Bioinformatics: Sequence Analysis

COMP 571

Luay Nakhleh, Rice University

Course Information

* Instructor: Luay Nakhleh (nakhleh@rice.edu); office hours by appointment (office: DH 3119)
* TA: Leo Elworth (DH 3121; ryan.a.leo.elworth@rice.edu); office hours by appointment
* Meeting time and place: T&TH 9:25-10:40, HZ 210
* Website: http://www.cs.rice.edu/~nakhleh/COMP571

Grading

* A set of homework assignments: 50%
* Midterm 1: 25%; in-class on 23 February 2017
* Midterm 2: 25%; in-class on 20 April 2017

Course Textbooks

Highly recommended, but not required

* Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids
    * Durbin et al., Cambridge University Press
* Algorithms on Strings, Trees, and Sequences
    * Gusfield, Cambridge University Press
* Genome-scale Algorithm Design
    * Makinen et al., Cambridge University Press
* A list of other recommended books is available on the course website

Intended Audience

* This is a computer science course!
* The course uses mathematics and algorithms, and homework assignments and exams will include these (and assume knowledge of programming).
* This is NOT a "programming for biologists" course!
* This is NOT a course about how to use bioinformatics tools or databases!
* Students are expected to have had (or are currently taking) an algorithms course, can program, and are not afraid of math.

Tentative List of Topics

* Pairwise sequence alignment
* Markov chains and HMMs
* Pairwise alignment using HMMs
* Profile HMMs for sequence families
* Multiple sequence alignment
* Phylogenetic tree inference
* Phylogenomics
* Suffix trees
* The Burrows-Wheeler transform
* Read alignment
* Genome compression
* Applications from genomics, transcriptomics, and metagenomics

* I teach COMP 182 immediately after this class (10:50 - 12:05 on TR)!

* So, I need to leave the classroom by 10:40.

* Please talk to Leo (the TA) first.

* Questions about administrivia?

Background

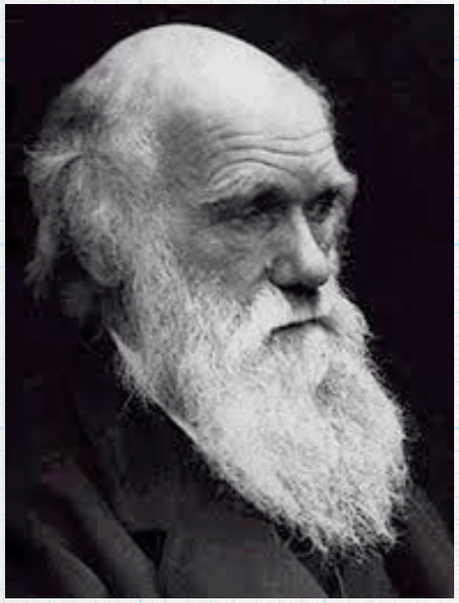
Life Through Evolution

* All living organisms are related to each other through evolution
* This means: any pair of organisms, no matter how different, have a common ancestor sometime in the past, from which they evolved
* Evolution involves inheritance, variation, and selection

Life Through Evolution

* Inheritance: passing of characteristics from parents to offsprings*
* Variation: process that leads to differences between parent and offspring
* Selection: favoring certain individuals over others doe to phenotypic differences

* this is "challenged" by horizontal gene transfer



I have called this principle, by which each slight variation, if useful, is preserved, by the term **Natural Selection**.

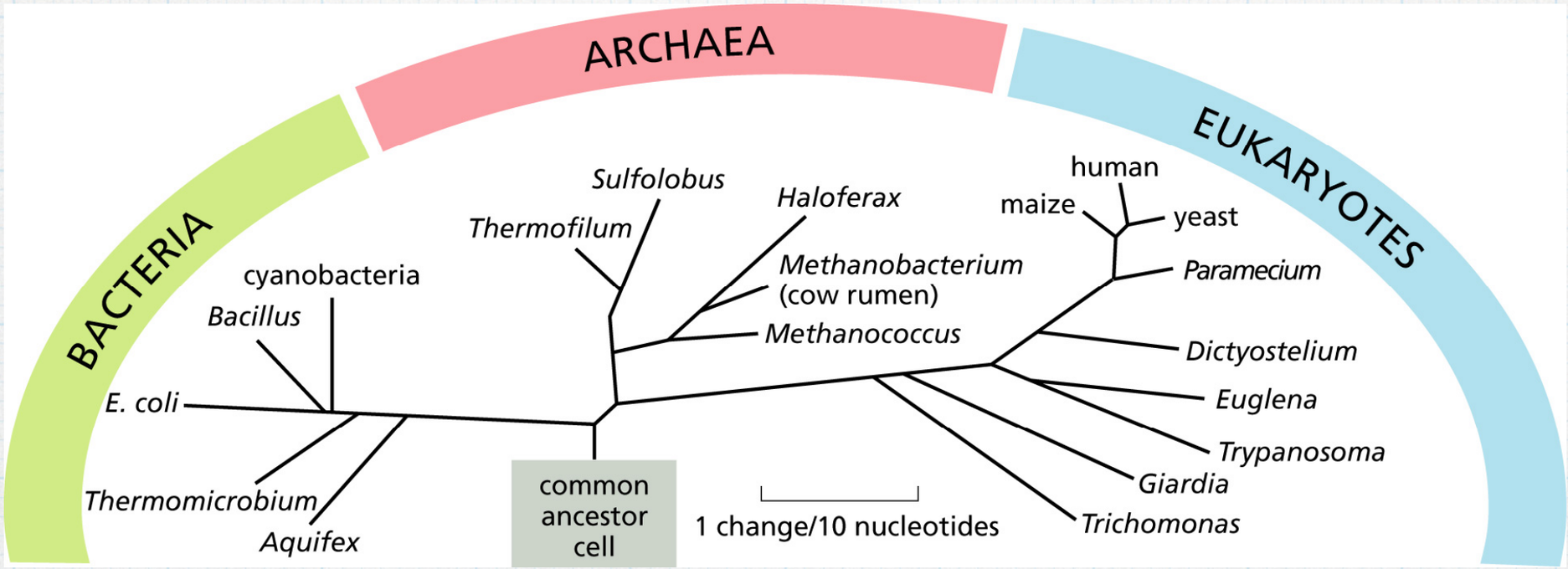
The **[neutral] theory** does not deny the role of natural selection in determining the course of adaptive evolution, but it assumes that only a minute fraction of DNA changes in evolution are adaptive in nature, while the great majority of phenotypically silent molecular substitutions exert no significant influence on survival and reproduction and drift randomly through the species.

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

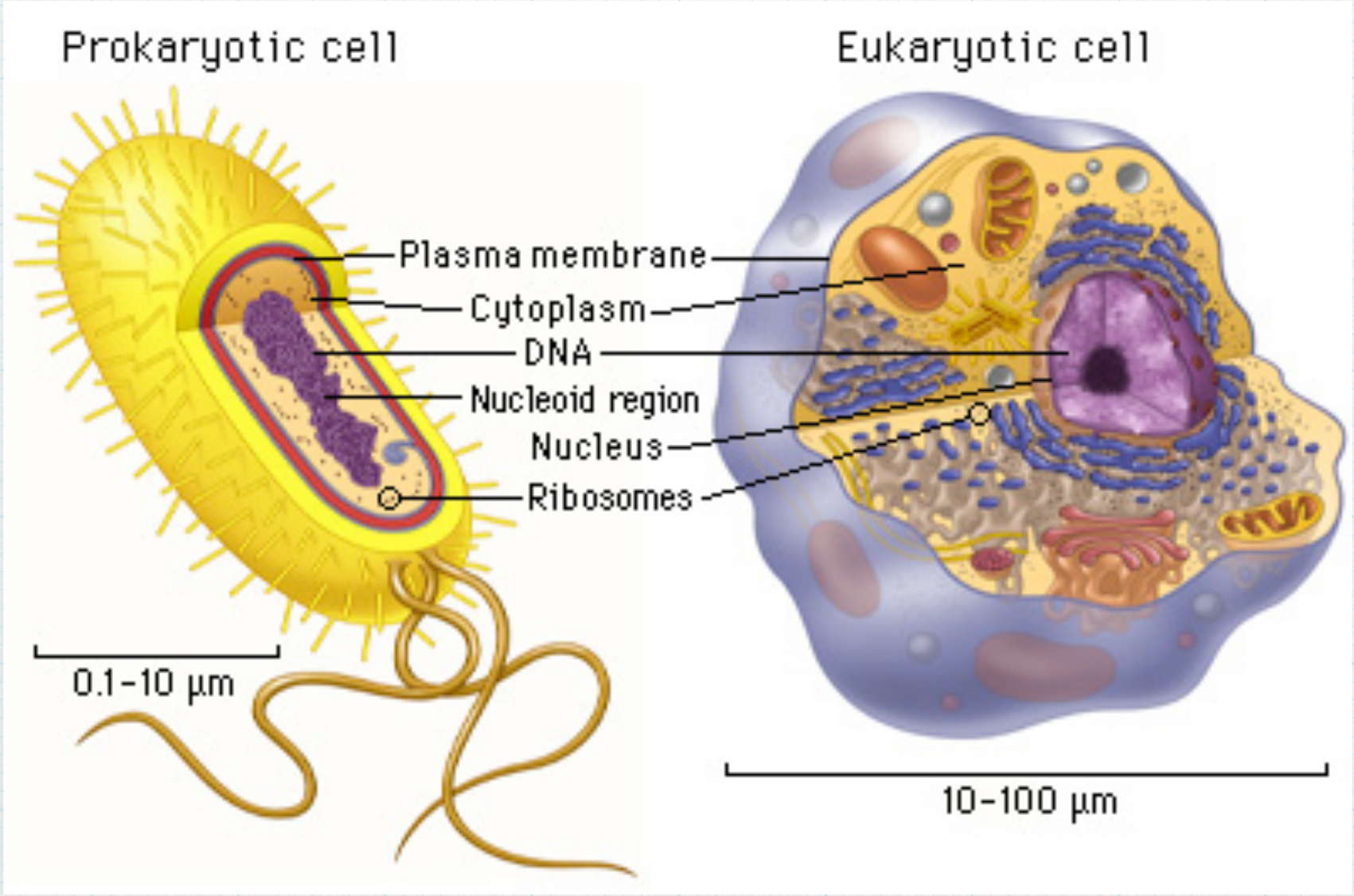
Evolution

* The accumulation of change over time in a population
* **Population genetics** mainly focuses on evolutionary analysis of changes within populations, whereas **phylogenetics** is mostly aimed at inter-species relationships
    * We will discuss later (under "phylogenomics") how these two disciplines are now coming together due to dense sampling of genomes.

The Tree of Life



Prokaryotic vs. Eukaryotic Cell Structure



Source: Pearson Education, Inc. The Biology Place

Prokaryotic vs. Eukaryotic Cells

	Prokaryotes	Eukaryotes
Size	1-10 μm in length	10-100 μm in length
Nucleus	does not exist	exists, and separated from the cytoplasm
Intracellular organization	no compartments	compartments (nucleus, cytosol, mitochondria, etc.)
Gene structure	no introns	introns and exons
Cell division	simple cell division	mitosis or meiosis
Ribosome	consists of a large 50S subunit and a small 30S subunit	consists of a large 60S subunit and a small 40S subunit
Reproduction	parasexual recombination	sexual recombination
Organization	mostly single cellular	mostly multicellular, and with cell differentiation

Source: Systems Biology in Practice, Klipp et al.

The Nucleic Acid World

* The full diversity of life on this planet—from the simplest bacterium to the largest mammal—is captured in a linear code inside all living cells.

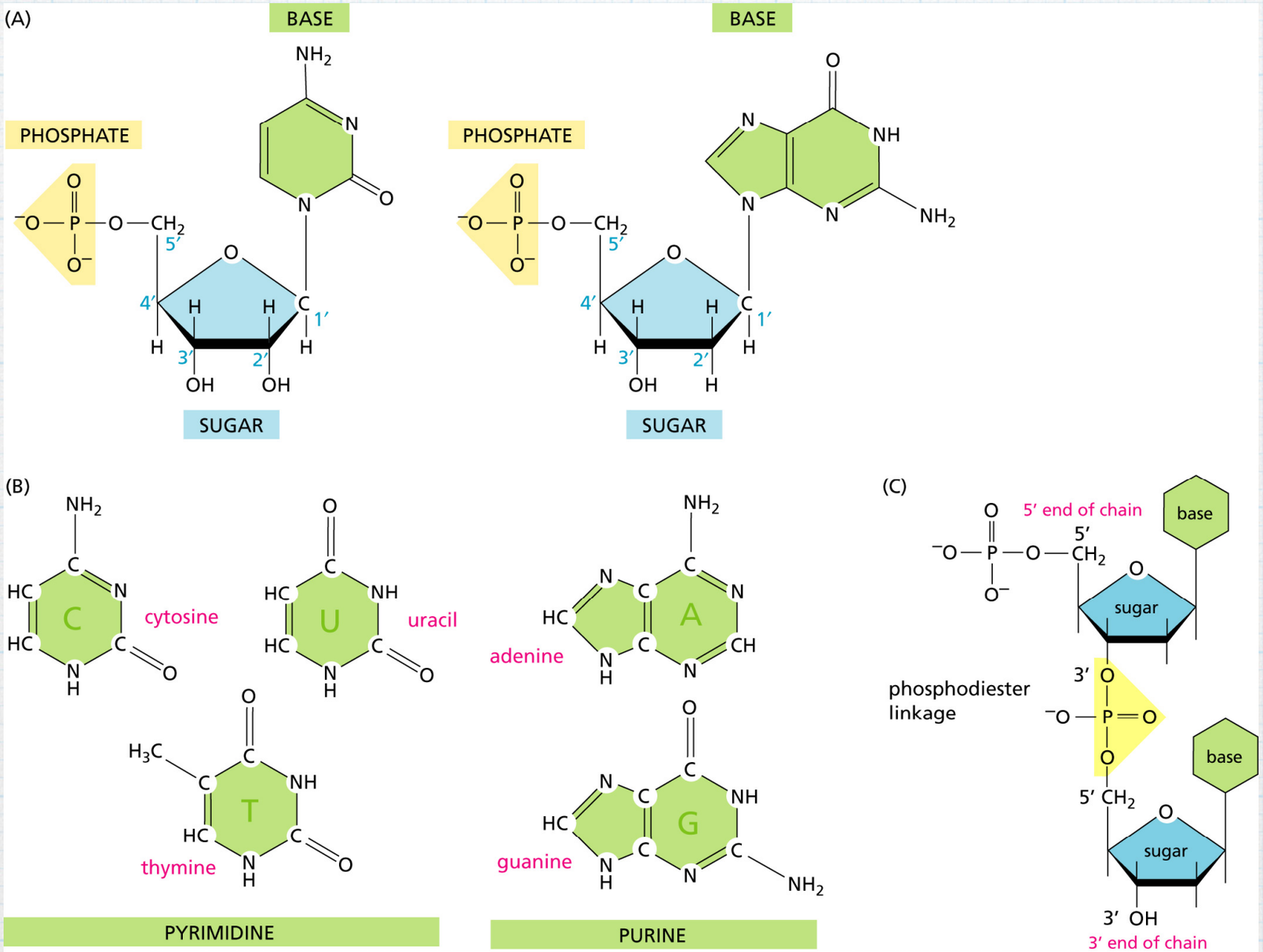
DNA

* Deoxyribonucleic Acid
* DNA molecules are linear polymers of just four different nucleotide building blocks.
* Genomic DNA molecules are immensely long, containing millions of bases each, and it is the order of these bases, the **nucleotide sequence** or **base sequence** of DNA, which encodes the information for making proteins.

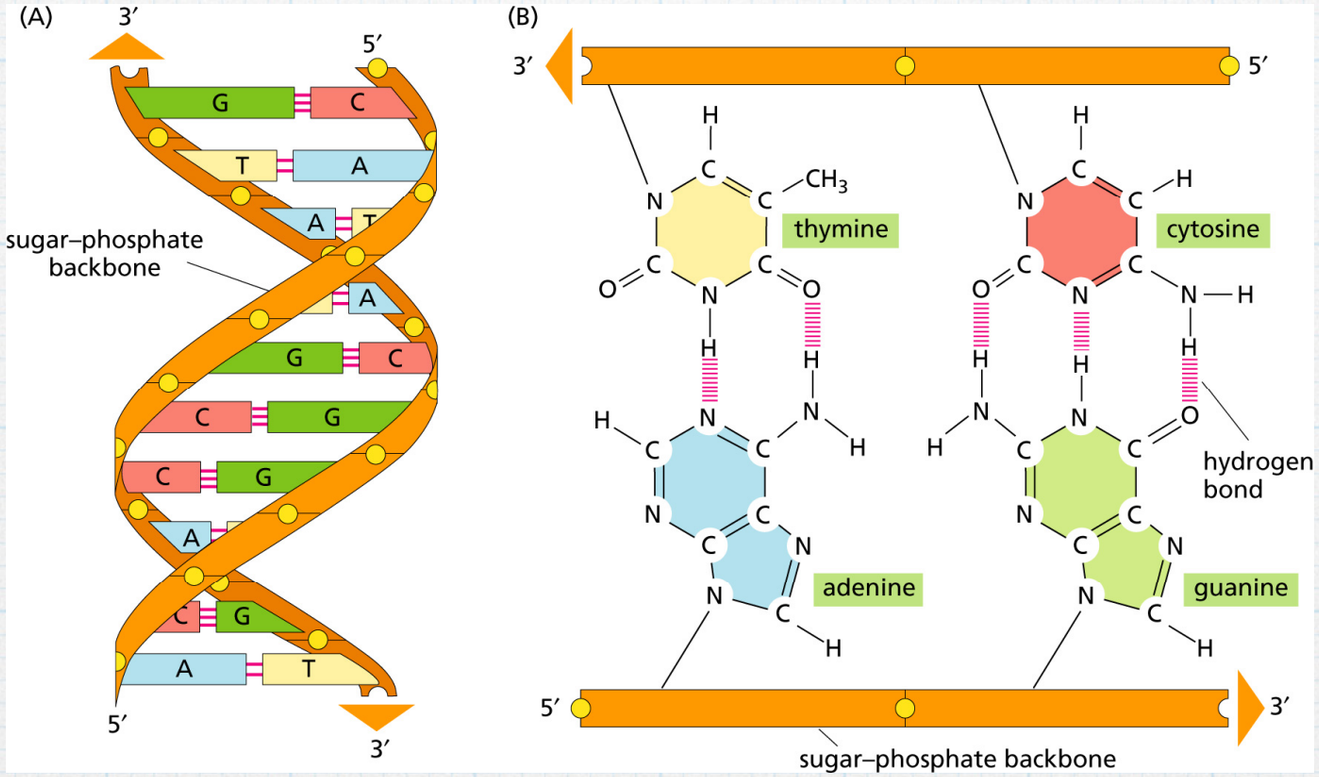
RNA

* Ribonucleic Acid
* RNA molecules are also linear polymers, but are much smaller than genomic DNA.
* Most RNA molecules also contain just four different base types.
* Several classes of RNA molecules are known, some of which have a small proportion of other bases.

The Building Blocks of DNA and RNA



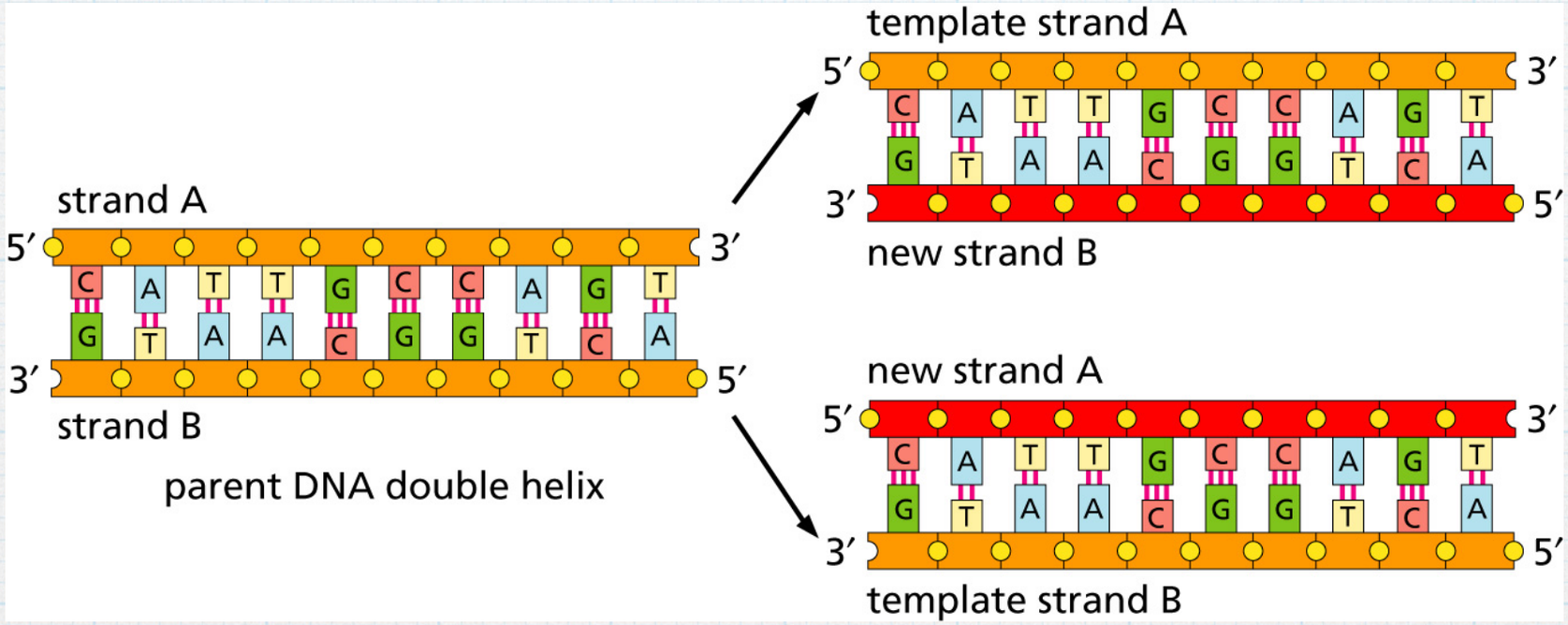
The Double Helix (DNA)



Watson-Crick base-pairing: A—T, C—G

Each strand of a DNA double helix has a base sequence that is **complementary** to the base sequence of its partner strand.

DNA Replication



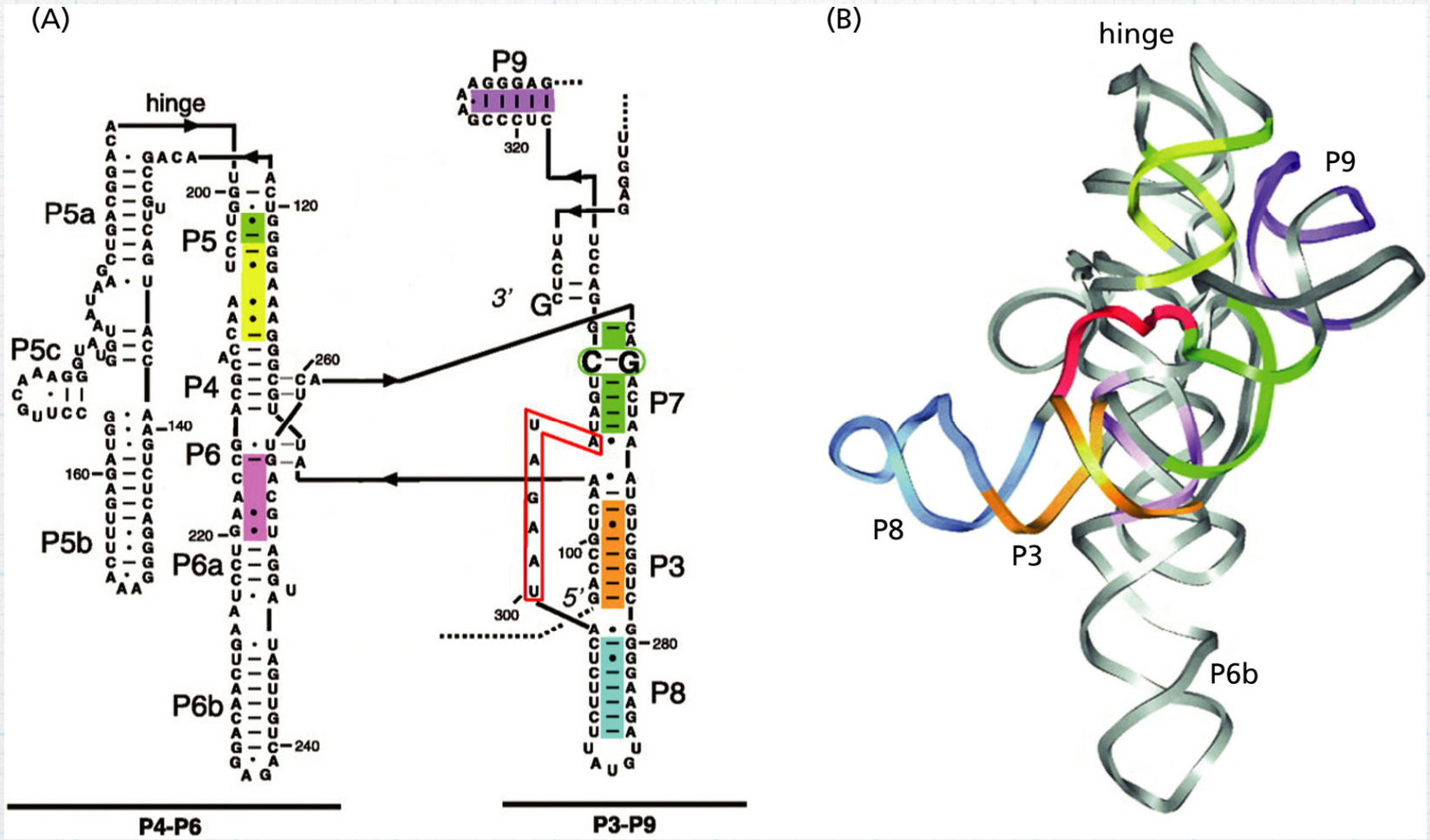
* Hydrogen bonds are noncovalent bonds: the two DNA strands can be easily separated.
* There are a number of processes in which strand separation is required.
* One such process is DNA replication, which is a necessary prelude to cell division.

RNA Structure

* Almost all RNA molecules in living systems are single stranded.

* As a result, RNA has much more structural flexibility than DNA, and some RNAs can even act as enzymes, catalyzing a particular chemical reaction.

Secondary and Tertiary Structures of RNA



The Tetrahymena ribozyme

The Central Dogma

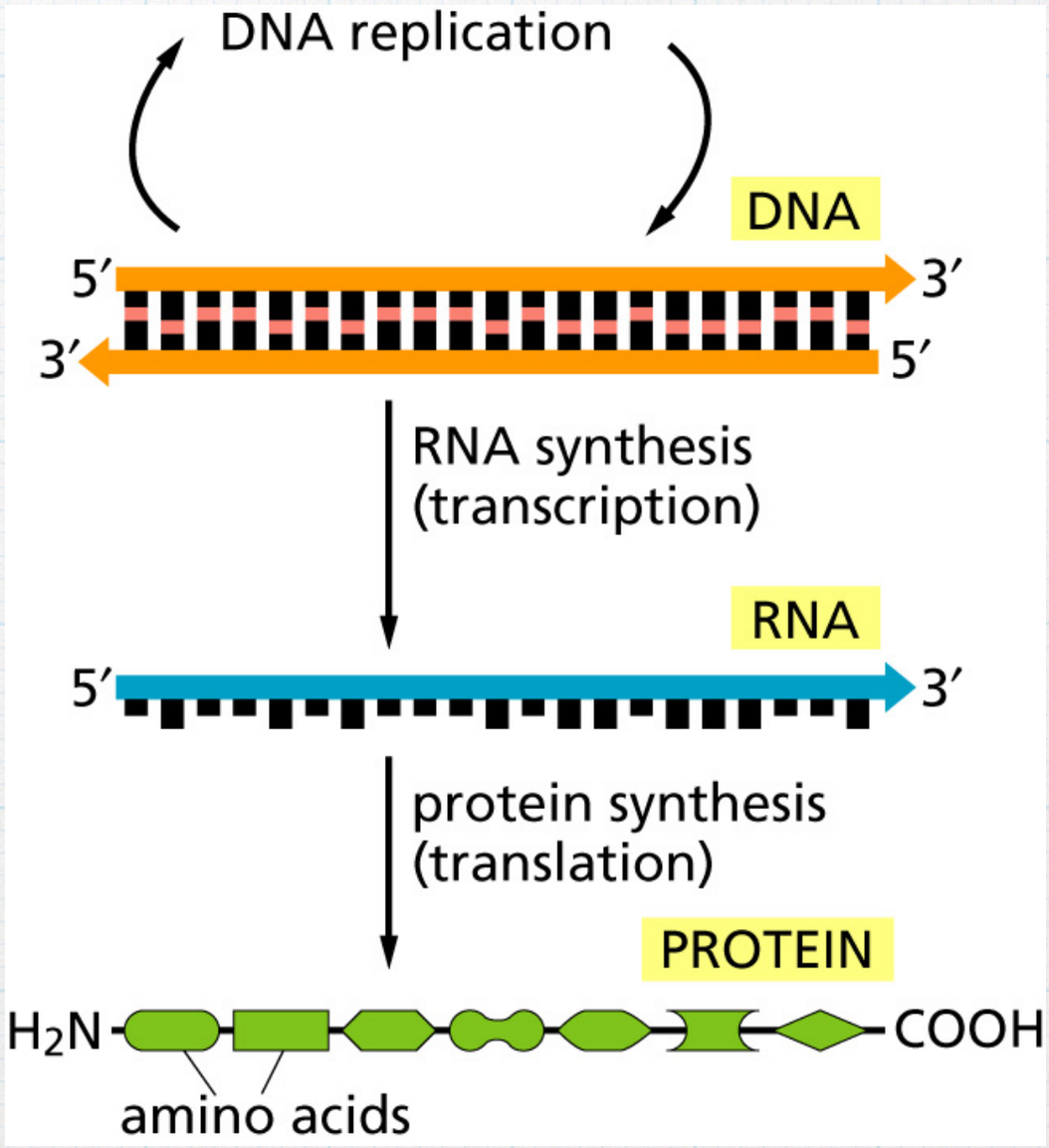
* A single direction of flow of genetic information from the DNA (information store), through RNA, to proteins

* This scheme holds for all known forms of life, with variations in the details of the processes involved in different organisms

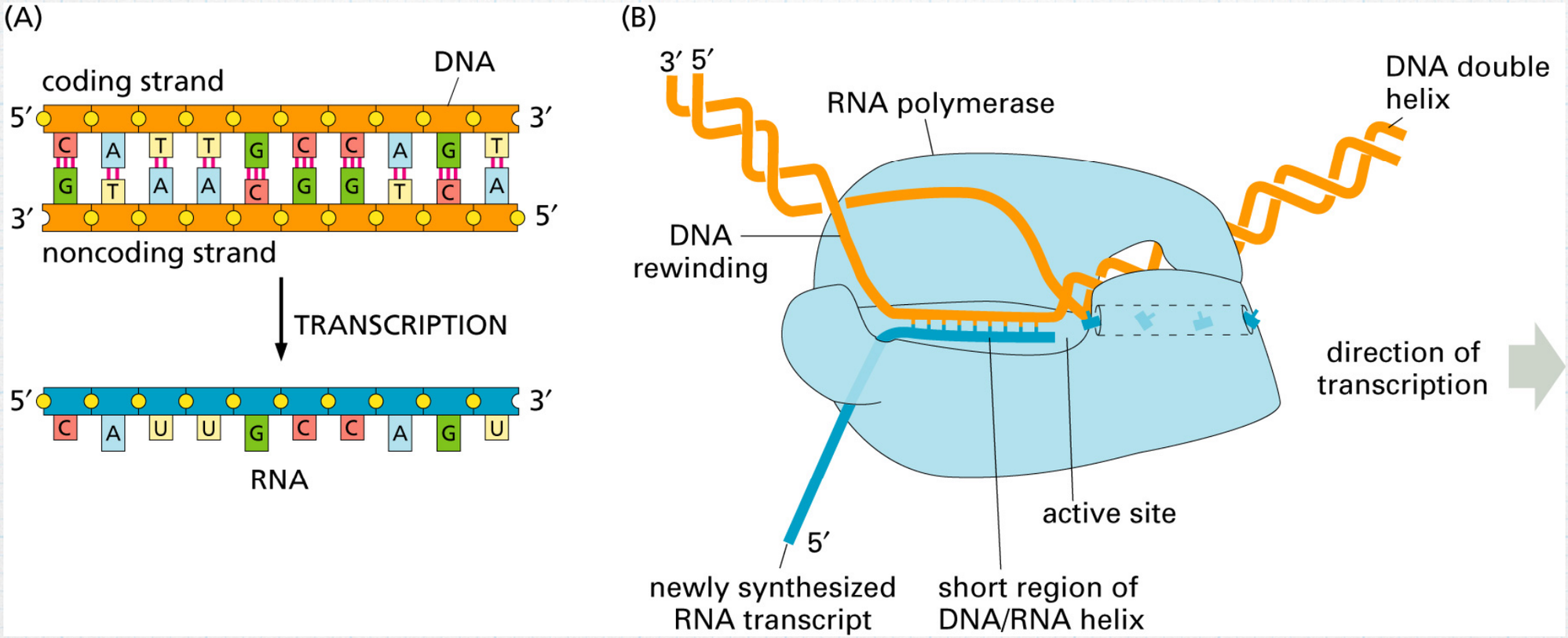
* Not all genetic information in the DNA encodes proteins

* RNA can also be the end product, and other regions of the genome have as yet no known function of product

* The genomic DNA encodes all molecules necessary for life, whether they are proteins or RNA or ...



Transcription



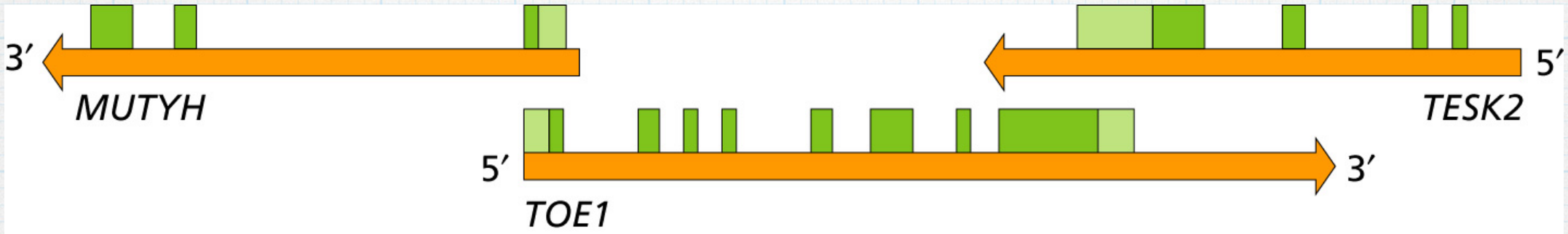
(A) One strand of the DNA is involved in the synthesis of an RNA strand complementary to the strand of the DNA

(B) The enzyme RNA polymerase reads the DNA and recruits the correct building blocks of RNA to string them together based on the DNA code

Terminology

* RNA transcribed from a protein-coding gene is called **messenger RNA (mRNA)**
* When a gene is being transcribed into RNA, the gene is said to be **expressed**

Overlapping Genes



Although only one segment of the DNA strand is transcribed for any given gene, it is also possible for genes to overlap so that one or both strands at the same location (locus) encode parts of different proteins.

This most commonly occurs in viruses as a means of packing as much information as possible into their very small genomes but it could also occur in mammals (the above figure shows overlapping genes in the human genome)

Regulated Gene Expression

* The genomic DNA sequence contains more information that just the protein sequences. The transcriptional apparatus has to locate the sites where gene transcription should begin, and when to transcribe a given gene. At any one time, a cell is only expressing a few thousand of the genes in its genome. To accomplish this **regulated gene expression**, the DNA contains control sequences in addition to coding regions (More on this in a few slides).

Translation

* mRNA is translated into protein according to the **genetic code**, which is the set of rules governing the correspondence of the base sequences in DNA or RNA to the amino acid sequence of a protein.
* Each amino acid is encoded by a set of three consecutive bases (**codon**)

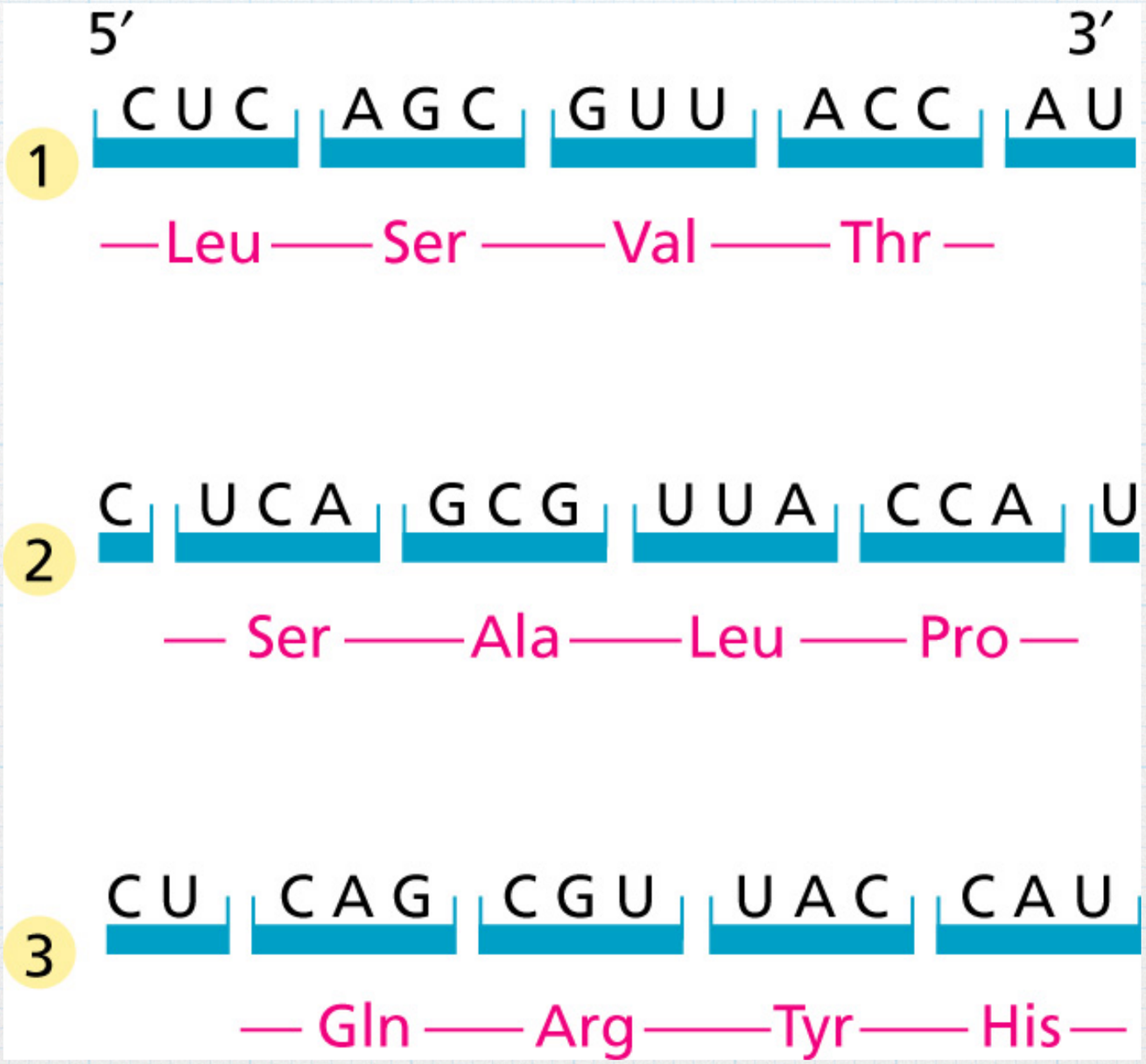
The Standard Genetic Code

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

Reading Frames

* Translation occurs in nonoverlapping sets of three bases.
* There are thus three possible ways to translate any nucleotide sequence, each of which is called a **reading frame**
* These three reading frames give three different protein sequences.
* In the actual translation process, the detailed control signals ensure that only the appropriate reading frame is translated into protein.

Reading Frames



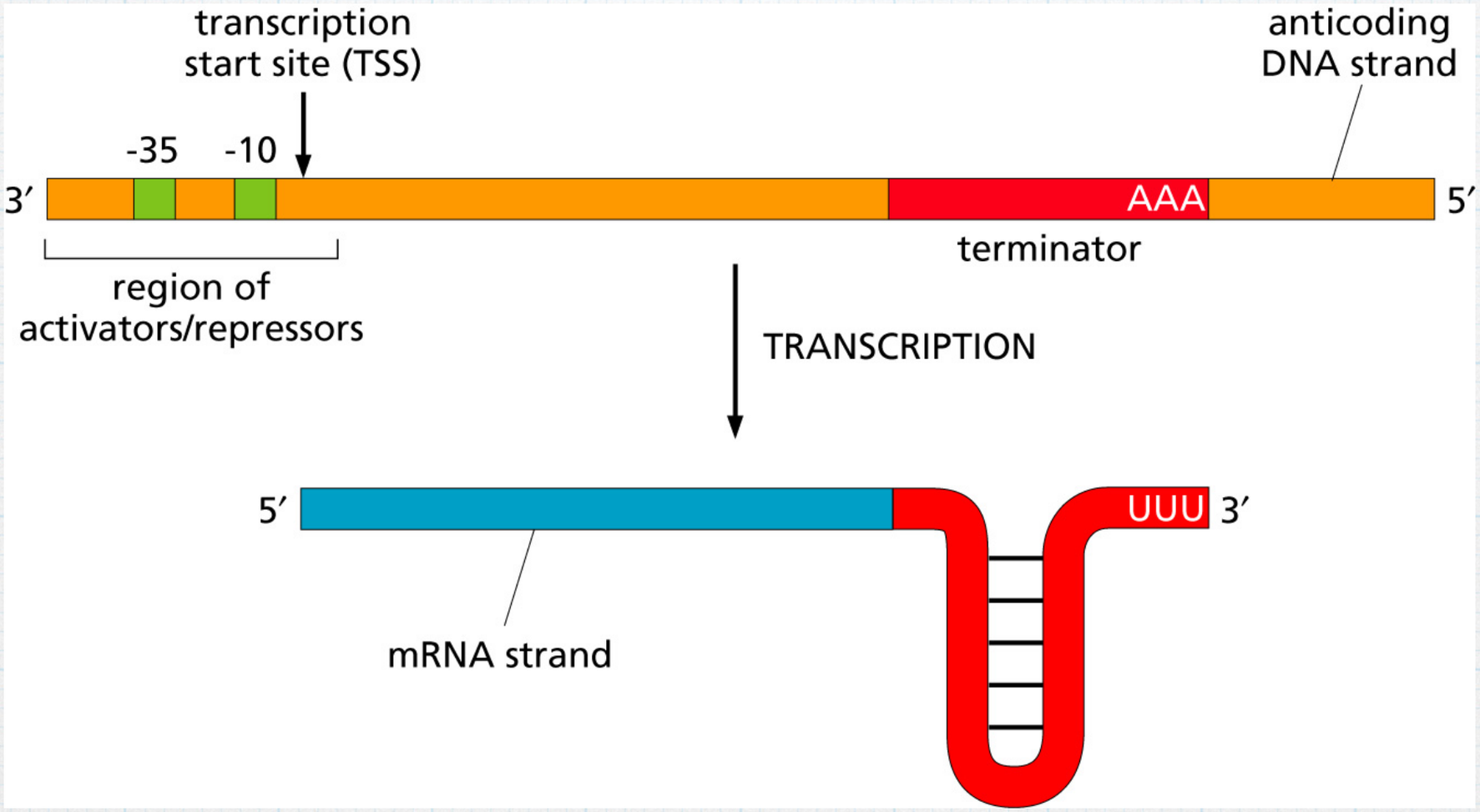
Gene Structure and Control

* The regulation of many processes that interpret the information contained in a DNA sequence relies on the presence of short signal sequences in the DNA.
* The general term for these signal sequences is **regulatory elements**.
* For example, the molecules involved in transcription and translation require signals to identify where they should start and stop.
* Gene structure and control differ between prokaryotes and eukaryotes

Transcription Regulation

* The control regions at which RNA polymerase binds to initiate transcription are called **promoters**.
* RNA polymerase binds more tightly to these regions than to the rest of the DNA and this triggers the start of transcription.

Gene Structure in Prokaryotes

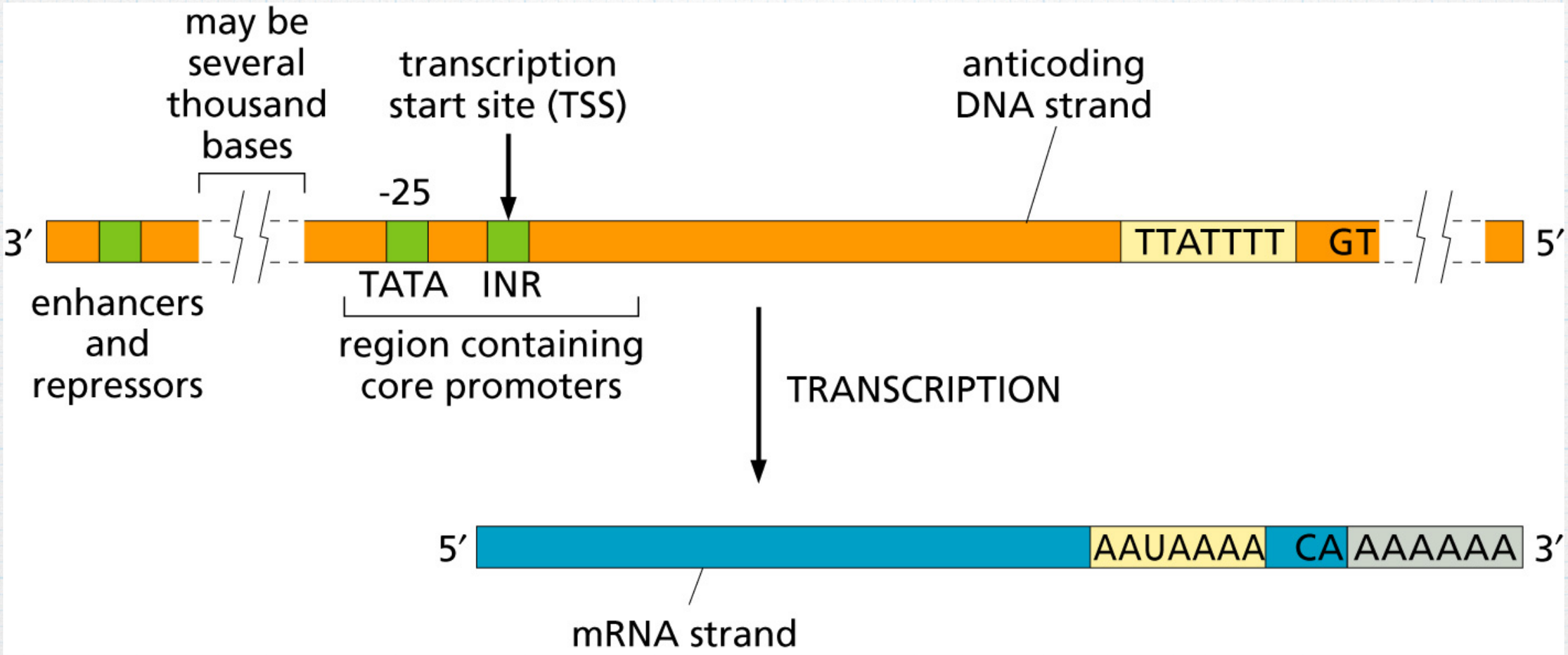


* Bacterial promoters typically occur immediately before the position of the **transcription start site** (TSS), and contain two characteristic short sequences, or motifs, that are almost the same in the promoters for different genes.

* The termination of transcription is controlled by the **terminator signal** which in bacteria differs from the promoter is that it is active when transcribed to form the end of the mRNA strand (forms a loop structure that prevents the transcription apparatus from continuing).

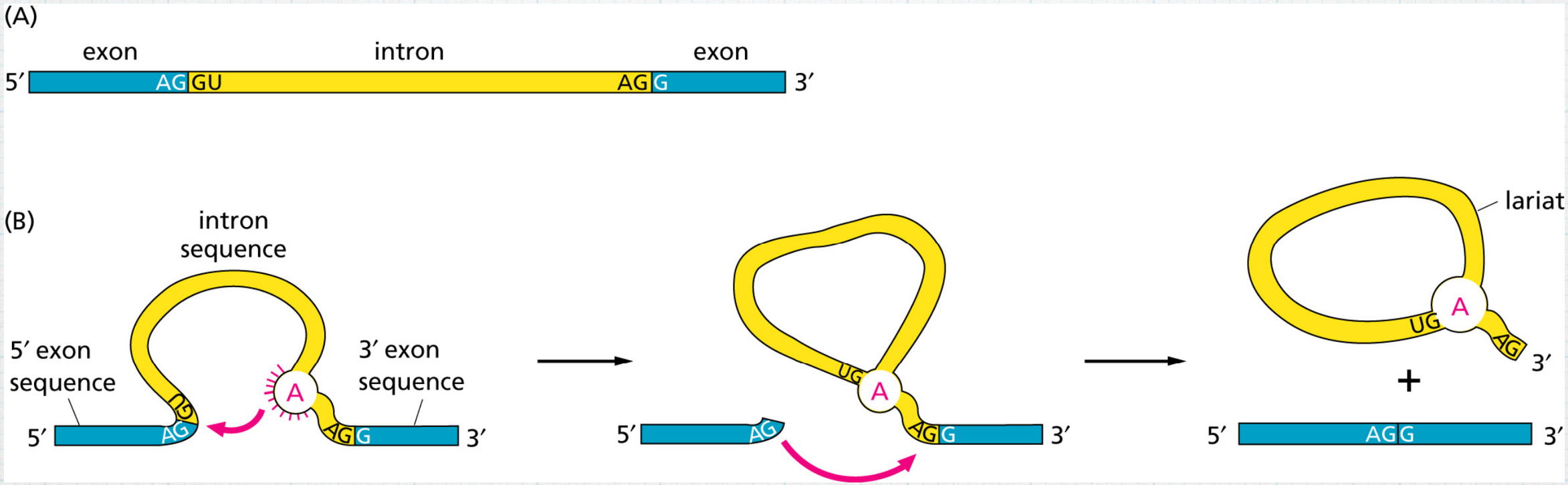
* Single type of RNA polymerase transcribes all genes.

Gene Structure in Eukaryotes



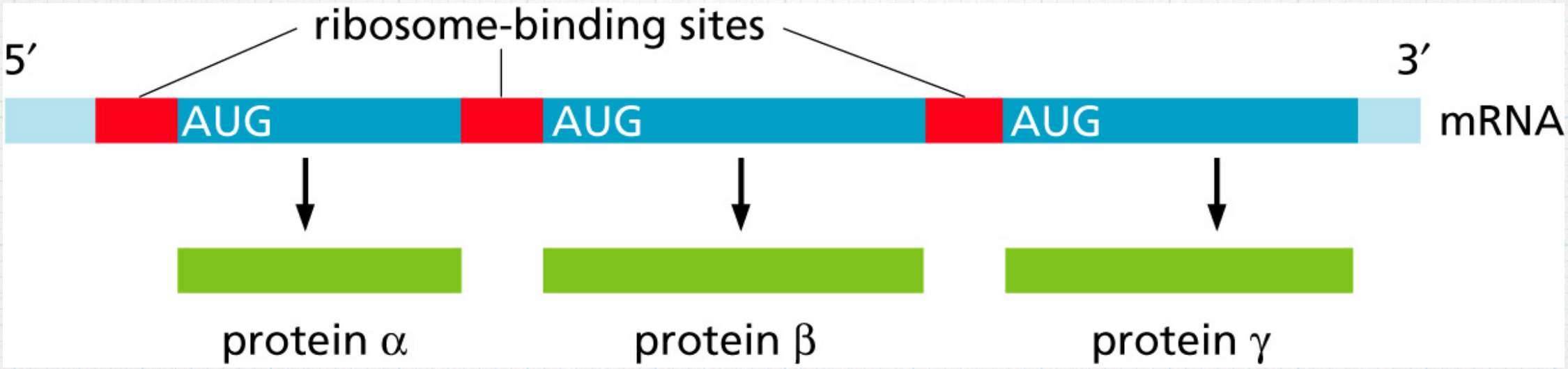
* Regulatory elements in eukaryotes are more complex.
* Three types of RNA polymerase transcribe genes: RNA polymerase II transcribes all protein coding genes, where other RNA polymerase tpyes transcribe genes for tRNAs, rRNAs and other types of RNA

Splicing of an Intron



* The existence of introns necessitates an extra step between transcription and translation, which is known as **RNA splicing**: (1) the complete gene is initially transcribed into RNA, and (2) the introns are then excised and the exons spliced together to provide a functional mRNA that gives the correct protein sequence when translated. In most protein coding genes, this process is carried out by the **spliceosome**, which consists of **small nuclear RNA** (**snRNA**) and proteins.

Operon Structure

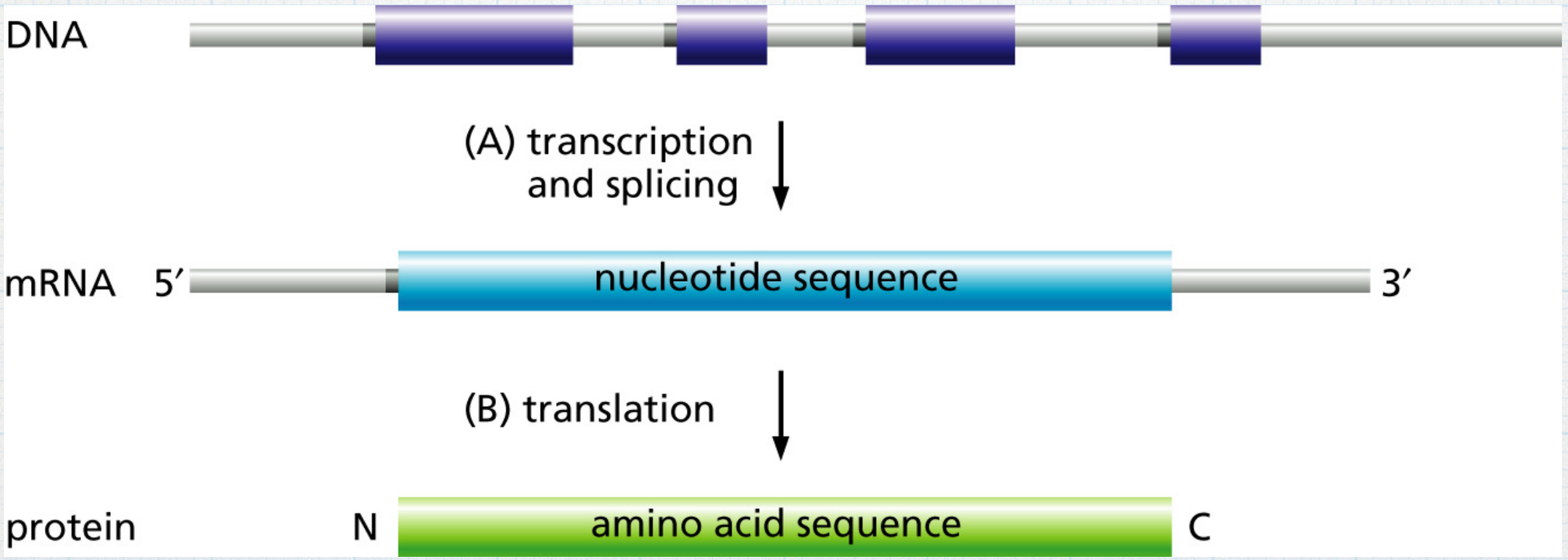


In bacteria, functionally related protein-coding sequences are often clustered together into **operons**. Each operon is transcribed as a single mRNA transcript and the proteins are then separately translated from this one long molecule.

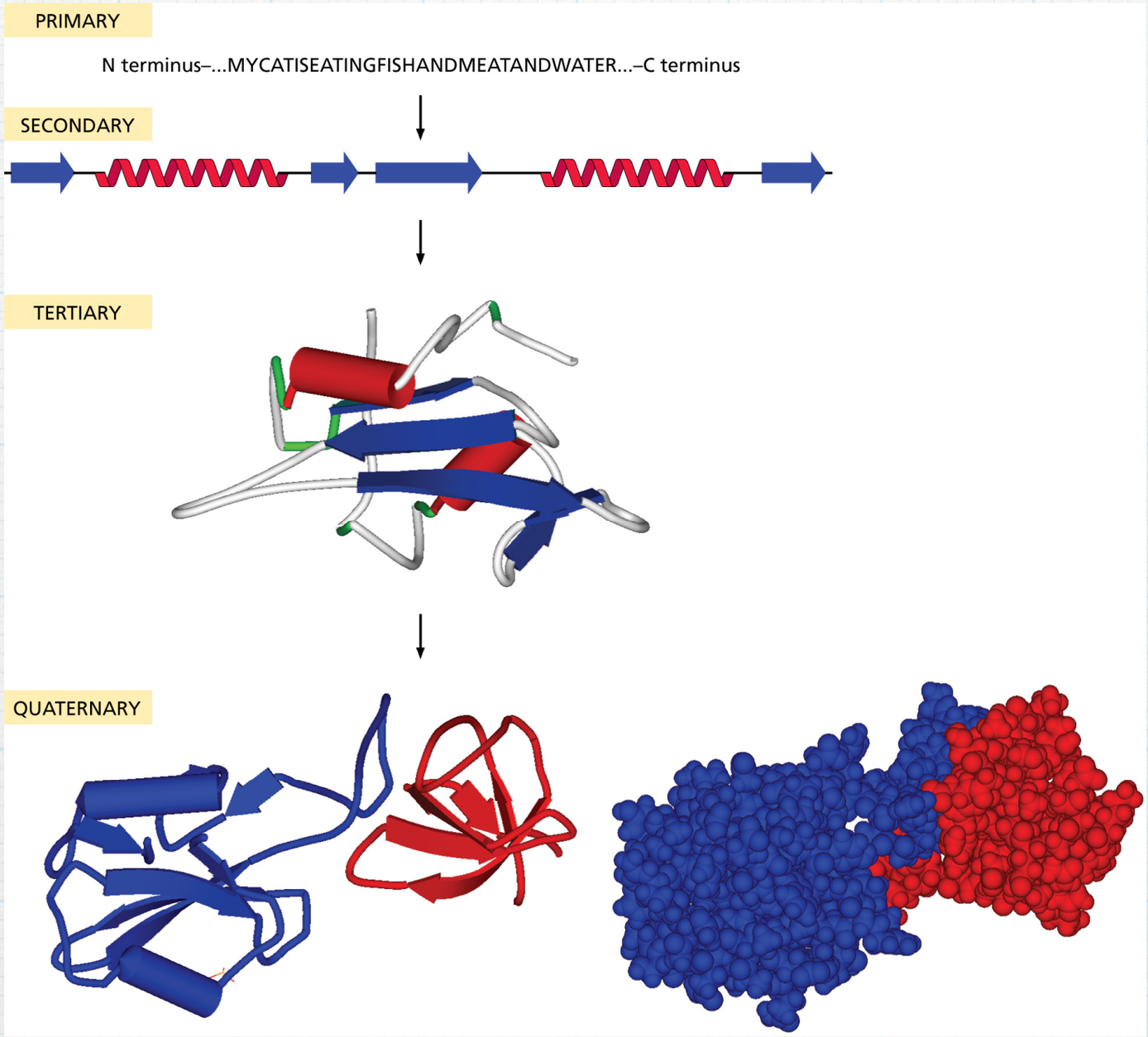
This has the advantage that only one control region is required to activate the simultaneous expression of all genes in the operon.

Not all bacterial genes are contained in operons; many are transcribed individually and have their own control regions.

Proteins


DNA
(A) transcription and splicing
mRNA
5′
nucleotide sequence
3′
(B) translation
protein
N
amino acid sequence
C

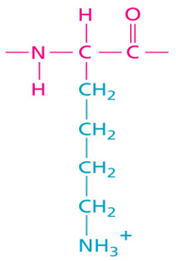
Levels of Protein Structure



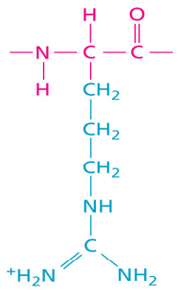
Side Chains of the Amino Acids

BASIC SIDE CHAINS

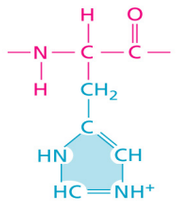
lysine
(Lys, or K)



arginine
(Arg, or R)



histidine
(His, or H)



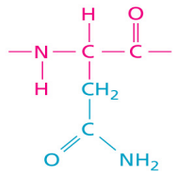
ACIDIC SIDE CHAINS

aspartic acid
(Asp, or D)

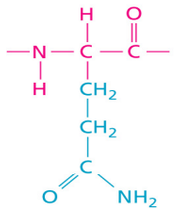
glutamic acid
(Glu, or E)

UNCHARGED POLAR SIDE CHAINS

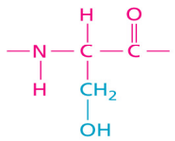
asparagine
(Asn, or N)



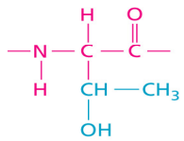
glutamine
(Gln, or Q)



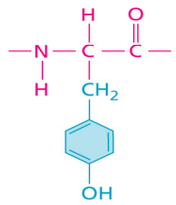
serine
(Ser, or S)



threonine
(Thr, or T)

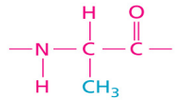


tyrosine
(Tyr, or Y)

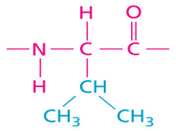


NONPOLAR SIDE CHAINS

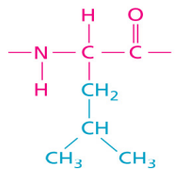
alanine
(Ala, or A)



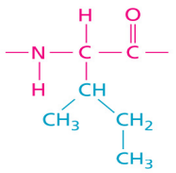
valine
(Val, or V)



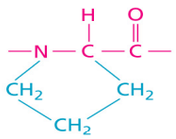
leucine
(Leu, or L)



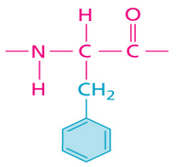
isoleucine
(Ile, or I)



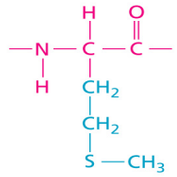
proline
(Pro, or P)



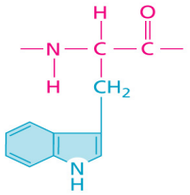
phenylalanine
(Phe, or F)



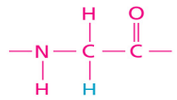
methionine
(Met, or M)



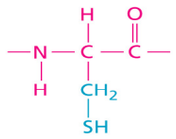
tryptophan
(Trp, or W)



glycine
(Gly, or G)



cysteine
(Cys, or C)



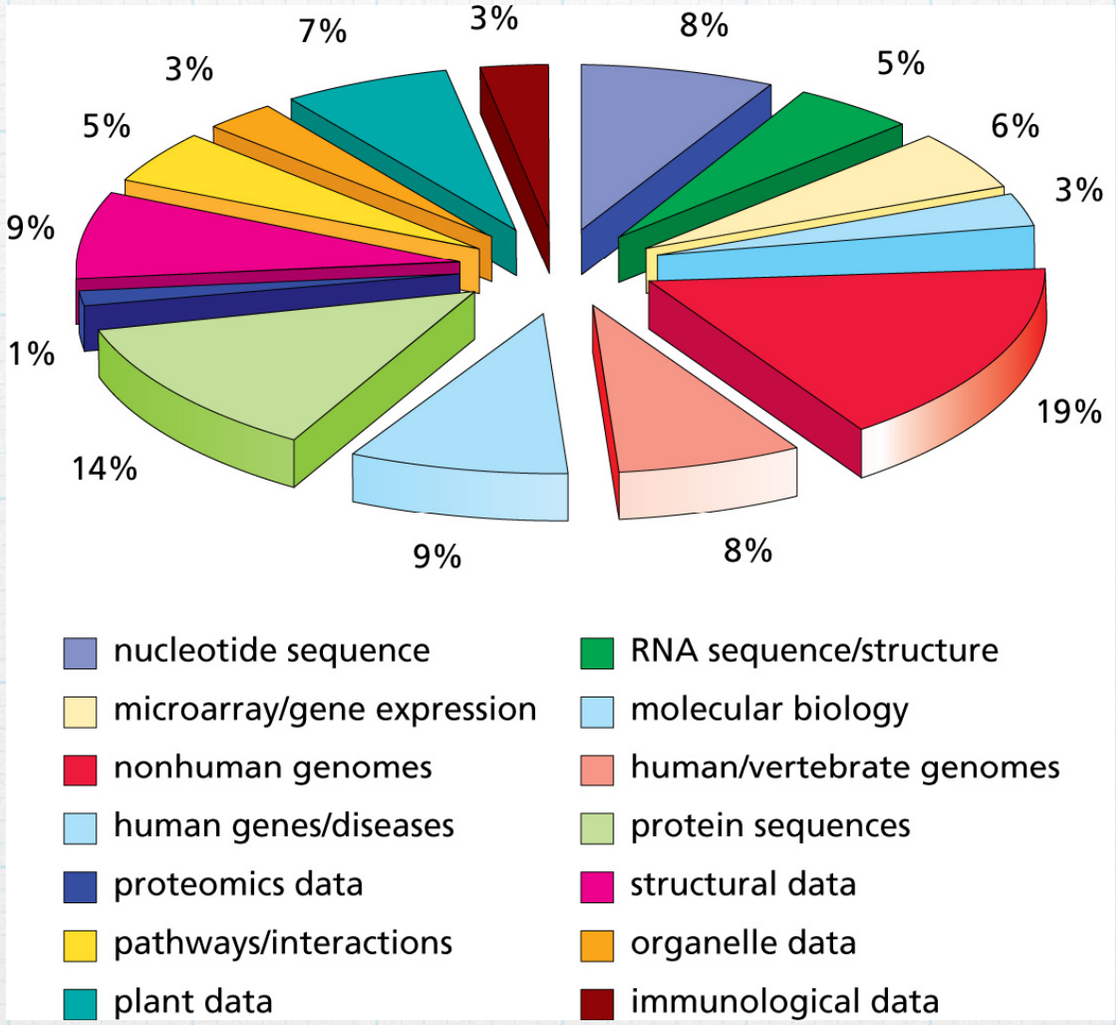
Organization of the DNA

* The **genome** is an organism's complete set of DNA
* Genomes vary widely in size
    * some bacteria have 600,000 base pairs
    * humans have about 3 billion base pairs
    * Except for mature red blood cells, all human cells contain a complete genome
* DNA in the human genome is arranged into 23 pairs of DNA molecules, called **chromosomes** (physically separate molecules, and vary widely in length)
* Each chromosome contains many **genes**

Gene, Locus, Allele

* A **gene** is a unit of heredity, and usually refers to a DNA sequence that encodes a protein or an RNA that has some function
* A **locus** is the specific location of a gene (or, more generally, a DNA sequence) in the genome
* Each of the different DNA sequences at a given locus is called an **allele**

Available Data



858 Databases in total

(as classified at the NAR Molecular Biology Database Collection Website, 2006)

Acknowledgments

* Understanding Bioinformatics, Zvelebil and Baum, Garland Science