WEEK 1 ASSIGNMENT

Data Systems in the Life Sciences – BIOL 51000 | Fall 2 2020

Christina Morgenstern

I. SEQUENCE ALIGNMENT

1. What are (biological) sequences?

The term sequence is used in different contexts, mathematics, musical compositions, literacy or movies and generally refers to a set of things that are arranged in a particular order [1]. In biology, a sequence is described as the one-dimensional arrangement of monomers that are linked to give rise to a polymer. This sequential ordering is known as the primary structure of a biological macromolecule. Biological sequences can refer to nucleic acid sequences, DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), as well as to protein sequences. [2]. Most of the time, when we think of biological sequences, we mean the sequence of a DNA molecule. DNA is composed of nucleotides that contain the four bases adenine (A), guanine (G), cytosine (C) and thymine (T), a sugar (deoxyribose) and a phosphate residue. Individual nucleotides are linked together via phosphate di-ester bonds. The structure of the DNA is double-stranded with two nucleotide chains twisted around each other and hold together using hydrogen bonding. RNA is the second nucleic acid in a cell and serves as information carrier and regulator. In contrast to DNA, RNA is composed of the four bases A, G, C and uracil (U), the latter is used instead of thymine. Structurally, RNA doesn´t form double-stranded molecules, it is single-stranded and folds back in complex three-dimensional structures [3]. When dealing with nucleic acids in bioinformatics, the information encoded is represented as a sequence of the four letters. Some of the information stored in the DNA is transcribed into mRNA and further translated into a sequence of amino acids to yield functional proteins, the work horses of the cell. For proteins, a sequence denotes the linear chain of amino acids. There are 20 (21 if counting the rare amino acid selenocysteine) amino acids in a cell that can be arranged to make up proteins. In both cases, nucleic acids and proteins, the sequence of the monomers is of great importance as misspellings can have an impact on the biological function of the molecule.

2. The importance of sequence alignment

Comparing and thus aligning sequences to each other is of great interest in biology and one of the main tasks of a bioinformatician. The goal of comparing DNA, RNA or protein sequences is to determine regions of similarity that may account for an evolutionary relationship between sequences or a functional and structural conservation. Generally, the more similar two sequences are, the more closely related they are in evolutionary history and the more similar their functions are. Apart from assessing evolutionary relationships between sequences and looking for functional domains, sequence alignments can be used to determine which genes are expressed in a certain cell or tissue by comparing the mRNA with its genomic region. Furthermore, polymorphisms and mutations between sequences can be identified by sequence alignment in order to determine variants for the study and treatment of diseases [4].

3. Utilization of sequence alignment

The search for similarity between two sequences is not an easy problem and several tools have been developed to address this question. In the light of the numerous applications of sequence alignment there is no single solution to an alignment problem. Local alignment, global alignment, synteny detection, multiple alignment, alignment of proteins, nucleotides or non-coding RNA structures and many other variants have arisen through the course of time and mediated by the technological advancement leading to an exponential accumulation of biological sequences [5]. In the global alignment problem, two sequences x and y are aligned and the optimal transformation from one sequence into the other is made out. Substitutions of one letter for another or deletion or insertion of a letter or more letters are made out in an edit process driven by a linear objective function that penalizes insertions, deletions and mismatches.

The local alignment such as the Smith-Waterman algorithm aims at identifying strong local similarities between two sequences. In synteny detection, all sufficiently similar pairs of substrings in two different genomes are being searched. This type of alignment is used for the identification of orthologues, pairs of regions that have evolved from the same region. Several sequences are aligned in a process called multiple sequence alignment. This process is used for example when working with next generation sequencing data and aligning the reads to a reference genome [5].

4. Description of pairwise sequence alignment

Pairwise sequence alignment refers to the process of comparing two sequences against each other and finding the best alignment [3]. Each position is scored in terms of match, mismatch or indels (*in*sertion or *del*etion). The process of comparing and evaluating each position is computationally costly. With the aim of finding the highest score and thus the best alignment, comparison matrices are used and a score for every position defined. A simple scoring scheme defines +8 points for a match, -12 points for a mismatch and -3 points for each gap symbol. The goal of an alignment algorithm is to find the alignment with the highest alignment score and thus the optimal alignment. For the global alignment algorithm, an initialization of the matrix with sequence A horizontally and sequence B vertically. Each letter is assigned one grid position and an initial score starting from -3 and decreasing by -3 for each letter. For each position of the matrix, and starting in the top left corner, three alignment scores are calculated. The first score takes the value to the left of the cell currently in and deduces 3 points. The second value takes the score right above and deduces 3 points. The third score takes the value up and left and compares the letters of the sequence at this position for match or mismatch and adds or deduces 8 or 12 points, respectively. From these three calculations, the highest score is the optimal alignment score for this position. This process is repeated for each pair in the sequence and the total score is calculated.

References:

[1] 'sequence - Wiktionary'. https://en.wiktionary.org/wiki/sequence (accessed Oct. 30, 2020).

[2] 'Sequence (biology)', *Wikipedia*. Apr. 10, 2019, Accessed: Oct. 30, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Sequence_(biology)&oldid=891790302.

[3] H. Husi, *Computational biology*. 2019.

[4] 'Sequence Alignment'. https://bioinf.comav.upv.es/courses/biotech3/theory/sequence_alignment.html (accessed Oct. 29, 2020).

[5] S. Batzoglou, 'The many faces of sequence alignment', *Brief Bioinform*, vol. 6, no. 1, pp. 6–22, Mar. 2005, doi: 10.1093/bib/6.1.6.

II. EXAMPLE FOR SEQUENCE ALIGNMENT CODE

The following pseudocode determines the best (global) alignment between two DNA sequences using the Needleman-Wunsch Algorithm. Considering two sequences: $x_1 \ldots\ldots x_m$ and $y_1 \ldots\ldots y_n$, there are three choices to get the best score *S(i,j)*.

x_i aligns to y_j : $S_{(i,j)} = S_{(i-1, j-1)} + w_{(}x_{i},y_{j)};$ where $w_{(}x_{i},y_{j)}$ is either a score for a match or mismatch
x_i aligns to gap: $S_{(i,j)} = S_{(i-1, j)}$ – gap_penalty (gp)
y_i aligns to gap: $S_{(i,j)} = S_{(i, j-1)}$ – gap_penalty (gp)

1. Declare inputs

# Define two sequences A and B with DNA letter code of defined length as strings

```
seq_x = ‘sequence_of_length_m’
seq_y = ‘sequence_of_length_n’
```

2. Initialize scoring matrix

```
S(0,0) = 0
S(0,j) = -j x gp
S(i,0) = -i x gp
```

3. Main iteration: Filling-in alignment scores

# loop through values *i* and *j*, the indices of rows and columns starting at 1.

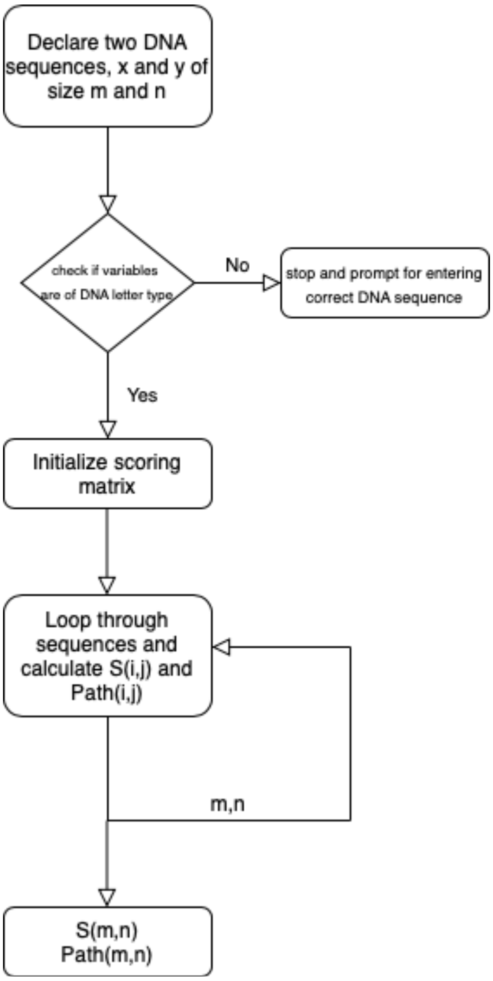
```
for each i=1…m
    for each j=1…n
        $S_{(i,j)} = max \begin{cases} S_{(i-1,j-1)} + w_{(xi,yi)} & [case\ 1] \\ S_{(i-1,j)} - gp & [case\ 2] \\ S_{(i,j-1)} - gp & [case\ 3] \end{cases}$

        $Path_{(i,j)} = \begin{cases} Diagonal, if & [case\ 1] \\ Up, if & [case\ 2] \\ Left, if & [case\ 3] \end{cases}$
```

4. Termination

`S(m,n)` is the optimal score, and from `Path(m,n)` trace back optimal alignment.

Flowchart describing execution of alignment program


Declare two DNA sequences, x and y of size m and n
check if variables are of DNA letter type
No
stop and prompt for entering correct DNA sequence
Yes
Initialize scoring matrix
Loop through sequences and calculate S(i,j) and Path(i,j)
m,n
S(m,n)
Path(m,n)