Final Report

(Submitted by- Shikha Saini)

Area of Research- is application of dynamic graphs in Pan-genomics. Pangenome is a term referring to whole set of genes within a species and is used in the field of molecular biology and genetics.

When it comes to 'Bioinformatics'- solving problems of biology using software tools and Information Technology, all these genetic information is represented in the form of graphs. 'Genome Assembly' is one such problem of this field, in which many small fragments of DNA sequences are combined together to form original chromosomes.

In this process a problem is to find out parts that are definitely needed to be present in the solution, called safe.

Problem Statement- Our problem is to dynamically verify the safety of a given walk for an edge covering walk of a strongly connected graph under edge insertions.

Progress- Firstly, we studied about the problem from research papers for proper understanding of the problem. For incremental verification of safe walks, our intent was to use 'The Hydrostructure' model as introduced in – 'The Hydrostructure: A Universal Framework for Safe and Complete Algorithms for Genome Assembly'.

I studied the model and then implemented (in C++) *the static algorithm* to compute 'the Hydrostructure' for a given walk in strongly connected graph, using the definitions of Cloud, River, Vapor and Sea as provided in the paper. Computation takes linear time of O(m) where, 'm' stands for number of edges in the graph.

After this, we have implemented the *trivial dynamic algorithm* for incremental verification of safety of walk under edge insertions. After each update, it computes the Hydrostructure and then verify for its safety by checking whether it is bridge-like case (when Vapor(W) is an open path and river is non-empty) or avertible case, giving an overhead of O(m) time complexity. It takes overall about $O(m^2)$ where, 'm' represents number of edges in the graph.

Then, we developed an improved *dynamic algorithm* that incrementally verify the safety of the specified walk in strongly connected graph using the components of the Hydrostructure, which can be maintained in overall linear time- O(m) where, 'm' stands for number of edges in the graph.

Approach used- From Hydrostructure model, we know that new edge can be inserted in ${}^4P_2$ ways (i.e., 16). We observed that among those sixteen cases, edge insertion in nine cases is safe and in three cases, it is unsafe. For remaining four cases, walk is safe however the Hydrostructure is modified by adding new reachable elements to sea or cloud.

Testing- I have tested the algorithm on a graph having 1000 nodes and following records were recorded on comparing trivial dynamic and dynamic algorithm: 13ms, 15ms, 17ms and 6ms, 8ms, 9ms respectively.

I thank my supervisor, Dr. Shahbaz khan, for his constant support and guidance in this project.