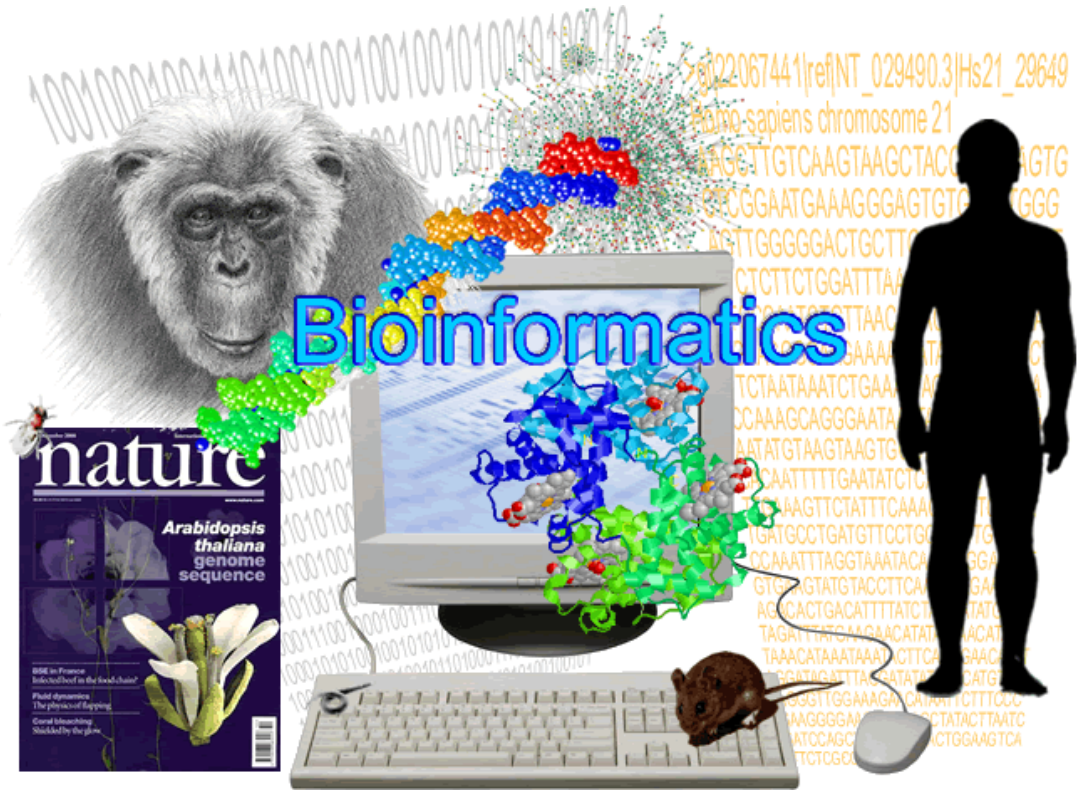
Lecture 1

INTRODUCTION



Outline

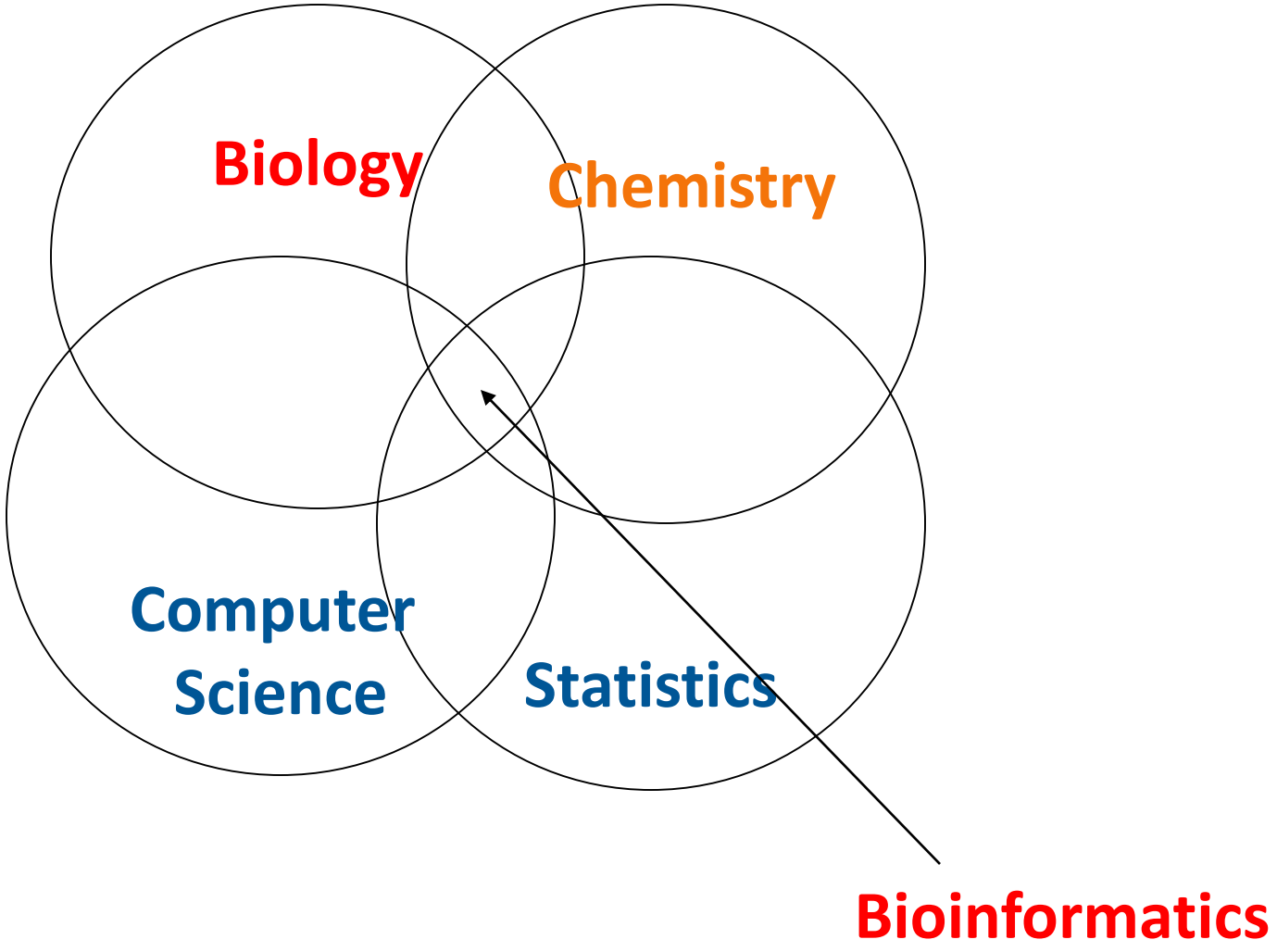
- History of Bioinformatics
- Biological foundations of bioinformatics
- Areas in Bioinformatics

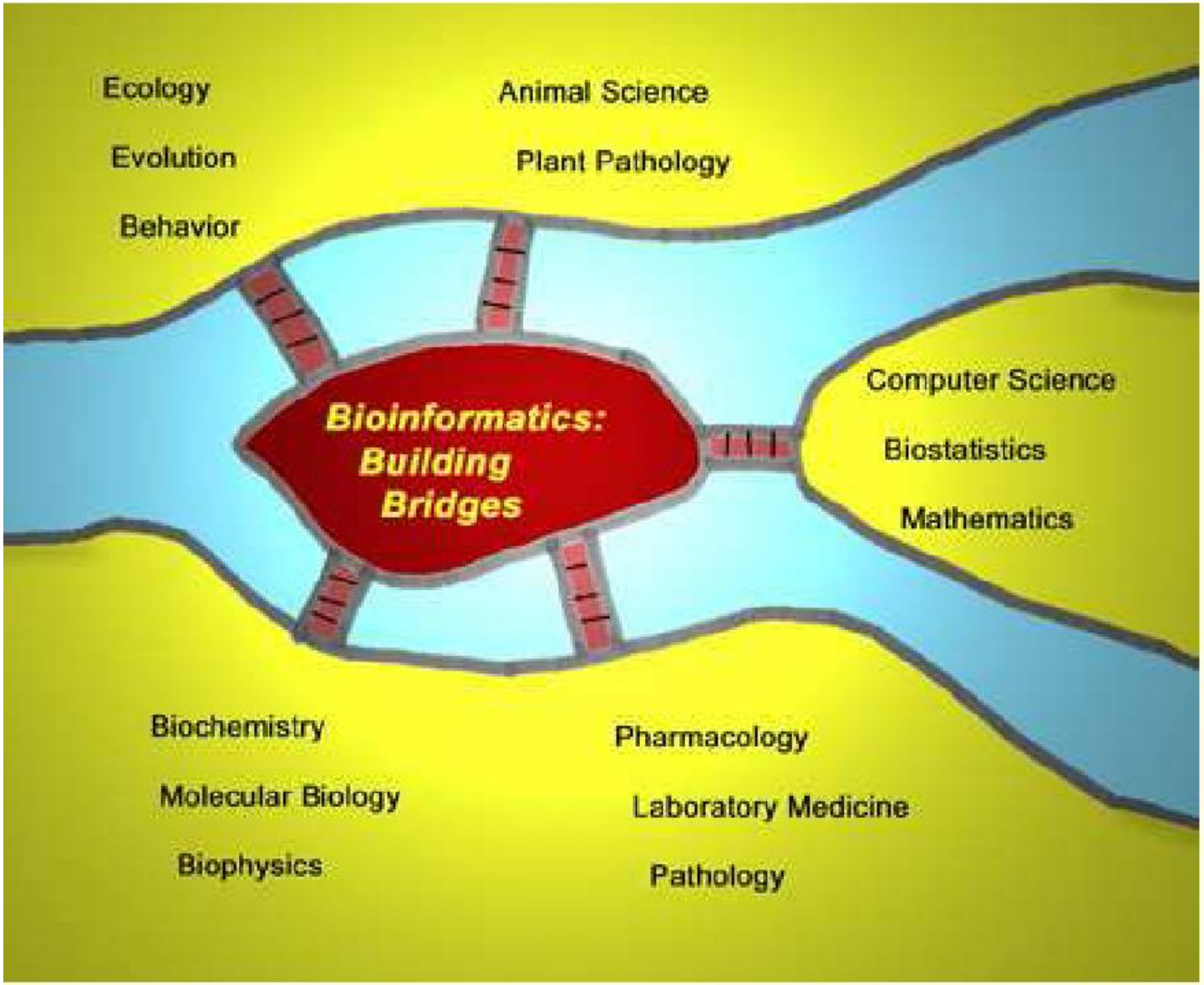
What is Bioinformatics?

"*Bioinformatics is an interdisciplinary research areas at the interface between computer science and biological science*" (Jin Xiong, 2006)

Bioinformatics involves

- The technology that used computers for storage, retrieval, manipulation and distribution of information related to
- Biological macromolecules such as DNA, RNA, proteins


Biology
Chemistry
Computer Science
Statistics
Bioinformatics


Ecology
Evolution
Behavior
Animal Science
Plant Pathology
Bioinformatics: Building Bridges
Computer Science
Biostatistics
Mathematics
Biochemistry
Molecular Biology
Biophysics
Pharmacology
Laboratory Medicine
Pathology

Related field: Computational Biology

The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

- Example: Molecular Dynamics Simulations Suggest that Electrostatic Funnel Directs Binding of Tamiflu to Influenza N1 Neuraminidases

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000939

Related field: Computational Biology

- Bioinformatics is limited to sequence, structural and functional analysis of genes and genomes and their corresponding products
- Computational biology encompasses all biological areas that involve computation
  - Mathematical modeling of ecosystems, population dynamics, phylogenetic structure using fossil records. These analyses do not necessarily involve biological macromolecules

Related field: Computational Biology

How the two terms relate:

- Bioinformatics: develop and apply computational tools in managing all kinds of biological data
- Computational biology is more defined to the theoretical development of algorithms used for bioinformatics

History of bioinformatics

Year	Who or Organization	What
1970	Needleman & Wunsch	First algorithm for comparing protein or DNA sequences
1978	Hogeweg & Hesper	Bioinformatics was defined as the "study of informatic processes in biotic systems"
1977	Sanger et al.	Complete gene sequence of the bacteriophage ϕX174
1980		IntelliGenetics Suite: the first software package for the analysis of DNA & protein sequences

History of bioinformatics

Year	Who or Organization	What
1981	IBM	First personal computer
1982	University of Wisconsin	Software package for molecular biology
1986	Mullis et al.	PCR: a milestone in molecular biology
1986		SWISS-PROT database
1988		National Center for Biotechnology Information (NCBI)
1990		Start of Human genome project

History of bioinformatics

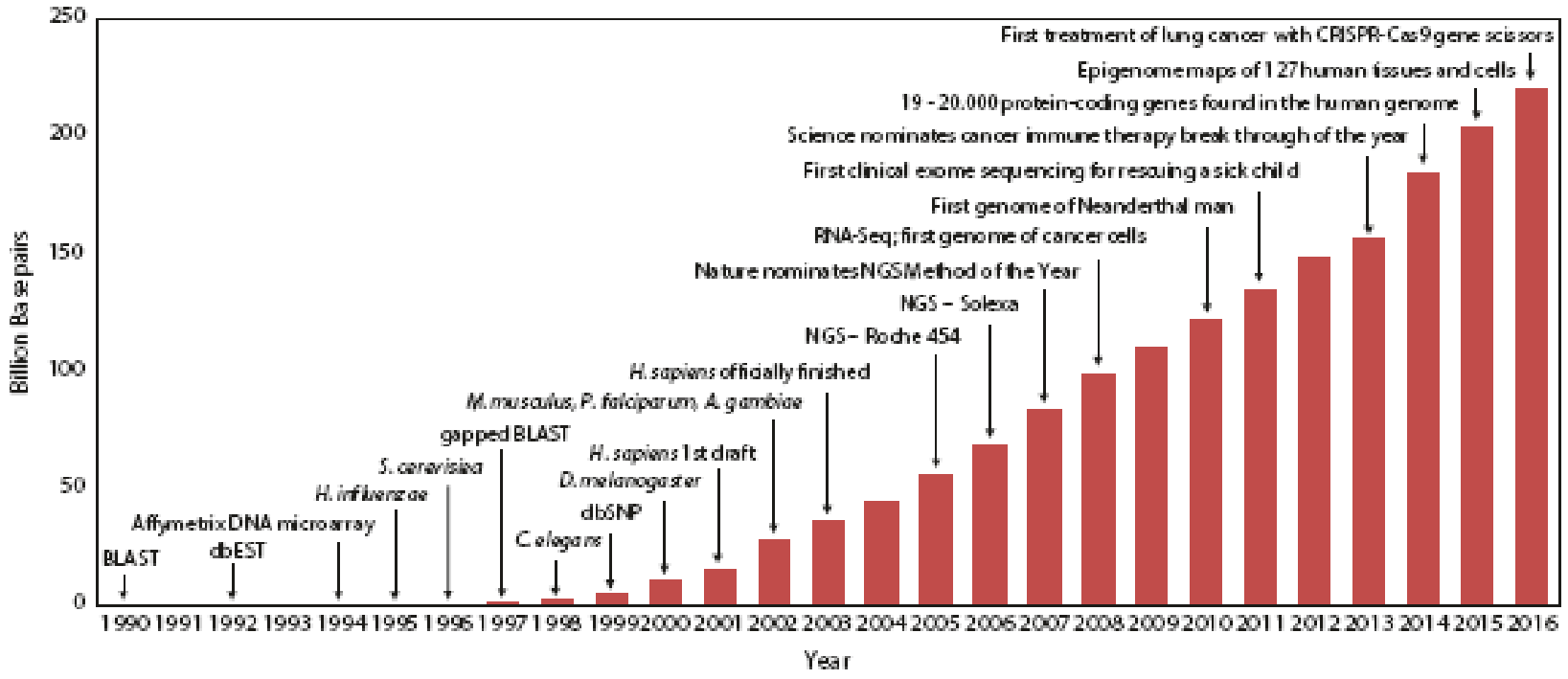
Year	Who or Organization	What
1991	CERN	World Wide Web
1995	Fleischmann et al.	First bacterial genome: *Haemophilus influenzae*
1998	*C. elegans* Sequencing Consortium	First animal genome: *Caenorhabditis elegans*
2002		• European Bioinformatics Institute (EMB-EBI) • Swiss Institute of Bioinformatics (SIB) • Protein Information Resource (PIR) → UniProt database

History of bioinformatics

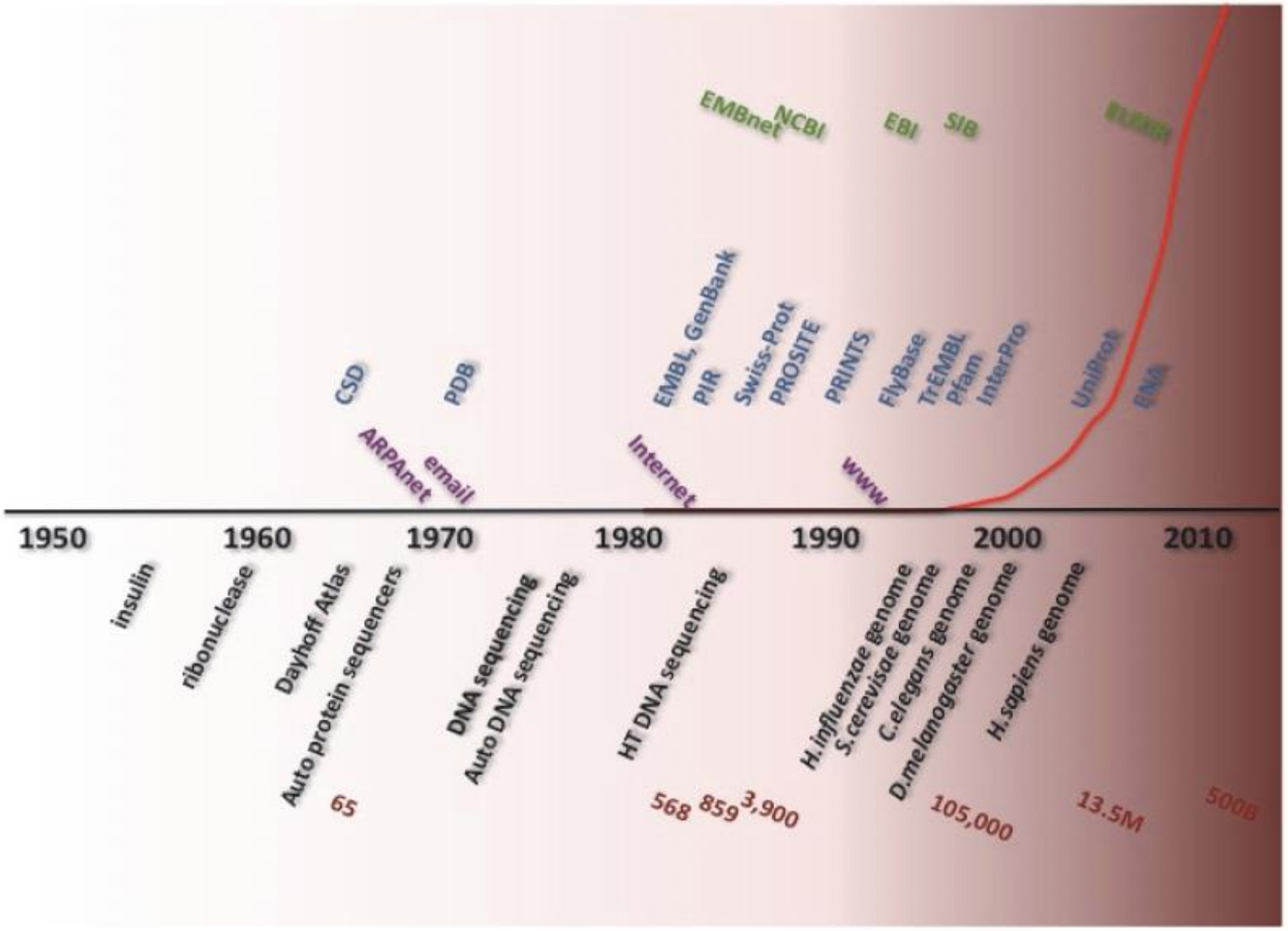
Year	Who or Organization	What
2003		Completion of Human genome project: *Homo sapiens*
2005		454 sequencing: First technique of the next generation sequencing
2006		Illumina sequencing

Why biologists need computer science?

- The advances in molecular biology techniques (such as genome sequencing and microarrays) generated a large amount of data that biologists can't deal with.
- The complexity of modern biological database needs bioinformatics tools.



Development of NCBI's Genbank database in connection with some milestones of bioinformatics



Historical milestones in connection with bioinformatics

Outline

History of Bioinformatics

Biological foundations of bioinformatics

Areas in Bioinformatics

DNA & RNA

- No nucleic acids, no life
- DNA stores information that makes and maintains organisms
  - RNA is the genetic material in some viruses named as riboviruses, including retroviruses
- All the information required to make and maintain a human being is stored in 46 DNA molecules
- Information in DNA guides the manufacture of proteins via RNA, proteins carry out the processes of essential to life
- Changes in DNA information change life forms - Evolution

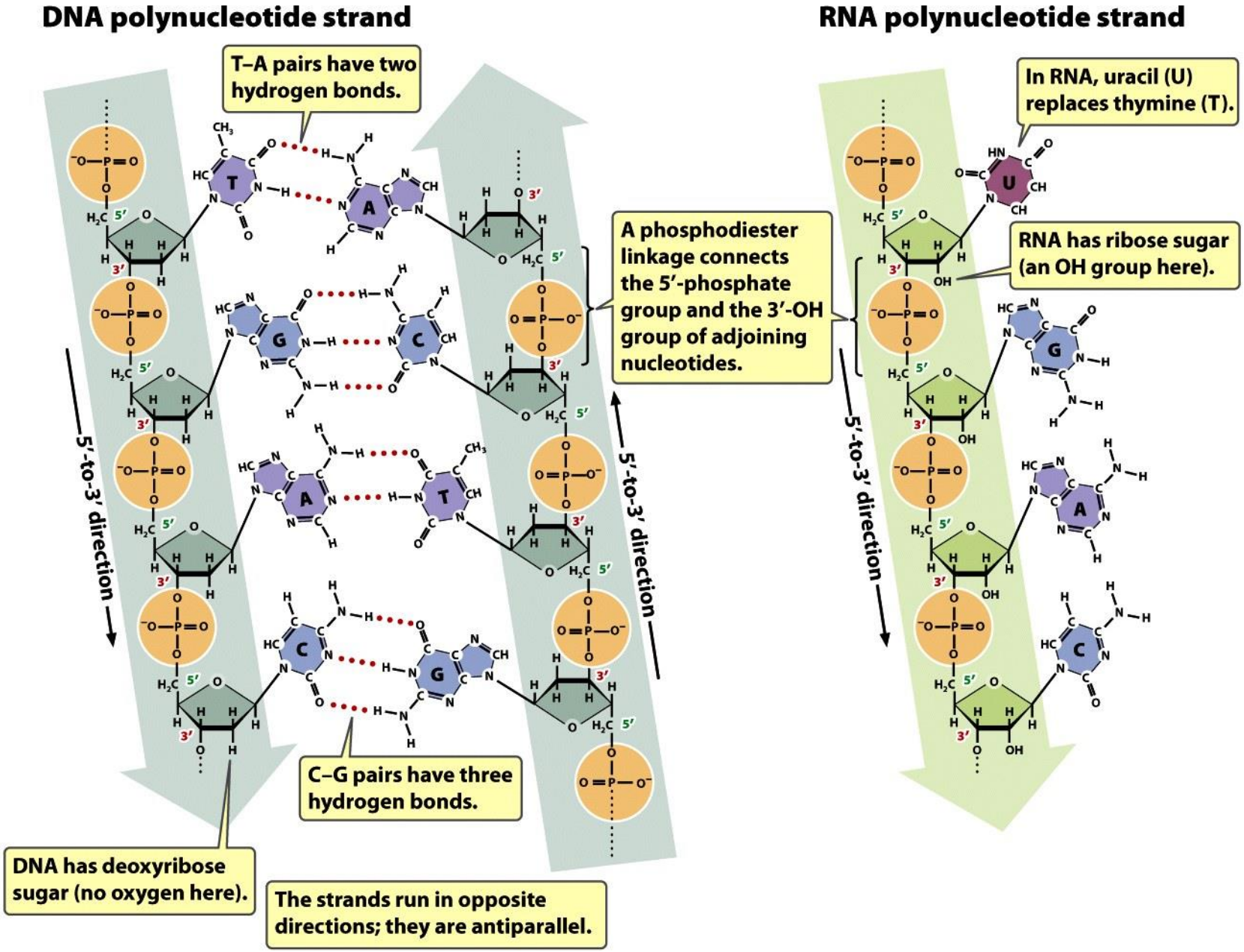
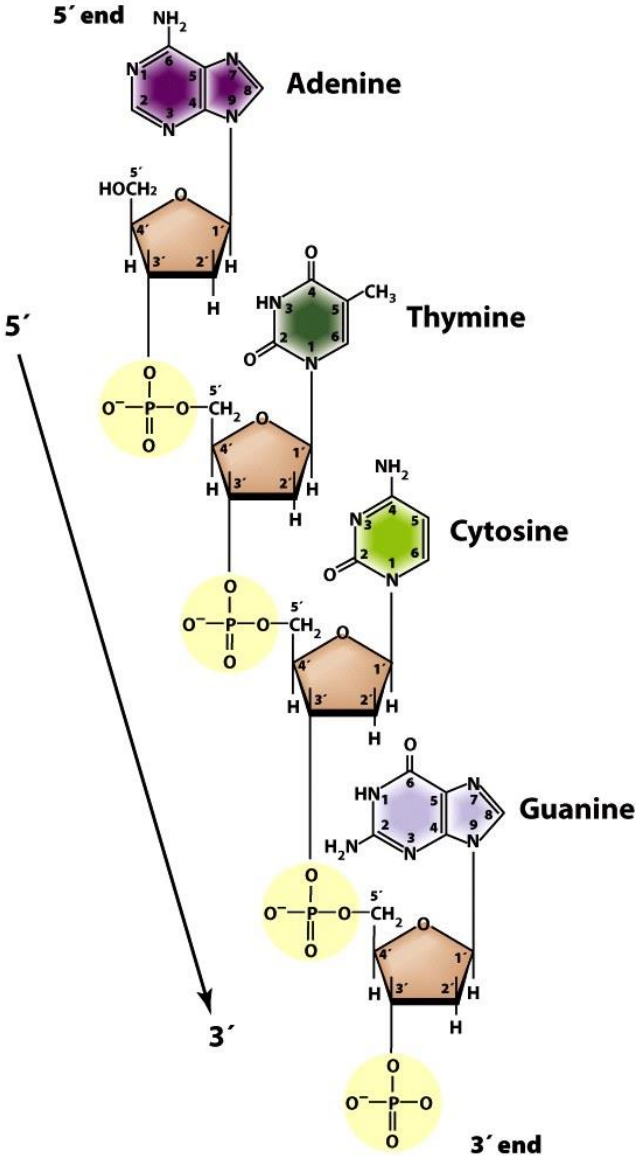


Figure 10-13
Genetics: A Conceptual Approach, Third Edition

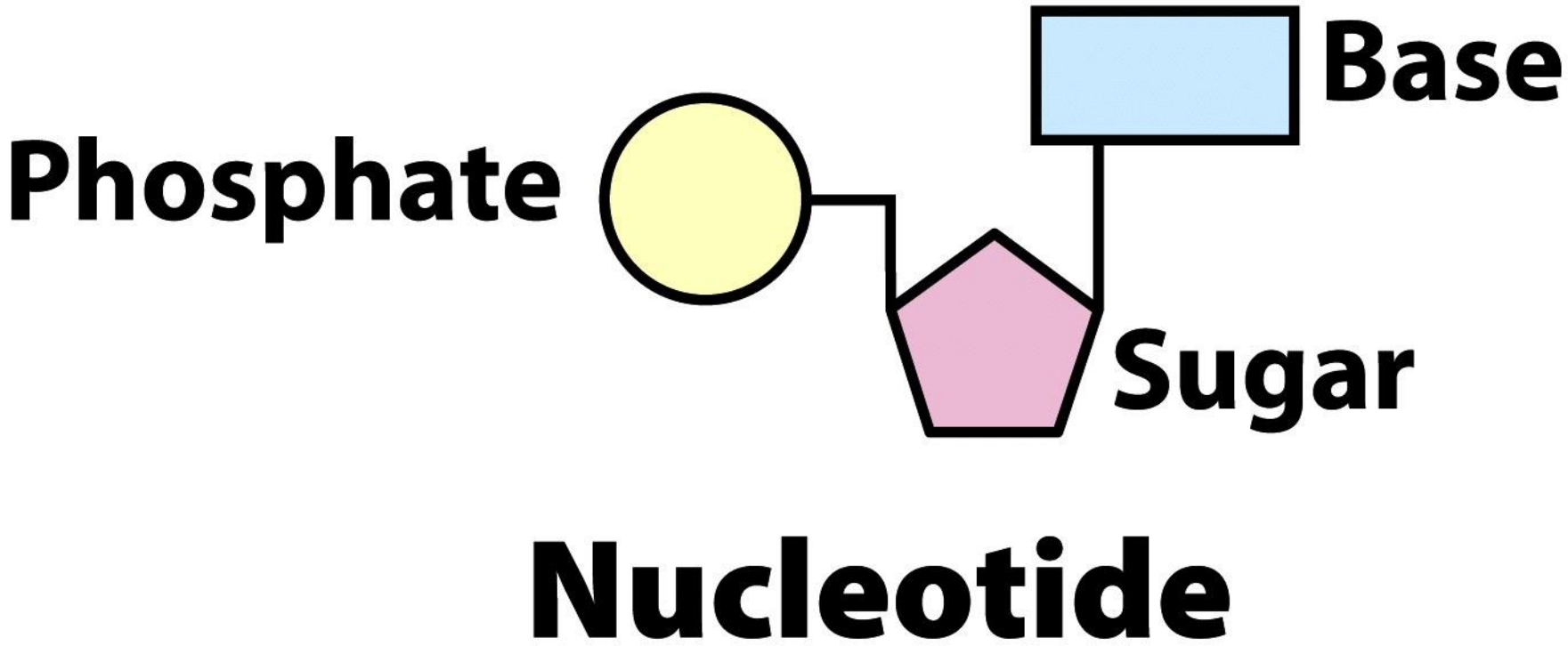
What is DNA?

- Deoxyribonucleic acid (DNA) is a polymer of nucleotides (‘building block’) and a macromolecule encoding the information of life



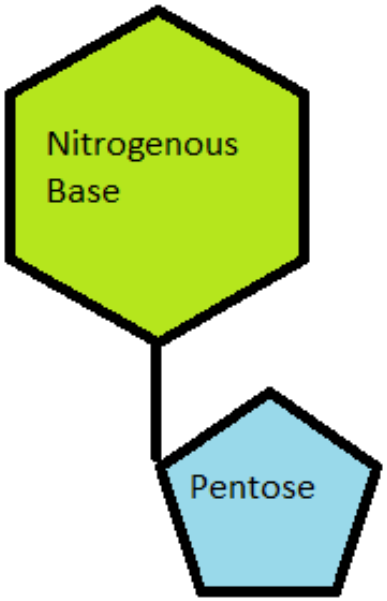
What is DNA?

- There are four different nucleotides based on the base structure: adenine (A), cytosine (C), guanine (G) and thymine (T).
- The bases can be separated into two different types: purines (A and G) and pyrimidines(C and T).
- What is dNTP?



Unnumbered 10 p269
Genetics: A Conceptual Approach, Third Edition

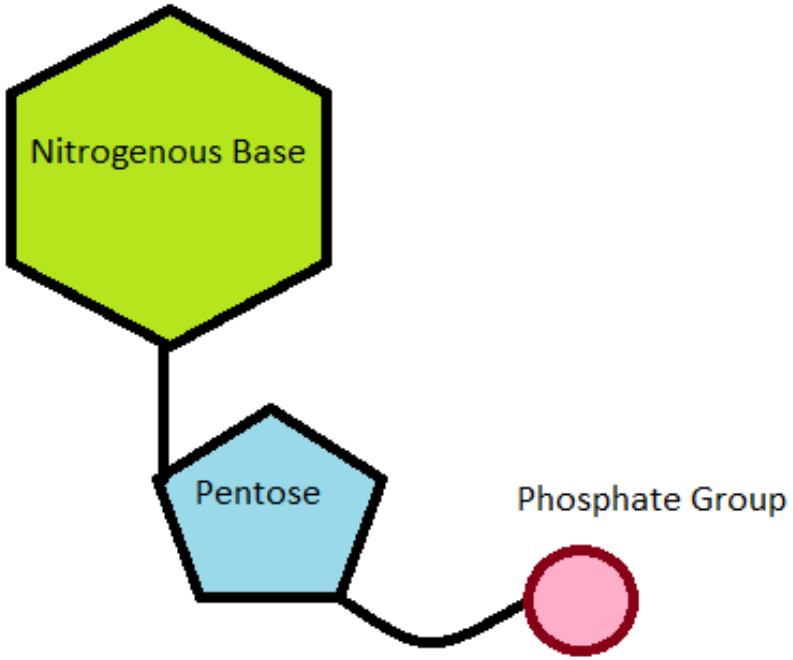
Nucleoside



=

Sugar + base

Nucleotide



=

Sugar + base + phosphate

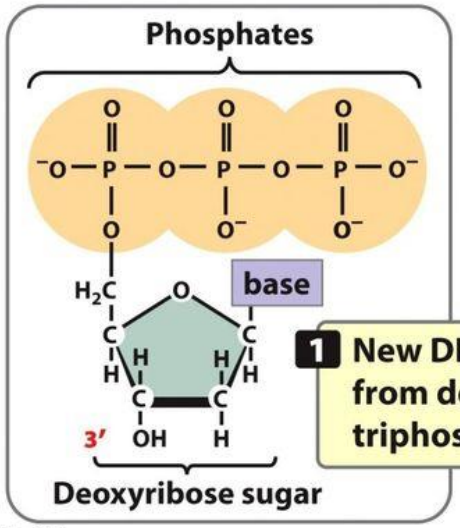


Figure 12.7a
Genetics: A Conceptual Approach, Fifth Edition
© 2014 W. H. Freeman and Company

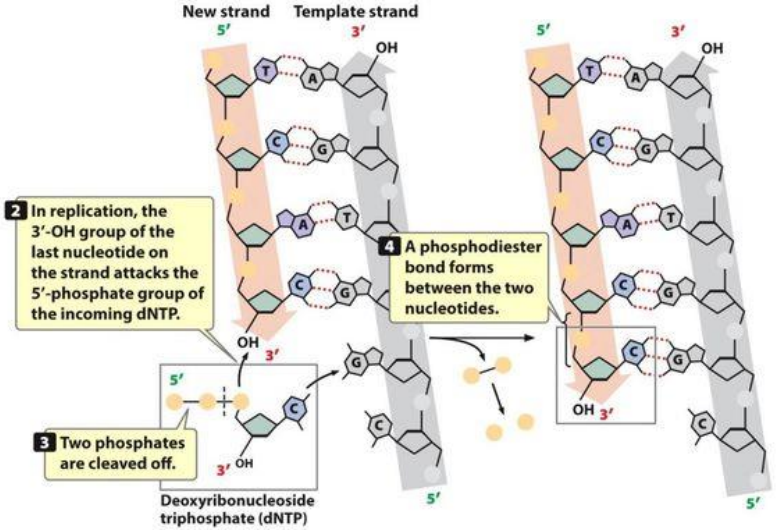
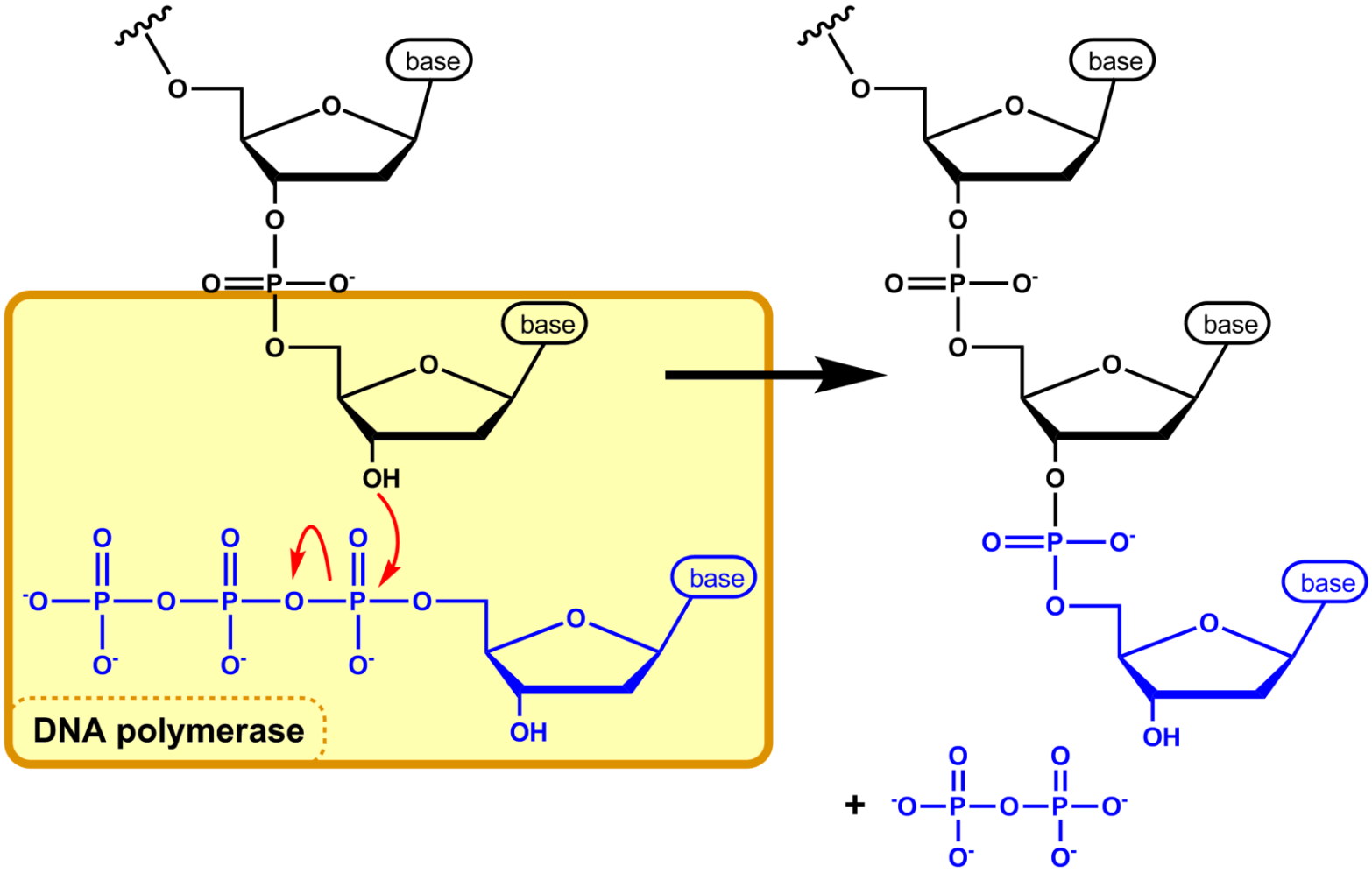
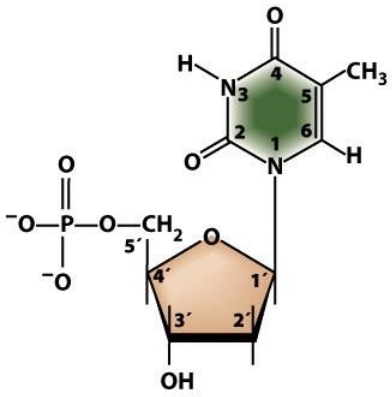


Figure 12.7b
Genetics: A Conceptual Approach, Fifth Edition
© 2014 W. H. Freeman and Company

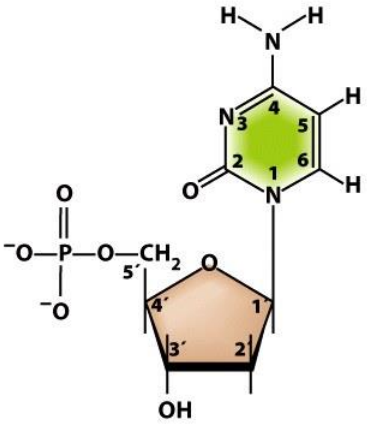


In DNA synthesis, the 3’ OH of a growing chain of nucleotides attacks the α-phosphate on the next NTP to be incorporated (blue), resulting in a phosphodiester linkage and the release of pyrophosphate (PPi).

Pyrimidine nucleotides

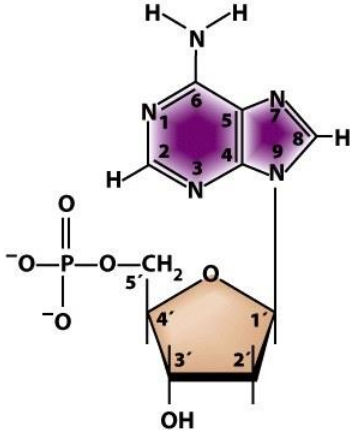


Deoxythymidine monophosphate, dTMP

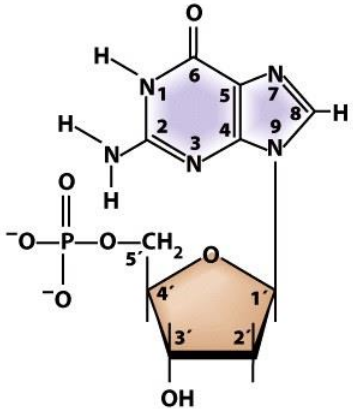


Deoxycytidine monophosphate, dCMP

Purine nucleotides

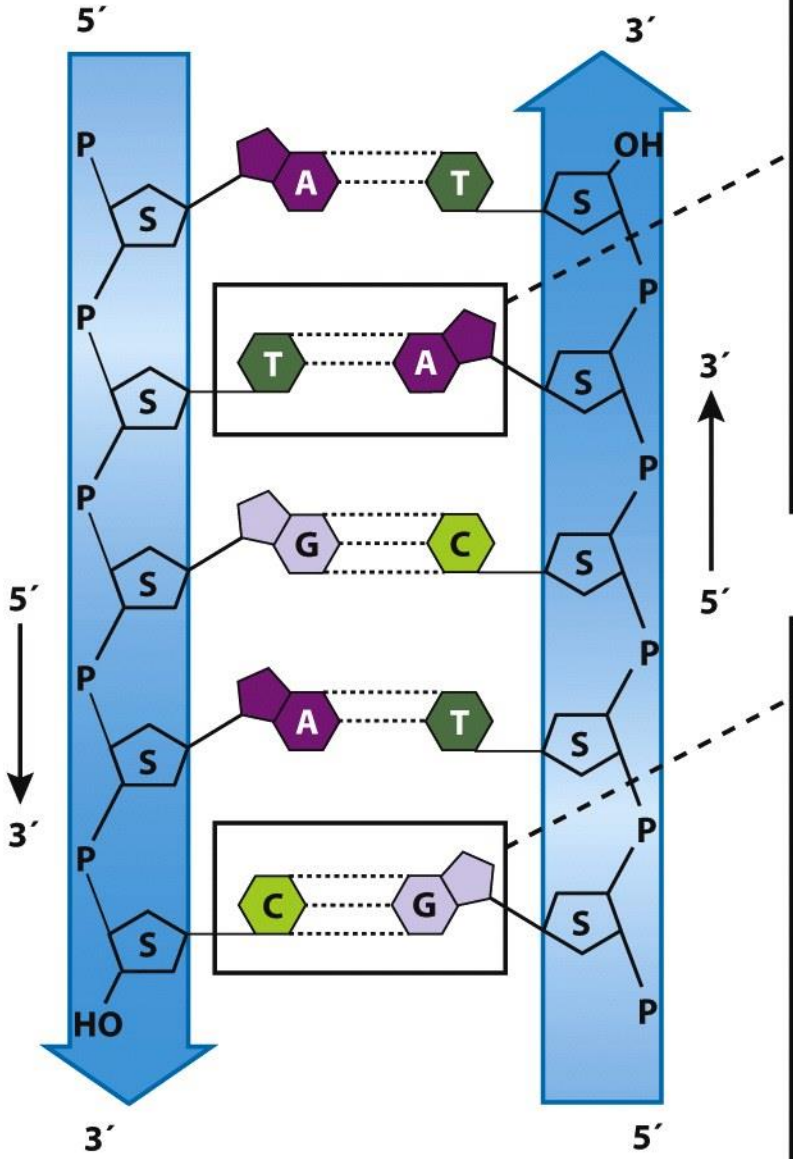


Deoxyadenosine monophosphate, dAMP

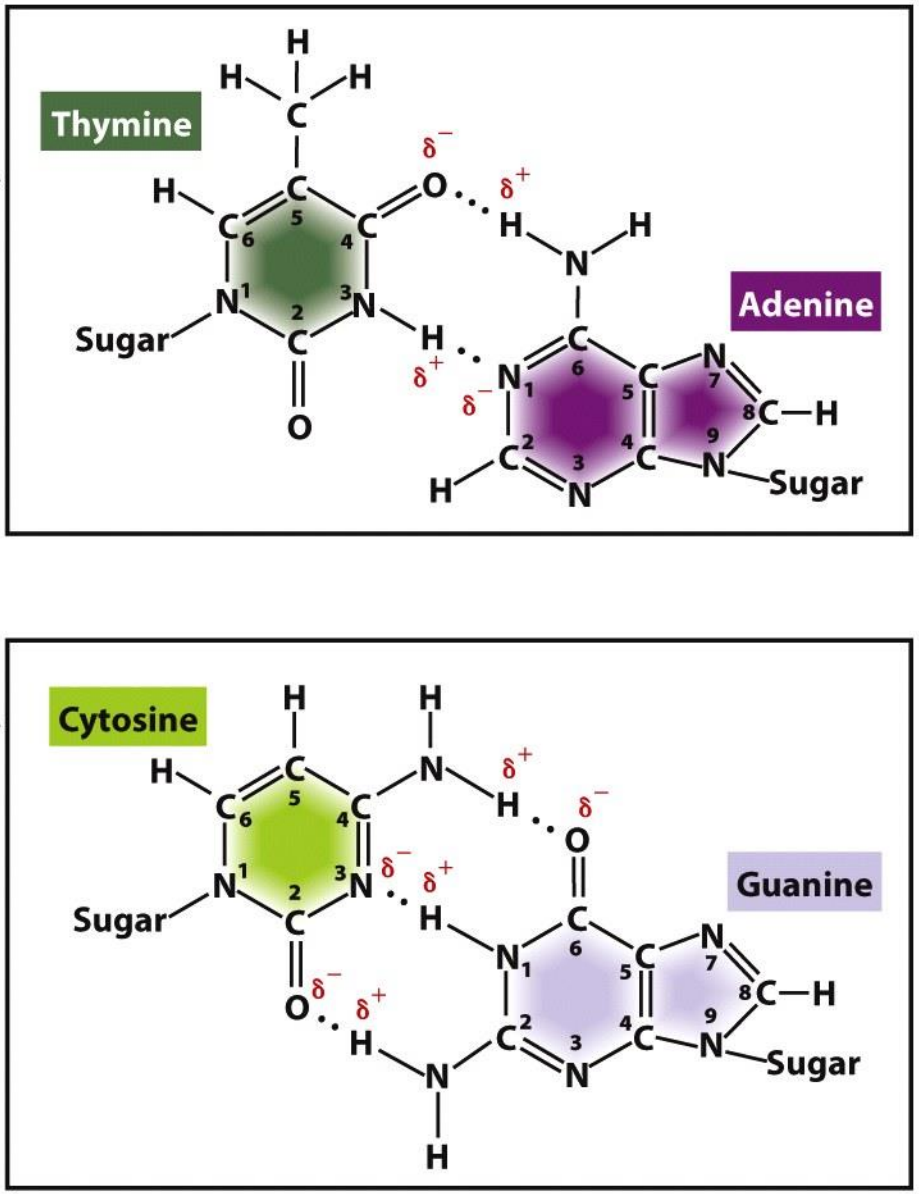


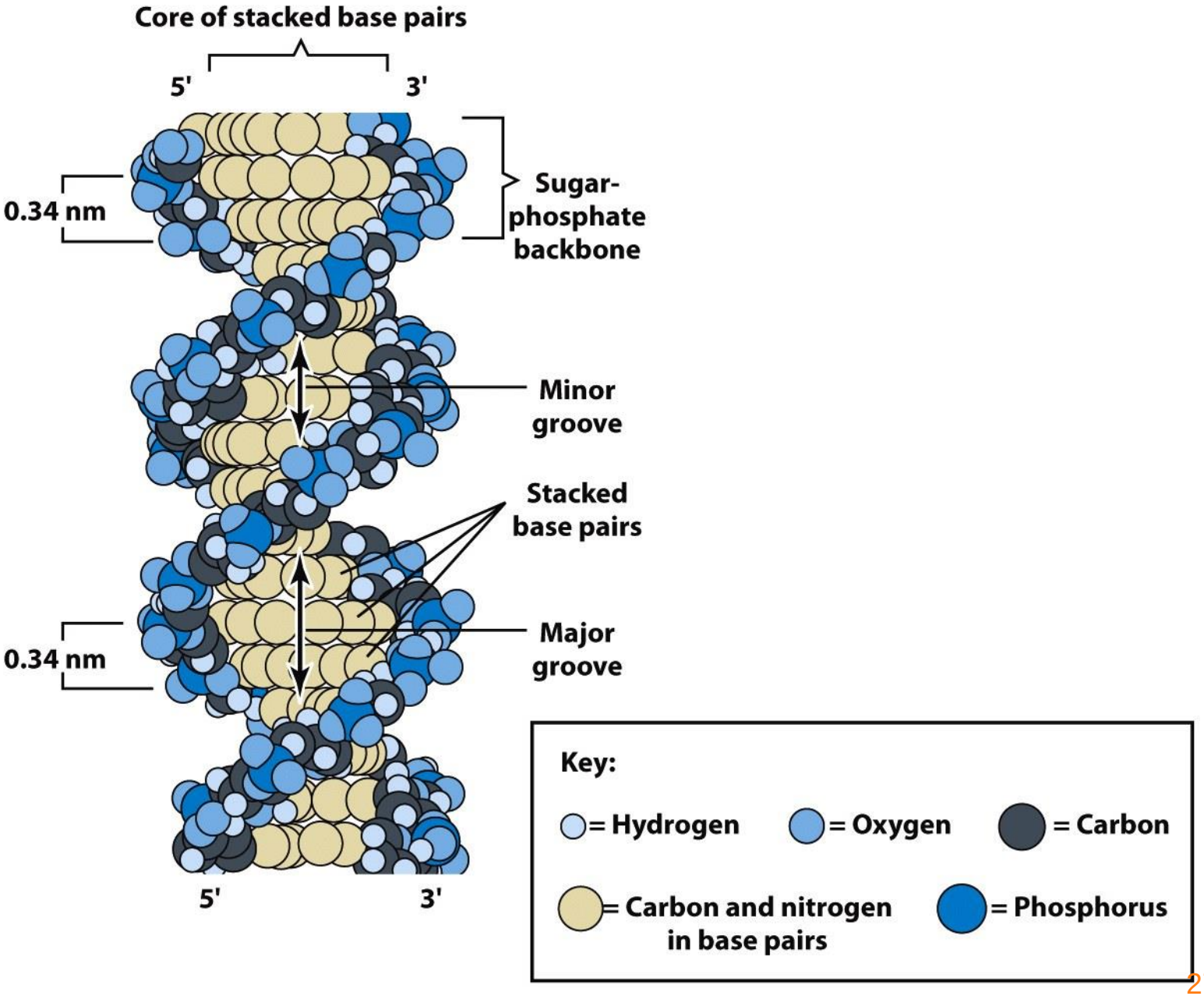
Deoxyguanosine monophosphate, dGMP

Opposite polarity of the two strands



Hydrogen bonding in A-T and G-C base pairs

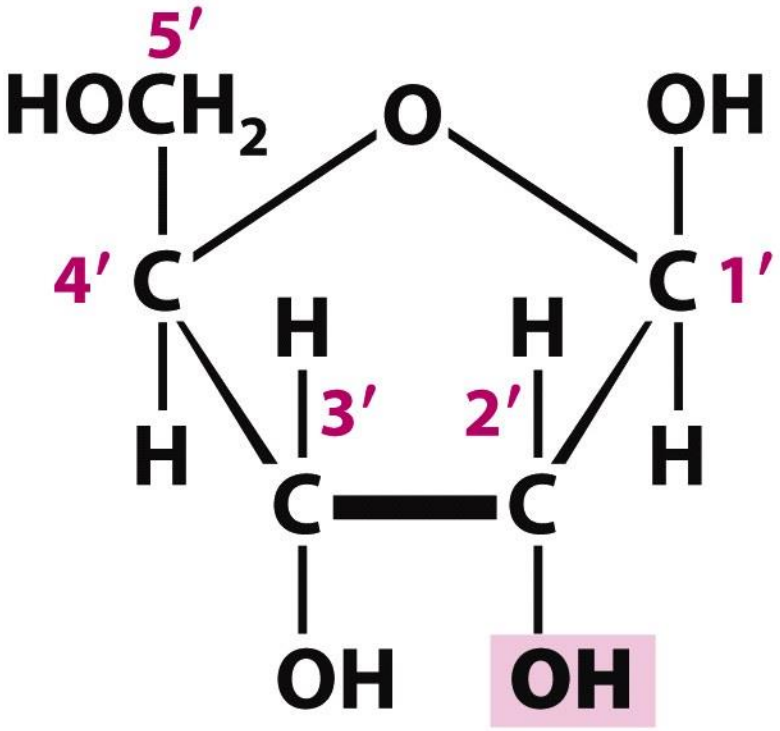



Core of stacked base pairs
5'
3'
0.34 nm
Sugar-phosphate backbone
Minor groove
Stacked base pairs
Major groove
0.34 nm
5'
3'
Key:
= Hydrogen
= Oxygen
= Carbon
= Carbon and nitrogen in base pairs
= Phosphorus

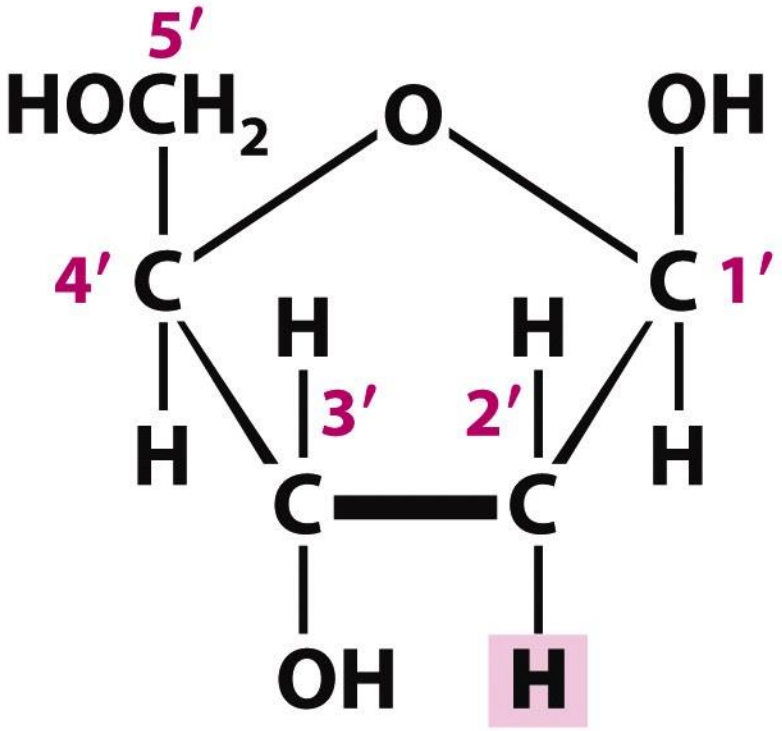
What is RNA?

- Ribonucleic acid (RNA) is a single stranded polymer that is composed of a long sequence of nucleotides.
- The sugar in RNA is ribose instead of deoxyribose in DNA
- RNA contains adenine (A), cytosine (C), guanine (G) and uracil (U) in place of thymine
- RNA is single stranded but can double back on itself via hydrogen bonds – RNA secondary structure

Why the published mRNA sequences in the nucleotide databases (NCBI) do not contain U but instead contain T?

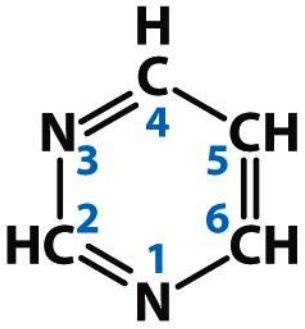


Ribose

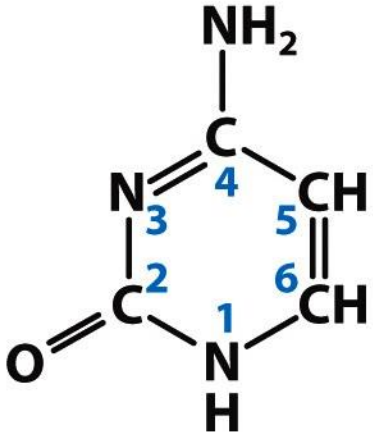


Deoxyribose

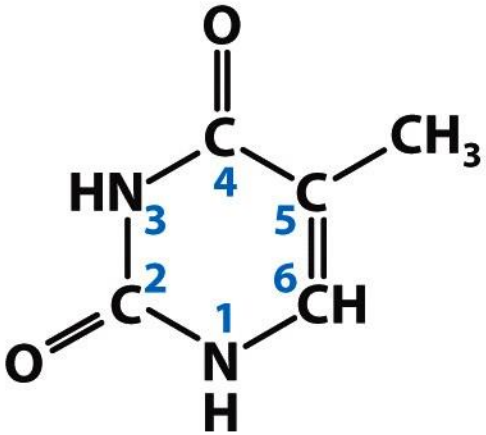
Figure 10-9
Genetics: A Conceptual Approach, Third Edition



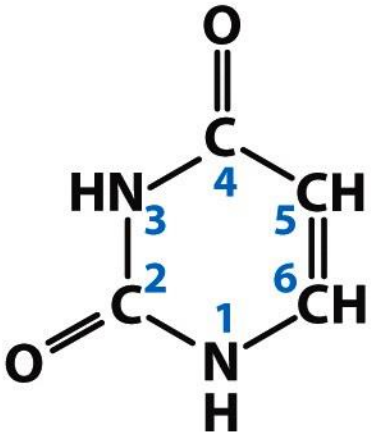
Pyrimidine
(basic structure)



Cytosine (C)



Thymine (T)
(present in DNA)



Uracil (U)
(present in RNA)

Figure 10-10 part 2
Genetics: A Conceptual Approach, Third Edition

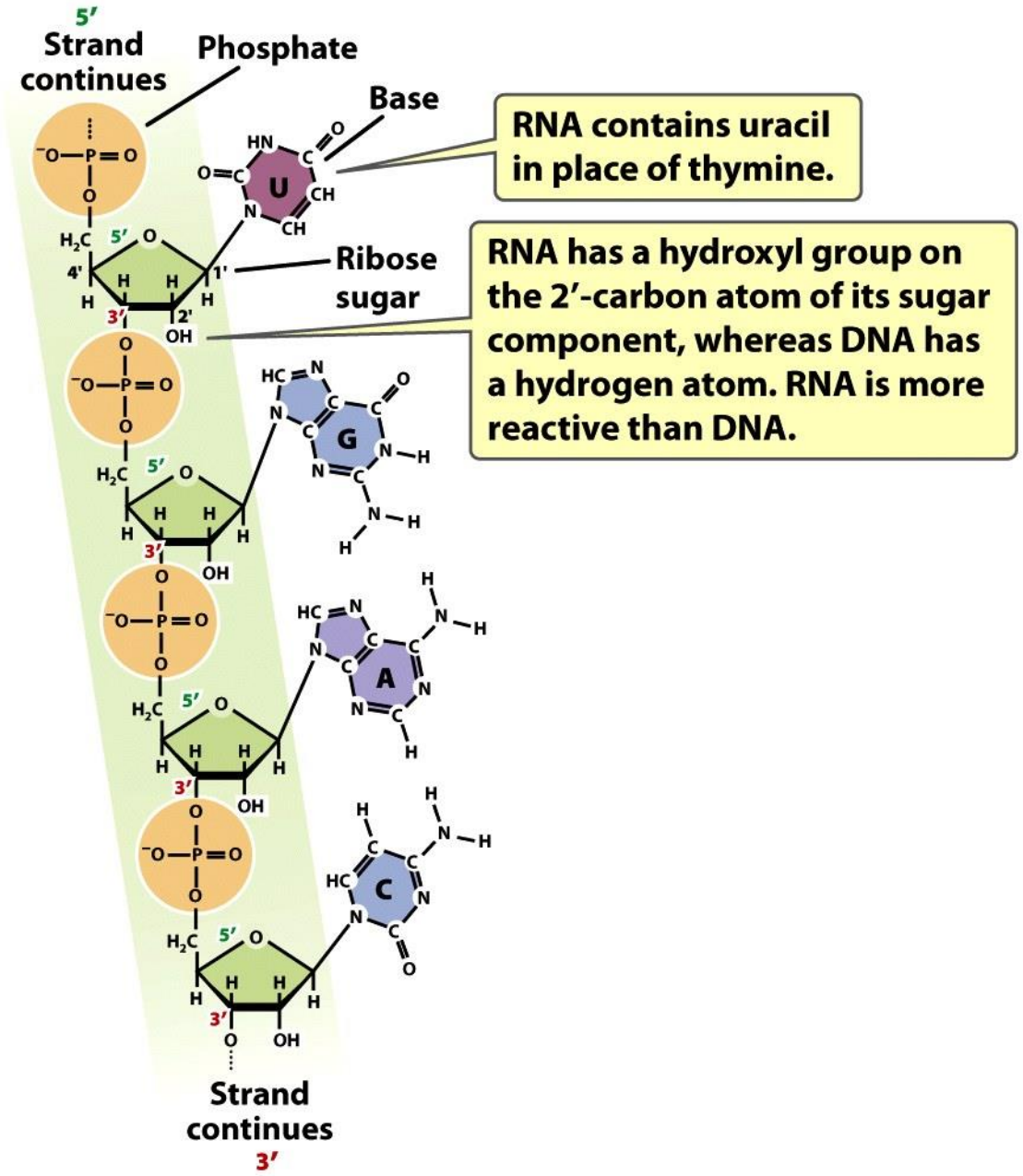


Figure 13-1a
Genetics: A Conceptual Approach, Third Edition

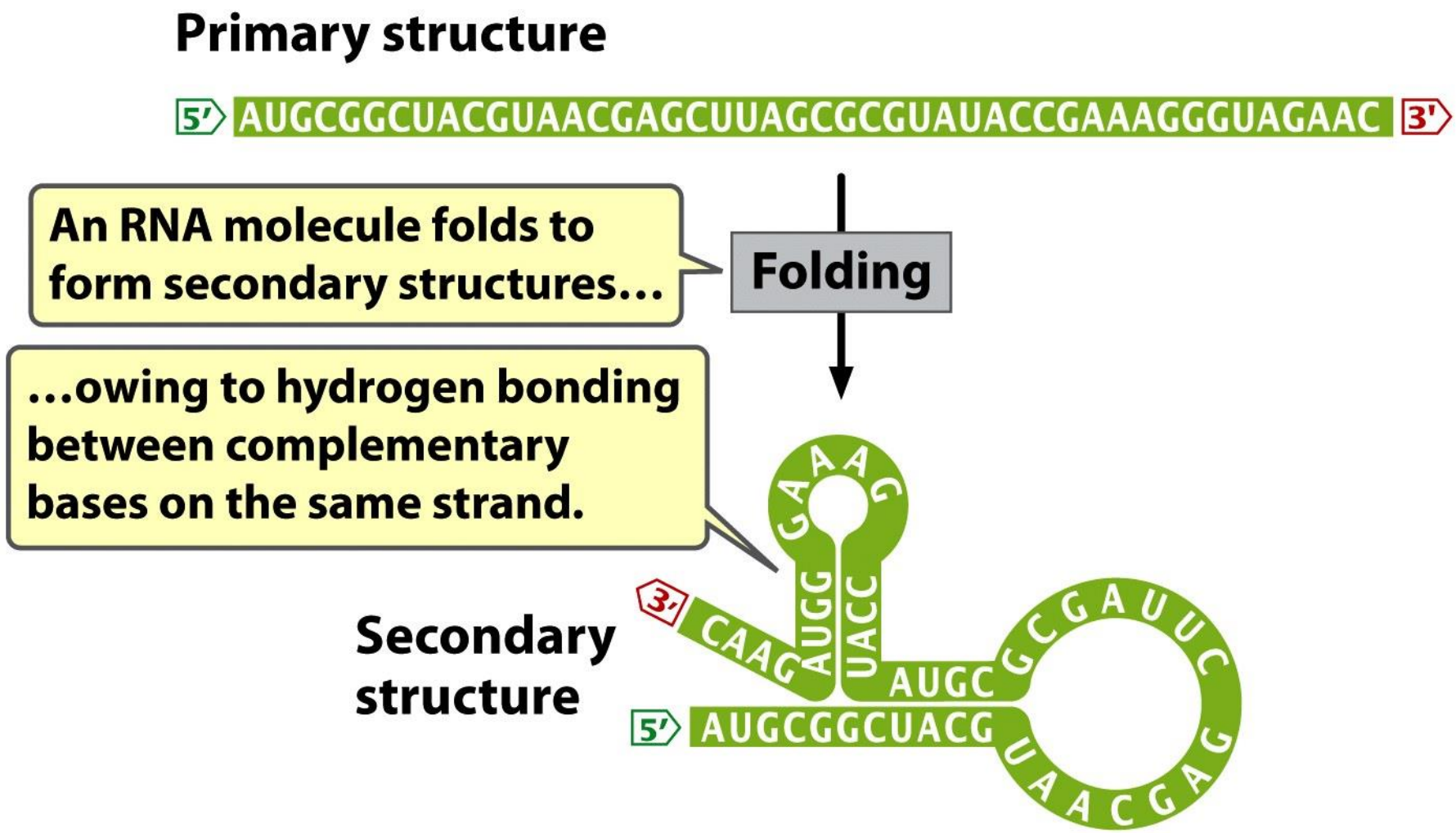


Figure 13-1b
Genetics: A Conceptual Approach, Third Edition

Proteins

- No proteins, no complex life
- Structure determines function
- Most drugs work by interacting with proteins
- Structure-function is vital to drug design
- Bioinformatics generates computer modelling for understanding structure-function interaction and speed up new drug design

Proteins

Proteins are polypeptides that have a three dimensional structure. They can be described through four different hierarchical levels:

- Primary structure: the amino acid sequence of a protein
- Secondary structure: the conformation of the polypeptide backbone
- Tertiary structure: its three-dimensional structure, that is further folding of the secondary structure in the three-dimensional space
- Quaternary structure: a structure achieved by proteins composed of more than one polypeptide chain

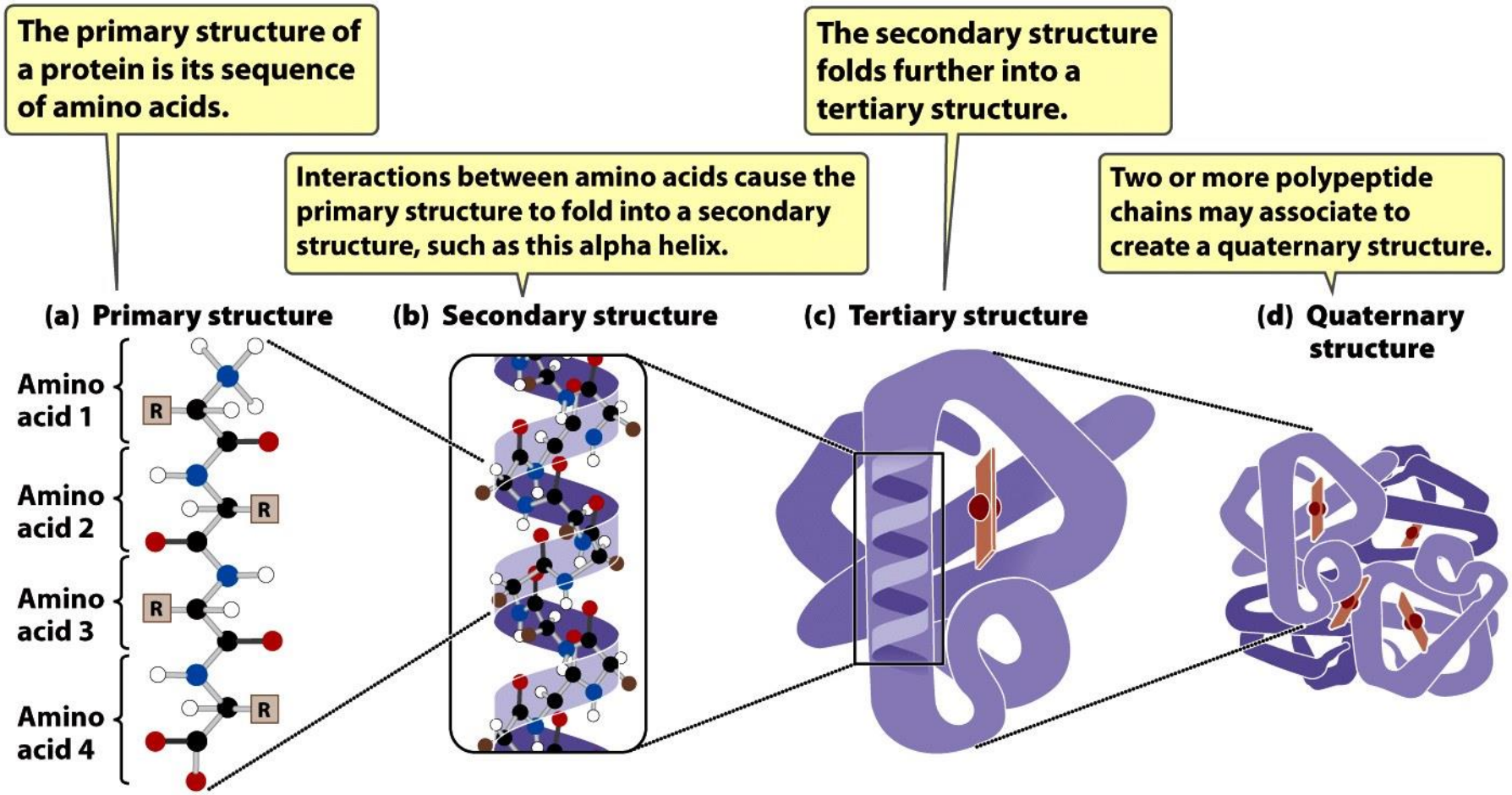
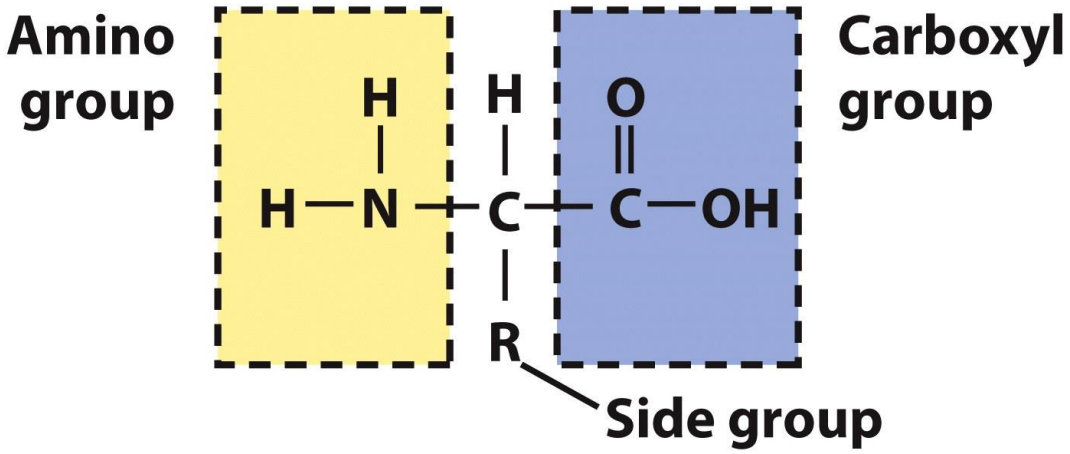


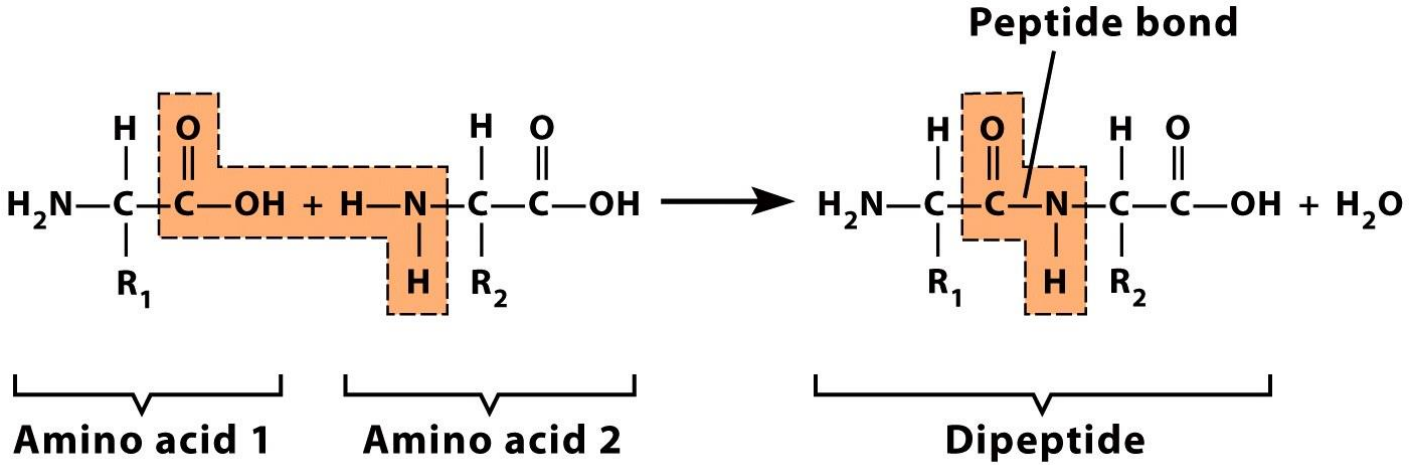
Figure 15-7
Genetics: A Conceptual Approach, Third Edition

Primary structure

- The primary structure is the protein sequence, the types & order of amino acids in the protein chain
- Amino acids are the building blocks of proteins
- Structure of an amino acid can be divided into a common main chain part & a side chain
  - Different chemical & physical properties of amino acids are due to their side chains
- Amino acids are linked by covalent peptide bonds



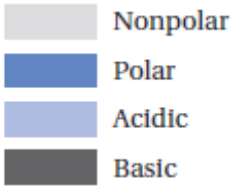
Amino acids have a free amino group, a free carboxyl group, and a side group (R)



Amino acids are joined by peptide bonds.

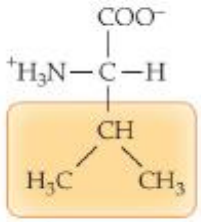
Amino acid	Three-letter code	One-letter code	Comment
Glycine	Gly	G	Only -H as side chain
Alanine	Ala	A	
Valine	Val	V	
Leucine	Leu	L	
Isoleucine	Ile	I	
Proline	Pro	P	Side chain to N bond
Phenylalanine	Phe	F	
Methionine	Met	M	
Tryptophan	Trp	W	
Cysteine	Cys	C	Forms disulfide bonds
Asparagine	Asn	N	Amide N polar
Glutamine	Gln	Q	Amide N polar
Serine	Ser	S	-OH group polar
Threonine	Thr	T	-OH group polar
Tyrosine	Tyr	Y	-OH group polar
Aspartic acid	Asp	D	
Glutamic acid	Glu	E	
Histidine	His	H	
Lysine	Lys	K	
Arginine	Arg	R	

20 amino acids colored according to their properties

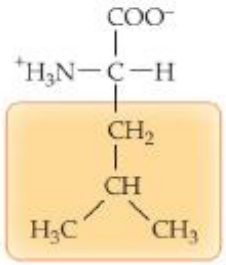


Classification on the basis of hydrophobicity

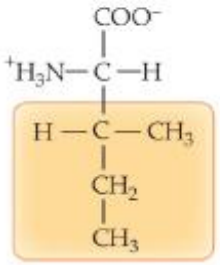
Amino acids with hydrophobic side groups



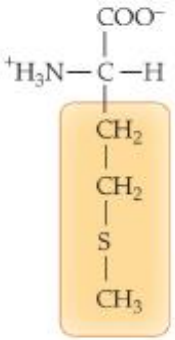
Valine
(Val) **V**



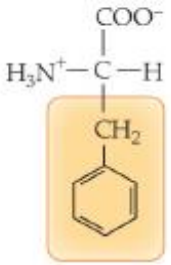
Leucine
(Leu) **L**



Isoleucine
(Ile) **I**



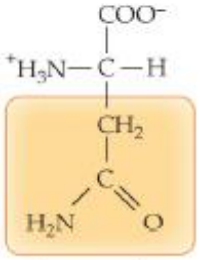
Methionine
(Met) **M**



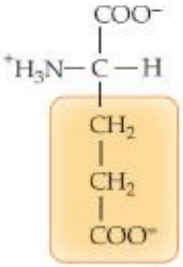
Phenylalanine
(Phe) **F**

Least able to dissolve in water

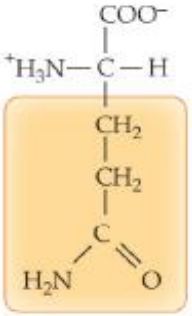
Amino acids with hydrophilic side groups



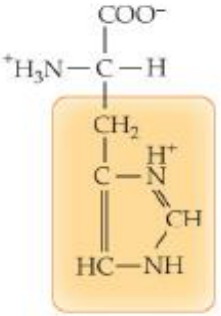
Asparagine
(Asn) **N**



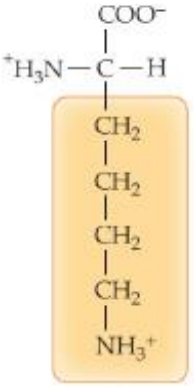
Glutamic acid
(Glu) **E**



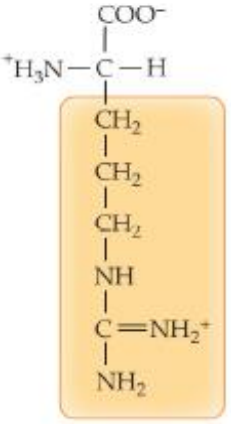
Glutamine
(Gln) **Q**



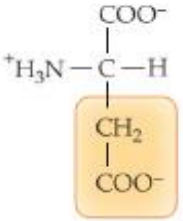
Histidine
(His) **H**



Lysine
(Lys) **K**



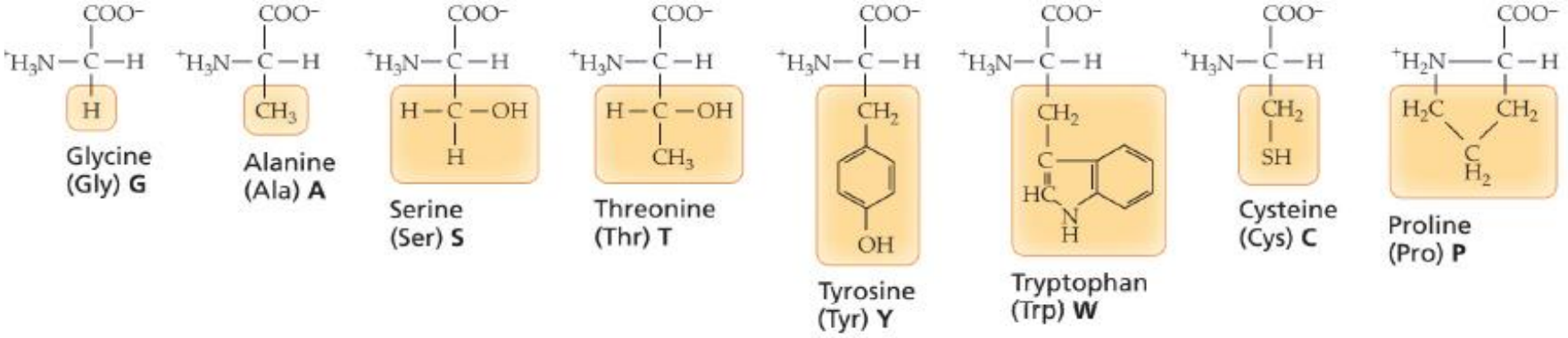
Arginine
(Arg) **R**



Aspartic acid
(Asp) **D**

Readily dissolve in water

Amino acids with side groups that are in between



Dissolve in water only slightly

Primary structure – Ionic character of amino acids

- In solution at physiological pH (7.4), amino acids exist as dipole ions or zwitterions
  - the amino group (NH2) exists as an ammonium ion (NH_3^+)
  - the carboxyl group (COOH) exists as a carboxylate ion (COO^-)
- At acidic pH < 7.4:
  - the amino group has a positive charge
  - the carboxyl is neutral
- At alkaline pH > 7.4:
  - the amino group is neutral
  - the carboxyl has a negative charge

Primary structure – Ionic character of amino acids

- Amino acids of proteins in solution accept or lose protons depending on the nature of the side chains
- The pK_a values of amino acids (i.e. the tendency of amino acids to lose protons) play an important role in determining the pH-dependent properties of a protein in solution
- Internal ionizable groups in proteins are essential for catalysis.
- During a cycle of function, these internal ionizable groups can experience different microenvironments, and their pKa values and charged states adjust accordingly.

Relationship between protein function & the location of amino acids in the polypeptide chain

- The location of amino acids in the folded conformation of a protein is relevant for the protein's function and its interaction with the environment
  - Proteins located in a hydrophobic environment, such as membrane, have nonpolar (hydrophobic) side chains on the surface interacting with the membrane lipids
  - Proteins located in an aqueous environment such as cytosol, have polar side chains (hydrophilic) on the surface interacting with the aqueous environment

Relationship between protein function & the location of amino acids in the polypeptide chain

- Arginine and lysine carry positive charges, and are often located on the interacting surface of proteins that interact with negatively charged molecules
  - Arginine and lysine are found on the surface of DNA-binding proteins that interact with the negatively charged phosphate group of DNA
- Serine, threonine, and tyrosine have hydroxyl groups (-OH) in their side chains. These OH groups can serve as phosphate attachment sites during phosphorylation

Relationship between protein function & the location of amino acids in the polypeptide chain

- The sulfhydryl (-SH) group in cysteine is ideal for binding metals through metal-thiolate bonds
  - In metallothionein, the intracellular metal-binding protein, one third of the amino acid residues are cysteines
- The -SH group is also ideal for forming strong covalent disulfide linkages that stabilize the conformation of proteins
  - Cysteines are found in many enzymes that function in harsh conditions of salt and pH, such as digestive enzymes like pepsin and chymotrypsin
  - The structure of many small proteins, such as insulin and ribonuclease, is stabilized by cysteine disulfide linkage

Secondary structure

- The secondary structure of protein sequence is made up of α-helices and β-strands
- Hydrogen bonding occurs within an α-helices and between different β-strands. A set of β-strands hydrogen bonded together side by side forms a β-sheet
- Most proteins have both α-helices and β-strands, but some have only α-helices or β-strands
- Functionally important regions are not the regular structures but the accessible regions, eg. the surface

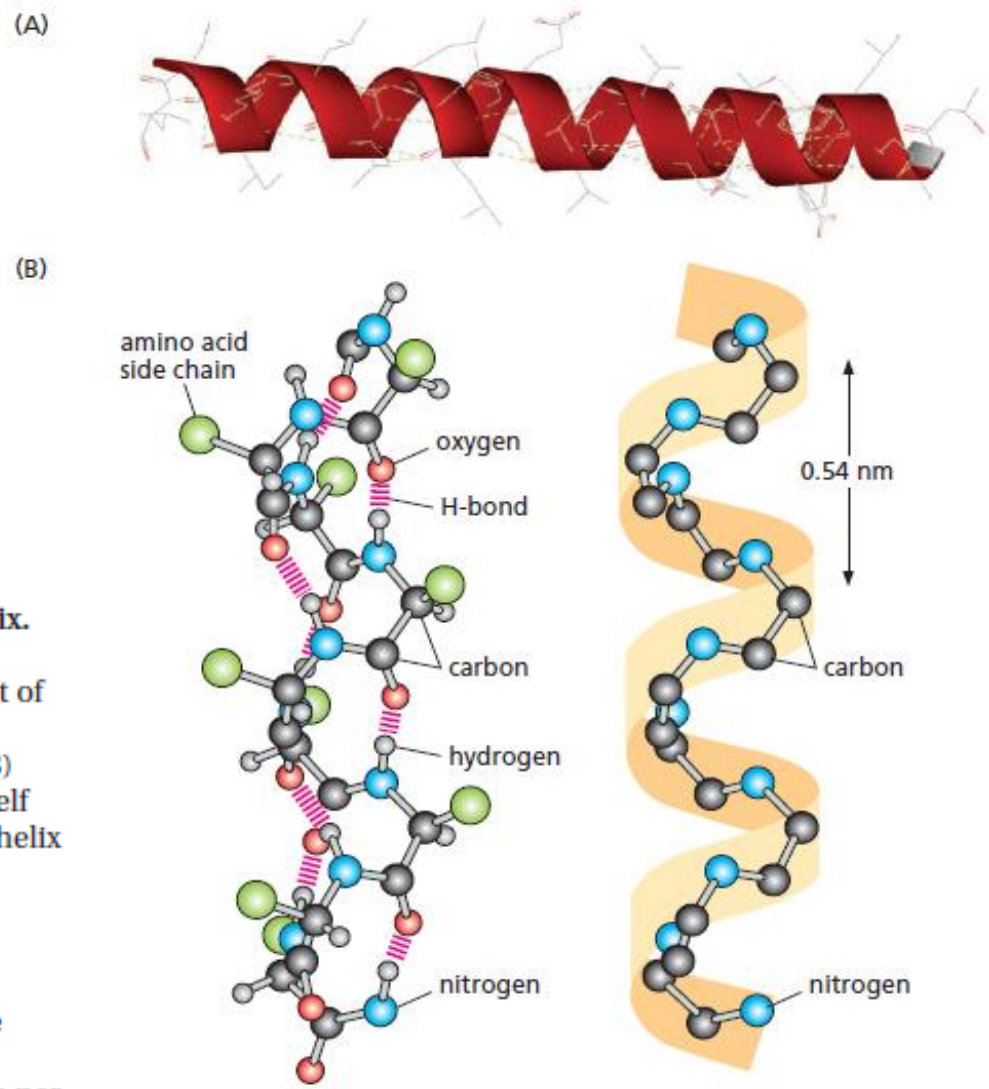
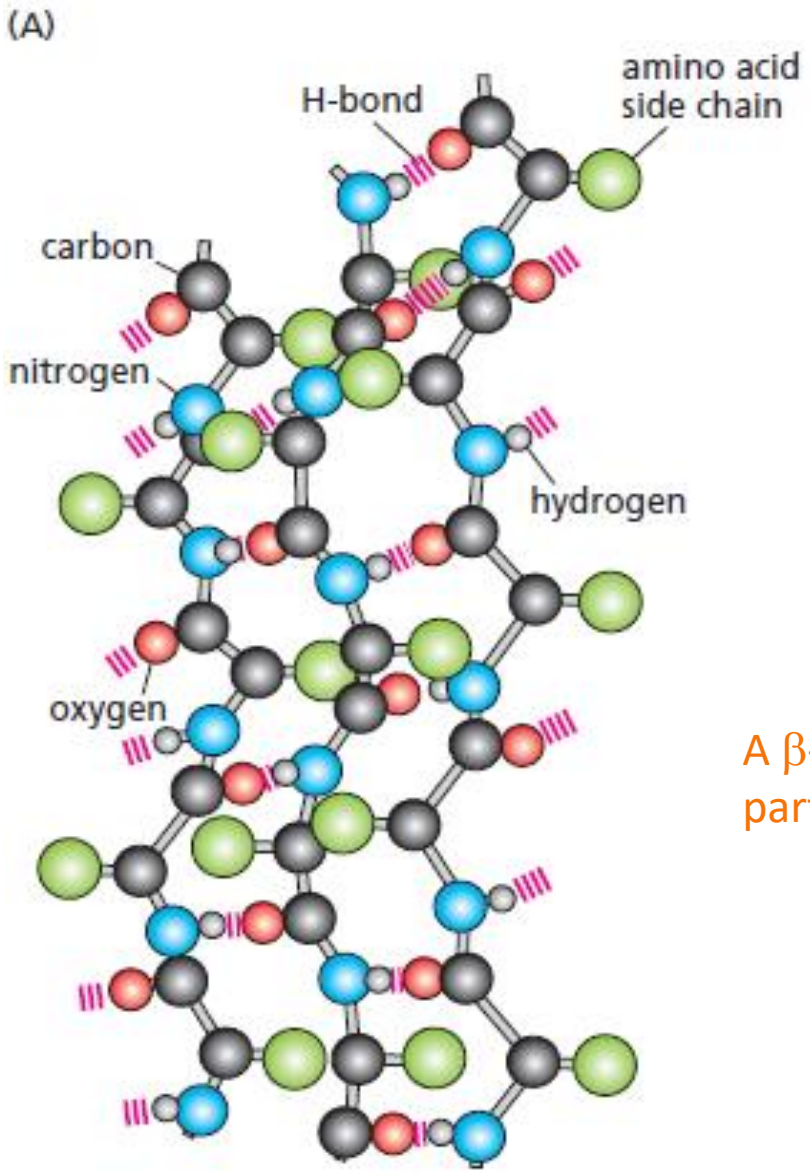


Figure 2.9
Hydrogen bonding in the α-helix. In the α-helix all hydrogen (H-) bonds involve the same element of secondary structure. (A) A representation of the α-helix. (B) The helical structure repeats itself every 0.54 nm (5.4 Å) along the helix axis, therefore we say that the α-helix has a pitch of 0.54 nm. α-Helices have 3.6 amino acid residues per helical turn. The separation of residues along the helix axis is 0.54/3.6 or 0.15 nm (1.5 Å); i.e., the α-helix has a rise per residue of 0.15 nm (1.5 Å).



A β-strand has 2 sides that can participate in hydrogen bonding

Tertiary structure

- Proteins are functional only when folding into tertiary structures
- Combination of secondary structure pack together to form tertiary structures
- Domain: a discrete tertiary structural unit (50-350 aa) with a particular biochemical or binding function
- The core of each domain has α-helices or β-strands or both tightly packed
- Many proteins have multiple domains
- Proteins that have different sequences can form similar tertiary structure
- Bioinformatics analysis are domain focused

Stability of Tertiary structure

- A protein chain starts to fold into tertiary structure as soon as it is synthesized; the final fold is a state of low free energy
- Most proteins start to unfold at 60°C when non-covalent bonds are broken; unfolded proteins are denatured and not functional
- Stability is a balance of multiple effects
  - Hydrophobicity of side chains
    - Hydrophobic interactions are more stable away from water contact
    - Hydrophobic groups tend to cluster inside proteins
    - Polar groups tend to be on the protein surface
  - Hydrogen bonding
  - Compactness: proteins with unoccupied interior space are less stable

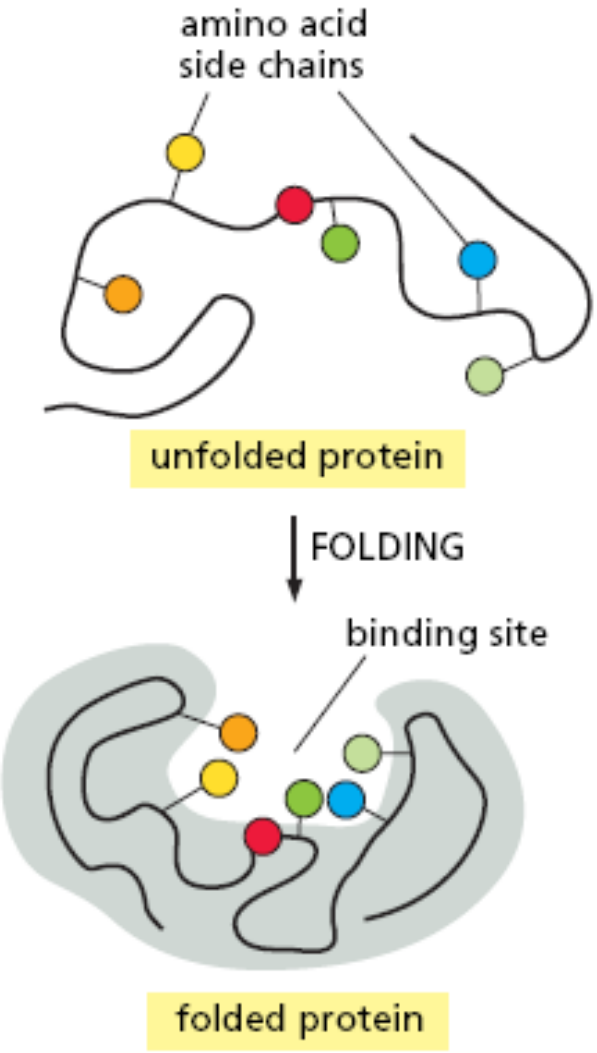
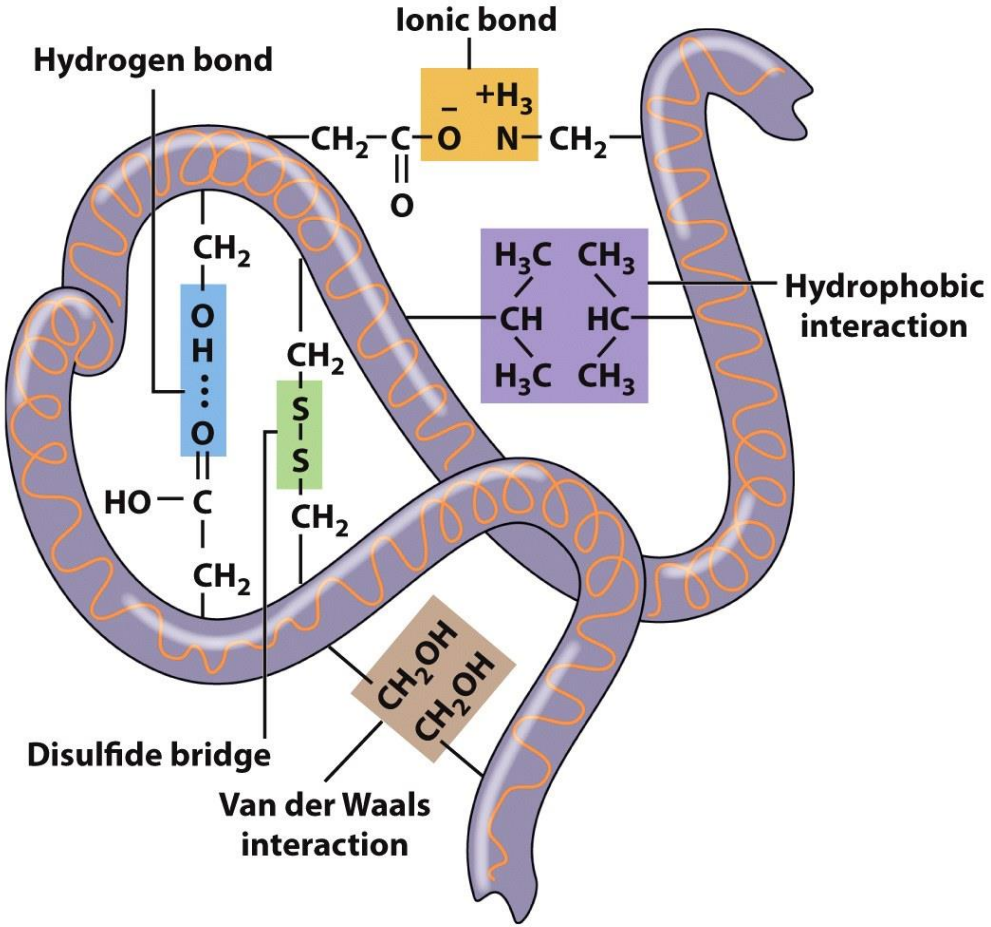


Figure 2.14
Distant residues can come close in the folded structure. A schematic which shows that when a polypeptide chain (primary structure) folds into a tertiary structure, residues that are far apart from each other in the sequence can come close together to form a functional unit, in this case a binding/catalytic site. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)



Molecular interactions determining tertiary structure

Quaternary structure

- Individually folded polypeptides (subunit) interact with each other to form a protein complex
- Formed by noncovalent interactions when the complex is active
- Hard to model or predict
- Many proteins contain multiple subunits
  - Oligomer = protein complex, monomer = subunit
  - Dimer – 2 subunits, trimer – 3 subunits, tetramer – 4 subunits
  - Homo – same subunit, hetero – different subunits.

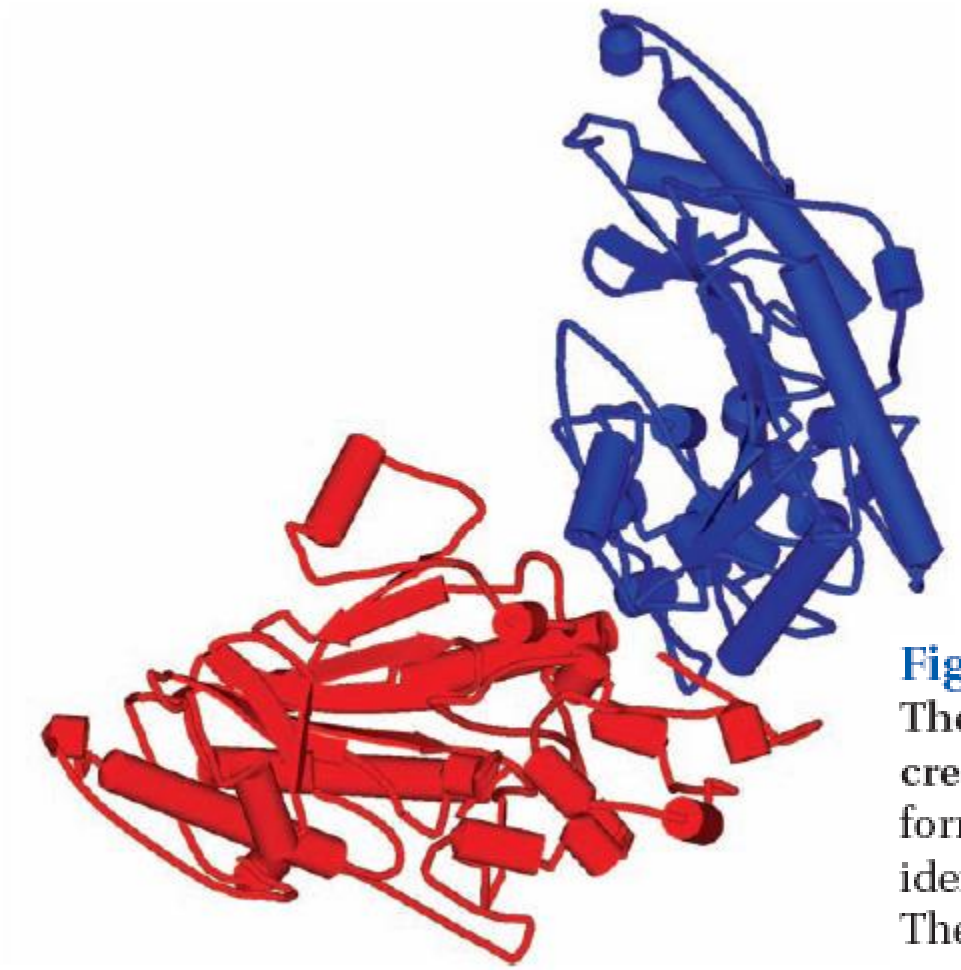
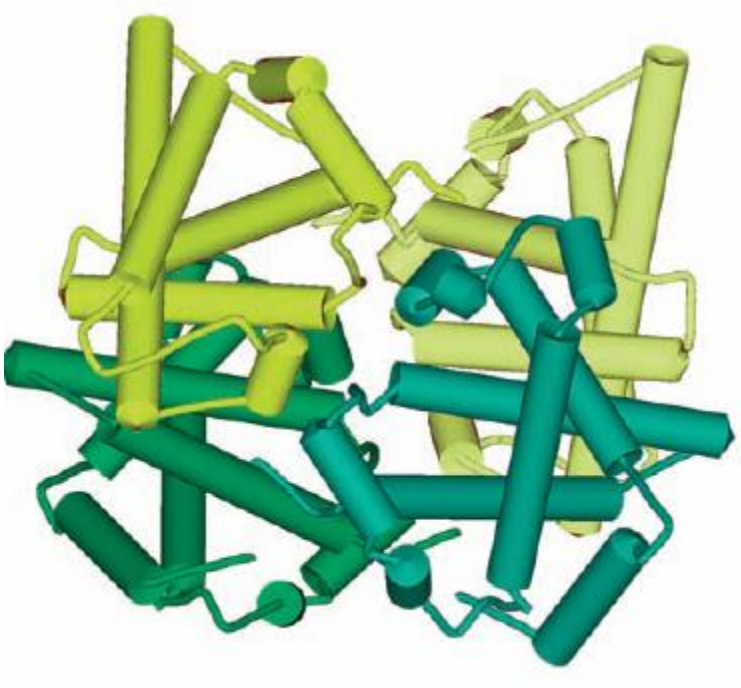


Figure 2.15
The quaternary structure of muscle creatine kinase. This structure is formed by the complex of two identical chains and structures. Therefore it is called a homodimer.



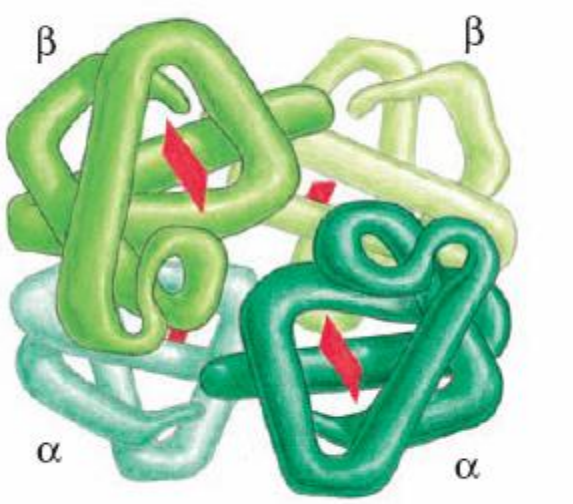
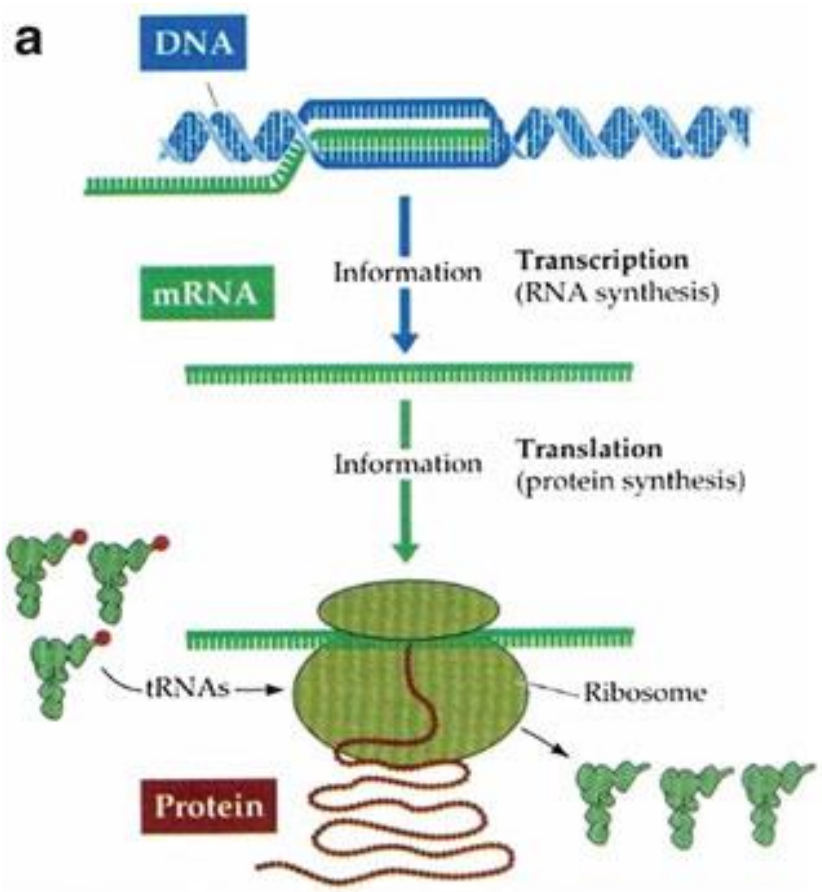
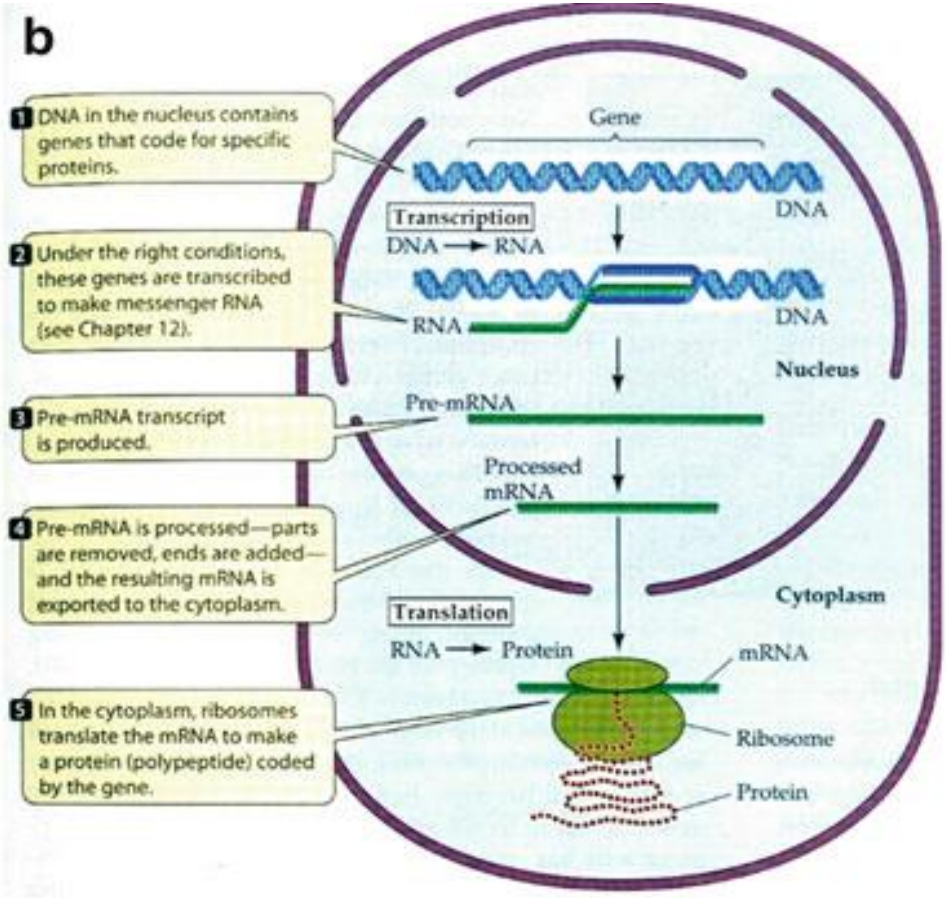


Figure 2.16
The quaternary structure of bovine deoxyhemoglobin which is a heterotetramer. In other words it is made up of different chains and structures from four folds. Two folds are the same therefore it consists of two homodimers. The heme groups are shown in red. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

The central dogma

- In 1958, Francis Crick introduced the phrase "central dogma" to represent the flow of genetic information
- DNA encodes all RNAs and proteins necessary for life
- Gene: a segment of DNA for an RNA or protein
- Transcription: DNA → RNA
- Translation: RNA → protein using genetic codes


a
DNA
Information
Transcription
(RNA synthesis)
mRNA
Information
Translation
(protein synthesis)
tRNAs
Ribosome
Protein


b
1 DNA in the nucleus contains genes that code for specific proteins.
2 Under the right conditions, these genes are transcribed to make messenger RNA (see Chapter 12).
3 Pre-mRNA transcript is produced.
4 Pre-mRNA is processed—parts are removed, ends are added—and the resulting mRNA is exported to the cytoplasm.
5 In the cytoplasm, ribosomes translate the mRNA to make a protein (polypeptide) coded by the gene.
Gene
DNA
Transcription
DNA → RNA
RNA
DNA
Nucleus
Pre-mRNA
Processed mRNA
Cytoplasm
Translation
RNA → Protein
mRNA
Ribosome
Protein

The genetic codes

- Initiation and termination codons
  - Initiation codon: AUG
  - Termination codons: UAA, UAG, UGA
- Degeneracy: partial and complete
- Ordered
- Nearly universal (exceptions: mitochondria and some protozoa)

▶ TABLE 12.1

The Genetic Code[a]

First (5′) letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third (3′) letter
U	UUU Phe (F)	UCU Ser (S)	UAU Tyr (Y)	UGU Cys (C)	U
	UUC Phe (F)	UCC Ser (S)	UAC Tyr (Y)	UGC Cys (C)	C
	UUA Leu (L)	UCA Ser (S)	UAA Stop (terminator)	UGA Stop (terminator)	A
	UUG Leu (L)	UCG Ser (S)	UAG Stop (terminator)	UGG Trp (W)	G
C	CUU Leu (L)	CCU Pro (P)	CAU His (H)	CGU Arg (R)	U
	CUC Leu (L)	CCC Pro (P)	CAC His (H)	CGC Arg (R)	C
	CUA Leu (L)	CCA Pro (P)	CAA Gln (Q)	CGA Arg (R)	A
	CUG Leu (L)	CCG Pro (P)	CAG Gln (Q)	CGG Arg (R)	G
A	AUU Ileu (I)	ACU Thr (T)	AAU Asn (N)	AGU Ser (S)	U
	AUC Ileu (I)	ACC Thr (T)	AAC Asn (N)	AGC Ser (S)	C
	AUA Ileu (I)	ACA Thr (T)	AAA Lys (K)	AGA Arg (R)	A
	AUG *Met* (M) (initiator)	ACG Thr (T)	AAG Lys (K)	AGG Arg (R)	G
G	GUU Val (V)	GCU Ala (A)	GAU Asp (D)	GGU Gly (G)	U
	GUC Val (V)	GCC Ala (A)	GAC Asp (D)	GGC Gly (G)	C
	GUA Val (V)	GCA Ala (A)	GAA Glu (E)	GGA Gly (G)	A
	GUG Val (V)	GCG Ala (A)	GAG Glu (E)	GGG Gly (G)	G

= Polypeptide chain initiation codon

= Polypeptide chain termination codon

[a]Each triplet nucleotide sequence or codon refers to the nucleotide sequence in mRNA (not DNA) that specifies the incorporation of the indicated amino acid or polypeptide chain termination. The one-letter symbols for the amino acids are given in parentheses after the standard three-letter abbreviations.

The central dogma

- Only messenger RNAs are translated into proteins
- Often one gene for one mRNA and one protein but many exceptions
- When a gene is transcribed into an RNA, it is "expressed" – gene expression
- The template DNA strand for translation is the noncoding strand, or anticoding/antisense strand
- Non-template DNA strand is the coding or sense strand

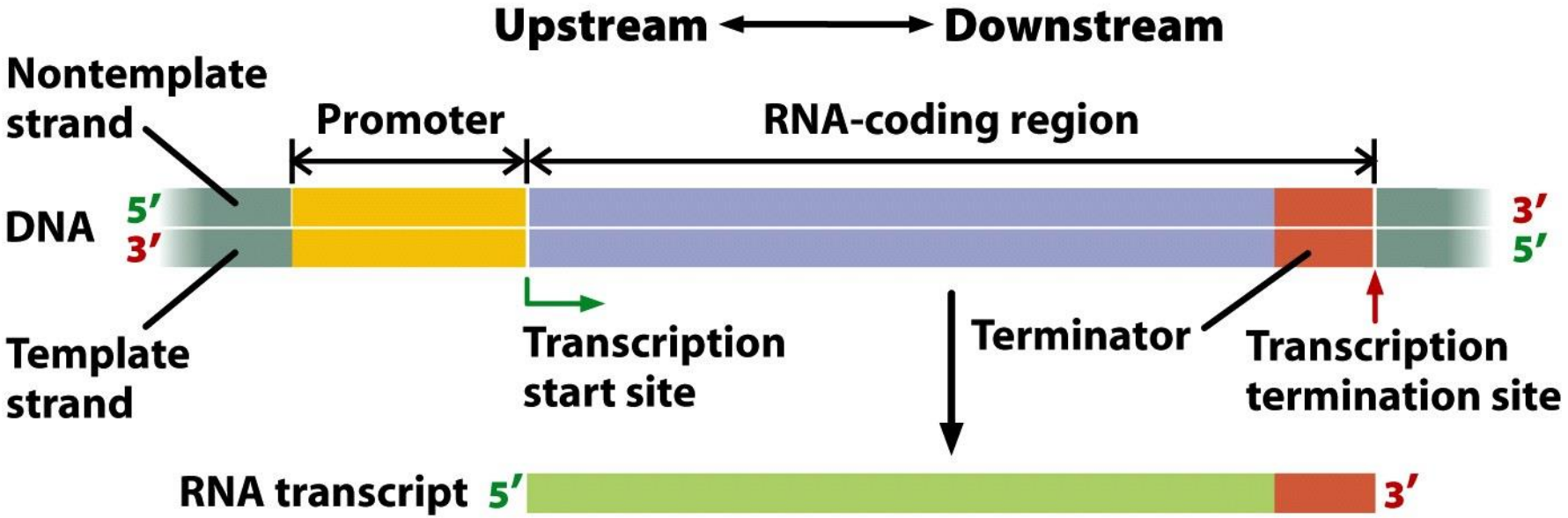
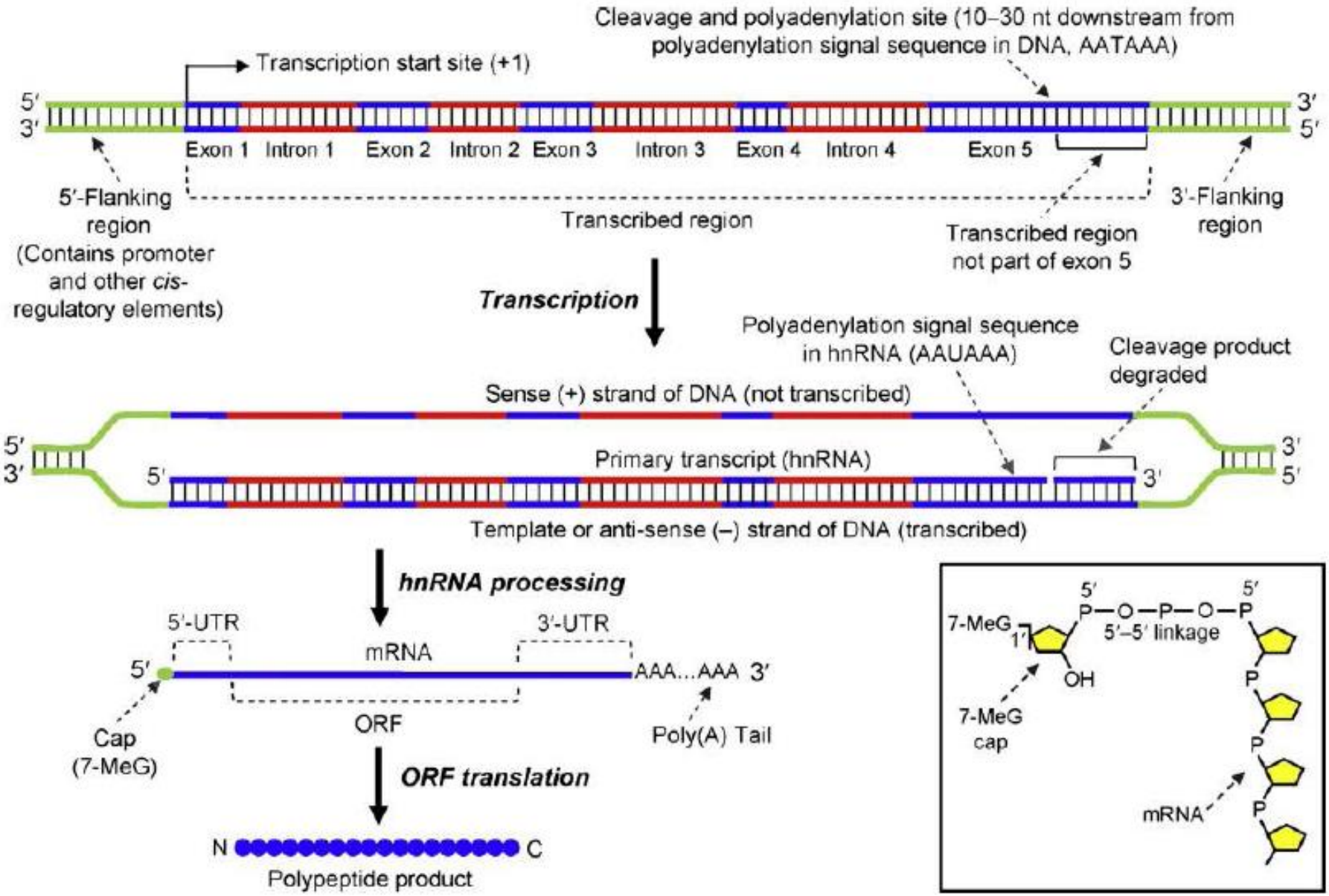


Figure 13-7
Genetics: A Conceptual Approach, Third Edition


Cleavage and polyadenylation site (10–30 nt downstream from polyadenylation signal sequence in DNA, AATAAA)
Transcription start site (+1)
5′
3′
3′
5′
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Intron 3
Exon 4
Intron 4
Exon 5
5′-Flanking region (Contains promoter and other cis-regulatory elements)
Transcribed region
Transcribed region not part of exon 5
3′-Flanking region
Transcription
Polyadenylation signal sequence in hnRNA (AAUAAA)
Cleavage product degraded
Sense (+) strand of DNA (not transcribed)
Primary transcript (hnRNA)
Template or anti-sense (–) strand of DNA (transcribed)
hnRNA processing
5′-UTR
3′-UTR
mRNA
AAA...AAA 3′
ORF
Cap (7-MeG)
Poly(A) Tail
ORF translation
N
C
Polypeptide product
7-MeG
1′
OH
5′–5′ linkage
7-MeG cap
mRNA

Gene structure

- Eukaryotic gene
  - Introns (10-100,000 bp) + exons (100-200, <1,000 bp)
  - RNA splicing: intron removal from mRNA
- Prokaryotic gene
  - No introns
  - No RNA splicing
- Regulatory element
  - Signal sequences in DNA that control transcription, RNA splicing and translation

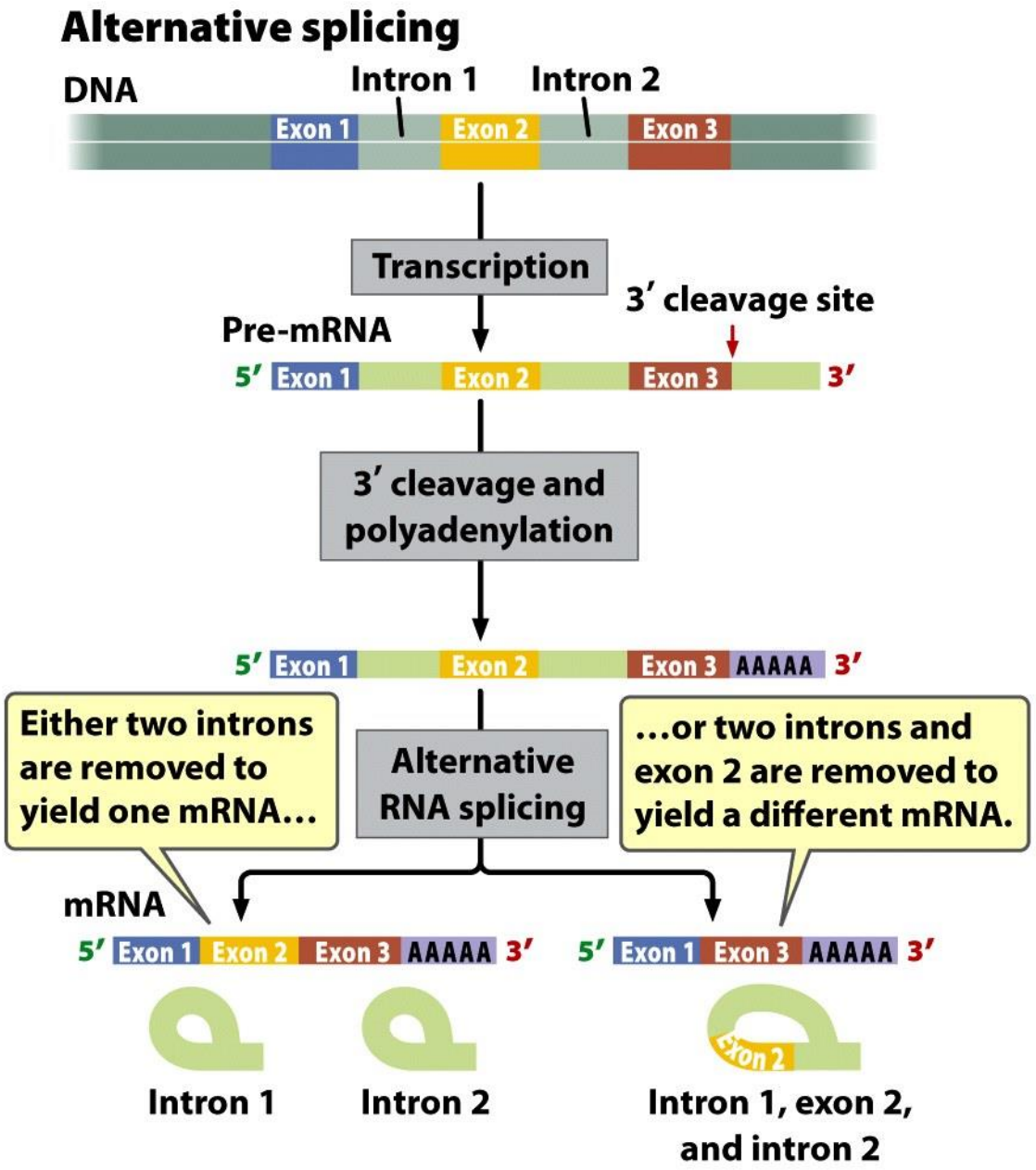


Figure 14-13a
Genetics: A Conceptual Approach, Third Edition

Gene structure

- The primary transcript that contains both exons and introns is called the heterogeneous nuclear RNA (hnRNA) or pre-mRNA
- In mRNAs, a few terminal exons are noncoding, whereas the internal exons code for amino acids. These terminal noncoding exons form the 5'- and 3'-untranslated regions (UTRs) of the mRNA
- The majority of internal exons in vertebrate genes are less than 300 bp; the average length being 135 bp; exons larger than 800 bp are rare.

Gene structure

- For most genes, the last exon (at the 3’-end) is the longest exon (could be well over 1 kb) and partially coding
- For most genes, the 5’-UTR is derived from more than one exon
- For most genes, the 3’-UTR is 3 to 5 times longer than the 5’-UTR, particularly in vertebrates
- In vertebrates, exons are small and introns are large. In contrast, in lower eukaryotes, the opposite is true

Genomes & Transcriptomes

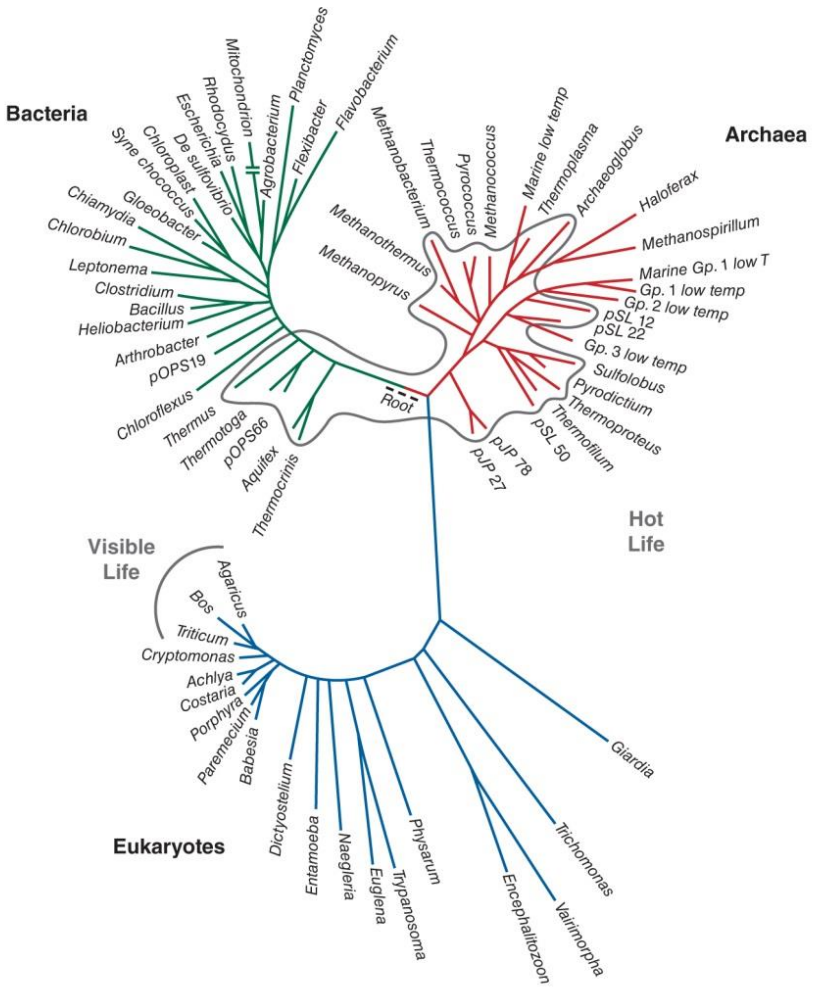
- Genome: total DNA in a single cell or an organism
- Genomics: a study of the DNA sequence, organization, function & evolution of genomes
- Transcriptome: a set of all RNA molecules, including mRNA, rRNA, tRNA, and non-coding RNA produced in one or a population of cells.
  - Unlike the genome, which is roughly fixed for a given cell line (excluding mutations), the transcriptome can vary with external environmental conditions.
- Transcriptomics (expression profiling) examines the expression level of mRNAs in a given cell population

Proteomes

- Proteome is the complete set of proteins encoded in the genome
  - Most importantly, while the genome is a rather constant entity, the proteome differs from cell to cell and is constantly changing through its biochemical interactions with the genome and the environment. One organism will have radically different protein expression in different parts of its body, in different stages of its life cycle and in different environmental conditions.
- Proteomics: a study of the complement of proteins to identify their localization, functions, interactions

The tree of life & evolution

- All existing life has evolved from a single common ancestor ~ 10 billion years ago
- Species of the present and the past are related to each other like the branches of a tree
- There are 3 domains of life



Origin of evolution

- DNA sequences change as a result of mutations
- Mutations can be nucleotide changes, deletions, insertions and duplications
- Mutations have different effects on gene expression, protein function and organism fitness
- Mutations can be retained or lost, depending on the process of natural selection
- The diversity of life is the result of evolution (DNA mutations + natural selection)
- Mutation mechanism & evolution model are the foundation for bioinformatics analysis of DNA & protein sequences

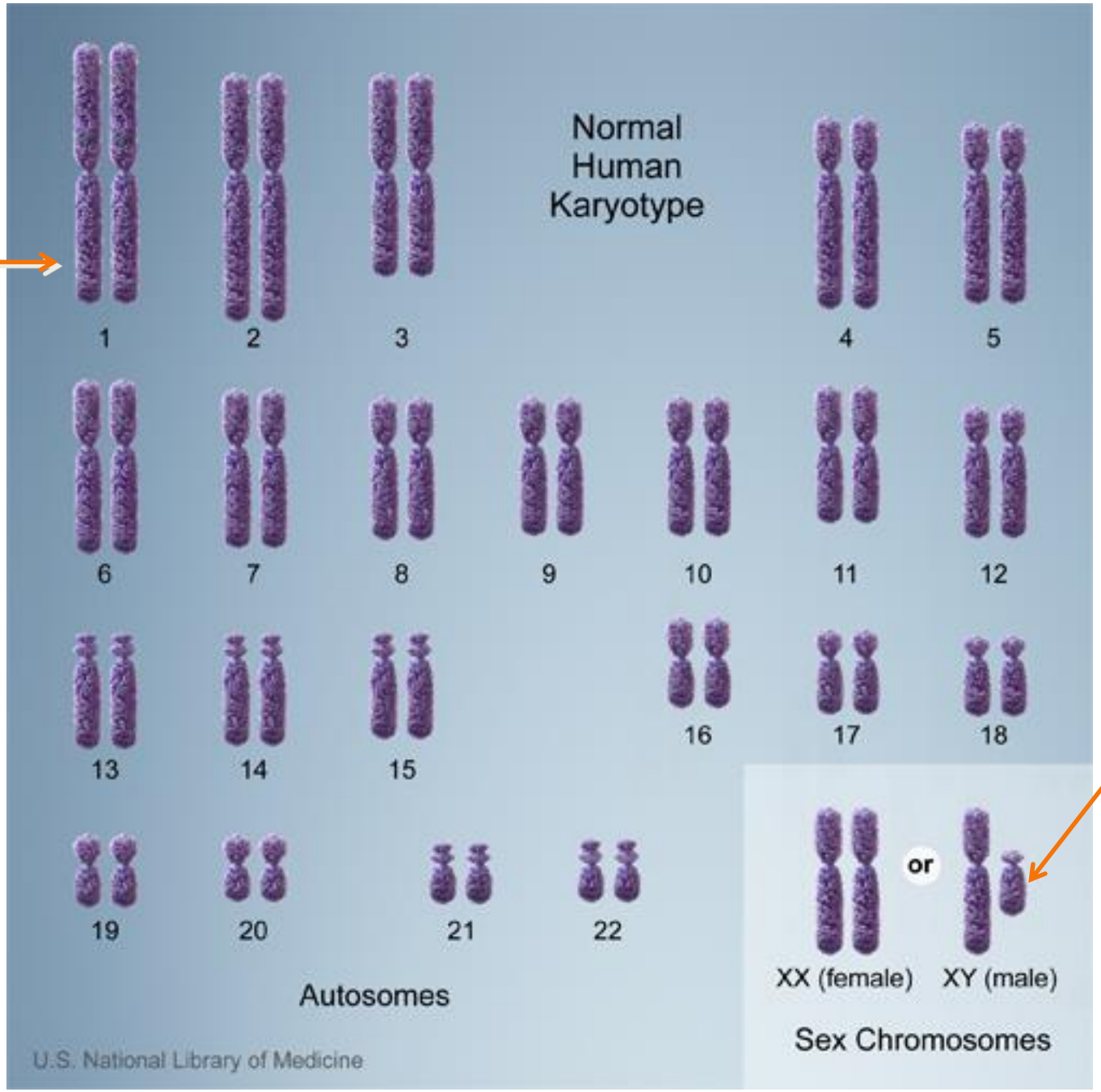
Human genome

- Genome size: 3.2 billion base pairs (3.2×10^9 bp or 3.2 Gb)
  - Arranged in 22 pairs of autosome & 1 pair of sex chromosome
  - Human genome is not the largest, some plants have genomes as large as 100 Gb
- About 20,000 – 25,000 protein-coding genes # 1.5-2% of the whole genome
  - Number of genes in a genome is not directly correlated with the complexity of organism: 6,000 genes in yeast, 15,000 genes in fruit fly and 40,000 genes in rice

Human genome

- Regulatory sequences constitute ~ 3- 3.5% of the genome
- Repeat sequences account for ~ 50% of the human genome. Hence repeat sequences constitute a significant source of genetic diversity
  - Simple repeats (e.g. (A)n, (CA)n, (CGG)n)
  - Tandem repeat blocks (e.g. centromeric repeats, telomeric repeats, ribosomal gene clusters),
  - Segmental duplications (e.g. blocks of 1200 kb or longer repeats copied from one region of the genome and integrated into another region of the genome)
  - Interspersed repeats (transposable element- derived)
- The genomes of 2 humans are about 99.9% identical

The largest # 250 Mb



The smallest # 50 Mb

Outline

History of Bioinformatics

Biological foundations of bioinformatics

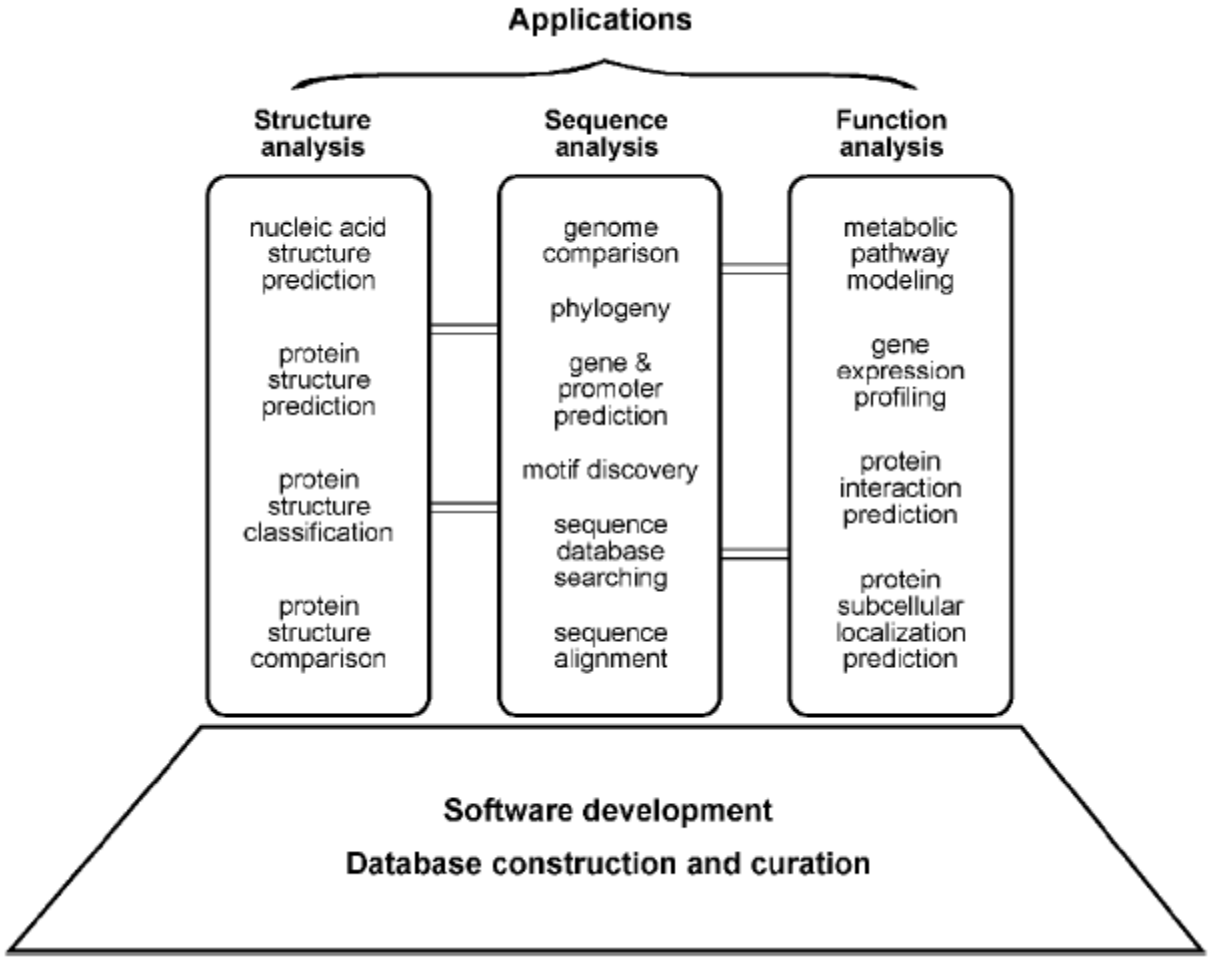
Areas in Bioinformatics

Two subfields of bioinformatics

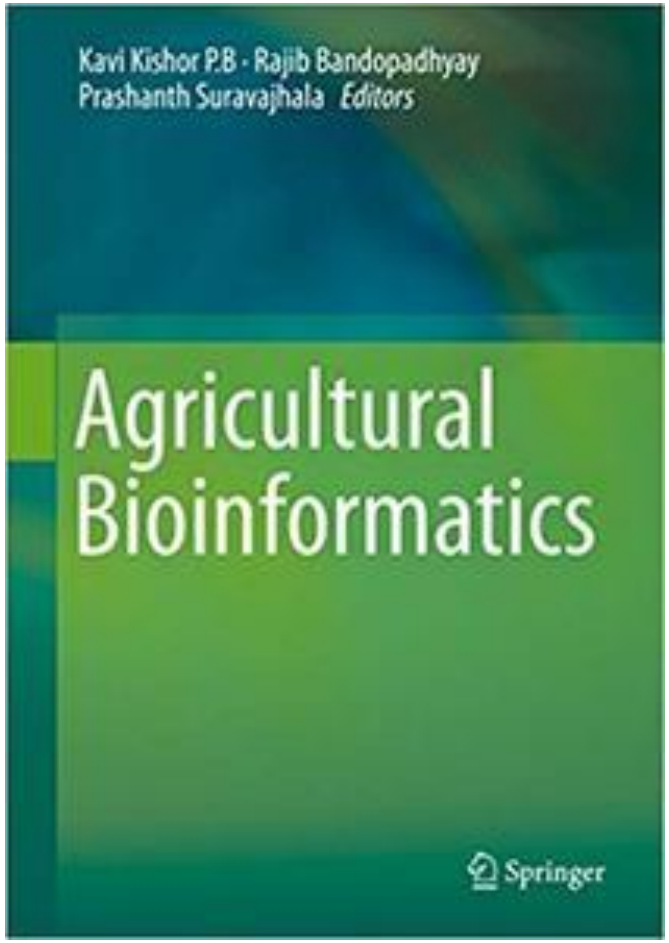
- Development of computational tools & databases
- Application of the tools and databases in generating biological knowledge to better understand living systems

These two subfields are complementary to each other

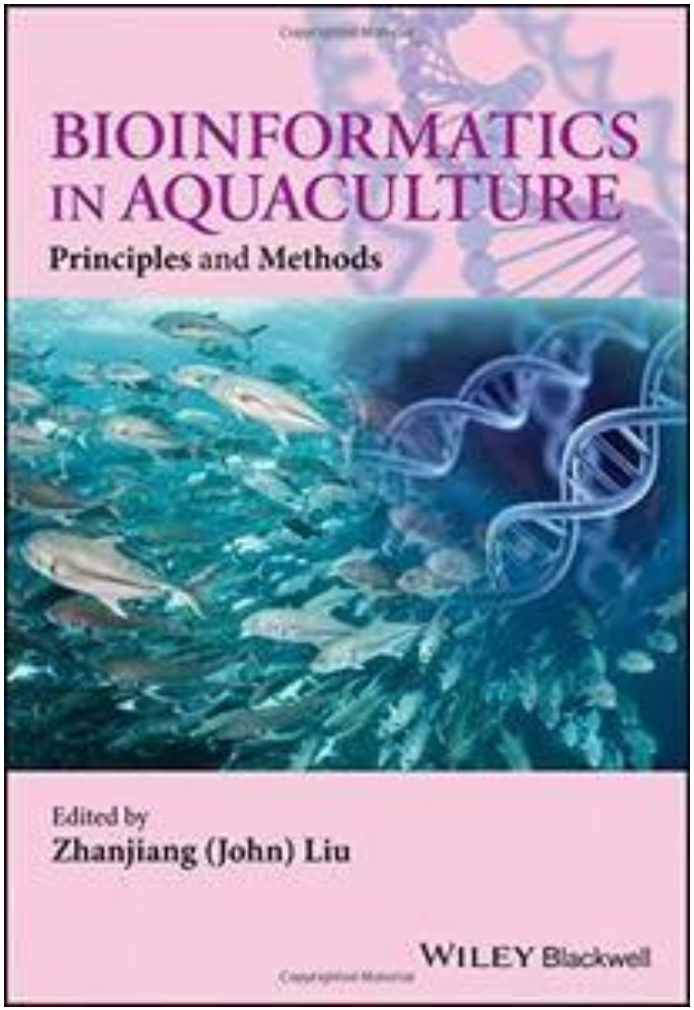
Three areas of genomics & molecular biological research



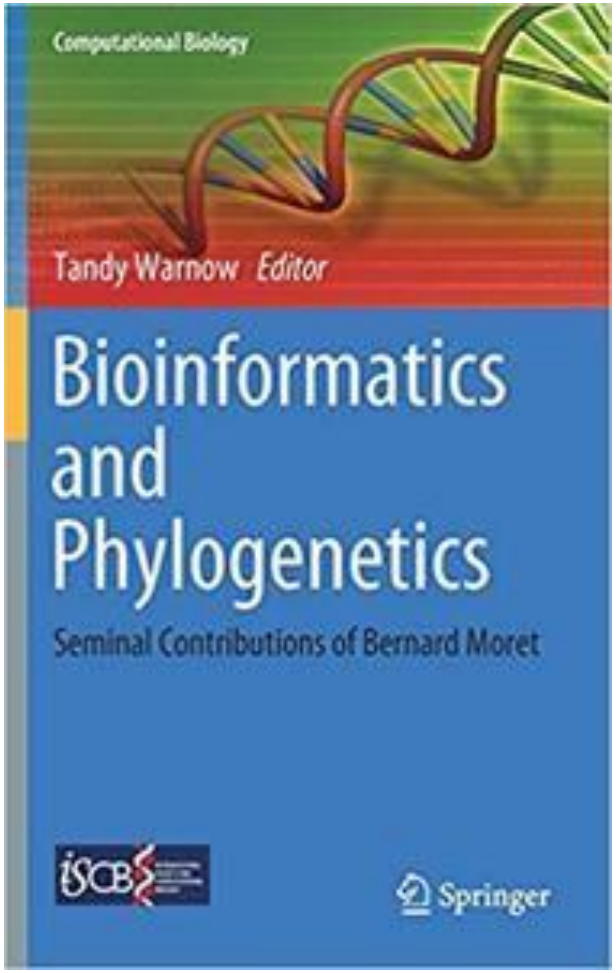
Applications

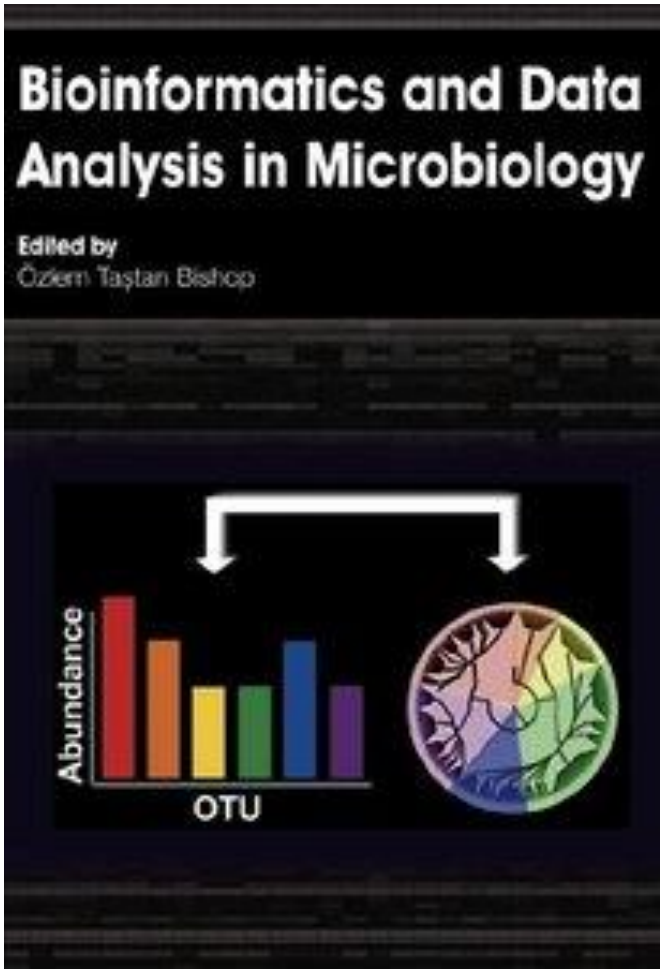


Applications

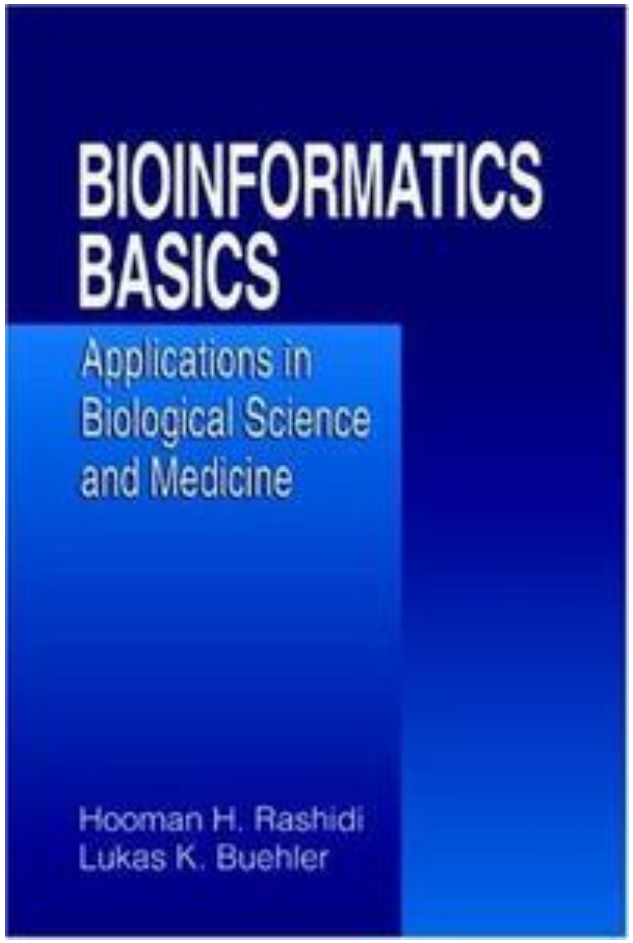


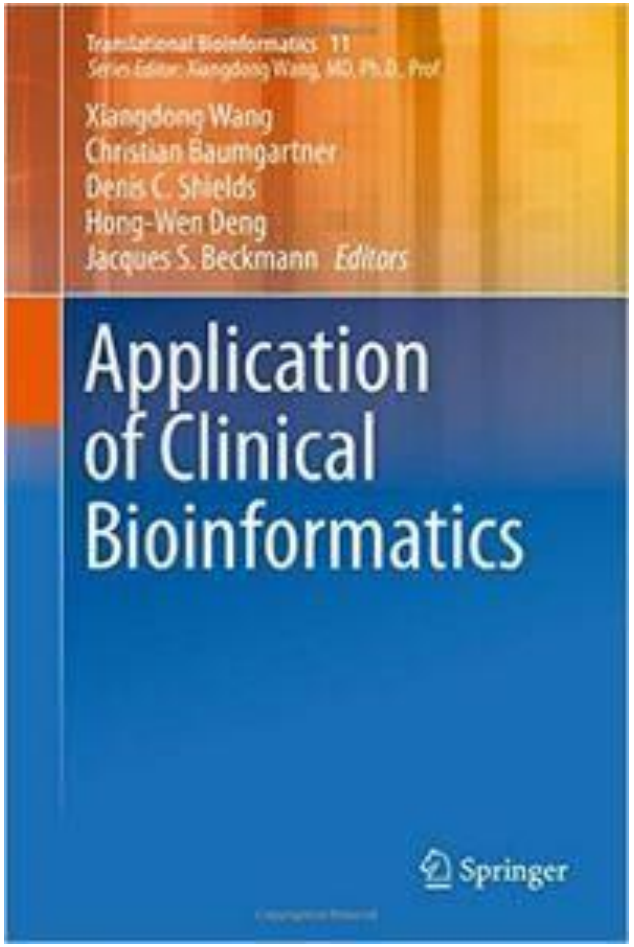
Applications



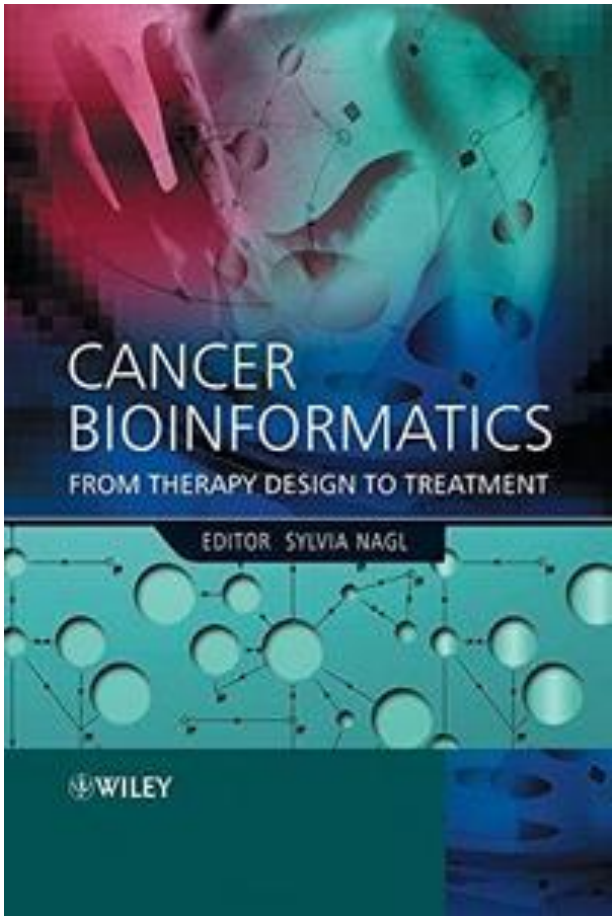


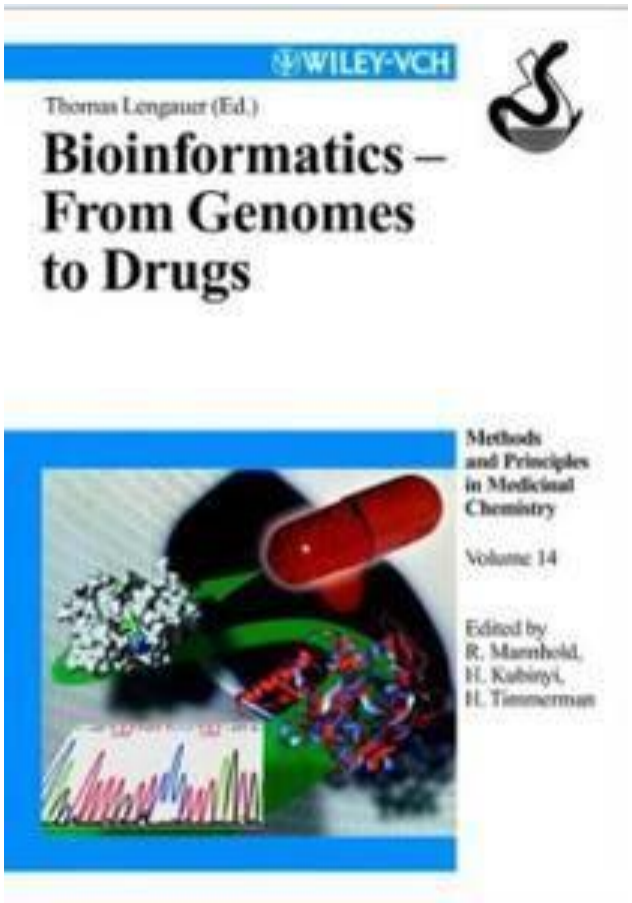
Applications





Applications



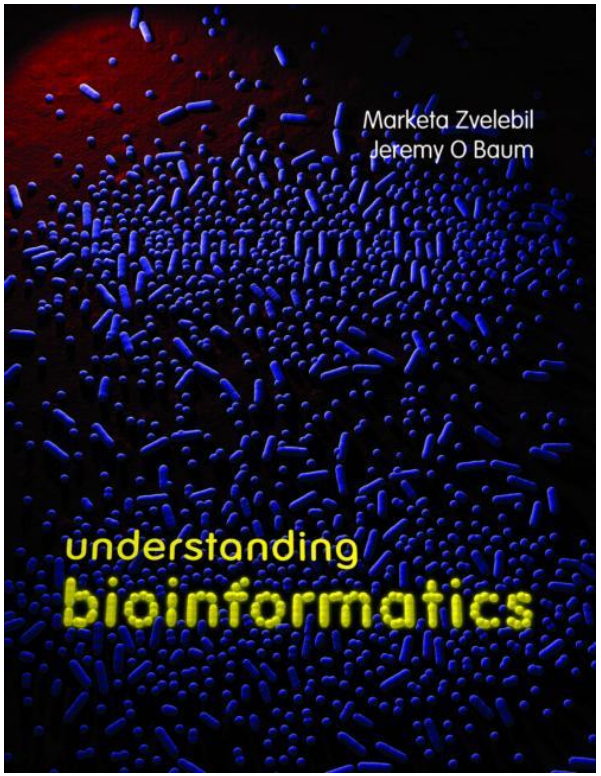


More in application

- Pharmacogenomics is the science that allows us to predict a response to drugs based on an individual's genetic makeup
- Pharmacogenetics is the study of genetic basis for variation in drug response
- Pharmacoinformatics concentrates on the aspects of bioinformatics dealing with drug discovery
- Personalized medicine: individualizing drug therapy in light of genomic information


It's all about your genome, and we have something JUST for you.

Background reading



Chapter 1 – Background basics

Chapter 12 – Bioinformatics – Research Applications