

SCOPE

The scope of this paper is to introduce the main principles of bioinformatics the coverage will include biological data base, sequence alignment and structure prediction.

OBJECTIVES

Understand the genomic data acquisitions and analysis comparative and predictive analysis of DNA and Protein sequence.

Unit I**Introduction to Bioinformatics**

Computer fundamentals - programming languages in bioinformatics, role of supercomputers in biology. Historical background. Scope of bioinformatics - genomics, proteomics, Computer aided drug design (structure based and ligand based approaches) and Systems Biology. Applications of bioinformatics.

Unit II**Biological databases and data retrieval**

Introduction to biological databases - primary, secondary and composite databases, NCBI, nucleic acid databases (GenBank, EMBL, DDBJ, NDB), protein databases (PIR, Swiss-Prot, TrEMBL, PDB), metabolic pathway database (KEGG, EcoCyc, and MetaCyc), small molecule databases (PubChem, Drug Bank, ZINC, CSD). Structure viewers (RasMol, J mol), file formats.

Unit III**Sequence alignment**

Similarity, identity and homology. Alignment – local and global alignment, pairwise and multiple sequence alignments, alignment algorithms, amino acid substitution matrices (PAM and BLOSUM), BLAST and CLUSTALW.

Unit IV**Phylogenetic analysis**

Construction of phylogenetic tree, dendrograms, methods of construction of phylogenetic trees - maximum parsimony, maximum likelihood and distance methods.

Unit V**Protein structure prediction analysis and gene prediction**

Levels of protein structure. Protein tertiary structure prediction methods –homology modeling, fold recognition and ab-initio methods. Significance of Ramachandran map. Introduction to genomics, comparative and functional genomics, gene structure in prokaryotes and eukaryotes, gene prediction methods and tools.

REFERENCES

1. Mount, D.W., (2001). Bioinformatics: Sequence and Genome Analysis, 1st ed., Cold Spring Harbor Laboratory Press (New York), ISBN: 0-87969-608-7.
2. Pevsner, J., (2003). Bioinformatics and Functional Genomics (2003), 1st ed., John Wiley & Sons, Inc (New Jersey), ISBN: 0-47121004-8.
3. Baxevanis, A.D., and Ouellette, B.F., (2005). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd ed., John Wiley & Sons, Inc. (New Jersey), ISBN: 0-47147878-4.
4. Ghosh, Z., and Mallick, B., (2008). Bioinformatics-Principles and Applications (2008). 1st ed. Oxford University Press (India), ISBN: 9780195692303.



KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University Established Under Section 3 of UGC Act 1956)

Coimbatore – 641 021.

LECTURE PLAN DEPARTMENT OF BIOCHEMISTRY

STAFF NAME: S. RAJAMANIKANDAN
SUBJECT NAME: BIOINFORMATICS
SEMESTER: IV

SUB.CODE:16BCU404A
CLASS: II B.Sc (BC)

Sl. No	LECTURE DURATION	TOPICS	BOOK REFERENCE	PAGE NO	WEB REFERENCE
Unit-1					
1	1	Computer Fundamentals-Programming languages in Bioinformatics	T1	16-20	
2	1	Role of supercomputers in biology, Historical Background			
3	1	Scope of Bioinformatics-genomics, proteomics	T1	52-55, 293-308	
4	1	Computer aided drug designing	T2 351-364		
5	1	Structure and ligand based approach			
6	1	System Biology, Application of Bioinformatics	T1	311-338	
7	1	Unit I revision			
8	1	Revision and Possible QP discussion			
9	1	Revision and Possible QP discussion			
Total: 9 hours					
Unit-2					
1	1	Introduction of biological databases- Primary, secondary and composite databases	T2	15-16	
2	1	NCBI, Nucleic acid databases (Genbank, EMBL,DDBJ, NDB)	T3	35-40	
3	1	Protein databases (PIR, Swissprot, TrEMBL, PDB)	T1	124-128	
4	1	Metabolic pathway database (KEGG, EcoCyc and Metacyc)	T1	138-139	
5	1	Small molecule database (Pubchem, Drug			Article 1

		Bank, Zinc, CSD)			
6	1	Structure viewers (Rasmol, J Mol)	T4	201-211, 195-196	
7	1	File formats	T3	29-34	
8	1	Unit II revision			
9	1	Revision and Possible QP discussion			
Total: 9 hours					
Unit-3					
1	1	Similarity identity and homology	T2 73-75		
2	1	Alignment-local and global alignment			
3	1	Pairwise and multiple sequence alignment	T2	75-77, 101-102	
4	1	Alignment algorithms	T2	77-82	
5	1	Amino acid substitution matrices (PAM and BLOSUM)	T2	83-91	
6	1	BLAST AND CLUSTALW	T1	133-144	Article 2
7	1	Unit III revision			
8	1	Revision and Possible QP discussion			
Total: 8 hours					
Unit-4					
1	1	Construction of phylogenetic tree, dendrograms	T2	103-105	
2	1	Methods of construction of phylogenetic trees	T2	105-112	
3	2	Maximum parsimony, maximum likelihood	T2	112-115	
4	1	Distance methods	T2	107-109	
5	1	Unit IV revision			
6	1	Revision and Possible QP discussion			
7	1	Revision and Possible QP discussion			
Total: 8 hours					
Unit-5					
1	1	Levels of protein structure	T2	236-247	
2	1	Protein tertiary structure prediction methods-homology modeling	T1	250-252	

3	1	Fold recognition and ab-initio methods	T1	252-259	
4	1	Significance of Ramachandran Map			W1
5	1	Introduction of genomics, comparative and functional genomics			Article 3
6	1	Gene structure in prokaryotes and eukaryotes	T2 171-190		
7	1	Gene prediction methods and tools			
8	1	Unit V revision			
9	1	Revision and Possible QP discussion			
Total: 9 hours					
PREVIOUS YEAR END SEMESTER EXAMINATION QUESTION PAPER DISCUSSION					
1	1	Previous year ESE question paper discussion			
2	1	Previous year ESE question paper discussion			
Total: 2 hours					
Grand Total: 45 hours					

REFERENCE

T1	:	Arthur M. Lesk, (2005). Introduction to Bioinformatics, 2 nd edition, Published by Oxford University Press, New Delhi-110001.
T2	:	Rastogi S.C., Mendiratta N., Rastogi.P. (2004). Bioinformatics Methods and Applications (Genomics, Proteomics and Drug Discovery). Published by Asoke K. Ghosh. Prentice Hall of India, Private Limited, M-97, Connaught Circus, New Delhi-110001.
T3	:	Westhead, D.R., Parish J.H., Twyman, R.M., (2003). Bioinformatics, Published by Vinod Vasishtha for Viva Books Private Limited, New Delhi-110 002.
T4	:	Mani K, Vijayaraj N, (2002). Bioinformatics for Beginners, Kalaikathir Achchagam, Coimbatore.
Article 1	:	Gozalbes, R., Pineda-Lucena, A (2011). Small molecule database and chemical descriptors useful in chemoinformatics: an overview. Comb Chem High Throughput Screen. 14:548-458.
Article 2	:	Chenna ,R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D. (2003). Multiple sequence alignment with the clustal series of programs. Nucleic Acids Res. 31:3497-3500.
Article 3	:	Haubold B, Wiehe T, (2004). Comparative genomics: methods and applications. Naturwissenschaften 91: 405-421.
W1	:	https://en.wikibooks.org/wiki/Structural_Biochemistry/Proteins/Ramachandran_Plot

UNIT-I

SYLLABUS

Introduction to Bioinformatics: Computer fundamentals- programming languages in Bioinformatics, role of supercomputers in Biology, Historical background. Scope of Bioinformatics-genomics, proteomics, computer aided drug design (structure based and ligand based approaches) and System Biology. Applications of Bioinformatics.

COMPUTER FUNDAMENTALS

Computer Definition

Computer is an advanced electronic device that takes raw data as an input from the user and processes it under the control of a set of instructions (called program), produces a result (output), and saves it for future use.

Functionalities of a computer

Step 1 – Takes data as input.

Step 2 – Stores the data/instructions in its memory and uses them as required.

Step 3 – Processes the data and converts it into useful information.

Step 4 – Generates the output.

Step 5 – Controls all the above four steps.

Advantages of Computers

High Speed

- Computer is a very fast device.
- It is capable of performing calculation of very large amount of data.
- The computer has units of speed in microsecond, nanosecond, and even the picosecond.
- It can perform millions of calculations in a few seconds as compared to man who will spend many months to perform the same task.

Accuracy

- In addition to being very fast, computers are very accurate.
- The calculations are 100% error free.

- Computers perform all jobs with 100% accuracy provided that the input is correct.

Storage Capability

- Memory is a very important characteristic of computers.
- A computer has much more storage capacity than human beings.
- It can store large amount of data.
- It can store any type of data such as images, videos, text, audio, etc.

Diligence

- Unlike human beings, a computer is free from monotony, tiredness, and lack of concentration.
- It can work continuously without any error and boredom.
- It can perform repeated tasks with the same speed and accuracy.

Versatility

- A computer is a very versatile machine.
- A computer is very flexible in performing the jobs to be done.
- This machine can be used to solve the problems related to various fields.
- At one instance, it may be solving a complex scientific problem and the very next moment it may be playing a card game.

Reliability

- A computer is a reliable machine.
- Modern electronic components have long lives.
- Computers are designed to make maintenance easy.

Automation

- Computer is an automatic machine.
- Automation is the ability to perform a given task automatically. Once the computer receives a program i.e., the program is stored in the computer memory, then the program and instruction can control the program execution without human interaction.

Reduction in Paper Work and Cost

- The use of computers for data processing in an organization leads to reduction in paper work and results in speeding up the process.
- As data in electronic files can be retrieved as and when required, the problem of maintenance of large number of paper files gets reduced.
- Though the initial investment for installing a computer is high, it substantially reduces the cost of each of its transaction.

Disadvantages of Computers

Dependency

- It functions as per the user's instruction, thus it is fully dependent on humans.

Environment

- The operating environment of the computer should be dust free and suitable.

No Feeling

- Computers have no feelings or emotions.
- It cannot make judgment based on feeling, taste, experience, and knowledge unlike humans.

Applications of Computer

Business

A computer has high speed of calculation, diligence, accuracy, reliability, or versatility which has made it an integrated part in all business organizations.

Computer is used in business organizations for –

- Payroll calculations
- Budgeting
- Sales analysis
- Financial forecasting
- Managing employee database
- Maintenance of stocks, etc.

Banking

Banks provide the following facilities –

KARPAGAM ACADEMY OF HIGHER EDUCATION

CLASS: II BSC BC

COURSE NAME: BIOINFORMATICS

COURSE CODE: 16BCU404A

UNIT: I BIOINFORMATICS INTRODUCTION BATCH-2016-2019

- Online accounting facility, which includes checking current balance, making deposits and overdrafts, checking interest charges, shares, and trustee records.
- ATM machines which are completely automated are making it even easier for customers to deal with banks.

Insurance

Insurance companies are keeping all records up-to-date with the help of computers. Insurance companies, finance houses, and stock broking firms are widely using computers for their concerns.

Insurance companies are maintaining a database of all clients with information showing –

- Procedure to continue with policies
- Starting date of the policies
- Next due installment of a policy
- Maturity date
- Interests due
- Survival benefits
- Bonus

Education

- The computer provides a tool in the education system known as CBE (Computer Based Education).
- CBE involves control, delivery, and evaluation of learning.
- Computer education is rapidly increasing the graph of number of computer students.
- There are a number of methods in which educational institutions can use a computer to educate the students.
- It is used to prepare a database about performance of a student and analysis is carried out on this basis.

Marketing

- **Advertising** – With computers, advertising professionals create art and graphics, write and revise copy, and print and disseminate ads with the goal of selling more products.
- **Home Shopping** – Home shopping has been made possible through the use of computerized catalogues that provide access to product information and permit direct entry of orders to be filled by the customers.

Healthcare

Computers have become an important part in hospitals, labs, and dispensaries. They are being used in hospitals to keep the record of patients and medicines. It is also used in scanning and diagnosing different diseases. ECG, EEG, ultrasounds and CT scans, etc. are also done by computerized machines.

Following are some major fields of health care in which computers are used.

- **Diagnostic System** – Computers are used to collect data and identify the cause of illness.
- **Lab-diagnostic System** – All tests can be done and the reports are prepared by computer.
- **Patient Monitoring System** – These are used to check the patient's signs for abnormality such as in Cardiac Arrest, ECG, etc.
- **Pharma Information System** – Computer is used to check drug labels, expiry dates, harmful side effects, etc.
- **Surgery** – Nowadays, computers are also used in performing surgery.

Engineering Design

Computers are widely used for Engineering purpose.

One of the major areas is CAD (Computer Aided Design) that provides creation and modification of images. Some of the fields are –

- **Structural Engineering** – Requires stress and strain analysis for design of ships, buildings, budgets, airplanes, etc.

- **Industrial Engineering** – Computers deal with design, implementation, and improvement of integrated systems of people, materials, and equipment.
- **Architectural Engineering** – Computers help in planning towns, designing buildings, determining a range of buildings on a site using both 2D and 3D drawings.

Military

Computers are largely used in defence. Modern tanks, missiles, weapons, etc. Military also employs computerized control systems. Some military areas where a computer has been used are –

- Missile Control
- Military Communication
- Military Operation and Planning
- Smart Weapons

Communication

Communication is a way to convey a message, an idea, a picture, or speech that is received and understood clearly and correctly by the person for whom it is meant. Some main areas in this category are.

- E-mail
- Chatting
- Usenet
- FTP
- Telnet
- Video-conferencing

Government

Computers play an important role in government services. Some major fields in this category are

- Budgets
- Sales tax department
- Income tax department

- Computation of male/female ratio
- Computerization of voters lists
- Computerization of PAN card
- Weather forecasting

Computers can be broadly classified by their speed and computing power.

S. No.	Type	Specifications
1	PC (Personal Computer)	It is a single user computer system having moderately powerful microprocessor
2	Workstation	It is also a single user computer system, similar to personal computer however has a more powerful microprocessor.
3	Mini Computer	It is a multi-user computer system, capable of supporting hundreds of users simultaneously.
4	Main Frame	It is a multi-user computer system, capable of supporting hundreds of users simultaneously. Software technology is different from minicomputer.
5	Supercomputer	It is an extremely fast computer, which can execute hundreds of millions of instructions per second.

Components of Computers

Input Unit

This unit contains devices with the help of which we enter data into the computer. This unit creates a link between the user and the computer. The input devices translate the information into a form understandable by the computer.

CPU (Central Processing Unit)

CPU is considered as the brain of the computer. CPU performs all types of data processing operations. It stores data, intermediate results, and instructions (program). It controls the operation of all parts of the computer.

CPU itself has the following three components –

- ALU (Arithmetic Logic Unit)
- Memory Unit
- Control Unit

Output Unit

The output unit consists of devices with the help of which we get the information from the computer. This unit is a link between the computer and the users. Output devices translate the computer's output into a form understandable by the users.

Central Processing Unit (CPU) consists of the following features –

- CPU is considered as the brain of the computer.
- CPU performs all types of data processing operations.
- It stores data, intermediate results, and instructions (program).
- It controls the operation of all parts of the computer.

CPU itself has following three components.

- Memory or Storage Unit
- Control Unit
- ALU(Arithmetic Logic Unit)

Memory or Storage Unit

This unit can store instructions, data, and intermediate results. This unit supplies information to other units of the computer when needed. It is also known as internal storage unit or the main memory or the primary storage or Random Access Memory (RAM).

Its size affects speed, power, and capability. Primary memory and secondary memory are two types of memories in the computer. Functions of the memory unit are –

- It stores all the data and the instructions required for processing.
- It stores intermediate results of processing.
- It stores the final results of processing before these results are released to an output device.
- All inputs and outputs are transmitted through the main memory.

Control Unit

This unit controls the operations of all parts of the computer but does not carry out any actual data processing operations.

Functions of this unit are –

- It is responsible for controlling the transfer of data and instructions among other units of a computer.
- It manages and coordinates all the units of the computer.
- It obtains the instructions from the memory, interprets them, and directs the operation of the computer.
- It communicates with Input/Output devices for transfer of data or results from storage.
- It does not process or store data.

ALU (Arithmetic Logic Unit)

This unit consists of two subsections namely,

- Arithmetic Section
- Logic Section

Arithmetic Section

Function of arithmetic section is to perform arithmetic operations like addition, subtraction, multiplication, and division. All complex operations are done by making repetitive use of the above operations.

Logic Section

Function of logic section is to perform logic operations such as comparing, selecting, matching, and merging of data.

Following are some of the important input devices which are used in a computer –

- Keyboard
- Mouse
- Joy Stick
- Light pen
- Track Ball

- Scanner
- Graphic Tablet
- Microphone
- Magnetic Ink Card Reader(MICR)
- Optical Character Reader(OCR)
- Bar Code Reader
- Optical Mark Reader(OMR)

Keyboard

Keyboard is the most common and very popular input device which helps to input data to the computer. The layout of the keyboard is like that of traditional typewriter, although there are some additional keys provided for performing additional functions.

Keyboards are of two sizes 84 keys or 101/102 keys, but now keyboards with 104 keys or 108 keys are also available for Windows and Internet.

The keys on the keyboard are as follows –

S. No	Keys & Description
1	Typing Keys These keys include the letter keys (A-Z) and digit keys (09) which generally give the same layout as that of typewriters.
2	Numeric Keypad It is used to enter the numeric data or cursor movement. Generally, it consists of a set of 17 keys that are laid out in the same configuration used by most adding machines and calculators.
3	Function Keys The twelve function keys are present on the keyboard which are arranged in a row at the top of the keyboard. Each function key has a unique meaning and is

	used for some specific purpose.
4	Control keys These keys provide cursor and screen control. It includes four directional arrow keys. Control keys also include Home, End, Insert, Delete, Page Up, Page Down, Control(Ctrl), Alternate(Alt), Escape(Esc).
5	Special Purpose Keys Keyboard also contains some special purpose keys such as Enter, Shift, Caps Lock, Num Lock, Space bar, Tab, and Print Screen.

Mouse

Mouse is the most popular pointing device. It is a very famous cursor-control device having a small palm size box with a round ball at its base, which senses the movement of the mouse and sends corresponding signals to the CPU when the mouse buttons are pressed.

Generally, it has two buttons called the left and the right button and a wheel is present between the buttons. A mouse can be used to control the position of the cursor on the screen, but it cannot be used to enter text into the computer.

Advantages

- Easy to use
- Not very expensive
- Moves the cursor faster than the arrow keys of the keyboard.

Joystick

Joystick is also a pointing device, which is used to move the cursor position on a monitor screen. It is a stick having a spherical ball at its both lower and upper ends. The lower spherical ball moves in a socket. The joystick can be moved in all four directions. The function of the joystick is similar to that of a mouse. It is mainly used in Computer Aided Designing (CAD) and playing computer games.

Light Pen

Light pen is a pointing device similar to a pen. It is used to select a displayed menu item or draw pictures on the monitor screen. It consists of a photocell and an optical system placed in a small tube. When the tip of a light pen is moved over the monitor screen and the pen button is pressed, its photocell sensing element detects the screen location and sends the corresponding signal to the CPU.

Track Ball

Track ball is an input device that is mostly used in notebook or laptop computer, instead of a mouse. This is a ball which is half inserted and by moving fingers on the ball, the pointer can be moved. Since the whole device is not moved, a track ball requires less space than a mouse. A track ball comes in various shapes like a ball, a button, or a square.

Scanner

Scanner is an input device, which works more like a photocopy machine. It is used when some information is available on paper and it is to be transferred to the hard disk of the computer for further manipulation.

Scanner captures images from the source which are then converted into a digital form that can be stored on the disk. These images can be edited before they are printed.

Digitizer

Digitizer is an input device which converts analog information into digital form. Digitizer can convert a signal from the television or camera into a series of numbers that could be stored in a computer. They can be used by the computer to create a picture of whatever the camera had been pointed at. Digitizer is also known as Tablet or Graphics Tablet as it converts graphics and pictorial data into binary inputs. A graphic tablet as digitizer is used for fine works of drawing and image manipulation applications.

Microphone

Microphone is an input device to input sound that is then stored in a digital form. The microphone is used for various applications such as adding sound to a multimedia presentation or for mixing music.

Magnetic Ink Card Reader (MICR)

MICR input device is generally used in banks as there are large number of cheques to be processed every day. The bank's code number and cheque number are printed on the cheques with a special type of ink that contains particles of magnetic material that are machine readable. This reading process is called Magnetic Ink Character Recognition (MICR). The main advantages of MICR is that it is fast and less error prone.

Optical Character Reader (OCR)

OCR is an input device used to read a printed text.

OCR scans the text optically, character by character, converts them into a machine readable code, and stores the text on the system memory.

Bar Code Readers

Bar Code Reader is a device used for reading bar coded data (data in the form of light and dark lines). Bar coded data is generally used in labelling goods, numbering the books, etc. It may be a handheld scanner or may be embedded in a stationary scanner. Bar Code Reader scans a bar code image, converts it into an alphanumeric value, which is then fed to the computer that the bar code reader is connected to.

Optical Mark Reader (OMR)

OMR is a special type of optical scanner used to recognize the type of mark made by pen or pencil. It is used where one out of a few alternatives is to be selected and marked. It is specially used for checking the answer sheets of examinations having multiple choice questions.

Following are some of the important output devices used in a computer.

- Monitors
- Graphic Plotter
- Printer

Monitors

Monitors, commonly called as Visual Display Unit (VDU), are the main output device of a computer. It forms images from tiny dots, called pixels that are arranged in a rectangular form. The sharpness of the image depends upon the number of pixels.

There are two kinds of viewing screen used for monitors.

- Cathode-Ray Tube (CRT)
- Flat-Panel Display

Cathode-Ray Tube (CRT) Monitor

The CRT display is made up of small picture elements called pixels. The smaller the pixels, the better the image clarity or resolution. It takes more than one illuminated pixel to form a whole character, such as the letter 'e' in the word help.

CRT Monitor

A finite number of characters can be displayed on a screen at once. The screen can be divided into a series of character boxes - fixed location on the screen where a standard character can be placed. Most screens are capable of displaying 80 characters of data horizontally and 25 lines vertically.

There are some disadvantages of CRT –

- Large in Size
- High power consumption

Flat-Panel Display Monitor

The flat-panel display refers to a class of video devices that have reduced volume, weight and power requirement in comparison to the CRT. You can hang them on walls or wear them on your wrists. Current uses of flat-panel displays include calculators, video games, monitors, laptop computer, and graphics display.

Flat Monitor

The flat-panel display is divided into two categories –

Emissive Displays – Emissive displays are devices that convert electrical energy into light. For example, plasma panel and LED (Light-Emitting Diodes).

Non-Emissive Displays – Non-emissive displays use optical effects to convert sunlight or light from some other source into graphics patterns. For example, LCD (Liquid-Crystal Device).

Printers

Printer is an output device, which is used to print information on paper.

There are two types of printers –

- Impact Printers
- Non-Impact Printers

Impact Printers

Impact printers print the characters by striking them on the ribbon, which is then pressed on the paper.

Characteristics of Impact Printers are the following –

- Very low consumable costs
- Very noisy
- Useful for bulk printing due to low cost
- There is physical contact with the paper to produce an image

These printers are of two types –

- Character printers
- Line printers
- Character Printers

Character printers are the printers which print one character at a time.

These are further divided into two types:

- Dot Matrix Printer(DMP)
- Daisy Wheel

Dot Matrix Printer

In the market, one of the most popular printers is Dot Matrix Printer. These printers are popular because of their ease of printing and economical price. Each character printed is in the form of pattern of dots and head consists of a Matrix of Pins of size (5*7, 7*9, 9*7 or 9*9) which come out to form a character which is why it is called Dot Matrix Printer.

Dot Matrix Printer

Advantages

- Inexpensive
- Widely Used

- Other language characters can be printed

Disadvantages

- Slow Speed
- Poor Quality

Daisy Wheel

Head is lying on a wheel and pins corresponding to characters are like petals of Daisy (flower) which is why it is called Daisy Wheel Printer. These printers are generally used for word-processing in offices that require a few letters to be sent here and there with very nice quality.

Daisy Wheel Printer

Advantages

- More reliable than DMP
- Better quality
- Fonts of character can be easily changed

Disadvantages

- Slower than DMP
- Noisy
- More expensive than DMP

Line Printers

Line printers are the printers which print one line at a time.

Line Printer

These are of two types –

- Drum Printer
- Chain Printer
- Drum Printer

This printer is like a drum in shape hence it is called drum printer. The surface of the drum is divided into a number of tracks. Total tracks are equal to the size of the paper, i.e. for a paper width of 132 characters, drum will have 132 tracks. A character set is

embossed on the track. Different character sets available in the market are 48 character set, 64 and 96 characters set. One rotation of drum prints one line. Drum printers are fast in speed and can print 300 to 2000 lines per minute.

Advantages

- Very high speed
- Disadvantages
- Very expensive
- Characters fonts cannot be changed

Chain Printer

In this printer, a chain of character sets is used, hence it is called Chain Printer. A standard character set may have 48, 64, or 96 characters.

Advantages

- Character fonts can easily be changed.
- Different languages can be used with the same printer.

Disadvantages

- Noisy

Non-impact Printers

Non-impact printers print the characters without using the ribbon. These printers print a complete page at a time, thus they are also called as Page Printers.

These printers are of two types –

- Laser Printers
- Inkjet Printers

Characteristics of Non-impact Printers

- Faster than impact printers
- They are not noisy
- High quality
- Supports many fonts and different character size

Laser Printers

These are non-impact page printers. They use laser lights to produce the dots needed to form the characters to be printed on a page.

Laser Printer

Advantages

- Very high speed
- Very high quality output
- Good graphics quality
- Supports many fonts and different character size

Disadvantages

- Expensive
- Cannot be used to produce multiple copies of a document in a single printing

Inkjet Printers

Inkjet printers are non-impact character printers based on a relatively new technology. They print characters by spraying small drops of ink onto paper. Inkjet printers produce high quality output with presentable features.

Inkjet Printer

They make less noise because no hammering is done and these have many styles of printing modes available. Color printing is also possible. Some models of Inkjet printers can produce multiple copies of printing also.

Advantages

- High quality printing
- More reliable

Disadvantages

- Expensive as the cost per page is high
- Slow as compared to laser printer

A memory is just like a human brain. It is used to store data and instructions. Computer memory is the storage space in the computer, where data is to be processed and

instructions required for processing are stored. The memory is divided into large number of small parts called cells. Each location or cell has a unique address, which varies from zero to memory size minus one. For example, if the computer has 64k words, then this memory unit has $64 * 1024 = 65536$ memory locations. The address of these locations Cache Memory varies from 0 to 65535.

Memory is primarily of three types –

- Cache Memory
- Primary Memory/Main Memory
- Secondary Memory

Cache memory is a very high speed semiconductor memory which can speed up the CPU. It acts as a buffer between the CPU and the main memory. It is used to hold those parts of data and program which are most frequently used by the CPU. The parts of data and programs are transferred from the disk to cache memory by the operating system, from where the CPU can access them.

Advantages

The advantages of cache memory are as follows –

- Cache memory is faster than main memory.
- It consumes less access time as compared to main memory.
- It stores the program that can be executed within a short period of time.
- It stores data for temporary use.

Disadvantages

The disadvantages of cache memory are as follows –

- Cache memory has limited capacity.
- It is very expensive.

Primary Memory (Main Memory)

Primary memory holds only those data and instructions on which the computer is currently working. It has a limited capacity and data is lost when power is switched off. It is generally made up of semiconductor device. These memories are not as fast as registers.

The data and instruction required to be processed resides in the main memory. It is divided into two subcategories RAM and ROM.

Characteristics of Main Memory

- These are semiconductor memories.
- It is known as the main memory.
- Usually volatile memory.
- Data is lost in case power is switched off.
- It is the working memory of the computer.
- Faster than secondary memories.
- A computer cannot run without the primary memory.

Secondary Memory

This type of memory is also known as external memory or non-volatile. It is slower than the main memory. These are used for storing data/information permanently. CPU directly does not access these memories, instead they are accessed via input-output routines. The contents of secondary memories are first transferred to the main memory, and then the CPU can access it. For example, disk, CD-ROM, DVD, etc.

Characteristics of Secondary Memory

- These are magnetic and optical memories.
- It is known as the backup memory.
- It is a non-volatile memory.
- Data is permanently stored even if power is switched off.
- It is used for storage of data in a computer.
- Computer may run without the secondary memory.
- Slower than primary memories.

RAM (Random Access Memory) is the internal memory of the CPU for storing data, program, and program result. It is a read/write memory which stores data until the machine is working. As soon as the machine is switched off, data is erased.

Access time in RAM is independent of the address, that is, each storage location inside the memory is as easy to reach as other locations and takes the same amount of time. Data in the RAM can be accessed randomly but it is very expensive.

RAM is volatile, i.e. data stored in it is lost when we switch off the computer or if there is a power failure. Hence, a backup Uninterruptible Power System (UPS) is often used with computers. RAM is small, both in terms of its physical size and in the amount of data it can hold.

RAM is of two types –

- Static RAM (SRAM)
- Dynamic RAM (DRAM)

Static RAM (SRAM)

The word **static** indicates that the memory retains its contents as long as power is being supplied. However, data is lost when the power gets down due to volatile nature. SRAM chips use a matrix of 6-transistors and no capacitors. Transistors do not require power to prevent leakage, so SRAM need not be refreshed on a regular basis.

There is extra space in the matrix, hence SRAM uses more chips than DRAM for the same amount of storage space, making the manufacturing costs higher. SRAM is thus used as cache memory and has very fast access.

Characteristic of Static RAM

- Long life
- No need to refresh
- Faster
- Used as cache memory
- Large size
- Expensive
- High power consumption

Dynamic RAM (DRAM)

DRAM, unlike SRAM, must be continually **refreshed** in order to maintain the data. This is done by placing the memory on a refresh circuit that rewrites the data several

hundred times per second. DRAM is used for most system memory as it is cheap and small. All DRAMs are made up of memory cells, which are composed of one capacitor and one transistor.

Characteristics of Dynamic RAM

- Short data lifetime
- Needs to be refreshed continuously
- Slower as compared to SRAM
- Used as RAM
- Smaller in size
- Less expensive
- Less power consumption

ROM stands for **Read Only Memory**. The memory from which we can only read but cannot write on it. This type of memory is non-volatile. The information is stored permanently in such memories during manufacture. A ROM stores such instructions that are required to start a computer. This operation is referred to as **bootstrap**. ROM chips are not only used in the computer but also in other electronic items like washing machine and microwave oven.

Let us now discuss the various types of ROMs and their characteristics.

MROM (Masked ROM)

The very first ROMs were hard-wired devices that contained a pre-programmed set of data or instructions. These kind of ROMs are known as masked ROMs, which are inexpensive.

PROM (Programmable Read Only Memory)

PROM is read-only memory that can be modified only once by a user. The user buys a blank PROM and enters the desired contents using a PROM program. Inside the PROM chip, there are small fuses which are burnt open during programming. It can be programmed only once and is not erasable.

EPROM (Erasable and Programmable Read Only Memory)

EPROM can be erased by exposing it to ultra-violet light for a duration of up to 40 minutes. Usually, an EPROM eraser achieves this function. During programming, an electrical charge is trapped in an insulated gate region. The charge is retained for more than 10 years because the charge has no leakage path. For erasing this charge, ultra-violet light is passed through a quartz crystal window (lid). This exposure to ultra-violet light dissipates the charge. During normal use, the quartz lid is sealed with a sticker.

EEPROM (Electrically Erasable and Programmable Read Only Memory)

EEPROM is programmed and erased electrically. It can be erased and reprogrammed about ten thousand times. Both erasing and programming take about 4 to 10 ms (millisecond). In EEPROM, any location can be selectively erased and programmed. EEPROMs can be erased one byte at a time, rather than erasing the entire chip. Hence, the process of reprogramming is flexible but slow.

Advantages of ROM

The advantages of ROM are as follows –

- Non-volatile in nature
- Cannot be accidentally changed
- Cheaper than RAMs
- Easy to test
- More reliable than RAMs
- Static and do not require refreshing
- Contents are always known and can be verified
- Memory unit is the amount of data that can be stored in the storage unit. This storage capacity is expressed in terms of Bytes.

The following table explains the main memory storage units –

KARPAGAM ACADEMY OF HIGHER EDUCATION

CLASS: II BSC BC

COURSE NAME: BIOINFORMATICS

COURSE CODE: 16BCU404A

UNIT: I BIOINFORMATICS INTRODUCTION BATCH-2016-2019

S. No	Unit & Description
1	Bit (Binary Digit) A binary digit is logical 0 and 1 representing a passive or an active state of a component in an electric circuit
2	Nibble A group of 4 bits is called nibble
3	Byte A group of 8 bits is called byte. A byte is the smallest unit, which can represent a data item or a character
4	Word A computer word, like a byte is a group of fixed number of bits processed as a unit, which varies from computer to computer but is fixed for each computer. The length of a computer word is called word-size or word length. It may be as small as 8 bits or may be as long as 96 bits. A computer stores the information in the form of computer words

The following table lists some higher storage units –

S. No	Unit & Description
1	Kilobyte (KB) 1 KB = 1024 Bytes
2	Megabyte (MB) 1 MB = 1024 KB
3	Gigabyte (GB) 1 GB = 1024 MB
4	Terabyte (TB) 1 TB = 1024 GB
5	Petabyte (PB) 1 PB = 1024 TB

Programming Languages in Bioinformatics

Programming in Perl For the most part, we have focussed on features in programming languages that add structure and discipline to the process of programming. The fundamental idea has always been to achieve suitable degree of abstraction. To ensure the integrity of the abstraction, we have seen different approaches to defining a strong notion of type combined with the flexibility to define polymorphic functions. At the other end of the spectrum we have scripting languages that impose minimal programming discipline but are ideal for rapid deployment of programs that attack “routine” tasks. One of the most successful scripting languages is Perl, an imperative language whose control flow is along the lines of C, but with several interesting features that make it ideal for text processing.

Scalar datatypes

Values associated with scalar variables in Perl are either numbers or character strings. Numbers correspond to double precision floating point numbers. Variable names begin with the special character \$. For instance, we could have the following assignments in a Perl program.

```
$num1 = 7.3;  
$str1 = "Hello123";  
$str2 = "456World";
```

Variables (and their types) do not have to be declared. Moreover, the type of a variable changes dynamically, according to the type of expression that is used to assign it a value. We use normal arithmetic operators to combine numeric values. The symbol . denotes string concatenation. If a numeric value is used in a string expression, it is converted automatically to the string representation of the number. Conversely, if a string is used in a numeric expression, it is converted to a number based on the contents of the string. If the string begins with a number, the resulting value is that number. Otherwise, the value is 0. Continuing our example above, we have the following (text after # is treated as a comment in Perl).

```
$num3 = $num1 + $str1;          # $num3 is now 7.3 + 0 = 7.3
$num4 = $num1 + $str2;          # $num4 is now 7.3 + 456 = 463.3
$str2 = $num4 . $str1;          # $str2 is now "463.3Hello123"
$str1 = $num3 + $str2;          # $str1 is now the number 470.6
$num1 = $num3 . "*" . $num4;    # $num1 is now the string "7.3*463.3"
```

By default, Perl values are made available to the program the moment they are introduced and have global scope and are hence visible throughout the program. Uninitialized variables evaluate to 0 or "" depending on whether they are used as numbers or strings. The scope of a variable can be restricted to the block in which it is defined by using the word `my` when it is first used.

```
my $a = "foo";
if ($some_condition) {
    my $b = "bar";
    print $a; # prints "foo"
    print $b; # prints "bar"
}
print $a; # prints "foo"
print $b; # prints nothing; $b has fallen out of scope
```

Arrays

Perl has an array type, which is more like a list in Haskell because it can grow and shrink dynamically. However, unlike arrays in Java or C and lists in Haskell, Perl arrays can contain a mixture of numeric and string values. Collectively, a Perl array variable starts with the symbol `@`, thus distinguishing it from a scalar. An array can be explicitly written using the notation `(element1,element2,...,elementk)`. Perl supports array like notation to access individual elements of an array—array indices run from 0, as in C or Java. The element at position `$i` in array `@list` is referred to as `$list[$i]`, not `@list[$i]`. A useful mnemonic is that an element of an array is a scalar, so we use `$`, not `@`. The `[...]` notation can be used after the `@` name of the array to denote a sublist. In this form, we provide a list of

the indices to be picked out as a sublist. This list of indices can be in any order and can have repetitions.

```
@list = ("zero",1,"two",3,4);      # Note the mixed types
$val1 = $list[3];                  # $val1 is now 3
$list[4] = "four";                  # @list is now ("zero",1,"two",3,"four")
@list2 = @list[4]; # @list2 is ("four")
@list3 = @list[4,1,3,4,0]; # @list3 is ("four",1,3,"four",0)
```

A key fact about Perl lists is that they are always flat—nested lists are automatically flattened out. Thus, we don't need to write functions such as `append` to combine lists.

```
@list1 = (1,"two");
@list2 = (3,4);
@list = (@list1,@list2);           # @list is (1,"two",3,4)
$list = ("zero",@list);            # @list is now ("zero",1,"two",3,4)
```

The example above shows that we can use a list variable on the left hand side of an assignment. We can also combine scalar variables into a list to achieve a multiple parallel assignment. If we use a list variable as part of the multiple assignment on the left hand side it “soaks” up all the remaining values.

```
($first,$second) = ($list[0],$list[1]);
($first,@rest) = @list;             # $first is $list[0]
                                   # @rest is @list[1,2,...]
($first,@rest,$last) = @list; # $first is $list[0]
                                   # @rest is @list[1,2,...]
                                   # $last gets no value
($this,$that) = ($that,$this); # Swap two values!
```

Perl makes a distinction between list context and scalar context. The same expression often yields different values depending on the context in which it is placed. One example of this is that a list variable `@list`, when used in scalar context, evaluates to the length of the list.


```
@n = @list;      # Copies @list into @n
$n = @list;      # $n is the length of @list
$m = $list[@list-1]; # $m is the last element of @list
```

Perl has several builtin functions to manipulate arrays. For instance, reverse @list reverses a list. The function shift @list removes and returns the leftmost element of @list. The function pop @list removes and returns the last element of @list. The function push(@list,\$value) adds \$value at the end of @list while unshift(@list,\$value) adds \$value at the beginning of @list.

Control flow

Perl supports the normal branching and looping constructs found in C or Java:

- if (condition) {statement-block} else {statement-block}
- while(condition) {statement-block}
- for (init; condition; step) {statement-block}
- if (condition) {statement-block} elsif {statement-block}
- elsif {statement-block} ... else {statement-block}
- unless (condition) {statement-block}
- until (condition) {statement-block}
- For stepping through a list, Perl provides a special loop called
- foreach. foreach \$value (@list){ # \$value is assigned to each item in
- print \$value, "\n"; # @list in turn }

Input/Output

Files can be opened for input or output using the open() function. As in languages like C, the open() statement associates a filehandle with a filename and a mode for the file, namely input, output or append. Here are some examples. In these examples the function die exits after printing a diagnostic. It is separated by an or whose semantics is that the second part of the or is executed only if the first part fails.

```
open(INFILE, "input.txt") or die "Can't open input.txt";
# Open in read mode -- could also write
```

open(INFILE," indicates open in (over)write mode open(LOGFILE, ">>my.log") or die "Can't open logfile"; # >> indicates open in append mode

Reading a file is achieved using the <> operator on a filehandle. In scalar context it reads a single line from the filehandle, while in list context it reads the whole file in, assigning each line to an element of the list:

```
$line = ; # Read one line from INFILE into $line
```

```
@lines = ; # Read all lines from INFILE into list @lines
```

The <> operator is most often seen in a while loop: while (\$line =) { # assigns each line in turn to \$line print "Just read in this line:", \$line; } The example above also introduces the print statement. The simplest form of the print statement takes a list of string arguments separated by commas. print can also take an optional first argument specifying which filehandle to print to:

```
print "Hello, world";
```

```
print "Hello ", $name, "\n";
```

```
print OUTFILE $record;
```

```
print LOGFILE $logmessage;
```

To close a file, we call the function close with the corresponding filehandle

Matching and regular expression

We mentioned in the beginning that Perl is well suited to text manipulation. A particularly useful feature in this regard is a powerful matching operator, that extends naturally to a search-and-replace operator. The match operator is =~ which allows us to match a string variable against a pattern, usually delimited by /. For instance, to print all lines from a file that contain the string CMI, we would write:

```
while ($line = )
```

```
{
```

```
if ($line =~ /CMI/){ print $line;
```

```
}
```

```
}
```

More generally, the pattern could be a regular expression. The syntax for regular expressions is similar to that in text editors like vi or emacs or in the grep command of Unix. In a regular expression, a character stands for itself. A sequence of characters in square brackets stands for a choice of characters. Thus, the pattern `/[Cc][Mm][Ii]/` would match any combination of lower and upper case letters that make up the string `cmi`—for instance, `CMI`, `CmI`, The character `.` is special and matches any character, so, for instance, the pattern `/[Cc].i/` would match any three letter pattern beginning with C or c and ending with i. We can specify a case-insensitive search by appending the modifier `i` at the end of the pattern.

if `($line =~ /CMI/i){ # Same as ($line =~ /[Cc][Mm][Ii]/)` Perl provides some special abbreviations for commonly used choices of alternatives. The expression `\w` (for word) represents any of the characters `_a.. . ,z,A.. . ,Z,0.. . ,9`. The expression `\d` (for digit) represents `0.. . ,9`, while `\s` represents a whitespace character (space, tab or newline). Repetition is described using `*` (zero or more repetitions), `+` (one or more repetitions) and `?` (zero or one repetitions). For instance the expression `\d+` matches a nonempty sequence of digits, while `\s*a\s*` matches a single a along with all its surrounding white space, if any. More controlled repetition is given by the syntax `{m,n}`, which specifies between m and n repetitions. Thus `\d{6,8}` matches a sequence of 6 to 8 digits. A close relative of the match operator is the search and replace operator, which is given by `=~ s/pattern/replacement/`. For instance, we can replace each tab (`\t`) in `$line` by a single space by writing `$line =~ s/\t/ /`; More precisely, this replaces the first tab in `$line` by a space. To replace all tabs we have to add the modifier `g` at the end, as follows. `$line =~ s/\t/ /g`; Often, we need to reuse the portion that was matched in the search pattern in the replacement string. Suppose that we have a file with lines of the form `phone-number name` which we wish to read and print out in the form `name phone-number`. If we match each line against the pattern `/\d+\s*\w.*`, then the portion `\d+` would match the phone number, the portion `\s*` would match all spaces between the phone number and the first part of the name (which could have many parts) and the portion `\w.*` would match the rest of the line, containing all parts of the name. We are interested in reproducing the phone number and the name,

corresponding to the first and third groups of the pattern, in the output. To do this, we group the portions that we want to “capture” within parentheses and then use \1, \2, ... to recover each of the captured portions. In particular, if \$line contains a line of the form phone-number name, to modify it to the new form name phone-number we could write `$line =~ s/(\d+)\s*(\w.+)/\2 \1/; # \1 is what \d+ matches, # \2 is what \w.* matches`. One thing to remember is that if we assigned a value to \$line using the <> operator, then it would initially have a trailing newline character. In the search and replace that we wrote above, this newline character would get included in the pattern \2, so the output would have a new line between the name and the phone number. The function `chomp $line` removes the trailing newline from \$line, if it exists, and should always be used to strip off unwanted newlines when reading data from a file.

Sorting

An extremely useful builtin function in Perl is `sort`, which sorts an array. The default behaviour of `sort` is to use alphabetic sort. So, for instance, to read a list of phone numbers where each line is of the form name:phone number and print them out in alphabetical order we could write:

```
while ($line = <PHONE> )
{ ($name,$phone) = split /:/$line;
$phonehash{$name} = $phone; # Store phone number in a hash
}
foreach $k (sort keys %phonehash){
print $k, ":", $phonehash{$k}, "\n";
}
```

Here, we sort the list generated by keys %phonehash before printout out the values in the hash. What if we want to supply a different basis for comparing values? We can supply `sort` with the name of a comparison function. Then comparison function we supply will be invoked by `sort`. Instead of using @_ to pass parameters to the comparison function (as with normal functions), `sort` will pass the values as \$a and \$b. The comparison function should return -1 if the first argument, \$a, is smaller than the second argument, \$b\$, 0 if the

two arguments are equal, and 1 if the second argument is smaller than the first. So, we can sort using a numeric comparison function as follows.

```
foreach $k (sort bynumber keys %somehash){ ... } sub bynumber { if ($a < $b)
{return -1}; if ($a > $b) {return 1}; return 0; }
```

In fact, this is so common that Perl supplies a builtin “spaceship” operator `<=>` that has this effect. `sub bynumber { return $a <=> $b; # Return -1 if $a < $b, # +1 if $a > $b, # 0 if $a == $b }` The operator `cmp` achieves the same effect as `<=>` but using string comparison instead of numeric comparison. To sort in descending order, we simply invert the position of the arguments in the comparison function.

```
foreach $k (sort bynumdesc keys %somehash){ ... } sub bynumdesc {return $b
<=> $a;}
```

We can also use the arguments `$a` and `$b` in more complex ways. For instance, to print out a hash based on the numeric sorted order of its values, we can write: `foreach $k (sort bystrvalue keys %somehash){ ... } sub bystrvalue {return $somehash{$a} cmp $somehash{$b};}` Finally, we can avoid using a separate named comparison function by just supplying the expression that is to be evaluated in braces, as follows: `foreach $k (sort {$a <=> $b} keys %somehash){ # Same as bynumber` `foreach $k (sort {$b <=> $a} keys %somehash){ # Same as bynumdesc` `foreach $k (sort {$somehash{$a} cmp $somehash{$b}} keys %somehash){ # Same as bystrvalue`

Bioinformatics

Bioinformatics is a new discipline that addresses the need to manage and interpret the data that in the past decade was massively generated by genomic research. This discipline represents the convergence of genomics, biotechnology and information technology, and encompasses analysis and interpretation of data, modeling of biological phenomena, and development of algorithms and statistics. Bioinformatics is by nature a cross-disciplinary field that began in the 1960s with the efforts of Margaret O. Dayhoff, Walter M. Fitch, Russell F. Doolittle and others and has matured into a fully developed discipline. However, bioinformatics is wide-encompassing and is therefore difficult to define. For many, including myself, it is still a nebulous term that encompasses molecular evolution, biological modeling, biophysics, and systems biology. For others, it is plainly

computational science applied to a biological system. Bioinformatics is also a thriving field that is currently in the forefront of science and technology. Our society is investing heavily in the acquisition, transfer and exploitation of data and bioinformatics is at the center stage of activities that focus on the living world. It is currently a hot commodity, and students in bioinformatics will benefit from employment demand in government, the private sector, and academia.

With the advent of computers, humans have become 'data gatherers', measuring every aspect of our life with inferences derived from these activities. In this new culture, everything can and will become data (from internet traffic and consumer taste to the mapping of galaxies or human behavior). Everything can be measured (in pixels, Hertz, nucleotide bases, etc), turned into collections of numbers that can be stored (generally in bytes of information), archived in databases, disseminated (through cable or wireless conduits), and analyzed. We are expecting giant pay-offs from our data: proactive control of our world (from earthquakes and disease to finance and social stability), and clear understanding of chemical, biological and cosmological processes. Ultimately, we expect a better life. Unfortunately, data brings clutter and noise and its interpretation cannot keep pace with its accumulation. One problem with data is its multi-dimensionality and how to uncover underlying signal (patterns) in the most parsimonious way (generally using nonlinear approaches).

Another problem relates to what we do with the data. Scientific discovery is driven by falsifiability and imagination and not by purely logical processes that turn observations into understanding. Data will not generate knowledge if we use inductive principles.

The gathering, archival, dissemination, modeling, and analysis of biological data falls within a relatively young field of scientific inquiry, currently known as 'bioinformatics', 'Bioinformatics was spurred by wide accessibility of computers with increased compute power and by the advent of *genomics*. Genomics made it possible to acquire nucleic acid sequence and structural information from a wide range of genomes at an unprecedented pace and made this information accessible to further analysis and experimentation. For example, sequences were matched to those coding for globular proteins of known structure

(defined by crystallography) and were used in high-throughput combinatorial approaches (such as DNA microarrays) to study patterns of gene expression. Inferences from sequences and biochemical data were used to construct metabolic networks. These activities have generated terabytes of data that are now being analyzed with computer, statistical, and machine learning techniques. The sheer number of sequences and information derived from these endeavors has given the false impression that imagination and hypothesis do not play a role in acquisition of biological knowledge. However, bioinformatics becomes only a science when fueled by hypothesis-driven research and within the context of the complex and ever changing living world.

The science that relates to bioinformatics has many components. It usually relates to biological molecules and therefore requires knowledge in the fields of biochemistry, molecular biology, molecular evolution, thermodynamics, biophysics, molecular engineering, and statistical mechanics, to name a few. It requires the use of computer science, mathematical, and statistical principles. Bioinformatics is in the cross roads of experimental and theoretical science. Bioinformatics is not only about modeling or data 'mining', it is about understanding the molecular world that fuels life from evolutionary and mechanistic perspectives. It is truly inter-disciplinary and is changing. Much like biotechnology and genomics, bioinformatics is moving from applied to basic science, from developing tools to developing hypotheses.

Definition

- Bioinformatics is the recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence (genes and RNAs), protein sequence and structural information. This includes databases of the sequences and structural information as well methods to access, search, visualize and retrieve the information.
- Sequence data can be used to make predictions of the functions of newly identified genes.

- Estimate evolutionary distance in phylogeny reconstruction, determine the active sites of enzymes, construct novel mutations and characterize alleles of genetic diseases to name just a few uses.
- Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline.
- The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

There are three important sub-disciplines within bioinformatics involving computational biology:

- The development of new algorithms and statistics with which to assess relationships among members of large data sets;
- The analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and
- The development and implementation of tools that enable efficient access and management of different types of information.

History of Bioinformatics

- The history of biology in general, B.C. and before the discovery of genetic inheritance by G. Mendel in 1865, is extremely sketch and inaccurate. This was the start of Bioinformatics history.
- G. Mendel is known as the "Father of Genetics". He did experiment on the cross-fertilization of different colors of the same species.
- Mendel illustrated that the inheritance of traits could be more easily explained if it was controlled by factors passed down from generation to generation.
- The understanding of genetics has advanced remarkably in the last thirty years. In 1972, Paul berg made the first recombinant DNA molecule using ligase.
- In that same year, Stanley Cohen, Annie Chang and Herbert Boyer produced the first recombinant DNA organism.

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

Chronological History of Bioinformatics

- 1953 - Watson & Crick proposed the double helix model for DNA based x-ray data obtained by Franklin & Wilkins.
- 1954 - Perutz's group develops heavy atom methods to solve the phase problem in protein crystallography.
- 1955 - The sequence of the first protein to be analyzed, bovine insulin is announced by F. Sanger.
- 1969 - The ARPANET is created by linking computers at Stanford and UCLA.
- 1970 - The details of the Needleman-Wunsch algorithm for sequence comparison are published.
- 1972 - The first recombinant DNA molecule is created by Paul Berg and his group.
- 1973 - The Brookhaven Protein DataBank is announced. Robert Metcalfe receives his Ph.D from Harvard University. His thesis describes Ethernet.
- 1974 - Vint Cerf and Robert Khan develop the concept of connecting networks of computers into an "internet" and develop the Transmission Control Protocol (TCP).
- 1975 - Microsoft Corporation is founded by Bill Gates and Paul Allen. Two-dimensional electrophoresis, where separation of proteins on SDS polyacrylamide gel is combined with separation according to isoelectric points, is announced by P.H.O'Farrel.
- 1988 - The National Centre for Biotechnology Information (NCBI) is established at the National Cancer Institute. The Human Genome Initiative is started (commission on Life Sciences, National Research council. Mapping and sequencing the Human Genome, National Academy Press: Washington, D.C.), 1988. The FASTA algorithm for sequence comparison is published by Pearson and Lipmann. A new program, an

Internet computer virus designed by a student, infects 6,000 military computers in the US.

- 1989 - The genetics Computer Group (GCG) becomes a private company. Oxford Molecular Group, Ltd.(OMG) founded, UK by Anthony Marchigton, David Ricketts, James Hiddleston, Anthony Rees, and W. Graham Richards. Primary products: Anaconds, Asp, Cameleon and others (molecular modeling, drug design, protein design).
- 1990 - The BLAST program (Altschul, et al) is implemented. Molecular applications group is founded in California by Michael Levitt and Chris Lee. Their primary products are Look and SegMod which are used for molecular modeling and protein design. InforMax is founded in Bethesda, MD. The company's products address sequence analysis, database and data management, searching, publication graphics, clone construction, mapping and primer design.
- 1991 - The research institute in Geneva (CERN) announces the creation of the protocols which make -up the World Wide Web. The creation and use of expressed sequence tags (ESTs) is described. Incyte Pharmaceuticals, a genomics company headquartered in Palo Alto California, is formed. Myriad Genetics, Inc. is founded in Utah. The company's goal is to lead in the discovery of major common human disease genes and their related pathways.

Major events in Computational Methods and Computational Biology

- 1993 - CuraGen Corporation is formed in New Haven, CT. Affymetrix begins independent operations in Santa Clara, California.
- 1994 - Netscape Communications Corporation founded and releases Navigator, the commercial version of NCSA's Mozilla. Gene Logic is formed in Maryland. The PRINTS database of protein motifs is published by Attwood and Beck.
- 1995 - *The Haemophilus influenza* genome (1.8) is sequenced. The *Mycoplasma genitalium* genome is sequenced.

KARPAGAM ACADEMY OF HIGHER EDUCATION

CLASS: II BSC BC

COURSE NAME: BIOINFORMATICS

COURSE CODE: 16BCU404A

UNIT: I BIOINFORMATICS INTRODUCTION BATCH-2016-2019

- 1996 - The genome for *Saccharomyces cerevisiae* (baker's yeast, 12.1 Mb) is sequenced. The prosite database is reported by Bairoch, et al. Affymetrix produces the first commercial DNA chips.
- 1997 - The genome for *E. coli* (4.7 Mbp) is published. Oxford Molecular Group acquires the Genetics Computer Group. LION bioscience AG founded as an integrated genomics company with strong focus on bioinformatics. The company is built from IP out of the European Molecular Biology Laboratory (EMBL), the European Bioinformatics Institute (EBI), the German Cancer Research Center (DKFZ), and the University of Heidelberg. Paradigm Genetics Inc., a company focused on the application of genomic technologies to enhance worldwide food and fiber production, is founded in Research Triangle Park, NC. decode genetics publishes a paper that described the location of the FET1 gene, which is responsible for familial essential tremor, on chromosome 13 (Nature Genetics).
- 1998 - The genomes for *Caenorhabditis elegans* and baker's yeast are published. The Swiss Institute of Bioinformatics is established as a non-profit foundation. Craig Venter forms Celera in Rockville, Maryland. PE Informatics was formed as a center of Excellence within PE Biosystems.
- This center brings together and leverages the complementary expertise of PE Nelson and Molecular Informatics, to further complement the genetic instrumentation expertise of Applied Biosystems. Inpharmatica, a new Genomics and Bioinformatics company, is established by University College London, the Wolfson Institute for Biomedical Research, five leading scientists from major British academic centers and Unibio Limited. Gene Formatics, a company dedicated to the analysis and predication of protein structure and function, is formed in San Diego. Molecular Simulations Inc. is acquired by Pharmacopeia.
- 1999 - deCode genetics maps the gene linked to pre-eclampsia as a locus on chromosome 2p13.

- 2000 - The genome for *Pseudomonas aeruginosa* (6.3 Mbp) is published. The *A. thaliana* genome (100 Mb) is sequenced. The *D. melanogaster* genome (180 Mb) is sequenced. Pharmacoepia acquires Oxford Molecular Group.
- 2001 - The human genome (3,000 Mbp) is published.

Genome sequencing

- 1990s did advances in sequencing technology make it feasible to sequence the entire genome of anything more complex than a bacterium.
- DNA sequencing includes several methods and technologies that are used for determining the order of the nucleotide base adenine, guanine, cytosine, and thymine in a molecule of DNA.

Scope of Bioinformatics in Genomics and Proteomics

Genomics

The study of genes and their function. Genomics aims to understand the structure of the genome, including the mapping genes and sequencing the DNA. Genomics examines the molecules mechanisms and the interplay of genetic and environmental factors in disease.

Genomics includes:

- **Functional genomics**- the characterization of genes and their mRNA and protein products.
- **Structural genomics**- the dissection of the architectural features of genes and chromosomes.
- **Comparative genomics**- the evolutionary relationships between the genes and proteins of different species.
- **Epigenomics (epigenetics)**- DNA methylation patters, imprinting and DNA packaging.
- **Pharmacogenomics**- new biological targets and new ways to design drugs and vaccines.

Human Genome Projects

Goals:

- Identify all the approximate 30,000 genes in human DNA
- Determine the sequences of the 3 billion base pairs that make up human DNA
- Sequence the genomes of other model organisms including *Escherichia coli*, yeast (*Saccharomyces cerevisiae*), the fruit fly *Drosophila melanogaster*, the worm *Caenorhabditis elegans* and the laboratory mouse,
- Store this information in databases
- Improve tools for data analysis
- Transfer related technologies to the private sector and
- Address the ethical, legal and social issues (ELSI) that may arise from the project.

Milestones:

- 1990: project initiated as joint effort of US Department of Energy and the National Institutes of Health
- June 2000: Completion of a working draft of the entire human genome
- February 2001: Analyses of the working draft are published
- April 2003: HGP sequencing is completed and project is declared finished two years ahead of schedule.

What does the draft human genome sequence tell us?

By the numbers

- The human genome contains 3 billion chemical nucleotide bases (A, C, T, and G).
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.
- The total number of genes is estimated at around 20,000- 25,000 much lower than previous estimates of approximately 100,000.
- Almost all (99.9%) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50% of discovered genes.

How it's arranged

- The human genome's gene-dense "urban centers" are predominantly composed of the DNA building blocks G and C.
- In contrast, the gene-poor "deserts" are rich in the DNA building blocks A and T. The GC and AT rich regions usually can be seen through a microscope as light and dark bands on chromosomes.
- Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between.
- Stretches of up 30,000 C and G bases repeating over and over often occur adjacent to gene rich areas, forming a barrier between the genes and the "junk DNA". These CpG islands are believed to help regulated gene activity.
- Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231).

Future Challenges:

What we still don't know

- Gene number, exact locations, and functions
- Gene regulation
- DNA sequence organization
- Chromosomal structure and organization
- Noncoding DNA types, amount, distribution, information content and functions
- Coordination of gene expression, protein synthesis, and post-translational events
- Interaction of proteins in complex molecular machines
- Predicted vs experimentally determined gene function
- Evolutionary conservation among organisms
- Protein conservation (Structure and function)
- Proteomes (total protein content and function) in organisms
- Correlation of SNPs (single-base DNA variations among individuals) with health and disease

- Disease susceptibility prediction based on gene sequence variation
- Genes involved in complex traits and multigene diseases
- Complex systems biology including microbial consortia useful for environmental restoration
- Developmental genetics, genomics

Anticipated Benefits of Genome Research

- Molecular medicine
- Improve diagnosis of disease
- Detect genetic predispositions to disease
- Create drugs based on molecular information
- Use gene therapy and control systems as drugs
- Design “custom drugs” based on individual genetic profiles

Microbial genomics

- Rapidly detect and treat pathogens (disease causing microbes) in clinical practice
- Develop new energy sources (biofuels)
- Monitor environments to detect pollutants
- Protect citizenry from biological and chemical warfare
- Clean up toxic waste safely and efficiently

Risk assessment

- Evaluate the health risks faced by individuals who may be exposed to radiation (including low levels in industrial areas) and to cancer causing chemicals and toxins.
- Bioarchaeology, Anthropology, Evolution and Human migration
- Study evolution through germline mutations in lineages
- Study migration of different population groups based on maternal inheritance
- Study mutations on the Y chromosome to trace lineage and migration of males
- Compare breakpoints in the evolution of mutations with ages of populations and historical events.

DNA identification

- Identify potential suspects whose DNA may match evidence left at crime scenes
- Exonerate persons wrongly accused of crimes
- Identify crime and catastrophe victims
- Establish paternity and other family relationships
- Identify endangered and protected species as an aid to wildlife officials
- Detect bacteria and other organisms that may pollute air, water, soil and food
- Match organ donors with recipients in transplant programs
- Determine pedigree for seed or livestock breeds
- Authenticate consumables such as caviar and wine

Agriculture, Livestock, Breeding, and Bioprocessing

- Grow disease, insect, and drought resistant crops
- Breed healthier, more productive, diseases resistant farm animals
- Grow more nutritious produce
- Develop biopesticides
- Incorporate edible vaccines incorporated into food products
- Develop new environmental cleanup uses for plants like tobacco
- Cellulosis biomass research for bioenergy

Anticipated benefits

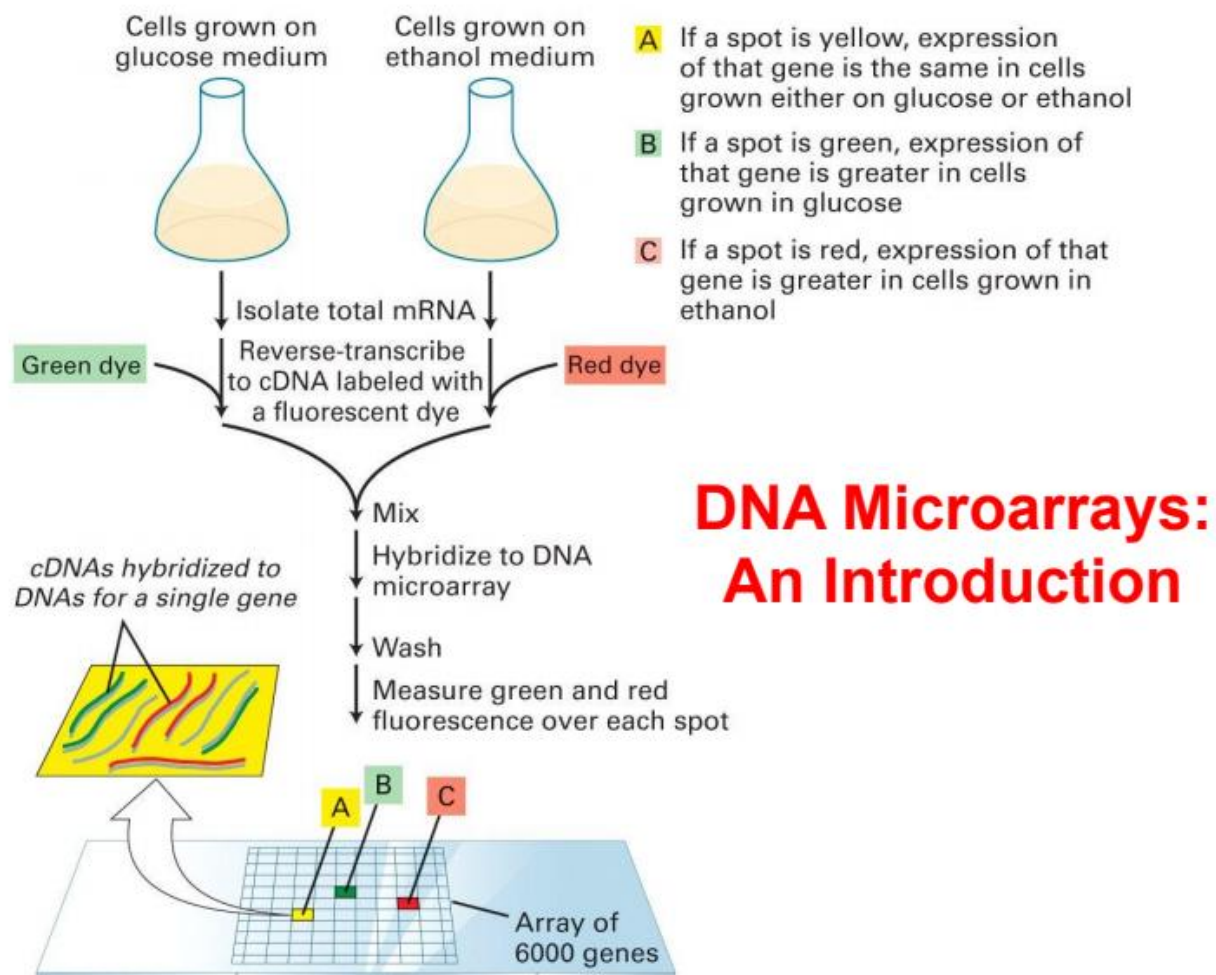
- Improved diagnosis of disease
- Earlier detection of genetic predispositions to disease
- Rational drug design
- Gene therapy and control systems for drugs
- Personalized, custom drugs

ELSI (Ethical, Legal, and Social Issues)

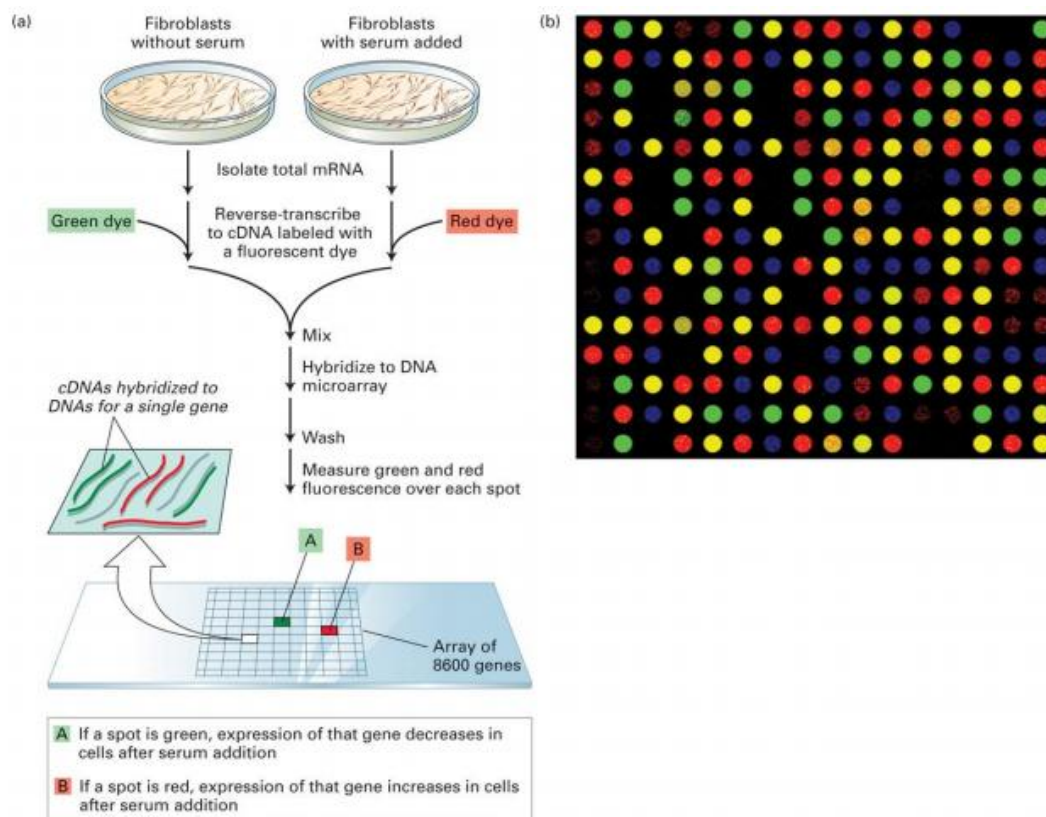
Privacy and confidentiality of genetic information

- Fairness in the use of genetic information by insurers, employers, courts, schools, adoption agencies, and the military, among others.
- Psychological impact, stigmatization and discrimination due to an individual's genetic differences.
- Reproductive issues including adequate and informed consent and use of genetic information in reproductive decision making.
- Clinical issues including the education of doctors and other health service providers, people identified with genetic conditions, and the general public about capabilities, limitations, and social risks; and implementation of standards and quality control measures.
- Uncertainties associated with gene tests for susceptibilities and complex conditions (e.g., heart disease, diabetes, and alzheimer's disease).
- Fairness in access to advanced genomic technologies.
- Conceptual and philosophical implications regarding human responsibility, free will vs genetic determinism and concepts of health and disease.
- Health and environmental issues concerning genetically modified (GM) foods and microbes.
- Commercialization of products including property rights (patents, copyrights, and trade secrets) and accessibility of data and materials.

DNA Microarray



DNA microarray analysis can reveal differences in gene expression in fibroblasts under different experimental conditions

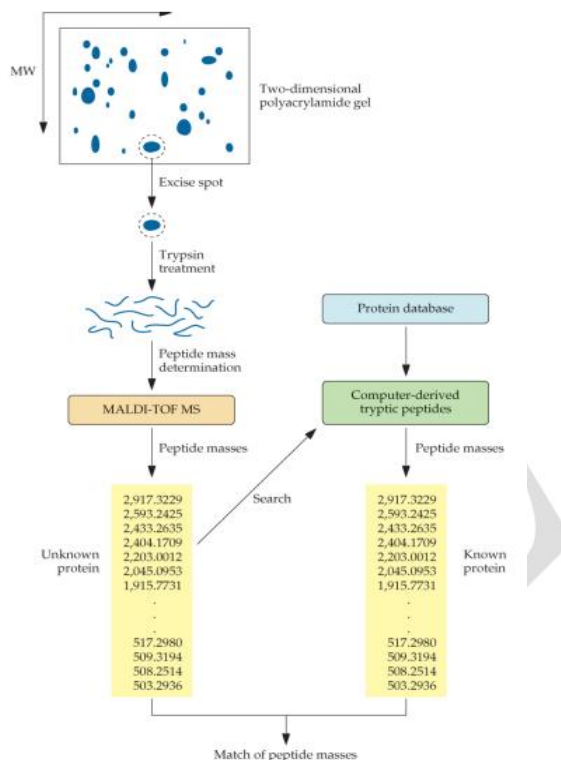


Proteomics

It is the study of proteome, which corresponds to all the proteins expressed in a given tissue under a particular set of conditions.

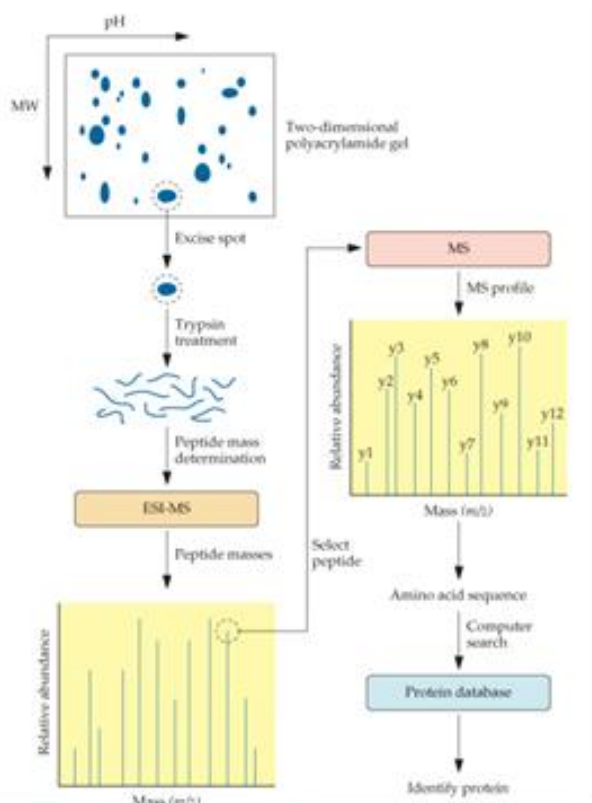
Proteomics

Fig. 5.11 One important aspect of proteomics is to identify proteins obtained from a biological sample. This can be done by peptide mass fingerprinting using a mass spectrometer and matching the data obtained to a gene/protein database.



Proteomics

Fig. 5.12 One important aspect of proteomics is to identify proteins obtained from a biological sample. This can be done by amino acid sequencing of peptides using a mass spectrometer and matching the data obtained to a gene/protein database.



Proteomics: Protein microarrays

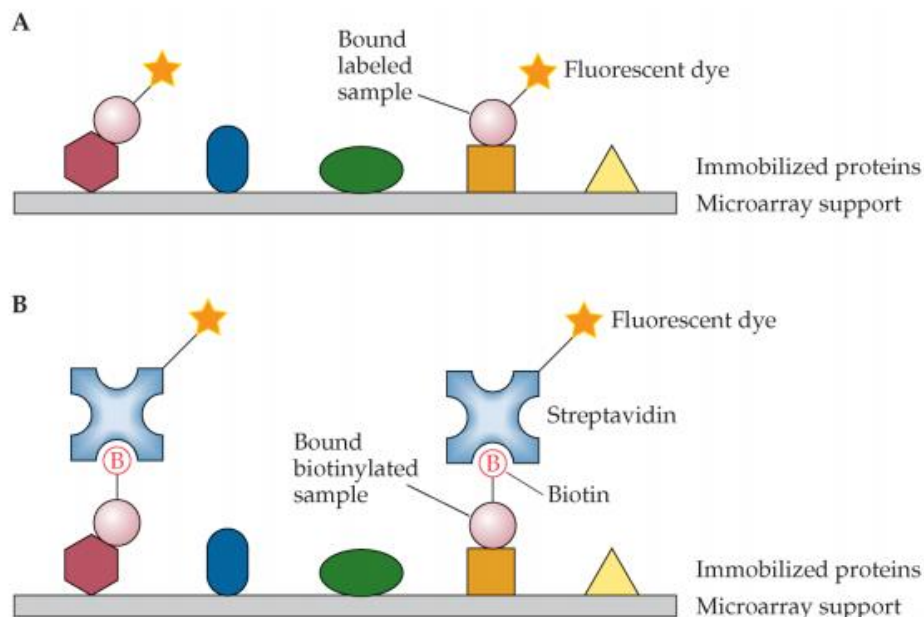


Fig. 5.15 Protein microarrays can be used to examine protein samples and elucidate differences between samples, much like DNA microarrays. Antibody probes are used to examine protein microarrays.

Systems Biology

Systems Biology can be defined as an approach to biology where organisms and biological processes should be analysed and described in terms of their components and their interactions in a framework of mathematical models.

In functional genomics one often uses statements such as 'gene or protein X performs function Y', for example 'the leptin protein regulates the amount of body fat'. But when one looks at this statement, it is clear that it is fundamentally misleading. In the given example, it is clear that the leptin protein is not a machine in itself that computes and performs the regulatory action. Rather, the leptin molecule is a component in a larger system, and it is that system that performs the regulatory function.

Systems Biology begins in the insight that biological processes must be understood in terms of the components that participate in the processes, and that the complexity of biological systems make it difficult, if not impossible, to understand the workings of the system by simple qualitative arguments. Mathematically strict models must be formulated. This is required both in order to be able to capture the actual behaviour of the system with acceptable precision, but also to be able to analyse the fundamental behaviour of the system.

The mathematical models may be very simple (Boolean on/off), or very complex (including detailed descriptions of interactions at a molecular level). The important issue is that it should be possible to analyse the model, either by some mathematical approach, or to simulate it, in order to evaluate its correspondence with the observed facts.

Paradoxically, the complexity of biology is the basis for the development of Systems Biology, at the same time as it is the main reason why computational approaches to biological processes have not been particularly successful in the past. However, the appearance of bioinformatics and functional genomics, and their results (complete genomes, microarray expression analysis, etc) has had a great impact. It now appears possible to obtain data that can be used to build sensible models, and to test them. This is probably the main reason why Systems Biology has become so popular in the last few years.

So far, only very limited results have been obtained. There are only a few, well-studied systems on which any deep analysis has been done. However, there are already some insights that may prove to be generally true. For instance, it seems clear that robustness is a very important factor in biological systems. This is the property that allows a system to absorb fairly large perturbations, and still function reasonably well. The functionally important behaviour of a system has a certain degree of resilience to damage. Some studies have pointed to different ways in which evolution have favoured systems that are robust in different ways.

One important goal of Systems Biology is to understand life processes in sufficient detail to make predictions about their behaviour. If we want to make a particular system

behave in a certain way, how should we change the system, or what type of perturbation should we apply? If we want to make a bacterium produce propanol instead of ethanol, then how should we change the metabolic network of the bacterium? Or, if we want to produce a pharmaceutical drug that can help with deficiencies of the insulin regulatory system that is the basis for diabetes type II (obesity-related diabetes), what components should we focus on? Which are the best drug targets?

Computer aided drug design

Computer-aided drug design, often called structure based drug design involves using the biochemical information of ligand-receptor interaction in order to postulate ligand refinements. For example, if we know the binding site the steric complementarity of the ligand could be improved to increase the affinity for its receptor. Indeed, using the crystal structure of the complex we can target regions of the ligand that fit poorly within the active site and postulate chemical modifications that lower the energetic potential by making more negative van der Waals terms, thus improving complementarity with the receptor. In a similar fashion, functional groups on the ligand can be changed in order to augment electrostatic complementarity with the receptor. When a target is selected for the design of new lead compounds three different situations can be faced regarding the amount of information of the system that is available:

- 1) The structure of the receptor is well known and the bioactive conformation of the ligand is not known,
- 2) Only the bioactive conformation of the ligand is known and
- 3) The target structure and the bioactive conformation of the ligand are unknown

The best possible starting point is an X-ray crystal structure of the target site. If the molecular model of the binding site is precise enough, one can apply docking algorithms that simulate the binding of drugs to the respective receptor site, like Autodock.²⁴ In the first step the program creates a negative image of the target site through the use of several atom probes that determine affinity potentials for each atom type in the substrate molecule

at different points in a grid, place the putative ligands into the site and finally they evaluate the quality of the fit. The program will try a set of different conformers of the ligand in order to obtain the best disposition of the atoms of the molecule for maximizing the scoring function that quantifies ligand receptor interaction. A different strategy for obtaining new lead compounds through rational drug design is the de novo design of ligands with the use of a builder program, like Ligbuilder. 25 This program also determines the shape and the electrostatic properties of the binding site cavity through the use of several atom probes and then it combines from a library of chemical fragments those that better fill the cavity based on steric and electrostatic complementarity.



Design of Drug candidates: An iterative process

The design of new ligands is carried out as step by step procedure

The state of the art design process is based in large part, on a good understanding of molecular recognition of protein-ligand complexes relying upon analogies to other systems and using advanced computerized molecular design programs.

Steps in structure based drug design

The steps used in structure based drug design for designing new lead compounds are

- Obtaining 3D structure of protein

- Active site identification
- Ligand-receptor fit analysis
- Design of new leads

Beginning the Design Phase

Once the phase of analysis is complete, the design phase can start

One has to identify candidate scaffolds with appropriate substituent's that can ensure enhanced interactions with selected sites of the protein

In the case of the optimization of a known series, the information is used to design new analogs.

Eight golden rules in receptor-based ligand design

The important considerations for receptor-based ligand design can be summarized into the following eight rules:

1. Coordinate to key anchoring sites
2. Exploit hydrophobic interactions
3. Exploit hydrogen bonding capabilities
4. Exploit electrostatic interactions
5. Favor bioactive form & avoid energy strain
6. Optimize vdW Contacts and avoid bumps
7. Structural water molecules and solvation
8. Consider entropic effect

Rule 1: Coordinate to Key Anchoring Sites

- When working with target proteins, first one has to consider the proper anchorage of the ligand to key elements of the catalytic site
- This anchorage not only positions the ligand in the active site but also counteracts the effect of de-solvating the two components when binding occurs. This is very important energetically.

Rule 2: Exploit Hydrophobic Interactions

- With hydrophobic pockets, placing a hydrophobic surface of the ligand in hydrophobic sites of the target protein provides an important driving force in complex formation because it reduces non-polar surface areas exposed to water
- Although individually small, the total contribution of hydrophobic forces to drug-receptor interactions is substantial
- Empirical data suggests that the free energy contribution due to hydrophobic forces is approximately 2.9 kJ/mol per methylene group and 8.4 kJ/mol for a benzene ring
- Unlike hydrogen bonds, the hydrophobic interactions are not directional

Rule 3: Exploit Hydrogen Bonding Capabilities

- Unsatisfied hydrogen bond donors and acceptors are rarely seen in proteins and protein-ligand complexes because this would be highly energetically unfavorable
- A carbonyl oxygen is optimally satisfied when it accepts two different hydrogen bonds with C=O --- H angles close to 120°. However hydrogen bonds to carbonyl oxygen atoms with a C=O --- H angle close to 180° form the basis for β -sheet formation and are quite favorable. The average N-H --- O angle is about 155° (with 90% lying between 140° and 180°).
- Almost all protein groups are capable of forming hydrogen bonds like this. Where groups are not explicitly hydrogen bonded, they are probably solvated.

Rule 4: Exploit Electrostatic Interactions

- The optimization of ligand-protein electrostatics can be achieved by placing a positive charge in close vicinity to an enzyme negative charge

Rule 5: Favor Bioactive Form & Avoid Energy Strain

- Conformational energy calculations are performed on each design idea in order to determine the internal penalty required for the new ligand to attain its bioactive binding conformation inside the protein. The internal energy that is required for the small molecule to reach its binding conformation is energy lost in binding.

- Restricting the conformation space of an inhibitor can be beneficial to binding when the conformation is biased towards the bioactive conformer.

Rule 6: Optimize VDW Contacts and Avoid Bumps

- Attractive van der Waals interactions occur over a short distance range and attraction decreases as $1/r^6$. As a result, optimization of attractive van der Waals interactions occurs as the shape of the protein binding site and the shape of the ligand match well.
- Calculations of steric fit are difficult because of possible flexing motions of the protein backbone and especially the residue side chains

Rule 7: Structural Water Molecules and Solvation

- Inhibitor design strategies have great potential when they target the displacement of water molecules tightly bound to the protein by incorporating elements of the water molecule within the inhibitor.
- When polar charged groups are considered in the design of a ligand, one should leave some room for other water molecules to solvate the charged center (except possibly when a salt bridge is formed).

Rule 8: Consider Entropic Effect

- A flexible molecule has a better chance of finding an optimal fit into a receptor, but this is achieved at the cost of large conformational entropy
- Sufficient conformational rigidity is essential to ensure that the loss of entropy upon ligand binding is acceptable
- A rigid molecule has little conformational entropy but is unlikely to fit optimally into the receptor
- An analysis of the contributions of various functional groups to protein-ligand binding demonstrates that each freely rotating bond in a ligand reduces binding free energy by about 2.9 kJ/mol
- Making a flexible molecule more rigid will lead to enhanced activity if the right conformation is maintained

- Example of Successful Structure-Based Design
- The use of the crystallographic structure of the HIV-1 protease in drug design represents one of the more impressive success stories in the structure-based drug design field. Structure-based design studies has resulted in the identification of distinct classes of inhibitors and several successful drug candidates have emerged from these studies and are used in the control of AIDS.
- The HIV-1 protease plays a crucial part in the life cycle of the HIV virus. Inhibitor drugs block the action of the protease and the virus perishes because it is unable to mature into its infectious form.
- The HIV-1 protease is a small dimer enzyme comprising two identical folded 99 amino-acid chains A and B

Ligand-Based Computer-Aided Drug Design

The ligand-based computer-aided drug discovery (LBDD) approach involves the analysis of ligands known to interact with a target of interest. These methods use a set of reference structures collected from compounds known to interact with the target of interest and analyse their 2D or 3D structures. The overall goal is to represent these compounds in such a way that the physicochemical properties most important for their desired interactions are retained, whereas extraneous information not relevant to the interactions is discarded. It is considered as an indirect approach to the drug discovery in that it does not necessitate knowledge of the structure of the target of interest. The two fundamental approaches of LBDD are (1) selection of compounds based on chemical similarity to known actives using some similarity measure or (2) the construction of a quantitative structure activity relationship (QSAR) model that predicts biological activity from chemical structure. The methods are applied for in silico screening for novel compounds possessing the biological activity of interest, hit-to-lead and lead-to drug optimization, and also for the optimization of DMPK/ADMET properties. LBDD is based on the similar property principle which states that molecules that are structurally similar are likely to have similar properties. LBDD approaches in contrast to SBDD approaches can also be applied when the structure of the biological target is unknown. Additionally, active

compounds identified by ligand based virtual high-throughput screening (LB-vHTS) methods are often more potent than those identified in SB-vHTS.

Molecular Descriptors

Molecular descriptors can include properties such as molecular weight, geometry, volume, surface areas, ring content, rotatable bonds, interatomic distances, bond distances, atom types, planar and nonplanar systems, molecular walk counts, electronegativities, polarizabilities, symmetry, atom distribution, topological charge indices, functional group composition, aromaticity indices, solvation properties, and many others. These descriptors are generated through knowledge-based, graph-theoretical methods, molecular mechanical, or quantum-mechanical tools and are classified according to the Chapter 1 Computer Aided Drug Design: An Overview 16 “dimensionality” of the chemical representation from which they are computed: 1- dimensional (1D), scalar physicochemical properties such as molecular weight; 2D, molecular constitution-derived descriptors; 2.5D, molecular configuration-derived descriptors; 3D, molecular conformation-derived descriptors. These different levels of complexity, however, are overlapping with the more complex descriptors, often incorporating information from the simpler ones.

Molecular Fingerprint and Similarity Searches

Molecular fingerprint-based techniques attempt to represent molecules in such a way as to allow rapid structural comparison in an effort to identify structurally similar molecules or to cluster collections based on structural similarity. These methods are fewer hypotheses driven and less computationally expensive than pharmacophore mapping or QSAR models. They rely entirely on chemical structure and omit compound with known biological activity, making the approach more qualitative in nature than other LBDD approaches. Additionally, fingerprint-based methods consider all parts of the molecule equally and avoid focusing only on parts of a molecule that are thought to be most important for activity. This is less error prone to overfitting and requires smaller datasets to begin with. Fingerprint methods may be used to search databases for compounds similar in structure to a lead query, providing an extended collection of compounds that can be

tested for improved activity over the lead. In many situations, 2D similarity searches of databases are performed using chemotype information from first generation hits, leading to modifications that can be evaluated computationally or ordered for in vitro testing.

Quantitative Structure-Activity Relationship Models

Quantitative structure-activity relationship (QSAR) models describe the mathematical relation between structural attributes and target response of a set of chemicals. Classic QSAR is known as the Hansch-Fujita approach and involves the correlation of various electronic, hydrophobic, and steric features with biological activity. In the 1960s, Hansch and others began to establish QSAR models using various molecular descriptors to physical, chemical, and biological properties focused on providing computational estimates for the bioactivity of molecules. In 1964, Free and Wilson developed a mathematical model relating the presence of various chemical substituents to biological activity (each type of chemical group was assigned an activity contribution), and the two methods were later combined to create the Hansch/ Free-Wilson method. A model is then generated to identify the relationship between those descriptors and their experimental activity, maximizing the predictive power. Finally, the model is applied to predict activity for a library of test compounds that were encoded with the same descriptors. Success of QSAR, therefore, depends not only on the quality of the initial set of active/inactive compounds but also on the choice of descriptors and the ability to generate the appropriate mathematical relationship. One of the most important considerations regarding this method is the fact that all models generated will be dependent on the sampling space of the initial set of compounds with known activity, the chemical diversity. In other words, divergent scaffolds or functional groups not represented within this “training” set of compounds will not be represented in the final model, and any potential hits within the library to be screened that contain these groups will likely be missed. Therefore, it is advantageous to cover a wide chemical space within the training set.

3D-QSAR

Comparative field molecular analysis (CoMFA) is a 3D-QSAR technique that aligns molecules and extracts aligned features that can be related to biological activity. This

method focuses on the alignment of molecular interaction fields rather than the features of each individual atom. CoMFA was established over 20 years ago as a standard technique for constructing 3D models in the absence of direct structural data of the target. In this method, molecules are aligned based on their 3D structures on a grid and the values of steric (van der Waals interactions) and electrostatic potential energies (Coulombic interactions) are calculated at each grid point. A comparative molecular similarity index (CoMSIA) is an important extension to CoMFA. In CoMSIA, the molecular field includes hydrophobic and hydrogen-bonding terms in addition to the steric and coulombic contributions. Similarity indices are calculated instead of interaction energies by comparing each ligand with a common probe and Gaussian-type functions are used to avoid extreme values.

Multidimensional QSAR

4D and 5D Descriptors Multidimensional QSAR (mQSAR) seeks to quantify all energy contributions of ligand binding including removal of solvent molecules, loss of conformational entropy, and binding pocket adaptation. 4D-QSAR is an extension of 3D-QSAR that treats each molecule as an ensemble of different conformations, orientations, tautomers, stereoisomers, and protonation states. The fourth dimension in 4D-QSAR refers to the ensemble sampling of spatial features of each molecule. A receptor-independent (RI) 4D-QSAR method was proposed by Hopfinger in 1997. This method begins by placing all molecules into a grid and assigning interaction pharmacophore elements to each atom in the molecule (polar, nonpolar, hydrogen bond donor, etc.). Molecular dynamics simulations are used to generate a Boltzmann weighted conformational ensemble of each molecule within the grid. Trial alignments are performed within the grid across the different molecules, and descriptors are defined based on occupancy frequencies within each of these alignments. These descriptors are called grid cell occupancy descriptors. A conformational ensemble of each compound is used to generate the grid cell occupancy descriptors rather than a single conformation. 5D-QSAR has been developed to account for local changes in the binding site that contribute to an induced fit model of ligand binding. In a method developed by Vedani and Dobler, induced fit is simulated by mapping a “mean

envelope” for all ligands in a training set on to an “inner envelope” for each individual molecule. Their method involves several protocols for evaluating induced-fit models including a linear scale based on the adaptation of topology, adaptations based on property fields, energy minimization, and lipophilicity potential. By using this information, the energetic cost for adaptation of the ligand to the binding site geometry is calculated. Vedani from the Biographics Laboratory developed a receptor modeling concept, Quasar, based on 6D-QSAR that explicitly allows for the simulation of induced fit. Quasar concept, previously 3,4,5D extended to six dimensions allows for the simultaneous consideration of different solvation models which can be achieved explicitly by mapping parts of the surface area with solvent properties (position and size are optimized by the genetic algorithm).

Pharmacophore Mapping

In 1998, the International Union of Pure and Applied Chemistry (IUPAC) formally defined a pharmacophore as “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response”. In terms of drug activity, it is the spatial arrangement of functional groups that a compound or drug must contain to evoke a desired biological response. Therefore, an effective pharmacophore will contain information about functional groups that interact with the target, as well as information regarding the type of noncovalent interactions and interatomic distances between these functional groups/interactions. A pharmacophore model of the target binding site summarizes steric and electronic features needed for optimal interaction of a ligand with a target. Most common properties that are used to define pharmacophores are hydrogen bond acceptors, hydrogen bond donors, basic groups, acidic groups, partial charge, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. Pharmacophore features have been used extensively in drug discovery for virtual screening, de novo design, and lead optimization. A pharmacophore model of the target binding site can be used to virtually screen a compound library for putative hits. Apart from querying database for active compounds, pharmacophore models can also be used by de novo design algorithms to guide the design of new compounds. Structure-based

pharmacophore methods are developed based on an analysis of the target binding site or based on a target-ligand complex structure. Ligand Scout uses protein-ligand complex data to map interactions between ligand and target. A knowledge based rule set obtained from the PDB is used to automatically detect and classify interactions into hydrogen bonds, charge transfers, and lipophilic regions. The algorithm creates regularly spaced grids around the ligand and the surrounding residues. Probe atoms that represent a hydrogen bond donor, a hydrogen bond acceptor, and a hydrophobic group are used to scan the grids. An empirical scoring function, SCORE, is used to describe the binding constant between probe atoms and the target. SCORE includes terms to account for van der Waals interactions, metal-ligand bonding, hydrogen bonding, and desolvation effects upon binding. A pharmacophore model is developed by rescoring the grids followed by clustering and sorting to extract features essential for protein-ligand interaction. The most common software packages used for ligand based pharmacophore generation include Phase, MOE, Catalyst, DISCO, and GASP.

Applications of Bioinformatics

- Bioinformatics is the use of IT in biotechnology for the data storage, data warehousing and analyzing the DNA sequences.
- In Bioinformatics knowledge of many branches are required like biology, mathematics, computer science, laws of physics & chemistry, and of course sound knowledge of IT to analyze biotech data.
- Bioinformatics is not limited to the computing data, but in reality it can be used to solve many biological problems and find out how living things works.

Bioinformatics is being used in following fields:

- Molecular medicine
- Personalised medicine
- Preventative medicine
- Gene therapy
- Drug development
- Microbial genome applications

- Waste cleanup
- Climate change Studies
- Alternative energy sources
- Biotechnology

Molecular medicine

- The human genome will have profound effects on the fields of biomedical research and clinical medicine.
- Every disease has a genetic component. This may be inherited (as is the case with an estimated 3000-4000 hereditary disease including Cystic Fibrosis and Huntingtons disease) or a result of the body's response to an environmental stress which causes alterations in the genome (e.g. cancers, heart disease, diabetes.).
- The completion of the human genome means that we can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly.
- This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

Personalised medicine

- Clinical medicine will become more personalised with the development of the field of pharmacogenomics.
- This is the study of how an individual's genetic inheritance affects the body's response to drugs.
- At present, some drugs fail to make it to the market because a small percentage of the clinical patient population show adverse affects to a drug due to sequence variants in their DNA.
- As a result, potentially life saving drugs never makes it to the marketplace.
- Today, doctors have to use trial and error to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment.

- In the future, doctors will be able to analyse a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning.

Preventative medicine

- With the specific details of the genetic mechanisms of diseases being unraveled, the development of diagnostic tests to measure a person's susceptibility to different diseases may become a distinct reality.
- Preventative actions such as change of lifestyle or having treatment at the earliest possible stages when they are more likely to be successful, could result in huge advances in our struggle to conquer disease.

Gene therapy

- In the not too distant future, the potential for using genes themselves to treat disease may become a reality.
- Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of a person's genes.
- Currently, this field is in its infantile stage with clinical trials for many different types of cancer and other diseases ongoing.

Drug development

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

Microbial genome applications

- Microorganisms are ubiquitous, that is they are found everywhere.
- They have been found surviving and thriving in extremes of heat, cold, radiation, salt, acidity and pressure.
- They are present in the environment, our bodies, the air, food and water.

- Traditionally, use has been made of a variety of microbial properties in the baking, brewing and food industries.
- The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications.
- For these reasons, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction.
- By studying the genetic material of these organisms, scientists can begin to understand these microbes at a very fundamental level and isolate the genes that give them their unique abilities to survive under extreme conditions.

Waste cleanup

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

Climate change Studies

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

Alternative energy sources

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

Biotechnology

- The archaeon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmental remediation.
- These microorganisms thrive in water temperatures above the boiling point and therefore may provide the DOE, the Department of Defence, and private companies with heat-stable enzymes suitable for use in industrial processes.
- Other industrially useful microbes include, *Corynebacterium glutamicum* which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine.
- The substance is employed as a source of protein in animal nutrition.
- Lysine is one of the essential amino acids in animal nutrition.
- Biotechnologically produced lysine is added to feed concentrates as a source of protein, and is an alternative to soybeans or meat and bone meal.
- *Xanthomonas campestris* pv. is grown commercially to produce the exopolysaccharide xanthan gum, which is used as a viscosifying and stabilizing agent in many industries.
- *Lactococcus lactis* is one of the most important micro-organisms involved in the dairy industry, it is a non-pathogenic rod-shaped bacterium that is critical for manufacturing dairy products like buttermilk, yogurt and cheese.
- This bacterium, *Lactococcus lactis* ssp., is also used to prepare pickled vegetables, beer, wine, some bread and sausages and other fermented foods.
- Researchers anticipate that understanding the physiology and genetic make-up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of *L. lactis* to serve as a vehicle for delivering drugs.

Antibiotic resistance

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

Forensic analysis of microbes

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

The reality of bioweapon creation

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

Evolutionary studies

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

Crop improvement

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

Insect resistance

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

Improve nutritional quality

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

Development of Drought resistance varieties

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

Veterinary Science

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

Possible Questions

1. Write short notes on input and output devices.
2. Briefly discuss about drug designing concepts.
3. Write about the storage devices of computer.
4. What is application of bioinformatics in medicine.
5. Enumerate the steps involved in designing a therapeutic drug.
6. Write in detail about the storage devices.
7. Define genome.
8. Discuss about scope of bioinformatics in proteomics.
9. Write a SQL program to create and delete database with example.
10. What is docking.
11. Give a detailed account on the structure based method to identify lead compounds.
12. Write a SQL statement to create an employee table with example.
13. Define Proteome.
14. Explain briefly about virtual screening.
15. Discuss in detail about human genome Project add a note on its current developments.

Karpagam Academy of Higher Education
Department of Biochemistry
II B.Sc., Biochemistry
17BCU404A- Bioinformatics

Question number	Unit	Question	Option I	Option II	Option III	Option IV	Answer
1	1	Second generation computers are made of	Vacuum tubes	Transistors	LSI	VLSI	Transistors
2	1	Which of the following memory is non-volatile	SRAM	DRAM	ROM	All of the above	ROM
3	1	The full form of EEPROM is	Electrically Erasable Programmable Read Only Memory	Electrically Erasable Programmable Read Only Memory	Electrically Erasable Programmable Read Only Memory	None of the above	Electrically Erasable Programmable Read Only Memory
4	1	Which is a valid statement	1 KB = 1024 Bytes	1 MB = 1024 Bytes	1 KB = 1000 Bytes	1 MB = 1000 Bytes	1 KB = 1024 Bytes
5	1	Perl stands for	Practical Extraction and Report Language	Practice for Exclusive and Report Language	Practical Extraction and Report Language	Practical Exclusive and Report Language	Practical Extraction and Report Language
6	1	Perl was developed in	1987	1977	1987	1997	1987
7	1	Proteomics is defined as	study of proteins	Sequencing proteins	Analysis of proteins	All of the above	All of the above
8	1	The acronym in SQL, which allows to change the definition of a table is	ALTER	UPDATE	DELETE	SELECT	ALTER
9	1	What is the full form of SQL	Structured Query Language	Structured Query List	Simple Query Language	None of the above	Structured Query Language
10	1	SQL data definition commands make up only	DDL	DML	DTML	NML	DDL
11	1	Which command is used to insert values from the table?	DELETE	INSERT	UPDATE	Both a and b	INSERT
12	1	The vector used for cloning the human genome	YAC	Bacterial	Plasmid vector	Plasmid vector	YAC
13	1	How clones make up the human genome	1,00,000	2 Billion	1 Billion	1000	2 Billion
14	1	Genes are "Open Regions" containing predominantly nucleotides	A and T	G and C	A and C	T and G	A and T
15	1	GC regions can be seen through microscope on chromosomes as	White bands	Dark bands	Light bands	Both b and c	Dark bands
16	1	Gene Pool "Delet Regions" are recombined with nucleotides	Chromosome X	Chromosome Y	Chromosome A	Chromosome 1	A and T
17	1	Genes are of genes are seen as	A and T	G and C	Both a and b	Chromosome 1	Chromosome 1
18	1	Repeated sequences that do not code for proteins is	RNA	rRNA	mRNA	junk DNA	Junk DNA
19	1	AT region can be seen through the microscope on the chromosome as	Dark bands	Light bands	Both a and b	Chromosome 1	Dark bands
20	1	Least number of genes are present in	Chromosome 1	X chromosome	Both a and b	Chromosome 1	X chromosome
21	1	The single base difference is referred as	SNP	SNR	NMR	GAAT	SNP
22	1	What is the difference in mutation rate in drug development process?	Genomics	Genetics	Proteomics	Metabonomics	C-100
23	1	First microarray is developed computer is the market	1985	1985	1981	1980	1980
24	1	The compact Disk is launched in	1981	1986	1974	1964	1980
25	1	CADD stands for	Computer Aided Drug Design	Computer Aided Drug Design	Computer Aided Drug Discovery	Computer Aided Drug Discovery	Computer Aided Drug Design
26	1	Identification of lead molecules based on the receptor features is	Pharmacophore drug design	QSAR	Structure based drug design	Molecular modelling	Structure based drug design
27	1	Identification of lead molecules based on the charge properties of the receptor binding site is	Pharmacophore drug design	QSAR	Structure based drug design	de novo drug design	de novo drug design
28	1	Nuclear magnetic Resonance for protein structure is published in	1980	1986	1985	1970	1980
29	1	Lead molecule screening from the database based of pharmacophore features is	Pharmacophore drug design	de novo drug design	Structure based drug design	de novo drug design	Pharmacophore drug design
30	1	Least human genome consist of	2.9 billion	3.1 billion	3.2 billion	3.3 billion	3.1 billion
31	1	NMR is published in the year	1985	1986	1974	1965	1980
32	1	The human genome consists of base	3 billion	3 billion	3.1 billion	3.2 billion	3.1 billion
33	1	Each field is controlled by	Chromosomes	Genes	Proteins	Metabonomics	Chromosomes
34	1	The genes have different function in human genome	SNP	SNR	NMR	GAAT	SNP
35	1	HTP is controlled on	SNP	SNR	NMR	GAAT	SNP
36	1	Sudden change in base sequence known as	Mutation	Variant	Contam	SNP, AR	Mutation
37	1	Single Base difference is called as	SNP	SNR	NMR	GAAT	SNP
38	1	Application of bioinformatics in medicine	Genetic disease and immunization	Drug discovery	Pharmacokinetic analysis	All the above	All the above
39	1	The human genome material consists of	2.9 billion nucleotides	3 billion	3.1 billion	3.2 billion	3.1 billion
40	1	Watson and Crick proposed the	Protein database	Double helix	Single helix	3-40	Double helix
41	1	Transformation of nucleotide is	Protein	Gene	SNP	Metabonomics	Protein
42	1	The computational methodologies that were used to find the best matches between the donor and the receiver is	molecular docking	molecular docking	molecular docking	molecular docking	molecular docking
43	1	The domain of 3D and molecular modelling are	Genetics	Genetic disease and immunization	Drug discovery	Pharmacokinetic analysis	All the above
44	1	The human genome project was completed in the year of	2003	2005	2004	2002	2003
45	1	The first sequenced protein was	Insulin	Melanin	Phenylalanine	Serotonin	Insulin
46	1	Dr. J. V. Neel proposed a new technique known as electrophoresis for sequencing proteins. In solution	Insulin	Phenylalanine	Phenylalanine	Phenylalanine	Insulin
47	1	Most amino acid sequence data has its availability of	INSR	INSR	INSR	INSR	INSR
48	1	Genbank stored at	NIH	NIH	NIH	NIH	NIH
49	1	Insulin consists of residues	51	51	51	51	51
50	1	Protein was first sequenced in	1954	1955	1956	1957	1955
51	1	A compound that has desirable properties to become a drug is called as	Lead	Lead	Lead	Lead	Lead
52	1	The genome for Biochemistry research	1.1 Mb	1.1 Mb	1.1 Mb	1.1 Mb	1.1 Mb
53	1	Expansion of TIGR	The Institute for Genomic Research	The Institute for Genomic Research	The International for Genomic Research	The International for Genomic Research	The Institute for Genomic Research
54	1	The sequence of the first protein to be analyzed, bovine insulin, is announced by	1955	1955	1956	1957	1955
55	1	The ARFANET is created by linking computers at Stanford and UCLA	1969	1979	1986	1970	1969
56	1	The first recombinant DNA molecule is created by	Clon	Clon	Clon	Clon	Clon
57	1	HTP stand for	Transmission Control Protocol	Transfer protocol	Transmission protocol	Transfer control protocol	Transmission Control Protocol
58	1	Discovery of GVCU	The genetics Computer Group	The genetics Group	Genetic computer	Genetic computer group	The genetics Computer Group
59	1	The FASTA algorithm for sequence comparison is published by	Tringler	Tringler	Tringler and L. Jones	Tringler	Tringler and L. Jones
60	1	Microsoft Corporation is founded by Bill Gates and Paul Allen	1975	1985	2005	2006	1975

UNIT-II

SYLLABUS

Introduction to Biological databases and data retrieval: Primary and secondary and composite databases, NCBI, Nucleic acid databases (GenBank, EMBL), DDBJ, NDB), Protein databases (PIR, Swiss-Prot, TrEMBL, PDB), Metabolic pathway database (KEGG, EcoCyc, and MetaCyc), small molecule database (Pubchem, Drug Bank, ZINC, CSD), Structure Viewers (RasMol, J Mol), File formats.

Data

Data is unprocessed facts and figures without any added interpretation or analysis

Information

Information is data that has been interpreted so that it has meaning for the user.

Database

Is a usually large collection of data organized especially for rapid search and retrieval.

There are many different types of database but for routine sequence analysis, the following are initially the most important

- Primary database
- Secondary database
- Composite database

Primary Database

- Primary databases are produced with experimentally derived data such as nucleotide sequence, protein sequence or macromolecular structure.
- Experimental results are submitted directly into the database by researchers, and the data are essentially archival in nature.
- Once given a database accession number, the data in primary databases are never changed: they form part of the scientific record.

Secondary Database

- Secondary databases comprise data derived from the results of analyzing primary data.

KARPAGAM ACADEMY OF HIGHER EDUCATION

CLASS: II BSC BC

COURSE NAME: BIOINFORMATICS

COURSE CODE: 16BCU404A

UNIT: II BIOLOGICAL DATABASES

BATCH-2016-2019

- Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary).
- They are highly curated, often using a complex combination of computational algorithm and manual analysis and interpretation to derive new knowledge from the public record of science.

	Primary Database	Secondary Database
Synonyms	Archival Database	Curated Database; Knowledgebase
Source of Data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary database
Examples	ENA, GenBank and DDBJ (Nucleotide sequence) Array Express Archieve and GEO (functional genomics data) Protein Data Bank (PDB coordinates of three dimensional macromolecular structure)	InterPro (protein families, motifs and domains) UniProt Knowledgebase (sequence and functional information on proteins) Ensemble (variation, function, regulation and more layered onto whole genome sequences)

Composite databases

- Collection of various primary database sequences
- Renders sequence searching highly efficient as it searches multiple resources
- Example: NRDB (non redundant database), OWL, MIPSX, SWISSPROT, TrEMBL

Nucleic acid Sequence databases

GenBank

- GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences

- GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.
- A GenBank release occurs every two months and is available from the ftp site.
- The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

Search GenBank for sequence identifiers and annotations with Entrez Nucleotide, which is divided into three divisions

CoreNucleotide (the main collection)

- The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB.

dbEST (Expressed Sequence Tags)

- The EST database is a collection of short single-read transcript sequences from GenBank. These sequences provide a resource to evaluate gene expression, find potential variation, and annotate genes.

dbGSS (Genome Survey Sequences)

- The GSS database is a collection of unannotated short single-read primarily genomic sequences from GenBank including random survey sequences clone-end sequences and exon-trapped sequences.

GenBank Data Usage

- The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information.

- Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted.
- NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

Confidentiality

- Some authors are concerned that the appearance of their data in GenBank prior to publication will compromise their work.
- GenBank will, upon request, withhold release of new submissions for a specified period of time.
- A date must be specified; we cannot hold a sequence indefinitely pending publication.
- However, if a paper citing the sequence or accession number is published prior to the specified date, the sequence will be released upon publication.
- In order to prevent the delay in the appearance of published sequence data, we urge authors to inform us of the appearance of the published data.
- As soon as it is available, please send the full publication data--all authors, title, journal, volume, pages and date--to the following address: update@ncbi.nlm.nih.gov

Submission to GenBank

There are several options for submitting data to GenBank:

BankIt, a WWW-based submission tool with wizards to guide the submission process

tbl2asn, a command-line program, automates the creation of sequence records for submission to GenBank using many of the same functions as Sequin. It is used primarily for submission of complete genomes and large batches of sequences and is available by FTP for use on MAC, PC and Unix platforms.

Submission Portal, a unified system for multiple submission types. Currently only ribosomal RNA (rRNA), rRNA-ITS or Influenza sequences can be submitted with the GenBank component of this tool.

Sequin, NCBI's stand-alone submission tool with wizards to guide the submission process is available by FTP for use on for MAC, PC, and UNIX platforms.

EMBL

- European molecular Biology Laboratory
- Nucleic acid database from EBI (European Bioinformatics Institute)
- Produced in collaboration with DDBJ and GenBank
- Search engine – SRS (sequence Retrieval System)
- Keeping with the tremendous growth in field of computational biology, a need was felt to establish an independent and parallel research institute that would act not just as a mirror housing the GenBank nucleotide resources of NCBI, but would also develop matching databases and analysis tools. The European Molecular Biology Laboratory (EMBL) was thus established in 1974 and is now supported with funding from 20 members states of the European Union, Israel and Australia. EMBL currently operates five research institutes in different countries with main institute at Heidelberg, Germany.

The Five institutes of EMBL with their core research activities are

- EMBL Heidelberg (Germany)
- EMBL Grenoble (France)- Structural Biology
- EMBL-European Bioinformatics Institute (Hinxton, UK)- Bioinformatics
- EMBL Hamburg (Germany)-Structural Biology
- EMBL Monterotondo (Italy)-Mouse Biology

The broad goals of EMBL are

- Basic research in Molecular biology
- Training manpower i.e. students, scientist and visitors
- Develop new tools, technologies and methods

- Offer service to the research community
- Transfer technology to industry for commercialization

The following are the broad categories of databases at EBI-EMBL

- Biological ontologies
- Literature
- Functional Genomics or microarray
- Nucleotides
- Pathways and networks
- Protein
- Proteomics
- Small molecules
- Structure

DDBJ

- DNA databank of Japan
- Started in 1986 in collaboration with GenBank
- Produced and maintained at NIG (National Institute of Genetics)
- DDBJ was established in the year 1986 at the National Institute of Genetics (NIG), Japan with support from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). Later on for its efficient functioning, Center for Information Biology (CIB) was established at NIG in 1995. In 2004, NIG was made a member of Research Organization of Information and Systems.
- The functioning and maintenance of DDBJ is monitored by an international advisory committee consisting of 9 members from Japan, Europe and USA. The committee reviews the functioning of DDBJ and reports the progress of DDBJ in database issue of Nucleic acid Research Journal every year. Since its inception there has been a tremendous increase in the number of sequence submitted to DDBJ.

Roles of DDBJ

As a member of INSDC, primary objective of DDBJ is to collect sequence data from researchers all over the world and to issue a unique accession number for each entry. The data collected from the submitters is made publically available and anyone can access the data through data retrieval tools available at DDBJ. Everyday data submitted at either DDBJ or EMBL or NCBI is exchanged, therefore at any given time these three databases contain same data. Following are the steps along with snapshots showing data retrieval from DDBJ using getentry.

- Open the homepage of DDBJ
- Click on the search/Analysis link on the menu bar
- Click on getentry link
- Type in the accession number in the search box and click on search
- Desired sequence will be retrieved.

Software development

DDBJ team continuously focuses on developing new software which can be used for data analysis. For example, WINA (A window Analysis Program for the number of synonymous and nonsynonymous nucleotide substitutions) has been developed by DDBJ. It is tool which helps in visualizing the difference in accumulation of both synonymous and nonsynonymous nucleotide substitutions.

Training courses

DDBJ also focuses on providing teaching assistance on bioinformatics. It conducts Bioinformatics training course which teaches analysis of data.

NDB database

- A repository of three dimensional structural information about nucleic acids
- The NDB is supported by funds from the national science foundation and the department of energy
- The NDB follows the dictionaries and formats used by the worldwide protein data bank

- Search the NDB by ID
- Enter an NDB ID or PDB ID
- Atlas, Deposit Data, Download Data, Search, Education, Standards, Tools, Links
- The NDB Atlas provides summary information and images for each structure in the database. The Atlas is first divided by experimental type and then by structure type.

Features include

- Image of the asymmetric and biological units, and crystal packing pictures for nucleic acid structures from X-Ray crystallographic experiments
- Image of the average and ensemble structure from NMR experiments
- Links to coordinate files, experimental data files
- Tables of derived data, including torsion angles and hydrogen bonding classifications
- Special features for RNA structures, including images of secondary and tertiary structure
- Each page in Atlas is generated directly from NDB using an XML translator which formats the data contained in each NDB files.
- Pictures present on the Atlas pages were generated by different software
- Blocview
- RNAview
- MaxIT
- In the nucleotide block models, adenine is red, thymine is blue, cytosine is yellow, guanine is green, and uracil is cyan. In the atom stick models, carbon is black, oxygen is red, nitrogen is blue, and phosphates are orange.
- The Atlas is first divided by how the structures were determined: by X-ray crystallographic or NMR experiments. Gallery index pages, which include images for each structure on the page, and plain text index pages are offered.

- The NDB processes data for the crystal structures of nucleic acids. Structures can be deposited to the NDB and PDB at the same time using ADIT (the AutoDep Input Tool).
- ADIT accepts coordinates in PDB or mmCIF format and structure factor files. All other information is entered into ADIT by the author.
- Coordinate files can be downloaded from download option
- Basic detail of DNA and RNA is given in education option
- X-plor parameters and geometries are given in standard option

Tools and Software's

RNA viewer- RNA 2-dimensional structure using the RNaview program

Base pair Viewer- RNA base pairs using the BPView program

DNA binding prediction for protein structures are HTHQuery and predictdnahth these predicts whether given three dimensional protein structure contains a DNA-binding Helix-turn-Helix (HTH) structural motif.

QPROF (Query of PROtein Features) A web utility for secondary similarity search of protein three dimensional structure.

RNAView Program Quickly generate display of RNA/DNA secondary structures with tertiary interactions.

RNAMLview Program Display and/or edit RNAView 2-dimensional diagrams

3DNA A software package for the analysis, rebuilding and visualization of three dimensional nucleic acid structures.

Freehelix98 described in "DNA bending: the prevalence of kinkiness and the virtues of normality" Richard E. Dickerson Nucleic Acid Research

Protein Databases**SwissProt**

- Annotated sequence database established in 1986
- Consists of sequence entries of different file formats
- Similar format to EMBL

SwissProt is an annotated protein sequence database which was formulated and managed by Amos Bairoch in 1986. It was established collaboratively by the Department of Medical Biochemistry at the University of Geneva and European Molecular Biology Laboratory (EMBL). Later it shifted to European Bioinformatics Institute (EBI) in 1994 and finally in April 1998, it became a part of Swiss Institute of Bioinformatics (SIB). In 1996, TrEMBL was added as an automatically annotated supplement to Swiss-Prot database. Since 2002, it is maintained by the UniProt consortium and information about a protein sequence can be accessed via the Uniprot website. The universal protein resource is the most widespread protein sequence catalog comprising of EBI, SIB and PIR.

There are four main features of Swiss-Prot**High quality annotation**

It is achieved through manually creating the protein sequence entries. It is processed through 6 stages.

Sequence curation: In this step, identical sequences are extracted through blast search and then the sequence from the related gene and same organism are incorporated into a single entry. It makes sure that the sequence is complete, correct and ready for further curation steps.

Sequence analysis: It is performed by using various sequence analysis tools. Computer predictions are manually reviewed and important results are selected for integration.

Literature curation: In this step, important publications related to the sequence are retrieved from literature databases. The whole text of each article is scanned manually and relevant information is gathered and supplemented to the entry.

Family based curation: Putative homolog's are determined by reciprocal Blast searches and phylogenetic resources which are further evaluated, curated, annotated and propagated across homologous proteins to ensure data consistency.

Evidence attribution: All information incorporated to the sequence entry during manual annotation is linked to the original source so that users can trace back the origin of data and evaluate its.

Quality assurance, integration and update: each completely annotated entry undergoes quality assurance before integration into Swiss-Prot and is updated as new data become available.

Minimum redundancy: during manual annotation, all entries belonging to identical gene and form similar organism are merged into a single entry containing complete information. This results in minimal redundancy.

Integration with other databases: Swiss-Prot is presently cross-referenced to more than 50 specialized documentation files. Documentation file section provides an updated descriptive list of all document files.

PIR

- Protein Information Resource
- A division of National Biomedical Research Foundation (NBRF) in U.S
- One can search for entries or do sequence similarity search at PIR site.

In year 1984, National Biomedical Research Foundation (NBRF) developed PIR for identification and interpretation of information on protein sequences. This database was actually derived from Atlas of Protein Sequence and Structure, which was developed by Margaret O Dayhoff in the year 1964. Four years later in 1988, PIR along with NBRF, Munich Information Centre for Protein Sequence (MIPS) and the Japan International Protein Information Database, developed an organization referred as **PIR – international with four main aims.**

- To create an organized, non redundant, comprehensive protein database to study structural, functional and evolutionary relationships.
- To generate information on biological origin of protein sequences
- To make database easily accessible in public domain
- To enable cross reference with other databases for presenting structural information of biomolecules.

TrEMBL

TrEMBL stands for automatic Translations of European Molecular Biology Laboratory nucleotide sequences. It is a protein sequence database consisting of unreviewed computer annotated translations of new DNA sequence in the nucleotide sequence databases. Swiss-Prot only includes entries validated by expert curators.

This database was created in 1996 as a computer-annotated supplementary database to Swiss-Prot. With the invent of high throughput sequencing techniques, there is an immense flow of new sequence data from the genome projects and Swiss-Prot is falling behind to provide quick database annotation. To address this problem, a Swiss-Prot buffer called TrEMBL was created. It allows very rapid access to sequence data from the genome projects, without having to compromise the quality of Swiss-Prot.

TrEMBL sequences are produced at the EBI from GenBank entries and annotated mostly computationally using sequence homology as a main principle. It also contains protein sequences selected from the literature and protein entries submitted directly by the researchers. TrEMBL unreviewed entries are kept separated from the Swiss-Prot manually annotated entries so as to maintain the high quality data of later.

The Key features of TEMBL are:

Automatic annotation: It is performed by transferring data from well-labeled entries of Swiss-Prot to unannotated entries in TrEMBL. This process raises the standard of annotation in TrEMBL next to the level of Swiss-Prot, thus improving the quality of data.

Redundancy removal: Full length sequence belonging to same organism and showing 100% identify are fused into a single entry to curtail redundancy.

Evidence attribution: Since TrEMBL contains data from a variety of sources, evidence attribution helps in identifying the source of individual data items. It allows automatic update of data if the underlying data source changes.

It has been dissected into two parts: SP-TrEMBL and REM-TrEMBL

SP-TrEMBL (Swiss-Prot TrEMBL) is a collection of sequences that will be finally upgraded to Swiss-Prot after their manual annotation is finished.

REM-TrEMBL (Remaining TrEMBL) stores those sequences that will never be incorporated in Swiss-Prot. E.g immunoglobulins and T-Cell receptors, fragments of fewer than 8 amino acids, synthetic sequences, patented sequences and coding sequences that do not code real proteins.

PDB

Protein Data Bank

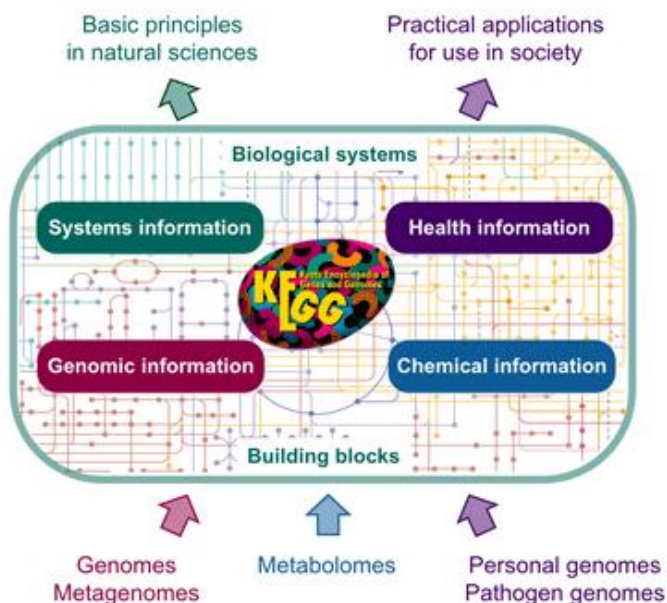
- Structural data from the PDB can be freely accessed at
- It is very large global repository for processing and distribution of 3D macromolecular structure data such as protein, nucleic acids
- Depositors to PDB have derived the structures using variety of tools and techniques like X-ray crystal structure determination, NMR, cryoelectron microscopy and theoretical modeling.
- The database provides access at no charge on internet to structural data as well as methods to visualize the structure and to download structural information.
- It is a primary data and databases derived from PDB are called secondary databases like SCOP and CATH.
- The PDB is overseen by an organization called world wide protein data bank (wwPDB): a consortium whose partners comprise: the research collaborator for structural bioinformatics, the macromolecular structure database at the European bioinformatics, the protein data bank Japan at Osaka university and more recently the BioMagResBank at the University of Wisconsin-Madison.
- In 1971 Walter Hamilton of BNL (Brookhaven National Laboratory) agreed to setup the data bank at Brookhaven and then he died in 1983.
- Then Tom Koeztle took over direction of PDB
- Then in 1998 PDB was transferred to RCSB (Research Collaboratory Structural Bioinformatics)
- Then in 2003, with formation of wwPDB, the PDB became an international organization

- Most structures are determined by X-ray diffraction and about 15% of structure by NMR and few by Cryo-electron microscopy. In the past, number of structures in PDB has grown nearly exponentially.
- The file format initially used by PDB was called PDB file format.
- Around 1996, mmCIF (Macro Molecular Crystallographic Information File) started to phased in.
- Then in 2005 XML version of above format called PDBML was described.
- The structure file can be downloaded in any of these three formats.
- Each structure published in PDB receives a four character alpha numeric identifier, its PDB ID
- The Structure files may be viewed using one of the several open source computer programs. Some other free but not open source programs include VMD, MDLChime, Swiss PDB Viewer, started Biochem and Sirius.
- PDB Wiki is a website for community annotation of PDB structures.

Metabolic Pathway databases




KEGG

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information. It is a computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information). It also contains disease and drug information (health information) as perturbations to the biological system.



The KEGG database has been in development by Kanehisa Laboratories since 1995, and is now a prominent reference knowledge base for integration and interpretation of large-scale molecular data sets generated by genome sequencing and other high-throughput experimental technologies.

KEGG is an integrated database resource consisting of eighteen databases (including computationally generated SSDB) shown below. They are broadly categorized into systems information, genomic information, chemical information and health information, which are distinguished by color coding of web pages.

Category	Database	Content	Color
Systems information	KEGG PATHWAY	KEGG pathway maps	
	KEGG BRITE	BRITE hierarchies and tables	
	KEGG MODULE	KEGG modules	
Genomic information	KEGG ORTHOLOGY (KO)	Functional orthologs	 
	KEGG GENOME	KEGG organisms (complete genomes)	
	KEGG GENES	Genes and proteins	

KARPAGAM ACADEMY OF HIGHER EDUCATION



CLASS: II BSC BC

COURSE NAME: BIOINFORMATICS

COURSE CODE: 16BCU404A

UNIT: II BIOLOGICAL DATABASES

BATCH-2016-2019

Chemical information	KEGG SSDB	GENES sequence similarity	
	KEGG COMPOUND	Small molecules	
	KEGG GLYCAN	Glycans	
	KEGG REACTION	Biochemical reactions	
	KEGG RCLASS	Reaction class	
	KEGG ENZYME	Enzyme nomenclature	
Health information	KEGG NETWORK	Disease-related network elements	
	KEGG VARIANT	Human gene variants	
	KEGG DISEASE	Human diseases	
	KEGG DRUG	Drugs	
	KEGG DGROUP	Drug groups	
	KEGG ENVIRON	Health-related substances	

Chemical information category is collectively called KEGG LIGAND
Health information category integrated with drug labels is called KEGG MEDICUS

These database contain various data objects for computer representation of the biological systems. Thus, the database entry of each database is called the KEGG object, which is identified by the KEGG object identifier consisting of a database-dependent prefix and a five-digit number

Release	Database	Object Identifier	Remark
1995	KEGG PATHWAY	map number	
	KEGG GENES	locus_tag / GeneID	
	KEGG ENZYME	EC number	
	KEGG COMPOUND	C number	
1998	KEGG REACTION	R number	
2000	KEGG GENOME	organism code / T number	

KARPAGAM ACADEMY OF HIGHER EDUCATION

CLASS: II BSC BC

COURSE NAME: BIOINFORMATICS

COURSE CODE: 16BCU404A

UNIT: II BIOLOGICAL DATABASES

BATCH-2016-2019

2002	KEGG ORTHOLOGY	K number	Ortholog IDs in 2000
2003	KEGG GLYCAN	G number	
2004	KEGG RPAIR	RP number	Discontinued in 2016
2005	KEGG BRITE	br number	
	KEGG DRUG	D number	
2006	KEGG MODULE	M number	
2008	KEGG DISEASE	H number	
2010	KEGG ENVIRON	E number	First called EDRUG
	KEGG RCLASS	RC number	
2014	KEGG DGROUP	DG number	
2017	KEGG NETWORK	N number	
	KEGG VARIANT	GeneID+variant number	

KEGG Molecular Networks

The most unique data object in KEGG is the molecular networks -- molecular interaction, reaction and relation networks representing systemic functions of the cell and the organism. Experimental knowledge on such systemic functions is captured from literature and organized in the following forms:

- Pathway map - in KEGG PATHWAY
- Brite hierarchy and table - in KEGG BRITE
- Membership (logical expression) - in KEGG MODULE
- Membership (simple list) - in KEGG DISEASE

genomes and high-throughput molecular datasets through the process of KEGG mapping

In 1995 the concept of mapping was first introduced in KEGG for linking genomes to metabolic pathways (metabolic reconstruction) using the EC number. Once the EC numbers were assigned to enzyme genes in the genome, organism-specific pathways could be generated automatically by matching against the enzyme (EC number) networks of the

KEGG reference metabolic pathways. The EC number is no longer used as an identifier in KEGG. The KEGG Orthology (KO) system is the basis for genome annotation and KEGG mapping.

Period	Identifier	Reference knowledge	Assignment
1995-1999	EC number	Metabolic pathways	Domain based
2000-2002	Ortholog ID	Metabolic and regulatory pathways	Domain based
2003-	KO	Pathways and BRITE hierarchies	Gene based

From a different perspective, individual instances of genes are grouped into KO entries representing functional orthologs in the molecular networks. There are two more types of such generalization in KEGG as shown below.

Network type	Class	Instance
All types	KO (gene ortholog)	Genes in KEGG GENES
Biochemical reaction	RC (reaction class)	Reactions in KEGG REACTION
Drug interaction	DG (drug group)	Drugs in KEGG DRUG

EcoCyc and MetaCyc

The EcoCyc and MetaCyc databases (DBs) are online reference sources for metabolic data. They are similar in describing metabolic pathways, reactions, enzymes and substrate compounds. Both DBs use the same DB schema. Both are accessed using the same software environment, called the Pathway Tools. Both are review-level DBs in that a given entry in either DB often integrates information from multiple literature citations. There is also overlap in the content of the DBs—both contain all pathways of *Escherichia coli* small-molecule metabolism.

The DBs do differ significantly in content. EcoCyc aims to describe the full biochemical network of *E.coli*. As well as describing the metabolism of *E.coli*, EcoCyc describes its signal-transduction pathways, its transporters, and all *E.coli* genes. MetaCyc does not focus on a single organism as EcoCyc does. It describes pathways from a wide

variety of species, with a focus on microorganisms, but includes some human and other mammalian pathways as well.

Intended uses of the DBs include the following.

MetaCyc is a general reference source on metabolic pathways for the scientific community. It also serves as a reference pathway DB for prediction of the pathway complement of an organism from the annotated genome of the organism.

EcoCyc is a resource for analysis of microbial genomes at the level of individual genes. Because the *E.coli* genome has a high fraction of genes whose functions were determined experimentally, it is an accurate reference for inferring gene function by sequence similarity.

Because of its links to sequence DBs such as SWISS-PROT, EcoCyc can be used to perform function-based retrieval of DNA or protein sequences, for example to prepare datasets for studies of protein structure–function relationships.

Both EcoCyc and MetaCyc are used as an aid in teaching biochemistry.

Small Molecule databases

ZINC

ZINC is now available for download (<http://zinc.docking.org>). It is currently built from the catalogs of ten major compound vendors and presently contains 727 842 purchasable compounds. The number of molecules in ZINC is growing, and the numbers reported here should be considered a representative snapshot; see the Web-page for up-to-date statistics. Of these 727842, 494 915 are Lipinski compliant, with the caveat that we have used Molinspiration's LogP as a surrogate for cLogP. Of these, 202 134 are "lead-like" molecules, which we define here as having molecular weight between 150 and 350, calculated LogP less than four, number of hydrogen-bond donors less than or equal to three, and number of hydrogen-bond acceptors less than or equal to six. A total of 34 224 molecules are "fragment-like", with calculated LogP values between – 2 and 3, less than three hydrogen-bond donors, less than six hydrogen-bond acceptors, less than three rotatable bonds, and molecular weight less than 250.

Database Subsets

Many users may only be interested in some of the molecules in ZINC. The ZINC Web pages allow the download of subsets by vendor and other criteria such as Lipinski-compliant, "lead-like", and "fragment-like" compounds. The search page may be used to download small subsets immediately or to create user-defined subsets using arbitrary criteria, including functional groups and molecular properties. Once prepared, each subset is available in SMILES, mol2, SDF, and DOCK flexibase format. Large files are broken into slices of approximately 20 to 100MB for easier download. In the limit, the entire ZINC database may be downloaded.

Process your Own Molecules

Users may upload their own molecules to the ZINC server in SMILES, SDF, or mol2 formats and have them processed using the same protocol we use to build ZINC. The uploaded molecules subsequently appear as a subset for download in the usual way and disappear from the server after a week.

Pubchem

PubChem (<https://pubchem.ncbi.nlm.nih.gov>) is a public repository for information on chemical substances and their biological activities, launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH). For the past 11 years, PubChem has grown to a sizable system, serving as a chemical information resource for the scientific research community. PubChem consists of three inter-linked databases, Substance, Compound and BioAssay. The Substance database contains chemical information deposited by individual data contributors to PubChem, and the Compound database stores unique chemical structures extracted from the Substance database. Biological activity data of chemical substances tested in assay experiments are contained in the BioAssay database. This paper provides an overview of the PubChem Substance and Compound databases, including data sources and contents, data organization, data submission using PubChem Upload, chemical structure standardization, web-based interfaces for textual and non-textual searches, and programmatic access. It also gives a brief description of PubChem3D, a resource derived from theoretical three-

dimensional structures of compounds in PubChem, as well as PubChemRDF, Resource Description Framework (RDF)-formatted PubChem data for data sharing, analysis and integration with information contained in other databases.

Drug Bank

The Drug Bank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.

The latest release of Drug Bank (version 5.0.11, released 2017-12-20) contains 10,970 drug entries including 2,390 approved small molecule drugs, 934 approved biotech (protein/peptide) drugs, 109 nutraceuticals and over 5,089 experimental drugs. Additionally, 4,899 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

CSD

The Cambridge Structural Database (CSD) is a highly curated and comprehensive resource.

Established in 1965, the CSD is the world's repository for small-molecule organic and metal-organic crystal structures. Containing over 900,000 entries from x-ray and neutron diffraction analyses, this unique database of accurate 3D structures has become an essential resource to scientists around the world.

With comprehensive and fully retrospective coverage of the published literature you can have full confidence that your CSD searches are returning all crystal structure matches. The CSD also contains data published directly through the CSD as CSD Communications that are not available anywhere else.

Each crystal structure undergoes extensive validation and cross-checking by expert chemists and crystallographers to ensure that the CSD is maintained to the highest possible standards. Every entry is enriched with bibliographic, chemical and physical property information, adding further value to the raw structural data. These editorial processes are vital for enabling scientists to interpret structures in a chemically meaningful way.

The CSD is continually updated with new structures (>50,000 new structures each year) and with improvements to existing entries. With regular web-updates and early online access to newly published structures you can keep fully informed of the latest research.

Structure Viewer

Visualization of Molecules

- Molecules are used two-dimensional (2D) structure and 3D structure.
- Mostly the molecules with interacting three dimensions.
- No. of tools are available for eg. Rotate, flip and otherwise manipulate virtual molecular models of chemicals and macromolecules.
- Small molecules in 3-D download and install on your computer.
- Scientific websites are Jmol, Java-based viewer for rendering molecules in 3-D.
- Chime is the program used most for viewing small molecules from websites.
- Macromolecules in 3D download and install.

CHIME

- Chime is a free downloadable
- Its chemical structure visualization Plug-in windows and Macintosh.
- It allows view chemical structure from within popular web browsers Java applets and Java applications.
- Chime already be installed for their graphics to work properly.
- Rasmol, Chime shows molecules within a webpage.

Cn3D

- Cn3D is a visualization tool for macromolecules.
- To view 3-dimensional structure from NCBI's Entrez → it's a retrieval service.
- Cn3D is able to correlate structure and sequence information.
 - **Example:** find the residues in a crystal structure that correspond to known
 - disease mutations.
 - = powerful annotations and editing features.
- = right click on the molecule to see the viewing options.

RASMOL

- It's free program
- Developed by Roger A Sayle (1993) University of Edinburgh's, Biocomputing Research unit and the Biomolecular structure Department at Glaxo Research and Development Green Ford, UK.
- Rasmol derived from **Raster** (the array of pixels on a computer screen) molecules.
- Molecular graphics program for visualization of proteins, nucleic acids and small molecules.
- Powerful program aimed at display, teaching and generation of publication quality image.
- Rasmol reads in molecular co-ordinate files in formats like Brookhaven Protein Databanks (PDB).
- Different parts of the molecules displayed and colored independently rest of the molecule or show in different representations simultaneously.
- Molecule may be shown Wire frame, cylinder (deriding), stick bonds, alpha-carbon trace, space filling (CPK) spheres, macromolecular ribbons (either smooth shaded solid ribbons or parallel strands) hydrogen bonding and dot surface.
- Molecule displayed may be rotated, translated, zoomed, z-clipped (slabbed) interactively using either the mouse, the scroll bars, command line or an attached dials box.
- Model can be rotated about the x, y and z axes interactively so that all parts of the molecule can be studied.
- Smaller molecule or layer and in addition it is possible to expand the viewing window up to the full size of the screen.
- Larger picture more elaborate the model, the longer it takes the computer to calculate the appearance of the drawing.

Color schemes are available

CPK

Carbon atoms → Pale grey Oxygen → red
Nitrogen → blue Sulphur → yellow

Group colouring

- Chain colored with color of the rainbow.
- Blue – N-terminus
- Red – C-terminus
- Useful for following the fold from one end of the chain to the other.
- Shapely and amino colours
- Backbone – Pale gray side chain atoms are all given a colour depends upon the size and the polarity of the side chain.
- Oxygen containing side chain (acids, amides and the hydroxy-amino acids Ser and Thr.
- Various shade of red and basic side chain Blue (Arg, Lys, His)
- Hydrophobic amino acids are mostly Grey; Ile is dark green and Val a Pale Magenta.
- Sulphur containing amino acids (Cys, Met) have muddy yellow colors.
- Trp is yellow and Grey, White

Residue colour list

A (ala) – Pale green	L (leu) - grey
C (cys) – Sandy yellow	M (met) – pale brown
D (Asp) – dark magenta	N (asn) - salmon
E (glu) – red	P (pro) - grey
F (phe) – grey	Q (gln) – flash pink
G (gly) – White	R (arg) – navy blue
H (his) – slate blue	S (ser) - tomato
I (ile) – dark green	T (thr) – orange red
K (lys) – royal blue	V (val) – pale magenta
W (trp) – yellow	Y (tyr) – clay grey

- Rasmol prepared by list of commands from a script file

- Rasmol works well for both small molecules and for large ones such as proteins, DNA, RNA.

Protein Explorer a Rasmol derivatives:

- Protein explorer (PE) enables to explore the 3D structure of any macromolecule.
- Proteins, DNA, RAN, carbohydrates and complexes such as between transcriptional regulatory explorers.
- It is not compatible with Internet Explorer
- Firebox free and is recommended for protein explorer.

Biomodel – 3:

- Developed by Angel Herraiez, Lecturer in Biochemistry and molecular Biology at the University of Alcala de Henares (Spain)
- Version V3
- Use J_{MOL}, Java applet to show manipulates the molecular models.

3D – Chemical libraries

- Use chime plug-in

3-D Virtual Chemistry Library

- Molecular database has about 150 molecules divided into six main groups.
- Simple molecule
- Polymers
- Senses
- Medical
- Horrible molecule and
- Interesting molecules
- I addition to structure it also has physical data, history and reactivity of the molecules.

3D Macromolecular structures

Using Cn3D

Entrez molecular modeling Databases

- It contains 3-D macromolecular structure
- Including proteins and polynucleotide
- MMDB contain over 40,000 structures and linked to the rest of the NCBI database
- Including sequences bibliographic citations, taxonomic classifications and sequence and structure neighbors.

Possible Questions

1. Define database.
2. Explain in brief about protein database
3. Explain about the types of file formats with example
4. What information can be obtained from metabolic pathway databases.
5. Give a short account on sequence retrieval system with examples.
6. Write short notes on Nucleic acid databases.
7. List out any five protein structure viewers.
8. What are small molecule databases? Mention about PubChem, DrugBank, ZINC & CSD.
9. Give a detailed account on structure viewers.
10. What does secondary database means.
11. Discuss primary and secondary databases.
12. Write a brief note on metabolic databases.
13. Write about composite database.
14. Explain NCBI and its uses.
15. Explain in detail about structure database PDB.

Karpagam Academy of Higher Education
Department of Biochemistry
II B.Sc., Biochemistry
17BCU404A- Bioinformatics

Question number	Unit	Question	Option I	Option II	Option III	Option IV	Answer
UNIT- II							
1	2	GDH1 began in the year	1960	1975	1986	2003	1986
2	2	Protein is	Primary Database	Secondary Database	Tertiary Database	All the above	Secondary Database
3	2	The NCBI is established at the national cancer institute	1986	1988	1974	1994	1988
4	2	The links for NCBI is	www.ncbi.nih.gov	www.ncbi.nih.gov	www.ncbi.com	www.ncbi.nih.gov	www.ncbi.nih.gov
5	2	The links for EMBL is	WWW_Embi.com	WWW.emb.uk	WWW.emb.ac.uk/embd	http://WWW.emb.ac.uk	WWW.emb.ac.uk/embd
6	2	Protein database	PIR	ExPASy	both a and b	GDH1	both a and b
7	2	The links for GDH1 is	WWW.gdh1.org.uk	www.gdh1.com	www.gdh1.ac.uk	www.gdh1.org	WWW.gdh1.org.uk
8	2	The link for SWISS-PROT is	www.expasy.org	www.expasy.org/ncbi	www.expasy.ac.uk	www.expasy.com	www.expasy.org/ncbi
9	2	Swiss prot is a sequence database	Nucleotide	Protein	both a and b	None of the above	Protein
10	2	is a retrieved DNA sequence database	NBRF	Gen bank	S12	None of the above	Gen bank
11	2	Translation of all codon sequences in EMBL is	Swissprot	tdh	TRIMBL	All of the above	TRIMBL
12	2	usually reflect some vital biological role	Protein	Motif	none of the above	EMBL	Motif
13	2	NCBI stands for	national centre for biotechnology information	national centre for bio information	national centre for biotechnology industry	national centre for bioinformatics info.	national centre for biotechnology information
14	2	Gen database is otherwise known as	Nucleotide database	Pattern Database	Structural database	Pattern Database	Pattern Database
15	2	OMIM stand for	Online Mendelian inheritance in man	Origin mutation In man	None	Both a and b	Online Mendelian inheritance in man
16	2	Transbl is used to convert	Protein sequence to nucleotide sequence	Translate Nucleotide Sequence into amino acids	Translate nucleotide to protein	both a and b	Translate Nucleotide Sequence into amino acids
17	2	The database stores 3d atomic coordinates of proteins and nucleic acid and the data is obtained by experimental methods like	Chromatography	colorimetric	X-ray crystallography	Electrophoresis	X-ray crystallography
18	2	Which electrophoresis used in nucleotide databases	1D gel	2D gel	3D gel	SDS	2D gel
19	2	Protein sequence database was developed at	NBRF	NCBI	EMBL	NCBS	NBRF
20	2	The number of chromosomes in Drosophila Melanogaster	11 pairs	4 pairs	4 pairs	6 pairs	13 pairs
21	2	The Drosophila melanogaster genome is sequenced using approach	Shot-gun	High resolution and physical mapping	Parallel	Vertical	Shot-gun
22	2	Saccharomyces cerevisiae, is also known as	Euge yeast	Baker's yeast	House mouse	BAC	Baker's yeast
23	2	Arabidopsis thaliana is a member of the family	Brassicaceae	Nematoide	Proteobacteria	Human simians	Brassicaceae
24	2	The number of chromosomes of Arabidopsis thaliana is	5	10	15	20	5
25	2	Arabidopsis thaliana contains bases	100 millions	150 millions	140 million	All the above	150 millions
26	2	Chromosomes information can cause the disease	Cancer	Malaria	Diabetes	All the above	Malaria
27	2	Protein sequence database was developed at	NBRF	NCBI	EMBL	NCBS	NBRF
28	2	Expansion of PIR	Protein Information Research	Protein Information Resource	Protein Integral research	Protein Information Results	Protein Information Resource
29	2	GD is the major repository of structures	DNA	RNA	Protein	All of the above	All of the above
30	2	The first secondary database have been developed was	PROSITE	PDB	PIR	GENBANK	PDB
31	2	With in PROSITE Motifs are encoded as	Regular expression	Patterns	motif	PIR	Patterns
32	2	The database stores 3d atomic coordinates of proteins and nucleic acid and the data is obtained by experimental methods like...	Chromatography	colorimeter	X-ray crystallography	electrophoresis	Chromatography
33	2	Protein can be classified using	Ball and stick	space fill	both a and b	normal	both a and b
34	2	SWISS-prot is sequence database	nucleotide	proteins	OMIM	OMIA	proteins
35	2	The Nucleic acid sequence databases are collections of	data	queries	entries	indices	data
36	2	Fields are used to create for relational databases	Entries	Queries	Indices	Custom	Queries
37	2	The first nucleic acid sequence was announced in the year	1964	1966	1956	1966	1964
38	2	The first secondary database have been developed was	PROSITE	PDB	PIR GENBANK	MAST	PROSITE
39	2	With in PROSITE Motifs are encoded as	Regular expression	Patterns	motif	All of the above	Patterns
40	2	Entries are deposited in	GENBANK	PROSITE	PDB	SWISS-PROT	PDB
41	2	Comments are list of	Accession of the numbers	Swiss Prot Identification	Codes of the true matches	Any possible matches which are often fragments	Codes of the true matches
42	2	The documentation files concludes with appropriate	Geographic references	Bibliographic References	Structural References	Statistical Refer	Bibliographic References
43	2	is beta is denatured by certain amino acid like	iodine	azurine	iodine	iodine	iodine
44	2	is beta is stabilized by	hydrogen bonds	disulfide bonds	Salt bond	electrostatic bonds	hydrogen bonds
45	2	Gene that have arisen from a common ancestor is called	homologous	orthologous	analogous	synonymous	homologous
46	2	is a short conserved pattern of aminoacids	motif	conting	discontinuous	blocks	motif
47	2	are collection of overlapping sequence that are obtained in a sequencing project	contig	contig	contig	contig	contig
48	2	Protein sequence database was developed at	NBRF	NCBI	EMBL	NCBS	NBRF
49	2	Expansion of PIR	Protein Information Research	Protein Information Resources	Protein Integral research	Protein Information Results	Protein Information Resources
50	2	GD is the major repository of structures	DNA	RNA	Protein	All of the above	Protein
51	2	The documentation files concludes with appropriate	Geographic references	Bibliographic References	Structural References	Statistical References	Bibliographic References
52	2	Motif finding, also known as profile analysis, constructs global MSAs that attempt to align short conserved among the sequences in the query set	DALI	msa	sequence motifs	Folded	sequence motifs
53	2	Derivation of identical cells or molecules derived from a single ancestor	clonal	sequence	clone	vector	clone
54	2	An organism's basic complement of DNA is called	Genome	genome	gene	genome	genome
55	2	Chromosome literally means	Coloured	colored body	colored protein	all of the above	colored body
56	2	The contains experimentally determined 3D protein structures	PIR	PIR	SWISSPROT	MMDB	MMDB
57	2	Local alignment was performed by	Deveron and Lipman	Smith and Waterman	Waterman	Cluck	Smith and Waterman
58	2	A protein sequence database formatted nucleotide sequences	TRIMBL	GENPASY	TRIPD	TRIPD	TRIMBL
59	2	The sequence of more than one identical item repetition	Aluminum	conserved region	nonconserved region	redundancy	redundancy
60	2	In the sequence database, each sequence is an	Files	Entry	Query	All the above	All the above

UNIT-III

SYLLABUS

Sequence alignment: Similarity, identity and homology, Alignment-local and global alignment, pairwise and multiple sequence alignments, alignment algorithms, amino acid substitution matrices (PAM and BLOSUM), BLAST and CLUSTALW

INTRODUCTION TO SEQUENCE ALIGNMENT

Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.

```

AAB24882      TYHMCQFHCRIYVNNHSGEKLIECNERSKAFSCPSHLQCHKRRQIGETHEHNQCGKAFPT 60
AAB24881      -----YECNQCGKAFAQHSSLKCHYRTHIGKPYECNQCGKAFSK 40
               ****: .***: * *:*** * :****.:* *****.,

AAB24882      PSHLQYHERHTHTGKPYECHQCGQAFKKCSLLQRHKRTHHTGKPYE-CNQCGKAFAQ- 116
AAB24881      HSHLQCHKRTHHTGKPYECNQCGKAFSQHGLLQRHKRTHHTGKPYMNVINMVKPLHNS 98
               **** *:*****:****:***.: *****:*****: *.:
    
```

A sequence alignment, produced by ClustalW, of two human zinc finger proteins, identified on the left by GenBank accession number.

Single letters: amino acids.

Red: small, hydrophobic, aromatic, not Y.

Blue: acidic. Magenta: basic.

Green: hydroxyl, amine, amide, basic.

Gray: others. "*": identical. "(": conserved substitutions (same colour group). ".": semi-conserved substitution (similar shapes).

Definition of sequence alignment

- Sequence alignment is the procedure of comparing two (pair-wise alignment) or more multiple sequences by searching for a series of individual characters or patterns that are in the same order in the sequences.
- There are two types of alignment: local and global. In global alignment, an attempt is made to align the entire sequence. If two sequences have approximately the same length and are quite similar, they are suitable for the global alignment.
- Local alignment concentrates on finding stretches of sequences with high level of matches.

L G P S S K Q T G K G S - S R I W D N

Global alignment

L N - I T K S A G K G A I M R L G D A

----- T G K G -----

Local alignment

----- A G K G -----

Interpretation of sequence alignment

- Sequence alignment is useful for discovering structural, functional and evolutionary information.
- Sequences that are very much alike may have similar secondary and 3D structure, similar function and likely a common ancestral sequence. It is extremely unlikely that such sequences obtained similarity by chance. For DNA molecules with n nucleotides such probability is very low $P = 4^{-n}$. For proteins the probability even much lower $P = 20^{-n}$, where n is a number of amino acid residues
- Large scale genome studies revealed existence of horizontal transfer of genes and other sequences between species, which may cause similarity between some sequences in very distant species

Methods of Sequence Alignment

- Dot Matrix analysis
- Dynamic Programming (DP) algorithm
- Word (or) K-tuple methods

Dot matrix analysis

- Comparing for two sequence
- One sequence (A) top of the matrix
- Other one sequence (B) down left side
- Any region of similarity is revealed by a diagonal row of dots.
- Five clear diagonals
- Diagonals are obtained by aligning genomic and cDNA.
- Five diagonals represent the five exons of the gene which was confirmed from the annotated entry of the gene.

Sequence alignment program is to align the two sequences

- To produce highest score a scoring matrix is used to add points to the score for each match and subtract them for each mismatch.
- Matrixes are used for nucleic acid alignment to involve fairly simple match/mismatch scoring schemes.

Parameters used for sequence alignment

1. scoring matrix
2. Substitution matrices
3. Gap penalty

Scoring matrices

- It is critical to have reasonable scoring schemes accepted by the scientific community for DNA and proteins and for different types of alignments
- The wealth of information accumulated in the gene/protein banks was utilised with dynamic programming procedure to create such matrices for scoring matches and separately penalties for gaps introduction and extensions

- Matrices for DNA are rather similar as there are only two options purine & pyrimidine and match & mismatch
- Proteins are much more complex and the number of option is significant
- PAM and other matrices are represented in log odds scores, which is the ratio of chance of amino acid substitution due to essential biological reason to the chance of random substitution
- There are many different PAMs, which are representing different evolutionary scenarios.
- PAM 250 represents a level of 250% of changes expected in 2500 MY
- PAM is more suitable for studying quite distant proteins, BLOSUM is for more conserved proteins of domains

Scoring matrices: PAM (Percent Accepted Mutation)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W

- Amino acids are grouped according to the chemistry of the side group: (C) sulfhydryl, (STPAG)-small hydrophilic, (NDEQ) acid, acid amide and hydrophilic, (HRK) basic, (MILV) small hydrophobic, and (FYW) aromatic.
- Log odds values: +10 means that ancestor probability is greater, 0 means that the probabilities are equal, -4 means that the change is random. Thus the probability of alignment YY/YY is $10+10=20$, whereas YY/TP is $-3-5=-8$, a rare and unexpected between homologous sequences.

Scoring matrices: BLOSUM62 (BLOcks amino acid SUBstitution Matrices)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	2	-1	-3	-1	-4	-3	3	3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-1	-3	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	-2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-2	-2	-4	-3	-2	-4	-3	-2	-2	-3	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Ideology of BLOSUM is similar but it is calculated from a very different and much larger set of proteins, which are much more similar and create blocks of proteins with a similar pattern.

Differences between PAM and BLOSUM

1. PAM matrices are based on an explicit evolutionary model (i.e. replacements are counted on the branches of a phylogenetic tree), whereas the BLOSUM matrices are based on an implicit model of evolution.
2. The PAM matrices are based on mutations observed throughout a global alignment, this includes both highly conserved and highly mutable regions. The BLOSUM matrices are based only on highly conserved regions in series of alignments forbidden to contain gaps.
3. The method used to count the replacements is different: unlike the PAM matrix, the BLOSUM procedure uses groups of sequences within which not all mutations are counted the same.
4. Higher numbers in the PAM matrix naming scheme denote larger evolutionary distance, while larger numbers in the BLOSUM matrix naming scheme denote higher sequence similarity and therefore smaller evolutionary distance. Example: PAM150 is used for more distant sequences than PAM100; BLOSUM62 is used for closer sequences than Blosum50.

Substitution matrices

- 210 score possibilities for any possible protein pair.

KARPAGAM ACADEMY OF HIGHER EDUCATION

CLASS: II BSC BC

COURSE NAME: BIOINFORMATICS

COURSE CODE: 16BCU404A

UNIT: III SEQUENCE ALIGNMENT

BATCH-2016-2019

- 20X20 matrix, where the diagonal gives 100% match between the amino acids.
- Main diagonal are of identical 20 amino acid scores and on each side of diagonal 190 scoring that are similar obtain 210 scoring terms for 20 amino acid combinations.
- Pair of amino acid is termed as log – odds values and these have been scaled and rounded to the nearest integer for computational efficiency known as score matrix or substitution matrix.
- The late Margaret Dayhoff pioneer in protein data basing and comparison
- Dayhoff, MDM [Mutation Data Matrix] (or) PAM [Point (or) Percent Accepted Mutation).
- PAM one such major amino acid scoring or substitution matrix
- BLOSUM series of matrices were created by Steve Henikoff and colleagues.
- Matrices are used BLOSUM 80, 62, 40 and 30.
- PAM matrices used are PAM 120, 160, 250 and 350 matrices.

80 – 100 %	Sequence	identity	BLOSUM80
60 – 80 %	Sequence	identity	BLOSUM62
30 – 60 %	Sequence	identity	BLOSUM45
0 – 30 %	Sequence	identity	BLOSUM30
80 – 100 %	Sequence	identity	PAM20
60 – 80 %	Sequence	identity	PAM60
40 – 60 %	Sequence	identity	PAM120
0 – 40 %	Sequence	identity	PAM350

Gap penalty

- Gap is any maximal consecutive run of spaces in a single string of a given alignment.
- Gap helps to create alignments that better conform to underlying biological models and more closely fit patterns that one expects to find in meaningful alignment.
- No. of continuous gaps and not only the number of spaces when calculating an alignment mark.

Example

X = a t t c - - g a - t g g a c c

Y = a - - c g t g a t t - - c c

- Four gaps containing a total of eight spaces
- 7 matches, no mismatch.
- No. of gaps in the alignment will be denoted as # gaps

Pairwise Sequence Alignment:

- The two sequences are homologous, i.e. they have evolved from a common ancestor.
- Differences between them are due to only two kinds of events, namely insertion
- deletions (*indels*) and substitutions (change of single elements of the sequence -
- amino acids if the sequence is a protein and nucleic acid, if the sequence is DNA).

Two types pairwise sequence alignment

- Needleman-Wunsch Algorithm (or) Global Alignment
- Smith-Waterman (or) Local Alignment

Needleman-Wunsch algorithm

- The **Needleman-Wunsch algorithm** performs a global alignment on two sequences (called *A* and *B* here).
- It is commonly used in bioinformatics to align protein or nucleotide sequences.
- The algorithm was published in 1970 by Saul B. Needleman and Christian D. Wunsch.
- The Needleman-Wunsch algorithm is an example of dynamic programming, and was the first application of dynamic programming to biological sequence comparison.

A modern presentation

Scores for aligned characters are specified by a similarity matrix. Here, $S(a,b)$ is the similarity of characters a and b . It uses a linear gap penalty, here called d .

For example, if the similarity matrix were then the alignment.

	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

then the alignment:

```
AGACTAGTTAC
CGA---GACGT
```

with a gap penalty of -5, would have the following score:

To find the alignment with the highest score, a two-dimensional array (or matrix) F is allocated. The entry in row i and column j is denoted here by F_{ij} . There is one column for each character in sequence A , and one row for each character in sequence B . Thus, if we are aligning sequences of sizes n and m , the amount of memory used is in $O(nm)$. (Hirschberg's algorithm can compute an optimal alignment in $\Theta(\min\{n,m\})$ space, roughly doubling the running time.

Dotplots

- The most intuitive representation of the comparison between two sequences uses dot-plots.
- One sequence is represented on each axis and significant matching regions are distributed along diagonals in the matrix.

Exercise: Making a dotplot

unix % **dottup**

DNA sequence dot plot

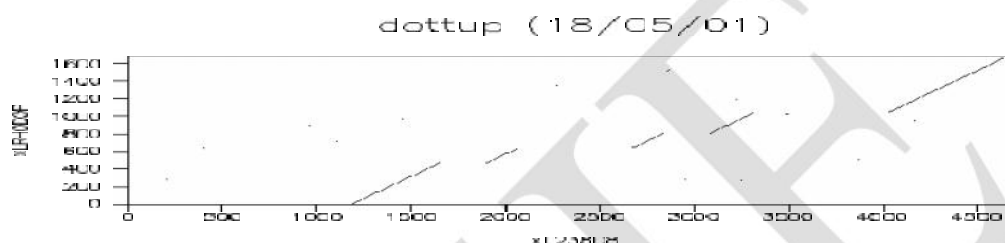
Input sequence: **embl:x123808**

Second sequence: **embl:xlrhodop**

Word size [4]: **10**

Graph type [x11]:

A window will pop up on your screen that should look something like this:



- The diagonal lines represent areas where the two sequences align well. You can see that there are five clear diagonals.
- Aligning genomic and cDNA - these five diagonals represent the five exons of the gene! If you look at the original EMBL entry for the genomic sequence using SRS, you will see that the annotated entry says that there are five exons in this gene. So our results are in agreement.
- The settings we have used for this example are those that give the best results. dottup looks for exact matches between sequences.
- As we expect the exon regions from the genomic sequence to exactly match the cDNA sequence we can use longer word lengths as we should still get exact matches.
- This gives a very clean plot. If you were to match the cDNA sequence against that of a related sequence, e.g. the rhodopsin from mouse (embl: m55171) then you wouldn't expect long exact matches so should use a shorter word length.

Smith-Waterman algorithm

- The Smith-Waterman algorithm is a well-known algorithm for performing local sequence alignment; that is, for determining similar regions between two nucleotide or protein sequences.

- Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure

Algorithm Explanation

A matrix H is built as follows:

if $a_i = b_j$ $w(a_i, b_j) = w(\text{match})$ or if $a_i \neq b_j$ $w(a_i, b_j) = w(\text{mismatch})$

Where:

a, b = Strings over the Alphabet Σ

$m = \text{length}(a)$

$n = \text{length}(b)$

$H(i, j)$ - is the maximum Similarity-Score between a suffix of $a[1...i]$ and a suffix of $b[1...j]$

, '-' is the gap-scoring scheme

Example

Sequence 1 = ACACACTA

Sequence 2 = AGCACACA

$w(\text{gap}) = 0$

$w(\text{match}) = +2$

$w(a, -) = w(-, b) = w(\text{mismatch}) = -1$

To obtain the optimum local alignment, we start with the highest value in the matrix (i, j) . Then, we go backwards to one of positions $(i-1, j)$, $(i, j-1)$, and $(i-1, j-1)$ depending on the direction of movement used to construct the matrix. We keep the process until we reach a matrix cell with zero value, or the value in position $(0, 0)$.

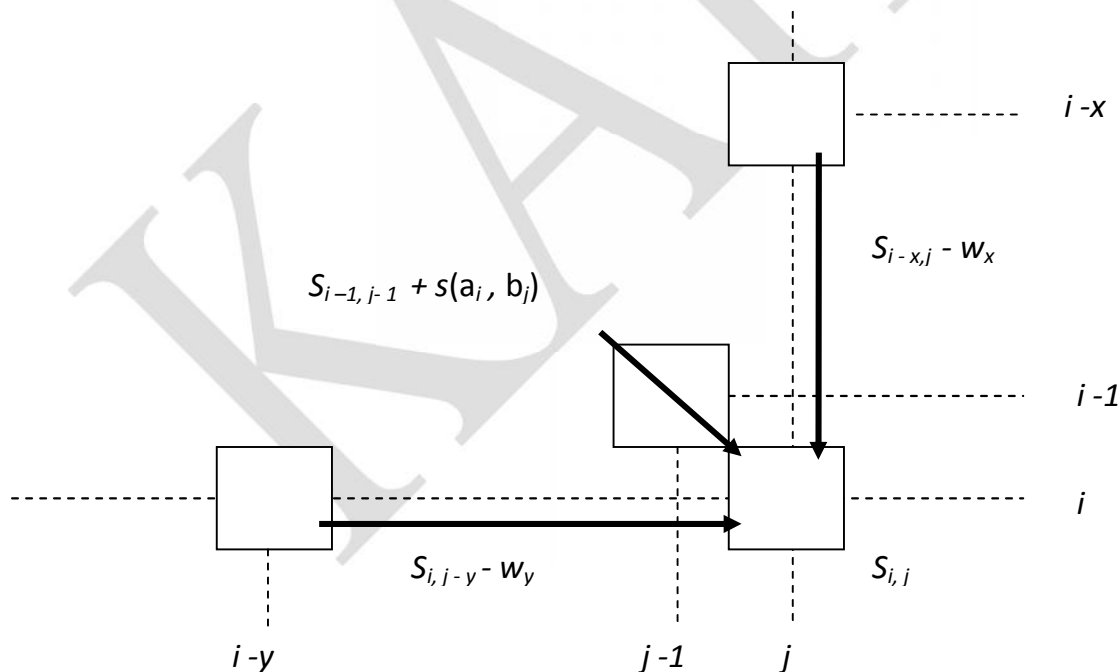
In the example, the highest value corresponds to the cell in position $(8, 8)$. The walk back corresponds to $(8, 8)$, $(7, 7)$, $(7, 6)$, $(6, 5)$, $(5, 4)$, $(4, 3)$, $(3, 2)$, $(2, 1)$, $(1, 1)$, and $(0, 0)$,

Once we've finished, we reconstruct the alignment as follows: Starting with the last value, we reach (i, j) using the previously-calculated path. A diagonal jump implies there is an alignment (either a match or a mismatch). A top-down jump implies there is a deletion. A left-right jump implies there is an insertion.

Description of the dynamic programming algorithm

- Consider building this alignment in steps, starting from the initial match (V/V) and then sequentially adding a new pair until the alignment is complete, at each stage choosing a pair from all the possible matches that provides the highest score for the alignment up to that point.
- If the full alignment has the highest possible (or optimal) score, then the old alignment from which it was derived (A) by addition of the aligned Y/Y pair must also have been optimal up to that point in the alignment.
- In this manner, the alignment can be traced back to the first aligned pair that was also an optimal alignment.
- The example, which we have considered, illustrates 3 choices: 1. Match the next character(s) in the following position(s); 2. Match the next character(s) to a gap in the upper sequence; 3. Add a gap in the lower sequence.

Formal description of dynamic programming algorithm



- This diagram indicates the moves that are possible to reach a certain position (i,j) starting from the previous row and column at position $(i-1, j-1)$ or from any position in the same row or column
- Diagonal move with no gap penalties or move from any other position from column j or row i , with a gap penalty that depends on the size of the gap

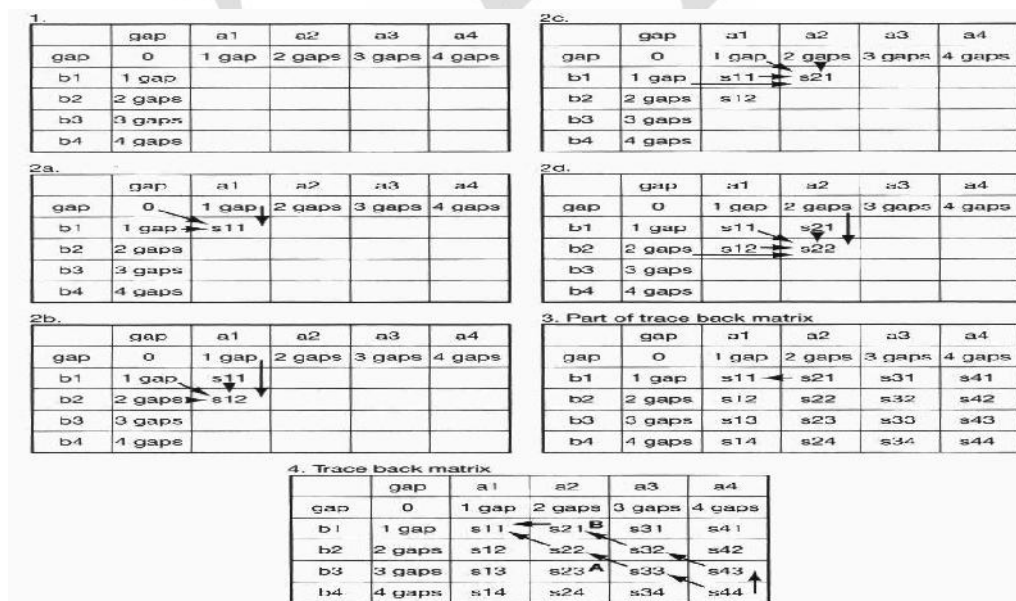
Formal description of dynamic programming algorithm

For two sequences $\mathbf{a} = a_1, a_2, \dots, a_i$ and $\mathbf{b} = b_1, b_2, \dots, b_j$, where $S_{ij} = S(a_1, \dots, a_i, b_1, \dots, b_j)$ then

$$S_{ij} = \max \{ S_{i-1, j-1} + s(a_i b_j), \\ \max (S_{i-x, j} - w_x), \\ x \geq 1$$

where S_{ij} is the score at position i in sequence \mathbf{a} and j in sequence \mathbf{b} , $s(a_i b_j)$ is score for aligning the character at positions i and j , w_x is the penalty for a gap of length x in sequence \mathbf{a} , and w_y is the penalty for a gap of length y in sequence \mathbf{b} .

Note that S_{ij} is a type of running best score as the algorithm moves through every position in the matrix



Alignment A: a1 a2 a3 a4

b1 b2 b3 b4

Alignment B: a1 a2 a3 a4 -

b1 - b2 b3 b4

The highest scoring matrix position is located (in this case s44) and then traced back as far as possible, generating the path shown.

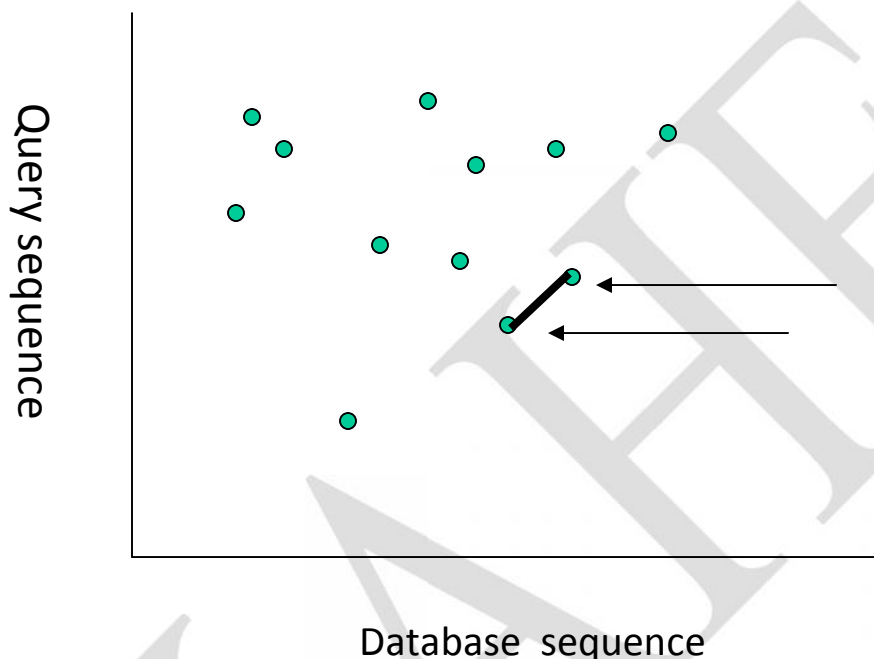
BLAST

- BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) comes under the category of homology and similarity tools.
- It is a set of search programs designed for the Windows platform and is used to perform fast similarity searches regardless of whether the query is for protein or DNA.
- Comparison of nucleotide sequences in a database can be performed. Also a protein database can be searched to find a match against the queried protein sequence.
- NCBI has also introduced the new queuing system to BLAST (Q BLAST) that allows users to retrieve results at their convenience and format their results multiple times with different formatting options.

BLAST procedure

- The steps used by the BLAST algorithm:
- The seq is optionally filtered to remove low-complexity regions (AGAGAG...)
- A list of words of certain length is made
- Using substitution scores matrixes (like PAM or BLOSUM62) the query seq. words are evaluated for matches with any DB seq. and these scores (log) are added
- A cutoff score (T) is selected to reduce number of matches to the most significant ones.
- The above procedure is repeated for each word in the query seq.
- The remaining high-scoring words are organised into efficient search tree and rapidly compared to the DB seq.

- If a good match is found then an alignment is extended from the match area in both directions as far as the score continue to grow. In the latest version of BLAST more time-efficient method is used
- The essence of this method is finding a diagonal connecting ungapped alignments and extending them



- The next step is to determine those high scoring pairs (HSP) of seq., which have score greater than a cutoff score (S). S is determined empirically by examining a range of scores found by comparing random seq. and by choosing a value that is significantly greater.
- Then BLAST determines statistical significance of each HSP score. The probability p of observing a score S equal to or greater than x is given by the equation: $p(S \geq x) = 1 - \exp(-e^{-\lambda(x-u)})$, where $u = [\log(Km'n')]/\lambda$ and K and λ are parameters that are calculated by BLAST for amino acid or nucleotide substitution scoring matrix, n' is effective length of the query seq. and m' is effective length of the database seq.

- On the next step a statistical assessments is made in the case if two or more HSP regions are found and certain matching pairs are put in descending order in the output file as far as their similarity/ score is concerned.

Depending on the type of sequences to compare, there are different programs:

- **blastp** compares an amino acid query sequence against a protein sequence database
- **blastn** compares a nucleotide query sequence against a nucleotide sequence database
- **blastx** compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

Multiple Sequence Alignment

- Often applied to proteins
- Proteins that are similar in sequence are often similar in structure and function.
- Sequence changes more rapidly in evolution than does structure and function

Overview of Methods

- Dynamic programming – too computationally expensive to do a complete search; uses heuristics
- Progressive – starts with pair-wise alignment of most similar sequences; adds to that
- Iterative – make an initial alignment of groups of sequences, adds to these (e.g. genetic algorithms)
- Locally conserved patterns
- Statistical and probabilistic methods

Dynamic Programming

- Computational complexity – even worse than for pair-wise alignment because we're finding all the paths through an n-dimensional hyperspace (We can picture this in 2 or 3 dimensions.)
- Can align about 7 relatively short (200-300) protein sequences in a reasonable amount of time; not much beyond that.
- Let's picture this in 3 dimensions (pp. 146-157 in book). It generalizes to n.
- Consider the pair-wise alignments of each pair of sequences.
- Create a phylogenetic tree from these scores.
- Consider a multiple sequence alignment built from the phylogenetic tree.
- These alignments circumscribe a space in which to search for a good (but not necessarily optimal) alignment of all n sequences.
- Create a phylogenetic tree based on pair-wise alignments (Pairs of sequences that have the best scores are paired first in the tree.)
- Do a "first-cut" msa by incrementally doing pair-wise alignments in the order of "aliveness" of sequences as indicated by the tree. Most alive sequences aligned first.
- Use the pair-wise alignments and the "first-cut" msa to circumscribe a space within which to do a full msa that searches through this solution space.
- The score for a given alignment of all the sequences is the sum of the scores for each pair, where each of the pair-wise scores is multiplied by a weight ϵ indicating how far the pair-wise score differs from the first-cut msa alignment score.
- Does not guarantee an optimal alignment of all the sequences in the group.
- Does get an optimal alignment within the space chosen.

Phylogenetic Tree

- Dynamic programming uses a phylogenetic tree to build a "first-cut" msa

- The tree shows how protein could have evolved from shared origins over evolutionary time.
- See page 143 in *Bioinformatics* by Mount.
- Chapter 6 goes into detail on this.

Progressive Methods

- Similar to dynamic programming method in that it uses the first step (i.e., it creates a phylogenetic tree, aligns the most-alike pair, and incrementally adds sequences to the alignment in order of “aliveness” as indicated by the tree.)
- Differs from dynamic programming method for MSA in that it doesn’t refine the “first-cut” MSA by doing a full search through the reduced search space. (This is the computationally expensive part of DP MSA in that, even though we’ve cut down the search space, it’s still big when we have many sequences to align.)
- Generally proceeds as follows:
 - Choose a starting pair of sequences and align them
 - Align each next sequence to those already aligned, one at a time
- Heuristic method – doesn’t guarantee an optimal alignment

ClustalW

- Based on phylogenetic analysis
- A phylogenetic tree is created using a pairwise distance matrix and nearest-neighbor algorithm
- The most closely-related pairs of sequences are aligned using dynamic programming
- Each of the alignments is analyzed and a profile of it is created
- Alignment profiles are aligned progressively for a total alignment
- W in ClustalW refers to a weighting of scores depending on how far a sequence is from the root on the phylogenetic tree.

Problems with Progressive Method

- Highly sensitive to the choice of initial pair to align. If they aren't very similar, it throws everything off.
- It's not trivial to come up with a suitable scoring matrix or gap penalties.

Iterative Methods for Multiple Sequence Alignment

- Get an alignment.
- Refine it.
- Repeat until one msa doesn't change significantly from the next.
- An example is genetic algorithm approach

Genetic Algorithms

- A general problem solving method modeled on evolutionary change.
- Create a set of candidate solutions to your problem, and cause these solutions to evolve and become more and more fit over repeated generations.
- Use survival of the fittest, mutation, and crossover to guide evolution.

Evolutionary Change in Genetic Algorithms

- survival of the fittest – the best solutions survive and reproduce to the next generation
- mutation – some solutions mutate in random ways (but they must always remain viable solutions)
- crossover – solutions “exchange parts”

Laying Out the Problem

- What would a candidate solution look like in a multiple sequence alignment program? (an MSA of ~20 proteins)
- How many candidate solutions should there be? (~100)

Evolving to a Next Generation

- Which candidate solutions should survive to the next generation?
 - First, take the top half based on best sum of pairs scores
 - Then randomly select second half, giving more chance to an MSA's being selected in proportion to how good its score is.

Possible question

1. What is sequence alignment? Explain in detail.
2. Write brief note on ortholog, paralog and homolog.
3. Explain about PAM and BLOSUM.
4. Explain the application of BLAST and its significance.
5. What is MSA? Illustrate about ClustalW algorithm.
6. Write a note on Block Substitution Matrices.
7. Mention the significance and methodology involved in Multiple Alignment.
8. Differentiate the tblastn and tblastx.
9. Explain the role of molecular evolution in the phylogeny analysis.
10. How can the unrooted tree be constructed from the sequence data? Explain.

Department of Biochemistry
II B.Sc., Biochemistry
17BCU404A- Bioinformatics

Question	Option I	Option II	Option III	Option IV	Answer
1. Expansion of BLAST is	Basic alignment	Basic, local alignment	Biological local alignment	Basic, local alignment with more	Basic, local alignment with more
2. Protein interaction with protein sequence	BLAST	BLAST	BLAST	BLAST	BLAST
3. Translated nucleic acid sequence with a protein	BLAST	BLAST	BLAST	BLAST	BLAST
4. Searches for all sequence with a given	BLAST	BLAST	BLAST	BLAST	BLAST
5. PSI-BLAST	BLAST	BLAST	BLAST	BLAST	BLAST
6. Expansion of the protein sequence and statistical sequence alignment sequence available for all alignments	BLAST	BLAST	BLAST	BLAST	BLAST
7. Database available for all alignments	BLAST	BLAST	BLAST	BLAST	BLAST
8. A sequence against which a protein and a nucleic acid sequence can be aligned	BLAST	BLAST	BLAST	BLAST	BLAST
9. Database available for all alignments	BLAST	BLAST	BLAST	BLAST	BLAST
10. BLAST	BLAST	BLAST	BLAST	BLAST	BLAST
11. BLAST	BLAST	BLAST	BLAST	BLAST	BLAST
12. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
13. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
14. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
15. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
16. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
17. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
18. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
19. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
20. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
21. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
22. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
23. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
24. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
25. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
26. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
27. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
28. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
29. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
30. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
31. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
32. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
33. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
34. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
35. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
36. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
37. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
38. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
39. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST
40. The use of computational data can be used for BLAST as a test sequence	BLAST	BLAST	BLAST	BLAST	BLAST

Crossing over occurs
Continue
C
B
B
A
C
B
C

UNIT-IV

SYLLABUS

Phylogenetic analysis: Construction of phylogenetic tree, dendrograms, methods of construction of phylogenetic trees-maximum parsimony, maximum likelihood and distance methods.

Phylogenetic tree

- A **phylogenetic tree** or **evolutionary tree** is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics.
- The taxa joined together in the tree are implied to have descended from a common ancestor.
- In a **rooted** phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of the descendants and the edge lengths in some trees may be interpreted as time estimates.
- Each node is called a taxonomic unit.
- Internal nodes are generally called hypothetical taxonomic units (HTUs) as they cannot be directly observed.
- Trees are useful in fields of biology such as systematics and comparative phylogenetics.

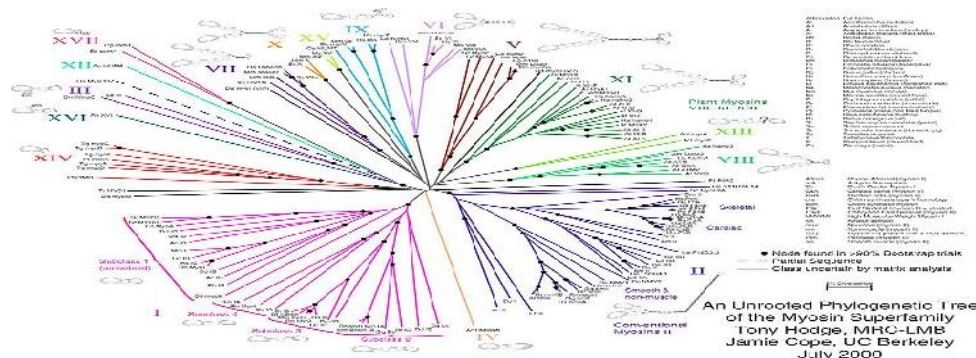
Types

A rooted phylogenetic tree

A rooted tree is used to make inferences about the most common ancestor of the leaves or branches of the tree. Most commonly the root is referred to as an "outgroup"

Unrooted tree:

An unrooted tree is used to make an illustration about the leaves or branches, but not make assumption regarding a common ancestor.



Total rooted trees and total unrooted trees, where n represents the number of leaf nodes. Among labeled bifurcating trees, the number of unrooted trees with n leaves is equal to the number of rooted trees with $n - 1$ leaves.

dendrogram is a broad term for the diagrammatic representation of a phylogenetic tree.

A **cladogram** is a tree formed using cladistic methods. This type of tree only represents a branching pattern, i.e., its branch lengths do not represent time.

A **phylogram** is a phylogenetic tree that explicitly represents number of character changes through its branch lengths.

A **chronogram** is a phylogenetic tree that explicitly represents evolutionary time through its branch lengths.

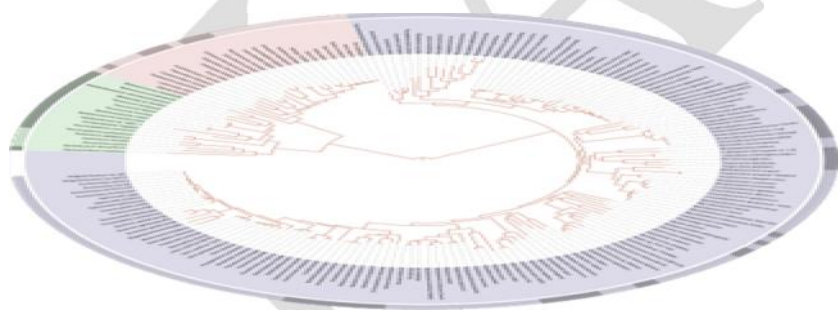
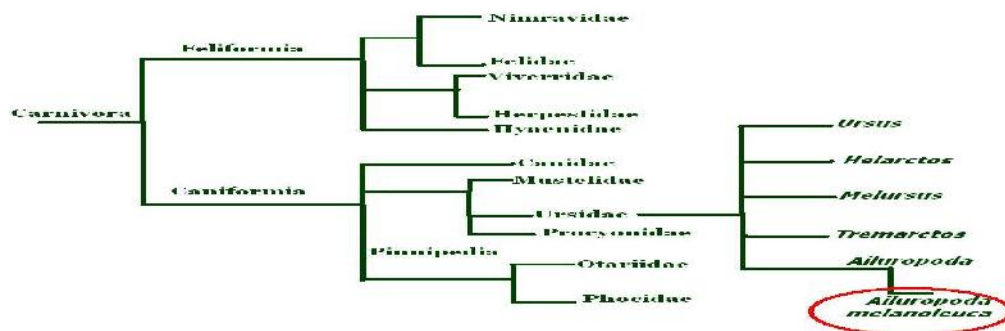
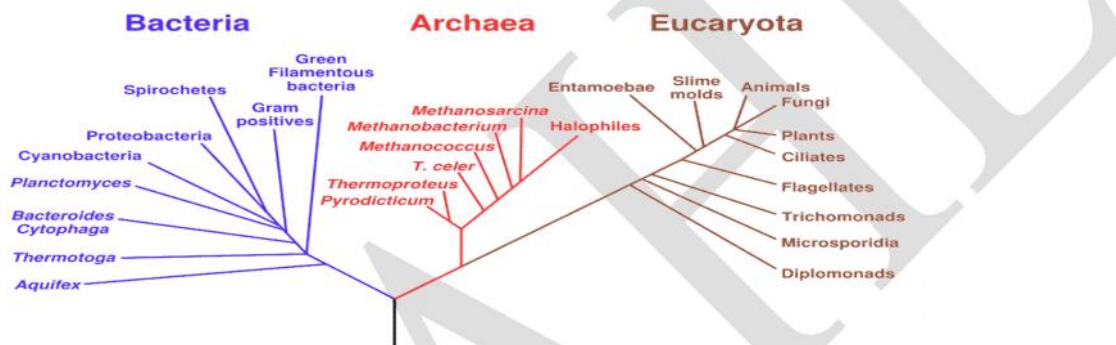


Fig. 2: A highly resolved, automatically generated Tree Of Life, based on completely sequenced genomes.

The agglomerative hierarchical clustering algorithms available in this program module build a cluster hierarchy that is commonly displayed as a tree diagram called a dendrogram. They begin with each object in a separate cluster. At each step, the two clusters that are most similar are joined into a single new cluster. Once fused, objects are never separated. The eight methods that are available represent eight methods of defining the similarity between clusters.



A phylogenetic tree, showing how Eukaryota and Archaea are more closely related to each other than to Bacteria, based on Cavalier-Smith's theory of bacterial evolution.



Tree-building methods can be assessed on the basis of several criteria:

- efficiency
- power
- consistency
- robustness
- falsifiability
- Tree-building techniques have also gained the attention of mathematicians. Trees can also be built using T-theory.

Limitations

- It is important to remember that trees do have limitations. For example, trees are meant to provide insight into a research question and not intended to represent an entire species history.
- Several factors, like gene transfers, may affect the output placed into a tree.

- All knowledge of limitations related to DNA degradation over time must be considered, especially in the case of evolutionary trees aimed at ancient or extinct organisms.

Construction

Phylogenetic trees composed with a nontrivial number of input sequences are constructed using computational phylogenetics methods. Distance-matrix methods such as neighbor-joining or UPGMA, which calculate genetic distance from multiple sequence alignments, are simplest to implement, but do not invoke an evolutionary model. Many sequence alignment methods such as ClustalW also create trees by using the simpler algorithms (i.e. those based on distance) of tree construction. Maximum parsimony is another simple method of estimating phylogenetic trees, but implies an implicit model of evolution (i.e. parsimony). More advanced methods use the optimality criterion of maximum likelihood, often within a Bayesian Framework, and apply an explicit model of evolution to phylogenetic tree estimation.[4] Identifying the optimal tree using many of these techniques is NP-hard,[4] so heuristic search and optimization methods are used in combination with tree-scoring functions to identify a reasonably good tree that fits the data.

Neighbor Joining or UPGMA

UPGMA and Neighbor Joining use a clustering procedure that is commonly found in data mining techniques. The method is simple and intuitive which makes it appealing. The method works by clustering nodes at each stage and then forming a new node on a tree. This process continues from the bottom of the tree and in each step a new node is added, and the tree grows upward. The length of the branch at each step is determined by the difference in heights of the nodes at each end of the branch. UPGMA has built in assumptions that the tree is additive and that all nodes are equally distance from the root. Since a “molecular clock” hypothesis assumption poses biological issues, UPGMA is not used much today, but gave way to a very common approach now termed “Neighbor Joining” Neighbor Joining (NJ) works like UPGMA in that it creates a new distance matrix at each step, and creates the tree based on the matrices. The difference is that NJ does not construct clusters but directly calculates distances to internal nodes. The first step in the NJ algorithm is to create

a matrix with the Hamming distance between each node or taxa. The minimal distance is then used to calculate the distance from the two nodes to the node that directly links them. From there, a new matrix is calculated and the new node is substituted for the original nodes that are now joined. The advantage here is that there is not an assumption about the distances between nodes since it is directly calculated.

Steps for building a tree

1. Construct distance matrix
2. Cluster the two shortest distance OTUs into an internal nodes
3. Recalculate the distance matrix
4. Repeat the process until all OTUs are grouped in a single cluster

Maximum Parsimony

Maximum Parsimony (MP) is probably the most widely and accepted method of tree construction to date. The method is different from the previously discussed distance based methods since it uses a character based algorithm. The method works by searching through possible tree structures and assigning a cost to each tree. Parsimony is based on the assumption that the mostly likely tree is the one that requires the fewest number of changes to explain the data in the alignment. The premise that taxa or nodes sharing a common characteristic do so because they inherited that characteristic from a common ancestor.

Conflicts with this major assumption are explained under the term homoplasy. There are three main ways to resolve conflicts: reversal (revert back to original state), convergence (unrelated taxa evolved the same characteristic completely independently) and parallelism (different taxa may have similar mechanisms that cause a characteristic to develop in a certain manner). The tree with the lowest tree score or length, as defined by the number of changes summed along the branches, becomes the most parsimonious tree and is taken as the tree that best represents the evolutionary pattern. Maximum Parsimony is also different from the other methods in that it does not find branch lengths but rather the total overall length in terms of the number of changes. Often MP, finds two or more trees that it deems equal and does not provide a definite answer in how to distinguish which tree represents the actual evolutionary tree. In most cases a strict (majority rule) consensus is used to solve this

dilemma Traditional parsimony uses recursion to search for the minimal number of changes within the trees. This done by starting at the leaf of a tree and working up towards the root, which is known as post-order traversal Another version of parsimony, weighted parsimony, adds a cost factor to the algorithm and weights certain scenarios accordingly. An artifact called long-branch attraction sometimes occurs in parsimony and should be handled. The branch length indicates the number of substitutions between two taxa or nodes. Parsimony assumes that all taxa evolve at the same rate and contribute that same amount of information. Long-branch is the phenomenon in which rapidly evolving taxa are placed together on a tree because they have many mutations. Anytime two long branches are present, they may be attracted to one another.

Steps for building a tree

- Start with multiple alignment
- Construct all possible topologies and base on evolutionary changes to score each of these topologies
- Choose a tree with the fewest evolutionary changes as the final tree

Maximum Likelihood

Proposed in 1981 by Felsenstein, Maximum likelihood (ML) is among the most computationally intensive approach but is also the most flexible ML optimizes the likelihood of observing the data given a tree topology and a model of nucleotide evolution Maximum Likelihood finds the tree that explains the observed data with the greatest probability under a specific model of evolution. ML is different from the other methods in that it is based on probability.

One of the big advantages to ML is the ability to make statistical comparisons between topologies and data sets. ML can return several equally likely trees – pro and con depending on the study Maximum Likelihood makes assumptions that the model used is accurate and if the model does not accurately reflect the underlying data set, the method is inconsistent. ML is designed to be robust, but breaching is assumptions can cause problems. A disadvantage of ML is the extensive computation as well as new evidence that suggest there can be multiple maximum likelihood points for a given phylogenetic tree.

Steps for building a tree

1. Start with a multiple alignment.
2. List all possible topologies of each data partition (i.e., column)
3. Calculate probability of all possible topologies for each data partition.
4. Combine data partitions
5. Identify tree with the highest over all probability at all partitions as most likely phylogeny

Possible questions

1. What is phylogenetic tree? Explain about their types and terminologies.
2. Explain in detailed about the distance methods.
3. Explain the representation of phylogenetic tree. Write short note on dendrograms.
4. Write brief note on maximum parsimony.
5. Write the applications of phylogeny analysis.
6. Write the dynamic programming of multiple sequence alignment.
7. Explain the steps in constructing a phylogeny tree.
8. What is the importance of maximum likelihood approach in phylogenetic tree construction.
9. Explain the role of molecular evolution in the phylogeny analysis.
10. How can be unrooted tree be constructed from the sequence data? Explain.

Karpagam Academy of Higher Education
Department of Biochemistry
II B.Sc., Biochemistry
17BCU404A- Bioinformatics

Question number	Unit	Question	Option I	Option II	Option III	Option IV	Answer
UNIT - IV							
1	2	The scientific discipline concerned with naming organisms is called	taxonomy	cladistics	binominal nomenclature	systematics	taxonomy
2	4	A phylogenetic tree that is "rooted" is one	that extends back to the origin of life on Earth	at whose base is located the common ancestor of all taxa depicted on that tree	that illustrates the rampant gene swapping that occurred early in life's history	with very few branch points	at whose base is located the common ancestor of all taxa depicted on that tree
3	4	The best classification system is that which most closely	unites organisms that possess similar morphologies.	conforms to traditional, Linnaean taxonomic practices	reflects evolutionary history	reflects the basic separation of prokaryotes from eukaryotes	reflects evolutionary history
4	4	A phylogenetic tree constructed using sequence differences in mitochondrial DNA would be most valid for discerning the evolutionary relatedness of	archaeans and bacteria	fungi and animals	Hawaiian silverswords	mooses and ferns	Hawaiian silverswords
5	4	The reason that paralogous genes can diverge from each other within the same gene pool, whereas orthologous genes diverge only after gene pools are isolated from each other, is that	having multiple copies of genes is essential for the occurrence of sympatric speciation in the wild	paralogous genes can occur only in diploid species; thus, they are absent from most prokaryotes	polyploidy is a necessary precondition for the occurrence of sympatric speciation in the wild	having an extra copy of a gene permits modifications to the copy without loss of the original gene product	having an extra copy of a gene permits modifications to the copy without loss of the original gene product
6	4	The Neighbor-Joining method is	Closely related method	Distance clustering method	Single sequence method	Maximum likelihood method	Distance clustering method
7	4	Phylogenetic system of classification is based on	Morphological features	Chemical relationship	Evolutionary relationship	Floral characters	Evolutionary relationship
8	4	Similarity between two short fragments results from the	Evolutionary convergence	Evolutionary Divergence	Common ancestor	All of the above	Evolutionary Divergence
9	4	PHI-EUP is the Multiple sequence alignment program which is a part of the	Genetics Computer Group	Computer Genetics Group	Group of Computer Genetics	None of the above	Genetics Computer Group
10	4	The two main features of any phylogenetic tree are the	clades and the nodes	topology and the branch lengths	clades and the root	alignment and the bootstrap	topology and the branch lengths
11	4	Reconstruction of Phylogenetic tree will be carried out from	Protein sequence	Nucleic acid sequence	Both a and b	None of the above	Both a and b
12	4	Phylogenetic of species can be reconstructed only by comparing	Orthologous genes	Paralogous genes	Both a and b	None of the above	Orthologous genes
13	4	_____ is a broad term for the diagrammatic representation of a phylogenetic tree	cladogram	Phylogram	cladogram	Phylogram	Dendrogram
14	4	A node in a phylogeny represents a	Common ancestor	Different ancestor	Both a and b	None of the above	Common ancestor
15	4	The groups showing similarities due to single ancestors are called	monophyletic	polyphyletic	triphyletic	polyphyletic	monophyletic
16	4	The study of kinds and diversity of organisms and the evolutionary relationships among them is called	systematics	cladistics	kinetics	mechanics	systematics
17	4	The group with unknown evolutionary relationship is	Monophyletic	Polyphyletic	Monophyletic	None	Polyphyletic
18	2	Which of the following systematics have traditional approach?	Evolutionary	Phyline	Numerical	All	Evolutionary
19	4	Why do scientists apply the concept of maximum parsimony?	to decipher accurate phylogenies	to eliminate analogous traits	to identify mutations in DNA codes	to locate homoplasies	to decipher accurate phylogenies
20	4	On a phylogenetic tree, which term refers to lineages that diverged from the same place?	sister taxa	basal taxa	rooted taxa	dichotomous taxa	sister taxa
21	4	What do scientists use to apply cladistics?	homologous traits	homoplasies	analogous traits	monophyletic groups	homologous traits
22	4	What does the trunk of the classic phylogenetic tree represent?	single common ancestor	root of ancestral organisms	new species	old species	single common ancestor
23	4	To apply parsimony to constructing a phylogenetic tree,	choose the tree that assumes all evolutionary changes are equally probable	choose the tree in which the branch points are based on as many shared derived characters as possible	choose the tree that represents the fewest evolutionary changes, either in DNA sequences or morphology	choose the tree with the fewest branch points	choose the tree that represents the fewest evolutionary changes, either in DNA sequences or morphology
24	4	Theoretically, molecular clocks are to molecular phylogenies as radiometric dating is to phylogenies that are based on the	fossil record	geographic distribution of extant species	morphological similarities among extant species	amino acid sequences of homologous polypeptides	fossil record
25	4	Concerning growth in genome size over evolutionary time, which of these does not belong with the others?	orthologous genes	gene duplications	paralogous genes	gene families	orthologous genes
26	4	Cladograms (a type of phylogenetic tree) constructed from evidence from molecular systematics are based on similarities in	morphology	biochemical pathways	habitat and lifestyle choices	mutations to homologous genes	mutations to homologous genes
27	4	Phylogenetic hypotheses (such as those represented by phylogenetic trees) are strongest when	they are based on amino acid sequences from homologous proteins, as long as the genes that code for such proteins contain no introns	each clade is defined by a single derived character	they are supported by more than one kind of evidence, such as when fossil evidence corroborates molecular evidence	they are based on a single DNA sequence that seems to be a shared derived sequence	they are supported by more than one kind of evidence, such as when fossil evidence corroborates molecular evidence
28	4	_____ is a term that is most appropriately associated with clade.	monophyletic	polyphyletic	monophyletic	monophyletic	monophyletic
29	4	A taxon, all of whose members have the same common ancestor, is	monophyletic	polyphyletic	monophyletic	monophyletic	monophyletic
30	4	Shared derived characters are most likely to be found in taxa that are	monophyletic	polyphyletic	monophyletic	polyphyletic	monophyletic
31	4	The importance of computers and of computer software to modern cladistics is most closely linked to advances in	light microscopy	fossil discovery techniques	Linnaean classification	molecular genetics	molecular genetics
32	4	The family that consists of related genes within an organism is called	orthologs	analog	xenologs	xenologs	paralog
33	4	The family that consists of related genes in another organism is called	orthologs	analog	xenologs	xenologs	orthologs
34	4	Which branching diagram is assumed to be an estimate of a phylogeny when branching lengths are proportional to the amount of inferred evolutionary change?	Phylogram	Cladogram	A guide tree	Cardiogram	Phylogram
35	4	Gene duplication results in	orthologs	analog	xenologs	xenologs	paralog
36	4	Two principal ways to construct guide tree in progressive alignment is	UPGMA and Neighbor joining method	Maximum Parsimony	Maximum Likelihood	all the above	UPGMA and Neighbor joining method
37	4	Which of these methods is a distance-based method in tree construction?	Unweighted pair group method with arithmetic mean	Jukes-Cantor	Minimum evolution	Maximum parsimony	Unweighted pair group method with arithmetic mean
38	4	Which one of the following is not a character-based method in tree construction?	Maximum parsimony	Maximum Likelihood	Neighbor joining	Neighbor joining	Maximum parsimony
39	4	A tree representation of a family showing the relationships between members and pattern of inheritance of a given trait is known as	pedigree	physical Map	genetic map	population studies	pedigree
40	4	The study of evolutionary relationships is	Phylogenetics	Molecular Evolution	Cladogenesis	Cladistics	Phylogenetics
41	4	A Minimum branch point in the phylogenetic tree is known as	node	clade	branch	node	node
42	4	Expand UPGMA	Unweighted Pair Group Method with Arithmetic Mean	Unweighted Pair Group Method with All Mean	Upregulated Gene Method with Arithmetic Mean	Unregulated Genome Method with All Mean	Unweighted Pair Group Method with Arithmetic Mean
43	4	One of the most common errors in making and analyzing phylogenetic tree is	trying to infer the evolutionary relationship of genes or proteins in the tree	trying to infer the evolutionary relationship of genes or proteins in the tree	assuming that clades are monophyletic	assuming that clades are monophyletic	using a had multiple sequence alignment as input
44	4	Which one of the following tool can be used to generate neighbor joining trees with or without bootstrap values?	ChustaX	BLAST	Swiss-PDB viewer	ChemSketch	ChustaX
45	4	Molecular phylogeny can be performed with _____ sequences	only DNA	only RNA	only protein	all the above	all the above
46	4	A phylogenetic tree that explicitly represents number of character changes through its branch lengths is	cladogram	Cladogram	Phylogram	Phylogram	Phylogram
47	4	Which of the following is the character based method?	UPGMA	Maximum Parsimony and Maximum Likelihood	Neighbor-Joining	Neighbor-Joining	Maximum Parsimony and Maximum Likelihood
48	4	Which of the following is not an algorithm for generating phylogenetic trees from molecular data?	Neighbor-joining	Parsimony	Maximum likelihood	Jukes & Cantor	Jukes & Cantor
49	4	_____ is a way to judge the reliability of the branches in a tree	bootstraping	clade	branch tree	cladogram	bootstraping
50	4	OTU stands for	operational taxonomic unit	Outgroups	Only translation units	Outlying units	operational taxonomic unit
51	4	A taxon _____	is a species	is a formal grouping at any given level	is a clade	of one type of organism at one level is comparable to another type of organism at the same level	is a formal grouping at any given level
52	4	What does a branch point in a phylogenetic tree represent?	A branch point represents a point at which two evolutionary lineages split from a common ancestor	A branch point represents a gene duplication event	A branch point represents a split between two phyla	A branch point represents a place where one species branches off from another	A branch point represents a point at which two evolutionary lineages split from a common ancestor
53	4	Which of the following methods to establish phylogenetic relationships among organisms has been developed most recently?	comparing physiology	comparing behavioral patterns	comparing the amino acid sequences of proteins and nucleotide sequences of nucleic acids	comparing morphology	comparing the amino acid sequences of proteins and nucleotide sequences of nucleic acids
54	4	Which statement below is true about an outgroup?	The outgroup should be from a lineage known to have diverged before the lineage that includes the ingroup	The outgroup would be found at one of the highest branches of a phylogenetic tree	Outgroup comparison is based on the assumption that homologies present in both the outgroup and ingroup must be derived characters	The outgroup and ingroup display a mixture of shared and derived characters	The outgroup should be from a lineage known to have diverged before the lineage that includes the ingroup
55	4	Unlike a regular phylogenetic tree, phylogenetic trees with branch lengths proportional to time can be used to _____	_____	reflect the rate of evolutionary change	hypothesize the relative relatedness between different taxa	represent the chronological time that has passed since two groups diverged from a common ancestor	represent the chronological time that has passed since two groups diverged from a common ancestor
56	4	Which statement below is true of parsimonious trees?	The most parsimonious tree requires the fewest evolutionary events to have occurred in the form of shared derived characters	The most parsimonious tree requires the fewest evolutionary events to have occurred in the form of shared derived characters	Given the rules of how morphological traits change over time, a tree can be found that reflects the most likely sequence of evolutionary events	The most parsimonious tree requires the fewest evolutionary events to have occurred in the form of shared ancestral characters	The most parsimonious tree requires the fewest evolutionary events to have occurred in the form of shared derived characters
57	4	Paralogous genes	result from gene duplication	are passed from generation to generation in a straight line	cannot diverge in the same gene pool	They increase the size of the genome and provide more opportunity for the evolution of novel characteristics	result from gene duplication
58	4	What is the evolutionary significance of paralogous genes?	They give the absolute time that two species diverged	They give the absolute time that the gene duplication occurred	None of the listed responses is correct	They increase the size of the genome and provide more opportunity for the evolution of novel characteristics	They increase the size of the genome and provide more opportunity for the evolution of novel characteristics
59	4						
60	4						



Crossing over mapwise
Contig
C
B
B
A
C
B
C

C
B
A
C
A
B
A

UNIT-V

SYLLABUS

Protein structure prediction analysis and gene prediction: Levels of protein structure. Protein tertiary structure prediction methods-homology modeling, fold recognition and ab-initio methods. Significance of Ramachandran map. Introduction to genomics, comparative and functional genomics, gene structure in prokaryotes and eukaryotes, gene prediction methods and tools. .

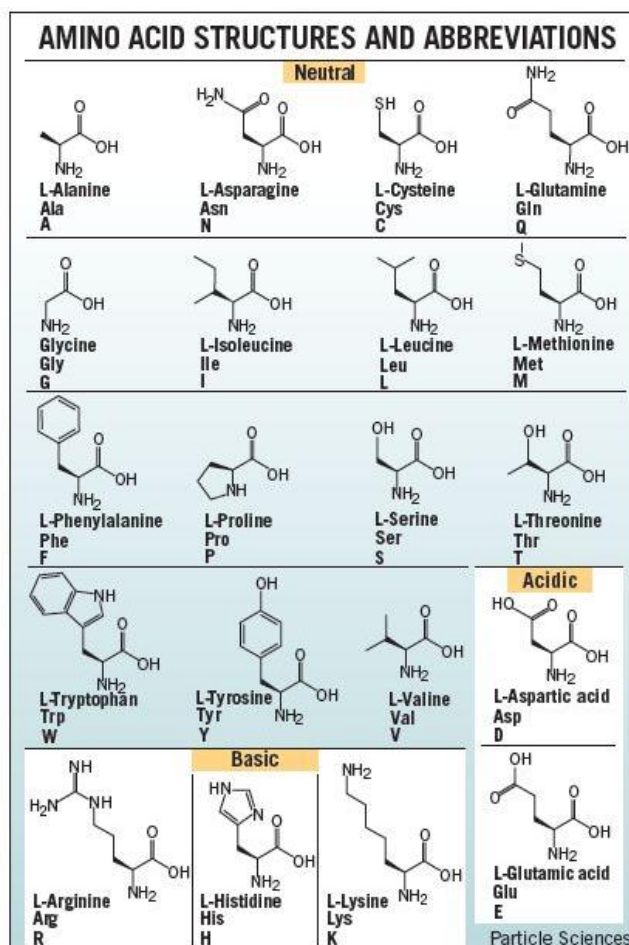
Levels of Protein Structure

The term structure when used in relation to proteins, takes on a much more complex meaning than it does for small molecules. Proteins are macromolecules and have four different levels of structure – primary, secondary, tertiary and quaternary.

Primary Structure

There are 20 different standard L- α -amino acids used by cells for protein construction. Amino acids, as their name indicates, contain both a basic amino group and an acidic carboxyl group. This difunctionality allows the individual amino acids to join together in long chains by forming peptide bonds: amide bonds between the -NH₂ of one amino acid and the -COOH of another. Sequences with fewer than 50 amino acids are generally referred to as peptides, while the terms protein or polypeptide are used for longer sequences. A protein can be made up of one or more polypeptide molecules. The end of the peptide or protein sequence with a free carboxyl group is called the carboxy-terminus or C-terminus. The terms amino-terminus or N-terminus describe the end of the sequence with a free α -amino group. The amino acids differ in structure by the substituent on their side chains. These side chains confer different chemical, physical and structural properties to the final peptide or protein. The structures of the 20 amino acids commonly found in proteins. Each amino acid has both a one-letter and three-letter abbreviation. These abbreviations are commonly used to simplify the written sequence of a peptide or protein.

Depending on the side-chain substituent, an amino acid can be classified as being acidic, basic or neutral. Although 20 amino acids are required for synthesis of various proteins found in humans, we can synthesize only 10. The remaining 10 are called essential amino acids and must be obtained in the diet. The amino acid sequence of a protein is encoded in DNA. Proteins are synthesized by a series of steps called transcription (the use of a DNA strand to make a complimentary messenger RNA strand - mRNA) and translation (the mRNA sequence is used as a template to guide the synthesis of the chain of amino acids which make up the protein). Often, post-translational modifications, such as glycosylation or phosphorylation, occur which are necessary for the biological function of the protein. While the amino acid sequence makes up the **primary structure** of the protein, the chemical/biological properties of the protein are very much dependent on the three-dimensional or tertiary structure.



Secondary Structure

Stretches or strands of proteins or peptides have distinct characteristic local structural conformations or secondary structure, dependent on hydrogen bonding. The two main types of secondary structure are the α -helix and the β -sheet. The α -helix is a right-handed coiled strand. The side-chain substituents of the amino acid groups in an α -helix extend to the outside. Hydrogen bonds form between the oxygen of the C=O of each peptide bond in the strand and the hydrogen of the N-H group of the peptide bond four amino acids below it in the helix. The hydrogen bonds make this structure especially stable. The side-chain substituents of the amino acids fit in beside the N-H groups. The hydrogen bonding in a β -sheet is between strands (inter-strand) rather than within strands (intra-strand). The sheet conformation consists of pairs of strands lying side-by-side. The carbonyl oxygens in one strand hydrogen bond with the amino hydrogens of the adjacent strand. The two strands can be either parallel or anti-parallel depending on whether the strand directions (N-terminus to C-terminus) are the same or opposite. The anti-parallel β -sheet is more stable due to the more well-aligned hydrogen bonds.

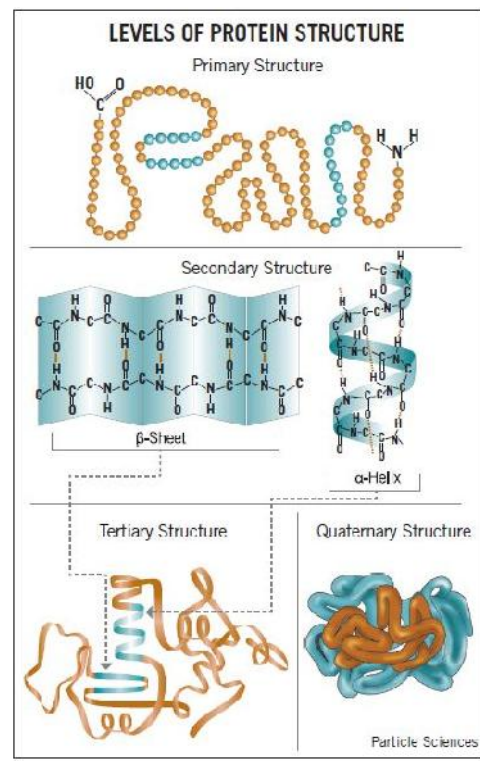
Tertiary Structure

The overall three-dimensional shape of an entire protein molecule is the tertiary structure. The protein molecule will bend and twist in such a way as to achieve maximum stability or lowest energy state. Although the three-dimensional shape of a protein may seem irregular and random, it is fashioned by many stabilizing forces due to bonding interactions between the side-chain groups of the amino acids. Under physiologic conditions, the hydrophobic side-chains of neutral, non-polar amino acids such as phenylalanine or isoleucine tend to be buried on the interior of the protein molecule thereby shielding them from the aqueous medium. The alkyl groups of alanine, valine, leucine and isoleucine often form hydrophobic interactions between one-another, while aromatic groups such as those of phenylalanine and tryptophan often stack together. Acidic or basic amino acid side-chains will generally be exposed on the surface of the protein as they are hydrophilic. The formation of disulfide bridges by oxidation of the sulfhydryl groups on cysteine is an important aspect of the stabilization of protein tertiary structure, allowing

different parts of the protein chain to be held together covalently. Additionally, hydrogen bonds may form between different side-chain groups. As with disulfide bridges, these hydrogen bonds can bring together two parts of a chain that are some distance away in terms of sequence. Salt bridges, ionic interactions between positively and negatively charged sites on amino acid side chains, also help to stabilize the tertiary structure of a protein.

Quaternary Structure

Many proteins are made up of multiple polypeptide chains, often referred to as protein subunits. These subunits may be the same (as in a homodimer) or different (as in a heterodimer). The quaternary structure refers to how these protein subunits interact with each other and arrange themselves to form a larger aggregate protein complex. The final shape of the protein complex is once again stabilized by various interactions, including hydrogen-bonding, disulfide-bridges and salt bridges.



Protein Stability

Due to the nature of the weak interactions controlling the three-dimensional structure, proteins are very sensitive molecules. The term native state is used to describe the protein in its most stable natural conformation in situ. This native state can be disrupted by a number of external stress factors including temperature, pH, removal of water, presence of hydrophobic surfaces, presence of metal ions and high shear. The loss of secondary, tertiary or quaternary structure due to exposure to a stress factor is called denaturation. Denaturation results in unfolding of the protein into a random or misfolded shape. A denatured protein can have quite a different activity profile than the protein in its native form, usually losing biological function. In addition to becoming denatured, proteins can also form aggregates under certain stress conditions. Aggregates are often produced during the manufacturing process and are typically undesirable, largely due to the possibility of them causing adverse immune responses when administered. In addition to these physical forms of protein degradation, it is also important to be aware of the possible pathways of protein chemical degradation. These include oxidation, deamidation, peptide-bond hydrolysis, disulfide-bond reshuffling and cross-linking. The methods used in the processing and the formulation of proteins, including any lyophilization step, must be carefully examined to prevent degradation and to increase the stability of the protein biopharmaceutical both in storage and during drug delivery.

Protein tertiary structure prediction Methods

The biological role of a protein is determined by its function, which is in turn largely determined by its structure. Thus there is enormous benefit in knowing the three dimensional structures of all the proteins. Although more and more structures are determined experimentally at an accelerated rate, it is simply not possible to determine all the protein structures from experiments. As more and more protein sequences are determined, there is pressing need for predicting protein structures computationally. Decades of intense research in this area brought about huge progress in our ability to predict protein structures from sequences only. The protein structure prediction methods can be broadly divided into three categories: 1) homology modeling, 2) threading or fold

recognition, and 3) Ab Initio. Essentially, the classification reflects the degree to which different methods utilize the information content available from the known structure database.

Comparative homology modeling:

So far protein prediction methods based on homology have been the most successful. Homology modeling is based on the notion that new proteins evolve gradually from existing ones by amino acid substitution, addition, and/or deletion and that the 3D structures and functions are often strongly conserved during this process. Many proteins thus share similar functions and structures and there are usually strong sequence similarities among the structurally similar proteins. Strong sequence similarity often indicates strong structure similarity, although the opposite is not necessarily true. Homology modeling tries to identify structures similar to the target protein through sequence comparison. The quality of homology modeling depends on whether there exists one or more protein structures in the protein structure databases that show significant sequence similarity to the target sequence. There are usually four steps in homology based protein structure prediction methods: (1) identify one or more suitable structural templates from the known protein structure databases; (2) align the target sequence to the structural template; (3) build the backbone from the alignment, including the loop region and any region that is significantly different from the template; and (4) place the side-chains. The first two steps, identification of structural templates and alignment of the target sequence onto the parent structures, are usually related. Sequence comparison methods determine sequence similarity by aligning the sequences optimally. The aligned residuals of the structure templates are used to construct the structural model in the second step. The quality of the sequence comparison thus not only determines whether a suitable structural template can be found but also the quality of the alignment between the target sequence and the parent structure, which in turn determines the accuracy of the structural model. Of critical importance is the ability for the sequence comparison to detect remote homologues and to correctly align the target sequence to and parent structure. In the following I discuss the various sequence comparison methods in relation to homology

modeling and their range of applicability, accuracy and shortcomings. For comparative modeling, local sequence comparison methods are usually used since the sequence similarity is most likely over segments of the two sequences. The local sequence comparison can either be pair wise or profile based. Pair wise comparisons, such as the widely used BLAST in the early days, can detect sequence similarities better than 30%. A number of tools have also been developed to detect weak homology relationships. Methods like profile and HMM use a statistical profile of a protein family. To further increase the chance of detecting remote homologues, PSI-BLAST and SAM-T98 build the profile or HMM by searching the database iteratively until no new hits are found. Methods such as PSI-BLAST encode the information about a whole protein family for the target sequence in a model to increase the chance of detecting remote homologies. To further increase the detection sensitivity, the sequences in the structure database can also be encoded in profiles. This forms the basis of the profile-profile based comparison methods (Koehl, 2002). With low sequence identities ($<20\%$), profile-profile methods clearly outperform the other two kinds of methods (Sauder, 2000): profile-profile methods identified more than 90% of homologous pairs, determined from structure-structure similarity comparison, with sequence identity better than 10% and an impressive 38% even for cases with sequence identities between 5% and 9%.

The structure models are constructed from the residuals of the structure template that are aligned to the target sequence in the sequence comparison. The quality of this alignment thus is critical for the accuracy achievable. The aligned residues from sequence comparison are generally different from that from structure-structure comparison though, especially when the sequence identity is low. To assess the ability of the sequence comparison methods to align the sequences correctly, it is instructive to compare the sequence-sequence alignment to the structure-structure alignment of the same pair of proteins. To determine how well the different similarity search methods can detect remote homologies and assess their ability in correctly aligning the sequences, Sauder et al. compared various sequence alignment methods to the CE structure alignment of the SCOP protein structures. For sequence identities less than 30%, profile-based comparison

methods, such as PSI-BLAST and profile-profile comparison, are all obviously better than the pair wise BLAST method. For example, at 10-15% sequence identity, BLAST aligns only 20% correctly while PSIBLAST and profile-profile comparison can correctly align 40% and 48% respectively. This also indicates that there is still large room for improvement in correctly aligning the target sequence to the target structure. One indication of the accuracy of comparative modeling is the sequence identity between the target and the template. It is believed that if two protein sequences have 50% or higher sequence identity, then the RMSD of the alignable portion between the two structures will normally be less than 1 \AA . In the so-called "twilight zone", with sequence identity between 20%~30%, 95% of the sequences with this level of identity have different structures though. When a structure template can indeed be found within the known protein structure databases in such cases, the backbone RMSD can be expected to be no better than 2 \AA . Structurally similar proteins can have low sequence identities in the 8~10% range and can still be identified with sensitive profile-profile based comparison, but the RMSD can be as large as 3~6. The error largely comes from the misalignment from sequence comparison. At such low sequence identity, comparison method that can detect the remote homology as well as align the sequences close to the optimal from structure alignment will be desirable.

Threading or fold recognition:

For evolutionally remotely related proteins, even if the sequence similarity is difficult to detect with sequence comparison methods, there could still be identifiable structural similarity. Structure alignments have been shown to be able to identify homologous protein pairs with sequence similarities less than 10%. When sequence comparison based methods are no longer sensitive enough to recognize the correct fold for the target sequence, fold recognition or threading can still be used to assign the correct fold to the target sequence. Threading or fold recognition is the method by which a library of unique or representative structures is searched for structure analogs to the target sequence, and is based on the theory that there may be only a limited number of distinct protein folds. For example, in an early paper, Chothia postulated that the number of unique protein folds would be on the order of only about 1000 unique protein folds. In estimation,

the number of distinct domains and folds were placed around 7000. Even though the number of new structures solved has been increasing at an accelerated rate (close to 3000 structures solved in 2002), the proportion of new folds, as determined by the CE algorithm (<http://cl.sdsc.edu/ce.html>), to the total number of new structures solved in a given year decreased from an average of ca. 30% in the 80's steadily down to only ca. 8% in year 2001 (<http://www.rcsb.org/pdb/holdings.html>). It is reasonable to expect that as more and more protein structures are determined experimentally, we will be able to find close structure analogues in the databases of known structures for almost any protein sequence in the near future.

Threading or fold recognition involves similar steps as comparative modeling. The difference is in the fold identification step. First of all, a structure library needs to be defined. The library can include whole chains, domains, or even conserved protein cores. Once the library is defined, the target sequence will be fitted to each library entry and an energy function is used to evaluate the fit between the target sequence and the library entries to determine the best possible templates. Depending on the algorithms to align the target sequence with the folds and the energy functions to determine the best fits, the threading methods can roughly be divided into four classes. (1) The earliest threading methods used the environment of each residue in the structure as the energy function and dynamical programming to evaluate the fit and the alignment. (2) Instead of using overly simplified residual environment as the energy function, statistically derived pair wise interaction potentials between residue pairs or atom pairs can be used to evaluate the best possible fits between the target sequence and library folds. In this method, for efficient optimal alignment between the target sequence and the folds, the potential for residual is obtained by summing over all the pair wise potentials involving i , and then "double dynamical programming" method can be used. (3) The third kind of methods does not use any explicit energy function at all. Instead, secondary structures and accessibility of each residue are predicted first and the target sequence and library folds are encoded into strings for the purpose of sequence-structure alignment. (4) Finally, sequence similarity and threading can be combined for fold recognition. For large-scale genome wise protein

structure prediction, sequence similarity can be first used for the initial alignments and the alignments can be evaluated by threading methods.

The threading methods are limited by the high computational cost since each entry in the whole library of thousands of possible folds needs to be aligned in all possible ways to select the fold(s). Another major bottleneck is the energy function used for the evaluation of the alignment. As these functions are drastically simplified for efficient evaluation, it is not reasonable to expect to be able to find the correct folds in all cases with a single form of energy function. Nevertheless, with the current functions, it is possible to reduce the thousands of possible folds to only a few. Similar to the comparative modeling case, for sequence similarities at protein family level, threading can produce alignments that are accurate to 1 to 3 $\%$ or in the case with low sequence similarity at the super-family level, alignment at the range of 3 to 6 $\%$ can still be expected. As more protein structures are determined and sequence comparison methods improve, more and more target sequences fold assignment can be achieved by comparative modeling though. Worth mentioning is the threading program PROSPECT, which performed best in its category in the CASP4 competition. What is unique to PROSPECT is that it is designed to find the globally optimal sequence-structure alignment for the given form of energy function. The divide-and-conquer algorithm is used to speed up the calculation by explicitly avoiding the conformation search space that is shown not to contain the optimal alignment. In several cases that have sequence identity as low as 17%, perfect sequence-structure alignment is still achieved for the alignable portions between the target and template structures. Even in cases that no fold templates exist for the target sequence, important features of the structure are still recognized through threading the target sequence to the structures.

Ab Initio methods:

When no suitable structure templates can be found, Ab Initio methods can be used to predict the protein structure from the sequence information only. Common to all Ab Initio methods are: 1) Suitably defined protein representation and corresponding protein conformation space in that representation; 2) Energy functions compatible with the protein representation; 3) Efficient and reliable algorithms to search the conformational

space to minimize the energy function. The conformations that minimize the energy function are taken to be the structures that the protein is likely to adopt at native conditions. The folding of the protein sequence is ultimately dictated by the physical forces acting on the atoms of the protein and thus the most accurate way of formulating the protein folding or structure prediction problem is in terms of all-atom model subject to the physical forces. Unfortunately the complexity of such a representation makes the solution simply impossible with today's computational capacity. For practical reasons, most Ab Initio prediction methods use reduced representations of the protein to limit the conformational space to manageable size and use empirical energy functions that capture the most important interactions that drive the folding of the protein sequence toward the native structures. Currently, many Ab Initio methods can predict large contiguous segments of the protein to accuracy within 6_ of RMSD and there are several reviews that highlight the success and failure of the current Ab Initio methods. (Hardin, 2002 and references therein). The ROSETTA Ab Initio method performed better than the other Ab initio methods in the recent CASP4 meeting and there are extensive literature covering this method so we concentrate on a brief discussion of method used in ROSETTA. The ROSETTA method also illustrates many features and techniques that are common to the majority of the Ab Initio methods based on reduced representation of the protein and empirical potentials. Discussion of other methods with empirical potentials can be found in Hardin's review.

The ROSETTA method, like many others, uses a reduced representation of the protein as short segments. This representation can be attributed to the observation by Go that local segments of the protein sequence have statistically important preferences for specific local structures and that the tertiary structure has to be consistent with this preference. In ROSETTA the protein is represented by short sequence segments and the local structures they can adopt are assumed to be those found in all the known protein structures. The energy function is defined as the Bayesian probability of structure/sequence matches and this forms the basis of the Monte Carlo sampling of the reduced protein conformational space. The non-local potential, which drives the protein

toward compact folded structure, includes terms that favor paired strands and buried hydrophobic residuals. The solvation effect can also be incorporated in the energy function. A problem intrinsic to the reduced representation of the protein and the simplified empirical potential is that the energy function is not sensitive enough to differentiate the correct native structures from conformations that are structurally close to the native state. The energy landscape calculated from such energy functions will not be properly funneled but flattened and caldera-like around the native structure. In fact, as the native state is approached, the correlation between the calculated energy and the measure of similarity between predicted and native structures are no longer valid. The usual practice is then to produce a large number of decoy structures and then use various filtering and clustering techniques to pick up the more native like structures. Filters can be used to eliminate structures with poorly formed secondary structures and low contact orders compared with that for sequences with compatible length. The other important technique is to use multiple sequences similar to the target sequence to generate decoy structures. Structures thus generated usually form dense clusters that are more compatible to the native structures of protein families of similar sequences than those obtained from a single sequence only.

Many Ab Initio methods now can predict long segments of the protein sequence with backbone atom RMSD less than 6 Å. The predicted local structures are usually right, with the correct contacts among residuals. One of the largest sources of errors was identified to be in the contacts between distant residuals in the sequence as measured by the contact order (CO).

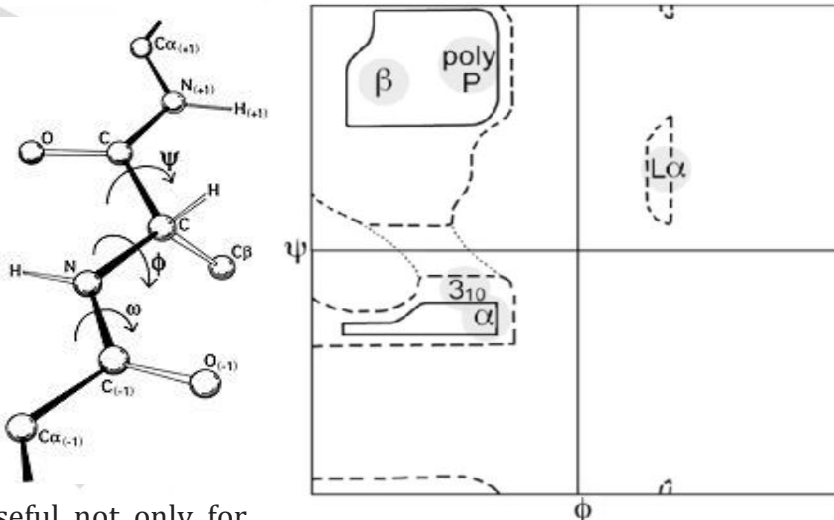
Significance of Ramachandran Plot

A **Ramachandran plot**, also known as a Ramachandran diagram or a $[\phi, \psi]$ plot, was originally developed by Gopalasamudram Ramachandran, an Indian physicist, in 1963. Ramachandran Plot is a way to visualize dihedral angles ψ against ϕ of amino acid residues in protein structure. Ramachandran recognized that many combinations of angles in a polypeptide chain are forbidden because of steric collisions between atoms. Ramachandran plots show the relationship between the phi and psi angles of a protein referring to

dihedral angles between the N and the C-alpha and the C-alpha and the C-beta. As an aside, the omega angle between the C-beta and the N tends to be fixed due to pi-pi interactions. The two-dimensional plot shows the allowed and disfavored values of ψ and ϕ : three-quarters of the possible combinations are excluded simply by local steric clashes. Steric exclusion is the fact that two atoms cannot be in the same place at the same time is the powerful organizing principle that propels the use of the Ramachandron plot.

Dihedral Angles

There are limits to possible distributions of phi and psi angles due to steric clashes between the side chains. Furthermore other limitations from higher order structure will result in the adoption of defined phi-psi angles. Using data from solved crystal structures, it can be seen that the dihedral angles will adopt specific conformations in a protein. Furthermore, it can be noted that some of these conformations relate to specific secondary structures. As seen above, peptides in alpha-helices and beta-sheets adopt a even more limited set of phi-psi angles. Certain amino acids like glycine and proline, which differ from canonical amino acids have an unique Ramachandran plot.



The angles from a Ramachandran plot are useful not only for determining a amino acids' role in secondary structure but can also be used to verify the solution to a crystal structure. Furthermore, it assists with constraining structure prediction simulations and helps with defining energy functions.

The Ramachandran Plot helps with determination of secondary structures of proteins.

- Quadrant I shows a region where some conformations are allowed. This is where rare left-handed alpha helices lie.

- Quadrant II shows the biggest region in the graph. This region has the most favorable conformations of atoms. It shows the sterically allowed conformations for beta strands.
- Quadrant III shows the next biggest region in the graph. This is where right-handed alpha helices lie.
- Quadrant IV has almost no outlined region. This conformation(ψ around -180 to 0 degrees, ϕ around 0 - 180 degrees) is disfavored due to steric clash.

Exception from the principle of clustering around the α -helix and β -strand regions is glycine. Glycine does not have a complex side chain, which allows high flexibility in the polypeptide chain as well as torsion angles, something normally not allowed for other amino acid residues. That is why glycine is often found in loop regions, where the polypeptide chain makes a sharp turn. This is also the reason for the high conservation of glycine residues in protein families, since the presence of turns at certain positions is a characteristic of a particular fold of a protein structure. Another residue with special properties in terms of its torsion angles is proline. Proline, in contrast to glycine, fixes the torsion angles at values, which are very close to those of an extended conformation of the polypeptide (like in a beta-sheet). Proline is often found at the end of helices and functions as a helix disruptor.

Introduction to Genomics

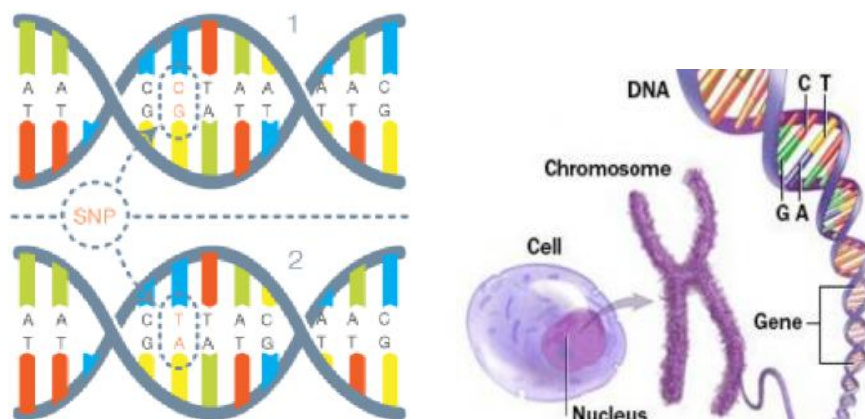
A genome is a complete set of genetic instructions. With 20,000–25,000 genes that contain over 3 billion base pairs of DNA, our genome provides the blueprint for everything that happens in our bodies.

Genomics identifies and analyzes variations in the DNA called single nucleotide polymorphisms (**SNPs**, pronounced “snips”). Over time, a small change can have a great impact on our biochemistry and our health. A SNP may have no effect at all, or it can modify the proteins created by a gene, making them more or less effective.

Over 150 SNPs have been identified as having a direct impact on health. Identifying and analyzing genetic variations provides clinicians with more information than ever before,

enabling personalized interventions that are specifically tailored to a patient's unique biochemical needs.

Genes provide the instructions for making proteins, hormones, immunoglobulins, and enzymes that control the transport nutrients and waste, and the essential communications that keep complex systems synchronized.



Occurrence of Variations: Every time a cell replicates itself to replace dying cells in an organ or tissue, or when a new protein needs to be produced, portions of the chromosome open to expose the needed genetic information. During this process, malfunctions can lead to errors or variations in newly created proteins. Typically, our own system detects and repairs DNA errors before they get translated into a protein, but sometimes they become a permanent part of the DNA, leading to alterations in a person's biochemistry and metabolism. This can result in a chronic disease, including cancer.

Genomic testing shines a light on a person's genetic blueprint, providing the scientific data and analytical information needed to develop a precise, individualized approach to:

- Support the health of our natural defense system
- Intervene with predispositions to certain diseases
- Inform prescription and dosage of drugs
- Boost the healing process
- Maximize longevity and athletic performance
- Improve health and well-being

Genomics vs Genetics: Genetics refers to the traits or characteristics a person inherits from their parents. A common misunderstanding before the 2003 completion of The Human Genome Project, was that genes were considered to be fixed—health was believed to be anchored solely in the DNA inherited from your mother and father. In fact, genes change. They adapt to environmental conditions, a characteristic referred to as genomic plasticity. These findings opened the door to the foundation of Nutrigenetics and Nutrigenomics, the study of the interaction of nutrition and genes, especially regarding the prevention or treatment of disease. This area of functional genomics is gaining increased attention and importance within the medical care process.

Comparative Genomics

At its most literal the term means comparing genomes. This immediately brings to mind DNA and protein sequences and inevitably comparison with the human genome. However, Comparative genomics is more than that. It applies to the comparison of any organism at a variety of levels: DNA or protein sequences, mapping positions and maps, function and evolution. The aim is to decipher how genes function and provide an understanding of the link between genotype and phenotype. Often this is with particular reference to a set of heritable characters or disease, as these are clearly more attractive funding possibilities (even more so when human studies enter into the experimental equation). With livestock, such as cattle, sheep, pigs, fish etc, which are of great economic importance to any country, there are clear commercial requirements to being able to understand the inheritance patterns of advantageous characters and also disease. However, any commercial applications are underpinned by a vast array of academic or "basic" research. When embarking on a research project, it is not always possible to decide categorically which organism to study and which set of genes or heritable characteristics within that organism. Not all organisms are amenable to experimentation, humans being the classic example! This is where "Model Organisms" enter into the subject. The term is self explanatory and an increasing number of different species are being used as tools in our attempts to understand how genes function and the interplay of complex factors such as control sequences, immediate gene environment, the importance of non-coding

elements (repeat sequences, retroelements etc.) and the macro environment surrounding the organism itself. For example; transgenics can be performed in mouse; mutation studies in yeast, *C. elegans*, *Drosophila*, *zebrafish*; analysis of quantitative traits in livestock, identification of evolutionary conserved control elements in Fugu and global comparisons of genome rearrangements in any number of species; the list is endless. As the worldwide sequencing capacity increases and high throughput functional assays are developed, the comparative approach will prove increasingly important, in terms of both sequence comparison and the use of biological models of function.

The aim is to give a brief overview of the subject, concentrating on some areas, such as the genome sequencing projects and the varied utility of model organisms, which can be used to help decipher gene function and evolution. The range of the subject matter and approaches in following, reflecting the wide variety of ways in which this subject is tackled. Comparative genomics will not only tell us much about how human genes function, but also the genotype-phenotype link in many other organisms and the process of evolution. So why are Comparative Genomics and model organisms so important, when, by the time this book is published, the completion of the human draft sequence will have been announced, with total sequence available (no gaps) by 2003.

Comparative genomics is based on collinearity and synteny of genes or chromosomes in diverse species descended from a common ancestor. Comparative genomics studies provide us with the information about orthologous gene functions from different species that are expected to produce similar phenotypes. With the progress of sequencing facilities and the availability of whole-genome sequences for major cereals such as rice, maize, and barley, it is now possible to identify genes and predict their functions in those cereal crops in which their sequencing information is still limited. Comparative genomics predicts the gene function by exploring genomics and postgenomic associations for the genes within plant species or between plants and prokaryotes. The transcriptomics and proteomics data provide important postgenomic evidences of similarity; thus coexpression data from microarray or ribonucleic acid (RNA)-seq can be utilized for

prediction of gene function. Biochemical functions can also be determined using 3D structures.

Availability of large-scale genomic information and conserved synteny between various grass species provides an opportunity to explore the gene function and structure. Comparative genomics has also emerged as an important tool for the identification of micro-RNA (mi-RNA) targets that are conserved during evolution and expected to play essential roles. Sequence comparison using online resources such as "gramene" (<http://www.gramene.org/>) is an important comparative functional genomics analysis tool for crop plants. Comparative analysis of RNA-seq expression profiling of watermelon resulted in the identification of genes homologous to tomato controlling carotenoid synthesis. Comparative analysis of Arabidopsis, rice, barley and maize genomes permitted identification of several important gene families including Sm and WAK.

Functional Genomics

"Functional genomics" -- words that resonate well and that lately appear frequently in headlines and titles of conferences. The uninitiated among us might, however, wonder what they are referring to. Genomics is the science that studies genomes, those unique ensembles of genes organized in higher structures that each living entity inherits from its generator(s) and carries around from birth to death. Such science stems from the astounding progress made in DNA sequencing and molecular biology, bioinformatics, and bioengineering, which has allowed decoding of the genome from entire species (the worm *Caenorhabditis elegans*, the fly *Drosophila*, mouse, humans, and zebrafish).

Functional genomics represents the second step. Now that we have all this information about which genes are present, say in a mouse, and how they are organized, how do these data translate in the generation of all the characteristics and functions that we see in the mouse itself? Are any of these genes or regulatory mechanisms involved in predisposition, development of a disease, or sensitivity to specific treatments? Elucidation of this type of data in humans represents one of the most challenging tasks in today's translational research, owing to the complexity of the biological systems and to the hard-

to-quantify interactions with the external environment. The rewards, however, may be equally amazing.

Functional genomics encompasses a number of different experimental approaches aimed at discovering the biological function of certain genes and defining how sets of genes and their products can interact in health and disease. Among these approaches are: expression profiling using microarrays, mutagenesis, fitness profiling of human homologues in yeast, high-throughput technologies for protein function discovery, the characterization of functional networks, and modeling of biological pathways. The availability of more sophisticated genetic maps inclusive of regulatory networks will then allow us to define the genetic basis underlying predisposition or development not only of single-gene-mutation diseases, but also of far more complex human disorders.

It attempts to describe the functions and interactions of genes and proteins by making use of genome-wide approaches, in contrast to the gene-by-gene approach of classical molecular biology techniques. It combines data derived from the various processes related to DNA sequence, gene expression, and protein function, such as coding and noncoding transcription, protein translation, protein–DNA, protein–RNA, and protein–protein interactions. Together, these data are used to model interactive and dynamic networks that regulate gene expression, cell differentiation, and cell cycle progression.

Studying cells at a systems level has been facilitated by recent technological advancements, as well as the availability of complete genome sequences. Since the landmark publication of the first draft of the human genome in 2001, the genomes of hundreds of organisms from all branches of the tree of life have been sequenced. This has lead to improved annotations of genes and their products, and has enabled genome-wide studies aimed at understanding interactions and molecular processes in the cell.

Gene Structure in Prokaryotes and Eukaryotes

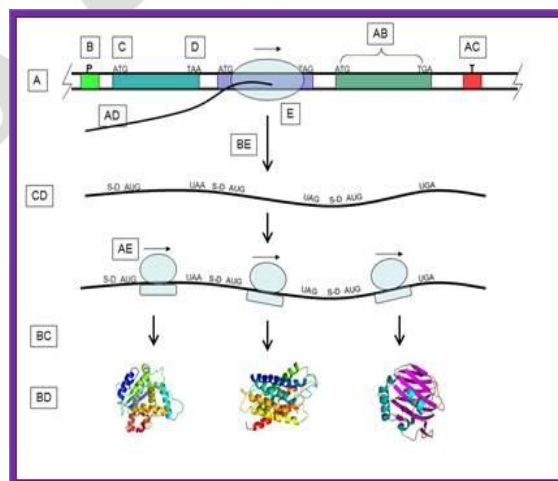
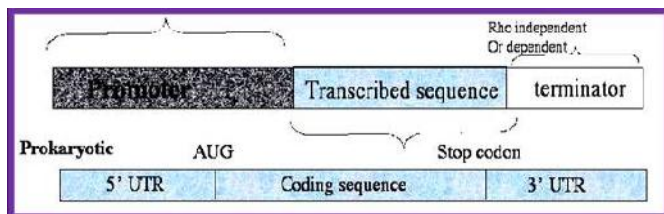
Prokaryotic Gene Structure:

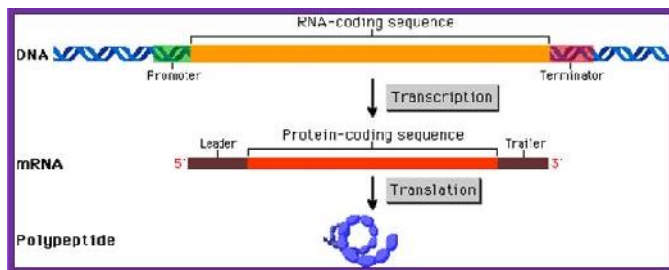
Genes, based on their activity, can be grouped as house keeping genes and others are classed as induced to express or express in stage specific or tissue specific manner.

- House keeping genes express all the time under all normal conditions.
- Most of the gene products of house keeping genes are involved in day-to-day metabolic activities responsible for the maintenance of the cell.
- But when cell confers with other signals such as, change in the temperature, change in the pH, and other environmental features such as exposure to toxic chemicals, or chemical inducers, light, non availability of nutrients, and any other factor that is not ambient to cells and not conducive to cells, specific genes respond to such changes or inductions and express to overcome such hostile or unfavorable situations.

Structural features of promoters of these genes, though basically have common features, but individually they vary slightly from one to the other. The RNA polymerase that is responsible for transcription of the gene is same with some variations in regulator sigma factors. Most of the housekeeping genes have following structural features in general.

The coding region starts with an initiator codon and the reading frame ends in a terminator codon.





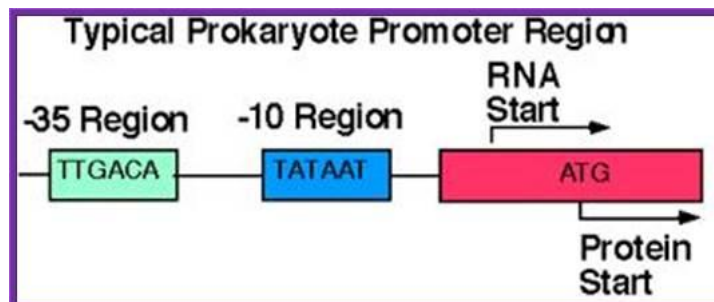
Prokaryotic coding region is collinear to its mRNA, which is collinear to polypeptide chain; The coding region is divided into cistrons separated by intergenic spacers; each cistron codes for a polypeptide chain: Typical Prokaryotic Gene structural elements. The coding region of structural genes is not split, but rRNA genes have spacers within them. The upstream elements from the start of the coding region include promoter elements. Nearly 50 to 100 ntds upstream of the start codon, is the first nucleotide at which transcription initiates, it means, it is at this site the first nucleotide is incorporated into the transcribed RNA.

- The site is called transcriptional initiation site or START.

Nearly 10 nucleotides upstream of the start, there is a sequence called TATAAT or Pribnow box.

- Any nucleotide present on the left of the start is denoted by (-) symbol and the region is called upstream element. The numbers are written as -10, -20, -35 etc.

The start site is the first ntds and symbolized by +1, any sequence to the right of the start is called downstream elements and numbered as +10, +35 and so on.



At -35 there is another consensus sequence TTGACA. These two sequences are the most important promoter elements, for if there is any change in their sequence and position, transcriptional initiation suffers.

- The meaning of a promoter essentially is a distinct sequence module recognized by transcription factors that recruits RNA polymerase (as a holozyeme) and bind to the sequence tightly and initiate transcription by unwinding the helically coiled DNA into transcriptional bubble.

The said sequences not only facilitate the binding of TFs and enzyme and also provide sequence information for the site at which the enzyme to initiate transcription. If any one of the consensus sequences is deleted or changed drastically, the enzyme won't bind, even if it binds, it initiates transcription at different positions.

PROMOTER

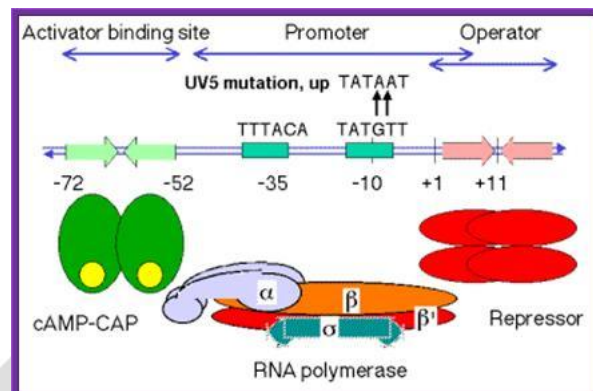
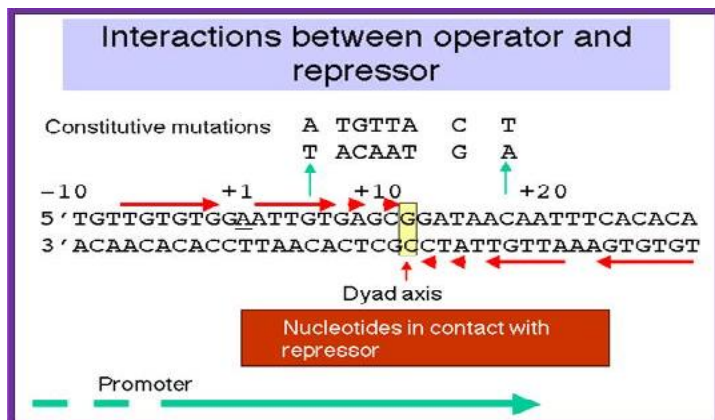
-200	-65	-60	-35	-10	+1
I-----I---//---I-----I-----I-----I-----I-----A					
Enhancers/ Activator			TTGACA	TATAT	
TGTGA--CTCACA					

Thèse séquences are used for the binding of RNA polymerases guided by specific sigma factors

In addition to -10 and -35 sequences one finds up Stream sequences such as activator and ehancer elements. One also finds another sequence called Operator elements to which Repressors/Activators bind.

Gene promoters contain upstream and downstream elements from the start site sequences (InR)





This is typical operator region of Lac operon, with a sequence of dyad axis, this facilitates the binding dimeric repressor proteins; the sequence extends from -10 to +10; this is identified by a dimeric Repressor and it binds and prevents RNA polymerase to act and transcribe. The operator sequences vary with other genes and other repressors.

In lac operon an upstream sequence to TATAAT and -72- 52 sequence one finds a sequence for the binding of an activator protein complex. That regulates transcription of the gene or genes.

The prokaryotic RNA-pol is a Holozyme, when it binds properly in a sequence context; it covers a length from -60 to +20 or little more. It is this segment of the gene that is called Promoter. Whether it is a house keeping gene or special gene, either from prokaryote or eukaryote, the meaning and the function of the promoter is same.

- So promoters act as defining set of sequence defined structural elements, which positions the transcriptional apparatus to initiate transcriptional process. Whether transcription is successfully initiated or not the criteria, but positioning and potentiality for initiating, is an important criteria, then only such sequence modules are called promoters.

For the RNA pol to recognize different genes, the promoter elements have recognition sequences (signature sequences), which are recognized by specific sigma factors that associate with RNA-Pol.

- In prokaryotes there are other sequences in the upstream of the promoter, beyond -35 sequences. Such sequences may present at -65 to -60 or they may be present at -200 or they may be present at -1000 bp upstream or they may present in downstream regions. They are called activator and enhancer sequences. Here enhancer means it increases the efficiency of transcription by 100 to 200%.

The -65 to -60 sequences, position certain factors, whose binding leads to the activation of the polymerase, which was hitherto remained inactive even though it is bound to correct promoter elements. This process is termed as activation and the element as activator elements.

The other sequence at -200 or -1000, is called enhancer, for it enhances the rate of transcription by 100 to 200 fold. This is achieved through certain proteins bind to enhancer elements, and then contacts RNA Holozyme by protein-protein interactions, by way of DNA bending or looping, and enhances the efficiency of enzyme.

Some proteins, after binding to their DNA sequences, they interact with transcriptional apparatus and activate the enzyme. The kind of sequences, however, and the position of the sequences vary from one gene to another.

It is important to realize, that the proteins that bind have specific motifs called DNA binding motifs, and also possess protein-protein interacting domains. The DNA also provides a structural motif in the form of sequence.

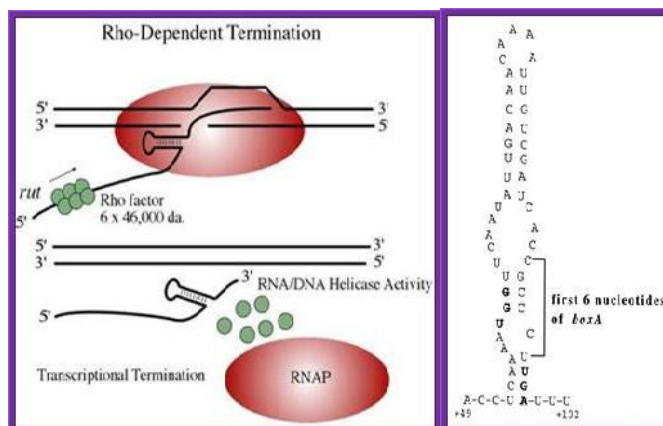
Understanding of DNA sequence context and 3-D structural organization of DNA binding protein is of great importance to appreciate the regulatory processes. Genes that are regulated in response to the needs, basically have the promoter components. In addition, they have operators, activators and enhancers in different positions, which require specific regulator proteins for operation.

Terminal Region of the Gene

In bacteria transcription termination takes place in two modes, one stem-loop Poly (U) mode called Rho independent mode and the other is called Rho dependent mode, it is also called intrinsic termination.

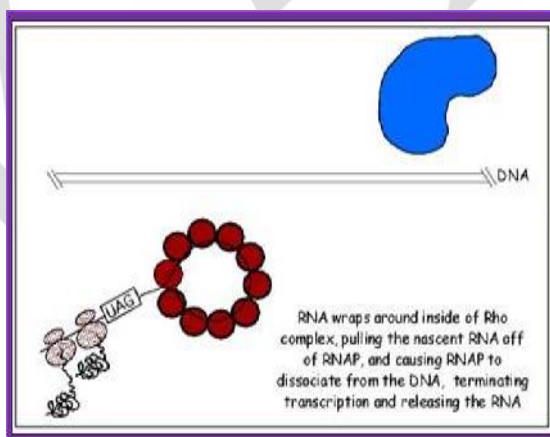
Rho dependent mode:

It uses a factor called Rho (hexamer) and a specific sequence in the terminal region of the gene.

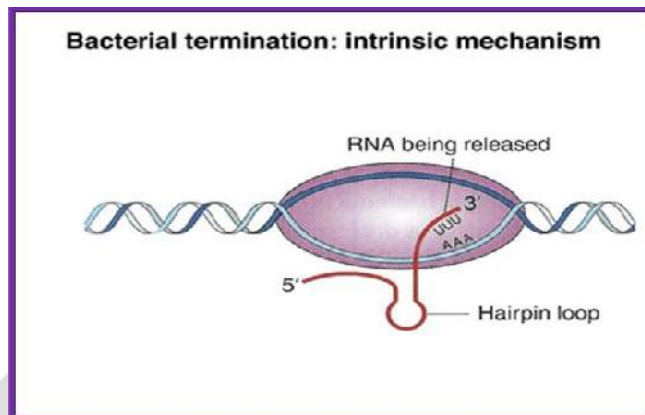
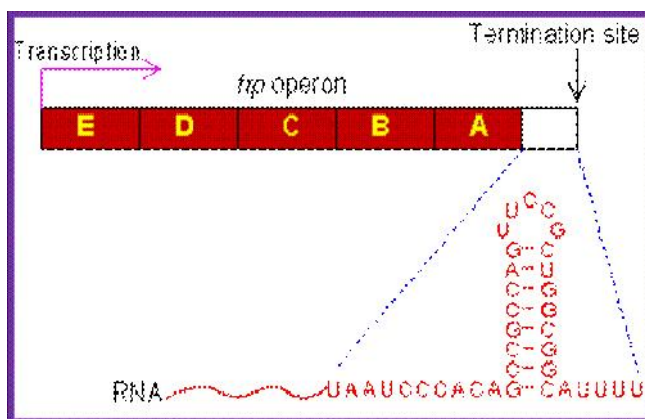


Rho Independent mode: At the end of coding region that is beyond the terminator codon or codons, there are certain sequences positioned which provide a sequence for the transcript to generate a secondary structure that facilitates the termination of transcription.

- One of the structural motifs that the sequence provides is the formation of stem with GC rich sequence and open loop and terminates in 2 to 4 U sequences.



- Transcription termination takes place in two modes one Rho independent manner and the second Rho dependent manner.



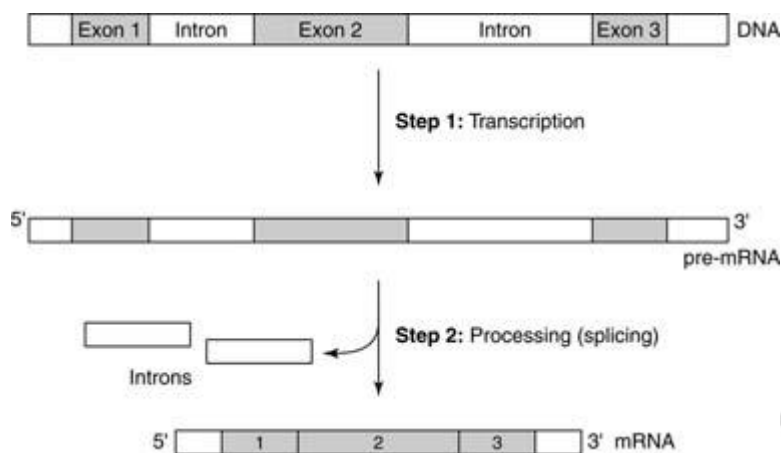
In some of the transcriptional terminator sequences there are specific sequences rich in Cs; they are little longer and far away from the TER codon.

Eukaryotic Gene Structure:

Although humans contain a thousand times more DNA than do bacteria, the best estimates are that humans have only about 20 times more genes than do the bacteria. This means that the vast majority of eukaryotic DNA is apparently nonfunctional. This seems like a contradiction. Why wouldn't more complicated organisms have more DNA? However, the DNA content of an organism doesn't correlate well with the complexity of an organism—the most DNA per cell occurs in a fly species. Other arguments suggest that a maximal number of genes in an organism may exist because too many genes means too many opportunities for mutations. Current estimates say that humans have about 100,000 separate mRNAs, which means about 100,000 expressed genes. This number is still lower than the capacity of the unique DNA fraction in an organism. These arguments lead to the conclusion that the vast majority of cellular DNA isn't functional.

Genes that are expressed usually have **introns** that interrupt the coding sequences. A typical eukaryotic gene, therefore, consists of a set of sequences that appear in mature mRNA (called **exons**) interrupted by introns. The regions between genes are likewise not

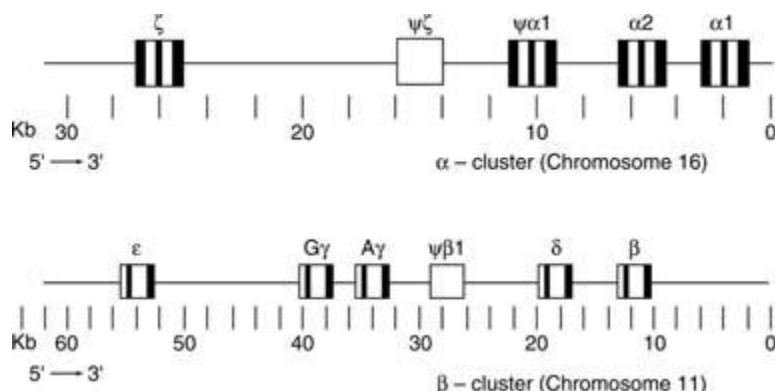
expressed, but may help with chromatin assembly, contain promoters, and so forth.



Intron sequences contain some common features. Most introns begin with the sequence GT (GU in RNA) and end with the sequence AG. Otherwise, very little similarity exists among them. Intron sequences may be large relative to coding sequences; in some genes, over 90 percent of the sequence between the 5' and 3' ends of the mRNA is introns. RNA polymerase transcribes intron sequences. This means that eukaryotic mRNA precursors must be **processed** to remove introns as well as to add the caps at the 5' end and polyadenylic acid (poly A) sequences at the 3' end.

Eukaryotic genes may be clustered (for example, genes for a metabolic pathway may occur on the same region of a chromosome) but are independently controlled. Operons or polycistronic mRNAs do not exist in eukaryotes. This contrasts with prokaryotic genes, where a single control gene often acts on a whole cluster (for example, *lacI* controls the synthesis of β -galactosidase, permease, and acetylase).

One well-studied example of a clustered gene system is the mammalian globin genes. Globins are the protein components of hemoglobin. In mammals, specialized globins exist that are expressed in embryonic or fetal circulation. These have a higher oxygen affinity than adult hemoglobins and thus serve to "capture" oxygen at the placenta, moving it from the maternal circulation to that of the developing embryo or fetus. After birth, the familiar mature hemoglobin (which consists of two alpha and two beta subunits) replaces these globins. Two globin clusters exist in humans: the alpha cluster on chromosome 16, and the beta cluster on chromosome 11.



These clusters, and the gene for the related protein myoglobin, probably arose by duplication of a primordial gene that encoded a single heme-containing, oxygen-binding protein. Within each cluster is a gene designated with the Greek letter Ψ . These are **pseudogenes**—DNA sequences related to a functional gene but containing one or more mutations so that it isn't expressed.

The information problem of eukaryotic gene expression therefore consists of several components: **gene recognition**, **gene transcription**, and **mRNA processing**. These problems have been approached biochemically by analyzing the enzyme systems involved in each step.

Gene Prediction Methods

There are two basic problems in gene prediction: prediction of protein coding regions and prediction of the functional sites of genes. A large number of researches working on this subject have accumulated, which can be classified into four generations in summary. The first generation of programs was designed to identify approximate locations of coding regions in genomic DNA. The most widely known programs were probably TestCode and GRAIL. But they could not accurately predict precise exon locations. The second generation, such as SORFIND and Xpound, combined splice signal and coding region identification to predict potential exons, but did not attempt to assemble predicted exons into complete genes. The next generation of programs attempted the more difficult task of predicting complete gene structures. A variety of programs have been developed, including GeneID, GeneParser, GenLang, and FGENEH. However, the performance of those programs remained rather poor. Moreover, those programs were all based on the assumption that

the input sequence contains exactly one complete gene, which is not often the case. To solve this problem and improve accuracy and applicability further, GENSCAN and AUGUSTUS were developed, which could be classified into the fourth generation.

There are mainly three main classes of methods for computational gene prediction.

A. Statistical or ab initio methods: These methods attempt to predict genes based on statistical properties of the given DNA sequence. Programs are e.g. Genscan, GeneID, GENIE and FGENEH.

B. Comparative methods: The given DNA string is compared with a similar DNA string from a different species at the appropriate evolutionary distance and genes are predicted in both sequences based on the assumption that exons will be well conserved, whereas introns will not. Programs are e.g. CEM (conserved exon method) and Twinscan.

C. Homology methods: The given DNA sequence is compared with known protein structures. Programs are e.g. TBLASTN or TBLASTX, Procrustes and GeneWise.

Sequence similarity searches

Sequence similarity search is a conceptually simple approach that is based on finding similarity in gene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome. This approach is based on the assumption that functional regions (exons) are more conserved evolutionarily than nonfunctional regions (intergenic or intronic regions). Once there is similarity between a certain genomic region and an EST, DNA, or protein, the similarity information can be used to infer gene structure or function of that region. EST-based sequence similarity usually has drawbacks in that ESTs only correspond to small portions of the gene sequence, which means that it is often difficult to predict the complete gene structure of a given region.

Local alignment and global alignment are two methods based on similarity searches. The most common local alignment tool is the BLAST family of programs, which detects sequence similarity to known genes, proteins, or ESTs. Two more types of software, PROCRUSTES and GeneWise, use global alignment of a homologous protein to translated ORFs in a genomic sequence for gene prediction. A new heuristic method based on pairwise genome comparison has been implemented in the software called CSTfinder. The biggest

limitation to this type of approaches is that only about half of the genes being discovered have significant homology to genes in the databases.

Ab initio gene prediction methods

The another class of methods for the computational identification of genes is to use gene structure as a template to detect genes, which is also called ab initio prediction. Ab initio gene predictions rely on two types of sequence information: signal sensors and content sensors. Signal sensors refer to short sequence motifs, such as splice sites, branch points, polypyrimidine tracts, start codons and stop codons. Exon detection must rely on the content sensors, which refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding non-coding sequences by statistical detection algorithms.

Possible Questions

1. Explain the steps involved in homology modeling.
2. Write in brief about the gene prediction methods?
3. Discuss about functional genomics.
4. Explain in detail about Ramachandran map.
5. Discuss about the protein tertiary structure prediction methods.
6. Define genomics? Write note on comparative genomics.
7. Describe about the ab-initio method for structure prediction.
8. Write about the gene structure in prokaryotes.
9. Discuss in detail about the threading approach.
10. Describe the tools available for gene prediction and its importance?

Karpagam Academy of Higher Education
Department of Biochemistry
II B.Sc., Biochemistry
17BCU404A- Bioinformatics

Question number	Unit	Question	Option I	Option II	Option III	Option IV	Answer
UNIT - V							
1	5	Hidden markov model is a	Statistical model	Computational method	Homology model	Sequence analysis	Statistical model
2	5	The secondary prediction method is	nearest neighbour method	hidden markov model	neural network	all the above	neural network
3	5	can be used for homology protein three dimensional structure	MEME	MODELLER	PDCCON	PROSITE	MODELLER
4	5	The secondary structure prediction was discovered in 1951 by	Brink-Rishling	Carey	both a and b	Michael Zhang's	Carey
5	5	Genome repository is	entire genetic material	exon	gene	Protein	entire genetic material
6	5	An interesting region in sequence	intron	exon	EST	all of the above	exon
7	5	Protein structure can be measured by	bond angles	torsion angles	Bond length	All the above	All the above
8	5	A function of position of two atoms in proteins is	Bond length	Bond angle	Torsion angles	transitional angle	Bond length
9	5	A function of position of three atoms in proteins is	Bond length	Bond angle	Torsion angles	transitional angle	Bond angle
10	5	A function of position of four atoms in proteins is	Bond length	Bond angle	Torsion angles	transitional angle	Torsion angles
11	5	Example for α -helical protein is	Keratin	Myoglobin	Collagen	Hemoglobin	Keratin
12	5	One of the below given amino acids is known as imino acid	proline	glycine	Ivaline	Leucine	proline
13	5	In X- crystallography, the diffraction pattern is converted in to electron density maps by	mathematical Fourier transform	fingerprinting	ESR	resonance	mathematical Fourier transform
14	5	The helical rotation of the DNA double helix is known as	axial rise	helix sense	helix pitch	rotation per residue	helix sense
15	5	The first significant macromolecular sequence database was created by	Pearson	M Darchoff	Thompson <i>et al</i>	Altsch <i>et al</i>	M Darchoff
16	5	The solution obtained after extracting the absorbed substances in chromatography is termed as	Solvent	Filterate	subvent	elute	Elute
17	5	A molecular graphics program intended for the visualisation of proteins	Mol mol	rasmol	both a and b	PDB	both a and b
18	5	are macromolecules composed of both RNA and several polypeptides	Chromosomes	ribosomes	autosomes	genes	ribosomes
19	5	Protein sequence determines	genetic variation	genetic disorders	protein structure	domain	protein structure
20	5	Sequence within a single species that arose by gene duplication is called	homologous	paralogous	homologous	orthologous	paralogous
21	5	A series of codes which can be translated into protein	ant codon	translation codon	initiation codon	ORF	ORF
22	5	Scattered X-rays cause positive and negative interference, generating an ordered pattern of signals called	reflections	interference	scattering	diffraction	reflections
23	5	The gene expression implies	gene function	protein	gene regulation	genetic material	gene function
24	5	In the past direct protein sequencing was carried out the process	Sanger method	Edman degradation	spectroscopy	both a & b	spectroscopy
25	5	Protein structure can be determined using spectroscopy	IR	UV	NMR	HPLC	NMR
26	5	The structure data from databank can be downloaded and fed into the	molecular visualization tool to visualize the of the molecules	one dimensional structure	two dimensional structure	three dimensional structure	three dimensional structure
27	5	Three subfields of Genomics are and	Structural, functional and comparative	clustering, chaotic and distance	maximum likelihood, parsimony and	Phylogenetic	Structural, functional and
28	5	Functional genomics is the study of the structure, expression patterns, interactions, and regulations of an encoded by genome.	RNAs and Proteins	DNAs and proteins	genes and proteins	Trns	RNAs and Proteins
29	5	Proteomics is the cataloging and analysis of to determine when a protein is expressed.	DNA	Gene	Proteins	Aminoacids	Proteins
30	5	The term proteomics indicates proteins expressed by a	DNA	Gene	Genome	Aminoacids	Genome
31	5	Proteomics can be divided into and	Expression proteomics and cell-map proteomics	Structural and functional proteomics	Both a and b	conserved regions	Expression proteomics and cell-map proteomics
32	5	The tool that calculates the isoelectric point and molecular weight of an input sequence.	Mw/PI	MI/Pw	PI/Mw	In/Wm	PI/Mw
33	5	Domain helps to attain	Stability of nucleotides	Stability of protein	Stability of gene	Stability of chromosomes	Stability of protein
34	5	The link for PDB	www.rcsb.org/pdb	www.pdb.com	www.pdb.org	www.pdb.ac.in	www.rcsb.org/pdb
35	5	In PDB protein code represents in	integer	alpha	alpha numeric	float	alpha numeric
36	5	Visualization tools	RASMOI	Deepviewer	PDB	RASMOI and Deepviewer	RASMOI and Deepviewer
37	5	The α helix has amino acids per turn with an H bond formed between every fourth residue	3.6	3.4	3.2	3.1	3.6
38	5	The activity of a gene is called	Gene function	gene expression	gene regulation	gene function	gene expression
39	5	β sheets are formed by H bonds between an average of consecutive amino acids	5-6	6-7	5-10	7-9	5-10
40	5	Expand FSSP	Families of structurally similar proteins	Families similar proteins	Function of structurally similar	function similar protein	Families of structurally similar proteins
41	5	A region of secondary structure that is not α helix, a β sheet	sheets	coils	Secondary structure	turns	coils
42	5	A type of architecture that also has a conserved loop structure	blacks	coils	folds	loop	folds
43	5	The Chou-Fasman method is usually accurate in predicting secondary structures	70%	90%	99%	50-60%	50-60%
44	5	is a widely used software for remote homology detection based on pairwise comparison of hidden Markov models.	J Hpredict	predict H	HPredict	HHpred	HHpred
45	5	PDB stands for	Protein Databank	Pattern Databank	Protein Database	Pattern Database	Protein Databank
46	5	Which one of the following is gene prediction tool?	Gene scan	Espay	NCBI	Uniprot	Gene scan
47	5	Which one of the following is gene prediction tool?	Swiss Prot	Procheck	Gen Mark	Trimmer	Gen Mark
48	5	Comparative modeling is also called as	Threading	Ab initio	Homology modeling	None of the above	Homology modeling
49	5	How many levels of protein structure are	Three	Four	Two	One	Four
50	5	Fold recognition is also called as	Threading	Ab initio	Homology modeling	None of the above	Threading
51	5	Biologically active protein structure is	Primary Level	Secondary Level	Both a and b	Tertiary Level	Tertiary Level
52	5	Linear sequence of amino acids is	Primary Level	Secondary Level	Both a and b	Tertiary Level	Primary Level
53	5	Hydrogen bonds stabilises the	Primary Level	Secondary Level	Both a and b	Tertiary Level	Secondary Level
54	5	Disorientation of protein monomers is	Primary Level	Secondary Level	Tertiary Level	Quaternary Level	Quaternary Level
55	5	Disulfide bonds playing a major role in stabilizing the protein structure at	Primary Level	Secondary Level	Tertiary Level	Both a and b	Tertiary Level
56	5	Which of the following amino acid is exceptional in Ramachandran plot?	Glycine	Methionine	Lysine	Alanine	Glycine
57	5	Which one of the following amino acids is Ramachandran plot exception?	Methionine	Lysine	Alanine	Proline	Proline
58	5	Which of the following technique is used in genomics?	SAGE	Protein Purification	Homology Modeling	Virtual screening	SAGE
59	5	Which one of the following technique is used in differential gene expression?	SAGE	Protein Purification	Homology Modeling	Virtual screening	SAGE
60	5	Which one of the following is not a gene prediction tool?	GENESH	GeneMark	PDB	GENEID	PDB

Crossine over machine

Contig

C
B
B
A
C
B
C

C
B
A
C
A
B
A

B
B

A
A
C
B
B
B
A

A
D
B
B
A
B
B
C
A
C
B

A
C
A
A

B
B
A
B
A
C
B
A
A
C
C
C

B
C
B
B
C
B