Lecture 1
Intro to Molecular Biology
Major Types of Molecules in a Cell

- Protein
- DNA
- RNA
Proteins

- A *protein* is a chain of molecules, called *amino acids*
- Every amino acid has a central carbon atom, known as *alpha carbon*, or $C_\alpha$, an *amino group* ($\text{NH}_2$), a *carboxyl group* (COOH) and a *side chain*
- The side chain is what distinguishes one amino acid from another
Amino acids

- Alanine: CH₃CH(NH₂)COOH
- Tyrosine: C₆H₄OH.CH₂.CH(NH₂)COOH
- Cysteine: SH.CH₂.CH(NH₂)COOH
- Glycine: NH₂.CH₂.COOH

- Amino group
- Carboxyl group
- Side chain

- Covalent bond
- Carbon atom
- Oxygen atom
- Hydrogen atom
- Nitrogen atom
- Sulphur atom
**Amino acids**

We (usually) find 20 amino acids:

- Alanine (A)
- Cysteine (C)
- Aspartic Acid (D)
- Glutamic Acid (E)
- Phenylalanine (F)
- Glycine (G)
- Histidine (H)
- Isoleucine (I)
- Lysine (K)
- Leucine (L)
- Methionine (M)
- Asparagine (N)
- Proline (P)
- Glutamine (Q)
- Arginine (R)
- Serine (S)
- Threonine (T)
- Valine (V)
- Tryptophan (W)
- Tyrosine (Y)

https://www.youtube.com/watch?v=qBRFIMcxZNM
• DNA carries genetic information
• DNA is a double helix of two complementary strands formed by four nucleotides (bases): **Adenine, Cytosine, Guanine and Thymine**
Prokaryotes and Eukaryotes

- According to the most recent evidence, there are three main branches to the tree of life
- Prokaryotes include Archaea ("ancient ones") and bacteria
- Eukaryotes are kingdom Eukarya and includes plants, animals, fungi and certain algae
{Proka|Euka}ryotes

- **Prokaryotes**: organisms lacking nuclear membrane
  - *E. coli, B. subtilis, H. influenzae, H.pylori, ...*
- **Eukaryotes**: organisms whose DNA is inside the nucleus
  - *A. thaliana, C. elegans, D. melanogaster, H. sapiens, M. musculus, S. cerevisiae (models)*
- Eukaryotic genes may have splicing sites
# Prokaryotes and Eukaryotes

<table>
<thead>
<tr>
<th>Prokaryotes</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single cell</td>
<td>Single or multi cell</td>
</tr>
<tr>
<td>No nucleus</td>
<td>Nucleus</td>
</tr>
<tr>
<td>No organelles</td>
<td>Organelles</td>
</tr>
<tr>
<td>One piece of circular DNA</td>
<td>Chromosomes</td>
</tr>
<tr>
<td>No mRNA post transcriptional modification</td>
<td>Exons/Introns splicing</td>
</tr>
</tbody>
</table>
DNA

- DNA is a *double stranded* chain of **nucleotides**, the building blocks of DNA.
- **Nucleotide** = sugar + phosphate + base
- possible bases: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T)
DNA

- Each base in one strand is paired to its *complement* on the other strand: A↔T   C↔G
- *Reverse-complementation* operation
  e.g. \( y = ATTCGGAT \)
  \( \tilde{y} = ATCCGCAAT \)
• Each sugar molecule contains five carbon atoms (labeled 1’ through 5’)
• Backbone bonds are between the 3’ carbon and the 5’ carbon
• Orientation of DNA is by convention 5’ to 3’
Orientation of DNA is by convention 5' to 3'

Overall Direction of Replication

5'

Parental Strand

3'

DNA Helicase (breaking bonds)

Replication Fork

Primer

DNA Pol III

Lagging Strand

Leading Strand

3'

5'

DNA Primase
RNA

- Single stranded
- Uracil (U) instead of thymine (T)
- Different types of RNA
  - mRNA (messenger RNA)
  - tRNA (transfer RNA)
  - rRNA (ribosomal RNA)
  - ...
- RNA is much less stable than DNA
Genes

- A **gene** is a segment of DNA that encodes for one or more polypeptide chain (protein)
- In Eukaryotes, a gene consists of alternating regions called exons and introns
- Exons correspond to coding regions
• **Gene expression** is the process by which DNA is transcribed into mRNA (eventually translated into proteins)

• Mechanisms controlling gene expression are not fully understood yet
Transcription is the first step of gene expression, in which a particular segment of DNA (gene) is copied into RNA by the enzyme RNA polymerase.
Transcription in Eukaryotes:
1. A gene is copied into a primary RNA, mRNA precursor
2. Introns are removed in a process called splicing
3. Poly-tail of A’s is attached at 3’ end of mature mRNA
Splicing is a process of removal of introns from the primary RNA.
**Translation** is the process of synthesizing a protein from mRNA

**Genetic code:** Triplets of mRNA bases (codons) associate with corresponding amino acids
Figure 7.1 All the triplet codons have meaning: 61 represent amino acids, and 3 cause termination (STOP).

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU</td>
<td>UUC</td>
<td>UUA</td>
<td>UUG</td>
</tr>
<tr>
<td></td>
<td>Phe</td>
<td>Leu</td>
<td>Leu</td>
<td>Arg</td>
</tr>
<tr>
<td>C</td>
<td>CUU</td>
<td>CUC</td>
<td>CUA</td>
<td>CUG</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>Ile</td>
<td>Met</td>
</tr>
<tr>
<td>A</td>
<td>AUU</td>
<td>AUC</td>
<td>AUA</td>
<td>AUG</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td>Met</td>
<td>Val</td>
</tr>
<tr>
<td>G</td>
<td>GUU</td>
<td>GUC</td>
<td>GUA</td>
<td>GUG</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Asp</td>
<td>Ala</td>
<td>Gly</td>
</tr>
</tbody>
</table>

https://www.youtube.com/watch?v=OEWOZS_JTgk
Alternative splicing

Example (mouse gene)

α-TM EXON GENE ORGANIZATION

α-TM mRNA TRANSCRIPTS

Striated muscle
Striated muscle'
Myoblast
Smooth muscle
Nonmuscle/fibroblast
Hepatoma
Brain
ORFs

- A stretch of DNA has six *reading frames*
- *Open reading frame* (ORF): start codon, **AUG** integral number of codons (none of which is a stop codon, **UAA, UAG, UGA**)
- Typically less than 10% of the DNA of eukaryotes codes for proteins
- The remaining 90% is called *junk DNA*
Figure 1.26  An open reading frame starts with AUG and continues in triplets to a termination codon. Blocked reading frames may be interrupted frequently by termination codons.
Figure 2.26  Two genes may share the same sequence by reading the DNA in different frames.
Genome

- In higher organisms, DNA is contained in *chromosomes*
- Number of chromosomes in characteristic of the specie (e.g., *H. sapiens* has 46, *S. cerevisiae* has 32, *C. elegans* has 12, *D. melanogaster* has 8)
- Eukaryotic chromosomes appear in pairs
- Corresponding genes in *homologous* chromosomes may differ (*alleles*)
Human Genome by numbers

• Contains $\approx 3,164.7$ billion nucleotide bases
• The average gene consists of 3,000 bases, but sizes vary greatly, with the largest known human gene being \textit{dystrophin} at 2.4 million bases
• The total number of genes is estimated at 30,000 to 35,000 much lower than initial estimates of 80,000-140,000 that had been based on previous extrapolations
• Next generation sequencing (NGS) technology revolutionized genomic research
• NGS technology allows fast and inexpensive DNA sequencing producing hundreds of millions of DNA sequenced reads
  – Sequence DNA fragments
  – Map reads back to a reference genome
  – Analyze the data