Cairo University
Faculty of Computers and Artificial Intelligence
Computer Science Department

Biological Sequence Analysis

Introduction

Dr. Amin Allam

Molecules of *nucleic acids* copy information about fundamental determinants of life. *DNA* is a string where each letter is a *nucleotides* $\{A, C, T, G\}$. A *genome* of a living organism is its entire *DNA*. Living organisms have up to trillions of *cells* (according to the organism type). Each *cell* of the same living organism contains the same *genome*. *DNA* varies in length from a few million *nucleotides* (bacteria) to a few billion *nucleotides* (mammals).

DNA is usually *double-stranded*, with one *strand* being the *Watson-Crick complement* of the other. The *complement* of A is T and the *complement* of C is G. Thus, a letter of *DNA* is usually called a *base-pair*, not a *nucleotide* since, it actually represents two *nucleotide*: one in a *DNA strand*, and one in its *reverse complement strand*.

A *DNA strand* has a start and an end, and such direction from start to end is important. For example, consider the following *DNA*:

```
ATGTC
TACAG
```

Each *nucleotide* is physically bound to its *complementary nucleotide* below it, forming one *base-pair*. Assuming that the start-to-end direction of the top *strand* is from left to right, the start-to-end direction of the bottom *strand* will be from right to left. Normally, we write *DNA* strings or substrings such that it starts at left and ends at right. Thus, the above *DNA* consists of two these two *strands*: `ATGTC` and its *reverse complement*: `GACAT`. When working with *DNA* sequences, both *strands* are equally important, and they are not equivalent.

DNA is a factory of *proteins*. *Proteins* are short strings (few hundred letters) where each letter is one of the *20 amino acids*. Bacteria make around *500* to *1500 proteins*, while *human genome* makes around *100,000 proteins*. Each *protein* is produced by a *gene*. A *gene* is a fragment of the *DNA*. Every three adjacent *nucleotides* of a *gene* produces one *amino acid* letter of the corresponding *proteins*. Three adjacent *nucleotides* are called a *codon*. There are $4^3 = 64$ possible *codons*. Since $64 > 20$, several different *codons* may produce the same *protein*. Also, there is a special type of *codons* called *stop codons* which indicate the end of *protein*.

RNA sequences is usually *single-stranded* and consist of letters from the *4*-sized alphabet of *nucleotides* $\{A, C, U, G\}$. It exists to initiate and regulate *protein* production from scattered *genes*. Also, the *genome* of many viruses is *RNA genome*.

So far we have discussed three types of *sequences*, or *strings*, in which we are interested:

- *DNA* sequences which consist of letters from the *4*-sized alphabet of *nucleotides* $\{A, C, T, G\}$.
- *Protein* sequences which consist of letters from the *20*-sized alphabet of *amino acids*.
- *RNA* sequences which consist of letters from the *4*-sized alphabet of *nucleotides* $\{A, C, U, G\}$.

In this course, we study some data structures and algorithms that facilitate performing important queries on such sequences, such as exact (and approximate) searching (and matching).