

Lecture Notes for INF281 Basics of Bioinformatics Sequence Analysis

Takaya Saito



This work is licensed under a Creative Commons Attribution 4.0 International License.

Contents

I	1
1 Introduction	1
1.1 Introduction to Molecular Biology	1
1.2 Introduction to Biotechnology	5

Part I

1 Introduction

1.1 Introduction to Molecular Biology

Molecular biology is the study of biology focusing on organisms and cells at the molecular level.

Five essential facts about cells

1. Two primary types of cells - eukaryotes and prokaryotes

- Eukaryote: animals & plants
- Prokaryote: bacteria & archaea

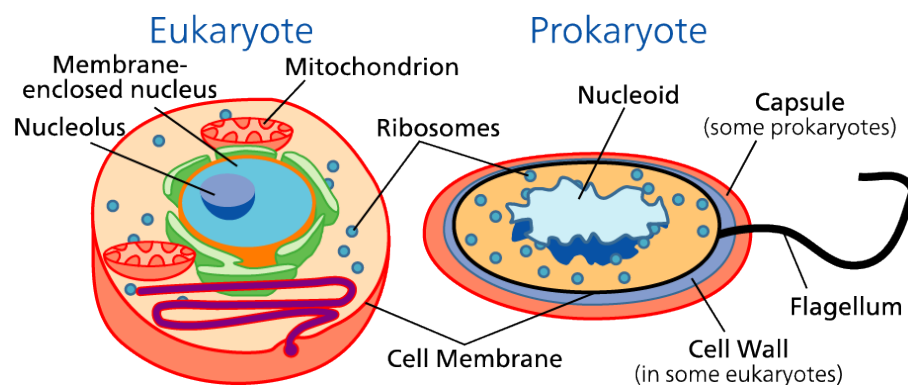


Figure 1.1: Eukaryotic and prokaryotic cells (source: Science Primer, Wikimedia Commons)

2. Cell size - around 1 to 100 micrometers

- Cell Size and Scale: <http://learn.genetics.utah.edu/content/cells/scale>

3. The number of cells

- Prokaryotes: 1 cell
- Human: Estimate of 15 trillion cells

4. An animal cell and cell organelles

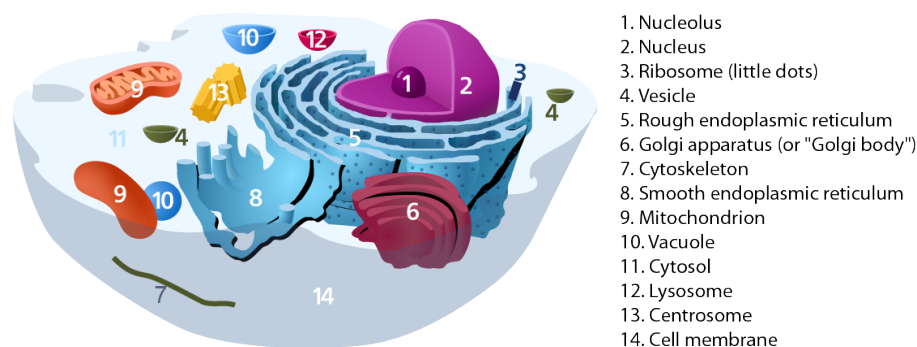


Figure 1.2: An animal cell and organelles (source: Kelvinsong, Wikimedia Commons)

5. Cellular processes

- Cell growth, cell development, cell signaling,
- Example: <http://www.nature.com/nrg/multimedia/rnai>

Central dogma of molecular biology

It describes the information flow within a cell.

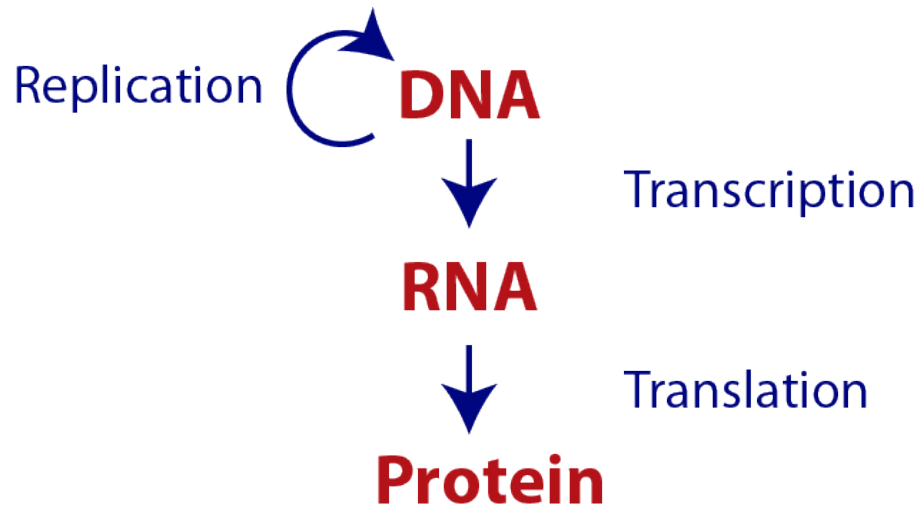


Figure 1.3: Central dogma of molecular biology

DNA (deoxyribonucleic acid)

DNA stores genetic information. It has four different bases: cytosine (C), guanine (G), adenine (A), and thymine (T).

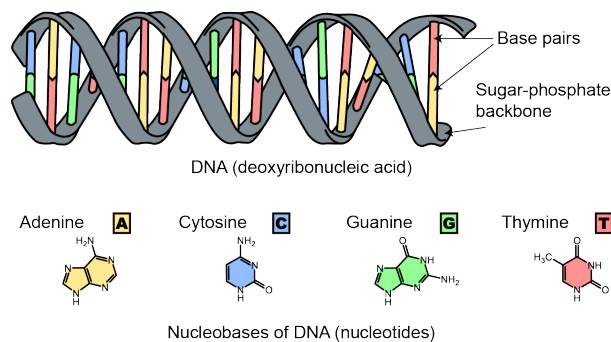


Figure 1.4: DNA double helix and base pairs
(modified from the original version by Sponk, Wikimedia Commons)

Base pair matching (Watson-Crick base pair)

Adenine (A) pairs with thymine (T), whereas cytosine (C) pairs with guanine (G).

```
DNA strand1: ACGT
              ||||
DNA strand2: TGCA
```

RNA (Ribonucleic acid)

RNA has various biological roles and several sub-classes. Messenger RNAs (mRNAs) convey genetic information. It has four different bases: cytosine (C), guanine (G), adenine (A), and uracil (U).

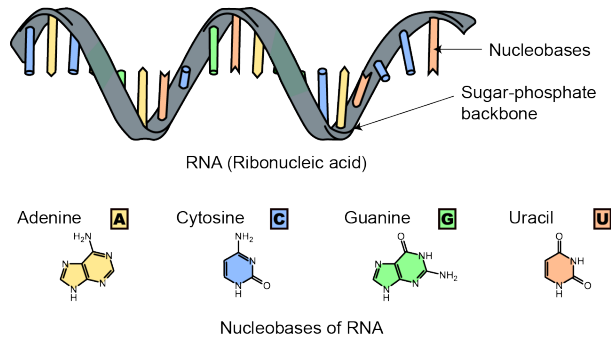


Figure 1.5: Single strand RNA
(modified from the original version by Sponk, Wikimedia Commons)

Transcription: mRNAs are transcribed from DNAs

DNA: ACGT -----> RNA: ACGU
Transcription

Protein

Proteins are large molecules consisting of amino acids. There are 20 common amino acids.

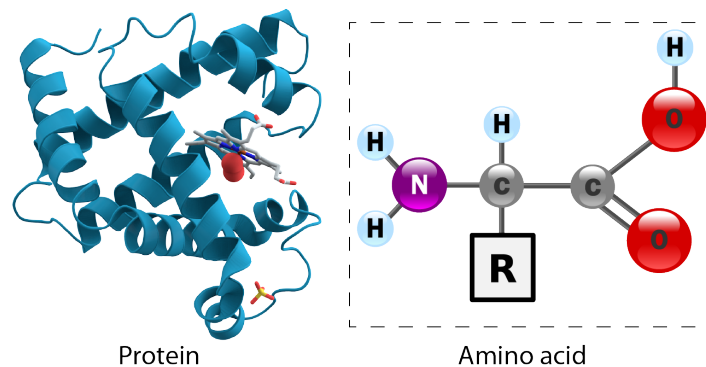


Figure 1.6: Protein 3D structure and amino acids
(sources: AzaToth, Wikimedia Commons, YassineMrabet, Wikimedia Commons)

Translation: Amino-acids are translated from mRNAs

mRNA: GUC -----> AA: Valine
Translation

Universal genetic code

A codon consists of three nucleic acids. Single-letter or three-letter names can be used for amino acids.

		Gentic code				Amino acids			
		2nd base				Basic	Acidic	Polar	Nonpolar (hydrophobic)
		U	C	A	G				
1st base	3rd base in each row	U	UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine			
			UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine			
			UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (Stop)	UGA Opal (Stop)			
			UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (Stop)	UGG (Trp/W) Tryptophan			
	C	CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine				
		CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine				
		CUA (Leu/L) Leucine	CCA (Pro/P) Proline	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine				
		CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine				
	A	AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine				
		AUC (Ile/I) Isoleucine	AAC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine				
		AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine				
		AUG (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine				
	G	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine				
		GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine				
		GUA (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine				
		GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine				

Figure 1.7: Universal genetic code
(modified from the original version by Häggström, Wikimedia Commons)

Cellular functions of proteins

- Enzymes: catalyze chemical reaction
- Cell signaling: hormone (e.g. insulin), antibodies,
- Structural: collagen, cartilage, keratin,

Exercises 1.1

1. Draw a simple diagram of the central dogma of molecular biology and briefly explain the information flow of the molecules.
2. What are the DNA sequences of the opposite strand for the following DNA sequences?

Seq1 CCGATT
Seq2 TTACGC
Seq3 ACGCGC

3. What are the mRNA sequences transcribed from the following DNA sequences?
4. What are the polypeptide sequences translated from the following mRNA sequences?
Answer them with both one-letter and three letter names.

Seq1 AUGUUUUA
Seq2 GCAGCAAAA

1.2 Introduction to Biotechnology

Biotechnology is the use of laboratory techniques to study living organism and cells.

Applications of biotechnology

Branches of biotechnology can be explained with different colors.

- Red: medical processes
- Green: agricultural processes
- White: industrial processes
- Blue: marine and aquatic applications

Laboratory tools and equipment

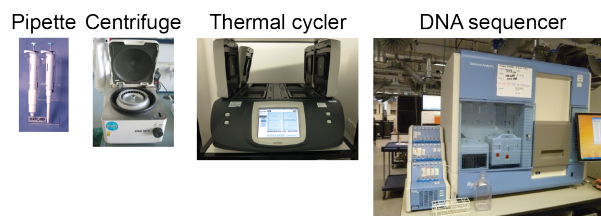


Figure 1.8: Pipette, centrifuge, thermal cycler, and DNA sequencer
(sources: Domain, Manske, Rror, RE73 via Wikimedia Commons)

Human genome project

It was a large-scale international research project to determine the whole DNA sequences of human.

- 1990 - 2003
- \$2.7 billion

Next generation sequencing

Sequence technologies have been rapidly advanced since the human genome project.

Example: sequence a whole human genome with Illumina HiSeq X Ten.

- One day
- \$1000

Protein sequencing

Proteins are generally more studied than DNAs and RNAs, but the whole proteome is generally harder to analyze than the whole genome. MS (mass-spectrometry) based technologies are widely used to sequence proteins.



Figure 1.9: Orbitrap mass spectrometer (source: Wiorkiewicz, Wikimedia Commons)