# CSI5180. Machine Learning for Bioinformatics Applications

Essential **Cellular Biology** (continued)

by

Marcel Turcotte

Version **November 25, 2019**

# Preamble

# Summary

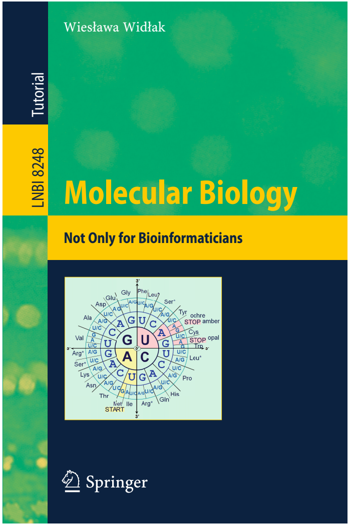
This lecture presents the **central dogma** and the **genetic code**, as well as the **structure macromolecules**. We will also briefly discuss concepts such as the **genome**, the **transcriptome**, the **proteome**, and the various **biological networks**. Throughout the presentation, we will highlight the importance of the concepts for bioinformatics.

**General objective**

- **Describe** the central dogma, transcription, translation, and genetic code.

**Reading**

- Lawrence Hunter, Life and its molecules: A brief introduction, *AI Magazine* **25** (2004), no. 1, 922.
- Wiesława Widłak (2013). **Molecular Biology: Not Only for Bioinformaticians** (Vol. 8248). Springer. Chapters 3, 4, 5, 6, and 9.



Tutorial

LNBI 8248

Wiesława Widłak

Molecular Biology

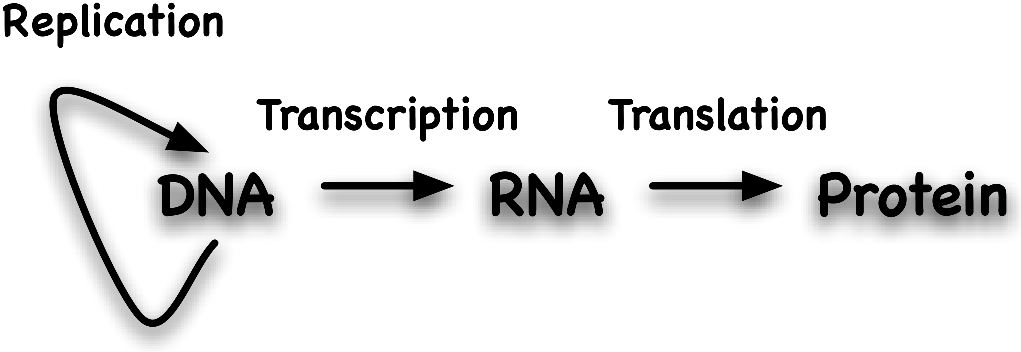
Not Only for Bioinformaticians

Springer

link.springer.com/book/10.1007/978-3-642-45361-8

# Central Dogma

# Central Dogma (1958)



Replication

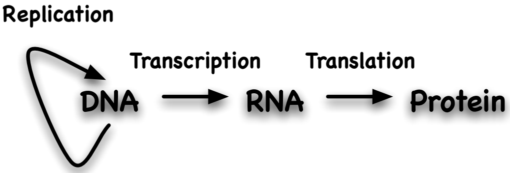
Transcription

Translation

DNA → RNA → Protein

Francis Crick (1958) *Symposium of the Society of Experimental Biology* **12**:138-167.

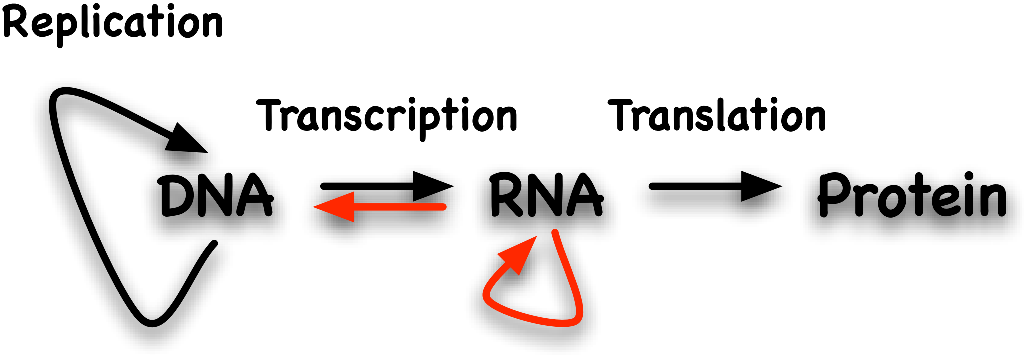
# Central Dogma (1958)



*The central dogma states that once "**information**" has passed into a protein it cannot get out again. The transfer of information from **nucleic acid to nucleic acid**, or from **nucleic acid to protein**, **may be possible**, but transfer from protein to protein, or from protein to nucleic acid, is impossible. Information here means the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.*

Francis Crick (1958) *Symposium of the Society of Experimental Biology* **12**:138-167.

# Central Dogma (2019)



Replication

Transcription

Translation

DNA

RNA

Protein

http://www.yourgenome.org/facts/what-is-the-central-dogma

# Central Dogma (contd)

**DNA**: stores genetic information (library of programs);

# Central Dogma (contd)

**DNA**: stores genetic information (library of programs);

**RNA**: stores a copy a gene during protein synthesis (mRNA),

# Central Dogma (contd)

**DNA**: stores genetic information (library of programs);

**RNA**: stores a copy a gene during protein synthesis (mRNA), adapter molecule involved proteins synthesis (tRNA),

# Central Dogma (contd)

**DNA**: stores genetic information (library of programs);

**RNA**: stores a copy a gene during protein synthesis (mRNA), adapter molecule involved proteins synthesis (tRNA), part of the ribosome (a ribo-protein complex),

# Central Dogma (contd)

**DNA**: stores genetic information (library of programs);

**RNA**: stores a copy a gene during protein synthesis (mRNA), adapter molecule involved proteins synthesis (tRNA), part of the ribosome (a ribo-protein complex), regulation/development (micro-RNAs, regulatory motifs, riboswitches, etc.);

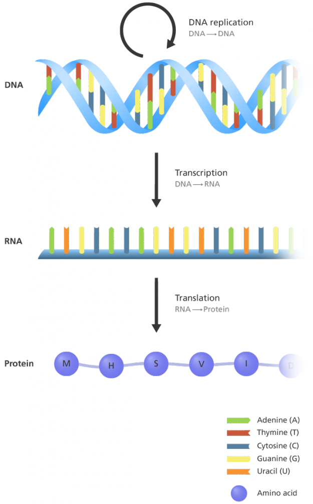
# Central Dogma (contd)

**DNA**: stores genetic information (library of programs);

**RNA**: stores a copy a gene during protein synthesis (mRNA), adapter molecule involved proteins synthesis (tRNA), part of the ribosome (a ribo-protein complex), regulation/development (micro-RNAs, regulatory motifs, riboswitches, etc.);

**Proteins**: catalyse reactions (modulator), communication (signalling), transport, structure, etc.

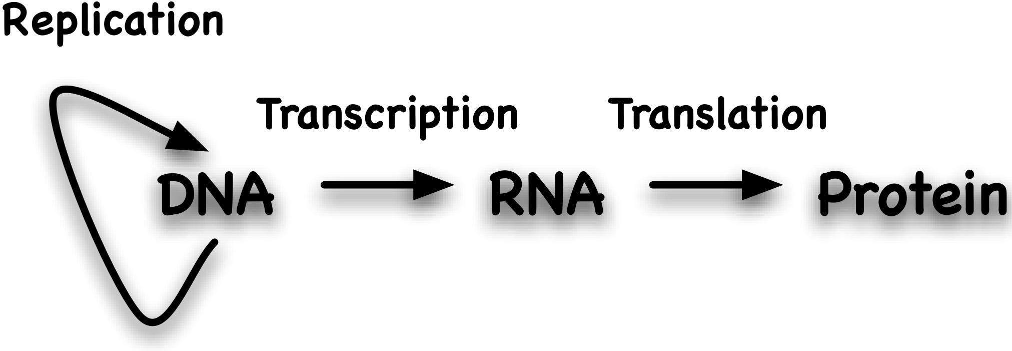
# Central Dogma



**Source:** `https://www.yourgenome.org`

# Replication

# Central Dogma (1958)



Francis Crick (1958) *Symposium of the Society of Experimental Biology* **12**:138-167.

# DNA and Heredity

- **DNA structure explains how information can be copied from one generation to the next**, or simply from one parent cell to its daughter cells during replication.

**Before replication**

```
5'  - GATACA -> 3' A
      ||||||
3' <- CTATGT -  5' B
```

⇒

**A** is as a template to produce **B’**

```
5'  - GATACA -> 3' A


5'  - GATACA -> 3' A
      ||||||
3' <- CTATGT -  5' B'
```

# DNA and Heredity

**Before replication**

```
5'  - GATACA -> 3' A
      ||||||
3' <- CTATGT -  5' B
```

⇒

**B** is as a template to produce **A'**

```
5' -  TGTATC -> 3' B


5' -  TGTATC -> 3' B
      ||||||
3' <- ACATAG -> 5' A'
```

# DNA and Heredity

### Parent cell (AB)

```
5'  - GATACA -> 3' A
      ||||||
3' <- CTATGT -  5' B
```

### Daughter cell AB'

```
5'  - GATACA -> 3' A
      ||||||
3' <- CTATGT -  5' B'
```

### Daughter cell A'B

```
5' -  TGTATC -> 3' B
      ||||||
3' <- ACATAG -> 5' A'
```

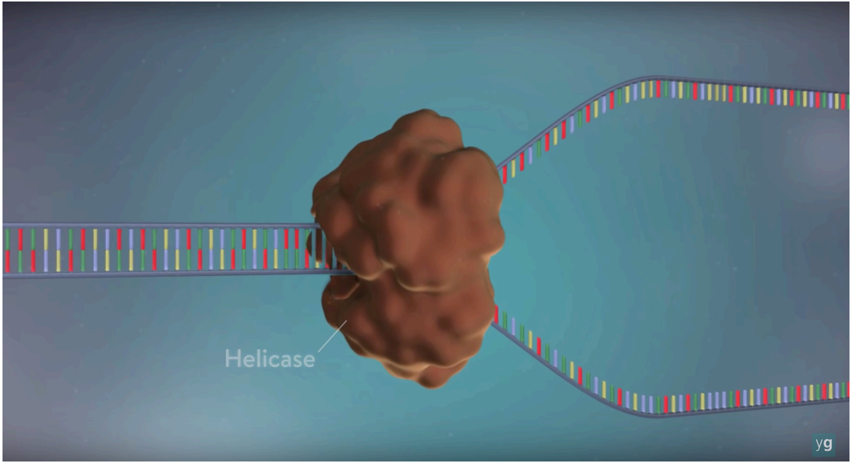
Two daughter cells, identical to their parent.
(**semi-conservative** process)

# Remarks

- Complex organisms are growing from a single cell to billions of cells. Each cell contains an **exact copy**[1] of the **DNA** of its **parent cell**.
- The information is redundant, the information on the second strand can be inferred from the information on the first strand. This is the basis of **DNA repair mechanisms**. A base that is deleted can be replaced. A mismatch can be detected.

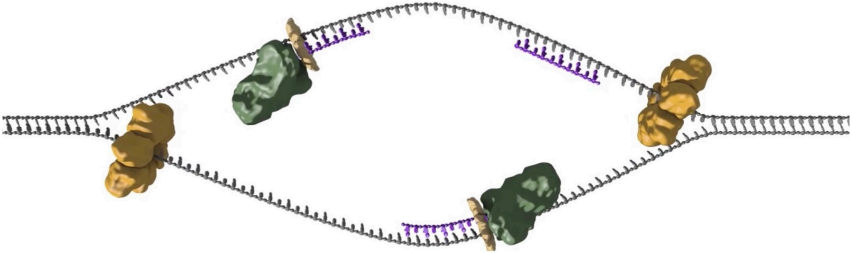
[1] With the exception of mature red blood cells (no DNA), germ cells (half of the DNA), or B cells.

# DNA Replication: DNA ⟶ DNA (basic)



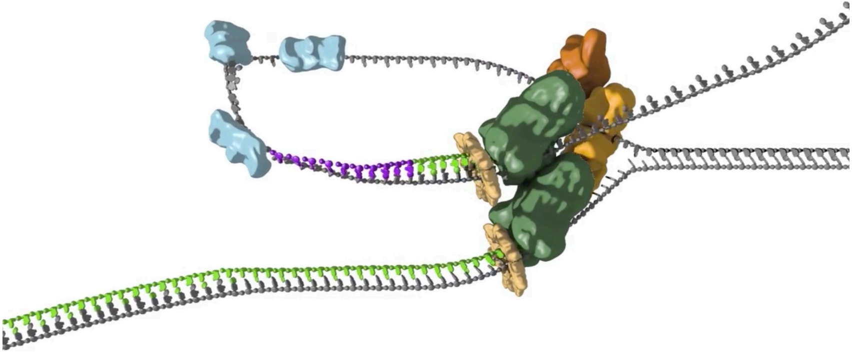
https://youtu.be/TNKWgcFPHqw

# DNA Replication: DNA ⟶ DNA (advanced)



https://youtu.be/0Ha9nppnw0c

# DNA Replication: DNA $\longrightarrow$ DNA (extreme)



https://youtu.be/QMX7IpME7X8

# Replication: Summary

- Replication is catalyzed by an enzyme (protein) called **DNA polymerase**.

# Replication: Summary

- Replication is catalyzed by an enzyme (protein) called **DNA polymerase**.
- The **complementarity of the base pairs is fundamental** to DNA replication mechanisms.

# Replication: Summary

- Replication is catalyzed by an enzyme (protein) called **DNA polymerase**.
- The **complementarity of the base pairs is fundamental** to DNA replication mechanisms.
- Each strand of a DNA molecule serves as a **template** for producing a complementary copy.

# Replication: Summary

- Replication is catalyzed by an enzyme (protein) called **DNA polymerase**.
- The **complementarity of the base pairs is fundamental** to DNA replication mechanisms.
- Each strand of a DNA molecule serves as a **template** for producing a complementary copy.
- The result is two double helices identical to their parent; each daughter molecule has one strand of its parent (this is called a **semi-conservative** system).

# Replication: Summary

- Replication is catalyzed by an enzyme (protein) called **DNA polymerase**.
- The **complementarity of the base pairs is fundamental** to DNA replication mechanisms.
- Each strand of a DNA molecule serves as a **template** for producing a complementary copy.
- The result is two double helices identical to their parent; each daughter molecule has one strand of its parent (this is called a **semi-conservative** system).
- It is a complex process (timing, topology, distribution to daughter cells). Some of its important steps were understood in the 1980s whilst the details are still an active research topic.

# Replication: Summary

- Replication is catalyzed by an enzyme (protein) called **DNA polymerase**.
- The **complementarity of the base pairs is fundamental** to DNA replication mechanisms.
- Each strand of a DNA molecule serves as a **template** for producing a complementary copy.
- The result is two double helices identical to their parent; each daughter molecule has one strand of its parent (this is called a **semi-conservative** system).
- It is a complex process (timing, topology, distribution to daughter cells). Some of its important steps were understood in the 1980s whilst the details are still an active research topic.
- Remember higher levels of organization of DNA!

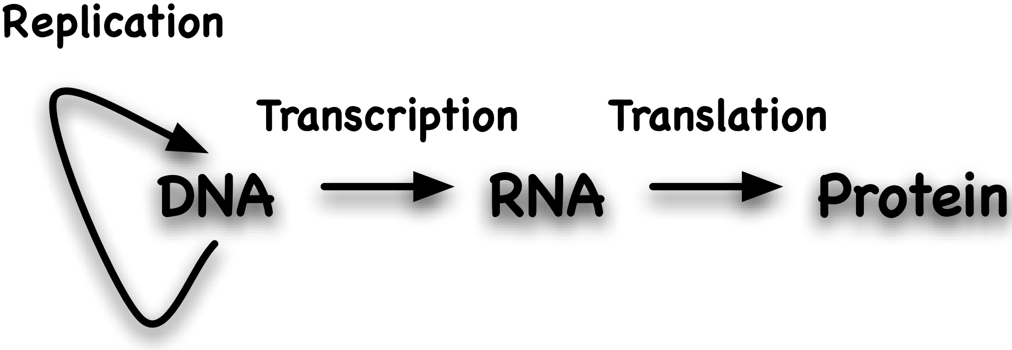
# Questions

Do not answer these questions right away. Keep them mind throughout the presentation.

- Replication is catalyzed by several enzymes, including **DNA polymerase**, **Primase**, **Ligase**, and **DNA helicase**.
- An **enzyme** is a macromolecule that **accelerate** a specific **chemical reaction**. Most enzymes are **proteins**. The ones above are.
- Where do protein come from?
- How are they regulated?

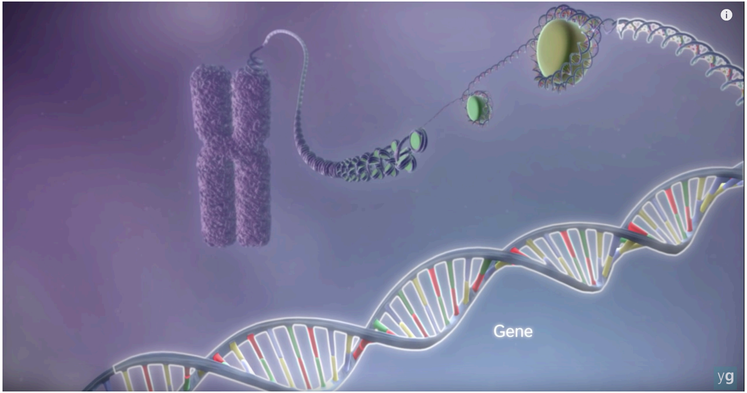
# Transcription

# Central Dogma (1958)



Francis Crick (1958) *Symposium of the Society of Experimental Biology* **12**:138-167.

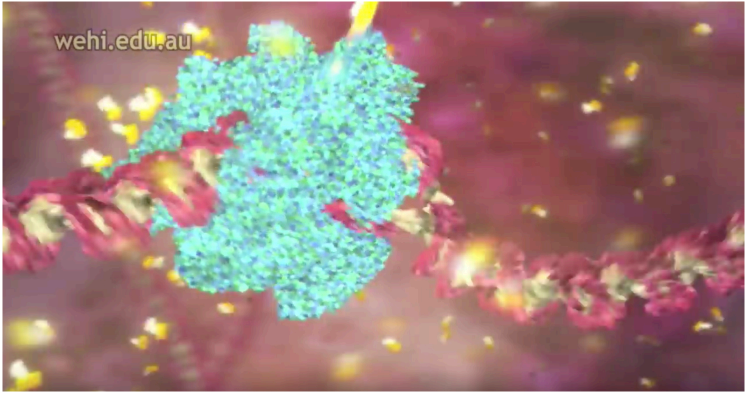
# Transcription: DNA ⟶ RNA (basic)



https://www.youtube.com/watch?v=gG7uCskUOrA [2]

[2] The video includes translation as well.

# Transcription: DNA ⟶ RNA (detailed)



`https://youtu.be/DA2t5N72mgw?list=PLD0444BD542B4D7D9` [3]

[3] The video includes translation as well.

# Genes

"(. . . ) a **gene** is a **sequence of genomic DNA** (. . . ) that **is essential for a specific function**." Li & Graur 1991.

There are three (3) kinds of genes:

1. **Protein-coding** genes
2. **RNA-coding** genes
3. **Regulatory** genes.

1 & 2 are called **structural gene** (only 1 for some authors).
The **genome** is the sum of all the genes.

# Transcription (continued)

- Transcription of **prokaryotic** genes is under the control of one type of RNA polymerase.
- While 3 are involved in this process for the **eukaryotic** genes (rRNA by RNA polymerase I, **protein-coding genes by RNA polymerase II**, while small cytoplasmic RNA genes, such as tRNA-specifying genes are under the control of RNA polymerase III, small nuclear RNA genes are transcribed by RNA polymerase II and/or III (U6 transcribed by II or III)).

# Transcription: DNA $\longrightarrow$ RNA

**The need for an intermediate molecule**. In Eukaryotes, it had been observed that proteins are synthesised in the cytoplasm (inside the cell but outside of the nucleus), whereas DNA is found in the nucleus.

- Carried out by a (DNA-dependent) **RNA polymerase**.

The collection of the transcripts is called the **transcriptome**.

# Transcription: DNA $\longrightarrow$ RNA

**The need for an intermediate molecule**. In Eukaryotes, it had been observed that proteins are synthesised in the cytoplasm (inside the cell but outside of the nucleus), whereas DNA is found in the nucleus.

- Carried out by a (DNA-dependent) **RNA polymerase**.
- Requires the presence of specific sequences (**called signals**) upstream of the start of transcription (in the case of protein-coding genes). This region is called the **promoter**.

The collection of the transcripts is called the **transcriptome**.

# Transcription: DNA $\longrightarrow$ RNA

**The need for an intermediate molecule**. In Eukaryotes, it had been observed that proteins are synthesised in the cytoplasm (inside the cell but outside of the nucleus), whereas DNA is found in the nucleus.

- Carried out by a (DNA-dependent) **RNA polymerase**.
- Requires the presence of specific sequences (**called signals**) upstream of the start of transcription (in the case of protein-coding genes). This region is called the **promoter**.
- In **Eukaryotes**, the messenger **RNA** contains non-coding regions, called **introns**, that are removed through various processes, called intron splicing. Before splicing the transcript is called a pre-mRNA.

The collection of the transcripts is called the **transcriptome**.

# DNA-RNA relationship

```
DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
```

# DNA-RNA relationship

```
DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...

DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
              |||||
RNA:          AUGGC
```

# DNA-RNA relationship

```
DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...

DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
              |||||
RNA:          AUGGC

DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
              ||||||
RNA:          AUGGCG ...
```

# DNA-RNA relationship

```
DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...

DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
              |||||
RNA:          AUGGC

DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
              ||||||
RNA:          AUGGCG ...
```

...

```
DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
              ||||||||||||||||||||||||||||||
RNA:          AUGGCGCCGAUAAUGUCGGUCCUUCCUUGA
```

# Transcription (continued)

Conceptually simple, **one to one relationship** between each nucleotide of the source and the destination.

- **G** pairs with **C**;
- **A** pairs with **U** (not T);
- Uses **ribonucleotides**; instead of deoxyribonucleotides;

The result (product) is called a **(pre-)messenger RNA** or **transcript**.

# Transcription (continued)

- I don't understand, **is it the whole of the genome that is transcribed?**

# Transcription (continued)

- I don't understand, **is it the whole of the genome that is transcribed?** No, translation is is not initiated randomly but at specific sites, called **promoters**.

Here is the consensus sequence for the **core promoter** in *E. coli* (*Escherichia coli*):

```
TTGACA(N){16,18}TATAAT
```

# Transcription (continued)

- I don't understand, **is it the whole of the genome that is transcribed?** No, translation is is not initiated randomly but at specific sites, called **promoters**.

Here is the consensus sequence for the **core promoter** in *E. coli* (*Escherichia coli*):

```
TTGACA(N){16,18}TATAAT
```

What is the likelihood of this motif to occur?

# Transcription (continued)

- Here size does matter, and it depends on your assumptions. How do you want to model the promoter sequence motif?

# Transcription (continued)

- Here size does matter, and it depends on your assumptions. How do you want to model the promoter sequence motif?
- The simplest model is *i.i.d.*, which stands for **independent and identically distributed**.

# Transcription (continued)

- Here size does matter, and it depends on your assumptions. How do you want to model the promoter sequence motif?
- The simplest model is *i.i.d.*, which stands for **independent and identically distributed**.
- What does it mean?

# Transcription (continued)

- Here size does matter, and it depends on your assumptions. How do you want to model the promoter sequence motif?
- The simplest model is *i.i.d.*, which stands for **independent and identically distributed**.
- What does it mean?
- First, since the positions are considered to be independent one from another, the probability of the motif is the **product of the probabilities** of occurrence of the nucleotides at each position.

# Transcription (continued)

- Here size does matter, and it depends on your assumptions. How do you want to model the promoter sequence motif?
- The simplest model is *i.i.d.*, which stands for **independent and identically distributed**.
- What does it mean?
- First, since the positions are considered to be independent one from another, the probability of the motif is the **product of the probabilities** of occurrence of the nucleotides at each position.
- Second, we also assume that the probability distribution for the nucleotides is the same for all the positions.

# Transcription (continued)

- Here size does matter, and it depends on your assumptions. How do you want to model the promoter sequence motif?
- The simplest model is *i.i.d.*, which stands for **independent and identically distributed**.
- What does it mean?
- First, since the positions are considered to be independent one from another, the probability of the motif is the **product of the probabilities** of occurrence of the nucleotides at each position.
- Second, we also assume that the probability distribution for the nucleotides is the same for all the positions.
- In general, the **maximum likelihood estimators** are used to estimated the probability distributions, which simply means that a large number of examples are collected and that the frequencies of occurrence are used as estimators.

# Simple probabilistic model

`TTGACA(N){16,18}TATAAT`

- To make the argument simple, we can assume the events to be equally likely, $p_A = p_C = p_G = p_T = \frac{1}{4}$, so that the probability of the motif is $\frac{1}{4^{12}} = 6 \times 10^{-8}$.

# Simple probabilistic model

`TTGACA(N){16,18}TATAAT`

- To make the argument simple, we can assume the events to be equally likely, $p_A = p_C = p_G = p_T = \frac{1}{4}$, so that the probability of the motif is $\frac{1}{4^{12}} = 6 \times 10^{-8}$.
- **How many promoters** would you expect to find in the *E. Coli* genome?
- $6 \times 10^{-8} \times 4.6 \ \mathrm{Mb} = 0.276 < 1$.

# Simple probabilistic model

`TTGACA(N){16,18}TATAAT`

- To make the argument simple, we can assume the events to be equally likely, $p_A = p_C = p_G = p_T = \frac{1}{4}$, so that the probability of the motif is $\frac{1}{4^{12}} = 6 \times 10^{-8}$.
- **How many promoters** would you expect to find in the *E. Coli* genome?
- $6 \times 10^{-8} \times 4.6 \text{ Mb} = 0.276 < 1$.
- Eukaryotic genomes are larger, often billions of bp, and accordingly their promoter sequence is more complex!

# Simple probabilistic model

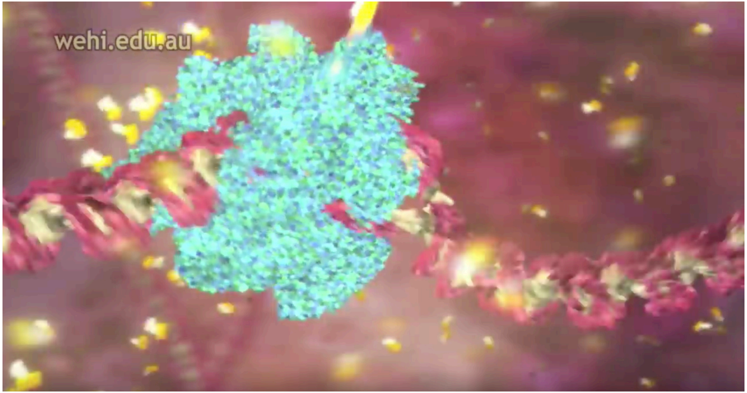
`TTGACA(N){16,18}TATAAT`

- To make the argument simple, we can assume the events to be equally likely, $p_A = p_C = p_G = p_T = \frac{1}{4}$, so that the probability of the motif is $\frac{1}{4^{12}} = 6 \times 10^{-8}$.
- **How many promoters** would you expect to find in the *E. Coli* genome?
- $6 \times 10^{-8} \times 4.6\ \mathrm{Mb} = 0.276 < 1$.
- Eukaryotic genomes are larger, often billions of bp, and accordingly their promoter sequence is more complex!
- Finally, other regulatory sequences exist, which are the binding site for regulatory proteins, which can enhance the transcription, positive regulation, or inhibit transcription, negative regulation.

# Bioinformaticist's point of view

- The discovery of (new) **regulatory motifs** (promotors, signals, etc.) is an active area of research.

# Transcription: DNA ⟶ RNA (detailed)



https://youtu.be/DA2t5N72mgw?list=PLD0444BD542B4D7D9 [4]

[4] The video includes translation as well.

# About the animation

- Transcription factors assemble at a DNA promoter region found at the start of a gene. Promoter regions are characterised by the DNA's base sequence, which contains the repetition TATATA and for this reason is known as the "TATA box".
- The TATA box is gripped by the transcription factor TFIID (yellow-brown) that marks the attachment point for RNA polymerase and associated transcription factors. In the middle of TFIID is the TATA Binding Protein subunit, which recognises and fastens onto the TATA box. It's tight grip makes the DNA kink 90 degrees, which is thought to serve as a physical landmark for the start of a gene.

# About the animation

- A mediator (purple) protein complex arrives carrying the enzyme RNA polymerase II (blue-green). It manoeuvres the RNA polymerase into place. Other transcription factors arrive (TFIIA and TFIIB - small blue molecules) and lock into place. Then TFIIH (green) arrives. One of its jobs is to pry apart the two strands of DNA (via helicase action) to allow the RNA polymerase to get access to the DNA bases.

# About the animation

- A mediator (purple) protein complex arrives carrying the enzyme RNA polymerase II (blue-green). It manoeuvres the RNA polymerase into place. Other transcription factors arrive (TFIIA and TFIIB - small blue molecules) and lock into place. Then TFIIH (green) arrives. One of its jobs is to pry apart the two strands of DNA (via helicase action) to allow the RNA polymerase to get access to the DNA bases.
- Finally, the initiation complex requires contact with activator proteins, which bind to specific sequences of DNA known as enhancer regions. These regions can be thousands of base pairs away from the initiation complex. The consequent bending of the activator protein/enhancer region into contact with the initiation-complex resembles a scorpion's tail in this animation.

# About the animation

- The activator protein triggers the release of the RNA polymerase, which runs along the DNA transcribing the gene into mRNA (yellow ribbon).

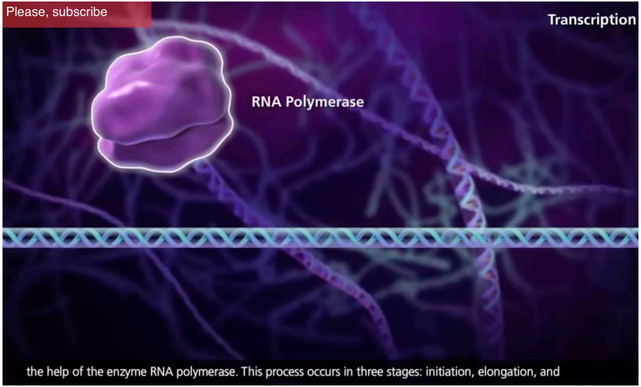
# About the animation

- The RNA polymerase unzips a small portion of the DNA helix exposing the bases on each strand. One of the strands acts as a template for the synthesis of an RNA molecule. The base-sequence code is transcribed by matching these DNA bases with RNA subunits, forming a long RNA polymer chain.

# Transcriptome and gene regulation

- Messenger RNA are degraded minutes (**prokaryotes**) or hours (**eukaryotes**) after synthesis.
- Furthermore, information stored in the untranslated regions of the transcript is involved in regulation and transport.

# Transcription: DNA ⟶ RNA (detailed)

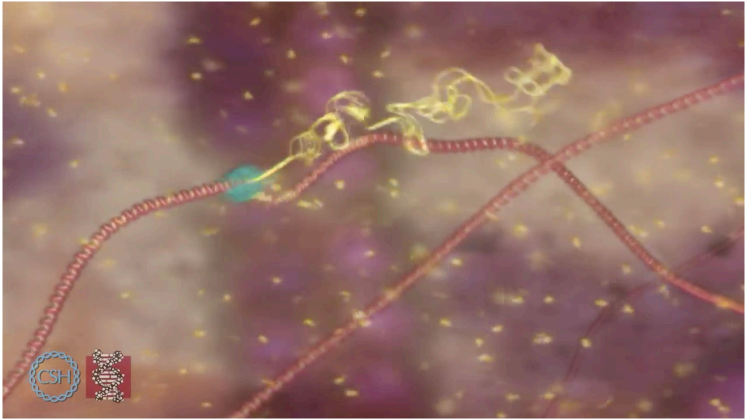


the help of the enzyme RNA polymerase. This process occurs in three stages: initiation, elongation, and

`https://youtu.be/-K8Y0ATkkAI`[5]

[5] The video includes translation as well.

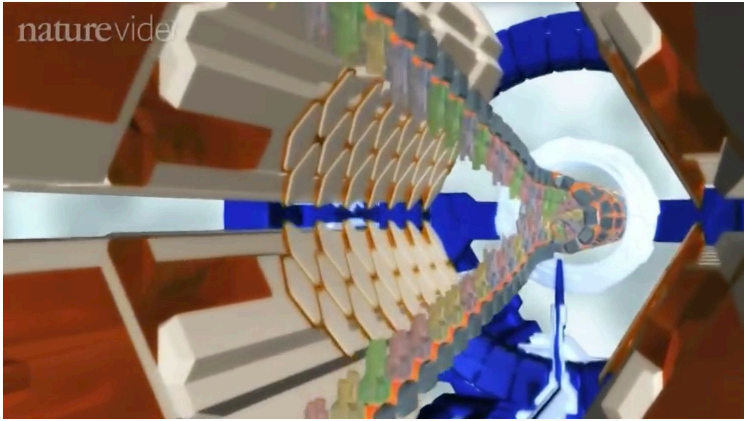
# Transcription: DNA ⟶ RNA (detailed)



https://youtu.be/9kOGOY7vthk [6]

[6] The video includes translation as well.

# Transcription: DNA ⟶ RNA (futuristic)



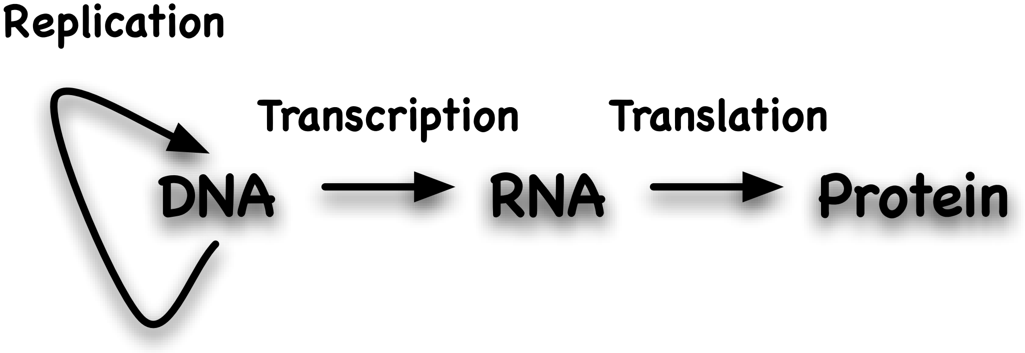
https://www.youtube.com/watch?v=J3HVVi2k2No

# Resources

- **Walter and Eliza Hall Institute of Medical Research** Videos
  - https://www.youtube.com/playlist?list=PLD0444BD542B4D7D9
- **Cold Spring Harbor Laboratory**'s DNA Learning Center
  - https://www.youtube.com/user/DNALearningCenter
- The Central dogma by **RIKEN Yokohama institute Omics Science Center**
  - https://youtu.be/ZNcFTRX9i0Y

# Translation

# Central Dogma (1958)



Replication

Transcription

Translation

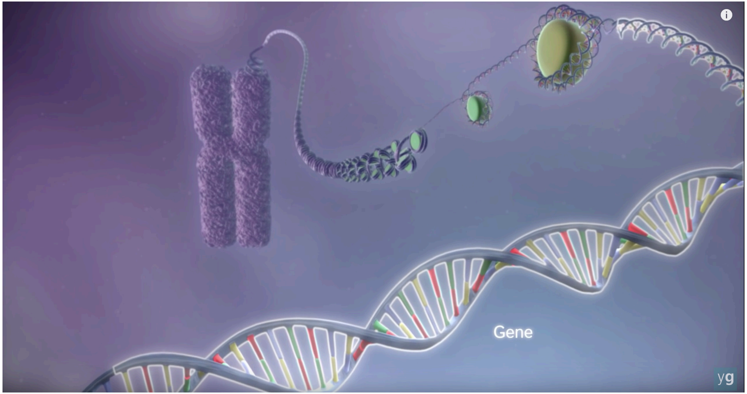
DNA

RNA

Protein

Francis Crick (1958) *Symposium of the Society of Experimental Biology* **12**:138-167.

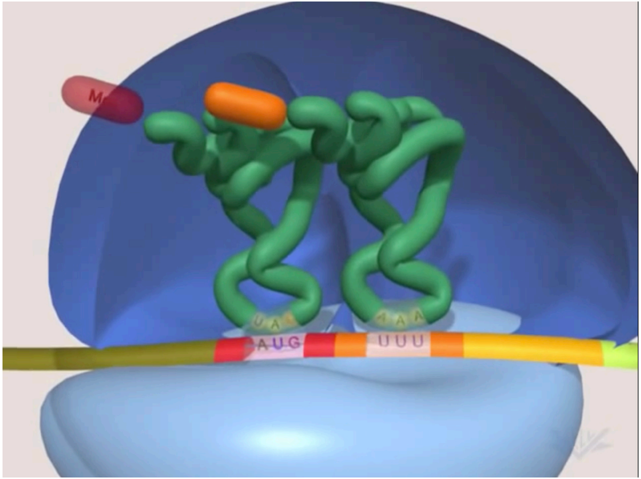
# Transcription: DNA $\longrightarrow$ RNA (basic)



https://youtu.be/gG7uCskU0rA?t=87 [7]

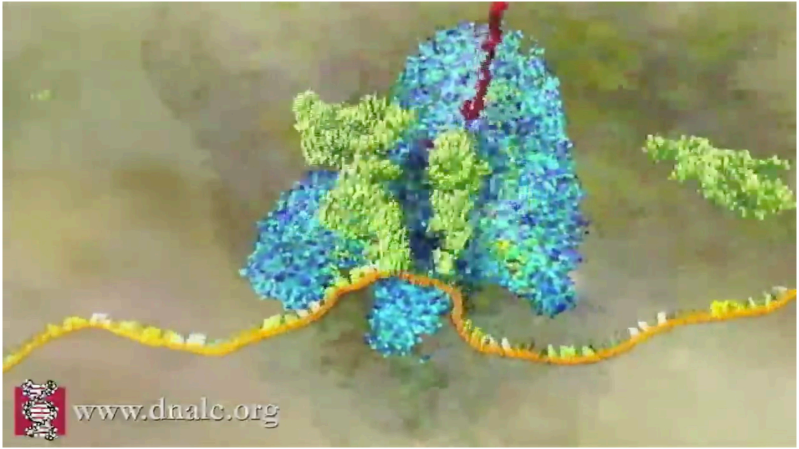
[7] The video includes transcription as well.

# Translation: RNA ⟶ Protein (basic)



https://youtu.be/5bLEDd-PSTQ

# Translation: RNA ⟶ Protein (detailed)



https://youtu.be/WkI_Vbwn14g?list=PLD0444BD542B4D7D9

# Translation: RNA $\longrightarrow$ Protein

- Translation is under the control of a riboprotein complex called the **ribosome**, adapter RNA molecules, called **tRNAs**, and several other proteins to control the regulation, charging tRNA molecules with the appropriate amino acids.

# Translation: RNA $\longrightarrow$ Protein

- Translation is under the control of a riboprotein complex called the **ribosome**, adapter RNA molecules, called **tRNAs**, and several other proteins to control the regulation, charging tRNA molecules with the appropriate amino acids.
- It is clear that what ever coding principle exists, there cannot be a one-to-one mapping!

# Translation: RNA $\longrightarrow$ Protein

- Translation is under the control of a riboprotein complex called the **ribosome**, adapter RNA molecules, called **tRNAs**, and several other proteins to control the regulation, charging tRNA molecules with the appropriate amino acids.
- It is clear that what ever coding principle exists, there cannot be a one-to-one mapping!

# Translation: RNA $\longrightarrow$ Protein

- Translation is under the control of a riboprotein complex called the **ribosome**, adapter RNA molecules, called **tRNAs**, and several other proteins to control the regulation, charging tRNA molecules with the appropriate amino acids.
- It is clear that what ever coding principle exists, there cannot be a one-to-one mapping! $4^1 < 20, 4^2 < 20, 4^3 > 20$!
- For each consecutive three nucleotide, this is called a codon (coding unit), correspond a unique amino acid.

$$4 \times 4 \times 4 = 64$$

# Translation: RNA $\longrightarrow$ Protein

- Translation is under the control of a riboprotein complex called the **ribosome**, adapter RNA molecules, called **tRNAs**, and several other proteins to control the regulation, charging tRNA molecules with the appropriate amino acids.
- It is clear that what ever coding principle exists, there cannot be a one-to-one mapping! $4^1 < 20, 4^2 < 20, 4^3 > 20$!
- For each consecutive three nucleotide, this is called a codon (coding unit), correspond a unique amino acid.

$$4 \times 4 \times 4 = 64$$

- **Contiguous**, **non-overlapping** triplets.

# Translation: RNA $\longrightarrow$ Protein

- Translation is under the control of a riboprotein complex called the **ribosome**, adapter RNA molecules, called **tRNAs**, and several other proteins to control the regulation, charging tRNA molecules with the appropriate amino acids.
- It is clear that what ever coding principle exists, there cannot be a one-to-one mapping! $4^1 < 20, 4^2 < 20, 4^3 > 20$!
- For each consecutive three nucleotide, this is called a codon (coding unit), correspond a unique amino acid.

$$4 \times 4 \times 4 = 64$$

- **Contiguous**, **non-overlapping** triplets.
- Since there are 64 possible codons, the code is said to be **degenerated**, i.e. several triples map onto the same amino acid.

# Universal Genetic Code

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

# DNA-RNA-Protein relationships

```
    DNA: TAC CGC GCC TAT TAC TGC CAG GAA GGA ACT
    RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein:  M   A   P   I   M   T   V   L   P   *
```

```
    DNA: TAC CGC GCC TAT TAC TGC CAG GAA GGA ACT
    RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein: Met Ala Pro Ile Met Thr Val Leu Pro Stop
```

⇒ Example from **Jones & Pevzner**, p. 65.

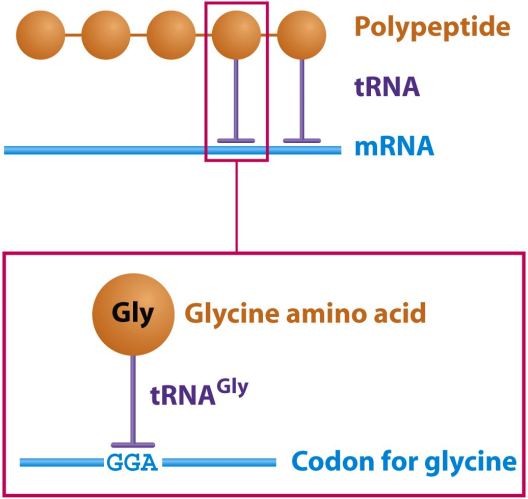
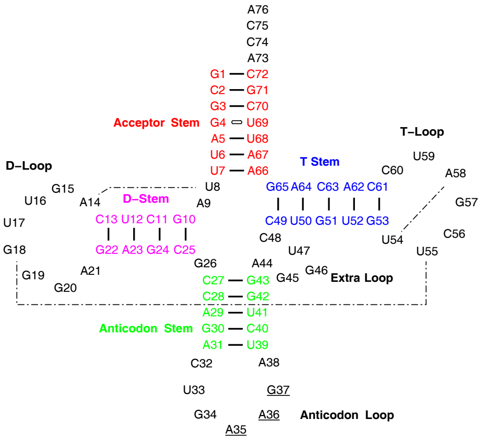


Figure 13-1 Genomes 3 (© Garland Science 2007)

# tRNA: 1, 2, 3

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

1 10 20 30 40 50 60 70



A76
C75
C74
A73
G1 — C72
C2 — G71
G3 — C70
**Acceptor Stem** G4 ⊂⊃ U69
A5 — U68
U6 — A67
U7 — A66

**D–Loop**
G15
U16
U17
G18
G19
G20
A21
A14
U8
A9

**D–Stem**
C13 U12 C11 G10
G22 A23 G24 C25

G26

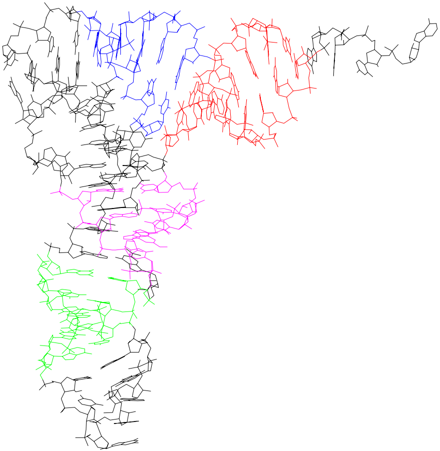
**T Stem**
G65 A64 C63 A62 C61
C49 U50 G51 U52 G53

**T–Loop**
C60
U59
A58
G57
C56
U55
U54

C48
U47
G46
G45
A44
**Extra Loop**

C27 — G43
C28 — G42
A29 — U41
**Anticodon Stem** G30 — C40
A31 — U39

C32
A38
U33
G37
G34
A36
A35
**Anticodon Loop**



# Transfer RNA (tRNA)

The transfer RNAs (tRNAs) are a

- **Adaptor molecules**.

# Transfer RNA (tRNA)

The transfer RNAs (tRNAs) are a

- **Adaptor molecules**.

# Transfer RNA (tRNA)

The transfer RNAs (tRNAs) are a

- **Adaptor molecules**. Bacteria have 30 to 45 different adaptors whilst some eukaryotes have up to 50 (48 in the case of humans).
- **Each tRNA is loaded (charged) with a specific amino acid at one end, and has a specific (triplet) sequence, called the anti-codon, at the other end.**

# Transfer RNA (tRNA)

The transfer RNAs (tRNAs) are a

- **Adaptor molecules**. Bacteria have 30 to 45 different adaptors whilst some eukaryotes have up to 50 (48 in the case of humans).
- **Each tRNA is loaded (charged) with a specific amino acid at one end, and has a specific (triplet) sequence, called the anti-codon, at the other end.**
- Notation: tRNA$^{\mathrm{Phe}}$ is a tRNA molecule specific for phenylalanine (one of the 20 amino acids).

# Transfer RNA (tRNA)

The transfer RNAs (tRNAs) are a

- **Adaptor molecules**. Bacteria have 30 to 45 different adaptors whilst some eukaryotes have up to 50 (48 in the case of humans).
- **Each tRNA is loaded (charged) with a specific amino acid at one end, and has a specific (triplet) sequence, called the anti-codon, at the other end.**
- Notation: tRNA$^{\mathrm{Phe}}$ is a tRNA molecule specific for phenylalanine (one of the 20 amino acids).
- The tRNA molecules are 70 to 90 nt long and virtually all of them fold into the same cloverleaf structure presented on the previous slide.

# Transfer RNA (tRNA)

- As will be seen next, it is quite important that **all the tRNAs have a similar structure** so that one molecular machine (the ribosome) can be used for the protein synthesis.

# Transfer RNA (tRNA)

- As will be seen next, it is quite important that **all the tRNAs have a similar structure** so that one molecular machine (the ribosome) can be used for the protein synthesis.
- The enzymes responsible for “charging” the proper amino acid onto each tRNA are called aminoacyl-tRNA synthetases.

# Transfer RNA (tRNA)

- As will be seen next, it is quite important that **all the tRNAs have a similar structure** so that one molecular machine (the ribosome) can be used for the protein synthesis.
- The enzymes responsible for “charging” the proper amino acid onto each tRNA are called aminoacyl-tRNA synthetases.

# Transfer RNA (tRNA)

- As will be seen next, it is quite important that **all the tRNAs have a similar structure** so that one molecular machine (the ribosome) can be used for the protein synthesis.
- The enzymes responsible for “charging” the proper amino acid onto each tRNA are called aminoacyl-tRNA synthetases. Most organisms have 20 aminoacyl-tRNA synthetases, meaning that a given aminoacyl-tRNA synthetase is responsible for the attachment of a specific amino acid on all the isoacepting tRNAs (different tRNAs charged with the same amino acid type).
- **Each tRNA also has unique features so that it gets loaded with the right amino acid.**

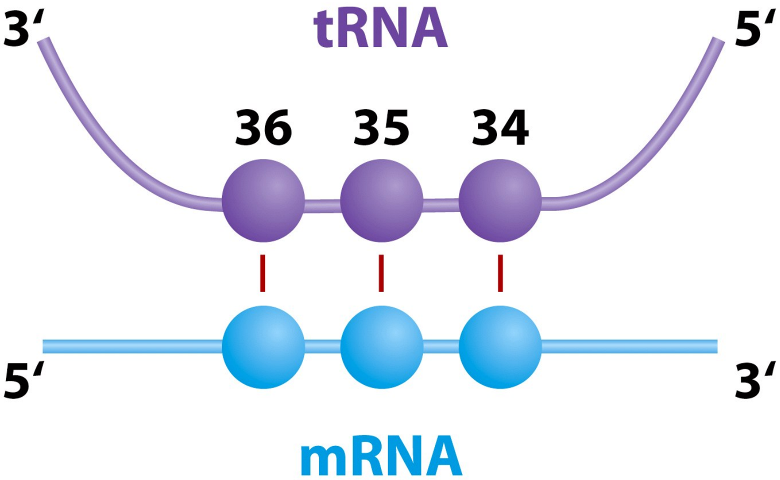


Figure 13-6 Genomes 3 (© Garland Science 2007)

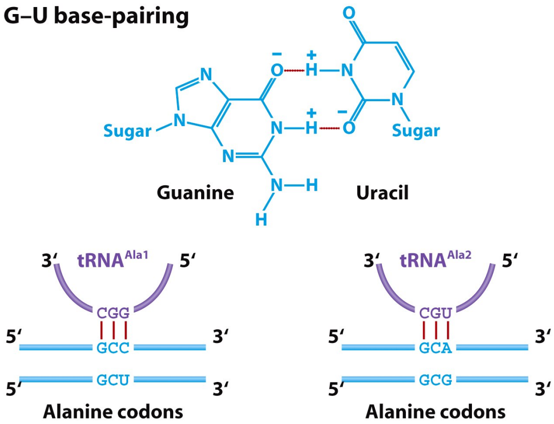


Figure 13-7a Genomes 3 (© Garland Science 2007)

**Wobble base** pairs are possible and **reduce the number of tRNAs needed** since the same tRNA binds 2 or possibly 3 codons.

# Ribosomes play an essential role in translation

Large RNAs + proteins complex (the result of the association of 3 to 4 RNAs + 55 to 83 proteins!).

# Ribosomes play an essential role in translation

Large RNAs + proteins complex (the result of the association of 3 to 4 RNAs + 55 to 83 proteins!). In bacteria, there are approximately 20,000 ribosomes at any given time (more in eukaryotes).

# Ribosomes play an essential role in translation

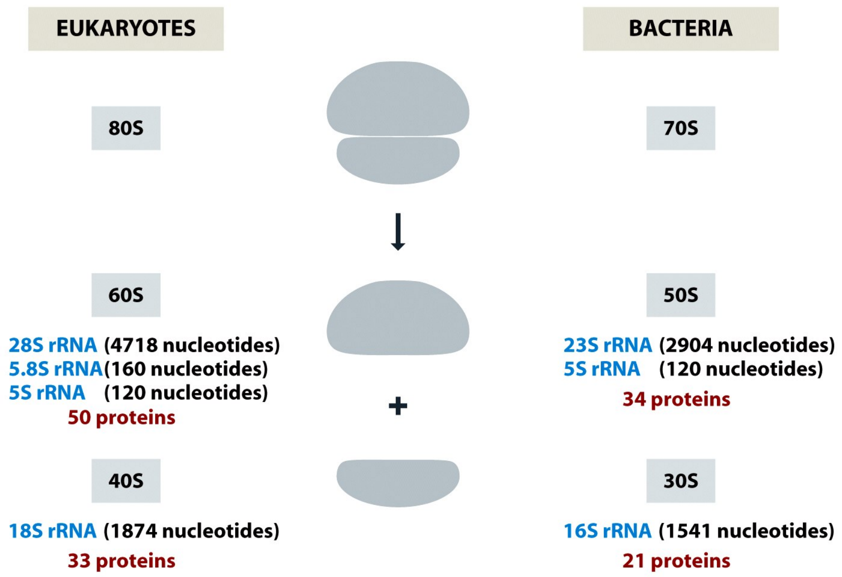
Large RNAs + proteins complex (the result of the association of 3 to 4 RNAs + 55 to 83 proteins!). In bacteria, there are approximately 20,000 ribosomes at any given time (more in eukaryotes).

- Coordinate protein synthesis by orchestrating the placement of the messenger RNAs (mRNAs), the transfer RNAs (tRNAs) and necessary protein factors;

# Ribosomes play an essential role in translation

Large RNAs + proteins complex (the result of the association of 3 to 4 RNAs + 55 to 83 proteins!). In bacteria, there are approximately 20,000 ribosomes at any given time (more in eukaryotes).

- Coordinate protein synthesis by orchestrating the placement of the messenger RNAs (mRNAs), the transfer RNAs (tRNAs) and necessary protein factors;
- Catalyze (at least partially) some of the chemical reactions involved in protein synthesis.



EUKARYOTES

80S

60S

28S rRNA (4718 nucleotides)
5.8S rRNA(160 nucleotides)
5S rRNA (120 nucleotides)
50 proteins

40S

18S rRNA (1874 nucleotides)
33 proteins

BACTERIA

70S

50S

23S rRNA (2904 nucleotides)
5S rRNA (120 nucleotides)
34 proteins

30S

16S rRNA (1541 nucleotides)
21 proteins

Figure 13-10 Genomes 3 (© Garland Science 2007)

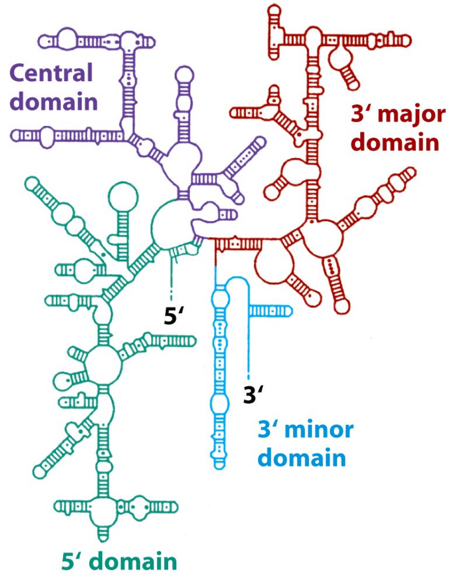


Figure 13-11 Genomes 3 (© Garland Science 2007)

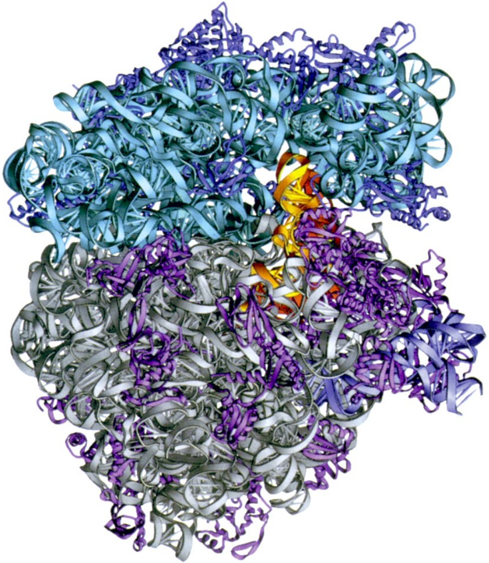
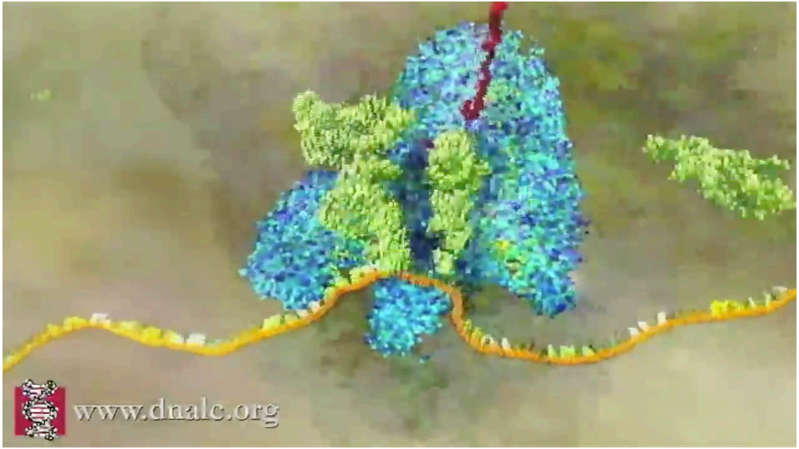


Figure 13-13 Genomes 3 (© Garland Science 2007)

# Translation: RNA $\longrightarrow$ Protein (detailed)



https://youtu.be/WkI_Vbwn14g?list=PLD0444BD542B4D7D9

# About the animation

- The message in mRNA (yellow) is decoded inside the ribosome (purple and light blue) and translated into a chain of amino acids (red).
- The ribosome is composed of one large (purple) and one small subunit (light blue), each with a specific task to perform. The small subunit's task is to match the triple letter code, known as a codon, to the anticodon at the base of each tRNA (green). The large subunit's task is to link the amino acids together into a chain. The amino acid chain exits the ribosome through a tunnel in the large subunit, then folds up into a three-dimensional protein molecule.

# About the animation

- As the mRNA is ratcheted through the ribosome, the mRNA sequence is translated into an amino acid sequence. The sequence of mRNA condons determines the specific amino acids that are added to the growing polypeptide chain. Selection of the correct amino acid is determined by complimentary base pairing between the mRNA's codon and the tRNA's anticodon. The codons are shown in this animation during the close up of the mRNA entering the ribosome. The codons are indicated as triplet groups of yellow-brown bases.
- tRNA (green) is a courier molecule carrying a single amino acid (red tip) as its parcel.

# Abous the animation

- During the amino acid chain synthesis, the tRNA steps through three locations inside the ribosome, referred to as the A-site, P-site and E-site. tRNA enters the ribosome and lodges in the A-site, where it is tested for a correct codon-anticodon match. If the tRNA's anticondon correctly matches the mRNA condon, it is stepped through to the P-site by a conformational change in the ribosome. In the P-site the amino acid carried by the tRNA is attached to the growing end of the amino acid chain.

# About the animation

The addition of amino acids is a three step cycle

1. The tRNA enters the ribosome at the A-site and is tested for a codon-anticodon match with the mRNA;
2. If it is a correct match, the tRNA is shifted to the P-site and the amino acid it carries is added to the end of the peptide chain. The mRNA is also ratcheted three nucleotides (1 codon);
3. The spent tRNA is moved to the E-site and then ejected from the ribosome.

# About the animation

- A typical eukaryotic cell contains millions of ribosomes in its cytoplasm.
- Many details, such as elongation factors (eg EFTu), have been omitted from this animation. This animation represents an idealised system with no incorrect tRNAs entering the ribosome, and consequently no error correction at the A-site.

Credit: **The Walter and Eliza Hall Institute of Medical Research**

# DNA-RNA-Protein relationships

```
    DNA: TAC CGC GCC TAT TAC TGC CAG GAA GGA ACT
    RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein:  M   A   P   I   M   T   V   L   P   *
```

```
    DNA: TAC CGC GCC TAT TAC TGC CAG GAA GGA ACT
    RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein: Met Ala Pro Ile Met Thr Val Leu Pro Stop
```

⇒ Example from **Jones & Pevzner**, p. 65.

# Remarks

- The translation starts at the **start codon**, ATG (AUG), and stops at a **stop codon**. The ATG codon determines the **reading frame** (phase).

# Remarks

- The translation starts at the **start codon**, ATG (AUG), and stops at a **stop codon**. The ATG codon determines the **reading frame** (phase).
- Most proteins start with a methionine. However, for certain mRNAs GUG or UUG are used as a start codon, or further processing removes the N-terminal part of the peptide (protein).
- 3 stop codons (non sense)
- 61 codons correspond to 20 aa (called sense codons) one of which is the start codon (codes for Met)
- The code is said to be **degenerated** because there are more than one code for each amino acid. Therefore, there is a unique translation, the same amino acid sequence can be encoded by more than one DNA sequence!

# Summary

- The code consists of triplets, called **codons**;

# Summary

- The code consists of triplets, called **codons**;
- The **start codon** is Met, which is the codon for amino acid Methionine;

# Summary

- The code consists of triplets, called **codons**;
- The **start codon** is Met, which is the codon for amino acid Methionine;
- There are **3 stop codons**; signifying the end of the chain, no amino acid is added;

# Summary

- The code consists of triplets, called **codons**;
- The **start codon** is Met, which is the codon for amino acid Methionine;
- There are **3 stop codons**; signifying the end of the chain, no amino acid is added;
- There are approximately **30 to 50 adapter molecules**, called transfer RNAs or **tRNAs** for short. Each tRNA is charged (loaded) with a specific amino acid, which correspond to its anti-codon. The tRNA molecules are nucleic acids and the recognition of the codon/anti-codon follows the normal base-pairing rules;

# Summary

- The code consists of triplets, called **codons**;
- The **start codon** is Met, which is the codon for amino acid Methionine;
- There are **3 stop codons**; signifying the end of the chain, no amino acid is added;
- There are approximately **30 to 50 adapter molecules**, called transfer RNAs or **tRNAs** for short. Each tRNA is charged (loaded) with a specific amino acid, which correspond to its anti-codon. The tRNA molecules are nucleic acids and the recognition of the codon/anti-codon follows the normal base-pairing rules;
- An **Open Reading Frame** (**ORF**) is a contiguous sequence of codons starting with Met (Start) and ending with a Stop codon;

# Summary

- Since the code is made of triplets, there are three possible translation frames in one strand, following that the start codon occurs at position $i$ mod $3 = 0, 1$ or 2;
- Since DNA is made of two complementary strands running anti-parallel, this makes a total of six translation frames.

A mutation occurring in a coding region will affect the gene product, the encoded protein.

# Genome sizes

| **Species** | **Size** |
|---|---:|
| Potato spindle tuber viroid (PSTVd) | 360 |
| Human immunodeficiency virus (HIV) | 9,700 |
| Bacteriophage lambda ($\lambda$) | 48,500 |
| *Mycoplasma genitalium* (bacterium) | 580,000 |
| *Escherichia coli* (bacterium) | 4,600,000 |
| *Drosophila melanogaster* (fruit fly) | 120,000,000 |
| *Homo sapiens* (human) | 3,000 000,000 |
| *Lilium longiflorum* (easter lily) | 90,000,000,000 |
| *Amoeba dubia* (amoeba) | 670,000,000,000 |

# Genome sizes

- *Haemophilus influenzae* (bacterium), dna = 1.8 Mbp
- *Escherichia coli* (baterium), dna = 4.6 Mbp
- *Saccharomyces cerevisiae* (yeast), dna = 12 Mbp
- *Caenorhabditis elegans* (worm), dna = 97 Mbp
- *Arabidopsis thaliana* (flowering plant), dna = 115 Mbp
- *Drosophila melanogaster* (fruit fly), dna = 137 Mbp
- Smallest Human chromosome (Y), dna = 50 Mbp
- Largest Human chromosome (1), dna = 250 Mbp
- Whole Human genome, dna = 3 Gbp
- *Mus musculus* (mouse), dna = 3 Gbp.

⇒ Mbp = million base pairs

# DNA is organized into chromosomes

> *The self-replicating genetic structures of cells containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular, and the entire genome is carried on one chromosome. Eukaryotic genomes consist of a number of chromosomes whose DNA is associated with different kinds of proteins.*

⇒ Work by Thomas Morgan in the 1920s established the connection between traits (genes) and chromosomes (DNA).

# Genome of multicellular animals (including

The human genome has two parts:

Nuclear genome: Consists of 23 pairs of chromosomes; for a total of 24 distinct linear molecules (22 autosomes and 2 sex chromosomes X and Y). The shortest chromosome consists of approximately 50 million nucleotides. The longest chromosome is more than 205 million nucleotides long. The sum of all the nucleotides is 3,2 billion nucleotides long. The nuclear genome encodes 20,000 to 25,000 protein genes.

# Genome of multicellular animals (including

The human genome has two parts:

Nuclear genome: Consists of 23 pairs of chromosomes; for a total of 24 distinct linear molecules (22 autosomes and 2 sex chromosomes X and Y). The shortest chromosome consists of approximately 50 million nucleotides. The longest chromosome is more than 205 million nucleotides long. The sum of all the nucleotides is 3,2 billion nucleotides long. The nuclear genome encodes 20,000 to 25,000 protein genes.

Mitochondrial genome: Consists of one circular molecule 16,569 nucleotides long, multiple copies of which are found in the organelles called mitochondria. The mitochondrial genome consists of 37 protein genes.

# Each cell has its own “identical” copy of the genome

- The adult human body consists of approximately $10^{13}$ cell.
- Each cell has its own copy of the genome.

# Human

- Most human cells are **diploid**, which means they have two copies of the 22 autosomes and two sex chromosomes (XX for females or XY for males).

# Human

- Most human cells are **diploid**, which means they have two copies of the 22 autosomes and two sex chromosomes (XX for females or XY for males).
- Diploid cells are also called **somatic** cells

# Human

- Most human cells are **diploid**, which means they have two copies of the 22 autosomes and two sex chromosomes (XX for females or XY for males).
- Diploid cells are also called **somatic** cells
- Sex cells (or **gametes**) are **haploid** and therefore have a single copy of the 22 autosomes as well as one sex chromosome.

# Bioinformaticist’s point of view

- The distinction between somatic and sex cells will be important for the discussion on evolutionary events, which is important for the comparison of molecular sequences, more later.

# Genes

What are the **genes**?

> *The fundamental physical and functional unit of heredity. A gene is an **ordered sequence of nucleotides** located in a **particular position** on a particular chromosome that **encodes a specific functional product** (i.e., a protein or RNA molecule).*

biotech.icmb.utexas.edu/search/dict-search.html

# Genes

What are the **genes**?

> *The fundamental physical and functional unit of heredity. A gene is an **ordered sequence of nucleotides** located in a **particular position** on a particular chromosome that **encodes a specific functional product** (i.e., a protein or RNA molecule).*

biotech.icmb.utexas.edu/search/dict-search.html

Can be several thousands nt (nucleotides) long.

# Genes

What are the **genes**?

> *The fundamental physical and functional unit of heredity. A gene is an* ***ordered sequence of nucleotides*** *located in a* ***particular position*** *on a particular chromosome that* ***encodes a specific functional product*** *(i.e., a protein or RNA molecule).*

biotech.icmb.utexas.edu/search/dict-search.html

Can be several thousands nt (nucleotides) long.

Occurs on either stand, not often but sometimes overlapping.

# Genome

What is a **genome**?

- **All the genetic material** in the chromosomes of a particular organism needed create and maintain the organism alive.

# Genome

What is a **genome**?

- **All the genetic material** in the chromosomes of a particular organism needed create and maintain the organism alive.

Can be several millions or even billion letters long.

# Genome

What is a **genome**?

- **All the genetic material** in the chromosomes of a particular organism needed create and maintain the organism alive.

Can be several millions or even billion letters long.
Most genomes consists of DNA (deoxyribonucleic acids) molecules.

# Genome

What is a **genome**?

- **All the genetic material** in the chromosomes of a particular organism needed create and maintain the organism alive.

Can be several millions or even billion letters long.
Most genomes consists of DNA (deoxyribonucleic acids) molecules. However, some pathogens (some viruses, viroids and sub-viral agents) are made up of ribonucleic acids (RNA).

# Genome organisation

Without going into to much details, in higher organisms, the genes are broken into subsegments that are called **exons**. The segments are separated by intervening sequences that are called **introns**.

# Genome organisation

Without going into to much details, in higher organisms, the genes are broken into subsegments that are called **exons**. The segments are separated by intervening sequences that are called **introns**.
Genomes are not packed with genes.

# Genome organisation

Without going into to much details, in higher organisms, the genes are broken into subsegments that are called **exons**. The segments are separated by intervening sequences that are called **introns**.

Genomes are not packed with genes.

Human genome organisation.

- Up to 60 % repetitive sequences
  - $\frac{1}{3}$ satellite DNA: low complexity, short and highly repeated
  - $\frac{2}{3}$ complex repeats: transposons, etc.
- Unique sequences;
  - 1.2 % protein-coding
  - 20 % introns

# Genome organisation

- “About one-half of the platypus genome consists of interspersed repeats derived from transposable elements.”
- Genome analysis of the platypus reveals unique signatures of evolution. Nature (2008) vol. 453 (7192) pp. 175-183

# Bioinformaticist's point of view

- **Repetitive sequences** are an **obstacle** for the algorithms involved in **sequence assembly**.
- **Repetitive sequences** are often **linked to diseases**, therefore, the detection of repetitive sequences is in itself an important study.

# Bioinformaticist's point of view

- **DNA Sequencing** (traditional or high-throughput)
- **Gene finding** (stochastic grammatical models)
- **Identifying signals** (pattern discovery)

# Proteome

- The collection of all the proteins is called the **proteome**; and **proteomics** studies the interactions of all the proteins.
- **The proteome is the sum of all the proteins at a given time. Just like the transcritome, the proteome is dynamic.**
- Proteins are the main players in the cell, constituting the structure of the cell, but more importantly by catalyzing most reactions.

# Proteome

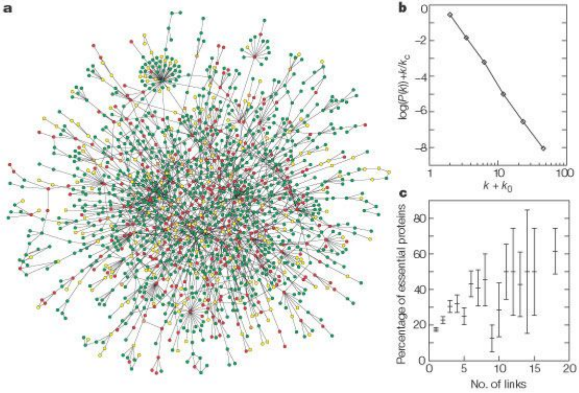
- The collection of all the proteins is called the **proteome**; and **proteomics** studies the interactions of all the proteins.
- **The proteome is the sum of all the proteins at a given time. Just like the transcritome, the proteome is dynamic.**
- Proteins are the main players in the cell, constituting the structure of the cell, but more importantly by catalyzing most reactions.
- "(. . . ) understanding how a genome specifies the biochemical capability of a living cell is one of the major research challenge of modern biology." [2]
- From hypothesis-driven reductionist approach to holistic, data-driven, systems-based approach.

# Interaction networks

- Protein-Protein interactions (PPI)
- Protein-DNA interactions
- Genetic interactions
- Metabolic networks
- Signaling network
- Transcription/regulatory network

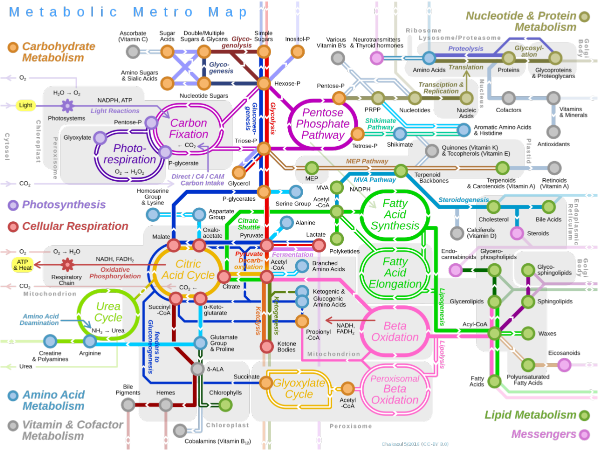
https://en.wikipedia.org/wiki/Biological_network

# Yeast proteome



H. Jeong, S. P. Mason, A.-L. Barabási & Z. N. Oltvai. Lethality and centrality in protein networks *Nature* **411**:4142 (2001)

# Metabolic network



**Source:** `https://en.wikipedia.org/wiki/File:Metabolic_Metro_Map.svg`

# Resources

- https://www.nature.com/scitable/ebooks/cntNm-14749010/
- https://www.nature.com/scitable/topic/genetics-5/
- https://www.khanacademy.org/test-prep/mcat/biomolecules
- https://www.nature.com/scitable/topic/cell-biology-13906536

# References

Wiesława Widła.
*Molecular Biology: Not Only for Bioinformaticians*, volume 8248.
Springer, 2013.

Terence A Brown.
*Genomes*.
Garland Science, 3 edition, 2006.

Marcel **Turcotte**

Marcel.Turcotte@uOttawa.ca

School of Electrical Engineering and **Computer Science** (EECS)

**University of Ottawa**