“Workhorse” or “machinery of life”

Functions:
- Structural elements (e.g. collagen fibers)
- Catalysis of chemical reactions (enzymes)
- Sensing (light, chemical signals)
- Immune system
- ...
Central Dogma of Molecular Biology

©️⃣❼⃣ Adapted from Central Dogma of Molecular Biochemistry with Enzymes by Dhorspool
C N U L C E I C A C I D S

- DNA and RNA
- Consist of many nucleotides
- Store and transmit genetic information
- Can adopt well defined structure

Image from the RCSB PDB of PDB ID 1BKV
RNA: typically single stranded but can adopt complex secondary structure with itself or other RNA molecules

DNA: typical double stranded (double helix) formed by a strand with its reverse complement
Nucleic Acids

- TCGA for DNA, UCGA for RNA
- T/U is complementary to A, C to G
- Complementarity is symmetric

©️ Madeleine Price Ball from DNA chemical structure

Adapted from DNA chemical structure by Madeleine Price Ball
Replication

- Duplication of DNA molecules
- Replication is highly controlled and starts at a defined point (called origin or ori)
- For each strand in the double helix there is a complementary strand synthesized
- The existing single strand serves as template
- Synthesis takes place in 5’→ 3’ direction
- Carried out by multi-protein complex

Generic structure of nucleotides
**Transcription/Translation**

**Transcription**
- “Creation of working copy of genes” (like making a copy of a protected file)
- Synthesis of single stranded RNA with DNA as template (starting after a region)

**Translation**
- Starts with first AUG/ATG codon
- Conversion of the information encoded in an RNA molecule into a protein (amino acid sequence)

©️ Adapted from *Central Dogma of Molecular Biochemistry with Enzymes* by Dhorspool
The basic physical unit of heredity; a linear sequence of nucleotides along a segment of DNA that provides the coded instructions for synthesis of RNA, which, when translated into protein, leads to the expression of hereditary character.

⇒ This comprises the protein coding regions as well as non-coding and regulatory regions
Adapted from Overview of the Central Dogma of Molecular Biology by Mike Jones
Adapted from Codon Wheel for translating genetic code from the Wellcome Trust Sanger Institute by Dunk
Biomolecules

- Single amino acid $\equiv$ residue
- Proteins are made up of multiple residues ($\equiv$ Polymer)
- Average number of residues is 200 - 300
- Lower end: Current research suggests that short proteins ($< 50$) are more important than previously anticipated
- Upper end: Titin (33k residues)
In biology, lipid is a loosely defined term for substances of biological origin that are soluble in nonpolar solvents. Include fats, waxes, sterols, fat-soluble vitamins (such as vitamins A, D, E, and K). Functions of lipids include storing energy, signaling and acting as structural components of cell membranes.
Adapted from *Cell membrane detailed diagram* by Dhatfield
Amino acids have different properties:
- Size
- Charge
- Aromaticity
- Hydrophobicity
- ...
Properties of AAs

Image from Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences by Rives et al.
Amino acids have different properties:
- Size
- Charge
- Aromaticity
- Hydrophobicity
- ...

Why do we care?
- Changes (i.e. mutations) in the DNA/RNA *can* lead to changes in the residues of a protein
- These changes *might* lead to malfunction (disease) if mutated AA has different properties
Changes lead to the rise of evolution (via the constant iteration of mutation and selection based on fitness)

**Single mutation types:**
- Change of a single nucleotide
- Insertion of a single nucleotide
- Deletion of a single nucleotide
Rearrangement of DNA segments (real molecule fragments) due to an undesired recombination process (wrong order, loss of fragment, ...)

Induced by: meiosis, DNA damage (chemical, radiation), virus infections
Substitutions

- **neutral**: residue changed but protein functional
- **silent**: residue yields a redundant codon
- **missense**: leads to the change of an amino acid
- **nonsense**: introduced a stop codon, further effect depends on the specific position
EXERCISE
ORF sequences consist of:

1. **Start codon** (ATG; also encoding Methionine)
2. Intermediate codons (encoding amino acids)
3. **Stop codon** (TAA, TAG, TGA; do not encode amino acids)

A single ORF can encode multiple protein sequences by

- Not cutting some introns
- Excluding some exons

**This is of no concern for the exercise!**

**We always assume 1 ORF = 1 protein!**
Since codons are triplets, we have three reading frames.

- ORFs stop at the first stop codon on the same frame.

- **GCTATGAGGTCATGGCTTCTGTAGTAACG**
  - ORF on F1: ATG AGGTCATGGCTTCTG TAG TAA CGT GAC
  - ORF on F3: ATG GCTTCTGTAGTAACG TGA

**DNA “Reading Frames”**
Since codons are triplets, we have three reading frames

- ORFs stop at the first stop codon on the same frame

<table>
<thead>
<tr>
<th>Frame</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1:</td>
<td>GCT ATG AGG TCA TGG CTT CTG TAG TAA CGT GAC</td>
</tr>
<tr>
<td>F2:</td>
<td>CTA TGA GGT CAT GGC TTC TGT AGT AAC GTG</td>
</tr>
<tr>
<td>F3:</td>
<td>TAT GAG GTC ATG GCT TCT GTA GTA ACG TGA</td>
</tr>
</tbody>
</table>
Since codons are triplets, we have three reading frames

ORFs stop at the first stop codon on the same frame

GCTATGAGGTCATGGCTTCTGTAACGTGAC

F1: GCT ATG AGG TCA TGG CTT CTG TAG TAA CGT GAC

F2: CTA TGA GGT CAT GGC TTC TGT AGT AAC GTG
Since codons are triplets, we have three reading frames. ORFs stop at the first stop codon on the same frame.

GCTATGAGGTCATGGCTTCTGTAGTAACGTGAC

F1: GCT ATG AGG TCA TGG CTT CTG TAG TAA CGT GAC
F2: CTA TGA GGT CAT GGC TTC TGT AGT AAC GTG
F3: TAT GAG GTC ATG GCT TCT GTA GTA ACG TGA
Since codons are triplets, we have three reading frames. ORFs stop at the first stop codon on the same frame.

GCTATGAGGTCATGGCTTCTGTAGTAACG TGAC

F1: GCT ATG AGG TCA TGG CTT CTG TAG TAA CGT GAC
F2: CTA TGA GGT CAT GGC TTC TGT AGT AAC GTG
F3: TAT GAG GTC ATG GCT TCT GTA GTA ACG TGA

One ORF on F1: ATGAGGTCATGGCTTCTGTAG
One ORF on F3: ATGGCTTCTGTAGTAACG TGAC
**Reverse Complementary**

- Complementary: A→T, T→A, G→C, C→G
- Reverse: ATGC → CGTA

<table>
<thead>
<tr>
<th>Strand</th>
<th>DNA sequence (read from 5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>5‘ GCTATGAGGTCATGCTTCTGTAGTAACGTGAC 3’</td>
</tr>
<tr>
<td>Reverse Complement (RC)</td>
<td>3‘ CGATACTCCAGTACCAGAAGACATCATTGCACCTG 5’</td>
</tr>
<tr>
<td>RC in reading direction</td>
<td>5‘ GTCACGTTACTACAGAAGCCATGACCTCATAGC 3’</td>
</tr>
</tbody>
</table>

---

ROSTLAB.
Indexing ORFs

- We start at index 0: first nucleotide (nt) in primary strand
- Rev. comp. strand is indexed according to primary strand
  - First nt in rev. comp. has same index as last nt in primary
  - Last nt in rev. comp. has same index as first nt in primary, i.e. 0
- Start of an ORF is the index of the first nt in start codon
- Stop of an ORF is the index of the last nt in stop codon
### DNA sequence

<table>
<thead>
<tr>
<th>Strand</th>
<th>DNA sequence (read from 5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary</strong></td>
<td>5’ GCTATGAGGTCATGGCTTCTGTAGTAACGTGAC 3’</td>
</tr>
<tr>
<td><strong>Reverse Complement (RC)</strong></td>
<td>3’ CGATACTCCAGTACCGAAGACATCATTGCACTG 5’</td>
</tr>
<tr>
<td><strong>RC in reading direction</strong></td>
<td>5’ GTCACGTTACTACAGAAGCCATGACCTCATAGC 3’</td>
</tr>
</tbody>
</table>
## Example

<table>
<thead>
<tr>
<th>Strand</th>
<th>DNA sequence (read from 5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>5’ GCTATGAGGTCATGGCTTCTGTAGTAACGTGAC 3’</td>
</tr>
<tr>
<td>Reverse Complement (RC)</td>
<td>3’ CGATACTCCAGTACCGAAGACATCATTGCACTG 5’</td>
</tr>
<tr>
<td>RC in reading direction</td>
<td>5’ GTCACGTTACTACAGAAGCCATGACCTCATAGC 3’</td>
</tr>
</tbody>
</table>

- **ATGAGGTCATGGCTTCTGTAG** – ORF on primary F1; Start: 3, Stop: 23
<table>
<thead>
<tr>
<th>Strand</th>
<th>DNA sequence (read from 5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>5’ GCTATGAGGTCATGGCTTCTGTAGTAACGTGAC 3’</td>
</tr>
<tr>
<td>Reverse Complement (RC)</td>
<td>3’ CGATACTCCAGTACCGAAGACATCATTGCACCTG 5’</td>
</tr>
<tr>
<td>RC in reading direction</td>
<td>5’ GTCACGTTACTACAGAAGCCATGACCTCATAGC 3’</td>
</tr>
</tbody>
</table>

- **ATGAGGTCATGGCTTCTGTAG** – ORF on primary F1; Start: 3, Stop: 23
- **ATGGCTTCTGTAGTAACGTGTA** – ORF on primary F3; Start: 11, Stop: 31
## Example

<table>
<thead>
<tr>
<th>Strand</th>
<th>DNA sequence (read from 5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary</strong></td>
<td>5’ GCTATGAGGTCA TGGCTTCTG TAGTAACG TGAC 3’</td>
</tr>
<tr>
<td><strong>Reverse Complement (RC)</strong></td>
<td>3’ CGAT ACTCCA GTA CCGAAGACATCATTTGCACTG 5’</td>
</tr>
<tr>
<td><strong>RC in reading direction</strong></td>
<td>5’ GTCACGT TACTACAGAAGCC ATGACCT CAT TAGC 3’</td>
</tr>
</tbody>
</table>

- **ATGAGGTCATGGCTTCTG TAG** – ORF on primary F1; Start: 3, Stop: 23
- **ATGGCTTCTGTAGTAACG TGA** – ORF on primary F3; Start: 11, Stop: 31
- **ATGACCT CATAG** – ORF on rev.comp. F3; Start: 12, Stop: 1
Circular DNA

File start

3'  5'

Start codon

End codon
Circular DNA

File start

3’ 5’

Start codon

End codon
Thank you!

QUESTIONS?