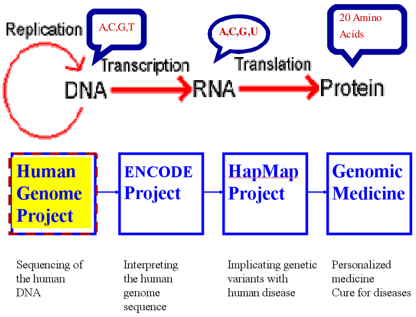
Computational Methods in Genomics

PART ONE

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

Outline

- Biology Review:
  - Central Dogma of Molecular Biology
- What is Bioinformatics?
  - Human Genome Project (HGP)
  - Importance of Model Organisms
  - Databases and Tools over the Internet
- Pairwise Sequence Alignment
  - Dynamic Programming (2008)
- Multiple Sequence Alignment
- DNA Fragment Assembly Problem
- Gene Prediction (2008)

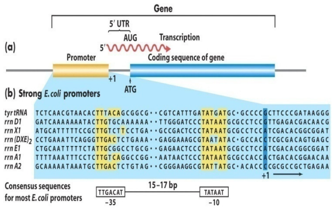


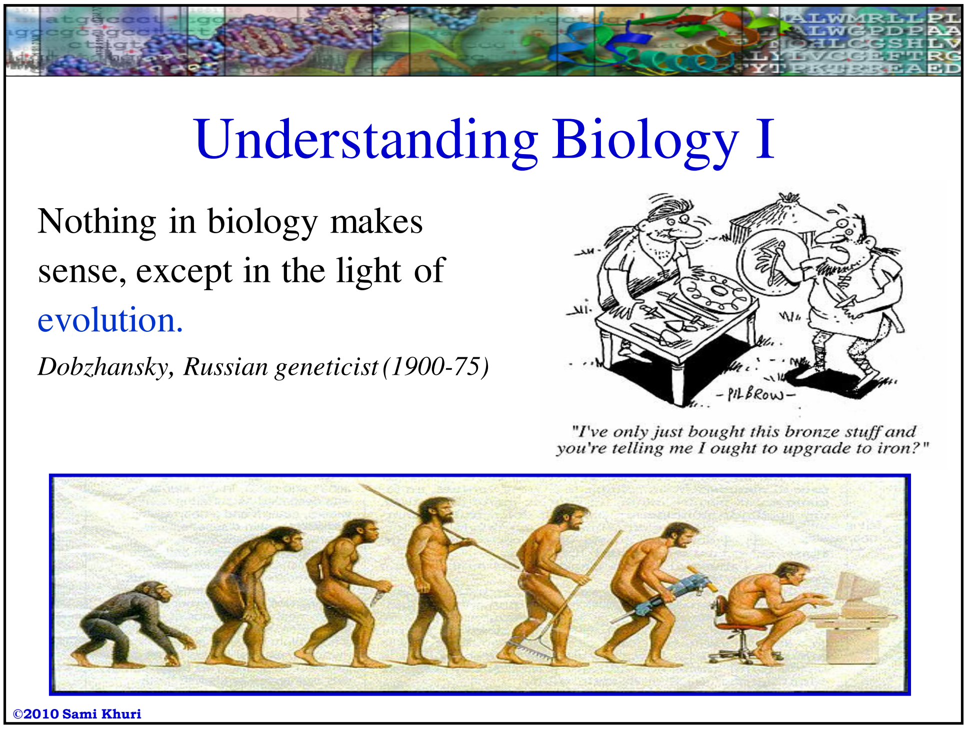
Human Genome Project: Sequencing of the human DNA

ENCODE Project: Interpreting the human genome sequence

HapMap Project: Implicating genetic variants with human disease

Genomic Medicine: Personalized medicine, Cure for diseases



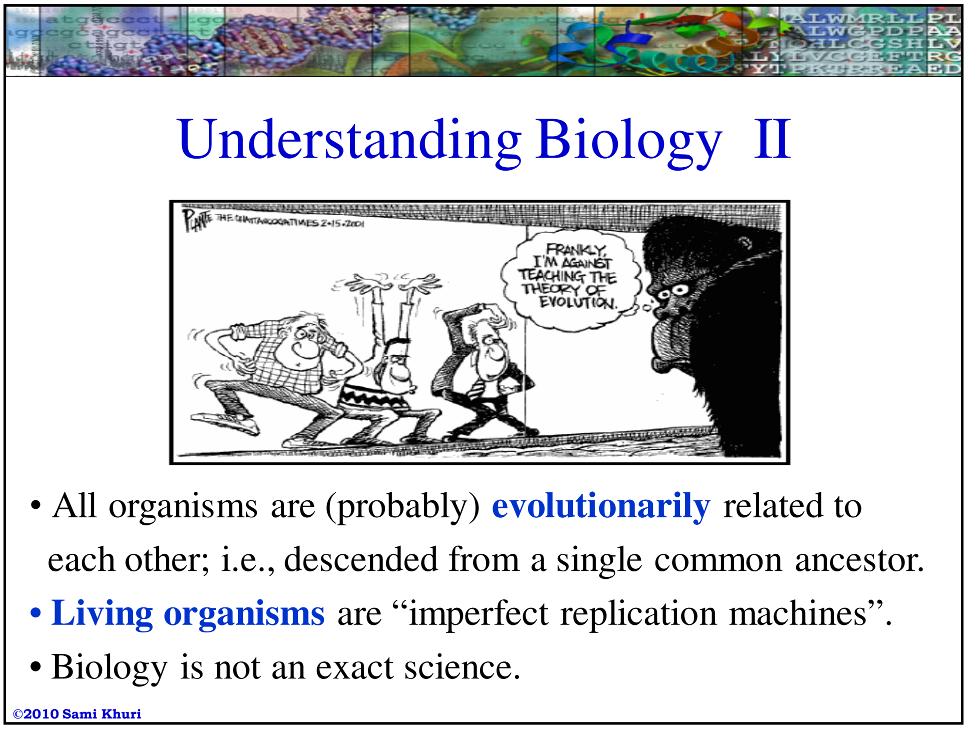


Understanding Biology I

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

Dobzhansky, Russian geneticist (1900-75)

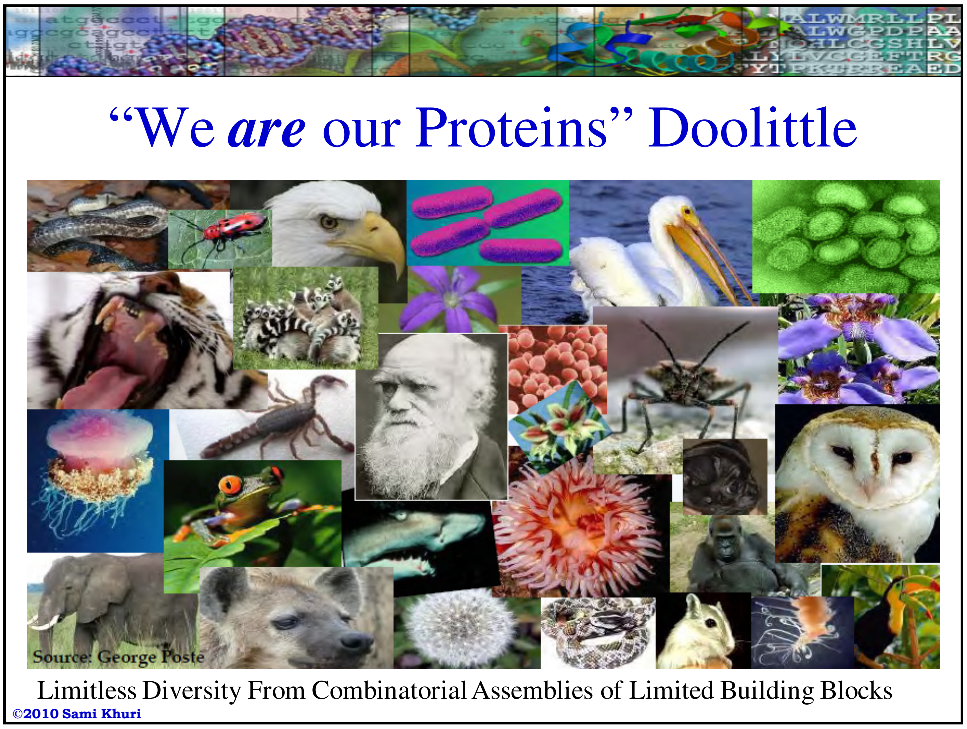
"I've only just bought this bronze stuff and you're telling me I ought to upgrade to iron?"



Understanding Biology II

PLANTE THE CHATTANOOGA TIMES 2-15-2001
FRANKLY, I'M AGAINST TEACHING THE THEORY OF EVOLUTION.

- 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.

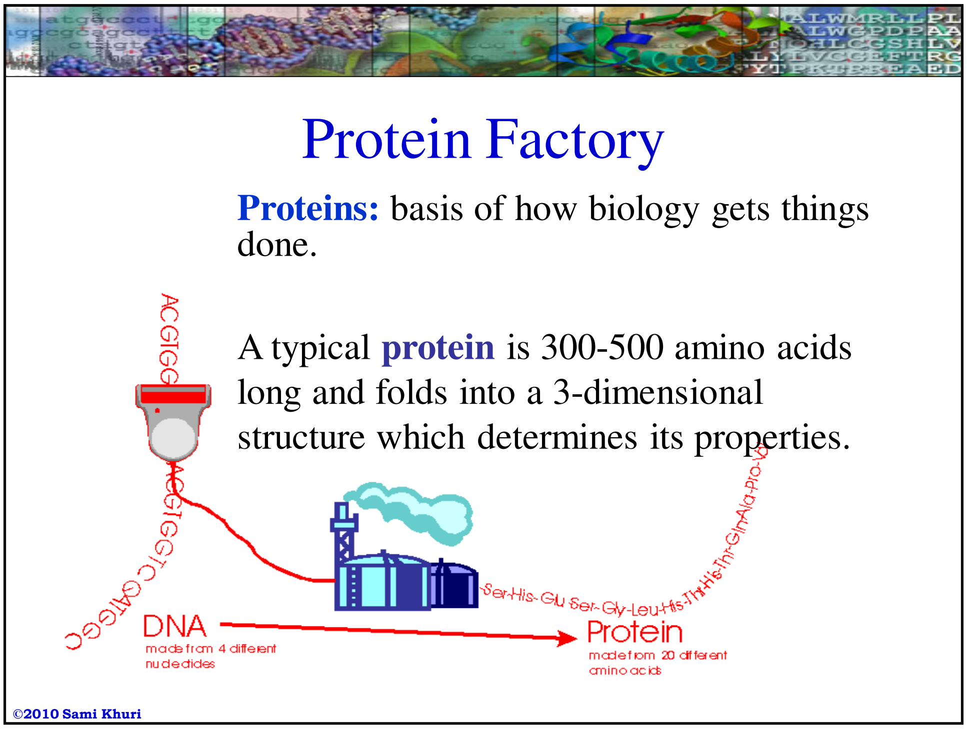


“We *are* our Proteins” Doolittle

Source: George Poste

Limitless Diversity From Combinatorial Assemblies of Limited Building Blocks

©2010 Sami Khuri



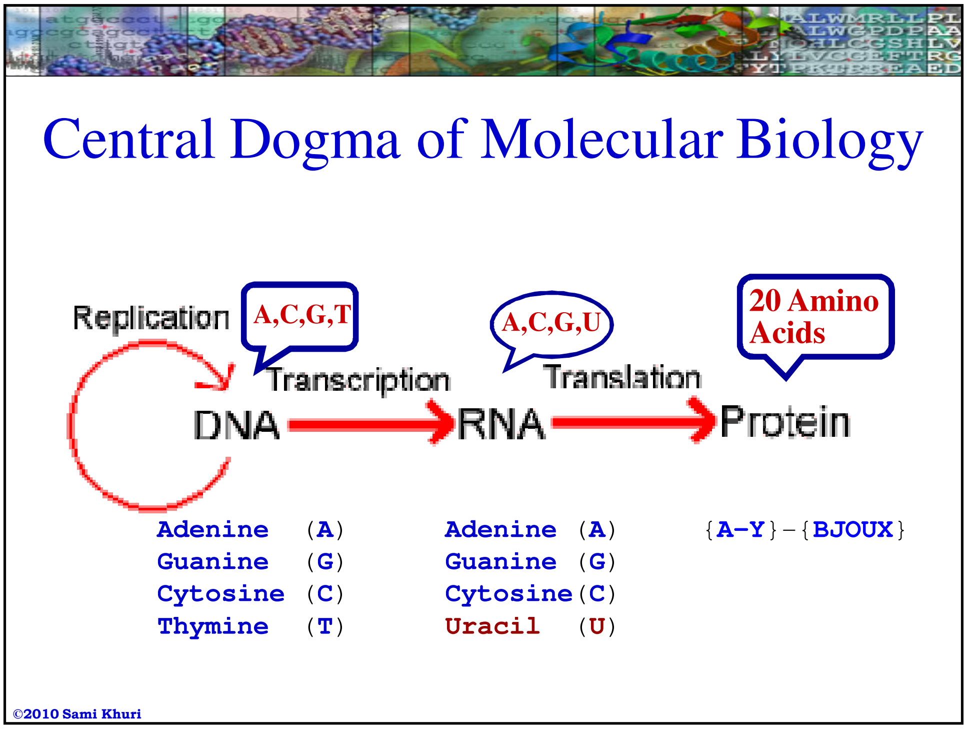
Protein Factory

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.

ACGIGG
ACGIGGIC GAIGGC
DNA
made from 4 different nucledides
Ser-His-Glu Ser-Gly-Leu-His-Thr-His-Thr-Gln-Ala-Pro-Val
Protein
made from 20 different amino acids

©2010 Sami Khuri



Central Dogma of Molecular Biology

Replication

A,C,G,T

Transcription

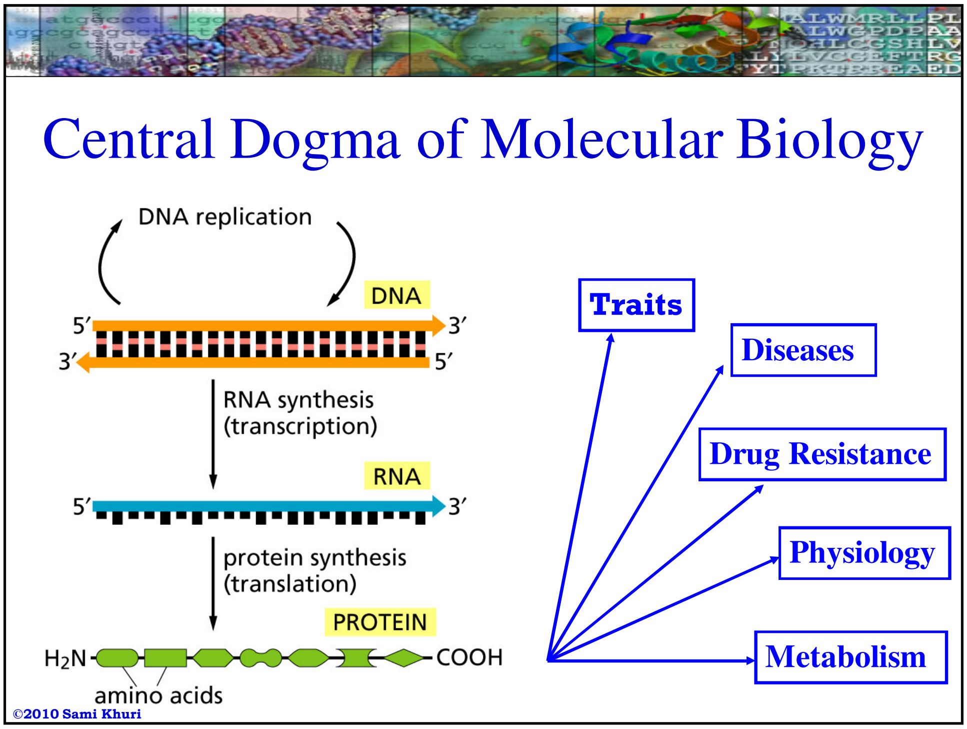
A,C,G,U

Translation

20 Amino Acids

DNA → RNA → Protein

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



Central Dogma of Molecular Biology

DNA replication

DNA

5′ 3′

3′ 5′

RNA synthesis (transcription)

RNA

5′ 3′

protein synthesis (translation)

PROTEIN

H_2N COOH

amino acids

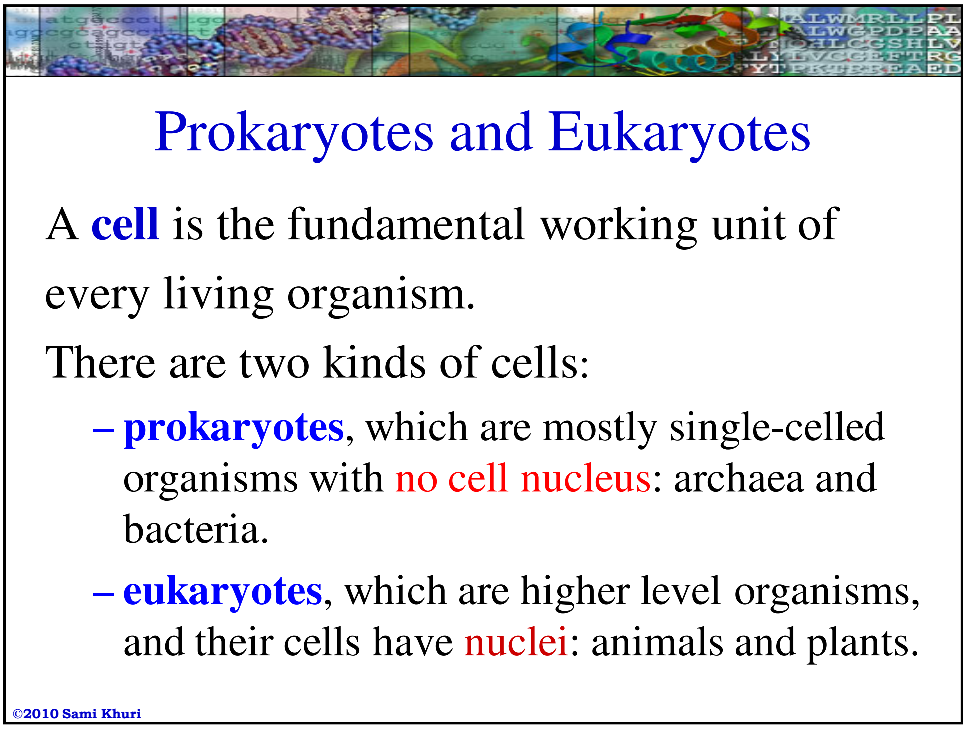
Traits

Diseases

Drug Resistance

Physiology

Metabolism

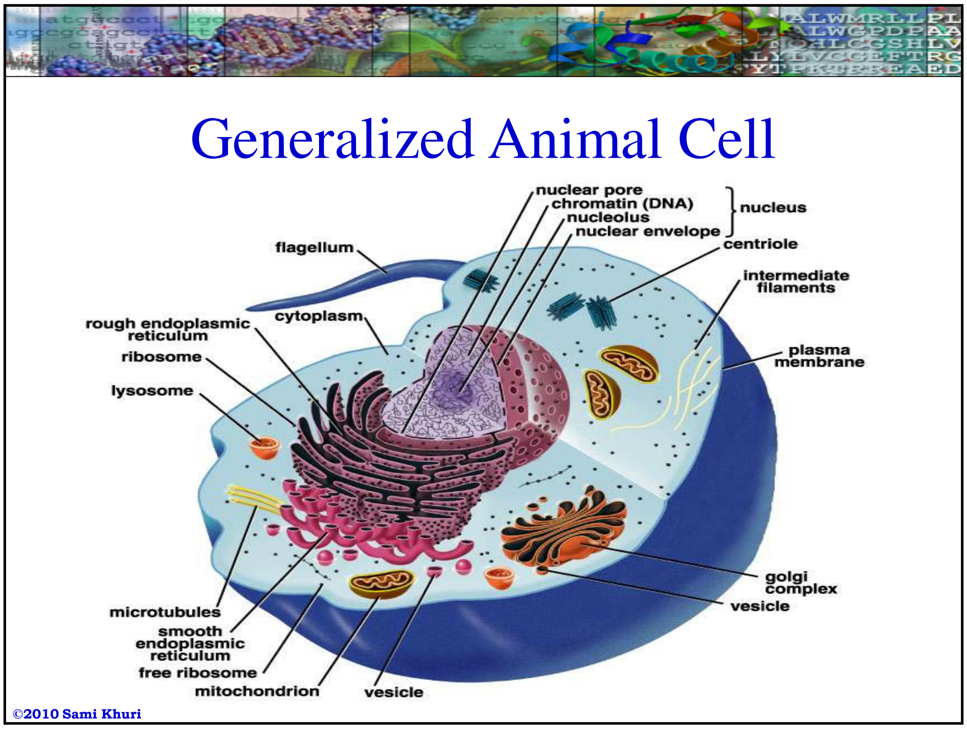


Prokaryotes and Eukaryotes

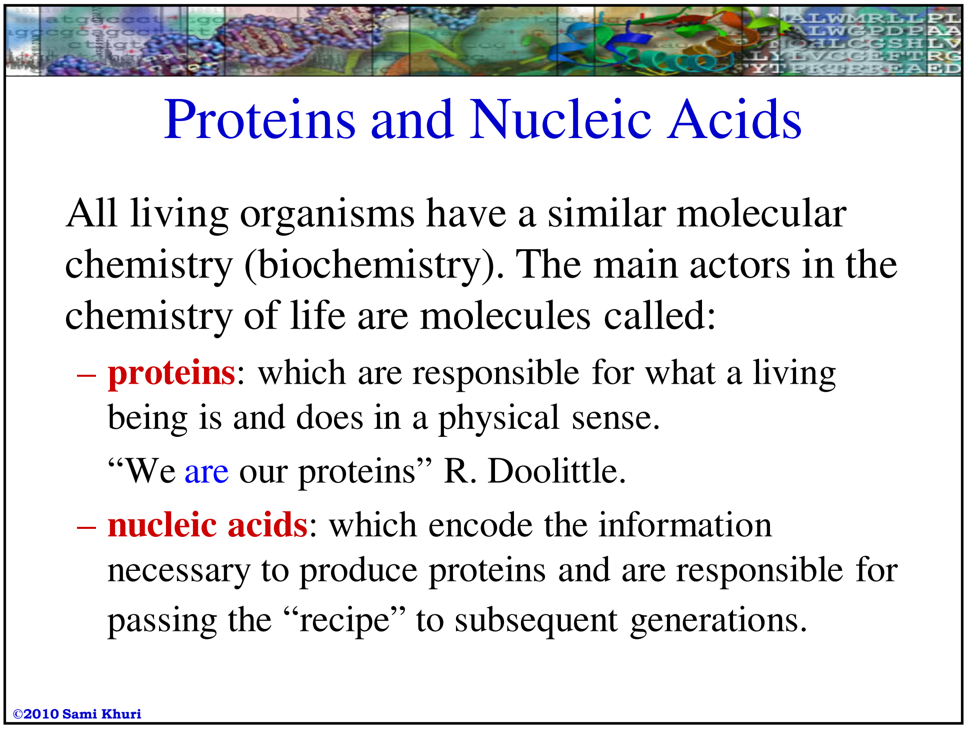
A **cell** is the fundamental working unit of every living organism.

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

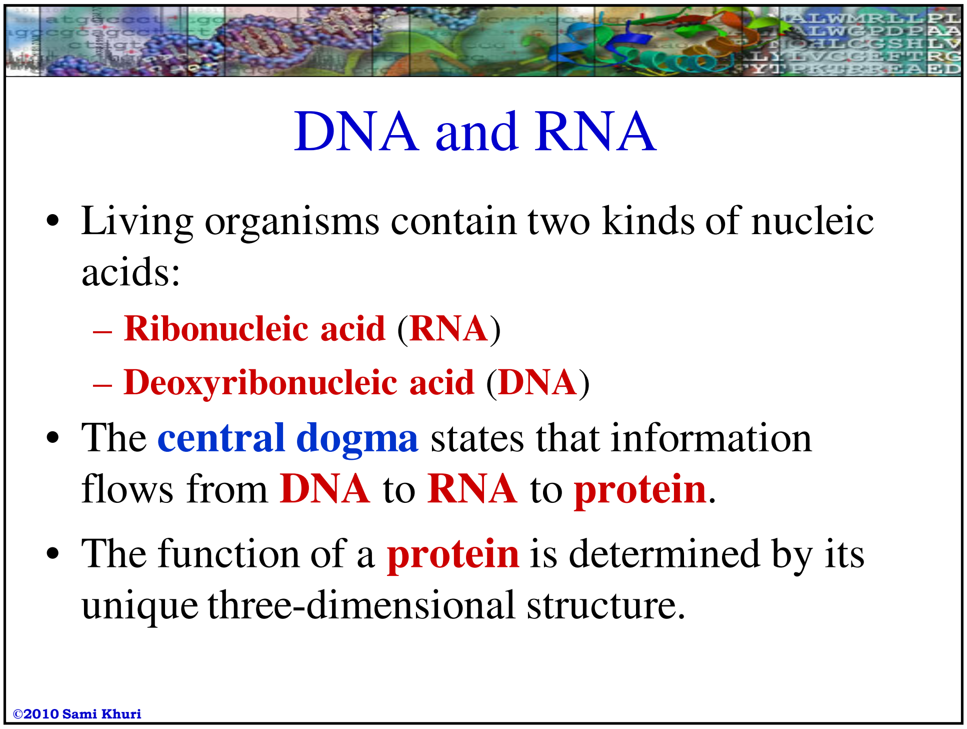


Proteins and Nucleic Acids

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

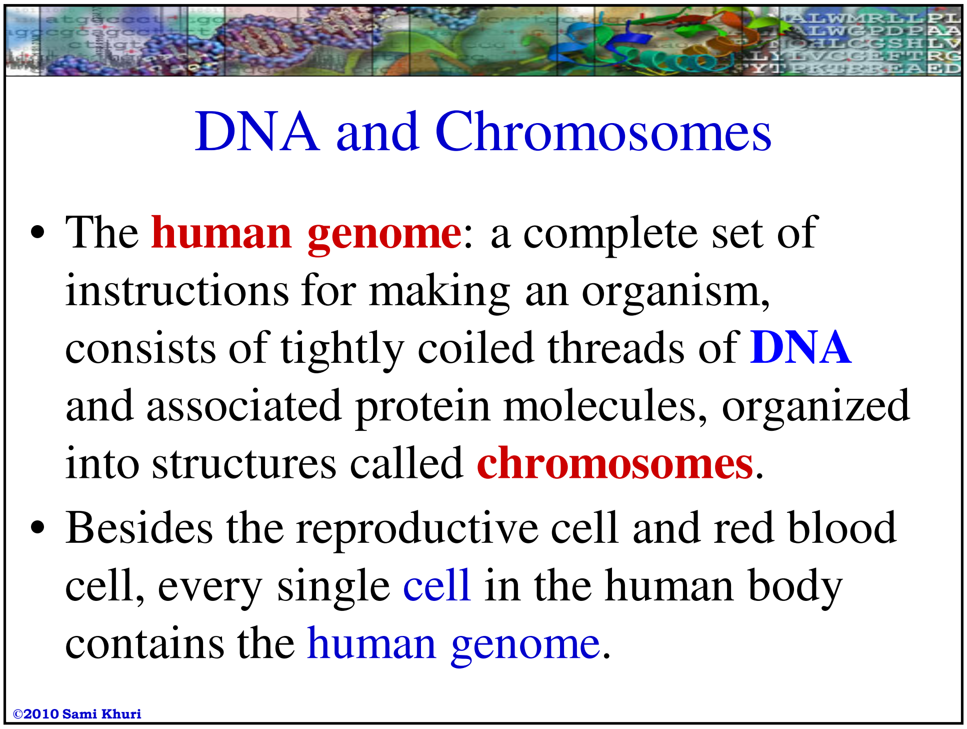
- **proteins**: which are responsible for what a living being is and does in a physical sense.

  "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.



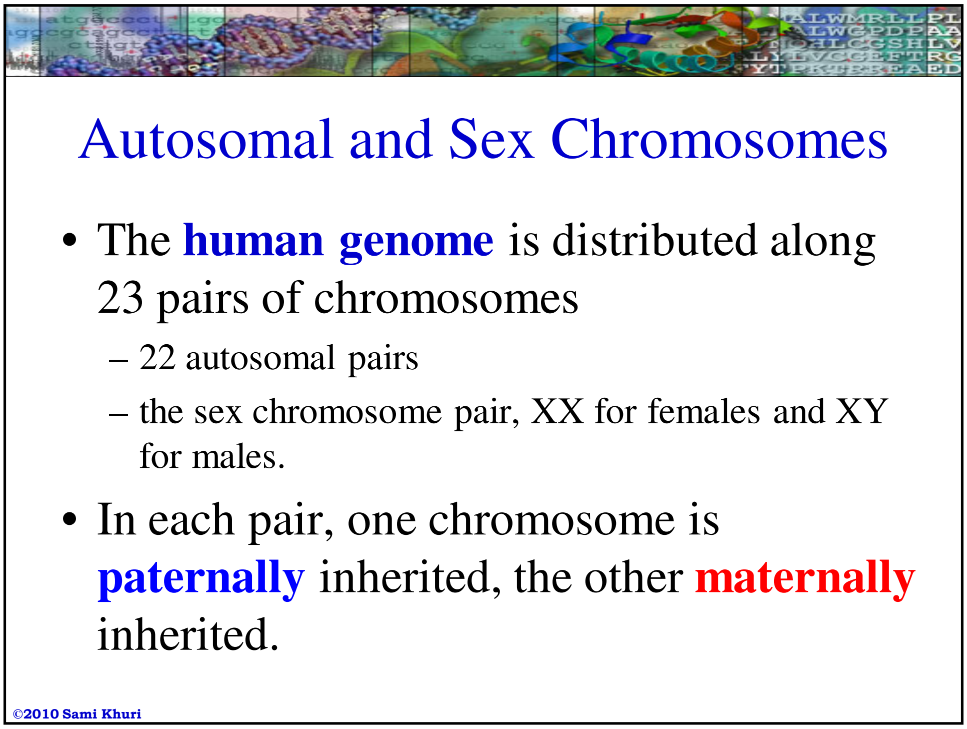
DNA and RNA

- 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.



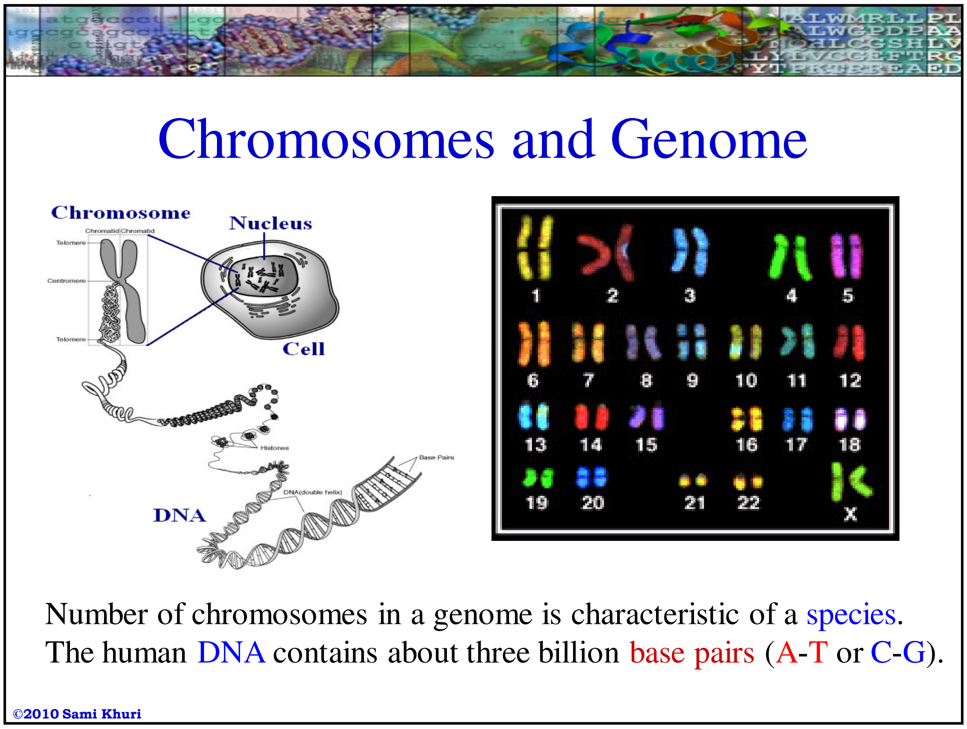
DNA and Chromosomes

- 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

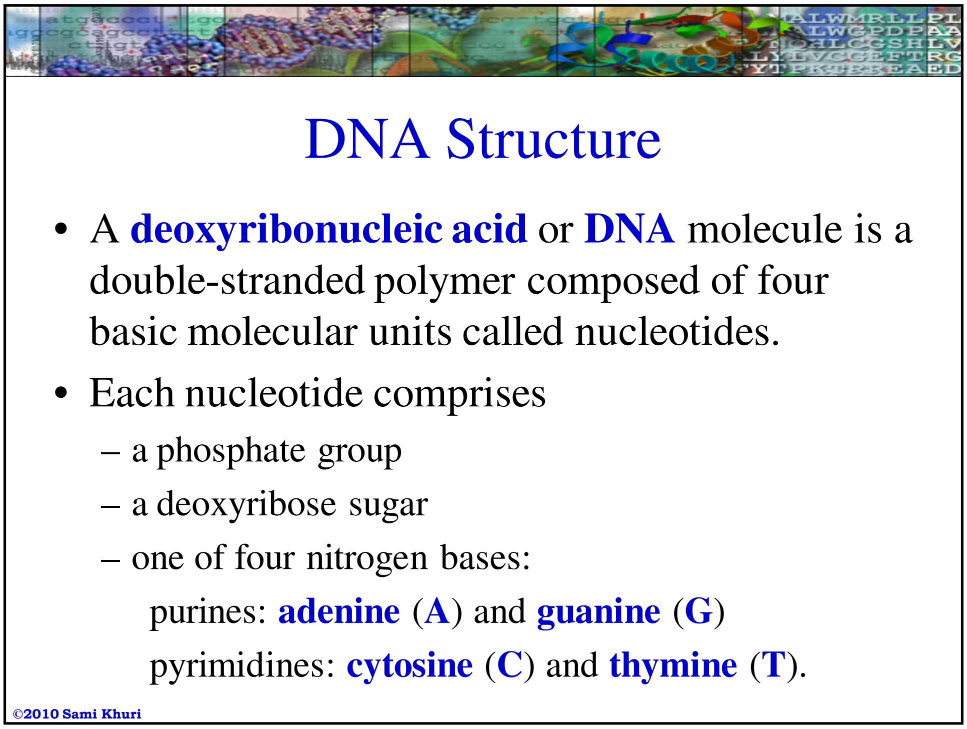
- The **human genome** is distributed along 23 pairs of chromosomes
  - 22 autosomal pairs
  - the sex chromosome pair, XX for females and XY for males.
- In each pair, one chromosome is **paternally** inherited, the other **maternally** inherited.



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).

©2010 Sami Khuri



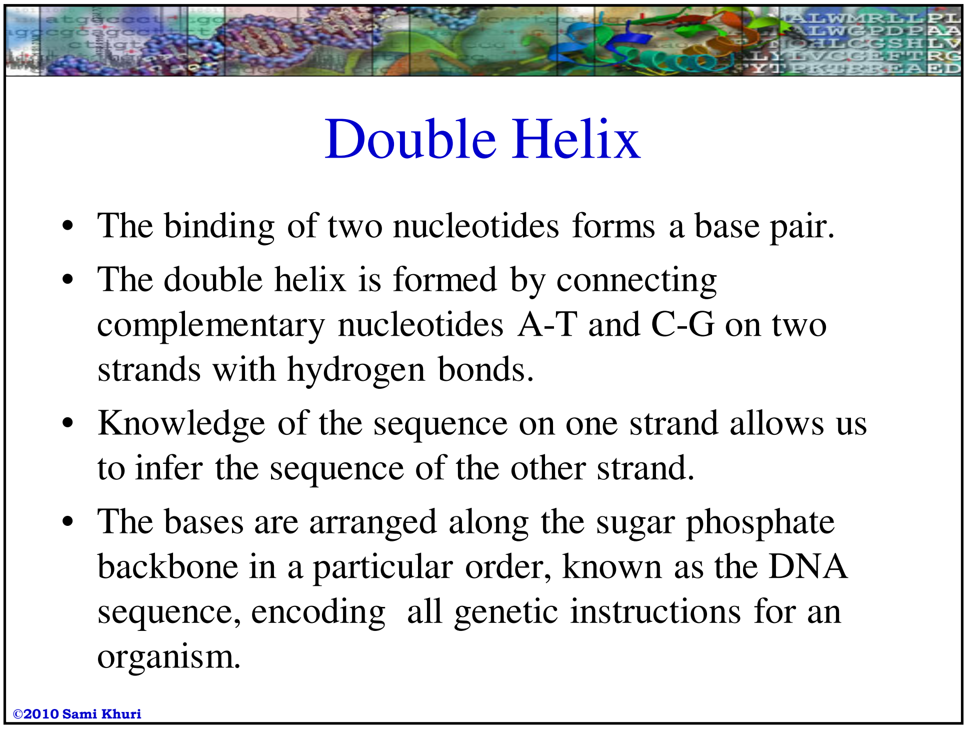
DNA Structure

- 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**)

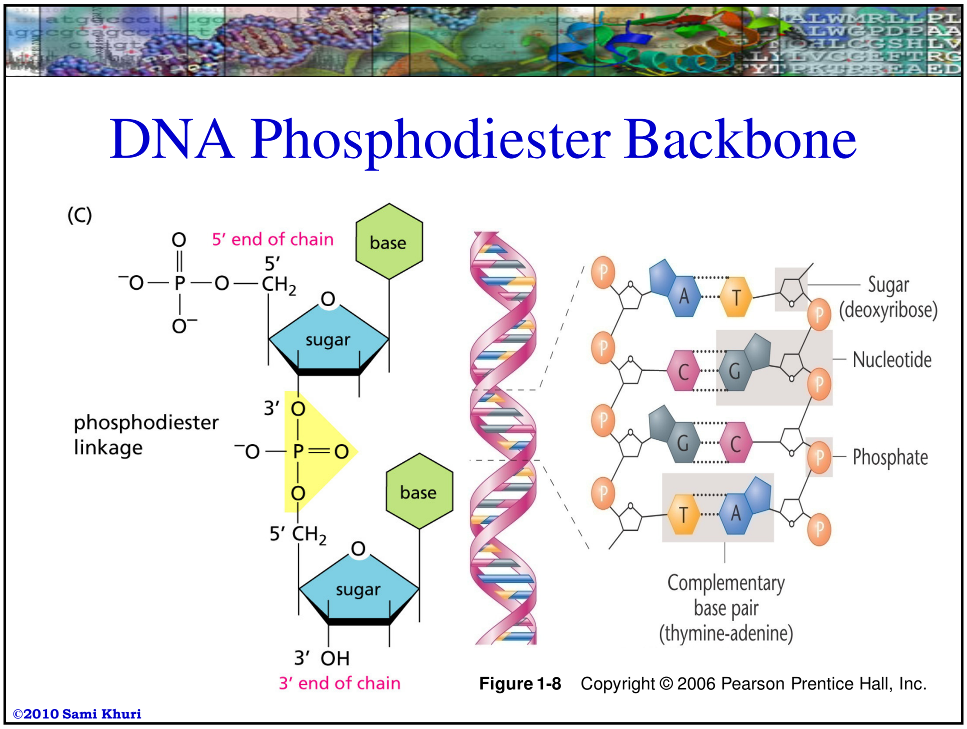
    pyrimidines: **cytosine** (**C**) and **thymine** (**T**).

©2010 Sami Khuri



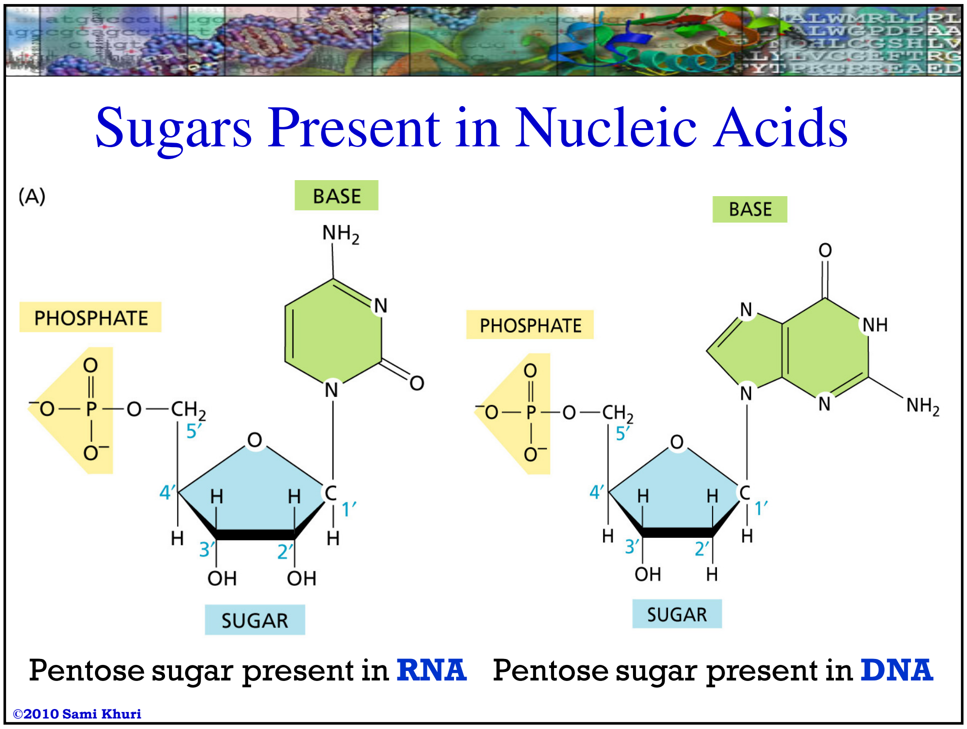
Double Helix

- 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.



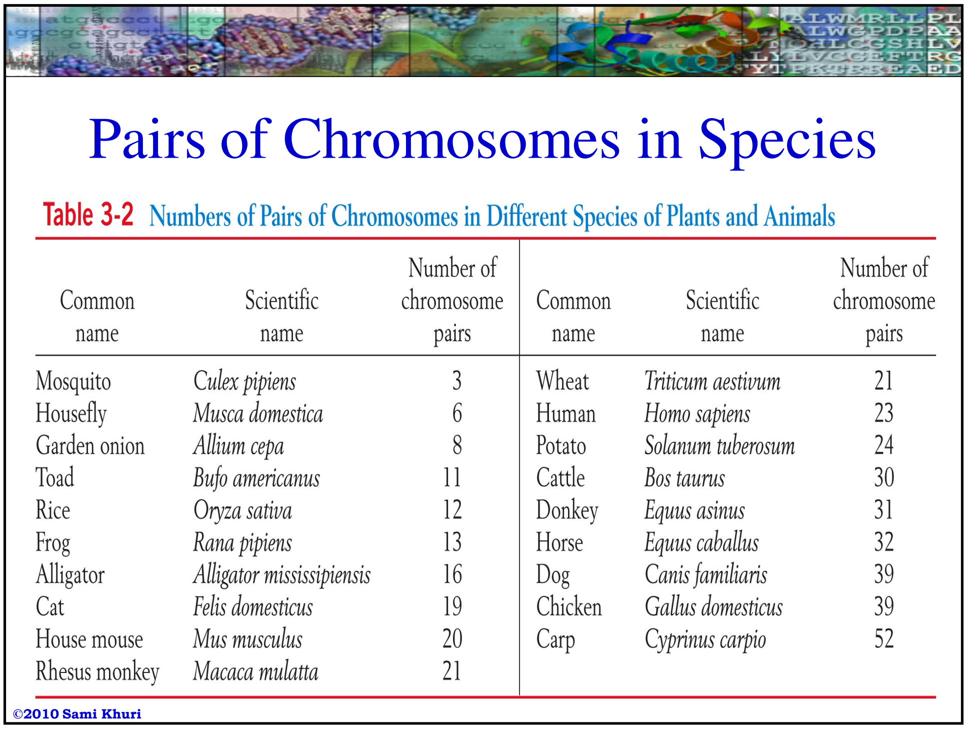
DNA Phosphodiester Backbone

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



Sugars Present in Nucleic Acids

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

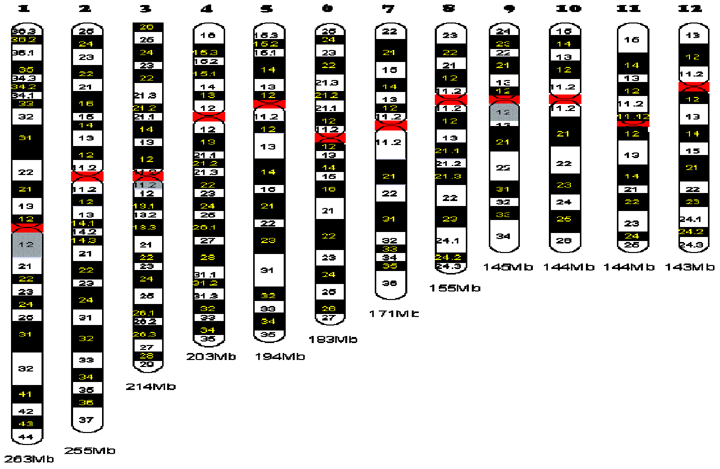


Pairs of Chromosomes in Species

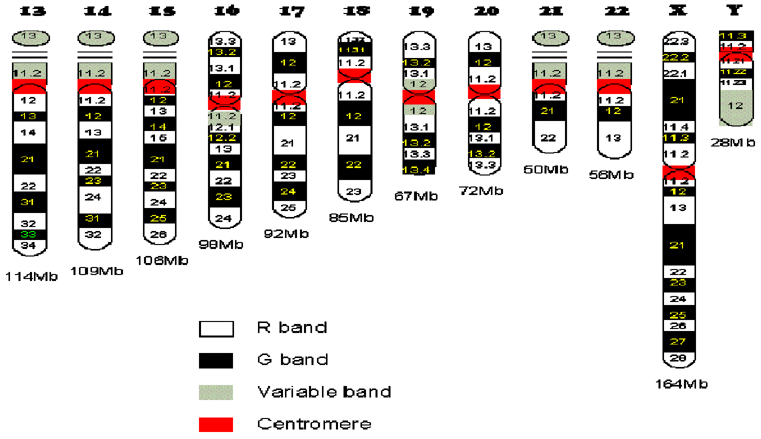
Table 3-2 Numbers of Pairs of Chromosomes in Different Species of Plants and Animals

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

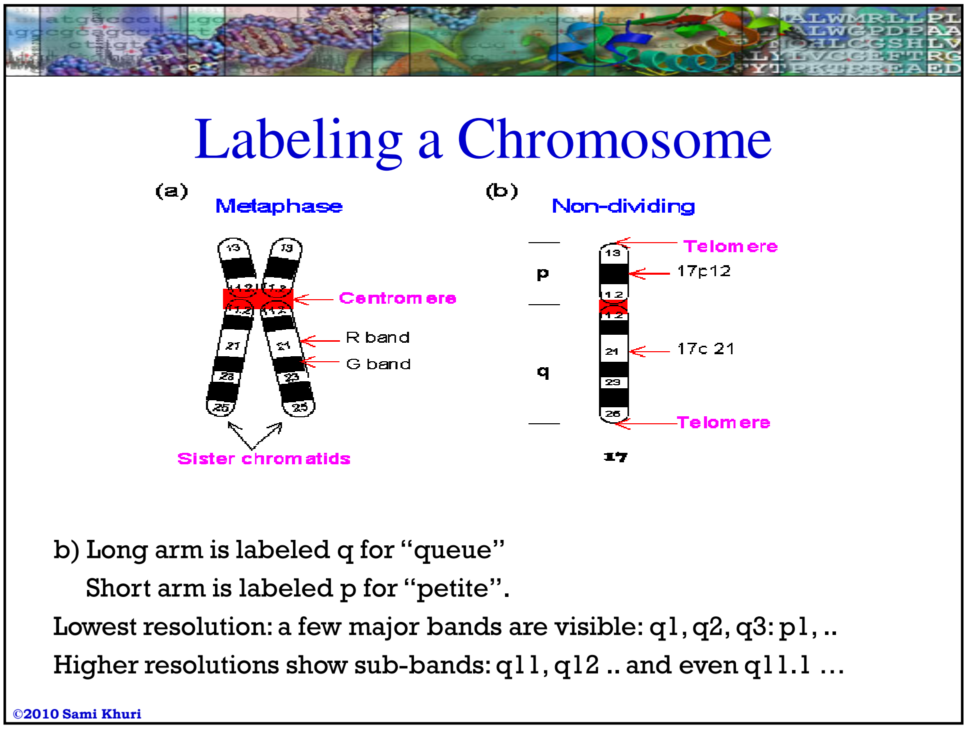


□ R band

■ G band

▨ Variable band

🟥 Centromere

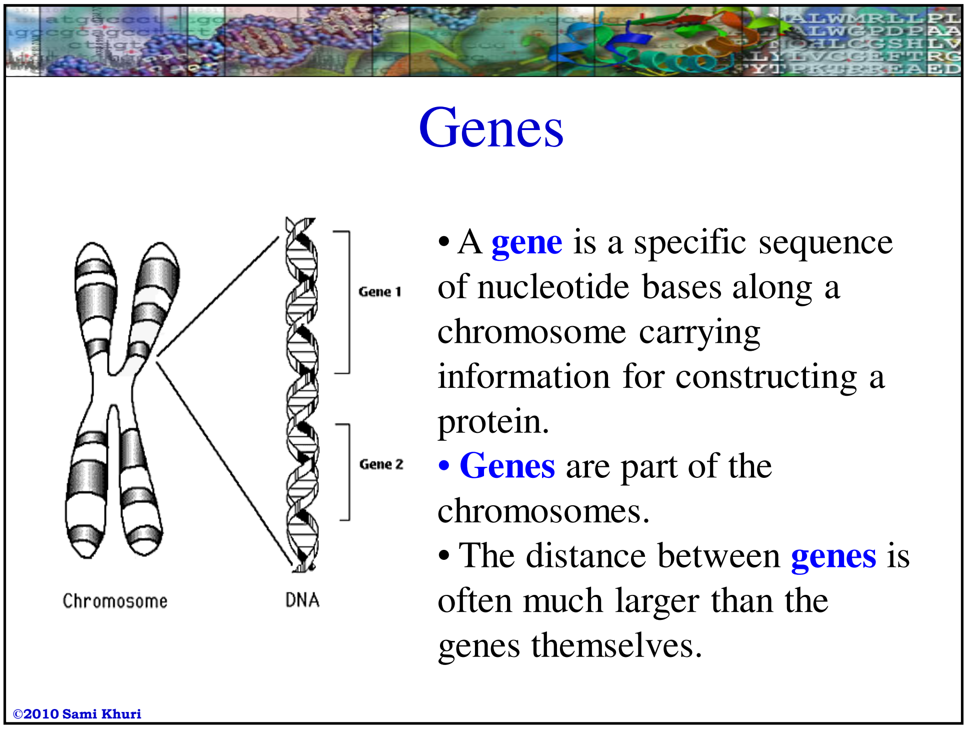


Labeling a Chromosome

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, ..

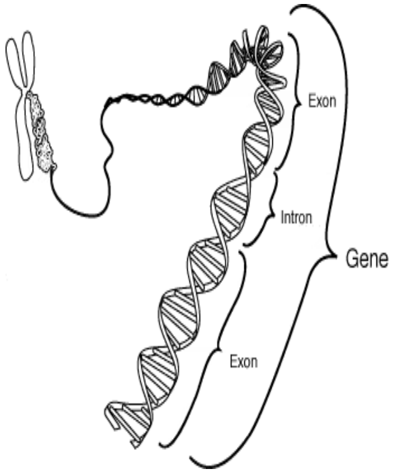
Higher resolutions show sub-bands: q11, q12 .. and even q11.1 …



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

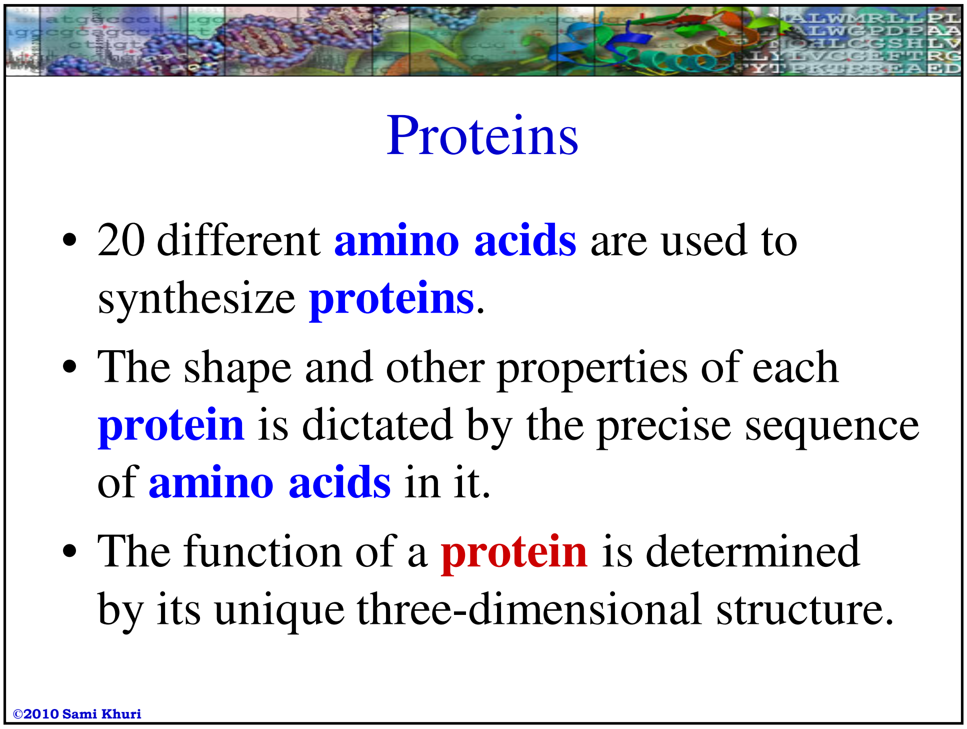
In eukaryotes, genes consist of:

- exons
  protein-coding regions
- introns
  noncoding regions.

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

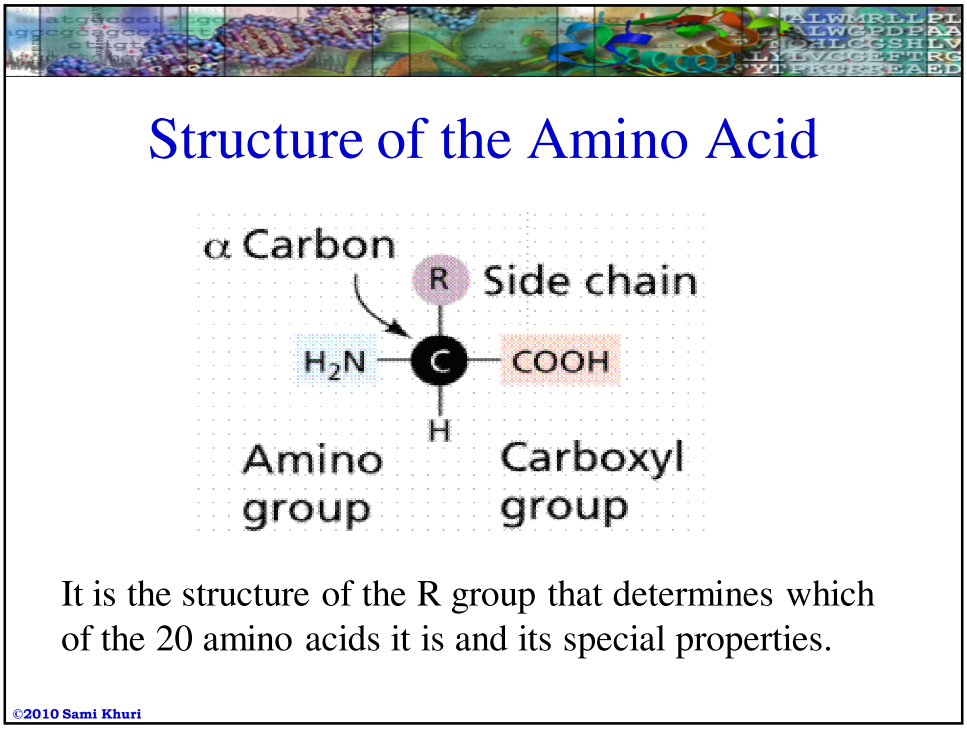
Ribonucleic Acid - RNA

- **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.



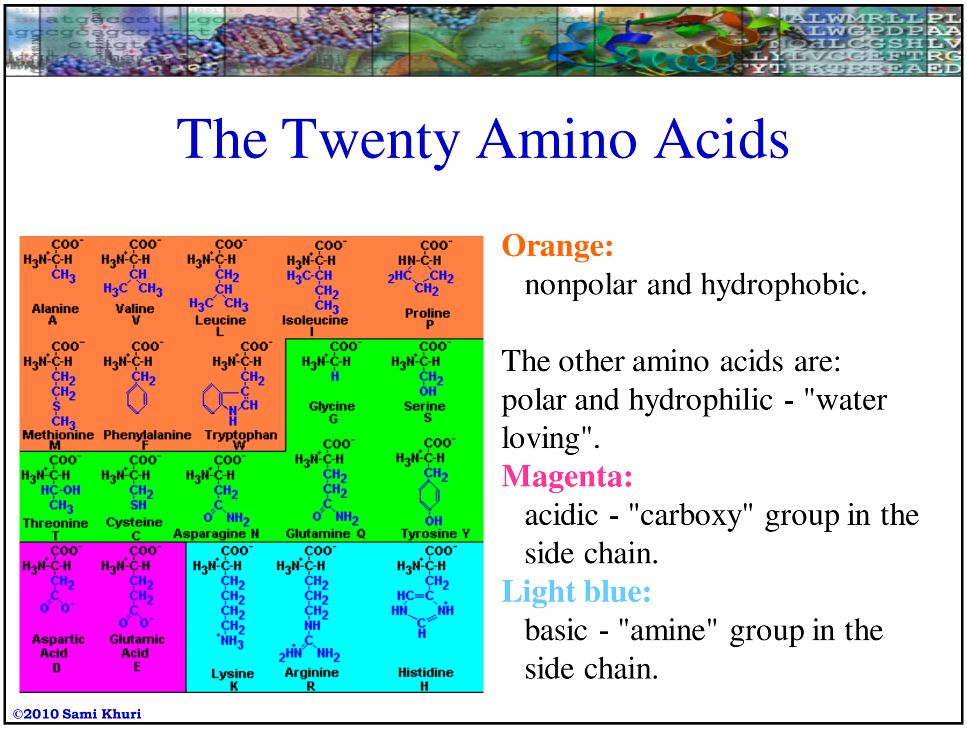
Proteins

- 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.



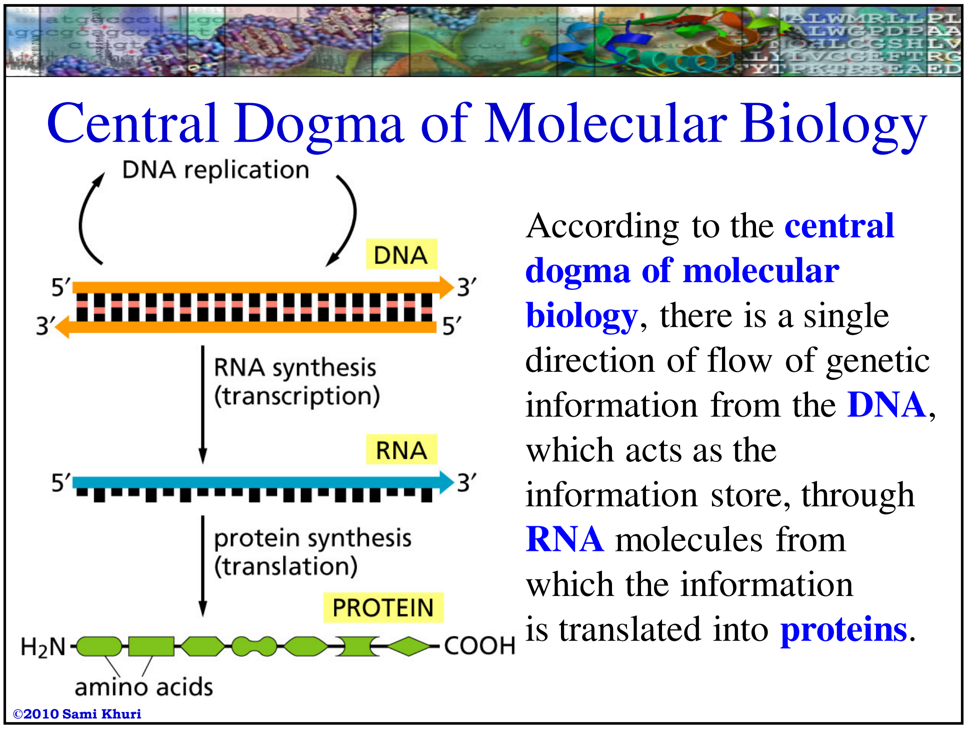
The Twenty Amino Acids

Orange:
nonpolar and hydrophobic.

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

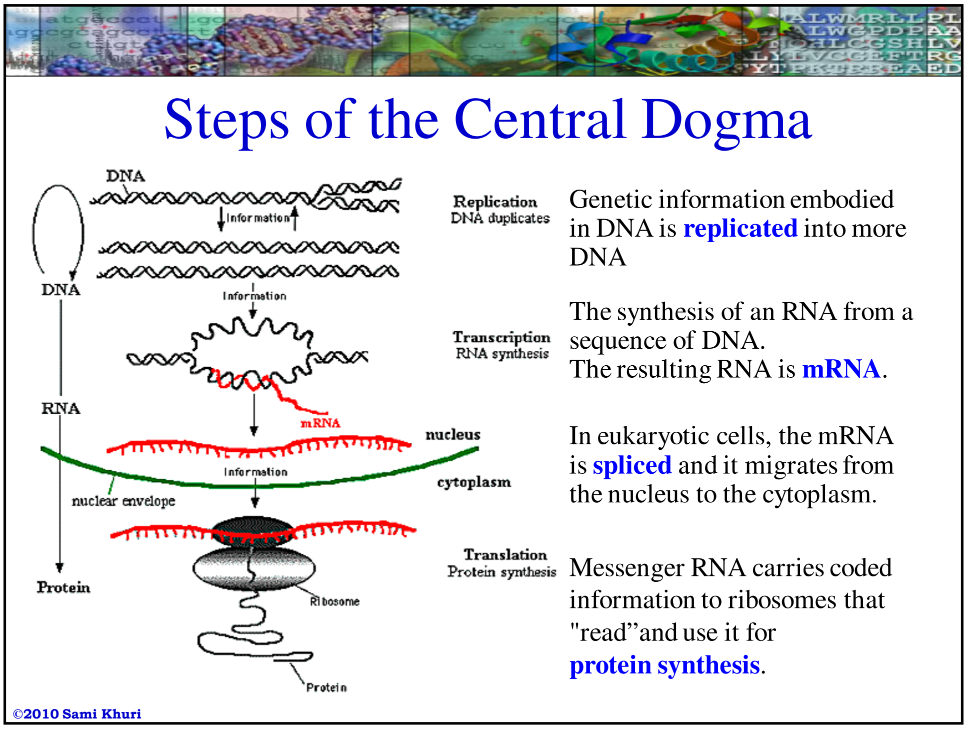
Magenta:
acidic - "carboxy" group in the side chain.

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



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**.



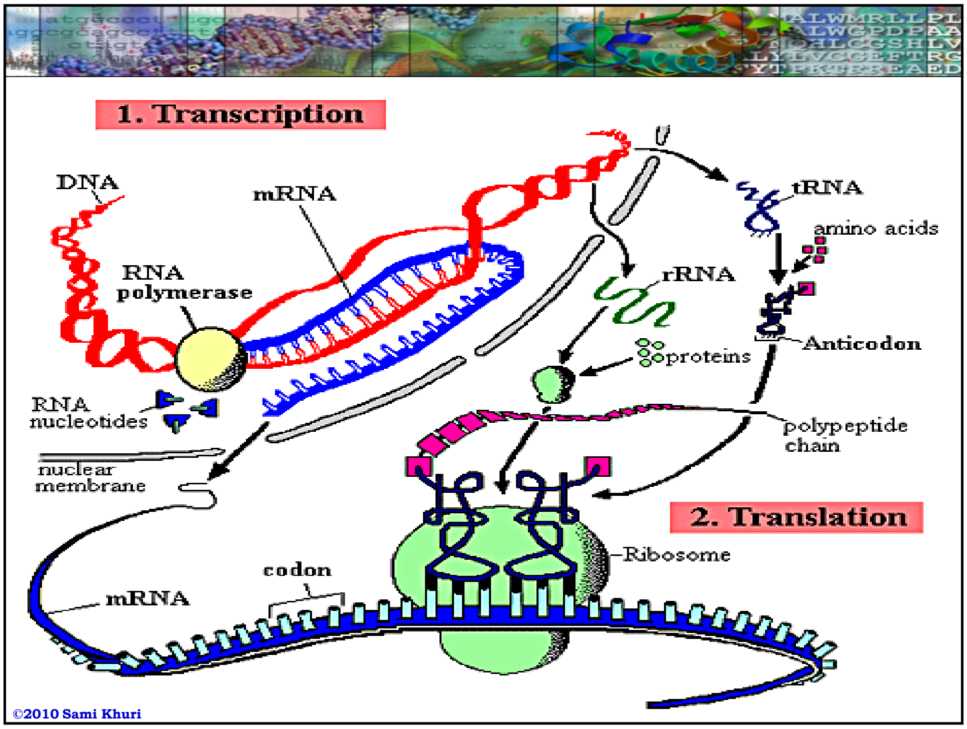
Steps of the Central Dogma

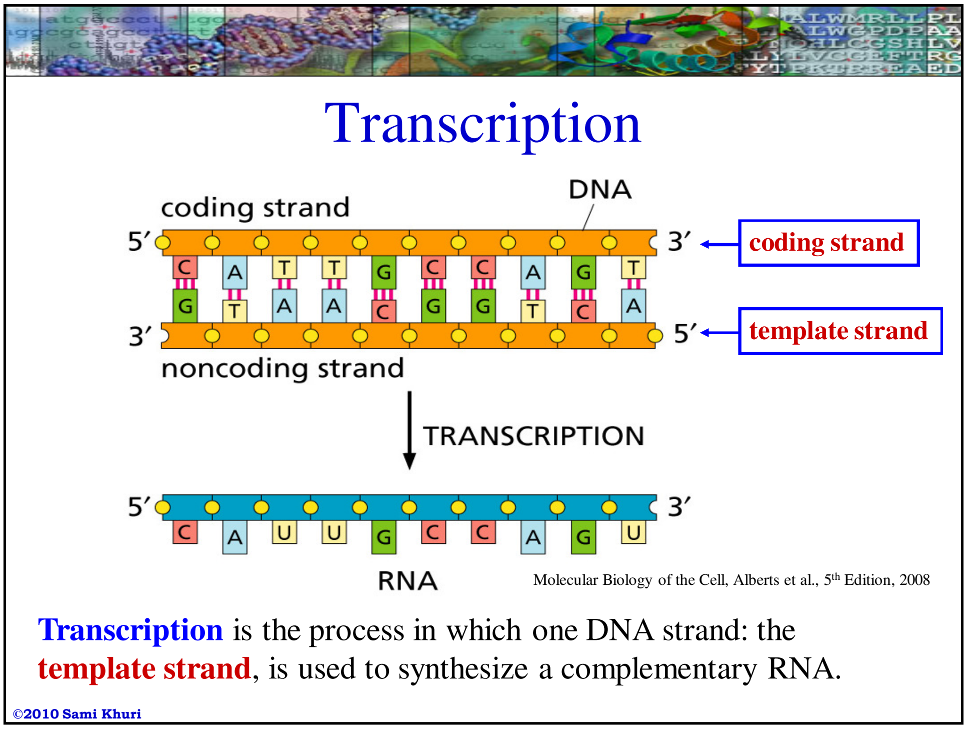
Genetic information embodied in DNA is **replicated** into more DNA

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.

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





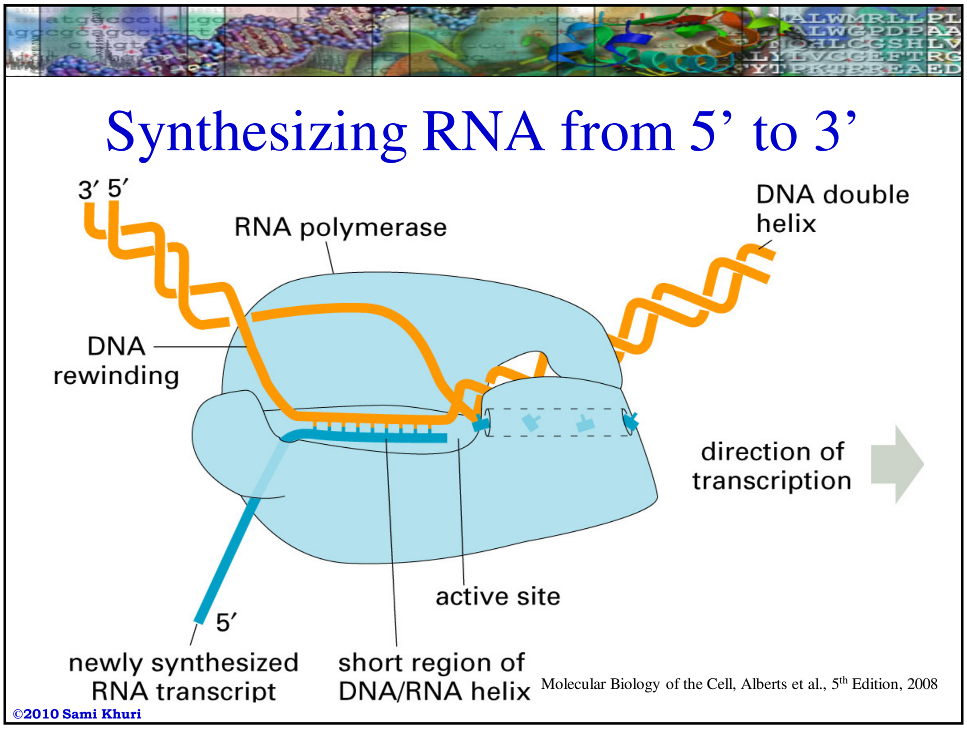
Transcription

coding strand

template strand

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

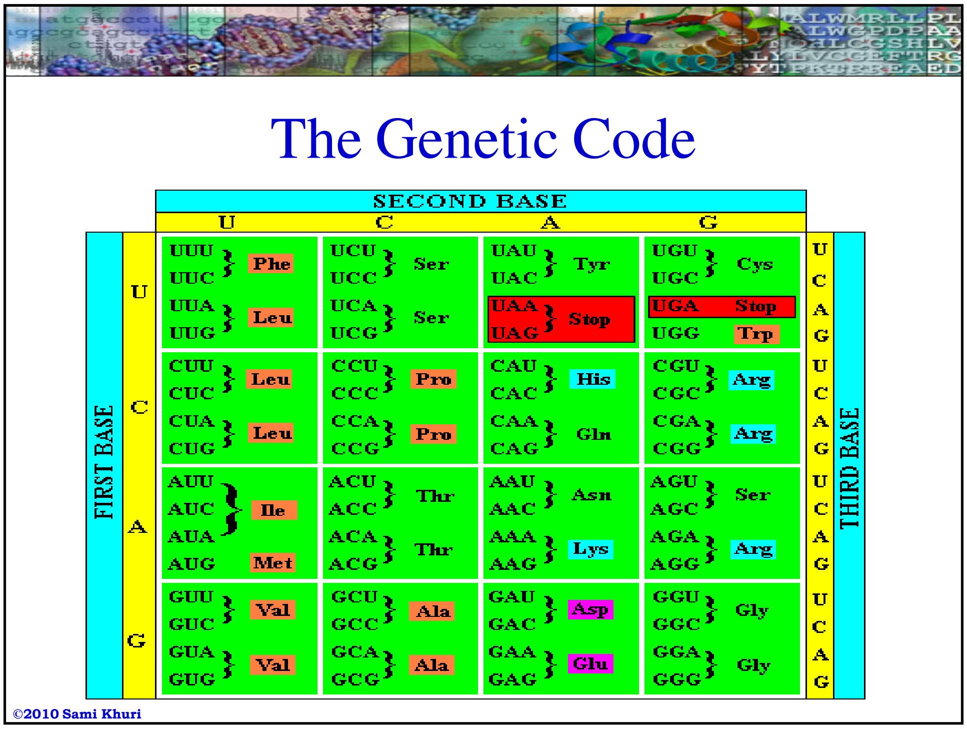
Transcription is the process in which one DNA strand: the **template strand**, is used to synthesize a complementary RNA.



Synthesizing RNA from 5’ to 3’

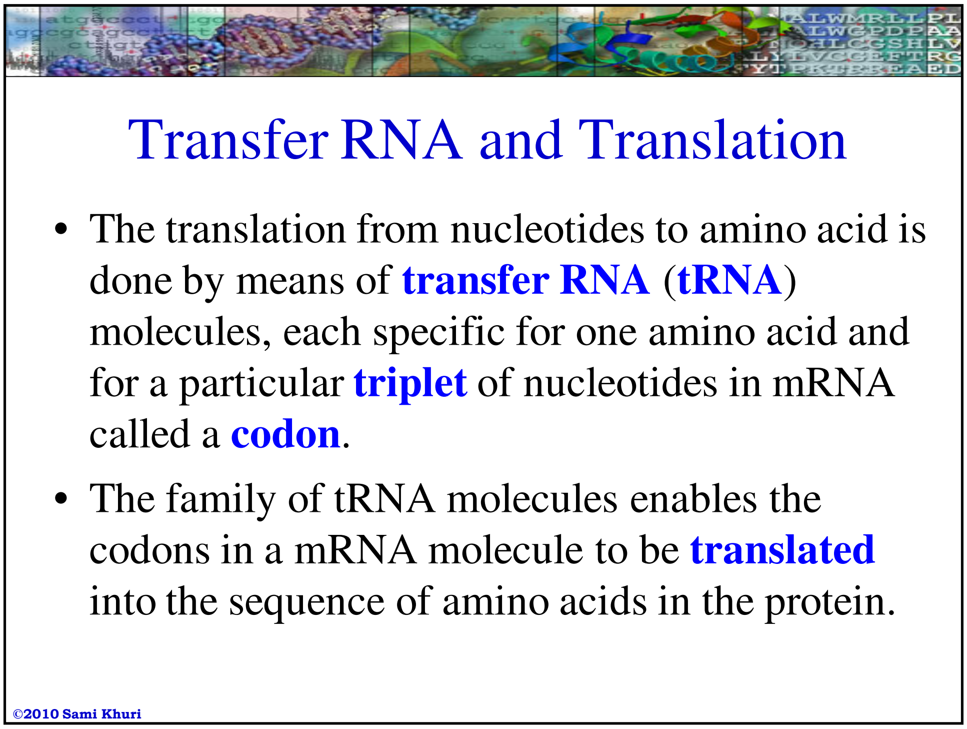
3′ 5′
RNA polymerase
DNA double helix
DNA rewinding
direction of transcription
active site
5′
newly synthesized RNA transcript
short region of DNA/RNA helix

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



The Genetic Code

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

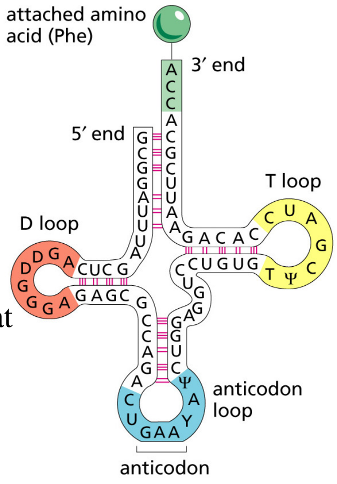
Codons and Anticodons

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.

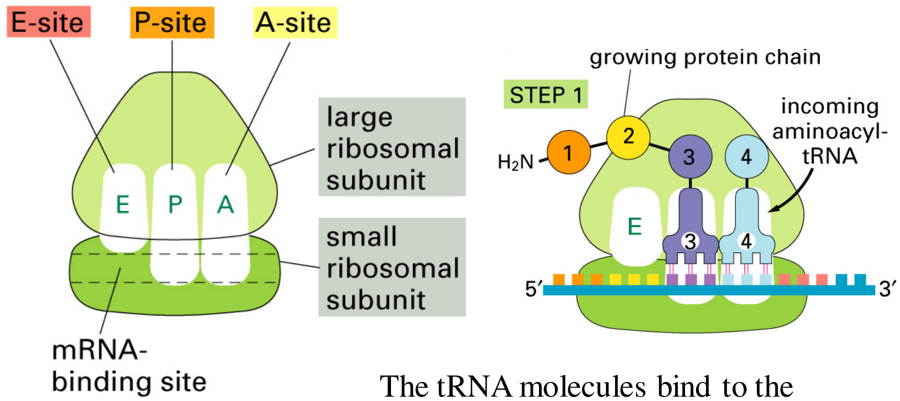
Anticodon: 3' AAG 5'
Codon: 5' UUC 3'



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.

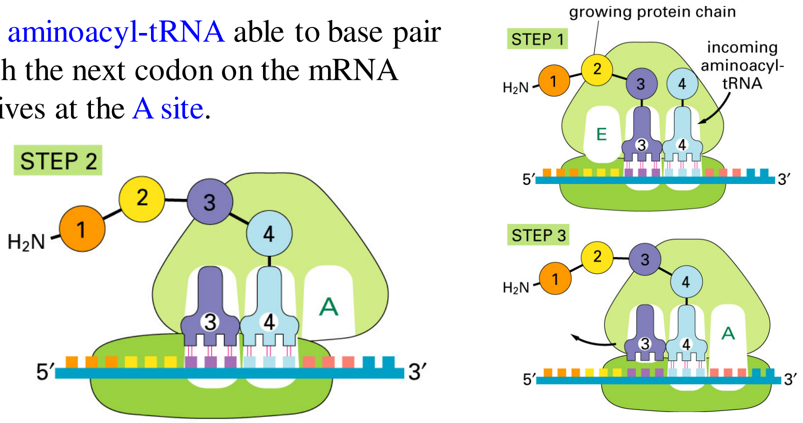
The tRNA molecules bind to the ribosome and are the physical link between the mRNA and the growing protein chain.

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

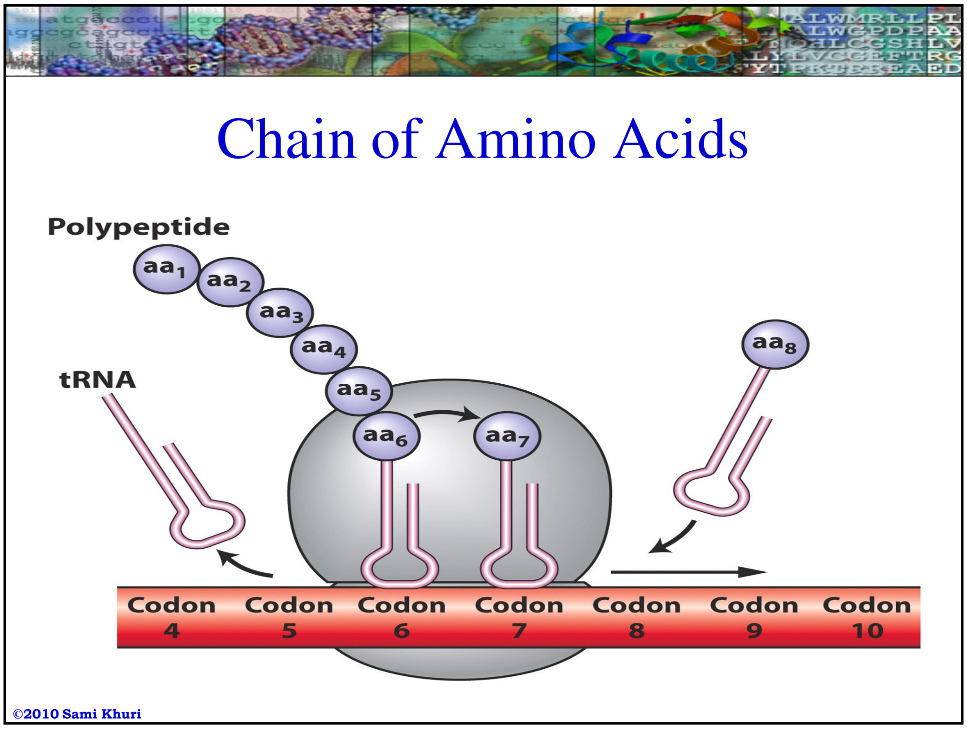
An aminoacyl-tRNA able to base pair with the next codon on the mRNA arrives at the A site.



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

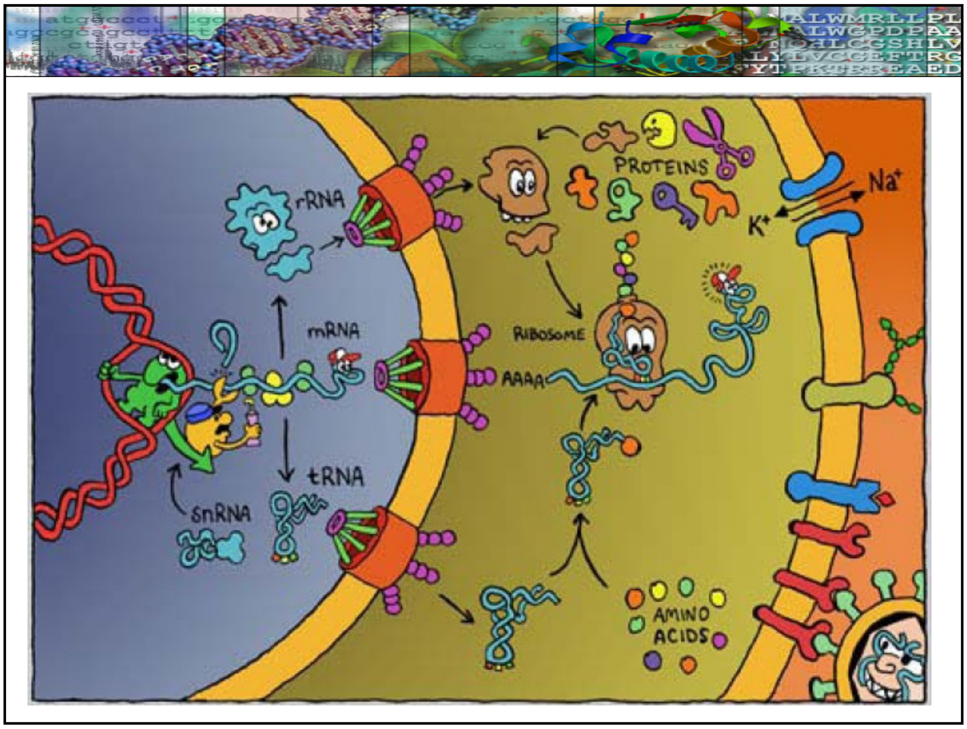
Steps of Translation: Termination

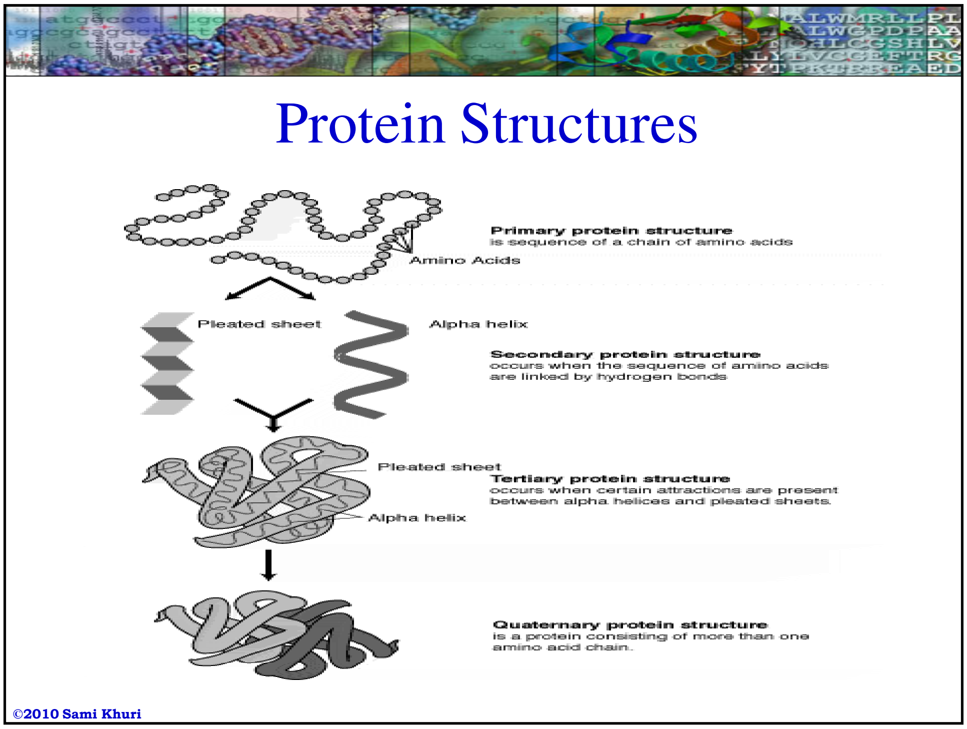
- The end of the message is marked by a STOP codon: UAA, UAG, UGG.
- No tRNA molecules have anticodons for STOP codons. A protein release factor recognizes these codons when they arrive at the A site.
- Binding of this protein releases the polypeptide from the ribosome.
- The ribosome splits into its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



Chain of Amino Acids

©2010 Sami Khuri





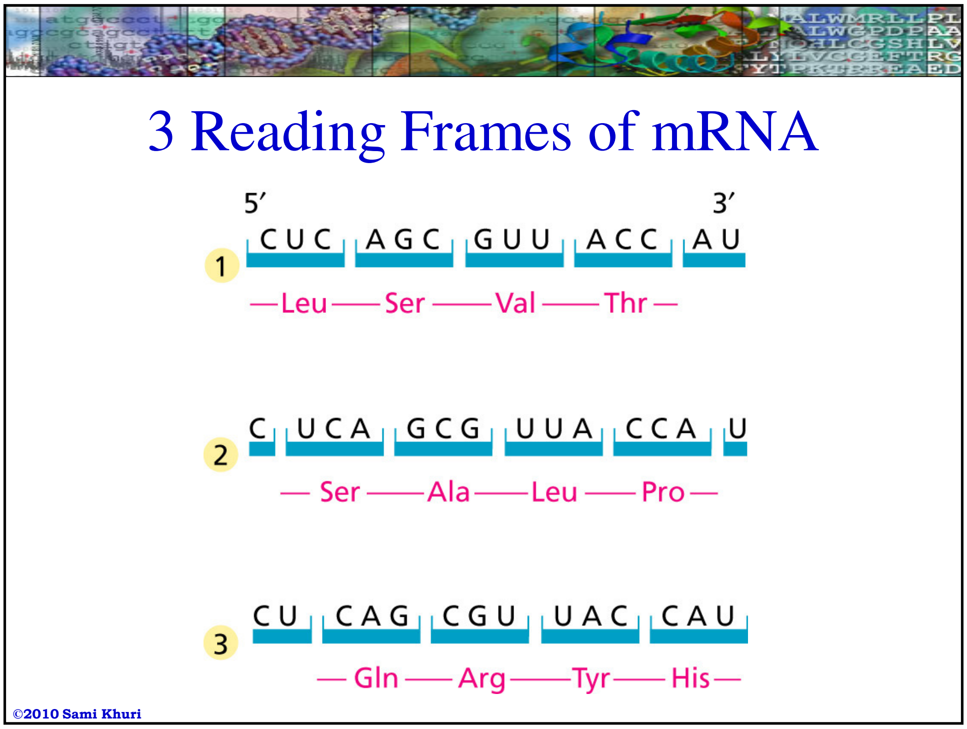
Protein Structures

Primary protein structure
is sequence of a chain of amino acids

Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds

Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.

Quaternary protein structure
is a protein consisting of more than one amino acid chain.



3 Reading Frames of mRNA

5′ 3′

1 CUC AGC GUU ACC AU

—Leu—Ser—Val—Thr—

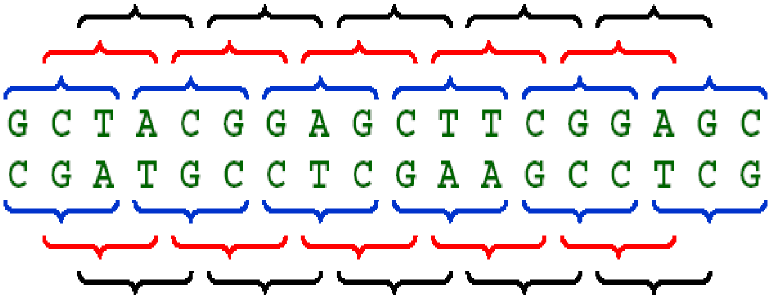
2 C UCA GCG UUA CCA U

—Ser—Ala—Leu—Pro—

3 CU CAG CGU UAC CAU

—Gln—Arg—Tyr—His—

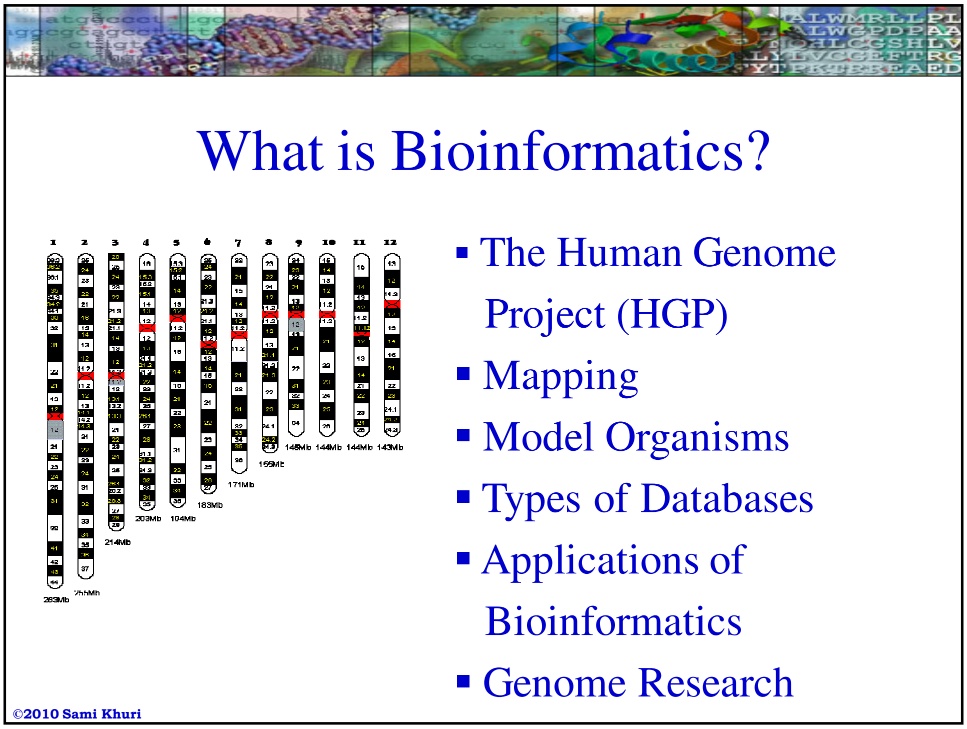
Six Reading Frames



```
G C T A C G G A G C T T C G G A G C
C G A T G C C T C G A A G C C T C G
```

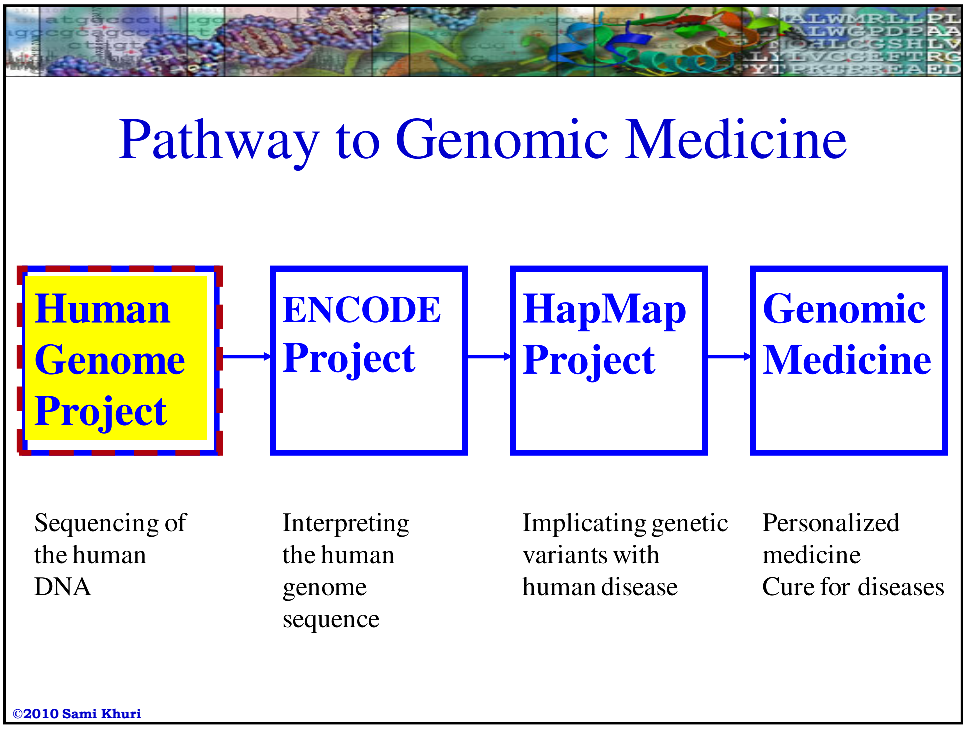
Sequencing SARS

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



What is Bioinformatics?

- The Human Genome Project (HGP)
- Mapping
- Model Organisms
- Types of Databases
- Applications of Bioinformatics
- Genome Research



Pathway to Genomic Medicine

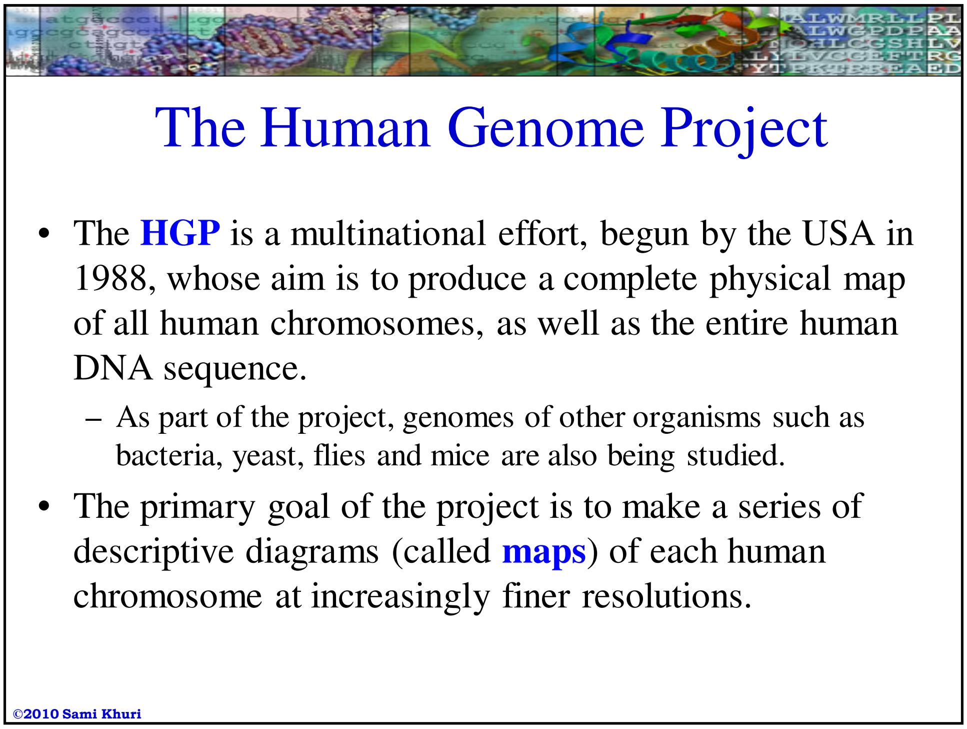
Human Genome Project → ENCODE Project → HapMap Project → Genomic Medicine

Sequencing of the human DNA

Interpreting the human genome sequence

Implicating genetic variants with human disease

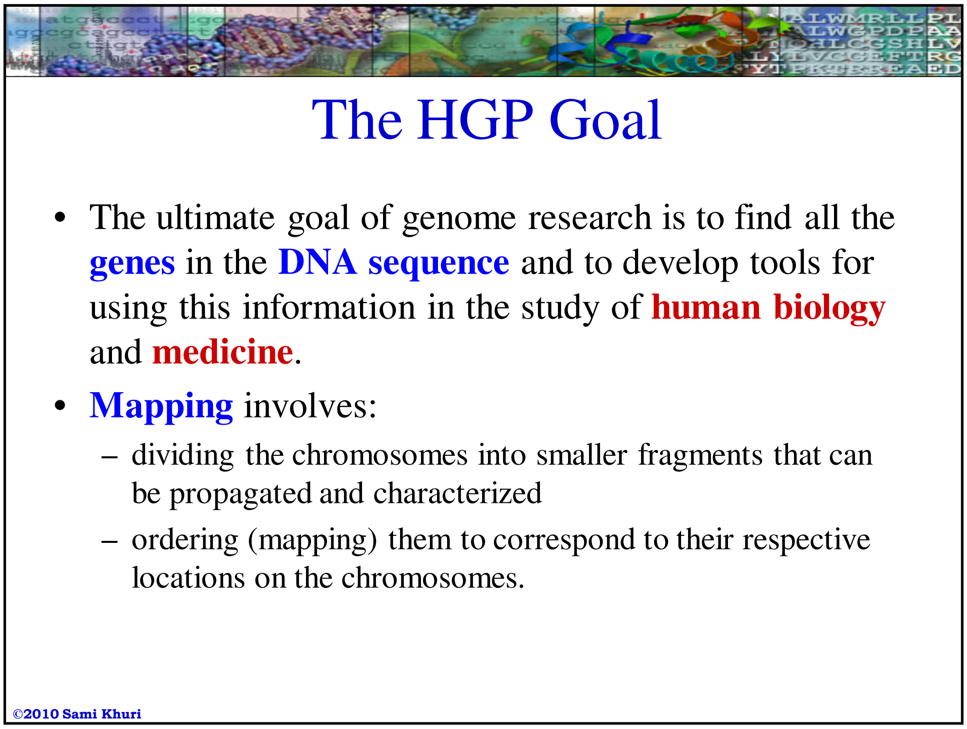
Personalized medicine
Cure for diseases



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.
  - As part of the project, genomes of other organisms such as bacteria, yeast, flies and mice are also being studied.
- The primary goal of the project is to make a series of descriptive diagrams (called **maps**) of each human chromosome at increasingly finer resolutions.

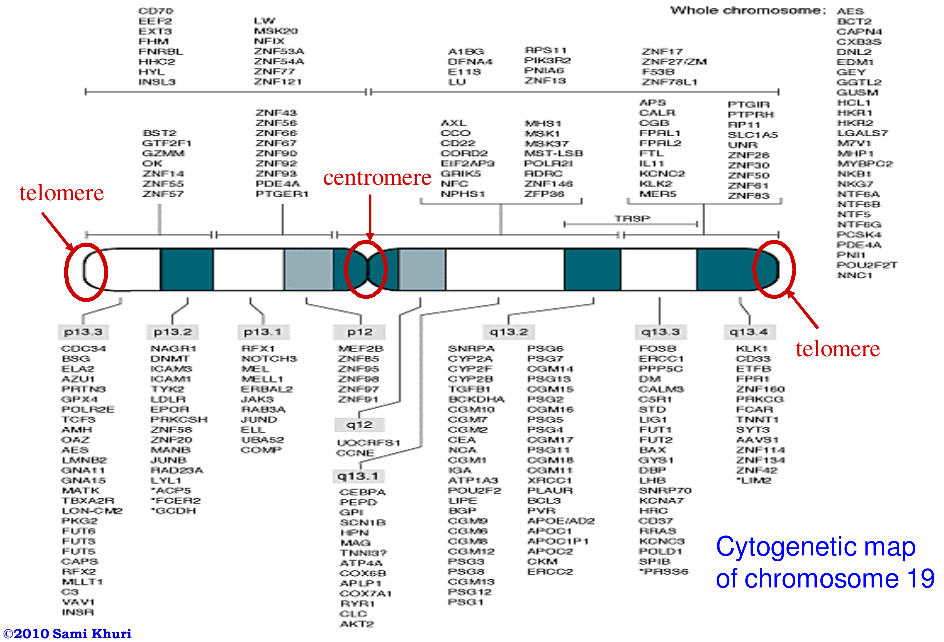
©2010 Sami Khuri



The HGP Goal

- 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**.
- **Mapping** involves:
  - dividing the chromosomes into smaller fragments that can be propagated and characterized
  - ordering (mapping) them to correspond to their respective locations on the chromosomes.

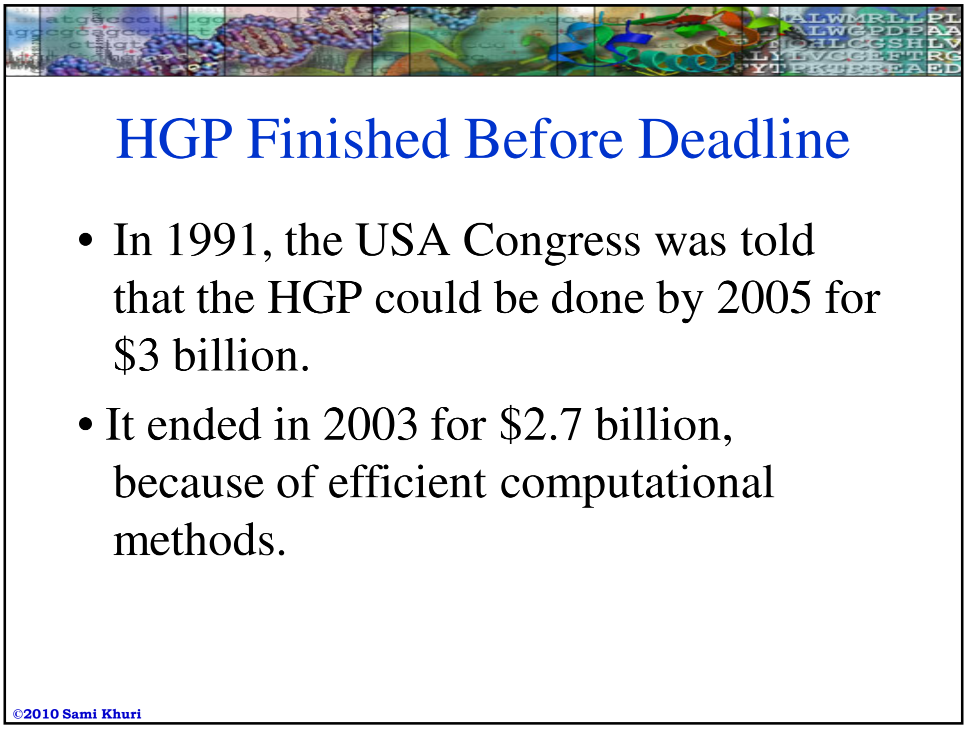
©2010 Sami Khuri



Cytogenetic map of chromosome 19

Goals of the HGP

- To ***identify*** all the approximately 20,000-25,000 genes in human DNA,
- To ***determine*** the sequences of the 3.2 billion chemical base pairs that make up human DNA,
- To ***store*** this information in databases,
- To ***improve*** tools for data analysis,
- To ***address*** the ethical, legal, and social issues (ELSI) that may arise from the project.



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.

©2010 Sami Khuri

Other Species

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

p53 gene
pax6 gene

C elegans

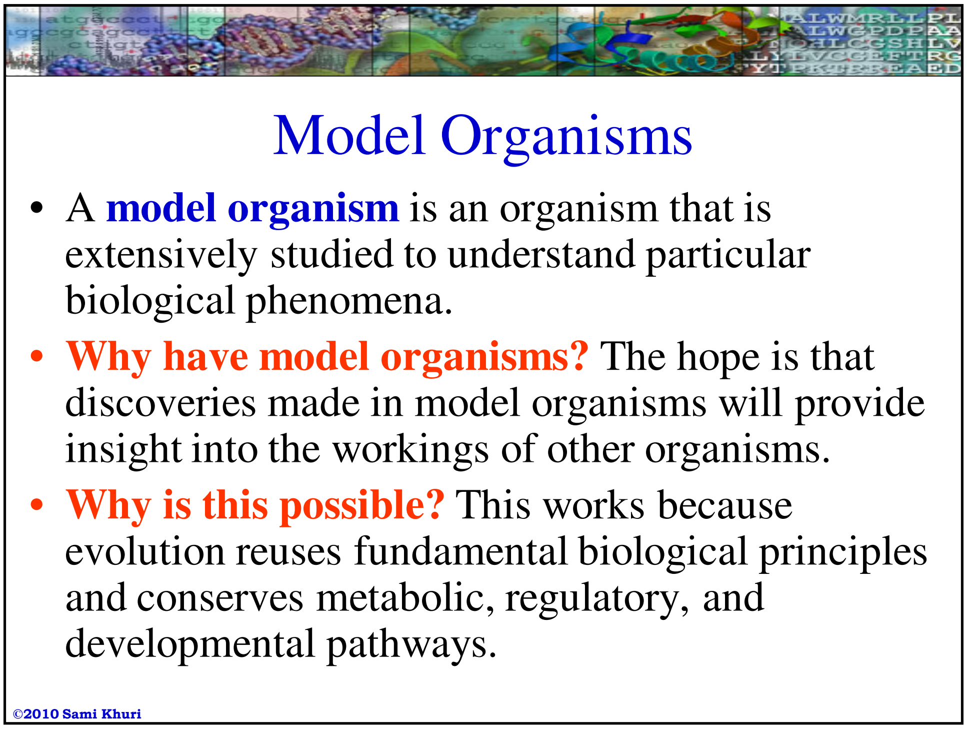
Diabetes

Baker's yeast

DNA repair
Cell division

Chimps are infected with SIV
Very rarely progress to AIDS

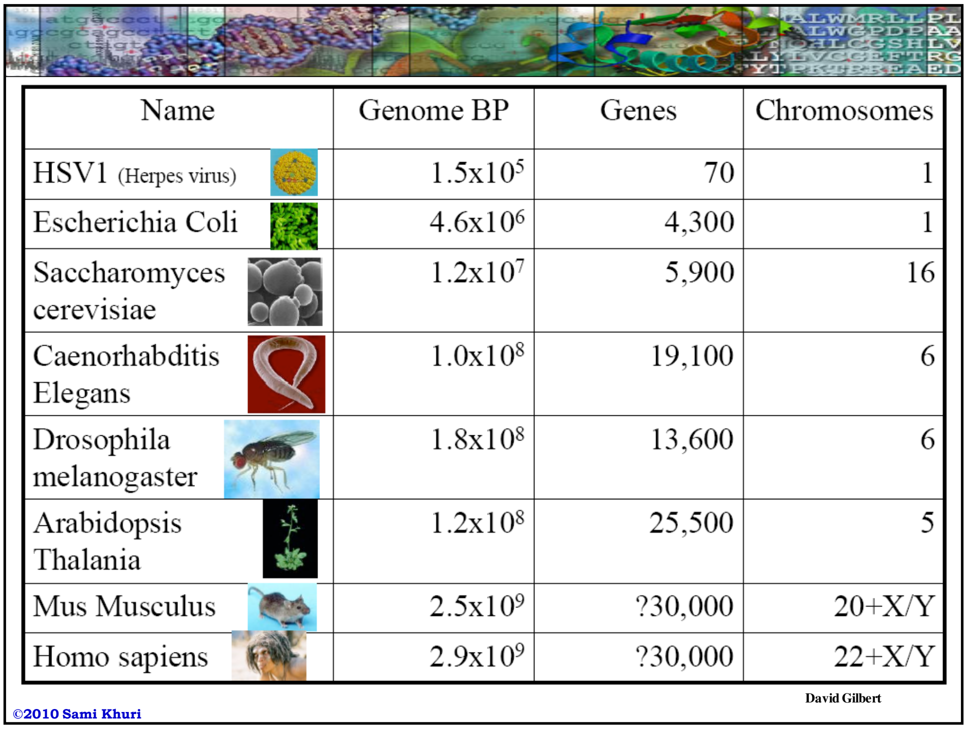
©2010 Sami Khuri



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.

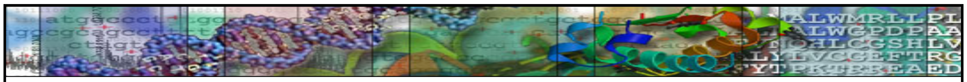
©2010 Sami Khuri



Name	Genome BP	Genes	Chromosomes
HSV1 (Herpes virus)	1.5x10^{5}	70	1
Escherichia Coli	4.6x10^{6}	4,300	1
Saccharomyces cerevisiae	1.2x10^{7}	5,900	16
Caenorhabditis Elegans	1.0x10^{8}	19,100	6
Drosophila melanogaster	1.8x10^{8}	13,600	6
Arabidopsis Thalania	1.2x10^{8}	25,500	5
Mus Musculus	2.5x10^{9}	?30,000	20+X/Y
Homo sapiens	2.9x10^{9}	?30,000	22+X/Y

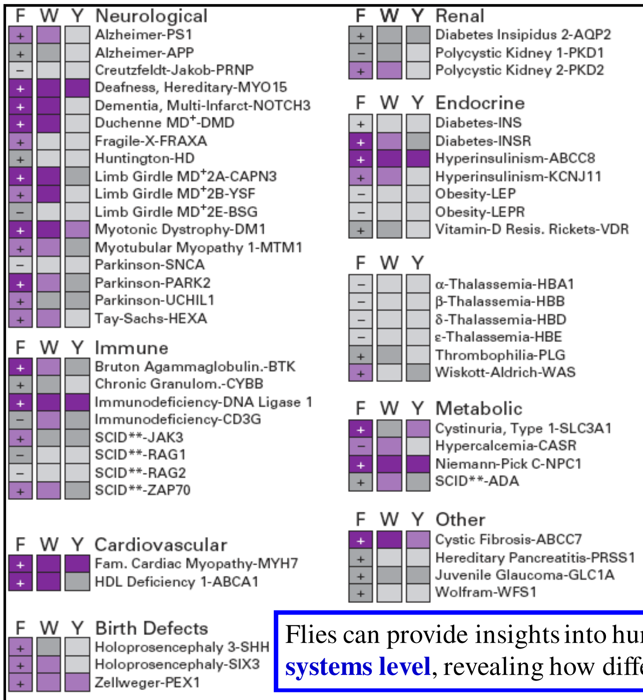
David Gilbert

©2010 Sami Khuri



Studying Human Diseases

Organism	Human Diseases
E. coli	DNA repair; colon cancer and other cancers
Yeast	Cell cycle; cancer, Werner syndrome
Drosophila	Cell signaling; cancer
C. elegans	Cell signaling; diabetes
Zebrafish	Developmental pathways; cardiovascular disease
Mouse	Gene expression; Lesch-Nyhan disease, cystic fibrosis, fragile-X syndrome, and many other diseases



F	W	Y	Neurological
▦ +	▦	□	Alzheimer-PS1
▒ +	▒	□	Alzheimer-APP
□ −	□	□	Creutzfeldt-Jakob-PRNP
■ +	■	■	Deafness, Hereditary-MYO15
■ +	■	□	Dementia, Multi-Infarct-NOTCH3
■ +	■	□	Duchenne MD$^+$-DMD
▦ +	□	□	Fragile-X-FRAXA
▒ +	□	□	Huntington-HD
■ +	■	▒	Limb Girdle MD$^+$2A-CAPN3
▦ +	■	□	Limb Girdle MD$^+$2B-YSF
▒ −	□	□	Limb Girdle MD$^+$2E-BSG
■ +	■	▦	Myotonic Dystrophy-DM1
▦ +	▦	▒	Myotubular Myopathy 1-MTM1
□ −	□	□	Parkinson-SNCA
■ +	▦	▒	Parkinson-PARK2
▦ +	▒	▒	Parkinson-UCHIL1
▦ +	▦	□	Tay-Sachs-HEXA

F	W	Y	Immune
■ +	▦	▒	Bruton Agammaglobulin.-BTK
▒ +	▒	□	Chronic Granulom.-CYBB
■ +	■	■	Immunodeficiency-DNA Ligase 1
▒ −	▦	▒	Immunodeficiency-CD3G
▦ +	▒	▒	SCID**-JAK3
▒ −	□	□	SCID**-RAG1
□ −	□	□	SCID**-RAG2
▦ +	▦	▒	SCID**-ZAP70

F	W	Y	Cardiovascular
■ +	■	■	Fam. Cardiac Myopathy-MYH7
■ +	■	▒	HDL Deficiency 1-ABCA1

F	W	Y	Birth Defects
▦ +	▒	□	Holoprosencephaly 3-SHH
▦ +	▦	□	Holoprosencephaly-SIX3
▦ +	▦	▦	Zellweger-PEX1

F	W	Y	Renal
▒ +	▒	▒	Diabetes Insipidus 2-AQP2
▒ −	▒	□	Polycystic Kidney 1-PKD1
▦ +	▦	□	Polycystic Kidney 2-PKD2

F	W	Y	Endocrine
□ +	□	□	Diabetes-INS
■ +	▦	▒	Diabetes-INSR
■ +	■	■	Hyperinsulinism-ABCC8
▦ +	▦	□	Hyperinsulinism-KCNJ11
□ −	□	□	Obesity-LEP
□ −	□	□	Obesity-LEPR
▒ +	▒	□	Vitamin-D Resis. Rickets-VDR

F	W	Y	
□ −	□	□	α-Thalassemia-HBA1
□ −	□	□	β-Thalassemia-HBB
□ −	□	□	δ-Thalassemia-HBD
□ −	□	□	ε-Thalassemia-HBE
▒ +	▒	□	Thrombophilia-PLG
▦ +	□	▒	Wiskott-Aldrich-WAS

F	W	Y	Metabolic
■ +	▒	▦	Cystinuria, Type 1-SLC3A1
▦ −	▦	□	Hypercalcemia-CASR
■ +	■	■	Niemann-Pick C-NPC1
▒ +	▦	▒	SCID**-ADA

F	W	Y	Other
■ +	■	▦	Cystic Fibrosis-ABCC7
▒ +	□	□	Hereditary Pancreatitis-PRSS1
▒ +	▒	▒	Juvenile Glaucoma-GLC1A
▒ +	□	□	Wolfram-WFS1

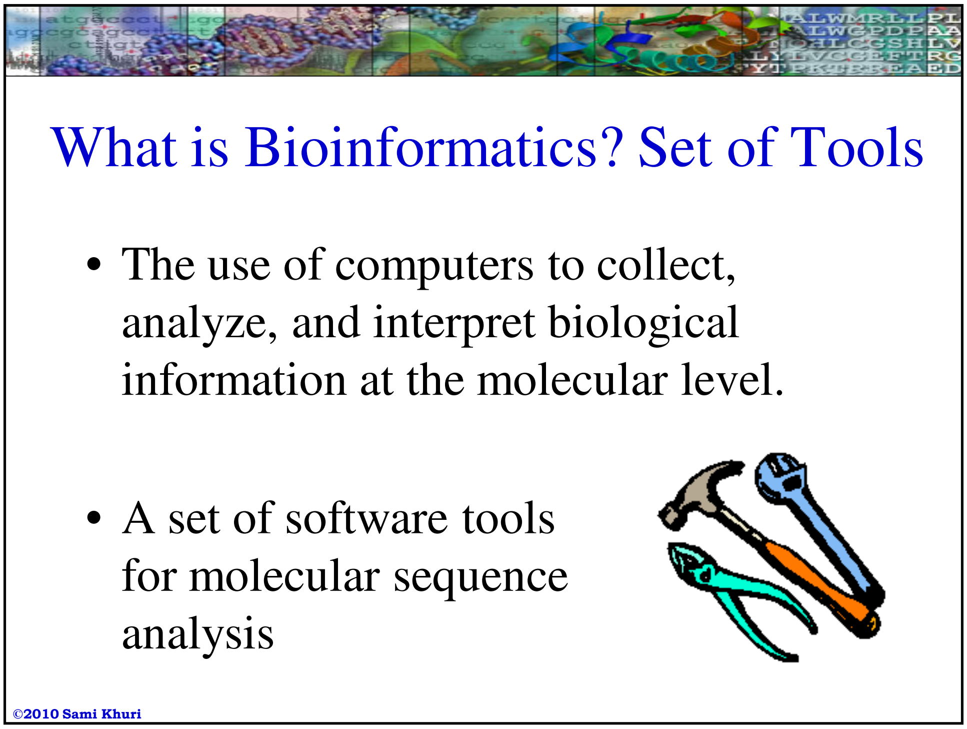
■ E-values <1x10^{-100}
▦ E-values of 1x10^{-40} to 1x10^{-100}
▒ E-values of 1x10^{-6} to 1x10^{-40}
□ E-values >1x10^{-6}

Flies have **orthologs** to humans disease-causing genes in categories such as:

- neurological
- renal
- immunological
- endocrine
- cardiovascular
- metabolic
- blood-vessel and
- cancerous disorders

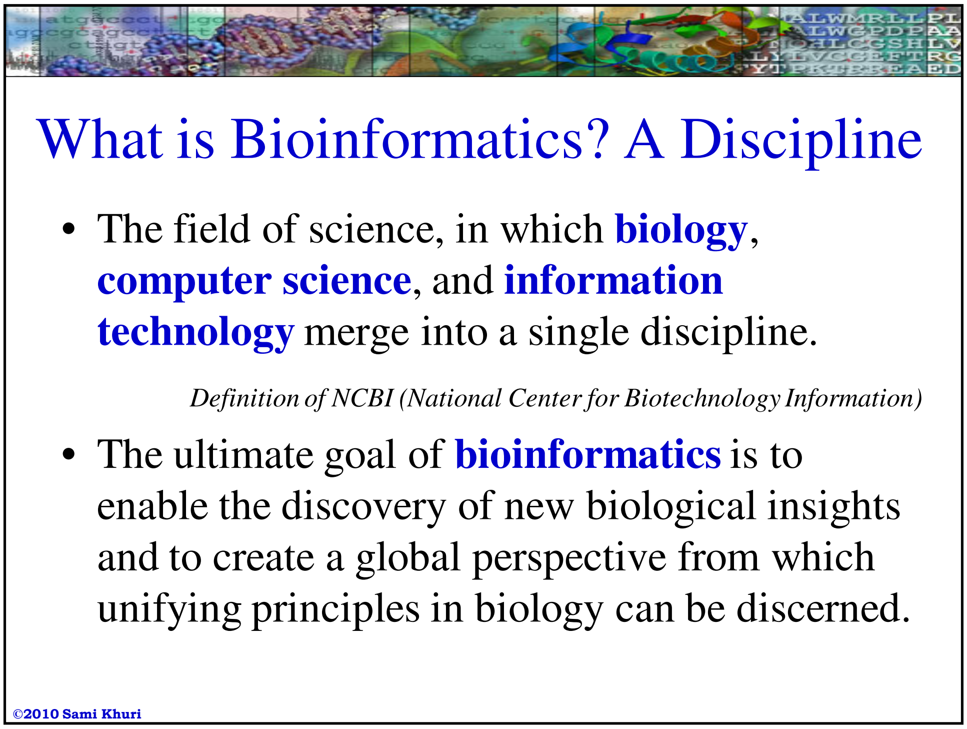
Flies can provide insights into human disease at the **systems level**, revealing how different genes interact in vivo

Discovering Genomics, Campbell, 2007



What is Bioinformatics? Set of Tools

- The use of computers to collect, analyze, and interpret biological information at the molecular level.
- A set of software tools for molecular sequence analysis

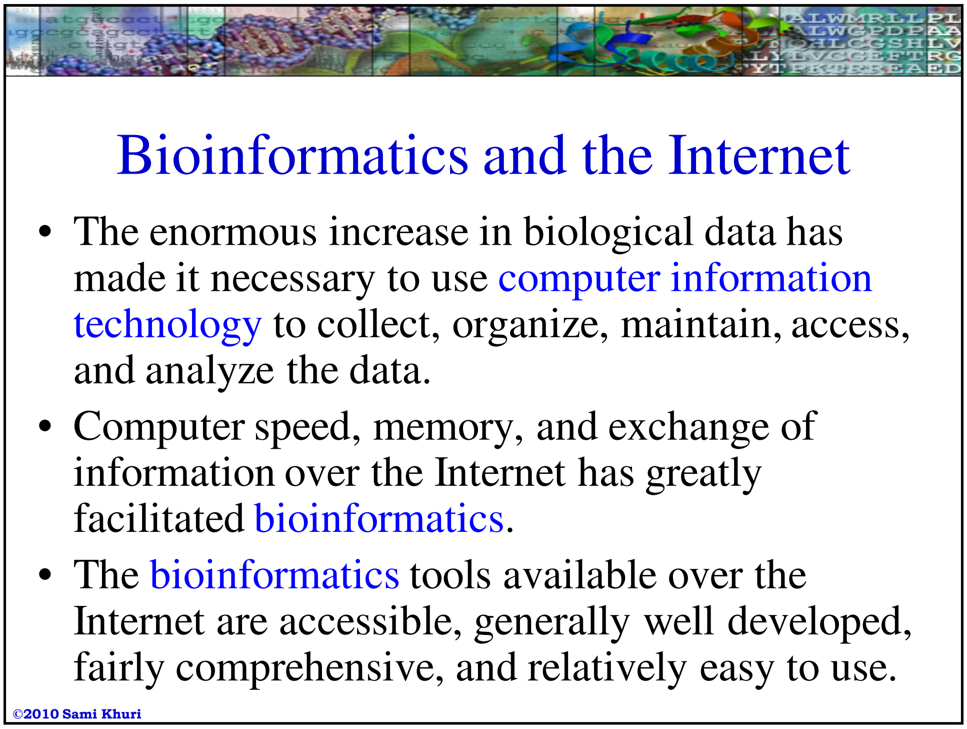


What is Bioinformatics? A Discipline

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

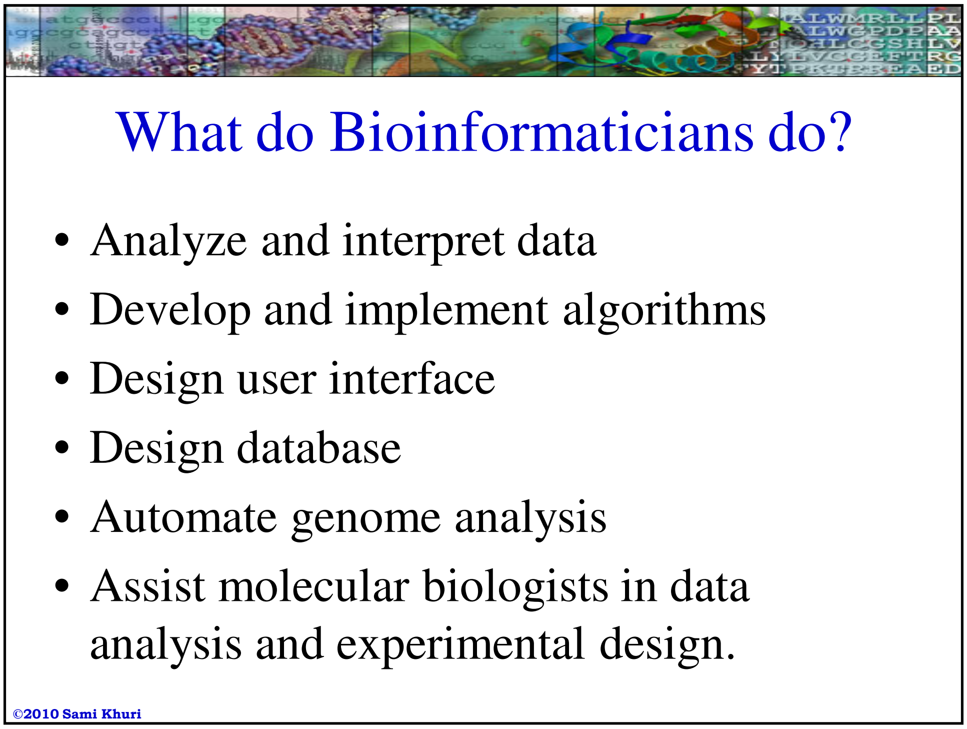
Definition of NCBI (National Center for Biotechnology Information)

- 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.



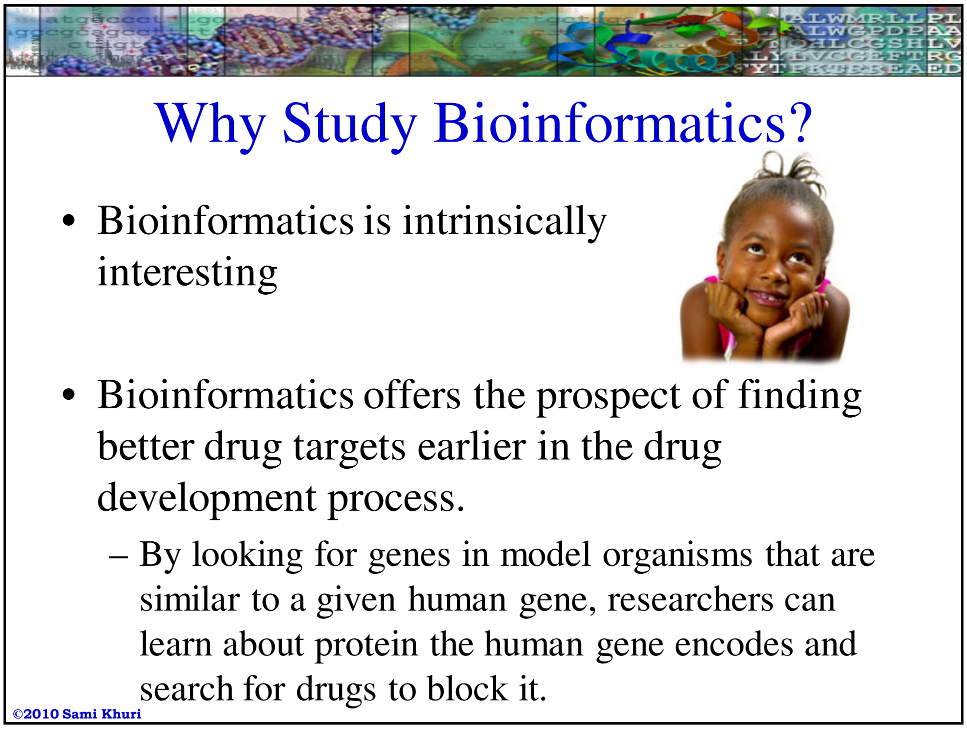
Bioinformatics and the Internet

- The enormous increase in biological data has made it necessary to use computer information technology to collect, organize, maintain, access, and analyze the data.
- Computer speed, memory, and exchange of information over the Internet has greatly facilitated bioinformatics.
- The bioinformatics tools available over the Internet are accessible, generally well developed, fairly comprehensive, and relatively easy to use.



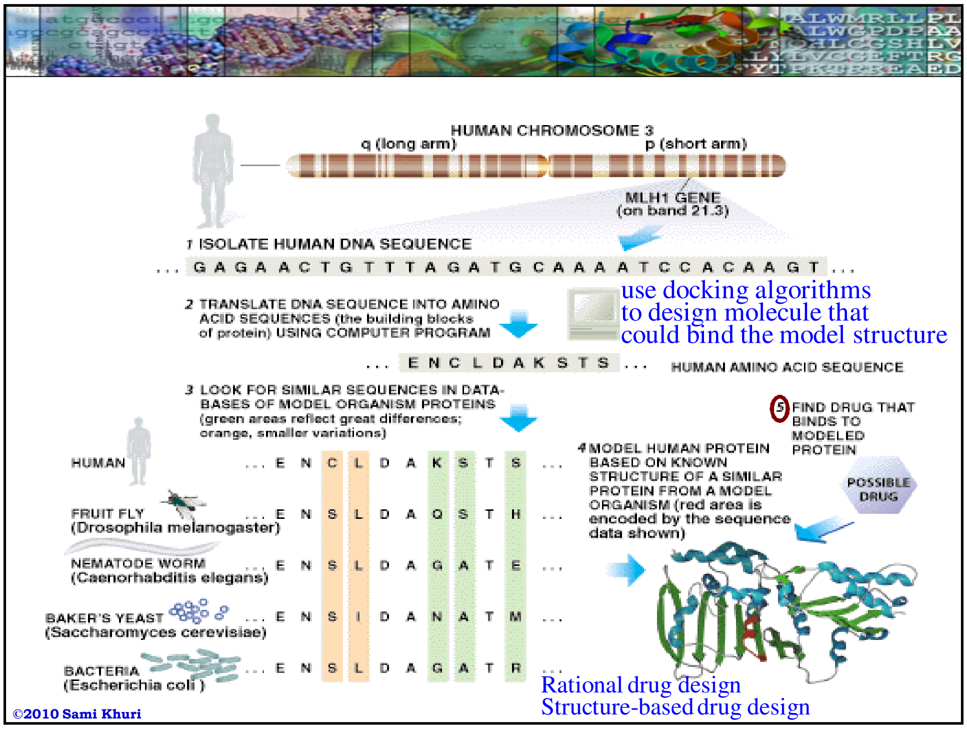
What do Bioinformaticians do?

- Analyze and interpret data
- Develop and implement algorithms
- Design user interface
- Design database
- Automate genome analysis
- Assist molecular biologists in data analysis and experimental design.



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 protein the human gene encodes and search for drugs to block it.



HUMAN CHROMOSOME 3

q (long arm)

p (short arm)

MLH1 GENE
(on band 21.3)

1 ISOLATE HUMAN DNA SEQUENCE

. . . G A G A A C T G T T T A G A T G C A A A A T C C A C A A G T . . .

2 TRANSLATE DNA SEQUENCE INTO AMINO ACID SEQUENCES (the building blocks of protein) USING COMPUTER PROGRAM

use docking algorithms
to design molecule that
could bind the model structure

. . . E N C L D A K S T S . . . HUMAN AMINO ACID SEQUENCE

3 LOOK FOR SIMILAR SEQUENCES IN DATABASES OF MODEL ORGANISM PROTEINS (green areas reflect great differences; orange, smaller variations)

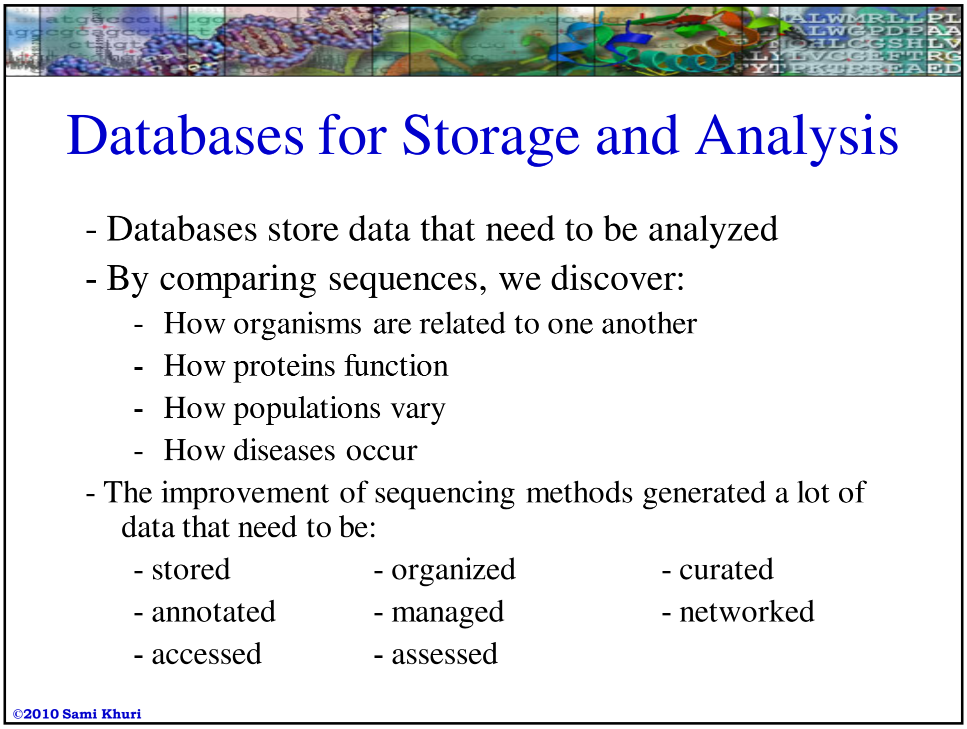
HUMAN	. . .	E	N	C	L	D	A	K	S	T	S	. . .
FRUIT FLY (Drosophila melanogaster)	. . .	E	N	S	L	D	A	Q	S	T	H	. . .
NEMATODE WORM (Caenorhabditis elegans)	. . .	E	N	S	L	D	A	G	A	T	E	. . .
BAKER'S YEAST (Saccharomyces cerevisiae)	. . .	E	N	S	I	D	A	N	A	T	M	. . .
BACTERIA (Escherichia coli)	. . .	E	N	S	L	D	A	G	A	T	R	. . .

4 MODEL HUMAN PROTEIN BASED ON KNOWN STRUCTURE OF A SIMILAR PROTEIN FROM A MODEL ORGANISM (red area is encoded by the sequence data shown)

5 FIND DRUG THAT BINDS TO MODELED PROTEIN

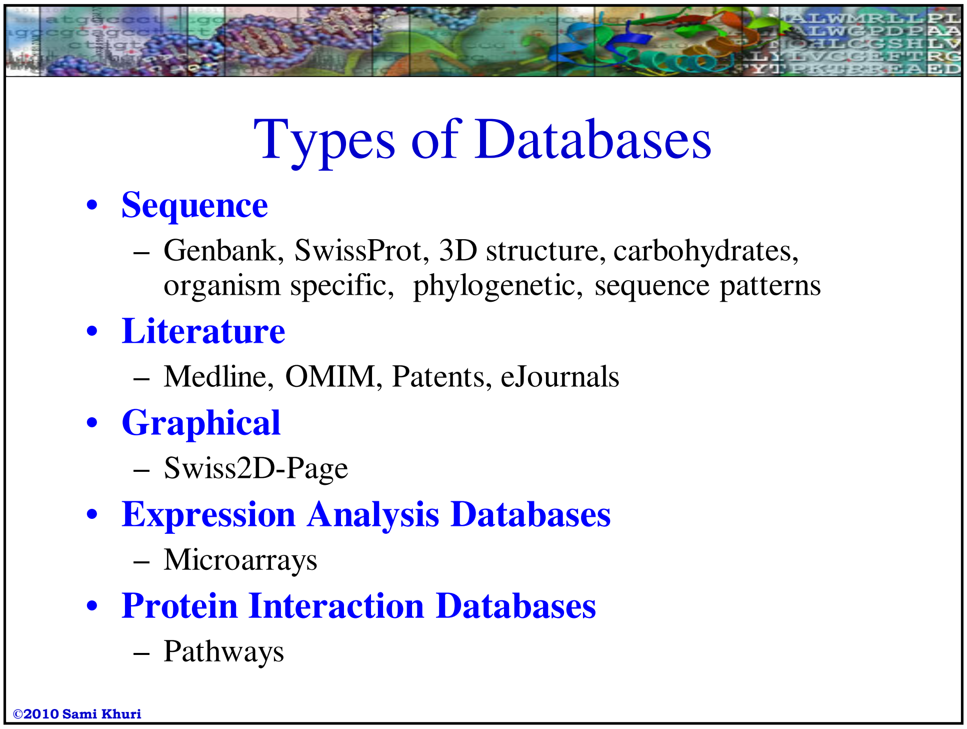
POSSIBLE DRUG

Rational drug design
Structure-based drug design



Databases for Storage and Analysis

- Databases store data that need to be analyzed
- By comparing sequences, we discover:
  - How organisms are related to one another
  - How proteins function
  - How populations vary
  - How diseases occur
- The improvement of sequencing methods generated a lot of data that need to be:
  - stored
  - annotated
  - accessed
  - organized
  - managed
  - assessed
  - curated
  - networked



Types of Databases

- **Sequence**
  - Genbank, SwissProt, 3D structure, carbohydrates, organism specific, phylogenetic, sequence patterns
- **Literature**
  - Medline, OMIM, Patents, eJournals
- **Graphical**
  - Swiss2D-Page
- **Expression Analysis Databases**
  - Microarrays
- **Protein Interaction Databases**
  - Pathways

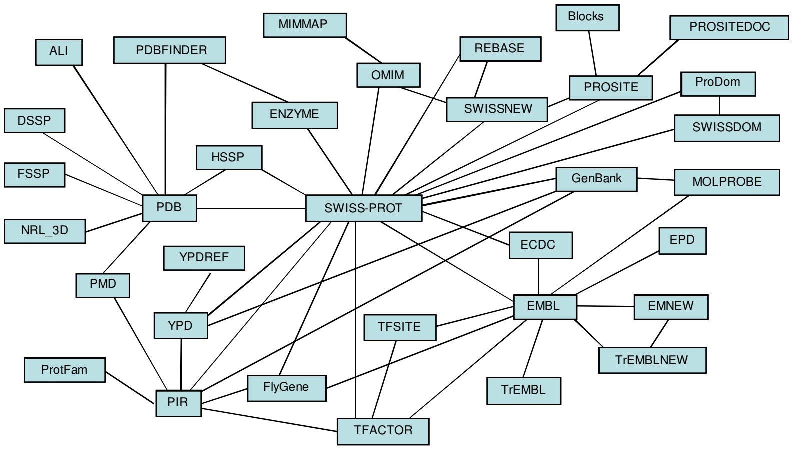
Three Major Databases

- GenBank from the NCBI (National Center of Biotechnology Information), National Library of Medicine http://www.ncbi.nlm.nih.gov
- EBI (European Bioinformatics Institute) from the European Molecular Biology Library http://www.ebi.ac.uk
- DDBJ (DNA DataBank of Japan) http://www.ddbj.nig.ac.jp

GenBank Taxonomic Sampling

Homo sapiens	62.1%
Mus musculus	7.7%
Drosophila melanogaster	6.1%
Caenorhabditis elegans	3.3%
Arabidopsis thaliana	2.9%
Oryza sativa	1.3%
Rattus norvegicus	0.8%
Danio rerio	0.6%
Saccharomyces cerevisiae	0.6%

Databanks Interconnection



MIMMAP
Blocks
PROSITEDOC
ALI
PDBFINDER
REBASE
OMIM
PROSITE
ProDom
ENZYME
SWISSNEW
DSSP
SWISSDOM
HSSP
FSSP
GenBank
MOLPROBE
PDB
SWISS-PROT
NRL_3D
ECDC
EPD
YPDREF
PMD
EMBL
EMNEW
YPD
TFSITE
TrEMBLNEW
ProtFam
FlyGene
TrEMBL
PIR
TFACTOR

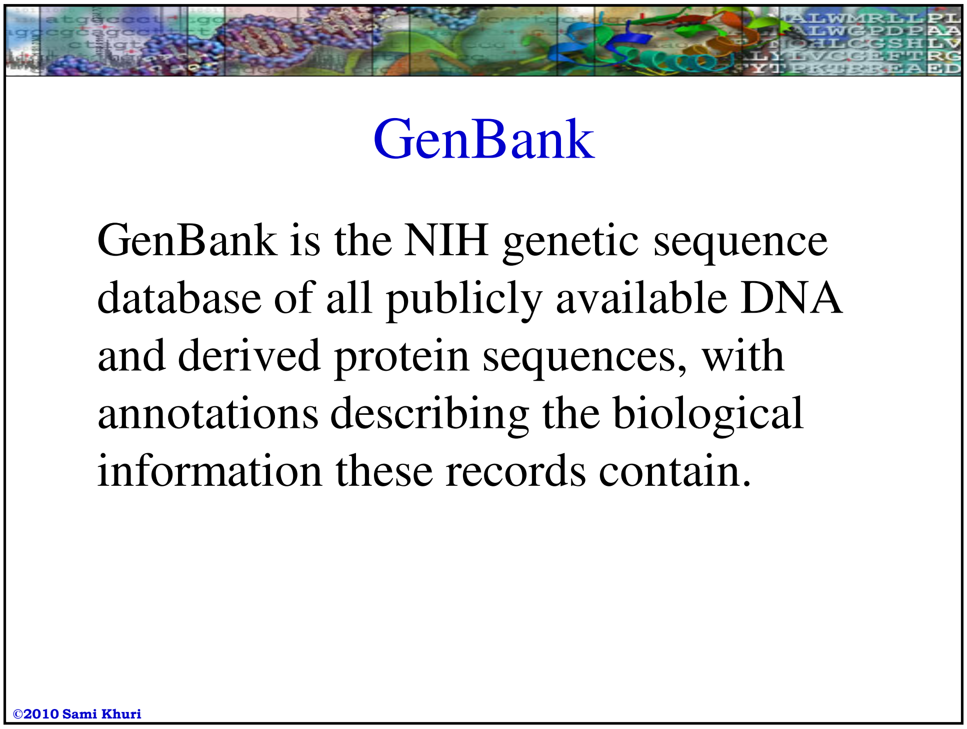
The major DNA databases are updated and synchronized daily.

What does NCBI do?

NCBI: established in 1988 as a national resource for molecular biology information.

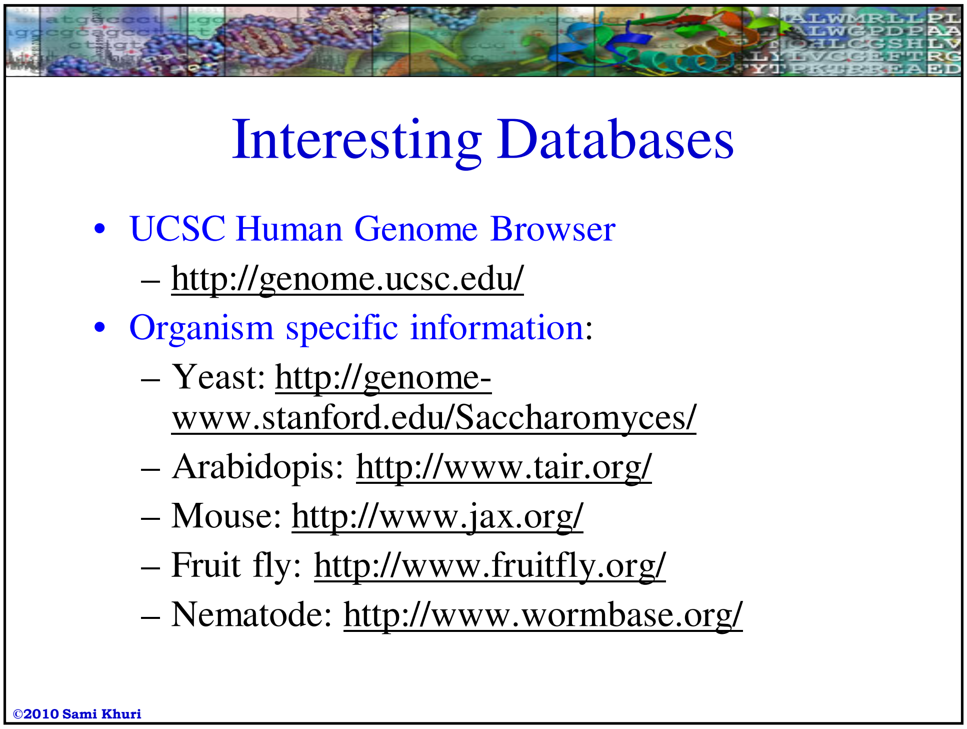
- it creates public databases,
- it conducts research in computational biology,
- it develops software tools for analyzing genome data, and
- it disseminates biomedical information,

all for the better understanding of molecular processes affecting human health and disease.



GenBank

GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.



Interesting Databases

- UCSC Human Genome Browser
  - http://genome.ucsc.edu/
- Organism specific information:
  - Yeast: http://genome-www.stanford.edu/Saccharomyces/
  - Arabidopis: http://www.tair.org/
  - Mouse: http://www.jax.org/
  - Fruit fly: http://www.fruitfly.org/
  - Nematode: http://www.wormbase.org/

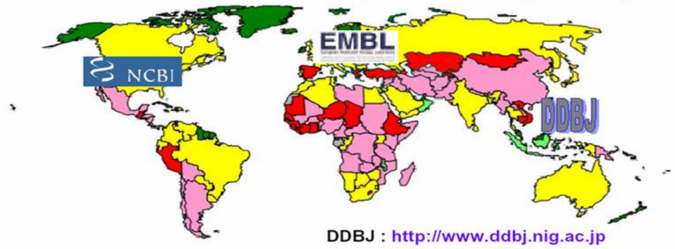
European Molecular Biology Laboratory

- The **European Molecular Biology Laboratory** (**EMBL**) was established in 1974.
- It is supported by sixteen countries.
- EMBL consists of five facilities:
  - The main Laboratory in Heidelberg (Germany),
  - Outstations in Hamburg (Germany), Grenoble (France) and Hinxton (the U.K.), and an external Research Programme in Monterotondo (Italy).

NCBI – EMBL - DDJB

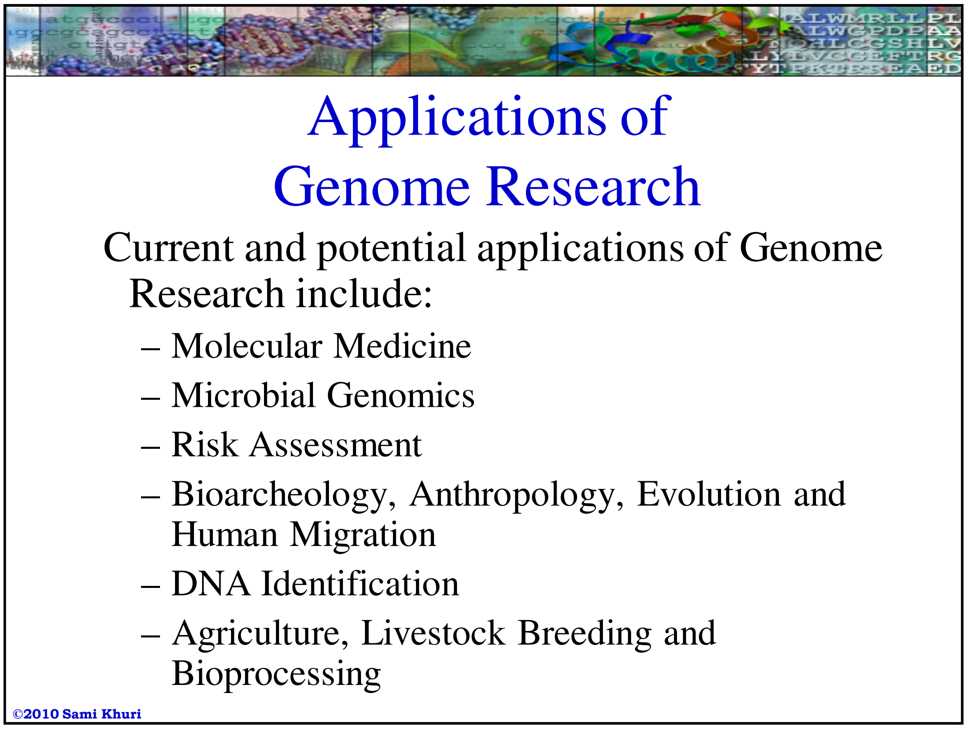
NCBI : http://www.ncbi.nlm.nih.gov/
NCBI, at the NIH campus, USA

EMBL : http://www.embl-heidelberg.de/
European Molecular Biology Laboratory, UK



DDBJ : http://www.ddbj.nig.ac.jp
DNA Databank of Japan

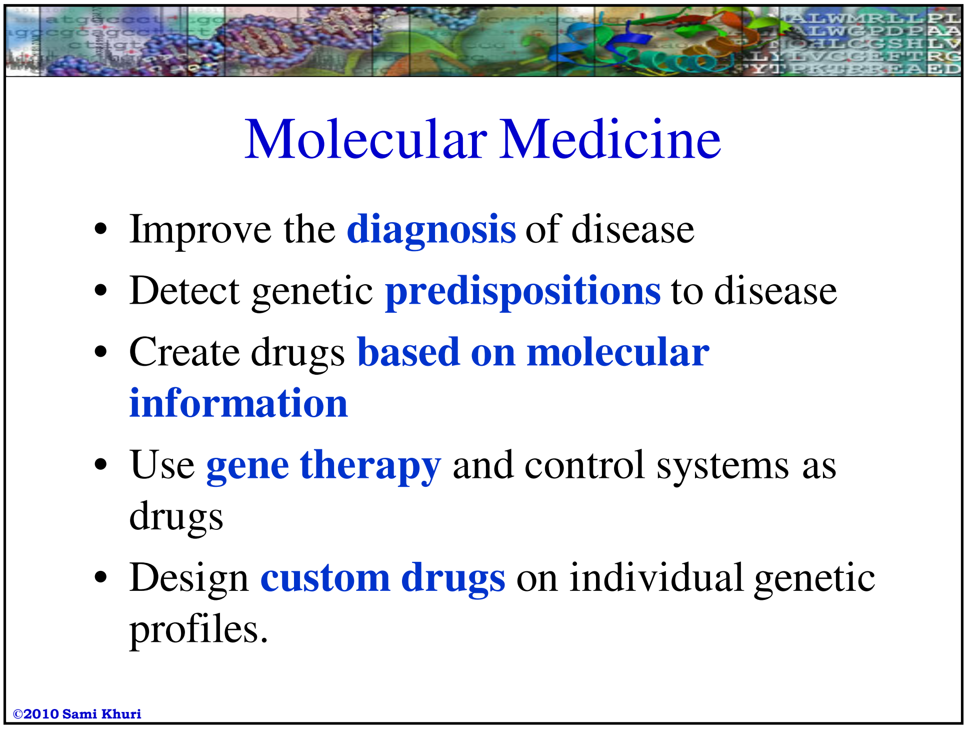
Nucleic acid Databases



Applications of Genome Research

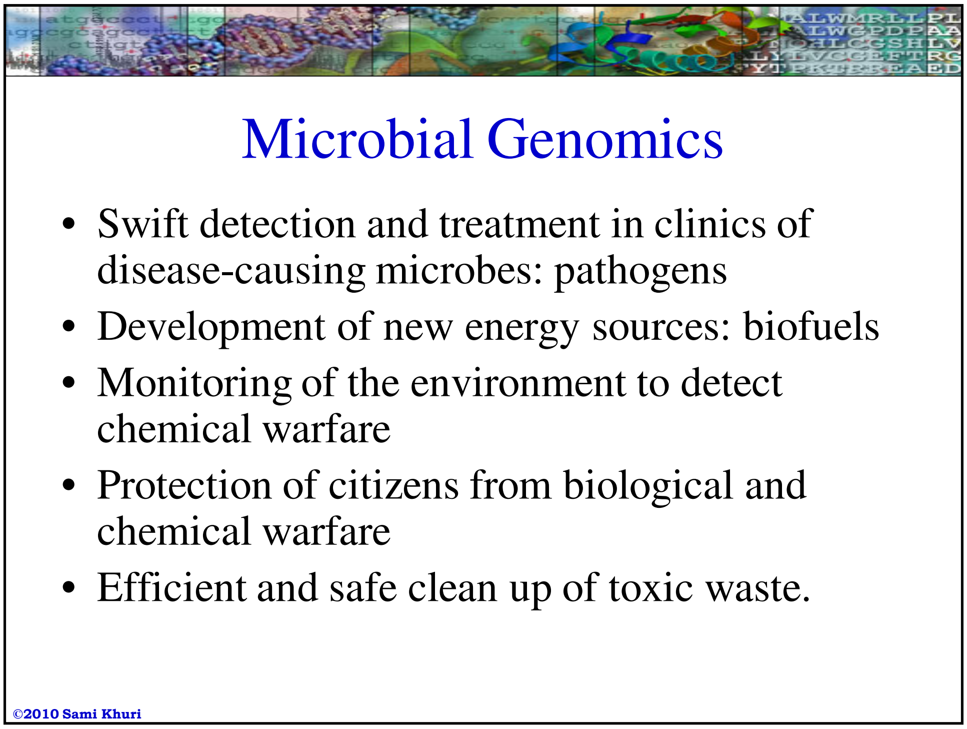
Current and potential applications of Genome Research include:

- Molecular Medicine
- Microbial Genomics
- Risk Assessment
- Bioarcheology, Anthropology, Evolution and Human Migration
- DNA Identification
- Agriculture, Livestock Breeding and Bioprocessing



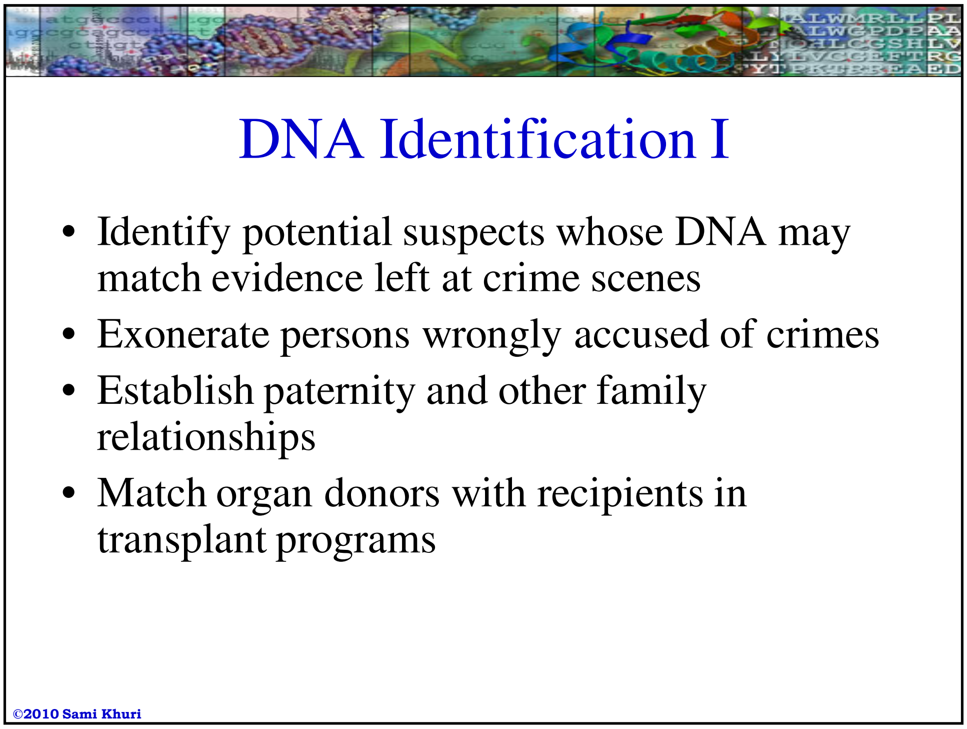
Molecular Medicine

- Improve the **diagnosis** of disease
- Detect genetic **predispositions** to disease
- Create drugs **based on molecular information**
- Use **gene therapy** and control systems as drugs
- Design **custom drugs** on individual genetic profiles.



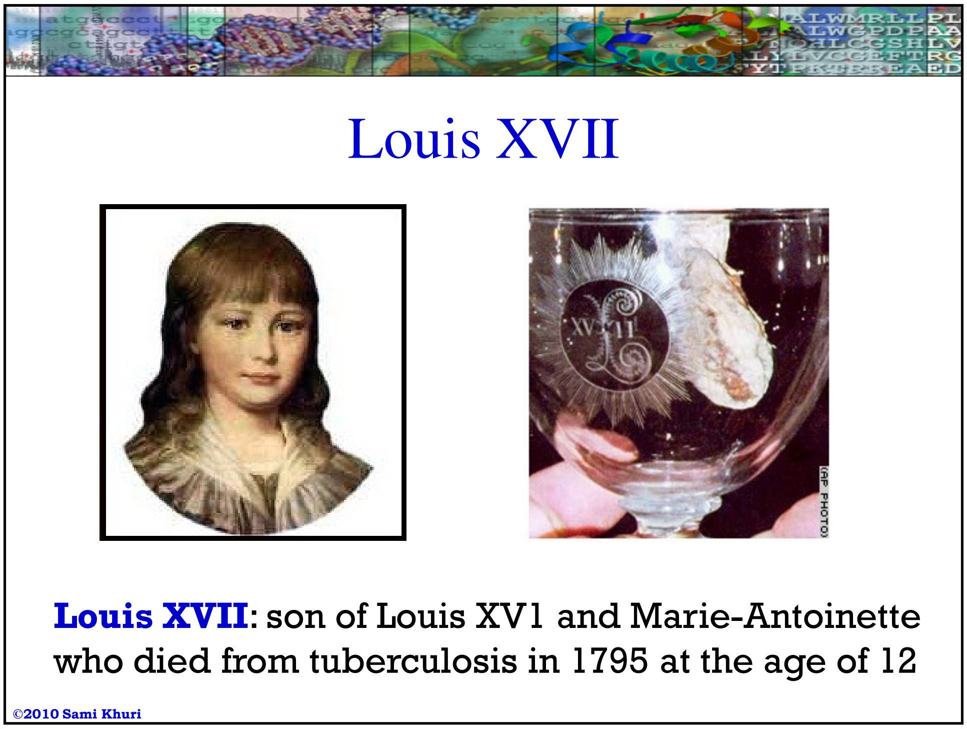
Microbial Genomics

- Swift detection and treatment in clinics of disease-causing microbes: pathogens
- Development of new energy sources: biofuels
- Monitoring of the environment to detect chemical warfare
- Protection of citizens from biological and chemical warfare
- Efficient and safe clean up of toxic waste.



DNA Identification I

- Identify potential suspects whose DNA may match evidence left at crime scenes
- Exonerate persons wrongly accused of crimes
- Establish paternity and other family relationships
- Match organ donors with recipients in transplant programs

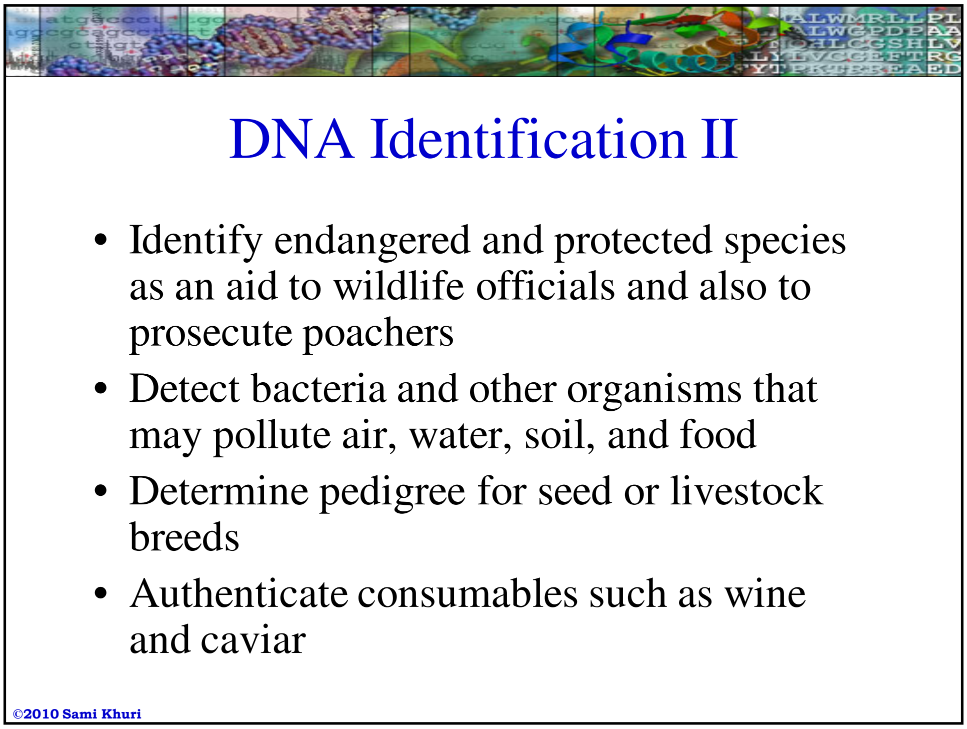


Louis XVII

(AP PHOTO)

Louis XVII: son of Louis XV1 and Marie-Antoinette who died from tuberculosis in 1795 at the age of 12

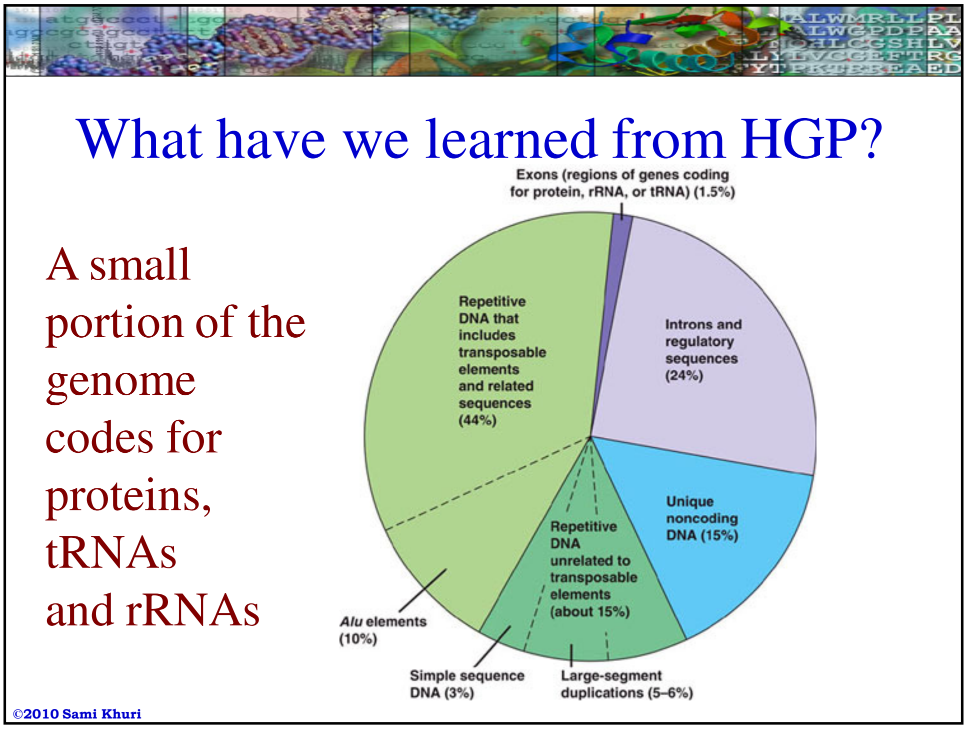
©2010 Sami Khuri



DNA Identification II

- Identify endangered and protected species as an aid to wildlife officials and also to prosecute poachers
- Detect bacteria and other organisms that may pollute air, water, soil, and food
- Determine pedigree for seed or livestock breeds
- Authenticate consumables such as wine and caviar

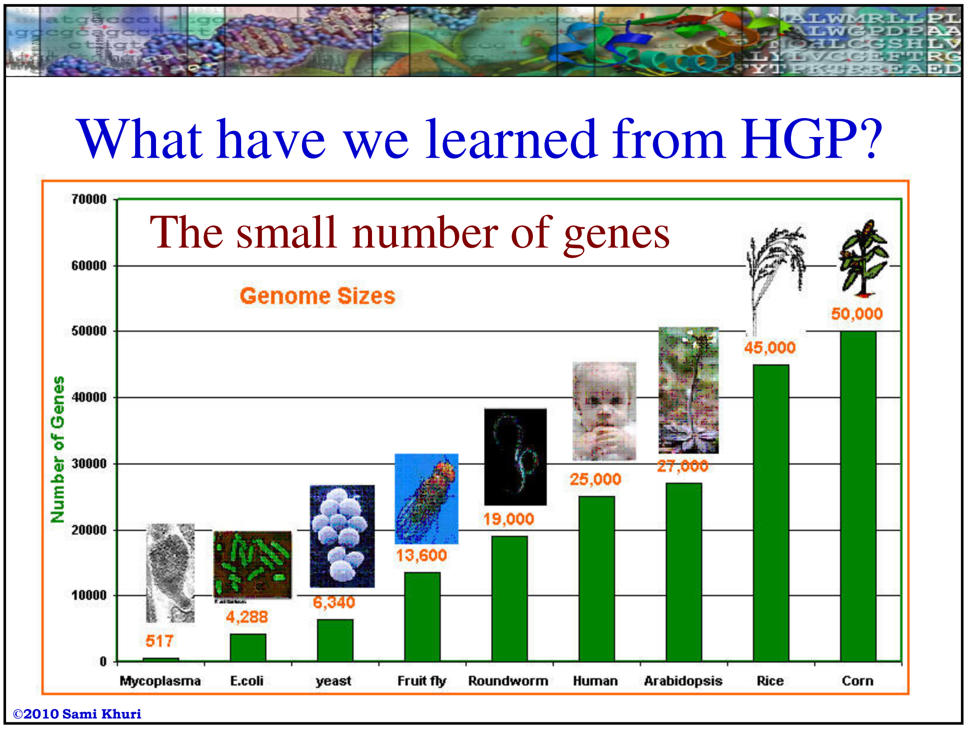
©2010 Sami Khuri



What have we learned from HGP?

A small portion of the genome codes for proteins, tRNAs and rRNAs

©2010 Sami Khuri

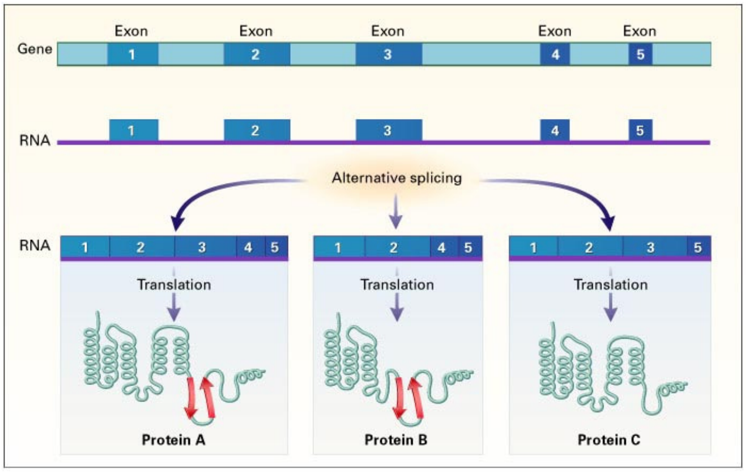


What have we learned from HGP?

The small number of genes

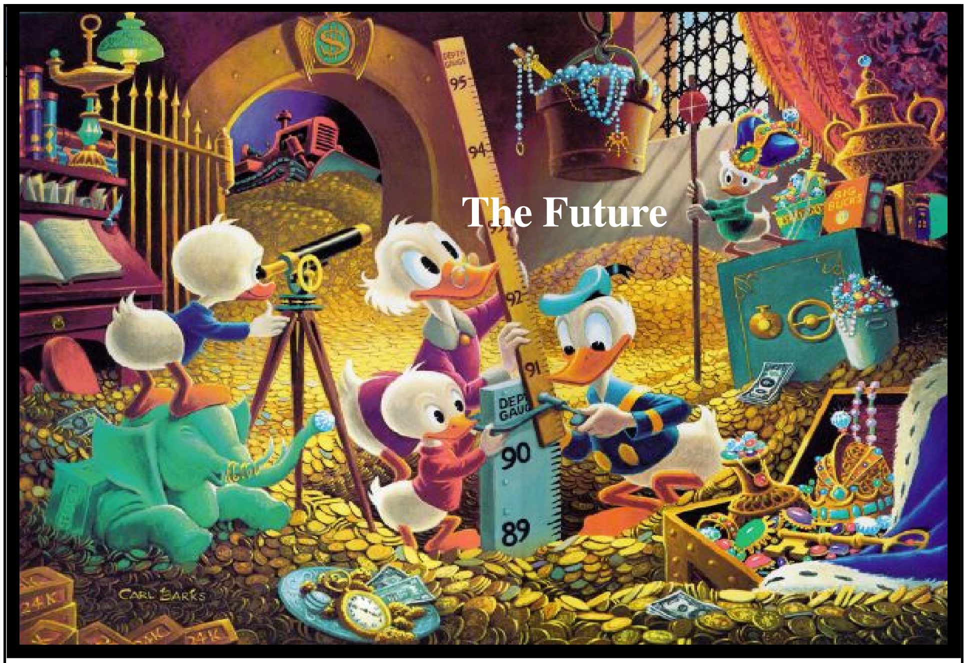
©2010 Sami Khuri

Alternative Splicing



Genomic Medicine by Guttmacher et al., NEJM, 2002

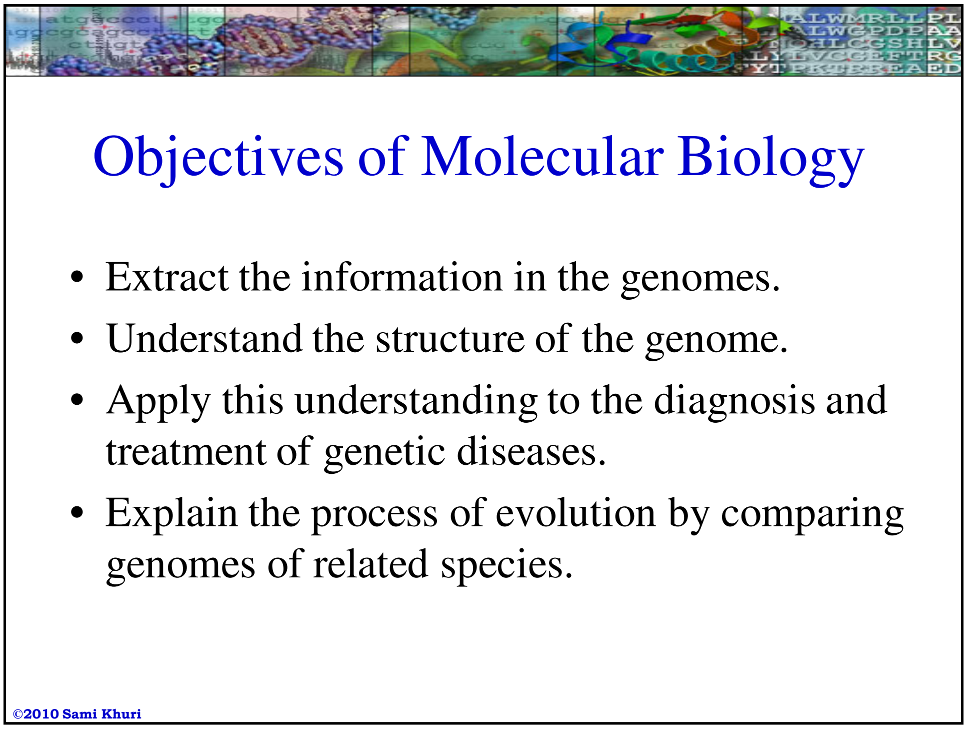
©2010 Sami Khuri



The Future

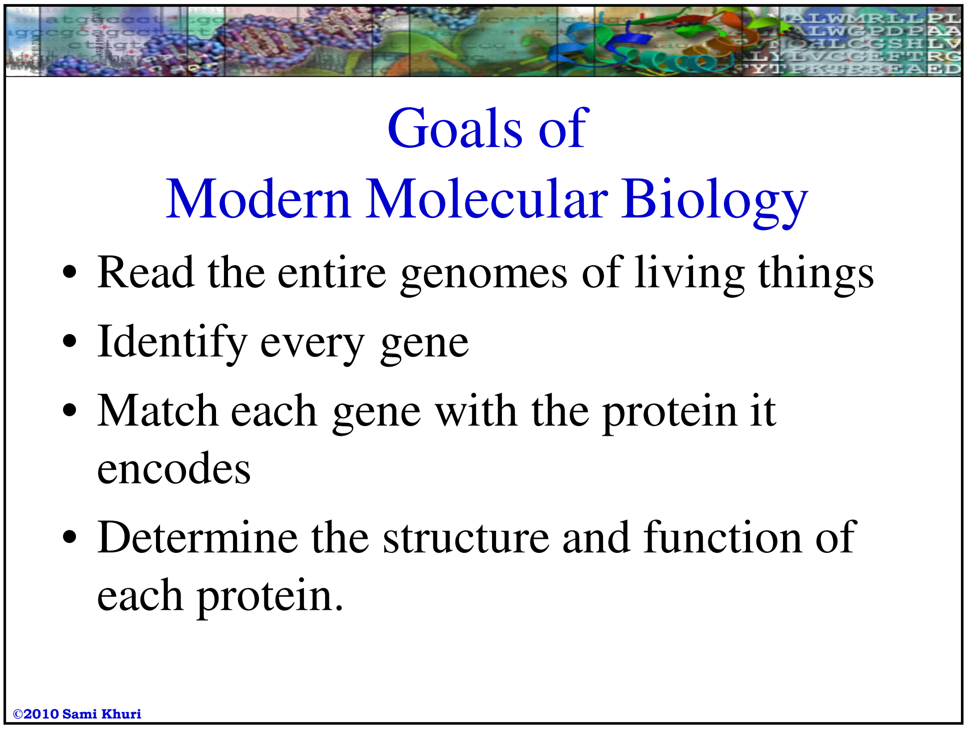
Convert all this progress into real riches for science, society, and patients

©2010 Sami Khuri



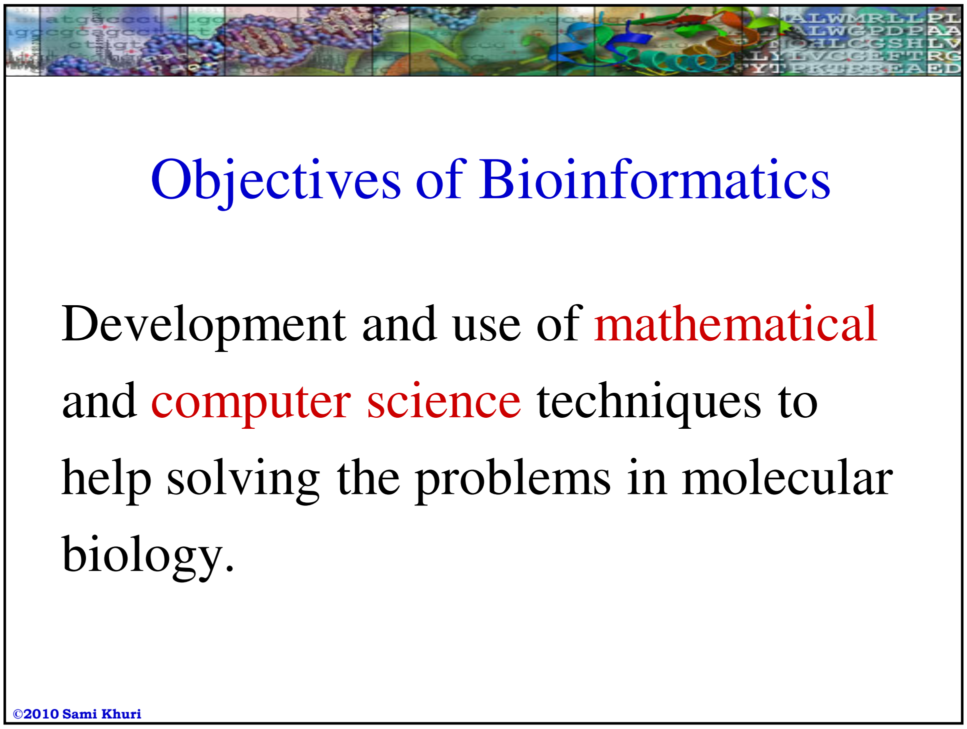
Objectives of Molecular Biology

- Extract the information in the genomes.
- Understand the structure of the genome.
- Apply this understanding to the diagnosis and treatment of genetic diseases.
- Explain the process of evolution by comparing genomes of related species.



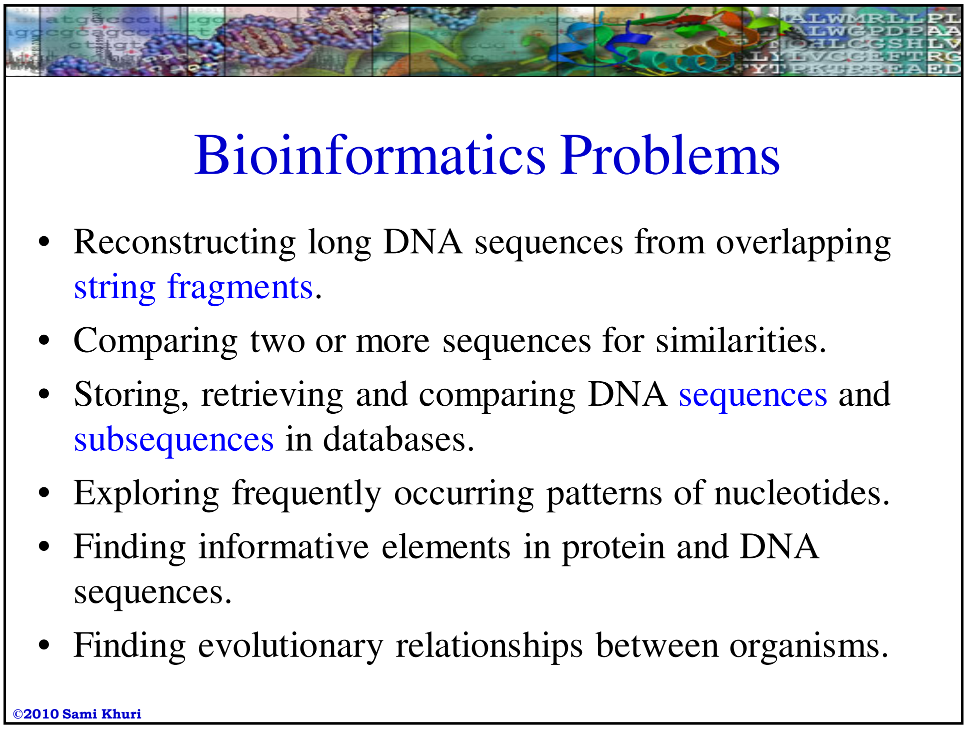
Goals of Modern Molecular Biology

- Read the entire genomes of living things
- Identify every gene
- Match each gene with the protein it encodes
- Determine the structure and function of each protein.



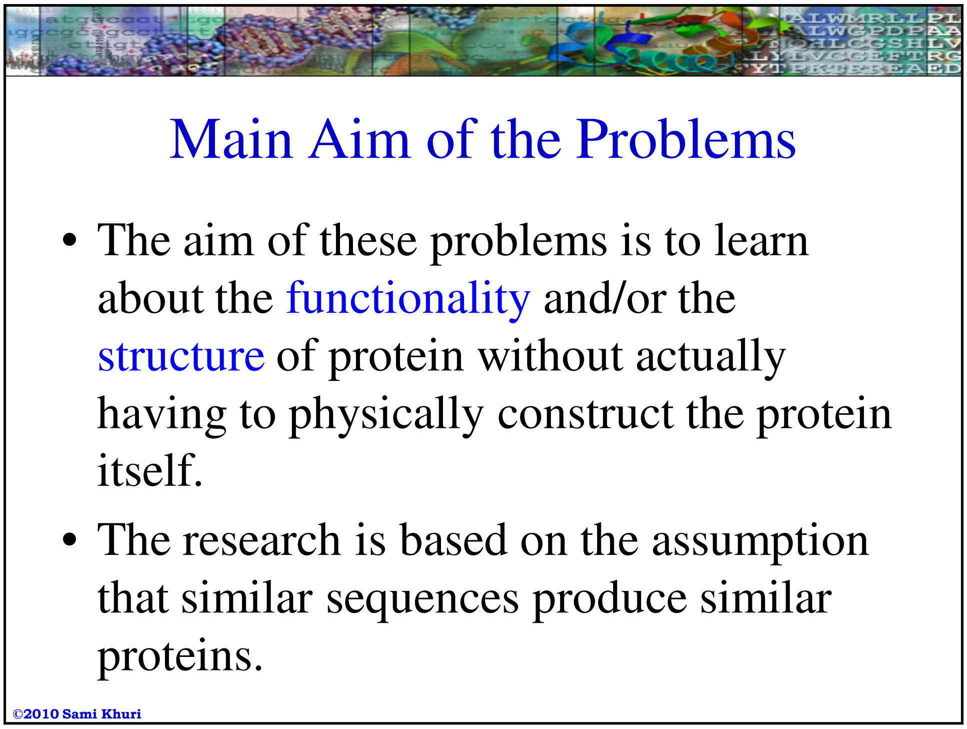
Objectives of Bioinformatics

Development and use of mathematical and computer science techniques to help solving the problems in molecular biology.



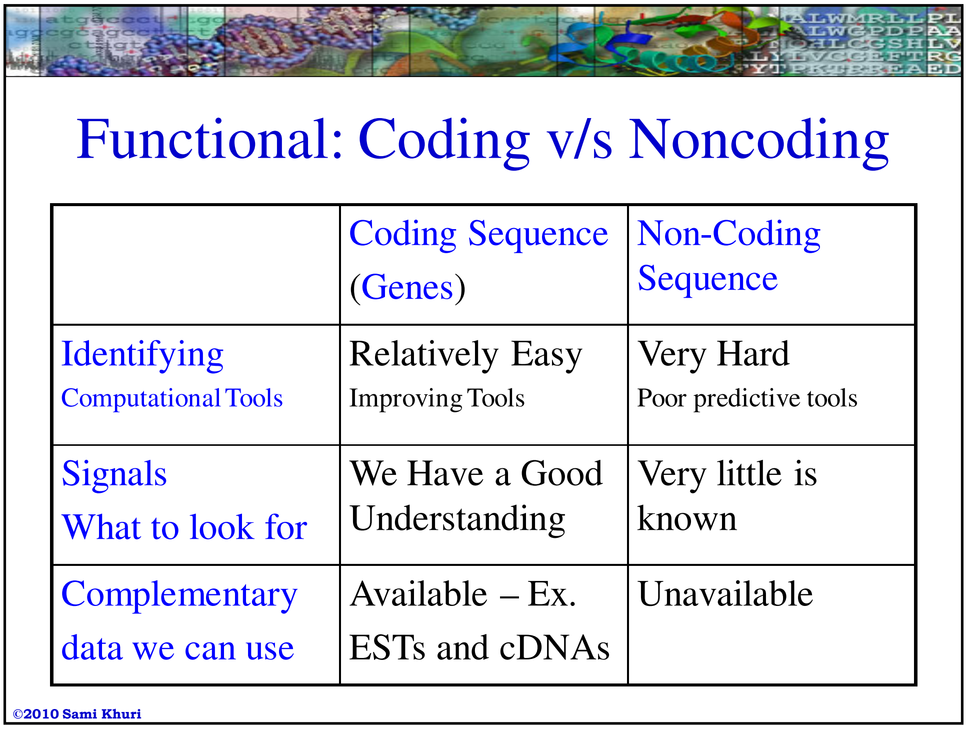
Bioinformatics Problems

- Reconstructing long DNA sequences from overlapping string fragments.
- Comparing two or more sequences for similarities.
- Storing, retrieving and comparing DNA sequences and subsequences in databases.
- Exploring frequently occurring patterns of nucleotides.
- Finding informative elements in protein and DNA sequences.
- Finding evolutionary relationships between organisms.



Main Aim of the Problems

- The aim of these problems is to learn about the functionality and/or the structure of protein without actually having to physically construct the protein itself.
- The research is based on the assumption that similar sequences produce similar proteins.



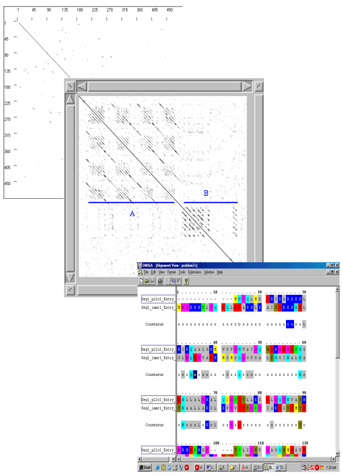
Functional: Coding v/s Noncoding

	Coding Sequence (Genes)	Non-Coding Sequence
Identifying Computational Tools	Relatively Easy Improving Tools	Very Hard Poor predictive tools
Signals What to look for	We Have a Good Understanding	Very little is known
Complementary data we can use	Available – Ex. ESTs and cDNAs	Unavailable

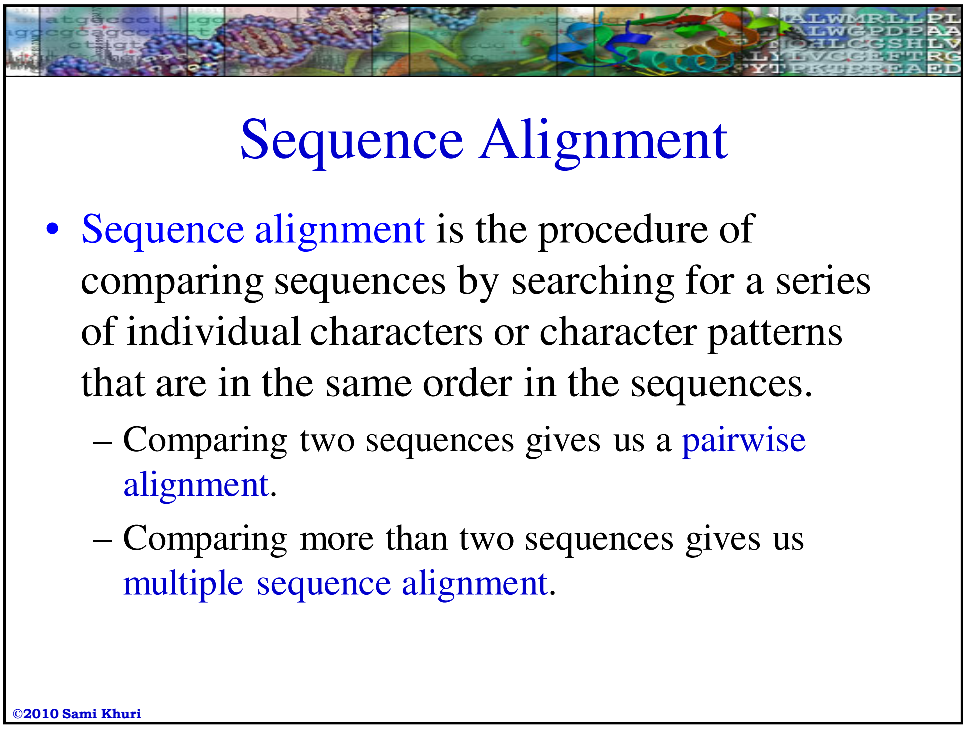
Post Human Genome Project

- Major role for comparative sequence analysis will be the identification of functionally important, non-coding sequences.
- Need to study the relation between Sequence Conservation and Sequence Function.
- Focus on the interpretation of the human genome.
- Learn the functional landscape of the human genome.
- **Challenge**: go from sequence to function
  - i.e., define the role of each gene and understand how the genome functions as a whole.

Pairwise and Multiple Sequence Alignment

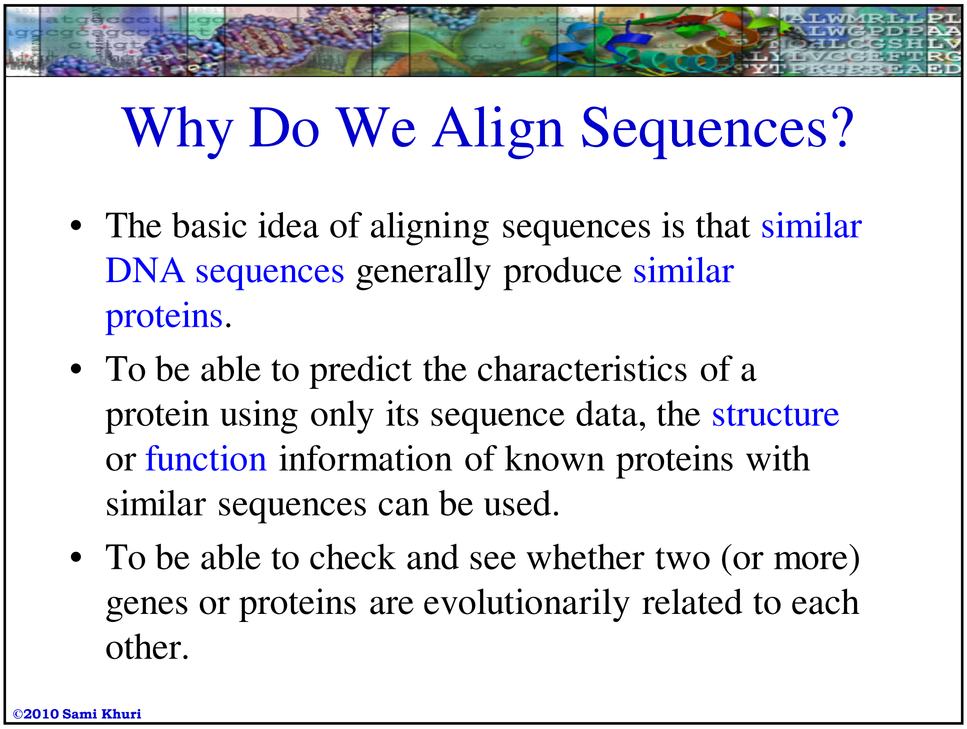


- **Homology**
- **Similarity**
- **Global string alignment**
- **Local string alignment**
- **Dynamic programming**
- **Scoring matrices:**
  PAM and BLOSUM
- **BLAST family**



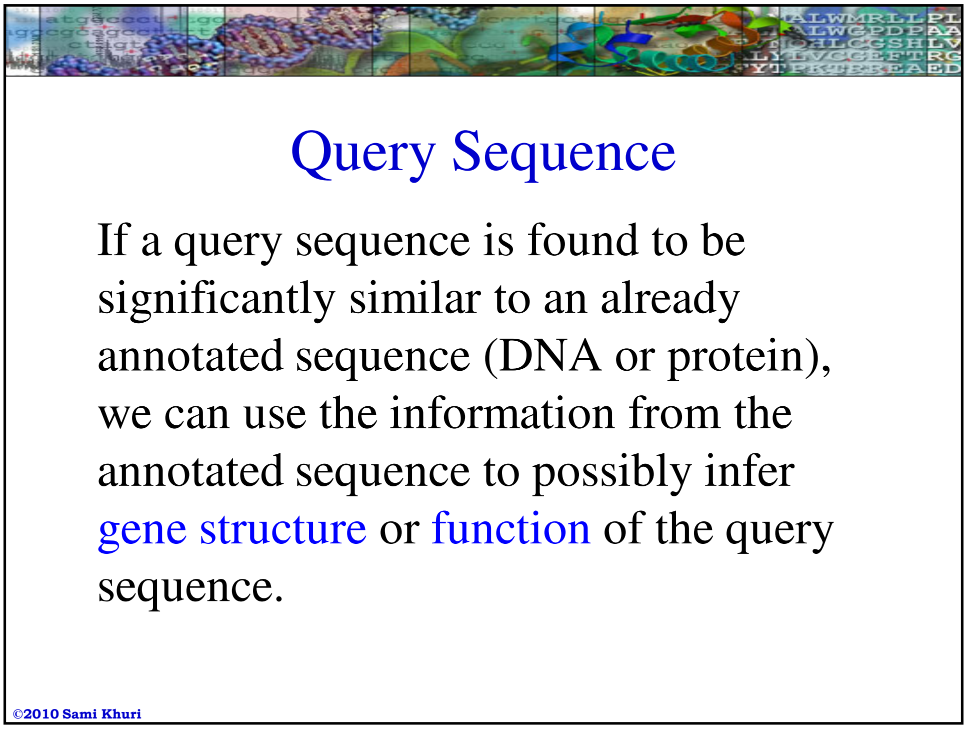
Sequence Alignment

- Sequence alignment is the procedure of comparing sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences.
  - Comparing two sequences gives us a pairwise alignment.
  - Comparing more than two sequences gives us multiple sequence alignment.



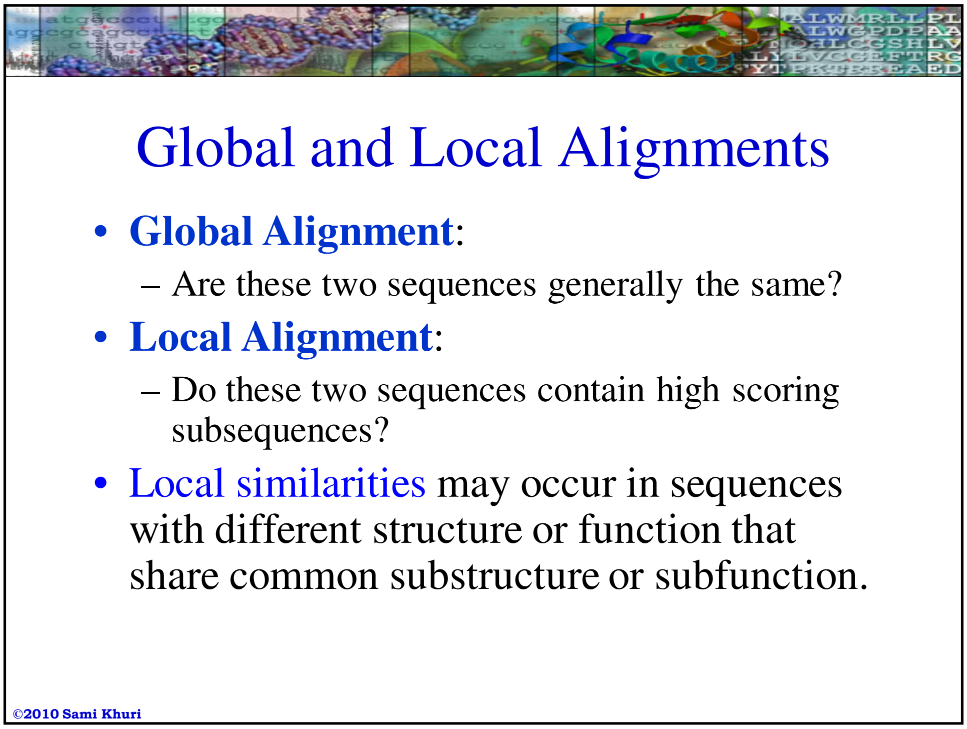
Why Do We Align Sequences?

- The basic idea of aligning sequences is that similar DNA sequences generally produce similar proteins.
- To be able to predict the characteristics of a protein using only its sequence data, the structure or function information of known proteins with similar sequences can be used.
- To be able to check and see whether two (or more) genes or proteins are evolutionarily related to each other.



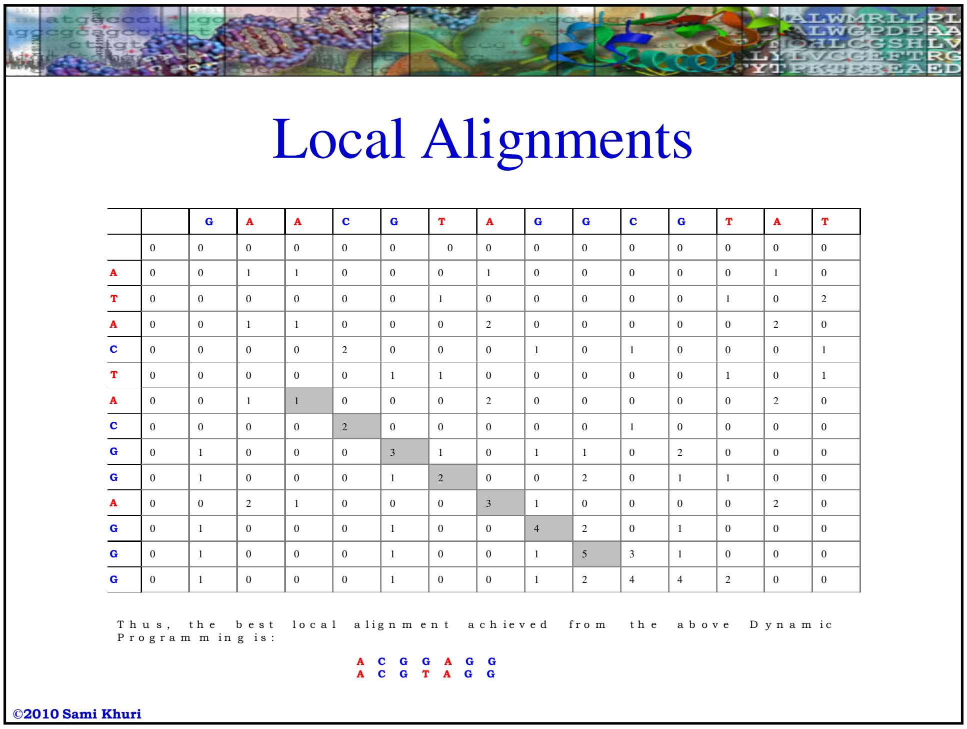
Query Sequence

If a query sequence is found to be significantly similar to an already annotated sequence (DNA or protein), we can use the information from the annotated sequence to possibly infer gene structure or function of the query sequence.



Global and Local Alignments

- **Global Alignment**:
  - Are these two sequences generally the same?
- **Local Alignment**:
  - Do these two sequences contain high scoring subsequences?
- Local similarities may occur in sequences with different structure or function that share common substructure or subfunction.

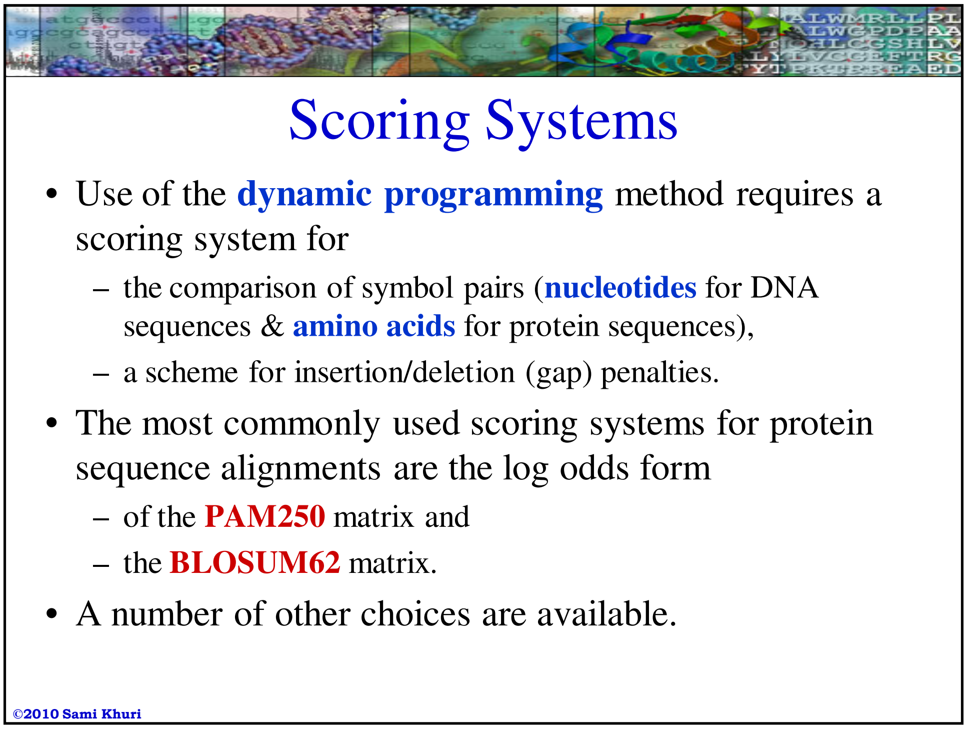


Local Alignments

		G	A	A	C	G	T	A	G	G	C	G	T	A	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0
T	0	0	0	0	0	0	1	0	0	0	0	0	1	0	2
A	0	0	1	1	0	0	0	2	0	0	0	0	0	2	0
C	0	0	0	0	2	0	0	0	1	0	1	0	0	0	1
T	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1
A	0	0	1	1	0	0	0	2	0	0	0	0	0	2	0
C	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0
G	0	1	0	0	0	3	1	0	1	1	0	2	0	0	0
G	0	1	0	0	0	1	2	0	0	2	0	1	1	0	0
A	0	0	2	1	0	0	0	3	1	0	0	0	0	2	0
G	0	1	0	0	0	1	0	0	4	2	0	1	0	0	0
G	0	1	0	0	0	1	0	0	1	5	3	1	0	0	0
G	0	1	0	0	0	1	0	0	1	2	4	4	2	0	0

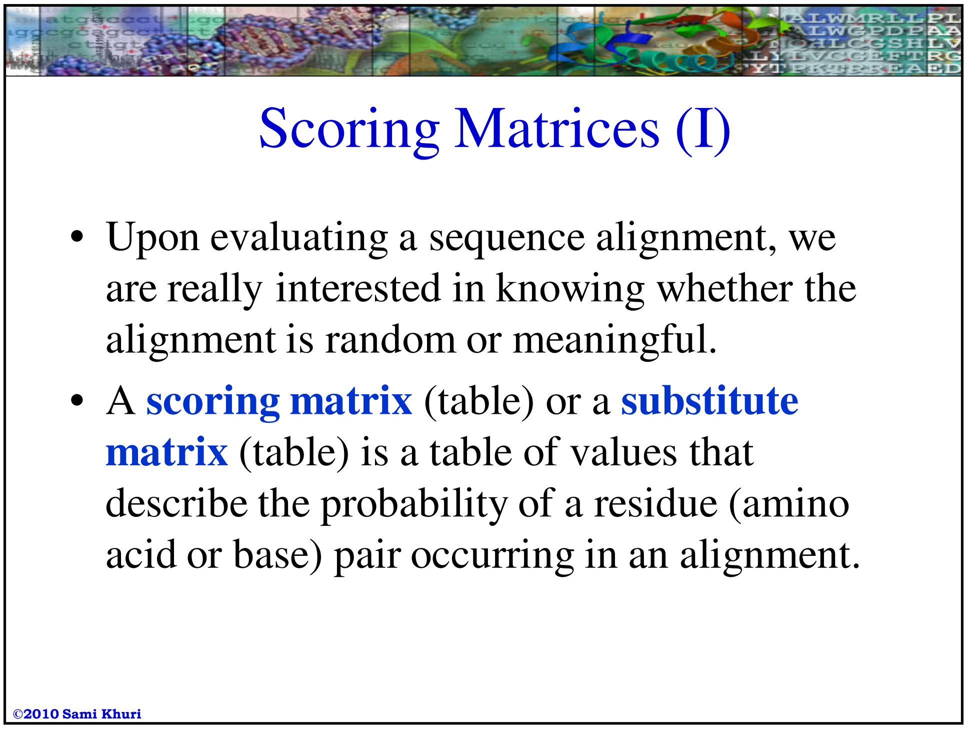
Thus, the best local alignment achieved from the above Dynamic Programming is:

```
A C G G A G G
A C G T A G G
```



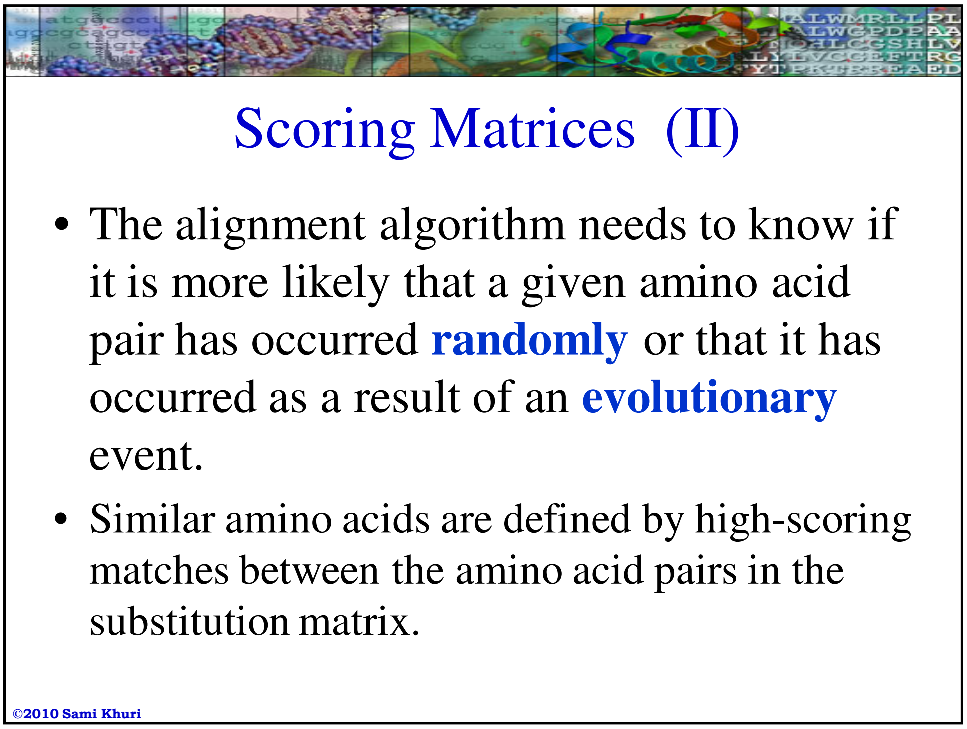
Scoring Systems

- Use of the **dynamic programming** method requires a scoring system for
  - the comparison of symbol pairs (**nucleotides** for DNA sequences & **amino acids** for protein sequences),
  - a scheme for insertion/deletion (gap) penalties.
- The most commonly used scoring systems for protein sequence alignments are the log odds form
  - of the **PAM250** matrix and
  - the **BLOSUM62** matrix.
- A number of other choices are available.



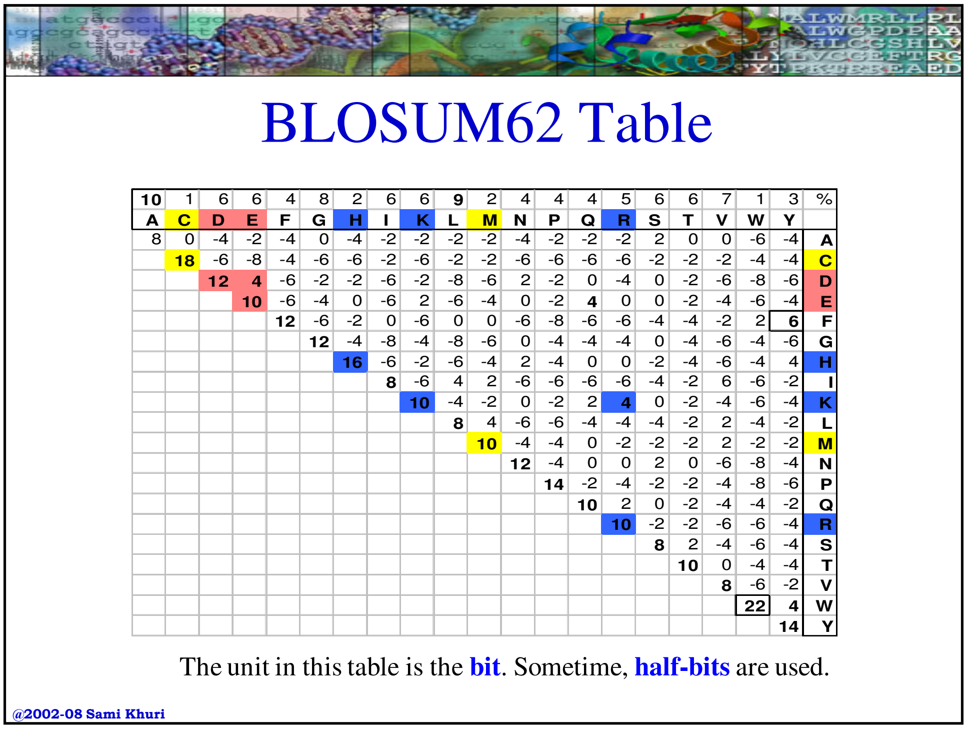
Scoring Matrices (I)

- Upon evaluating a sequence alignment, we are really interested in knowing whether the alignment is random or meaningful.
- A **scoring matrix** (table) or a **substitute matrix** (table) is a table of values that describe the probability of a residue (amino acid or base) pair occurring in an alignment.



Scoring Matrices (II)

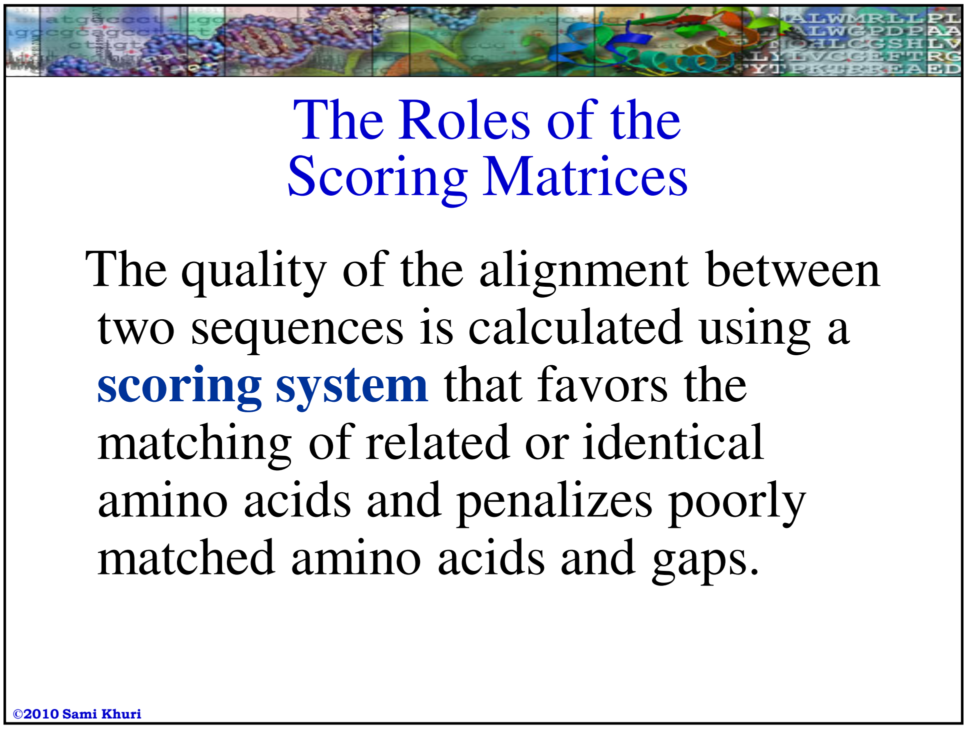
- The alignment algorithm needs to know if it is more likely that a given amino acid pair has occurred **randomly** or that it has occurred as a result of an **evolutionary** event.
- Similar amino acids are defined by high-scoring matches between the amino acid pairs in the substitution matrix.



BLOSUM62 Table

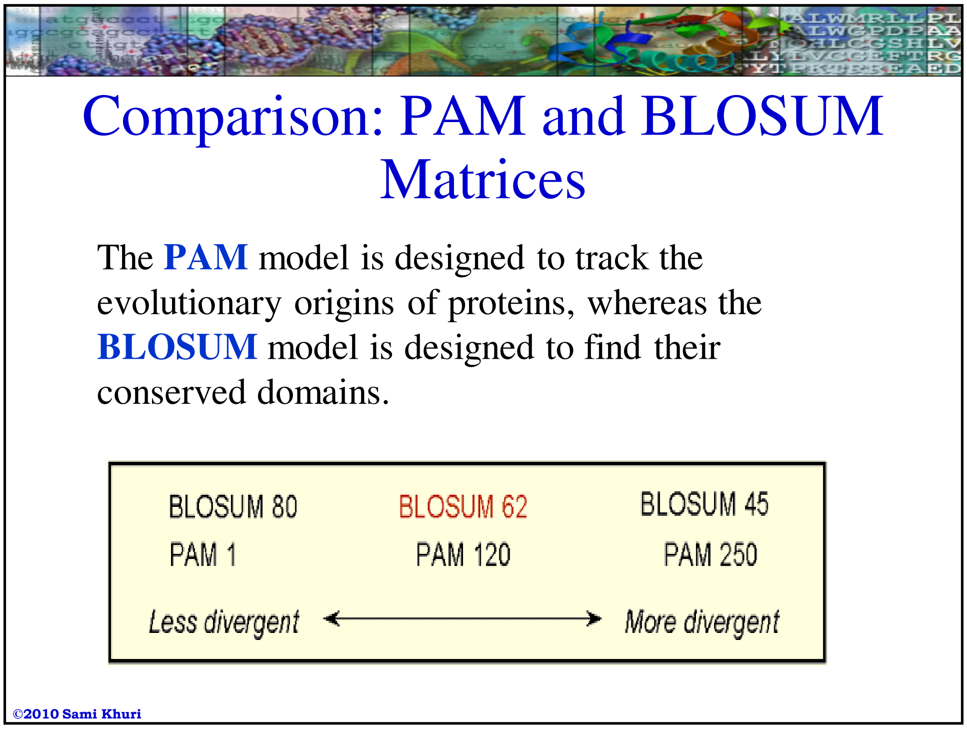
10	1	6	6	4	8	2	6	6	9	2	4	4	4	5	6	6	7	1	3	%
A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
8	0	-4	-2	-4	0	-4	-2	-2	-2	-2	-4	-2	-2	-2	2	0	0	-6	-4	A
	18	-6	-8	-4	-6	-6	-2	-6	-2	-2	-6	-6	-6	-6	-2	-2	-2	-4	-4	C
		12	4	-6	-2	-2	-6	-2	-8	-6	2	-2	0	-4	0	-2	-6	-8	-6	D
			10	-6	-4	0	-6	2	-6	-4	0	-2	4	0	0	-2	-4	-6	-4	E
				12	-6	-2	0	-6	0	0	-6	-8	-6	-6	-4	-4	-2	2	6	F
					12	-4	-8	-4	-8	-6	0	-4	-4	-4	0	-4	-6	-4	-6	G
						16	-6	-2	-6	-4	2	-4	0	0	-2	-4	-6	-4	4	H
							8	-6	4	2	-6	-6	-6	-6	-4	-2	6	-6	-2	I
								10	-4	-2	0	-2	2	4	0	-2	-4	-6	-4	K
									8	4	-6	-6	-4	-4	-4	-2	2	-4	-2	L
										10	-4	-4	0	-2	-2	-2	2	-2	-2	M
											12	-4	0	0	2	0	-6	-8	-4	N
												14	-2	-4	-2	-2	-4	-8	-6	P
													10	2	0	-2	-4	-4	-2	Q
														10	-2	-2	-6	-6	-4	R
															8	2	-4	-6	-4	S
																10	0	-4	-4	T
																	8	-6	-2	V
																		22	4	W
																			14	Y

The unit in this table is the **bit**. Sometime, **half-bits** are used.



The Roles of the Scoring Matrices

The quality of the alignment between two sequences is calculated using a **scoring system** that favors the matching of related or identical amino acids and penalizes poorly matched amino acids and gaps.

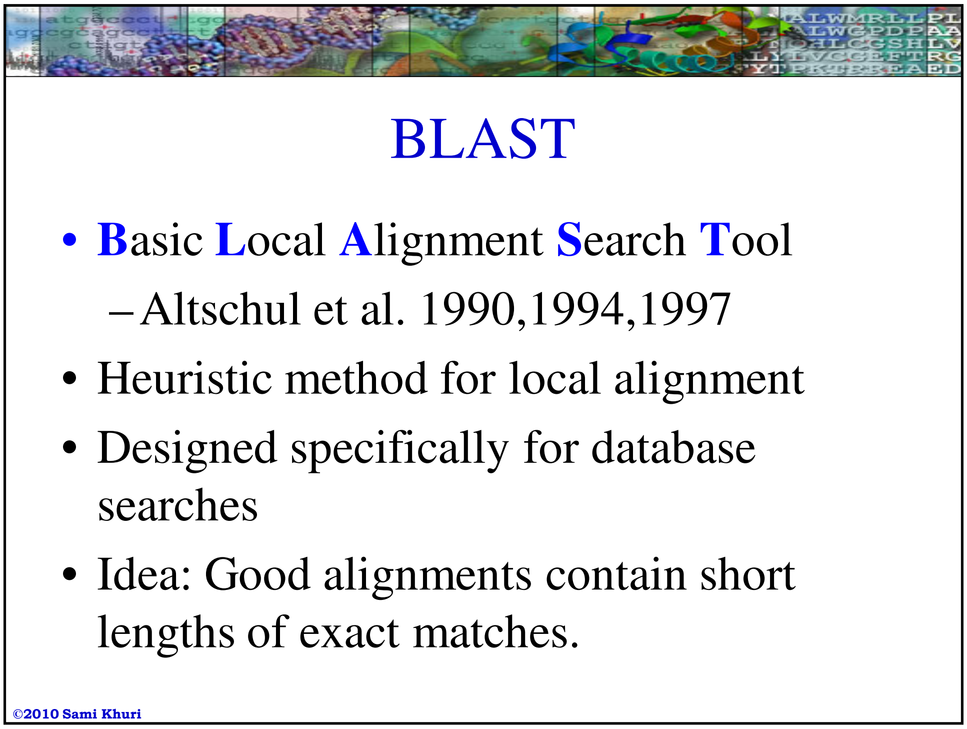


Comparison: PAM and BLOSUM Matrices

The **PAM** model is designed to track the evolutionary origins of proteins, whereas the **BLOSUM** model is designed to find their conserved domains.

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250

Less divergent ← → *More divergent*



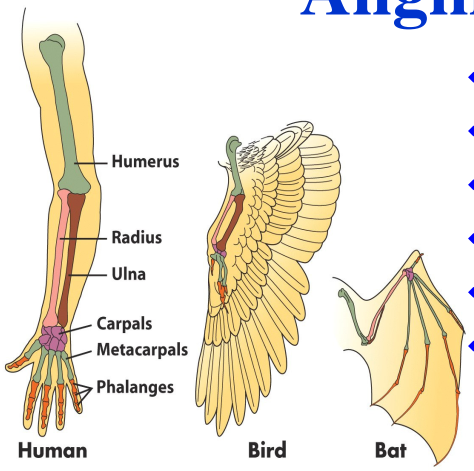
BLAST

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Altschul et al. 1990,1994,1997
- Heuristic method for local alignment
- Designed specifically for database searches
- Idea: Good alignments contain short lengths of exact matches.

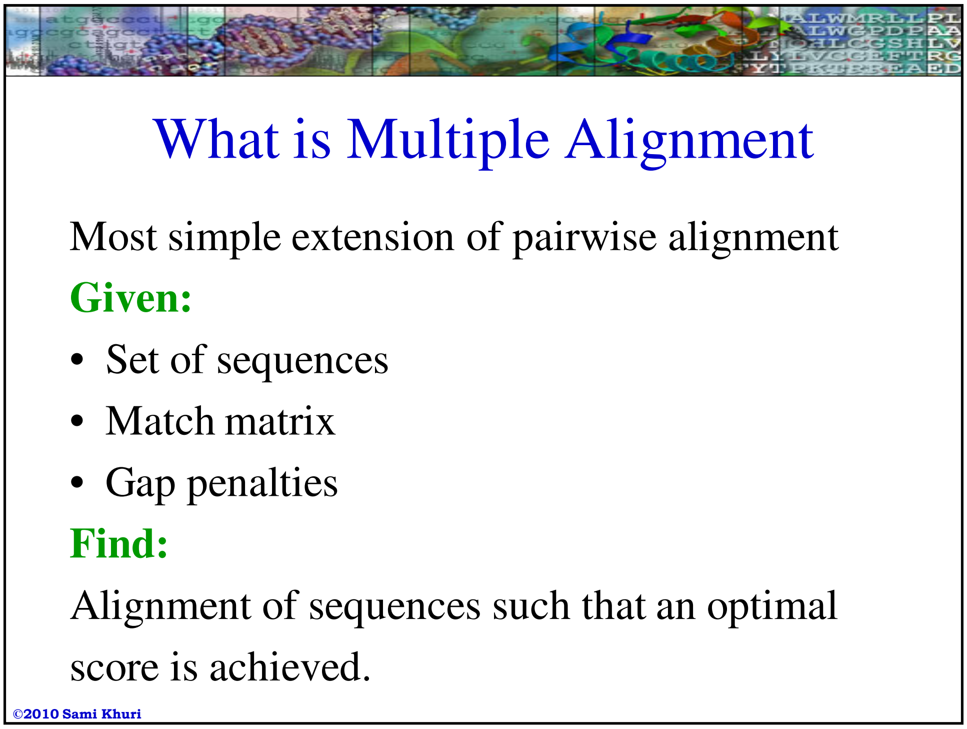
The BLAST Family

- **blastp**: compares an amino acid query sequence against a protein sequence database.
- **blastn**: compares a nucleotide query sequence against a nucleotide sequence database.
- **blastx**: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

Multiple Sequence Alignment



- ❖ Progressive Alignment
- ❖ Iterative Pairwise
- ❖ Guide Tree
- ❖ ClustalW
- ❖ Co-linearity
- ❖ Multiple Sequence Alignment Editors



What is Multiple Alignment

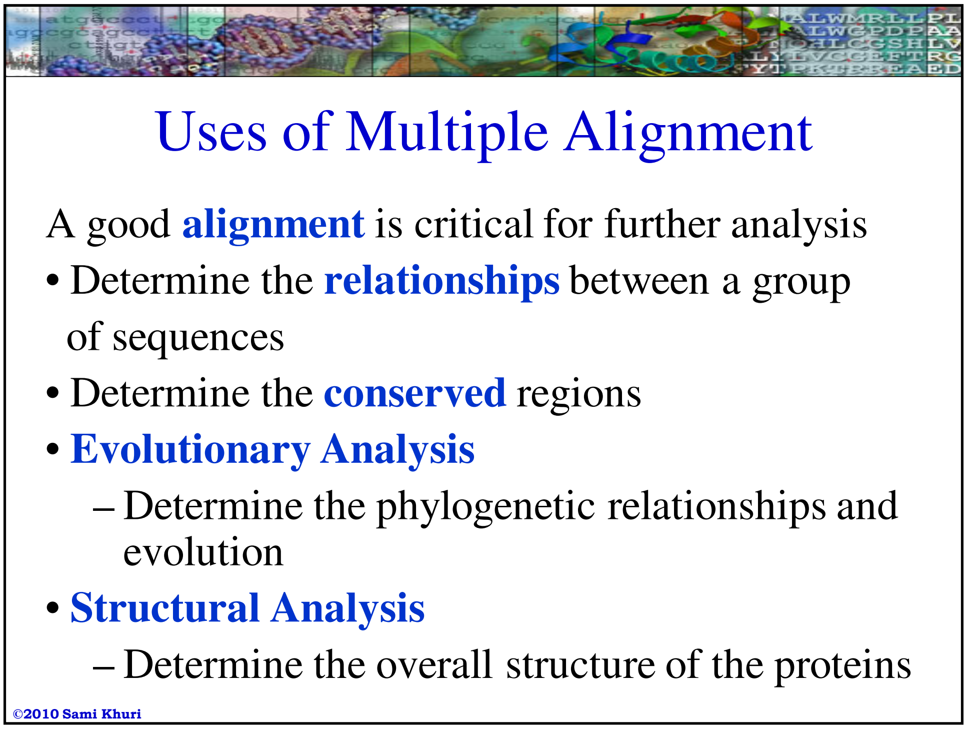
Most simple extension of pairwise alignment

Given:

- Set of sequences
- Match matrix
- Gap penalties

Find:

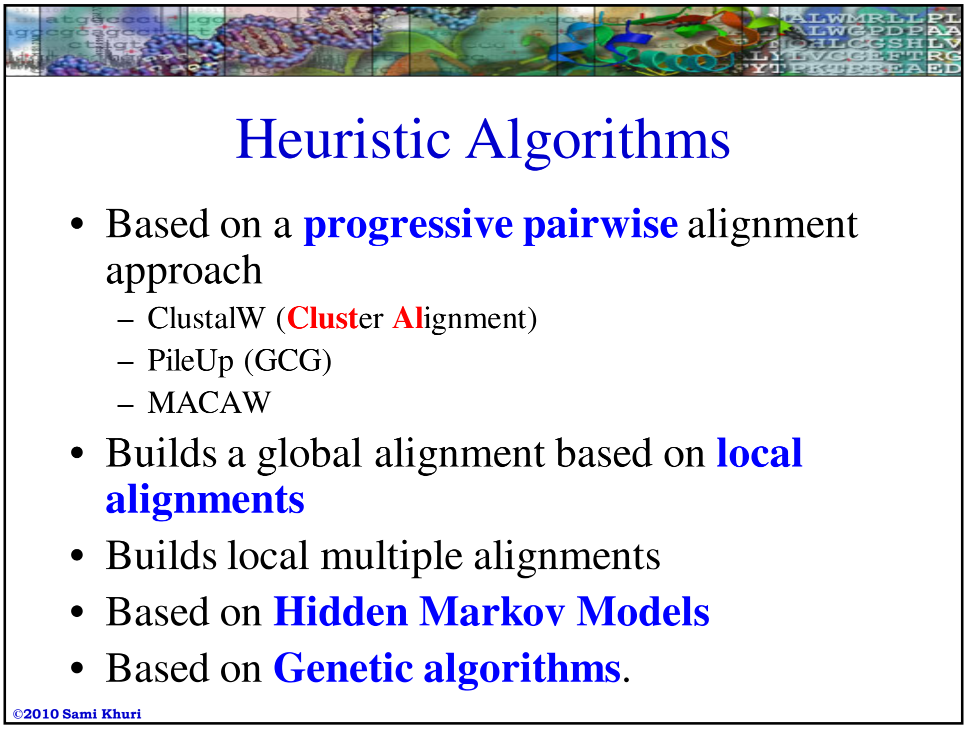
Alignment of sequences such that an optimal score is achieved.



Uses of Multiple Alignment

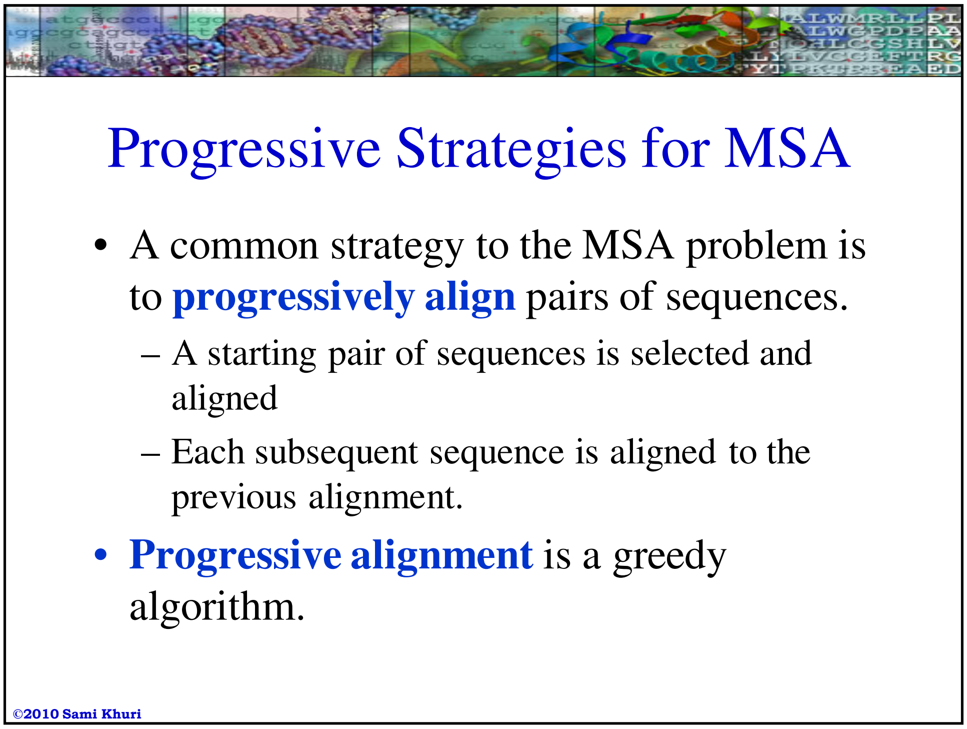
A good **alignment** is critical for further analysis

- Determine the **relationships** between a group of sequences
- Determine the **conserved** regions
- **Evolutionary Analysis**
  - Determine the phylogenetic relationships and evolution
- **Structural Analysis**
  - Determine the overall structure of the proteins



Heuristic Algorithms

- Based on a **progressive pairwise** alignment approach
  - ClustalW (**Clust**er **Al**ignment)
  - PileUp (GCG)
  - MACAW
- Builds a global alignment based on **local alignments**
- Builds local multiple alignments
- Based on **Hidden Markov Models**
- Based on **Genetic algorithms**.



Progressive Strategies for MSA

- A common strategy to the MSA problem is to **progressively align** pairs of sequences.
  - A starting pair of sequences is selected and aligned
  - Each subsequent sequence is aligned to the previous alignment.
- **Progressive alignment** is a greedy algorithm.

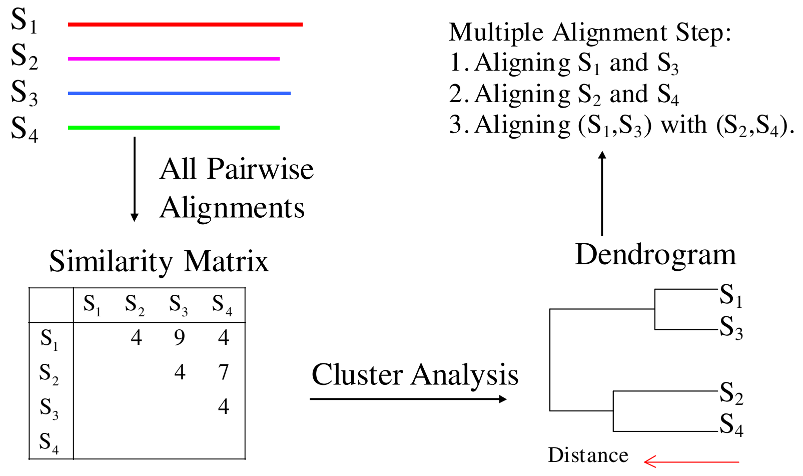
Iterative Pairwise Alignment

- The greedy algorithm:

```
align some pair
while not done
    pick an unaligned string "near"
        some aligned one(s)
    align with the previously aligned group
```

- There are many variants to the algorithm.

Steps of ClustalW



Similarity Matrix

	S_1	S_2	S_3	S_4
S_1		4	9	4
S_2			4	7
S_3				4
S_4				

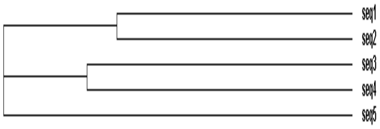
Multiple Alignment Step:
1. Aligning S_1 and S_3
2. Aligning S_2 and S_4
3. Aligning (S_1,S_3) with (S_2,S_4).

ClustalW: An Example

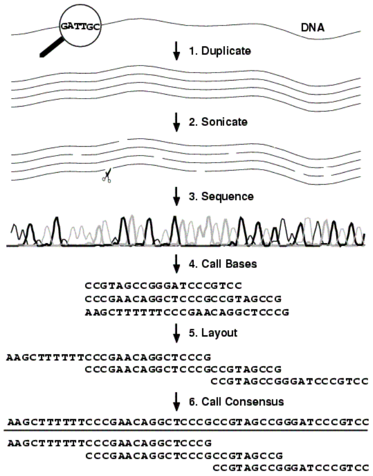
```
CLUSTAL W (1.82) multiple sequence alignment


seq3              FEGGILVEAL 10
seq4              FDG-ILVQAV 9
seq5              YEGGAVVQAL 10
seq1              YDG-GAVEAL 9
seq2              YDG-G--EAL 7
                  ::*    :*:
```

* = identity
: = strongly conserved
. = weakly conserved

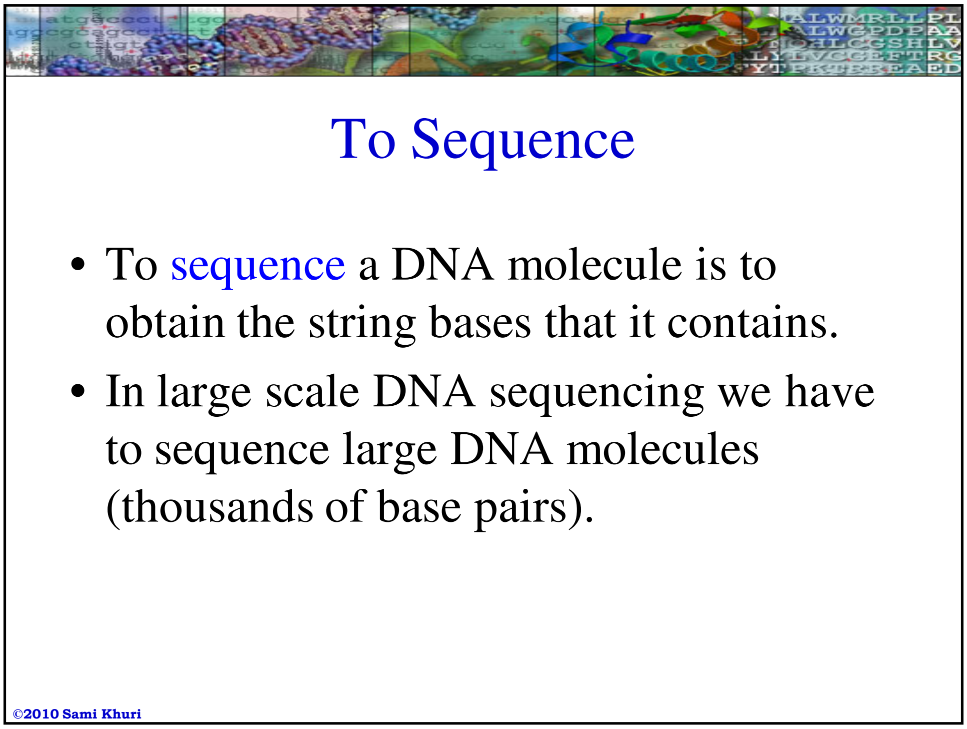
By using the same five sequences and aligning them with CLUSTALW, we get the illustrated results.



DNA Fragment Assembly

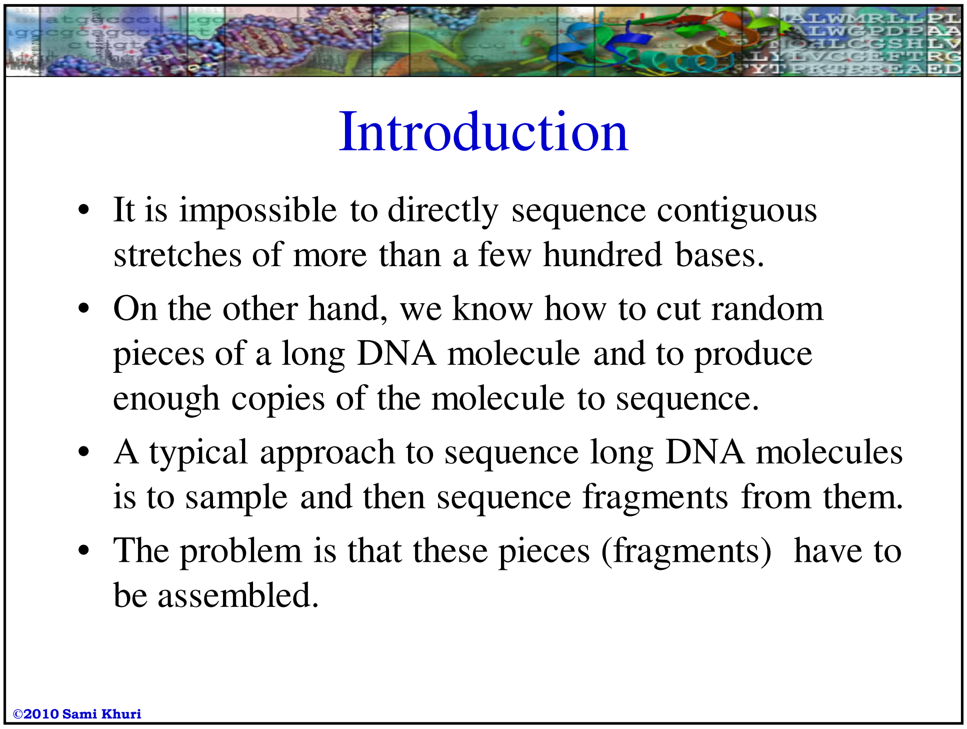

GATTGC
DNA
1. Duplicate
2. Sonicate
3. Sequence
4. Call Bases
CCGTAGCCGGGATCCCGTCC
CCCGAACAGGCTCCCGCCGTAGCCG
AAGCTTTTTTCCCGAACAGGCTCCCG
5. Layout
AAGCTTTTTTCCCGAACAGGCTCCCG
CCCGAACAGGCTCCCGCCGTAGCCG
CCGTAGCCGGGATCCCGTCC
6. Call Consensus
AAGCTTTTTTCCCGAACAGGCTCCCGCCGTAGCCGGGATCCCGTCC
AAGCTTTTTTCCCGAACAGGCTCCCG
CCCGAACAGGCTCCCGCCGTAGCCG
CCGTAGCCGGGATCCCGTCC

- **Overlap Graphs**
- **Shotgun Sequencing**
- **Repeated Regions**
- **Sequencing by Hybridization**
- **Hamiltonian Cycle**
- **Euler Path**



To Sequence

- To sequence a DNA molecule is to obtain the string bases that it contains.
- In large scale DNA sequencing we have to sequence large DNA molecules (thousands of base pairs).



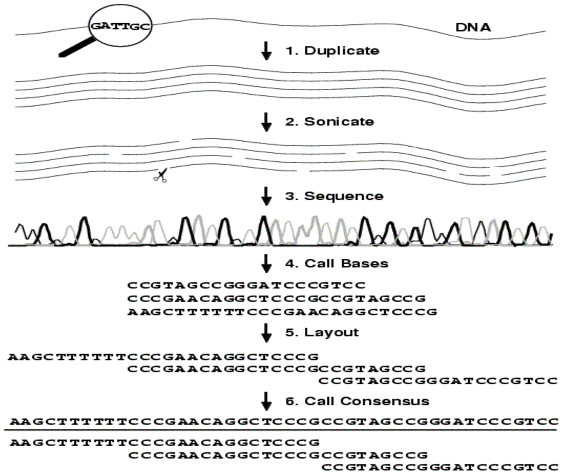
Introduction

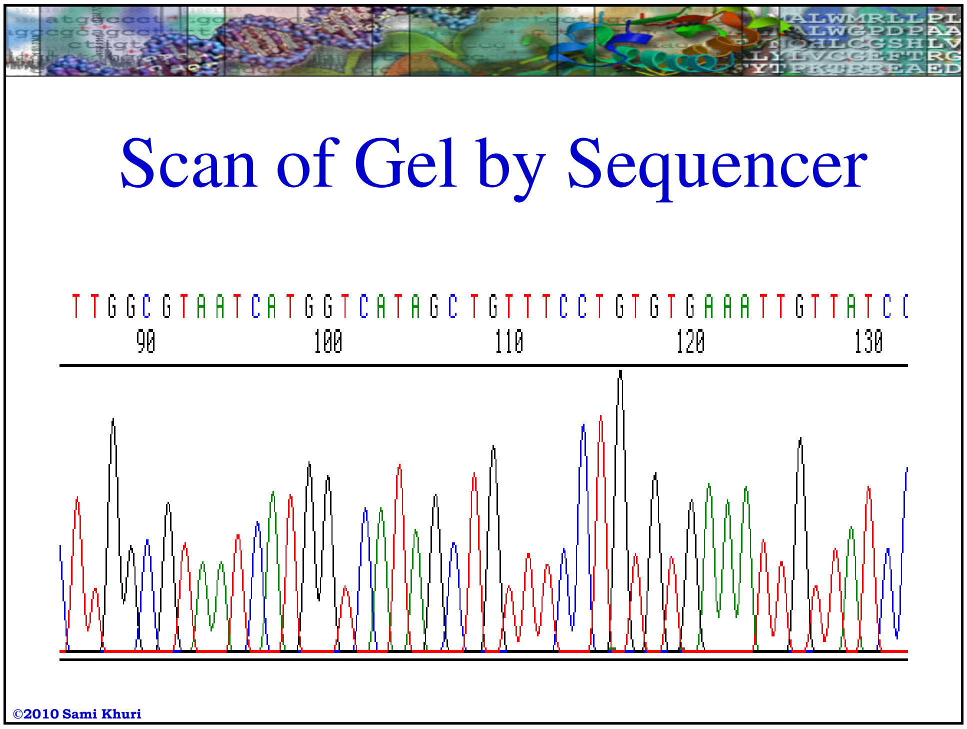
- It is impossible to directly sequence contiguous stretches of more than a few hundred bases.
- On the other hand, we know how to cut random pieces of a long DNA molecule and to produce enough copies of the molecule to sequence.
- A typical approach to sequence long DNA molecules is to sample and then sequence fragments from them.
- The problem is that these pieces (fragments) have to be assembled.

Fragment Assembly Problem

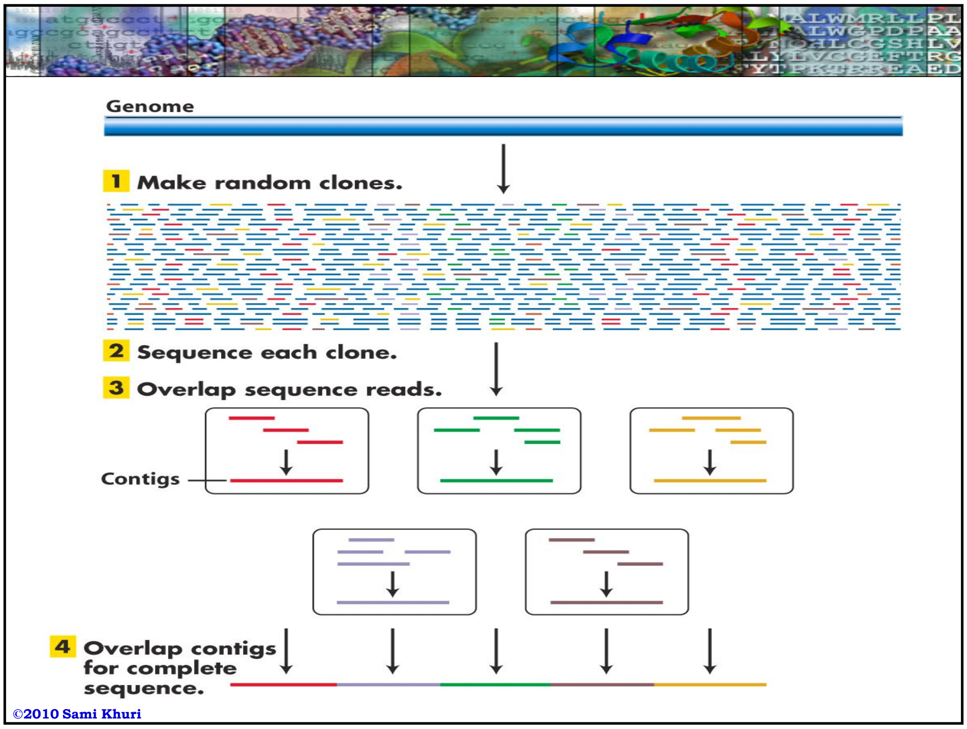
- In large scale DNA sequencing, we are given a collection of many fragments of short DNA sequences.
- The fragments are approximate substrings of a very long DNA molecule.
- The Fragment Assembly Problem consists in reconstructing the original sequence from the fragments.

Steps of Fragment Assembly


GATTGC
DNA
1. Duplicate
2. Sonicate
3. Sequence
4. Call Bases
CCGTAGCCGGGATCCCGTCC
CCCGAACAGGCTCCCGCCGTAGCCG
AAGCTTTTTTCCCGAACAGGCTCCCG
5. Layout
AAGCTTTTTTCCCGAACAGGCTCCCG
CCCGAACAGGCTCCCGCCGTAGCCG
CCGTAGCCGGGATCCCGTCC
6. Call Consensus
AAGCTTTTTTCCCGAACAGGCTCCCGCCGTAGCCGGGATCCCGTCC
AAGCTTTTTTCCCGAACAGGCTCCCG
CCCGAACAGGCTCCCGCCGTAGCCG
CCGTAGCCGGGATCCCGTCC



Scan of Gel by Sequencer



Genome

1 Make random clones.

2 Sequence each clone.

3 Overlap sequence reads.

Contigs

4 Overlap contigs for complete sequence.

Consensus Sequence Building

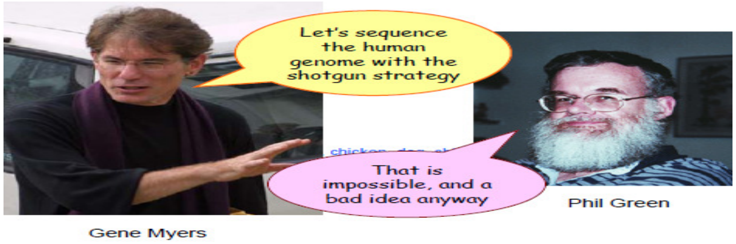
```
TAGAGCTGCTTA    CTGCT          TTACTGACTA     TATATCGAGCTAC
GTCGTAACG       AC             TTAGT          ATCGAGC
CGAGCTACTA      GCTGCTTAGTC    ATATCGAGCTAC   GAGCTGCT
AGGACGCTGCTA    GTCGT          GACGCTGCTAGT   CGAGCTAT
GTCG            CTAT           GCTGCTA        TTCGAG
TAGGACG         ATACT          GATACTAC
TCGT            TCGTAACGAT     GCTTAG
CTAT            GCGGCG
```

↓ Fragment Assembly

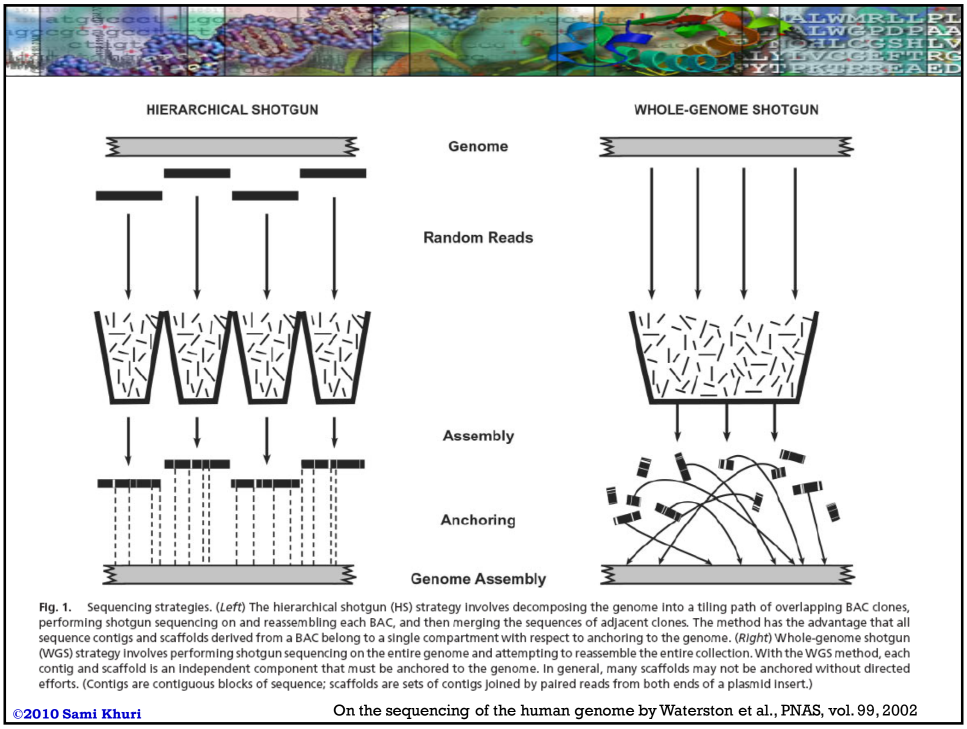
```
GCGGCG        CTAT  CGAGCTACTA    GCTGCTA       ACTAGAGCTGCTTA         ATACT
   GCGTATTC   CTAT           AGGACGCTGCTA              CTGCT  GTCGTAACG
        TTCGAG  ATCGAGC     TAGGACG     AGTTACTGACTA          GTCG    GATACT
  GGCGTATT      ATATCGAGCTAC   GACGCTGCTAGTTAC              TAGTCGTAA
GCGGCG       GCTATATCGAGCTAC                          GCTGCTTAGTC
          CGAGCTAT                                  GAGCTGCT   TCGTAACGAT
----------------------------------------------------------------------------
CGGCGATATTCGAGCTATATCGAGCTACTAGGACGCTGCTAGTTACTGACTAGAGCTGCTTAGTCGTAACGATACT
```

Original Sequence

Genome Sequencing Strategies


Let's sequence the human genome with the shotgun strategy
That is impossible, and a bad idea anyway
Gene Myers
Phil Green

- Human Genome Project: map-based strategy
  - individual clones subjected to shotgun sequencing
  - shotgun fragments then reassembled
- Celera: whole genome sequence strategy
  - shotgun sequencing



HIERARCHICAL SHOTGUN

WHOLE-GENOME SHOTGUN

Genome

Random Reads

Assembly

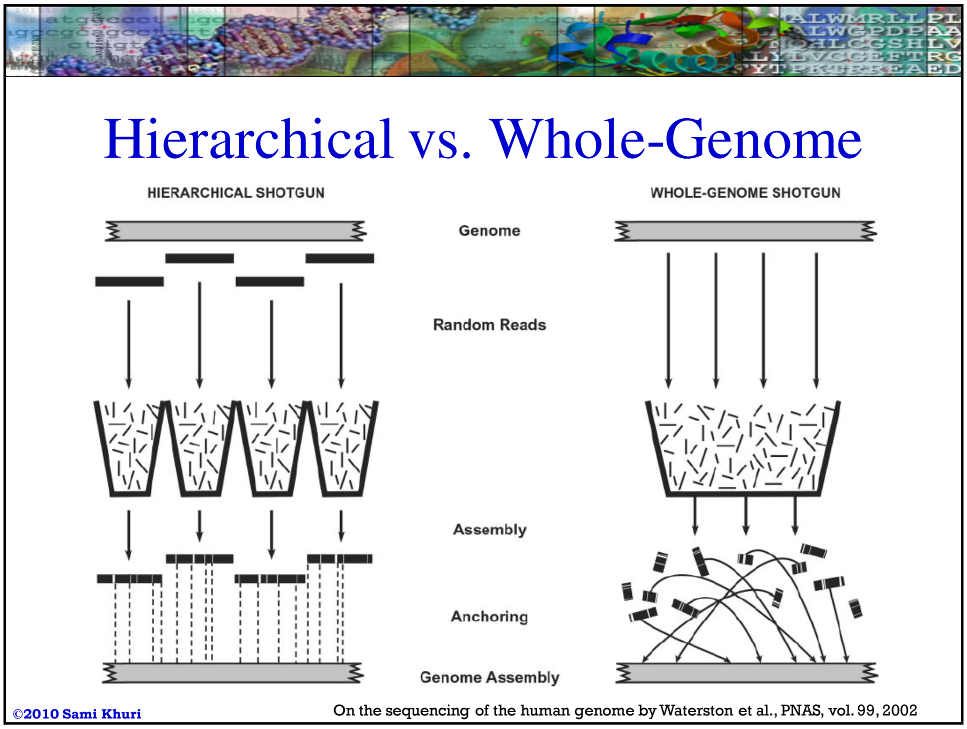
Anchoring

Genome Assembly

Fig. 1. Sequencing strategies. (*Left*) The hierarchical shotgun (HS) strategy involves decomposing the genome into a tiling path of overlapping BAC clones, performing shotgun sequencing on and reassembling each BAC, and then merging the sequences of adjacent clones. The method has the advantage that all sequence contigs and scaffolds derived from a BAC belong to a single compartment with respect to anchoring to the genome. (*Right*) Whole-genome shotgun (WGS) strategy involves performing shotgun sequencing on the entire genome and attempting to reassemble the entire collection. With the WGS method, each contig and scaffold is an independent component that must be anchored to the genome. In general, many scaffolds may not be anchored without directed efforts. (Contigs are contiguous blocks of sequence; scaffolds are sets of contigs joined by paired reads from both ends of a plasmid insert.)

©2010 Sami Khuri

On the sequencing of the human genome by Waterston et al., PNAS, vol. 99, 2002



Hierarchical vs. Whole-Genome

HIERARCHICAL SHOTGUN

WHOLE-GENOME SHOTGUN

Genome

Random Reads

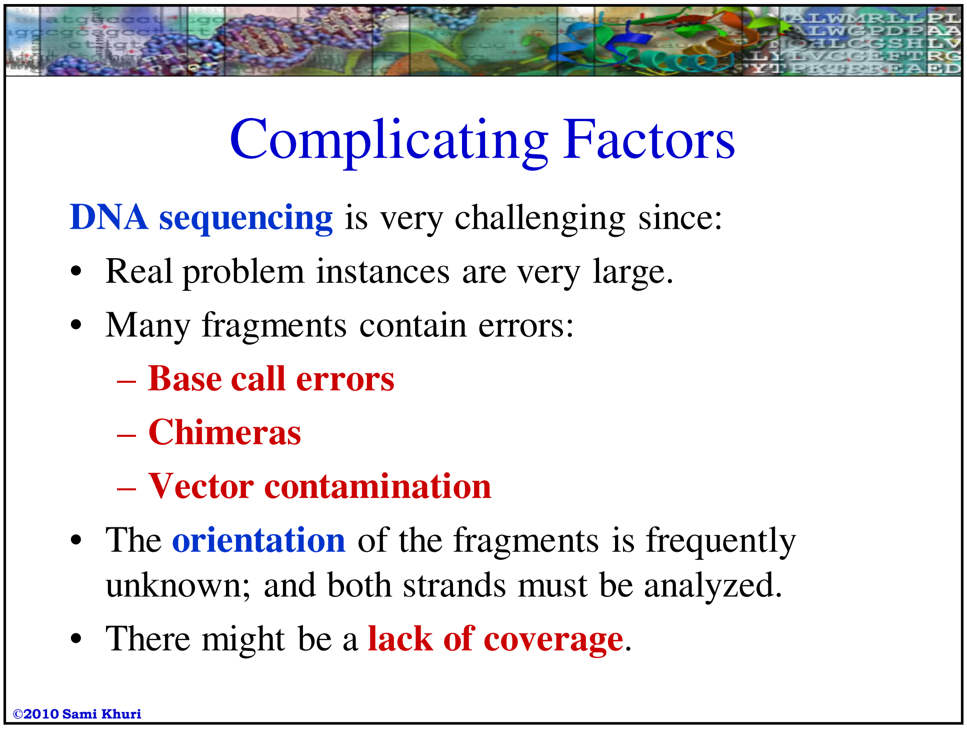
Assembly

Anchoring

Genome Assembly

©2010 Sami Khuri

On the sequencing of the human genome by Waterston et al., PNAS, vol. 99, 2002

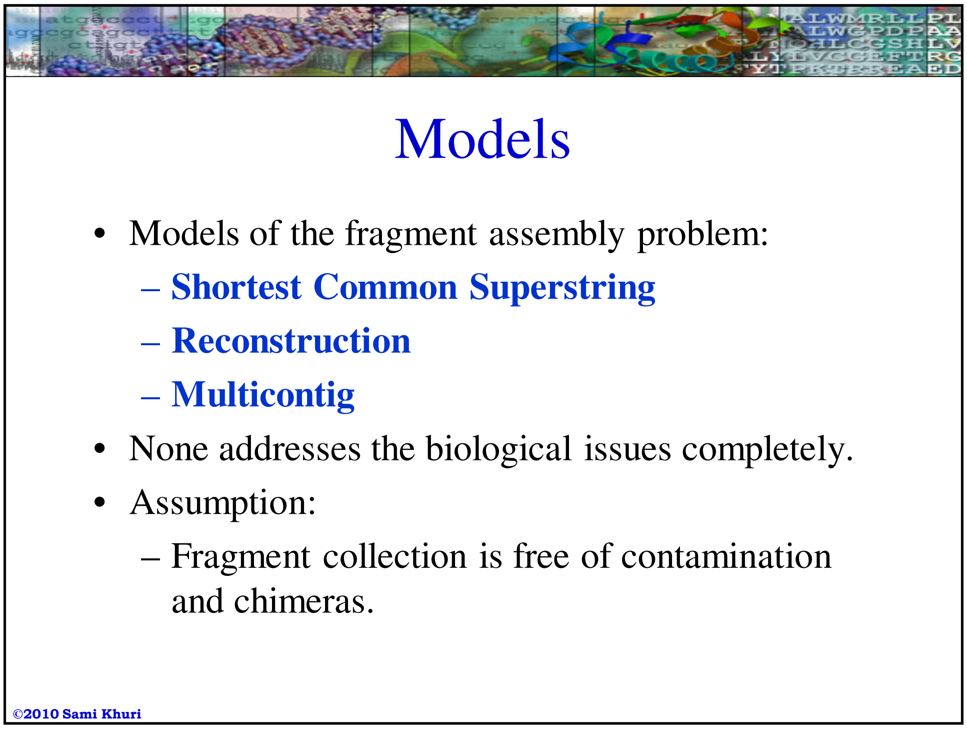


Complicating Factors

DNA sequencing is very challenging since:

- Real problem instances are very large.
- Many fragments contain errors:
  - **Base call errors**
  - **Chimeras**
  - **Vector contamination**
- The **orientation** of the fragments is frequently unknown; and both strands must be analyzed.
- There might be a **lack of coverage**.

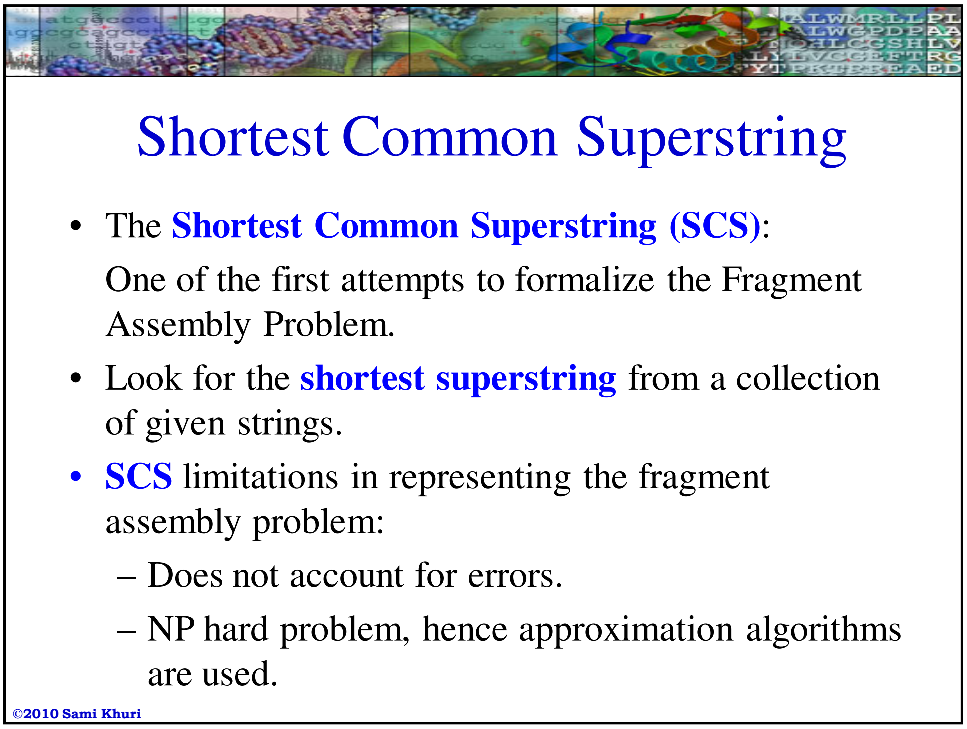
©2010 Sami Khuri



Models

- Models of the fragment assembly problem:
  - **Shortest Common Superstring**
  - **Reconstruction**
  - **Multicontig**
- None addresses the biological issues completely.
- Assumption:
  - Fragment collection is free of contamination and chimeras.

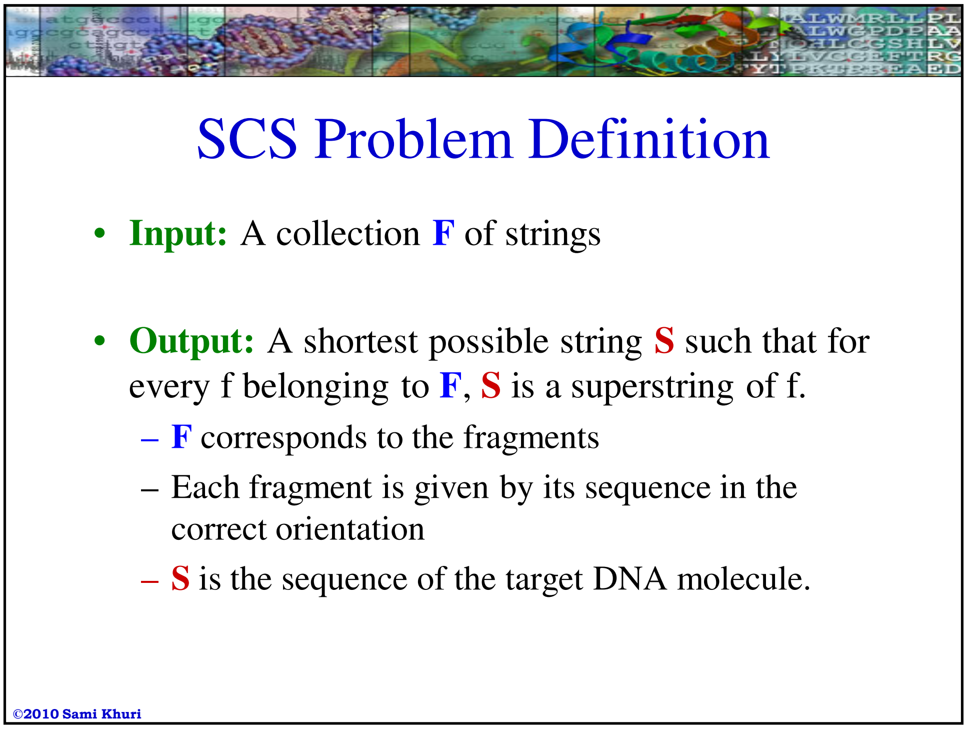
©2010 Sami Khuri



Shortest Common Superstring

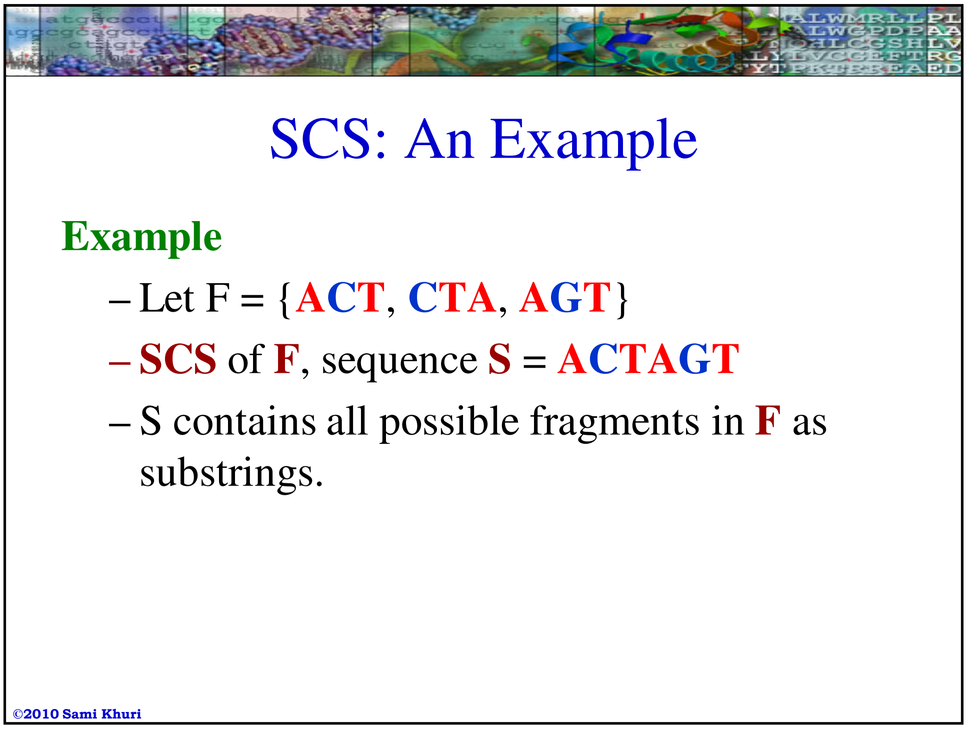
- The **Shortest Common Superstring (SCS)**:

  One of the first attempts to formalize the Fragment Assembly Problem.
- Look for the **shortest superstring** from a collection of given strings.
- **SCS** limitations in representing the fragment assembly problem:
  - Does not account for errors.
  - NP hard problem, hence approximation algorithms are used.



SCS Problem Definition

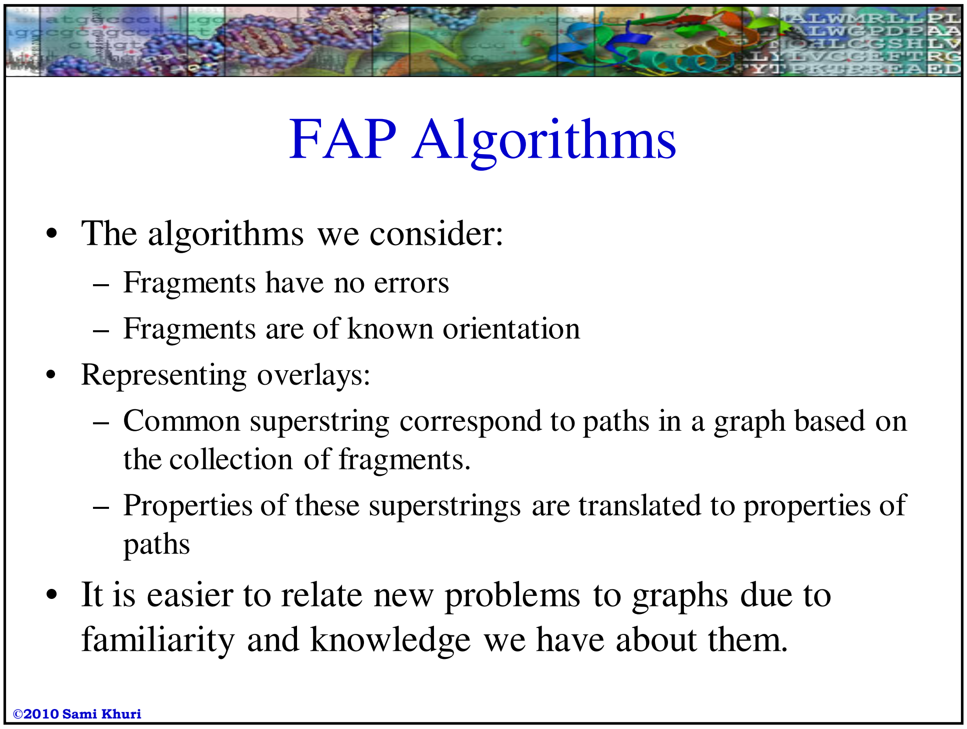
- **Input:** A collection **F** of strings
- **Output:** A shortest possible string **S** such that for every f belonging to **F**, **S** is a superstring of f.
  - **F** corresponds to the fragments
  - Each fragment is given by its sequence in the correct orientation
  - **S** is the sequence of the target DNA molecule.



SCS: An Example

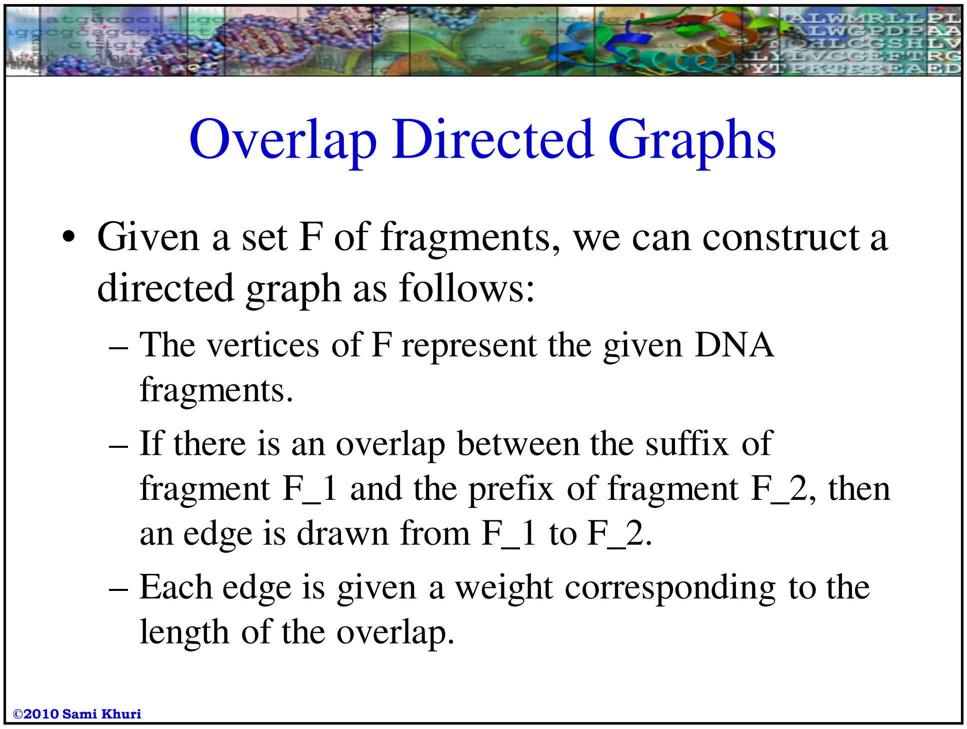
Example

- Let F = {**ACT**, **CTA**, **AGT**}
- **SCS** of **F**, sequence **S** = **ACTAGT**
- S contains all possible fragments in **F** as substrings.



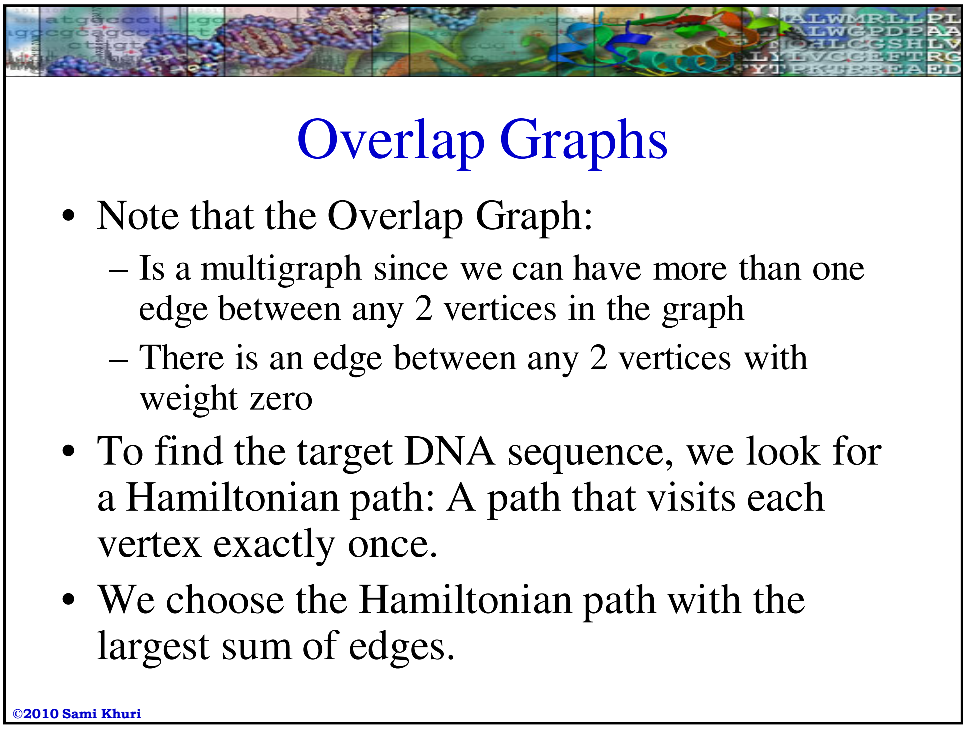
FAP Algorithms

- The algorithms we consider:
  - Fragments have no errors
  - Fragments are of known orientation
- Representing overlays:
  - Common superstring correspond to paths in a graph based on the collection of fragments.
  - Properties of these superstrings are translated to properties of paths
- It is easier to relate new problems to graphs due to familiarity and knowledge we have about them.



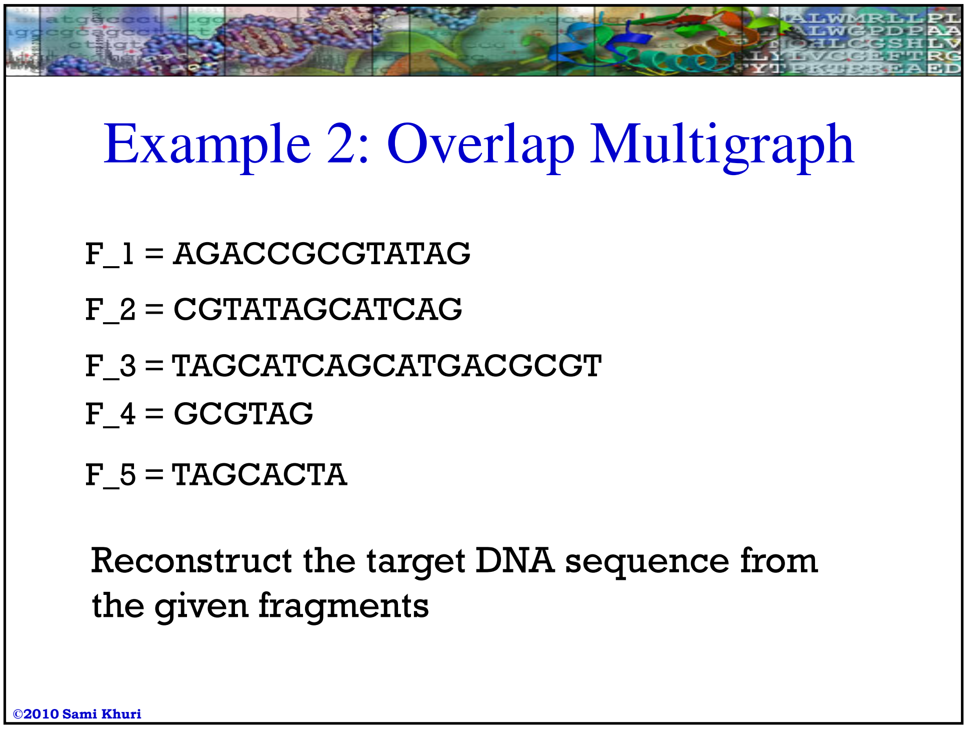
Overlap Directed Graphs

- Given a set F of fragments, we can construct a directed graph as follows:
  - The vertices of F represent the given DNA fragments.
  - If there is an overlap between the suffix of fragment F_1 and the prefix of fragment F_2, then an edge is drawn from F_1 to F_2.
  - Each edge is given a weight corresponding to the length of the overlap.



Overlap Graphs

- Note that the Overlap Graph:
  - Is a multigraph since we can have more than one edge between any 2 vertices in the graph
  - There is an edge between any 2 vertices with weight zero
- To find the target DNA sequence, we look for a Hamiltonian path: A path that visits each vertex exactly once.
- We choose the Hamiltonian path with the largest sum of edges.



Example 2: Overlap Multigraph

F_1 = AGACCGCGTATAG

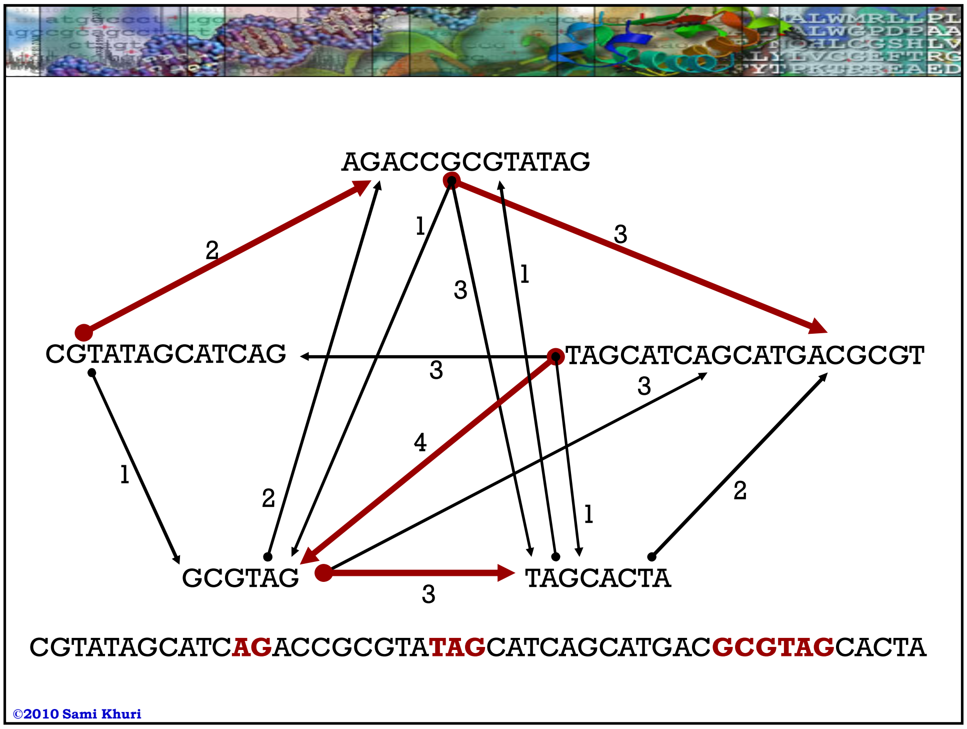
F_2 = CGTATAGCATCAG

F_3 = TAGCATCAGCATGACGCGT

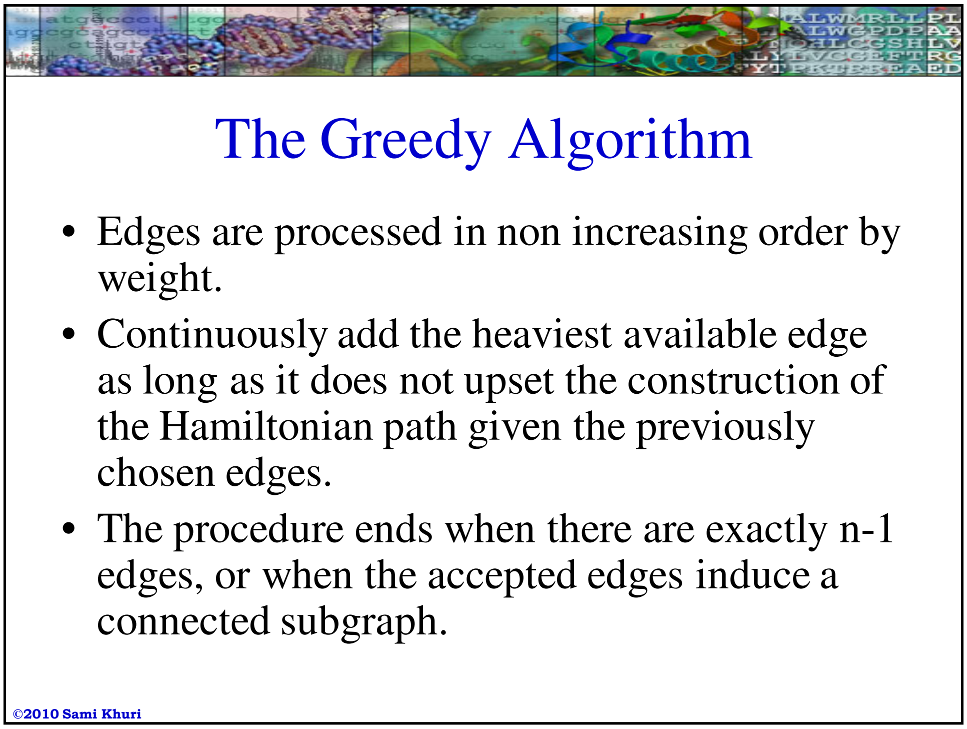
F_4 = GCGTAG

F_5 = TAGCACTA

Reconstruct the target DNA sequence from the given fragments

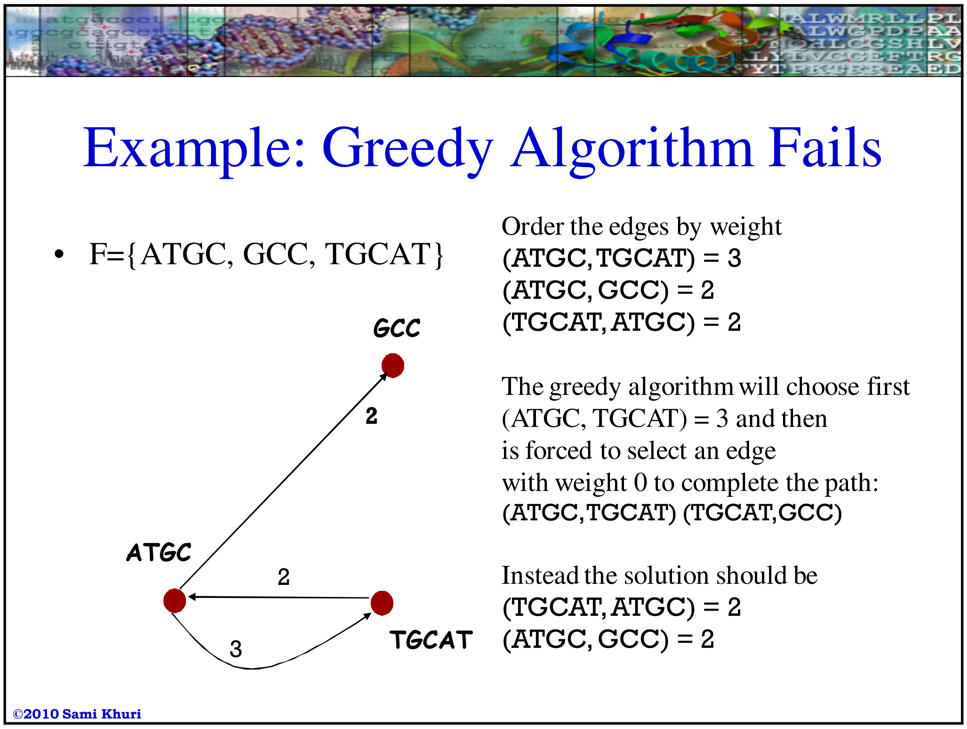


CGTATAGCATC**AG**ACCGCGTA**TAG**CATCAGCATGAC**GCGTAG**CACTA



The Greedy Algorithm

- Edges are processed in non increasing order by weight.
- Continuously add the heaviest available edge as long as it does not upset the construction of the Hamiltonian path given the previously chosen edges.
- The procedure ends when there are exactly n-1 edges, or when the accepted edges induce a connected subgraph.



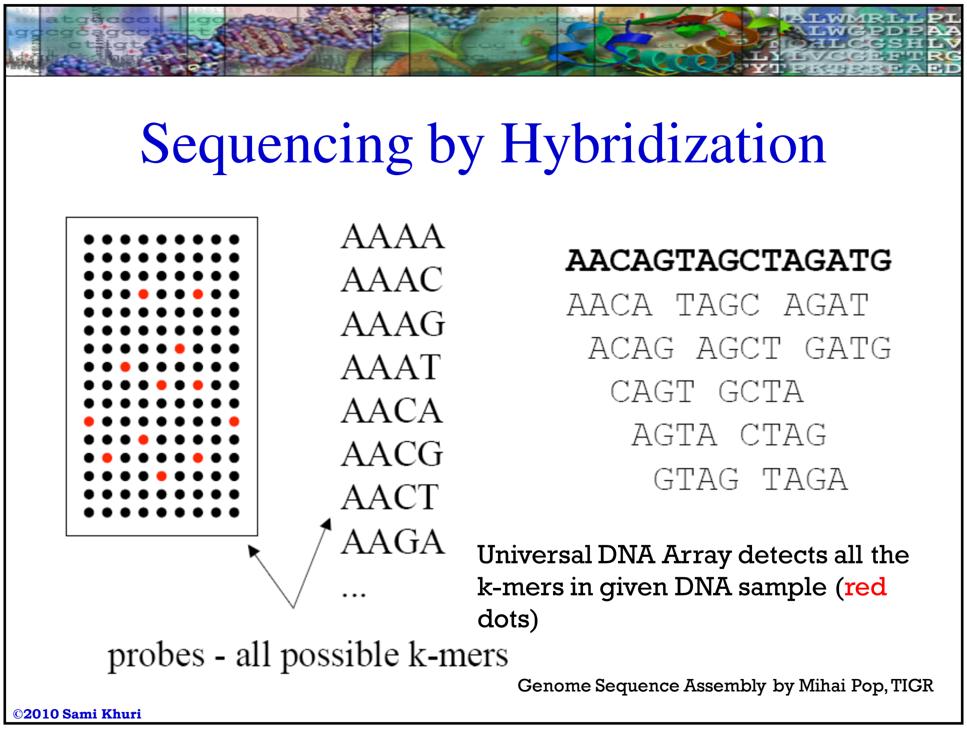
Example: Greedy Algorithm Fails

- F={ATGC, GCC, TGCAT}

Order the edges by weight
(ATGC, TGCAT) = 3
(ATGC, GCC) = 2
(TGCAT, ATGC) = 2

The greedy algorithm will choose first (ATGC, TGCAT) = 3 and then is forced to select an edge with weight 0 to complete the path:
(ATGC,TGCAT) (TGCAT,GCC)

Instead the solution should be
(TGCAT, ATGC) = 2
(ATGC, GCC) = 2



Sequencing by Hybridization

AAAA
AAAC
AAAG
AAAT
AACA
AACG
AACT
AAGA
…

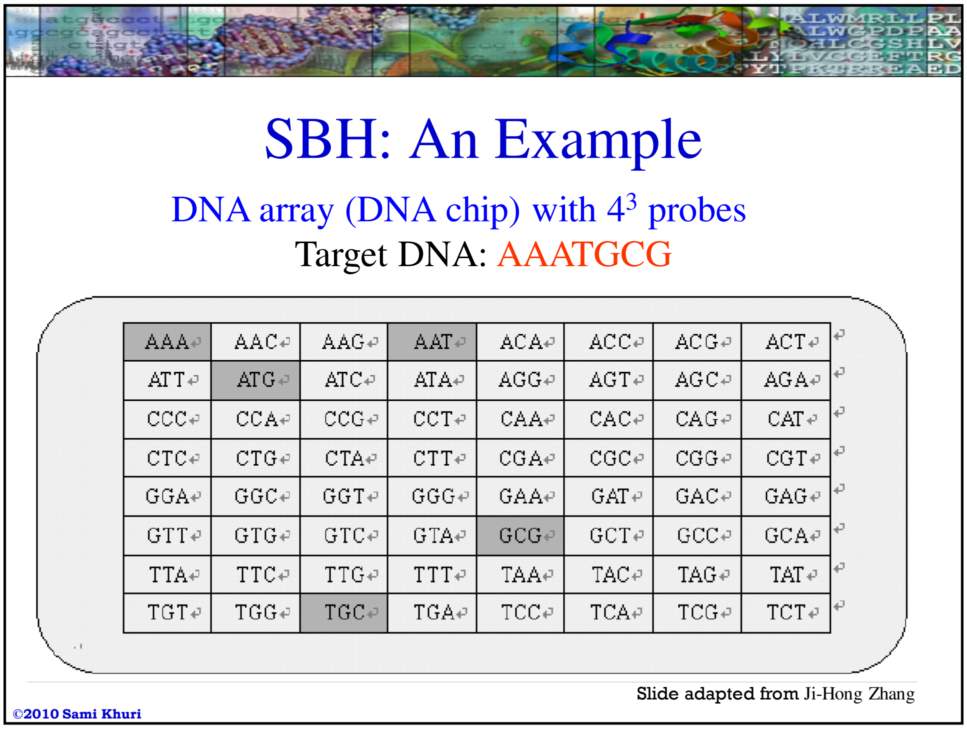
probes - all possible k-mers

AACAGTAGCTAGATG

```
AACA TAGC AGAT
 ACAG AGCT GATG
  CAGT GCTA
   AGTA CTAG
    GTAG TAGA
```

Universal DNA Array detects all the k-mers in given DNA sample (red dots)

Genome Sequence Assembly by Mihai Pop, TIGR



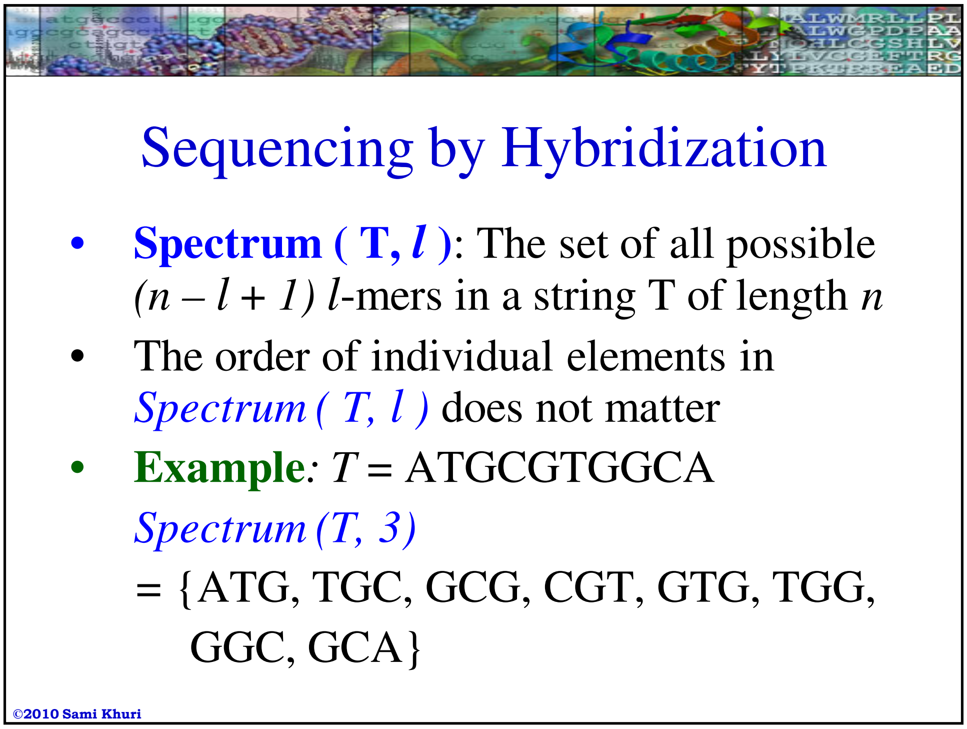
SBH: An Example

DNA array (DNA chip) with 4^3 probes

Target DNA: AAATGCG

AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT
ATT	ATG	ATC	ATA	AGG	AGT	AGC	AGA
CCC	CCA	CCG	CCT	CAA	CAC	CAG	CAT
CTC	CTG	CTA	CTT	CGA	CGC	CGG	CGT
GGA	GGC	GGT	GGG	GAA	GAT	GAC	GAG
GTT	GTG	GTC	GTA	GCG	GCT	GCC	GCA
TTA	TTC	TTG	TTT	TAA	TAC	TAG	TAT
TGT	TGG	TGC	TGA	TCC	TCA	TCG	TCT

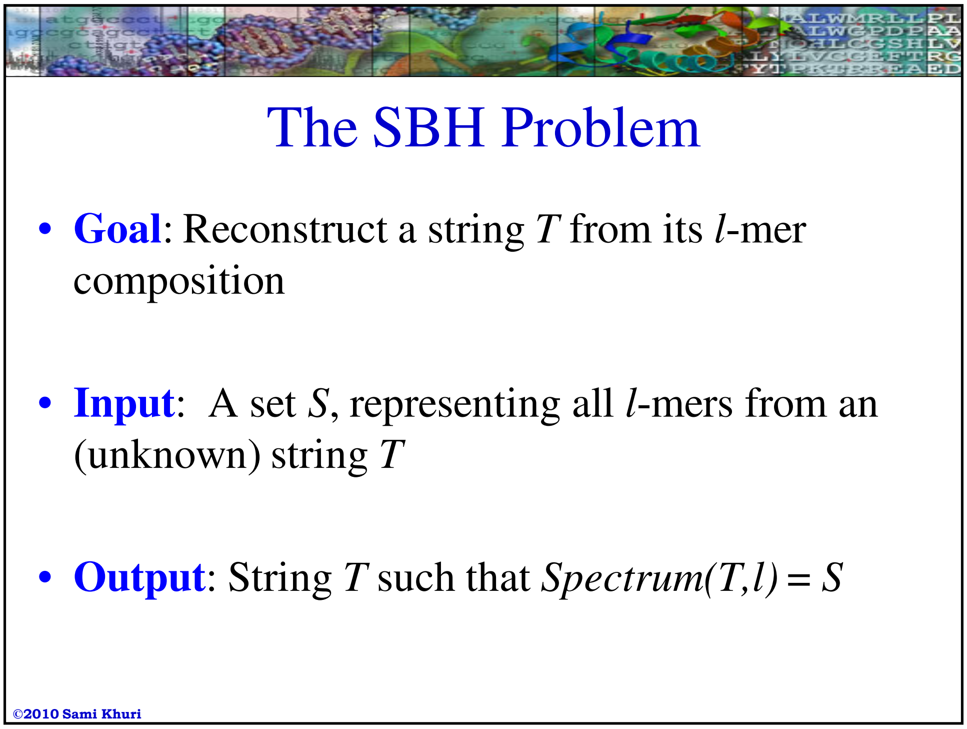
Slide adapted from Ji-Hong Zhang



Sequencing by Hybridization

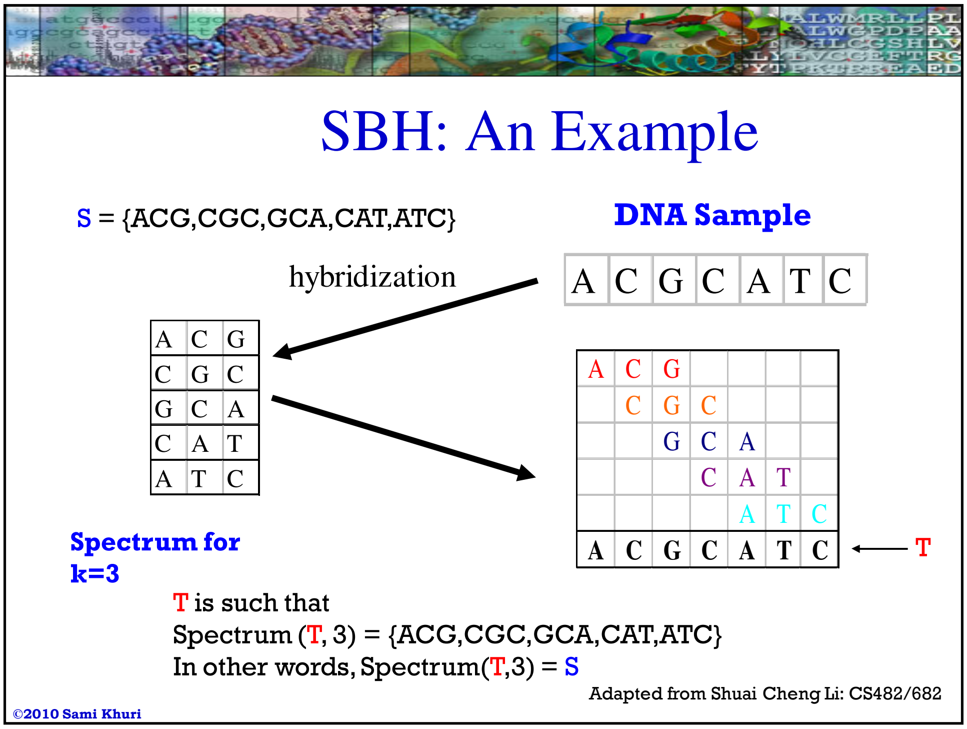
- **Spectrum (T, *l*)**: The set of all possible *(n – l + 1) l*-mers in a string T of length *n*
- The order of individual elements in *Spectrum (T, l)* does not matter
- **Example***: T* = ATGCGTGGCA

  Spectrum (T, 3)
  = {ATG, TGC, GCG, CGT, GTG, TGG,
  GGC, GCA}



The SBH Problem

- **Goal**: Reconstruct a string *T* from its *l*-mer composition
- **Input**: A set *S*, representing all *l*-mers from an (unknown) string *T*
- **Output**: String *T* such that *Spectrum(T,l)* = *S*



SBH: An Example

S = {ACG,CGC,GCA,CAT,ATC}

DNA Sample

A	C	G	C	A	T	C

hybridization

A	C	G
C	G	C
G	C	A
C	A	T
A	T	C

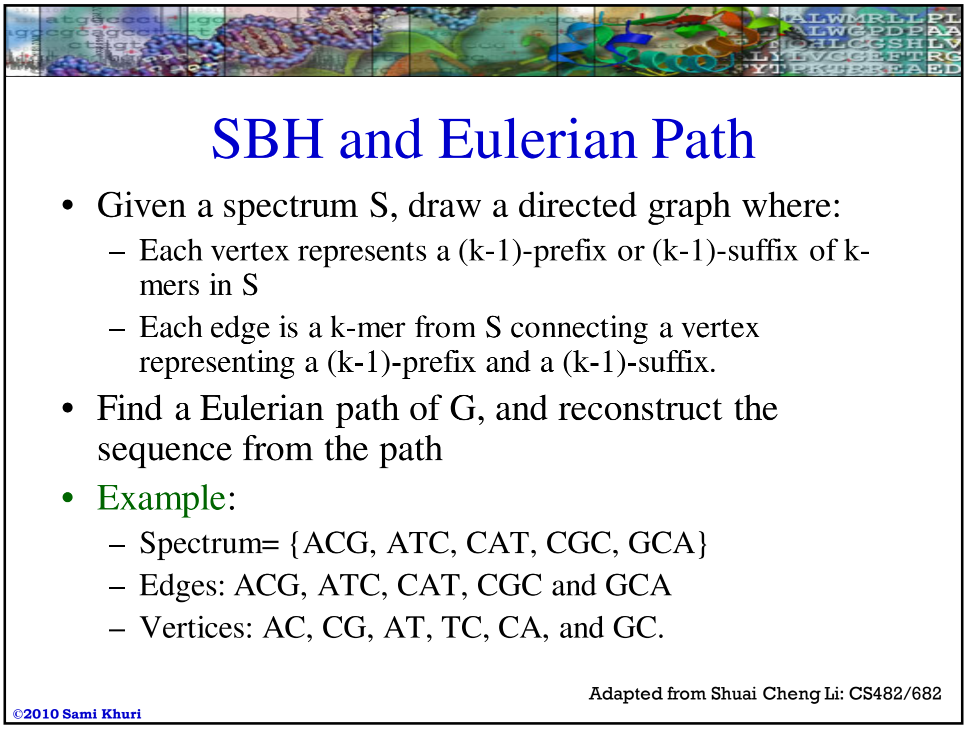
A	C	G					
	C	G	C				
		G	C	A			
			C	A	T		
				A	T	C	
A	**C**	**G**	**C**	**A**	**T**	**C**	← **T**

Spectrum for k=3

T is such that
Spectrum (T, 3) = {ACG,CGC,GCA,CAT,ATC}
In other words, Spectrum(T,3) = S

Adapted from Shuai Cheng Li: CS482/682

©2010 Sami Khuri

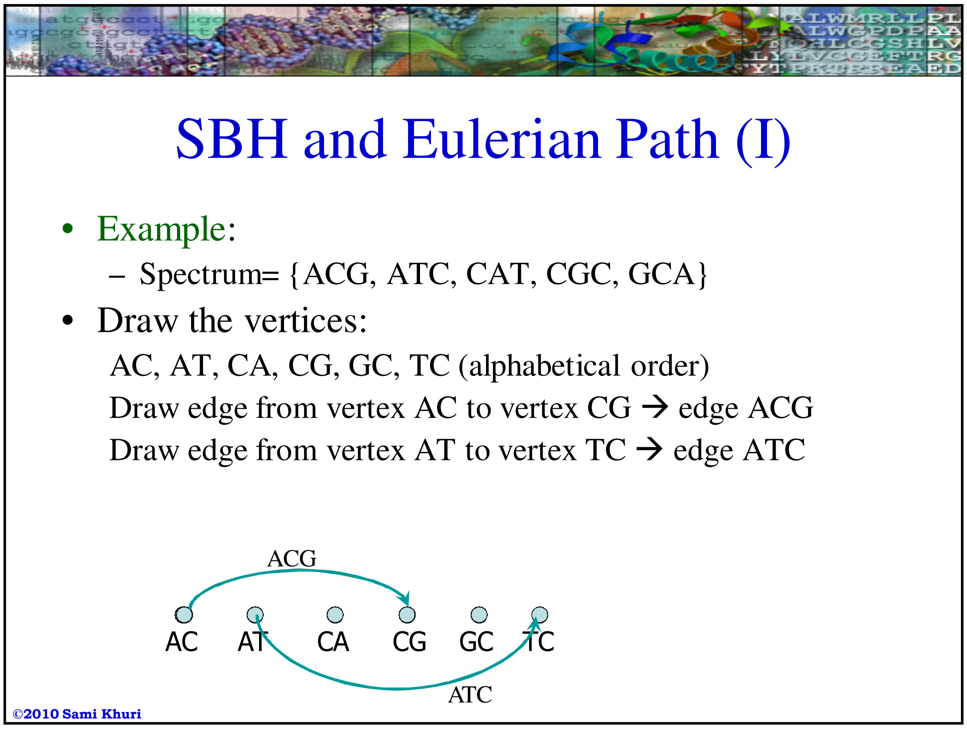


SBH and Eulerian Path

- Given a spectrum S, draw a directed graph where:
  - Each vertex represents a (k-1)-prefix or (k-1)-suffix of k-mers in S
  - Each edge is a k-mer from S connecting a vertex representing a (k-1)-prefix and a (k-1)-suffix.
- Find a Eulerian path of G, and reconstruct the sequence from the path
- Example:
  - Spectrum= {ACG, ATC, CAT, CGC, GCA}
  - Edges: ACG, ATC, CAT, CGC and GCA
  - Vertices: AC, CG, AT, TC, CA, and GC.

Adapted from Shuai Cheng Li: CS482/682

©2010 Sami Khuri



SBH and Eulerian Path (I)

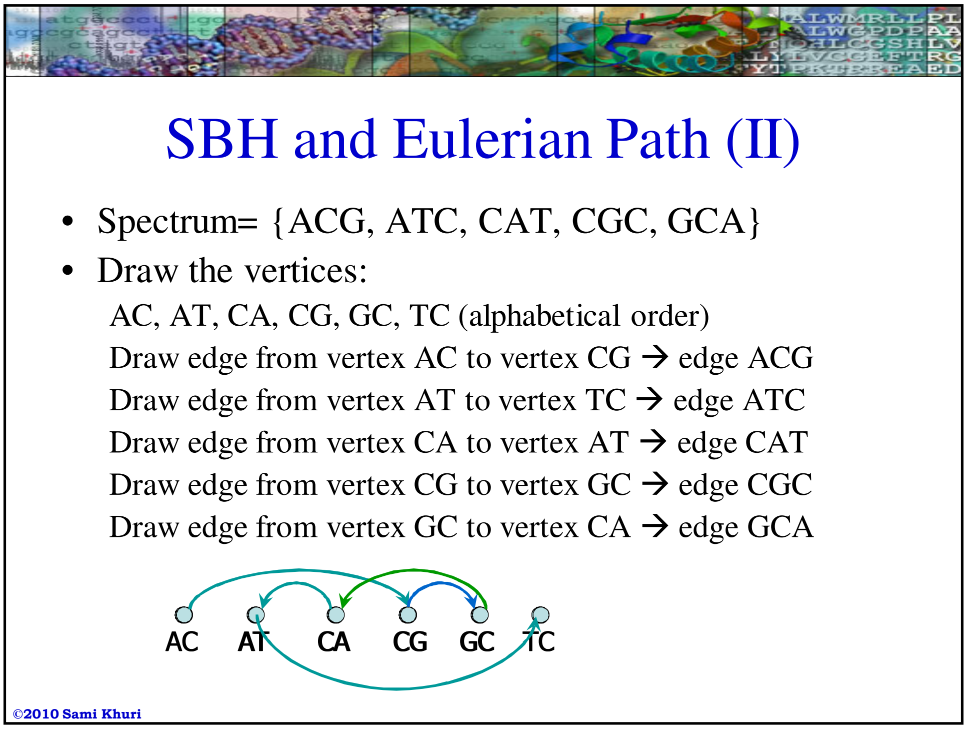
- Example:
  - Spectrum= {ACG, ATC, CAT, CGC, GCA}
- Draw the vertices:

  AC, AT, CA, CG, GC, TC (alphabetical order)

  Draw edge from vertex AC to vertex CG → edge ACG

  Draw edge from vertex AT to vertex TC → edge ATC

ACG
AC
AT
CA
CG
GC
TC
ATC



SBH and Eulerian Path (II)

- Spectrum= {ACG, ATC, CAT, CGC, GCA}
- Draw the vertices:

  AC, AT, CA, CG, GC, TC (alphabetical order)

  Draw edge from vertex AC to vertex CG → edge ACG

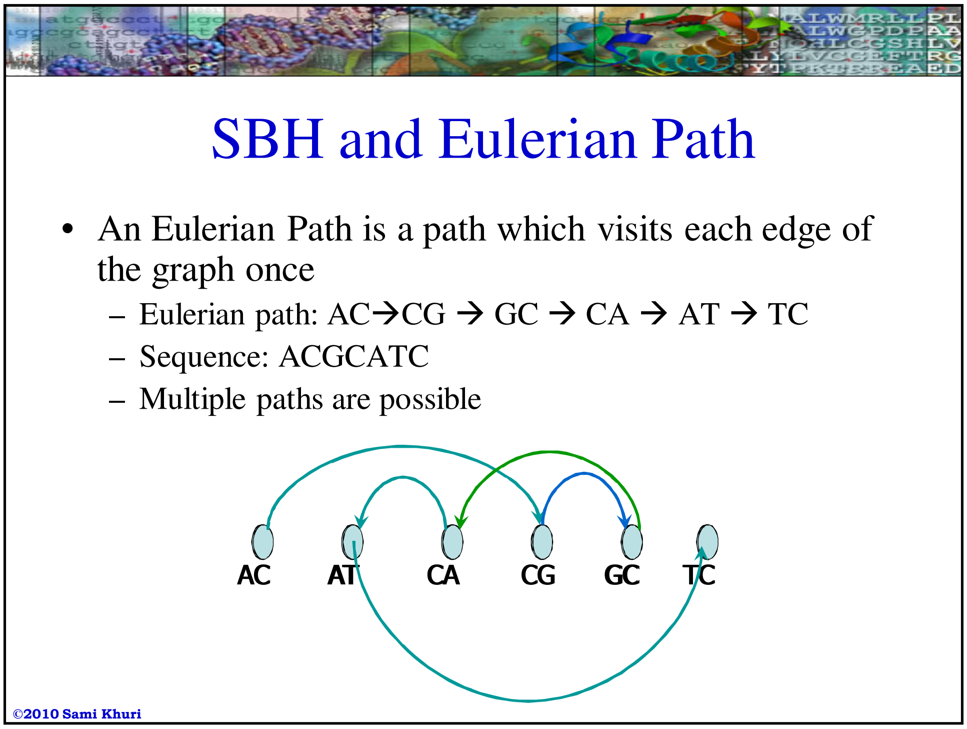
  Draw edge from vertex AT to vertex TC → edge ATC

  Draw edge from vertex CA to vertex AT → edge CAT

  Draw edge from vertex CG to vertex GC → edge CGC

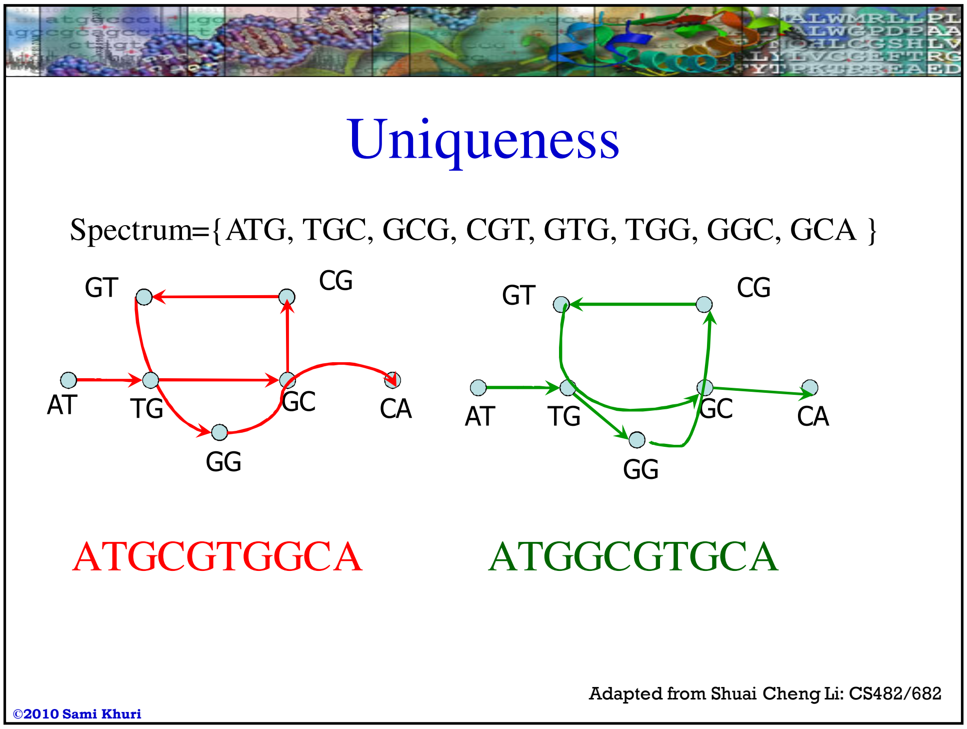
  Draw edge from vertex GC to vertex CA → edge GCA

AC
AT
CA
CG
GC
TC



SBH and Eulerian Path

- An Eulerian Path is a path which visits each edge of the graph once
  - Eulerian path: AC→CG → GC → CA → AT → TC
  - Sequence: ACGCATC
  - Multiple paths are possible

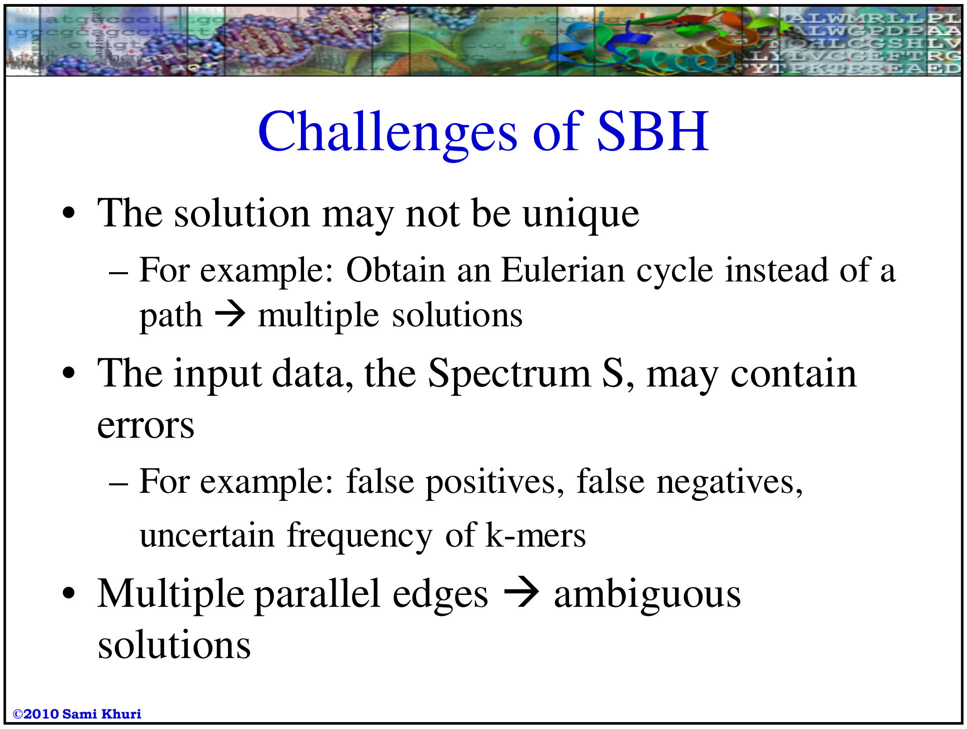


Uniqueness

Spectrum={ATG, TGC, GCG, CGT, GTG, TGG, GGC, GCA }

ATGCGTGGCA ATGGCGTGCA

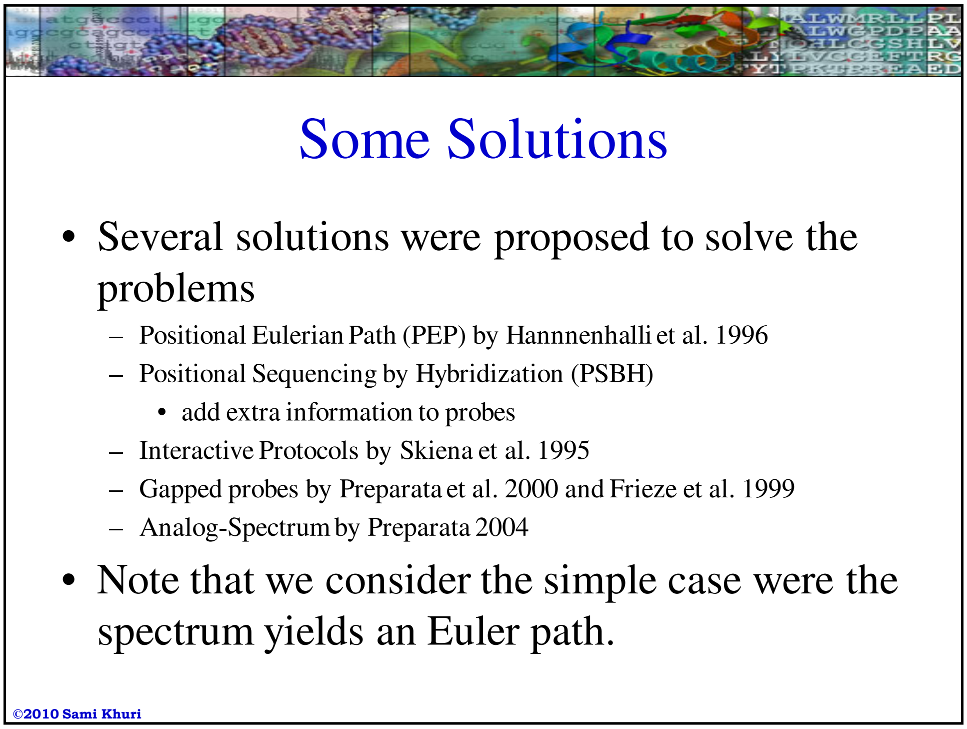
Adapted from Shuai Cheng Li: CS482/682



Challenges of SBH

- The solution may not be unique
  - For example: Obtain an Eulerian cycle instead of a path → multiple solutions
- The input data, the Spectrum S, may contain errors
  - For example: false positives, false negatives, uncertain frequency of k-mers
- Multiple parallel edges → ambiguous solutions

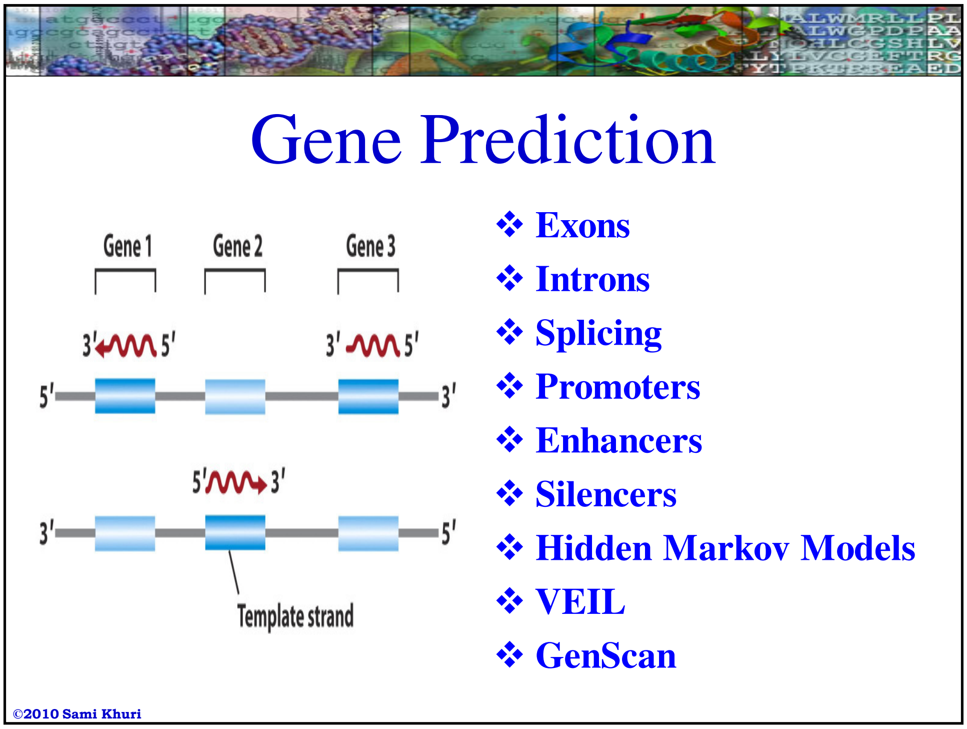
©2010 Sami Khuri



Some Solutions

- Several solutions were proposed to solve the problems
  - Positional Eulerian Path (PEP) by Hannnenhalli et al. 1996
  - Positional Sequencing by Hybridization (PSBH)
    - add extra information to probes
  - Interactive Protocols by Skiena et al. 1995
  - Gapped probes by Preparata et al. 2000 and Frieze et al. 1999
  - Analog-Spectrum by Preparata 2004
- Note that we consider the simple case were the spectrum yields an Euler path.

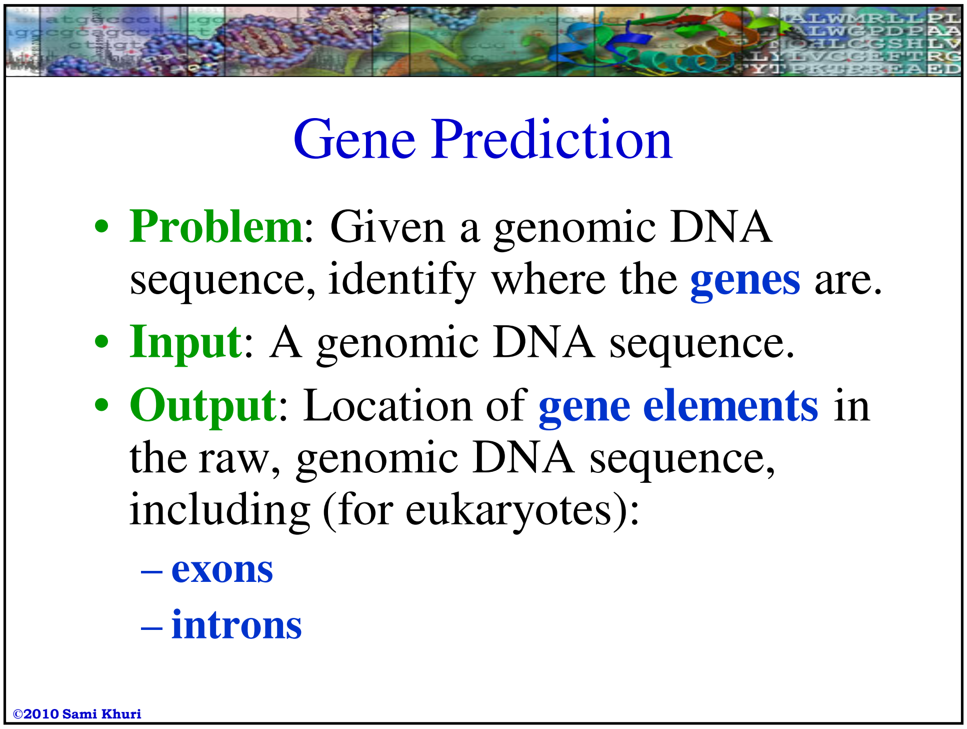
©2010 Sami Khuri



Gene Prediction

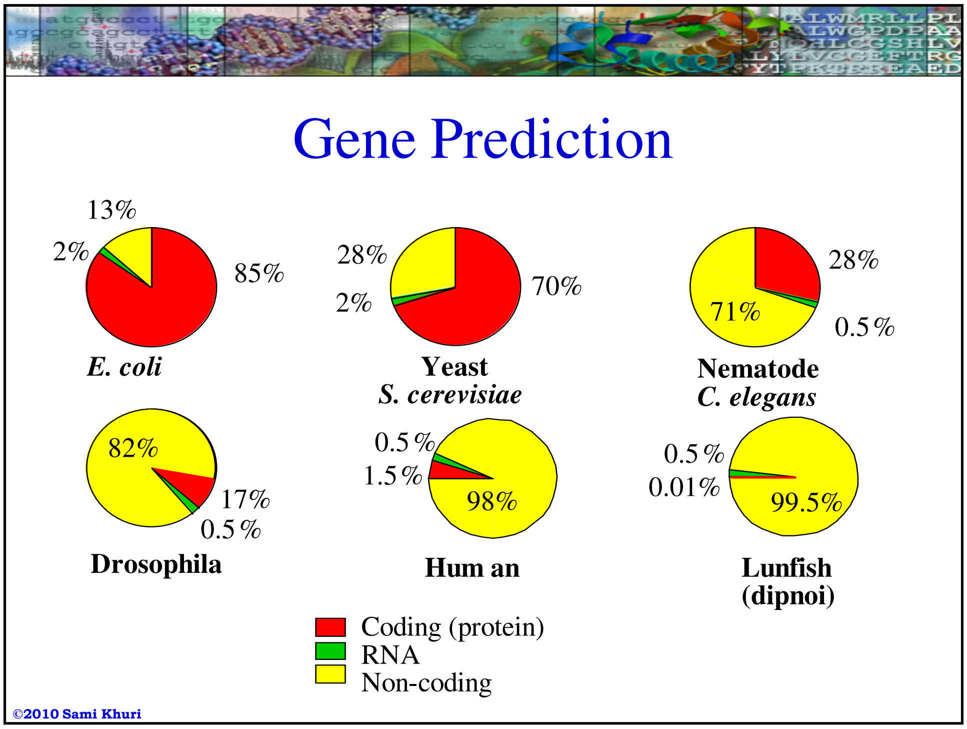
Gene 1
Gene 2
Gene 3
3′
5′
3′
5′
5′
3′
5′
3′
3′
5′
Template strand

- ❖ Exons
- ❖ Introns
- ❖ Splicing
- ❖ Promoters
- ❖ Enhancers
- ❖ Silencers
- ❖ Hidden Markov Models
- ❖ VEIL
- ❖ GenScan



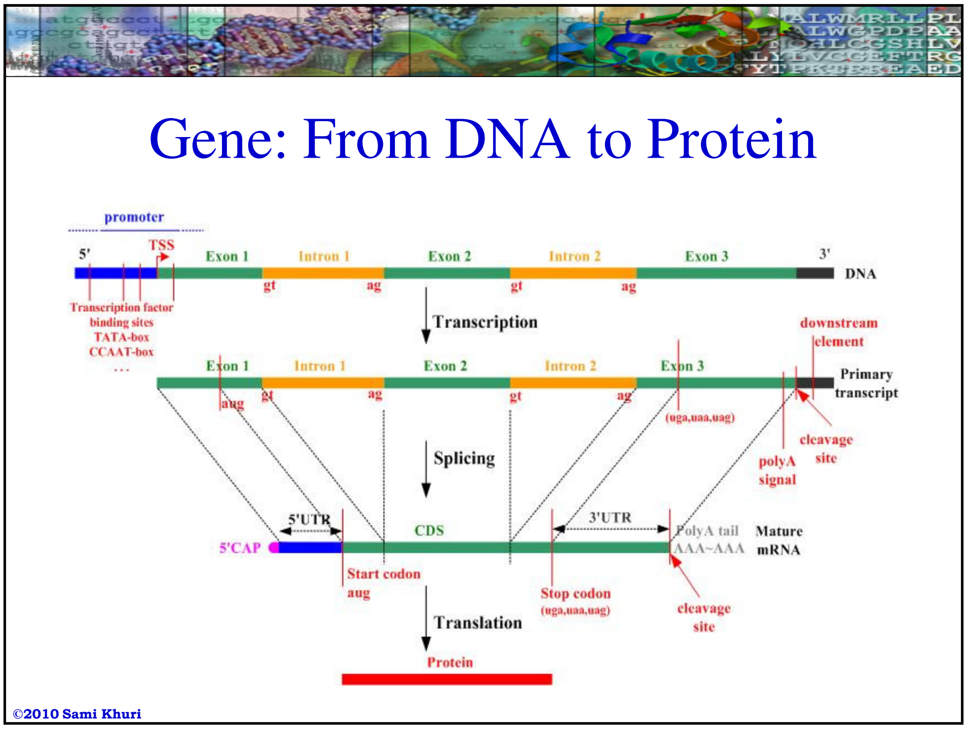
Gene Prediction

- **Problem**: Given a genomic DNA sequence, identify where the **genes** are.
- **Input**: A genomic DNA sequence.
- **Output**: Location of **gene elements** in the raw, genomic DNA sequence, including (for eukaryotes):
  - **exons**
  - **introns**

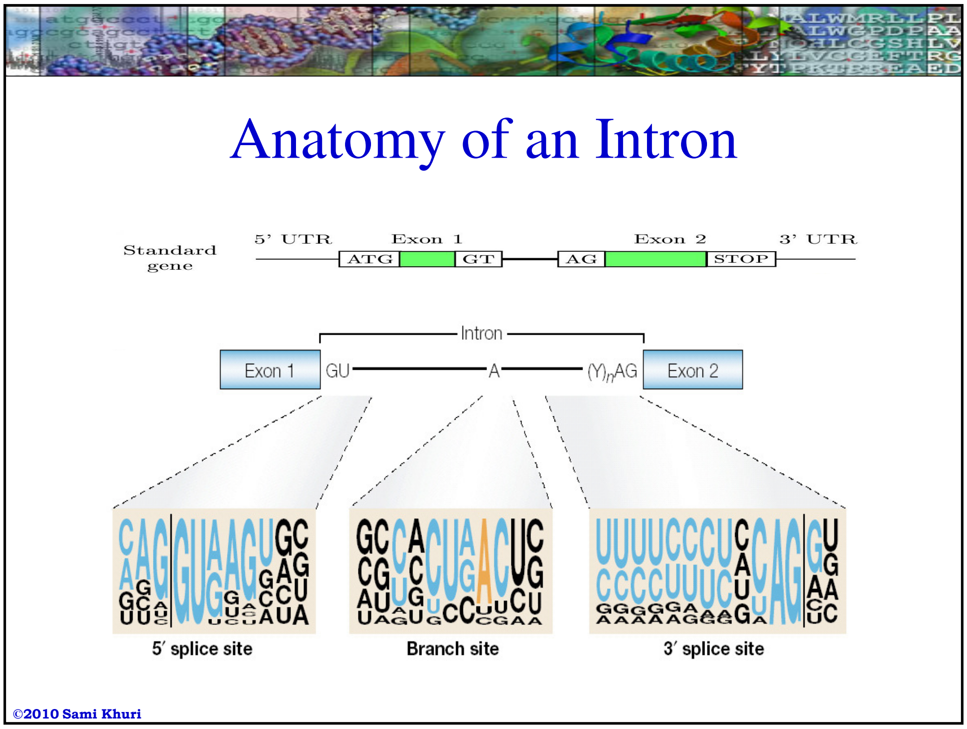


Gene Prediction

🟥 Coding (protein)
🟩 RNA
🟨 Non-coding

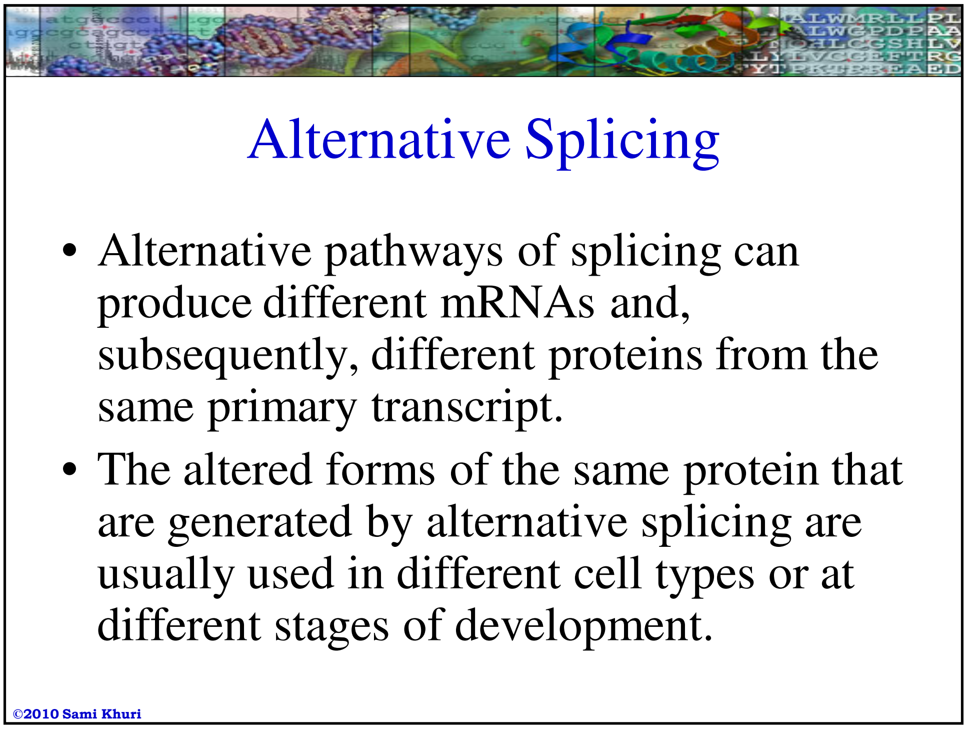


Gene: From DNA to Protein



Anatomy of an Intron

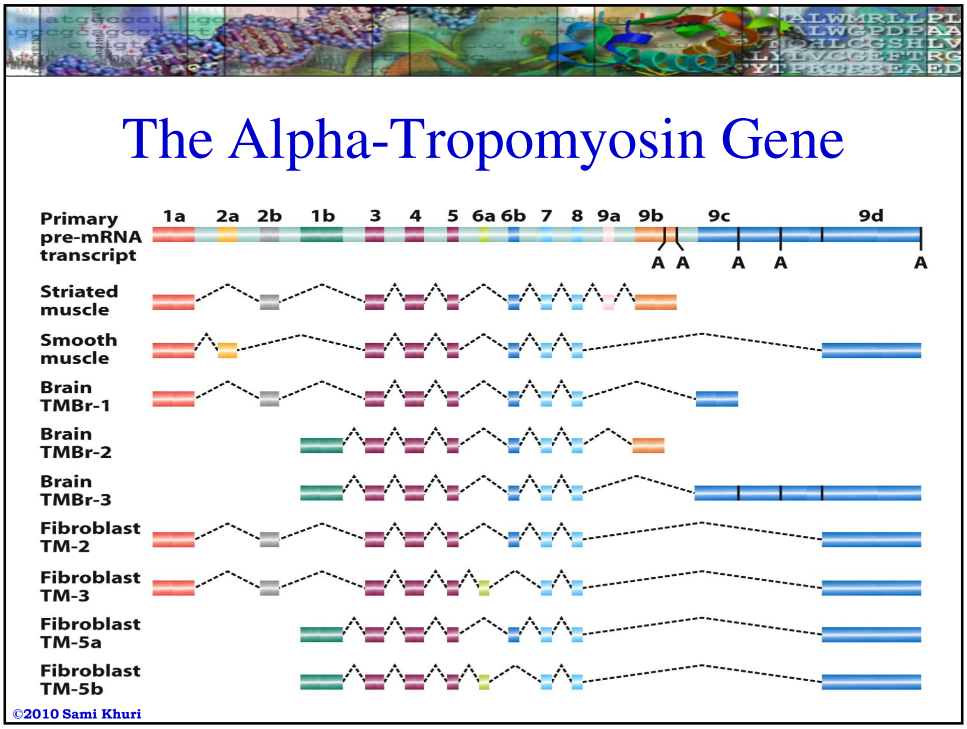
©2010 Sami Khuri



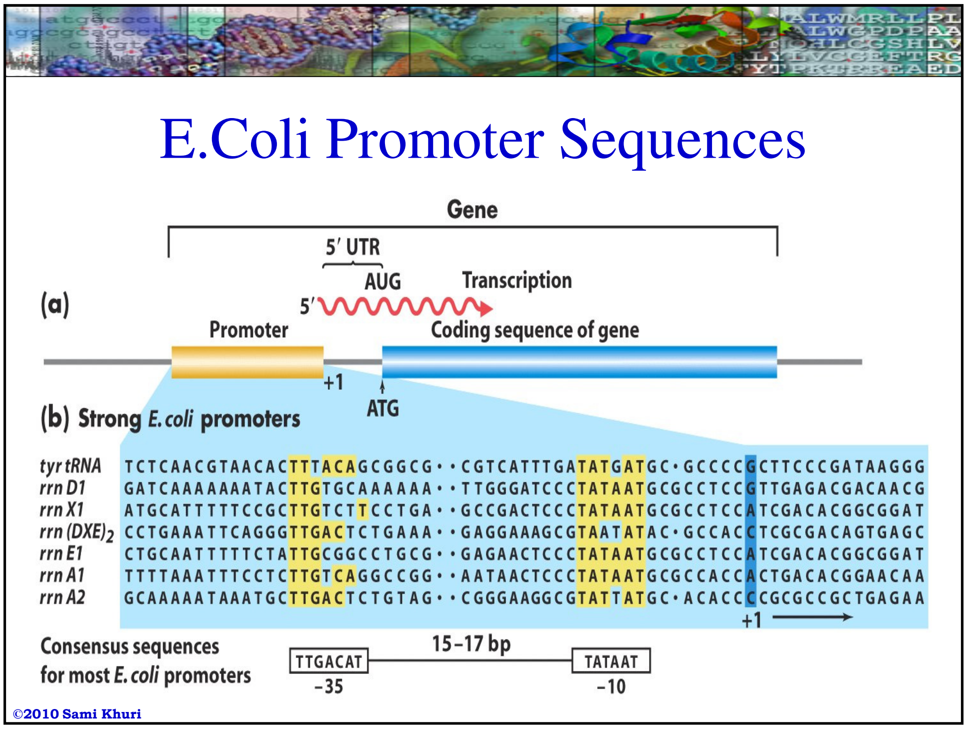
Alternative Splicing

- Alternative pathways of splicing can produce different mRNAs and, subsequently, different proteins from the same primary transcript.
- The altered forms of the same protein that are generated by alternative splicing are usually used in different cell types or at different stages of development.

©2010 Sami Khuri



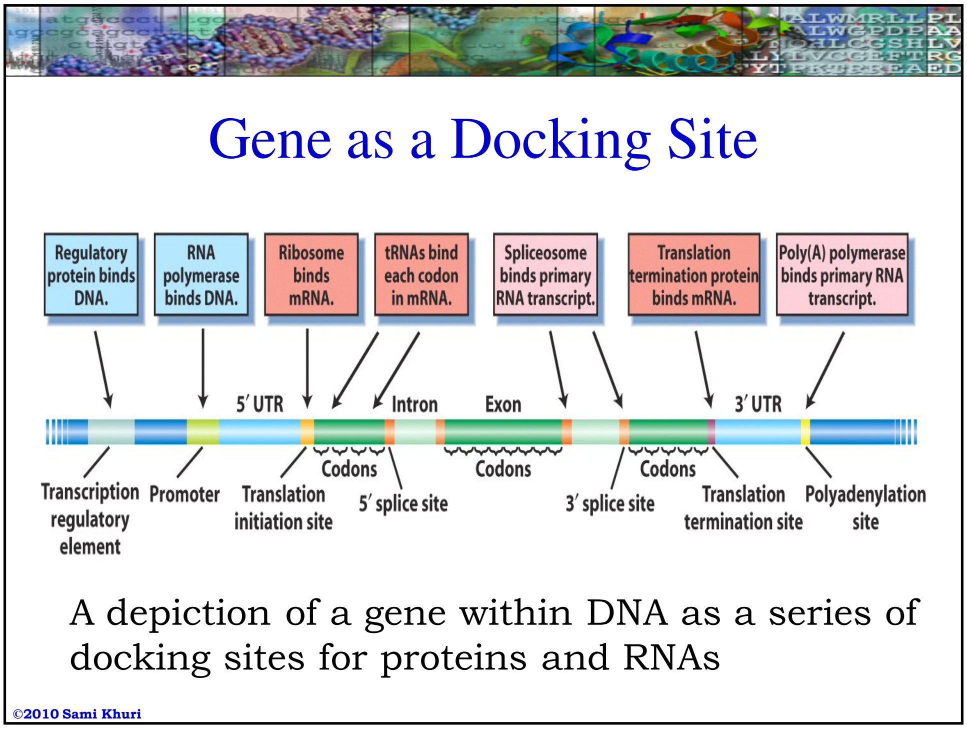
The Alpha-Tropomyosin Gene



E.Coli Promoter Sequences

```
tyr tRNA    TCTCAACGTAACACTTTACAGCGGCG••CGTCATTTGATATGATGC•GCCCCGCTTCCCGATAAGGG
rrn D1      GATCAAAAAAATACTTGTGCAAAAAA••TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
rrn X1      ATGCATTTTTCCGCTTGTCTTCCTGA••GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn (DXE)2  CCTGAAATTCAGGGTTGACTCTGAAA••GAGGAAAGCGTAATATAC•GCCACCTCGCGACAGTGAGC
rrn E1      CTGCAATTTTTCTATTGCGGCCTGCG••GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn A1      TTTTAAATTTCCTCTTGTCAGGCCGG••AATAACTCCCTATAATGCGCCACCACTGACACGGAACAA
rrn A2      GCAAAAATAAATGCTTGACTCTGTAG••CGGGAAGGCGTATTATGC•ACACCCCGCGCCGCTGAGAA
```

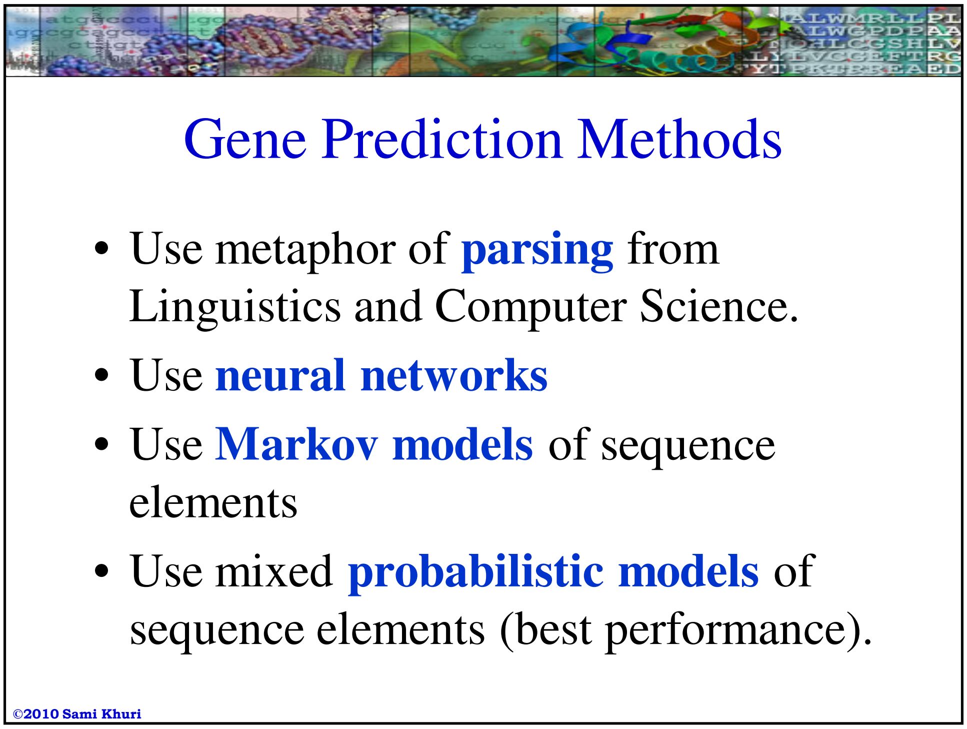
Consensus sequences for most *E. coli* promoters: TTGACAT (−35) — 15–17 bp — TATAAT (−10)



Gene as a Docking Site

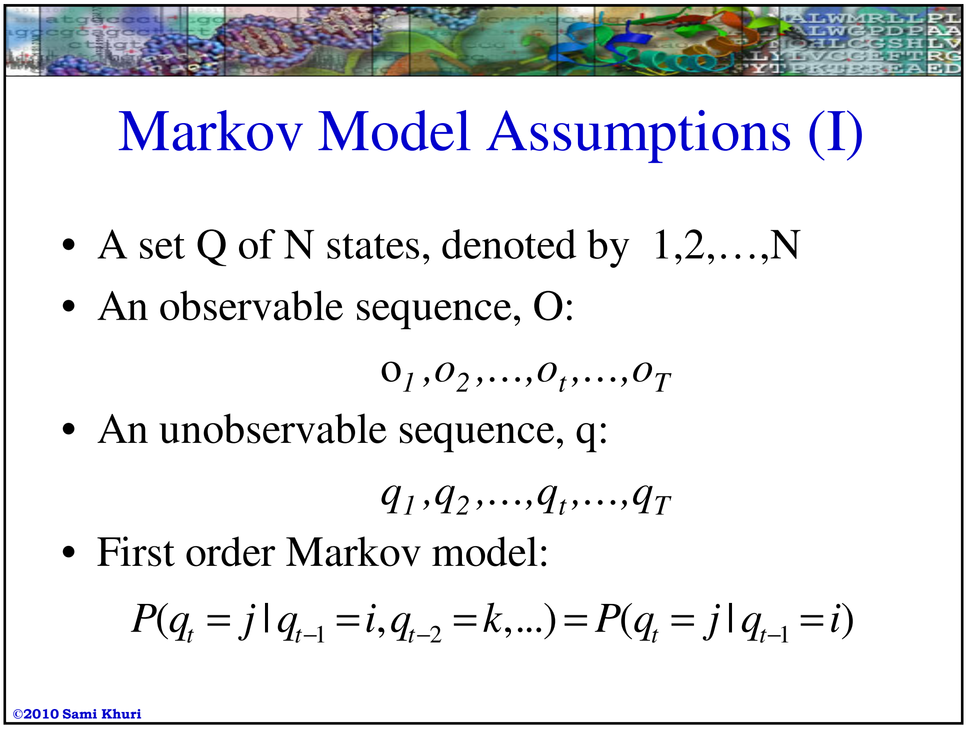
Regulatory protein binds DNA.
RNA polymerase binds DNA.
Ribosome binds mRNA.
tRNAs bind each codon in mRNA.
Spliceosome binds primary RNA transcript.
Translation termination protein binds mRNA.
Poly(A) polymerase binds primary RNA transcript.
5′ UTR
Intron
Exon
3′ UTR
Codons
Codons
Codons
Transcription regulatory element
Promoter
Translation initiation site
5′ splice site
3′ splice site
Translation termination site
Polyadenylation site

A depiction of a gene within DNA as a series of docking sites for proteins and RNAs



Gene Prediction Methods

- Use metaphor of **parsing** from Linguistics and Computer Science.
- Use **neural networks**
- Use **Markov models** of sequence elements
- Use mixed **probabilistic models** of sequence elements (best performance).



Markov Model Assumptions (I)

- A set Q of N states, denoted by 1,2,…,N
- An observable sequence, O:

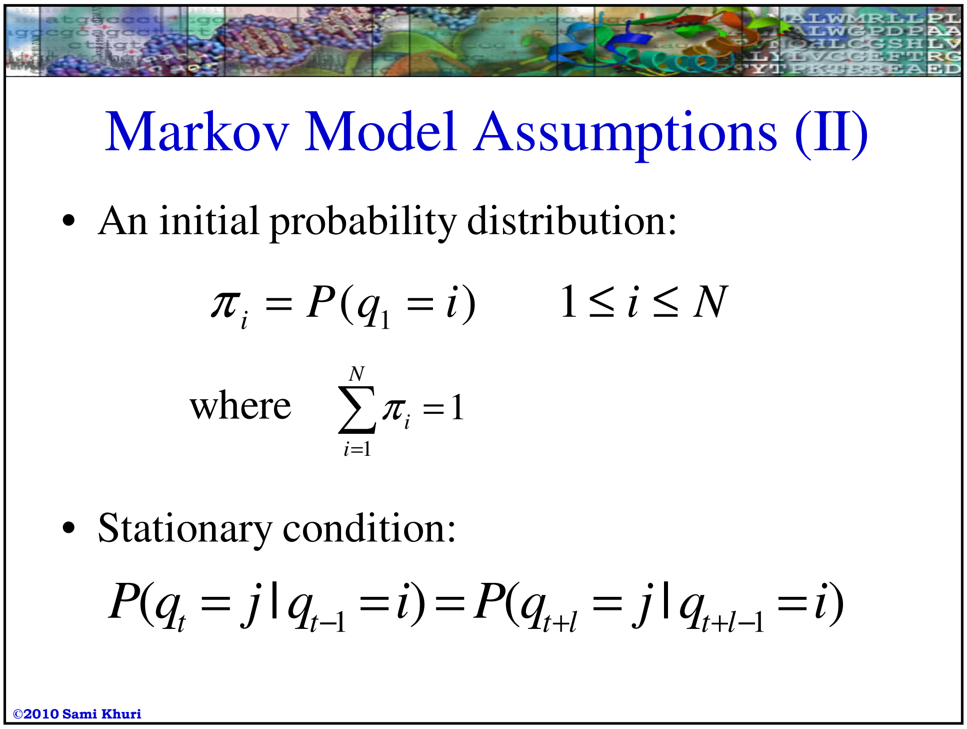
$$o_1, o_2, \ldots, o_t, \ldots, o_T$$

- An unobservable sequence, q:

$$q_1, q_2, \ldots, q_t, \ldots, q_T$$

- First order Markov model:

$$P(q_t = j \mid q_{t-1} = i, q_{t-2} = k, \ldots) = P(q_t = j \mid q_{t-1} = i)$$



Markov Model Assumptions (II)

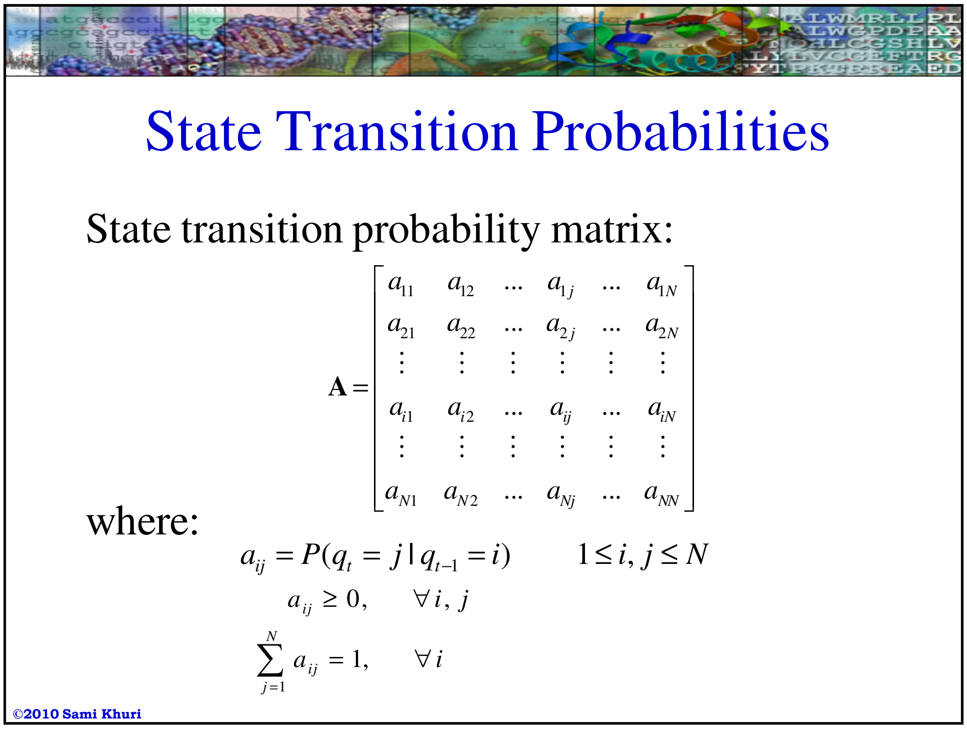
- An initial probability distribution:

$$\pi_i = P(q_1 = i) \qquad 1 \le i \le N$$

where $\sum_{i=1}^{N} \pi_i = 1$

- Stationary condition:

$$P(q_t = j \mid q_{t-1} = i) = P(q_{t+l} = j \mid q_{t+l-1} = i)$$



State Transition Probabilities

State transition probability matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{iN} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{Nj} & \dots & a_{NN} \end{bmatrix}$$

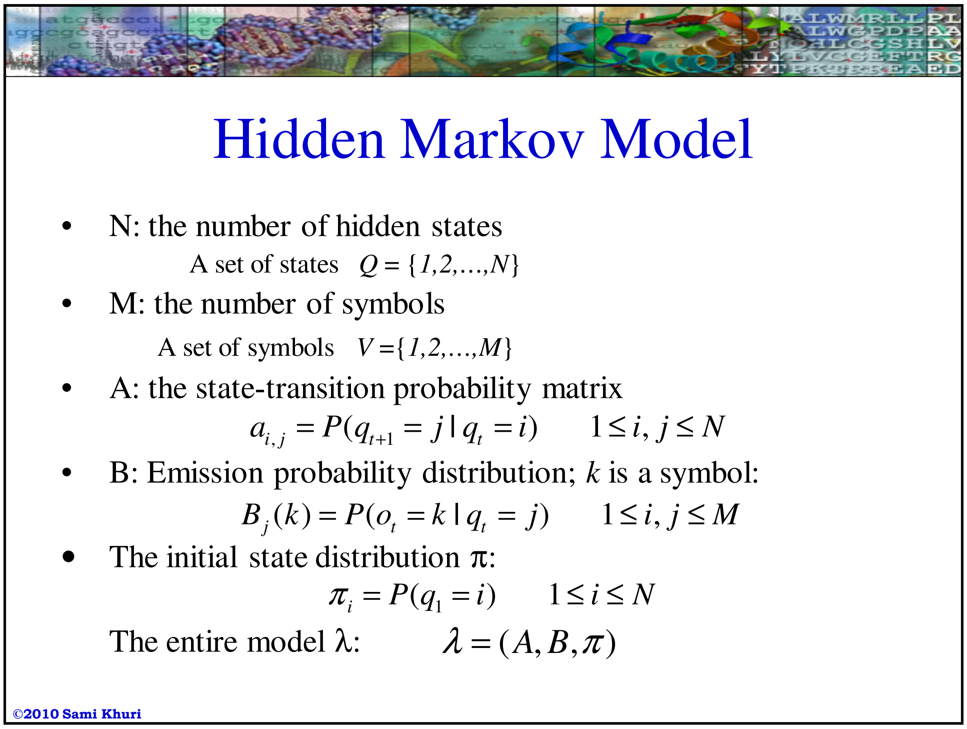
where:

$$a_{ij} = P(q_t = j \mid q_{t-1} = i) \qquad 1 \le i, j \le N$$

$$a_{ij} \ge 0, \qquad \forall\, i, j$$

$$\sum_{j=1}^{N} a_{ij} = 1, \qquad \forall\, i$$

©2010 Sami Khuri



Hidden Markov Model

- N: the number of hidden states

  A set of states $Q = \{1,2,\ldots,N\}$
- M: the number of symbols

  A set of symbols $V = \{1,2,\ldots,M\}$
- A: the state-transition probability matrix

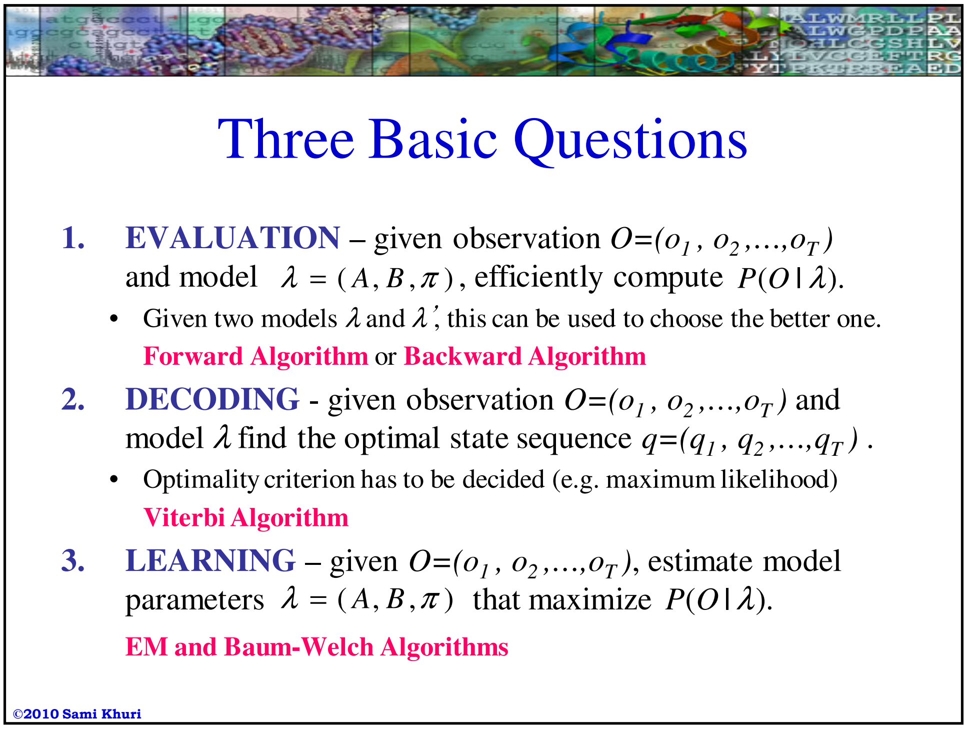
  $$a_{i,j} = P(q_{t+1} = j \mid q_t = i) \qquad 1 \le i, j \le N$$
- B: Emission probability distribution; *k* is a symbol:

  $$B_j(k) = P(o_t = k \mid q_t = j) \qquad 1 \le i, j \le M$$
- The initial state distribution π:

  $$\pi_i = P(q_1 = i) \qquad 1 \le i \le N$$

  The entire model λ: $\lambda = (A, B, \pi)$

©2010 Sami Khuri



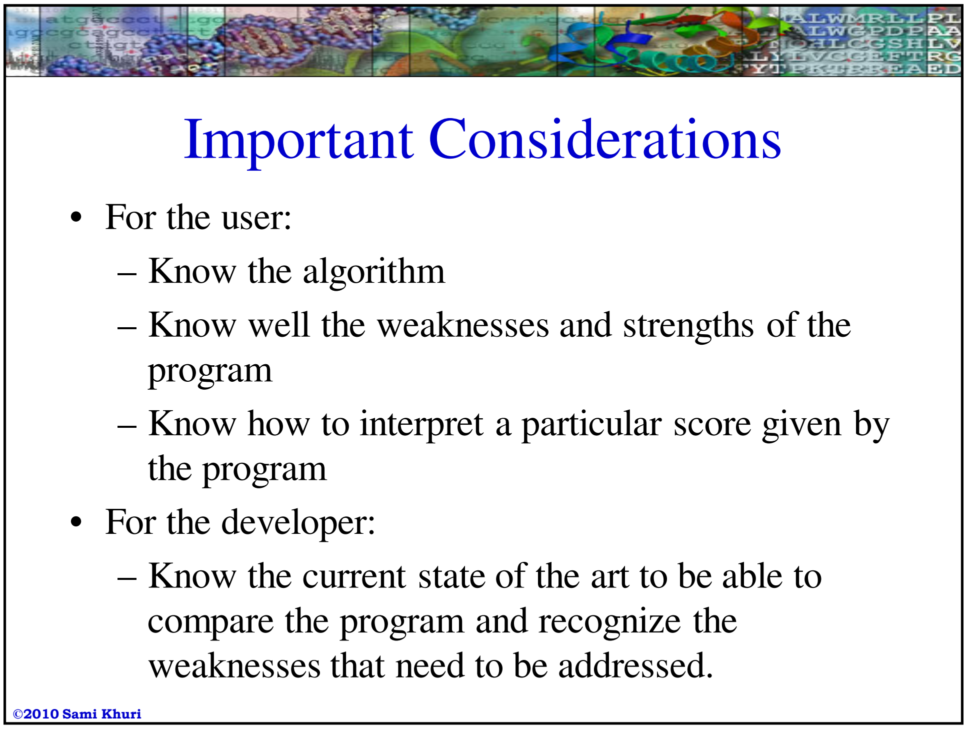
Three Basic Questions

1. **EVALUATION** – given observation $O=(o_1, o_2,\ldots,o_T)$ and model $\lambda = (A, B, \pi)$, efficiently compute $P(O\mid\lambda)$.
   - Given two models λ and λ', this can be used to choose the better one.

   Forward Algorithm or **Backward Algorithm**
2. **DECODING** - given observation $O=(o_1, o_2,\ldots,o_T)$ and model λ find the optimal state sequence $q=(q_1, q_2,\ldots,q_T)$.
   - Optimality criterion has to be decided (e.g. maximum likelihood)

   Viterbi Algorithm
3. **LEARNING** – given $O=(o_1, o_2,\ldots,o_T)$, estimate model parameters $\lambda = (A, B, \pi)$ that maximize $P(O\mid\lambda)$.

   EM and Baum-Welch Algorithms



Important Considerations

- For the user:
  - Know the algorithm
  - Know well the weaknesses and strengths of the program
  - Know how to interpret a particular score given by the program
- For the developer:
  - Know the current state of the art to be able to compare the program and recognize the weaknesses that need to be addressed.