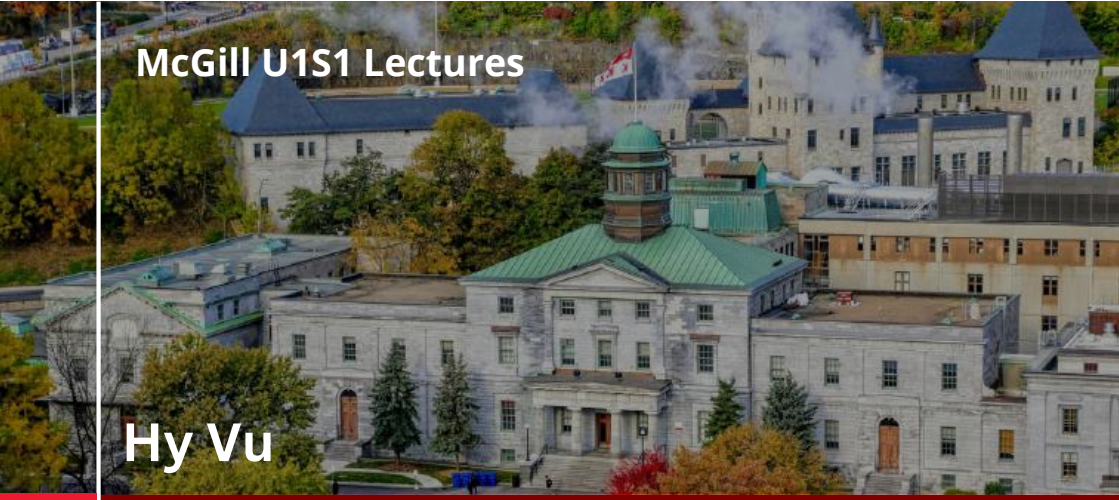


McGill U1S1 Lectures

Hy Vu

Molecular Biology

Lecture Notes

Prof: Dr. Hasting, et al.

McGill

Molecular Biology

Hy Vu

Prof: Ken Hastings, Paul Lasko
and Richard Roy

August 2023

Contents

Foreword

This is all the lectures turned into notes from my courses of McGill, more specifically BIOL 200. I want to note that by no mean, these are of my owns but of professor Ken Hastings, Paul Lasko and Richard Roy. I divide their lecture into according order: Part I (All of Paul Lasko and Ken Hasting's Lecture, which are typically all of basic molecular biology), Part II (Richard Roy's lecture on transcription) and Part III (Richard Roy's lecture on application of these techniques).

To be more specific: chapter 1 (introduction), chapter 2 (every lectures about proteins), chapter 3 (every lectures about DNA and genetic) and chapter 4 (every lectures lectures about detecting nucleic acid), chapter 5 (every lectures about transcription), chapter 6 (every lectures about translation) and chapter 7 (2 final lectures about techniques of system biology and gene targeting).

Prerequisites: General Biology I + II, adequate knowledge of general and organic chemistry.

Part I

General Principle of Molecular Biology

Chapter 1

Introduction

1.1 Building Blocks of Cells

Almost all of biological system are made up tiny identical building block which can combined together to create a bigger building block.

Definition 1.0 Identical biological subunit (mentionned above) are called **monomers**. Monomers is a further classified into 3 types of molecules: monosaccharide, amino acids and nuleotides.

Remark 1.0 *Although all polypeptide can be considered as macromolecules, the inverse is not so true. Not all macromolecules can be polypeptides since macromolecules can be made other molecules than monomers.*

Definition 1.1 When ≥ 2 identical monomers covalently bonded together they form **polymer (macromolecule)**. The prior classification of monomer when turned into polymers would be: polysaccharide, protein (polypeptide) and nucleic acids (polynucleotides).

The general structure of monomers of same grouping is that they have a *characteristics elements* and a *common elements*. The charateristics elements helps each monomers of the same group to differentiate/individuate themselves. While common elements helps each monomers linked together via polymerization to form polymers. The linkage of monomers varies from 1 group to the next i.e. if monomer have 1 linkage site, then only 2 monomers can link with each other; if they have 2 linkage site, then

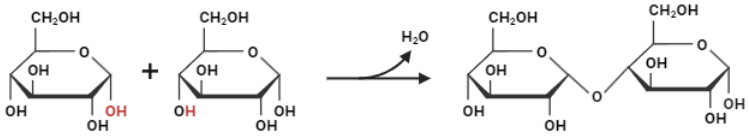


Figure 1.1: Dehydration reaction between 2 monosaccharide, specfically glucose. 1 glucose lose a hydrogen atom, the other an OH group.

the polymer would be *linear* and if ≥ 3 linkage site, then the polymer would be *branched*.

INSERT IMAGE

Remark 1.1 for amino acids and nucleotides, the monomers can only be linear for simplicity and packing. Their linkage site is often time asymmetric therefore the can only polymerize from 1 end in 1 direction (*unidirection*) linearly and end at another (chemically) distinct end.

Each covalent bond that hold the polymer structure would differs: for polysaccharide (sugar), the covalent bond is called **glycosidic bonds**; for polypeptide, the bond would be **peptide bonds**; and for nucleic acids, it would be **phosphodiester bonds**.
Even when these covalent bond are differs for each polymers, polymerization is the same for all. The reaction that create covalent bond (within a polymer) is called **dehydration reaction**, of which result in a net loss of an $H-$atom from on monomer and an $OH-$group from the other (see Figure 1.1). The same is true with the decomposition of polymer back to its monomers, but instead of losing H_2O, the 2 monomers involved gained H_2O; the decomposition of polymer is called **hydrolysis**. The covalent bond of these monomers are generally stable under normal biological condition thus making its polymers useful for normal cellular works

Definition 1.3 Polymers that are made and maintain in the body are called **biopolymer**. Any polymers that are not made by the body are called *synthetic polymers.*

Remark 1.2 *One misconception that students have is classified fats with polymers. They're not polymers but they're an important type of macro-*

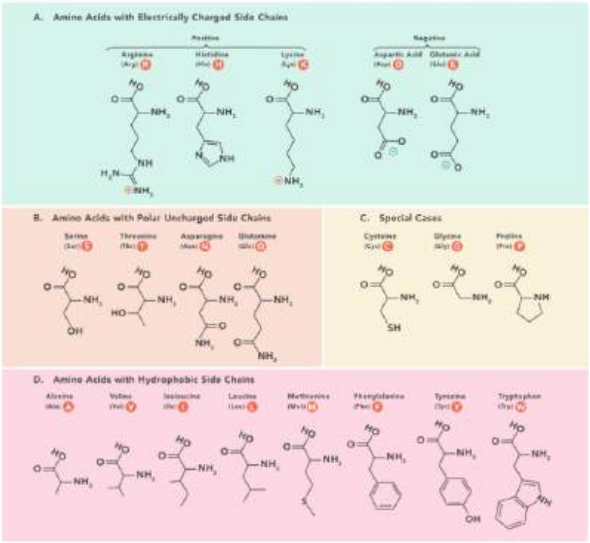


Figure 1.2: Table of amino acid with skeletal structural formula (the H-atom won't be apparent; to know more, review organic chemistry: skeletal formula).

molecules that can form the lipid bilayer of all cells.

The following portions will look more at the mentioned monomers and its roles.

1.1.1 Amino Acids and Side Chains

The building blocks of proteins are called amino acids; in fact, there is not just 1 type of amino acids, the monomers consists of 20 different amino acids. (see Figure 1.2). Amino acids is made up of 5 components: α-carbon, amino group ($-NH_2$), carboxylic acid group ($-COOH$), hydrogen atom and side chains (R-group). Like we've said before, amino acid has asymmetric linkage site which means it would have 2 separate ends called: N-terminus (at the amino group) and C-terminus (at the carboxylic terminus). When amino acids polymerize, it always starts from the N-terminus and add amino acids on the C-terminus.

INSERT IMAGE

All amino acids are identical, the only component that varies is the R-group. Although there are a lot of amino acids, we can always classify them according to their size, shape, charge, water solubility and R-group reactivity.
We begin with **hydrophobic amino acids**; the main characteristic with their R-group is that they're non-polar thus is hydrophobic. The key indication is that it contains mainly carbon, hydrogen or a phenyl group (benzene). We can see these key points in *alanine, valine, isoleucine, leucine* and *methionine*, where they consists of only linear chain of H and C (Except for Met with sulfur). The rest such as *phenylalanine, tyrosin* and *tryptophan* have a phenyl group (which is very nonpolar).

Remark 1.3 *Sometimes it's confusing to see if a molecule is polar or charged but the key different is that: a polar molecule has a net charge of 0 but the* $+$ *and* $-$ *is scattered through out the molecule so that 1 side is more* $-$ *than the other v.v., a charged molecule is any molecule with a net charge of* $-$ *or* $+$*.*

Contrarily, **hydrophilic amino acids** are all amino acids that are polar. For theese polar amino acids, we can divide them into 3 kinds: acidic, basic and uncharged. For **basic amino acids**, their R-group have a net charge of $+$; their main characteristic is having many amino group $-NH_2$ which are *lysine, arginin* and *histidine*. Meanwhile **acidic amino acids** have net charge of $-$ and the ket identification is having many carboxylic group $-COOH$ which are *aspartate* and *glutamate*. Then for **uncharged amino acids**, their R=group is uncharged but are polar due to the presence of hydroxyl ($-OH$) and amide group ($-CONH_2$).

Observation 1.0 Histidine has a ring containing 2 nitrogens are called **imidazole** that can shift from positively charged to uncharged according to changes in conditions (such as pH) via deprotonation.

Finally, the last 3 special amino acids that does not fit in any classification, the only key identifiers are **glycine** is the smallest amino acid thus their R-group would be the smallest atom, hence H. *Cysteine* is the only one that possess a **sulfhydryl group** ($-SH$). And *proline* has the almost the same ring as histidine but with missing double bond and an N atom.

Observation 1.1 Cysteine can deprotonate its sulfhydryl group to become a thiolate ($-S^-$) which are essential for most **proteases** (enzyme that destroy proteins). When being part of a protein, that same sulfhydryl group can de-

protonate into thiolate which can form a covalent bond with another thiolate. This covalent bond is strong and stable and is called **Disulfide bond**.

Remark 1.4 *Although we need almost all 20 amino acids, our body is only capable of producing 11 of the 20 amino acids. The other 9 are called* **essential amino acids** *that must be consume via certain plants or animals. These 9 amino acids include phenylalanine, valine, threonine, trytophan, isoleucine, methionine, leucine, lysine and histidine.*

Polymerization components

Aminod acids cannot simply polymerize together on its own. It needs an template that contains a specific sequence of polypeptide and a mediator that connect amino acids together. The template that contains this sequence is called *mRNA* and the mediator that can catalyze a bond is an **enzyme** and for amino acids it is a **ribosome**.

1.1.2 Structure of Nucleotides and Nucleic Acids

There are 2 types of nucleic acids: DNA (deoxyribosenucleic acid) and RNA (ribosenucleic acid), and as the name suggested, the only different between these 2 is that DNA lack an oxygen group while RNA doesn't. The monomer of both DNA and RNA are *nucleotides* (as mentioned above) that composed of 3 structure: a phosphate group ($-PO_4^{2-}$), a pentose (5C) sugar and a nitrogenous base. In RNA structure, the pentose is called *ribose* and for DNA is *deoxyribose* (no oxygen on OH group on $2'$ carbon). This pentose sugar is connected with the phosphate group by the **phosphodiester bond** at the $5'$ carbon and also connected with the 5 available bases at $1'$ carbon (via *N-glycosidic bond*): *adenine (A)* and *guanine (G)* which both has 2 ring structure and are classified under **purines**; *uracil (U), thymine (T)* and *cytosine (C)* which are single ring and are classified under **pyrimidines**. A, G and C bases can be found in both RNA and DNA structure while T is uniquely DNA and U is uniquely RNA.

Remark 1.5 *In all cases of DNA and RNA, we number the carbon that connect to the bases as* $1'$ *then the carbon right next to it is* $2'$ *and so on till the carbon that connect to the phosphate which is* $5'$. *When nucleotides polymerize, they always start from the* $5'$ *end and add the next nucleotide at the* $3'$ *end*

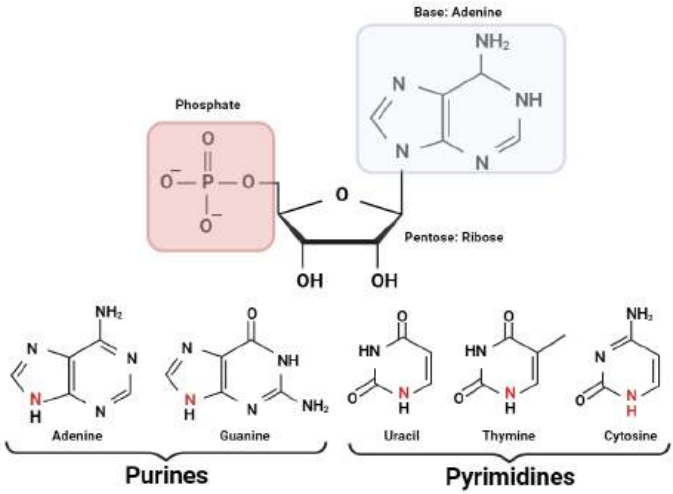


Figure 1.3: The red box is the phosphate group, blue box represent the base that would connect to the pentose sugar. Purines are all bases with 2 ring structure and pyrimidines are 1 ring structure.

Definition 1.4 When only the pentose sugar and the bases are connected without the phosphate group, we call such structure a **nucleoside**.

All nucleosides can become nucleotides via the process of **esterification** which is the formation of ester. In such reaction a covalent bond is formed between phosphoric acid with the $5'$ hydroxyl group on the pentose of the nucleoside. A nucleoside with 1 phosphate group is called *nucleoside monophosphate* and the highest energy monomeric form is with 3 phosphate group thus making **nucleoside triphosphate**. If we want to be specific on the bases, we simply add in the base name e.g. **Adenosine monophosphate (AMP), adenosine triphosphate (ATP), etc.**

INSERT IMAGE

These nucleotides are highly energized thus are good building block in DNA replication and mRNA transcription.

Like proteins, RNA polymerization is made via a template and an enzyme. The template in this case would be DNA, and the enzyme that catalyzes the phosphodiester bond

INSERT IMAGE

DNA Polymerization and Complementary Anti-Parallel Strands

Unlike proteins and RNAs, DNA are formed by not 1 growing chain but 2 growing chains of opposing direction forming a double helix i.e. if we were to line the DNA up on a flat surface, we would see 1 chain grow from left to right and the other from right to left; such peculiarity is called **antiparallel**.

Remark 1.6 *DNA in real life are 3D so there won't be a sense of left or right but just remember that they both would grow from 5' to 3' and antiparallel of each other.*

The 2 strands having pentose-phosphate backbone (exterior) are held together by hydrogen bonding between their respective complementary bases (interior). These complementary bases form the **Watson-Crick base pair** (only some base can be complementary with the other): $A-T$ and $G-C$.

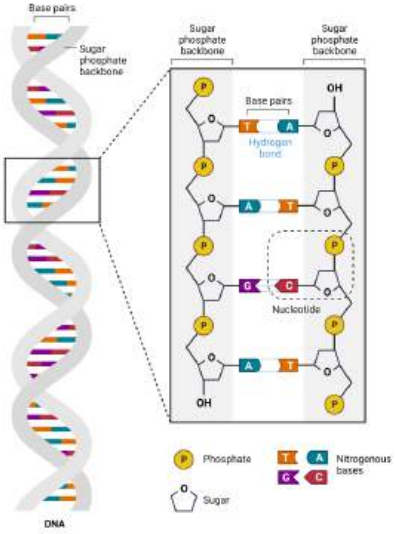


Figure 1.4: Watson-Crick base pairing

There are many conformation to DNA but the main 3 forms that are the

variation of the double helical shape that a DNA can have: A, B and C-DNA. An *A-DNA* is similar to a B-DNA however it is shorter, rare and its complementary bases are not perpendicular to the line of axis of the double helix. *B-DNA* is the most abundant type, it adopts the right-hand helical shape and its complementary bases are perpendicular to the line of axis. *Z-DNA* is another type that adopt the left handed helical shape and have its strands in a zig-zag pattern.

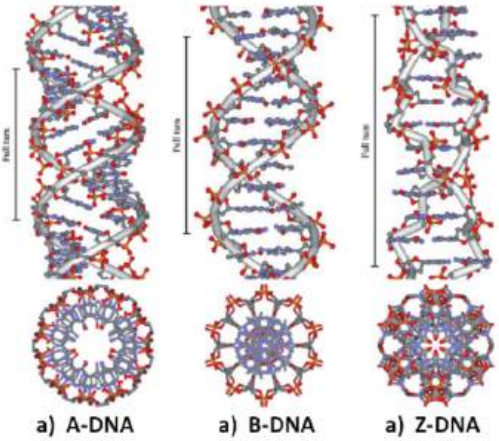


Figure 1.5: Caption

Remark 1.7 *left and right hand helix refers to which direction the helix is revolving around its axis. If it spiral from left and up toward the right it is right handed helix, the contrary would be left handed.*

DNA are held together by weak hydrogen bond. The process of breaking this hydrogen bond is called **denaturation** and the inverse is thus **renaturation**.

Observation 1.2 It is harder to denature $G-C$ base pair than $A-T$ since $G-C$ has 3 H-bonds unlike $A-T$ that only has 1.

Observation 1.3 DNA is in fact very easy to bend, it may be hard when separating the 2 strand but folding it and twisting it is possible (which is also why it can have the double helix shape) and is also important since DNA is a long sequence and would need to be compact for storage.

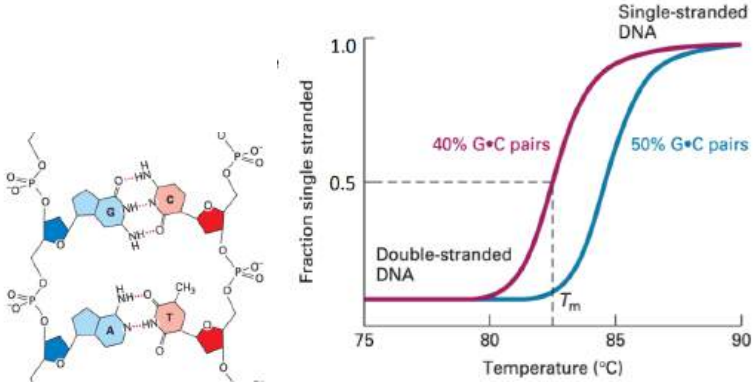


Figure 1.6: The denaturation of G-C compared to A-T

1.2 Overview of Flow of Informations

In general, the main focus of molecular biology is information flow. As we have previously seen on section 1.1, information are stored in biopolymers: DNA, RNA and proteins as a sequence chain. The flow of such information is given as the following

$$\text{DNA} \longrightarrow \text{RNA} \longrightarrow \text{Protein}$$

In the nucleus of a cell, DNA can use itself as a template and with the help of DNA polymerase and dNTPs monomers, it can self-replicate into another DNA (maybe for cell division). For RNA replication, it uses RNA polymerase with rNTPs to read from a DNA template.

Definition 1.5 The process of which RNA is made from reading a DNA template by a RNA polymerase and built with rNTPs monomers is called **transcription**.

Remark 1.7 *RNA polymerase will read the DNA template from 3' to 5' (antiparallel to the template) and synthesize RNA from 5' to 3'.*

Although transcribing RNA is good but it does not have much a use one it leaves cells. However, prior to it leaving the nucleus, it would be matured into mRNA. This mRNA would be used as a template for ribosomes to read and use aminoacyl tRNA (a type of RNA that carry an amino acid) to create

protein (at first polypeptide).

Definition 1.6 The process of which protein is made by ribosome using aminoacyl tRNA and read from an mRNA template is called **translation**.

1.2.1 DNA Replication

To perform DNA replication, as said above, it uses itself as a template. During the process, DNA template will be locally unwound, exposes the initiation site. The 2 strands will be separated and DNA polymerase will bind to the initiation site. It will then use dNTPs (include dATP, dGTP, dCTP and dTTP) monomers to create a direct base pairing with the template but also linking monomers together via polymerization. During polymerization of the monomers, the 3'-OH group of the previous monomer will attack the α-phosphate of the incoming monomer and the β,γ-phosphate is disconnected. In theory, DNA replication does not have an end site thus no ending.

End Result: formation of 2 new DNA, each has 1 original DNA strand (the template) and 1 new DNA strand (synthesized).

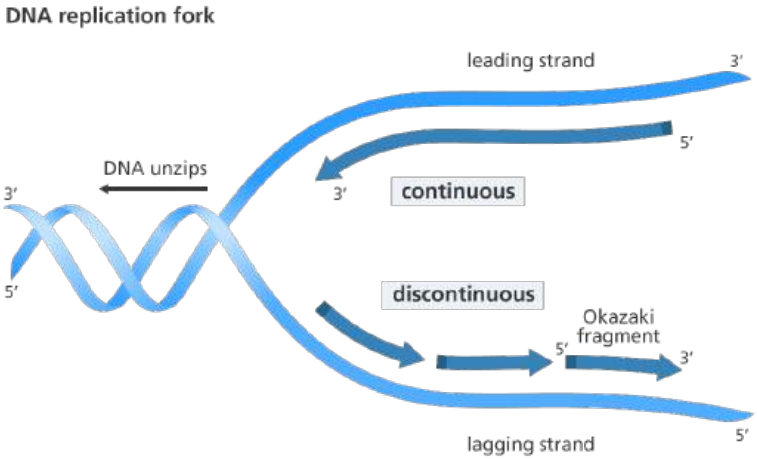


Figure 1.7: DNA replication

1.2.2 Transcription and Translation

In layman term, the process of transcription would rewrite the DNA in a different nucleotide and the process translation, we're changing the those nucleotide into amino acids. Think of a language, transcription is simply writing DNA in a different font (like cursive to block form) and translation is writing DNA is a different language (like from English to French).

In transcription, RNA polymerase will bind to the initiation site (called **promoter**) separating the template and the non-template strand and start reading. It will then use rNTPs monomers to pair with the template strand and also link with the next monomers. Like DNA replication, β, γ-phosphate will be disconnected during polymerization. The newly formed RNA will be displaced out of the RNA polymerase when the non-template reform its H-bond with the template strand. Transcription will end when it reaches the stop site and RNA will leave the DNA.

End Result: 1 strand of unmatured RNA (separated from the DNA template).

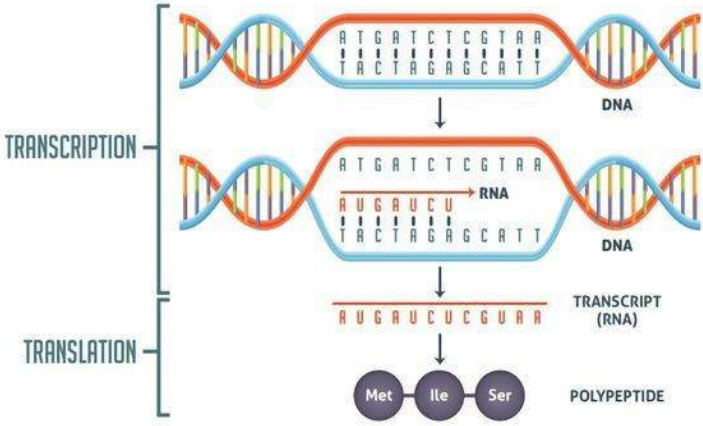


Figure 1.8: Transcription and translation.

When RNA is modified and matured it becomes mRNA and will leave the nucleus. This mRNA will then be used for the process of translation.

In translation, mRNA will be bound by ribosomes that would brings in aminoacyl tRNA (acts like an adaptor between nucleotides and amino acid) that carry amino acid monomers in the form of high-energy amino acyl-tRNA esters. Inside the ribosome, peptidyl transferase reaction will be catalyzed by the large rRNA (large subunit of the ribosome) hence we would call it ribozyme.

End Result: 1 long chain of amino acids which would later fold into proteins.

1.3 Extra Stuff Added

Because of the DNA non-symmetric double helix shape, it would form **major** and **minor grooves**. The purpose of these grooves are for binding protein to recognize a specific DNA sequence.

Chapter 2

Protein Structures and Functions

Proteins and polypeptides are 2 almost synonymous words however, for our definition, proteins are more complex structure of 1 or more polypeptides while polypeptide is a long chain of amino acids join together by peptide bond. Hence we can say that all proteins are polypeptides but not the opposite.

INSERT IMAGE HERE

On average, proteins range from having 300−400 amino acids; the smallest protein known is approximately 40 amino acids in length while the largest (called *titin*) is 30000 amino acids in length.

2.1 Hierarchy of Protein Structures

When amino acids first arranged into linear 1 dimensional sequence (regardless of shape or size) and this structure is called **primary structure**. Proteins will have more functions as it folds and twists into a more 3D shape.

2.1.1 Secondary, Tertiary and Quaternary Structure

This long chain of amino acids can fold onto itself to create the **secondary structure**, its shape is held together by hydrogen bonding. We also call it conformation of the peptide backbone since the main thing we're focusing on is the backbone and not the $R-$group. Even then, the folding process is also dependent on the types of amino acid.

Definition 2.1. The tendency of non-polar (hydrophobic) molecules coalesce/aggegrate together is called the **hydrophobic effect**.

Example 2.1.1. After translation, some of the amino acids monomer on the polypeptide would be hydrophobic which mean they would coalesce together. Indeed this is the case as proteins have be found having their hydrophilic amino acid in the exterior while the hydrophobic ones coalesce in the interior.

INSERT IMAGE HERE

Remark 2.1. *An amino acid residue an older way of saying amino acid monomer, they're interchangeable.*

With that being said, this simple concept underlies the entire way of protein folding; it would dictate an endless possibility of shapes that protein can adopt. Nevertheless, each protein would only adopt 1 or a number of similar structure called **local conformations.** For the secondary structure, there are mainly 2 local conformations which make up of around 60% of the average polypeptide length (the rest are non-conforming chain or random folding): α-helix and β-sheet. Both of these structure is held in shape by hydrogen bonding between carbonyl oxygen of 1 amino acid residue bind to the amino group hydrogen of the other amino acid residue.

α-Helix Conformation

α-helix conformation is characterized by its single helical shape and H-bonding periodicity i.e. we know where the hydrogen bonding is between 2 amino acid residues. It is observed that for an n^{th} amino acid residue, it would form a hydrogen bond (of 2 atom mentioned in highlighted section above) with the amino acid residue in the $n+4^{th}$ position

$$n \longleftrightarrow n+4$$

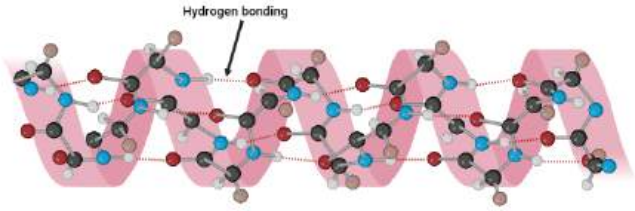


Figure 2.1: Alpha helix conformation

As we can see though, the H-bond is not perfectly aligned to the helix axis which allow it to have 4 residues. If it was perfectly aligned and parallel, then the periodicity would be roughly 3.6. Furthermore, because there are so much of these H-bond, it gives a gross structure of a straight rod i.e. it is hard to bend it.

β-Sheet Conformation

β-sheet conformation is characterized by many β-strands (seesaw/zigzag-like polypeptide chain) connected laterally by H-bonding. One can visualize the total geometry of it as a multi-folding piece of paper. Each of the

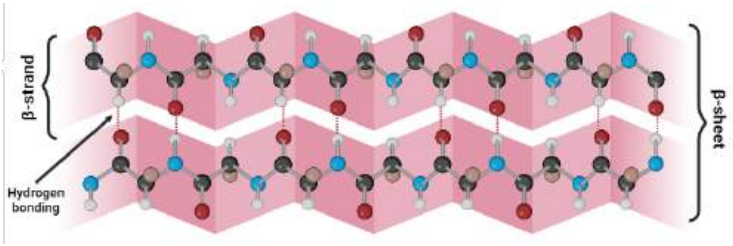


Figure 2.2: Top view of beta sheet conformation

β-strand can run parallel or antiparallel to each other. Figure 2.2 is in fact the top view of the β-sheet, if we were to turn it sideway, we would see the R-group pointing up and down periodically. The protruding R-group would later determine the interactions of the β-sheet with other part of the protein.

Ribbon Diagram

Another aspect we will need to look at is **ribbon diagram** which is the most used way to represent secondary and tertiary structure. In this diagram, α-helix is shown as a helical ribbon while β-sheet are β-strands represent as a long flat arrow (correspond to the directionality) chaining together by unspecified amino acids residue chain.

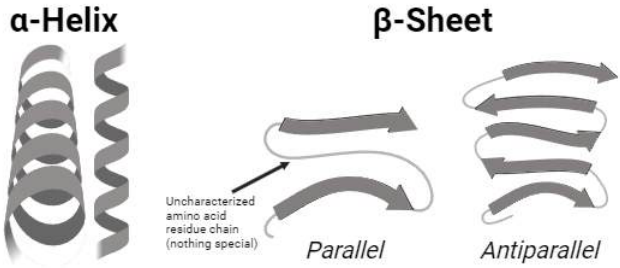


Figure 2.3: Ribbon Diagram

This type of diagram exclude the R-group.

Tertiary structure is the total conformation of the polypeptide; that is when the α-helix, β-sheet and other randomized chain starts to interact and form bonds with each other i.e. it is the organization of multiple secondary structure conformation. They interact via R-group which again from ribbon diagram won't be shown. When a bond is form on a particular section so that when isolated that particular section, it can still form it shape; that section is called a **domain**.

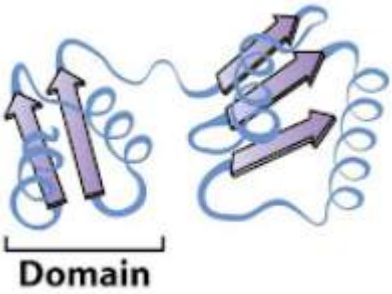


Figure 2.4: Tertiary structure and a domain example

Most of the structure we've talked about are held together by non-covalent bond e.g. hydrogen bond, Van der Waals interaction, etc. With tertiary structure however, we get to have some covalent bond such as the *disulfide bridge* (review observation 1.1)

When many secondary structures combine together to form a distinct 3D structure, we call that a **motifs**.

Example 2.1.2. A **coiled coil motif** have 2 α-helix coil around each other. In the interior of this coil, we can find repeating pattern of hydrophobic R-group at around every 3.5 amino acid residue.

Example 2.1.3. An **EFhand or helix-loop-helix motifs** have 2 perpendicular α-helix connect by a loop with a characterized sequence of R-group that serve as a sort of binding site to a Ca^{2+} ion. The proteins shown is called **calmodulin**.

Example 2.1.4. A **Zinc-finger motifs** have an α-helix and a β-sheet connect with each other. On these to structure, there exists R-group such as cysteine and histidine that serve as binding site for Zinc.

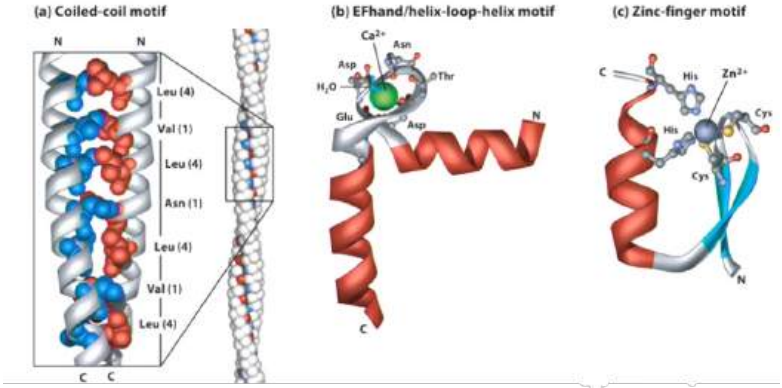


Figure 2.5: Motifs as described by the 3 previous examples

The main difference between motifs and domain is that motifs is not as stable as compared to a domain. What we mean by that is...theoretically if we were to cut out a motifs and drop it in water, it will fall apart as compared to a domain that would stay put. These newly protein shape (mainly domain) has 4 classes (in term of their shape): fibrous, globular, transmembrane and intrinsically disordered proteins (i.e. they have no particular

shape until they interact with another protein with a definite shape)

We could technically stop at tertiary structure and its form would be protein but let's take our final step to the final structure. The **quaternary structure** or **multimeric protein** is when multiple different tertiary structure (or simply polypeptide) comes together.

Example 2.1.5. Influenza virus have hemagglutinin which are proteins that can bind to RBCs' receptor to initiate viral attachment. When looking at the secondary structure of hemagglutinin, we can see a glubolar (round shape) and fibrous domain called HA2 interact with a subunit called HA1. This quaternary structure alone form a *heterodimer* (dimer = 2 monomer

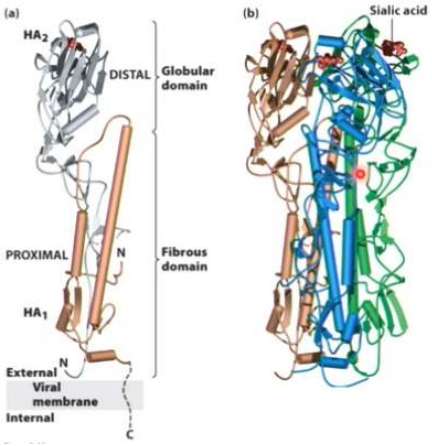


Figure 2.6: Tertiary (a) and quaternary (b) structure of hemagglutinin in influenza virus.

= 2 subunits [HA1, HA2], hetero = different), the dimer structure however does not functions alone. When 3 hetero dimers are connected via non-covalent bond, they form a trimer which is a larger quaternary structural form of hemagglutinin.

A higher level than quaternary structure is called **supramolecular complexes** where their molecular weight can exceed $1MDa$ (an average protein molecular weight is $12.3-450kDa$). The supramolecular complex consists of multiple distinct protein which in itself may contain multiple subunit. Nevertheless, unless it is activated, the supramolecular complex will not have a rigid form, which is unlike quaternary structure that comes together rigidly.

2.2 Proteins Folding, Misfolding and Degredation

Ideally, if a protein have hydrophobic amino acid side chains, those must be located in the interior during protein folding. If patches of hydrophobic side chains exists at the surface of the protein final form then there's a potential misfolding.

Definition 2.2. A **misfolding protein** is a result when protein follow the wrong folding pathway (could be during the secondary or tertiary structure). This could potentially caused by a change in external environment.

As we know from before, the sequencing of proteins would determine its overall conformation. When proteins are first synthesized by rRNA, it forms a long linear unfolded polypeptide call **nascent polypeptide**. When the nascent polypeptide are folded correctly into a specific shape, we call it the **native state/conformation** (the “correct" state). When proteins are misfolded or denatured, they can either be degraded by the cell, or can adopt **toxic conformation** which has wrong functional uses, insoluble and tend to form long linear or fibrillar aggregates called **amyloid deposists**.

Sometimes however, we could observed denatured proteins to spontaneously refold itself to a native conformation. This spontaneity of refolding is though to be the “folding pathway". Furthermore, one thing to look at when studying about protein is that the N-terminus is synthesize before the C terminus which result in the folding of protein before the C-terminus is made.

2.2.1 Protein Folding and Chaperones

Although some proteins, such as ribonuclease, can fold all by themselves. The majority of proteins would required assistance to get to their native conformation or refold from a misfold conformation. Through evolution, the cell has develop these essential proteins that can help with this job and they are *chaperones*.

Definition 2.3. **Chaperones** are proteins that guide protein folding by letting partially misfolded proteins to return to its folding pathway. Not only that it can also disassemble toxic conformations that can cause aggregates or protein complexes. Finally, it can acts as a mediate transformer between inactive and active proteins.

In order guide partially misfolded protein back to its original folding pathway, chaperones must have the ability to recognized exposed hydrophobic patches. The number and the production of chaperones is also proportional to the condition that cause misfolding i.e. chaperones are upregulated under conditions that would increase the chance of protein misfolding e.g. heat shock, stress, etc.

Definition 2.4. Proteins that are being guided by chaperones are called **client protein**.

Chaperones use cycles of ATP hydrolysis to work through the misfolded molecules at the exposed hydrophobic patches. By blocking these patches, they keep the (re)folding proteins away from the misfolded section. There are mainly 2 classes of chaperones: molecular chaperones and chaperonins.

Definition 2.5. **Molecular chaperones (mainly Hsp70)** are proteins that can help misfolded or nascent polypeptide to return to the correct folding pathway.

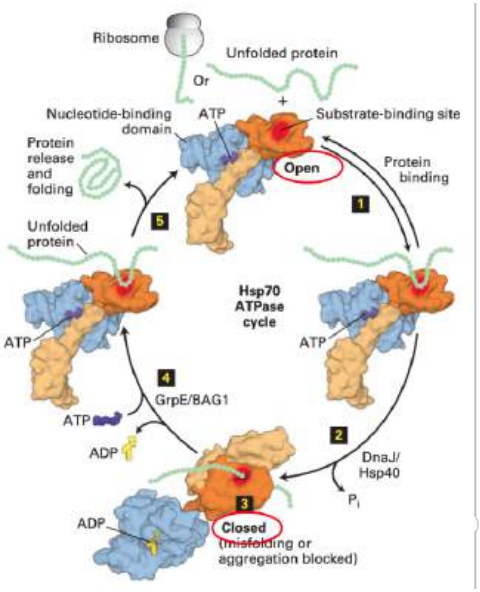


Figure 2.7: The cycle of Hsp70 in protein refolding and control

The main type of molecular chaperones are Hsp70 stands for “heat shock protein 70" thus they are upregulated during high temperature. They function mainly by binding to the exposed hydrophobic residues of the nascent polypeptide, protecting it from aggregation and thus allowing the entire polypeptide folds properly.These chaperones do their jobs through a cycle with their client protein by binding and conformational changing that use ATP.

Although technically a class on its one, it is in fact belong to the superclass *molecular chaperones* classification, this is the chaperonins.

Definition 2.6. **Chaperonins (Hsp60)** are a class of molecular chaperones that has a molecular mass of around $60kDa$.

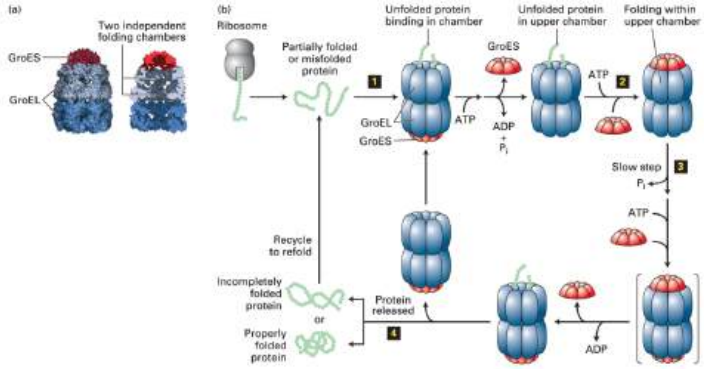


Figure 2.8: The cycle of Hsp60 in protein refolding and control

Like Hsp70, Hsp60 also prevent misfolding of protein during stressful situation such as high heat. They mainly form enclosed chamber made up of protein binding subunits that face to the chamber's interior. They also goes through a cycle with their client proteins through binding and conformational changes that use ATP. We can further classify Hsp60 into 2 classfication: Group I and II

- **Group I** are Hsp60 that functions in bacteria or organelles that have endosymbiotic origin e.g. mitochondria and chloroplast.

- **Group II** are Hsp60 that functions in eukaryotic cells

2.2.2 Ubiquitin/Proteasome Degradation System

Although chaperones can help with refold misfolded proteins, sometimes there are misfolded proteins that is irretrievable. In such case, these proteins must be destroyed before aggregating or produce any bad functions. There are 2 parts to this degradation system: ubiquitin labeling and proteasome cleavage.

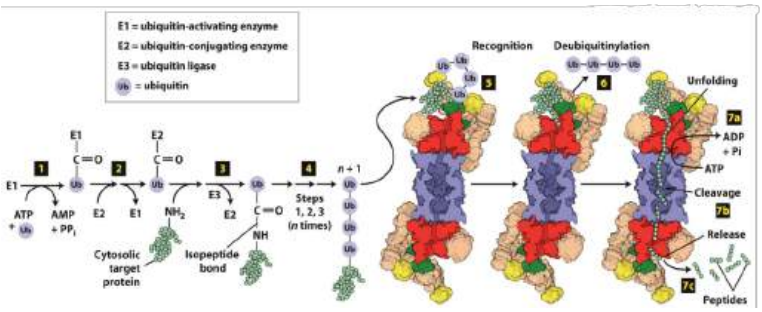


Figure 2.9: Process of protein degradation using ubiquitin and proteasome.

Ubiquitin Labeling

Firstly, ubiquitin-activative enzyme (E1) will use ATP to bind with ubiquitin. E1 will then be swapped with ubiquitin-conjugating enzyme (E2). E2

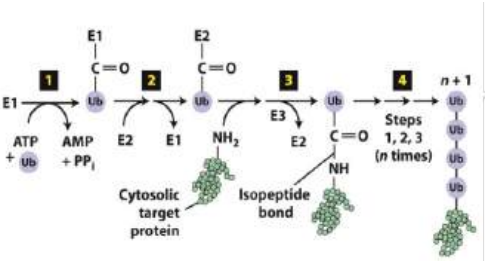


Figure 2.10: The process of ubiquitin labelling

then swap with ubiquitin ligase E3; E3 is especially important since it can target misfolded proteins, exposed hydrophic patches and even oxidized amino acids. A failure in making E3 can be detrimental.

Definition 2.7. **Ubiquitin** is a 76 monomeric protein that can covalently bond with *lysines monomer* on the misfolded protein.

Nevertheless, after E3 is bound to ubiquitin, the process will restart till there is a long connection of ubiquitin with the misfolded protein. After the misfolded protein is labelled with ubiquitin, it will be brought to proteasome to be destroyed.

Proteasome Cleavage

Once arrived, proteins from the 19S cap recognized and bind with the polyubiquitin. Using hydrolysis, it removes the polyubiquitin while keeping the misfolded protein, this process is called **deubiquitinylation**. It then use ATP to unfold the protein and feed it down to the 20S core protein. The

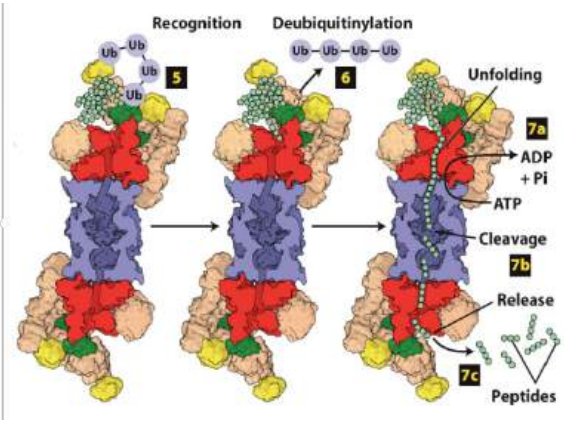


Figure 2.11: Process of proteasome degradation of proteins

20S core consists of protease subunits (enzymes that breakdown proteins) facing inward to the core. This design is to isolate the dangerous protease from the cytoplasm (otherwise it could breakdown other important protein). When the protein is being fed down into the chamber, the protease will degrade it by cleaving the peptide bond between monomeric units into amino acids or **oligopeptides** (polypeptides with small amount of amino acids monomer). These degraded residue will then be fed through the exit 19S cap to be release back to the cytoplasm.

2.2.3 Quality Control Failure

Although misfolded protein can be first go through control by chaperones, then if all fail it would be degraded by proteasomes; this process is not perfect. When this system along with its 2 controls fails...misfolded proteins will aggregate nevertheless this process would take a long time. Like we've said before, the aggregation of these proteins would form an *amyloid* which is an important aspect for some **neurodegenerative diseases** such as Parkinson's, Alzheimer's and "mad cow" disease. Misfolded proteins follow

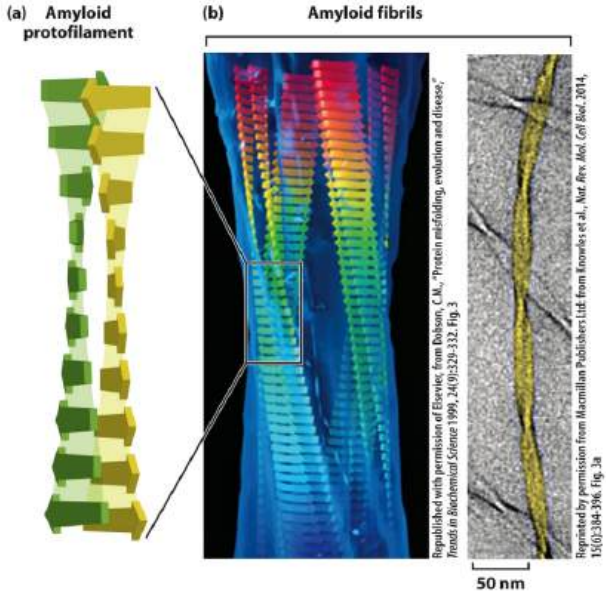


Figure 2.12: Amyloid protofilaments

the same structure as native ones, that is they'd have α-helix and β-sheet then these structure aggregate into long filaments called **amyloid protofilament**. These long filaments are resistant to proteolysis (protein degradation by enzymes).

Observation 2.1 Amyloid deposits, shown as plaques and tangles, can be seen in the brain tissues from patients that have Alzheimer's disease under microscopes.

2.3 Proteins Functions and Regulations

Proteins can have a variety of functions but the most common for all of them is **binding**. Proteins can bind to themselves, larger and molecules, and even ions. These molecules that portein bind to is called **ligand** and when bounded they're called the *ligand-protein complex*.

$$\text{Protein} + \text{ligand} \longrightarrow \text{ligand-protein complex}$$

The binding of proteins and its ligand is depending on 2 main factor: *specificity and affinity.*

Definition 2.8. **specificity** defines a binding site that is specific for a type of ligands even when there is a high amount of other ligands present.

Definition 2.9. **affinity** defines the strength of the binding site i.e. how strong the connection between the protein is with its ligand. Affinity is usually expressed as the constant of dissociation K_d which means that if $K_d \uparrow$ the interaction is weakened (since the ligand-protein complex can dissociate into 2 easily).

Binding is simply the interaction/bonding at the molecular suffaces. The specificity of protein binding comes from the high amount of weak interaction between the 2 surfaces but if there's many interaction, it'll become strong.

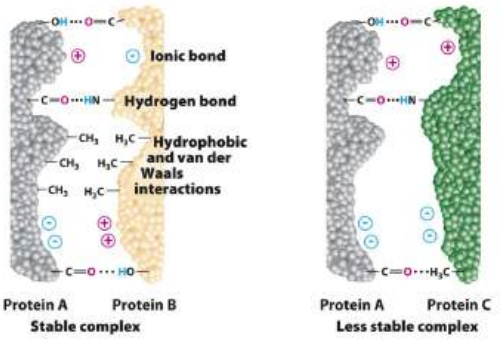


Figure 2.13: Binding of 2 different proteins at protein A. Protein A is more specific in term of binding to B than C.

Example 2.3.1. **Antibodies**, are proteins that protect your body from foreign substances called **antigens**, can bind to antigens with high specificity

and affinity. To know why, we would have to look at the structure of them. Antibodies are made of 2 pair of identical heavy chains and identical light chains. One important region of the antibodies is called **antigen-binding**

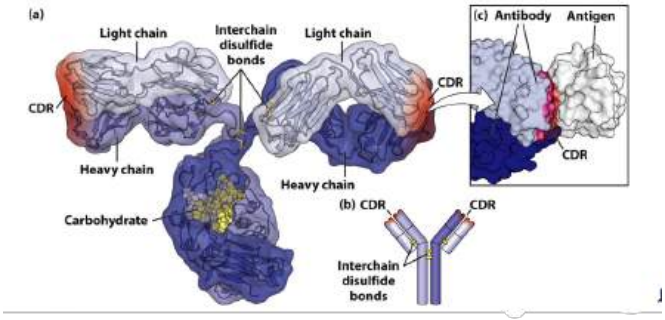


Figure 2.14: Structure of antibodies and CDR region

surface or **complemetary determining region (CDR)** that can bind strongly to antigens. The CDR is made from many proteins loops from the heavy and light chain; each loops are highly variable in amino acids, due to gene encoding, which makes the combination of amino acids that could bind to the antigen tremendously high.

2.3.1 Enzymes

An important (arguably the most essential) and most abundant class of protein are *enzymes*

Definition 2.10. **Enzymes** are catalytically active proteins, proteins used to speed up a chemical process. The ligands for enzymes includes the substrate for the chemical reaction that they catalyze.

Enzymes does not breakdown after it catalyze a reaction, the substrates that it catalyes may change but itself is not. The general mechanism of enzyme catalysis is as follows

$$\text{substrates} + \text{Enzyme} \longrightarrow \text{S-E complex} \xrightarrow{\text{catalysis}} \text{Enzymes} + \text{products}$$

As you can see, enzyme would stay the same at the beginning and end of a reaction but the substrates may not.
Substrates can bind to an enzyme at its **active site**. This site can include

the **substrate binding site** which is where substrate specificity will come, and **catalytic sites** which are amino acids monomers that can catalyze the reaction at the binding site.

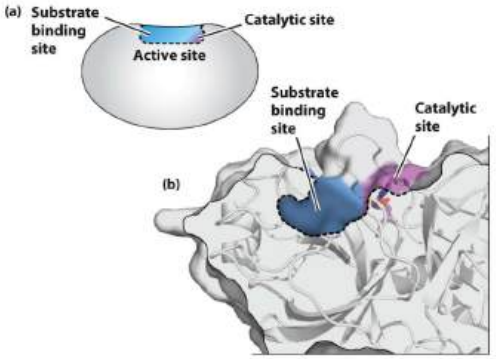


Figure 2.15: Active site of most enzymes are divided into the substrate binding site and catalytic site.

Example 2.3.2. **Proteases** are enzymes that hydrolyze peptide bonds in the polypeptides. **serine protease** is a large family of proteases whose catalytic mechanism involves serine monomer in the catalytic site. We're now going to focus on 1 types of serine protease called **serine protease trypsin.**

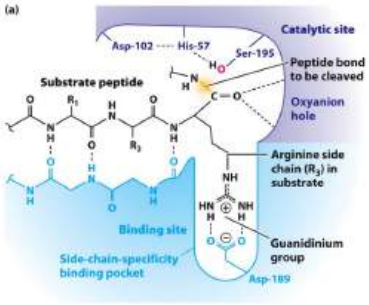


Figure 2.16: Theoretical cross section of serine protease trypsin. The trypsin binding site made from the 189th amino acid monomer in the structure called asparagine (Asp-189).

The reason that it has "trypsin" is because within the substrate binding site, there is a site called *trypsin* that has a negative charge which would allow correspond protein with a positive charge residue to be held in place

for catalysis on the serine site. After binding to trypsin, the cataytic site with serine will begin catalysis.

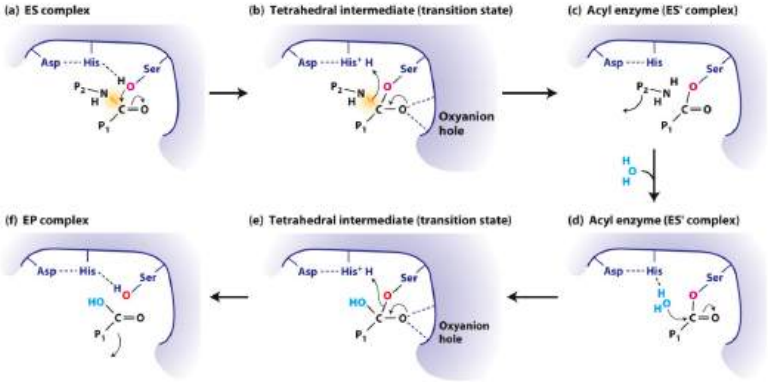


Figure 2.17: Total mechanism of catalysis by serine protease on a polypeptide.

In general, serine will cleave the peptide bond (separating the polypeptide into 2 chain) with the formation of the substrate-enzyme complex to itself (1 of the 2 chain will stay because of this covalent bond). Then the complex will be hydrolyzed to break apart the substrate enzyme complex thus releasing the remainding polypeptide and restoring the native state to the catalytic site.

An important thing to know about enzymes is their rate of catalysis. The study of the rates of enzymed catalyzed reactions is called **enzyme kinetics**; the study can range from the most complicated catalytic reactions to the simplest, for our sake, we would only need to know briefly about the simplest case. The simplest case of enzyme kinetics is called **Michaelis-Menten Kinetics**. The equation that describe the Michaelis-Menten kinetics model is

$$v = \frac{dp}{dt} = \frac{V_{\max}}{K_m + a} \tag{2.1}$$

where v is the rate of reaction that turn and product P with concentration p into a product A with concentration a. $V_{\max}$ is called the **limiting rate** which is the maximum catalyzing rate of enzymes once the concentration reached saturation (no more enzymes can accept substrate). K_m is the substrate concentration such that it supports the rate of reaction half of $V_{\max}$. We can plot equation (2.1) to give the following curve

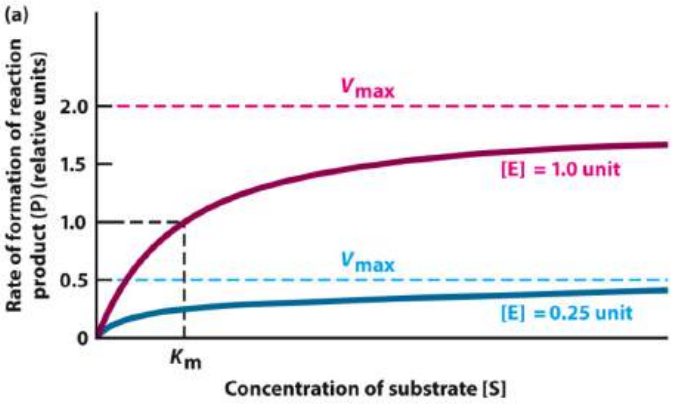


Figure 2.18: Michealis-Menten Kinetics curves.

Remark 2.2. *The following curve has 2 curve in it, the reason for the 2 curves is to show that if there was an increase/decrease in the number of enzymes the maximum rate of reaction would increase. However, their K_m should be the same.*

One interesting thing about enzyme is that it requires an optimum pH to functions to its highest effectiveness. The pH that it functions best is determined by its active site acid-base chemistry, protein conformation and charge distribution. In fact, serine protease trypsin requires optimal pH of around 7 to functions fully. If the pH is lower than the optimum, the enzyme will not be as effective; and if the pH is higher than the optimum, the enzyme structure and bond would be denature thus making it also less effective.

Example 2.3.3. Chymotrypsin is a enzyme that is similar to trypsin that require an acid-base reaction between the 2 active site. If the pH is lower than 7, the reaction cannot occur however if pH is higher than 9 the molecular structure of chymotrypsin will be denatured.

Enzymes do not only catalyze a products but these products will be used as substrate for another enzyme to catalyze into other products etc. These interconnected enzymes form a **pathway**, for a biological system, the largest pathway is the **metabolic pathway**. As time went on however, the body realized that having these enzymes in separate compartment of

the body would be difficult for substrates to navigate. The best case is to have these enzymes sits side by side to each other or in a **scaffolding**.

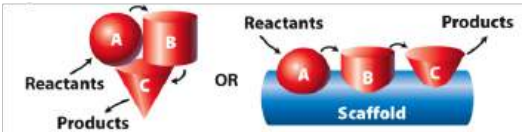


Figure 2.19: Enzymes can sits next to each other or in a scaffolding.

Through further evolutionary changes, these enzymes are now altogether a single unit of proteins i.e. each of the section may perform the enzymatic activities of different enzymes yet they're part of the same proteins (only 1 mRNA to create this complex!).

Figure 2.20: Each of the A, B, C sections domains of the whole protein unit.

Allosteric Effects

Allosteric effect is when the binding of a ligand on 1 side of a protein can lead to a conformational change that block a binding site from another side of that same protein.

Example 2.3.4. We can look back at Hsp70, when there is ATP bound to it, it has the "open" conformation that allow misfolded protein to enter, this binding also prevent ADP from binding. When ATP removed and ADP starts binding, which change the Hsp70 conformation to "closed" and also the ADP binding prevent the binding of ATP v.v.

A common type of allosteric effects that mainly used to control protein activity is the binding of Ca^{2+} and GTP.

Example 2.3.5. **Calmodulin** is a type of protein that can bind to Ca^{2+} . Once Ca^{2+} bind to the EF-hand of calmodulin, its conformation will be changed allowing it to target other proteins in order to change its structure and functions.

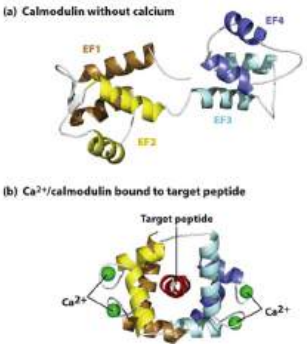


Figure 2.21: Calmodulin with and without calcium

Example 2.3.6. **GTPases** are a type of hydrolase enzyme that works with GTP and GDP. When GTP is bind to GTPase, it will be adopt the "on"/active conformation and begin to hydrolyze GTP; the process of GTP hydrolysis is sped up by **GTPase-activating proteins (GAP)**. Once GTP is hydrolyzed to GPD, release a phosphate in the process, which bind to GTPase, the enzyme adopt the "off"/inactive conformation. GDP however must the release from GTPase, and this process is aided by a protein called **guanine nucleotide exchange factor (GEF)**; GEF help GTPase release GDP and allow GTP to bind to GTPase returning it to its active conformation and v.v.

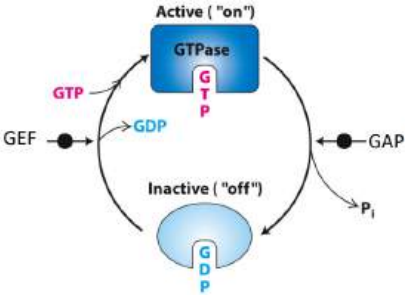


Figure 2.22: Mechanism of GTPase

2.3.2 Phosphorylation and Dephosphosphorylation

Sometimes, proteins need a certain conformational change that would result in its functions changing. There are many way to do this but the most

used way is *phosphorylation*

Definition 2.11. **phosphorylation** is the process of adding a phosphate group PO_3^- to any molecules and even ions. The reversal process, removing a phosphate group, is called **dephosphorylation**.

Phosphorylation are typically done by **protein kinase** to side chains of the protein (this process, although made from covalent bond, is completely reversible) and happens right after **translation**. Human genome encodes more than 500 variation of protein kinase and each is for a specific protein.

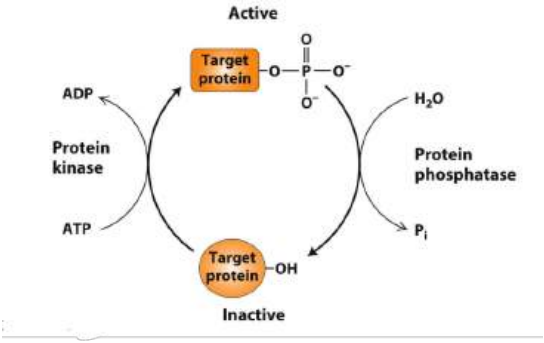


Figure 2.23: Caption

Remark 2.3. *One of the probable reason of phosphorylation is that phosphate groups are negatively charged so by introducing it to the protein, this would change its conformation which later on its functions.*

Often than not, we would encounter kinases or phosphotases of which their targets are not proteins but other kinases or phosphotases too. This would cause a cascade which allow for signal amplification.

2.4 Protein Analysis I

We can characterize proteins in many ways: functions, density, shape, charge etc. There are many ways to characterize these proteins, we begins with the simplest.

2.4.1 Centrifugation

Centrifugation is a method to separating particles by their shape and size however before getting into centrifugation, we need to establish to foundation that lay the method

Definition 2.12. According to Newtonian mechanics, centrifugal force is the force that act on object in the rotating frame of reference. It's a *pseudo*-force that seem to push the object experiencing it out of the rotating frame of reference.

Remark 2.4. *This force is measured with respect to earth gravity (1g) e.g. if the centrifugal force create an effect on an object experiencing twice the earth graviational pull, the object is experiencing 2g.*

Essential when a tube full of a particles (suspended in a liquid medium) is spinning rapidly, the centrifugal force will pull on that object. If the particle is denser than the medium, the centrifugal force will pull it toward to the bottom of the tube (away from the spinning axis). The reverse would occur if the particles are less dense and nothing would happen if the particle is as dense as the medium.

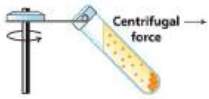


Figure 2.24: Simple centrifugation.

The majority of the time, the particles of interest is denser than its medium. When the particles moves down to the bottom of the tube, it forms a **pellet** and the liquid above it is called **supernatant**.
The rate at which the particles sediment to the bottom from the supernatant depending on the size and shape of the particle. In general, the larger to particles is, the longer the rate of sedimentation, this indirect relationship is calculated using the **Svedberg (S)** unit which is $10^{-23}s$. This unit quantify the size of particles by relating it to its rate of sedimentation.

Example 2.4.1. The rate of sedimentation of a particle is $20 \times 10^{-23}s$ would have size of $23S$. Therefore for this particle would have a smaller size than a particle of $30S$.

Nevertheless, this unit has lots of variant (not just size but centrifugal speed, shape, density of solution etc.) so be mindful when using it.

Remark 2.5. *It is important to realize that size and mass is a little different from each other. A particle can have a large size but small mass (compared to other particles) while another would be opposite of any of the combination.*

Differential Centrifugation

A common technique used in centrifugation is called **differential centrifugation** where a tube of substances is ran through multiple round of centrifugation of different speed and at each speed, variety of pellet will be separated from the supernatant.

Example 2.4.2. We can apply differential centrifugation on a cell. We begin by lysing a group of homogeneous cells and place it in a centrifugal tube. We then start centrifuging the tube at low g; all the organelles with higher mass would form a pellet such as the nucleus leaving organelles such as mitochondria in supernatant. The pellet is removed and centrifugation at higher g will begin; this time, mitochondria will form a pellet. This process will continue will all required particles are removed from the supernatant.

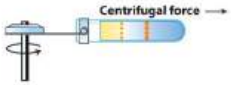


Figure 2.25: Differential Centrifugation

Differential centrifugation is a contrasting with another technique of centrifugation called *equilibrium density gradient.*

Equilibrium Density Gradient Centrifugation

In this type centrifugation, the tube is will be filled with sucrose of different concentration (pipette in to avoid mixture), then the tube will be centrifuged. What will be the result in a density gradient sucrose tube. Using that same tube, we can put proteins pellet and start centrifugation at high speed. The proteins will begin to move down until it stop at a point, and that point is where the protein density will match with the point on the tube's density gradient.

2.4.2 Eletrophoresis

Electrophoresis is a separation method electric charges to separate proteins by *charge:mass ratio.* We first prepare a substance called **agarose**

gel and pour in into a holder. We then create indentation into the gel where we can put our sample proteins in. A electric gradient will be induce through the gel body. The direction of protein migration is depending on its net charge, while the speed that it goes through is depending on its charge:mass ration and the length of migration will be depedning on the size, i.e.

Figure 2.26: Bands on agarose gel electrophoresis.

large protein will be impeded while smaller ones can goes through (due to the molecular structure of agarose gel).

Remark 2.6. *The factor of direction of migration, speed and distance during electrophoresis is a sample's* ***electrophoretic mobility****.*

SDS-PAGE

A biggest breakthough in gel electrophoresis and is the discovery of an anionic detergen called **sodium docecyl sulfate (SDS)**. This detergent was discovered in the 1960s, and is still used in every lab.
SDS denatures proteins by binding to its hydrophobic side chains and tails forming a coat of SDS, thereby disrupting the oil-drop structure of it (hyrophobic inside, the rest outside). Not only that, SDS is a bit negatively charged so after binding to the polypoptide body and straighten it out, it prevents the polypeptide from refolding (due to the negative charge of SDS would repel each other), Furthermore, it also break complex multimer proteins structures to each individual polypeptide therefore better for experimentation.

This allow us to perform a new electrophoresis technique called **SDS-PAGE** (stands for Sodium docecyl sulfate - polyacrylamide gel electrophoresis). This technique use a protein sample that was treated with SDS, which means that during the electrophoresis, its shape has no effect (SDS dena-

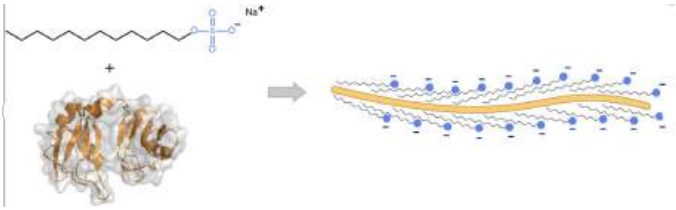


Figure 2.27: SDS binding and forming a coat around a polypeptide

tured) and its all uniformly charged with negative (due to SDS negativity). This means that the electrophoretic mobility of all protein would be the same. Using polyacrylamide (instead of agarose) also allow smaller molecules (short polypeptide) to pass through while impeding the larger ones.

Remark 2.7. *Some post-translational modification can change the electrophoretic mobility of proteins.*

Example 2.4.3. Performing SDS-PAGE on phosphorylated and unphosphorylated phospholamban would result in phosphorylated phospholamban having a higher molecular weight reading than unphosphorylated.

Isoelectric Focusing

When the $COOH$ group of the amino acid is subjected under basic environment it will become negatively charge while NH_3 group would become positively charge under acidic environment. This means that a group of amino acid will become negatively charge in acidic environment and as we increase the pH (making it more basic), all of them will become positively charge
If all of them can go from negative to positive there must be a point where the net charge is 0, and that point is the isoelectric point.

Definition 2.13. **Isoelectric point (pI)** is the pH at which the sum of charges on an amino acid/protein is 0.

Remark 2.8. *pI depends on the amino acid composition of that protein.*

This allows us to use a technique called **isoelectric focusing (IEF)**. First we create a pH gradient using a special buffer (an ampholytes) immobilized in acrylamide gel. We then mix protein into this pH gradient and subjected it to an electric field. After the electric field is passed through, all of

the proteins would have moved to their respective pI which form narrow strips on the in the gel (each strips is a pI). This was perform in "1 dimension" (along a line).

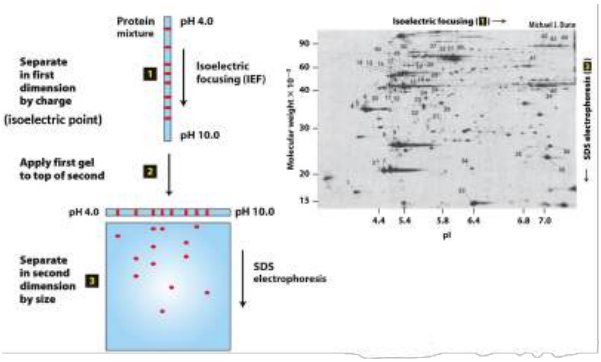


Figure 2.28: Process of IEF and electrophoresis

We then use this same acrylamide gel with its proteins (separated by pI) and connect it with another acrylamide gel (without pH gradient) and perform electrophoresis. We use IEF because there is no correspondance between pI and molecular weight so at the end, we would get simultaneously a wide range of different proteins.

2.4.3 Mass Spectrometry

Mass spectrometry (MS) is a technique that can obtain high precision determination of charge:mass ratio of ionized molecules. Unlike the centrifugation and electrophoresis, mass spectrometry is analytical not preparatory. There are 3 concepts that mass spectrometry follow: 1) create ionized gaseous molecules 2) measure the acceleration of ions in an electric/magnetic field and 3) acceleration of these molecules depends on its mass:charge ratio (m/z).

Remark 2.9. *Each amino acids and oligopeptide would have its own molecular weight. If the amino acid has a charge of* ± 1 *then its m/z is equal to its molecular weight (MW)*

For the first part of ionizing the gas, we can use a technique called **electrospray ionization**. The ions from the electrospray will be separated into

population differing in m/z via the mas spectrometry (we will not get into the mechanism since it's out of the scope of this course).

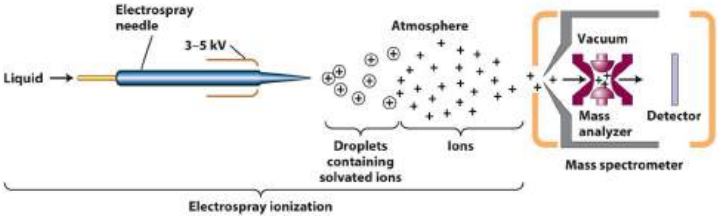


Figure 2.29: Electrospray ionization in mass spectrometry

And at the end, we'll have an unuasual graph where the $x-$axis represent m/z while the $y-$axis represents the relative abundance of each of those ions populations.

Example 2.4.4. A mixture of 3 major peptides from mouse H-2 class I histocompatibility antigen is ionized and analyze by the mass spectrometer, the graph yield from the analysis is

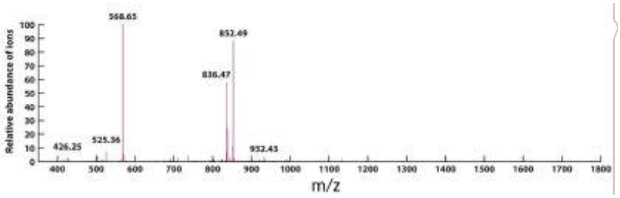


Figure 2.30: Abundance of the above mixture.

MS/MS

After analyzing the result of all the ions, we can take one of the ions and perform a high energy collision with a nobel gas which would result into more ionic fragments; we then take those fragments and run through the MS. This technique of analysing and fragmenting then analysing again is called **MS/MS** or **tandem MS**.

Remark 2.10. *The fragmentation of this ion molecules (of the same population) does not break all of its peptide bond. The fragmentation is random*

and partial, which means that it only a small amount of peptide bond is broken per molecule (on average).

Example 2.4.5. Continue from example 2.4.4, we will use one of its fragment of m/z 836.47 (called FIIVGYVDDTQFVR) and run it through MS/MS. As you can see, the FIIVGYVDDTQFVR would break into either VDDTQFVR

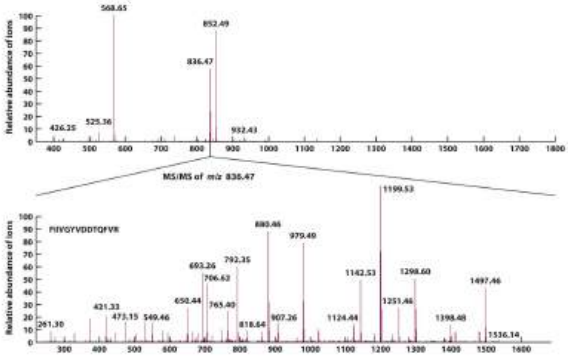


Figure 2.31: FIIVGYVDDTQFVR through MS/MS

(979.03), GYVDDTQFVR (1129.29), etc.

These information from MS/MS, which is analyzed computationally by computer with known protein sequences, can give the amino acid sequence on the peptide ion. The analysis of proteins samples by MS and bioinformatics to identify a population of protein in subcellular organelles is called **proteomics**.

2.5 Protein Analysis II

We're going to look at more interesting techniques to separate proteins.

2.5.1 Chromatography

Chromatography is a seperation technique where gravity and is used to separate proteins of different size.

There is 2 phases in chromatography: mobile and solid phase. The mobile phase moves past the solid phase and typically it is aqueous. We first filled the chromatography column (like a buret) with the **solid matrix** (solid

phase). The aqueous phase will be then added to the top of the solid matrix and let gravity pull it down. As it moves down, we will add some solvent to facilitate the movement. The rate as the mobile phase moves through the solid matrix is depending on their interactions.

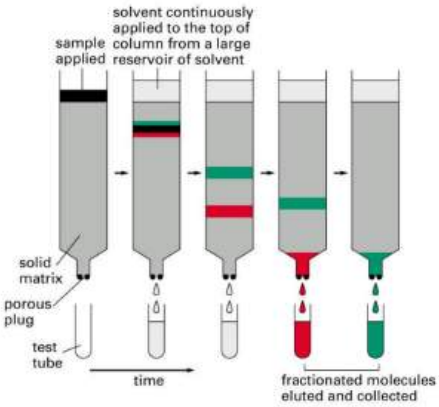


Figure 2.32: General mechanism of chromatography.

Gel Filtration Chromatography

Gel filtration chromatography is a type of chromatography that filter proteins by size. The solid matrix now is made out of polymer gel bead that is porous. When the mobile phase is added, it begins to separate by size.

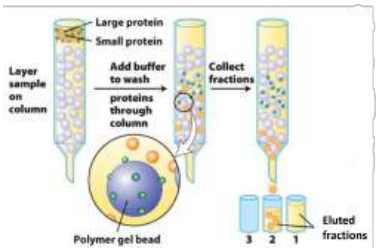


Figure 2.33: Mechanism of gel filtration chromatography

Smaller molecules (proteins) can go into the pore and through the channel of the gel bead which takes more time. On the other hand, the bigger

molecule (proteins) cannot enter thus will pass across the gel bead taking less time.

Ion-Exchange Chromatography

Ion-Exchange Chromatography separates proteins based on electric charge. The solid matrix are are now made from beads that are positively charges. When the mobile phase is added, it would begins to separate by charge (positive moves down faster).

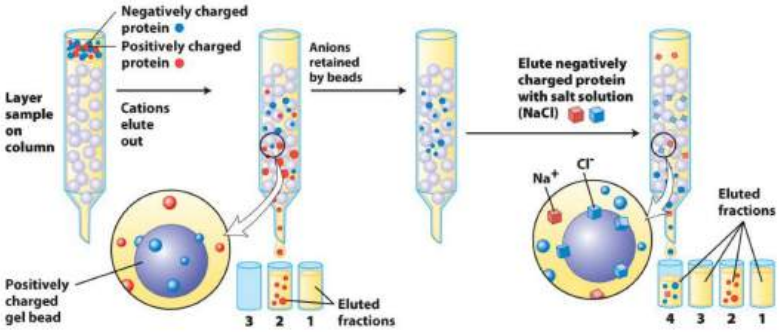


Figure 2.34: Mechanism of ion-exchange chromatography

The negatively charged proteins will be attracted to the positive beads while the positive proteins will be repelled thus making it travel down the column faster.
To get the rest of the negative proteins, we will wash the solid matrix with high concentrated NaCl which will displace the negative protein.

Brief Overview of Antibodies

Before moving on to the next type of of separation, we would like to have a look at 2 concept of antibodies.

Antibodies recognized an antigen by only the **epitope** surface. Each antibody will recognized an individual epitope via its **paratope**. Antibodies can be raise to go against any type of chemical antigen including proteins.

Antibodies has 2 region: **constant region** and **variable region** (where the paratope is located). The constant region is the same among homogeneous species of antibodies; however they're different from 1 species to the

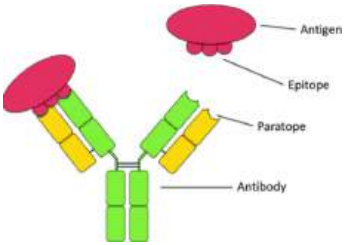


Figure 2.35: Epitope and paratope surface

next. e.g. mouse antibodies is familiar to rabbit antibodies.Using this understanding, we can have a system of 2 antibodies: primary (bind to antigen epitode) and secondary (bind to the primary constant region)

Antibody-Affinity Chromatography

In this type, the bead would be attached with an antibody that recognized a specific types of proteins. We then allow the proteins go through the solid matrix, any proteins that is targeted by the antibody will be stopped while all the rest will flow through. To unbind that specific proteins, we elute the matrix with a low pH mixture to denature the antibody.

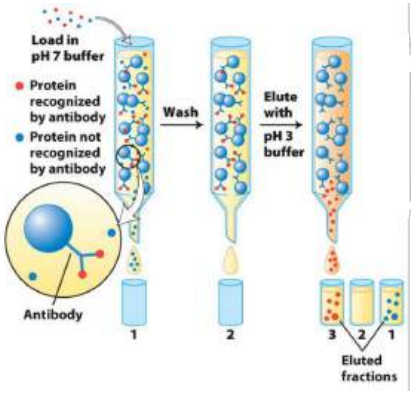


Figure 2.36: Mechanism of antibody-affinity chromatography.

2.5.2 Antibody Separation Method

The following method utilize primary and secondary antibodies (sometimes) for the process of separation. Some of these are combined method of previous method we've seen.

Immunoblot

Immunoblot or **Western blot** is a type of electrophoresis for us to identify a specific protein using its corresponding antigen. It is an analytical method

We first prepare the typical electrophoresis setup. Now instead of running the electric current parallel to the ground through the gel, we'll be running it perpendicular to the ground thus making proteins going to the bottom of the gel plate (top down). At the bottom is a membrane which would catch the antigen (protein). After catching this, We will then incubate for the first round with the primary antibodies for it to bind to the specified antigen. Then on the second round of incubation, the secondary antibodies will bind to the primary and allow us to detect them (There is an enzyme complex linked with the secondary for detection).

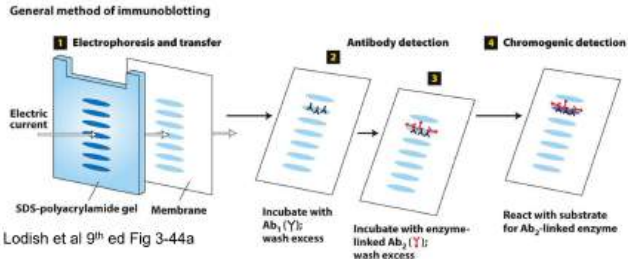


Figure 2.37: Mechanisms of immunoblotting.

Remark 2.11. *we use 2 round of incubation since the cost of making an antibodies complex along with enzyme is costly for most researcher i.e. the antibodies complex of primary and and secondary is more expensive than the technique.*

Immunoprecipitation and Co-Immunoprecipitation

Immunoprecipiation (IP) and **Co-Immunoprecipitation (Co-IP)** is a separation method where a protein complex is separated by just the separation

of 1 protein in that complex.

A mixture of proteins is prepared. Antibodies that bind to a specific protein will be dropped into the mixture. The antibodies will bind to that protein and separate it from the complex. We then add special protein (G coupled bead) that can bind to this antibody then goes through 1 round of centrifugation.

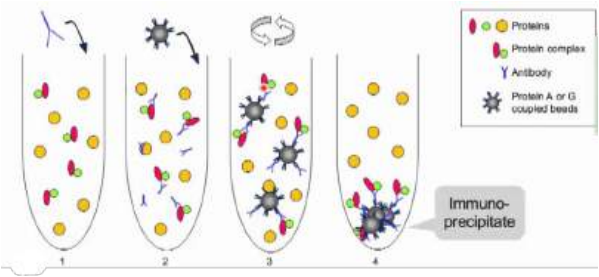


Figure 2.38: Mechanism of immunoprecipitation.

Not only that we'd find the protein-antibody complex in pellet, we would find that the entire protein complex is now attached to the antibody. The protein-antibody complex in pellet is also called the **immunoprecipitate** and the protein that previously attached to the protein now forming a complex along the immunoprecipitate is the **co-immunoprecipitate**.

Immunofluorescence Microscopy

Immunofluorescence Microscopy is a separation technique where fluorescent antibodies are attached to a specific protein in which it will glow under fluoresence microscope.

We prepare a sample on a microscope slide, then introduce the primary antibodies which would bind to a specific proteins. After the binding is done, a secondary fluorescent antibody will be introduce which will glow under fluorescence microscope.
Method of Making Fluorescence Antibodies: Mouse's purified antibodies are injected (acting as an antigen) into a rabbit. The rabbit will start producing antibodies that would bind to the mouse antibodies. The rabbit's antibodies are thus secondary and the mouse's antibodies are primary. We can then chemically couple a *fluorochrome* to the secondary antibodies

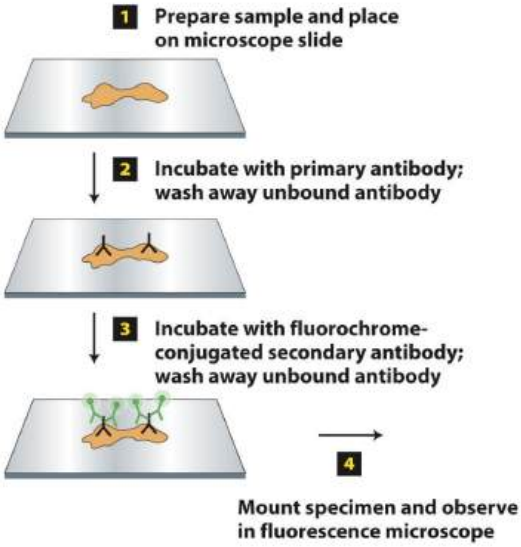


Figure 2.39: Mechanism of immunofluorescence microscopy

which allow detection under fluorescence microscope.

Example 2.5.1. Glucose transporter (GLUT2) is labelled with yellow-green fluorescence in a section of rat intestinal wall.

As we can see from this, GLUT2 only on the lateral and basal surface of the intestinal membrane and not at the apical surface (brush border).

We can perform a double or triple immunofluorescence microscopy etc. by first bind different primary antibodies to proteins then bind different secondary fluorescent antibodies with those primary antibodies then it will glow in different colour under the fluorescence microscope.

Example 2.5.2. Double immunofluorescence can be used to study the distribution of 2 proteins such as vinculin and actin in a cell. From the immunofluorescence, we can see that actin and vinculin has some association with each other; however, vinculin is more at the cell surface while actin is all over the cell.

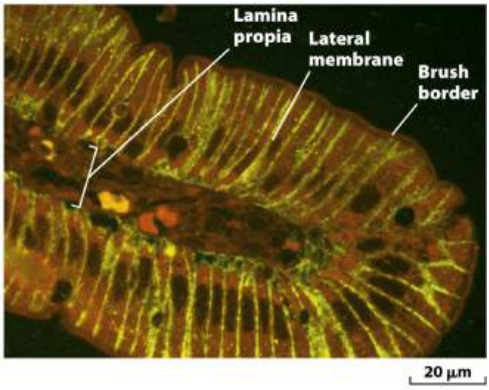


Figure 2.40: Immunofluorescence of GLUT2

Green Fluorescent Proteins (GFP)

GFP is a jelly fish (*Aequorea victoria*) protein that glow green. It is an enzyme that modify its own amino acid side chains to create fluorochrome. If we can take this jelly fish genes for GFP and introduce it to other organisms, GFP will also be produced in that organism body. This GFP can be used as reporter gene for transcriptional control or fusion proteins for intracellular protein localization studies.

Definition 2.14. **reporter gene** is a gene that is attached to a sequence of gene of interest. It is used as an indicator whether the cell or organisms has taken up or location of gene of interest.

Definition 2.15. **Fusion protein** is a protein that can fuse with another functional protein (without altering its functions) which make that protein detectable.

For acting as reporter gene, we take a piece of genes called the genes promoter and insert GFP while replacing the gene we wanted to expressed. For acting as fusion protein, we can even tag GFP along with the gene sequence which would illuminate that protein after translation.

Example 2.5.3. ODR10 is an odorant receptor (found in roundworm: *Caenorhabditis Elegans*) expressed only in sensory neurons and specifically targeted to the tips of those neurons.
We first use GFP as reporter gene by replacing the ODR10 gene from the promoter and insert the GFP. This results in the GFP expressed in cells

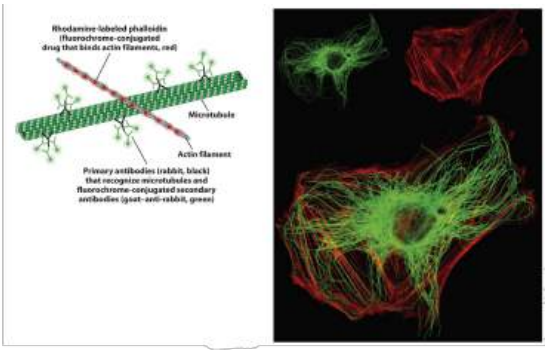


Figure 2.41: Double immunofluorescence microscopy. Actins are labelled with red and microtubules are labelled with green

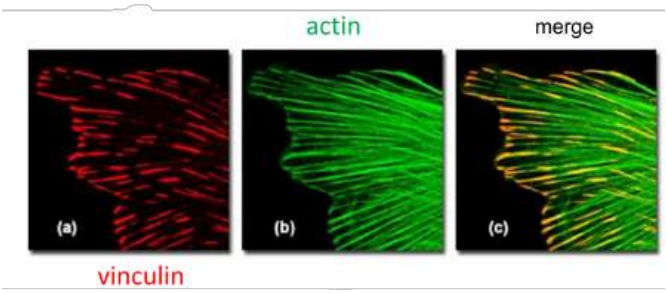


Figure 2.42: Immunofluorescence of vinculin and actin

where ODR10 gene would be active.
For the second method, we use the GFP as fusion protein by inserting it along with the ODR10 gene. As a result, ODR10-GFP fusion protein would be found where ODR10 is normal targeted (most active).

Remark 2.12. *Not only the colour green is available but researchers have engineered to create different colour variant.*

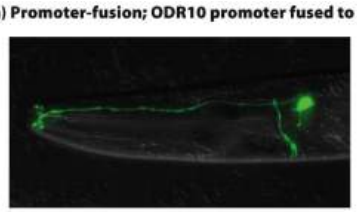


Figure 2.43: ODR10 gene is replaced with GFP at the promoter region.

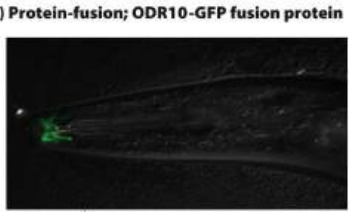


Figure 2.44: ODR10 gene is connected along with GFP.

2.6 Review Questions

Does misfolding affect only quaternary structure of proteins, or can it affect other structure?
It affects allstructure except primary

Does SDS discrupt secondary structures in addition to tertiary and quaternary structure of proteins?
Yes, it would discrupt the typical bond and make the entire protein becoming lienar again.

How is it high pH and low pH interfere with catalytic activity of trypsin and chymotrypsin?
At low pH, we cannot initiate catalysis of chymotrypsin. To pick up the 1 proton from ser-195, the his-57 must be unprotonated (via acidic environment).

When separating by size when would use differential centri or gel filtration chrom?
They work best on differetn size ranges. diff centri are good for larger objects like subcellular organelles while gel fil is best for smaller object.

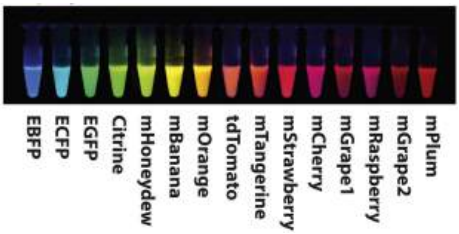


Figure 2.45: Colour Variants

In tandem MS, doesnt the first MS effectively destroy sample? how can sth is recovered from it?
On the second run, the ion of interested will not be collided with the detector hence it won't be destroy. This ionic beam of interest will be go through mass spectrometry again through argon gas. All in all, we don't recover it, what we do is performing a second run.

In MS, does the liquid passing through the needle have 1 frag of protein or it can be a mixture of different protein? It can be either, proteomics.

What is the difference between antibo-antige chrom and immunopre and co-IP?

Chapter 3

DNA and Genetic Mechanisms

Since the dawn of time, human has always wanted to keep a record of information of their time, like the Kish table (35th century BC) or Dead Sea Scrolls (3th century BC). Nevertheless, there are lots of issues arise with this kind of method of recording information "permenantly". These issues can be: requirement of physical space, material degradation or error during copying and translating (from 1 language to the next).
Just the way genes are stored in long stretchy DNA but 1 thing for sure is that DNA and information record (made by human) cannot compete with each other. Take electronic information storage, the highest theoretical storing time so far is more than 100 years but DNA has been forever and lasted roughly 4 billions years.

This is our introduction to the topic of DNA and its mechanism of information storage and replicate.

3.1 Principles of DNA Replication

Whenever a cell under goes division, the information stored in its DNA will be replicated.

During replication, the double stranded DNA is denatured and leaving 2 disjoint parental strands. Each of them will be used as *template* for the

formation of 2 new daughter strands (which is complementary to the parent strands). The formation of the daughter strands is aided with the DNA polymerase; where the $\alpha-$phosphate group of the incoming dNTPs reacts with the OH group in the growing DNA chain unidirectionally from 5' to 3'. Nevertheless, all DNA polymerase can do is adding dNTPs and nothing else. In order to perform a sucessful replication, we need other structure.

3.1.1 Mechanism of Proteins and Replication

Many fundamental processes involved many different proteins and nucleic acid coming together with different roles creating a *molecular machine*.

Remark 3.1. *We'll be using the term "dsDNA" which is short for double strand DNA, while "ssDNA" is single strand DNA.*

The first 2 proteins structures for DNA replication are: **topoisomerase** and **helicase**. In order to the dsDNA to replicated, each of its strand need to be separated, which is where helicase come into play. It will break the hydrogen bonds between complementary base pair thus separating it. As the dsDNA get denatured which create torsion and straints, topoisomerase relieve this supercoiling.

Definition 3.1. The point at which the 2 strands of the dsDNA begin to separate is called the **DNA fork**. At this point, helicase is still continuously going through the DNA and denatures its complementary bonds.

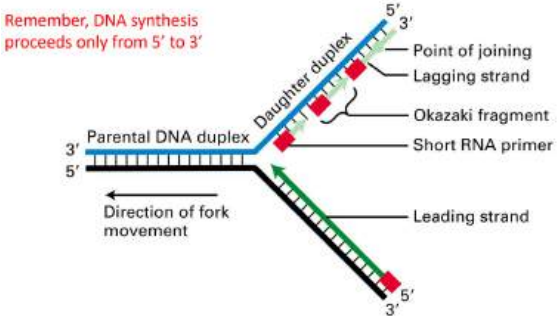


Figure 3.1: DNA replication.

DNA polymerases cannot initiate the synthesis a new strand but they can elongate an existing strand. A specialized RNA polymerase called **primase** will forms a short RNA molecule (called **primer**) complementary to the ssDNA region which allow DNA polymerase to attach and extend.

On the 3'OH strand, synthesis goes in a continuous motion because primase and polymerase will read from 3' to 5' which would synthesize from 5' to 3'. The daughter strand of this continous faction is also called **leading strand** (in the direction of the replication fork). Contrarily, on the 5'phosphate strand, synthesis are fragmented since polymerase reads in 1 direction but the strand run the other direction hence they can only read and and synthesize in small fragments, these fragments are called **Okazaki fragment**. daughter strand made from these small fragments is called the **lagging strand**. When fragments are completed, the RNA components is replaced with DNA and adjacent DNA will be "glued" together via **DNA ligase**.

3.1.2 Replisome

Definition 3.2. **Replisome** is the machinary comprised of multiple protein complexes that sre involved in the replication of DNA.

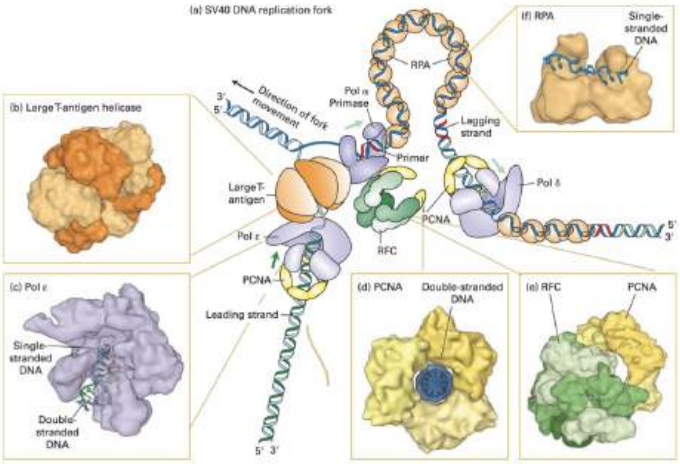


Figure 3.2: Replisome of SV40

Many of our understand of DNA replication is from prokaryotic viral infection studies. **But why would we study viral infection for DNA replication?** Because a virus' main job is to make a whole bunch of itself via replication thus they have a high replication rate. Therefore it is easier to study something with a higher replication rate for DNA replication.

Remark 3.2. *Viral replication is interesting since it* ***replicates using the host machinery*** *(only different in helicase).*

We will now look at the components of replisome and their involvement with DNA replication of **simian vacuolating virus 40 (SV40)**.

SV40's helicase are known as **large T-antigen helicase** is a hexameric protein structure (made from 6 polypeptide molecules, 3 pair of different kinds). As the name implied, they act as helicase which unwind the dsDNA at the replication fork and is made by the viral genome.
Replication protein A (RPA) is a protein that binds to the ssDNA and stabilize it. It keeps ssDNA template in an optimal shape for DNA polymerase to read and synthesize as well as increases accuracy of its reaction (more precision in base pairing).
DNA polymerase ε (Pol ε) is a DNA polymerase that synthesize the leading strand.
Proliferating cell nuclear antigen (PCNA) is a homotrimeric proteins ring-like structure that wrap itself around the complex of Pol ε to unable it from dissociating from the template and stabilize it as it synthesize the daughter strands.
Primase/Polymerase α complex consists of a primase which would form the RNA component of the primer and a DNA polymerase α (pol α) that extends the RNA primer with DNA.
Pol δ/RFC/PCNA complex is a protein complex consists of polymerase δ which would replace RNA with DNA, and an RFC (replication factor C) PCNA complex where the RFC would open the PCNA ring and load it at the primer of the DNA where it can also goes with the pol δ. Essentially, this entire complex replace the primase/polymerase α complex as well as completes the synthesis of the Okazaki fragments.

Remark 3.3. *The biggest difference between pol α, δ and ε is that pol α lacks exonucleases (discuss later on) that can proofread the replication process which means that it cannot carry replication as long as pol δ and ε.*

Finally, it is the **Ribonuclease H and flap structure-specific endonuclease 1 (FEN-1)** which would displace the RNA component at the 5' ends

of the Okazaki fragments. Then when everything is in place, **DNA ligase** which would ligate the adjacent DNA together.

Origin of rep happens at specific place of the region, and they're mostly AT-rich. The reason it is AT rich is because it's easier to denature it.

3.1.3 Rigorous Description of DNA Replication Steps

We will now look more in-depth at the steps of DNA replication. Before that, we must know that all DNA replication happens specific point(s) the dsDNA. The region(s) where DNA replication starts is called **origin(s) of replication**.

Remark 3.4. *One special thing about the origin of replication is that* ***it is A-T base pair rich****. The reason is that AT base pair is easier to denature than CG base pair.*

Helicases will get attracted to the replication origin and unwind (hydrolysis of ATP) a segment of the double helix DNA. RPA will bind to each ssDNA. RNA primer are synthesize by primase-pol α. The leading strands then get extended from 5' to 3' from both side of the the DNA.

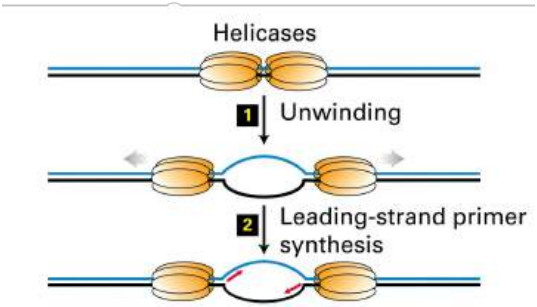


Figure 3.3: Helicase attachment and primer synthesis.

As helicase moves along, the replication is happening bidirectional. Pol ε/RFC/PCNA complex will continue to synthesize the leading strands on both directions.

But as we keep synthesizing along, there will be regions with no replication...well because this is the lagging strand portion: primase-pol α will

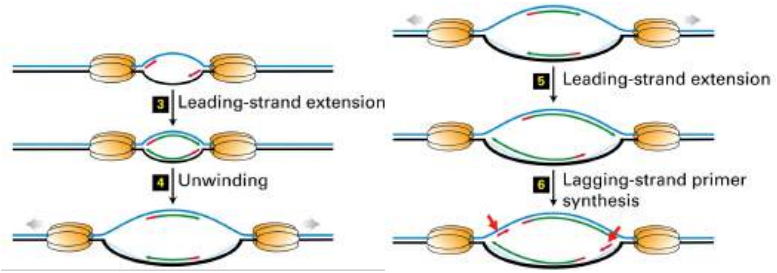


Figure 3.4: Helicase extends and leading strand is extended.

make the primer then will be replaced by Pol ε/RFC/PCNA complex to extend the Okazaki fragments, etc.

This process will be repeated over and over then when this is done, FEn1 and DNA Pol 1 will remove the primers and ligase will link all of the fragments together.

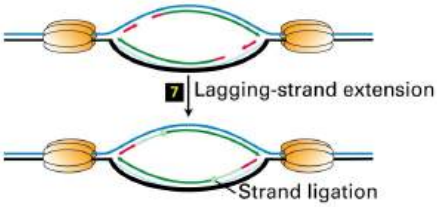


Figure 3.5: Separate strands are ligated together via DNA ligase.

Remark 3.5. *The new dsDNA has 1 parental (original) strand and 1 daughter (synthesized) strand and this structure of 1 new and 1 old is called* ***semi-conservative****.*

3.2 DNA Repair and Recombination

DNA is constantly subjected to stress and damage. On average, each cell experiences around $2,000 - 10,000$ hydrolytic depurination per day,

Definition 3.3. **Depurination** is a disruption of the Purine's backbone causing it to breakaway from the DNA.

It also experiences roughly 1 cytosine **denamination** (removal of NH_2), 1 guanine **oxidation** every 5 days and 600 adenine **methylation** per days.

DNA is especially important since it holds all genetic information and by subjecting it to this much stress, it can lead to mutation. Nevertheless our cells has mechanisms to fix and repair the damages; before that, let us understand some terms.

Definition 3.4. A **mutation** is a change in an organism' DNA sequence. It can be permanent, transmissible and occurs spontaneously.

Definition 3.5. **Mutagens** are chemicals and/or substances that can increase the frequency of mutation e.g. UV radiation, ionizing radiation, etc.

Definition 3.6. **Carcinogens** are agents (sometimes mutagens) that causes cancer.

The following are human diseases and cancerous syndrome that are caused by DNA mutation where its repair system is defected.

Disease	DNA-Repair System Affected	Sensitivity	Cancer Susceptibility	Symptoms
Hereditary nonpolyposis colorectal cancer	DNA mismatch repair	UV irradiation, chemical mutagens	Colon, ovary	Early development of tumors
Xeroderma pigmentosum	Nucleotide excision repair	UV irradiation, point mutations	Skin carcinomas, melanomas	Skin and eye photosensitivity, keratoses
Bloom's syndrome	Repair of double-strand breaks by homologous recombination	Mild alkylating agents	Carcinomas, leukemias, lymphomas	Photosensitivity, facial telangiectases, chromosome alterations
Fanconi anemia	Repair of double-strand breaks by homologous recombination	DNA cross-linking agents, reactive oxidant chemicals	Acute myeloid leukemia, squamous-cell carcinomas	Developmental abnormalities including infertility and deformities of the skeleton, anemia
Hereditary breast cancer, BRCA1 and BRCA2 deficiency	Repair of double-strand breaks by homologous recombination		Breast and ovarian cancer	Breast and ovarian cancer

Figure 3.6: Human diseases that link with DNA repair system defect.

3.2.1 DNA Polymerase Proofreading

DNA repair begins with the **proofreading by DNA polymerase**. When performing experience with DNA polymerase alone, we see that the error oc-

curs at a rate of 1 error per 10,000 nucleotides but when we actually measure error rate in cells, we found it is of 1 error per 1,000,000,000 nucleotides. **How can this be?**...well, this is thanks to the DNA pol proofreading mechanism.

In Eukaryotes, DNA pol ε and δ (except α) have an **exonuclease site** that can proofread DNA from $3' - 5'$. When there is an incorrect base pairing during replication, DNA pol will pause and the incorrect strands will be inserted into the exonuclease site. The exonuclease will hydrolyze the phosphodiester bond thus remove the mispaired base.

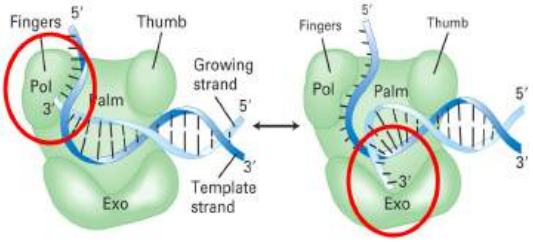


Figure 3.7: Exonuclease site on the DNA polymerase

Remark 3.6. *The reason it is called exonuclease is because it hydrolyzes at each end of DNA strands (during replication).*

3.2.2 Base Excision Repair

Deamination of cytosine (base C) can also produce thymine (base T) which mean that during the replication instead of creating the base pair of GC, it will create a base pair of AT.

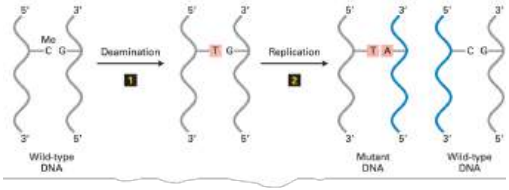


Figure 3.8: Deamination of cytosine to thymine.

When there is a base change at a single base pair, we call it a **point mutation** and when the base pair of the original DNA isn't complemetary to

each other, we call it a **mismatched base pair**. A problem arises before doing repair is that **which base in the mismatched base pair is the one that has the point mutation?**...The majority of TG mismatches stems from deamination of C to T therefore T is incorrect.

Now that we know which base to repair, we can begin to look at the mechanism. First **DNA glycosylase** will break the bond between the the thymine base with the sugar back bone. **APEI endonuclease** will cut the DNA strand where there is a missing base. AP ligase will the remove the sugar backbone leaving a blank point where DNA pol β will increase the C and ligase to repair the sugar backbone.

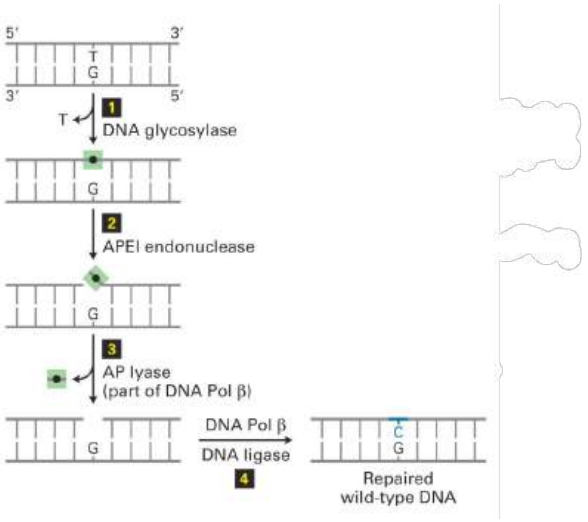


Figure 3.9: Mechanism of Base Excision Repair

3.2.3 Mismatch Excision Repair

Now we will generalize the previous problem into mismatches of multiple nucleotides (this includes false insertion/deletion of nucleotides) now we will ask the same question as before, **which bases in the mismatched during mutation?** In this situation, the newly synthesized strand is the wrong one. Once the synthesized strand with mismatches is recognized, mismatch exicision repair will initiated.

The mechanism begins with **MSH2 and MSH6 protein** recognized the mismatched daughter strand. This recognition will trigger the activation of MLH1 endonuclease along with PMS2. DNA helicase will come and unwind this section, the MLH1 endonuclease will cut away the newly synthesized strand and DNA exonuclease will come and digest nucleotides to the daughter strand.

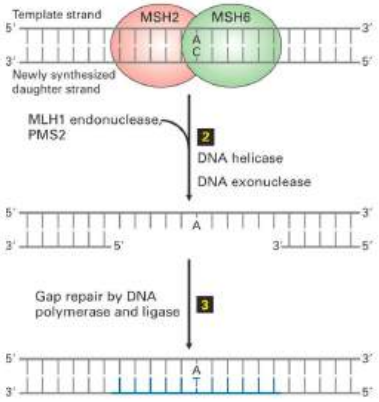


Figure 3.10: Caption

DNA polymerase δ will come and fill in with correct nucleotides in the gap where the endonuclease cut.

3.2.4 Nucleotide Excision Repair

DNA excision repair is a mechanism that fixes DNA regions where there is distortion or chemically modified. Chemicals including many carcinogens can cause this distortion or changes. One example of this is the distortion of thymine thymine dimer. In this situation, UV radiation causes 2 adjacent thymine base pair bond with each other instead of their adenine complementary base pair. (see Figure 3.11)
When there is no repair to this region, the normal DNA pol will pause and a special DNA pol called DNA pol η will read through the T-T dimer but this polymerase lack proofreading which at the end, the daughter strand will have a high concentration of error.

Now we've understand what this thymine-thymine dimer can do to the later synthesized strands, we will look at the mechanism used to counter

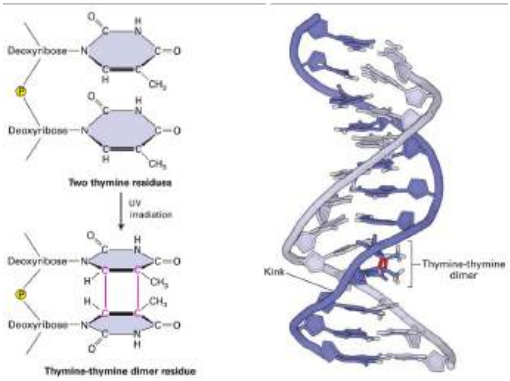


Figure 3.11: Thymine-Thymine dimer.

this dimer.
We begin with $XP-C$ and $23B$ proteins recognized the distorted double helix. After this recognition, TFIIH, XP-G, and RPA will be activated and unwind the DNA creating a "bubble" of 25 nucleotides.

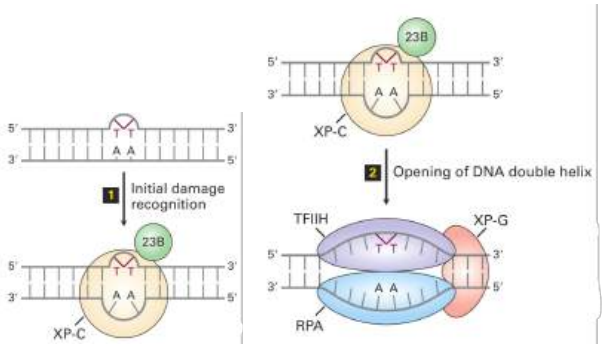


Figure 3.12: Mechanims of nucleotide excision repair.

XP-F and XP-G endonucleases will cut out the damaged strand.

Remark 3.7. *XP-n's name is derived from the xeroderma pigmentosum, which is a genetic disease that cause a high disposition of UV-induced cancer. In such disease, protein XP is damaged.*

After the XP endonuclease cut out the damaged strand, DNA pol will fill in the damaged strand gap while DNA ligase will repair the phosphate

backbone.

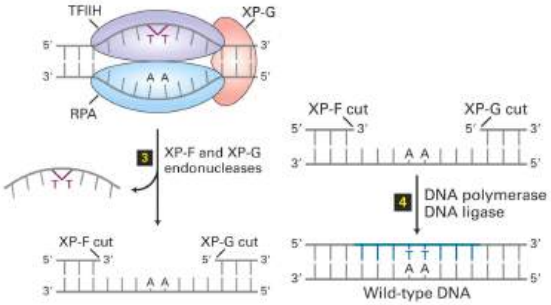


Figure 3.13: Mechanisms of nucleotide excision repair (part 2).

3.2.5 Double Strand Break Repair: End-joining

Radiation and certain anticancer drugs (such as bleomycin) can cause a dsDNA break (we will abbreviate that to **DSB**). If this DSB is not fixed, any from the distal location will be lost which can be lethal for cells. End-joining mechanism can rejoin 2 broken dsDNA but most of the time, certain bases are lost at the joining points which can introduce deletion mutation. Nevertheless, we have a mechanism against deletion mutation (explained above) so we can still proceed with endjoining.

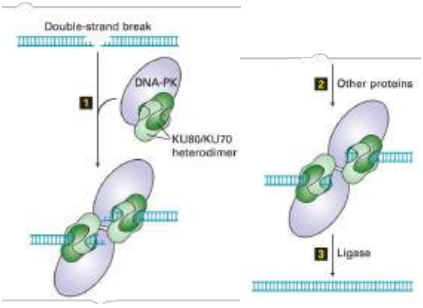


Figure 3.14: Mechanism of end-joining for DSB.

First **KU80/70 heterodimer** along with **DNA-PK** will bind to the end of each DSB. Once DSB-protein bounded come together, they will recruit

some nucleases to remove certain bases. After this, the 2 strands will be ligated back together.

Remark 3.8. *End-joining mechanism does not ensure the ligated part was part of the original DSB i.e. 2 joining part can be from 2 different DSBs come together; this would produce a* ***chromosomal rearrangements****.*

3.2.6 Double Strand Break Repair: Homologous Recombination

When DNA sequence is damaged, it is replaced with an undamaged copy of it (same sequence) on the homologous chromosome. This requires and exchange of strands between 2 DNA which is called **recombination**. The homologous recombination repair proteins are encoded by the genetic sequence $BRCA1$ and $BRCA2$. Take an example of damaged DNA such as the collaspsed of replication fork.

Remark 3.9. *To make the explanation smoother, we will assign colour to each of the strand we'll be talking about: parental strands are light and dark blue while the daughter strands are light and dark green.*

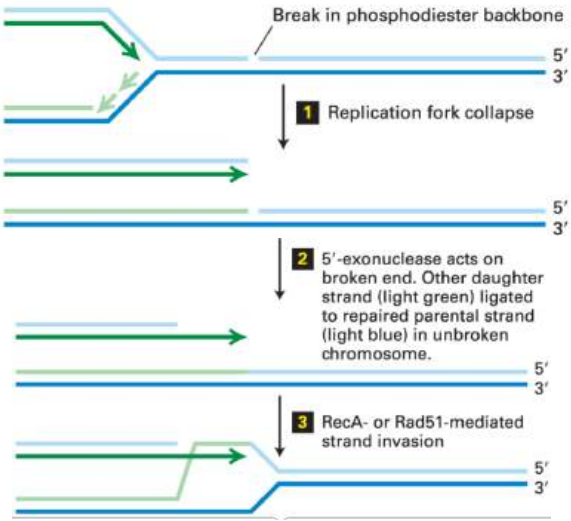


Figure 3.15: Mechanism of homologous recombination (part 1).

Replication of the light and dark blue strand will stop when there is a break occured. The light blue and dark green strand pair will be partially removed because of the break while the light green strand will be ligated to the other end of the light blue strand. Then **strand invasion** will occur where the new ligated light green strand will be anneal with the dark green strand that was partially removed. (See Figure 3.15)

Remark 3.10. *For the sake of visualization, it is put as 2D but in fact, the diagonal of the light green strand in step 3 has only 1 phosphodiester bond.*

After strand invasion is done, **branch migration** is initiated where the partially removed light blue strand can migrate to the dark blue strand and anneal with it. The crossing of the light blue and light green strand during invasion and migration is called the **Holliday structure**.

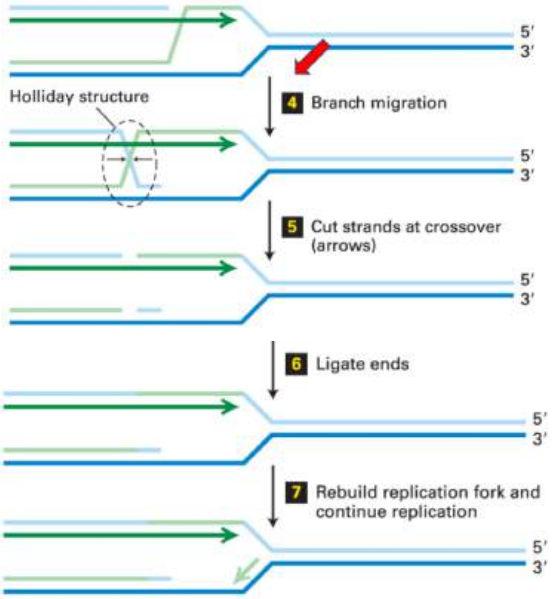


Figure 3.16: Mechanisms of homologous recombination (part 2).

The rest of the structure is ligated together which resembles the original replication fork. DNA polymerase is recruited long with the rest of the replisome and replication resume.

Remark 3.11. *Certain organism has highly efficient DNA repair mechanism/system which allow them to have some resistance toward radiation e.g.*

cockroaches, tardigrades, etc.

3.3 PCR and DNA sequencing

The following are laboratory method to either amplify a certain DNA sequence or determine the sequence itself.

3.3.1 Polymerase Chain Reaction

Definition 3.7. **Polymerase Chain Reaction (PCR)** is a laboratory technique where a specific DNA sequence is rapidly produced (amplification). Many other technique are based on the frame work of PCR.

PCR is essentially good for sequencing, DNA cloning, pathogen detection and even gene editing. PCR however requires 1 to know the nucleotide at the ends of the region to be amplified. For a single PCR, the following are also required: DNA template (can be a complex mixture), DNA polymerase that can function at high T^o, DNA primers which are designed to be complementary to the ends of regions of amplification, and dNTPs (which acts as monomers for the template).

PCR Mechanism: 3 Step Cycle

We begins with apply heat shock ($\sim 98^oC$) to the DNA template which would destabilize and separate it into 2 ssDNA [*denaturation*]. We then decrease the temperature down to $48-72^oC$, at which we will add DNA primers, and they will pair complementarily to the end region of the sequence to be amplified [annealing].

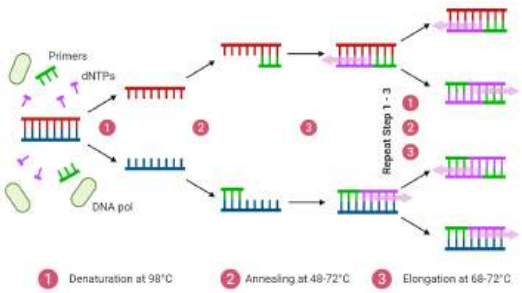


Figure 3.17: Process of PCR

We then increase the temperature back to $68-72^oC$ where dNTPs will be added and along side the heat resistant DNA polymerase and the region to be amplified will be made [extension]. The heat resistance DNA pol that we typically used in the lab is called **Taq polymerase**.

Remark 3.12. *Taq polymerase got its name for being the DNA pol of a thermophilic bacterium* ***thermus aquaticus****. We use it because at high T^o, it won't be denatured*

Remark 3.13. *Taq polymerase is cheap but it doesn't have proofreading activity therefore, it is good for small fragments amplification.*

To reach high level of amplification, we typically perform PCR of 20-40 repetitions of these 3-step cycles (denaturation-annealing-extension). This introduce problems with normal laboratory procedure as it can be repetative and long. Engineering advances has designed the **thermo cycler** which can rapidly change T^o and help with the repetative nature of PCR.

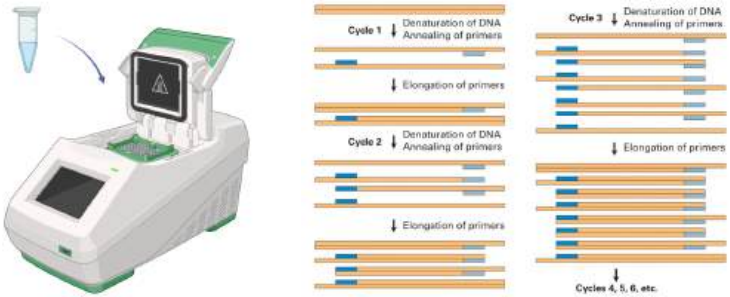


Figure 3.18: A thermo cycler aiding in the repetition of a 3 steps PCR

The amplification of DNA sequence using PCR is very sensitive and exponential, meaning that with just a small amount of template, it can produce a extremely large amount of the region to be amplified. Every cycle would amplify twice as much DNA sequence as the previous (just until the concentration of dNTPs) reduces.

Remark 3.14. *As more cycle of PCR is ran through, there would be high suseptibility to error (present as noise in the amplification).*

PCR Analysis: Gel Electrophoresis

This final step to PCR is not part of amplification of a DNA sequence. This step is just to analyze the amplified sequence using gel electrohporsis which can determine its size.

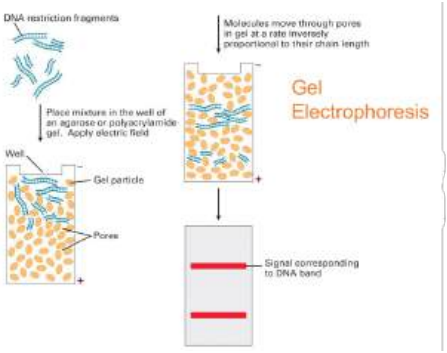
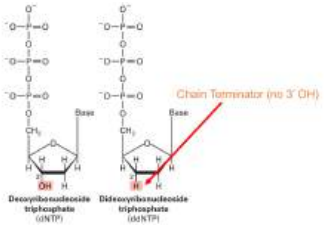


Figure 3.19: Electrophoresis of PCR.

3.3.2 DNA sequencing: Deoxy Chain-Termination Method

Definition 3.8. **DNA sequencing** is a process to determine the order of nucleic acids in a DNA sequence.

DNA sequencing is good to figure out the genome of an organism (which we will look later on), determine important genes and regulation etc. There are many method of DNA sequencing but one of the method we'll be looking which were primarily used in the 1970s is called **deoxy chain-termination DNA sequencing** or simply **Sanger sequencing**.
For Sanger sequencing, we are required: 4 test tubes (mostly Eppendorf tubes) containing DNA pol, primer (oligonucleotide), DNA template and dNTPs. In addition to these 4 tubes are added low concentration ($\sim 1mM$) of 4 dideoxyribose nucleotides (ddNTPs). Each tube will receive only 1 of 4 types ddNTPs. The components in these 4 tubes are similar to that of PCR but the different is the addition of ddNTPs. One special feature of ddNTPs is that the OH group on the $3'$ end of the dNTPs is deoxygenated and become only H. The reason for adding them is that ddNTPs during polymerization, can bind to the template at random length. After binding to the growing DNA strand, polymerization stop since ddNTPs does not have the OH to dehydrate to form a bond.



Because of the low [ddNTPs], the insertion of these ddNTPs will vary in sequencing length. When looking at 1 of the tube containing only the ddGTPs, synthesis would be stop at every *GC* pairing on the template.

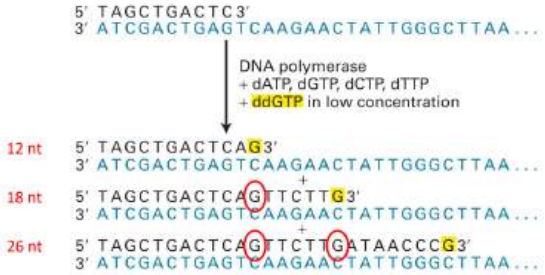


Figure 3.20: DNA sequencing in G-tube

We can then run them through agarose gel electrophoresis for all 4 ddNTPs tubes which it will form bands depending on the length of the nucleotide sequence.

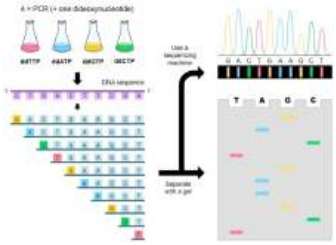


Figure 3.21: DNA sequencing Electrophoresis.

Computer and analyitcal method will be used to analyze and sequence the DNA.

Limitations

Even if DNA sequencing method is great, polymerase can only runs for 300-500 nucleotide base at most and gel can only create a limited amount of visible band (it will be hard to see bands of 500nt vs 501nt). Furthermore, the sequencing a large region, we would have to break it down into smaller sequencing. Because of this, the rate of sequence production is limited by the amount of reaction needed but as well as running this amount of reactions it could drive up the cost.

Remark 3.15. *At the beginning, it took researchers 13 years and using over 13 billions dollar just to sequence all of human's DNA (genome). As of today with better and more efficient method, it cost around* 600$.

3.3.3 Next Generation Sequencing

Definition 3.9. **Next generation sequencing (NGS)** or even **massive parallel sequencing** is a sequencing technique where millions of sequencing reaction can be carried out at the same time.

The process of NGS is fairly lengthy (as it also use a different technique) but we will break it down in each step.

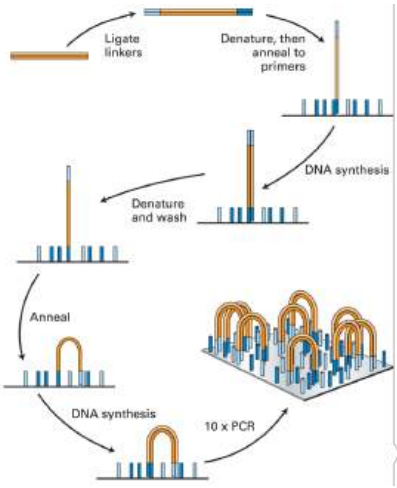


Figure 3.22: General mechanism of NGS.

To begin with, dsDNA fragments are isolated and will have "adapters"

attached to each of its ends by **ligate linkers**. This new dsDNA will be denatured then anneal to primers anchored on a support that is complementary to the linkers/adapters. PCR will be then conducted to amplify all of these dsDNA fragments in a fixed spatial arrangement.
Next, all of the newly synthesized dsDNA will be denatured and washed, left only 1 ssDNA attached to the support. Then, fluorescent dNTPs will be sequenced on to this ssDNA; however only 1 dNTPs would bind at a time. After 1 fluorescent dNTPs bind, computer will image the fluorescing dNTPs. We then chemically remove this fluorescent ability while still attached the base then start the process all over again.

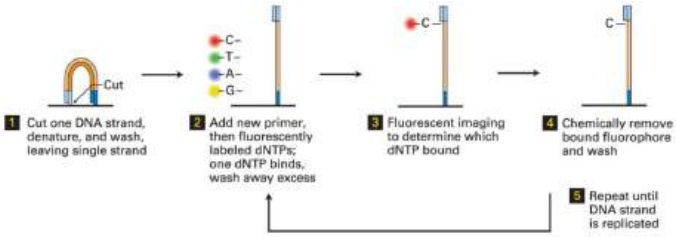


Figure 3.23: Fluorscent imaging DNA sequence.

At the end, we will have a full fluorescent image sequence of the entire DNA fragements and any other fragments.

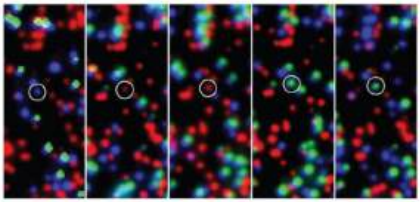


Figure 3.24: Imaging of NGS

Each of the coloured dot is 1 nucleotide of different DNA sequences. The entire procedure takes an entire day but at the end, you'd be able to sequence over 10^{11} bases. After sequencing all of the necessary DNA sequence, we can use computer recombine these sequences becoming a genomic sequence (process is called genome assembly).

Currently there are even newer techniques in development that takes

even less time while sequencing longer DNA chain. Some of these method are: **pacific bioscience sequencing** that uses a DNA sequence and record the colour change accoding to each base pair; or **Oxford nanopore technology** that uses a DNA sequence and feed it through a current induced pore and record the reading.

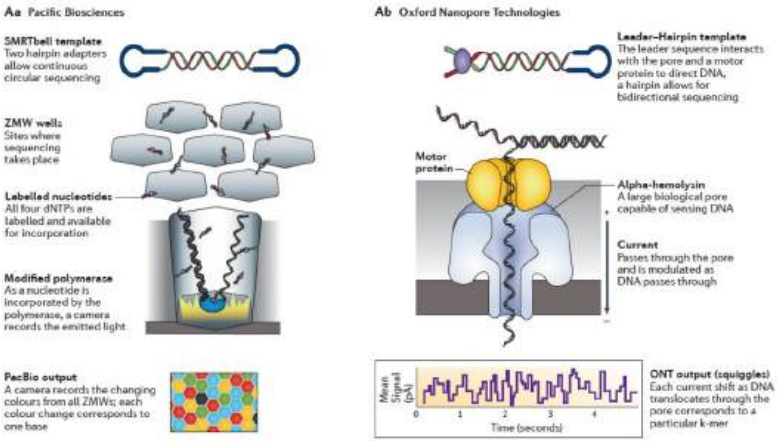


Figure 3.25: New techniques for DNA sequencing.

nanopore technique is very good since it is portable and able to sequence without genome assembly.

3.4 DNA Cloning and Expression

Definition 3.10. **DNA cloning** or **DNA recombinant technology** is a way to isolate a particular DNA sequence (gene) and amplify/propagate it.

Definition 3.11. A **vector** is any particle that can act as a vehicle that can carry a foregin DNA into another cell of interest.

The general process of DNA cloning begins with getting a DNA fragment of interest and combining it with a vector called **plasmid**. The combination of this DNA and vector will form a **recombinant DNA**. This recombinant DNA will be inserted into the host cells (since we're using plasmid, the host would be bacteria). When the bacteria replicates, the recombinant DNA also replicate along side it. Once this is done, we have a large sample of recombinant DNA that we can separate, sequence, and even manipulate.

3.4.1 Integration of DNA to Plasmid

Definition 3.12. A **plasmid** is a small circular extrachromosomal dsDNA. They're usually found in bacteria.

The reason for this is that it is easier to integrate a DNA fragment into a plasmid than integrate it into a chromosome. It can also replicate before cell division. In order to integrate the DNA fragment, we must first cut the plasmid open (since it is circular). The enzyme that carry out this cutting is called **restriction endonuclease (enzyme)**. This restriction enzyme will make a *staggered/symmetric cut* at a specific sequence.

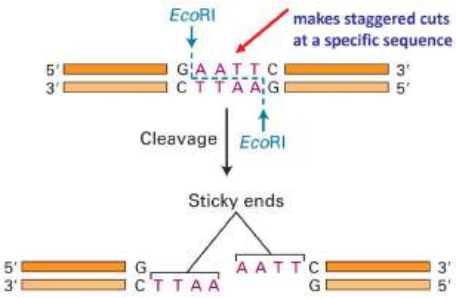


Figure 3.26: A type of restriction enzyme is called EcoRI that can cut the plasmid at the sequence AATT

The specific DNA sequence that restriction enzymes would be cutting is called the **restriction site**.

Definition 3.13. The resulting 2 ends of the restriction enzymes' cutting is called **sticky-ends**.

Remark 3.16. *All sticky-ends cut by the same restriction enzyme are complementary (can adhere to each other).*

This also means that the DNA fragment would have to be cut with the same restriction enzyme to have the same sticky ends. After this, the DNA fragment and the plasmid's sticky-ends would anneal with each other; we introduce T4 ligase to catalyze the formation of backbone between the 2 again.

Definition 3.14. A **polylinker** or **multiple cloning site** is a DNA sequence on the plasmid that the DNA can be inserted.

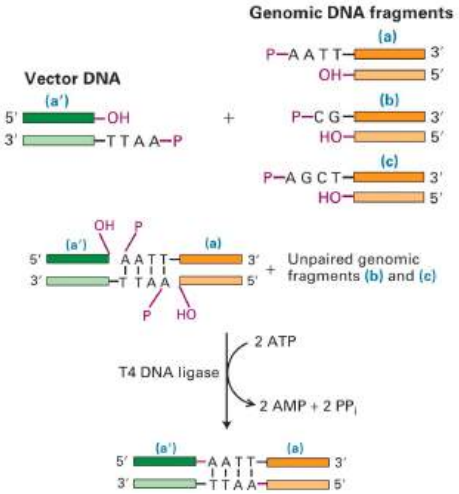


Figure 3.27: Only the sticky ends that was cut with the same restriction enzyme can be complementary with each other.

Because the polylinker is the place that DNA can be inserted, it is also the place that contain the restriction site; and not just 1 site but multiple (hence poly-). Certain plasmid won't have this polylinker but we can engineer them to have 1. Another important aspect that the plasmid must have is an origin of replication; this is quite evident because we need this plasmid to replicate along side the bacteria.

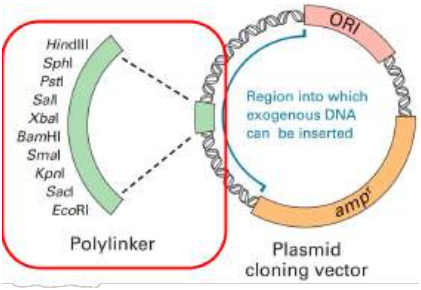


Figure 3.28: Plasmid's polylinker, origin of replication and antibiotic resistance gene (in this case is ampicillin).

Finally before allowing bacteria to take up this plasmid, we need to ask

ourselves **how can we be sure that the bacteria actually have the plasmid in them?** Well...to solve this, we also introduce a new antibiotic resistance gene to the plasmid. Once any bacteria has the plasmid, they would also have this gene which means that when we introduce them to antibiotics, all of the bacteria without the plasmid would die while the other would live.

3.4.2 Integration of Recombinant Plasmid to Bacteria

We can integrate the recombinant plasmid to bacteria (in this case: E. Coli) through the process of **transformation**. In transformation, the bacteria is mix with the plasmid and $CaCl_2$. Then we allow heat shock in the mixture, after which the bacteria will uptakes the plasmid. Once transformation is done, we culture the bacteria on nutrient agar plates containing the antibiotic that the plasmid's resistance gene is carrying (in this case: ampicillin). The transformed bacteria will survive while the rest is killed.

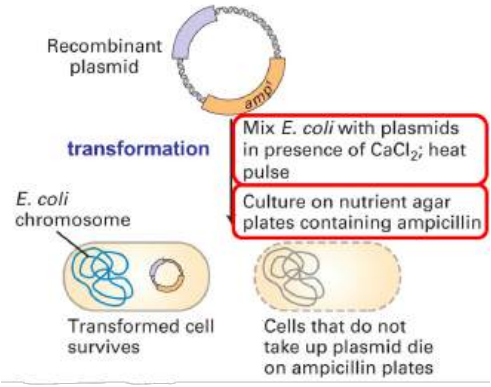


Figure 3.29: Bacteria is subjected through transformation then ampicillin.

Once the plasmid is inside the bacteria, due to its nature, it will begin to replicate first. After which, the bacteria will multiply thus creating a colony of bacterial cells having at least 1 copy of the recombinant plasmid.

Now that we know how to clone a gene, the real question to ask is **what can we do with cloning?** Well...we can use them to make a DNA library.

3.4.3 DNA Libraries

Definition 3.15. **DNA libraries** are permanent collections of genes that can be obtained and maintained.

There are mainly 2 types of DNA libraries: genomic library and cDNA library.

Genomic libraries

Definition 3.16. **Genomic libraries** are DNA libraries that consists of all of the chromosomal DNA of an organism.

It involves first digesting/cutting up the genomic libraries into smaller segments. The restriction enzyme used is called Sau3A which would cut at a restriction site of 4 nucleotide long. These restriction sites are scattered through the genomic DNA which means that the cutting would be randomized but inclusive. On the side of the vector, we use a restriction enzyme called BamHI which would cut at a restriction site of 6 nucleotide.

Remark 3.17. *We don't use BamHI for genomic DNA since cutting restriction site of 6 nucleotides would miss certain DNA fragment as compared to Sau3A with only 4 nucleotides.*

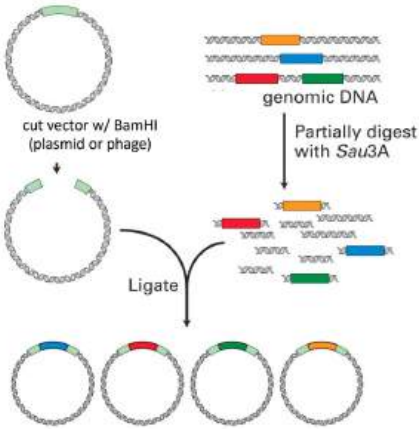


Figure 3.30: The process of making genomic library

Remark 3.18. *Sticky-ends of BamHI cut is the same as those from Sau3A cut.*

Once the cutting is done, the genomic DNA fragments are ligated to the vector and let bacteria propagate them creating the genomic library. In fact, in the early days, in order to sequence any genome, we must first create a genomic library which we would take 1 of the DNA from it to sequence.

Remark 3.19. *Genomic library is not used anymore since the we already have genomes sequences available which we can use PCR to amplify our gene of interest.*

cDNA library

Unlike genomic library, cDNA library is much more common and is still used widely till this day.

Definition 3.17. **cDNA libraries** are DNA libraries that represents mRNA that's present in a given sample.

The very first step into making cDNA library is hybridize mRNA with **oligo-dT primer** (an RNA sequence full of T). The reason that we can add a long sequence of T since mRNA will go through a process of maturation where poly-A tails are added. This oligo-dT primer will acts as a primer for an enzymes called **reverse transcriptase** to make cDNA.

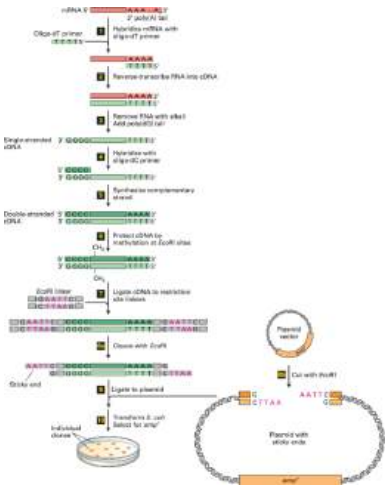


Figure 3.31: Process of making cDNA library.

We then remove the RNA template making a ss"cDNA" with added poly-G tails. Then a complementary strand will be synthesize which then make a ds"cDNA". This new double stranded cDNA will have EcoRI linkers added and spliced to make the sticky-ends. Plasmid vectors would also be cut with EcoRI as well. Lastly, the process of DNA cloning is perform thus creating a cDNA library.

3.4.4 Application of recombinant DNA

We can use them to determine when a certain mRNA is expressed, which genes is regulated together and even where those genes/mRNA are expressed.

In situ hybridization

Definition 3.18. In molecular biology, **hybridization** is a phenomenon where ssDNA or RNA can anneal complementary with their corresponding cDNA or RNA

In situ hybridization is a type of hydridization that used *labelled* recombinant DNA to localize a specific DNA sequence.

The process begin with making the specimen permeable to the recombinant DNA or RNA (which is called **probe**). These probes have been labelled and can be recognized by antibodies. The specimen will be treated with protease and detergent to exposed the mRNA to the probes. We then let the specimen incubate with the probes in environment that promotes hybridization. Then the specimen will be incubate again with antibodies that once bind to the probes can produce a coloured reaction.

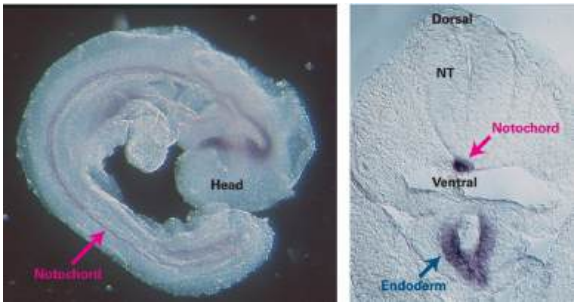


Figure 3.32: We can use *in situ* hybridization to see the expression of the Sonic hedgehog genes in the notochord, head and endoderm of a mouse embryos of 10 days.

DNA Microarray and Cluster Analysis

Using the understanding of hybridization, we can now monitor multiple gene expression simultaneously. This monitoring of many gene expression at one is called a **DNA microarray analysis**.

It uses the hybridization of labelled (fluorescent) cDNA to genes of interest on an array. When these cDNA anneal to the genes on the array, depending on the how much the gene is expressed, the colour would reveal the amount of expression.

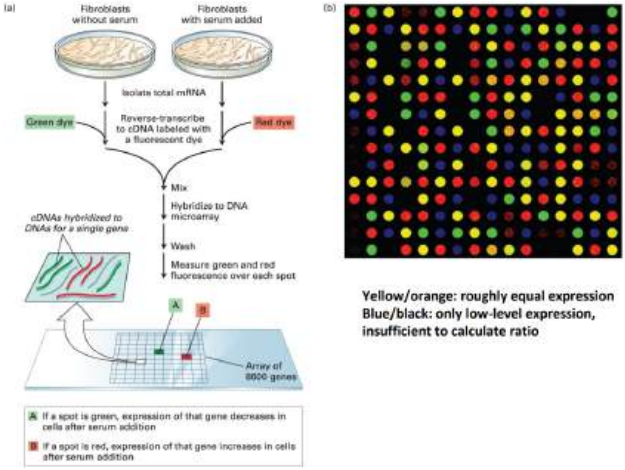


Figure 3.33: Microarray analysis with 2 different cDNA of different conditions.

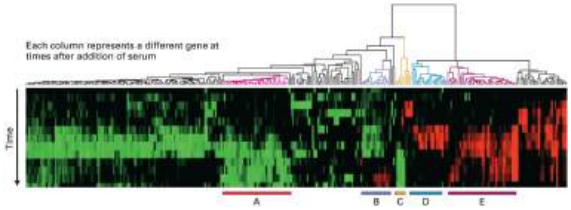


Figure 3.34: Cluster analysis

Cluster analysis is a statistical method of grouping objects that are similar to each other. In this case, these objects are similar genes and is expressed as different colour.

Production of eukaryotic proteins

We can also incorporate eukaryotic genes into the plasmid to make the bacteria make the eukaryotic protein. In this case, we can replace the LacZ gene that lies next to the Lac promoter (when exposed to lactose, transcription of LacZ gene begins) of the bacterial plasmid (make $\beta-$galactosidase) with G-CSF cDNA of eukaryotes. After this recombination, when this bacteria is exposed to lactose, the bacteria won't produce $\beta-$galactosidase but it will produce G-CSF proteins.

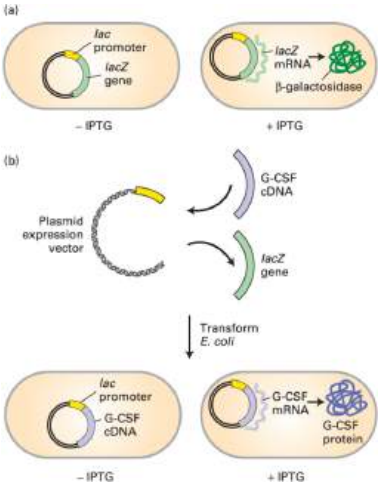


Figure 3.35: Production of eukaryotic proteins

Plasmid Expression and Stability

The expression of plasmid can be **transient** (short-term) or **stable**.
For short term, during the process of cell division, the plasmid is designed so that some of the daughter cells won't have the plasmid. This would cause a diminished plasmid later down the generation. The process of plasmid expression that is short term is called **transient transfection**
For long term, during the process of cell division, all of the plasmid is distribute to all of the daughter cells (normal DNA cloning). This process is called **Stable transfection**.

Finally, as we've said before, integration of DNA into chromosomal DNA is quite difficult however, this process can be mediated with **retroviral vec-**

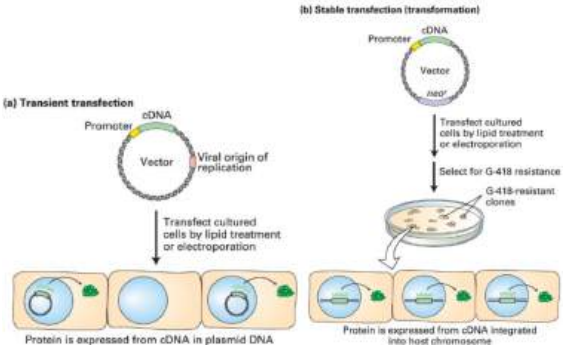


Figure 3.36: Transient and stable transfection.

tors. The retroviral vector can take a clone DNA and incorporate to mammalian genomes (chromosomal DNA).

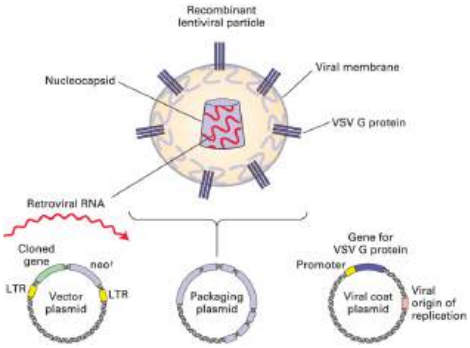


Figure 3.37: Retroviral vector and DNA integration

3.5 Genomes and Transposable Elements

Definition 3.19. Genome is an organism's entire hereditary information. They're made of mostly DNA (except certain virus that has RNA genome).

Remark 3.20. *The total of all DNA encoding for the organism is the* ***genomic library****.*

When talking about the size of genome, we would like to mention a

measurement that's commonly use: the base-pair (bp) count. As the name implies, it is the amount of base pair e.g. 1000bp = $1kbp = 1kb$ etc.

Remark 3.21. *Biological complexity is not related to the amount of bp e.g. human has around* $3300Mb$ *while a simple tulip would have* $33000Mb$.

Sequencing an animal genome allow scientist to study about their past species, the environment and other species that live along side it. As of 2021, we were able to sequence the genome of over 3278 animal species! This looks like a big number but compared to > 2 millions species on Earth, this number if miniscule (~ 0.2%).

Remark 3.22. *The largest sequenced genome is that of the* ***Australian lungfish Xiphophorus****. Its genomic size (DNA content) is around* $43Gb$ *but even then most of it is transposable elements instead of genes.*

3.5.1 Genes

Definition 3.20. **Genes** are DNA sequences that encodes the synthesis of a functional products (polypeptides or RNA).

Example 3.5.1. Certain genes can encode from rRNA and tRNA that does not have to go through translation

With the above definition, we can think of genes as **transcription units** which includes all of components for an mRNA to be made. This transcription units consists of 2 parts: control regions and the open reading frame (ORF).

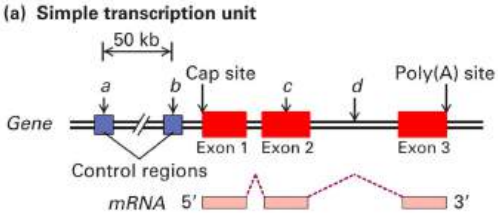


Figure 3.38: A simple transcription unit and its matured mRNA

The control regions measure around 50kb and involve in controlling gene expression. One of the sequence in them is called the **promoter** (where transcription start); the other control regions have **cis-regulatory factors**

that can regulates the gene right next to the control region; while **trans-regulatory factors** regulates distal gene expression.

The ORF are region has the information that specifies the amino acids sequence. It has in it 2 sub regions: introns and exons. **Introns** are DNA sequences that are not used in the translation and is spliced out when mRNA is matured, while **exons** is the opposite. Introns however can have regulatory elements as well as alternative splicing. The matured mRNA will be then used to make amino acids sequence which in turn make proteins.

Remark 3.23. *Proteins with similar functions often contain similar amino acids sequences that encode the functional domain.*

Using the above remark, we can in fact compare the amino acid sequence for a protein from different species. There's a tools used for this purpose called: **basic local alignment search tool (BLAST)**. It will align amino acid sequences with already known amino acid sequence on the database in the most optimal way. BLAST will give it a high score if amino acids matche and a low score if the amino acids are related.

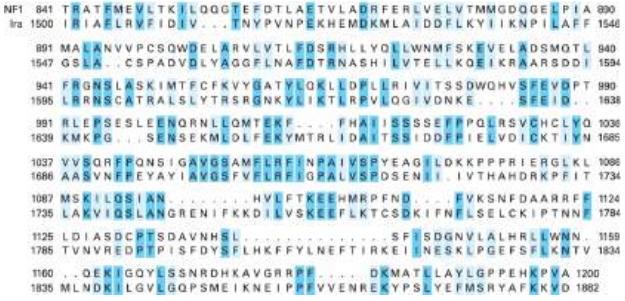


Figure 3.39: Comparison of regions of the NF1 proteins of human with *S. cerevisiae* Ira proteins using blast. As we can see, there are many similarities.

Remark 3.24. *We don't use DNA sequence for this because of degeneracy in genetic code.*

Although the difference in genome size between species is enormous yet the difference in proteins coding gene is very small (smaller variation). This also means that the difference in genome size is due to the different amount of non-coding genes and transposable elements. Therefore the density of genes is much more greater in lower eukaryotes than complex

ones because the genome size of lower eukaryotes are tends to be smaller than complex ones. (as well as not having non-coding RNA regions)

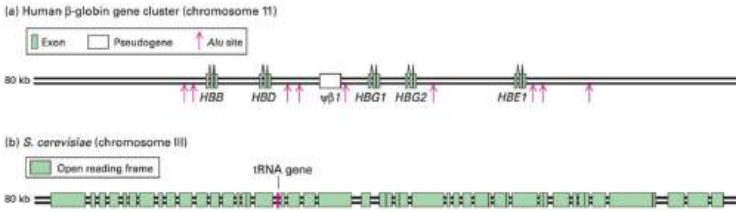


Figure 3.40: Gene density of human vs S. cerevisiae

We can in fact compare the gene density of human and the S. cerevisiae. What we find is that human has little OFR, lots of non-coding gene and **pseudogene**, which is a gene that consists of mostly stop codon so it isn't expressed much. On the other hand, the S. cerevisiae has much more closely packed ORF.

Evolutionary Discovery

Definition 3.21. **Orthologs** are same proteins in different species while **paralogs** are closely related proteins in the same species.

We can use BLAST to compare genes of different species to see their evolutionary relationship. In this example, we compare the genes of different species that encode for tubulin α and β.

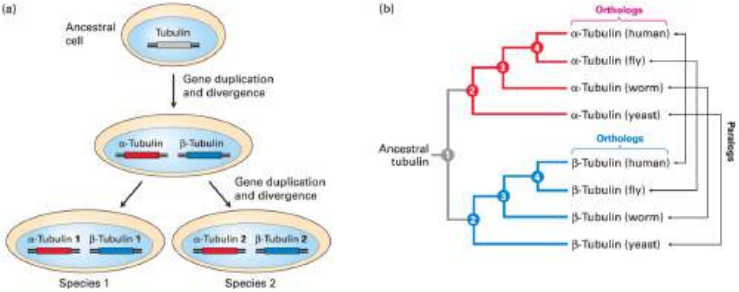


Figure 3.41: Tubulin of different species and their evolution.

What we find is that the tubulin (not α nor β) has originally came from 1 common ancestor. As the first evolutionary event occur, the tubulin protein split into α and β which are paralogs to each other. Then each of the

2 branches of tubulin will split into tubulin of different species that are orthologs to each other.

As for proteins that are paralogs to each other, tubulin isn't the only one. In fact, we have many genes that can encode proteins that are homologous

Definition 3.22. A set of duplicated genes that encodes for homologous proteins with similar yet non-identical amino acid sequence is called a **gene family**. If a protein only have 1 gene that encodes for it, it is said to be **solitary**.

Remark 3.25. $25-50\%$ *of protein-coding genes are solitary while the rest is part of a gene family*

There could be several genes of different species that can encode for the same proteins (with minor changes here and there), due to evolution.

3.5.2 Satellite DNA

Alongside with introns, there are also non-coding sequences that are long tandem of short repetition of sequences.

Definition 3.23. **Satellite DNA** or **simple-sequence DNA** are repetitive DNA sequence

There are 2 types of non-coding genes: microsatellite DNA and minisatellite DNA.
The first type of non coding gene is **microsatellite DNA**, which are repeating units of 1-4bp in length and can tandem up to 600bp that can be found in transcription units. The number of repeat can expand because due to its instability during replication as well as **backward slippage**. The expansion can cause many neuromuscular diseases such as *myotonic dystrophy* and *spinocerebellar ataxia*, etc.

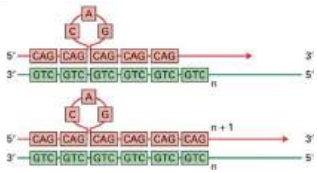


Figure 3.42: Backward slippage and addition of repetition.

Remark 3.26. ***backward slippage*** *is a mutation where the nascent daughter strand "slips" backward by 1 repetition which cause and extra repetition to be added.*

The second type is **minisatellite DNA** which is a longer microsatellite DNA, with 14-100bp in each repeating units with 20-50 repeating units. they're located mostly in the centromeres and telomeres (center and ends of a chromosome respectively). Minisatellite varies extensively in length for different individual which means they are useful for identification such as paternity determination or criminals since they vary from individual to individuals.

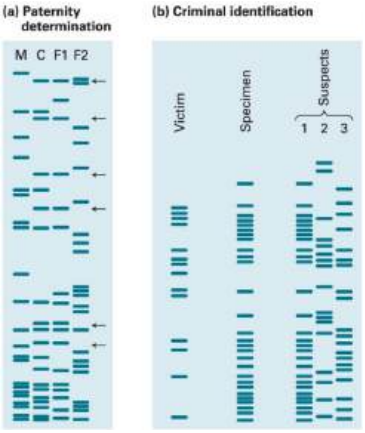


Figure 3.43: Minisatellite and their use for identification.

3.5.3 Transposable Elements

Definition 3.24. **Transposable elements** or **transposons** are DNA segments that can move from 1 place of the genome to the next.

They can influced evolution and cause mutation that lead to disease. There are 2 classes: DNA transposon and retrotransposon. (see Figure 3.44)

DNA transposons

In the DNA transposons, they will cut themselves out of the DNA sequence and reinsert else where in the genome (think of cut [CTRL-X] and paste

[CTRL-V] on a computer). The number of DNA transposon can increase when it migrates during DNA replication.

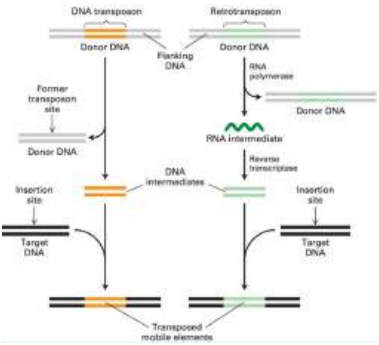


Figure 3.44: 2 types of transposable elements

If the transposon migrate from a region that was already translated to the region in front of the replication fork, then the transposon will be replicate the second time thus increase the amount of transposons.

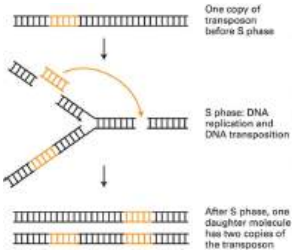


Figure 3.45: Migration in front of the replication fork increase the amount of transposons.

Retrotransposons: LTR transposons

A type of eukaryotic retrotransposon we will be first looking at is **LTR retrotransposon**. It consists of of a **long terminal repeated (LTR)** non-coding strands at each ends, and a protein-coding region. These coding regions encodes for proteins that are shared with **retrovirus** such as reverse transcriptase. The only different it has from retrovirus is that it does not encodes for envelope proteins. Because of this close relation to retrovirus, it sometimes has the name *retrovirus-like elements*.

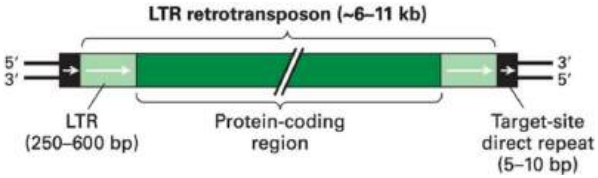


Figure 3.46: LTR retrotransposons

The way transposons make its RNA is similar to the making of retroviral genomic RNA from an integrated retroviral DNA. First, its LTR will initiate the process o transcription. Then the new RNA will be processed and matured (by adding poly A tails and cut certain elements out) to become retroviral RNA genome.

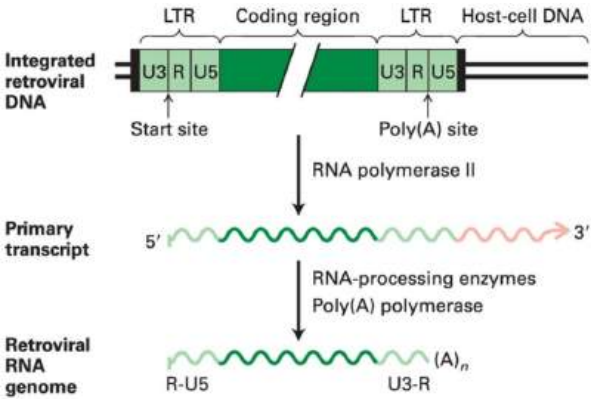
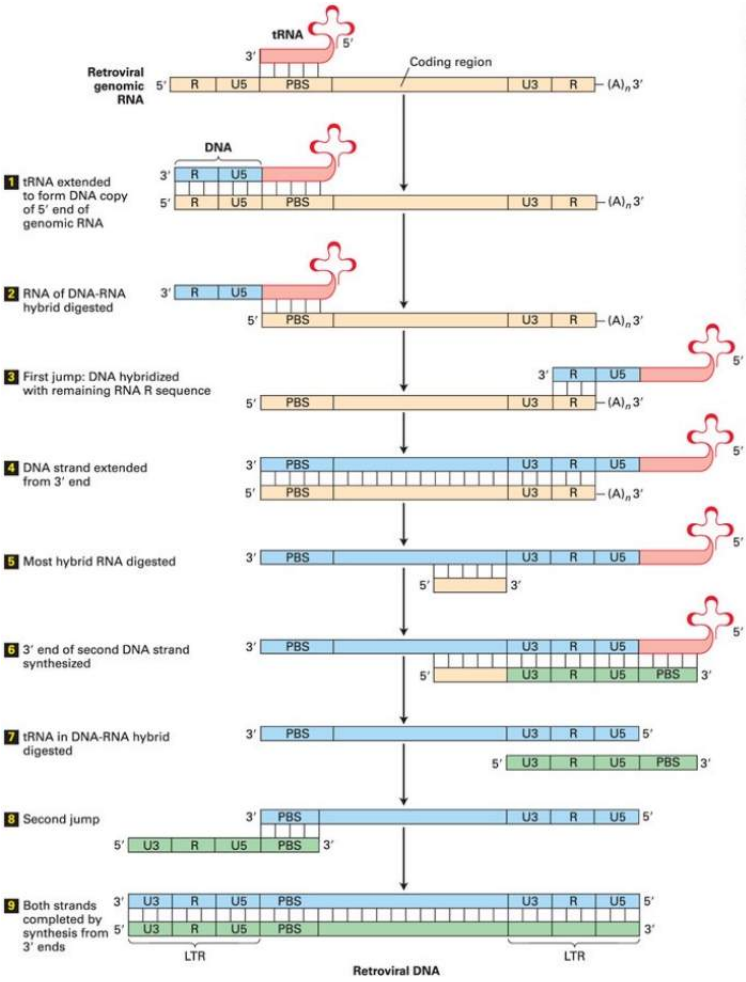


Figure 3.47: Making retroviral genomic RNA.

The process of LTR transposon reverse transcription is also similar to that of retrovirus. The reverse transcription is long and take multiple step to ensure the correct creation of the retroviral DNA which would be inserted to the host's genome later on.

The following page describes the entire process of reverse transcription of retrovirus.


tRNA
Retroviral genomic RNA
Coding region
DNA
1 tRNA extended to form DNA copy of 5′ end of genomic RNA
2 RNA of DNA-RNA hybrid digested
3 First jump: DNA hybridized with remaining RNA R sequence
4 DNA strand extended from 3′ end
5 Most hybrid RNA digested
6 3′ end of second DNA strand synthesized
7 tRNA in DNA-RNA hybrid digested
8 Second jump
9 Both strands completed by synthesis from 3′ ends
R
U5
PBS
U3
(A)n
LTR
Retroviral DNA

Retrotransposons: LINEs

The last type of retrotransposons that is not a viral-related is called **long interspersed elements (LINEs)** along with its subtype **short interspersed elements (SINEs)**. In humans, there are roughly 900,000 LINEs, each is 6kb in length, creating a total of 54Mb (2% of the human genome). On each LINEs are 2 regions of ORF: ORF 1 and 2. **ORF 1** encodes the proteins that helps to transport LINEs RNA while **ORF 2** encodes reverse transcriptase and nucleases.

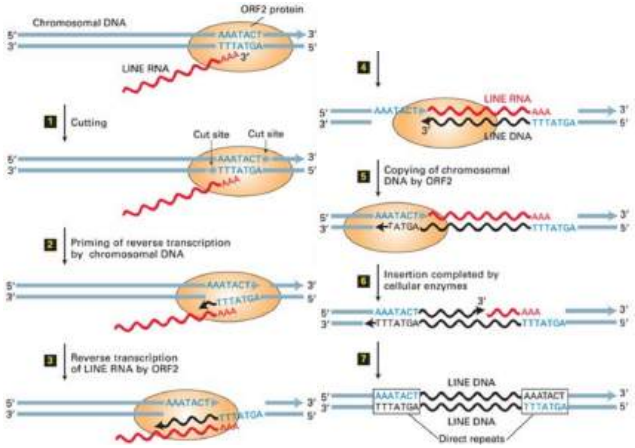


Figure 3.48: Reverse transcription and integration

3.5.4 Transposable Elements and Their Effects on Evolutions

The elements are repetitive and their movement would lead to genomic changes. Rarely, there would be recombination event between repeated elements which means that exons can be shuffle around making new genes with the same combination.

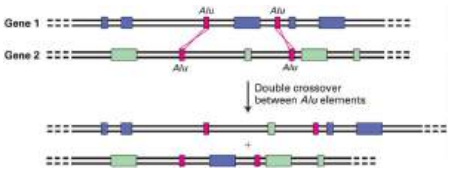


Figure 3.49: Recombination events

Another way transposons can drive evolution is through **unrelated flanking**. This is when the transposons splice itself from the DNA sequence and accidentally take some unrelated DNA sequence with it and insert it to a different gene creating a new gene sequence.

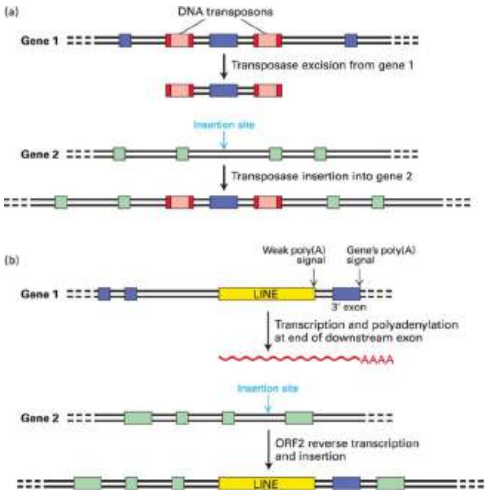


Figure 3.50: Mechanism of unrelated flanking.

3.6 Chromosomes

Definition 3.25. A **chromosome** is a long characteristic linear DNA sequence of a species' genome.

Remark 3.27. *If we were to relate genome to be a book then chromosomes are chapters while genes are each paragraph of the chapter and DNA are each words.*

Chromosomes are never alone in its DNA form but are packed as a DNA protein complex called **chromatin**. The special feature of chromatin is enabling the long chromosome to be condensed. What we meant by that is: a typical chromosom in its linear form (not condensed) is 5cm long. This might not be significant for us but this is 5000x longer than the width of a cell! Having chromatin allow chromosomes to condensed into the nucleus of a cell.

Definition 3.26. **Interphase** is a mitotic phase where the cell spend most of its time in. At this phase, the cell grows, replicates its chromosomes and prepares for cell division.

One would think that the chromosome is not packed tightly during interphase but actually it would still have some level of condensation. However, during this time, the chromosome is still usable and can be replicate.

Definition 3.27. **Metaphase** is a mitotic phase where the cell lines up its chromosomes and divide into 2 (for cell division).

Chromosomes are very densely packed during metaphase and are generally unusable (too dense for anything to occur). The tightly packing during metaphase is to faciliate an equal division of chromosomes for the 2 daughter cells. The chromatin will unwind a bit during the transition of metaphase to interphase.

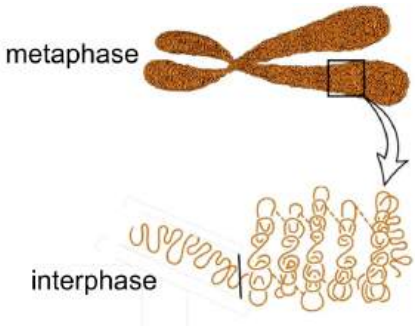


Figure 3.51: Chromosome during metaphase and interphase.

Definition 3.28. Like we've said before, **chromatin** is a complex of eukaryotic DNA and proteins.

When we look at a chromatin fiber, it is organized in loops called **topological domain** with are separated by **boundary elements**, which are non-looped chromatin fiber. Each of these topological domain has regulatory sequence tend interact with other regulatory elements in the same domain. If we open it more, we can see DNA molecules organized into characteristic **nucleosomes**, which are DNA in association with a protein complex called **histone** (an octamer). Histone is therefore the major protein of the formation chromatin.

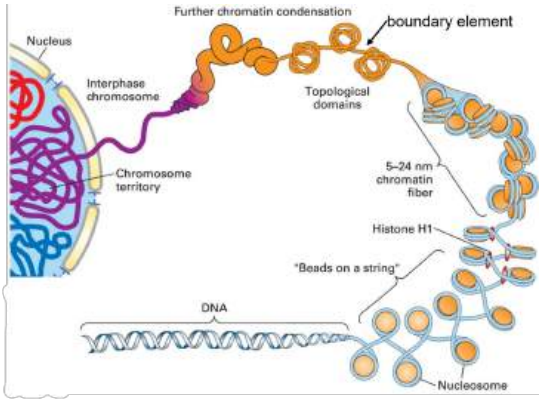


Figure 3.52: Organization levels of chromosome.

Most chromosome are a single dsDNA molecules. Some organism have DNA replication without separation which in turn create giant and branched chromosomes. One of these chromosomes we will look at are polytene chormosomes.

3.6.1 Polytene Chromosomes

Definition 3.29. **polytene chromosomes** are large chromosomes found in the salivary glands of most insects.

They have 1000x the normal DNA content in a normal chromosome, which is evident by the size different. They're also thick and long and have pattern of dark stripes. These stripes are folded sections forming **dark bands** (topological domain) and **light bands** (boundary elements). [See Figure 3.53]

They form through incomplete replication which is when the replication fork stop near the end (both direction) instead of going all the way through. After this happened, more and more replication generation happens each time the replication fork stop short, which creates many parallel identical chromatids. At the end we would get a **multi-strand chromosomes**. In fact with only 10 generation/cycles of replication, the chromosome would have 1024 different copies (strands). [see Figure 3.54]

Interphase chromosomes is very dynamics. Polytene chromosomes would puff up to show transcriptional activation. We can look at the polytene

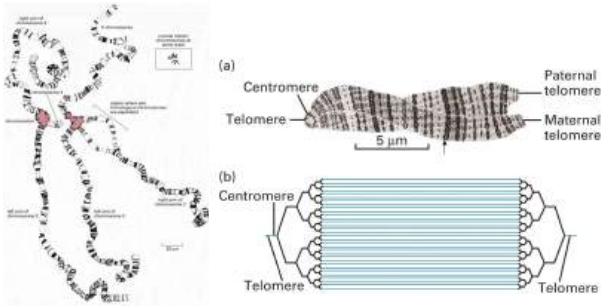


Figure 3.53: Polytene chromosome of *drosophila* (fruit fly); boxed region is the normal chromosome (left). Incomplete replication and parallel chromatids (right).

chromosomes over 22h period. When we look at the gene 74EF, after 6-7 hours, they puffed out then later they get back to normal. We can thus theorize that there are genes that are expressed at a particular time for that larvae then after the gene isn't expressed thus return to normal. In fact, the puffing up of 74EF gene is associated with RNA polymerase II activation, which means an active transcription.

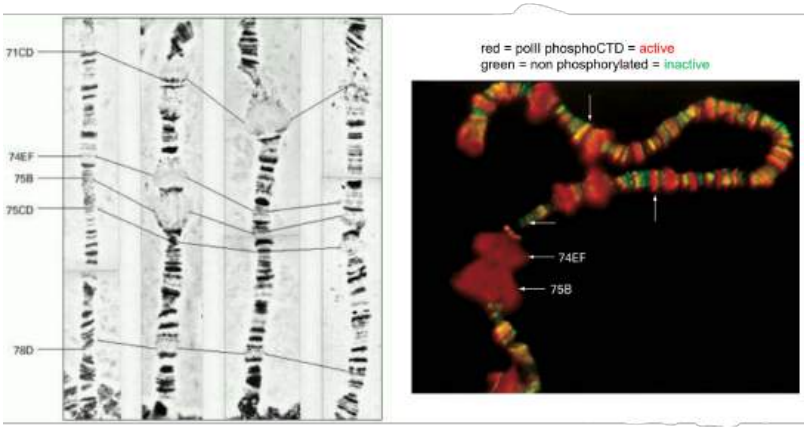


Figure 3.54: Chromosome puffing up and transcriptional activation.

3.6.2 Karyotype

Definition 3.30. **Karyotype** is the complete set of chromosomes of a species.

When looking at karyotypes of different species, we realize that chromosomes of different species would have different shape, size and number. We can look at a human karyotype, with coloration thanks to fluorescent in situ hybridization (FISH). The process of "chromosomal painting" is mediated through a panel of probes with different colour which would colour different chromosome regions in a karyotype.

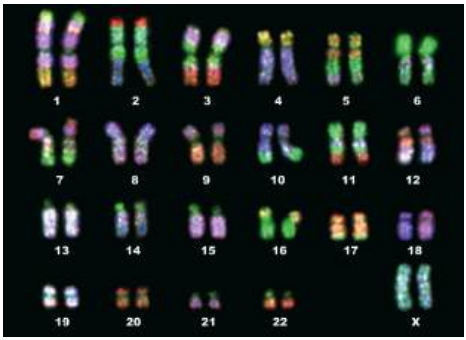


Figure 3.55: Fluorescent *in situ* hybridization.

Remark 3.28. *Each of the chromosomes pair is formed by 1 chromosome from the maternal and 1 from the paternal side.*

Remark 3.29. *It's not evident but each sister chromatid (from a chromosome pair) forms an X. This cannot be seen from fig 3.56 because typically they're very condensed.*

Chromosomal Rearrangement and Evolution

Chromosomes cannot change shape but it can break and rejoin. The process of breaking and rejoing is called **translocation** and can happens in somatic cells in the body.

Definition 3.31. **Somatic cells** are cells that reproduce asexually by mitosis while **germ cells** are cells that reproduce sexually by meiosis forming gametes.

Translocation can also induce a disease to human such as **chronic myelogenous leukemia (CML)**. The condition CML happens when the body merge accidentally chromosome 9 and 22. this would give rise to a different gene creating **oncogenic** proteins (proteins that can cause cancer).

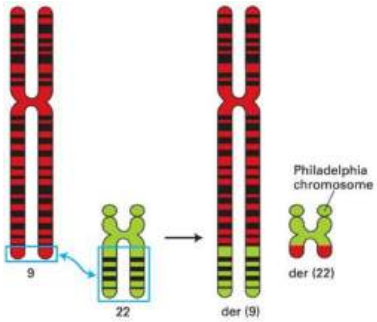


Figure 3.56: Chromosome 9 and 22 translocation.

During diagnostic testing and screening of CML, a chromosome called **Philadelphia chromosome** are actively looked for (is also the main cause of the oncogenic proteins).

Translocation can occur in germ cells which would create gametes with variation of chromosomes. The offspring from such gamete has low fertility creating a **dead-end** to the organism. However, occasionally, the translocation of chromosome creating a rearrangement variant survive the dead-end and pass on to the next generation. These rare events of successful translocated chrosomes is the basis that helps karyotypes evolve overtime. In fact, that is how human chromosome came to life. The chromosomes of primates and homo sapiens, are similar but we can see that the homo sapiens chromosomes is simply the translocation and even break into new chromosomes from the original primates chromosomes.

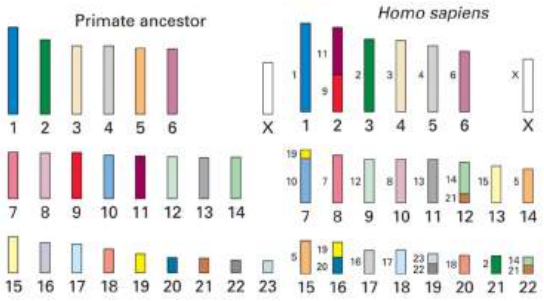


Figure 3.57: Primate ancestor chromosomes and the evolved homosapiens chromosomes

3.6.3 Elements of Linear Chromosomes Replication and Stability

We will now look at the process of making artificial chromosomes from already existed DNA. To sucessfully perform this, we would be requiring the following important elements: origin of replication, centromere and telomeres.

In this experiment, yeast leu$^-$ cells have LEU gene inactivated by a mutation thus require external recombination to re-activate them.

Origin of Replication (OR)

We've cloned the normal yeast LEU gene onto a plasmid. We then integrate this plasmid into the yeast for replication. It turns out that it doesn't replicate well in yeast. This is because origin of replication from bacteria (prokaryotes) is different from that of yeast (eukaryotes).

Definition 3.32. The yeast's origin of replication is also called **autonomously replicating sequence (ARS)**.

Supposed that we can insert a random piece of a yeast DNA that contains the yeast's OR. Then when the plasmid is reintroduced into the yeast, the plasmid starts to replicate!

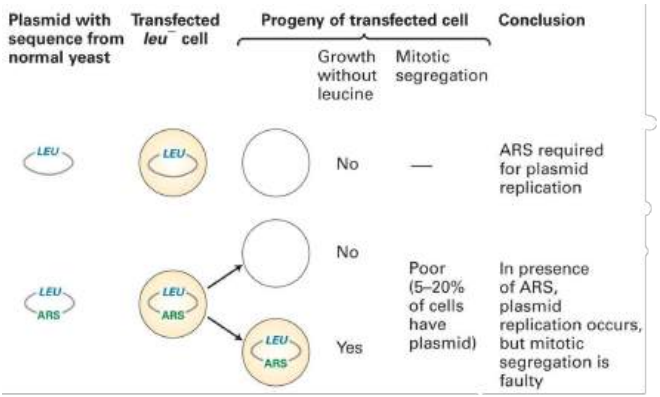


Figure 3.58: Yeast origin of replication and troubles

However...there is a problem, the cell is replicating but it doesn't distribute the plasmid to the daughter evenly. **What could potentially cause**

that?...well, from our understand of chromosomes, during metaphase, the division of chromosomes must be even for both daughter cells to receive the same. Therefore there must be a problem with *mitotic segregation*. How can we improve that?

Centromere

We can insert random strand of the DNA that was replicated in both daughter cells to our plasmid. When we inserted this DNA in, both daughter cells were able to have the LEU genes. Turns out that DNA is the centromere sequence (CEN) of the yeast.

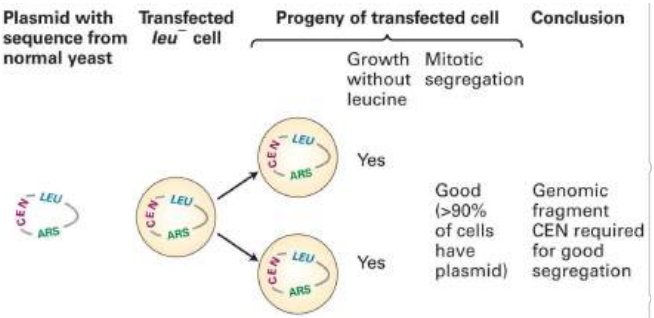


Figure 3.59: Yeast centromeres

Mechanism: The centromere has an apparatus called **kineticore** that the **microtubules spindles** can attached to and do mitotic segregation. When several centromeres are cloned, they all have generally the same sequence. A histone variant called **CENP-A** is only presence in 1 special type nucleosomes called **centromere nucleosomes**. The CENP-A in the centromere nucleosomes allow it to recruit the CBF3 complex which in turn recruit the Ndc80 that is attached to the microtubules.

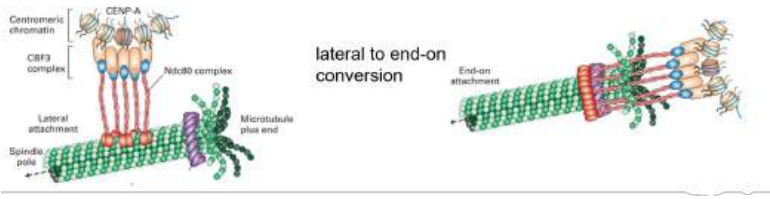


Figure 3.60: CENP-A recruit CBF3 which recruits Ndc80+microtubules.

Back to the yeast however...its chromosomes are not circular like plasmid but linear. It's possible to cut open the plasmid but once cut, the plasmid won't work like a linear DNA. how can we solve this?

Telomeres

We need to add yeast **telomeres (TEL)** to each ends for the LEU plasmid. Once added, we see that the cut plasmid will now function like a normal chromosome, restoring the yeast ability to produce leucine.

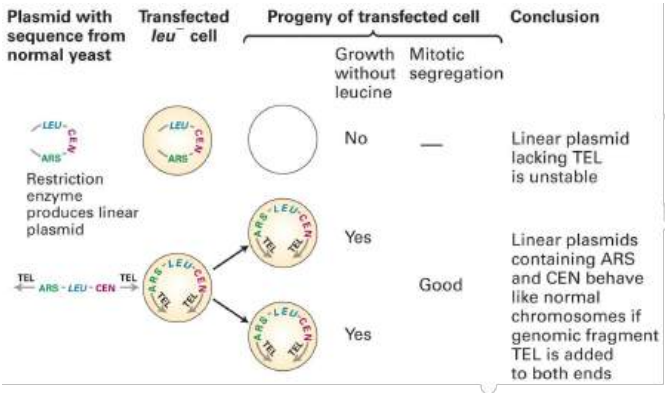


Figure 3.61: Yeast telomeres and functional chromosome.

Telomeres have many function such as protection from exonuclease, end-to-end fusion and solve a replication problem.

Telomeres and Replication Problem

When replication reach the end of a chromosome, the leading strand is made by DNA pol in 1v continuous fashion thus the leading strand is complete. On the lagging strand however, there will be a missing DNA strand at the end; this is because when we made the RNA primer at the end, it will be later removed. Telomeres can be used to solve this problem. [see Figure 3.62]

Telomeres has sequences that is simple and repeated. It works with a reverse transcriptase called **telomerase** which use RNA template that is complement to the telomeric DNA repeating sequence. It will extend the template strand then will "slip" over to the new extension; in which the

old extension would extends the template strand with the telomerase sequence then it would slip again then the process repeat so on. [see Figure 3.63]

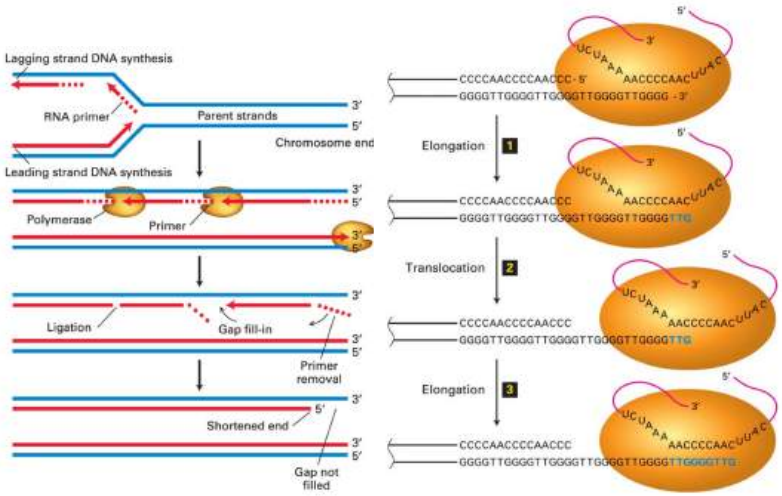


Figure 3.62: Telomere replication problem (left). Telomerase and resolution to the replication problem (right).

Telomerase are only active in germ and stem cells. Somatic cells doesn't require a lot of telomeric repeats since it only divide a few time.

Remark 3.30. *Telomeres are often re-activated in cancer cells. They're thus one of the target for cancer therapy.*

Telomerase are important but **what would be the consequence of not having them?** Well...Mice lacks telomerase are fine for 3 generations, after which fecundity/fertility begins to decline.

3.7 Review Questions

Does RPA have proofreading capabability?
No it doesn't,

Is there any reason δ and ε can't extend RNA primer? Yes, we won't talk about it.

What is the different roles of alpha, delta and epsilon polymerase Pol alpha is can synthesize from the RNA primer, delta would replace RNA with DNA as well as is important for Okazaki frag and epsilon is the main pol that synthesize the leading strand. Delta and Epsilon has proof reading.

How is RNA primer removed from the leading strand? FEn1 and ribonuclease H are only in the lagging strand?

Remark 3.31. *There is no real single leading or lagging strand but it is an either leading or lagging strand mechansim.*

Because of this, FEn1 and ribonuclease H would remove all primers.

Part II

Eukaryotic Transcription and Translation

Chapter 4

Nucleic Acid Detection and Quantification

The use of biological tech has evolved and pushed our understanding down to the nucleic acid level. We now have methods to analyze and characterize DNA and RNA. These methods look at the qualitative aspect (sequence, types of molecules, structures, etc.) and quantitative aspect (level of gene products, tumour growth, etc.).

Remark 4.1. *When we can separate and treats diseases according to its quantitative aspect, we've made the first step into personalized medicine (treating is more specific instead of general average).*

Typically, these molecules that we're actively finding for an analysis are presence in minute quantities and is hard to detect. Nevertheless, researcher has designed method to detect these molecules at very weak level (concnetration). 1 important discovery is the **molecular probe** which are nucleic acid that can complementary base pair to a nucleic acid we want to investigate. The probe would allow us to localize and numerize this nucleic acid.

4.1 General Procedure of Probes

Using electrophoresis (or even immunobot), we can separate the nucleic acid and transfer it to a **solid state membrane** (nylon, nitrocellulose, etc.). We then hybridize the membrane with the probe allow it to bind to the target nucleic acid.

Definition 4.1. Remember, **Hybridization** is the process of incorporate a probe into a target DNA or RNA.

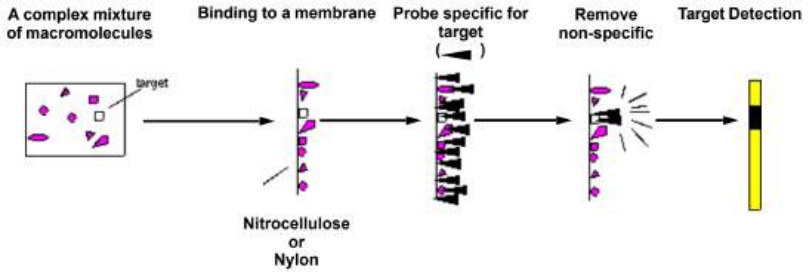


Figure 4.1: Mechanism of Probes

After, we wash to remove all the unrelated (non-specific) binding and we'll be able to detect the target nucleic acid.

4.1.1 Labelling Probes with PNK

Supposed we want to localized a gene sequence that we know, **how would we do so?** Well, we can use probing again but this time, along with the probe, we will label it

Definition 4.2. **Labelling** is the process of making a target molecule detectable, visually (most of the time).

In our case, the probe is the target molecule to be labelled.
To make this, we first need to synthesize an oligonucleotide that is reverse complement to the sequence.

Remark 4.2. *Complement is when the base-pairing match while reverse meaning running on the opposing direction (antiparallel).*

Then, we use **polynucleotide kinase (PNK)**. PNK will phosphorylate the oligonucleotide by transfer the $\gamma-$phosphate from ATP to the free OH-group at the 5' end. This would render our oligonucleotide radioactive hence the method is **isotopic radiolabelling**. In this case, the probe is for ssDNA.

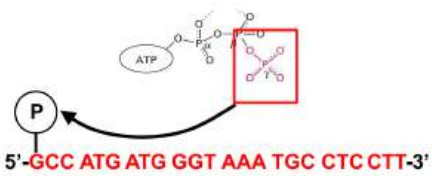


Figure 4.2: Labelling Probes by phosphorylation at the 5' end.

4.1.2 Labelling Probes with PCR

Sometimes we need dsDNA as a probe. To do so, We can synthesize dsDNA in a PCR reaction. The change that we'll make is addition of low concentration of a radiolabeled dNTP (1 specific type out of the 4). Then at the end of the PCR, we would have a probe with various radiolabelling on the dsDNA sequence.

Figure 4.3: Radiolaballing dCTPs and incorporate it in PCR of probes.

Remark 4.3. *We must be labelled at the α instead of γ-phosphate since γ will be broken off during polymerization.*

Remark 4.4. *This probe need to be broken off into single stranded to be useful.*

4.2 Southern Blot

In order to analyse DNA, we must first cut DNA into fragments using restriction enzyme. We then separate them by size using electrophoresis. At the same time, this electrophosis must be ran under **alkaline solution** which would render the dsDNA becoming ssDNA. We need to this because this allow us to insert a probe later on.

After separating we will transfer them to a solid state membrane where they can adhere strongly to. This adhesion can be made permanent by UV crosslinking. This membrane will then be hybridize with any specific probe (ssDNA or RNA) to detect a sequence of interest.

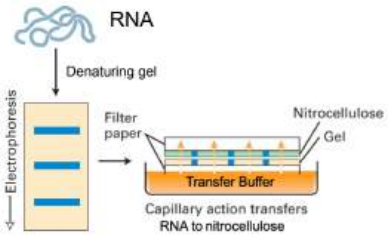


Figure 4.4: Electrophoresis under alkaline solution

After a while, the probe would be able to find its complementary base and bind to that; washing at the end would remove all the non-specific base pair. At the end, through **Autoradiography**, we will see a dark band on the membrane.

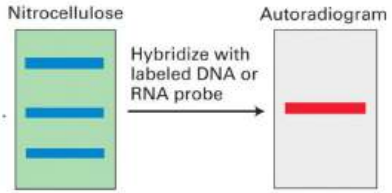


Figure 4.5: Hybridization of probes and autoradiograph

4.2.1 Southern Analysis and Polymorphism

Definition 4.3. **Polymorphism** is the existence of more than 1 form of a gene. *poly* is many and *morphism* refers to morphology or shape.

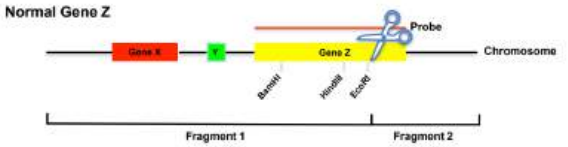


Figure 4.6: Normal Gene, no polymorphism

We can use Southern blotting along with probes (Southern analysis) to detect polymorphism. First, we would need to recognize a specific gene

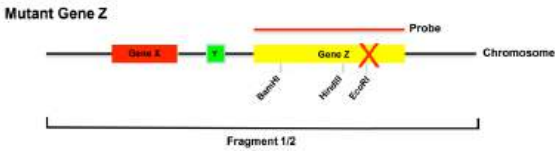


Figure 4.7: Mutant gene, polymorphism

that we want to detect for polymorphism and synthesize a complementary probe. Then we will cut the restriction site on this gene (typically is EcoRI) using restriction enzyme.
If there is no mutation on the gene site, EcoRI will cut it into 2 fragment which after southern blot and hybridization allow us to 2 band. 2 bands are caused by the probe complement the gene and even when cut, the 2 parts of the gene is still complement to at least 1 section of the probe. (see Figure 4.6)

If there was a mutation, EcoRI won't cut at the restriction site but at somewhere down the DNA that has similar restriction site. This leaves the gene being intact which means after Southern blot, we will only have 1 big band.

This method of polymorphism detection also allow us to trace back if a specific disease is genetically transmitted or not. Or better, we can look at how a specific genes are transmitted genetically.

Definition 4.4. **Alleles** are alternative variation of a gene. i.e. polymorphism of a gene.

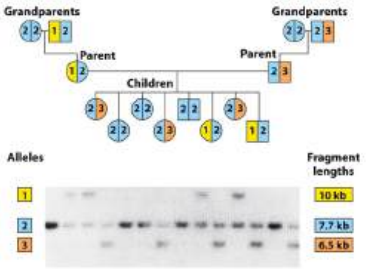


Figure 4.8: Alleles and Southern analysis

We can compare different alleles found on individual using southern

blot. What we found is that certain alleles are expressed more in individuals while others are less expressed. We tend to call the alleles that is expressed the most **dominant alleles** while the opposite is the **recessive alleles**. (see Figure 4.8)

4.3 Northern Blot

In order to analyse RNA, we first have to denature the shape of the RNA while performing the electrophoresis. The denaturation must be continuous hence we will run it in a transfer buffer. After this, we will then perform the same solid state transfer and probes hybridizatation.

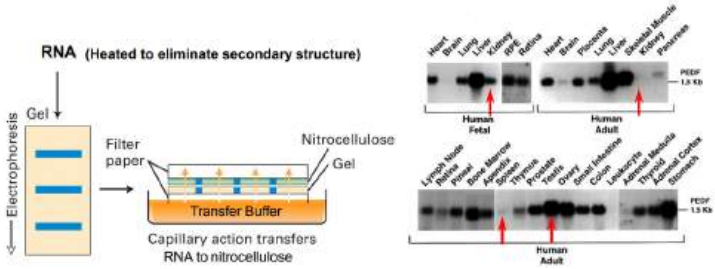


Figure 4.9: Northern blot (right) and PEDF expression (left)

Analysing RNA (typically are mRNA that will be translated into proteins) would allow us to see where the gene is expressed (spatial) but also when it would be expressed (temporal). The reason for temporal since certain gene would stop expressing (or slow down) as we reach adulthood.

Example 4.3.1. Take the proteins PEDF, we can takes its mRNA form through different stages of growth in the human body and analyze it with Northern blot to see where and when it is expressed. (see Figure 4.9, left)
What we found is that this gene is expressed a lot in the kidney during fetal/infancy stage and stop in adulthood (thus temporal); also, the gene is expressed a lot in the testis instead of the spleen (thus spatial)

Remark 4.5. *Using southern blot analysis, we can see variants of gene (alleles) related to a disease as well as its origin. Northern blot analysis tells us the size of the molecule of the sequence, RNA* ***isoform*** *(functional proteins with different gene) that is tissue specific or even state specific.*

This is great because Make quantitative conclusion but also qualitative conclusion!

4.4 RT-qPCR Method

There are new method that are good for lab and allow us to quantify specific transcript level. Most efficient quantifying method is using **RT-qPCR**. The process is divided into RT and PCR. RT or *reverse transcribe* is the process of making DNA from RNA; after, we can then coupled this process with PCR to amplify the DNA (which is equivalent to the gene).

Remark 4.6. *PCR usually exponentiate very fast then plateau overtime due to lower amount of "ingredient". The amount of cycle (threshold "time") that PCR can go through before plateau is direct dependent (or proportional) to the amount of starting material.*

We begins with the RT reaction. mRNA is converted to cDNA (complementary) by priming the poly A tail with poly T primer. We then remove the original mRNA and anneal a poly G primer adapter to the 3' end to the sscDNA. A poly C primer is used to initiate the synthesis of the second DNA strand making it dscDNA again.

The above is the general mechanism of how RT reaction. What is interesting about such reaction is that the primer used can change which mRNA were making cDNA i.e. If we carry RT reaction with poly T primer, it would prime all mRNAs (because of poly A primer). But if we carry RT reaction with a specific sequence primer, then only some of the mRNAs with that specific sequence will be RT to cDNA.

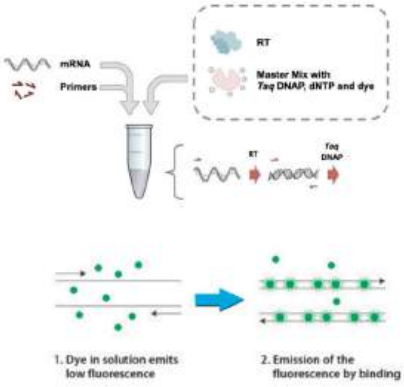


Figure 4.10: Mechanism of RT-qPCR

We can now perform PCR reaction with a dye. This dye will allow emission of low fluorescence. Every cycle, the dye will be integrated into the DNA more and more making it more visible i.e. The more DNA is amplified through PCR, the more fluorescence signal is detected.

Remark 4.7. *This is also the reason there is a "q" in the RT-qPCR. The q stands for "quantification cycle" i.e. we would be able to quantify how much DNA is made each PCR cycle.*

4.5 RNA-seq

RNA-seq is a short form of *RNA sequencing*. As the name suggested, this new method allow us to sequence RNA. Not only that it allow us to determine the entire global gene expression. It is fairly simple, first, we will extract and purify the total RNA from a tissue population or a cell. Then, we will convert these RNAs to cDNA to make cDNA library.

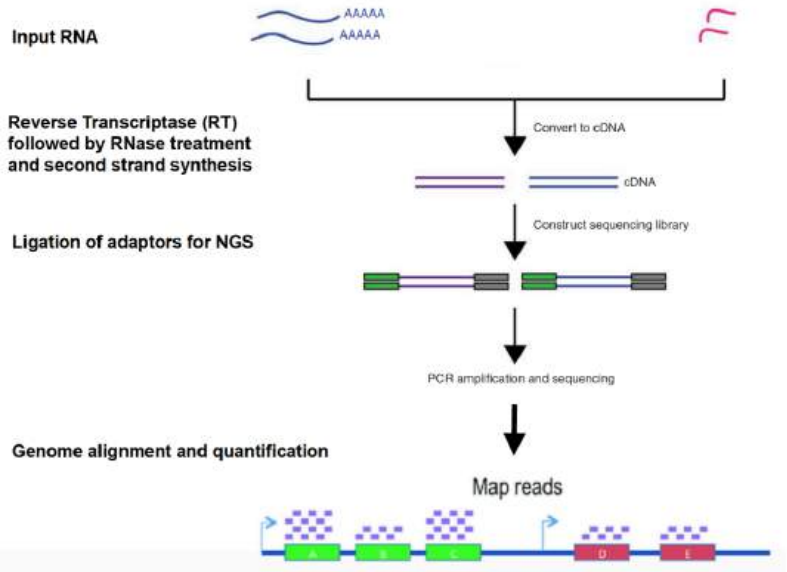


Figure 4.11: RNA-seq mechanism.

NGS adaptors are added to each end of the cDNA; after NGS (next-generation sequencing) is carried out.At the end, we will get a myriad of

sequencing from each RNA we've extracted from the original cell or tissue population. Finally we will recombine these sequencing from NGS, with the help of computer, for the entire genome of the cell/tissue.

Remark 4.8. *The only "downside" to this method is that it isn't as specific as the other 3 above method.*

Chapter 5

Eukaryotic Transcription and Control

Definition 5.1. **Transcription** is a process of making RNA from a DNA template.

Most of gene expression encompassed by rates of transcription i.e. the amount of gene expression control is from the regulation of transcription.

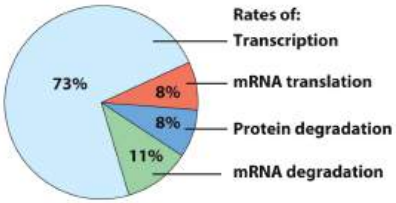


Figure 5.1: Transcription is the major process

Remark 5.1. *Out of all the cellular processes, transcription is the most common for all organisms.*

5.1 An Overview of Transcription

In transcription, RNA polymerase (pol) localizes, denatures the DNA's base pairing and uses one of its strand as template. It will then read the template

from 3' to 5' and synthesize RNA from 5' to 3' using rNTPs. The polymerization process is energetic so it will favour the α-phosphate group from the rNTPs thus removing the β and γ.

Convention: When talking about transcription, we need some sort of coordination along the DNA strand. The point at which the transcription starts (begin using rNTPs) is used as the *point of reference*. We call the direction that the transcription is heading toward **downstream** while the opposite direction is called **upstream**. On the downstream direction, the base-pair will be indicated by a positive number e.g. +10bp, +1kb, etc. while on the upstream it is negative e.g. -10bp, -1kbp etc.

Definition 5.2. A **promoter** is a region (usually upstream) that "promote" and activate transcription.

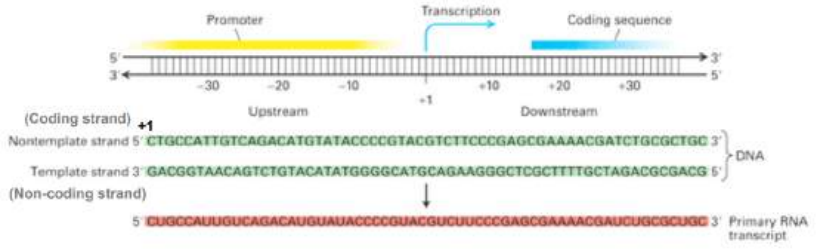


Figure 5.2: Transcription and direction.

Promoters can be really big and they don't have to be necessarily upstream, they can have elements that are kbs away. On the other hand, usually downstream is the location of the **coding sequence** which would form mRNAs after transcription and then translated to proteins. The coding sequence is made from 2 DNA strand, one of which is called the **template strand (coding strand)**. RNA pol read this strand from 3' to 5' and RNA will be formed from 5' to 3'. The initial RNA formed is called a **pre-RNA**. The main RNA pol used for such process is called **RNA pol II**

Remark 5.2. *RNA pol II can advance at a rate of 1000-3000nt/min. The long gene transcribed by them is **dystrophin** which takes a day.*

We can divide transcription into 3 stages:

I **Initiation:** RNA pol will bind to the promoter region, denature it locally and catalyze the first phosphodiester bond.

II **Elongation:** RNA pol will move downstream (3' to 5'), keep denature the DNA in front of it and polymerize a longer RNA strand (5' to 3').

III **Termination:** RNA pol recognizes a stop site, releases RNA and dissociate from the DNA template.

5.1.1 Prokaryotic vs Eukaryotic Transcription

Transcription between eukaryotes and prokaryotes can differs vastly but are also similar. The main differences and similarities we'll be looking at is: activation of transcription, cistronic, RNA polymerases.

Activation of Transcription

In Eukaryotes: RNA pol cannot do stuffs on its own, and require proteins to help them to locate the where to start. These proteins are also called **(general) transcription factors (TFs)**. The promoter region, along with TFs, will enhance the recruitment RNA pol to the site.

Definition 5.3. ***Cis*-regulatory elements** are non-coding gene that can regulate the transcription of the gene right next to (on the same side as) it. They includes: promoter regions, enhancers, operator, etc.

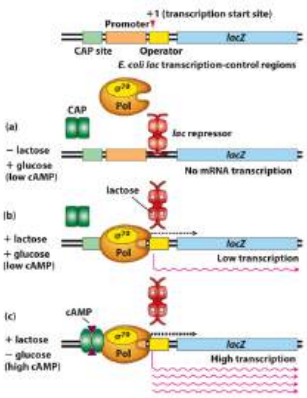


Figure 5.3: Operators (lac operon) for bacterial transcription of LacZ gene.

In Prokaryotes (bacteria): it is the σ factor that helped these RNA pol to locate the promoter region. One of the cis-regulatory elements that is unique for bacteria are **operators**. This is where other factors (proteins, etc.) can bind to that would lead to transcription to be repressed or enhanced (see Figure 5.3). They also have different sites that is *trans* (on a different side) to the coding gene that allow it to work with the *cis* and do regulation.

Remark 5.3. *Both prokaryotes have both trans and cis-regulatory elements.*

More differences between them: Bacterial transcription changes rapidly due to **allosteric factor** that can change the conformation of the transcriptional apparatus. This conformational change allow a rapid change in gene expression (genes can be enhanced or repressed quickly). So we can think of the transcription of bacteria is a mean to adapting to its environment. In eukaryotes, transcription are essential to create distinguished cell types (during embryogenesis) that give rise to us. Essentially, transcription for more complex eukaryotes is a a mean to differentiate into a complex system of cells that work together.

Remark 5.4. *we made up of 30 trillions cells with 200 different cell type and each of these cells are different from each other.*

Cistronic

Definition 5.4. **Cistronic** refers to a DNA segment that corresponds to a gene that can be transcribed.

In Eukaryotes: generally, eukaryotes are *monocistronics* which means that 1 DNA fragment would be transcribed into 1 mRNA which would be translated into 1 protein type.

In Prokaryotes: contrary to eukaryotes, they are *polycistronics*. This means that 1 DNA segment would be transcribed into 1 mRNA (that has many subsite on it) and then this mRNA can be translated into many differen protein.

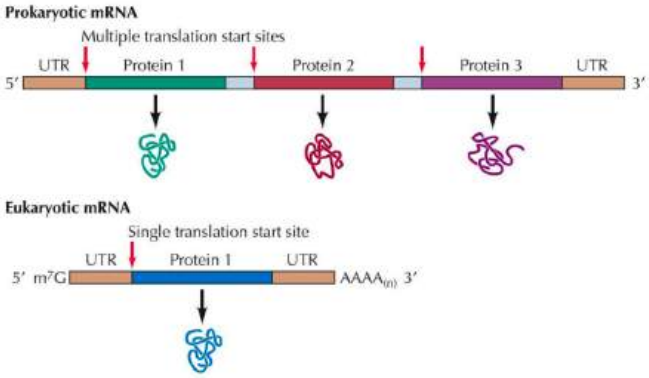


Figure 5.4: Cistronics between prokayotes and eukaryotes.

RNA Polymerase

In general, prokaryotes (bacteria) have only 1 types of RNA pol while eukaryotes have 3 types (RNA pol I, II and III).

Polymerase	RNA Transcribed	RNA Function
RNA polymerase I	Pre-rRNA (28S, 18S, 5.8S rRNAs)	Ribosome components, protein synthesis
RNA polymerase II	mRNA snRNAs siRNAs miRNAs	Encodes protein RNA splicing Chromatin-mediated repression, translation control Translation control
RNA polymerase III	tRNAs 5S rRNA snRNA U6 7S RNA Other small stable RNAs	Protein synthesis Ribosome component, protein synthesis RNA splicing Signal recognition particle for insertion of polypeptides into the endoplasmic reticulum Various functions, unknown for many

Figure 5.5: Different RNA pol in eukaryotes.

Even when they differs in number, but the general structure is sort of similar. Both of them exists in a multimeric complexes (multiple proteins complex) of βs, αs and ωs-like subunits and other extra components.

From the figure, you can see that we highlighted a particular structure on the RNA pol II: the **C-terminal domain (CTD)**, which is an important structure for the viability of cells. This structure is **intrinsically disordered** i.e. we can never exactly spot them. Another interesting thing about them is that they have many peptide-peptide repeats. (see Figure 5.8) Another structure you can spot is the **clamp domain**. This structure is important to keep the RNA pol from dissociate away from the DNA as they start transcribing. First, at initiation, the clamp domain is free/open thus won't do much; but as we transition toward elongation, the clamp domain will close and tightly bound to the DNA, disable any possibility of breaking apart. The newly synthesize RNA will exit where the CTD is roughly located. It is not apparent right now but in subsequent lectures, CTD is very important.

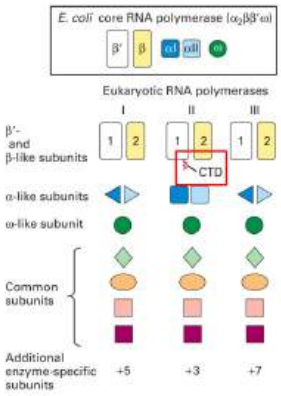


Figure 5.6: Similarities of RNA pol between prokaryotes and eukaryotes.

Remark 5.5. *A toxin called* α*-Amanitin found in many mushroom species has impacts on different RNA pol of eukaryotes i.e. they have different sensitivity to the toxin. RNA pol I is insensitive, pol II displays high sensitivity and pol III is moderately sensitive to this toxin.*

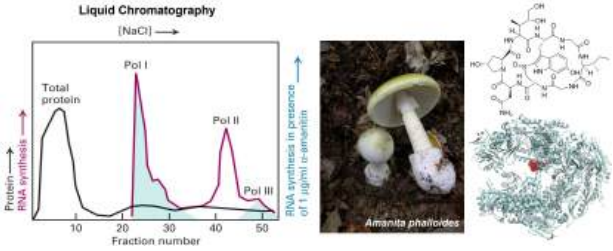


Figure 5.7: Chromatography of different RNA pol producing proteins under α-Amanitin.

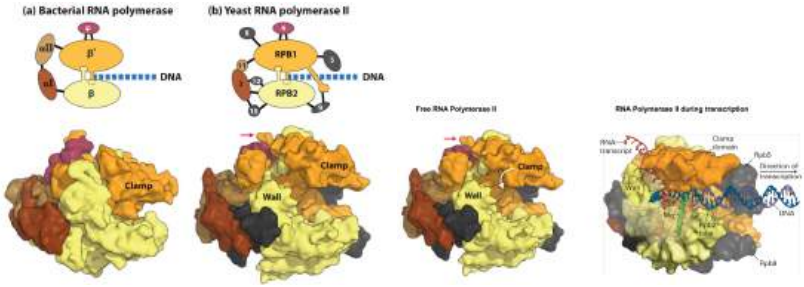


Figure 5.8: RNA pol between eukaryotes, prokaryotes and RNA pol II clamp domain.

5.1.2 Brief on CTD Phosphorylation

CTD can be phosphorylated on the subunits and they are phosphorylated in at a specific time: during or after initiation stage of transcription. We can perform experiment with the polytene chromosomes of drosophila melanogaster's salivary gland. We will stain it with antibodies that can either recognize phosphorylated CTD (stain it red) or unphosphorylated CTD (stain it green). What we found is that during transcription, the chromosome will form these morphologically distinct puffs which is coloured red hence it has a high amount of phosphorylated CTD while the unphosphorylated CTD are

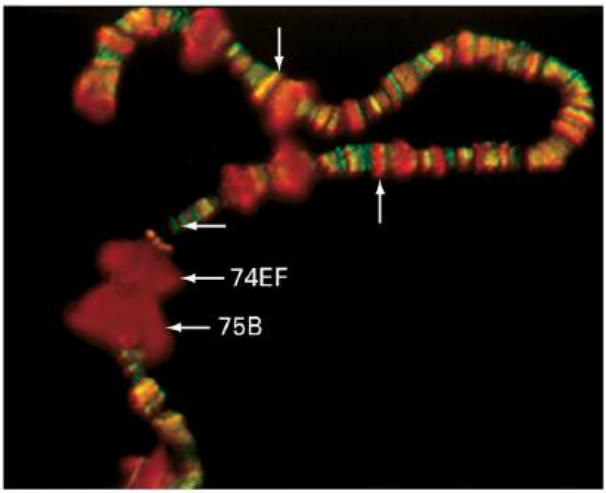


Figure 5.9: Phosphorylated CTD on polytene chromosomes of drosophila melanogaster's salivary gland.

simply scattering around. This is the best *in vivo* evidence that CTD associates with active transcription. transcription is an important process and is the predominant means of gene express and modification.

5.2 *Cis*-Regulatory Elements

Definition 5.5. **Regulatory elements** are elements that can effects transcription by either enhance it or suppress it (same as definition 5.3). When such elements is *cis*, it is close by otherwise it is *trans* (far away).

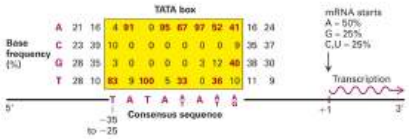


Figure 5.10: The sequence TATA appears very often in promoter lead scientist to develop the TATA box.

One of the *cis*-regulatory elements through investigation at the promoter region is the **TATA box** (named by its nucleotide sequence: TATA).

This TATA box is found in a number of genes, they're found upstream (around -30bp) to the gene sequence and is relatively the same spot. We can theorize that this is where the RNA pol will start transcription. A lot of gene have TATA box especially those who are expressed regularly.

Nevertheless not all genes have TATA box, but they would have other elements that can either be *cis* or *trans*-reg elements such as **TFIIB recognition elements** (roughly -35bp upstream), initiator (~ −1bp), DPE (~ +30bp). Again, not all gene have these elements.

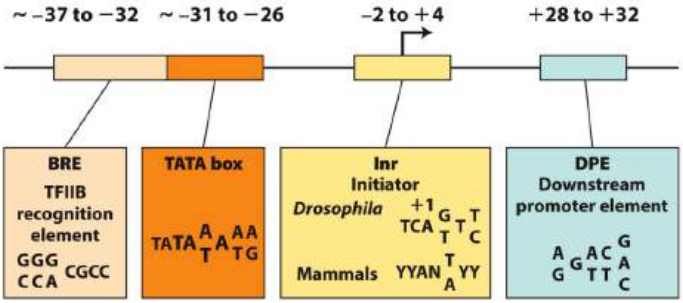


Figure 5.11: Other *cis* and *trans*-regulatory elements

TATA box is not the only elements that can influence; there are **promoter proximal elements** that can work with TATA box (around +1bp up to -200bp). They somehow effect the transcription of genes. There are **distal elements** (very far away ±50kb) that can enhance transcription. The mystery of this elements is how can it affect genes far away and we still haven't had an answer. However not all genes looks like this.

The majority of mammalian genes have **CpG islands** in between instead of TATA box, because of this genes in these regions are not produced in large quantity. They're typically associated with *housekeeping genes* (genes that are essential and are always expressed in all cells of an organism). CpG islands have *divergence transcription* where transcription is going both way (up and downstream) and have multiple starting site. They're probably responsible for a phenomenon that over 80% of the entire genome is transcribe in the first few RNA-seq series.

In yeast (simple eukaryotes), the genome is more compact and have less introns. The promoter site will have a TATA box at ~-90bp and *cis*-reg

elements are upstream and can be activated by other *trans*-acting factors. These elements are also known as **upstream activating sequence (UAS)**. UAS are like yeast-enhancers but not like typical mammalian enhancers where they can be located very far away.

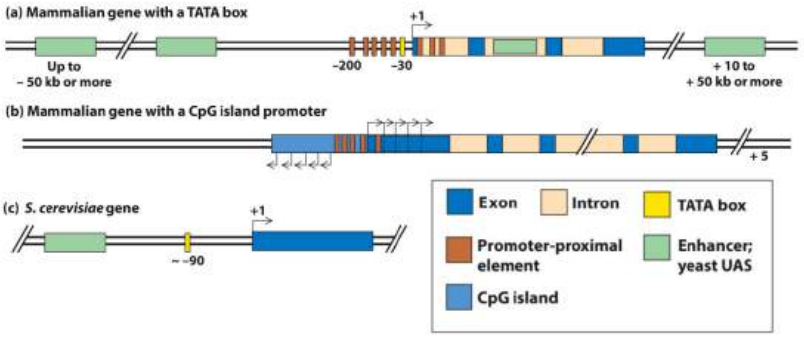


Figure 5.12: Regulation of gene via *cis*-regulatory elements of mammalians and yeasts.

With all of the above about cis-reg elements, **How do we know that the cis-reg elements actually play some roles in the regulation of a gene transcription?**To do this, we'll perform some experiment.

5.2.1 5'-Deletion Series

Transformation is the process of introducing vectors into a bacteria. Allow bacteria to make more of that DNA. In mammalian cells, the process is called **transfection**; while yeast is **transgenics**.

Supposed we isolate a large portion of the upstream region of a gene interest (thus include the promoter and some *cis*-reg elements). The gene we'll be using in this experiment is *transthyretin (TTR)* which is responsible for transport proteins that carry Vit A or thyroxine. The technique/ experiment used to analyze the upstream region of this gene is called **5'-deletion series**.

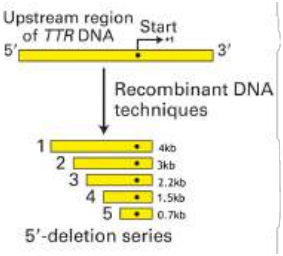


Figure 5.13: First step to 5' deletion

Procedure: We start with the TTR DNA and cut out the upstream region. If we're doing this in high quantity, the cut out segment should be equal in length. Then we start cutting away the length from the 5' and end up with 5 different DNA segment length (that could have a *cis*-acting elements). We then insert them onto an expression vector that have reporter gene (like a proxy). When we grow them, we will have 5 different genes construct, we then transfect these vectors into mammalian cells. Assuming that the mammalian cells have all the required machinery for transcription, we can then detect the level of transcription.

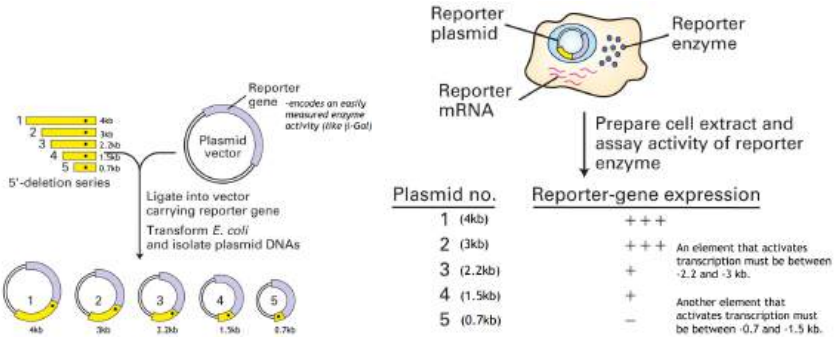


Figure 5.14: Transforming the 5 DNA fragment (on a plasmid) into bacteria. Then transfect them into mammalian cells for transcription.

As a result, the longer the fragment, of 4kb in length, the better the gene expression (via reporter gene) of the vector would be. As the fragment getting slightly shorter, from 4kb to 3kb, the gene expression is worsened but still relatively fine. However, at a certain point, from 3kb down to 2.2kb, it would jump down to very low transcriptional level. This must mean that between -2.2kb and -3kb there are elements that regulates transcription. Then it stays this way when the fragment is cut down to 1.5kb. However for the final fragment of 0.7kb there's no more transcription. This meant that between -1.5kb and -0.7kb there are also elements that activates transcription.

5' deletion series allow us to identify where those transcription regulatory elements lie within a large gene expression. This is not an ideal way since the more you cut the DNA down, for a better localization of the ele-

ments, the more you change the reporter gene position.

5.2.2 Linker Scanning Mutation

Linker scanning mutation is another way to localize regulatory element. It has similar procedure to that of 5' deletion

> **Procedure:** We take that small segment of the upstream DNA and insert it into an expression vector. These vectors will be transformed into bacteria. We then use restriction enzyme that can cut a specific small spot that we would know. As we have different restriction enzyme, it will cut at different spot on the expression plasmid's integrated regulatory segment. After each cut, we will measure the level of transcription and see the changes in gene expression (via reporter gene) correspond to whichever segment we've cut.

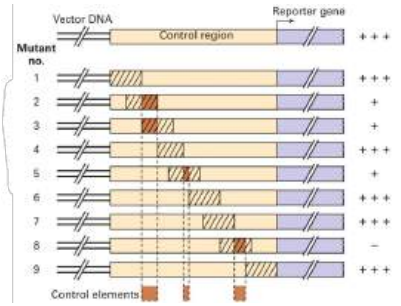


Figure 5.15: Mutant scanning linker with multiple restriction enzymes

As a result, we would find that some of the segment once cut, there will be no changes while others there'll be a big drop in gene expression; which means they have an important role in the transcription of the downstream gene. This should behave the same in mammalian cells.

5.2.3 Enhancer

Now let's go back to enhancer and take a deeper look at them.
Enhancer are regulatory elements that can work close or even very far away from the gene. **So what is it about enhancer that make them special? mechanism?** To answer this, let's look at the PAX6 gene.

PAX6 gene is an important gene for the development of the eyes and pancreatic tissues. When examining PAX6 gene in mice, we found there are at least 3 different transcript made from 3 different promoter. Not only that, there are numbers of elements that influence its expression in various tissues. These control elements can be, located between introns, for the retina, lens, cornea, pancreas etc. When we look at PAX6 for human, it is pretty much conserve i.e. these genes are generally stay the same all through out the organism life time as well as present on organism of similar lineage.

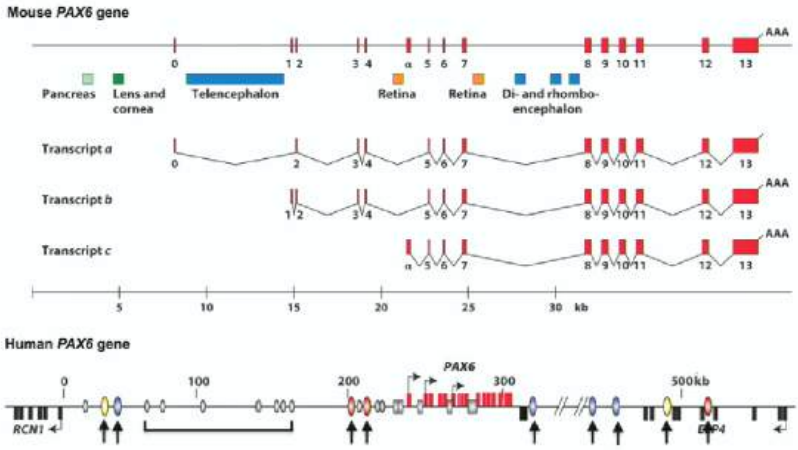


Figure 5.16: PAX6 gene of mouse and human.

Remark 5.6. *These control elements that are tissue-specific or state-specific are the enhancers.*

Generally, when comparing nucleotide sequences, there won't be much similarities between organisms thus high fluctuation. But once in a while, we run into sequences that don't fluctuate between organisms. This could be that they share such sequences but as well as a fluctuation in these sequence could lead to the organism death. To see this, we need to inspect another gene call SAU1.

SAU1 gene is a gene important for the dev of limbs. If you compare its nucleotide sequence, there will this conservation at a peak, this meant that the regions are under constraints and being conserved for a reason. **what**

is this reason?
If you take this DNA segment alone and introduce to an expression vector and then transfect it to a mouse fertilized egg. The reporter gene will be expressed in limb bud as the egg grow. These genes are thus said to correspond to an enhancer to genes expression, in this case limb development.

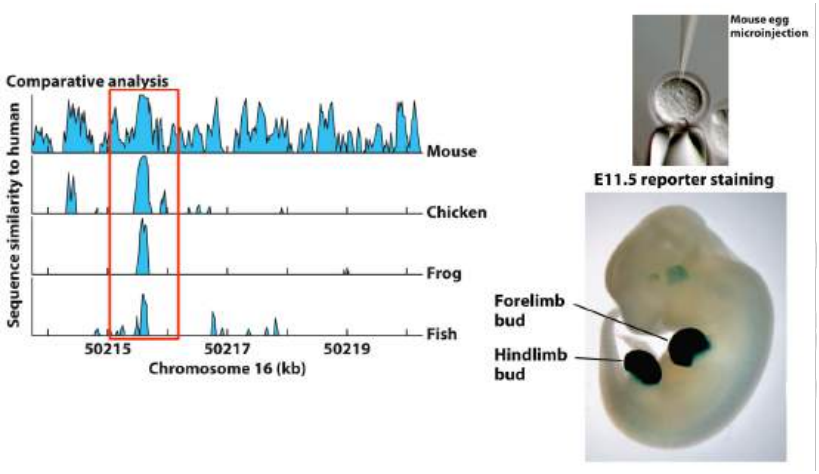


Figure 5.17: SAU1 gene and enhancer with limb bud.

One last peculiar thing about enhancer is that **they tend to act at great distances but how...?** This is where our linear model of gene need to be reconfigured. These DNAs are not simply long and straight but would fold and loop in these **topological associated domains** during transcription. Therefore, enhancers can be far away from the gene expression linearly, but they're very close with each other topologically.

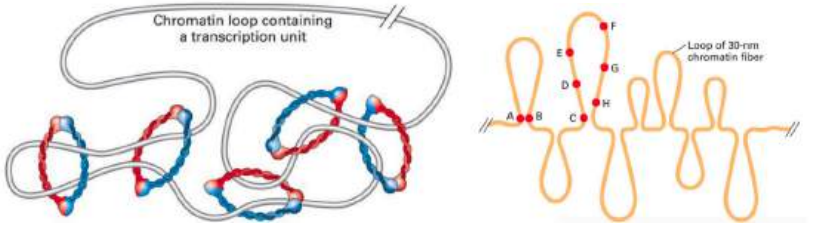


Figure 5.18: Topological associated domains.

5.3 Transcription Factors

What is it that DNA molecules are using to initiate or enhance transcription? In eukaryotes, it would be the general transcription factors.

Definition 5.6. **General transcription factors (TFs)** are proteins that are required to initiate eukaryotic transcription.

In the 1980s, there was a race to figure out what protein they were. Essentially, they were all trying to answer: **what are the critical proteins that work with RNA pol to initiate transcription?** We know that a promoter could assemble an acquired RNA. We can then insert this promoter into a plasmid, cut it and inject it with a **nuclear extract** (an extract from a cell nucleus)...transcription starts to happen! This implies there are proteins in the nuclear extract for transcription.

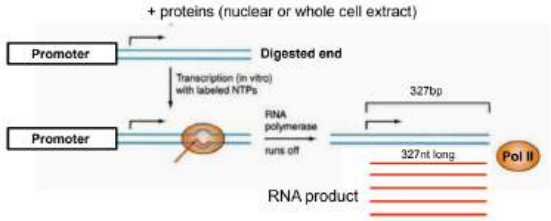


Figure 5.19: Nuclear extract and the discovery of TFs.

We can use liquid chromatography to purify nuclear extract. After purification, we get some TFs, that are TFIID, TFIIB, E, A, F and H. By examining these TFs and combining with RNA pol II, we can recreate transcription *in vitro*. TFIID is the only one that were characterized easily and a lot.

Definition 5.7. **Transcription factor II D (TFIID)** is a TF that make up the RNA pol **pre-initiation complex (PIC)** i.e. It is a protein complex that prepare the DNA template for transcription by RNA pol.

TFIID binds directly to the TATA-box. It does it through a single subunit called **TATA-box binding protein (TBP)**. As the name suggested, it binds to the TATA-box. What TBP can do is bind tightly to the minor grooves of the dsDNA (double helix) and cause a slight distortion. This distortion would then cause the 2 strands to separate. Although TBP mainly bind to the TATA-box, it is still very much required for any promoter site without a TATA-box for initiation of transcription.

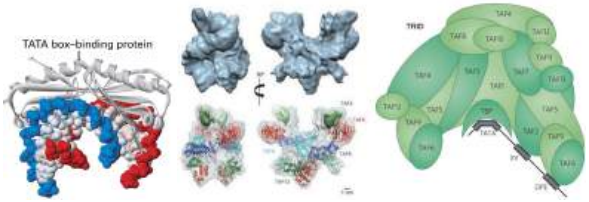


Figure 5.20: TAFs and TBP binding.

Remark 5.7. *TBP is important and required by RNA pol I and III for an efficient transcription process.*

There are many other elements such as TAFs which would enhance the binding of TBP called the **TBP associated factors (TAFs)**. These TAFs are also the reason why TFIID can bind to site without a TATA-box since it can interact with different site on the promoter.

5.3.1 Pre-Initiation Complex Formation

After years of researches and studies, we were able to drawn out a general mechanism of all of the TFIIs.

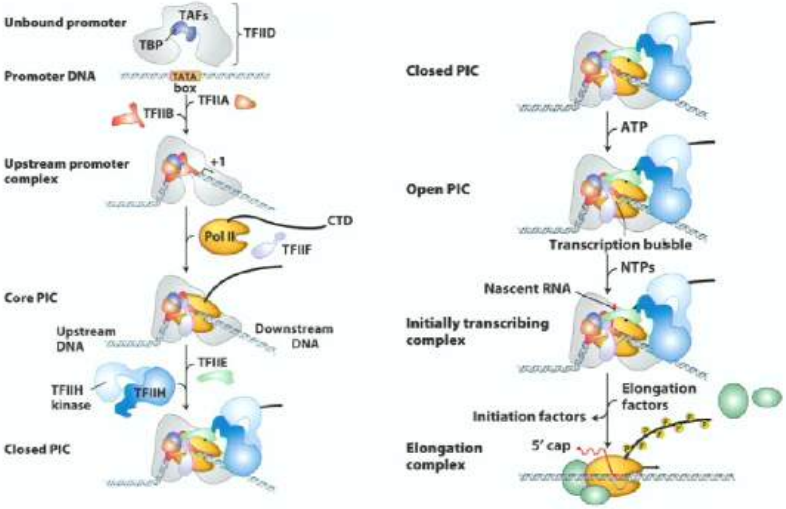


Figure 5.21: General mechanism of forming a PIC.

Firstly, TFIID interacts with the promoter site (mainly TATA-box). This will create a kink on the dsDNA. TFIIA and B will come and stabilize TFIID creating an **upstream promoter complex**. RNA pol II and TFIIF will come and interact with the complex creating the "core PIC". Finally, TFIIE will come and recruit TFIIH with its kinase forming a **closed PIC**. However, we must know that this can never happens *in vivo* because the moment the PIC detect ATP in the environment, it will become **opened PIC**, create a transcription bubble. If there were dNTPs in the environement, transcription will initiate. After initiation, elongation happens and almost all of the TFIIs will dissociate. It is presumed that the only TFs that remain at the promoter is TBP. As for RNA pol II, other factors will come to join it called **elongation factors** (we will see it later on).

5.3.2 Closed to Opened PIC Transition

The main TFII that help with the transition from a closed to an opened PIC is the TFIIH. TFIIH has 2 DNA ATP-dependent helicases that are involved in a disease called **Xeroderma Pigmentosum**: XPB and D. XPB will use ATP and melt the DNA base-pairing which allow the PIC complex to interact with it and begins initiation.

TFIIH is really hard to purified. In the original experiment, 900L of packed HeLa cells were used through liquid chromatography just to get a very minute amount of TFIIH.

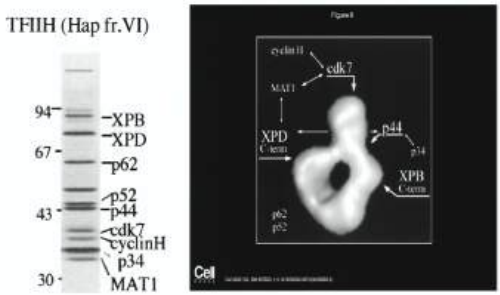


Figure 5.22: Post-purification and structure of TFIIH and its constituents.

TFIIH (along with XPD and XPB) is also involved in nucleotide excision repair. In fact, this finding was astounding since it like DNA reparation with

transcription factors.

Remark 5.8. *Xeroderma pigmentosum patients has bad nucleotide excision repair which means that any exposure to mutagens. They cannot step out to the sun as the UV radiation can potential damage them.*

Researchers has noticed that specific lesion on the DNA are more actively repaired than other. These lesions arise around the transcription promoter site and are very rapidly repaired because TFIIH are presence more. TFIIH doesn't accompany RNA pol during elongation therefore lesions further away will have less repair. Nevertheless, TFIIH still contributes to transcription-coupled DNA repair.

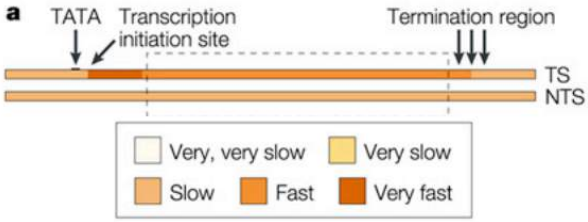


Figure 5.23: TFIIH reparation at the promoter site.

To end this lecture, when we look at the CTD of the RNA pol II, we found that it is heavily phosphorylated. The phosphorylation is carried out by proteins kinase that are associated with TFIIH: CDK7, cyclin H and Mat1. These 3 kinase will make the serine residue on the CTD be phosphorylated.

Remark 5.9. *Phosphorylation of CTD happens during the transition of initiation to elongation due protein kinase presence in TFIIH. This type of kinase is called modular kinase*

Remark 5.10. *I want to note to that TAF makes up the TFIID along with TBP. Furthermore, TFIID would help RNA pol becoming more efficient.*

5.4 Transcription Activators

Enhancer are those very far away, what about those that are closed together? i.e. promoter-proximal elements. Remember back to the mutation linker scanning, we figure out that certain region, when cut out of the DNA, would cause transcription level to diminish while other have no changes.

5.4.1 Electrophoretic Mobility Shift Assays and Activator

We can use those segment from the linker scanning to identify the proteins that are associated to it.

To do so, we can use **electrophoretic mobility shift assays (EMSA)**.

> **Procedure:** A segment of DNA will be labelled and then mixed with a nuclear to evaluate there is a protein or protein complex that would interact with that DNA substrate. This mixture of solution is them ran through EMSA. Any DNA and unbound protein woud be able to travel through the gel complex of the EMSA. The bounded protein-DNA complex would be hinder by the gel.

As a result, we would see a long band at the end of the EMSA which represents the unbound proteins and DNA. These bands are called **shifts**. Closer to the starting site, we can find small bands that represents the bounded protein and DNA.

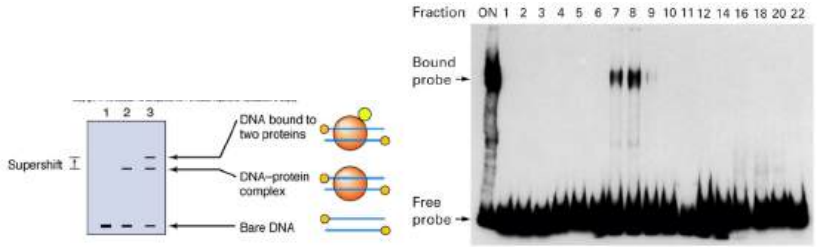


Figure 5.24: EMSA and results. On the 7 and 8th fraction, we can see that the proteins has bounded with the DNA while the rest has no effects.

Remark 5.11. *EMSA cannot reveal the sequence that is bound to the protein.*

Note: There are atleast 1 complex of bound protein and DNA.

The main thing we want to know is which proteins bind to that DNA segment. To do so, we simply need to perform a liquid chromatography on the nuclear extract to gets all of the composition.
Supposed that after doing the nuclear extract, you came up with multiple proteins candidate, **how would we come determine which candidate is best?** We can find the cDNA that correspond to those proteins protein. We

then insert that cDNA in a vector. At the same time, we will prepare an expression vector containing an **identified control element** coupled with a reporter gene (e.g. LacZ). We then **co-transfect** a cell with both of these vectors.

> **Mechanism of Action:** The machinery of the cell will turn the protein's cDNA to a functional protein. This functional proteins will bind to the control element on the other cDNA thus recruit the cell's machinery to transcribe the reporter gene.

As a result, if the candidate protein matches with the control element, there should be a strong expression of the reporter gene, otherwise it is a mismatch. To further test this, we can even mutate them to see a decrease in expression.

This is a test for to determine DNA transcription factor or more specifically *activator*.

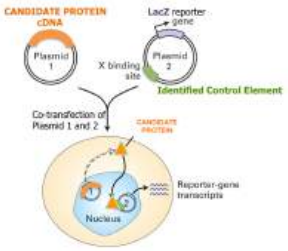


Figure 5.25: Co-transfection mechanism.

Definition 5.8. Proteins that can bind to DNA is called **DNA-binding protein (DBP)**. A DPB or TF that can bind to either enhancer or the promoter-proximal elements and increases transcription of a gene is called **activator**.

Remark 5.12. *All activator are TFs and DPB but the antithesis isn't true. Furthermore, not all DPB are TFs but all TFs are DPB.*

By examining proteins with similar functions. We found out a specific proteins' motifs that are used the majority of the time for DNA recognition (i.e. used by these DBPs). This motif is called **recognition helix (helix-turn-helix)** which comprises of 2 α-helical domain forming a dimer. This structure was first discover through bacteriophage repressor 434 homodimer of *E. Coli*. (see Figure 5.26)
The recognition helix bind mainly to the major grooves of the DNA using non-

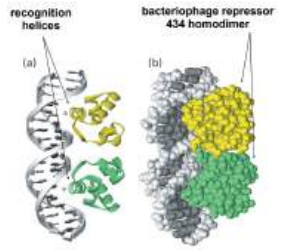


Figure 5.26: recognition helix motifs and bacteriophage repressor 434 homodimer.

covalent interaction such as hydrogen bonding, Van der Waals interaction, etc.

5.4.2 Modular Structure of Activators

Definition 5.9. A proteins is said to be a**modular structures** meant that they have different discrete part on them that can perform different and independent functions. These parts are called **modular domains**.

Deletion of these modular domains could lead to a loss of function for the activators themselves. To see if this is true, we will perform an experiment.

GAL4 is an important DNA binding transcription factor (activator) that can activate transcription of LacZ gene which gives rise to β-galatosidase. To be more precise, GAL4 has DNA-binding domain that can bind to the **upstream activating sequence of galactosidase (UAS_{gal})**. Not only that it contains **activation domain** that allow other co-activating proteins to bind to.

> **Procedure:** First, insert the UAS_{gal} next to a reporter gene, then filled its environment with GAL4. Now we will do some modification on the GAL4, that is making a series of the deletion from the N-terminal, C-terminal and its middle.

If we were to make deletions on the N-terminal, we found that deletion of the first 50 peptide (pt) would disable the ability of GAL4 from binding to the DNA which lead to a lower β-galatosidase activity. But if we start deletion on the C-terminal, we found that transcription will be high at the beginning but will start to diminish until it cannot transcribe (from 881 to 691 pt). We can then perform another deletion of the middle and found that no matter the length (as long as it does not lead to the deletion of the previous ends), the DNA-binding properties stays the same as well as the transcription of β-galatosidase.

The reason for this is because of modular domains. All activators has at least 1 DNA-binding domain, 1 activation domain and 1 flexible protein domain. In the experiment, when we delete in the N terminal, we cut away its DNA-binding domain while on the C terminal deletion, we get rid of its activation domain hence lowering transcription. (see Figure 5.27)

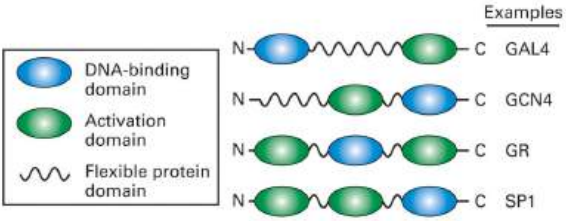


Figure 5.27: Different modular domains for GAL4 and other TFs.

Remark 5.13. *These 3 domains are only some of the possible domains. There could also be transcription repression domain, chromatin remodelling domain, nuclear import domain and protein interaction domain.*

Remark 5.14. *DNA binding domain is universal i.e. if we were to take 1 domain of 1 activator and insert it in another activator, it would still function.*

5.4.3 Homeodomain and DNA Binding Domain Subtypes

Definition 5.10. A **homeodomain** is a *conserved* DNA binding domain of around 60 amino acid long, consisted of many recognition helix and is found in many TFs.

Remark 5.15. *A structure is said to be **conserved** is when that structure stays relatively the same through evolution as well as interspecies.*

Remark 5.16. *Homeodomain got its name because of its presence to many TFs that give rise to **homeotic transformation** (a mutation where 1 body part gives rise to a different one.)*

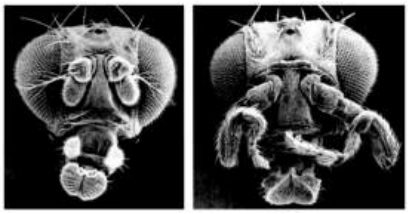


Figure 5.28: Homeotic transformation of a fly head where its antennas is replaced by its legs.

Homeodomain is 1 type of DNA binding domain, but there are other structure such as the Zn**-finger binding domain** . The Zn-finger can be

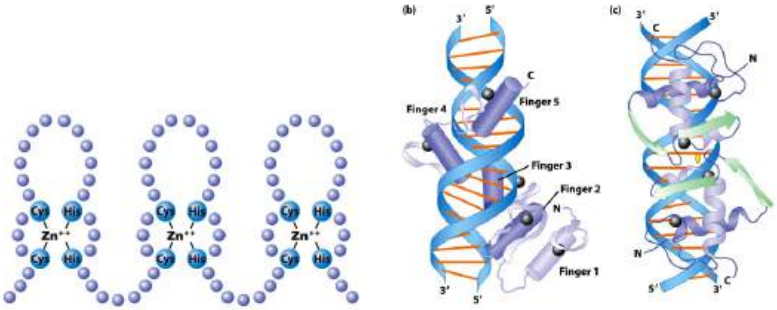


Figure 5.29: Different subtype of Zn-finger.

divided into 3 types that made up the DBP: C_2H_2, C_4 and C_6 class. $C_2H_2 - Zn$ finger consists of 3 or more Zn finger motif bind to DNA as a single monomer. $C_4 - Zn$ finger consists of 2 Zn finger motif binding together to a DNA as a homodimer or heterodimer. $C_6 - Zn$ finger consists of 6 cysteine ligands bind to 2 Zn^{+2} ions.

Leucine-Zipper Proteins and Helix-Loop-Helix Proteins

The final 2 structure as DBP (activator) are leucine-zipper proteins and helix-loop-helix proteins.
Leucine-zipper proteins consist of uniquely 2 α-helix with extensions that can bind to the major groove of the DNA. When observing this DBP more closely, we can see a repetitive leucine amino acid as well as hydrophobic ones, all of which can be found on the DNA binding domain. The hydrphobic amino acid is especially important for the formation of the coiled-coil domain between the 2 α-helix hence create a dimer.

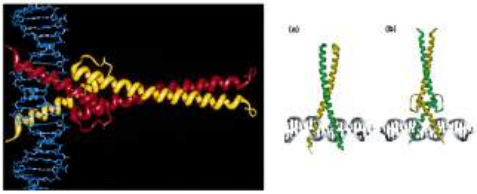


Figure 5.30: leucine-zipper (a) and HLH proteins structure (right and b).

Helix-loop-helix (HLH) proteins have similar structure to that of leucine-zipper proteins, that is they have 2 α-helix. Instead of having a simple ex-

tension to bind to the DNA, they form a loop of amino acid before extend out which create a kink between the α-helix and the extension. Not only that, on the extension (which is the DNA-binding domain), they have hydrophobic amino acid spaced in intervals similar to that of an amphipathic α-helix.

Remark 5.17. *DNA binding domains can interact with unrelated classes to create a cooperative DNA binding complex. They somehow reinforce each other on that complex.*

Example 5.4.1. NFAT1 and AP1 are DNA binding proteins that would loosely (weak) bind to DNA individually but when they interact together, they form a cooperative strong bind to the DNA.

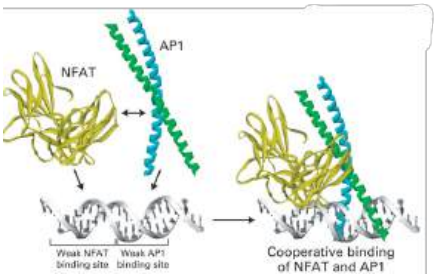


Figure 5.31: Cooperative binding of NFAT1 and AP1.

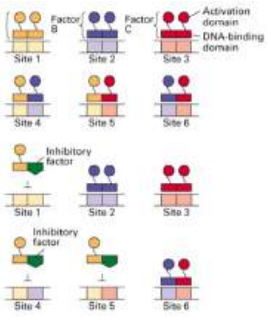


Figure 5.32: Different combination of DBPs.

This DNA binding proteins cooperation would lead to homo and heterodimerization of 2 proteins monomeric unit. This would diversify the transcription activity output since there is more way to bind to the DNA thus we get more gene product activated in different ways. e.g. 3 TFs can homo/heter-dimerize into 6 different gene regulation combination.

Remark 5.18. *Once again, to remind you, The DNA binding domain will interact directly to the regulatory sequences.*

5.4.4 ChIP-seq

Now that we went through all of knowing about these activator, the main question remain **to what DNA sequence are activator or TFs bind to?** Well...to answer this, we need to introduce a new experiment.

At the beginning, we will start with a technique called **Epitode tagging**. It is a technique that allow you to express a protein of interest by allowing antibody to bind to it.

> **Procedure:** Begin by inserting the gene of interest into a vector. Next to it, an epitode tag will be attached to it. This expression vector will then be introduced to cell for transcription. After transcription, the mRNA (then become proteins) will have a point where specific antiabodies can recognize (due to the epitode tag).

The process of tagging can depend on the experimenter e.g. protein can be put through PCR or chIP-seq.

ChIP-seq is a method to perform epigenomic mapping

> **Procedure:** First, cells will be cross-linked with formaldehyde to stabilize the chromatin bound protein. Then the isolated chromatin will be fragmented into smaller pieces, accomplished by enzymatic digestion and monitored by gel electrophoresis. The chromatin was previously epitode tagged so then it would be under incubation with a specific primary antibody. The antibody-bound chromatin will be isolated through immunoprecipitation. The chromatin will then be purified with proteinase and RNase to destroy any RNA and proteins leaving only DNA. This DNA can then be ran through PCR, qPCR, and NGS.

At the end, we would get the DNA sequence corresponds to the gene that is bound by the TFs of which we've epitode tagged previously (which was recognized by the incubation of the primary antibody).

5.4.5 Transciptional mediator

There are large proteins complexes interact with the RNA pol II referred to as the **holoenzyme**. When you purify them, these large complexes stuck

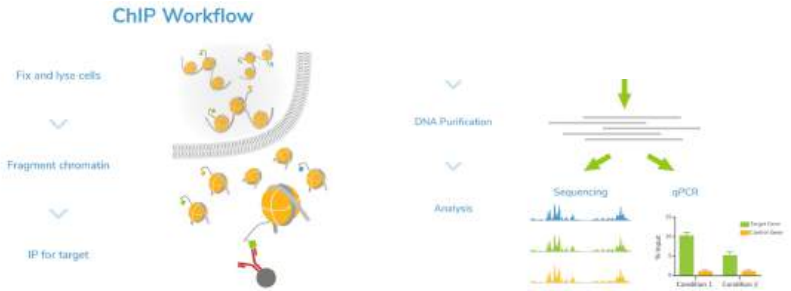


Figure 5.33: Process of chIP-seq (left to right).

together and is called the *mediator complex*.

Definition 5.11. A **mediator** is a multisubunit protein complex bridges the function of TFs and RNA pol II. It also facilitate the initiation phase.

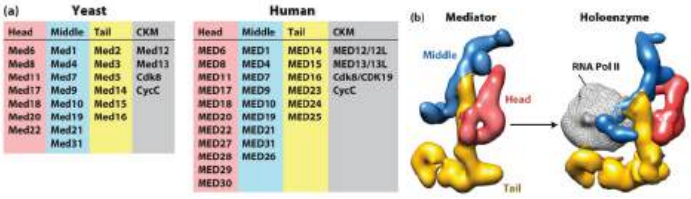


Figure 5.34: Mediator domains of yeast and human is conserved.

Mediators typically form 3 major domain: head, middle and tail. Another part called the **cyclin kinase module** which consists of kinases but we won't look at them.
What's interesting about mediators is their 31 subunits are conserved through evolution. We can see this when comparing yeast and human's mediator. (see figure 5.34)

Remark 5.19. *1 thing to look back and realize is that the activation domain on TFs are not well characterized. It is unstructured or what we called them intrinsically disordered domains.*

In certain condition, the head and middle domain will take on conformation that allow them to interact with RNA pol II's interfaces. Many of the mediator subunits are found to interact specifically to DNA binding activator. When there is a mutation on these subunit, it will not compromise

the entire function of the mediator; instead, it will disable the transcription activation by a activator.

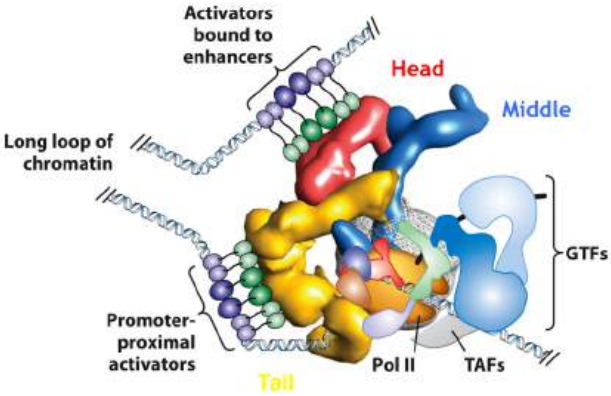


Figure 5.35: Mediator illustration 1

When look at a mediator in a chromatin loop, we will see that the mediator will site in the middle presenting the activator to the *trans*-regulatory factors (enhancers etc.) or even proximal-promoter elements. Doing so it enhances to overall transcription reaction i.e. They position activators in a correct location to better enhance the activation of a specific gene.

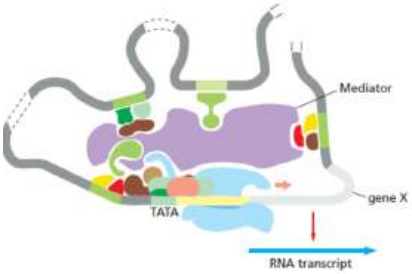


Figure 5.36: Mediator illustration 2

We can think of mediator like a glue that hold everything together and bridge the effects of TFs to the DNA elements which then bridge to the RNA pol II. Essentially, it's the reason behind having the topological loop as discussed before how enhancer far away can have an effect on the transcription.

5.5 Transcriptional Activation and Repression

Remember, transcription is not linear but just for now we will return to this model. The main question we want to answer today is that **What does transcription looks like? what does it mean to transcribe a gene at a high rate?** Commonly, what students and people originally think is that a highly transcribed gene is when the transcription of that gene has a higher *flux* of RNA pol II going through it at a high constant rate.

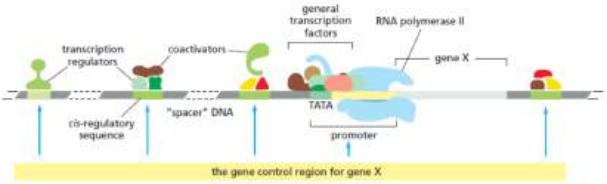


Figure 5.37: Linear model of transcription

This thought process is indeed true for bacterial or prokaryotic gene; that is, if a gene is highly transcribed, there will be a large number of RNA pol going through and transcribing said gene...this is not so true with eukaryotic genes. To see why this is the case, we need to look at some experiment

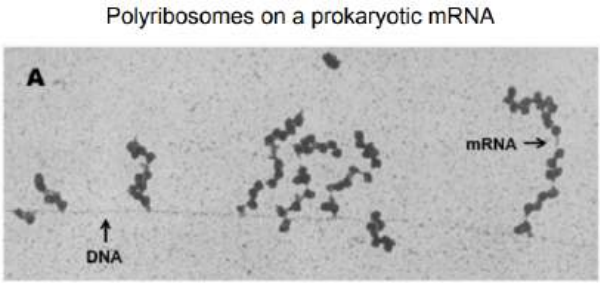


Figure 5.38: Transcription of bacteria happens with high flux of RNA pol that transcribe 1 after the other. Every RNA transcript coming out will be bound by ribosomes immediately.

5.5.1 GFP-labelled RNA Stem-loop Binding Proteins

When analysing the RNA transcription levels with all the previously mentioned method like RT-qPCR, Northern blot, etc. we're measuring the **steady**

state level which is the combination of RNA synthesized and degraded i.e. it doesn't tell you which gene is activated, how long, how much etc.

Researchers then thought of a clever way to somehow "tag" the RNA as it is transcribed so that they can detect the level of transcription through time. This way is through introduction of a structure at the 5' end of the gene of interest. This structure can then be recognized by labelled proteins.

Example 5.5.1. Bacteriophage's coat protein can recognize stem loop structure.

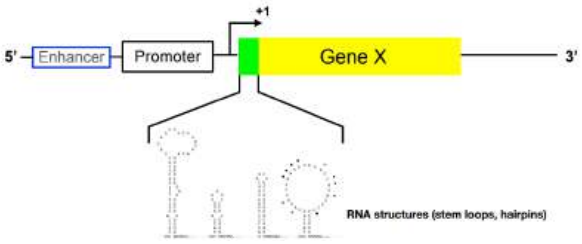


Figure 5.39: Stem loop structure can be inserted to the 5' end of the gene of interest to be recognized by coat protein.

We can thus introduce stem-loop structure to the gene of interest; at the same time, we introduce that recognition proteins that has been labelled, supposedly GFP. Once the gene transcribed by the RNA pol II, the GFP labelled RNA binding protein will come and regnize the stem loop on that RNA and bind to it. When the binding is established, there would be a GFP focus point.

Theoretically, we could insert this stem loop for every RNA transcript.

5.5.2 Transcription Burst

In the researchers' studies, they mainly focused on a gene that is important for the gastrulation of *Drosophila* and it is called **Snail**. Essentially they want to understand the transcription of Snail during gastrulation. They also identified a **shadow enhancer** that is down stream to the gene ($\sim 7.5kb$) which would also be used in the experiment.

Remark 5.20. *It is called shadow enhancer because it is at a far away distance from the actual gene it is regulating.*

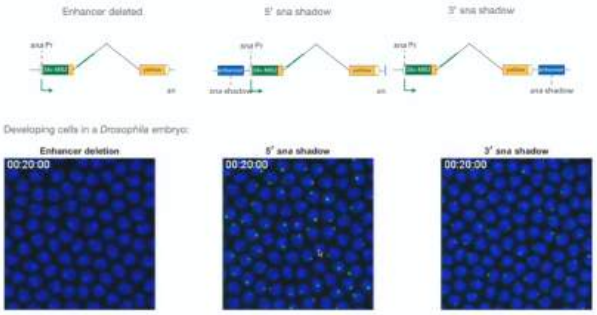


Figure 5.40: Snail gene transcription.

In this experiment, researchers set up 3 independent constructs for the gene: 1, the shadow enhancer will be at the same spot; 2, the enhancer is moved toward the 5' end; and 3, the enhancer will be completely removed. Along side these constructs, we will also insert **MS2 stem loop** that can be recognized by *MS2co proteins* (GFP labelled).
The experiment is then carried on, allow transcription (during gastrulation) to happen what interesting was that transcription of Snail gene happened in burst!

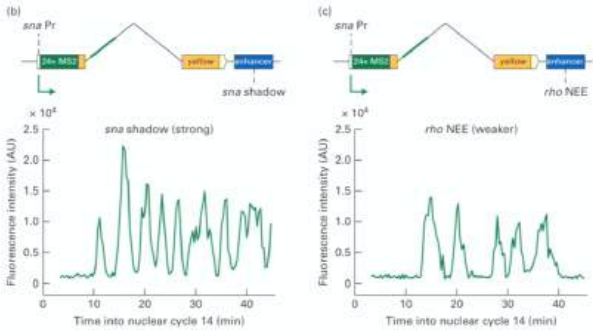


Figure 5.41: Different burst frequency when enhancer is moved.

i.e. It did not follow the typical flux model of the bacteria! What they also found is that transcriptional efficiency is proportional to the burst frequency i.e. transcription of the gene is higher as the burst frequency increased. Not only that but enhancers also plays a role in burst frequency too.

The main questions we would have right now is **why does the RNA pol II stop...isn't it inefficient to be in bursts? what is the biochemical reasoning behind this?** To answer this, we need to understand a concept that is of a distance to to this.

5.5.3 P-Granules and Transcription Condensates

Definition 5.12. **P-granules** are structures found in cytoplasm of germ cells of *Caernohabditis elegans* (roundworm) and is especially important for its development.

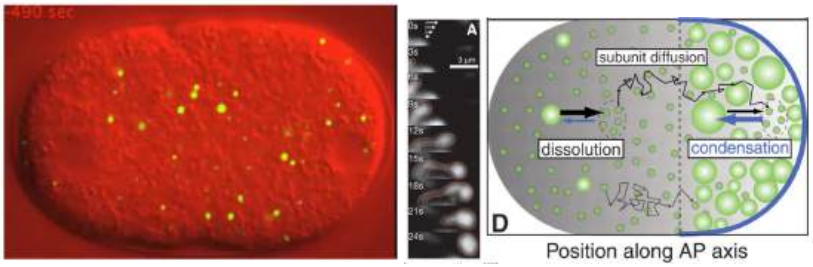


Figure 5.42: P-granules (in green) on *C. elegans* zygote (left). Formation of the liquid-liquid condensates (right).

2 scientists looked P-granules of a developping *C. elegans* zygote. What they found is that P-granules will move to the posterior end of the zygote as it divides. This posterior end will develop into the germ line of the zygote. They want to further see how this P-granules could give rise to the germ line so they begin to purify it. Nevertheless, It is impossible to purify P-granules because they're not really granules but more of a droplets. They're more known as **liquid-liquid condensates**.
A further investiagation led them to find that the constituents of these condensates are actually soluble to the cytoplasm; but in the posterior it isn't soluble and form condensates. We will stop at this concepts because we're not interested in embryology but the transcription mechanism.

Transcription Condensate

How would understanding this P-granules link with transcription? Well...it seems to us, through many experiment, that the condensates also formed

and is important for a transcriptional reaction. To see this is the case, we will need to perform our own experiment.

We begin by labelling 1 of the subunits of the mediator called **MED1**. MED1 has intrinsically disordered regions (IDR) that seems to be important for its function. A labelled protein called **mCherry** can recognize the IDR and bind to it. Another component that would be important is the transcriptional activator (chromal domain activator) called **BRD4** that also have an IDR that allow GFP to bind to it.

> **Procedure:** We will set up 6 independent experimental mixture. Mixture 1 would have the mCherry and the MED1-IDR together, mixture 2 would have the MED1-IDR and GFP and mixture 3 would have MED1-IDR, mCherry and GFP. Mixture 4 would have the BRD4, MED1-IDR and mCherry, mixture 5 would have BRD4, MED1-IDR and GFP and finally a mixture 6 have all of the component together.

As a result, we found that in mixture 1, the MED1-IDR along with the mCherry will form these "puncta" (condensates) which we can detect with a red colour. mixture 2 doesn't have anything since MED1-IDR isn't associate with GFP. Mixture 3 is the same as mixture 1 due to association of mCherry only. Mixture 4 is similar to 1 since the BRD4-IDR isn't associated to mCherry. Mixture 5 has similar condensates to 1 but will give off green since BRD4-IDR is associated with GFP (and not MED1-IDR). Finally mixture 6 would have condensates that is the merge of green and red.

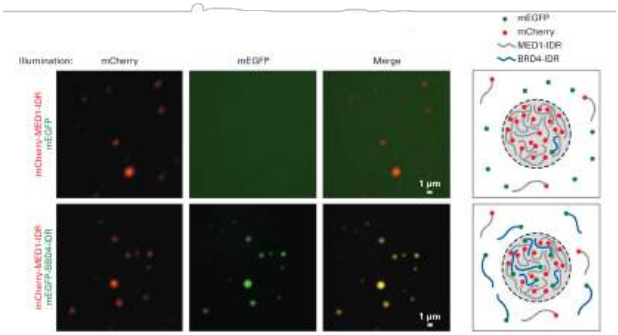


Figure 5.43: Results of the experiments.

What we can draw from this experiment is that as the IDR favours the interactions they have on one another and start to associates concentrate together. Once they're concentrated together, they form large liquid-liquid condensates that are no longer soluble in the nuclei.

We can think of these condensates as the place where important transcriptional regulators can come together. The formation of the condesantes is depending on the concentration of the macromolecules found in it such as DNA, RNA, proteins etc. but also its valency such as electrostatic interactions, IDRs interactions, modifications, etc.

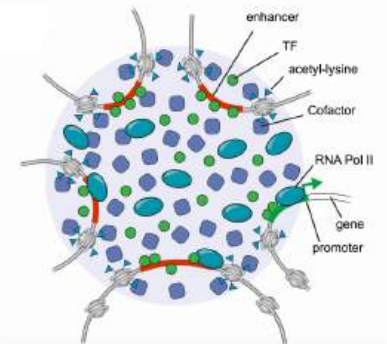


Figure 5.44: Constituents of a typical condensate.

To really drive this point home of a condensate and a transcriptional burst, another experiment and observation was made. In this experiment, mediator and RNA pol II was labelled with **JF646** and **Dendra2** respectively. Researchers then look at how they would interact *in vivo*. What they found is that the RNA pol II and the mediator would come together then move away.

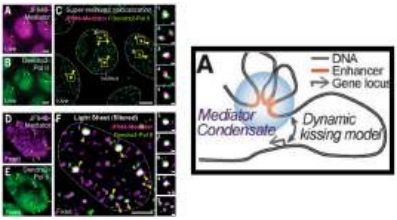


Figure 5.45: Experiment and the dynamic kissing model.

From this finding they theorized a model called the **dynamic kissing model** where mediator would form a condensate with the enhancer and will come together with the RNA pol II and initiate transcription. Once RNA pol II is transcribing, the condensate will dissociate then associate again for transcription.

Finally, researchers came up with the best hypothesis so far (but still questionable) is that the bursting nature of transcription is due to the association and dissociation of the condensates. At the beginning, the condensate will form with all of the necessary components to initiate a transcriptional reaction. Once this reaction takes place, more RNA are being produced and the interactions of this RNA will change the valency of the condensate leading it to dissociates. This RNA-dependent formation and dissociation of the transcriptional condensates may explain the bursting nature and the periodicity of highly transcribed genes.

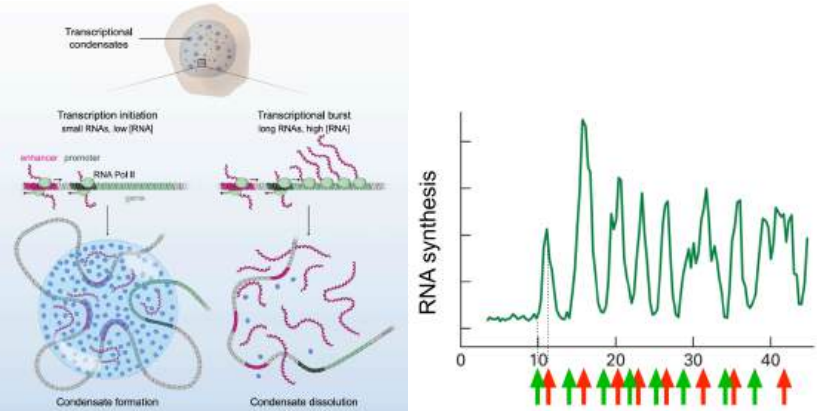


Figure 5.46: The hypothetical model of transcriptional condensate (left). Bursting and periodicity of transcription (right) where green arrow is where the condensate form and red is where it begins to dissociate.

5.6 Chromatin, Epigenetics and Histone Code

As we've known from previous chapters that DNA are not free flowing in the nucleus but are packed tightly together. They're associated with histone proteins that bind strongly to. The histone will wind up the DNA into a

higher structure call **nucleosomes**. Nucleosomes are important structures that are required to pack DNA into chromatin then finally chromosomes.

Definition 5.13. A **heterochromatin** is a highly compact form of a chromatin that can be found often at the nuclear envelope near the pore. On the other hand, **euchromatin** is a less compact form of chromatin as compared to heterochromatin.

The main problem with heterochromatin is that it is really dense and compact, to the points that TFs cannot access the DNA template. We considered heterochromatin as transcriptionally inactive.

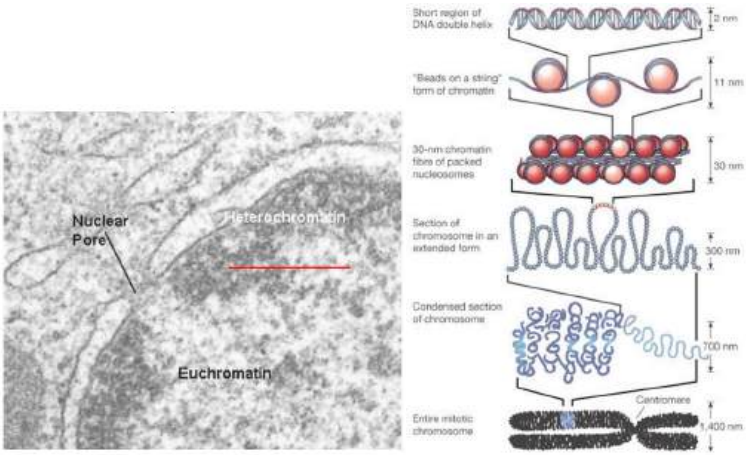


Figure 5.47: Heterochromatin and euchromatin.

Remark 5.21. *Transcriptionally inactive regions of genome are maintained in a heterochromatinized state.*

Euchromatin on the other hand is less compact and is more easily accessible by TFs which makes it transcriptionally active. The majority of understanding the functionality of chromatin and transcription is through studying yeast.

5.6.1 Yeast's Mating Type and Silencing

Saccharomyces cerevisiae is a yeast variant that can exist in either diploid or haploid form. In general, 2 haploid yeast can mate to give rise to a diploid

one that is also more stable; on the contrary, under stressful condition, a diploid yeast can under go meiosis to become haploid again.

Remark 5.22. *2 haploid yeast can only mate with each other if they have different* ***mating type****, which is the equivalent of "sex" in higher organism.*

Each of the haploid cell would have a gene that express this mating type, which can be either a or α. A diploid yeast must be a combination of mating type a and α, aa nor $\alpha\alpha$ will not work. The way that a yeast can identify themselves as 1 mating type (either a or α) is based on a complex series of events where 1 section of the chromosome will move into its actively expressed region of that yeast.
In this case, we will look at the **chromosome III** and its **locus** (another way of saying gene but on chromosome).

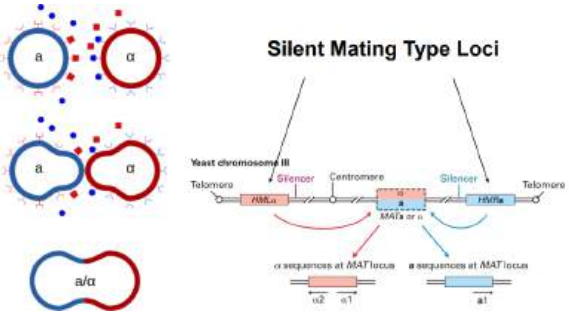


Figure 5.48: Mating of 2 haploid yeast to diploid (left). Recombination of either HMRa or HMLα to the MAT locus (right).

> **Mechanism of Action:** Chromosome III has 2 main loci which a yeast's mating type would be **homothallic mating left α (HMLα)** and **homothallic mating right a (HMRa)**. These 2 loci encode for a yeast to become either **a** or α. The yeast can take on 1 of these mating type by transfering either HMRa or HMLα into the actively expressed region called **MAT locus**. Once this recombination event is made, the chromatin must silence both the HMR**a** and HMLα or else the yeast would be "confused" because there should only be 1 region that express the mating type of the yeast that is the MAT locus.

This is a generalization of the entire process. However, the recombination or transfer event to the MAT locus is irrelevant to our study of chro-

matin. What we're more interested is that chromatin can silence or shut down a specific locus. Essentially we're asking ourselves **how can you shut down the expression of a gene on a chromosome even when it's "expressable"?**

5.6.2 Silencing Initiation Proteins

From what we see in yeast, the middle region (MAT) would be highly expressed while other is silenced so we know that it has something to do with the region on the chromosome. We also know that this silencing mechanism is made through a physical process instead of a biochemical ones. We can see this by introducing a bacterial enzyme that will methylate residues around the G-A-C-T motifs to the yeast.

Remark 5.23. *Although the enzyme is heterologous from the yeast cell, we still express it in the yeast anyway.*

What we found from this is that certain regions at the silencing region is not methylated which to show that the enzyme cannot access the DNA of that region to methylate. Essentially there's a physical mechanism that make those DNA regions inaccessible for proteins and TFs. What's even more interesting is that this same silencing happens in the telomeres i.e. if you introduce a gene in the telomere, it is silenced.

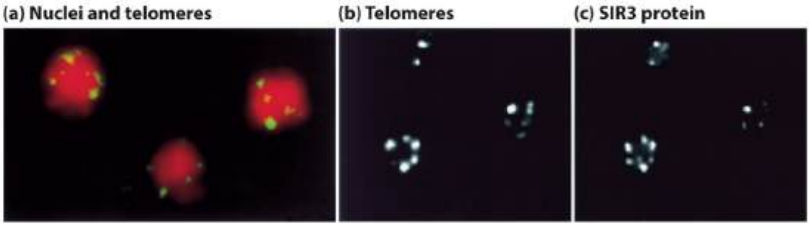


Figure 5.49: *in situ* hybridization of SIR3 found that it maily located at the telomeres which is also important for silencing

Geneticist studying this was intrigued by such finding so what they did next was to do a genetic analysis, that is, introducing a mutation such that they can find something that doesn't follow the normal rule. In such case they found that a mutation on histone protein would disable the silencing mechanism. A further analysis and genetic screening led them to be able to identify these proteins that are **RAP1** then many other **silent information**

regulator 1-4 (SIR1-4). Each of these protein would have different function that can help histone increating the silencing.

- **RAP1:** can bind to DNA silencer region or repetitive sequences of telomeres.
- **SIR1:** can coorperate with RAP1 and is important for silencing of mating types loci.
- **SIR2,3 and 4:** SIR3 and 4 can bind to the **hypoacetylated histone tail** and recruits SIR2 . These 3 SIRs together would form the higher conformation that lead to silencing.

Do these proteins act independently from each other? Well...It seems like there only 1 proteins that initiate everything that is the RAP1. After the recognition, it will recruits SIRs to form higher order complexes.

> **Mechanism of Action:** RAP1 first identifies the sequences to be silenced, supposedly the telomeres. This initial recognition will then recuit SIR2, 3 and 4 to the site where they will interact with the RAP1 to form large complexes. Not only SIR2-4 can be recruited to telomere by RAP1 but it can come because it can regonize the hypoacetylated region of the histone. What's even more interesting is that SIR2 have enzymatic activities that can bind to the hypoacetylated histone tail and perform changes on it and ensure that all the region around it is also hypoacetylated. As there's more hypoacetylated histone, the size of the complex will grow and tightly compacted together.

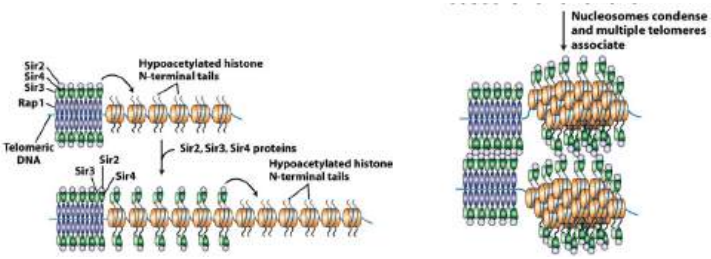


Figure 5.50: Building a silencing complex with RAP1 and SIR2-4.

5.6.3 Histone Tail Modifications

The majority of N-terminal histone tail is unstructured but they can be modified through post-translational modification e.g. phosphorylation, acetylation, ubiquination etc. Modifying these tails would results in a different output made by the histone.

Example 5.6.1. Acetylation of the lysines on the histone tail would lead to the neutralizing the electrostatic interaction between the histone tail hence it would loosen its grip on the DNA. Because of this loosening, acetylation of lysine is associated with an increase in transcription.

Remark 5.24. *Histone is multisubunit complexes that made up of H1, H2B, H2A, H3 and H4 proteins subunits.*

We won't look into much or acetylation and ubiquination but what we're more interested in this lecture is the methylation on the lysines. Methylation of lysines can certainly change the functionality of the histone, but also the amount of added methyl group makes.

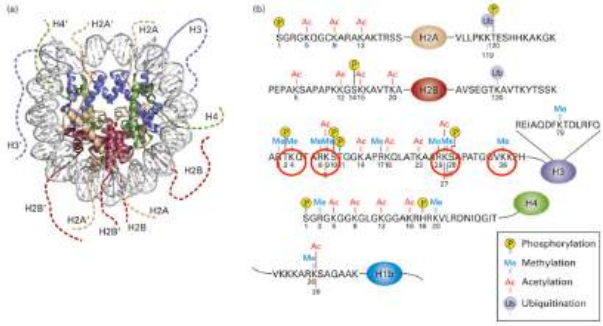


Figure 5.51: Histone complex and methylation at different sites.

Looking at the H3 histone subunits, there are important methylation that we can spot. On the lysine at position 4 of H3 (H3K4), we can perform a mono-, di- and even tri-methylated and this methylation would have so aspect with the transcriptional activation and euchromatin. Moving toward the C-terminal, we can spot the H3K9 (lysine at position 9 on H3 histone) which is important for our study since mono-, di- or trimethylate it, is observed to have associations with heterochromatin i.e. acetylate a lysine cause chromatin to loosen while methylate the H3K9 would lead to tightening.

Remark 5.25. *Methylation of lysines cannot be generalize like acetylation. The reason for is that methylation of H3K4 is completely opposite to that of H3K9.*

Another important methylation event is that of H3K27, mediated by a **polycomb group**, which is also associated with heterochromatin formation. Finally, methylation of H3K36 is associated with transcription activation. Not only H3 that has these markings but also H4, etc. that can alter the transcription output. What we're now interested is **Which region of the genome is affected by these markings (histone modification)?** Well...we can carry out a ChIp-seq experiment to find out.

In this experiment, we will have antibodies that can recognize mono-, di- and trimethylated H3K4 of chromosome 17. These H3K4 are written as H3K4me1, H3K4me2 and H3K4me3 (for mono-, di- and trimethylated H3K4). We then run a chIP-seq with these 3 antibodies, then get a reading from NGS to see which region of the genome is are being affected. Not only would we carry chIP-seq and NGS with the H3K4, we will also carried out another one with TFs to see correlation between the 2 in order to activate transcription.

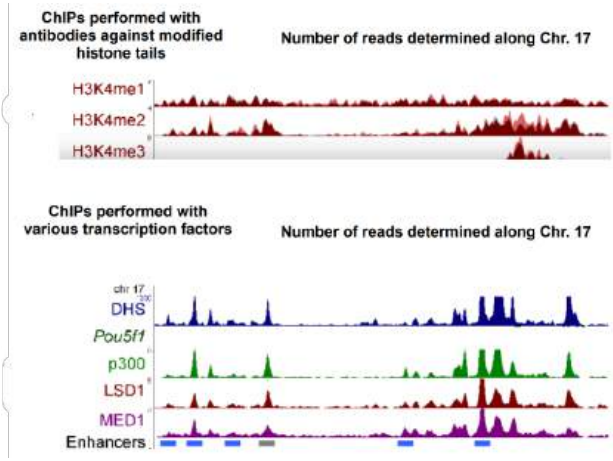


Figure 5.52: H3K4me1-3 chIP-seq along with various TFs.

As we can see H3K4me1 is varied through out the chromosome but also has peaks that are shared with H3K4me2. As for H3K4me2, it is mostly associated with enhancers and active regions around the proximal promoter

elements. Finally for H3K4me3, it is associated mostly with active promoters.

5.6.4 Co-Activators and Repressor

Over the year, scientist found that certain activators works alone and even sometimes work with other activators to get a maximal output. These additional factors could changes the conformation of the chromatin or histones around it which somehow enhance the effect of the original activator.

GCN4 is an important transcription activator in yeast. GCN4 has transcriptional activation domain that can interact with UAS of of yeast but what's more interesting is that GCN4 forms interactions with another complex called that **SAGA complex**. The SAGA complex comprises of multiple subunits that GCN4 can interact with which then facilitate transcription.

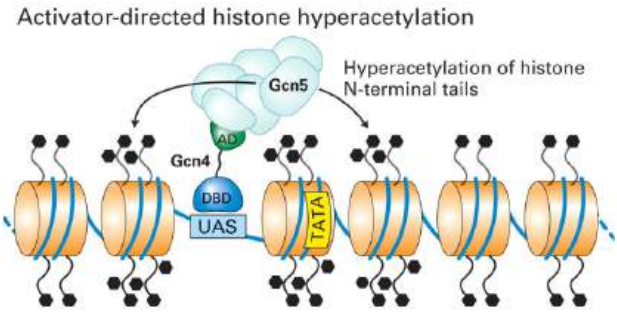


Figure 5.53: GCN5 works with GCN4 as a co-activator and add acetyl group to the histone tail.

It was later determine that main subunit that GCN4 interact is **GCN5** which later analysis shown was to be a histone acetyltransferase (HATs)! This is a big result because GCN5 is part of this SAGA complex which interact with GCN4, an activator, which then affects the chromatin around it through modification of the histone tails thereby enhances transcription. When an activator works along side another activators, we call it **co-activators**. (see Figure 5.53)

Remark 5.26. *GCN5 is an acetyltransferase which add in acetyl group while SIR2 is a deacetylase which remove acetyle groups.*

The reciprocal is also true that is having an additional repressor to enhance the effect of the original repressor. In this case, a repressor called

Ume1 can bind to the **upstream repressing sequence 1 (URS1)** that lead to repressing of a gene. Ume1 is found to also interact with a subunit called **Rpd3** on a large protein complex. Rpd3 is a **histone deacetylase complex (HDAC)** (like SIR2) that can remove acetyle group from the histone tails which increases the electrostatic interaction hence making the chromatin more compact. Rpd3 can also recruit *Sin3* that can help with setting up the complex. So we can see that when a repressor works along side another repressor, which enhances the output, is called **co-repressor**.

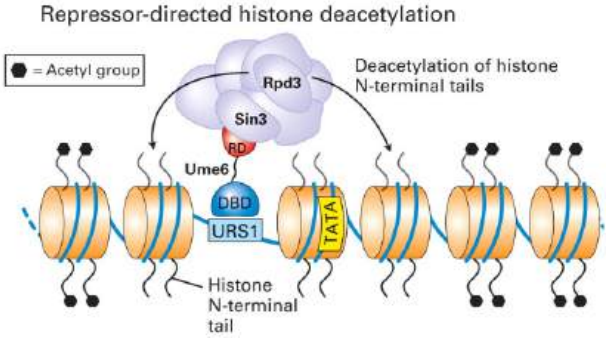


Figure 5.54: Rpd3 works with Ume1 as a co-repressor and remove acetyl group from the histone tail.

Activation Domain and Chromatin Decondensation

Investigators are now curious and wanted to see if the transcription activation domain (found in activator like GCN4) could have any effects on the chromatin configuration.
They decided to introduce a labelled **lac repressor (LacI)** into a cell with a highly heterochromatinized DNA. What they found was the LacI mostly settle or condense at a specific spot due to the heterochromatin. Now, investigators modified the LacI by adding a very well known and strong transcription activation domain VP16 which form **LacI-VP16**. They then introduce this LacI-VP16 (also labelled) and see what would happened to the same condensed chromatin.
Unlike from the previous 1 where the LacI would settle at a specific spot, the LacI-VP16 now scatters all through out the nucleus. This to show that the chromatin was modified and is no longer condensed, hence euchromatinized. This is due to the VP16 recruitement of other HATs that add acetyl group into the histone tail which loosen it. In addition to that, VP16

also recruited **chromatin remodelers**. All you need to know is that these chromatin remodelers are ATP-dependent proteins that can push chromosomes in a way that create a new configuration. Not only that, it can also push histone (in turns pushes a nucleosome) which allow the access of a specific region.

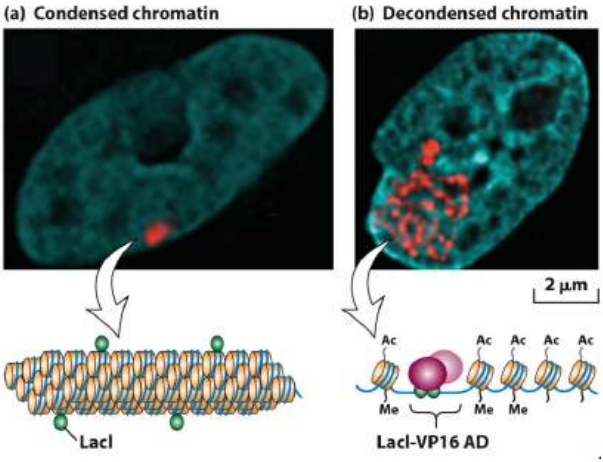


Figure 5.55: Decondensation of chromatin when adding LacI with addition of VP16, a strong transcription activation domain.

5.6.5 Epigenetic

Noticed that what we've done above was simply changing how genes are epxressed in a genome, we've never change the actual sequence of the genome itself.

Definition 5.14. An **epigenetic trait** is a change in genomic expression that does not result in altering the DNA sequence (unlike genetic traits).

The epigenetic trait are heritable which means it can be transmitted from the parental cells to the daughter cells.

Example 5.6.2. Inactive X chromosome is present in all female mammals. This is an epigenetic trait that is passed down from the maternal to the daughter where the X chromosome is silenced by heterochromatin.

Example 5.6.3. Developmental restriction is a silencing mechanism that ensure limbs or body part does not grow at a different spot. This is also

an epigenetic trait since there's no change in the DNA sequence but the expression of a certain gene that give rise to that limb is silenced at a different position.

Example 5.6.4. (genomic) Imprinting is a phenomenon where a gene is expressed or not depending if such gene is inherited by the paternal or maternal side. Once again, it is epigenetic since we do not modify this inherited gene, we simple either express or silence it via DNA methylation of cytosines.

The methylation of cytosines create a DNA mark that can be recognized by specific proteins which then allow the recruitment of Sin3. Sin3 can come in and modify the histone in its proximity so we can see that these DNA marks are not only methylation but also as a recruitment point for transcriptional output.
Many of these marks from the parental must be recognized and inherit by the offsprings. This is done through specific complexes that can either recognize or rewrite the epigenetic marks. This epigenetic modification can be done through either an **epigenetic writers** which can put down those marks e.g. histone methyltransferase can methylate H3K9; or **epigenetic readers** which can read recognize those mark. Although we made a distinction between them, it is possible that an epigenetic reader becomes a writer.

Example 5.6.5. Genes that are heterochromatinized will remain heterchromatinized after cell division due to epigenetic reader acting as a writer too. Take the trimethylation of H3K9 on a histone of specific genes. After the replication, the trimethylation will be dispersed and isn't always present on the histone.

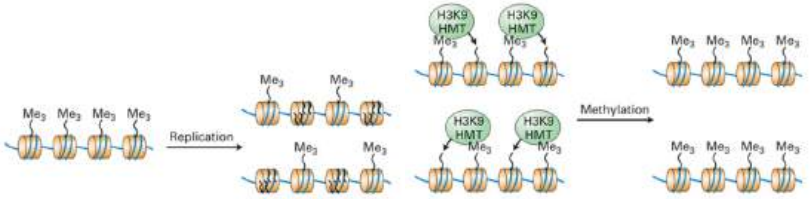


Figure 5.56: H3K9 HMT acts as epigenetic reader and writer to ensure post-translation H3K9 is trimethylated.

Epigenetic reader such as **H3K9 histone methyltransferase** can be recruited to histone with the H3K9me3 marks. Once the H3K9 HMT read and

recognized these marks, they will act as an epigenetic writer and ensure that all the neighbouring H3K9 will be trimethylated.

Pioneer Transcription Factors

There are TFs that take advantage of the modification of proximal histones mechanism and they are especially important for the for the initial stage of embryogenesis. They can activate transcription of specific genes that will distinguish various cell types. These TFs are called **pioneer transcription factors**.

> **Mechanism of Action:** Pioneer TFs can comes and interact to the DNA sequence even when it is heterochromatinized. This is probably because they can bind to DNA that are exposed on the outside of the nucleosome. The interaction of pioneer TFs is so strong that it lead to the unwinding of the chromatin. At the same time, they can also recruit in co-activators, mainly HATs, which can lead to histone acetylation thereby open up the chromatin even further.

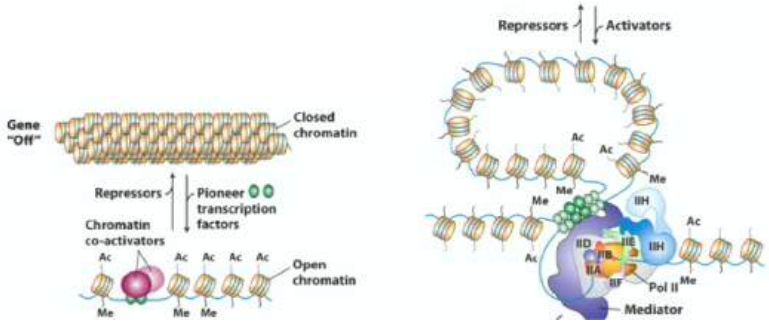


Figure 5.57: Pioneer TFs opens up heterochromatin along with HATs which allow general TFs come and further activate transcription. (left to right)

After this is mediotor would be recruited because of all the transcription activation domain. RNA polymerase would also comes and begin transcription. That's how we theorize the very early transcription reaction during embryogenesis where the chromatin seems to be tightly wound and condensed.

5.7 RNA Processing I

To understand RNA processing, we would need to go back and take a look at RNA pol. We know that eukaryotes have 3 different RNA pol (I, II and III) that shares the following features:

- They're made from multimeric proteins complexes (subunits).
- Some of these subunits are **homologous** (similar in structure and physiology) with the bacterial RNA pol.
- All of these subunits are essential

Nevetheless, they have some distincts features such as: RNA I transcribe rRNA, RNA III transcribe tRNA and RNA pol II is transcribe mRNA (which is the main transcripts) and some other important RNAs. We'll focus mostly on the RNA pol II and we can see that there are some distinguish features that we've been mentioning over and over again, that is the C-terminal domain (CTD). CTD is high important for a number of down stream reaction which we'll be talking about.

5.7.1 Structure and Functions of CTD

The CTD has a number of these heptapeptide repeats (YSPTSPS) and it's is repeating 52 times in humans and 26 in yeast. Interestingly enough, if you were to remove 1 of these repeats in yeast, it won't be able to grow and die. So not only it's unique to the RNA pol II, it's essential for life.

Figure 5.58: Enter Caption

Each of the repeats of YSPTSPS (left to right $\Longleftrightarrow$ position 1 to 7) form a heptapeptide but what's important to know is its phosphorylation by TFIIH. At initiation, TFIIH will phosphorylate all of the serine at position 5 on all of those repeats in the CTD. Strangely enough, a second phosphorylation event will take place on the serine at position 2 and is not caused by TFIIH during elongation. This event is mediated by another protein kinase which plays a critical role in switching from post-initiation to elongation.

In the initiation, all the serine-5 will be phosphorylated due to the action of TFIIH. Then the RNA pol II will move away from the the promoter site, leaving only maybe TBP behind. RNA pol II will continue to transcribe for around 100nt then it **STOPS** then it continues again. For years, this pausing confused researchers and they thought it was a faulty of an RNA pol II but in fact, it isn't i.e. this pausing event is essential.

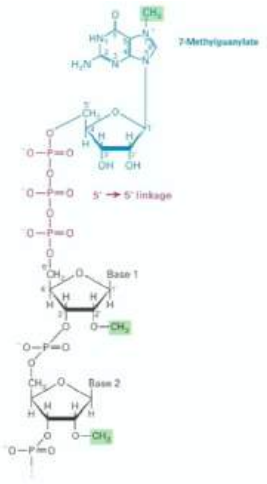


Figure 5.59: Addition of 7' methylguanylate CAP and 2'OH methylation.

Going back to the first phosphorylation event, the phosphorylated 5-serine on the CTD creates a platform for the recruitment of different enzymes for modification of the nascent mRNA. This enzyme is called the **capping enzyme**. This capping enzyme will add on a *7' methylguanylate CAP* to the 5' end of the nascent mRNA (as it leaves the RNA pol II) via a **5'-5' triphosphate linkage**. Not only that, the 2'OH group is also methylated and in some vertebrates, the second nt 2'OH group will also be methylated.
The addition of CAP would protect the pre-mRNA from ribonuclease but also facilitate nuclear export and recognition by translation factors. So we can see that the reason that RNA pol II stalled is allowing the capping enzyme to modify the emerging nascent mRNA.

Well then **what causes or how does RNA pol II stop?** The following will describe the entire mechanism of pausing and restarting of RNA pol II.

Mechanism of Action: First, after RNA pol II have left the PICs, a protein called **negative elongation factor (NELF)** and its associate **DRB sensitive inhibitor factor (DSIF)** will bind to the RNA pol II. NELF acts like an antagonist that block rNTPs from going into the catalytic site which stall the RNA pol II. After stalling, a protein kinase called **CDK9** coupled with T-cyclin to form **P-TEFb** which can come and phosphorylate DSIF, NELF and the serine-2. Phosphorylation of NELF will make it leave and **PAF** and **SPT6** will substitute

in. Phosphorylation of DSIF will change its conformation and making it clamp on to the RNA pol II which makes it more processive to the template.

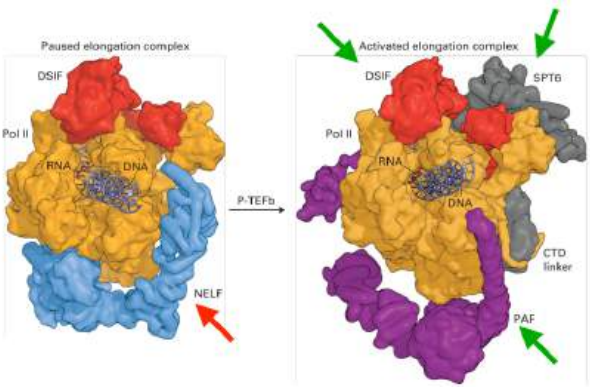


Figure 5.60: RNA pol II and its associated proteins during and after pausing.

Remark 5.27. *DRB is an important drug inhibitor of elongation.*

The phosphorylation of serine-2 is essential as it will lead to the recruitment of lots of proteins such as splicing factors, phosphoadenylation factors and export factors.

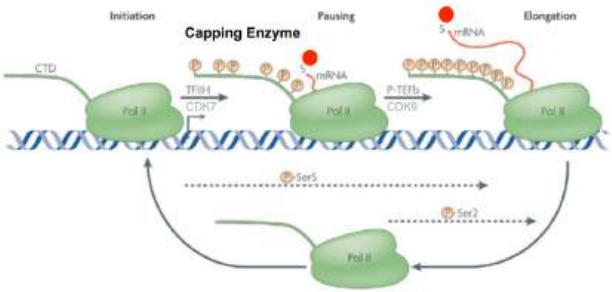


Figure 5.61: Mechanism of stalling and restarting of RNA pol II.

After this stalling and starting, the RNA pol II will become highly efficient and it will fully turn into elongation.

5.7.2 Discoveries of Introns

The newly made *nascent* RNA or **pre-mRNA** must be modified before it'll be translated by rRNAs. Typically, the process of maturation is through removing bits of the mRNA called *introns* while leaving the *exons*.

Remark 5.28. *Most eukaryotic genes can have introns which is opposed to bacterial gene who has none. Mammalian has lots of large introns as compared to yeasts who have more compact ones.*

Remark 5.29. *Introns are not junks. When sitting on the DNA template, they harbour information such as enhancers and regulatory elements!*

Through some experimentation and comparison gene and its correspondent cDNA; they realized something really odd, that is **why is the cDNA smaller and has less sequence as compared to its template gene?** Well...this is due to the presence of introns. The discovery of introns due to the discrepancy between the mRNA (in experimental case cDNA) and its gene.

To perform this experiment and see what introns are like, we'll be needing a few component. First, we would need a viral gene, in such as it's the *adenovirus hexon gene*, an mRNA that correspond to its gene.

> **Procedure:** First, we extract the viral gene with formamide at high temperature. This extracted viral dsDNA will be hybridized with its corresponding mRNA. Then hybridization structure will be capture under electronmicroscope (EM micrograph).

As a result, you would expect to see the mRNA binds strongly to its complementary DNA segment of the viral gene. You would also see that there are these large segment of DNA that lies in the middle of the mRNA that are not complementary to it. Essentially, the mRNA interacts with the DNA regions that encodes for its exons and the non-complementary DNA regions encodes for the mRNA introns that were spliced out.

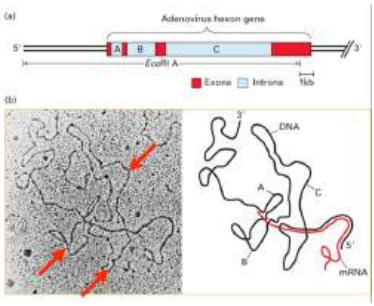


Figure 5.62: Discovery of introns from RNA-DNA hydridization.

Now the main question remain is **where are those introns?** well...researchers

in the early days were able to identify regions that are a bit "off" and don't corresponds to anything, they are the **introns boundaries**. We can pick out these boundaries according to the cDNA and saw that it was found there are some important sequences/structures that are conserved and essential for the splicing of introns.

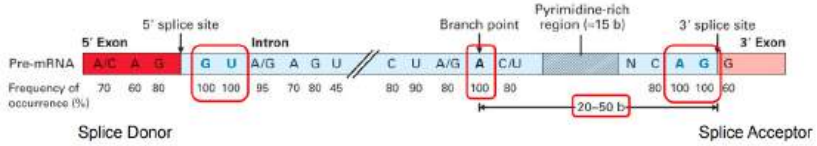


Figure 5.63: Conserved structures (red boxes) of introns.

The 5' and 3' ends has nucleotides that are highly conserved. On the 5' end of introns, it is always a dinucleotide **GU sequence** while 3' end is always **AG sequence**. There's also a pyrimidine rich regions called **polypyrimidine region** (downstream to 3') and upstream from it is always a conserved adenosine nt and it is called the **branch point**.

5.7.3 Spliceosomes and Splicing Reaction

Definition 5.15. **Spliceosomes** are large RNA and proteins complexes whose main function is to carry out introns splicing.

The RNA that are part of the spliceosome is called the **small nuclear RNA (snRNA)**. These snRNAs have associated proteins (splicing factors, 6-10 proteins) and the association of these 2 create the **small nuclear ribonucleoproteins particles (snRNPs)**. Each spliceosome consists of 5 snRNPs which in turn consists of snRNAs (U1, 2, 4, 5 and 6) and 6-10 other associated proteins.

The U1 snRNA is important for the interaction with the GU sequence (introns) as well as some upstream sequences (exons). U2 snRNA will bind to the pyrimidine rich region but it will leave out the branch point **(see Figure 5.64)**. There must be at least to snRNAs that bind to the introns in order for it the be excised (in this case U1 and 2 snRNAs).

Base pairing of U1 snRNA are cirtical for the region because experimentation shown that if you introduced a mutation to that region, splicing cannot occur. An important experiment was done by transfecting those regions with a mutation which would block the splicing activity of snRNA.

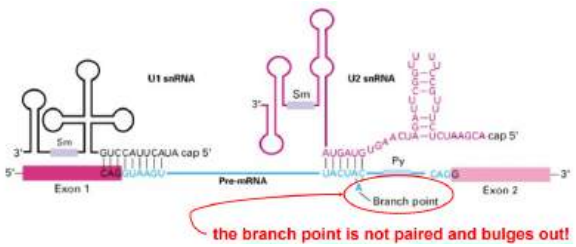


Figure 5.64: Base pairing of U1 and 2 snRNAs

However, if you *co-transfect* a cDNA that gives rise to a snRNA that is complementary to the mutation, splicing occurs.

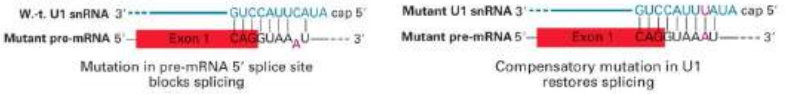


Figure 5.65: Mutation of the pre-mRNA and recompensation of the snRNA.

Now we will look at the general splicing reactions of an introns. The splicing of introns typically involved in 2 *trans*-**esterification** (forming an ester on opposing side).

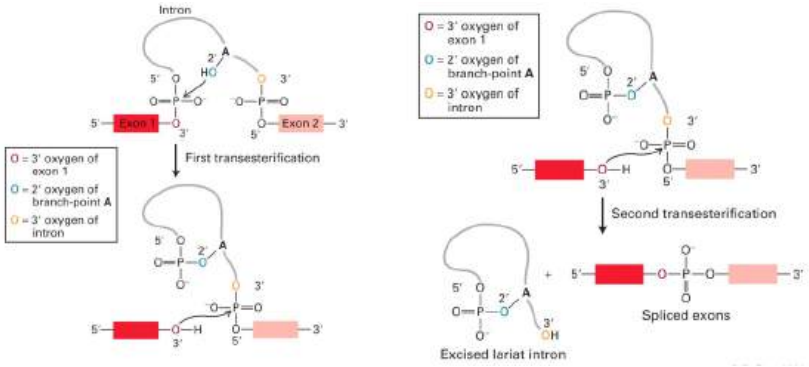


Figure 5.66: General mechanism of introns splicing.

Mechanism of Action (Splicing Reaction): First the OH group of the branch point will do a nucleophylic attack on the PO_3^- of the 5' G residue of the introns (first esterification). This attacking lead to the formation of a ***lariat***. The free up OH of the 3' will then attack the PO_3^- of the first residue of the exons (second esterification). This attack leads to the 2 exons bind together via a phosphodiester bond while the introns lariat is released.

In fact, using radiolablled RNA substrate, we can extract or separate each intermediate products of this reaction *in vitro*.

Now the above description was just the general reaction itself but the majority of the splicing reaction need to incorporate the spliceosomes i.e. what we see above is a chemical reaction but what really happened is a biochemical reaction which is a chemical reaction carried out by these snRNPs thus form a spliceosome.

The pathway of which a spliceosome is form and splice out the introns is called the **spliceosome cycle**.

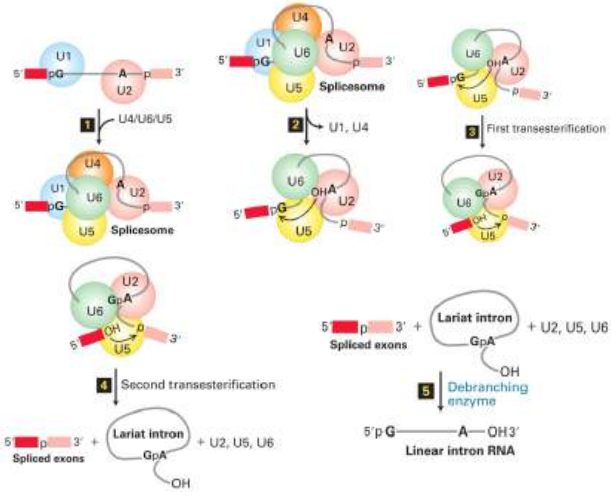


Figure 5.67: Spliceosome cycle.

Mechanism of Action (Spliceosome Cycle): First U1 and U2 snRNPs will come and interact with their respective boundary. This interaction also recruit U4,5 and 6 snRNPs. This form the spliceosome complex but also the rearrangement of RNAs interaction lead to its activation. During its activation stage, U1 and 4 snRNPs leave the complex. The rest of the snRNPs are now free and active to carry out the splicing reaction (described above). Once splicing is done, the U3, 5 and 6 snRNPs will dissociate from the exon joining site while the lariat introns will degraded.

Remark 5.30. *Degradation of lariat introns cannot be carried out right away since it's in a loop conformation. The lariat introns must be first linearize by* ***debranching enzyme*** *after which it will be degraded by exonucleases.*

Self-Splicing Introns

Previous studies have shown that under certain condition, the splicing reaction can carry out the splicing reaction by the RNA themselves i.e. no help from the snRNPs.
In the study, researchers try to detect linear and cyclic introns lariat under many conditions. First is the control with and without Mg^{2+} and it was found that the splicing reaction must have Mg^{2+} or else no cycling nor linear introns would form.

Remark 5.31. *This is not much news since we know that the majority of biological reactions requires not only enzymes but also* ***cofactors*** *like Mg^{2+}, Cu^{2+}, etc.*

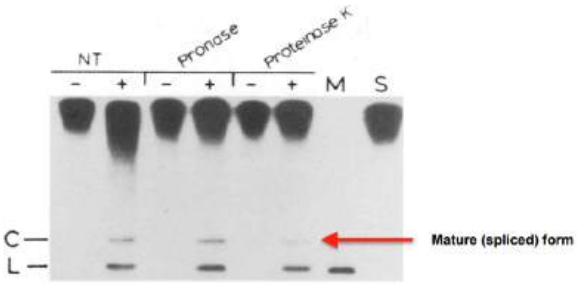


Figure 5.68: Experiment shows that spliced introns are presence without snRNPs.

What news for the researchers was that if they put proteinase (enzymes that degrade proteins) subsequently destroy the snRNPs, both cyclic and linear introns could still be detected. This shows us that some of these RNA can be catalytic and carry out the splicing reaction by themselves.

These introns that can carry out the splicing is called **self-splicing introns**, they're also divided into group I and II. We won't look at the mechanism of the self-splicing introns since it's identical to that of the spliceosomes. Nevertheless, the main difference we can take home is that splicing reaction carried out by spliceosome is much more efficient that self-splicing introns.

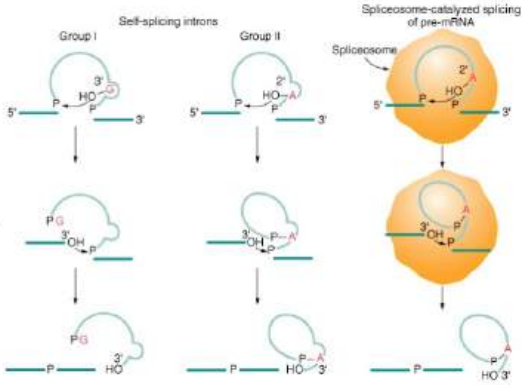


Figure 5.69: Introns can be spliced out by the spliceosomes and self-splicing introns group I and II.

Nevertheless, we can look at some structural similarities between the spliceosome and the self-splicing introns. On the side of the of the spliceosomes, we can see the snRNAs of the snRNPs form these branches where the RNAs self-complement to one another. On the side of the self-splicing introns, we can see the introns themselves form similar branching structures that are also self-complement to itself. Essentially it forms secondary structures and loops

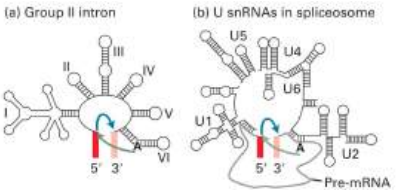


Figure 5.70: Self-splicing introns vs spliceosomes structures.

Remark 5.32. *Group II self-splicing introns are only represent in mitochondria and chloroplast which is also thought to the be predecessor of other introns.*

5.8 RNA Processing II

The splicing of mRNA would lead to the its maturation (pre-mRNA to mRNA). However, it's not only mRNA that will be matured or processed, other RNA such as rRNAs and tRNAs would also be processed.

5.8.1 Processing of rRNAs and tRNAs

RNA pol I is responsible for the transcription of rRNAs in the *nucleolus*. rRNAs are formed from a cluster of DNA called **rDNA**. This rDNA is recognized by RNA pol I and its general TFs. The transcription will give rise to a **large pre-ribosomal RNA**. This large pre-rRNA would be processed by splicing out specific segments.

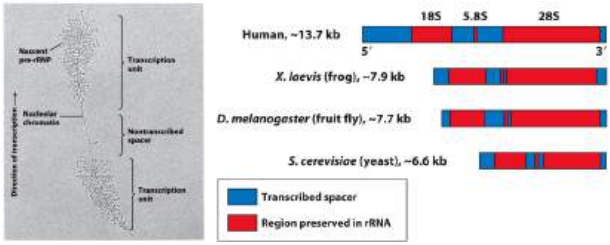


Figure 5.71: rRNAs transcription and cleavage

Remark 5.33. *Interestingly enough, RNA pol I will transcribe from the rDNA in such a way that its sequence is conserved for all eukaryotic organsm e.g. the pre-rRNA transcript of human is similar to that of the yeast. From the image we can see that the 18S, 5.8S and 28S, in that order, for all eukaryotes.*

Strangely, if you were to introduce 1 copies of the rDNA into a cell, supposedly that of *Drosophila*, the RNA pol I will recognize them and start transcribing so much RNA that it form a nucleolus-like structure.

Not only rRNAs was processed but also that of tRNAs. They won't go through splicing like that of rRNA but they're heavilty processed. Certain

regions on the pre-tRNAs will be removed, such as the 5' residues, and modified, such as the di-uridine residue on the 3' to a ACC tri-nucleotides. There are also chemical modification of internal bases such as chemically change a U to a Ψ (pseudo-uridine).

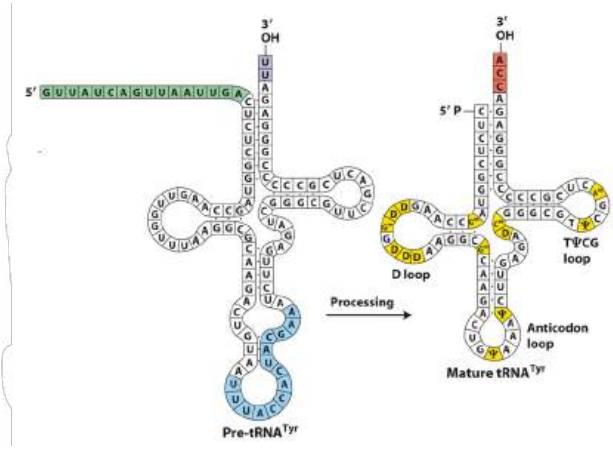


Figure 5.72: Processing of tRNA.

Remark 5.34. *These modification would turn the pre-tRNA to a matured tRNA that is important for the synthesis of protein.*

5.8.2 Proteins and RNA Splicing

Now we will return to the splicing event. We know from before that splicing reaction is carried out by these snRNAs (or self-splicing) but what we've noticed from years to years is that some of these events would require other proteins to help.

These proteins would have different domains one of which can bind and interact with RNAs called the **RNA binding proteins**. There are many of these domains such as KH and RRG domain.

Example 5.8.1. RRM domain has positively charge $\beta-$sheet that can bind and interact with the negatively charged RNA (due to the phosphate).

An important RNA binding proteins is called the **U2AF protein** that participate in the splicing reaction. It has 2 subunits: 35 and 65kDa. The 35kDa

subunit interact with the 3' end of the intron (AG sequence) that's going to be spliced out while the 65kDa subunit will bind to polypyrimidine of the introns. We can see that U2AF proteins are helpful to distinguish the 3' end of an intron which helps with splicing efficiency.
Nevertheless...there's a problem. Introns can be enormous (up to 0.5 million nt) and is much larger than exons, **so how can these RNA binding proteins find these sequences in these enormous introns?** Well...Cells depend on other proteins that can facilitate this process and they are called the **SR proteins**.

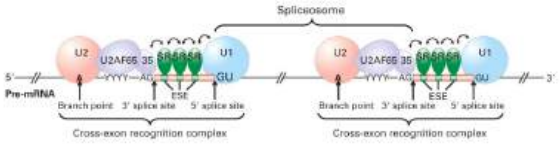


Figure 5.73: Proteins that help facilitate splicing

Remark 5.35. *They're called SR proteins is because they're rich in serine and arginine.*

SR proteins can recognize sequences and structures on the exons called the **exonic splicing enhancers (ESE)**. RNA regions that are ESE will be covered with these SR proteins. Other proteins such as U1 snRNPs and U2AF has a high association with these SR proteins which means that they will bind to the RNA sequences that is the closest to the SR proteins which is where the true splicing site is. Furthermore because U2AF has found its binding site, U2 snRNPs also interact with it to bind its branching point.

Remark 5.36. *All of these proteins interactions form the **cross-exon recognition complex** and they help with the spliceosome to carry out a splicing reaction.*

5.8.3 Alternative Splicing and Sex Determination

The story that lead to alternative splicing originate from looking at genes from genomic mapping
The first organism that has genome mapped out (*Haemophilus Influenzae*) was predicted to have 6,000 genes which is roughly correct. We then sequenced for *C. Elegans* (roundworm) which result around 20,000 genes and it sort of makes sense since they're pretty simple. We then theorize

that human would have up to or even more than 100,000 genes! Unfortunately this is not the case, we only have around 20,000 genes which is the equivalent to roundworms! This is bizarre because **how can we build our complex morphology with this much gene?** Well...this is due to the fact 1 gene can give rise to multiple different RNA which is different proteins.

Example 5.8.2. Fibronectin gene is a gene that has sequences that encode for the stciky domain on proteins (EIIIA and EIIIB). It is expressed in fibroblast and hepatocytes (liver cells) but the resulting mRNA are different.

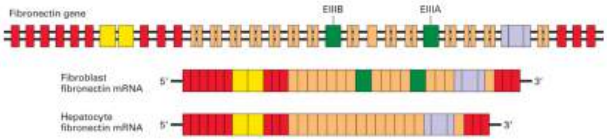


Figure 5.74: Fibronectin alternative splicing in hepatocytes.

In fibroblast, EIIIA and B are expressed which makes sense since it will stick cells together but in hepatocytes EIIIA and B are spliced out, which also makes sense since you do not want these proteins be sticky in the blood stream. The resulting EIIIA and B spliced out from hepatocytes is due to alternative splicing.

Remark 5.37. *Alternative splicing isn't the same as polycistronic arrangment. In polycistronics 1 mRNA would give rise to multiple protein species at once however alternative splicing will give rise to 1 protein species only.*

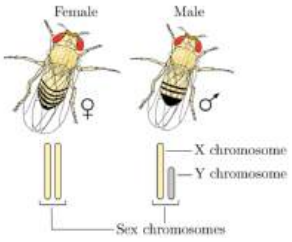


Figure 5.75: Sex chromosome difference between male and female *Drosophila*.

There are sexual dimorphic characteristic of male and female *Drosophila*. The differentiation between male and female is due to a cascade of alternative splicing. Before talk about alternative splicing that lead to the sexual determination of *Drosophila*, we need to address some understand and terminology.

Female *Drosophila* (and even mammals) have 1 pair of X chromosomes while the male have 1X and 1Y chromosomes (these are sex chromosomes). This disparity in X chromosomes in the male *Drosophila* would lead to a reduced gene expression. Through evolution, a mechanism was developed to counter this

problem by which the X chromosome in the male would be upregulated (produce twice the gene expression) to match that of female.

Definition 5.16. The biological process where an organisms equalize their gene expression due to a chromosomal disparity is called **dosage compensation**

Remark 5.38. *Dosage compensation's mechanism can vary from species to species. In human, dosage compensation works by down regulate the 1 of the X chromosome in the female while in Drosophila is upregulated the X chromosome in the male.*

Definition 5.17. **Sex lethal (Sxl)** is an RNA binding proteins that is important for the cascade of sex determination in *Drosophila*.

Mechanism of Action (Sex Determination of *Drosophila*): In early embryogenesis of female *Drosophila*, the promoter of the Sxl gene is activated lead to the transcription of Sxl (male doesn't make early Sxl).
1. In later development of female, the Sxl will bind to the site and block U2AF on the mRNA (that would give rise to Sxl). Then, during alternative splicing, for female, the splicing will skip an exon due to the binding of Sxl and this mRNA product will give rise to more Sxl. For male, because there's no sxl, the mRNA encodes for an early stop codon (in the exon that female excluded) thus making no sxl.
2. The Sxl will then bind to another pre-mRNA that was encoded by **transformer (Tra)** gene. In female, this binding will lead to 1 of the exon being excluded during alternative splicing and this lead to the mRNA translated into Tra proteins. In male, the pre-mRNA will go through alternative splicing with all of its exon intact, the mRNA encodes for an early stop codon (in the exon that female excluded) thus making no Tra proteins.
3. Finally, in female, the Tra proteins along with Tra2 and RBP1 will bind to the ESE of 1 of the exon on the pre-mRNA, that is encoded by **double sex (dsx)** gene. This binding, in female, will lead to the exon being *included* during alternative splicing that give rise to the female dsx proteins form. In male, there are no Tra proteins which means that the exon (that female included) will be excluded and after translation it will give rise to a male dsx protein form.

Remark 5.39. *When zygotic transcription is activated, both sexes will make Sxl.*

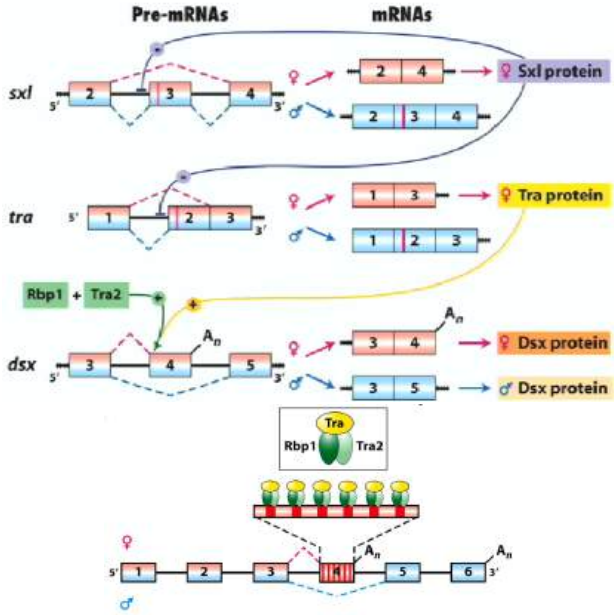


Figure 5.76: Mechanism of sex determination of *Drosophila* and alternative splicing.

5.8.4 RNA Editing

Sometimes RNA is modified in a way that is different from alternative splicing, and it is called **RNA editing**.
RNA editing is unique for plastids such as mitochondria or chloroblast but is rare in normal nuclear genome. RNA editing depends on an enzyme called *deaminase* that carry out a **deamination reaction** that convert A to inosine (I) and C to U.

Remark 5.40. *The DNA sequence gives rise to normal mRNA but the RNA itself when interact with these deaminase would lead to alteration.*

You might wonder **then what's the different between RNA editing and modification on the tRNA? are they the same?** Well...no they're not the same, RNA editing always involved in changing A to I and C to U but tRNA modification can change stuffs like U to Ψ.

There are situation, where RNA editing arises, that are regulated in eukaryotes even when it's uncommon such as the editing of mammalian apolipopro-

tein B.
Alipoprotein B (apoB) is the primary component of a *chylomicron* which is used to carry lipids in the blood stream to the liver.

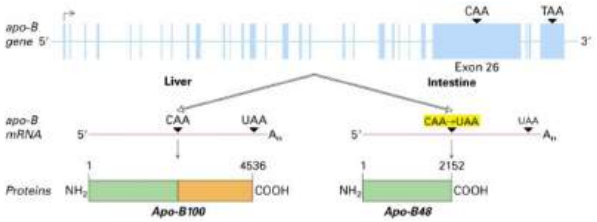


Figure 5.77: ApoB transcript made by liver vs intestine.

ApoB is synthesized both in the liver and the intestine but the amino acids component of it is different. The apoB synthesized in the liver are large apo that consists of ~ 4536 amino acids residues. They function as the transport for lipid. The apoB synthesized by the intestine is half its size. The reason for this is because there's a frame stop caused by RNA editing (turns $CAA \longrightarrow UAA$ [stop]) in the middle of the mRNA. This stop codon generation in the middle of the mRNA transcript will lead to translation stop midway thereby reduce the apoB in the intestine ~half the size.

5.8.5 Polyadenylation

A final modification event happens for mRNA transcript is **polyadenylation**.

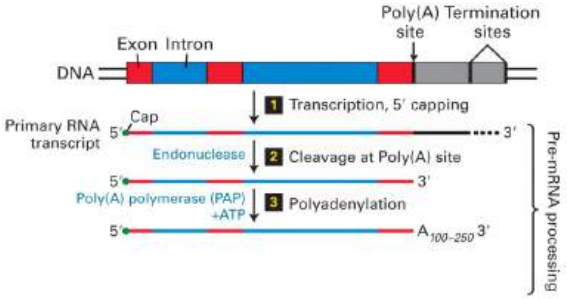


Figure 5.78: Poly A site from gene to mRNA

Through examination of viral transcripts, we typically found that transcription of a gene would pass through a poly-adenosine (poly A) site before it reaches the termination site where the mRNA is released for translation. What researchers found too is that the matured mRNA misses these transcripts beyond the poly A site...this must mean that there is a cleavage a the poly A site on the mRNA.

When we examine the 3' untransltered region, you'll notice a specific a region, 5' to where the poly A tail would be, has the sequece **AAUAAA**. This is the **poladenylation signal** which combine to some other G or U sequences downstream will lead to the recruitment of proteins called the **cleavage and polyadenylation signal factors (CPSF)**.

> **Mechanism of Action (Polyadenylation):** The CPSF, as well as other proteins, will come to the poly A signal and cleave at the poly(A) site.
> **1.** Immediately after the cleavage is established, the poly(A) tail is added on by **poly-A polymerase (PAP)**. PAP is not efficient so it will add around 12nt only.
> **2.** Then, another protein called **Pol-A binding proteins (PABPN1)** is recruited and change PAP into a more efficient one. This conformational change would make the PAP add up to 200 A residues.

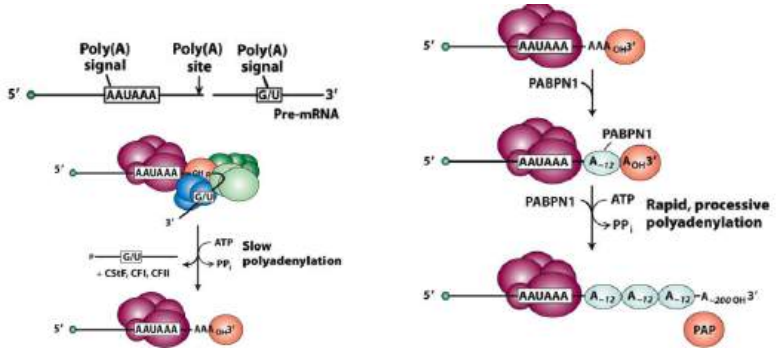


Figure 5.79: Process of polyadenylation.

Remark 5.41. *PABN1 works exclusively in the nucleus (hence it has N in its*

name), there's another variance of pol(A) binding proteins that can work in the cytoplasm.

After polyadenylation, the mRNA is now less susceptible to 5'-3' exonuclease. Essentially polyadenylation help to prolong the degradation of the mRNA. This nascent mRNA has gone through many modification and finally become a matured mRNA.

Remark 5.42. *Not all mRNAs are polyadenylated. Histone mRNAs are not polyadenylated but they have secondary sequences that fold at the 3' end.*

5.8.6 Divergent Transcription

If we recall a long ago, we've said that transcription can go in 2 direction due to *divergent promoter*. We specify the "correct" direction as the *sense* direction while the other is *antisense*.
What we found is that the level of transcript in the sense direction accumulate much more than that of the antisense i.e. more transcript is made in the sense direction. Not only that, the sense transcript is much more stable than the antisense's i.e. doesn't get degraded easily. But then **why? why would the sense direction better than the antisens?** Well...further investigation on the transcripts itself reveal many reasons.

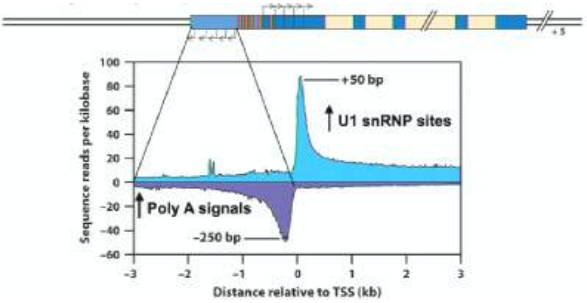


Figure 5.80: Divergent transcript level of sense and antisense direction.

When look at the antisense, researchers found that it has more early poly A signals. This early poly A signals would lead cell to detect it as a faulty transcript and degrade it. On the sense direction, the poly A signals are at the very end where the terminal signals would be. It also has an enrichment of U1 snRNPs signals. Because of the correct signalling in the sense direction, its transcript's stability is higher.

5.9 Nucleo-Cytoplasmic Transport

All of the mRNA post-modification are in the nucleus but they need to be transported out of the nucleus to be translated in the cytoplasm.

Definition 5.18. A cell's **nucleus** is a membrane-bound organelle that house the genetic materials of the cell.

As the definition suggested that it's membrane-bound, the nucleus blocks the movement of macromolecules using its double envelope. The only way to access through the nucleus is via the **nuclear pores**.

When we look at it under electron-microscope, we can see these holes on the nuclear envelope which is the nuclear pores. On the cytoplasmic side, the nuclear pores looks like a hole with filaments hanging out from it. On the nuclear plasmic side, the nuclear pores look like a basket, almost like a champagne's *muselet*.

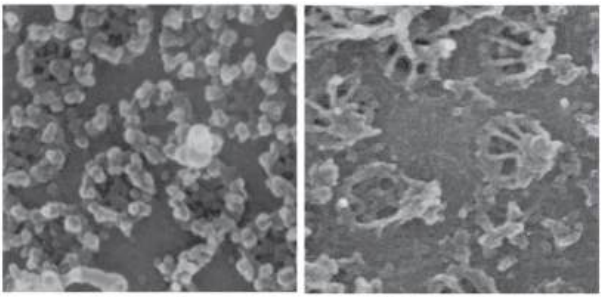


Figure 5.81: Nuclear pore on the cyto- and nuclear-plasmic side.

The pores are held up by the **nuclear pores complexes (NPCs)**. They're highly ordered structured and is 30x bigger than typical ribosome (MW of around 156MDa). Molecules of MW from 40-60kDa can freely travel through the proteins however larger molecules must be transported. This NPC composes a variety of proteins, around 50 for yeast and up to 100 for vertebrates. Out of all of these proteins, the main class that we'll look into is called **nucleoporins** more specifically **FG nucleoporins**

FG Nucleoporins are distinct since they have lots of FG (phenylalanine and glycine) repeats. FG nucleoporins will make up the middle of the pore where molecular traffic will move. Because of the FG repeats, the interaction they have on one another lead to them having an instrincally disor-

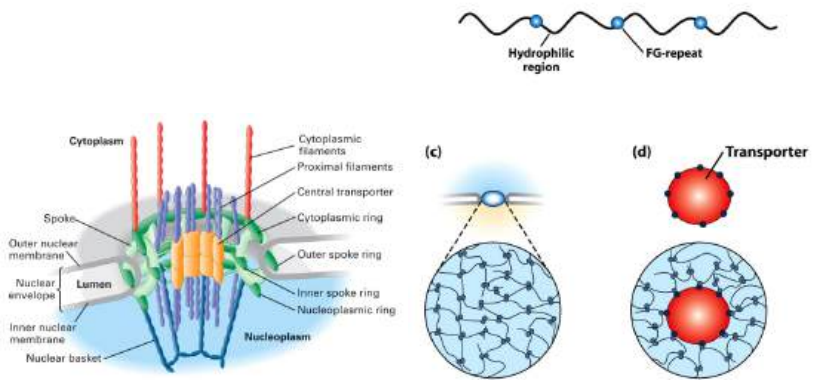


Figure 5.82: Nuclear pore complex and nucleoporins.

dered state. They form this unstructured gel-like matrix. Proteins that have domains that can interact with these disordered domains can pass through the pore with ease.

5.9.1 Proteins Transport

It was found that proteins synthesized in the cytoplasm that subsequently move into the nucleus has specific sequences which are referred as **nuclear localization signals (NLSs)**. These NLSs somehow lead to the transport of these cytoplasmic proteins into the nucleus.

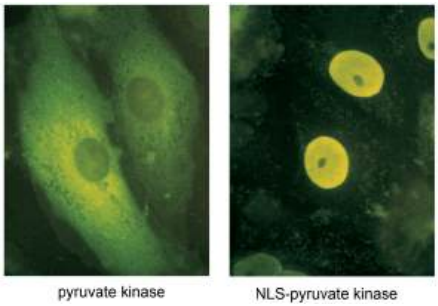


Figure 5.83: NLS discovery via pyruvate kinase.

The discovery of NLS is through investigation of SV40's (virus) T-antigen, that can enter the nucleus and cause havoc. Researchers found that mu-

tation in some proteins of the T-antigen would disable it from entering the nucleus. They decided to takes these proteins sequences and inserted onto a cytoplasmic proteins (pyruvate kinase). This insertion lead to the pyruvate kinase migrates into nucleus, this means that these proteins sequences has to do something with transporting through the nucleus (NLS). Through investigation, researchers were able to find the proteins that interact with these NLS of the transporting proteins (we refer as "*cargo*"). First is the **RAs-related nuclear proteins (RAN)**, which are G-proteins that can exist in 2 conformation depending on whether it's GTP or GDP bound. Another important types of proteins that will participate is **nuclear transport receptors**, also referred as *importins* or *exportins*,which are proteins that bind to the cargo's NLS and faciliate the transport through the nucleoporins by associate with its disordered domains.

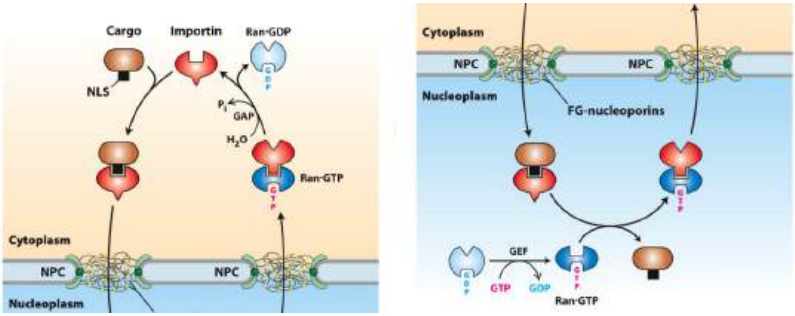


Figure 5.84: Mechanism of proteins import

Mechanism of Action (Nucleoplasmic Import): We first begin in the cytoplasm.
1. In the cytoplasm, the importin will bind with the cargo at the NLS domain and bring it through the NPC into the nucleoplasm. This movement is driven by the concentration gradient of the import complex, $[\text{importin}]_{\text{out}} > [\text{importin}]_{\text{in}}$. In the nucleoplasm, RAN-GDP will become RAN-GTP via with the help of GEF. RAN-GTP will interact with the import complex and displace the cargo from the importin in place of itself.
2. Now, in the nucleoplasm, the RAN-GTP importin complex, driven by concentration gradient, will move through the NPC into the cytoplasm. In the cytoplasm, this complex will undergo GTP-

hydrolysis faciliated by GAP. This reaction dissociate the complex into the importin, which can go back to bind to the cargo's NLS, and the RAN-GDP.

Remark 5.43. *The binding of GTP to RAN and hydrolysis of RAN-GTP to RAN-GDP both causes a conformational changes which is why it can bind or release the importin.*

Remark 5.44. *Both movement is caused concentration gradient which is maintained mainly by the GTP hydrolysis in the cytoplasm and the generation of GTP in the nucleoplasm.*

The similar mechanism of RAN-mediation is used by the nucleoplasmic export but the nuclear transport receptor that it will use is **exportin 1**

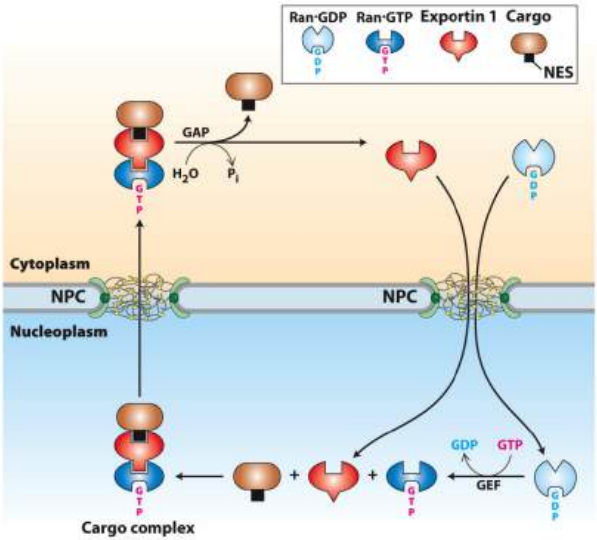


Figure 5.85: Mechanism of proteins export

Mechanism of Action (Nucleplasmic Export): Unlike import, we first being in the nucleoplasm.
1. In the nucleoplasm, RAN-GDP turns into RAN-GTP thanks to

GEF. RAN-GTP, exportin 1 and the cargo will form a tertiary complex and export the cargo through the NPC into the cytoplasm.
2. In the cytoplasm, the tertiary complex will be GTP-hydrolyzed, facilitated by GAP. This hydrolyzation will lead to the RAN conformational change into RAN-GDP which dissociate the entire complex. The RAN-GDP and exportin 1 will then move back into the nucleoplasm to be used for export again.

5.9.2 mRNP Transport

tRNAs and rRNAs use the similar mechanism of exportation to get out to the cytoplasm for translation (exportin t is the main exportin of tRNAs). Some mRNAs will use this similar mechanism e.g. mRNAs that are associated with hnRNP proteins but most of them will be exported by a RAN-independent mechanism.

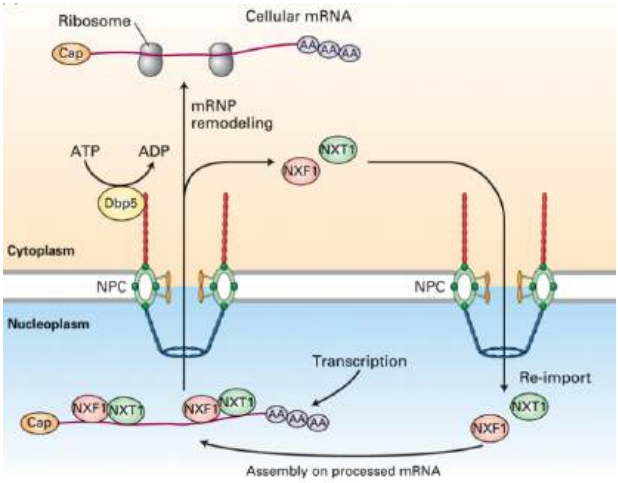


Figure 5.86: Mechanism of mRNP transport.

Mechanism of Action (mRNP Export): The main exporter used in this are 2 subunits called **NXF1** and **NXT1**. When meant mRNP as the mRNA and its associated proteins. In the nucleoplasm, NXF1

and NXT1 will recognized the matured mRNA and interact with its associated proteins such as SR proteins. These 2 exporter will bring the mRNP to through the FG nucleuporins by interact directly to its intrinsically disordered domains.

To visualize what happened during this mRNP transport, we took micrograph of an important mRNP that will develop into a sticky substance that help to glue the eggs of *Chironomous tentans* onto leaves and such. The transcription of this gene would lead to the final mRNP takes of a conformation similar to that of a croissant. This croissant-like mRNP will then go to the nuclear pore and move through it. What we've found interesting is that immediately as the mRNP move to the cytoplasm, ribosomes form and begin translation of the mRNP immediately. This means that there's a sense of directionality, 5' end comes out first, when the mRNP is transported through the nuclear pore.

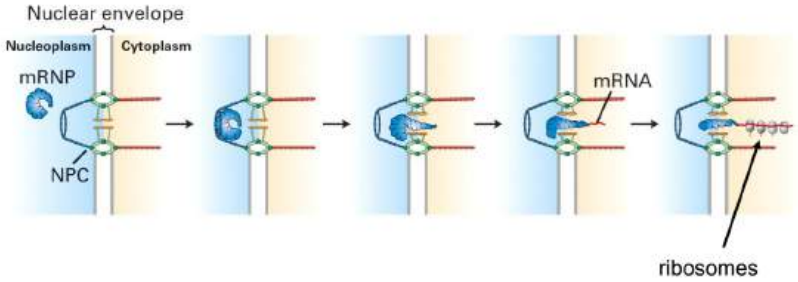


Figure 5.87: mRNP direct translation when exiting the nuclear pore.

Essentially what happened is that the mRNA with its associated proteins will fold into each other into an mRNP (croissant-like structure). This mRNP will move through the nuclear pore where it will be immediately translated.

5.9.3 Cytoplasmic Remodelling

Coming back to the idea of the mRNP, observations have been made that as these mRNP moves to the nuclear more, most if not all of the proteins will be removed e.g. the exporter proteins as it moves the 5' end to the cytoplasmic side, RNA helicase (found on the cytoplasmic filament) will cause it to dissociate; even the proteins that are associated with the poly A

tails are removed.
As it moves all the way through the cytoplasm, the nuclear factors (proteins) will be replaced with cytoplasmic factors. The nuclear factors will be sent back into the nucleus while the mRNA, now has a cytoplasmic coat, is ready to be translated.

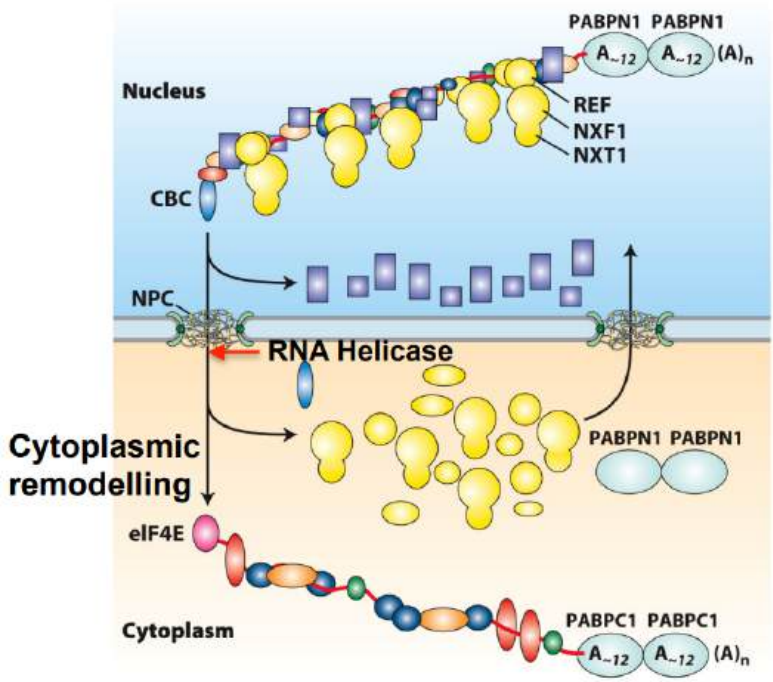


Figure 5.88: Cytoplasmic Remodelling.

Remark 5.45. *mRNAs never exists in its transcript only form but they're usually coated or associated with proteins and factors forming mRNPs.*

Chapter 6

Eukaryotic Translation and Regulation

In this final chapter, we will look at the translation of a transcript to a proteins and how it is regulated in the cell.

6.1 Principle of Translation

All 3 classes of RNA: mRNA, tRNA and rRNA are essential for protein synthesis. The main mediator of translation is a structure called ribosome. A **ribosome** is a combination of proteins and a number of rRNAs together.

6.1.1 Functional Ribosome Translation Machine

You'll recall that rRNAs are made from rDNA clusters whose transcription is highly efficient. Furthermore, the rRNA alone makes up 80% of the total RNA in the cell. This makes sense since rRNAs are critical for formation of ribosome which is the essential cellular machinery for protein synthesis. We also recall that transcription of rRNA are highly conserved between eukaryotes and bacteria. This is also shown through its structure in forming the ribosomes:

In bacteria, a 23S rRNA, a 5S rRNA and 31 different proteins will come together to form the **large ribosomal subunit** (50S). In addition, a 16S rRNA and 21 other proteins will give rise to the **small ribosomal subunit** (30S).

Together, the 50S large subunit will come together with the 30S small subunit to form a **functional ribosome** that can carry out translation.

Remark 6.1. *The 30S, 23S and etc. are units of **Svedbergs** which is the rate of sedimentation. In general, the larger the S is, the bigger the molecule is*

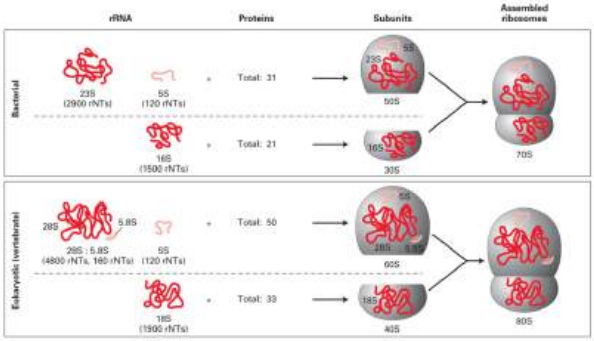


Figure 6.1: Ribosome unit of bacteria vs eukaryotes (vertebrates).

Because these structures are conserved through organism, we can see a similar build up of the functional ribosome in vertebrates where instead of 70S, we're 80S and other minor changes as well (see Figure 6.1)
Essentially rRNA has sequences that are conserved in all organisms but there are minor changes in the rest. Due to sequences similarity, the secondary structure, when the rRNA fold into stem loops, helix etc., is also conserved.

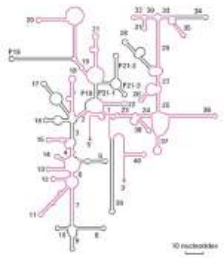


Figure 6.2: Structural conservation between 16S bacterial rRNA and 18S eukaryotic rRNA highlighted in the pink region.

The structural conservation through evolution also tells us that rRNA serves an important roles and once again, that role is proteins synthesis.

6.1.2 tRNA and Amino Acid

The tRNA is also critical for protein synthesis as it serves as the mediator between the codon on the mRNA and the corresponding amino acid. Orig-

inally, pre-tRNAs are transcribed by RNA pol III. They the goes on and mature through many modification (see 5.8.1 Processing of rRNAs and tRNAs). Although many cartoon representation of a tRNA shows it looks like a cross, the actual structure of tRNA is like an upside down "L". Hence the anticodon loop would be at the top of the L while the end hold the amino acid.

Remark 6.2. *tRNAs do not act independently but they need to be coupled with amino acids to be functional.*

This attachment process of amino acid to tRNAs is faciliated by an enzyme called **aminoacyl-tRNA synthetase (aaRS)**. aaRS has 2 substrates binding site, 1 for the amino acid and 1 for the appropriate tRNA. It will then use ATP to form a strong covalent bond between the amino acid and its appropriate tRNA.

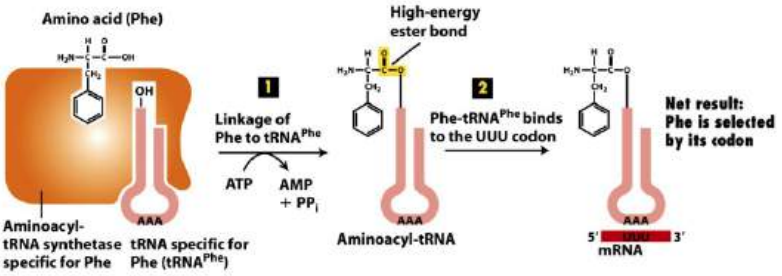


Figure 6.3: aaRS covalently bond amino acid to its appropriate tRNA.

Remark 6.3. *1 aaRS can bind to* > 1 *unique tRNA and 1 tRNA can bind to* > 1 *1 codon on the mRNA.*

Because of this, we typically say that the genetic code is generate. **Why would it be degenerate?** Well...we need to understand the nature of base pairing of the tRNA's anticodon and mRNA's codon. Due to the environment created between them, it permits a *non-Watson-Crick* base pair between the codon and the anticodon (e.g. TTA can bind with AAC). As a result, 1 amino acids is encoded by multiple tRNA (anticodon).

Example 6.1.1. Leucine is encoded by 6 different anticodons, serine is encoded by 4 anticodons, etc.

Nevertheless, there are 2 amino acids that only have 1 codon: methionine (met) and tryptophan (trp). Trp is not so important for us however met is, because it is the main amino acid that will start translation.

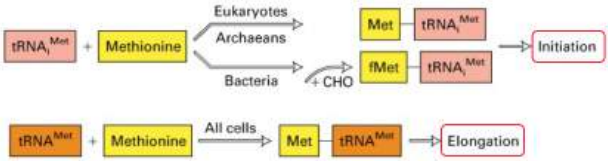


Figure 6.4: 2 types and tRNA^{met} and 2 subtypes of tRNA$_i^{met}$.

There are 2 tRNAs that will encoded for met (universal for all organism): tRNA$_i^{met}$ (for initiation) and tRNA^{met} (for elongation). The main differences between them is that tRNA^{met} is used exclusively in elongation and cannot start translation and the same is true for tRNA$_i^{met}$ (exclusively initiation). tRNA$_i^{met}$-amino acid complex is slightly different between bacteria and eukaryotes since the methionine will be added with a *formyl* group in bacteria.

Now that we got all of the machinery behind protein synthesis we can begin looking at translation. The process of translation can be divided into 3 main steps: initiation, elongation and termination.

6.1.3 Initiation and Pre-Initiation Complex

Initiation is the most highly controlled process of all of translation. In order to initiate translation, the small ribosomal subunit must be dissociate from the large one. This dissociation will disable translation from happening until the structure is in the right condition.

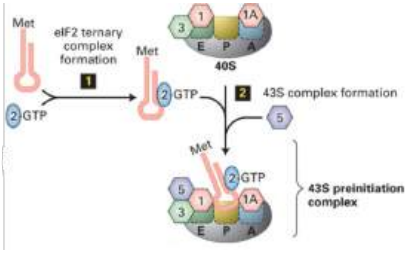


Figure 6.5: Formation of 43S PIC.

The dissociation for small ribosomal subunit is made by associating it with **eukaryotic initiation factors (eIF)** 1, 1A and 3. Then eIF2, forming a ternary complex with the tRNA$_i^{met}$ and GTP, will bring tRNA$_i^{met}$ to the

P-site on the small ribosomal subunit. This combination along with eIF5 will form the **43S pre-initiation complex (PIC)** that will be cruicial to start translation.

Remark 6.4. *When cells are in sub-optimal condition for protein synthesis, it will phosphorylate eIF2. This phosphorylation will take away the GTP and thus no formation of 43S PIC hence no translation.*

What about the mRNA? Does it has proteins bind to it too?

Well researchers through out the years have hypothesized that there are proteins that bind to the 5' cap (7' mGDP structure [cap]) that can greatly enhance translation. Through affinity chromatography, they were able to show that there are such proteins that when bound to the mRNA's cap will drive up transcription while unbound will have almost no translation.

Remark 6.5. *Only class II mRNA transcript would have the 5' cap which is why they're the most highly translated.*

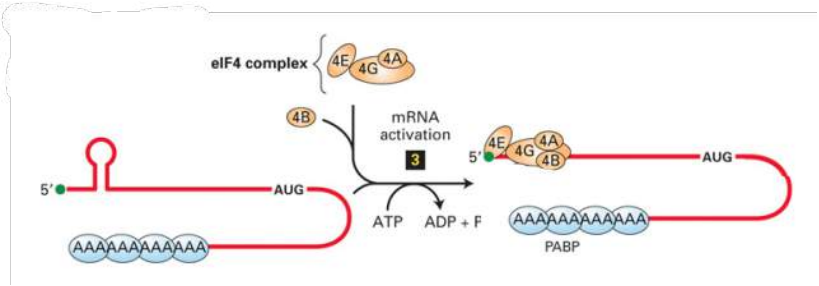


Figure 6.6: Binding of eIF4 complex to mRNA that lead to its translation

These cap-binding proteins were purified and isolated to be eIF4E, G, A and B which are subunits that make up the **eIF4** complex. The main proteins that binds directly to the 5' cap is eIF4E which can ensure the correct positioning of the rest of the subunits, then allow them to carry their normal function.

Remark 6.6. *eIF4E level is highly limited as an overexpression of it could lead to cell overgrowth and tumour development.*

eIF4G is the main protein that brings the 43S PIC to the mRNA through a protein-protein interaction with eIF3.

Researchers also found that eIF4G interacts with the poly A tail and its associated proteins PABPC to form a loop. This loops increases translation efficiency since once a ribosome dissociate, it can form a PIC again to initiate another translation process.

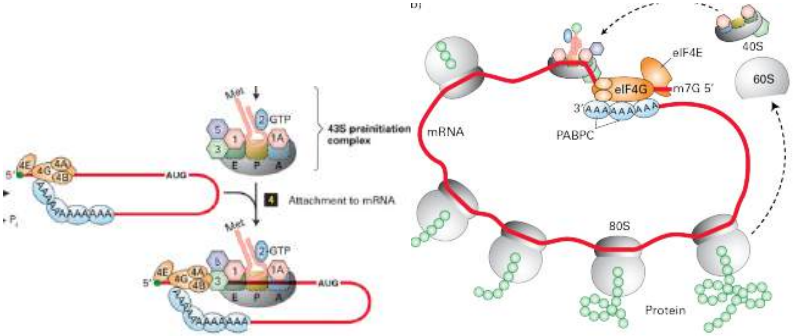


Figure 6.7: Association of mRNA and 43S PIC through eIF4G and eIF3 (left). eIF4G association with poly A tail forming a loop.

There's an important enzymatic activity of this new PIC is scanning. eiF4a, a helicase, will use ATP and associate with the 5' end; it will then scan (5' to 3') through the entire mRNA, with the enhancement of eIF4B, until it hits an AUG start codon. At the same time, it will remove any unnecessary mRNA transcript before the AUG codon.

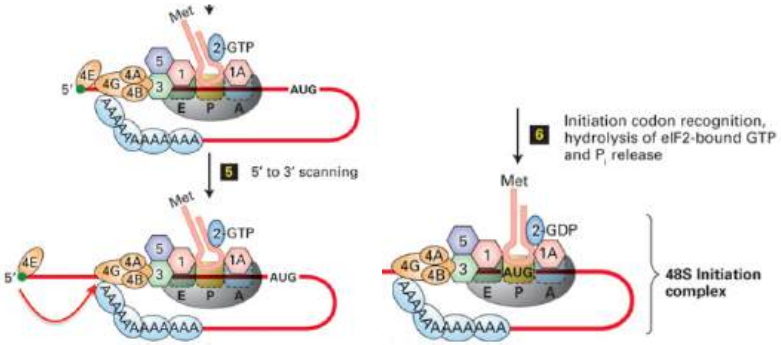


Figure 6.8: Scanning complex and hydrolysis eIF2-GTP to GDP.

Remark 6.7. *1 debate is that when this entire complex moves, does eIF4E*

goes with it or stay at the 5' cap. Well...majority of researchers agree that eIF4E will go along with that scanning complex.

Once AUG is associated with the $tRNA_i^{met}$, it will go through a conformational change caused by the hydrolysis of eIF2-GTP to eIF2-GDP and this change will be accelerate by eIF5, a GAP. Then, the entire PIC will dissociate from the small ribosomal subunit and the large ribosome subunit is recruited and associate in, forming a ribosome translational machine called **80S initiation complex**.

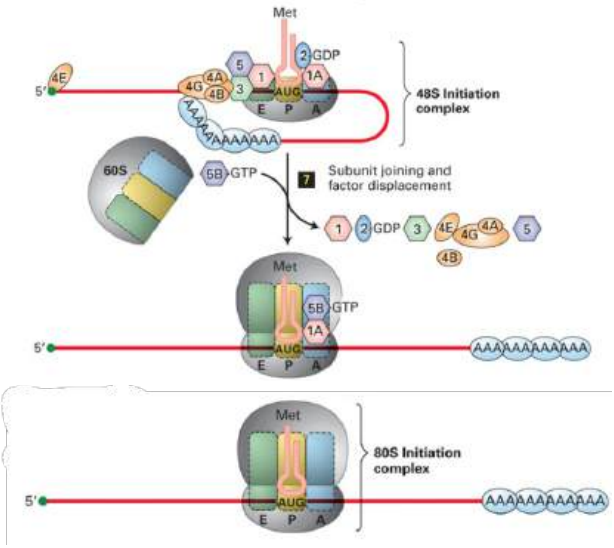


Figure 6.9: Dissocation of the PIC and formation of the 80S initiation complex

Looking at the 80S initation complex, we can spot different sites on it. There are the **aminoacyle (A) site** that can accept new anticodon in, the **peptide (P) site** that hold the peptide chain, and the **exit (E) site** that discard the used anticodon.

Remark 6.8. *$tRNA_i^{met}$ is present at the ribosome which leave the A and E site free. Its position is important since it needs to bind to the AUG but also leave out a position for coming tRNA.*

6.1.4 Elongation

$tRNA_i^{met}$ is associated with the P-site on the ribosome while the A site is free for tRNA to come a bind. The diffusion of tRNA into the A-site is carried out by **elongation factor 1 α (EF1α)** coupled with GTP. This EF1α will bring tRNA and diffuse it to the A-site and if the codon doesn't match, the tRNA simply dissociate out. This will continue till the correct tRNA goes into the A site and matches with the codon.

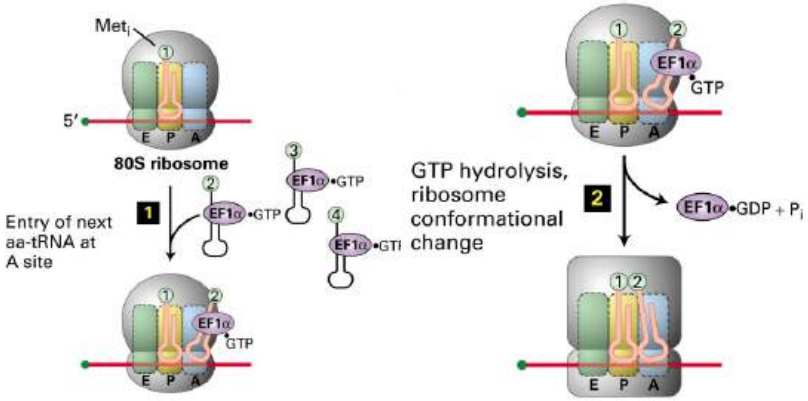


Figure 6.10: tRNA moves into A-site, match and lead to its conformational change.

Once there is a match, EF1α-GTP will go through GTP hydrolysis which lead to a conformational change to the tRNA (in the A-site) in a way the 2 tRNAs in the A and P-site are closer together. EF1α also leaves.

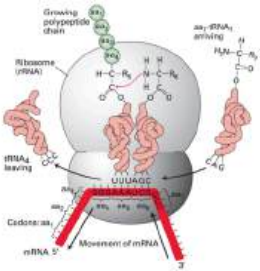


Figure 6.11: peptidyltransferase activity of ribosomes.

The positioning of 2 amino acids in proximity with each other would introduce enzymatic activity. In this case, **Peptidyltransferase** activity will initiate which will catalyze a peptide bond between the P-site amino acid's carboxyl group and the A-site amino acid's amino group.

Remark 6.9. *This peptidyltransferase activity is not dependent on other proteins.*

When investigators look at where the bond catalysis takes place, very little to no proteins was found however lots

of rRNA was available. This finding lead them suspect that this bond catalysis is made directly by the ribosome.
By taking a 23S ribosome and strip off all of its proteins, researchers found that it can still catalyze peptide bond with no problems. These ribosome are called **ribozyme** i.e. catalytic ribosome or ribosome that has enzymatic activity.

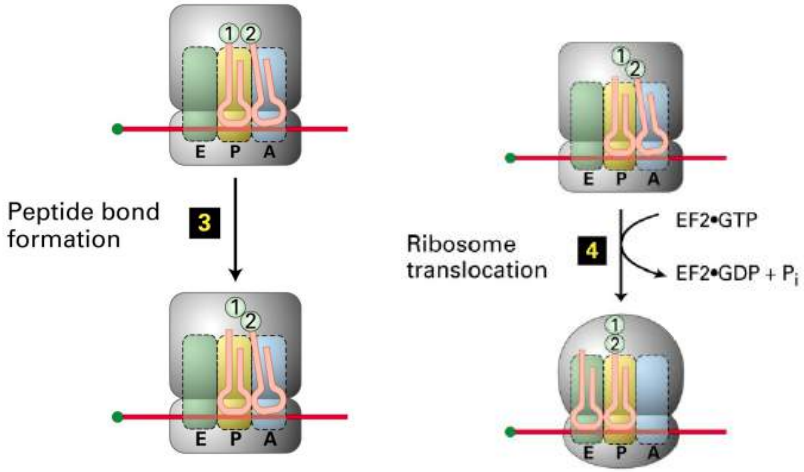


Figure 6.12: Catalysis of a new peptide bond and translocation of ribosome

Once the peptide bond is formed between amino acids, the ribosome will translocate a distance of 1 codon (3 nt), which is driven by the hydrolysis of EF2 i.e. if EF2 isn't hydrolyzed, translocation cannot happen nor ribosome can slip backward. In translocation, the tRNA that bound to the amino acids chain, found in the A-site, will now be at the P-site; while the tRNA found in the P-site will be at the E-site because it's no longer attached to the amino acid.
The tRNA at the E-site will be forced out due to the ribosome's conformational change via the hydrolysis of EF1α when a new tRNA matches in the A-site (and will be recycled). The cycle of elongation will keep repeating previously mentioned steps.

6.1.5 Termination

This elongation cycle keeps going until it hits a stop codon. There's no tRNA that can match with the codon however there is a proteins called **eukary-**

otic release factor 1 (eRF1) associated with **eRF3**, a GTP-bound factor. The eRF1, that has similar shape to the tRNA, will go and recognize the stop codon which also release the tRNA in the E-site and the peptide chain.

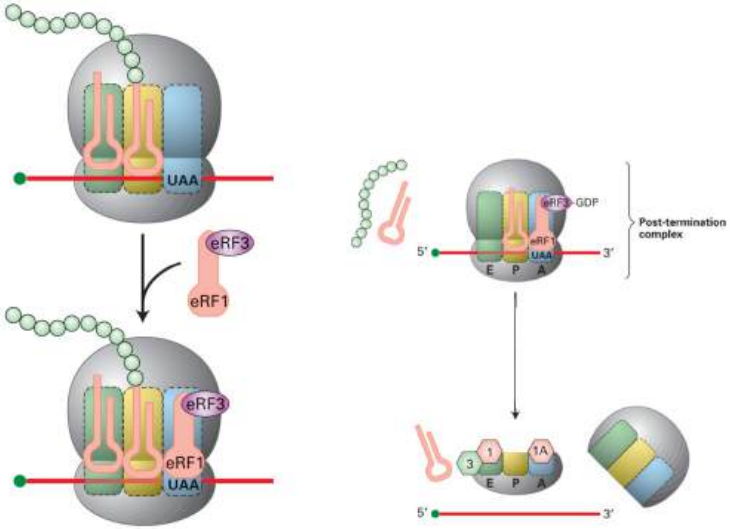


Figure 6.13: Termination of translation.

In this critical moment, eRF3 goes through GTP hydrolysis which lead to the dissocation of the large ribosomal subunit from the small ribosomal subunit. The tRNAs will be recycled while the small rib sub will be associated with the other eIFs so that the large will not associate. These 2 subunits will ready to initiate another cycle because of the mRNA loop conformation.

Remark 6.10. *A highly efficient transcription is correlate the amount of ribosome translating it.*

Using ultra centrifugation, we can identify these complexes along different stages of translation. You first lyse a batch of cells and take its cellular content and put it through a ultra gradient centrifugation. We can see different amount of protein bound and produce as we read along the sedimentation. Near the top, are the small 40S then large 60S, then 80S functional ribosome, then all the highly translated message will be at the

bottom. The very last conformation of ribosome and mRNA nearing the bottom is called **polyribosome**.

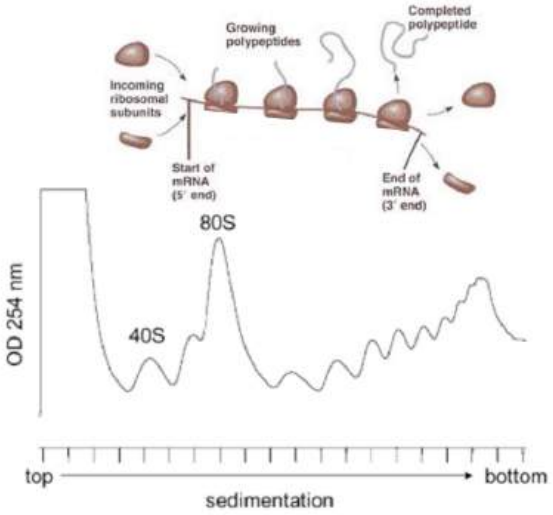


Figure 6.14: Sedimentation rate of different ribosome.

6.2 Post-Transcriptional/Translational Regulation

Remember back when a matured mRNA is sent out to the cytoplasm and there are these helicases on the nuclear pore that can "knock off" nuclear protein and replace for a cytoplasmic protein. It is important to know that this process sometimes has failure i.e. the helicase may not knock all of the nuclear protein off.

Once this mRNA gets to the cytoplasm, it will begin translation with the ribosome. The first ribosome to translate the mRNA plays an important role to not only translate but remove all of the proteins that is associated with the mRNA. This step is called **pioneer round of replication**. Typically after this pioneer round, the mRNA is now free from its proteins.
Nevertheless, if there was a *nonsense mutation* which lead an in-frame stop in the middle of the mRNA, after the pioneer round, 1 side would be free from proteins while the other isn't and typically proteins found on this side are nuclear. Its important to see that proteins arise from this mutation, that

is half the size of the typical one, can often lead to a *negative dominant effect*.

Definition 6.1. A **negative dominant effect** is a gain of function that is poisonous to the cell.

Example 6.2.1. Estrogen receptors are proteins that has a DNA binding domain that can activate the transcription of estrogen mRNA transcript. If there is an in-frame stop during translation of estrogen receptor, the product would have the DNA binding domain but cannot activate the DNA. Essentially, it blocks the normal function of transcribing mRNA.

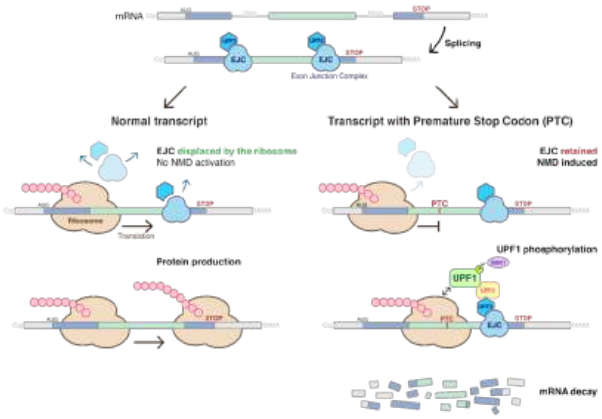


Figure 6.15: Activation of NMD as compared to normal mRNA transcript.

Cells have evolved mechanism that prevent the production of these deadly protein variant. After the pioneer round, if the mRNA transcript still has proteins associated with it, this will alarm the cell and it activates the **nonsense mediated decay** which is a mechanism that will find these in-frame stop mRNA and degrade them.

6.2.1 mRNA Destabilization

If we look back at bacteria, we know that they can switch their transcription very quickly to adapt to a new environment. This also means that the stability of bacteria's mRNA transcript has to be low since mRNA transcript that is good for 1 environment may not be good for another.
However...if we were to compare this with human, we can see that our

Cell	Cell Generation Time	mRNA Half-Lives*	
		Average	Range Known for Individual Cases
Escherichia coli	20–60 min	3–5 min	2–10 min
Saccharomyces cerevisiae (yeast)	3 h	22 min	4–40 min
Cultured human or rodent cells	16–24 h	10 h	30 min or less (histone and *c-myc* mRNAs) 0.3–24 h (specific mRNAs of cultured cells)

Figure 6.16: Table of mRNA stability (half-life) between bacteria, yeast and human.

mRNA are much more stable. This is because we can maintain the internal environment of the body (hence the environment of the cell) at a relatively constant rate via *homeostasis*.

As you would've noticed from the table that there's a column of cases where the stability of mRNA, even in human, isn't as high; **why is that the case?** Well...these are mRNA that are required to produce a specific function for a short period of time. Once the task is complete, it's useless to have the mRNA to translate even more protein.

Remark 6.11. *Not only that it would be useless but having these mRNA around could lead to tumour growth or leukemia.*

Example 6.2.2. **Granulocyte macrophage colony stimulating factors (GM-CSF)** are cytokines that can lead to the proliferation of white blood cells. If after an immune response and GMCSF is still activated, this could lead to the body develop leukemia.

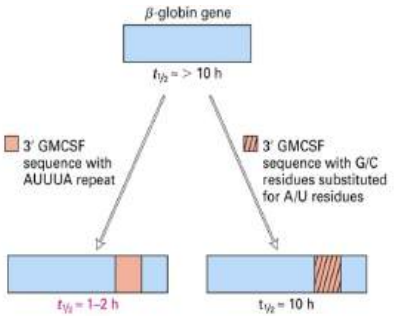


Figure 6.17: AUUUA sequence and mRNA destabilization.

To ensure that this translation doesn't occur, the mRNA must be destabilized. What researcher found is that along 3' UTR of the GMCSF's mRNA is an AUUUA sequence that is responsible for the degradation of that mRNA. Turns out, majority of short-lived mRNA also has this sequence.
To test this, researchers hybridize this sequence onto a much more stabilized mRNA (> 10h). They found that this hybrid mRNA is degraded within just 1-2h. They further drive home this point by mod-

ify the AUUUA sequence which shows that the mRNA is now stable for a long while.

6.2.2 mRNA Degradation

The degradation of mRNA are carried out very efficiently. When an mRNA has sequences (like AUUUA) on the 3' UTR, it will recruit in the first machinery called **deadenylase complex** that would degrade the poly A tail. Once the mRNA is deadenylated, the degradation or *decay* can happen in 2 different directions 5' to 3' and 3' to 5'.

Decay from 3' to 5' is mediated by **exosomes** immediately after deadenylation. Exosome is quite special in its structure as it has a site dedicated for exonucleation, and if this site fail, it has another endonuclease site to ensure a proper decay.

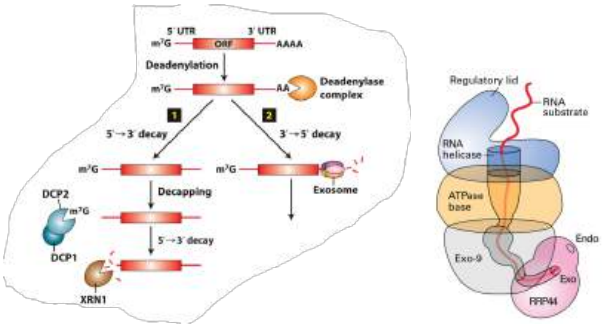


Figure 6.18: Deadenylation dependent decay of mRNA and exosome structure.

On the other hand, decay from 5' to 3' is carried out a little differently. First the 7'm cap on the 5' end is removed by **mRNA decapping enzyme** then it will be degraded by **XRN1 (exonuclease)**.

Remark 6.12. *The degradation activity was found to happen in a section of the cytoplasm called* ***P-bodies*** *(condensate-like).*

Furthermore, the decay above is *deadenylation-dependent* i.e. it requires the removal of the poly A tail to be degraded. Another type of decay that is *deadenylation-independent* can happen without adenylation. In this decay, endonuclease will cleave the middle of the mRNA allow XRN1 to decay 1 half from 5' to 3' while letting exosome decay the other half from 3' to 5'.

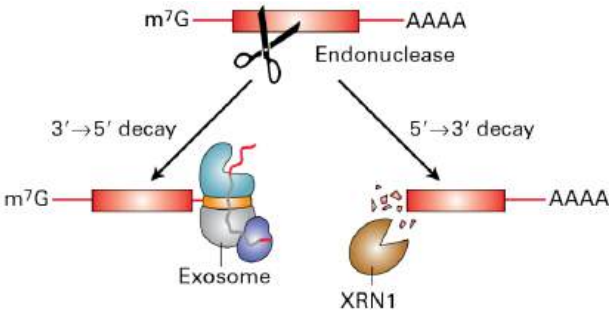


Figure 6.19: Deadenylation independent decay of mRNA.

6.2.3 mRNA regulation

Iron (Fe) is an important component for some enzymatic and physiological activity in the body. Nevertheless, Fe is very toxic for the body which means that its [Fe] is high regulated. This regulation is carried out by **iron responsive element-binding protein (IRE-BP)**. To understand the mechanism of IRE-BP we must understand how Fe is transported in the body.

Iron responsive element is a short stem-loop found in most 3' UTR of mRNA whose products are involved in the metabolism of Fe such as *ferritin* (store Fe) or *transferrin receptor* (transport ferritin). Essentially, in normal condition, these mRNA would give rise to, let's say, transferrin receptor. This receptor would bind and transport intracellularly.

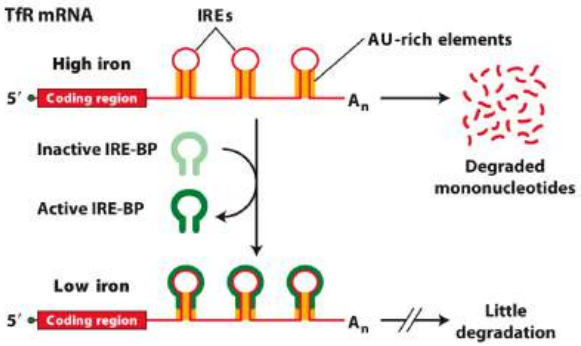


Figure 6.20: Iron regulation by transferrin IRE and IRE-BP.

Coming back to IRE-BP and now we know, as the name suggested, IRE-BP would bind to the IRE stem-loop under certain condition. When there is a high [Fe], the IRE-BP is inactive and doesn't bind to the IREs of transferrin receptor, which has AU rich region and lead to its degradation. This degradation decrease the amount of transferrin receptor made to uptake of Fe.
When there is a low [Fe], IRE-BP takes an active conformation and bind to the IREs. This binding lower its rate of degradation hence allowing production of more transferring to uptake the Fe.

6.2.4 mRNA Translation Regulation

It turns out that there are regulatory mechanism that does not rely on regulating the mRNA but regulating its translation. This discovery was done by looking at the northen and western blot of the polypeptide synthesis under strict control. On the northern blot, just as expected, amount of mRNA through time does not change. We would expect that on the western blot would be the same however, through time, the amount of polypeptide change i.e. mRNA level does not correlate with protein level. This also meant that there are some regulatory mechanism on translation.

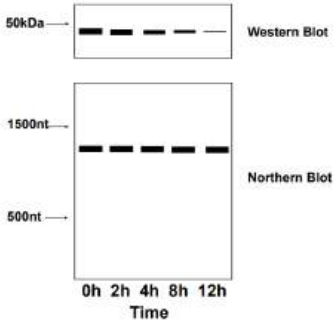


Figure 6.21: Northern and western blot discrepancy.

Example 6.2.3. We can see this translation regulatory mechanism presents in *Drosophila* embryo development. To begin with, **Hunchback (HB)** is a special gene that is crucial for the trunk development of *Drosophila*. During normal development, the maternal HB mRNA is distributed all through out the egg. Another gene called **Nanos** has its mRNA localized mainly in the posterior region. Nanos mRNA functions as translational inhibitor of HB mRNA in the posterior area. This inhibition lead to a steep concentration gradient between HB protein and Nanos protein which lead to the establishment of anterior-posterior body pattern.

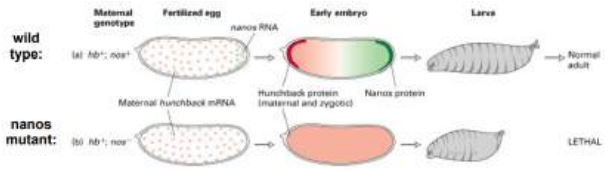


Figure 6.22: Wild-type and mutant *Drosophila* embryo development.

If we were to remove the nanos proteins and observe its development, we would find that HB mRNA can now translate everywhere in the embryo. However, the larva that it develops into has no functions essential it's a *lethal* larva that cannot develop further to a matured *Drosophila*.

Example 6.2.4. Going back to the IRE-BP, we've only looked at its regulation on the transferrin receptor, now we will look at ferritin. When there's a high [Fe] in the cell, IRE-BP will be inactive and cannot bind to the IRE on the 5' UTR of ferritin mRNA. This allows the mRNA to be translated into ferritin and store Fe.

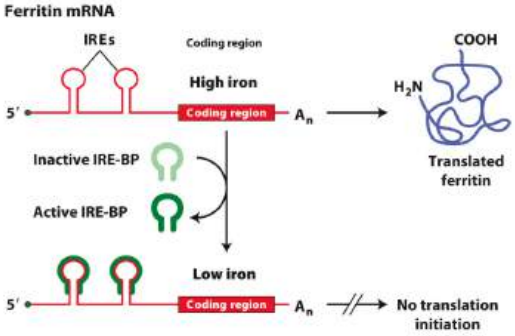


Figure 6.23: Iron regulation by ferritin IRE and IRE-BP.

On the other hand, when there's a low [Fe] in the cell, IRE-BP will be activated and bind to the IRE stem-loop. This binding will disable the ferritin mRNA to be translated and thus will not store additional Fe.

Example 6.2.5. *C. Elegans* like other worms, would go through 4 different larval stage post-embryonic development: $L1 \longrightarrow L4$.
Researchers found the mutation on a gene called **Lin-4** would lead to the worm having repetitive larval stage i.e. after it goes through L1 it will go

through L1 and again and etc. At the same time, researcher found a gene called **Lin-14** that give rise to a nuclear proteins whose regulation is important for the larva to reach L2. They found that once the larva reaches L2, lin-14 is downregulated; they also found that mutation on its 3' UTR would lead to a similar repetitive phenotype.

Does it mean that lin-14 is linked to lin-4? Well...apparently yes! Through testing, researcher was able to sequence the RNA product of lin-4 and they found that it has no protein coding sequence (no ORF)...This is not bad news because apparently, this small lin-4 RNA is a non-coding RNA that is partially complement to regions of the lin-14 mRNA's 3' UTR. The binding of lin-4 RNA to the 3'UTR *downregulates* the expression of lin-14 which allow the larva to reach L2.

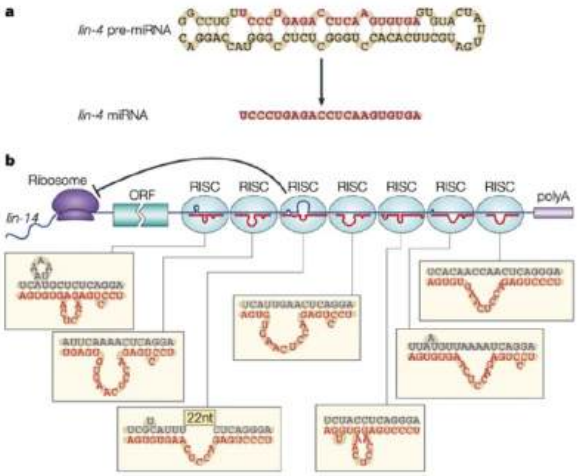


Figure 6.24: Lin-4 and its role in downregulating lin-14 through partial complementary.

To sum it up, in order to reach L2, lin-14 mRNA requires to have proper 3'UTR and functional lin-4 to downregulate its translation.

This was an astounding result but not exactly because there's a chance that this is only present in *C. Elegans*...well, turns out, around 10 years after this discovery, researchers found a similar conserved mechanism/structure in human. This discovery opens up a new era to biology that is *microRNA*.

6.3 MicroRNAs and Regulation

Definition 6.2. **MicroRNAs (miRNAs)** are small non-coding RNA of 21-23nt and are involved in the post-transcriptional gene expression regulation.

They can bind and block translation of an mRNA through their partial antisense complementary and limited homology. Not only that they can also deadenylate RNA through complex formation.

miRNA genes are transcribed by RNA pol II which gives rise to a **primary miRNA (pri-miRNA) transcript**. The pri-miRNA takes on a "hairpin" conformation with self-complementary. This pri-miRNA is then capped and slightly processed by nuclear protein called **Drosha and DGCR8**. This modification will turn pri-miRNA to pre-miRNA which is transport to the cytoplasm via *exportin 5* in the nuclear pore complex.

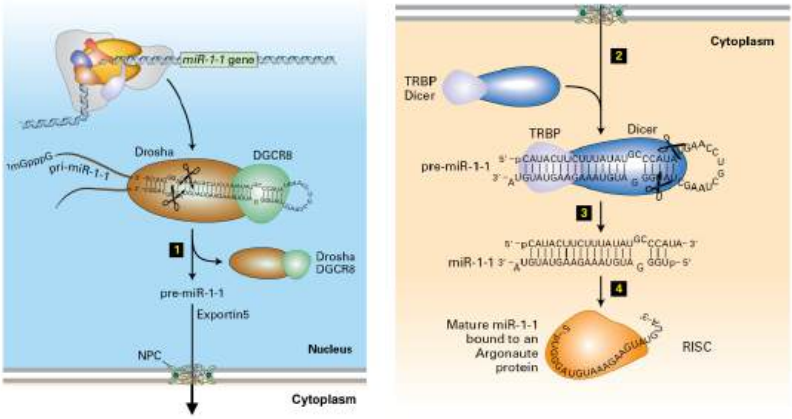


Figure 6.25: miRNA creation and formation of the miRISC.

Remark 6.13. *miRNA can control a population of gene i.e. 1 miRNA can control/complementary to 100 different genes.*

In the cytoplasm, the pre-miRNA is further modified by an enzyme called **Dicer**, RNAse III-like. Then this pre-miRNA goes to an **RNA-induced silencing complex (RISC)** where a helicase protein called **argonaute** would unwind the dsRNA into ssmiRNA using ATP. The RISC along with a miRNA

will form the **miRNA-induced silencing complex (miRSIC)**. This miRISC will find the appropriate target for miRNA to Watson-Crick base pair with the 3'UTR and shut down that mRNA.

So far, we're able to characterized different over thousand of different miRNA, some of which are conserved in human. Not only that, these miRNAs has a very close ties with many metabolic processes in our body as well as tissue growth, neural development, stem cell pluripotency and even cancer.

Remark 6.14. *It is estimated that around* 60% *of our known coding gene in the genome as under miRNA-mediated control.*

6.4 RNA and Gene Silencing

researcher was trying to figure out a phenomenon where injection of dsRNA (complementary to a gene) can cause the animal to recapitulate a phenotype that results in a loss of function of that gene. **So why is this the case?** Well...it was found that this recapitulation mechanism is a conserved way of dealing with invasion of maybe virus or transposable elements.

Works by these individuals create a new era of understanding RNA controlling via *RNAi pathway* to create a phenotype with a loss of function

Definition 6.3. An **RNA interference (RNAi) pathway** is a biological process where sequence-specific gene suppression is carried out dsRNA.

Example 6.4.1. Wild-type flies would typically have red coloration for their eyes due an specific gene expression. We could introduce a dsRNA that correspond to the coloration gene which trigger the RNAi pathway. The RNAi pathway would block the expression of the red coloration gene which lead to the flies to have a white eyes mutation.

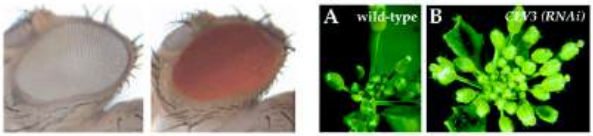


Figure 6.26: RNAi experiment with *Drosophila*'s eyes color and flowers' CLV3 gene.

This does not only present in flies but also in other species as well such as introduction of this dsRNA to interfere with CLV3 gene would result in plants having multiple flowers and shoots.

6.4.1 RNAi Pathway

Our prior knowledge to miRNAs and its pathway really helps us in developing and understanding of the pathway of RNAi or RNA mediated interference.

A precursor dsRNA, when introduced into the cell, will be recognized by the RNAse-III like enzyme called Dicer. It will cut the dsRNA into smaller segment of around 21-23nt called **small interfering RNA (siRNA)**. This siRNA will be handed off to the RISC that has the agronaute protein to separate the 2 strands from each other using ATP.

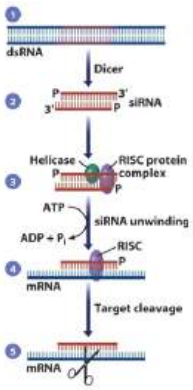


Figure 6.27: RNAi pathway

Remark 6.15. *This siRNA helicase agronaute proteins is not the same agronaute as the miRNA.*

After it will then find a target mRNA that's complement to the siRNA and this complementary can be anywhere, not just 3'UTR. This special agronaute for siRNA has another special function is that it have endonuclease activity i.e. once Watson-Crick base pair between the siRNA and the mRNA has established, the agronaute protein will cleave the mRNA via endonuclease activity which then allow its rapid decay via XRN1 and exosomes.

miRNA vs siRNA

There are lots of similarities between them however there are 2 main distinct differences: complementary and regulation.
miRNA can base pair with the mRNA mostly on that 3'UTR only and it's only partially complementary. On the other hand, siRNA is fully complementary to the mRNA it binds not only at the 3'UTR.
miRNA regulate the expression of a gene by supressing its translation while siRNA regulate gene expression by directly destroying the transcript.

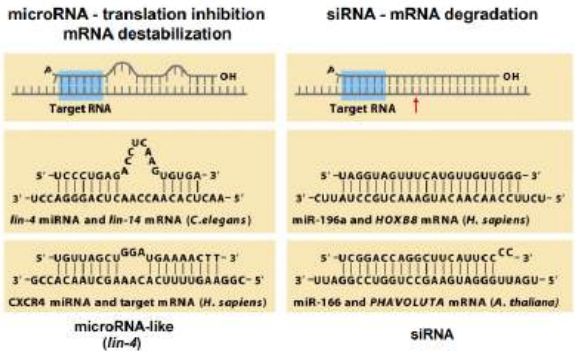


Figure 6.28: Comparison between miRNA and siRNA.

The mechanism of siRNA is in fact preferable for viral invasion since we need to effectively destroy the transcript instead of leaving it floating aorund.

The mechanism of miRNA is favourable when it comes to adaptation since its best to not destroy the transcript right away since it could be useful for another environment.

6.4.2 Other Non-Coding RNA

It was found the dsRNA and siRNAs could also correspond to specific regions on a genome and shut it down.

Example 6.4.2. *Schizosacromyes Pombe* is a yeast variant where its centromere has to be maintained at a transcriptionally inactive state. The reason for this is that it's the region where microtubules will attach and separate the 2 chromtide; so having RNA pol going through them is not ideal. Nevetheless, in identifying products that at this centromeric region, researcher was able to isolate lots of dsRNA and siRNAs products that are complementary to centromeric regions.
This was astounding as it was thought that these dsRNAs and siRNAs works only in the cytoplasm.

From this example, we can see that not only does siRNAs can destroy mRNA thus reduce gene expression, they can also induces conformational changes on the chromatin.

Remark 6.16. *Later discoveries shows that this is not unique to only S. Pompe but to many vertebrates too.*

It does not stop there, we were able to further discover a different class of small RNAs called **piRNA** that interact with and agronaute protein called **PIWI**. Its main function is to maintain the integrity of germ cells by the elimintation of abnormally expression RNA and chromatin in it.

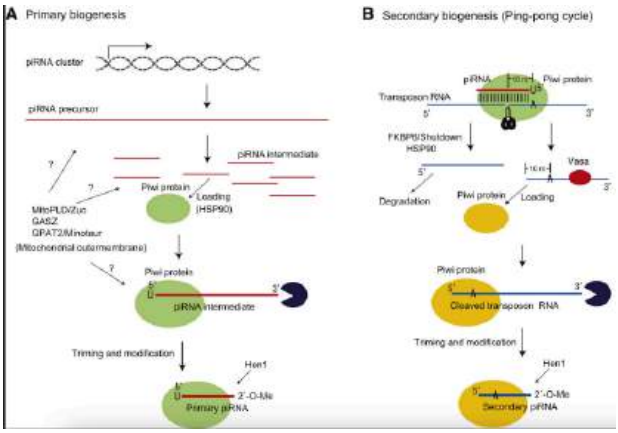


Figure 6.29: piRNA and PIWI interaction

piRNAs are generated from a large cluster of disable old transposable element. RNA pol II would go through and transcribe a long precursor piRNA. This precursor will bind to PIWI which would modify it. Together PIWI and piRNA form a ribonuclear protein complexes that can track down mutated and overly expressed transposon RNA. Once its target is identify, the piRNA will base pair with it while the PIWI will activate its endonuclease activity.

6.4.3 Long Non-Coding RNA and Gene Expression

Viruses are constantly evolving to effectively uses its host replication machinery. Due to the development of siRNA and miRNA, it was hard for them do such infection. Nevertheless, viruses has build a mechanism by which they would recognize miRNAs and generate complementary long sequence of RNA. This RNA sequence would base pair to the miRNA and compromise its normal function. Another interesting RNA product they use are these **circular RNA (cirRNA)** that can "sponge" up all the miRNA from ever performing their usual function.

All in all, these sequences of RNA that viruses developed to combat the miRNAs are called **long non-coding RNA (lncRNA)**. Not only that lncRNA

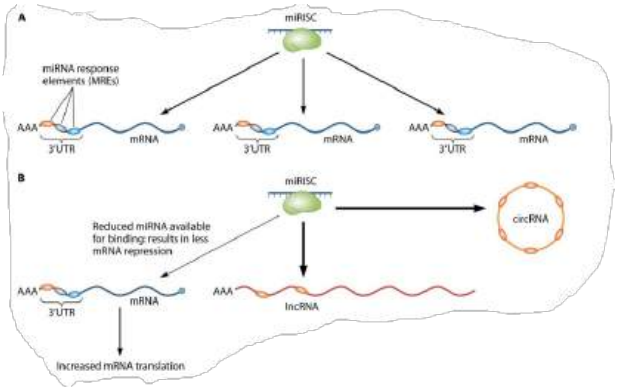


Figure 6.30: Virus development of lncRNA and cirRNA to combat miRNA.

are important for us to develop new strategy to fight viruses but also it contributes to a beautiful physiological development which we will look at.

Dosage Compensation

We know from before that dosage compensation in *Drosophila* is mainly driven by the Sxl gene however we did also said that this mechanism isn't the same in mammal. So **How does mammalian do dosage compensation?** Well...instead of having the male upregulate its X chromosome, evolutions have found that it's beneficial for the female mammals to inactivate 1 of its 2 X chromosome.

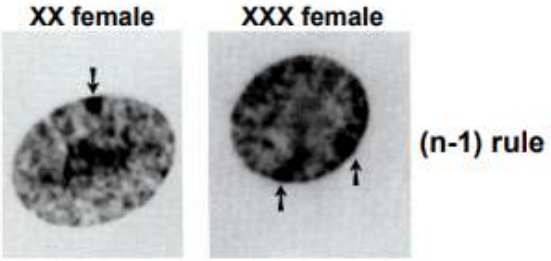


Figure 6.31: Barr bodies electron micrograph

Remark 6.17. *In peculiar cases where there's a mutation lead to the development of 3X chromosomes, the female would then inactivate 1 leaving 2X chromosomes and etc. following the (n-1) rule.*

If we were to look through and electron micrograph, we can see that these inactive X chromosome is represented by these electron dense region called **Barr bodies** (see Figure 6.31). Similar thing if you were to look at a chromosome spread, you would see that there's 1 X chromosome that is not H4 acetylated.

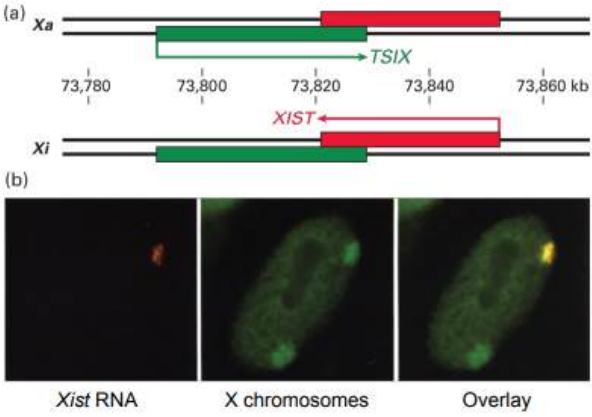


Figure 6.32: Xist and Tsix

All of these are carried out by a gene called **Xist**. Xist encodes for a lncRNA that can bind to the X chromosome and inactivate it. **But then...how can you justify shutting down 1 X chromosome rather than the other one?** Well...This has to do with the transcription of Xist in the antisense called **Tsix**. It turns out, Tsix is Xist antagonist i.e. it disables to the expression of Tsix.
The decision of which chromosome to inactivate is a stochastic process (random) during embryogenesis, where by a higher expression of Xist than Tsix would lead to that chomosome inactivation and v.v.

The Xist lncRNA, when bind to the chromosome, will recruit factors that lead to a tighter chromosomal conformation. It will interact directly with a protein called **SHARP** which brings in a host of other factors that can methylate the histone lysine 27 or deacetylate the histone tail (HDAC3). All in all, it will lead the chromosome to have heterochromatin structure

spread through out it thus create **transcription interference**. It's also observed that these Xist lncRNA and its recruited factors form a sort of liquid-liquid condensate which bring back the idea that RNA acts like a glue in forming such structure.

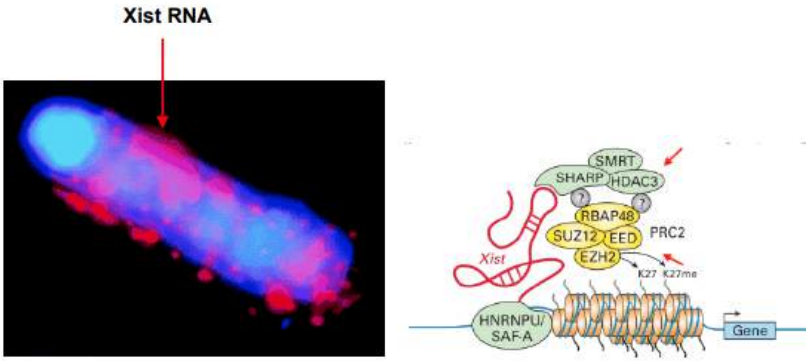


Figure 6.33: Xist and factors recruited lead to the formation of heterochormatin condensates around the inactive X chromosome

Remark 6.18. *This inactive X chromosome is an epigenetic trait. As the cell replicate, this inactive X chromosome will replicate and thus its daughter cell will also have the same X chromosome inactivation*

Part III

Brief Application of Molecular Biology

Chapter 7

An Approach to Systems Biology and Gene Targeting

In this final chapter, we will see how we can utilize these theory and apply it for research and even further discoveries.

Definition 7.1. **Systems biology** is a biomedical research approach that put pieces of biological information together to get an understanding of a "larger picture" (how stuff functions with each other).

Definition 7.2. **Gene targeting** is biotechnical tool/method used to target and alter a DNA sequence.

7.1 Systems Biology

When we first sequence out our genome, many thought that it was the "end" of science and discovery. They thought that having these sequences would essentially allow us to understand everything about life. Well...Not really, we cannot tell what those sequences really do. To really fully utilize the genomic sequence, we must look at how genes works with each other to give a final picture.

By comparing genomes of organisms, we can see that there's a **basic cellular toolkit**. They're used for cellular metabolism, transcription and translation regulation etc. Now, if we were to look at the most predominant common gene (almost half of the genome), we would find that it's full of

unknown sequences i.e. we don't even know what's the purpose of those genes.

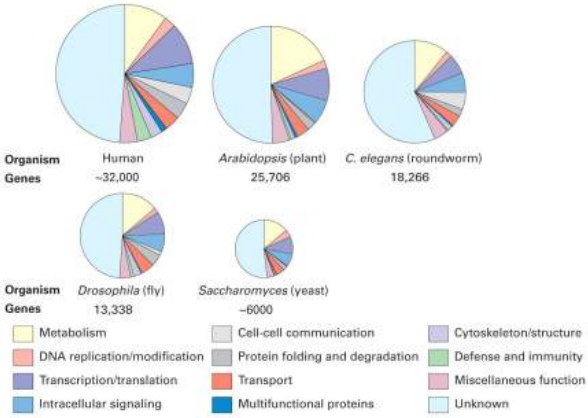


Figure 7.1: Different organism with common cellular toolkit.

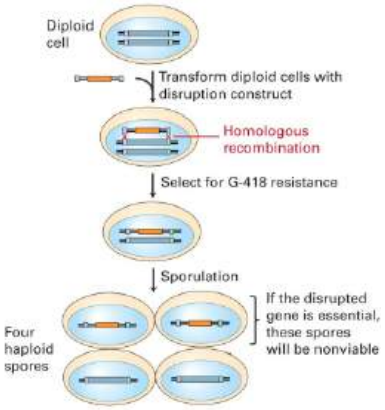


Figure 7.2: Yeast gene modification

The simplest way to look at the functions of these unknown gene is to carry out gene modification via homologous recombination. This experiment would be carried out with *S. Cerevisiae* (Brewer's yeast) where under good conditions, when its gene is disrupted and another gene with homologous structure is around, it will recombine it into its own genome.
Using this theoretical understanding, we can perform gene modification with *S. Cerevisiae*.

Procedure: we begin by taking *S. Cerevisiae* in its diploid form. A gene with 100% homologous flanking end of a yeast's gene is prepared and amplified through PCR. This new gene will allow the

yeast to have a specific drug resistance (G-418). After recombination takes place, the yeast will be incubated in the G-418 to eliminate all variant that didn't take in the gene. Then the rest of the modified yeast will be allowed to **sporulate** (diploid back to haploid).

Remark 7.1. *We used its diploid form because just in case the replacement is fatal, we won't kill the yeast right away.*

Essentially, we replace the yeast gene with another one that confer to only its survival in G-418. When the yeast sporulate, it will return back to its haploid progeny which mean that the yeast only have 1 set of chromosome either from the wild-type or the modified. If at this point the yeast didn't die, we would know that this replaced gene isn't essential for the yeast survival. To find out what the gene does, we would have to put the yeast into different trial test.

7.1.1 Functional Genomics

Yeast can tell us about 1 single cell and its genes but it doesn't tell us about the developmental growth of organism, reproduction nor adaptation. Therefore we need a better method which is *Functional genomics*.

Definition 7.3. **Functional genomics** or **Genome-wide functional analysis** is the study and gathering information about the expression of a gene in a genome.

You recall about RNAi mechanism that if you were to inject a *C. Elegans* with a dsRNA, it can activate the RNAi pathway and block the gene that is complementary to the dsRNA. So using this understanding, we could potentially create many dsRNA that is complementary to the unknown genes of *C. Elegans* (roughly 19,000) and test for it.

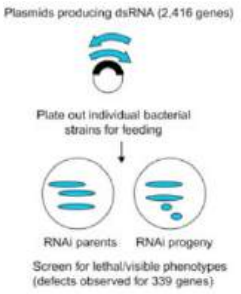


Figure 7.3: Generating dsRNA plasmid from bacteria

We first need to make the dsRNA which can be done using plasmid engineering (previously talked about). We would have 1 RNA strand with a **T7 promoter** driving the expression in the sense direction and the other stand with T7 promoter driving in the antisense direction. The T7 promoter will recruit **T7 RNA polymerase** and synthesize the dsRNA. All of these will be done in bacteria **but T7 RNA pol isn't normally expressed in them so how can transcription be done?** Well...it turns out a reagent called **IPTG** would induce the expression of T7 RNA pol.

The plasmid is now made, we can then infect the *C. Elegans* with each of these bacteria to see any changes.

Remark 7.2. *Each variant of these bacteria would carry 1 dsRNA that is complementary to 1 unknown gene. For this examination, there were 19,000 different variant to test all of the C. Elegans unknown gene.*

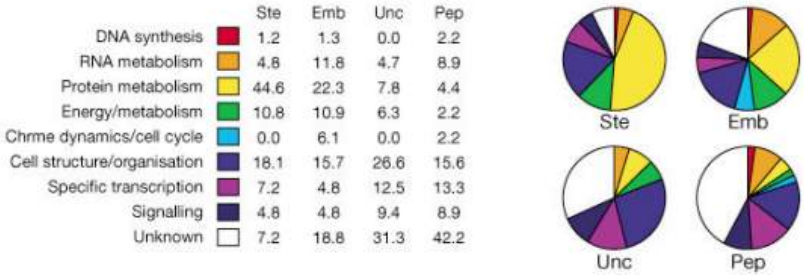


Figure 7.4: Gene categories from *C. Elegans* testing.

As a result of this transformation, you could put a lot of these genes into categories. Gene gives rise to sterility and lethal embryogenesis will obviously be categorized as part of the basic cellular toolkit.
Genes fall under category of uncoordination typically make proteins that are vital for neuromuscular controls.
There is also the category of post-embryonic phenotypes which are genes that help with the timing and localizing of organ development. However, this category is more *C. Elegans* specific and not other organism hence it's not part of the basic cellular toolkit.

7.1.2 2-Hybrid Screening

Sometimes, functional genomics isn't favourable to carry out systematic biology so...A method, proteomic approach, was developped based on the idea that transcription factors (TF) are modular. That is, they have a DNA binding (DB) and transcription activation (TA) domain.

Remark 7.3. *It does not matter the origin of these 2 domains, as long as they're put on the gene, downstream transcription will occur.*

Definition 7.4. A **2-hybrid screening** is a technique to detect protein-protein interaction.

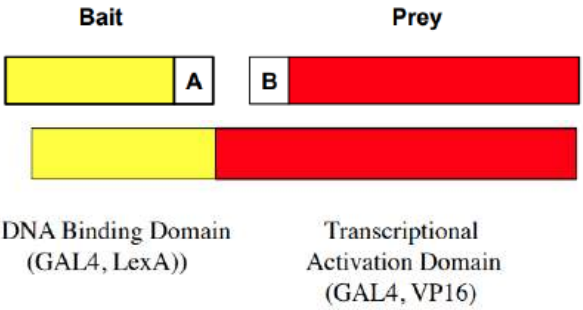


Figure 7.5: *A*-DB domain and *B*-TA domain fusion.

> **Mechanism of Action (2-Hybrid Screening):** A protein *A* is fused with a DB domain which can be GAL4, etc. If we know protein *A* interact with protein *B*, then we can fuse a TA domain onto *B* which can be GAL4, VP16 etc. So by bring this fused protein *A* and *B* together, they will interact and initiate transcription of the which ever gene reconized by the DB domain.

We can even use this alongside a reporter gene or a selectable marker e.g. UAS along with histidine synthesizing gene, UAS with LacZ. We typically refers the protein *A*-DB domain as the **bait** and protein *B*-TA domain as the **prey**.
The best is that you're testing for the interaction between these 2 so protein *B* can be any protein in the entire **proteome** (complete set of proteins expressed by an organism).

Essentially, you can make libraries that are all fused to transcriptional activation and then you can co-transform these things and evaluate which cells grow; then you can go back and figure out what the gene product was that was in the prey.

7.1.3 Protein Fragment Complementation

Sometimes, 2-hybrid screening does not work or not viable so another method can be used in replacement called **protein fragment complementation**.

In this method, instead of having prey and bait interact which lead to transcription, the prey and bait interaction will lead to a reconstitution of a functional protein

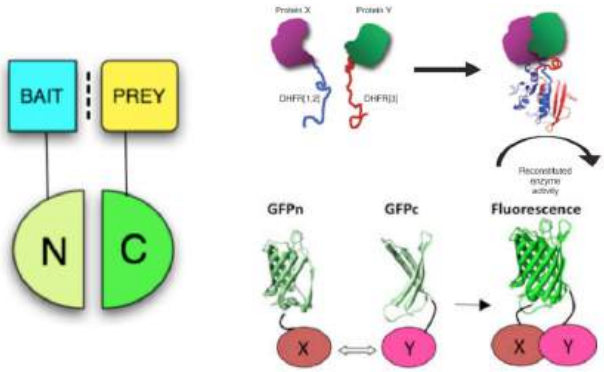


Figure 7.6: Different form of protein fragment complementation.

Example 7.1.1. A protein called dihydrofolate reductase is important for puridine synthesis that lead to cellular growth. So by having 2 protein that if they interact, they'll form dihydrofolate reductase which lead to cell growth otherwise nothing happens.
Same experiment can be done with 2 other protein that lead to reconstituation od GFP which if interaction does happened, cell will glow green.

7.1.4 BioID and Proximity Labelling

Once again, there is still a number of proteins where by the above techniques will not work, especially membrane protein. Another technique was

introduced whereby proteins in proximity to the protein of interest is labelled, and it's called **Proximity labelling**.

> **Mechanism of Action (BioID and Proximity Labelling):** A bait will be set up through the fusion of a target protein and a mutant biotin ligase in a plasmid. This plasmid is then grown with bacteria after which they're extracted and injected into cells. Once it's translated, proximal proteins will come and interact with the target protein bound to the biotin ligase. The biotin ligase will indiscriminatly catalyze all of the proximal protein thus label them with biotin.
> Then the cell is lysed and its constituent will be ran through an affinity chromatography where **avidin** can bind to the biotin conjugate thereby isolating the protein interacted with target protein. This isolated mixture is then diluted with biotin solution which allow the biotin-protein conjugate release from avidin.

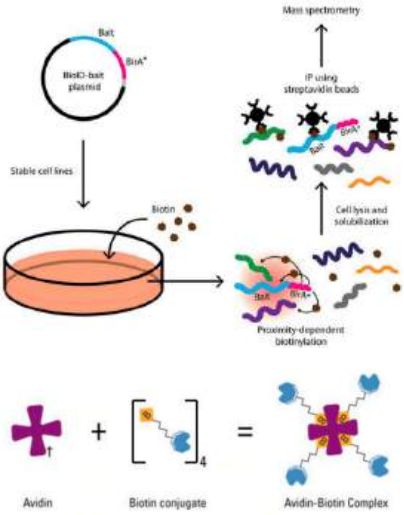


Figure 7.7: BioID mechanism of action

This final mixture of proximal labelled protein will then be ran through a mass spectrometry to be identified.

Essentially, doing these experiment allow us to enhance our understanding of how proteins interact with each other hence how gene are related to

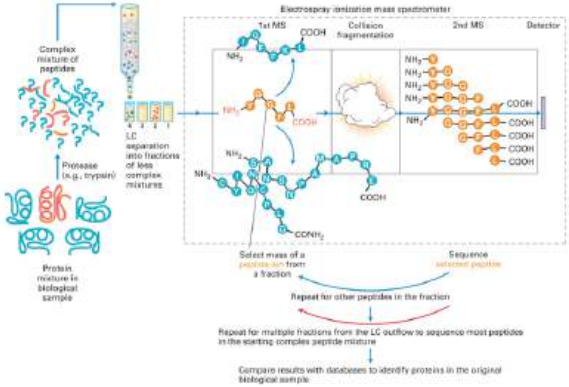


Figure 7.8: Summary of Mass Spectrometry.

one another.

Example 7.1.2. **Oct3/4** is an important TF for pluripotency. They can interact with another gene called **nanog** which can lead to embryo development.

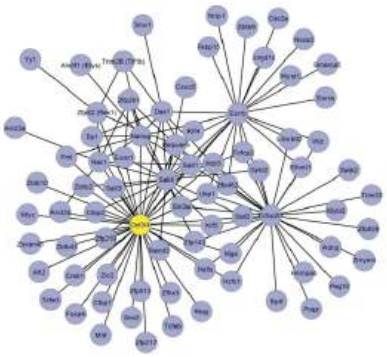


Figure 7.9: Oct-3/4 interaction with many other gene especially nanog which is important for embryogenesis

Not only that, we can see that there are lots of interaction between oct-3/4 and other genes.

7.2 Gene Targeting

We can all agree from this point that to see the function of a gene, the best way to do is to look at its phenotypes whose genes were eliminated i.e. abnormal phenotypes denote disruption of a pathway originated from that gene.

Definition 7.5. The process of mutating an organism and look at its phenotype to then later figure out genes affected is called **forward genetic analysis**. One the other hand, the process of eliminating a gene and look for its function through the abnormal phenotype is called **backward genetic analysis**.

Example 7.2.1. Backward genetic analysis can be observed through the RNAi experimentation of yeast described above where elimination of a gene lead to a potential loss of function. The same with with homologous recombination of yeast gene.

Now some researchers thought that **if we can do homologous combination on yeast, what about other animals?** Well...it was possible
Long ago, researcher was thought to use [flanking] homology directed recombination to engineer a gene (part of a gene) and replace it with a selectable marker (reporter gene of sort).

In this studies/experiment, they used cell that constitute the inner cell mass of mouse early embryo.

Remark 7.4. *These cell can build up an entire organism by themselves except placenta and supportive structure i.e. they're pluripotent cells.*

Because they're pluripotent and found in the embryo, we obviously called them **embryonic stem cell (ESC)**. What we can do with these ECS is to replace or disrupt a gene in it and the grow an entire new organism from it. We will perform a similar experiment as the homologous recombination of *S. Cerevisiae*.

> **Mechanism of Action:** A replacement construct is set up where a gene resistance for *neomycin (G-418)* will be bound by flanking sequences (20-40nt) that is 100% homologous to the that which bound the gene we want to replace on the ESC. Further upstream, we also add a gene for **thymidine kinase** (a *herpes simplex virus* gene) which can interact lethally with antiviral drugs. This gene

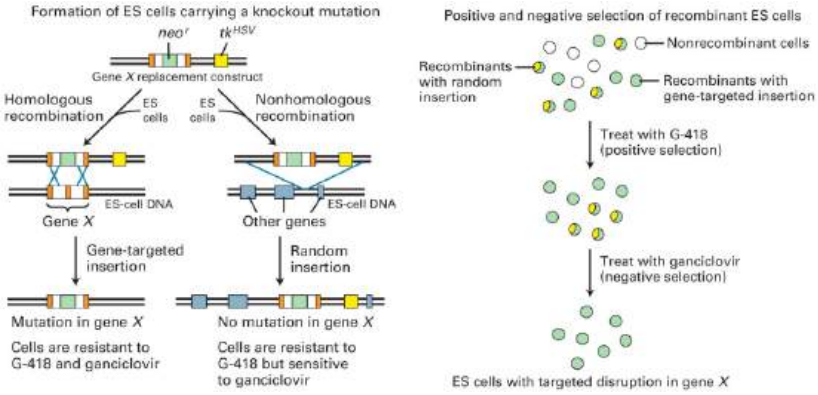


Figure 7.10: Process of making targeted knock out mice.

will then be incubated with the ESC and allow it to uptake the gene. 2 situations can arise from this uptake

1. (Preferable and Rare) ESC homologously recombine replacement construct where the neomycin resistance gene will replaced in to its gene while the thymidine kinase gene is lost.

2. ESC would non-homologously recombine the replacement construct where it randomly insert the entire construct into its genome.

At the end we would get 2 populations experiment which we would go through 2 round of selection. First is the positive selection round where neomycin is added to remove all of the ESC that has not uptake replacement construct. Second is the negative selection where an antiviral called **ganciclovir** is added to remove all the ESC that non-homologously recombine the replacement construct. The thymidine kinase, that isn't lost from this recombination, will react with the antiviral and subsequently kill the cell.

After, these ESC with a selectable marker will be isolated and then inject to a mouse blastocyst where it will contribute to the mouse embryonic formation. One thing about this blastocyst is that it has to encodes for a different coat coloration from the ESC e.g. the ESC can give rise to a mouse with brown coat (dominant trait) while the blastocyst give rise to a black coat (recessive trait).
The progeny that comes out of this manipulation is be the normal blasto-

cyst growth without ESC which is homozygous and give rise to mice with black coat. It can also be the heterozygeous which give rise to mice with brown coat and black stripes, we call it **chimeric** which is the possession of 2 different genotypes. If we then cross these chimeric mice, we would end up with a mouse with a single genotype (homozygous) where it would lack the gene that was previously replaced called **targeted knockout mouse**.

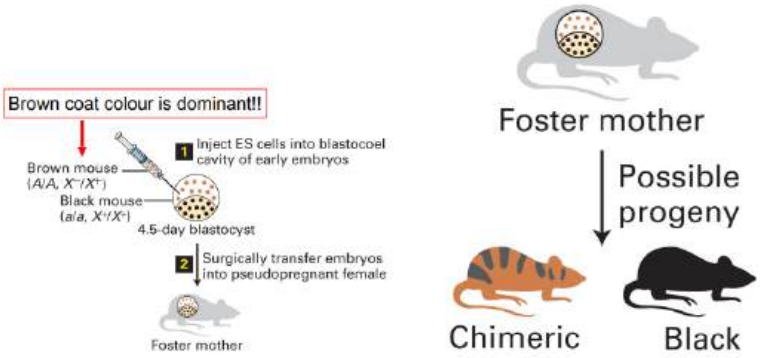


Figure 7.11: Injected the modified ESC to form chimeric mouse

The main problem with this gene editing technique is that it's long, expensive and even sometimes we get a homozygous targeted knockout mouse that die.

7.2.1 Transgenic Mice

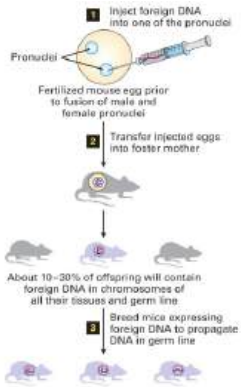


Figure 7.12: Transgenic mice

A better method is to use **transgenic mice**. In this case, fertilized mouse eggs would have its *pronuclei* injected with a transgene which would randomly integrate itself in the chromosome. This fertilized egg is transferred into a foster mother which would give rise to off spring where there's a $10-30\%$ chance that they will contain the original transgene. This first transgenic mouse is then crossed with other transgenic mouse

to create a colonies of numerous transgenic mice.

Remark 7.5. *Unlike targetted knockout mice, transgenic mice do not involved in surgery of embryo into the mice.*

7.2.2 CRISPR-CAS9

This lead to a somewhat beautiful and simpler technique called **CRISPR-CAS9**. This new technique utilize our understanding of bacteria to genome/gene editing.

CRISPR-CAS9 was first observed in the 1980s identify some peculiar repetitive sequences in *E. Coli* and then later (20 years), researchers was able to drawn out that these sequences were in a **clustered regularly interspaced short palindromic regions (CRISPR)**. Sequnces found the these CRISPR has the same homology as bacteriophage's sequences. This was very odd since that means that the bacteria must somehow acquired these DNA sequences but **why would the bacteria need these sequences?** Well...it turns out that this is an acquired immune response of bacteria against virus or bacteriophage

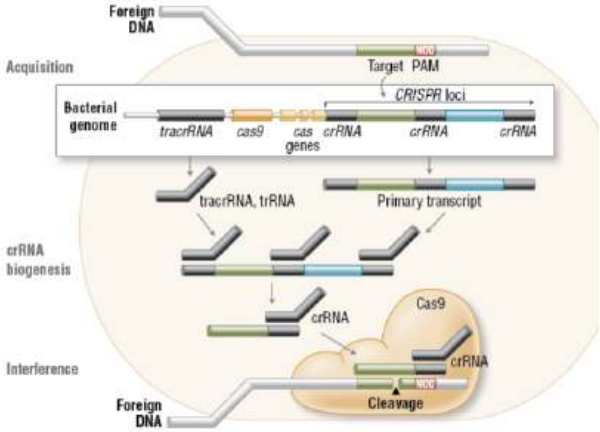


Figure 7.13: Bacteria's acquired immune response using CRISPR-CAS9.

Mechanism of Action (Bacteria Acquired Immune Response): When bacteriophage inject a foreign DNA into bacteria, it would activate bacteria's immune response where it will generate a type of RNA called ***trans*-active CRISPR RNA (tracrRNA)**. This tracrRNA will comes to the CRISPR regions and interact complementary these region. The interaction leads to its maturation and ultimately gives rise to **CRISPR RNA (crRNA)**. The crRNA will recognize and bind to a complex called **CAS9**. Along with this crRNA-CAS9 complex will find the foreign DNA and bind to it. crRNA will search for the target region which is upstream from a conserved region of around 3nt (typically has diguanine) called **protospacer adjacent motif (PAM) sequence**. Once the target regions is found, the CAS9 will activate its endonuclease activity create a double strand break (DSB) 3nt from the PAM region.

Remark 7.6. *CAS9 complex have 3 different domain: RuvC domain for 1 DSB, HNH domain for complementary binding and 1 DSB and a C-terminal domain.*

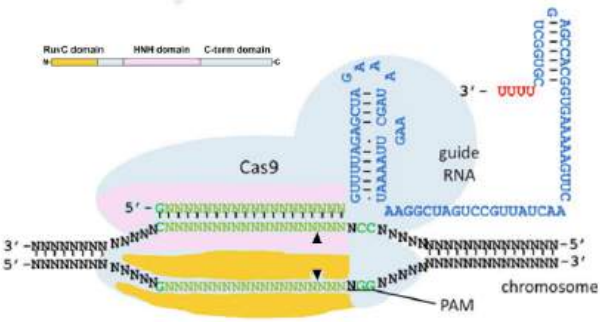


Figure 7.14: Replacing tracrRNA and crRNA with 2 sgRNA.

This new simple discovery lead scientist to question **if we can harness this CRISPR-CAS9 mechanism to do genomic editing or not?** Well...turns out we can although, we need to make some modification. The main modification would be combining the tracrRNA and crRNA together to form a **single guide RNA (sgRNA)**.

To do this gene editing, we first need to have transgenic animals or at least a transgenic cell. Essentially, we need to introduce all of the above

construct into a cell.
In the first transgene, we would have an sgRNA with 20nt that is complement to the target gene region we would want to remove. In the second transgene, we would have it transcribed and translated into a CAS9 complex. We then transfect a cell with these 2 transgenes where it would lead to a DSB of the target gene.

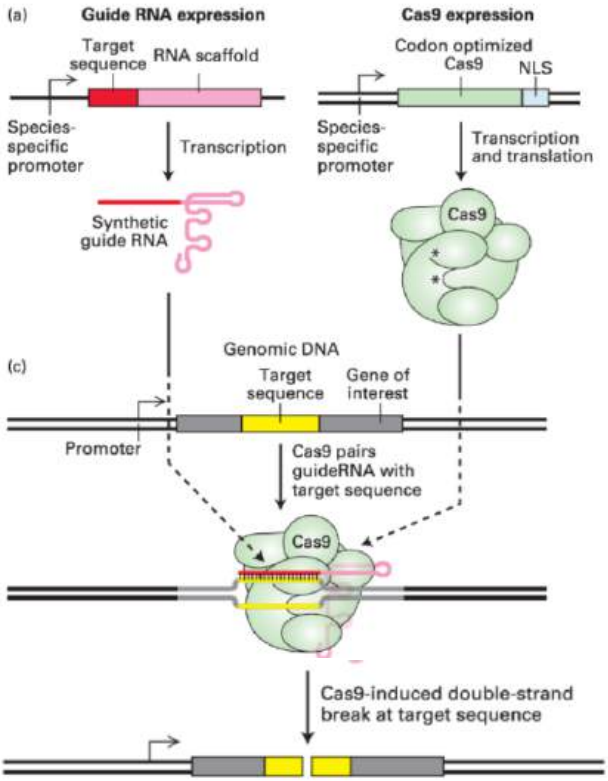


Figure 7.15: Full Mechanism of CRISPR-CAS9 gene editing

Once the DSB happens, the body would naturally fix it and in this case, it would perform a **non-homologous end-joining** where a few bases from the DSB would be randomly added back together. This would subsequently lead to a permature stop codon in the transcript hence shutting down the target gene.

Another way we can also do this is to engineer another transgene repair template that has around 100% homology at the flanking to the target gene. This would allow the cell to perform **homolgous-directed repair** where the target DSB gene would be replaced by the homologous transgene.

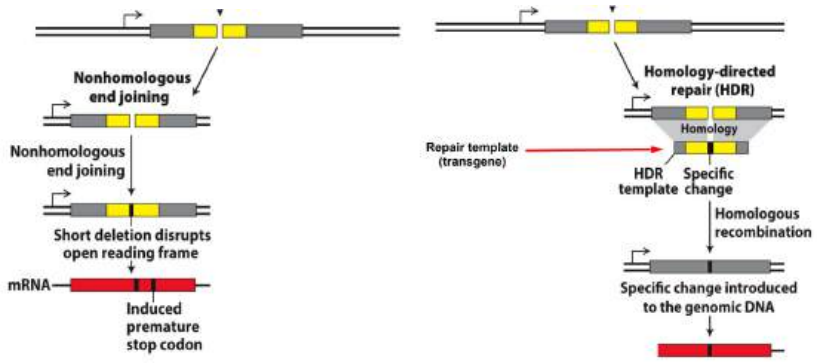


Figure 7.16: DSB rejoining via non-homologous end-joining and homologous-directed repair.

Remark 7.7. *This homologous transgene repair template can even have some modification at non-flanking regions.*

Remark 7.8. *The 2 scientists that develop this CRISPR-CAS9 gene editing technique was later awarded the Nobel prize in chemistry.*

The discovery and development of genome/gene editing revolutionized biology. The next step for the applications of these techniques comes down to ethical reasoning i.e. **Should we allow genetic editing so that the next generation of human becomes more advance etc.?** And this question would be for you, the students to answer.

Note to Author: *Good luck on your final exam on December* 13^{th}*, 2023!!!*

Remark 7.9. *I TRUST RICK ROYYYYY!!!*

Good luck!

Hopefully this notebook can help you!

The physical and chemical properties of the cell and its components in relation to their structure and function. Topics include: protein structure, enzymes and enzyme kinetics; nucleic acid replication, transcription and translation; the genetic code, mutation, recombination, and regulation of gene expression.

Library of Aves' Bindery