Abstract

In recent years, development in molecular biology and innovations in DNA sequencing technology has led to large scale sequencing of genomes, metagenomes and transcriptome. As a result sequencing data is growing exponentially. This mounting data needs secured storage and frequent analysis to resolve several mysteries of nature. Although, cost of storage capacities in computers is reducing and processing power enhancing with time, there still arises the need of storing Big-data of genomics efficiently and processing various analysis techniques optimally.

The research work presented in this thesis comprises of two aspects of computing:

1) Using distributed computing for secured storage, retrieval, analysis of bioinformatics data

2) Applying signal processing for optimizing pattern matching in bioinformatics applications.

In the **first portion of research work**, the distributed computing concept has been applied to Bioinformatics applications. The Web-based application is developed to store and retrieve DNA sequencing data using resources distributed across the network. The effort in this research work is to use distributed computing for secured storage of DNA sequencing data and process this data for further analysis.

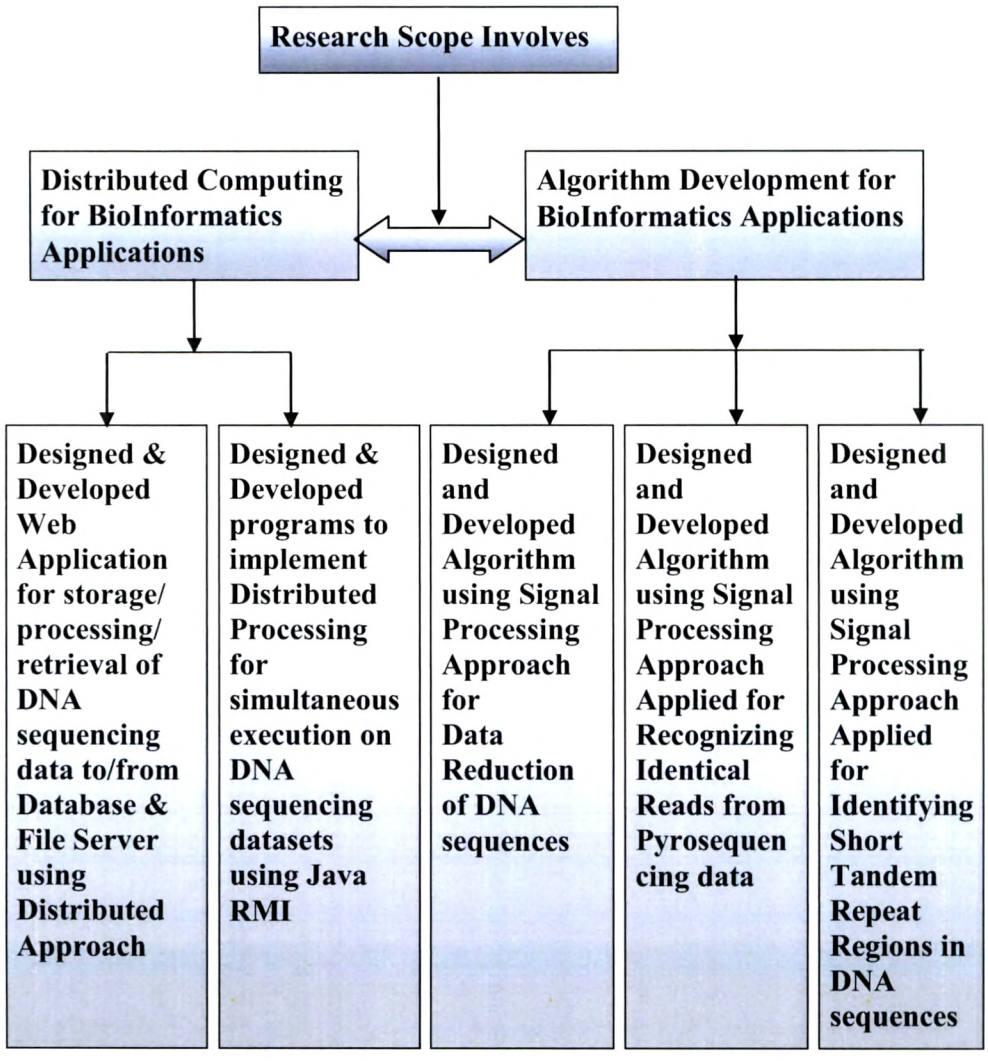

Figure 1. Scope of Research Work

To meet all these challenges, the research work discussed in the thesis, involves the development of Web-based application for storage and retrieval of data into the Database. The key features of this application are that, the application uses distributed approach, where the client, the Web-Server and the Database Server all are located on different computers. The client submits the request; the Web-Server provides the interface to the client, it also acts as a controller and contains the business logic; the Database Server is used to store DNA sequencing data

which is accessible across Internet/Intranet. Besides, the Web-Server also communicates to one more server, and that is the File Server which is used to store raw data generated through sequencing process. Storage of DNA data into Database provides, structured format of storing, which is easier to store and retrieve the information. Database also provides security of data. Besides, Web-based application also provides the transparency of data location, transparency of database management software, control of concurrent access and scalability of resources. The facility of uploading and retrieving the data is independent of the location of the user who wants to store the data to the database.

The facility of uploading and retrieving the Fasta file to and from the centralised File Server, is also provided. This facility of storing Fasta files to File Server is provided as parallel implementation, to acquire originally generated data through expensive process of Genome sequencing.

The features of Distributed Application are:

- Web-based application
- M-V-C Architecture
- Client-Server Model
- Distributed approach
- Open Standards
- Transparency
- Heterogeneous environment

- Operating System Independent
- Database Management System Independent
- File System Independent
- Location Independent
- Open Source

*This part of research work is published as a paper titled "***Distributed Computing for Structured Storage, Retrieval and Processing of DNA Sequencing Data"** *in an* International Journal of Internet and Web Technology, Recent Science Publications, London

Article Id: 27702633 Vol.38, Issue 1, July 2013, Pg. 1113-1118,

ISSN: 2051-6878

IMPACT FACTOR: 1.89

Site: http://recentscience.org/article/uploadfiles/27702633.pdf

To emphasize on maximized utilization of existing resources in an optimum time and enhancing processing capabilities, Distributed processing is carried out. To achieve this, in addition to Web-based application, the distributed computing is applied for simulatneous processing of an algorithm used to Recognize Identical Reads. The processing is done for recognizing identical reads using multiple datasets, stored in multiple FASTA files. The RMI programming in Java and multi-threading concept of Java is used to implement distributed processing.

The **later part of the research work** involves development of new pattern matching algorithms for DNA sequence analysis. The algorithms use Signal Processing approach for optimization.

The three new algorithms designed for addressing the Bioinformatics issues are:

- 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 pattern matching algorithms are developed to deal with the above mentioned issues. Signal processing approach using Wavelet Transforms has been applied for data reduction, in these pattern matching algorithms, to optimize time involved in searching.

Algorithm – 1) Data Reduction for Data Analysis/Transmission

The purpose of the algorithm is to apply data reduction at two levels, for the DNA sequences. Given a sequence Si of nucleotides, initially, a group of four nucleotide bases is converted into a single byte through binary representation. Each nucleotide base is represented in just two bits. Thus, a group of four nucleotides can be stored in a single byte, against each nucleotide requiring one byte for storage. This reduces the data to one-fourth i.e. $1/2^2$. Further, applying wavelet transforms for four times reduces the number of elements in data to $1/2^4$ times i.e. one-sixteenth of the length of numerical representation of the original nucleotide sequence. Thus, the proposed algorithm applies reduction twice.

So, 64 times data reduction can be achieved using the proposed algorithm, which can be used for various types of speedy processing as well as optimal transmission of the data during distributed computing.

The research work of developing an Algorithm-1, is published as a **paper titled "Dimensionality Reduction of DNA Sequences Using Wavelet Transforms"** and published in Conference Proceedings of 6th WSEAS World Congress: Applied Computing Conference 2013, held at Nanjing, China from 17th to 19th Nov. 2013, Organized by: WSEAS (World Scientific and Engineering Academy and Society).

The paper is published on Pg. 145-152 of Conference Proceedings in Recent Advances in Computer Engineering Series - 18,

ISSN:1790-5109, ISBN: 978-960-474-354-4

Conference Proceedings will be indexed by SCOPUS, ISI (Thomson Reuters), ACS, British Library etc.

Site: http://www.wseas.us/e-library/conferences/2013/Nanjing/ACCIS/ACCIS-23.pdf

Algorithm - 2) Recognizing Identical Reads

The process of DNA sequencing generates large number of identical reads. These potential identical reads which are nearly 3% to 20% of total reads, may introduce biases in analyzing quantification and transcriptome profiling of sequence data. It will also consume unnecessary time in assembling the same reads to form contigs. The research work has introduced an *ab initio method*, which uses the *apriori* approach of signal processing as a recognition criterion, for recognizing identical reads. Detecting identical reads from DNA sequencing, includes exact and near exact identical reads. The proposed algorithm does not require any mapping reference for recognizing identical read, as is the case in existing algorithms. Several other existing algorithms use string comparision which has computational complexity of exponential order

and hence is slow, or use seeds for clustering. This research work makes the use of efficient Haar Wavelet Transforms having computational complexity of *O*(log n) for searching. The feature of Wavelet Transform is used for data reduction and then comparison of reduced, transformed sequences to recognize identical reads, is performed using this proposed algorithm.

This part of research work, the results of Algorithm-2, has been published in form of **two papers**.

1) A paper titled "**Signal Processing Approach for Recognizing Identical Reads From DNA Sequencing of Bacillus Strains**" in an IOSR Journal of Computer Engineering (IOSR-JCE), Volume 10, Issue 1 (Mar. 2013), PP 19-24 e-ISSN: 2278-0661, p- ISSN: 2278-8727.

Published by International Organisation of Scientific Research

The **paper is indexed** by **CrossRef DOI 10.9790/0661-01011924** as well as is indexed by Google Scholar

No. Of Citation: 1

Site:

http://www.iosrjournals.org/iosr-jce/pages/v10i1.html

http://search.crossref.org/?q=10.9790%2F0661-01011924

http://dx.doi.org/10.9790/0661-01011924

2) A paper titled "Recognizing Artificial Duplicate Reads in 454 Pyrosequencing Using Wavelet Transforms" in an International Journal of Advanced Computing

Article Id: 27702498 Vol.46, Issue 2, May 2013, Pg 1205-1211, ISSN:2051-0845

IMPACT FACTOR: 2.31

No. of Citation: 1

Site:

http://recentscience.org/article/uploadfiles/2770249827702498.pdf

Algorithm –3) Recognizing Short Tandem Repeat Regions (STRs)

DNA sequences exhibit ubiquitously distributed patterns of repeat regions, known as Short Tandem Repeats (STRs). Short tandem repeat regions are portions of DNA sequence which have same set of nucleotide bases repeating several times. They are nucleotide sequences in DNA of 1–6 bp unit length, contiguously placed, with multiple copies, distributed randomly in eukaryotic and prokaryotic genomes. Through this algorithm, Short tandem repeat region is identified using Signal Processing. Haar Wavelet Transforms are applied for decomposing the signals to four levels of transforms. The similar consecutive nucleotide bases in a given DNA sequence, after applying wavelet transforms, result into zero values in detailed co-efficients. If the sequence contains the homopolymer of mono-nucleotide, the region in sequence containing Short tandem repeat, would have its detailed-coefficients set to zeros, in its first level of transform. This set of zeros is actually a repeat region. Identifying the start and end position of this set of zeros, is sufficient to identify the region of STR. The repeat region in the sequence, which contains repeat pattern in form of di-nucleotide or multi-mers, can be identified by recursively doing multi-level discrete wavelet transform (DWT). Detailed explanation follows in Chapter 5.

This part of research work has been communicated for publication in an International Journal.

Thus, to briefly mention the research work involves two aspects of computer science, the distributed computing, and the pattern matching

algorithms using signal-processing approach, for DNA sequencing data. The pattern matching algorithms use data reduction techniques by Wavelet Transforms to optimize time of transmission/processing /storage, which is also one aspect of distributed computing. In brief, the research work deals with an inter-disciplinary context of study.

Layout of Thesis Report

Chapter 1 provides the Introduction to research work and describes in brief about general concepts of Distributed Computing, BioInformatics, its data and applications, and concepts of Signal Processing and Wavelet Transforms.

Chapter 2 explains the use of Distributed Computing in developing Web-Application to store and retrieve the DNA sequencing data to secured Database Server.

The process of storing and retrieving of raw Fasta files to File Server are also described in this chapter. The chapter also provides details on features and implementation issues of the Web-Application. Distributed Processing implemented using Java RMI and multithreading is also described in this chapter.

The **List of Java programs developed** and used in implementing Distributed Computing is also mentioned in this chapter. It also illustrates the Data Dictionary designed, developed and used to store sequencing and analysis data in database. Both the data dictionary and some of the developed Java programs are stated in Appendix.

Chapter 3 discusses BioInformatics, BioInformatics Data and Application, DNA sequencing concepts and various issues related to sequencing data.

Chapter 4 describes in detail about Signal Processing, Wavelet Transforms, Haar Wavelets and its characteristics and applications.

Chapter 5 explains the purpose and design of the proposed algorithms used for Data reduction, Identifying Identical Reads in DNA sequencing data and Finding Short Tandem Repeats in DNA sequences. The analysis and results of each of these algorithms are also presented in this chapter. The algorithms and **Matlab programs** developed for each of these issues are presented in this chapter. Since, it is not possible to present all the programs in the thesis, some of the Matlab programs are presented in Appendix.

Chapter 6 concludes the research work and mentions possible future enhancements to this research study.

The studies of literature review, research work and results of the algorithms have been published as papers in International Journals or Conference Proceedings. The details of which, are mentioned in the List of Publications, List of Conference Presentations, List of Technical Lectures/Talks stated at the end of thesis, on Page No. 264 onwards.