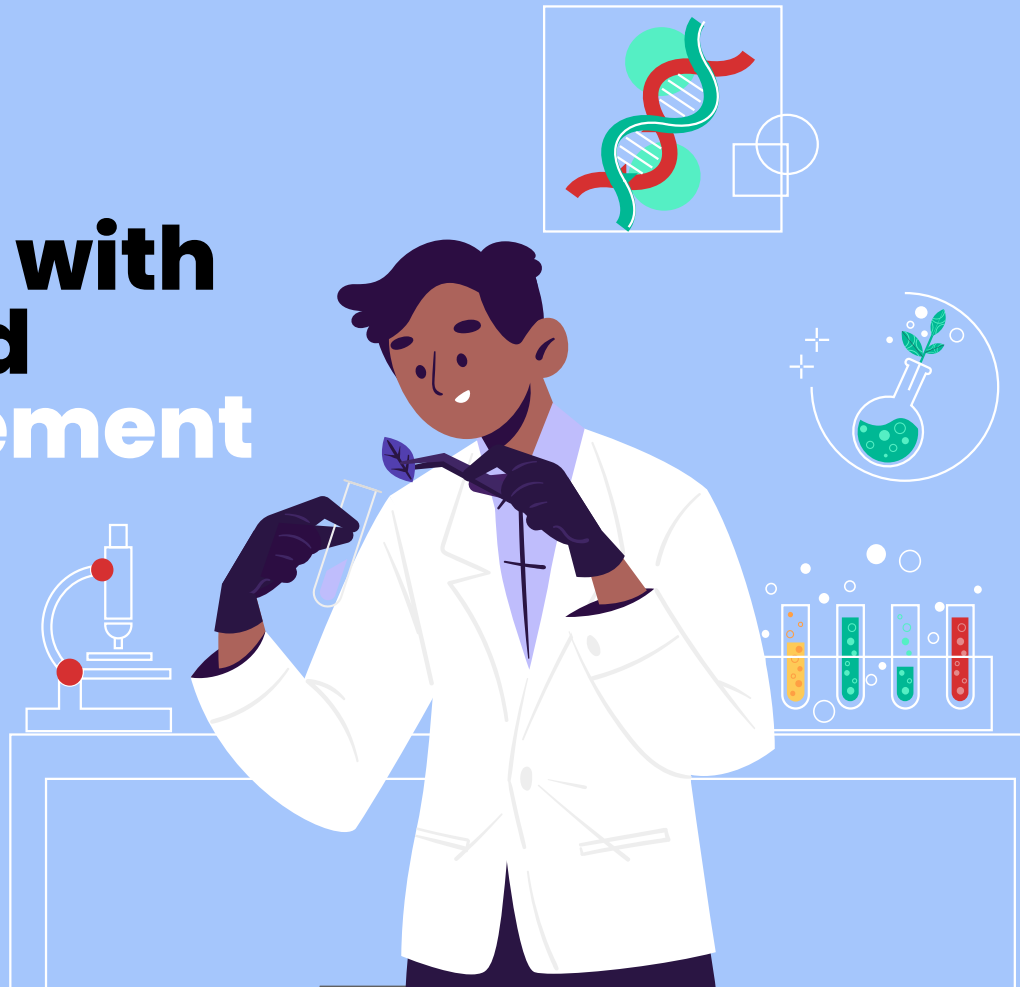
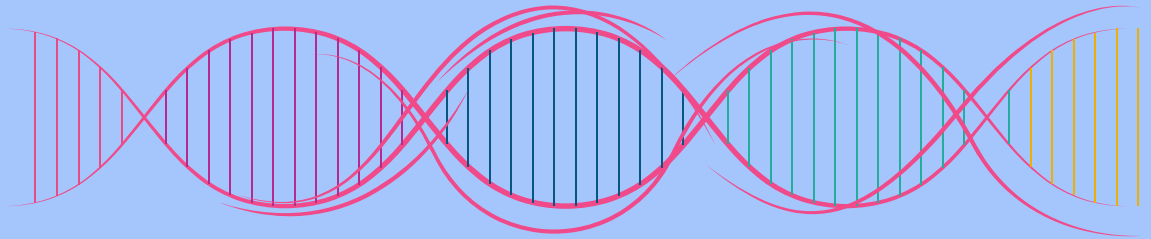
Frequent Words with Mismatches and Reverse Complement Problem

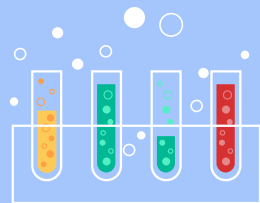
K.SRI SAI HARSHITH



Objective

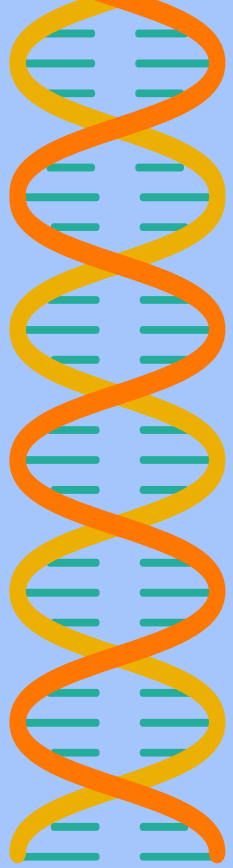


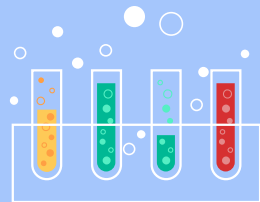
- Our objective is to find DNA A boxes with problem solving algorithm called frequent word with mismatches and reverse complements.
- Biologically, DNA A proteins can't only bind to perfect DNA A boxes but also bind to their slight variations like Mismatches and reverse complement.
- By using the concepts of Reverse complement and finding frequent words with mismatches in a pattern algorithm.
- An input of a DNA sequence and the required hamming distance 'd' and the k value for k-mer are given.
- We can find the DNA A boxes for the given sequence by finding all neighbours of a k-mer and doing their reverse complement.
- Then the output string from the neighbour method will be passed to the frequent words algorithm which then return the k-mers with atmost 'd' mismatches as the output.



Introduction

- Bioinformatics is a field of computational science that has to do with the analysis of sequences of biological molecules. It usually refers to genes, DNA, RNA, or protein, and is particularly useful in comparing genes and other sequences in proteins or any other sequences within an organism.
- DNA replication is a process by which DNA makes a copy of itself.
- The region where DNA replication begins is known as oriC.
- DNA A protein is a factor that promotes the unwinding of DNA in the oriC.
- The initiation phase of DNA replication is determined by the concentration of DNA A protein.
- Replication begins with active DNA A binding to the upstream of oriC called DNA Box.
- Binding of DNA A to DNA A Box leads to strand separation.
- Locating oriC represents an important task for a various biomedical problem.





Methods

For finding DnaA boxes that bind to their slight variations like Mismatches and reverse complements, we redefine the Frequent Words Problem algorithm to account for both mismatches and reverse complements.

1. **Neighbours:** Generates all k-mers with at most ‘d’ mismatches from given pattern.
2. **Reverse Compliments:** Basically takes the complement of each nucleotide (‘A’, ‘C’, ‘G’, ‘T’) in pattern, then reverses the resulting string. We find the reverse compliments of the neighbours.

   If, Pattern = p1 ⋯ pn

   Then its reverse complement will be, (Pattern)’ = pn’ ⋯ p1’
3. **Hamming Distance:** The number of mismatches between two strings, p and q. From step 2 we make sure that the k-mers have at most ‘d’ hamming distance with the original DNA sequence given
4. Then we calculate the most frequent k mers and print them along with its reverse complement

Results

```
1  GAGTCTAACTTTCCTGAGTCTAGCCCTGTGCTACAAGCCCTGTACTTTCCTGAGTCTAACTTTCCTGAGTCTAAACTGAGCACTTTCCTGCCCT
   GTGCTACAAAACTGAGCAACTGAGCGCCCTGTGAGTCTAGCTACAAGCTACAAACTTTCCTGCTACAAGCCCTGTGCCCTGTAACTGAGCGCTA
   CAAACTTTCCTGCTACAAGCTACAAAACTGAGCGCCCTGTGAGTCTAGCCCTGTACTTTCCTGCTACAAGAGTCTAGAGTCTAGCCCTGTAACT
   GAGCACTTTCCTACTTTCCTGCTACAAAACTGAGCGCTACAAACTTTCCTGCCCTGTGAGTCTAGCTACAAGAGTCTAAACTGAGCGCTACAAG
   AGTCTAGCCCTGTGCCCTGTAACTGAGCGCTACAAAACTGAGCACTTTCCTGCTACAAGCTACAAGAGTCTAAACTGAGCGCTACAAGCTACAA
   AACTGAGCGCTACAAGCTACAAAACTGAGCGCCCTGTGCTACAAGCCCTGTAACTGAGCGCCCTGTAACTGAGCAACTGAGCGAGTCTAGCTAC
   AAGAGTCTAGCCCTGTACTTTCCTGCTACAAGAGTCTAGAGTCTAACTTTCCTGAGTCTAGAGTCTAACTTTCCTGCTACAAGCCCTGTACTTT
   CCTGCTACAAAACTGAGCGCCCTGTGCCCTGTGAGTCTAACTTTCCTGCTACAAACTTTCCTACTTTCCTGCTACAAGCTACAAGCCCTGTGCT
   ACAAACTTTCCTGCCCTGTGCCCTGTACTTTCCTACTTTCCTGAGTCTAGCCCTGTGCTACAAGCTACAAGAGTCTAGCCCTGTGAGTCTAAAC
   TGAGCAACTGAGCGAGTCTAGCCCTGTGCCCTGT
2  5 3
```

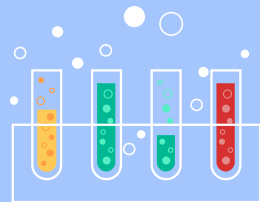
Input:

When an input of the DNA sequence along with the length of the k-mer ‘k’ and the number of at most mismatches ‘d’.

Output:

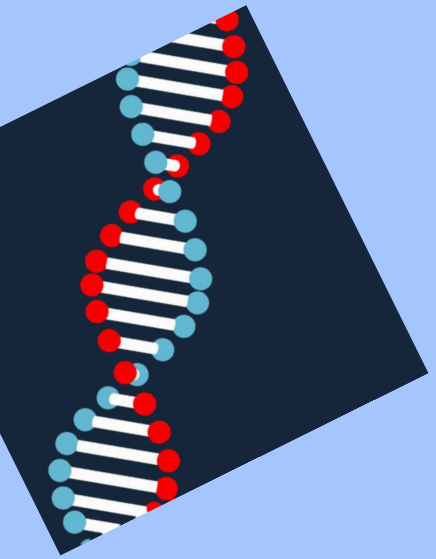
The DnaA boxes obtained from the input DNA sequence, “AATAC” is the most frequent 5-mer with 3 mismatch, along with its reverse complement “GTATT” is obtained as the output.

```
AATAC GTATT
```



Conclusion

By improvising the frequent words problem, i.e. the Frequent Words with Mismatches and Reverse Complement Problem, we were able to locate DnaA boxes of the DNA sequence. As a result of derived DnaA boxes, we obtain the most frequent k-mer with ‘d’ mismatch, along with its reverse complement.



Thank You