SULEYMAN DEMIREL UNIVERSITY

Distributed computing CSS440

Parallelization of Smith Waterman Local Sequence Alignment Algorithm

Merekeyev Raiymbek-190103157
Bektursyn Azamat - 190103190

01

Introduction:

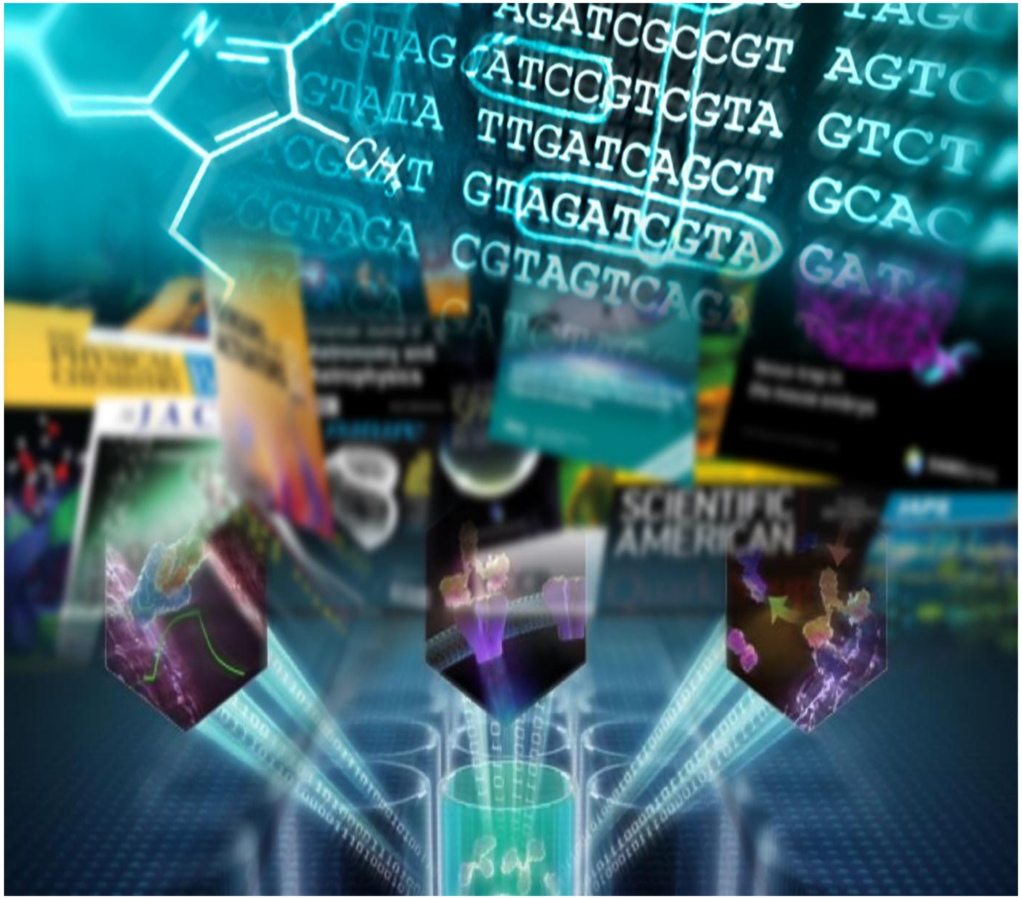
In modern world, there are a lot of organisms that evolved time to time since the beginning of the life on the earth. A lot of diseases have appeared since then. Current COVID-19 can be an example of that. But human beings discovered a vaccine which destroys the life cycle of that viruses and injects to the organism a new antibodies. You may be wondering how scientists discover such things like vaccine. The answer is Genome.

02

How do we calculate and process the Genome?

The genome is a developing field that continuously sets many new challenges for scientists in both biological and computational applications.

Scientists refer to the difference and similarities between virus genes and open ups a new vaccine based on characteristics. This method is called as sequence alignment.

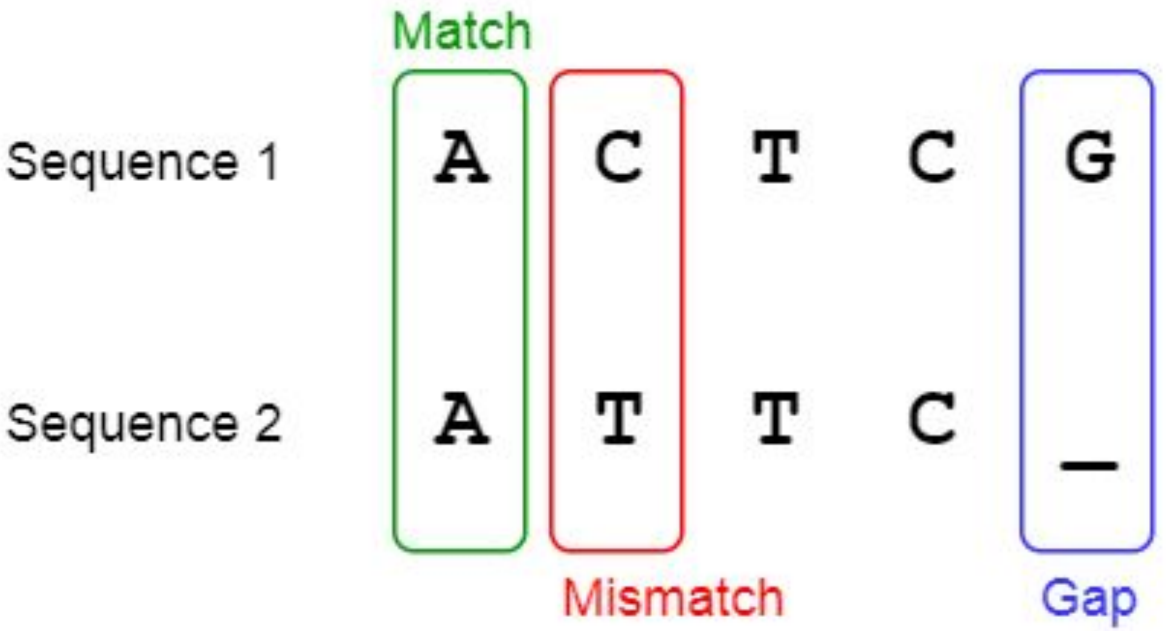


03

What is sequence alignment?

- In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences
- Sequencing is used in molecular biology to study genomes and the proteins they encode. Information obtained using sequencing allows researchers to identify changes in genes, associations with diseases and phenotypes, and identify potential drug targets.
- is used in evolutionary biology to study how different organisms are related and how they evolved
- involves identification of organisms present in a body of water, sewage, dirt, debris filtered from the air, or swab samples from organisms. Knowing which organisms are present in a particular environment is critical to research in ecology, epidemiology, microbiology, and other fields. Sequencing enables researchers to determine which types of microbes may be present in a microbiome, for example.
- As most viruses are too small to be seen by a light microscope, sequencing is one of the main tools in virology to identify and study the virus.
- Medical technicians may sequence genes (or, theoretically, full genomes) from patients to determine if there is risk of genetic diseases.

04



05

Methods of sequence alignments

The most dominant parts of alignments are local and global alignments.

In local alignment we match only parts of whole sequence, whereas in global alignment we try to match entire series. As we know genes of organisms so large to compute and compare to map similarities between them.

Everybody would agree that sequential computing is easier to understand and easier to implement rather than parallel computing, but simply not as efficient as the last one. Nowadays there are a lot of tools that can optimize our work. To deal with efficiency, in this project we'll try to parallelize our solution to get faster result.

06

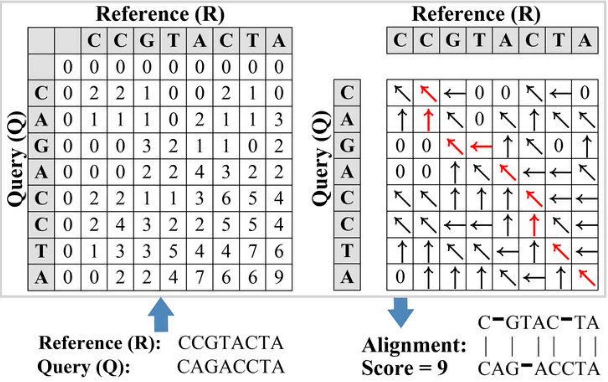
How we are going to implement sequence alignments

In out project we are going to implement Smith-Waterman algorithm(performs local sequence alignment) in sequential and in parallel manner for comparison and analysis. In Parallel manner we will use tools like OpenMP to divide the whole problem into subproblems and give each subprolmes to each processor to compute them. At the final stage, we will compare the results derived from subproblems to obtain global optimal solution.

07

Smith-Waterman algorithm

The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings of nucleic acid sequences or protein sequences. Instead of looking at the entire sequence, the Smith–Waterman algorithm compares segments of all possible lengths.



Reference (R)

Query (Q)

		C	C	G	T	A	C	T	A
	0	0	0	0	0	0	0	0	0
C	0	2	2	1	0	0	2	1	0
A	0	1	1	1	0	2	1	1	3
G	0	0	0	3	2	1	1	0	2
A	0	0	0	2	2	4	3	2	2
C	0	2	2	1	1	3	6	5	4
C	0	2	4	3	2	2	5	5	4
T	0	1	3	3	5	4	4	7	6
A	0	0	2	2	4	7	6	6	9

Reference (R)

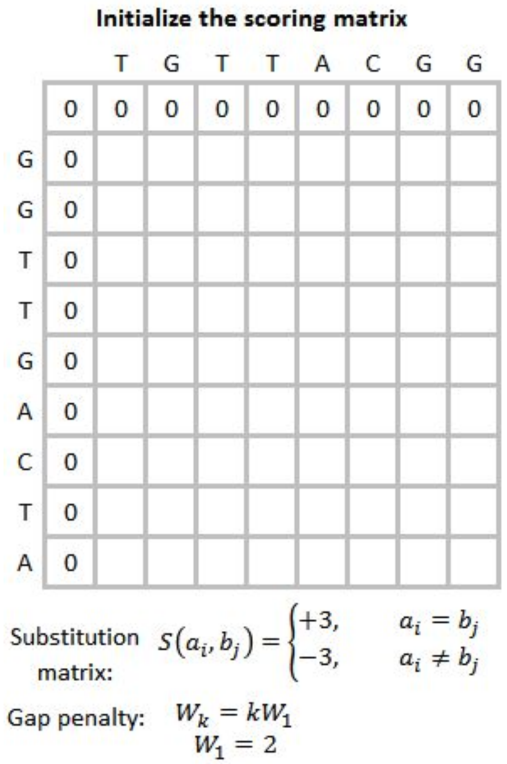
Query (Q)

	C	C	G	T	A	C	T	A
C	↖	↖	←	0	0	↖	←	0
A	↑	↑	↖	0	↖	↑	↖	↖
G	0	0	↖	←	↑	↖	0	↑
A	0	0	↑	↖	↖	←	←	↖
C	↖	↖	↑	↑	↑	↖	←	←
C	↖	↖	←	←	↑	↑	↖	←
T	↑	↑	↖	↖	←	↑	↖	←
A	0	↑	↑	↑	↖	←	↑	↖

```
Reference (R):  CCGTACTA
Query (Q):      CAGACCTA
```

```
             C-GTAC-TA
Alignment:   |  | || ||
Score = 9    CAG-ACCTA
```

Representation of algorithm



Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Substitution matrix: $s(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$
$W_1 = 2$

Keywords: Local alignment, gene sequencing, smith-waterman algorithm, parallelisation, openMP

References:

1.https://www.health.ny.gov/prevention/immunization/vaccine_safety/science.htm

2.https://en.wikipedia.org/wiki/Sequence_alignment

3.https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/sequence-alignment

Thanks for your attention!