BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Statistical Methods for Bioinformatics

Course Code: 23ST08DEC

Unit-IV

Overview of bioinformatics

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. I. Angel Agnes Mary
Guest Faculty
Department of Statistics

UNIT – IV

OVERVIEW OF BIOINFORMATICS

Bioinformatics

Bioinformatics is an emerging branch of biological science that emerged by the combination of both biology and information technology. It is an interdisciplinary field of study that uses Biology, Chemistry, Mathematics, Statistics, and Computer Science that are merged to form a single discipline. This sector is mainly involved in analyzing the biological data, developing new software using biological tools.

According to the NCBI- National Center for Biotechnology Information, the branch of NLM- National Library of Medicine and NIH- National Institutes of Health, defines Bioinformatics as the analysis, collection, classification, manipulation, recovery, storage and visualization of all biological information using computation technology.

The term Bioinformatics was first coined in the year 1960 by the two Dutch biologists named Paulien Hogeweg and Ben Hesper. According to their research and discoveries, Bioinformatics was defined as the study of information processes in biotic systems.

Application of Bioinformatics

Bioinformatics is mainly used to extract knowledge from biological data through the development of algorithms and software. Bioinformatics is widely applied in the examination of Genomics, Proteomics, 3D structure modelling of Proteins, Image analysis, Drug designing and a lot more. A significant application of bioinformatics can be found in the fields of precision and preventive medicines, which is mainly focused on developing measures to prevent, control and cure dreadful infectious diseases.

The main aim of Bioinformatics is to increase the understanding of biological processes.

Listed below are a few applications of Bioinformatics.

- In Gene therapy.
- In Evolutionary studies.
- In Microbial applications.
- In Prediction of Protein Structure.
- For the Storage and Retrieval of Data.
- In the field of medicine, used in the discovery of new drugs.
- In Biometrical Analysis for identification and access control for improvising crop management, crop production and pest control.

Scope

Current biological and medical labs use methods that produce extremely large data sets, which cannot be analyzed by hand - for instance sequencing human genomes. Thus modern biological and medical research and development cannot be done without bioinformatics. Future applications in biology, chemistry, pharmaceuticals, medicine, and agriculture.

Aims of Bioinformatics

In general, the aims of bioinformatics are three-fold.

1. The first aim of bioinformatics is to store the biological data organized in form of a database. This allows the researchers an easy access to existing information and submits new entries. These data must be annoted to give a suitable meaning or to assign its functional characteristics. The databases must also be able to correlate between different hierarchies of information. For example: GenBank for nucleotide and protein sequence information, Protein Data Bank for 3D macromolecular structures, etc.

2. The second aim is to develop tools and resources that aid in the analysis of data. For example: BLAST to find out similar nucleotide/amino-acid sequences, ClustalW to align two or more nucleotide/amino-acid sequences, Primer3 to design primers probes for PCR techniques, etc.

3. The third and the most important aim of bioinformatics is to exploit these computational tools to analyze the biological data interpret the results in a biologically meaningful manner.

Goals

The goals of bioinformatics thus is to provide scientists with a means to explain

1. Normal biological processes

2. Malfunctions in these processes which lead to diseases

3. Approaches to improving drug discovery

Human Genome Project

Human genome project (HGP) was an international scientific research project which got successfully completed in the year 2003 by sequencing the entire human genome of 3.3 billion base pairs. The HGP led to the growth of bioinformatics which is a vast field of research. The successful sequencing of the human genome could solve the mystery of many disorders in humans and gave us a way to cope up with them.

Goals of the human genome project

Goals of the human genome project include:

- Optimization of the data analysis.
- Sequencing the entire genome.
- Identification of the complete human genome.
- Creating genome sequence databases to store the data.
- Taking care of the legal, ethical and social issues that the project may pose.

Methods of the human genome project

In this project, two different and significant methods are typically used.

1. Expressed sequence tags wherein the genes were differentiated into the ones forming a part of the genome and the others which expressed RNAs.
2. Sequence Annotation wherein the entire genome was first sequenced and the functional tags were assigned later.

The process of the human genome project

- The complete gene set was isolated from a cell.
- It was then split into small fragments.
- This DNA structure was then amplified with the help of a vector which mostly was BAC (Bacterial artificial chromosomes) and YAC (Yeast artificial chromosomes).
- The smaller fragments were then sequenced using DNA sequences.
- On the basis of overlapping regions, the sequences were then arranged.
- All the information of this genome sequence was then stored in a computer-based program.
- This way the entire genome was sequenced and stored as genome database in computers. Genome mapping was the next goal which was achieved with the help of microsatellites (repetitive DNA sequences).

Features

Features of the Human genome project include:

- Our entire genome is made up of 3164.7 million base pairs.
- On average, a gene is made up of 3000 nucleotides.
- The function of more than 50 percent of the genes is yet to be discovered.
- Proteins are coded by less than 2 percent of the genome.
- Most of the genome is made up of repetitive sequences which have no coding purposes specifically but such redundant codes can help us better understand of genetic development of humanity through the ages.

Applications of HGP

As the goals of the human genome project were achieved, it led to great advancement in research. Today, if any disease arises due to some alteration in a certain gene, then it could be traced and compared to the genome database that we already have. In this way, a more rational step could be taken to deal with the problem and can be fixed with more ease.

Internet

The **Internet** is a global system of interconnected computer networks that use the standard Internet protocol suite (TCP/ IP) to serve billions of users worldwide. It is a *network of networks* that consists of millions of private, public, academic, business, and government networks, of local to global scope, that are linked by a broad array of electronic, wireless and optical networking technologies. The Internet carries a vast range of information resources and services, such as the inter- linked hypertext documents of the World Wide Web (WWW) and the infrastructure to support electronic mail.

Types of Internet Connections

There are five types of internet connections which are as follows:

(i) Dial up Connection

(ii) Leased Connection

(iii) DSL connection

(iv) Cable Modem Connection

(v) VSAT

World Wide Web

The Internet is an international network (a collection of connected, in this case, computers) – networked for the purpose of communication of information. The Internet offers many software services for this purpose, including:

- World Wide Web
- E-mail
- Instant messaging, chat
- Telnet (a service that lets a user login to a remote computer that the user has login privileges for)
- FTP (File Transfer Protocol) – a service that lets one use the Internet to copy files from one computer to another

The Web was originally designed for the purpose of displaying "public domain" data to anyone who could view it. Although this is probably the most popular use of the Web today, other uses of the Web include:

- Research, using tools such as "search engines" to find desired information.
- A variety of databases are available on the Web (this is another "research" tool). One example of such a database: a library's holdings.
- Shopping – most sizable commercial organizations have Web sites with forms you can fill out to specify goods or services you wish to purchase. Typically, you must include your credit card information in this form. Typically, your credit card information is safe – the system is typically automated so no human can see (and steal) your credit card number.
- We can generalize the above: Web forms can be filled out and submitted to apply for admission to a university, to give a donation to a charity, to apply for a job, to become a member of an organization, do banking chores, pay bills, etc.
- Listen to music or radio-like broadcasts, view videos or tv-like broadcasts.
- Some use the Web to access their e-mail or bulletin board services such as Blackboard.

Use of Internet in Bioinformatics

Internet is the most potential tool of this information age and it is serving as a platform for Bioinformatics tool. It provides the opportunity to search that information, which was available only by reaching to the information centre.

Areas of Services

The Internet provides various facilities for Bioinformatics, such as;

- Bioinformatics research
- Courses
- Resources
- Biological databases
- Construction tools
- Software resources
- WWW search tools
- Courses of Bioinformatics
- Advanced topics in Bioinformatics
- Scientific databases
- Electronic journals
- Asking queries from the librarian in online manner
- News events and activities such as; announcement for Bioinformatics interest group, meetings on federated databases, molecular biosciences and technology seminars.

World Wide Web Virtual Library: Biotechnology

This directory, provided by Cato Research Ltd., contains over 1000 URLs specific to biotechnology, pharmaceutical development, and related fields. The emphasis is on product development and the delivery of products and services [2].

Subject Specific Sites

These sites are more likely to concentrate on a particular area of Bioinformatics. These sites are further divided into the various areas of Bioinformatics e.g., Codon usage, and Genome analysis/Genomic comparisons, Phylogenetics etc.

General Bioinformatics Web Sites

Many of the sites are offering the same sorts of links and many to other Bioinformatics sites; many have links to a Sequence Retrieval System or other facilities for sequence retrieval. These are categorized as under:

- Academic Sites
- Corporate/Government Sites

Access to Journals

Providing access to journals such as; Nature, Science, Molecular biology and Evolution, Nucleic Acids Research, Bioinformatics, The Journal of Molecular Biology, Genetics, New Scientist, Online Journal of Bioinformatics, Internet Science Journal.

As a Centre for Biotechnology Information

One can explore extensive sites of resources and including newsletters, Bioinformatics databases, and links to the major medical bibliographic databases. It not only connects to textual databases but also to Protein Structure Servers. These include 3DB browser, biomolecular modeling and structural classification of proteins etc. Biotechnologists can reach any Bioinformatics Centers on Internet. From DNA Databank Japan to European Bioinformatics Centre can be reached by using Internet. One can search databases on protein, nucleotides, and protein structure.

Computer

A computer is an electronic device, operating under the control of instructions stored in its own memory that can accept data (input), process the data according to specified rules, produce information (output), and store the information for future use1.

Functionalities of a computer

Any digital computer carries out five functions in gross terms:

- Takes data as input.
- Stores the data/instructions in its memory and use them when required.
- Processes the data and converts it into useful information.
- Generates the output
- Controls all the above four steps.

Computer Components

Any kind of computers consists of **HARDWARE** AND **SOFTWARE**.

Hardware:

Computer hardware is the collection of physical elements that constitutes a computer system. Computer hardware refers to the physical parts or components of a computer such as the monitor, mouse, keyboard, computer data storage, hard drive disk (HDD), system unit (graphic cards, sound cards, memory, motherboard and chips), etc. all of which are physical objects that can be touched.

Software

Software is a generic term for organized collections of computer data and instructions, often broken into two major categories: system software that provides the basic non- task-specific functions of the computer, and application software which is used by users to accomplish specific tasks.

Software Types

A. System software is responsible for controlling, integrating, and managing the individual hardware components of a computer system so that other software and the users of the system see it as a functional unit without having to be concerned with the low-level details such as transferring data from memory to disk, or rendering text onto a display. Generally, system software consists of an operating system and some fundamental utilities such as disk formatters, file managers, display managers, text editors, user authentication (login) and management tools, and networking and device control software.

B. Application software is used to accomplish specific tasks other than just running the computer system. Application software may consist of a single program, such as an image viewer; a small collection of programs (often called a software package) that work closely together to accomplish a task, such as a spreadsheet or text processing system; a larger collection (often called a software suite) of related but independent programs and packages that have a common user interface or shared data format, such as Microsoft Office, which consists of closely integrated word processor, spreadsheet, database, etc.; or a software system, such as a database management system, which is a collection of fundamental programs that may provide some service to a variety of other independent applications.

Comparison Application Software and System Software

	System Software	Application Software
Definition	Computer software or just software is a general term primarily used for digitally stored data such as computer programs and other kinds of information read and written by computers. App comes under computer software though it has a wide scope now.	Application software, also known as an application or an "app", is computer software designed to help the user to perform specific tasks.
Example:	1) Microsoft Windows 2) Linux 3) Unix 4) Mac OSX 5) DOS	1) Opera (Web Browser) 2) Microsoft Word (Word Processing) 3) Microsoft Excel (Spreadsheet software) 4) MySQL (Database Software) 5) Microsoft PowerPoint (Presentation) 6) Adobe Photoshop (Graphics Software)
Interaction:	Generally, users do not interact with system software as it works in the background.	Users always interact with application software while doing different activities.
Dependency:	System software can run independently of the application software.	Application software cannot run without the presence of the system software.

Characteristics of Computer

To understand why computers are such an important part of our lives, let us look at some of its characteristics –

- Speed − Typically, a computer can carry out 3-4 million instructions per second.
- Accuracy − Computers exhibit a very high degree of accuracy. Errors that may occur are usually due to inaccurate data, wrong instructions or bug in chips – all human errors.

- Reliability − Computers can carry out same type of work repeatedly without throwing up errors due to tiredness or boredom, which are very common among humans.
- Versatility − Computers can carry out a wide range of work from data entry and ticket booking to complex mathematical calculations and continuous astronomical observations. If you can input the necessary data with correct instructions, computer will do the processing.
- Storage Capacity − Computers can store a very large amount of data at a fraction of cost of traditional storage of files. Also, data is safe from normal wear and tear associated with paper.

Computers are now classified on the basis of their use or size

- Desktop
- Laptop
- Tablet
- Server
- Mainframe
- Supercomputer

Computer Operating System

System software that is responsible for functioning of all hardware parts and their interoperability to carry out tasks successfully is called operating system (OS). OS is the first software to be loaded into computer memory when the computer is switched on and this is called booting. OS manages a computer's basic functions like storing data in memory, retrieving files from storage devices, scheduling tasks based on priority, etc.

Bioinformatics: Software

JAVA in Bioinformatics:

Since research centers are scattered all around the globe ranging from private to academic settings, and a range of hardware and OSs are being used, Java is emerging as a key player in bioinformatics. Physiome Sciences' computer-based biological simulation technologies and Bioinformatics Solutions' PatternHunter are two examples of the growing adoption of Java in bioinformatics.

Perl in Bioinformatics:

String manipulation, regular expression matching, file parsing, data format interconversion etc are the common text-processing tasks performed in bioinformatics. Perl excels in such tasks and is being used by many developers. Yet, there are no standard modules designed in Perl specifically for the field of bioinformatics. However, developers have designed

several of their own individual modules for the purpose, which have become quite popular and are coordinated by the BioPerl project.

Bioinformatics Projects:

BioJava:

The BioJava Project is dedicated to providing Java tools for processing biological data which includes objects for manipulating sequences, dynamic programming, file parsers, simple statistical routines, etc.

BioPerl:

The BioPerl project is an international association of developers of Perl tools for bioinformatics and provides an online resource for modules, scripts and web links for developers of Perl-based software.

BioXML:

A part of the BioPerl project, this is a resource to gather XML documentation, DTDs and XML aware tools for biology in one location.

Biocorba:

Interface objects have facilitated interoperability between bioperl and other perl packages such as Ensembl and the Annotation Workbench. However, interoperability between bioperl and packages written in other languages requires additional support software. CORBA is one such framework for interlanguage support, and the biocorba project is currently implementing a CORBA interface for bioperl. With biocorba, objects written within bioperl will be able to communicate with objects written in biopython and biojava (see the next subsection). For more information, see the biocorba project website at http://biocorba.org/ . The Bioperl BioCORBA server and client bindings are available in the bioperl-corba-server and bioperl-corba-client bioperl CVS repositories respecitively. (see http://cvs.bioperl.org/ for more information).

Ensembl :

Ensembl is an ambitious automated-genome-annotation project at EBI. Much of Ensembl\'s code is based on bioperl, and Ensembl developers, in turn, have contributed significant pieces of code to bioperl. In particular, the bioperl code for automated sequence annotation has been largely contributed by Ensembl developers. Describing Ensembl and its capabilities is far beyond the scope of this tutorial The interested reader is referred to the Ensembl website at http://www.ensembl.org/.

bioperl-db:

Bioperl-db is a relatively new project intended to transfer some of Ensembl's capability of integrating bioperl syntax with a standalone Mysql database (http://www.mysql.com) to the bioperl code-base. More details on bioperl-db can be found in the bioperl-db CVS directory at http://cvs.bioperl.org/cgi-bin/viewcvs/viewcvs.cgi/bioperl-db/?cvsroot=bioperl. It is worth mentioning that most of the bioperl objects mentioned above map directly to tables in the bioperl-db schema. Therefore object data such as sequences, their features, and annotations can be easily loaded into the databases, as in $loader->store($newid,$seqobj) Similarly one can query the database in a variety of ways and retrieve arrays of Seq objects. See biodatabases.pod, Bio::DB::SQL::SeqAdaptor, Bio::DB::SQL::QueryConstraint, and Bio::DB::SQL::BioQuery for examples.

Biopython and biojava:

Biopython and biojava are open source projects with very similar goals to bioperl. However their code is implemented in python and java, respectively. With the development of interface objects and biocorba, it is possible to write java or python objects which can be accessed by a bioperl script, or to call bioperl objects from java or python code. Since biopython and biojava are more recent projects than bioperl, most effort to date has been to port bioperl functionality to biopython and biojava rather than the other way around. However, in the future, some bioinformatics tasks may prove to be more effectively implemented in java or python in which case being able to call them from within bioperl will become more important. For more information, go to the biojava http://biojava.org/ and biopython http://biopython.org/ websites.