Bioinformatics

a.a. 2018-2019

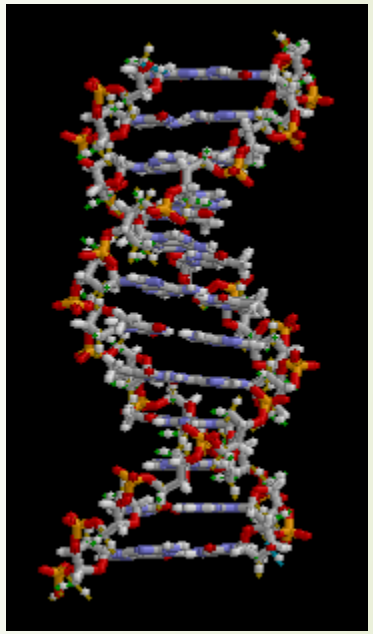
Prof. Elisa Ficarra

Introduction to Molecular biology

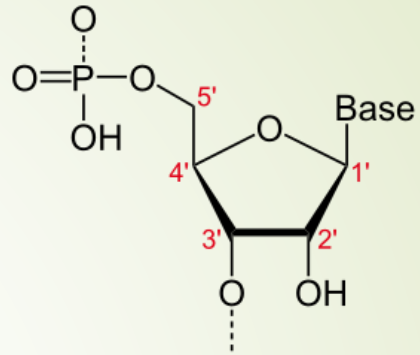
DNA molecule (in Human beings)

- The **Deoxyribonucleic acid** is a molecule that carries the genetic instructions used in the growth, development, functioning and reproduction of all known living organisms and many viruses.

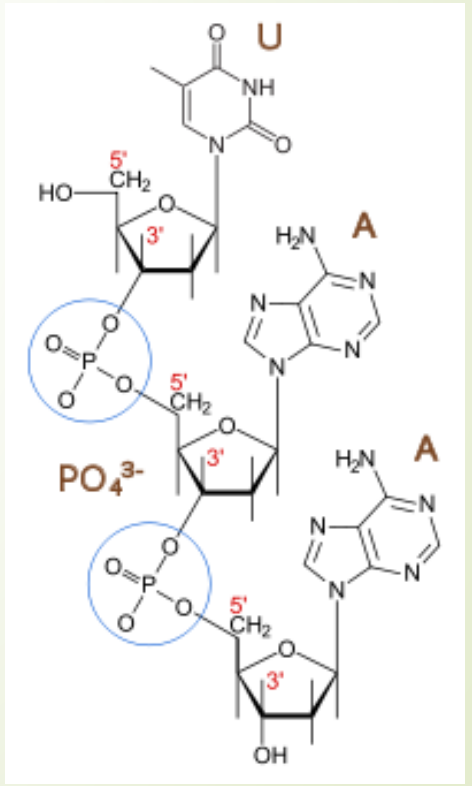
- DNA molecules consist of two biopolymer strands (i.e. large molecules composed of many repeated subunits) coiled around each other to form a double helix.



DNA

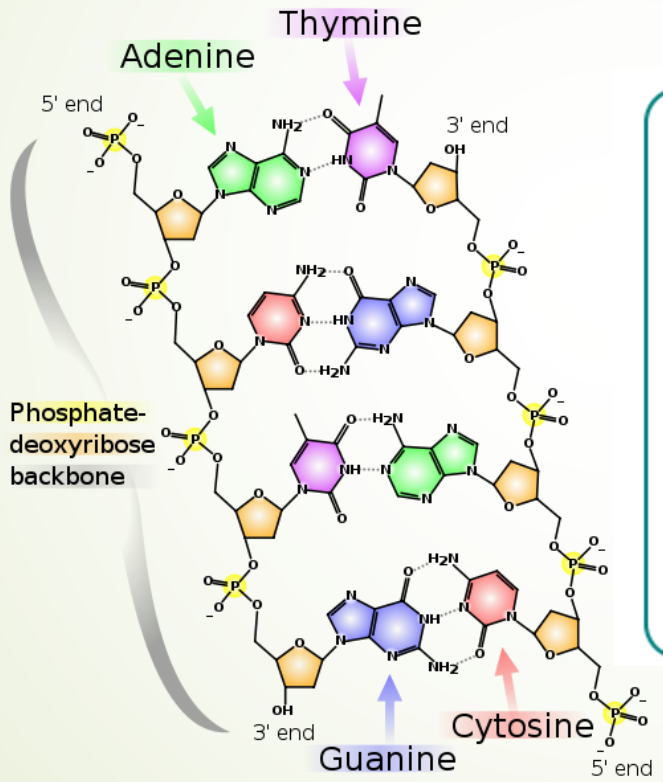


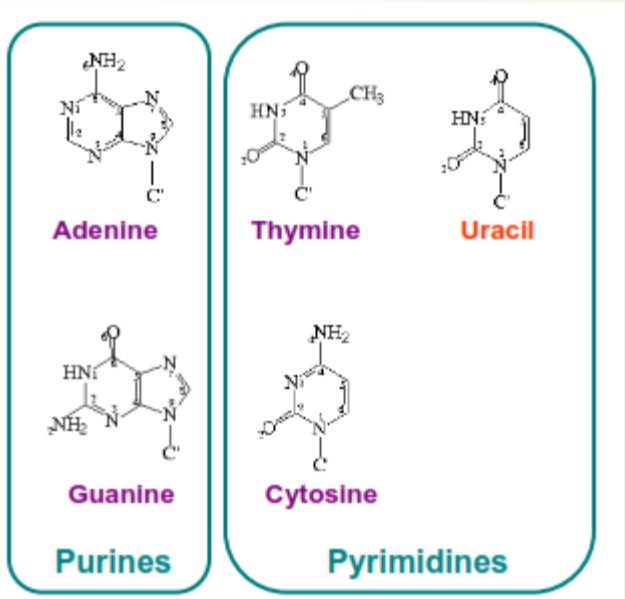
- The backbone of the DNA strand is made from alternating phosphate and sugar residues.
- The sugar in DNA is 2-deoxyribose, which is a pentose (five-carbon) sugar.
- The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings.



- These asymmetric bonds mean a strand of DNA has a direction. In a double helix the strands are antiparallel (i.e. the direction of the nucleotides in one strand is opposite to their direction in the other strand).
- The asymmetric ends of DNA strands are called the 5′ (five prime) and 3′ (three prime) ends, with the 5' end having a terminal phosphate group and the 3' end a terminal hydroxyl group.

DNA - bases


Thymine
Adenine
5' end
3' end
Phosphate-
deoxyribose
backbone
3' end
Guanine
Cytosine
5' end


Adenine
Thymine
Uracil
Guanine
Cytosine
Purines
Pyrimidines

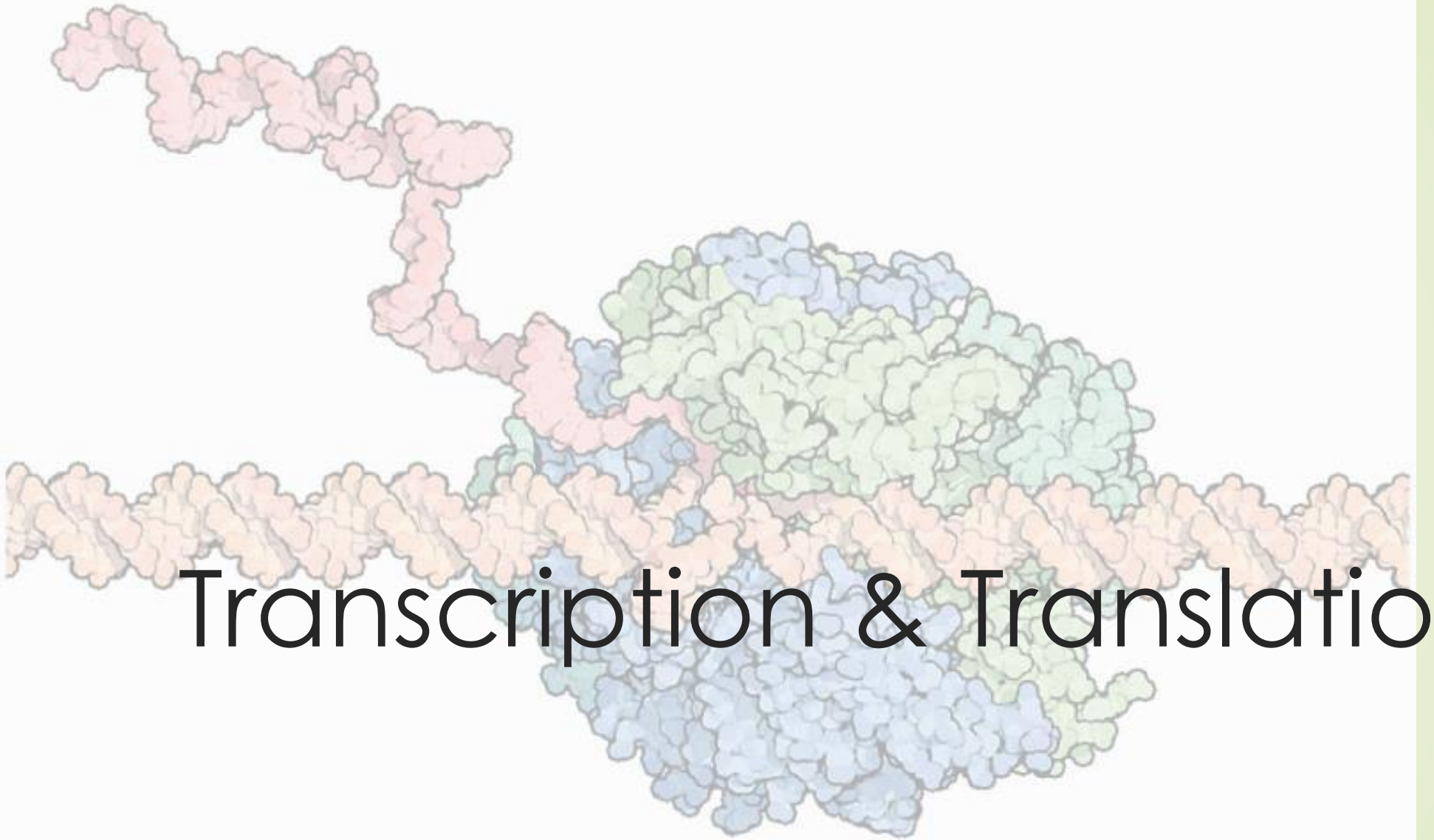
Central Dogma of Molecular Biology

DNA synthesis (replication) -> Cell duplication

DNA transcription (activation) -> Gene Expression

Gene Expression

- There are 4 major events that occur during the process of gene expression
  - Transcription
  - RNA processing
  - Translation
  - Protein processing



Transcription & Translation

Replication, Transcription and translation at a glance

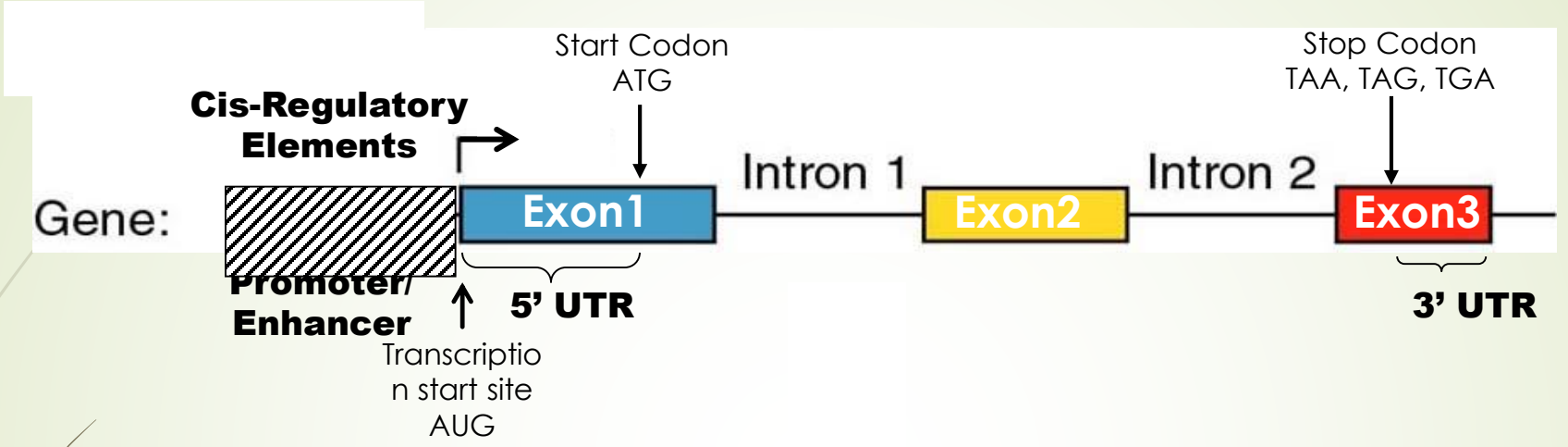
DNA synthesis (replication) maintains the genetic information and passes this to the next generation.

RNA synthesis (transcription) is a transfer of the information from the DNA where it is stored into RNA which can be transported and interpreted.
The language (of nucleic acids) is essentially the same.

Messenger RNA (mRNA) moves the information from the DNA to the ribosomes to enable the **production of protein (translation).**

Translation represents a change in the language from the nucleotide letters in the RNA to the amino acid letters in the protein.

Eukaryotic Gene Structure



Regulatory Sequences

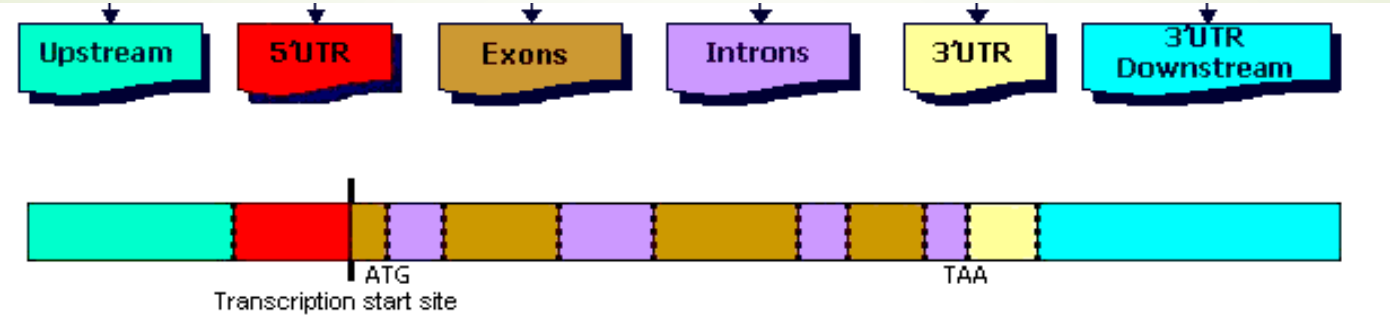
Structural or Coding Sequences

Termination sequence

Regulatory and Coding Sequence Unit = Operon

The gene is a Transcription Unit.
One gene -> One protein (with possible several isoforms)

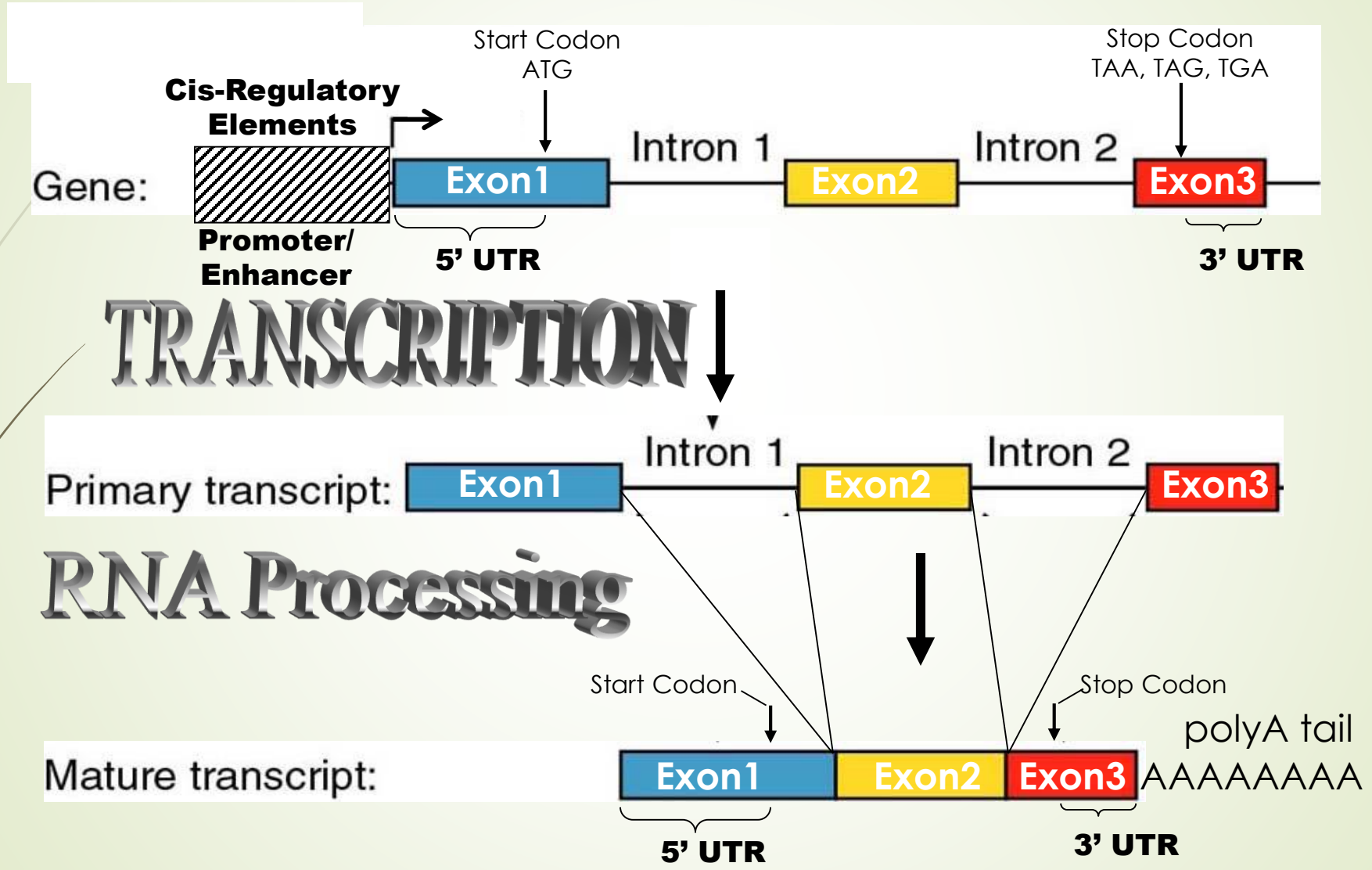
Cis-regulatory elements - UTRs



• **Untranslated region (UTR) –** UTRs may affect gene regulation & expression. Normally untranslated. In Humans is partially translated because the start codon is part of the 5'UTR

• ***Cis*-regulatory elements (CREs) –** sequences that control the transcription of a gene. Cis-elements may be located in 5' or 3' untranslated regions or upstream the 5'UTR or within introns. Promoters facilitate the transcription of a particular gene and are typically upstream of the coding region (and the 5'UTR).

Eukaryotic Gene Structure



Major classes of RNA

All three major classes of RNA (mRNA, tRNA, & rRNA) are synthesized by transcription of the appropriate genes and are involved in protein synthesis.

All three major types of RNA are involved in directing the formation of protein.

1) mRNA carries the message from the DNA to the ribosome.

2) rRNAs are major structural components of the protein-synthesizing ribosome.

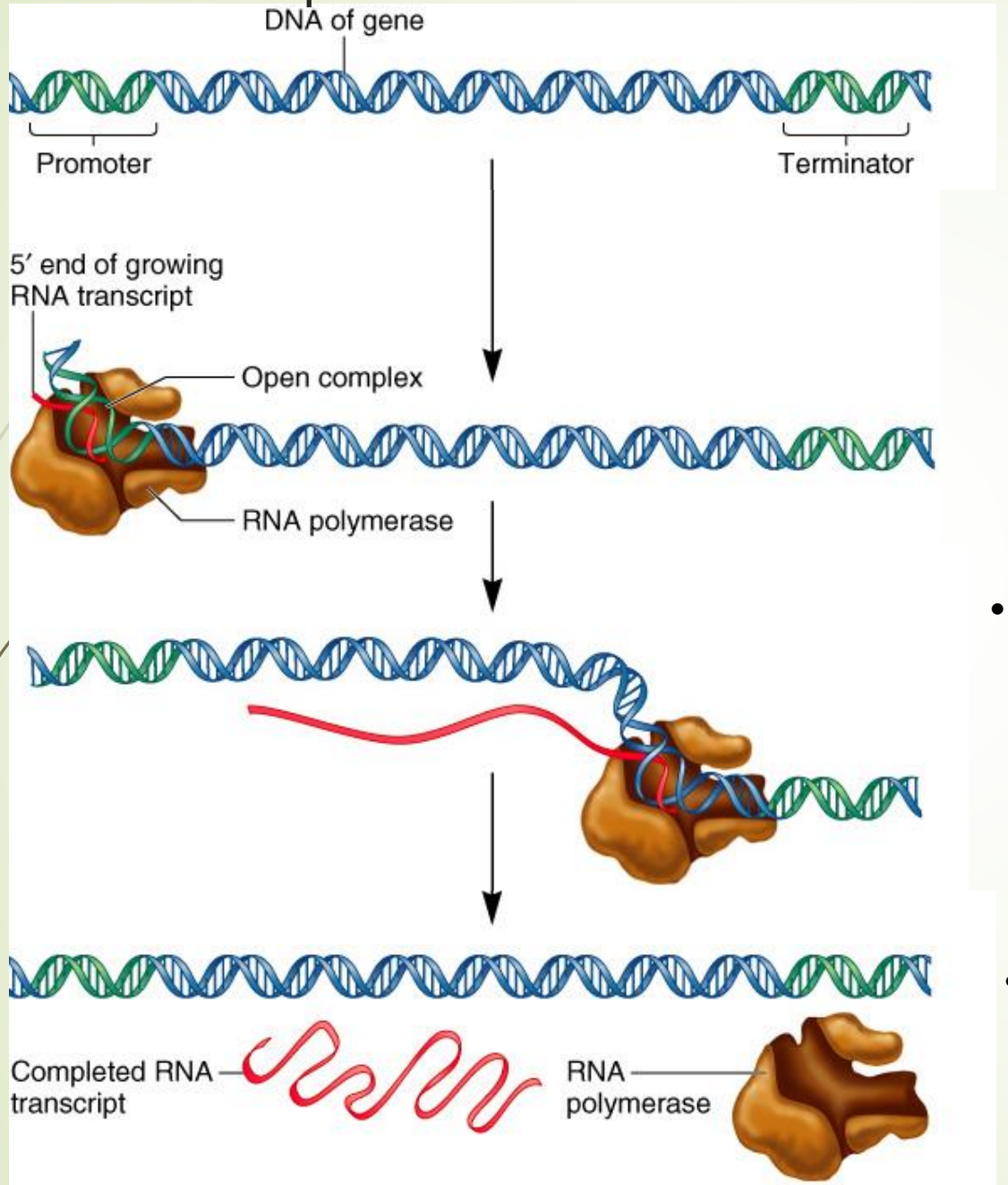
3) tRNAs act as adaptor molecules in aligning the amino acids according to the sequence present in the mRNA.

Non coding RNA – functional RNA molecules that are not translated into protein, e.g. ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA (microRNAs are predicted to control the translational activity of approximately 30% of all protein-coding genes in

Transcription and Translation

https://www.youtube.com/watch?v=2BwWavExcFI

Transcription Proceeds Through 3 Steps



Initiation

- **Transcription factors & RNA polymerase recognize & bind the promoter**
- **DNA adjacent to the promoter is denatured forming the open promoter complex**

Elongation

- **RNA polymerase moves along the DNA in synthesizing a RNA transcript. Synthesis is 5'→3' – Only 1 strand of DNA is read as a template.**

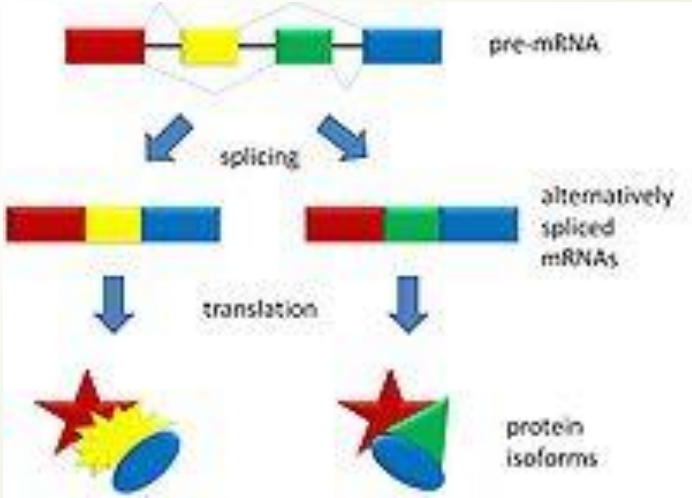
Termination

- **A termination signal is reached causing RNA polymerase to dissociated from the DNA**

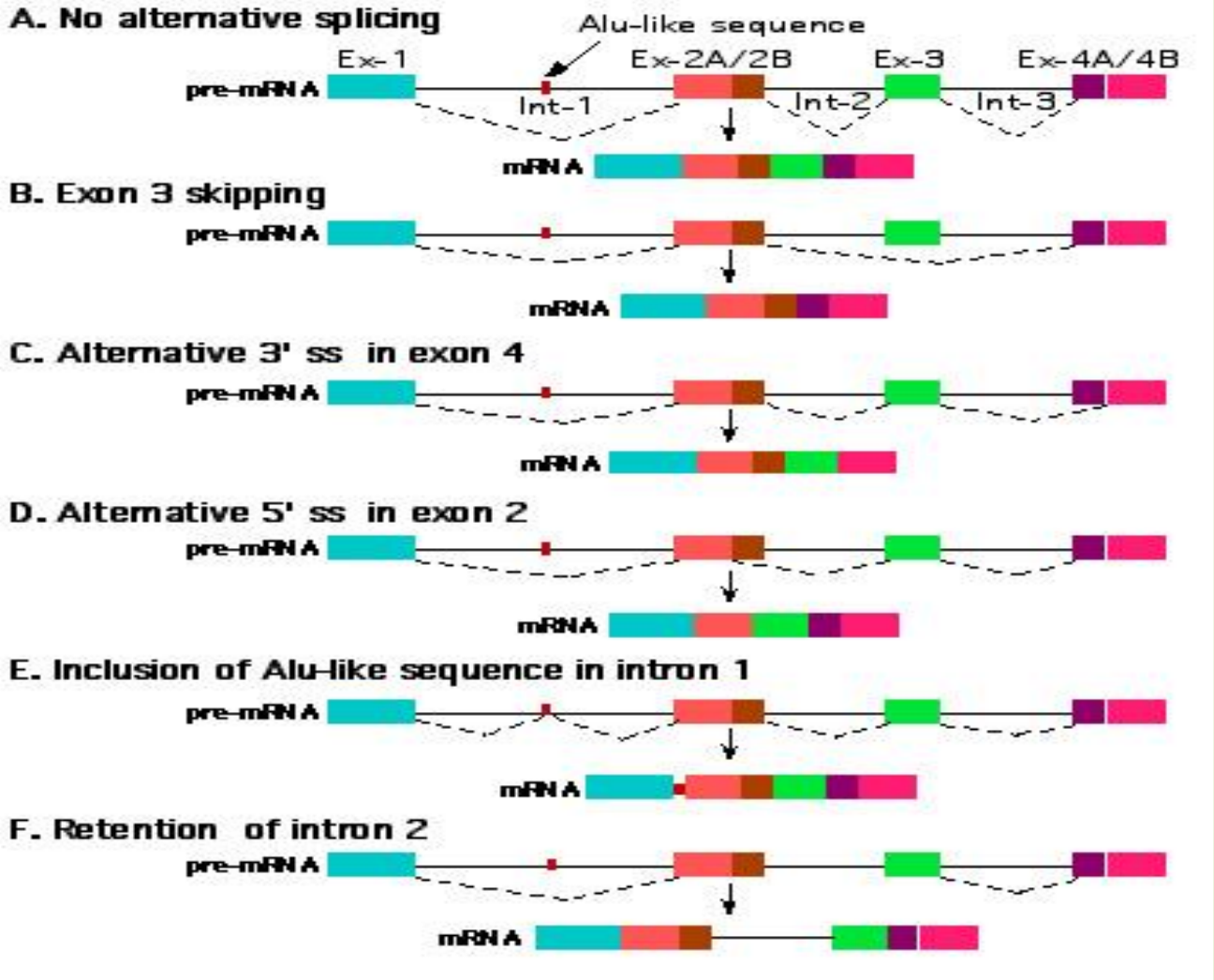
Alternative Splicing

- Process by which the exons of the RNA produced by transcription of a gene are reconnected in multiple ways during RNA splicing. The resulting different mRNAs may be translated into different protein isoforms; thus, a single gene may code for multiple proteins.

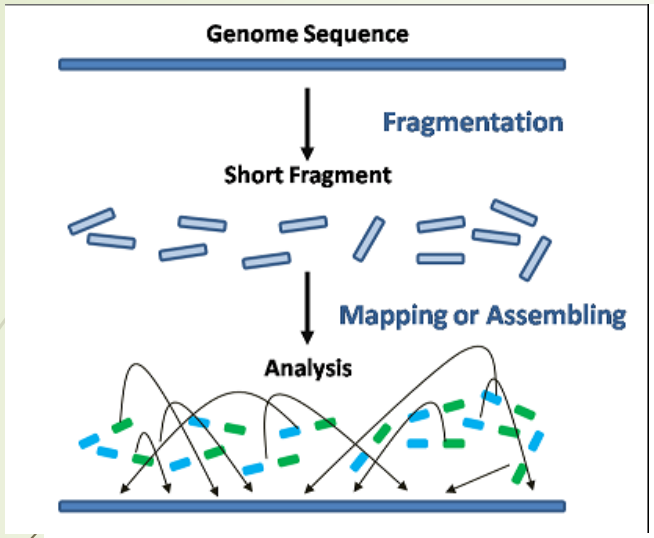
- It greatly increases the diversity of proteins that can be encoded by the genome; in humans, over 80% of genes are alternatively spliced.
- There are numerous modes of alternative splicing observed. The most common is exon skipping. In this mode, a particular exon may be included in mRNAs under some conditions or in particular tissues, and omitted from the mRNA in others

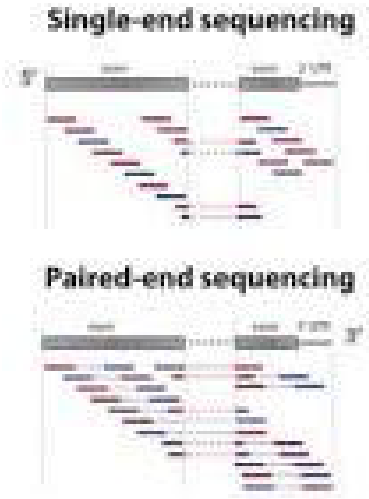


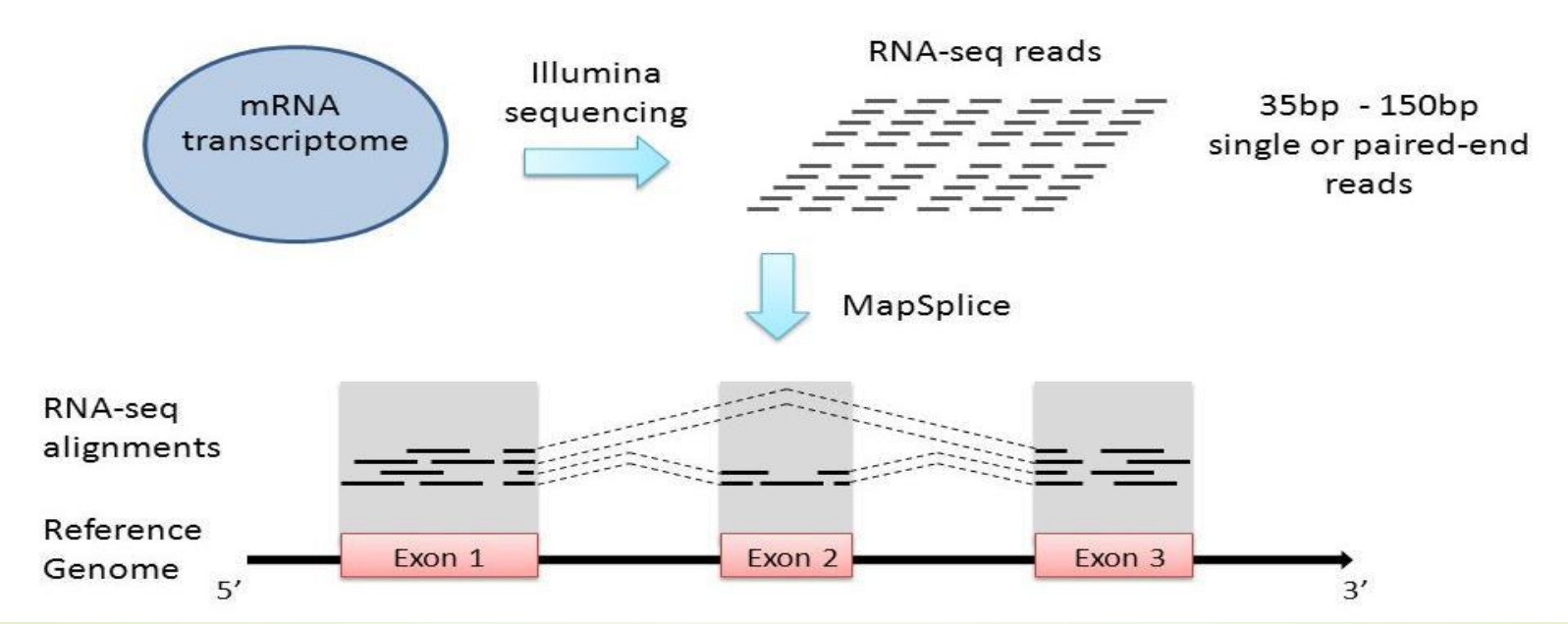
Examples of Alternative splicing



A glance to *Sequencing*

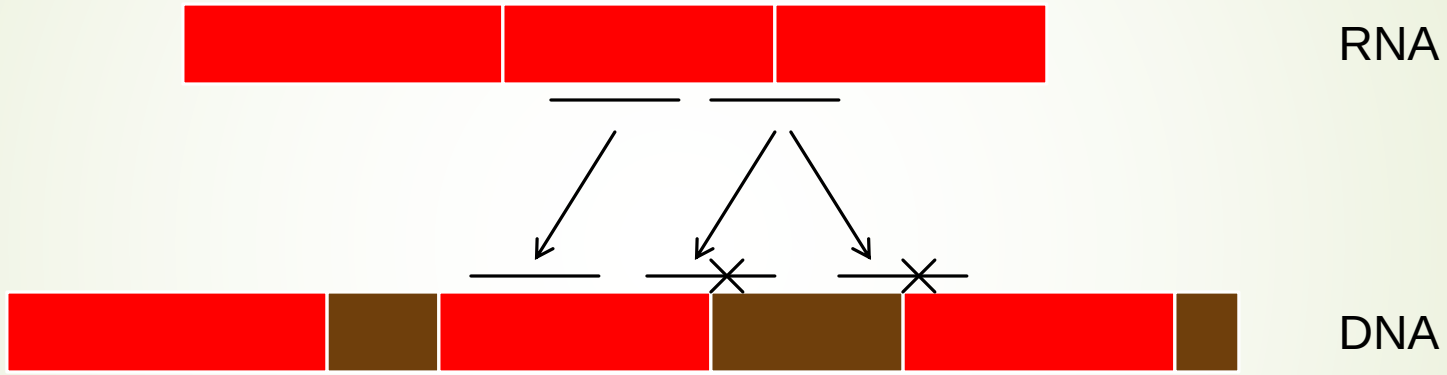






Splicing Effects on Alignment

- Considering splicing enables a more precise read mapping and thus expression analysis

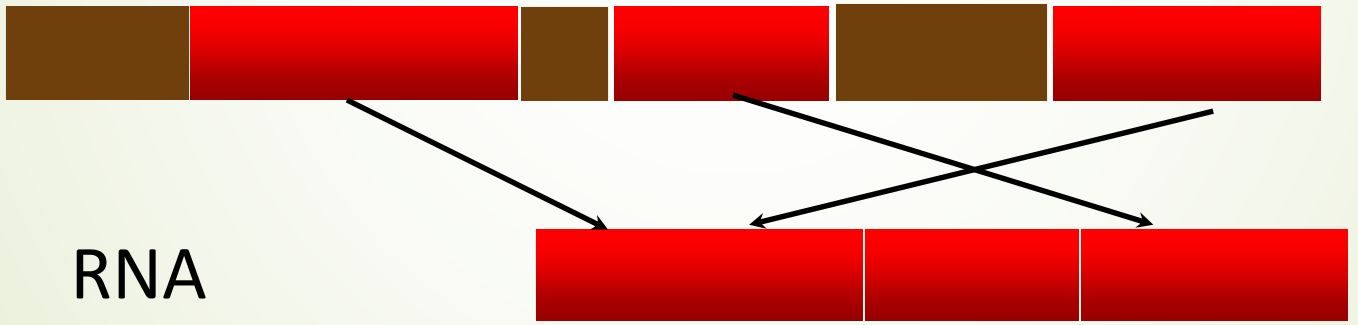


- Alternative slicing needs to take into account also all the possible combination between exons, in terms of order and number.
- A large number of reads across the real junctions would not be mapped without splicing-aware alignment. As an alternative there are tools (like BLAST) that trim the reads in small sequences and try to align them.

Alternative Splicing (2)

DNA Reference

DNA: Exons and Introns



Sequencing RNA

Mutations

- **By effect on structure**
  - **Small-scale mutations, such as those affecting a small gene in one or a few nucleotides** (e.g. point mutations, insertions, deletions, duplications)
  - **Large-scale mutations in chromosomal structure** (e.g. Gene duplication, Chromosomal translocation, Chromosomal inversions, Interstitial deletions)
- **By effect on function**
  - **Loss-of-function mutations** are the result of gene product having less or no function. When the allele has a complete loss of function (null allele) it is often called an amorphic mutation. Phenotypes associated with such mutations are most often recessive
  - **Gain-of-function mutations** change the gene product such that it gains a new and abnormal function. These mutations usually have dominant phenotypes
- **By effect on fitness**
  - In applied genetics it is usual to speak of mutations as either harmful or beneficial

Fusion transcripts (*Chimeras*)

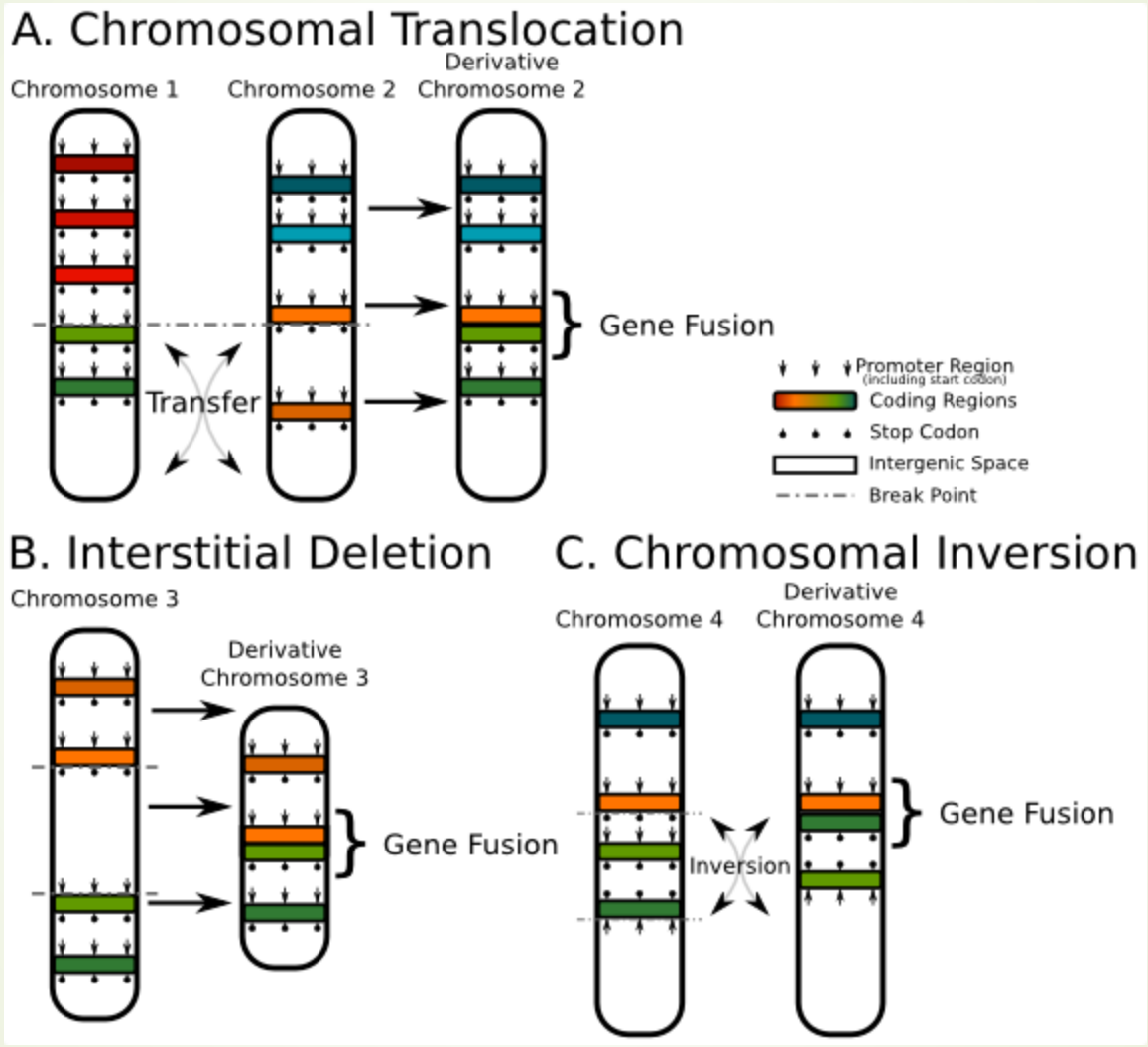
- Chimeric RNA encoded by
  - a fusion gene at DNA level
  - a fused mRNA coming from multiple RNA polymerases

- A fusion gene is a hybrid gene formed from two previously separate genes. It can occur as the result of a translocation, interstitial deletion, or chromosomal inversion. Often, fusion genes are oncogenes

- "normal" (cis-)splicing processes a single molecule. In contrast, trans-splicing results in an RNA transcript that comes from multiple RNA polymerases on the genome. Exons from two different primary RNA transcripts are joined

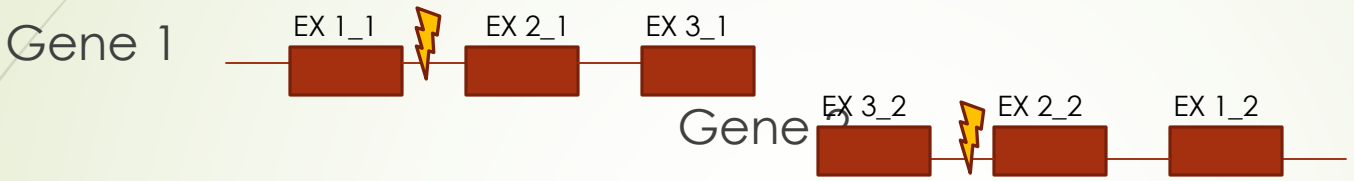
- When the RNA is affected by a pathology, some intron can be retained and it can be part of the fusion too (as UTRs and other usually non coding regions)

Gene fusion mechanisms

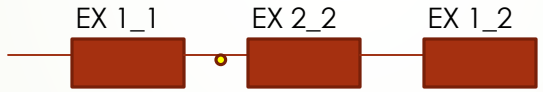


Example of gene fusion in mRNA

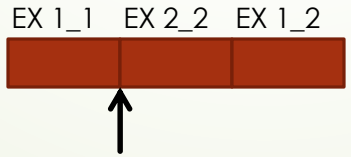
- DNA



- Fused DNA



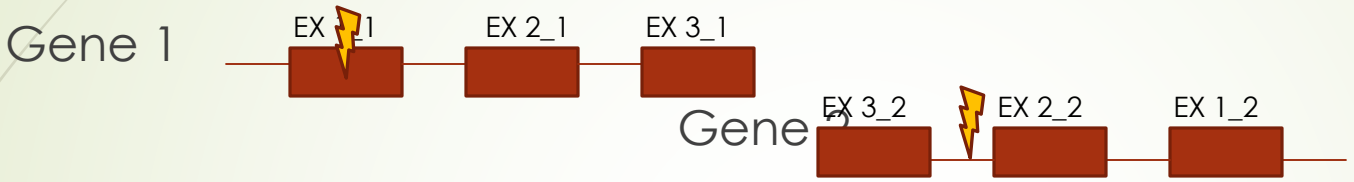
- mRNA



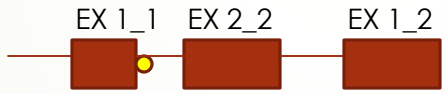
Common case: on the mRNA the breakpoints are at the exon's boundaries whereas on the DNA they are within the introns.

Example of gene fusion in mRNA (2)

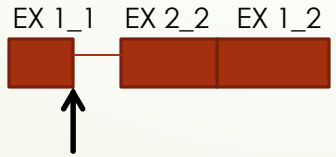
- DNA



- Fused DNA



- mRNA



Less common case: on the mRNA the breakpoints are inside an exon (in Gene 1) and within an intron of the second gene, as it happens in the DNA because of the loss of a splicing donor in Gene 1

Computational Issues

- Alternative splicing and chimerism makes analysis of RNA expression or alteration much more complex
- The problem is to map the reads on the correct location on the genome avoiding mismatches and, vice versa, multiple matching
- The identification of alternative splicing and chimerism it is very hard because splicing and fusion points are not known
- In particular, fusion breakpoints are often not-canonical, i.e. genes can be broken inside exons, keeping subsequences of them and loosing the others. Moreover, the genes involved in the fusion can be, often, located in different chromosomes and thus far away from each others.