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.

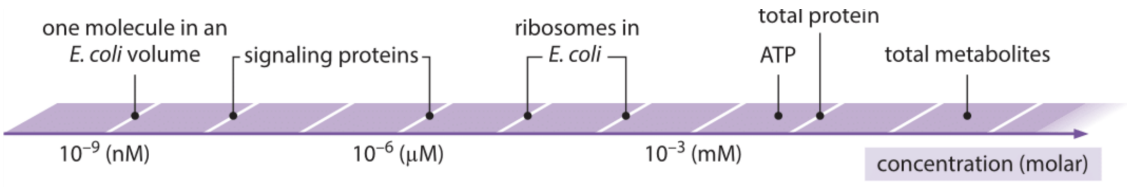


Figure 1.1: Concentrations in biology.

  - 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:

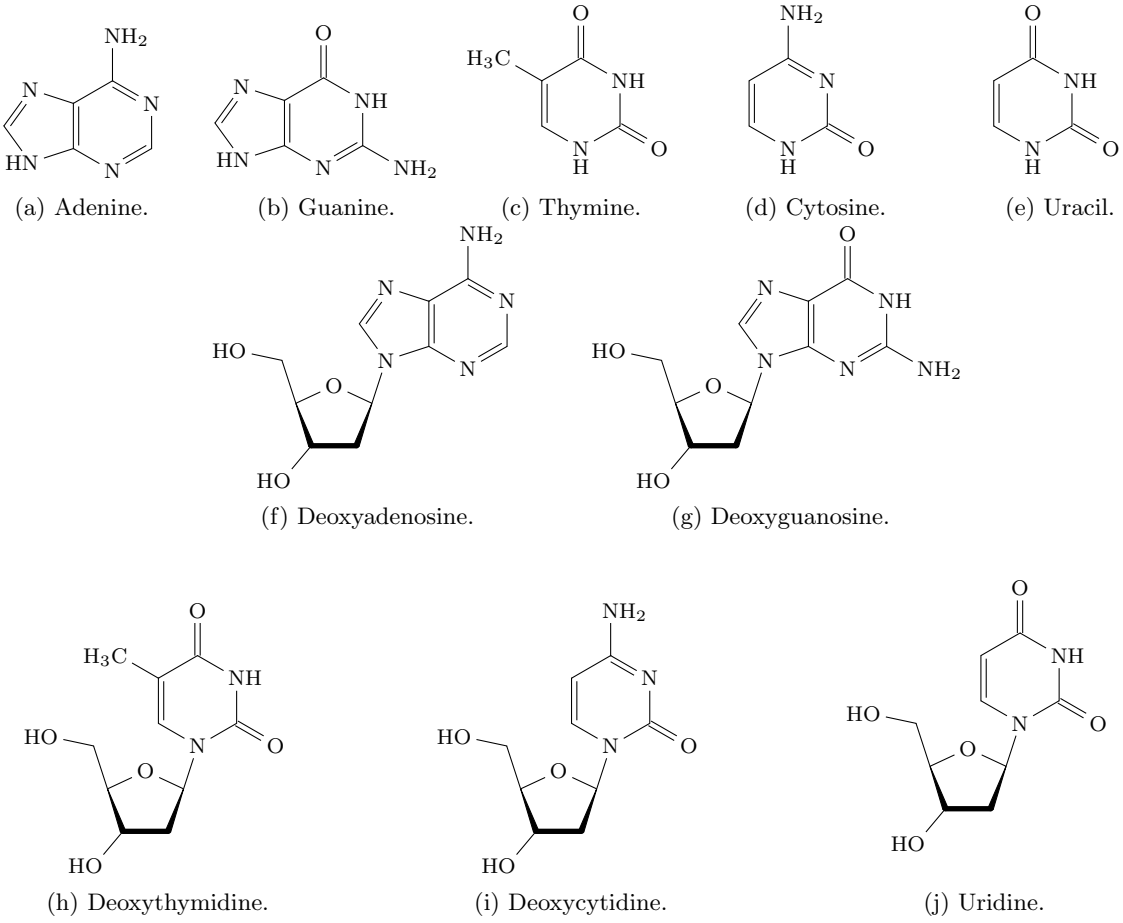
$$\tau = \frac{R^2}{6D}$$

  - 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.

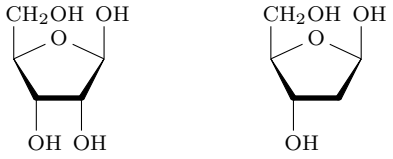
(f) Deoxyadenosine. (g) Deoxyguanosine.

(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**.

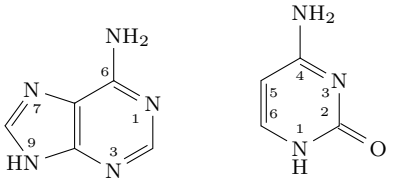
- **Ribose**: A 5-carbon monosaccharide, a derivative of which is a component of DNA.
- **Deoxyribose**: A molecule identical to ribose but without the 2′ hydroxyl group.



(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*

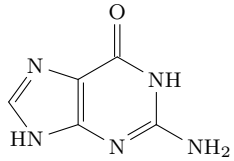
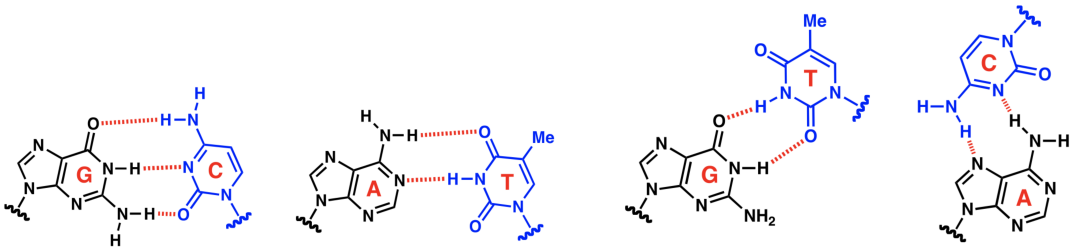


Figure 1.5: Inosine.

- 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.

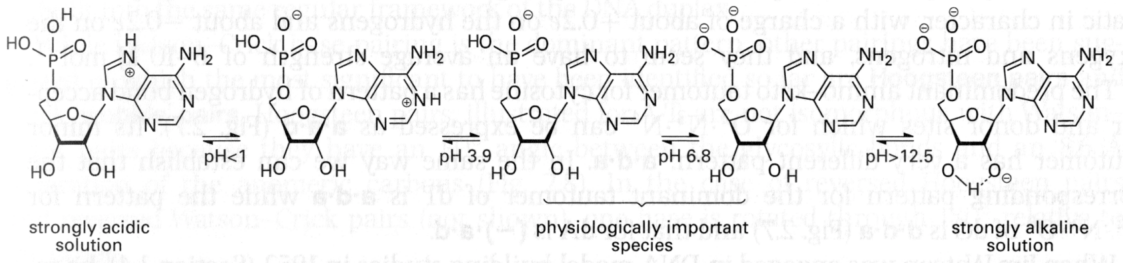


Figure 1.7: DNA ionization states.

  - 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.

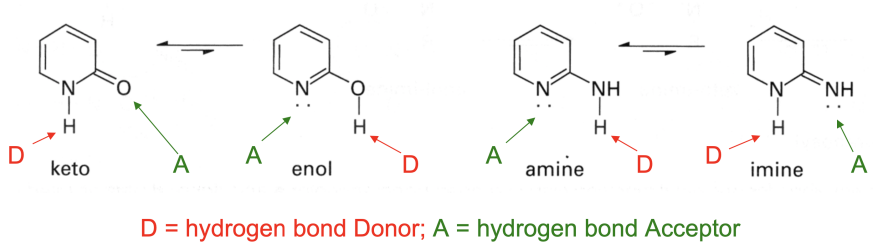


Figure 1.8: Base tautomerization.

  - 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.

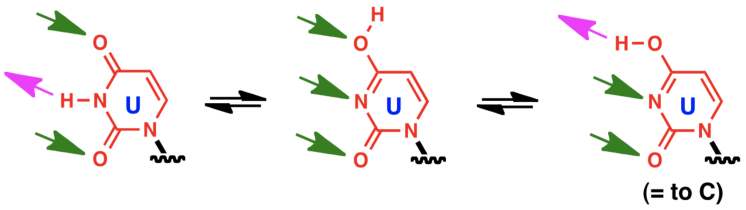
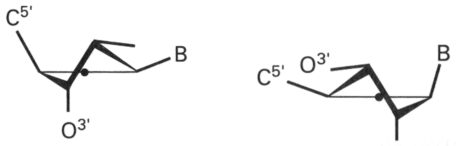


Figure 1.9: Uracil tautomerization.

  - 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.

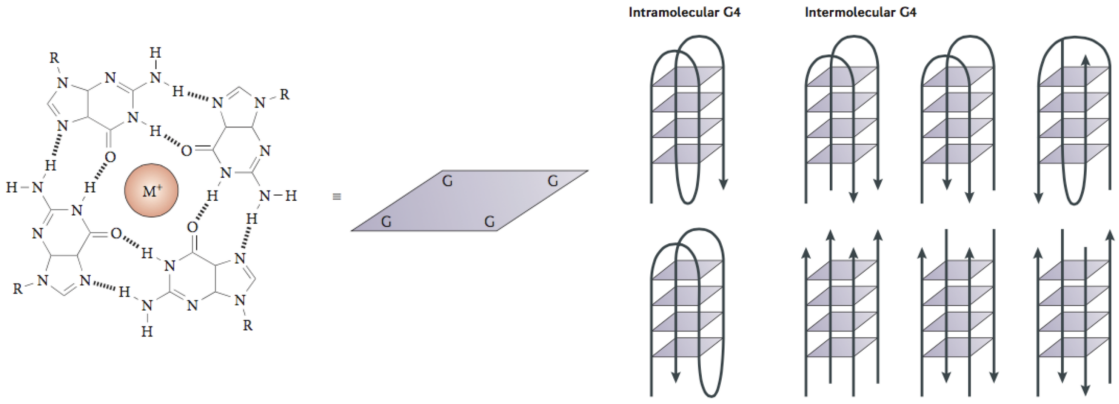


Figure 1.11: G-quadruplex structure.

  - 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.