

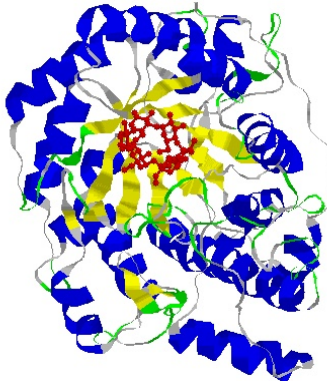
Algorithms in Bioinformatics ONE Transcription Translation

Sami Khuri
Department of Computer Science
San José State University

sami.khuri@sjsu.edu

©2018 Sami Khuri

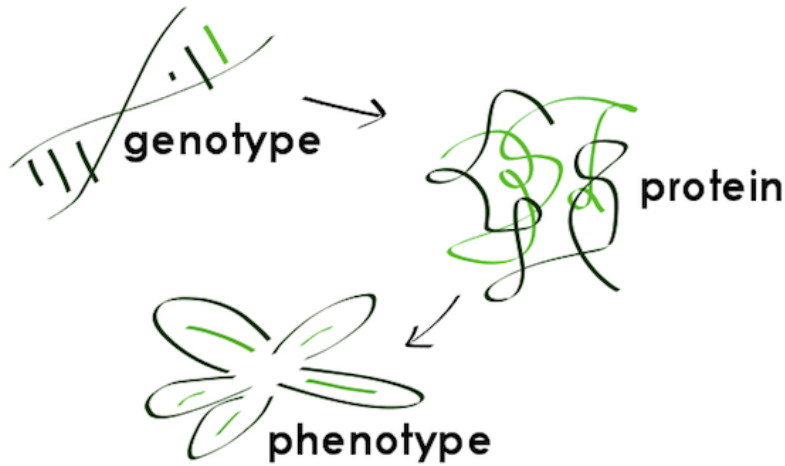
Biology Review



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

©2018 Sami Khuri

Genotype to Phenotype

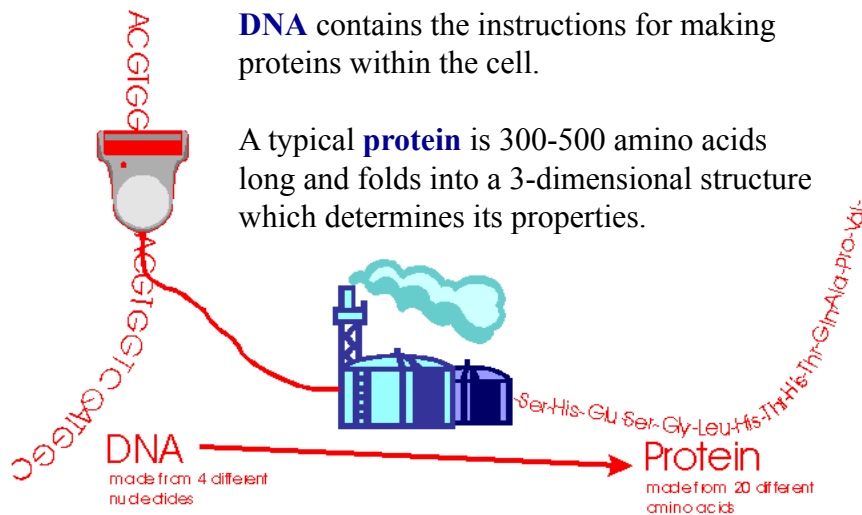


©2018 Sami Khuri

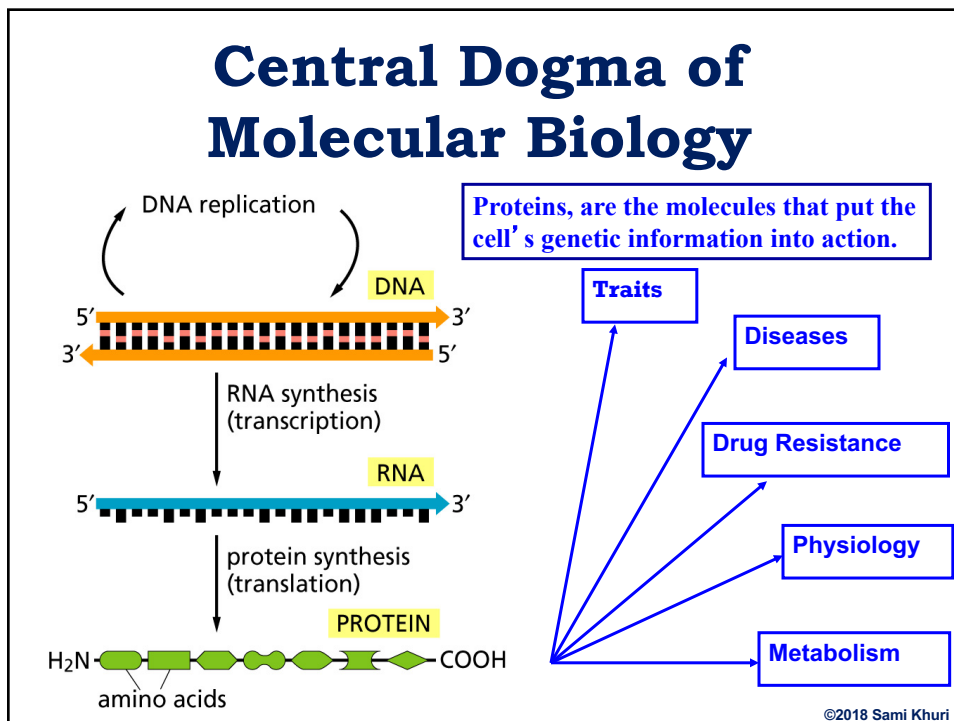
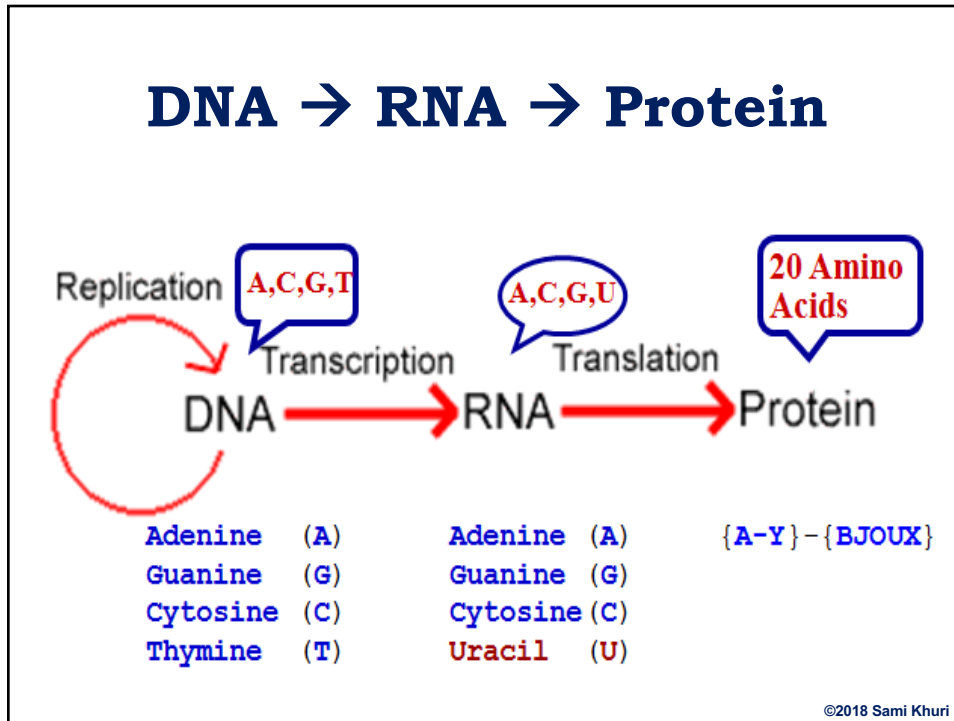
Protein Factory

DNA contains the instructions for making proteins within the cell.

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



©2018 Sami Khuri



Prokaryotes and Eukaryotes

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

There are two kinds of cells:

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

©2018 Sami Khuri

Proteins and Nucleic Acids

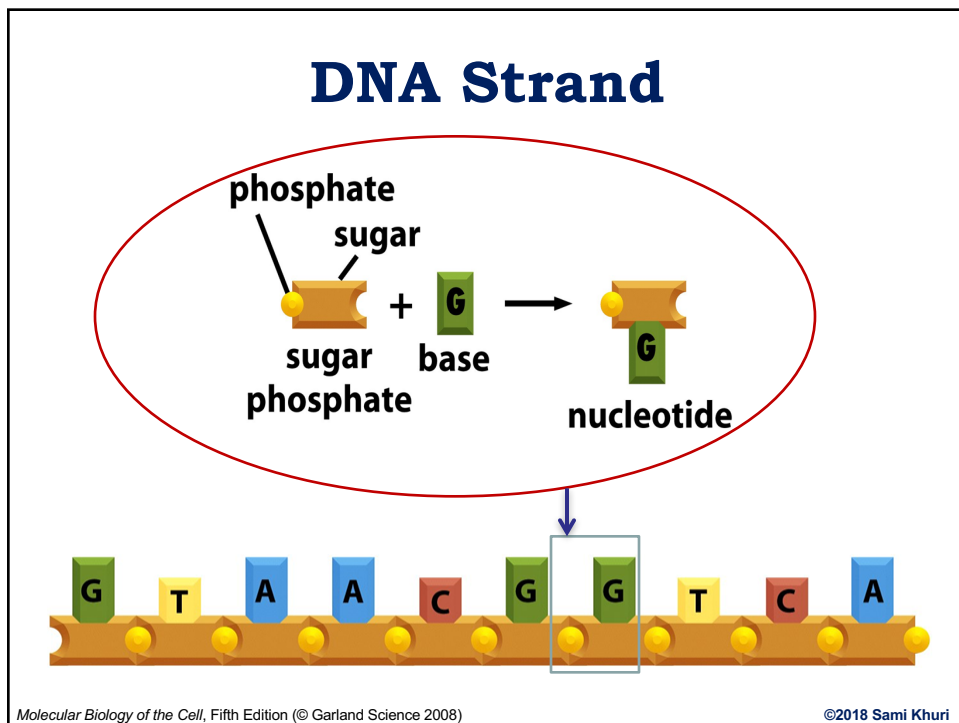
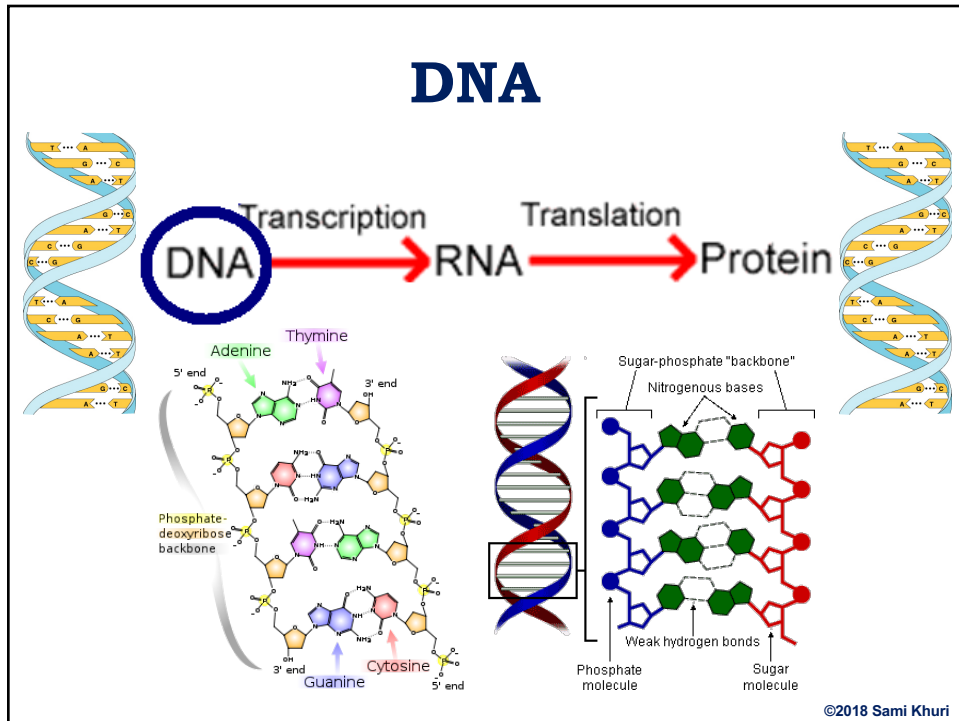
All living organisms have a similar molecular chemistry. The main actors in the chemistry of life are molecules:

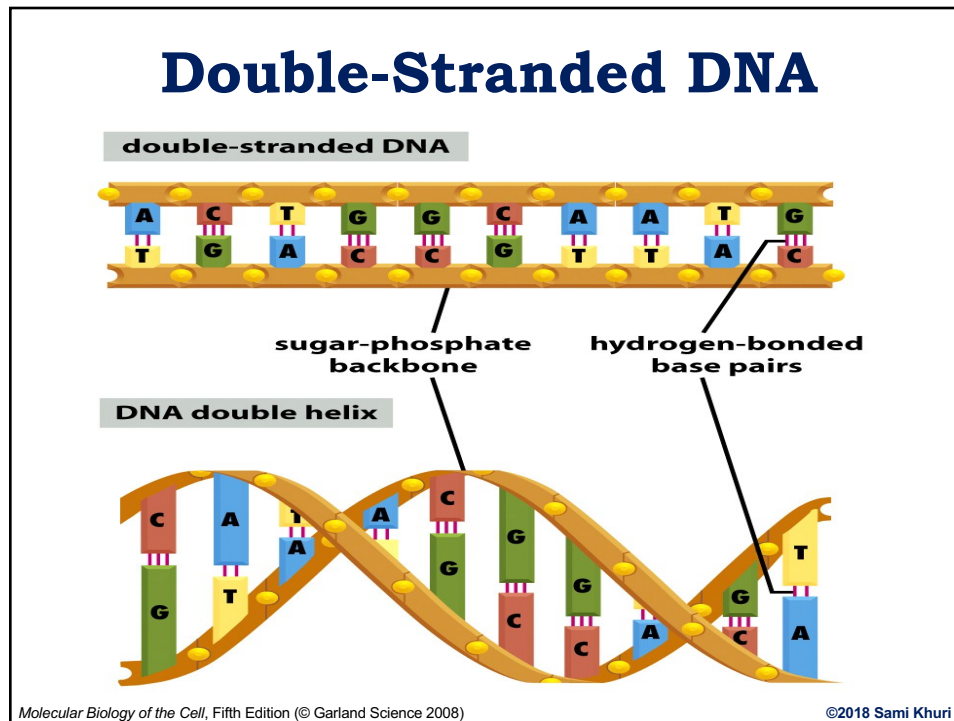
- **proteins**: which are responsible for what a living being is and does in a physical sense.
“We **are** our proteins” R. Doolittle.
- **nucleic acids**: which encode the information necessary to produce proteins and are responsible for passing the “recipe” to subsequent generations.

Living organisms contain 2 kinds of nucleic acids:

- **Ribonucleic acid (RNA)**
- **Deoxyribonucleic acid (DNA)**

©2018 Sami Khuri



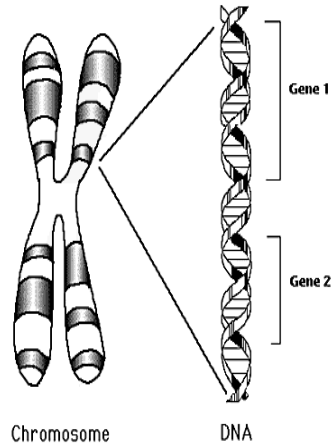


Double Helix

- 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.

©2018 Sami Khuri

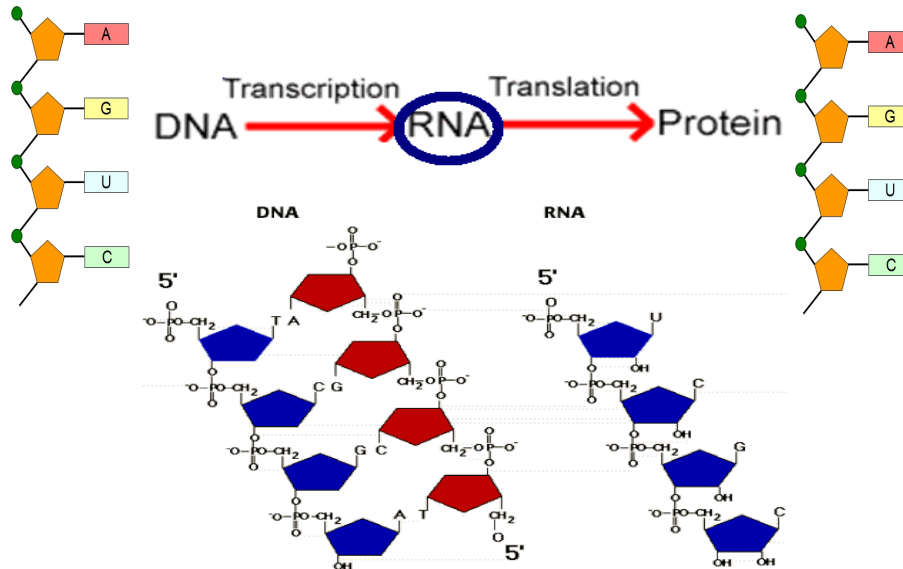
Genes



- A **gene** is a specific sequence of nucleotide bases along a chromosome carrying information for constructing a protein. A gene encodes a protein (or an RNA).
- The distance between **genes** is often much larger than the genes themselves.
- The human genome has around 23,500 genes.

©2018 Sami Khuri

RNA



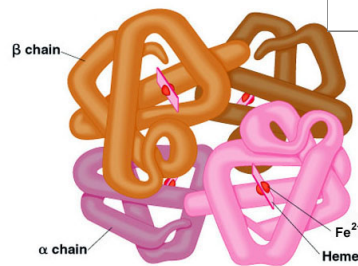
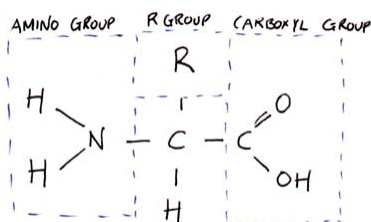
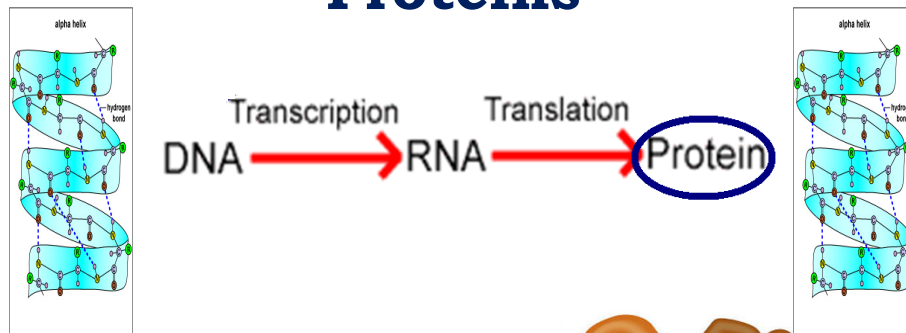
©2018 Sami Khuri

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**.
- **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**.
- **RNA** molecules tend to have a less-regular three-dimensional structure than DNA.

©2018 Sami Khuri

Proteins



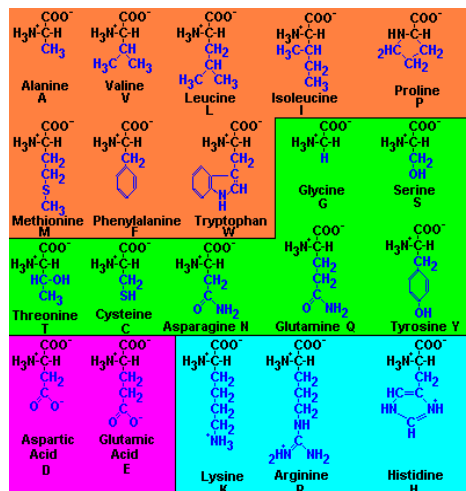
©2018 Sami Khuri

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.

©2018 Sami Khuri

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.

©2018 Sami Khuri

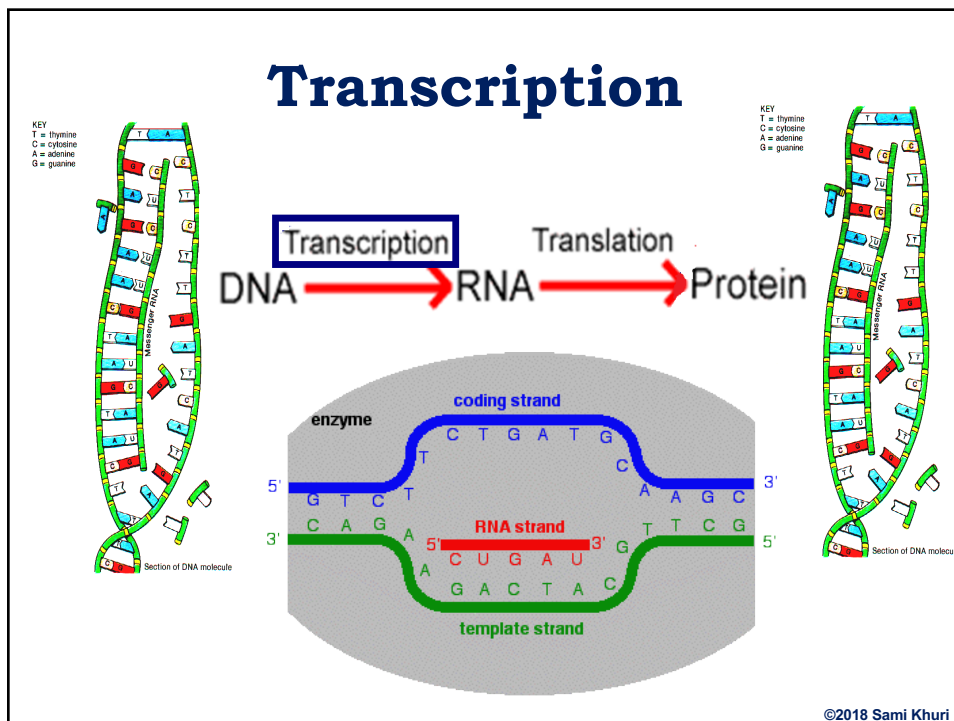
The 20 Amino Acids

1-letter	3-letter	Amino acid	1-letter	3-letter	Amino Acid
A	Ala	Alanine	M	Met	Methionin
C	Cys	Cysteine	N	Asn	Asparagine
D	Asp	Aspartic Acid	P	Pro	Proline
E	Glu	Glutamic Acid	Q	Gln	Glutamine
F	Phe	Phenylalanine	R	Arg	Arginine
G	Gly	Glycine	S	Ser	Serine
H	His	Histidine	T	Thr	Threonin
I	Ile	Isoleucine	V	Val	Valine
K	Lys	Lysine	W	Trp	Tryptophan
L	Leu	Leucine	Y	Tyr	Tyrosine

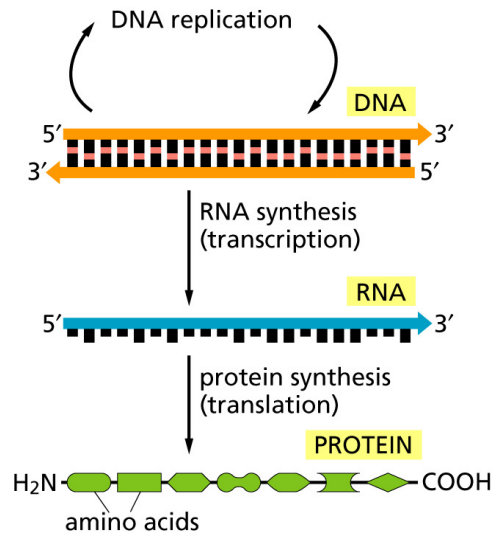
Patrice Koehl

©2018 Sami Khuri

Transcription



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**.

©2018 Sami Khuri

Synthesizing RNA: 5' to 3'

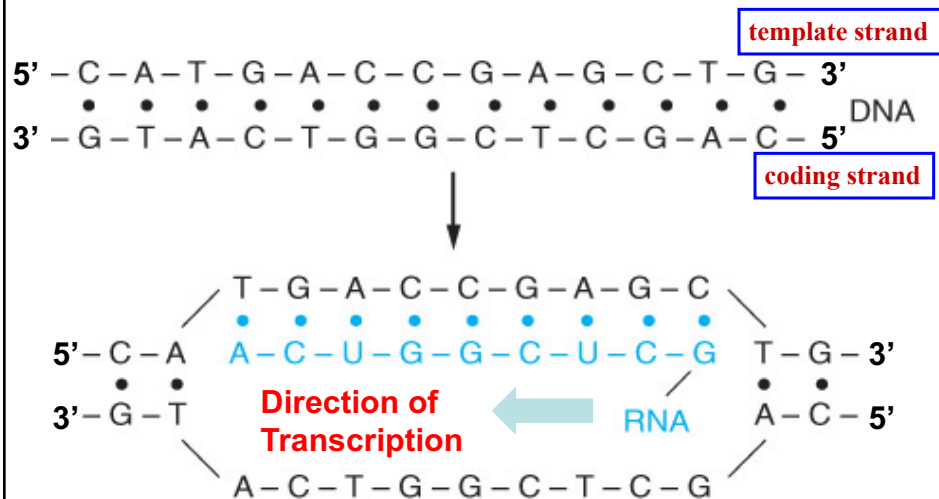
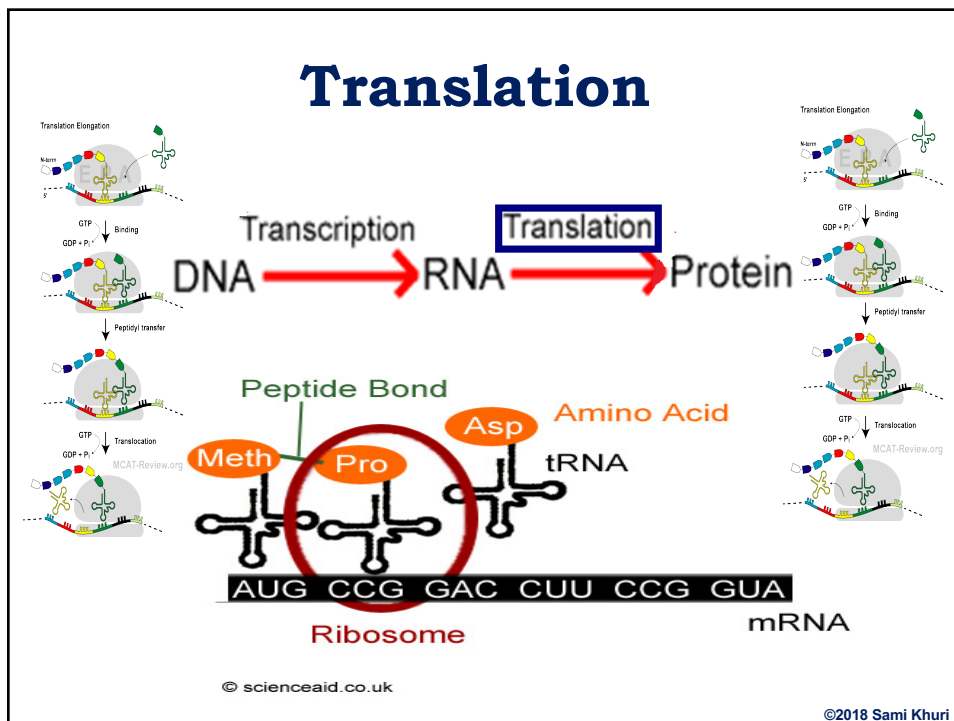
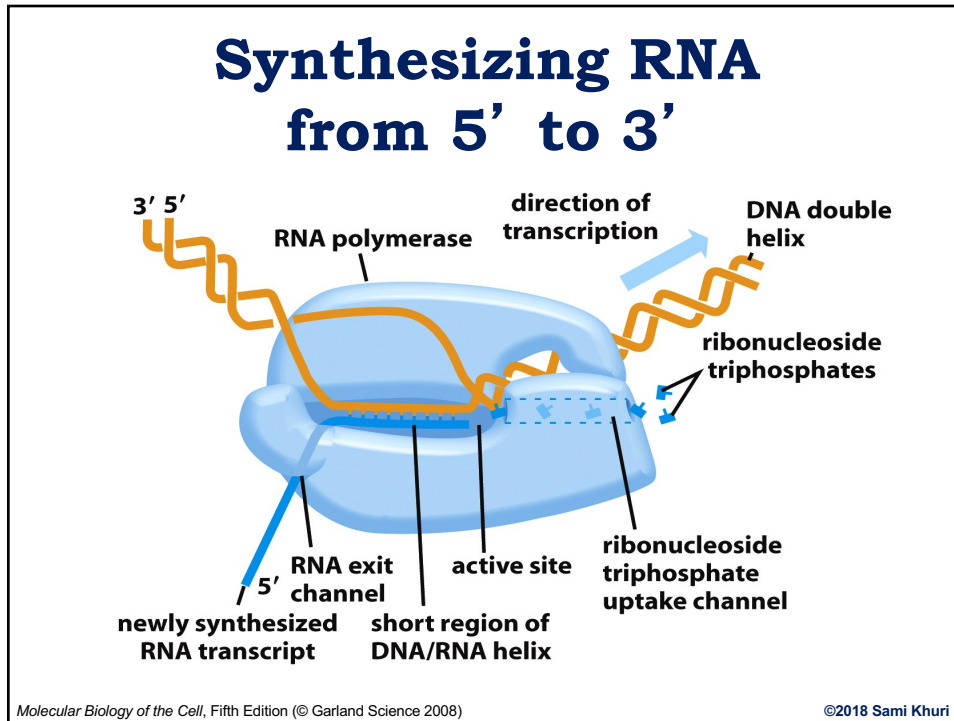


Figure 2.6: Synthesis of RNA, a single-stranded molecule complementary to one of the two DNA strands

©2018 Sami Khuri

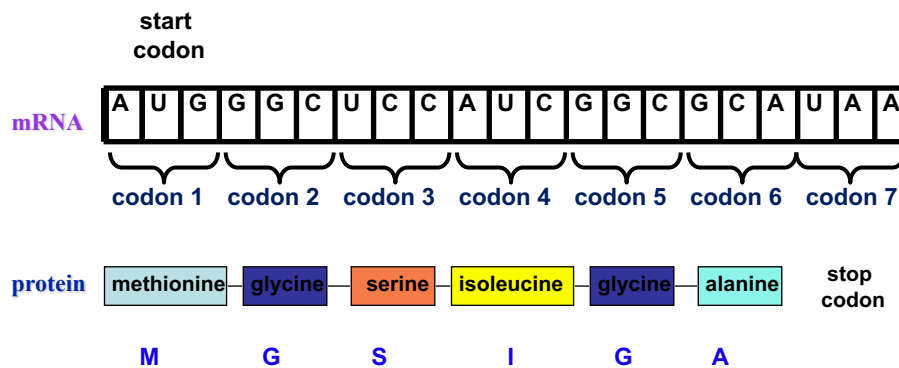


The Genetic Code

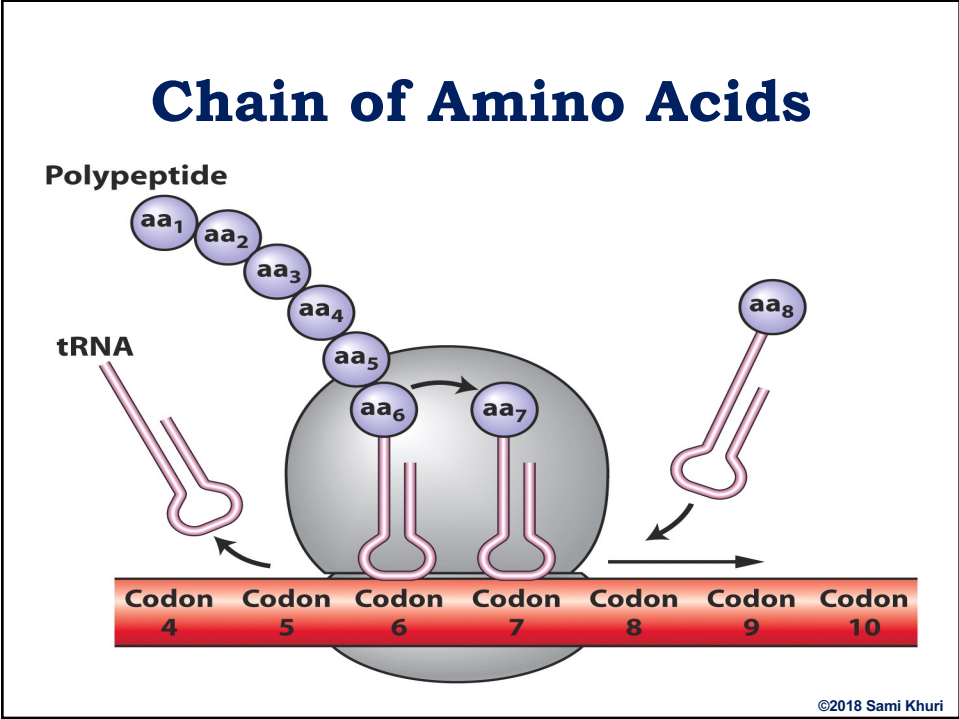
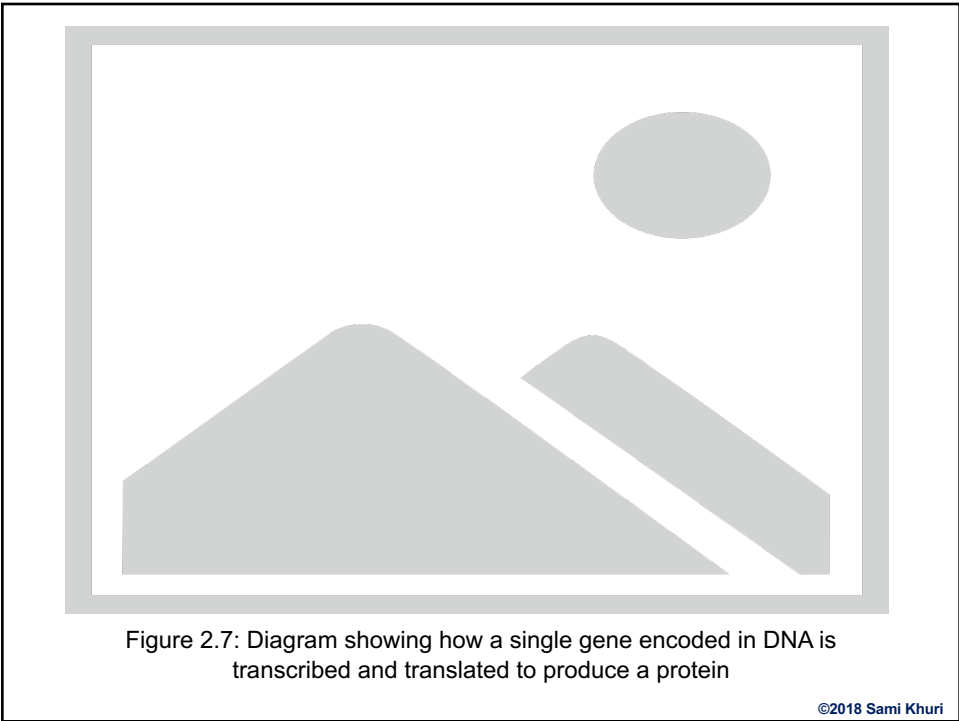
The Genetic Code						
First Codon Position (5' End)		Second Codon Position				Third Codon Position (3' End)
		U	C	A	G	
U	U	UUU Phe (F)	UCU Ser (S)	UAU Tyr (Y)	UGU Cys (C)	U
		UUC Phe (F)	UCC Ser (S)	UAC Tyr (Y)	UGC Cys (C)	C
		UUA Leu (L)	UCA Ser (S)	UAA Stop	UGA Stop	A
		UUG Leu (L)	UCG Ser (S)	UAG Stop	UGG Trp (W)	G
	C	CUU Leu (L)	CCU Pro (P)	CAU His (H)	CGU Arg (R)	U
		CUC Leu (L)	CCC Pro (P)	CAC His (H)	CGC Arg (R)	C
		CUA Leu (L)	CCA Pro (P)	CAA Gln (Q)	CGA Arg (R)	A
		CUG Leu (L)	CCG Pro (P)	CAG Gln (Q)	CGG Arg (R)	G
	A	AUU Ile (I)	ACU Thr (T)	AAU Asn (N)	AGU Ser (S)	U
		AUC Ile (I)	ACC Thr (T)	AAC Asn (N)	AGC Ser (S)	C
		AUA Ile (I)	ACA Thr (T)	AAA Lys (K)	AGA Arg (R)	A
		AUG Met (M)	ACG Thr (T)	AAG Lys (K)	AGG Arg (R)	G
	G	GUU Val (V)	GCU Ala (A)	GAU Asp (D)	GGU Gly (G)	U
		GUC Val (V)	GCC Ala (A)	GAC Asp (D)	GGC Gly (G)	C
		GUA Val (V)	GCA Ala (A)	GAA Glu (E)	GGA Gly (G)	A
		GUG Val (V)	GCG Ala (A)	GAG Glu (E)	GGG Gly (G)	G

©2018 Sami Khuri

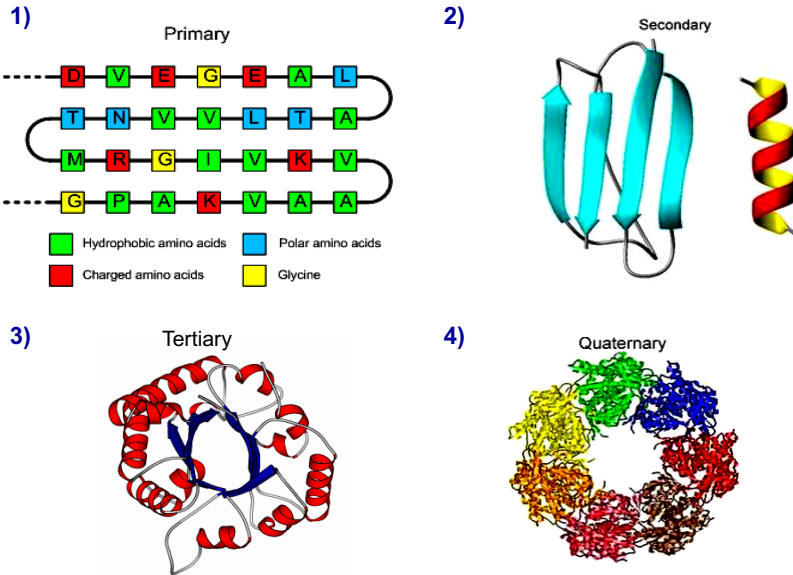
Translation: An Example



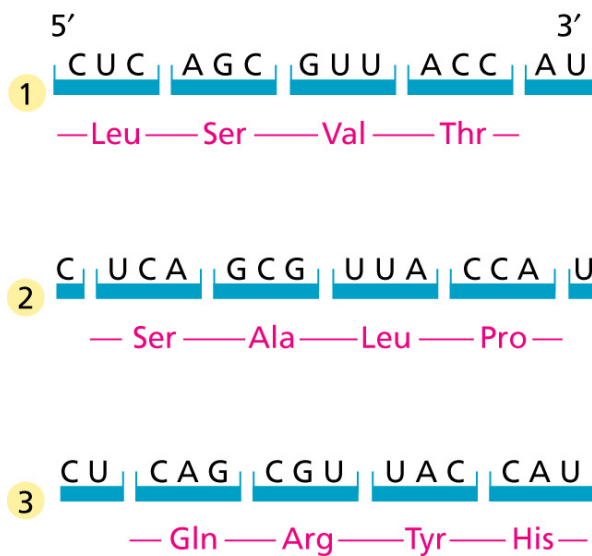
©2018 Sami Khuri



Four Structures of Proteins

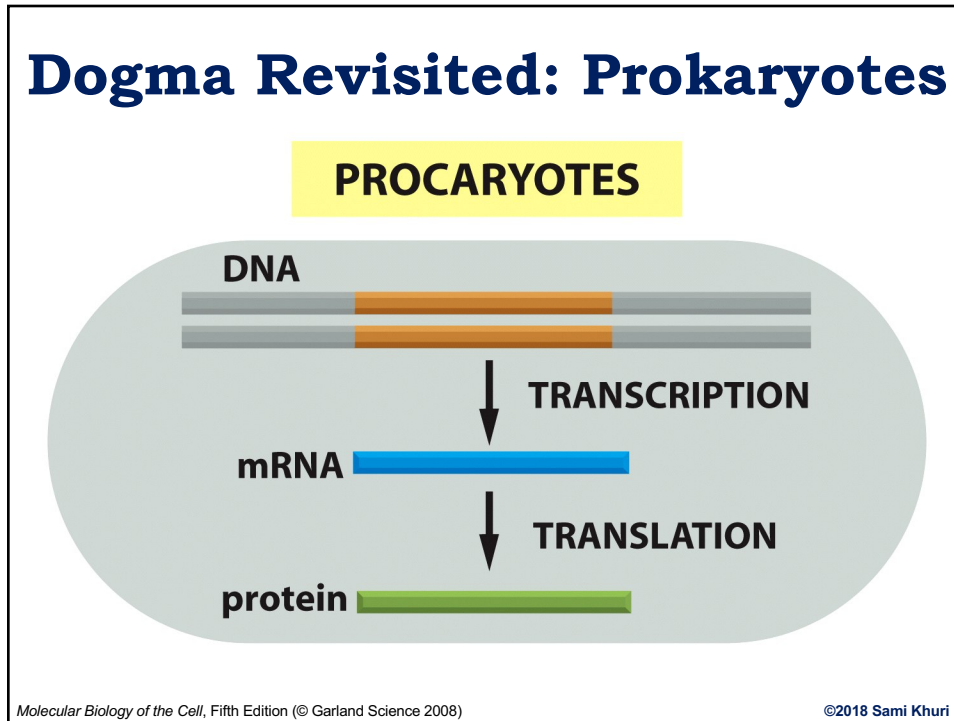


3 Reading Frames of mRNA

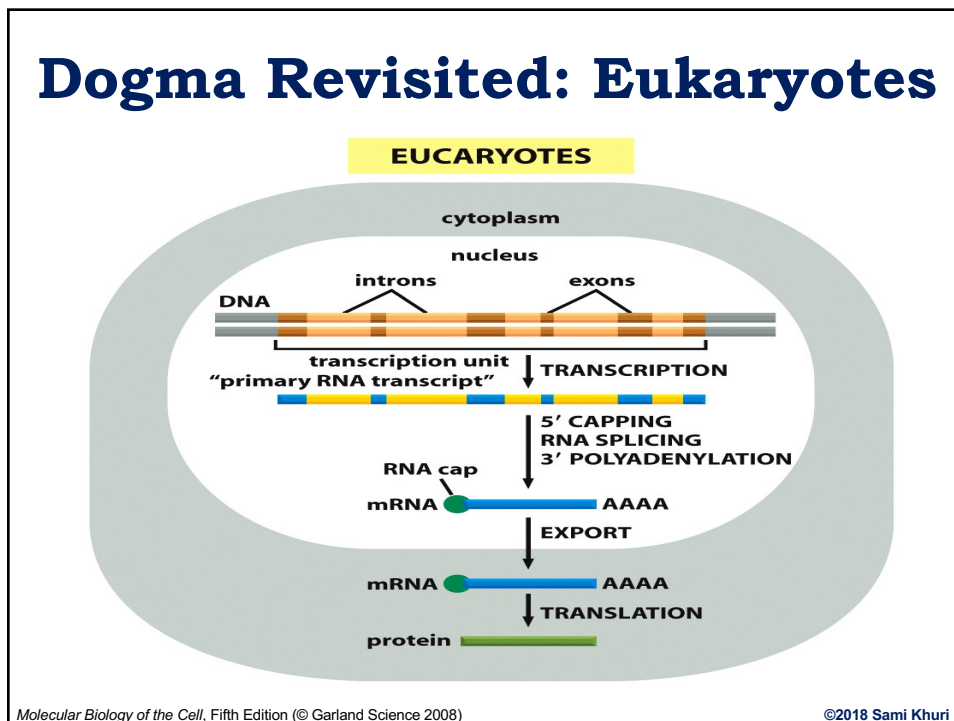


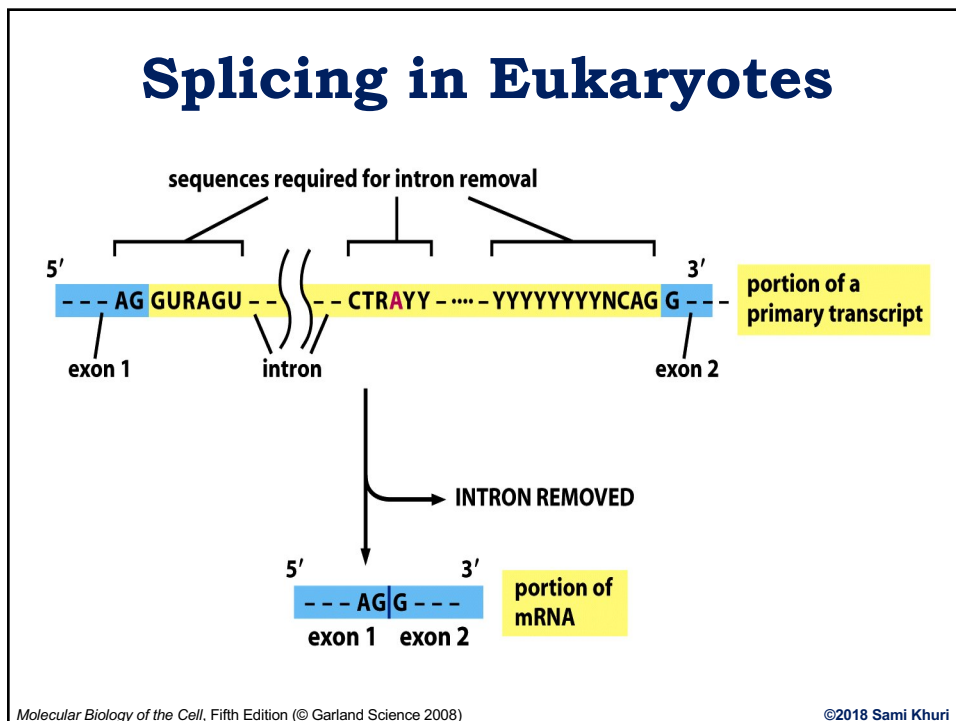
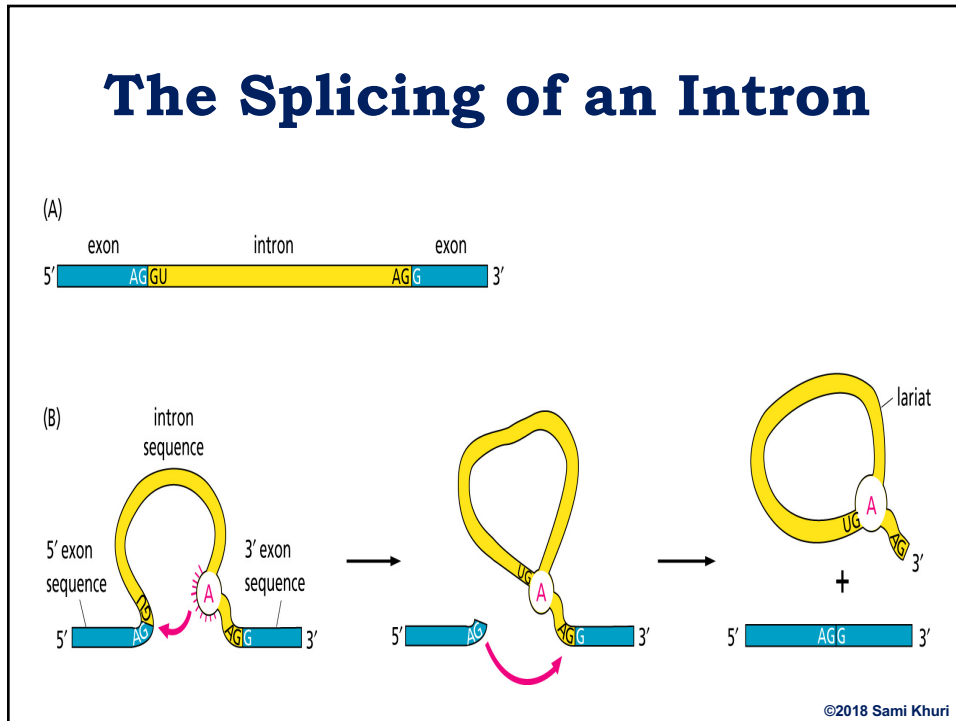
©2018 Sami Khuri

Dogma Revisited: Prokaryotes



Dogma Revisited: Eukaryotes

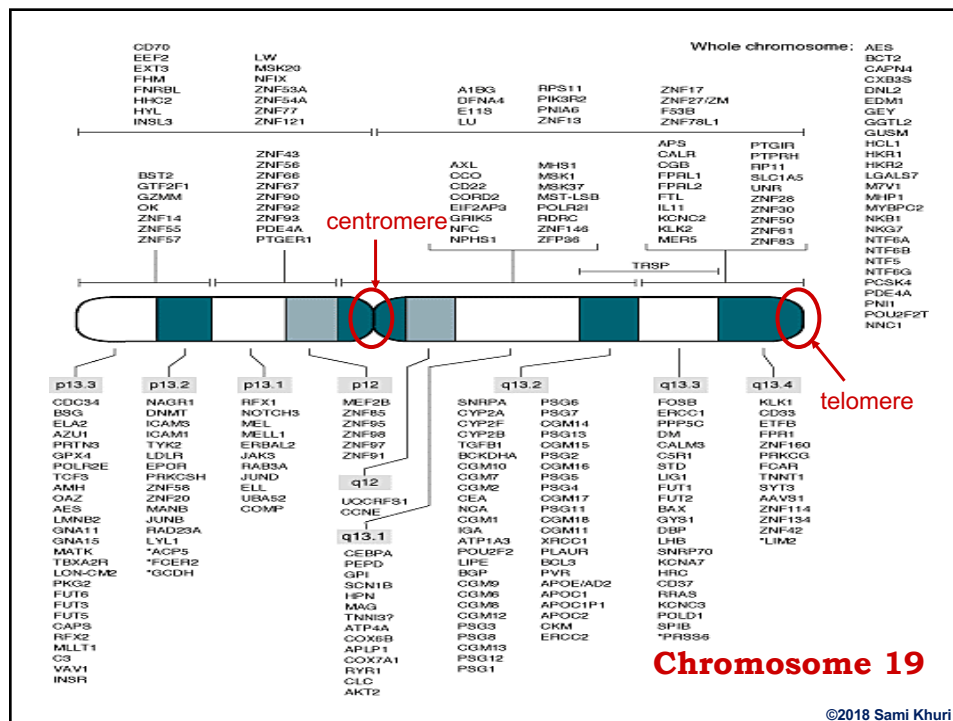




The Human Genome Project

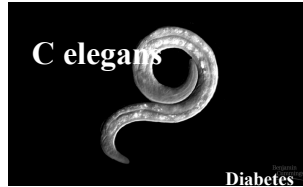
- The **HGP** is a multinational effort, begun by the USA in 1988, whose aim is to produce a complete physical map of all human chromosomes, as well as the entire human DNA sequence.
- The ultimate goal of genome research is to find all the **genes** in the **DNA sequence** and to develop tools for using this information in the study of **human biology** and **medicine**.
- The primary goal of the project is to make a series of descriptive diagrams (called **maps**) of each human chromosome at increasingly finer resolutions.

©2018 Sami Khuri



Other Species

As part of the HGP, genomes of other organisms, such as bacteria, yeast, flies and mice were also being studied.



Baker's yeast



DNA repair
Cell division



Chimps are infected with SIV
Very rarely progress to AIDS

©2018 Sami Khuri

Model Organisms

- A **model organism** is an organism that is extensively studied to understand particular biological phenomena.
- **Why have model organisms?** The hope is that discoveries made in model organisms will provide insight into the workings of other organisms.
- **Why is this possible?** This works because evolution reuses fundamental biological principles and conserves metabolic, regulatory, and developmental pathways.

©2018 Sami Khuri

HGP Finished Before Deadline

- In 1991, the USA Congress was told that the HGP could be done by 2005 for \$3 billion.
- It ended in 2003 for \$2.7 billion, because of efficient computational methods.

©2018 Sami Khuri

What is Bioinformatics? A Discipline

- The field of science, in which **biology**, **computer science**, and **information technology** merge into a single discipline.

Definition of NCBI (National Center for Biotechnology Information)

- The ultimate goal of **bioinformatics** is to enable the discovery of new biological insights and to create a global perspective from which unifying principles in biology can be discerned.

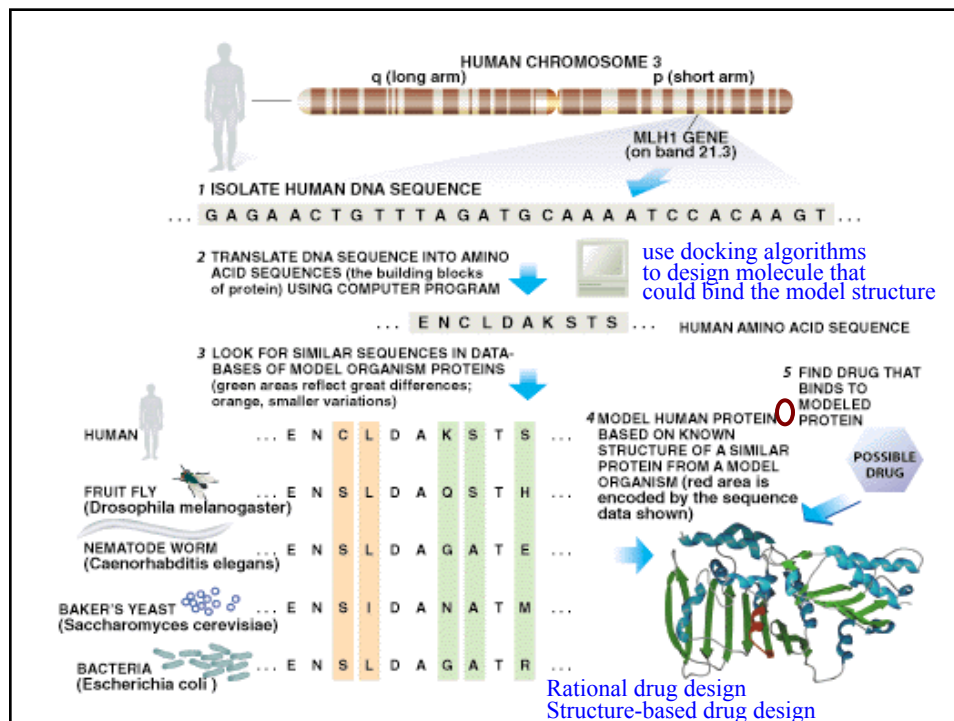
©2018 Sami Khuri

Why Study Bioinformatics

- Bioinformatics is intrinsically interesting
- Bioinformatics offers the prospect of finding better drug targets earlier in the drug development process.
 - By looking for genes in model organisms that are similar to a given human gene, researchers can learn about the protein the human gene encodes and search for drugs to block it.



©2018 Sami Khuri





Concluding Remarks

- Biology is becoming an information science
- Progression: **in vivo** to **in vitro** to **in silico**
- Are natural languages adequate in predicting quantitative behavior of biological systems?
 - Need to produce biological knowledge and operations in ways that natural languages do not allow
- “Biology easily has 500 years of exciting problems to work on”. Donald Knuth
- The role that mathematics and computer science can play in Bioinformatics is still in an embryonic stage.

©2018 Sami Khuri