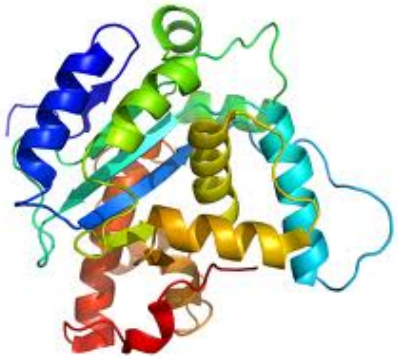
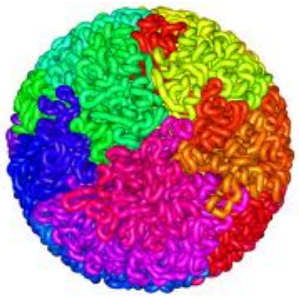





Computational Modeling of Molecular Structure

Jianlin Cheng, PhD

Department of EECS

Informatics Institute

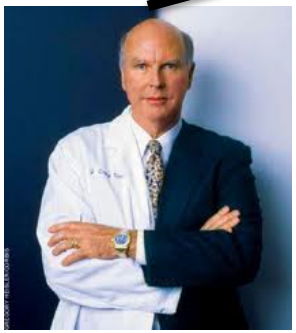
University of Missouri, Columbia

Spring, 2018

The Genomic Era

Collins, Venter, Human Genome, 2000

DNA Sequencing Revolution



Scientists

Government

Company

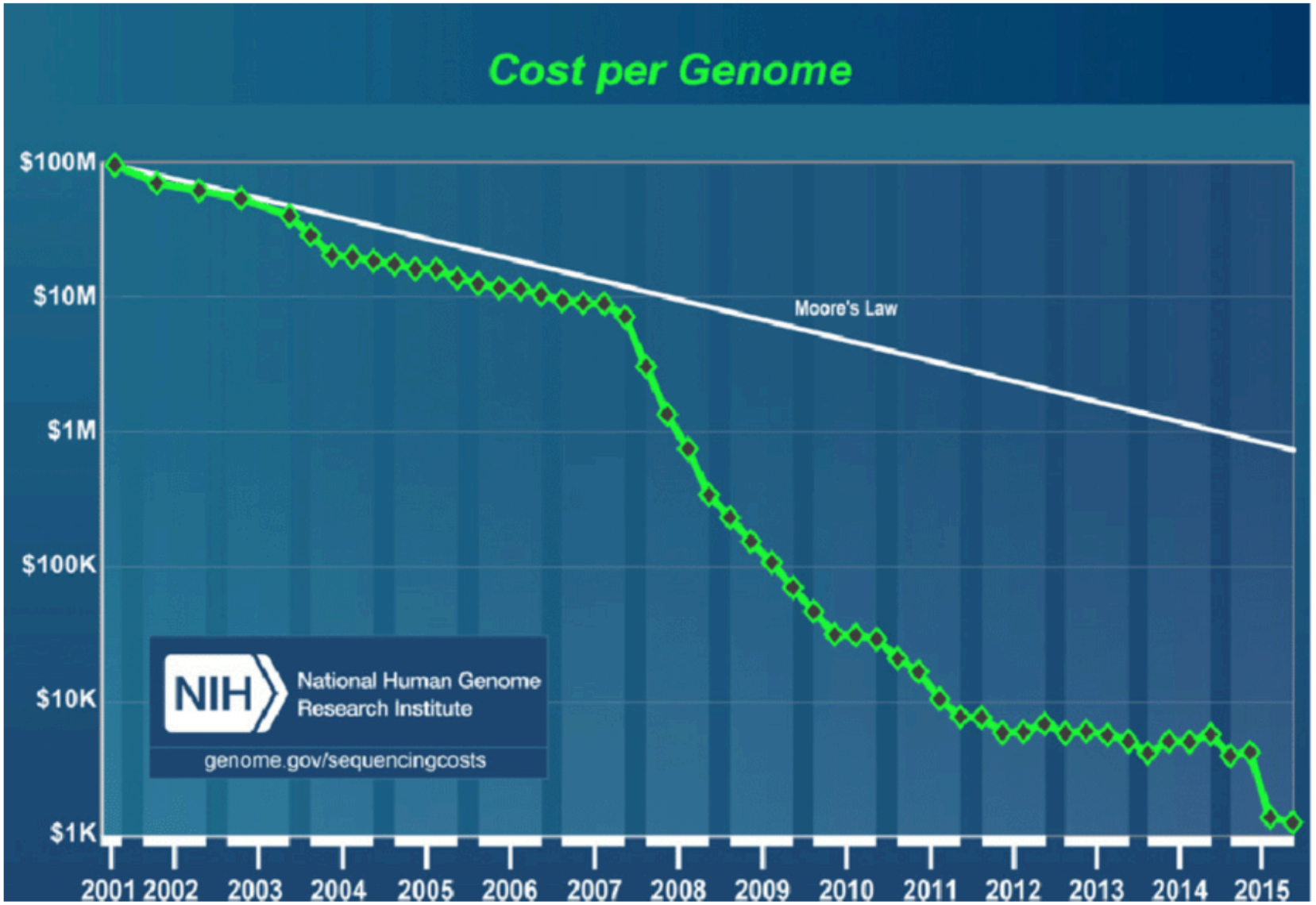
$1000 Personal Genome

A Topic of Big Bio Data Analysis

Science enters $1,000 genome era

By Paul Rincon
Science editor, BBC News website

The HiSeq X Ten is capable of sequencing five human genomes a day, Illumina claims


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

Objectives

- Properties of molecular structures (proteins, RNA, genome / DNA)
- Computational representation of molecular structures
- Data-driven computational modeling of molecular structures
- Application of modeling of molecular structures such as drug design

Significance of Studying Molecular Structures

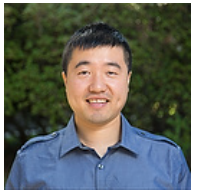
- One foundation of life sciences
- Personal healthcare and medicine
- One major topic of bioinformatics and computational biology – an important field of computer science
- **A great application area of computer algorithms and data science**
- **A very interdisciplinary field (CS, data science, stats, math, biology, chemistry, physics)**

A Good Career for CS Graduates

- Five PhD graduates are assistant / associate professors of bioinformatics in CS departments
- One PhD student secured a scientist position in a bioinformatics company
- One PhD student works for Microsoft
- Numerous other graduate students received good training and worked in data-intensive fields.

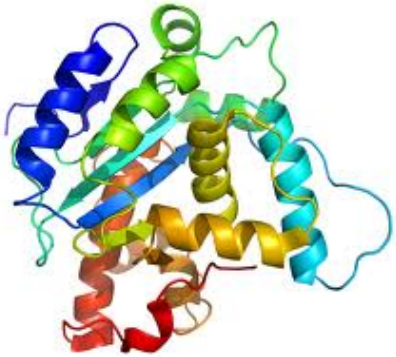




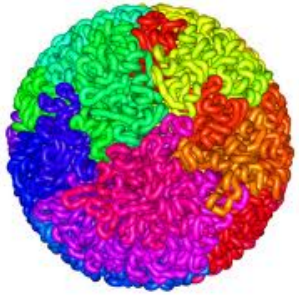




Three Kinds of Structures

- **Protein Structure**



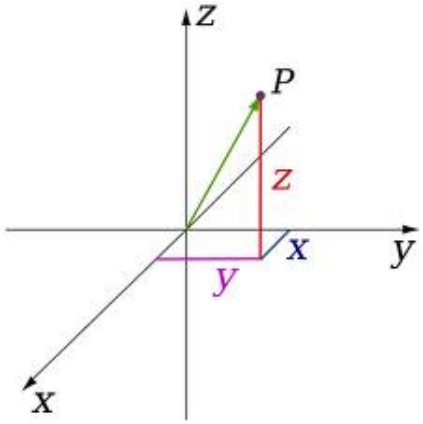
- **Genome Structure**

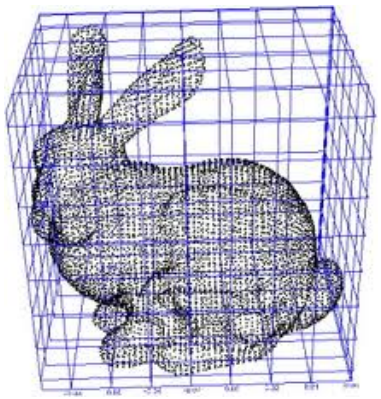


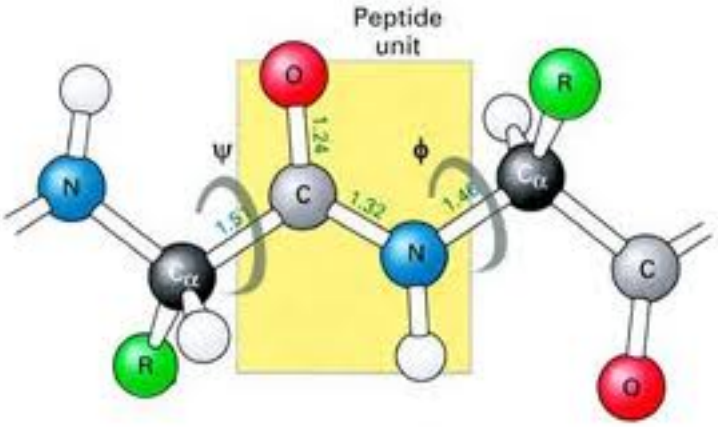
- **RNA Structure**

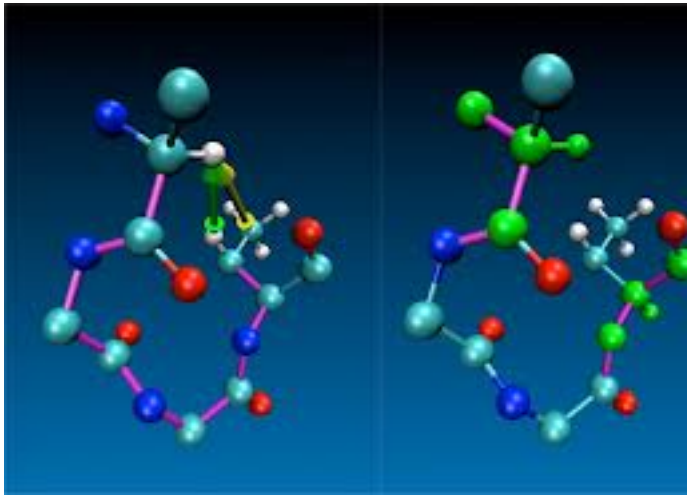
Representation of Molecular Structures

- X, Y, Z coordinates
- Euclidean grid
- Vector and angles
- Computer graphics









Algorithms

- Grid-based simulation (random walk)
- Vector-based simulation
- Angular-based simulation
- Gradient descent simulation and variants
- Simulated annealing
- Markov Chain Monte Carlo
- Probabilistic modeling
- Constraint-based optimization
- Machine learning (e.g., deep learning)

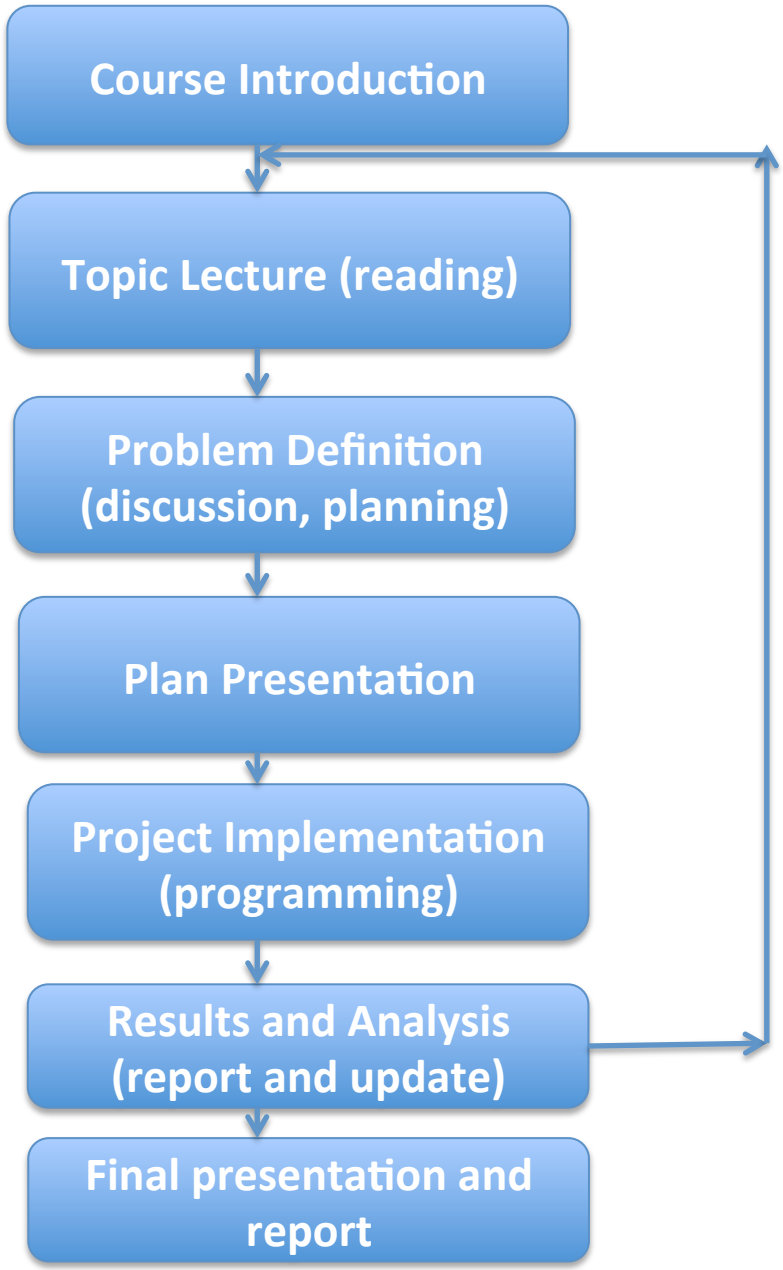
Software Packages

- RasMol, Jmol, PyMol, Chimera
- Modeller, Rosetta, I-TASSER, MULTICOM, MTMG, CNS, CONFOLD, Zdock, MOGEN, LorDG, AutoDock
- Your own algorithm, implementation, and practice

Course Format

- Course web site: http://calla.rnet.missouri.edu/cheng_courses/cscmms2018/
- Problem solving
- Active learning by practicing
- Syllabus (see details)

Teaching Format of Each Topic



Group:

3-4 students per group

Rotate as topic coordinator

Each member participates in every topic

All members present the whole project

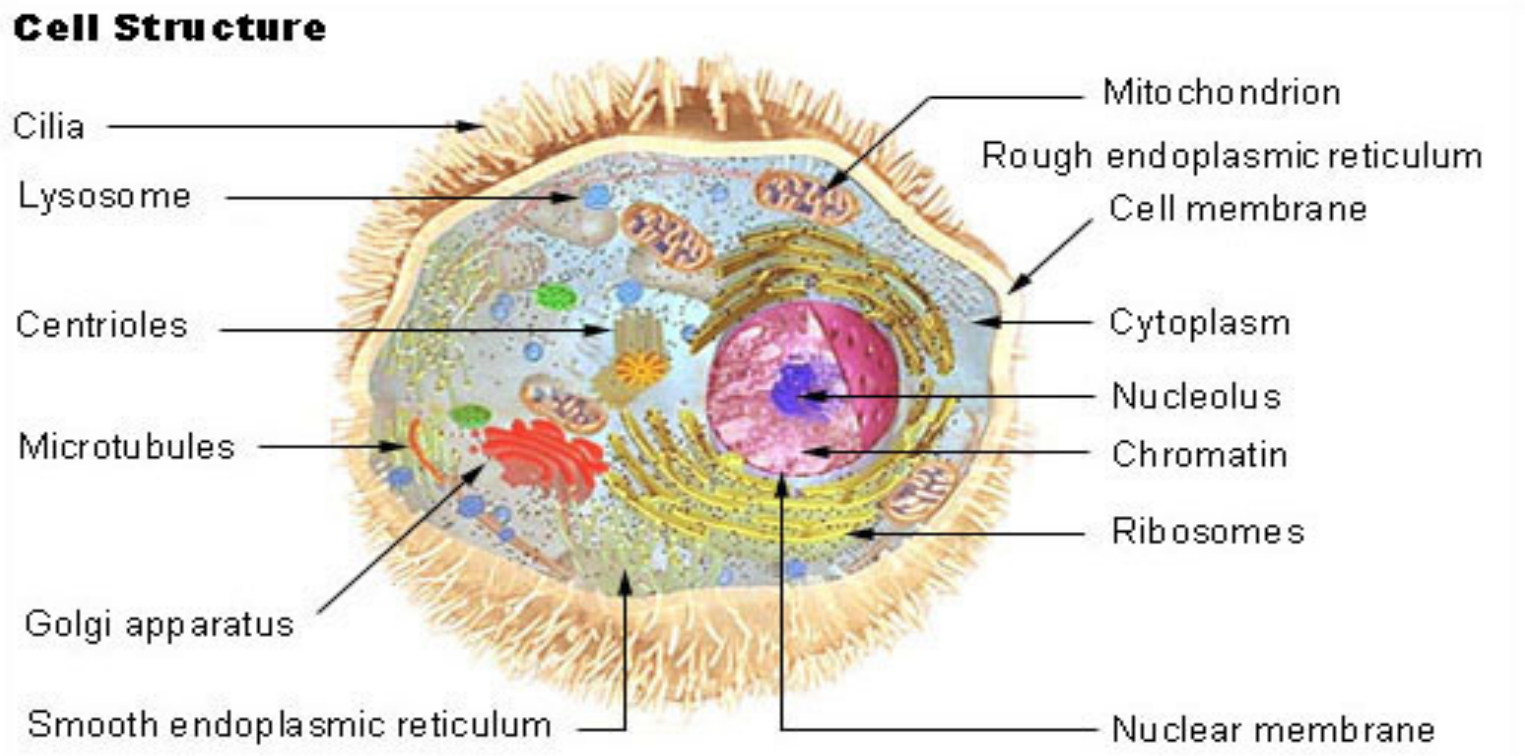
Grading

- Class discussions (15%)
- Literature reviews (10%)
- Topic plan presentation (20%, group)
- Topic implementation and report (45%, group)
- Final presentation (10%, group)
- Grade scale: A+, A, A-, B+, B, B-, C+, C, C-, and F.

Introduction to Molecular Biology for Computer Science and Engineering Students

Introduction to Molecular Biology

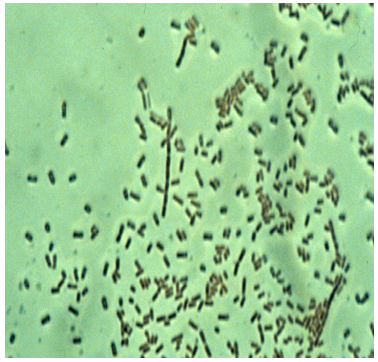
- Cell is the unit of structure and function of all living things.

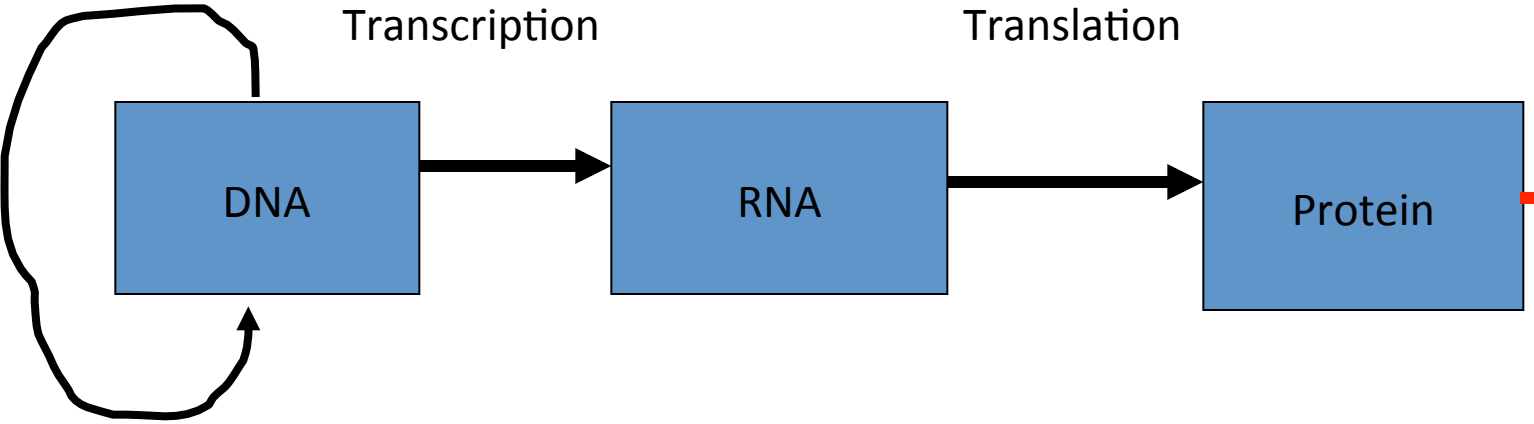


Two types of cells: eukaryote (higher organisms) and prokaryote (lower organisms)

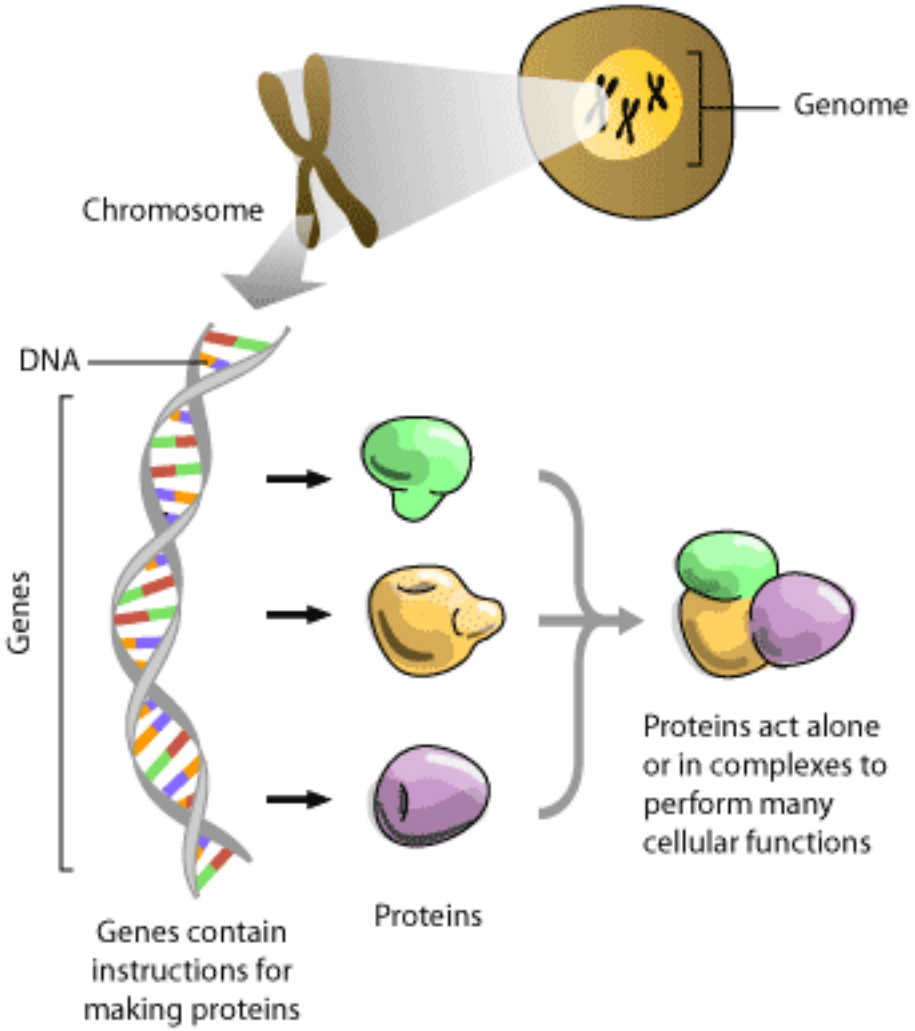
Central Dogma of Molecular Biology

Phenotype

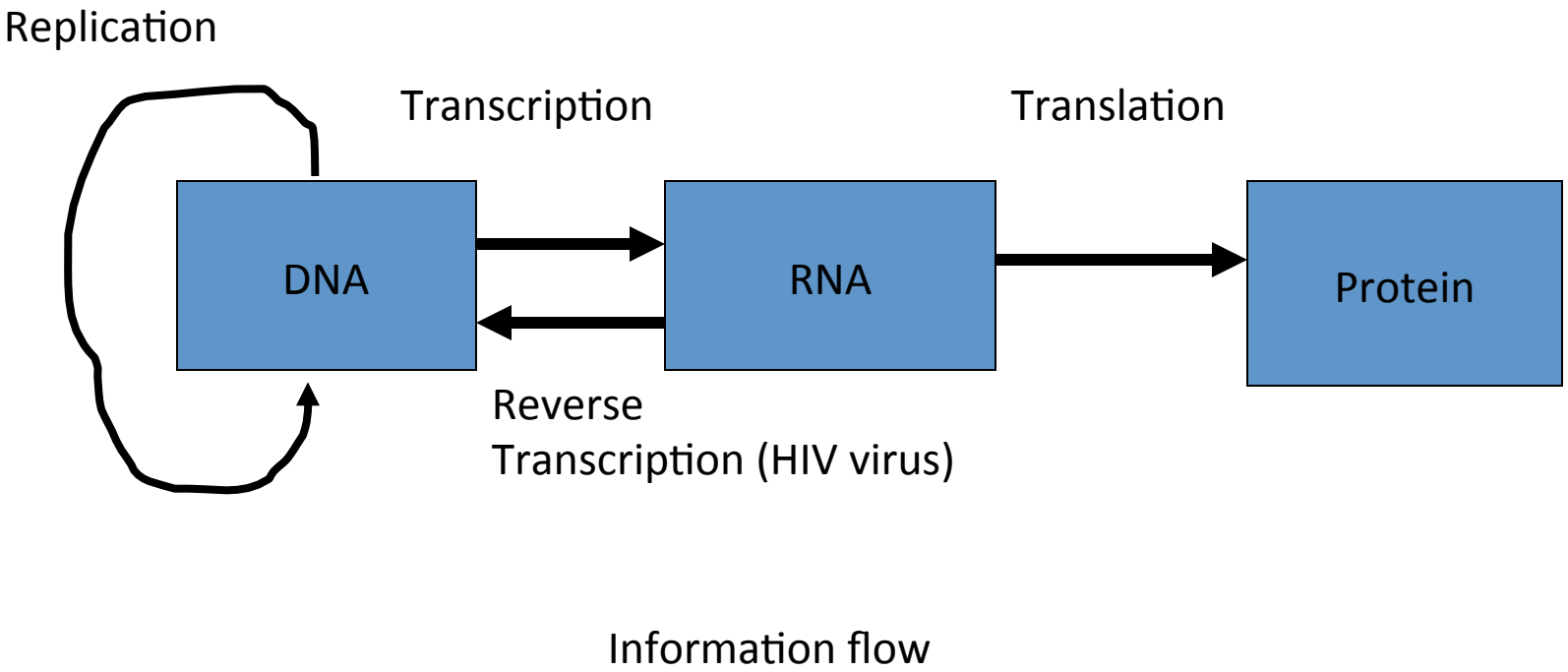


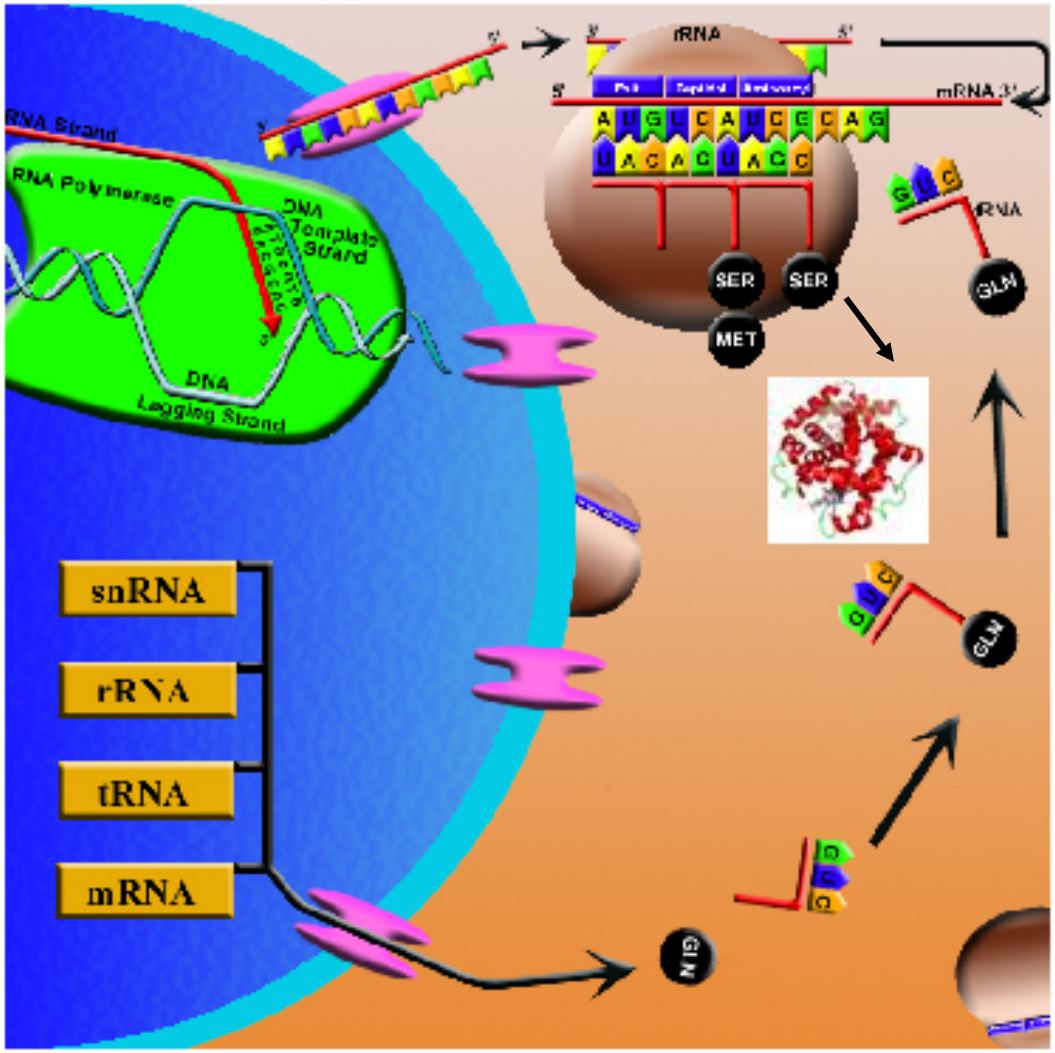


Genotype

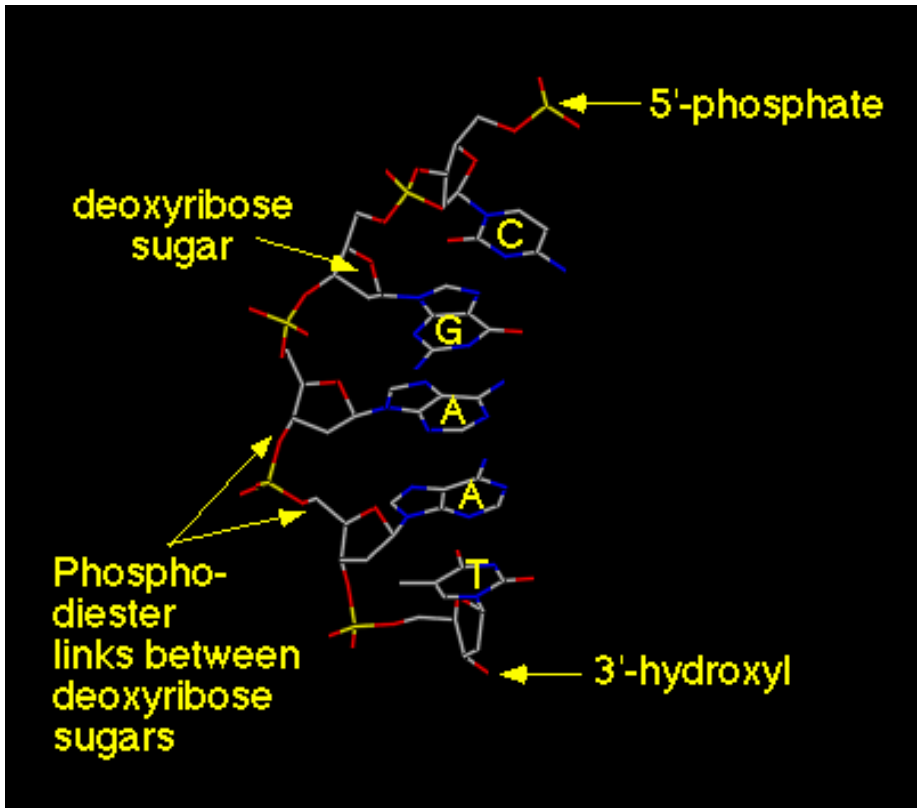

Genome
Chromosome
DNA
Genes
Genes contain instructions for making proteins
Proteins
Proteins act alone or in complexes to perform many cellular functions

Central Dogma of Molecular Biology




RNA Strand
RNA Polymerase
DNA Template Strand
DNA Leggling Strand
rRNA
mRNA 3'
SER
SER
MET
GLN
snRNA
rRNA
tRNA
mRNA

DNA (Deoxyribose Nucleotide Acids)

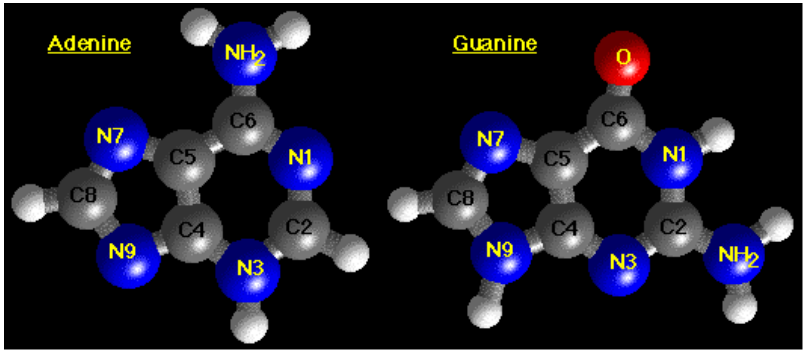


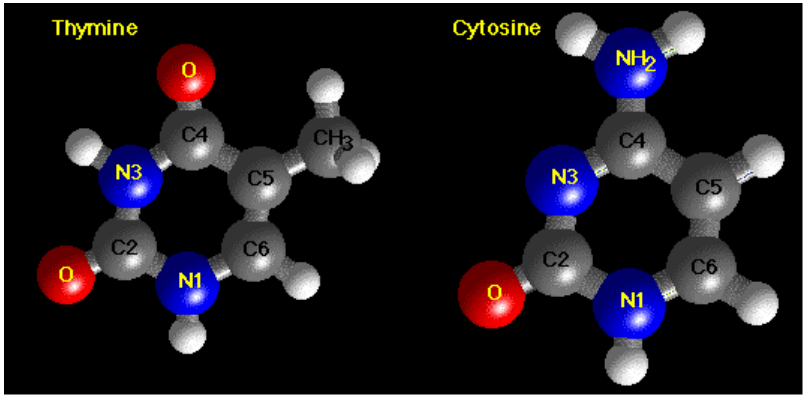
DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a "polynucleotide." Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group.

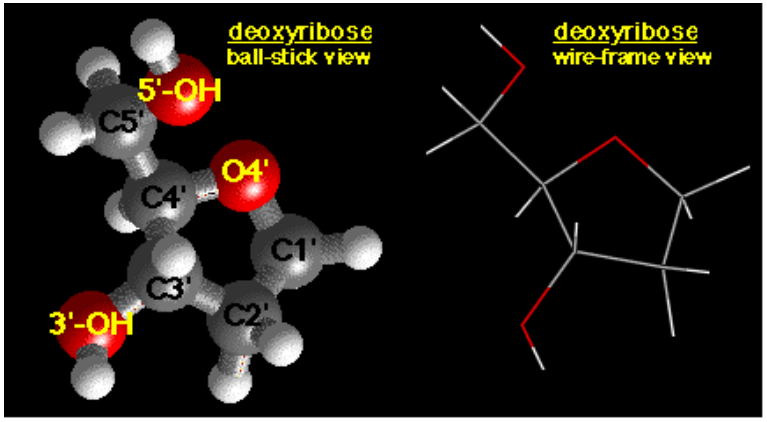
CGAATGGGAAA……

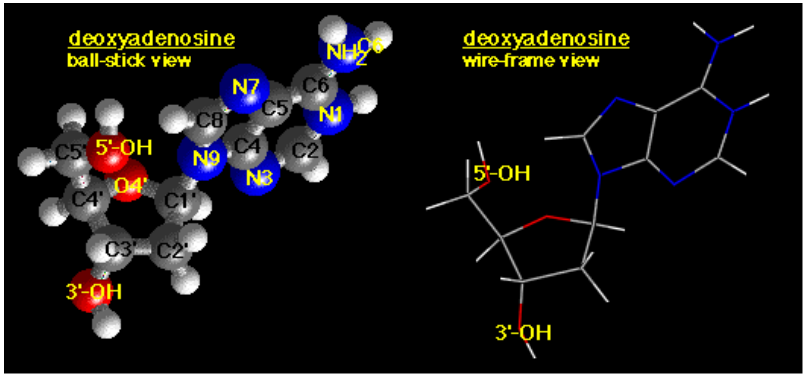
A is for adenine
G is for guanine
C is for cytosine
T is for thymine

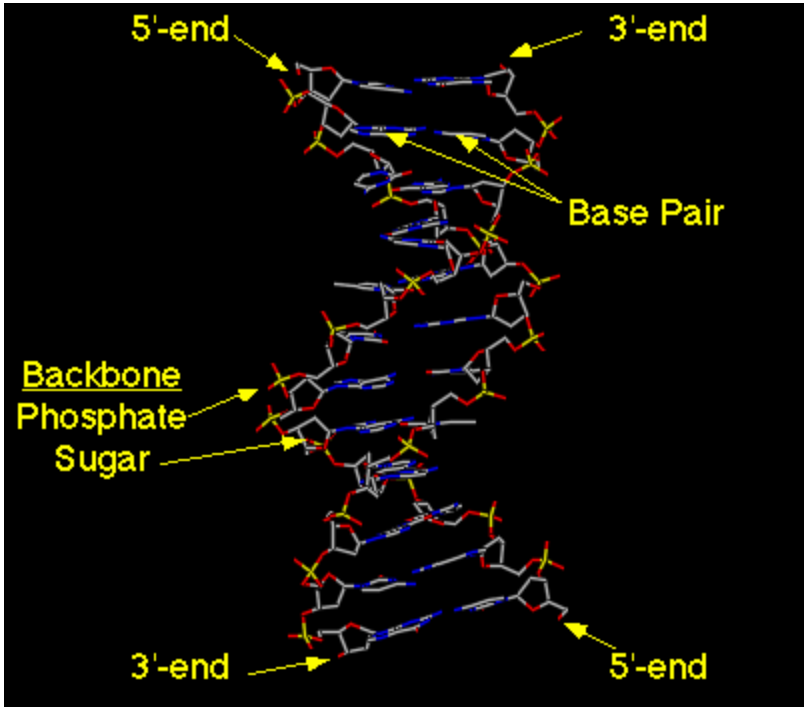
Introduction to DNA structure, Richard B. Hallick, 1995

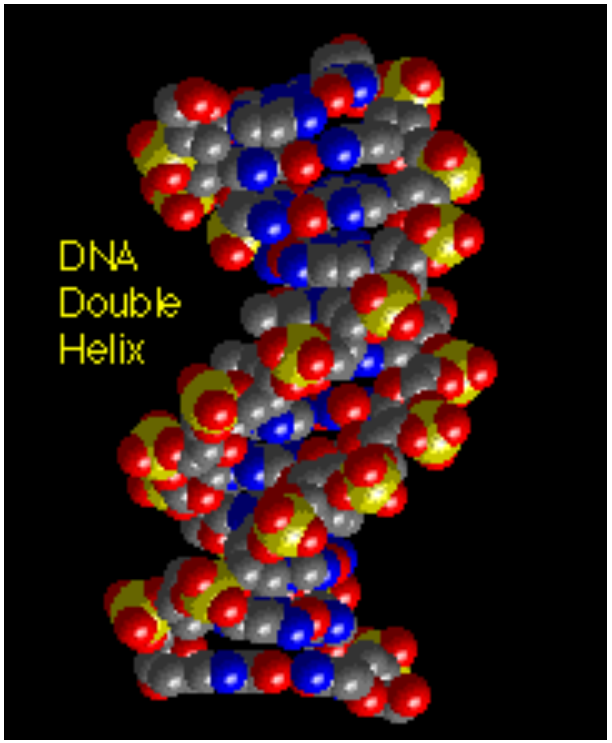

Adenine
NH2
N7
C6
C5
N1
C8
C4
C2
N9
N3
Guanine
O
N7
C6
C5
N1
C8
C4
C2
N9
N3
NH2


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

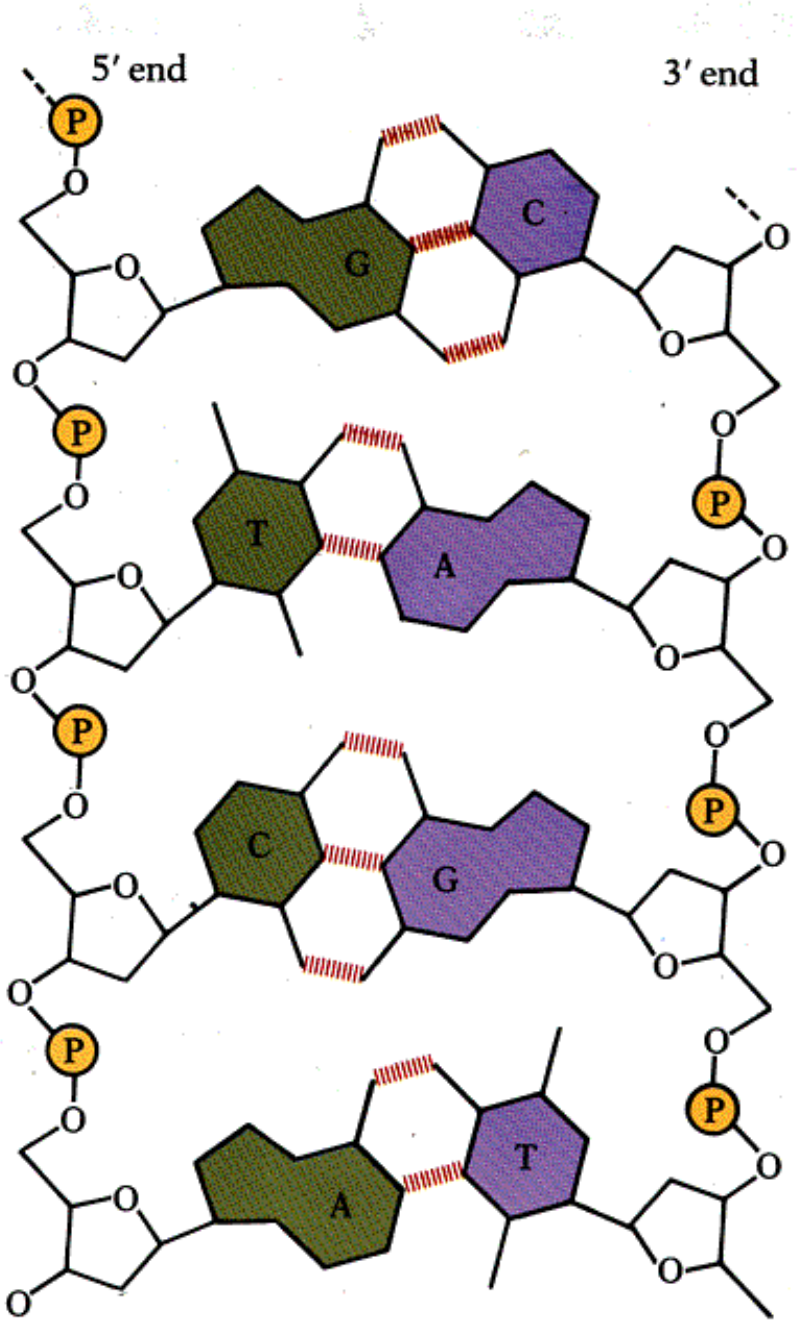

deoxyadenosine
ball-stick view
NH2
N7
C6
C8
C5
N1
C5'
5'-OH
N9
C4
C2
O4'
N3
C4'
C1'
C3'
C2'
3'-OH
deoxyadenosine
wire-frame view
5'-OH
3'-OH





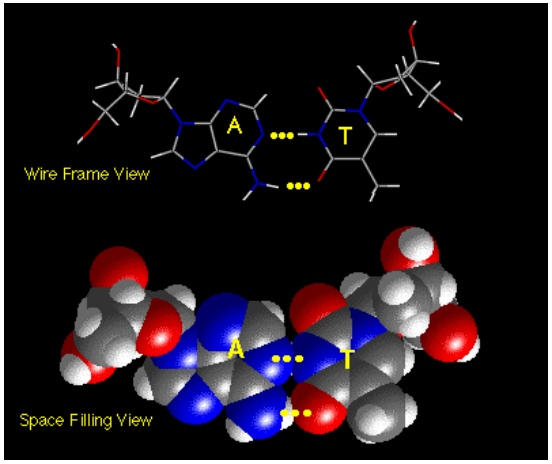
Base Pairs:
A-T (2 H-bonds)
C-G (3 H-bonds)

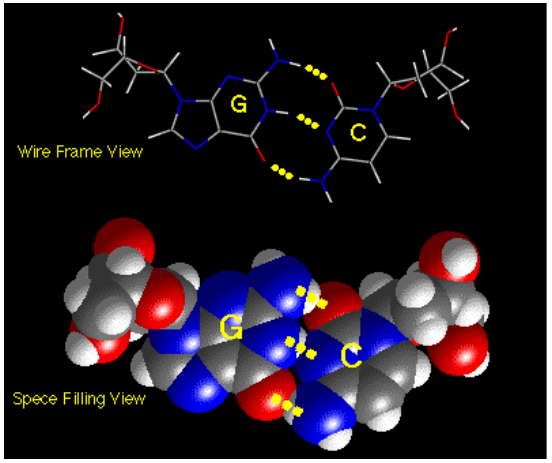
Hydrogen bonds: non-covalent bonds mediated by hydrogen atoms

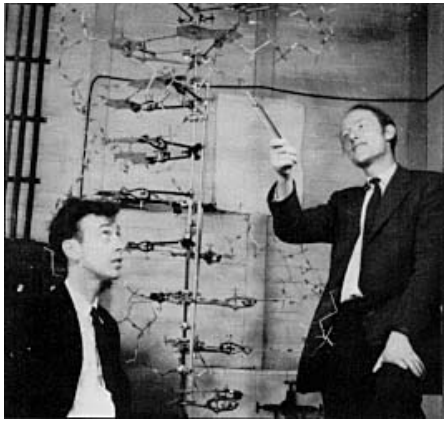


Uncoiled DNA Molecule

Source: Dr. Gary Stormo, 2002







James Watson & Francis Crick

Maurice Wilkins

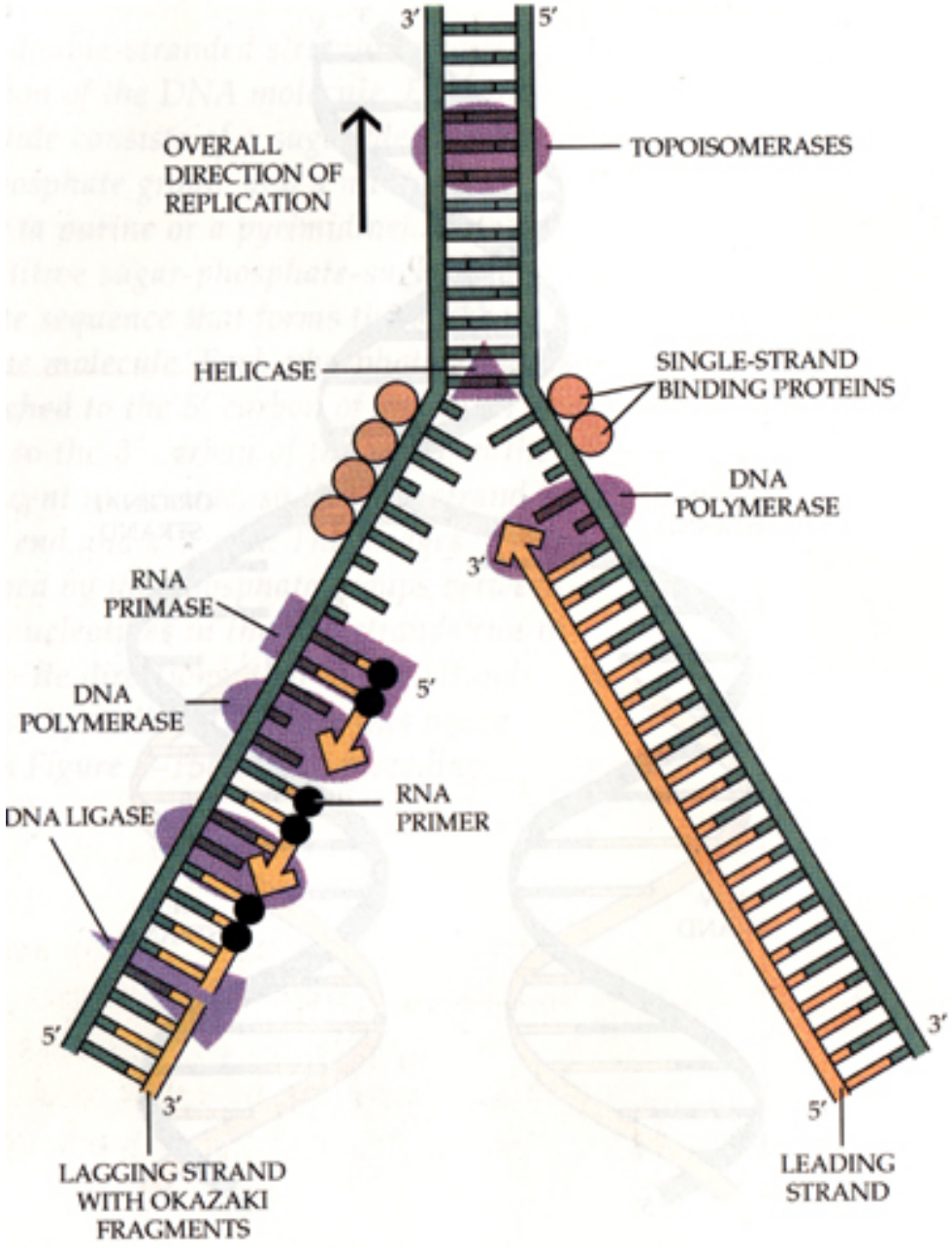
Rosalind Franklin

Linus Pauling

Erwin Chargaff

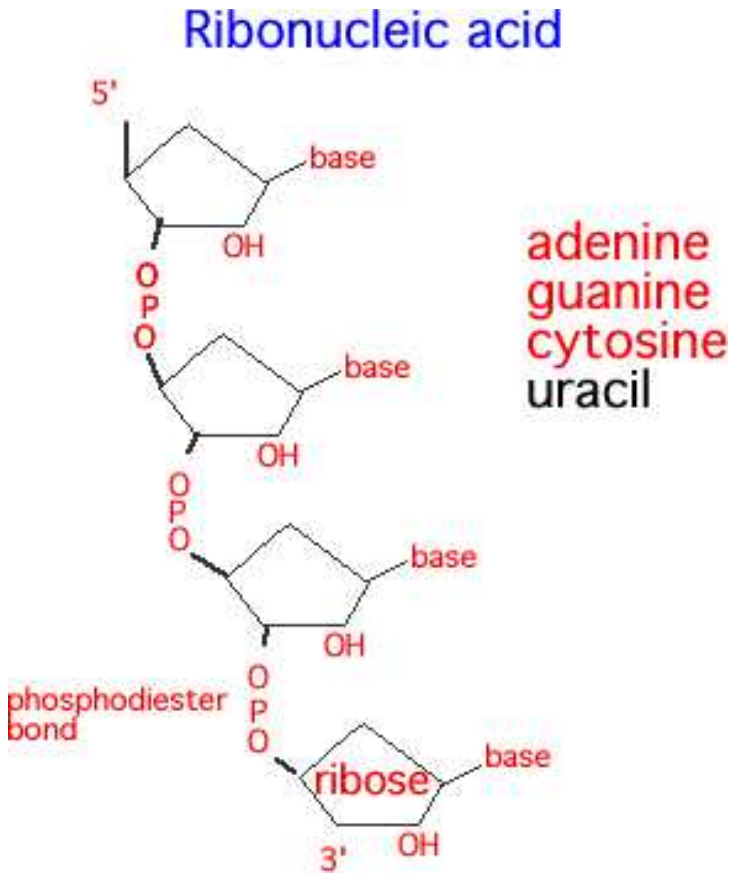
Fundamental Problems: How genetic information pass from one cell to another and from one generation to next generation

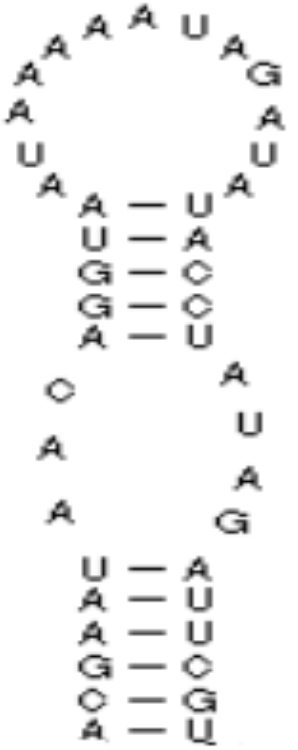
DNA Replication



DNA Polymerase

RNA (Ribose Nucleotide Acids)





ACGAAUAACAGGUAAUAAAAAUAGAUAUACCUAUAGAUUCGU

Different Kinds of RNA

- mRNA: messager RNA

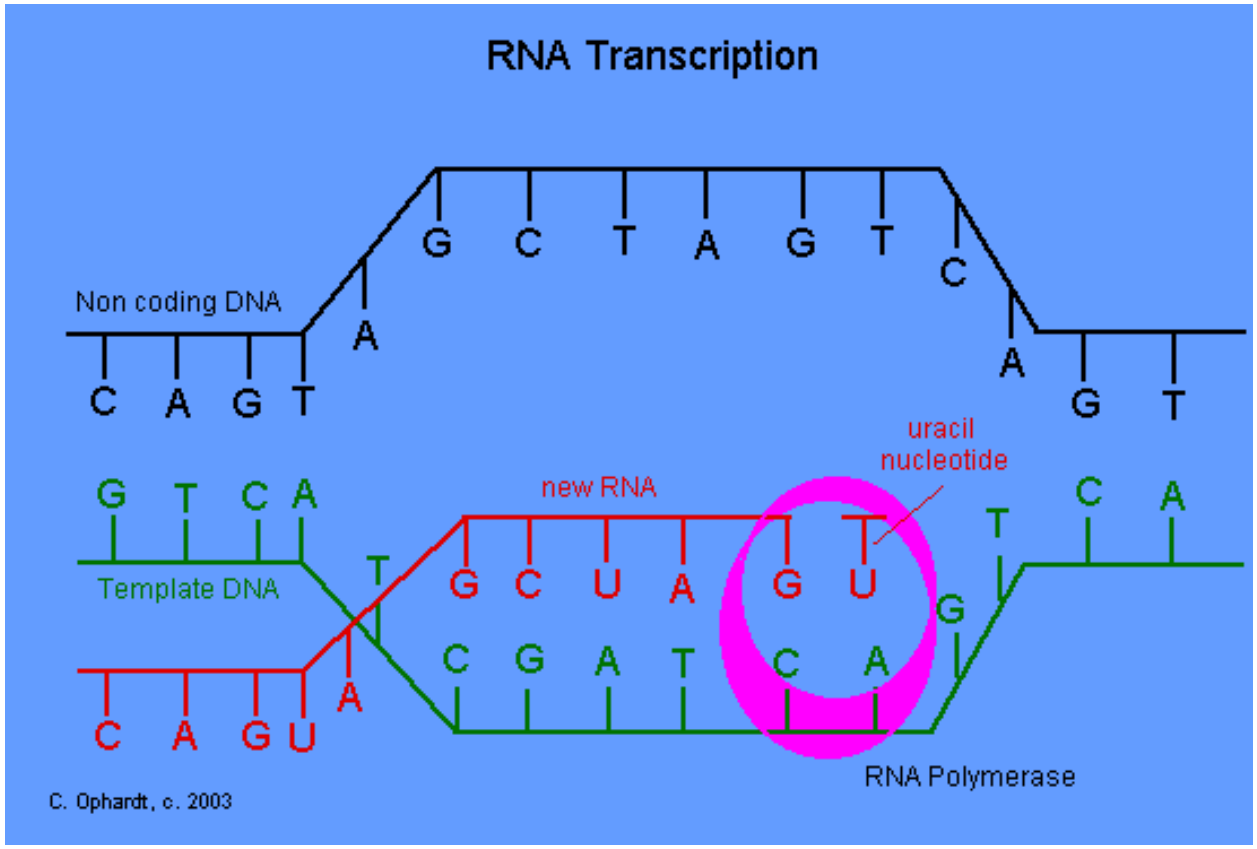
  carry genetic information out of nucleus for protein synthesis (transcription process: RNA polymerase)
- rRNA: ribosomal RNA

  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

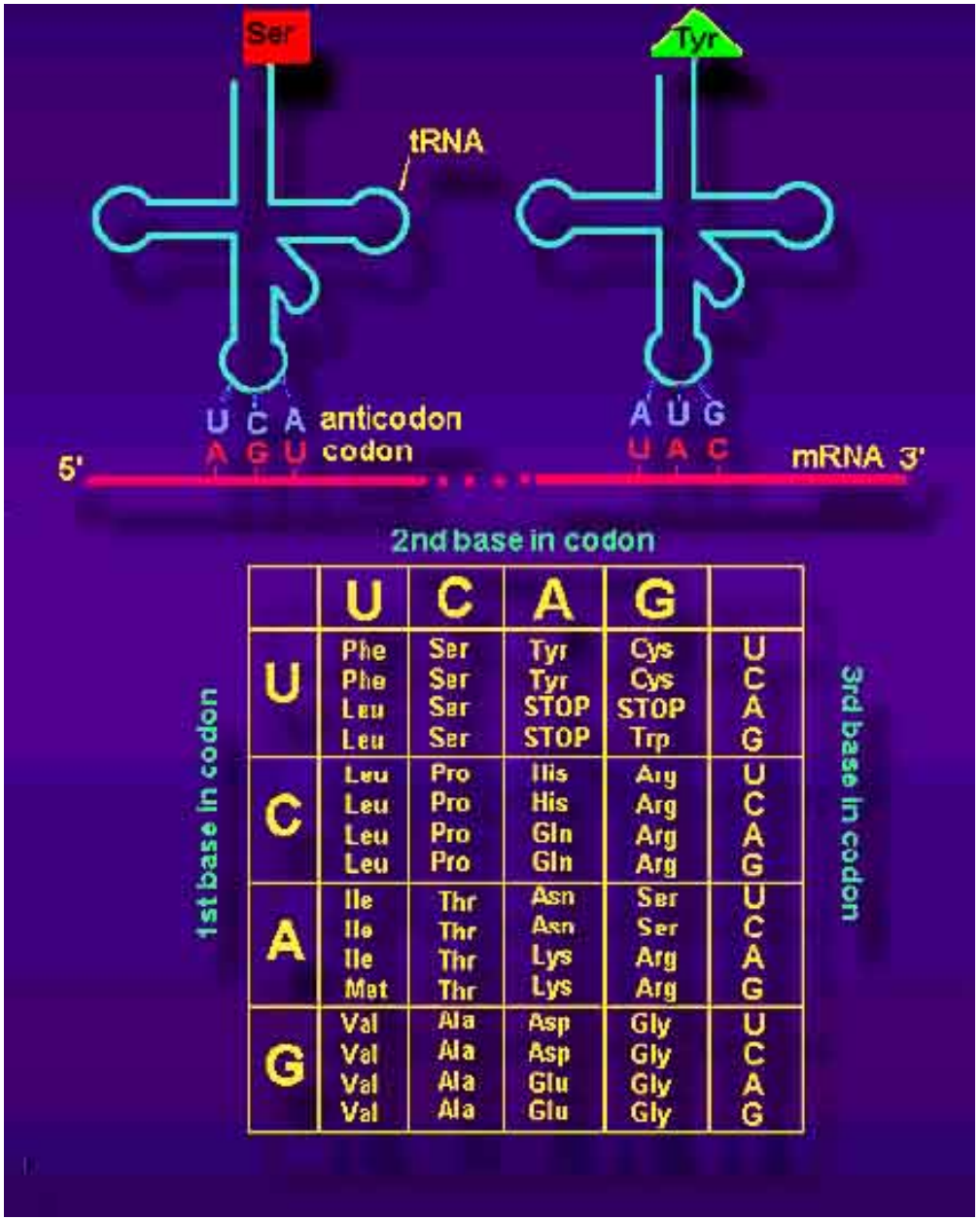
  involve RNA splicing
- Non-coding RNA

Transcription of Gene into RNA



Genetic Code and Translation

Three Nucleotides is called a codon.

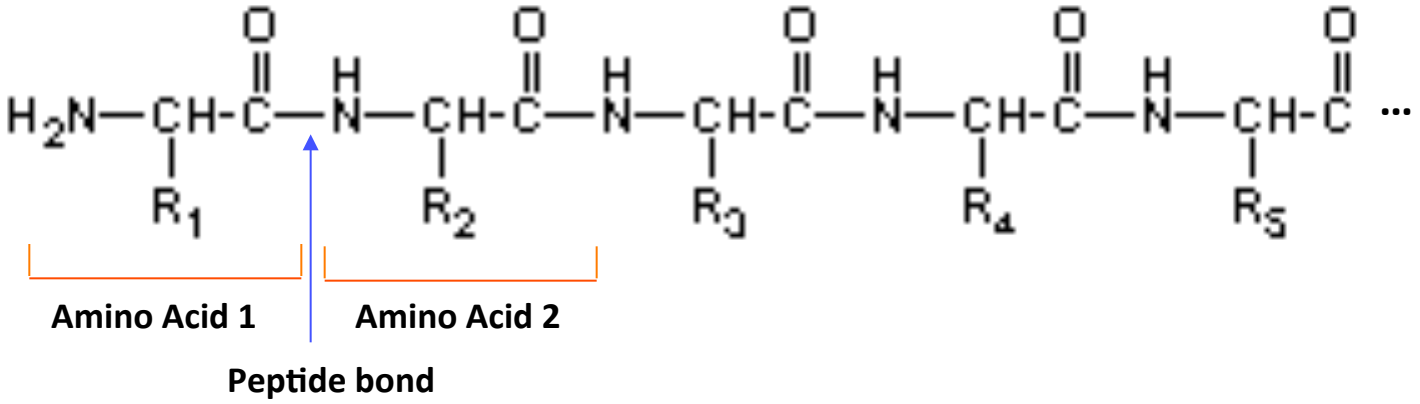


	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A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

Protein Sequence

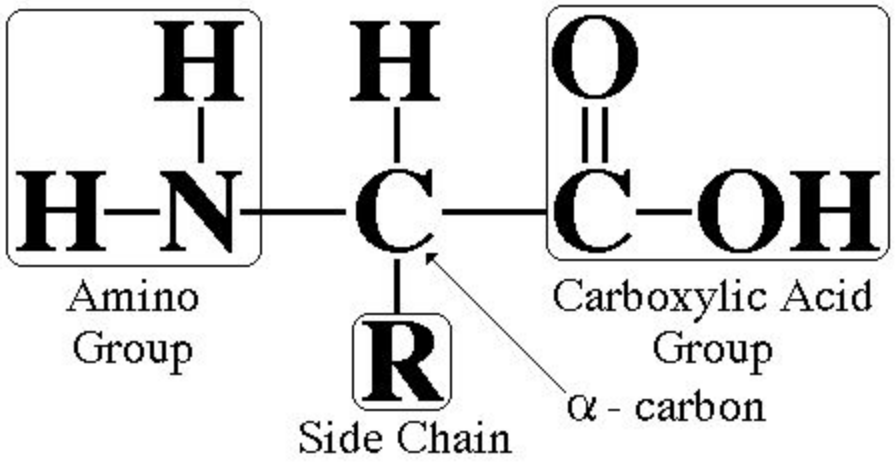
A directional sequence of amino acids/residues

N

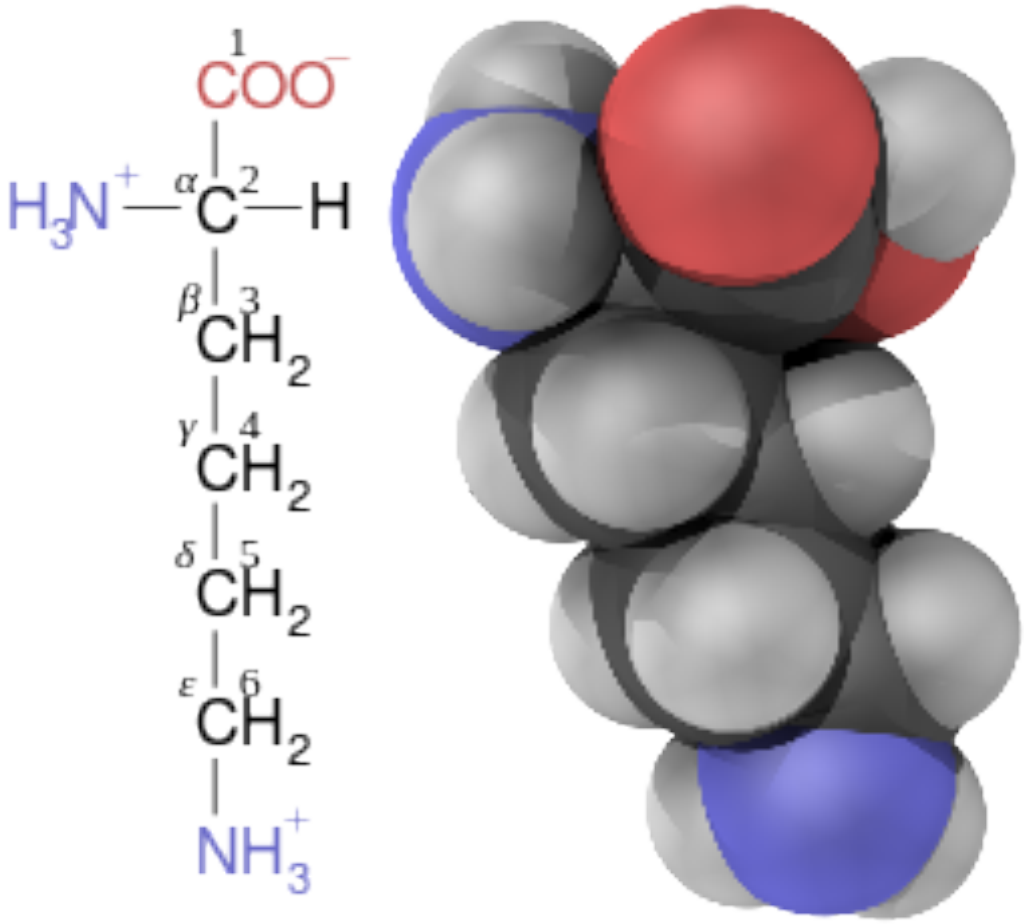


C

Amino Acid Structure



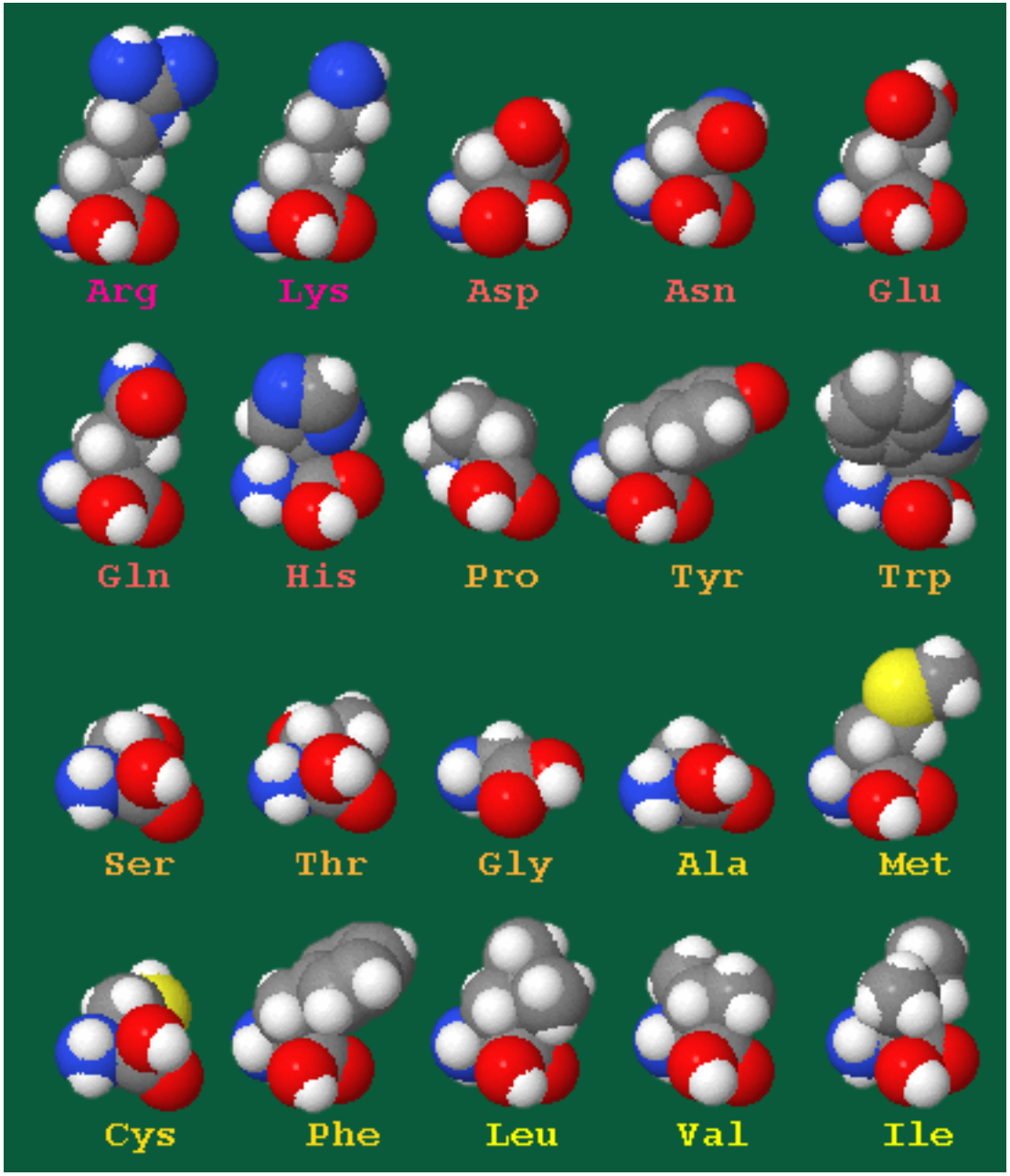
Lysine



Amino Acids

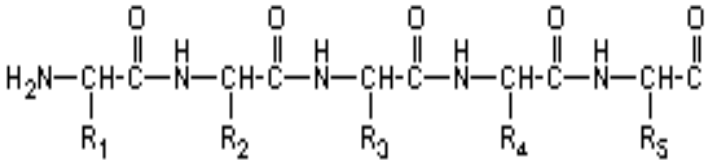
Amino acid	Abbrev.	Side chain	Hydro-phobic	Polar	Charged	Small	Tiny	Aromatic or Aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Alanine	Ala, A	$-CH_3$	X	-	-	X	X	-	67	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	$-CH_2SH$	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	$-CH_2COOH$	-	X	negative	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

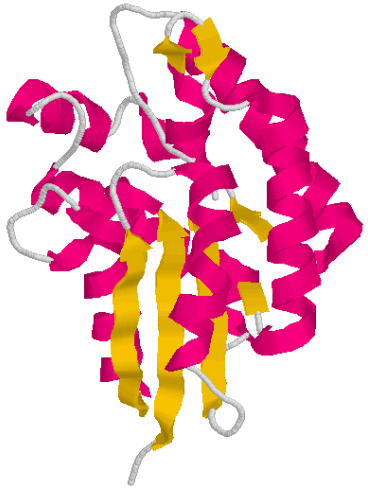
Hydrophilic

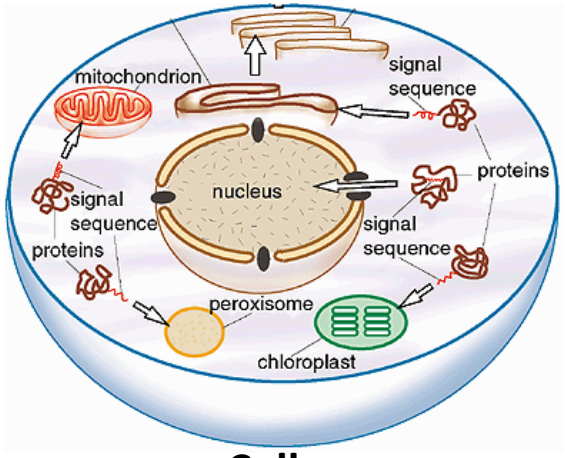

Arg
Lys
Asp
Asn
Glu
Gln
His
Pro
Tyr
Trp
Ser
Thr
Gly
Ala
Met
Cys
Phe
Leu
Val
Ile

Central Dogma of Proteomics

AGCWY……



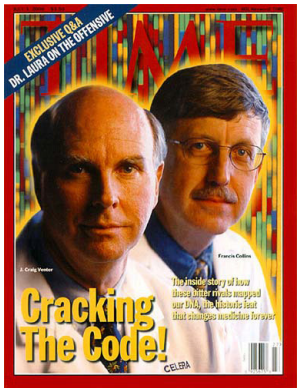




Cell

Sequence **Structure** **Function**

The Genomic Era

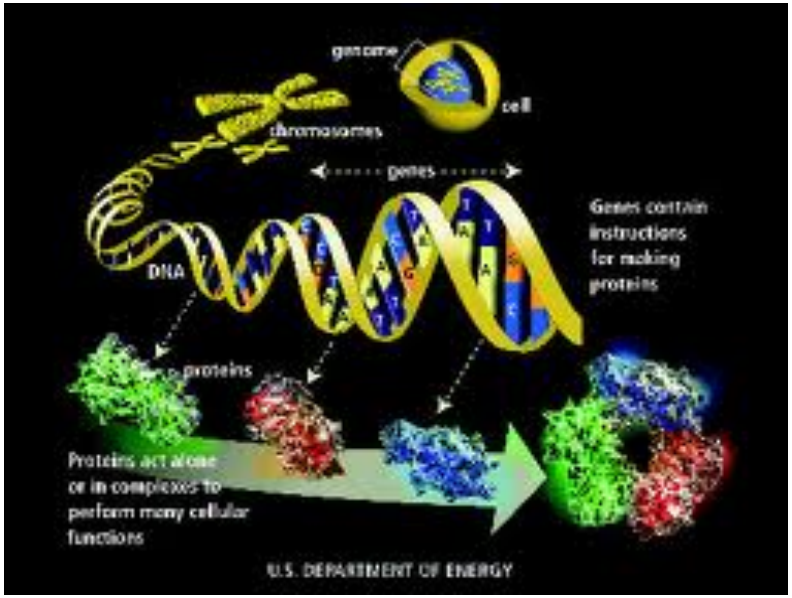


Collins, Venter, Human Genome, 2000

Personal Genome's Implications

- Personalized Disease Prevention
- Personalized Disease Diagnosis
- Personalized Medicine
- Personalized Health Care
- Precision Medicine

Genome Implications to Information Sciences and Life Sciences



Elements and Systems

Assignment One

Read one of the two articles and write a half page summary:
A. Sali. T. Blundell. Comparative Protein Modeling by Satisfaction of Spatial Restraints. JMB, 1993.
J. Li, J. Cheng. A Stochastic Point Cloud Sampling Method for Multi-Template Protein Comparative Modeling. Scientific Reports, 2016.

Submit your review summary (half page)
to mumachinelearning@gmail.com .
Due by Feb. 3 (Saturday).

Form your group (3 to 4 students per group)

Acknowledgements

images.google.com and all the authors providing valuable images