

Application of Internet in Biological Research

16-18 January 2002

INTERNET & EMAIL

Madhu Kumar S.D.

*Department of Computer Engineering,
Regional Engineering College, Calicut*

This lecture notes covers some basic topics in Internet. It includes discussions about Domain Name System Addressing, Name Sever and the email utility. The material is taken from the Internet sites mentioned in the references.

What is Internet?

Internet is a network of networks. Internet is a rapidly growing global digital library built on a remarkably flexible communication technology. Internet is a platform where millions of people communicate electronically on a one-to one basis or in groups and share ideas of information. Internet is a vast collection of information that can be searched and retrieved electronically.

There are four basic building blocks to the Internet, Hosts, Routers and Clients and Connections. In most cases your computer falls under the "Client" category. Data is sent from your computer in the form of a "packet". It is similar to an envelope; it surrounds your data and contains both a return and destination address. Your computer handles the packets for you; it's all done in the background, without your knowledge.

Internet Protocol (IP)

The Internet Protocol, IP, specifies the basic rules that a computer must follow to communicate across the Internet. IP defines the format of Internet packets, which are called IP datagrams. IP also defines an address scheme that assigns each computer a unique number used in all communications. IP software makes an interconnected set of networks and routers operate like a single, large network.

Each computer on the Internet must have IP software that allows it to create and send IP datagrams. Each router also has IP software that knows how to forward datagrams to their destination. When a datagram arrives at a router, the IP software chooses the path that will lead to the datagram's destination.

Transmission Control Protocol (TCP)

To handle communication errors, a computer must also have TCP software. TCP eliminates duplicate data, ensures that data is reassembled in exactly the order it was sent and resends data when a datagram is lost.

TCP uses acknowledgements and timeouts to handle the problem of data loss. The sender retransmits data unless an acknowledgement arrives before the timer expires. TCP's scheme of timeout works across the Internet because TCP changes the timeout automatically depending on whether the destination is close or far away from the source.

Domain Name System

What is a DNS ?

The Domain Name System is a system which helps internet users to use the net more easily by allowing them to specify a meaningful names to web sites and / or other users they want to communicate with.

Why do we need a DNS ?

When computers talk to each other via the internet, they use the IP protocol. The Internet Protocol distinguish hosts from each other by an IP address which is a string of numbers appended to each other and separated by periods. An example of such a string might be 197.15.3.2 .

The IP addresses are unique, which means that each host has its own IP address which is different from all other IP addresses in the world. However, these IP addresses are hard for us, humans, to remember and we prefer to use meaningful names, which we are used to refer to in the everyday life.

The role of the DNS is derived from that difference between humans and computers:

Humans :	Computers:
Prefer Names	Prefer Numbers
ibm.com	197.15.3.6

The DNS is needed by applications to convert humans meaningful names into computer meaningful names (IP addresses) and provide the final user an easier way to communicate via the Internet.

Why do computers prefer addresses based on numbers?

Computers prefer number addresses because they perform better with numbers, which require less computation.

For example, let's say that an address named 'com' is represented by the number 231. In binary representation, the computer needs at least 3 Bytes to represent the name 'com', because each character is represented with one byte at least. However, representing the number 231 requires only 8 binary bits (a single byte). As a result comparing the name 'com' with other names will require comparing at least 3 bytes while comparing the number 231 will require only a single byte comparison.

What is a Domain Name, and how does a computer name look like?

Each computer name consists of a sequence of alpha-numeric segments separated by periods. For example, a computer name might be : `www.lingo.com`

A computer name is also called a 'Domain Name'.

Domain names are hierarchical, with the most significant part of the name on the right. The left most segment of a name (`www` in the example) is the name of an individual computer. Other segments of the full name identify the group that owns the individual name. In the example, the individual name `www` is owned by the `lingo` group of names, which is itself owned by the `com` group (which holds a lot of company names).

TCP/IP Internet Domain Names

The mechanism that implements a machine name hierarchy for TCP/IP Internet is called the Domain Name System (DNS).

DNS has two independent aspects:

It specifies the name syntax and rules for delegating authority over names. It specifies the implementation of a distributed computing system that efficiently maps names to addresses. The domain name system uses a hierarchical naming scheme known as domain names. A domain name consists of a sequence of sub-names separated by a delimiter character, the period. The domain calls each section a label.

The domain name `cs.purdue.edu` contains three labels: `cs`, `purdue`, and `edu`. Any suffix of a label in a domain name is also called a domain. In the above example the lowest level domain is `cs.purdue.edu`, the second level domain is `purdue.edu` and the top level domain is `edu`.

Writing domain names with the local label first and the top domain last makes it possible to compress messages that contain multiple domain names.

Official And Unofficial Internet Domain Names

The domain name standard specifies an abstract hierarchical name-space with arbitrary values for labels. Therefore it is possible for any group that builds an instance of the domain system to choose labels for all parts of its hierarchy.

For example, a private company can establish a domain hierarchy in which the top level labels specify corporate subsidiaries, the next level labels specify corporate divisions, and the lowest level specify departments.

However most users of the domain technology follow the hierarchical labels used by the official Internet domain system. Most sites follow the Internet scheme so they can attach their TCP/IP installations to the global Internet without changing names.

The Internet authority has chosen to partition its top level into the following domains:

<u>Domain Name</u>	<u>Meaning</u>
COM	Commercial organizations
EDU	Educational institutions
GOV	Government institutions
MIL	Military groups
NET	Major network support centres
ORG	Organizations other than those above
ARPA	Temporary ARPANET domain
INT	International organizations
Country code	Each country (geographical scheme)

The top level names permit two completely different naming hierarchies: geographic and organizational. The geographic scheme divides the universe of machines by country. Each country is assigned its own top level domain with the country's international 2 letter identifier as its label.

Mapping Domain Names To Addresses

The domain name scheme includes an efficient, reliable, general-purpose, distributed system for mapping names to addresses. The system is distributed in the technical sense, meaning that a set of servers operating at multiple sites co-operatively solve the mapping problem. It is efficient in the sense that most names can be mapped locally; only a few require Internet traffic. It is general purpose because it is not restricted to machine names.

The domain mechanism for mapping names to addresses consists of independent, co-operative systems called name servers. A name server is a server program that supplies name-to-address translation, mapping from domain names to IP addresses. The client software, called a name resolver, uses one or more name servers when translating a name.

The easiest way to understand how domain servers work is to imagine them arranged in a tree structure that corresponds to the naming hierarchy.

The root of the tree is a server that recognizes the top level domains and knows which server resolves each domain. Given name to resolve, the root can choose the correct server for that name. At the next level, a set of name servers each provide answers for one top level domain (e.g., edu). A server at this level knows which servers can resolve each of the sub domains under its domain.

Domain Name Resolution

Domain name resolution proceeds top-down, starting with the root name server and proceeding to servers located at the leaves of the tree. There are two ways to use the domain server system:

- Contacting name servers one at a time

- Asking the name server system to perform the complete translation

In either case, the client software forms a domain name query that contains the name to be resolved, a declaration of the class name, the type of answer desired, and a code that specifies whether the name server should translate the name completely. It sends the query to a name server for resolution.

When a domain name server receives a query, it check to see if the name lies in the sub domain for which it is an authority. If so, it translates the name to an address according to its database, and appends an answer to the query before sending it back to the client.

If the name server cannot resolve the name completely, it checks to see what type of interaction the client specified. If the client requested complete translation, the server contacts a domain name server that can resolve the name and returns the answer to the client.

If the client requested non-recursive resolution, the name server cannot supply an answer. It generates a reply that specifies the name server the client should contact next to resolve the name.

A client must know how to contact at least one name server. To ensure that a domain name server can reach others, the domain system requires that each server know the address of at least one root server. In addition, a server may know the address of a server for the domain immediately above it.

Domain name servers use a well-known protocol port for all communication, so clients know how to communicate with a server once they know the IP address of the machine in which the server executes. There is no standard way for hosts to locate a machine in the local environment on which a name server runs; that is left to whoever designs the client software.

In some systems, the address of the machine that supplies domain name service is bound into application programs at compile time, while in others, the address is configured into the operating system at start-up. In others, the administrator places the address of a server in a file on secondary storage.

Internet Access

In the late 1980s and early 1990s, about the only option for Internet access was some kind of dedicated or leased line connection. However, individual dial-up accounts became available from a number of commercial Internet access providers as early as 1992. Usually access to Internet is available in two forms: Dial-up access and dedicated access.

Dialup access

In this case a user modem dials a host computer to access the Internet. This access is of two types-terminal dial-up and TCP/IP based. Terminal dial-up access doesn't permit graphics and thus restricts the range of available services. However it allows E-mail, FTP, and other non-graphic applications. Dial-up connections based on serial line Internet Protocol (SLIP) and point-to-point protocol (PPP) can connect one's computer directly to Internet and one can run networking applications.

Dedicated access

Dial up users must wait for their modems to dial and connect while users with dedicated access have Internet connectivity all the time. In case of dedicated access in LAN, at least one computer on the LAN is configured with TCP/IP software the programme that instructs the computer how to communicate with the Internet, however dedicated access is not right for every one, besides being expensive it take time to set it up and requires specialized manpower for operations and maintenance.

World Wide Web (WWW)

WWW has become the most widely used information service on the Internet. WWW is a large set of hypermedia documents stored on computers throughout the Internet. A given WWW

Document can contain embedded menu items that refer to WWW documents stored on other computers, a user can follow a link from a document on one computer to a document on another computer as easily as a user can follow a link between documents on the same computer.

A browser is a computer program used to access the WWW. To make it possible to record the location of a particular piece of information, each WWW document is assigned a unique name known as a Uniform Resource Locator (URL) that contain the information a browser needs to obtain the document.

Web pages are written in a computer language known as Hyper Text Markup Language (HTML). There are advanced web technologies like forms, frames, CGI programs and Java.

Website

Website is an addressed location, which consists of a web server and a set of web pages. It is a combination of textual, graphics, movie, clips and sound files. The resources in a website are usually located on one server, but they don't need to be in all cases. A Web page has links, so that user can move easily between pages in a website, or jump to another related site.

URL

Internet addresses are URLs (Uniform Resource Locations) and understanding them can help the user navigate the Internet better. Most URLs consist of four basic parts: 1. The protocol, 2. The server or domain name, 3. The directory path, and 4. The file name. Each part of a URL helps direct browser through the internet to the file, the user wants to retrieve.

URLs begin by designating the protocol to be used. The WWW protocol is hypertext transport protocol (HTTP), which enables hypertext transmissions across the network. Most Web clients can also connect to resources using other protocols such as gopher, FTP and Telnet.

BROWSERS

Most internet services operate according to a scheme called client/server. A server on one computer starts a programme that contacts another remote computer system, the client is the programme the user is running, and server is running on a remote system. The user gives command to the client, which passes them on to the server. The server interprets those commands and returns information to client, which passes information to the user.

Browser allows the user to view Web sites. There are many types of browsers available today. Depending on the type of content a browser can handle, they are divided into two types.

- Graphic-based
- Text-based

Among graphical browsers, Netscape Communicator and Internet Explorer are the most popular. In the US some people use the America Online (AOL) browser while several others use Mosaic, HotJava and other browsers. Text-based browsers do not allow much to be seen on the web page other than the plain text. Hence they are not so popular, though some people use these browsers, as they are the cheapest. Among other text-based browsers, Lynx is by far the most popular.

Lynx

Lynx is one of the Web browsers, which can only display the textual data. The graphical or multimedia elements cannot be delivered through this shareware web browser. Lynx was developed at the University of Kansas for their own campus, then was released to the Internet community. Lynx uses full-screen capabilities, and it can easily move back and forth through hypertext links.

Mosaic

Mosaic is the first popularly used graphical Web browser designed and developed at the National Centre for Super Computing Application (NCSA). Mosaic provides a powerful user-friendly graphical interface where hyper textual links appear to connect other web pages, or highlighted words in the text. Colour is used to highlight text links and other elements.

Netscape Navigator

Netscape navigator is the most widely used graphical web browser and popularly known as Netscape. The salient features of Netscape are:

1. It combines text, pictures, graphics, and animation in lots of attractive ways.
2. It runs reasonably well even on inexpensive dialup connections.
3. Consecutive version are being updated to add new internet services and
4. its authors give it away for free.

Netscape is the best organized, best performing and most efficient web tool. Netscape recognizes the advanced HTML tags. Netscape Navigator supports a host at widely used enhancements to HTML. A clear interface, easy-to-use controls and a straightforward set-up also add to this browser's appeal. The latest version is named as Netscape Communicator, which has the facilities of both editing the HTML document and browsing it.

Internet Explorer

The Microsoft goes long way towards its goal to integrate all Internet tools with Windows 95 under the folder name Internet Explorer for browsing the web. It is distributed with Windows 95. This web browser provides all the features needed to explore the Internet thoroughly.

Offline Operations and File Downloading

The Internet represents a vast storehouse of information and software. Some people are simply overwhelmed by the amount of data(information) contained on the Internet. Some ISPs do not provide unlimited time accounts, so people may tend to be 'stressed' when surfing because they know the clock is ticking. This chapter will explore ways to lessen the impact of on-line surfing by showing how it's possible to save something for viewing at a later point in time. This chapter will also discuss downloading, what is meant by downloading, and some of the guidelines in regard to downloading.

Saving information

You've found a site that contains information, which you find extremely useful, but do you have to return to the site everytime you need to refer the information? No! A little known feature of most browsers is the ability to load and view web pages off-line. While you can view a web page off-line, unless you specifically reconstruct the directory structure, saving all of the associated graphics, you won't see the images for the page. With that in mind, here are some the steps you would need to follow in order to save this page and view it off-line in either Navigator or Explorer.

Click on File Menu.

Select the option "Save as..." (Explorer users! Use "Save as File..." option.

Save the file, but remember the file name!

This will allow you to save the file to your local hard drive in a place other than the cache directory. The cache directory is used by the browser to speed loading of the web sites if you make a return visit, but you are not assured that the files will be safe there as the browsers periodically clean out the cache directory to prevent it from growing too large. Normally the files are named using a four-character extension (.html). In the case of both browsers, the program will know, and use the file name under which it was stored on the remote system.

Viewing a Document Off-line

Once you are disconnected from the Internet, activate your browser.

Netscape

Click on the "Open File..." menu option under the "File" menu and move to the appropriate directory where you saved the html file.

Select the file that you wish to view from the list supplied and it will appear in the main display window.

Explorer

Click on the "Open..." menu option under the "File" menu.

Either type the filename, including the full path of where you saved it, or use the "Browse" feature in this dialog box to move around your harddrive until you locate the file.

Select the file that you wish to view from the list supplied and it will appear in the main display window.

Viewing a file, which you have saved to your harddrive, as opposed to viewing it on-line poses a few unique problems. While you may have saved all the pages in a website, the odds are good that you did not save any pages from the offsite links. An offsite link is a link which takes you to another host computer, usually containing documents which the author of the saved document feels are worthy, but did not write himself/herself. Since you didn't save these offsite files any links to them will simply not work. Clicking on them in an off-line mode of operation will not harm your computer, nor damage any of the saved files in your computer.

Saving a Graphic

You've found a beautiful picture or just a photo and want to save that image on your computer, maybe even use it for a Windows background. It's not hard to save a graphic file (a picture) that loaded with a page when you accessed that page. It does help to understand a little about the process of loading a web page into your computer first.

Whenever you visit a webpage, your browser begins a special communications session with the server upon which the page resides (or is "Hosted"). The host computer (also called a server) gives your browser the requested file (this is usually an HTML file), which is usually all text and the browser loads it into your computer's memory. The HTML file is a text file containing text and "mark-up" commands which instruct the browser on how to display the file. Basically the browser reads the HTML file after loading it. Locates all the commands to load graphics, then retrieves the graphics from the server, again, storing it locally on your computer. Finally it assembles the web page by combining the graphics and the text contained in the HTML file. In essence, all these graphics and text you are seeing are loaded from your own computer, not from the remote site. This process is called "Caching", which is a technical term for local, temporary disk storage.

The real trick, and it's not really a trick, is moving the graphic image from one directory to another directory. In essence, moving the image from your cache directory to another, more permanent location. Navigator and Explorer users have it easy, the designers of these programs foresaw the desire to easily save graphics.

Navigator

Position the cursor over the image and click on the right mouse button.

Select "Save image as..."

Enter an appropriate file name for the image.

Click on the "OK" Button.

Explorer

Position the cursor over the image and click on the right mouse button.

Select "Save picture as..."

Enter an appropriate file name for the image.

Click on the "OK" Button.

Downloading Programs

Downloading is the term used to describe the process of moving software from someone else's computer into your own local computer. Generally there are four major types of software you can download from the Internet. These are Demoware, Freeware, Shareware and patches/upgrades.

Demoware - This is a crippled version of a retail copy of software. By crippled we mean that in some way the functionality of the program has been reduced enough to make the program marginally useful. The primary purpose of Demoware is to let you "test drive" software before making a decision to purchase it.

Freeware - Freeware is free, completely, no strings attached. Usually these types of programs have been written by government agencies or companies that want you to use their programs and perhaps be induced into buying something from their mainstream product line.

Shareware - Typically shareware products are a mix of freeware and demoware. In many cases today you can get a fully functional program which is shareware, but has a built-in self destruct mechanism causing it to become unusable after so many days of use. Many shareware authors rely on the honor system for people to pay them.

Patches/Upgrades - Patches and Upgrades are not usually complete programs. Basically a patch or upgrade requires that you already have a particular piece of software on your computer. The patch will alter the software, either correcting a mistake in the system, or adding additional functionality to the software.

Downloading software is going to consume a significant portion of your on-line time. Whether it's getting new software, or upgrades, or just obtaining demo copies of the latest programs, everyone needs to know how to download. It's a sound idea, before you begin, to create a directory or folder and call it "Download". This way you will never have to scramble to figure out where to put a file you want to get from the net. Additionally, if you always save the downloaded files in the same directory, then removing them will be easier since they are all in one location on your harddrive.

Download Identification

Downloading files is only one step in a series of steps usually required to make a file useable on your system.

Once you have downloaded a file (the easy part), the next problem is figuring out what to do with it. In many cases, the download will be an executable file, so simply running the file will install the software on your machine. In some cases you won't be so lucky.

The key thing to remember about the file is the extension. The extension is the part following the period ("."). For example, some files may have a ".zip" extension. This identifies the file as a particular type of archive file. File extensions viz. .zip, .z, .gz, .pak, .arc, .arj, .uae are generally used to identify archives. What is an archive? An archive is a single file in which one or more files have been stored. Usually compression is performed, during the process of storing the files. In order to deal with archives you need to obtain a special program which is capable of decompressing the archive. You do not need to own one program for each type of archive. There are several highly capable programs available today which handle multiple archive formats with considerable ease.

The biggest problem with downloading software is that often many people do not realize that once the file is loaded onto the computer, it almost always requires additional steps before you can use the software. The process of downloading and using a downloaded program follows these basic steps, regardless of the program doing the download.

EMAIL

Electronic mail is the fastest and cheapest way of communication today. Any type of files can be send to anybody through email. Computer networks are used to transmit electronic mails. Now people use the internet to send and receive emails. Following is a very brief discussion about the working of emails.

How Does E-Mail Reach its Destination?

Let us assume that you are at home, writing an electronic mail message to your uncle, who works on the other side of the world. You know that once you send your mail, it will somehow make itself to you uncle's computer. How exactly does this happen?

Step 1: Sending (uploading) mail

To connect to the Internet, you must have an account with an Internet Service Provider (ISP), a company that connects people to the Internet. Connection from your computer to the ISP's computers will probably be made through the phone, using a device called a modem. The modem enables your computer to transmit digital data over the phone lines, which are analog. The data passed between the two computers is decoded using computer protocols. In this case, various modem specific protocols (such as X.25) are used, as well as protocols from the Internet protocol suite (such as PPP and SLIP). Your mail program sends your mail through the use of yet another protocol - the Simple Mail Transfer Protocol (SMTP).

Step 2: At the ISP's

The ISP's modem converts data from the phone line back to digital pulses, and it server interprets the information. The server figures out the nature of your data, by decoding the SLIP/PPP data (in this case, the data is your e-mail). The ISP has a local area network (LAN) (or specifically, an Ethernet LAN). The LAN connects computers in the ISP's office space. Once the data is interpreted, the server's first role is to figure out if the data you had just sent is destined for a computer in the ISP's LAN. Data not destined for a computer connected to the ISP's LAN, is forwarded to the Wide Area Network (WAN) - the Internet. For security reasons, the ISP had installed a firewall, a device that keeps unwanted people from tampering with data on it LAN. The firewall job is to receive all data going to/from the Internet, and filtering out it "bad" part.

Step 3: The WAN

Wide Area Network (WAN) is a general term that refers to a network that exceeds local boundaries. A WAN usually includes many LANs in it, which together form a huge network. The world's most famous WAN is the Internet.

Step 4: At your uncle's office

Your uncle's office has a local area network that's connected directly to the internet through a router. A router is a device that serves as a kind of electronic "yellow-pages" of computers. Every computer on the internet has an "address", and routers are the devices that know how to get data to each computer using this address.

Email and your Browser

As discussed earlier, email is a means of sending messages from one person to another using the Internet as the transmission mechanism, much like using the postal service as your carrier.

How many times have you been surfing a web site and wanted to contact the author or webmaster and the only way to do so was via an email address? If you are like most people, your answer would be "Very often". So what do you do? You grab a piece of paper and write down the email address. Then you switch to your email program, enter the address, write your email and send it, then you return to your browser. Tedious to say the least.

This chapter will cover setting up Navigator and Explorer for email and newsgroups. In order to provide a clear distinction between the two browsers, we have separate sections for each of the browsers.

Email Addressing

Like real mail, sending email requires that you know the address of the individual or company you wish to send email to. Unlike real mail, an email address is fairly simple, composed of only two parts. The first part of the email address is the user's account name that you are sending the email to. Some ISPs will allow the user to pick a user name, others will follow a set pattern, like the first name followed by the first letter of their last name, or perhaps even a number.

eg. ravi
ravi001
ravi_s

You also need to know the name of the server. In Ravi's case he logs into the Internet via a server called yahoo.com. Since we can't just merge the two pieces together, we need something to keep them apart, we use the "@" character to

separate the user name from the server name. Hence Ravi's address now becomes one of these;

ravi@yahoo.com

ravi001@yahoo.com

ravi_s@yahoo.com

Server names can be longer and more complex but basically it's safe to say that anything to the right of the "at" sign is the server name and everything to the left is the user name.

Setting up your browser for email

Before you begin setting up your browser to handle email, there are certain questions which only you and your Internet Service Provider (ISP) must answer. You must have this information on hand before you begin, so talk to your ISP. The information you need is;

Your Email Directory.*

Your News Directory.*

Your Email Address.*

Your POP User name.**

Your Outgoing Mail (SMTP) Server Name.**

Your Incoming Mail (POP) Server Name.**

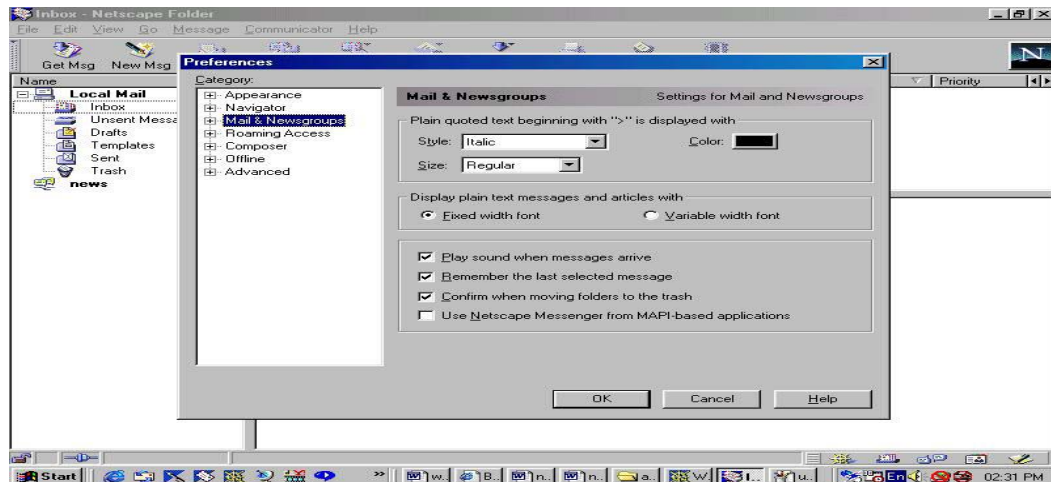
Your News (NNTP) Server Name.**

*You supply this information

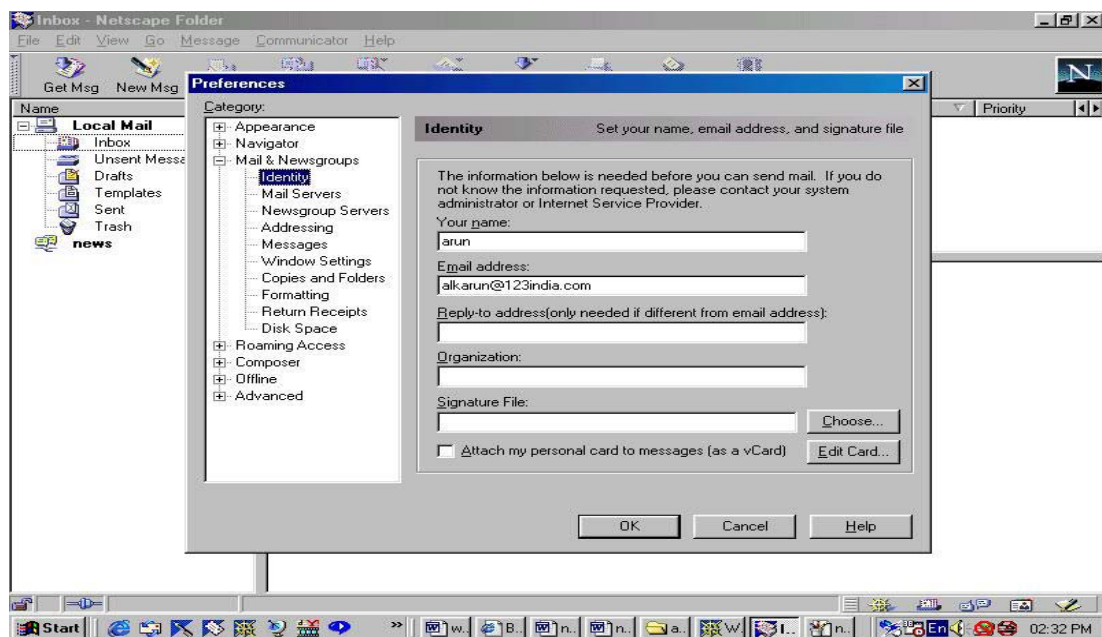
** Your ISP must supply this information

Netscape

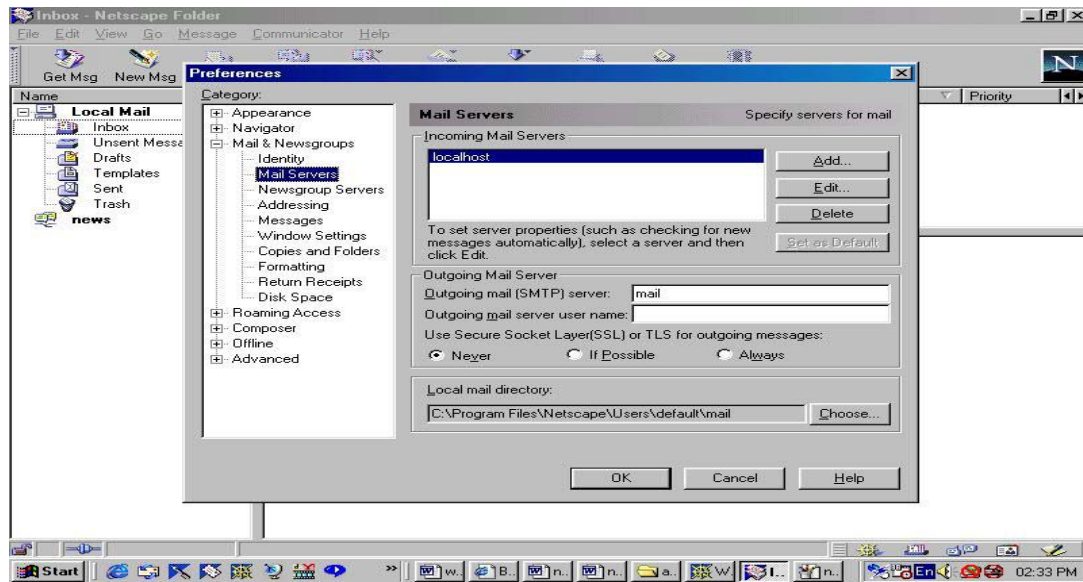
What many people don't know is that Netscape contains an excellent email program built right into itself. The trick is to configure this email subprogram properly. Once you have this information on hand, you are ready to configure your browser for running email. Start up your Navigator and select the "Communicator" menu and click "messenger". Then select "Edit" and choose "Preferences". One of the options under this menu will be "Mail and Newsgroups", select that one.



Click on the tab to select the "Identity" dialog and enter your real name, or a handle if you wish, and your email address in the appropriate position. This section is where the email program will look to find all the information it needs to build the email header, which always includes a name and email address. Clicking on the image to the left will bring up a sample representation of the "Identity" dialog box.



Now click on the "Servers" tab to bring up the dialog box for setting up the servers. This box is where you enter the information provided by your ISP. It informs the email program about facilities on the server which it needs in order to send or receive email. Clicking on the image to the right will bring up a sample representation of the "Servers" dialog box.



One important feature to note is in the Server's Dialog box. When the server name is entered you are prompted to enter more details of your POP server. Then click "POP" and check the option "Leave messages on server". These options are critically important. If Navigator is going to copy your email to the local disk, then delete them from the server, you run the risk of losing email if your connection gets severed during that transfer. For safety's sake, you would like to have the messages remain on the server, but depending upon the server, this might not always be possible. When in doubt, consult with your ISP.

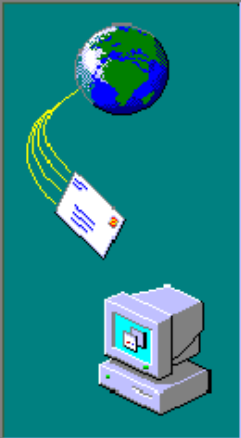
Explorer

Unlike Navigator, Explorer doesn't contain built in software. However if you downloaded the full package from Microsoft, you have not only downloaded mail, but a Newsgroup reader as well. Please note this applies to Windows95 users only. Setting up the email program in Explorer is very different than that of Navigator. In some ways it's easier, because Microsoft has included a setup wizard, but even the best wizard in the world cannot answer the questions you have to answer.

Upon running mail for the first time, either from the task bar, or by clicking on "Send Message" in the Explorer menu, you will be presented with a series of wizard windows which are designed to walk you through setting up email. We have included the important screens here. The others which are not included contain only information and do not require user interaction.

The first screen you will be presented with is shown below. (It's actually the second screen, after one telling you that you need information from your ISP, but you already knew that!)

Internet Mail Configuration



When you send email, your name will appear in the From field of the outgoing message. Enter your name as you would like it to appear (e.g. John Smith).

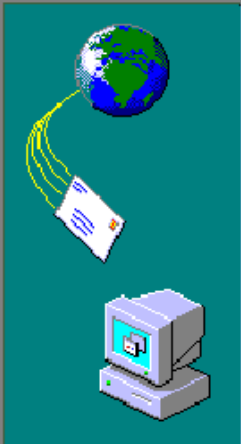
Name:

Your email address is the address other people use to send you email messages. It may consist of your POP3 account name, an "@" sign, and the domain name of your server. Enter your email address (e.g. j-smith@isp.com).

Email Address:

This screen contains two fields you need to enter information into. The first field is your name, the second field is your full email address. Enter both in the window and press the "Next" button to advance to the next wizard screen.

Internet Mail Configuration




A POP3 server is the server that is used for your incoming mail. Enter the name of your POP3 server.

Incoming Mail (POP3) Server:

An SMTP server is the server that is used for your outgoing mail. Your SMTP server may have the same name as your POP3 server. Enter the name of your SMTP server.

Outgoing Mail (SMTP) Server:

This contains information which we said, you would need. If you don't have this information at the moment, then hit "Cancel" and contact your ISP to get the information you need. Otherwise fill in the appropriate information and hit the "Next" button to advance to the next wizard screen.



The image shows a Windows-style dialog box titled "Internet Mail Configuration". On the left is a graphic of a globe with a mail icon and a computer monitor. The main text area contains two paragraphs: "Your email account is the account that you log on to on your POP3 server. This is the account that your incoming messages are sent to so that you can receive them. Enter your email account (e.g. j-smith)." and "When you log on to your POP3 server, you must provide a password. If you do not enter your password below, Internet Mail will prompt you for your password each time you check the server for mail." Below the first paragraph is a label "Email Account:" followed by a text box containing "bobmi". Below the second paragraph is a label "Password:" followed by a text box containing masked characters "*****". At the bottom are three buttons: "< Back", "Next >", and "Cancel".

Internet Mail Configuration

Your email account is the account that you log on to on your POP3 server. This is the account that your incoming messages are sent to so that you can receive them. Enter your email account (e.g. j-smith).

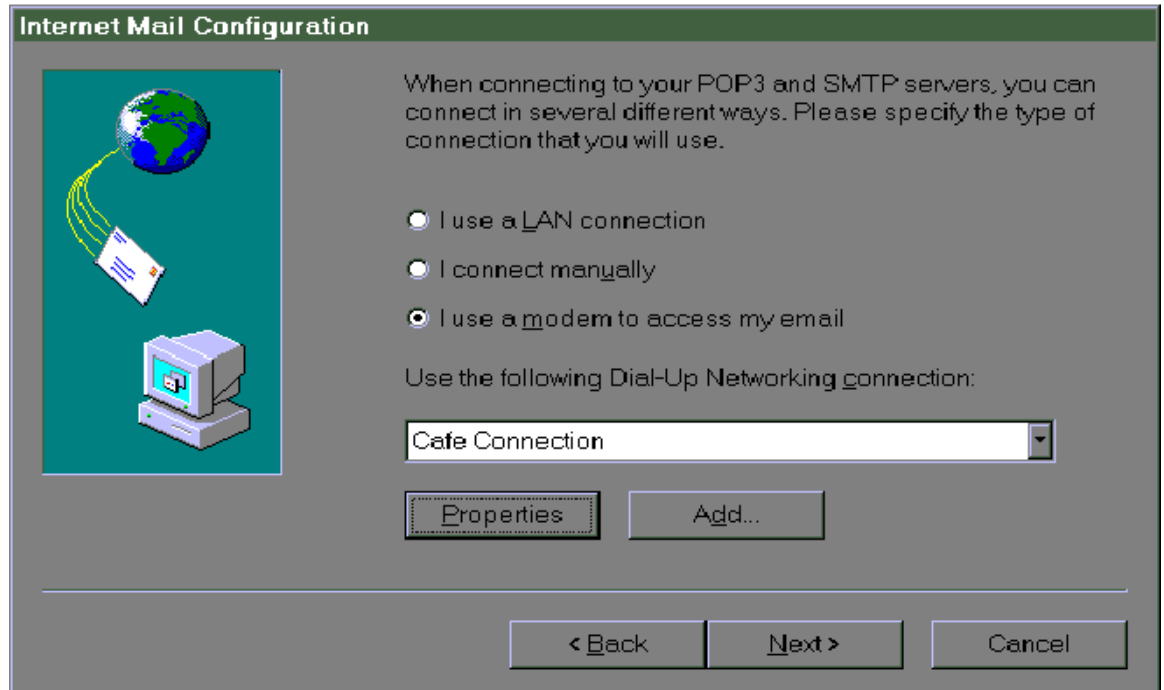
Email Account:

When you log on to your POP3 server, you must provide a password. If you do not enter your password below, Internet Mail will prompt you for your password each time you check the server for mail.

Password:

< Back Next > Cancel

This is the third screen of the email wizard. Here you enter your account name, not your email address (just the portion to the left of the "@" sign). Then enter your password. If you don't know your password, contact your ISP for it. Pressing "Next" will advance you to the final wizard screen.



In this screen you will need to select your method of connection. In most cases this is "I use a modem to access my email". The Dial-Up networking connection is the name of the connection you double click on to start your Internet connection. If you do not know this information, or your ISP setup the connection on your computer, then contact your ISP. They will be able to walk you through finding out this data. Completing this screen will complete the setup for the Explorer Email. Press "Next" to advance to the last wizard screen.

Email, the basics

Email Revisited

Shortly after connecting to your ISP, you are informed that you have email waiting. What to do? Basically your ISP has provided you with a means of sending and receiving email. Now it's up to you to learn how to do it. Before we begin let's look at some of the issues involved in email at a general level.

Receiving Email - General

Can I send email to just anyone? No. In order to send email to someone, you must know their email address. Just like you need to know a street address in order to send a letter to someone, email functions in much the same way. Additionally, you cannot send email to someone who does not have an email account somewhere. A majority of the people who are using email started off using it for business or schooling purposes, but this is rapidly changing also. Email is quickly becoming as valuable a tool as the telephone, or the normal mail system. In order to be useful, your typical email program needs to be capable of;

Receiving Email.

Sending Email.

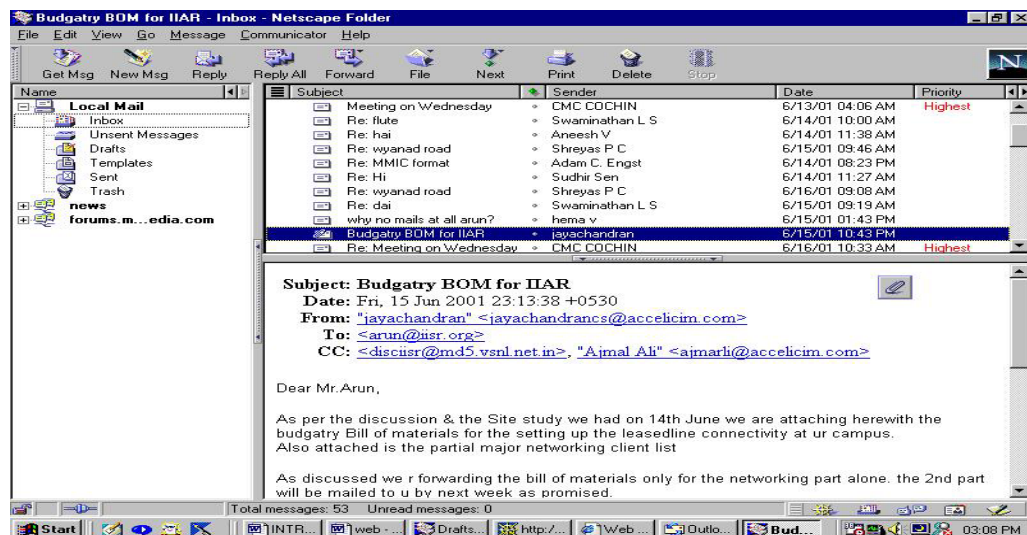
Working with attached files.

Managing an address book.

Receiving Email - Navigator Email Program

Assuming you have properly configured your email program, you can run your Navigator and it will periodically check to see if you have new mail. Unlike Explorer, the mail program for Navigator is tightly integrated into the browser. They even set it up in such a way as to have your browser inform you when new email comes in. by flashing a small icon of an envelope in the lower right hand corner of the browser window.

You can get to the Navigator email reader by using the Window | Communicator –Messenger menu or by clicking once on the Inbox icon in the lower right hand corner of the browser window. No matter which method you choose, selecting the Navigator email program will bring up a program which looks like the image given below.



This program contains several major components. These are the button bar, the folder window (top left window), the message title window (top right window) and the message window (bottom window).

Get Msg - Pressing this button will tell the email reader to open a connection to your server and check to see if you have any new email. If so, it will download the emails into the program for you to view.

New Msg - Takes you to Navigator's Email Authoring program for a new email message.

Reply - Reply to the person that sent you email.

Reply All - Reply to all of the recipients of an address list. Sometime you will get an email where you are merely one of a list of people. Use this option to reply to all of the people on that list.

Forward - Copy the text of an email you received into another email and send it to a totally unrelated individual.

Next - Pressing this button will advance to the next message in your message window. (The one below it)

Print - This will print the currently selected message to your printer.

Delete - Delete the currently selected email(s) from the currently selected folder. You can also use this to delete folders.

Stop - By pressing the STOP button you can interrupt the communications link between your email reader and the mail server. This is useful if your program becomes stuck trying to get mail from a server that might have gone off-line.

The Folder Window. The left window is called the folder window. It contains all of the folders in which you can store your email.

The purpose behind folders is simple. Folders allow you to maintain a minimal amount of order in sorting your mails. For example, you could have a folder for email from friends and another folder for email from family. Most people end up working with a couple of folders in order to maintain some control of how they save their email.

You can create a new folder by right clicking the 'Local Mail'. You can use the DELETE key or the DELETE button to remove the selected folder and all its contents.

A single click on any folder icon will open that folder. Thus causing all of the messages in that folder to be displayed in the Message Title Window.

The Message Title Window: The upper right window is the Message Title Window. It contains the titles and authors of the messages in the currently selected folder.

Reading a message is a just simple click. Click on a message title to view the contents of the message in the message window. Previously read messages are displayed in normal text, while new, unread messages are displayed with their Titles and Authors in bold text.

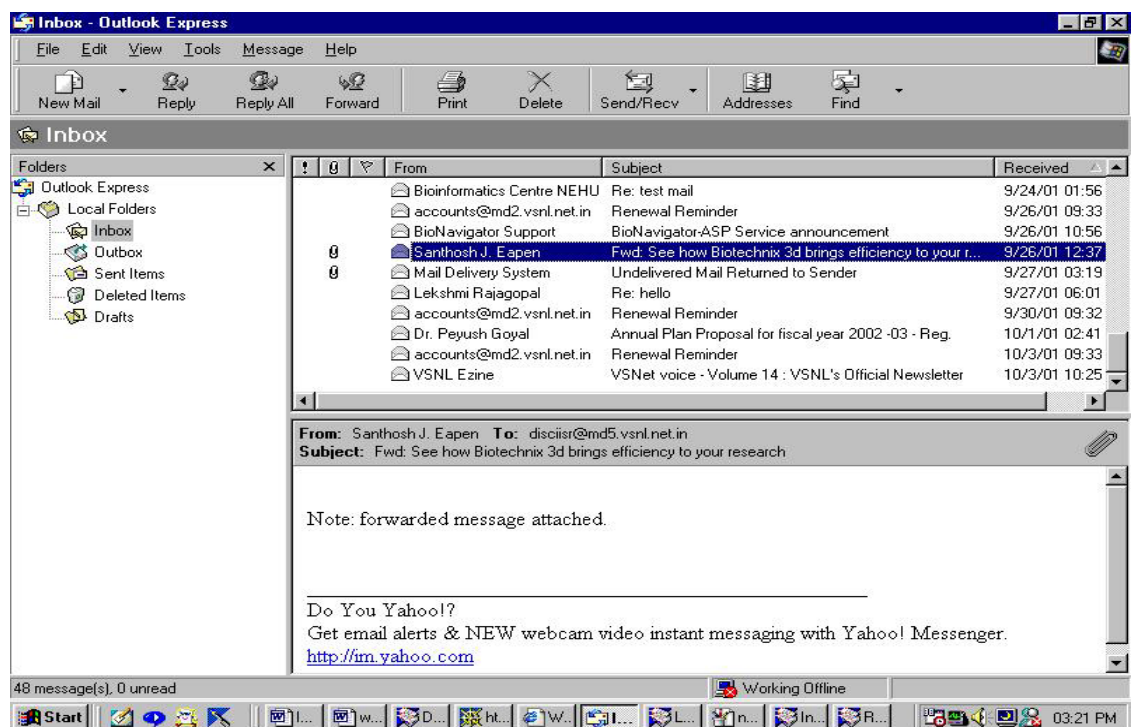
You may move a message from one folder to another by selecting the appropriate message in the message folder window. Then, drag and drop the message onto the desired folder in the folder window.

The Message Window: The window which fills the lower most portion of the screen is the message window. Here the contents of the currently selected message are displayed.

Click on a hyper link and within a few seconds the browser will be running on your screen and you will be off to that location.

Receiving Email - Explorer Email Program

Like Navigator users, Explorer users have a variety of ways to get to their email reader program. From Explorer itself you can click on the MAIL icon on the button bar. Windows95/98 users can also use the start button on the task bar. Using any of these methods will invoke the screen shown below



Again, your screen has been subdivided into sections, four of them in this case. The button bar, a folder selector, the message title window and the message window.

The Button Bar: The button bar basically contains shortcuts to certain menu functions.

New Mail - Pressing this button allows you to write a new email for someone by invoking the email authoring program.

Reply - This button will allow you to send a reply to someone that sent you email.

Reply All - Sometimes you will receive an email in which you are not the only recipient. Pressing this button allows you to reply to all of the email addresses from that email.

Forward - This button will help you to send a message you received to someone else.

Print – Prints the message selected.

Delete - Allows you to delete the currently selected message.

Send and Receive - Pressing this button will serve two functions. It will query your mail server and get any new email that might be waiting for you, and it will send any pending messages, that you have authored, to your email server.

Addresses – Opens the address book that contains a list of addresses entered by you.

Find – A search option for your messages.

The Folder Selector: Like other email programs, the Explorer email reader is capable of handling folders. This control allows you to switch folders with ease. Simply click on the down arrow and then select the folder you desire.

The purpose behind folders is simple. Folders allow you to maintain a minimal amount of order in sorting your emails. For example, you could have a folder for email from friends and another folder for email from family. Most people end up working with a couple of folders in order maintain some control of how they save their email.

You can create a new folder, or delete an old one using the FILE|FOLDER menu options. To create a folder, use FILE|FOLDER|CREATE, to delete use FILE|FOLDER|DELETE or directly by right clicking the “Local Folder” button. In order to move a message from one folder to another you can use the EDIT|MOVE TO or EDIT|COPY TO menu options. This will enable you to relocate the currently selected email from one folder to another.

The Message Title Window: The message title window contains all of the titles of email in the currently selected folder. To view any one message, click on the message and it will be displayed in the lower message window.

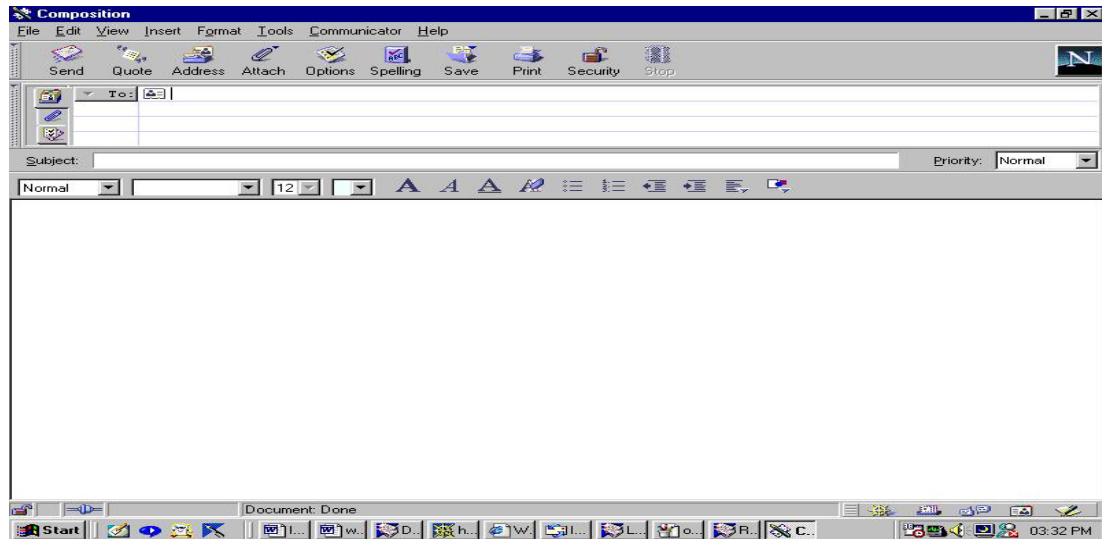
The Message Window: The message window will display the contents of the currently selected message. Like the Navigator Email reader, this area is capable of displaying graphics and is HTML capable. This means that clicking on a link in the message window will invoke your web browser and take you to that location.

Sending Email - Navigator Email Program

Having the ability to read email without the ability to send/reply to email would be like owning a bicycle with only one wheel. There are several ways of reaching the Email Authoring Program. You can access it from the Navigator FILE|NEW

MESSAGE menu or press NEW MSG button or from the email reader by pressing the TO:Mail, REPLY:MAIL, REPLY:ALL or FORWARD buttons on the button bar.

Using any of these methods will bring up the Email Authoring program. Shown below is a typical screen for the Navigator email authoring program.



There are three basic components to the Navigator email program, the button bar, the envelope and the message area.

The Button Bar: The button bar contains several one-shot functions. By one shot we mean that you should press these buttons only once. Pressing any button more than once would be pretty meaningless.

The button bar contains the following buttons;

SEND - The SEND button is exactly what it says. Pressing this button will cause the email program to connect to your server and attempt to transmit your email program.

QUOTE - The process of QUOTING is simple. In reply to someone else's email, you decide you wish to either quote something, or perhaps rebut what has been said. Pressing QUOTE will transfer the entire contents of their email, into the edit window for you to use.

ADDRESS - Pressing this button will call up the address book for the authoring program.

ATTACH - This button allows you to attach a file to a email/newsgroup posting.

OPTIONS – Gives options for the format of your mail.

SPELLING – Checks the spelling of your message.

SAVE – Saves the message.

PRINT – Prints the message you've typed.

SECURITY – Shows security information.

STOP - Halts the communications between the authoring program and the mail/news server.

The Envelope: Like real mail, email has an envelope. This envelope is, however, unlike any other mail envelope you have ever encountered. We know that in order to send email to someone, we need to know their address. Assuming I want to send email to my Aunt Mary, I might put "AuntMary@nyc.com" in the "To:" field of the email envelope. The "To:" field is where you place the email address for the person or persons you want this message to go to. If you want the email to be received by more than one person you can also include additional email addresses here. Just separate them by commas. Like this;

addr1@anyserver.com, addr2@myserver.com, etc.

The "CC:" field is the carbon copy field. It's major use is to send a duplicate copy of the mail to one or more secondary individuals. For example lets say you wanted to send an email to your boss, reminding him about your vacation. Your envelope might look like this;

To: BigBoss@mycorp.com

CC: payroll@mycorp.com, departmentX@mycorp.com

The subject field is very important. The authoring program will complain if you do not fill in this field. Basically, it's a good idea to have a subject field. It's the title that others will see in their Message Title Window. So lets add a line to our email to the boss.

To: BigBoss@mycorp.com

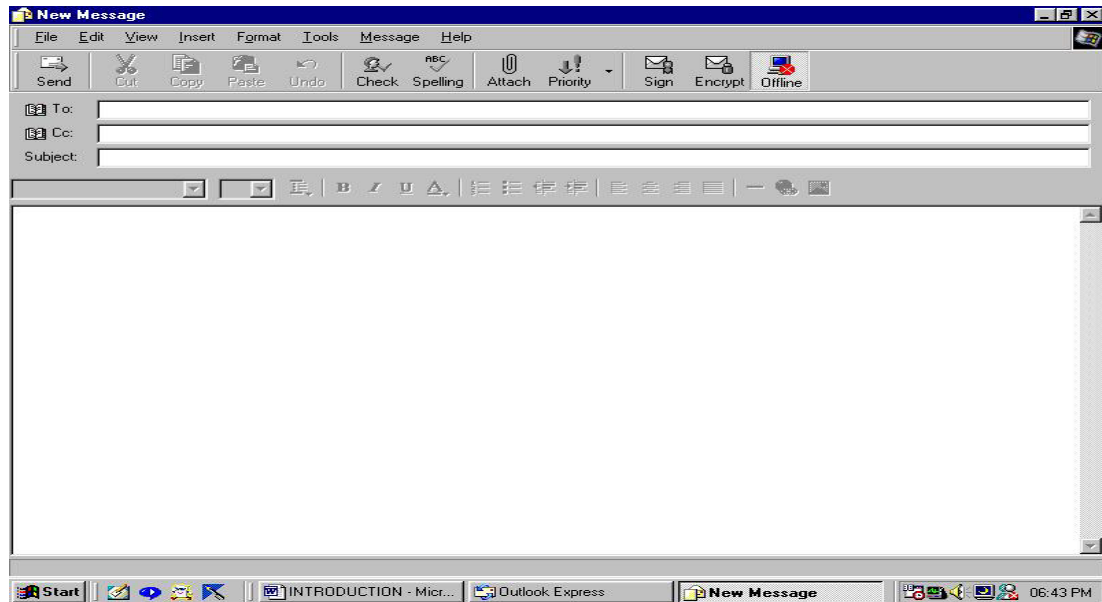
CC: payroll@mycorp.com, departmentX@mycorp.com

Subject: Reminder: Scheduled Vacation begins Monday.

Notice that the subject clearly informs the recipient as to the content of the message. The lesson here is simple. Try to be concise and informative with your subject line.

The "attach" button allows you to attach a file or web page to your message.

Sending Email - ExplorerEmail Program



In each of these programs, there is considerable similarities. So if something isn't explained for one program, it will be explained for the other. The basic layout of these screens comprise three major sections, a button bar, the envelope and the message entry area.

The primary difference lie in the button bars and the envelope areas. We shall address the email authoring program first. The button bar for the Explorer Email program is more extensive. Looking from left to right we have the following buttons.

Send - Send the current message to your email server for transmission out on the net.

The next four buttons Undo, Cut, Copy, Paste are windows functions taking advantage of your windows clipboard. For more information on using these functions, we suggest you look them up in your windows manual or your windows help files.

Address Book - Pressing this button calls up the address book for the Explorer email program.

Check Names - This button is somewhat misleading. The purpose behind the button is simple. It provides a means of checking that the addresses you have entered are in the correct format. It does not check to see if the address is a valid one.

Attach - Click on this button if you want to send this email with a file attached to it.

Priority – Options for priority of your message

Sign - Pressing this button will insert your signature file into the email message.

Encrypt – For encrypting the message.

Explorer Email Envelope Area

The envelope region of the screen (below the button bar) contains three basic fields. These are;

To:

CC:

Subject:

We know that in order to send email to someone, we need to know their address. Assuming I want to send email to my Aunt Mary, I might put "AuntMary@nyc.com" in the "To:" field of the email envelope.

The "To:" field is where you place the email address for the person or persons you want this message to go to. If you want the email to be received by more than one person you can also include additional email addresses here. Just separate them by commas. Like this;

addr1@anyserver.com, addr2@myserver.com, etc.

The "CC:" field is the carbon copy field. It's major use is to send a duplicate copy of the mail to one or more secondary individuals. For example, let's say you wanted to send an email to your boss, reminding him about your vacation. Your envelope might look like this;

To: BigBoss@mycorp.com

CC: payroll@mycorp.com, departmentX@mycorp.com

The subject field is very important. The authoring program will complain if you do not fill in this field. Basically it's a good idea to have a subject field. It's the title that others will see in their Message Title Window. So lets add a line to our email to the boss.

To: BigBoss@mycorp.com

CC: payroll@mycorp.com, departmentX@mycorp.com

Subject: Reminder: Scheduled Vacation begins Monday.

Notice that the subject clearly informs the recipient as to the content of the message. The lesson here is simple. Try to be concise and informative with your subject line.

The Message Area: By far the largest region in this program, the message area is where you type your message in. We suggest learning the keystrokes for an application like Notepad, since windows uses the same keystroke set across the board for navigating in large text entry areas.

FTP (File Transfer Protocol) and Telnet

FTP - An Introduction

Computers work with data and programs. The ability to transfer data and programs is an integral part of the standard suite of Internet functions. One of the basic abilities provided by most ISP's is UPLOADING/DOWNLOADING of files. There are many different ways of uploading/downloading of files, but in this chapter we are going to review the single most popular method available on the Internet. Simply put, uploading is the process of moving a file from your computer to some remote computer. Downloading is the exact opposite, that is, moving a file from some remote computer to your computer.

One of the oldest Internet methods for moving files from one computer to another is called FTP. FTP stands for File Transfer Protocol. The name "FTP" can represent both a file transfer program, and the underlying protocol used to send information. It's perfectly natural to hear a person say "I used FTP to send a program," or, "Just FTP the files I need." It's a generic catch-all phrase for both the program being used and for how the information is transmitted from one computer to another.

Navigator/Explorer users know about FTP. A partial version of FTP is built right into their program, but using it as such is cumbersome and difficult. Unlike email and network newsgroups, there is no distinct interface for FTP within Navigator or Explorer browsers. One way of using FTP from within your browser is from a web page. Often you will find a link which says "Download Now!" or something similar. Clicking on that link will automatically tell the browser to take the next data being sent to your computer and save it as a file on your local harddrive. This is by no means complete FTP. For example the direction of the link is all one way. Your browser doesn't have an easy way of sending files, only receiving them.

Given the fact that the browsers do not have a decent capability to send files, what other options are open to us? There are several extremely powerful programs available on the Internet right now that can perform the task of FTP in both directions. All we need to do is go get one of these programs. Most of the programs available on the Internet are the result of many hours of hard labor by one or more individuals. These programs are usually called shareware, which means you can download a fully working copy of the program and if you like it enough to keep it, the authors would like you to pay for it. Most shareware programs are extremely inexpensive and worth the money the author is asking for. Show your appreciation for the author and his effort and preserve your honor by paying the requested amount to the author!

Now for the list of locations from which you can download files. These locations include a lot more information and programs than just FTP programs (This is a more expanded list than the one given in chapter II).

Queue ftp (<http://www.eesoft.com/qftp>)

Cuteftp (<http://www.cuteftp.com>)

ftppro (<http://www.ftppro.com>)

turboftp (<http://www.turboftp.com>)

FTP Basics

If you have got a good FTP program you will have quite a lot of capability, most of which you will never use. There are six basic commands/functions common to all FTP programs which you will use, over and over again. These are;

OPEN
CLOSE
CD
DIR
GET
PUT

OPEN - FTP is both a communications protocol and a file transferring program. Therefore the first thing you need to do is OPEN a communications channel from your computer to the designated FTP server.

This is the same as logging into a server or bbs. Typically a command to open a FTP server could look like anyone of these samples.

open ftp.blah.com

open moreblah.com

open 87.123.56.92

open 87.123.56.92 5000

If you are lucky enough to have downloaded a windowing version of ftp (for Macs or Windows), then you would have a dialog box to enter the name, or address, of the server. Like the URL used in your browser, this name is an important piece of information as it is the only item needed to locate the server on the Internet.

CLOSE - Perhaps the easiest of the FTP commands. This command simply tells the remote computer that you are finished and wish to close the communications channel.

CD - This is the Change Directory command. Like most computers, FTP servers and their contents are stored in directories. The CD command gives you a means of navigating these directories.

DIR - DIR works just like the MSDOS version. It performs a directory listing of the currently selected directory.

GET - Now for the heart of FTP. GET is the command used to retrieve a file from the remote computer and store it on your computer. Some windowing

FTP programs allow an implied "GET" by letting the user click and drag a file from one window to another. Other FTP programs require you to type the command out directly. Some typical typed "GET" commands could be;

```
get stuff.txt
```

```
get /pub/b/stuff.txt
```

```
get /pub/c/programs.obsolete/stuff.txt
```

It's important to note that the file is going to be stored in whatever YOUR current directory is. For this reason it's a good idea to start your FTP program from a download directory. Alternatively, if you have a windowing version of FTP, you can navigate your local directories much like you can with file manager.

PUT - The reverse of "GET", this command allows you to send one or more files from your computer to a remote computer. There is a note of caution that should be aired now. Make sure that you have prior permission from the system owner before placing anything into a /pub directory. And remember, IT IS ILLEGAL TO USE THESE PROGRAMS TO TRANSFER COPYRIGHTED SOFTWARE TO UNAUTHORIZED INDIVIDUALS.

Telnet

In this section we will explore using Telnet as a means of connecting to various services on the Internet. Telnet is a character based terminal application and is one of the granddaddy services available on the Internet for what seems like eons. Before we begin, let's backtrack a little and talk about the Internet first.

The Internet is basically a network of thousands of computer systems. There are two basic types of computers connected to the Internet, Hosts and Clients. Clients are machines which access the Internet on a part time basis. They have no permanent connection to the Internet, but instead are capable of hooking to the Internet via a dial up modem, or some similar method. The odds are very good that you are reading this tutorial right now using a Internet client computer.

It's the HOST computers which we are going to be concerned about. Host computers are always connected to the Internet, and in most cases rarely are disconnected from the network. Host computers are where web pages reside, where your email is stored until you retrieve it, where chats and mud's are played. When the Internet was first designed, Telnet was just about the only means of people accessing information on another computer from their own. In effect, Telnet turns your computer into a dumb terminal. What's a dumb terminal? Well in the early days of home computers (back in the mid 70's), a dumb terminal was a common way of talking to your computer. It was a video display and keyboard, nothing more. It had the ability of displaying the information your computer put out and allowing you to enter new information into your computer via the keyboard. Very limited or no graphics, extremely limited functionality. Obviously, we have come a long way since then, but some things haven't changed all that much.

Once you are connected to the Internet, you can start a Telnet session. If you don't have a Telnet program, you can obtain one from any of the ftp site archives. Windows95/98 users should also note that there is a Telnet application on your distribution CD for windows95/98. This program is full featured and contains a built-in ftp program which is especially nice to use. Despite the enormity of the web & the richness of the ftp sites, there is still some things which you can only get to via Telnet.

IRC

MUDS/MOOS

Public Access Servers. (i.e. libraries, stores etc..)

The above list is only a partial listing of things you can do with Telnet.

IRC - IRC stands for Internet Relay Chat. It's a system which allows you to talk "live" with other people around the world. Even such applications as mIRC, a very popular windows based client for IRC's, is nothing more than a fancy, dedicated Telnet front end with IRC extensions built right in. Warning!!! IRC can have an addictive effect on both your working and social life! The ability to talk real time with people from anywhere in the world has proven to have a strong effect on people, and yet, IRC can be a wonderful resource for people looking for information.

MUDS - MUDS, or Multi-User Dungeons, are on-line games with a fantasy role playing theme.

Public Access Servers - There are a considerable number of public access servers available from the Internet.

Nearly every Internet service provider has the ability to accept Telnet sessions. First of all in order to Telnet into a system you will either need to have an existing account or the system will need to be a public access system. Attempting to gain entry to a system which you do not have an account on is called "Hacking" and is against the law. If you have a legitimate need to access something on a system, an email to the root/sysop/sysadmin may get you an account, or at least the data you require. Never, ever, attempt to access a system to which you do not have an account.

References

1. RFC's
2. Yahoo - DNS
3. <http://www.knowhereglobal.com>

EXERCISES FOR INTERNET & E-MAIL SESSION

- 1 Create your own email accounts if you don't have. (If it is possible in IISR's local area network, the authority can already create one for the participants for the time being.)
- 2 Create new address books and folders in your email id. & try aliases.
- 3 Practise to use a search engine.
- 4 Using WWW visit a website and try to get into the connected links.
- 5 Try FTP facility; try *get*, *put* & *ls* commands.
- 6 Try *finger*, *ping*, and *talk* facilities.

INTERNET RESOURCES IN BIOLOGICAL SCIENCES

M.G. Sreekumar

Indian Institute of Management, Kozhikode

Introduction

The explosive growth of the Internet and the World Wide Web (WWW) in recent years have its impacts on almost every field of human activity, and Biological Sciences is no exception. The secret to understanding the Internet is that it is populated by two types of computer programs: servers and clients. Servers are programs that provide resources and Clients are programs that we use to access these resources. E-Mail, LISTSERV/Mail lists, USENET/News groups, FTP, Telnet, Gopher, Archie, WWW (World Wide Web) etc., are among the prominent services of the Internet. Each type of service in the Internet has its own client. For example, to access the WWW, we need to use a Web client such as "Netscape" or "Internet Explorer". As already mentioned, Internet is a collection of networks covering the world. These networks contain many different types of computers and there should be a tool to hold the whole thing together and that is "TCP/IP", which stands for Transmission Control Protocol/Internet Protocol. A protocol is a set of rules describing, in technical terms, how something should be done. TCP/IP is the common name for a collection of more than 100 protocols used to connect computers and networks.

The WWW is almost synonymous with the Net. It is a large system of servers which offers all kinds of information on the Net. The information can be in the form of regular text, as well as pictures, sounds, video clips, and other types of data. Invented in early 1990s by physicists under the leadership of Tim Berners-Lee, at the European Particle Physics Laboratory (CERN), Switzerland, the Web has rapidly become the Graphical User Interface (GUI) of the Net. In fact, the Web is just one Internet application, a way of using the vast interconnected network and view information from around the world. The main use of the Web is for information retrieval, whereby multi-media documents are copied over the Net for local viewing. It uses a protocol called HTTP (Hypertext Transfer Protocol). The Web is in fact a multi-media hypertext - any sort of digital data can be distributed inside a Web document, and each document contains links to other documents, as shown by highlighted or underlined text. Simply click on the word or the highlighted area (where the palm appears) and we can travel to the document in question. Web documents are written using a markup language called Hypertext Markup Language (HTML). HTML is independent of platforms. So it does not matter if you are using a Macintosh, a PC or a Sun workstation to access them. The Massachusetts Institute of Technology's (MIT) World Wide Web Consortium (W3C) is the international standardising agency for HTML and HTML 4.0 is the latest standard. HTML has tags for providing references to other Web pages, which can be on the same server or any other server on the Net irrespective of the geographical location. This facilitates hypertext links across the documents on the Net. Web pages can contain references of images (GIF and

JPEG formats), audio files (AU and WAV formats) and video files (MPEG format) which thus adds the multi-media dimension to the information provided on the Web. The clients or browsers access the Web pages on the servers, renders and formats them according to the HTML tags to display on the client's system. When the user selects a hypertext link (indicated by an underscore) on a Web page, client can follow the link and fetch the referred document irrespective of the location of the document on the Net. Every Web document has a unique address, called the Uniform Resource Locator (URL). It is a simple way of describing almost any information resource, using a standard format for locating information on the Net as:

<protocol>://<host.domain>:<port>/<protocol-specific>

It is an Internet address which is unique to that resource. It consists of a protocol, a host name, a port (optional), a directory and a file name. That is, the URL consists of the computer on which the document is stored and a file name. For example, the URL of the Indian Institute of Management at Kozhikode, is: "http://www.iimk.org". Please note the file name is missing in the example, as it refers to the home page.

The key to the Web is a browser program, which is used to retrieve and display Web documents. The browser (HTML viewer) is an Internet compatible program that runs on our local computer, whether it is a Macintosh, IBM PC, UNIX Workstation, and does three things for the Web documents : 1. it uses the Internet to retrieve documents from other computers, called servers; 2. it displays these documents on our screen, using the formatting specified in the document; and 3. it makes the displayed documents active, so that pointing and clicking on a cross-referenced item in a document will take us to the reference. There are two popular Web browsers, viz., the Navigator of Netscape and the Internet Explorer, from Microsoft. However, Netscape has become the de facto standard for browsers.

Internet Information Infrastructure (III)

Look closer, the Internet has three major tasks in the handling of information. There is no second opinion, it is pro publishing and it is a content publisher. It acts as a content holder too. The Internet presents itself as the most logical repository for content because the Internet is accessible to applications, laptop and desktop computers, Internet appliances, cellular telephones, and personal digital assistants. Fortunately, the Internet can provide the information infrastructure necessary to provide secure, collaborative communication and access. Hence it also acts as a content communicator, with its hosts of protocols and standards.

The Internet Information Infrastructure can replace the basic functionality of a corporate extranet - minus an extranet's management cost and complexity. The Internet has become more than merely a platform for intercommunicating - the

Internet with a set of federated services *is* the computer. The Internet Information Infrastructure provides the federated service set. With user devices acting as I³ nodes, the "Internet-computer" connects users and applications, replacing the extranet. On a global perspective, the Unicode could play wonderful role in taking the fruits of technology across platforms and languages. The emerging metadata standards and systems are again rays of hope towards resource discovery in the Net.

Unicode is a *lingua franca* that enables a single software product or a single Website to be targeted across multiple platforms, languages, and countries without re-engineering, and for data to be transported through many different systems without corruption. Data processing industry leaders have adopted the Unicode standard. Unicode is required by modern standards such as XML and Java, and is supported in many operating systems and all modern browsers. Incorporating Unicode into client/server or multi-tiered applications and Web sites offers significant cost savings over the use of legacy character sets. The Unicode Consortium operates with the International Standards Organization (ISO).

Metadata - that is, data about data - is information that describes content. Metadata works like a card catalog in a library, an entry in a book catalog, or the information in an online index. Metadata allows a user or application program to identify precisely what it is that the associated data represents and show the relationships among different pieces of content, even accommodating changing relationships. The best metadata system would be extensible, so that metadata types that can be anticipated easily - e.g., author, version, size - can be defined as files are uploaded; moreover, new types of metadata can be added after file creation, as needs dictate. The W3C's (World Wide Web Consortium) DCMI (Dublin Core Metadata Initiative) is a major development in this front.

Biological Literature

Finding the vast amounts of biological data related to the management, study, and use of biological resources cached in hundreds of databases around the world is difficult at best. Mostly a major part of this information sits unused because of a lack of efficient means to retrieve and derive comprehensible information from it. A solution is in the midst of development: The National Biological Information Infrastructure, an initiative of the US Geological Survey's Biological Resources Division (BRD), will eventually allow researchers, land and natural resource managers, students, and the general public to locate these databases, query multiple databases simultaneously, and retrieve data.

When NBII is completed, users will be able to draw from any number of databases to assess biological and environmental processes and conditions. NBII will also provide software to help users analyze these data. A "distributed electronic federation," NBII, in its current form (www.nbii.gov/index.html), has

two components. The first is a set of links to a variety of government and non government sites. Most of these sites do not provide direct access to biological information but instead direct the user to potential sources of information.

The second component is a "metadata clearinghouse," in which research products such as databases, published papers, and technical reports are cataloged using a standard set of descriptors known as "metadata". These metadata describe the subject matter, taxonomic subject, time span and spatial limits of the research, how the data were collected, who collected the data, and where the research product can be found on the Internet or elsewhere. In essence, the clearinghouse is an electronic card catalogue that allows users to find information across a wide variety of sources.

The Association of Systematics Collections (ASC) and the United States Geological Survey, Biological Resources Division (USGS-BRD), established a cooperative agreement to develop two resource databases. The databases were developed to support the National Biological Information Infrastructure (NBII), the electronic gateway to biological data and information, and the Integrated Taxonomic Information System (ITIS), a partnership dedicated to provide an on-line, scientifically credible list of biological names.

The Taxonomic Resources and Expertise Database (TRED) lists taxonomic experts for North America, north of Mexico, with their taxonomic, geographic, and habitat expertise, along with address and related information. The second database, the Directory of Research Systematics Collections (DRSC) was developed via the *Survey of Research Systematics Collections (SRSC)*. This database serves as a directory of research natural history collections and their databases.

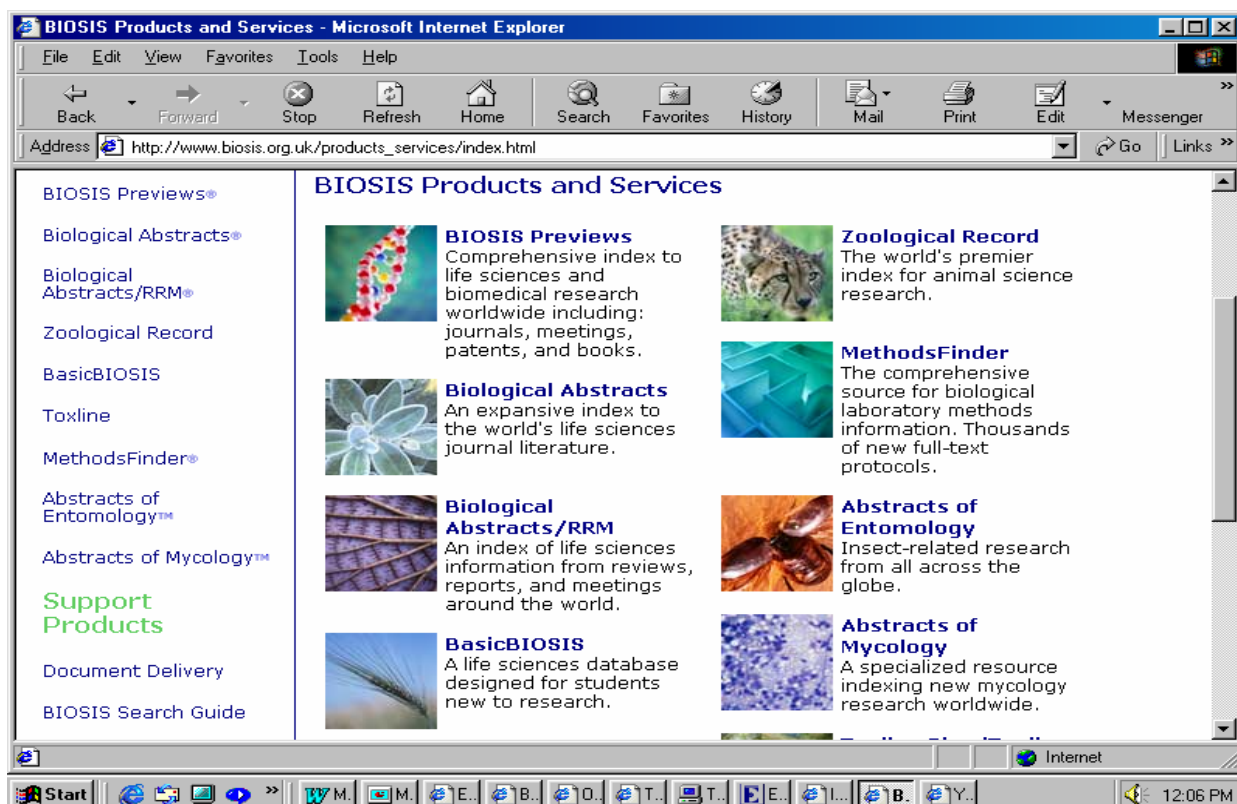
Bioinformatics

Bioinformatics harnesses the power of computational and information technologies to organize and analyze biological data from research collections, experiments, remote sensing, modeling, database searches and instrumentation – and deliver them to users throughout the world. Today, the Internet and World Wide Web are powerful tools for linking and utilizing the extraordinary assets of natural history and research institutions. In India, under the DBT, Bioinformatics Centres located in different parts of the country are catering to the needs of thousands of scientists in meeting their ever increasing information and computational needs.

BIOSIS

BIOSIS, established in 1926, is a not-for-profit organization whose mission is to foster the growth, communication and use of biological knowledge for the common good. By accumulating the world's largest collection of abstracts and bibliographic references to worldwide biological and medical literature and

organizing this information for easy access, BIOSIS provides vital sources of information for life scientists.



BIOSIS processes approximately 550,000 items each year, from primary research and review journals, books, monographs, and conference proceedings. The BIOSIS information system, a pool of material totalling over 13 million citations at the end of 1999, is available in several formats. These include *Biological Abstracts*[®] (*BA*), the leading reference database for life science information and *Biological Abstracts/RRM*[®] (*Reports, Reviews, Meetings*) (*BA/RRM*), the companion reference to *Biological Abstracts*. BIOSIS also publishes *Zoological Record*, supplying a comprehensive collection of references to zoological research articles worldwide, for which it processes an additional 72,000 items each year. The *Zoological Record* database includes approximately 1.3 million records from 1978 to the present. BIOSIS' print products are also available in CD-ROM and Web formats, along with *BIOSIS Previews*[®] on CD-ROM and Web. In addition to print, compact disc, and Web formats, BIOSIS' extensive database of citations can be accessed via an online service. Online products consist of *BIOSIS Previews*[®], *Zoological Record Online*[®], *BasicBIOSIS*[®], and TOXLINE.

Biological Abstracts (BA)

The Biological Abstracts is the Key to the World's Life Sciences Journals. Comprehensive coverage and context-sensitive indexing make the information in *BA* essential for all life sciences researchers. *BA* directs users to information on life science topics from botany to microbiology to pharmacology, serving to connect researchers with critical journal coverage. Whether the topic of study is botany, pharmacology, biochemistry, or evolutionary ecology, *BA* has the journal articles that one's research depends on. Total Journal Coverage *BA* indexes articles from over 4,000 serials each year. This publication also offers Over 360,000 new citations each year, Nearly 90% of citations include an abstract by the author, and almost 5.8 million archival records are available back to 1980. *BA* articles originate from journals all around the world, and cover topics in every life sciences discipline. If the information you need lies in the life sciences, *BA* should be part of your information solution. Plus, BIOSIS indexing, MESH® disease terminology, and CAS Registry Numbers® all help electronic *BA* users find the most relevant records efficiently. Subscribers can access *BA* on CD, through the Web, or in print. Electronic versions of *BA* are updated quarterly. Print versions are issued twice each month.

AGRIS

AGRIS, the international information system for the agricultural sciences and technology, was created by the Food and Agriculture Organization of the United Nations (FAO) in 1974, to facilitate information exchange and to identify world literature dealing with all aspects of agriculture. AGRIS is a cooperative system in which participating countries input references to the literature produced within their boundaries and, in return, draw on the information provided by the other participants. 199 national, international and intergovernmental centres participate and submit about 14,000 items per month. The system collects bibliographic references (to date, about 3 million) to either conventional (journal articles, books) or non-conventional materials (sometimes called "grey literature" e.g. theses, reports, etc.), not available through ordinary commercial channels. One of the main reasons for AGRIS' existence is to encourage the exchange of information among developing countries, whose literature would not be covered by other international systems. The AGRIS database is now on line at this site, as well as AGROVOC, the Multilingual Agricultural Thesaurus, which is directly accessible at "<http://www.fao.org/agrovoc>" Coming under WAICENT's "umbrella mechanism" will bring numerous advantages to participating countries such as a.) access to full text in electronic format; participant control of scope, coverage, description and distribution of local information; b.) faster access to information following elimination of centralized conversion and processing; and c.) greater exposure of national information and access to world agricultural data via the WAICENT server. FAO's primary role will be to assist countries in reaching the goal of self-sufficiency in information management. The Organization will continue, through WAICENT, to develop tools and methodologies which can be used at national, regional and international levels, as part of the global information exchange process; sensitize decision-makers about the importance of

investing in information systems; and establish training courses both at FAO and in developing countries.

AGRICOLA

Compiled by the National Agricultural Library of the U.S. Department of Agriculture, and published on the Web by Community of Science (COS), AGRICOLA (Agricultural Online Access) is the most comprehensive source of U.S. agricultural and life sciences information. AGRICOLA serves as the document locator and bibliographic control system for the National Agricultural Library (NAL) collection. It contains over 3.3 million bibliographic records of journal articles, monographs, theses, patents, software, audio-visual materials, and technical reports related to agriculture from 1979 to the present. Indexers regularly scan over 1,400 journals for input into the database. Since 1984, the database also has included some records produced by cooperating institutions for documents not held by the NAL. Data is provided to Community of Science under contract with the National Technical Information Service. AGRICOLA provides comprehensive coverage of newly acquired worldwide publications in agriculture and related fields, covering the field of agriculture in the broadest sense. Records are catalogued using the controlled vocabulary of Library of Congress Subject Headings and, since 1985, of the CAB Thesaurus.

AGRICOLA through Community of Science, enables you to:

- Select a search interface designed for your level of search experience.
- Search for abstracts in specific journals.
- Download search results in citation formats for use with most citation management software packages.
- Track your queries so that you can repeat past searches against current data.
- Search by MARC record field.

A database search produces a customized bibliography. Within seconds, the system provides a table of references found on a given topic. You can then specify which references you want to see.

CAB Abstracts

CAB ABSTRACTS[®] is a bibliographic database compiled by CAB (Commonwealth Agricultural Bureau) International. It covers the significant research and development literature in the fields of agriculture, forestry, aspects of human health, human nutrition, animal health and the management and conservation of natural resources. Over three million records have been added to

the database since its computerization in 1973. These records are made available through a wide range of products and services both printed and electronic.

CAB International's Internet Products.

CABDirect® - A new combined CAB ABSTRACTS and CAB HEALTH database accessible via the Internet with links to full text via Ingenta Journals. Online Communities include i. Animalscience.com - A new online community for specialists in animal health and production; ii. Nutritiongate.com - A new online community providing access to information and facilities in the areas of human and animal nutrition; iii. Leisuretourism.com - A new online community for the leisure, tourism and recreation sectors; iv. organic-research.com - A new online community providing access to information on organic farming and worldwide research; v. Primary Journals - The content of CAB's primary and review journals is available with at least a one year backfile at no extra cost for organizations subscribing to the printed version; vi. CABI Online Abstract Journals - All the abstract journals are available on the Internet as fully searchable archival datasets - minimum 10 year back file; vii. Agricultural Engineering on the Web - an online service providing access to the world's published research literature in agricultural engineering; viii. AgBiotechNet - provides easy access to information on subjects such as genetic engineering, *in vitro* culture, biosafety, intellectual property rights and other key issues in agricultural biotechnology; ix. Animal Health and Production Compendium.

CABWeb®s

HORT CABWeb® - providing desktop access to the world's published scientific research literature covering all aspects of horticulture and seed science. **PEST CABWeb®** - providing desktop access to the world's literature on crop protection, including: entomology, nematology, weeds, biological control, pathology and other aspects of pest management. **TREE CABWeb®** - providing desktop access to the world's published scientific research literature in forest science, agroforestry and forest products.

The Agricultural Information Scenario in India

Under the auspices of the ICAR, the project "Agricultural Research Information System (ARIS)" is being implemented to bring information management culture to National Agricultural Research System (NARS) so that agricultural scientist can carry out research more effectively by having systematic access to research information available in India as well as in other countries, better project management of agricultural research, and modernization of the office tools. The basic infrastructure required for linking all ICAR institutes has already been created. The E-mail connectivity has been established to 72 out of 86 ICAR institutes by linking through dial-up including six institutes with VSAT connectivity using NICNET and ERNET services. ARIS has four information

modules namely Agricultural Research Personnel Information System (ARPIS); Agricultural Research Financial Information System (ARFIS); Agricultural Research Library Information System (ARLIS) and Agricultural Research Management Information System (ARMIS). The ICAR is also running an Agricultural Research Information Centre (ARIC) as the Central source of information on all research projects and schemes financed by the ICAR. Databases on A.P. Cess Fund Schemes, Deputation reports and research projects of institutions were computerized and updating of the databases is taken up. The ARIC is the national input centre for the AGRIS and CARIS agricultural databases of the FAO, which is the largest information system of its kind in the world. The ARIC is also the national focal point for the SAARC Agricultural Information Centre (SAIC). It has published several directories in addition to a half-yearly Directory of conferences, seminars, symposia and workshops in agriculture. It is being upgraded to provide on line uplinking and downlinking facilities to the ICAR system, and to the agricultural information system of the entire world.

Agri Informatics and India

Agricultural Marketing Information: *A key step towards globalisation of Indian Agriculture* The availability of prompt and reliable information about what is happening in the market, what quantities are arriving and what prices are quoted for different commodities considerably improve the decision making capability of the farmers and strengthens their bargaining power. At present, the information is disseminated through various media like radio, newspapers, blackboard display and public address system at market yards. The information provided by these methods is stale and does not help the farmers sufficiently in taking decisions in marketing their produce. The farmers are also not able to know about the prices prevailing in other markets, as the Market Committees are able to disseminate information in respect of their own markets. The farmers are, therefore, left with no alternative but to dispose of their produce in the nearest market, even at uneconomic prices. Improvement in the present market information system by linking all important agricultural produce markets in the Country, the State Agricultural Marketing Boards & Departments, under Agricultural Marketing Information Network (AGMARKNET) will enable farmers to take market based decisions and benefit other market users as well.



Access to scholarly literature through other sources

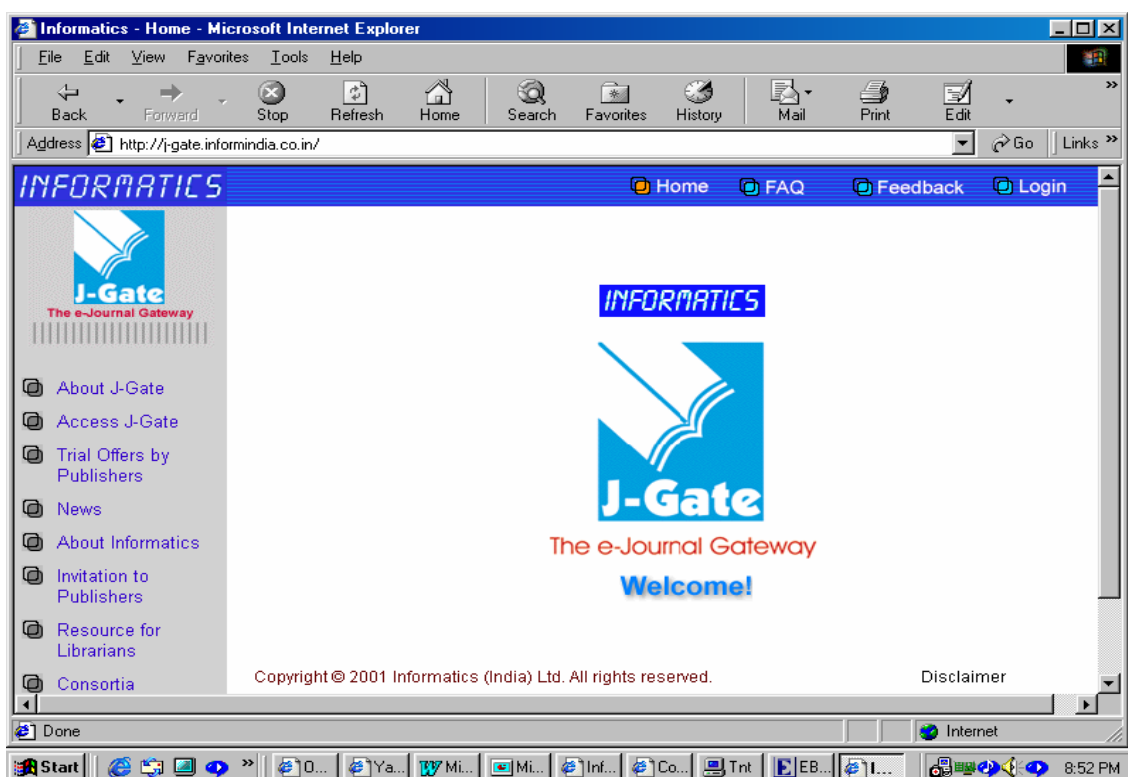
i. JSTOR

JSTOR, a Journal Archive Service, has announced the release of its new Ecology & Botany Collection. Developed in cooperation with the Ecological Society of America (ESA), the collection will enable ecologists, conservationists, and scholars in related fields to access research, over the Internet, from journals that go back 130 years and which represent 1,500 cumulative years of academic literature in the biological sciences. The complete back-runs of 29 academic journals critical to researchers in the field have been converted by JSTOR from print to a full-text, searchable electronic format. In addition to serving as an important archive, the Ecology & Botany Collection creates a new level of access to a significant collection of valuable scientific literature. The historical material contained in this collection will be valuable to scholars and professionals around the world.

ii. J-Gate: The Electronic Journal Gateway

J-Gate is an Internet gateway set up by Informatics for integrating e-content and e-commerce for journal literature. It envisages providing seamless access to journal articles at publisher's site, local site of the libraries, or at the J-gate archive, through a table-of-content (TOC) and abstract database as the search and link interface. It will also support online subscription to journals, and other related services. J-Gate is being implemented in a phased manner and when fully operational, it will offer several integrated services related to e-journals. For more details visit the J-Gate site j-gate.informindia.com, or Contact J-Gate team at j-gate@infomindia.com.

iii. Ingenta



Ingenta is the global research gateway serving the online information needs of over 1.4 million visitors a month. It provides a free online search service of published content from reliable research sources and is one of the UK's top 20 Web services. For publishers of professional and academic periodicals and journals, Ingenta provides a suite of services including data conversion, secure online hosting and distribution of this material to subscribers and pay-per-view customers. As well as providing Web users with a broad based article search and delivery service through www.ingenta.com, Ingenta also develops subject-focused websites, built in conjunction with societies, publishers and university presses.

Ingenta library services

For libraries and information professionals, Ingenta provides an extensive range of free and customized online services, from enabling authenticated campus-wide access to subscribed full text articles to creating branded library gateways incorporating management tools and deposit account facilities.

Ingenta.com: *The world's largest website for the search and delivery of research articles, ingenta.com offers you access to article summaries from over 25,000 publications linked to the full text of over 5,200 titles. Full text access is set by each of our 170 publisher partners, but typically individual subscribers, or individuals within institutions that subscribe, can access the full text of their publications for free. Non-subscribers can, for the majority of articles, pay-per-view choosing from electronic, fax or Ariel delivery.*

References

1. World Wide Web Consortium - See: www.w3c.org
2. BIOSIS - See: http://www.biosis.org.uk/free_resources/resource_guide.html
3. See: <http://biology.usgs.gov>
4. See: <http://www.nbii.gov>
5. See: <http://www.itis.usda.gov>
6. See: <http://www.nbii.gov/datainfo/syscollect/tred>
7. See: <http://www.nbii.gov/datainfo/syscollect/drsc/index.html>
8. See: <http://www.fao.org/agrovoc>
9. See: <http://agmarknet.nic.in/agmarknet1.htm>
10. See: <http://icar.org.in/icar4.htm>
11. See: <http://www.cabi-publishing.org/Products/Internet/Index.asp>
12. See: <http://www.animalscience.com/>
13. See: <http://www.nutritiongate.com/>
14. See: <http://www.leisuretourism.com/>
15. See: <http://www.organic-research.com/>
16. See: <http://asae.cabweb.org/>
17. See: <http://www.agbiotechnet.com/>
18. See: <http://www.cabdirect.org/>
19. JSTOR - See: <http://www.jstor.org>
20. J-Gate - See: <http://j-gate.informindia.com>
21. See: <http://www.ingenta.com>

THE ART OF FINDING INFORMATION ON THE WEB

Mr. Santhosh J Eapen

Indian Institute of Spices Research

Calicut – 673 012, Kerala.

The Internet is a tremendously useful information source for biological research. Data is now digitized and distributed from centralized databases. There are online and e-journals for scientific publications. Any research group has a web page giving details on their research activities, publications, software downloads to automated data-processing services.

Search Engines

These are softwares that help us to find our way around the billion or more pages that make up the Web. It is a service that indexes, organizes, and often rates and reviews web sites. There are significant differences in how search engines build their databases and rank sites. Some rely on people to maintain a catalog of web sites or pages. They are called General Topical Search Engines. Some use software called a spider or crawler to identify key information on sites across the Internet. They are known as Keyword Search Engines. Some others combine both types of services.

These differences make some search engines far more useful than others for searching science and technology websites. Key features to look at in a web search engines database building and indexing strategies are free URL submission, full-text indexing, automated, comprehensive web crawling, a fast 'refresh' rate and a sensible ranking strategy for results.

AltaVista: You can search for keywords or type your query in the form of a question. Other features include a facility to translate sites in several languages; photo and media finder which searches for images; and family filter which helps you to limit the types of results produced by the search.

Lycos: It is a search engine with a directory, organized by subject.

Google: Google is one of the best search engines, which is extremely comprehensive. Pages are ranked based on how many times they are linked from other pages. Google catches copies of web pages, so pages can be accessible even if the server is offline. It returns only pages that contain all the relevant search terms.

HotBot: For a new user HotBot is probably the best search engine. It is also relatively comprehensive, regularly updated and offers form-based query tools.

Excite: A hybrid search engine. It has the traditional search capacity because of using a software program but also has a directory.

Yahoo: This is actually a directory rather than a search engine, meaning human compile and categorize the sites it searches.

If you are looking for a broad, common topic use a search engine that tends to give you fewer results of higher quality (usually the type that relies on people to catalog sites) eg. Yahoo.

If you are looking for a rare topic a site that may be less discriminating but yields more results (usually the type compiled by a software program) eg. Excite, Google.

As scientists we should know how to structure a query to weed out the majority of the junk that will come up in a search. Each search engine makes different default assumption when we enter a few keywords into its little box. If your first search doesn't produce the results you want, try searching with at least one more search engine. Many search engines have a link that leads to detailed information about how it compiles and searches through information, and how to get the best results from it. Reading this page can save you a lot of time.

Effective searching is done when most of the planning, analyzing and keyword refining is done away from the computer. In order to find information effectively, you should plan your search by creating word webs or concept maps or idea maps on the topic. Combine these concepts or ideas depending on the kind of search engine that you will be using. Include distinctive names, acronyms, and abbreviations within your strategies. Include synonyms and variant spellings within your strategies.

Topical Search

With this kind of search, you begin with a broad topic, then you narrow it down. Eg. Science-Biology-Molecular Biology. This search is used with topical search engines such as Yahoo.

Keyword Search (boolean searching)

In order to search effectively, you need to use Boolean logic, which is an extremely simple way of stating how a group of things should be divided or combined into sets. Boolean queries restrict the results that are returned from a database by joining a series of search terms with the operators AND, OR and NOT; and mathematical signs + or -. The word **and** and the sign + will connect your concepts together by requiring that both of these words appear in the results. The word **or** will allow either word to appear in the results. The words **and not** and the minus sign (-) will exclude words to appear in the results. The word **adj** means "adjacent" and will generate results showing where those two words are next to each other in any order. The word **near** will generate results showing where those two words are close to each other. Quotation marks will generate results showing those web pages that show the exact phrasing contained within the quotation marks. If you want to broaden your search or you are unsure of the spelling for a particular word, many search engines will allow by suffixing the word with signs like *, ? and \$.

Joining two keywords with AND finds documents that contain only *keyword 1* and *keyword 2*; using OR finds documents that contain either *keyword 1* or *keyword 2* (or both); and using NOT finds documents that contain *keyword 1* but not *keyword 2*. However, search engines differ in how they interpret a space or an implied operator. Some consider a space an OR, while others defaults to AND. Therefore, find out how the search engine you are using works before you formulate your query.

Boolean queries are read from left to right, just like text. Parentheses can structure more complex boolean queries. If you look for documents that contain *keyword 1* and one of either *keyword 2* or *keyword 3*, but not *keyword 4*, the query should be like : (*keyword 1* AND (*keyword 2* OR *keyword 3*)) NOT *keyword 4*. By using quotation marks you can specify a phrase in many search engines.

Finding Scientific Articles

Print journals in the sciences are increasingly publishing their content in an electronic format in addition to hardcopy. Almost all journals have their own website, most of which are accessible only to subscribers. But abstracts and table of contents usually are free. E-journals have no print counterpart.

There are several scientific literature databases available on the Net, but most of them are not free. Medline database is an excellent resource for searching scientific literature in the biological sciences. It is a free server sponsored by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine.

Technical reports of several research groups are available on their web sites. These are not peer reviewed or published elsewhere, and your only clue to their quality is the reputation and expertise of the authors.

PubMed (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>) is one of the most valuable web resources available to biologists. It indexes over 4000 journals in cell and molecular biology, biochemistry, genetics and related fields. It uses a keyword-based search strategy and allows boolean operators in query statements. Users can specify which database fields to check for each search term by following the search term with a field name enclosed in square brackets. In addition, users can search PubMed using Medical Subjects Heading (MeSH) terms. MeSH is a library of standardized terms that may help manuscripts that use alternate terms to refer to the same concept. PubMed again allows you to use a web interface to narrow your search. By using the Limits form below the query box you can limit your search to particular fields in the PubMed database, such as the Author Name or Substance Name field or language, content and date. The Preview/Index menu allows you to build a detailed query interactively. You can also use the options in the History form to access results from earlier searches and to narrow a search by adding new terms to the query. If you want to collect results from multiple queries and save them into one big file, the Clipboard will help you. If you find a search strategy that works for you, you can save that strategy in the form of a URL, and repeat the same search at any time. The URL has to be bookmarked for this.

Public Biological Databases

Storage of macromolecular data like a gene, protein, an enzyme or a functional RNA in electronic databases is a complicated task. At present each entry into the database is given a serial number and then stored in a relational database that

knows the proper linkages between that serial number and all information pertaining to that entry. DNA sequence, protein sequence and protein structure, though well related are currently maintained as separate database projects and in unconnected data formats, because biological sequence and structure databases have evolved in parallel.

The first molecular biology database was the Protein Data Bank (PDB), the central repository for X-ray crystal structures of protein molecules (at the Brookhaven National Laboratory in 1971). From 15 sets of co-ordinates in 1973, now it has about 14,000 entries. Journals that publish crystallographic results now require submission to the PDB as a condition of publication.

Sequence databases generally specialize in one type of sequence data : DNA, RNA or protein. There are major sequence data collections and deposition sites in Europe, Japan and the United States. The first DNA sequence database was the Gene Sequence Database (GSDB) at Los Alamos National Lab in 1979. Now up-to-date gene sequence information is available from GSDB through the National Centre for Genome Resources. The European Molecular Biology Laboratory, the DNA Database of Japan and the National Institutes of Health co-operated to support the database Gene Bank. Various biological database and information sources available on the web are summarized in Table 1.

Table: Major Biological Data and Information Sources

Subject	Source	Link
Biomedical literature	PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
Nucleic acid sequence	GenBank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide
	SRS at EMBL/EBI	http://srs.ebi.ac.uk
Genome sequence	Entrez Genome	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
	TIGR databases	http://www.tigr.org/tdb/
Protein sequence	GenBank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Protein
	SWISS-PROT at ExPASy	http://www.expasy.ch/spro/
	PIR	http://www-nbrf.georgetown.edu
Protein structure	Protein Data Bank	http://www.rcsb.org/pdb/
Entrez Structure DB		
Protein and peptide mass spectroscopy	PROWL	http://prowl.rockefeller.edu
Post-translational modifications	RESID	http://www-nbrf.georgetown.edu/pirwww/search/textresid.html
Biochemical and bio-physical information	ENZYME	http://www.expasy.ch/enzyme/
	BIND	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Structure
Biochemical pathways	PathDB	http://www.ncgr.org/software/pathdb/
	KEGG	http://www.genome.ad.jp/kegg/
	WIT	http://wit.mcs.anl.gov/WIT2/
Microarray	Gene Expression Links	http://industry.ebi.ac.uk/~alan/MicroArray/
2D-PAGE	SWISS-2DPAGE	http://www.expasy.ch/ch2d/ch2d-top.html
Web resources	The EBI Biocatalog	http://www.ebi.ac.uk/biocat/
	IUBio Archive	http://iubio.bio.indiana.edu

Finding Software

Journals like Bioinformatics, Nucleic Acids Research, Journal of Molecular Biology and Protein Science cover new developments in the field of Bioinformatics. Proven software for a particular application can be obtained from reliable web resources that link to computational biology software sites. Most of the major biological databases have listed their softwares eg. PDB, TIGR, NCBI, etc. Many organizations and groups provide web implementations of their software. They are also available as web servers. You can use the web server versions to get you started and understand the inputs, outputs and options for the program. If you are good in programming, you can develop scripts that allow you to hit a web server with multiple requests without entering them manually into a form. By this you will be able to download a local copy of the software and run it on your own machine so that you avoid slow data transfer to and from remote sites.

Judging the Quality of Information

Some simple guidelines can help you to screen the information you find on web. Always approach the information offered on the web with healthy skepticism.

Authority: One of the first things to consider when evaluating a software, data or information formed on the Internet is the source. The authors, their reputation, their affiliation and credentials, expertise etc. are all important in judging a site. Many commercial companies offer a stripped-down version of their software or services to make you buy their more complete package. A site of an individual researcher may not always all inclusive as a publicly funded database does.

Transparency: You should have access to the source code (the human readable version of a computer program) for whatever the web & server is doing. Others may have online help pages and other documentation. Read them and understand the method and its results you use it.

Timeliness: Many sites on the web are not out-of-date sites and so be skeptical. But timeliness is not always an issue with software. If they are no longer supported or code libraries have ceased to be developed then you should be careful.

THE BIOLOGIST IN THE COMPUTER AGE

Santhosh J. Eapen

*Indian Institute of Spices Research
Calicut – 673 012, Kerala*

Biology as a science has been dealing with problems of information management since the 17th century. The cataloguing of species was the preoccupation of biologists for nearly three centuries. Beginning with animals, plants and microscopic life upon the invention of the compound microscope. New forms of life and extinct life forms are still being discovered even today. Taxonomy was the first informatics problem in biology. He solved it with a great innovation, naming species with a series of one-word names of increasing specificity which is now known as the “Tree of Life”.

The biologist of the modern times is overloaded with vast information on individual genes in addition to his regular tasks of indenting for chemicals, recording lab observations, ordering books and searching literature. We realize today, that our knowledge of biology has exploded in such a way, that we need powerful tools to organize that knowledge itself and that the questions we ask of biological systems today may be so complex, that we cannot hope for answers if we limit ourselves to the capabilities of our natural brains.

The Internet is a tremendously useful information source for any biologist. Data is now digitized and distributed from centralized databases. There are online and e-journals for scientific publications. Any research group has a web page giving details on their research activities, publications, software downloads to automated data-processing services. The information available is overwhelming and it is readily available at greater speed than ever. But do we have the skills to find the right information efficiently? We should use the right tools to get all these jobs done in the changing scenario.

Search Engines

These are softwares that help us to find our way around the billion or more pages that make up the Web. It is a service that indexes, organizes, and often rates and reviews web sites. There are significant differences in how search engines build their databases and rank sites. Some rely on people to maintain a catalog of web sites or pages. They are called General Topical Search Engines. Some use software called a spider or crawler to identify key information on sites across the Internet. They are known as Keyword Search Engines. Some others combine both types of services.

These differences make some search engines far more useful than others for searching science and technology websites. Key features to look at in a web search engines database building and indexing strategies are free URL submission, full-text indexing, automated, comprehensive web crawling, a fast ‘refresh’ rate and a sensible ranking strategy for results.

AltaVista: You can search for keywords or type your query in the form of a question. Other features include a facility to translate sites in several languages; photo and media finder which searches for images; and family filter which helps you to limit the types of results produced by the search.

Lycos: It is a search engine with a directory, organized by subject.

Google: Google is one of the best search engines, which is extremely comprehensive. Pages are ranked based on how many times they are linked from other pages. Google catches copies of web pages, so pages can be accessible even if the server is offline. It returns only pages that contain all the relevant search terms.

HotBot: For a new user HotBot is probably the best search engine. It is also relatively comprehensive, regularly updated and offers form-based query tools.

Excite: A hybrid search engine. It has the traditional search capacity because of using a software program but also has a directory.

Yahoo: This is actually a directory rather than a search engine, meaning human compile and categorize the sites it searches.

If you are looking for a broad, common topic use a search engine that tends to give you fewer results of higher quality (usually the type that relies on people to catalog sites) eg. Yahoo.

If you are looking for a rare topic a site that may be less discriminating but yields more results (usually the type compiled by a software program) eg. Excite, Google.

As scientists we should know how to structure a query to weed out the majority of the junk that will come up in a search. Each search engine makes different default assumption when we enter a few keywords into its little box. If your first search doesn't produce the results you want, try searching with at least one more search engine. Many search engines have a link that leads to detailed information about how it compiles and searches through information, and how to get the best results from it. Reading this page can save you a lot of time.

Effective searching is done when most of the planning, analyzing and keyword refining is done away from the computer. In order to find information effectively, you should plan your search by creating word webs or concept maps or idea maps on the topic. Combine these concepts or ideas depending on the kind of search engine that you will be using. Include distinctive names, acronyms, and abbreviations within your strategies. Include synonyms and variant spellings within your strategies.

Topical Search

With this kind of search, you begin with a broad topic, then you narrow it down. Eg. Science-Biology-Molecular Biology. This search is used with topical search engines such as Yahoo.

Keyword Search (boolean searching)

In order to search effectively, you need to use Boolean logic, which is an extremely simple way of stating how a group of things should be divided or combined into sets. Boolean queries restrict the results that are returned from a database by joining a series of search terms with the operators AND, OR and NOT; and mathematical signs + or -. The word **and** and the sign + will connect your concepts together by requiring that both of these words appear in the results. The

word **or** will allow either word to appear in the results. The words **and not** and the minus sign (-) will exclude words to appear in the results. The word **adj** means “adjacent” and will generate results showing where those two words are next to each other in any order. The word **near** will generate results showing where those two words are close to each other. Quotation marks will generate results showing those web pages that show the exact phrasing contained within the quotation marks. If you want to broaden your search or you are unsure of the spelling for a particular word, many search engines will allow by suffixing the word with signs like *, ? and \$.

Joining two keywords with AND finds documents that contain only *keyword 1* and *keyword 2*; using OR finds documents that contain either *keyword 1* or *keyword 2* (or both); and using NOT finds documents that contain *keyword 1* but not *keyword 2*. However, search engines differ in how they interpret a space or an implied operator. Some consider a space an OR, while others defaults to AND. Therefore, find out how the search engine you are using works before you formulate your query.

Boolean queries are read from left to right, just like text. Parentheses can structure more complex boolean queries. If you look for documents that contain *keyword 1* and one of either *keyword 2* or *keyword 3*, but not *keyword 4*, the query should be like : (*keyword 1* AND (*keyword 2* OR *keyword 3*)) NOT *keyword 4*. By using quotation marks you can specify a phrase in many search engines.

BIOINFORMATICS

The science of informatics is concerned with the representation, organization, manipulation, distribution, maintenance, and use of information, particularly in digital form. Bioinformatics is a newly emerging interdisciplinary research area which may be defined as the interface between biological and computational sciences. Bioinformatics is a rather young discipline, bridging the life- and computer sciences. The need for this interdisciplinary approach to handling biological knowledge is not trivial. It underscores the radical changes in quantitative and qualitative terms, that the biosciences have seen in the last two decades. This scientific field deals with the computational management of all kinds of biological information, whether it may be about genes and their products, whole organisms or even ecological systems. Most of the bioinformatics

work that is being done deals can be described as analyzing biological data, although a growing number of projects deal with the organization of biological information. As a consequence of the large amount of data produced in the field of molecular biology, most of the current bioinformatics projects deal with structural and functional aspects of genes and proteins. Thus we can identify the two main areas of current activities in bioinformatics: the organization and the analysis of biological data.

Bioinformatics commonly is perceived as the handling of sequence data in large databases. Indeed, algorithms to create, maintain and access the sequence databases are among the most important contributions that bioinformatics has made for the life-sciences. But there is important information other than biological sequences available, like citation databases, and in the end it takes a lot more than collecting and storing information to progress from data to knowledge.

Electronic journals and databases

Print journals in the sciences are increasingly publishing their content in an electronic format in addition to hardcopy. Almost all journals have their own website, most of which are accessible only to subscribers. But abstracts and table of contents usually are free. E-journals have no print counterpart.

There are several scientific literature databases available on the Net, but most of them are not free. Medline database is an excellent resource for searching scientific literature in the biological sciences. It is a free server sponsored by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine.

Technical reports of several research groups are available on their web sites. These are not peer reviewed or published elsewhere, and your only clue to their quality is the reputation and expertise of the authors.

PubMed (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>) is one of the most valuable web resources available to biologists. It indexes over 4000 journals in cell and molecular biology, biochemistry, genetics and related fields. It uses a keyword-based search strategy and allows boolean operators in query statements. Users can specify which database fields to check for each search term by following the search term with a field name enclosed in square brackets. In addition, users can search PubMed using Medical Subjects Heading (MeSH) terms. MeSH is a library of standardized terms that may help manuscripts that use alternate terms to refer to the

same concept. PubMed again allows you to use a web interface to narrow your search. By using the Limits form below the query box you can limit your search to particular fields in the PubMed database, such as the Author Name or Substance Name field or language, content and date. The Preview/Index menu allows you to build a detailed query interactively. You can also use the options in the History form to access results from earlier searches and to narrow a search by adding new terms to the query. If you want to collect results from multiple queries and save them into one big file, the Clipboard will help you. If you find a search strategy that works for you, you can save that strategy in the form of a URL, and repeat the same search at any time. The URL has to be bookmarked for this.

Public Biological Databases

Storage of macromolecular data like a gene, protein, an enzyme or a functional RNA in electronic databases is a complicated task. At present each entry into the database is given a serial number and then stored in a relational database that knows the proper linkages between that serial number and all information pertaining to that entry. DNA sequence, protein sequence and protein structure, though well related are currently maintained as separate database projects and in unconnected data formats, because biological sequence and structure databases have evolved in parallel.

The first molecular biology database was the Protein Data Bank (PDB), the central repository for X-ray crystal structures of protein molecules (at the Brookhaven National Laboratory in 1971). From 15 sets of co-ordinates in 1973, now it has about 14,000 entries. Journals that publish crystallographic results now require submission to the PDB as a condition of publication.

Sequence databases generally specialize in one type of sequence data : DNA, RNA or protein. There are major sequence data collections and deposition sites in Europe, Japan and the United States. The first DNA sequence database was the Gene Sequence Database (GSDB) at Los Alamos National Lab in 1979. Now up-to-date gene sequence information is available from GSDB through the National Centre for Genome Resources. The European Molecular Biology Laboratory, the DNA Database of Japan and the National Institutes of Health co-operated to support the database Gene Bank. Various biological database and information sources available on the web are summarized in Table 1.

Table: Major Biological Data and Information Sources

Subject	Source	Link
Biomedical literature	PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
Nucleic acid sequence	GenBank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide
	SRS at EMBL/EBI	http://srs.ebi.ac.uk
Genome sequence	Entrez Genome	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
	TIGR databases	http://www.tigr.org/tdb/
Protein sequence	GenBank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Protein
	SWISS-PROT at ExPASy	http://www.expasy.ch/spro/
	PIR	http://www-nbrf.georgetown.edu
Protein structure	Protein Data Bank	http://www.rcsb.org/pdb/
Entrez Structure DB Protein and peptide mass spectroscopy	PROWL	http://prowl.rockefeller.edu
Post-translational modifications	RESID	http://www-nbrf.georgetown.edu/pirwww/search/textresid.html
Biochemical and bio- physical information	ENZYME	http://www.expasy.ch/enzyme/
	BIND	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Structure
Biochemical pathways	PathDB	http://www.ncgr.org/software/pathdb/
	KEGG	http://www.genome.ad.jp/kegg/
	WIT	http://wit.mcs.anl.gov/WIT2/
Microarray	Gene Expression Links	http://industry.ebi.ac.uk/~alan/MicroArray/
2D-PAGE	SWISS-2DPAGE	http://www.expasy.ch/ch2d/ch2d-top.html
Web resources	The EBI Biocatalog	http://www.ebi.ac.uk/biocat/

	IUBio Archive	http://iubio.bio.indiana.edu
--	---------------	---

Finding Software

Journals like Bioinformatics, Nucleic Acids Research, Journal of Molecular Biology and Protein Science cover new developments in the field of Bioinformatics. Proven software for a particular application can be obtained from reliable web resources that link to computational biology software sites. Most of the major biological databases have listed their softwares eg. PDB, TIGR, NCBI, etc. Many organizations and groups provide web implementations of their software. They are also available as web servers. You can use the web server versions to get you started and understand the inputs, outputs and options for the program. If you are good in programming, you can develop scripts that allow you to hit a web server with multiple requests without entering them manually into a form. By this you will be able to download a local copy of the software and run it on your own machine so that you avoid slow data transfer to and from remote sites.

Judging the Quality of Information

Some simple guidelines can help you to screen the information you find on web. Always approach the information offered on the web with healthy skepticism.

Authority: One of the first things to consider when evaluating a software, data or information formed on the Internet is the source. The authors, their reputation, their affiliation and credentials, expertise etc. are all important in judging a site. Many commercial companies offer a stripped-down version of their software or services to make you buy their more complete package. A site of an individual researcher may not always all inclusive as a publicly funded database does.

Transparency: You should have access to the source code (the human readable version of a computer program) for whatever the web & server is doing. Others may have online help pages and other documentation. Read them and understand the method and its results you use it.

Timeliness: Many sites on the web are not out-of-date sites and so be skeptical. But timeliness is not always an issue with software. If they are no longer supported or code libraries have ceased to be developed then you should be careful.

Conclusion

It is becoming apparent, that the interplay of signals and pathways in complex biological systems - such as the regulation of transcription, cellular activation, developmental organization or cellular communication - may be so complex and sensitive to slight variations of interaction energies, that it may be impossible in principle for our human brains to understand such phenomena quantitatively. As ever more biological phenomena are being described, our concepts of biological knowledge and understanding will change, and we will need to recruit ever more computer tools to organize such knowledge and to extract and present the relevant information to us in a comprehensible way. The development of concepts and models that integrate such complex knowledge and allow its visualization to make it accessible, is the grand future challenge of bioinformatics.

References

<http://www.med.nyu.edu/rcr/rcr/btr/index.html>

Developing Bioinformatics Computer Skills by Cynthia Gibas and Per Jambeck

Bioinformatics & Internet Glossary

Please get in touch if there are any Bioinformatics terms you would like to have explained or added to this list. More importantly, if you would like to suggest amendments to any of these definitions, please also get in touch.

Accession Number

constant from one database to the next

Algorithm

Any sequence of actions (eg computational steps) that perform a particular task.

APPLET

The name given to a small binary file which is executed on your computer as part of a Web Page loading. Applets are created by JAVA. Applets should not be confused with pages which have JAVASCRIPT in them.

BANDWIDTH

The maximum amount of data that can travel a communications path in a given time, usually measured in seconds. If you think of the communications path as a pipe, then bandwidth represents the width of the pipe that determines how much data can flow through it all at once.

BAUD

A term used to refer to the transmission speed of a modem. The higher the value, the faster the modem. The value is usually measured in Bits Per Second.

BIT

Short for binary digit, a bit is the smallest unit of data a computer can handle. Bits are used in various combinations to represent different kinds of data. Each bit has a value of 0 or 1.

Browser

Program used to access sites on the World Wide Web. Hypertext markup language (HTML) enables browsers to represent a Web page the same way regardless of computer platform.

CGI

Common Gateway Interface. This is a programming standard which defines how a Web Page can interact with a user, i.e. fill out a form. Typically CGI applications are written in either PERL or C, both are computer languages of varying complexity. CGI applications run on the server, not on the visiting client.

Characters and character states

In phylogenetics, characters are homologous features in different organisms. The exact condition of that feature in a particular individual is the character state. For example, the character "hair colour" might have the character states "gold", "red", and "yellow". In molecular biology, the character states can be one of the four nucleotides (A, C, T, G) or one of the 20 amino acids. Some authors define "character" to mean the character state as defined here.

Client

A computer, or the software running on a computer, that interacts with another computer at a remote site (server). Note the differences between client and user.

Cluster analysis

A process of assigning data points (sequences) into groups (clusters), starting from pair-wise distances. Useful for identifying outliers and weak links between groups. Fairly easy to do by hand for small data sets.

COMPUTER VIRUS

A computer program created specifically to invade computers and networks and wreak havoc on them. The mischief caused can be very minor: it may cause a funny image or cryptic message to be displayed on your screen, or it can do some serious damage by altering or even destroying files.

COOKIES

A cookie is a file sent to a web browser by a web server that is used to record one's activities on a website. For instance, when you buy items from a site and place them in a so-called virtual shopping cart, that information is stored in the cookie. When the browser requests additional files, the cookie information is sent back to the server. Cookies can remember other kinds of personal information --your password, so you don't have to re-enter it each time you visit the site; your preferences, so the next time you return to a site, you can be presented with customized information. Some people regard cookies as an invasion of privacy; others think they are a harmless way to make websites more personal.

Most cookies have an expiration date and either reside in your computer's memory until you close your browser or they are saved to your hard drive. By the way, cookies cannot read information stored in your computer.

You can use a text editor to view cookie files. Internet Explorer creates separate files for each cookie and stores them in folders named "Cookies" or "Temporary Internet Files." For Windows users of Navigator, the file is called cookies.txt and is located in the the same folder as Netscape.

Descriptor

Information about a sequence or set of sequences whose scope depends on its placement in a record. A descriptor is placed on a set of sequences to reduce the need to save multiple redundant copies of information.

DNS

Domain Name Server. This is a program running on a server which automatically translates domain names into their correct TCP/IP addresses. A busy DNS server is the most common reason for Navigator's frequent "Unable to find Domain Name or invalid entry in the DNS" error.

Domain name

Refers to one of the levels of organization of the Internet; used to both classify and identify host machines. Top-level domain names indicate the type of site or the country in which the host is located.

Download

To transfer a file from a remote host to a local machine via FTP.

EBI

The European Bioinformatics Institute, an outstation of EMBL in the UK. For more information, see <http://www.ebi.ac.uk/>.

e-mail

Electronic mail. Refers to messages that can be composed on the computer and transmitted via the Internet to a remote location within seconds. [Ant: snail mail, postal mail.]

EMBL

the European Molecular Biology Laboratory in Heidelberg, Germany. For more information, see <http://www.embl-heidelberg.de/>.

EST

Expressed sequence tag. ESTs are usually short (300-500 bp) single reads from mRND (cDNA) which are usually produced in large numbers. They represent a snap-shot of what is expressed in a given tissue, and/or at a given developmental stage. They represent tags (some coding, others not) of expression for a given cDNA library. These records usually are very poor in annotation and have only library and BioSource information. They are represented in a variety of databases, notably DDBJ/EMBL/GenBank, dbEST, and Unigene. For more information, see: <http://www.ncbi.nlm.nih.gov/dbEST/>

E-value

For a given score, the number of hits in a database search that we expect to see by chance with this score or better. The E-value takes into account the size of the database that was searched. The lower the E-value, the more significant the score is. See also [P-value](#).

FAQ

A computer file of frequently asked questions. Exactly what it sounds like: a compiled list of questions and answers intended for new users of a computer-based resource, such as a mailing list or a newsgroup.

Feature

Annotation on a specific location on a given sequence.

Firewall

Refers to the separation of a company or organization's internal network from the public part, if any, of the same network. Intended to prevent unauthorized access to private computer systems. See also [Internet](#) and [Intranet](#).

FTP

File transfer protocol. The method by which files are transferred between hosts.

Gene family

Two or more genes that are related by divergent evolution from a common ancestor, either by speciation or gene duplication.

Gopher

A document delivery system allowing the retrieval and display of text-based files.

Graphical user interface

Software that allows a user to interact via "user-friendly" menu and mouse-drive commands, as is typical of Macintosh and Windows applications, and less common for UNIX applications; as opposed to a "command line interface" of typed or scripted commands.

GSS

Genome survey sequences. This DDBJ/EMBL/GenBank division is similar in nature to the EST division, except that its sequences are genomic in origin, rather than cDNA (mRNA). The GSS division contains (but will not be limited to) the following types of data: random "single-pass read" genome survey sequences; single-pass reads from cosmid/BAC/YAC ends (these could be chromosome specific, but need not be); exon-trapped genomic sequences; Alu PCR sequences. For more information, see <http://www.ncbi.nlm.nih.gov/dbGSS/>.

GUI

Graphical user interface. Refers to software front ends that rely on pictures and icons to direct the interaction of users with the application.

Heuristic algorithm

An economical strategy for deriving a solution to a problem for which an exact solution is computationally impractical or intractable. Consequently, a heuristic approach is not guaranteed to find the optimal or "true" solution.

Hidden Markov model

A kind of formal probabilistic model that is well suited to providing a mathematical framework for profile analysis.

Homologous

In phylogenetics, describing particular features in different individuals that are genetically descended from the same feature in a common ancestor. In molecular biology, often "homologous" simply means similar, regardless of genetic relationship.

Homoplasy

Similarity that has evolved independently and is not indicative of common phylogenetic origin.

Host

Any computer on the Internet that can be addressed directly through a unique IP address.

HTGS (HTG)

High-throughput genome sequences (HTG is the HTGS division in DDBJ/EMBL/GenBank). Various genome sequencing centers worldwide have begun the large-scale sequencing of human and other higher

eukaryotic genomes. The databases have deemed it beneficial to put the unfinished sequences that are the result of such sequencing efforts in a separate division. These unfinished records, in most cases, are notable for important numbers of gaps in the nucleotides, low accuracy, and no annotations on the record. These sequences do not achieve the high standard expected DDBJ/EMBL/GenBank records. For more information, see: <http://www.ncbi.nlm.nih.gov/HTGS/>

HTML

Hypertext markup language. The standard, text-based language used to specify the format of World Wide Web documents. HTML files are translated and rendered through the use of Web browsers.

Hyperlink

A graphic or text within a World Wide Web document that can be selected by means of a mouse. Clicking on a hyperlink transports the user to another part of the same Web page or to another Web page, regardless of location.

Hypertext

Within a Web page, text that is differentiated by colour or by underlining and functions as a hyperlink.

Indel

Acronym for "Insertion of DEletion". Applied to length-variable regions of a multiple alignment when it is not specified whether sequence length differences have been created by insertions or deletions.

Internet

A system of linked computer networks used for the transmission of files and messages between hosts. A network of networks. See also [Intranet](#).

Intranet

Intranets use Internet technology and protocols over a private network. They are often not connected to the Internet or are protected from the Internet by a firewall. See also [Internet](#) and [Firewall](#).

IP address

The unique, numeric address of a computer host on the Internet.

ISDN

An acronym for *Integrated Services Digital Network*. ISDN lines are connections that use ordinary phone lines to transmit digital instead of analog signals, allowing data to be transmitted at a much faster rate than with a traditional [modem](#).

ISP

This is a company through which you can access the Internet.

Iterative search

A search procedure that is repeated, usually with increasing sensitivity in each round. For instance, taking all the significant hits from an initial BLAST search and using each of them as a query for a new round of BLAST searches would be one form of iterative search.

Java

A programming language developed by Sun Microsystems that allows small programs (applets) to be run on any computer. Java applets are typically invoked when a user clicks on a hyperlink on a Web page.

LAN

Local area network. A network that connects computers in a small, defined area, such as the offices in a single wing or a group of buildings.

Linux

A freely available but commercial-strength clone of the UNIX operating system. A godsend for starting bioinformaticians on a budget. Easily installed alongside Windows on a PC, so the same machine can be booted into either Linux or Windows.

MODEM

A device which translates computer signals into acceptable format and then transmits those signals over a phone line to another computer. Bidirectional, capable of sending and receiving data via the telephone line.

Molecular clock

The hypothesis that nucleotide or amino acid substitutions occur at more or less fixed rate over evolutionary time, like the slow ticking of a clock. It has been proposed that given a calibration data and a constant molecular clock, the amount of sequence divergence can be used to calculate the time that has elapsed since two molecules diverged.

Multiple alignment

An alignment of three or more sequences, with gaps (spaces) inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column of the multiple alignment.

Mutation studies

In Sequin, a set of sequences for the same gene in the same species, perhaps the same individual, in which several different induced mutations are isolated and sequenced.

NCBI

the US National Center for Biotechnology Information

NIH

the US National Institute of Health

Orthologous (also Orthologus)

Homologous sequences are said to be orthologous when they are direct descendants of a sequence in the common ancestor (ie without having undergone a gene duplication event). See also [Homologous](#) and [Paralogous](#).

PAM matrix

PAM (percent accepted mutation) and BLOSUM (blocks substitution matrix) are matrices that define scores for each of the 210 possible amino acid substitutions. The scores are based on empirical substitution frequencies observed in alignments of database sequences and in general reflect similar physicochemical properties (eg a substitution of leucine for isoleucine, two amino acids of similar hydrophobicity and size, will score higher than a substitution of leucine for glutamate),

Paralogous (also Paralogus)

Two homologous sequences (eg sequences that share a common evolutionary ancestor) that diverged by gene duplication, as opposed to orthologs, which diverged by speciation. A gene family within a single organism is necessarily composed only of paralogs (barring horizontal transmission of genes from another species). Homologous sequences in two organisms A and B that are descendants of two different copies of a sequence that was created by a duplication event in the genome of the common ancestor. Paralogs diverged by gene duplication, as opposed to orthologs, which which diverged by speciation. See also [Homologous](#) and [Orthologous](#).

PDF (PORTABLE DOCUMENT FORMAT)

An acronym for *Portable Document Format*. A file type created by Adobe Systems Incorporated that allows fully formatted, high-resolution, postscript documents to be easily transmitted across the Internet and viewed on any computer that has Adobe Acrobat Reader software.]

PERL

Another commonly used programming language used to code server based web applications. The most commonly used programming language in Bioinformatics.

Phylogenetic studies

In Sequin, a set of sequences for the same gene in individuals of different species. The presumption is that the individuals cannot interbreed. Sequin does not allow a single organism name, but expects the organism to be encoded in the Definition line. It does, however, present a control for setting the proper genetic codes.

Platform

Properly, the operating system running software on a computer (eg Unix or Windows95). More often used to refer to the type of computer, such as a Macintosh or PC-compatible.

Population studies

In Sequin, a set of sequences for the same gene in individuals of a single species. The presumption is that the individuals can interbreed. Sequin allows entry of a single organism name, through some distinguishing source information, such as strain, clone, or isolate, must be entered for each sequence if the program is to function properly.

Profile

A linear model of the consensus of a multiple alignment For each column of a protein alignment, a profile assigns 20 residue scores (one per amino acid), and one or more gap penalties for insertions of extra residues adjacent to this column or a deletion of the consensus residue at this

column. Profiles are also called "position specific scoring matrices" (PSSMs). Profiles that don't allow insertions and deletions are also called "weight matrices".

Protein description

In a sequence record, used if the protein name is not known.

Protein name

In a sequence record, the preferred field for a Protein feature.

P-value

Like an E-value, but a P-value is the probability of a hit occurring by chance with this score or better, as opposed to the expected number of hits. A P-value has a maximum of 1.0, while an E-value has a maximum of the number of sequences in the database that was searched. For small (significant) P-values, P and E are approximately equal, so the choice of one or the other in a software package is arbitrary. NCBI BLAST 2.0, FASTA, and HMMER report E values. WU-BLAST 2.0 reports P-values. See also [E-value](#).

ROUTER

A piece of hardware or software that connects two or more networks. A router functions as a sorter and interpreter as it looks at addresses and passes bits of information to their proper destination. Software routers are sometimes referred to as [gateways](#).

Server

A computer that processes requests issued from remote locations by client machines.

Site

An individual column of residues in an amino acid or nucleotide alignment. The residues at a site are presumed to be homologous.

SMTP

Simple Mail Transfer Protocol. One of the protocols used by your server to send/receive email

Spam

Postings to newsgroups or mail broadcast to a large number of e-mail accounts that usually are irrelevant or not of interest to the recipients. Analogous to postal junk mail.

STS

Sequences tagged site. STSs are operationally unique sequences that identify the combinations of primer pairs used in PCR assays that generate mapping reagents, each of which maps to a single position within the genome. Variations on this definition are also present in this division. This division of GenBank is intended to facilitate cross-comparison of STSs with sequences in other divisions for the purpose of correlating map positions of anonymous sequences with known genes. For more information, see: <http://www.ncbi.nlm.nih.gov/dbSTS>

TCP/IP

Transmission Control Protocol/Internet Protocol. This term actually describes two software mechanisms used to allow multiple computers to talk to each other in an error free fashion.

Telnet

An Internet protocol or application that allows users to connect to computers at remote locations and use these computers as if they were physically operating the remote hardware.

Training Set

A collection of trusted sequences (amino acid or nucleotide) from which a multiple sequence alignment, profile or profile-HMM is built.

URL

Uniformed resource locator. Used Web browsers, URLs specify both the type of site being accessed (FTP, Gopher, or Web) and the address of the Web site.

UNIX

One of the most popular disk operating systems for the internet hosts.

WAN

An acronym for *Wide Area Network*. A network that connects computers over long distances via telephone lines or satellite links. In a wide area network, the computers are physically and sometimes geographically far apart. See [LAN](#).

WEB PAGE

A web page is a document created with [HTML](#) (*HyperText Markup Language*) that is part of a group of [hypertext](#) documents or resources available on the World Wide Web. Collectively, these documents and resources form what is known as a website.

WEB SITE

The Web Site is a location on the Internet, specifically it refers to the POP location in which it resides. All Web Sites are referenced using a special addressing scheme called a URL. A Web Site can mean a single HTML file put up by an individual (called a personal page or vanity page) or hundreds of files placed on the net by a commercial venture.

World Wide Web

A document delivery system capable of handling non-text-based media of various types.

