CHAPTER 1

INTRODUCTION

1.1: INTRODUCTION
1.2: STATEMENT OF THE PROBLEM
1.3: BACKGROUND OF THE PROBLEM
1.4: PURPOSE OF THE STUDY
1.5: SIGNIFICANCE OF THE STUDY
1.6: RESEARCH QUESTIONS
1.7: SCOPE OF THE STUDY
1.8: LIMITATIONS OF THE STUDY
1.9: THESIS OVERVIEW
1.1: Introduction

Bioinformatics and computational biology are rooted in life sciences as well as computer and information sciences (Konopka A.K. (2005)) and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science. Bioinformatics and computational biology each maintain close interactions with life sciences to realize their full potential. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.
The computational approach has now a days a major impact on biomolecular research (Adami, C, 2004). One of the great impediments is that we are still heavily dependent on powerful resources. However, a new class of small and highly specialized applications is developed, that work very well on personal computers. Data compression is concerned with how information is organized in data (G. Held 1991). There has been extensive research in Data compression algorithms and techniques (T. C. Bell et al. 1990, R. M. Gray, Springer (1990), R. M. Gray 1990). The primary aim of the data compression algorithms is to minimize the storage space and the transmission cost of large size data (J. Ziv et al 1977, J. Ziv 1978).

Efficient storage means removal of redundancy from the data being stored in the DNA molecule. The DNA (Deoxyribonucleic acid) molecule residing in the cell nucleus encodes information conventionally represented as a symbolic string over the alphabet {A, C, G, T} (David Loewenstern et al 1997).

Data compression algorithms (Grumbach S et al 1994, Storer J.A. 1988, Szpankowski W. 1991) removes redundancy and is used to understand biologically important molecules. The starting point is the suggestion made by Chen, Kwong & Li (Chen X et al, 1999) regarding the possibility of a global analysis of nucleotide
sequences. Similar ideas were developed, to a lesser extent, in (Sato H et al, 2001, Matsumoto T, 2000).


Textual data compression, (Bell T.C et al 1990, Nelson M.R, 1989, Raita T et al, Rubin F 1976) and the associated techniques coming from information theory (Billingsley P 1965) are often perceived as being of interest for data communication and storage. However, they are also deeply related to classification (Loewenstern D et al, 2003) and data mining and analysis (Faloutsos C et al, 2007).

In recent years, a substantial effort has been made for the application of textual data compression techniques to various computational biology tasks (Apostolico A et al, 2006, Cunial F et al, 2008) ranging from storage and indexing of large datasets to comparison and reverse engineering (Hartemink A et al, 2005) of biological networks (Basso K et al, 2003). Phylogenetic analysis (Steel M, 1992) elucidate functional relationship within living cells. The problem of inferring the evolutionary history and constructing the phylogenetic tree is a major task in computational biology. Phylogenies are reconstructed using data of all kinds (Tabus I et al, 2003) from molecular data, metabolic data, morphological data to geographical and geological data.

There are three major methods for performing a phylogenetic analysis, distance method, maximum parsimony, (Swofford D.L, 2003) and maximum likelihood methods (Strimmer K, 1996). The alignment methods seem inadequate for post-genomic studies since they do not perform well with data set size and they seem to be confined only to genomic and proteomic sequences. Therefore, alignment-free similarity measures (Li M et al, 2003) are actively pursued. Normalized Compressed Distance (Paul M et al, 2008) is used in construction of evolutionary trees (Badger M et al, 2001) and classify biological sequences and for anomaly detection, and competitiveness in clustering domain data and
phylogeny construction. NCD(Nykter.M et al.,2005) is based on construction of a matrix of pairwise distances between objects (files) that indicate how similar they are (Rudi L. Cilibrasi et al 2005). These distances are based on comparing compressed file sizes. This can be applied to files of widely different types, such as music pieces or genetic codes as well as many other specialized domains.

Communications engineering as well as genetics have both experienced a major breakthrough in the mid 20th century. Until the discovery of the molecular basis of genetics, the research was concentrating on classical genetics, based on the rules of Mendelian inheritance of traits. Shannon introduced the concept of information(Shannon 1940) based solely on the statistical characteristics of the information source (shannon 1948)

The definition of information based solely on statistical characteristics of the information source also applies to genetic data(Whelan S et al 2001). Recent advances in DNA sequencing technology supply enough data to apply Shannon’s general information concept to molecular biology(Mueller J,2003). The possibility of using mutual information for classification and content recognition of genetic sequences is exploited.(Dawy Z et al,2005) Mutual information (Butte.AJ et al ,2003) describes the amount of information shared by stochastic processes. It can be used to derive distance measures quantifying the similarity of the processes. Mutual information based distance measures can be used to compare texts written by different authors or to build phylogenies of different species.

Many computational resources are also available via the World Wide Web. Many Web sites dedicated to the delivery of biomedical data provide significant computational capacity available without cost(Figure:1.2). Such facilities may provide significant help for smaller institutions in meeting the computational needs of biomedical researchers. Furthermore, much of the software widely used in academic research is available as open source, and scripts that facilitate installation of several such packages are available. As a result, given some investment in knowledgeable support staff, it is possible to deliver useful services in support of life sciences research at many levels of investment in computing systems.
1.2 Statement of the Problem

The problem addressed in this research is to develop a novel compression algorithm to achieve the best compression ratios for Entire Genome (DNA sequences). The current proposed algorithm is used to infer the relationships between different organisms through Normalized Compression Distance.

Further the same Algorithm can be used to get evolutionary relationships among different species using shannon’s Information theory. Compressing data, of necessity, involves understanding the way information is structured and, if possible, the mechanism by which the information was generated or is destined to be used. Compression algorithms attempt to extract the way information is organized in the data in an adaptive fashion, in a sense learning the structure which allows for compression. The conceptual tools developed in the field of source coding that have guided the development of data compression algorithms are thus useful instruments for the analysis of how information is organized in general, and in biological systems in particular.

Compression tools can be used to construct an evolutionary tree. Compression ratios signify a great deal of important statistical information. The Normalized Compression
Distance, an efficiently computable and thus practically applicable form of the normalized information distance is used to calculate Distance Matrix. This new distance matrix is proposed to reconstruct Phylogenetic tree. Phylogenies are the main tool for representing the relationship among biological entities.

The DNA sequencing efforts of the past years together with rapid progress in sequencing technology have generated a huge amount of sequence data available in public molecular databases. This recent development makes it statistically feasible to apply universal concepts from Shannon’s information theory to problems in molecular biology, e.g. to use mutual information for gene mapping and phylogenetic classification.

### 1.3: Background Of The Problem

Data compression reveals certain theoretical ideas such as entropy, mutual information and complexity between sequences of different genomes. In the present scenario, many methods like quartet methods, neighbor joining, UPGMA, Fitch’s least squares are in use for constructing phylogenetic trees. These methods fail to handle simple changes like reversal and translocation in DNA sequences and do not apply to more than one gene. A tree constructed using one gene is often different from that using another. Distances such as minimum alignment score works only for closely related genes. These distances are not expected to be general distance measures. Hence we define a proper measure between a pair of DNA sequences. Conditional Compression is newly proposed to evaluate the distance or relatedness between two DNA sequences. Further demonstration is given on how to use compression efficiently to construct an evolutionary tree even for larger Genomes.

### 1.4 Purpose Of The Study

The primary goal of this thesis is to present a compression algorithm, “DNABIT Compress” for DNA sequences based on a novel algorithm of assigning binary bits for smaller segments of DNA bases. This proposed new algorithm could achieve the best compression ratio as much as 1.58 bits/bases where the existing best methods could not achieve a ratio less than 1.72 bits/bases. Data compression also plays a vital role in
analyzing biological sequences to discover hidden patterns, infer phylogenetic relationship between organisms which are areas of active research in bioinformatics.

1.5: Significance of the study

One of the important challenges faced by aligning uncompressed data is the low productivity even after spending huge amounts of space and time. This situation can be overcome with compressed data by using efficient compression techniques and tools. The simplicity and flexibility of “DNABIT Compress algorithm” could make it an invaluable tool for DNA compression in clinical research. This study will help in constructing evolutionary relationship among different species automatically that is without multiple sequence alignment.

1.6: Research Questions

1) What is the effectiveness of the Compression Algorithm specifically designed to compress large DNA Genomes?
2) Can a sequence distance using the information theoretical concept of Normalised Compression Distance be applicable to the problem of whole Genome phylogeny?
3) Can we construct phylogenies from unaligned complete genomes using compressed DNA sequences of genomes?
4) Is it statistically feasible to apply universal concepts from Shannon’s information theory to problems in molecular biology i.e., to use mutual information for phylogenetic classification?
1.7 Scope Of The Study

This study focuses on Compressing large DNA genomes. Genomes of any species can be extracted from the databases that are publicly available to all. The present study uses a new proposed Normalised Compression Distance tool to construct Phylogeny of large DNA genomes which can be downloaded and accessed in any Personal Computer. Biological aspects are not vividly used in this study except to explain the examples and to understand the basics of the problem. Concepts like mutual information based distance measures based on Shannon Entropy combined with source coding can be applied to phylogenetic classification.

1.8 Limitations Of The Study

This study is completely based on efficient and sharp approximate shared information between two DNA sequences, as this information is not computable. However to improve our results, we need a better conditional estimator of DNA sequences, as distances between highly divergent sequences tend to be similar to each other. Although this method is based on information theory rather than a biological model, it is worth stating that the alignment algorithms that Biologists use today are in fact also based on information theory.

1.9 Thesis Overview

Our Goal in this work is not to confirm or refute previous phylogenetic studies but rather to bring a new methodology and a new tool to the comparative genomics research community. Our new method for whole genome comparison and phylogeny does not require gene identification nor any human intervention. This thesis concerns a remarkable new scientific development that advances the state of the art in the field of DNA sequence analysis, or searching for previously unknown but
meaningful similarity among species in fully or semi-automatic ways. A new way of looking at data compressors and information flow in DNA (deoxyribonucleotide) allows us to use compression programs for a wide variety of problems.

In this thesis a few themes are important.

The first is the design of new DNA compress algorithm, “DNABIT Compress” for DNA sequences based on a novel algorithm of assigning binary bits for smaller segments of DNA bases to compress DNA sequences. This proposed new algorithm achieves the best compression ratios for large DNA Genomes. Significantly better compression results show that the approximate repeats are one of the main hidden regularities in DNA sequences.

The second is a new technique for measuring the relatedness between two DNA sequences. Using our proposed DNABIT compress tool, we present strong experimental support for this theory, and demonstrate its application in comparing genomes and constructing evolutionary trees.

And the third is applying universal concepts from shannon’s information theory to problems in molecular biology. To use mutual information for phylogenetic classification.

In the first chapter, a general background, overview of the problem statement, background of the problem, and significance of the study is presented. In chapter two, an introduction to the bioinformatics, role of Computer science in bioinformatics, role of phylogenetics, bioinformatics algorithms and tools, integration of bioinformatics and life sciences are presented and concluded with the importance of Bioinformatics.

In chapter three, brief discussion about the Computer applications in genomics, Normalised Compression Distance, information theory in Molecular Biology, Mutual Information based Distance Measures and databases used by the bioinformatician is discussed. This chapter is concluded with the importance of computer applications in DNA compression of genomes and their functional studies using bioinformatics tools.

In chapter four, briefly discussed the role of compression algorithms to compress repetitive and non-repetitive DNA sequences. The proposed new algorithms GENBIT
compress and Huffbit Compress is discussed. The methodology proposed in new algorithms could achieve a better compression ratio than the existing world wide COMPRESSION algorithms.

In the fifth chapter, explained the methodology, the whole process of construction of the DNABIT compress algorithm and its tool. Described the process of loading data from the genomic data bank/protein data bank, then the data is introduced to DNABIT tool to compress large genomes. Normalised Compressed Distance tool which takes as input the compressed distances of DNABIT tool is also narrated. Phylogeny construction of Genomes using Normalised Compression Distance and Shannon’s mutual information theory is discussed.

In the sixth chapter, outline of different compression ratios of large genomes with the new proposed DNACOMPRESS tool is discussed. The role of similarity metrics and mutual information theory in phylogeny using Distance based method of unaligned DNA sequences rather than traditional multiple alignment methods is discussed for whole genome phylogeny.

The Chapter seven summarizes the conclusions of the key reports and discusses prospects for further work in these areas.

In summary, the present study has made an attempt to show that a substantial amount of mathematical theory is presented as well as very many (though not yet enough) experiments. The results serve to test, verify, and demonstrate the power of this new technology. The core ideas of this thesis relate substantially to data compression programs. The DNA compress algorithm and the computational tools developed in java are useful in analysing the data. This study can be useful to generate new biological information. With the exponential growth of sequence and structure data, Bioinformatics will continue to play an important role in new biological discovery and in formulating intelligent questions for designing experiments.