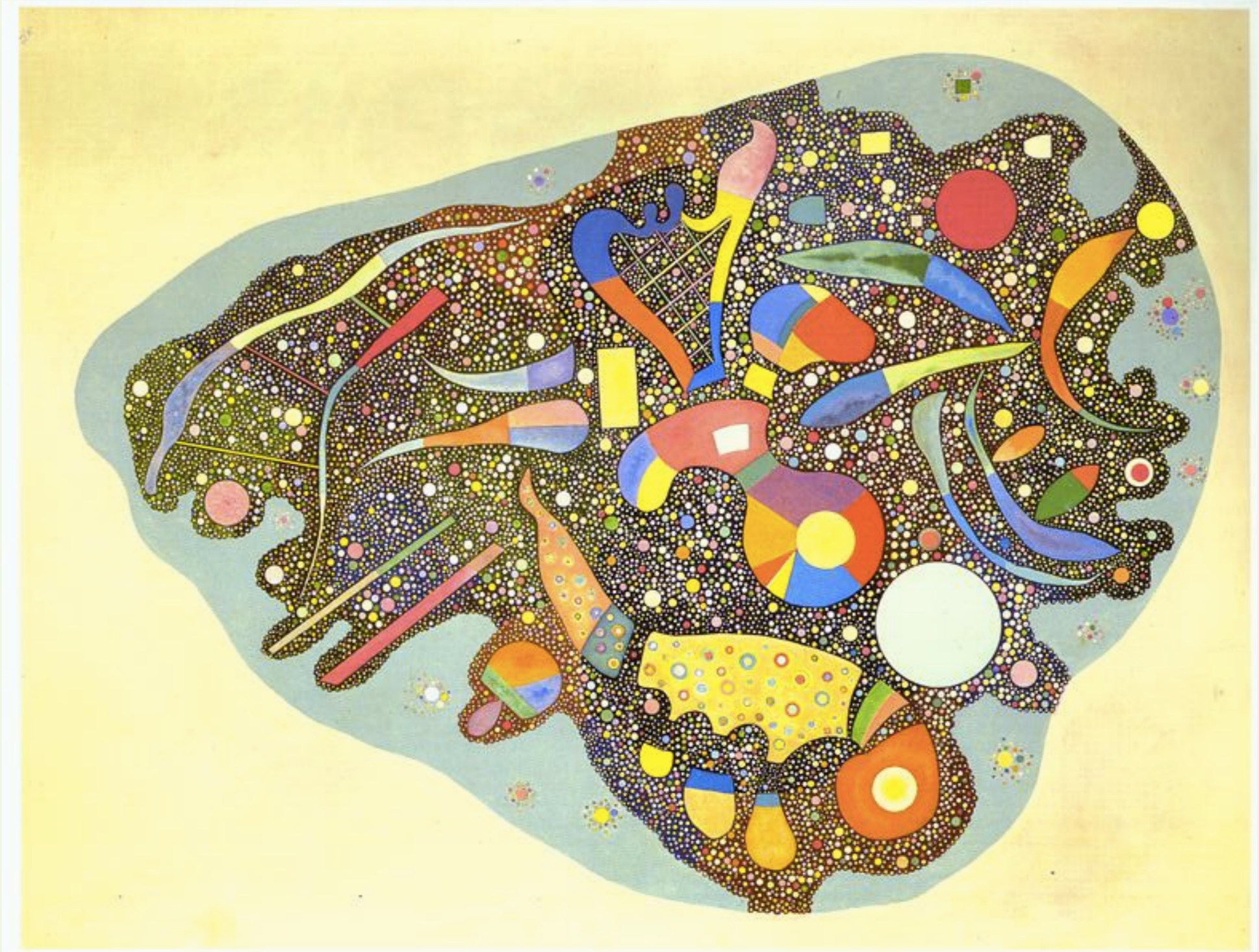
Stat540
Molecular Biology Introduction



https://commons.wikimedia.org/wiki/File:Colorful_Ensemble.jpg

The audience

Thomas Rowlandson [Public domain], via Wikimedia Commons

Molecular biology is a huge huge huge field

Molecular biology is a huge huge huge field

In Bioinformatics, it is important to know:

Molecular biology is a huge huge huge field

In Bioinformatics, it is important to know:

- What you are dealing with,

Molecular biology is a huge huge huge field

In Bioinformatics, it is important to know:

- What you are dealing with,
- What you need to learn,

Molecular biology is a huge huge huge field

In Bioinformatics, it is important to know:

- What you are dealing with,
- What you need to learn,
- What to look for,

Molecular biology is a huge huge huge field

In Bioinformatics, it is important to know:

- What you are dealing with,
- What you need to learn,
- What to look for,
- Where to find it.

Molecular biology is a huge huge huge field

Terminology matters

Molecular biology is a huge huge huge field

Terminology matters

- Don’t panic, know that it takes time.

Molecular biology is a huge huge huge field

Terminology matters

- Don’t panic, know that it takes time.
- Use wikipedia.

Molecular biology is a huge huge huge field

Terminology matters

- Don’t panic, know that it takes time.
- Use wikipedia.
- Have a reference book (I like Albert’s Molecular Biology)

Molecular biology is a huge huge huge field

Terminology matters

- **Don’t panic, know that it takes time.**
- **Use wikipedia.**
- **Have a reference book (I like Albert’s Molecular Biology)**
- **Ask your biology questions from your fellow students with biology background. This is the beauty of this field. They will probably ask you later about stat and cs or math too.**

Molecular biology is a huge huge huge field

Terminology matters

- **Don’t panic, know that it takes time.**
- **Use wikipedia.**
- **Have a reference book (I like Albert’s Molecular Biology)**
- **Ask your biology questions from your fellow students with biology background. This is the beauty of this field. They will probably ask you later about stat and cs or math too.**
- **Study it seriously and put time on it, if you wish to continue in Bioinformatics.**

Molecular biology is a huge huge huge field

Terminology matters

- **Don’t panic, know that it takes time.**
- **Use wikipedia.**
- **Have a reference book (I like Albert’s Molecular Biology)**
- **Ask your biology questions from your fellow students with biology background. This is the beauty of this field. They will probably ask you later about stat and cs or math too.**
- **Study it seriously and put time on it, if you wish to continue in Bioinformatics.**

The point: **you will probably never be as good as a biochemist or a biologist in biology. But you will be a good Bioinformatician if you learn more biology and get to know the terminology.**

- The cell
- The molecules
- The data

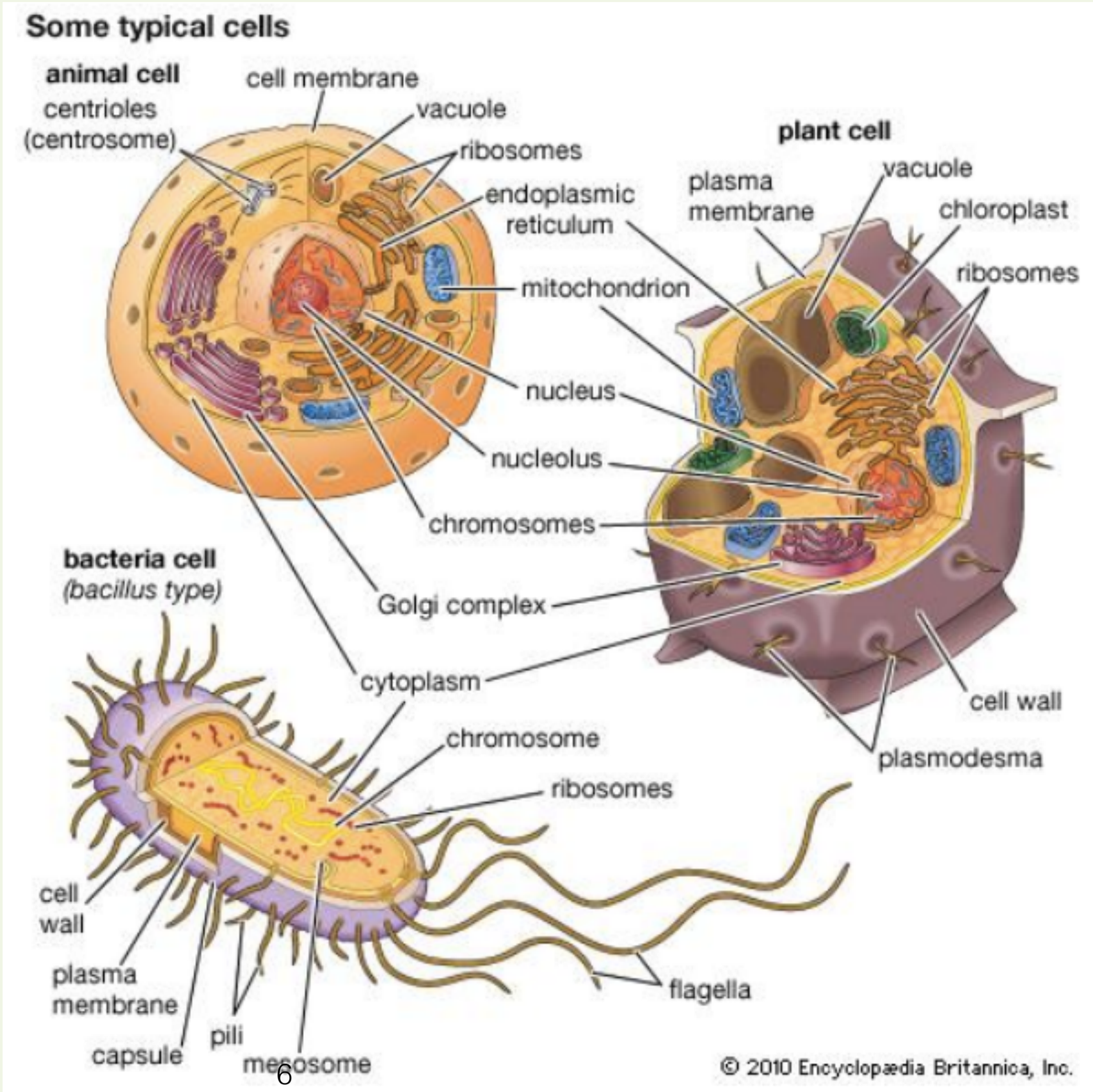
The cell

Basic structural, functional and biological unit of all known living organisms

They can replicate independently

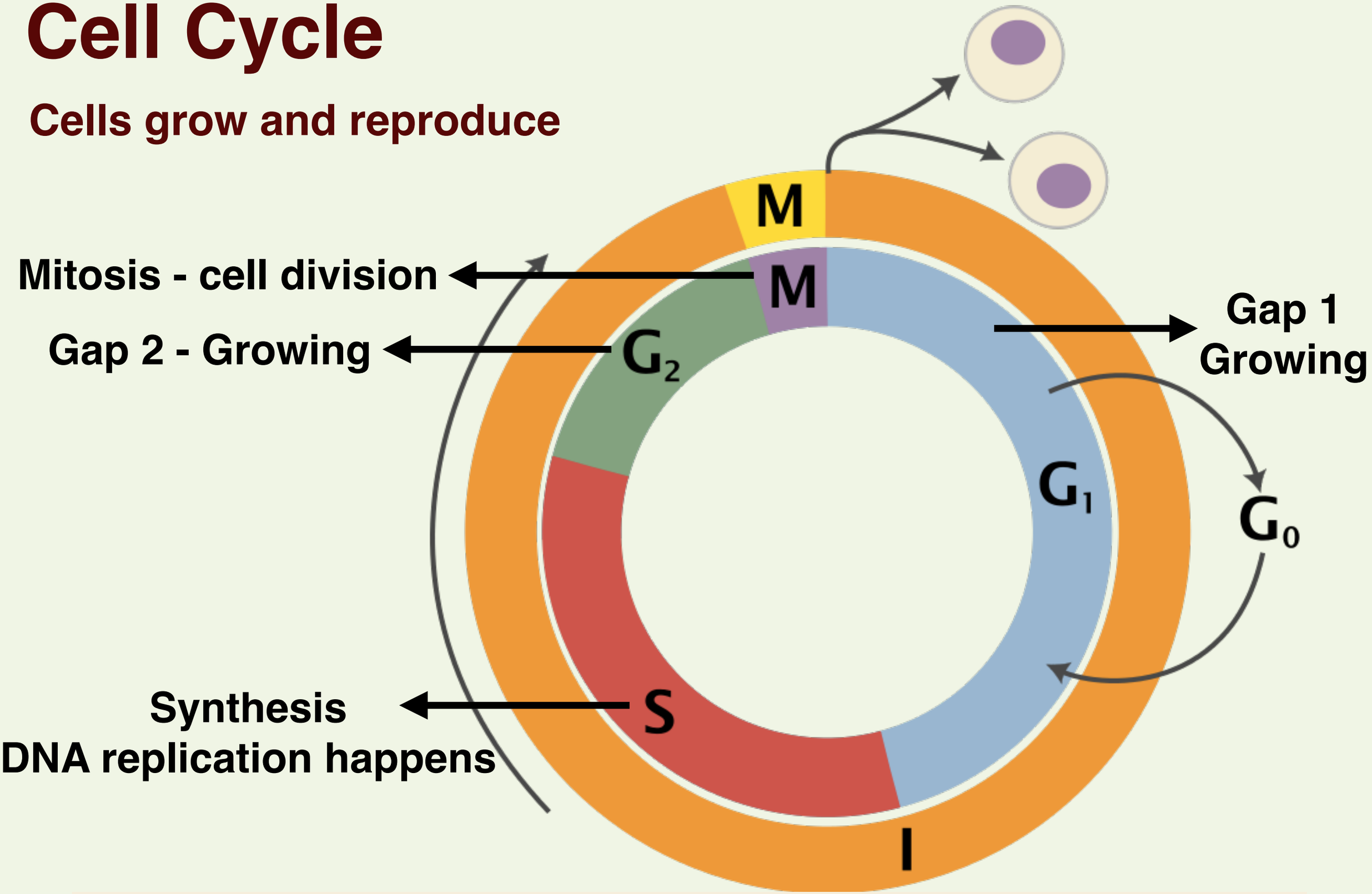
They contain organelles and biomolecules, such as proteins and nucleic acids.

You can refresh your high school cell biology knowledge by looking at wikipedia page on cell: https://en.wikipedia.org/wiki/Cell_(biology)



Cell Cycle

Cells grow and reproduce



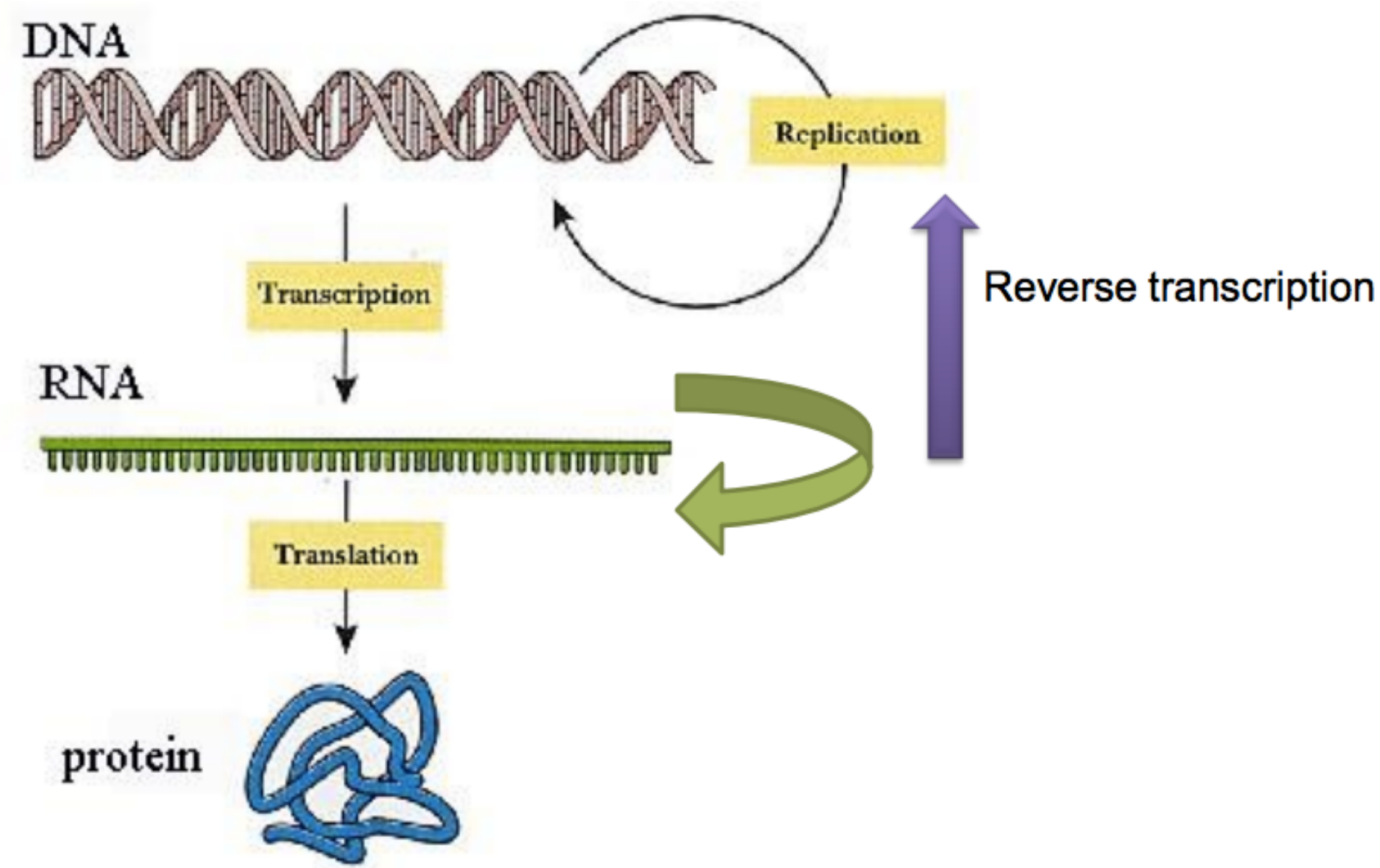
Nice read on cell cycle: http://www.ncbi.nlm.nih.gov/books/NBK9876/

What are the molecules *and* how do they fit in the picture?

What are the molecules *and* how do they fit in the picture?

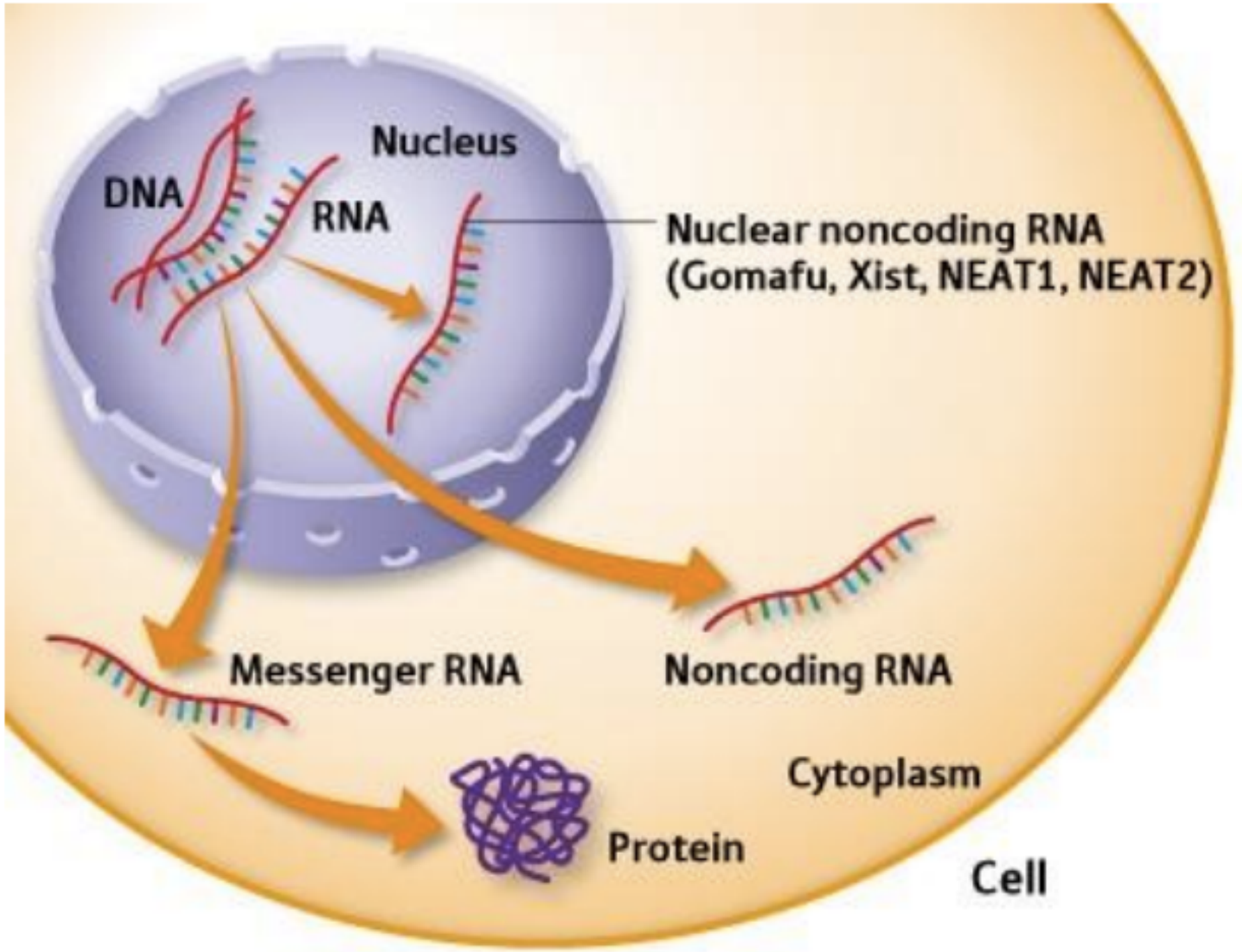
DNA & RNA
Protein

The **Central Dogma**: Flow of heritable information



http://www.angelfire.com/dc/apgenetics/central.dogma.jpg

The **Central Dogma**: Flow of heritable information

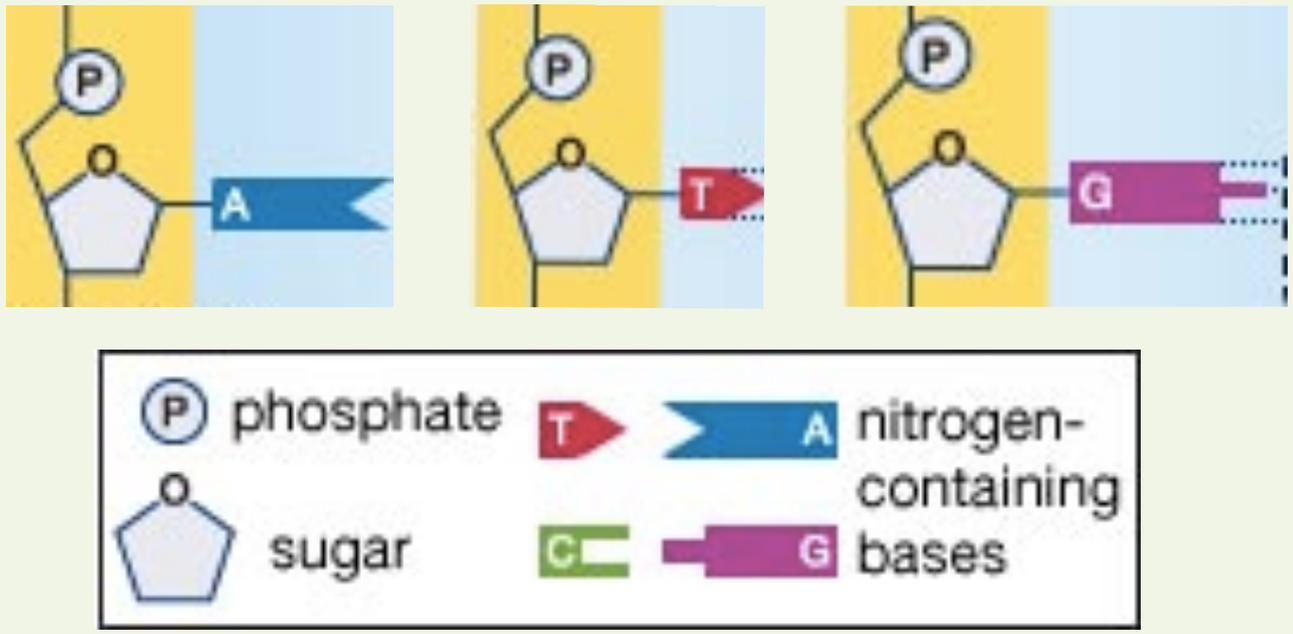


http://www.rikenresearch.riken.jp/eng/frontline/5920

DNA and RNA

DNA and RNA

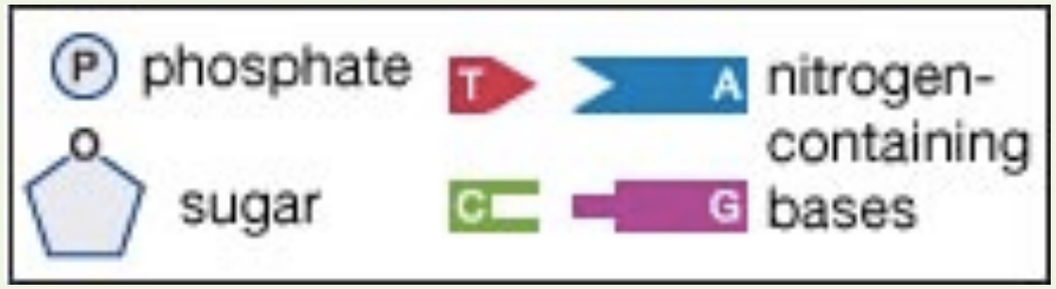
- **Nucleotide**



DNA and RNA

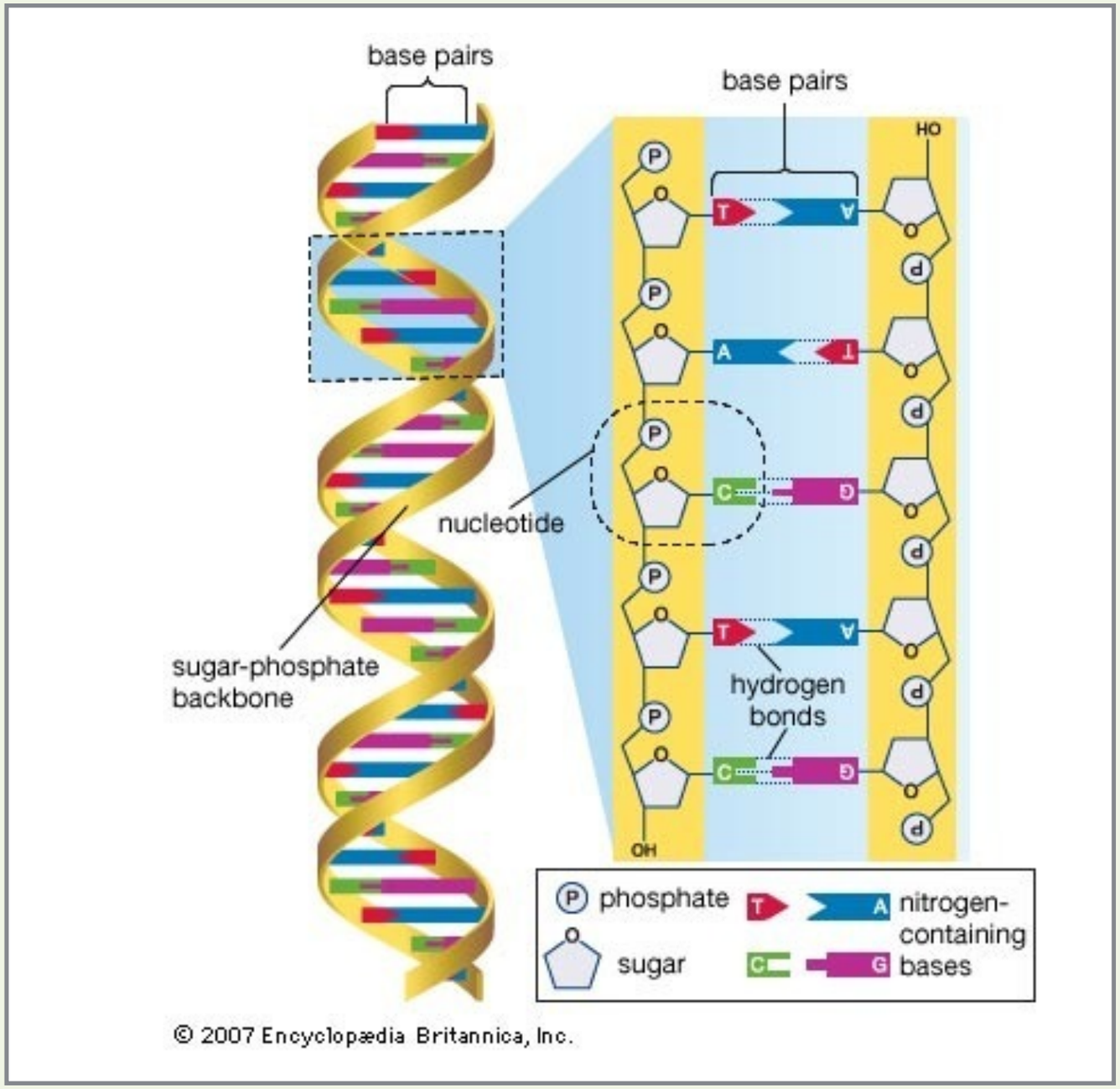
- **Nucleotide**

- **Nucleic acid**



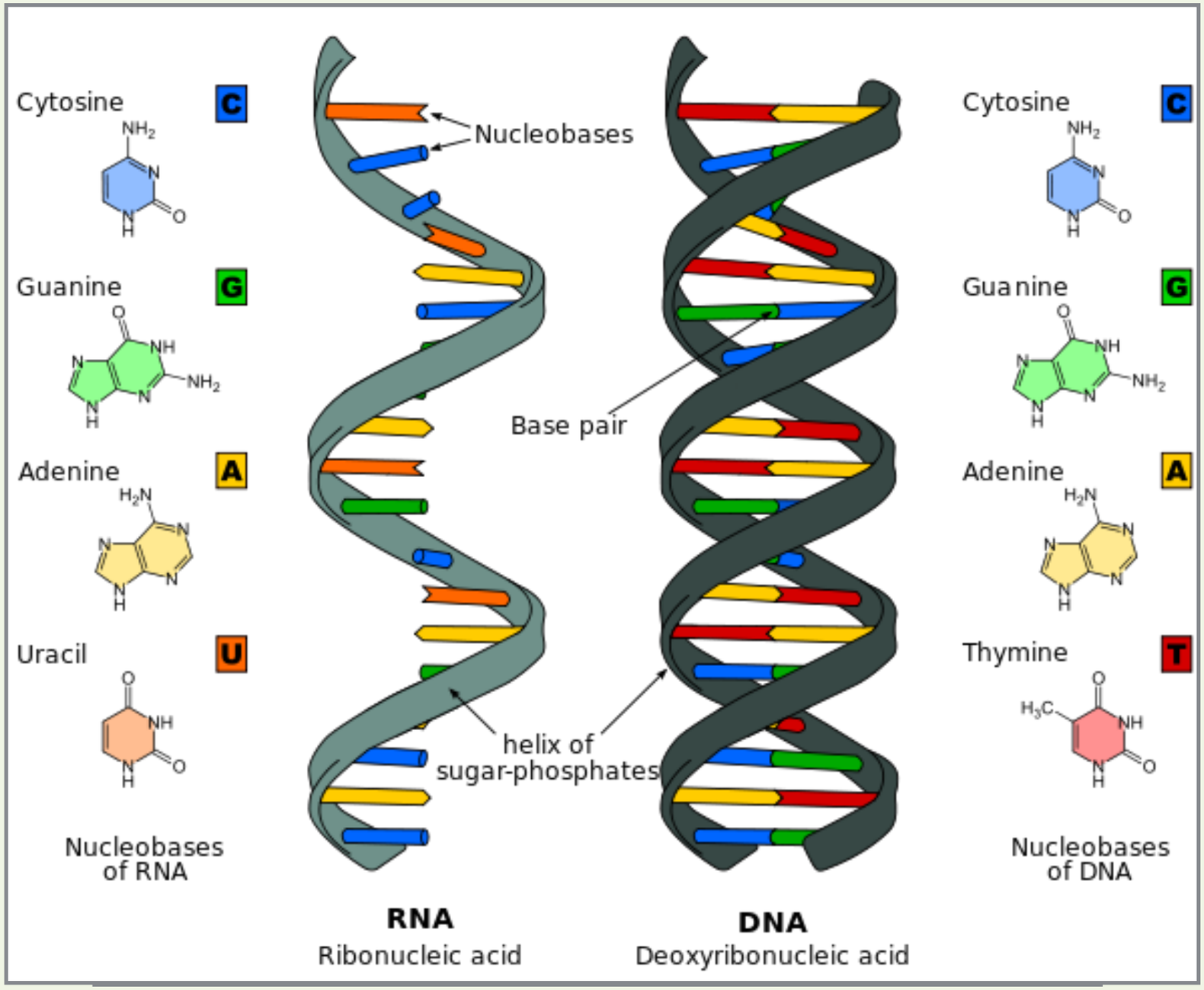
DNA and RNA

- **Nucleotide**
- **Nucleic acid**
- **DeoxyriboNucleic Acid**

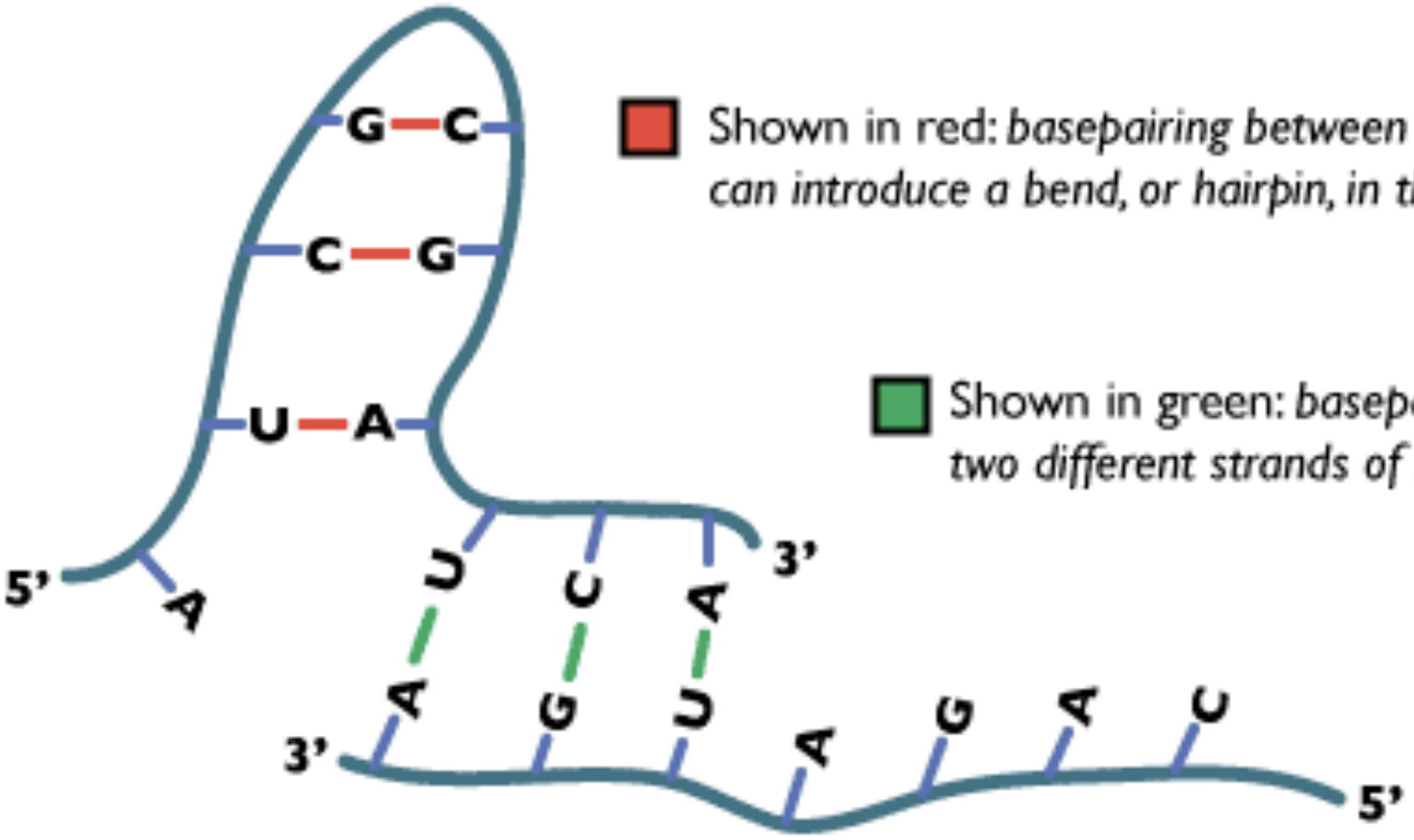


DNA and RNA

- **Nucleotide**
- **Nucleic acid**
- **DeoxyriboNucleic Acid**
- **RiboNucleic Acid**



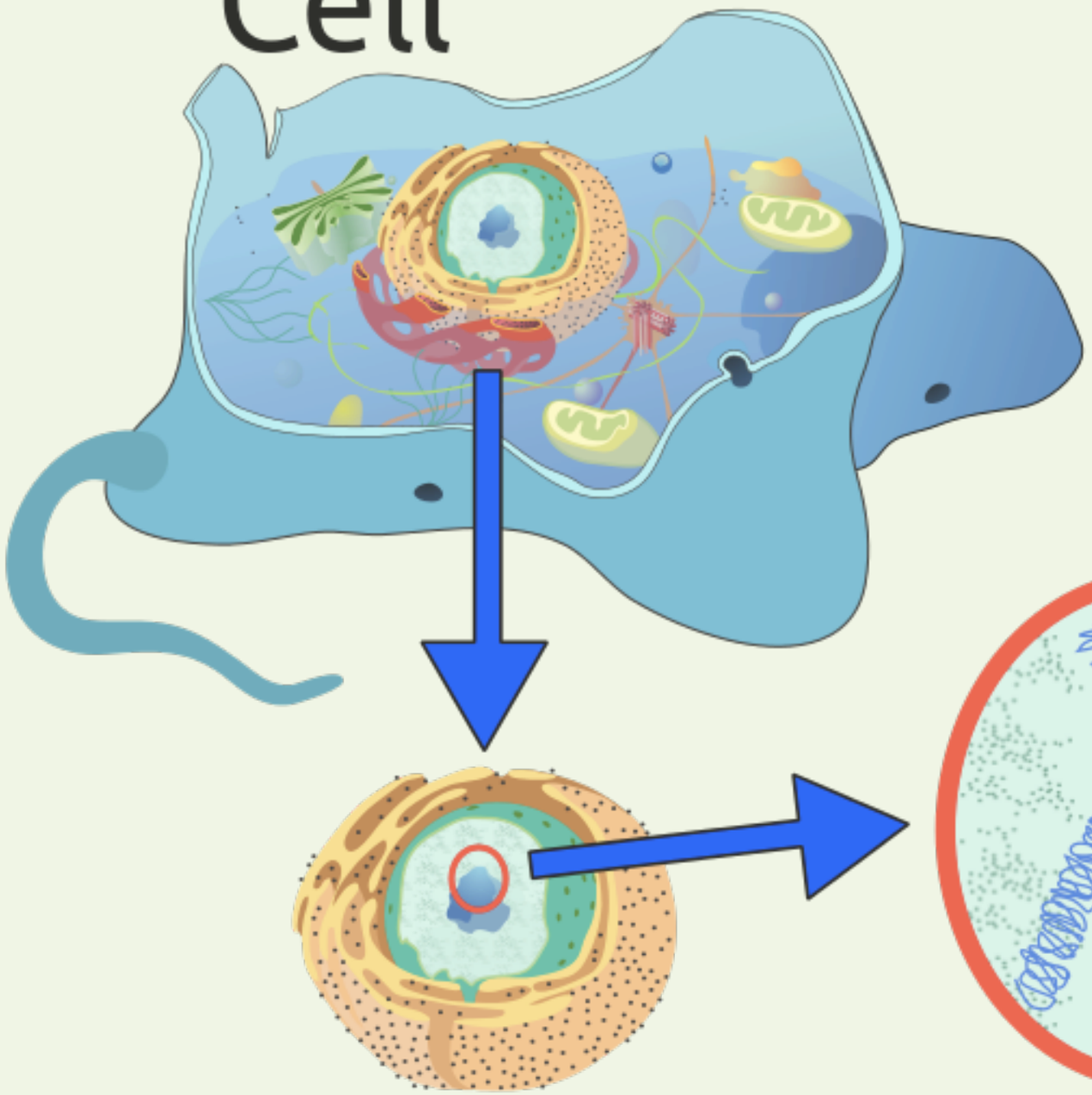
RNA folding



Shown in red: *basepairing between bases on the same strand can introduce a bend, or hairpin, in the molecule.*

Shown in green: *basepairing can also occur between two different strands of RNA.*

Cell

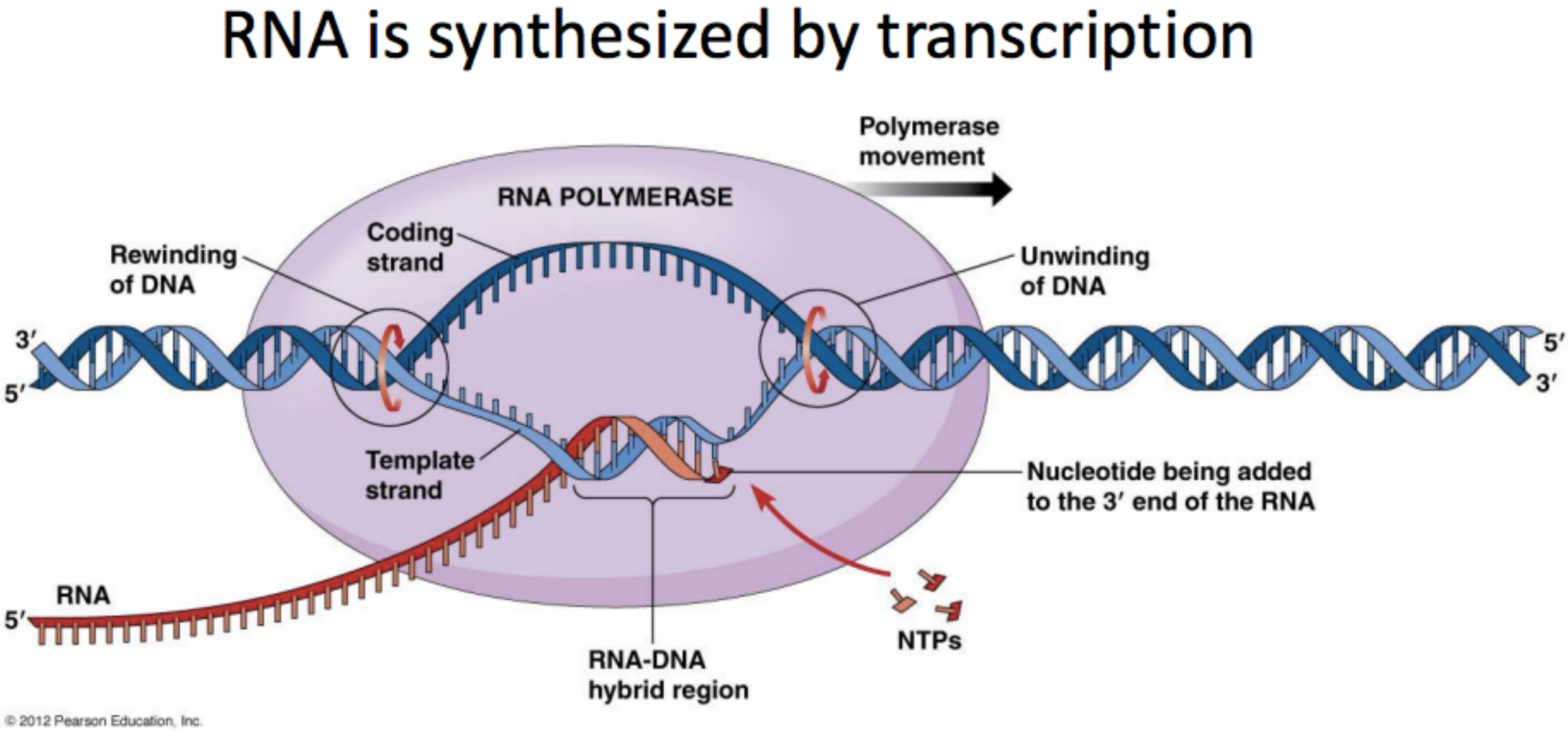


DNA

Nucleus

Chromosome

Transcription

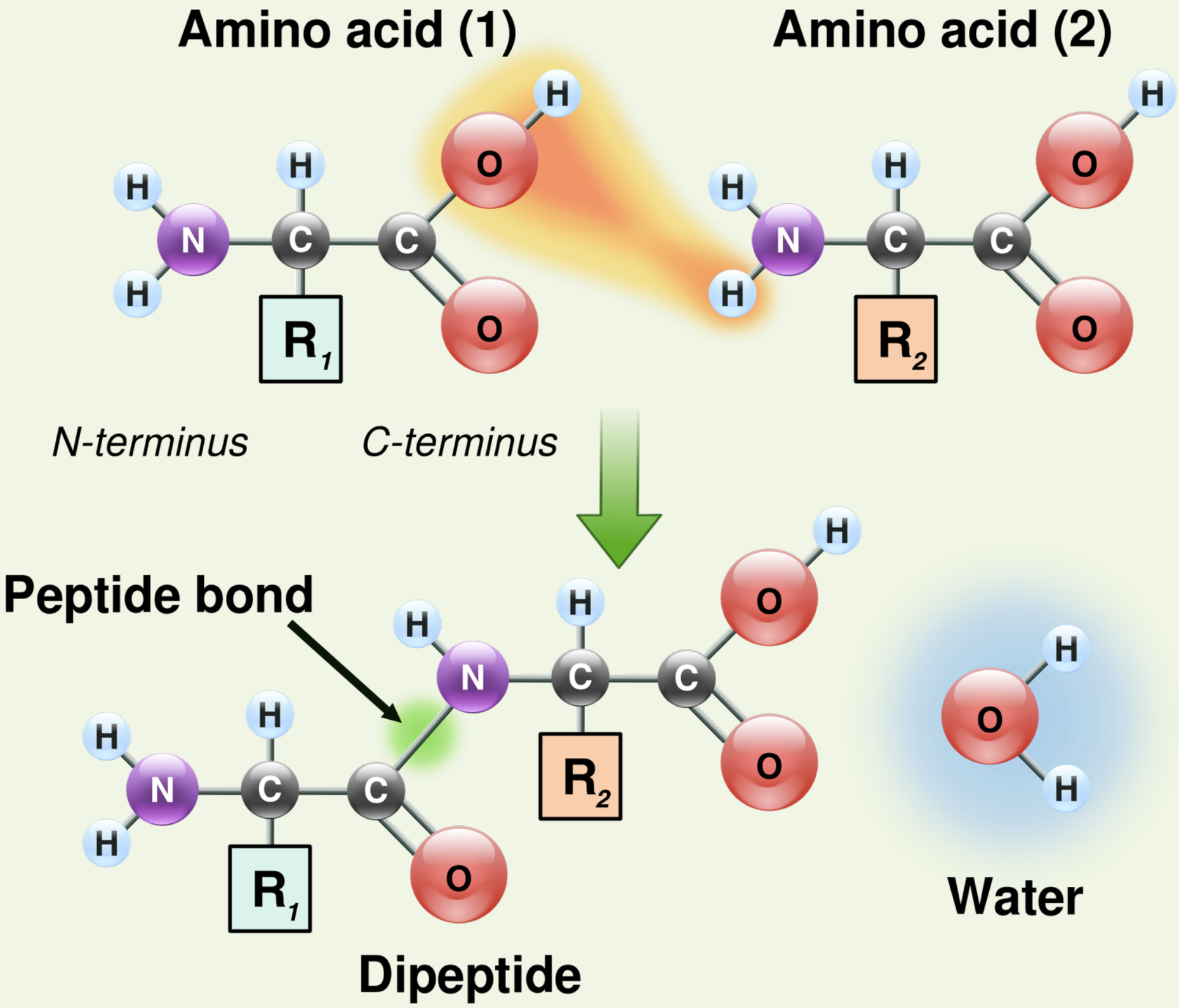


Transcription proceeds from 5’ to 3’ on the RNA, *i.e.* from 3' to 5' on the **complementary** template DNA

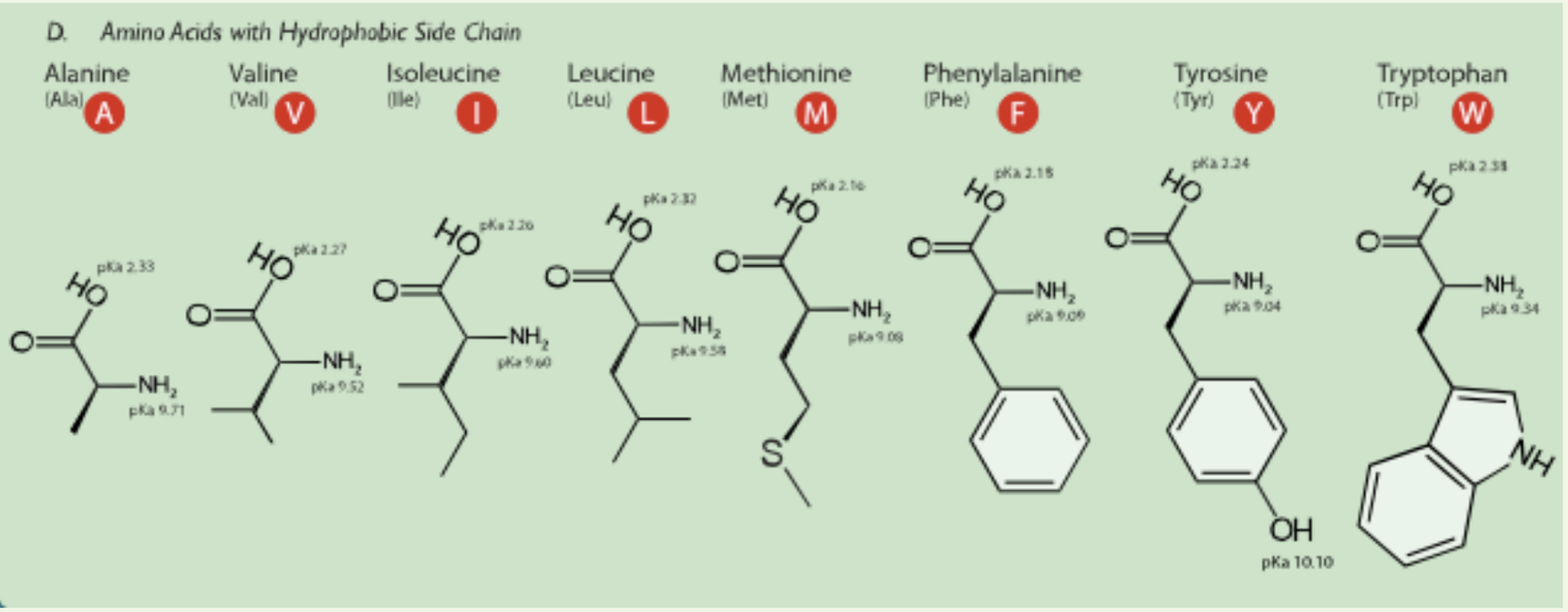
http://www.rikenresearch.riken.jp/eng/frontline/5920

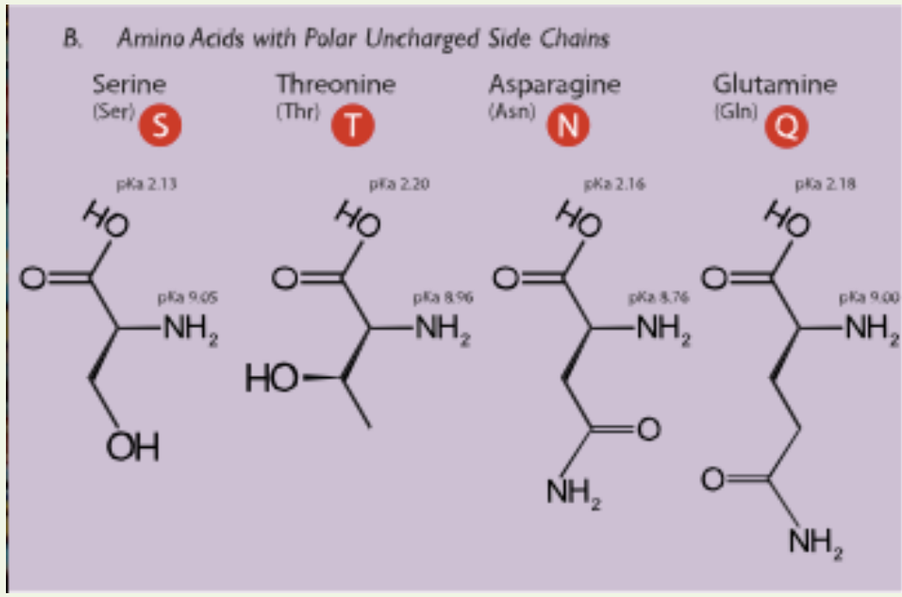
Proteins

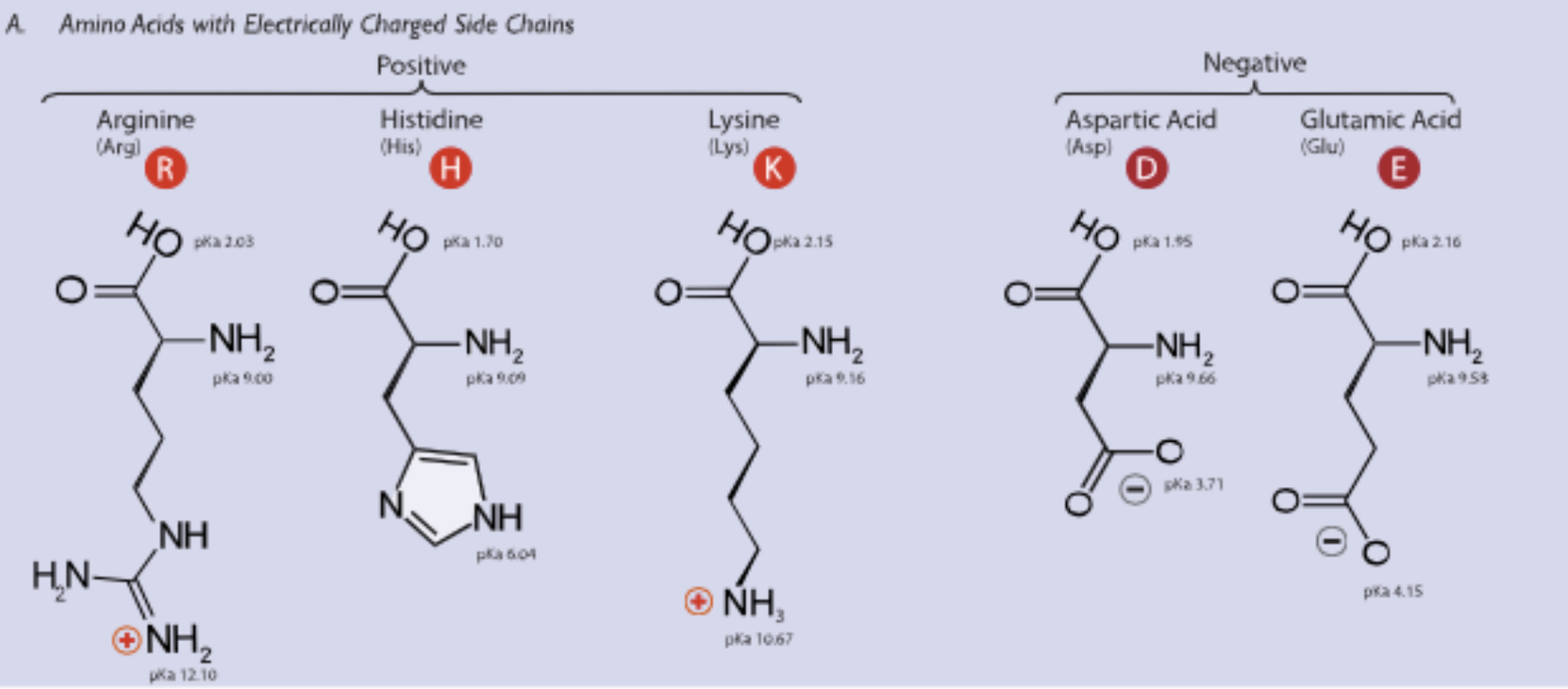
sequences of amino acids with side chains

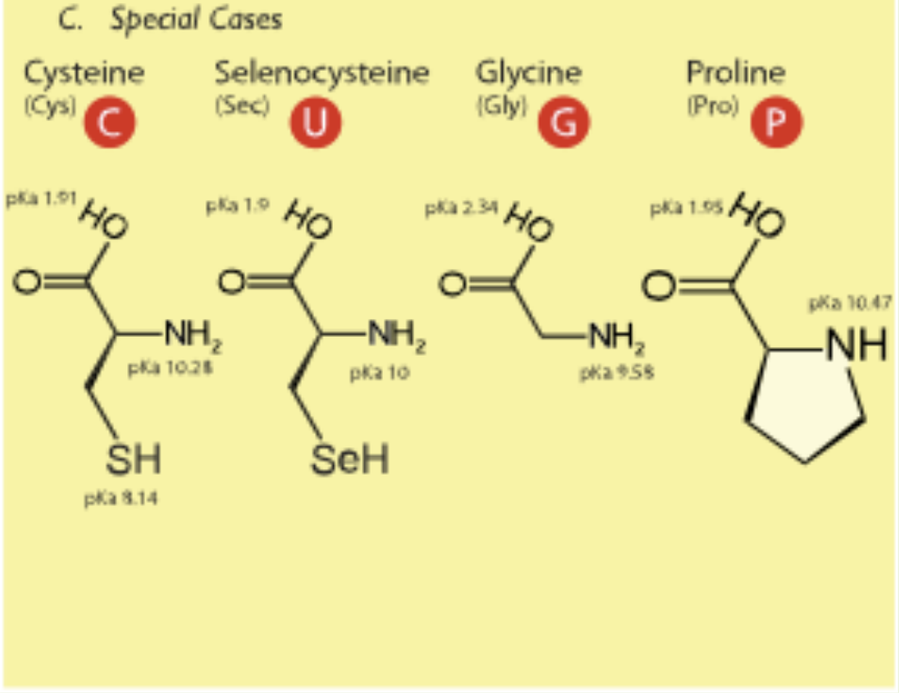


The 21 amino acids in eukaryote

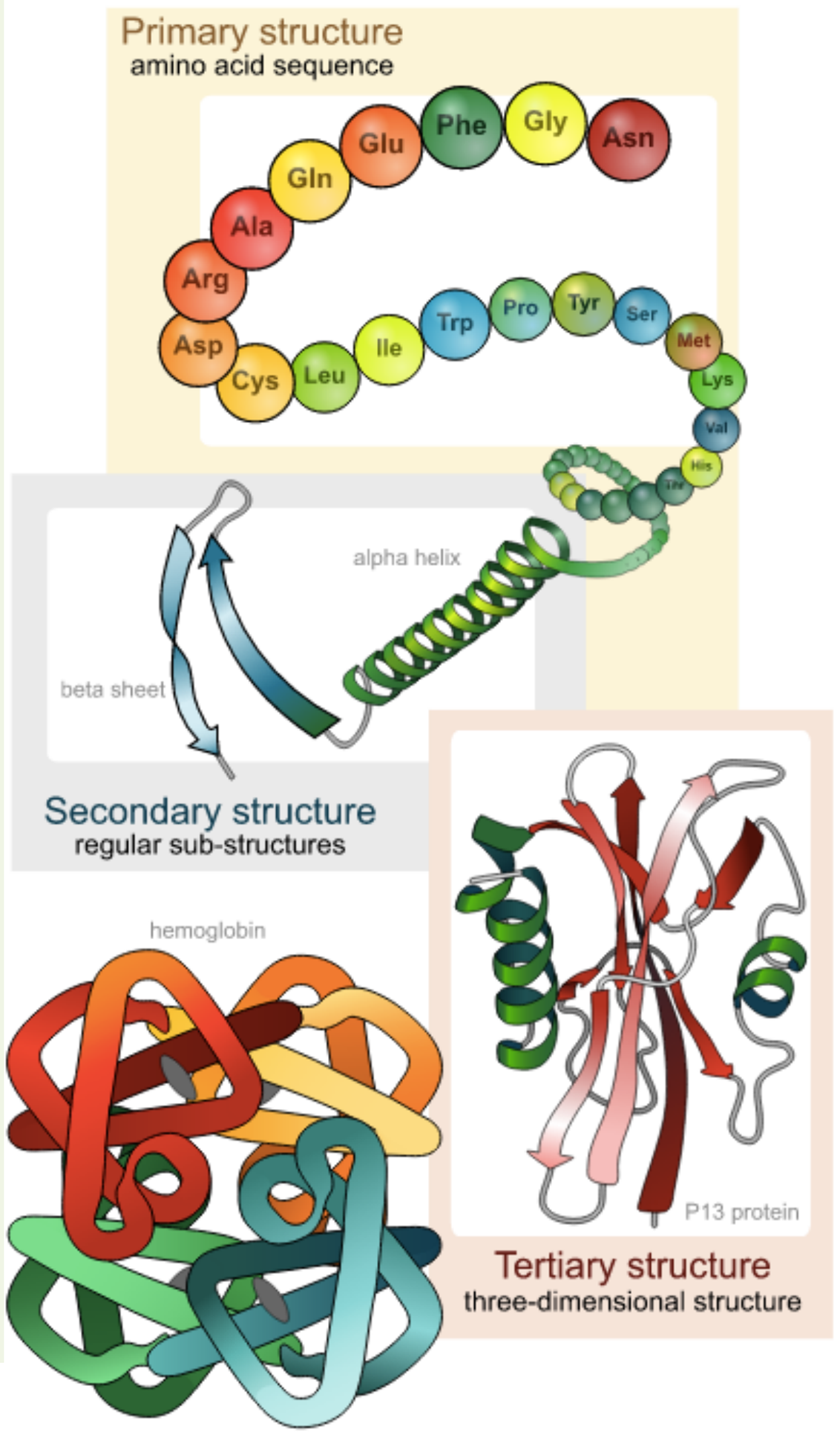






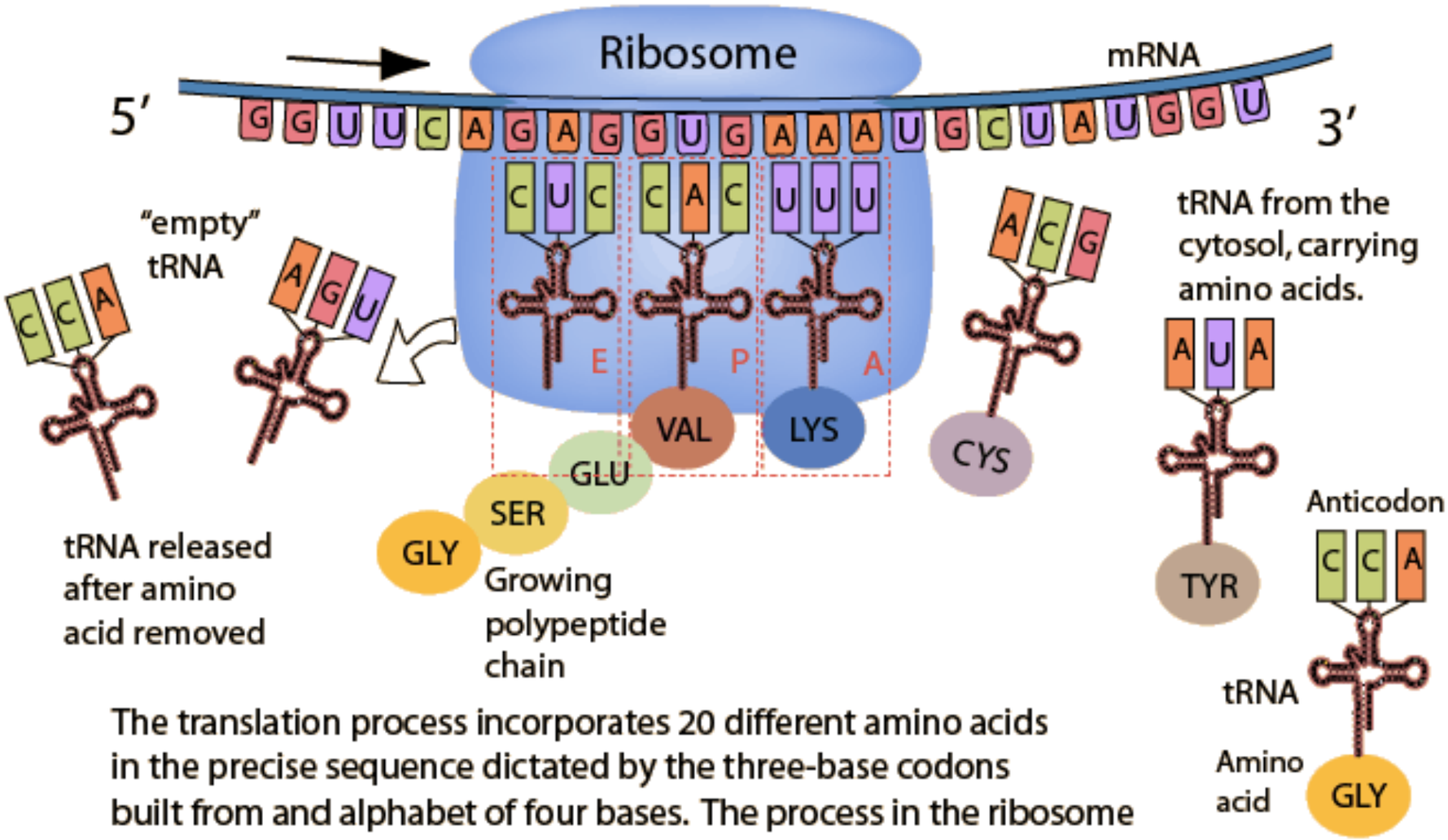


Protein structure at different levels



Proteins - translation

Protein is synthesized by translation



The translation process incorporates 20 different amino acids in the precise sequence dictated by the three-base codons built from and alphabet of four bases. The process in the ribosome builds the polypeptide chains tha will become proteins.

http://hyperphysics.phy-astr.gsu.edu/hbase/organic/imgorg/translation2.gif

Proteins - translation

Translation is interpreted via Codon

<u>Codon Degeneracy</u>:
Many distinctive codons can redundantly map onto the same amino acid

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phe; UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr; UAA: Stop; UAG: Stop	UGU, UGC: Cys; UGA: Stop; UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His; CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile; AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn; AAA, AAG: Lys	AGU, AGC: Ser; AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp; GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

http://www.mun.ca/biology/scarr/MGA2_03-20.html

What is a gene?

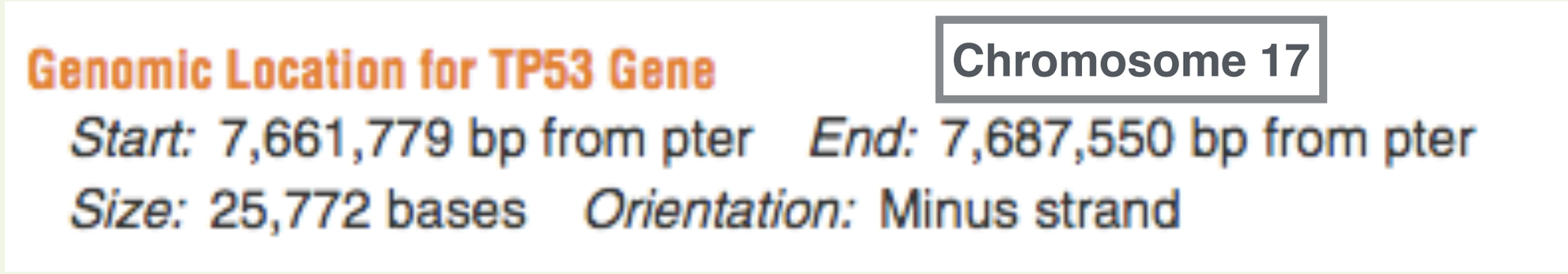
The gene

A gene is a locus (or region) of DNA that encodes a functional RNA or protein product and is the molecular unit of heredity

The gene

A gene is a locus (or region) of DNA that encodes a functional RNA or protein product and is the molecular unit of heredity

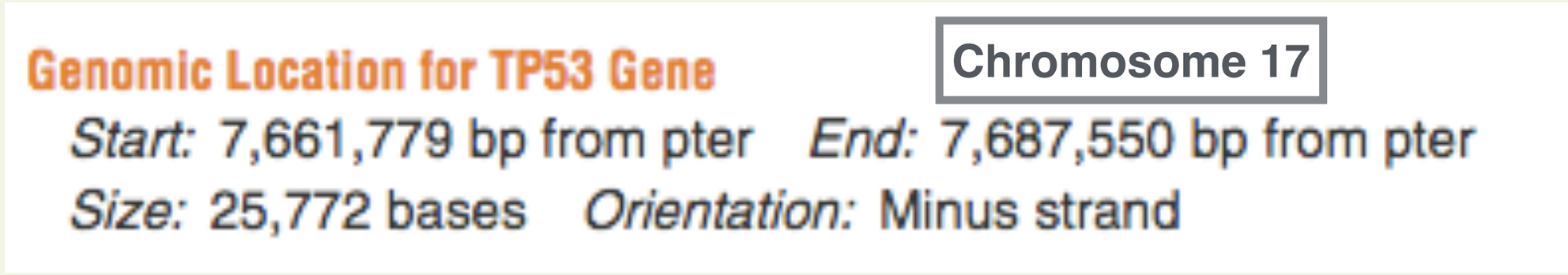
It is a DNA strand and is identified by its location on the Genome



The gene

A gene is a locus (or region) of DNA that encodes a functional RNA or protein product and is the molecular unit of heredity

It is a DNA strand and is identified by its location on the Genome



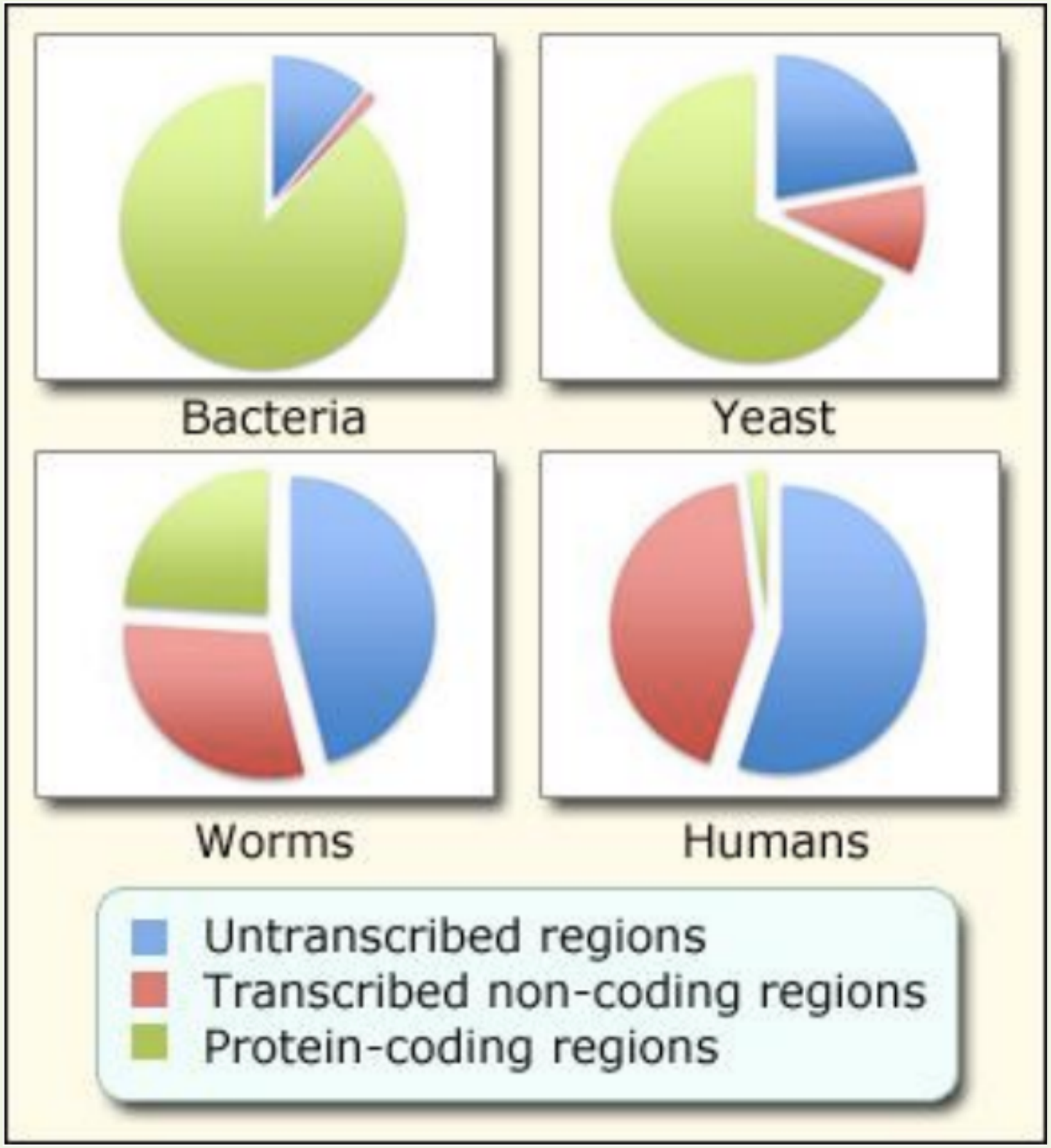
Genes could be protein coding or non-coding RNA. Non-coding RNA are still functional.

The gene

Genes are only a fraction of the genome

Human protein coding gene count:
~20000 - 30000

Yeast protein coding gene count:
~5000 - 6000

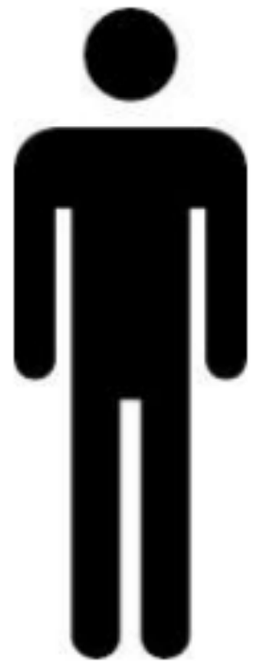


The gene

However, the noncoding DNA is *probably* not *all junk* DNA

Human and Fugu fish share roughly the same genes and regulatory sequences, but the Fugu fish has a more condensed form with little junk region.

VS.



Fugu fish: 400 million base

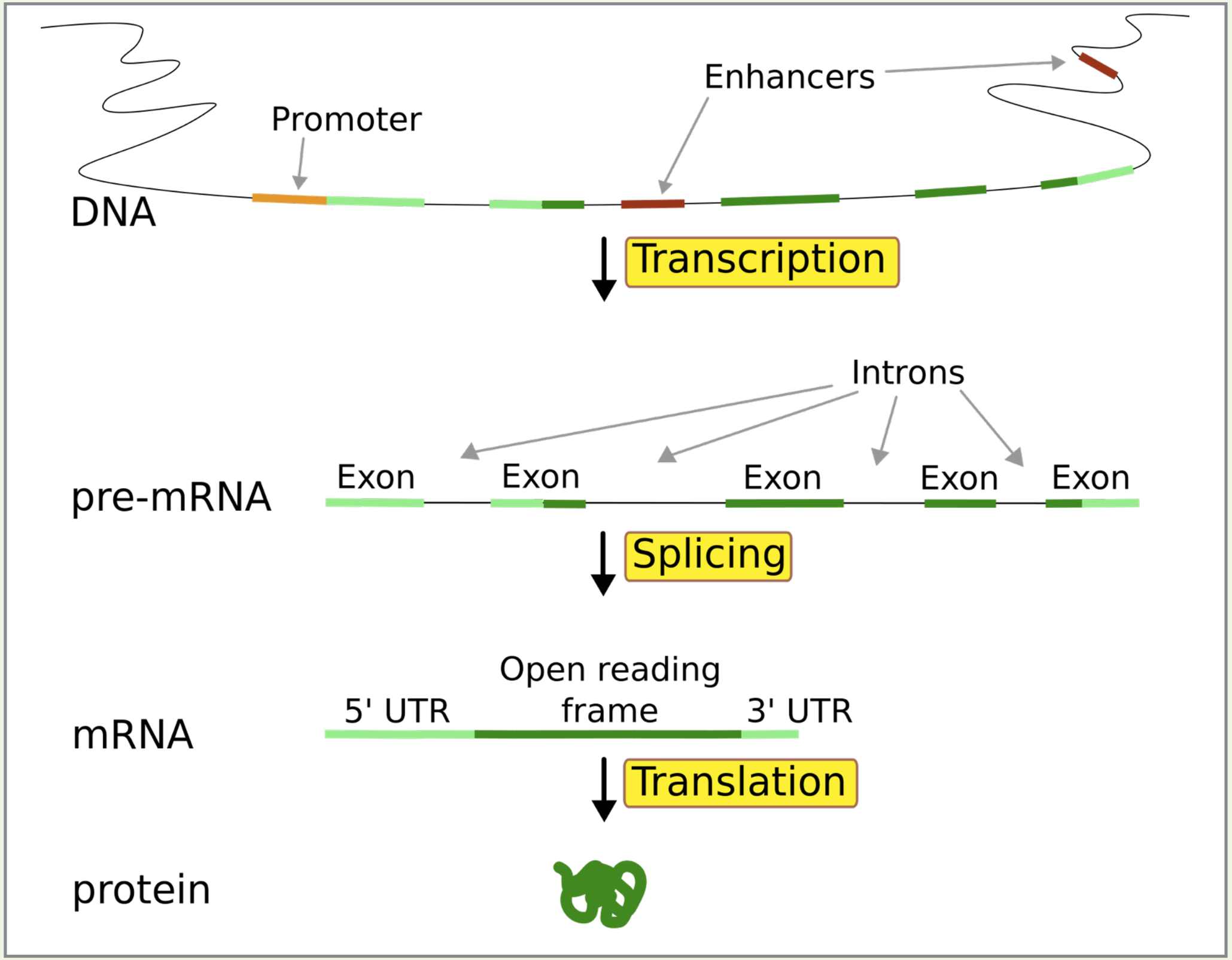
Human: 3 billion base

But really,
What is a gene?

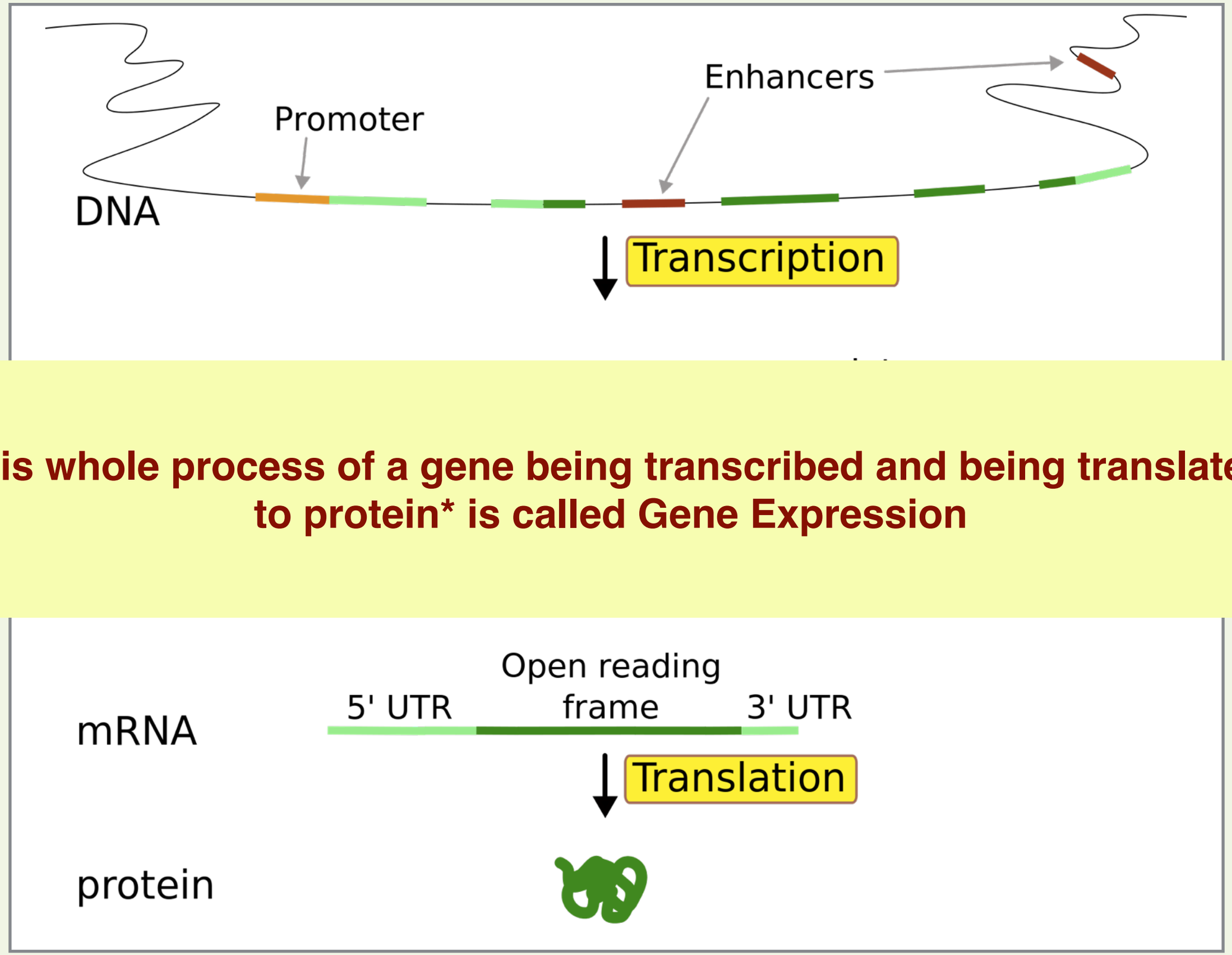
But really,
What is a gene?

Transcription
Translation
Exon
Intron
Gene Expression

The gene


Enhancers
Promoter
DNA
Transcription
Introns
Exon
Exon
Exon
Exon
Exon
pre-mRNA
Splicing
Open reading
5' UTR
frame
3' UTR
mRNA
Translation
protein

The gene



This whole process of a gene being transcribed and being translated to protein* is called Gene Expression

Gene Expression

All the cells of a multicellular organism are copied from one single Zygote cell and therefore share the same DNA.

Gene Expression

All the cells of a multicellular organism are copied from one single Zygote cell and therefore share the same DNA.

Gene expression is different in different cell types and also based on various conditions for the same cell type.

Gene Expression

All the cells of a multicellular organism are copied from one single Zygote cell and therefore share the same DNA.

Gene expression is different in different cell types and also based on various conditions for the same cell type.

The expression level of each gene (or if it is expressed at all), in each cell is regulated by various cellular mechanisms affected by the cell state or signals from the environment.

Gene Expression

All the cells of a multicellular organism are copied from one single Zygote cell and therefore share the same DNA.

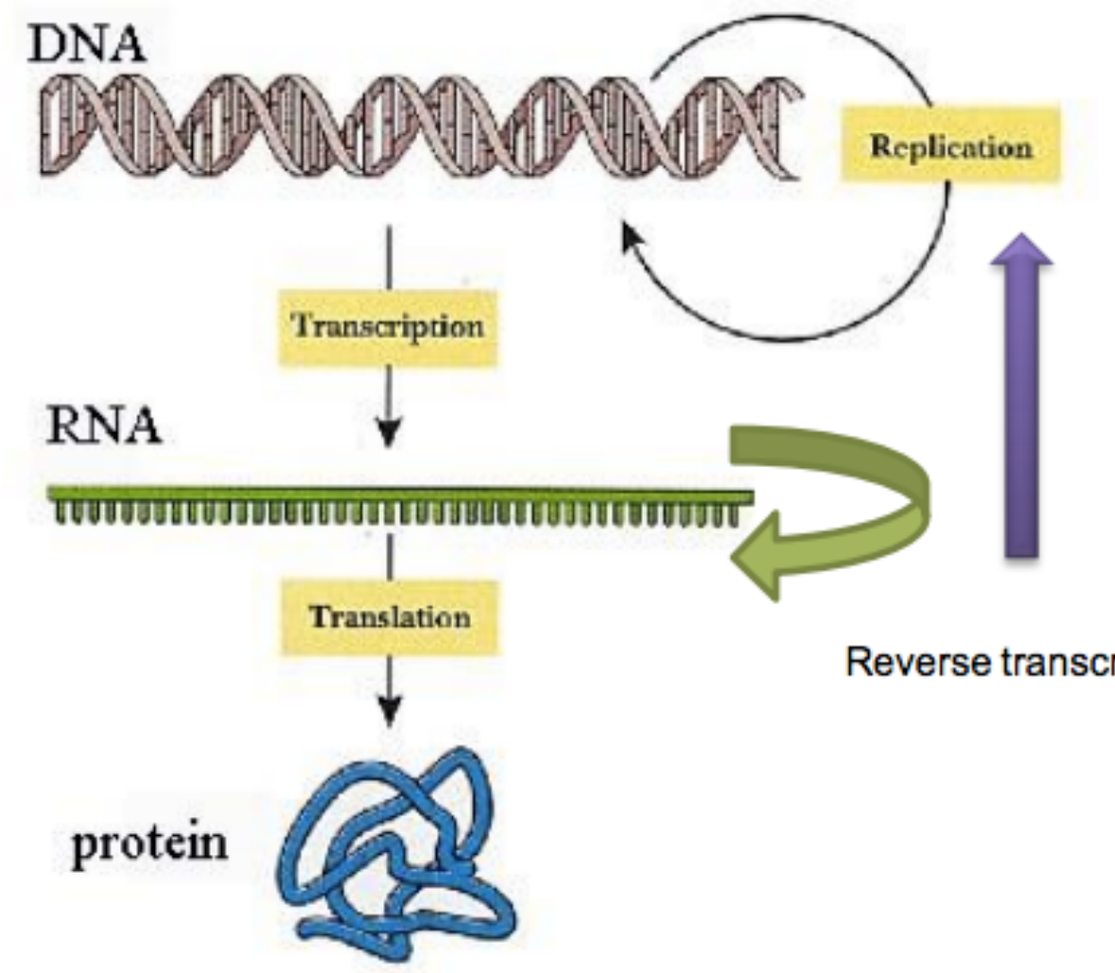
Gene expression is different in different cell types and also based on various conditions for the same cell type.

The expression level of each gene (or if it is expressed at all), in each cell is regulated by various cellular mechanisms affected by the cell state or signals from the environment.

We have different populations of RNA and Protein in each cell, based on its type or the condition it comes from

The **Central Dogma**: Summary

http://www.angelfire.com/dc/apgenetics/central.dogma.jpg



- The inherited information is stored in DNA
- RNA carries the inheritance information from nucleus to cytoplasm
- The inheritance information is expressed in proteins with RNA being the mediator
- The proteins come in versatile structures equipped with diverse functions to meet cellular demand

Example for the flow of heritable information, quantitatively:
in human,
DNA: 3 billion bp (haploid) → RNA transcripts: 62.1-74.7% of DNA transcribed
(Djebali, Davis et al, 2012) → Proteins: about 20,000-25,000

The data

The data

Why do we collect molecular biology data?

What can we measure?

How does it look like?

It is just data. Why should I know Biology? Why should I care?

The data

Why do we collect molecular biology data?

Usually, to learn about the genetic causes of diseases and cellular mechanisms. In the disease case, we can compare the genetic material of healthy and affected people and find the common differences.

What can we measure?

How does it look like?

It is just data. Why should I know Biology? Why should I care?

The data

Why do we collect molecular biology data?

Usually, to learn about the genetic causes of diseases and cellular mechanisms. In the disease case, we can compare the genetic material of healthy and affected people and find the common differences.

What can we measure?

We can measure the abundance of proteins and RNA. Also, we can screen the complete DNA of an organism or some parts of it.

How does it look like?

It is just data. Why should I know Biology? Why should I care?

The data

Why do we collect molecular biology data?

Usually, to learn about the genetic causes of diseases and cellular mechanisms. In the disease case, we can compare the genetic material of healthy and affected people and find the common differences.

What can we measure?

We can measure the abundance of proteins and RNA. Also, we can screen the complete DNA of an organism or some parts of it.

How does it look like?

It comes in variety of shapes, depends on the data. The most common feature might be that, genetic data is BIG. You usually need computational and statistical techniques to learn biological facts from it.

It is just data. Why should I know Biology? Why should I care?

The data

Why do we collect molecular biology data?

Usually, to learn about the genetic causes of diseases and cellular mechanisms. In the disease case, we can compare the genetic material of healthy and affected people and find the common differences.

What can we measure?

We can measure the abundance of proteins and RNA. Also, we can screen the complete DNA of an organism or some parts of it.

How does it look like?

It comes in variety of shapes, depends on the data. The most common feature might be that, genetic data is BIG. You usually need computational and statistical techniques to learn biological facts from it.

It is just data. Why should I know Biology? Why should I care?

It is not just data. It is usually, noisy data. You might need to apply several steps of pre-processing before you can even study the data. You need to know about the nature of the data to be able to pick the right pre-processing methods, or invent ones. Also, you usually have to choose your computational/statistical methods to analyze the data, for that, you must know what is the biological question you want to answer.

The DNA data

Genomics data: Screening the genome of an organism.
Examples: DNA sequencing, Exom sequencing, SNP arrays

This kind of study is done to learn about genetic variation of an organism and learning about disease causing mutations by comparing *genomes* of different individuals. Genomes could be compared at a whole level or only at particular nucleotides of interest (SNP arrays), or only at the protein coding parts (exom sequencing)

The RNA data

The RNA data

Transcriptomics data: Screening the RNA of a cell sample.
Examples: RNA sequencing, microarrays

This data is known as gene expression data or gene expression profiling. While genomic data helps to learn about genetic variations and their relations to phenotypes, gene expression data helps to find the genes which are causal for or affected by a condition/disease by comparing their RNA levels in the sample.

The RNA data

Transcriptomics data: Screening the RNA of a cell sample.
Examples: RNA sequencing, microarrays

This data is known as gene expression data or gene expression profiling. While genomic data helps to learn about genetic variations and their relations to phenotypes, gene expression data helps to find the genes which are causal for or affected by a condition/disease by comparing their RNA levels in the sample.

Shortcoming of RNA data

Biological shortcoming: The RNA abundance of a gene is a noisy representation of the Protein abundance of the gene. Not all the transcribed RNA's are going to translate to Protein, and a single RNA could be used to generate multiple proteins.

Technical shortcomings: Despite the advances of the technology, both microarray data and RNA-seq data are still noisy

The Portein data

Proteomics data: Screening the abundance of proteins in a sample
Examples: Mass spectrometry

Mass spectrometry screens the abundance of *all* the proteins in a sample.

Some shortcomings: expensive and still very noisy.

The Portein data

Proteomics data: Screening the abundance of proteins in a sample
Examples: Mass spectrometry

Mass spectrometry screens the abundance of *all* the proteins in a sample.

<u>Some shortcomings:</u> expensive and still very noisy.

Protein Interaction data: Screening protein interaction data for a targeted set of proteins.
Examples: Yeast two hybrids, affinity purification coupled to mass spectrometry

<u>Some shortcomings:</u> it could be only done for a selected set of proteins

Note:
Gene regulation slides from last year were not covered in this presentation, please check out the last years slides.

A good reference:

Molecular Biology for Computer Scientists
https://www.cs.princeton.edu/~mona/IntroMaterials/hunter-bio-for-CS.pdf