Gene Regulation & Epigenetic Processes

Harsh Agrawal

March 16, 2024

Abstract

This notebook compiles content from Molecules Cells and Processes module delivered by Prof. Claire Higgins (2nd Year Imperial College London MBE). A total of 6 lectures are covered in these notes. Please suggest any relevant changes/improvements at ha1822@ic.ac.uk.

Contents

1 Structure and Organization of DNA

1.1 Structure of DNA

We begin our discussion by first establishing the basics of genetics which is the study of genes, and heredity. As we know, genetic information is captured in the form of DNA which is present in all cells of the body.

DNA (or De-oxy Ribonucleic Acid) is a polymer made up of monomeric units which are called nucleotides. Each nucleotide comprises of three parts: a ribose sugar backbone, a nitrogenous base, and three phosphate groups.

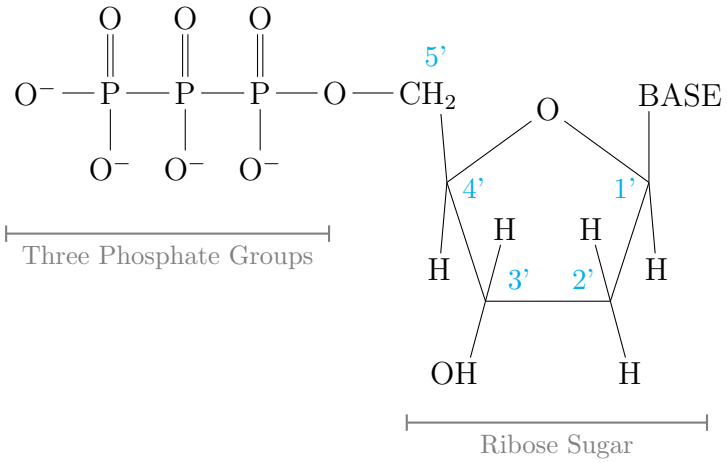

Figure 1: Structure of a Nucleotide

Some important details to note about the structure of DNA are:

- The Carbons in the Ribose sugar are named with the convention of 1' Carbon, 2' Carbon, so on and so forth. The purpose of putting the 'prime' after the Carbon number is to particularly specify the Carbon of the sugar ring as opposed to the Carbon of the nitrogenous base (attached at the 5' end).
- At the 2'C, there is a Hydrogen atom as apposed to a Hydroxyl group (-OH). This is why DNA is called 'De-Oxy' Ribonucleic Acid. In RNA, the 2'C has a Hydroxyl group.
- A single nucleotide has three phosphate groups (PO_4^{3-}) whereas a nucleotide part of the DNA chain only contains one phosphate group. The other two phosphate groups are lost during the formation of the DNA chain. The negative charge on the Phosphate group gives DNA its overall negative charge.

There are four nitrogenous bases (or nucleobases) that are found in DNA and are classified into two groups:

- Purines: Adenine and Guanine. Purines have two N-C rings.

- Pyrimidines: Cytosine and Thymine. Pyrimidines have only one N-C ring. *Note: In RNA, Thymine is replaced by Uracil.*

A nitrogenous base, part of the nucleotide, is referred by its full name which for example, in case of Adenine, would be 2' de-oxyadenosine triphosphate or **A**. For ease of writing, these nucleotides are often referred to by their first letter shortcodes: A, T, C, and G.

Two nucleotides are covalently linked together by a phosphodiester bond between an Oxygen of Phosphate group (at 5' C) of one nucleotide and the 3' Carbon of the sugar of the other nucleotide. This forms the sugar-phosphate backbone of the DNA. A DNA sequence is read from the 5' end (free Phosphate group) to the 3' end (free Hydroxyl group).

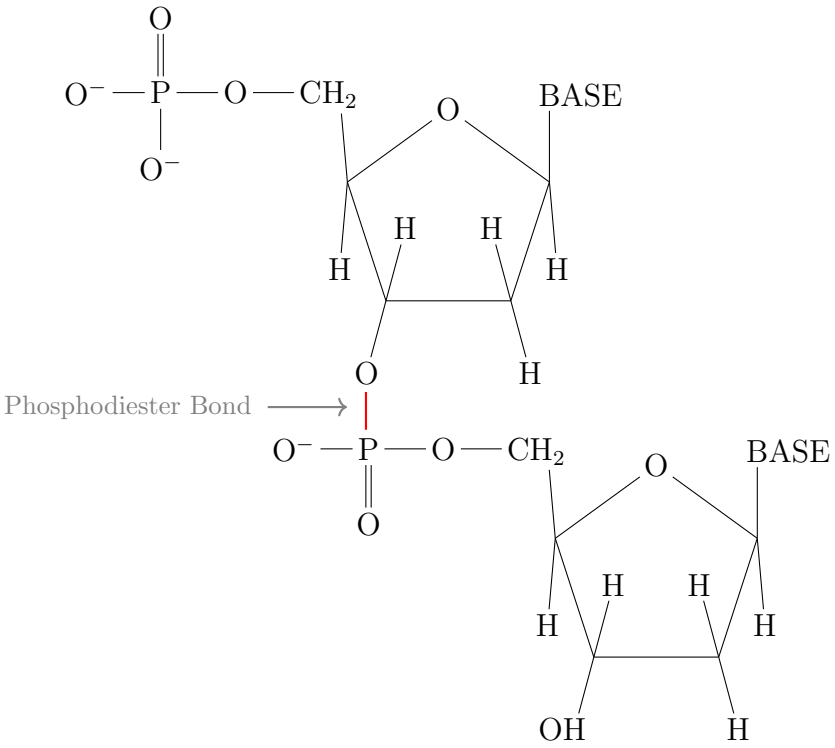

Figure 2: Formation of a Phosphodiester Bond

Nuclebases in the nucleotides also tend to form Hydrogen bonds with an opposing DNA strand. 'A' double bonds to 'T' whereas 'C' triple bonds to 'G'. This is called complementary base pairing. The two strands of DNA are anti-parallel to each other, meaning that the 5' end of one strand is opposite to the 3' end of the other strand.

To maximize the efficiency of base-pair packing, the sugar phosphate backbone wind around each other to give DNA a helical structure. In this structure, the non-polar bases occupy the interior whereas the negatively charged phosphate group occupy the exterior. This gives DNA its famous double helix structure. The length of the DNA is referred by base-pairs (bps). The human genome has around 3 billion base pairs.

1.2 Chromosomes and Organization of DNA

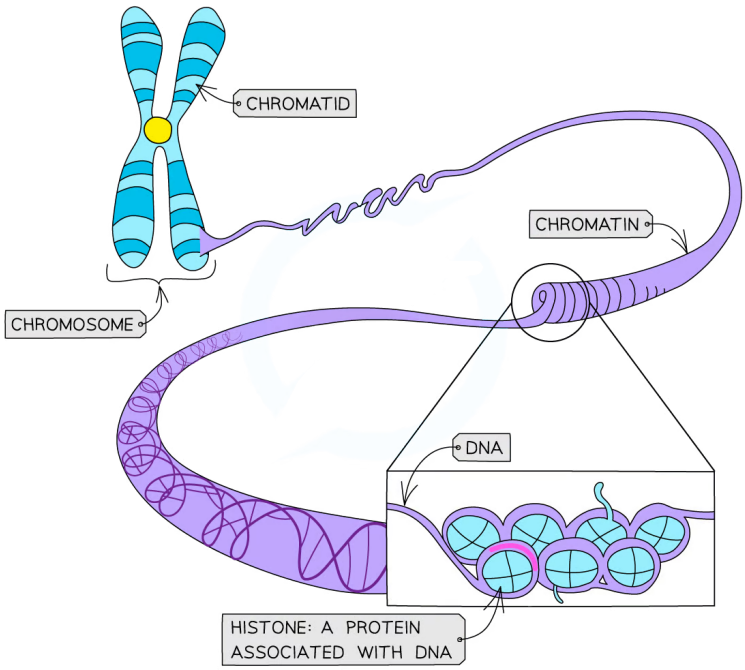

Figure 3: Compression of DNA into Chromosomes. Taken from https://www.savemyexams.com/a-level/biology/cie/22/revision-notes/

DNA is present in a compressed form in the nucleus of the cell called chromosomes.

- Around 146 Base Pairs of DNA winds around 8 'Histone' proteins to form a **nucleosome**. This is the basic unit of DNA packaging. The positive charge on Histones and the negative charge on DNA facilitates their binding. These nucleosomes are often referred as 'beads on a string' (DNA=string, Histones=beads).

- Nucleosomes pack into a coil called **chromatin** (or chromatin fibres). These chromatin fibres further coil into loops and form a **chromatid**. In a somatic cell, there are 46 chromatids. The above image shows two chromatids forming a **chromosome**.

- A chromosome can have one or two chromatids (called sister chromatids). In the growth phase, each chromosome in a somatic cell contains one chromatid but while cell division, each chromatid makes another copy and thus there are 96 chromatids (albiet only 46 chromosomes) in the cell.

De-condensed chromatin is found within the nucleus in tightly packed regions called **Heterochromatin** and loosely packed regions called **Euchromatin**. This can be observed in Figure 4. Euchromatin is more abundant (92%) than heterochromatin (8%)Euchromatin is transcriptionally active.

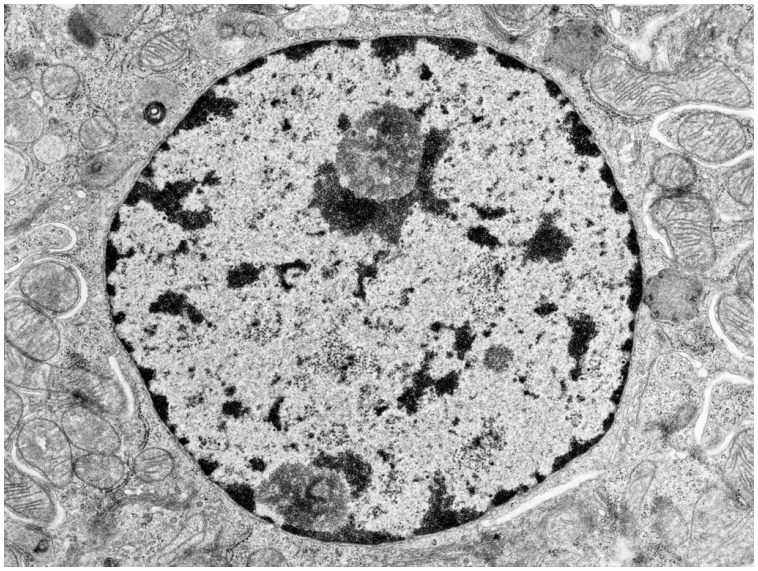

Figure 4: Figure showing euchromatin and heterochromatin. Taken from https://medcell.org/histology/cell_lab/euchromatin_and_heterochromatin.php. The dark regions are heterochromatin whereas the lighter region are euchromatin

1.3 Epigenetic Modifications on Histones & DNA

A Nucleosome is formed by wrapping of DNA around a histone octamer. Two copies of 4 different histone proteins (H2A, H2B, H3, H4) form this octamer. Approximately 146 base pairs of DNA wrap around the octamer. The DNA that links two nucleosomes is called **linker DNA** and is around 20 base pairs long.

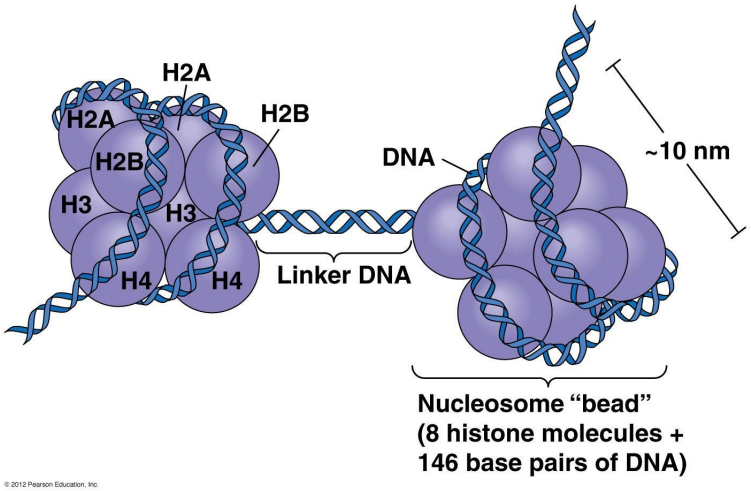

Figure 5: Figure showing DNA wrapped around histones. Taken from https://www.extremetech.com/extreme/213582-new-findings-shed-light-on-fundamental-process-of-dna-repair.

The size of DNA wrapped around the histone octamer was found via MNAse digest enzyme tests which were used to digest the linker DNA and keep the wrapped DNA uncleaved.

This DNA was then run on a gel to obtain the size of the DNA. The original structure estimate of nucleosomes ('beads on a string') was obtained by an electron micrograph.

Histones play a crucial role in gene expression as modifications to histones have the ability to make the DNA more active, or repressed for transcription, as well as affect chromosomal packing. These changes include: methylation (most common), phosphorylation, acetylation, ubiquitination, and sumoylation.

Acetylation of Histones

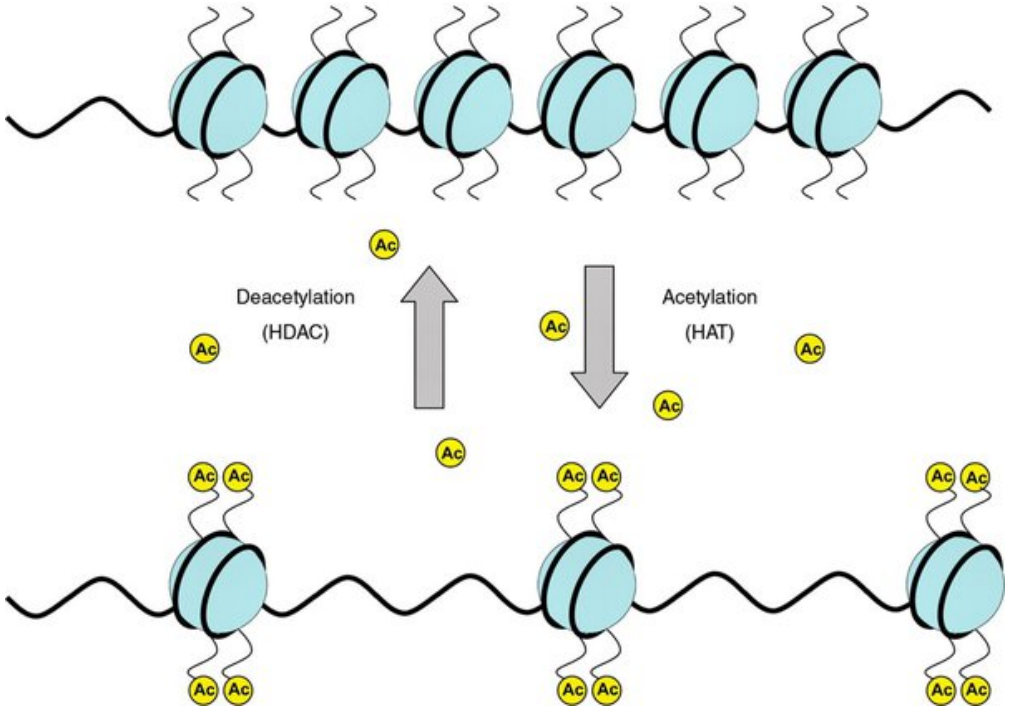

Figure 6: Histone Acetylation. Taken from Eslaminejad, Mohamadreza & Fani, Nesa & Shahhoseini, Maryam. (2013)

One of the most common ways to modify histones is by acetylation. The process occurs as follows:

- The Enzyme group **Histone Acetyl Transferases (HATs)** transfer the acetyle group from acetyl CoA to the lysine side-chain of the histone protein which reduces the overall positive charge on the histone [1].
- This reduction of positive charge weakens the interaction between the histone and the DNA (as DNA is negatively charged), making the DNA more unwound and therefore more accessible for transcription.
- This relaxed chromatin can be reversed back to tightly wound chromatin by another class of enzymes called **Histone Deacetylase (HDAC)** which remove the acetyl group to bring back the original positive charge on the histones.

Mutations in genes correlated to HDAC have been known to cause cancer as it impacts chromosomal packing. This is why HDAC inhibitors are also being explored as a potential cancer treatment.

Methylation of Histones

Methylation is another method of modifying histones.

- The enzyme group responsible for methylation is **Histone Methyl Transferases (HMTs)** which adds a methyl group to the histone protein whereas the enzyme group **Histone Demethylases (HDMs)** are responsible for removing the methyl group.
- There are different methylation marks that are named as 'H3K27me3', 'H3K20me3', etc.
  - 'H3' refers to the third histone protein.
  - 'K' refers to the lysine residue.
  - '20' refers to the position of the lysine residue in the histone protein.
  - 'me3' refers to the addition of three methyl groups on the lysine residue[1]. There can be one, two, or three methyl groups added per lysine residue.
- Some of these marks are repressive while some are activating (can differ from cell to cell).

Methylation of DNA

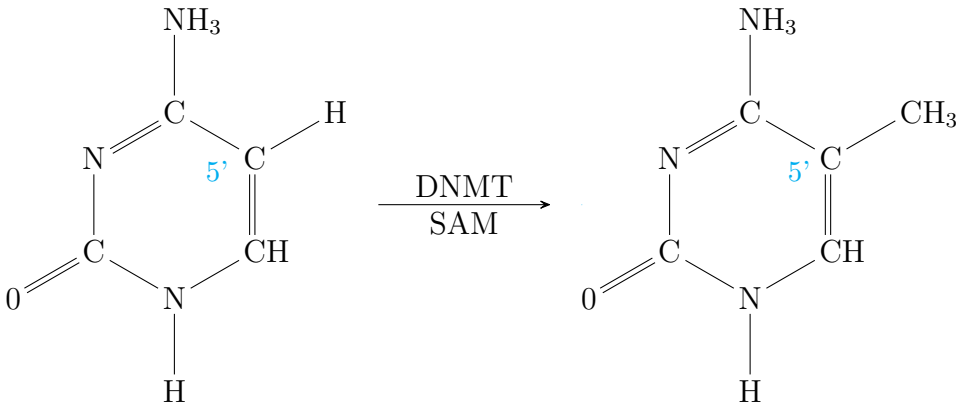

Figure 7: Methylation of Lysine Residue

Apart from occuring on histones, methylation can also occur in DNA by the addition of a methyl group to the 5th Carbon of the **Cytosine** base (only) on the DNA strand (figure 7). This is catalyzed by a class of enzymes called **DNA Methyl Transferases (DNMTs)**. DNMTs transfer a methyl group from S-adenosyl methionine (SAM) to the cytosine base.

[1]Amino Acids as part of proteins are also called *residues*.

This only occurs on a Cytosine base adjacent to a Guanine base, also referred to as a **CpG** site[2]; eg: TAT**C**GTGCT (*only the bold C can be methylated*).

DNA methylation is a repressive mark and it makes the DNA more packed and less accessible for transcription. Thus, there is an increase in DNA methylation in heterochromatin as compared to euchromatin.

1.4 Gene Regulation due to Epigenetic Modifications

Gene expression is highly regulated by these epigenetic changes. When the chromatin is in a repressed state (methylated), the DNA is inaccessible for the transcription factors to bind to the promotor regions of the respective genes, thus hindering the expression of the gene.

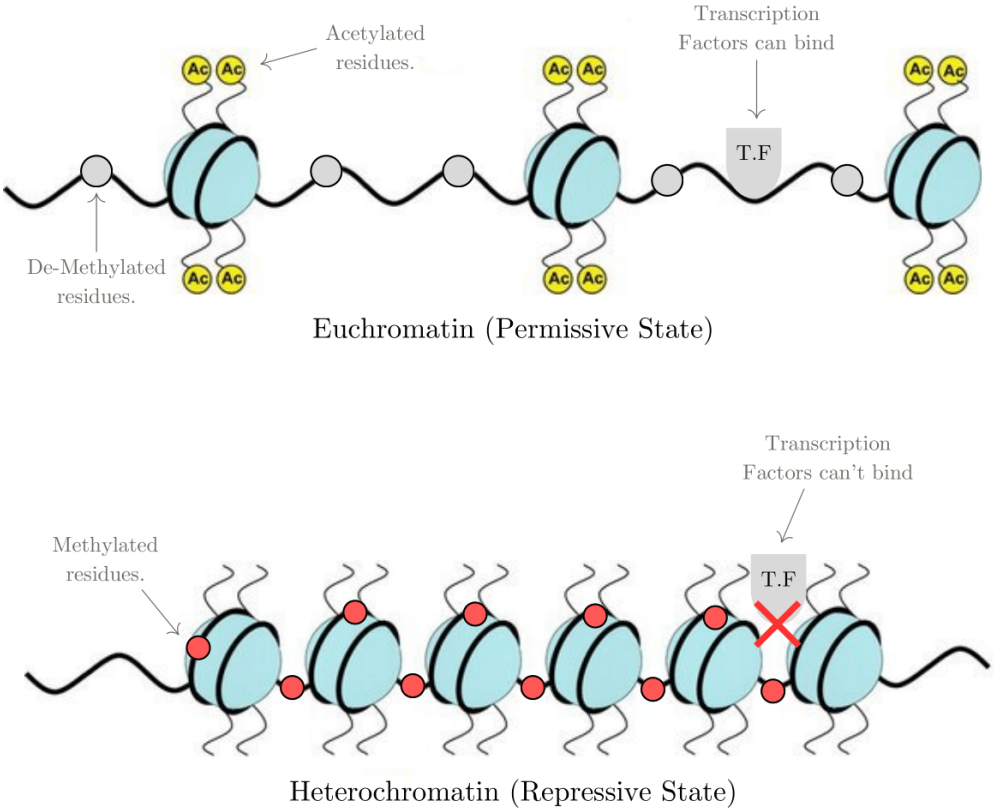

Figure 8: Annotations showing how repressed chromain is inaccessible for transcription factors to bind. Source image taken from Eslaminejad, Mohamadreza & Fani, Nesa & Shahhoseini, Maryam. (2013)

Environmental factors play an important role in epigenetic modifications to chromatin. For example, a study conducted in plants revealed that cold stress (an environmental factor) resulted in the degredation of Histone Deacetylase (HDAC), which in turn increased acetylation, and increase transcription [2].

[2]CpG stands for Cytosine-Phosphate-Guanine.

The transcriptional state of chromatin can be more presicely stated as:

- **Active Chromatin**: Active chromatin is accessible for transcription factors to bind and the genes can thus be transcribed.
- **Posed Chromatin**: In the poised state, the chromatin is physically closed and the majority of the transcription factors can't physically bind. However, its found that a set of transcription factors, called the the **pioneer factors**, can bind to this poised state. These pioneer factors can then recruit enzymes such as HATs etc. to acetylate and open up the chromatin for transcription.
- **Repressed Chromatin**: In the repressed state, the chromatin is tightly packed and none of the transcription factors (including pioneer factors) can bind to the promotor regions of the genes.

Histone marks distinguish between repressed and permissive states. For example, the presence of a trivalent motif: H3K4me1, H3K27ac, and H3K9me3 on the chromatin makes it permissive whereas it's absence makes the chromatin repressive.

Direct Reprogramming of Fibroblasts

- Direct Reprogramming is carried out by adding a few essential transcription factors to Fibroblasts that then swtich on multiple hundreds or more genes required for the fibroblast to be converted to a Neuron.
- One such transcription factor is **Ascl1** which when added to the fibroblasts is able to convert them into neurons, whereas when the same transcription factor is added to Keratinocytes, conversion to neurons isn't observed.
- This is because Fibroblasts have permissive chromatin (presence of **the trivalent motif**) at the Ascl1 binding sites allowing transcription of Ascl1 and all further genes, to occur. The absence of the trivalent motif in the Ascl1 binding regions of Keratinocytes prevent downstream transcription and thus the conversion to Neurons isn't observed.

2 RNA Transcription

2.1 Transcription process

The transcription of gene (DNA → RNA) is carried out in three stages: initiation, elongation, and termination.

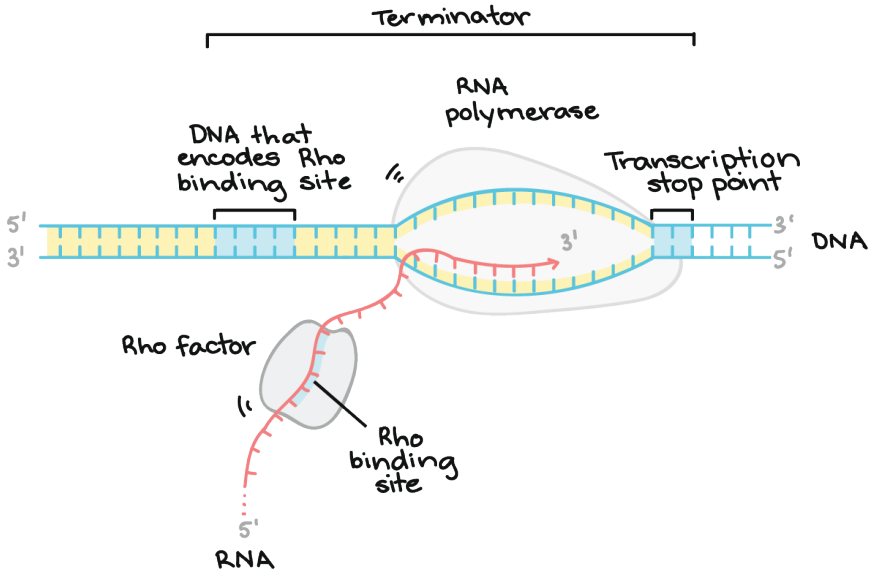

Figure 9: Image showing transcription. Source: https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/eukaryotic-pre-mrna-processing

Initiation

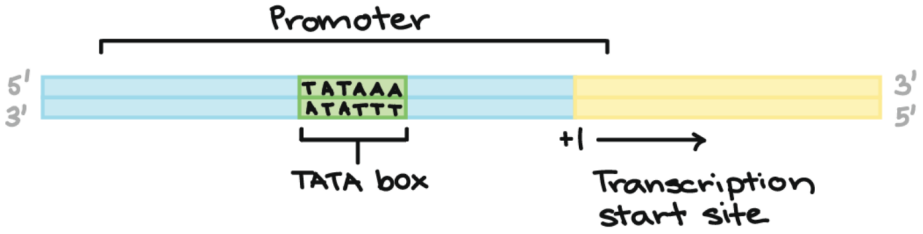

Figure 10: Image showing the TATA Box. Source: https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/eukaryotic-pre-mrna-processing

- The enzyme responsible for the conversion of DNA to RNA is called **RNA Polymerase**. binds to a region of the DNA (in the 5' end) called the **promotor region**. The promotor region is a sequence of DNA that signals the start of the gene.

- The site on the DNA from which the first RNA nucleotide is transcribed is called the +1 site, or the **initiation site**. Nucleotides that come before the initiation site are given negative numbers and said to be upstream. Nucleotides that come after the initiation site are marked with positive numbers and said to be downstream[3].

- In eukaryotes, this promoter sequence consists of a region called the **TATA box** (figure 10) which consists of A, T repeats[4] and is found 25bp upstream of the transcription site. The TATA box is recognized by **TATA binding protein** (TBP) that is essential to forming the preinitiation complex which then recruits the RNA polymerase to attach and start transcription.

- In Cancer, the TATA box is often found to be mutated. In case of Tumor Suppressor Genes, the TATA box is mutated for reduced TBP binding $\longrightarrow$ reduced transcription of Tumor Suppressor Genes $\longrightarrow$ decreased suppression of cancer cells $\longrightarrow$ higher cancer malignancy. In case of Oncogenes, the TATA box might be mutated to increase TBP binding thus promoting the transcription of the oncogenes.

- The DNA strand that runs from 5' $\rightarrow$ 3' is called the **coding strand** or the **sense** strand. This strand isn't used for transcription. The antiparallel strand (that runs from 3' $\rightarrow$ 5') is called the **template strand** or the **anti-sense strand** and is used for transcription.

Elongation, Termination, and Post-Processing

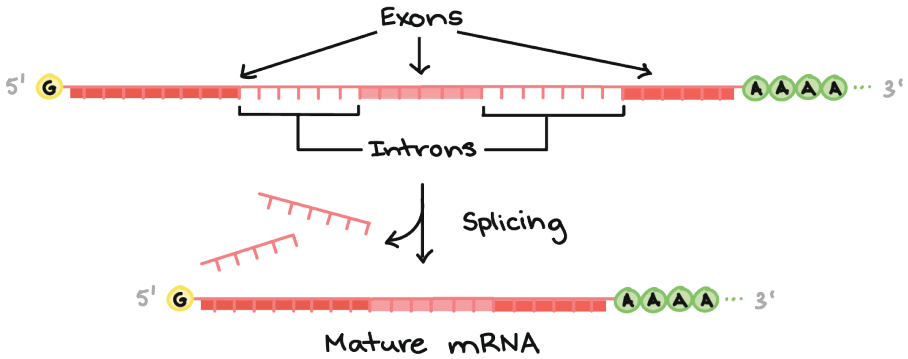

Figure 11: Image showing Splicing of RNA transcript to mRNA. Source: https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/eukaryotic-pre-mrna-processing

[3]Sentence taken directly from: https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/stages-of-transcription

[4]Identified in 1978 and is found in ~30% of human gene promoters

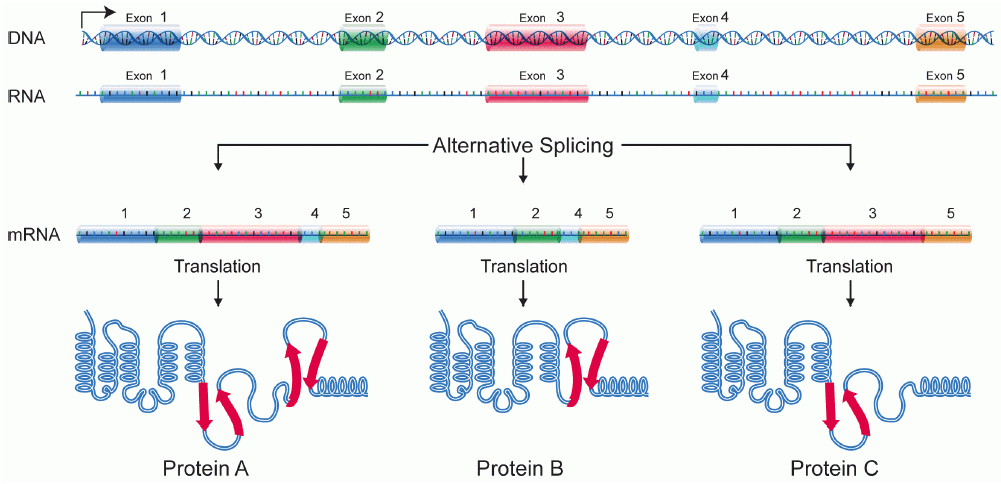

Figure 12: Image showing Alternative Splicing to create different mature mRNA transcripts. Source:
https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/eukaryotic-pre-mrna-processing

- The process of Elongation is the extension of the newly synthesized RNA strand. The RNA polymerase moves along the DNA template strand and adds nucleotides to the 3' end of the RNA strand. The RNA polymerase, unlike DNA polymerase, doesn't require a primer to start the elongation process.
- The synthesized RNA strand is complementary to the anti-sense strand is thus identical to the sense strand (except for the Uracil base switch in place of Thymine).
- This RNA transcript is then processed by the following the addition of a 5' cap and a 3' poly-A tail. This is added to prevent RNA from being degraded.
- It's followed by splicing the RNA to remove introns and join exons to form the mature mRNA (figure 11). Every 5' splice site on an intron usually starts 'GT' whereas the 3' splice site ends with 'AG'.
- The region of the transcript that codes for the protein is called the **Open Reading Frame (ORF)**. The rest of the transcript is called the **Untranslated Region (UTRs)**[5].
- Initially, it was presumed that humans had *sim* 100,000 genes. However, it was later found that humans only have 20,000 but the number of proteins coded by these genes is 100,000. This was found to be due to **Alternative Splicing** (figure 12) where different exons are joined together to form different mature mRNA transcripts.
- It's estimated that *sim*60% of the genes in the human genome have multiple splice variants, thus coding for slightly different proteins.

[5]Start codon of the ORF is usually 'ATG' whereas the stop codon is usually 'TAA' / 'TAG'.

- **5' Promoter Region**: This is the region where relevant proteins such as RNA polymerase and transcription factors bind to initiate transcription of that gene.
- **3' Poly-A Tail**: This is a sequence of Adenine bases added to the 3' end of the RNA transcript. This is important for the stability of the RNA transcript.
- **Protein Coding Region**: This is the region that comprises of **introns** and **exons**. Introns are spliced away and exons are joined together to form the mature mRNA that is translated into a protein[6].

Introns comprise of over 99% of the entire genome (not translated to proteins). However, around 90% of observed phenotype-related mutations are found in exons.

[6]Whole genome sequencing is quite expensive and in-efficient. Since only the exome codes for proteins, whole exome sequening has proven to be a much more efficient alternative

References

[1] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, Feb 2011.

[2] Junghoon Park and Chae Jin Lim. Epigenetic switch from repressive to permissive chromatin in response to cold stress. *Proceedings of the National Academy of Sciences*, 115(23), May 2018.