CHAPTER 1

INTRODUCTION

Chapter 1. Introduction

Human Genome Project[1] was completed in April 2003. The main purpose of Human Genome Project was to empower the researchers with details, to recognize the genetic factors causing human disease and identify innovative strategies for their diagnosis, treatment, and prevention. The task of Human Genome Project was to map all the genes of a single human being. For the first time the set of all genes, together known as genome, were identified. These set of all genes provide a blue-print of a building of a human being. The human genome comprises of 3 billion[2] chemical "*letters*". These 3 billion "*letters*" also known as *nucleotide bases* or *base-pairs* compose a complete set of DNA in a single human body. *DNA,* an abbreviation for deoxyribonucleic acid, is the chemical compound that contains the genetic information for building, running, and maintaining living beings[3]. The geniticists investigate the DNA molecule to identify chemical sequence causing fatal diseases like Cancer, HIV, and Neurological Disorders etc. It took decades of international research work by hundreds of scientists, to complete the DNA sequencing of a single human genome. Many other organisms, plant or animal are much more complex and have much large DNA sequence than the human being (See Figure 2). The size of DNA sequence is also known as the *Genome Size*. DNA sequencing is the process of identifying the number and order of nucleotide bases in a

[1] http://www.genome.gov/10001772

[2] ttp://www.nlm.nih.gov/medlineplus/magazine/issues/summer13/articles/summer13pg15.html - Understanding the Human Genome Project - A Fact Sheet

[3] ibid

DNA sequence, of any given organism. The differences in number and order of nucleotide bases, distinguish one organism from the other or more precisely, one human being from another.

There is an unbelievable drop in cost of sequencing a single human genome from several millions of dollars to a very few thousand dollars. The high-capacity DNA sequencing machines has now cut down the time to just about 10 days to sequence half a dozen of genomes, which earlier took many years to sequence just one human genome. It is not far off, when DNA sequencing will become an everyday tool for research in life-sciences and medicine. Several benchtop sequencers are available in laboratories and hospitals across the globe, to facilitate personalized medicine that will enhance health-care at lower costs.

DNA sequencing process is given too much consideration. It is not sequencing, but computing that has become slower and more expensive aspect of genomics research today. The duration from 2008 to 2013, demonstrated the increase in performance of a single DNA sequencer from about three to five fold per year[4]. Considering Moore's Law[5] as a benchmark, computer processors are assessed to double in speed every two years during the same period. Sequencers have demonstrated faster improvement rate than computers. Therefore, the necessary computational techniques need to be invented, else the need for putting off vital genomics research on hold, will arise. The necessity of optimized computational techniques arises, because DNA sequencers do not generate a single sequence of an entire genome that researchers can read like one book. It even does not generate the continuous or

[4] The DNA Data Deluge, Michael C. Schatz, Ben Langmead, IEEE Spectrum, Jun 2013

[5] http://en.wikipedia.org/wiki/Moore'sLaw

organised fragments of a vast sequence. Instead, it generates something similar to a giant heap of uncomprehensive and shredded pieces of newspapers. The stack of fragments cannot be dealt with manually, so the problem of arranging all the fragments is assigned to computer programs that accompany the sequencers. The computational focal point should consider both, improvising the algorithms and a refurbish focus on such "big data" approaches as distributed data storage, availability of data, fault tolerance, parallelization, and economies of scale[6].

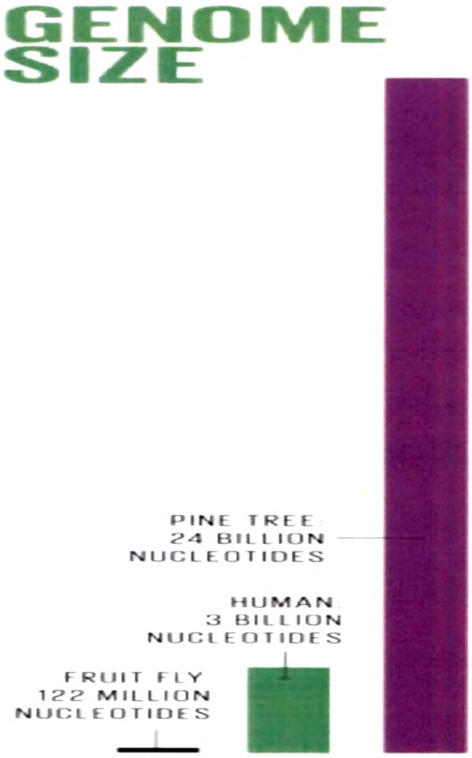

Figure 2. Varying Genome Sizes of different organisms
(Source: The DNA Data Deluge, Michael C. Schatz & Ben Langmead, IEEE Spectrum, Jun 2013)[7]

[6] Ibid, The DNA Data Deluge, Michael C. Schatz & Ben Langmead, IEEE Spectrum, Jun 2013

[7] http://spectrum.ieee.org/biomedical/devices/the-dna-data-deluge

The wide usage of High-Throughput DNA Sequencing Technology for studying various biological mysteries in nature, has led to generation of large data sets. The process of generating this data is usually very complex, is difficult to reproduce, needs skillful work force, requires specialized technical support system, entails efforts, and is financially intensive. Hence, it is necessary to store, retrieve, and process the Big-data efficiently. Moreover, the access to the large data should be viable with the use of freely available softwares or computationally inexpensive techniques. It should be feasible to perform the incremental updates whenever need arises for various experiments. The researchers in biology who are the end-users, expect support for genomic analysis tools and prefer to use these tools without getting involved in installation or configuration of these tools for variety of platforms, at different locations or clients. Moreover upgrading or reconfiguring the many available tools at each location is practically very difficult.

Large sets of data, leads to the need of high configuration computers. Acquiring high-configuration computers for special-purpose usage is not cost effective and technically viable, for all organizations. Moreover, due to confidentiality issues, disclosure laws and regulations, it may not be possible to store Big-data of genomics in global centralized repositories. Besides, if the data is stored in global central repository like NCBI, frequent use of the Internet to download the data is required. To perform online processing of data requires large bandwidth network resources, which tend to be expensive. Hence, the need arises of developing an in-house repository of storage and processing environment for this Big-data.

Nevertheless, storing large data at data generation site usually becomes difficult due to the need of high-tech, special-purpose computer. To make this probable, a single high configuration computer, can be substituted by the cumulative storage and processing environment provided by clusters of general-purpose computers, using *Distributed Computing*. Economy of scale and sufficient processing power can be availed through cumulative in-expensive, off-the-shelf network or cluster of personal computers and workstations. Hence, working with network of general-purpose computers is preferable to procuring special-purpose, high-configuration, expensive super-computer. This concept of using low capacity, inexpensive cluster of computers for storage and processing of large datasets is feasible through Distributed computing.

The research work in this thesis represents the use of Distributed Computing for BioInformatics data storage, retrieval, and processing. Moreover, in the urge to optimize the processing time, the research work also involves development of various algorithms using signal-processing approach for DNA sequence analysis. So, the three broad aspects of research study involve : 1) Distributed Computing for storing and retrieving Bioinformatics data 2) Pattern Matching algorithms to process Bioinformatics data 3) Signal processing approach used to design algorithms for processing and reducing Bioinformatics data.

1.1. Distributed Computing (Applied to BioInformatics)

Distributed Computing can be defined as computing or performing the processing using spatially distributed systems. Thus, Distributed

Computing makes use of Distributed system. Distributed system is "A collection of independent computers that appear to users of the system as a single computer."[8]

The research deals with development of Web-application for storage and retrieval of DNA sequencing data using Distributed Systems.

The research work under-taken deals with development of Web-Application with various components of the application like Client, Web Server, Database Server and File Server, all put on different computers across the network and can be accessed using Internet/Intranet. The developed application maintains complete transparency from the user, heterogeneity of platform in terms of Database Management System, File system, Operating System, Web Server and several other aspects, which are discussed in consecutive chapter on Distributed Computing.

The Web-Application is developed for storing and retrieving BioInformatics data and particularly, the DNA sequencing data. In addition, distributed processing is applied for executing an algorithm applied for bioinformatics simulataneously on multiple machines for multiple datasets. Java RMI and multithreading is used for distribution of processing.

[8] Andrew S. Tanenbaum et al., Distributed Systems: Principles and Paradigms, NJ, Prentice-Hall, 2003.

1.2. BioInformatics data and algorithms (Data acquired from NCBI[9])

"Bioinformatics is a sub-discipline of biology and computer science concerned with the acquisition, storage, analysis, and dissemination of biological data, most often DNA and amino acid sequences. Bioinformatics uses computer programs for a range of applications, including determining gene and protein functions, establishing evolutionary relationships, and predicting the three-dimensional shapes of proteins".[10]

Of all the different types of data available in BioInformatics, most of this research work deals with DNA sequencing data and in particular, the genomic sequencing data.

Importance of DNA sequencing or genomic sequencing is that it gives the better knowledge about the structure or content of DNA of an organism. The sequence of nucleotides that form a particular DNA can be identified using the process of *DNA sequencing*.

DNA sequences can be represented as linear sequences of strings for computational purposes. These linear sequences need to be searched for some similar patterns, the results of which are analyzed for biological interpretation and represented to the end user for further study.

Several Pattern Matching algorithms have been applied for numerous sequence analysis processes. The research study involves three areas, where different approach of pattern matching for DNA sequences is taken into consideration.

[9] http://www.ncbi.nlm.nih.gov

[10] http://www.genome.gov/glossary/index.cfm?id=17

Algorithms developed for the following aspects of DNA sequence data:

1) Data Reduction: Data reduction is the process of minimizing the amount of data that needs to be proccessed. Data reduction can enhance processing and storage efficiency and condense costs. The purpose of Data Reduction is usually to reduce the number of data records by removing the irrelevant or redundant data[11] and produce summary data at various aggregation levels based on applications. DNA sequences are usually very large. Storing, Transmitting or Processing of DNA sequences in its original string form becomes space intensive and at times computationally slow. Reduced form of data also reduces time involved in transmission or processing and hence facilitates during distributed computing. Therefore, the algorithm to apply reduction at two levels, for the DNA sequences has been designed which results in achieving 64 times data reduction. The proposed algorithm, have proved to be efficient in processing of DNA sequences.

2) Duplicate Reads in DNA Sequencing: The process of DNA sequencing, that involves the process of DNA amplification, causes the creation of several duplicate reads. These duplicate reads comprises of 3% to 20% of total reads. Allowing these reads unidentified, creates biases in DNA analysis. These unidentified duplicate reads would unnecessarily ascend time while assembling the reads to form contigs. Hence, the research study involves identifying these duplicate reads using ab initio signal-processing approach.

[11] http://www.fhwa.dot.gov/ohim/handbook/chap7.pdf : Travel Time Data Collection Handbook

3) Identifying Tandem Repeats: "Tandem Repeat[12] is a sequence of two or more DNA base pairs that is repeated in such a way that the repeats lie adjacent to each other on the chromosome. Tandem repeats are generally associated with non-coding DNA". In some instances, the number of times the DNA sequence repetition is variable. Such variable tandem repeats are useful in DNA fingerprinting procedures. Identification of Tandem repeats is essential because they play a role in mutational dynamics, which can cause genetic diseases. In this study, identification of these repeat regions using Haar Wavelet Transforms is applied.

1.3. Signal Processing and Wavelet Transforms

A transform maps an input signal into an output signal.

Processing of large DNA sequences in the raw form becomes space and time intensive. Hence, the DNA sequences can be processed after applying transformations, which in turn reduces amount of data processing, without changing the fundamental behaviour or property of the DNA sequence. This transformation of data helps in reducing space and time of processing.

In the undertaken research work, the signal processing approach is applied for performing transforms on DNA sequences. The DNA sequences are originally represented in linear form as strings. These string form of DNA, are first converted into signals using Binary Indicators, EIIP values or dipole moments, such that the biological inference remains unaltered. Once the sequence is converted into signal,

[12] http://www.genome.gov/glossary/index.cfm?id=193

it is further processed using Wavelet transforms. Wavelet transforms convert a signal into a series of Wavelets, small waveforms. They provide a way for analyzing waveforms, bounded in both frequency and duration. The Wavelet transforms are used to reduce the total size or dimension of the sequence. Wavelet Transformations convert the signals from one form to another form, without altering the original meaning of the signal.

Wavelet transforms are known to provide lossless compression. Besides, Wavelet transforms provides scale-analysis i.e. converts the signal from time-amplitude domain into time-scale domain[13] (or time-frequency domain), thus, preserving the time information. This helps in maintaining the positional information of the nucleotides, which is an integral aspect in DNA sequence processing. Signal Processing and Wavelet Transforms have been used in all the three algorithms which have been developed as part of the research work for DNA data processing. The use of Wavelet transforms for all the three algorithms is discussed in detail in Chapter 5.

Thus, Distributed computing is applied to Web-Application develoment, distributed processing using java RMI. Moreover, use of data reduction techniques using signal processing approach to solve BioInformatics related problems, consecutively attributes to distributed computing in terms of overall time of processing and securing data as the DNA sequences are ecoded to apply digital signal processing.

[13] Wavelet Toolbox User's Guide Version 1, Michel Misiti,Yves Misiti,Georges Oppenheim, Jean-Michel Poggi, 1996-97 The MathWorks Inc, http://www.mathworks.com