Parallel implementation of Sequence Alignment

Final project

Course 10324, Parallel and Distributed Computation

2021 Fall Semester

Sequence Alignment – a way to estimate a similarity of two strings of letters - is an important field in bioinformatics[1]. Sequence is a string of capital letters including hyphen sign (-), for example

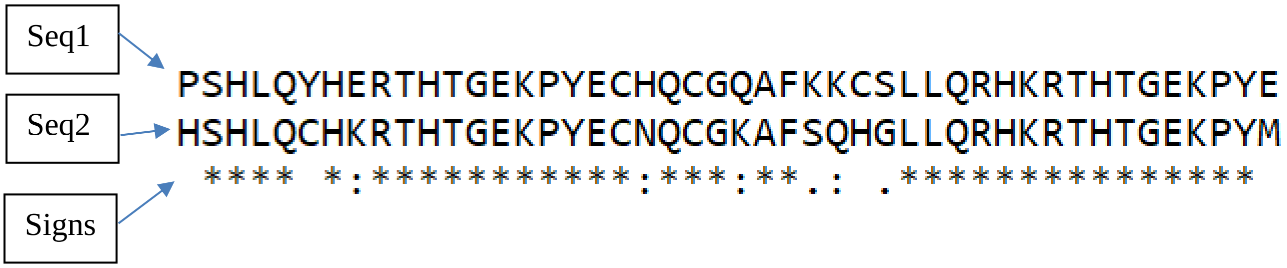
```
PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE
```

Each letter in the sequence represents DNA, RNA, or protein. Identification of region of similarity of set of Sequences is extremely time consuming.

Alignment Score Definition with pair-wise comparison

1. Similarity of two sequences Seq1 and Seq2 of equal length is defined as follows:

- Two Sequences are places one under another:

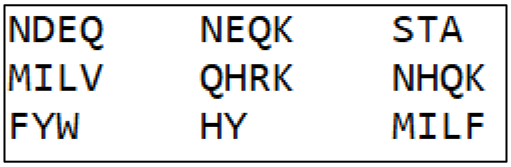


```
PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE
HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYM
 **** *:***********:***:**.: .*************** 
```

- Each letter from Seq1 is compared with the correspondent letter from Seq2. If these letters are identical the pair is marked with Star sign (*).

[1] Mount DM. *Bioinformatics: Sequence and Genome Analysis* (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY., 2004

- Otherwise the additional check is provided. The letters are checked if they both present at least in one of 9 groups called Conservative Groups:

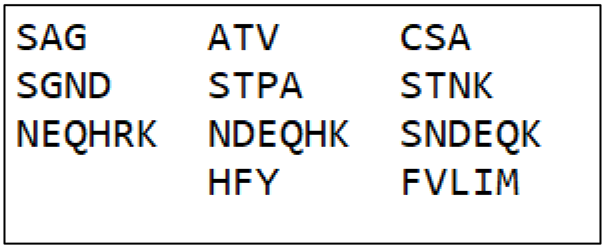


NDEQ	NEQK	STA
MILV	QHRK	NHQK
FYW	HY	MILF

In case that the pair is found in one of Conservative Group it is marked with Colon sign (**:**).

For example, the pair (E, K) is marked with sign **:** because they both were found in group **NEQK**

- If no Conservative Group is found, the pair is checked against 15 Semi-Conservative Groups



SAG	ATV	CSA
SGND	STPA	STNK
NEQHRK	NDEQHK	SNDEQK
	HFY	FVLIM

If the pair do presents in one of Semi-Conservative Groups, it is marked with Point sign (**.**).

For example, the pair (K,S) is marked with sign **.** because they both were found in group **STNK**

- If the letters in the pair are not equal, do not present both not in Conservative nor in Semi-Conservative groups – the pair is marked with Space sign (‘ ‘).

At the end of the check process the whole Sequence of Signs is obtained. This Sequence is used to estimate the similarity of two sequences – Seq1 and Seq2. For this project following formula is used to estimate the Alignment Score:

$S = W_1$*NumberOfStars - W_2*NumberOfColons – W_3*NumberOfPoints - W_4*NumberOfSpaces

where W_i are the given weight coefficients.

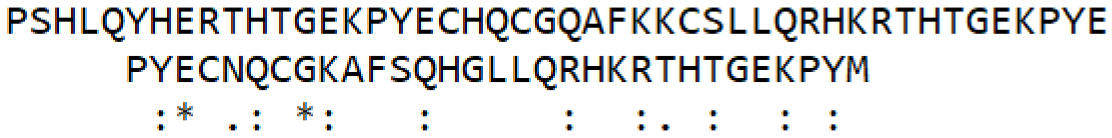
2. Similarity of two sequences Seq1 and Seq2 in case that Seq2 is shorter than Seq1, is defined as follows:

- The Sequence Seq2 is places under the Sequence Seq1 with offset **n** from the start of the Sequence Seq1. The Sequence Seq2 do not allowed to pass behind the end of Seq1.
- The letters from Seq1 that do not have a corresponding letter from Seq2 are ignored.
- The Alignment Score is calculated according the pair-wise procedure described above.

For example, Sequence Seq2 is placed at different offsets under Seq1:

Score = -27

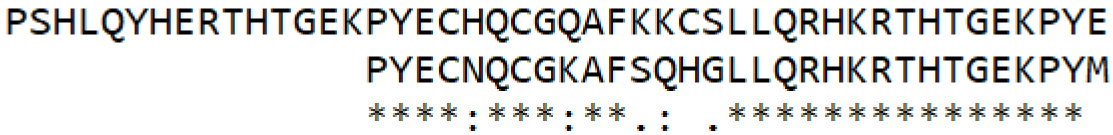
Offset n = 5



```
PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE
     PYECNQCGKAFSQHGLLQRHKRTHTGEKPYM
      :* .: *:   :     :  :. :  : :
```

Score = 17

Offset n = 15



```
PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE
               PYECNQCGKAFSQHGLLQRHKRTHTGEKPYM
               ****:***:**.: .***************
```

Mutant Sequence Definition

For a given Sequence S we define a Mutant Sequence MS(n) which is received by substitution of one or more letter by other letter defined by Substitution Rule.

We will define the Substitution Rules as follows:

1. The original letter is allowed to be substituted by another letter if there is no Conservative Group that contains both letters. For example,
   - **N** is not allowed to be substituted by **H** because both letters present in Conservative Group **NHQK**
   - **N** may be substituted by **W** because there is now Conservative Group that contains both **N** and **W**
2. It is not mandatory to substitute all instances of some letter by same substitution letter, for example the sequence P**S**HL**S**P**S**Q has Mutant Sequence P**F**HL**S**P**L**Q

Project Defitintion:

For two given sequences Seq1 and Seq2, find a mutant of Seq2 and its offset that produce a maximum / minimum Alignment Score.

Requirements

- Implement the Simplified Sequence Alignment algorithm explained in the class (see above). Each Mutant sequence is created from the original sequence Seq2 by substituting only one letter in Seq2 with another letter according to above rules.
- The input file **input.txt** initially is known for one machine only. The results must be written to the file **output.txt** on the same machine. Both files has to be in the same directory with the executable file.
- The computation time of the parallel program must be faster than sequential solution.
- Be ready to demonstrate your solution running on VLAB (**on two computers** if MPI is used).
- **No code sharing between students is allowed.** Each part of code, if any, which was incorporated to your project must be referenced according to the academic rules.
- Be able to explain each line of the project code, including those that was reused from any source.
- **The project that is not created properly (missing files, build or run errors) will be not accepted. The whole project must be uploaded to Moodle.**

Structure of the file input.txt and output.txt

input.txt

The first line of the file contains **W1 W2 W3 W4** for computation of Alignment Score

The second line of the file contains a sequence **Seq1** (not more than 10000 letters)

The third line of the file contains a sequence **Seq2** (not more than 5000 letters)

The forth line contains word **maximum** or **minimum** which defines the goal of the search

output.txt

The first line of the file contains a **Mutant** of the Seq2 that produced an answer to the problem

The second line of the file contains its **Offset** and the **Alignment Score** found

בהצלחה