Week 2

DNA

2.1 DNA Synthesis and Transcription

10/4:

- DNA binding proteins.
  - Interact with major grooves of DNA to achieve sequence specificity.
  - Example: Transcription factors that have to turn a gene on or off.
  - Such proteins often do this with two primary motifs: The leucine zipper and the zinc finger.
    - Leucine zipper: Less programmable.
    - Zinc finger: More programmable. Contains 1+ zinc ion coordinated by cysteine and histidine. One zinc finger interacts with three base pairs.
  - With the scale of our genome, you typically need a sequence of 15-18 base pairs to achieve specificity. That's 5-6 zinc fingers.
  - When they were discovered, zinc fingers were thought to be a possibility for genome editing.
- DNA structure and binding modes.
  - DNA binding proteins may also interact with the minor group (to achieve some specificity), electrostatically with the phosphate backbone (usually not sequence-specific), and intercalation (a flat molecule splits two base pairs a bit and inserts itself in).
- Minor groove: Deep and narrow.
  - Minor groove binding can involve electrostatics, hydrophobic burial, and hydrogen bonding.
  - Minor groove binding can be site specific. There are some features you can take advantage of, e.g., being flat and flexible.
- Pyrrole-Imidazole-Hydroxypyrrole (Py/Im/Hp) Polyamides.
  - Pioneered by the Dervan lab at CalTech.
  - Minor-groove binding polyamides consisting of three aromatic ring amino acids.
  - Flexible because of the amides.
  - Eight-ring pyrrole-imidazole polyamides achieve affinities and specificities comparable to DNA-binding proteins.
  - Cell-permeable molecules for gene-specific regulation *in vivo*.
  - Still not specific enough, though.
- Minor groove-binding small molecules.
  - Examples (not testable material): Hoechst 33258, DAPI, Distamycin, and Berenil.

    - Distamycin in the minor groove. Distamycin is an antibiotic and it fits very well into the minor groove. Preference for A/T sequences.
    - Hoechst 33258 in the minor groove. Also fits very well; uset to some extent to dye DNA, but more often in flow cytometry.
  - Common features (testable material):
    - Flat (to slip into the minor groove).
    - Small linked aromatic ring systems (to allow ring systems to make local adjustments).
    - Curved (to match curvature of the minor groove).
    - Positively charged (to interact with the phosphates).
    - H-bond donors on concave face (to H-bond with acceptors on base pairs).

- Phosphate backbone binding.
  - Driven by electrostatics.
    - Ligands that bind this way are always cationic, binding depends strongly on salt concentration.
    - Ions (Na^+, Mg^{2+}, etc.)
  - Example: Biogenic polyamines (involved in a lot of biological processes).
    - For example, putrescine is a positively charged polyamine. It is responsible for bad breath!
  - **Transfaction reagents** wrap around DNA and neutralize some of its negative charge to help it get into cells.

- Intercalators.
  - Example: Ethidium bromide (a toxic molecule used to stain DNA).
  - Features in common:
    - Extended aromatic systems (to provide extensive overlap with base pairs).
    - Electron deficient (to complement regions of high electron density in base pairs).

- Intercalation requires structural rearrangement.
  - The intercalator does disturb DNA structure a bit as it pushes base pairs farther apart, causing buckling in adjacent base pairs and a tilt in the helical axis.
  - Because of this, intercalators can alter DNA replication.

- Ethidium bromide as a DNA dye.
  - Biochemical analysis of DNA (gel electrophoresis).
    - Agarose is used for DNA strands over 300 bp. Bands greater than 100,000 bp are not resolved.
    - Concentration of agarose can be increased to create a denser matrix, or decreased to create a less dense matrix.
    - Larger molecules need a less dense matrix.
    - Because DNA is negatively charged, it migrates to the cathode (+) of the electrophoresis system.
  - Ethidium bromide is toxic: It can act as a mutagen because it intercalates double-stranded DNA and, as mentioned, affects replication.
    - If you add too much ethidium bromide into your PCR, it may not work (the polymerase may not be able to overcome it).
  - Safer options are offered by many biotech companies: sybr-green, sybr-gold, etc.
    - Tang isn't sure how much safer these actually are.

    - The common design is bigger molecules: Will still intercalate DNA, but will not penetrate the skin as easily.

- Summary.
  - Several forms of DNA/RNA (most important ones: A-RNA and B-DNA).
  - Unusual forms may also play an important role (e.g., G-quadruplex and tRNA).
    -
  - Molecules that interact with the DNA.
    - Major grove interactions are sequence specific.
    - Minor grove interactions can be sequence specific.
    - Phosphate backbone/intercalation interactions are (typically) not sequence specific.
- This is the end of the previous lecture's slides.
- Now: Replication, transcription, translation, and nucleic acid catalysis.
- Overview.
  - DNA replication in cells: DNA-templated synthesis of DNA.
  - DNA repair in cells: Enzymatic repair of DNA mutations
  - DNA transcription in cells: DNA-templated synthesis of RNA.
  - Translation: Nucleic acid catalysis (difficult) and RNA-templated synthesis of proteins.
- If you find the early topics above difficult, review the relevant sections of Nelson and Cox (2021).
- DNA is synthesized/replicated by DNA polymerase.

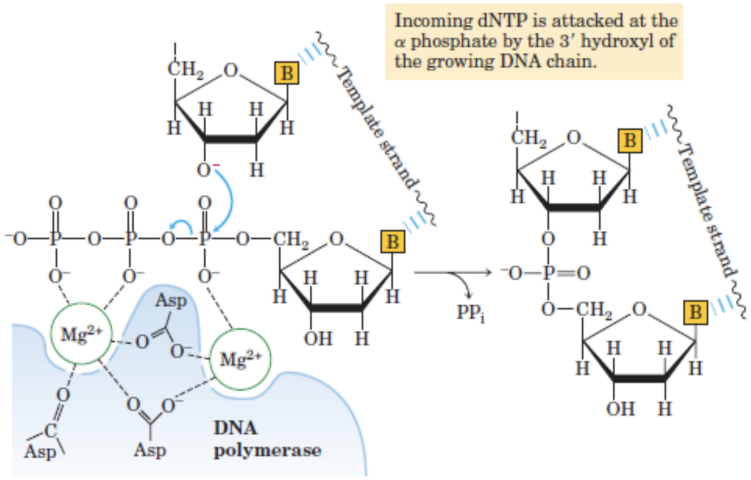


Figure 2.1: Mechanism of DNA synthesis.

  - DNA polymerases require three components: A template, a primer, and dNTPs[1] (N = A, T, G, or C).
  - Mechanism: The 3' hydroxyl of the growing DNA chain attacks the α phosphate of the incoming dNTP via nucleophilic acyl substitution.
    - Most textbooks will draw the electron pushing as a substitution reaction, but in reality, the double bond gets resolved, and then kicks back down to get rid of the β and γ phosphates.

[1] Deoxynucleoside triphosphate

    - Energetically driven by the BDE of the dNTP, so as long as dNTP is abundant, the reaction can proceed.
  - In bacteria, this process can occur at a rate of 1000 bp/second.
    - Our cells are slower.
  - Notice the presence here of magnesium (partially coordinated to aspartic acid) catalyzing the reaction as part of DNA polymerase.
    - One Mg^{2+} coordinates with the β and γ phosphates to stabilize them.
    - The other coordinates with the α phosphate to stabilize the highly negatively charged nucleophilic acyl substitution intermediate.
  - DNA replication in *E. coli* is highly accurate (error rate 10^{-9}-10^{-10}).
    - Two reasons: Templated synthesis (error rate of 10^{-4}-10^{-5} *in vitro*) and error correction mechanisms.
  - Templated synthesis is equivalent to having a 99.999% yield in an organic reaction (which never happens).
  - One error-correction mechanism bridging the gap from 10^{-5} to 10^{-10}: DNA polymerase "proofreads" even as it synthesizes DNA.
  - Example procedure:
    - Let C* be a rare tautomer of cytosine that pairs with A and is incorporated into the growing strand.
    - Before the polymerase moves on, the C* reconverts to C and is now mispaired.
    - The mispaired 3'-OH end of the growing strand blocks further elongation. DNA polymerase slides back to position the mispaired base in the $3' \rightarrow 5'$ exonuclease active site.
    - The mispaired nucleotide is removed.
    - DNA polymerase slides forward and resumes its polymerization activity.
  - Not every polymerase has this feature, but most high-fidelity ones do.

- DNA replication in cells.
  - DNA replication is semiconservative.
    - Meselson-Stahl experiment, "the most beautiful experiment in biology."
  - DNA replication begins at an **origin** and proceeds **bidirectionally**.
  - Bacterial chromosomes have a single point of origin; most other cells have multiple such points.

- DNA replication requires many enzymes and protein cofactors.
  - DNA replication in cells requires much more than solely polymerases.
  - DNA replication in *E. coli* requires 20 or more different enzymes and proteins, each performing a specific task.
  - DNA replicase system (replisome): The entire complex is required for DNA replication.

- Three identifiable phases of DNA replication in *E. coli*.

  1. Initiation.
     - Five repeats of 9 bp (R sites) for DnaA binding; A = T rich DNA unwinding element (DUE).
     - DnaA binding, DUE denaturing, DnaB helicase loading, then ready for the next phase.
     - Initiation is the only phase of DNA replication that is known to be precisely regulated (replication occurs only once in each cell cycle). You don't want the daughter cells to have multiple unneeded copies of the genome.
  2. Elongation.

  - Single-stranded DNA-binding protein (SSB) stabilizes the regions denatured by helicase.
  - Two distinct but related operations: Leading strand synthesis and lagging strand synthesis.
  - Lagging strand synthesis requires RNA primers (synthesized by primase) to form Okazaki fragments.
    - Regardless of whether it's DNA- or RNA-templated synthesis, it's always primed.
  - After completion of an Okazaki fragment, RNA primer is removed and replaced with DNA by DNA polymerase I, and the remaining nick is sealed by DNA ligase.

3. Termination.
  - Ter sequences: Trap for the bidirectional replication of DNA by forming Tus-Ter complex — prevent overreplication by one replication folk when the other folk is abnormally delayed or halted.
  - **Catenane** formation: Bidirectional replication of DNA meets at the end.
  - DNA topoisomerase IV: Separate the catenated chromosomes into normal chromosomes for the daughter cells.

- **Catenane**: A mechanically interlocked molecular architecture consisting of two or more interlocked macrocycles[2].

- We don't have to remember every protein; Tang just wants to show us.
  - Note that the ones that she did show us, though, are all involved in elongation; initiation and termination require additional proteins.
  - Also, new proteins are still being discovered.
  - DNA replication in eukaryotic cells is both similar and more complex than in *E. coli*.

- Mutations happen constantly in DNA.
  - A race between mutation and repair.
  - Why mutations don't generally affect us: Mutations could occur in DNA that is not used in a given cell (most DNA in any given cell is dormant), the cell could die, etc. Also, only 1% of our genome is protein-coding. And most amino acids in our proteins are not fully **conservative**. Significant ones are very rare (and usually lead to apoptosis anyway, so no problem).
  - Transition (purine to purine, or pyrimidine to pyrimidine), transversion (purine to pyrimidine or vice versa), and frameshift (insertion/deletion by $3n \pm 1$) mutations.
    - Frameshift typically corresponds to early stop codon.
  - Mutation locations and effects.
    - Promoter: Reduced or increased gene expression.
    - Regulatory sequence: Alteration of regulation of gene expression.
    - $3'$ of protein-coding region: Defective transcription termination or alternation of mRNA stability.
    - Certain locations within intron: Defective mRNA splicing.
    - Origin of DNA replication: Defect in initiation of DNA replication.
  - Many disease-causing mutations in humans are non-coding.
  - Mutations in one place can interact with other base pairs a few away because they may be close in the 3D structure of DNA. Tang studies this and other noncoding mutations.

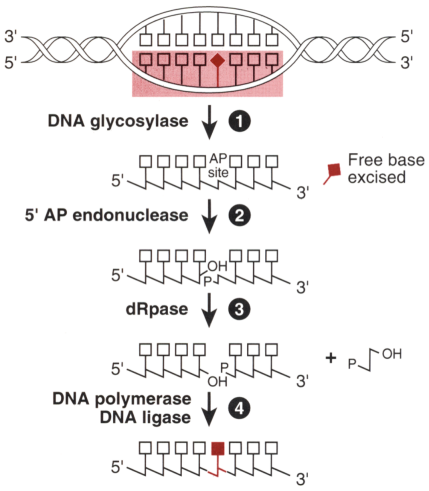
- **Conservative** (amino acid): An amino acid in a protein, the identity of which is critical to the form and/or function of the full protein.

- The causes of DNA mutations in cells.

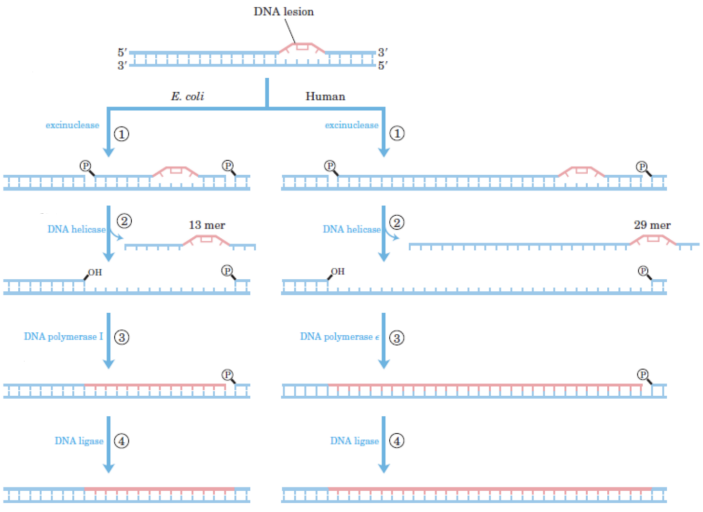
[2] Recall the discussion of this in *The Knot Book*!

  - Natural mismatching and tautomerization — know this!
    - Natural mismatching-induced mutation: 10^{-9}-10^{-10}.
    - Tautomerization ($< 0.01\%$ frequency): Mutation if the rare tautomer is paired during DNA replication.
  - Deamination of exocyclic amines[3] (C, A, G) — know this!
    - Adenine to hypoxanthine.
    - Guanine to xanthine.
    - Cytosine to uracil (500 times per day per genome, which is a significant amount). Hence T in DNA but U in RNA.
    - Cells that are uracil N-glycosylase deficient (ung^-) show a higher rate of transitions.
    - Note: Deamination of A is more common in single-stranded DNA, but deamination of C is exponentially more common in double-stranded DNA.
  - Depurination (A, G).
    - Protonation of purines can lead to cleavage of the glycosyl bond (creating an abasic site).
    - Abasic sites undergo a retro-Michael-like reaction leading to a phosphodiester bond cleavage. $t_{1/2} \approx 400\,\text{h}$ at $37°\text{C},\ \text{pH} = 7$.
    - Mammalian cells can lose as many as 10,000 purines per cell per generation ($k = 3 \times 10^{-11}$ per second at $37°\text{C},\ \text{pH} = 7$).
    - Depyrimidination occurs 20 times slower.
  - Oxidants, radicals, radiations.
  - Chemicals: Alkylating reagents, nucleophiles, crosslinking reagents, and intercalating reagents.

- The reactions of OChem III are not the focus of this course! Tang may occasionally reference such content, but it will be minor and not (directly) tested.
- It may seem like our DNA lives a hard life, but in practice, our genome is very stable.
- Four strategies of DNA repair in cells.

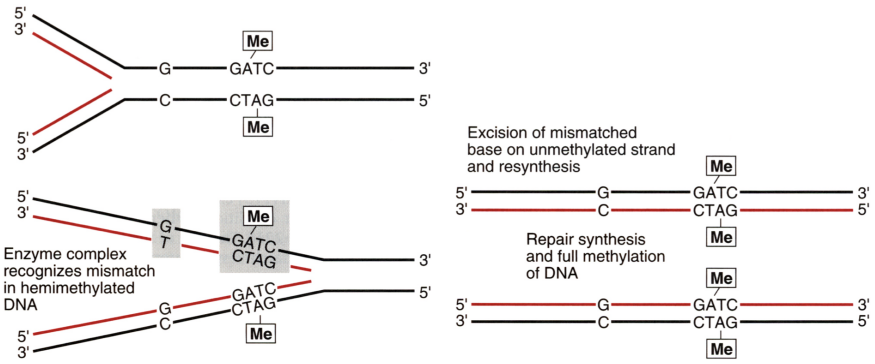


(a) Base excision repair.



(b) Nucleotide excision repair.

[3]Literally: Getting rid of the amine group which lies outside the ring and replacing it with a carbonyl group.



(c) Mismatch repair.

Figure 2.2: DNA repair strategies.

– Direct reversal/repair (DR): Enzymes catalyze the reverse reaction; no removal or replacement of the base is needed.
  - The detailed mechanism of DNA phytolyases is not testable material.
– Base excision repair (BER): DNA glycosylases cleave the ribose-base bond to produce apurinic/apyrimidinic (AP) sites.
  - See Figure 2.2a.
  - Procedure:
    1. DNA glycosylases hydrolyze the N-glycosyl bond of damaged bases.
    2. This creates an "AP" site.
    3. AP endonucleases recognize the AP site and hydrolyze the phosphodiester 5′ or 3′ of each AP site (mostly 5′).
    4. Exonucleases remove the backbone at free ends.
    5. DNA polymerase and ligase fill in and seal the gap.
  - Most DNA glycosylases recognize a specific damaged base.
  - In general, < 30 kD monomeric proteins.
  - No requirement for cofactors.
– Nucleotide excision repair (NER): Enzymes remove a segment of DNA including the lesion and several nucleotides on either side.
  - See Figure 2.2b.
  - Create nicks at two sites. Remove the DNA with lesion. Fill the gap with polymerase. Ligation via ligase.
– Mismatch repair (MR): A subset of BER and NER systems that can discriminate an improper base among two normal nucleotides forming a non W-C pair.
  - Most repair mechanisms are good for recognizing obvious abnormalities (e.g., uracil in DNA, paired pyrimidines, etc.). MR deals with cases when you have two pairs that don't match and it's not immediately clear which is the error.
  - Origin of mismatched (non-W-C) natural DNA base pairs:
    - ➢ DNA polymerase errors: 10^{-4} (intrinsic) $\times$ 10^{-3} (proofreading) $= 10^{-7}$ per base per generation.
    - ➢ Heteroduplex DNA arising from homologous recombination.
    - ➢ Deamination of 5-Me-C to T (forming G:T pairs).
  - The challenge in repairing a mismatch is distinguishing the "incorrect" base among two natural bases.

- Methyl-directed MR.
  - See Figure 2.2c.
  - *E. coli* methylates N^6 of A in GATC ("dam" methylation).
  - Methylation lags behind DNA replication (which always makes non-methylated DNA).
  - 1976: B. Wagner and M. Meselson hypothesized that the lack of methylation in a newly synthesized strand allows strand discrimination during mismatch correction.
- Experimental support:
  - No PCR, none of today's routine bio experiments were available in the 1980s.
  - The key experiments: Introduce into cells hemimethylated heteroduplex DNA and allow mismatch repair to take place.
  - This occurred even if the nearest methylation site was > 1000 bp from the mismatch!
  - Neither strand methylated: Correction of either strand.
  - One strand methylated: Correction of unmethylated strand.
  - Both strands methylated: Slow correction of either strand.
- Current MR model (not testable):
  - MutS binds the mismatch or frameshift loop.
  - MutS/L/H complex brings the mismatch and GATC together.
  - MutH nicks the nonmethylated strand 5′ of the GATC.
  - ExoVII or RecJ degrades 5′-3′ from GATC to the mismatch or Exol degrades 3′-5′ from GATC to the mismatch.

- No translation today; will be next time. The next lecture will have less content.