IN4030
Introduction to bioinformatics

Week 1
Introduction to basic molecular biology and to comparison of sequences

Torbjørn Rognes
torognes@ifi.uio.no

Department of Informatics, UiO
13 January 2021

UiO : Universitetet i Oslo

Overview

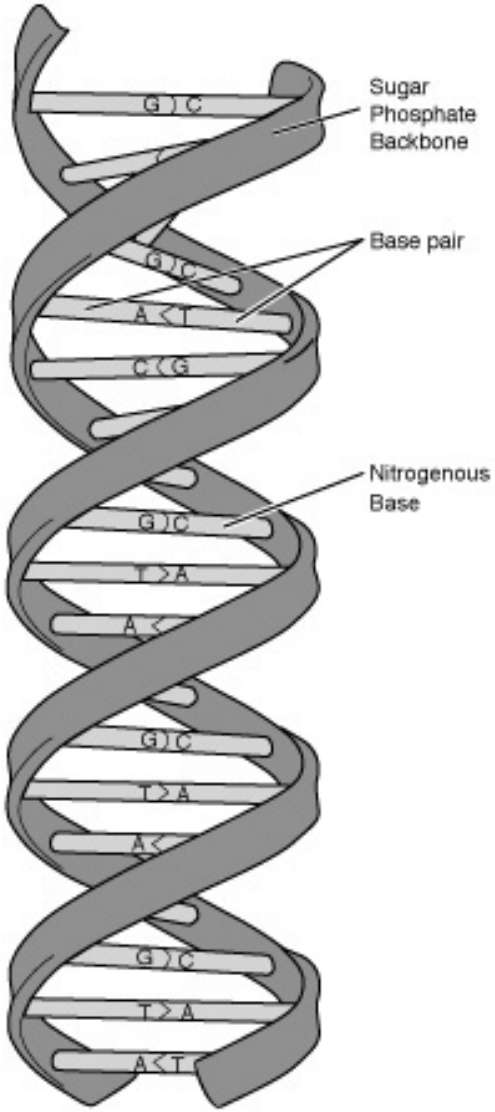
- DNA, genome, chromosome, gene
- The ”Central Dogma” (from DNA to protein)
- RNA
- Protein – amino acids

- A little more biology will be introduced later in the course as necessary

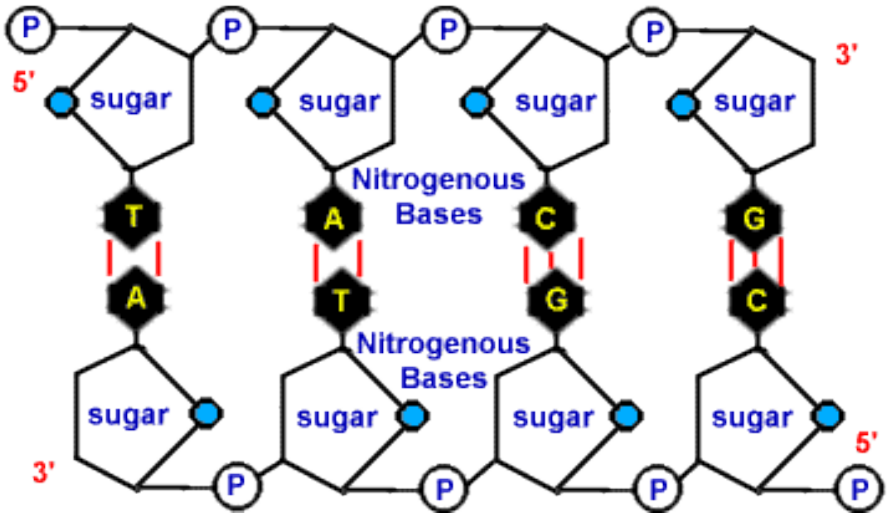
DNA

- DNA can store information
- DNA stores the instructions needed by the cell to live
- It consists of two strands which are interwoven together to form a double helix.
- Each strand is a chain of small molecules called nucleotides.

Double stranded DNA

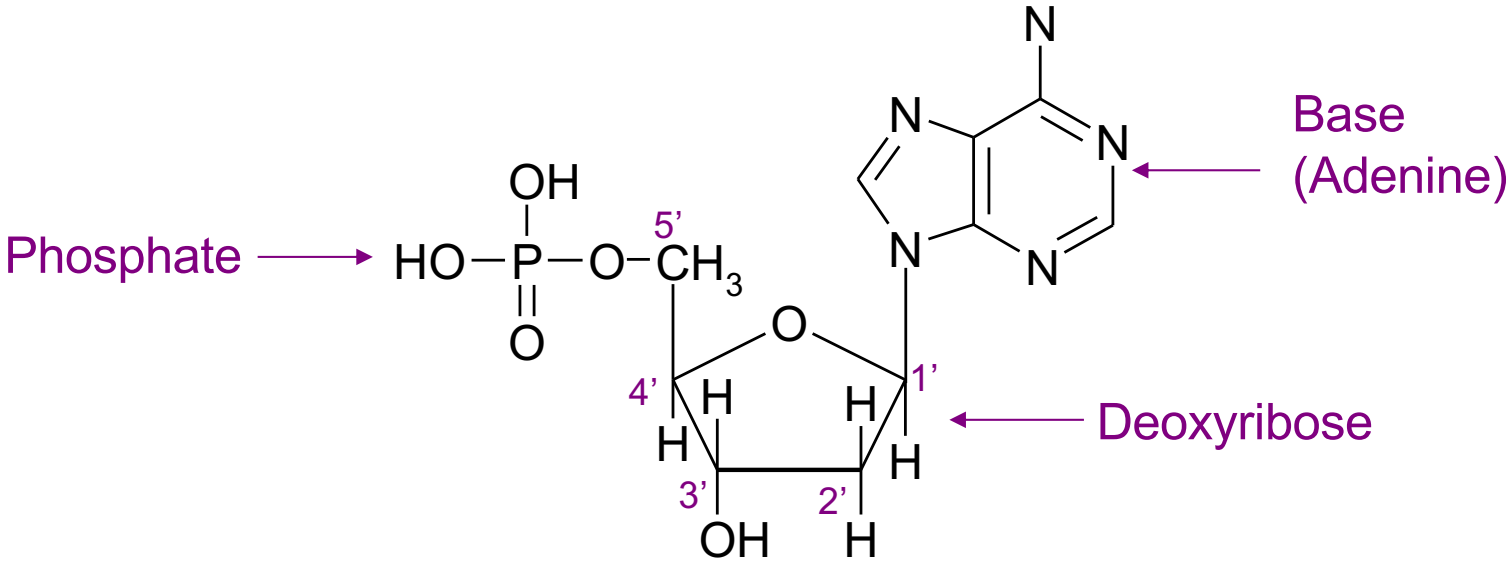


- Normally, DNA is double stranded within a cell. The two strands are antiparallel. One strand is the reverse complement of the other.
- The double strands are interwoven together and form a double helix.
- Double strandedness facilitates DNA replication and provides a "RAID 1" like mirror (backup)



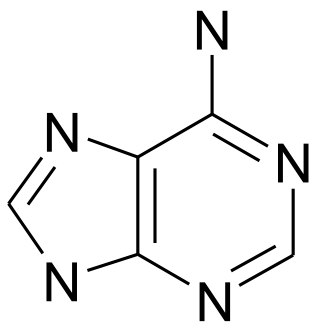
Nucleotides for DNA

- A nucleotide consists of three parts:
  - Deoxyribose
  - Phosphate (bound to the 5’ carbon)
  - Base (bound to the 1’ carbon)

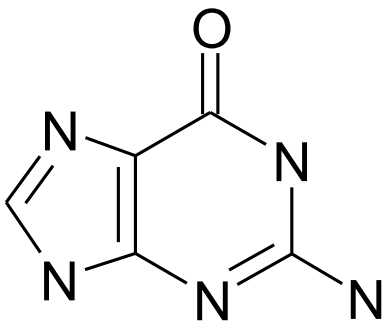


More on bases

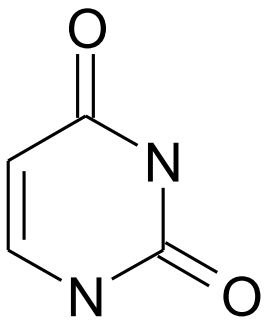
- There are 5 main nucleotides: adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U).
- A, G are called purines. They have a 2-ring structure.
- C, T, U are called pyrimidines. They have a 1-ring structure.
- DNA uses A, C, G, and T
- RNA uses A, C, G and U



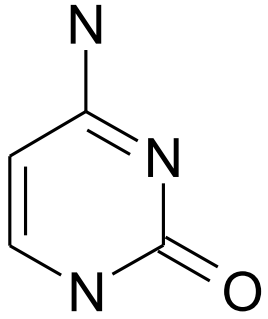
Adenine



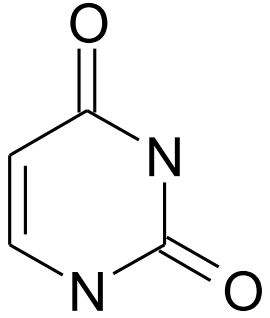
Guanine



Thymine



Cytosine



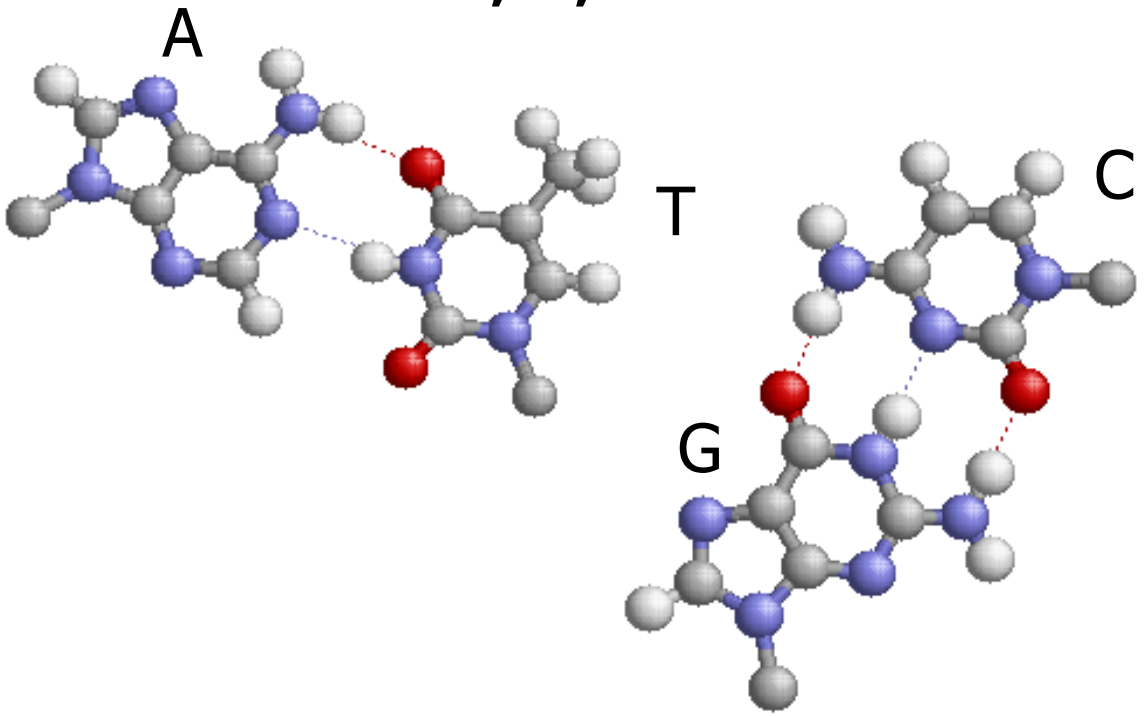
Uracil

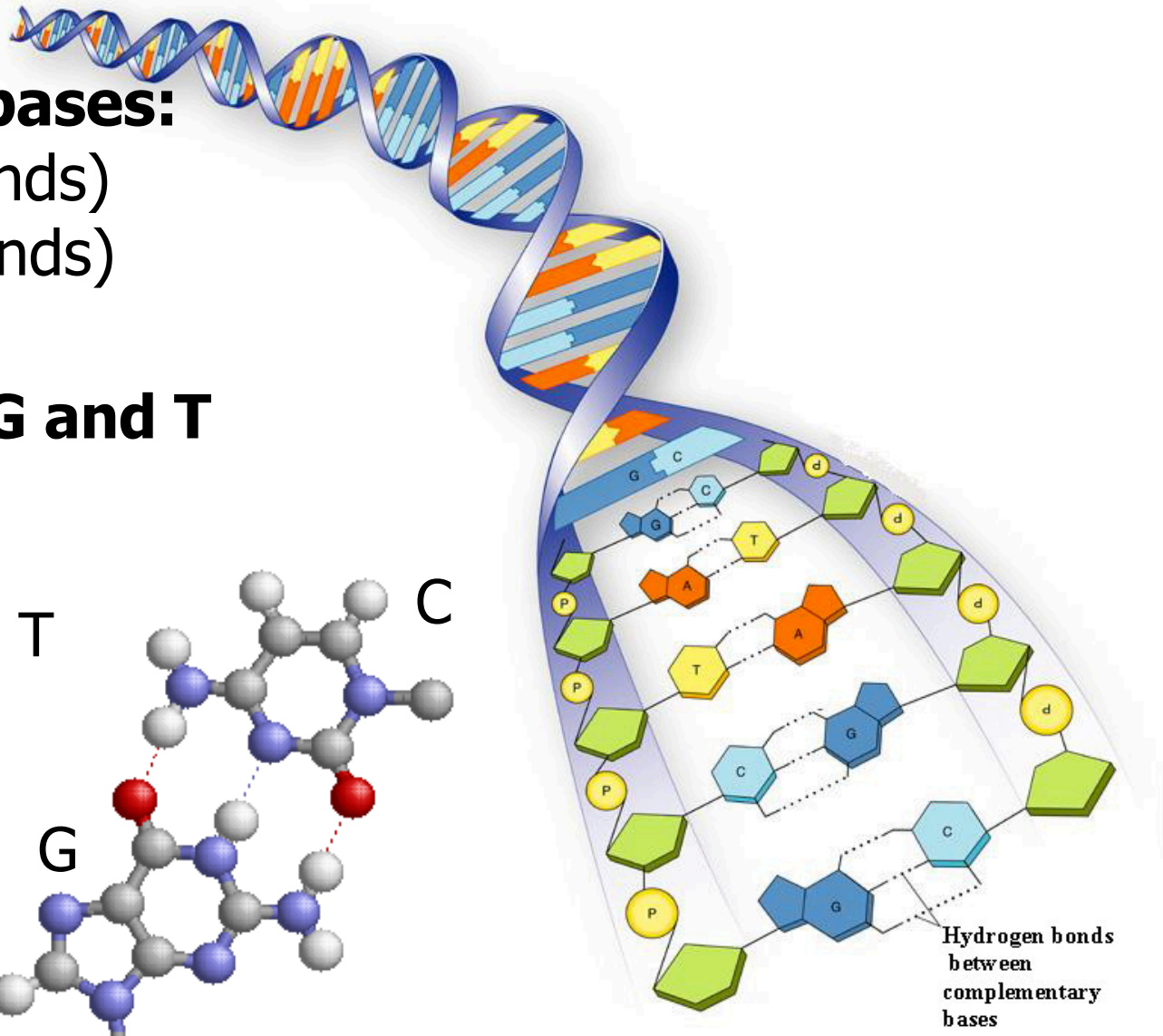
Four nucleotides form 2 pairs

Complementary bases:

- A with T (2 H-bonds)
- C with G (3 H-bonds)

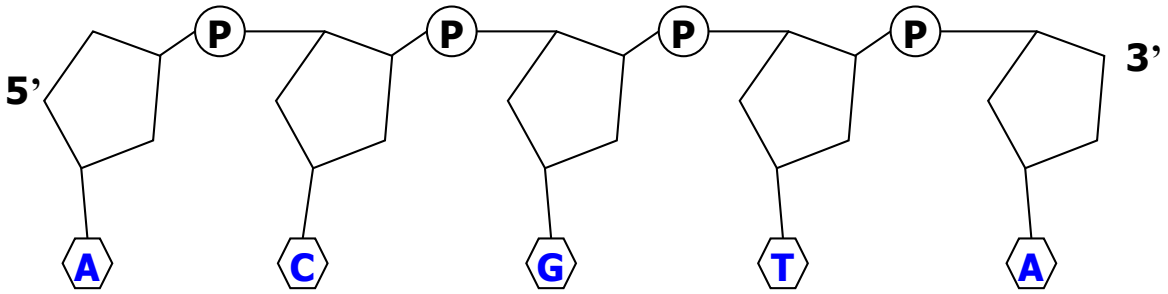
Four bases: A, C, G and T





Orientation of a DNA

- One strand of DNA is generated by chaining together nucleotides.
- It forms a phosphate-sugar backbone.
- It has direction: from 5’ to 3’. (Because DNA always extends from the 3’ end.)
- Upstream: from 5’ to 3’
- Downstream: from 3’ to 5’



Chromosomes

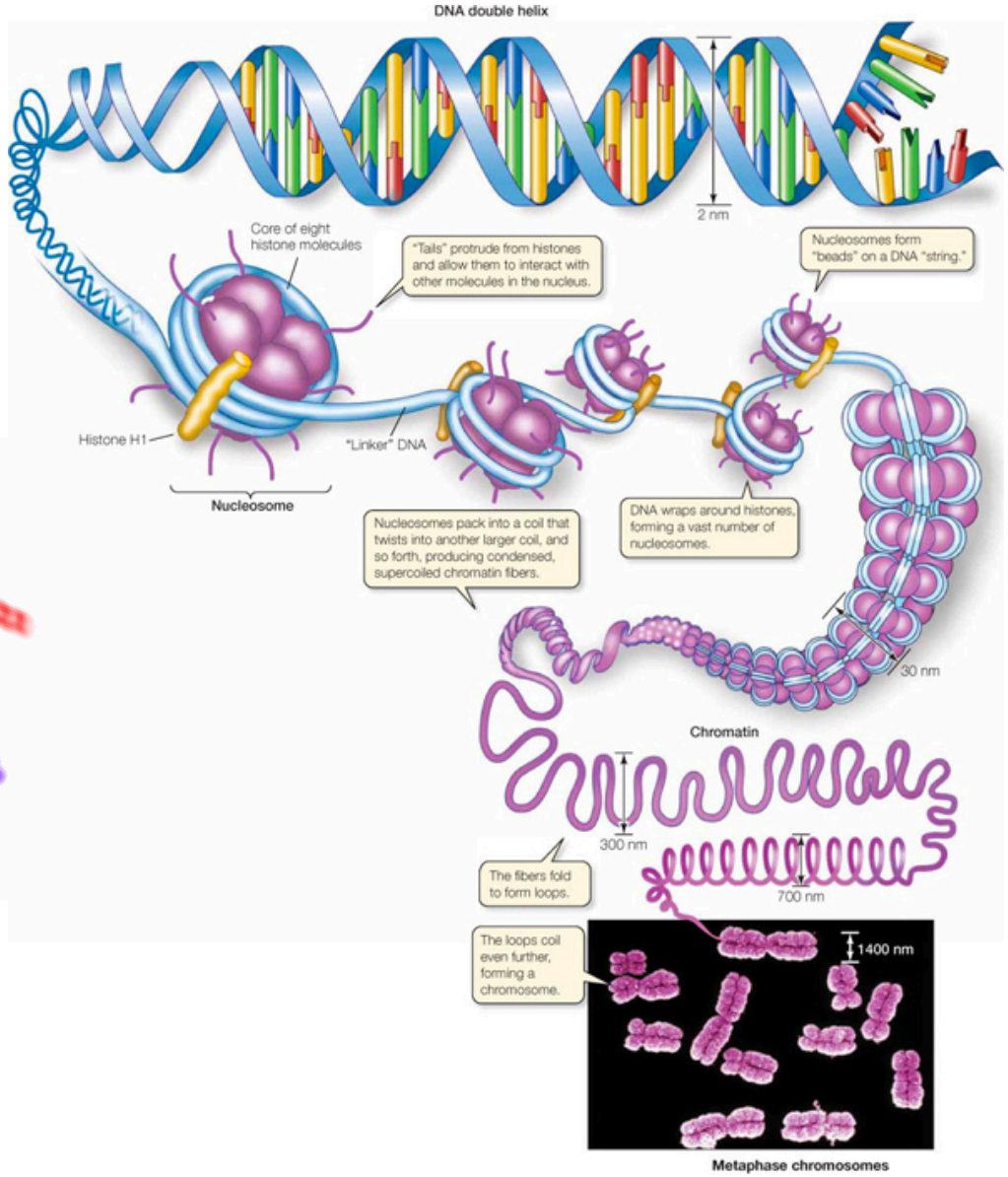
- Usually, DNA is tightly wound around histone proteins and forms a chromosome.
- The total information stored in all chromosomes constitute a genome.
- In most multi-cell organisms, every cell contains the same complete genome.
  - May have some small differences due to mutations
- Example:
  - Human Genome: has 3Gbp (3 billion base pairs), organized in 23 pairs of chromosomes

Genomes and chromosomes

The genome is our genetic material. It consists of DNA. From ~2 to ~150 000 million nucleotides (base pairs).

Human genome with 23 pairs of chromosomes (22 + XX or XY)

ca 3 000 000 000 bp



Genes and genomes

- Prokaryotic genome: E.g. E. coli
  - Number of base pairs: 5M
  - Number of genes: 4k
  - Average length of a gene: 1000 bp

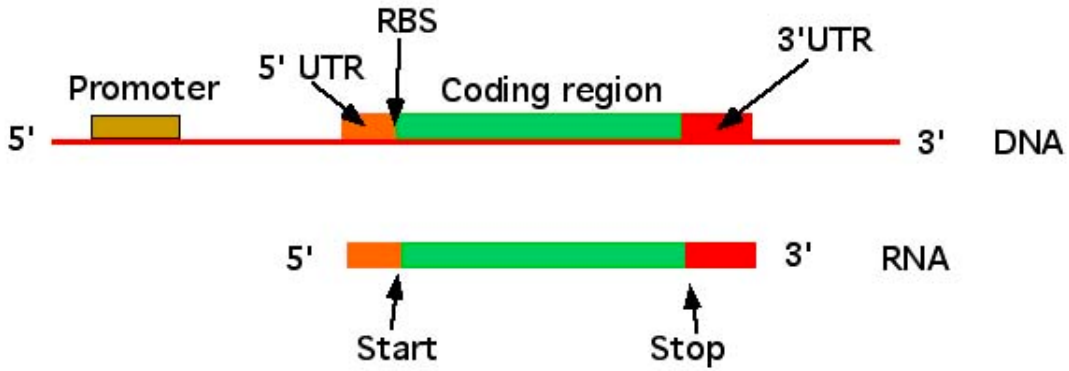
- Eukaryotic genome: E.g. Human
  - Number of base pairs: 3G
  - Estimated number of genes: 20k – 30k
  - Estimated average length of a gene: 1000-2000 bp

- Note that 90% of the E. coli genome consists of coding regions.

- About 1-2% of the human genome is believed to be coding regions. The rest is sometimes called junk DNA.

Gene

- A gene is a sequence of DNA that encodes a protein or an RNA molecule.
- In the human genome, it is expected there are about 20,000 genes encoding proteins
- For a gene that encodes a protein,
  - In Prokaryotic genomes, one gene corresponds to one protein
  - In Eukaryotic genomes, one gene may correspond to more than one protein because of alternative splicing



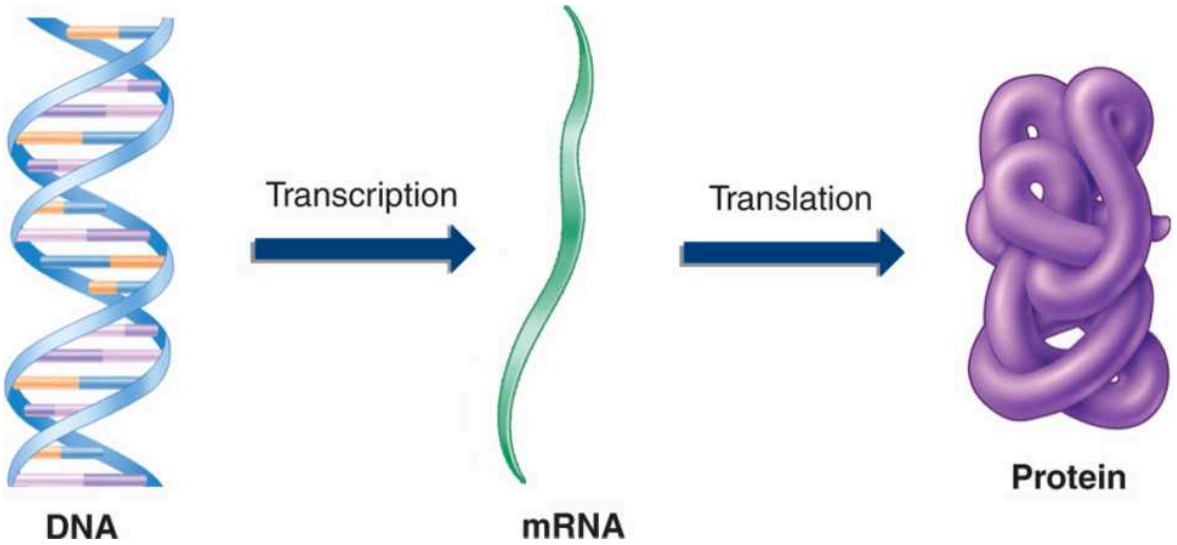
DNA -> mRNA -> Protein

Genes can be turned on and *expressed* (produced) at certain times and places.

The expression of gene consists of at least two steps

- Transcription: DNA → mRNA
- Translation: mRNA → Protein

This is known as "the Central Dogma".



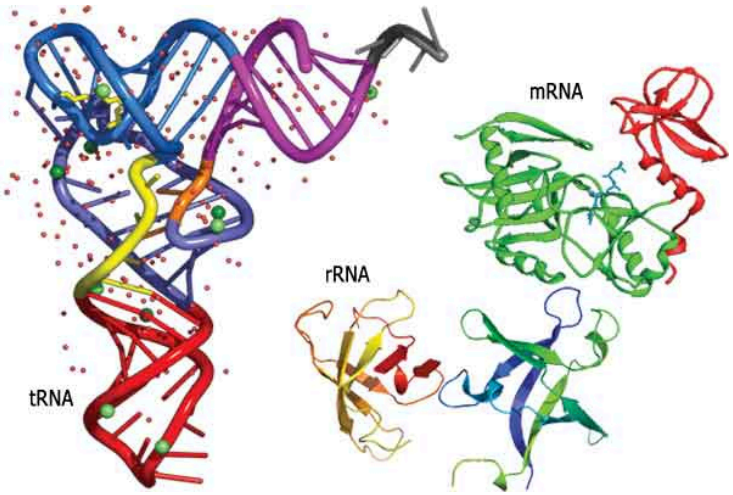
Transcription (Prokaryotes)

- Synthesize a piece of RNA (messenger RNA, mRNA) from one strand of the DNA gene.
  1. The enzyme RNA polymerase temporarily separates the double-stranded DNA
  2. It begins the transcription at the transcription start site.
  3. A → A, C→C, G→G, and T→U
  4. Once the RNA polymerase reaches the transcription termination site, transcription stop.

RNA

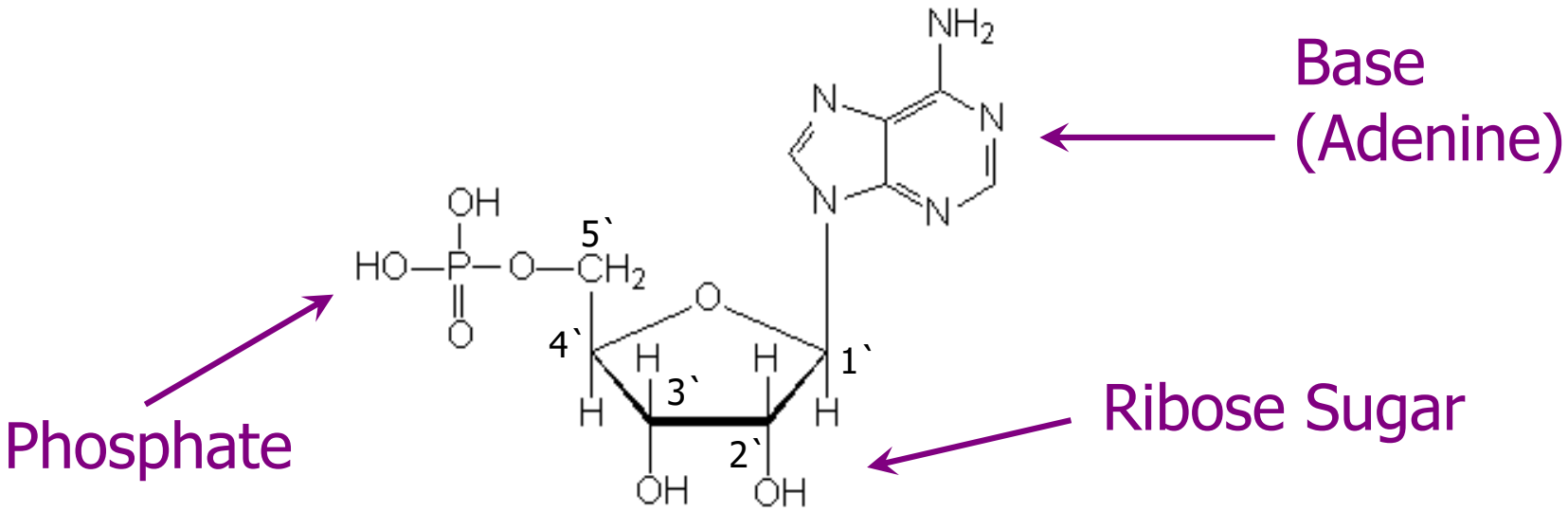
RNA has the properties of both DNA and proteins

- Similar to DNA, it can store and transfer information
- Similar to proteins, it can form complex 3-dimensional structure and perform some functions.



Nucleotides for RNA

- Nucleotides consists of three parts:
  - Ribose Sugar (has an extra OH group at 2')
  - Phosphate (bound to the 5' carbon)
  - Base (bound to the 1' carbon)



RNA vs DNA

- RNA is single stranded
- The nucleotides of RNA are quite similar to those of DNA, except that it has an extra OH at position 2'.
  - Due to this extra OH, it can form more hydrogen bonds than DNA. Thus, RNA can form complex 3-dimensional structures.
- RNA use the base U instead of T.
  - U is chemically similar to T. In particular, U is also complementary to A.

Translation

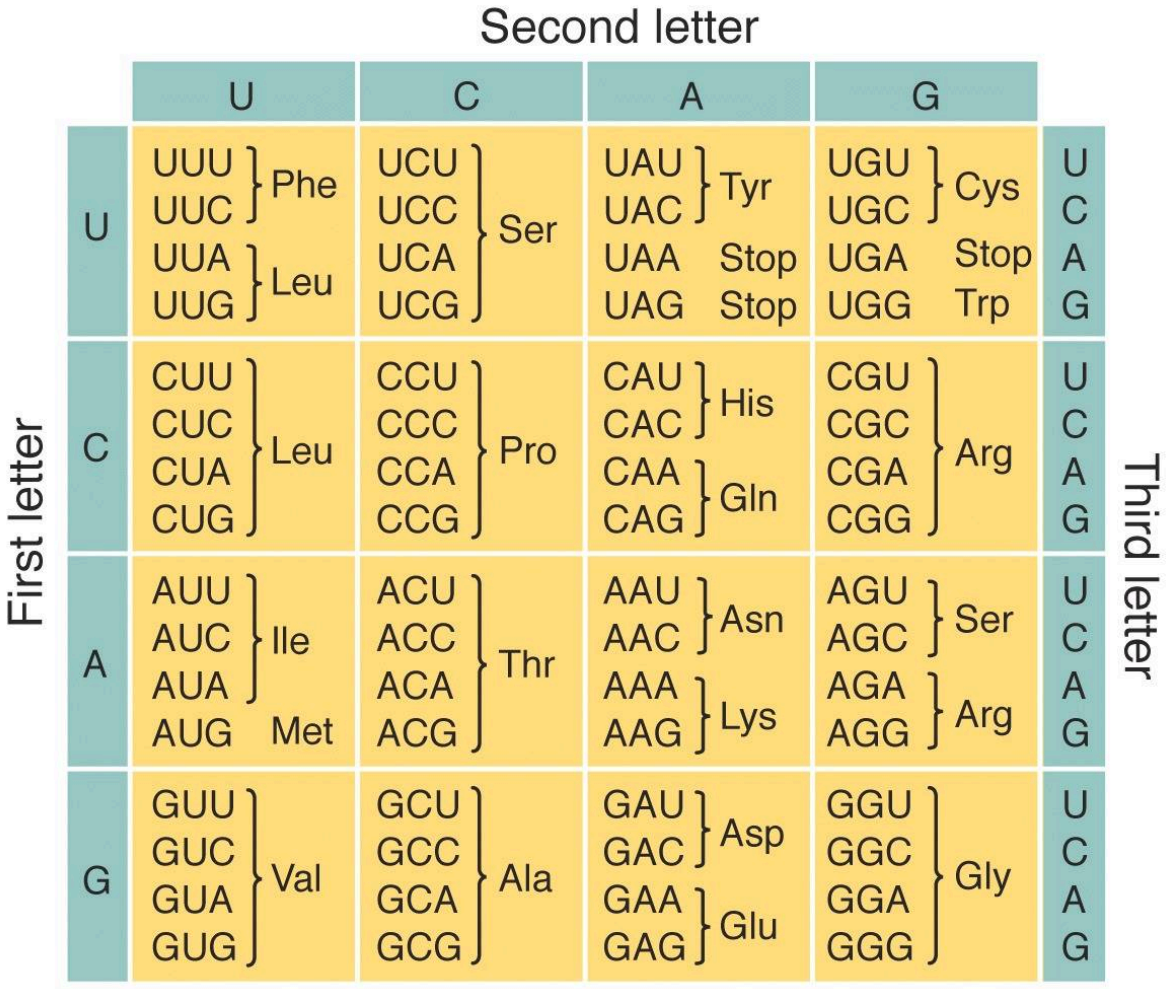
- Translation synthesizes a protein from a mRNA.
- Each amino acid is encoded by consecutive sequences of 3 nucleotides, called a codon.
- The decoding table from codon to amino acid is called the genetic code.
- Note:
  - There are 4^3=64 different codons. Thus, the codons are not one-to-one corresponding to the 20 amino acids.
  - Almost all organisms use the same "universal" decoding table!
  - The codons that encode the same amino acid tend to have the same first and second nucleotide.
  - Recall that amino acids can be classified into 4 groups. A single base change in a codon is usually not sufficient to cause a codon to code for an amino acid in a different group.

The universal genetic code

During translation, groups of 3 nucleotides are read from the mRNA. These *codons* selects new amino acids to be added to the protein chain.

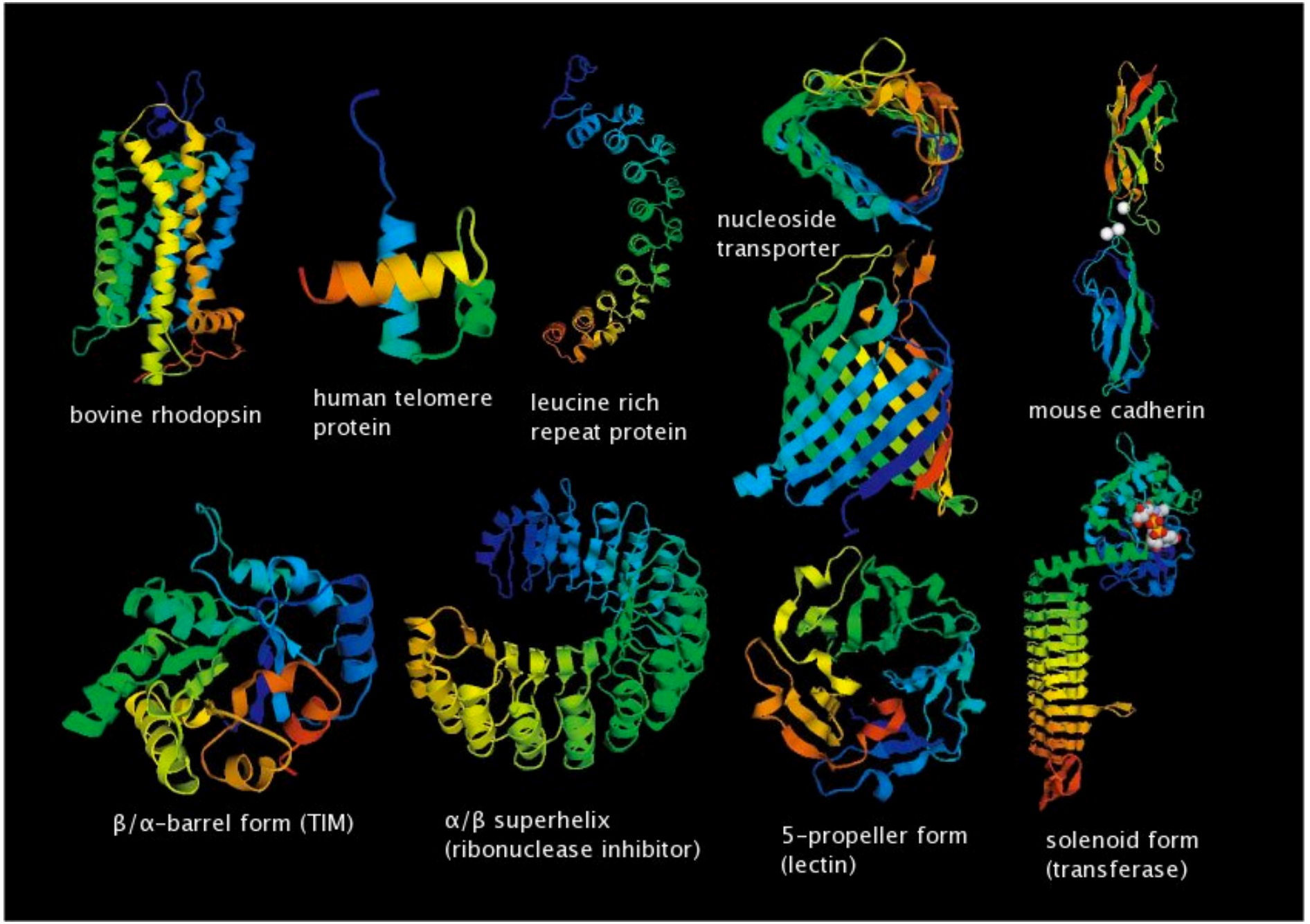
Start codon:
AUG

Stop codons:
UAA,
UAG,
UGA



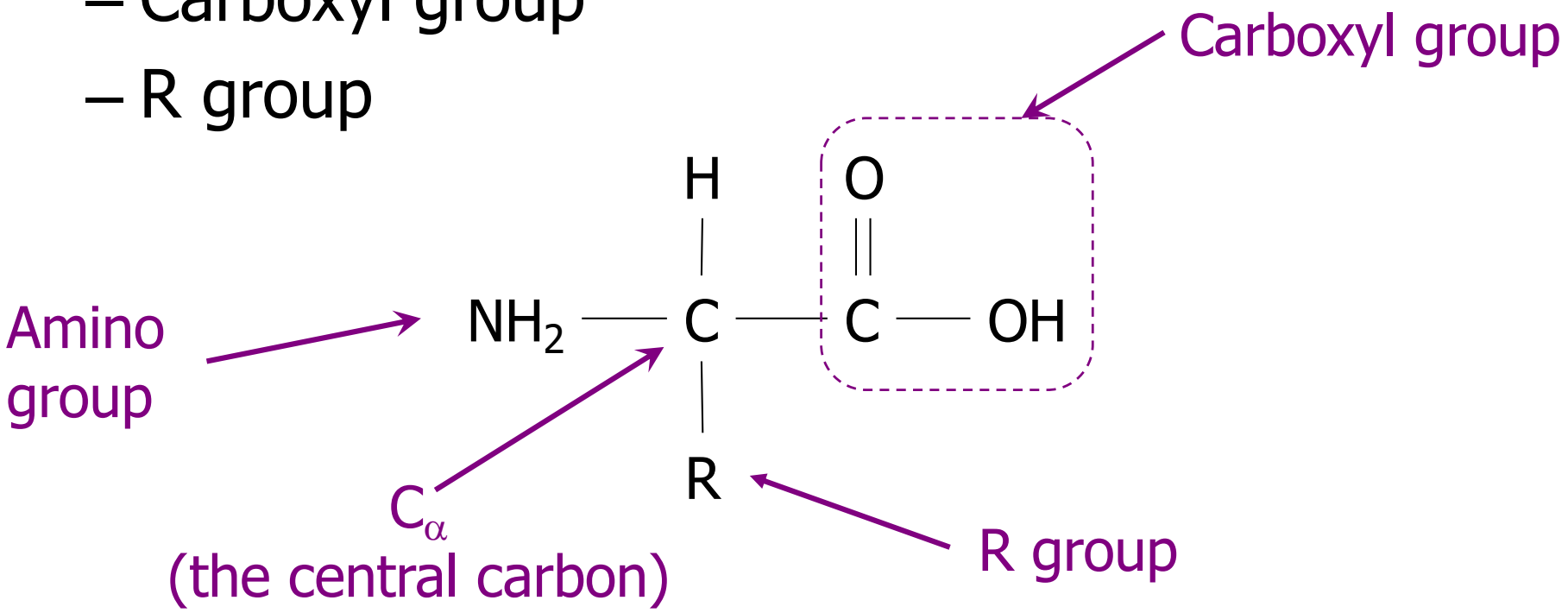
Proteins

- A protein is a sequence composed from an alphabet of 20 amino acids.
  - The length is in the range of 20 to more than 5000 amino acids.
  - On average, proteins contain around 350 amino acids.
- Proteins fold into three-dimensional shapes, which form the building blocks and perform most of the chemical reactions within a cell.

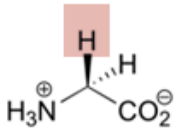

bovine rhodopsin
human telomere protein
leucine rich repeat protein
nucleoside transporter
mouse cadherin
β/α-barrel form (TIM)
α/β superhelix (ribonuclease inhibitor)
5-propeller form (lectin)
solenoid form (transferase)

Amino acids

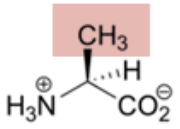
- Each amino acid consist of
  - Amino group
  - Carboxyl group
  - R group



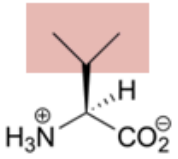
Nonpolar, aliphatic side groups



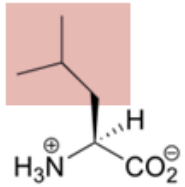
Glycine
Gly, G



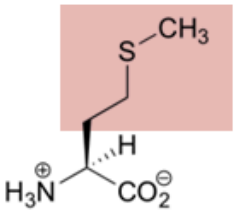
Alanine
Ala, A



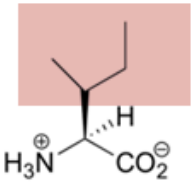
Valine
Val, V



Leucine
Leu, L

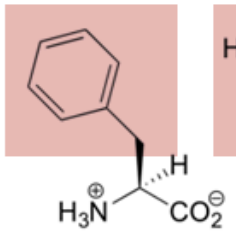


Methionine
Met, M

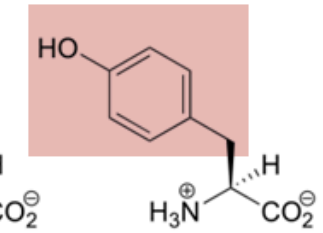


Isoleucine
Ile, I

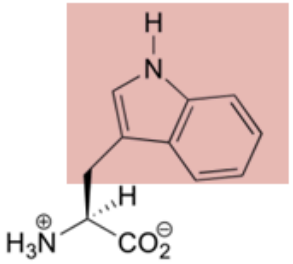
Aromatic side groups



Phenylalanine
Phe, F

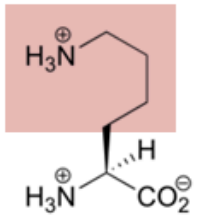


Tyrosine
Tyr, Y

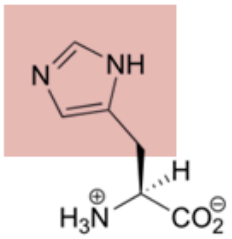


Tryptophan
Trp, W

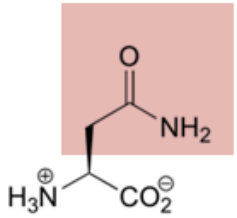
Positively charged side groups



Lysine
Lys, K

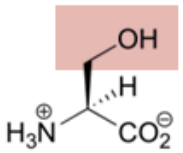


Histidine
His, H

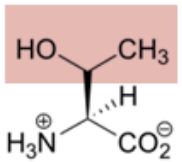


Asparagine
Asn, N

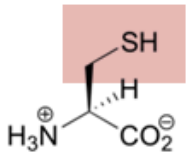
Polar, uncharged side groups



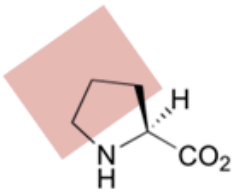
Serine
Ser, S



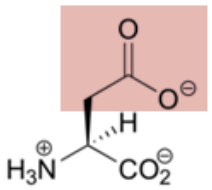
Threonine
Thr, T



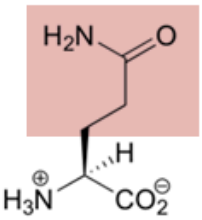
Cysteine
Cys, C



Proline
Pro, P

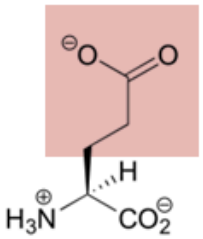


Aspartate
Asp, D

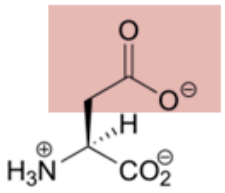


Glutamine
Gln, Q

Negatively charged side groups



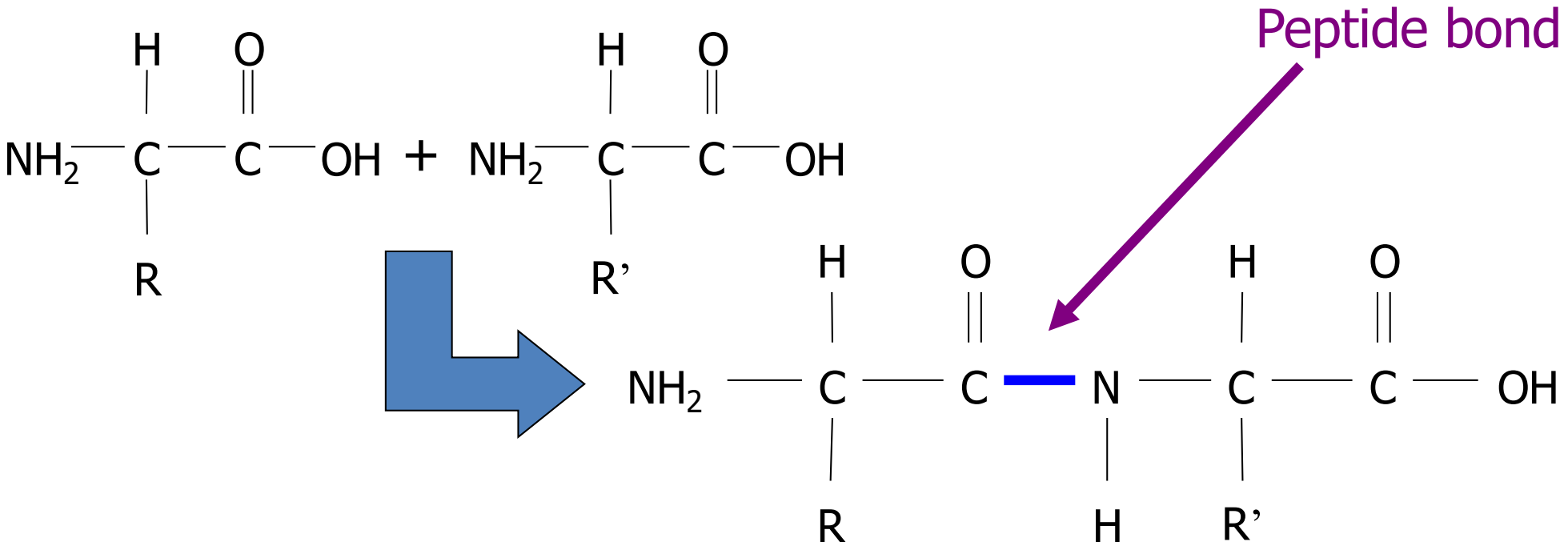
Glutamate
Glu, E



Aspartate
Asp, D

Polypeptides

- A protein or a polypeptide chain is formed by joining the amino acids together via a peptide bond.
- One end of the polypeptide is the amino group, which is called N-terminus. The other end of the polypeptide is the carboxyl group, which is called C-terminus.



Protein structure

- Primary structure
  - The amino acid sequence
- Secondary structure
  - The local structure formed by hydrogen bonding: α-helices and β-sheets.
- Tertiary structure
  - The interaction of α-helices and β-sheets due to hydrophobic effect
- Quaternary structure
  - The interaction of more than one protein to form protein complex