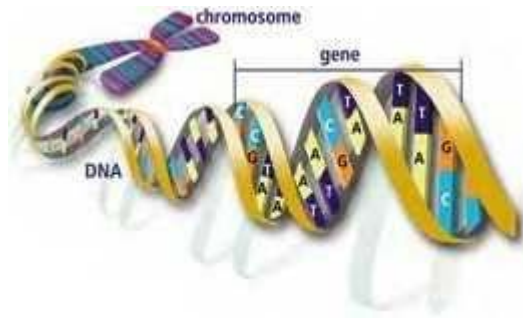


A. Title

Automatic DNA Retrieval and Storage

B. Introduction

Organism contains genetic material that governs an individual's characteristics and are transferred from parent to progeny. The genetic material of all living organisms, both eukaryotes and prokaryotes, is DNA (Deoxyribonucleic acid). DNA is made up of two strands; each strand consists of building blocks called nucleotides. There are four different nucleotides (Adenine, Thymine, Guanine and Cytosine) in a DNA strand and the sequence of these bases within the strand determines the genetic information stored in that strand [1].



DNA: Molecule of Life

(Source: http://www.ornl.gov/TechResources/Human_Genome/graphics/slides/images1.html)

According to the Human Genome Project (HGP) the genome is an organism's complete set of DNA. Genomes vary widely in size. Bacterium, the smallest known genome for a free-living organism contains about 600,000 base pairs. Human and mouse genomes have approximately 3 billion base pairs. Today researchers and scientists are working to:

- (1) Improve content and utility of the National Center for Biotechnology Information (NCBI) database;
- (2) Develop better tools for data generation;
- (3) Capture and annotation;
- (4) Develop and improve tools and databases for comprehensive functional studies; and

(5) Create mechanisms to support effective approaches for producing robust, exportable software that can be shared widely.

The National Center for Biotechnology Information (NCBI) in the United States and the European Bioinformatics Institute (EBI) in England are two major life science servers supporting databases containing a wide variety of genomic information. The National Center for Biotechnology Information (NCBI) being a national resource for molecular biology information has a mission to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control body function. More specifically, the National Center for Biotechnology Information (NCBI) has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules [1].

As new entries are submitted each day, the exponential growth in the database makes it tedious and time consuming for a researcher to retrieve the latest DNA sequences from the servers in order to perform further analysis. With the large volume of biological data, it is important to automate and organize the retrieval of the latest data from the National Center for Biotechnology Information (NCBI) database.

The goal of this project is to develop and implement a user-friendly, general-purpose web-based application for downloading DNA sequences automatically from the National Center for Biotechnology Information (NCBI) servers. The application also stores the data in a local database that can be used for pattern matching algorithms in bioinformatics. This application is based on the new Microsoft technology called the .NET Framework. This web

application uses ASP.NET to retrieve data stored on the National Center for Biotechnology Information (NCBI) website. As this new technology supports a large number of programming languages, it will allow future enhancements and additions to this application.

C. Literature Review And Current State-Of-The-Art

After studying numerous textbooks and conducting an extensive search through the World Wide Web, I found no application to support automatic DNA retrieval to a local database. Most web sites have their own software tools for DNA retrieval from the general web sites like the National Center for Biotechnology Information (NCBI) and EBI, but they do not have the option of storing those sequences into the local database for further analysis [5].

There are many tools developed for the analysis of DNA sequence but I found no tool for retrieval of DNA from other publicly accessed sites [6][7][8]. It seems that a majority of researchers depend on the data stored on other sites, but they do not have local access to it.

D. Methodology

Bio-informatics is an art of organizing and analyzing a large volume of data (3 billion base pairs in human genome) resulting from the molecular and biological techniques like isozyme electrophoresis and PCR amplification. With the constant update of the data in the database servers (The National Center for Biotechnology Information and EBI), it is difficult to find the latest updated data. This application is an attempt to download and organize the DNA sequence automatically as per the user specifications from the National Center for Biotechnology Information (NCBI) databases.

This web-based application will enable the user to:

- (1) Select a database from a set of databases provided in the National Center for Biotechnology Information (NCBI) site;
- (2) Specify the search criteria;

- (3) Retrieve DNA sequences related to the search criteria;
- (4) Organize the retrieved sequence; and
- (5) Store the DNA sequences in an easily accessible local database.

Users and programmers can access the sequences stored in local database for further processing.

The application will start with the introduction page, which describes the application in detail and will also have links to other pages. Figure 1.1 shows a sample introductory page.

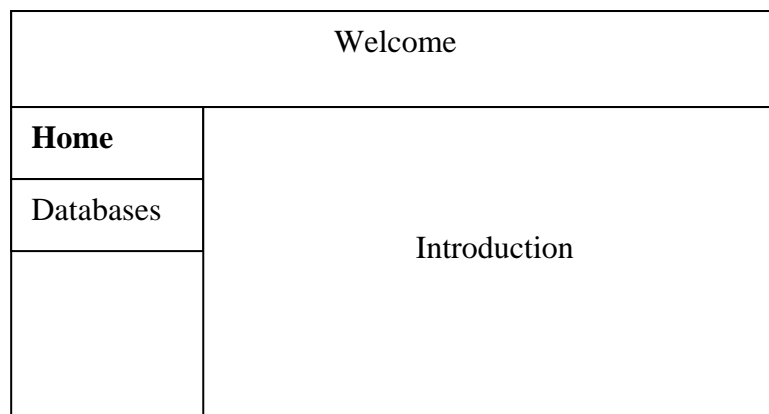


Figure 1.1

One of the links would be a database link, which when clicked would give a list of databases that this application will support. Figure 1.2 shows various links when the database would be selected from left-hand menu.

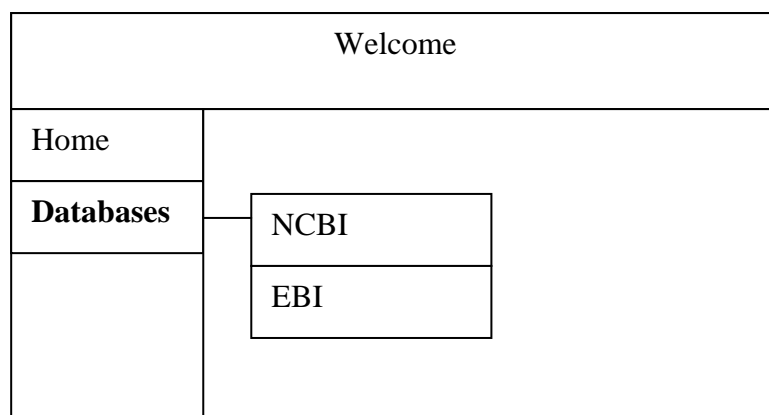


Figure 1.2

When the mouse is passed over each line, a brief description of the database will be shown. The user would be allowed to select one of the databases from that list. Figure 1.3 shows mouse-over action on any database link.

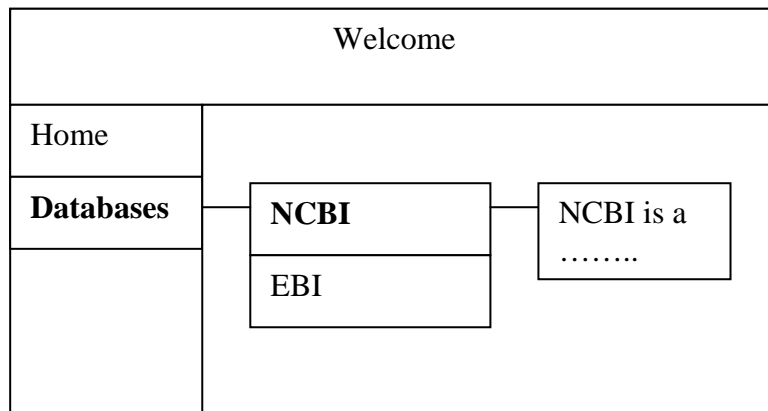


Figure 1.3

When the user clicks on any of the links, a detailed description of the database will be displayed on the page. Figure 1.4 shows the details of the database selected.

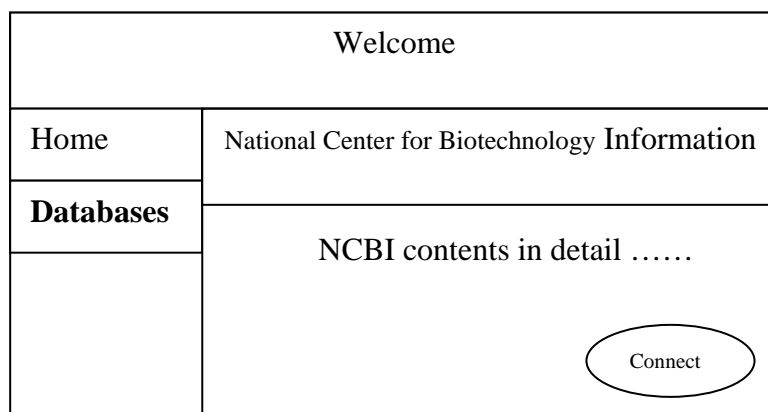


Figure 1.4

After the user selects the database, she/he selects the connect button. The connect button will take the user to another page that will allow the user to type the criteria on the

basis of which DNA sequences will be retrieved. The user has the option of downloading all available DNA sequences matching the criteria or only the latest DNA sequences published. The application will check existing publications in the local database and download only those not already in the database. Figure 1.5 shows the sample page to enter the search criteria.

NCBI Database	
Home	Category: <input type="text"/> For: <input type="text"/> Download: <input type="checkbox"/> Recent Publications <input type="checkbox"/> All <input type="button" value="Download"/>
Databases	

Figure 1.5

Once the user selects the download option, the application will:

- (1) Download the file from the database site;
- (2) Check for the files that are stored in a local database;
- (3) Display the number of files found for that query; and
- (4) Display the number of files extracted.

Figure 1.6 shows the final result of a search.

National Center for Biotechnology Information Database	
Home	<u>RESULT OF SEARCH FOR <search criteria></u> Total number of files found: _____ Total number of files extracted: _____
Databases	

Figure 1.6

The files extracted from the above steps will be stored in a local database. The database will use a unique ID to store the information on extracted files. The DNA sequence along with the other information will be stored in separate fields. The application will check for the DNA sequences already stored in the database and will extract only the new sequence. When no record for the query is found in the local database, all extracted files will be stored in the database.

The database will have the following fields:

1. File ID
2. DNA sequence
3. Header Info

The file ID is a unique number that is extracted from the DNA file. This is a unique ID assigned to articles published in the NCBI database. The file ID will be used as an identifier in the process of search on the local database.

E. Contributions of the Project

The goal of this project is to automatically retrieve and store the DNA sequences of interest in the local database. This DNA data can serve as an input for further analysis in the field of research of the users. Researchers will have immediate access to the DNA sequences of their interest from the local database. Immediate access to the local database will improve the efficiency and effectiveness of algorithms that will analyze the existing relevant strands.

This project focuses on automating data retrieval and storage for researchers interested in the analysis of DNA sequence. The project not only will save enormous time, it also will eliminate errors and repetitions. The user-friendly interface of this web application will make the task simpler for unsophisticated users.

F. References

- [1] Peter J. Russell. (2000). *iGenetics*. Benjamin Cummings.
- [2] Hooman H. Rashidi, Lukas K. Buehler. (1999). *Bioinformatics Basics: Applications in Biological Science and Medicine*. CRC press.
- [3] Department of Energy and the National Institutes of Health. *Human Genome Project*. Source retrieved March, 2003 from http://www.ornl.gov/TechResources/Human_Genome/research/informatics.html
- [4] James Tisdall. (2001). *Beginning Perl for Bioinformatics*. O'Reilly & Associates.
- [5] National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD, Source retrieved February, 2003 from <http://www.ncbi.nih.gov/>
- [6] Sequence Retrieval – Find the nucleotide sequence for a gene of interest, Source retrieved March, 2003 from <http://www.colorado.edu/chemistry/bioinfo/BioinformaticsApplications.htm>
- [7] Dr. Arthur Carty, (Ottawa, 3 February, 1999), Canadian Bioinformatics Resource (CBR); North America's largest DNA sequence retrieval system, Source retrieved March, 2003 from http://www.nrc-cnrc.gc.ca/newsroom/news/cbr_e.html
- [8] BCM Search Launcher - at the Human Genome Center, Baylor College of Medicine, Houston TX, Source retrieved March, 2003 from http://www.dur.ac.uk/biological.sciences/Bioinformatics/DNA_corner.htm