7.012: Introductory Biology (Draft)[1]

Robert Koirala

Massachusetts Institute of Technology

Last Updated: December 18, 2020

[1]This document is evolving like all of us are.

Preface

These notes grew out of an introductory course that I took at MIT in Fall 2020. Being a math major, I can't imagine writing biology if it was not for the amazing lectures given by the course instructors: Cathy Drennan, Eric Lander, Summer Morril, and guest lecturer Ayce Yesilaltay.

I wrote these notes for two reasons. Firstly, I wanted to learn the material by writing instead of memorizing. I could only remember bigger pictures of gazillion terms introduced in the class. Writing allowed me to go back and find small pictures whenever I had to. Secondly, I wanted to share the appreciation for biology. I used to love biology until my eleventh grade and then memorization happened. Memorization was part of the reason I avoided biology in twelfth grade. In fact, I waited until my junior year to take an introductory class in biology, but my plan was to take it senior year. Given a weird online semester, I decided to get done with my science GIRs (General Institute Requirements). It turned out that the class was more than just memorization (at least for the exams). It was about us. The instructors allowed me to go through the mind of Mendel. And most importantly, it was about the ongoing pandemic. I am thankful to the instructors for restoring my love for biology.

I have to admit that I have failed to keep my spirit in typing notes for some lectures. Partly because some lectures did not have moments of discovery. It got back to memorization of terms. On the other hand, there were times when I got tired of ZOOM. Finally, I was not motivated to learn some topics. Not everything appealed to me.

The organization of material in these notes is based on the lectures. In fact, what I have written is more or less what the instructors said in the class. Also, most of pictures that are not cited are derived from the handouts (and where applicable I have tried to track the source of images). There is no formal bibliography section, but I have embedded links to the sources within the text. I have to confess that these notes are not free of mistakes both technical and mechanical. In some places, I have failed to fill logical holes. But, like every math book says, I expect you to fill those holes yourselves. And if you have any comments or mistakes to point out, you can email me at r[lastname][at]mit[dot]edu. Regardless of the mistakes, I have managed to give a flavor of math-writing. In particular, I have used definition, lemma and proposition where applicable. However, I don't suppose it to be a math book where one sets up definitions and axioms and infers theorems logically.

Like the course, these notes are divided into eight modules focused on areas of current

research in molecular and cell biology, immunology, neurobiology, human genetics, biochemistry, and evolution.

Most of the times, studying biological phenomena boils down to figuring out the chemistry behind it. Module 1 introduces the basic chemistry that is used in this class. Module 2 develops genetics from a historical point of view. This was where I went through a process of discovery done by Mendel, Morgan and others. Module 3 is about DNA, its structure, and its production. The essence of inheritance of genetic information lies in DNA. Furthermore, the module talks about mutations which are the basis of evolution of species. Module 4 brings biochemistry, genetics and molecular biology together. Module 5 gives an overview of genetics from a wide angle. Rather than focusing on one gene, it zeros in on genomes which consist of thousands of gene. Module 6 is relevant to the ongoing pandemic. It is based on microbes (viruses and bacteria) and how our immune system battles against them. There is a lecture specific to Covid-19. Module 7 discusses how the structure introduced in Module 1 is determined in practice. Studying the protein structure will shine a light into the mechanisms involved in different diseases (cancer and heart disease), so that we can come up with cures. Lastly, Module 8 puts biology as a field in a social context. It brings up issues like climate change and ethics behind gene editing.

Acknowledgements: I am grateful to my grandma, mother, father, and sister for who I am today.

The other Cambridge
Massachusetts

Robert Koirala
December, 2020

Contents

Module 1

Biochemistry

1.1 September 2

Welcome to 7.012!

Instructors:

Eric Lander: He hated biology in high school because he thought biology consisted of a lot of memorization. In fact, he has a PhD in pure mathematics (Topics in algebraic coding theory). "Sorry, I am not qualified in paper to be a biologist." However, in mid 90s a human genome project caught his attention. The project consisted of reading out every species on the planet. He also realized that biology was relevant to so much that was going on in the world. It did not take that long for him to fall in love with genetics. Soon he with his other colleagues launched the first genome center, Broad Institute, at MIT.

Now, he studies diseases at Broad institute. In spare time, Prof. Lander deals with public policies. In particular, he works with the White House.

Fun Fact 1.1.1. Covid-19 test is done at Broad Institute.

Cathy Drennan: For 14 years, she taught 5.111 (an introductory chemistry class).

"How did you end up in biology?"

"This is MIT, where we do interdisciplinary research. It does not matter if we have degrees from other departments."

She took an indirect path to become a biology professor. In college, she was forced to take chemistry but fell in love with it. Chemistry is about reactions of chemicals. And a lot of interesting reactions occur in living cells. Therefore, her love for reactions brought to biology.

"It is an important time to study biology. We are in the middle of a pandemic," says

Prof. Drennan. She hopes that this class will motivate us to pursue research in biology. There are a lot of technological developments that are changing biology: CRISPR and electron microscope are some of them.

Prof. Drennan loves to wear geeky T-shirts. Today, her T-shirt says, "Peace, love & vaccine. Life is good."

Summer Morril: Unlike other professors, she has studied biology most of her life. In fact, she holds a PhD (MIT) in genetics. She also loves mathematics and probability in a biological context. Dr. Morril thinks about the best ways to communicate biology. In addition, she supervises 20+ lovely TAs.

Resources:

Look at Canvas for a detailed information about psets, exams, grades and recitation. We will use Piazza to answer questions. Recorded lectures will be posted online in Canvas. Enrolled students have access to MITx: 7.00X. It is similar to lectures that Prof. Lander will give but is less identical to the lectures given by Prof. Drennan. Her lecture will focus on biochemistry. There is a study guide called *The Secret of Life* which might turn into a textbook. Check the syllabus on how to access the guide. There is an optional textbook, *Life: The Science of Biology* by Sadava. But it does not cover everything that is taught in this class.

Now that we are done with logistics, let's talk about what's coming next in this class and what's not.

1.1.1 Levels of Organization of Life on the Earth

Biology is a hot pot of a lot of topics and one of them is levels of organization of life on the Earth. We can study life in the following hierarchy:

- **Biosphere:** It consists of an entire world, organisms, and environment. An example is the Earth. As of now, we don't know about other biospheres.
- **Ecosystem:** Prof. Lander loves to go to New Hampshire. There are forests with trees, musk deer, bacteria, and people, all of them interacting together to create an ecosystem.
- **Organism:** There are more than one million species of organisms in the Earth including humans. In fact, MIT undergraduates consist of a species.
- **Organs:** We can focus down to organs of an organism. Let's try eyes. Walking through a forest in New Hampshire, we can look at forest through our eyes.
- **Tissues:** Going further down the hierarchy, we can study tissues of an organ say retina of our eyes. Retina helps in forming images in our eyes.
- **Cells:** We can study individual cells. There are tiny cells, large cells, funny shaped cells, rod cells that are found in retina.

- **Organelles:** We can push our boundary to organelles. Within cells, there are organelles like nucleus and mitochondria (energy powerhouse). We can talk about what is common in different kinds of cells.
- **Molecules:** And everything boils down to molecules. ATP (adenosine triphosphate) is a molecule found in mitochondria.
- **Physics:** If we go further, we will encounter physics. Because this is a biology class, we won't be talking about physics. That way, we don't have to talk about mathematics. Neither about philosophy. Nor about linguistics.

Because cells, organelles, and molecules are common building blocks of all living species, we will focus just on them. It does not mean other topics are boring.

1.1.2 Evolution

Definition 1.1.2. *Evolution* is an iterative process in which a monkey turns into a person after a reasonable amount of time.

We don't have time to study every single topic (including evolution) in a semester. However, we acknowledge the studies of evolution and may make brief references. Here is a brief outline of evolution of life on the Earth:

- **4.5 Billion years ago (BYA):** The Earth forms. It is hot without oxygen but toxic gases.
- **4 BYA:** It cools down.
- **3.7 BYA:** The first life in a form of bacteria (prokaryotes) evolves.
- **1.5 BYA:** Around this time, the first nucleated cell (eukaryotic cells eg: fungi) evolves.
- **0.5 BYA:** This is when true multicellular organisms evolve. Nature starts to experiment with different body plans. Bodies with scales, tails, and wings. Some have limbs. Every body plan that exists today comes from that period.
- **0.005 BYA:** Humans evolve.
- **0.0000002 BYA:** MIT is founded. And we start to evolve into a species.

Let's now overview what we will discuss in this class.

1.1.3 Types of Cells

There are two fundamental types of cells:

- **Prokaryotes** lack true nucleus and organelles. They are about 1-2 microns large. eg. bacteria.

- **Eukaryotes** have a nucleus and organelles like mitochondria. They are about 10-40 microns. eg. cells in our body.

1.1.4 Background on Biology

Biology is a young subject. However, it was a topic of interest for a long period. A general outline of what happened is given below:

- Dissection revealed anatomy. Greeks regarded brains as radiator.
- Microscopy revealed that living matter was made of cells. These were the finest structures in living bodies that we could see.
- We learned a lot just by "looking," but without experimental methods we could not prove much. For instance, Aristotle believed that sex distinction depended on whether the conception occurred in north wind or the south wind. It was based on faith not on experiments and did not help to predict sex distinction.

1.1.5 Analytical Biology

It is the experimental methods that drive this course. Biologists want to understand the functions of organisms. Why does a butterfly flap its wings in a forest in New Hampshire? Some people found that functionality is related to genes while other studied functionality by studying proteins.

- **Genetics:** It is about genes. Geneticists study an organism minus an individual component. They choose naturally occurring mutants for that. Suppose that a butterfly can't fly. Flightlessness of the butterfly helps in figuring out the components that are important for the flight.
- **Biochemistry:** It is about proteins. Biochemists are interested in what stuffs are made up of. First they isolate one component from the whole organism. Let's go back to butterfly. Biochemists are happy if they purify two substances that slide on each other which will answer the muscle movement during a flight. They explain the flight in terms of a state of some molecules, sugars, proteins (polymers of amino acids), nucleic acids (DNA and RNA), lipids (they are hydrophobic: they don't like water), and phospholipids (they have hydrophobic end that hides and hydrophilic ends that stick out). To explain the state of these things, they use chemistry and learn about biochemical reactions. Both geneticists and biochemists use different approach to study functions of organisms.

These fields did not have anything in common until the rise of molecular biology in 1960s. A gene reduces to DNA and RNA which are interconnected with proteins. When people discovered genes they thought that they knew everything that was to functionality. But a new generation thought about reading out individual genes and express them to understand everything about functionality. They went from theory to practice revolutionizing

the field. When they thought more they figured that reading out one gene at a time would take centuries.

Another generation (of Prof. Lander) considered to look at entire genome (thousands of genes) at a time that led to an inception of the genome project at MIT. The project brought centuries to hours. Throughout the period of this class, six entire genome at MIT are being read out. These days people at Broad study cellular reprogramming, CRISPR, genome editing, and self-cell biology. We will talk about these topics in the course.

This overview can be summarized in Figure 1.1. Don't worry if it was to fast. Prof. Drennan will be back Friday with the first lecture on biochemistry.

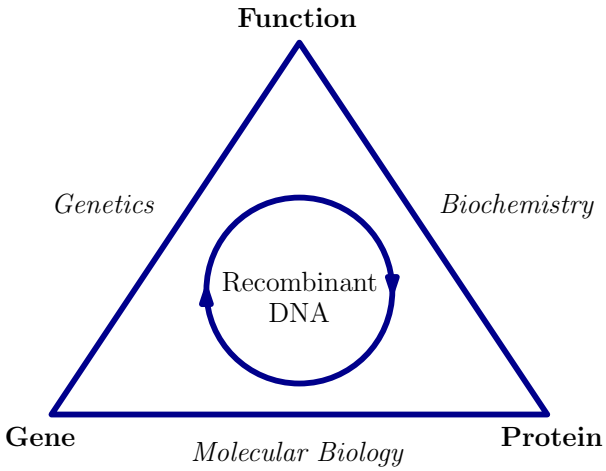


Figure 1.1: Triangle

1.2 September 4

"Noble Gases," says Prof. Cathy's shirt.

In this half of the class, we attempt to unravel the secrets of life. We will start simple with (bio)chemistry and build up on that.

Fun Fact 1.2.1. At MIT, if we take an intro bio class, we first learn chemistry. But if we take an intro chemistry class, we are supposed to learn physics. Guess what they teach in an intro physics class. Math. But if we take math, we just do math.

Moral: You should be a math major like me.

1.2.1 Elements of Life

If we look at a "fundamental" level of our body, we find a lot of chemistry going on. This section is a glimpse of what carries on those chemistry.

- **Elements:** Some common elements in our body are C, H, I, P, S, and O. In addition, d-block elements (transition elements) also play a vital role in our body. They constitute metallic proteins. For instance, iron is found in haemoglobin.

  Poll 1.2.2. What does not naturally occur in our body?
  - Fe
  - Zn
  - Au
  - Mn

  Answer: All but gold (Au).

  Other important d-block elements are Mo and Co. D-block elements are typically metals. Metals are avengers: Iron Man and Thor. They are powerful but dangerous.

- **Ions:** Elements we listed above are found in uncharged state. In contrast, ions are charged elements. Positively charged ions are called *cations.* Some common cations in our body are Na^+, K^+, Mg^{2+}, and Ca^{2+}. On the other hand, negatively charged ions are called *anions.* Cl^- is the most common anion found in our body.

1.2.2 Functional Groups of the Molecules of Life

These are small bonded units of elements and ions that play common functional roles.

- Hydroxyl: $-O-H$
- Carbonyl: $-\overset{\overset{O}{\|}}{C}-$
- Carboxyl: $-\overset{\overset{O}{\|}}{C}-OH$
- Amino group: $-NH_2$
- Phosphate: $-PO_4^{3-}$
- Sulfhydryl: $-SH$

These groups can adapt different charged state depending on their pKa and pH scale of the ambient environment.

Definition 1.2.3. *pH* is a measure of acidity and basicity of a system (solutions). It ranges from 0 to 14. If the pH value of a solution is below 7 it is acidic whereas it is neutral and basic at higher values.

Recall that an acid donates protons (hydrogen) and base accepts protons (hydrogen).

Definition 1.2.4. *pKa* is the pH value at which a molecule/functional group is half protonated and half deprotonated.

Question 1.2.5. If the pKa of a carboxyl group is 3.7, will you expect the group to be negatively charged or neutral at pH 7.4?

1.2.3 Chemical bonds

Now that we have overviewed elements and groups that are found in our body, let's discuss what keeps them together. The elements bond because they want to have a noble gas configuration. Noble gases are so happy with eight electrons in their outermost orbit (except Helium which has two). They are dressed. They have swords. To achieve this configuration, elements tend to donate or accept electrons.

Definition 1.2.6. The tendency to attract electrons is called *electronegativity.*

Depending on the difference in electronegativity, there are two types of bond.

- **Ionic bond:** Some atoms have negligible electronegativity, so they donate electrons while others with high electronegativity accept electrons together forming ionic bonds through the transfer of electrons.

  Example 1.2.7. Na has to get seven electrons to be a noble gas. It has low electronegativity, 0.93. However, getting rid of one electron is easier than accepting seven. Donating an electron it becomes Na^+. In contrast, Cl has seven electrons and has high electronegativity, 3.16. It gets an electron from Na and becomes Cl^-. With eight electrons in the outermost orbit it lives with Na^+ happily ever after.

- **Covalent bond:** Most of the bonds in biology are covalent bonds. When the difference in electronegativity is not that big, elements achieve a noble gas configuration by sharing electrons forming covalent bonds. Depending on the difference in electronegativity there are two types of covalent bond:
  - **Nonpolar:** There is an equal sharing of electrons. eg. CH_4
  - **Polar:** There is an unequal sharing of electrons.

    Rule to keep in mind: If the electronegativity difference between bonded atoms is greater than 0.4 *then the bond is considered polar.*

    Example 1.2.8. Methane (CH_4): The electronegativity of H is 2.20 and that of C is 2.55. Therefore, the bond is non-polar. C and H live happily ever after sharing the electrons equally.

    Example 1.2.9. Water: The electronegativity of oxygen is 3.44 and the difference is 1.4 which is greater than 0.4. Therefore, the bond is polar.

    Fun Fact 1.2.10. In class, two dogs taught us about ionic bond, covalent bond, and valence electrons to understand the biology.

Polar versus Nonpolar Molecules

In the previous section, we talked about the polarity of bonds. But molecules overall can also have polarity:

- **Polar molecules** have polar bonds due to unequal electronic distribution.
- **Nonpolar molecules** have nonpolar bonds and a uniform electronic distribution

Remark 1.2.11. A molecule with polar bonds can be nonpolar if the bonds are symmetric so that the dipoles (polarity of bonds) cancel out. For eg. carbon dioxide is a linear molecule with a structure O=C=O. Even though each bond is polar they are aligned in opposite direction and cancel out each other making the molecule nonpolar.

Water makes a huge portion of our body. Therefore, it is important to get a sense of how molecules interact with water. The concept of polarity gives rise to the following notions:

- **Hydrophilic** molecules are water loving. They are polar molecules and dissolve in water.
- **Hydrophobic** molecules, on the other hand, hate water. They are non polar, so they don't dissolve in water.

1.2.4 Non-covalent Interaction

In the last section, we studied some bonds that hold elements together. However, in our body, there are weaker forces that help in forming bonds. Recall that the strength is measured in terms of energy required to break these bonds. We list these weaker forces that hold molecules together in the order of strength of interaction: from strongest (top) to weakest (bottom).

- **Ionic bond:** The strength of ionic bonds are typically greater than 100kcal/mol. For eg. NaCl.
- **Covalent bonds:** Strenghth: 110 kcal/mol. Some common covalent bonds are C–C, C–N, C–O, C–H, N–H, and O–H.
- **Electrostatic interactions (Ionic interaction):** Strength: 3-8 kcal/mol. We should NOT get confused it with ionic bond. It is just an interaction between positively and negatively charged groups. For instance, the interaction between amino group, $-\overset{\overset{H}{|}}{\underset{\underset{H}{|}}{N^{+}}}\cdot H$, and carboxyl group $-\overset{\overset{O}{\|}}{C}\cdot O^{-}$ is electrostatic.
- **Hydrogen bond:** Strength: 3-7 kcal/mol. This is an interaction between electronegative atom and hydrogen in a polar bond. For instance, there is a hydrogen bond between carbonyl group $-\overset{\overset{O}{\|}}{C}-$ that acts as acceptor and hydroxyl $-O-H$ that acts as donor. Although the strength of hydrogen bonds is low they are powerful because there are a lot of them.
- **van der Waal's interaction:** Strength: 0.1-1 kcal/mol. It is an interaction due to transient polarization of molecules as a result of non uniform distribution of electron cloud. In non polar compounds, partial charge might form. These polarized molecules when come close to each other can have interaction and. stay together. This force keeps noble gases bonded together.

1.2.5 Unfavorable Interaction

So far, we have only talked about interactions/bond that are favorable in the nature. However, there are unfavorable interactions that are important in biochemistry.

- There is electrostatic repulsion because like charges like to repel.
- Similarly, there is repulsion between hydrophobic molecule and water. It helps in structure formation for protein and lipids.

If every reaction was favourable there would just be a chunk of particles in the world instead of individual beings.

1.3 September 9

Prof. Drennan's shirt has prints of amino acids which is what we are touching upon today.

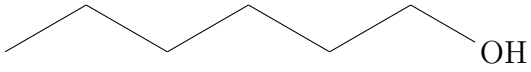
Previously, we studied elements, ions, and bonds that hold them together. Today, we will talk about the molecules of life.

1.3.1 Carbohydrates/Sugars/Saccharides

When we think of carbohydrates we think about food. But carbohydrates are also part of structures like DNA, RNA, cellulose, and coating of sugar.

Research filling: Laura Kiessling and Barbara Imperiali are two world leaders of sugar. They work on sweeter sides of sciences. We should reach out to them if we want to work on a research project.

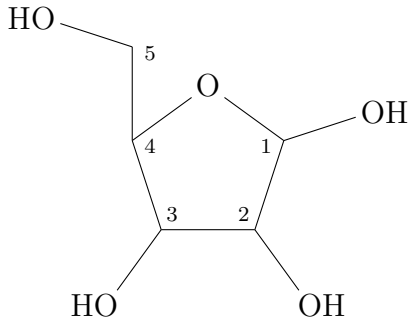
Digression on line drawings: We don't draw carbon and hydrogen associated to a molecule. It is mean to say that carbon is boring but we could say that carbon is predictable. It forms four bonds. For instance, the molecule below



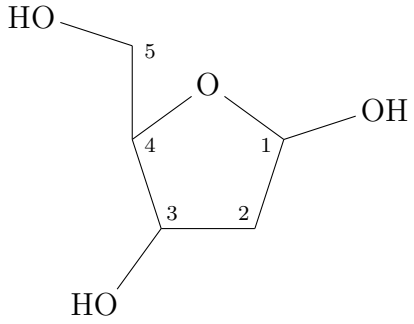
is $C_6H_{13}OH$.

Based on the number of fundamental blocks (sugars), carbohydrates are categorized as:

- **Monosaccharides** are building blocks of bipolymers. The two types of sugar that are important in our discussion are ribose and deoxyribose. Both of them have five carbons and hydroxyl groups. The structure for ribose is



. In deoxyribose, the carbon at 2nd position has no oxygen justifying the term deoxy. Its structure is



- **Oligosaccharides** have only two or three sugars. For eg. sucrose and lactose.
- **Polysachharides** have more sugars. For eg. starch and cellulose.

There are ways to join and break saccharides.

Definition 1.3.1. *Condensation* is the removal of two hydrogen and one oxygen to make water to form a glycosidic linkage between saccharides. This linkage will help in forming polysaccharides from monosachharides.

Definition 1.3.2. In contrast, *hydrolysis* is an attack of water. It is the reverse process of condensation and helps in breaking bonds in polysaccharides.

1.3.2 Nucleotides/Nucleic Acids

Genetics discussed in second half of the course boils down to DNA and RNA. We need to discuss about nucleic acids to make sense of DNA and RNA. They are biopolymers made up of nucleotides.

Definition 1.3.3. A *nucelotide* in RNA has three components in it: base, sugar (ribose or deoxyribose), and one to three phosphates groups.

Definition 1.3.4. A *nucleoside*, in contrast, has a base and a sugar but no phosphate group.

There are five types of bases:

- Adenine (A): Both in RNA and DNA

- Guanine (G): Both in RNA and DNA
- Cytosine (C): Both in RNA and DNA
- Thyamine (T): Only in DNA
- Uracil (U): Only in RNA

We do not need to memorize structures of bases and nucleotides. But we need to be able to distinguish things with the help of information provided: charts of bases, sugars, and functional groups.

In class, we did the nomenclature of ribonucleotide (ATP), see Figure 1.2 and of deoxyribonucleotide (dATP), see Figure 1.3.

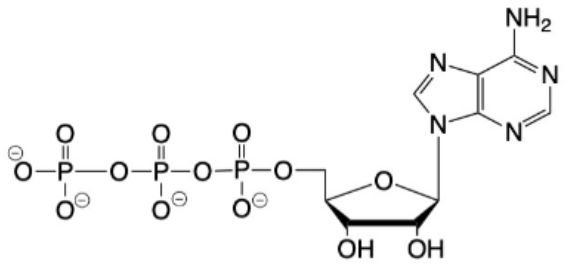


Figure 1.2: Ribonucleotide Adenosine Triphosphate (ATP)

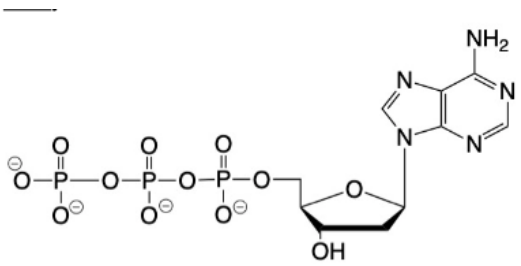


Figure 1.3: Deoxyribonuceotide Adenosine Triphosphate (ATP)

Here are some important abbreviations that will come up a lot in biochemistry.

- NTPs make up RNA. N stands for base (AGCU) and TP for triphosphate.
- dNTPs make DNA. N stands for base (AGCT) and TP for triphosphate
- NMP stands for nucleoside monophosphate

Definition 1.3.5. *Oligonuleotides* are long chains of DNA and RNA.

Definition 1.3.6. *Phosphodiester linkages* connect nucleotides between the 3 carbon of one sugar molecule and 5 carbon of another sugar molecule.

1.3.3 Lipids/Fats

Let's talk about fats and lipids. They are greasy and insoluble in water.

Law 1.3.7. *Like dissolve like.*

Therefore, lipids are "fat soluble" only. Because of their greasy nature, they play a great role in forming cell membranes. In addition, they also store energy. Some of the important lipids are: phospholipids (phosphatidylcholine), steroids (cholesterol) and vitamins (D, E, K, and A). They have non-polar bonds. Source of Vitamn A: spinach, leafy greens (folic acids), and carrots.

Fun Fact 1.3.8. The term Folic has connection to foliage. In fall foliage, go to Vermont. Prof. Drennan likes to go there.

Let's focus on phospholipid in particular.

- They are components of cell membranes
- They contain fatty acids, which are carboxylic acid (polar and hydroplilic) attached to a long hydrocarbon chain (non-polar and hydrophobic). Hydrocarbon chain can be saturated (all carbon are singly bonded with as many hydrogen as they can hold) or unsaturated (with one or more double/triple bond).
- Phospholipids also contain glycerol, phosphate, and small molecules like choline, serine or ethanolamine.
- Phospholipids are *amphipatic:* Red head (polar and hydrophilic) and yellow tails (non-polar and hydrophobic). The polarity helps in formation of lipid bilayers.
  - Extracellular: Heads face towards solvent. Tails face inside and are buried away from water.
  - Intracellular: Head groups are facing towards the solvent (water).

.

1.3.4 Amino Acids/Peptides/Proteins

Amino aids are building blocks of peptides (short chain) and proteins (longer chain). The structure of an amino acid is divided into α- carbon that is attached to carboxyl group, amine group, and a chain of hydrocarbons (shorthand notatiion R). There are two types of isomers L and D. Only L-isomers are found in proteins. L-isomers are non-superimposable to their mirror counterpart.

Proteins are the workhorses of our cells. In next class, we will talk more about proteins.

1.4 September 11

"Keep calm and talk about Thermodynamics," says Prof. Cathy's T-shirt. That's what we will be doing today. First, we will complete our discussion on amino acids from last time and talk about proteins.

Prof. Drennan's cousin and his wife had a nine pound baby boy in San Francisco. Baby boy does not have a name (could name him Thermodynamic). His brother's name is Julian and is 3 years old.

1.4.1 Peptide Bonds

Amino acids are the fundamental part of proteins. When they undergo condensation (lose water molecule) a peptide bond is formed. Different amino acids might have different R side chain (hydrocarbon chain). There are 20 types of amino acids. Clycine is the smallest one. Depending upon the polarity of molecules, we can categorize them into:

- **Nonpolar:** Alanine, valine, cysteine (sometime it is polar and sometime it is nonpolar), and proline (unusual one: the R group comes around and joins back to the amine group).
- **Polar:** Hydroxyl or carboxyl group make them polar. Serine and tyrosine.

1.4.2 Introduction to Protein Structure

Proteins are chains of amino acids. The chain can have different structures.

- **Primary structures** have a sequence of amino acids from N terminus to C terminus. There is a beginning (amino acid) and an end (last amino acids.) See Figure 1.4.

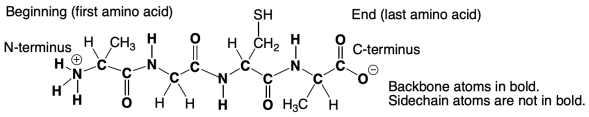


Figure 1.4: Primary Strucutre

- **Secondary sequences:** They appear as α helix, β strand, β sheet and turns and loops. The structures are stabilized by hydrogen bonds between backbone atoms.
  - α **helix** is a right handed helix.
  - β **strand** is a short extended stretch of (typically 3-10) residues (amino acids).
  - β **sheet** has two or more β strands running next to each other with H-bond between carbonyl and amino groups. The sheets can have parallel or antiparallel β strands. The numbering of atoms starts from nitrogen to α carbon

and carbonyl carbon. If the numbering is parallel in a H bond, the strand is parallel otherwise it is anti-parallel.

- **Turns and Loops** connect units of secondary structures.

- **Tertiary:** It is a three dimensional arrangement of primary and secondary structures. It can have one or more *domain* where a domain is defined as a self folding unit of structure.

- **Quaternary structures:** These structures describe the number, type and arrangement of polypetide chains.
  - **Monomer:** 1 polypeptide chain
  - **Dimer:** 2 polypeptide chain
  - **Trimer:** 3 polypeptide chain
  - **Tetramer:** 4 polypeptide chain

  Remark 1.4.1. Chains can be identical (homo) or different (hetero) from each other.

Question 1.4.2. How are these structures stabilized?

We can answer it by studying the types of interaction that keep the structures intact.

- **Primary:** They have only peptide bonds which are covalent.

- **Secondary:** They have H-bond between back bone atoms.

- **Tertiary:** They have disulfide bond that are covalent bond. Eg. Cystine (sulfur between two cystine can form covalent bond). In addition, they also have electrostatic interaction. For instance, Glu (Glutamine) (-) and Arg (Argenine)(+) can interact with each other. Tertiary structures also have hydrogen bond (between backbone atoms) and van der Waals interactions.

- **Quaternary:** They have disulfide bonds between chains, electrostatic interactions, H-bond as well as van der Waals Interaction.

Question 1.4.3. But why do protein fold at all?

It is because of hydrophobic effect. Hydrophobic residues cluster together, away from water, forming the protein interior. Water molecules form a cage around unfolded protein chain. But they are free to move around after the protein adopts a folded structure. The release of water molecules from their cage is entropically favorable because this state has more disorder.

Fun Fact 1.4.4. Dog likes to teach chemistry. Picture of a dog with and without hair cut explains the concept of entropy. See Figure 1.5

Figure 1.5: Dog likes to teach chemistry

1.4.3 Introduction to Biological Catalyst

In this section, we will talk about the role of protein. Energy is all around us but we need to be able to use it at a rate that is compatible with maintenance of life. Indeed, we have glucose ($C_6H_{12}O_6$) in our body but we can't use it if there were no proteins.

Question 1.4.5. What makes for a good energy source?

- **Availability:** Glucose comes from photosynthesis
- **Favorability of reaction:**
  - **Exergonic:** Products are lower in energy than reactants. Therefore, exergonic reactions are favorable.
  - **Endergonic:** Products are higher in energy than reactants making endergonic reactions unfavorable. However, these reactions can be driven forward by coupling them with favorable reactions.

  When we talk about favorability we will associate to a reaction its ΔG°_{rxn} (Gibbs free energy). If the free energy is less than zero, the reaction is exergonic and spontaneous whereas the reaction is endergonic and nonspontaneous if $\Delta G^\circ_{rxn} > 0$.

Here is a reaction that occurs in our body for a release of energy:

$$\overbrace{C_6H_{12}O_2}^{eat} + \overbrace{6O_2}^{breathe} \rightarrow \overbrace{6CO_2}^{exhale} + \overbrace{6H_2O}^{pee}$$

For this reaction, $\Delta G^\circ_{rxn} = -673\text{kcal/mol}$, so the reaction is spontaneous. But this is a slow process. Spontaneity and slowness bring us to concepts:

- **Thermodynamics** addresses the issue of stability and energy of reactants and products.
- **Kinetics** studies the rate of formation of products.

  Remark 1.4.6. The conversion of a diamond to graphite is thermodynamically favorable. "A diamord is forever" is therefore a kinetic statement.

Catalysts are superheroes of kinetics. Most biological catalysts are proteins called enzymes, but sometimes RNA can act as a catalyst.

Definition 1.4.7. All reaction have a barrier called *activation energy barrier.*

The rate of all reactions can be increased by a catalyst as they lower the activation energy. It can be represented in a reaction activation diagram via energy barrier curve. There is a transition state that will be formed. Catalysts lower the energy barrier curve. See Figure 1.6.

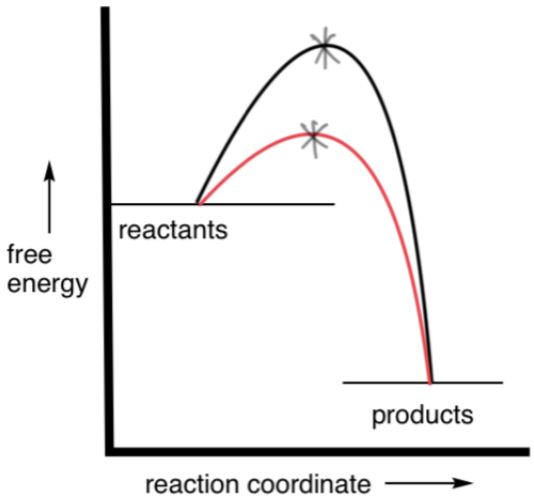


Figure 1.6: Stared regions correspond to transition states. Black curve is the energy diagram of reaction without catalyst and red curve is that with catalyst.

Question 1.4.8. Why do we care about catalysts?

Because they

- lower energy of transition states.
- lower E_a barrier of forward reaction.
- increase the rate of forward reaction
- lower E_a barrier for reverse reaction
- increase the rate of reverse reaction
- but they do not change ΔG of a reaction

1.5 September 14

Fire is still going on in California and Oregon. Prof. Cathy misses the baby and her husband's family in Oregon. Anyway, we should join Biology Undergraduate Student Association. They have cool zoom events. We don't have to be a bio major.

Today, Prof. Cathy is wearing a T-shirt that she will explain later.

1.5.1 Enzyme Catalysis and Inhibition

Last time, we touched upon the fact that enzymes (cataylst) lower and increase the rate of reactions. Today, we will see how they actually do it. Briefly, enzymes bind to reactants (substrates in its active site) that get converted into products with the help of residues (amino acids) and go into another round of enzyme catalysis also known as *turnover.*

Fun Fact 1.5.1. Prof. Eric Lander likes to put cells in grinders and purify enzyme molecules.

Poll 1.5.2. In a graph of substrates vs. rate, the rate increases with substrate but levels offs. Why?

Answer: It is because active sites become full. Adding enzyme will increase the number of active sites.

Definition 1.5.3. *Enzyme inhibitors* bind to enzymes and slow or stop reactions. They are used to treat diseases and pain. For instance, chemotherapeutic drugs and aspirin can be enzyme inhibitors.

There are two types of inhibitors:

- Ones that block substrate from binding. They are found at active sites.
- The second type seal active sites but are not present in active sites. They are at *allosteric site* which is a fancy name for a site remote from active site.

  Fun Fact 1.5.4. A lot of pharmaceutical companies in the past thought about the first type of inhibitors. But recently, they have started to look into allosteric site type inhibitors. There are so many startups at MIT which study the second types.

  Fun Fact 1.5.5. Prof. Cathy has a neighbor who has a house in New Hampshire. It also means that it is her house. Her neighbor learned biochemistry from her.

  Moral: Remember to donate money to Prof. Cathy if any of us earn money off of this material. This also means if you pass the exams or ASE based off of this material, please donate me, XD.

1.5.2 Introduction to Enzyme Pathways

Enzymes work together in reaction to produce molecules. They often act in particular order to accomplish their function, which brings us to a topic of enzyme pathways.

Fun Fact 1.5.6. Dorothy Hodgkin won a Nobel prize in '64 for determining the structure of vitamin B-12 and penicillin using X-ray crystallography.

Question 1.5.7. Why do we care about enzyme pathways?

- To make more natural product that have some particular pathways. Microbes might not make a ton of a product because it might not use it. But if we are interested in the product, we need to study pathways.

- To make a new product that nature doesn't make. We can change stability of products. We might want product to last longer or be slippery.
- To make a natural product from different ingredients. A lot of pharmaceuticals are made from different ingredients. On a different note, we can get use something else than glucose to get energy.

Research filling: We should UROP with Krish Prather. She is a chemical engineer interested in making low carbon fuels. She also manipulates microbes to make cancer drugs.

Question 1.5.8. What do we need to know about enzyme pathways if we want to do any sorts of engineering like Prof. Krish? Unfortunate news: Most enzyme reactions are reversible and reach equilibrium at some point, so we need to learn about equilibrium.

Reaction Equilibrium

Let us consider a reaction,

$$A + B \rightleftharpoons C. \tag{1.5.1}$$

At equilibrium, the rate of forward reaction equals to the rate of reverse reaction. We associate an equilibrium constant (K) to a reaction at equilibrium defined as:

$$K = \left.\frac{[\text{Products}]}{[\text{Reactants}]}\right|_{eq}$$

where $[X]$ is the concentration of X. In our case (1.5.1),

$$K = \left.\frac{[C]}{[A][B]}\right|_{eq}.$$

We should take a chemistry class to understand these things in detail. But intuitively, if we want more product we want large K. Under standard condition, the equilibrium constant is related to $\Delta G^{\circ'}$ by the equation

$$\Delta G^{\circ'} = -RT \ln K$$

where R is the universal gas constant and T is temperature and $^{\circ'}$ indicates standard state (room temperature, normal pressure) and pH 7.

Under nonstandard state, $\Delta G = RT \ln(Q/K)$ where Q is the reaction quotient

$$Q = \left.\frac{[\text{Products}]}{[\text{Reactants}]}\right|_{t}$$

where t represents the time of the reaction not necessarily the equilibrium.

With Q and K together, we can get information on the direction of reactions.

- For $Q < K$, ΔG is negative. It means that the reaction is spontaneous in forward direction towards products.
- For $Q > K$, ΔG is positive which implies that the reaction is spontaneous in reverse direction towards reactants.

Fun Fact 1.5.9. Prof. Drennan, finally explained her T-Shirt. It looks like green lantern symbol. But if we look at it carefully we will see Q. The shirt says $Q > K$. She says that our knowledge is our superpower. We are superhero. Like Tony Stark, we can use our knowledge of Q to change the rate of reaction. This brings us to the ways to shift the direction of a reaction.

Theorem 1.5.10 (Le Chatelier)**.** *A system in equilibrium subject to stress tends to minimize the effect of the stress.*

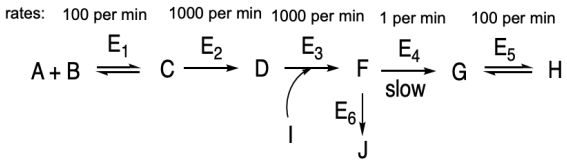
Proof. Remember that this isn't a math class. We should not always expect proofs. □

Question 1.5.11. Based on the above principle, how can we get reaction (1.5.1) to shift to make more C?

Answer: Add more A and B. Could also remove C.

Let's now talk about the main point of this section, enzyme pathways.

- These are the paths of reactions. The paths can have *branch points*. Consider a reaction



  where E_i are enzymes. The formation of J is a branch point.
  - A **disadvantage** of having a branch a branch point is that we won't get as much H or J.
  - In contrast, an **advantage** of having a branch point is that we don't need two separate sets of enzymes to form H and J. Enzymes $1-3$ do our work.
- Enzyme pathways have rate limiting steps. Given the rate shown in the scheme above , a cell can't make product H faster than 1 per min.

  Proposition 1.5.12. *The overall rate of a pathway can't be faster than its slowest step.*

  Moral: Don't be the limiting person.
- Enzyme pathways are also subject to *feedback inhibition.* It happens when a product of a downstream enzyme binds to an allosteric site of an upstream enzyme inhibiting the upstream enzyme.
- Enzyme pathway flux can be limited by reagent/substrate.

1.6 September 16

Today, Prof. Drennan's shirt has ATP.

We have been studying enzymes and enzyme pathways. There are number of people who are interested in manipulating pathways to understand human disease. They ask how reactions reach equilibrium and how to shift them. Moreover, they are interested in branch points, particularly in slow steps and feedback inhibition.

A graduate student at MIT uses organisms to convert pollutants to bio fuel. She can't change the temperature but can add substrates and enzymes to increase the rate of reaction that converts pollutants. Happy ending of the graduate student: works in Vertex Pharmaceutical. Let's think along the line of the graduate student to address the issue of speed of reaction.

- To fix an issue with a slow enzyme, we can put in a faster enzyme derived from a different organisms.

- We can also evolve our enzyme to be faster. It involves directed evolution.

  Fun Fact 1.6.1. Francis Arnold at Caltech won Noble prize in 2018 for her work in directed evolution.

1.6.1 Important Use of Enzyme Pathways: Cellular Metabolism

Going forward, we will focus on energy. After all, enzyme pathways are there to produce energy. And energy is really about ATP and about *metabolism*, which is life sustaining chemical processes in organisms. Some purposes of metabolism are:

- Conversion of food to energy for cellular processes.

- Conversion of food into building blocks: sugar molecular, parts of lipids, phosphates, and amino acids. We need to replenish our building blocks from time to time.

- Metabolism is also about conversion of waste into exportable waste. Remember that a box needs to be squeezed to be thrown in trash room.

There are two types of metabolism:

- **Catabolic:** It is a breaking down of complex molecules (eg. glucose, which has six carbon to carbon dioxide, which has one carbon) to make ATP.

- **Anabolic:** It involves making complex molecules from simple ones. This process usually requires ATP.

1.6.2 Metabolism of Glucose with O_2 (Aerobic Respiration)

From now on wards, we will discuss energy generating metabolism. These processes usually require a molecule with stored energy to be oxidized and a terminal electron acceptor that gets reduced.

Definition 1.6.2. *Oxidation* is the loss of electrons and *reduction* is the gain of electrons.

Most microorganisms and humans like to use glucose, which leads us to a lovely equation

$$\overbrace{C_6H_{12}O_2}^{eat} + \overbrace{6O_2}^{breathe} \rightarrow \overbrace{6CO_2}^{exhale} + \overbrace{6H_2O}^{pee}. \tag{1.6.1}$$

The net energy produced in this reaction is 32 ATP which is a billionaire status kinda thing. 32 ATP is a lot. Here, the *stored energy* is glucose (oxidized) and the *terminal electron (e^-) acceptor* is oxygen. The glucose turns into waste products: carbon dioxide and water.

Fun Fact 1.6.3. We have been thinking about reaction (1.6.1) in different contexts. In the present context, we study it because it requires a lot of enzymes. Studying this reaction in detail can take more than a third of a semester. But we will take only one lecture. Dang!

Let's get back to metabolism. They are of two types

- Aerobic: Involves air (oxygen)
- Anaerobic (Fermentation): Does not involve air (oxygen)

Aerobic respiration

Aerobic respiration involves multiple pathways and enzymes. The reaction (1.6.1) is an example of aerobic respiration. Here, glucose (energy source) that contains six carbon atoms is cleaved down to six carbon dioxides with the help of enzymes. In practice, this reaction involves

- **glycolysis:** has to do with conversion of glucose (six C) to two pyurvates (three C each)
- **pyruvate oxidation:** pyruvate gets oxidized releasing two CO_2
- **citric acid cycle:** releases two acetyl CoA (two carbon).

We can also think about enzymatic steps of this reaction. Recall that they can be endergonic (requires energy) or exergonic (releases energy).

- ATP cleavage can drive endergonic enzymatic reaction and exergonic reactions lead to ATP generation.

- Glycolysis requires two ATP to drive endergonic reactions during the first steps of glycolysis, but gets four ATPs back near the end (net two ATPs). This is a pretty good deal.

  Moral: Early investment is good.

- The citric acid cycle also generates two ATPs.

Question 1.6.4. Now that we are done with the first three steps and have four ATPs in our bag, where do the rest of the 32 ATPs come from?

To address this question, we need to look at what other processes happen during the breakdown of glucose, Figure 1.7.

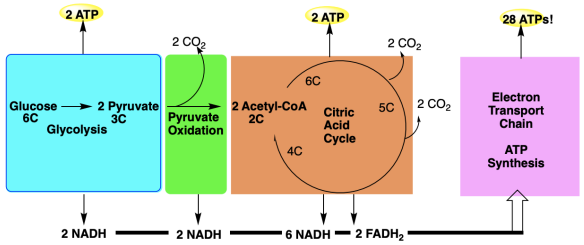


Figure 1.7: Energy formation

Let's dissect the name of NADH for a minute. N stands for nicotinamide. A for adenine and D for dinucleotide. Recall that a nucleotide has a base, sugar (deoxy and oxy), and 1-3 phosphates. Therefore it is a two electron carrier.

Question 1.6.5. What makes *NADH* a two electron courier?

Let NAD^+ be A^+, so NADH = AH. Recall that every bond has two electrons. Therefore, if we break a bond of AH, there are three possibilities. We can have A^{2-} and H^+. That's not happening here. (Why?) We can also have $\dot{A}$ and $\dot{H}$ (hydrogen atom and radical species): the dots represent electrons. This is also unfavorable. But we observe A^+ and H^- (hydride) (Why?). As it turns out H^- is same as $H^+ + 2e^-$. Therefore, NADH process involves these two electron.

Back to Figure 1.7, there is a generation of electrons in oxidation. Pyruvate oxidation generates electrons that are picked up by NADH. The citric acid cycle is also oxidation that generates elections which are accepted by six NADH and $FADH_2$. They act as electron couriers and deliver the electrons to the *Electron Transport Chain* (ETC) where electrons are coupled to ATP synthesis.

Let's now talk about Electron Transport Chain and ATP synthesis.

- ETC and the enzyme that synthesizes ATP (ATP synthase) are embedded in a membrane in mitochondria (that's why it is called the powerhouse of the cell).

- NADH and $FADH_2$ deliver electrons to ETC and those electrons are transported through the membrane of mitochondria.

- Quinones are lipid bases that help in electron transport.

  Fun Fact 1.6.6. Someone in the chat commented that human are so wack. We should use copper wire to transport electrons instead.

- The electrons end up in oxygen (terminal electon accetpor) which gets reduced to water. At this time, protons in mitochondrial matrix are pumped out of the matrix. There is a electrochemical gradient that drives ATP synthesis. In fact, when eletrons flow in (H^+) are pumped out. When these protons again come in, electrons are coupled with (Adenosine diphosphate) ADP+P_i to form ATP.

  Fun Fact 1.6.7. There is at least one Noble prize for figuring out how things worked.

Poll 1.6.8. Which of the following is an example of anabolic metabolism?

- Fatty acids to acetyl CoA
- Amino acids to protein
- Glucose to pyurvate
- Lipids to glycerol
- all of the above

Answer: Amino acids to proteins.

1.6.3 Metabolism of Glucose without O_2 (Fermentation)

Fun Fact 1.6.9. Prof. Drennan hasn't been able to win herself. During her practice lectures, she finishes way ahead of time. For some reason, she speaks slowly during real lecture and runs over time. Nothing different happened today, so we will cover this section very quickly.

Definition 1.6.10. The breakdown of glucose without oxygen is called *fermentation.*

Fermentation starts out the same as aerobic respiration. Namely, glucose is oxidized in both cases. In aerobic reaction, we go from Pyurvate to acetyl CoA. But in fermentation, we get acetaldehyde after the breakdown of glucose. Acetaldehyde is a terminal electron acceptor in contrast to oxygen in aerobic process. Here, waste products are ethanol and carbon dioxide. We use all of NADH to make ethanol and no electron go in ETC. Therefore, all the energy we get is just from glycolysis. Poor fermentation process. But hey, beer is made from fermentation.

Moral: Living with oxygen is a great thing.

This is the end of biochemistry material. There is an exam in a week from today on this material. We will start genetics in the next class.

Module 2

Genetics

2.1 September 18

From today, Prof. Lander is teaching the class. He is wearing a blue shirt.

2.1.1 Mendelian Inheritance and the Chromosomal Basis of Inheritance

In the first lecture, we saw that biology has two independent way of thinking:

- **Genetics:** study of organism minus an individual
- **Biochemistry:** study of an individual component minus an organism.

In the first Module, we talked about biochemistry. Now, we will shift our gear to genetics. The hero of genetics is Gregor Mendel. People usually don't talk about his life when they talk about genetics. In this class, we will study his life. He was like an MIT guy. Mendel was like a saint. Actually, he was a monk.

History of Heredity

European were explorers. They explored places and came back with plants and animals. They were interested in breeding creatures. Tomatoes were introduced in Europe by explorers. Italian did not have tomato sauces before that. Soon after they started building roads, they grew apples not just for family but to sell them over a larger region. They started making money. And they thought about the best types of apples.

There was a rise of commercial agriculture. Bakewell started breeding sheep for their wool. The best sheep are Marino sheep. Spaniards knew that. Folks were interested in understanding the hybridization and variations.

Let's focus on the Europeans of Austro-hungarian empire in Moravia. Bruno was a learning place and a civic boosters. Bruno was Kendall square of MIT where lots of advances in biology are made. Broad Institute is at Kendall square. People in Bruno were on top of organizing societies like Moravia society for the Improvement of Agriculture and Natural Sciences. Now, it is turned into a museum. Prof. Lander has been in the museum. The museum is small but has a terrific collection of scientific works.

In 1815, Karl Andre started laying out plans for these societies. He imagined that there will be interesting science coming out of Bruno. One of the members of the society would become a botanist who would explain things clearly. CH Nack succeeded Karl Andre. Nack had a day job as in Augustinian monasteries. He started looking for monks who could crack some codes. Monks who knew math and physics and stuff. In search of a monk, Nack found Gregor Mendel and recruited him.

Let's look at what our monk, Mendel, did as a scientist to answer this question.

2.1.2 Mendel's Experiment

Set up

He set up an experiment. People that time were working on willow trees. But Mendel decided to study pea plants. He loved pea plants because

- They are small.
- Pea plants grow quickly
- Their flowers can self fertilize because they have both pollen that produce sperm and carpel that produce eggs. We can open it up and take out the pollen with a paint brush and put it in another plant. It is easy to do breeding.
- They have a lot of variation. Seed dealers in Bruno were selling all sorts of pea plants. Mendel decided to get a lot of seeds. He bought 34 varieties and grew them. After a while he picked 22 varieties from 34. He grew them for two years and eventually settled on seven. He spent a huge amount of time setting up an experiment.
  - flower color
  - flower position
  - pod color
  - pod shape
  - pea shape
  - pea color
  - stem length

Experiment

After setting up an experiment, Mendel realized that he should have a good control. If he wasn't careful enough, the experiment would fail and he might misinterpret it. Before interpreting the result of crossing two plants, he self crossed peas multiple times. He noticed same traits in offspring after multiple generation and called them *true breed.* It allowed him to carefully interpret what happened next.

Figure 2.1: A version of cartoon by J. Chase that appeared in the New Yorker

Even after years of experiment, brother Mendel did not grow tired of peas. He enjoyed crossing strains of peas. He observed that offsprings (first generation F_1) of peas with round and wrinkled seeds (parent generation F_0) were round. That was a big deal. At a first glance, he thought that some traits disappeared. For instance, babies can be tall or short regardless of their parents. The traits blend. However, Mendel had discrete traits and they did not blend.

Definition 2.1.1. In genetics, *phenotype* refers to the traits of an organism: roundness, blue eyedness, and height.

Definition 2.1.2. The phenotype that shows up in next generation is called *dominant* and the one that does not is called *recessive.* In simple Mendel type experiment where there are discrete traits, phenotype can be dominant and recessive. When there is a continuum of phenotype (height), the words dominant and recessive become irrelevant.

Mendel did not stop in the first generation. He continued to self cross plants. And wrinkles came back in the second generation (F_2). He also got some round peas in F_2. Didn't wrinkledness completely go away in F_1? Mendel was a math and physics monk. He did one thing that nobody had done.

He counted. He counted how many were round. He counted how many were wrinkled. He got 5474 of round and 1850 of wrinkled. That was a ratio of $2.96 : 1$. Close to 3. At first, he thought that 2.96 was his answer. But he got suspicious and did not settle in one experiment. In his other experiments, he got the numbers $3.01, 2.95, 2.85, 3.15, 3.14, 2.84$

and so on. But he never got 3.00. A statistician would also have guessed it to be 3. However, it was only after decades, statistics was developed in England. It was motivated by agriculture problem. Even without statistics, he decided the number to be 3 using his mathematical intuition.

Modelling

After doing experiments, Mendel wanted to make a model that would explain the data. He also wanted to explain why he stopped seeing traits in some generation.

Mendel claimed that there are particles of inheritance. Every plant for a given trait has two particles that come with flavors. For shape of the seeds, say plants have particle R for roundness and r for wrinkledness. Mendel started with true breeds, so he assumed that round peas have RR whereas wrinkled peas have rr. In the first generation, RR gives one particle of inheritance R and rr give one r to their offspring. After the fertilization of sperm and eggs coming together, the offspring has Rr. Mendel claimed that R should dominate r. With this idea of dominance, he was able explain the roundness of peas in F_1. He continued self crossing Rr, and got Rr, Rr, rR and rr in F_2. In that model, he got a ratio of $3:1$. (round : wrinkled) in F_2.

Definition 2.1.3. From now on wards, we will refer to the particles of inheritance as *gene.*

The word was invented in the the 21st century but being asynchronous on this part of history does not do much harm.

Definition 2.1.4. The two flavors of genes are called *alleles.*

R is an allele for roundness and r for roundness. Individuals get two alleles.

Definition 2.1.5. The type of alleles individuals have are referred to as *genotypes.*

Definition 2.1.6. If individuals have same copies of alleles, they are called *homozygotes.* For instance, RR and rr are homozygotes. In contrast, if they have different copies, they are called *heterezygotes.* For instance, Rr and rR are heterzygotes.

Note: Dominant and recessive refer to phenotype not to alleles. Heights have more than four phenotypes and two words don't give the whole picture.

Publication

Mendel got his model. He explained his observations by claiming that plants have genes in a manuscript that no one would read for decades. In modern days, he would write up a manuscript and email it to a journal like *Nature.* However, there was neither *Nature* nor email.

Question 2.1.7. Anyway, what would have *Nature* done with his manuscript?

Peer review. Someone said pea review. The journal would send out the manuscript to experts who would advise whether to publish it.

Fun Fact 2.1.8. Prof. Lander showed us a picture of Mendel's manuscript and asked us if we would accept his paper. A lot of us said that we would reject them. I can't read German. Some don't like peas. There are some statistician who don't see hypotheses and test results. Others said that there is not enough data. Neither testable predictions. Tough reviewers.

Making and Testing Prediction

Although some of us said that there were no testable predictions, Mendel had made and tested predictions. Remember, we said that, in F_2, a quarter of the plants are homozygotes with both alleles R or r and half of them are heterogygotes with one allele R and the other r. We can continue to self cross to third generation and do similar calculation. In fact, we can predict that a third of them will have round offspring. For wrinkled seeds in F_2, their offspring will have $0:1$ ratio of round : wrinkled. In practice, it turns out to be true. In addition, if we are to take heterozygotes from that generation and instead of self crossing them, we can do back cross (crossing it to wrinkled stains from F_0). Suddenly, we can predict all sorts of new things. What happens if we cross some phenotypes? Our monk, Mendel, did all of that and formulated his first law.

Law 2.1.9. *For any trait, each individual has two alleles that gets randomly transmitted to their offspring.*

Mendel also considered two traits of peas (roundness, color of flowers) at a time. Each of the two traits had dominant phenotype (uppercase letters) and recessive phenotype (lower case letters). When Mendel crossed them, he got double heterozygotes. They inherited alleles from both round and wrinkled parents and alleles from both parents that have different color of flowers. Mendel crossed these heterozygotes and saw that the alleles from two traits did not interfere with each other. See Figure 2.2. When we do math we get 1/4th of peas with different traits. Different genes are independently inherited. With this observation, Mendel came up with his second law.

Law 2.1.10. *Independent assortment of multiple traits.*

After these seminal contributions, Mendel never wrote good papers. He stopped contributing to science. Some say that he studied hogweeds. Other say that he spent his life as a monk. He also sank like stone in water. Nobody read his paper that would be the foundation for genetics when people rediscovered doing some weird genetics.

2.1.3 Chromosome Theory

Decades after Mendel wrote down his laws, people in Germany were looking in cells through microscope. Dye industries were flourishing that time. People threw dyes in

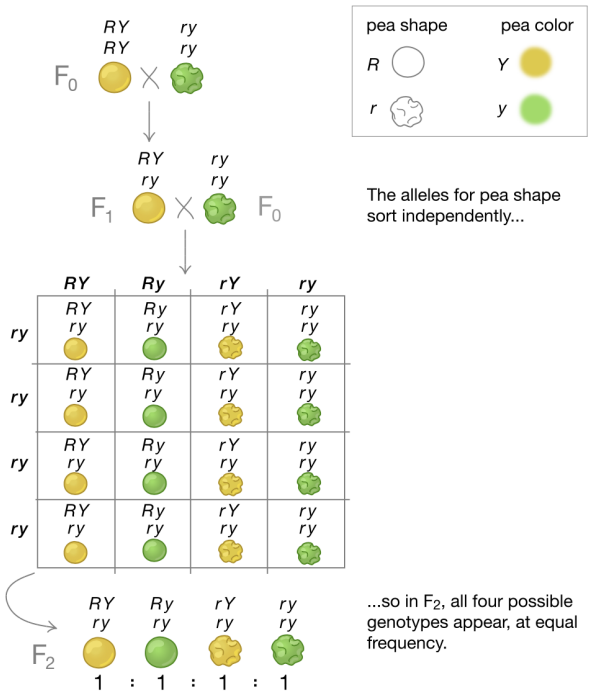


Figure 2.2: Mendel's Second Law

cells and saw all sorts of things. In particular, they found funny colored things. These things absorbed dyes but nobody knew what they were. People called them colored things. It did not sound scientific. A remedy was to choose some fancy Greek word. Color means chromo and thing or body means somes. Chromosomes. These people were smart enough to cover up their ignorance. They had no idea what they were looking at. The chromosomes underwent a choreography. They lined up into two daughter cells. In 1882, Wlather Flemming drew pictures of these choreography. The pictures allowed to see that there was a process of cell division:

- A cell is still intact with X shaped chromosomes. Initially, they are not lined up.
- The chromosomes line up in the mid axis.
- They get pulled apart form daughter cells.
- Cells that start with $2n$ chromosomes end up with $2n$ chromosomes

This process is called *mitosis*. It occurs in our body cells. There is another process called *meiosis*. It is usually in sperm cells and gametes. Something different happens in this process.

- There is no replication.
- The chromosomes line up in pairs instead of in the mid axis. Chromosomes have different sizes.
- Those pair divide into 2 XX.
- Then they undergo second division like in mitosis.

- Cells that start with $2n$ chromosomes end up with n of them.

Chromosomes come in pairs. When we make gametes each sperm gets only one of the pair. Didn't that dead monk say something about it? Yeah am I am talking about our hero, Mendel. He vaguely talked about particles of inheritance coming in pair. However, everything Mendel said was abstract nonsense. What should we expect from a mathematician? He talked about particles of inheritance. But what are they? He did not show us. It was only decades later, people saw chromosomes. If genes were in chromosomes, it would explain Mendel's first law through mitosis. People were excited about chromosomes theory.

What about the second law? Well, that's great because they just have to be different chromosomes. Which one lines up in one side or other side during meiosis is independent. But there is a problem.

Question 2.1.11. What if the genes are in the same chromosome?

They have to travel together. And the lining up would be dependent. And there is no little d-big D, little e-big E.

Question 2.1.12. Mendel's laws or chromosome theory?

They are beautiful partners if genes were in different chromosomes. But who's right in general? We don't have time to find out today. Mendel is our hero. He might be right. Chromosomes could also be right. Tune it for the next episode of 7.012.

2.2 September 21

Welcome to the second episode with Prof. Eric Lander. We ran into a deep and serious contradiction between our hero and chromosome theory last time. We will try to resolve that contraction today.

2.2.1 Mendel's Law vs the Chromosome Theory

Mendel used abstract nonsense to make sense of results of his pea experiment. In particular, he used a vague notion of particles of inheritance called genes. They come in different alleles. Each organism has two alleles: one from dad and other from mom. Dad and mom pass one of their alleles to their gametes: sperm and egg. When these sperm and egg meet, we get a diploid number of alleles. This resolves Mendel's first law.

Question 2.2.1. How do we take independent assortment into account in Chromosome theory?

To answer this question, we need to look at chromatids (limbs) of chromosomes experimentally and ask what this question means in terms of chromosomes. During meiosis, the chromosomes might line in any way. The chance of each line up is 50-50 in every meiosis.

It means that if the genes are in separate chromatids, the assortment is independent. But what will happen if alleles of two genes say $A - B$ are on the same chromosome of dad and $a - b$ are on the same chromosome of mom? They will be transmitted together and we will only get AB and ab in offspring. There is no other way to line up, so we never see non-parents. In particular, aB and Ab will disappear in the next generation.

2.2.2 Thomas Hunt Morgan and Fruit Flies

Remember that this class is driven by experimental methods. It requires new organisms and new experiments to resolve the question we posed in the previous section.

Thomas Hunt Morgan was a famous embryologist. He preferred empirical and experimental method over abstract nonsense. He studied all sorts of embryology but he detested Mendelian genetics which was rediscovered in 1900s. Folks were excited about Mendelian inheritance. But Morgan shunned it as he said what the 'factors' were in Mendelian experiment.

> In the modern interpretation of Mendelism, facts are being transformed into factors at a rapid rate. If one factor will not explain the facts, then two are involved; if two prove insufficient, three will sometimes work out ... that the results are often so excellently 'explained' because the explanation was invented to explain them. We work backwards from the facts to the factors, and then, presto! explain the facts by the very factors that we invented to account for them ... So long as we do not lose sight of the purely arbitrary and formal nature of our formulae, little harm will be done; and it is only fair to state that those who are doing the actual work of progress along Mendelian lines are aware of the hypothetical nature of the factor-assumption.

Although he disliked Mendelian theory of inheritance, Thomas H. Morgan became the greatest geneticist of his time within the next couple years of its discovery. He brought rigor. Morgan was suited in upper Manhattan. He chose fruit flies. Drosophila. Because he could get generations of flies within two weeks. Flies are very small but have interesting bodies. They come in different body color. Wings. His goal was to experiment with these traits.

In order to study genetics we need to study mutants. Mendel had peas. And Thomas had flies. There was no market for fruit flies. Now, there is one in Bloomington. Morgan had stock of flies. He began to notice mutants. White mutants when he crossed. Normally, fruit flies are dark brown colored. He discovered the mutant about the same time he and his wife had their first child. When Thomas returned from his lab, Lilian (also a biologist) would ask, "How is the white fly?"

"How is the baby?" he would ask.

Both of them raised mutants and children together. A typical body is in Figure 2.3. But mutants have pathetic wings, curled wings, and white colored body. Some eyes are dull

reddish in color. Other have cinnabar eyes, vermilion eyes. Shapes of the eyes are also different. Thomas crossed them like Mendel crossed his peas.

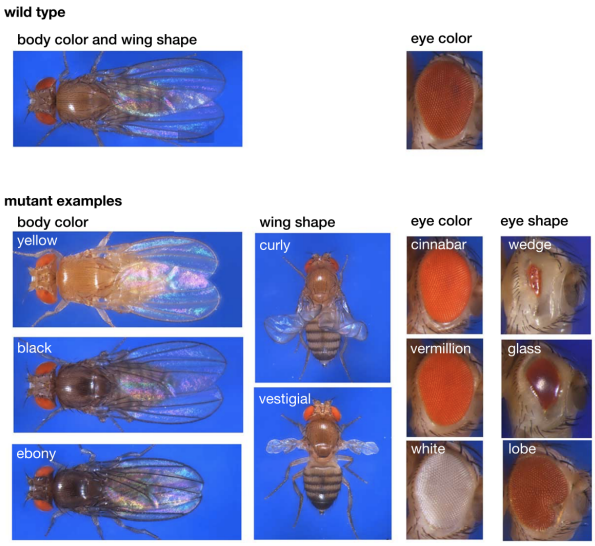


Figure 2.3: Fruit flies by Holtzman S and Kaufman T (2013). Flybase.

2.2.3 Fruit Flies Cross

When Morgan crossed flies with vestigial wings and wild type wings (normal phenotype wings), he found that F_1 were all wild type. In other words, the vestigial wings were recessive. We choose vg for vestigial and $+$ for normal.

Question 2.2.2. What happens when we cross two genes?

The first generation consists of all the white type wild color. When we back cross with black body (b) and vestegial wings (vg), based on Mendel's theory, we get the following phenotypes: ++ (normal body, normal wings), +vg (normal body, vestigial wings), b+ (black body, normal wings), and bvg (black body, vestigial wings). What if they were on the same chromosome? We would only get parental types. We would never see a non parental.

2.2.4 Hypotheses: Recombination

People saw chromosomes squeezed on top of each other. They went back to Greek words and named that phenomenon *chiasma.* Further, they hypothesized it involved exchange of material between two *chromatids* (arms of chromosomes). It was hard to see the exchange because chromosomes did not have distinguishable colors.

Question 2.2.3. Did Thomas get overwhelmed with the idea?

Morgan hated random explanations. He discarded Mendel's particle of inheritances. He though that they brought wacky ideas of exchange to support inheritance. They could come up with more explanation and *ad infinitum.*

When we have a really hard problem we need a young person. A person who can think fresh. Our young person was Alfred Sturtevant. He was the greatest sophomore UROP ever working with Morgan in Columbia, New York. Sturtevant computed the recombinant frequency (RF)

$$RF = \frac{NP}{T}$$

where NP is the number of non parental type (recombinant) offspring and T is the total number of offspring. As an answer to Question 2.2.2, Alfred got 17% RF in F_2 instead of 50% as predicted by Mendel's theory. Sometimes it was 5%, and other times it was 10%. "Could you just give me the data of all the crosses that you have done?" Alfred would go around everybody and ask. He pulled the greatest all nighter after collecting the data that he would later write in his autobiography. He cracked the problem blowing off all of his assignments.

Fun Fact 2.2.4. Discoveries will be granted excuse at MIT from submitting the assignments.

2.2.5 Linkage Map

Alfred came up with an idea that there is a recombination of chromosomes but they follow a rule. There is a line (linkage map) of alleles and the distance between them gives a degree of recombination and the degrees add up. Assuming that chromosome theory is right, b is somewhere and vg is somewhere in the line. They have 17% cross over, so they are 17 units apart in the linkage map.

When vg and cinnabar were crossed together, Alfred got 8% RF. "Where would I put cinnabar?" Suppose that cinnnabar is over right then the recombination between b-cn would be 25%. He checked in the notebooks and found that a combination of 9%. He concluded that cinnabar lies in between b and cn in the linkage map. He noticed that lobe was 5% away from vg. He checked the lab note and it was 13 from cn. The black side was also 22. He looked at curved. It was 3%. Based on a small amount of data he predicted the recombination of everything else. First, he claimed that genes lay on a line. Then, he became a geneticist.

Why linkage mapping?

Now we might be wondering why linkage mapping is useful at all. It is because

- It establishes chromosome theory of inheritance.
- If we have new mutants, we can figure out where they lie in the line.

Later, when molecular biology comes along, we will see that we can find a DNA segment in chromosomes if we want to find genes of scarlet eyes. Knowing the position of DNA segment will allow us to clone genes. The same method works for cloning in humans to cure diseases.

Digression: When we talk about 10% RF, we just mean the relative recombination. In the line, they could be at 20% and 30%. In practice, if we had 12 intervals we might sometime get a recombination of 120% between the genes at the end points. But we assume that 20% is due to statistical error. When we have two 10% and 10% it could add up to 18%. For practical purpose, significantly small errors are not relevant. But we can't ignore it. The recombination frequency maxes out at 50% in Poisson fashion. It is illegal to have 120%.

2.2.6 Sex Chromosomes and Sex Linkage

Alfred showed that pairwise recombination cross checked each other with the pairwise distances in the linkage map. With this strong evidence, we have a ground for chromosome theory.

In fact, there are more evidences. Chromosomes in human beings come in pair. There are 23 pairs of chromosomes of which 22 are identical in males and females. Males have one big chromosome lined up with a tiny chromosome, Xx. In contrast, females have both big chromosomes, XX. Chromosomes are numbered by sizes. We will learn in genomics that whoever numbered according to the size was wrong. Some smaller ones turned out to be slightly larger than the others.

Fun Fact 2.2.5. Before 90's, we used to have 24 pairs of chromosomes. Everyone confirmed it until some legit guy discarded it as he did not see the 24th. "I was having trouble with 24th as well," said those who confirmed at first and went into a bandwagon.

In contrast to human beings, male in birds are *homogametic* (XX) and females are *heterogametic* (Xx). In worm, females have XX and males have X only.

Question 2.2.6. What does this have to do with genetics? Do these chromosomes determine sex?

This sounds obvious at first but there is no reason to believe that chromosomes determine the sex of an organism. There are other sexual dimorphism. Bodies of males and females are different. They play different reproductive roles. It might be the case that males chew up their bigger chromosome, so the difference in chromosome could be a result of sex instead. This will bring us to our last topic of the day, mutants.

Remember, we said that genetics is about studying organisms minus a gene (mutants.) Morgan studied mutants in fruit flies. He crossed a white eyed male fly with chromosomes say X^W/Y and a normal female with chromosome X^+/X^+. Had this trait been a normal autosomal recessive—autosomal means that it is not on sex chromosomes—the female offspring would be normal. And if we crossed them with normal male, all their offspring

would be normal. It would mean that there is no sex linkage with chromosomes. However, Morgan found that half of the male offspring were affected and none of the females were affected. The trait was linked with the genes.

Males in F_1 get X^+ from mom and Y from dad. Meanwhile, females in F_1 get X^+ from mom and X^W from dad. Now, if we cross the female with a normal male X^+/Y, half of the offspring will have white phenotype. In particular, half of the males will have X^W/Y (hemizygous) and half of them will have X^+/Y (homozygous). Moreover, half of the females will have X^+/X^+ and half of them will have X^W/X^+. It means that all the females are normal but half of them are carrier. And the hemizygous males are white. It was an evidence that genotype was linked with sex.

Finally, people settled in chromosome theory but accepted that recombination occurred. They also saw chromosome differences in males and females and hemizygotes. These oddities convinced people that chromosome theory should be right. Next time, we will come with a strange organism. Not peas. Neither fruit flies. But Human. That will be the beginning of the study of our genetics.

2.3 September 25

We should check Splash! It is a great opportunity to teach.

Announcement: Yom Kippur is a Jewish festival. Prof. Lander will be out on Monday but there is an online lecture posted on Canvas that covers the material. Dr. Morril will review the content on Monday lecture time.

Today, we will focus on human genetics which is what Prof. Lander works on. In fact, we will dive into the mind of human geneticists.

2.3.1 Rediscovery of Mendel in 1900

We have been talking about genes and proteins. Genetics and biochemistry did not have a common language until the beginning of 20th century. Mendel was so far ahead of time that his paper got lost. But the discovery of chromosomes in plants and fruit flies brought us back to Mendel.

In the previous sections, we talked about peas and flies. We might be wondering if scientists wondered about humans at all. There is a wide range of normal variation in human appearance: height, weight, hair color etc. Scientist got interested when they saw how traits clustered in families. They vaguely explained it until Mendel came up with his theory. Soon they started looking at extreme variations.

- **Familial Gigantism:** The sons are 7.5 ft tall. And sisters are very short. See Figure 2.4.
- **Polydactyly:** Poly means many. Dactyl means finger. This also seemed to cluster

Figure 2.4: Familial Gigantism by W. W. de Herder

in family, see Figure 2.5. “Wait, you guys don’t have six?”

Figure 2.5: Polydactyly derived from Wikimedia

- **Hemophilia:** It means lack of blood clotation. People knew that hemophilia was a familial transmission in the family of Queen Victoria way before genetics. Frederick William died at 3. Henry died at 4. Gonzalo bled to death after accidents. Note that only males have hemophilia.

- **Musical Ability:** People started to see Mendelism everywhere. People collected data on musical ability of people. Almost all of the offspring of musically inclined were musical. They thought that musical ability was well passed on. However, musical ability is both about nature and nurture.

- **Thallasophilia:** Thallaso means sea. Philia means liking. There were families where members liked sea. Some were naval officers. Some sea travellers. People took this as an "evidence" of Mendelism.

- **Insanity:** In medical parlance, insanity is called neuropathy. There was a paper with Rr's and Dd's by a physician in New York Prof. Lander found while he was reading a paper on fruit flies by T. H Morgan. The physician concluded that insanity was passed on to offspring.

- **Pellagra:** It means skin lesions. A congressional committee in 1920 concluded that the disease was inherited. But it turned out to be caused by malnutrition. In the US, malnourishment was common: malnourished parents had a chance of having malnourished children.

- **Eugenics:** Mendelism started to proliferate so much that people started eugenics movement. They started family fitter contest in Kansas in 1920. On the negative side, people did pro-eugenic demonstrations in Wall street. "I cannot read the sign. By what right have I children?"

  "I must drink alcohol to sustain life. Shall I transfer the craving to others?"

2.3.2 Recognizing Inheritance in Human

In practice, the study of genetics in human is different from that in flies and peas. Flies have hundreds of offspring and a very short lifetime. In contrast, the world record so far in human being is 69 children. In general, we have small families. Moreover, we can't arrange crosses like that in flies. It is unethical to manipulate people.[1] Genetics in human is therefore observational instead of experimental.

Complexities of Real Human inheritance

There are a lot of complications in studying genetics in human beings:

- Some traits are limited to particular sex. For instance, males have breast tissues but it is uncommon for them to have breast cancers. Females have 60% risk of getting breast cancer while it is 1% in male.

[1] But why is it ethical to manipulate flies and peas?

- Often times, we have incomplete penetrance. The mutant genotype does not always cause disease. Female only have 60% chance to have breast cancer not 100%. Later in the course, we will see that some other cells have to undergo change for this mutation to give rise in breast cancer. Factors like age, environment, and diet might also play a role.
- Finally, there are polygenic inheritance: multiple genes are responsible for a phenotype. For a complex diseases, alleles of many genes contributes to the risk factor. However, in this course, our stories will be Mendelian where a phenotype is dependent only on one gene.

2.3.3 Inferring Inheritance: Example 1

Notations: Squares and circles represent male and female respectively. Filled symbol denotes an affected person while unfilled symbols mean unaffected. If a person is deceased or their phenotype is unknown we use a slash.

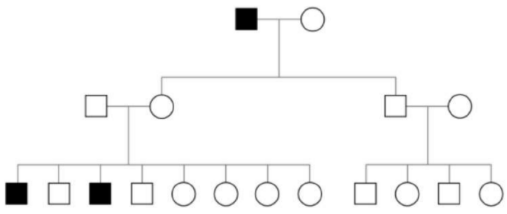


Figure 2.6: Example 1

Consider a pedigree in Figure 2.6. Three males are affected by a trait.

Question 2.3.1. Is the trait Y linked? In other words, is the trait a consequence of Y alleles?

This could probably be X linked recessive. Suppose we start with X^M/Y (M for mutant) on the left and X^+/X^+ (+ for normal) on the right. Then female offspring on the left has X^M/X^+. If we cross her with normal male X^+/Y, we get X^M/Y and X^+/Y and so on. We can try it for male offspring on the right and see that it explains the pedigree. We should check that there is no other possibility (Y linked, autosomal recessive etc.). Therefore, the trait has to be X- linked.

Question 2.3.2. What does reading out pedigree mean to us?

In example 1, we can infer that the trait is predominantly in male. In fact, the chance of getting the trait in male X^M/Y is high as he just needs X^M whereas a female should have X^M/X^M. Further, if a female is affected so does her dad.

In general, affected male transmit the alleles but offspring might not have the trait. For instance, when a male X^M/Y is crossed with X^+/X^+, the offspring are unaffected. Nevertheless, 100% of daughters are carriers whereas son aren't.

Let's look at hemophilia in Queen Vicotria's family. From the pedigree in Figure 2.7 we can infer that Queen Victoria was a carrier X^M/X^+ and Albert had X^+/Y.

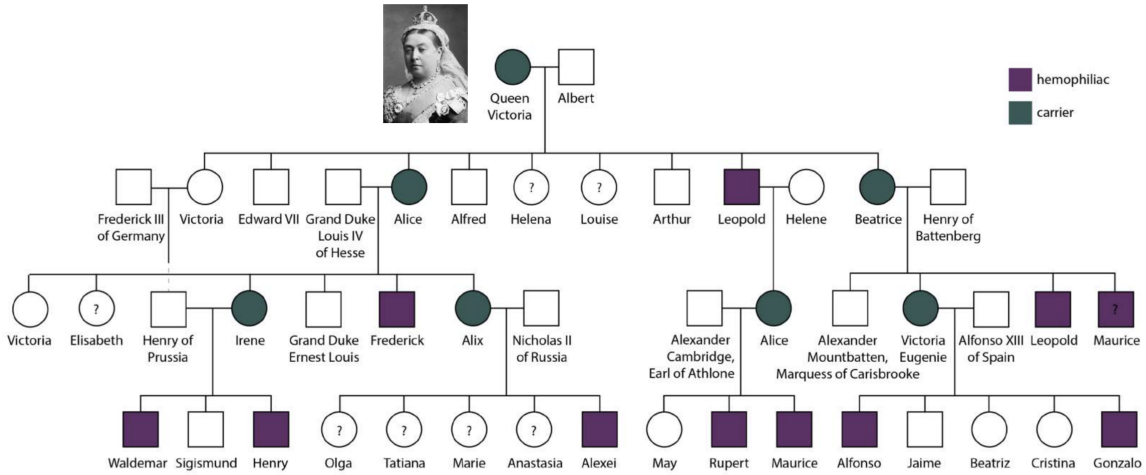


Figure 2.7: Pedigree of Queen Victoria

2.3.4 Inferring Inheritance: Example 2

Consider the pedigree in Figure 2.8. The ratio of infected male to infected female is same. In fact, in every generation, half the offspring are unaffected and the other half are infected. Therefore, this trait could be autosomal (not linked to sex) dominant. Suppose the male on the top left has chromosomes $m/+$ (m for allele of mutant autosomal chromosome and $+$ for normal) and the female on the top right has $+/+$. Half of the offspring will be affected if we follow rules for Mendelian inheritance. However, for our inference to be statistically significant, we need to increase our sample size (number of families/generation).

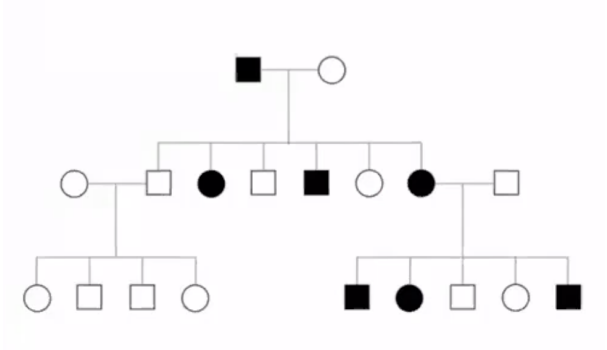


Figure 2.8: Example 2

Similarly, consider a pedigree (2.9) of Huntington Disease, brain degeneration that leads to motor disorder in the fifth-sixth decade of life. Nancy Wexler went to Venezuela to collect information about Huntington disease inspired after her mother succumbed to

the disease. She made inferences about the disease like we did. Later in the course, we attempt to find the gene for Huntington Disease.

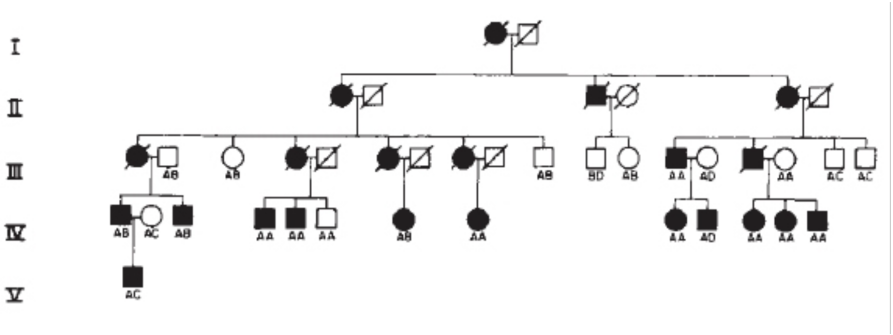


Figure 2.9: Pedigree of Huntington Disease

2.3.5 Inferring Inheritance: Example 3

Finally, let's look at a pedigree of cystic fibrosis, Figure 2.10. Mucous build up in lungs and cysts developed in pancreas of an infected child. They die early in their thirties. In fact, 1 in 2000 children in the US of European descent are infected by this disease. It is clear that the trait has to be autosomal recessive.

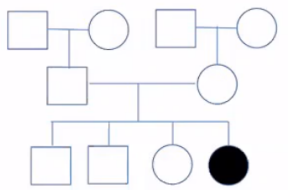


Figure 2.10: Cystic Fibrosis

Actually, Figure 2.10 isn't that of cystic fibrosis but of a phenotype associated to getting hit by a truck. The daughter on far right in F_2 was hit by a truck and nobody else was.

Moral: We need to look at a lot of families and look at the proportion of affected children to actually use Mendelian theory.

Anyway, let's get back to cystic fibrosis (see Figure 2.11). If the trait is autosomal recessive, the proportion of children getting affected should be a quarter. When we look at a lot of families, it turns out to be a third. Human geneticists were bamboozled why the numbers did not match. It is because, they were missing small families with zero cases. But when they accounted for statistical bias, they got 1/4.

Moral: In a textbook, genetics and inference is simple. But, when we collect data there are biases which we need to take into account.

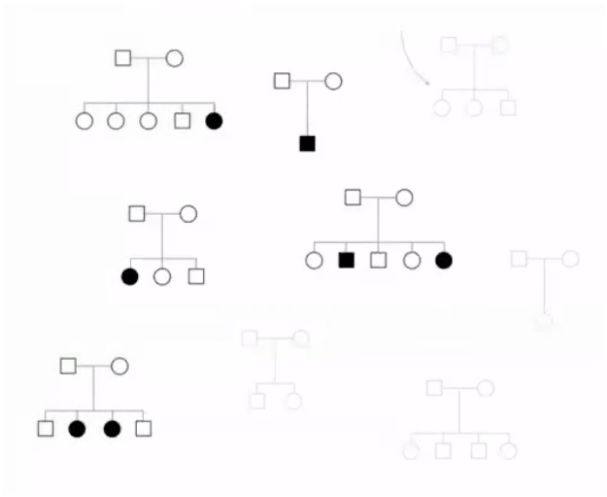


Figure 2.11: Real Cystic Fibrosis: Light ones are the families that did not have cystic fibrosis.

2.3.6 Population Genetics

We can think about population as a whole to study genetics. For instance, we mentioned that 1/2000 children have cystic fibrosis. Similarly, the rate for hypercholesterolemia is one in a million (recessive disorder).

Question 2.3.3. Why do we care about population genetics?

Say, we randomly draw two families from a population. We can estimate the frequencies of having mutant alleles among family members. In particular, we can calculate the chance of having homozygotes ((AA) or (aa)) or heterozygotes Aa offspring. Let the probability of getting A be p and getting a be q where $p + q = 1$. Then, the chance of getting A/A is p^2 and a/a is q^2. Moreover, the chance of getting heterozygotes is $2pq$. In the case of cystic fibrosis, note that $q^2 = 1/2000$. Using this, we can compute frequencies of heterozygotes as well.

2.3.7 Archibald Garrod and Alkaptonuria

In the previous section, we studied a pedigree of cystic fibrosis (alkaptonuria) but people were confused about it until 1980s.

Question 2.3.4. Why did people believe that diseases like cystic fibrosis were genetic?

To answer this question, we have to study about Archibald Garrod. If our monk Mendel is hero number one, Garrod is hero number two. He was a physician in London in the early part of the twentieth century interested in a disease called alkaptonuria. In London, when mannies washed diapers of children in basic solutions the urine turned black. It wasn't just urine turning black. Children got tendinitis, spinal problems and kidney stones.

As Archibald noted that cystic fibrosis was enriched in inbred marriage (see Figure 2.12), he used Mendelian mechanism to infer that the disease had to be genetic. In the figure, assume that asymptomatic parents passed it onto both children who passed it onto their children. The probability of two carriers of disease coming back together is higher in the first cousin marriage than that in general marriage (it is q^2).

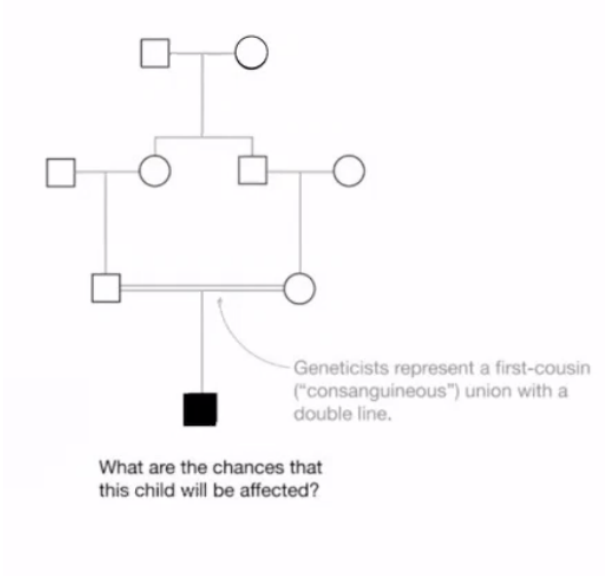


Figure 2.12: Cousin Marriage

Actually, he did more than claiming that diseases was genetic. At that time, People knew chemically that homogentisic acid (HGA) was associated with affected children. After noticing aromatic rings in HGA (see Figure 2.13) and amino acids, Archibald got an insight that there was biochemistry going on not just genetic.

"Maybe HGA are break down products of amino acids."

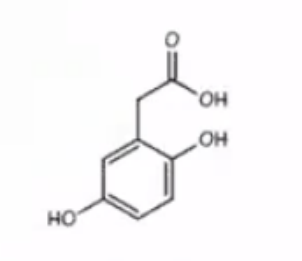


Figure 2.13: Homogentistic acid

Now it is unethical, but he fed babies with alkaptonuria lots of proteins: tyrosine, phenylalanie and even HGA. In all cases, the babies peed out HGA. He inferred that the genetic disease must affect some biochemical pathways to break down HGA.

In his lecture in 1911, our hero Garrod made the first ever connection between genes and proteins. It was a brilliant lecture but people forgot like they forgot Mendel's paper. He was so far ahead of time. People did not realize that he had just connected genetics and biochemistry.

In the next lecture, we will see how 30 years later people came to really understand genes and proteins. That they are two sides of the same coin.

2.4 September 28

Welcome to the last lecture of this module by Prof. Lander.

Decades had nothing to say about genes and proteins together. But, in 1908, Garrod made the connection by providing genetic basis for enzymes. In particular, he demonstrated that alkaptonuria was genetic by explaining its high frequency in cousin marriage. At first, his lecture had no impact in the field until the next 20 years.

Two decades after Garrod gave his lecture, Beadele and Tatum started studying how genetic mutations could affect biochemical pathways. Although it was a terrible choice, they worked with fruit flies. Despite the difficulty to purify the mutated region, they managed to show that some eye color mutants could be rescued if the ambient tissue of eye was wild type. Most of the times, the eye color was autonomous. It depended only on the genotype of the cells in the eye. However, it wasn't the case for some of them. The neighboring regions rescued the eye through a biochemical process.

2.4.1 Yeast as a Model Eukaryote

As it was hard to purify just the eyes, people started working on single cell fungus called Neurospora. Nowadays, they don't use Neurospora. They work in yeast (Baker's yeast). Today, we will talk about what Beadle and Tatum did but taking an example of yeast.

Recall that eukaryotes have nucleus and undergo mitosis. As an eukaryote, yeast has a diploid cell with $2n$ chromosomes and undergoes mitosis. The number of chromosomes pair n is 16. Yeast undergoes sporulation (meiosis) to form haploid cells which have n chromosomes instead of $2n$. Haploid cells (gametes) have two mating types a and α. They mate to form diploid cell which can undergo mitosis. Note that human gametes (sperm and eggs) can't undergo mitosis. As haploid and diploid can fend for itself, we can study both stages of yeast.

Suppose we have carbon source, nitrogen, phosphorous and salt in the petri plate (but no amino acids and lipids). Though an elaborate biochemical pathways, yeast can make lipids and acids for itself.

Definition 2.4.1. The medium that has insufficient nutrients for the yeast to grow is called *minimal medium* whereas the medium that has enough nutrients for the yeast to grow is called *rich medium*.

Yeast can grow in minimal medium does not mean that it is stupid. If it is in rich medium (with a lots of amino acids) it avoids the synthesis of amino acids. It means that the presence of amino acids inhibit the enzymes synthesis pathways. Inhibition of enzyme pathways is the basis of connection between genes and proteins.

Now let's carry out the experiment:

- Grow a test tube of yeast in (rich) liquid medium.
- Take a petri plate consisting of agar forming a soft solid support with various nutrients.
- Dilute yeast appropriately and pour in petri plates so that single cells will land in different places.
- Incubate them for some time each of the cell will grow up into a pile of identical cells called colonies.
- Transfer each colonies to rich medium/minimal medium and find the mutants.

2.4.2 Mutant Hunt

Beadle and yeast were excited when yeast grew in minimal medium. But they got interested in finding mutants that would not survive in minimal medium. This brings us to mutant hunts.

Question 2.4.2. How do we find yeasts that are unable to carry out synthesis?

- Plate a bunch of yeast in a rich medium. Many of them will make colonies some of which will be defective.
- Transfer the colonies in a minimal medium. Originally, people used tooth picks to transfer. Some of the transferred colonies stop growing.
- Add different nutrients to see if the mutant grows. For instance, we can add and Arg. suppose the yeast does not grow when we add Phe, Tyr, and Leu but grows when we add Arg. From this we ca infer that the defect in the mutant was the inability to make Arg on its own.

We can make a collection of mutants which can't grow in certain medium but in other medium by this procedure.

Definition 2.4.3. *Prototophs* can grow on minimal medium. Proto-basis. *Auxotrophs* need help of supplements. Aux-help. Auxiliary.

Our mutant was auxiliary that required Arg.

Definition 2.4.4. *Genetic screen* is a process of testing loss of ability (to do something) in each colony one at a time.

Definition 2.4.5. Suppose we want to figure out if a mutant gained ability (eg. drug resistant). We put the drug in a plate and a mutant is the only one that grows. This process is called *genetic selection.*

Tricks of Mutant Hunting

The first trick is tooth pick. Esther Lederberg being bored with toothpick went to her make up kit:

- Took out a velvet pad.
- Sterilized it in an autoclave.
- Put it down on the petri plate containing colonies.
- Picked up and put it on a petri plate of interest and repeated.

The process described is called *replica plating.*

Question 2.4.6. Do we grow haploids or diploids?

Diploids have more copies of chromosomes. If the mutation is recessive we might not see the effect. In diploids, both of the alleles have to be affected. In contrast, we can find both recessive and dominant in haploids even if one of the alleles to be affected.

2.4.3 Characterizing Mutants: Dominance Test

Suppose we have collected mutants after mutant hunt that are rescued by Arg. In the following section we will discuss the process of characterizing mutants in groups.

Question 2.4.7. Are they recessive or dominant?

We mate mutant haploids (m) with normal haploids (wild type) of the opposite mating type (+). We get a diploid cell with genotype $m/+$. If the diploid can grow in minimal medium, the phenotype is recessive.

2.4.4 Characterizing Mutants: Complementation Test

Suppose four of the yeasts (say m_1, m_2, m_3 and m_4) have recessive phenotype. Note that different mutation could hit the same gene.

Question 2.4.8. Are the mutations in four different gene or in different gene?

We can cross the mutants and wild type (wt), see Table 2.1. If the mutations are in different gene they complement each other (the daughter cell will not have mutation and is represented by + in the table.) In fact, the diploid will be doubly heterozygous. If the haploids fail to complement each other (represented by $-$) then we know that the mutations are in same gene.

Definition 2.4.9. The process of figuring out whether mutants complement each other is called *complementation test.*

Remark 2.4.10. Complementation test can be used to find recessive traits only.

Definition 2.4.11. *Complementation groups* are the mutants that are unable to complement each other.

Complementation groups define the number of genes. In particular, the number of blocks in the table is the number of genes. In our case, there are two blocks, red and blue. Therefore, there are at least two strains of yeast. To count the number we did not require any molecular biology or DNA sequencing.

	wt	m_1	m_2	m_3	m_4
wt	+	+	+	+	+
m_1	+	−	−	+	+
m_2	+	−	−	+	+
m_3	+	+	+	−	−
m_4	+	+	+	−	−

Table 2.1: Complementation group

Question 2.4.12. Suppose a congenitally deaf woman marries a congenitally deaf man. Will their child be deaf?

Maybe. In a case that Prof. Lander knows, the babies weren't deaf. It says that the mutation in parents was in different genes and they complemented each other. We can infer that there are different mutation that can result in deafness.

2.4.5 Characterizing Mutants: Epistasis Test

In this section, we will look at the biochemical pathways to synthesize Arg. Suppose α is transformed to Arg in the following pathway

$$\alpha \xrightarrow{E_1} \beta \xrightarrow{E_2} \gamma \xrightarrow{E_3} Arg$$

with the help of enzymes E_1, E_2 and E_3.

If we know the biochemical pathway a priori we can rescue a mutant by finding where the pathway is blocked. Imagine that we have mutations in enzyme E_1 or E_2 or E_3, but we don't know. Suppose the mutation is in E_1. When we give a lot of γ or β the mutant will grow well. Suppose that the mutation is in E_2. The mutant won't grow if we give α but will grow in γ.

Two mutants could have same phenotypes. Suppose we have homozygous mutant, E_1E_2, with mutation in both E_1 and E_2. It will grow if we give γ but not in α or β. In some sense, the mutation looks exactly like that in E_2. Suppose the mutant E_1E_3 has mutation in E_1 and E_3. It will grow only with Arg. Therefore, the mutation looks the same as that in E_3.

We can distinguish between the same phenotype by a process called *epistasis. Epi* means on top of and *stasis* means standing. E_1E_2 looks like E_2. Note that the later step in the biochemical pathways stands out. In E_1E_3, E_3 stands out. Therefore, by crossing

mutants and getting double mutants we can tell the order of enzyme pathways. In fact, we can also study the order of effects in some developmental processes just by figuring out which mutation is epistatic.

Fun Fact 2.4.13. This is what Beadle and Tatum did and got a Nobel prize in 1958.

> In this long, roundabout way, first in Drosophila and then in *Neurospora,* we had rediscovered what Garrod had seen so clearly so many years before. By now we knew of his work and were aware that we had added little if anything new in principle. We were working with a more favorable organism and were able to produce, almost at will, inborn errors of metabolism for almost any chemical reaction whose product we could supply through the medium. Thus we were able to demonstrate that what Garrod had shown for a few genes and a few chemical reactions in man was true for many genes and many reactions in *Neurospora.*

2.4.6 Class by Dr. Morril

Today, Dr. Summer Morril is teaching the class instead of Prof. Lander. He is celebrating Yom Kippur. We covered polling questions and reviewed some topics.

- Recombination frequency (RF) is lower the closer two genes are on a chromosomes.
- To find the mode of inheritance, we need to look at a larger pedigree for statistical reason.

Module 3

Molecular Biology

3.1 September 30

Prof. Cathy is back with her geeky T-shirts. Today, there are double helical structure of DNA all over her shirt. She also has a bag pack that has prints of elements that are important for life. We should contact her if we want to know more about her geeky dresses.

In the previous modules, we learned about biochemistry and genetics. Today, we will move on to molecular biology, the field that connected biochemistry and genetics. In particular, we will address three main questions:

- What is the structure of DNA?
- How do we replicate DNA?
- How are DNA translated into proteins?

3.1.1 Structure of DNA

Let's briefly talk about the discovery of DNA.

Erwin Chargaff had found that the amount of adenine (A) equals the amount of thymine (T) and the amount of guanine (G) equals amount of cytosine (C). Moreover, Rosalind Franklin had used X-ray diffraction to get information about the DNA. It had a turn of 3.4 nm and width of 2nm. This was the key information that led to Watson and Crick's double helix model.

In 1953, Linus Pauling and Robert Corey, unaware of the data of Chargaff or Rosalind, proposed a triple helix model. In their model, phosphates are in the middle of triple helix and bases on the outside.

- Pros: Bases are accessible to read out.
- Cons: Negatively charged phosphates that point towards each other would be repulsive. DNA is unstable.

Watson and Crick model

Triple helix model along with other models turned out to be wrong. The right one is double helix model due to Watson and Crick (1953).[1] They knew about Chargaff rule and X-ray data of Rosalind Franklin.

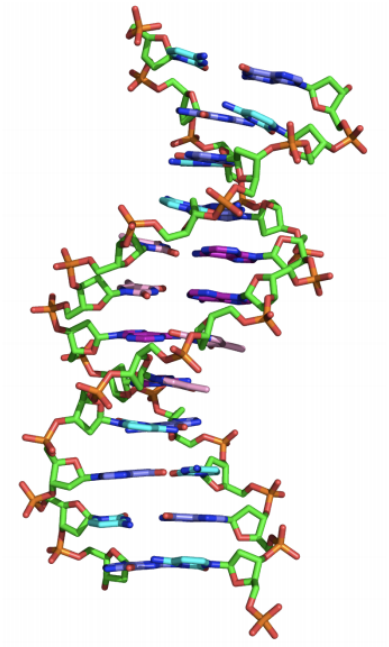


Figure 3.1: Double Helix Model

- Outside, there are phosphates (red and orange) and sugars (green) whereas bases (pink and purple blues) are on the inside.
- There are turns that give rise to major and minor grooves.
- Helices are generally right handed and are 2nm wide.
- DNA has two antiparallel strands where each strand has 5' and 3' end.
- The distance between deoxyribonulcleotides is 0.34 nm, and there are 10 deoxyribonucleotides per turn.

[1]There is a debate that they used the data of Rosalind Franklin unethically. Watson and Crick's treatment of women was also controversial. Prof. Cathy recommends us to watch *Picture a Scientist*, where Nancy Hopkins tells a story of how she first met Crick.

- There is complimentary base pairing (A with T and G with C) that stabilizes the DNA. Complementary base pairing means bases on opposite sides have hydrogen bond with each other. It explains the fact that A and G are in equal amount. The H bond stabilizes the DNA structure. Moreover, there is stacking (van der Waals interactions) between bases on the same strands that contributes to stability.

Poll 3.1.1. How many hydrogen bond is formed between AT and GC?

- AT makes 3 H bond and GC makes 3
- AT 2 and GC 3
- AT 2 and GC 2.

Answer: AT 2 and GC 3. It also depends on the orientation of the molecules.

The hydrogen bond pattern differs depending on the orientation of bases. But we have unique hydrogen bonding patterns. Them pattern is our genetic information. We can read out these hydrogen bonding to copy, transcribe, and translate.

The structure communicated so much about how life can be copied:

- Knowing about complementary base pairing allow us to read, copy, and transcribe DNA.
- We can store lots of information in the long chains of deoxyribonucleotides.
- We can understand mutation. In fact, mutations arise because of changes in base pair matching. Mutation is often harmful (cancer), but it is a built-in mechanism for adaptation. A basis for our evolution.

3.1.2 DNA Replication

To keep our life going on, we need to replicate our DNAs. In this section, we will see how the replication is carried out.

- A replication process produces two DNA molecules from one:
  - **Semi conservative:** Each strand is a template. Both new DNA molecules have one original DNA strand and a new one.
  - **Conservative** replication preserves the original DNA molecule and make a totally new one.

  Nature uses semi conservative process.
- Replication starts at an origin of replication *(ori)*. In prokaryotes, chromosomes are circular.
- To access bases and read a DNA, it has to be unwound. An enzyme called *DNA helicase*[2] helps to unwind DNA to form *replication fork.*

[2]In biology, ase is used at the end of the name of enzymes.

- A *primer* (a short starter strand made up of RNA) is produced by an enzyme called a primase. The sequence of the primer is complementary to that of DNA at the DNA's 3' end. In addition, the primer provides a free 3'OH to attach deoxyribnucleotide.

  Example 3.1.2. Suppose a DNA is 5'TAAATCGTACGCT 3'. Then the primer is 3'OH ATTTAGCATGCGA5'.

- The primase is displaced by an enzyme, *DNA polymerase*, that catalyzes polymerization reaction and a new strand is made.

- New strand grows from their 5' end to their 3' end with dNTPs added at the 3' end. DNA polymerase reads out the template strand while the new strand grows.

- Pryophosphate (PP) is lost from dNTP as dexoyribonucleotide is added. Cleavage of bonds linking phosphate is favorable and drives process.

  Suppose that there is G in the template. Then, (5' end of) C, as a triphosphate (dNTP), comes in at 3' end of the new strand forming a *phosphate diester linkage* (diester bond). When C is added two extra phosphates come off.

Question 3.1.3. What is happening with the second strand of original DNA?

The two strands grow differently at the replication fork:

- **Leading strand** is generated from the leading strand template and grows continuously from 5' to 3' when dNTP is added at 3' end of growing strand. This is what we described previously.

- **Lagging strand** is generated from the lagging strand template. It grows discontinuously from the 3' end of multiple primers. It is the rate limiting step.

  The primase puts down multiple primers and DNA polymerase adds deoxynucleotide in between the primers generating fragments of DNA called *Okazaki fragment.* Okazaki fragments that are generated during lagging strand replication are ligated together by an enzyme called *ligase.*

Fun Fact 3.1.4. The rate of this process in bacteria is 100 base per second (bps) while it is 50 bps in human.

3.2 October 2

"I stop for UAA," says Prof. Cathy's T shirt that she will explain later.

Last time, we saw how DNA replicates. Today, we study how well the process is carried out.

3.2.1 Fidelity of DNA Replication

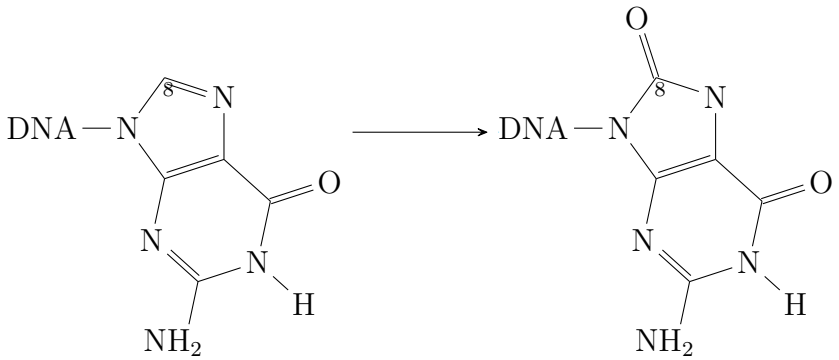
During the replication process, 1 in every 10^{10} bases are mismatched. DNA polymerase detects the mismatch and corrects it. If DNA polymerase detects a mismatch it

- slows down to allow dissociation of the incorrect nucleotides.

- If slowing down does not work, the DNA strand is flipped from the *polymerase domain* to a *repair domain* (exonuclease domain): these are the functional domain unlike the structural domains we mentioned in protein structure. In the repair domain, nucleotide is cleaved off through hydrolysis. Then strand is flipped back and polymerization (condensation) continues.

3.2.2 Fidelity of DNA Itself

Bases are chemically unstable and can undergo oxidation causing mutation. For instance, when G is oxidized, a double bond oxygen is added to carbon 8. Before the oxidation, there is a double bond between N and C. Because the nitrogen has extra electrons it forms bond with hydrogen. The resulting compound is called 8-oxo-G.



Guanine 8-oxo-G

8-oxo-G can form a hydrogen bond with C regularly. However, the DNA polymerase can get fooled and put A next to 8-oxo-G instead of C. In fact, the extra hydrogen in 8-oxo-G allows it to form a hydrogen bond with A unlike normal G which has low affinity for A. Moreover, 8-oxo-G does not fit well in the double helix because it has an extra oxygen.

Therefore, if we oxidize G in GC, 8-oxo-G can bind with A to form GA in the first round of replication. If GA goes unchecked before the second round of replication, it becomes TA changing our genetic information permanently.

Moral: Do NOT smoke. The smoke forms reactive oxygen species which can oxidized the bases damaging our DNA.

3.2.3 Mistakes in DNA Can Be Repaired

However, there are enzymes that repair the mistakes:

- **Excision repair:** Enzyme comes in and removes abnormal base and replaces it with normal base.
- **Mismatch repair:** Repair enzymes scan DNA after replication for base pairing mistakes and fix them.

3.2.4 From DNA to RNA (Transcription)

Recall that DNA undergoes replication, transcription, and translation. Previously, we learned about replication. In this section, we will study about transcription. In this process, a DNA sequence is copied to a complementary RNA sequence called mRNA (m for messenger). Then mRNA is tranlated to make proteins. There are differences in eukaryotic and prokaryotic cells but a lot of the process in transcription are same.

Step 1-Initiation

RNA polymerase binds to a special DNA sequence called *promoter* and unwinds DNA (10bps). As RNA polymersae moves, the DNA rewinds.

The promoter informs RNA polymerase

- which strand to transcribe (template)
- where to start transcription

The template strand is transcribed and the non template strand is called the *complementary strand*. *Promoter region* contains an initiation site where transcription begins. Transpciption *factors* (proteins) can help the RNA polymerase to bind activator or can stop it from binding *repressors*. This section is just a preview of what we will talk about in half a lecture next week.

Step 2-Elongation and Termination

- An RNA polymerase reads the template strand from 3' to 5' end. Unlike in DNA replication, no primer is needed.
- First nucleotide added becomes the 5' end of mRNA.
- Each nucleotide is then added at the 3' end of the mRNA. Nucleotides used are ribonucleotides NTPs–ATP, GTP, CTP, and UTP.
- Then two phosphate groups are removed because it is energetically favorable.

- When the RNA polymerase reaches the termination side the RNA transcript and polymerase are released.

  The final products of this process are:

  - the mRNA that is produced is complementary to the strand that is transcribed.
  - The non template strand (coding strand) has the same sequence as the mRNA (except U is used instead of T).

Unlike in DNA replication, there is no proofreading because mRNA has a short life time. A mistake of mRNA is erased once it starts to transcribe a new message.

Example 3.2.1. If the nontemplate coding strand is 5' ATG, DNA templates is 3' TAC and mRNA is 5' AUG. Note that coding strand and mRNA have same sequence except at U.

3.2.5 Genetic Code

Genetic code is the secret of life where the magic happens. It specifies which amino acids will be used to build a protein.

Definition 3.2.2. *Codon* is a sequence of 3 bases (3 letter word). Each codon specifies one amino acid exception stop codon.

Question 3.2.3. Why does codon come in triplet?

We have four different kinds of bases and 20 amino acids. If we have one base per codon we will get one amino acid which in total gives just 4 amino acids. If we had 2:1 that we would have 16 options of amino acids. However, a ration of 3:1 gives us 64 codons. We can have hydrophobic and hydrophilic amino acids.

There are two types of codons.

- **Start codons:** AUG.
- **Stop codons:** UAA (I stop for UAA), UAG, and UGA.

The ratio of $3:1$ arises some redundancy. Most amino acids have more than one codon. But each codon specifies one thing. For instance, UUU always specifies Phe.

The genetic code is nearly universal i.e shared by all organisms. A universal code allows for production of viral proteins in human cells. Therefore, viruses can survive in our body.

Moral: We should wear mask to prevent viruses from attacking us. It is for the same reason we should get the flu shot. We don't want viruses to hijack our body. At worst, we don't want to transmit the virus.

3.3 October 5

Prof. Cathy's T-Shirt has genetic codes and amino acid. Today, we will talk more about genetic code.

3.3.1 Types of Mutation in Genetic Code

DNA stores information in the form of genetic code while RNA communicates the message from DNA to protein. Previously, we studied the translation of messages. In this section, we will look into mutations (errors in translation) in terms of genetic code.

- **Silent Mutation:** It is a result of change in codon base but no change in amino acid. If we change the last base U to C in UGU (cys) we will have UGC which is still a cystine.
- **Missense Mutation:** Codon bases change which results in change in amino acid. For instance, in UGU (cys) the first U changes to C and UGU becomes CGU which is Arg which is positively charged.
- **Nonsense Mutation:** It is change in base that generates a stop codon. Say UGU (cys) forms UGA (stop codon). Stop codon truncate the protein. Most proteins have 200 amino acids. If we have truncated protein, we will have mutation.
- **Frameshift Mutation:** Bases are deleted or added because of change in amino acid or introduction of a stop codon. Consider AUGUGUCYSGAU where AUG (met), UGU (CYS) GAU (Asp). If we introduce G between met and cys, we get AUG GUG UGA. Now our protein sequence becomes Met Val Stop.

Poll 3.3.1. A mutation results in ca change in a codon from GGU to GGG, the resulting type of mutation is

- Silent
- Missense
- Nonsense
- Frameshift

3.3.2 Translation of Genetic Code

Definition 3.3.2. *Translation* is a process of synthesizing proteins, so it is also called protein synthesis.

There are three types of RNA involved in the synthesis:

- The message is in the form of mRNA.

- Translation is done by tRNA (t is for transfer).
- The translational machinery is rRNA (r is for ribosomes).

3.3.3 tRNA

In the last lecture, we learned about mRNA. Today, we will focus on tRNA and rRNA. The structured piece of RNA (70-90 ribonucleoties) is stabilized by hydrogen bond. Its 3D structure looks like an upside down L often depicted as a cloverleaf, Figure 3.2

Figure 3.2: tRNA

The important functions of tRNA are:

- to bind an amino acid at its 3' end.
- to bind an mRNA codon with its 3 base anticodon loop. Anti codon will read the mRNA.

Nomenclature of RNA: $tRNA^{Ser}$ or tRNAser is the tRNA for Ser.

Question 3.3.3. How many different tRNAs does a cell need?

Different species have different number of tRNA molecules. The number might be different from the number of amino acids or that of codons.

Definition 3.3.4. *Wobble position* in anticodon allows for one tRNA to recognize multiple codons for the same amino acid, so cell doesn't need 61 tRNAs.

For example codons for Ser are UCU, UCC, UCA, and UCG. The different base in third position is recognized by wobble.

Remark 3.3.5. 5' position of mRNA is read by 3' base of anticodon and 3' base of mRNA codon is read by 5' base of anticodon.

Poll 3.3.6. A tRNA with the following anticodon loop would read out the codon for aspartic acid

- 3' CUA 5'
- 3' AUC 5'
- 3' GUC 5'

Attaching the amino acid to tRNA

Aminoacyl-tRNA synthetases attach amino acids to tRNA in a process called *charging*. It is called so because the cell acquires charges. These enzymes require ATP to attach amino acids for the transfer to tRNA, see Figure 3.3.

The enzymes are specific. For instance, $tRNA^{Ala}$ is charged with Ala and not Leu or Ser, etc. If these enzymes weren't specific the genetic code could not be successfully translated.

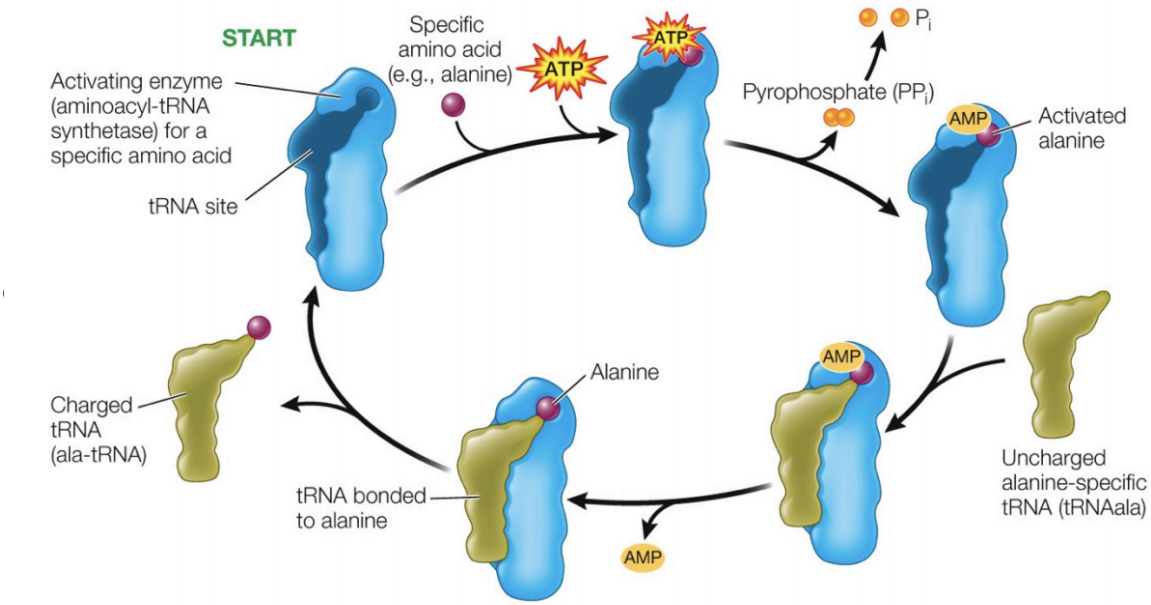


Figure 3.3: Charging of tRNA

3.3.4 Introduction to Ribosome (the Translational Machinery)

Let's now talk about ribosomes that are involved in translation. Ribosomes have both rRNA and proteins. They have two sub-units. These units are separate but are giant.

Fun Fact 3.3.7. *Omes* is used for something big. Biomes. Genomes. MIT Domes.

Although they are present both eukaryotes and prokaryotes, there are some differences. In prokaryotoes (bacteria),

- **Small subunit:** 1500 ribonucleoties and 20 proteins
- **Large subunit:** 3000 ribonucleotides and 30 proteins.

In eukaryotes ribosomes are even bigger.

Fun Fact 3.3.8. The differences in prokaryotic and eukaryotic ribosomes have been exploited in the design of antibiotics. Antibiotics target ribosome of bacteria.

3.3.5 Step 1: Initiation

Now that we have described the components that are involved, let's get into translational machinery. It starts with the assembly of ribosome with our tRNA.

- Small subunit binds mRNA on start codon (AUG).
- tRNA with an amino acid (say tRNA^{Met}) binds mRNA. The anticodon base of tRNA for our case is 3UAC 5'.
- Large ribosomal subunit binds such that tRNA is in P site. There are also E site and A site. We will clarify the terms in the next section.

3.3.6 Step 2: Elongation

The next tRNA with another amino acid say Ser attached binds to A-site in the large subunit. The Met in tRNA in P-site is released after it forms a peptide bond with Ser such that Met is on N terminus. The first tRNA, which is now free of Met, moves to E-site and is ejected. Now the second tRNA which has both Met and Ser moves to P-site.

The third tRNA (attached to say Gly) binds to A site. Ser in the second tRNA gets detached and forms a peptide bond with Gly. Similarly, the tRNA in Psite is ejected. This will form N-Met-Ser-Gly-tRNA (on Psite). Here, N emphasizes that Met is on N terminus. In this way, the process will continue to elongate the chain of amino acids.

Summary of sites:

- Incoming charged tRNA binds to A (A for amino)
- The tRNA with the growing polypeptide chain is bound to P (P for peptide)
- tRNA exists at E (E for exit).

3.3.7 Step 3: Termination

The termination involves the stop codon coming in from the Asite. The polypetide chain is released and all components dissociate.

Next time, we will talk about the differences between prokaryotic and eukaryotic cells in terms of RNA and DNA.

3.4 October 7

Fun Fact 3.4.1. Today, Jennifer Doudna and Emmanuelle Charpentier have won this year's Nobel Prize in chemistry for their discovery of gene editing technology CRISPR. Prof. Cathy knows Jennifer Doudna. Doudna started her resaerch as a side project as she got curious how bacteria protect themselves from viruses.

Moral: This is the best time to be studying biology. We don't know what else is out there.

Although we talked about gene translation in general, we did not get into detail of the differences in the process in eukaryotes and prokaryotes. Today, will touch upon some differences.

3.4.1 Genetic Material Differences

- **Location of DNA**
  - Eukaryotes: Nucleus.
  - Prokaryotes: Cytoplasm/Nucleiod.
- **Type of Chromosome**
  - Eukaryotes: Linear Double Stranded (ds) DNA that is wrapped around proteins. 2m of linear DNA, packed around the nucleus.

    Question 3.4.2. How do virus get so compact DNA? This is an open question in this field.
  - Prokaryotes: Circular double stranded DNA.
- **Genomes** of
  - Eukaryotes have *introns* (non coding regions).
  - Prokaryotes lack introns. For the purpose of this class, we won't think about exceptions.
- **Chromosomes ends** in
  - Eukaryotes have *telomeres.*
  - Prokaryotes lack telomeres.

What are Telomeres?

Definition 3.4.3. *Telomeres* are repeated sequences at chromosome ends. Repeated deoxyribonucleotide sequence is $(\text{TTAGGG})_n$3'. Enzyme that makes telomeres is called a *telomerase.*

Question 3.4.4. Does a telomerase need an accompanying primase?

Answer: No because it carries its own RNA template/primer sequence is always the same.

These repeating sequences are added because it is difficult for a DNA polymerase to duplicate the 3' end of a linear DNA Chromosomes, so the DNA might becomes shorter on each round of replication.

Question 3.4.5. Why would it be a problem for a chromosome to become shorter?

It would lose important information/lose parts of genes.

Question 3.4.6. How does adding repeat sequence help?

It provides a buffer between genes and ends of chromosomes.

Question 3.4.7. Is the repeat sequence valuable information?

No, it is a buffer.

Fun Fact 3.4.8. People have raised a link between telomerase and cancer. Recall that cancer is uncontrolled cell division. When telomerase is too active, cancer cells can keep dividing without genes being lost. It allows the spreading of cancer.

Research filling: Lots of cancer research @MIT. Angelika Amon. Sangeeta Bhatia. Paula Hammond. Engineering flavor to the research. They do differently than those in traditional medical school.

3.4.2 Transcriptional Differences

- **Location of Transcription**
  - Eukaryotes: Nucleus.
  - Prokaryotes: Cytoplasm.
- **mRNA processing**
  - Eukaryotes: A *precursor* (pre) mRNA (initial transcript) is not same as *mature* (processed) mRNA. There are three types of mRNA processing:
    - **5' cap added:** We add a cap 7-methyl-G at 5' end of mRNA. It protects the mRNA from degradation and facilitates mRNA to bind to the ribosome.

    - ∗ **3' polyA tail added:** At 3' end of premRNA we add lots of A (Adenine). The tails protect mRNA when it is exported from the nucleus to the cytoplasm for translation. It is also important for the stability of the mRNA.
    - ∗ **RNA splicing:** *Introns* are removed from pre mRNA.
  - Prokaryotes: There is no mRNA processing.

Definition 3.4.9. *Exons* are coding regions while *introns* are non coding regions that are intervening between exons.

Exons remain in mRNA and introns are removed by splicing. BEWARE! The nomenclature is pretty bad. Think of introns as intervening regions. We need to remove intervening sequences.

Lemma 3.4.10. *Not all non-coding regions are introns but all introns are noncoding.*

Proof by counter example. 5'UTR and 3'UTR are untranslated regions. These regions are noncoding regions but are not introns. These are not between exons. □

Digression: Introns have several different purposes. Sometimes they have regulatory sequence, that helps regulate whether the gene is expressed or not. Also, by splitting the gene into introns and exons, it is possible for cells to create different versions of the same protein, called *isoforms*, by just choosing to put together some of the exons but not all of them. The introns help separate out the exons so we can "pick and choose" which exons to put together to make our final protein. A great way to get protein variety from the same gene.

Definition 3.4.11. Catalyst that splices out the introns is called the *spliceosome.*

The basic mechanism of splicing is shown in Figure 3.4.

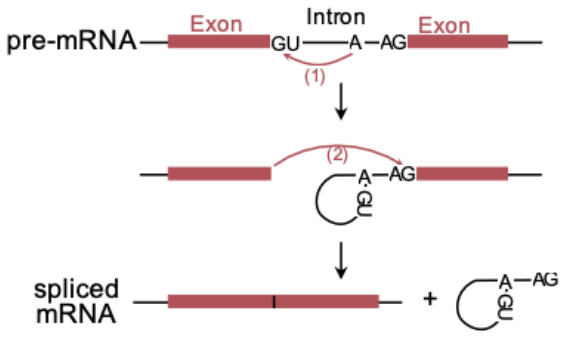


Figure 3.4: Splicing

Question 3.4.12. Why is splicing important?

- If intron (non-coding region) are translated, we could get non-functioning proteins.

- Alternative splicing: one transcript can generate more than one proteins variant by splicing different introns.

In summary, RNA processing in eukaryotes can be described in the diagram:

DNA $\overset{transcription}{\rightarrow}$ (capped adenylate) premRNA $\overset{splice}{\rightarrow}$ mature mRNA $\overset{translation}{\rightarrow}$ protein.

3.4.3 Tranlational Differences

- **Location of Translation**
  - Eukaryotes: Free ribosomes and ribosomes attached to endoplasmic reticulum.
  - Prokaryotes: Free ribosomes.
- **Ribosome size**
  - Eukaryotes: Big.
  - Prokaryotes: Smaller.
- **Translation initiation** involves
  - Eukaryotes: Riosome binding to 5' cap on mRNA.
  - Prokaryotes: There is no 5' cap. Instead there are specific sequences in 5' UTR (untranslated) region of mRNA.
- **Length of 5' UTR**
  - Eukaryotes: Longer.
  - Prokaryotes: Shorter 5' UTR in prokaryotes (3-10 nucleotides).
- **Number of polypetide chains encoded per mRNA**
  - Eukaryotes: One (*monocistronic*).
  - Prokaryotes: One or more (*polycistronic*).

  Polycistronic mRNA has multiple start and stop codon to make multiple polypeptide chains.

  Question 3.4.13. What could be the advantage of multiple proteins on one strand of mRNA?
  - Coordination: Make all polypeptide chains of heterotrimer at once.
  - Make all proteins in one pathway

Module 4

Gene Regulation and Recombinant DNA

4.1 October 9

Yesterday, Prof. Cathy's internet went down and had to come to MIT. She brought her dog for the first time after the pandemic. After seeing some grad students it whimpered with delight. Grad students treat it better than Prof.

"AC⚡GT DNA Rocks," says Prof. Cathy's shirt. Today, we will talk about gene regulation.

4.1.1 Gene Regulation

There are many kinds of regulations involved in the process

Genes (DNA) $\overset{transcription}{\longrightarrow}$ mRNA $\overset{translation}{\longrightarrow}$ Protein:

- We can regulate how much gene is transcribed to mRNA. The regulation of transcription is known as *gene regulation* (it is also said gene expression.)
- We can regulate mRNA after it is produced. This is known as *post transcriptional regulations*. Eg. splicing (can have premRNA to mature mRNA).
- We can have *translational regulation* that regulates the production of protein
- And finally, we can regulate the activity of protein after its production. It is called *post translational regulation*. For instance, we can regulate feedback inhibition.

Research filling: Gene Wei Li (biologist, uses maths a lot). Ibrahim Cisse (physicist interested in making of mRNA using single molecule techniques watching it happen in cells not in test tuve) and Seychelle Vos (biophysicist interested in biophysical methods to understand how gene that are wrapped up are able to transcribe).

Question 4.1.1. Why should we regulate the amount and/or kind of active proteins in a cell?

Recall that tRNA are charged when proteins are formed. Attaching amino acids to tRNA requires ATP. If we are making too much of protein, a lot of ATP would be wasted. Regulation will help a cell produce the right amount of proteins.

Poll 4.1.2. Which of the following could be a reason that prokaryotic cells regulate the amount and/or kind of protein they are making?

1. Food sources changes. There is no longer any glucose available so the bacteria no longer need to make proteins that metabolize glucose
2. Oxygen is no longer available. Cells need to make proteins that function during anaerobic respiration
3. Cell differentiation. A skin cell needs different proteins than a liver cell
4. All of the above
5. 1 and 2

Answer: 5

A purpose of the regulation in

- prokaryotes is to respond to changes in food/environment.
- eukaryotes is to account cell differentiation (liver cells are different than skin cells).

In terms of gene expression (regulation) there are two kinds of proteins:

- **Inducible proteins:** They are made when needed only.
- **Constitutive proteins:** They are always made in all types of cells.

There are two main methods of gene regulations that involves

- use of *reversible modifications.*
- use of *regulatory proteins.*

4.1.2 Reversible Modification

Reversible modification can affect DNA or proteins that package DNA. There are several types of reversible modification but we will focus on methylation only.

Definition 4.1.3. *Methylation* means adding methyl group to molecules.

Methylation of DNA occurs in cystine. It is common in both eukaryotes and prokaryotes. Heavily methylated genes aren't expressed. It means that we can silence a gene as the methylated gene won't send out code and therefore mRNA are not formed.

Definition 4.1.4. Enzyme that put methyl group on DNA is a DNA *methylase.* Enzyme that takes methyl group off DNA is a DNA *demethyalase.*

The function of methylation of DNA are:

- Both in prokaryotes and eukaryotes: protection from foreign DNA. Foreign DNA is often methylated to silence it.

  This could be a problem for bioengineers. Often, bioengineers want to add new genes to organisms and want those genes to be expressed. For instance, one can introduce enzymes that would speed up some pathways. Suppose an organism can convert green house gas CO2 to acetyl CoA and another organism can convert acetyl CoA to high value chemical. We can put gene of the second organisms in first organism. If this works out, we can control carbon emission in a very cheap way. Moreover, we don't produce a lot of secondary waste in contrast to organic synthesis.

  Fun Fact 4.1.5. Prof. Cathy had a grad student in chemical engineering department who tried to do this, but all the genes were silenced. He tried to figured out the methylation and demethylation but did not work out well. Eventually, he got a slightly different project for his PhD.

- In prokaryotes, methylation helps in identification of strands (that has mismatch after replication) to be repaired. Older strand is methylated. Say, G was in original DNA and got mismatched to form AG. G is methylated which helps prokaryotes figure out that G is older and correct it to GC.

- In eukaryotes, methylation helps in taking account of cell differentiation. Genes that are not needed are methylated. In cancer cells, some genes that should be off (like telemorase) are not turned off.

Definition 4.1.6. *Genetic change* is a modification of DNA that changes the DNA sequence.

Definition 4.1.7. An *epigenetic change* is a modification of DNA that does not affect the DNA sequence but can still affect gene expression.

Methylation is an *epigentic* change.

4.1.3 Regulatory Proteins in Gene Expression

In this section, we will study gene expression in prokaryotes. They are easier to study. Their genes are organized into control regions and structural genes (both found in a single *operon*) which allows whole pathways to be tuned on or off by the action of regulatory proteins.

Definition 4.1.8. *Operon* is a cluster of genes with one promoter (RNA polymerase binds to the promoter to initiate transcription).

Definition 4.1.9. *Operator* is a short sequence between promoter and structural genes (coding region) where regulatory proteins bind.

Types of protein regulations:

- **Negative regulation:** A *repressor protein* binds to the operator and blocks RNA polymerase from binding to the promoter turning off the transcription.
- **Positive regulation:** An *activator protein* binds to the operator and facilitates RNA polymerase to bind to promoter turning on the transcription.
- **Inducible gene expression:** Gene expression can be turned on and dialed up through use of inducers.

  Example 4.1.10. LAC operon: a prokaryotic inducible system for metabolizing lactose. Prokaryotes don't need lactose metabolizing enzyme when they don't have lactose, while they need it when they are running out of other sources like glucose. LAC is turned off in the first case and turned on in the second case.

  – From off to on: Initially, when the lactose is absent, a repressor is bound to the operator and transcription is off. But when there is lactose after sometime, an inducer binds the repressor and inactivates it. The repressor leaves the DNA and transcription of genes that encodes lactose-metabolizing enzymes is turned on.

  – Dialing up gene expression: When lactose is present and other sources are low, an activator can bind the operator and facilitate binding of RNA polymerase to promoter and dial up expression (transcription).

  Example 4.1.11. An example of inducible system in human is of protein called P450 in liver. They are induced when someone drinks alcohol. They metabolize alcohols. These enzymes are not specific. In fact, they can bind to tylenol and produce some toxic compounds.

  Moral: Don't drink alcohol and tylenol at the same time.

- **Repressible gene expression:** In contrast to inducible expression, we can turn the transcription off or dial down with the help of *co-repressors.*

  Example 4.1.12. NikR operon: a repressible operon that encodes proteins that are responsible for importing nickel ions into bacterial cells. *E. coli* wants nickel but too much can be toxic.

  – From on to off: A repressor binds in operator but has high affinity of binding to operator only when there is excess of Nickel. In that case, Nickel binds to repressor and acts as co-repressor stopping the transcription of Ni-importing proteins.

  – From off to back on: In absence of excess nickel, nickel dissociates from the repressor, causing the repressor to leave the operator site. It allows RNA polymerase to bind to the promoter which turns transcription back on.

Poll 4.1.13. An inducer is

- a molecule that binds to RNA polymerase, increasing its affinity for DNA
- a molecule that binds to a repressor, decreasing its affinity for DNA
- a molecule that induces repressor to repress transcription

Answer: second one

4.2 October 13

Today is a review of exam 2.

4.2.1 Genetics

Dr. Morril is reviewing this material. We will focus in two types of genetic tests:

- **Genetic Linkage:** Are two genes linked together on a chromosomes? The type of information we might be given in exam are:
  - A true breeding purple eyed (pr) dumpy (dp) is crossed with a true breeding WT fly.
  - All offspring are normal. (This implies mutant is recessive)
  - One of the normal offspring is crossed with a true breeding purple, dumpy strain.
  - All combination of phenotype are seen among the offspring: 175 purple normal winged, 175 normal eyed dumpy, 325 normal eyed normal winged and 325 dumpy.

  The types of question that we might be asked are:
  - Figure out F_i.
  - Are these traits linked?
  - Which allele are on the same chromosomes?
  - What is the recombinant frequency?
  - We should pay special attention to alleles that are travelling together. Which generation tells us this? F_0.
  - What if we wanted to see how 3 genes are linked, i.e. how do we make a linkage map?
- **Test of Epistasis:** This test allows us to order enzyme pathways.
  - What happens when you mutate 2 genes at once? What will be the phenotype of an offspring of single mutants?
  - What is the order of compounds formed in a pathway?
  - Which intermediates when supplemented will help the mutant to grow?

4.2.2 Molecular Biology

Poll 4.2.1. Prof. Drennan is wearing a Tshirt that says "I am a 5'GGC GAA GAG AAG 3'". What does that mean?

Answer: I am a GEEK.

We should make sure that we understand molecular biology because this exam is longer than previous exam and genetics takes a long time. We should

- **Replication:** DNA to DNA. Start: Unwind the DNA and copy both strands. Leading and Lagging. Primase to make primers. Machinery and End: Lagging strands form Okazaki fragments.
- **Transcription:** DNA to mRNA: Start: Transcribing includes a single strand. Machinery and End. Termination site. Gene regulation.
- **Translation:** mRNA to Protein: Start: Translating one strand. Machinery and End: Stop and Start codon. It involves ribosomes.

Types of mutations:

Poll 4.2.2. Which is a possible missense mutation for codon CGA?

- UGA
- AGA
- AGU
- CGC

Answer: AGU.

Predict the likely effect on 300-amino-acid sequence if we add one base in codon for an amino acid at position 10 of the protein sequence.

Exons and Introns: Remember, if we are talking about premRNA, it will have exons, introns, 5'UTR and 3'UTR. Mature mRNA will have exons and 5'UTR and 3'UTR. Translated part of mRNA includes only the exons.

4.3 October 16

Prof. Lander is back with a podcast called Brave New Planet. We should check it out if we are interested. Today, we will talk about recombinant DNA and a map between function, gene, and protein, see Figure 4.1.

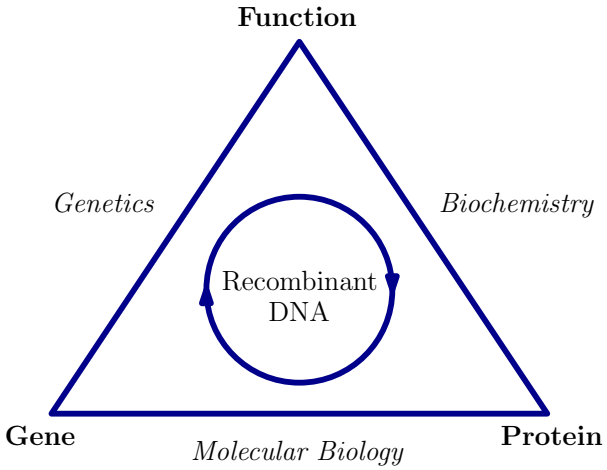


Figure 4.1: Triangle

Question 4.3.1. How did recombinant DNA technology arise?

By 1960, people declared a victory over molecular biology and moved on to solve the brain. In late 60s, the next generation of scientists noted that nobody had read an actual gene. Unlike proteins, genes look the same. Moreover, the human genome has 3 billion base pairs. A typical gene might have thirty thousand base pairs. And an individual mutation is just a base pair. It was impossible to purify one gene based on biochemical properties. We needed an entirely new method (recombinant DNA technology) of biochemical purification.

4.3.1 Cloning Overview

Definition 4.3.2. A *vector*[1] is a piece of DNA that knows how to replicate in the cell.

- **Cut DNA:** We cut human DNA into lots of fine pieces.
- **Paste into a vector:** Then we attach those pieces to a vector. Depending on the cell it could be bacterial vector or yeast vector. Pasting happens in a test tube.
- **Transform:** Now we transfer the vector to a bacterial cell. We can transform DNA into cells because bacteria slurp their environment. The vector replicates.
- **Select Cells:** Finally, we take those replicated individual bacteria into petri plate. Bacteria grow in colonies. We choose the ones that only have vectors. We are left with thousands of bacterial cells each with one piece of human DNA. In other words, we have purified each piece of human DNA.

4.3.2 Cutting DNA

Question 4.3.3. Who are the experts in cutting DNA?

[1] An actual vector has length and direction.

MIT Engineers? It turns out that bacteria are smarter than we are.[2] Consider a part of DNA:

$$AGTA\underline{GAATTC}TTACC$$
$$TCAT\underline{CTTAAG}AATGG.$$

Say we are interested in the underlined part. There is a naturally occurring restriction enzymes (ECoRI) in bacteria (*E. coli* of strain *R*) that recognize the six letter sequences and breaks the DNA chain (at *G* in both top and bottom strand). The broken double strand of DNA are held just by hydrogen bond which is very weak. Therefore, there is a double stranded break. The enzyme leaves 5' overhang with extra unpaired letters.

Remark 4.3.4. A piece of DNA should be palindromic for an enzyme to cut both top and bottom strands.

Question 4.3.5. Why in the world would bacteria build up this method?

To get rid of foreign DNA. Bacteria are unicellular and the mechnism to cut DNA is their cellular immune system. Suppose a *bacteria phage lambda* (virus that attacks bacteria) injects its genome into bacteria infecting it. Some bacteria eventually evolved so that they could recognize phage DNA which would be chopped off.

Question 4.3.6. The probability of having *GAATTC* in DNA is 1/4000. If *E. coli* has four million bases why doesn't it chop off its own DNA?

Bacteria differentiate their own DNA from foreign DNA with the help of *EcoRI methylase.* Methylase puts on methyl group on bacteria's DNA that prevents bacteria from cutting it. In contrast, phase lambda's DNA lacks methyl group and gets chopped off. Occasionally, they escape.

Restriction enzymes evolved over billions of year. Some restriction enzyme can recognize six base, other can recognize four base. In early days at MIT, biologists would have to purify restriction enzymes. Nowadays, they go to a 400 page catalog online of New England biolab. The lab sells lots of FatI and FauI. In fact, we can get ten thousands of EcoRI just for $50.

Later in the class, we will talk about CRISPR which is "programmable" restriction enzyme.

4.3.3 Pasting DNA

We need to paste the DNA back together in a vector. EcoRI adds phosphate on 5' end and OH in 3' end of the broken pieces allowing us to reseal the sugar phosphate bond in DNA to paste the broken strands.

"Bacteria do you happen to have enzyme that can seal a broken DNA?"

"Yeah I do. How many do you want?"

[2] I wish I were a bacteria.

Definition 4.3.7. The enzyme that ligates the broken DNA is called *ligase.*

Ligase is used to ligate Okazaki fragments.

Ancient people had to purify their own ligase. But nowadays, there is capitalism. We can go to catalog and buy it with $$$.

Using ligase, we ligate the human DNA piece into a circular thing called *origin of replication (ori).* Ori can replicate.

Question 4.3.8. Why do bacteria have circular pieces that replicate?

These are called *plasmids.* They carry genes for anti-biotic resistance. When a bacterium undergoes mating it shares the gene to its partner. When bacteria die, other bacteria slurp plasmid.

There are thousands of different plasmids with different features: sizes and restriction site. Some of the plasmids are natural and some are synthetic. We can get them in the catalog.

We cut the plasmids open and get source of human DNA (cut it into pieces by EcoRI). Then we mix the vector and the pieces of human DNA in a test tube. Finally, we add ligase. It seals it up the vector with human DNA.

Question 4.3.9. But what if the vectors bind to each other? What if ligase reconnects chopped stuff we didn't intend? What if multiple DNA bind?

There are all sorts of tricks of the trade. We won't answer all of them. At least, we can prevent the vector from reclosing itself. Remember that we need phosphate groups to form a peptide bond between amino acids in vector. If we take off the phosphate groups, the vector can't close. There is an enzyme called *phosphotase* that takes off phosphate group. There are lots of engineering tricks to arrange ligase to do the right things. We just need to go to catalog and find the right enzyme.

4.3.4 Transformation

Question 4.3.10. How do we transform plasmids that has human gene and restriction sites? How do we teach a bacteria to take up DNA?

Bacteria slurp stuff for a living. We can persuade them by using heat shocks to the plasmids.

We can change the concentration of plasmids relative to the number of bacteria so that a bacterium gets one plasmid. Using jargon, we want to achieve *multiplicity of infection* less than 1. In other words, the number of plasmids picked up by a bacterium has to be less than 1.

Initially, we don't know which bacteria took what parts of DNA: some bacteria get hemoglobin and some get actin. But we have purified hemoglobin and actin. Now we want the bacteria to have babies. First, we isolate them because all of them will start dividing and it will be hard for us to clone particular part of DNA.

4.3.5 Selection

We pour out the test tube containing bacteria in a petri plate and spread it out. The bacteria will make babies because the plate has nutrients. A single colony is a descendent of single bacteria most of which have 0 pieces of DNA (we can do that by changing concentration). Some of them will have hemoglobin, some actin.

Question 4.3.11. How are we going to select the colonies that have acquired the vector?

Recall that vector includes antibiotic resistance gene. If it it AmpR, it will become resistance to Ampicilin. In that case, if we add ampicilin, any bacteria that don't have AmpR in their plasmid die.

Now we have a plate that has ampicilin resistant bacteria with different pieces of gene. This is a bacterial library of human DNA. If we take a bacteria and purify the plasmid we get a sequence of piece of human gene.

Stay tuned for the next episode to know how to find the gene that we are looking for.

4.4 October 19

Last time, we formed a library of cloned DNA. But we did not solve the issue of finding the gene that we are looking for. Recall our code of arms 4.2. Studying the functions of genes will allow us to carry out the search for the gene of interest.

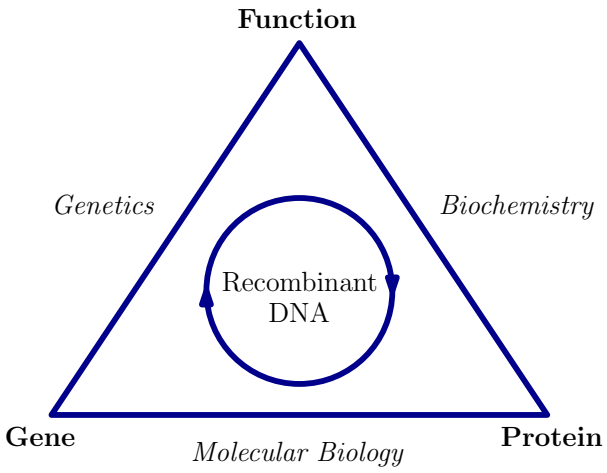


Figure 4.2: Triangle

An overview of what we did last times is:

- Cut DNA.
- Paste into a vector and replicate cells.
- Transform plasmids into bacterial cells.
- Pour bacteria onto a petri plate and select the ones that acquired plasmids.

Before we dive into finding our genes, notice that there are different ways to clone DNA depending on the types of vector, shape of the vector, where it can grow and the size of the cloned DNA (in base pairs), see Table 4.1

Vector	Molecule	Grows in	Inser size
Bacterial Plasmid	Circular	Bacteria	100-5000bp
Yeast Plasmid	Circular	Yeast	100-50000bp
Bacterial Virus	Circular or linear	Bacteria	15000-40000bp
Mammalian virus	Circular or linear	Mammals	1000-5000bp
Yeast Artificial Chromosome (YAC)	Linear	Yeast	1M+bp

Table 4.1: Ways of cloning DNA

In addition, there are many ways to prepare our DNA of desirable size:

- We can chop our DNA at every EcoRI site to completion. Sometimes we might end up chopping genes too much (say up to 4kb) that we won't be able to get the whole gene (say 15kb). In that case, we can use restriction enzyme and methylase at the same time. The methyl group at restriction site will prevent the DNA from breaking.
- We can physically shear DNA into random fragments and put *adapters* at the end that have restriction sites.

4.4.1 Expression Cloning

In this section we will talk about expressing genes. This happened in early days at MIT and Harvard. In particular, we will talk about insulin.

Definition 4.4.1. *Insulin* is a hormone that regulates the amount of glucose.

Patients with diabetes Type I don't have insulin. In the past, people purified insulin from animal pancreas. Sometimes, the insulin might have some viruses. And it is also difficult to purify. However, with recombinant DNA, they produced insulin using bacteria:

- First "find" the insulin gene.
- Put it into the bacteria.
- Ask the bacteria to make the insulin.

Question 4.4.2. Will bacteria transcribe the gene?

Recall that promoter tells the RNA to transcribe the gene. However, bacterial promoter and human promoter are so different that bacteria only recognize bacterial promoter. Therefore, we should include bacterial promoter to transcribe gene.

Question 4.4.3. How about slicing?

No, bacteria don't do splicing. Yeast would also not do correct splicing. Instead we could clone mature mRNA (it does not have introns). But this is not DNA. In fact, it is single stranded.

However, we can reverse transcribe DNA from mRNA using *reverse transcriptase* to get complementary DNA (cDNA) and make a library of cDNA instead of genomic DNA. If we add a primer with reverse transcriptase at the 3' end of mRNA, the reverse transcription starts.

Remark 4.4.4. We can get reverse transcriptase from retroviruses or capitalism (go to catalog).

4.4.2 Finding Our Gene: Penicillin Resistance Gene

For the rest of today's class, we will talk about reading out the gene. It depends on what kind of gene we are looking for.

In this section, we plan to read out penicillin resistance gene in *E. coli*. First, we put penicillin in a petri plate (gene library). All but the cells that have penicillin resistance gene will die or won't grow.

Later in the semester, we will see how we can "locate" (sequence) the resistant gene in the genome of *E. coli* containing 4M bp. For now, we will stick with what ancient people ten years ago did:

- Insert: Bacteria DNA (from resistant strain)
- Vector: Bacterial vector
- Host: Transform the plasmids into *E.coli* with sensitive strain. One of them will acquire resistant gene.
- Identification: Now we can look for something to grow in a penicillin-rich medium.

  Remark 4.4.5. We implicitly assumed that resistance is dominant over sensitivity.

4.4.3 Finding Our Gene: Based on Yeast Mutation

Suppose, we identified a yeast mutant, an arginine auxotroph (it requires Arg to grow). Let ArgA be the gene associated with the mutation. How are we going to find it?

- Insert: Wild type yeast.
- Vector: Yeast plasmid.
- Host: ArgA auxotroph.
- Selection: Plate them in a minimal medium (without arg). ArgA auxotrophs that gain gene will grow. If selection does not work, we can use screening (look at one colony at a time).

Definition 4.4.6. This process is called *cloning by function* or *cloning by complementation.*

4.4.4 Finding Our Gene: Based on a Protein

Suppose, we have purified a protein (say hemoglobin or tyrosinase) biochemically, and we want to find the gene that encodes the protein.

- Insert: Human cDNA.
- Vector: Bacteria, yeast, human cells, chinese hamster cells.
- Host: Chinese hamster cell (anything that can express the function would work).
- Identification: Put a little piece of filter paper to make a replica plate. Note that we can get antibodies that can bind to specific proteins. Therefore, if we put the antibodies the recognize tyrosinase, they will sticks to cells that transcribe protein.

4.4.5 Finding Our Gene: Based on Human Disease

So far, we figured out our gene of interest based on function. In the case of Huntington disease[3] where

- We do not know proteins that are responsible for brain degeneration.
- Huntington disease is seen in human but not in bacteria.
- We also don't know how to recognize the disease in human cells. On the other side, infecting human deliberately will be unethical.

We will come back later in the course to answer how to find genes without knowing cellular function or protein. It involves DNA sequencing.

[3]It leads to a brain degeneration in the fifth decade of life which leads to death.

4.5 October 21

Last time, we found a gene of interest when we knew its function or protein associated to it. Today, we will analyze the gene in depth.

Suppose we have a library of genome (human-Hemoglobin β/yeast ArgA). For hemoglobin, we knew the associated protein. In particular, we used antibodies against the protein to find the gene. When we have an entire sequence of genome there are other ways to find gene. In particular, we can run it through a computer. On the other hand, we did complementation test in minimal media to find yeast ArgA. It required molecular biology.

Today, we will go in another direction. In particular, we will link protein and function with gene in our triangle 4.3. We have to use recombinant to read a DNA sequence.

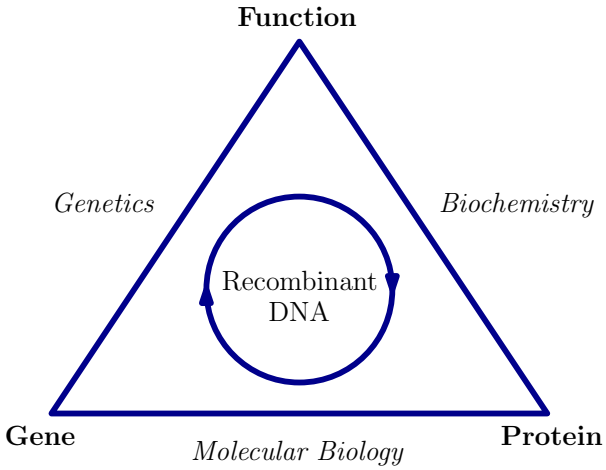


Figure 4.3: Triangle

4.5.1 Initial Analysis: Measuring Fragment Size

In this section, we will measure the size of the gene that we found. Recall that we have bacteria colonies that can grow in minimal media. Each bacterium has a plasmid. Using biochemical techniques, we can get plasmid out of the bacteria to get a test tube with plasmids and bacteria. The chemical peculiarities arising from double helical structure of plasmid will allow us to separate it from bacteria using a centrifuge.

Lemma 4.5.1. *In an electric field, a larger charged particle moves slower than a smaller particle with same charge*

Proof. This is proved in an electromagnetism class. □

Each plasmid could have different insert. The first clue to find the difference is length of insert. First, we use EcoR1 at restriction sites of plasmids to cut insert out of vector. Then, we use *gel electrophoresis* and Lemma 4.5.1 to separate molecules based on their length. The scheme of electrophoresis (see Figure 4.4) is as follows:

- Pour agarose into a little tray. Agarose will form a cross link, so a molecule has to sneak its way out of agarose.
- Pipette the inserts into a depression. We can dye (ethidium bromide) inserts, so that we can follow them.
- Charge the bottom of a tray positively. DNA are negatively charged (phosphate groups). Therefore, our inserts migrate downwards. Note that Lemma 4.5.1 implies larger fragments will move slower than smaller fragments.

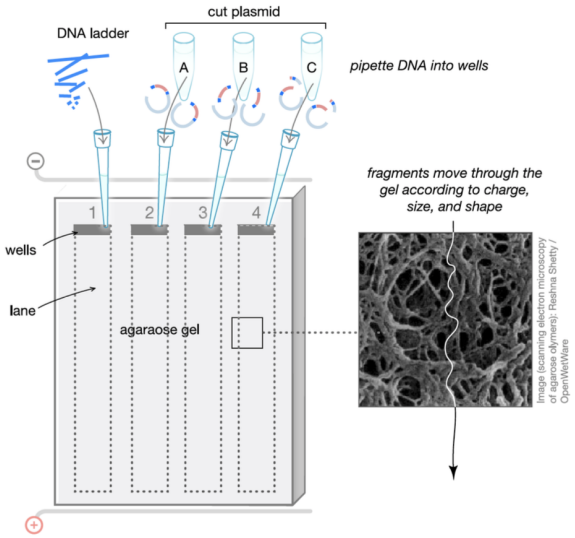


Figure 4.4: Electrophoresis

Definition 4.5.2. *DNA ladder* is molecular weight standards in gel electrophoresis.

There are fragments of known length that we can get in the catalog and see how they move in the gel. Comparing our inserts with molecular standards, we can find the size of our DNA.

Suppose we want to study more about our insert of size say 4.7kb. We can use enzymes to get restriction sites and cut our insert into fragments and measure their length. Based on the length we can figure out where those restriction sites were. This process is called *restriction site mapping.*

4.5.2 DNA Sequencing: Basic Idea

Now that we have purified our DNA based on their size, we just need to read it out. Nowadays, there are companies that will sequence DNA for us. In this section, we will learn the fundamentals of DNA sequencing.

Proposition 4.5.3. *There exists a way to read out DNA.*

Proof. The ideas goes back to Fredrick Sanger. Suppose that we know the beginning sequence of our DNA. At least for the plasmid, we can use catalog.

- First, heat the DNA to separate it into two strands.
- Add a primer that binds to the known sequence. (We can synthesize it in a synthesizing machine.)
- Add DNA polymerase and nucleotides (dATP, dCTP, dTTP and dGTP). Then, we can read what's being added.

□

However, the DNA polymerase fills in the nucleotides so fast that we can't see what it adds. Fred Sanger came up with an idea to add *defective* version of nucleotides (say T^*). By defective, we mean that the chain can't be extended. Suppose the ratio between T^* and T is 99 : 1. There is one percent chance that the polymerization stops at each T. By measuring the length of pieces formed in the polymerization, we can figure out the position of A in our DNA. We can do similar analysis with A^*, G^* and C^* in different lanes.

Remark 4.5.4. The polymerization reaction of nucleotide continues from 5' to 3' end as each nucleotide is added at 3' hydroxyl group. If we take off the hydroxyl group at 3' end the polymerization stops. Therefore, 2'3' di-deoxynucleotide is called a *defective nucleotide.*

Example 4.5.5. Suppose the template strands to be read out have a sequence 3'GATACTGGACGA5' and their corresponding primer is 5'CTA3'. We add polymerase, T^* and T (the latter two at a 99:1 ratio). Then we get the following possible fragments: 5'CTATGACCTGCT*3', 5'CTATGACCT*3', and 5'CTAT*3'. Now measuring the length of these pieces (using gel electrophoresis), we can figure out where A's are located in our DNA.

4.5.3 Radioactive Labelling

So far, we threw in primer, polymerase, dNTP and some dNT^*P in the gel. It turns out that agarose is not very good for single base resolution. Instead, we use polyacrylamide. Prehistoric people in the 1980s used radioactive nucleotides in polyacrylamide. To visualize it, they would

- take the gel containing glass into a dark room.
- Pry off the glass.
- Wrap it up in Saran Wrap.
- Put a piece of X-ray film.
- Put in the freezer.

After couple days, when they developed the X-ray film, they would see radioactive bands and sequence the DNA with about 500bp.

Fun Fact 4.5.6. Fredrick Sanger won a Nobel prize for his idea to sequence DNA.

4.5.4 DNA Sequencing: Fluorescent Sequencing

Radioactive labelling is painfully slow as we can read one type of nucleotide at a time. Instead, we can use fluorescent labelling in the defective nucleotides.

An advantage is that we can label different nucleotides with different colors (say A^* is red, T^* is yellow, G^* is green and C^* is blue). It allows us to carry our analysis in a single lane. We need a laser detector to see the colors going by.

Using this process, we can read around 700-1000 letters at a time. To read 100 thousands letter, we just need 100 samples with about 1000 letters. If we have 10 machines, we could do million letters. We will come back and talk about techniques that sped up this process dramatically.

Question 4.5.7. How do we read a DNA sequence that is 4000bp long?

- Shear it into four fragments.
- Read them out.
- Use a computer to reassemble it.

This process is called *shotgun sequencing.*

4.5.5 Cloning Revisited

For the last part of this module, we will see how we can do cloning faster.

Question 4.5.8. Suppose we know a sequence for hemoglobin gene. How do we sequence the hemoglobin gene in 500 other people?

A long round about is to make a cDNA library and carry out the analysis described in the previous sections. However, there is a way to read it without making a cDNA library. In some sense, it is cloning without cloning. The only challenge is that there might be small variation in the gene in different person but not so much that we can't recognize it.

Consider the known hemoglobin gene:

- Melt them apart.
- Choose a primer on either side of the regions we want to sequence.
- Add polymerase and nucleotides. The polymerization starts at each primer. And we will get two double helices from a double helix.
- Repeat.

Definition 4.5.9. The aforementioned process is called *polymerase chain reaction* (PCR).

We can take any DNA and carry out a PCR reaction.

Remark 4.5.10. This process has an exponential growth of 2^n where n is the number of iteration done.

Note that heating up might inactivate polymerase. However, there are *extremophiles* (bacteria like *Thermos aquaticus* that live in hot springs or volcano) whose (Taq) polymerase stays active in high temperature. Now we can just heat-cool, heat-cool and heat-cool.

Question 4.5.11. How do we sequence an entire genome? What do we learn from it?

Stay tuned for the next part of this class: Genomics.

Module 5

Genomics

5.1 October 23

In this module, we will explore genomics, a topic of great interest to Prof. Lander. We will cover the motivation for Human Genome Project and what goes on in the project nowadays.

So far, we have made the diagram 5.1 functional by interlinking function, genes and protein. At least, we can find gene if we know the function or protein. Now, the last intellectual unit of this discussion is to look at a big picture of the triangle. We will shortly discuss people got interested in genomics.

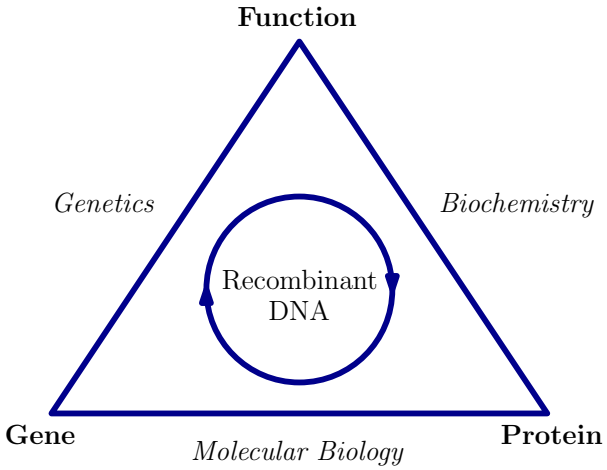


Figure 5.1: Triangle

5.1.1 Review of Recombinant DNA

Earlier in recombinant DNA, we saw how to:

- **Make a library:**

  - Cut DNA into pieces.
  - Clone the pieces into vector.
  - Transform vector into cells.
  - Plate them out and select the cells that acquired vectors based on resistance marker.

- **Find a gene of interest:** Depending on prior information about gene we can do:
  - Complementaion test based on phenotype.
  - Antibody test if we know the protein.

- **Sequence the gene:**
  - Take DNA fragments.
  - Put primer, polymerase, and nucleotides some of which are defective.
  - Measure the length of molecules of different length.
  - Infer the position of bases based on the length.

Once we know a sequence, we can clone faster using PCR primers.

5.1.2 Finding Our Gene: Based on a Human Disease Revisited

But we can't find the gene in which mutation give rise to Huntington disease because we don't know protein that causes the disease. In fact, it is challenging to study a degenerated brain to look for particular protein. We also don't know which cell to look at. It brings up an ethical question to infect a healthy person to study this disease. So we can't do complementation test. Further, it is a dominant disease.

Remark 5.1.1. We can also ask the question of finding gene for cystic fibrosis.

Locating Genes Based on Recombinational mapping

Alfred Strutevent had a solution to the problem posed earlier. Recall that he "located" genes of flies (eye color, wings and body shape) based on recombinant frequency. In fact, people made a map of genes based on visible phenotype (markers). We can do similar analysis in human.

Recall that we have have 23 pairs of chromosomes. First, we find on which chromosome the gene for Huntington disease is located. In theory, we cross infected people with a healthy person and calculate the recombinant frequency between genes for Huntington disease and visible markers. If RF between visible marker and the Huntington disease is 50% then the genes are not linked. And if they are linked, it means that the gene for Huntington disease and our visible marker are in the same chromosome.

But we will encounter the following problems:

- Ethical concern.

- Statistical issues (human have few children).
- Lack of visible markers (we don't have curly wings, and blue eye is a multigene trait).

5.1.3 Positional Cloning in Humans: Genetic Mapping

Instead, it was an MIT professor David Botstein's (yeast geneticist) idea to

- Study pedigree of existing families with Huntington disease. Remember that we can infer the genotypes at a disease level from a big pedigree.
- Look for genetic markers and compute the recombinant frequency between the marker and Huntington disease.

Regarding markers, Botstein had found that DNA in yeast are *polymorphic* (a particular gene in two yeast have different nucleotides in some positions). It turns out that human genes are also polymorphic.[1]

Definition 5.1.2. A DNA sequence variation in single nucleotide in the genome of pairs of chromosomes or across individuals of same species is called single nucleotide polymorphism (SNPs). See Figure 5.2.

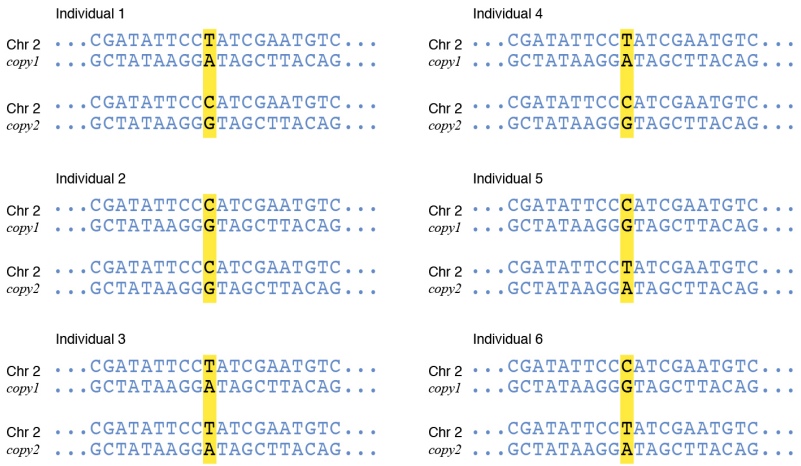


Figure 5.2: SNPs in genome of chromosome 2 (Chr 2) in different individuals. National Human Genome Research Institute

In fact, 1/1000 letter in our genome is hetereozygous. Since our genome has three billion sites, we have three million markers. In 1983, Researchers in MGH, inspired by Botstein, chose SNPs as markers arbitrarily and computed RF to no vail until their twelfth attempt. It turns out that gene for Huntington disease is on chromosome 4.

In 1985, a similar analysis was carried out to hunt for gene associated to cystic fibrosis. It turned out that the gene for fibrosis is located on chromosome 7.

[1] Polymorphism occurs naturally because of mutation and natural selection.

5.1.4 From Gene Mapping to Gene Discovery

Now that we have located the gene in a specific chromosome, let's discuss how we are actually going to find the gene. In this section we will consider cystic fibrosis.

With a lot of effort, people found markers that had a recombinant frequency of 1%. However, in human, 1% RF means that the genes are one million base-pair apart. It took five years of work to go from initial linkage to finding cystic fibrosis gene.

Feel people's pain. People would

- Make a piece of DNA into a radioactive probe.
- Wash it over a library.
- See where it sticks. The DNA would stick to an overlapping piece of DNA. That's how they would find next piece of DNA.

This process is called *chromosome walking.* Each step would take weeks. Using this technique, they found SNPs closer to the gene associated to fibrosis. And after five years, they figured out the cystic fibrosis gene. In fact, the triple TTT that enconde Phenylalanine at 508th position was missing in a lot of patients. Thus the genetic name for cystic fibrosis became $\Delta F508$ (Δ for deletion and F for the F sound in Phe[2]).

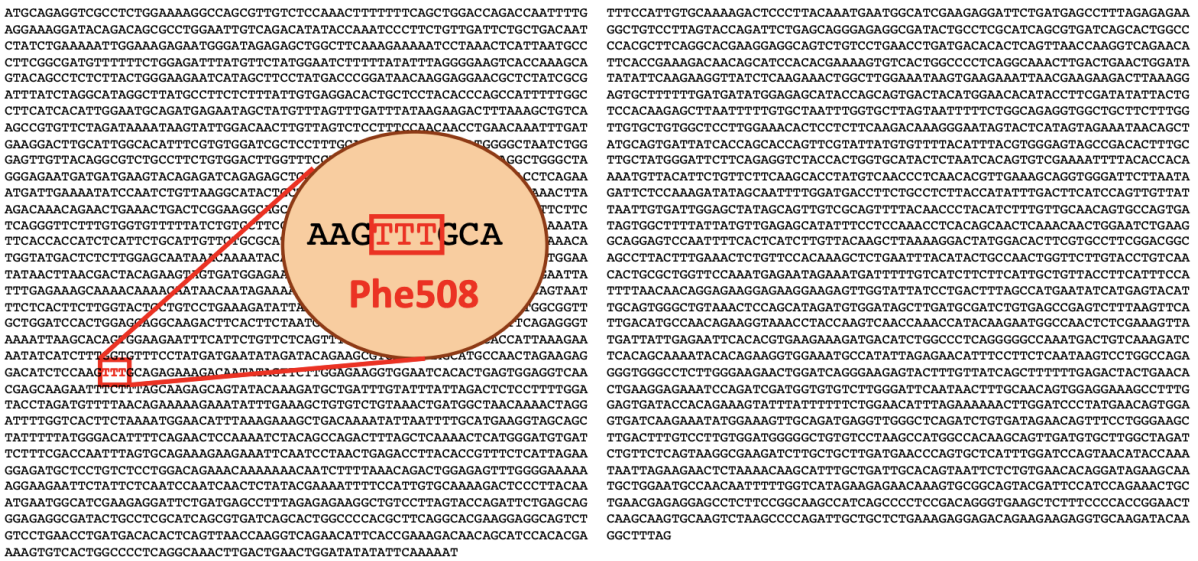


Figure 5.3: TTT is missing in cystic fibrosis patient

Now that we know the genetic basis for cystic fibrosis, we can run genetic diagnostics to see if a person is infected. In fact, we can

- Carry out polymerization chain reaction (PCR) of our DNA.
- Sequence it.

[2] A bit of linguistics here.

- Look if TTT at 508 position is missing.

In addition, we can also look for mutations in other places. In fact, we can take the whole sequence, translate it into amino acid and analyze it in the computer. People with cystic fibrosis will have similar sequence. From a functional point of view, they all encode ion transport which turns out to be chloride transporter. This ion transporter is the basis for cystic fibrosis. There is even a molecular model for the disease.

5.1.5 Human Genome Project: Goals

The analysis in the previous section gave rise to one observation: we can find genes but it is lot of work hoping around to test random markers until they show linkage. We needed to have

- Genetic map: Genetic landmarks to trace inheritance.
- Physical map: We don't want to do chromosome walking to figure out the DNA fragments covering the chromosomes.
- Sequence: DNA sequence 3 billion bases at a time
- Gene list: Identify all genes.

It was a great idea address the issues. But it would take two centuries at the rate at which everything was done. Human Genome Project which was started in October 1990 made it possible in about thirteen years. By April 25, 2003, the project was completed (99.3%).

Fun Fact 5.1.3. The team planned to complete by this date because it was half a century after Watson and Crick published their double helical model of DNA.

5.1.6 Human Genome Project

By now, it is clear that to make a genetic map, we collect random markers and map them in big families (40). People wrote computer codes to map all the markers. In fact, one of the first things that Prof. Lander did in biology was to create a genetic map.

To make a physical map, instead of starting at a point in genome and walking across, a large number of pieces (100,000) of DNA and small pieces (100 bp from PCR) were tested against each other. By testing, we can figure out which points on small fragments were contained in which markers. Therefore, we can reconstruct from a huge number of PCR and bunch of large clones how they overlapped with each other.

Finally, to sequencing all DNA we need to:

- Take the overlapping 100,000 bp fragments that have been cloned
- Break each fragments into tiny fragments (2000 bp)

- Sequence the DNA fragments from both ends
- Ask the computer to reassamble the 100,000 bp from lots of these fragments. This is done even today.

At the end, the project was successful in reading out essentially a whole human genome. There were about 300 gaps including centromeres but for practical use it meant reading out everything.

Fun Fact 5.1.4. The project was an international collaboration (16 labs around the world). It took 13 years and three billion dollars to sequence one human genome.

Fun Fact 5.1.5. The first person whose genome was sequenced was called RPI11 (Rochester Polytechnic Institute). 30 people signed up but their information about identity was destroyed.

5.1.7 Improvements since Human Genome Project

Question 5.1.6. What about genotyping all of the genes?

In theory, we do PCR of each SNPs site and read them out. Consider individual 1 in Figure 5.2.

- Take capillary tubes that have DNA fragments with same sequence to that of the individual except at SNPs.
- Wash the individual's (fluorescent labelled) fragments of copy1 and copy 2 through each capillary tubes.

The tube sends fluorescent signals when fragments stick while washing. Suppose copy1 sticks well in a capillary tube in which DNA fragments at SNPs have AT. Then we can infer that copy1 has TA. Similarly, we can pass copy 2 through capillaries.

In practice, we use *photolithography* to carry out the process we outlined. It turns out that we can use multiple capillaries and sequence million SNPs at a time.

Definition 5.1.7. *A reversible defective nucleotide* is a nucleotide that can reversibly terminate the PCR.

Question 5.1.8. What about sequencing?

Initially, the rate of sequencing was about 1 million letter base per day per machine. Nowadays, we can use a chip containing multiple capillaries where we carry out PCR of fragments but with fluorescent labelled reversible defective nucleotides. Suppose that the next base that is supposed to be added is G. The capillary sends out colored signals when a reversible defective nucleotide G^* is added. We can just read the signals to read the sequence.

As of today, the rate is three thousand billion base pairs per day per chip. At Broad Institute, every nine minutes a whole sequence is read out. it takes $600 to sequence a whole human genome.

Moral: Everything that seems impossible today becomes possible soon. Things that seems possible but hard become easy. And they become so cheap that it is suitable for high school experiments.

5.2 October 26

Last time, we set up a ground for why we wanted to sequence genome. Today, Prof. Lander will give a world wind tour of the human genome. In particular, we delve into human genome and see what we can infer from it.

5.2.1 Contents of the Human Genome: Coding Regions

In a rough draft of the Human Genome Project published in Feb 2001 in *Nature*, we could already see 90% of the human genome sequence. By April 2003, we we 99.99% information.

Before the project, people thought that our genome had hundred thousand genes with about 3×10^9 base pairs. It turned out others were pseudo genes, and broken pieces. There are about 21 thousand protein coding gene constituting 1.3% of the genome. These genes come in families:

- Protein kinase: 500 genes.
- Lipid kinase: 20 genes.
- Phosphatase:[3] 200 genes.
- Olfactory receptors: 400 genes.[4]
- Many others: 100 genes, 4 genes, 1 etc.

We have lots of copies of genes because either locally extra copies are made due to inaccurate recombination or genes get broken and there are erroneous phenomena of *duplication* and *divergence.* There is no point for originality of genes but for functionality of an organism. The extra copies are free to mutate.

5.2.2 Contents of the Human Genome: Transposons

Definition 5.2.1. *Transposons* are DNA segments that are able to copy themselves to different parts of the genome.

[3]These take off phosphates that kinases put on.

[4]Mice have more olfactory receptor genes. They rely on smell. In human, a larger fraction is non-functional.

Remember we talked about retrovirus: it can make copy of its RNA and transpose into human genome. In fact, we got transposons when our ancestors got infected by all sorts of transposable elements. Transposon consist of more than half of our human genome and come in different flavor:

- RNA transposons: They transpose through an RNA intermediate.

	Size	# of copies	% of Genome
LINEs	6000bp	100,000	$> 21\%$
SINEs	330bp	1000000	$> 10\%$
Endogenous retroviruses	2000bp	100000	>7-8%

Table 5.1: RNA Transposons

- DNA transposons: They transpose through a DNA intermediate. There are 380,000 each with a size of 300bp. They constitute about 3-50% of our genome. As of now, we can no longer recognize them because of mutation.

Question 5.2.2. How do LINEs work?

- Each of the full length LINE element encodes a transcript, that encodes protein like any other gene.
- The protein binds to the RNA that made it.
- Then it reverse transcribes that RNA.
- Makes cDNA and puts it into a new site.

That's how we are able to get thousands of copies of genes in genomes.

Question 5.2.3. How do SINEs work?

If a LINE element is a parasite (copying itself around the genome) the SINE element is parasite on the parasite. It makes a much smaller message that has same little sequence at the end that can be recognized by the LINE elements to reverse transcribe it and put is somewhere.

Sometimes, transposable elements evolve into functional genes. In addition, when an element copies itself into RNA and it gets reverse transcribed into genome, with some probability, it picks up extra DNA. Usually, the picked up DNA is useless but not always. Suppose the sequence has a regulatory sequence that turns on genes in liver cells. When the transposon lands into some other parts of the body, the sites are regulated by liver cells. This is how new regulatory elements evolve.

5.2.3 Evolutionary Comparison across Species

The transposons end up covering genomes. And every element that goes into our genome will stay there and pile up mutation. We can infer how old the transposons are by

looking at the number of mutations they have undergone. In addition, we can compare transposons in different species. In fact, we can form a clock to track evolution.

For instance, when we sequenced the genome of human, mouse, cat and dog we found 10 identical base pairs in all of them. We can infer that they might have same ancestors and some genes got preserved[5] (an example of what is known as *evolutionary conservation.*) In fact, we can track human evolution and infer when creatures moved away from each other.

```
Human   GCCTGGCCGAAAATCTCTCCCGCGCGCCTGACCTTGGGTTGCCCCAGCCA
Mouse   ----------AAGCCTGTGGCGCGC-CGTGACCTTGGGCTGCCCCAGGCG
Rat     ----------AAGTTTCT---CTGC-CCTGACCTTGGGTTGCCCCAGGCG
Dog     GGCTGC----AGACCTGCCCTGAGGGAATGACCTTGGGCGGCCGCAGCGG
                  *    *       *    **********  *** ***
```

Figure 5.4: Evolutionary conservation

5.2.4 Evolutionary Conservation: Patterns in Coding vs Non Coding Regions

We can take a chunk of DNA across of all of the organism (us, dog, and mice) and ask what kind of mutation tend to occur.

We find that protein coding region are highly conserved, with mutation (change in base pairs) and some *gaps* (deletion of nucleotides) that result in loosing function.

In the intergenic region, we see highly conserved non coding elements (regulatory elements), gaps, and frameshifts.

The graph below shows the amount of sequence conservation in horizontal axis: Red is the amount of conservation in transposoons, blue is that of the whole genome. There are more conservations in transposoons in the background.

For instance, Satb1 is a single gene involved in early development. We might imagine that they have to be regulated correctly. In fact, genes involved in early development have a huge number of conserved non coding elements around them that are controlling their regulation.

Evolutionary conservation allowed us to infer that regulatory elements have huge number of conserved elements. In fact, there are probably about 3 million regulatory sequences in the genome, but we don't know what they do. This is a task for a next generation of biologists.

Fun Fact 5.2.4. Functional genes are conserved while non-coding regulatory elements tend to evolve faster. The coding genes across the mammals are more or less the same

[5]However, there is a chance that same environmental condition (and need for similar functionality) might have given rise to same genes.

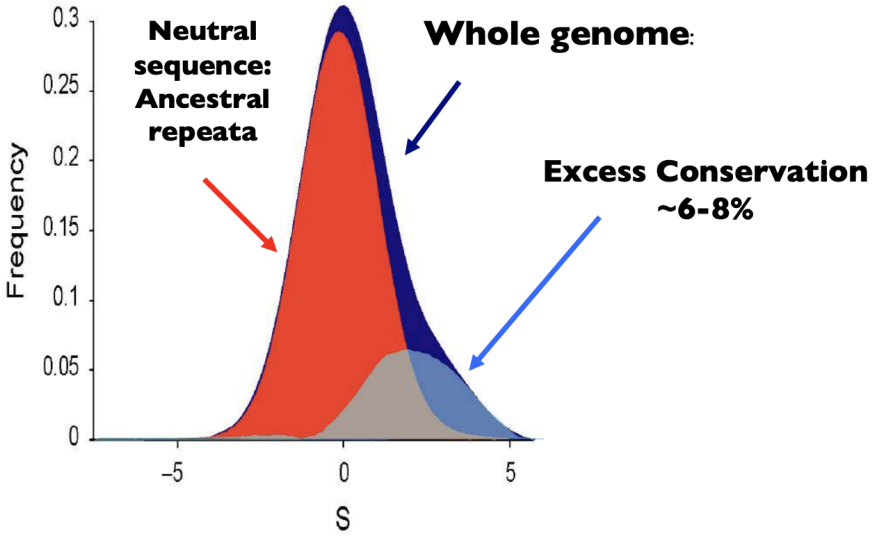


Figure 5.5: Evolutionary Conservation of Non Coding Elements

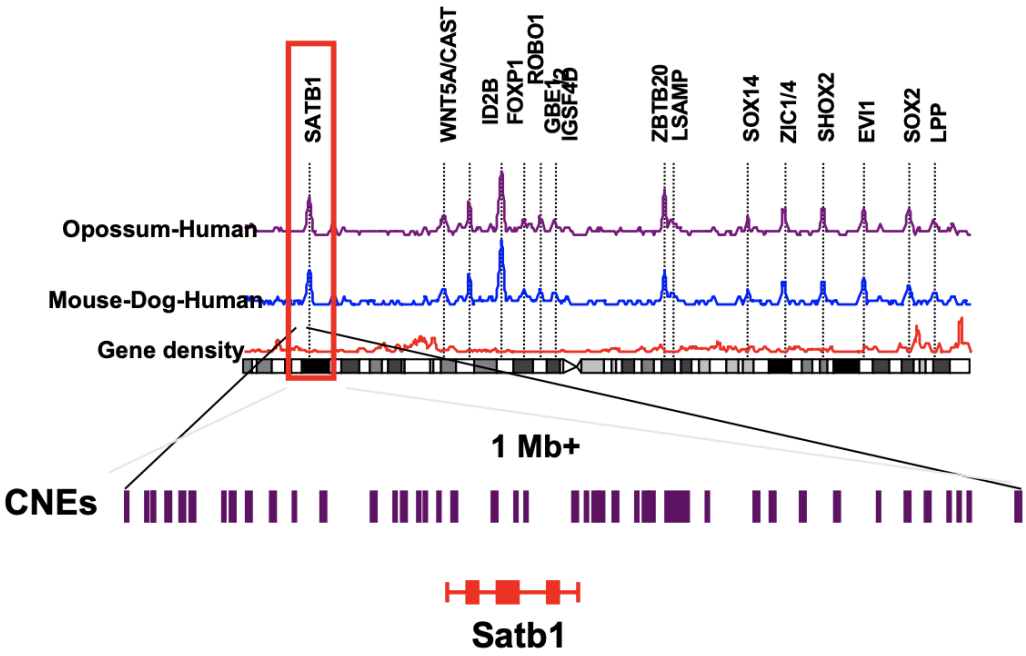


Figure 5.6: Evolutionary Conservation

but they differ in terms of regulatory elements. For instance, they all have bones and brains but the difference is in how long they let their bone develop and how their brain folds.

So far, we have found that in our genome there are:

- 21,000 protein coding genes in many families (due to duplication and divergence with exons covering 1.3% of the genome). This is largely similar across mammals

- Some RNA encoding genes: tRNA, and rRNA.
- Conserved regulatory regions (millions), covering 6% of the genome; evolve faster than protein coding regions.
- Transposons, covering greater than 50% of the genome; mostly selfish parasites but not entirely.

5.2.5 Comparison among Human Populations

Our history is written in our genome. When we trace back our history, we find that we (*Homo sapiens*) come from East Africa. People have argued using bones that our ancestors left Africa around 50-100 thousand years ago. Indeed, we can use DNA mutations to reconstruct the family tree. We can look at the genomes found in mitochondrial/nuclear sequence and see how they are related.

Using human genome, we can also infer that we diverged from Neanderthals about 500,000 years ago. We might have bred with them or killed them. However, there are segments of DNA in Neanderthals that are present in us. So, we might have mated with Neanderthals.

5.2.6 DNA Variation

Question 5.2.5. What can we learn from DNA variation if we take human population?

Recall that we can trace inheritance of Mendelian diseases using SNPs. But it works only for single gene trait. In fact, most of the diseases like diabetes, alzheimer's, heat disease, schizophrenia etc are *polygenic.*

However, we now have a tool to carry out gene mapping. In particular, we can hunt for SNPs associated to a disease. For instance, in the case of diabetes, we compare the genome of thousands of infected people and healthy people and see which SNPs are significantly common in infected people but not in healthy. Suppose, we find that at a SNP of people with diabetes (but not in healthy people) have GC in copy 1 of chromosome and TA in copy 2. It means that peple with GC/TA at that SNP have higher risk of having diabetes (12%) while people with TA/GC have lower risk of diabetes (10%).

This works for Schizophrenia (which is characterized by delusions). When researchers collected 82,000 cases, they found 245 SNPs (non-coding regions) associated to the disease. In fact, people have found more than 100,000 such associations between SNPs polymorphism and different common diseases. In additoin, we can also map height, traits of red blood cells and eye color. We just need to collect tons of data.

Moral: If we know the sequence of a genome, we can get a lot of information by looking at the big picture: transposons, evolutionary comparison, conservation.

5.3 October 28

This is the last class on genomics. Today, we will look at the whole genome from a bird' eye and also talk about perturbing the genome.

By now, in the Triangle 5.7, we can go from function to gene (complementation), gene to protein (DNA sequence encode protein) and protein to gene (use antibodies). We just need to find a way from gene to function to complete our triangle. Today, we will study genome-wide RNA expression and complete the triangle.

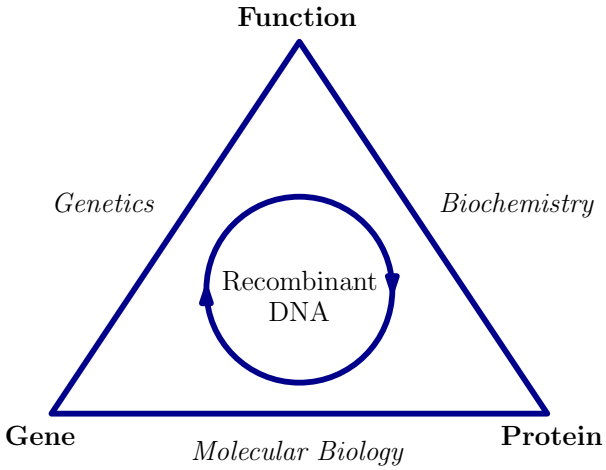


Figure 5.7: Code of Arms

5.3.1 Genes Differ in RNA Expression across Cells and Circumstances

Last time, we ended our class by describing DNA variation. In this section, we will talk about RNA variation. Genes are expressed differently across many different tissues or cell types.

Definition 5.3.1. *House keeping gene* are expressed in most of the cells. We need it to run a cell. For example, gene in mitochondria and DNA polymerase.

Definition 5.3.2. Genes that are expressed only in certain cells are called *specialized genes.* For example, genes expressed in Neurons (ion channels) and genes moderately expressed in macro-phage.

Question 5.3.3. What can we learn from expression pattern of all 21,000 genes in a tissue?

We can study diseases.

Proposition 5.3.4. *There are two types of Leukemia.*

Sketch of the proof: When Sidney Farber worked on childhood cancer, leukemia, he noticed that that there were different kinds of leukemia that needed different treatment.

After 40 years of staring at microscopes tying to spot differences in enzymes, proteins, cell surface markers, he found the differences between **acute myelocytic leukemia** (AML) and **acute lymphocytic leukemia** (ALL). However, we can prove this without spending 40 years.

Proof. For now, assume that we know how to see to what extent genes are expressed in a patient. Given that we can collect the gene expression in thousands of patients and look for coherence like we did for SNPs. Suppose we get an unlabelled data, Figure 5.8, that has information on how all the 21000 genes are expressed in the patients. We can consider each row (corresponding to one patient) to be a point in 21000 dimensional space (corresponding to the number of genes). If there are n flavors of leukemia, the points form n clusters in the space.[6] It turns out that the data form two clusters, see Figure 5.9. Therefore, there are two flavors of leukemia.

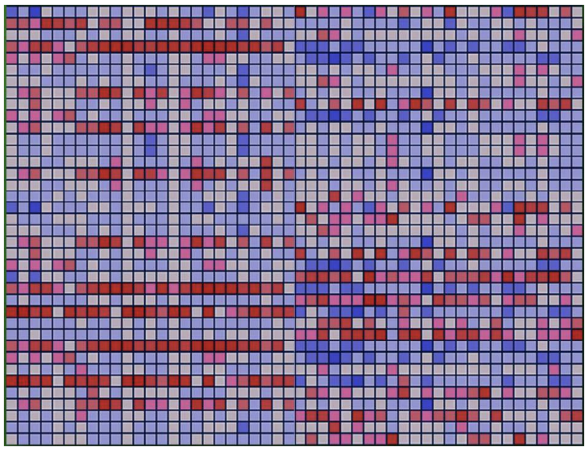


Figure 5.8: Unlabelled data. Color represents how well the gene is expressed (red is high and blue is low)

.

Now let's address the question of how we actually measure the gene expression:

- Form a *micro array* with little patches of oligonucleotides corresponding to all 21000 genes.
- Take RNA from tumor.
- Label it with fluorescent.
- Wash it over the micro array and "see" where it sticks.
- Using laser, measure (and note) the gene expression levels based on stickiness: x_{11} $x_{12} \dots x_{1n}$ for $1, 2, \dots n$ genes for patient1 and in general $x_{n1}, x_{n2} \dots$ for patient$n$.

□

Corollary 5.3.5. *We can diagnose whether a person has AML or ALL.*

[6] This analysis is called cluster analysis and done using computers in practice.

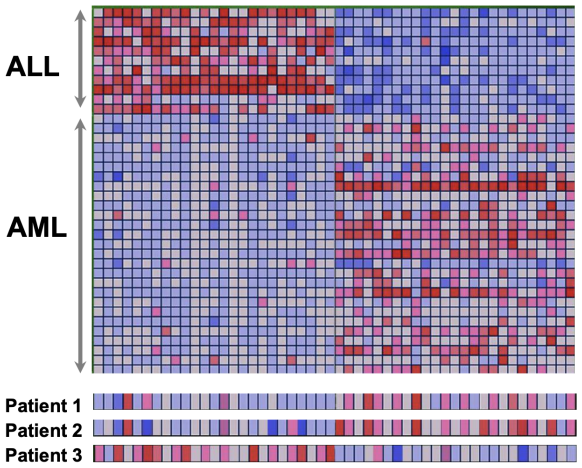


Figure 5.9: Gene Expression in Leukemia Patients: Patients along vertical direction and the gene numbers in horizontal direction

Proof. Measure the gene expression of the person and compare it with the labelled database (see Figure 5.9). □

5.3.2 Human Cell Atlas: Single Cell RNA Sequencing

The analysis we did in the previous section is not limited to tumor. In fact, we can look at gene expression in each individual cell and classify all human cells. Within the last ten years, researchers at MIT have made progress. In this section, we will briefly mention how the methods work.

There are beads. Each bead has a large number of copies of given piece of DNA (molecule). And it consists of:

- Constant region.
- Bar code for bead: every molecule in the bead have same sequence but it it unique to each molecule.
- Random sequence (unique molecular identifier).
- Poly-T (TTTT).

Proposition 5.3.6. *We can read cell types.*

Proof. The proof to read out cells is by construction of a scheme:

- Little bubbles are formed in a device where beads are flowing in one way and cells are flowing in other.
- Each cell breaks open and (poly-A sites of) all the mRNAs attach to one bead at poly T site.

- Transcribe cDNA coming off all of these mRNA. Each beads contains all the cDNA that encodes the information of mRNAs and the barcode in the beads encodes which cell the mRNAs are from.

- Throw the beads into a parallel DNA sequencer and read out RNAs.

Now that we have constructed how to sequence mRNAs of a cells, it is clear that every cell is a point in 21,000 dimensional space. If we want to find out the cell types, we can look at how many clusters these data form. □

Fun Fact 5.3.7. Using the construction in Proposition 5.3.6, people studied retina for 40 years and thought that they figured out every types of cell in it. It turns out there were more than double the number of them.

Fun Fact 5.3.8. There is a project called Human Cell Atlas that is working to classify cells in human body.[7] It involves groups in more than 70 countries around the world.

5.3.3 Completing the Triangle

In the previous section, we saw a global picture of gene expression. Here, we resolve the issue of going from gene to function and complete the triangle.

Proposition 5.3.9. *If we know a gene, we can infer its function in an organism.*

Sketch of the proof: We have to edit gene and observe a loss/gain of function in an organism. Before giving a detailed proof, we will explain how to edit genes in the following section and the statement follows easily.

5.3.4 Modifying the Genome: Adding Genes (Transgenic Mice)

We saw how to mess with bacterial DNA by throwing in plasmids. In this section, we will see how to add genes in multicellular organisms (in particular mice). We can inject DNA with a needle into a newly fertilized egg of mice and implant it into pseudo pregnant mouse. Note that where the DNA is inserted is random and the number of copies of DNA is uncontrolled. Therefore, this process is not very optimal.

5.3.5 Modifying the Genome: Subtracting Genes in ES Cells

We can also grow up a mouse with a knocked out gene but we prove two things:

Proposition 5.3.10. *1. There exists a process with a high precision to knock a gene out of 3 billion bases of DNA.*

[7] Classification problems are the interesting ones in math: classification of semisimple Lie groups and finite simple groups are some of them.

2. There exists a mouse that can grow with a messed up gene.

Proof. Let's first prove the second statement by constructing a process in which a mouse can grow given that we can knock out a gene.

Note that, in a mouse embryo, there are cells at certain stages called blastocysts (also called *inner cell mass*). We can culture these cells to form *embryonic stem (ES) cells.*

It is easy to see that ES cells are *pluripotent.*[8] In other words, if we

- Grow the ES cells (of say black mouse) for a while
- Inject the grown mass into a new blastocyst (say of albino mouse)
- Implant them into uterus of pseudo pregnant mouse,

the ES cells will grow into a mouse. Note that the grown mouse will have cells from black mouse as well as from albino mouse. Therefore, it will be *chimeric.*

However, we have no guarantee that the black cells that are injected into mouse are part of the cells that will produce sperm and eggs. To confirm that some of the chimeric mice acquire ES cells that can produce sperm and eggs, we can breed chimeric offspring with an albino mice and see if we get black and albino offspring. When we do experiment, we get black and albino offspring. Note that the black mouse will have gene that comes from the black ES cells that we started with. Therefore, we can grow a mouse with a knocked out gene.

Now let's go back to the first statement. To disrupt a gene, we can

- Take a piece of DNA that is same as the gene of interest but is interrupted by something that can break the gene and can be positively selected (like in antibiotic resistance markers).
- Just with bacteria, transform some of these DNA.
- Hope that by genetic recombination, the cell will insert our DNA in just the right place to recombine and replace the cell's own copy of gene.

Despite the odds of recombination is low (1 in 300), people did it for years and were lucky enough to find cells that acquired the disrupted gene. This completes the proof of the proposition. □

However, there are better ways to prove the first statement. In particular, it was Mario Capecchi (Nobel laureate) idea to not just select for the recombination based on *positive marker* but use *negative markers.* Note that the disrupted DNA that a cell acquires can land anywhere in the cell. To ensure that the only cells that come out are the ones that have DNA in the right place, we can attach negative markers. Therefore, cells that picked up DNA randomly but are not of interest won't undergo recombination. Then we can grow up a mouse with a knocked out gene in the right place.

[8] This is an example of proof by experiment.

Given a mouse where we have knocked out the genes of interest, we can see a partial proof of Proposition 5.3.9. In particular, we can study the function that the mouse loses. When we experiment, very often the mouse will die which means that the genes are essential for life.[9]

5.3.6 Modifying the Genome:: Subtracting Genes in Specific Tissues

In the previous section, we knocked out genes in all cells. Instead of knocking out in all cells, we can knock out in specific cells (say in liver).

There is a certain sequence called *lox P site* that is used by a bacterial virus. The virus comes in as a piece of liner DNA but when it enters the cell an enzyme called *Cre recombinase*[10] causes a recombination between these sites that give rise to circular DNA.

If we had made a mouse with ES cells without destroying the gene but put lox P site around it, the gene would still be functioning. But, in addition to lox p site, if we have inserted Cre gene attached to cell (liver) specific promoter, then in liver cells Cre enzymes is made. Now Cre enzymes circularize lox P sites and make the gene non-functional. This gives us tissue specific knock outs. Given a tissue specific knock out, we can study what a gene does in specific tissue (liver). We can also study where the gene product goes. We can do it by adding green fluorescent protein (GPP)[11] to our gene we can track the products.

5.3.7 Modifying the Genome: Knocking in Genes in ES cells

We can also study where the gene product is being made. We attach gene to make a protein fusion with a green fluorescent protein (GPP). GPP is found in jelly fish (this is why they glow green). And then we can see where protein glows green.

[9] An amateur biologist got interested in figuring out where the ears of dragonflies were. On a warm July day, he went near a pond and caught a dragonfly. He tied a rope in its tail gently so that the dragonfly would not run away. Then, the biologist started his experiment:

"Fly, little dragonfly," instructed the biologist and it flew. He concluded that dragonflies had ears but did not know where.

"Fly, little dragonfly." This time he took off its one leg but it was able to fly. So, he concluded that the ears were not in that leg.

He did the same with other legs but the dragonfly flew when he asked it to. After so much effort, he got exhausted but curious at the same time to figure out where the ears actually were.

"Fly, little dragonfly." This time he took off its wings. Sadly, the dragonfly could not fly.

"Eureka! Dragonflies have ears in their wings."

[10] The name was a molecular joke for cream cheese and lox bagel

[11] Jellyfish have GPP.

5.3.8 Modifying the Genome: Knocking Down RNA

Previously, we messed around with gene. In this section, we will do the same with RNA. Note that RNAs have short life (they are short term messages). Therefore, we don't have to wait for many generation to study the gain/loss of function because of RNA knock out.

Definition 5.3.11. *Antisense RNA* is an RNA that comes and blocks the RNA being translated.

People used antisense RNA to degrade other RNA but this process was not very optimal. A better idea sprung out of an accident.

In worms *C. elegans,* people tried injecting antisense RNA against a gene called *twitcher gene* (that makes the worm twitch). As a control, they injected the sense version of RNA. And for no reason, they threw both sense and antisense RNAs at a same time. The sense and antisense RNAs formed double stranded RNAs. They had a significant effect on knocking down the RNA of interest leading to mutant phenotype.

It turns out *C. elegans* have adefense mechanism against double stranded (RNA) viruses called *RNA inhbition (RNAi).* The cell asks certain protein to go around and destroy matching copies of double stranded RNAs.

RNAi and lox P site opened a possibility of looking into natural systems that can be used to modify genes. CRISPR is one of them.

5.3.9 Modifying the Genome: 2G-CRISPR

In this section, we will overview what CRISPR does.

Fun Fact 5.3.12. CRIPSR was recognized by Nobel prize a few weeks ago.

Recall that a *restriction enzyme* is a protein that recognizes a certain (palindromic) sequence say $GAATTC$ and cuts that piece. But a restriction enzyme knows only one sequence, so we need thousands of them if we want to knock out thousands of gene at a time.

Instead we can imagine a programmable restriction enzyme that can instructed by giving instruction that cuts any DNA that matches it (just like RNAi cuts RNA).

Remember that restriction enzymes are defense mechanism each one designed to cut particular sequence. In contrast, a programmable restriction enzyme does not have to decide in advance what it is going to attack. Instead, we can have a library of sequences that gets turned into RNA and become instructions for our programmable restriction enzyme. This way we can get one restriction enzyme with dozens of instructions.

Lemma 5.3.13. *There exists a programmable restriction enzyme in nature.*

Proof. They exist in bacteria.[12] □

We will give a basis for why they exist. In the past, when some virus infected bacteria, bacteria would chop up the virus but would keep it as information (also called *memories of past genetic aggression*). The information were passed on to a programmable restriction enzyme in bacteria so that they could protect themselves from future infections.

With some effort, we can adapt these programmable restriction enzymes from bacteria to any eukaryotic cells including human cells. If we put in a protein called *Cas9* and give instructions in the form of RNA, the enzyme will search for DNA that matches the instruction and cut. When it cuts the DNA, either the cell will chew back and glue the hole or we can hand it a recombination template and insert a new sequence.

CRISPR is just a logical extension of programmable restriction enzyme.

Remember how we made mice before. It takes two years to grow a mouse. With CRISPR we can do it in couple weeks. In fact, we put a protein and a guide RNA to knock out genes. This process works even in living cells with high efficiency that we don't have to select manually.

Indeed, we can create a virus that carries in the instruction for the Cas9 protein and the guide RNA and edit our own cells. We can use gene editing to repair genetic diseases (an example of *gene therapy*). For instance, we can insert TTT in genes of patient with cystic fibrosis.

Proof of Propositon 5.3.9. We can add/delete gene and study the gain/loss of function in an organism (tissues). This completes our Triangle 5.10. □

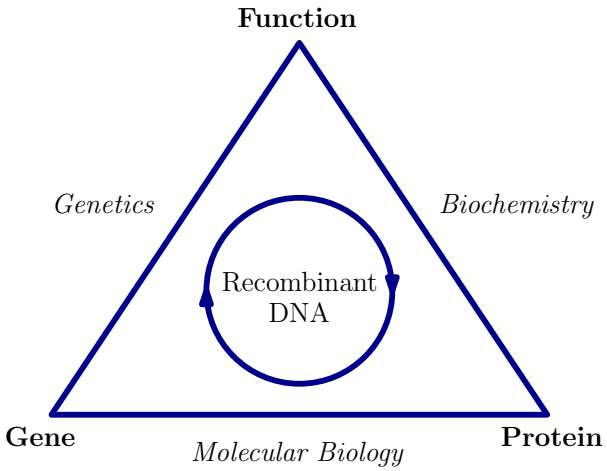


Figure 5.10: Triangle

[12]This is an example of proof by example.

Module 6

Microbes and Immunology

6.1 October 30

"World's Most Wanted." Prof. Drennan is looking for microbes.

In first half of this module, we will study microbes. And in the second half we will talk about defense mechanism against microbes. Today, we will start with viruses and later we will examine bacteria.

6.1.1 Viruses

Definition 6.1.1. *Viruses* are small infectious agents that replicate inside a host cell but not on their own.

They are so small that there are 10 million *viral particles* in one drop of sea water. In fact, in an ordering by size: lipid < protein < virus < bacterial cell < human cell < ant. However, viruses can be dangerous: HIV (770,000 people died of AIDS-related diseases in 2018), Spanish flu (50 million people died) and Covid-19 (as of this morning 45 million cases and 1.2 million death world wide).

6.1.2 Components of a Viral Particle

We have to understand viruses better if we want to come up with a treatment. A viral particle has

- Protein coat.
- (Often) lipid envelope around the protein coat.
- Protein spikes on the lipid envelope. The spikes help virus to attach to a host.

- Viral genome inside the protein coat.
- (Sometimes) encapsulated proteins within protein coat.

Most of the information is stored in viral genome.

- Genomes can be DNA or RNA.
- They are very small (around 2000 bases).
- They encode as few as 2-4 proteins.
- They are overlapping: two or more proteins are encoded by the same nucleotide sequence.

6.1.3 Types of Hosts

Recall that viruses need hosts to replicate. But they can infect all types of cell: bacterial, animal, and plants.

Definition 6.1.2. Viruses that infect many hosts are called *generalists*. For instance, influenza virus. In contrast, *specialists* attack only few hosts For instance, HIV (it attacks immune cells in animal cells).

It is uncommon for viruses to switching from one kind of host to other kind of host. If they switch, it can be serious in a population that lacks immunity. For example, bird flu (initially in birds) and ebola (started with bats) affected human seriously.

On the other hand, switching within the same population can results in the spread of viruses. If the first infected patient dies before passing it to other, there is no spread. But when it does, there types of spread are:

- **Outbreak:** Local spread.
- **Epidemic:** Wide spread (but not global) occurrence of disease.
- **Pandemic:** Global spread.

Fun Fact 6.1.3. "There is an MIT professor who predicted the pandemic. What did she know?" After the outbreak of Covid-19, FBI interrogated Prof. Drennan at her door step.

Last year, when Prof. Drennan was giving this lecture, she stood in front of the classroom and said that the conditions were right in the world for another pandemic. She had no specific information but the knowledge of biology and history.

6.1.4 How Do Viruses Replicate?

In this section, we will study the life cycle of a virus that infects bacteria (a.k.a *bacteriophage* or phage).

Definition 6.1.4. *Horizontal gene transfer* is the transfer of genetic material between organisms whereas *lateral gene transfer* is from parent to offspring.

There are two cycles in a phage:

- **Lytic cycle:** Viral genome is replicated by host to make more bacteriophage, killing the host. The infection process includes:
  - Phage genome replication.
  - Phage protein synthesis.
  - Phage particle assembly.
  - Phage particle release (*cell lysis*).

  It takes about 30 minutes.
- **Lysogenic cycle:** Unlike in lytic cycle, there is a horizontal gene transfer between viral genome combines and its host genome (source of new genes). The viral genome can stay in host for long time, increasing the diversity of hosts' genome which can be good sometimes. For instance, a gene that affords antibiotic resistance is good acquired from virus is good for bacteria.

  Often under stressful conditions, the viral genes excise from the host genome and undergo lytic cycle.

In both cycles, the process starts when phage attaches to bacterial cell and injects phage DNA.

Now let's focus just on lytic cycle.

Question 6.1.5. Which/when are proteins are made from viral DNA?

Phage DNA have

- Promoter (P).
- Operator (O).
- Transcription start site.
- Genes associated to P and O ("early" genes).
- Another set of P, O, and transcription site.
- Genes associated to second P and O ("late" genes).

Recall that transcription of gene involves making mRNA which is translated to proteins. In the case of virus, early proteins form from early genes. One of the early proteins combine to second operator activating the late genes. The late genes will then transcribe late proteins.

Poll 6.1.6. Guess the function of early proteins.

1. Viral genome replication

2. Repress host gene expression
3. Lyse bacterial cells
4. 1 and 2
5. All of the above

Answer: 1 and 2. We don't want to lyse the cell before it is ready. The virus would not have time to assemble.

In addition, the functions of early proteins include

- Cleave host DNA to make viral parts.
- Activate late genes.

In contrast, late gene proteins

- Make viral coat proteins.
- Lyse bacterial cells.

6.1.5 Types of Viruses

So far, we have talked about virus (with DNA) that infect bacterial cell. But in general they are other types of viruses:

- **RNA virus:** They can also be categorized further on the basis of
  - Genomic replication cycle: RNA $\Longleftrightarrow$ RNA.[1]
  - Virion content:
    - (−) RNA: Ebola and Influenza.
    - ds (double stranded) RNA: Reoviruses and rotaviruses.
    - + RNA: Hepatitis C and SARS (SARS-COV-2).
- **Reverse transcribing viruses:**
  - Genomic replication cycle: RNA $\Longleftrightarrow$ DNA.
  - Virion Content:
    - +RNA: HIV and other retroviruses.
    - ds DNA: Hepatitis viruses.
- **DNA virus:**
  - Genomic replication: DNA $\Longleftrightarrow$ DNA.
  - Virion content:
    - ss (single stranded) DNA: Parvoviruses.

[1] Here, $\Longleftrightarrow$ sign means that RNA is transcribed to RNA and vice versa during the replication cycle.

    * dsDNA: Herpes viruses.

Note that RNA is worse genetic material than DNA because of its instability. Moreover, RNA polymerases do not proofread like DNA polymerases do which can lead to higher mutation. Viruses "survive" by adapting in new environment by evolving. Fast evolution of viruses makes the design of antivirals challenging.

+ Sense and - Sense RNA Viruses

- + sense RNA is ready for translation. Therefore, it does not require an RNA-dependent RNA polymerase encapsulated in the virus particle.
- In contrast, -sense RNA must be converted to +sense RNA before translation. It needs an encapsulated RNA-dependent RNA polymerase.[2]

Comparison of + sense RNA virus with a + sense RNA retrovirus

Recall that corona virus is + sense RNA virus and HIV virus is a + sense RNA retrovirus. In this section, we will see how they get into the cell and how they replicate. This is important if we want to prevent them from getting into our cell and if we want to create vaccines.

Regarding the first point, the process by which viruses enter cells are:

- + sense RNA virus: *Endocytosis*. The virus eats up the cell membrane and goes in.
- +sense RNA retrovirus: *Membrane Fusion*. The virus particle fuses in cell membrane and gets in.

As for the second point

- + sense RNA virus makes -sense RNA to make more +sense RNA.
- Meanwhile, +sense RNA retrovirus carries an encapsulated reverse transcriptase in its virus particle to convert its RNA genome into cDNA. Then the cDNA is inserted into the genome of a host cell. The integrated DNA gets transcribed making more + sense RNA for the new viral particles.

6.2 November 2

"You can't *B. cereus*."

Lately, we have been talking about microbes and viruses. Today, we will learn about bacteria. Then we will talk about treatments against both virus and bacteria.

[2] X dependent Y polymerase means that X is being read and Y is being made. Recall that when DNA is translated to mRNA it needs DNA dependent RNA polymerase.

6.2.1 Bacteria

Bacteria are prokaryotes. There are about 10^{30} bacterial cells on earth (for comparison there are 10^{31} viral particles). The convention to write names of bacteria is *Genus species.* For instance, *Escherichia coli*. Often, genus is abbreviated as in *E. coli*.

Definition 6.2.1. *Strains* are genetic variant or sub-type of species. For instance, the strains of *E. coli* are $O157 : H7$ (food borne pathogen that causes disease) and $K - 12$ (common lab strain and is harmless).

Poll 6.2.2. Most species are not characterized yet: estimate is that we have characterized % of all bacteria and have % left to characterize.

- 0%
- 5%
- 15%
- 25%
- 40%

Answer: 0%

Fun Fact 6.2.3. The full extent of genetic diversity of bacteria is unknown. Most species aren't characterized yet. An open problem for next generation of biologists.

Bacteria live everywhere: in our gut, lettuce, acidic hot springs, and radioactive waste. They can live in hosts or freely (different from viruses). Moreover, they can live with oxygen (aerobically) and even without oxygen (anaerobically).

6.2.2 Bacteria Friend or Foe?

Based on the relation between host and bacteria, there are different types of bacteria:

- **Symbiotic bacteria:** Host and bacteria mutually benefited.
- **Commensal bacteria** gets nutrients from host but host is unharmed.
- **Parasitic** bacteria get nutrients form host and host is harmed.

Definition 6.2.4. A *parasite* is a microorganism that lives at host's expense and decreases fitness of hosts.

Definition 6.2.5. A *pathogen* is a microorganism that can cause disease.

However, most bacteria are not parasites and many parasites are not bacteria.

6.2.3 What Do Bacterial Parasites Do That Is so Bad?

- Bacteria produce *toxins* (poisons) that can be peptide-based or are compounds made by enzymes. For instance, *Bacillus cereus* is a food-borne pathogen that produces a peptide based toxin called *cereulide* that destroys mitochondria and causes extreme distress (*food poisoning.*)

  Moral: Bacteria multiply faster in room temperature, so we should refrigerated leftovers.

- When bacteria colonize areas like stomach lining and internal tissues, they cause infection. *Helicobacter pylori* colonizes stomach lining leading to gastric ulcers. However, bacteria living in skin, oral and nasal cavities, gastrointestinal and urogenital region are not infectious.

Fun Fact 6.2.6. People initially thought that ulcers were caused by stress. Barry Marshal injected *H. pylori* in his body and infected himself with ulcer to prove that *H. pylori* causes the disease. For this discovery, Barry Marshall and Robin Warren got Nobel prize in 2005.

6.2.4 Not Always Born to Be Bad

Definition 6.2.7. *Virulence* is a parasite's ability to damage a host.

We can measure "badness" of bacteria with its virulence. Note that parasites are not virulent all the time. For instance, *C. difficile* is often found in the gut of a healthy person but can also cause inflammation in colon.

Definition 6.2.8. *Virulence factors* are proteins or small molecules made by enzymes that increase the destructive nature of a pathogen.

An active area of research concerns the identification of virulence factors and how to inhibit them. We don't have to kill microbes to remain healthy. We just need to control their destructive ability.

6.2.5 Getting Infected

There are various modes of transmission for viral and bacterial infections. It is important to know the modes of transmission to protect ourselves. Note that asymptomatic people can be infectious.

Fun Fact 6.2.9. In 1800s, a Mary Mallon was asymptotic to typhoid but transferred *Salmonella typhi* to a lot of people. People did not know how it was transmitted.

The ways in which we can get infected by microbes are:

- **Direct contacts**
  - Person to person (skin to skin and sexually transmitted)
  - Droplet Spread (sneezing and coughing).
- **Indirect Contacts**
  - Airborne (transmission over longer distances than 1 meter). Covid-19 is airborne diseases.
  - Vehicles (contaminated objects, food, and drinking water).
  - Vectors[3] (insect bites and ticks).

6.2.6 Fighting Infections: Comparing Viral and Bacterial Strategies

Now that we have studied about microbes and how we get infected, we can study how to protect ourselves from them with preventive measure and treatments:

- **Immune Response:**
  - Natural Response: Our body has immune system that can fight against bacteria and viruses. We will talk about it in detail in the second part of this module.
  - Vaccine-elicited immune response: We expose a person to attenuated microbes and when we get exposed to them the next time, our immune system can fight against them.
    * Virus: Vaccines are major way of fighting against viruses. We use live-attenuated virus or killed viral proteins to make vaccines. Examples of success includes Polio (99-100% effective 3 doses), and measles (98% effective 2 doses; 90% of unvaccinated will get measles if exposed). For HIV, it is still an active area of research), and so for SARS-CoV-2.[4]
    * Bacteria: However, it is a minor way of tackling bacteria. Partly because it requires boosters in addition to live-attenuated bacteria to make a vaccine. Typhoid (first report 1896) vaccines are good only for 2-3 years. Tuberculosis vaccines are 50% effective (as of 2019) but still saves millions of lives.

    Remark 6.2.10. Vaccines prevent disease. They DON'T treat or cure disease.

- **Drug Therapies:** These are ways of treating infections with drug compounds:
  - Antivirals: These are secondary methods battle viral infections.
    * Antibody therapies (we will discuss it in detail later in the course).
    * Use of small molecules that target viral encoded enzymes. Recall that virus have enzymes in their protein coat like RNA dependent RNA polymerase and reverse transcriptase.

[3] This is different than the vector in Definition 4.3.2.

[4] Update on Dec 2020: We have vaccines that are around 94% effective.

An example of an antiviral pro drug (no phosphates) is Aciclovir, which targets some viral polymerase (pro-drugs are turned into active agents inside cell).

– Antibiotics: These are primary methods for combating bacterial infections.
  * Antibody therapies.
  * Small molecules target things that are different between human and bacteria: ribosomes and cell wall.

  An example of antibiotic is penicillin, which targets cell wall biosynthesis.

Fun Fact 6.2.11. When Dorothy Hodgkin confirmed the structure of penicillin (proposed by Edward Abraham), see Figure 6.1, it looked so weird that one of her colleague promised to quit his job, go to farm, and grow mushrooms if it was correct. She won a Nobel for it and he never quit his job.

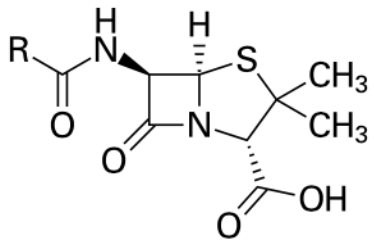


Figure 6.1: Structure of Penicillin

6.2.7 Antibiotic Resistance

Although we have treatments against bacteria, we need to be aware that using it recklessly can cause problems. Remember that antibiotics work by either killing bacteria or stopping them from reproducing. However, because of natural selection bacteria that develop resistance to antibiotic (or even multiple drugs) to survive. Then they pass along their antibiotic resistance genes to the next generation.

Definition 6.2.12. *Antibiotic resistance* is a resistance built up in a bacteria against an antibiotic whereas *multi-drug resistance* is against multiple drugs.

In fact, antibiotics are blunt instruments because they kill both good and bad bacteria which can be harmful:

- **Scenario 1:** Antibiotics to treat an infection of *Y. ellow* bacteria kills it and commensal bacteria. However, it does not kill the antibiotic resistant *C. difficile,* which lives asymtomatically in gut. There will be more nutrients available for *C. difficile*, so they grow rapidly causing infections in colon. We can eat yogurt with probiotics (a less scary name for live bacteria) while we are on antibiotic so that there is not enough nutrient for *C. difficile.*

- **Scenario 2:** Commensal bacteria can develop resistance to the antibiotic. Although they won't do any harm on their own, when there is a *horizontal gene transfer* between them and pathogenic bacteria, it can be troublesome.

6.3 November 4

"I am 10% human and 90% bacteria."

Last time, we started a discussion on antibiotic resistance. Today, we will complete it and talk about Human Microbiome Project that plans to study microbes found in our body.

6.3.1 Factors That Leads to the Development of Antibiotic Resistance

- When antibiotics (not necessarily the same) keep attacking the same processes (or target) like cell wall biosynthesis and the protein synthesis at the ribosome, bacteria come up with a defense mechanism for that process (target).
- If we keep using the same antibiotics due to lack of new ones,[5] bacteria can recognize the antibiotics and fight against it.
- If we use antibiotics frequently for treating human and animals disease, bacteria have more opportunities to develop resistance.

6.3.2 On a Molecular Level, How Does Antibiotic Resistance Develop?

- The first mechanism is inactivation of a drug through drug modification. For instance,
  - Mutation can alter an enzyme so that it phosphorylates a ribosome targeting drug, inactivating the drug.
  - Hydrolysis of a drug by an enzyme hydrolase.
- Bacteria can change the target inhibiting the ability of drug to inhibit its target. For instance,
  - Often times, mutations alter an enzyme target so that the drug no longer binds.
  - Mutations alter a methylase so that it can now methylate rRNA altering the ribosome so that the antibiotic cannot bind to ribosome.
- Finally, bacteria can change their cell wall and/or membrane that it is hard for a drug to get in/stay. For example,

[5] Unfortunately, there has been a "discovery void" for new antibiotics. It is expensive to bring drugs to markets (because of capitalism and bureaucracy).

Around MIT, there are tons of startups pharmaceutical companies. In Starbuck near MIT, Prof. Drennan meets people who share their pain. Once a company came up with a drug that worked but they were going bankrupt. It turns out that clinical trials need to up the dose and need to do the trial again until they fail.

  - Mutation in cell wall proteins can lead to impermeability.
  - Mutations can alter membrane protein pumps so that drug is actively pumped out the cell. This increases efflux of the drug. In fact, the increase in efflux might lead to multi-drug resistance.

;

Poll 6.3.1. Identify the mechanism of antibiotic resistance acquiring a gene that encodes a methylase that methylates small molecules changing the small molecule's shape and binding properties

1. drug modification
2. target modification
3. cell membrane/cell wall protein modification

Answer: 1

6.3.3 How Does Antibiotic Resistance Spread?

Fighting against bad bacteria involves figuring out how antibiotic resistance spread. It also reminds us to be careful in misusing antibiotics (recall how antibiotic resistant *C. difficle* infect an asymptotic person seriously).

Proposition 6.3.2. *There exists a mechanism antibiotic resistant bacteria can use to spread the resistance.*

Proof. We will prove the statement by constructing the mechanisms. Suppose that we have antibiotic resistant (AR) bacteria.

- **Mechanism 1:** They will multiply and infect new hosts. For instance,
  - Bacteria in pigs that are given antibiotics prophylactically develop AR bacterial strain.
  - That strain ends up in pig feces.
  - Feces go into ground water.
  - The ground water used to water lettuce.
  - Other animals (including us) eat lettuce and get infected.
- **Mechanism 2:** There can be a horizontal gene transfer between AR bacteria and another bacteria of different species/strain. Often times, bacteriophages help in the transfer. Recall that in lysogenic cycle, we have a phage and a host DNA.
  - The phage DNA is combined with host DNA.
  - Under stress condition, the integrated phage DNA can undergo exicision taking some pieces of host DNA to start a lytic cycle.
  - When this phage infects other bacteria, it can transfer resistance genes.

□

6.3.4 How Big a Problem Is Antibiotic Resistance?

Spread of antibiotics spreading combined with reckless use of antibiotics is a big problem. Bacteria have developed resistance to almost every class of antibiotics (like vancomycin). In fact some are displaying multi-drug resistance. It is estimated that there will be 10 million deaths per year due to antibiotic resistance if no major changes/advances are made by 2050.[6] The number of death is equivalent to an international plane crash every 15 minutes. And it was a pre-covid estimates.

To tackle the problem, we need to

- Go after new antibiotic targets.
- Create more specific antibiotics that target pathogenic bacteria and not all bacteria.
- Consider methods that lower virulence rather than kill virulent bacteria. For that, we need to understand more about how bacteria that live within us which brings us to the last topic of this class: Human Microbiome Project.

6.3.5 NIH Human Microbiome Project

In 2004, Human Microbiome Project was launched with a vision to:

- Learn about bacteria that cannot be easily *cultured.*[7] Approximately, 99% of bacterial species cannot be easily cultured in a laboratory because we don't know what they survive on and what their metabolism mechanism looks like. Therefore, we have to study communities instead.
- Probe the microbial communities including bacteria, viruses, archaea, protists and fungi that live within us.

To study the whole community, the plan was to

- Isolate total DNA from microbes living in particular regions of the human body (259 people were used as test subjects).
- Carry out whole genome sequencing for over 1000 microbes that are isolated from each body part.
- Develop new tools for DNA sequence analysis.

Definition 6.3.3. *Metagenomics* is the study of genes from whole communities as opposed to a single organism.

Definition 6.3.4. *Mircobiome* is a combined genetic material of all microorganisms in a particular environment.

Definition 6.3.5. *Microbiota* is a community of microogranisms.

[6] Nepalese including me have no chance to make these advances.

[7] To culture bacteria means to grow single species of bacteria under defined conditions in lab.

Some Initial Results of HMP

- Bacterial cells outnumbered human cells by 10:1 ratio.[8]
- Based on the extrapolation of sample data, there are 5000 times more microbes than there are people on Earth as of today.
- The microbiome of everyone in the sample was unique.

  Remark 6.3.6. Babies in wombs are not associated with microbes. Meanwhile, newborns get their first microbiome from their mother during birth. It changes with age.

- There were more bacterial genes than human genes (23,000 genes in human compared to 1,000,000+ genes in our microbiome.) We still don't know a lot about it.
- ∼80% genes in the HMP metagenomes could not be assigned a specific function.
- ∼50% of genes in the HMP meta-genomes could not be given any *annotation.*

Analyzing the HMP Data-Annotating Genes

As mentioned previously, HMP consists of sequencing human microbiome. Part of analyzing the sequenced data is to know the function of the microbiome.

Definition 6.3.7. *Gene annotation* is a process of assigning a function to that gene in a public database.

Note that an annotation can be *generic* (eg. methytransferase can methylate any molecule) or *specific* eg. vitamin-B_{12} dependent homocysteine methyltransferase or they can be completely wrong.

Gene annotation can be made by:

- Biochemical characterization of the encoded proteins. This process is ideal but has low throughput.
- By identifying sequence homology between the encoded protein and a protein of known function using computer programs like BLAST.

When we compare our protein against other protein sequence in a database using BLAST, we get something like in Table 6.1.

In the table, identity indicates the percent of identical residues between our sequence and the database sequences. In addition, the closer the E value to zero, the more significant the match.

[8]This ratio might turn out to be overestimation. There are other papers which say that the ratio could be 1.3:1. We should take the data as a pinch of salt for now.

Description	E value	Identity
Unknown protein	0	100%
Biotin Synthase	7e-107	55%

Table 6.1: Toy result of gene annotation using BLAST

In case there is no match, we can look at genes next to our gene of unknown function. Recall that bacterial genes are often in operons. The function of gene next door might provide a clue.

6.3.6 Future

In 2016, HMP was "completed." However, from previous sections, it is clear that we need to annotate a lot of genes and study our microbiome better. Some possible projects in the future would:

- Use HMP data to design more specific antibiotics and/or alter the microbiome to remove harmful microogranisms. and/or lower their virulence. For instance,
  - Get rid of *C. difficile* to prevent recurrent infections.
  - Manipulating human gut microbiome could be a viable therapeutic strategy for reccurring *C. difficile* infection.
- Investigate the connections between human microbiome and human health using HMP data:
  - "Does the human microbiome influence the brain and behavior?" (NY Times article by Carl Zimmer Jan 28, 2019)
  - Understand the role of gut microbiome in colon cancer to see if we can manipulate it to prevent cancer.

6.4 November 9

Now that we have seen our villians, it is time for us to study about our defense mechanism. Prof. Ayce will be with us for the next two classes on immunology. We will talk about immune system of vertebrates.

6.4.1 Immune System

Definition 6.4.1. An *immune system* is a vast network of cells/tissues that defend body against invaders like bacteria, viruses and toxins. The goal is to keep out, destroy and/or neutralize these invaders (*antigens*) using soluble proteins called *antibodies.*

Our body has two types of immune system:

- **Innate immunity:** It is an inborn, immediate and fixed immunity. The response of innate system is static. The components of innate immunity are:
  - **Skin and mucous membranes:** Skin works as physical barrier. Its low pH, dryness, and presence of commensal bacteria prevent harmful bacteria from invading. On the other hand, mucous membranes have *lysozyme* that keeps invaders out. Lysozyme is an enzyme that is present in our bodily secretion that breaks down bacterial cell wall. It also has hydrophobic peptides that poke holes on membranes of invaders.
  - **Innate cells and the complement system:** Innate cells are underneath the skin. Innate cells are *phagocytic*. They
    * Have receptors that bind molecular motifs unique to bacteria and viruses.
    * Engulf invaders by breaking down their sugar, lipids, and protein.
    * Secrete *cytokines* to attract other immune cells. Cytokines are signaling molecules of the immune system.

    Meanwhile, complement systems poke holes on membrane cells that are covered by antibodies.

- **Adaptive immunity:** Only vertebrates have adaptive immunity which is unique each infection. If we get infected the second time, the immune system can recognize invaders. The response is delayed but targeted. The two responses are:
  - **Humoral immunity** is mediated by antibodies secreted by B cells. B cells mature in bone marrow.
  - **Cell mediated immunity** consists of cellular part of adapted immunity that are mediated by T cells. T cells mature in *thymes.*

Organism who don't have adaptive immunity are prone to many disease. In human, Severe Combined Immune Deficiency (SCID) is one of genetic defect that is a result of lack of adaptive immunity. David Vetter who had SCID lived in a plastic for six years. NASA engineers built him a special suit so that he could come out of bubble.

6.4.2 Hematopoeitic Stem Cells

Definition 6.4.2. The innate cells, B cells, and T cells come from one single cell type found in bone marrow called *hematopoietic stem cells* (hema means blood and poietic means making in Greek).

Remark 6.4.3. Babies make these cells in all of their bones. In contrast, adults use only some bones to form cells.

These cells can renew themselves. Some offspring will change their genetic profile and form all types of other cells (See Figure 6.2).

Remark 6.4.4. AML cancer occurs in common myeloid progenitor whereas ALL cancer in common lymphoid progenitor.

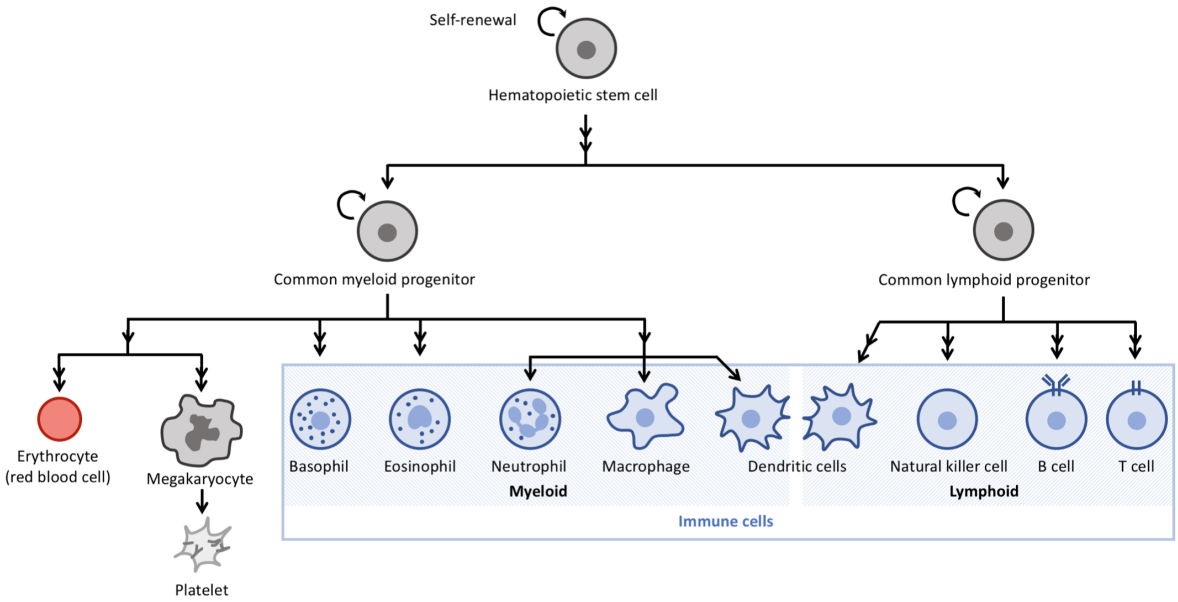


Figure 6.2: Hematopoietic stem cells

6.4.3 Battle against Antigens

In this section, we study the role of macrophages, dendritic cells, B cell (antibody), and T cell in our fight against antigens.

Definition 6.4.5. B cells with membrane bound receptors (antibodies) that have have not seen antigen yet *naive B cells*. An antigen can bind in the binding sites of B cell receptors.

Remark 6.4.6. Adaptive cell have a repertoire of binding cells. Every B cell has specific type of antibody, see Figure 6.3.

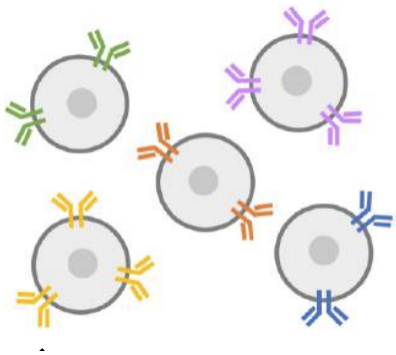


Figure 6.3: B Cells (each color represent different specificity of receptors)

Remark 6.4.7. T cell is also specific to the same antigen from the same bacterium. The specificity allows two factor authentication when we are deploying our troops of B cells and T cells.

Suppose bacteria attack us. These bacteria will have proteins or special sugars that a specific type of B cell receptors can recognize.

Definition 6.4.8. *Epitope* is the actual part of the protein of antigen that binds to B cells.

Our immune response is as follows:

- B cell binds to bacteria at epitope, see Figure 6.4.
- These B cells travel to lymphoid tissues.
- Binds to helper T cells (also called CD4+) specific to the antigen, see Figure 6.4.
- Costimulatory molecule binds to B cells activating the B cells.
- The activated B cell proliferates by undergoing *clonal expansion.* The offspring will be better at recognizing bacteria.
- They will also undergo *differentiation*, see Figure 6.6, to produce:
  - Plasma cells: They have lots of endoplasmic reticulum and secrete antibodies.
  - Memory cells: They are long lived and form a quick defense mechanism during the next exposure to the same antigen.

    Remark 6.4.9. Babies have passive immunity acquired through mom's antibodies that pass through their gut but don't have memory cells.
- Antibodies secreted by plasma cells cover bacteria making it nonfunctional.
- Macrophages have special receptors that recognizes antibodies and engulfs bacteria neutralized by antibodies.

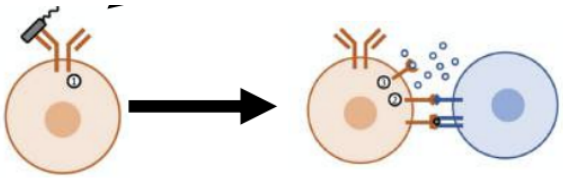


Figure 6.4: (Left) B cells attached to a bacterium travels to lymph node (Right) and attaches to T cells

.

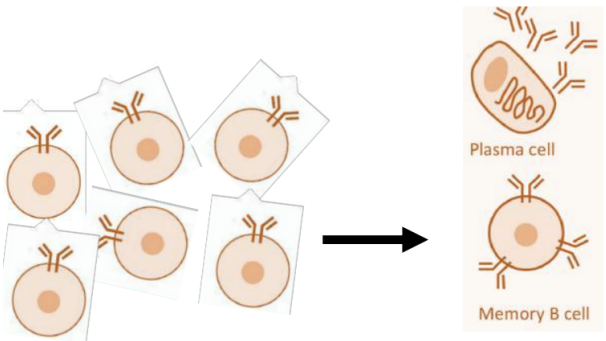


Figure 6.5: Differentiation

Remark 6.4.10. In our body, antibodies are produced at a rate of 2000 per seconds.

Remark 6.4.11. Lymph nodes closest to the infection are battle field except when the infection is all over the body. In the later case, lymph nodes all over the bodies are utilized. In this battle, some of the bacteria can also raise our temperature.

In short,

- Secreted antibodies bind and neutralize antigens.
- Bound antibodies mark target.
- Macrophages destroy the target.

Time of the Response

In the first exposure to antigen, there is lag in activation of antibodies (5-10 days). However, the response to the second exposure is faster (1-3 days), stronger and better (higher affinity) response because of memory cells. The bacteria will not have a chance to harm us the second time although there might be some infection.

Vaccines exploit secondary responses. They expose us to inactivated (non functional) pathogens via injection and when we get infected we will have our antibodies ready to fight the pathogens.

6.4.4 Antibodies Structure

There is no doubt that antibodies are our soldiers. They are also called immunoglobulin. Antibodies have (see Figure 6.6):

- **Two identical heavy chains and light chains.** Therefore, antibodies are also called *heterotetramer*
- **Variable regions** that can bind to two identical antigens. In other words, antibodies are *bivalent.*
- **Constant region** determines where in our body (mucosal membrane, blood etc) the antibody will go and which arm of the immune system to alert.

6.4.5 Antibodies in Research and Medicine

- Antibodies can be used for
  - Research: cloning and protein purification.
  - Diagnosis of diseases.
  - Therapeutic uses: biologicals (fancy way of saying medicine).
  - Forensics.

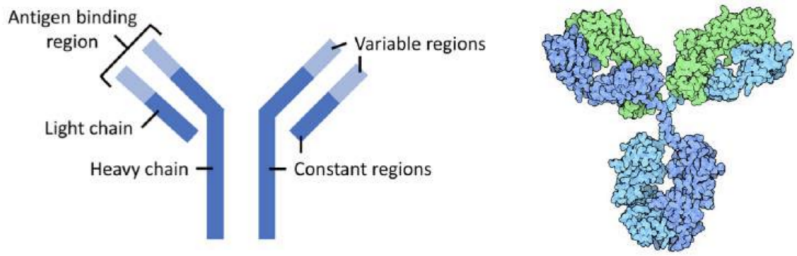


Figure 6.6: Structure of antibody

- We can immortalize and grow a B cell that makes antibody (momoclonal antibodies) by injecting it in animals.[9]
- We can also collect plasma of animal that has been infected and recovered (polyclonal antibodies).
- Given all the advantages, antibodies may be generated against body's own proteins (autoimmune disease). This is an active area of research.

Next time, we will try to answer the following questions:

Question 6.4.12. Why are there millions of B cells specific antibodies although we have twenty one thousand genes?

Question 6.4.13. What if the antigen hides in the cell rather than going to blood (where B cells can attack)?

6.5 November 13

Prof. Ayce Yesilaltay is back with her second lecture on immunology. Today, we will try to address two question that we posed at the end of the class.

6.5.1 VDJ Recombination

Earlier, we stated that antibodies (proteins) come in millions of flavor which is good for our immune system as we can recognize so many antigens. However, there are only about twenty one thousand genes. Recall that genes encode the information of proteins. Susumu Tonegawa (MIT) discovered a recombination event called *VDJ recombination* (variable, diversity and joining) that solves the riddle for which he got a Nobel prize in 1987.

Researchers found that variable region in antibodies of B cells genome are different that those in embryonic genome. In the latter, there are

[9] Using animal to produce antibodies might be unethical?

- Variable region: 20-50 in number.
- Diversity regions: 15-20.
- Joining regions: 6.
- Constant region: 1.

When a B cells develop from embryonic cells, unlike in other cells (say liver), there are

- **DNA rearrangement:** A B cell gets one of each region (variable, diversity, joining and constant).
- **Sloppy end joining:** Some nucleotides are inserted or deleted in different regions to get junctional diversity.
- **Mutation:** After the activation of B cells, an enzyme called *activation induced cytidine deaminase* can mutate B cells. Mutations can increase or decrease the affinity of B cells to binders.

Accounting the above events in light chains (VJ), heavy chains (VDJ), we can get millions of B cells with only twenty one thousands gene.

Proposition 6.5.1. *The affinity of B cells to bind to antigens increases with time.*

Proof. Recall that B cells bind to T cells in lymph node. When a bacteria attaches to B cell, it breaks down bacteria into peptides. Then B cells will show *mHC (major histocompatibility complex)* on their surface that T cells recognizes and binds to B cells. B cells that have high affinity to bacteria because of mutation will attach have broken down peptides and therefore mHC. Therefore, more T cells are likely to bind to them. When they undergo cloning the affinity is transferred. And after multiple iteration we will end up with high affinity binders. This is called *affinity maturation.* □

Remark 6.5.2. There are two classes of MHC:

- **MHC Class I:** Found in every cell.
- **MHC Class II:** Only antigen presenting cells (dendritic cells, macrophages, and B cells) have it.

6.5.2 T Cells

Note that everything we talked about our defense mechanism against invaders works when everything is free floating in blood or lymph. The complement system, innate cells, or B cells can tag the invaders so that macrophages can engulf them.

"You can't see me," say some bacteria hiding inside cells. This is when T cells happen.

There are there types of T cells:

- **Helper T cells:** The function of helper T cells are

  - Activate B cells
  - Secrete cytokines that helps in the proliferation of cytotoxic T cells and differentiation.

- **Cytotoxic T cells:** They kill infected cells.
- **Regulatory T cells:** They help to dampen our immune system. For instance, even though our gut bacteria are checked each day, the immune system is dampened so that commensal bacteria can stay in our guts.

T cells are similar to B cells because they have

- Unique T cell receptors each with one type of specificity.
- Variable region that antigen can bind.

6.5.3 Activation of T Cells

However, T cells don't directly bind an antigen. They recognize peptides that B cells present to them in the form of mHC (major histocompatibility complex) proteins. But they can only recognize certain kind of peptides. To understand the specificity, we need to look closer into how T cells are work:

- Dendritic cell (arbage collectors) engulf bacteria and break them into peptides.
- Dendritic cells show MHC to helper Tcell CD4+.
- T cells that can recognize the mHC binds to dendritic cells.
- Cytokine released from dentritic cells bind to T cells and activate them.
- T cells proliferate and differentiate to *effector* and *memory cells.*
- T cells go to lymph nodes and activate B cells.

6.5.4 Killer T Cells

The mechanism we described in previous section does not say what happens to infected body cells. There are *cytoxic* T cells (CD8+ a.k.a called killer T cells) that bind to the denditric cells and differentiate. But they can go around and find infected cells. Recall that there are *MHC class I* in every cell. Cytoxic T cells bind to MHC class I and see if it has bacterial peptide. If there is, killer cells stop the infection from spreading by releasing two compounds:

- One to poke holes in the infected cell.
- Proteases that will kill the infected body cells.

Remark 6.5.3. This mechanism works to kill cancer cell as well. But cancer cells can get out of control when they start to express proteins that are not recognized by T cells.

Fun Fact 6.5.4. MHC is the main reason for organ/tissue rejection. When MHC class I does not match, T cells can destroy the transplanted cells. Identical twins and siblings have higher chance of tissue matching.

6.5.5 T Cells in Medicine

- **Vaccines:** Recall that when are infected the second time, there is a quicker and better response from memory cells. Using that idea, we can expose live attenuated microbes to generate antibody response (smallpox vaccine and BCG for tuberculosis). It is because live microorganisms
  - Engage all arms of the immune system (innate cells and both types of adaptive cells)
  - Have more antigens
  - Spread more quickly in the host,

  therefore we build up a good immune system the first time we are exposed.

  In contrast, when we inject just the viral (bacterial) protein as a vaccine, antibodies, dendritic cells, and helper T cell will be activated. However, killer T cell won't be able to see those proteins. Therefore, the vaccine will generate less effective immune response.

- **Cancer Therapy:** For cancer therapy, killer T cells are targeted to kill cancer cells via CAR T (Chimeric antigen receptor T cell) against specific tumor protein markers (effective in skin cancer).

- **Self Tolerance:** T cells that recognize body's own proteins are eliminated during their maturation in thymus. For example, clonal deletion.

- **Organ rejection:** Mismatch of MHC results in organ rejection.

6.6 November 16

Prof. Lander is back with a timely topic today: Covid 19. Prof. Ayce will also join us to explain immune response to corona virus.

6.6.1 From an Atypical Pneumonia to a Pandemic

Timeline

- Dec 31: Chinese authorities treated dozens of cases of pneumonia of unknown cause.

- Jan 11: China reported its first death.
- Jan 30: WHO declared a global health emergency.
- Feb 2: The first coronavirus death was reported outside China.
- Feb 23: Italy saw major surge in cases.
- Feb 29: The United States reported a death.
- March 10: MIT asked undergraduates to move out by March 17.

As of today, 11.1 million people are infected and 246 thousand death are attributable to covid in the US. Meanwhile, around the world, more than 70 million people are infected while 1.5 million people have succumed to covid.

6.6.2 Structure of SARS-CoV-2 Virus

The virus has

- RNA genome: 30kb wrapped nucleocapsid protein (compared to 20kb in Ebola and $\sim$ 9kb in HIV).
- Envelope consisting of proteins and lipids. The surface proteins are the virial material.
- Spike protein (S). It looks like crown, hence the name corona.
- M, N, E proteins in envelope assembly.

6.6.3 Epidemiology: Going Viral

Definition 6.6.1. *Reproduction number, R0*, is an average number of people that an infected person can infect.

Definition 6.6.2. *Serial interval* is the time it takes to transmit from one person to another.

R0 for covid is estimated to be ~ 3 and serial interval is 1 week.

Note that the growth of infected people at least in the beginning (assuming that there is no vaccine and immunity) is exponential. Therefore, to stop the spread we have to decrease R0 until it is less than 1. To do that we can

- Wear mask.
- Physical distance.
- Come up with a vaccine.

There are different types of viral spreading:

- **Silent spread:** A large fraction (more than half) of infected people are asymptomatic or presymptomatic but can spread the virus.
- **Superspreading:** Some people (in large gatherings, who comes in touch with a lot of people and who have more virus) spread it a lot while other don't.

6.6.4 Viral Testing: RT-PCR

Fun Fact 6.6.3. Broad Institute that Prof. Lander leads runs the testing program for MIT,[10] 108 colleges in the North East, and 500 other organizations.

Once a sample is collected from nose, see 6.7, it is taken to Broad. To see if there are any virus particles in the swab, we look for the RNA of the virus:

- Convert the RNA to DNA using reverse transcriptase.
- Measure how much of DNA corresponding to the virus is present.

  To do a quantitative measurement we do (qPCR). In particular, we use nucleotides that have fluorescent molecules attached by a matching piece of nucleic acid to a clincher that keeps the fluorescent molecule from fluorescing. And when polymerase comes along, its exonuclease cuts the molecule which then fluoresces. The amount of fluorescence increases at each cycle of PCR. We can use the number of cycles required for the fluorescence to have certain intensity to quantify how much viral particles are there.

Fun Fact 6.6.4. As of this weekend, Broad Institute has done about 5.2 million test (1 in 14 test in the United States). At its peak, 99k test were done in a day.

6.6.5 Virus Life Cycle

- **Binding:**
  - Spike proteins bind to a receptor called ACE2 at *receptor binding domain.*

    Remark 6.6.5. Using Human Atlas Project, people figured out that olfactory express high amount of these receptors. Therefore, an infected person is likely to lose smell. Smell receptor express high level of ACE2 protein.
  - The viruses also have an enzyme called *protease* that cuts off the receptor binding domain.
  - There are hydrophobic element in their spike that is grips on our cell membrane and opens it up.
- **Fusion:** They have a spring mechanism that they use to throw in their genome in our cell.

[10] People who are staying on campus including me have to test twice a week.

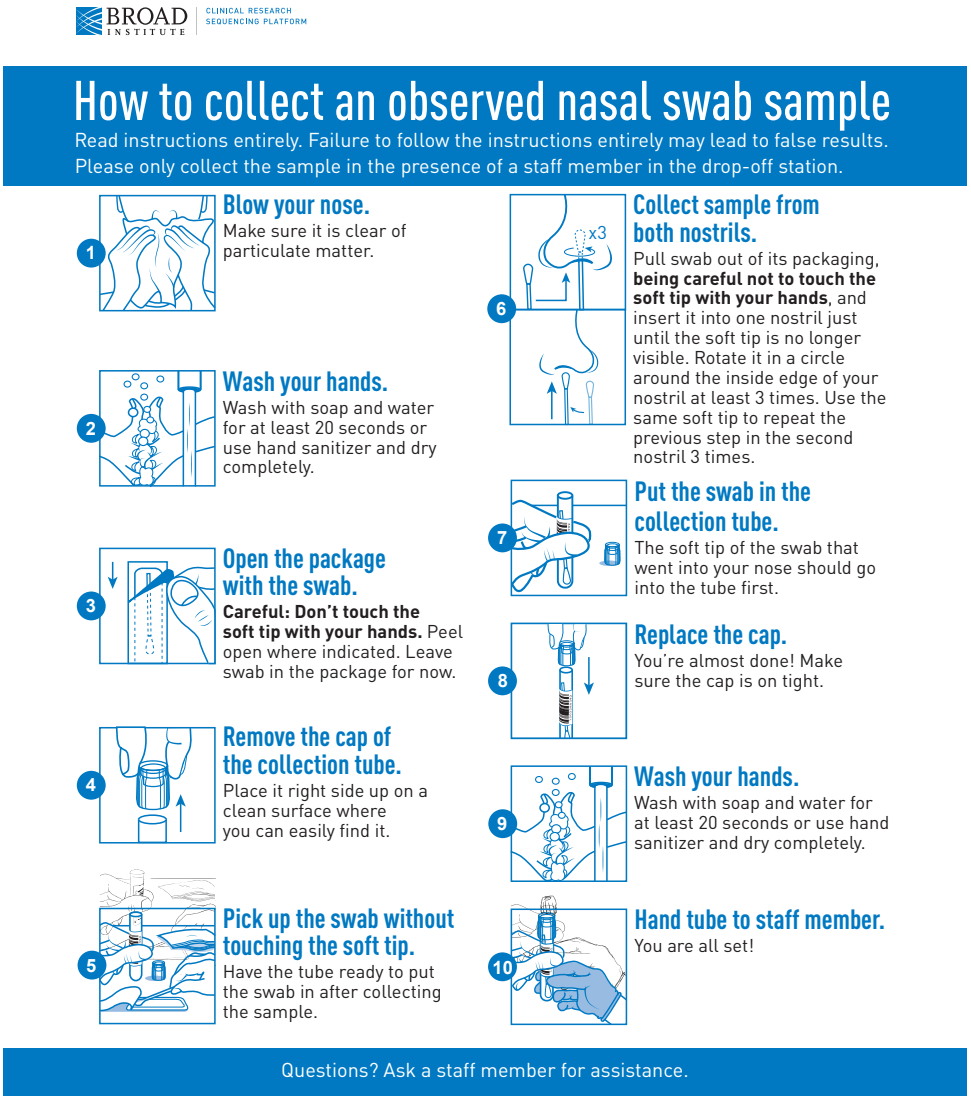


Figure 6.7: Swab test

- **Replication:**
  - Starts by translating certain protein.
  - Those proteins help the virus to create its own little compartments
  - The compartments take endoplasmic reticulum which causes it to create pinch off little spheres within ER.[11]
  - These sphere become a basis form basis for replication
- **Release:** The replicated viruses are released in the cell.

[11] Cells are not supposed to have double stranded RNA in cytoplasm. There are cellular mechanism to check it. And if the mechanism finds dsRNA, it alerts the immune system. There are hypothesis that corona virus forms little sphere to protect itself from the surveillance mechanism. Another reason could be to have high concentration.

6.6.6 Viral Genome

Recall that corona virus is RNA virus therefore it has RNA genome. RNA genome is a mRNA. It has 13 different genes which make 27 proteins.

- **Nonstructural proteins:** They help the virus keep processes going when it infects the cell.

  Remember that proteins are formed from self respecting gene in eukaryotes that have (P and O). In contrast, the gene of corona virus has a huge *open reading frame* that produces a big *poly-protein.* There can also be ribosome frame shift to form even larger protein. Then the virus uses protease (it is in the poly-protein) to cut it into pieces to get 16 proteins.

- **Structural and accessory proteins:** These consist of E, M, N, S etc.

  The virus copies the mRNA back and stops at different places to produce different messages (proteins) that can read different places of open reading frame.

Note that the virus is very long. Recall that RNA polymerase are error prone resulting in high mutation. For instance, HIV mutates at a high rate because it is error prone. But high mutation rate might lead to a less defective progenies. However, corona viruses have proofreading so it can have long RNA.

6.6.7 Body's Response

Now that we have overviewed the epidemiology and virus biology, we are ready to discuss our body's immune response. We will compare healthy response and dysfunctional immune response, see Figure 6.8.

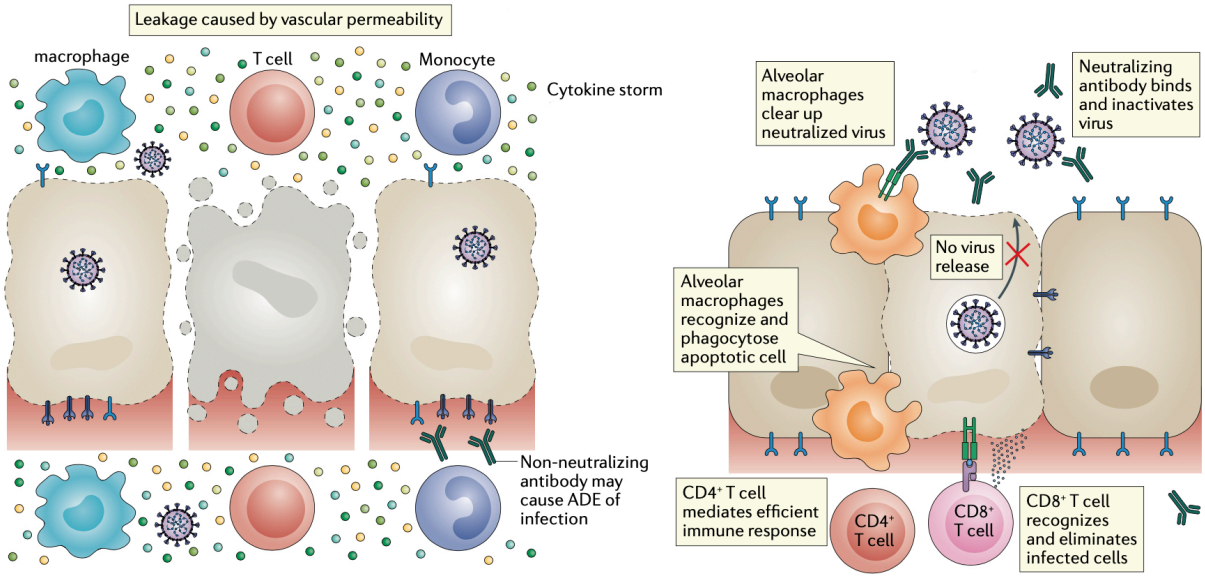


Figure 6.8: Immunity Response (Dysfunctional on left and healthy on right) derived from Tay et. al

- **Healthy Response:**
  - Killer cells and macrophages fight against the virus to clear the infected cells rapidly.
  - B cells in coordination with helper T cells secrete antibodies that neutralize the virus.
  - There is minimal inflammation and lung damage.
- **Dysfunctional Response:**
  - The immune system (T cells, macrophages and B cells) are infiltrated with viral content.
  - Cytokine will fight against our own cells leading to *cytokine storm*.
  - There are chances of getting pneumonia.
  - There is widespread inflammation and damages in multiple organs.

6.6.8 Vaccines: Strategy

- **Classical ways:** These are the vaccines that we described previously. They expose our body to SARS CoV-2 so that we develop a strong immune response.
  - Inactivated Vaccine: These contain chemically inactivated SARS-CoV-2.
  - Recombinant proteins vaccines: These consist of spike proteins purified from the virus.
- **Second generation vaccines:**
  - Viral vector vaccines: We inject another virus modified to express spike proteins.
- **Third generation vaccines:** These vaccines are not approved yet. The idea is to deliver to our cells genetic code to express spike proteins so that corona virus can't enter our cell.
  - RNA vaccines: It consists of RNA of the virus packed in lipid nanoparticles.
  - DNA vaccines: A circular DNA that encodes the spike protein is injected in our body.

Fun Fact 6.6.6. A week ago Pfizer reported about its vaccine that is 90% effective. The design of the trial was double blind and placebo controlled.

- 44K people participated.
- 90 individuals developed infection and showed symptoms.
- Prof. Lander guesses 86 got placebo and 8 got vaccine.

Although the result is promising, as of today, we don't know the details of

- Protection in subgroups (age, gender, race etc)
- Long lasting.

- Side effects.
- B cell and T cell response.

This morning, Moderna announced their RNA vaccine that has 94.5% effectiveness. In a press release, they said that 40K people were involved in the trial. Out of them 95 got infected: 90 got placebo and 5 got vaccine, 6.2. Note that in both vaccines the numbers

	Any	Severe
Total	95	11
Placebo	90	11
Vaccine	5	0

Table 6.2: Trial

of infected people who got vaccine are very small. Therefore, we need to grasp the data with care.

"If we play our cards right, we can get the pandemic under control by the end of 2021. But we have to worry about getting them under control equitably in parts of the world that don't have public health infrastructure and financial resources," says Prof. Lander.

Module 7

Protein Structure and Function

7.1 November 18

We will have Prof. Cathy for the next two classes on protein structure on their function. Today, her shirt has lots of bands in a gel which she plans to explain later. The classes will build on biochemistry, molecular biology (particularly gene expression) and give an introduction to recombinant genetic technology.

7.1.1 Protein Characterization

Protein (enzyme) characterization usually means determining:

- Enzyme reaction: substrates and products.
- Whether a *cofactor* is required (B_{12} and Fe).
- Kinetics: The speed of the enzyme.
- The binding affinity of substrates.
- Concentration of enzyme.
- Enzyme mechanism: How many steps are involved and role of catalysis.
- Enzyme Structure: structure of protein fold, active site location and composition.

We characterize proteins to

- Understand the basis of genetic diseases arising from errors in DNA that encodes proteins.
- Design inhibitors (chemotherapeutic agents, antibiotic, and antiviral agents).
- Engineer enzymes to be faster or make a different product.

Fun Fact 7.1.1. People are making covid protein and characterizing them to see how they work.

Proposition 7.1.2. *There exists a process to characterize a protein (strucutre).*

7.1.2 Purifying Protein

Before proving the statement, we need to construct a way to get pure protein. In the past, people went to slaughter house and use blenders to purify it. However, we will construct the purification process using recombinant technology.

First, we get

- a gene that encodes the protein (clone or buy).
- *Expression vector* (buy).
- *Expression system* that makes our protein for us (e.g *E coli*).

To set up an (*E. coli*) expression system,

- **Grow bacterial cells** to a certain density before inducing/starting expression. Cells won't have enough energy to divide when the density is very low. We can use
  - *Inducer concentrations* to tune expression
  - An *inducible system* to control density of cells and inducer.
- **Purify the protein:** We want to separate proteins from all other proteins that *E. coli* makes.

  We can
  - Modify an end of a gene to include purification tag like *histidine tag* (His-Tag).
  - Use a system that recognizes His-Tags to purify the proteins. In practice, we
    * Use metals like nickel in the form of beads in a column.
    * Run a solution containing proteins, so only proteins will attach to the column.

Definition 7.1.3. A *his-tag* is a tag of multiple histidine codons added to an end of a gene, so multiple that histidine residues are added to an end of a protein.

Remark 7.1.4. We don't use a single single histidine as a tag because of its low affinity.

Earlier, we hand waved how to use histag and a nickel column to purify proteins. In practice, we use multiple columns to carry out the process:

- **Load column:** We put in proteins from a cell.
- **Flowthrough:** It has solution containing proteins (without His-tag) that flows through column.

- **Wash column:** In this column, we remove proteins (without His-tag or weakly bound to Ni) are easily washed out of the column.
- **Elute column:** After we have washed out multiple times to remove proteins with low affinity, we elute the column to remove proteins that are tightly bound to Ni to get His-tagged protein. We use *imidazole buffer* to elute the column.

The most common method to make sure that our process worked is sodium dodecyl sulfate (SDS) Page (polyacyrl amide gel electrophoresis). SDA page (a gel) separates proteins by molecular weight.

Recall that we used Lemma 4.5.1 and gel to sequence DNA. The idea is similar when we separate protein:

- Proteins are *denatured.*
- They are covered with SDS to provide a uniform negative charge.
- Negatively charged denatured SDS coated protein travel through gel when electric current is applied.
- Using Lemma 4.5.1, we know that larger proteins move more slowly (top of gel) and smaller proteins more quickly (bottom of gel).

Remark 7.1.5. 1. A gel has multiple lanes. In addition to (multiple) flowthrough, wash and elute lanes, there is a *marker lane* where we put molecules of known size for comparison, see Figure 7.1

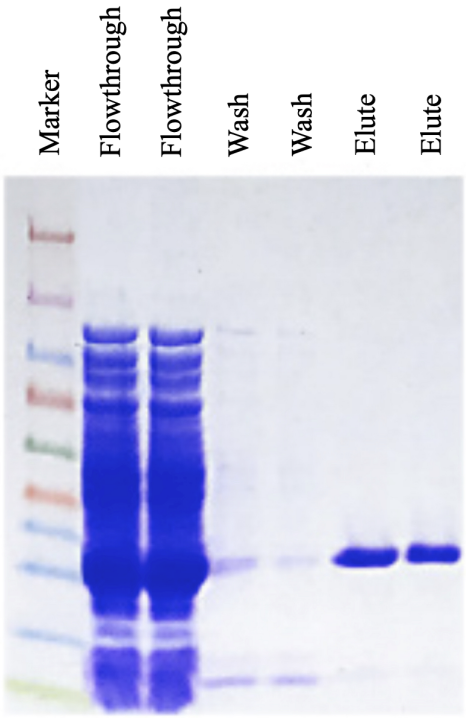


Figure 7.1: Results of protein characterization in a gel

2. A protein of a particular molecular weight will appear as a band that is visualized by a stain. Note that a pure protein has a single band in a lane.

7.2 November 20

"Biochemists do it for the proteins."

Today, we will complete the discussion of protein structure.

7.2.1 Enzyme Strategies

Last time, we wanted to learn how enzymes work as they can be used for chemotherapy, antibiotics etc. Recall that enzymes catalyze reactions by lowering the activation energy barrier. In particular, the ways in which an enzyme can catalyze a reaction are:

- **General acid base catalysis:** Enzymes stabilize transition states by accepting/donating protons. Recall that an acid donates a proton and a base accepts a proton
- **Covalent catalysis:** The substrate forms transient covalent bond with transition states and stabilize it.
- **Metal ion catalysis:** Metal ions activate water molecules for hydrolysis reactions.

Example 7.2.1. A protease speeds up the hydrolysis of a peptide bond. One protease has a catalytically essentials Ser residue at position 195.

Poll 7.2.2. If the pKa of a Ser side chain is typically 13, would you expect the Ser side chain to be neutral or negatively charged at pH 7.4?

- (-) charged: pH is below pKa so Ser is deprotonated
- (-) charged: pH is below pKa so Ser is protonated
- Neutral: pH is below pKa so Ser is protonated
- Neutral: pH is below pKa so Ser is deprotonated

7.2.2 Determining Protein Structure

Now that we have described how to purify proteins and how they work, let us prove Proposition 7.1.2. Note that we have a partial answer to this question because we at least know how an enzyme can catalyze a reaction. Now we will prove that there is a way to determine the structure of a protein.

Definition 7.2.3. *Resolution* of is a measure of how precisely atomic position are known.

At a resolution of

- Low: 15-20 Å, we can determine overall shape of protein molecule.

- 6Å, we can model α helices.
- 4-5 Å, we can model β sheets.
- 2.5-4 Å we can model side chains.
- 2 Å or better, we can model atoms with high precision.

Proof of Proposition 7.1.2. The proof is by giving examples. Recall that we can use light microscope to look at single celled organism but not proteins. Therefore, if we can find a way that has high resolution, we can "see" proteins and determine its structure.

In fact, we can use *(cyro)-electron microscope (EM)* or *X-ray crystallography* to determine the structure. The resolution of EM is 2.8-3.5Å and crystallography is 1.5-2.0Å. □

Remark 7.2.4. There is an open source protein data bank. As of November, there are 170 thousand structures in the database.

We describe the basis of electron microscopy below:

- Uses an electron beam.
- Protein molecules are coated onto a grid. The protein specimen must be preserved (stained or frozen hydrated).
- EM collects thousands of images .
- Images of particles are sorted, aligned, classified and averaged.
- EM maps are calculated and protein structure model is built into the map.

  Remark 7.2.5. Few years ago we would do the calculation and build a map by hand. Now we can use techniques from machine learning.

On the other hand, the basis of X-ray crystallography is:

- Use of x ray beams.
- Protein molecules are crystallized. Crystals must be protected (cyro cooled).
- Collect images of X-ray diffraction pattern generated by crystal.
- X-ray waves to generate an electron density map
- Structure is validated by back calculating diffraction

Remark 7.2.6.

1. Fibrils and membrane proteins don't crystallize well therefore cyro-EM is better for them.
2. It is harder to get structures of multiple *conformational states* in X ray crystallograpy but easier in EM.
3. We can validate structure by back calculating diffraction pattern in crystallography but it is hard to validate in EM.

4. X-ray crystallography is cheaper than EM.

Fun Fact 7.2.7. Cyro-EM microscopes cost 3-19 million dollars and requires buildings that have low vibrations. And MIT.nano has low vibration buildings.

7.3 November 30

Today, Prof. Lander will explain how we can use molecular biology to cure diseases. In the next 50 minutes, we will theoretically cure a heart disease.

7.3.1 Heart Disease

Assuming that heart pumps blood to and from tissue, the purpose of pumping blood are:

- Provide oxygen to cells.
- Remove waste products (carbondioxide).
- Pumps cells (white blood cells/red blood cells).
- Distributes nutrients.
- Distributes hormones and platelets.

Shutting down our circulation system will lead to a failure of aforementioned purposes.

Lemma 7.3.1. *If there is a failure in blood pumping, a person can get brain strokes or heart attacks.*

Proof. Recall that a heart pumps blood to body, brain, and itself through *blood vessels*. There is a chance that we develop atherosclerotic plaques (protein, lipids, and cholesterol) in our vessel. It will makes those vessels stiffer and decreases the volume available for pumping. Sometimes, a piece of plaque gets knock off and might block a vessels with narrower hole which will block the blood flow. If there is a blockage in the vessel that supplies blood to the brain, we will get brain stroke. If there is a blockage in vessel in heart, we will get heart attack. □

7.3.2 Cholesterol

Cholesterol is a hydrocarbon, see 7.2. It is highly hydrophobic and is also waxy. It can turn into cholesteryl linoleate (a cholestol ester).

The functions of cholesterol are

- **Structural role in cell membranes:** Fifty percent of lipid bilayers in plasma membranes is cholesterol.

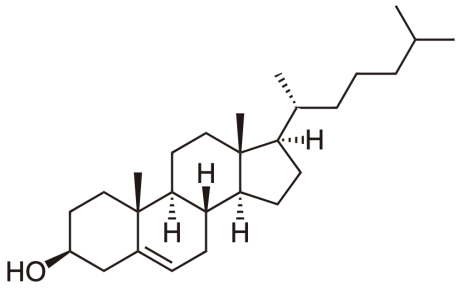


Figure 7.2: Structure of cholesterol

- **Precursor** of
  - Steroid hormones (estrogen).
  - Vitamin D
  - Bile acid (these emulsify fats in the digestive system).

We get cholesterol from

- **Diet:** Eggs, Butter
- **(Bio)synthesis:** We make cholesteroal in our own body starting from acetyl CoA.

  Fun Fact 7.3.2. Konrad Bloch got a Nobel prize for figuring out the pathways involved in the synthesis.

  In short, the process is as follows:

  acetylCoA $\to\to$HMGCoA $\to$ mevalonate $\to\to \cdots \to$ cholesterol.

  For our discussion, we will just look at HMG-CoA$\to$ mevalonate. HMG-CoA reductase carries out this transformation.

Remark 7.3.3. 8 hundred thousand people die of cardiovascular disease. There is a correlation between the chance of infection and the amount of cholesterol in the plaques.

7.3.3 Transport of Cholesterol

Cholesterol is made in every cell but a large portion comes from liver. Before it is distributed around the body, cholesterol gets esterified because cholesterol does not dissolve in blood. Then the cholesterol is put in lipoprotein particles. Based on their density, lipoproteins are classified into:

- Very Low Density Liprotein (VLDL).
- Low Density Lipoprotein (LDL)
- Intermediate Density Lipoprotein (IDL)
- High density Lipoprotein (HDL)

- Chylomicrons

In summary, liver secretes cholesterol which goes to $VLDL \to IDL \to LDL \iff HDL$.

7.3.4 Connection to Heart Disease

Proposition 7.3.4. *Cholesterol is connected to heart disease. In other words, we can predict heart attack based on cholesterol level.*

It 1856, Rudolf Virichow observed lipid accumulation in arteries of people who died of heart attack. In 1913, people experimented with rabbits by feeding high cholesterol diets. As a result, they developed symptoms like atherosclerosis.

However, we can't feed human cholesterol. In 1950s, there was an epidemiological study called Framingham Heart study. They followed people over time and measured LDL and HDL level. As a result, higher level of LDL was a correlated to heart disease. In contrast, higher the HDL level, the less the risk of heart attack, see 7.3

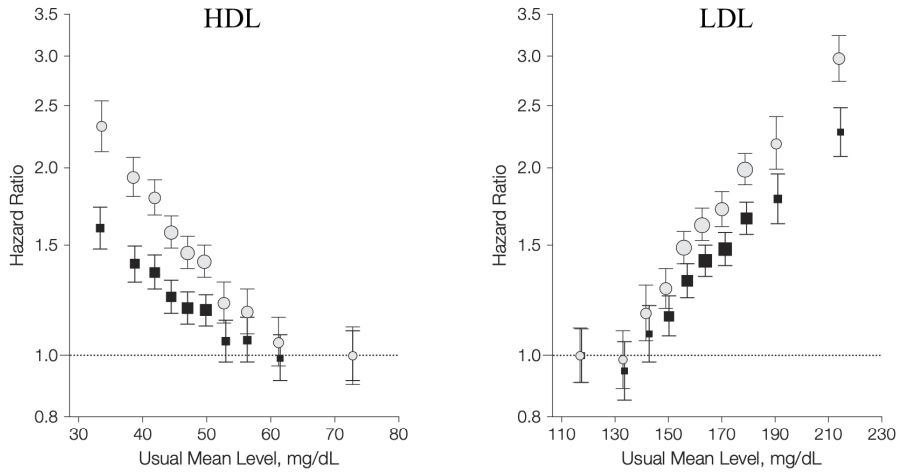


Figure 7.3: Correlation between LDL and HDL with chance of getting heart disease. Derived from JAMA

However, the correlation does not prove Proposition 7.3.4. We would have to manipulate people. We need to see if decreasing LDL decreases heart disease.

7.3.5 Genetics of Heart Diseases

In 1970s, Michael Brown and Joseph Goldstein studied patients with familial hypercholesterolemia (FH): high level of LDL in blood. It turned out that homozygotes (fh/fh) people with high level of cholesterol (greater than 600mg/dl) died to heart attack before getting 20 years old compared to normal person (+/+) with low cholesterol level (100mg/dl) who lived a normal life. And heterozygotes (fh/+) had cholesterol level of 250 mg/dl and

died 10-20 years earlier than normal people. The frequency of homozygotes was $q^2 = \frac{1}{10^6}$. (Check what p is).

With a supportive result, Brown and Goldstein decided to find a molecular basis of heart attacks. They

- Cloned genes for FH using *fibroblasts*.
- Radioactively labelled cholesterol and fed it to fibroblast cells.

They observed that fibroblast of heterozygotes took half as much cholesterol as those of normal people. They hypothesized (later proved experimentally) the existence of receptor in fibroblasts that takes up cholesterol. Further, they claimed that FH patients had mutations in the gene that encodes the LDL receptor (LDLR). When they put ordinary fibroblast in low cholesterol environment, they upregulated LDLR and HMGCoA reductase.

Proof of Proposition 7.3.4. The preceding discussion implies the proposition. □

Fun Fact 7.3.5. In 1985, Brown and Goldstein won Nobel prize for figuring out the connection we described.

7.3.6 Rational Therapy

Now that we know what causes, let's cure the heart disease. In particular, we want to reduce LDL levels:

- Eat less cholesterol: It reduces LDL level but the reduction is only about 10%.
- Consume cholesterol with resins that bind with bile acid[1] and prevent our body from taking up the resins: It is clear that the bile acid gets excreted, so our body starts to use up cholesterol to produce bile acids. This process will reduce cholesterol level by 20%.
- Inhibit HMG-CoA reductase: People found a inhibitor called *lovastatin* that resulted in 60% reduction in LDL levels. In fact, it decreased the rate of heart attacks.

However, increasing HDL did not do anything. In fact, millions of dollars were spent but it turned out to be just a correlation.

7.4 December 2

Last time, we used genetics, chemistry, and molecular biology to solve heart disease. In general, if we understand our system well we can come up with therapies. Today, we

[1] Bile acids emulsify fats and are formed from cholesterol.

won't be able to solve all types of cancer in 50 minutes, but we will see why people have cancer.

7.4.1 Cancer

Definition 7.4.1. *Cancer* is a condition when there is unregulated growth of cells. When cells divide our of control, they form an abnormal mass known as *neoplasm* or *tumor*.

Definition 7.4.2. If the abnormal mass stays in same place, we say that the tumor is *benign*. In contrast, if cells invade nearby tissues we call the tumor *malignant*. If they distribute through the body (blood or lymph) to form secondary tumors in other locations, we call them *metastatic*.

Remark 7.4.3. Benign does not mean something harmless. But they can be removed surgically.

Example 7.4.4. The most common cancer in the states are in lungs, stomach, breast, colon, uterus etc. There are 1.2M new cases of cancer per year in US. It accounts for a quarter of death in the US.

Remark 7.4.5. Cancers (usually) derive from a single cell. We don't become aware of it until we get 10^8 cells. It can be detectable on X ray. It takes about 10^9 cells for it to be palpable (1cm). If it reaches 10^8 cells we are dead.

Cancer usually arise from mutation. There are about 10^{16} cell division in our lifetimes with 10^-8 mutation rate per base accounting for 10^-6 mutation in gene. Note that mutation arise by chance, inheritance, and *mutagens* (smoking).

However, it takes multiple mutation to cause a cancer. For instance, a cell will undergo some mutation resulting in slightly higher replication rate. But nothing that would cause tumor will kill us. However, some daughter cells might mutate to have higher replication rate and so on. Finally, tumor can gain ability to affect the nearby tissues. Meanwhile, other mutation can cause cells to produce abnormal proteins. There is surveillance system that destroys those proteins (even cells sometimes), but it does not always work.

Today, we have tools to know what these mutations are. We can even compute the replication rate. If we know the gene, we know mutation and then we can make therapies.

7.4.2 Regulation of Cell Growth: Growth Factors and Receptors

Recall that cancer arise because of uncontrolled growth. In a normal person, cell division is strictly regulated.

Definition 7.4.6. *Growth factors* are the signals given by neighboring cells to regulate the cell division.

Example 7.4.7. For instance, wound cells upregulate cell division of neighboring cells. There are epidermal growth factor (EGF, nerve growth factors (NGF) etc.

These growth factors pass the signals to cells:

- Bind to growth factor receptors on the cell surface.
- Causes the a pair of growth factor receptors to dimerize so that their *cytoplasmic domain* come together inside the cell.

The receptors have tyrosine in their cytoplasmic domain and undergo a kinase activity:

- The pair of receptors phosphorylate each other.
- Phosphates change the structure of receptors which is recognized by an adapter protein called Grb2.
- Grb2 binds to another protein Sos.
- Sos binds to RAS.

7.4.3 Regulation of Cell growth: RAS

A major player in the signal processing we just described is RAS. It has two states:

- **Off state:** Binds to guanine diphosphate (GDP).
- **On state:** Binds to guanine triphosphate (GTP).

There are proteins called guanine nucleotide exchange factors (GEFs) that stimulate the transition from on state to off state.

"SOS is a GEF."

Remark 7.4.8. Note that RAS can go from on state to off state itself with the help of GTPase that can remove a phosphate group. There are proteins called GAPs that can stimulate GTPase.

When RAS activates, it turtles all the way down:

- RAS stimulates a protein kinase RAF.
- RAF (protein kinase) phosphorylates another protein called MEK.
- MEK phosphyraltes protein called ERK.
- ERK turns on transcription factor.
- Transcription factor turns on genome that turn on cell growth program.

Definition 7.4.9. The process of activating RAF, MEK, and ERK is amplification.

7.4.4 Mutations That Cause Cancer

If the program we described before has a bug we will get cancer:

- **RAS stays on:** If there is a mutation in active site that prevents the GTPase activity will result in constitutively on RAS that will always send a signal turning on RAF and the cell growth is uncontrolled.
- **Mutations in GAPs** might prevent GAPs from stimulating GTPase activity.
- **Mutation in MEK, RAF and ERK** can turn them constitutively on.
- **Over expression of transcription factors.**
- **Mutations in growth factor receptors** can lead to a constitutive dimerization.
- **Self stimulation of growth factor:** Recall that growth factors are supposed to come from neighboring cells. But a cell itself can make a lot of growth factors.

Remark 7.4.10. If the genes that encode Grb, RAS, and MEK are deleted there would be no signal at all and therefore no cell growth.

7.4.5 Rational Therapy

Now to cure cancer, our goal is to stop the mutations. In particular, we should control the kinase activities of the proteins (RAS, MEK, ERK etc). In fact, we can make kinase inhibitors (RAF inhibitors, ERK inhibitors etc.)

Example 7.4.11. Sorafenib (drug) are used for skin cancer, mutations in BRAF (type of RAF).

Remark 7.4.12. We can't make sure that only cancer cells get the inhibitors. But cancer cells turn out to be sensitive to kinase activities. We can manipulate it by giving appropriate doses of inhibitors.

In the past, people used to give random drugs to see if the cell growth would stop. But now they haev *targeted therapy*:

- Pick a point in signaling process for cell division.
- See if there is mutation.
- Find a way to stop the mutation.

Remark 7.4.13. We can sequence DNA to see where the mutation is and give medicine accordingly.

Module 8

Big Picture Biology

8.1 December 6

In the first lecture, we said that this class won't go farther than chemistry. Nevertheless, we made brief references to physics (for instance Lemma 4.5.1) and math (almost everywhere). It is evident that we can't run away from philosophy now. Maybe from linguistics.

8.1.1 Science and Society

This is the last class with Prof. Lander.[1] It is about reflecting on how taking 7.012 will contribute us to make important decisions that we will confront to. As MIT graduates and responsible members of society (assuming that we live in a "society"), we need to engage in important questions of society. Today, we will open up a discussion on ethics and CRISPR germline editing.

CRISPR gene editing has been around for about only seven years. But there are already handful of companies that are trying to cure diseases by editing or restoring genes:

- Dominant Progressive blindness: We can take a virus that carries CRISPR protein and guide RNA and inject it into eye. The virus will cut the gene (retinoid pigmentosa) that causes blindness.
- Repair muscle in duchenne muscular dystrophy: It worked in dogs and mice.

The applications that we have discussed so far involve editing somatic genomes. In the past, CRISPR has been used to edit egg or fertilized eggs (germline) in mice. Note the effect of editing embryo is in every cell. We can also edit our germline. But if we do so, our gene pool will eventually change. There will be no way back. For better or for worse.

[1] For this class, we read the following articles: https://www.nejm.org/doi/full/10.1056/nejmp1506446, https://www.nap.edu/read/25665/chapter/2 (optional) and https://www.nature.com/articles/d41586-019-00726-5.

In 2018, germline editing went from being a theoretical topic to a feasible process. He Jiankoi claimed that he edited a genome of two twin embryos. In particular, he deleted CCR5, a receptor for HIV virus that causes AIDS. It created a firestorm in international communities.

In march 2019, Prof. Lander et.al called for a moratorium for five years. Further, US National academy and British Royal Society created a commission that Prof. Lander served on to discuss about gene editing. They have published a report in 2020.

Question 8.1.1. Should germline editing be done? Whether it is safe or good is another question?

Remark 8.1.2. The question of safety might fade away soon given a rapid technological advancements. However, is it *good* or *bad* has to be discussed.

Conjecture 8.1.3. *Assuming that there is a free market in a "society" and status quo in terms of regulations, companies that will exploit public to sell "best" genes will emerge. Genes for height, color of skin and intelligence.*[2] *A notion of "best" gene will marginalize of certain group of people. Further, our gene pool will be less diverse.*

Remark 8.1.4. Diversity is a concept that geneticist have understood for decades. For example, diversity of wheat is helpful when some insects infect wheat. In the name of gene enhancing or improving human species, we might start lacking diversity in genes. It won't be good for us as a species.

Conjecture 8.1.5. *Assuming that there is no effort in solving economic inequality, gene editing will increase the gap.*

Remark 8.1.6. Unless there is a free (cheap) health system, gene editing is going to be expensive. Note that in vitro fertilization is already inaccessible to poor people/countries. People who will get "good" genes will be from affluent background. And poor people will fall into poverty trap. But maybe gene editing will be cheaper with time like penicillin.

The floor is ours to prove/disprove conjectures and find a ground to make our world a better place.

8.2 December 9

"Keep calm and ~~call~~ [become] a biologist," says Prof. Cathy's shirt.

Congratulation! We have made through this semester together. Today, Prof. Cathy will talk about saving our planet as a biologist.

Proposition 8.2.1. *Assume that our planet is facing issues like: climate change, air pollution (carbon emission), land pollution (oil spills, heavy metals), water pollution (lead in drinking water) etc. There are roles a biologist can play to solve the problems.*

[2] Maybe this is good because we can come up with mathematicians who can solve all the open problems.

8.2.1 Using Bacteria to Save the Planet

In most of our discussion on bacteria, we pictured them as bad players. However, we can be friends:

- Bacteria can metabolize compounds that are harmful to the environment and remove them from our environment.
- Bacteria or bacterial enzymes or bacterial pathways can be used to replace environmentally unfriendly chemical or processes to limit pollution and lower energy usage.

Bacteria can help by metabolizing pollutants

Recall that metabolism requires two things:

- A molecule with stored energy that can be oxidized.
- Terminal electron acceptor which is reduced.

Glucose is a good source of energy but a lot of the times bacteria don't have it. In that cases, bacteria (*Thauera aromatica*) survive anaerobically and can metabolize components of oil spills like toluene, see Figure 8.1.

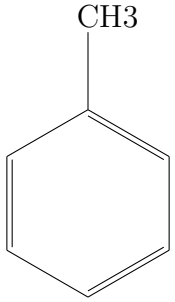


Figure 8.1: Toluene

During the anaerobic respiration,

- **Energy source** is toluene.
- **Terminal electron acceptors** are NO_3^-, Fe^{3+} or SO_4^{2-}.

Therefore, we can deposit *T. aromatica* cultures in areas with crude oil spills.

Question 8.2.2. How can an enzyme bind toluene although it has no functional groups?

Poll 8.2.3. Which of the following:

- Hydroxyl: $-\mathrm{O}-\mathrm{H}$
- Carbonyl: $-\overset{\overset{\mathrm{O}}{\|}}{\mathrm{C}}-$

- Carboxyl: $\mathrm{—\overset{\overset{\mathrm{O}}{\|}}{\mathrm{C}}-OH}$
- Amino group: $\mathrm{—NH_2}$
- Phosphate: $\mathrm{—PO_4^{3-}}$
- Sulfhydryl: $\mathrm{—SH}$

are common functional groups in biology?

1. All of the above
2. None of the above
3. Only the groups with oxygen atoms are functional groups
4. Only charged groups are functional groups.

Note that the substrate is hydrophobic. The potential ways in which it can form bond with toluene are (from strongest to weakest):

- **Covalent bond:** Link atoms together. They are formed transiently in covalent catalysis but not involved in binding substrates. Therefore, this bond is not formed.
- **Electrostatic interaction:** $+$ and $-$ attract each other. But toluene is not charged.
- **Hydrogen bonds:** Electronegative atom with an H in polar bond will make contact with an electronegative atom with a lone pair of electrons. In toluene, there the electronegativity difference $\Delta\chi = 2.55 - 2.20 = 0.35$ which is less than 0.4. Nay for hydrogen bond.
- **Van der Waals:** Weak packing interactions. Hydrophobic molecules like toluene can bind enzymes using van der Waals ineraction. Transient partial charges form and these charge will bind things.

We can use x ray structure to see how it is packed in enzyme:

- Hydrophobic residues and hydrophopic parts or polar residues form direct contacts to toluene via vander waal ineractions.
- Second substrate fumarate is held in place by making an electrostatic interaction with Arg and a hyrdogen bond with Asn.

Moral: Bonding is really important in biology.

Putting Fewer Pollutants into Our Environment

We can replace an environment-unfriendly chemical process that produce chemicals harmful for environment with a biological process. For instance, Vitamin biotin is made (globally 10-30 tons and several hundred million US dollars per year) by 15 step chemical synthesis which is not commercially optimal. Instead, we can make biotin using bacterial enzymes. The present synthesis involves a very slow enzymes. We use biotin as feedstock for pigs.

Poll 8.2.4. Recall that enzymes work by:

- Making the reaction spontaneous
- Speeding up reactions by lowering the activation energy barrier
- Changing ΔG of the reaction
- All of the above

Answer: Lowers the activation energy.

We can also replace environmentally unfriendly chemical process that consume high energy or release of greenhouse gas with a biological process. For instance, the Haber-Bosch process uses 1% of the total world's energy annually to synthesize ammonia:

$$N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g).$$

1.6×10^{10}kg of ammonia are produced by the process in US per year. The process uses a lot of energy because it is hard to break the nitrogen bond.

Ammonia is used for fertilizers and explosives. Haber developed *phosgene* as chemical warfare in the World War II.

Remark 8.2.5. We should use science for good.[3]

Instead of using Haber Bosch process, we can use bacterial enzyme (nitrogenase) to make ammonia.

Proof of Proposition 8.2.1. From the above discussion, it is clear that there are ways in which knowledge of biology can be used to protect and repair our planet. □

[3]But good means different things to different people. Maybe we should not view science from a good/bad perspective.