COMPUTER PROGRAMMING, MOLECULAR MODELING AND SEQUENCE ANALYSIS

PRACTICAL-I

(DIBL01)

(PG DIPLOMA)

ACHARYA NAGARJUNA UNIVERSITY

CENTRE FOR DISTANCE EDUCATION

NAGARJUNA NAGAR,

GUNTUR

ANDHRA PRADESH

CONTENTS

1. **Structure and organization of Computers**

2. **Familirization with Windows NT & MS Office**

3. **Use of internet, World Wide Web**

4. **Operating systems such as DOS, WINDOWS**

5. **Programming with C**

6. **Search on MEDLINE, CD, and Bibliographic Databases**

7. **Biological Sequence Analysis**

8. **Pairwise Sequence Alignment**

9. **Multiple Sequence Alignment**

   Protein and nucleic acid sequence analysis

10. **EMBOSS**

11. **RASMOL**

12. **3-D Structure prediction and anlysis**

1. STRUCTURE AND ORGANISATION OF COMPUTERS

INTRODUCTION

How different parts of a computer are organized and how various operations are performed between different parts to do a specific task. The internal architecture of computer may differ from system to system, but the basic organization remains the same for all computer systems.

To understand digital signal processing systems, we must understand a little about how computers compute. The modern definition of a *computer* is an electronic device that performs calculations on data, presenting the results to humans or other computers in a variety of (hopefully useful) ways.

The generic computer contains *input* devices (keyboard, mouse, A/D (analog-to-digital) converter, etc.), a *computational unit*, and output devices (monitors, printers, D/A converters). The computational unit is the computer's heart, and usually consists of a *central processing unit* (CPU), a *memory*, and an input/output (I/O) interface. What I/O devices might be present on a given computer vary greatly.

OBJECTIVES

- understand basic organization of computer system
- understand the meaning of Arithmetic Logical Unit, Control Unit and Central Processing Unit
- differentiate between bit , byte and a word
- define computer memory
- differentiate between primary memory and secondary memory
- differentiate between primary storage and secondary storage units
- differentiate between input devices and output devices

BASIC COMPUTER OPERATIONS

A computer performs basically five major operations or functions irrespective of their size and make. These are 1) it accepts data or instructions by way of input, 2) it stores data, 3) it can process data as required by the user, 4) it gives results in the form of output, and 5) it controls all operations inside a computer. We discuss below each of these operations.

1. Input: This is the process of entering data and programs in to the computer system. You should know that computer is an electronic machine like any other machine which takes as inputs raw data and performs some processing giving out processed data. Therefore, the input unit takes data from us to the computer in an organized manner for processing.

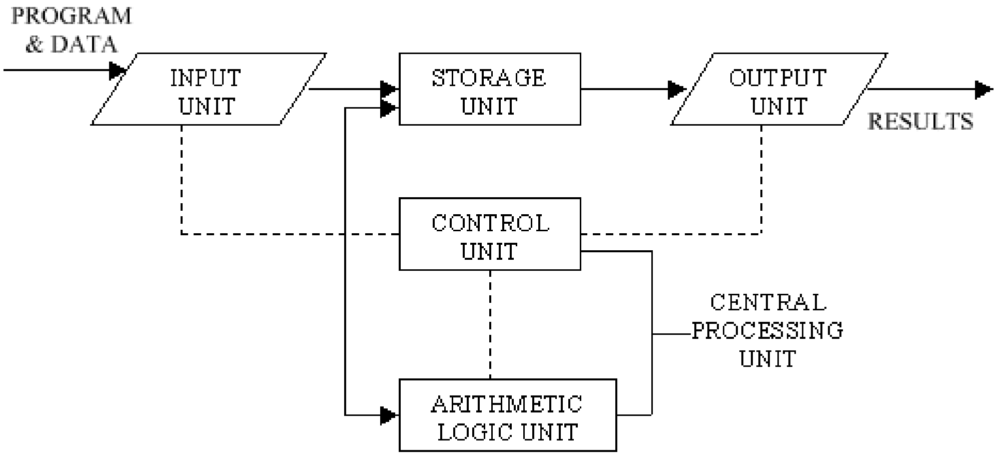


Fig. 2.1 Basic computer Operations

2. Storage: The process of saving data and instructions permanently is known as storage. Data has to be fed into the system before the actual processing starts. It is because the processing speed of Central Processing Unit (CPU) is so fast that the data has to be provided to CPU with the same speed. Therefore the data is first stored in the storage unit for faster access and processing. This storage unit or the primary storage of the computer system is designed to do the above functionality. It provides space for storing data and instructions.

The storage unit performs the following major functions:

- All data and instructions are stored here before and after processing.
- Intermediate results of processing are also stored here.
-

3. Processing: The task of performing operations like arithmetic and logical operations is called processing. The Central Processing Unit (CPU) takes data and instructions from the storage unit and makes all sorts of calculations based on the instructions given and the type of data provided. It is then sent back to the storage unit.

4. Output: This is the process of producing results from the data for getting useful information. Similarly the output produced by the computer after processing must also be kept somewhere inside the computer before being given to you in human readable form. Again the output is also stored inside the computer for further processing.

5. Control: The manner how instructions are executed and the above operations are performed. Controlling of all operations like input, processing and output are performed by control unit. It takes care of step by step processing of all operations in side the computer.

FUNCTIONAL UNITS

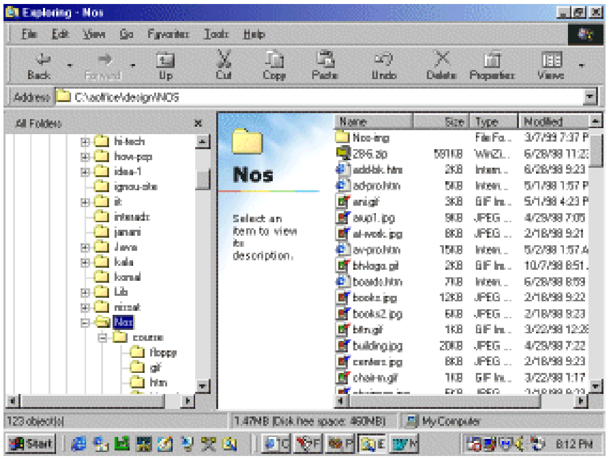
In order to carry out the operations mentioned in the previous section the computer allocates the task between its various functional units. The computer system is divided into three separate units for its operation. They are 1) arithmetic logical unit, 2) control unit, and 3) central processing unit.

Arithmetic Logical Unit (ALU)
After you enter data through the input device it is stored in the primary storage unit. The actual processing of the data and instruction are performed by Arithmetic Logical Unit. The major operations performed by the ALU are addition, subtraction, multiplication, division, logic and comparison. Data is transferred to ALU from storage unit when required. After processing the output is returned back to storage unit for further processing or getting stored.

Control Unit (CU)
The next component of computer is the Control Unit, which acts like the supervisor seeing that things are done in proper fashion. The control unit determines the sequence in which computer programs and instructions are executed. Things like processing of programs stored in the main memory, interpretation of the instructions and issuing of signals for other units of the computer to execute them. It also acts as a switch board operator when several users access the computer simultaneously. Thereby it coordinates the activities of computer's peripheral equipment as they perform the input and output. Therefore it is the manager of all operations mentioned in the previous section.

Central Processing Unit (CPU)
The ALU and the CU of a computer system are jointly known as the central processing unit. You may call CPU as the brain of any computer system. It is just like brain that takes all major decisions, makes all sorts of calculations and directs different parts of the computer functions by activating and controlling the operations.

HARDWARE



Exploring - Nos
File Edit View Go Favorites Tools Help
Back Forward Up Cut Copy Paste Undo Delete Properties Views
Address C:\aoffice\design\NOS
All Folders
hi-tech
how-pop
idea-1
ignou-site
it
intesads
janani
Java
kala
komal
Lib
nirzat
Nos
course
floppy
gif
htm
Nos
Select an item to view its description.
123 object(s)
1.47MB (Disk free space: 460MB)
My Computer
Start
8:12 PM

Name	Size	Type	Modified
Nos-img		File Fo...	3/7/99 7:37 P
286.zip	581KB	WinZi...	6/28/98 11:2
addbk.htm	2KB	Intern...	6/28/98 9:23
adpro.htm	5KB	Intern...	5/1/98 1:57 P
ani.gif	3KB	GIF Im...	5/1/98 4:23 P
aup1.jpg	9KB	JPEG ...	4/29/98 7:05
al-work.jpg	8KB	JPEG ...	2/18/98 9:21
avpro.htm	15KB	Intern...	5/2/98 1:57 A
bhloga.gif	2KB	GIF Im...	10/7/98 8:51
boards.htm	7KB	Intern...	6/28/98 8:59
books.jpg	12KB	JPEG ...	2/18/98 9:22
books2.jpg	6KB	JPEG ...	2/18/98 9:23
btn.gif	1KB	GIF Im...	3/22/98 12:2
building.jpg	20KB	JPEG ...	4/29/98 7:22
centers.jpg	8KB	JPEG ...	2/18/98 9:23
chair-n.gif	1KB	GIF Im...	3/22/98 1:17

SOFTWARE

Computer Architecture

Personal Computer Configuration

Now let us identify the physical components that make the computer work. These are

1. Central Processing Unit (CPU)
2. Computer Memory (RAM and ROM)
3. Data bus
4. Ports
5. Motherboard
6. Hard disk
7. Output Devices
8. Input Devices

All these components are inter-connected for the personal computer to work.

MEMORY SYSTEM IN A COMPUTER

There are two kinds of computer memory: *primary* and *secondary*. Primary memory is accessible directly by the processing unit. RAM is an example of primary memory. As soon as the computer is switched off the contents of the primary memory is lost. You can store and retrieve data much faster with primary memory compared to secondary memory. Secondary memory such as floppy disks, magnetic disk, etc., is located outside the computer. Primary memory is more expensive than secondary memory. Because of this the size of primary memory is less than that of secondary memory.

Computer memory is used to store two things: i) instructions to execute a program and ii) data. When the computer is doing any job, the data that have to be processed are stored in the primary memory. This data may come from an input device like keyboard or from a secondary storage device like a floppy disk.

As program or the set of instructions is kept in primary memory, the computer is able to follow instantly the set of instructions. For example, when you book ticket from railway reservation counter, the computer has to follow the same steps: take the request, check the availability of seats, calculate fare, wait for money to be paid, store the reservation and get the ticket printed out. The program containing these steps is kept in memory of the computer and is followed for each request.

But inside the computer, the steps followed are quite different from what we see on the monitor or screen. In computer's memory both programs and data are stored in the binary form. You have already been introduced with decimal number system, that is the numbers 1 to 9 and 0. The binary system has only two values 0 and 1. These are called *bits*. As human beings we all understand decimal system but the computer can only understand binary system. It is because a large number of integrated circuits inside the computer can be considered as switches, which can be made ON, or OFF. If a switch is ON it is considered 1 and if it is OFF it is 0. A number of switches in different states will give you a message like this: 110101....10. So the computer takes input in the form of 0 and 1 and gives output in the form 0 and 1 only. Is it not absurd if the computer gives outputs as 0's & 1's only? But you do not have to worry about. Every number in binary system can be converted to decimal system and vice versa; for example, 1010 meaning decimal 10. Therefore it is the computer that takes information or data in decimal form from you, convert it in to binary form, process it producing output in binary form and again convert the output to decimal form.

The primary memory as you know in the computer is in the form of IC's (Integrated Circuits). These circuits are called Random Access Memory (RAM). Each of RAM's locations stores one *byte* of information. (One *byte* is equal to *8 bits*). A bit is an acronym for *binary digit*, which stands for one binary piece of information. This can be either 0 or 1. You will know more about RAM later. The Primary or internal storage section is made up of several small storage locations (ICs) called cells. Each of these cells can store a fixed number of bits called *word length***.**

Each cell has a unique number assigned to it called the address of the cell and it is used to identify the cells. The address starts at 0 and goes up to (N-1). You should know that the memory is like a large cabinet containing as many drawers as there are addresses on memory. Each drawer contains a word and the address is written on outside of the drawer.

Capacity of Primary Memory

Each cell of memory contains one character or 1 byte of data. So the capacity is defined in terms of byte or words. Thus 64 kilobyte (KB) memory is capable of storing 64 X 1024 = 32,768 bytes. (1 kilobyte is 1024 bytes). A memory size ranges from few kilobytes in small systems to several thousand kilobytes in large mainframe and super computer. In your personal computer you will find memory capacity in the range of 64 KB, 4 MB, 8 MB and even 16 MB (MB = Million bytes).

The following terms related to memory of a computer are discussed below:

1. **Random Access Memory (RAM):** The primary storage is referred to as random access memory (RAM) because it is possible to randomly select and use any location of the memory directly store and retrieve data. It takes same time to any address of the memory as the first address. It is also called read/write memory. The storage of data and instructions inside the primary storage is temporary. It disappears from RAM as soon as the power to the computer is switched off. The memories, which loose their content on failure of power supply, are known as **volatile** memories .So now we can say that RAM is volatile memory.
2. **Read Only Memory (ROM):** There is another memory in computer, which is called Read Only Memory (ROM). Again it is the ICs inside the PC that form the

ROM. The storage of program and data in the ROM is permanent. The ROM stores some standard processing programs supplied by the manufacturers to operate the personal computer. The ROM can only be read by the CPU but it cannot be changed. The basic input/output program is stored in the ROM that examines and initializes various equipment attached to the PC when the switch is made ON. The memories, which do not loose their content on failure of power supply, are known as **non-volatile** memories. ROM is non-volatile memory.

3. **PROM** There is another type of primary memory in computer, which is called Programmable Read Only Memory (PROM). You know that it is not possible to modify or erase programs stored in ROM, but it is possible for you to store your program in PROM chip. Once the program are written it cannot be changed and remain intact even if power is switched off. Therefore programs or instructions written in PROM or ROM cannot be erased or changed.

4. **EPROM:** This stands for Erasable Programmable Read Only Memory, which over come the problem of PROM & ROM. EPROM chip can be programmed time and again by erasing the information stored earlier in it. Information stored in EPROM exposing the chip for some time ultraviolet light and it erases chip is reprogrammed using a special programming facility. When the EPROM is in use information can only be read.

5. **Cache Memory:** The speed of CPU is extremely high compared to the access time of main memory. Therefore the performance of CPU decreases due to the slow speed of main memory. To decrease the mismatch in operating speed, a small memory chip is attached between CPU and Main memory whose access time is very close to the processing speed of CPU. It is called CACHE memory. CACHE memories are accessed much faster than conventional RAM. It is used to store programs or data currently being executed or temporary data frequently used by the CPU. So each memory makes main memory to be faster and larger than it really is. It is also very expensive to have bigger size of cache memory and its size is normally kept small.

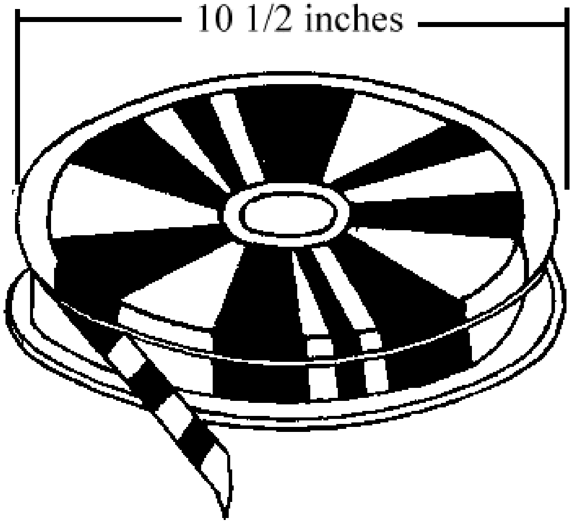
6. **Registers:** The CPU processes data and instructions with high speed, there is also movement of data between various units of computer. It is necessary to transfer the processed data with high speed. So the computer uses a number of special memory units called *registers.* They are not part of the main memory but they store data or information temporarily and pass it on as directed by the control unit.

SECONDARY STORAGE

The operating speed of primary memory or main memory should be as fast as possible to cope up with the CPU speed. These high-speed storage devices are very expensive and hence the cost per bit of storage is also very high. Again the storage capacity of the main memory is also very limited. Often it is necessary to store hundreds of millions of bytes of data for the CPU to process. Therefore additional memory is required in all the computer systems. This memory is called *auxiliary memory* or *secondary storage.*

In this type of memory the cost per bit of storage is low. However, the operating speed is slower than that of the primary storage. Huge volume of data are stored here on permanent basis and transferred to the primary storage as and when required. Most widely used secondary storage devices are *magnetic tapes* and *magnetic disk*.

1. **Magnetic Tape:** Magnetic tapes are used for large computers like mainframe computers where large volume of data is stored for a longer time. In PC tapes can also be used in the form of cassettes. The cost of storing data in tapes is inexpensive. Tapes consist of magnetic materials that store data permanently. It can be 12.5 mm to 25 mm wide plastic film-type and 500 meter to 1200 meter long which is coated with magnetic material. The deck is connected to the central processor and information is fed into or read from the tape through the processor. It similar to cassette tape recorder.



Magnetic Tape

Advantages of Magnetic Tape:

- **Compact:** A 10-inch diameter reel of tape is 2400 feet long and is able to hold 800, 1600 or 6250 characters in each inch of its length. The maximum capacity of such tape is 180 million characters. Thus data are stored much more compactly on tape.
- **Economical**: The cost of storing characters is very less as compared to other storage devices.
- **Fast**: Copying of data is easier and fast.
- **Long term Storage and Re-usability**: Magnetic tapes can be used for long term storage and a tape can be used repeatedly with out loss of data.

2. **Magnetic Disk:** The gramophone record, which is circular like a disk and coated with magnetic material. Magnetic disks used in computer are made on the same principle. It rotates with very high speed inside the computer drive. Data is stored on both the surface of the disk. Magnetic disks are most popular for *direct access* storage device. Each disk consists of a number of invisible *concentric circles* called *tracks*. Information is recorded on tracks of a disk surface in the form of tiny magnetic spots. The presence of a magnetic spot represents *one bit*

and its absence represents zero bit. The information stored in a disk can be read many times without affecting the stored data. So the reading operation is non-destructive. But if you want to write a new data, then the existing data is erased from the disk and new data is recorded.

3. **Floppy Disk:** It is similar to magnetic disk discussed above. They are 5.25 inch or 3.5 inch in diameter. They come in single or double density and recorded on one or both surface of the diskette. The capacity of a 5.25-inch floppy is 1.2 mega bytes whereas for 3.5 inch floppy it is 1.44 mega bytes. It is cheaper than any other storage devices and is portable. The floppy is a low cost device particularly suitable for personal computer system.

Floppy Disk

4. **Optical Disk:**

With every new application and software there is greater demand for memory capacity. It is the necessity to store large volume of data that has led to the development of optical disk storage medium. Optical disks can be divided into the following categories:

1. *Compact Disk/ Read Only Memory (CD-ROM)*: CD-ROM disks are made of reflective metals. CD-ROM is written during the process of manufacturing by high power *laser beam.* Here the storage density is very high, storage cost is very low and access time is relatively fast. Each disk is approximately 4 1/2 inches in diameter and can hold over 600 MB of data. As the CD-ROM can be *read only* we cannot write or make changes into the data contained in it.
2. *Write Once, Read Many (WORM)*: The inconvenience that we can not write any thing in to a CD-ROM is avoided in WORM. A WORM allows the user to write data permanently on to the disk. Once the data is written it can never be erased without physically damaging the disk. Here data can be recorded from keyboard, video scanner, OCR equipment and other devices. The advantage of WORM is that it can store vast amount of data amounting to gigabytes (10^9 bytes). Any document in a WORM can be accessed very fast, say less than 30 seconds.
3. *Erasable Optical Disk*: These are optical disks where data can be written, erased and re-written. This also applies a laser beam to write and re-write the data. These disks may be used as alternatives to traditional disks. Erasable optical disks are based on a technology known as *magnetic optical* (MO). To write a data bit on to the erasable optical disk the MO drive's laser beam heats a tiny, precisely defined point on the disk's surface and magnetises it.

INPUT OUTPUT DEVICES

A computer is only useful when it is able to communicate with the external environment. When you work with the computer you feed your data and instructions through some devices to the computer. These devices are called Input devices. Similarly computer after processing gives output through other devices called output devices.

For a particular application one form of device is more desirable compared to others. Various types of I/O devices are used for different types of applications. They are also known as peripheral devices because they surround the CPU and make a communication between computer and the outer world.

Input Devices

Input devices are necessary to convert our information or data in to a form which can be understood by the computer. A good input device should provide timely, accurate and useful data to the main memory of the computer for processing followings are the most useful input devices.

1. **Keyboard: -** This is the standard input device attached to all computers. The layout of keyboard is just like the traditional typewriter of the type QWERTY. It also contains some extra command keys and function keys. It contains a total of 101 to 104 keys. A typical keyboard used in a computer is shown in the figure. Correct combination of keys to input data have to be pressed. The computer can recognize the electrical signals corresponding to the correct key combination and processing is done accordingly.

2. **Mouse:** - Mouse is an input device shown in Figure that is used with your personal computer. It rolls on a small ball and has two or three buttons on the top. When you roll the mouse across a flat surface the screen censors the mouse in the direction of mouse movement. The cursor moves very fast with mouse giving you more freedom to work in any direction. It is easier and faster to move through a mouse.

3. **Scanner:** The keyboard can input only text through keys provided in it. If we want to input a picture the keyboard cannot do that. Scanner is an optical device that can input any graphical matter and display it back. The common optical scanner devices are Magnetic Ink Character Recognition (MICR), Optical Mark Reader (OMR) and Optical Character Reader (OCR).

   Magnetic Ink Character Recognition (MICR): - This is widely used by banks to process large volumes of cheques and drafts. Cheques are put inside the MICR. As they enter the reading unit the cheques pass through the magnetic field which causes the read head to recognise the character of the cheques.

   Optical Mark Reader (OMR): This technique is used when students have appeared in objective type tests and they had to mark their answer by darkening a square or circular space by pencil. These answer sheets are directly fed to a computer for grading where OMR is used.

Optical Character Recognition (OCR): - This technique unites the direct reading of any printed character. Suppose you have a set of hand written characters on a piece of paper. You put it inside the scanner of the computer. This pattern is compared with a site of patterns stored inside the computer. Whichever pattern is matched is called a character read. Patterns that cannot be identified are rejected. OCRs are expensive though better the MICR.

Output Devices

Visual Display Unit: The most popular input/output device is the Visual Display Unit (VDU). It is also called the monitor. A Keyboard is used to input data and Monitor is used to display the input data and to receive massages from the computer. A monitor has its own box which is separated from the main computer system and is connected to the computer by cable. In some systems it is compact with the system unit. It can be *color* or *monochrome*.

Terminals: It is a very popular interactive input-output unit. It can be divided into two types: hard copy terminals and *soft copy* terminals. A *hard copy* terminal provides a printout on paper whereas soft copy terminals provide visual copy on monitor. A terminal when connected to a CPU sends instructions directly to the computer. Terminals are also classified as dumb terminals or intelligent terminals depending upon the work situation.

Printer: It is an important output device which can be used to get a printed copy of the processed text or result on paper. There are different types of printers that are designed for different types of applications. Depending on their speed and approach of printing, printers are classified as *impact* and *non-impact* printers. Impact printers use the familiar typewriter approach of hammering a typeface against the paper and inked ribbon. *Dot-matrix printers* are of this type. Non-impact printers do not hit or impact a ribbon to print. They use electro-static chemicals and ink-jet technologies. *Laser printers* and *Ink-jet printers* are of this type. This type of printers can produce color printing and elaborate graphics.

In this lesson we discussed five basic operations that a computer performs. These are input, storage, processing, output and control. A computer accepts data as input, stores it, processes it as the user requires and provides the output in a desired format. The storage unit of a computer is divided into two parts: primary storage and secondary storage. We have discussed the devices used for these two types of storage and their usefulness.

2. Familiarization with Windows NT, MS Office

Introduction

The Microsoft Windows NT operating system was designed and built with fully integrated networking capabilities. These networking capabilities differentiate Windows NT from other operating systems, such as MS-DOS, OS/2, and UNIX, in which network capabilities are installed separately from the core operating system.

This chapter introduces the Windows NT networking architecture. It provides you with descriptions of the following topics.

•The design goals and rationale for the Windows NT operating system.

•The basic components of the Windows NT operating system architecture.

•The basics of networking architecture in general. This includes a detailed description of the model on which Windows NT was designed, as well as the industry standards and specifications.

•The Windows NT vertical layers and the interfaces for communication between layers.

•The Windows NT network protocols, which enable layers on two different computers to communicate with each other.

•Distributed processing of applications across the network and the mechanisms Windows NT uses to create connections between servers and workstations.

•The mechanisms for sharing resources across the network, including Multiple Universal Naming Convention Provider (MUP) and Multi-Provider Router (MPR).

•The workstation and server services.

•How binding options work, enabling communications between network layers.

•How Remote Access Service (RAS) works to connect remote or mobile clients to corporate networks.

•How Services for Macintosh are built into Windows NT, allowing Apple Macintosh clients to connect to a Windows NT Server as if it were any other AppleShare server.

Windows NT Operating System Design and Basics

Two primary forces shaped the design of the Windows NT operating system: market requirements and prudent, vigorous design.

Microsoft customers around the world provided the market requirements. Customers wanted the following features.

•Portability across families of processors, such as the Intel x86 line

•Portability across different processor architectures, such as complex instruction set computing (CISC), such as the Intel x86 processors, and reduced instruction set computing (RISC), such as MIPS, DEC, and PowerPC

•Transparent support for single-processor and multiprocessor computers

•Support for distributed computing

•Built-in networking

•Industry standards compliance, such as POSIX

•Certifiable security, such as C2, Functional C2, and E3

Leading-edge thinkers in operating system theory and design developed the design goals, complementing the market requirements. The following features have been built into the Windows NT design.

•*Extensibility*, or modularity of Windows NT. The modular design allows Microsoft to add new modules to all levels of the operating system without compromising its existing stability.
•*Portability*, or the ability of Windows NT to run on both CISC and RISC processors.
•*Scalability*, or the ability to take full advantage of symmetric multiprocessing hardware.
•*Reliability and robustness*, which means that the architecture protects the operating system and its applications from damage. Applications run in their own processes and cannot read or write outside of their own address space. The operating system, in the kernel, is isolated from applications, which interact with the kernel using only well-defined user-mode application programming interfaces (APIs).
•*Performance*, or speed of activity. By running its high-performance subsystems in kernel mode where they interact with the hardware and with each other without thread and process transitions, Windows NT 4.0 improves performance, particularly for graphics-intensive applications, such as Microsoft PowerPoint®, by as much as 20 percent.
•*Compatibility*, which means that Windows NT 4.0 continues to support MS-DOS, OS/2, Windows 3.x, and POSIX applications, as well as the FAT file system and a wide variety of devices and networks.
Windows NT continues to blend together real-world experience in operating systems with some of the best ideas from the computing industry and academia on operating system theory.

Introduction into Microsoft Word

Microsoft Word is a powerful tool to create professional looking documents.

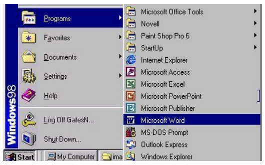
This tutorial will help you get started with Microsoft Word and may solve some of your problems, but it is a very good idea to use the Help Files that come with Microsoft Word , or go to Microsoft's web site located at http://microsoft.com/office/word/default.htm for further assistance.

Starting Microsoft Word

Two Ways

Double click on the Microsoft Word icon on the desktop.

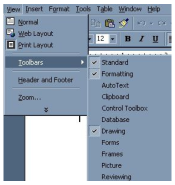
Click on Start --> Programs --> Microsoft Word



Viewing the toolbars

The toolbars in Microsoft Word provide easy access and functionality to the user. There are many shortcuts that can be taken by using the toolbar. First, make sure that the proper toolbars are visible on the screen.

Click **View**
Select **Toolbars**
Select **Standard, Formatting, and Drawing**
Other toolbars can be selected if you wish



Name	**Icon**	**Description**
New Blank Document	🗋	Creates a new, blank file based on the default template.
Open (File menu)	📂	Opens or finds a file.
Save (File menu)	💾	Saves the active file with its current file name, location, and file format.
Mail Recipient	✉	Sends the contents of the document as the body of the e-mail message.
Print (File menu)	🖨	Prints the active file or selected items. To select print options, on the File menu, click Print.
Print Preview (File menu)		Shows how a file will look when you print it.
Spelling and Grammar (Tools menu)	ABC✓	Checks the active document for possible spelling, grammar, and writing style errors, and displays suggestions for correcting them. To set spelling and grammar checking options, click Options on the Tools menu, and then click the Spelling and Grammar tab.
Cut (Edit menu)	✂	Removes the selection from the active document and places it on the Clipboard.
Copy (Edit menu)	⧉	Copies the selection to the Clipboard.
Paste (Edit menu)	📋	Inserts the contents of the Clipboard at the insertion point, and replaces any selection. This command is available only if you have cut or copied an object, text, or contents of a cell.
Format Painter (Standard toolbar)	🖌	Copies the format from a selected object or text and applies it to the object or text you click. To copy the formatting to more than one item, double-click 🖌, and then click each item you want to format. When you are finished, press ESC or click 🖌again to turn off the Format Painter.

Undo (Edit menu)	↶	Reverses the last command or deletes the last entry you typed.
Redo (Edit menu)	↷	Reverses the action of the Undo command.
Hyperlink	🔗	Inserts a new hyperlink or edits the selected hyperlink.
Tables and Borders	⊞	Displays the Tables and Borders toolbar, which contains tools for creating, editing, and sorting a table and for adding or changing borders to selected text, paragraphs, cells, or objects.
Zoom	75%	Enter a magnification between 10 and 400 percent to reduce or enlarge the display of the active document.
Office Assistant	?	The Office Assistant provides Help topics and tips to help you accomplish your tasks.

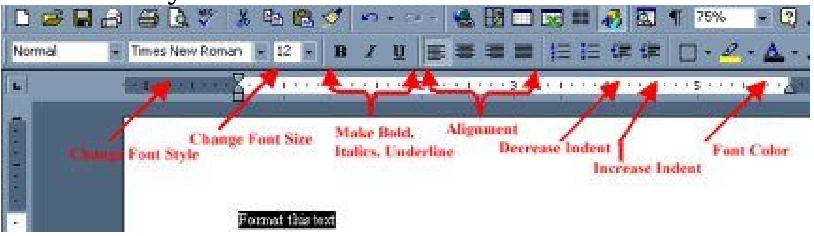
Creating A New Document

Click on File

Select New

To create a blank document, simply select **Blank Document**. To create a document based on one of the templates provided in Microsoft Word, select which one you would like to create and select **OK**

Formatting Text

1. Highlight the text that you want to format by dragging your mouse over while holding down the left mouse button
2. Change the text to your desire
    - 

Inserting a Table

1. Click where you want your table to go
2. Click **Table** at top of screen
3. Select **Insert**
4. Select Table
5. Give your table dimensions

Inserting a Picture

1. Click where you want your picture to go
2. Click **Insert** at top of screen
3. Select **Picture**
4. Select **Clip Art or From File**
5. Select picture and click **Insert**

Inserting Page Numbers and Date/Time

1. Click **Insert** at top of screen
2. Select **Page Numbers and/or Date & Time**

Spell Checking Your Document

Click **Tools** at top of screen
Select **Spelling and Grammar**

Introduction into Microsoft PowerPoint

Microsoft PowerPoint is a powerful tool to create professional looking presentations and slide shows. PowerPoint allows you to construct presentations from scratch or by using the easy to use wizard.

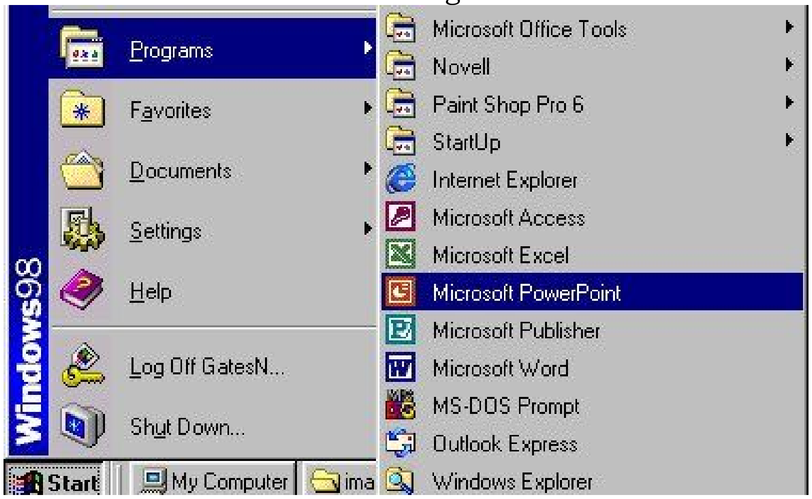
This tutorial will help you get started with Microsoft PowerPoint and may solve some of your problems, but it is a very good idea to use the Help Files that come with Microsoft PowerPoint, or go to Microsoft's web site located at http://microsoft.com/ office/powerpoint/default.htm for further assistance.

Starting Microsoft PowerPoint

Two Ways

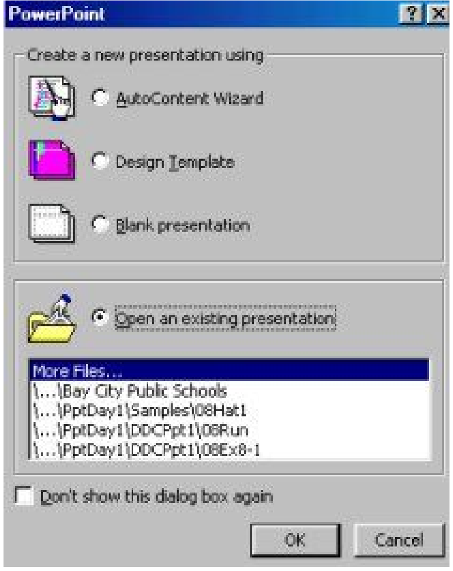
Double click on the Microsoft PowerPoint icon on the desktop.

Click on Start --> Programs --> Microsoft PowerPoint



Creating & Opening a Presentation

After you open up Microsoft PowerPoint, a screen pops up asking if you would like to create a New Presentation or Open An Existing Presentation.


PowerPoint

Create a new presentation using

○ AutoContent Wizard

○ Design Template

○ Blank presentation

● Open an existing presentation

More Files...
\...\Bay City Public Schools
\...\PptDay1\Samples\08Hat1
\...\PptDay1\DDCPpt1\08Run
\...\PptDay1\DDCPpt1\08Ex8-1

☐ Don't show this dialog box again

AutoContent Wizard

Creates a new presentation by prompting you for information about content, purpose, style, handouts, and output. The new presentation contains sample text that you can replace with your own information. Simply follow the directions and prompts that are given by Microsoft PowerPoint.

Design Template

Creates a new presentation based on one of the PowerPoint design templates supplied by Microsoft. Use what is already supplied by Microsoft PowerPoint and change the information to your own.

Blank Presentation

Creates a new, blank presentation using the default settings for text and colors. Go to next step: Creating A Blank Presentation

Opening An Existing Presentation

Select **Open An Existing Presentation** from the picture above

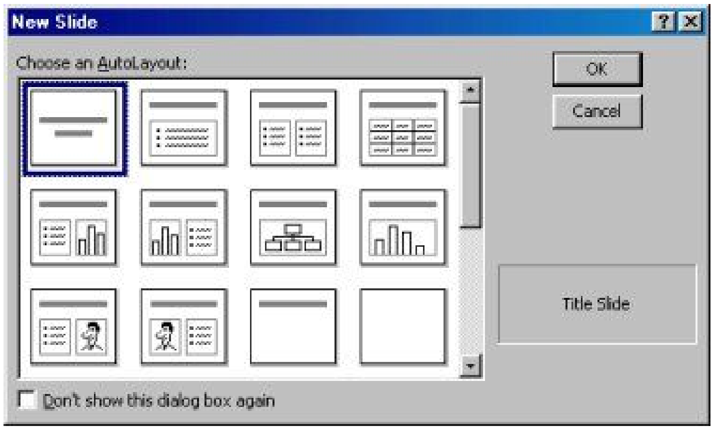
Click on your presentation in the white box below step 1

If you do not see your presentation in the white box, select **More Files** and hit OK.

Locate you existing Presentation and hit the **Open** button

Create a Blank Presentation

After you select Blank Presentation a window pops up asking you to select the layout of the first slide.



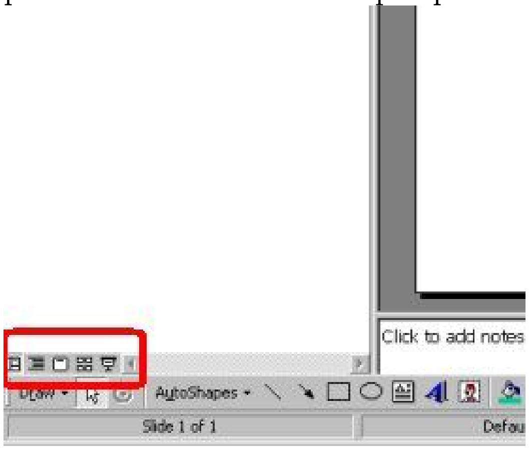
Pre-Designed Slide Layouts (Left to Right)

Title Slide
Bulleted List
Two Column Text
Table
Text & Chart
Chart & Text
Organizational Chart
Chart
Text & Clip Art
Clip Art & Text
Title Only
Blank Slide

NOTE:*If you already know what you want in your next slide, it is a very good idea to choose one of the pre-designed layouts from above. However if you do not, then you can* *still insert what you want in throughout your Presentation anytime you desire. Just choose Blank Slide and insert items as you see fit.*

Different Views That PowerPoint Demonstrates

There are different views within Microsoft PowerPoint that allow you to look at your presentation from different perspectives.



Normal View

Switches to normal view, where you can work on one slide at a time or organize the structure of all the slides in your presentation

Outline View

Switches to outline view, where you can work with the structure of your file in outline form. Work in outline view when you need to organize the structure of your file.

Slide View

Switches to slide view, where you can work on one slide at a time

Slide Sorter View

Displays miniature versions of all slides in a presentation, complete with text and graphics. In slide sorter view, you can reorder slides, add transitions, and animation effects. You can also set the timings for electronic slide shows.

Slide Show View

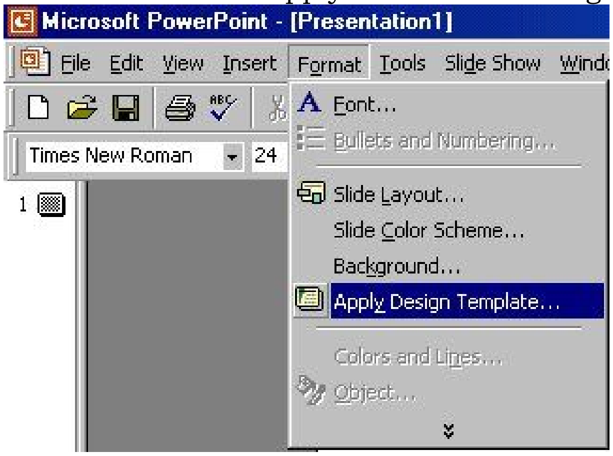
Runs your slide show in a full screen, beginning with the current slide if you are in slide view or the selected slide if you are in slide sorter view. If you simply want to view your show from the first slide:

Click **Slide Show** at the top of the screen
Select **View Show**

Slide Manipulation

- Inserting A New Slide
  1. Click **Insert** at top of screen
  2. Select New Slide

- Formatting A Slide Background
  - You can format your slide to make it look however you would like, whether it be a background color, picture, or a design template built into Microsoft PowerPoint. The next step will show you how to apply a Design Template, but the other items mentioned above can be accomplished the same way.
    1. Click **Format** at the top of the screen
    2. Select Apply Design Template



    3. Select Design you wish to apply
    4. Click Apply Button

- Inserting Clipart & Pictures

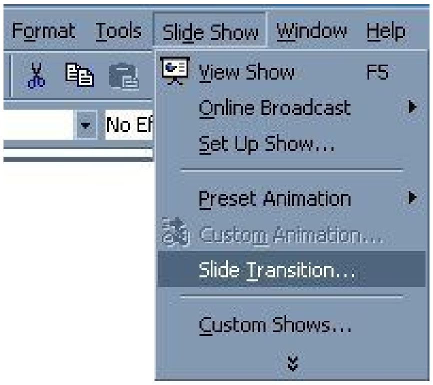
  Display the slide you want to add a picture to.
  Click **Insert** at the top of the screen
  Select **Picture**
  Select **Clip Art**
  Click the category you want
  Click the picture you want
  Click Insert Clip on the shortcut menu
  When you are finished using the Clip Gallery, click the Close button on the Clip Gallery title bar

Steps 1-4 are very similar when inserting other Pictures, Objects, Movies, Sounds, and Charts

Adding Transitions to a Slide Show

You can add customized transitions to your slide show that will make it come alive and become appealing to your audience. Follow these steps when adding Slide Transitions.

1. In **slide or slide sorter view**, select the slide or slides you want to add a transition to.
2. On the **Slide Show** menu at the top of the screen, click **Slide Transition**



3. In the Effect box, click the transition you want, and then select any other options you want
4. To apply the transition to the selected slide, click Apply.
5. To apply the transition to all the slides, click Apply to All.
6. Repeat the process for each slide you want to add a transition to.
7. To view the transitions, on the Slide Show menu, click Animation Preview.

Viewing The Slide Show

You can view your slide show by any of the following ways:

1. Click Slide Show at the lower left of the PowerPoint window.
2. On the Slide Show menu, click View Show.
3. On the View menu, click Slide Show.
4. Press F5 on the keyboard

Navigating While In Your Slide Show

- **Forward Navigation**
  - Simply click on the left Mouse Button or hit the Enter Button on your keyboard
- **Reverse Navigation**
  - Hit the Backspace on the keyboard
- **Exiting the show**
  - Hit the Esc Button on the keyboard

Pack up a presentation for use on another computer

1. Open the Presentation you want to pack
2. On the **File** menu, click **Pack and Go**
3. Follow the instructions in the Pack and Go Wizard.

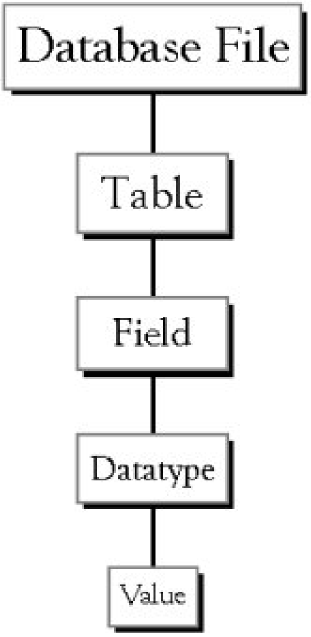
Unpack a presentation to run on another computer

1. Insert the disk or connect to the network location you packed the presentation to
2. In My Computer, go to the location of the packed presentation, and then double-click **Pngsetup**
3. Enter the destination you want to copy the presentation to

Microsoft Access Description

- Microsoft Access is a powerful program to create and manage your databases. It has many built in features to assist you in constructing and viewing your information. Access is much more involved and is a more genuine database application than other programs such as Microsoft Works.
- This tutorial will help you get started with Microsoft Access and may solve some of your problems, but it is a very good idea to use the Help Files that come with Microsoft Access, or go to Microsoft's web site located at http://microsoft.com/office/access/default.htm for further assistance.

  First of all you need to understand how Microsoft Access breaks down a database. Some keywords involved in this process are: *Database File, Table, Record, Field, Data-type*. Here is the Hierarchy that Microsoft Access uses in breaking down a database.



Database File: This is your main file that encompasses the entire database and that is saved to your hard-drive or floppy disk. Example) StudentDatabase.mdb

Table:A table is a collection of data about a specific topic. There can be multiple tables in a database. Example #1) Students Example #2) Teachers

Field: Fields are the different categories within a Table. Tables usually contain multiple fields. Example #1) Student LastName Example #2) Student FirstName

Datatypes: Datatypes are the properties of each field. A field only has 1 datatype. FieldName) Student LastName Datatype) Text

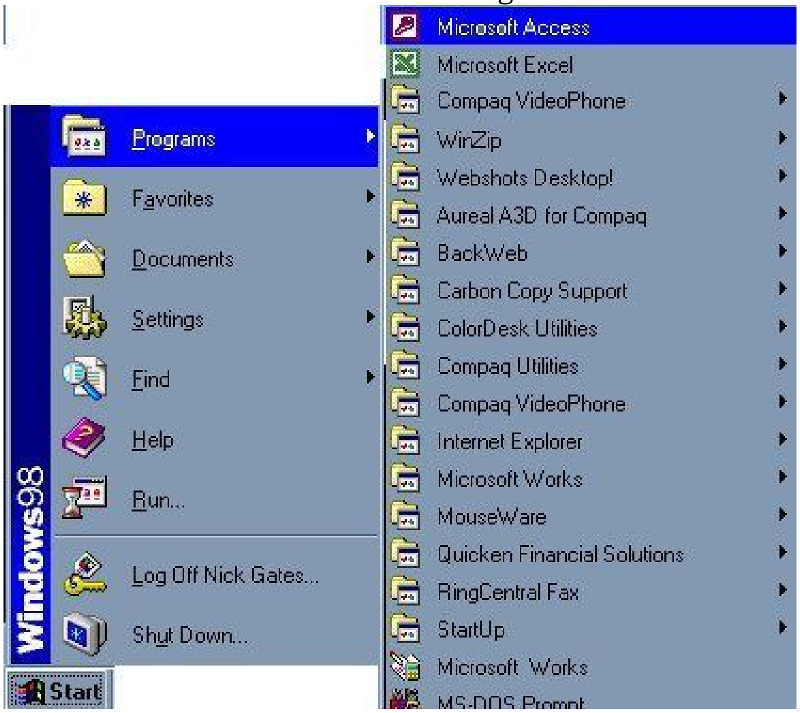
-

This tutorial will help you get started with Microsoft Access and may solve some of your problems, but it is a very good idea to use the Help Files that come with Microsoft Access (or any program you use for that matter), or go to Microsoft's web site located at http://microsoft.com/office/access/default.html for further assistance.

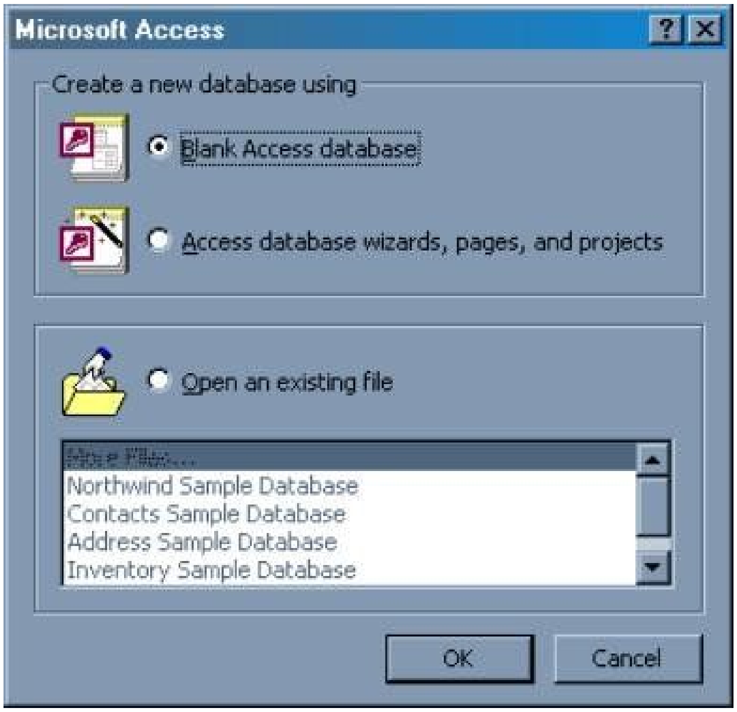
Starting Microsoft Access

- Two Ways
    1. Double click on the Microsoft Access icon on the desktop.

    2. Click on Start --> Programs --> Microsoft Access



Creating New, and Opening Existing Databases



Create a new database using

● Blank Access database

○ Access database wizards, pages, and projects

○ Open an existing file

More Files...
Northwind Sample Database
Contacts Sample Database
Address Sample Database
Inventory Sample Database

The above picture gives you the option to:

- Create a New Database from scratch
- Use the wizard to create a New Database
- Open an existing database
    - The white box gives you the most recent databases you have used. If you do not see the one you had created, choose the More Files option and hit OK. Otherwise choose the database you had previously used and click OK.

Create a database using the Database Wizard

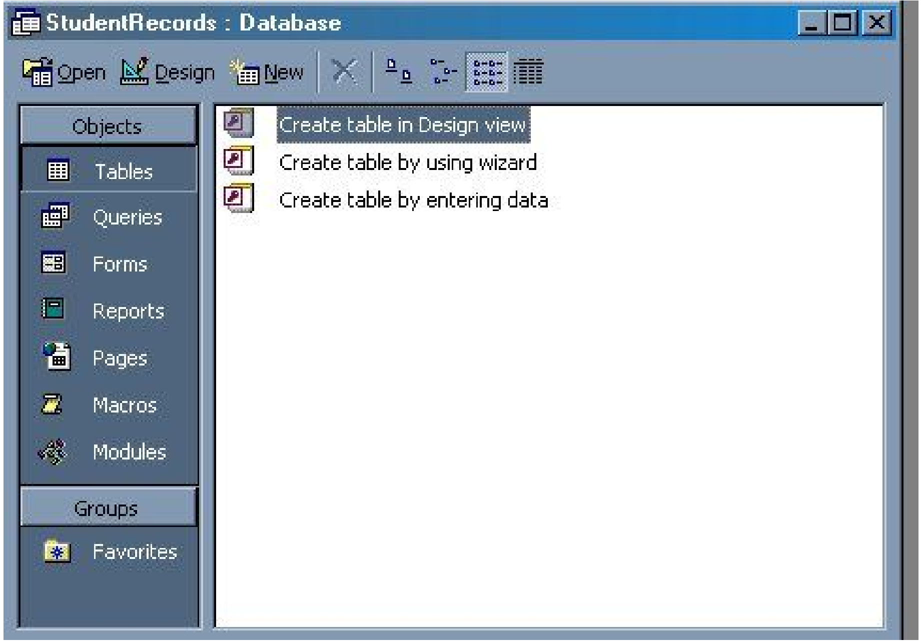
1. When Microsoft Access first starts up, a dialog box is automatically displayed with options to create a new database or open an existing one. If this dialog box is displayed, click **Access Database Wizards, pages, and projects** and then click **OK.**

   If you have already opened a database or closed the dialog box that displays when Microsoft Access starts up, click **New Database** on the toolbar.
2. On the **Databases** tab, double-click the icon for the kind of database you want to create.
3. Specify a name and location for the database.
4. Click **Create** to start defining your new database

Create a database without using the Database Wizard

1. When Microsoft Access first starts up, a dialog box is automatically displayed with options to create a new database or open an existing one. If this dialog box is displayed, click **Blank Access Database**, and then click **OK**.

2. If you have already opened a database or closed the dialog box that displays when Microsoft Access starts up, click **New Database** on the toolbar, and then double-click the **Blank Database** icon on the **General** tab.
3. Specify a name and location for the database and click **Create**. (Below is the screen that shows up following this step)

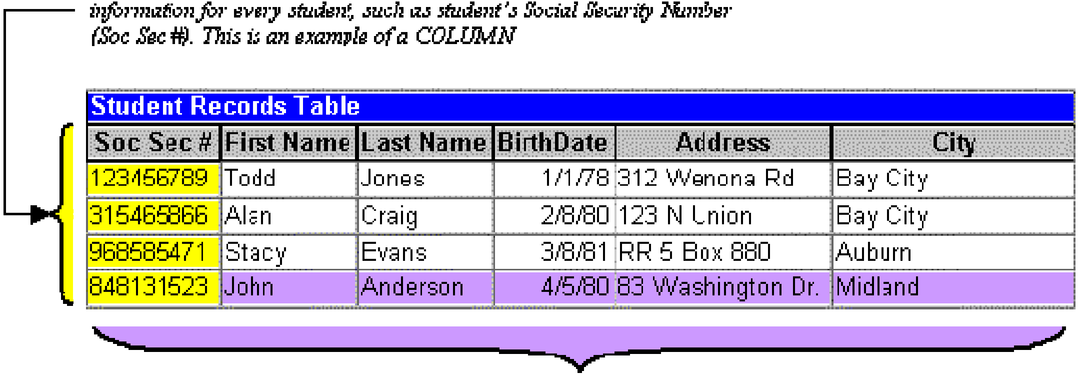

StudentRecords : Database
Open
Design
New
Objects
Tables
Queries
Forms
Reports
Pages
Macros
Modules
Groups
Favorites
Create table in Design view
Create table by using wizard
Create table by entering data

Tables

A table is a collection of data about a specific topic, such as students or contacts. Using a separate table for each topic means that you store that data only once, which makes your database more efficient, and reduces data-entry errors.

Tables organize data into columns (called **fields**) and rows (called **records**).

Each field in the Student Records table contains the same type of information for every student, such as student's Social Security Number (Soc Sec #). This is an example of a COLUMN



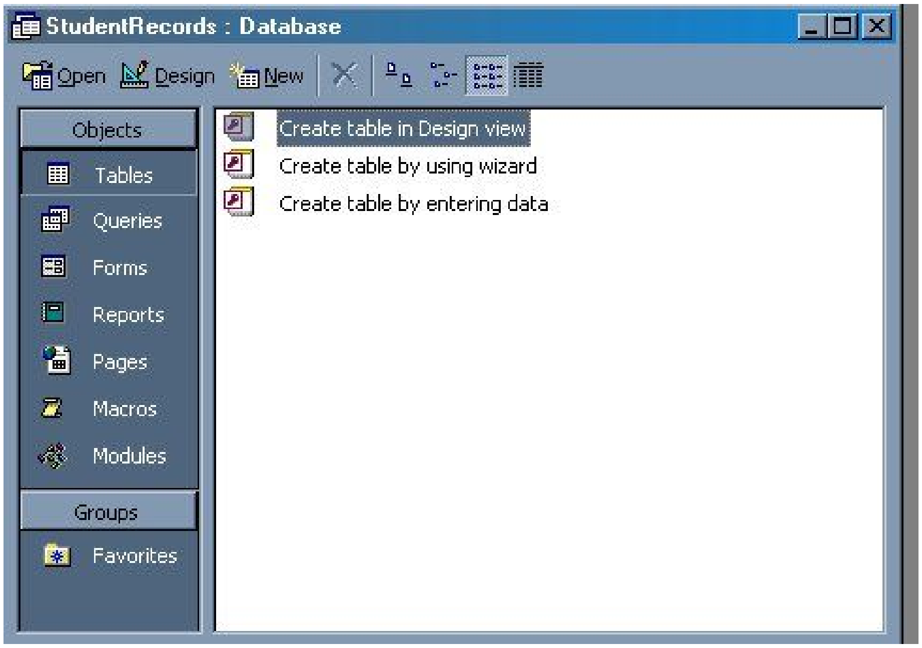
Student Records Table

Soc Sec #	First Name	Last Name	BirthDate	Address	City
123456789	Todd	Jones	1/1/78	312 Wenona Rd	Bay City
315465866	Alan	Craig	2/8/80	123 N Union	Bay City
968585471	Stacy	Evans	3/8/81	RR 5 Box 880	Auburn
848131523	John	Anderson	4/5/80	83 Washington Dr.	Midland

Each record in a Student Records table contains all of the information about one student, such as their First Name, Last Name, Birthday, Address, and City, etc... This is an example of a ROW.

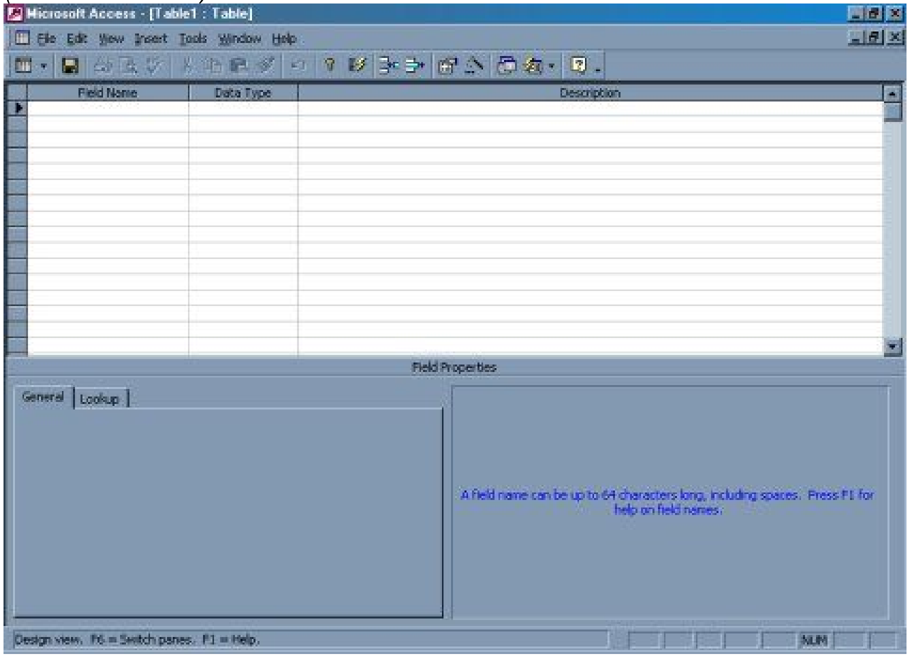
Create a Table from scratch in Design view

If you haven't already done so, switch to the Database Window You can press F11 to switch to the Database window from any other window.


StudentRecords : Database
Open
Design
New
Objects
Tables
Queries
Forms
Reports
Pages
Macros
Modules
Groups
Favorites
Create table in Design view
Create table by using wizard
Create table by entering data

Double-Click on "**Create table in Design view**".

(DESIGN VIEW)



A field name can be up to 64 characters long, including spaces. Press F1 for help on field names.

Design view. F6 = Switch panes. F1 = Help.

Define each of the fields in your table.

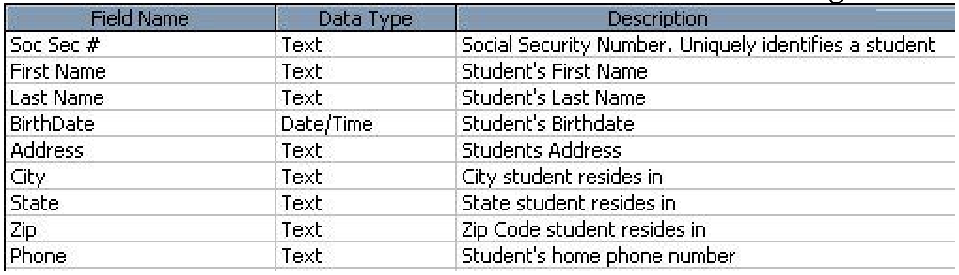
Under the Field Name column, enter the categories of your table.

Under Data Type column, enter the type you want for you categories.

The attribute of a variable or field that determines what kind of data it can hold. For example, in a Microsoft Access database, the Text and Memo field data types allow the field to store either text or numbers, but the Number data type will allow the field to store numbers only. Number data type fields store numerical data that will be used in mathematical calculations. Use the Currency data type to display or calculate currency values. Other data types are Date/Time, Yes/No, Auto Number, and OLE object (Picture).

Under the Description column, enter the text that describes what you field is. (This field is optional).

For our tutorial enter the following items:



Field Name	Data Type	Description
Soc Sec #	Text	Social Security Number. Uniquely identifies a student
First Name	Text	Student's First Name
Last Name	Text	Student's Last Name
BirthDate	Date/Time	Student's Birthdate
Address	Text	Students Address
City	Text	City student resides in
State	Text	State student resides in
Zip	Text	Zip Code student resides in
Phone	Text	Student's home phone number

Primary Key

- One or more fields (columns) whose value or values uniquely identify each record in a table. A primary key does not allow Null values and must always have a unique value. A primary key is used to relate a table to foreign keys in other tables.
- **NOTE:** You do not have to define a primary key, but it's usually a good idea. If you don't define a primary key, Microsoft Access asks you if you would like to create one when you save the table.
- For our tutorial, make the **Soc Sec #** field the primary key, meaning that *every* student has a social security number and no 2 are the same.
    - To do this, simply select the Soc Sec # field and select the primary key button
    - After you do this, Save the table

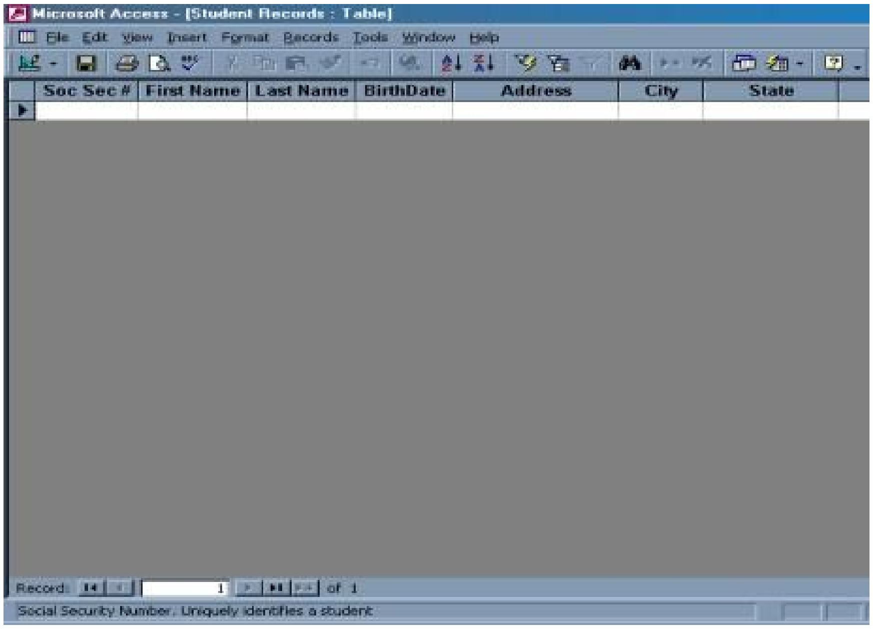
Switching Views

- To switch views form the datasheet (spreadsheet view) and the design view, simply click the button in the top-left hand corner of the Access program.

Datasheet View	Design View
Displays the view, which allows you to enter raw data into your database table.	Displays the view, which allows you to enter fields, data-types, and descriptions into your database table.

Entering Data

- Click on the Datasheet View and simply start "chugging" away by entering the data into each field. **NOTE:** Before starting a new record, the **Soc Sec #** field must have something in it, because it is the Primary Key. If you did not set a Primary Key then it is OK.


Microsoft Access - [Student Records : Table]
File Edit View Insert Format Records Tools Window Help
Soc Sec #
First Name
Last Name
BirthDate
Address
City
State
Record: 1 of 1

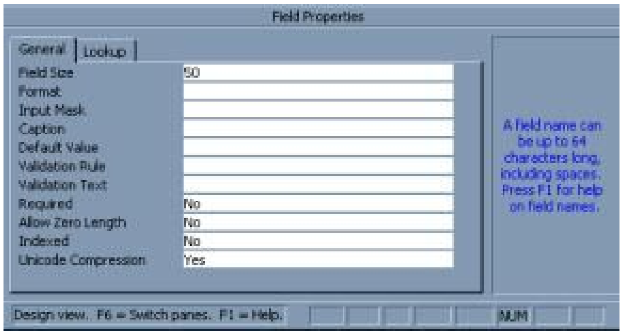
Social Security Number. Uniquely identifies a student

Manipulating Data

- **Adding a new row**
  - o Simply drop down to a new line and enter the information
- **Updating a record**
  - o Simply select the record and field you want to update, and change its data with what you want
- **Deleting a record**
  - o Simply select the entire row and hit the Delete Key on the keyboard

Advanced Table Features w/Microsoft Access

- **Assigning a field a specific set of characters**
  - o Example) Making a Social Security Number only allows 9 characters.
    1. Switch to Design View
    2. Select the field you want to alter
    3. At the bottom select the General Tab



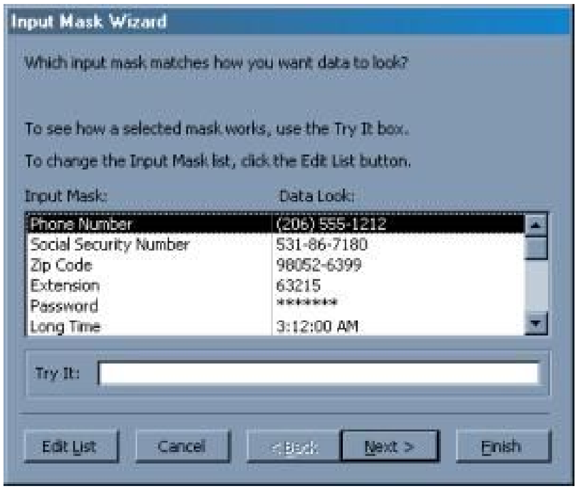
Field Size	50
Format	
Input Mask	
Caption	
Default Value	
Validation Rule	
Validation Text	
Required	No
Allow Zero Length	No
Indexed	No
Unicode Compression	Yes

A field name can be up to 64 characters long, including spaces. Press F1 for help on field names.

Design view. F6 = Switch panes. F1 = Help.

4. Select **Field Size**
5. Enter the number of characters you want this field to have

- **Formatting a field to look a specific way (HINT: You do not need to assign a field a specific set of characters if you do this)**
    - Example) Formatting Phone Number w/ Area Code (xxx) xxx-xxxx
        1. Switch to Design View
        2. Select the field you want to format
        3. At the bottom select the General Tab
        4. Select **Input Mask Box** and click on the **…** button at the right.
        5. Select Phone Number option



Which input mask matches how you want data to look?

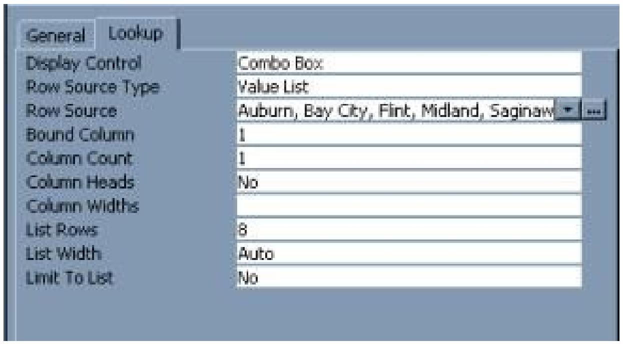
To see how a selected mask works, use the Try It box.

To change the Input Mask list, click the Edit List button.

Input Mask:	Data Look:
Phone Number	(206) 555-1212
Social Security Number	531-86-7180
Zip Code	98052-6399
Extension	63215
Password	********
Long Time	3:12:00 AM

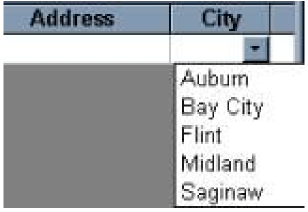
6. Click on Next
7. Leave *!(999) 000-0000 the way it is*. This is a default.
8. Click Next
9. Select which option you want it to look like
10. Click Next
11. Click Finish

- **Selecting a value from a dropdown box with a set of values that you assign to it. This saves you from typing it in each time**
    - Example)Choosing a city that is either Auburn, Bay City, Flint, Midland, or Saginaw
        1. Switch to Design View
        2. Select the field you want to alter (City)
        3. At the bottom select the Lookup Tab
        4. In the **Display Control** box, select **Combo Box**
        5. Under **Row Source Type**, select **Value List**
        6. Under **Row Source**, enter the values how you want them displayed, separated by a comma. (*Auburn, Bay City, Flint, Midland, Saginaw*)
            - **NOTE:**This will not alphabetize them for you, so you will have to do that yourself. It should look something like this:



Display Control	Combo Box
Row Source Type	Value List
Row Source	Auburn, Bay City, Flint, Midland, Saginaw
Bound Column	1
Column Count	1
Column Heads	No
Column Widths	
List Rows	8
List Width	Auto
Limit To List	No

7. Select in the datasheet view and you should see the change when you go to the city field.



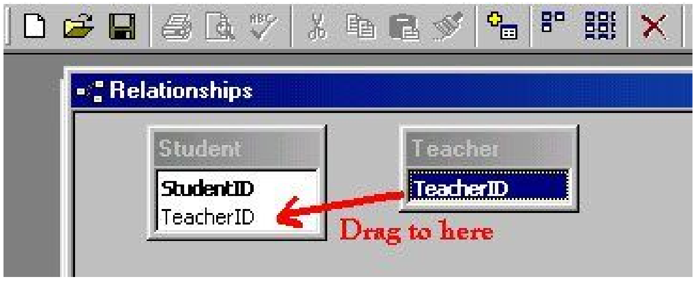
Relationships

After you've set up multiple tables in your Microsoft Access database, you need a way of telling Access how to bring that information back together again. The first step in this process is to define relationships between your tables. After you've done that, you can create queries, forms, and reports to display information from several tables at once.

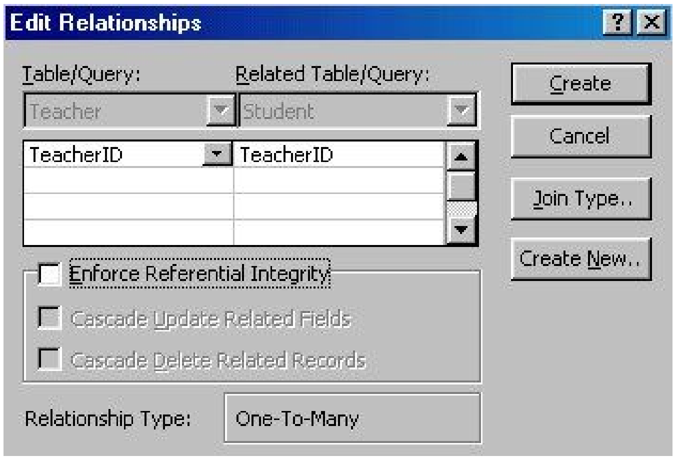
A relationship works by matching data in key fields - usually a field with the same name in both tables. In most cases, these matching fields are the primary key from one table, which provides a unique identifier for each record, and a foreign key in the other table. For example, teachers can be associated with the students they're responsible for by creating a relationship between the teacher's table and the student's table using the TeacherID fields.

Having met the criteria above, follow these steps for creating relationships between tables.

1. In the database window view, at the top, click on Tools ---> Relationships
2. Select the Tables you want to link together, by clicking on them and selecting the Add Button
3. Drag the primary key of the Parent table (Teacher in this case), and drop it into the same field in the Child table (Student in this case.)



4. Select **Enforce Referential Integrity**



Edit Relationships

Table/Query: Teacher

Related Table/Query: Student

Table/Query	Related Table/Query
TeacherID	TeacherID

☐ Enforce Referential Integrity

☐ Cascade Update Related Fields

☐ Cascade Delete Related Records

Relationship Type: One-To-Many

   - When the Cascade Update Related Fields check box is set, changing a primary key value in the primary table automatically updates the matching value in all related records.
   - When the Cascade Delete Related Records check box is set, deleting a record in the primary table deletes any related records in the related table
5. Click Create and Save the Relationship

Forms

A form is nothing more than a graphical representation of a table. You can add, update, delete records in your table by using a form. **NOTE:** Although a form can be named different from a table, they both still manipulate the same information and the same exact data. Hence, if you change a record in a form, it will be changed in the table also.

A form is very good to use when you have numerous fields in a table. This way you can see all the fields in one screen, whereas if you were in the table view (datasheet) you would have to keep scrolling to get the field you desire.

Create a Form using the Wizard

It is a very good idea to create a form using the wizard, unless you are an advanced user and know what you are doing. Microsoft Access does a very good job of creating a form using the wizard. The following steps are needed to create a basic form:

1. Switch to the Database Window. You can do this by pressing F11 on the keyboard.
2. Click on the **Forms** button under **Objects** on the left side of screen
3. Double click on **Create Form Using Wizard**
4. On the next screen select the fields you want to view on your form. Most of the time you would select all of them.
5. Click Next
6. Select the layout you wish
7. Click Next
8. Select the style you desire...**HINT**: if you plan on printing your form, I suggest you use a light background to save on printer toner and ink
9. Click Next

10. Give you form a name, and select **Open the Form and enter information**
11. Select **Finish**
12. You should see your form. To adjust the design of your form, simply hit the design button (same as with the tables), and adjust your form accordingly

Reports

A report is an effective way to present your data in a printed format. Because you have control over the size and appearance of everything on a report, you can display the information the way you want to see it.

Create a Report using the Wizard

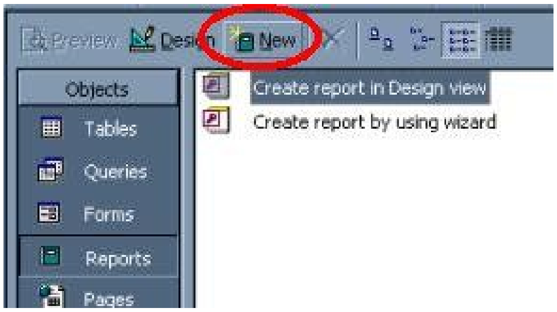
As with the Form, it is a very good idea to create a report using the wizard, unless you are an advanced user. Microsoft Access does a very good job using the wizard to create reports.

1. Switch to the Database Window. You can do this by pressing F11 on the keyboard.
2. Click on the **Reports** button under **Objects** on the left side of screen
3. Double click on **Create Report Using Wizard**
4. On the next screen select the fields you want to view on your form. Most of the time you would select all of them.
5. Click Next
6. Select if you would like to group your files. Keep repeating this step for as many groupings as you would like.
7. Click Next
8. Select the layout and the paper orientation you desire
9. Click Next
10. Select the style you desire...**HINT**: if you plan on printing your report, I suggest you use a light background to save on printer toner and ink
11. Click Next
12. Give you report a name, and select **Preview the Report**
13. Select **Finish**
14. You should see your report. To adjust the design of your report, simply hit the design button (same as with the tables), and adjust your report accordingly

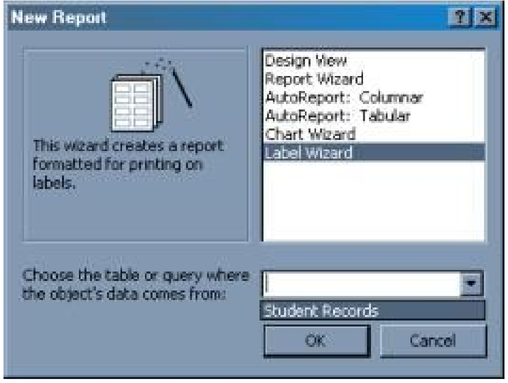
Creating Mail Merge Labels using a Wizard

Microsoft Access lets you create Mailing Labels for your database that you have. To do this do the following:

1. Switch to the Database Window. You can do this by pressing F11 on the keyboard.
2. Click on the **Reports** button under **Objects** on the left side of screen
3. Click on **New**



4. Select **Label Wizard** and the table you would like to get your information from.



This wizard creates a report formatted for printing on labels.

Choose the table or query where the object's data comes from: Student Records

5. Click OK
6. Select the layout of your labels
7. Click Next
8. Select the font size and color you want on each label
9. Click Next
10. Select how you want your label to look
11. Click Next
12. Select how you want your labels sorted

Give your label report a name and preview it.

3. Use of Internet, World Wide Web

Use of Internet

The Internet is unlike all the other communications media anyone has ever encountered. People of all ages, colors, creeds, and countries freely share ideas, stories, data, opinions, and products. Increasingly, news gets out on the Internet before it's available on other media, and the cyber-deprived are losing ground in keeping current on the world's happenings.

Here are some of the ways the Internet is being used:

- **Finding people:** If you've lost track of your childhood sweetheart, now's your chance to find him or her anywhere in the country. You can use one of the directory services to search the phone books of the entire United States.

- **Finding businesses, products, and services:** New yellow page directory services enable you to search by the type of company you're looking for. You can indicate the area code or zip code to help specify the location. People are shopping for that hard-to-find, special gift item.

- **Research:** Law firms are realizing that a great deal of information they formerly paid $600 an hour to find from commercial services can be found for almost nothing when they go directly to the Net. Real estate appraisers use demographic data available on the Net, including unemployment statistics, to help assess property values. Genetics researchers and other scientists download up-to-date research results from around the world. Businesses and potential businesses research their competition over the Internet.
- **Education:** Schoolteachers coordinate projects with classrooms all over the globe. College students and their families exchange e-mail to facilitate letter writing and keep down the cost of phone calls. Students do research from their home computers. The latest encyclopedias are online.
- **Travel:** Cities, towns, states, and countries are using the Web to put up (post) tourist and event information. Travelers find weather information, maps, transportation schedules and tickets, and museum hours online.
- **Marketing and sales:** Software companies are selling software and providing updates via the Net. (The folks making money from the manufacture of floppy disks are looking for new products. Aside from the large pile of AOL disks we now use as coasters, most software distribution is migrating to the Net.) Companies are selling products over the Net. Online bookstores and music stores enable people to browse online, choose titles, and pay for stuff over the Net.
- **Job seaches:** Not just for students, the Internet is an incredible tool for finding a job. It's especially good for students because it provides a powerful, economical way to conduct a real job search. You can publish your résumé online for prospective employers. You can check out the Monsterboard, an impressive compilation of job-related information that enables you to search by discipline (the area of study — all searches need the other kind) or geography or a host of other criteria. You can find the Monsterboard at Monster.com. If you're not just getting out of school, that is, if you're already employed, you might want to use caution when using monster.com. If you register and your employer uses monster.com, your résumé might show up where you least desire it.

- **Love:** People are finding romance on the Net. Singles ads and matchmaking sites vie for users. Contrary to Internet lore, the Net community is no longer just a bunch of socially challenged male nerds under 25.
- **Healing:** Patients and doctors keep up-to-date with the latest medical findings, share treatment experience, and give one another support around medical problems. Some practitioners exchange e-mail directly with their patients.
- **Investing:** People do financial research, buy stock, and invest money. Some companies are online and trade their own shares. Investors are finding new ventures, and new ventures are finding capital.
- **Organizing events:** Conference and trade-show organizers are finding that the best way to disseminate information, call for papers, and do registration is to do it on the Web. Information can be updated regularly, and paper and shipping costs are dramatically reduced. Registering online saves the cost of on-site registration staff and the hassle of on-site registration lines.
- **Nonprofits:** Churches, synagogues, and other community organizations put up pages telling about themselves and inviting new people. The on-line church newsletter always comes before Sunday.

World Wide Web

Introduction

Tools such as Gopher and WWW [tbl94] have been developed largely to facilitate sharing of technical and corporate information at colleges, universities and research institutions. But very quickly, potential educational applications of these technologies became apparent, especially for open and distance learning. In this regard, WWW has more appeal, because of its hypermedia foundation. Its use is growing very rapidly, and it can now be used to access an immense volume of rapidly evolving information. Thus, another strength of WWW is its access to the latest version of a document. Because of all these characteristics, pedagogical uses of the web can evolve along two major axes:

1. use of the technology on a closed corpus of educational material, for the hypermedia and distance delivery capabilities of the web, on one hand, and
2. use of this technology on an organized structure of links for an open corpus of material that was not necessarily meant initially for pedagogical use, but which can be "redirected" and exploited in guided educational explorations.

These two axes are not antagonistic but can be alternatively or complementarily followed.

Long before WWW had reached wide acceptance, the Internet was being used for educational purposes, mostly via mailing lists and bulletin boards. In spite of its tremendous growth, WWW is still only marginally used for teaching and learning activities, probably even more so in Europe, where very few primary or secondary schools have access to the Internet. In this regard, WWW is acting as a driving force, since its ease of use makes the Internet trivially accessible to pupils with no prior knowledge of computers, programming or even networking, making it more tempting for teachers to take advantage of it.

This access by young pupils to the Internet raises some issues, both ethical and technical: can one allow minors free use of a facility that can potentially give them access to pornographic material? Or that makes it very easy for them to pull huge amounts of data accross long distances without their realizing the costs and consequences of their actions? Should one therefore limit their access to the Internet to class work, where a teacher can ensure they are not misusing the facility?

Educationally attractive WWW features

Needless to say that the first features that come to mind are:

- the ability to have multimedia documents [sch94],
- the hypertext/hypermedia capability [bus45],
- WWW networked basis, allowing for distance learning,

Of course, the availability of public-domain client and server software for many different platforms, the relative simplicity of the HTML syntax and the availability of free editing tools are also essential elements to the interest that school teachers find in the WWW.

These features and their attractiveness, for education in general, as well as for distance learning, have already been discussed at length in the past and will therefore not be discussed any further in this paper.

A new feature that has been added recently to HTML is the possibility to include, in a document, input fields that the reader can fill out and the contents of which will be transmitted to the server when the reader pushes a "submit" button. This fill-out form capability has been essential in allowing new uses of the web, where the user can be more active than just clicking on sensitive areas of a document.

One potential difficulty for teachers in using forms to create "interactive" documents is that the use of forms requires some programming on the server to handle the requests (generated by the forms these teachers would like to include in their WWW-based courses) and process the information entered by the user. As a help to people who are not knowledgeable about cgi-bin scripts programming, one can find CGI-compliant programs (e.g. the Zot-Dispatch automatic form handling script generator, available from the University of California, Irvine) that will automatically generate the cgi-bin script appropriate to handle the results sent via a form. Unfortunately, these tools are not always general enough to allow sophisticated handling of the information provided by the user. They typically allow field substitution in a template document that will either be emailed to an address or list of addresses, or sent back to the user, or appended to a file. Some have gone further than this, developing tools to automatically convert applications with GUIs to fill-out forms coupled with cgi-bin programs [thr94].

In an advanced educational setting, one will want more complex handling. For instance, we are currently working on the design of a tool that will allow teachers to not only specify interactively their forms via a regular WWW viewer, but also specify, with a symbolic formalism, free-text answer analysis criteria that will be applied to some user-entered form fields to select appropriate response documents, in addition to modifying these response documents on-the-fly, based on some other user-entered form fields as well. This tool will also generate automatically the cgi-bin scripts that will handle the information sent via the forms, and possibly pass that information via sockets to programs that would be running on the WWW server, a mechanism we have used in a former project, in which we allowed students to control remotely, via WWW, example programs that were described in their *Data Structure* class.

Another, even more recent, feature that is of interest in an educational setting is the ability to run a WWW viewer in "kiosk mode", by which the browsing can be limited to some specific set of documents. This is obviously one way to solve the problem of giving pupils access to the Internet without having to fear questionable uses that could provide arguments to those people who are opposed to giving Internet access to schools under the pretext that it would give minors uncontrolled access to pornographic or violence-related documents, while the school is paying for these resources.

Use of WWW for pedagogical purposes

Just browsing through the web is already an educational experience in itself. Many people have lived this phenomenon by which they start browsing through the web with something specific in mind and end up diverted from their initial goal for a while because they found something interesting on the way, that they were not explicitly looking for. This is what one could call "accidental learning", that is, learning that happens at an unexpected moment, about an unexpected subject.

There are however structured ways [1] in which WWW can be used for education. Two main approaches come to mind: on one hand, using the technology on a closed corpus of educational material, mostly for the hypermedia and distance delivery capabilities of WWW and, on the other hand, using the technology to access, in a structured way, an open corpus of material that was not necessarily meant initially to be used for specific educational purposes.

Closed corpus

In this first category, one can often find "hyper-courses" that take advantage of WWW's hypertext capabilities. This, combined with the fill-out form capability, can allow courseware designers to create educational material that has most of the characteristics of regular courseware that runs on a stand-alone machine, with the additional advantage of being easily accessible from remote locations.

One characteristic of WWW that can appear as a limitation is the fact that the HTTP protocol is stateless, that is, there is no direct relationship of any kind between two consecutive requests to the same server, even if the queries come from the same user. The server treats every request it receives independently from any other request it received in the past or that it will receive in the future. This statelessness allows the HTTP server software to impose very little overhead on the server machine, and keeps the protocol between the client and server very simple.

However, good educational material should take into account the background of each learner to tailor its behavior to the learner's capabilities and past history, and to provide appropriate remediation to learners who experience difficulties with some concepts. To achieve this, the educational software has to keep track of the users' states and actions, which seems to contradict the statelessness of WWW. The solution we developed to overcome this difficulty has been to combine the fill-out form capability with the possibility for the HTTP server to execute external programs or shell scripts in response to a request.

Normally, when the HTTP server executes an external program or shell script, it waits until the program is finished to send the program's output as a virtual HTML document. In our case [ibr94], to maintain a state between consecutive requests, the external shell script spawns a child process that will continue executing during the whole educational sequence. The output of the child process is passed to the HTTP server as a virtual HTML document. The hyper-links embedded in this document contain the process Id of the child process. With this information, the shell script invoked by the HTTP server on a later request can reconnect to the spawned child process and pass it the rest of the request data that will contain information provided by the learner. This mechanism is described in Fig. 1:

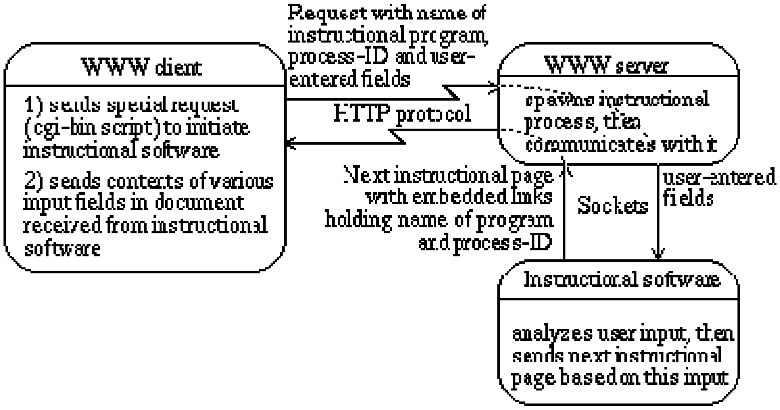

Request with name of instructional program, process-ID and user-entered fields
WWW client
1) sends special request (cgi-bin script) to initiate instructional software
2) sends contents of various input fields in document received from instructional software
HTTP protocol
WWW server
spawns instructional process, then communicates with it
Next instructional page with embedded links holding name of program and process-ID
Sockets
user-entered fields
Instructional software
analyzes user input, then sends next instructional page based on this input

Fig. 1. Interaction scheme between the remote user and some instructional software via WWW.

The approach we have just described has its advantages and its drawbacks. Among the advantages we find the portability achieved by using HTML for the spawned process output. Whatever the machine the user is working on, the viewer should be able to handle HTML documents and thus be able to display the output of the educational process, regardless of the hardware of the machine. There is therefore only one executable version of the courseware, that resides on the server machine. As a consequence, anybody can engage very simply in the educational process, from anywhere in the (Internet) world, at any time, without any bootstrapping procedure. One drawback is that this server machine can become overloaded and thus unable to serve new users. Another drawback is that the user interactions and the program output are limited by the HTML syntax.

Another approach to the closed corpus option is to define a new document type and have a special dedicated viewer handle the "document" sent by the HTTP server. This viewer can be an interpreter for some special purpose language and it will use the document sent by the server as a program to interpret/execute on the client machine, thus removing the load from the server. This external viewer can send a specific system signal to the main WWW viewer to tell it to display a new document, the URL of which has been put in a temporary file. When used in combination with fill-out forms, this allows bidirectional communication (via the WWW server) between the main WWW viewer and the external viewer that are both running on the same client machine. A more effective scheme is currently being proposed as a Client Communication Interface (CCI) protocol, that should allow bidirectional communication between a WWW viewer and an external application, without necessarily transiting throught a WWW server.

The main advantage of this approach is that it can potentially use all the graphical and computational capabilities of the client machine, thus resulting in more powerful user interfaces and faster response times. The drawbacks are that one needs to have a dedicated viewer on the client machine (this viewer has to be downloaded by the users/learners before they can start the educational process, or it has to be preinstalled by the teacher or the local system manager), and that a different executable version of the interpreter has to be produced for every different kind of hardware that the learners can potentially use. Another possible drawback could be related to security issues if the interpreter can potentially execute instructions that would harm the local computer environment (software viruses).

Open corpus

The other approach to educational use of the web is to exploit the enormous amount of information that is accessible via the Internet, whether it has been put there for educational purposes or not. It is the authors' belief that nobody yet has been able to evaluate, even very approximately, the volume of information that is now available to every Internet user. From the statistics of the NSFNet backbone traffic, one could conjecture that this volume is likely to be in the order of magnitude of thousands of billions of characters (terabytes).

Given its gigantic size, this ocean of information has to be harnessed to make it manageable and useful to users and learners. In this regard, WWW appears to be very useful in that its network-based hypertextual capabilities make it very appropriate to organize into a hierarchy this huge volume. CERN [2] has started a very interesting project called the WWW Virtual Library. Its purpose is to create a distributed catalogue of all the Internet resources accessible via WWW. It is maintained on a voluntary basis by people who are specialists in a specific domain and who are willing to share and maintain a document, or a hierarchy of documents, that reflect a certain structure of their domain, giving access to all the major Internet-accessible resources that they are aware of. The subjects of the WWW Virtual Library range over a very wide spectrum, covering most regular subjects of study, as well as recreational subjects or more esoteric ones like "Paranormal Phenomena" or UFOs.

Electronic journals and online magazines also constitute interesting sources of information that are probably going to develop and grow as people get more accustomed to the electronic media.

As useful as they may be for researchers or post-graduate students, these structuring resources are probably not directly usable as is for more basic educational purposes. They are nevertheless a good example of what can be done to present a subject to people who want to learn about it. One can easily imagine developing more elaborate structural documents that will guide learners in their exploration of a subject domain. The educational material would thus consist of a combination of explanatory text, pointers to more in-depth material publicly available on the Internet and, possibly, some features of the closed corpus approach by which the learner could be evaluated and proposed complementary or remedial material.

An interesting pedagogical strategy [wra94] one can use in the *open corpus* approach is to have pupils use WWW in their process of knowledge appropriation [3] by giving them the task to create their own documents that will tie together the information they have collected in a more constructive way than just enumerate pointers to existing documents [4]. These *synthetic* documents (in the sense of documents synthesizing the content of other documents) created by pupils can also help the teacher have a better grasp of the pupils' mental structures. Even more important, the pupils' documents can help the teacher offer appropriate remediation to those pupils who have not been able to go far enough in their discovery of a subject.

4. Operating systems such as DOS, Windows

The software that the rest of the software depends on to make the computer functional. On most PCs this is Windows or the Macintosh OS. Unix and Linux are other operating systems often found in scientific and technical environments.

An operating system (sometimes abbreviated as "OS") is the program that, after being initially loaded into the computer by a boot program, manages all the other programs in a computer. The other programs are called *applications* or application programs. The application programs make use of the operating system by making requests for services through a defined application program interface (API). In addition, users can interact directly with the operating system through a user interface such as a command language or a graphical user interface (GUI).

The most important program that runs on a computer. Every general-purpose computer must have an operating system to run other programs. Operating systems perform basic tasks, such as recognizing input from the keyboard, sending output to the display screen, keeping track of files and directories on the disk, and controlling peripheral devices such as disk drives and printers.
An operating system performs these services for applications:

- In a multitasking operating system where multiple programs can be running at the same time, the operating system determines which applications should run in what order and how much time should be allowed for each application before giving another application a turn.
- It manages the sharing of internal memory among multiple applications.
- It handles input and output to and from attached hardware devices, such as hard disks, printers, and dial-up ports.
- It sends messages to each application or interactive user (or to a system operator) about the status of operation and any errors that may have occurred.
- It can offload the management of what are called *batch* jobs (for example, printing) so that the initiating application is freed from this work.
- On computers that can provide parallel processing, an operating system can manage how to divide the program so that it runs on more than one processor at a time.
-

MS DOS

Microsoft DOS (Disk Operating System) is a command line user interface. MS-DOS 1.0 was released in 1981 for IBM computers and the latest version of MS-DOS is MS-DOS 6.22, which was released in 1994. While MS-DOS is not commonly used by itself today, it still can be accessed from every version of Microsoft Windows by clicking Start / Run and typing "command" or by typing "CMD" in Windows NT, Windows 2000 or Windows XP.

Additional information on other computer Operating Systems can also be found on our Operating Systems page.

HOW TO LOAD

This device driver must be loaded by a device or devicehigh command in your CONFIG.SYS file.

To load this file within Windows 95 / Windows 98 the config.sys you must have:
device=c:\windows\command\ansi.sys
To load this file within Windows 3.x / Windows NT you must have:
device=c:\dos\ansi.sys

Syntax Notes

To be functional, each DOS command must be entered in a particular way: this command entry structure is known as the command's "syntax." The syntax "notation" is a way to reproduce the command syntax in print.

For example, you can determine the items that are optional, by looking for information that is printed inside square brackets. The notation [d:], for example, indicates an optional drive designation. The command syntax, on the other hand, is how YOU enter the command to make it work.

Command Syntax Elements

1. Command Name

The DOS command name is the name you enter to start the DOS program (a few of the DOS commands can be entered using shortcut names). The DOS command name is always entered first. In this book, the command is usually printed in uppercase letters, but you can enter command names as either lowercase or uppercase or a mix of both.

2. Space

Always leave a space after the command name.

3. Drive Designation

The drive designation (abbreviated in this book as "d:") is an option for many DOS commands. However, some commands are not related to disk drives and therefore do not require a drive designation. Whenever you enter a DOS command that deals with disk drives and you are already working in the drive in question, you do not have to enter the drive designator. For example, if you are working in drive A (when the DOS prompt A> is showing at the left side of the screen) and you want to use the DIR command to display a directory listing of that same drive, you do not have to enter the drive designation. If you do not enter a drive designation, DOS always assumes you are referring to the drive you are currently working in (sometimes called the "default" drive).

4. A Colon

When referring to a drive in a DOS command, you must always follow the drive designator with a colon (:) (this is how DOS recognizes it as a drive designation).

5. Pathname

A pathname (path) refers to the path you want DOS to follow in order to act on the DOS command. As described in Chapter 3, it indicates the path from the current directory or subdirectory to the files that are to be acted upon.

6. Filename

A filename is the name of a file stored on disk. As described in Chapter 1, a filename can be of eight or fewer letters or other legal characters.

7. Filename Extension

A filename extension can follow the filename to further identify it. The extension follows a period and can be of three or fewer characters. A filename extension is not required.

8. Switches
Characters shown in a command syntax that are represented by a letter or number and preceded by a forward slash (for example, "/P") are command options (sometimes known as "switches"). Use of these options activate special operations as part of a DOS command's functions.

9. Brackets
Items enclosed in square brackets are optional; in other words, the command will work in its basic form without entering the information contained inside the brackets.

10. Ellipses
Ellipses (...) indicate that an item in a command syntax can be repeated as many times as needed.

11. Vertical Bar
When items are separated by a vertical bar (|), it means that you enter one of the separated items. For example: ON | OFF means that you can enter either ON or OFF, but not both.

MS DOS Commands

APPEND
(External)
APPEND ;
APPEND [d:]path[;][d:]path[...]
APPEND [/X:on|off][/path:on|off] [/E]
Displays or sets the search path for data files. DOS will search the specified path(s) if the file is not found in the current path.

ASSIGN
(External)
ASSIGN x=y [...] /sta
Redirects disk drive requests to a different drive.

ATTRIB
(External)
ATTRIB [d:][path]filename [/S]
ATTRIB [+R|-R] [+A|-A] [+S|-S] [+H|-H] [d:][path]filename [/S]
Sets or displays the read-only, archive, system, and hidden attributes of a file or directory.

BACKUP
(External)
BACKUP d:[path][filename] d:[/S][/M][/A][/F:(size)] [/P][/D:date] [/T:time] [/L:[path]filename]

Makes a backup copy of one or more files. (In DOS Version 6, this program is stored on the DOS supplemental disk.)

BREAK
(Internal)
BREAK =on|off
Used from the DOS prompt or in a batch file or in the CONFIG.SYS file to set (or display) whether or not DOS should check for a Ctrl + Break key combination.

BUFFERS
(Internal)
BUFFERS=(number),(read-ahead number)
Used in the CONFIG.SYS file to set the number of disk buffers (number) that will be available for use during data input. Also used to set a value for the number of sectors to be read in advance (read-ahead) during data input operations.

CALL
(Internal)
CALL [d:][path]batchfilename [options]
Calls another batch file and then returns to current batch file to continue.

CHCP
(Internal)
CHCP (codepage)
Displays the current code page or changes the code page that DOS will use.

CHDIR
(Internal)
CHDIR (CD) [d:]path
CHDIR (CD)[..]
Displays working (current) directory and/or changes to a different directory.

CHKDSK
(External)
CHKDSK [d:][path][filename] [/F][/V]
Checks a disk and provides a file and memory status report.

CHOICE
(Internal)
CHOICE [/C[:]keys] [/N][/S][/T[:]c,nn] [text]
Used to provide a prompt so that a user can make a choice while a batch program is running.

CLS (Clear Screen)

(Internal)

```
CLS
```

Clears (erases) the screen.

COMMAND

(External)

```
COMMAND [d:][path] [device] [/P][/E:(size)] [/MSG][/Y [/C (command)|/K (command)]
```

Starts a new version of the DOS command processor (the program that loads the DOS Internal programs).

COMP

(External)

```
COMP [d:][path][filename] [d:][path][filename] [/A][/C][/D][/L][/N:(number)]
```

Compares two groups of files to find information that does not match. (See FC command).

COPY

(Internal)

```
COPY [/Y|-Y] [/A][/B] [d:][path]filename [/A][/B] [d:][path][filename] [/V]
or
COPY [/Y|-Y][/A][/B] [d:][path]filename+[d:][path]filename[...][d:][path][filename] [/V]
```

Copies and appends files.

COUNTRY

(Internal)

```
COUNTRY=country code,[code page][,][d:][filename]
```

Used in the CONFIG.SYS file to tell DOS to use country-specific text conventions during processing.

CTTY

(Internal)

```
CTTY (device)
```

Changes the standard I/O (Input/Output) device to an auxiliary device.

DATE

(Internal)

```
DATE mm-dd-yy
```

Displays and/or sets the system date.

DBLSPACE

(External)

```
DBLSPACE / automount=drives
DBLSPACE /chkdsk [/F] [d:]
DBLSPACE /compress d: [/newdrive=host:] [/reserve=size] [/F]
DBLSPACE /create d: [/newdrive=host:] [/reserve=size] [/size=size]
DBLSPACE /defragment [d:] ]/F]
```

DBLSPACE /delete d:
DBLSPACE /doubleguard=0|1
DBLSPACE /format d:
DBLSPACE [/info] [d:]
DBLSPACE /list
DBLSPACE /mount[=nnn] host: [/newdrive=d:]
DBLSPACE /ratio[=ratio] [d:] [/all]
DBLSPACE /size[=size] [/reserve=size] d:
DBLSPACE /uncompress d:
DBLSPACE /unmount [d:]
A program available with DOS 6.0 that allows you to compress information on a disk.

DEBUG
(External)
DEBUG [pathname] [parameters]
An MS-DOS utility used to test and edit programs.

DEFRAG
(External)
DEFRAG [d:] [/F][/S[:]order] [/B][/skiphigh [/LCD|/BW|/GO] [/H]
DEFRAG [d:] [/V][/B][/skiphigh] [/LCD]|/BW|/GO] [/H]
Optimizes disk performance by reorganizing the files on the disk.

DEL (ERASE)
(Internal)
DEL (ERASE) [d:][path]filename [/P]
Deletes (erases) files from disk.

DELOLDOS
(External)
DELOLDOS [/B]
Deletes all files from previous versions of DOS after a 5.0 or 6.0 installation.

DELTREE
(External)
DELTREE [/Y] [d:]path [d:]path[...]
Deletes (erases) a directory including all files and subdirectories that are in it.

DEVICE
(Internal)
DEVICE=(driver name)
Used in the CONFIG.SYS file to tell DOS which device driver to load.

DISKCOMP
(External)
DISKCOMP [d:] [d:][/1][/8]

Compares the contents of two diskettes.

DISKCOPY

(External)

DISKCOPY [d:] [d:][/1][/V][/M]

Makes an exact copy of a diskette.

DOS

(Internal)

DOS=[high|low],[umb|noumb]

Used in the CONFIG.SYS file to specify the memory location for DOS. It is used to load DOS into the upper memory area and to specify whether or not the upper memory blocks will be used.

DOSKEY

(External)

DOSKEY [reinstall] [/bufsize=size][/macros][/history][/insert|/overstrike] [macroname=[text]]

Loads the Doskey program into memory which can be used to recall DOS commands so that you can edit them.

DOSSHELL

(External)

DOSSHELL [/B] [/G:[resolution][n]]|[/T:[resolution][n]]

Initiates the graphic shell program using the specified screen resolution.

EDIT

(External)

EDIT [d:][path]filename [/B][/G][/H][/NOHI]

Starts the MS-DOS editor, a text editor used to create and edit ASCII text files.

EXIT

(Internal)

EXIT

Exits a secondary command processor.

EXPAND

(External)

EXPAND [d:][path]filename [[d:][path]filename[. . .]]

Expands a compressed file.

FASTHELP

(External)

FASTHELP [command][command] /?

Displays a list of DOS commands with a brief explanation of each.

FDISK
(External)
FDISK [/status]
Prepares a fixed disk to accept DOS files for storage.

FILES
(Internal)
FILES=(number)
Used in the CONFIG.Sys file to specify the maximum number of files that can be open at the same time.

FIND
(External)
FIND [/V][/C][/I][/N] ÒstringÓ [d:][path]filename[...]
Finds and reports the location of a specific string of text characters in one or more files.

FOR
(Internal)
FOR %%(variable) IN (set) DO (command)
or (for interactive processing)
FOR %(variable) IN (set) DO (command)
Performs repeated execution of commands (for both batch processing and interactive processing).

FORMAT
(External)
FORMAT d:[/1][/4][/8][/F:(size)] [/N:(sectors)] [/T:(tracks)][/B|/S][/C][/V:(label)] [/Q][/U][/V]
Formats a disk to accept DOS files.

GOTO
(Internal)
GOTO (label)
Causes unconditional branch to the specified label.

HELP
(External)
HELP [command] [/B][/G][/H][/NOHI]
Displays information about a DOS command.

INCLUDE
(Internal)
INCLUDE= blockname
Used in the CONFIG.SYS file to allow you to use the commands from one CONFIG.SYS block within another.

INSTALL
(Internal)
INSTALL=[d:][\path]filename [parameters]
Used in the CONFIG.SYS file to load memory-resident programs into conventional memory.

JOIN
(External)
JOIN d: [d:path]
JOIN d: [/D]
Allows access to the directory structure and files of a drive through a directory on a different drive.

LABEL
(External)
LABEL [d:][volume label]
Creates or changes or deletes a volume label for a disk.

LASTDRIVE
(Internal)
LASTDRIVE=(drive letter)
Used in the CONFIG.SYS file to set the maximum number of drives that can be accessed.

MEM
(External)
MEM [/program|/debug|/classify|/free|/module(name)] [/page]
Displays amount of installed and available memory, including extended, expanded, and upper memory.

MIRROR
(External)
MIRROR [d:]path [d:] path [...]
MIRROR [d1:][d2:][...] [/T(drive)(files)] [/partn][/U][/1]
Saves disk storage information that can be used to recover accidentally erased files.

MKDIR
(MD) (Internal)
MKDIR (MD) [d:]path
Creates a new subdirectory.

MODE
(External)
MODE n
MODE LPT#[:][n][,][m][,][P][retry]
MODE [n],m[,T]
MODE (displaytype,linetotal)
MODE COMn[:]baud[,][parity][,][databits][,][stopbits][,][retry]

MODE LPT#[:]=COMn [retry]
MODE CON[RATE=(number)][DELAY=(number)]
MODE (device) CODEPAGE PREPARE=(codepage) [d:][path]filename
MODE (device) CODEPAGE PREPARE=(codepage list) [d:][path]filename
MODE (device) CODEPAGE SELECT=(codepage)
MODE (device) CODEPAGE [/STATUS]
MODE (device) CODEPAGE REFRESH
Sets mode of operation for devices or communications.

MORE
(External)
MORE < (filename or command)
(name)|MORE
Sends output to console, one screen at a time.

MOVE
(Internal)
MOVE [/Y|/-Y] [d:][path]filename[,[d:][path]filename[...]] destination
Moves one or more files to the location you specify. Can also be used to rename directories.

MSBACKUP
(External)
MSBACKUP [setupfile] [/BW|/LCD|/MDA]
Used to backup or restore one or more files from one disk to another.

MSCDEX
(External)
MSCDEX /D:driver [/D:driver2. . .] [/E][/K][/S][/V][/L:letter] [/M:number]
Used to gain access to CD-ROM drives (new with DOS Version 6).

MSD
(External)
MSD [/B][/I]
MSD [/I] [/F[d:][path]filename [/P[d:][path]filename [/S[d:][path]filename
Provides detailed technical information about your computer.

NUMLOCK
(Internal)
NUMLOCK=on|off
Used in the CONFIG.SYS file to specify the state of the NumLock key.

PATH
(Internal)
PATH;
PATH [d:]path[;][d:]path[...]
Sets or displays directories that will be searched for programs not in the current

directory.

PAUSE
(Internal)
PAUSE [comment]
Suspends execution of a batch file until a key is pressed.

POWER
(External)
POWER [adv:max|reg|min]|std|off]
Used to turn power management on and off, report the status of power management, and set levels of power conservation.

PRINT
(External)
PRINT [/B:(buffersize)] [/D:(device)] [/M:(maxtick)] [/Q:(value]
[/S:(timeslice)][/U:(busytick)] [/C][/P][/T] [d:][path][filename] [...]
Queues and prints data files.

PROMPT
(Internal)
PROMPT [prompt text] [options]
Changes the DOS command prompt.

RECOVER
(External)
RECOVER [d:][path]filename
RECOVER d:
Resolves sector problems on a file or a disk. (Beginning with DOS Version 6, RECOVER is no longer available).

REM
(Internal)
REM [comment]
Used in batch files and in the CONFIG.SYS file to insert remarks (that will not be acted on).

RENAME (REN)
(Internal)
RENAME (REN) [d:][path]filename [d:][path]filename
Changes the filename under which a file is stored.

REPLACE
(External)
REPLACE [d:][path]filename [d:][path] [/A][/P][/R][/S][/U][/W]
Replaces stored files with files of the same name from a different storage location.

RESTORE

(External)

RESTORE d: [d:][path]filename [/P][/S][/B:mm-dd-yy] [/A:mm-dd-yy][/E:hh:mm:ss] [/L:hh:mm:ss] [/M][/N][/D]

Restores to standard disk storage format files previously stored using the BACKUP command.

RMDIR (RD)

(Internal)

RMDIR (RD) [d:]path

Removes a subdirectory.

SCANDISK

(External)

SCANDISK [d: [d: . .
.]||/all][/checkonly|/autofix[/nosave]|/custom][/surface][/mono][/nosummay]

SCANDISK volume-
name[/checkonly|/autofix[/nosave]|/custom][/mono][/nosummary]

SCANDISK /fragment [d:][path]filename

SCANDISK /undo [undo-d:][/mono]

Starts the Microsoft ScanDisk program which is a disk analysis and repair tool used to check a drive for errors and correct any problems that it finds.

SELECT

(External)

SELECT [d:] [d:][path] [country code][keyboard code]

Formats a disk and installs country-specific information and keyboard codes (starting with DOS Version 6, this command is no longer available).

SET

(Internal)

SET (string1)=(string2)

Inserts strings into the command environment. The set values can be used later by programs.

SETVER

(External)

SETVER [d:]:path][filename (number)][/delete][/quiet]

Displays the version table and sets the version of DOS that is reported to programs.

SHARE

(External)

HARE [/F:space] [/L:locks]

Installs support for file sharing and file locking.

SHELL

(Internal)

SHELL=[d:][path]filename [parameters]

Used in the CONFIG.SYS file to specify the command interpreter that DOS should use.

SHIFT

(Internal)

```
SHIFT
```

Increases number of replaceable parameters to more than the standard ten for use in batch files.

SORT

(External)

```
SORT [/R][/+n] < (filename)
SORT [/R][/+n] > (filename2)
```

Sorts input and sends it to the screen or to a file.

STACKS

(Internal)

```
STACKS=(number),(size)
```

Used in the CONFIG.SYS file to set the number of stack frames and the size of each stackframe.

SYS

(External)

```
SYS [source] d:
```

Transfers the operating system files to another disk.

TIME

(Internal)

```
TIME hh:mm[:ss][.cc][A|P]
```

Displays current time setting of system clock and provides a way for you to reset the time.

TREE

(External)

```
TREE [d:][path] [/A][/F]
```

Displays directory paths and (optionally) files in each subdirectory.

TYPE

(Internal)

```
TYPE [d:][path]filename
```

Displays the contents of a file.

UNDELETE

(External)

```
UNDELETE [d:][path][filename] [/DT|/DS|/DOS]
UNDELETE [/list|/all|/purge[d:]|/status|/load|/U|/S[d:]|/Td:[-entries]]
```

Restores files deleted with the DELETE command.

UNFORMAT
(External)
UNFORMAT d: [/J][/L][/test][/partn][/P][/U]
Used to undo the effects of formatting a disk.

VER
(Internal)
VER
Displays the DOS version number.

VERIFY
(Internal)
VERIFY on|off
Turns on the verify mode; the program checks all copying operations to assure that files are copied correctly.

VOL
(Internal)
VOL [d:]
Displays a disk's volume label.

XCOPY

(External)
XCOPY [d:][path]filename [d:][path][filename] [/A][/D:(date)] [/E][/M][/P][/S][/V][/W][Y\-Y]
Copies directories, subdirectories, and files.

Introduction to windows

Windows is a computer operating system and graphical user interface (GUI) which enables you to work with a wide variety of programs on your computer - often simultaneously.

Designed by Microsoft Corporation to improve upon and replace Windows 3.1, Windows is faster, more powerful and easier to use than its predecessor. While Windows 3.1 is a 16-bit graphical user interface relying entirely upon the old computer conventions of the ms-dos operating system, Windows is a complete 32-bit operating system - which means it has power, speed and capacities not available in Windows 3.1.

Some of the new features incorporated into Windows include:

- 32 bit multitasking - the ability to run two or more programs (such as your word processor, your calculator and your printer) at once, switching between them at your convenience.
- plug and play - automatic detection and configuration of new hardware

- long file names - the use of lengthy file names, not limited to the 8 letters or numbers followed by three letter extensions of the previous Windows ms-dos operating systems.

In Windows, everything starts and ends on your DESKTOP - your computer screen covered with a solid color background, a patterned background, or a picture (wallpaper). Atop the desktop are icons which represent programs or collections of programs. By double clicking with your mouse on these icons, you can open programs or files, or collections of programs and files.

Apart from two or three icons which are installed with Windows, you may choose and create the icons which appear on your desktop.

At the bottom of your Windows screen is your TASKBAR, with a START menu. When you click on START, you open cascading panels where you can access hundreds of your most frequently used programs and documents. Those which you have opened are indicated by rectangular icons on your taskbar. Such programs will either be visible on your screen or minimized - not visible, but open and readily accessible.

On the right side of your TASKBAR are small icons which access programs that are automatically available when you turn on Windows. In the above illustration, the icons at right on the TASKBAR are gateways to a Norton antivirus program, your sound volume control, the Microsoft Plus program scheduler, Symantec's Crashguard and WindoWindows's system resource monitor.

TITLE BAR

The top line of your open window is the TITLE BAR, and lists the name of the program. You can move a window around your screen by dragging its title bar. If two programs or windows are open on screen, you can make one active by clicking its title bar (or an empty space within itself window); the title bar of an activated window is darkened.

On the top right are three control buttons. The left one is MINIMIZE, which keeps your window or program open and active, but reduces it to a rectangular icon on your TASKBAR. Open programs are indicated on your taskbar, but their names appear indented when they are visible on your screen.

The button on the middle RESTORES a full-screen window to window-size, or MAXIMIZES a window-size window to full-screen. Clicking the X button on the right closes your window or program.

MENU BAR AND VIEWING OPTIONS

In a window you will often see a variety of icons representing programs, files or facets of your computer. Double clicking these icons opens them.

The row beneath the title bar is the MENU BAR. When you clicking words on the MENU BAR, you open up a menu of additional choices.

Next to some of the commands on the MENU BAR are KEYBOARD SHORTCUTS (i.e. *control f4*) - keystroke alternatives for accessing the same command or action that is on the menu bar. Next to other menu bar commands, a right arrow indicates more options.

If you have mistakenly opened a menu and wish to close it without making a choice, simply click on the title bar or on a blank space within your window.

Most windows have a VIEW menu with a variety of choices for viewing their icons - SMALL ICONS, LARGE ICONS, LIST and DETAILS. They usually also have a HELP menu, with choices like CONTENTS, TOPICS, SEARCH and ABOUT...(which provides information on the program and current system resources).

The Start Menu

At the bottom left of your TASKBAR and at the bottom of your screen is your START MENU. By clicking on it and holding, you can choose *programs, documents, settings, find, help, run* or *shutdown,* as well as any icons that you place atop START.

You can also open the START MENU by holding down the *control* key, then pressing *escape*. By holding down your cursor on *programs*, you can view the programs initially installed on your computer. Holding down your cursor on one of those program names will open the program.

Holding your cursor on *documents* will reveal the last ten Windows 95 (not 16-bit Windows 3.1) documents which you opened. By holding your cursor on a document name, you can open it, provided that it is still at same location on your computer.

SETTINGS gives you access to the *Control Panel,* where you can reconfigure your computer and its hardware, as well as alter settings for the software which enables your computer to operate.

You also have access to your PRINTERS folder here, as well as options for your taskbar and programs menu.

In FIND, you can search for programs, files and folders.

HELP provides you with a variety of means for getting help with Windows and its accessories.

Using RUN, you can RUN a program from your hard drive or a floppy disk.

Using SHUT DOWN, you have several options for shutting down and restarting your computer.

SHUTDOWN

You should always use shutdown when exiting Windows. Simply turning off your computer without shutting down properly can damage your programs.

In SHUT DOWN, you can choose *restart in Ms-Dos mode* to restart your computer in MsDos 7.0 - Windows's Ms-Dos. Choose RESTART to reboot, or to use an earlier Ms-Dos or Windows 3.1.

NOTE: In the event of a crash (often a frozen computer screen and mouse), you may not be able to access SHUT DOWN. Instead, press and hold the *control, alt* then *delete* keys, and choose *end task* to shut down the offending program or shut down to exit Windows altogether. If this doesn't work, press the *reset* or *reboot* button (not the power button) on your computer.

HOW TO CHANGE DATE AND TIME

From the START button, choose *settings, control panel, date and time.* On the date and time screen, change date and time. Be sure to note A.M. then P.M. Choose *apply*, then *o.k.*. (Once you install the clock on your taskbar, you can double click the time to reach these screens).

On the time zone screen, change the time zone if necessary. Indicate also if you want automatic time changes to occur when daylight savings time changes in your locality. Then close *Control Panel* by clicking on the X control at the top right of its screen.

NOTE: If your date and/or time are inaccurate, your computer may need a new battery.

The Taskbar

The TASKBAR at the bottom of your screen indicates your minimized programs -programs you have opened but not closed, and which are still active. If you don't like the location of your taskbar, move it to another location; you can also hide it and resize it.

To MOVE your taskbar, grab it with your cursor on a blank area and drag it to another side of your screen. To RESIZE it, place your cursor along its border and drag it outward.

By right clicking the taskbar, you can arrange open windows on your screen, tiling them horizontally or vertically or cascading them (If you hold your mouse briefly on an incomplete name on a minimized application, the entire name will pop up.)

You can access taskbar options via the START MENU (*settings, taskbar*) or by right clicking the taskbar, and clicking *properties*. Your choices in *Taskbar, properties/ options* include:

ALWAYS ON TOP - When chosen, your taskbar will be visible no matter what program you have open, giving you less screen space but making minimized programs accessible.

AUTOHIDE - When you choose *autohide*, the taskbar disappears, giving you more screen space. It reappears when you move your cursor to the very bottom of your screen.

SHOW SMALL ICONS - You can make the icons on your START menu smaller, so that you can add more programs to your START Menu.

SHOW CLOCK - When chosen, a small clock will appear at the bottom right of your taskbar. By double clicking the time, you access the *date and time* dialog box.

Click the checkbox next to each of these options to turn it on, then click *o.k.*, then *apply*. Click again next to the option to turn it off, and click *o.k.*, then *apply*.

About My Computer and Explorer

In Windows, you can locate and manage your files and directories (folders) in either *My Computer* or *Explorer*. *My Computer* is most convenient for viewing one directory or window at a time; *Explorer* is your best choice when you want to view or reorganize files in different parts of your computer.

In *My Computer*, you choose a drive, and by default, view icons or a list of folders and files in that drive. Your view is one window at a time, unless you choose *explore* in *My Computer* to move to *Explorer*, which makes it easy to work in a particular folder, but difficult to move between folders.

When you open a folder in *My Computer*, a new window opens; when you open a folder within that window, another window opens. Soon your screen is cluttered with windows, unless you initially choose, in *view, options* to use a single window.

In *Explorer*, you have a dual pane view (like in Windows 3.1 File Manager). You view your drive and root folders in the left pane, and the contents of chosen folders in the right pane. You also have two choices not included in *My Computer* - a *tools* menu which enables you to choose *go to* if you wish to open a file, and *find* to locate a file.

5. Programming with C

→ Write a program to calculate simple interest.

```
# include < stdio .h>
main(  )
{
int p = 20,000,  n=10, r=12, I;
clrscr ();
{
I = p* n * r/100;
Print f (“simple interest;, % d”, I);
}
getch () ;
}
```

→ Find the students average in 5 subjects. If average is greater than or equal to 60- I class

        >= 50 –II class
        >= 40-III class
        <40 fail using if & else.

```
# include <stdio.h>
main ( )
{
in t S1, S2, S3,S4, S5, T, avg;
clrscr  ( );
printf (“enter the marks of S1, S2, S3, S4, S5”);
scanf (“ % d % d % d %d %d”, &S1  &S2 &S3 &S4 &S5);
T = S1 + S2 + S3 + S4 + S5;
Avg = T/5 ;
if (avg >=50)
printf (“second class”);
else
if (avg >=40)
print f (“third class”);
else
printf (“fail”);
}
```

→ write a program to print the Fibonacci series < = 100.

```
# include <stdio.h>
# include < conio.h>
main ( )
{
int a=0, b=1, c=0;
clrscr ( );
print f (“%d %d”, a, b);
while (c<=100)
{
        c = a+b ;
        a = b;
        b=c;
        printf (“%d \ t”, c);
}
        getch (  );
}
```

→ write a program to print the biggest number.

```
# include <stdio.h>
# include <conio.h>
main ( )
{
int a,b,c;
clrscr ( );
printf (“enter 3 numbers”);
scanf (%d %d %d, &a, &b, &c);
if (a>b) && (a>c)
{
printf (“the biggest is a% d”, a);
}
else if (b>c)
{
printf (“the biggest is c%d”, c);
}
getch ( );
}
```

→ write a program to print 5th table.

```
# include <stdio.h>
main ( )
{
int n=5, I, t;
clrscr (  );
for (I=1; I<=10, I++)
{
t = n * I;
printf (“%d %d = %d \n”, n, I, t);
}
getch (  );
}
```

→ Write a program to check whether the given number is Armstrong or not

```
# include <stdio.h>
# include <conio.h>
void main ( )
{
int r=153, r, sum=0, n1=n;
clrscr ( );
while (n>0)
{
```

```
r = n% 10;
sum = sum + (r * r * r);
n = n/10;
}
if (sum = n1)
printf (“Armstrong”);
else
printf (“not Armstrong”);
getch ( );
}
```

→ Write a program to generate current bill

```
# include <stdio.h>
# include <conio.h>
void main ( )
{
int a, t;
clrscr ( );
scanf (“%d”, &a);
if (a<=100)
{
t = a *2;
printf (“the bill is %d”, t);
}
else if (a>100 && a <=300)
{
t = a*3;
printf (“the bill is %d”, t);
}
else
{
t = a*5;
printf (“the bill is %d”, t);
}
getch (   );
}
```

6. Search on Medline, CD, Bibliographic databases

The utility of a thorough literature search is not limited to a student preparing for a seminar or dissertation, but is also required for those engaged in patient care. It has been shown that use of on-line literature searches early in the course of hospitalisation significantly reduces the duration of hospital stay[1] and traditional paper-based references play only a limited role in clinical problem solving[2]. However, with limited resources in terms of time and money and the ever-expanding literature volume, one would like to perform a search that is not only comprehensive or complete, but also do it in the shortest possible time. If the way we perform a search is not proper, we could well be a victim of the most fundamental computer adage: garbage in, garbage out[3]. This happens because computer based literature databases make it possible for us to just punch in a few words and get some relevant results. We get the satisfaction of obtaining an instantaneous result. However, it is worth pondering whether this is the best response.

In this series, we will try to provide information about the medical databases and learn to conduct a comprehensive search in an efficient manner. We begin the series with the most widely used database, the MEDLINE. We will discuss the search features at Entrez PubMed (http://www.ncbi.nlm.nih.gov/entrez/query) based on PubMed Help[4]. A simple tutorial is also available on http: //www.library.health.ufl.edu/pubmed/pubmed2/. However, only search-related issues will be discussed and topics related to displaying, copying and printing would not be dealt with.

:: What is medline?

The National Library of Medicine (NLM), USA and its predecessor, the Surgeon-General's office of the U.S. Army, have been indexing medical literature since 1879, when the predecessor to today's Index Medicus was first produced[5]. MEDLINE (Medical Literature Analysis and Retrieval System Online) is the electronic version of the Index Medicus. It consists of over 10 million articles published in 4300 journals in 40 different languages from 75 countries[6]. Approximately 8000 completed references are added to this database every Saturday, January through October amounting to the annual increment of over 400,000 articles.

To sift through this massive load for what one needs at a particular moment could be worse than looking for a pin in the haystack. We will, therefore, briefly review the basics of indexing and retrieval systems.

:: 'text word' search versus "concept based" search

In text word search, the query words are matched word to word with those in the database. Such a search has some limitations. The words must exactly match in order for the article to be retrieved. Thus, a difference in spelling may lead to exclusion of an article. Secondly, as MEDLINE includes only the title and abstract of an article, incomplete or inadequate abstracts may preclude inclusion of the article. However, the most fundamental limitation is that the "text word" search does not pay any attention to the idea or concept behind an article. For example while searching for drugs used in parkinsonism use of key words 'drug' and 'parkinsonism' will retrieve articles related to drugs for treatment of parkinsonism as well as drug induced parkinsonism. In addition, this kind of search may lead to inappropriate results as the searcher and the author differ in the way they represent or understood a concept[7]. For example one can represent coagulopathy as bleeding disorder, bleeding diathesis or by individual factor deficiencies. “Concept-based” search overcomes few of these limitations.

:: Medline’s “concept based” search

NLM has designed a “concept-based” search methodology using Medical Subject Heading (MeSH). Over 19,000 standardised medical terms constitute the thesaurus of MeSH[7]. The arrangement of the MeSH is in the form of a tree where subject headings are arranged under one another with increasing specificity [Table - 1]. It also contains a group of 82 subheadings [Table - 2]. Trained indexers scan published articles, interpret the findings, identify the thrust or themes of these articles, and assign 10-12 MeSH terms and subheadings to each article[7]. In an example shown in [Table - 3] ‘MH’ denotes the allotted MeSH terms for the given article and the terms after the forward slash (/) are the subheadings. Etiology and complications (marked using asterisk*) are the MeSH major topics for the given article. Major topics relate to the main concept conveyed by the article, as judged by the indexing experts.

:: The basic search

The most basic form of search would be to type a word or two in the query box without any specification. PubMed handles the query word(s) in the following sequential manner (called automatic term mapping):

a. A query term without any specification (e.g. trichobezoars) is first mapped to the ‘MeSH Translation Table’, which contains MeSH Terms, and other systems to check for equivalent synonyms or lexical variants in English. If a match is found in this translation table, the term will be searched as MeSH and as a text word. For example, trichobezoar will be translated to bezoars and the resulting search would be ("bezoars"[MeSH Terms] OR trichobezoar[Text Word]). Similarly, ischemia is also searched as ischaemia and ANA as antinuclear antibody.

b. The query term is then mapped to the ‘Journals Translation Table’, which contains the full title of the journals, the MEDLINE abbreviation, and the ISSN number. It tries to map a search term to the journal abbreviation. For example, journal of postgraduate medicine will be translated to: "J Postgrad Med"[Journal Name].

c. The ‘Phrase List’ is consulted next whereby PubMed translates separate query words into as a single phrase. For example giant cell tumour is matched to the phrase “giant cell tumour.”

a.

d. If the query word is not found in these three mappings, and is a word with one or two letters after it, PubMed then checks the Author Index. For example roy ak will be searched in the list of authors.

The results generated thus from this automatic mapping are displayed in a chronological order of entry of the citations in the MEDLINE.

:: The advanced search

One may be tempted to question the need for an advanced search if the basic search itself involves automatic mapping and also looks for the variations in languages. The answer to this question is simple, to enhance sensitivity and specificity of the search. For example, text words HIV therapy would generate more than 30,000 articles. Text words hospital infection control will get mapped onto the journal title Hospital Infection Control. Hence, we need to learn more than just typing text words as keywords. We will now review few of the advance search facilities available at Entrez PubMed.

:: Use of search field tags

Each search field tag is abbreviated in two characters and can be typed in brackets, e.g. [au] for author's name or [ti] for title words. Use of these tags directs the search in the specified field alone e.g. roy ak [au] will retrieve articles authored by roy ak; hernia [ti] will get the articles containing hernia in the article titles alone. A list of some common tags is provided in [Table - 4].

:: Use of mesh terms

To search a term as a MeSH term, one may use search field tag [mh]. For example, if search is requested for therapeutics with the field tag [mh], MeSH search being a concept-based search would not retrieve articles that contains word therapeutics in title or abstract but do not have therapeutics as the main theme of the article. In this manner, the advanced search focuses on articles based on understanding of the main concept underlying the articles. This is almost akin to a search made on the basis of thinking or cognitive ability.
The search can be further fine tuned by using major topics [majr] and subheadings [sh]. While searching for articles related to aetiology of duodenal diseases, use of aetiology as one of the key word with the tag [majr] would retrieve the article cited in [Table - 3] and other articles focusing specifically on the aetiology of duodenal diseases. Subheadings can be used with MeSH terms to describe a particular aspect of the subject in a more comprehensive manner. Tag [sh] is used for such a search e.g. hernia [mh] AND surgery [sh]. The two character abbreviation of subheadings can also be used e.g. su for surgery [Table - 2].
Subheadings can be directly attached to the MeSH term in the following format: MeSH Term/Subheading-e.g. hernia/surgery or hernia/su, which would give even better results than henia [mh] AND surgery [sh]. Only one appropriate subheading can be attached to a MeSH term. Of course, not all MeSH term/subheading combinations are valid (e.g. hernia/toxicity).

:: From where does one obtain mesh terms?

We have seen the importance of MeSH terms and subheadings. The question is how does one know if a particular word or term is indeed a MeSH term. The MeSH Browser (http://www.ncbi.nlm.nih.gov:80/entrez /meshbrowser.cgi) can be used for this or one can download the entire list for personal use from http:// www.nlm.nih.gov/mesh/filelist.html. Another way to find out the MeSH terms is to check these from the MEDLINE citation of a known article.

:: Use of boolean characters

'AND', 'OR', and 'NOT' are the Boolean characters, which can used to get more precision in a search. Please note that the Boolean characters should be typed in capital or upper case. More than one Boolean character can be used at a time. For example, if one were to find articles that deal with the effects of cough or constipation on inguinal hernia, he can use the syntax: (cough OR constipation) AND inguinal hernia. If search is required for review articles that discuss the treatment of hernia in all patients except children, syntax (hernia/therapy [mh] AND review [pt]) NOT child [mh] will give the appropriate results.

Please note that use of NOT can lead to unsatisfactory results. For example if one is interested in knowing about non-surgical treatment options for hernia and use hernia/therapy NOT surgery as the search, he will miss articles discussing the both surgical and non-surgical treatment.

:: Use of limits

PubMed search has a feature called “Limits.” By using this feature, one can literally limit the search to a specific age group, gender, human or animal studies, a specific language, type of articles (review, studies, etc), publication or Entrez date. One can even limit the search to a specific subset of citations within PubMed e.g. AIDS-related citations or in-process citations i.e. Pre-MEDLINE citations. These limits can be set from the Features Bar and can also be used in the form of tags as shown in the [Table -5]. It should be borne in mind that the use of “limits” for publication type, age, gender, human or animal studies will restrict the retrieval to MEDLINE citations. The Pre-MEDLINE citations get excluded depriving some of the most recently entered citations that are undergoing the indexing process.

:: Use of some additional features of pubmed

Truncation

With the use of asterisk at end of a term one can search all the words (up to 150) beginning with that term. For example, search query with polyp* will include additional terms such as polyps, polyposis, polypus, polypectomy, etc. However, this kind of search will not allow the searcher to use the automatic term mapping. In addition, if there is a phrase involving term with a space after the specified term (e.g. polyp surgery), it will not be included in the result.

Phrase Searching

We have seen that PubMed uses a Phrase Index and maps two or more text words into a logical phrase. However, for a phrase that is recognised by PubMed as separate words, double quotes around the phrase, e.g. "medical literature” will force PubMed to search it as a single phrase. Please note that using this function one will switch off the automatic term mapping.

Journal Subsets

The following journal subsets (jsubset) are available for a search: jsubseta - Abridged Index Medicus, jsubsetd - Dental, jsubsetn - Nursing. For example: hernia/su NOT jsubsetn will not retrieve articles from nursing journals.

Preview

This feature allows to have a peek at the number of citations that will be retrieved. One can then add or delete more terms to refine the search. For example, search for HIV/therapy may result in about 30,000 retrievals, if word child is added it will give about 6,000 citation, addiing review [pt] will result in a manageable 800 articles being retrieved. Preview can thus be used to know the result of the search prior to actual display of the citations, thus saving the on-line time.

History
This feature allows to combine searches or add more terms to an existing search. Up to 100 searches are kept in 'memory' by PubMed for one hour and are numbered in the order they were performed. By using the sign (#) before the search number, e.g., #1 OR #2, or #1 AND (etiology OR pathology) one can improve the results.

Details
This feature gives the strategy based on which the results were generated. From Details, one can save a search strategy or edit the search strategy and resubmit it. Studying this can help to improve the search abilities.

Related to
Each citation in PubMed has a link that will retrieve a pre-calculated set of PubMed citations that are closely related to the selected article. So if a 'good' article is found use this facility to get more 'good' articles.

Saving a search strategy
In the Details window, clicking URL translates and embeds search strategy as part of the URL. Then using web browser's bookmark function one can save it as a bookmark.

:: How to search for a journal?

One may search by the full journal title, e.g. journal of postgraduate medicine; the MEDLINE abbreviation, e.g. j postgrad med; or the ISSN number, e.g. 0022-3859. However, searching with ISSN number alone may fail to retrieve older citations. If a journal name is also the same as one of MeSH headings or is a single word, (e.g., heart) PubMed will search the unqualified term as a MeSH heading or a text word, hence it is better to qualify the journal title with the journal title search field tag, e.g., heart [ta].

:: How to search for a person?

To search for a person use last name with the initials in small letters without any punctuations (e.g. roy ak). To find only Roy A use roy a@ or "roy a". Whereas, roy a* or roy a will retrieve even roy aa, roy ab, etc.

:: How to search for a single known article?

PubMed's Single Citation Matcher allows to locate a specific single article using any or all of the following bibliographic elements: journal title, date, volume, issue, page, or author.

:: What is the best search strategy?

Indexing is done by humans who, in spite of their excellence and training, have different views of what is most important in the articles and not every researcher would view a paper in exactly the same way as the indexer has. It is therefore difficult to guarantee that all articles will always be indexed with every appropriate MeSH heading. Hence, MeSH alone is not sufficient when creating a powerful and comprehensive search and it is not advisable to rely solely on the MeSH[9],[10],[11],[12]. So neither the text based search nor the MeSH based search is complete, if performed in isolation. These two methodologies complement each other. Many researchers agree that a combination of free text searching and concept-based searching is the only solution to adequate searching of MEDLINE[12],[13],[14],[15]. The new feature of PubMed whereby the search query is converted automatically to a combination of text word and MeSH term should help to solve this problem. The matter does not end here. Federiuk found that use of abbreviations in the titles and abstracts hampers the search results[16]. She showed that three different search strategies retrieved pools of unique articles: concept based search with the abbreviation, text word based search with the abbreviation and the text word search with the definition of the abbreviation. Hence, to make sure that one gets the best results, one may have to use combinations of search methodologies.

One must also remember that MEDLINE is a date limited manually indexed database containing selected portions of selected journals. The journals listed in MEDLINE are 'the best' but not all the best[17]. Hence any claim of 'thorough' literature search based solely on MEDLINE, is a loud claim. For a comprehensive search one need to review other databases, few of which will be covered in the forthcoming articles.

:: Final tips for better search results

One of the most important issues is to formulate a flexible search strategy using combinations of text words and MeSH terms taking care of abbreviations, off-line. Preview the results, add or delete terms and then get the results. Use the feature 'related to'. Use history feature to combine various search results. One will have to vary his methodology depending on the need. In a clinical setting, a specific search using MeSH terms and major concepts should suffice. Randomised controlled trial [pt] or drug therapy or dt [sh] or therapeutic use or tu [sh] or all random [tw] should be used for treatment related studies[14]. If one is writing a review or claiming to be the 'first' for something, use of all the possible combination of search methodologies is recommended. Text word searches can be used if the search is related to a very new subject and has not yet been given an indexing term, is a brand name, or is obscure[15]. It is a good strategy to check MeSH terms of one's own articles once it is in MEDLINE. This will help to understand how indexing is done. May be the authors will be able to suggest better MeSH terms for their articles to NLM indexers!

:: Medline sites

There are numerous sites providing free access to MEDLINE. However, the number of relevant citations varies from site to site even though a similar query approach is used. There is difference in the ease of use, the interface, the level of details and additional facilities at various sites. Hence, depending on the need, one should be aware of the strengths and limitations of a particular site. One can find a list and comparison of these sites at http://www.docnet.org.uk/drfelix/ and http://www.muhealth.org/~library/docs/mla97.html.

:: How to make your own articles easily retrievable?

A few researchers grade their work by the number of times it is cited by other workers. For others to cite your work, it should first be retrievable by and accessible to the interested individuals. Hence, it is important to pay special attention to the following: Abstract: The content of the article is very important. But it is not accessible fully to the searchers on the MEDLINE. What is available to them is the abstract of the article. So, it is essential to take utmost care that the abstract includes all the key elements, and is adequate to make the article easily and frequently retrievable. Abbreviations[16]: The results of both subject and text word searches may be affected by the use of abbreviations, hence abbreviations should be avoided. If an abbreviation is used but not defined in the title or abstract, the article will not be retrieved by text word search using the definition.

:: Do we need to teach literature search?

Medical educators have recognised the importance of teaching information retrieval skills to the medical students (Steering Committee on Medical Evaluation of Medical Information Sciences in Medical Education 1986)[18]. Proud et al have shown that when students were taught the skills of accessing MEDLINE, they could formulate a question, retrieve current information, critically review relevant articles, communicate effectively, and use these skills to contribute to patient care[19]. If a newly learned skill is taught at the point of need, there is a greater likelihood of it being retained[20]. Hence, it is worth including literature search using databases as one of the teaching-learning activities.

CD databases

CDDB

From Wikipedia, the free encyclopedia.

CDDB (which stands for **C**ompact **D**isc **D**ata**b**ase) is a database for software applications to look up CD (compact disc) information over the Internet. This is performed by a client which calculates a (nearly) unique disc ID and then queries the database. As a result, the client is able to display the artist name, CD title, track list and some additional information.

The database is used primarily by media player and CD ripper software.

Contents

- 1 History
- 2 Technical
- 3 Alternatives
- 4 External links

History

CDDB was invented by Ti Kan and Steve Scherf. The source code was released under the GNU General Public License, and thus many people submitted CD information believing that the contributions, too, would remain freely available to others. Later, however, the project was sold and the license conditions were changed and it was no longer a free service, requiring commercial developers to pay an "initial fee", as well as a license fee based on the usage of the servers and support. It also included terms that many programmers felt were unacceptable: no other similar database (such as freedb) could be accessed in addition to CDDB, and the CDDB logo was required to be displayed while the database was being accessed.

In March 2001, CDDB, now owned by Gracenote, banned all unlicensed applications from accessing their database. New licenses for CDDB1 (the original version of CDDB) were not available anymore, as they wanted to force programmers to switch to CDDB2 (a new version incompatible with CDDB1 and hence with freedb).

After the unpopular commercialization of CDDB as Gracenote, most media player applications switched to freedb, but continued to refer to the service as 'CDDB' as a generic term. It is still common to see many applications refer to CDDB in their documentation when in fact the application is using freedb.

Technical

CDDB was designed around the task of identifying entire CDs, not merely single tracks. The identification process involves creating a 'discid', a sort of "fingerprint" of a CD created by performing calculations on the track duration information stored in the table-of-contents of the CD. This discid is used with the internet database, typically either to download track names for the whole CD or to submit track names for a newly-identified CD.

Since identification of CDs is based on the length and order of the tracks, CDDB cannot identify playlists in which the order of tracks has been changed, or compilations of tracks from different CDs. CDDB also cannot distinguish different CDs that have the same number of tracks and the same track lengths.

Alternatives

The licence change motivated a new project, freedb, which is intended to remain free.

An alternative project that aims to enhance CDDB beyond a mere database of CDs is called MusicBrainz. Their site also contains more information on CDDB and some database statistics of CDDB and freedb.

Another commercial alternative to CDDB is the AMG LASSO service. LASSO was launched by All Media Guide in late 2004 and includes recognition technology for CDs, DVDs, and digital audio files. The AMG metadata database is generally recognized to be more comprehensive and of higher quality, because of quality controls that CDDB lacks. Microsoft's Windows Media Player, Musicmatch Jukebox, and the Virgin Digital Megastore are licensees.

Bibliographic Databases

Search bibliographic databases on the Internet:

- Using EndNote's *Connect...*and *Search...* commands, you can search Internet databases just as easily as you can search your EndNote database on your computer.
- Simply open any of more than 560 predefined connection files and you're online and searching.
- Access hundreds of remote bibliographic databases, including Web of Science, Ovid, PubMed, the Library of Congress, and university card catalogs from EndNote.
- Connect to data sources worldwide—EndNote provides MARC formats that support native language libraries around the world.
- Search remote bibliographic databases using EndNote's simple search window—great for locating specific references.
- Export reference directly from Web of Science, Highwire Press, Ovid, OCLC, ProQuest and more.

- Save and load search strategies at the click of a button.
- Drag and drop references instantly to your own EndNote database in one simple step. No additional importing required.

ENDNOTE:

Bibliographic Databases on the World Wide Web
Information that about print and electronic journal articles or articles in periodicals can generally be found in bibliographic databases. Examples of information types found in bibliographic databases generally include title, author, abstract; and may also include links to full-text content. For searches relating to biomedical subject material, the National Library of Medicine (NLM) databases provide access to peer-reviewed bibliographic citations.
Before describing the various types of databases, it might be useful to distinguish between several terms that are sometimes confused when people discuss databases.

Databases

- **Providers**
- **Interfaces**
- **Types**

Databases are simply collections of data, organized into files (often called tables) that contain records (e.g. a row of data about a specific individual). Records may be further delimited into specific fields (may be classified on the basis of several different criteria (e.g. last name, first name, SS#, street address, city, state, zip, etc.). The files in the database can be searched (queried) through search interfaces that facilitate construction of queries, or directly by using specialized languages (e.g. SQL).

Database providers are companies that provide access to information in groups of databases, generally for a fee. One example of a major database provider is Dialog. Dialog provides access to hundreds of databases through proprietary telnet, dialup, or login via the WWW. Dialog provides information about the contents of each of the databases (or files) in its Blue Sheets, which can be accessed for no cost. Dialog also provides several proprietary interfaces for searching its databases. Ovid is a database provider accessible at UIC to search the full text core collections.

Database search interfaces link the user to the search engine that search the databases. They facilitate searching by allowing the use of natural language terms, by mapping user-generated search terms to appropriate subject headings (indexed databases), and provide user-friendly tools such as menus, check boxes, buttons, and check lists to define search parameters, eliminating much typing of terms in appropriate search syntax. A number of databases provide interfaces that are accessible via the WWW. An example is the Ovid interface used at UIC to search the Core Biomedical Collection (CBC) full text collections and IBIS. Another is the PubMed or Internet Grateful Med (IGM) interfaces used to search the National Library of Medicine's (NLM) MEDLINE bibliographic database.

- **Internet** Grateful Med provides free access to MEDLINE, AIDSLINE, HealthSTAR, AIDSDRUGS, AIDSTRIALS, DIRLINE, HISTLINE, HSRPROJ, OLDMEDLINE and SDILINE. Search features: Utilize full range of Medical Subject Heading (MeSH) search features using UMLS Metathesaurus. Ability to limit searches by language, publication type, age groups, etc., using pull-down menus.

- **PubMed** provides free access to MEDLINE. Sets of related articles pre-computed for each article cited in MEDLINE Choice of Web search interfaces from simple keywords to advanced Boolean expressions. Field restrictions and MeSH index terms (main topics and subheadings) supported. Linkages to publishers' sites for full-text journals. Approximately 100 journals available, some by subscription only. Clinical query form with built-in search filters for diagnosis, etiology therapy, and prognosis. Links to molecular biology databases of DNA/protein sequences and 3-D structure data.

- **Illinois Bibliographic Information Service (IBIS)** is a collection of databases which you can use to find references to recent articles in journals and magazines in many subject areas. Among the available databases are Reader's Guide to Periodical Literature, Social Sciences Index, PsychINFO,ERIC, Art Index, Current Contents, and other databases. UIC owns a selection of the journals that are cited in IBIS. To find out if UIC owns a particular magazine or journal, use UICCAT, UIC's online catalog. Do a title search for the name of the magazine or journal. IBIS is only accessible to affiliates of institutional subscribers.

Types of databases can be classified arbitrarily on the basis of structure, accessibility, content or purpose:

Structure

- **Flat file** A database whose data is organized into a single table or tables that must be searched separately for information in specific records or fields.
- **Relational** A database whose tables are linked together by a linking table that contains records or fields common to both or pointers to fields in other tables. This allows advanced searching across multiple tables.

Accessibility

- **Free** Databases that can be accessed without charge like library catalogs or many bibliographic databases (e.g. UICCAT, MEDLINE).
- **Proprietary** Databases whose contents can be accessed only by paying a fee or subscribing to an organization or database provider (e.g. Chemical Abstracts).

Content or purpose

- **Library or union catalogs** Online Catalogs are the primary means of access to a library's collection in a machine-readable format. They are specialized types of databases whose data are bibliographic records standardized into fields, linked to a holdings database containing the contents of the library collection. This allows library users to search by subject, title, author, ISBN, etc. to find information about the book or magazine they are seeking. The holdings data tells the circulation status of the item (i.e., checked out or not) and the library call number. A library's catalog is cumulative, including all materials (books, journals, audiovisual, etc.) held in a collection at any particular geographic location or locations. A union catalog links to the holdings of a consortium of libraries. UIC's Illinet Online is an example of a union catalog (your browser must have a telnet application associated with it).
    - UICCAT is the electronic "card catalog" for UIC libraries. It contains information about books, magazines, journals, government publications,

and other materials owned by the UIC libraries. You can search the collections of the five libraries on UIC's Chicago campus -- the Main Library, Library of the Health Sciences, Architecture and Art Library, Math Library, and Science Library. UICCAT also contains information about materials at the Library of the Health Sciences' regional sites at Peoria, Rockford and Urbana.

  - ILLINET Online (IO - previously called LCS/FBR or MILO) is the combined online catalog of over 800 Illinois libraries. It also includes the circulation status (whether an item is checked out) for materials at UIC and 44 other libraries, known collectively as ILCSO (Illinois Library Computer Systems Organization) libraries. UIC students, faculty, and staff have borrowing privileges at all ILCSO libraries.

- **Bibliographic** A bibliographic database contains bibliographic information (title of article, journal name, author, date of publication, volume #, issue, page #, etc.) about various types of publications and formats (print, video, audio, software, etc.). Bibliographic databases are machine-readable form of indexes and abstracts. In bibliographic databases, the base record is a citation to an article, book, chapter, or paper. The citation may include an abstract or summary of the item, subject headings Author tile, publication type, date of publication, and language of the material may also be available.
  - **MEDLINE** Subject: Biomedicine Type: Bibliographic citations Coverage: All languages; publications from 1966 to the present. Recent references are contained in the current file (MEDLINE); segmented MEDLINE Backfiles (MED90, MED85, MED80, MED75, and MED66) contain older material. The file contains over 8.5 million records. Document Types: Articles from more than 3,700 international biomedical journals (some chapters and articles from selected monographs are found in earlier years). Special Features: MEDLINE is NLM's premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, and the preclinical sciences. Journal articles are indexed for MEDLINE, and their citations are searchable, using NLM's controlled vocabulary, MeSH (Medical Subject Headings). MEDLINE contains all citations published in Index Medicus, and corresponds in part to the International Nursing Index and the Index to Dental Literature. Citations include the English abstract when published with the article (approximately 74% of the current file).
  - **HealthSTAR** (Health Services, Technology, Administration, and Research) Subject: Clinical (emphasizes the evaluation of patient outcomes and the effectiveness of procedures, programs, Journal of Bone and Joint Surgery (Am. Vol) Journal of Clinical Investigation Lancet New England Journal of Medicine Pediatrics Science Some special features of CBC include: full text and graphics; ability to view an outline, move to a selected section of a document such as abstract or methods hyperlinks to references and graphics from within the text ability to retrieve all articles from a particular journal or journal issue in the database limiting to original articles, reviews, and publication year
- **Full text** Full text databases are often linked to bibliographic database citations. They contain the articles full text, but may be missing graphics like photos and

charts. Generally there is a fee for accessing them, unless access is gained through a university like UIC's Core Biomedical Collection.

- **IDEAL** is an online electronic library containing all 175 Academic Press journals. Abstracts and tables of contents are presented in HTML and full-text articles are delivered in Adobe Acrobat format.

- **Data databases** This is a catchall term to include a myriad of database types. Examples are databases containing genetic base sequences, amino acid structures, graphic images, patents, business information, statistics, and many other types of information.

More than 200,000 researchers, scholarly writers, students, and librarians use EndNote to search online bibliographic databases, organize their references, and create bibliographies instantly and automatically. Instead of spending hours typing bibliographies, or using index cards to organize their references, they do it the easy way -- by using EndNote! EndNote for Windows and Macintosh is a valuable all-in-one tool that integrates the following tasks into one program.

Features

Create an unlimited number of databases

In EndNote, bibliographic data is entered into a record by filling out a simple template that has the proper fields displayed for a given type of reference. EndNote has 17 customizable reference types including journal article, book, electronic source, and conference proceedings, with up to 30 fields for entering data, including abstract and notes. You can customize the reference types to add fields of your own. Each EndNote database can store up to 32,000 references.

Includes more than 100 import filters

There are many sources of bibliographic data that can be used to help you build your EndNote database of references. EndNote includes more than 100 customizable import filters for a variety of online and CD-ROM database providers (e.g. Dialog, SilverPlatter). You can create your own import filters as well. Once you've downloaded the references in a tagged format, choose the appropriate import filter to import the data into your EndNote database.

Please Note

If you have access to a library catalog or bibliographic database that supports the Z39.50 protocol, you can use EndNote to search and retrieve references directly into EndNote. This removes the need to learn a separate program to access online databases or go through extra steps of saving the references to a text file and importing them into EndNote. EndNote includes more than 100 connection files to access and search online databases.

Launch a web browser

EndNote records include a URL field to store World Wide Web addresses. This allows you to store a web page within your EndNote record, providing a virtual map to full-text online information. By selecting the Launch URL... command, you can automatically start your web browser (e.g. Navigator, Internet Explorer) and link to online journals, full-text articles, or any other web address stored within your EndNote record. Your EndNote database now becomes the card catalog of the electronic library.

Term Lists

For consistent data entry in your EndNote database, you can maintain lists of important terms such as journals, authors, and keywords. Term Lists store glossaries of keywords, author names, or any other terms that are important to managing your references. You can use these lists when entering reference information to ensure that terms are entered in a consistent way. Term Lists can be linked to different fields and then called up instantly when you are typing into a particular field. This feature also speeds up data entry since terms can be instantly transferred into a record.

Term lists can also be accessed when using EndNote's Search command so that you will be sure to search for the same terms that you have entered in your references. You can create up to 31 Term Lists, and each list can hold thousands of names or terms. A special case term list for journals is designed to handle journal names and abbreviations.

Searching

The Search... command in EndNote offers a high degree of flexibility and control in designing searches. You can limit your searches to specific fields such as Author name, Year, or Keywords. You can also choose a general search, one that searches on all fields in a database.

EndNote allows you to create unlimited length search strings using the boolean connectives AND, OR, and NOT. With the QuickFind feature EndNote will display your search results almost instantaneously, even if your database has more than 10,000 records in it. EndNote also allows you to use comparative operators (greater than, less than, equal, etc.) and to search for a range of references between two values. There are also options to combine search results and restrict searches to only the references currently showing in the Library window.

Global Editing

Use the Change Text and Change Field commands to modify existing references. EndNote's Change Text and Change Field features help you to keep a clean and organized library. These commands can automatically add a term to all or some of your references. These are useful tools for labeling groups of references as well as for fixing common typos or spelling mistakes.

The Change Text ... command searches for text in your references, and either deletes or replaces it with other text. For example, you can use this command to search for a misspelled word and replace it with the correct spelling. The Change Field... command can insert text at the beginning or end of the field, replace the contents of the field with other text, or delete the entire contents of the field. For example, you can use this command to add "Reprint on File" to the Notes field of a set of references.

Sorting

In EndNote, you can sort references on any fields you choose such as Year, Journal, Title, or Label. You can sort on up to five fields in either ascending or descending order. You can also specify a custom sort order for a bibliographic style and for multiple in-text citations.

Display Font

EndNote allows you to choose a font and size for the Library window display. You may choose a smaller font if you would like to see more of the titles in the Library window, or select a font that makes the references more readable for you. In addition, you can choose the font and size for text typed into records. The font and size selected for display will not affect the actual font of the bibliographies created from word processing documents.

Create Bibliographies

if you use Microsoft Word for Windows or Macintosh or WordPerfect for Windows, please read about the EndNote Add-in. The Add-in lets you create one-step bibliographies from within these word processors. For all other word processors: To use EndNote to cite references in a paper and then create a bibliography for the paper, you simply insert the necessary citations from your EndNote database into the text of your word processing document. When you have completed the paper, you select a bibliographic style (e.g. Chicago, APA, JAMA) and tell EndNote to "format" a bibliography for the paper. EndNote scans the paper, finds the citations you pasted, modifies the in-text citations and dds a formatted bibliography to the end of your paper.

Bibliographic Styles

EndNote comes with more than 300 predefined bibliographic styles for the leading journals. You can also easily create an unlimited number of your own styles. You simply create a template that displays the reference fields and punctuation in the proper order for your style [for example: Author (Year). Title, Volume...]. Other settings let you adjust the format of the author names, page numbers, journal names, and the sort order for the references.

7 Biological Sequence Analysis

A. BLAST

1. Using the NCBI Entrez database, retrieve the following files:
   - AF390557 (GenBank nucleotide Accession number)
   - *Burkholderia pseudomallei* ABC transporter
   - 1BOUa (PDB protein structure)
   - AAC59692 (NCBI protein Accession number)
   - NP_388106 (NCBI protein Accession number)

2. Submit AF390557 using blastn, blastx. Submit the sequence using the FASTA format and a text only format. Scrutinize the results. Are the blastn and blastx results the same? Save your results in a webpage to be discussed later. (NB. These BLAST exercises use sequences already present in GenBank; therefore the first result is irrelevant to the discussions in the context of this workshop)

3. Submit NP_388106 to BLAST. Use 2 different blast programs for this exercise. Scrutinize the results and save them in a webpage.

4. For the results from these exercises, you can add your own notes to the html file for discussion. Some of the aspects that can be discussed are significance, possible gene, homology etc.

5. Data mining scenario: You are working on annotating a genome project (prokaryotic). A protein family of interest to you is available from the AF390557 GenBank file. Using a suitable blast program, attempt to find similar genes from the DNA database that your research group have just sequenced. Use the *Bacillus subtilis, Pseudomonas aeruginosa* or *Escherichia coll* genomes for this exercise.

NOTE: This approach would be useful if you have known protein sequences but do not know yet the corresponding gene encoding similar proteins in a newly sequenced genome.

6. Using the online documentation (at NCBI) as a guide: change the settings for the Advanced BLAST options. Observe differences in the results and discuss your conclusions in a html page. Explore other formatting options for the BLAST results. i.e. Increase the number of descriptions, alignments or change the alignment view.

7. Try further exercises with the sequences provided from the 'Downloads page' of the course website (Course Line).

B. Gene Prediction / Domain Assignment

1. Access the file fgidseq.txt from the CourseLine 'Downloads' page. This sequence was the result of automated DNA sequencing work on a recombinant clone carrying a genomic insert. The insert DNA is of prokaryotic origin.

2. Predict the possible genes within this sequence of DNA. Use an appropriate alignment tool to aid you in this i.e. BLAST (http://www.ncbi.nlm.nih.gov/BLAST/), HMMs (GeneMark http://ebi.ac.uk/genemark/). Tips: for a more comprehensive result, use a combination of tools. Determine the coding regions of the genes, the translated protein sequences, the protein domains and any other information that can be gleaned from the results of your analysis. Use applications such as text or html editors to manage your results.

8. Pair Wise Sequence Alignment

Aim

To perform pair wise sequence of a proteins using BLAST search tool.

Objectives

Access the BLAST sews via NCBI and NTH

Choose the current BLAST search tool to compare an unknown sequence to sequences in the database at NCBI.

To identify the probable sequence from the sequence data provided.

BLAST (Basic Local Alignment Search Tool) algorithm was described by Altschul et.al; It is very efficient pair wise sequence alignment tool. It has been optimized work on sequences from various public segment pair and defined as a pair of sub sequences of the same length from an ungapped alignment.

For a given query sequence BLAST calculations are all segment pairs b/w the query and database sequences above the scoring threshold. The algorithm searches for fixed length hits which are the resulting high scoring pairs. HSPS are produced from the optical alignments.

The best high scoring pairs are listed from BLAST 2 programmes.

Procedure

After going to NCBI website, Aelert, BLAST, select standard nucleotide BLAST or protein BLAST. Then paste query sequence. After the search has been completed, the results would be displayed. The search may take few seconds or minutes. Scroll down to distribution of BLAST hits in the query sequence. This shows graphical view of all sequences that are similar to query sequence.

Observe the sequence producing significant alignments. On the left side of the list are the links that will take you to sequence file in the database. Both the names of database and accession number are presented.

On the right hand side there is a score and e-value for each sequence. The score indicates the degree of similarity between the sequence and the query sequence; the higher the score the better the match. The e-value for a good match is usually expressed as a negative exponent which means that there is a low probability that the sequence match only by random match / chance.

II Aim

To perform alignment by using FASTA search tool

Principle

The FASTA algorithim described by Lipmann & Pearson is based on identification of short key words of K-tupples, common to both sequences under comparison. K-tupples sizes of 1-2 residues are ued in protein search, while larger k.tupples are used in DNA searcher FASTA is a local alignment tool.

Procedure

For the given query sequence, sequence comparison of k-tupples and then relative off sets between the two sequences can be viewed as focusing a diagonal matches in a dynamic programming matrix. The program uses heuristic approach to join k-tupples that lie close together on the same diagonal. The regions formed in this way contain mismatches lying b/w matches or matching k-tupples. If a significant number of matches are found, FASTA uses o dynamic programming algorithm to compute gapped alignment that incorporate the un-gapped regions.

III Aim

To perform pair wise sequence alignment using dynamic programming by S. search tool.

Function

S search does a rigorous Smith – Waterman search for similarity between a query sequence and a group of sequences of the same type (nucleic acids proteins). This may be the most sensitive method available for similarity searches. Compared to BLAST & FASTA it can be very slow.

Description

S Search uses William-Pearson implementation of Smith & Waterman to search for similarities between types as the query sequence.

9. Multiple Sequence Alignment

A. Formatting Sequences

1. BLAST the NCBI GenBank nucleotide sequence AF390557 using an appropriate BLAST program to enable you to search for similar protein sequences. (NOTE: Objective – find similar or homologous protein sequences for multiple sequence alignment)

2. Scrutinise the results. Choose results which are for proteins from the same protein family or with similar function. Pick at least 10 protein sequences by accessing the respective GenBank files via the links on the BLAST results page.

3. Prepare the chosen 10 sequences in FastA format and compile all the 10 sequences into 1 text file only. Name the file formsa_hostname.txt or any name which seems appropriate.

B. ClustalW-WWW

1.Open a site running a ClustalW server such as at BCM SearchLauncher (Baylor College of Medicine) or at EBI (European Bioinformatics Institute).

2. Fill the form provided with the FASTA formatted sequences compiled from the AF390557 BLAST results. Make sure to fill the sequences into just one form.

3. Click to submit the query 'only once'. (Clicking more than once will cause the server to be burdened with redundant jobs).

4. The results will appear as a webpage. The webpage can be saved as a html file.

C. ClusterW-Unix

1. Copy the BLAST pairwise alignments chosen from Part A. Choose 10 and save them into a single file using a text editor.

2. Using a text editor, edit the files: remove the words 'query', long annotation etc. These may interface with the alignment process. Save the file before quitting the editor.

3. Open a Unix Shell.

4. Type Align123 (or ClusterW is installed). Note: Align123 is a distribution of ClusterW by Accelrys Inc., Which is distributed with its Life Sciences package. It has a few extra features which the free ClustalW distribution does not have. However the operating procedures are essentially the same.
5. Choose "Sequence. Input from Disk" and type in the sequence name-remember that Unix is case sensitive. Select the default 'Do no remove gaps' option. (To simplify things, make sure that you are in the same Unix directory as the sequence for input)

6. Proceed to aligning the sequences by following the menu instructions and queries. Accept the default names give. The format of the output can be changed by using the

change output format option in the alignment menu. Having results in the. aln format will suffice for this exercise.

7. After the run has completed, exit Align123.

8. View the alignment file. Check whether the alignments are 'biologically sound'. If there are any contaminating sequences or input, remove them by the editing the input sequence file and run the alignment again. Refine the alignment until an optimal alignment is achieved.

D. Alignment Presentation

1. Open the BoxShade website.
2. Input the alignment sequence by using 'copy and paste' into the form provided. Select the corresponding format (for Align123 output chooses MSF, for BCM- Search Launcher results use other).

3. Select the front size as 10.

4. Select the output format. (Postscript for Unix machines, RTF_new for windows based machines).

5. Choose whether a consensus line is required, select consensus line with letter for this exercise.

6. The results can be downloaded after the submission is processed.

E. Alignment Analysis

1. Scrutinize the alignments. Are they biologically sound? Are all the aligned sequences significant?

2. Check for regions of high conservation?

3. Are there any motifs of interest? If so, what are their relevance to biological function.

10. Protein and Nucleic acid Sequence analysis with EMBOSS

Contents

- Overview
- EMBOSS key features
- What can I use EMBOSS for?
- How are the applications organized?
- EMBOSS Frequently Asked Questions
- How to cite EMBOSS
- Licensing

Introduction

EMBOSS is "The European Molecular Biology Open Software Suite ".

EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology (e.g. EMBnet) user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. EMBOSS breaks the historical trend towards commercial software packages.

The EMBOSS suite

Within EMBOSS you will find around 100 programs (applications). These are just some of the areas covered:

- Sequence alignment
- Rapid database searching with sequence patterns
- Protein motif identification, including domain analysis
- Nucleotide sequence pattern analysis, for example to identify CpG islands or repeats.
- Codon usage analysis for small genomes
- Rapid identification of sequence patterns in large scale sequence sets.
- Presentation tools for publication
- And much more.

How useful is EMBOSS?

There have been over 14 000 unique downloads of EMBOSS in the short time it has been available - many of these downloads have been performed by sysadmins setting up one installation for users all over an individual site.

The uses and interfaces to EMBOSS have long grown beyond our ability to keep track of them. EMBOSS is used extensively in production environments rather than being the sort of "research project" code that gets presented at conferences, but never actually deployed.

EMBOSS is a properly constructed toolkit for creating robust bioinformatics applications or pipelines for your work.

EMBOSS has several important advantages:

- It is free as in 'no charge'.
- It is free as in 'free speech'.

- It runs on practically every UNIX you can think of and some that you can't - plus MacOSs.
- It provides a comprehensive set of sequence analysis programs (approximately 100).
- It handles all sequence and many sequence alignment and structure formats.
- It integrates other publicly available packages.
- Encourages the use of EMBOSS in sequence analysis training.
- Encourages developers elsewhere to use the EMBOSS libraries.
- It is free of arbitrary size limits - notoriously difficult memory management in C is handled by the system, not the programmer, and memory for sequences or matrices is allocated dynamically; the only restriction is the hardware.
- It contains library functions for general string handling, pattern-matching, sorting, iteration and (imminently) extremely fast indexing.
- It contains library functions for all common sequence analysis tasks.
- It has a consistent API for interface designers to program to.

EMBOSS key features

There have been tens of thousands of unique downloads in the short time it has been available including site-wide installations by administrators catering for hundreds or even thousands of users.

The uses and interfaces to EMBOSS have long grown beyond our ability to keep track of them. EMBOSS is used extensively in production environments rather than being the sort of "research project" code that gets presented at conferences, but never actually deployed.

EMBOSS has several important advantages:

- A properly constructed toolkit for creating robust bioinformatics applications or workflows.
- A comprehensive set of sequence analysis programs.
- All sequence and many alignment and structural formats are handled.
- Extensive programming library for common sequence analysis tasks.
- Additional programming libraries for many other areas including string handling, pattern-matching, list processing and database indexing.
- It is free-of-charge.
- It is an open-source project.
- It runs on practically every UNIX you can think of and some that you can't, plus M$ Windows and MacOS.
- Each application has the same style of interface so master one and you've mastered them all.
- The consistent user interface facillitates GUI designers and developers.
- It integrates other popular publicly available packages.
- It is free of arbitrary size limits: there are no limits on the amount of data that can be processed. For the programmer, memory management for objects such as sequences and arrays is simplified.

EMBOSS is mature and stable. A major new version of EMBOSS is released each year.

What can I use EMBOSS for?

Within EMBOSS you will find around hundreds of programs (applications) covering areas such as:

- Sequence alignment,
- Rapid database searching with sequence patterns,
- Protein motif identification, including domain analysis,
- Nucleotide sequence pattern analysis---for example to identify CpG islands or repeats,
- Codon usage analysis for small genomes,
- Rapid identification of sequence patterns in large scale sequence sets,
- Presentation tools for publication,

and much more.

Popular applications include:

prophet	Gapped alignment for profiles.
infoseq	Displays some simple information about sequences.
water	Smith-Waterman local alignment.
pepstats	Protein statistics.
showfeat	Show features of a sequence.
palindrome	Looks for inverted repeats in a nucleotide sequence.
eprimer3	Picks PCR primers and hybridization oligos.
profit	Scan a sequence or database with a matrix or profile.
extractseq	Extract regions from a sequence.
marscan	Finds MAR/SAR sites in nucleic sequences.
tfscan	Scans DNA sequences for transcription factors.
patmatmotifs	Compares a protein sequence to the PROSITE motif database.
showdb	Displays information on the currently available databases.
wossname	Finds programs by keywords in their one-line documentation.
abiview	Reads ABI file and display the trace.
tranalign	Align nucleic coding regions given the aligned proteins.

EMBOSS applications

Program name	Description
aaindexextract	Extract data from AAINDEX
abiview	Reads ABI file and display the trace
acdc	Tests definition files for any EMBOSS application.
antigenic	Finds antigenic sites in proteins
backtranambig	Back translate a protein sequence to ambiguous codons
backtranseq	Back translate a protein sequence
banana	Bending and Curvature Plot in B-DNA
biosed	Replace or delete sequence sections

btwisted	Calculates the twisting in a B-DNA sequence
cai	CAI codon usage statistic
chaos	Create a chaos plot for a sequence.
charge	Protein charge plot
checktrans	ORF property statistics
chips	Codon usage statistics
cirdna	Draws circular maps of DNA constructs
codcmp	Codon usage table comparison
coderet	Extract CDS, mRNA and translations from feature tables
compseq	Counts the composition of dimer/trimer/etc words in a sequence
cons	Creates a consensus from multiple alignments
cpgplot	Plot CpG rich areas
cpgreport	Reports CpG rich regions
cusp	Create a codon usage table
cutgextract	Extract data from CUTG
cutseq	Removes a specified section from a sequence.
dan	Plot melting temperatures for DNA.
dbiblast	Database indexing for BLAST 1 and 2 indexed databases
dbifasta	Index a fasta database
dbiflat	Database indexing for flat file databases
dbigcg	Database indexing for GCG formatted databases
dbxfasta	Database b+tree indexing for fasta file databases
dbxflat	Database b+tree indexing for flat file databases
dbxgcg	Database b+tree indexing for GCG formatted databases
degapseq	Removes gap characters from sequences
descseq	Alter the name or description of a sequence.
diffseq	Find differences between nearly identical sequences
digest	Protein proteolytic enzyme or reagent cleavage digest

distmat	Creates a distance matrix from multiple alignments
dotmatcher	Produces a dotplot of two sequences.
dotpath	Displays a non-overlapping wordmatch dotplot of two sequences
dottup	DNA sequence dot plot
dreg	Regular expression search of a nucleotide sequence
einverted	Finds DNA inverted repeats
embossdata	Finds or fetches the data files read in by the EMBOSS programs
embossversion	Writes the current EMBOSS version number
emowse	Protein identification by mass spectrometry
emma	Multiple alignment program
entret	Reads and writes (returns) flatfile entries
epestfind	Finds PEST motifs as potential proteolytic cleavage sites
eprimer3	Picks PCR primers and hybridization oligos
equicktandem	Finds tandem repeats
est2genome	Align EST and genomic DNA sequences
etandem	Looks for tandem repeats in a nucleotide sequence.
extractfeat	Extract features from a sequence
extractseq	Extract regions from a sequence.
findkm	Calculates Km and Vmax for an enzyme reaction
freak	Residue/base frequency table or plot
fuzznuc	Nucleic acid pattern search
fuzzpro	Protein pattern search
fuzztran	Protein pattern search after translation
garnier	Predicts protein secondary structure
geecee	Calculates the fractional GC content of nucleic acid sequences
getorf	Finds and extracts open reading frames (ORFs)
helixturnhelix	Finds nucleic acid binding domains.
hmoment	Hydrophobic moment calculation

iep	Calculates the isoelectric point of a protein
infoalign	Information on a multiple sequence alignment
infoseq	Displays some simple information about sequences
isochore	Plots isochores in large DNA sequences
jembossctl	Jemboss Authentication Control
lindna	Draws linear maps of DNA constructs
listor	Writes a list file of the logical OR of two sets of sequences
marscan	Finds MAR/SAR sites in nucleic sequences
maskfeat	Mask off features of a sequence
maskseq	Mask off regions of a sequence.
matcher	Local alignment of two sequences
megamerger	Merge two large overlapping nucleic acid sequences
merger	Merge two overlapping sequences
msbar	Mutate sequence beyond all recognition
mwcontam	Shows molwts that match across a set of files
mwfilter	Filter noisy molwts from mass spec output
needle	Needleman-Wunsch global alignment.
newcpgreport	Report CpG rich areas
newcpgseek	Reports CpG rich regions
newseq	Type in a short new sequence.
noreturn	Removes carriage return from ASCII files
notseq	Excludes a set of sequences and writes out the remaining ones
nthseq	Writes one sequence from a multiple set of sequences
octanol	Displays protein hydropathy
oddcomp	Finds protein sequence regions with a biased composition.
palindrome	Looks for inverted repeats in a nucleotide sequence.
pasteseq	Insert one sequence into another.
patmatdb	Matching a Prosite motif against a Protein Sequence Database.

patmatmotifs	Compares a protein sequence to the PROSITE motif database.
pepcoil	Predicts coiled coil regions
pepinfo	Plots simple amino acid properties in parallel
pepnet	Protein helical net plot
pepstats	Protein statistics
pepwheel	Shows protein sequences as helices
pepwindow	Displays protein hydropathy
pepwindowall	Displays protein hydropathy of a set of sequences
plotcon	Plots the quality of conservation of a sequence alignment
plotorf	Plot potential open reading frames
polydot	Multiple dotplot
preg	Regular expression search of a protein sequence
prettyplot	Displays aligned sequences, with colouring and boxing.
prettyseq	Output sequence with translated ranges
primersearch	Searches DNA sequences for matches with primer pairs
printsextract	Preprocesses the PRINTS database for use with the program PSCAN
profit	Scan a sequence or database with a matrix or profile
prophecy	Creates matrices/profiles from multiple alignments
prophet	Gapped alignment for profiles
prosextract	Extracts ID, AC, and PA lines from the PROSITE motif database.
pscan	Locates fingerprints (multiple motif features) in a protein sequence.
psiphi	Calculates phi and psi torsion angles from cleaned EMBOSS-style protein co-ordinate file
rebaseextract	Extract data from REBASE
recoder	Find and remove restriction sites but maintain the same translation
redata	Isoschizomers, references and Suppliers for Restriction Enzymes
remap	Display a sequence with restriction cut sites, translation etc..
restover	Finds restriction enzymes that produce a specific overhang

restrict	Finds Restriction Enzyme Cleavage Sites
revseq	Reverse and complement a sequence.
seealso	Finds programs sharing group names
seqmatchall	Does an all-against-all comparison of a set of sequences
seqret	Reads and writes (returns) a sequence.
seqretsplit	Reads and writes (returns) sequences in individual files
showdb	Displays information on the currently available databases
showalign	Display a multiple sequence alignment
showfeat	Show features of a sequence.
showorf	Pretty output of DNA translations
showseq	Display a sequence with features, translation etc
shuffleseq	Shuffles a set of sequences maintaining composition
sigcleave	Predicts signal peptide cleavage sites
silent	Silent mutation restriction enzyme scan
sirna	Finds siRNA duplexes in mRNA
sixpack	Display a DNA sequence with 6-frame translation and ORFs
skipseq	Reads and writes (returns) sequences, skipping the first few
splitter	Split a sequence into (overlapping) smaller sequences.
stretcher	Global alignment of two sequences.
stssearch	Searches a DNA database for matches with a set of STS primers
supermatcher	Finds a match of a large sequence against one or more sequences
syco	Synonymous codon usage Gribskov statistic plot
tcode	Fickett TESTCODE statistic to identify protein-coding DNA
textsearch	Search sequence documentation text. SRS and Entrez are faster!
tfextract	Extract data from TRANSFAC
tfm	Displays a program's help documentation manual
tfscan	Scans DNA sequences for transcription factors.
tmap	Predict transmembrane proteins

tranalign	Align nucleic coding regions given the aligned proteins
transeq	Translates nucleic acid sequences.
trimest	Trim poly-A tails off EST sequences
trimseq	Trim ambiguous bits off the ends of sequences
twofeat	Finds neighbouring pairs of features in sequences
union	Reads sequence fragments and builds one sequence
vectorstrip	Strips out DNA between a pair of vector sequences
water	Smith-Waterman local alignment.
whichdb	Search all databases for an entry
wobble	Wobble base plot
wordcount	Counts words of a specified size in a DNA sequence.
wordmatch	Finds all exact matches of a given size between 2 sequences
wossname	Finds programs by keywords in their one-line documentation.
yank	Reads a range from a sequence, appends the full USA to a list file

Summary

EMBOSS is "The European Molecular Biology Open Software Suite". EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology (e.g. EMBnet) user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. EMBOSS breaks the historical trend towards commercial software packages.

11. Analysis of Protein Structure Using RasMol

Goal

The goal of this lab is to examine all levels of protein structure and, in particular, the structure of an α-amylase protein using the computer program RasMol.

Objectives

After completing Lab 4, you will be able to

Identify which amino acid reduce contributes the carboxyl group and which contributes the amino group to a peptide bond.

Determine the approximate mass of given polypeptide.

identify and discuss the four levels of protein structure

use the RasMol computer software to visualize a given protein and identify the components of its three-dimensional structure

Use the RasMol computer software to visualize an α-amylase protein and identify its active site, inhibitor, and associated ions.

Background protein structure

Proteins are polymers, or chains, whose subunits comprise a particular class of organic molecules called amino acids. There are typically 20 different amino acids that form amino acid polymers. These amino acids are joined together by strong covalent bonds called peptide bonds, and the protein polymers are called polypeptides. A peptide bond forms between the carboxyl group on one amino acid and the amino group on the other (Fig). They symbol “R” represents the side chains of the amino acids. This polypeptide chain constitutes the primary structure of the protein, and the particular amino acid composition of each polypeptide determines the final biochemical characteristics of the protein. Polypeptide chains can have only a few amino acid residues or as many as several thousand amino acid residues. The mean molecular weight of an amino acid is 110; thus, the molecular weight of a polypeptide that has 1000 amino acid residues would be about 110,000. This is usually expressed in units called Daltons, which are equal to one atomic mass unit. Therefore, the mass of a protein with 1000 amino acid residues is about 110,000 daltons, or 110 kd (kilodaltons).

Polypeptides do not exist as long, straight chains, however. Rather, each polypeptide chain folds into a distinct three-dimensional shape or conformation, which is determined by its primary amino acid sequence. Chemical interactions between amino acids, which may be near each other but may also be distant from each other in the primary sequence, contribute to the folding of a protein. The regular repeating patterns of twists or kinks of a polypeptide chain constitute the protein’s secondary structure. Two common secondary structures are the α-helix and the β-sheet. These structures arise from regular hydrogen bonding between carbonyl and amino groups of particular amino acid residues within a contiguous stretch of polypeptide chain.

Additional interactions between nonadjacent amino acid side chains cause the protein to fold into other three-dimensional shapes. This third level of protein structure is called the tertiary structure. In addition to hydrogen bonds, there are electrostatic interactions between charged side chains and hydrophobic interactions between non-polar side chains. For example, amino acid residues with hydrophobic side chains will generally interact with each other to avoid the aqueous environment in which they are dissolved. Conversely, amino acids with hydrophilic side chains will tend to remain on the outside of the protein, hydrogen bonding with the water molecules of the solution. All these weak, non-covalent bonds contribute to the specific shape a protein assumes. Covalent bonds can also stabilize the protein structure. The amino acid cysteine contains sulfur, and the sulfurs in two cysteine residues can form a covalent bond called a disulfide bridge.

In addition to chemical interactions within a polymer, there may also be chemical interactions between different polypeptides allow proteins to form more complex structures, called the quaternary structure. Like the bonds that form secondary and tertiary structures, the bonds that form the quaternary structure are typically weak chemical bonds. The chemical bonds that form the quaternary structure are between amino acids on different polypeptide chains.

The α-amylase protein you will examine is from pig (porcine) pancreas; it consists of a single polypeptide chain containing 496 amino acid residues. The amino acid sequence of porcine α-amylase is provided in Fig.4.2. Each amino acid residue is identified by its position in the polypeptide chain (e.g., glnl, tyr2, ala3, pro4, etc.). A polypeptide chain is always written starting with the terminal amino acid that has a free amino group, called the amino terminal end of the protein, and ends with the terminal amino acid that has a free carboxyl group, called the carboxyl terminal end of the protein.

Enzyme function

Enzymes are usually proteins that function as metabolic catalysts; they enhance the rate of a reaction by lowering the activation energy. Reactions between molecules occur only when substrate atoms collide with some minimum amount of energy. Enzymes act to decrease this necessary energy by interacting with substrate molecules at the active site on the enzyme. The binding of the substrate to the active site is very specific and precise. Particular amino acids in the active site interact chemically with the substrate, straining the bonds between substrate atoms and making it easier to break and make bonds.

Alpha-amylase enzymes are part of a super family of enzymes that have a distinctive conformation called an $(\alpha/\beta)_8$ barrel. In these proteins, eight parallel strands of β-sheet coil to form a central barrel, which is surrounded by eight α-helices (Fig 4.3). The active site of α-amylase lies at one end of the barrel, and five glucose residues of a starch polymer fit into the active site (cleavage occurs between the third and fourth residues). Alpha-amylase also has a chloride ion and a calcium ion that are important for its function. The calcium ion helps the protein fold and maintain its shape while the chloride ion interacts with the substrate at the active site.

Background questions

1. Give the abbreviated name and position number of the amino acid residue at the amino terminal end and the amino acid residue at the carboxyl terminal end of porcine α-amylase

2. Amino acids pro4 and gln5 of α-amylase form a peptide bond. Which amino acid contributes the amino group and which contributes the carboxyl group of the peptide bond between them?
3. The amino acid cysteine contains sulfur. The sulfurs in two cysteine residues can form a covalent bond called a disulfide bridge that contributes to the tertiary structure of a protein. Identify, by number, two amino acids in porcine α-amylase that might form a disulfide bridge.
4. What is the approximate mass of porcine α-amylase?
5. Define active site. Describe the active site of α-amylases.

Laboratory overview

RalMol is program that converts the crystallographic data used to determine the three-dimensional structure of a protein into a computer image. This public domain software is available at http://www.umass.edu/microbio/rasmol/. The protein files required for this lab can be downloaded from http://www.rcsb.org/pdb or from this manual's website at http://www.mhhe.com/theil. The script file, which contains specific information regarding the α-amylase protein, must be downloaded from this manual's website (see preceding address).

The RasMol program uses a combination of menu items (chosen with the mouse) and typed commands to view particular aspects of protein structure. In Parts I, II, and III you will explore the primary, secondary, and tertiary structure of crambin, a very small protein from plants. In Parts IV, V, and VI, you will analyze the structure, active site, and disulfide bridges in the porcine α-amylase protein.

Two other protein analysis programs, Chime and Protein Explorer, are also available at http://www.umass.edu/microbio/rasmol/. Both are RasMol derivatives and display macromolecular structures in three dimensions.

Timeline

This computer lab takes 2-3 hours; it can be done in class or assigned as homework.

Procedure

Using a Macintosh to view proteins with RasMol

1. The RasMac icon has three circles arranged in a triangle. Follow directions given by your instructor for opening the program from the desktop.
2. There should be two windows: one with a black background titled "RasMol", the other a white window titled "RasMol Command Line". Adjust the size of the windows by clicking on and dragging (up, down, left, right) the small box on the lower right-hand corner of each window. The white RasMol Command Line window should cover about one-third of the left of the screen, and the black RasMol window should cover about two-thirds of the right side of the screen.
3. Go to step 4 of the PC instructions that follow.

Using A PC to view proteins with RasMol

From the Start Menu, open RasMol. The black "RasMol" window with the menu items above it will be displayed.

On the Task Bar at the bottom of the screen, the item "RasMol Command" will be displayed. Open this by clicking on it. This will open a white window with a prompt line that accepts typed commands.

Adjust the size of the two windows so that both can be seen at the same time. You may need to move the RasMol Command Line window slightly to the left inn order to reduce it from the right side. Move the mouse to the outside line of the window that you want to change until you see a double-ended arrow. Click and hold the mouse while you drag side of the window to the size desired. The RasMol Command Line window should cover about one-third of the left of the screen, and the black RasMol window should cover about two-thirds of the right side of the screen.

In the instructions that follow, the left side of the instruction table presents the commands that should be typed in the RasMol Command Line window. The right side of the table comments on (explains) the purpose of each command. Type only the command, not the comment. After typing the command, push the "Enter" key on the keyboard. To minimize errors, type commands carefully, and make sure spaces are placed only where shown in the commands. bulleted items direct you to use the mouse to choose menu items or manipulate the image.

By the end of these exercises, you will be very familiar with many of the RasMol commands. If you want more information, consult the RasMol User Manual, available from the Help menu. You can also type "help", followed by the topic name in the Command Line window, to get information presented there.

Part-I Primary Protein Structure

Commands-type text below and press "enter	Comments
Load crambin.pdb	Loads the protein data bank (pdb) file for crambin

- View the crambin protein using the different options under the Display menu on the black "RasMol" window. A good choice from the Color menu is "structure".

1. Which view option in the Display menu gives the most chemical detail? ______________________________
2. Which view option gives the best display of the secondary structure, such as helices (corkscrew-like structure)?_________________________
3. Which view option gives the most realistic display in terms of the actual relative size of the atoms of the molecule ?___________

- Select "ball & stick" from the Display menu and "cpk" from the Colors menu.

You now have a chemical view of crambin. Every knob represents an atom, and the sticks represent covalent bonds. Each atom is colored by chemical convention. In cpk, gray=carbon; red=oxygen; blue=nitrogen; and yellow=sulfur (hydrogen are not shown). You can use the mouse to click on an atom. The Command Line will tell you the atom and the molecule in which the atom is located. The atom letters are C,O,N and S (sometimes these letters are followed by other letters, such as A, B, D1, etc.). The molecule is also called the "group" and has the three-letter abbreviation for the amino acid and its number. The amino acids in crambin will now be studied in more detail.

Restrict leu	Displays only leucine residues

- Use the mouse to click on each atom and identify it.

4. How many carbons are there in leucine? How many oxygens? How many nitrogens? How many sulfurs? _________________

* Next you will view the arginine residue next to leucine 18.

Select arg 17	Selects arginine 17

- To view the molecule, choose “ball & stick” from the Display menu. Use the mouse to click on each atom and identify it.

5. How many carbons are there in arginine? How many oxygens? How many nitrogens? How many sulfurs? _________________

- Next you will look at both arg 17 and leu 18; note that they are not joined by a peptide bond. The following commands will join them.

Select arg17, leu18	Selects arginine 17 and leucine 18

- To see the peptide bond form, you need to have a good view of both molecules and a good view of the small gap between them. Rotate the molecules until there is a good view of the gap.
- To see the peptide bond form, choose “ball & stick” from the Display menu and watch the gap between the two amino acids disappear as they are joined by a peptide bond. (If you did not see it, go back to the command “restrict leu” and try again)

6. Which two specific atoms (and from which amino acids) were joined to make the peptide bond? _____________________

Select all	Selects all the amino acids

- Choose “ball & stick” from the Display menu.

Families of amino acids can be selected with RasMol. They include acidic, basic, aromatic (side chains having ring structures), and aliphatic (side chains without ring structures) amino acids.

- Identify the amino acids in crambin that belong to each family, and place your results in Table 4.1. The commands that must be typed in the Command Line window are shown as follows. Select, color, and restrict one family of amino acids at a time. Display each using the “ball & stick” view from the Display menu. It is not necessary to list the group numbers or to list a particular amino acid more than once per family.

Select acidic (or basic, or aromatic, or aliphatic)	Selects only acidic (basic, aromatic, or aliphatic) amino acid residues
Color orange (or blue, or magenta, or green)	Colors acidic (basic, aromatic, or aliphatic) amino acid residues organge (blue, magenta, or green)
Restrict acidic (or basic, or aromatic, or aliphatic)	Selects only acidic (basic, aromatic, or aliphatic amino acid residues.

AMINO ACIDS OF EACH FAMILY PRESENT IN CRAMBIN

Amino Acid Family	Number of Amino Acids in Each Family Present in Crambin	Names of the Amino Acids in Each Family Present in Crambin
Acidic		
Basic		
Aromatic		
Aliphatic		

- To view the entire protein, do the following.

Select all	Selects all the amino acids

- Choose “ball & stick” from the Display menu.
- Choose “cpk” from the Colors menu.

Part II : Secondary protein structure

The following instructions assume that Part II directly follows Part I. If this is a new session, the first command to be typed should be “load crambin. Pdb,” as in Part I. Colors should be “cpk”. When typing commands, be sure to leave a space where a space is indicated.

Two major types of secondary structures can be identified in the crambin protein: the α-helix and the β-sheet.

Select helix	Selects all the helical regions of crambin
Color blue	Colors the helices blue

- The α-helix is most easily seen in “ribbons” or “strands” views. Select “strands” if your computer is slow.
- Rotate the molecule by clicking and dragging the mouse or by using the scroll bars.

How many helices do you see?

- Rotate the molecule to see how the helices are joined together at one end by a short stretch of amino acids that is at right angles to the helices.
- Select “ball & stick” from the Display menu. The amino acids of the α-helices will remain blue.
- Find amino acids 19 and 22 (19 is at the end of the longer α-helix, and 22 is in the loop between the two helices).

What is the name of the amino acid at these locations?
Describe the side chain of this amino acid.

Select pro	Selects only proline residues
Restrict pro	Shows only proline residues

How many proline residues are found in crambin ?

How many are part of an α-helix (colored blue) ?
How many are in the middle of an α-helix ?
Proline residues tend not to be found in the middle of α-helices.

Select helix	Selects all the α-helices of crambin
Restrict helix	Shows only the α-helices

- Select "ball & stick" from the Display menu. Select "group" from the Colors menu. This colors the residues of the α-helices according to their position in the α-helix and will make it easier to distinguish the other residues in each α-helix. Click on the amino acid residues of the longer α-helix, one at a time.

How many amino acids are in this longer α-helix ? _________________
Give the name and number of any three amino acids in the longer α-helix. Alert: Make sure you have the Command Line window open wide enough or you may not see the entire group number. ______________________________

Select all	Selects all the atoms

- Select "ball & stick" from the Display menu.

Select sheet	Selects all sheet structures of crambin
Color yellow	Colors the sheets yellow

- Select "ribbons" from the Display menu.

Select helix	Selects all the helices of crambin
Colour blue	Colors the helices blue

- Select "ribbons" from the Display menu.

How many sheet structures do you see ?

Select helix, sheet	Selects both the helices and the sheets
Hbonds on	Turns on the hydrogen bonds in the helices and sheets

Where are the hydrogen bonds in the β-sheets, among the atoms of one strand or between the atoms of parallel strands ?
Where are the hydrogen bonds in the α-helices?

Select all	Selects all the amino acids
Hbonds off	Turns off the hydrogen bonds

- Select "ball & stick" from the display menu and "cpk" from the Colors menu.

Part III: Tertiary Protein Structure

The following instructions assume that Part III directly follows Part II. If this is a new session, the first command to be typed should be “load crambin.pdb”, as in Part I. Colors should be “cpk”. When typing commands, be sure to leave a space where a space is indicated.

The sulfur atom in one cysteine residue can join with the sulfur atom in another cysteine residue to form a disulfide bond. Disulfide bonds contribute to the tertiary structure of proteins.

- Use the Display menu to select “wireframe” structure and the Colors menu to select “cpk” colors.

Select cys	Selects all the cysteine residues

- Use the Display menu to select “ball & stick”.
- Rotate the molecule. Note the position of each cysteine residue.

s-s bonds on	Makes the disulfide bonds between cysteine residues visible

- Look for the yellow dashed lines showing the disulfide bridges between cysteine residues. You can try coloring the disulfide bonds other colors to make them more visible (the command is “color s-s bonds”; then type the name of the color). The cysteine residues and their disulfide bridges can be seen more easily if the rest of the molecule is removed and the cysteine residues are enlarged.

Restrict cys	Shows only the cysteine residues and disulfide bonds
Zoom 150 (or zoom 125)	Magnifies the image

- Identify the cysteine residues that are paired by disulfide bonds and identify the sulfur atoms in each disulfide bond. Use the mouse to click on each molecule or atom, and the corresponding name and position number will be displayed in the Command Line window. Enter the position numbers of each cysteine molecule pair in the first column of Table 4.2, and enter the position numbers of the sulfur atoms of the pair in the second column of Table 4.2.

Table 4.2
Cysteine Molecule Pairs and Their Sulfur Atoms

Cysteine Molecule Pairs	Sulfur Atoms of S-S Bonds

- Choose one cysteine residue from each pair that is joined by a disulfide bond, and color that one green. An example follows. Do this for each of the three pairs.

Select cys40	Selects cys40
Color green	Colors cys40 green

1. Two of the cysteine residues in crambin are joined by a covalent peptide bond rather than by a disulfide bond. Which two are these? ______________________
2. Identify the atom (by name and position) in each of these cysteine molecules that is part of the peptide bond. ______________________________

Part IV: Protein Structure of α-Amylase

Now you will view the porcine α-amylase protein.

zap	Removes all the information, including the file
Load amypig.pdb	Loads the protein data bank file for α-amylase
Script amylase.txt	Loads the script file (contains information specific to α-amylase protein)
Select helix	Selects all the helical regions of α-amylase
Color blue	Colors the helices blue
Hbonds on	Displays the hydrogen bonds in the helices

- Select “ribbons” from the Display menu; if your computer is slow, choose “strands”.

Select sheet	Selects all the sheet structure regions of α-amylase
Color yellow	Colors the sheets yellow
Hbonds on	Displays the hydrogen bonds between sheets
Restrict helix, sheet	Removes everything but the helices and sheets

- Select “ribbons” from the Display menu. Rotate the molecule so that you see down the barrel of the $(\alpha/\beta)_8$ configuration.

Eight β-strands should be in the center of the barrel, with eight helices around the outside. (Some β-strands remain on the outside of the barrel.)

Select active	Selects the active site of α-amylase; defined in the amylase.txt file (based on published information)

- Select “ball & stick” from the Display menu.

Cpk 200	Displays the active-site residues in ball & stick form at size 200
Color green	Colors the active-site residues green
Zoom 150	Magnifies the region of interest

- Select “spacefill” from the Display menu.

Select all	

- Select “spacefill” from the Display menu. Use the Options menu to turn off the heteroatoms, mostly water that obscures the view. Just click on “heteroatoms” with the mouse.
- Try other views under the Display menu to help see the active site in relation to the structure of the protein.

Part V: Active-Site Region of α-Amylase

The following instructions assume that Part V directly follows Part IV. If this is a new session, replace the first three commands that follow with the first three commands of Part IV.

reset	Resets to normal view
Select all	Selects everything
Hbonds off	Turns off the hydrogen bonds

- Select “wireframe” from the Display menu.

One way to study the active site is to use an inhibitor molecule that binds to the enzyme. Since the inhibitor cannot be cleaved, it becomes trapped in the active site. The inhibitor present in the active site of α-amylase is a nonhydrolyzable disaccharide called daf (daf=1,4-deoxy-4-(5-hydroxymethyl-2,3,4-trihydroxycyclohex-5, 6-enyl) amino fructose). The non-hydrolyzable disaccharide has two glucose (glc) residues attached to one end and one glc molecule attached to the other end. Thus, the nonhydrolyzable disaccharide occupies the third and fourth positions of the active site.

Select glc.daf	Selects the inhibitor molecule (which contains glc and daf) blocking the active site.

- Select “ball & stick” from the Display menu.

In addition to certain amino acids that are important for enzyme activity, α-amylase contains two ions, Ca^{2+} and Cl^{-}, which are required for proper function. Because these atoms are difficult to see, they will be enlarged here to help you find them.

Cpk200	Sets the inhibitor size at 200
Color cyan	Colors the inhibitor cyan
Select Cl	Selects the chloride ion
Cpk300	Sets the chloride ion size at 300
Color magenta	Colors the chloride ion magenta
Select Ca	Selects the calcium ion
Cpk 300	Sets the calcium ion size at 300
Color red	Colors the calcium ion red
Zoom 150	Magnifies the region of interest
Select amino	Selects all amino acid residues

- Rotate the image to help see the position of the heteroatoms (i.e., inhibitor, Cl^{-}, Ca^{2+}, water) in the protein.

Restrict hetero	Restricts the atoms to heteroatoms
Select active	Selects the active site

- Select “ball & stick” from the Display menu.

Cpk 200	Sets the size at 200

- Use the mouse to click on the amino acid groups on the screen; the Command Line will identify each amino acid residue by number.

1. Identify the 13 amino acid residues at the active site by name and number. Alert: Make sure the Command Line window is open wide enough to show the entire group number. ____________________
2. Using Table 4.3, determine how many of the 13 amino acids at the active site are non-polar, polar, and electrically charged.
   Non-polar : _______________
   Polar : _______________
   Charged : _______________

TABLE – 4.3
Properties of Amino Acid Side Chains

Properties of Side Chains (R groups)	Amino Acids
Nonpolar	Gly,ala,val,leu,ile,met,phe,trp,pro
Polar	Ser, thr, cys, tyr, asn, gln
Electrically charged	Asp, glu, lys, arg, his

*To see the $(\alpha/\beta)_8$ barrel formation along with the cleft in the active site, do the following steps.

Restrict dom-a	Restricts view to only the (α/β)8 barrel formation
Color yellow	Colors $(\alpha/\beta)_8$ barrel yellow

- Select “ribbons” from the Display menu. The helices and sheets of the $(\alpha/\beta)_8$ barrel formation are now visible. Rotate the image to see down the middle of the barrel, through the hole in the center.
- Select “spacefill” from the Display menu and notice how the hole in the center disappears.

Select active	Selects the active-site amino acids
Color red	Colors the active site red

- Rotate the image until the cleft of the active site is clearly visible. (The active site may be on the opposite side).

Select glc, daf	Selects inhibitor molecule
Color green	Colors the inhibitor green

- Select “spacefill” from the Display menu and watch as the inhibitor fills the active-site cleft. Rotate the image to see how the inhibitor fits within the active-site region.

Part VI: Disulfide Bridges in α-Amylase
The following instructions assume that Part VI directly follows Part V. If this is a new session, replace the first two steps that follow with the first three steps of Part IV.

Reset	Resets normal view
Restrict none	Unselects everything
Select all	

*Use the menus to select “wireframe” structure and “cpk” colors.

Select cys	Selects all the cysteine residues
Color yellow	Colors the cysteine residues yellow
Cpk 150	Enlarges the cysteine residues
Zoom 150	Magnifies

- Rotate the molecule. Note the position of each cysteine residue by clicking on it to see the group number.

1. Are cysteine residues that are close in the tertiary structure always close together in the primary structure ? ___________________________

Select helix	Selects all the α-helices of crambin

Ssbonds on	Makes the disulfide bonds visible
Color ssbonds green	Colors the disulfide bonds green
Zoom 200	Magnifies

- Look for the green dashed lines showing the disulfide bridges between cysteine residues. (If green does not show up well on your screen, color the disulfide bonds another color with the command “color s-s bonds yellow”, for example)
- The cysteine residues and the disulfide bridges can be seen more easily if you remove the rest of the molecule with the following steps.

Restrict cys	Shows only cysteine residues and disulfide bonds
Color cpk	Colors the atoms with conventional colors
Zoom 125	Magnifies

- Note: You may try other values for “zoom” to make the cysteine residues easier to see and count.
- Use the mouse and Command Line window to identify each cysteine residue by its position number. Choose one cysteine residue from each pair, and color that one yellow. Leave the other one in the pair in cpk colors. An example follows.

Select cys 115	Selects cys 115
Color yellow	Colors cys 115 yellow

- Continue as in the preceding example for each pair.

2. How many cysteine residues are in porcine α-amylase ?
3. Which atom of the cysteine molecule participates in disulfide bond formation?
   ____
4. Are all the cysteine molecules in pairs held by disulfide bonds ? ______ If not, which cysteine residues (by number) are not ?
   ____
5. Do you think it is possible to predict from just the primary amino acid sequence which cysteine residues will form disulfide bonds ? Why?
   ____

11. ANALYSIS OF DNA STRUCTURE USING RASMOL

Goal

The goal of this laboratory is to examine the structure of DNA using the computer program RasMol.

Objectives

After completing Lab 6, you will be able to

1. identify the nucleotides of DNA as purine or pyrimidine
2. describe the complementary and anti-parallel nature of two strands of DNA.
3. describe the formation of phosphodiester bonds.
4. identify the major and minor grooves on a DNA molecule
5. describe one type of protein-DNA interaction.

Background

The hereditary material of most organisms is made of DNA, or deoxyribonucleic acid, which is a polymer of nucleotides. A nucleotide comprises a purine base (adenine, guanine) or a pyrimidine (cytosine, thymine, uracil) base linked to a sugar residue (ribose or deoxyribose) on which one or more phosphate groups are attached (Fig.6.1). In DNA, the deoxynucleotides, adenosine, thymidine, guanosine, and cytidine, are joined together by covalent bonds, called phosphodiester bonds. This is a phosphate ester linkage between the 5'-phosphate group on the sugar residue of one nucleotide and the 3'-hydroxyl group on the sugar residue of the next nucleotide (Fig.6.2). DNA exists as a double-stranded molecule in which the bases of one strand are hydrogen bonded, or base paired, with the bases of the other strand. This base pairing is complementary; adenine (A) always pairs with thymine (T), and guanine (G) always pairs with cytosine (C). the A-T base pair has two hydrogen bonds, and the G-C base pair has three hydrogen bonds (Fig. 6.3). The two strands of DNA form a helical structure, the double helix described by Watson and Crick nearly 50 years ago. The two strands of the double helix are organized in opposite orientations, such that the 5' end of one strand is aligned with the 3' end of the other strand. RNA, or ribonucleic acid, is similar in structure to DNA, except ribose is the sugar moiety the nucleotide uridine is used rather than thymidine, and RNA is single stranded.

Backround question

What is a nucleotide?

How does the structure of purine bases differ from that of pyrimidine bases?

How does the base pairing between G and C differ from the pairing between A and T ?

Explain why the two strands of DNA are complementary.

Explain the anti-parallel nature of double-stranded DNA.

Laboratory overview

In this laboratory, the structure of DNA will be explored using the computer program RasMol. You will use many of the same commands and operations you used to view protein structure in Lab 4. The pdb files needed for this lab can be down-loaded from this manual's website at http://www.mhhe.com/thiel. (Refer to the Bioinformatics section of Appendix I for information regarding other DNA manipulation programs.

Also, the Chime and Protein Explorer programs mentioned in Lab 4 can be used to visualize DNA structure).

In Parts I through IV, you will explore nucleotide structure, phosphodiester linkages, and the complementary and anti-parallel nature of double-stranded DNA. In Part V, you will examine interactions between DNA and a DNA-binding protein.

Timeline

Procedure

Using a Macintosh to view DNA with RasMol

The RasMac icon has three circles arranged in a triangle. Follow directions given by the instructor for opening the program from the desktop.

There should be two windows: one with a black background titled “RasMol”; the other with a white background titled “RasMol Command Line”. Adjust the size of the windows by clicking on and dragging (up, down, left, right) the small box on the lower right-hand corner of each window. The white Command Line Window should cover about one-third of the left of the screen, and the black RasMol window should cover about two-thirds of the right side of the screen.

Go to step 4 of the PC instructions that follow.

Using A PC to view DNA with RasMol

1. From the Start Menu, open RasMol. The black “RasMol” window with the menu items above it will be displayed.
2. On the Task Bar at the bottom of the screen, “RasMol Command” will be displayed. Open this by clicking on it. This will open a second white window with a prompt line that accepts typed commands.
3. Adjust the size of the two windows so that both are seen at the same time. You may need to move the Command Line window slightly to the left in order to reduce it from the right side. Move the mouse to the outside line of the window that you want to change until you see a double-ended arrow. Click and hold the mouse while you drag the side of the window to the size desired. The white Command Line window should cover about one-third of the left of the screen, and the black RasMol window should cover about two-thirds of the right side of the screen.
4. In the instructions that follow, the left side of the instruction table presents the commands that should be typed in the Command Line window. The right side of the table gives comments on (explains) the purpose of each command. Type only the command, not the comment. After typing the command, push the “Enter” key on the keyboard. TO MINIMIZE ERRORS, TYPE COMMANDS CAREFULLY, AND MAKE SURE SPACES ARE PLACED ONLY WHERE SHOWN INN THE COMMANDS. BULLETED ITEMS DIRECT YOU TO USE THE MOUSE TO CHOOSE MENU ITEMS OR MANIPULATE THE IMAGE.

Part I : DNA structure

- Open the file “dnal.pdb”. Turn the DNA molecule to view it from the side rather than from the end. Use the Display and Colors menu items to view the molecule in different forms and colors.
- Select “ball & stick” from the Display menu and “cpk” from the Colors menu. (In cpk colors, red=O, blue=N, gray=C, and yellow=P.) Turn off the heteroatoms by clicking “heteroatoms” under the Options menu.

Commands – type text below and press “enter”	Comments
Select a	Selects all of the deoxyadenosine residues
Color blue	Colors them blue
Select g	Selects all of the deoxyguanosine residues
Color green	Colors them green
Select t	Selects all of the deoxythymidine residues
Color yellow	Colors them yellow
Select c	Selects all of the deoxycytidine residues
Color red	Colors them red

1. How many A residues are there in this strand of DNA? __________
2. How many G residues are there in this strand of DNA? _________
4. How many T residues are there in this strand of DNA? __________

Select nucleic	Identifies the molecule as nucleic acid for the next step
Select backbone	Identifies the sugar-phosphate backbone

- Display as “ribbons”.

5. How many sugar-phosphate backbones are there in DNA? ________

Recall that the backbones contain alternating phosphate and sugar residues linked with phosphodiester bonds.

- Select “ball & stick” from the Display menu and “cpk” from the Colors menu.

Restrict 1,17,18,24	Identifies two pairs in the DNA helix
Zoom 150	Magnifies the image
Hbonds on	Shows the hydrogen bonds for each pair

6. Identify the two nucleotides in each pair by letter and number ____

- If you cannot remember which color represents which base, click on the base with the mouse, and letter and number of the molecule (e.g., G24) will appear in the command line. It is also possible to identify each atom in the molecule by clicking directly on the atom. This may be easier to see if you zoom to 200.

Reset	Clears everything
Select all	Selects everything

- View the molecule as “sticks” from the Display menu.

Hbonds on	Turns on all of the hydrogen bonds

Zoom 150	Magnifies the image

- Rotate the molecule to see the hydrogen bonds.

7. Is there a consistent number of hydrogen bonds between certain pairs of nucleotides? _________ If so, what general pattern is seen?
    - View using "chain" under the Colors menu.
    - Find and identify the major and minor grooves. These appear as hollows or curve where there are no molecules. If you trace each hollow around the three-dimensional structure, you will find that both the large hollow (the major groove) and the smaller hollow (the minor groove) trace a helical path. It may help to change the display to "spacefill" to help see the grooves. Rotate the molecule until they are visible. Turn off the hetero atoms under the Options menu (you may need to turn them on and then off to get rid of them).

Zap	Closes the file

Part II: Nucleotide Structure

- Open the file "dnal. Pdb". Display it in "ball & stick" with "cpk" colors and the heteroatoms off (under the Options menu). (In cpk colors, red=O, blue = N, gray=C, and yellow=P).

First you will examine the chemical structure of adenosine.

Restrict a22	Shows only nucleotide A22
Zoom 200	Magnifies the image

1. Identify the purine base that contains two nitrogenous rings. Draw a picture of this purine molecule (adenine).
    - Rotate the nucleotide to clearly view the 5-carbon deoxyribose ring.
2. Draw a picture of the deoxyribose ring and label each carbon (1'-5').
3. Adenine is attached to which carbon in the deoxyribose ring? ___________________________________
4. Find the 3' carbon of the deoxyribose ring. Does it have oxygen attached? ______________________________
5. Find the 2' carbon of the deoxyribose ring. Does it have oxygen attached?
6. Find the 5' carbon of the deoxyribose ring. What chemical group is attached? ____________________________

Now you will examine the chemical structure of thymidine.

Restrict t3	Restricts view to nucleotide T3

- To see the nucleotide, choose "ball & stick" from the Display menu. (Keep "cpk" colors and the heteroatoms off.)

7. Identify the pyrimidine base that contains one nitrogenous ring. Draw a picture of this pyrimidine base (thymine).

Select all	

- Display in “ball & stick”. Turn off the heteroatoms.

Restrict 1,7,18,24	Identifies two nucleotide pairs in the DNA helix
Hbonds on	Shows the hydrogen bonds for each pair

8. Which two nucleotides are purines? ___________________________
9. Which two nucleotides are pyridimines? _______________________
10. What can you say about nucleotide pairs with regard to the purine-pyrimidine content? ____________________________________

Zap	Closes the file

Part III: Phosphodiester Bond Formation

Reopen the file “dnal. Pdb.” Display in “ball & stick” with “cpk” colors and heteroatoms off. (In cpk colors, red=O, blue=N, gray=C, and yellow=P.)

Restrict 2	Shows only nucleotide 2
Select 3	Selects nucleotide 3
Zoom 200	Magnifies

To see this nucleotide, select “ball & stick” from the Display menu. Ignore the water molecules, in red that also appear.

Rotate the molecule until the gap between the deoxyribose rings of nucleotides 2 and 3 can be seen (P is in yellow).

Restrict 2,3	Restricts to nucleotides 2 and 3 in preparation for viewing phosphodiester bond formation.

Now view the formation of a phosphodiester bond between nucleotides 2 and 3 by choosing “ball & stick” from the Display menu. The bond will form in the gap between the two deoxyribose rings of nucleotides 2 and 3 as soon as “ball & stick” is selected, so make sure the pull-down menu is not covering the gap.

Select 4	

To view, choose “ball & stick” from the Display menu. Ignore the water molecules, in red, that also appear.

Rotate the molecule until the gap between the deoxyribose rings of nucleotides 3 and 4 can be seen (P is in yellow).

Restrict 2,3,4	Restricts view to nucleotides 2, 3, and 4 in preparation for viewing phosphodiester bond formation.

Now view the phosphodiester bond formation between nucleotides 3 and 4 by using the Display menu to choose “ball & stick”. The bond will form in the gap between the two deoxyribose residues of nucleotides 3 and 4.

Between which two atoms does the phosphodiester bond form ? ____

To which two carbons of the deoxyribose rings is the phosphodiester bond attached? ______________________________

Zap	Closes the file.

Part IV : The Antiparallel Nature of DNA strands

Open the file “dnal.pdb” and view in “ball & stick” with “cpk” colors and the heteroatoms turned off.

Restrict 2,3,4,21,22,23	Restrict the view to these nucleotides
Zoom 150	Magnifies the view
Select purine, pyrimmidine	This will help distinguish the bases from the backbone
Color yellow	
Select backbone	

Select “cpk” from the Color menu (the bases will remain yellow and the sugar-phosphate backbones will have cpk colors).

Examine the terminal nucleotides on each chain, C21 and C23 on one chain and G4 and G2 on the other chain. Look at the terminal atoms of each, the ones that are not bonded to anything. One will be a phosphate group, and the other will be a hydroxyl group. (Note that the H of the hydroxyl group is not shown, only the O).

Identify the terminal group at each end of each chain and the carbons on the deoxyribose ring to which they are attached. The phosphate and hydroxyl groups (PO_4, OH) are named according to the number of the carbon (1’, 2’, 3’, 4’, 5’) to which they are attached (e.g., 2’-OH, 3’-OH, or 5’-PO_4).

What is the terminal group on the C21 end? ______________________

What is the terminal group on the C23 end? ______________________

What is the terminal group on the G4 end? ______________________

What is the terminal group on the G2 end? ______________________

The differences in the terminal groups result from the orientation of the two strands. One strand is oriented in the opposite direction to the other strand. Rotate this short strand of DNA on the screen so that the purines and pyrimidines are oriented 90^o to the plane of the screen (so only the edge of them is seen). Note that the oxygen (red) of the deoxyribose ring is oriented “up” on one strand and “down” on the other strand.

Zap	Closes the file

Part V: Interaction Between DNA and the Cro Repressor Protein

Open the file “cro.pdb”. This is a short stretch of DNA with two molecules of Cro protein attached to it.

Select "ribbons" under the Display menu. Rotate the molecule until you see the double helix of DNA on one side and the two proteins on the other side. The proteins are small globular (round structures with α-helices.

Identify the major groove and the minor groove of the DNA. These appear as hollows or curve where there are no molecules. If each hollow around the three-dimensional structure is traced, you will find that both the large hollow (the major groove) and the smaller hollow (the minor groove) trace a helical path.

Select "sticks" under the Display menu.

Select a	Selects A residues
Color cyan	Colors them cyan
Select g	Selects G residues
Color green	Colors them green
Select t	Selects T residues
Color red	Colors them red
Select c	Selects C residues
Color magenta	Colors them magneta

The nucleotides are now colored so that they can be easily identified.

Select protein	Selects the Cro proteins
Color blue	Colors them blue

Select "sticks" under the Display menu again.

Select all	
Hbonds on	Selects all of the hydrogen bonds
Color hbonds yellow	Colors all of them yellow
Zoom 150	Magnifies the view

Table
Properties of Amino Acid Side Chains

Properties of Side Chains (R Groups)	Amino acids
Nonpolar	Gly,ala,val,leu,ile,met,phe,trp,pro
Polar	Ser, thr,cys,tyr,asn,gln
Electrically charged	Asp,glu,lys,arg,his

Rotate the molecule. Look for close interactions between the amino acids in the protei, colored blue, with the nucleotides of DNA, colored red, cyan, green, or magenta.

1. Identify an amino acid that is in close proximity to and may interact with a nucleotide. Click on each with the mouse and identify the name and number of each molecule of the pair (one amino acid and one nucleotide). ______________________________________
2. Move to another region of protein / DNA interaction and find another possible pair of amino acid and nucleotide that may interact and give the name and number of each.

____

3. What generalization (i.e., polar, non-polar, electrically charged) can be made about the amino acids that are closest to the DNA ?

_____________________________________-

Cro interacts with DNA as a dimer, i.e., two identical polypeptides that bind together. The two Cro polypeptides that are bound to the DNA are examined next.

Hbonds off	Selects α helices of the protein
Restrict protein	Removes the nucleic acid
Select helix	Selects α helices of the protein (not DNA)
Color yellow	Colors them yellow

Display as “ribbons”.

Rotate the two molecules of Cro. Try to see the orientation of each protein molecule relative to the other. If this cannot easily be seen, imagine that your closed fists represent the two Cro proteins. This represents the orientation of the two Cro proteins relative to each other. They are identical but rotated 180o from each other.

To see this more easily, some amino acids are next identified and colored in each Cro protein.

Select lys14	Selects amino acid 14, which is a lysine
Color red	Colors lys14 red in each Cro protein
Select lys40	
Color magenta	Colors lys40 in each protein

Note the relative position of these amino acids in each molecule of Cro. Try selecting and coloring asp55, cys54, thr39, glu35, and val26, or click on amino acids with the mouse and look on the Command Line to get their name and number. (Available colors: red, green, blue, cyan, yellow, orange, white, purple, magenta).

Next, the interaction of some of these amino acids with the DNA is examined.

Select nucleic	Selects the DNA
Color white	Colors the DNA white

Display as “spacefill”. The amino acids and their interaction with the major and minor groove should now be visible.

To which groove does Cro bind ? ________________________________

Select all	

Display in “spacefill” to see more clearly how the Cro protein fits into the grooves of the DNA. Heteroatoms can be switched off.

Restrict protein	
Colour blue	

Display as “ribbon”.

Select 28-36	Selects amino acids 28-36 in Cro proteins
Color green	
Select nucleic	Select the DNA

Color white	

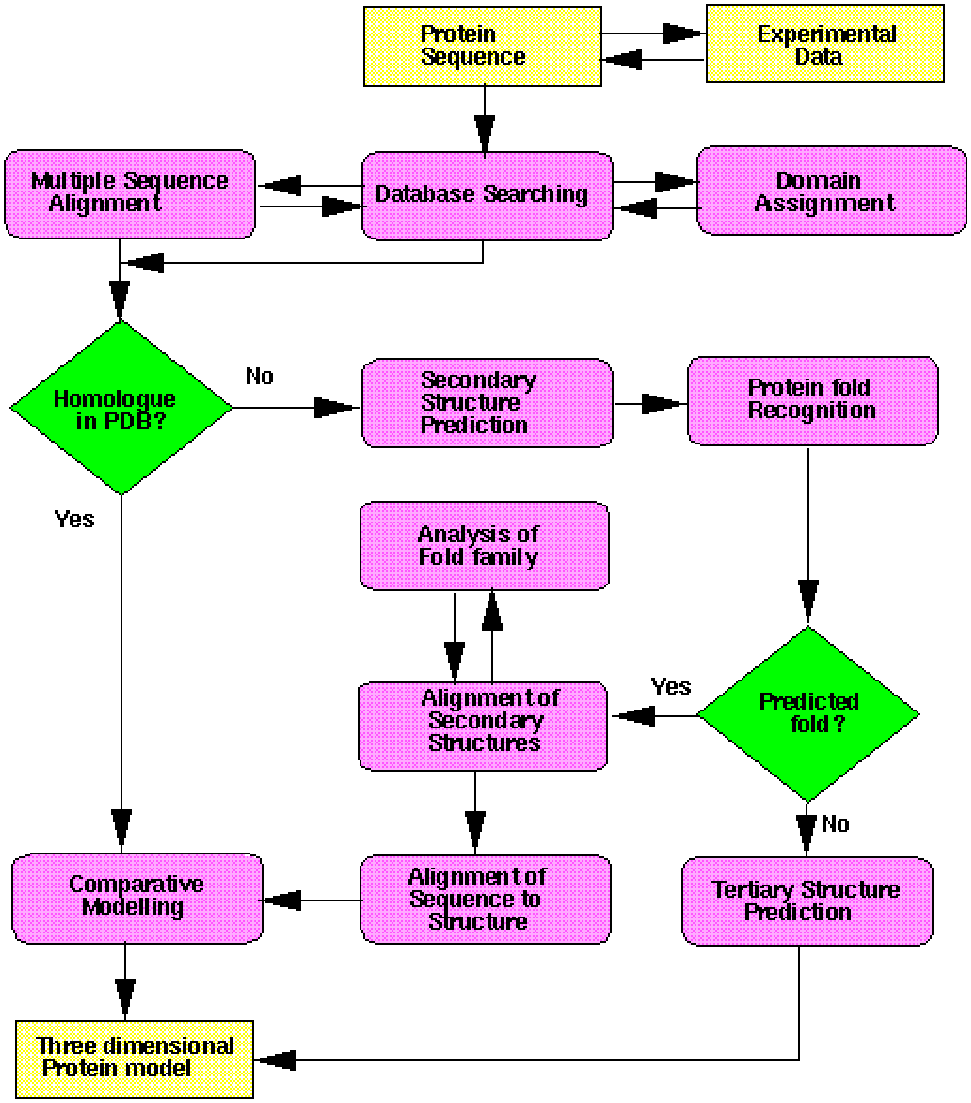
Display as “spacefill”.

Rotate the molecule to see how one of the helices of each molecule of Cro fits in the major groove. To see the helix in the groove, you should be looking down the spiral center of the green helix of Cro, not from the side of the helix.

Display as “spacefill”. Heteroatoms can be switched off.

3-D Structure Prediction and Analysis

STRUCTURE PREDICTION FLOWCHART


Protein Sequence
Experimental Data
Multiple Sequence Alignment
Database Searching
Domain Assignment
Homologue in PDB?
No
Yes
Secondary Structure Prediction
Protein fold Recognition
Analysis of Fold family
Alignment of Secondary Structures
Yes
Predicted fold ?
No
Comparative Modelling
Alignment of Sequence to Structure
Tertiary Structure Prediction
Three dimensional Protein model

The individual tasks of above chart are discussed briefly below.

Protein sequence data

There is some value in doing some initial analysis on your protein sequence. If a protein has come (for example) directly from a gene prediction, it may consist of multiple domains. More seriously, it may contain regions that are unlikely to be globular, or soluble. This flowchart assumes that your protein is soluble, likely comprises a single domain, and does not contain non-globular regions.

Things to consider are:

- Is your protein a transmembrane protein, or does it contain transmembrane segments? There are many methods for predicting these segments, including:
    - TMAP (EMBL)
    - PredictProtein (EMBL/Columbia)

  - TMHMM (CBS, Denmark)
  - TMpred (Baylor College)
  - DAS (Stockholm)
- Does your protein contain coiled-coils? You can predict coiled coils at the COILS server or you can download the COILS program (recently re-written by me of all people; note that a version of COILS is contained within the GCG suite of programs).
- Does your protein contain regions of low complexity? Proteins frequently contain runs of poly-glutamine or poly-serine, which do not predict well. To check for this you can use the program SEG (a version of SEG is also contained within the GCG suite of programs).

If the answer to any of the above questions is yes, then it is worthwhile trying to break your sequence into pieces, or ignore particular sections of the sequence, etc. This is related to the problem of locating domains.

Experimental Data

Much experimental data can aid the structure prediction process. Some of these are:

- Disulphide bonds, which provide tight restraints on the location of cysteines in space
- Spectroscopic data, which can give you and idea as to the secondary structure content of your protein
- Site directed mutagenesis studies, which can give insights as to residues involved in active or binding sites
- Knowledge of proteolytic cleavage sites, post-translational modifictions, such as phosphorylation or glycosylation can suggest residues that must be accessible
- Etc.

Remember to keep all of the available data in mind when doing predictive work. Always ask yourself whether a prediction agrees with the results of experiments. If not, then it may be necessary to modify what you've done.

Sequence database searching

The most obvious first stage in the analysis of any new sequence is to perform comparisons with sequence databases to find homologues. These searches can now be performed just about anywhere and on just about any computer. In addition, there are numerous web servers for doing searches, where one can post or paste a sequence into the server and receive the results interactively:

There are many methods for sequence searching. By far the most well known are the BLAST suite of programs. One can easily obtain versions to run locally (either at NCBI or Washington University), and there are many web pages that permit one to compare a protein or DNA sequence against a multitude of gene and protein sequence databases. To name just a few:

- National Center for Biotechnology Information (USA) Searches
- European Bioinformatics Institute (UK) Searches
- BLAST search through SBASE (domain database; ICGEB, Trieste)
- and others too numerous to mention.

One of the most important advances in sequence comparison recently has been the development of both gapped BLAST and PSI-BLAST (position specific interated BLAST). Both of these have made BLAST much more sensitive, and the latter is able to detect very remote homologues by taking the results of one search, constructing a *profile* and then using this to search the database again to find other homologues (the process can be repeated until no new sequences are found). It is essential that one compares any new protein sequence to the database with PSI-BLAST to see if known structures can be found prior to doing any of the other methods discussed in the next sections.

Other methods for comparing a single sequence to a database include:

- The FASTA suite (William Pearson, University of Virginia, USA)
- SCANPS (Geoff Barton, European Bioinformatics Institute, UK)
- BLITZ (Compugen's fast Smith Waterman search)
- and others.

It is also possible to use multiple sequence information to perform more sensitive searches. Essentially this involves building a *profile* from some kind of multiple sequence alignment. A profile essentially gives a score for each type of amino acid at each position in the sequence, and generally makes searches more sentive. Tools for doing this include:

- PSI-BLAST (NCBI, Washington)
- ProfileScan Server (ISREC, Geneva)
- HMMER Hidden Markov Model searching (Sean Eddy, Washington University)
- Wise package (Ewan Birney, Sanger Centre; this is for protein versus DNA comparisons)
- and several others.

A different approach for incorporating multiple sequence information into a database search is to use a MOTIF. Instead of giving every amino acid some kind of score at every position in an alignment, a motif ignores all but the most invariant positions in an alignment, and just describes the key residues that are conserved and define the family. Sometimes this is called a "signature". For example, "H-[FW]-x-[LIVM]-x-G-x(5)-[LV]-H-x(3)-[DE]" describes a family of DNA binding proteins. It can be translated as "histidine, followed by either a phenylalanine or tryptophan, followed by an amino acid (x), followed by leucine, isoleucine, valine or methionine, followed by any amino acid (x), followed by glycine,... [etc.]".

PROSITE (ExPASy Geneva) contains a huge number of such patterns, and several sites allow you to search these data:

- ExPASy
- EBI

It is best to search a few different databases in order to find as many homologues as possible. A very important thing to do, and one which is sometimes overlooked, is to compare any new sequence to a database of sequences for which 3D structure information is available. Whether or not your sequence is homologous to a protein of known 3D structure is not obvious in the output from many searches of large sequence databases. Moreover, if the homology is weak, the similarity may not be apparent at all during the search through a larger database.

One last thing to remember is that one can save a lot of time by making use of pre-prepared protein alignments. Many of these alignments are hand edited by experts on the particular protein families, and thus represent probably the best alignment one can get given the data they contain (i.e. they are not always as up to date as the most recent sequence databases). These databases include:

- SMART (Oxford/EMBL)
- PFAM (Sanger Centre/Wash-U/Karolinska Intitutet)
- COGS (NCBI)
- PRINTS (UCL/Manchester)
- BLOCKS (Fred Hutchinson Cancer Research Centre, Seatle)
- SBASE (ICGEB, Trieste)

Generally one can compare a protein sequence to these databases via a variety of techniques. These can also be very useful for the domain assignment.

Next Homologue in PDB? or Secondary structure prediction or Domain assignment

Locating domains

If you have a sequence of more than about 500 amino acids, you can be nearly certain that it will be divided into discrete functional domains. If possible, it is preferable to split such large proteins up and consider each domain separately. You can predict the locatation of domains in a few different ways. The methods below are given (approximately) from most to least confident.

- If homology to other sequences occurs only over a portion of the probe sequence and the other sequences are whole (i.e. not partial sequences), then this provides the strongest evidence for domain structure. You can either do database searches yourself or make use of well-curated, pre-defined databases of protein domains. Searches of these databases (see links below) will often assign domains easily.
    - SMART (Oxford/EMBL)
    - PFAM (Sanger Centre/Wash-U/Karolinska Intitutet)
    - COGS (NCBI)
    - PRINTS (UCL/Manchester)
    - BLOCKS (Fred Hutchinson Cancer Research Centre, Seatle)
    - SBASE (ICGEB, Trieste)

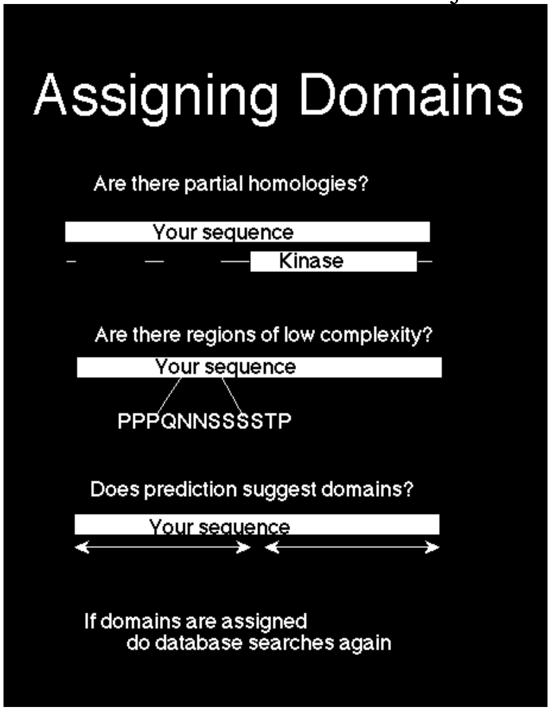
  You can also find domain descriptions in the annotations in SWISSPROT.
- Regions of low-complexity often separate domains in multidomain proteins. Long stretches of repeated residues, particularly Proline, Glutamine, Serine or Threonine often indicate linker sequences and are usually a good place to split proteins into domains.

  Low complexity regions can be defined using the program SEG which is generally available in most BLAST distributions or web servers (a version of SEG is also contained within the GCG suite of programs).
- Transmembrane segments are also very good dividing points, since they can easily separate extracellular from intracellular domains. There are many methods for predicting these segments, including:
    - TMAP (EMBL)
    - PredictProtein (EMBL/Columbia)
    - TMHMM (CBS, Denmark)
    - TMpred (Baylor College)
    - DAS (Stockholm)

- Something else to consider are the presence of *coiled-coils*. These unusual structural features sometimes (but not always) indicate where proteins can be divided into domains. You can predict coiled coils at the COILS server or you can download the COILS program (recently re-written by me of all people; a version of SEG is also contained within the GCG suite of programs).
- Secondary structure prediction methods (see below) will often predict regions of proteins to have different protein structural classes. For example one region of sequence may be predicted to contain only lpha helices and another to contain only beta sheets. These can often, though not always, suggest likely domain structure (e.g. an all alpha domain and an all beta domain)

If you have separated a sequence into domains, then it is very important to repeat all the database searches and alignments using the domains separately. Searches with sequences containing several domains may not find all sub-homologies, particularly if the domains are abundent in the database (e.g. kinases, SH2 domains, etc.). There may also be "hidden" domains. For example if there is a stretch of 80 amino acids with few homologues nested in between a kinase and an SH2 domain, then you may miss matches found when searching the *whole* sequence against a database.

Anyway, here is my slide from the talk related to this subject:



Assigning Domains

Are there partial homologies?

Your sequence

Kinase

Are there regions of low complexity?

Your sequence

PPPQNNSSSSTP

Does prediction suggest domains?

Your sequence

If domains are assigned
do database searches again

Multiple Sequence Alignment

Regardless of the outcome of your searches, you will want a multiple sequence alignment containing your sequence and all the homologues you have found above.

Some sites for performing multiple alignment:

- EBI (UK) Clustalw Server
- IBCP (France) Multalin Server
- IBCP (France) Clustalw Server

- IBCP (France) Combined Multalin/Clustalw
- MSA (USA) Server
- BCM Multiple Sequence Alignment ClustalW Sever (USA)

If you are going to do a lot of alignments, then it is probably best to get your own copy of one of many programs, some FTP sites for some of these are:

- HMMer (HMM method, Wash U)
- SAM (HMM method, Santa Cruz)
- ClustalW (EBI,UK)
- ClustalW (USA)
- MSA (USA)
- AMPS (UK)

Note that PileUp is contained within the GCG commercial package. Most institutions with people doing this sort of work will have access to this software, so ask around if you want to use it.

Probably the most important advance since these pages first appeared are Hidden Markov Models for sequence alignment. Several methods are listed above.

Alignments can provide:

- Information as to protein domain structure
- The location of residues likely to be involved in protein function
- Information of residues likely to be buried in the protein core or exposed to solvent
- More information than a single sequence for applications like homology modelling and secondary structure prediction.

Some tips

- Don't just take everything found in the searches and feed them directly into the alignment program. Searches will almost always return matches that do not indicate a significant sequence similarity. Look through the output carefully and throw things out if they don't appear to be a member of the sequence family. Inclusion of non-members in your alignment will confuse things and likely lead to errors later.
- Remember that the programs for aligning sequences aren't perfect, and do not always provide the best alignment. This is particularly so for large families of proteins with low sequence identities. If you can see a better way of aligning the sequences, then by all means edit the alignment manually.

Is your protein homologous to a known structure?

If yes, then you can proceed to comparative or homology modelling
otherwise you will probably need to perform a secondary structure prediction

Comparative or Homology Modelling

If your protein sequence shows significant homology to another protein of known three-dimensional structure, then a fairly accurate model of your protein 3D structure can be obtained via homology modelling. It is also possible to build models if you have found a suitable fold via fold recognition and are happy with the alignment of sequence to structure (Note that the accuracy of models constructed in this manner has not been assessed properly, so treat with caution).

It is possible now to generate models automatically using the very useful SWISSMODEL server.

Some other sites useful for homology modelling include:

- WHAT IF (G. Vriend, EMBL, Heidelberg)
- MODELLER (A. Sali, Rockefeller University)
- MODELLER Mirror FTP site

Sequence alignments, particularly those involving proteins having low percent sequence identities can be inacurrate. If this is the case, then a model built using the alignment will obvious be wrong in some places. I would suggest that you look over the alignment carefully before building a model.

Note that when using SWISSMODEL it is possible to send in a protein sequence only. I would only recommend doing this if the degree of sequence homology is high (50% or greater) for the above reasons. It is best, particularly if one has edited an alignment, to send an alignment directly to the server.

Once you have a three-dimensional model, it is useful to look at protein 3D structures. There are numerous free programs for doing this, including:

- GRASP Anthony Nicholls, Columbia, USA.
- MolMol Reto Koradi, ETH, Zurrich, C.H.
- Prepi Suhail Islam, ICRF, U.K.
- RasMol Roger Sayle, Glaxo, U.K.

Most places with groups studying structural biology also have commercial packages, such as Quanta, SYBL or Insight, which contain more features than the visualisation packages described above. Crystallographers also tend to use O and FRODO, though these require a lot of experience to use with ease.

Secondary Structure Prediction methods and links

There are now many web servers for structure prediction, here is quick summary:

- PSI-pred (PSI-BLAST profiles used for prediction; David Jones, Warwick)
- JPRED Consensus prediction (includes many of the methods given below; Cuff & Barton, EBI)
- DSC King & Sternberg (this server)
- PREDATORFrischman & Argos (EMBL)
- PHD home page Rost & Sander, EMBL, Germany
- ZPRED server Zvelebil et al., Ludwig, U.K.
- nnPredict Cohen et al., UCSF, USA.
- BMERC PSA Server Boston University, USA
- SSP (Nearest-neighbor) Solovyev and Salamov, Baylor College, USA.

With no homologue of known structure from which to make a 3D model, a logical next step is to predict secondary structure. Although they differ in method, the aim of secondary structure prediction is to provide the location of alpha helices, and beta strands within a protein or protein family.

Methods for single sequences

Secondary structure prediction has been around for almost a quarter of a century. The early methods suffered from a lack of data. Predictions were performed on single sequences rather than families of homologous sequences, and there were relatively few known 3D structures from which to derive parameters. Probably the most famous early methods are those of Chou & Fasman, Garnier, Osguthorbe & Robson (GOR) and Lim. Although the authors originally claimed quite high accuracies (70-80 %), under careful examination, the methods were shown to be only between 56 and 60% accurate (see

Kabsch & Sander, 1984 given below). An early problem in secondary structure prediction had been the inclusion of structures used to derive parameters in the set of structures used to assess the accuracy of the method.

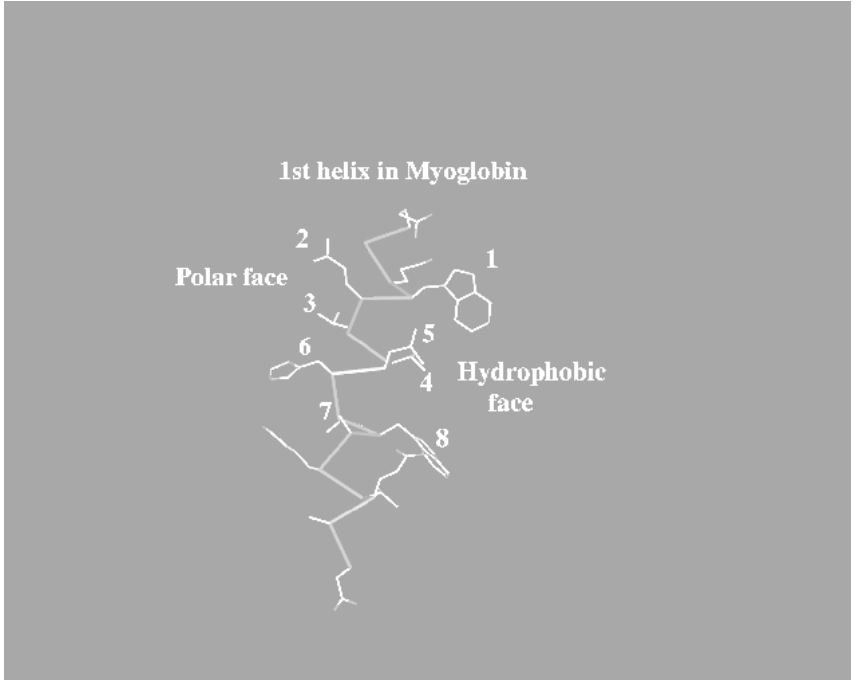
Recent improvments

The availability of large families of homologous sequences revolutionised secondary structure prediction. Traditional methods, when applied to a family of proteins rather than a single sequence proved much more accurate at identifying core secondary structure elements. The combination of sequence data with sophisticated computing techniques such as neural networks has lead to accuracies well in excess of 70 %. Though this seems a small percentage increase, these predictions are actually much more useful than those for single sequence, since they tend to predict the core accurately. Moreover, the limit of 70-80% may be a function of secondary structure variation within homologous proteins.

Manual intervention

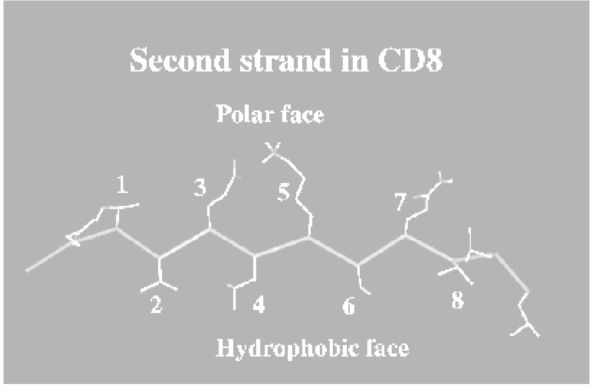
It has long been recognised that patterns of residue conservation are indicative of particular secondary structure types. Alpha helices have a periodicity of 3.6, which means that for helices with one face buried in the protein core, and the other exposed to solvent, will have residues at positions i, i+3, i+4 & i+7 (where i is a residue in an a helix) will lie on one face of the helix. Many alpha helices in proteins are amphipathic, meaning that one face is pointing towards the hydrophobic core and the other towards the solvent. Thus patterns of hydrophobic residue conservation showing the i, i+3, i+4, i+7 pattern are highly indicative of an alpha helix.

For example, this helix in myoglobin has this classic pattern of hydrophobic and polar residue conservation (*i = 1*):



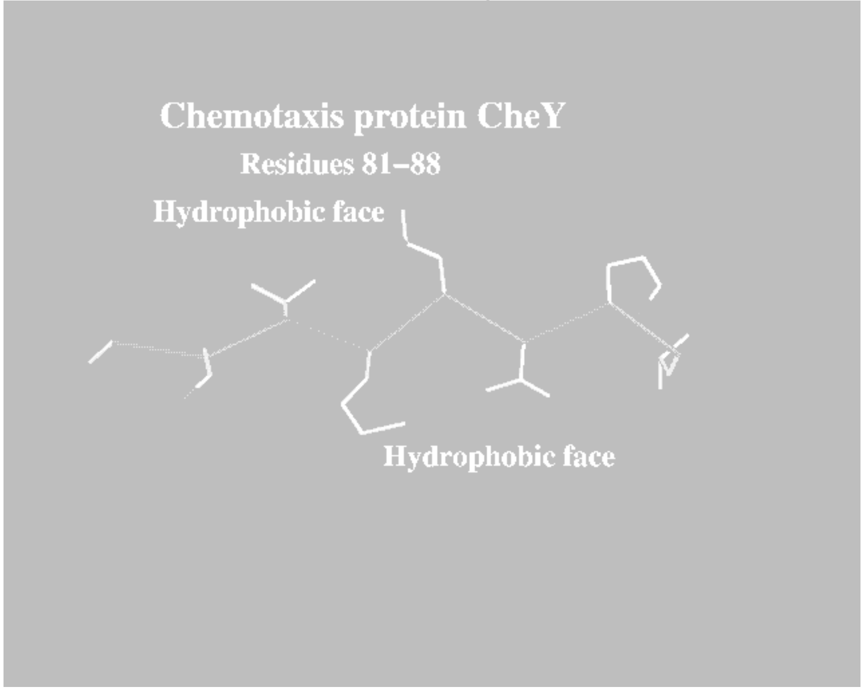
Similarly, the geometry of beta strands means that adjacent residues have their side chains pointing in oppposite directions. Beta strands that are half buried in the protein core will tend to have hydrophobic residues at positions i, i+2, i+4, i+8 etc, and polar residues at positions i+1, i+3, i+5, etc.

For example, this beta strand in CD8 shows this classic pattern:



Beta strands that are completely buried (as is often the case in proteins containing both alpha helices and beta strands) usually contain a run of hydrophobic residues, since both faces are buried in the protein core.

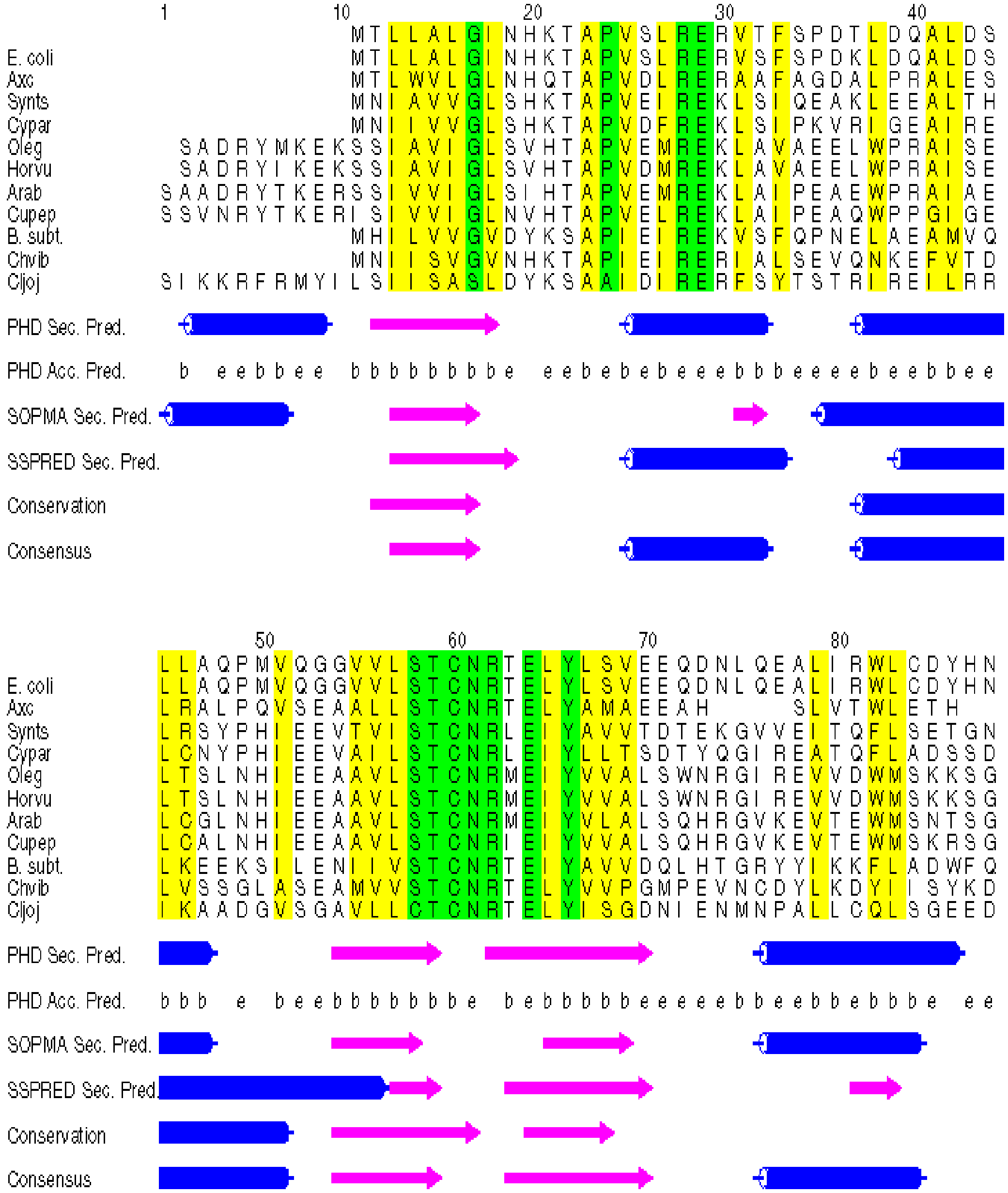
This strand from Chemotaxis protein CheY is a good example:



- The principle behind most manual secondary structure predictions is to look for patterns of residue conservation that are indicative of secondary structures like those shown above. It has been shown in numerous successful examples that this strategy often leads to nearly perfect predictions.

A strategy for secondary structure prediction

In practice, I recommend getting as many state-of-the-art prediction approaches as possible and combining this with some human insight to give a consensus prediction for the family. If you then align all of your predictions (including ideas you have based on residue conservation) with your multiple sequence alignment you can get a consensus picture of the structure. For example, here is part of an alignment of a family of proteins I looked at recently:



In this figure, three automated secondary structure predictions (PHD, SOPMA and SSPRED) appear below the alignment of 12 glutamyl tRNA reductase sequences.

Positions within the alignment showing a conservation of hydrophobic side-chain character are shown in yellow, and those showing near total conservation of non-hydrophobic residues (often indicative of active sites) are coloured green.

Predictions of accessibility performed by PHD (PHD Acc. Pred.) are also shown (b = buried, e = exposed), as is a prediction I performed by looking for patterns indicative of the three secondary structure types shown above. For example, positions (within the alignment) 38-45 exhibit the classical amphipathic helix pattern of hydrophobic residue conservation, with positions i, i+3, i+4 and i+7 showing a conservation of hydrophobicity, with intervening positions being mostly polar. Positions 13-16 comprise a short stretch of conserved hydrophobic residues, indicative of a beta-strand, similar to the example from CheY protein shown above.

By looking for these patterns I built up a prediction of the secondary structure for most regions of the protein. Note that most methods - automated and manual - agree for many regions of the alignment.

Given the results of several methods of predicting secondary structure, one can build up a *consensus* picture of the secondary structure, such as that shown at the bottom of the alignment above.

Fold recognition methods and links

Some links for methods of FOLD recognition:

- Some links for methods that run via the WWW:
    - 3D-pssm (this server)
    - TOPITS (EMBL)
    - UCLA-DOE Structre Prediction Server (UCLA)
    - 123D
    - UCSC HMM (UCSC)
    - FAS (Burnham Institute)
- Methods where an executable or code is available:
    - THREADER(Warwick)
    - ProFIT CAME (Salzburg)
- Other relevant links:
    - Protein Structure Prediction Centre (US)
    - CASP1
    - CASP2
    - CASP3
    - UCLA-DOE Fold-Recognition Benchmark Home Page

Even with no homologue of known 3D structure, it may be possible to find a suitable fold for you protein among known 3D structures by way of *fold recognition methods*

3D structural similarities

Ab initio prediction of protein 3D structures is not possible at present, and a general solution to the protein folding problem is not likely to be found in the near future. However, it has long been recognised that proteins often adopt similar folds despite no significant sequence or functional similarity and that nature is apparently restricted to a limited number of protein folds.

There are numerous protein structure classifications now available via the WWW:

- SCOP (MRC Cambridge)
- CATH (University College, London)
- FSSP (EBI, Cambridge)

- 3 Dee (EBI, Cambridge)
- HOMSTRAD (Biochemistry, Cambridge)
- VAST (NCBI, USA)

Thus for many proteins (~ 70%) there will be a suitable structure in the database from which to build a 3D model. Unfortuantely, the lack of sequence similarity will mean that many of these go undetected until after 3D structure determination.

The goal of fold recognition

Methods of protein fold recognition attempt to detect similarities between protein 3D structure that are not accompanied by any significant sequence similarity. There are many approaches, but the unifying theme is to try and find folds that are compatable with a particular sequence. Unlike sequence-only comparison, these methods take advantage of the extra information made available by 3D structure information. In effect, the turn the protein folding problem on it's head: rather than predicting how a sequence will fold, they predict how well a fold will fit a sequence.

Some papers on the subject:

The realities of fold recognition

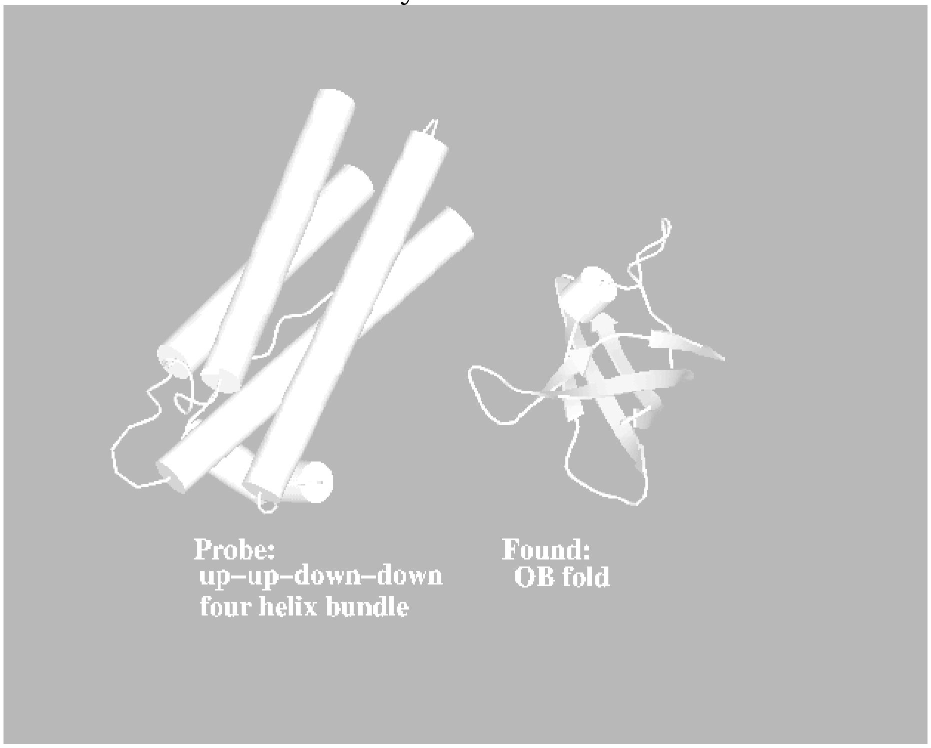
Despite initially promising results, methods of fold recognition are not always accurate. Guides to the accuracy of protein fold recognition can be found in the proceedings of the Critical Assessment of Structure Predictions (CASP) conferences. At the first meeting in 1994 (CASP1) the methods were found to be about 50 % accurate at best with respect to their ability to place a correct fold at the top of a ranked list. Though many methods failed to detect the correct fold at the top of a ranked list, a correct fold was often found in the top 10 scoring folds. Even when the methods were successful, alignments of sequence on to protein 3D structure were usually incorrect, meaning that comparative modelling performed using such models would be inaccurate.

The CASP2 meeting held in December 1996, showed that many of the methods had improved, though it is difficult to compare the results of the two assessments (i.e. CASP1 & CASP2) since very different criteria were used to assess correct answers. It would be foolish and over-ambitious for me to present a detailed assessment of the results here. However, and important thing to note, was that Murzin & Bateman managed to attain near 100% success by the use of careful human insight, a knowledge of known structures, secondary structure predictions and thoughts about the function of the target sequences. Their results strongly support the arguments given below that human insight can be a powerful aid during fold recognition. A summary of the results from this meeting can be found in the *PROTEINS* issue dedicated to the meeting (*PROTEINS*, Suppl 1, 1997).

The CASP3 meeting was held in December 1998. It showed some progress in the ability of fold recognition methods to detect correct protein folds and in the quality of alignments obtained. A detailed summary of the results will appear towards the end of 1999 in the *PROTEINS* supplement.

For my talk, I did a crude assessment of 5 methods of fold recognition. I took 12 proteins of known structure (3 from each folding class) an ran each of the five methods using default parameters. I then asked how often was a correct fold (not allowing trival sequence detectable folds) found in the first rank, or in the top 10 scoring folds. I also asked how often the method found the correct folding class in the first rank. The results are summarised in here in a PostScript file.

Perhaps the worst result from this study is shown below:



One method suggested that the sequence for the Probe (left) (a four helix bundle) would best fit onto the structure shown on the right (an OB fold, comprising a six stranded barrel).

The results suggest that one should use caution when using these methods. In spite of this, the methods remain very useful.

A practical approach:

Although they are not 100 % accurate, the methods are still very useful. To use the methods I would suggest the following:

- Run as many methods as you can, and run each method on as many sequences (from your homologous protein family) as you can. The methods almost always give somewhat different answers with the same sequences. I have also found that a single method will often give different results for sets of homologous sequences, so I would also suggest running each method on as many homologoues as possible. After all of these runs, one can build up a consensus picture of the likely fold in a manner similar to that used for secondary structure prediction above.
- Remember the expected accuracy of the methods, and don't use them as black-boxes. Remember that a correct fold may not be at the top of the list, but that it is likely to be in the top 10 scoring folds.
- Think about the function of your protein, and look into the function of the proteins that have been found by the various methods. If you see a functional similarity, then you may have detected a *weak sequence homologue*, or *remote homologue*. At CASP2, as said above, Murzin & Bateman managed to obtain

remarkably accurate predictions by identification of remote homologues. Their paper appeard in the *PROTEINS* supplement for the CASP2 experiment:
Murzin AG, Bateman A (1997) Distant homology recognition using structural classification of proteins *Proteins*, Suppl 1, 105-112.
and provides some key insights into protein fold recognition using humans rather than computers.

- Don't trust the alignments that are output by the programs. They can be used as a starting point, but the best alignment of sequence on to tertiary structure is still likely to come from careful human intervention. One strategy for doing this is discussed in the next section

Analysis of protein folds and alignment of secondary structure elements

---

If you have predicted that your protein will adopt a particular fold within the database, then an important thing to consider to which fold your protein belongs, and other proteins that adopt a similar fold. To find out, look at one of the following databases:

- SCOP (MRC Cambridge)
- CATH (University College, London)
- FSSP (EBI, Cambridge)
- 3 Dee (EBI, Cambridge)
- HOMSTRAD (Biochemistry, Cambridge)
- VAST (NCBI, USA)

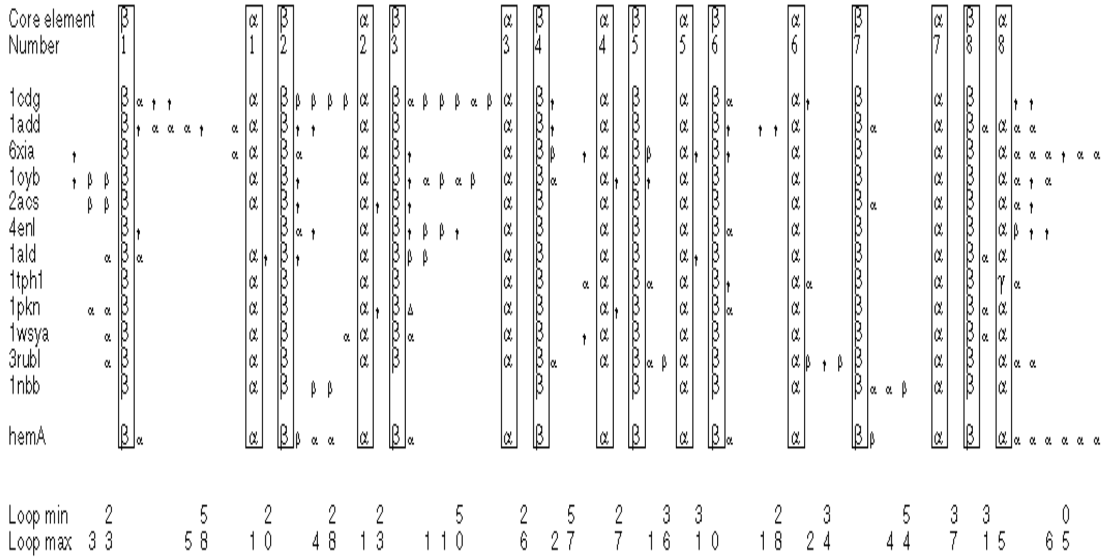
(Note that these databases don't always agree as to what constitutes a similar fold, so I would recommend looking at as many of them as possible).

If your predicted fold has many "relatives", then have a look at what they are. Ask:

- Do any of members show functional similarity to your protein? If there is any functional similarity between your protein and any members of the fold, then you may be able to back up your prediction of fold (possibly by the conservation of active site residues, or the approximate location of active site residues, etc.)
- Is this fold a *superfold*? If so, does this superfold contain a *supersite*? Certain folds show a tendancy to bind ligands in a common location, even in the absense of any functional or clear evolutionary relationships. For an explanation of this, please see our work on supersites.
- Are there *core* secondary structure elements that should really be present in any member of the fold?
- Are there non-core secondary structure elements that might not be present in all members of the fold?

Core secondary structure elements, such as those comprising a beta-barrel, should really be present in a fold. If your predicted secondary structures can't be made to match up with what you think is the core of the protein fold, then your prediction of fold may be wrong (but be careful, since your secondary structure prediction may contain errors). You can also use your prediction together with the core secondary structure elements to derive an alignment of of predicted and observed secondary structures.

For example, we predicted that the glutamyl tRNA reductases (hemA family) would adopt an alpha-beta barrel fold using a combination of fold recognition and secondary structure prediction methods. We aligned the secondary structures of diverse members of the alpha-beta barrel fold using a structural alignment program, and aligned the secondary structures to the *core* (boxed below) secondary structure elements.



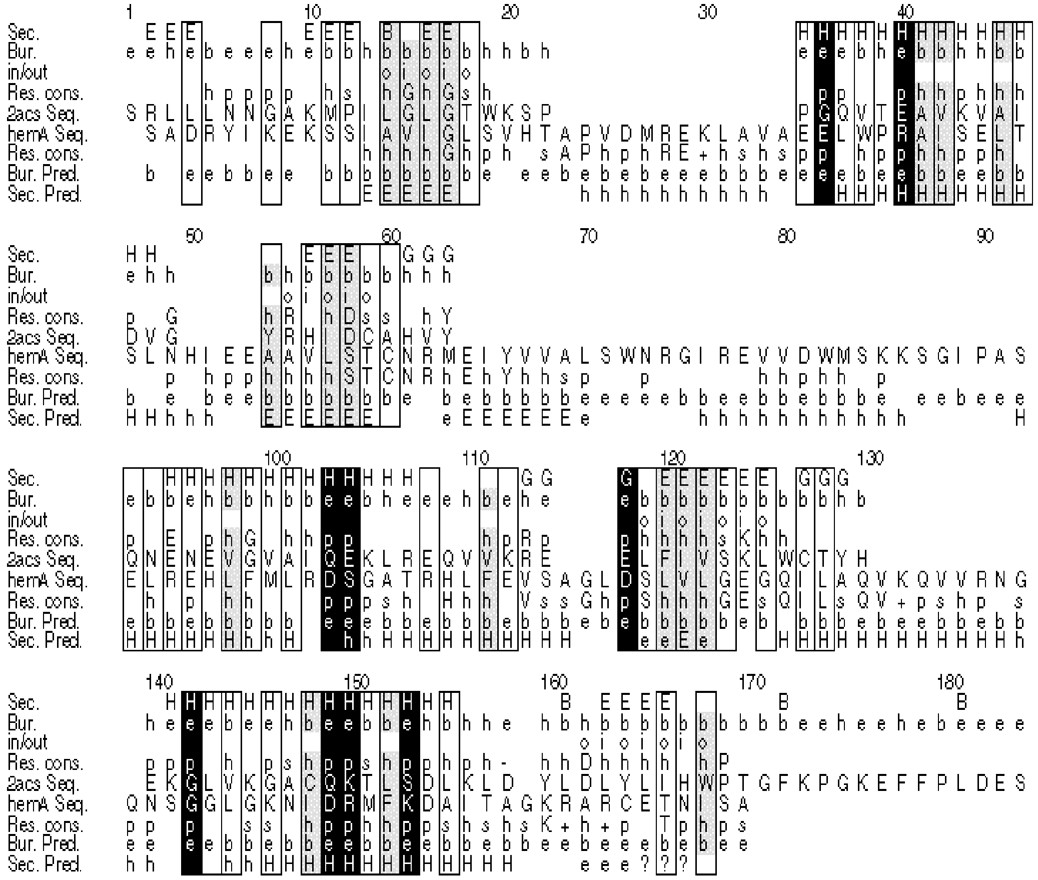
In the alignment above, each alpha and beta character refers to an entire secondary structure element. Those that are boxed are *core* secondary structure elements found in most members of the fold. The alignment of predicted secondary structures to the core elements appears at the bottom of the figure. Note that I have had to delete several alpha helices and beta strands from our prediction to allow for alignment. This is not surprising, because insertions or deletions of secondary structure elements are common across the diverse set of proteins that adopt this fold.

Alignment of sequence to tertiary structure

Remember that the alignments of sequence on to tertiary structure that one gets from fold recognition methods may be inaccurate. In instance where one has identified a *remote homologue*, then the fold recognition methods can sometimes give a very accurate alignment, though it is still sometimes fruitful to edit the alignment around variable regions (see the Multiple Sequence Alignment for ways of doing this). In other cases, it may be wise to create your own alignment by starting with the alignment from the fold recognition method, and considering the alignment of secondary structures.
There is no generally accepted way for doing this, though one method (ie. mine) involves:

- Ensuring that residues predicted to be buried/exposed align to those *known* to be buried or exposed in the template structure. Note that *conserved* hydrophobic/polar residues are more likely to be buried/exposed than non-conserved residues, which could simply be anomalies. One can predict residue accessibility manually, or by use of an automated server like PHD.
- Ensuring that critical hydrogen bonding patterns are not disrupted in beta-sheet structures.
- Trying to conserve residue properties (i.e. size, polarity, hydrophobicity) as best as possible across known and unknown structure.

For example, in trying to align the prediction of the glutamyl tRNA reductases (hemA) with one alpha/beta barrel structure (2acs):


1 10 20 30 40
Sec. E E E E E E B E E H H H H H H H H H H H H
Bur. e e h e b e e e h e b b h b b b b b h h b h e e e b h e b b h h b b
in/out o i o i o
Res. cons. h p p p p h s h G h G s h p p p h h p h h h
2acs Seq. S R L L L N N G A K M P I L G L G T W K S P P G Q V T E A V K V A I
hemA Seq. S A D R Y I K E K S S I A V I G L S V H T A P V D M R E K L A V A E E L W P R A I S E L T
Res. cons. h h h h G h p h s A P h p h R E + h s h s p p h p p h h p p h
Bur. Pred. b e e b b e e b b b b b b b b b e e e b e b e b e e e b b b e e e e b e e b b e e b b
Sec. Pred. E E E E E h h h h h h h h h H H H H H H H H H H
50 60 70 80 90
Sec. H H E E E G G G
Bur. e h h b h b b b b b h h h
in/out o i o i o
Res. cons. p G h R h D s s h Y
2acs Seq. D V G Y R H I D C A H V Y
hemA Seq. S L N H I E E A A V L S T C N R M E I Y V V A L S W N R G I R E V V D W M S K K S G I P A S
Res. cons. p h p p h h h h S T C N R h E h Y h h s p p h h p h h p
Bur. Pred. b e b e e b b b b b b b e b e b b b b b e e e e b b e e b b e b b b e e e b e e e
Sec. Pred. H H h h h E E E E E E e E E E E E E e h h h h h h h h h h H
100 110 120 130
Sec. H H H H H H H H H H H H H G G G E E E E E E G G G
Bur. e b b e h b b h b b e e b h e e e h b e h e e b b b b b b b b b b h b
in/out o i o i o i o
Res. cons. p E p h G h h p p h p R p p h h h h s K h h
2acs Seq. Q N E N E V G V A I Q E K L R E Q V V K R E E L F I V S K L W C T Y H
hemA Seq. E L R E H L F M L R D S G A T R H L F E V S A G L D S L V L G E G Q I L A Q V K Q V V R N G
Res. cons. h p h h p p p s h H h h V s s G h p S h h h G E s Q I L s Q V + p s h p s
Bur. Pred. e b b e b b b b b e e e b b e b b b e b b b e b e b b b b b e b b b e b e e b b e b b
Sec. Pred. H H H H H H h h H h h H H H H H H H H H H e e E e H H H H H H H H H H H H h
140 150 160 170 180
Sec. H H H H H H H H H H H H H H H B E E E E B B
Bur. h e e e b e e h b e e b b e h b h h e h b h b b b b b b b b b e e h e e h e b e e e e
in/out o i o i o i o
Res. cons. p p p h p s h p p s h p p h p h - h h D h h h h h P
2acs Seq. E K G L V K G A C Q K T L S D L K L D Y L D L Y L I H W P T G F K P G K E F F P L D E S
hemA Seq. Q N S G G L G K N I D R M F K D A I T A G K R A R C E T N I S A
Res. cons. p p p s s h p p h h p p s h s h s K + h + p T p h p s
Bur. Pred. e e e e b b e b b e e b b e b b e b b e e b e e e b e b e e
Sec. Pred. h h h h H H H H H H H H H H H H H e e e ? ? ?

[Sec.= known secondary structure from PDB code 2ACS (E = extended, H = alpha helix, G = 3-10 helix, B = beta-bridge); Bur. = known residue exposure for 2ACS (b = buried, h = half-buried, e = exposed); in/out = positioning of residues in the beta-barrel (i = pointing inwards, o = pointing outwards); Res. cons = conservation of residues (totally conserved = UPPER CASE, h = hydrophobic, p = polar, c = charged, a = aromatic, s = small, - = negaitve, + = positive) Pred denotes predicted burial and secondary structure for the glutamyl tRNA reductase family; boxed positions are those with the same known/predicted burial. Shaded positions show a conservation of hydrophobic character in BOTH families of proteins, and positions in inverse text show a conservation of polar character in BOTH families.]

In the construction of this alignment, several things were considered:

- The observed residue burial or exposure
- The predicted residue burial or exposure
- The conservation of residue properties in known and unknown structures
- Whether or not the side chains on the core beta-strands pointed in towards the barrel or out towards the helices

- The hydrogen bonding pattern of the beta-strands comprising the core beta-barrel.

By using an initial alignment from one of the fold recognition methods as a guide, the alignment above was created by trying to optimise the match of features described above.

Remember that proteins having similar three-dimensional structures with little or no sequence similarity can differ substantial with respect to the finer details of their structures (i.e. loops, precise orientation of side chains, orientation of secondary structures, etc.). See here for some work I did with Geoff Barton on this subject.