Chapter 1 Introduction

1.1 Introduction

With the advent of Internet, it has become a common scenario in information society to access to information made available in the World Wide Web. The main goal of this study is to explore and introduce the field of Bioinformatics and the main resources available in the Internet, to enable both biological and computer scientist to access information for research. There are a few main primary public databases in the internet such as NCBI (National Center for Biotechnology Information), EBI (European Bioinformatics Institute) and Genome Net that provide search tools, software and databases for the field of Bioinformatics. Bioinformatics is an interdisciplinary field, which merges information technology, and biology that has skyrocketed in recent years.

This study is to look into the computer-based methods to mine the genomic data that are being uncovered in large volumes within the diverse genome sequencing projects and other new experimental technology in molecular biology by using a data mining tool called BLAST. BLAST (Basic Local Alignment Search Tool) is a sequence similarity search tools that compares the nucleotide protein sequence. Bioinformatics will provide the integrated approach of mining of the voluminous data generated by the Human Genome Project (HGP) to identify and understand human diseases and to develop corresponding drug and gene based therapies.
In this chapter, an introduction to the field of Bioinformatics is discussed. The introduction to The Human Genome Project, which is the sequencing of the whole human genome to unearth the hidden data in the genomic text along the efforts made by the Department and the National Institutes of Health in 1990, is explained. It is followed by a statement of problem, project objectives, motivation and finally on the project organization.
1.1.1 Introduction to Bioinformatics

In the last few years a new biological discipline, Bioinformatics or computational biology has emerged. It is an emerging and rapidly growing area of science and is a fast-growing field within the biological sciences that was developed because of the need to handle large amounts of genetic and biochemical data [1].

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. At the beginning of the "genomic revolution," a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues, but also the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data [2].

![Figure 1.2 Interaction of disciplines that have contributed to the formation of Bioinformatics](image)
Bioinformatics also means the application of mathematics, statistics, and information technology, including computers and the theory surrounding them, to the study and analysis of very large biological, particularly genetic, data sets [3]. It is also defined as a combination of Computer Science, Information Technology and Genetics to determine and analyze genetic information [4]. This field had been fueled by the increase in DNA data generation, particularly from the human genome project. It is a field that has exploded into the awareness of biomedical pharmaceutical and agricultural researchers in the last few years, in parallel with the equally new field of genomics. The central idea of genomics, first articulated by a few visionary biologists and funding officials at the US Department of Energy and National Institutes of Health in the 1980’s was to scale up the laboratory techniques of molecular biology, and to bring to biology a degree of completeness that seemed revolutionary at the time [5].

Over 1,000, 000 DNA and protein sequences have now been determined and deposited within computerized database. This data contains a wealth of information hidden within it on everything from protein structure to the location of drug target sites and disease mechanisms. The challenge facing research biologists and the pharmaceutical industries is how to extract biologically useful information from these tens of thousands of sequences. The explosion in genomic information has created an unprecedented need for scientists with the ability to use computers and information science to study large amounts of genetic information and biological structures. The amount of genomic information to process is staggering.

The goal of computational biology in the early twenty-first century is to link the various genome sequencing projects to a high-throughput effort in complete structural and functional annotation of whole genomes or biological pathways. The first level of the biological hierarchy is comparative analysis of the rapidly emerging genomic data at the sequence level. Complete annotation will include the determination of structure and function of proteins, and a move from analysis of these individual macromolecules to their complex interactions that make up the processes of cellular decisions [6].
The field of Bioinformatics encompassing quantitative and computer methods in Molecular Biology is becoming increasingly important at a time when an enormous wealth of genetic information has been brought to researchers by revolutionary techniques and DNA sequencing. The Human Genome Project contains approximately 100,000 genes out of which about 5,000 are thought to be targeted in disease situations, using small-molecule drugs. This fact lies in the center of the new drug-discovery paradigm that relies on Bioinformatics research and analysis generating massive quantities of data need to be managed and disseminated. Genomics-based approaches to drug design are now the R&D focus at pharmaceutical companies. Basic research in genomics is growing in academic laboratories around the world.

1.1.2 Bioinformatics: From the Pre-genomics to the Post-Genomics Era

Biologists and geneticists already rely on computers to compile, analyze and integrate biological and genetic information in various research databases. With the explosion of data being generated from the Human Genome Project (HGP) as well as from High Throughput Screening programs and related genomic information emanating from microbial and plant resources, the need for high computing power and specialized software for mining such data has led to the emergence of a new and challenging domain in Information Technology which is appropriately labeled BIOINFORMATICS. Bioinformatics will provide the integrated approach of mining the voluminous data generated by the HGP to identify and understand human diseases and to develop corresponding drug and gene based therapies. Bioinformatics will increasingly replace conventional methodologies in the area of drug discovery and drug delivery.
1.1.3 The Human Genome Project (HGP)

The Human Genome Project (HGP) is an International 13-year effort formally begun in October 1990 to discover all the estimated 30,000-35000 human genes and make them accessible for further biological study. Another project goal is to determine the complete sequence of the 3 billion DNA subunits (bases in the human genome) [7]. One of the goals of the HGP is the development of analysis algorithms and integration of genetic databases (informatics) for managing and interpreting genome data. These resources have been made available to health professionals and researchers through the Internet on the World Wide Web (WWW). Universities, research centers, and health organizations have also generated numerous databases that have been made readily available for searching at no charge. Some have made registration a requirement to allow access only to health professionals, and some require a fee for registration. Searching the Internet is rewarding because many databases are linked to each other and offers access to various other related sites [8]. As part of the project, genomes of other organisms such as bacteria, yeast, flies, and mice are also being studied [9]. Other organisms are part of the project and its purpose is to perfect the sequencing methods so that they can be applied to the human genome.

The Human Genome Project (HGP) was undertaken, with the goal to advance fundamental biological understanding and provide the basis for future advances in biotechnology, agricultural, environmental remediation and quality, and health and medical practice. The successes in the analysis of entire genomes have dramatically changed how the biochemistry of living cells is viewed, and provide clear directions for the future of mathematical and computational modeling of molecular, cellular, development and physiological behavior that in turn will open new experimental horizons [6].
Almost everyone will be affected by applications of information and technologies derived from the HGP era of the late 20th century. Among the fields that HGP research will impact are engineering, computer science, mathematics, counseling, pharmaceuticals, law and religion.

1.2 Statement of Problems

Computers have become an essential component of modern biology. They help to manage the vast and increasing amount of biological data and continue to play an integral role in the discovery of new biological relationships. This in silico approach to biology has helped to reshape the modern biological sciences [10]. Molecular biology and computer science are among the most rapidly evolving disciplines in the contemporary scientific scene. In each field, there have been both theoretical advances and practical applications with important social consequences [11]. With the emergence of World Wide Web, it had also revolutionized genomics and bioinformatics research among the biologist and computer scientist. The web will turn out to be an ideal vehicle for delivering genomic data to scientific community.

There is a burgeoning volume of information and data arising from the rapid research and unprecedented progress in molecular biology. For this, the number of journals, reports and research papers and tools required for the analysis of these sequences has also increased. The life sciences today needs tools in information technology and computation to prevent degradation of this data into an inchoate accretion of unconnected facts and figures [12].
1.3 Project Objectives

This project undertakes a study in the area of Bioinformatics, with emphasizes on web-based bioinformatics research via Internet. BLAST (Basic Local Alignment Search Tool), a data mining tool is used in this project. The objectives are summarized as below.

- To explore the possible field of research at the interface of biology and information technology which may lead to important new information systems, such as algorithms and software, and computer hardware technologies.
- To study the available public databases which can be publicly access via Internet.
- To study the web-based bioinformatics resources available in the Internet.
- To expose this emerging field to researchers and students.
- To get an overview of the current state of bioinformatics research.
- To explore BLAST, a data mining tool.

1.4 Motivation and Contribution

Bioinformatics had became an important field that promise to lead to advances in understanding basic biological processes, diagnosis, treatment, and prevention of many genetic diseases. Bioinformatics resources and tools become essential to all aspects of biological research, it is also critical for computer scientist and biologist to aware of the availability of resources in the Internet. The WWW facilitates and the sharing of this treasure, has changed the nature of learning by providing increased access to resources. That is why it is important exploring the data available in the databases for research. By getting the information needed from the databases, analysis of the sequence from the wet laboratory can be interpreted. Identifying genes in DNA sequences and determining the three-dimensional structure of proteins, given the protein sequence will be essential for an accurate and effective analysis of diseases processed by the computer. Computational
tools play a major role in this field of research. Research collaboration between information technology and molecular biology will also lead to a new breakthrough in drug discovery in future.

The rationale for applying computational approaches to facilitate the understanding of various biological processes includes:

- A more global perspective in experimental design; and

- The ability to capitalize on the emerging technology of database-mining—the process by which testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms. [2]

1.5 Project Organization

There are total of six chapters in this project. The project is organized as follow.

- Chapter 1 Introduction
  This chapter gives an introduction to the project. Here, bioinformatics and Human Genome Project are introduced. The objectives and organization of the project are defined.

- Chapter 2 Literature Review: Web-based Bioinformatics resources on the Internet (WWW)
  This chapter provides a discussion on databases, search tools, web-based resources available via Internet to conduct bioinformatics research. In this chapter we discuss the primary public domain bioinformatics resources available on the world-wide-web (WWW). These major databases servers are National Center for
Biotechnology Information (NCBI), European Bioinformatics Institute (EBI) and Genome Net (KEGG & DDBJ).

- **Chapter 3 Data Mining Using BLAST**
  BLAST (Basic Local Alignment Search Tool), a sequence similarity search program via Internet, the five main types of BLAST: blastn, blastp, tblastn, tblastx is defined here. Why we would use one over another and how to set the parameters to get what we want out of a search are also defined here.

- **Chapter 4 Methodology and Results**
  In this chapter, a brief introduction to the concept of molecular biology and DNA sequencing is explained. Then, the details steps in submitting BLAST query is displayed and finally analysis of output of the results obtained from the samples query sequence.

- **Chapter 5 Discussion and Conclusion**
  This chapter summarizes the study of the data-mining tool used in this project. It also discusses the future research of Bioinformatics.

**1.6 Chapter Summary**

With the human genome project underway and the promise of a revolutionary approach to understanding diseases, bioinformatics emerged as one of the most critical scientific disciplines today. It is an area heavily funded for many years, and major areas of computer science are likely to have impact on research development in molecular biology. Bioinformatics will dominate the field within a decade. It is also the intensive interdisciplinary cooperation between the computer scientist and biologist communities that is the motor of progress in this key-technology for the 21st century.