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 groove (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.

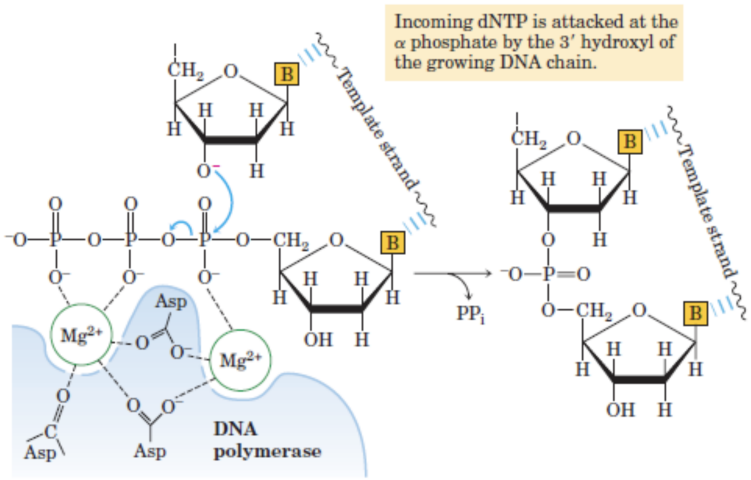


Figure 2.1: Mechanism of DNA synthesis.

  - 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.

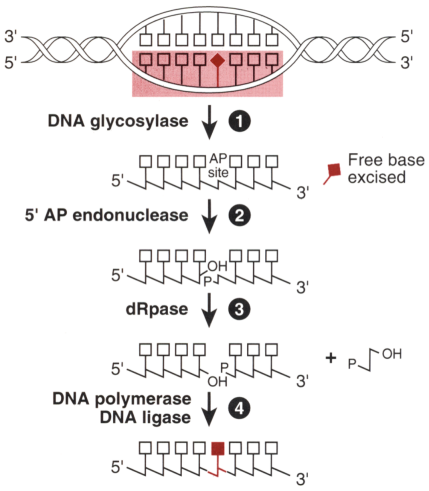
- **Conservative** (amino acid): An amino acid in a protein, the identity of which is critical to the form and/or function of the full protein.

- The causes of DNA mutations in cells.

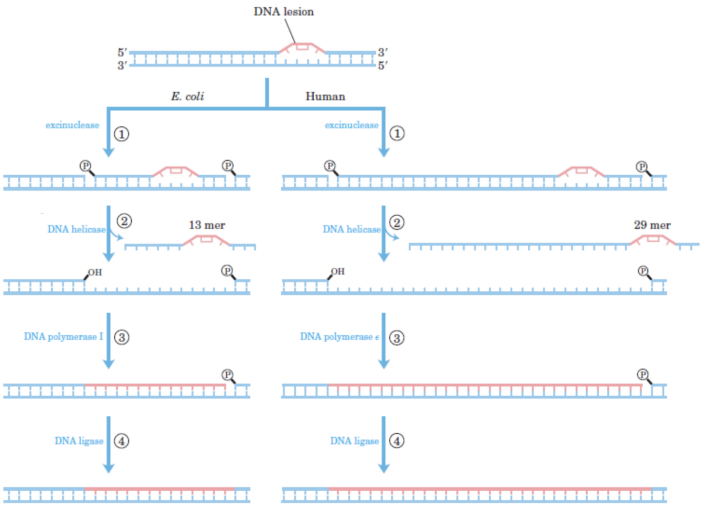
[2] Recall the discussion of this in *The Knot Book*!

  - 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.

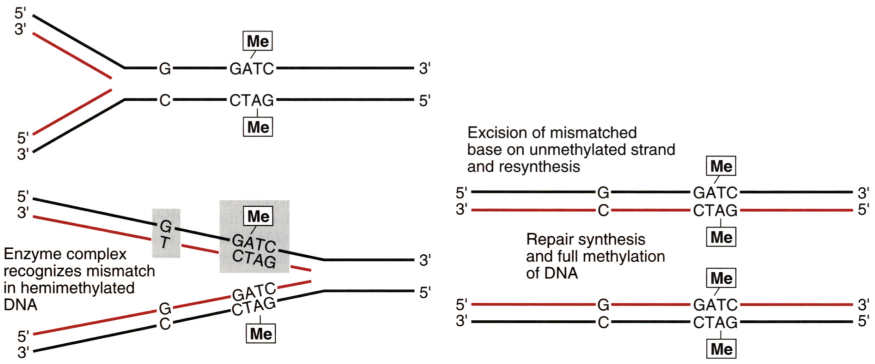


(a) Base excision repair.



(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. Chemically modifies what's already there (doesn't replace it with new bp's).
  - 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.

2.2 Chemical Modifications of DNA and RNA Bases

10/6:

- On this year's Nobel prize in chemistry.
  - Awarded for the **click reaction**.
  - This is the *key* reaction that biologists use today in their research. If you asked any chemical biologist who should have gotten this year's Nobel prize, Carolyn Bertozzi would have been on their short list.
- **Click reaction**: A reaction between an azide and an alkyne that may or may not need to be accelerated by a copper iodide catalyst (depending on how strained the alkyne is).

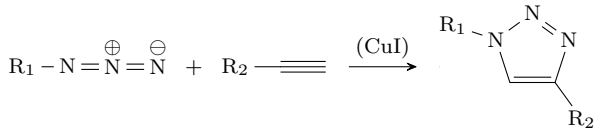


Figure 2.3: Click reaction.

  - Useful in biology because it's *orthogonal* to the entire biological system.
  - Two keys:
    1. It will occur in aqueous solution (unlike many organic reactions we learn about).
    2. It will not impact anything else in the biological system.
  - Thus, you can use it to manipulate a specified thing in your system.
- Tang cut out bioorthogonal chemistry from this year's syllabus but reconsidered the day before the Nobel was awarded; we will now have a lecture on it.

  - Tang brags about predicting Nobel prizes.
  - K. Barry Sharpless is the second chemist ever to receive two Nobel prizes. The faculty at UChicago used to debate whether or not he was worth a second Nobel; Tang bet he was, and now he is.

- On the style of this class.
  - Not as much of an attention to detail; instead of using 3 quarters to cover 1 book, we're building a bookcase. Every lecture is about a different topic, each of which has books written on it.
  - In terms of testing, Tang will not test a tiny thing on her slides; it's more about the concept.
  - She wants us to know that such research exists so that we know to read more about it if we ever need to in our research.

- **DNA transcription**: A process that passes genetic information from DNA to RNA.
  - RNA is synthesized by RNA polymerases using DNA as a template and NTPs as building blocks (instead of dNTPs).
  - RNA polymerases do not require a primer (vs. DNA polymerase).
  - RNA polymerases also elongate RNA in the $5'$-$3'$ direction.
  - RNA polymerases lack proofreading mechanisms (vs. DNA polymerases).
    - Synthesis is fairly accurate, a mistake here will likely not be repeated, and mRNA doesn't really matter.
  - Defining the template and nontemplate strand: The template strand does all of the heavy lifting/is directly involved in synthesis. The nontemplate/coding strand is what's replicated (i.e., what gets "all the publicity").

- RNA synthesis begins at promoters.
  - Similar to DNA synthesis, initiation is what's most controlled. The speed of transcription is determined by how strong the promoter, i.e., how strongly RNA polymerases are attractedl
  - RNA polymerase binds to specific sequences in DNA (promoters), which direct the transcription of adjacent segments of DNA (genes).
  - Consensus sequences in promoters: Affect the efficiency of RNA polymerase binding and transcription initiation.
  - Promoter sequence establishes a basal level of expression that can vary greatly from one *E. coli* gene to the next.
  - This bacterial example is completely different from how eukaryotes operate.

- Ribosome binding site (prokaryotes)/Kozak sequence (eukaryotes) is a **promotor** at the start of RNA that binds it to the ribosomer. There's also a **terminator** site.

- Tang is skipping the historical exploration of translation.

- Translation.

- Nucleic acid catalysis.
  - Some RNA can actually catalyze chemical reactions.
  - Ribozymes were a hot topic in the 80s and 90s.
  - Why study ribozimes?
    - Implications for the origin of life: Prebiotic soub to RNA to proteins to simple life (links *amplifiable information* to *function*).
    - RNA and proteins were a chicken-and-egg problem; this discovery suggests that RNA came first.

  - We haven't found natural examples of nucleic acid catalysis yet; all known examples were developed in the lab.
    - Tang suggests there may be some examples in basic forms of life.
  - How do nucleic acids catalyze reactions compared with proteins?
    - All known all-RNA catalysts in nature accelerate phosphoryl transfer reactions (forming or breaking phosphodiester bonds).
  - More info in slides.

- **Ribozyme**: Catalytic RNA; short for ribonucleic acid enzyme.
- **Aptamer**: A receptor; the analogous function in proteins is antibodies (binding but not causing a reaction).
- *Tetrahymena* Catalytic RNA.

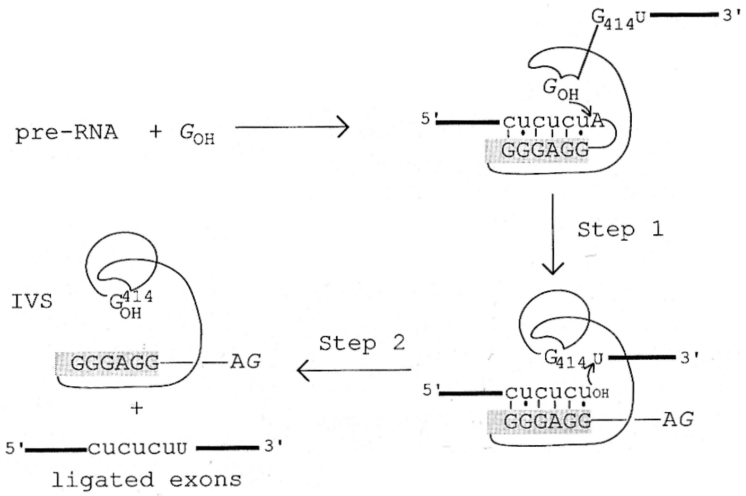


Figure 2.4: Catalytic RNA.

  - Discovered 1981; Nobel Prize 1989.
  - *Tetrahymena thermophilia* is a highly heat resistant form of life. Their pre-ribosomal RNA (pre-rRNA) can catalyze its own splicing, yielding an interveining sequence (IVS) and a mature rRNA via two phospho-transesterifications.
  - More details on the mechanism:
    - Catalyzed by GTP (or GMP, but slower this way).
    - The RNA folds, engages GTP. This enables GTP to insert itself, cleaving the RNA. Now there is a sequence that is only attached to the rest of the RNA by hydrogen bonding.
    - The partially free RNA pice adds more of the original RNA to itself and finally detaches.
  - Tom Cech (discoverer) believed that protein was catalyzing the reaction.
    - Tried denaturing enzymes and heat, but the reaction still proceeded. Strongly suggested it wasn't a protein, but couldn't confirm it wasn't a heat-resistant or otherwise very stable protein (or trace amounts of a highly efficient protein catalyst).
    - Final clue: Went to an *in vitro* system. Did *in vitro* transcription to get the RNA and RNA only (the material is no longer distilled from the organism), dumped that into the reaction, added the substrate, and watched it occur.
    - Took them a year.

- *Tetrahymena* Ribozyme Structure.
  - The X-ray crystal structure was solved to a decent resolution (2.8 Å — atomic resolution) for one domain of the catalytic core, and to 5 Å for the entire core.
    - Nowadays, you will not be able to publish resolution as low as 5 Å.
    - This was Jennifer Doudna's first paper as an independent researcher!
  - Combination of structural and biochemical studies suggests a mechanism mediated by several bound magnesium cations.
    - Emphasizes that nucleic acid reactions often need to be catalyzed by ions.
  - Take home lesson: RNA can fold into compact, protein-like structures.
- Structure of (most of) the ribosome.
  - Nobel prize (2009) — many scientists tried and failed for years, but they finally got it in 2009.
  - The ribosome is the cell's way of converting genetic information into molecular structure and chemical function.
  - Bacterial ribosome: Huge — 2.6 million Da (far bigger than glycosylase), 2/3 RNA (3 total), 1/3 protein (55 total), two subunits (50S and 30S).
    - You can delete some of the proteins and it will still function, but you cannot delete any of the RNA.
- Video that Tang saw as a grad student that really impressed her and she wants to share with us (link).
- The ribosome is a ribozyme.
  - The reaction that the ribosome catalyzes is carried out almost entirely by RNA (that's what the ribosome active site is made of).
  - Details of the protein building reaction.
    - There are three sites in the ribosome: The exit, peptodyl, and aminoacyl site. The tRNA comes in at the A site and exits at the P site. At the E site, the tRNA has already been utilized.
    - The incoming amine group does a nucelophilic attack on the tRNA-bound ester group of the amino acid added just before.
    - This kicks out the ester-bound tRNA, and everything shifts down a site.
    - A new codon is exposed at the aminoacyl site, and a new tRNA plust amino acid binds to it.
    - This process repeats over and over again, 3 RNA at a time, building a longer and longer peptide chain.
  - A transition state mimic that scientists use to get the crystal structure of the active state of the ribosome is CCdAp-Puromycin.
    - Puromycin is a useful antibiotic used to inhibit protein synthesis.
    - Puromycin doesn't break, so we can make the ribosome get stuck in the transition state.
  - Tang goes over the electron pushing of the protein building reaction, as seen in Figure **??**.
- Mechanism of the ribosome.
  - Recall from lecture 2 that at physiological pH, no nucleobases are charged. You have to go to pH $\approx$ 3 for protonation or pH $\approx$ 10 for deprotonation. This implies that nucleobases are really terrible acid-base catalysts.
  - Yet we do have a proton transfer occuring from the incoming amino acid to the exiting tRNA.
  - The XPS crystal structure implies that A2486 catalyzes the proton-transfer reaction.
  - N1 is the site on adenine that can most easily be protonated, but N3 is closest to the carbonyl O and the incoming amine, so it is active as the catalyst.

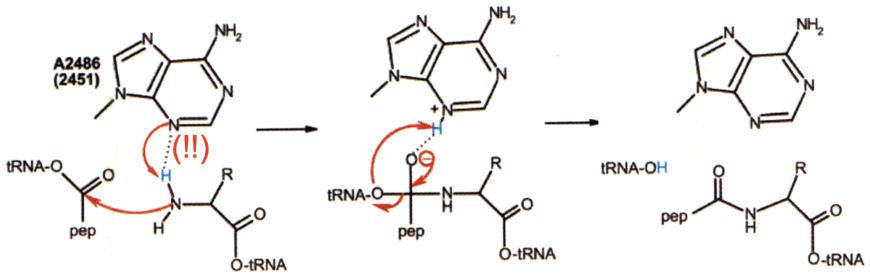


Figure 2.5: Mechanism of the ribosome.

  - But N3 has p$K_a \approx 1.5$, implying that if you have an isolated adenine, you need to drop the pH to 1 before you can get protonation.
  - How does this work? We have a complex hydrogen-bonding network that delocalizes a negative phosphate charge to a neighboring guanine that, in turn, passes it to A2486 to aid in its deprotonation effort. The effect is that actual pK_a of AdeN3 is 7.6, up six orders of magnitude due to the H-bonding interactions.
  - If tested on this, we'll be given information on the hydrogen bonding network. We likely won't be tested on it, though.

- Slides on RNA ligase will not be tested: It's one of Tang's favorite topics, but it falls under directed evolution.
- Now for lecture 4 content.
- Will modification always change DNA or RNA always affect W-C interactions?
  - The most frequent modification (5-methylation) does not change base pairing, but it does affect interaction with proteins.
  - Today: Modifications that show up naturally but don't necessarily cause modifications.
  - This lecture will be shorter: Tang shoots to finish today.
- Overview.
  - Diverse natural base modifications in DNA and RNA and their biological functions.
  - Epigenetics: All cells have the same set of DNA, but different cells can behave very differently. This is caused by epigenetics.
  - Epitranscriptomics describes modifications on mRNA. One example: N^6-methyladenosine (m^{6}A).
  - If time allows: The arms race of base modifications in bacteria and phages.
- Natural DNA base modifications.
  - Most installed by enzymes (sometimes, phages directly use modified dNTP to synthesize their genome, but that's beyond the scope of this class; for now, most means always).
  - Not always required for survival, but can lead to an evolutionary advantage (recall from last time the example of mismatch correction based on methylation; bacteria that can't do this have a much higher mutation rate).
  - Modifications occur at specific locations on the four cananonical bases.
    - Adenine: C2 and N^6.
    - Cytosine: C5 and N^4.
    - Guanine: N7.
    - Thymine: C^5.

    - Exceptions exist, but we won't discuss them.
  - 1.5% of our genome (5% of cytosine) is 5-methylcytosine.
  - 5-(hydroxymethyl)cytosine, 5-formylcytosine, and 5-carboxycytosine are also possible in humans, in decreasing frequency.
  - Other stranger modifications (such as bonding sugars to C5) can occur in lower organisms.
  - Uracil can also be hydroxymethylated and formylated. Base J is uracil with a sugar at C5.
  - N^6-methyladenine is very abundant in bacteria, but there is a huge contraversy over whether or not it is in humans.
  - We don't need to memorize any of these save 5-methylcytosine.
- Detecting DNA modifications: Use LC-MS/MS.
  - Harvest the DNA, digest it into individual nucleotides, get rid of the phosphate, shoot it into a mass spectrometer, and see if you can detect the modified base.
  - Restriction: Rare modifications can fall below the detection limit.
  - Point of contraversy: The second and third steps above are accomplished using enzymes from prokaryotes, but these can leech bacterial DNA nucleotides.
    - Errors regarding this can account for some of the false positive detections of N^6-methyladenine in eukaryotic DNA.
  - We will see beter methods of sequencing later.
- Epigenetics and DNA methylation.
  - Epigenetics is the study of heritable phenotype changes that do not involve alterations in the DNA sequence. Two big areas: Modification of DNA and modification of histone proteins.
  - 5mC is the "5th richest" base in human DNA. Happens primarily within CpG islands in promoters. Most often correlated with gene suppression.
- DNA methylation is dynamic.

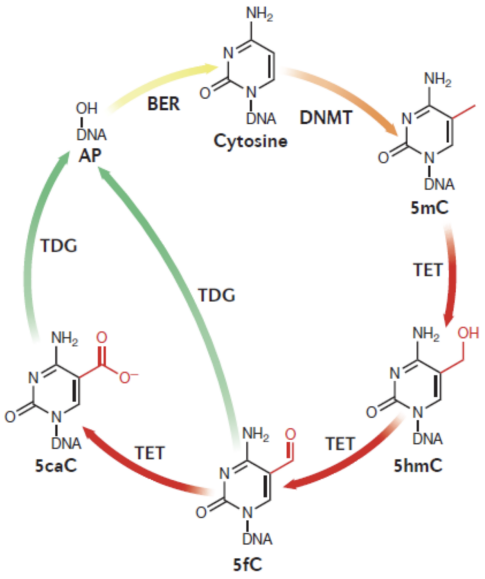


Figure 2.6: Active methylation cycle of cytosine.

  - Two ways to get rid of modified bases:

    - The passive way, i.e. if we never replace the methylation marks after DNA replication (recall from the discussion associated with Figure 2.2c that newly synthesized DNA is unaltered). If modifications are never regenerated, they will slowly become less and less common, only occuring in the original strand that has now been duplicated and "diluted" many times.
    - Active demethylation: Catalyzed by the TET family in eukaryotes.
  - 5mC and 5hmC have different effects on gene transcription (5hmC is more activating than repressing).
  - 5mC and 5hmC are viewed as modifications; 5fC and 5caC are viewed as lesions and will be fixed. We will not be tested on this, though.
  - Histone marks can have very different functions (acetylation vs. methylation).
  - Epigenetics is a huge research field and waiting for a Nobel prize.
- 5mC/5hmC in early embryonic development.
  - Two scenarios: Skin/liver cells are still dividing but are at their terminal epigenetic state; a fertilized egg is still diversifying. The fertilized egg has more motivation to change its epigenetics.
  - Thus, throughout development, you see a quick decrease in 5mC and some waves in 5hmC.
- DNA methylation on aging and cancers.
  - Not tested, but interesting.
  - DNA methylation maps change with chronological age. When you are born, you have the most beautiful epigenetics; it gets messed up as you age.
- How to detect 5mC/5hmC sites in DNA?

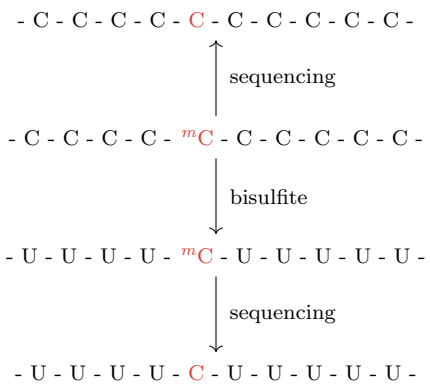


Figure 2.7: Bisulfite chemistry.

  - Bisulfite chemistry.
    - One of Tang's favorite topics in biochemistry; will definitely be tested.
  - When you mix DNA with bisulfite and heat it up, cytosine converts to uracil. Since uracil pairs with thymine, you will detect thymine when you should detect guanine if you do the bisulfite treatment.
  - If you have 5mC, bisulfite won't attack for steric reasons, so 5mC remains unchanged.
  - Thus, between the original strand and the bisufited strand, you get differentiation.
  - Whichever cytosines don't change before and after bisulfiting are your methylated C's.
    - Notice how in Figure 2.7, the only cytosine which doesn't change in between the two rounds of sequencing is the methylated one, indicated in bright red.

  - Assuming the reaction yield of bisulfite chemistry is 100%. What if the yield is 50%? We are lucky here: Bisulfite chemistry is 99.9% efficient, so the number of false positives is very low.

- If you want to detect beyond 5mC, there are more complex methods; she won't test us on these though.
- Natural RNA base modifications.

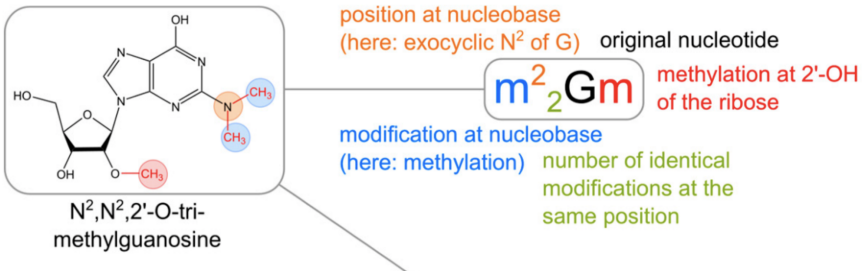


Figure 2.8: Naming convention for RNA base modifications.

  - If you compare between DNA and RNA base modifications, you will find DNA boring.
  - There are 20 known DNA base modifications; there are over 150 known RNA ones, and many look very weird.
  - Naming convention for base modifications: See Figure 2.8.
  - RNA base modificatinos occur in all three major RNA species (tRNA, mRNA, and rRNA) and in other RNA species such as snRNA and miRNA.
  - They are found in all three domains (archaea, bacteria, and eukarya). Some modifications are unique to a single domain.

- tRNA is heavily modified.
  - $> 75\%$ of RNA modifications are present in tRNA.
  - tRNA are typically less than 100 nucleotides long, so the density of modifications is very high.
  - These modifications enhance translation.

- **Transcriptome**: The set of all RNA within a cell.
- **Epitranscriptome**: The set of all biochemical modifications to all RNA.
- Epitranscriptome and mRNA methylation.
  - Epitranscriptomics defines the half-life of RNA and determines how strongly mRNA gets translated.
  - In many papers about base modifications, it's evident that the figures are not drawn by chemists (there are obvious mistakes such as missing charges, wrong atoms, etc.).