

Bioinformatics: Sequence Analysis

COMP 571 - Fall 2010
Luay Nakhleh, Rice University

Course Information

- * Instructor: Luay Nakhleh (nakhleh@cs.rice.edu); office hours by appointment (office: DH 3119)
- * TA: Natalie Yudin (natalieyudin@rice.edu); office hours by appointment (office: DH 3111)
- * Meeting time and place: T&TH 2:30-3:50, KH 107
- * Website: <http://www.cs.rice.edu/~nakhleh/COMP571>

Grading

- * **Class participation: 20%**
- * **A set of homework assignments: 80%**

Course Textbooks

Highly recommended, but not required

- * Understanding Bioinformatics
 - * M. Zvelebil and J.O. Baum, Garland Science (2008)
- * Population Genetics
 - * M.B. Hamilton, Wiley-Blackwell (2009)
- * A list of other recommended books is available on the course website

Calendar

- * Aug 23 (M): first day of class
- * Sep 6 (M): Labor Day, no classes (we don't have a class anyway 😞)
- * Sep 21 (T): Instructor out of town (no class)
- * Oct 11-12 (M&T): Midterm recess (no classes)
- * Nov 18 (Th): Instructor out of town (no class)
- * Nov 25-26 (Th&F): Thanksgiving Recess (no classes)
- * Dec 3 (F): last day of class
- * Total number of class meetings: 26

Background

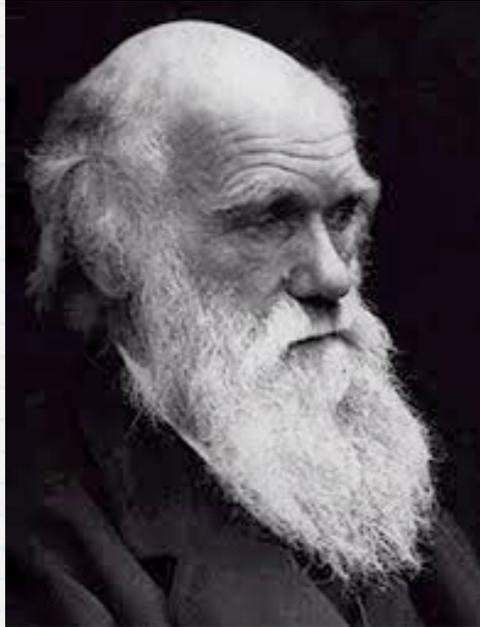
Life Through Evolution

- * 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 some organisms over others

* 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

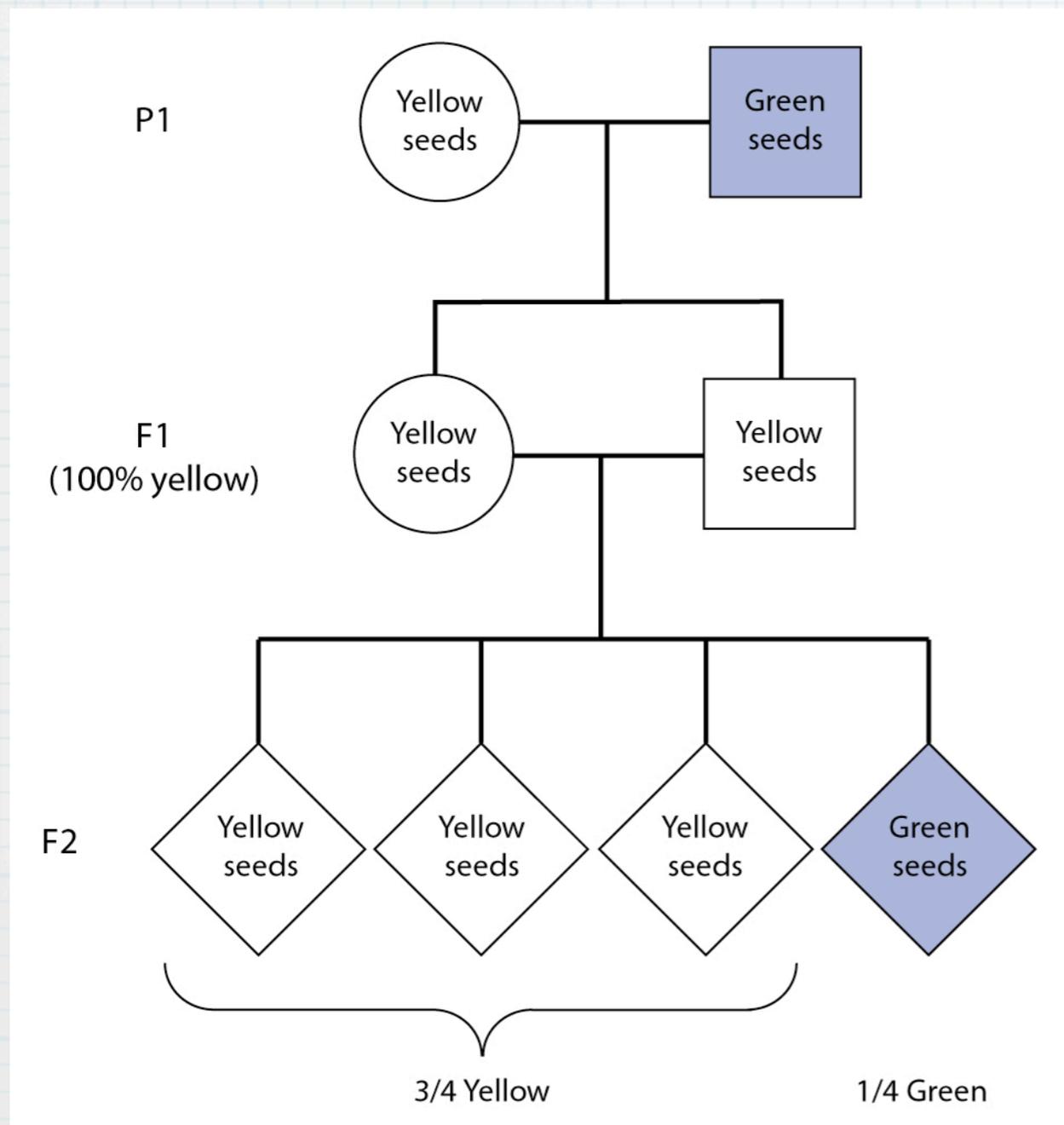
- * The accumulation of change over time in a population
- * **Population genetics** mainly focuses on evolutionary analysis of changes within populations, whereas **phylogenetics** is mostly aimed at inter-species relationships

Population Genetics

Mendel's Model of Particulate Genetics

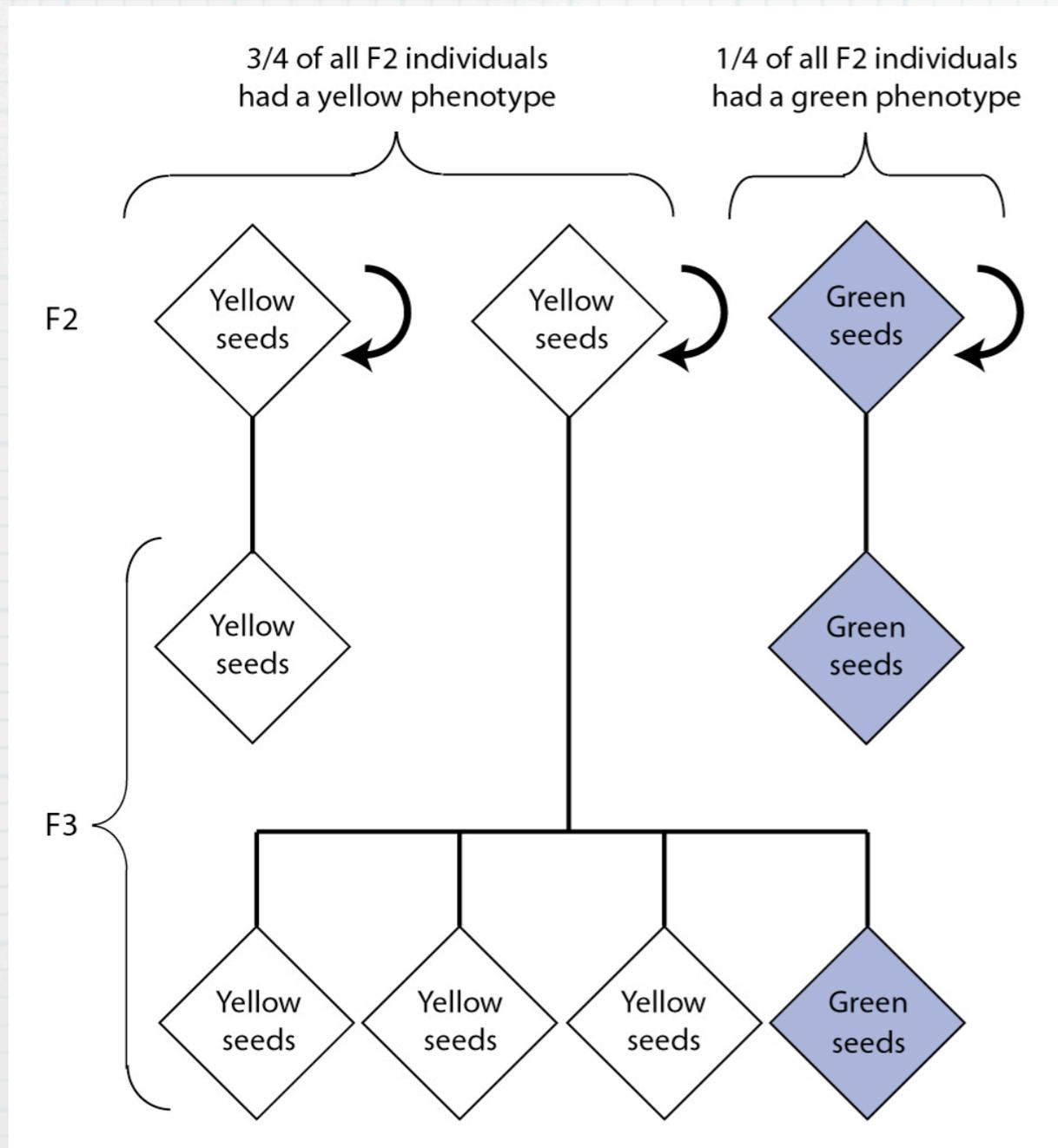
- * Mendel used experiments with pea plants to demonstrate independent assortment of both alleles within a locus and of multiple loci
- * Mendel used pea seed coat color as a phenotype, and his goal was to determine, if possible, the general rules governing inheritance of this phenotype

Mendel's Model of Particulate Genetics



Mendel's crosses to examine the segregation ratio in the seed coat color of pea plants. The parental plants (P1 generation) were pure breeding, meaning that if self-fertilized all resulting progeny had a phenotype identical to the parent. Some individuals are represented by diamonds since pea plants are hermaphrodites and can act as a mother, a father, or can self-fertilize.

Mendel's Model of Particulate Genetics

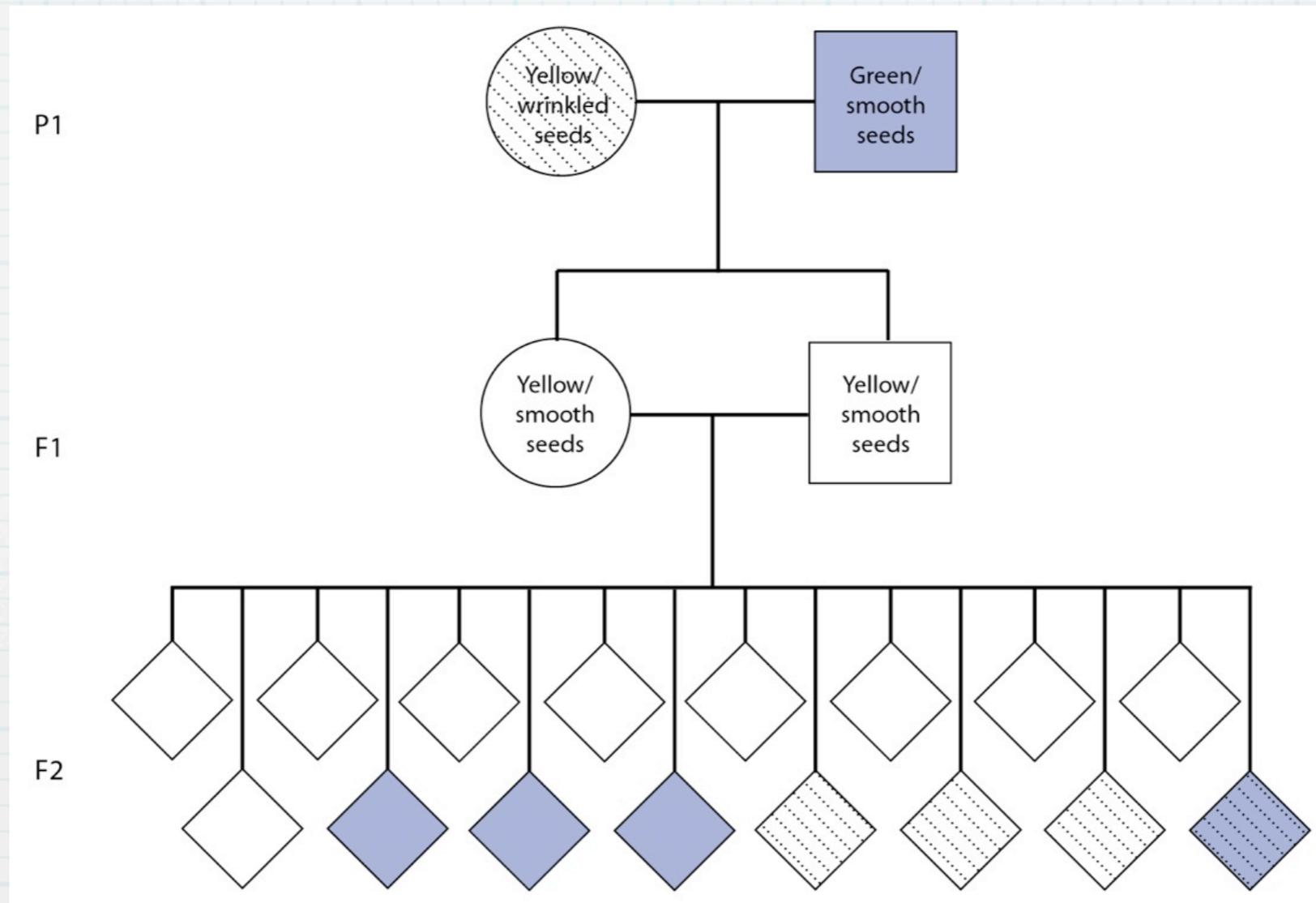


Mendel's self-pollinated (indicated by curved arrows) the F2 progeny produced by the cross shown in the figure on the previous slide. Of the F2 progeny that had a yellow phenotype (three-quarters of the total), one-third produced all progeny with a yellow phenotype and two-thirds produced progeny with a 3:1 ratio of yellow and green progeny.

Mendel's Model of Particulate Genetics

- * Mendel's first law predicts independent segregation of alleles at a single locus:
- * Two members of a gene pair (alleles) segregate separately into gametes so that half of the gametes carry one allele and the other half carry the other allele

Mendel's Model of Particulate Genetics



Mendel's crosses to examine the segregation ratios of two phenotypes, seed coat (yellow or green) and seed coat surface (smooth or wrinkled) in pea plants. The hatched pattern indicates wrinkled seeds while white indicates smooth seeds. The F2 individuals exhibited a phenotypic ratio of 9 round/yellow : 3 round/green : 3 wrinkled/yellow : 1 wrinkled/green

Mendel's Model of Particulate Genetics

- * Mendel's second law predicts independent assortment of multiple loci:
- * during gamete formation, the segregation of alleles of one gene is independent of the segregation of alleles of another gene

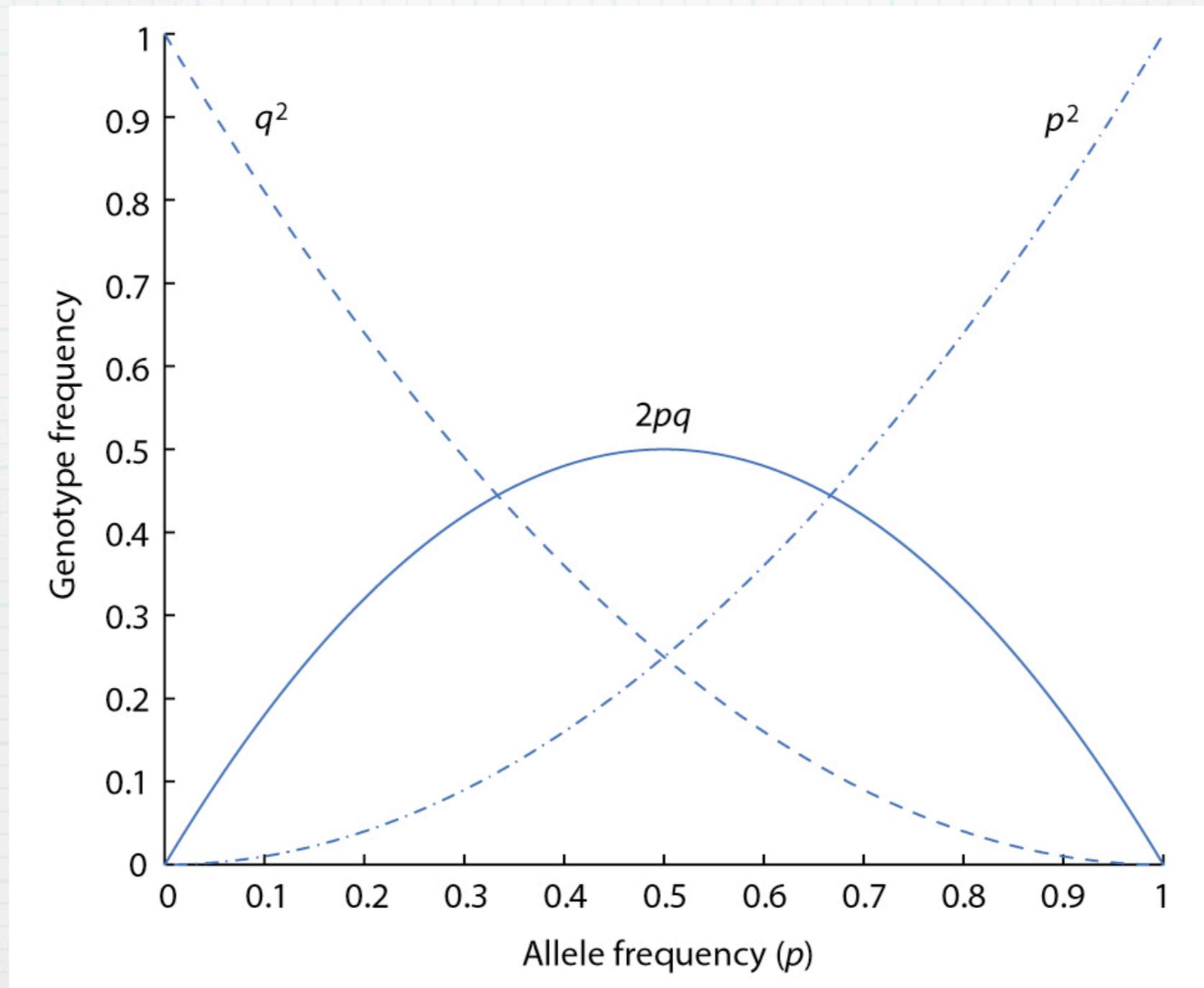
Hardy-Weinberg Expected Genotype Frequencies

- * In 1908, Hardy and Weinberg formulated the relationship that can be used to predict allele frequencies given genotype frequencies, or predict genotype frequencies given allele frequencies
- * This relationship is the well-known Hardy-Weinberg equation

$$p^2 + 2pq + q^2 = 1$$

where p and q are allele frequencies for a locus with two alleles

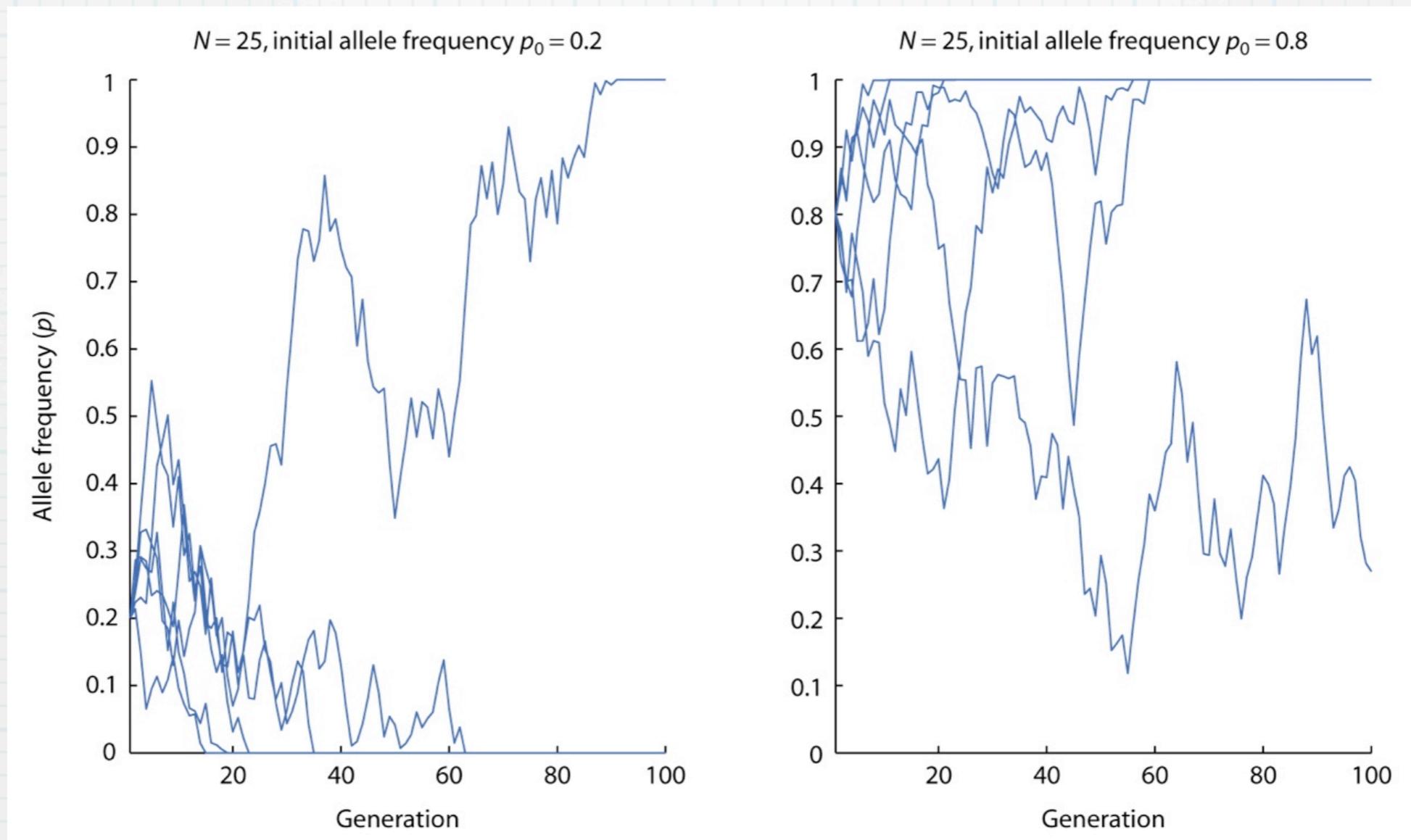
Hardy-Weinberg Expected Genotype Frequencies



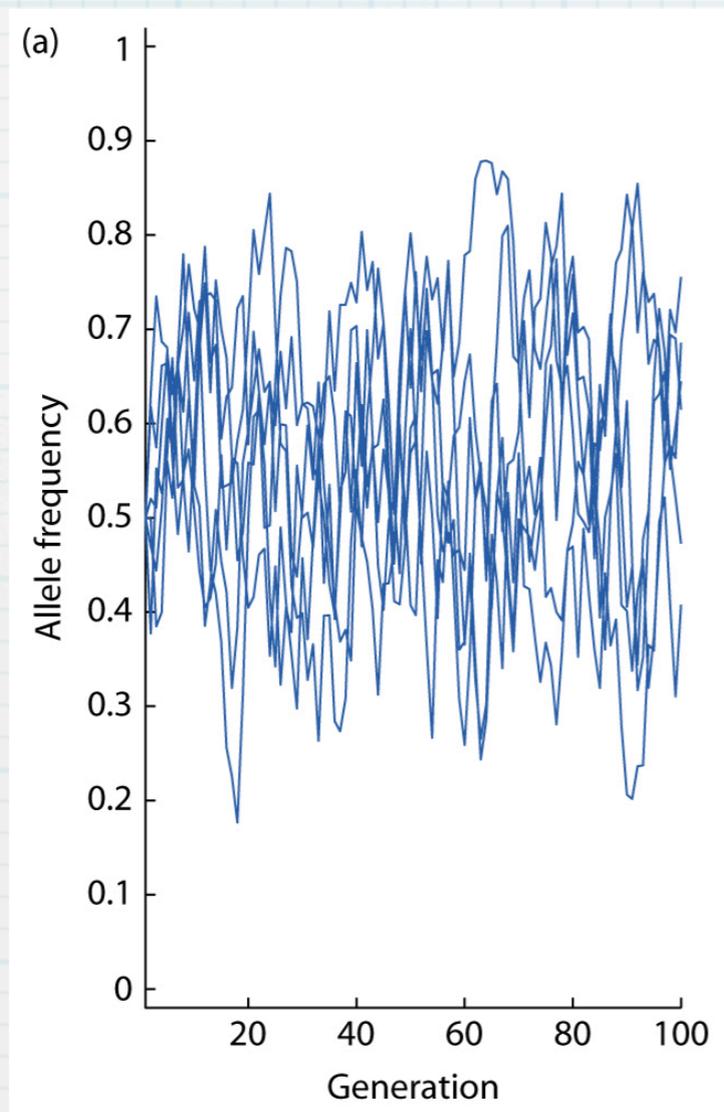
Hardy-Weinberg Expected Genotype Frequencies

- * A single generation of reproduction where a set of conditions, or assumptions, are met will result in a population that meets Hardy-Weinberg expected genotype frequencies, often called Hardy-Weinberg equilibrium (HWE)
- * The list includes: **infinite population size**, **no migration**, **no mutation**, **no selection**

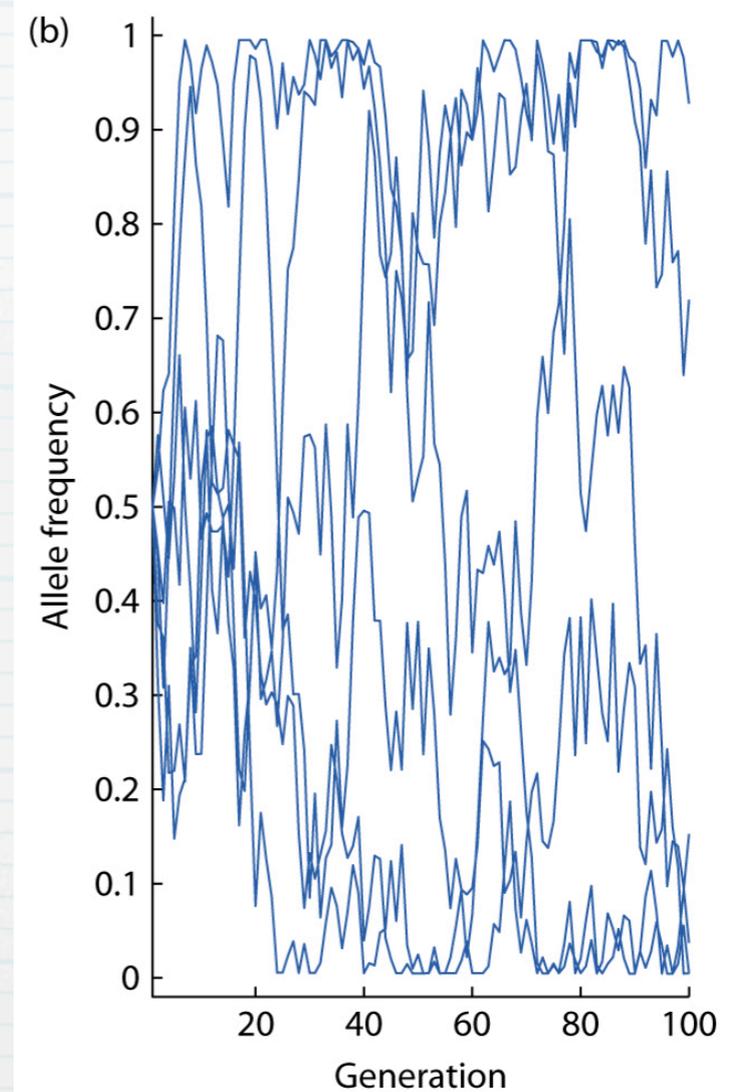
Deviation 1 from HWE: Finite Population Size



Deviation 2 from HWWE: Migration



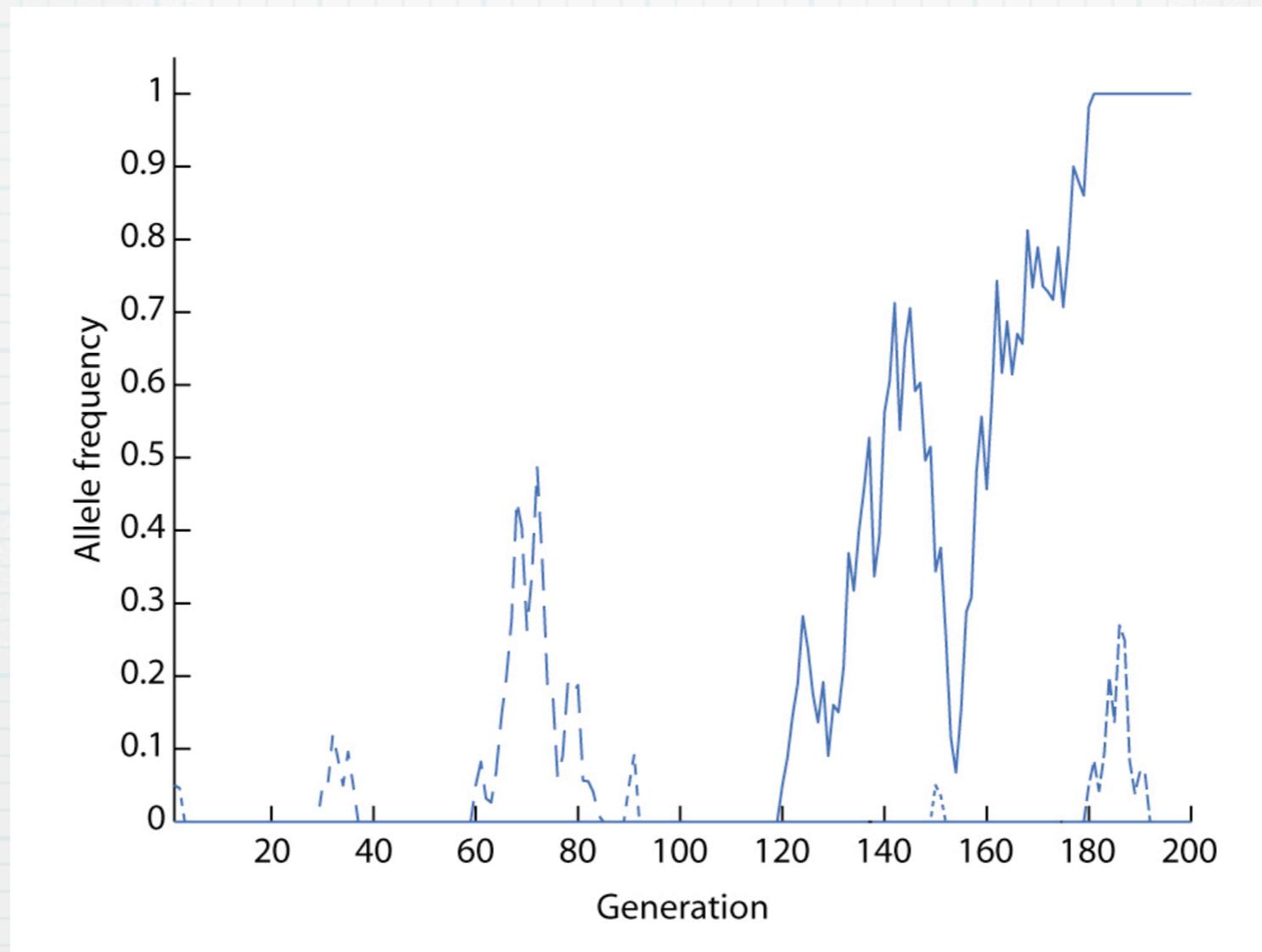
$m=0.2$



$m=0.01$

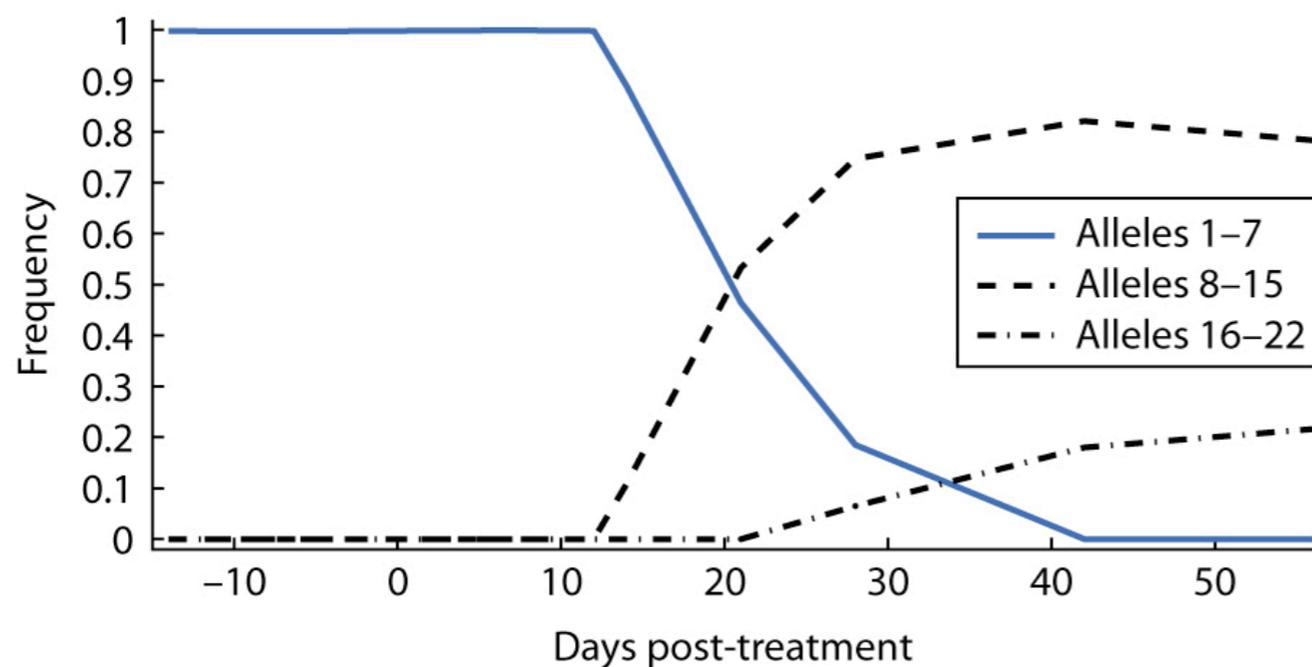
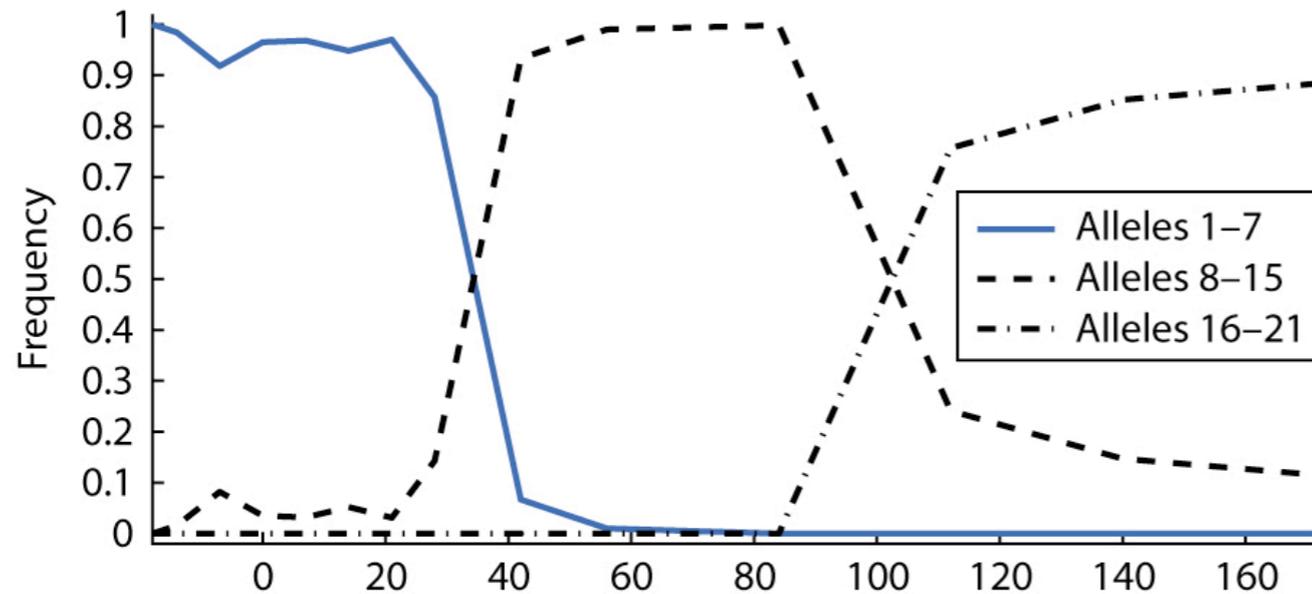
Allele frequencies for six randomly chosen subpopulations out of 200.
Each subpopulation contains 10 individuals.

Deviation 3 from HWWE: Mutation



One new mutation is introduced into the population every 30 generations,
and $N_e=10$.

Deviation 4 from HWE: Natural Selection



Allele frequencies at the protease locus over time in the HIV population in two patients undergoing protease inhibitor treatment.

Alleles found at very low frequencies before drug treatment come to predominate in the HIV population after drug treatment, due to natural selection among HIV genotypes for drug resistance.

Molecular Evolution

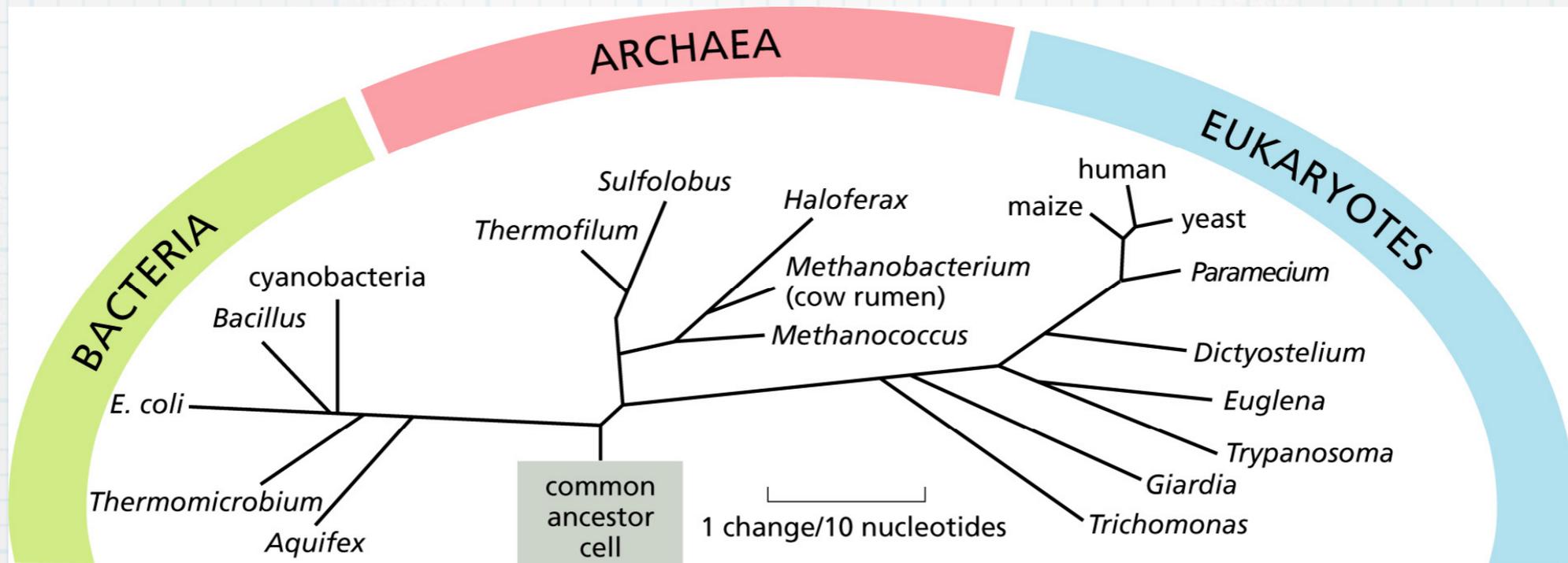
- * A primary focus of molecular evolution (or, molecular population genetics) is to make inference about the contribution of each of the aforementioned evolutionary forces (genetic drift, migration, mutation, and selection) to generate the patterns of molecular sequence variation we see today

Part I of the Course: Population Genetics

- * Genotype frequencies
- * Genetic drift
- * Population structure
- * Mutation
- * Natural selection
- * Molecular evolution

Phylogenetics

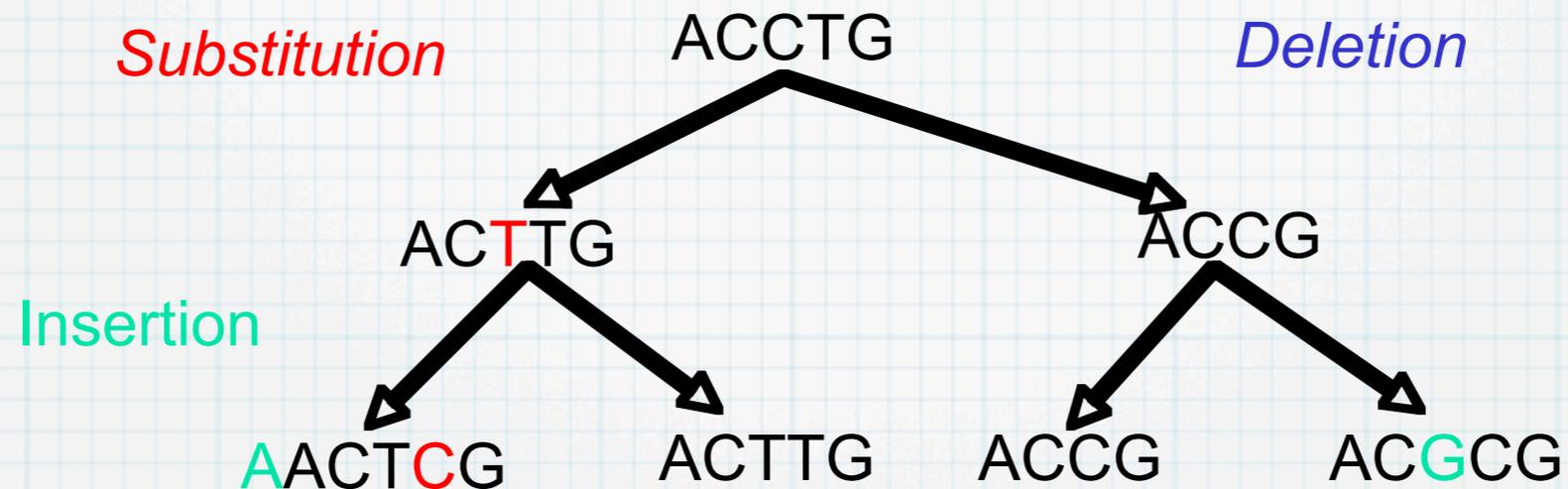
The Tree of Life



Sequence Variations Due to Mutations

- * Mutations and selection over millions of years can result in considerable divergence between present-day sequences derived from the same ancestral sequence.
- * The base pair composition of the sequences can change due to point mutation (substitutions), and the sequence lengths can vary due to insertions/deletions

Sequence Evolution



The observed sequences
(today's sequences)

```
AACTCG
ACTTG
ACCG
ACGCG
```

A major task in biology: reconstruct the evolutionary history of these sequences

This typically entails: (1) sequence alignment, and then (2) phylogeny reconstruction

Sequence Alignment

Alignment is the task of locating “equivalent” regions of two or more sequences to maximize their similarity

Mismatches

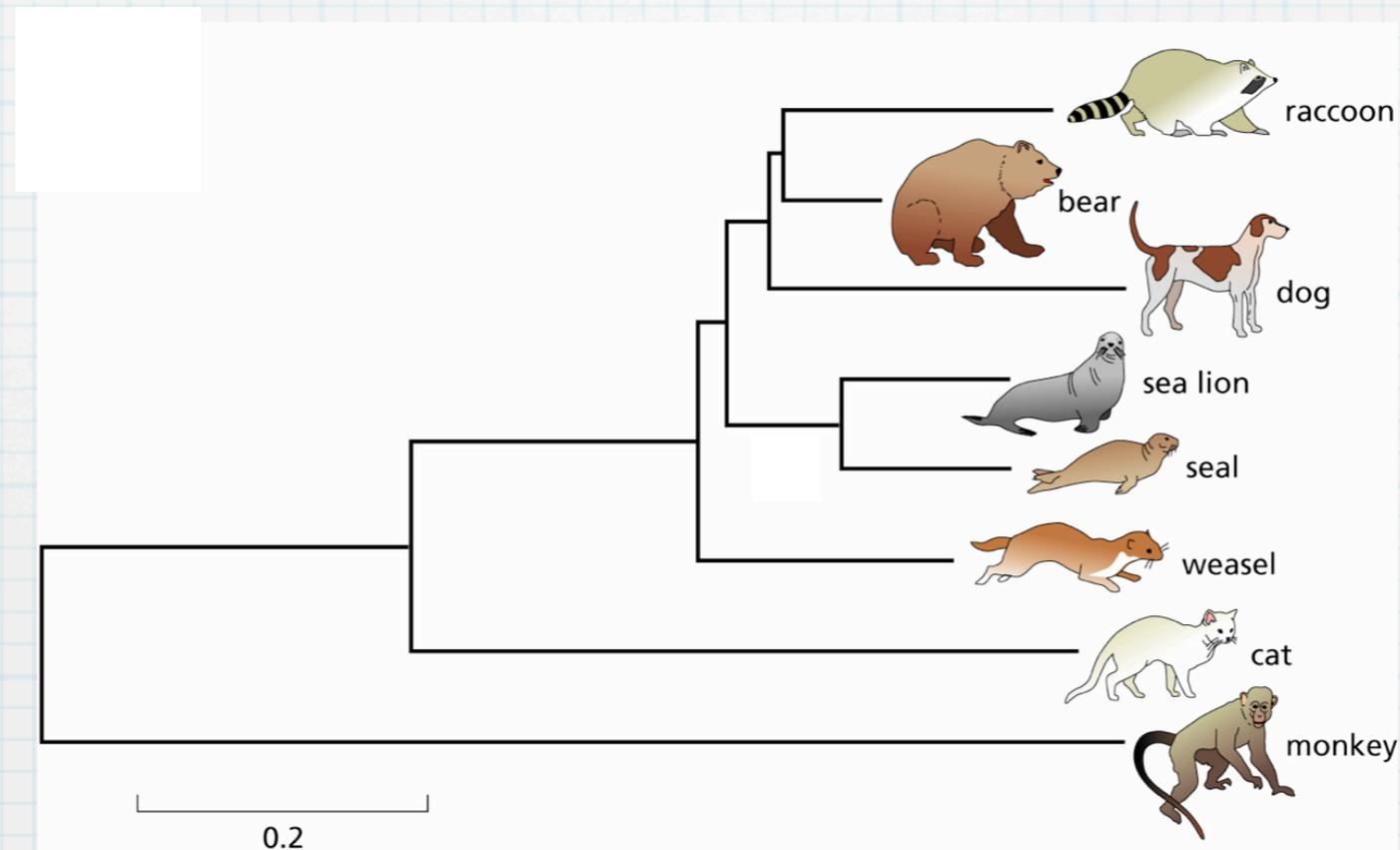
T	H	A	T	S	E	Q	U	E	N	C	E
T	H	I	S	S	E	Q	U	E	N	C	E

T	H	I	S	I	S	A	-	S	E	Q	U	E	N	C	E
T	H	-	-	-	-	A	T	S	E	Q	U	E	N	C	E

gap (indels: insertions/deletions)

Phylogeny Reconstruction

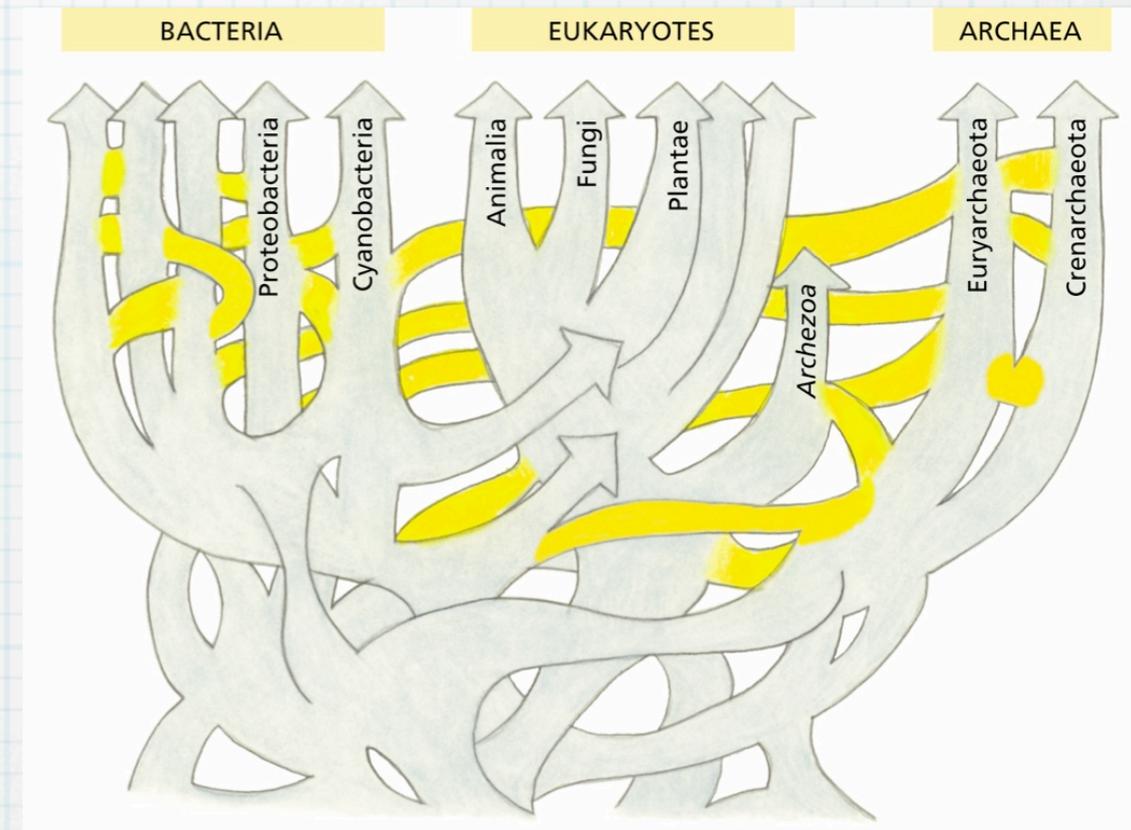
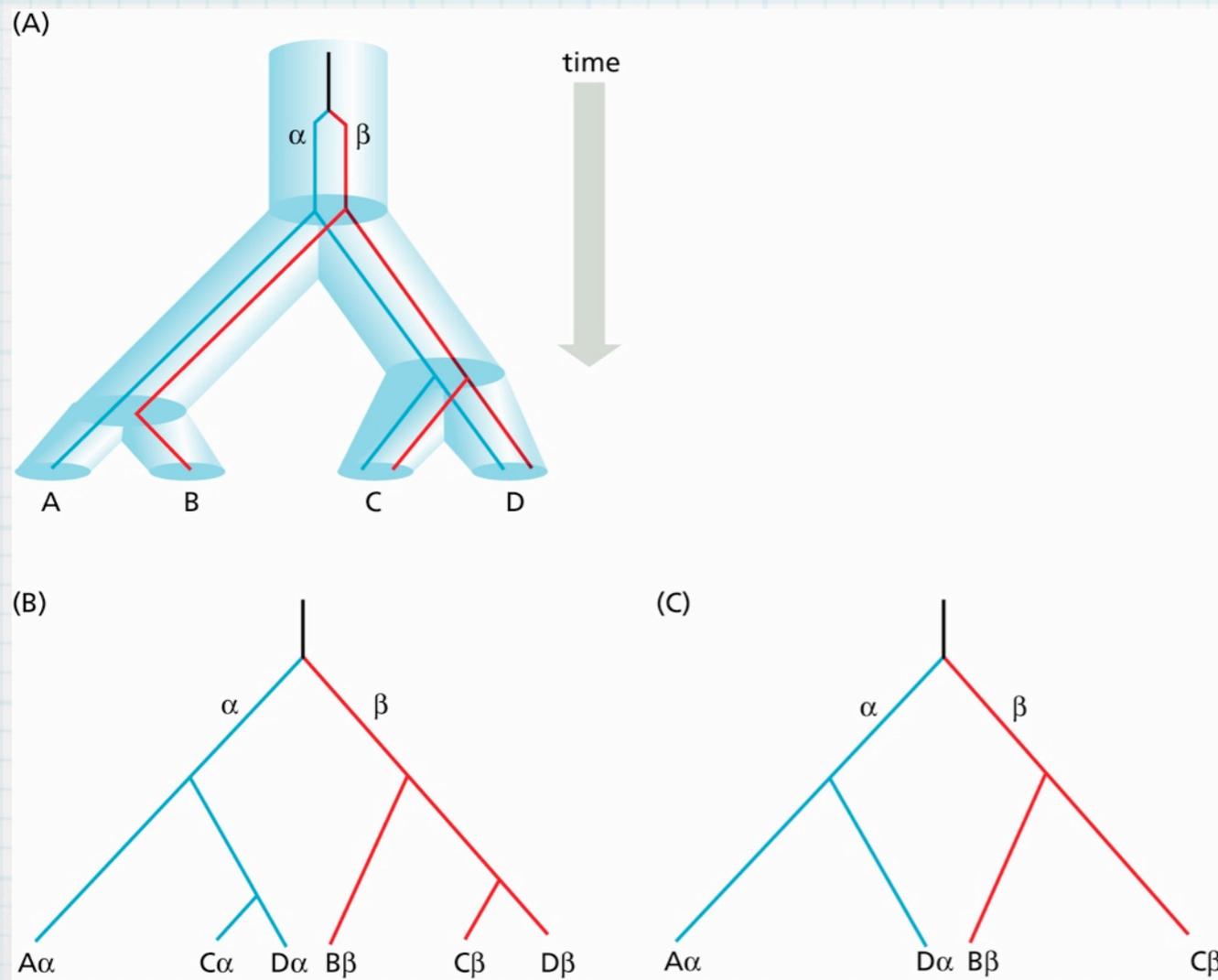
Phylogeny reconstruction is the task of inferring the evolutionary history of a set of taxa (species, genes, proteins, etc.)



The Genomic Era

- * Technologies today allow us to sequence whole genomes of organisms
- * Two significant tasks:
 - * Understanding the evolution of genomes (mutations at this level differ from those at the nucleotide level)
 - * Annotation of genomes (genes, regulatory elements, etc.)

Trees in Phylogenomics

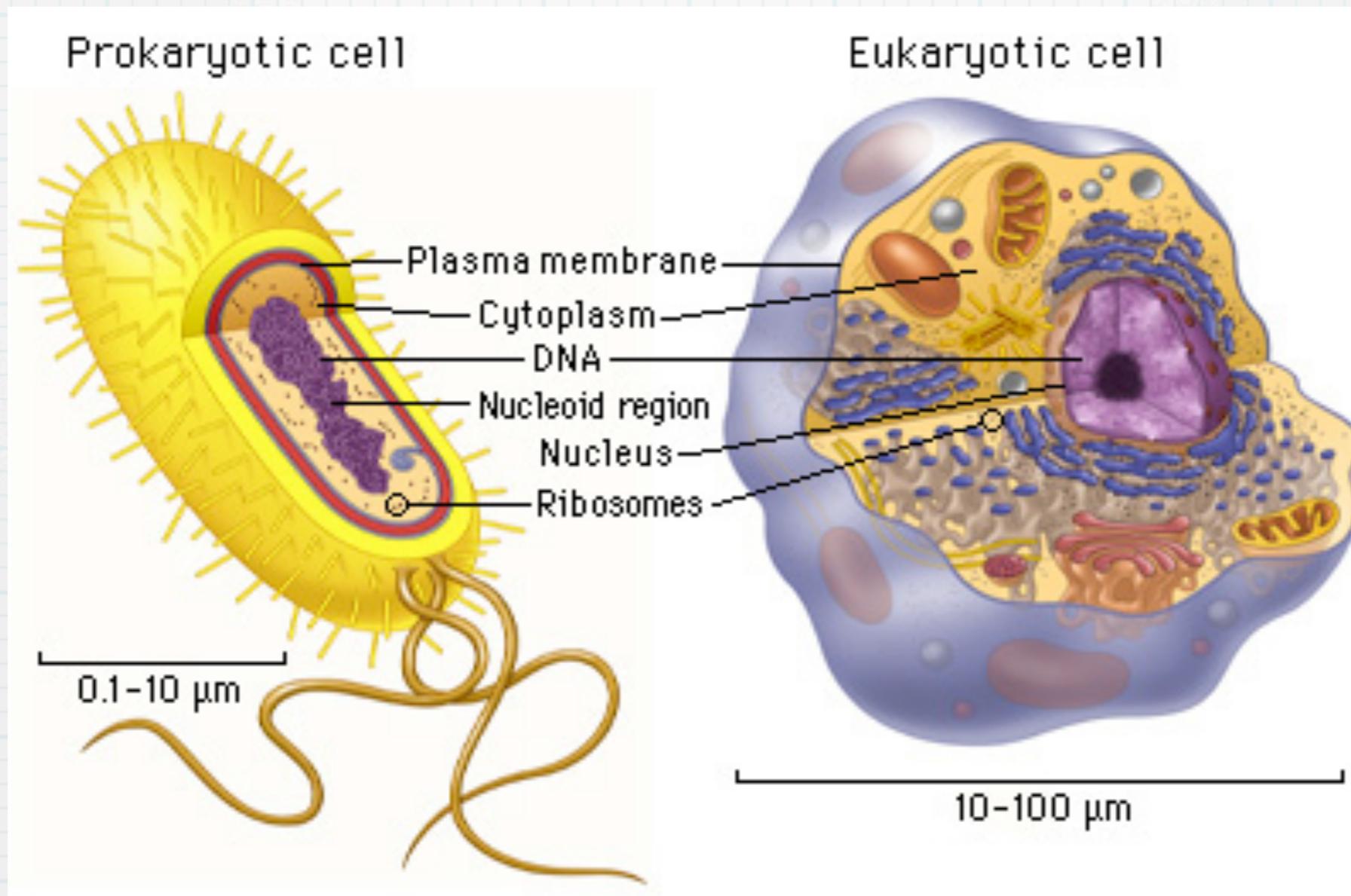


Part II of the Course: Phylogenetics

- * Sequence alignment
- * Database search through efficient pairwise alignment heuristics
- * Multiple sequence alignments
- * Phylogenetic trees
- * Phylogenomics (mainly, gene tree reconciliation)

A Little More Biology

Prokaryotic vs. Eukaryotic Cell Structure



Source: Pearson Education, Inc. The Biology Place

Prokaryotic vs. Eukaryotic Cells

	Prokaryotes	Eukaryotes
Size	1-10 μ m in length	10-100 μ m 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.

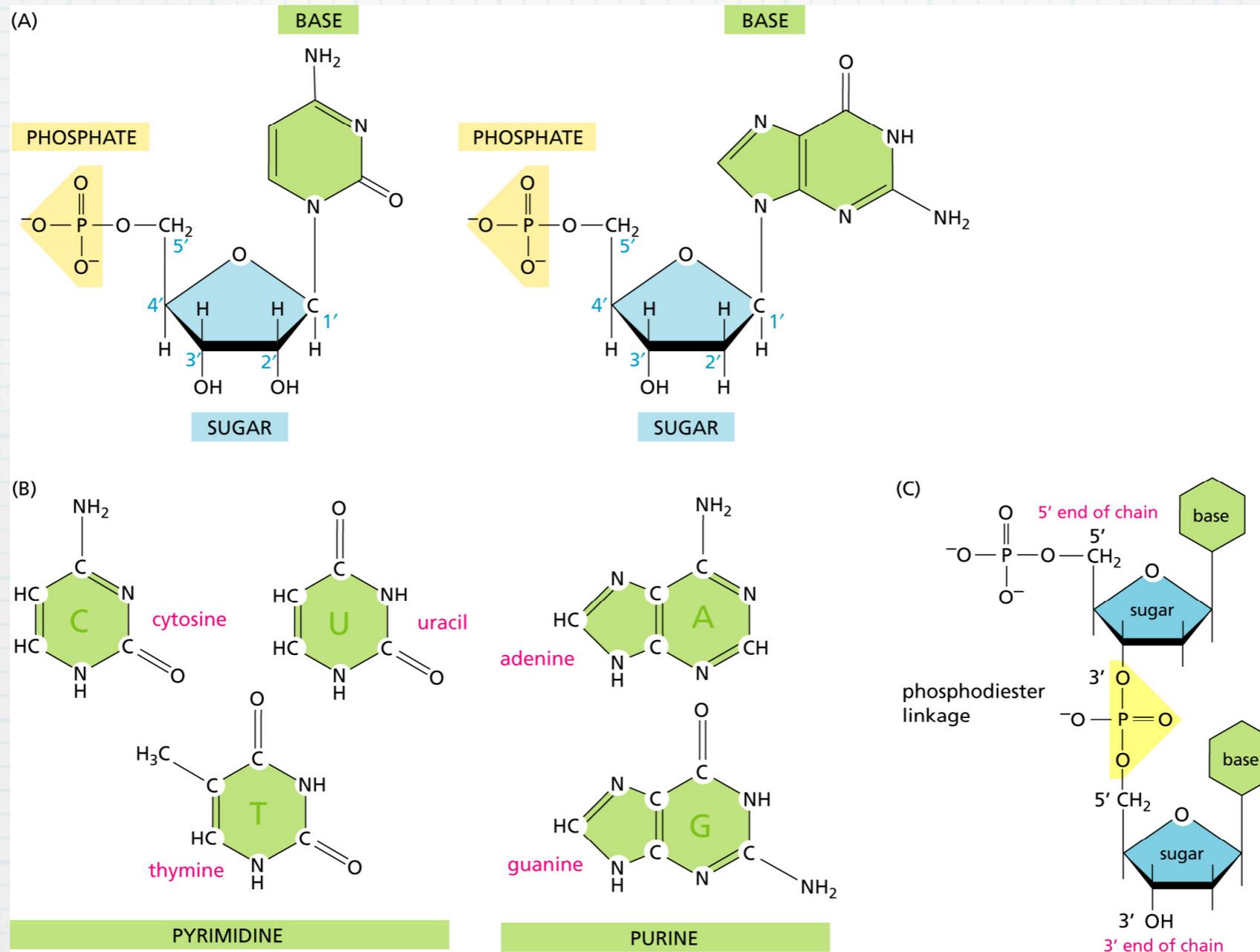
DNA

- * 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.

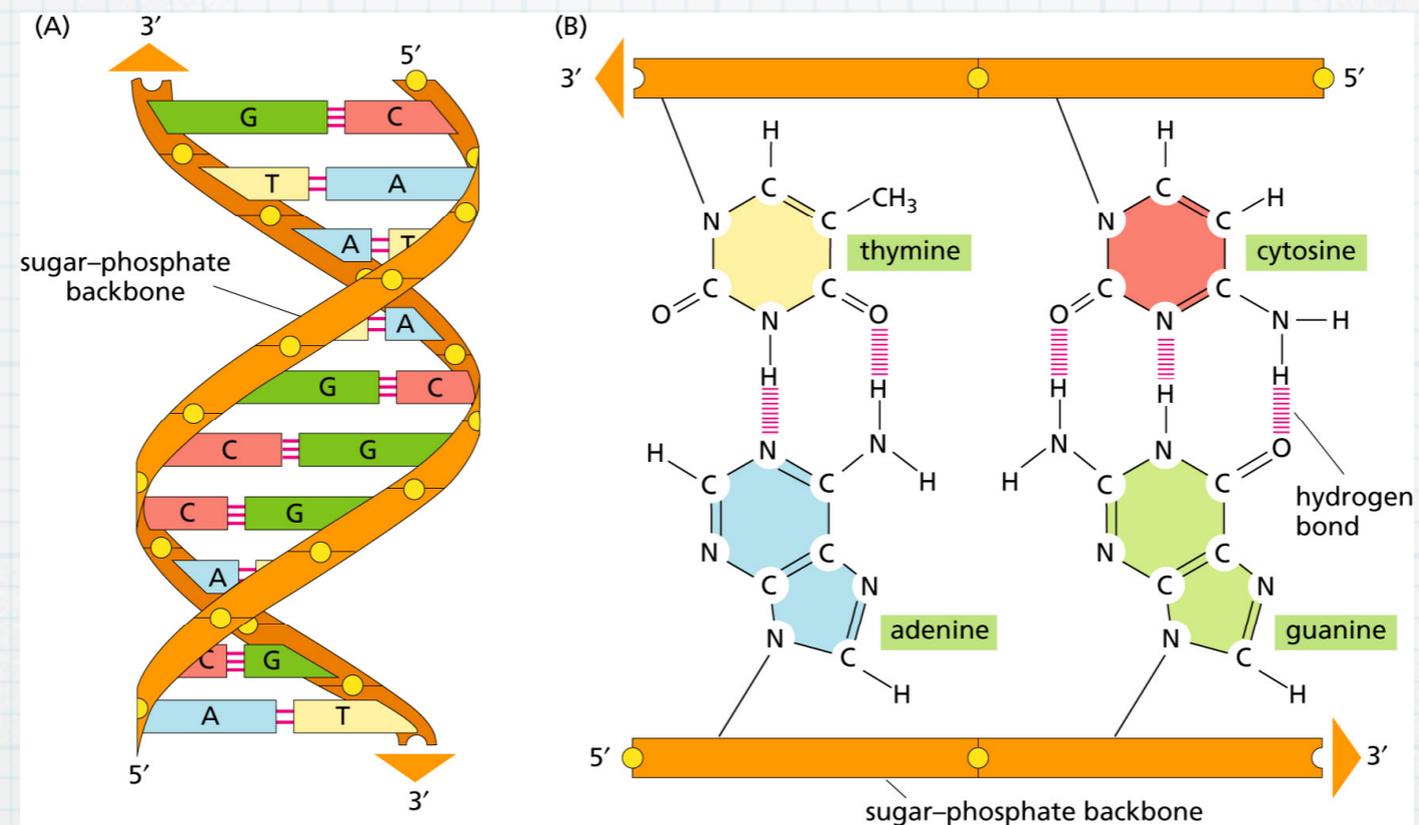
RNA

- * **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



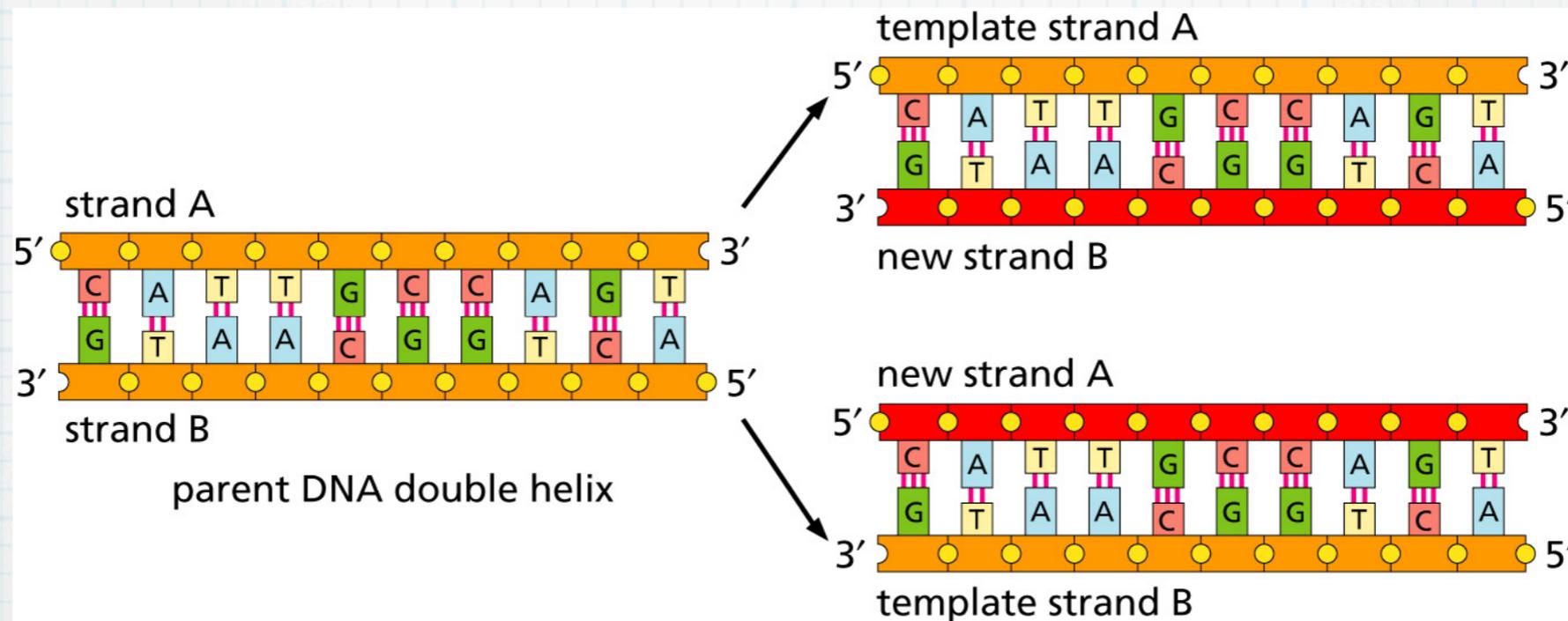
The Double Helix (DNA)



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

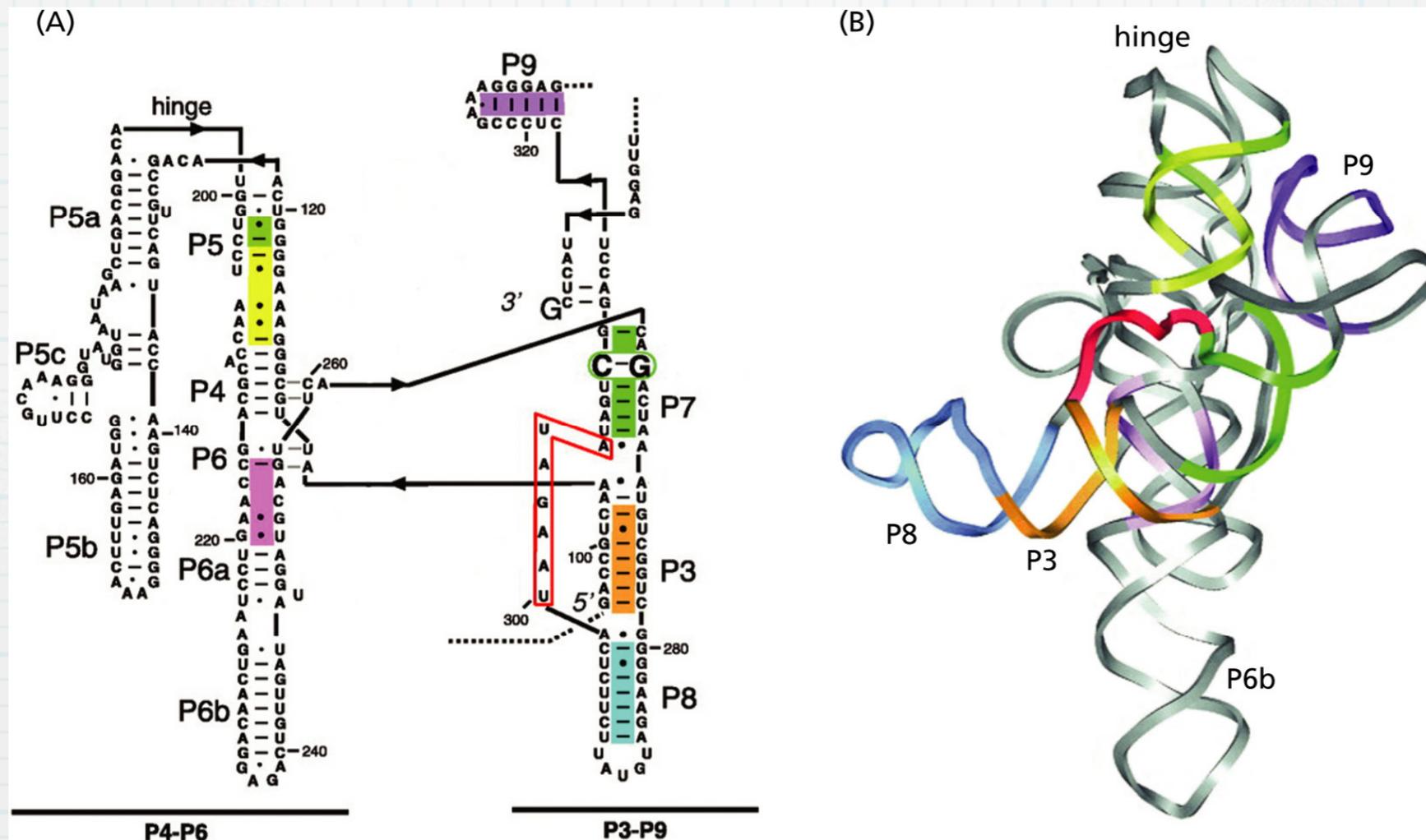


- * Hydrogen bonds are noncovalent bonds: the two DNA strands can be easily separated.
- * There are a number of processes in which strand separation is required.
- * One such process is DNA replication, which is a necessary prelude to cell division.

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.

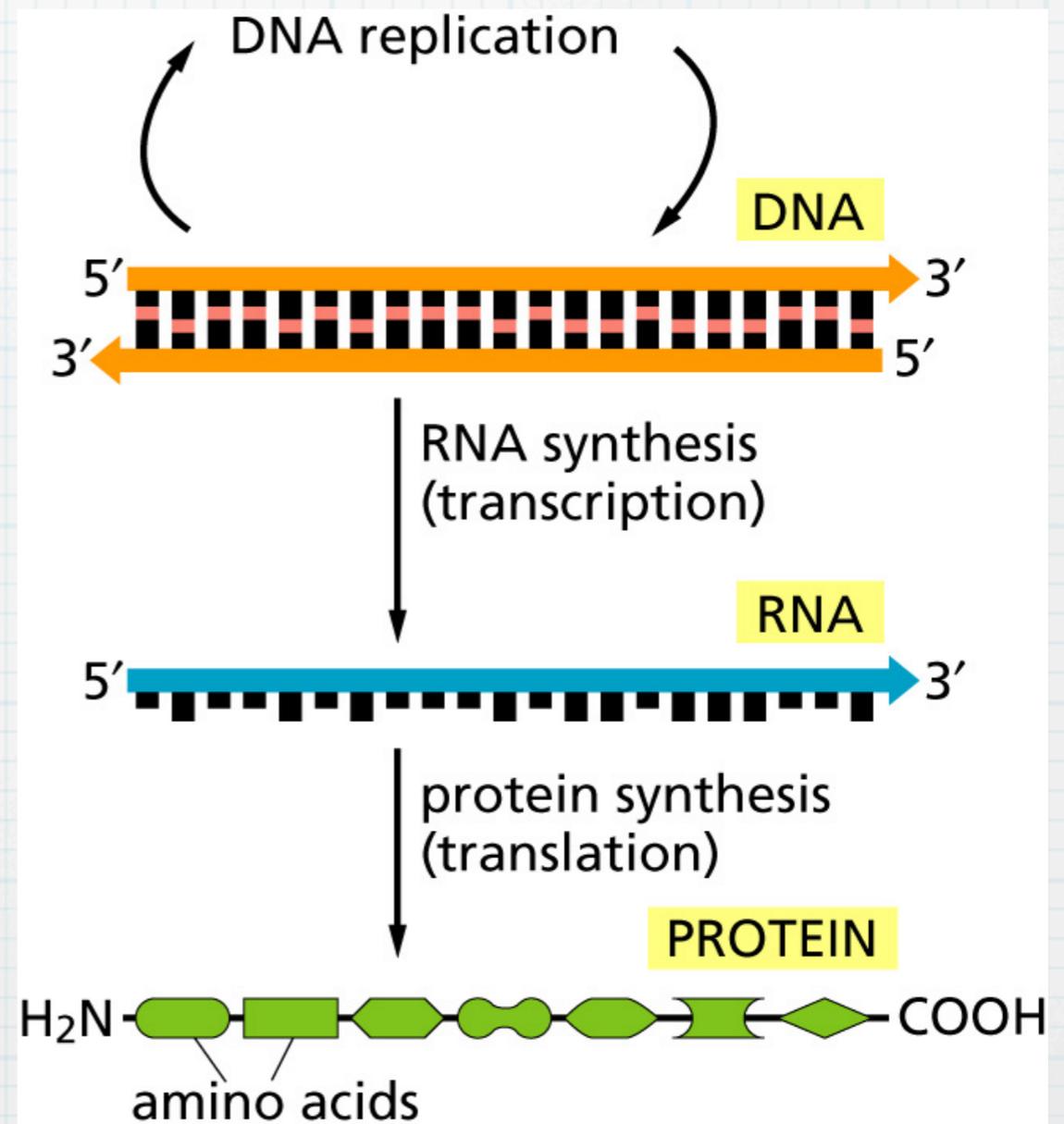
Secondary and Tertiary Structures of RNA



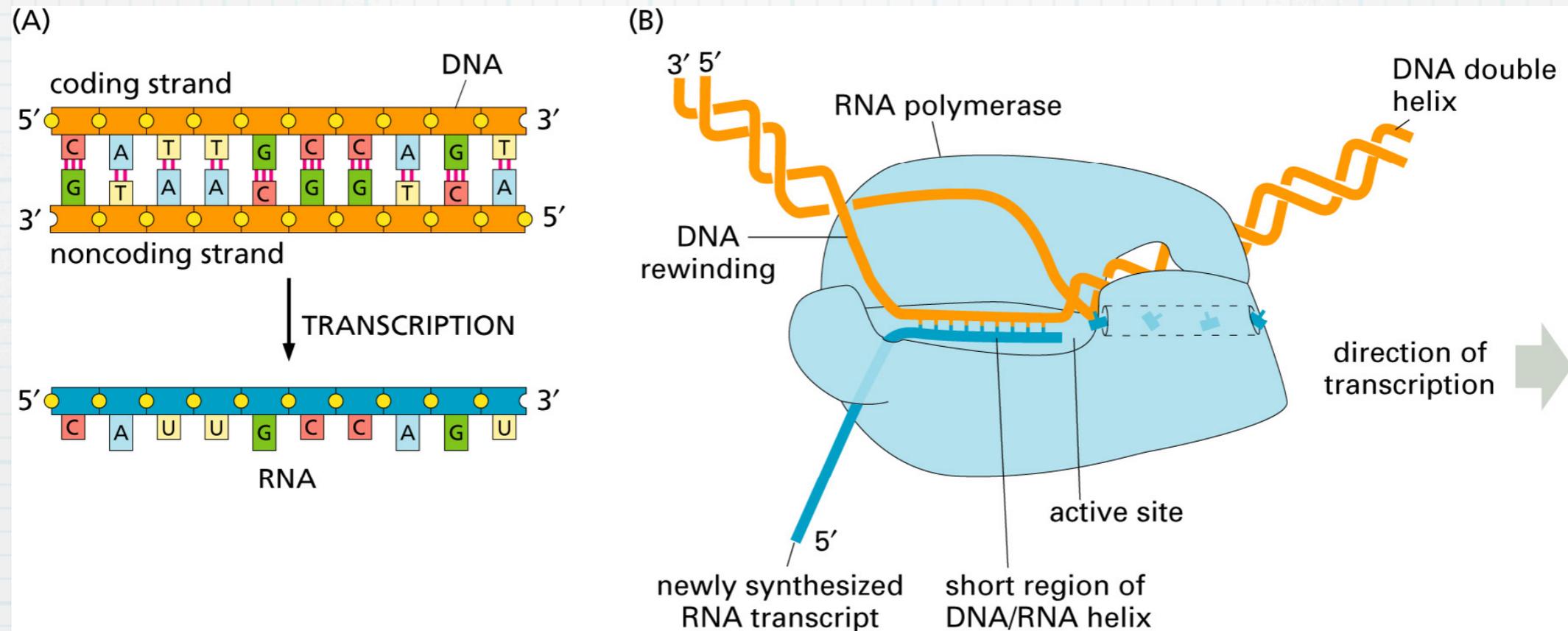
The Tetrahymena ribozyme

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



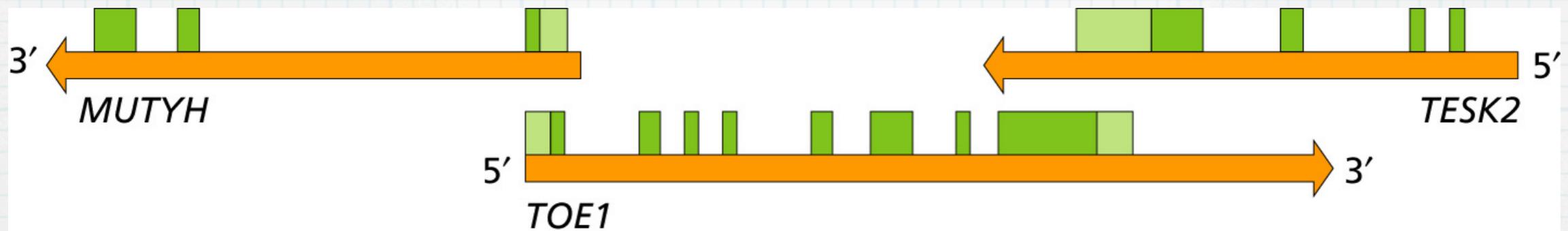
(A) One strand of the DNA is involved in the synthesis of an RNA strand complementary to the strand of the DNA

(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 than 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**)

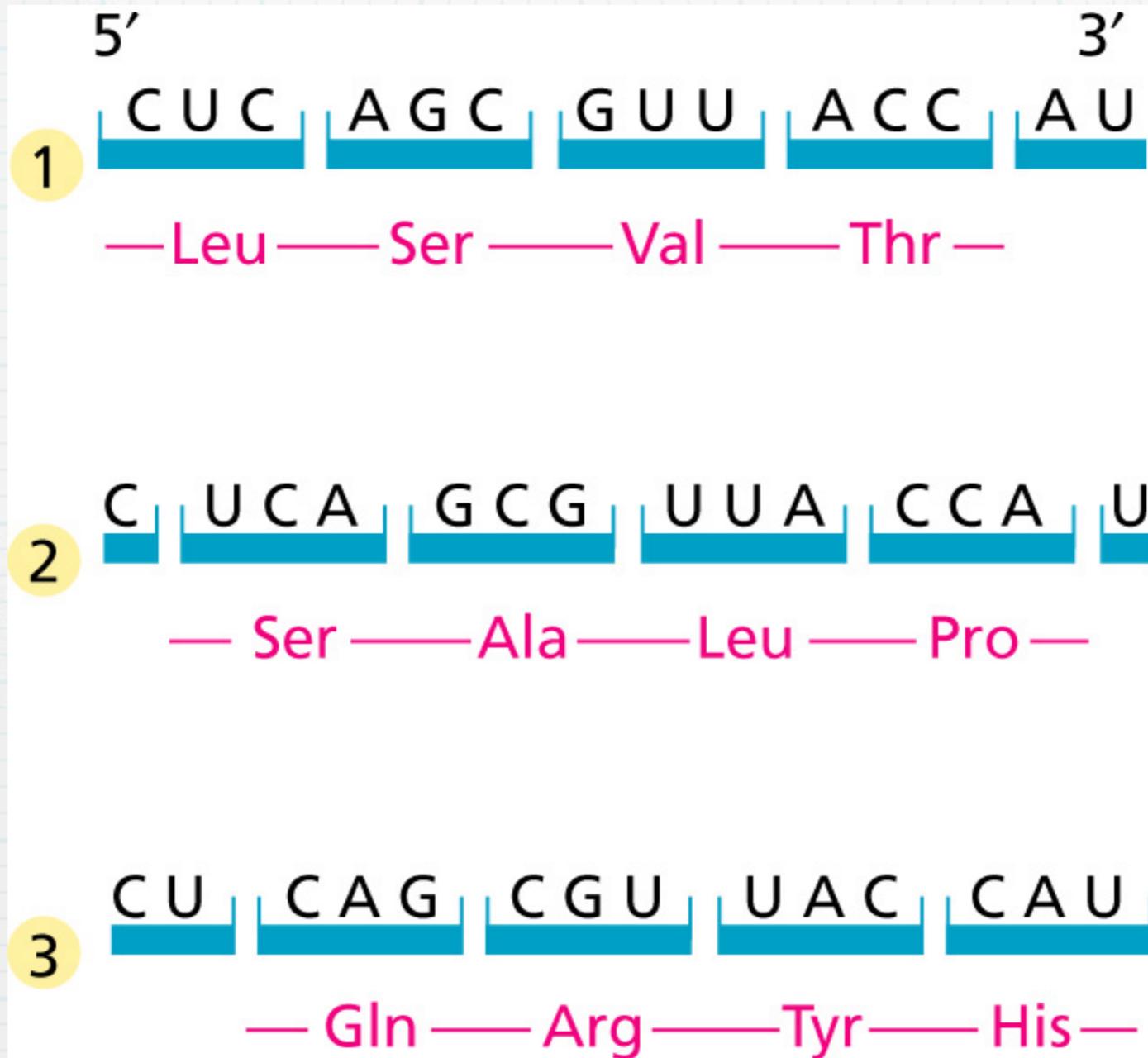
The Standard Genetic Code

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

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



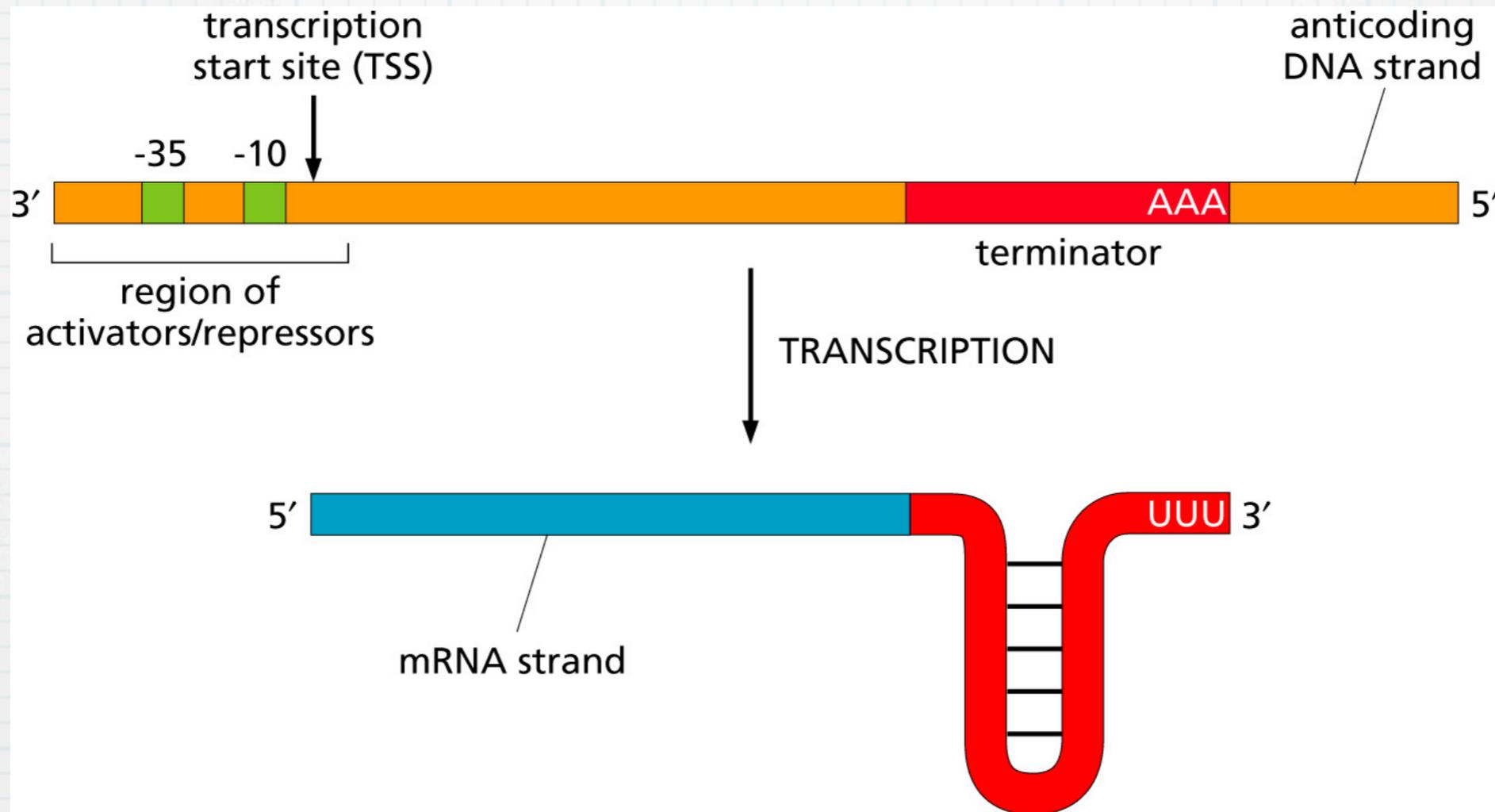
Gene Structure and Control

- * 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

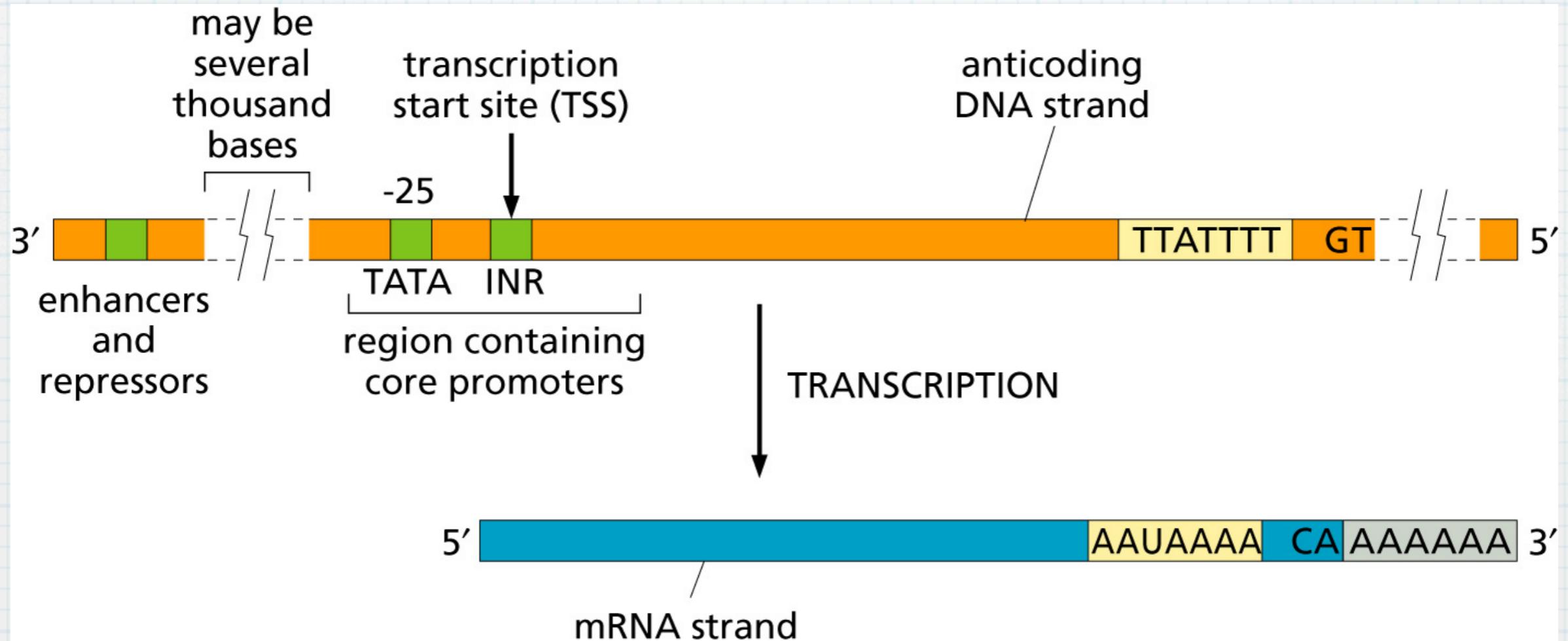


* Bacterial promoters typically occur immediately before the position of the **transcription start site (TSS)**, and contain two characteristic short sequences, or motifs, that are almost the same in the promoters for different genes.

* 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.

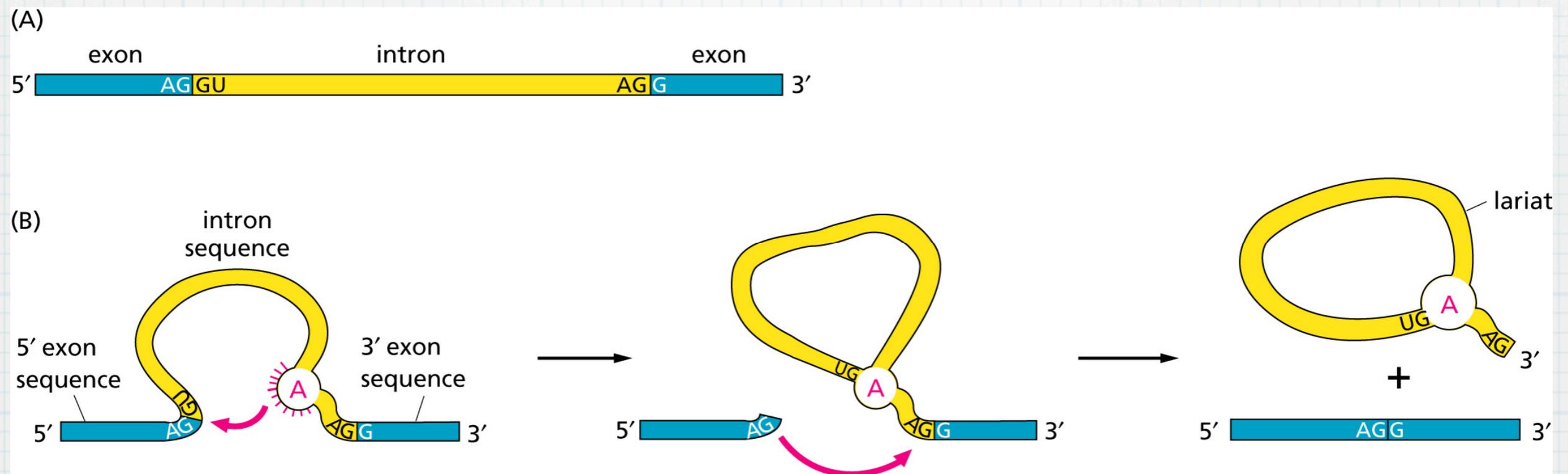
Gene Structure in Eukaryotes



* Regulatory elements in eukaryotes are more complex.

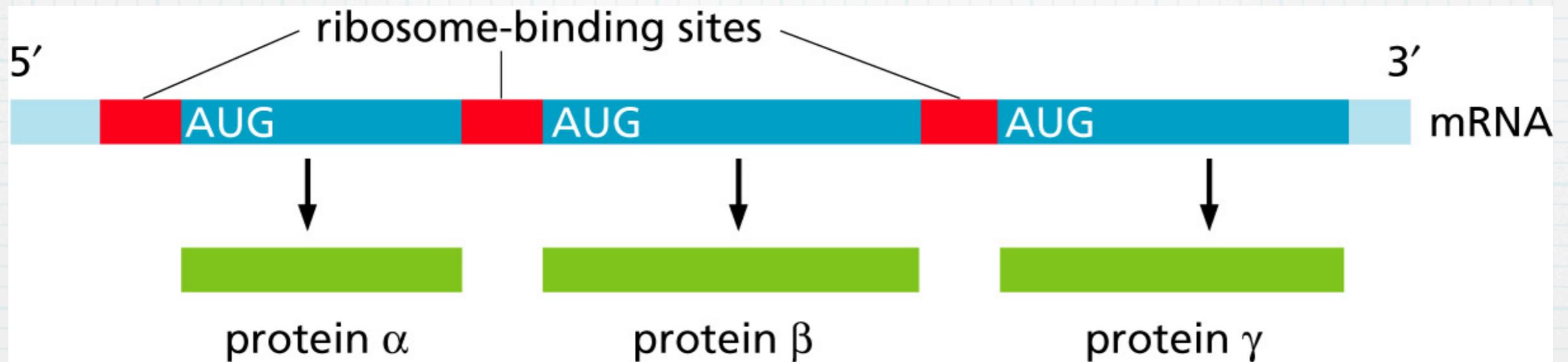
* Three types of RNA polymerase transcribe genes: RNA polymerase II transcribes all protein coding genes, where other RNA polymerase types transcribe genes for tRNAs, rRNAs and other types of RNA

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

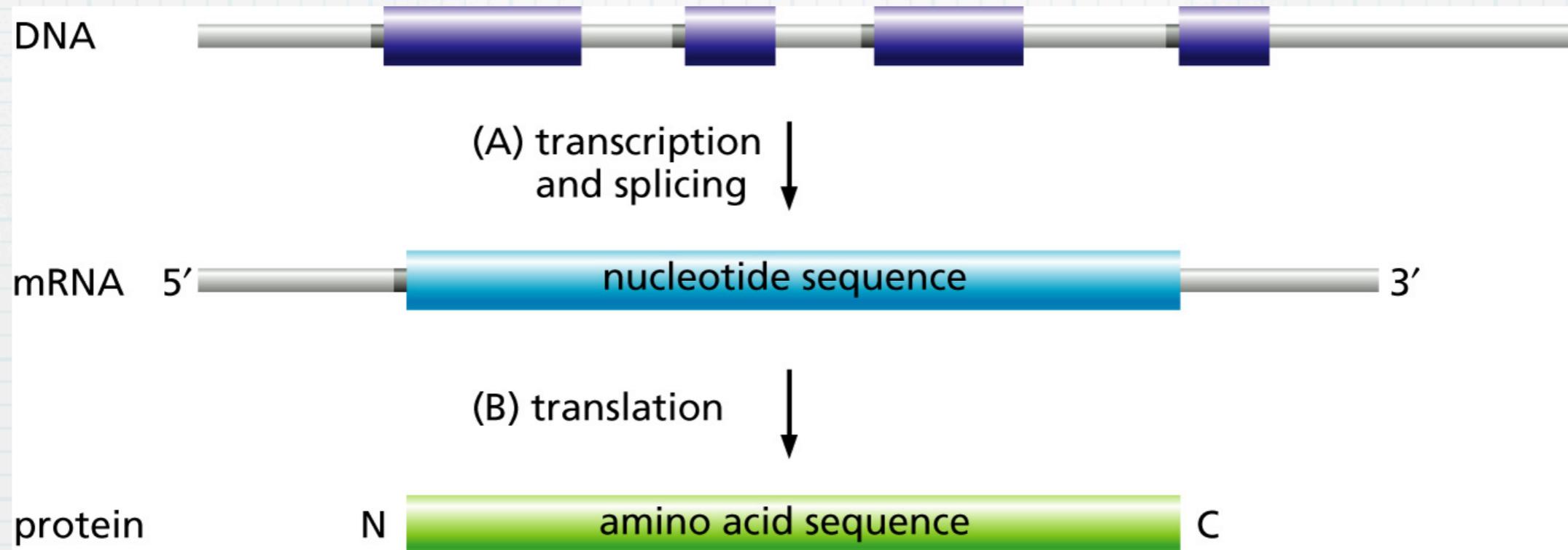


In bacteria, functionally related protein-coding sequences are often clustered together into **operons**. Each operon is transcribed as a single mRNA transcript and the proteins are then separately translated from this one long molecule.

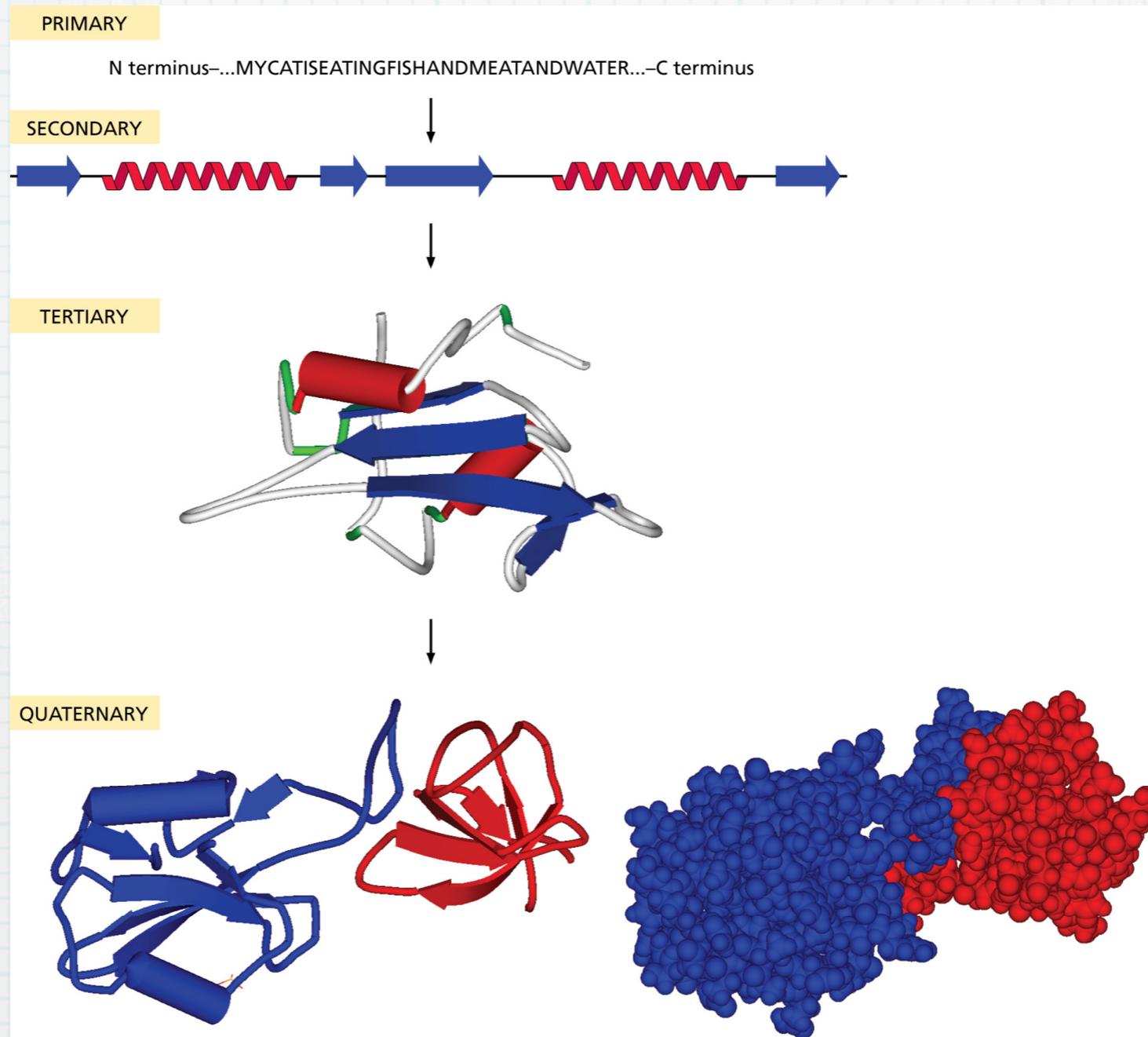
This has the advantage that only one control region is required to activate the simultaneous expression of all genes in the operon.

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

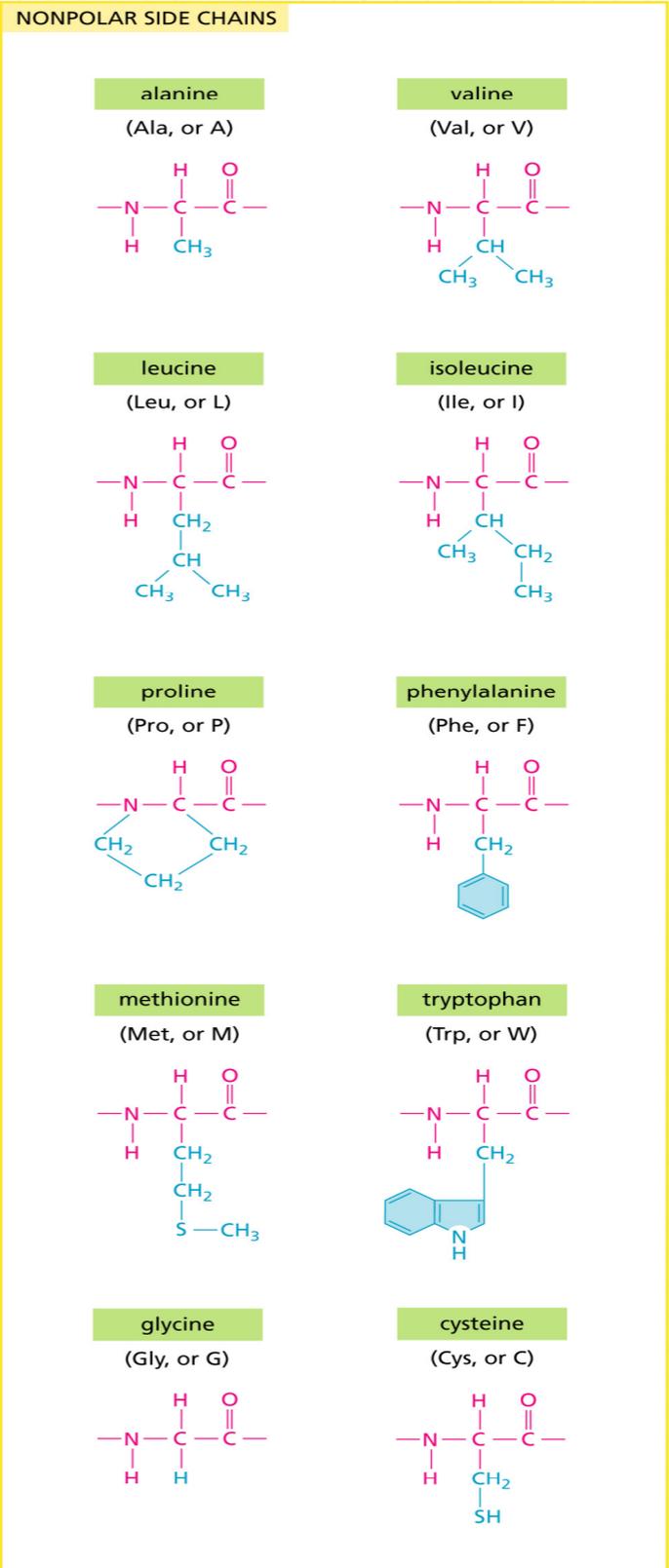
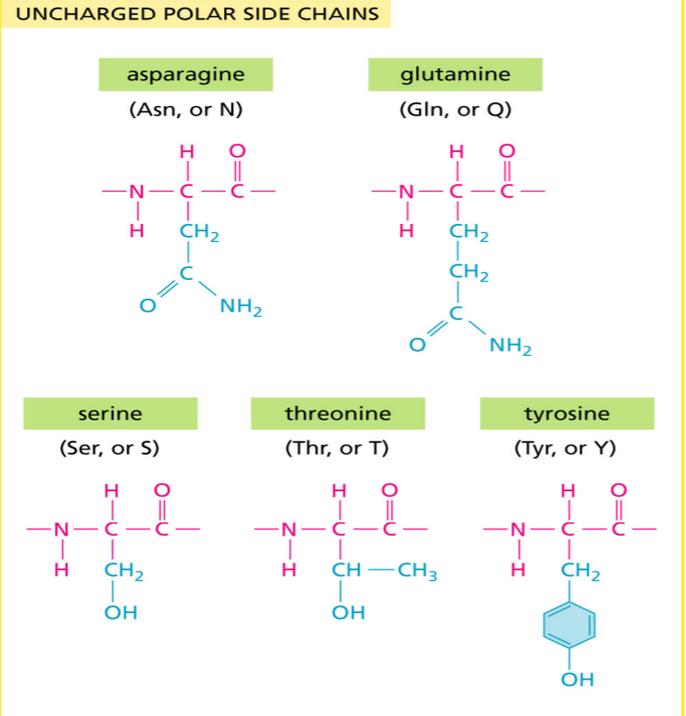
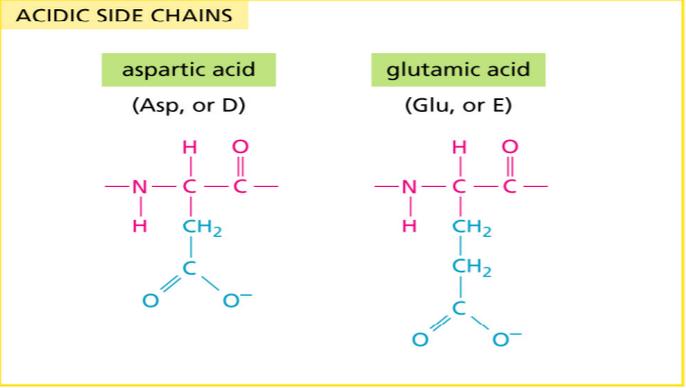
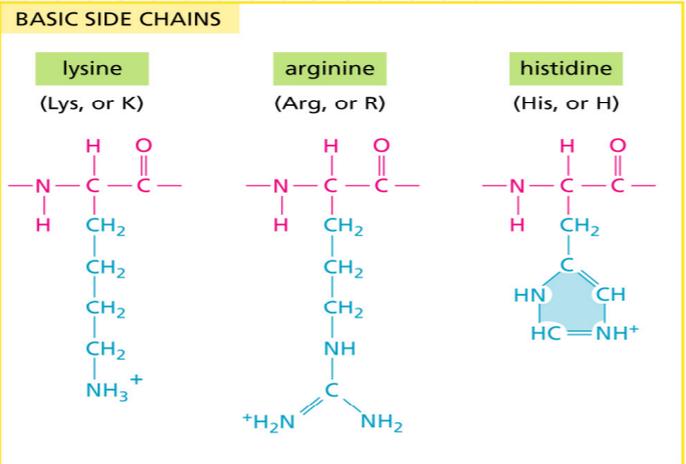
Proteins



Levels of Protein Structure



Side Chains of the Amino Acids



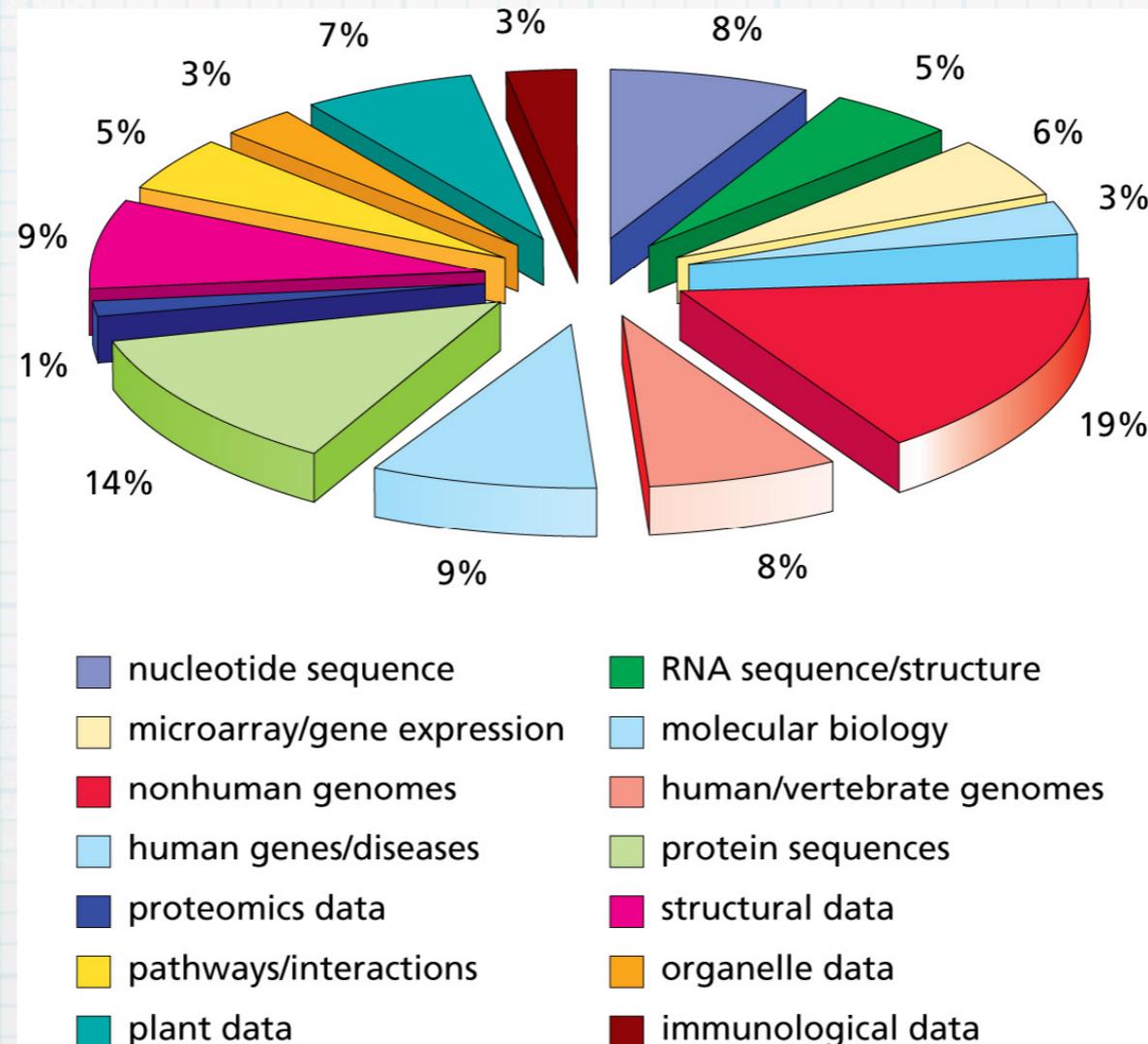
Organization of the DNA

- * 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)