

BIOINFORMATICS AND DNA SEQUENCE

ALIGNMENT

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY IN

[INFORMATION TECHNOLOGY]

Submitted by:

UTSAV SINGH

2K20/IT/156

SAMHEL BODH

2K20/IT/127

Under the supervision of:

Dr. Ritu Agarwal



INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(FORMERLY Delhi College of Engineering)

CONTENTS

1. Candidate's Declaration
2. Abstract
3. Acknowledgement
4. Introduction
 - What is Bio-informatics?
 - What is a DNA and its Sequence?
 - What is DNA Sequence Alignment?
 - Applications of DNA Sequence Alignment
5. Problem Statement and Engineering Constrains
6. Solution using Brute Force
7. Solution using Dynamic Programming
8. Screen-shots of Working Program
9. Conclusion
10. Future Scope
11. References

1.Candidates Declaration

We, Utsav Singh (2K20/IT/156) and Samphel Bodh (2K20/IT/127) , students of B. Tech. (INFORMATION TECHNOLOGY), hereby declare that the project Dissertation titled “BIOINFORMATIC AND DNA SEQUENCE ALIGNMENT” which is submitted by us to the Department of INFORMATION TECHNOLOGY, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the CGPA of 4th Semester for the degree of Bachelor of Technology, is a python based program which is used to dynamically align two selected DNA sequences and the code is original and not copied from any source.

Place: Delhi

Date: 17th April, 2022

Signed by:

Utsav Singh
2K20/IT/156

Samphel Bodh
2K20/IT/127

2.Abstract

Bio-informatics is a field of science where we use computational powers and algorithms to analyze data extracted from nature.

One of the major topics of Bio-informatics is DNA sequence alignment which is very important in providing scientists with lot of evolutionary data of different organisms , DNA mutations and can be used to predict different diseases that one is likely to have so that person can take precautions to delay or prevent those diseases.

In our project we will be using Dynamic Programming to align two chosen DNA sequences that can be used for future purposes.

The goal of this project is to create a penalty based DNA sequence alignment program.

3.Acknowledgement

In performing our major project, We had to take the help and guideline of some respected persons, who deserve my greatest gratitude. The completion of this assignment gives me much pleasure. We would like to show my gratitude to Dr. Ritu Agarwal, Mentor for this major project. Giving us a good guideline for my report throughout numerous consultations. We would also like to extend my deepest gratitude to all those who have directly and indirectly guided me in writing this assignment.

Many people, especially my classmates , have made valuable comment suggestions on this proposal which inspired us to improve my assignment. I thank all the people for their help directly and indirectly to complete my assignment. Also, We would like to thank, Department of Information Technology, Delhi Technological University for allowing me to work on this topic.

4.Introduction

BIOINFORMATICS

Bioinformatics is a relatively new branch of scientific studies. It involves using computational power and algorithms to analyze data extracted from nature.

Despite being a new field it has various applications. It is a field which has a significant impact in the world of drug development and the treatment of chronic diseases.

Bioinformatics focuses on developing new technologies in the fields of medicine, research, and biotechnology. This subject is interdisciplinary and requires thorough knowledge of both engineering as well as life sciences. This sector draws from a well of biological data and uses this information to create new tools and software which will be relevant in the world of biological research.

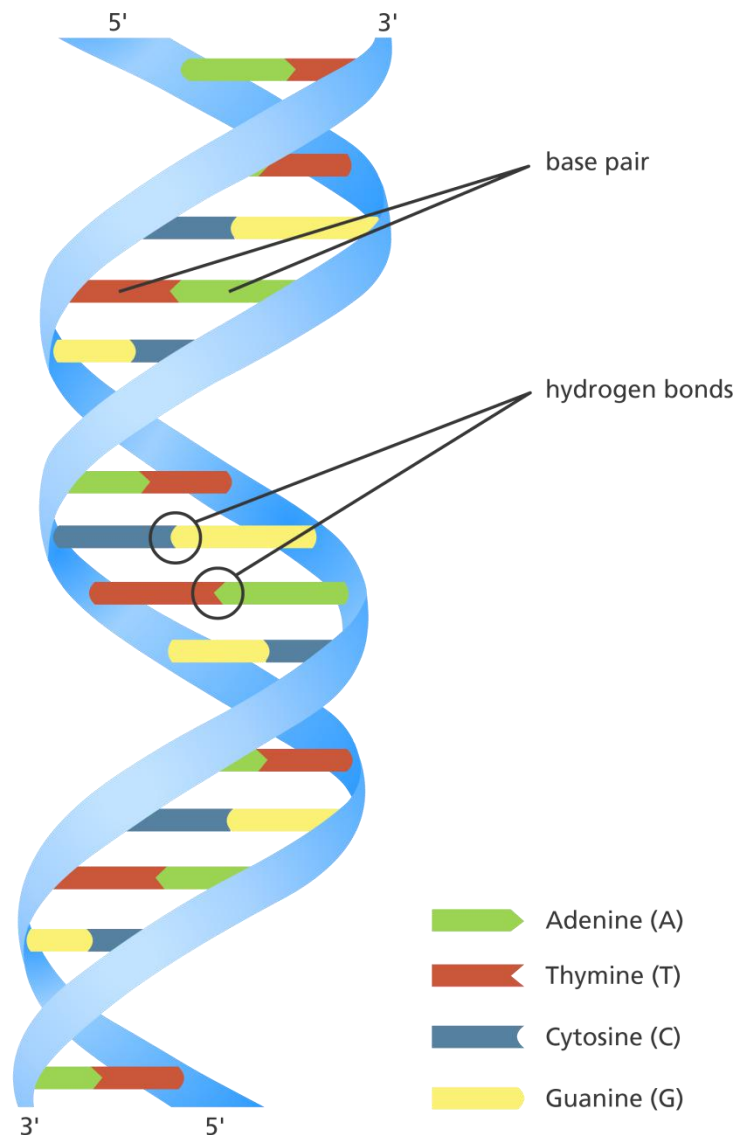
Bioinformatics Applications List

Listed below are the various examples of the application of bioinformatics:

1. We heavily employ bioinformatics in gene therapy.
2. This branch finds application in evolutionary theory.
3. Microbial analysis.
4. Understanding of protein structure.
5. Storage and revival of biotechnological data.
6. In the discovery of new drugs.
7. In agriculture to understand crop patterns, pest control, and crop management.

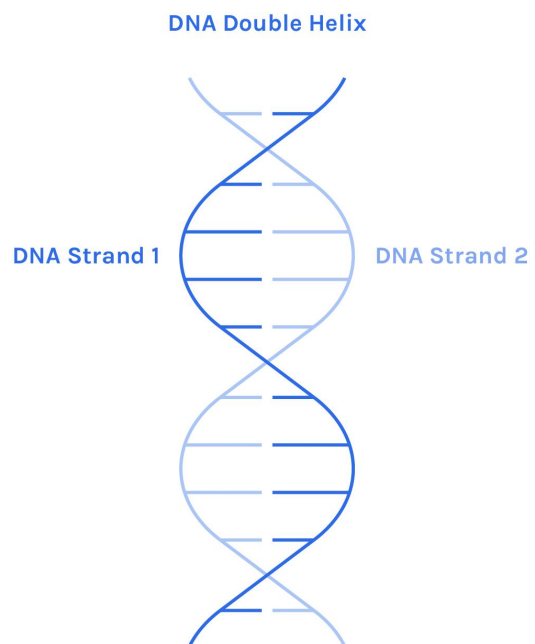
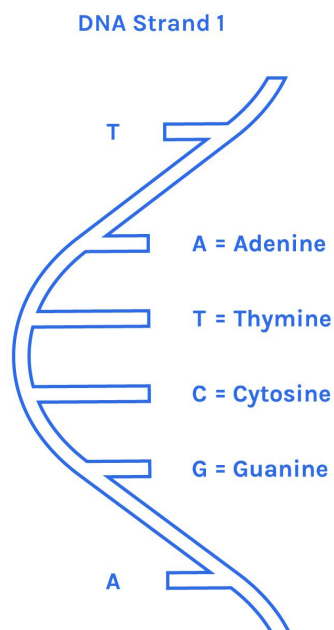
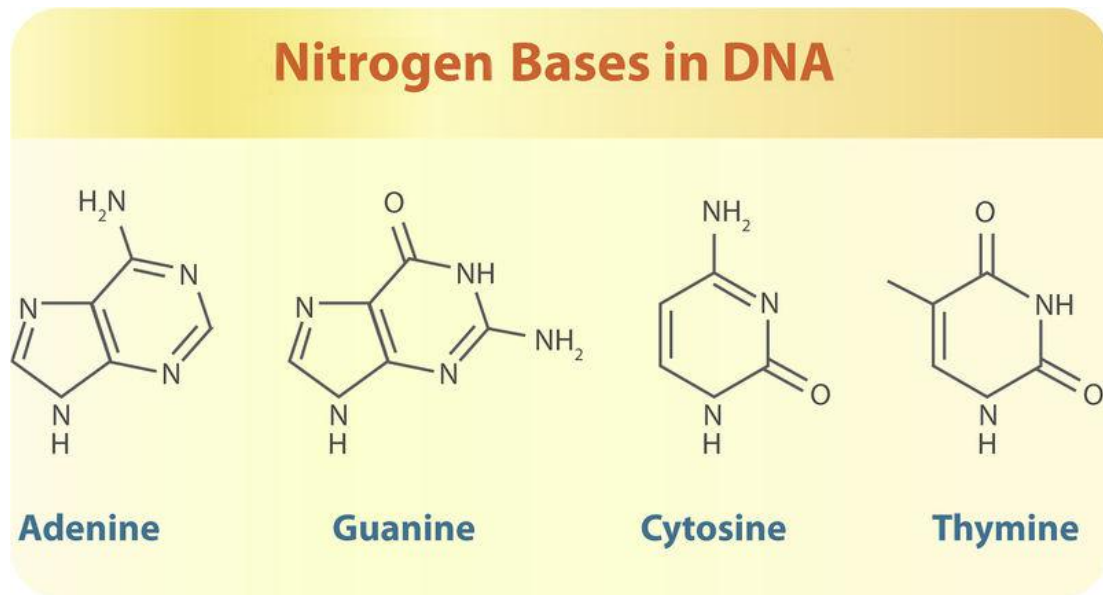
DNA

Deoxyribonucleic acid is a polymer composed of two polynucleotide chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses. DNA and ribonucleic acid are nucleic acids. Alongside proteins, lipids and complex carbohydrates, nucleic acids are one of the four major types of macromolecules that are essential for all known forms of life.



DNA SEQUENCE

DNA sequence consists of a series of nucleotides .
Nucleotides are nitrogenous bases Adenine (A) , Guanine (G) , Cytosine (C) and Thymine (T).



DNA SEQUENCE ALIGNMENT

Evolution of sequences

time= t_0

r TATACATTAG

s TATACATTAG



r T^TATA^GATTAG
 [^]

s TATACATT~~AG~~



time= t

r TATTAGATTAG

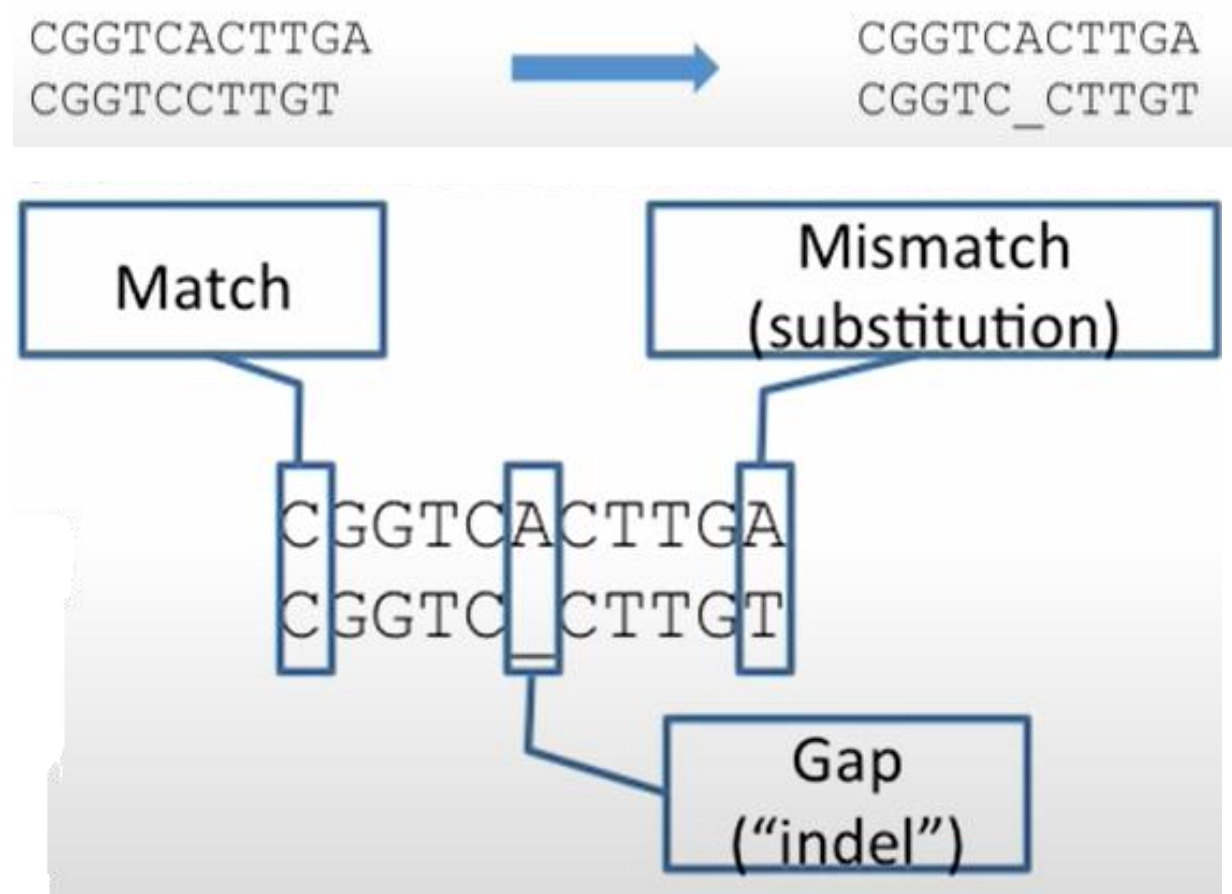
s TATACATTG

To understand the concept of sequence alignment we need to first look at the evolution of DNA sequences, initially at time = t_0 we have two identical sequences *r* and *s* overtime through millions of years there are going to be some mutations and some insertions that will happen and we will end up with two similar but different sequence *r* and *s* at time = t .

In real life case scenario we never know the starting sequence we only have present data and our aim is to find the original sequence from which the current sequence is evolved from .

We know for a fact that genome mutations i.e. changes in the DNA sequence are very very rare and take a very long time to occur naturally.

So when we have two sequences and we are trying to deduce whether they have evolved from the same source we will first try to align them by adding some blank space to account for the mutations and such that the characters in the same position correspond to the same character in the source sequence and both sequence are of equal length with the help of added blanks.



As mentioned before that DNA mutations are rare so when aligning we will try to minimize the the number of indels and mismatch and find out the alignment with the highest alignment score .

The higher the alignment score higher the possibility of the alignment being correct.

We will assign Penalties that will affect the alignment score.

We can set these penalties with any value we want depending on how we want to align our sequence.

For E.g :-

Seq 1 : ACG

Seq 2 : AGG

Penalty Values:-

Match: +1 alignment score;

Otherwise (mismatch or indel): -1 alignment score

Possible Alignments:-

AC _ G

AGG _

Alignment score = -2

ACG

AGG

Alignment score = 1

ADVANTAGES AND USES OF SEQUENCE ALIGNMENT

Alignments are a powerful way to compare related DNA or protein sequences. They can be used to capture various facts about the sequences aligned, such as common evolutionary descent or common structural function.

There are companies such as 23andMe that will analyze a large part of your genome and give a detailed report about your ancestry and characteristics that you have or will have according to your DNA. As written in your genetics, these include hair, eye colour, height etc.

These reports also include various possible diseases that you are likely to have; this information is very helpful because if you are prone to a certain disease or problem according to your genetics, you can take action early and take the required precautionary steps and possibly prevent or at least delay the problem for as long as possible.








5.Problem Statement

We have to create a program that we let us select two difference sequences from the database and align the sequence according to assigned penalties.

Engineering Constrains:-

1. This program must be able to align two selected sequences within reasonable time and print the aligned sequences to screen.
2. Should be able to read the sequence from the .txt files save the user in the sequence folder.

Directory Structure :-

 sequences	16-04-2022 12:27	File folder	
 ADA PROJECT REPORT	18-04-2022 16:49	DOCX File	0 KB
 alignerFunctions	18-04-2022 07:32	Python File	2 KB
 DNasequencer	18-04-2022 07:28	Python File	1 KB
 menuFunctions	18-04-2022 07:32	Python File	2 KB

We are storing all the sequence txt files in the sequences folder if the user wants to add more sequence they can simply add the txt file to this folder.

6.Solution using Brute Force

First obvious way of solving this problem is by using brute force.

We will create a function that will create all the possible alignments for both sequences then we will calculate the alignment score for each combination and compare them.

The combination with the highest alignment score will be the required alignment.

Problem with this solution - DNA sequence can be very long and creating all the possible alignment will take a very long time or might not even be feasible.

Moreover the recursive function will take too much space and we might face stack overflow before the time limit is exceeded.

So we have to solve this problem smartly to reduce the time complexity and improve the recursive function to create a smarter function i.e. this problem will require dynamic programming.

7.Solution using Dynamic Programming

In this solution we will try to use Dynamic Programming to build up the best alignment by using optimal alignments of smaller sub-sequences.

We will first start with the right-most character in both sequences and assuming that the entire sub problem in the left is solved . Therefore we will have three different cases possible (match , mismatch and indel).

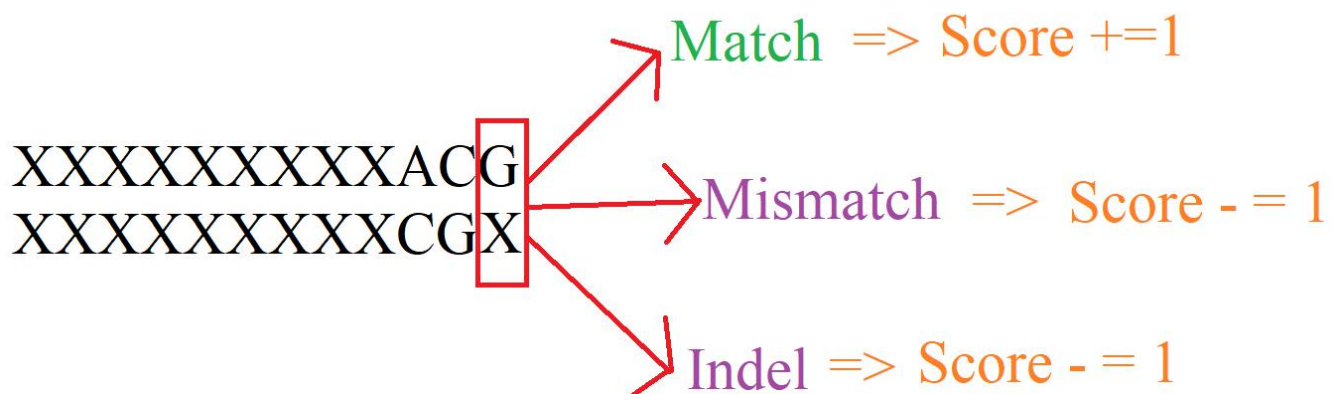
Taking two random sequences as an example :-

Seq 1 : XXXXXXXXXXXXACG

Seq 2 : XXXXXXXXXXXXCGX

X : Any Random Possibility

Looking at the right most



So we will end up with the following algorithm :-

$$\text{Alignment}(i,j) = \max \{$$

$$\quad \text{Alignment}(i-1,j-1)+S,$$

$$\quad \text{Alignment}(i,j-1)-1,$$

$$\quad \text{Alignment}(i-1,j)-1$$

$$\}$$

With this we can draw the matrix :-

Seq 1 : ACG

Seq 2 : AG

		A	C	G	
		-1	0	1	2
Seq 1	A	0	-1	-2	-3
	G	-1			
		-2			

Now we can fill other values :-

		A	C	G	
		-1	0	1	2
A	-1	0	-1	-2	-3
	0	-1	0	-1	-2
	1	-2	-1	-1	-1

After the alignmentMatrix is filled we will trace the path with the lowest cost and get the required alignment.

		A	C	G	
		-1	0	1	2
-1	0	-1	-2	-3	
A 0	-1	0	-1	-2	
G 1	-2	-1	-1	-1	

With this we get the required alignment :-

A	C	G
A	-	G

8.Screen-Shot of Working Program

```
Welcome to the Sequence Alignment Tool

1.Select sequences for Alignment.
2.Close.

Please select your option: 1

Please select sequences from the following choices :-

1.seq1
2.seq2
3.seq3
4.seq4

Please enter the first sequence number :2
Please enter the second sequence number :1
Sequences Before Alignment :-
Sequence 1 : AACGTTCTGATAAGGCGACGGCTGA
Sequence 2 : AAAAATTCTCGGCGTCCGCGCTAA

The two aligned sequences and their alignment score are :-
Alignment 1 : -AACGTTCTGATAAGGCGAC-G-GCTGA
              ||  ||  |  |  ||||  |  ||||  |
Alignment 2 : AAAAATT-TC-TC-GGCGTCCGCGCTAA

The alignment score : 4

The Percentage Match was : 57.14285714285714

Press Enter to return to Main Menu..._
```

9.Conclusion

In this project we went into the details of Bioinformatics and DNA sequence Alignment , we discussed its advantages and uses and were able to implement a program that is able to align two selected DNA sequence in reasonable time.

10. Future Scope

There are different types of sequence alignment , In our project we have implemented pairwise sequence alignment i.e. two sequences are aligned at a time . In the future to we can expand this project to implement multiple sequence alignment in which we align more then two sequence all at the same time.

We would also work on the making a proper UI for the program that is pleasing to the eye and show a better representation of the process of DNA Sequence alignment.

11. References

1. NPTEL VIDEO - https://www.youtube.com/watch?app=desktop&v=vzOoLMCyG4w&ab_channel=NPTEL-NOCIITM
2. Kevin Yip Chinese University of Hong Kong Video - https://www.youtube.com/watch?v=9bCkAsaP_z4&t=1s&ab_channel=Gene%26Tech
3. INFO ABOUT BIOINFORMATICS - [Bioinformatics – Definition, History and Application \(vedantu.com\)](#)
4. GENOME SEQUENCE DATABASE - [Home - Genome - NCBI \(nih.gov\)](#)