DNA stands for Deoxyribonucleic Acid. DNA has a unique double helix shape, like a twisted ladder.

What are DNA Motifs?
- DNA motifs are recurring short segments with expected length of 8-30 nucleotides. Motifs may occur with mutations.
- They often indicate binding sites for proteins such as transcription factors (TF).
- Motifs or Transcription Factor Binding Sites (TFBS) located upstream of genes.

Motif Finding Problem
The planted (l, d) Motif Finding Problem can be described as follows:

**Input:**
1. Set of t sequences \( s_1, s_2, ..., s_t \) of length n over alphabet \( R = \{A, C, G, T\} \).
2. Two integers \( l \) and \( d \) where \( 0 \leq d < l \leq n \).

**Output:**
1. Consensus motif (original motif without mutations).
2. All subsequences (motifs) of length \( l \) at hamming distance \( d \) from the consensus motif.

Motif Finding Algorithms

**Exact Algorithm**
Consumes a long time. Always obtains correct solution.

**Examples:**
- WINNOWER \[2000\]
- SP-STAR \[2000\]
- MITRA \[2002\]
- PMS series \[2005\]
- RISOTTO \[2003\]

**Approximate Algorithm**
Faster. Does not necessarily return the correct solution.

**Examples:**
- MEME \[1994\]
- CONSENSUS \[1993\]
- PROJECTION \[2001\]
- MULTIPROFILE \[2002\]
- Pattern Branching \[2001\]

Computational Results
- Total number of \( d \)-neighbors \( N \) for a given subsequence \( x \) of length \( l \) with allowed mutation \( d \) is \( |N(x,y)| = \sum_{d=0}^{d} \binom{n}{l}\binom{l}{d} \)
- If \( A, B \) are two strings of length \( l \) with \( d(A, M) \leq d \) and \( d(B, M) \leq d \) then \( d(A, B) \leq 2d \) and expected Hamming distance \( E_{d} = 2d \)
- In this research, matrix profile for neighbors can be calculated directly by applying equations (1 & 2) without the need to generate the whole set.
- Simple example: suppose \( x = "AAA" \) with \( d=1 \) (\( x \) is character present in \( x \) and \( O \) is all other characters)

**Proposed Algorithm**

```
Input
M: consensus motif
n: length of each DNA sequence
d: maximum allowed mutations
t: number of DNA sequences
DNA[t][n]: t\textsuperscript{th} DNA sequence

Output
Consensus motif

Algorithm
For each subsequence \( S \) of length \( l \) in the first sequence \( DNA[0] \) from starting positions 0 to \( n+1 \)
If \( S \) is found in all other sequences within specified distance range then add it to the Nominated Motifs
Else Ignore \( S \)

Conclusions and Future Work
- Implement the computational results in the designed algorithm to make it faster.
- Compare it with other approximate algorithms to test its performance.
- Test it on real DNA datasets such as TRANSFAC dataset.