**ADVANCED DATA STRUCTURES PROJECT**

**PROJECT TITLE:** DNA SEQUENCING USING SUFFIX TREES

**PROJECT BY:** SHANIA JOB (19PD32)

S SUBHIKSHAA (19PD36)

**ABSTRACT:**

The human DNA is a combination and repetition of the bases (ATGC) and it has the length of around 3.5 billion characters.

Therefore, analyzing such a long string demands a well thought of data structure in order to find patterns accurately in a short period of time.

The suffix tree data structure, which basically represents all the suffixes, is built in linear time (using Ukkonen's algorithm).

Our project 'DNA SEQUENCING USING SUFFIX TREES' aims to accomplish the following:

1. Given a certain combination of compounds in a genome(string) find if there exists a certain sub combination (a particular substring) within the main one (Pattern matching).

2. Given a certain combination of compounds in a genome and finding if there are repeats of a sequence(substring) is present within the genome(string).

3. Finding the longest repeated substring. This process involves finding the longest sequence(substring) in the given genome(string).