Introduction to Bioinformatics and Molecular Biology

Jianlin Cheng, PhD

School of Electrical Engineering and Computer Science
University of Central Florida

2006

Goals

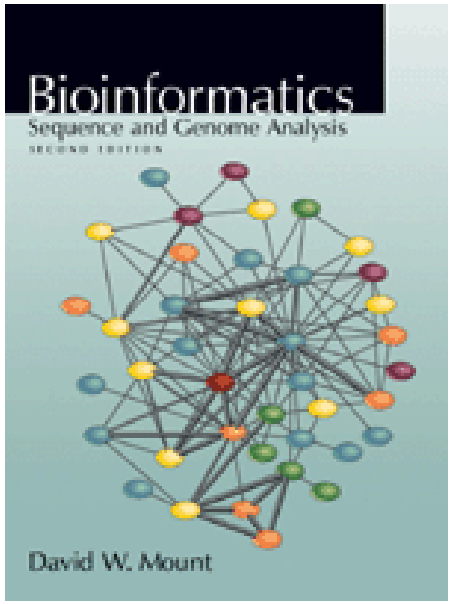
- Introduce fundamental problems, concepts, methods, and applications in Bioinformatics
- Emphasize both the methods and the practical use of bioinformatics tools and databases.
- Audience: computer scientists, engineers, biologists, statisticians, …
- Prerequisite: some background in programming OR molecular biology.

Ten Topics

1. Introduction to Molecular Biology and Bioinformatics
2. Pairwise Sequence Alignment Using Dynamic Programming
3. Practical Sequence/Profile Alignment Using Fast Heuristic Methods (BLAST and PSI-BLAST)
4. Multiple Sequence Alignment
5. Gene and Motif Identification
6. Phylogenetic Analysis
7. Protein Structure Analysis and Prediction
8. RNA Secondary Structure Prediction
9. Clustering and Classification of Gene Expression Data
10. Search and Mining of Biological Databases, Databanks, and Literature

Resources

- Slides (**main materials**, http://www.eecs.ucf.edu/~jcheng/cheng_courses.html)
- A general introduction book (David Mount. Bioinformatics: Sequence and Genome Analysis, 2004. additional reading materials)



- Other online Bioinformatics courses (additional reading materials)
- Reference books (complementary)

Grading

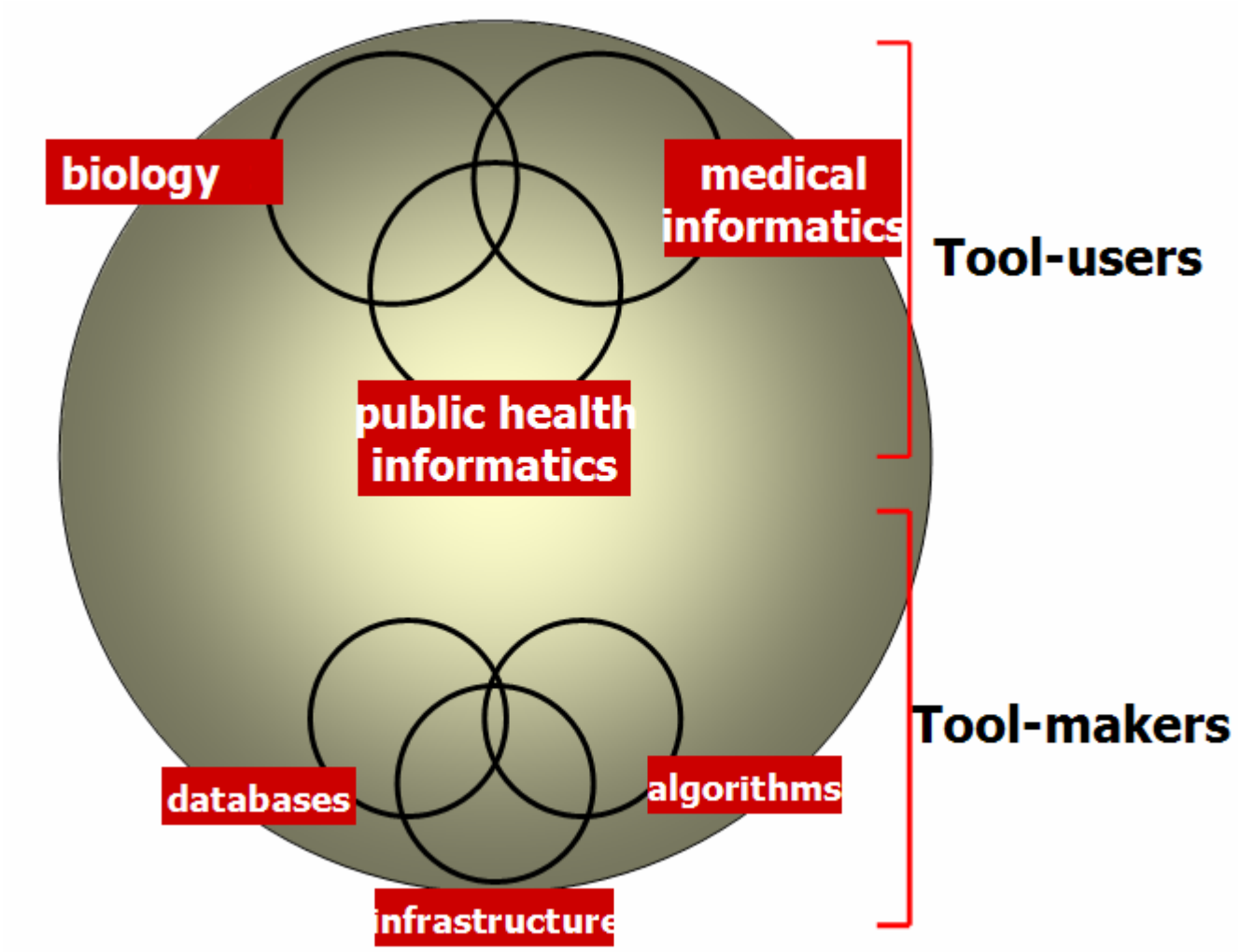
- Homework (20%)
- Project (20%)
- Midterm (30%)
- Final (30%)

What's Bioinformatics?

An interdisciplinary science of developing and applying computational techniques to address problems in molecular biology (or more broadly biology: computational biology).

Two main fields:

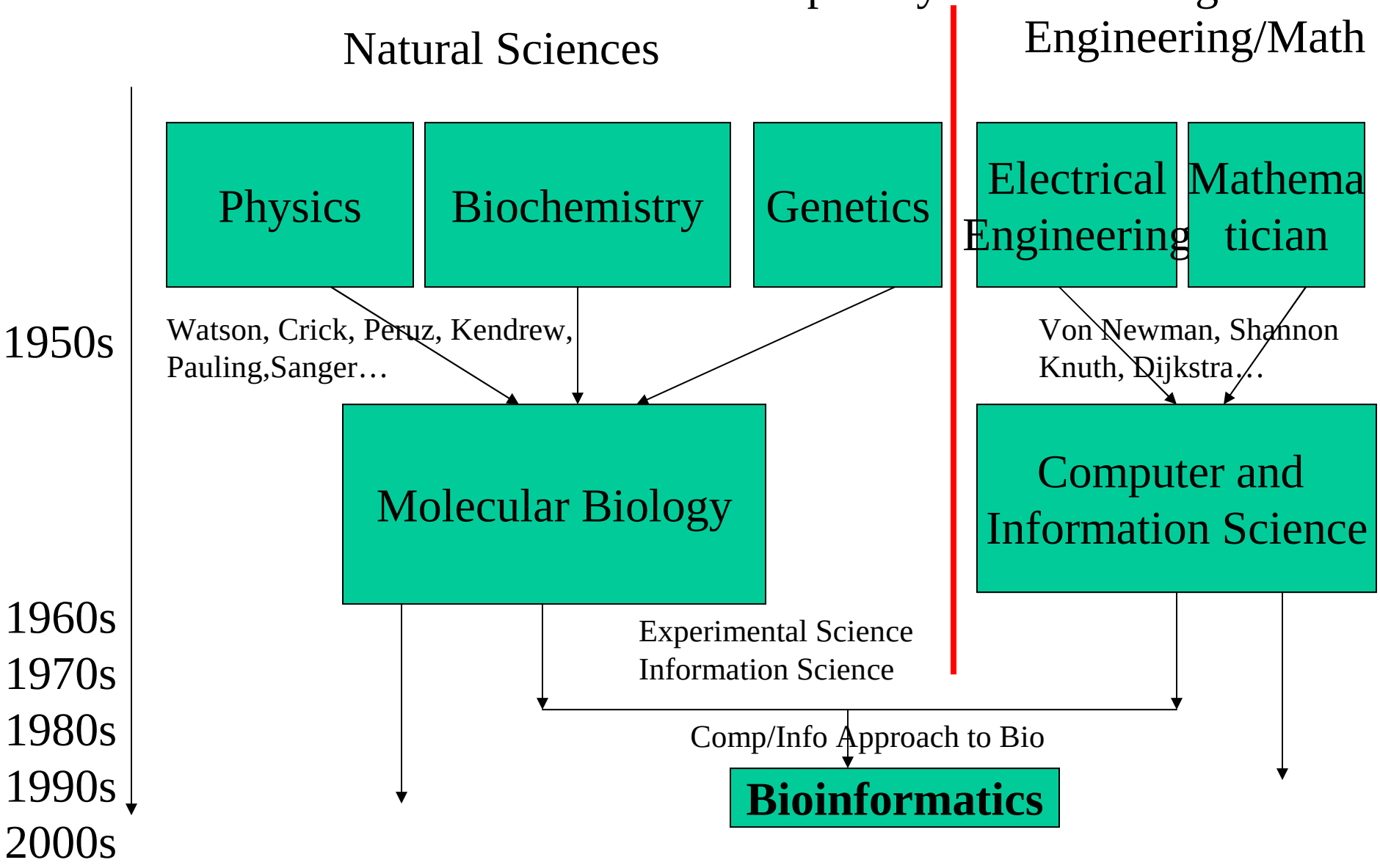
- Develop bioinformatics algorithms and tools
- Apply bioinformatics tools to address biological problems



Adapted from J. Pevsner, 2005

History of Bioinformatics

How does a new interdisciplinary science emerge?



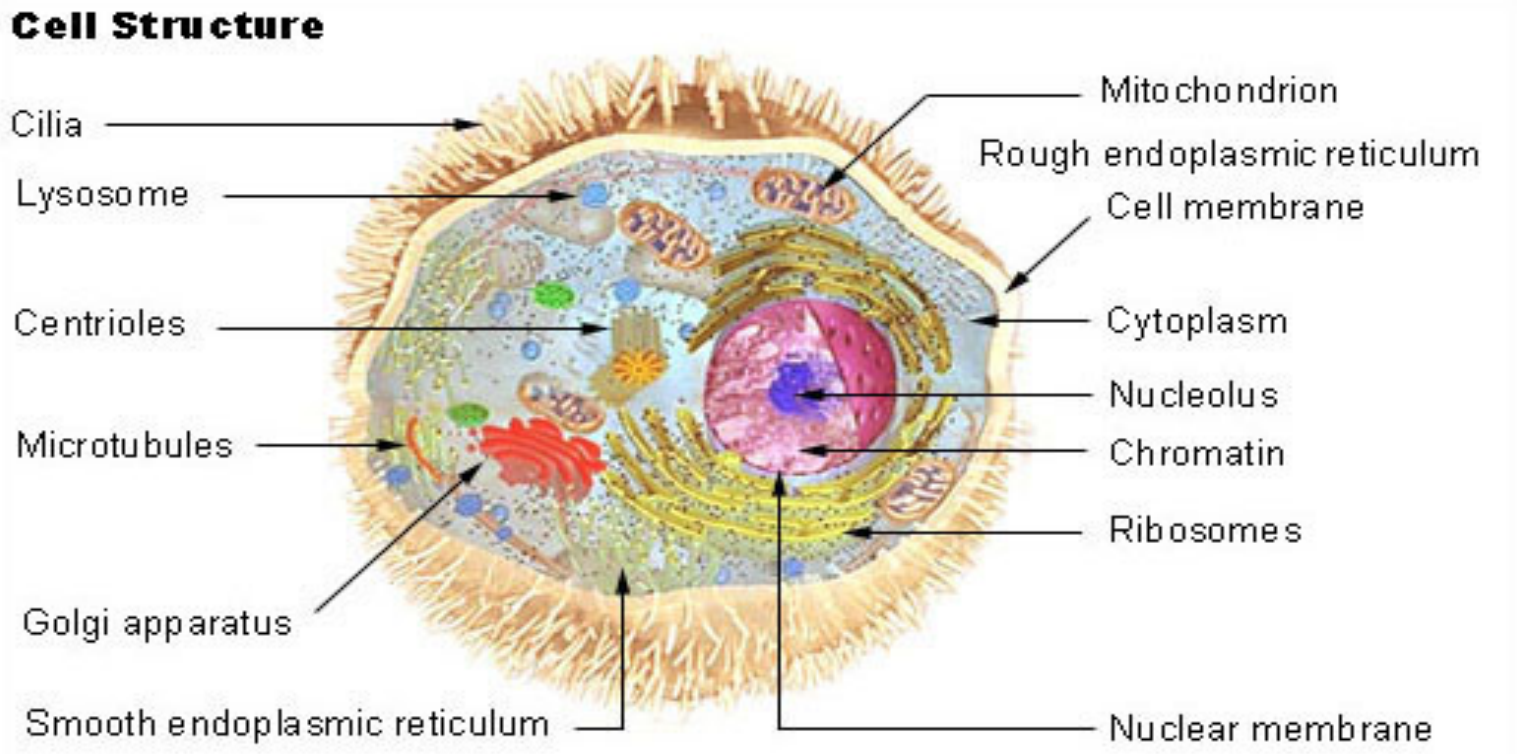
Major Events in Bioinformatics History

http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/milestones.html

1962	**Pauling's theory of molecular evolution**
1965	**Margaret Dayhoff's Atlas of Protein Sequences**
1970	**Needleman-Wunsch algorithm**
1977	**DNA sequencing and software to analyze it (Sateen)**
1981	**Smith-Waterman algorithm developed**
1981	**The concept of a sequence motif (Doolittle)**
1982	**GenBank Release 3 made public**
1982	**Phage lambda genome sequenced**
1983	**Sequence database searching algorithm (Wilbur-Lipman)**
1985	**FASTP/FASTN: fast sequence similarity searching**
1988	**National Center for Biotechnology Information (NCBI) created at NIH/NLM**
1988	**EMBnet network for database distribution**
1990	**BLAST: fast sequence similarity searching**
1991	**EST: expressed sequence tag sequencing**
1993	**Sanger Centre, Hinxton, UK**
1994	**EMBL European Bioinformatics Institute, Hinxton, UK**
1995	**First bacterial genomes completely sequenced**
1996	**Yeast genome completely sequenced**
1997	**PSI-BLAST**
1998	**Worm (multicellular) genome completely sequenced**
1999	**Fly genome completely sequenced**
2004	**Human genome sequenced**

Introduction to Molecular Biology

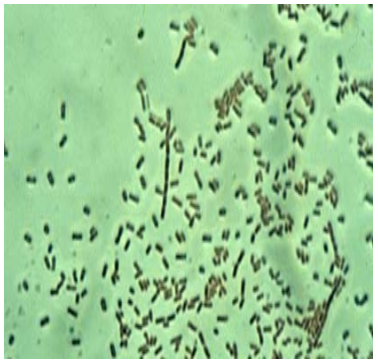
- Cell is the unit of structure and function of all living things.



Two types of cells: eukaryote (higher organisms) and prokaryote (lower organisms)

Central Dogma of Molecular Biology

Phenotype



Replication

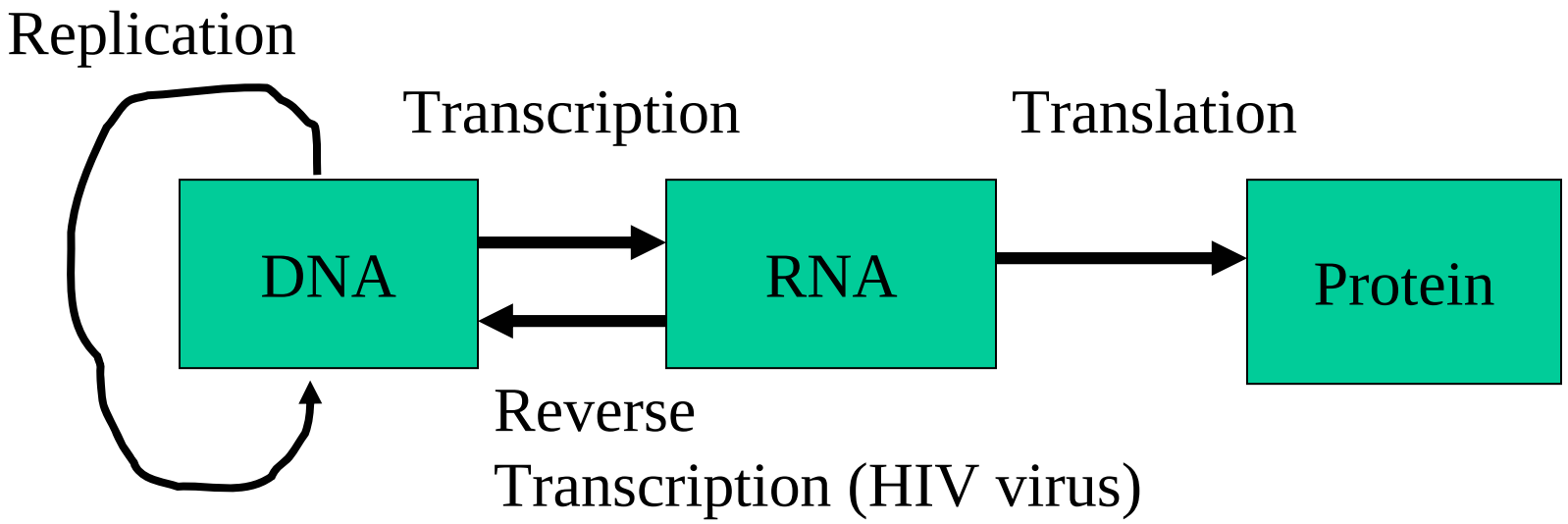
Transcription

Translation

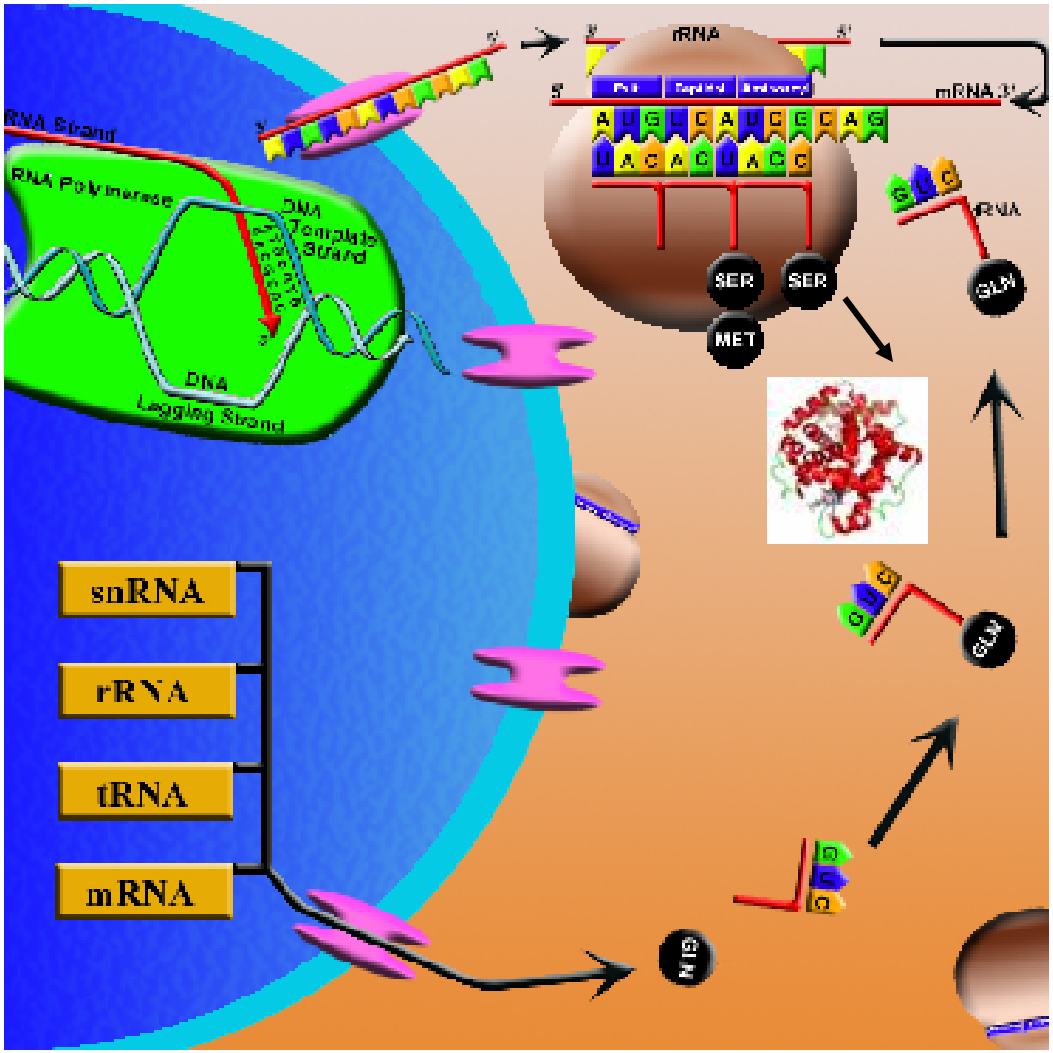
DNA → RNA → Protein

Genotype

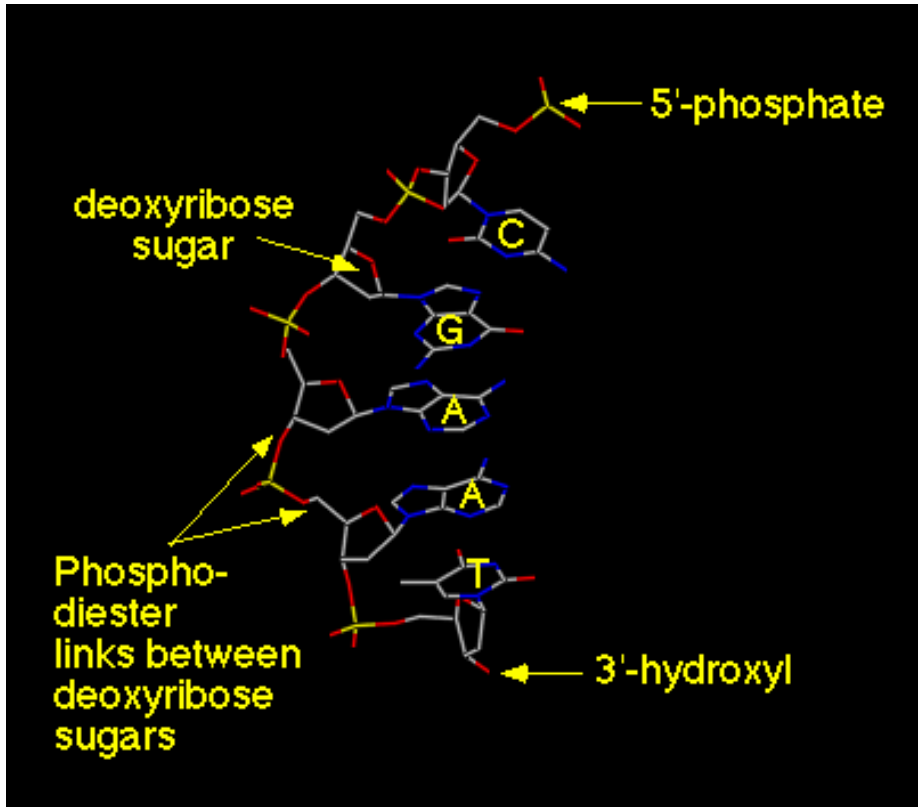
Central Dogma of Molecular Biology



Information flow


RNA Strand
RNA Polymerase
DNA Template Strand
DNA Leading Strand
rRNA
mRNA
SER
SER
MET
GLN
snRNA
rRNA
tRNA
mRNA

DNA (Deoxyribose Nucleotide Acids)

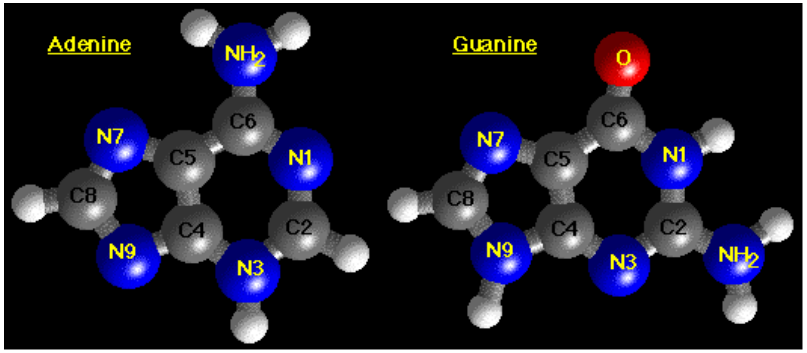


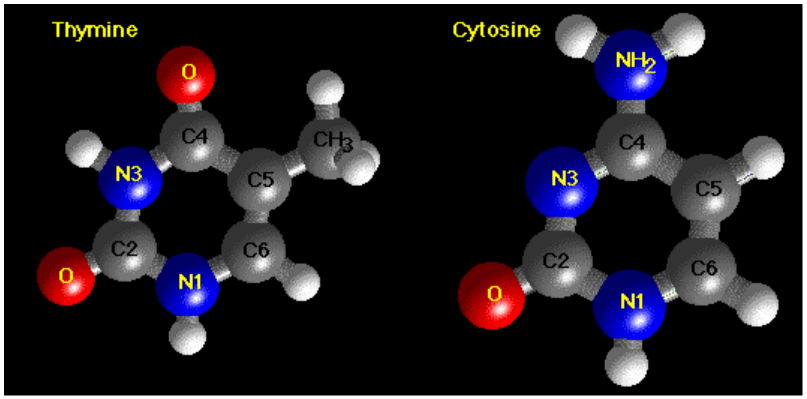
CGAATGGGAAA……

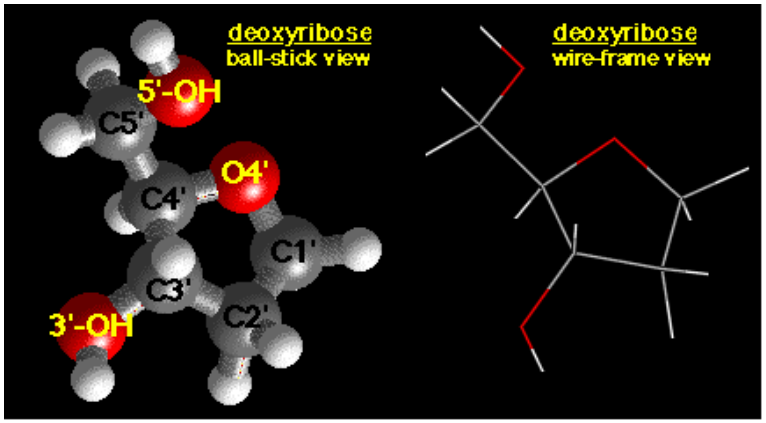
DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a "polynucleotide." Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group.

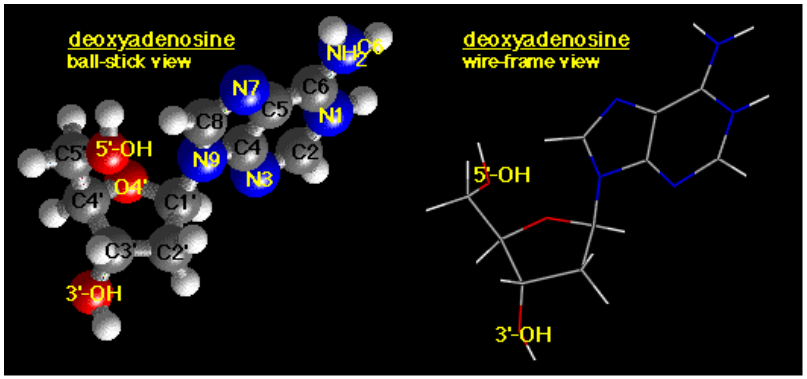
A is for adenine
G is for guanine
C is for cytosine
T is for thymine

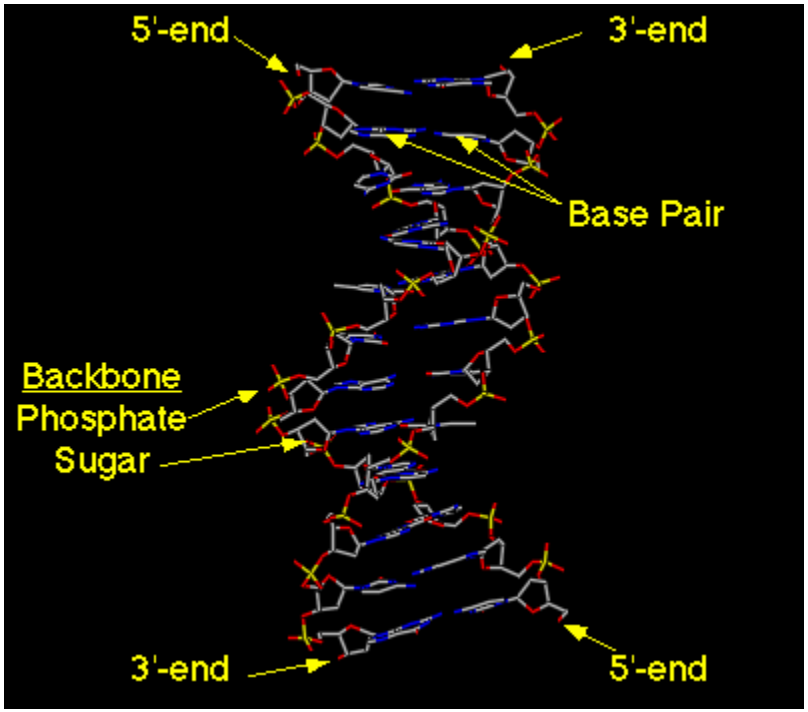
Introduction to DNA structure, Richard B. Hallick, 1995

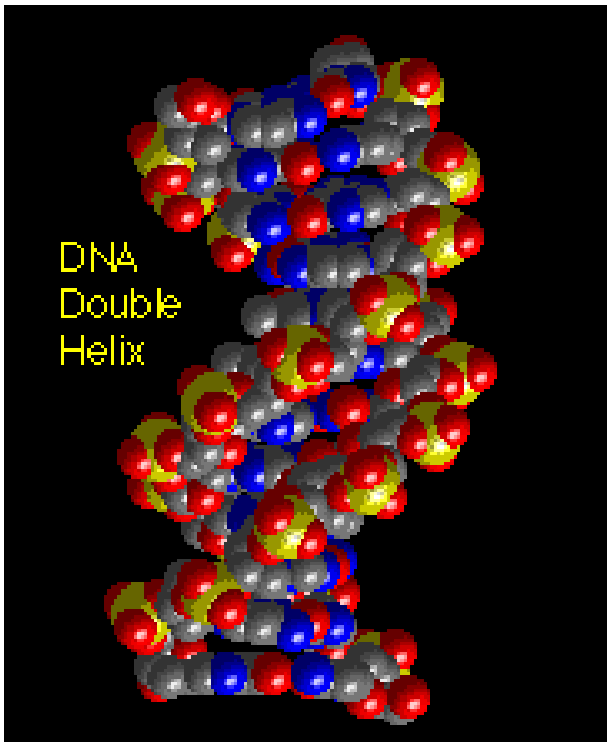

Adenine
NH2
C6
N7
C5
N1
C8
C4
C2
N9
N3
Guanine
O
C6
N7
C5
N1
C8
C4
C2
N9
N3
NH2


Thymine
O
C4
CH3
N3
C5
C2
C6
O
N1
Cytosine
NH2
C4
N3
C5
C2
C6
O
N1


deoxyribose
ball-stick view
5'-OH
C5'
O4'
C4'
C1'
C3'
C2'
3'-OH
deoxyribose
wire-frame view

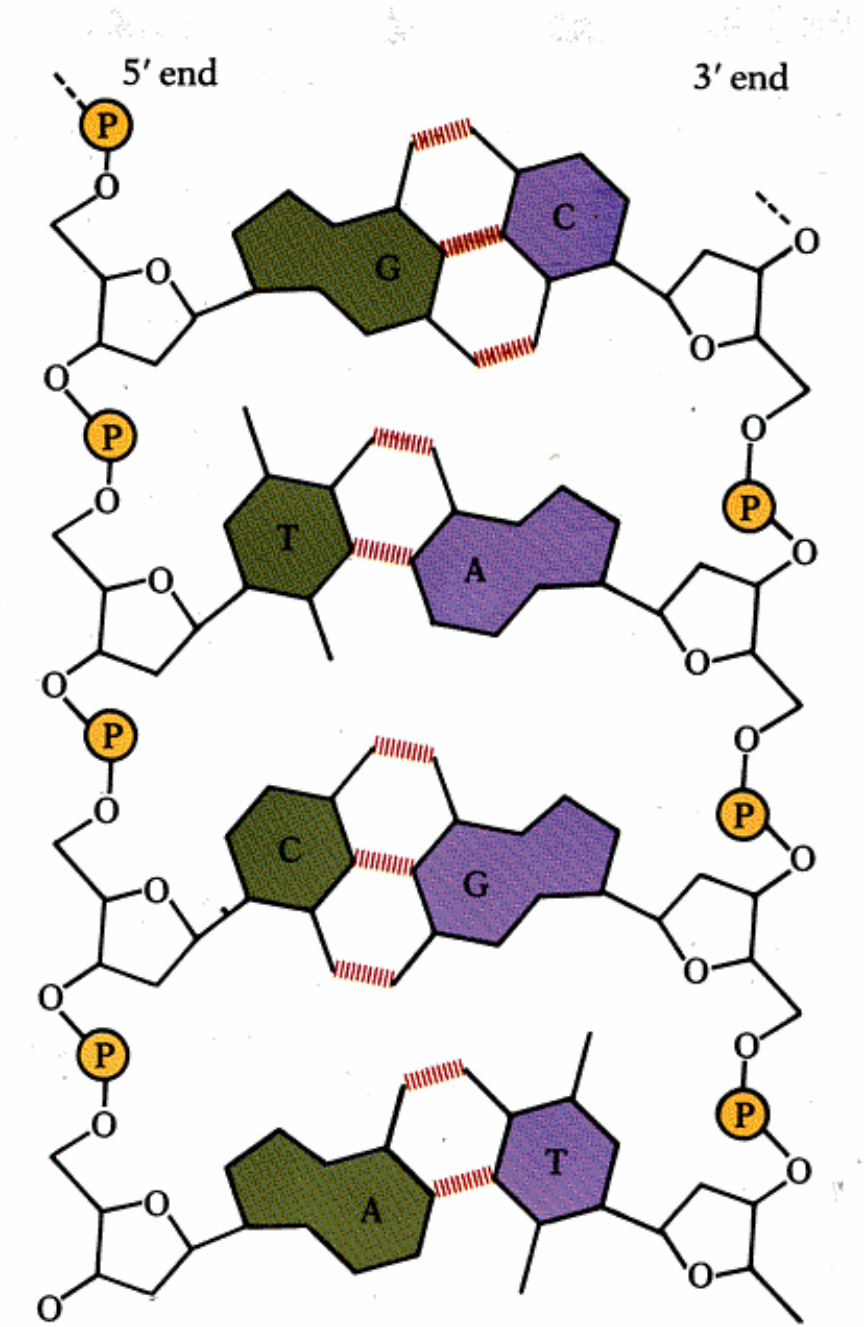

deoxyadenosine
ball-stick view
NH2
C6
N7
C5
N1
C8
C4
C2
5'-OH
C5'
N9
N3
O4'
C4'
C1'
C3'
C2'
3'-OH
deoxyadenosine
wire-frame view
5'-OH
3'-OH





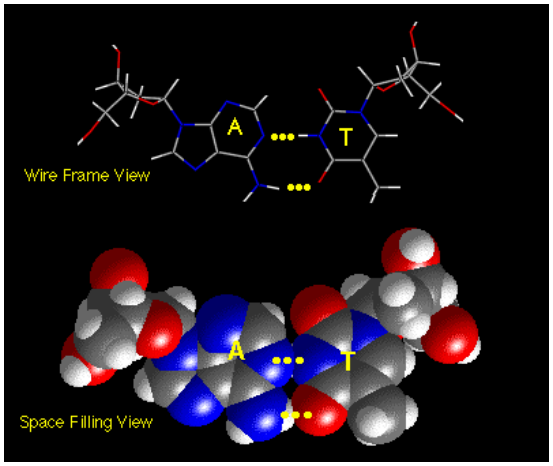
Base Pairs:
A-T (2 H-bonds)
C-G (3 H-bonds)

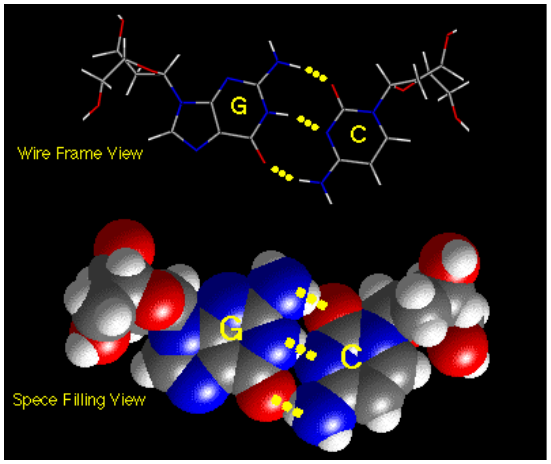
Hydrogen bonds: non-covalent bonds mediated by hydrogen atoms

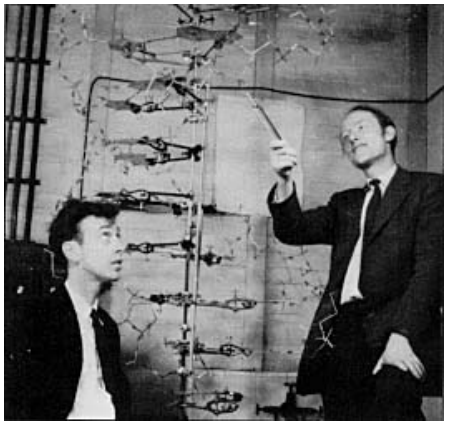


Uncoiled DNA Molecule

Source: Dr. Gary Stormo, 2002







James Watson & Francis Crick

Maurice Wilkins

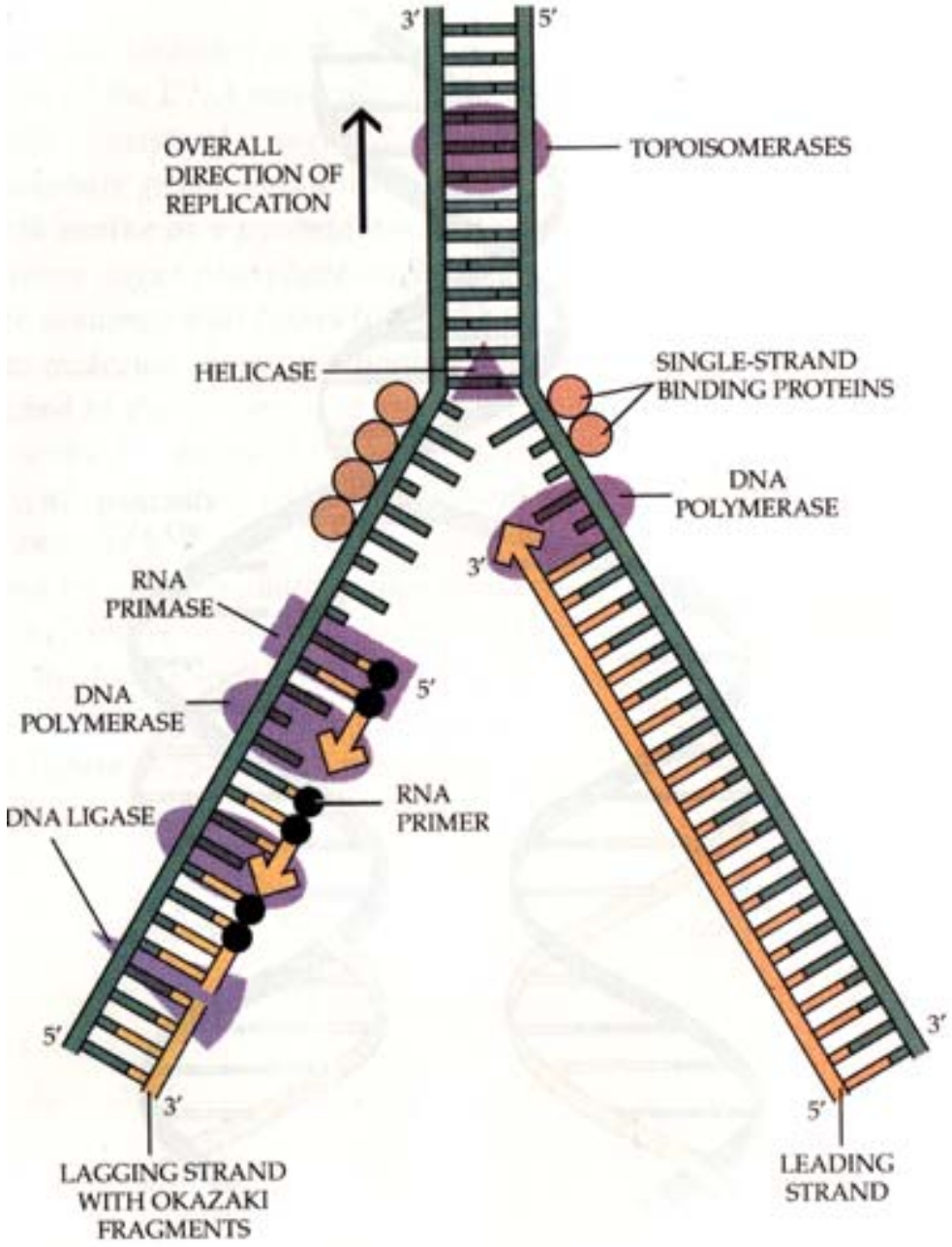
Rosalind Franklin

Linus Pauling

Erwin Chargaff

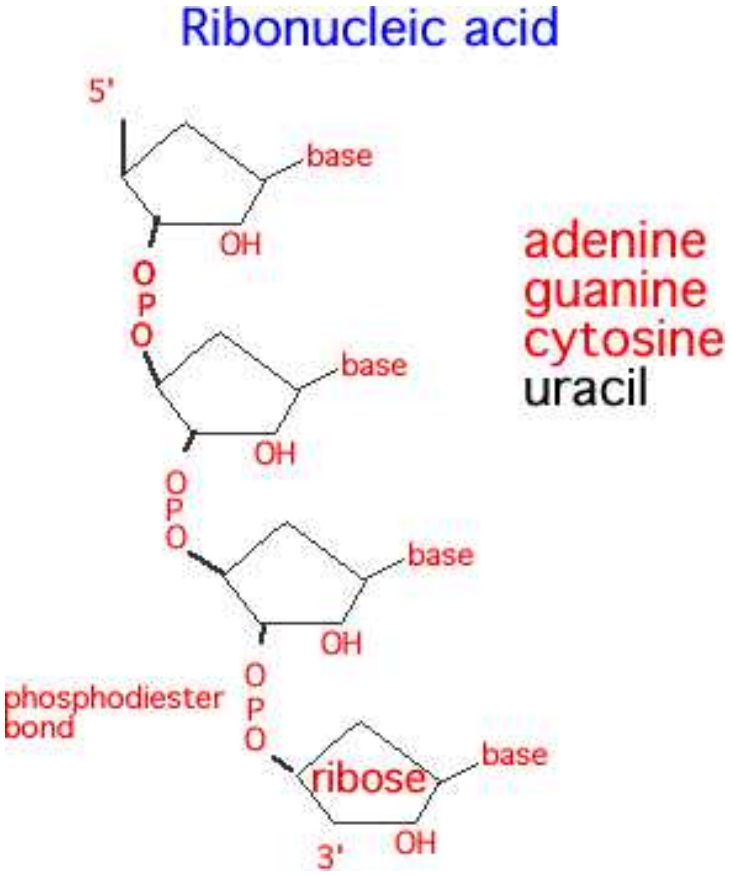
Fundamental Problems: How genetic information pass from one cell to another and from one generation to next generation

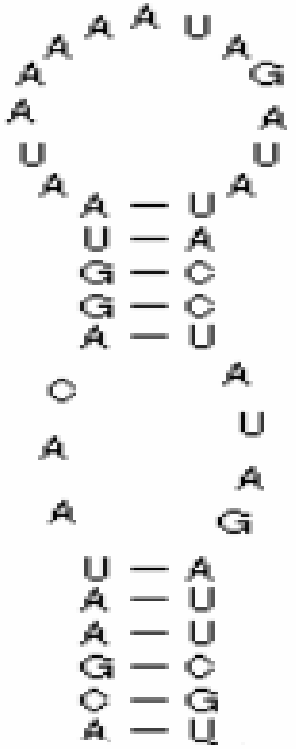
DNA Replication



DNA Polymerase

RNA (Ribose Nucleotide Acids)





ACGAAUAACAGGUAAUAAAAAUAGAUAUACCUAUAGAUUCGU

Different Kinds of RNA

- mRNA: messager RNA

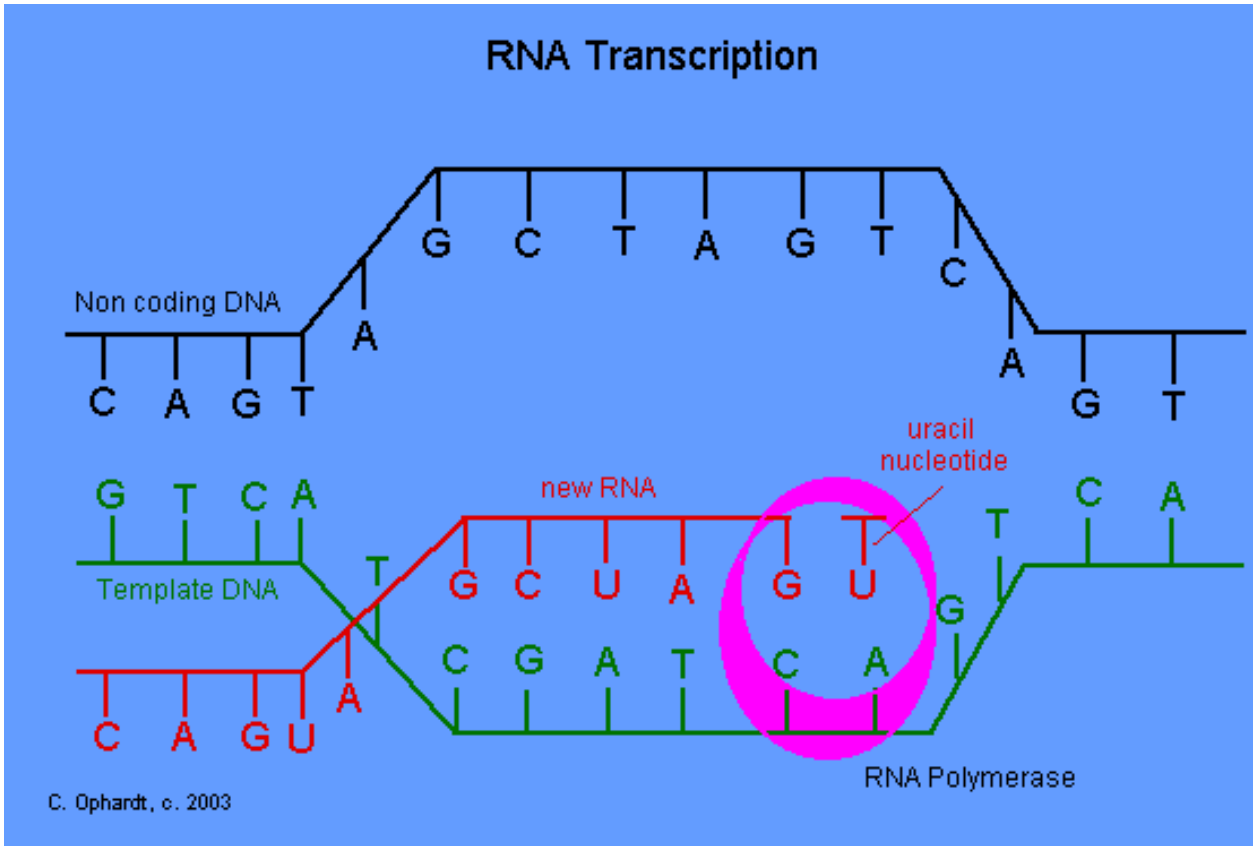
  carry genetic information out of nucleus for protein synthesis (transcription process: RNA polymerase)
- rRNA: ribosomal RNA

  constitute 50% of ribosome, which is a molecular assembly for protein synthesis
- tRNA: transfer RNA

  decode information (map 3 nucleotides to amino acid); transfer amino acid
- snRNA: small RNA molecules found in nucleus

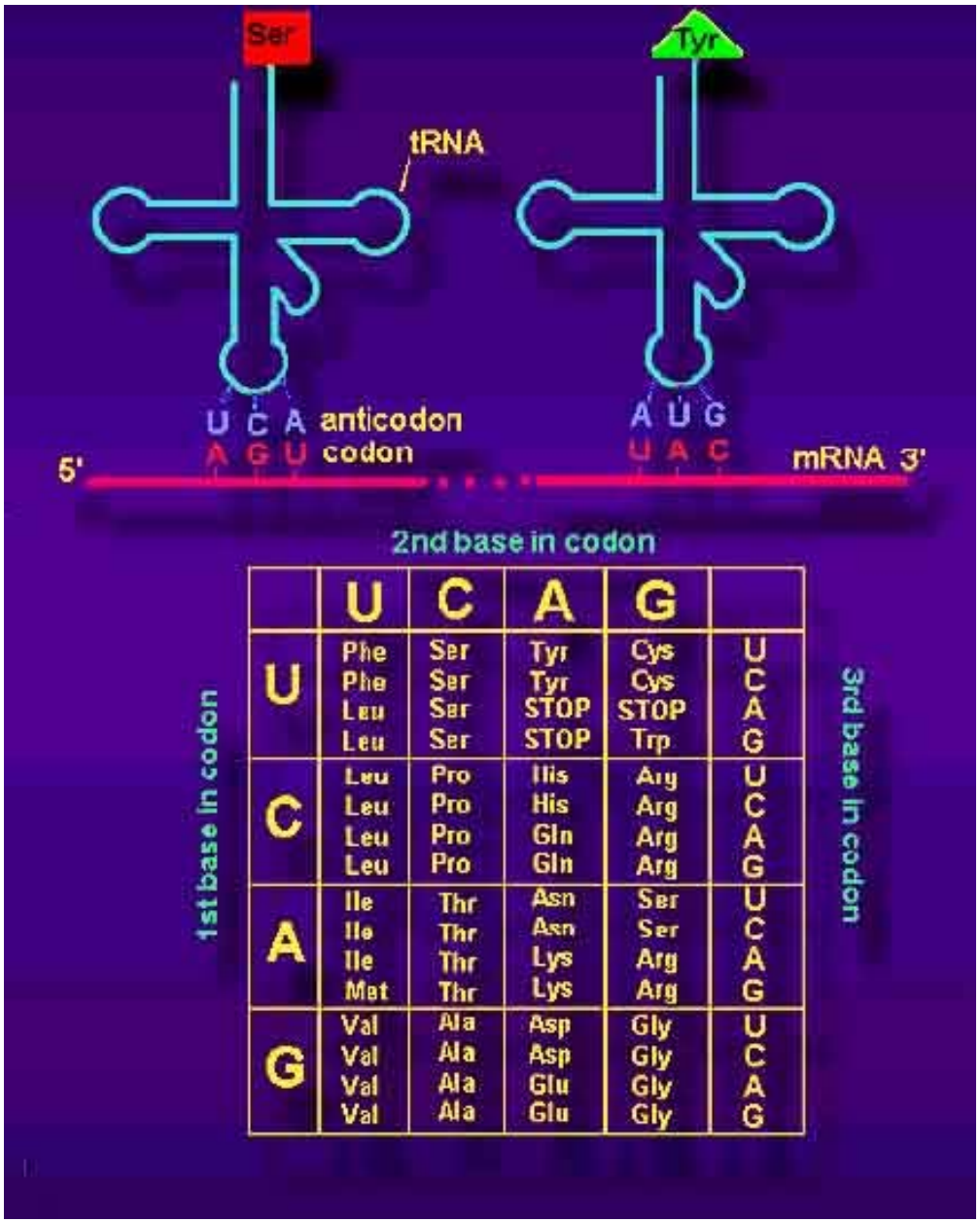
  involve RNA splicing

Transcription of Gene into RNA



Genetic Code and Translation

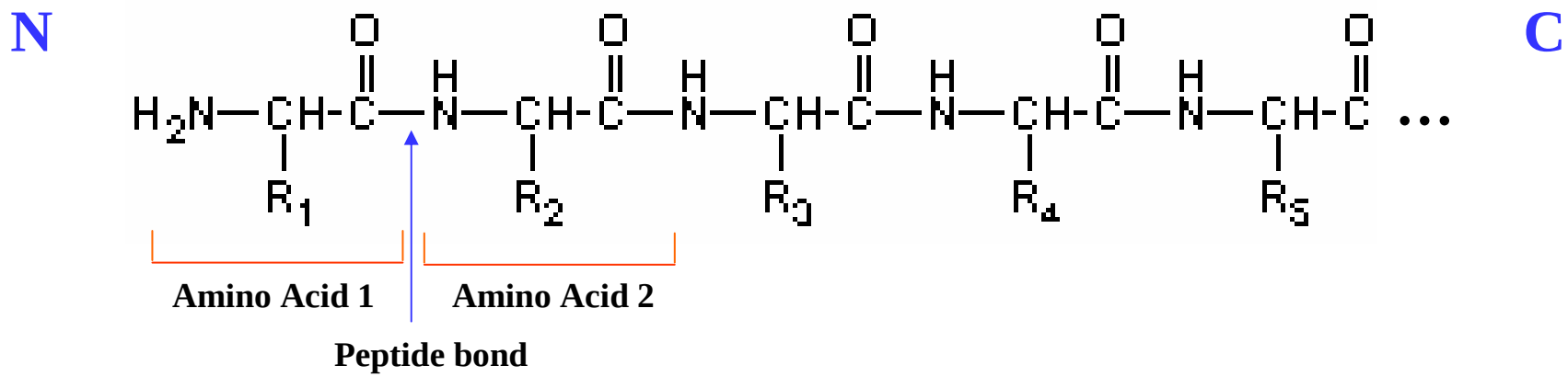
Three Nucleotides is called a codon.



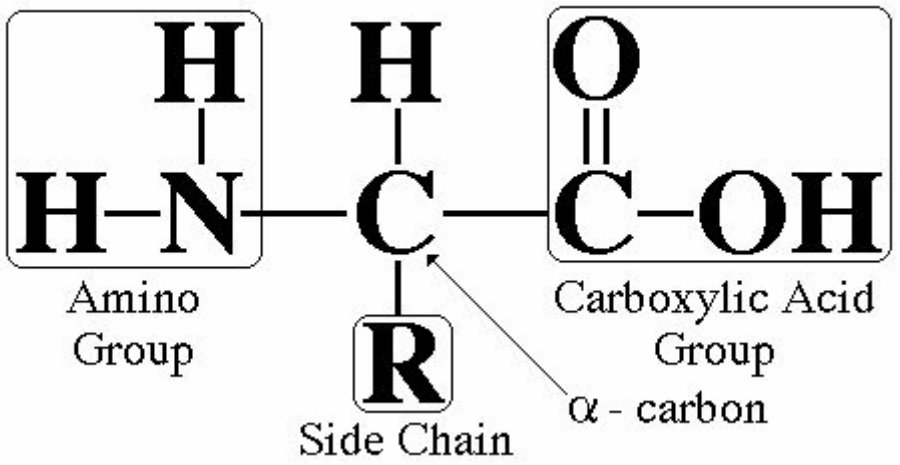
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	Ilis	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Protein Sequence

A directional sequence of amino acids/residues

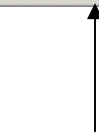


Amino Acid Structure



Amino Acids

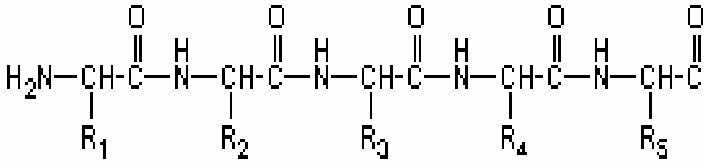
Amino acid	Abbrev.	Side chain	Hydro-phobic	Polar	Charged	Small	Tiny	Aromatic or Aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Alanine	Ala, A	$-CH_3$	X	-	-	X	X	-	67	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	$-CH_2SH$	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	$-CH_2COOH$	-	X	negative	X	-	-	91	GAU, GAC	5.3
Glutamate	Glu, E	$-CH_2CH_2COOH$	-	X	negative	-	-	-	109	GAA, GAG	6.3
Phenylalanine	Phe, F	$-CH_2C_6H_5$	X	-	-	-	-	Aromatic	135	UUU, UUC	3.9
Glycine	Gly, G	-H	X	-	-	X	X	-	48	GGU, GGC, GGA, GGG	7.2
Histidine	His, H	$-CH_2-C_3H_3N_2$	-	X	positive	-	-	Aromatic	118	CAU, CAC	2.3
Isoleucine	Ile, I	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	Aliphatic	124	AUU, AUC, AUA	5.3
Lysine	Lys, K	$-(CH_2)_4NH_2$	-	X	positive	-	-	-	135	AAA, AAG	5.9
Leucine	Leu, L	$-CH_2CH(CH_3)_2$	X	-	-	-	-	Aliphatic	124	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	124	AUG	2.3
Asparagine	Asn, N	$-CH_2CONH_2$	-	X	-	X	-	-	96	AAU, AAC	4.3
Proline	Pro, P	$-CH_2CH_2CH_2-$	X	-	-	X	-	-	90	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	$-CH_2CH_2CONH_2$	-	X	-	-	-	-	114	CAA, CAG	4.2
Arginine	Arg, R	$-(CH_2)_3NH-C(NH)NH_2$	-	X	positive	-	-	-	148	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	$-CH_2OH$	-	X	-	X	X	-	73	UCU, UCC, UCA, UCG, AGU,AGC	6.8
Threonine	Thr, T	$-CH(OH)CH_3$	X	X	-	X	-	-	93	ACU, ACC, ACA, ACG	5.9
Valine	Val, V	$-CH(CH_3)_2$	X	-	-	X	-	Aliphatic	105	GUU, GUC, GUA, GUG	6.6
Tryptophan	Trp, W	$-CH_2C_8H_6N$	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	Tyr, Y	$-CH_2-C_6H_4OH$	X	X	-	-	-	Aromatic	141	UAU, UAC	3.2

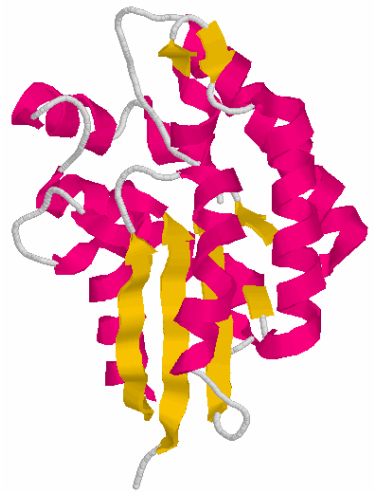


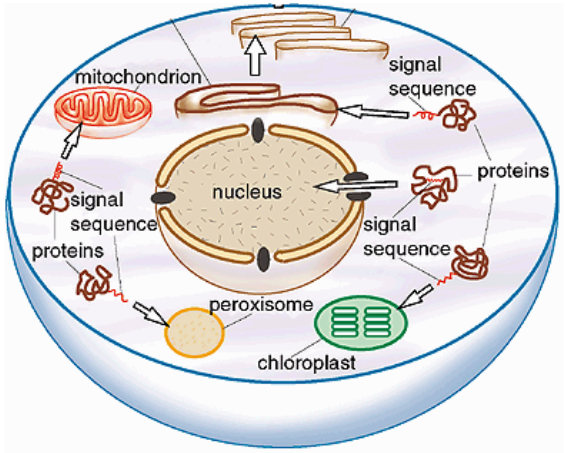
Hydrophilic

Central Dogma of Proteomics

AGCWY……







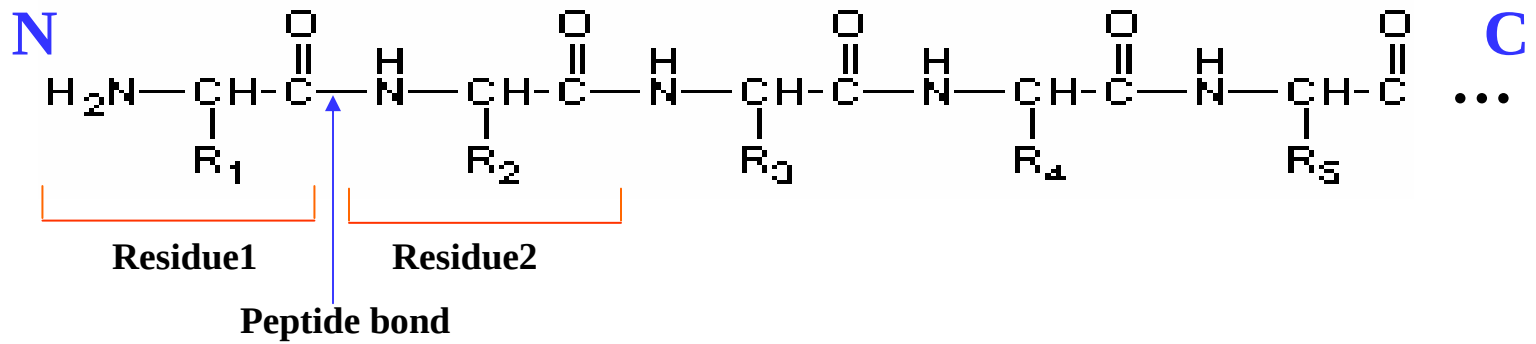
Cell

Sequence **Structure** **Function**

From Sequence to Consequence

Four Levels of Protein Structure

Primary Structure (a directional sequence of amino acids/residues)



Secondary Structure (helix, strand, coil)

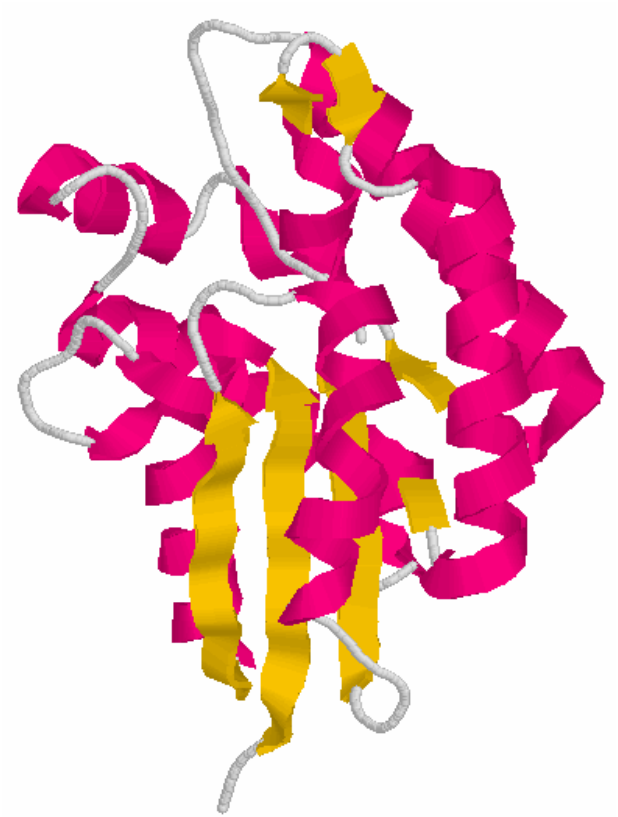
Alpha Helix

Beta Strand / Sheet

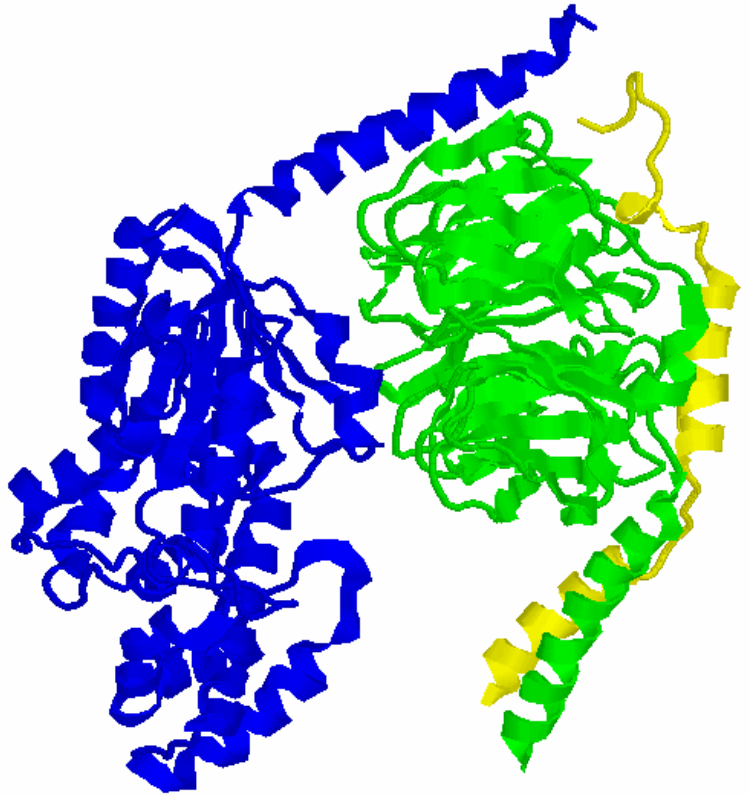
Coil

Four Levels of Protein Structure

Tertiary Structure



Quaternary Structure (complex)



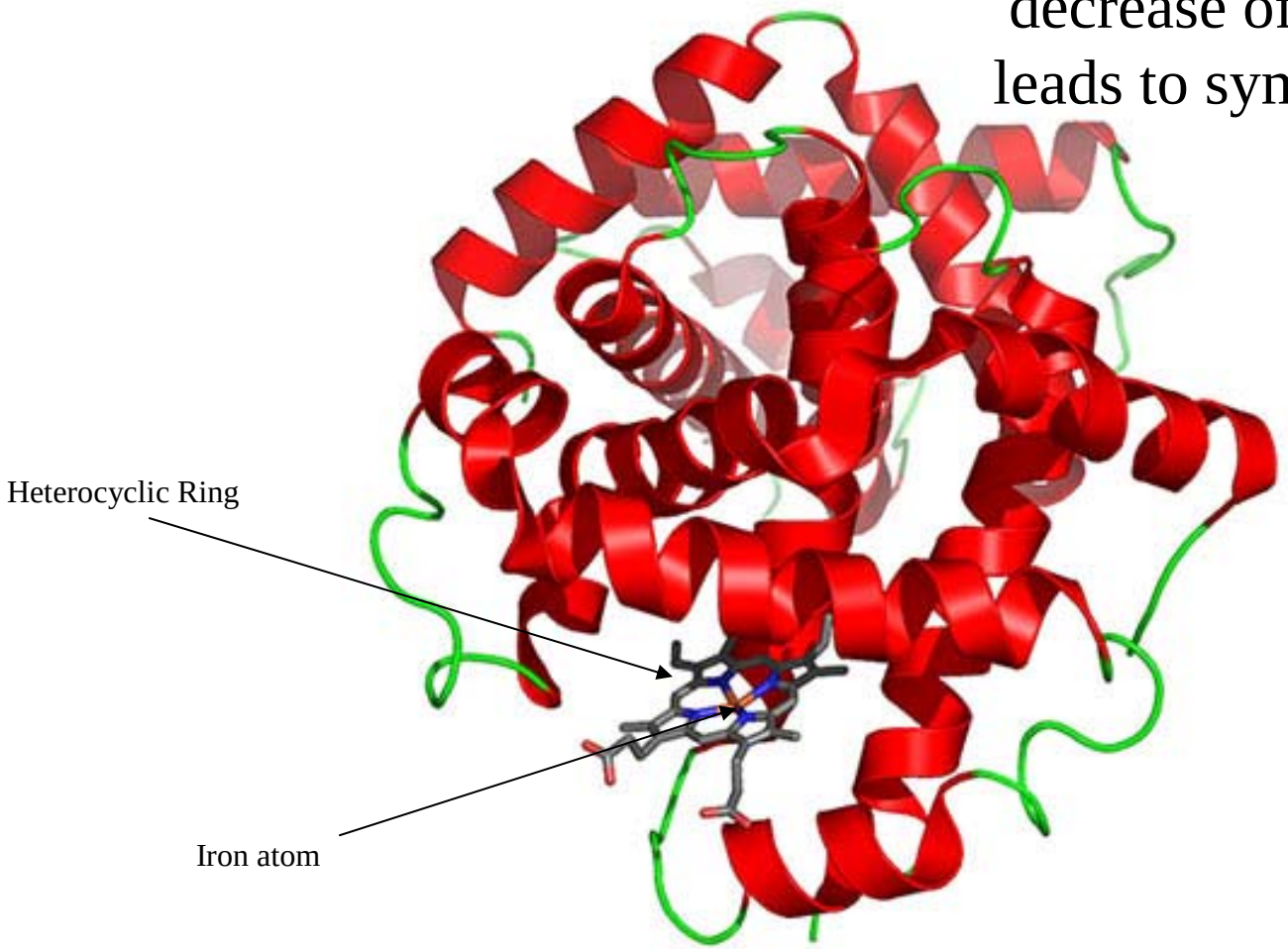
G Protein Complex

Protein Function

- Enzymatic catalysis
- Transport and storage
- Coordinated motion
- Immune protection
- Generation and transmission of nerve impulses
- Control of growth and differentiation
- Structural materials

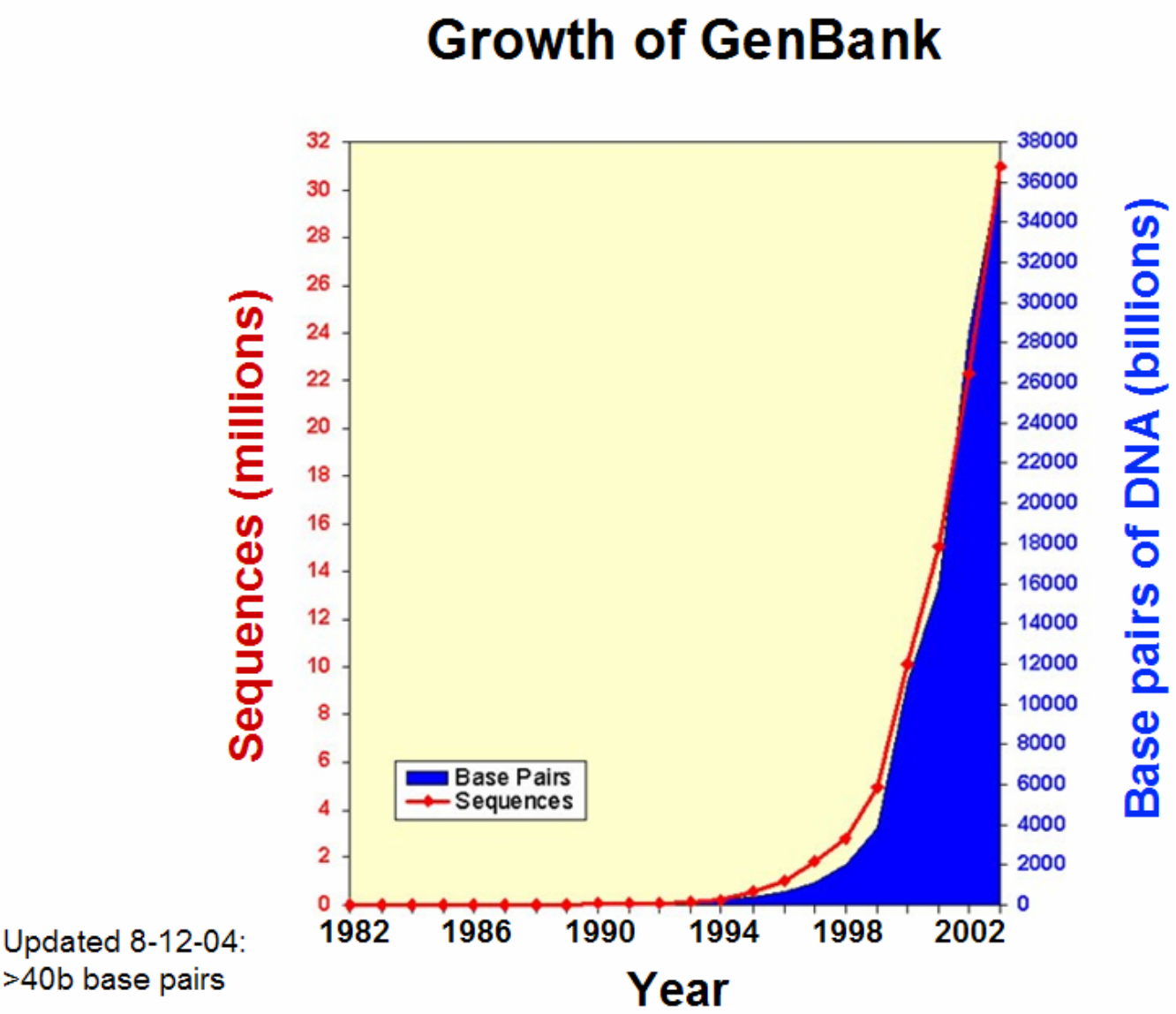
Example: Hemoglobin Transports Oxygen

Decreased levels of hemoglobin, with or without an absolute decrease of red blood cells, leads to symptoms of anemia.



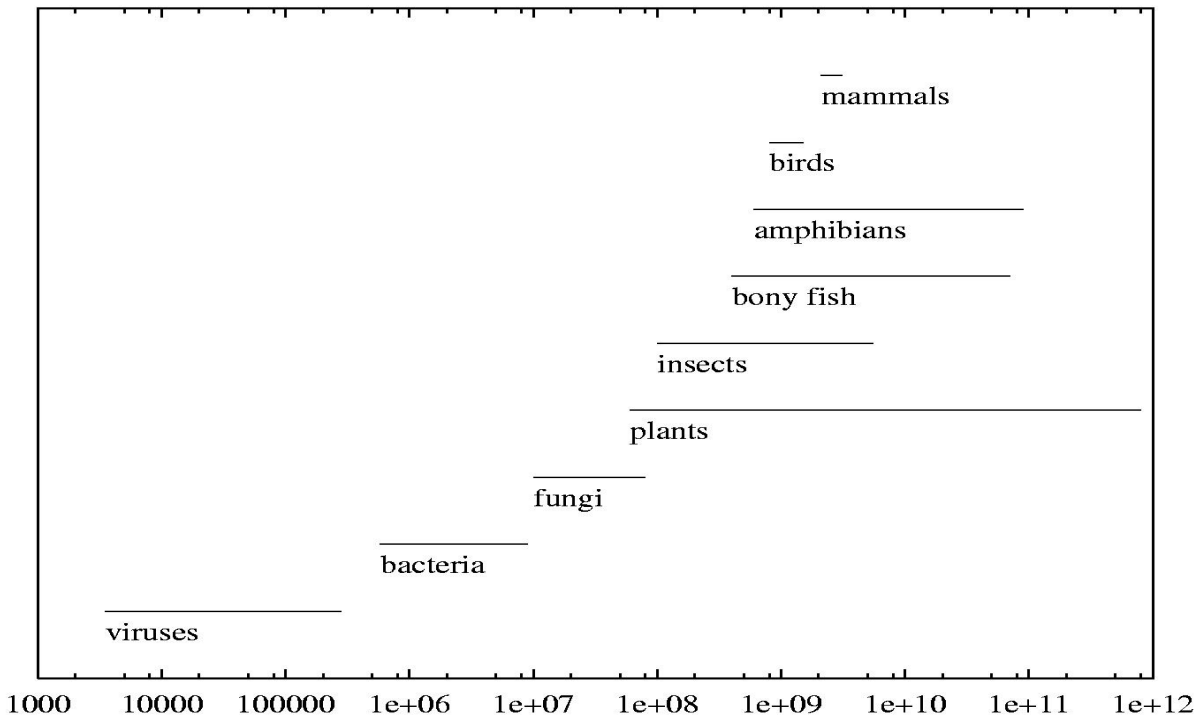
Huge Amount of Biological Data

- DNA/RNA Sequence Data (GeneBank) (human genome project)
- Protein Sequence Data (SwissProt and PIR)
- Protein Structure Data (Protein Data Bank)
- Gene Expression Data (Gene expression omnibus)
- Protein Function (Gene Ontology)
- And many more….

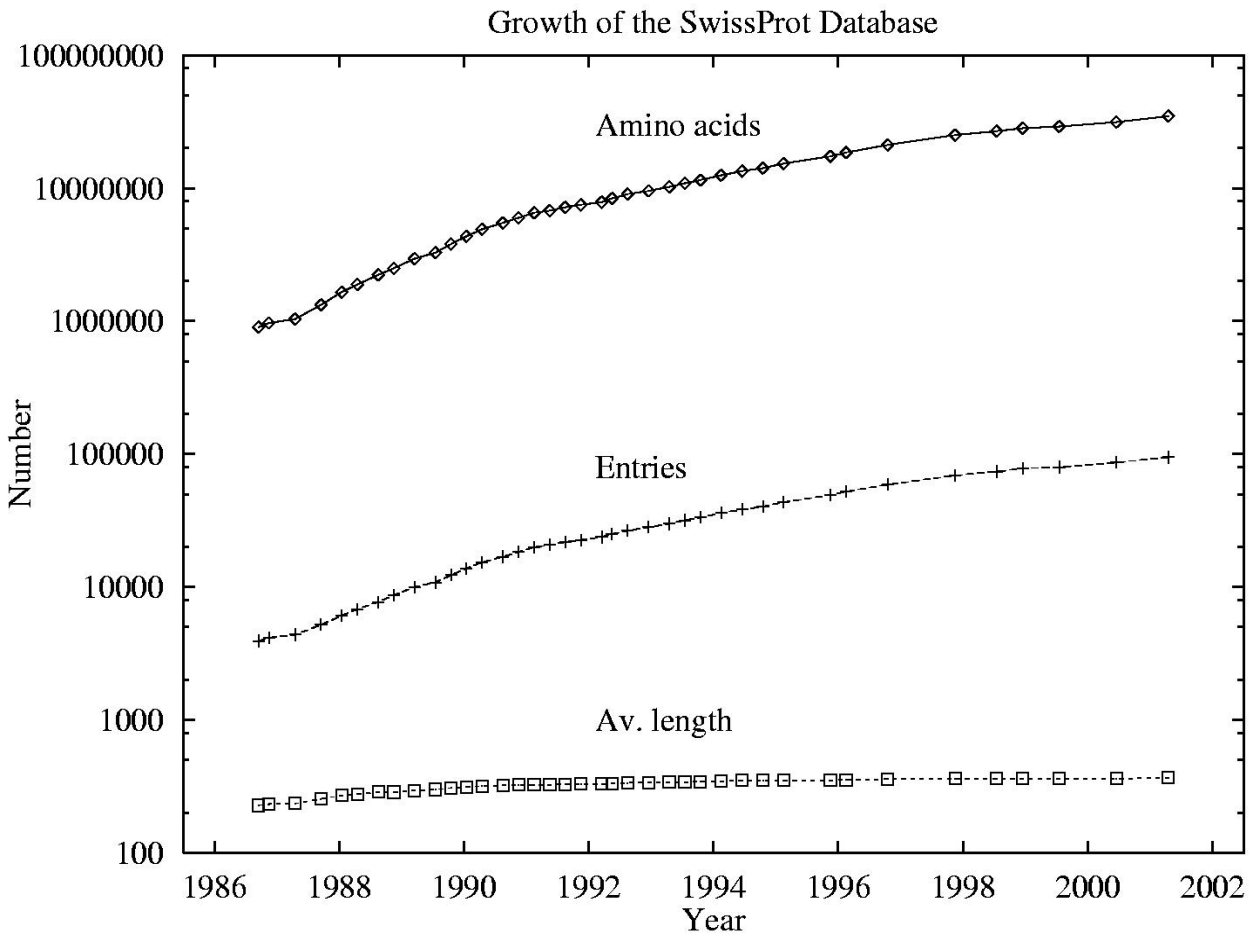


Source: J. Pevsner, 2005

Genome Size



Source: P. Baldi, 2003



Source: P. Baldi, 2003

What can we do with these huge amount of data?

Find buried treasure - Doug Brutlag, 1999.

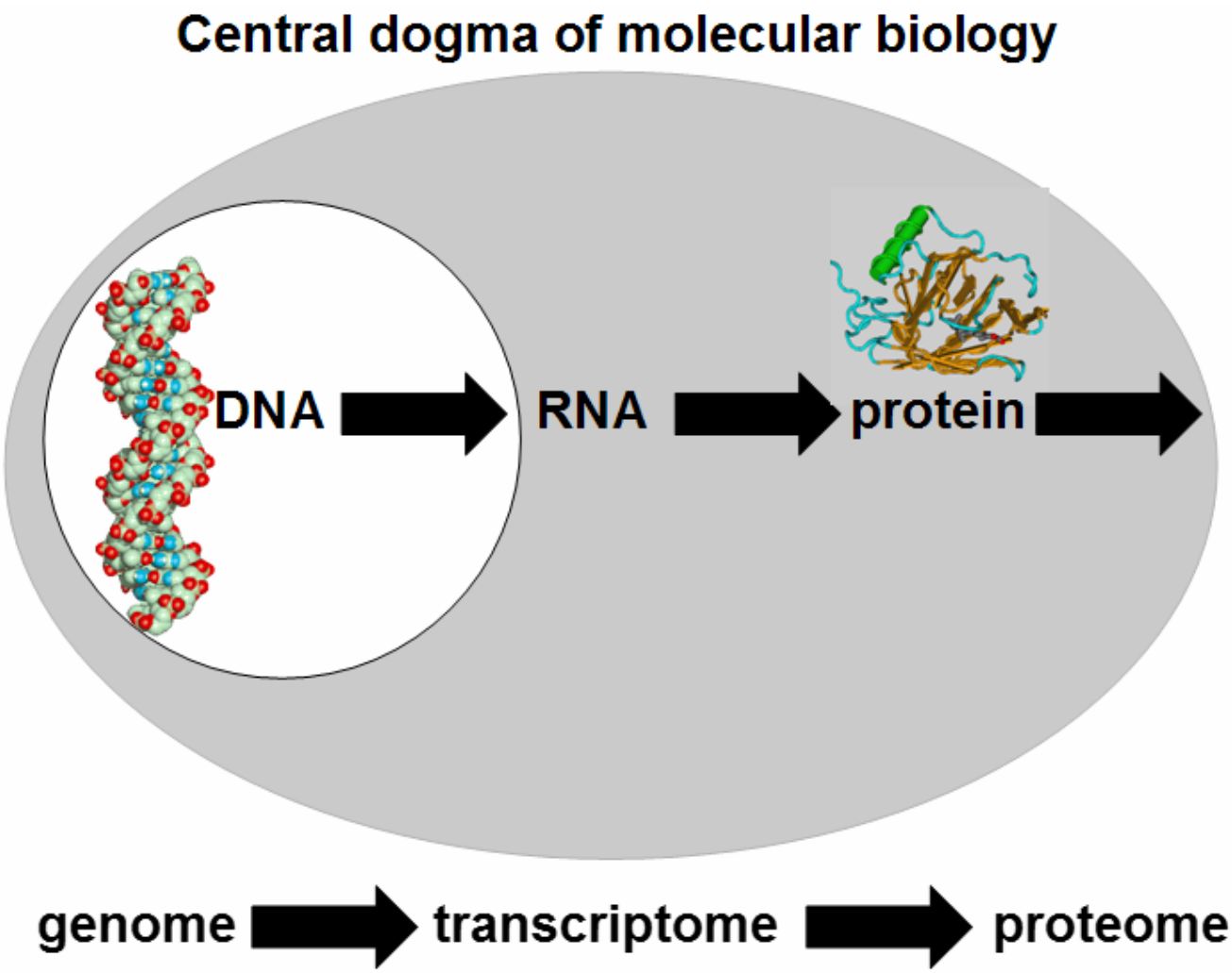
What can we do with these huge amount of data?

- Store (databases)
- Search (PSI-BLAST)
- Analyze / Annotate (visualization, interpretation, classification, pattern recognition)
- Generate new biological knowledge and build biological models (prediction)

Typical Problems

- Where is a gene? (about 1.5% human DNA codes gene)
- Are two protein sequences similar? Implication?
- Find protein sequences similar to my protein
- What does a protein’s structure look like?
- Does this gene mutation relate to a disease (breast cancer)? How?

Where are genes?



Central dogma of bioinformatics and genomics

Genomics, transcriptomics, and proteomics

Source: J. Pevsner, 2005

Ten Topics

- 1. Introduction to Molecular Biology and Bioinformatics
- 2. Pairwise Sequence Alignment Using Dynamic Programming
- 3. Practical Sequence/Profile Alignment Using Fast Heuristic Methods (BLAST and PSI-BLAST)
- 4. Multiple Sequence Alignment
- 5. Gene and Motif Identification
- 6. Phylogenetic Analysis
- 7. Protein Structure Analysis and Prediction
- 8. RNA Secondary Structure Prediction
- 9. Clustering and Classification of Gene Expression Data
- 10. Search and Mining of Biological Databases, Databanks, and Literature