CHEM 23300 (Introduction to Chemical Biology) Notes

Steven Labalme

October 13, 2022

Weeks

List of Figures

Week 1

The Molecular Level

1.1 Chemical Biology Introduction and the Central Dogma

9/27:

- Questions:
  - What edition(s) of the textbook(s) should we have?
    - Doesn't matter.
  - Will there be TA office hours?
    - No.
- CHEM 233 used to be Intermediate Organic Chemistry, and CHEM 332 was the grad class. They have been merged this year because of the overlap in content.
- Krishnan weeks 5-7; Tang otherwise.
- We will not be going through reactions. The format is slides; don't try to copy them down, just make some notes. Copy them down ahead of time!
- Goes over the syllabus.
  - No fixed textbook. Lehninger is recommended though. Whatever edition you can find.
  - No office hours (ask questions in class or ask her to meet outside).
    - Tang will show up early and stay late.
  - Midterms are 1 hour; final is 2 hours.
  - Three problem sets.
  - One in-class quiz:
    - Krishnan will give us cutting-edge literature to read one week before the quiz and 5 questions.
    - We can form study groups to discuss the questions.
    - Multiple choice quiz on that day.
  - We're not supposed to memorize things in this course; the problems won't be like that.
  - Tang may lower the exam difficulty levels from previous years.
  - Tang doesn't want us to have to fight for points; is trying to give us a big curve so that we can just focus on learning.
  - Since this is now only a twice a week class, Tang is cutting material on carbohydrates and protein design. May try to squeeze in orthogonal chemistry, though.
- The central dogma in biology. *picture*
  - DNA $\rightarrow$ RNA $\rightarrow$ protein $\rightarrow$ needed chemical transformations.

- Size in biology.
  - An activity matching biological entities (e.g., E. coli, cells, RNA) to their sizes in microns.
  - Uses the world zoom website.
  - We may be tested on sizes, but only relative not exact (e.g., E. coli vs. a ribosome).
- Red blood cells are smaller than normal cells because they don't have nuclei, and they don't need meat to divide.
- Concentrations in biology.

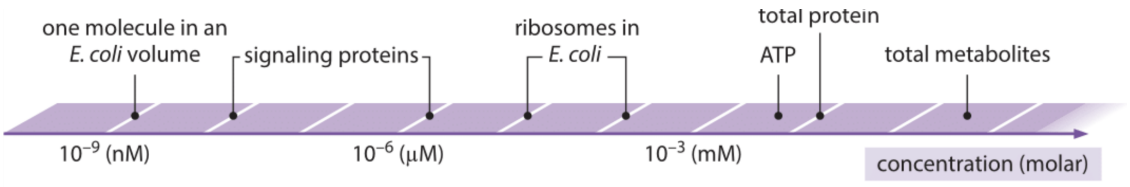


Figure 1.1: Concentrations in biology.

  - You need a couple of copies of signaling proteins.
  - Cells dedicate a lot of resources to building ribosomes.
  - Different ions have different concentrations in different parts of the body. Additionally, different types of cells have different concentrations.
- *Bound* divalent ions such as Mg^{2+} help cancel the charge of ATP; that's why we need them in solution.
- The materials left after we remove all of the water from our cells.
  - Largely protein, lipid, rRNA.
  - Far more mRNA and proteins in mammalian cells than bacterial cells.
- Time for protein diffusion within a cell.
  - Time scale τ to traverse distance R given diffusion coefficient D:

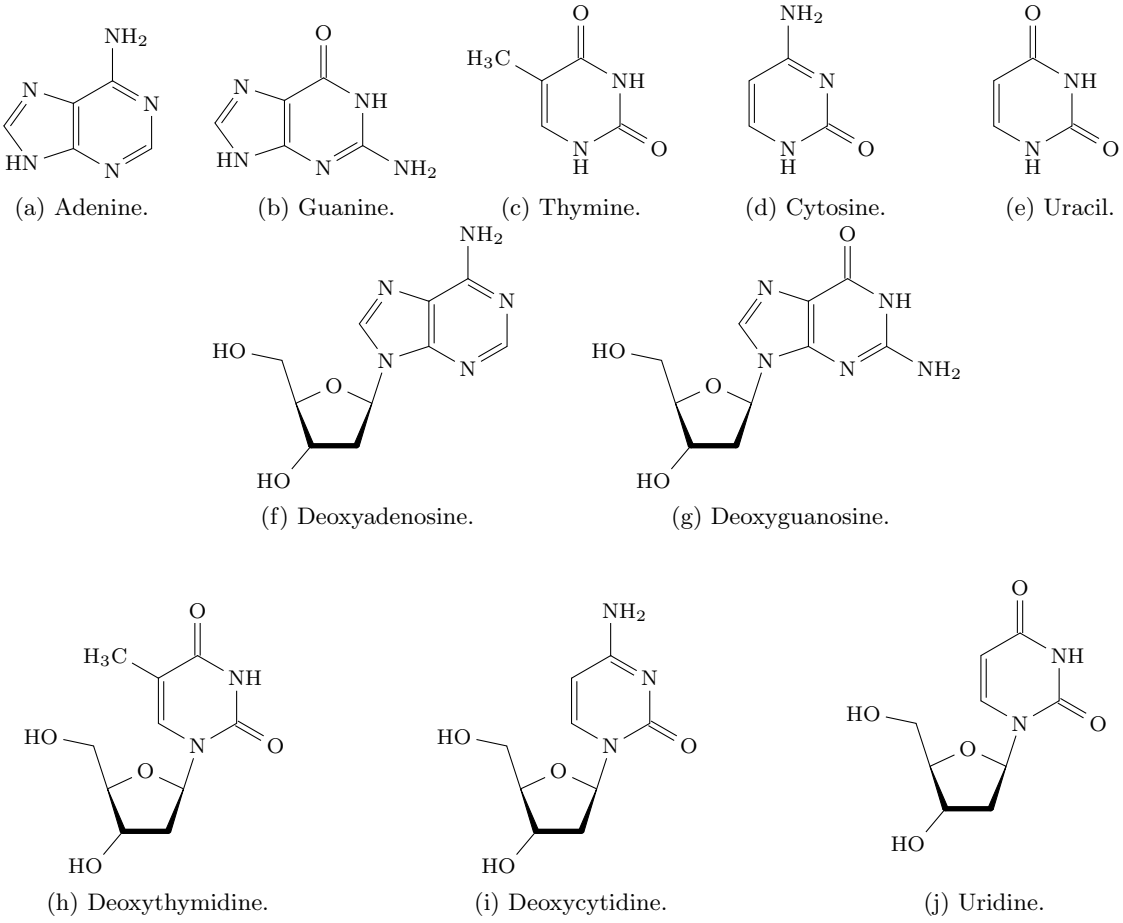
$$\tau = \frac{R^2}{6D}$$

  - For a protein in cytoplasm, $D \approx 10\,\mu m^2/s$.
- The molecular hierarchy of structure.
  - The cell and its organelles are made of supramolecular complexes (e.g., the plasma membrane, chromatin, and the cell wall), which are made up of macromolecules (e.g., DNA, proteins, cellulose), which are made up of monomeric units (e.g., nucleotides, amino acids and sugars).
- We will be expected to know how to draw the amino acids and nucleic acid bases.
- We will not talk much about lipids and sugars.
- Chirality and isomers review.
- Thalidomide.
  - Was only distributed in Germany; the FDA is very proud of having picked up on the scientific malpractice and barred it from ever entering the US.
  - Just selling one isomer doesn't work because it racemizes so quickly.
  - Now used to treat cancer; you have to sign a bunch of paperwork saying that you won't get pregnant before you use it.

1.2 Chemistry and Biophysics of Nucleic Acids

9/29:

- Feel free to come by and introduce yourself now that the class is a more manageable size.
- DNA and RNA basics.
- Humans have on the order of 3×10^{13} cells and on the order of 1 m of DNA in each cell.
  - Calculated by multiplying the number of base pairs per cell ($\approx 3 \times 10^9$) by the length of each base pair ($\approx 3.3 - 3.4\,\text{Å}$).
  - Some people say 2 m because we have two copies of our genome.
  - DNA wraps around histone proteins to form chromosomes to fit into such a tiny space.
- DNA is ACTG. RNA is ACUG.
- **Bases** and their corresponding **nucleosides**.



(a) Adenine. (b) Guanine. (c) Thymine. (d) Cytosine. (e) Uracil.

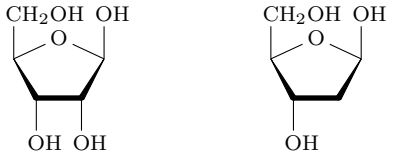
(f) Deoxyadenosine. (g) Deoxyguanosine.

(h) Deoxythymidine. (i) Deoxycytidine. (j) Uridine.

Figure 1.2: Bases and nucleosides.

  - Notice that deoxyribose is joined with the base at its $2'$ carbon.
  - If we use ribose instead of deoxyribose, we get adenosine, guanosine, etc.
  - Memorize these structures!
- Nomenclature.
- **Base**: The heterocycle. *Also known as* **nucleobase**.

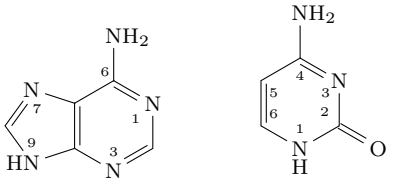
- **Ribose**: A 5-carbon monosaccharide, a derivative of which is a component of DNA.
- **Deoxyribose**: A molecule identical to ribose but without the 2′ hydroxyl group.



(a) Ribose. (b) Deoxyribose.

Figure 1.3: DNA sugars.

- **Nucleoside**: Base + sugar.
- **Nucleotide**: Base + sugar + phosphate(s).
- Base numbering.



(a) Adenine. (b) Cytosine.

Figure 1.4: Base numbering.

  - Generalize from the above two examples.
- Sugar numbering: Start to the right of the oxygen and move clockwise. Use primes do distinguish from the numbered base carbons.
  - Remember that DNA and RNA run 5' to 3' with phosphate groups linking the deoxyribose groups.
- Listing features common to all or some of the bases.
  - E.g., heterocycles, on the way to being or already aromatic, nitrogen in the ring, oxygen only ever outside the ring, etc.
- **Inosine**: An intermediate between adenine and guanine. *Also known as* **hypoxanthine**. *Structure*

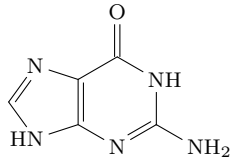
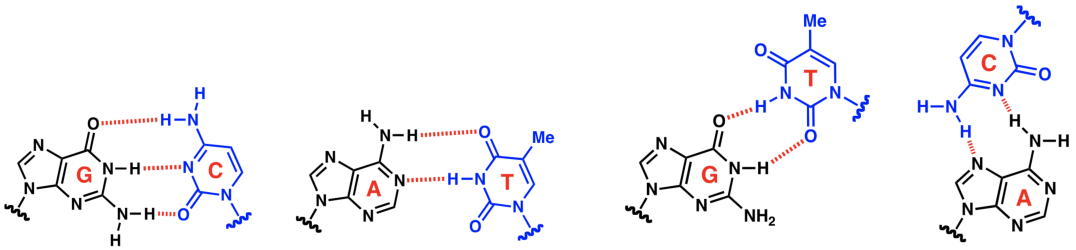


Figure 1.5: Inosine.

- Common modifications to adenine: Methylation at the 1 or 6 position.
- Five percent of cytosine exists in its methylated form; important epigenetically in determining which genes get turned on and off.
  - Methylation of cytosine occurs at the 5-carbon.

- Hydrogen bonding between bases.



(a) Watson-Franklin-Crick. (b) Wobble.

Figure 1.6: Hydrogen bonding in bases.

  - Hydrogen on a heteroatom serves as a donor; heteroatoms serve as acceptors.
  - There are many types, but we're only responsible for Watson-Crick-Franklin and Wobble interactions.
  - Watson-Crick-Franklin is the standard interaction found in double helix DNA.
  - Wobble:
    - G/T is more common and important than C/A in nature.

- A triplet codon NNN yields an amino acid. There are $4^3 = 64$ possible codons but only 20 amino acids. Thus, some codons code for the same base. For example, NNC and NNT always encode for the same base since C normally pairs with G and T can be paired with G via a wobble interaction.
  - Something about the pairing of strands of DNA with lots of G's and T's.

- pK_a review.
  - $MeNH_2$'s protonated form has p$K_a \approx 10.6$.
  - Aniline's protonated form has p$K_a \approx 4.6$ because aniline is a weaker base.
  - Pyridine's protonated form has p$K_a \approx 5$ because it is basic, but it is also sp^2.
  - An amide has p$K_a \approx 18$.
    - Did Tang switch from doing the pK_a of the conjugate acid to doing the pK_a of the molecule itself here? Why?
  - Ethanol has p$K_a \approx 16$.

- Predicting DNA ionization states.

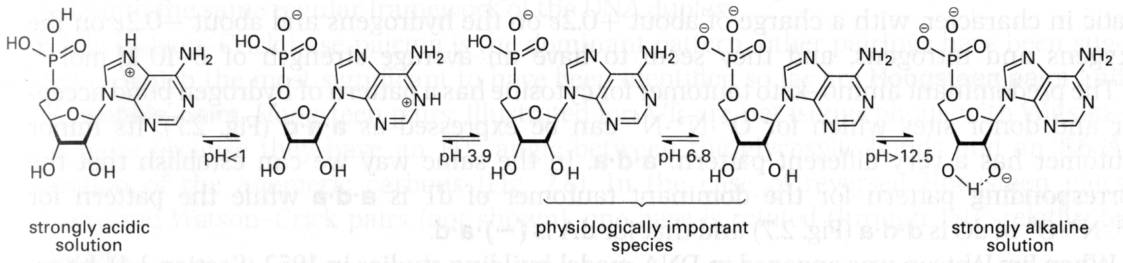


Figure 1.7: DNA ionization states.

  - At physiological pH (5-9), only phosphates are charged (as desired).

  - Phosphate pK_as: About 1-2 and 7.
  - Ribose with free 2′ and 3′ –OH groups: p$K_a \approx 12.4$ (vs. 15-16 for an isolated secondary alcohol).
  - Why are the heteroatoms on adenine that get protonated the ones that do?
  - These numbers can change a lot after polymerization.

- Anti and syn base conformations.
  - We have free rotation about the glycosidic $C^{1'}$-N bond, subject only to the whims of sterics.
  - This leads to **anti** and **syn** conformations.
  - Anti is preferred among natural nucleotides for steric reasons.
  - Exceptions:
    - G prefers syn in mononucleotides, in alternating CpGpCpG oligonucleotides, and in Z-DNA.
    - Non-natural nucleotides can shift the equilibrium towards syn.
      - Examples: 8-bromoguanosine (N^3 of the now-electron-deficient heterocycle seeks stabilization through an H-bonding interaction with the 5′ hydroxyl group, but this requires a syn conformation to be most efficient [i.e., to bring the involved atoms close together]) and 6-methyluridine (Me is more bulky than =O, so *it* sits anti to the sugar).

- **Bulk** (of the base): O^2 (the oxygen attached to the 2-carbon) in pyrimidines or the whole six-membered ring in purines.
  - See Figure 1.2.

- **Anti** (base conformation): The bulk of the heterocycle points away from the sugar.

- **Syn** (base conformation): The bulk of the heterocycle is over the sugar.

- Base tautomerization basics.

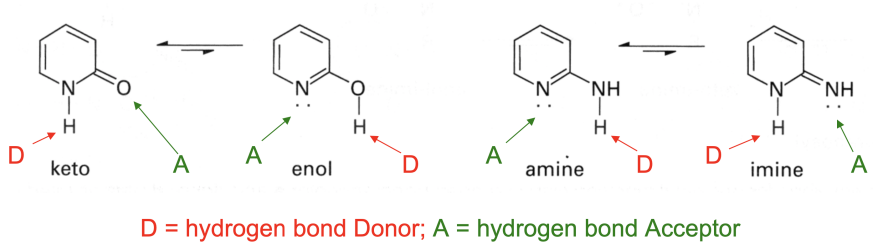


Figure 1.8: Base tautomerization.

  - Recall that tautomerization involves movement in atoms whereas resonance does not.
  - Bases exist in equilibrium between keto and enol forms, and between amino and imino forms.
  - Tautomerization changes which groups function as hydrogen bond donors and acceptors in base pairing.
  - The keto and amino forms among natural bases are preferred by more than 99.99%, according to X-ray and NMR analyses.
  - It is difficult to determine what form a base is in just via organic chemistry first principles.
    - Sometimes, tautomerization will do something highly unfavored like breaking aromaticity. But other times, making a system aromatic will generate an unstable enol. Confounding factors like this make it hard to tell.
    - In fact, when Watson and Crick were originally solving the structure of DNA, they had it backwards until a physical chemist wrote to them with a calculation suggesting the right form, and that allowed Watson and Crick to solve the structure right away.

- Enol and imino tautomers lead to mutagenic H-bonding patterns.

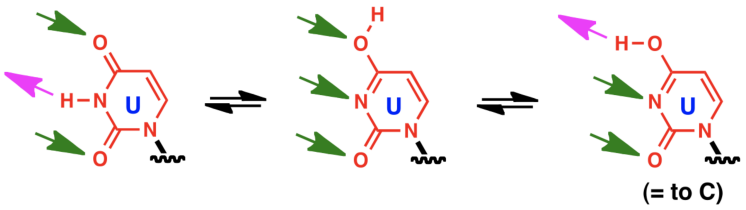
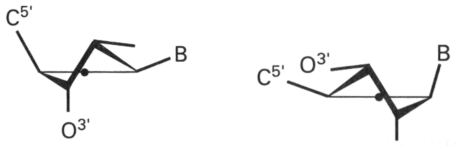


Figure 1.9: Uracil tautomerization.

  - Because donor/acceptor dynamics are shifted, tautomers of one base can look like *another* base from a hydrogen-bonding perspective.
  - The tautomerization equilibrium can be shifted by functionalizing the base pairs. This is why bromine is a mutagen — it makes it far more likely for U to be read as C, for instance.
  - Tang goes over the tautomers for the other bases, too (see slides).

- Ribose exists in many conformers.



(a) $C^{2'}$-*endo*. (b) $C^{3'}$-*endo*.

Figure 1.10: Puckomers of ribose.

  - The furanose ring is nonplanar/puckered to minimize non-bonded interactions between substituents.
  - **Endo** and **exo** atoms.
  - **Puckomers** are in rapid equilibrium with an energy barrier of less than 5 kcal/mol (higher in polymeric DNA/RNA).
  - Crystallography and NMR suggest two preferred puckomer groups: $C^{2'}$-*endo* and $C^{3'}$-*endo*.
    - In the former, the 2′-carbon of (deoxy)ribose is endo (and the 3′-carbon exo; all others lie in the plane).
    - In the latter, the 3′-carbon of (deoxy)ribose is endo (and the 2′-carbon exo; all others lie in the plane).
  - Sugar conformation dramatically changes the shape of the duplex.
    - B-DNA favors $C^{2'}$-*endo* sugars.
      - ➢ Since DNA runs 5′ to 3′, as we can see in Figure 1.10a, $C^{2'}$-*endo* gives us a more stretched out/relaxed form of the polymer. B-DNA is the most common form of DNA.
    - RNA and A-DNA favor $C^{3'}$-*endo* sugars.
      - ➢ Conversely, $C^{2'}$-*endo* gives us a much more compact/bunched up form of the polymer.

- **Endo** (atoms): Atoms on the same side of the furanose as $C^{5'}$.
- **Exo** (atoms): Atoms on the opposite side of the furanose from $C^{5'}$.
- **Puckomer**: A ribose conformer.

- Why RNA contains uracil and DNA contains thymine.
  - There is a slow but appreciable rate of hydrolysis of C to U (500 times per cell per day).
  - When C $\rightarrow$ U in DNA, because uracil is not a typical constituent of DNA and can easily be distinguished from T (T has an additional methyl group), our DNA correction mechanisms can easily repair the error, preventing our DNA from mutating long-term.
  - RNA, on the other hand, cannot be edited. However, RNA is transient but DNA is not, and it is highly unlikely to have the same mutation at the same position every time. Thus, a few proteins are liable to get messed up in a variety of different ways from mutated mRNA, but long term, the base genetic code in DNA is preserved.
  - Note that this is just a hypothesis (and a hard one to test), but it seems reasonable.
- Why nature chooses phosphates.
  - Requirements any possible linking group for nucleic acid monomers must satisfy.
    1. Multivalent (so it can connect two monomers).
    2. Cannot cross biological barriers (e.g., the nuclear membrane).
    3. Kinetically stable to hydrolysis (we don't want our DNA strand to be breaking at random all the time).
    4. Thermodynamically unstable/must exist in high energy forms (we want the synthesis of the polymer [which involves cleaving some phosphate groups] to be thermodynamically favorable).
    5. Kinetically unstable with catalyst: modulated reactivity (so an enzyme can hydrolyze it; we don't want it to be so stable that we can't work with it).
  - Phosphate groups satisfy these requirements since they...
    1. Are divalent.
    2. Are polar.
    3. Are negatively charged (nucleophiles that might hydrolyze it are Coulombically repelled).
    4. Can exist in high energy forms (such as ATP).
    5. Are more reactive in the presence of magnesium.
  - Some possible alternatives include citric acid, arsenate esters, silyl esters, and amides.
    - Citric acid is abundant, but ester bonds are unstable in biological systems and the negative charges are quite far apart (so nucleophilic attack is not as hindered).
    - Arsenate and silyl esters are also too labile.
    - Amides are too stable; we can't hydrolyze it easily with any sort of catalyst.
      - Scientists have used amides to connect nucleobases in the lab, though.
  - This is another hard-to-test hypothesis that seems reasonable.
- What binds two strands of DNA.
  - Not hydrogen bonds.
    - These only decide specificity; there is no thermodynamic preference for two-stranded DNA over single-stranded DNA hydrogen bonded unspecifically to a bunch of water molecules, for instance.
  - Stacking, on the other hand, is key.
    - It excludes water and maximizes van der Waals interactions.
    - More explanation?
    - Not testable material.
- DNA and the double helix.
  - DNA can occur in different 3D forms.

  - Nucleic acids in higher order structures (e.g., tRNA and G-quadruplex).
  - Geometric parameters.

- DNA and RNA polymorphism.
  - Various forms exist and are interchangeable; we don't need to know most of them.
  - Determinants of DNA and RNA forms.
    1. Sequence (not only composition).
    2. Counter ion and [salt].
    3. Humidity (crystals).
    4. Temperature.
  - Not testable material.

- Major nucleic acid forms.
  - We are responsible for A-DNA, B-DNA, and Z-DNA.
    - A- and B-DNA are most important; then Z-DNA.
  - A- and B-DNA are right-hand double helices; Z-DNA is a left-hand double helix.
  - The number of base pairs per turn of...
    - A-DNA is 11;
    - B-DNA is 10;
    - Z-DNA is 12.
  - The rise per base pair of...
    - A-DNA is 2.9 Å;
    - B-DNA is 3.3-3.4 Å;
    - Z-DNA is 3.7 Å.
  - Other important numbers?
  - In B-DNA, the base pairs are relatively centered within the strand; in A-form, they rotate around.

- B-DNA.
  - Base pairs on center of helical axis.
  - Major and minor is an accurate descriptor.
    - What are these grooves and what is their significance?
  - Both grooves have similar depths.
  - Sugar pucker is $C^{2'}$-*endo* (2′ and 5′ on the same side).

- A-DNA.
  - Base pairs are displaced from center of helical axis.
  - Major groove less wide than minor.
  - Sugar pucker is $C^{3'}$-*endo* (3′ and 5′ on the same side).
  - DNA/RNA hybrids are A-like (transcription, reverse transcription, and DNA replication).

- Enzymes recognize the 3D helical structure of DNA, not just their individual substrate. Like reading a word instead of letter by letter.

- Higher order structures.
  - The structure of tRNA provides a wealth of information.

    - Until the early 1990s, we could only crystallize tRNA, so we primarily learned from it for a long time.
  - L-shape: Two perpendicular A-RNA helices.
  - Tons of fun H-bonding interactions provide structure. Even three nucleobases can interact all together in some cases.

- G-quadruplex.

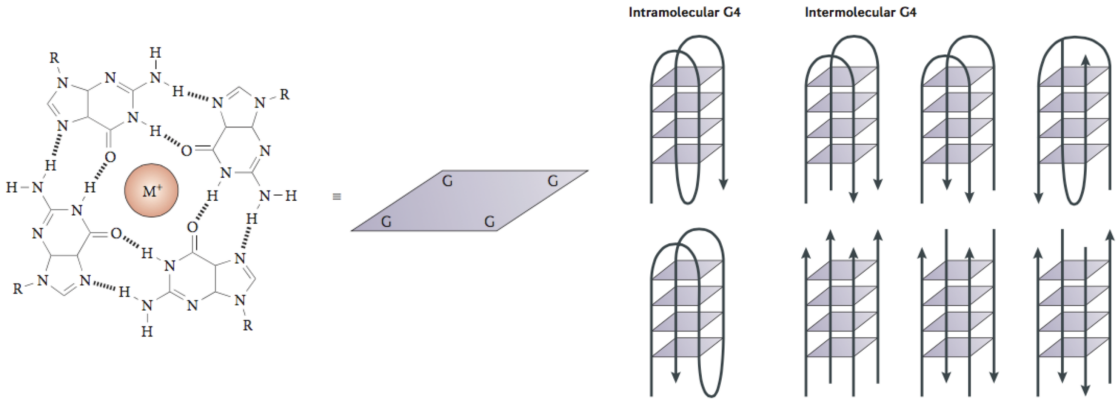


Figure 1.11: G-quadruplex structure.

  - Helical structure containing guanine tetrads from one, two, or four strands.
  - Hoogsteen hydrogen bonding.
  - Stabilized by the presence of a cation, especially potassium.
  - Importance of G-quadruplexes.
    - Chromosomes have a structure called a telomere at both ends. During replication, some of the telomere is lost each time. When the whole telomere is gone, the cell will not be able to divide any more. This is the aging process, and the discovery of telomeres received the Nobel prize in the early 2000s.
    - Telomerase is an enzyme that fights against this loss, trying to extend the DNA post-replication using RNA templates.
    - If telomerase is overactivated, the cells are immortalized and they become cancerous.
    - Telomeric quadruplexes decrease the activity of telomerase; they moderate telomerase so that it's active enough so that we don't die early, but not so active that we become big balls of cancer.

- The spinach aptamer.
  - For a long time, we've been able to tag any *protein* we want with GFP (green fluorescent protein) and follow it.
  - It would be very beneficial to be able to do the same thing with RNA.
  - Thanks to the Jaffrey lab, now we can with DFHBI.
  - The spinach aptamer binds to DFHBI and becomes fluorescent in the presence of RNA. DFHBI's π structure too bendy to flouresce on its own, but when it is stabilized by insertion into RNA, it can fluoresce.
  - We still need a lot of work before this is as good as GFP.

- The remaining slides will be covered next lecture.

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.

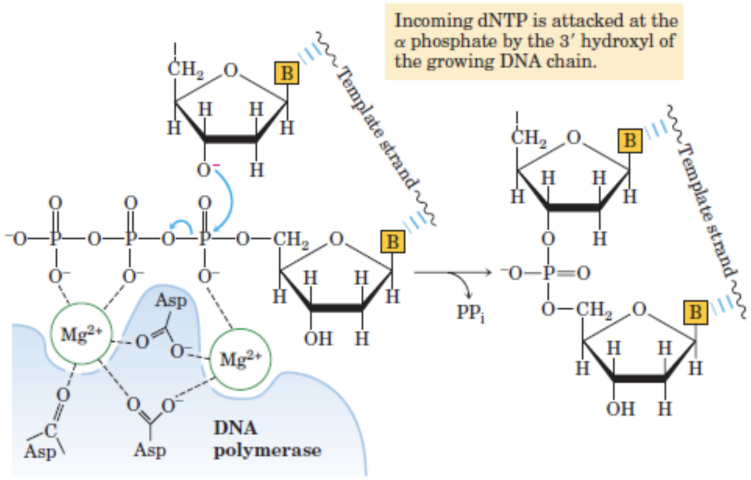


Figure 2.1: Mechanism of DNA synthesis.

  - 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.

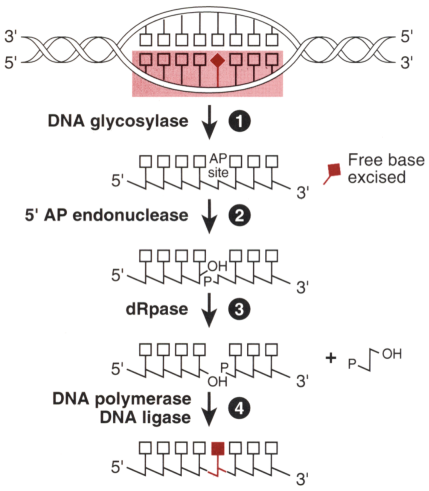
- **Conservative** (amino acid): An amino acid in a protein, the identity of which is critical to the form and/or function of the full protein.

- The causes of DNA mutations in cells.

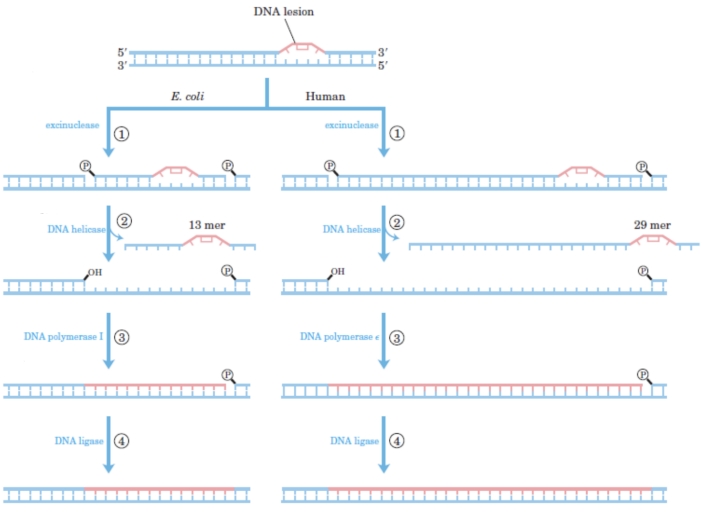
[2] Recall the discussion of this in *The Knot Book*!

  - 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.

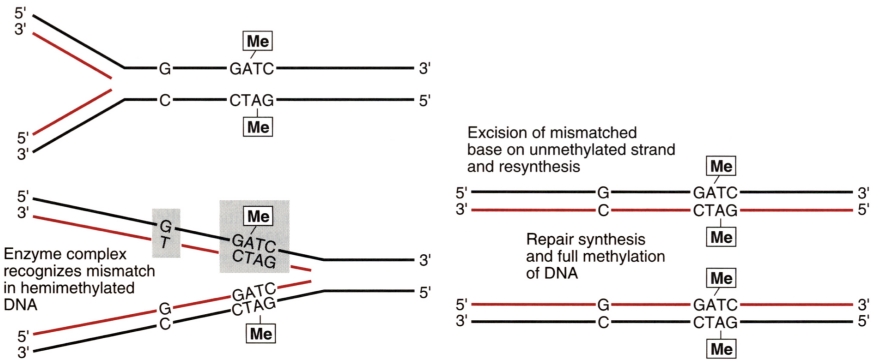


(a) Base excision repair.



(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.

References

Nelson, D. L., & Cox, M. (2021). *Lehninger principles of biochemistry* (eighth). W.H. Freeman.