Lecture Notes for INF281 Basics of Bioinformatics Sequence Analysis

Takaya Saito

Contents

I **1**

1 Introduction **1**
1.1 Introduction to Molecular Biology 1
1.2 Introduction to Biotechnology 5
1.3 Bioinformatics in INF281 6

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

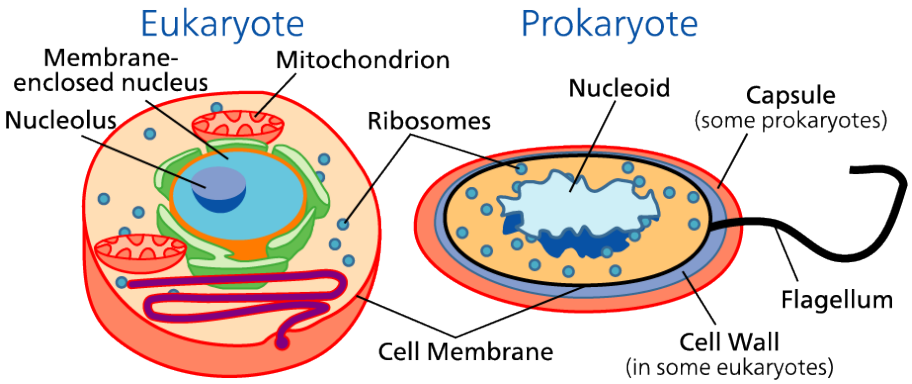

Figure 1.1: Eukaryotic and prokaryotic cells (source: Science Primer, Wikimedia Commons)

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

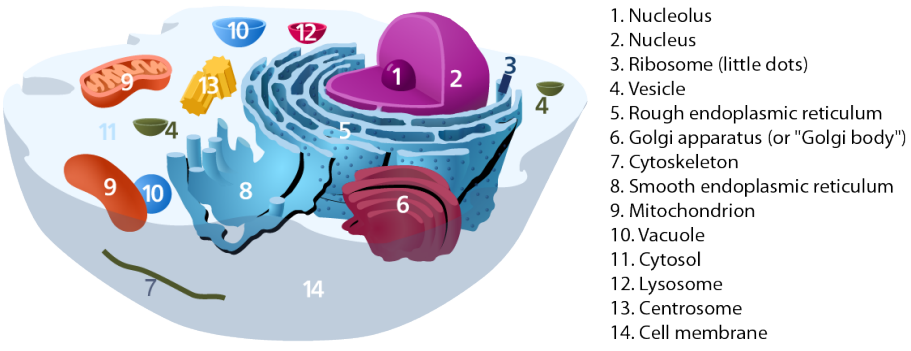

Figure 1.2: An animal cell and organelles (source: Kelvinsong, Wikimedia Commons)

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.

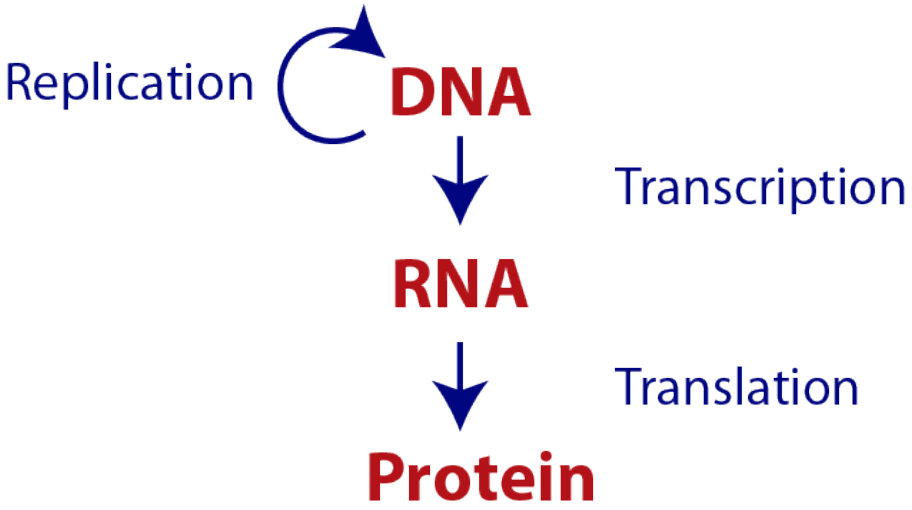

Figure 1.3: Central dogma of molecular biology

DNA (deoxyribonucleic acid)

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

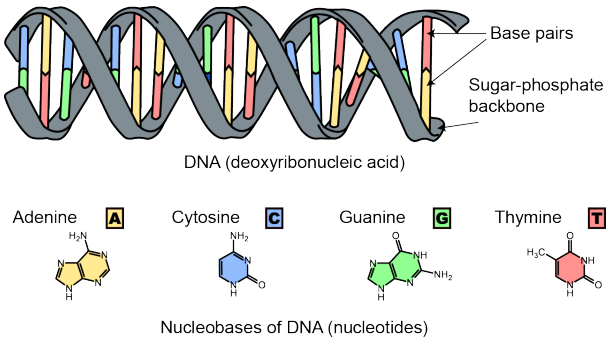

Figure 1.4: DNA double helix and base pairs
(modified from the original version by Sponk, Wikimedia Commons)

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).

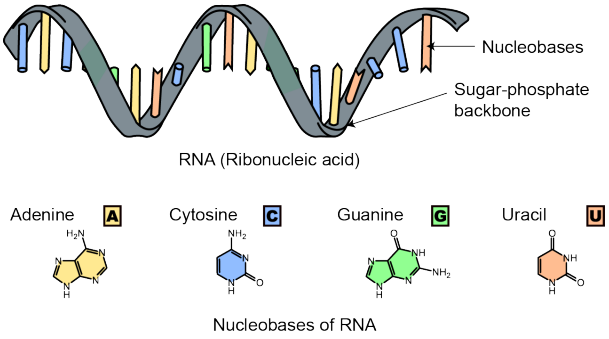

Figure 1.5: Single strand RNA
(modified from the original version by Sponk, Wikimedia Commons)

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.

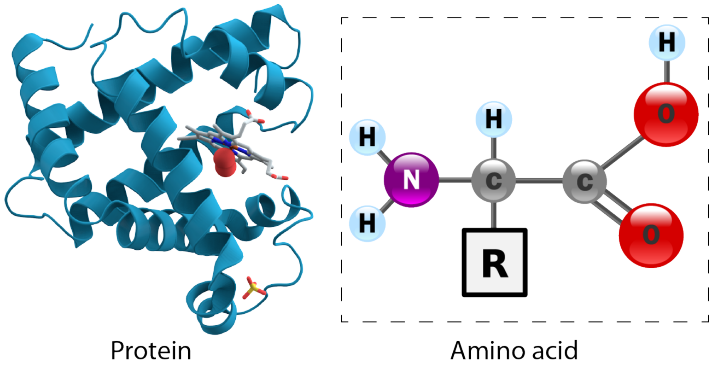

Figure 1.6: Protein 3D structure and amino acids
(sources: AzaToth, Wikimedia Commons, YassineMrabet, Wikimedia Commons)

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.

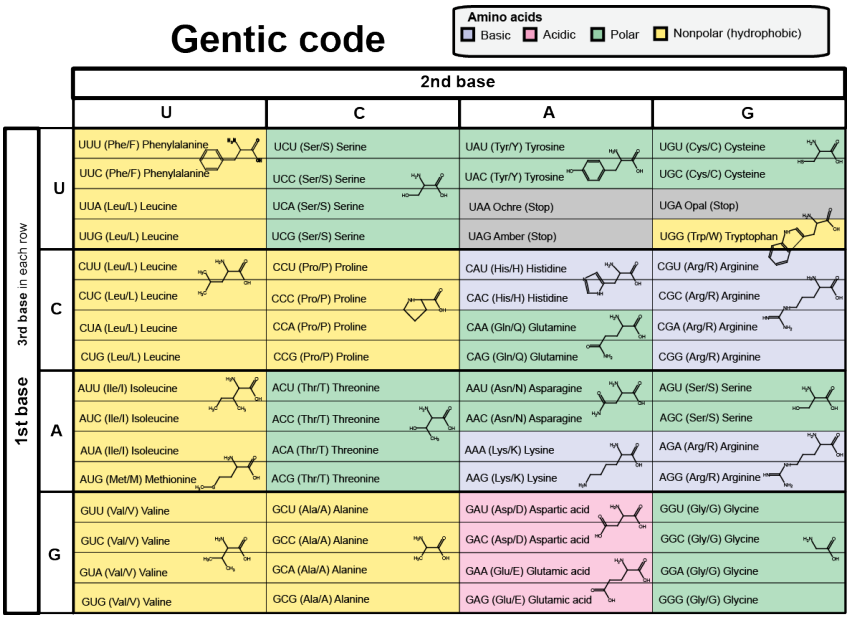

Figure 1.7: Universal genetic code
(modified from the original version by Häggström, Wikimedia Commons)

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 AUGUUUUAA
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.

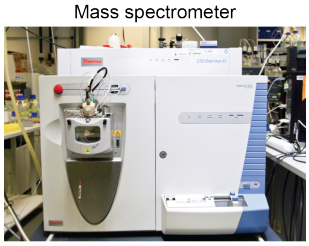

Figure 1.9: Orbitrap mass spectrometer (source: Wiòrkiewicz, Wikimedia Commons)

1.3 Bioinformatics in INF281

Bioinformatics uses computational approaches to solve problems in life sciences. It is based on computer science.

Similar or almost equivalent disciplines

- Biostatistics
- Biophysics
- Systems biology
- Computational biology

Not much related with bioinformatics

- Health informatics
- Forensic science

Scope of INF281

We mainly cover the following fields of bioinformatics in this course.

- Pairwise alignment
- Database search
- Statistical evaluation
- Multiple alignment
- Phylogenetic tree
- Scoring scheme
- Sequence patterns

Popular bioinformatics programs

BLAST and ClustalW are popular tools for sequence analysis.

- BLAST: a program for database search

  URL: `http://blast.ncbi.nlm.nih.gov`

- ClustalW: a program for multiple alignments

  URL: `http://www.ch.embnet.org/software/ClustalW.html`

Rank	Title	Times cited
1	Protein measurement with the folin phenol reagent	305148
2	Cleavage of structural proteins during the assembly of the head of bacteriophage T4	213005
3	A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding	155530
4	DNA sequencing with chain-terminating inhibitors	65335
5	Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction	60397
6	Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications	53349
7	Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density	46702
8	Density-functional thermochemistry. III. The role of exact exchange	46145
9	A simple method for the isolation and purification of total lipides from animal tissues	45131
10	**Clustal W**: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice	40289
11	Nonparametric estimation from incomplete observations	38600
12	**Basic local alignment search tool**	38380
13	A short history of SHELX	37978
14	**Gapped BLAST and PSI-BLAST**: A new generation of protein database search programs	36410
15	A revised medium for rapid growth and bio assays with tobacco tissue cultures	36132

Table 1.1: The 15 most cited papers of all time
(The top 100 papers, Van Noorden, Maher, and Nuzzo, *Nature*, 2014)