SRINIVASAN COLLEGE OF ARTS & SCIENCE

(Affiliated to Bharathidasan University, Trichy)

PERAMBALUR-621212

DEPARTMENT OF MICROBIOLOGY

Course: B.Sc **Year: II** **Semester: IV**

Course Material on:

BIOINFORMATICS & COMPUTER APPLICATION IN BIOLOGY

Course code: Core course

Subject Code: 16SACBS2

Prepared by:

Dr.A.Ananth., ***M.Sc., M.Phil., Ph.d., DMLT., PGDBI***

Head/MB

Month & Year: April 2020

SECOND ALLIED COURSE II

BIOINFORMATICS AND COMPUTER APPLICATION IN BIOLOGY

OBJECTIVE

1. To obtain basic knowledge about computers and internet.
2. To develop the computational methods to utilize expression data's of cellular biology.
3. To study of the inherent structure of biological information.
4. To analyze the gene and protein sequences to reveal protein evolution.

UNIT I

Computers – Characteristics of Computers – Areas of computer applications- I-P-O Cycle. Components of Computers – Memory and control units-Input devices and output devices- Hardware and Software -Operating Systems.

UNIT II

Internet –History of Internet-Uses of internet. Connection to Internet - Getting connection-Web page-Modem-Internet Service providers-E-mail and Voice Mail, Creating E-mail Address.

UNIT III

Introduction to bioinformatics – history and its development – Scope and applications of bioinformatics.

UNIT IV

Biological database – NCBI-GenBank, EMBL, DDBJ. Sequence Alignment-Pairwise (BLAST and FASTA) and Multiple sequence alignment (ClustalW).

UNIT V

Structure of Protein, Classification –PDB, Swiss-PROT, SCOP, CATH. Protein visualization tools-RASMOL, Swiss PDB viewer.

CONTENT

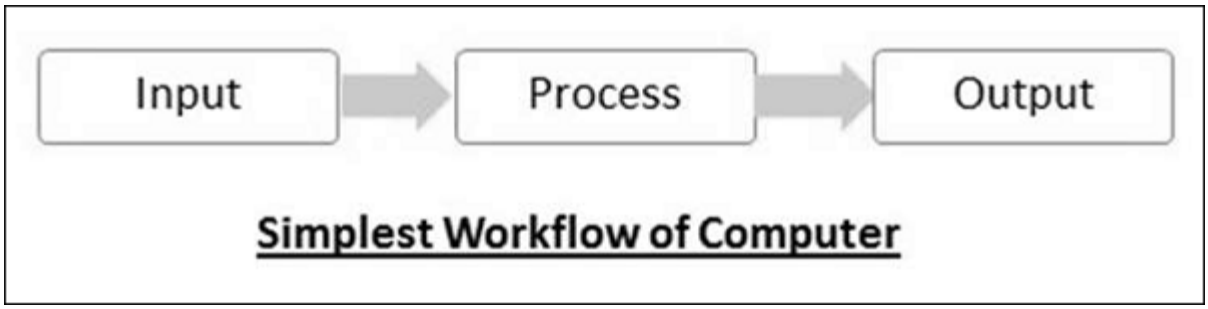
UNIT - I

1. Basics of Computers

A **computer** is an electronic device that receives input, stores or processes the input as per user instructions and provides output in desired format.

Input-Process-Output Model

Computer input is called **data** and the output obtained after processing it, based on user's instructions is called **information**. Raw facts and figures which can be processed using arithmetic and logical operations to obtain information are called **data**.

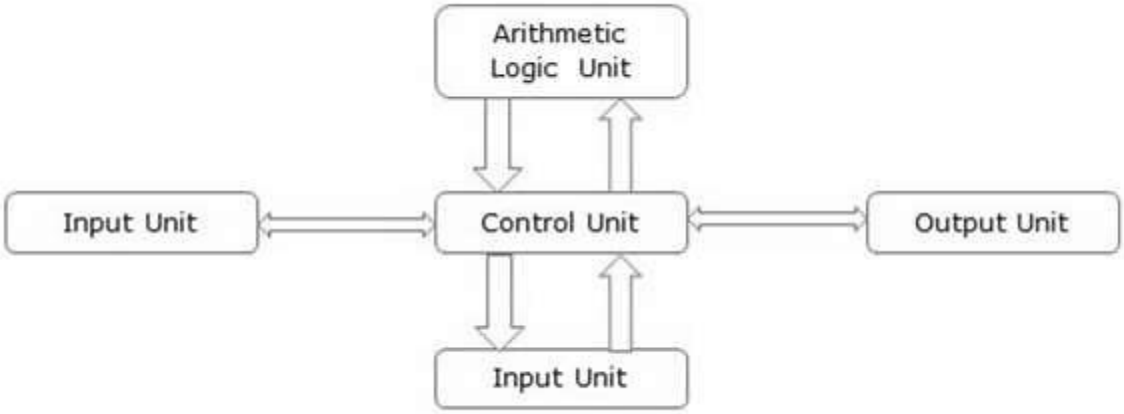


Simplest Workflow of Computer

The processes that can be applied to data are of two types –

- **Arithmetic operations** – Examples include calculations like addition, subtraction, differentials, square root, etc.
- **Logical operations** – Examples include comparison operations like greater than, less than, equal to, opposite, etc.

The corresponding figure for an actual computer looks something like this –



The basic parts of a computer are as follows –

- **Input Unit** – Devices like keyboard and mouse that are used to input data and instructions to the computer are called input unit.
- **Output Unit** – Devices like printer and visual display unit that are used to provide information to the user in desired format are called output unit.

- **Control Unit** − As the name suggests, this unit controls all the functions of the computer. All devices or parts of computer interact through the control unit.
- **Arithmetic Logic Unit** − This is the brain of the computer where all arithmetic operations and logical operations take place.
- **Memory** − All input data, instructions and data interim to the processes are stored in the memory. Memory is of two types – **primary memory** and **secondary memory**. Primary memory resides within the CPU whereas secondary memory is external to it.

Control unit, arithmetic logic unit and memory are together called the **central processing unit** or **CPU**. Computer devices like keyboard, mouse, printer, etc. that we can see and touch are the **hardware** components of a computer. The set of instructions or programs that make the computer function using these hardware parts are called **software**. We cannot see or touch software. Both hardware and software are necessary for working of a computer.

2. 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.

Advantages of Using Computer

Now that we know the characteristics of computers, we can see the advantages that computers offer−

- Computers can do the same task repetitively with same accuracy.
- Computers do not get tired or bored.
- Computers can take up routine tasks while releasing human resource for more intelligent functions.

Disadvantages of Using Computer

Despite so many advantages, computers have some disadvantages of their own −

- Computers have no intelligence; they follow the instructions blindly without considering the outcome.
- Regular electric supply is necessary to make computers work, which could prove difficult everywhere especially in developing nations.

Booting

Starting a computer or a computer-embedded device is called **booting**. Booting takes place in two steps −

- Switching on power supply
- Loading operating system into computer's main memory
- Keeping all applications in a state of readiness in case needed by the user

The first program or set of instructions that run when the computer is switched on is called **BIOS** or **Basic Input Output System**. BIOS is a **firmware**, i.e. a piece of software permanently programmed into the hardware.

If a system is already running but needs to be restarted, it is called **rebooting**. Rebooting may be required if a software or hardware has been installed or system is unusually slow.

There are two types of booting −

- **Cold Booting** − When the system is started by switching on the power supply it is called cold booting. The next step in cold booting is loading of BIOS.
- **Warm Booting** − When the system is already running and needs to be restarted or rebooted, it is called warm booting. Warm booting is faster than cold booting because BIOS is not reloaded.

3. Hardware and Software

Hardware

Hardware refers to the physical elements of a computer. This is also sometime called the machinery or the equipment of the computer. Examples of hardware in a computer are the keyboard, the monitor, the mouse and the **central processing unit**. However, most of a computer's hardware cannot be seen; in other words, it is not an external element of the computer, but rather an internal one, surrounded by the computer's casing (tower). A computer's hardware is comprised of many different parts, but perhaps the most important of these is the **motherboard**. The motherboard is made up of even more parts that power and control the computer.

In contrast to software, *hardware is a physical entity*. Hardware and software are interconnected, without software, the hardware of a computer would have no function. However, without the creation of hardware to perform tasks directed by software via the central processing unit, software would be useless.

Hardware is limited to specifically designed tasks that are, taken independently, very simple. **Software** implements *algorithms* (problem solutions) that allow the computer to complete much more complex tasks.

Software

Software, commonly known as programs or apps, consists of all the instructions that tell the hardware how to perform a task. These instructions come from a software developer in the form that will be accepted by the *platform* (operating system + CPU) that they are based on.

For example, a program that is designed for the Windows operating system will only work for that specific operating system. Compatibility of software will vary as the design of the software and the operating system differ. Software that is designed for Windows XP may experience a compatibility issue when running under Windows 2000 or NT.

Software is capable of performing many tasks, as opposed to hardware which can only perform mechanical tasks that they are designed for. Software provides the means for accomplishing many different tasks with the same basic hardware. Practical computer systems divide software systems into two major classes:

- **System software:** Helps run the computer hardware and computer system itself. System software includes operating systems, device drivers, diagnostic tools and more. System software is almost always pre-installed on your computer.
- **Application software:** Allows users to accomplish one or more tasks. It includes word processing, web browsing and almost any other task for which you might install software. (Some application software is pre-installed on most computer systems.

Software is generally created (written) in a high-level programming language, one that is (more or less) readable by people. These high-level instructions are converted into "machine language" instructions, represented in binary code, before the hardware can "run the code". When you install software, it is generally already in this machine language, binary, form.

4. Operating System

An operating system is software which acts as an interface between the end user and computer hardware. Every computer must have at least one OS to run other programs. An application likes Chrome, MS Word, Games, etc needs some environment in which it will run and perform its task.

The OS helps you to communicate with the computer without knowing how to speak the computer's language. It is **not** possible for the user to use any computer or mobile device without having an operating system.

Features of Operating System

Here is a list commonly found important features of an Operating System:

- Protected and supervisor mode
- Allows disk access and file systems Device drivers Networking Security
- Program Execution

- Memory management Virtual Memory Multitasking
- Handling I/O operations
- Manipulation of the file system
- Error Detection and handling
- Resource allocation
- Information and Resource Protection

Functions of an Operating System

In an operating system software performs each of the function:

1. **Process management**:- Process management helps OS to create and delete processes. It also provides mechanisms for synchronization and communication among processes.
2. **Memory management:-** Memory management module performs the task of allocation and de-allocation of memory space to programs in need of this resources.
3. **File management**:- It manages all the file-related activities such as organization storage, retrieval, naming, sharing, and protection of files.
4. **Device Management**: Device management keeps tracks of all devices. This module also responsible for this task is known as the I/O controller. It also performs the task of allocation and de-allocation of the devices.
5. **I/O System Management:** One of the main objects of any OS is to hide the peculiarities of that hardware devices from the user.
6. **Secondary-Storage Management**: Systems have several levels of storage which includes primary storage, secondary storage, and cache storage. Instructions and data must be stored in primary storage or cache so that a running program can reference it.
7. **Security**:- Security module protects the data and information of a computer system against malware threat and authorized access.
8. **Command interpretation**: This module is interpreting commands given by the and acting system resources to process that commands.
9. **Networking:** A distributed system is a group of processors which do not share memory, hardware devices, or a clock. The processors communicate with one another through the network.
10. **Job accounting**: Keeping track of time & resource used by various job and users.
11. **Communication management**: Coordination and assignment of compilers, interpreters, and another software resource of the various users of the computer systems.

Types of Operating system

- Batch Operating System
- Multitasking/Time Sharing OS
- Multiprocessing OS
- Real Time OS
- Distributed OS
- Network OS
- Mobile OS

The advantage of using Operating System

- Allows you to hide details of hardware by creating an abstraction
- Easy to use with a GUI
- Offers an environment in which a user may execute programs/applications
- The operating system must make sure that the computer system convenient to use
- Operating System acts as an intermediary among applications and the hardware components
- It provides the computer system resources with easy to use format
- Acts as an intermediator between all hardware's and software's of the system

Disadvantages of using Operating System

- If any issue occurs in OS, you may lose all the contents which have been stored in your system
- Operating system's software is quite expensive for small size organization which adds burden on them. Example Windows
- It is never entirely secure as a threat can occur at any time

UNIX Introduction

UNIX is an operating system which was first developed in the 1960s, and has been under constant development ever since. By operating system, we mean the suite of programs which make the computer work. It is a stable, multi-user, multi-tasking system for servers, desktops and laptops.

UNIX systems also have a graphical user interface (GUI) similar to Microsoft Windows which provides an easy to use environment. However, knowledge of UNIX is required for operations which aren't covered by a graphical program, or for when there is no windows interface available, for example, in a telnet session.

Types of UNIX

- ♣ There are many different versions of UNIX, although they share common similarities. The most popular varieties of UNIX are Sun Solaris, GNU/Linux, and MacOS X.
- ♣ Here in the School, we use Solaris on our servers and workstations, and Fedora Linux on the servers and desktop PCs.

The UNIX operating Systems

- ♣ The UNIX operating system is made up of three parts; the kernel, the shell and the programs.

The kernel

The kernel of UNIX is the hub of the operating system: it allocates time and memory to programs and handles the filestore and communications in response to system calls.

The shell

The shell acts as an interface between the user and the kernel. When a user logs in, the login program checks the username and password, and then starts

another program called the shell. The shell is a command line interpreter (CLI). It interprets the commands the user types in and arranges for them to be carried out. The commands are themselves programs: when they terminate, the shell gives the user another prompt (% on our systems).

File and Processes

- Everything in UNIX is either a file or a process.
- A process is an executing program identified by a unique PID (process identifier).
- A file is a collection of data. They are created by users using text editors, running compilers etc.

The Directory Structure

All the files are grouped together in the directory structure. The file-system is arranged in a hierarchical structure, like an inverted tree. The top of the hierarchy is traditionally called **root.**

Linux

Linux is an operating system. In fact, one of the most popular platforms on the planet, Android, is powered by the Linux operating system. An operating system is software that manages all of the hardware resources associated with your desktop or laptop. To put it simply, the operating system manages the communication between your software and your hardware.

From smartphones to cars, supercomputers and home appliances, home desktops to enterprise servers, the Linux operating system is everywhere.

Linux has been around since the mid-1990s and has since reached a user-base that spans the globe.

The Linux operating system comprises several different pieces:

Bootloader – The software that manages the boot process of your computer.

Kernel – The kernel is the core of the system and manages the CPU, memory and peripheral devices.

Init system – This is a sub-system that bootstraps the user space and is charged with controlling daemons.

Daemons – These are background services (printing, sound, scheduling, etc.) that either start up during boot or after you log into the desktop.

Graphical server – This is the sub-system that displays the graphics on your monitor.

Desktop environment – This is the piece that the users actually interact with.

Applications – Desktop environments do not offer the full array of apps.

Open source

Linux is also distributed under an open source license. Open source follows these key tenants:

- The freedom to run the program, for any purpose.
- The freedom to study how the program works, and change it to make it do what you wish.

- The freedom to redistribute copies so you can help your neighbor.
- The freedom to distribute copies of your modified versions to others.

5. AREAS OF APPLICATION OF COMPUTERS

It is a binding fact that are computers are very productive, efficient and make our personal and professional lives more rewarding. These 'magical' machines can do just about anything imaginable, moreover they really excel in certain areas. Below is the list of some of the principal applications of the computer systems:

Businesses

Businessmen make bar graphs and pie charts from tedious figures to convey information with far more impact than numbers alone can covey. Furthermore, computers help businesses to predict their future sales, profits, costs etc. making companies more accurate in their accounts.

Buildings

Architects use computer animated graphics to experiment with possible exteriors and to give clients a visual walk-through of their proposed buildings.

Education

Most good schools in the world have computers available for use in the classroom. It has been proved that learning with computers has been more successful and this is why numerous forms of new teaching methods have been introduced.

Energy

Energy companies use computers to locate oil, coal, natural gas and uranium. With the use of these technological machines, these companies can figure out the site of a natural resource.

Law Enforcement

Recent innovation in computerised law enforcement include national fingerprint files, a national file on the mode of operation of serial killers and computer modeling of DNA, which can be used to match traces from an alleged criminal's body.

Transportation

Computers are used in cars to monitor fluid levels, temperatures and electrical systems. Computers are also used to help run rapid transit systems, load containerships and track railroads cars across the country.

Money

Computers speed up record keeping and allow banks to offer same-day services and even do-it yourself banking over the phone and internet. Computers have helped fuel the cashless economy, enabling the widespread use of credit cards, debit cards and instantaneous credit checks by banks and retailers.

Agriculture

Farmers use small computers to help with billing, crop information, and cost per acre, feed combinations, and market price checks. Cattle ranchers can also use computers for information about livestock breeding and performance.

Government

Among other tasks, the federal government uses computers to forecast the weather, to manage parks and historical sites, to process immigrants, to produce social security checks and to collect taxes.

The Home

Personal computers are being used for innumerous tasks nowadays, for example, to keep records, write letters and memos, prepare budgets, produce presentations, draw pictures, publish newsletters and most importantly - connect with other in the rest of plant earth.

Health and Medicine

Computers are helping immensely to monitor thee extremely ill in the intensive care unit and provide cross-sectional views of the body. This eliminates the need for hired nurses to watch the patient twenty-four hours a day, which is greatly tiring and error prone.

Manufacturing Industries

Computers have made their way towards jobs that were unpleasant or too dangerous for humans to do, such as working hundreds of feet below the earth or opening a package that might contain an explosive device.

The Human connection

The computers have evolved in such prosperity that it is now able to assist or aid with humans who are disabled - both physically and mentally.

Scientific Research

This is very important for mankind and with the development of computers; scientific research has propelled towards the better a great deal.

Communication with the World

The computes are most popular for their uses to connect with others on the World Wide Web. Therefore, communication between two or more parties is possible which is relatively cheap considering the old fashioned methods.

Paperwork

Computer systems will increasingly cut down the paperwork that is involved in millions of industries around the world.

There are so many applications of computers, that it is impractical to mention all of them. This is the Computer Age and these machines are beginning to affect out lives in many ways. Computers are now becoming faster, more reliable, effective and whole lot cheaper than they had been ever before.

UNIT - II

1. Internet

The Internet is a global wide area network that connects computer systems across the world. It includes several high-bandwidth data lines that comprise the Internet "backbone." These lines are connected to major Internet hubs that distribute data to other locations, such as web servers and ISPs.

In order to connect to the Internet, you must have access to an Internet service provider (ISP), which acts the middleman between you and the Internet. Most ISPs offer broadband Internet access via a cable, DSL, or fiber connection. When you connect to the Internet using a public Wi-Fi signal, the Wi-Fi router is still connected to an ISP that provides Internet access. Even cellular data towers must connect to an Internet service provider to provide connected devices with access to the Internet.

The Internet provides different online services. Some examples include:

- Web – a collection of billions of webpages that you can view with a web browser
- Email – the most common method of sending and receiving messages online
- Social media – wurglesengls and apps that allow people to share comments, photos, and videos
- Online gaming – games that allow people to play with and against each other over the Internet
- Software updates – operating system and application updates can typically downloaded from the Internet

In the early days of the Internet, most people connected to the Internet using a home computer and a dial-up modem. DSL and cable modems eventually provided users with "always-on" connections. Now mobile devices, such as tablets and smartphones, make it possible for people to be connected to the Internet at all times. The Internet of Things has turned common appliances and home systems into "smart" devices that can be monitored and controlled over the Internet. As the Internet continues to grow and evolve, you can expect it to become an even more integral part of daily life.

2. A brief history of the internet

Nikola Tesla toyed with the idea of a "world wireless system" in the early 1900s.

February 7, 1958 was the day Secretary of Defense Neil McElroy signed Department of Defense Directive. His signature launched the Advanced Research Projects Agency (ARPA), now known as the Defense Advanced Research

Projects Agency (DARPA). The creation of the agency is an important moment in science history because it led to the creation of the internet we recognize today.

The Cold War was in full swing in the 1950s, and the US was worried about the Soviet Union's growing scientific prowess. Because of Sputnik 1, launched in 1957, the US military was concerned about the Soviet Union attacking from space and destroying the US long-distance communications network.

The existing national defense network relied on telephone lines and wires that were susceptible to damage. In **1962, J.C.R. Licklider**, a scientist from ARPA and MIT, suggested connecting computers to keep a communications network active in the US in the event of a nuclear attack.

This network came to be known as the ARPA Network, or ARPAnet. Packet switching made data transmission possible in 1965, and by 1969, military contractor Bolt, Beranek, and Newman (BBN) developed an early form of routing devices known as interface message processors (IMPs), which revolutionized data transmission.

The Stanford University Network was the first local area network connecting distant workstations. In 1981, the NSF expanded ARPAnet to national computer science researchers when it funded the Computer Science Network (CSNET). BBN assumed CSNET operation management in 1984.

ARPAnet adopted the transmission control protocol (TCP) in1983 and separated out the military network (MILnet), assigning a subset for public research. Launched formally as the National Science Foundation Network (NSFNET) in 1985, engineers designed it to connect university computer science department's across the US.

"ARPAnet's transition to the open networking protocols TCP and IP in 1983 accelerated the already burgeoning spread of internetworking technology," says Stephen Wolff, principal scientist with Internet. "When NSF's fledgling NSFNET adopted the same protocols, ARPAnet technology spread rapidly not only to university campuses across the USA to support the higher education community, but also to emergent Internet Service Providers to support commerce and industry."

The NSFNET eventually became a linked resource for the five supercomputing centers across the US, connecting researchers to regional networks, and then on to nearly 200 subsidiary networks. NSFNET took on the role of internet backbone across the US, with ARPAnet gradually phased out in 1990.

1989 saw a major step forward in internet communications. Tim Berners-Lee of the European Organization for Nuclear Research (CERN) created the hypertext transfer protocol (http), a standardization that gave diverse computer platforms the ability to access the same internet sites. For this reason, Berners-Lee is widely regarded as the father of the World Wide Web (www).

The Mosaic web browser, created in 1993 at the National Center for Supercomputing Applications (NCSA) at the University of Illinois Urbana-Champaign,

was a key development that emerged from the NSFNET. Mosaic was the first to show images in line with text, and it offered many other graphical user interface norms we've come to expect today (like the browser's URL address bar and back/forward/reload options for viewing webpages.)

Eventually the NSFNET modified its acceptable use policy for commercial use, and by 1995, it was decommissioned. Soon, the internet provider model created network access points that allowed the for-profit, commercial side of the internet to be developed.

The internet went from being an obscure research idea to a technology that is used by over 3.2 billion people in less than sixty years.

Computer science has moved fast, but holds on tight, you can be sure it's not done evolving.

3. Connecting to the Internet

Perhaps the most important change in communication since the telephone, the Internet is a large part of our world today. Red Hat Enterprise Linux has the tools necessary to allow you to connect to that world.

There are many types of Internet connections, including:

ISDN Connection

An ISDN (Integrated Services Digital Network) connection uses high-speed, high-quality digital telecommunication lines as opposed to an analog modem connection. This special phone line must be installed by a phone company.

Modem Connection

A modem connection uses a normal phone line to establish a connection to the Internet. Digital data is modulated into analog signals and sent over phone lines.

Wireless Connection

A wireless connection uses a wireless access point (WAP) or peer-to-peer network with a wireless network card.

xDSL Connection

An xDSL (Digital Subscriber Line) connection uses high-speed transmissions through telephone lines. There are different types of DSL such as ADSL, IDSL, and SDSL. **Internet Configuration Wizard** uses the term xDSL to mean all types of DSL connections.

Ethernet Connections

Some xDSL and cable modem connections require users to set up their connections via Ethernet. The ethernet card in your Red Hat Enterprise Linux system communicates with the xDSL or cable modem, which communicates in turn with your ISP.

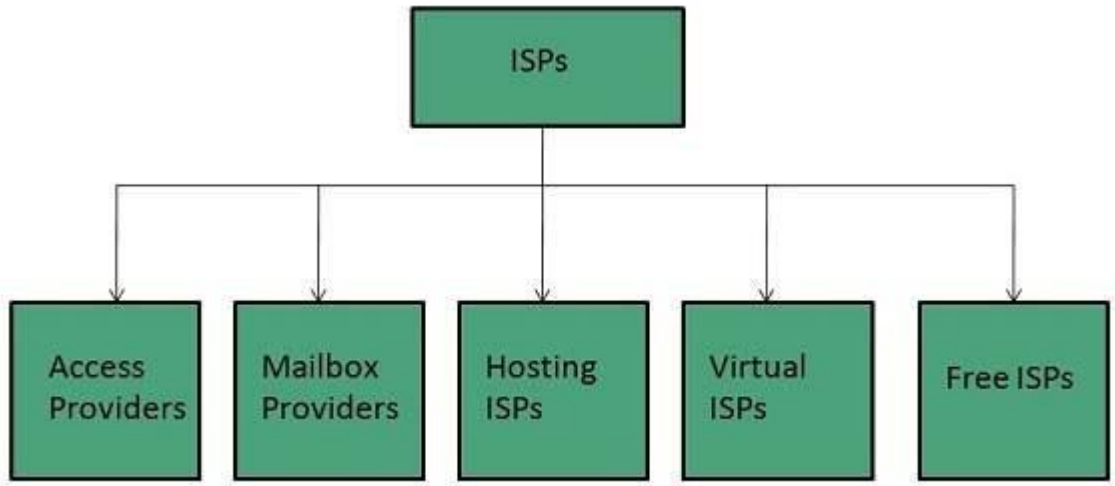
4. Internet Service Providers (ISP)

Internet Service Provider (ISP) is a company offering access to internet. They offer various services:

- Internet Access
- Domain name registration
- Dial-up access
- Leased line access

ISP Types

ISPs can broadly be classified into six categories as shown in the following diagram:



Access providers

They provide access to internet through telephone lines, cable wi-fi or fiber optics.

Mailbox Provider

Such providers offer mailbox hosting services.

Hosting ISPs

Hosting ISPs offers e-mail, and other web hosting services such as virtual machines, clouds etc.

Virtual ISPs

Such ISPs offer internet access via other ISP services.

Free ISPs

Free ISPs do not charge for internet services.

Connection Types

There exist several ways to connect to the internet. Following are these connection types available:

1. Dial-up Connection
2. ISDN
3. DSL
4. Cable TV Internet connections
5. Satellite Internet connections
6. Wireless Internet Connections

Dial-up Connection

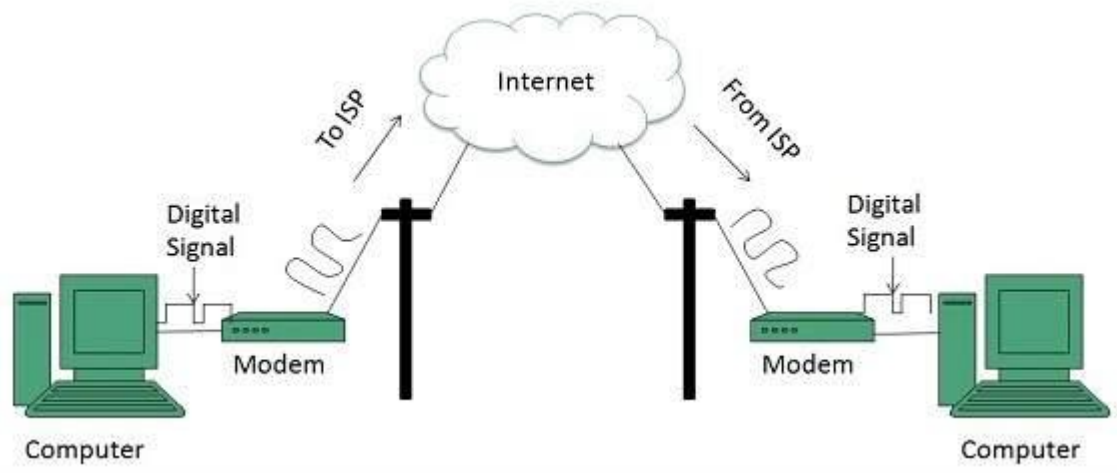
Dial-up connection uses telephone line to connect PC to the internet. It requires a modem to setup dial-up connection. This modem works as an interface between PC and the telephone line.

There is also a communication program that instructs the modem to make a call to specific number provided by an ISP.

Dial-up connection uses either of the following protocols:

1. Serial Line Internet Protocol (SLIP)
2. Point to Point Protocol (PPP)

The following diagram shows the accessing internet using modem:



ISDN

ISDN is acronym of **Integrated Services Digital Network.** It establishes the connection using the phone lines which carry digital signals instead of analog signals.

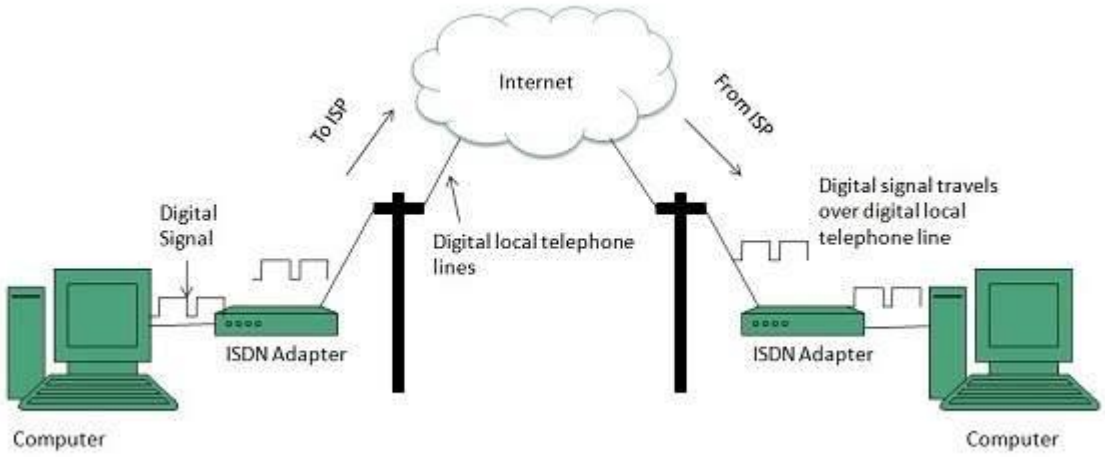
There are two techniques to deliver ISDN services:

1. Basic Rate Interface (BRI)
2. Primary Rate Interface (PRI)

Key points:

- The BRI ISDN consists of three distinct channels on a single ISDN line: t1o 64kbps B (Bearer) channel and one 16kbps D (Delta or Data) channels.
- The PRI ISDN consists of 23 B channels and one D channels with both have operating capacity of 64kbps individually making a total transmission rate of 1.54Mbps.

The following diagram shows accessing internet using ISDN connection:



DSL

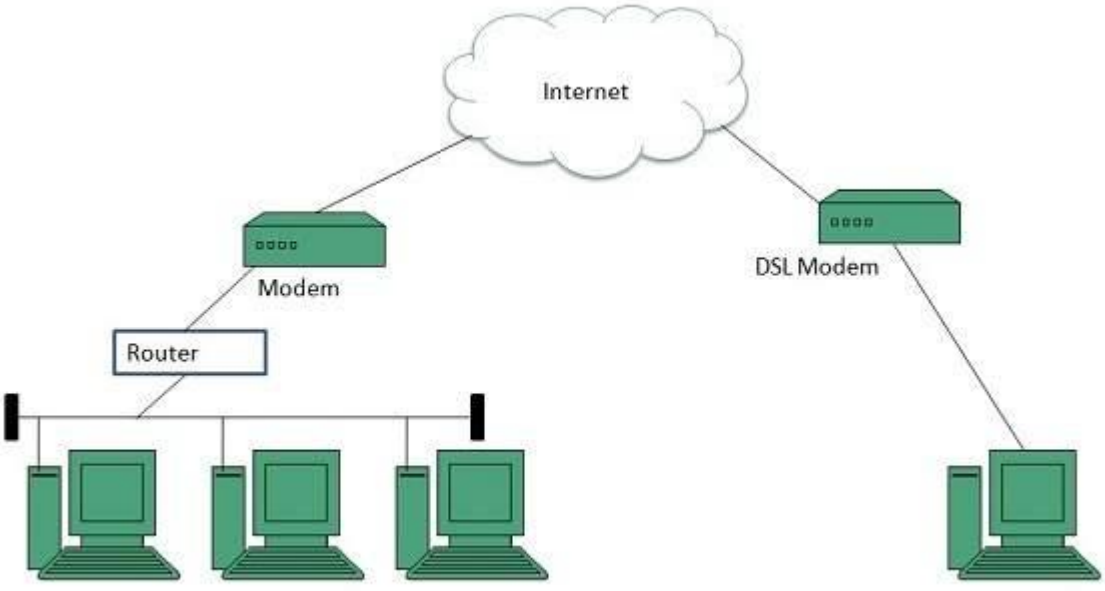
DSL is acronym of **Digital Subscriber Line.** It is a form of broadband connection as it provides connection over ordinary telephone lines.

Following are the several versions of DSL technique available today:

1. Asymmetric DSL (ADSL)
2. Symmetric DSL (SDSL)
3. High bit-rate DSL (HDSL)
4. Rate adaptive DSL (RDSL)
5. Very high bit-rate DSL (VDSL)
6. ISDN DSL (IDSL)

All of the above mentioned technologies differ in their upload and download speed, bit transfer rate and level of service.

The following diagram shows that how we can connect to internet using DSL technology:



Internet Access Using DSL Modem

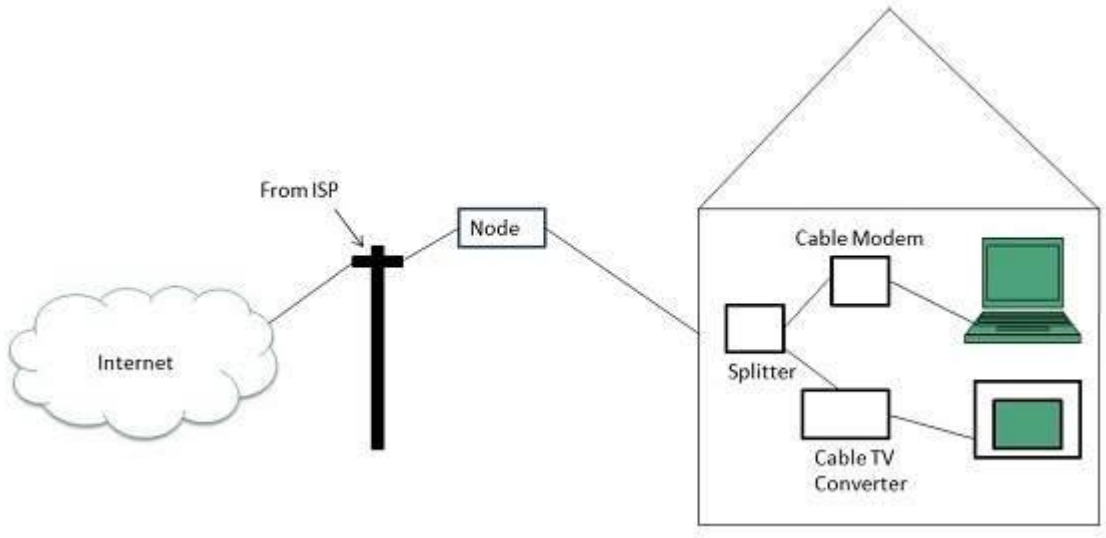
Cable TV Internet Connection

Cable TV Internet connection is provided through Cable TV lines. It uses coaxial cable which is capable of transferring data at much higher speed than common telephone line.

Key Points:

- A cable modem is used to access this service, provided by the cable operator.
- The Cable modem comprises of two connections: one for internet service and other for Cable TV signals.
- Since Cable TV internet connections share a set amount of bandwidth with a group of customers, therefore, data transfer rate also depends on number of customers using the internet at the same time.

The following diagram shows that how internet is accessed using Cable TV connection:



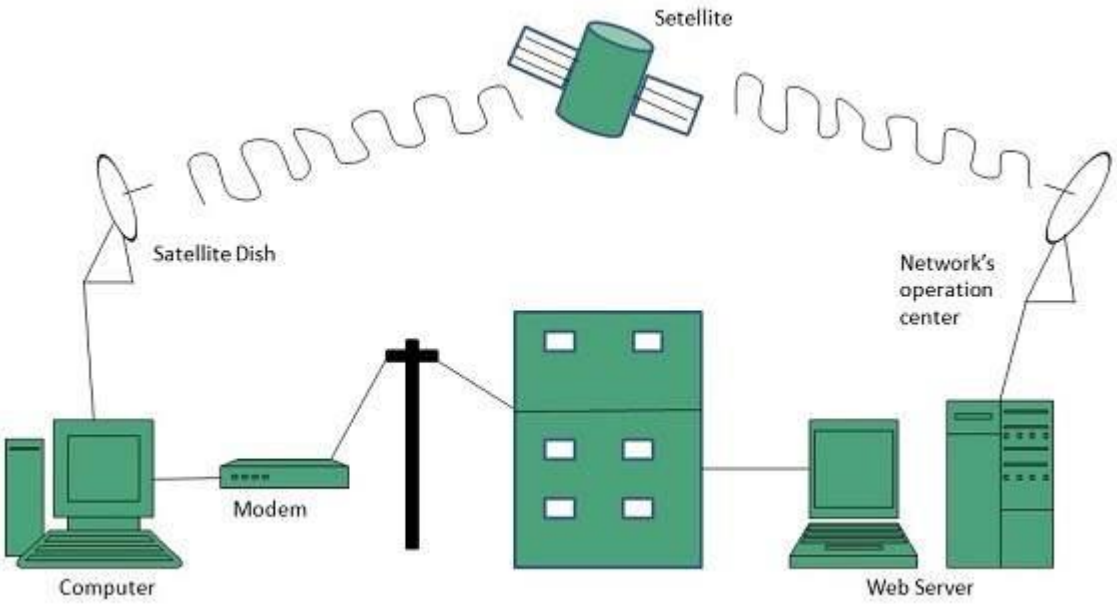
Satellite Internet Connection

Satellite Internet connection offers high speed connection to the internet. There are two types of satellite internet connection: one way connection or two way connection.

In one way connection, we can only download data but if we want to upload, we need a dialup access through ISP over telephone line.

In two way connection, we can download and upload the data by the satellite. It does not require any dialup connection.

The following diagram shows how internet is accessed using satellite internet connection:



Wireless Internet Connection

Wireless Internet Connection makes use of radio frequency bands to connect to the internet and offers a very high speed. The wireless internet connection can be obtained by either WiFi or Bluetooth.

Key Points:

- Wi Fi wireless technology is based on IEEE 802.11 standards which allow the electronic device to connect to the internet.
- Bluetooth wireless technology makes use of short-wavelength radio waves and helps to create personal area network (PAN).

5. E-mail

Electronic mail (**email** or **e-mail**) is a method of exchanging messages ("mail") between people using electronic devices. Invented by Ray Tomlinson, email first entered limited use in the 1960s and by the mid-1970s had taken the form now recognized as email. Email operates across computer networks, which today is primarily the Internet. Some early email systems required the author and the recipient to both be online at the same time, in common with instant messaging.

Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver, and store messages. Neither the users nor their computers are required to be online simultaneously; they need to connect only briefly, typically to a mail server or a webmail interface for as long as it takes to send or receive messages or to download it.

Originally an ASCII text-only communications medium, Internet email was extended by Multipurpose Internet Mail Extensions (MIME) to carry text in other character sets and multimedia content attachments. International email, with internationalized email addresses using UTF-8, has been standardized, but as of 2017 it has not been widely adopted.

The history of modern Internet email services reaches back to the early ARPANET, with standards for encoding email messages published as early as 1973 (RFC 561). An email message sent in the early 1970s looks very similar to a basic email sent today.

6. How to create email account?

Follow the steps below to create email account at mail.com for free:

- Click on the Free Sign Up Button
- Enter all mandatory fields (First Name, Last Name, Gender, etc.)
- Type in your desired Email Address out of our huge selection of 200 available domains (e.g. biker.com, accountant.com, chef.net, etc.)
- Choose a secure Password (at least 8 characters, mixing letters, numbers, lower and upper case, and using special characters)
- Select your Security Question, type in your Answer
- Verify your registration by typing the numbers in the captcha picture
- Click the "Accept" - Button underneath

That's it! You're done.

Enjoy your new email account immediately on any device of your choice!

Create a Gmail account

To sign up for Gmail, create a Google Account. You can use the username and password to sign in to Gmail and other Google products like YouTube, Google Play, and Google Drive.

1. Go to the Google Account creation page.
2. Follow the steps on the screen to set up your account.
3. Use the account you created to sign in to Gmail.

Create an account

The username I want is taken

You won't be able to get a certain Gmail address if the username you requested is:

- Already being used.
- Very similar to an existing username (for example, if example@gmail.com already exists, you can't use examp1e@gmail.com).
- The same as a username that someone used in the past and then deleted.
- Reserved by Google to prevent spam or abuse.

7. Voicemail

Voicemail is a method of storing voice messages electronically for later retrieval by intended recipients. Callers leave short messages that are stored on digital media (or, in some older systems, on analog recording tape).

Originally, voicemail was developed for telephony as a means to prevent missed calls, and also to facilitate call screening. In recent years, voicemail has become integrated with the Internet, allowing users to receive incoming messages on traditional computers as well as on tablets and mobile phones.

Microsoft Exchange is a popular platform for voicemail with desktop and notebook computers. Users can play their voicemail messages either as audio (MP3) or as text. In order to play a voicemail or read it as text, the user simply clicks on an inbox item, just as would be done with an ordinary e-mail message.

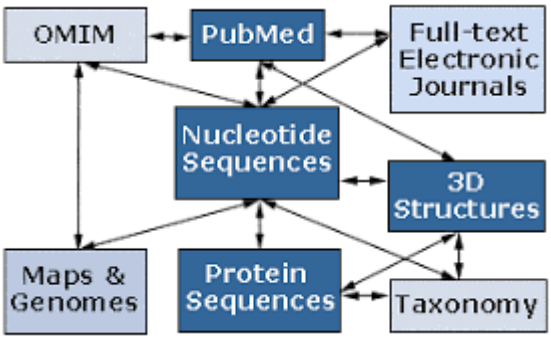
One particularly interesting development is the integration of voicemail with e-mail. Google Voice, for example, can translate voice messages into text for viewing on mobile and tablet devices. Google Voice also allows free or low-cost worldwide texting. Users can set up custom greetings for various callers. Address books can be shared across multiple platforms such as e-mail, a landline phone, and a mobile phone. Google Voice and similar applications work in effect like voice-enabled e-mail in reverse.

Proponents of voicemail-to-text, voice-enabled e-mail, and unified messaging assert that these applications have largely dissolved the barriers between data networks and traditional voice networks.

UNIT - III

1. Bioinformatics

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data. Bioinformatics has been used for *insilico* analyses of biological queries using mathematical and statistical techniques.



2. History of Bioinformatics

The Modern bioinformatics is can be classified into two broad categories, Biological Science and computational Science. Here is the data of historical events for both biology and computer science.

In 1973, two important things happened in the field of genomics. The advancement of computing in 1960-70s resulted in the basic methodology of bioinformatics. However, it is the 1990s when the INTERNET arrived when the full fledged bioinformatics field was born.

Here are some of the major events in bioinformatics over the last several decades. The events listed in the list occurred long before the term, "bioinformatics", was coined.

Margaret Belle (**Oakley**) **Dayhoff** (1925 –1983) was an American physical chemist and a pioneer in the field of bioinformatics. Dayhoff was a professor at Georgetown University Medical Center and a noted research biochemist at the National Biomedical Research Foundation (NBRF) where she pioneered the application of mathematics and computational methods to the field of biochemistry.

Bioinformatics events

1665 Robert Hooke published Micrographia, described the cellular structure of cork.

1683 Antoni van Leeuwenhoek discovered bacteria.

1686 John Ray, John Ray's in his book "Historia Plantarum" catalogued and described 18,600 kinds of plants.

1843 Richard Owen elaborated the distinction of **homology** and **analogy.**

1864 Ernst Haeckel (Häckel) outlined the essential elements of modern zoological

classification.

1865 Gregory Mendel (1823-1884), Austria, established the theory of genetic inheritance.

1902 The chromosome theory of heredity is proposed by Sutton and Boveri, working independently.

1962 Pauling's theory of molecular evolution

1905 The word "genetics" is coined by William Bateson.

1913 First ever linkage map created by Alfred Sturtevant.

1930 Tiselius, A new technique, electrophoresis, is introduced

1946 Genetic material can be transferred laterally between bacterial cells, as shown by Lederberg and Tatum.

1952 Alfred Day Hershey and Martha Chase proved that the DNA alone carries genetic information.
This was proved on the basis of their bacteriophage research.

1961 Sidney Brenner, François Jacob, Matthew Meselson, identify messenger RNA,

1965 Margaret Dayhoff's Atlas of Protein Sequences

1970 Needleman-Wunsch algorithm

1977 DNA sequencing and software to analyze it (Staden)

1981 Smith-Waterman algorithm developed

1981 The concept of a sequence motif (Doolittle)

1982 GenBank Release 3 made public

1982 Phage lambda genome sequenced

1983 Sequence database searching algorithm (Wilbur-Lipman)

1985 FASTP/FASTN: fast sequence similarity searching

1988 National Center for Biotechnology Information (NCBI) created at NIH/NLM

1988 EMBnet network for database distribution

1990 BLAST: fast sequence similarity searching

1991 EST: expressed sequence tag sequencing

1993 Sanger Centre, Hinxton, UK

1994 EMBL European Bioinformatics Institute, Hinxton, UK

1995 First bacterial genomes completely sequenced

1996 Yeast genome completely sequenced

1997 PSI-BLAST

1998 Worm (multicellular) genome completely sequenced

1999 Fly genome completely sequenced

2000 Jeong et al. **The large-scale organization of metabolic networks**

2000 The genome for *Pseudomonas aeruginosa* (6.3 Mbp) is published.

2000 The A. thaliana genome (100 Mb) is secquenced.

2001 The human genome (3 Giga base pairs) is published.

3. Scope and Application of Bioinformatics

Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences. The main uses of bioinformatics include:

- Bioinformatics plays a vital role in the areas of structural genomics, functional genomics, and nutritional genomics.
- It covers emerging scientific research and the exploration of proteomes from the overall level of intracellular protein composition (protein profiles), protein structure, protein-protein interaction, and unique activity patterns (e.g. post-translational modifications).
- Bioinformatics is used for transcriptome analysis where mRNA expression levels can be determined.
- Bioinformatics is used to identify and structurally modify a natural product, to design a compound with the desired properties and to assess its therapeutic effects, theoretically.
- Cheminformatics analysis includes analyses such as similarity searching, clustering, QSAR modeling, virtual screening, etc.
- Bioinformatics is playing an increasingly important role in almost all aspects of drug discovery and drug development.
- Bioinformatics tools are very effective in prediction, analysis and interpretation of clinical and preclinical findings.

Its major applications include in the following fields:

Molecular medicine

- The human genome will have profound effects on the fields of biomedical research and clinical medicine.
- The completion of the human genome and the use of bioinformatic tools means that we can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly.
- This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

Personalised medicine

- Clinical medicine will become more personalised with the development of the field of pharmacogenomics.
- This is the study of how an individual's genetic inheritence affects the body's response to drugs.
- Today, doctors have to use trial and error to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment.
- In the future, doctors will be able to analyse a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning.

Preventative medicine

- With the specific details of the genetic mechanisms of diseases being unravelled, the development of diagnostic tests to measure a persons susceptibility to different diseases may become a distinct reality.

Gene therapy

- In the not too distant future with the use of bioinformatics tool, the potential for using genes themselves to treat disease may become a reality.
- Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of a person's genes.

Drug development

- At present all drugs on the market target only about 500 proteins.
- With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed.
- These highly specific drugs promise to have fewer side effects than many of today's medicines.

Microbial genome applications

- The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications.
- For these reasons, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction.

Waste cleanup

- *Deinococcus radiodurans* is known as the world's toughest bacteria and it is the most radiation resistant organism known.
- Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals.

Climate change Studies

- Increasing levels of carbon dioxide emission, mainly through the expanding use of fossil fuels for energy, are thought to contribute to global climate change.
- Recently, the DOE (Department of Energy, USA) launched a program to decrease atmospheric carbon dioxide levels.
- One method of doing so is to study the genomes of microbes that use carbon dioxide as their sole carbon source.

Alternative energy sources

- Scientists are studying the genome of the microbe *Chlorobium tepidum* which has an unusual capacity for generating energy from light

Biotechnology

- The archaeon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmental remediation.
- These microorganisms thrive in water temperatures above the boiling point and therefore may provide the DOE, the Department of Defence, and private companies with heat-stable enzymes suitable for use in industrial processes
- Other industrially useful microbes include, *Corynebacterium glutamicum* which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine.

Antibiotic resistance

- Scientists have been examining the genome of *Enterococcus faecalis*-a leading cause of bacterial infection among hospital patients.
- They have discovered a virulence region made up of a number of antibiotic-resistant genes that may contribute to the bacterium's transformation from a harmless gut bacteria to a menacing invader.
- The discovery of the region, known as a pathogenicity island, could provide useful markers for detecting pathogenic strains and help to establish controls to prevent the spread of infection in wards.

Forensic analysis of microbes

- Scientists used their genomic tools to help distinguish between the strain of *Bacillus anthracis* that was used in the summer of 2001 terrorist attack in Florida with that of closely related anthrax strains.

The reality of bioweapon creation

- Scientists have recently built the virus poliomyelitis using entirely artificial means.
- They did this using genomic data available on the Internet and materials from a mail-order chemical supply.
- The research was financed by the US Department of Defence as part of a biowarfare response program to prove to the world the reality of bioweapons.
- The researchers also hope their work will discourage officials from ever relaxing programs of immunisation.
- This project has been met with very mixed feelings.

Evolutionary studies

- The sequencing of genomes from all three domains of life, eukaryota, bacteria and archaea means that evolutionary studies can be performed in a quest to determine the tree of life and the last universal common ancestor.

Crop improvement

- Comparative genetics of the plant genomes has shown that the organisation of their genes has remained more conserved over evolutionary time than was previously believed.
- These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops.
- At present the complete genomes of Arabidopsis thaliana (water cress) and Oryza sativa (rice) are available.

Insect resistance

- Genes from *Bacillus thuringiensis* that can control a number of serious pests have been successfully transferred to cotton, maize and potatoes.
- This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced and hence the nutritional quality of the crops is increased.

Improve nutritional quality

- Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients.
- This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively.
- Scientists have inserted a gene from yeast into the tomato, and the result is a plant whose fruit stays longer on the vine and has an extended shelf life.

Development of Drought resistance varieties

- Progress has been made in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminium and iron toxicities.
- These varieties will allow agriculture to succeed in poorer soil areas, thus adding more land to the global production base.
- Research is also in progress to produce crop varieties capable of tolerating reduced water conditions.

Veterinary Science

- Sequencing projects of many farm animals including cows, pigs and sheep are now well under way in the hope that a better understanding of the biology of these organisms will have huge impacts for improving the production and health of livestock and ultimately have benefits for human nutrition.

Comparative Studies

- Analysing and comparing the genetic material of different species is an important method for studying the functions of genes, the mechanisms of inherited diseases and species evolution.
- Bioinformatics tools can be used to make comparisons between the numbers, locations and biochemical functions of genes in different organisms.

UNIT - IV

1. Biological database

A database is an organized collection of data.

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

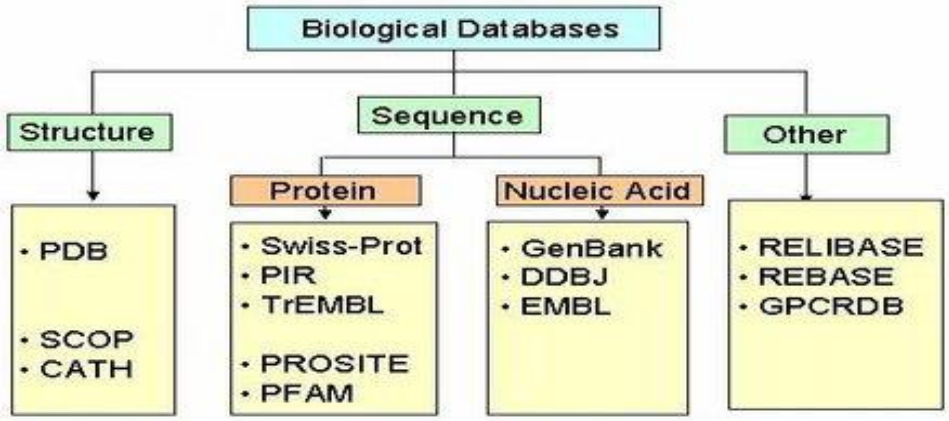
Biological databases can be broadly classified into

1. Sequence,
2. Structure and
3. Functional databases.

Nucleic acid and protein sequences are stored in **sequence databases** and **structure databases** store solved structures of RNA and proteins. **Functional databases** provide information on the physiological role of gene products.

Databases are important tools in assisting scientists to analyze and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications, predicting certain genetic diseases and in discovering basic relationships among species in the history of life.

Biological knowledge is distributed among many different general and specialized databases. This sometimes makes it difficult to ensure the consistency of information. Integrative bioinformatics is one field attempting to tackle this problem by providing unified access. One solution is how biological databases cross-reference to other databases with accession numbers to link their related knowledge together.



Relational database concepts of computer science and Information retrieval concepts of digital libraries are important for understanding biological databases. Biological database design, development, and long-term management is a core area of the discipline of bioinformatics. Data contents include gene sequences, textual descriptions, attributes and ontology classifications, citations and tabular data. These are often described as semi-structured data and can be represented as tables, key delimited records and XML structures.

Terms:

- *Entities*: The kind of things that we want to store in a database. E.g.: Genes, DNA sequences, bibliographical references.
- *Records*: The particular things stored in the database. E.g.: The gene BRCA1
- *Identifiers* or *key*: The unique name that identifies a record
- *Fields*: The properties that an entity has. E.g.: The name, sequence and mutations of the gene

The data repositories more relevant to the biological sciences include:

- nucleotide and protein sequences
- protein structures
- genomes
- genetic expression
- bibliography

Main sequence databases:

- NCBI
- EMBL

Main protein databases:

- Uniprot
- PDB
- MMDB

Some genome databases:

- ENSEMBL (Human, mouse and others)
- SGD (Yeast)
- TAIR (Arabidopsis)

Bibliography:

- Pubmed
- Web of Science

Human diseases:

- OMIM

Metabolic pathways:

- KEGG

Sequence databases

A sequence database is a collection of DNA or protein sequences with some extra relevant information. The main sequence databases are Genbank and EMBL.

The sequences submitted to any of those databases are shared between them, so any sequence could be retrieved in the european or the american database. But they differ in the tools to search and browse the data and in some databases that provide extra information to the raw sequences like: mutations, coded proteins, bibliographical references, etc.

The sequences are split in these databases in different sections to ease the search. Among others, there are sections for mRNAs, publised nucleotide sequences, genomes, and genes.

Genbank

Genbank is a public collection of annotated sequences hosted by the NCBI. Among other kinds of sequences Genbank includes messenger RNAs, genomic DNAs and ribosomic RNA.

Some characteristics:

- It is a public repository, anyone can send sequences to it.
- There are sequences of different qualities, anything submitted is stored.
- There could be multiple sequences for the same gene or for the same mRNA
- A sequence can have several versions that represent the modifications done by the authors.

Due to the huge amount of sequences stored to ease the search the databases are split in different divisions.

These divisions follow two criteria:

- The species and type of sequence. Among the taxonomical divisions you can find: primate, rodent, other mammalian, invertebrate an others.
- The other divisions are related to the kind of sequences like: EST, WGS, HTGS, and many others.

If you are looking for reads comming from the Next Generation Sequencing Technologies they are stored in a special division called SRA.

RefSeq

RefSeq is a reference database curated by NCBI.

In RefSeq there are only well annotated and good quality sequences. It stores genomic, transcript and protein sequences and links the sequences that belong to a gene.

UniProt

UniProt is a protein database that includes information divided in two sections: Swiss-Prot and TrEMBL. UniProt aims to store sequence and functional information for the proteins.

TrEMBL is automatically annotated while Swiss-Prot is reviewed manually by humans that add information by reviewing the literature. Due to this effort Swiss-Prot has information of a higher quality, but it has less sequences than TrEMBL.

PubMed

PubMed is a bibliographical database that comprises biomedical literature (MEDLINE), life science journals and on-line books. It is a good collection of publications related to biochemistry, cellular biology and medicine. As of 2016 PubMed stores 26 million citations.

For each record it stores:

- title
- authors
- abstract

There is a related database named PubMed Cental (PMC) that only includes citations of Free Access Journals. These citations include the complete text for the papers stored.

PDB, Protein Data Bank

PDB stores 3D structures for proteins and nucleic acids.

2. National Center for Biotechnology Information (NCBI)

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). The NCBI is founded in 1988 through legislation.

The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine.

GenBank

NCBI has had responsibility for making available the GenBank DNA sequence database since 1992. GenBank coordinates with individual laboratories and other sequence databases such as those of the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ).

Since 1992, NCBI has grown to provide other databases in addition to GenBank. NCBI provides Gene, Online Mendelian Inheritance in Man, the Molecular Modeling Database (3D protein structures), dbSNP (a database of single-nucleotide polymorphisms), the Reference Sequence Collection, a map of the human genome, and a taxonomy browser, and coordinates with the National Cancer Institute to provide the Cancer Genome Anatomy Project. The NCBI assigns a unique identifier (taxonomy ID number) to each species of organism.

The NCBI has software tools that are available by WWW browsing or by FTP. For example, BLAST is a sequence similarity searching program. BLAST can do sequence comparisons against the GenBank DNA database in less than 15 seconds.

NCBI Bookshelf

The "NCBI Bookshelf is a collection of freely accessible, downloadable, on-line versions of selected biomedical books.

The Bookshelf covers a wide range of topics including molecular biology, biochemistry, cell biology, genetics, microbiology, disease states from a molecular and cellular point of view, research methods, and virology.

Some of the books are online versions of previously published books, while others, such as Coffee Break, are written and edited by NCBI staff.

The Bookshelf is a complement to the Entrez PubMed repository of peer-reviewed publication abstracts in that Bookshelf contents provide established perspectives on evolving areas of study and a context in which many disparate individual pieces of reported research can be organized.

3. European Molecular Biology Laboratory (EMBL)

The **European Molecular Biology Laboratory** (**EMBL**) is a molecular biology research institution supported by 25 member states, four prospect and two associate member states.

EMBL was created in 1974 and is an intergovernmental organisation funded by public research money from its member states. Research at EMBL is conducted by

approximately 85 independent groups covering the spectrum of molecular biology. The list of independent groups at EMBL can be found at www.embl.org.

EMBL groups and laboratories perform basic research in molecular biology and molecular medicine as well as training for scientists, students and visitors. The organization aids in the development of services, new instruments and methods and technology in its member states. Israel is the only full member state located outside Europe.

EMBL goal was to create an international research centre, similar to CERN, to rival the strongly American-dominated field of molecular biology.

Each of the different EMBL sites have a specific research field. The EMBL-EBI is a hub for bioinformatics research and services, developing and maintaining a large number of scientific databases, which are free of charge.

The EMBL Rome site is dedicated towards the study of epigenetics and neurobiology. Scientists at EMBL Barcelona will explore how tissues and organs function and develop, in health and disease. At the headquarters in Heidelberg, there are units in Cell Biology and Biophysics, Developmental Biology, Genome Biology and Structural and Computational Biology as well as service groups complementing the aforementioned research fields.

Many scientific breakthroughs have been made at EMBL. The first systematic genetic analysis of embryonic development in the fruit fly was conducted at EMBL for which they were awarded the Nobel Prize in Physiology or Medicine in 1995. In the 1980s, developed cryogenic electron microscopy for biological structures at EMBL.

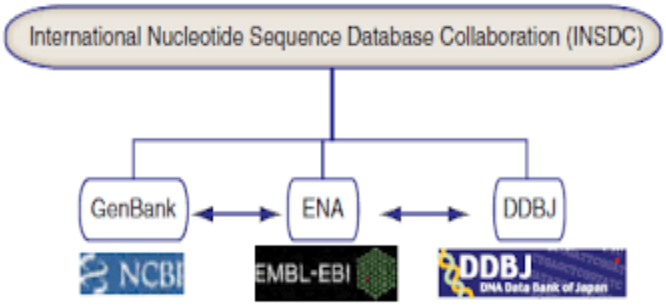
4. DNA Data Bank of Japan (DDBJ)

The DNA Data Bank of Japan (DDBJ) is a biological database that collects DNA sequences. It is located at the National Institute of Genetics (NIG) in Japan. It is also a member of the International Nucleotide Sequence Database Collaboration or INSDC. It exchanges its data with European Molecular Biology Laboratory at the European Bioinformatics Institute and with GenBank at the National Center for Biotechnology Information on a daily basis. Thus these three databanks contain the same data at any given time.

DDBJ began data bank activities in 1986 at NIG and remains the only nucleotide sequence data bank in Asia. Although DDBJ mainly receives its data from Japanese researchers, it can accept data from contributors from any other country. DDBJ is primarily funded by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). DDBJ has an international advisory committee which consists of nine members, 3 members each from Europe, US, and Japan. This committee advises DDBJ about its maintenance, management and future plans once a year. Apart from this DDBJ also has an international collaborative committee which advises on various technical issues related to international collaboration and consists of working-level participants.

Construction and Operation of INSDC

In Japan, DDBJ Center internationally contributes as a member of INSDC to collect and to provide nucleotide sequence data with ENA/EBI in Europe and NCBI in USA.



DDBJ Center is officially certified to collect nucleotide sequences from researchers and to issue the internationally recognized accession number to data submitters. The accession number issued for each sequence data is unique on the database and internationally recognized to guarantee the submitter the property of the submitted and published data. Since DDBJ Center exchanges the released data with ENA/EBI and NCBI on a daily basis, the three data centers share virtually the same data at any given time.

The virtually unified database is called INSD; International Nucleotide Sequence Database. DDBJ collects sequence data mainly from Japanese researchers, but of course accepts data and issue the accession numbers to researchers in any other countries. 99% of INSD data from Japanese researchers are submitted through DDBJ.

Management and operation of the NIG Supercomputer System

The National Institute of Genetics Supercomputer System (NIG Supercomputer) is a large-scale computer utilization site with genome analysis as its primary focus. The system provides Supercomputing System Services comprising leading-edge, large-scale cluster-type computers, large-scale memory-sharing computers, and high-capacity, high-speed disk devices.

We provide databases maintained by DDBJ and others through web services or on NIG Supercomputer.

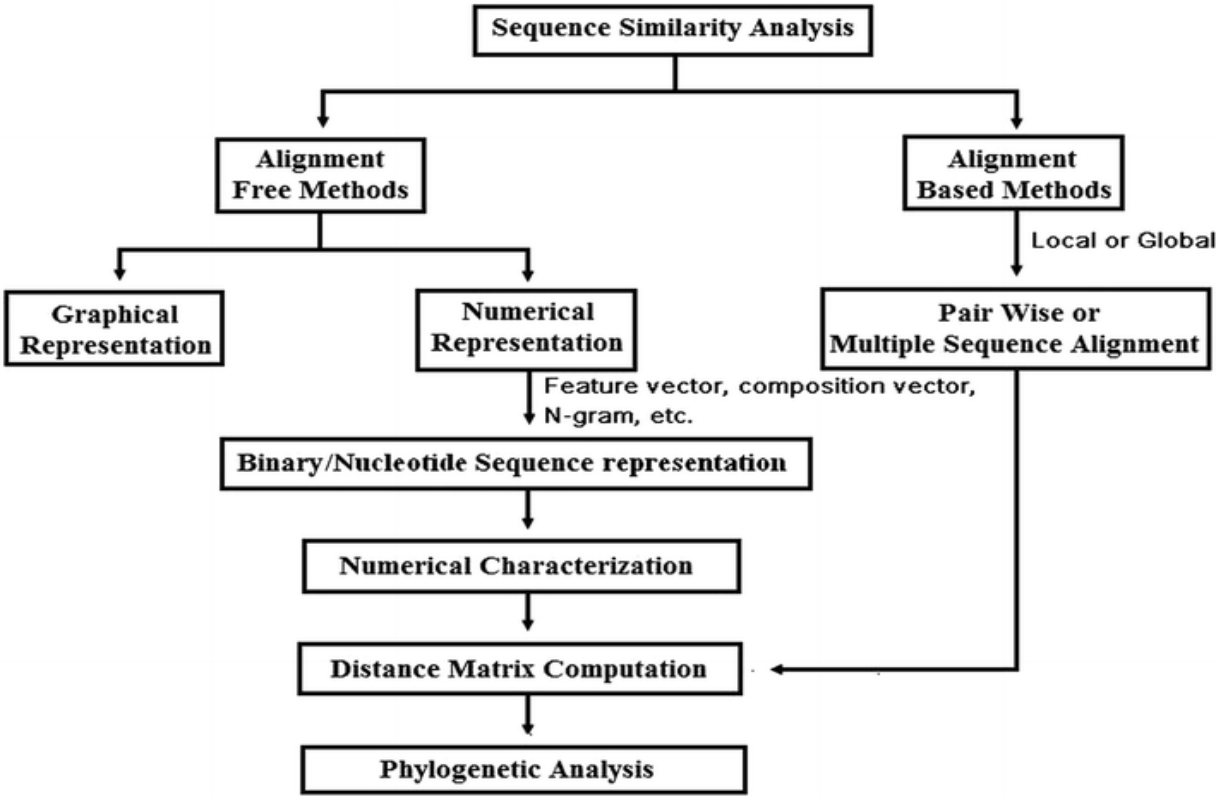
You can collectively download databases from our FTP site.

We provide software tools for data analyses developed by DDBJ and others through web services or on NIG Supercomputer.

5. Sequence alignment

In bioinformatics, a **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix.

Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. Sequence alignments are also used for non-biological sequences, such as calculating the distance cost between strings in a natural language or in financial data.



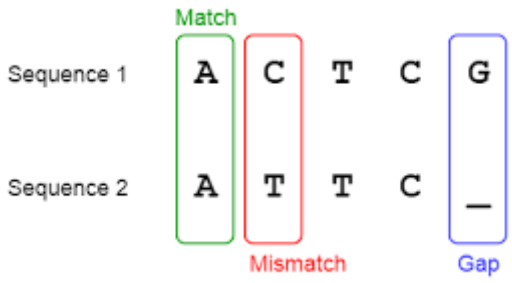
5.1 Pairwise Sequence Alignment

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

Pairwise sequence alignment methods are used to find the best-matching piecewise (local or global) alignments of two query sequences.

The three primary methods of producing pairwise alignments are

1. Dot-matrix methods,
2. Dynamic programming, and
3. Word methods



1. Dot-matrix methods

The dot-matrix approach implicitly produces a family of alignments for individual sequence regions. It is qualitative and conceptually simple, though time-consuming to analyze on a large scale.

In the absence of noise, it can be easy to visually identify certain sequence features—such as

- Insertions,

- Deletions,
- Repeats, or inverted repeats—from a dot-matrix plot.

To construct a dot-matrix plot, the two sequences are written along the top row and leftmost column of a two-dimensional matrix and a dot is placed at any point where the characters in the appropriate columns match—this is a typical recurrence plot.

Some implementations vary the size or intensity of the dot depending on the degree of similarity of the two characters, to accommodate conservative substitutions.

Problems with dot plots as an information display technique include:

- Noise,
- Lack of clarity,
- Non-intuitiveness,
- Difficulty extracting match summary statistics and
- Match positions on the two sequences.

Dot plots can also be used to assess repetitiveness in a single sequence. A sequence can be plotted against itself and regions that share significant similarities will appear as lines off the main diagonal. This effect can occur when a protein consists of multiple similar structural domains.

2. Dynamic programming

The technique of dynamic programming can be applied to protein alignments use a substitution matrix to assign scores to amino-acid matches or mismatches and a gap penalty for matching an amino acid in one sequence to a gap in the other.

DNA and RNA alignments may use a scoring matrix, but in practice often simply assign a positive match score, a negative mismatch score, and a negative gap penalty.

Dynamic programming can be useful in aligning nucleotide to protein sequences, a task complicated by the need to take into account frameshift mutations (usually insertions or deletions).

The BLAST and EMBOSS suites provide basic tools for creating translated alignments More general methods are available from open-source software such as Genewise.

3. Word methods

Word methods are especially useful in large-scale database searches of a large proportion of the candidate sequences will have essentially no significant match with the query sequence. Word methods are best known for their implementation in the database search tools FASTA and the BLAST family.

Word methods identify a series of short, nonoverlapping subsequences ("words") in the query sequence that are then matched to candidate database sequences. The relative positions of the word in the two sequences being compared are subtracted to obtain an offset; this will indicate a region of alignment if multiple distinct words produce the same offset. Only if this region is detected do these

methods apply more sensitive alignment criteria; thus, many unnecessary comparisons with sequences of no appreciable similarity are eliminated.

5.1.1 Basic Local Alignment Search Tools (BLAST)

BLAST is an algorithm used for calculating sequence similarity between biological sequences such as nucleotide sequences of DNA and amino acid sequences of proteins.

BLAST is a powerful tool for finding sequences similar to the query sequence within the same organism or in different organisms. It searches the query sequence on NCBI databases and servers and post the results back to the person's browser in chosen format.

Input sequences to the BLAST are mostly in FASTA or Genbank format while output could be delivered in variety of formats such as HTML, XML formatting and plain text.

HTML is the default output format for NCBI's web-page. Results for NCBI-BLAST are presented in graphical format with all the hits found, a table with sequence identifiers for the hits having scoring related data, along with the alignments for the sequence of interest and the hits received with analogous BLAST scores for these.

Entrez

The Entrez Global Query Cross-Database Search System is used at NCBI for all the major databases such as Nucleotide and Protein Sequences, Protein Structures, PubMed, Taxonomy, Complete Genomes, OMIM, and several others.

Entrez is both indexing and retrieval system having data from various sources for biomedical research. NCBI distributed the first version of Entrez in 1991, composed of nucleotide sequences from PDB and GenBank, protein sequences from SWISS-PROT, translated GenBank, PIR, PRF, PDB and associated abstracts and citations from PubMed.

Entrez is specially designed to integrate the data from several different sources, databases and formats into a uniform information model and retrieval system which can efficiently retrieve that relevant references, sequences and structures

PubChem database of NCBI is a public resource for molecules and their activities against biological assays. PubChem is searchable and accessible by Entrez information retrieval system.

5.1.2 FASTA

FASTA is a DNA and protein sequence alignment software package. Its legacy is the FASTA format which is now ubiquitous in bioinformatics.

The original FASTP program was designed for protein sequence similarity searching in the 1980s heuristic methods.

FASTA, published in 1987, added the ability to do DNA:DNA searches, translated protein:DNA searches, and also provided a more sophisticated shuffling program for evaluating statistical significance.

There are several programs in this package that allow the alignment of protein sequences and DNA sequences.

Uses

The current FASTA package contains programs for protein:protein, DNA:DNA, protein:translated DNA (with frameshifts), and ordered or unordered peptide searches.

Recent versions of the FASTA package include special translated search algorithms that correctly handle frameshift errors when comparing nucleotide to protein sequence data.

In addition to rapid heuristic search methods, the FASTA package provides SSEARCH, an implementation of the optimal Smith–Waterman algorithm.

A major focus of the package is the calculation of accurate similarity statistics, so that biologists can judge whether an alignment is likely to have occurred by chance, or whether it can be used to infer homology.

The FASTA package is available from the University of Virginia and the European Bioinformatics Institute.

The FASTA file format used as input for this software is now largely used by other sequence database search tools (such as BLAST) and sequence alignment programs (Clustal, T-Coffee, etc.).

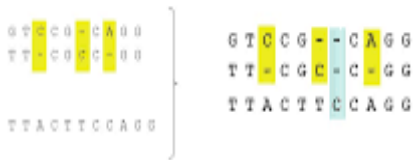
FASTA format

- In bioinformatics and biochemistry, the **FASTA format** is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes.
- The format also allows for sequence names and comments to precede the sequences.
- The format originates from the FASTA software package, but has now become a near universal standard in the field of bioinformatics.
- The simplicity of FASTA format makes it easy to manipulate and parse sequences using text-processing tools and scripting languages.

5.2 Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set.



Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related. Such conserved sequence motifs can be used in conjunction with structural and mechanistic information to locate the catalytic active sites of enzymes.

Alignments are also used to aid in establishing evolutionary relationships by constructing phylogenetic trees. Multiple sequence alignments are computationally difficult to produce and most formulations of the problem lead to NP-complete combinatorial optimization problems. Nevertheless, the utility of these alignments in bioinformatics has led to the development of a variety of methods suitable for aligning three or more sequences.

1. Dynamic programming

The technique of dynamic programming is theoretically applicable to any number of sequences; however, because it is computationally expensive in both time and memory, it is rarely used for more than three or four sequences in its most basic form.

This method requires constructing the *n*-dimensional equivalent of the sequence matrix formed from two sequences, where *n* is the number of sequences in the query.

Standard dynamic programming is first used on all pairs of query sequences and then the "alignment space" is filled in by considering possible matches or gaps at intermediate positions, eventually constructing an alignment essentially between each two-sequence alignment.

2. Progressive methods

Progressive, hierarchical, or tree methods generate a multiple sequence alignment by first aligning the most similar sequences and then adding successively less related sequences or groups to the alignment until the entire query set has been incorporated into the solution.

The initial tree describing the sequence relatedness is based on pairwise comparisons that may include heuristic pairwise alignment methods similar to FASTA. Progressive alignment results are dependent on the choice of "most related" sequences and thus can be sensitive to inaccuracies in the initial pairwise alignments.

Many variations of the Clustal progressive implementation are used for multiple sequence alignment, phylogenetic tree construction and as input for protein structure prediction. A slower but more accurate variant of the progressive method is known as T-Coffee.

3. Iterative methods

Iterative methods attempt to improve on the heavy dependence on the accuracy of the initial pairwise alignments, which is the weak point of the progressive methods. Iterative methods optimize an objective function based on a selected alignment scoring method by assigning an initial global alignment and then realigning sequence subsets.

The realigned subsets are then themselves aligned to produce the next iteration's multiple sequence alignment.

4. Motif finding

Motif finding (profile analysis) constructs global multiple sequence alignments that attempt to align short conserved sequence motifs among the sequences in the query set.

This is usually done by first constructing a general global multiple sequence alignment, after which the highly conserved regions are isolated and used to construct a set of profile matrices.

The profile matrix for each conserved region is arranged like a scoring matrix but its frequency counts for each amino acid or nucleotide at each position are derived from the conserved region's character distribution rather than from a more general empirical distribution.

5. Techniques inspired by computer science

A variety of general optimization algorithms commonly used in computer science has also been applied to the multiple sequence alignment problem. Hidden Markov models have been used to produce probability scores for a family of possible multiple sequence alignments for a given query set

Genetic algorithms and simulated annealing have also been used in optimizing multiple sequence alignment scores as judged by a scoring function like the sum-of-pairs method.

5.3 Phylogenetics alignment

Phylogenetics and sequence alignment are closely related fields due to the shared necessity of evaluating sequence relatedness.

The field of phylogenetics makes extensive use of sequence alignments in the construction and interpretation of phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species.

The degree to which sequences in a query set differ is qualitatively related to the sequences' evolutionary distance from one another.

Roughly speaking, high sequence identity suggests that the sequences in question have a comparatively young most recent common ancestor, while low identity suggests that the divergence is more ancient.

This approximation, which reflects the "molecular clock" hypothesis that a roughly constant rate of evolutionary change can be used to extrapolate the elapsed time since two genes first diverged (that is, the coalescence time), assumes that the effects of mutation and selection are constant across sequence lineages.

Therefore, it does not account for possible difference among organisms or species in the rates of DNA repair or the possible functional conservation of specific regions in a sequence.

More statistically accurate methods allow the evolutionary rate on each branch of the phylogenetic tree to vary, thus producing better estimates of coalescence times for genes.

Progressive multiple alignment techniques produce a phylogenetic tree by necessity because they incorporate sequences into the growing alignment in order of relatedness.

Other techniques that assemble multiple sequence alignments and phylogenetic trees score and sort trees first and calculate a multiple sequence alignment from the highest-scoring tree.

Commonly used methods of phylogenetic tree construction are mainly heuristic because the problem of selecting the optimal tree, like the problem of selecting the optimal multiple sequence alignment, is NP-hard.

PHYLIP

PHYLogeny Inference Package (PHYLIP) is a free computational phylogenetics package of programs for inferring evolutionary trees (phylogenies).

It consists of 35 portable programs, i.e., the source code is written in the programming language C. As of version 3.696, it is licensed as open-source software; versions 3.695 and older were proprietary software freeware. Releases occur as source code, and as precompiled executables for many operating systems including Windows (95, 98, ME, NT, 2000, XP, Vista), Mac OS 8, Mac OS 9, OS X, Linux (Debian, Red Hat); and FreeBSD from FreeBSD.org. Full documentation is written for all the programs in the package and is included therein.

The author of the package is Professor Joseph Felsenstein, of the Department of Genome Sciences and the Department of Biology, University of Washington, Seattle.
Methods (implemented by each program) that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees.

Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.
Each program is controlled through a menu, which asks users which options they want to set, and allows them to start the computation.

The data is read into the program from a text file, which the user can prepare using any word processor or text editor (but this text file cannot be in the special format of that word processor, it must instead be in *flat ASCII* or *text only* format).

Some sequence analysis programs such as the ClustalW alignment program can write data files in the PHYLIP format.

Most of the programs look for the data in a file called *infile*. If they do not find this file, they then ask the user to type in the file name of the data file.

Output is written onto files with names like `outfile` and `outtree`. Trees written onto `outtree` are in the Newick format, an informal standard agreed to in 1986 by authors of seven major phylogeny packages.

UNIT - V

1. Protein structure

Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule. Proteins are polymers – specifically polypeptides – formed from sequences of amino acids, the monomers of the polymer. A single amino acid monomer may also be called a *residue* indicating a repeating unit of a polymer.

By physical size, proteins are classified as nanoparticles, between 1–100 nm. Very large aggregates can be formed from protein subunits. For example, many thousands of actin molecules assemble into a microfilament.

A protein generally undergoes reversible structural changes in performing its biological function. The alternative structures of the same protein are referred to as different conformational isomers, or simply, conformations, and transitions between them are called conformational changes.

Configuration of Protein Structure

Primary structure

The primary structure of a protein refers to the sequence of amino acids in the polypeptide chain. The primary structure is held together by peptide bonds that are made during the process of protein biosynthesis. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity.

Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is not involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein.

A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The sequence of amino acids in insulin was discovered by Frederick Sanger, establishing that proteins have defining amino acid sequences.

Secondary structure

Secondary structure refers to highly regular local sub-structures on the actual polypeptide backbone chain. Two main types of secondary structure, the α-helix and the β-strand or β-sheets, were suggested in 1951 by Linus Pauling et al.

These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. They have a regular geometry, being constrained to specific values of the dihedral angles ψ and φ on the Ramachandran plot.

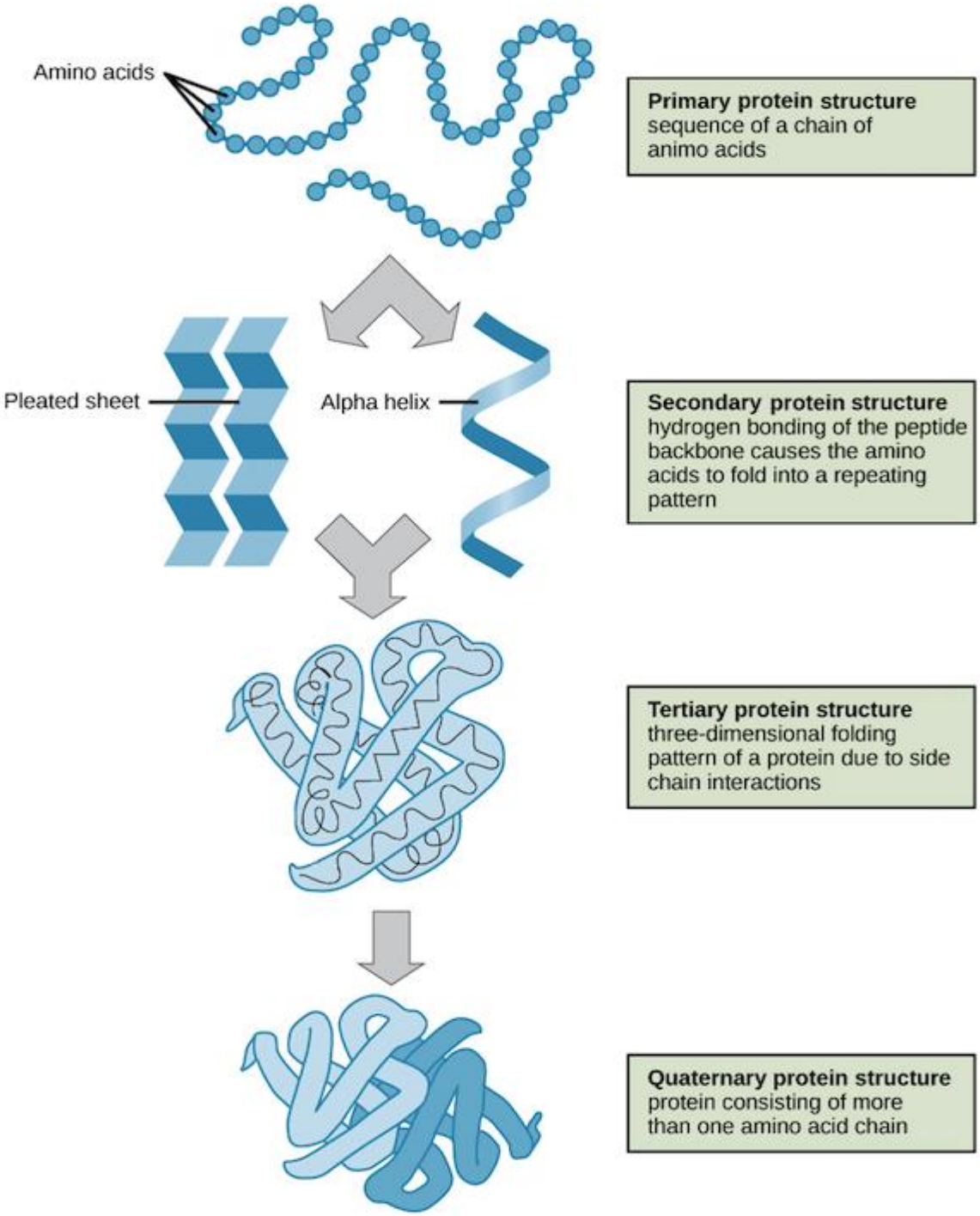
Both the α-helix and the β-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone. Some parts of the protein are ordered but do not form any regular structures.

Tertiary structure

Tertiary structure refers to the three-dimensional structure of monomeric and multimeric protein molecules. The α-helixes and β-pleated-sheets are folded into a compact globular structure.

The folding is driven by the *non-specific* hydrophobic interactions, the burial of hydrophobic residues from water, but the structure is stable only when the parts of a protein domain are locked into place by *specific* tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds.

The disulfide bonds are extremely rare in cytosolic proteins, since the cytosol (intracellular fluid) is generally a reducing environment.



Quaternary structure

Quaternary structure is the three-dimensional structure consisting of the aggregation of two or more individual polypeptide chains (subunits) that operate as a single functional unit (multimer).

The resulting multimer is stabilized by the same non-covalent interactions and disulfide bonds as in tertiary structure. There are many possible quaternary structure organisations. Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers.

Specifically it would be called a dimer if it contains two subunits, a trimer if it contains three subunits, a tetramer if it contains four subunits, and a pentamer if it contains five subunits.

The subunits are frequently related to one another by symmetry operations, such as a 2-fold axis in a dimer.

Multimers made up of identical subunits are referred to with a prefix of "homo-" and those made up of different subunits are referred to with a prefix of "hetero-", for example, a heterotetramer, such as the two alpha and two beta chains of hemoglobin.

2. Protein Data Bank

The **Protein Data Bank** (**PDB**) is a database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations. The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

The PDB is a key in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology.

Two forces converged to initiate the PDB:

1) A small but growing collection of sets of protein structure data determined by X-ray diffraction; and
2) The newly available (1968) molecular graphics display, to visualize these protein structures in 3-D.

In 1969, began to write software to store atomic coordinate files in a common format to make them available for geometric and graphical evaluation.

By 1971, SEARCH, enabled researchers to remotely access information from the database to study protein structures offline. SEARCH was instrumental in enabling networking, thus marking the functional beginning of the PDB.

The Protein Data Bank was announced in October 1971.

In October 1998, the PDB was transferred to the Research Collaboratory for Structural Bioinformatics (RCSB); the transfer was completed in June 1999.

In 2003, with the formation of the wwPDB, the PDB became an international organization. The founding members are PDBe (Europe), RCSB (USA) and PDBj (Japan).

Content

The PDB database is updated weekly. Likewise, the PDB holdings list is also updated weekly. As of 4 July 2019, the breakdown of current holdings is as follows:

- 126,949 structures in the PDB have a structure factor file.
- 10,022 structures have an NMR restraint file.
- 3,774 structures in the PDB have a chemical shifts file.
- 3,484 structures in the PDB have a 3DEM map file deposited in EM Data Bank

File format

The file format initially used by the PDB was called the PDB file format. This original format was restricted by the width of computer punch cards to 80 characters per line. Around 1996, the "macromolecular Crystallographic Information file" format, mmCIF, which is an extension of the CIF format started to be phased in. mmCIF is now the master format for the PDB archive. In 2019, the wwPDB announced that depositions for crystallographic methods would only be accepted in mmCIF format.

The structure files may be viewed using one of several free and open source computer programs. Currently UniParc contains protein sequences from the following publicly available databases;

3. Swiss-Prot

Swiss-Prot is a manually annotated, non-redundant protein sequence database. It combines information extracted from scientific literature and biocurator-evaluated computational analysis.

The aim of Swiss-Prot is to provide all known relevant information about a particular protein. Annotation is regularly reviewed to keep up with current scientific findings. The manual annotation of an entry involves detailed analysis of the protein sequence and of the scientific literature.

Sequences from the same gene and the same species are merged into the same database entry. Differences between sequences are identified, and their cause documented.

A range of sequence analysis tools is used in the annotation of Swiss-Prot entries. Computer-predictions are manually evaluated, and relevant results selected for inclusion in the entry.

These predictions include post-translational modifications, transmembrane domains and topology, signal peptides, domain identification, and protein family classification.

Relevant publications are identified by searching databases such as PubMed. The full text of each paper is read, and information is extracted and added to the entry. Annotation arising from the scientific literature includes, but is not limited to:

- ♣ Protein and gene names
- ♣ Function
- ♣ Enzyme-specific information such as catalytic activity, cofactors and catalytic residues
- ♣ Subcellular location
- ♣ Protein-protein interactions
- ♣ Pattern of expression
- ♣ Locations and roles of significant domains and sites
- ♣ Ion-, substrate- and cofactor-binding sites
- ♣ Protein variant forms produced by natural genetic variation, RNA editing, alternative splicing, proteolytic processing, and post-translational modification

Annotated entries undergo quality assurance before inclusion into Swiss-Prot. When new data becomes available, entries are updated.

4. Structural Classification of Proteins (SCOP) database

The **Structural Classification of Proteins (SCOP) database** is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences.

A motivation for this classification is to determine the evolutionary relationship between proteins.

Proteins with the same shapes but having little sequence or functional similarity are placed in different superfamilies and are assumed to have only a very distant common ancestor.

Proteins having the same shape and some similarity of sequence and/or function are placed in "families" and are assumed to have a closer common ancestor. The SCOP database is freely accessible on the internet. SCOP was created in 1994 in the Centre for Protein Engineering and the Laboratory of Molecular Biology.

Similar to CATH and Pfam databases, SCOP provides a classification of individual structural domains of proteins, rather than a classification of the entire proteins which may include a significant number of different domains.

The levels of SCOP are as follows;

1. Class: Types of folds, e.g., beta sheets.
2. Fold: The different shapes of domains within a class.
3. Superfamily: The domains in a fold are grouped into superfamilies, which have at least a distant common ancestor.
4. Family: The domains in a superfamily are grouped into families, which have a more recent common ancestor.
5. Protein domain: The domains in families are grouped into protein domains, which are essentially the same protein.
6. Species: The domains in "protein domains" are grouped according to species.
7. Domain: part of a protein. For simple proteins, it can be the entire protein.

5. CATH

CLASS **A**RCHITECTURE **T**OPOLOGY **H**OMOLOGOUS SUPERFAMILY

The CATH Protein Structure Classification database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains. It was created in the mid-1990s by Professor Christine Orengo and colleagues and continues to be developed by the Orengo group at University College London. CATH shares many broad features with the SCOP resource, however there are also many areas in which the detailed classification differs greatly.

The four main levels of the CATH hierarchy;

S.No	LEVEL	DESCRIPTION
1	Class	The overall secondary-structure content of the domain. (Equivalent to the SCOP Class)
2	Architecture	High structural similarity but no evidence of homology. (Equivalent to the 'fold' level in SCOP)
3	Topology/fold	A large-scale grouping of topologies which share particular structural features
4	Homologous superfamily	Indicative of a demonstrable evolutionary relationship. (Equivalent to SCOP superfamily)

CATH is an open source software project, with developers developing and maintaining a number of open source tools.

CATH maintains a todo list on GitHub to allow external users to create and keep track of issues relating to the CATH protein structure classification.

Additional sequence data for domains with no experimentally determined structures are provided by CATH's sister resource, Gene3D, which are used to populate the homologous superfamilies.

Protein sequences from UniProtKB and Ensembl are scanned against CATH HMMs to predict domain sequence boundaries and make homologous superfamily assignments.

The CATH team aim to provide official releases of the CATH classification every 12 months. This release process is important because is allows for the provision of internal validation, extra annotations and analysis.

In order to address this issue: CATH-B provides a limited amount of information to the very latest domain annotations

6. Protein visualization tools

Protein structure visualization tools

Introduction:

Before computer visualization software was developed, molecular structures were presented by physical models of metal wires, rods and spheres. With the development of computer hardware and software technology and computer graphics programs were developed to visualizing and manipulating three-dimensional structures. The computer graphics help to analyze and compare protein structure to gain the functions of protein.

Molecular visualization helps the scientists to bioengineer the protein molecules. User-friendly graphic interface makes this area of Bioinformatics a full filled, scientific thrill to the bioscientists.

Tools for molecular visualization:

There are a number of software's both free and commercial are available to visualize the biomolecules. The most commonly used free software are :

- **RasMol**
- **Chime**
- **MolMol**
- **Protein explorer**
- **Kinemage Cn3D**

RasMol

RasMol is a molecular graphics program intended for the visualisation of proteins, nucleic acids and small molecules. It is created by **Roger Sayle in 1992**. The program is aimed at display, teaching and generation of publication quality images. RasMol runs on wide range of architectures and operating systems including Microsoft Windows, Apple Macintosh, UNIX and VMS systems.

RasMol Features:

The program consists of two windows:

- ✓ For the command line
- ✓ For providing the graphics.

1. **Input file format:**

The input file can be in **PDB format** and can be downloaded from the PDB structure database.

- Protein Data Bank (PDB) files can be downloaded for visualization from members of the Worldwide Protein Data Bank (wwPDB). These have been uploaded by researchers who have characterized the structure of molecules usually by X-ray crystallography, protein NMR spectroscopy, or cryo-electron microscopy.

2. Display:

There are different ways of displaying coordinates. These include:- wireframe, sticks, spacefill, strands & cartoons.

- The initial image is shown as a **„wire"** model.

  ✓ -from the Display menu one can choose other visualisation styles such as „spacefill", „stick", „ball and stick" as well as the visually most attractive „ribbon" and „cartoon" models.

- In the last two styles, alpha helices are rendered as helical ribbons and beta structures as **flat arrows pointing** in the direction of the polypeptide chain.

3. Colour:

- The atoms of the model can be coloured by the standard CPK (named after Corey, Pauling and Koltun)

  To color by atom type : **Colours/CPK**

  - Carbon: gray
  - Hydrogen: white
  - Oxygen: red
  - Nitrogen: blue
  - Sulfur: yellow
  - Iron: yellow

- The protein can be coloured based on polypeptide chains, the chemical property of the amino acids.
- To color by the protein-secondary structure: **Colours/Structure**
  - α -helices: magenta
  - β -sheets: yellow
  - turns: pale blue
  - all other residues: white
- The structure can be cut in the **z-dimension**
- The left and right mouse buttons can be used to rotate the protein along the „x" and „y" axes.

4. For moving the molecule(s):

Action	PC
Rotation	Left-mouse button (Click & Hold)
Translation	Right-mouse button (Click & Hold)
Zoom	<alt><SHIFT> and left-mouse button
Z-Rotation	<alt><SHIFT>and right-mouse button

5. Atom Selection and view:

- By clicking any **part of the structure**, the residue number of the given chain and the particular atom will be shown in the command window.

- Carbon: gray
- Hydrogen: white
- Oxygen: red
- Nitrogen: blue
- Sulfur: yellow
- Iron: yellow

- The protein can be coloured based on polypeptide chains, the chemical property of the amino acids.

- To color by the protein-secondary structure: **Colours/Structure**
  - α -helices: magenta
  - β -sheets: yellow
  - turns: pale blue
  - all other residues: white
- The structure can be cut in the **z-dimension**
- The left and right mouse buttons can be used to rotate the protein along the „x" and „y" axes.

4. For moving the molecule(s):

Action	PC
Rotation	Left-mouse button (Click & Hold)
Translation	Right-mouse button (Click & Hold)
Zoom	<alt><SHIFT> and left-mouse button
Z-Rotation	<alt><SHIFT>and right-mouse button

5. Atom Selection and view:

- By clicking any **part of the structure**, the residue number of the given chain and the particular atom will be shown in the command window.

RasMol

RasMol is a computer program written for molecular graphics visualization intended and used mainly to depict and explore biological macromolecule structures, such as those found in the Protein Data Bank in the early 1990s.

Historically, it was an important tool for molecular biologists since the extremely optimized program allowed the software to run on (then) modestly powerful personal computers. Before RasMol, visualization software ran on graphics workstations that, due to their cost, were less accessible to scholars.

RasMol continues to be important for research in structural biology, and has become important in education.

RasMol has a complex licensing version history. Starting with the version 2.7 series, RasMol source code is dual-licensed under a GNU-General Public

License (GPL). Starting with version 2.7.5, a GPL is the only license valid for binary distributions.

RasMol includes a scripting language, to perform many functions such as selecting certain protein chains, changing colors, etc. Jmol and Sirius software have incorporated this language into their commands.

Protein Data Bank (PDB) files can be downloaded for visualization from members of the Worldwide Protein Data Bank (wwPDB). These have been uploaded by researchers who have characterized the structure of molecules usually by X-ray crystallography, protein NMR spectroscopy, or cryo-electron microscopy.

Rasmol can communicate with other programs via Tcl/Tk on Unix platforms, and via Dynamic Data Exchange (DDE) on Microsoft Windows. With a multiple sequence alignment program, the responsible Java class can be freely used in other applications.

Swiss-PDB Viewer

Introduction:

Swiss-PdbViewer is an application that provides a user friendly interface allowing to analyze several proteins at the same time. Swiss-PdbViewer has been developed on 1994 by Nicolas Guex. Swiss-PdbViewer is linked to SWISS-MODEL, an automated homology modeling server developed within the Swiss Institute of Bioinformatics (SIB) at the Structural Bioinformatics Group at the Biozentrum in Basel.

SWISS-MODEL consists of three tightly integrated components:

1. **The SWISS-MODEL pipeline** – a suite of software tools and databases for automated protein structure modeling

SWISS-MODEL pipeline comprises the four main steps that are involved in building a homology model of a given protein structure:

- Identification of structural template(s). BLAST and HHblits are used to identify templates. The templates are stored in the SWISS-MODEL Template Library (SMTL), which is derived from PDB.
- Alignment of target sequence and template structure(s).
- Model building and energy minimization. SWISS-MODEL implements a rigid fragment assembly approach for modelling.
- Assessment of the model's quality using QMEAN, a statistical potential of mean force.

2. **The SWISS-MODEL Workspace**(A web-based graphical user workbench)

In this mode the input is a project file that can be generated by the DeepView (Swiss Pdb Viewer) visualization and structural analysis tool, to allow the user to examine and manipulate the target-template alignment in its structural context.

3. **The SWISS-MODEL Repository**

The SWISS-MODEL Repository provides access to an up-to-date collection of annotated three-dimensional protein models for a set of model organisms of high general interest. SWISS-MODEL Repository is integrated with several external resources, such as UniProt, InterPro, STRING, and Nature PSI SBKB.

Uses:

- To find hydrogen bonds within proteins and between proteins and ligands.
- To view several protein structures simultaneously and superimpose them to align their structures and sequences.
- To examine electron-density maps from crystallographic structure determination
- To judge the quality of maps and models, and to identify many common problems in protein models.
- It computes electrostatic potentials and molecular surfaces, and carries out energy minimization.