SUMMARY

Summary

The colossal of short fragments of DNA sequences known as *Reads* generated after the process of DNA sequencing cannot be left scattered on several machines due to vast amount of efforts and funds required to generate these sequences. Hence, it is required to store this DNA sequencing data in a secured form such that one can refer to it easily.

This research work involve the development of Web-based application with Distributed approach for storing the actual data in raw form on a central File Server as well as store the sequencing data and its analysis data in the structured format on a Database server. This Database Server and File Server are invoked by the client from remote machine. The client never accesses the Database or File Server directly but, instead this request is managed by the Web-Server. The client invokes the Web-Server which validates the request and forwards this request, if appropriate, to the Database or File Server depending on the type of request. The four components of the Web-application, that is, the client, the Web-server containing the business logic, the Database Server containing the structured data and the File Server containing the raw data are all on different machines across the Internet/Intranet, and the entire communication is happening transparent to the user. Each of these components may be having heterogeneous computational environment. Thus, Distributed computing is involved in storage and retrieval of DNA sequencing data and its analytical data.

To further enhance the computational capacity in terms of storage and processing, ***three new algorithms have been developed as a part of this***

research work. The developed algorithms use Haar Wavelet Transforms, a signal processing approach to deal with the DNA sequences. The algorithms are developed to take care of three issues in BioInformatics such as:

- Data Reduction for Analysis or Transmission of DNA sequences
- Finding identical reads from a DNA sequencing data
- Recognizing Short Tandem Repeats in DNA sequences

The data reduction of DNA sequences is achieved upto 64 times using the developed algorithm, which is a good result for DNA sequences. The search for identical reads is achieved for exact and near exact DNA reads using this newly developed algorithm using signal processing. The Short Tandem repeat regions can be identified for tetra-mers, the results of which are comparable with the other repeat finding algorithms, even without the need for supplying several parameters like repeat pattern or reference sequence etc. Thus, the three issues handled using signal processing approach and Haar Wavelet Transforms have proved to be efficient in terms of processing time as well as space. All the algorithms have proved to work in linear time of computational complexity, excluding other overheads.

The tasks undertaken as part of this research, the algorithms developed and the results acquired, all been published, either in International Journals or in International Conference Proceedings.

The research in the field of Distributed Computing and BioInformatics is still evolving, hence a great deal of work is in progress, and a lot more need to be accomplished.