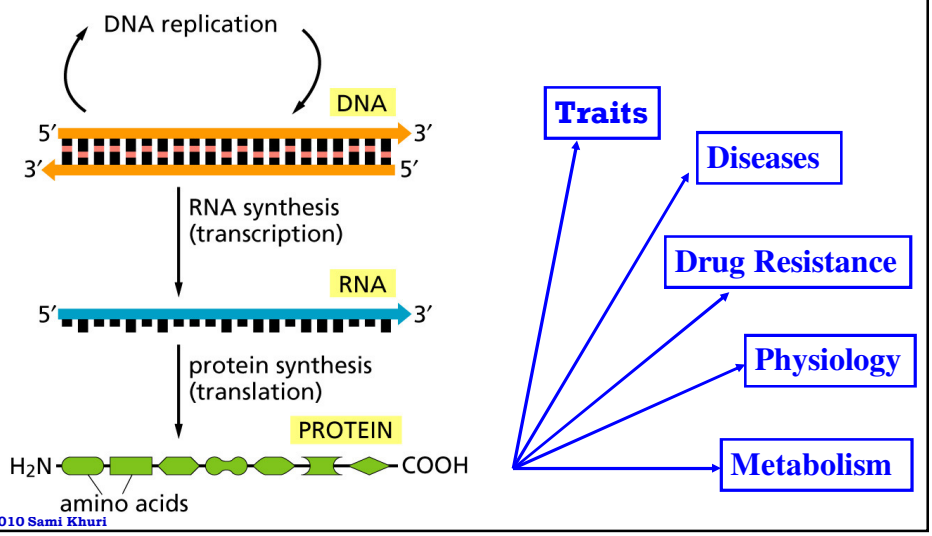


Central Dogma of Molecular Biology



Adenine	(A)	Adenine	(A)	{A-Y}-{BJOUX}
Guanine	(G)	Guanine	(G)	
Cytosine	(C)	Cytosine	(C)	
Thymine	(T)	Uracil	(U)	

Central Dogma of Molecular Biology



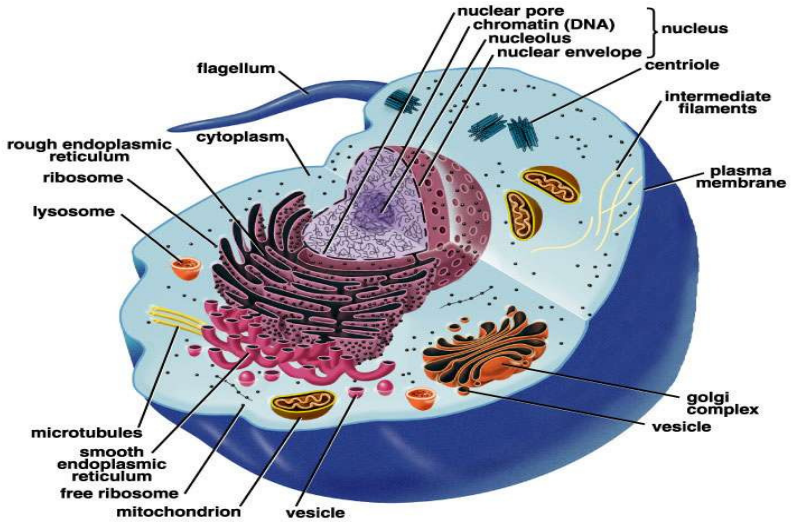
Prokaryotes and Eukaryotes

A **cell** is the fundamental working unit of every living organism.

There are two kinds of cells:

- **prokaryotes**, which are mostly single-celled organisms with no cell nucleus: archaea and bacteria.
- **eukaryotes**, which are higher level organisms, and their cells have nuclei: animals and plants.

Generalized Animal Cell



Proteins and Nucleic Acids

All living organisms have a similar molecular chemistry (biochemistry). The main actors in the chemistry of life are molecules called:

- **proteins**: which are responsible for what a living being is and does in a physical sense.

  "We are our proteins" R. Doolittle.
- **nucleic acids**: which encode the information necessary to produce proteins and are responsible for passing the "recipe" to subsequent generations.

DNA and RNA

- Living organisms contain two kinds of nucleic acids:
  - **Ribonucleic acid** (**RNA**)
  - **Deoxyribonucleic acid** (**DNA**)
- The **central dogma** states that information flows from **DNA** to **RNA** to **protein**.
- The function of a **protein** is determined by its unique three-dimensional structure.

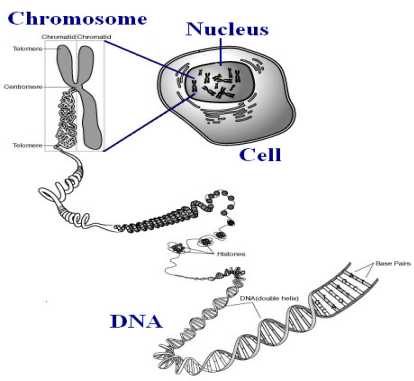
DNA and Chromosomes

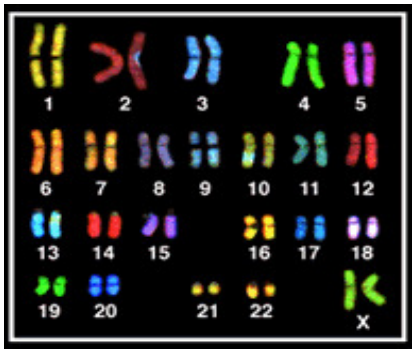
- The **human genome**: a complete set of instructions for making an organism, consists of tightly coiled threads of **DNA** and associated protein molecules, organized into structures called **chromosomes**.
- Besides the reproductive cell and red blood cell, every single cell in the human body contains the human genome.

Autosomal and Sex Chromosomes

- The **human genome** is distributed along 23 pairs of chromosomes
  - 22 autosomal pairs
  - the sex chromosome pair, XX for females and XY for males.
- In each pair, one chromosome is **paternally** inherited, the other **maternally** inherited.

Chromosomes and Genome





Number of chromosomes in a genome is characteristic of a species.
The human DNA contains about three billion base pairs (A-T or C-G).

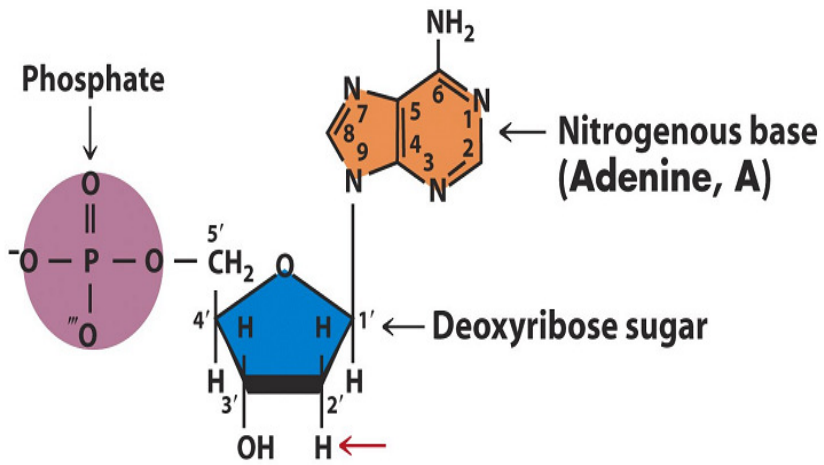
DNA Structure

- A **deoxyribonucleic acid** or **DNA** molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.
- Each nucleotide comprises
  - a phosphate group
  - a deoxyribose sugar
  - one of four nitrogen bases:

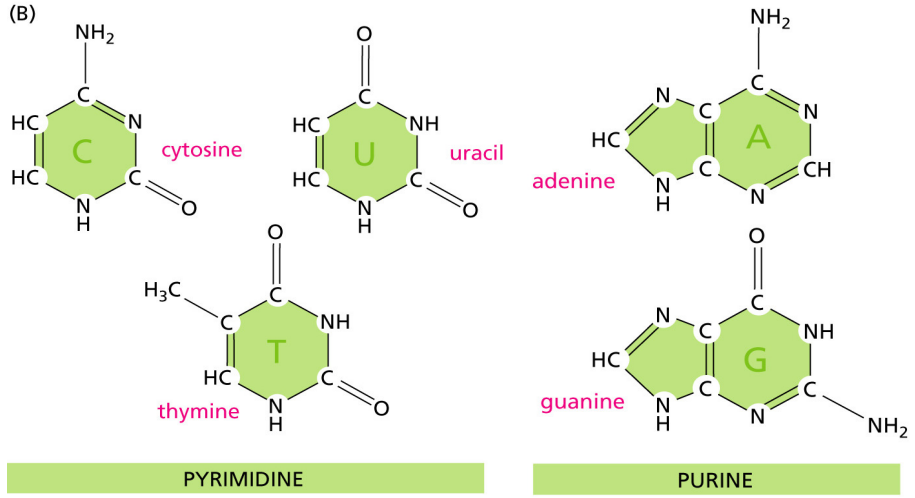
    purines: **adenine** (**A**) and **guanine** (**G**)

    pyrimidines: **cytosine** (**C**) and **thymine** (**T**).

A Nucleotide



Purines and Pyrimidines



Double Helix

- The binding of two nucleotides forms a base pair.
- The double helix is formed by connecting complementary nucleotides A-T and C-G on two strands with hydrogen bonds.
- Knowledge of the sequence on one strand allows us to infer the sequence of the other strand.
- The bases are arranged along the sugar phosphate backbone in a particular order, known as the DNA sequence, encoding all genetic instructions for an organism.

DNA Phosphodiester Backbone

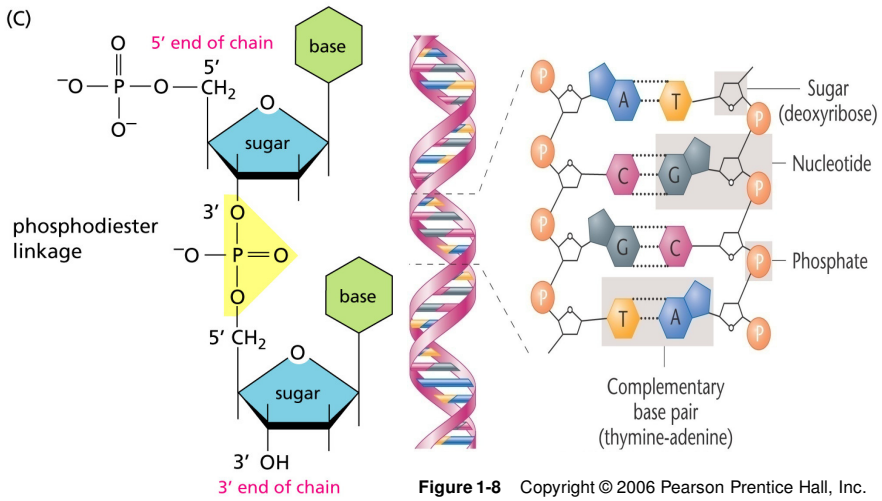
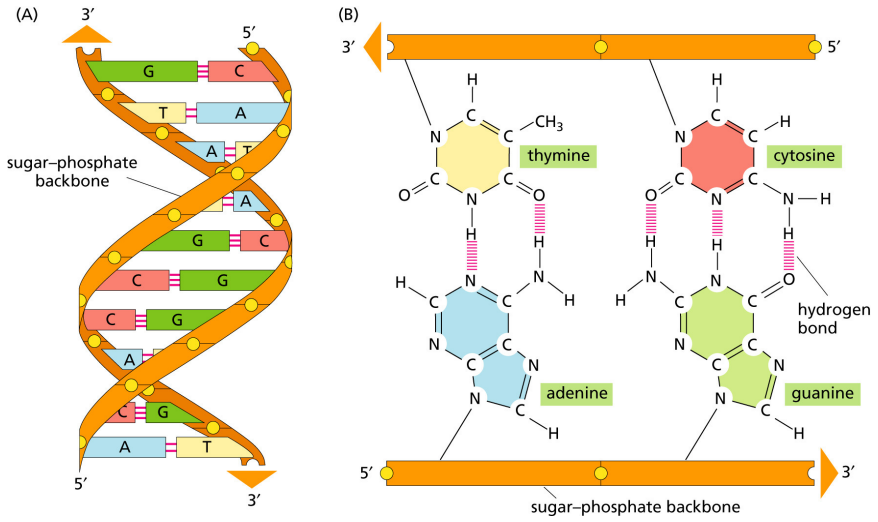


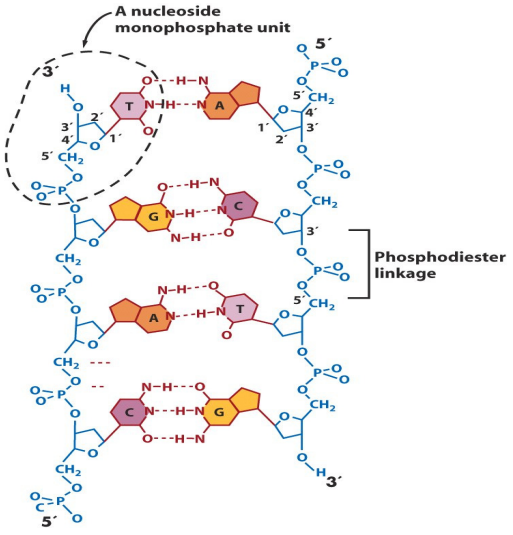
Figure 1-8 Copyright © 2006 Pearson Prentice Hall, Inc.

The Double Helical Structure of DNA



The Two Backbones of DNA

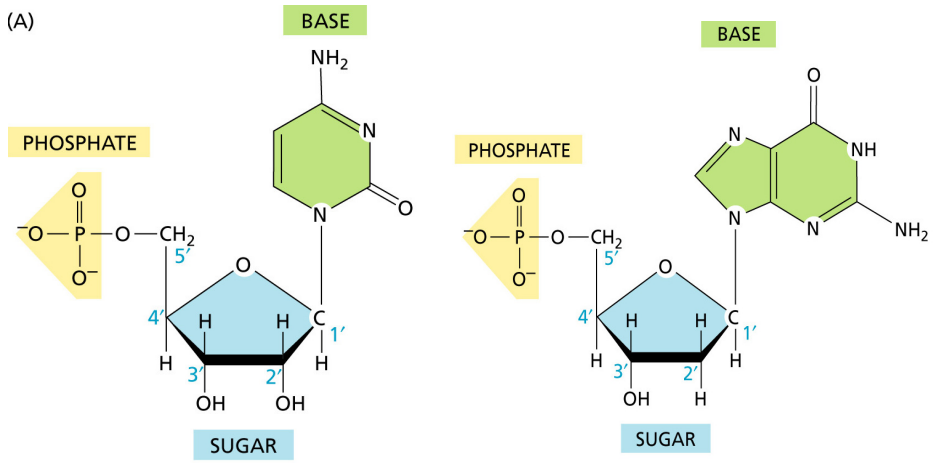
The DNA backbones have alternating sugar-phosphate components.
The backbones run in opposite directions.



DNA and Chromosomes

- The **genome** is a complete set of instructions for making an organism, consists of tightly coiled threads of **DNA** organized into structures called **chromosomes**.
- Besides the reproductive cell and red blood cell, every single cell in the human body contains the human genome.

Sugars Present in Nucleic Acids



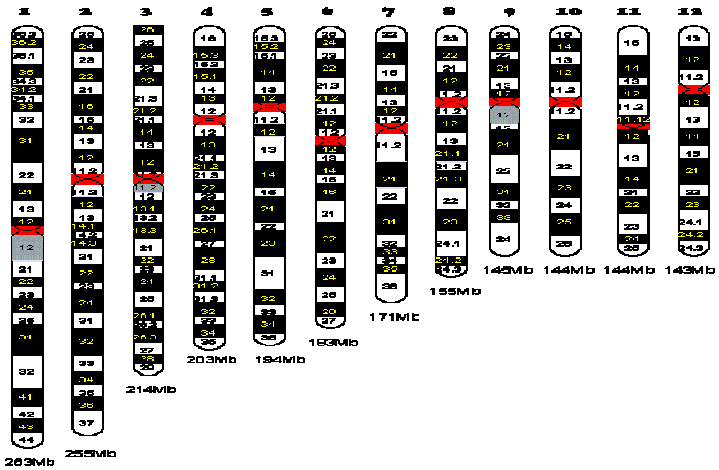
Pentose sugar present in **RNA** Pentose sugar present in **DNA**

Pairs of Chromosomes in Species

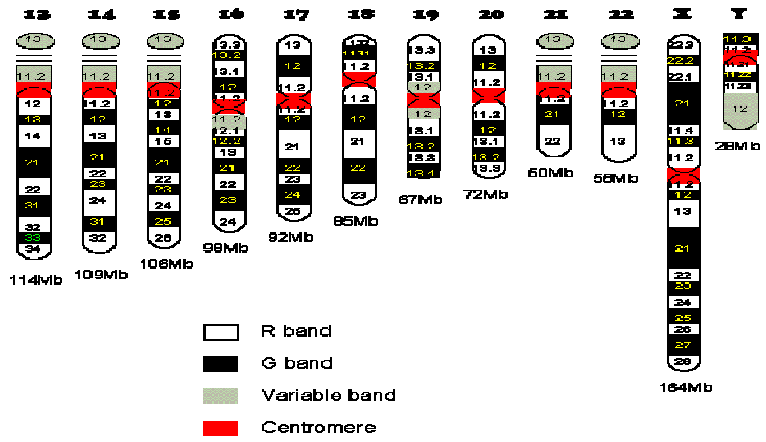
Table 3-2 Numbers of Pairs of Chromosomes in Different Species of Plants and Animals

Common name	Scientific name	Number of chromosome pairs	Common name	Scientific name	Number of chromosome pairs
Mosquito	*Culex pipiens*	3	Wheat	*Triticum aestivum*	21
Housefly	*Musca domestica*	6	Human	*Homo sapiens*	23
Garden onion	*Allium cepa*	8	Potato	*Solanum tuberosum*	24
Toad	*Bufo americanus*	11	Cattle	*Bos taurus*	30
Rice	*Oryza sativa*	12	Donkey	*Equus asinus*	31
Frog	*Rana pipiens*	13	Horse	*Equus caballus*	32
Alligator	*Alligator mississipiensis*	16	Dog	*Canis familiaris*	39
Cat	*Felis domesticus*	19	Chicken	*Gallus domesticus*	39
House mouse	*Mus musculus*	20	Carp	*Cyprinus carpio*	52
Rhesus monkey	*Macaca mulatta*	21			

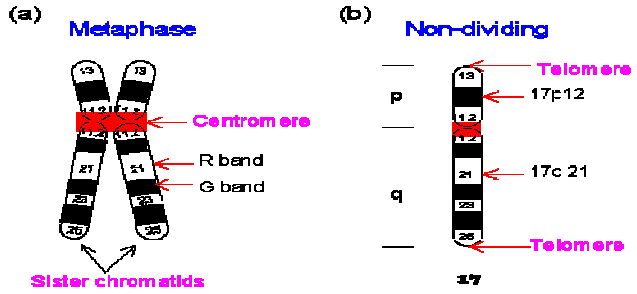
Banding Pattern of Human Chromosomes 1 to 12



Chromosomes 13 to 22
Chromosomes X and Y



Labeling a Chromosome

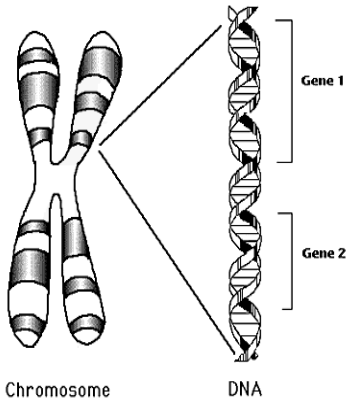


b) Long arm is labeled q for "queue"
Short arm is labeled p for "petite".

Lowest resolution: a few major bands are visible: q1, q2, q3: p1, ..

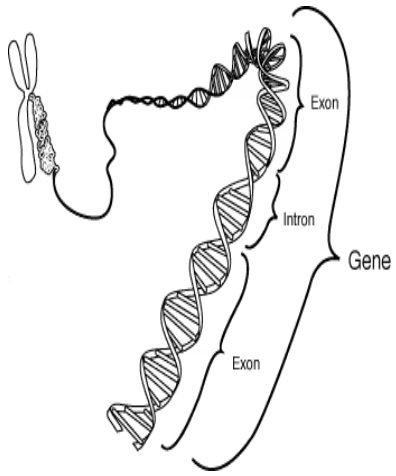
Higher resolutions show sub-bands: q11, q12 .. and even q11.1 …

Genes



- A **gene** is a specific sequence of nucleotide bases along a chromosome carrying information for constructing a protein.
- **Genes** are part of the chromosomes.
- The distance between **genes** is often much larger than the genes themselves.

Exons and Introns



www.accessexcellence.org/AB/GG/gene.html

In eukaryotes, genes consist of:

- exons
  protein-coding regions
- introns
  noncoding regions.

Approximately 5-10% of the gene is made up of exons while the rest are introns.

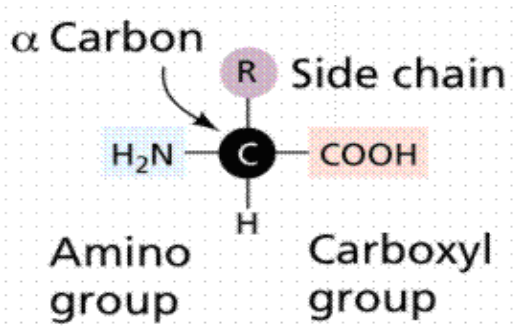
Ribonucleic Acid - RNA

- **RNA** is found in the cell and can also carry genetic information.
- While DNA is located primarily in the nucleus, **RNA** can also be found in the cytoplasm: the cellular liquid outside the nucleus.
- **RNA** is built from the nucleotides cytosine, guanine, adenine and uracil (U) (instead of thymine).
- **RNA** has its sugar phosphate backbone containing ribose.
- **RNA** forms a single strand.

Proteins

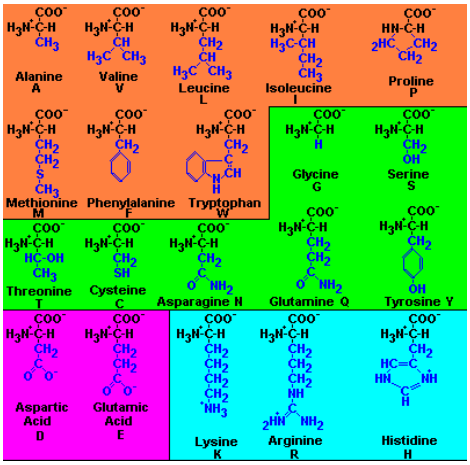
- 20 different **amino acids** are used to synthesize **proteins**.
- The shape and other properties of each **protein** is dictated by the precise sequence of **amino acids** in it.
- The function of a **protein** is determined by its unique three-dimensional structure.

Structure of the Amino Acid



It is the structure of the R group that determines which of the 20 amino acids it is and its special properties.

The Twenty Amino Acids



Orange:
nonpolar and hydrophobic.

The other amino acids are:
polar and hydrophilic - "water loving".

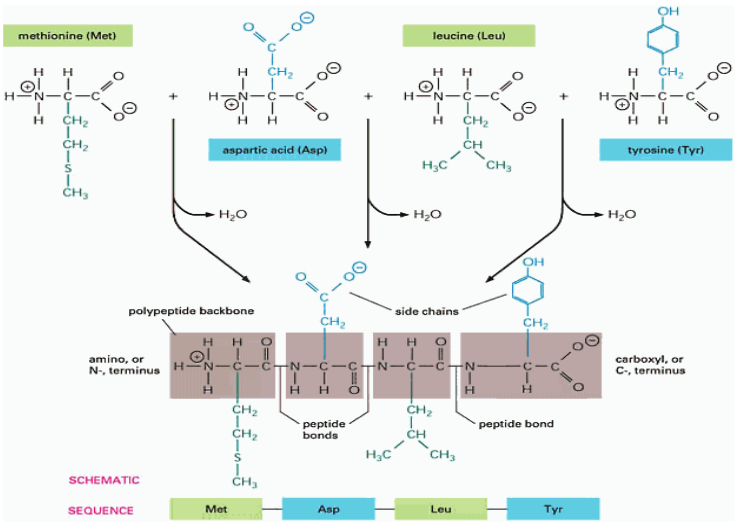
Magenta:
acidic - "carboxy" group in the side chain.

Light blue:
basic - "amine" group in the side chain.

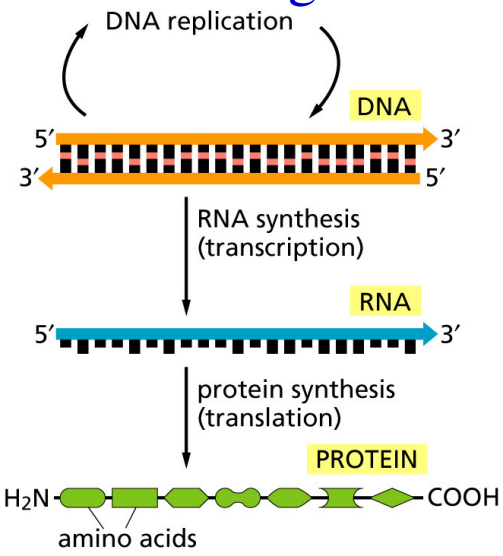
	One-letter code	Three-letter-code	Name
1	A	Ala	Alanine
2	C	Cys	Cysteine
3	D	Asp	Aspartic Acid
4	E	Glu	Glutamic Acid
5	F	Phe	Phenylalanine
6	G	Gly	Glycine
7	H	His	Histidine
8	I	Ile	Isoleucine
9	K	Lys	Lysine
10	L	Leu	Leucine
11	M	Met	Methionine
12	N	Asn	Asparagine
13	P	Pro	Proline
14	Q	Gln	Glutamine
15	R	Arg	Arginine
16	S	Ser	Serine
17	T	Thr	Threonine
18	V	Val	Valine
19	W	Trp	Tryptophan
20	Y	Tyr	Tyrosine

David Gilbert

Protein Structure

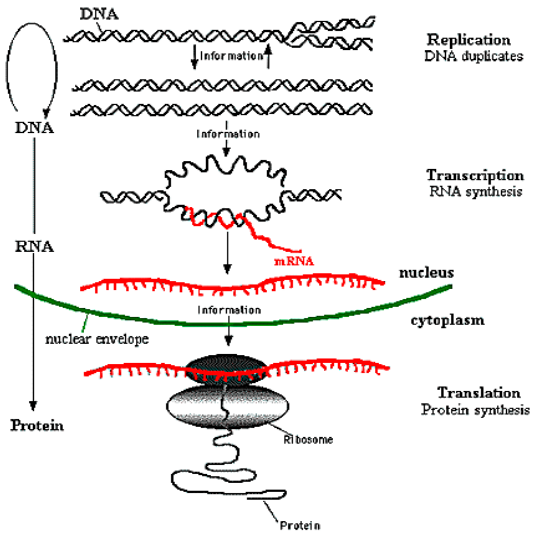


Central Dogma of Molecular Biology



According to the **central dogma of molecular biology**, there is a single direction of flow of genetic information from the **DNA**, which acts as the information store, through **RNA** molecules from which the information is translated into **proteins**.

Steps of the Central Dogma



Replication
DNA duplicates

Genetic information embodied in DNA is **replicated** into more DNA

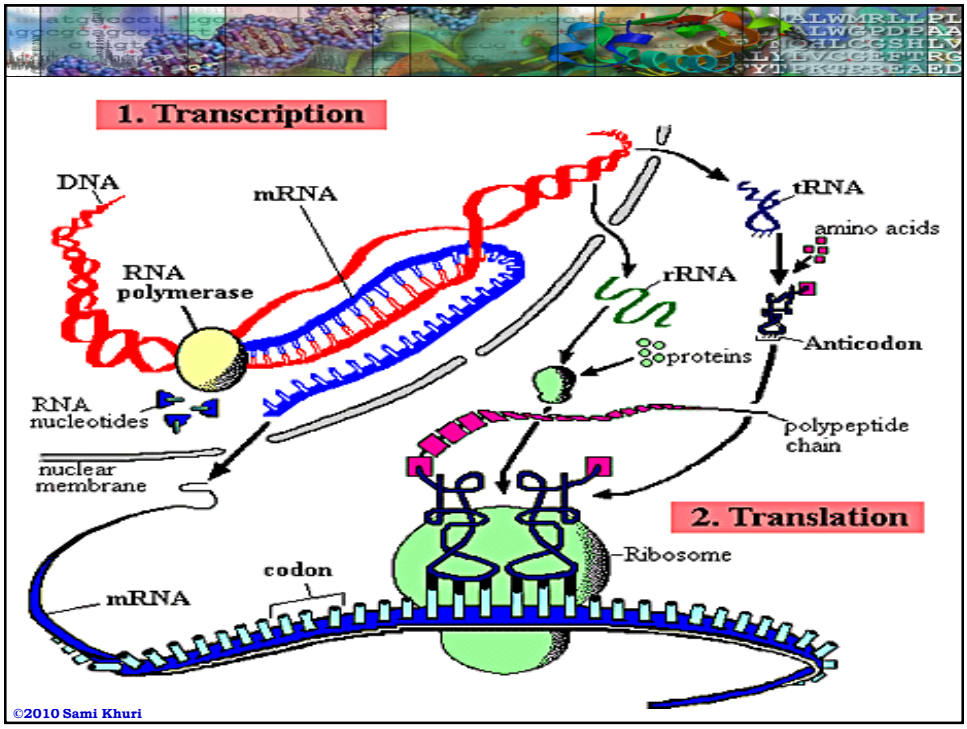
Transcription
RNA synthesis

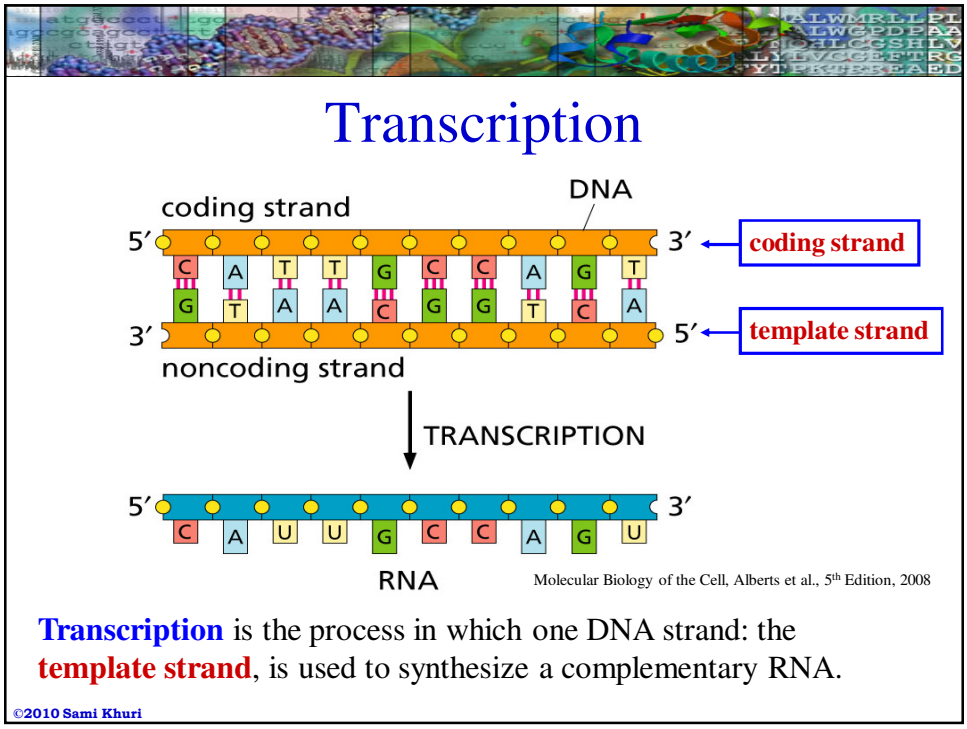
The synthesis of an RNA from a sequence of DNA.
The resulting RNA is **mRNA**.

In eukaryotic cells, the mRNA is **spliced** and it migrates from the nucleus to the cytoplasm.

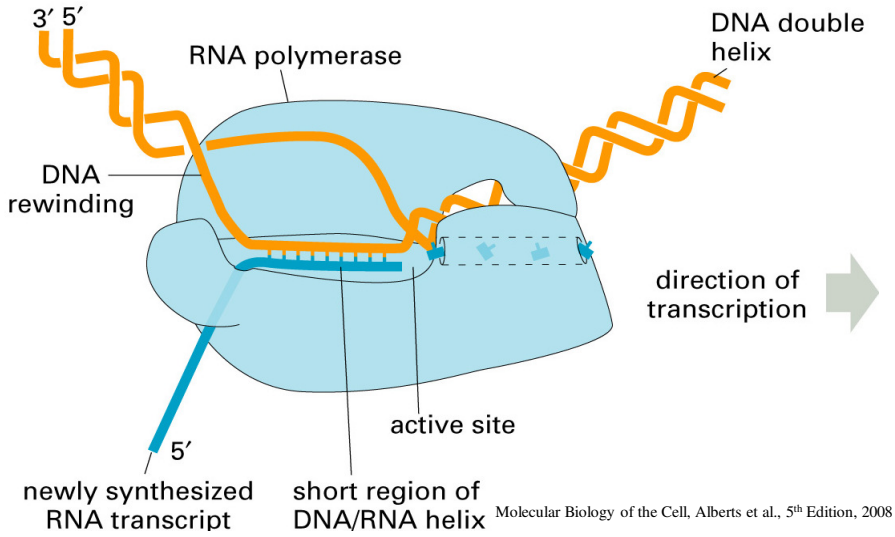
Translation
Protein synthesis

Messenger RNA carries coded information to ribosomes that "read"and use it for **protein synthesis**.


1. Transcription
DNA
mRNA
RNA polymerase
RNA nucleotides
nuclear membrane
tRNA
amino acids
rRNA
proteins
Anticodon
polypeptide chain
2. Translation
Ribosome
codon
mRNA
©2010 Sami Khuri


Transcription
DNA
coding strand
5′
3′
C A T T G C C A G T
G T A A C G G T C A
3′
5′
noncoding strand
coding strand
template strand
TRANSCRIPTION
5′
3′
C A U U G C C A G U
RNA
Molecular Biology of the Cell, Alberts et al., 5th Edition, 2008
Transcription is the process in which one DNA strand: the template strand, is used to synthesize a complementary RNA.
©2010 Sami Khuri

Synthesizing RNA from 5' to 3'



Molecular Biology of the Cell, Alberts et al., 5th Edition, 2008

The Genetic Code

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

Transfer RNA and Translation

- The translation from nucleotides to amino acid is done by means of **transfer RNA** (**tRNA**) molecules, each specific for one amino acid and for a particular **triplet** of nucleotides in mRNA called a **codon**.
- The family of tRNA molecules enables the codons in a mRNA molecule to be **translated** into the sequence of amino acids in the protein.

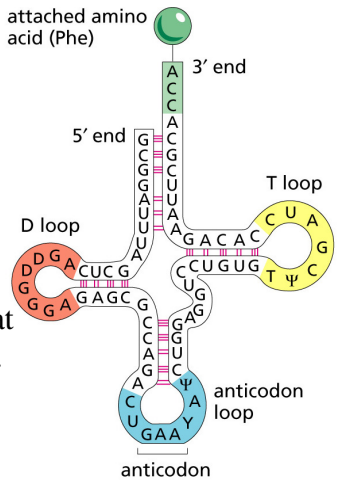
Codons and Anticodons

At least one kind of tRNA is present for each of the 20 amino acids used in protein synthesis.

Each kind of tRNA has a sequence of 3 unpaired nucleotides - the anticodon - which can bind to the complementary triplet of nucleotides - the codon - in an mRNA molecule.

The reading of codons in mRNA requires that the anticodons bind in the opposite direction.

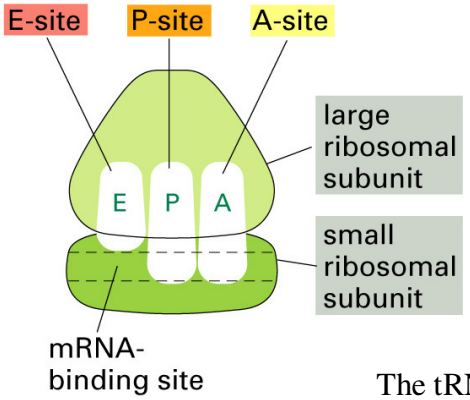
Anticodon: 3' AAG 5'
Codon: 5' UUC 3'



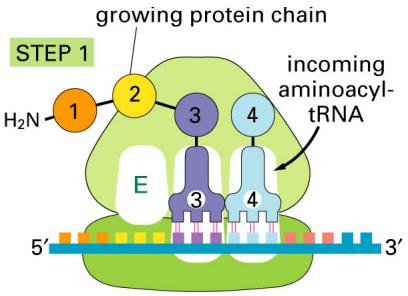
Start and Stop Codons

- The codon AUG serves two related functions
  - It begins most messages; that is, it signals the start of translation placing the amino acid methionine at the amino terminal of the polypeptide to be synthesized.
  - When it occurs within the message, it guides the incorporation of methionine.
- Three **codons**, UAA, UAG, and UGA, act as signals to terminate translation. They are called **STOP codons**.

Translation



Binding site of ribosome for the mRNA and the three tRNA binding sites.



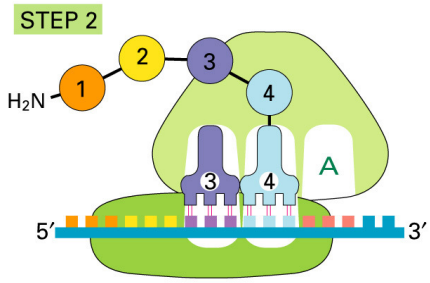
The tRNA molecules bind to the ribosome and are the physical link between the mRNA and the growing protein chain.

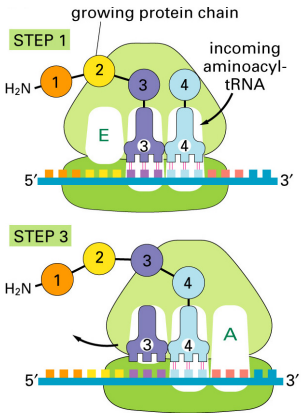
Steps of Translation: Initiation

- The small subunit of the ribosome binds to a site "upstream" of the start of the message.
- It proceeds downstream until it encounters the start codon AUG.
- It is then joined by the large subunit and a special initiator tRNA. The initiator tRNA binds to the P site on the ribosome.
- In eukaryotes, initiator tRNA generally carries methionine (Met).

Steps of Translation: Elongation

An aminoacyl-tRNA able to base pair with the next codon on the mRNA arrives at the A site.



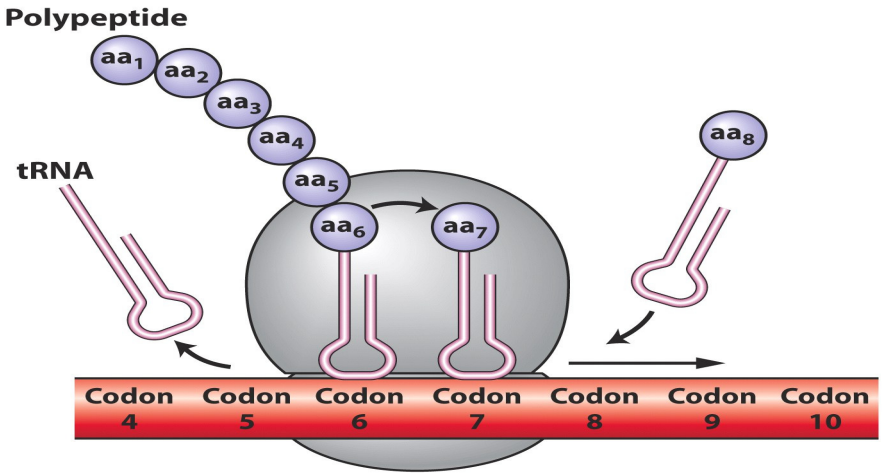


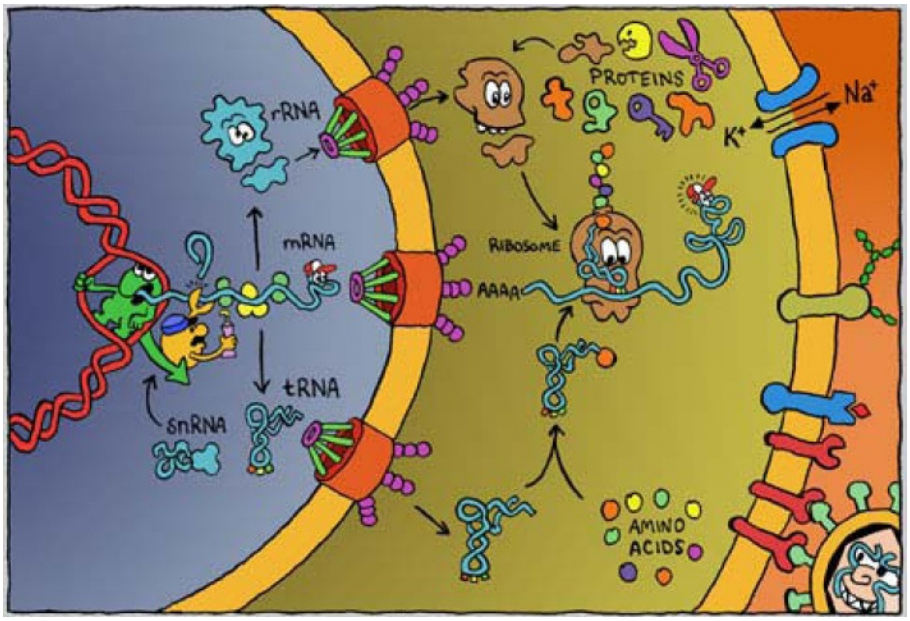
The preceding amino acid is linked to the incoming amino acid with a peptide bond.

Steps of Translation: Termination

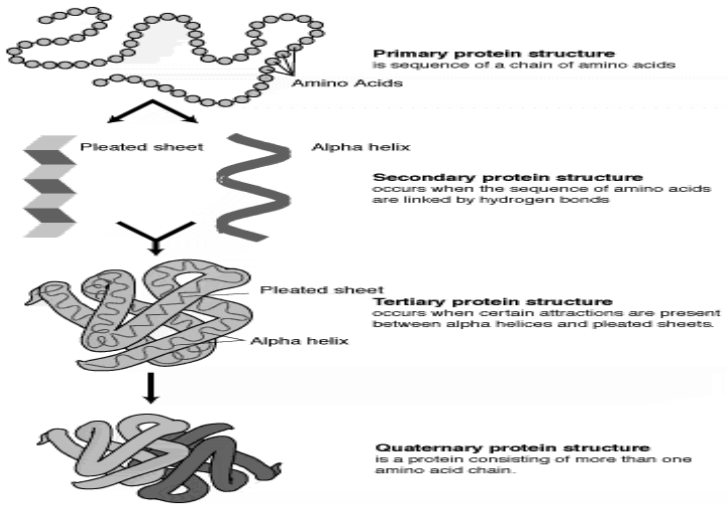
- The end of the message is marked by a STOP codon: UAA, UAG, UGG.
- No tRNA molecules have anticodons for STOP codons. A protein release factor recognizes these codons when they arrive at the A site.
- Binding of this protein releases the polypeptide from the ribosome.
- The ribosome splits into its subunits, which can later be reassembled for another round of protein synthesis.

Chain of Amino Acids

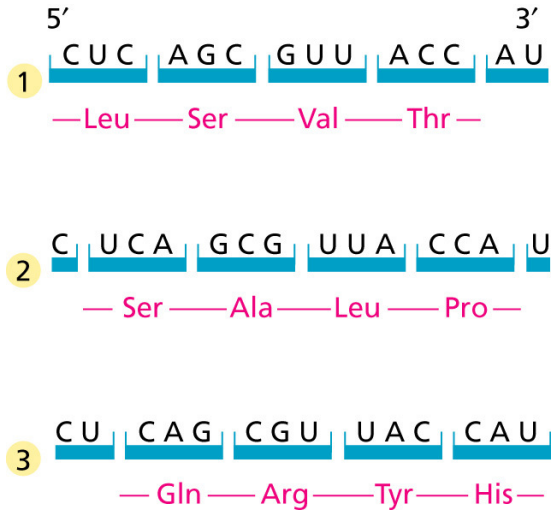




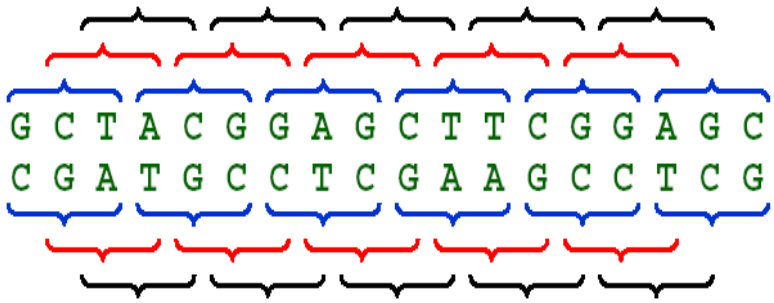
Protein Structures



3 Reading Frames of mRNA



Six Reading Frames



Sequencing SARS

in vivo → in vitro → in silico http://www.bcgsc.ca/bioinfo/SARS