Assignment 1: Sequence alignment

Míriam Méndez

November 8, 2023

1 Introduction

The document is divided in three sections Global Alignment, Local Alignment, Levenshtein distance. Each subsection shows the results of running the sequences (titled in subsection) with said program (titled in section).

2 Global alignment

2.1 X = "ATCGAT" Y = "ATACGT"

This is the default test we were given when downloading the program and the results were as follows:

```
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ gcc -o global global_alignment.c
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ ./global
Score matrix:
          A    T    A    C    G    T
     0   -2   -4   -6   -8  -10  -12
A   -2    2    0   -2   -4   -6   -8
T   -4    0    4    2    0   -2   -4
C   -6   -2    2    3    4    2    0
G   -8   -4    0    1    2    6    4
A  -10   -6   -2    2    0    4    5
T  -12   -8   -4    0    1    2    6

AT-CGAT
|| || |
ATACG-T

Percent Identity:
  - Divided by the shortest sequence (6): 83.33%
  - Divided by the length of the alignment (7): 71.43%
  - Divided by the average length of the sequence (6.00): 83.33%
  - Divided by the number of non-gap positions (5): 100.00%

Hamming Distance : 2
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$
```

2.2 X = "ACGATAGCGAAACCAAAA" Y = "CACGTAGCCGATGTC"

This test was used to check if the percent identity metric was calculated correctly. It has been obtained the same results found on page 25 from the slides of Lecture 2.

```
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ gcc -o global global_alignment.c
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ ./global
Score matrix:
          C    A    C    G    T    A    G    C    C    G    A    T    G    T    C
     0   -2   -4   -6   -8  -10  -12  -14  -16  -18  -20  -22  -24  -26  -28  -30
A   -2   -1    0   -2   -4   -6   -8  -10  -12  -14  -16  -18  -20  -22  -24  -26
C   -4    0   -2    2    0   -2   -4   -6   -8  -10  -12  -14  -16  -18  -20  -22
G   -6   -2   -1    0    4    2    0   -2   -4   -6   -8  -10  -12  -14  -16  -18
A   -8   -4    0   -2    2    3    4    2    0   -2   -4   -6   -8  -10  -12  -14
T  -10   -6   -2   -1    0    4    2    3    1   -1   -3   -5   -4   -6   -8  -10
A  -12   -8   -4   -3   -2    2    6    4    2    0   -2   -1   -3   -5   -7   -9
G  -14  -10   -6   -5   -1    0    4    8    6    4    2    0   -2   -1   -3   -5
C  -16  -12   -8   -4   -3   -2    2    6   10    8    6    4    2    0   -2   -1
G  -18  -14  -10   -6   -2   -4    0    4    8    9   10    8    6    4    2    0
A  -20  -16  -12   -8   -4   -3   -2    2    6    7    8   12   10    8    6    4
A  -22  -18  -14  -10   -6   -5   -1    0    4    5    6   10   11    9    7    5
A  -24  -20  -16  -12   -8   -7   -3   -2    2    3    4    8    9   10    8    6
C  -26  -22  -18  -14  -10   -9   -5   -4    0    4    2    6    7    8    9   10
C  -28  -24  -20  -16  -12  -11   -7   -6   -2    2    3    4    5    6    7   11
A  -30  -26  -22  -18  -14  -13   -9   -8   -4    0    1    5    3    4    5    9
A  -32  -28  -24  -20  -16  -15  -11  -10   -6   -2   -1    3    4    2    3    7
A  -34  -30  -26  -22  -18  -17  -13  -12   -8   -4   -3    1    2    3    1    5
A  -36  -32  -28  -24  -20  -19  -15  -14  -10   -6   -5   -1    0    1    2    3

-ACGATAG-CGAAACCAAAA
 ||| ||| |||   |
CACG-TAGCCGATGTC----

Percent Identity:
  - Divided by the shortest sequence (15): 66.67%
  - Divided by the length of the alignment (20): 50.00%
  - Divided by the average length of the sequence (16.50): 60.61%
  - Divided by the number of non-gap positions (13): 76.92%

Hamming Distance : 10
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$
```

Note: the hamming distance value has to be ignored, since the two sequences don't have the same length.

3 Local alignment

3.1 X = "ATCGAT" Y = "ATACGT"

This is the default test we were given when downloading the program. The results were as follows:

```
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ gcc -o local local_alignment.c
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ ./local
Score matrix:
          A    T    A    C    G    T
     0    0    0    0    0    0    0
A    0    2    0    2    0    0    0
T    0    0    4    2    1    0    2
C    0    0    2    3    4    2    0
G    0    0    0    1    2    6    4
A    0    2    0    2    0    4    5
T    0    0    4    2    1    2    6

AT-CGAT
|| || |
ATACG-T

Percent Identity:
  - Divided by the shortest sequence (6): 83.33%
  - Divided by the length of the alignment (7): 71.43%
  - Divided by the average length of the sequence (6.00): 83.33%
  - Divided by the number of non-gap positions (5): 100.00%

Hamming Distance : 2
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$
```

3.2 X = "PAWHEAE" Y = "HDAGAWGHEQ"

This test was used to check if the score matrix was calculated correctly. It has been obtained the same results found on page 29 from the slides of Lecture 2. And also the solution can be found on the web site provided by the teacher.

```
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ gcc -o local local_alignment.c
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ ./local
Score matrix:
          H    D    A    G    A    W    G    H    E    Q
     0    0    0    0    0    0    0    0    0    0    0
P    0    0    0    0    0    0    0    0    0    0    0
A    0    0    0    2    0    2    0    0    0    0    0
W    0    0    0    0    1    0    4    2    0    0    0
H    0    2    0    0    0    0    2    3    4    2    0
E    0    0    1    0    0    0    0    1    2    6    4
A    0    0    0    3    1    2    0    0    0    4    5
E    0    0    0    1    2    0    1    0    0    2    3

----PAW-HEAE
     || ||
HDAG-AWGHE-Q

Percent Identity:
  - Divided by the shortest sequence (7): 57.14%
  - Divided by the length of the alignment (12): 33.33%
  - Divided by the average length of the sequence (8.50): 47.06%
  - Divided by the number of non-gap positions (5): 80.00%

Hamming Distance : 8
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$
```

Note: the hamming distance value has to be ignored, since the two sequences don't have the same length.

3.3 X = "ACGATAGCGAAACCAAAA" Y = "CACGTAGCCGATGTC"

This test was also carried out, which showed a large difference in the results compared to those obtained with the global alignment:

```
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ gcc -o local local_alignment.c
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ ./local
Score matrix:
          C    A    C    G    T    A    G    C    C    G    A    T    G    T    C
     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
A    0    0    2    0    0    0    2    0    0    0    0    2    0    0    0    0
C    0    2    0    4    2    0    0    1    2    2    0    0    1    0    0    2
G    0    0    1    2    6    4    2    2    0    1    4    2    0    3    1    0
A    0    0    2    0    4    5    6    4    2    0    2    6    4    2    2    0
T    0    0    0    1    2    6    4    5    3    1    0    4    8    6    4    2
A    0    0    2    0    0    4    8    6    4    2    0    2    6    7    5    3
G    0    0    0    1    2    2    6   10    8    6    4    2    4    8    6    4
C    0    2    0    2    0    1    4    8   12   10    8    6    4    6    7    8
G    0    0    1    0    4    2    2    6   10   11   12   10    8    6    5    6
A    0    0    2    0    2    3    4    4    8    9   10   14   12   10    8    6
A    0    0    2    1    0    1    5    3    6    7    8   12   13   11    9    7
A    0    0    2    1    0    0    3    4    4    5    6   10   11   12   10    8
C    0    2    0    4    2    0    1    2    6    6    4    8    9   10   11   12
C    0    2    1    2    3    1    0    0    4    8    6    6    7    8    9   13
A    0    0    4    2    1    2    3    1    2    6    7    8    6    6    7   11
A    0    0    2    3    1    0    4    2    0    4    5    9    7    5    5    9
A    0    0    2    1    2    0    2    3    1    2    3    7    8    6    4    7
A    0    0    2    1    0    1    2    1    2    0    1    5    6    7    5    5

-ACGATAG-CGAAACCAAAA
 ||| ||| |||   |
CACG-TAGCCGATGTC----

Percent Identity:
  - Divided by the shortest sequence (15): 66.67%
  - Divided by the length of the alignment (20): 50.00%
  - Divided by the average length of the sequence (16.50): 60.61%
  - Divided by the number of non-gap positions (13): 76.92%

Hamming Distance : 10
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$
```

Note: the hamming distance value has to be ignored, since the two sequences don't have the same length.

4 Levenshtein distance

4.1 X = "ATCGAT" Y = "ATACGT"

This is the default test we were given when downloading the program. The results were as follows:

```
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ gcc -o levenshtein levenshtein.c
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ ./levenshtein
ATCGAT
ATACGT

Levenshtein Distance : 2
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$
```

4.2 X = "KITTEN" Y = "SITTING"

This was a dummy test extracted from Wikipedia [1] in order to check that the program has been implemented correctly.

```
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ gcc -o levenshtein levenshtein.c
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ ./levenshtein
KITTEN
SITTING

Levenshtein Distance : 3
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$
```

4.3 X = "AGTCC" Y = "CGCTCA"

This test was used to check if the score matrix was calculated correctly. It has been obtained the same results found on page 5 from the slides of Lecture 2.

```
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ gcc -o levenshtein levenshtein.c
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$ ./levenshtein
AGTCC
CGCTCA

Levenshtein Distance : 3
miriam@MSI:/mnt/c/Users/miriam/Documents/Bioinformatics/lab$
```

References

[1] Levenshtein distance, 2023. https://en.wikipedia.org/wiki/Levenshtein_distance.