DNA, RNA, replication, translation, and transcription

Overview

Recall the central dogma of biology:

DNA (genetic information in genes) → RNA (copies of genes) → proteins (functional molecules)

DNA structure

One monomer unit = deoxyribonucleic acid

- composed of a base, a sugar (**deoxy**ribose), and a phosphate
- directionality along the backbone → 5' (phosphate) to 3' (OH)

Double-strand pairing:

- complementary **base-matching**: A-T, C-G
- base-matching achieved by H-bonding and geometry (long vs short nucleotides)
- antiparallel (one strand 5'→3', the other 3'→5')

Helical shape

- 10.4 nucleotides per turn
- diameter = 2 nm
- both major and minor grooves
- called **B-DNA**. The helix twist and diameter can also change under dehydrating conditions and methylation to **A-DNA** and **Z-DNA**

Base-pairing and strand interactions

- A, G are long (double ring purines)
- C,T are short (single ring pyrimidines)
- need one long and one short nucleotide per pair
- C-G have three hydrogen bonds (slightly stronger matching)
- A-T have two hydrogen bonds (slightly weaker matching)

- **base stacking** of aromatic rings allows sharing of pi electrons and adds stability to interior structure of DNA → some hydrophobic driving force as well
- pair structure allows template for **semi-conservative copying**

Information in DNA sequence is the **genome**

- **genes** are stretches of information in the sequence that encode for particular function (usually a particular protein, but sometimes also an RNA sequence)
- about 20,000 genes in humans
- typically 1000s of nucleotides long
- genes can be **expressed** (use to make proteins) or **repressed** (not used)
- regions of DNA are divided into **coding and non-coding** segments
- over 50% of human DNA is non-coding
- genes can be spliced together
- genes are organized in the large-scale structure of the DNA in the nucleus

In bacteria, genome usually **circular**

The genome in eukaryotes is organized into **chromosomes**

- each chromosome a separate DNA molecule
- human cells contain 46 chromosomes (22 each from mother and father)
- chromosomes are extended and replicated during **interphase** portion of the cell cycle → extended allows for gene expression
- chromosomes are condensed, visible with light during cell division (**M phase**)

Special DNA sequences exist in each chromosome

- **replication origins** – multiple locations where the replication machinery first binds to start replication
- **centromere** – center "pinch point" of a chromosome that allows one copy of each to be pulled apart into two daughter cells during division

- **telomere** – specialized sequences at the chromosomes end that facilitate replication there

Higher-order DNA structure

- How do cells efficiently store very long chains of DNA?
- DNA wraps around protein "spools" to form **nucleosomes**
- Nucleosomes are made of **histone** proteins
- Spools organize into **chromatin fibers** that pack in regular ways, on different length scales

Replication

DNA replication is **semi-conservative** → one strand from each of the initial two strands end up in a daughter strand

Each strand serves as a template for a new strand

New strand is formed by complementary base-pairing of the correct nucleotide plus formation of a phosphodiester bond

Synthesis begins at replication origins

- about 100 nucleotides long → rich in A-T, which are easier to pull apart because have 2 rather than 3 hydrogen bonds
- ~1 in bacteria
- ~10000 in humans

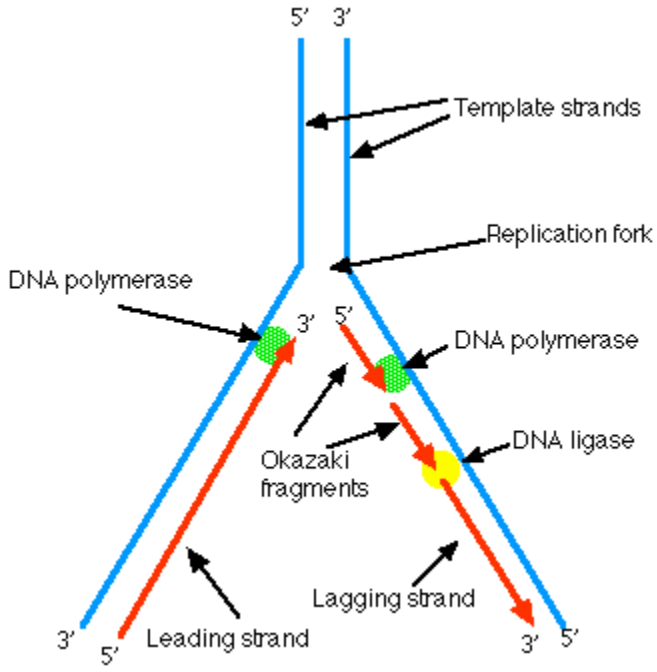
Initiator proteins bind at replication origins and recruit DNA replication machinery proteins

- **DNA polymerase** is responsible for catalyzing synthesis of new strands

Replication forks form and involve a **leading** and a **lagging** strand

- DNA is directional; two strands are antiparallel
- DNA polymerase can only synthesize from 5' to 3' direction, adding new nucleotides to the 3' end

- lagging strand must be synthesized by first spooling out some template strand and then synthesizing in reverse



Error-correction machinery

- mutations occur 1 in 10^9 nucleotides copied → evolution, cancer
- much better error rates than expected simply from base-pairing energetics
- DNA polymerase **proofreads** to make sure correct nucleotide is added → if not, it excises and goes back to add the correct one
- **Mismatch repair** machinery fixes incorrectly added nucleotides not found by DNA polymerase → detects nicks in newly created strand

Damage to DNA continuously occurs

- **Homologous recombination** uses similar sequences in nearby strands in order to fill in excised damaged DNA
- also the basis of heredity

Transcription

Messenger RNA, or mRNA, is the RNA "copies" of genes ultimately used to synthesize proteins, although some RNA are the final product themselves

RNA has some distinctions from DNA

- ribose rather than deoxyribose sugar (differs in an OH group)
- **uracil** instead of thymine (loss of a methyl group)
- single-stranded, and typically folds into unique shapes, like proteins
- less chemically stable

Other kinds of RNA

- **Ribosomal RNA, rRNA**, is RNA that becomes part of the ribosome, the big molecular machine responsible for synthesizing proteins
- **Transfer RNA, tRNA**, is used to bring correct amino acids to the ribosome during protein synthesis
- **Micro RNAs (mRNAs)** are important in regulating gene expression
- others

Transcription involves the synthesis of rRNA from DNA using **RNA polymerase**

- RNA polymerase must unpair and unwind DNA as it is reading it
- much less accurate than replication → errors of 1 in 10^4
- protein synthesis can tolerate more errors
- multiple RNAs can be sequenced from the same gene at the same time

In bacteria:

- RNA polymerase binds to specific regions of the DNA called **promoters,** specific nucleotide sequences
- Promoters orient polymerase in a specific direction
- RNA polymerase binds to the promoter with the help of an accessory protein, called a **sigma factor**
- RNA transcript is synthesized by ribonucleotide triphosphate additions
- Synthesis stops at a **terminator** sequence, typically of poly A-T stretches of DNA

In eukaryotes, the situation is different in a number of ways:

1. Different kinds of RNA polymerases, depending on whether the product is protein or RNA
2. RNA polymerase requires a number of helper proteins to bind to DNA and initiate RNA synthesis → **transcription factors**
3. Transcription factor TFIID binds to a specific DNA sequence upstream 25 nucleotides from the region coding for the protein → **TATA sequence or TATA box**
4. Other proteins assemble to form a large **transcription complex**
5. **Chromatin-remodeling proteins** are involved to make DNA accessible from the wound histone structure
6. RNA chemically modified with a **methylated guanine** at the 5' end and a **poly-A** sequence at the 3' end → these help the ribosome later ensure that the complete recipe for a protein is there
7. RNA is processed after synthesis → **splicing** to remove **noncoding** regions called **introns**
8. The nuclear pore complex selectively exports complete, spliced mRNA molecules to the cytosol

Splicing has evolutionary advantages

- promotes diversity of proteins produced from a single gene → **alternative splicing**
- enables formation of new proteins from combinations of different genes separated by long noncoding regions

Translation

Information transmission

- 4 bases in DNA/RNA to 20 amino acids in proteins
- "translation" since the chemical language is different
- How many nucleotides needed to specify each amino acid?
- Two = 16 combinations → not enough!
- Three = 64 combinations → plenty!

Processed (e.g., spliced) mRNA is read in groups of three nucleotides

- called **codons**
- redundancy of codons for different amino acids → typically the last nucleotide is variable
- three possible **reading frames** depending on starting nucleotide

transfer RNAs (tRNAs) are the intermediates between nucleotides and amino acids

- about 80 nucleotides long
- have specific 3D shape, like an "L"
- at one end: **anticodon** that base-pairs with mRNA
- at the other end: covalently coupled amino acid
- different tRNAs for each amino acid type
- base pairing weakest at third "wobble" nucleotide (why it is most variable)
- tRNAs are **charged** with an amino acid by **aminoacyl tRNA synthetases** that ensure correct addition of individual amino acids to corresponding tRNA

The synthesis of proteins is choreographed by large molecular machines called **ribosomes**

- large and small subunits = ~82 proteins (1/3) plus **4 ribosomal RNA (rRNA)** strands (2/3)
- overall structure and catalytic activity dictated by RNA → **ribozyme**
- huge! ~several million Daltons (vs 30k for a typical protein)
- Nobel prize in 2009!

General process for synthesis of proteins

- binding sites for mRNA and three tRNA amino acid carriers in ribosome
- correct tRNA binds at A site through base pairing with mRNA
- high-energy covalent bond attaching amino acid is added to growing chain
- ribosome shifts over one tRNA unit, placing the tRNA in the P and then E site

- synthesized protein exits the **ribosome exit tunnel**, opposite the mRNA
- about 2 AA/seconds in eukaryotes (20 AA/seconds in some bacteria)

Initiation of protein synthesis

- all proteins begin with a methionine → **start codon** that signals initiation of protein synthesis (methionine typically removed in post-translational processing)
- **initiation factor tRNA** binds first with small subunit to mRNA
- large subunit then binds
- synthesis continues until a **stop codon** is reached, which bind **release factor** proteins

Many proteins can be made at once from the same RNA transcript by having multiple ribosomes bound to it → **polyribosomes**

After synthesis

Proteins can fold as they are exiting the ribosome tunnel, but can also receive the help of **chaperones** to fold without danger of aggregation

Many proteins undergo **post-translational** modifications

- disulfide bond formation
- phosphorylation
- binding of small molecule **cofactors**
- association with other protein subunits into large functional structures
- **glycosylation** = addition of sugars to the surface to create glycoproteins

The protein life cycle ends with degradation into constituent amino acids that can be used to build new proteins

- **proteases** are proteins that degrade other proteins
- the **proteosome** is a large cylindrical protein complex that is responsible for degrading most proteins in eukaryotic cells in its interior

- the proteosome recognizes proteins that need to be degraded because they are "tagged" with **ubiquitin** – a protein that can be attached to proteins to signal that they are destined for degradation

The origins of life?

The central dogma is so central to all living things, but one wonders how it may have evolved

Life requires both storage and replication of genetic information, and the ability to catalyze specific reactions → RNA has both of these abilities

RNA thought to be the original molecule of life, carrying both genetic info and performing chemical reactions (ribozymes)

Life then shifted to a DNA platform for the storage of the genetic information because of its increased chemical stability and double-stranded format that enables proofreading

Life then shifted to a protein platform for chemical processes → broader chemical functionality

Manipulating DNA, proteins, cells

Ability to isolate and manipulate DNA has made possible **recombinant DNA technology**, or genetic engineering

Dramatic impact on our lives

- detection of genetic disorders
- forensic science
- everyday products, e.g., laundry detergents (enzymes that digest 'dirt')
- emerging new medicines

Manipulation of DNA

- **restriction nucleases** – proteins that cleave DNA at particular locations, enabling fragmentation into smaller parts in a predictable way
- **gel electrophoresis** – separation of DNA fragments of different sizes
- **hybridization** – double strands can be separated by heating just below boiling temperature → cooling then allows re-association to correct base pairing strands

- **probe sequences** of nucleotides can be used to hybridize to certain sequences of DNA
- **recombinant DNA** can be joined to existing DNA using **DNA ligase**

Cloning DNA using bacteria

- foreign DNA can be introduced into bacterial cells (**transformation**) that then will naturally replicate this DNA alongside their own during division → typically it integrates into the bacterial genome itself by **homologous recombination**
- **plasmids** are typically used → circular DNA molecules that exist independently of the bacterial chromosome
- plasmids also occur naturally, but those carrying foreign DNA are termed **vectors**

Cloning DNA using chemistry

- **polymerase chain reaction (PCR)** is a much quicker tool for duplicating DNA without cells, developed in 1980s
- a molecule of DNA must be heated up to separate the strands
- uses a heat resistant **DNA polymerase** isolated from hot-springs bacteria to then make copies
- repeated cycles of cooling and heating then enable repeated replication
- **amplicifation** from one to billions of the same molecule

Sequencing DNA

- DNA polymerase used to make copies of a sequence
- **dideoxy DNA sequencing** → use of small amounts of dideoxyribonucleotides that terminate a growing copy (no way to add subsequent nucleotides)
- four separate experiments use either an A, T, C, G dideoxy base in addition to the four usual ones
- each experiments produces many copies of different lengths, but each terminated at a specific kind of nucleotide (either A, T, C, G)
- gel electrophoresis gives the molecular weight distribution in each of the four cases and can be used to show the sequence

- highly automated

Coaxing cells to make proteins

- **expression vectors** can be introduced with transcription, translation signals to ensure a gene (protein) is made
- introduced into cells, protein will be produced
- protein production can be monitored by making a fusion protein with **green fluorescent protein** (GFP), from the luminescent jellyfish

Manipulating the genome of existing animals

- genes can be knocked out or modified using hybridization techniques
- homologous recombination can be used to incorporate foreign DNA into a genome

Making antibodies

- immunize mouse with antigen
- mouse produces antibodies-manufacturing B-lymphocytes (white blood cells)
- hybridoma cells are made by fusing the B-lymphocytes with a myeloma cell (cancerous white blood cells)
- resultant cells produce antibody but also divide indefinitely (property of cancerous cells)
- culture in appropriate medium, grow cells
- lyse the cells
- fractionation, centrifugation

Phage display technologies used to find proteins that bind to a specific target molecule

- **phage =** bacterial virus = DNA inside protein coat
- phage DNA designed to encode for a particular protein in its coat, i.e., so that the protein is replicated on the surface of new copies of the phage
- amplification in E coli and random mutagenesis to build a library of different phages encoding for different proteins
- exposed to immobilized target binding molecule (e.g., protein or DNA)

- wash → binding proteins are kept
- re-infect E coli to amplify DNA sequence
- repeated rounds involving different levels of mutagenesis will **sort** to find proteins with good binding properties
- equivalent to a form of directed evolution on the lab bench

Protein purification using **column chromatography** uses the properties of solid matrix (beads) to selectively separate based on specific properties:

- size – gel filtration
- charge – ion exchange
- hydrophobicity
- affinity – specific interactions (adsorbed proteins, ligands, etc)