Chapter 2 Literature Review

With the globalization of the Internet and the data deluge from the Genome sequencing projects, bioinformatics is going through a period of explosive growth and development. The WWW facilitates and the sharing of this treasure, has changed the nature of learning by providing increased access to resources in variety types of media. Over the course of the last decade, molecular biology has become one of the most rapidly expanding sciences. The overwhelming and ever-growing pool of knowledge makes it virtually impossible to follow new and recent discoveries using conventional methods such as reviewing journals or scientific textbooks. Computer based resources have therefore become critical for both the biologist and computer scientist [3].

The purpose of this chapter is to provide with examples of such resources and covers the review of the current advancement in bioinformatics and the widely available resources in the Internet for the research community worldwide.

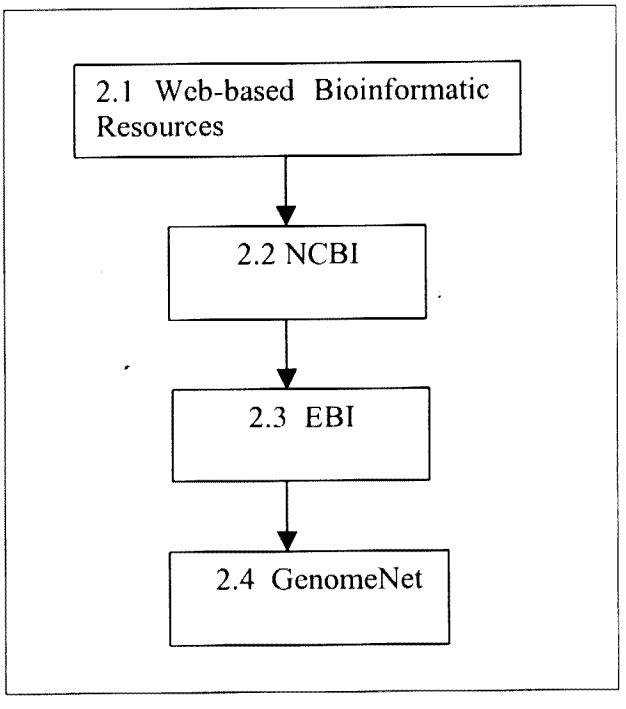

Figure 2.1 Overview of Chapter 2

2.1 Web-based Bioinformatics Resources

Access to and use of the Internet has become so ubiquitous in our society, especially among scientists. Since the 1980s, free software for biosciences has been made available most widely through the Internet [5]. The PC explosion that followed and the increasing popularity of the World Wide Web are witnesses of the computer revolution, the fastest growing technology in man's history [8].

Computational tools and databases are essential to the management and identification of subtle patterns found by using this exponentially growing volume of biological data. The NCBI in the United States and the EBI in England are two main life science servers responsible for dealing with this staggering volume of data. They both maintain reliable database and analytical software that serve as valuable tools for scientific community. The services are made possible by fast, elaborate computers that can perform the necessary analytical tasks, and the Internet that facilitates the electronic communication efforts [10].

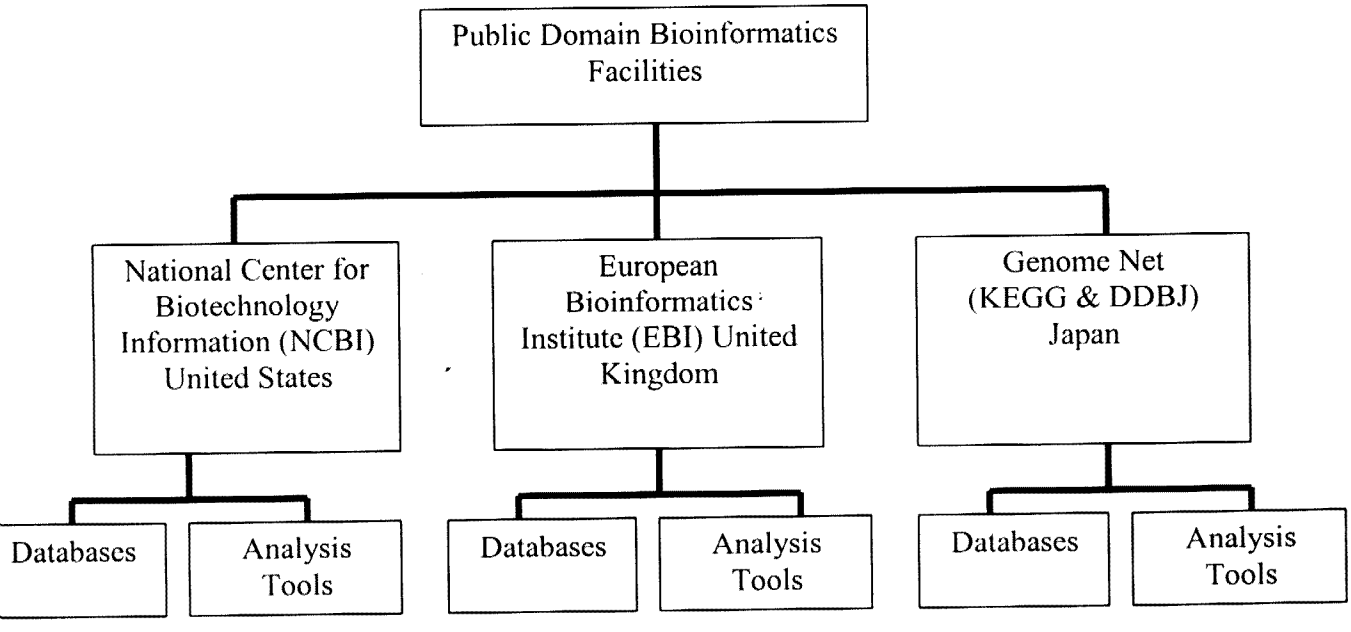

Figure 2.2 Primary Public Domain Bioinformatics Servers

2.2 National Center for Biotechnology Information (NCBI)

The National Center for Biotechnology Information (NCBI) was established in November 1988, at the National Library of Medicine (NLM) in the United States. The NLM was chosen because it had experience in creating and maintaining biomedical databases and as part of the National Institutes of Health (NIH), it could establish an intramural research program in computational molecular biology. The mission of the NCBI is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. It was set up to perform these four major tasks as quoted from the NCBI web site (http://www.ncbi.nlm.nih.gov/): [13]

1. Create automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics.
2. Perform research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules.
3. Facilitate the use of databases and software by biotechnology research and medical personnel.
4. Coordinate efforts to gather biotechnology information worldwide. [4]

NCBI has provided integrated access to all public genetic sequence information and its associated annotation, as well as the citations and abstracts from the published literature referenced by the genetic sequence records. It defines interconnections between genetic sequence data, structure data, taxonomic data, and literature references [9]. These links between literature articles, between genetic sequence records based on Blast sequence comparisons and citations in the genetics sequence records that determine links between the literature and genetic sequence information. Besides, biological literature is available in PubMed that includes all MEDLINE records and some additional records.

At NCBI, databases are linked through a unique search and retrieval system, called Entrez. Entrez allows a user to not only access and retrieve specific information from a single database, but to access integrated information from many NCBI databases [2].

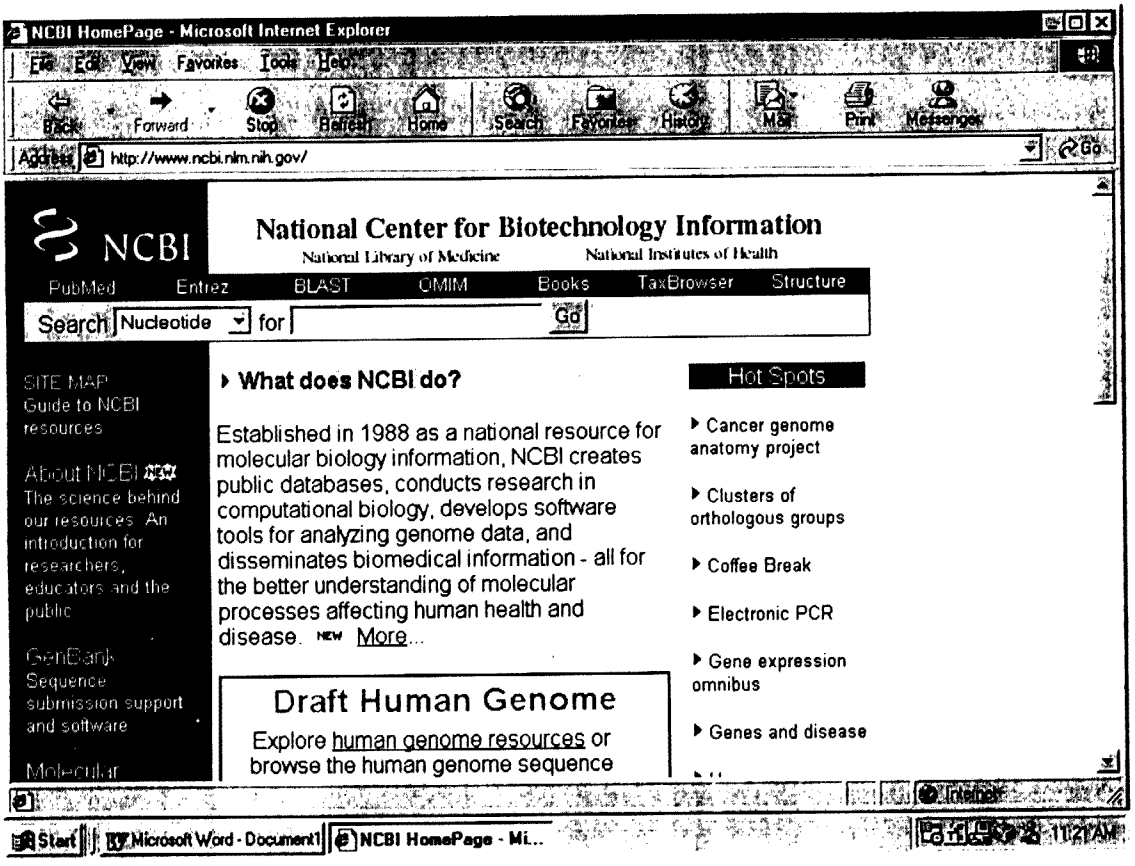

Figure 2.3 National Center for Biotechnology Information (NCBI)

There are seven main databases and analysis tools supported by the NCBI server at their web site:

1. PubMed (Public MEDLINE)
2. BLAST: Basic Local Alignment Search Tool
3. Entrez
4. BankIt (World Wide Web submission)
5. OMIM (Online Mendelian Inheritance in Man)
6. Taxonomy
7. Structure

In the following study, five of the main databases: PubMed, BLAST, Entrez and OMIM are discussed.

2.2.1 BLAST

Researchers at the NCBI have developed a program called the Basic Local Alignment Search Tool (BLAST) to rapidly compare an amino acid or nucleotide query sequence to databases of sequences. BLAST is not guaranteed to find the best hit between a query sequence and a database. BLAST is available on the WWW by going to the NCBI home page. (http://www.ncbi.nlm.nih.gov/BLAST) [4].

BLAST is a program for sequence similarity searching, and is instrumental in identifying genes and genetic features. BLAST can execute sequence searches against the entire DNA database in less than 15 seconds. NCBI's databases and software tools are available from the WWW or by FTP. NCBI also has e-mail servers that provide an alternative way to access the databases for text searching or sequence similarity searching [14].

2.2.2 Entrez

Entrez is NCBI's search and retrieval system that provides users with integrated access to sequence, mapping, taxonomy, and structural data. Entrez also provides graphical views of sequences and chromosome maps. A powerful and unique feature of Entrez is the ability to retrieve related sequences, structures, and references. The journal literature is available through PubMed, a Web search interface that provides access to over 11 million journal citations in MEDLINE and contains links to full-text articles at participating publishers' Web sites [14].

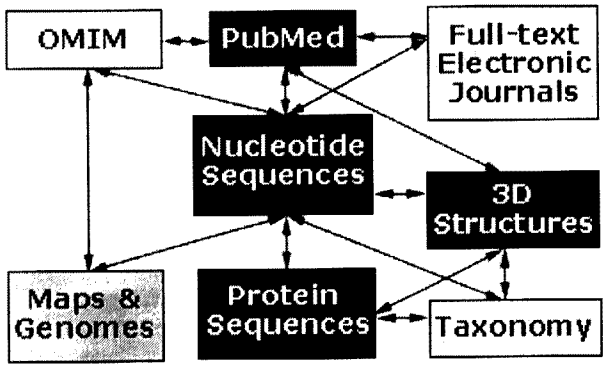

Figure 2.4 Entrez database Browser System

Entrez provides molecular biology data and bibliographic citations from the NCBI's integrated databases. These include: DNA sequences from GenBank, EMBL, and DDBJ. Protein sequences from SwissProt, PIR, PRF, PDB; and translated protein sequences from the DNA sequences databases; genome and chromosome mapping data; three-dimensional protein structures derived from PDB, and incorporated into NCBI's Molecular Modeling Database (MMDE). In addition, a bibliographic database (PubMed) containing citations for nearly 9 million biomedical articles is available via the National Library of Medicine's MEDLINE and Pre-MEDLINE databases. The Internet address for Entrez is http://www.ncbi.nlm.nih.gov/Entrez [8]

2.2.3 OMIM

Online Mendelian Inheritance in Man (OMIM) is a Web-based catalog that contains thousands of entries for genes and genetic disorders and serves as a phenotypic companion to the Human Genome Project. The OMIM cytogenetic and morbid maps present cytogenetic locations for those genes with published locations and provide an alphabetical list of all the diseases described in OMIM.

To validate the findings generated through computer-based comparative analysis, it is essential to consider the results of wet-bench biology reported in the scientific literature. Therefore, the integration of scientific data with the literature is a necessary step for creating a unified information resource in the life sciences. To this end, individuals are

provided with a direct link from OMIM to PubMed, NCBI's literature retrieval system [14].

This is a catalog of human genes and genetic disorders, compiled and indexed by Victor A. McKusick and others at Johns Hopkins University and elsewhere. The effort was initiated in the early 1960s with the Catalog of X-Linked Traits in Man. After numerous printed editions, it has been developed by the National Center for Biotechnology Information (NCBI) for the World Wide Web. The database contains abstracted information from numerous published sources, references, and links to other Internet resources such as Entrez: the NCBI's MEDLINE and GenBank Retrieval systems [8]. The Internet address for OMIM is http://www.ncbi.nlm.nih.gov/omim

2.2.4 PubMed

PubMed provides Web-based access to over 11 million citations, abstracts, and indexing terms for journal articles in the biomedical sciences. It also includes links to full-text journals. Currently, approximately 20 million searches are conducted per month and as many as 140,000 different users seek information daily via PubMed [14].

PubMed is the search service of the National Library of Medicine (NLM). It allows the user to gain access to citations in MEDLINE and PRE-MEDLINE, and is linked to participating online journals and related databases enabling the user to retrieve pertinent information in a speedy and efficient manner. Keywords may be used to retrieve journals articles that contain relevant topics. Multiple keywords may be used to increase the specificity of the search. Other search criteria such as author names and journal titles are also available for the user's convenience [1].

2.3 European Bioinformatics Institute (EBI)

EBI is an outstation of the European Molecular Biology Laboratory (EMBL) located at Hinxton, England. Fourteen European countries and Israel support EMBL and its outstations. EBI's main purpose is to conduct research and provide information about bioinformatics to the world's scientific community. The EBI is comparable to the NCBI in the United States and is the main bioinformatics server for the European community. Its tasks and goals are similar to those of NCBI and include:

- Bioinformatics tracking technology
- Research and development of bioinformatics software
- Training and supporting its subscribers
- Relevant bioinformatics services

The European Bioinformatics Institute (EBI) is a non-profit academic organization that forms part of the European Molecular Biology Laboratory (EMBL). The EBI is a center for research and services in bioinformatics. The Institute manages database of biological data including nucleic acid, protein sequence and macromolecule structure. The mission of the EBI is to ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is accessible freely to all facets of the scientific community in ways that promotes scientific progress [16].

The EBI also provides network services that allow access to the most up-to-date data collection via the Internet through its World Wide Web interfaces and ftp services, also providing database and sequence similarity searches facilities. The services offered by EBI also include databases, data submission, query databases and similarity searches, online applications, FTP archives and research and development [1].

The staffs at EBI are involved in many facets of the bioinformatics world. Their research tasks include:

- Developing more robust comparison algorithms
- Creating more elaborate, but user-friendly, networked information systems
- Designing more-efficient databases

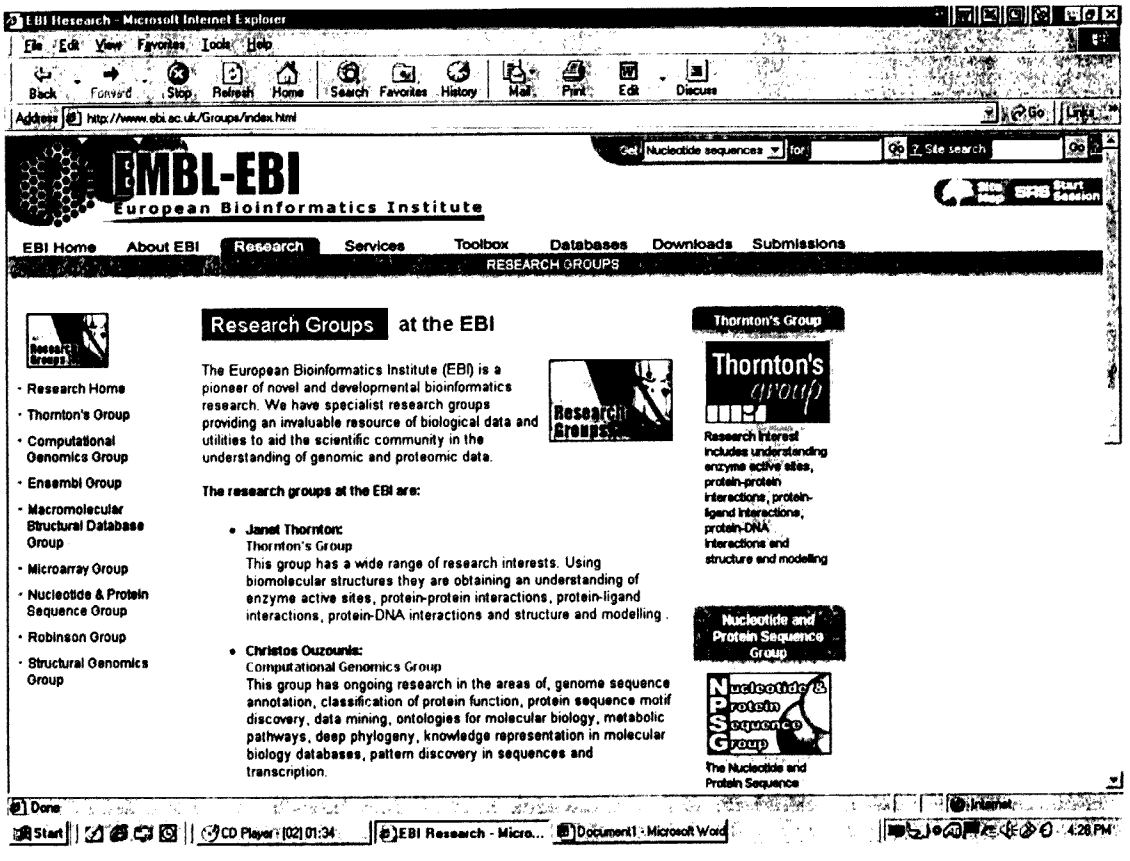

Figure2.5 European Bioinformatics Institute

2.3.1 Databases at EBI

Following are the main databases supported by EBI's Web Server:

1. EMBL database
2. SWISS-Prot database
3. Radiation hybrid database

4. dbEST and dbSTS
5. PDB (Brookhaven Mirror)
6. IMGT database
7. Databases on EBI ftp server
8. NBD (Mirror site)
9. Flybase archives
10. MitBASE
11. Subtilist Web server
12. Software Bio Catalog

Only the first two are discussed below. Other EBI services and tools can be accessed through their website at www.ebi.ac.uk/ebi_home.html. [1]

2.3.2 EMBL - The Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database is a central activity of the EBI. This is a comprehensive database of nucleotide sequences (e.g., DNA and RNA). The nucleotide sequences at EMBL are from a variety of sources. Some are from scientific literature and patent applications; large portion of the database includes sequences submitted directly by the sequencing source. The database is collaboration between the American Genbank nucleotide database at NCBI and the DNA database of Japan (DDBJ). The EMBL database communicates with the other two databases through its daily exchange program and constantly updates its contents. This allows EMBL to offer the worldwide scientific community an updated nucleotide database of all known public domain nucleotide sequences.
[15]

2.3.3 The SwissProt Database

SwissProt is an annoted protein sequence database, produced in collaboration between the EBI and the Department of Medical Biochemistry of the University of Geneva (Switzerland). It contains high-quality annotated data, is nonredundant, and is cross-referenced to many other databases. For standardization purposes, the format of a *SwissProt* entry follows as closely as possible the format of an *EMBL* entry [1].

2.4 GenomeNet- Japanese Bioinformatics Servers

GenomeNet is a Japanese network of database and computational services for genome research and related research areas in molecular and cellular biology. It was established in September 1991 under the Human Genome Program (HGP) of the Ministry of Education, Science, Sports, and Culture (MESSC). GenomeNet services are operated jointly by the Supercomputer Laboratory (SCL), Institute for Chemical Research (ICR), Kyoto University and the Human Genome Center (HGC), in the Institute of Medical Science, of the University of Tokyo. GenomeNet can be accessed at GenomeNet Services at http://www.genome.ad.jp/ and provides the following services:

DBGET/LinkDB Integrated Database Retrieval System

- DBGET/LinkDB/KEGG database links diagram
- Database release information (updated daily)
- Growth of major databases (since 1982)

The DBGET/LinkDB can be accessed at http://www.genome.ad/dbget and http://www.genome.ad.jp/dbget/dbget.links.html

KEGG: Kyoto Encyclopedia of Genes and Genomes

- KEGG table of contents
- Complete genomes in KEGG
- LIGAND chemical database for enzyme reactions
- BRITR: biomolecular relations in information transmission and expression
- IUPAC/IUBMB nomenclature recommendations

Sequence Interpretation Tools

- IDEAS interface to DBGET/BLAST/FASTA

Genome Databases in Japan (Anonymous FTP of the GenomeNet)

- How to download the KEGG system

GenomeNet was also developed for the interpretation of sequence information and provides one of the most useful collections of analysis tools for a variety of biological questions. GenomeNet's tools include BLAST and FASTA, which are sequence similarity search programs, and MOTIF, a sequence motif search program developed at Kyoto University [1].

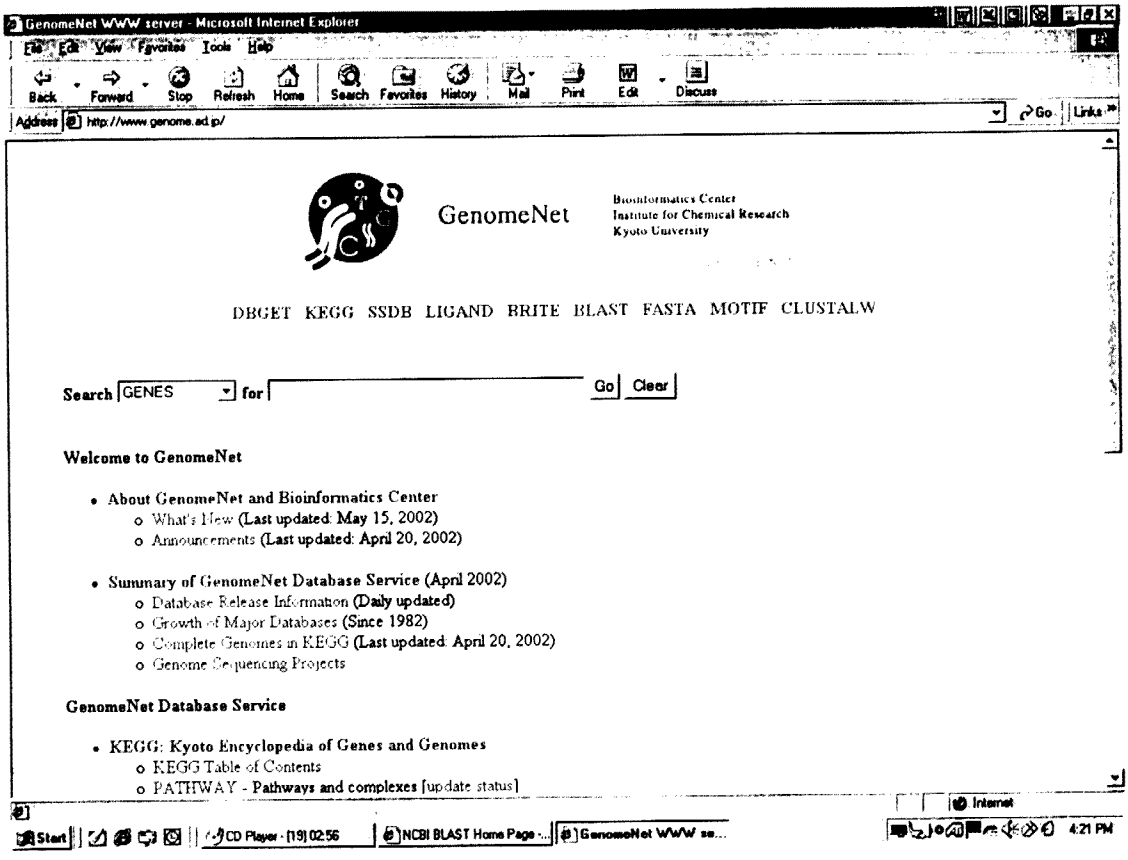

Figure 2.6 GenomeNet Japan

2.4.1 DBGET/LinkDB Integrated Database Retrieval System

The integrated database retrieval system DBGET/LinkDB is the backbone of the Japanese GenomeNet service. DGET is used to search and extract entries from a wide range of molecular biology databases, while LinkDB is used to search and compute links between entries in different databases. DBGET/LinkDB is designed to be a network distributed database system with an open architecture, which is suitable for incorporating local databases or establishing a specialized server environment. It also has an advantage of simple architecture allowing rapid daily updates of all the major databases. The WWW version of DBGET/LinkDB at GenomeNet is integrated with other search tools, such as BLAST, FASTA and MOTIF, and with local helper applications, such as RasMol [17].

2.4.2 KEGG (Kyoto Encyclopedia of Genes and Genomes)

KEGG is an effort to make links from the gene catalogs generated by the genome sequencing projects to the biochemical pathways.

The objectives of KEGG are the following:

1. To computerize all aspects of cellular functions in terms of the pathway of interacting molecules or genes.
2. To maintain gene catalogs for all organisms and link each gene product to a pathway component.
3. To organize a database of all chemical compounds in the cell and link each compound to a pathway component, and
4. To develop computational technologies for pathway comparison, reconstruction, and analysis. [18]

The KEGG can be accessed at http://www.genome.ad.jp/kegg/ and http://www.genome.ad.jp/kegg/kegg2.html

2.5 Chapter Summary

NCBI, EBI, DBGET/LinkDB and KEGG had provided useful information, software and links to other useful sites related to bioinformatics research. All these three main Web-based Bioinformatics resources had benefited the bioinformatics research community in many ways. Researchers obtained reliable databases and analytical software that served as valuable tools for their findings. This has amazingly shortened the period of time in finding the results of unknown data and had created an inspired research community. Researchers are able to create and share their knowledge and findings in the Internet. This had helped to create a community of shared knowledge workers. Better education in Bioinformatics is required so that users know what is available and how to access and use the resources. Only then, will we all be able to exploit the true potential of Bioinformatics.