KARPAGAM ACADEMY OF HIGHER EDUCATION

(Established Under Section 3 of UGC Act 1956)
Coimbatore – 641 021.
(For the candidates admitted from 2016 onwards)

DEPARTMENT OF MICROBIOLOGY

CLASS: III B.Sc (MB) **BATCH – 2016 -2019**
SUBJECT NAME: BIOINFORMATICS **SUB.CODE: 17MBU502B**
SEMESTER: V **4H – 4C**

SYLLABUS

Instruction Hours / week:L: 4 T: 0 P: 0 **Marks:** Internal: **40** External: **60** Total: **100**
End Semester Exam: 3 Hours

Unit I

RDBMS - Definition of relational database. Mode of data transfer (FTP, SFTP, SCP), advantage of encrypted data transfer.

Unit II

Biological databases – nucleic acid, genome, protein sequence and structure, gene expression databases, Database of metabolic pathways, Mode of data storage - File formats - FASTA, Genbank and Uniprot, Data submission & retrieval from NCBI, EMBL, DDBJ, Uniprot, PDB.

Unit III

Local and Global Sequence alignment, pairwise and multiple sequence alignment. Scoring an alignment, scoring matrices, PAM & BLOSUM series of matrices. Types of phylogenetic trees, Different approaches of phylogenetic tree construction - UPGMA, Neighbour joining, Maximum Parsomony, Maximum likelihood.

Unit IV

Diversity of Genomes: Viral, prokaryotic & eukaryotic genomes Genome, transcriptome, proteome, 2-D gel electrophoresis, Maldi Tof spectroscopy. Major features of completed genomes: *E.coli, S.cerevisiae, Arabidopsis,* Human.

Unit V

Hierarchy of protein structure - primary, secondary and tertiary structures, modeling. Structural Classes, Motifs, Folds and Domains. Protein structure prediction in presence and absence of structure template Energy minimizations and evaluation by Ramachandran plot Protein structure and rational drug design.

SUGGESTED READINGS

1.Saxena Sanjay (2003) A First Course in Computers, Vikas Publishing.
2.Pradeep and Sinha Preeti (2007) Foundations of Computing, 4th ed., BPB Publications.
3.Lesk M.A.(2008) Introduction to Bioinformatics . Oxford Publication, 3rd International Student Edition.
4. Rastogi S.C., Mendiratta N. and Rastogi P. (2007) Bioinformatics: methods and applications, genomics, proteomics and drug discovery, 2nd ed. Prentice Hall India Publication.
5.Primrose and Twyman (2003) Principles of Genome Analysis & Genomics. Blackwell.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore - 641021.
(For the candidates admitted from 2016 onwards)
DEPARTMENT OF MICROBIOLOGY

SUBJECT : BIOINFORMATICS

SEMESTER : V

SUBJECT CODE: 17MBU502B CLASS : III BSc Microbiology

Duration Hours	Topics to be covered	Support materials
	Unit I	
1 hour	Introduction to Computer Fundamentals	T- 35
1 hour	RDBMS – Definition of relational database	T, 350-357
1 hour	Mode of data transfer	W1
1 hour	FTP	W1
1 hour	SFTP, SCP	W2
1 hour	Encrypted data	W3
1 hour	Advantages of encrypted data	W3
1 hour	Question paper discussion	
1 hour	Revision	
9 hours	Total no. of hours planned for Unit I	

References

T1- Gibas and Jambeck (2001).Developing Bioinformatics Computer Skills. 1st edition.Shroff Publishers

W1- www.itprotoday.com

W2 - https://www.ssh.com/ssh/sftp/

W3- www.wired.co.uk/article/encryption-software-app-private-data-safe

Duration Hours	Topics to be covered	Support materials
	Unit I1	
1 hour	Introduction to Biological Database	T1-Pg. 10-17
1 hour	Biological Database- protein	T2- Pg. 36- 65
1 hour	Nucleic acid and structure	T1- pg. 14
1 hour	Gene expression databases, database of metabolic pathways	T1- pg. 21
1 hour	Mode of Data storage - FASTA	T3-pg.468-469

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore - 641021.
(For the candidates admitted from 2016 onwards)
DEPARTMENT OF MICROBIOLOGY

SUBJECT : BIOINFORMATICS

SEMESTER : V

SUBJECT CODE: 17MBU502B CLASS : III BSc Microbiology

1 hour	Mode of Data storage- Genbank and Uniprot	T3- pg 52, 63
1 hour	Data submission and retrieval from NCBI	T1- pg. 18
1 hour	Data submission and retrieval from EMBL	T3-pg. 47-50
1 hour	Data submission and retrieval from DDBJ	T3- pg. 56-57
1 hour	Data submission and retrieval from Uniprot, PDB	T3- pg. 52
10 hours	Total no. of hours planned for Unit II	

References

T1- Attwood TK and Parry- Smith DJ (2006) Introduction to Bioinformatics. Pearson Education Ltd.

T2- Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press

T3- Orpita Bosu and Thukral, Bioinformatics Databases, Tools and Algorithms.Oxford University Press

Duration Hours	**Topics to be covered**	**Support materials**
	Unit III	
1 hour	Local alignment	T1- pg. 31-33
1 hour	Global sequence alignment	T1- pg. 34-40
1 hour	Pairwise and multiple sequence alignment	T1- pg. 34-40, 63
1 hour	Scoring an alignment and scoring matrices	T1- pg. 64-73
1 hour	PAM and BLOSUM	T1- pg. 64-73
1 hour	Phylogenetic tree	T1- pg. 127-130
1 hour	Different approaches of phylogenetic tree construction	T1- pg. 131-133 pg. 163-168
1 hour	UPGMA	T1- pg. 127-130
1 hour	Neighbour joining	T1- pg. 127-130
1 hour	Maximum parsomony and maximum likelihood	T1- pg. 127-130
10 hours	Total no. of hours planned for Unit III	

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore - 641021.
(For the candidates admitted from 2016 onwards)
DEPARTMENT OF MICROBIOLOGY

SUBJECT : BIOINFORMATICS

SEMESTER : V

SUBJECT CODE: 17MBU502B CLASS : III BSc Microbiology

References

T1 - Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press

Duration Hours	Topics to be covered	Support materials
	Unit IV	
1 hour	Viral genome	W1
1 hour	Prokaryotic and Eukaryotic genome	W1 and W2
1 hour	Transcriptome, Proteome and genome	W3
1 hour	2-D gel electrophoresis, MALDI-TOF	W4, W5 and W6
1 hour	Major features of completed genome- *E.coli*	W7
1 hour	*S.cerevisae, Arabidopsis*	W8
1 hour	Human	W9
1 hour	Revision	
1 hour	Class Test	
1 hour	Possible question discussion	
10 hours	Total no. of hours planned for Unit IV	

References

W1 -www.nature.com/scitable/topicpage/genome-packaging-in-prokaryotes-the-circular-chromosome-9113

W2 - http://www2.csudh.edu/nsturm/CHEMXL153/GenomeOrganization.htm

W3 - https://www.ncbi.nlm.nih.gov/books/NBK21125/

W4 - https://www.creative-proteomics.com/technology/maldi-tof-mass-spectrometry.htm

W5 - https://www.shimadzu.com/an/lifescience/maldi/princpl1.html

W6- http://www.premierbiosoft.com/tech_notes/mass-spectrometry.html

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore - 641021.
(For the candidates admitted from 2016 onwards)
DEPARTMENT OF MICROBIOLOGY

SUBJECT : BIOINFORMATICS

SEMESTER : V

SUBJECT CODE: 17MBU502B CLASS : III BSc Microbiology

W7 -http://utminers.utep.edu/rwebb/html/the e.coli genome.html

W8 - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308767/

W9-– http://en.wikipedia.org/wiki/Human_Genome_Project

Duration Hours	Topics to be covered	Support materials
	Unit V	
1 hour	Hierarchy of protein structure-primary, secondary	T1- pg. 173-177 T1- pg. 178-180
1 hour	Tertiary structure, modeling	T1- pg. 180-182
1 hour	Structural classes, motifs	T1- pg. 200-212
1 hour	Folds and domains	T1- pg. 200-212
1 hour	Protein structure prediction –presence of structural template	T1- pg. 212-213
1 hour	Absence of structural template	T1- pg. 212-213 Pg. 213-214
1 hour	Energy minimizations	T1- pg. 200
1 hour	Ramachandran plot	T1- pg. 187-201
1 hour	Protein structure and rational drug design	T1- pg. 214-226
1 hour	Discussion of Previous Question papers	
10 hours	Total no. of hours planned for Unit V	

References

T1 - Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press

SUGGESTED READINGS

1. Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore - 641021.
(For the candidates admitted from 2016 onwards)

DEPARTMENT OF MICROBIOLOGY

SUBJECT : BIOINFORMATICS

SEMESTER : V

SUBJECT CODE: 17MBU502B CLASS : III BSc Microbiology

2. Gibas and Jambeck (2001).Developing Bioinformatics Computer Skills. 1st edition.Shroff Publishers
3. Attwood TK and Parry- Smith DJ (2006) Introduction to Bioinformatics. Pearson Education Ltd.

Unit I: Introduction to Computer Fundamentals

RDBMS - Definition of relational database. Mode of data transfer (FTP, SFTP, SCP), advantage of encrypted data transfer.

A **relational database management system** (**RDBMS**) is a database management system (DBMS) based on the relational modelinvented by Edgar F. Codd at IBM's San Jose Research Laboratory.

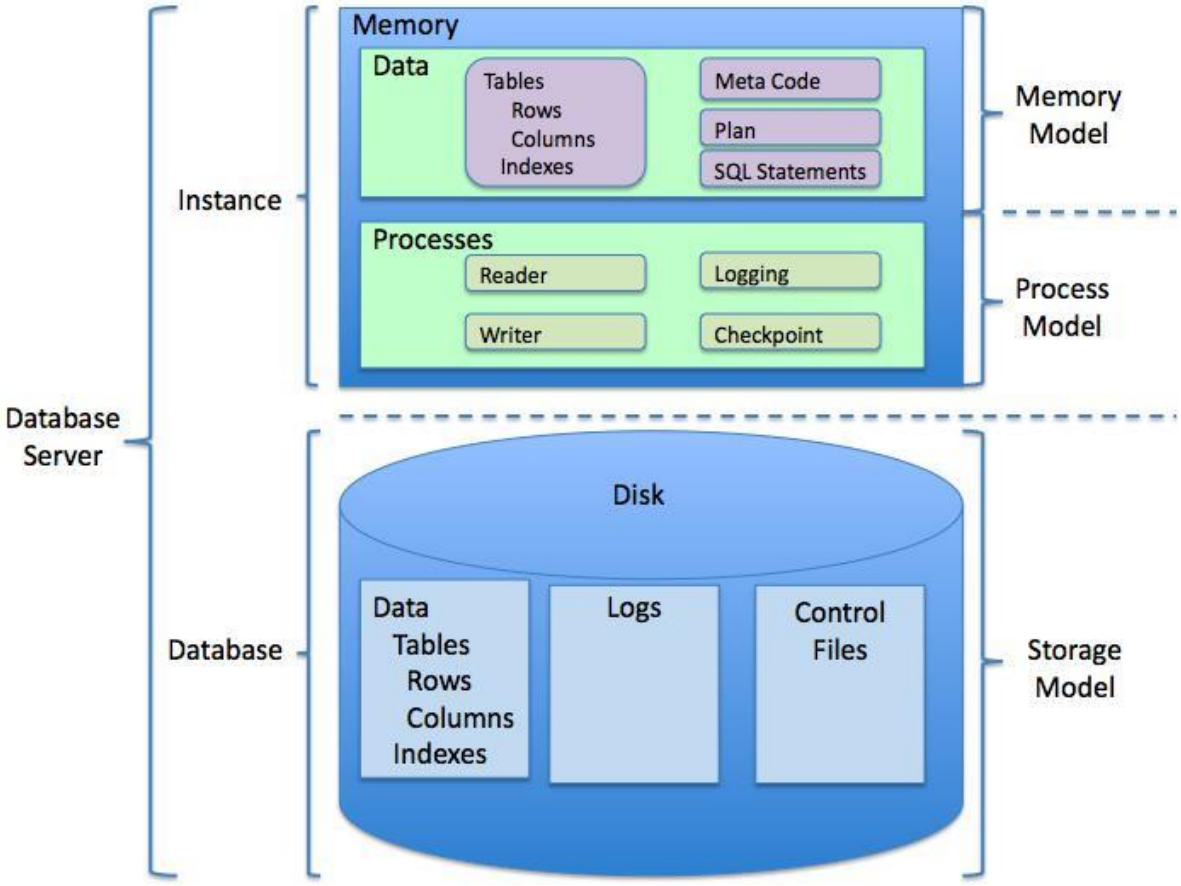
- Most databases in widespread use today are based on his relational database model.

RDBMSs have been a common choice for the storage of information in databases used for financial records, manufacturing and logistical information, personnel data, and other applications since the 1980s.

- Relational databases have often replaced legacy hierarchical databases and network databases because they were easier to implement and administer.

- Nonetheless, relational databases received continued, unsuccessful challenges by object database management systems in the 1980s and 1990s, (which were introduced in an attempt to address the so-called object-relational impedance mismatch between relational databases and object-oriented application programs), as well as by XML database management systems in the 1990s.

- However, due to the expanse of technologies, such as horizontal scaling of computer clusters, NoSQL databases have recently become popular as an alternative to RDBMS databases.



FTP: File Transfer Protocol

Introduction

FTP is another commonly used application. It is the Internet standard for file transfer. We must be careful to differentiate between *file transfer*, which is what FTP provides, *and file access*, which is provided by applications such as NFS.

- The file transfer provided by FTP copies a complete file from one system to another system. To use FTP we need an account to login to on the server, or we need to use it with a server that allows *anonymous FTP*/

- Like Telnet, FTP was designed from the start to work between different hosts, running different operating systems, using different file structures, and perhaps different character sets. Telnet, however, achieved heterogeneity by forcing both ends to deal with a single standard: the NVT using 7-bit ASCII.
- FTP handles all the differences between different systems using a different approach. FTP supports a limited number of file types (ASCII, binary, etc.) and file structures (byte stream or record oriented).
- RFC 959 [Postel and Reynolds 1985] is the official specification for FTP. This RFC contains a history of the evolution of file transfer over the years.

FTP Protocol

FTP differs from the other applications that we've described because it uses two TCP connections to transfer a file.

1. The *control connection is* established in the normal client-server fashion. The server does a passive open on the well-known port for FTP (21) and waits for a client connection. The client does an active open to TCP port 21 to establish the control connection. The control connection stays up for the entire time that the client communicates with this server. This connection is used for commands from the client to the server and for the server's replies.

   The IP type-of-service for the control connection should be "minimize delay" since the commands are normally typed by a human user. A *data connection* is created each time a file is transferred between the client and server. (It is also created at other times, as we'll see later.)

   The IP type-of-service for the data connection should be "maximize throughput" since this connection is for file transfer.

Figure shows the arrangement of the client and server and the two connections between them.

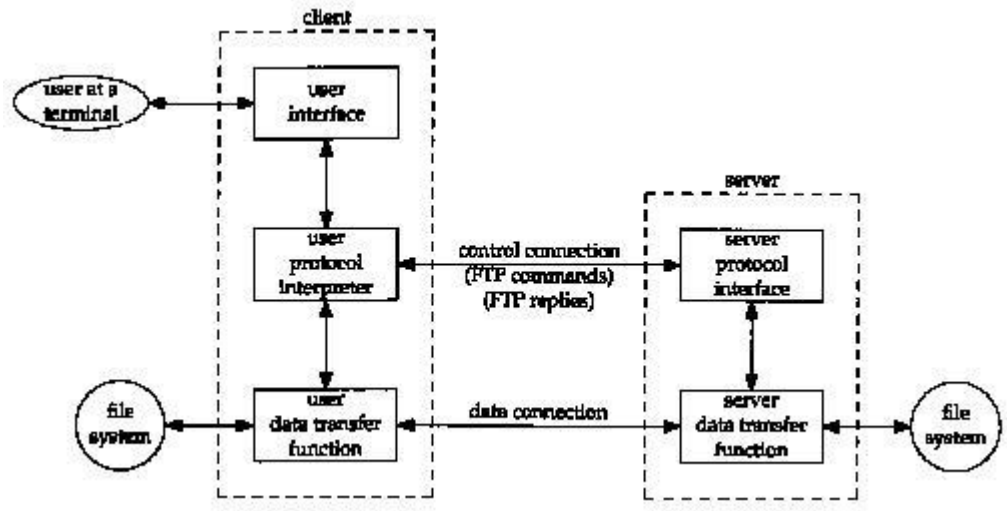


Figure Processes involved in file transfer.

This figure shows that the interactive user normally doesn't deal with the commands and replies that are exchanged across the control connection. Those details are left to the two protocol interpreters.

- The box labeled "user interface" presents whatever type of interface is desired to the interactive user (full-screen menu selection, line-at-a-time commands, etc.) and converts these into FTP commands that are sent across the control connection. Similarly the replies returned by the server across the control connection can be converted to any format to present to the interactive user.
- This figure also shows that it is the two protocol interpreters that invoke the two data transfer functions, when necessary.

Data Representation

- Numerous choices are provided in the FTP protocol specification to govern the way the

file is transferred and stored. A choice must be made in each of four dimensions.

1. **File type.**
    a. ASCII file type.

    (Default) The text file is transferred across the data connection in NVT ASCII. This requires the sender to convert the local text file into NVT ASCII, and the receiver to convert NVT ASCII to the local text file type. The end of each line is transferred using the NVT ASCII representation of a carriage return, followed by a linefeed. This means the receiver must scan every byte, looking for the CR, LF pair.

    b. EBCDIC file type.

    An alternative way of transferring text files when both ends are EBCDICsystems.

    c. Image file type. (Also called binary.)

    The data is sent as a contiguous stream of bits. Normally used to transfer binary files.

    d. Local file type.

    A way of transferring binary files between hosts with different byte sizes. The number of bits per byte is specified by the sender. For systems using 8-bit bytes, a local file type with a byte size of 8 is equivalent to the image file type.

2. Format control. This choice is available only for ASCII and EBCDIC file types.
    a. Nonprint.

    (Default) The file contains no vertical format information.

    b. Telnet format control.

    The file contains Telnet vertical format controls for a printer to interpret.

    c. Fortran carriage control.

    The first character of each line is the Fortran format control character.

3. Structure.
    a. File structure.

       (Default) The file is considered as a contiguous stream of bytes. There is no internal file structure.
    b. Record structure.

       This structure is only used with text files (ASCII or EBCDIC).
    c. Page structure.

       Each page is transmitted with a page number to let the receiver store the pages in a random order. Provided by the TOPS-20 operating system. (The Host Requirements RFC recommends against implementing this structure.)
4. Transmission mode. This specifies how the file is transferred across the data connection.
    a. Stream mode.

       (Default) The file is transferred as a stream of bytes. For a file structure, the end-of-file is indicated by the sender closing the data connection. For a record structure, a special 2-byte sequence indicates the end-of-record and end-of-file.
    b. Block mode.

       The file is transferred as a series of blocks, each preceded by one or more header bytes.
    c. Compressed mode.

       A simple run-length encoding compresses consecutive appearances of the same byte. In a text file this would commonly compress strings of blanks, and in a binary file this would commonly compress strings of 0 bytes. (This is rarely used or supported. There are better ways to compress files for FTP.)

If we calculate the number of combinations of all these choices, there could be 72 different ways to transfer and store a file. Fortunately we can ignore many of the options, because they are either antiquated or not supported by most implementations.

Common Unix implementations of the FTP client and server restrict us to the following choices:

- Type: ASCII or image.
- Format control: nonprint only.
- Structure: file structure only
- Transmission mode: stream mode only.

This limits us to one of two modes: ASCII or image (binary).

SFTP (SSH File Transfer Protocol)

is a secure file transfer protocol. It runs over the SSH protocol. It supports the full security and authentication functionality of SSH.

SFTP has pretty much replaced legacy FTP as a file transfer protocol, and is quickly replacing FTP/S. It provides all the functionality offered by these protocols, but more securely and more reliably, with easier configuration. There is basically no reason to use the legacy protocols any more.

SFTP also protects against password sniffing and man-in-the-middle attacks. It protects the integrity of the data using encryption and cryptographic hash functions, and autenticates both the server and the user.

SFTP PORT NUMBER

SFTP port number is the SSH port 22 (follow the link to see how it got that number). It is basically just an SSH server. Only once the user has logged in to the

server using SSH can the SFTP protocol be initiated. There is no separate SFTP port exposed on servers. No need to configure another hole into firewalls.

SFTP CLIENT FOR WINDOWS AND MAC

Many SFTP client implementations are available. Many SSH clients support SFTP.

- Tectia SSH Client
- WinSCP
- FileZilla
- PuTTY
- Cyberduck

SFTP SERVER FOR LINUX, WINDOWS, AND MAC

SFTP server usually comes as part of an SSH implementation. Most organizations use either Tectia SSH or OpenSSH as the server; both come with SFTP server implementations out-of-the-box.

- Tectia SSH Server for Windows
- Tectia SSH Server for IBM z/OS mainframes
- OpenSSH - open source server for Linux & Unix
- FileZilla - a free sftp server for Windows

SCP COMMAND ON LINUX

The scp command is a file transfer program for SFTP in Linux. The scp command line interface was designed after the old rcp command in BSD Unix.

A Definition of Data Encryption

Data encryption translates data into another form, or code, so that only people with access to a secret key (formally called a decryption key) or password can read it.

- Encrypted data is commonly referred to as ciphertext, while unencrypted data is called plaintext. Currently, encryption is one of the most popular and effective data security methods used by organizations.
- Two main types of data encryption exist - asymmetric encryption, also known as public-key encryption, and symmetric encryption.

The Primary Function of Data Encryption

The purpose of data encryption is to protect digital data confidentiality as it is stored on computer systems and transmitted using the internet or other computer networks.

- The outdated data encryption standard (DES) has been replaced by modern encryption algorithms that play a critical role in the security of IT systems and communications.
- These algorithms provide confidentiality and drive key security initiatives including authentication, integrity, and non-repudiation.
- Authentication allows for the verification of a message's origin, and integrit y provides proof that a message's contents have not changed since it was sent. Additionally, non-repudiation ensures that a message sender cannot deny sending the message.

The Process of Data Encryption

Data, or plaintext, is encrypted with an encryption algorithm and an encryption key. The process results in ciphertext, which only can be viewed in its original form if it is decrypted with the correct key.

- Symmetric-key ciphers use the same secret key for encrypting and decrypting a message or file. While symmetric-key encryption is much faster than asymmetric encryption, the sender must exchange the encryption key with the recipient before he can decrypt it.

Challenges to Contemporary Encryption

The most basic method of attack on encryption today is brute force, or trying random keys until the right one is found.

- Of course, the length of the key determines the possible number of keys and affects the plausibility of this type of attack. It is important to keep in mind that encryption strength is directly proportional to key size, but as the key size increases so do the number of resources required to perform the computation.
- Alternative methods of breaking a cipher include side-channel attacks and cryptanalysis. Side-channel attacks go after the implementation of the cipher, rather than the actual cipher itself. These attacks tend to succeed if there is an error in system design or execution. Likewise, cryptanalysis means finding a weakness in the cipher and exploiting it. Cryptanalysis is more likely to occur when there is a flaw in the cipher itself.

Data Encryption Solutions

- Data protection solutions for data encryption can provide encryption of devices, email, and data itself.
- In many cases, these encryption functionalities are also met with control capabilities for devices, email, and data. Companies and organizations face the challenge of protecting data and preventing data loss as employees use external devices, removable media, and web applications more often as a part of their daily business procedures.
- Sensitive data may no longer be under the company's control and protection as employees copy data to removable devices or upload it to the cloud. As a result, the best data loss prevention solutions prevent data theft and the introduction of malware from removable and external devices as well as web and cloud applications.

- In order to do so, they must also ensure that devices and applications are used properly and that data is secured by auto-encryption even after it leaves the organization.
- As we mentioned, email control and encryption is another critical component of a data loss prevention solution. Secure, encrypted email is the only answer for regulatory compliance, a remote workforce, BYOD, and project outsourcing.
- Premier data loss prevention solutions allow your employees to continue to work and collaborate through email while the software and tools proactively tag, classify, and encrypt sensitive data in emails and attachments. The best data loss prevention solutions automatically warn, block, and encrypt sensitive information based on message content and context, such as user, data class, and recipient.

ADVANTAGES

Encryption Provides Security for Data at All Times

Generally, data is most vulnerable when it is being moved from one location to another. Encryption works during data transport or at rest, making it an ideal solution no matter where data is stored or how it is used. Encryption should be standard for all data stored at all times, regardless of whether or not it is deemed “important”.

Encrypted Data Maintains Integrity

Hackers don’t just steal information, they also can benefit from altering data to commit fraud. While it is possible for skilled individuals to alter encrypted data, recipients of the data will be able to detect the corruption, which allows for a quick response to the cyber-attack.

Encryption Protects Privacy

Encryption is used to protect sensitive data, including personal information for individuals. This helps to ensure anonymity and privacy, reducing opportunities for surveillance by both criminals and government agencies. Encryption technology is so powerful that some governments are attempting to put limits on the effectiveness of encryption—which does not ensure privacy for companies or individuals.

Encryption is Part of Compliance

Many industries have strict compliance requirements to help protect those whose personal information is stored by organizations. HIPAA, FIPS, and other regulations rely on security methods such as encryption to protect data, and businesses can use encryption to achieve comprehensive security.

Encryption Protects Data across Devices

Multiple (and mobile) devices are a big part of our lives, and transferring data from device to device is a risky proposition. Encryption technology can help protect store data across all devices, even during transfer. Additional security measures like advanced authentication help deter unauthorized users.

The Future of Encryption

As hackers continue to become more savvy and sophisticated, encryption technology must evolve as well. Security professionals are working on a few different exciting technological advances in the encryption field, including Elliptic Curve Cryptography (ECC), homomorphic encryption, and quantum computation.

ECC is a method of cryptography that isn't so much an improvement of the encryption method itself, but a method that allows encryption and decryption to take place much faster, without any loss of data security.

Homomorphic encryption would be a system allowing calculations on encrypted data without decrypting it. This method would allow encryption across cloud systems, and ensure greater privacy for users. As an example, a financial institution could make assessments for individuals without revealing personal information.

POSSIBLE QUESTIONS

UNIT-I

PART-A (1 MARKS)

(Q.NO 1 TO 20 Online Examination)

PART-B (2 MARKS)

1. Define RDBMS?
2. Define encrypted data
3. Define FTP
4. Define SFTP
5. Define SCP
6. Define Flat file
7. Advantages of encrypted data
8. Difference between Linux and Windows
9. Explain file databases
10. Define Bioinformatics

PART-C (8 MARKS)

1. Describe mode of data transfer?
2. Write notes on the role of relational database?
3. Write the advantages of encrypted data
4. Describe Computer Fundamentals
5. Write short notes on SFTP and SCP

Unit I Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
The speed of optical fiber than twisted pair	100 times	1000 times	1500 times	2000 times	1000 times
The actual machinery in a computer is called the	machinery	hardware	software	fleshware	machinery
The major components in computers are	joystick	CPU	pendrive	Bluetooth	CPU
IBM S/390 is a	microcomputer	laptop	mainframe	supercomputer	mainframe
Modern web browers use	XHTML	HTTL	MTML	NTML	XHTML
Computer network that spans a city	Metropolitan area network	Mail area network	Messenger area network	Merge area network	Metropolitan area network
Speed of transmission in optical fiber cable	trillion bits/sec	million bits/sec	billion bits/sec	Giga bits/sec	trillion bits/sec
Communication satellites present in space, about	20000 miles above equator	22000 miles above equator	10000 miles above equator	15000 miles above equator	22000 miles above equator
Which is the part that transmits the data from one part of the computer to another?	Bus	CPU	hard disk	software	BUS
OAN means	Open area network	Object area network	Oriented area network	Office area network	Office area network
VPN means	Virtual private network	Vertical position network	Visible private network	Vertical spectrum network	Virtual private network
Primary storage is	software	hardware	external	Internal	hardware
Speed of Wireless LAN	upto 100 Mbs/s	upto 100 kbs/s	upto 1 Gbs/s	upto 10 Mbs/s	upto 1Gbs/s
PAN means	Private area network	Personnel area network	Public area network	Private access network	Personnel area network

HTTP means	Hypertext transfer protocol	High transfer protocol	Horizontal transfer protocol	High transmit protocol	Hypertext transfer protocol
URI means	Universal resource information	Uniform Resource Identifier	Uniform twisted Identifier	Useage resource identifier	Uniform Resource Identifier
Modern web browers use	XHTML	HTTL	MTML	NTML	XHTML
The hardware component that is responsible for executing the instructions is called as	CPU	IPU	PPU	RPU	CPU
Primary storage also known as	Main memory	storage memory	Secondary memory	last memory	Main memory
The place where the programs and data are stored temporarily during processing is called as	Primary storage	Secondary Storage	Tertiary storage	Network storage	Primary storage
IRC stands for	Internet Relay Chat	Internet relative choice	Intranet Relay chat	Intranet relative choice	Internet Relay Chat
Emails, Usenet news, IRC are referred as	Internet suite	Intranet suite	Internet source	Intranet source	Internet suite
FTP means	File transfer protocol	File transfer process	File transmit program	File divert protocol	File transfer protocol
System utilities and other operating services are provides by	System support software	System support hardware	System support service	System surround service	System support software
LAN stands for	Local Area Network	Local internet network	Local wide network	Local access news	Local Area Network
WAN stands for	Wide Area Network	Local internet network	Local wide network	Local access news	Wide Area Network

URI stands for	Uniform Resource Identifier	Unified Resource Identifier	Unidentified Resource Identifier	Universal Resource Identity	Uniform Resource Identifier
URIs retrieved by	HTTP	HIIP	CPU	HTML	HTTP
Worldwide web is called as	Hypermedia database	External database	End-user database	Distributed database	Hypermedia database
RDBMS includes	Data definition language	Dot definition language	Duty definition language	Data decision language	Data definition language
RDBMS stands for	Relational DBMS	Object DBMS	Oriented DBMS	Other DBMS	Relational DBMS
ODBMS stands for	Object DBMS	Oriented DBMS	Other DBMS	Official DBMS	Object DBMS
Amino acids are	building block of carbohydrates	buiding block of vitamins	building block of minerals	building block of proteins	building block of proteins
Amino acids have	both amino and carboxyl group	both amino and keto group	amino group only	carbonyl and keto group	both amino and carboxyl group
Which of the following statements is true about a peptide bond	It is non planar	It is capable of forming a hydrogen bond	The cis configuration is favoured over the trans configuration	hydroxy group	It is capable of forming a hydrogen bond
Glycine and proline are the most abundant amino acids in the structure of	Hemoglobin	Myoglobin	Insulin	Collagen	Collagen
Which out of the following amino acids carries a net positive charge at the physiological p H	Valine	Leucine	Isoleucine	amino acids	Leucine

Which out of the following amino acids is a precursor for a mediator of allergies and inflammation	Histidine	Tyrosine	Phenyl Alanine	Tryptophan	Histidine
All of the below mentioned amino acids can participate in hydrogen bonding except one	Serine	Cysteine	Threonine	Valine	Valine
All of the following amino acids are both glucogenic as well as ketogenic except	Isoleucine	Leucine	Tyrosine	Phenyl alanine	Leucine
Which out of the following amino acid is a precursor of niacin (Vitamin)?	Tyrosine	Threonine	Tryptophan	Phenylalanine	Tryptophan
Which of the following peptides is cyclic in nature-?	Glutathione	Gramicidin	Met encephalin	Leuencephalin	Gramicidin
Which out of the followings is not a fibrous protein?	Carbonic anhydrase	Collagen	Fibrinogen	Keratin	Carbonic anhydrase
Which out of the following is not a haemo protein	Catalase	Myeloperoxidase	Glutathione peroxidase	Aconitase	Aconitase
All the below mentioned proteins are metalloproteins except	Carbonic anhydrase	Xanthine oxidase	Lactate dehydrogenase	Superoxide dismutase	Lactate dehydrogenase
Which out of the following is a peptide antibiotic	Erythromycin	Gramicidin	Ciprofloxacin	Tetracycline	Gramicidin
Which of the following amino acids is most compatible with an α-helical structure?	Tryptophan	Alanine	Leucine	Proline	Alanine

The highest concentration of cystine can be found in	Melanin	Keratin	Collagen	Myosin	Keratin
Which of the amino acids below is the uncharged derivative of an acidic amino acid?	Cystine	Tyrosine	Glutamine	Serine	Glutamine
Which of the following amino acids is sweet in taste?	Glycine	Alanine	Valine	Glutamic acid	File transfer protocol
Sulphur containg amino acids are	Cystine and Methionine	Methionine and threonine	Cysteine and threonine	Cysteine and serine	Cystine and Methionine
The 21st amino acid is	hydroxy lysine	hydroxy proline	Selenocysteine	citruline	Selenocysteine
The amino acid used in the SDS PAGE electrophoresis	Aspartic acid	Glutamic acid	Glycine	Aspartic acid and Lysine together	Glycine
Single letter code of pyrrolysine	B	J	O	U	O
The primary structure of protein represents	sequence of amino acid	Helical strucutre	3 D structure	Sub unit of protein	sequence of amino acid
Enzymes are	proteins	carbohydrates	nucleic acids	DNA molecule	Proteins
The first protein squenced by F.Sanger is	Haemoglobin	myoglobin	Insulin	myosin	Insulin
Myoglobin is a	Protein with primary structure	Protein with secondary structure	Protein with tertiary structure	Protein with quartenary structure	Protein with tertiary structure
Haemoglobin has	primary	secondary	tertiary	quarternery structure	quarternery structure
The 3 D structure of protein can be determined by	NMR	X Ray	Spectroscopy	PAGE	X-Ray
SMTP was published in the year	1980	1982	1985	1990	1982

ODBC means	open database connectivity	open direct byte connectivity	open district base connectivity	open dense base connectivity	open database connectivity
Transmission speed range (million bits per second) of coaxial cable ranges from	200 to 500	100 to 200	300 to 600	500 to 1000	200 to 500
UTP means	unshielded twisted pair	unit twisted pair	uniform twisted pair	un twisted pair	unshielded twisted pair
STP means	Shielded twisted pair	Shared twisted pair	Smart twisted pair	Small twisted pair	Shielded twisted pair

Unit II:

Introduction to Bioinformatics and Biological Databases

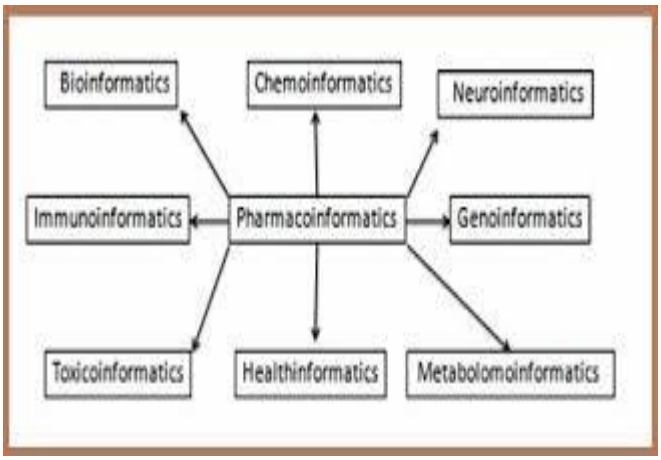
Biological databases – nucleic acid, genome, protein sequence and structure, gene expression databases, Database of metabolic pathways, Mode of data storage - File formats - FASTA, Genbank and Uniprot, Data submission & retrieval from NCBI, EMBL, DDBJ, Uniprot, PDB.

. Introduction to concepts of Bioinformatics

Bioinformatics is an interdisciplinary research area at the interface between computer science and biological science.

A variety of definitions exist in the literature and on the World Wide Web; According to **Luscombe *et al***.

- Bioinformatics is a union of biology and informatics:
- The study involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins.



Bioinformatics is the application of statistics and computer science to the field of molecular biology.

Bioinformatics differs from a related field, *computational biology*.

- **Bioinformatics** is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products and is often considered *computational molecular biology*.

- But, **computational biology** includes all biological areas that involve computation.

For example, mathematical modeling of ecosystems, population dynamics, application of the game theory in behavioral studies, and phylogenetic construction using fossil records all employ computational tools, but do not necessarily involve biological macromolecules.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.

The term ***bioinformatics*** was coined by **Paulien Hogeweg and Ben Hesper** in **1978** for the study of informatic processes in biotic systems.

Common activities in bioinformatics include

- mapping and analyzing DNA and protein sequences,
- aligning different DNA and protein sequences to compare them and
- creating and viewing 3-D models of protein structures.

Primary goal or objective of bioinformatics is

- ✓ To better understand a living cell and how it functions at the molecular level.
- ✓ To increase the understanding of biological processes.
- ✓ To analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures.
- ✓ By analyzing raw molecular sequence and structural data, bioinformatics research can generate new insights and provide a "global" perspective of the cell.
- ✓ To understand functions of a cell by analyzing sequence data because the flow of genetic information is dictated by the "central dogma" of biology in which DNA is transcribed to RNA, which is translated to proteins.
- ✓ Cellular functions are mainly performed by proteins whose capabilities are ultimately determined by their sequences. Therefore, solving functional problems using sequence and structural approaches are important.
- To focus on developing and applying computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms, and visualization).

Scope of bioinformatics:

Bioinformatics consists of **two subfields** and are complementary to each other

1. the development of computational tools and databases

2. application of these tools and databases in generating biological knowledge to better understand living systems.

1. The **tool development** includes

- writing software for sequence, structural, and functional analysis,
- as well as the construction and curating of biological databases.

2. **Application of these tools** in three areas of genomic and molecular biological research:

- molecular sequence analysis,
- molecular structural analysis, and
- molecular functional analysis.

The analyses of biological data often generate new problems and challenges that in turn spur the development of new and better computational tools.

Sequence analysis includes

- sequence alignment,
- sequence database searching,
- motif and pattern discovery,
- gene and promoter finding,
- reconstruction of evolutionary relationships,
- genome assembly and comparison.

Structural analyses includes

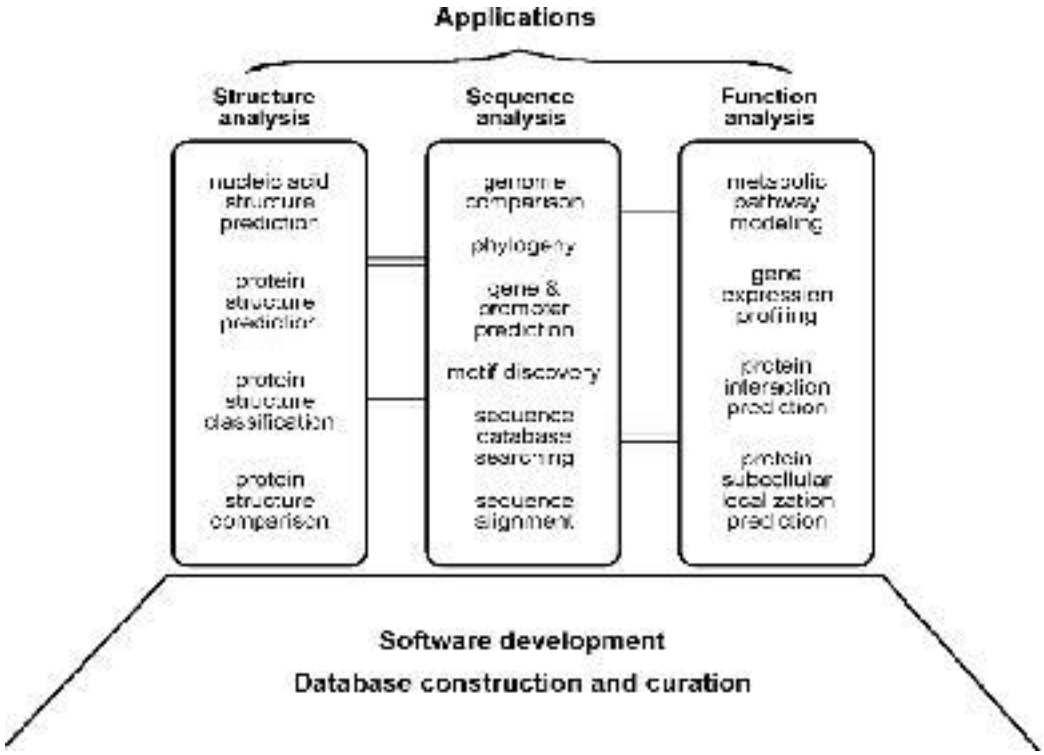
- protein and nucleic acid structure analysis,
- comparison, classification, and prediction.

Functional analyses includes

- gene expression profiling,
- protein– protein interaction prediction,
- protein subcellular localization prediction,
- metabolic pathway reconstruction, and simulation.

These three aspects of bioinformatics analysis are not isolated but often interact to produce integrated results.

Overview of various subfields of bioinformatics



Application of Bioinformatics in various Fields

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

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
- Antibiotic resistance
- Forensic analysis of microbes
- Bio-weapon creation
- Evolutionary studies
- Crop improvement
- Insect resistance
- Improve nutritional quality
- Development of Drought resistance varieties
- Vetinary Science
- Forensic DNA analysis
- Knowledge- based drug design

Major research areas of bioinformatics includes:

- sequence alignment, gene finding, genome assembly,
- protein structure alignment, protein structure prediction,
- prediction of gene expression and protein-protein interactions,
- genome-wide association studies and the modeling of evolution.
- drug design, drug discovery.

1. Sequence analysis

Sequence alignment and Sequence database

- Since the Phage Φ-X174 was sequenced in 1977, the DNA sequences of thousands of organisms have been decoded and stored in databases.
- This sequence information is analyzed to determine genes that encode polypeptides (proteins), RNA genes, regulatory sequences, structural motifs, and

repetitive sequences.

- A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees).
- Today, computer programs such as BLAST are used daily to search sequences from more than 260 000 organisms, containing over 190 billion nucleotides.
- ✓ Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.
- Another aspect of bioinformatics in sequence analysis is annotation, which involves computational gene finding to search for protein-coding genes, RNA genes, and other functional sequences within a genome.
- Bioinformatics helps to bridge the gap between genome and proteome projects — for example, in the use of DNA sequences for protein identification.

2. Genome annotation

- In the context of genomics, **annotation** is the process of marking the genes and other biological features in a DNA sequence.
- The first genome annotation software system was designed in **1995** by **Dr. Owen White**, for the first genome of a free-living organism, the bacterium *Haemophilus influenzae*. *Dr.* White built a software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes.
- ✓ Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

3. Computational evolutionary biology

Evolutionary biology is the study of the origin and descent of species, as well as

their change over time.

Bioinformatics has enabled to:

- trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,
- compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation,
- build complex computational models of populations to predict the outcome of the system over time
- track and share information on an increasingly large number of species and organisms

4. Analysis of gene expression

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including

- microarrays,
- expressed cDNA sequence tag (EST) sequencing,
- serial analysis of gene expression (SAGE) tag sequencing,
- massively parallel signature sequencing (MPSS), or
- various applications of multiplexed in-situ hybridization.

Bioinformatics have been applied to develop statistical tools for separating signal from noise in high-throughput gene expression studies.

5. Analysis of regulation

- Regulation is the complex orchestration of events starting with an extracellular signal such as a hormone and leading to an increase or decrease in the activity of one or more proteins.
- Bioinformatics techniques have been applied to explore various steps in this process. For example, promoter analysis involves the identification and study of sequence motifs in the DNA surrounding the coding region of a gene.

6. Analysis of protein expression

- Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample.

✓ Bioinformatics is very much involved in protein microarray and HT MS data.

7. Analysis of mutations in cancer

✓ In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways.

- Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer.
- Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms.
- New physical detection technologies are employed, such as oligonucleotide microarrays to identify chromosomal gains and losses (called comparative genomic hybridization), and single-nucleotide polymorphism arrays to detect known *point mutations*.
- These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate terabytes of data per experiment.

8. Comparative genomics

- The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms.

✓ A multitude of evolutionary events acting at various organizational levels shape genome evolution.

At the lowest level, point mutations affect individual nucleotides.

At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation.

- Complexity of genome evolution helps to developers of mathematical models and algorithms, based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.

9. Modeling biological systems

- Systems biology involves the use of computer simulations of cellular subsystems to both analyze and visualize the complex connections of these cellular processes (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks).
- Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

10. High-throughput image analysis

✓ Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-

content biomedical imagery.

- Modern image analysis systems helps an observer's ability to make measurements from a large or complex set of images, by improving accuracy, objectivity, or speed.
- ✓ A fully developed analysis system may completely replace the observer.

Biomedical imaging is becoming more important for both diagnostics and research. Some examples are:

- high-throughput and high-fidelity quantification and sub-cellular localization (high-content screening, cytohistopathology, Bioimage informatics)
- morphometrics
- clinical image analysis and visualization
- determining the real-time air-flow patterns in breathing lungs of living animals
- quantifying occlusion size in real-time imagery from the development of and recovery during arterial injury
- making behavioral observations from extended video recordings of laboratory animals
- infrared measurements for metabolic activity determination
- inferring clone overlaps in DNA mapping, e.g. the Sulston score

11. Structural Bioinformatic Approaches

- ✓ Protein structure prediction is another important application of bioinformatics.
- The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it.
- ✓ Knowledge of this structure is important in understanding the function of the protein.

History of Bioinformatics

The Modern bioinformatics is can be classified into **two** broad categories, Biological Science and computational Science.

However, it is the 1990s when the INTERNET arrived when the full fledged bioinformatics field was born.

Here are some of the major events in bioinformatics over the last several decades.

The events listed in the list occurred long before the term, "bioinformatics", was coined.

BioInformatics Events	
1843	Richard Owen elaborated the distinction of **homology** and **analogy.**
1961	Sidney Brenner, François Jacob, Matthew Meselson, identify messenger RNA,
1965	Margaret Dayhoff's Atlas of Protein Sequences
1970	Needleman-Wunsch algorithm
1977	DNA sequencing and software to analyze it (Staden)
1981	Smith-Waterman algorithm developed
1981	The concept of a sequence motif (Doolittle)
1982	GenBank Release 3 made public
1982	Phage lambda genome sequenced
1983	Sequence database searching algorithm (Wilbur-Lipman)
1985	FASTP/FASTN: fast sequence similarity searching
1988	National Center for Biotechnology Information (NCBI) created at NIH/NLM
1988	EMBnet network for database distribution
1990	BLAST: fast sequence similarity searching

1991	EST: expressed sequence tag sequencing
1993	Sanger Centre, Hinxton, UK
1994	EMBL European Bioinformatics Institute, Hinxton, UK
1995	First bacterial genomes completely sequenced
1996	Yeast genome completely sequenced
1997	PSI-BLAST
1998	Worm (multicellular) genome completely sequenced
1999	Fly genome completely sequenced
2000	Jeong et al. **The large-scale organization of metabolic networks.**
2000	The genome for *Pseudomonas aeruginosa* (6.3 Mbp) is published.
2000	The A. thaliana genome (100 Mb) is secquenced.
2001	The human genome (3 Giga base pairs) is published.

Milestones in bioinformatics:

Listed below are some of the major events in bioinformatics over the last several decades. Most of the events in the list occurred long before the term, "bioinformatics", was coined.

1962	Pauling's theory of molecular evolution
1965	Margaret Dayhoff's Atlas of Protein Sequences
1970	Needleman-Wunsch algorithm
1977	DNA sequencing and software to analyze it (Staden)
1981	Smith-Waterman algorithm developed
1981	The concept of a sequence motif (Doolittle)

1982	GenBank Release 3 made public
1982	Phage lambda genome sequenced
1983	Sequence database searching algorithm (Wilbur-Lipman)
1985	FASTP/FASTN: fast sequence similarity Searching
1988	National Center for Biotechnology Information (NCBI) created at NIH/NLM
1988	EMBnet network for database distribution
1990	BLAST: fast sequence similarity searching
1991	EST: expressed sequence tag sequencing
1993	Sanger Centre, Hinxton, UK
1994	EMBL European Bioinformatics Institute, Hinxton, UK
1995	First bacterial genomes completely sequenced
1996	Yeast genome completely sequenced
1997	PSI-BLAST, ***Escherichia coli*** **genome completely sequenced**
1998	Worm (multicellular) genome completely sequenced
1999	Fly genome completely sequenced
2000	First plant genome sequenced – *Arabidopsis*
2001	Draft of human genome sequence
2002	Draft of mouse genome sequence, Japanese puffer fish genome, rice genome sequence
2003	Sequence of human chromosome 14
2005	Rice genome sequence

Introduction to Biological Databases

The very first challenge in the genomics is to store and handle the accumulating volume of raw sequence data and informations through the establishment and use of computer databases.

The development of databases to handle the large amount of molecular biological data is a fundamental task of bioinformatics.

WHAT IS A DATABASE?

A *database* is a computerized archive used to store and organize data in such a way that information can be retrieved easily by different search criteria.

Databases are composed of computer hardware and software for data management.

Objective of the database development

- to organize data in a set of structured records to enable easy retrieval of information.

- Each record, also called as *entry*, should contain a number of fields that hold the actual data items, for example, fields for names, phone numbers, addresses, dates.

- To retrieve a particular record from the database, a user can specify a particular piece of information, called *value*, to be found in a particular field and expect the computer to retrieve the whole data record. This process is called *making a query.*

- Biological databases have a higher level of requirement, known as *knowledge discovery*, which refers to the identification of connections between pieces of information that were not known when the information was first entered.

For example, databases containing raw sequence information can perform extra computational tasks to identify sequence homology or conserved motifs.

- These features facilitate the discovery of new biological insights from raw data.

Types of databases:

- Originally, databases all used a flat file format, which is a long text file that contains many entries separated by a *delimiter*, a special character such as a vertical bar (|).
- Within each entry are a number of fields separated by tabs or commas.
- To facilitate the access and retrieval of data, sophisticated computer software programs for organizing, searching, and accessing data have been developed.
- They are called as ***database management systems***.
- These systems contain not only raw data records but also operational instructions to help identify hidden connections among data records.
- The purpose of establishing a data structure is for easy execution of the searches and to combine different records to form final search reports.

Depending on the types of data structures, these database management systems can be classified into two types:

1. Relational database management systems
2. Object-oriented database management systems

Databases employing these management systems are known

as Relational databases

Object-oriented databases.

Relational Databases

- Instead of using a single table as in a flat file database, relational databases use a set of tables to organize data.
- Each table, also called a *relation*, is made up of columns and rows.
- Columns represent individual fields.
- Rows represent values in the fields of records.
- The columns in a table are indexed according to a common feature called an *attribute*, so they can be cross-referenced in other tables.
- Relational databases can be created using a special programming language called *structured query language* (SQL).

Object-Oriented Databases

- Object-oriented databases have been developed that store data as objects.
- In an object-oriented programming language, an object can be considered as a unit that combines data and mathematical routines that act on the data.
- Programming languages like C++ are used to create object-oriented databases.
- The object-oriented database system is more flexible; data can be structured based on hierarchical relationships.

BIOLOGICAL DATABASES

Current biological databases use all three types of database structures:

- flat files,
- relational,
- object oriented.

Based on their contents, biological databases can be divided into **three categories**:

1. primary databases,
2. secondary databases,
3. specialized databases.

Primary databases contain

- Original biological data.
- They are archives of raw sequence or structural data submitted by the scientific community.
- Examples: GenBank and Protein Data Bank (PDB).
- There are three major public sequence databases that store raw nucleic acid sequence data produced and submitted by researchers worldwide which are freely available on internet.
  - GenBank,
  - European Molecular Biology Laboratory (EMBL) database
  - DNA Data Bank of Japan (DDBJ).
- Most of the data in the databases are contributed directly by authors with a minimal level of annotation.

- A small number of sequences, especially those published in the 1980s, were entered manually from published literature by database management staff.
- Presently, sequence submission to either GenBank, EMBL, or DDBJ is a precondition for publication in most scientific journals to ensure the fundamental molecular data to be made freely available.
- These three public databases closely collaborate and exchange new data daily.
- They together constitute the **International Nucleotide Sequence Database Collaboration.**

This means that by connecting to any one of the three databases, one should have access to the same nucleotide sequence data

- **PDB -** is a only one centralized database for the three-dimensional structures of biological macromolecules.
- This database archives atomic coordinates of macromolecules (both proteins and nucleic acids) determined by x-ray crystallography and NMR.
- It uses a flat file format to represent protein name, authors, experimental details, secondary structure, cofactors, and atomic coordinates.
- The web interface of PDB also provides viewing tools for simple image manipulation.

Secondary databases contain

- To turn the raw sequence information into more sophisticated biological knowledge, much post processing of the sequence information is needed.
- Thus secondary databases contains computationally processed sequence information derived from the primary databases.
- The amount of computational processing work varies greatly among the secondary databases;

some are simple archives of translated sequence data from identified open reading frames in DNA,

whereas others provide additional annotation and information related to higher levels of information regarding structure and functions.

Example of secondary databases is

SWISS-PROT,

- which provides detailed sequence annotation that includes structure, function, and protein family assignment.
- The sequence data are mainly derived from TrEMBL, a database of translated nucleic acid sequences stored in the EMBL database.
- The annotation of each entry is carefully curated with good quality by human experts. The protein annotation includes function, domain structure, catalytic sites, cofactor binding, post translational modification, metabolic pathway information, disease association, and similarity with other sequences.
- Much of this information is obtained from scientific literature and entered by database curators.
- The annotation provides significant added value to each original sequence record.
- The data record also provides cross referencing links to other online resources of interest. Other features such as very low redundancy and high level of integration with other primary and secondary databases make SWISS-PROT very popular among biologists.

UniProt database

- A recent effort to combine **SWISS-PROT, TrEMBL, and PIR** led to the creation of UniProt database
- It has larger coverage than any one of the three databases while at the same time maintaining the original SWISS-PROT feature of low redundancy, cross-references, and a high quality of annotation.

Pfam and Blocks databases

- contain aligned protein sequence information, motifs and patterns, which can be used for classification of protein families and inference of protein functions.

DALI database

- is a protein secondary structure database that is vital for protein structure classification and threading analysis to identify distant evolutionary relationships among proteins.

Specialized databases contain

- Serve a specific research community or focus on a particular organism.
- The content of these databases may be sequences or other types of information.
- The sequences in these databases may overlap with a primary database, but may also have new data submitted directly by authors.
- Because they are often curated by experts in the field, they may have unique organizations and additional annotations associated with the sequences.
- Many genome databases that are taxonomic specific fall within this category.
- **Examples:** Flybase, WormBase, AceDB, TAIR, GenBank EST database and Microarray Gene Expression Database.
- Some of these deal with particular classes of sequence:

RDP - the 'Ribosomal Database Project' provides ribosome related data services to the scientific community, including online data analysis, rRNA derived phylogenetic trees, and aligned and annotated rRNA sequences.

HIV-SD - the 'HIV Sequence Database' collects, curates and annotates HIV and SIV sequence data and provides various tools for analysing this data.

IMGT - the 'ImMunoGeneTics database' is a database specialising in Immunoglobulins, T cell receptors and the Major Histocompatibility Complex (MHC) of all vertebrate species.

- Others nucleotide sequence databases are focussing on particular features such as:

TRANSFAC - contains sequence information on transcription factors and transcription factor binding sites.

EPD - the 'Eukaryotic Promoter Database' is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally.

REBASE - for restriction enzymes and restriction enzyme sites.

GOBASE - is a specialised database of organelle genomes.

INFORMATION RETRIEVAL FROM BIOLOGICAL

DATABASES There are a number of retrieval systems for biological data.

The most popular retrieval systems for biological databases are

- Entrez
- Sequence Retrieval Systems (SRS)

To perform complex queries in a database often requires the use of Boolean operators. Most search engines of public biological databases use some form of this Boolean logic. This is to join a series of keywords using logical terms such as AND, OR, and NOT to indicate relationships between the keywords used in a search. *AND* means that the search result must contain both words;

OR means to search for results containing either word or both;

NOT excludes results containing either one of the words.

Parentheses ( ) to define a concept if multiple words and relationships are involved, so that the computer knows which part of the search to execute first. Items contained within parentheses are executed first.

Quotes can be used to specify a phrase.

Entrez

- The NCBI developed and maintains Entrez, a biological database retrieval system.

- It is a gateway that allows text-based searches for a wide variety of data, including annotated genetic sequence information, structural information, as well as citations and abstracts, full papers, and taxonomic data.
- The key feature of Entrez is its ability to integrate information, which comes from cross-referencing between NCBI databases based on preexisting and logical relationships between individual entries.
- This is highly convenient: users do not have to visit multiple databases located in disparate places.

For example, in a nucleotide sequence page,

one may find cross-referencing links to the translated protein sequence, genome mapping data, or to the related PubMed literature information, and to protein structures if available.

There are **several options common to all NCBI databases** that help to narrow the search.

One option is **"Limits,"** which helps to restrict the search to a subset of a particular database (e.g., the field for author or publication date) or a particular type of data (e.g., chloroplast DNA/RNA).

Another option is **"Preview/Index,"** which connects different searches with the Boolean operators and uses a string of logically connected keywords to perform a new search.

"History" option provides a record of the previous searches so that the user can review, revise, or combine the results of earlier searches.

"Clipboard" that stores search results for later viewing for a limited time.

To store information in the Clipboard, the **"Send to Clipboard"** function should be used.

One of the databases accessible from Entrez is a biomedical literature database known as **PubMed,** which contains abstracts and in some cases the full text articles from nearly 4,000 journals.

An important feature of PubMed is the retrieval of information based on medical subject headings (MeSH) terms.

The MeSH system consists of a collection of more than 20,000 controlled and standardized vocabulary terms used for indexing articles.

PubMed uses a word weight algorithm to identify related articles with similar words in the titles, abstracts, and MeSH. By using this feature, articles on the same topic that were missed in the original search can be retrieved.

GenBank

is the most complete collection of annotated nucleic acid sequence data for almost every organism.

The content includes genomic DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms. GenBank is a relational database.

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

There are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in the WGS division as of August 2009.

There is also a **GenPept database** for protein sequences, the majority of which are conceptual translations from DNA sequences, and amino acid sequences derived using peptide sequencing techniques.

There are two ways to search for sequences in GenBank.

Text-based keywords similar to a PubMed search

Molecular sequences to search by sequence similarity using BLAST.

The following are the informations obtainable from the GenBank 1. Submissions to GenBank

Many journals require submission of sequence information to a database prior to publication so that an accession number may appear in the paper. There are several options for submitting data to GenBank:

- **BankIt**, **a** WWW-based submission tool for convenient and quick submission of sequence data
- **Sequin**, **NCBI's** stand-alone submission software for MAC, PC, and UNIX platforms, is available by FTP. When using Sequin, the output files for direct submission should be sent to GenBank by e-mail.
- **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.
- **Barcode Submission Tool**, **a** WWW-based tool for the submission of GenBank sequences and trace data for Barcode of Life projects.

Currently, only mitochondrial cytochrome c oxidase subunit I (COI) genes are being accepted with this tool.

There are specialized, streamlined procedures for batch submissions of sequences, such as EST, STS, and GSS sequences.

2. Submissions of Sequence Reads

- Reads of Sanger-style sequencing can be submitted to the Trace Archive.
- Runs of next-generation sequencing, for example 453 or Solexa, can be submitted to the Short Read Archive (SRA).

3. Updating or Revising a GenBank Sequence

Revisions or updates to GenBank entries can be made by the submitters at any time and can be accepted through the Update option on the BankIt page.

4. 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),

dbEST (Expressed Sequence Tags),

dbGSS (Genome Survey Sequences).

- Search and align GenBank sequences to a query sequence using **BLAST** (Basic Local Alignment Search Tool).
- Search, link, and download sequences programatically using NCBI e-utilities.

5. 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.

Structure of GenBank - Sequence Format

- ✓ To search GenBank effectively using the text-based method requires an understanding of the GenBank sequence format.
- ✓ Search output for sequence files is produced as flat files for easy reading.
- ✓ The resulting flat files contain three sections –

Header, Features, and Sequence entry.

LOCUS	The LOCUS field contains a number of different data elements, including locus name, sequence length, molecule type, GenBank division, and modification date. Each element is described below.
• Locus Name	The locus name in this example is SCU49845. The locus name was originally designed to help group entries with similar sequences: the first three characters usually designated the organism; the fourth and fifth characters were used to show other group designations, such as gene product; for segmented entries, the last character was one of a series of sequential integers.

	the locus name is usually the first letter of the genus and species names, followed by the accession number.
• Sequence Length	Number of nucleotide base pairs (or amino acid residues) in the sequence record. Example, the sequence length is 5028 bp. There is no maximum limit for sequence size that can be submitted to GenBank. The minimum length required for submission is 50 bp.
• Molecule Type	The type of molecule that was sequenced. Example, the molecule type is DNA or RNA or protein The various molecule types are described in the Sequin documentation and can include genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA, and small cytoplasmic RNA.
• GenBank Division	The GenBank division to which a record belongs is indicated with a three letter abbreviation. In this example, GenBank division is PLN. The GenBank database is divided into 18 divisions: 1. PRI - primate sequences 2. ROD - rodent sequences 3. MAM - other mammalian sequences 4. VRT - other vertebrate sequences 5. INV - invertebrate sequences 6. PLN - plant, fungal, and algal sequences 7. BCT - bacterial sequences 8. VRL - viral sequences 9. PHG - bacteriophage sequences 10. SYN - synthetic sequences 11. UNA - unannotated sequences

	12. EST - EST sequences (expressed sequence tags) 13. PAT - patent sequences 14. STS - STS sequences (sequence tagged sites) 15. GSS - GSS sequences (genome survey sequences) 16. HTG - HTG sequences (high-throughput genomic sequences) 17. HTC - unfinished high-throughput cDNA sequencing 18. ENV - environmental sampling sequences
Modification Date	The date in the LOCUS field is the date of last modification. The sample record shown here was last modified on 21-JUN-1999. In some cases, the modification date might correspond to the release date.
DEFINITION	Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding). If the sequence has a coding region (CDS), description may be followed by a completeness qualifier, such as "complete cds".
ACCESSION	The unique identifier for a sequence record. An accession number applies to the complete record and is usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345) or two letters followed by six digits (e.g., AF123456). Accession numbers do not change, even if information in the record is changed at the author's request. Records from the RefSeq database of reference sequences have a different accession number format that begins with two letters followed by an underscore bar and six or more digits, for example: NT_123456 constructed genomic contigs

	NM_123456 mRNAs NP_123456 proteins NC_123456 chromosomes
VERSION	A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database. If there is any change to the sequence data (even a single base), the version number will be increased, e.g., U12345.1 → U12345.2, but the accession portion will remain stable. The accession.version system of sequence identifiers runs parallel to the GI number system, i.e., when any change is made to a sequence, it receives a new GI number AND an increase to its version number.
GI	"GenInfo Identifier" sequence identification number, in this case, for the nucleotide sequence. If a sequence changes in any way, a new GI number will be assigned. A separate GI number is also assigned to each protein translation within a nucleotide sequence record, and a new GI is assigned if the protein translation changes in any way (see below). GI sequence identifiers run parallel to the new accession.
KEYWORDS	Word or phrase describing the sequence. The Keywords field is present in sequence records primarily for historical reasons, and is not based on a controlled vocabulary.
SOURCE	Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type.
• Organism	The formal scientific name for the source organism (genus and species) and its lineage, based on the phylogenetic classification scheme used in the NCBI Taxonomy Database.

REFERENCE	Publications by the authors of the sequence that discuss the data reported in the record. References are automatically sorted within the record based on date of publication, showing the oldest references first. Some sequences have not been reported in papers and show a status of "unpublished" or "in press". Various classes of publication can be present in the References field, including journal article, book chapter, book, thesis/monograph, proceedings chapter, proceedings from a meeting, and patent. The last citation in the REFERENCE field usually contains information about the submitter of the sequence, rather than a literature citation. It is therefore called the "submitter block" and shows the words "Direct Submission" instead of an article title. The various subfields under References are searchable in the Entrez search fields noted below.
• AUTHORS	List of authors in the order in which they appear in the cited article. Enter author names in the form: Lastname AB (without periods after the initials).
• TITLE	Title of the published work or tentative title of an unpublished work. Sometimes the words "Direct Submission" instead of an article title.
JOURNAL	MEDLINE abbreviation of the journal name. (Full spellings can be obtained from the Entrez Journals Database.)
• PUBMED	PubMed Identifier (PMID). References that include PubMed IDs contain links from the sequence record to the corresponding PubMed record.
• Direct	Contact information of the submitter, such as institute/department and

Submission postal address.	
FEATURES	Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features.
source	Mandatory feature in each record that summarizes the length of the sequence, scientific name of the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone, tissue type, etc., if provided by submitter.
Taxon	A stable unique identification number for the taxon of the source oganism. A taxonomy ID number is assigned to each taxon (species, genus, family, etc.) in the NCBI Taxonomy Database.
CDS	Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons). The CDS feature includes an amino acid translation. Submitters are also encouraged to annotate the mRNA feature, which includes the 5' untranslated region (5'UTR), coding sequences (CDS, exon), and 3' untranslated region (3'UTR).
<1..206	Base span of the biological feature indicated to the left, in this case, a CDS feature. (The CDS feature is described above, and its base span includes the start and stop codons.) Features can be complete, partial on the 5' end, partial on the 3' end, and/or on the complementary strand. Examples: 1. complete feature is simply written as *n..m* Example: 687..3158 The feature extends from base 687 through base 3158 in the sequence shown.

	2. < indicates partial on the 5' end Example: <1..206 The feature extends from base 1 through base 206 in the sequence shown, and is partial on the 5' end 3. > indicates partial on the 3' end Example: 4821..5028> The feature extends from base 4821 through base 5028 and is partial on the 3' end. 4. (complement) indicates that the feature is on the complementary strand Example: complement(3300..4037) The feature extends from base 3300 through base 4037 but is actually on the complementary strand.
protein_id	A protein sequence identification number, similar to the Version number of a nucleotide sequence. Protein IDs consist of three letters followed by five digits, a dot, and a version number. If there is any change to the sequence data (even a single amino acid), the version number will be increased, but the accession portion will remain stable (e.g., AAA98665.1 will change to AAA98665.2).
GI	"GenInfo Identifier" sequence identification number, in this case, for the protein translation. The GI system of sequence identifiers runs parallel to the accession.version system, which was implemented by GenBank, EMBL, and DDBJ in February 1999.
translation	The amino acid translation corresponding to the nucleotide coding sequence (CDS). In many cases, the translations are conceptual and can

	indicate whether the CDS is based on experimental or non-experimental evidence.
• gene	A region of biological interest identified as a gene and for which a name has been assigned. The base span for the gene feature is dependent on the furthest 5' and 3' features.
complement	Indicates that the feature is located on the complementary strand.
• Other Features	Examples of other records that show a variety of biological features; a graphic format is also available for each sequence record and visually represents the annotated features: • AF165912 (gene, promoter, TATA signal, mRNA, 5'UTR, CDS, 3'UTR) GenBank flat file • AF090832 (protein bind, gene, 5'UTR, mRNA, CDS, 3'UTR) GenBank flat file • L00727 (alternatively spliced mRNAs) GenBank flat file A complete list of features is available from the resources noted above.
ORIGIN	The ORIGIN may be left blank, may appear as "Unreported," or may give a local pointer to the sequence start, usually involving an experimentally determined restriction cleavage site or the genetic locus (if available). The sequence data begin on the line immediately below ORIGIN. To view/save the sequence data only, display the record in FASTA format.

Protein sequence database: -

Primary protein sequence databases:

1. PIR - Protein Information Resource
2. MIPS – Martinsried Institute for protein sequences
3. SWISS – PROT
4. TrEMBL – Translated EMBL

5. NRDB – Non-Redundant Database

6. OWL

7. SWISS-PROT + TrEMBL

Secondary protein sequence databases:

1. PROSITE

2. Profiles

3. PRINTS

4. Pfam

5. BLOCKS

6. IDENTIFY

Primary protein sequence databases:

1. PIR –

- Developed by Margaret Dayhoff during 1960s at National Biomedical Research Foundation (NBRF).
- Maintained by PIR-International consortium
- This consortium includes

- Protein Information Resource (PIR) at NBRF

- International Protein Information Database of Japan (JIPID)

- Martinsried Institute for Protein Sequences (MIPS)

- Based on data quality and annotation level, PIR database divided into four sections:

1. PIR 1 – contains fully classified and annotated entries

2. PIR 2 – contains preliminary entries, not completely reviewed and may contain redundancy

3. PIR 3 – contains unverified entries

4. PIR 4 - four catagories

(i) Conceptual translations of artefactual sequences

(ii) Conceptual translations of sequences that are not transcribed or translated

(iii) Conceptual translations of genetically engineered

(iv)sequences that are not genetically encoded and not produced on ribosomes.

SWISS-PROT

- established by the Department of Medical Biochemistry at the University of Geneva and EMBL during 1986.
- Now this database is maintained collaboratively by SIB (Swiss Institute of Bioinformatics) and EBI/EMBL.
- Minimally redundant database
- Interlinked to many other resources.
- This database provides high-level annotations including the descriptions of function of protein, structure of its domains, its post-translational modifications, varients etc.
- Now contains ..... entries from more than ..... different species.

Structure of SWISS –PROT

The quality of annotations and structure of database made SWISS-PROT as choice of most research purposes than the other databases.

- Each entry in this database consists of following:
- Each line is flagged with a two-letter code – which helps to present the information in a structured way.
- Entries begins with an Identification line (ID)

Ends with a // terminator.

ID line – informs the entry name, length of the protein name.

Contains ID code – designed to be informative and people-friendly in the form of PROTEIN_SOURCE – indicates the organism name, PROTEIN part of the code denotes the type of protein.

AC line – denotes Accession number – remain static between database releases.

DT line – provide information about the date of entry of the sequence to the database And details of when it was last modified.

DE line – informs the name by which the protein is known.

GN line – gives the gene name

OS line – organism species name

OC line – Organism classification within the biological kingdoms.

The next section of the database provides a list of supporting references.

Following the **references comment lines** (CC) are present and divided into themes which tells about

- FUNCTION of protein

- Its post-transcriptional modifications (PTM),

- Its TISSUE SPECIFICITY, SUB CELLULAR LOCATION.

Database **cross-reference lines (DR)** follow the comment field – provides links to other databases including primary sequence, secondary databases, specialized databases etc.

KEYWORD line (KW) - provides the keyword related the entries

Feature Table line (FT) – highlights regions of interest in the sequences, including local secondary structure (transmembrane domains), Ligand binding sites, post-translational modifications.

The final section of the database entry includes

SQ line (SQ) – sequence information in single letter amino acid code,
each line contains 60 residues.

Sequence data in SWISS-PROT, contains precursor form of protein, therefore informations related to size, molecular weight, region of signal sequence (SIGNAL), transit (TRANSIT) or pro-peptide (PROPEP) respectively. The keys CHAIN and PEPTIDE are used to denote the location of the mature form.

Secondary protein databases:

- These secondary protein sequence databases have become important tools for identifying distant relationships in novel sequences and for inferring protein function.

- These databases have developed by using signature-recognition methods to address different sequence analysis problems, resulting in rather different and independent databases.
- To perform a comprehensive analysis, a user therefore has to know several important things. For example,

- what are the resources and where can they be found?

- What is the difference between them in terms of diagnostic performance and family coverage?

- What do the different search outputs mean?

- Is it sufficient to use just one of the databases, and if so, which one?

The sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall sequence alignment,

but it can be identified by the occurrence in its sequence of a particular cluster of residue types which is commonly known as a pattern, motif, signature, or fingerprint. These motifs arise because of particular requirements on the structure of specific region(s) of a protein, which may be important, for example, for their binding properties or for their enzymatic activity.

There are a few databases available, which use different methodology and a varying degree of biological information on the characterised protein families, domains and sites.

A brief description of some of specialised protein sequence databases:

Secondary database	Primary source	Stored information
PROSITE	SWISS-PROT	Regular expressions (patterns)
Profiles	SWISS-PROT	Weighted matrics (profiles)
PRINTS	OWL	Aligned motifs (fingerprints)
Pfam	SWISS-PROT	Hidden Markov Models (HMMs)
BLOCKS	PROSITE/PRINTS	Aligned motifs (Blocks)

IDENTIFY	BLOCKS/PRINTS	Fuzzy regular expressions (patterns)

Examples of secondary protein databases include:

- **PROSITE** –

First secondary database developed and maintained by Swiss Institute of Bioinformatics.

is the extensive documentation on many protein families, as defined by sequence domains or motifs.

PROSITE contains biologically significant sites and patterns using computational tools and it can rapidly and reliably identify to which family of proteins the new sequence belongs.

The profile structure used in PROSITE is similar to but slightly more general than the one introduced by Gribskov and co-workers (Gribskov et al.,1987). Generalised profiles are remarkably similar to the specific type of Hidden Markov Models (HMMs) used in Pfam.

Structure of PROSITE:

ID (IDentification) line - is always the first line of an entry.

The general form of the ID line is:

ID ENTRY_NAME; ENTRY_TYPE.

The first item on the ID line is the entry name. This name is a useful means of identifying an entry.

The entry name consists of from 2 to 21 uppercase alphanumeric characters. The characters that are allowed in an entry name are: A-Z, 0-9, and the underscore character "_".

The second item on the ID line indicates the type of PROSITE entry. Currently this can be one the following:

PATTERN

MATRIX

RULE

AC (ACcession number) line –

- It is always the second line of an entry.
- lists the accession number associated with an entry.
- Accession numbers provide a stable way of identifying entries from release to release.
- Accession numbers allow unambiguous citation of database entries.
- Researchers who wish to cite a PROSITE entry in their publications should always cite the accession number of that entry in order to ensure that readers can find the relevant data in a subsequent release.

The format of the AC line is:

AC PSnnnnn;

Where 'PS' stands for PROSITE and 'nnnnn' is a five digit number.

DT (DaTe) line –

- It is always the third line of an entry.
- shows the date of entry or last modification of the entry.

The format of the DT line is:

DT MMM-YYYY (CREATED); MMM-YYYY (DATA UPDATE); MMM-YYYY (INFO UPDATE).

where:

MMM is the month and YYYY the year.

First date indicates when the entry first appeared in the database.

Second date indicates when the 'primary' data of the entry was last modified.

Third date indicates when any data other then the 'primary' data has been modified.

Example:

DT APR-1990 (CREATED); JUL-1990 (DATA UPDATE); JUL-1998 (INFO UPDATE).

DE (DEscription) line -

- It is always the fourth line of an entry.

- provides descriptive information about the content of the entry.

The format of the DE line is:

DE Description.

The description is given in ordinary English and is free-format. Examples:

DE Myb DNA-binding domain repeat signature 1.

PA (PAttern) lines –

- contains the definition of a PROSITE pattern.

The patterns are described using the following conventions:

- The standard IUPAC one-letter codes for the amino acids are used.

- The symbol 'x' is used for a position where any amino acid is accepted.

- Ambiguities are indicated by listing the acceptable amino acids for a given position, between square parentheses '[ ]'. For example: [ALT] stands for Ala or Leu or Thr.

- Ambiguities are also indicated by listing between a pair of curly brackets '{ }' the amino acids that are not accepted at a given position.

For example: {AM} stands for any amino acid except Ala and Met.

- Each element in a pattern is separated from its neighbor by a '-'.

- Repetition of an element of the pattern can be indicated by following that element with a numerical value or a numerical range between parenthesis. Examples: x(3) corresponds to x-x-x, x(2,4) corresponds to x-x or x-x-x or x-x-x-x.

- When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a '<' symbol or respectively ends with a '>' symbol. In some rare cases (e.g. PS00267 or PS00539), '>' can also occur inside square brackets for the C-terminal element. 'F-[GSTV]-P-R-L-[G>]' means that either 'F-[GSTV]-P-R-L-G' or 'F-[GSTV]-P-R-L>' are considered.

A period ends the pattern.

Examples:

PA [AC]-x-V-x(4)-{ED}.

This pattern is translated as: [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

PA <A-x-[ST](2)-x(0,1)-V.

This pattern, which must be in the N-terminal of the sequence ('<'), is translated as: Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val

MA (MAtrix) lines –

contain the definition of a PROSITE profile (or matrix) entry.

PP line -

PROSITE profiles normally use two cut-off levels,

a reliable cut-off (LEVEL=0) and

a low confidence cut-off (LEVEL=-1).

The low level cut-off usually covers the twilight zone where few true positives, that cannot be separated from false positives, might be present.

The output of the *pfsearch* and the *pfscan* programs indicate strong matches (level 0) with '!' and weak matches (level -1) with '?'.

This specific tagging in the match list can be used in post-processing, to validate some true positives present in the twilight zone or to eliminate some false positives detected with significant score.

NR (Numerical Results) lines –

contain information relevant to the results of the scan with a pattern on the complete Swiss-Prot knowledgebase.

The format of the NR line is:

NR /QUALIFIER=data; /QUALIFIER=data; .....

The qualifiers that are currently defined are:

/RELEASE	Swiss-Prot release number and total number of sequence entries in that release.
/TOTAL	Total number of hits in Swiss-Prot.
/POSITIVE	Number of hits on proteins that are known to belong to the set in consideration.
/UNKNOWN	Number of hits on proteins that could possibly belong to the set in consideration.
/FALSE_POS	Number of false hits (on unrelated proteins).
/FALSE_NEG	Number of known missed hits.
/PARTIAL	Number of partial sequences which belong to the set in consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences.

CC (Comments) lines –

contains various types of comments.

The format of the CC line is:

CC /QUALIFIER=data; /QUALIFIER=data; .....

The qualifiers that are currently defined are:

/TAXO_RANGE Taxonomic range.

/MAX-REPEAT	Maximum known number of repetitions of the pattern or profile in a single protein.
/SITE	Indication of an `interesting' site in a pattern.
/SKIP-FLAG	Indication of an entry that can be, in some cases, ignored by a program (because it is too unspecific).
/VERSION	The version number of a pattern or a profile.

There are 5 qualifiers specific to profile entries:

/MATRIX_TYPE	Describes the region of the protein identified by the profile.
/SCALING_DB	Scaling database used to calibrate the profile.
/AUTHOR	Author of the profile.
/FT_KEY	Feature key to describe the region covered by the profile.
/FT_DESC	Feature description of the region covered by the profile.

- **PRINTS** - A different approach to pattern recognition, termed "fingerprinting" is used by this database. Within a sequence alignment, it is usual to find several motifs that characterise the aligned family. The ability to tolerate mismatches, both at the level of residues within individual motifs, and at the level of motifs within the fingerprint as a whole, renders fingerprinting a powerful diagnostic technique.
- **Pfam** - Another important secondary protein database is Pfam. The methodology used by Pfam to create protein family or domain signatures is Hidden Markov Models (HMMs). HMMs are closely related to profiles, but are based on probability theory methods. These allow a direct statistical approach to identifying and scoring matches, and also to combining information from a multiple alignment with prior knowledge. These databases are useful for analysing multidomain proteins. The biggest drawback of Pfam is its lack of biological information (annotation) of the protein families.

- **BLOCKS** - Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks for the Blocks Database are made automatically by looking for the most highly conserved regions in groups of proteins documented in InterPro.

- **SBASE** - This is a protein domain library sequences database that contains annotated structural, functional, ligand-binding and topogenic segments of proteins, cross-referenced to all major sequence databases and sequence pattern collections.

Structural databases:

Structure database is a database that is modeled around the various experimentally determined macromolecular structures.

- This Database contains Structures of Protein, DNA, and RNA Molecules. Most coordinates were obtained from X-Ray or NMR studies. The Database is maintained at the Brookhaven National Laboratory.
- The aim of most protein structure databases is to organize and annotate the protein structures, providing the biological community access to the experimental data in a useful way.
- The number of known protein structures is increasing very rapidly and these are available through the Protein Data Bank (PDB).
- The Nucleic Acid Database (NDB) is the database for structural information about nucleic acid molecules.
- The Cambridge Crystallographic Data Centre (CCDC) provides a database of structures of 'small molecules', of interest to biologists concerned with protein-ligand interactions.

Examples of MACROMOLECULAR 3D STRUCTURE DATABASES

Protein Data Bank (PDB):

- The Protein Data Bank (PDB) was established in 1971 as the central archive of all experimentally determined protein structure data.
- Today the PDB is maintained by an international consortia collectively known as the Worldwide Protein Data Bank (wwPDB).
- Aim of the wwPDB is to maintain a single archive of macromolecular structural data that is freely and publicly available to the global community.

RCSB PDB : (http://www.rcsb.org/pdb/home/)

- The RCSB PDB contains 3-D biological macromolecular structure data from X-ray crystallography, NMR, and Cryo-EM.
- It is operated by Rutgers, The State University of New Jersey and the San Diego Supercomputer Center at the University of California, San Diego.
- The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

MMDB: http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

- NCBI's structure database is called MMDB (Molecular Modeling DataBase),
- it is a subset of three-dimensional structures obtained from the Protein Data Bank (PDB) excluding theoretical models.
- MMDB is a database of ASN.1-formatted records.
- It was designed for flexibility, and as such, is capable of archiving conventional structural data as well as future descriptions of biomolecules, such as those generated by electron microscopy (surface models).

EBI structure databases: (http://www.ebi.ac.uk/Databases/structure.html)

1) **MSD (macromolecular structure databases)**:

The Macromolecular Structure Database is a European project for the collection, management and distribution of data about macromolecular structures. It is responsible for the deposition and validation of new protein structures. It includes PDB search tools.

2) **CSA(Catalytic Site Atlas)**:

The Catalytic Site Atlas is a resource of catalytic sites and residues identified in enzymes using structural data.

3) **DSSP**:

The DSSP database is a database of secondary structure assignments (and much more) for all of the entries in the Protein Data Bank (PDB).

4) **HSSP(homology-derived structures of proteins)**:

HSSP is a derived database merging structural (2-Dimensional and 3-Dimensional) and sequence information (1-Dimensional).

5) **PDBsum**:

PDBsum is a pictorial database providing an at-a-glance overview of every macromolecular structure (nucleic acids and proteins) deposited in the Protein Data Bank (PDB).

PSdB: (http://www.daviddeerfield.com/PSdb/)

- The Protein Structure Database (PSdb), is a protein database, derived from the information available in the Protein Databank and NRL-3D database,
- It relates secondary (e.g. Helix, Sheet, Turn, Random Coil) and tertiary information (e.g. Solvent accessibility, internal relative distances, and ligand interactions) to the primary structure.

CATH: (http://www.cathdb.info/)

- CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels :

Class (C), Architecture (A), Topology (T) and Homologous superfamily (H).

- The boundaries and assignments for each protein domain are determined using a combination of automated and manual procedures which include computational techniques, empirical and statistical evidence, literature review and expert analysis.

SCOP: (http://Scopes-lmb.cam.ac.uk/scop/)

- The SCOP database, created by manual inspection and by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- It provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.

SWISS-3D IMAGE: (http://expasy.org/sw3d/)

- It is an image database which strives to provide high quality pictures of biological macromolecules with known three-dimensional structure.
- The database contains mostly images of experimentally elucidated structures,
- but also provides views of well accepted theoretical protein models.

SWISS-MODEL: (http://swissmodell.expasy.org//SWISSMODEL.html)

- SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer).

ModBase: (http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi)

- It is a database of annotated comparative protein structure models, and associated resources. MODBASE contains theoretically calculated models, not experimentally determined structures.

Bibliographic databases

- Services that produced abstracts of scientific literature began to make their data available in machine-readable form in the early 1960's.
- A **bibliographic database** is a database of bibliographic records, an organized digital collection of references to published literature, including journal and newspaper articles, conference proceedings, reports, government and legal publications, patents, books, etc.
- In contrast to library catalogue entries, a large proportion of the bibliographic records in bibliographic databases describe analytics (articles, conference papers, etc.) rather than complete monographs,
- and they generally contain subject descriptions in the form of keywords, subject classification terms, or abstracts.

- A bibliographic database may be general in scope or cover a specific academic discipline.
- A significant number of bibliographic databases are still proprietary, available by licensing agreement from vendors, or directly from the abstracting and indexing services that create them.
- Many bibliographic databases evolve into digital libraries, providing the full-text of the indexed contents.
- Others converge with non-bibliographic scholarly databases to create more complete disciplinary search engine systems, such as Chemical Abstracts or Entrez.

➢ The tools that enable scientificaly coherent and efficient use of the resources described below are an important activity in the field of bioinformatics called Text Mining, which is described by Dr. Dietrich Rebholz-Schuhmann, from the EBI.

The best known bibliographic databases:

1. MEDLINE - accessible through EBI's SRS.

2. PUBMED -accessible through NCBI's ENTREZ.

EMBASE is a commercial product for the medical literature.

BIOSIS, the inheritor of the old Biological Abstracts, covers a broad biological field; the Zoological Record indexes the zoological literature.

CAB International maintains abstract databases in the fields of agriculture and parasitic diseases.

AGRICOLA is for the agricultural field what MEDLINE is for the medical field.

The bibliographical databases are with the exception of MEDLINE/PUBMED only available through commercial database vendors.

Organism specific databases:

- There are countless organism-specific databases present.

Below list includes only organisms that are of direct interest to researchers.

Virus

organism	database	description
Virus	VIDA	Organizes open reading frames ORFs) from viral genomic sequences into virus-specific homologous protein families.

Bacteria and Archaea

organism	database	description
Microbial	CMR	TheComprehensiveMicrobialResourcedisplays information on all of the publicly available, complete prokaryotic genomes.
Escherichia coli	EcoCyc	A database for the bacterium Escherichia coli K-12 MG1655.
Escherichia coli	EcoGene	Contains updated information about the E. coli K-12 genome and proteome sequences, including extensive gene bibliographies. A major EcoGene focus has been the re-evaluation of translation start sites.
Bacillus subtilis	SubtiList	A reference database for the Bacillus subtilis genome.
Bacillus subtilis	DBTBS	A database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information.

Protists

organism	database	description
The Plasmodium Genome Resource hosts genomic and *proteomic* data	*Plasmodium* PlasmoDB	(and more) for different species of the parasitic eukaryote Plasmodium. It brings together data

		provided by numerous laboratories worldwide, and adds its own data analysis.
Plasmodium falciparum	GeneDB P.falcipatum whole range	The GeneDB is a project of the Sanger Institute Pathogen Sequencing Unit's and aims to provide reliable access to the latest sequence data and annotation/curation for the of organisms sequenced by the Unit.
Tetrahymena thermophila	TGD	Provides information on the genome, genes, and proteins of Tetrahymena collected from the scientific literature, research community, and many other sources.

Fungi

organism	database	Description
Saccharomyces cerevisiae	SGD	Saccharomyces Genome Database is a scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae, which is commonly known as baker's or budding yeast.
Schizosaccharomyces pombe(*fission yeast*)	S.pombe GeneDB	Contains all S. pombe known and predicted protein coding genes, pseudogenes, transposons, tRNAs, rRNAs, snRNAs, snoRNAs and other known and predicted non-coding RNAs.
Neirospora crassa	MNCDB	The MIPS Neurospora crassa Genome Database aims to present information on the molecular structure and functional network of the entirely sequenced, filamentous fungus Neurospora crassa.

Animals - Invertebrates

organism	database	description
Drosophila	FlyBase	A comprehensive database for information on the

melanogaster		genetics and molecular biology of Drosophila. It includes data from the Drosophila Genome Projects and data curated from the literature.
Caenorhabditis elegans	Wormbase	Repository of mapping, sequencing and phenotypic information about the C. elegans and some related nematodes.

Animals - Vertebrates

organism	**database**	**description**
Homo sapiens	GDB	The Human Genome Database, GDB is the official central repository for genomic mapping data resulting from the Human Genome Initiative. Holds data on human gene loci, polymorphisms, mutations,probes, genetic maps, GenBank, citations and contacts.
Homo sapiens	HPRD	Human Protein Reference Database - is a comprehensive collection of protein features, post-translational modifications (PTMs, protein-protein interactions and disease association for each protein in the human proteome.
Homo sapiens	mtDB	Human Mitochondrial Genome Database provides a comprehensive database of complete human mitochondrial genomes.
Mus musculus	MGI	Mouse Genome Informatics provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse.
Rattus	RGD	The Rat Genome Database curates and integrates rat genetic and genomic data and provides access to this data to support research using the rat as a genetic

		model for the study of human disease.

Plants

organism	database	description
Arabidopsis thaliana	TAIR	The Arabidopsis Information Resource maintains a database of genetic and molecular biology data that includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.
Arabidopsis thaliana	MATDB	MIPS Arabidopsis thaliana database is the www access to data of Arabidopsis sequences and annotation produced by the Arabidopsis Genome Initiative, plus the mitochondrial and chloroplast genomes.

Tree of Life

database	description
Tree of Life	Provides identification keys, figures, phylogenetic trees, and other systematic information for a group of organisms; and provides information about the evolutionary history and characteristics of creatures, from frogs and flowers to dinosaurs and protists. The project presents the evolutionary tree of life as an integrated whole.

POSSIBLE QUESTIONS

UNIT-I

PART-A (1 MARKS)

(Q.NO 1 TO 20 Online Examination)

PART-B (2 MARKS)

1. Define a database.
2. What is database management system?
3. Differentiate primary and secondary databases?
4. What is a relational database?
5. Expand PDB and NCBI.
6. What is Uniprot database? What is its special feature?
7. What is Entrez?
8. Name any two options to submit data to Genbank?
9. Define Accession number? What is its role in sequence retrieval?
10. Expand PIR and MIPS.
11. What are the categories of PIR?
12. What are structural databases?
13. What is the difference between PDB and NDB?
14. Differentiate SCOP and CATH?
15. Define bibliographic database?
16. Give two examples of specialized database?
17. What is ModBase?

18. Expand MMDB? What is its application?
19. What is FASTA?
20. Explain Data retrieval

PART-C (8 MARKS)

1. Describe in detail the classification of biological databases based on their contents?
2. Write notes on the different methods of information retrieval from biological databases?
3. Write notes on primary protein sequence databases?
4. What is secondary protein Sequence database? Describe with examples.
5. Give an account on 3D structure databases?
6. Write notes on Bibliographic databases?

Unit II Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
The computerized language used to store and organize data is called as	Database	Databasic	Data store	Warehouse	Database
Databases are compose of	Hardware and software	Information	Value	Knowledge	Information
Databases enable the retrieval of	Information	Internal	External	Resource	Information
The original biological containing database is called as	Primary database	Secondary database	Tertiary database	Structure database	Primary database
Expansion of EMBL is	European Molecular Biology Laboratory	DNA Data Bank of Japan	INSDC	Protein information resource	European Molecular Biology Laboratory
DDBJ means	DNA Data Bank of Japan	European Molecular Biology Laboratory	INSDC	Protein information resource	DNA Data Bank of Japan
Three primary abase together constitutes	INSDC	SWISS-PROT	TrEMBL	PIR	INSDC
INSDC stands for	International Nucleotide Sequence Database Collaboration	International Protein Sequence Database Collaboration	International Nucleotide Structure Database Collaboration	Internal protein Sequence Database Collaboration	International Nucleotide Sequence Database Collaboration
The secondary abase which provides	SWISS-PROT	TrEMBL	PIR	INSDC	SWISS-PROT

uence annotation is					
TrEMBL contains	translated nucleic acid database	truncated nucleic acid database	translated protein database	truncated protein database	translated nucleic acid database
PIR stands for	Protein information resource	SWISS-PROT	TrEMBL	Primary Internal resource	Protein information resource
The primary source of SITE is	SWISS-PROT	UNI-PROT	OWL	PROSITE	SWISS-PROT
The primary source of ile is	SWISS-PROT	UNI-PROT	OWL	PROSITE	SWISS-PROT
The primary source of NTS is	OWL	UNI-PROT	SWISS-PROT	PROSITE	OWL
The primary source of CKS is	PROSITE	UNI-PROT	OWL	BLOCKS	PROSITE
The primary source of NTIFY is	BLOCKS	UNI-PROT	OWL	PROSITE	BLOCKS
The database maintains abstracts of agriculture parasitic diseases is	CAB International	MEDLINE	AGRICOLA	EMBASE	CAB International
Agricultural field ographic database is	AGRICOLA	MEDLINE	BIOSIS	EMBASE	AGRICOLA
Medical field ographic database is	MEDLINE	BIOSIS	AGRICOLA	EMBASE	MEDLINE
Publically available ographic database is	PUBMED	MEDLINE	AGRICOLA	EMBASE	PUBMED
The database contains nformations of specific nisms is	Organism specific database	Bibliographic database	Biology database	Protein database	Organism specific database
TAIR database contains nformations of	Arabidopsis	Rice	Maize	Sorghum	Arabidopsis
The links for NCBI is _____	WWW.ncbi.co.in	www.ncbi.nih.gov	www.ncbi.com	www.ncbi.nih.gov	www.ncbi.nih.gov

The links for EMBL is _____	WWW. Embl.com	WWW.ebi.uk	WWW.ebi-ac.uk/embl	http://WWW.embl.ac.uk	WWW.ebi-ac.uk/embl
The links for DDBJ is _____	WWW.ddbj.nig.ac.jp	www.ddbj.com	www.ddbj.ac.in	www.ddbj.nlm.com	WWW.ddbj.nig.ac.jp
The link for SWISS T is _____	www.swissprot.com	www.expansy.org.ncbi	www.swiss.ac.in	www.expansy.com	www.expansy.org.ncbi
Print database is erwise known as _____	Nucleotide database	Pattern Database	protein database	Structural database	Pattern Database
The software programs asy access and retrieval ata is called as	Database management system	Knowledge management system	Database maintainence service	Data managing system	Database management system
Type of database uses f tables to organize the is called as	Relational database	Operational database	Object database	Resource databae	Relational database
The language programm to create relational base called as	Structured query language	Sequence query language	Secondary quary language	Stored query language	Structured query language
Type of database eloped to store data as ct is called as	Object-oriented database	Relational database	Operational database	Oriental database	Object-oriented database
The language program to create object-oriented base is	C++	UNIX	WINDOWS	EXCEL	C++
PDB stands for	Protein Data Bank	Structure database	Nucleic acid Data Bank	Cambridge Crystallographic Data Centre	Protein Data Bank
NDB stands for	Nucleic acid Data Bank	Protein Data Bank	Structure database	Cambridge Crystallographic Data Centre	Nucleic acid Data Bank
The database of small	CCDC	NDB	PDB	EMBASE	CCDC

ecule structure					
CCDC stands for	Cambridge Crystallographic Data Centre	Nucleic acid Data Bank	Protein Data Bank	Structure database	Cambridge Crystallographic Data Centre
CCDC database contains	Protein-Ligand Interactions	Protein-Protein Interactions	Protein-Lipid Interactions	Protein-DNA Interactions	Protein-Ligand Interactions
SWISS-PROT was established by	University of Geneva	University of Florida	University of Singapore	University of Delhi	University of Geneva
SWISS-PROT was maintained by	SIB	PIR	NCBI	EMBL	SIB
SWISS-PROT was established during	1986	1996	2006	2011	1986
NCBI's stand alone sequence submission software is	Sequin	BankIt	tbl2asn	Barcode submission tool	Sequin
The database combines	SWISS-PROT, TrEMBL, PIR	EMBL	NCBI	EBI	SWISS-PROT, TrEMBL, PIR
GSS stands for	Genome Survey Sequences	Barcode submission tool	Expressed sequence tags	Protein information resource	Genome Survey Sequences
Rice genome sequences released at	1990	1995	2000	2005	2005
EMBL created at	1990	1992	1994	1996	`1994
Commercial product for the medical literature is	EMBASE	BIOSIS	AGRICOLA	MEDLINE	EMBASE
The inheritor of the old Biological Abstracts database is	BIOSIS	MEDLINE	AGRICOLA	EMBASE	BIOSIS
PDB is the major iratory of --------- tures	DNA	RNA	Protein	cDNA	Protein
Which electrophoresis d in proteome	1D gel	2D gel	3Dgel	4 D gel	2 D

abases					
Protein sequence ermines ……	genetic variation	genetic disorders	protein structure	Sequence similarity	Protein structure
Taxonomy database deals with	Genetic sequence	Protein structure	DNA structure	Taxonomy of organisms	Taxonomy of organisms
HIV database deals	HIV virus	STD	Virus	infectious diseases	HIV virus
GOLD is a	Protein database	Genome database	Domain	Motif	Genome database
Which of the wing is a characterstics atabase	Expensive	Complexity	unidentity	Specificity	Specificity
Proteomics is the study of	study of mRNA	study of cDNA	study of siRNA	study of proteins	study of proteins
Genomics is the study	DNA c	RNA	Gene	Protein	Gene
The database contains rimentally determined romolecular structures is	Structure database	PDB	NDB	CCDC	Structure database

Unit III – Sequence Alignments, Phylogeny and Phylogenetic trees

Local and Global Sequence alignment, pairwise and multiple sequence alignment. Scoring an alignment, scoring matrices, PAM & BLOSUM series of matrices. Types of phylogenetic trees, Different approaches of phylogenetic tree construction - UPGMA, Neighbour joining, Maximum Parsomony, Maximum likelihood.

The objective of a sequence alignment is, usually, to align the homologous positions of the two sequences. The homologous positions are the ones that come from the same position in the ancestral sequence. We don't know the ancestral sequence, so we won't be completely sure that we have succeeded. Another complementary objective could be to align protein regions that have the same structure or function.

Aligning similar sequences by any algorithm usually creates alignments that are usually correct, but when sequences are very different aligning them could be a challenge. Once a long time has passed since the split of the species the sequences can be so changed by the mutations that any meaningful similarities could have been lost an creating a meaningful alignment could be very difficult.

Sequence alignment

- It is an important first step toward structural and functional analysis of newly determined sequences.

  As new biological sequences are being generated at exponential rates, sequence comparison is becoming increasingly important to draw functional and evolutionary inference of a new protein with proteins already existing in the database.

- Sequence comparison lies at the heart of bioinformatics analysis.

The most fundamental process in this type of comparison is sequence alignment.

- This is the process by which sequences are compared by searching for common character patterns and establishing residue–residue correspondence among related sequences.

Pairwise sequence alignment

- is the process of aligning two sequences
- is the basis of database similarity searching and multiple sequence alignment.

EVOLUTIONARY BASIS of sequence alignment

- DNA and proteins are products of evolution.
- The building blocks of these biological macromolecules, nucleotide bases, and amino acids form linear sequences that determine the primary structure of the molecules.
- These molecules can be considered molecular fossils that encode the history of millions of years of evolution.
- During this time period, the molecular sequences undergo random changes, some of which are selected during the process of evolution.
- As the selected sequences gradually accumulate mutations and diverge over time, traces of evolution may still remain in certain portions of the sequences to allow identification of the common ancestry.
- The presence of evolutionary traces is because some of the residues that perform key functional and structural roles tend to be preserved by natural selection; other residues that may be less crucial for structure and function tend to mutate more frequently.

For example, active site residues of an enzyme family tend to be conserved because they are responsible for catalytic functions. **Therefore, by comparing sequences through alignment, patterns of conservation and variation can be identified**.

When a sequence alignment is generated correctly,

- it reflects the evolutionary relationship of the two sequences:
- regions that are aligned but not identical represent residue substitutions;
- regions where residues from one sequence correspond to nothing in the other represent insertions or deletions that have taken place on one of the sequences during evolution.

The **degree of sequence conservation** in the alignment reveals evolutionary relatedness of different sequences, whereas the **variation** between sequences reflects the changes that have occurred during evolution in the form of substitutions, insertions, and deletions.

- Identifying the evolutionary relationships between sequences helps to characterize the function of unknown sequences.

When asequence alignment reveals *significant* similarity among a group of sequences, they can be considered as belonging to the **same family**.

If one member within the family has a known structure and function, then that information can be transferred to those that have not yet been experimentally characterized.

- Therefore, **sequence alignment can be used as basis** for prediction of structure and function of uncharacterized sequences.
- Sequence alignment provides inference for the relatedness of two sequences under study.

If the two sequences share significant similarity, meaning that the two sequences must have derived from a common evolutionary origin.

SEQUENCE HOMOLOGY VERSUS SEQUENCE SIMILARITY

An important concept in sequence analysis is sequence homology.

Sequence homology

- is an inference or a conclusion about a common ancestral relationship drawn from sequence similarity comparison when the two sequences share a high enough degree of similarity.
- homology is a qualitative statement.

Sequence similarity

- is the percentage of aligned residues that are similar in physiochemical properties such as size, charge, and hydrophobicity.
- is a direct result of observation from the sequence alignment.
- Sequence similarity can be quantified using percentages;

For example, two sequences share 40% similarity.

Generally, the sequence similarity level depends on

- ✓ the type of sequences being examined
- ✓ sequence lengths.

Nucleotide sequences consist of only four characters,

therefore, unrelated sequences have at least a 25% chance of being identical.

For protein sequences, there are twenty possible amino acid residues,

therefore, two unrelated sequences can match up 5% of the residues by random chance.

If gaps are allowed, the percentage could increase to 10–20%.

Sequence length is also a crucial factor.

shorter the sequence, the higher the chance that some alignment by random chance.

longer the sequence, the higher the chance that some alignment by random chance.

For determining a homology relationship of two protein sequences, for example, if both sequences are aligned at full length, which is 100 residues long, an identity of 30% or higher can be safely regarded as having close homology. They are referred to as **"safe zone"**

If their identity level falls between 20% and 30%, determination of homologous relationships in this range becomes less certain. This is the area regarded as the **"twilight zone"**.

Below 20% identity, where high proportions of nonrelated sequences are present, homologous relationships cannot be reliably determined and this area regarded as **"midnight zone"**.

SEQUENCE SIMILARITY VERSUS SEQUENCE IDENTITY

- Sequence similarity and sequence identity are synonymous for nucleotide sequences.
- For protein sequences, however, the two concepts are very different.

In a protein sequence alignment,

Sequence identity refers to the percentage of matches of the same amino acid residuesbetween two aligned sequences.

Sequence similarity refers to the percentage of aligned residues that have similarphysicochemical characteristics and can be more readily substituted for each other.

Methods of Pair-wise sequence alignment

The overall goal of pair-wise sequence alignment is to find the best pairing of two sequences, such that there is maximum correspondence among residues.

To achieve this goal, one sequence needs to be shifted relative to the other to find the position where maximum matches are found.

There are two different alignment strategies that are often used:

- global alignment
- local alignment.

Global Alignment

- In *global alignment*, two sequences to be aligned are assumed to be generally similar over their entire length.
- Alignment is carried out from beginning to end of both sequences to find the best possible alignment across the entire length between the two sequences.

- This method is more applicable for aligning two closely related sequences of roughly the same length.
- For divergent sequences and sequences of variable lengths, this method may not be able to generate optimal results because it fails to recognize highly similar local regions between the two sequences.

```
seq1   EARDF-NQYYSSIKRSGSIQ
          .  :  .:::::::::.  .  .
seq2   LPKLFIDQYYSSIKRTMG-H
```

In this figure, the region with the highest similarity is highlighted in a box.

Local alignment

- does not assume that the two sequences in question have similarity over the entire length.
- It only finds local regions with the highest level of similarity between the two sequences and aligns these regions without regard for the alignment of the rest of the sequence regions.
- This approach can be used for aligning more divergent sequences with the goal of searching for conserved patterns in DNA or protein sequences.
- The two sequences to be aligned can be of different lengths.
- This approach is more appropriate for aligning divergent biological sequences containing only modules that are similar, which are referred to as *domains* or *motifs*.

```
seq1   NQYYSSIKRS
          .:::::::::.
seq2   DQYYSSIKRT
```

In the line between the two sequences, ":" indicates identical residue matches and "."indicates similar residue matches.

Alignment Algorithms

Alignment algorithms, both global and local, are fundamentally similar and only differ in the optimization strategy used in aligning similar residues.

Both types of algorithms can be based on one of the three methods:

1. dot matrix method,
2. dynamic programming method,
3. word method.

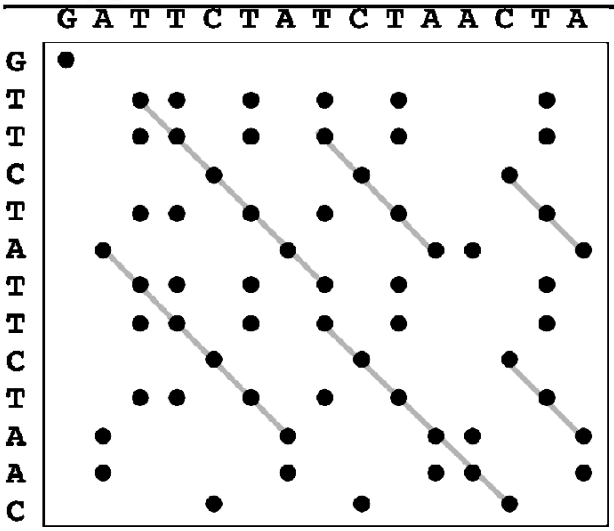
Dot Matrix Method

The most basic sequence alignment method is the dot matrix method, also known as the *dot plot method.*

- It is a graphical way of comparing two sequences in a two dimensional matrix. In a dot matrix,

- two sequences to be compared are written in the horizontal and vertical axes of the matrix.
- The comparison is done by scanning each residue of one sequence for similarity with all residues in the other sequence.
- If a residue match is found, a dot is placed within the graph. Otherwise, the matrix positions are left blank.
- When the two sequences have substantial regions of similarity, many dots line up to form contiguous diagonal lines, which reveal the sequence alignment.
- If there are interruptions in the middle of a diagonal line, they indicate insertions or deletions.
- Parallel diagonal lines within the matrix represent repetitive regions of the sequences (Figure).

Figure : Example of comparing two sequences using dot plots. Lines linking the dots indiagonals indicate sequence alignment. Diagonal lines above or below the main diagonal represent internal repeats of either sequence.



A problem exists when comparing large sequences using the dot matrix method due the high noise level.

For DNA sequences,

- the problem is particularly acute because there are only four possible characters in DNA and each residue therefore has a one-in-four chance of matching a residue in another sequence.
- There are many variations of using the dot plot method.

For example, a sequence can be aligned with itself to identify internal repeat elements.

- In the self comparison, there is a main diagonal for perfect matching of each residue.
- If repeats are present, short parallel lines are observed above and below the main diagonal.
- Self complementarily of DNA sequences (also called *inverted repeats*) – for example, those that form the stems of a hairpin structure – can also be identified using a dot plot.

- In this case, a DNA sequence is compared with its reverse-complemented sequence.
- Parallel diagonals represent the inverted repeats.

For comparing protein sequences,

- a weighting scheme has to be used to account for similarities of physicochemical properties of amino acid residues.
- The dot matrix method gives a direct visual statement of the relationship between two sequences and helps easy identification of the regions of greatest similarities.

Advantage of this method is

- Identification of sequence repeat regions based on the presence of parallel diagonals of the same size vertically or horizontally in the matrix.
- The method has some applications in genomics.
- It is useful in identifying chromosomal repeats
- comparing gene order conservation between two closely related genomes.
- used in identifying nucleic acid secondary structures through detecting self-complementarity of a sequence.

Limitation of this visual analysis method

- it lacks statistical rigor in assessing the quality of the alignment.
- The method is also restricted to pairwise alignment.
- It is difficult for the method to scale up to multiple alignment.

The following are examples of webservers that provide pairwise sequence comparison using dot plots.

- ✓ Dotmatcher (bioweb.pasteur.fr/seqanal/interfaces/dotmatcher.html)
- ✓ Dottup (bioweb.pasteur.fr/seqanal/interfaces/dottup.html)
- ✓ Dothelix (www.genebee.msu.su/services/dhm/advanced.html)
- ✓ MatrixPlot (www.cbs.dtu.dk/services/MatrixPlot/)

Dotmatcher

- aligns and displays dot plots of two input sequences (DNA or proteins) in FASTA format.
- A window of specified length and a scoring scheme are used.
- Diagonal lines are only plotted over the position of the windows if the similarity is above a certain threshold.

Dottup

- aligns sequences using the word method
- capable of handling genome-length sequences.
- Diagonal lines are only drawn if exact matches of words of specified length are found.

Dothelix

- is a dot matrix program for DNA or protein sequences.
- The program has a number of options for length threshold (similar to window size) and implements scoring matrices for protein sequences.
- In addition to drawing diagonal lines with similarity scores above a certain threshold,
- program displays actual pairwise alignment.

MatrixPlot

- is a more sophisticated matrix plot program for alignment of protein and nucleic acid sequences.
- The user has the option of adding information such as sequence logo profiles and distance matrices from known three-dimensional structures of proteins or nucleic acids.
- Instead of using dots and lines, the program uses colored grids to indicate alignment or other user-defined information.

Dynamic Programming Method

Dynamic programming is a method that determines optimal alignment by matching two sequences for all possible pairs of characters between the two sequences.

It is fundamentally similar to the dot matrix method in that it also creates a two dimensional alignment grid.

- it finds alignment in a more quantitative way by converting a dot matrix into a scoring matrix to account for matches and mismatches between sequences.
- By searching for the set of highest scores in this matrix, the best alignment can be accurately obtained.
- Dynamic programming works by first constructing a two-dimensional matrix whose axes are the two sequences to be compared.
- The residue matching is according to a particular scoring matrix.
- The scores are calculated one row at a time.
- This starts with the first row of one sequence, which is used to scan through the entire length of the other sequence, followed by scanning of the second row. The matching scores are calculated.
- The scanning of the second row takes into account the scores already obtained in the first round. The best score is put into the bottom right corner of an intermediate matrix. This process is iterated until values for all the cells are filled.
- Thus, the scores are accumulated along the diagonal going from the upper left corner to the lower right corner.
- Once the scores have been accumulated in matrix, the next step is to find the path that represents the optimal alignment.
- This is done by tracing back through the matrix in reverse order from the lower right-hand corner of the matrix toward the origin of the matrix in the upper left-hand corner.
- The best matching path is the one that has the maximum total score.

If two or more paths reach the same highest score, one is chosen arbitrarily to represent the best alignment.

Most commonly used pairwise alignment web servers apply the local alignment strategy, which include SIM, SSEARCH, and LALIGN.

SCORING MATRICES

SIM (http://bioinformatics.iastate.edu/aat/align/align.html)

SSEARCH (http://pir.georgetown.edu/pirwww/search/pairwise.html)

LALIGN (www.ch.embnet.org/software/LALIGN form.html)

SIM

- is a web-based program for pairwise alignment using the Smith–Waterman algorithm that finds the best scored nonoverlapping local alignments between two sequences.

- It is able to handle tens of kilobases of genomic sequence.
- The user has the option to set a scoring matrix and gap penalty scores.
- A specified number of best scored alignments are produced.

SSEARCH

- is a simple web-based programs that uses the Smith–Waterman algorithm for pairwise alignment of sequences.
- Only one best scored alignment is given.
- There is no option for scoring matrices or gap penalty scores.

LALIGN

- is a web-based program that uses a variant of the Smith–Waterman algorithm to align two sequences.
- gives a specified number of best scored alignments.
- The user has the option to set the scoring matrix and gap penalty scores.
- The same web interface also provides an option for global alignment performed by the ALIGN program.

Multiple sequence alignment

- Multiple sequence alignment is an essential technique in many bioinformatics applications.
- A natural extension of pairwise alignment is multiple sequence alignment, which is to align multiple related sequences to achieve optimal matching of the sequences.

- Related sequences are identified through the database similarity searching.
- It is theoretically possible to use dynamic programming to align any number of sequences as for pairwise alignment. However, the amount of computing time and memory it requires increases exponentially as the number of sequences increases.
- As a consequence, full dynamic programming cannot be applied for datasets of more than ten sequences. In practice, heuristic approaches are most often used.
- As the process generates multiple matching sequence pairs, it is often necessary to convert the numerous pairwise alignments into a single alignment, which arranges sequences in such a way that evolutionarily equivalent positions across all sequences are matched.

Advantages of multiple sequence alignment

- it reveals more biological information than many pairwise alignments can.
- it allows the identification of conserved sequence patterns and motifs in the whole sequence family, which are not obvious to detect by comparing only two sequences.
- Many conserved and functionally critical amino acid residues can be identified in a protein multiple alignment.
- essential prerequisite to carrying out phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures.
- has applications in designing degenerate polymerase chain reaction (PCR) primers based on multiple related sequences.

SCORING FUNCTION

- Multiple sequence alignment is to arrange sequences in such a way that a maximum number of residues from each sequence are matched up according to a particular scoring function.

- The scoring function for multiple sequence alignment is based on the concept of sum of pairs (SP).

it is the sum of the scores of all possible pairs of sequences in a multiple alignment based on a particular scoring matrix.

- In calculating the SP scores, each column is scored by summing the scores for all possible pairwise matches, mismatches and gap costs.
- The score of the entire alignment is the sum of all of the column scores.
- The purpose of most multiple sequence alignment algorithms is to achieve maximum SP scores.

Many algorithms have been developed to achieve optimal alignment.

- Some programs are exhaustive in nature; some are heuristic.
- Because exhaustive programs are not feasible in most cases, heuristic programs are commonly used.

These include

- progressive,
- iterative,
- block-based approaches.

Progressive method

- is a stepwise assembly of multiple alignment according to pairwise similarity.

Example is **Clustal**, - which is characterized by adjustable scoring matrices and gap penalties as well as by the application of weighting schemes.

T-Coffeeand**DbClustal**have been developed that combine both global and localalignment to generate more sensitive alignment.

Pralineis profile based and has the capacity to restrict alignment based on proteinstructure information and is thus much more accurate than Clustal.

Iterative approach

- works by repetitive refinement of suboptimal alignments.

Block-based method - focuses on identifying regional similarities.

EXHAUSTIVE ALGORITHMS

The exhaustive alignment method involves examining all possible aligned positions simultaneously.

Similar to dynamic programming in pair-wise alignment, which involves the use of a two-dimensional matrix to search for an optimal alignment,

To use dynamic programming for multiple sequence alignment,

extra dimensions are needed to take all possible ways of sequence matching into consideration.

This means to establish a multidimensional search matrix.

For example, for three sequences, a three-dimensional matrix is required to account for all possible alignment scores.

For aligning *N* sequences, an *N*-dimensional matrix is needed to be filled with alignment scores.

- As the amount of computational time and memory space required increases exponentially with the number of sequences, it makes the method computationally difficult to use for a large data set.

For this reason, full dynamic programming is limited to small datasets of less than ten short sequences.

DCA(Divide-and-Conquer Alignment, http://bibiserv.techfak.uni-bielefeld.de/dca/)

- is a web-based program that is in fact semi-exhaustive because certain steps of computation are reduced to heuristics.
- It works by breaking each of the sequences into two smaller sections.
- The breaking points are determined based on regional similarity of the sequences.
- If the sections are not short enough, further divisions are carried out.
- When the lengths of the sequences reach a predefined threshold, dynamic programming is applied for aligning each set of subsequences.
- The resulting short alignments are joined together head to tail to yield a multiple alignment of the entire length of all sequences.

HEURISTIC ALGORITHMS

Because the use of dynamic programming is not feasible for routine multiple sequence alignment, faster and heuristic algorithms have been developed. The heuristic algorithms fall into three categories:

- ✓ progressive alignment type,
- ✓ iterative alignment type,
- ✓ block-based alignment type.

Progressive Alignment Method

- depends on the stepwise assembly of multiple alignment and is heuristic in nature.
- It speeds up the alignment of multiple sequences through a multistep process.
- It first conducts pairwise alignments for each possible pair of sequences using the Needleman–Wunsch global alignment method and records these similarity scores from the pairwise comparisons.
- The scores can either be percent identity or similarity scores based on a particular substitution matrix.
- Both scores correlate with the evolutionary distances between sequences.
- The scores are then converted into evolutionary distances to generate a distance matrix for all the sequences involved.
- A simple phylogenetic analysis is then performed based on the distance matrix to group sequences based on pair-wise distance scores.

As a result,

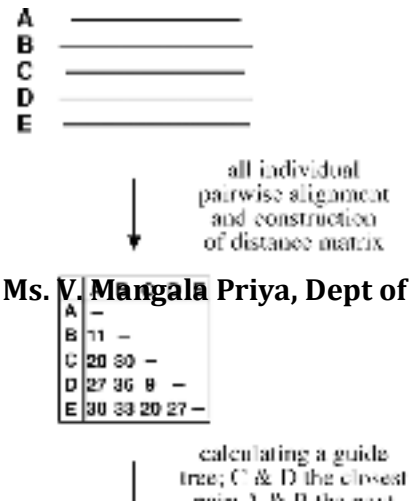
- a phylogenetic tree is generated using the neighbor-joining method.
- The tree reflects evolutionary proximity among all the sequences.

the resulting tree is an approximate tree and the tree can be used as a guide for directing realignment of the sequences called as a *guide tree.*

According to the guide tree,

- The two most closely related sequences are first re-aligned using the Needleman– Wunsch algorithm.
- To align additional sequences, the two already aligned sequences are converted to a consensus sequence with gap positions fixed.
- The consensus is then treated as a single sequence in the subsequent step.

- the next closest sequence based on the guide tree is aligned with the consensus sequence using dynamic programming.
- More distant sequences or sequence profiles are subsequently added one at a time in accordance with their relative positions on the guide tree.
- After realignment with a new sequence using dynamic programming, a new consensus is derived, which is then used for the next round of alignment.
- The process is repeated until all the sequences are aligned (as shown in the Figure



Schematic of a typical progressive alignment procedure (e.g., Clustal).

Angled wavy lines represent consensus sequences for sequence pairs A/B and C/D.

Clustal- (www.ebi.ac.uk/clustalw/)

is a progressive multiple alignment program available either as a stand-alone or on-line program.

The stand-alone program, which runs on UNIX and Macintosh, has two variants, ClustalW and ClustalX.

The W version provides a simple text-based interface and the X version provides user-friendly graphical interface.

Features of Clustal:

1) this program is the flexibility of using substitution matrices.

2) Clustal does not rely on a single substitution matrix.

Instead, it applies different scoring matrices when aligning sequences, depending on degrees of similarity.

3) The choice of a matrix depends on the evolutionary distances measured from the guide tree. For example,

for closely related sequences that are aligned in the initial steps,

Clustal automatically uses the BLOSUM62 or PAM120 matrix.

When more divergent sequences are aligned in later steps of the progressive alignment,

BLOSUM45 or PAM250 matrices may be used.

4) Clustal is the use of adjustable gap penalties that allow more insertions and deletions in regions that are outside the conserved domains, but fewer in conserved regions.

For example, a gap near a series of hydrophobic residues carries more penalties than the series of hydrophilic or glycine residues, which are common in loop regions.

In addition, gaps that are too close to one another can be penalized more than gaps occurring in isolated loci.

5) The program also applies a weighting scheme to increase the reliability of aligning divergent sequences (sequences with less than 25% identity).

This is done by down weightingredundant and closely related groups of sequences in the alignment by a certain factor.

6) This scheme is useful in preventing similar sequences from dominating the alignment.

The weight factor for each sequence is determined by its branch length on the guide tree.

The branch lengths are normalized by how many times sequences share a basal branch from the root of the tree. The obtained value for each sequence is subsequently used to multiply the raw alignment scores of residues from that sequence so to achieve the goal of decreasing the matching scores of frequent characters in a multiple alignment and thereby increasing the ones of infrequent characters.

DbClustal(http://igbmc.u-strasbg.fr:8080/DbClustal/dbclustal.html)

is a Clustalbased database search algorithm for protein sequences that combines local and global alignment features.

It first performs a BLASTP search for a query sequence.

The resulting sequence alignment pairs above a certain threshold are analyzed to obtain *anchorpoints*, which are common conserved regions, by using a program called Ballast.A global alignment is subsequently generated by Clustal, which is weighted toward the anchor points.

Poa(Partial order alignments,www.bioinformatics.ucla.edu/poa/)

is a progressive alignment program that does not rely on guide trees.

Instead, the multiple alignment is assembled by adding sequences in the order they are given.

Phylogenetic analysis

- Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized as phylogenetic trees.
- Thus, molecular phylogenetics is a fundamental aspect of bioinformatics.

MOLECULAR EVOLUTION AND MOLECULAR PHYLOGENETICS

"What is evolution?"

Evolution can be defined in various ways under different contexts.

In the biological context,

Evolution can be defined as the development of a biological form from other preexisting forms or its origin to the current existing form through natural selections and modifications.

The driving force behind evolution is natural selection in which "unfit" forms are eliminated through changes of environmental conditions or sexual selection so that only the fittest are selected.

The underlying mechanism of evolution is genetic mutations that occur spontaneously. The mutations on the genetic material provide the biological diversity within a population; hence, the variability of individuals within the population to survive successfully in a given environment.

Genetic diversity thus provides the source of raw material for the natural selection to act on.

Phylogenetics

is the study of the evolutionary history of living organisms using tree like diagrams to represent pedigrees of these organisms.

The tree branching patterns representing the evolutionary divergence are referred to as *phylogeny.*

Phylogenetics can be studied in various ways.

- studied using **fossil records**, which contain morphological information about ancestors of current species and the timeline of divergence.

- studied using **molecular data** that are in the form of DNA or protein sequences can also provide very useful evolutionary perspectives of existing organisms because, as organisms evolve, the genetic materials accumulate mutations over time causing phenotypic changes.

Because genes are the medium for recording the accumulated mutations, they can serve as ***molecular fossils***. Through comparative analysis of the molecular fossils from a number of related organisms, the evolutionary history of the genes and even the organisms can be revealed.

Advantage of using molecular data:

- Molecular data are more numerous than fossil records and easier to obtain.
- There is no sampling bias involved, whichhelps to mend the gaps in real fossil records.
- More clear-cut and robust phylogenetic trees can be constructed with the molecular data.
- Therefore,they have become favorite and sometimes the only information available for researchers to reconstruct evolutionary history.

The advent of the genomic era with tremendous amounts of molecular sequence data has led to the rapid development of molecular phylogenetics.

Molecular phylogeneticscan be defined as the study of evolutionary relationships of genes and other biological macromolecules by analyzing mutations at various positions in their sequences and developing hypotheses about the evolutionary relatedness of the biomolecules.

Based on the sequence similarity of the molecules, evolutionary relationships between the organisms can be inferred.

Major Assumptions:

To use molecular data to reconstruct evolutionary history requires number of reasonable assumptions.

[1] molecular sequences used in phylogenetic construction are homologous, meaning that they share a common origin and subsequently diverged through time.

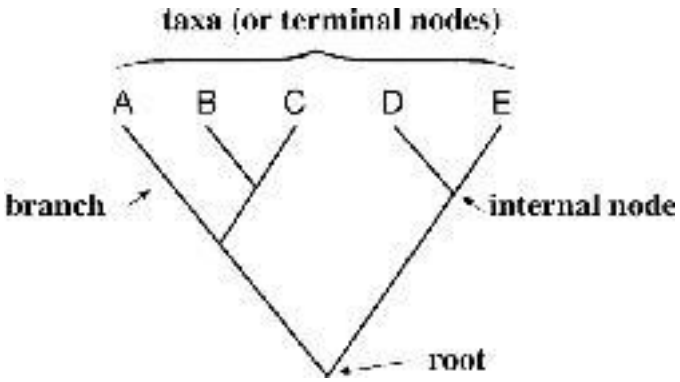
Phylogenetic divergence is assumed to be bifurcating, meaning that a parent branch splits into two daughter branches at any given point.

[2] each position in a sequence evolved independently.

[3] The variability among sequences is sufficiently informative for constructing unambiguous phylogenetic trees

TERMINOLOGY USED IN PHYLOGENETIC TREE

A typical bifurcating phylogenetic tree is a graph shown in Figure.



- ✓ The lines in the tree are called *branches.*
- ✓ At the tips of the branches are present-day species or sequences known as ***taxa*** (the singular form is *taxon*) or operational taxonomic units.
- ✓ The connecting point where two adjacent branches join is called a ***node***, which represents an inferred ancestor of extant taxa.
- ✓ The bifurcating point at the very bottom of the tree is the ***root node***, which represents the common ancestor of all members of the tree.
- ✓ A group of taxa descended from a single common ancestor is defined as a ***clade*** or ***monophyletic****group*.

In a monophyletic group,

- ✓ two taxa share a unique common ancestor not shared by any other taxa.
- ✓ They are also referred to as ***sister taxa*** to each other (e.g., taxa B and C).
- ✓ The branch path depicting an ancestor–descendant relationship on a tree is called a ***lineage***, which is often synonymous with a tree branch leading to a defined monophyletic group.
- ✓ When a number of taxa share more than one closest common ancestors, they do not fit the definition of a clade. In this case, they are referred to as ***paraphyletic*** (e.g., taxa B, C, and D).
- ✓ The branching pattern in a tree is called ***tree topology***.

- When all branches bifurcate on a phylogenetic tree, it is referred to as ***dichotomy***.
- In this case, each ancestor divides and gives rise to two descendants.
- Sometimes, a branch point on a phylogenetic tree may have more than two descendents, resulting in a ***multifurcating node***.
- The phylogeny with multifurcating branches is called ***polytomy*** (as shown in the following Figure).

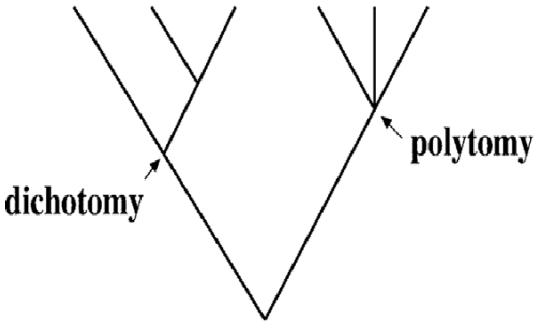


Figure showing an example ofbifurcation and multifurcation.

Multifurcation is normally a result of insufficient evidence to fully resolve the tree or a

- A polytomy can be a result of either an ancestral taxon giving rise to more than two immediate descendants simultaneously during evolution, a process known as ***radiation,*** or an unresolved phylogeny in which the exact order of bifurcations cannot be determined precisely.
- A phylogenetic tree can be either rooted or unrooted (as shown the following Figure).

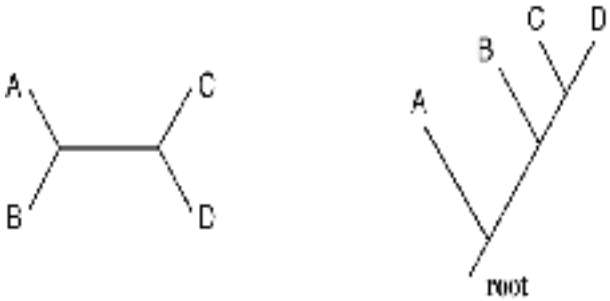


Figure shows rooted versusunrooted trees.

A phylogenetic tree without definition of a root is unrooted (*left*).

- ✓ An ***unrooted phylogenetic tree*** does not assume knowledge of a common ancestor, but only positions the taxa to show their relative relationships.

Because there is no indication of which node represents an ancestor, there is no direction of an evolutionary path in an unrooted tree.

- ✓ To define the direction of an evolution path, a tree must be rooted.
- ✓ In a ***rooted tree***, all the sequences under study have a common ancestor or root node from which a unique evolutionary path leads to all other nodes.
- ✓ Rooted tree is more informative than an unrooted one.
- ✓ To convert an unrooted tree to a rooted tree,one needs to first determine where the root is.
- ✓ Strictly speaking, the root of the tree is not known; the common ancestor is already extinct.

There are two ways to define the root of a tree.

1. ***outgroup approach***, - which is a sequence that is homologous to the sequences underconsideration, but separated from those sequences at an early evolutionary time.

- Outgroups are generally determined from independent sources of information.

For example, a bird sequence can be used as a root for the phylogenetic analysis of mammals based on multiple evidence that indicate that birds branched off prior to all mammalian taxa in the ingroup.

- Outgroups are required to be distinct from the ingroup sequences, but not too distant from the ingroup.

- Using too divergent sequences as an outgroup can lead to errors in tree construction.

2. ***midpoint rooting approach***

In the absence of a good outgroup, a tree can be rooted using the *midpoint rootingapproach*, in which the mid point of the two most divergent groups judged by overallbranch lengths is assigned as the root.

This type of rooting assumes that divergence from root to tips for both branches is equal and follows the “molecular clock” hypothesis.

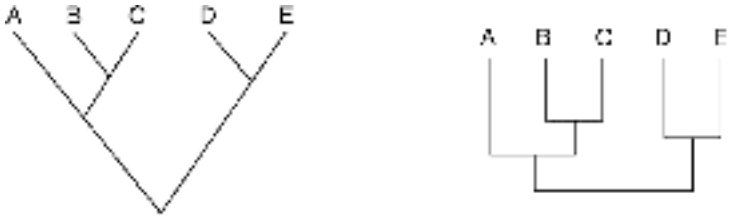
Molecular clock is an assumption by which molecular sequences evolve at constantrates so that the amount of accumulated mutations is proportional to evolutionary time. Based on this hypothesis, branch lengths on a tree can be used to estimate divergence time.

FORMS OF TREE REPRESENTATION

The topology of branches in a tree defines the relationships between the taxa.

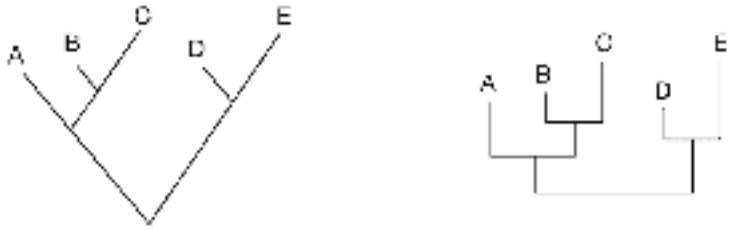
The trees can be drawn in different ways as shown the following figure

1. cladogram
2. phylogram



Cladogram

The branch lengths are unscaled in the cladograms



Phylogram

The branch lengths scaled in the phylograms.

In each of these tree representations, the branches of a tree can freely rotate without changing the relationships among the taxa.

In a *phylogram*,

- ✓ the branch lengths represent the amount of evolutionary divergence. Such trees are said to be scaled.

- ✓ The scaled trees have the advantage of showing both the evolutionary relationships and information about the relative divergence time of the branches.

In a *cladogram,*

- ✓ the external taxa line up neatly in a row or column.
- ✓ Their branch lengths are not proportional to the number of evolutionary changes and thus have no phylogenetic meaning.
- ✓ In these unscaled tree, only the topology of the tree matters, which shows the relative ordering of the taxa.

Procedure of phylogenetic tree:

Molecular phylogenetic tree construction can be divided into five steps:

(1) choosing molecular markers;

(2) performing multiple sequence alignment;

(3) choosing a model of evolution;

(4) determining a tree building method;

(5) assessing tree reliability.

(1) Choice of Molecular Markers

- For constructing molecular phylogenetic trees, one can use either nucleotide or protein sequence data.
- The choice of molecular markers is an important factor because it can make a major difference in obtaining a correct tree.
- The decision to use nucleotide or protein sequences depends on the properties of the sequences and the purposes of the study.
- For studying very closely related organisms, nucleotide sequences, which evolve more rapidly than proteins, can be used.

For example,

- ✓ For evolutionary analysis of different individuals within a population, non coding regions of mitochondrial DNA are often used.
- ✓ For studying the evolution of more widely divergent groups of organisms, choose either slowly evolving nucleotide sequences, such as ribosomal RNA or protein sequences.
    - If the phylogenetic relationships to be delineated are at the deepest level, such as between bacteria and eukaryotes, using conserved protein sequences makes more sense than using nucleotide sequences.

(2) Alignment

- The second step in phylogenetic analysis is to construct sequence alignment.
- This is the most critical step in the procedure because it establishes positional correspondence in evolution.
- Only the correct alignment produces correct phylogenetic inference because aligned positions are assumed to be genealogically related.
- Incorrect alignment leads to systematic errors in the final tree or even a completely wrong tree.
- For that reason, it is essential that the sequences are correctly aligned.
- Multiple state-of-the-art alignment programs such as T-Coffee should be used.
- The alignment results from multiple sources should be inspected and compared carefully to identify the most reasonable one.
- Automatic sequence alignments almost always contain errors and should be further edited or refined if necessary.
- Manual editing is often critical in ensuring alignment quality.
- It is also often necessary to decide whether to use the full alignment or to extract parts of it. Truly ambiguously aligned regions have to be removed from consideration prior to phylogenetic analysis.
- Which part of the alignment to remove is often at the discretion of the researcher. It is a rather subjective process.

- In extreme cases, someresearchers like to remove all insertions and deletions (indels) and only use positions that are shared by all sequences in the dataset.
- The clear drawback of this practice is that many phylogenetic signals are lost.
- In fact, gap regions often belong to *signature indels* unique to identification of a subgroup of sequences and should to be retained for treeing purposes.
- In addition, there is an automatic approach in improving alignment quality.

Rascal and NorMD - can help to improve alignment by correcting alignment errors andremoving potentially unrelated or highly divergent sequences.

Gblocks (http://woody.embl-heidelberg.de/phylo/) - can help to detect and eliminatethe poorly aligned positions and divergent regions so to make the alignment more suitable for phylogenetic analysis.

Multiple Substitutions

- A simple measure of the divergence between two sequences is to count the number of substitutions in an alignment.
- The proportion of substitutions defines the observed distance between the two sequences.
- However, the observed number of substitutions may not represent the true evolutionary events that actually occurred.
- When a mutation is observed as A replaced by C, the nucleotide may have actually undergone a number of intermediate steps to become C, such as A→T→G→C.
- Similarly, a back mutation could have occurred when a mutated nucleotide reverted back to the original nucleotide. This means that when the same nucleotide is observed, mutations like G→C→G may have actually occurred.
- Moreover, an identical nucleotide observed in the alignment could be due to parallel mutations when both sequences mutate into T, for instance.
- Such multiple substitutions and convergence at individual positions obscure the estimation of the true evolutionary distances between sequences. This effect is

known as ***homoplasy***, which, if not corrected, can lead to the generation of incorrect trees.

- To correct homoplasy, statistical models are needed to infer the true evolutionary distances between sequences.

(3) Choosing Substitution Models

The statistical models used to correct homoplasy are called ***substitution models*** **or** ***evolutionary models.***

For constructing DNA phylogenies, there are a number of nucleotide substitution models available.

- These models differ in how multiple substitutions of each nucleotide are treated.

1. Jukes–Cantor Model

The simplest nucleotide substitution model is the Jukes–Cantor model, which assumes that all nucleotides are substituted with equal probability.

A formula for deriving evolutionary distances that include hidden changes is introduced by using alogarithmic function.

dAB =−(3⁄4) ln[1−(4⁄3)pAB]

where dAB is the evolutionary distance between sequences A and B and

pAB is the observed sequence distance measured by the proportion of substitutions overthe entire length of the alignment.

2. KimuraModel

Another model to correct evolutionary distances is called the Kimura two-parameter model. This is a more sophisticated model in which mutation rates for transitions and transversion are assumed to be different, which is more realistic.

According tothis model, transitions occur more frequently than transversions, which, therefore,provides a more realistic estimate of evolutionary distances.

The Kimura model uses the following formula:

dAB =−(1⁄2) ln(1−2$p$ti− $p$tv)−(1⁄4) ln(1−2$p$tv)

where dAB is the evolutionary distance between sequences A and B,

*p*ti is the observed frequency for transition,

*p*tv the frequency of transversion.

(4) Phylogenetic Tree Construction Methods and Programs

There are currently two main categories of tree-building methods, each having advantages and limitations.

1. Distance based method

is based on distance, which is the amount of dissimilarity between pairs of sequences, computed on the basis of sequence alignment.

The distance-based methods assume that all sequences involved are homologous and that tree branches are additive, meaning that the distance between two taxa equals the sum of all branch lengths connecting them.

2. Character based method

is based on discrete characters, which are molecular sequences from individual taxa. The basic assumption is that characters at corresponding positions in a multiple sequence alignment are homologous among the sequences involved. Therefore, the character states of the common ancestor can be traced from this dataset. Another assumption is that each character evolves independently and is therefore treated as an individual evolutionary unit.

DISTANCE-BASED METHODS

- True evolutionary distances between sequences can be calculated from observed distances after correction using a variety of evolutionary models.
- The computed evolutionary distances can be used to construct a matrix of distances between all individual pairs of taxa.
- Based on the pairwise distance scores in the matrix, a phylogenetic tree can be constructed for all the taxa involved.
- The algorithms for the distance-based tree-building method can be subdivided into either clustering based or
  - optimality based.

Clustering-type algorithms

- compute a tree based on a distance matrix starting from the most similar sequence pairs.

- These algorithms includes

(i) Unweighted pair group method using arithmetic average (UPGMA) algorithm

(ii) neighbor joining algorithm.

Optimality-based algorithms

- compare many alternative tree topologies and select one that has the best fit between estimated distances in the tree and the actual evolutionary distances.

- This category includes

(i) Fitch–Margoliash algorithms

(ii) minimum evolution algorithms.

Clustering-Based Methods

(i) Unweighted Pair Group Method Using Arithmetic Average (UPGMA) algorithm

- ✓ The simplest clustering method is UPGMA, which builds a tree by a sequential clustering method.
- ✓ Given a distance matrix, it starts by grouping two taxa with the smallest pairwise distance in the distance matrix.
- ✓ A node is placed at the midpoint or half distance between them.
- ✓ It then creates a reduced matrix by treating the new cluster as a single taxon.
- ✓ The distances between this new composite taxon and all remaining taxa are calculated to create a reduced matrix.
- ✓ The same grouping process is repeated and another newly reduced matrix is created.
- ✓ The iteration continues until all taxa are placed on the tree.
- ✓ The last taxon added is considered the outgroup producing a rooted tree.

- ✓ The basic assumption of the UPGMA method is that all taxa evolve at a constant rate and that they are equally distant from the root, implying that a molecular clock is in effect.
- ✓ However, real data rarely meet this assumption. Thus, UPGMA often produces erroneous tree topologies.
- ✓ However, owing to its fast speed of calculation, it has found extensive usage in clustering analysis of DNA microarray data.

(ii) Neighbor Joining algorithm

- ✓ The UPGMA method uses unweighted distances and assumes that all taxa have constant evolutionary rates.
- ✓ Since this molecular clock assumption is often not met in biological sequences, to build a more accurate phylogenetic trees, the neighbor joining (NJ) method can be used, which is somewhat similar to UPGMA.
- ✓ It builds a tree by using stepwise reduced distance matrices.
- ✓ the NJ method does not assume the taxa to be equidistant from the root.
- ✓ It corrects for unequalevolutionary rates between sequences by using a conversion step.
- ✓ This conversion requires the calculations of "*r*-values" and "transformed *r*-values" using the following formula:

$d\_AB = dAB - 1/2 \times (rA + rB)$

where *d*_AB is the converted distance between A and B and

*d*AB is the actual evolutionary distance between A and B.

The value of *r*A (or *r*B) is the sum of distances of A (or B) to all other taxa.

Optimality-Based Methods

optimality-based methods have a well-defined algorithm to compare all possible tree topologies and select a tree that best fits the actual evolutionary distance matrix. Based on the differences in optimality criteria, there are two types of algorithms,

(i) Fitch–Margoliash algorithms

(ii) minimum evolution algorithms

(i) Fitch–Margoliash algorithms

- ✓ The Fitch–Margoliash (FM) method selects a best tree among all possible trees based on minimal deviation between the distances calculated in the overall branches in the tree and the distances in the original dataset.
- ✓ It starts by randomly clustering two taxa in a node and creating three equations to describe the distances, and then solving the three algebraic equations for unknown branch lengths.
- ✓ The clustering of the two taxa helps to create a newly reduced matrix.
- ✓ This process is iterated until a tree is completely resolved.
- ✓ The method searches for all tree topologies and selects the one that has the lowest squared deviation of actual distances and calculated tree branch lengths.

(ii) Minimum Evolution algorithm

Minimum evolution (ME) constructs a tree with a similar procedure, but uses a different optimality criterion that finds a tree among all possible trees with a minimum overall branch length.

Searching for the minimum total branch length isan indirect approach to achieving the best fit of the branch lengths with the original dataset.

Pros and Cons

The most frequently used distance methods are clustering based.

Major advantage is that

- they are computationally fast and are therefore capable of handling datasets that are deemed to be too large for any other phylogenetic method.
- The overall advantage of all distance-based methods is the ability to make use of a large number of substitution models to correct distances.

Drawback is that

- The actual sequence information is lost when all the sequence variation is reduced to a single value.

ancestral sequences at internal nodes cannot be inferred.

CHARACTER-BASED METHODS Character-based methods (also called *discrete methods*) are

- based directly on the sequence characters rather than on pairwise distances.

- They count mutational events accumulated on the sequences and may therefore avoid the loss of information when characters are converted to distances.

- This preservation of character information means that evolutionary dynamics of each character can be studied.

- Ancestral sequences can also be inferred.

- The two most popular character-based approaches are

(i) maximum parsimony (MP) mehod

(ii) - (ML) method.

(i) Maximum Parsimony method

- ✓ The parsimony method chooses a tree that has the fewest evolutionary changes or shortest overall branch lengths.
- ✓ It is based on a principle related to a medieval philosophy called *Occam's razor*.
- ✓ The theory was formulated by William of Occam in the thirteenth century
- ✓ For phylogenetic analysis, parsimony seems a good assumption.

By this principle,

- ✓ a tree with the least number of substitutions is probably the best to explain the differences among the taxa under study.
- ✓ This view is justified by the fact that evolutionary changes are relatively rare within a reasonably short time frame.
- ✓ This implies that a tree with minimal changes is likely to be a good estimate of the true tree.
- ✓ By minimizing the changes, the method minimizes the phylogenetic noise owing to homoplasy and independent evolution.

(ii) Maximum Likelihood Method

- ✓ uses probabilistic models to choose a best tree that has the highest probability or likelihood of reproducing the observed data.
- ✓ It finds a tree that most likely reflects the actual evolutionary process.
- ✓ ML is an exhaustive method that searches every possible tree topology and considers every position in an alignment, not just informative sites.
- ✓ By employing a particular substitution model that has probability values of residue substitutions, ML calculates the total likelihood of ancestral sequences evolving to internal nodes and eventually to existing sequences.
- ✓ ML works by calculating the probability of a given evolutionary path for a particular extant sequence.
- ✓ The probability values are determined by a substitution model (either for nucleotides or amino acids).

(5) PHYLOGENETIC TREE EVALUATION

After phylogenetic tree construction, the next step is to statistically evaluate the reliability of the inferred phylogeny.

There are two questions that need to be addressed.

(i) how reliable the tree or a portion of the tree is;

(ii) whether this tree is significantly better than another tree.

To **answer the first question**,

We need to use analytical resampling strategies such as **bootstrapping** and **jackknifing**, which repeatedly resample data from the original dataset.

For the second question,

conventional statistical tests are needed.

What Is Bootstrapping?

- *Bootstrapping* is a statistical technique that tests the sampling errors of a phylogenetictree.

- The rationale for bootstrapping is that a newly constructed tree is possibly biased owing to incorrect alignment or chance fluctuations of distance measurements.

To determine the robustness or reproducibility of the current tree,

- trees are repeatedly constructed with slightly perturbed alignments that have some random fluctuations introduced.

A truly robust phylogenetic relationship should have enough characters to support the relationship even if the dataset is perturbed in such away.

Otherwise, the noise introduced in the resampling process is sufficient to generate different trees, indicating that the original topology may be derived from weak phylogenetic signals. Thus, this type of analysis gives an idea of the statistical confidence of the tree topology.

Parametric and Nonparametric Bootstrapping

- Bootstrap resampling relies on perturbation of original sequence datasets. There are two perturbation strategies.

Nonparametric bootstrapping - is through random replacement of sites.

parametric bootstrapping - new datasets can be generated based on a particular sequence distribution.

Both types of bootstrapping can be applied to the distance, parsimony, and likelihood tree construction methods.

In nonparametric bootstrapping,

a new multiple sequence alignment of the same length is generated with random duplication of some of the sites (i.e., the columns in an alignment) at the expense of some other sites.

- In other words, certain sites are randomly replaced by other existing sites.

- This process is repeated 100 to 1,000 times to create 100 to 1,000 new alignments that are used to reconstruct phylogenetic trees using the same method as the originally inferred tree.

- The new datasets with altered the nucleotide or amino acid composition and rate heterogeneity may result in certain parts of the tree having a different topology from the original inferred tree.

- All the bootstrapped trees are summarized into a consensus tree based on a majority rule.

- The most supported branching patterns shown at each node are labeled with bootstrap values, which are the percentage of appearance of a particular clade.

Thus,

the bootstrap test provides a measure for evaluating the confidence levels of the tree topology. Analysis has shown that a bootstrap value of 70% approximately corresponds to 95% statistical confidence.

In parametric bootstrapping

- uses altered datasets with random sequences confined within a particular sequence distribution according to a given substitution model.

- The parametric bootstrapping method may help avoid the problem of certain sites being repeated too many times as in nonparametric bootstrapping resulting in skewed sequence distribution.

- If a correct nucleotide/amino acid distributionmodel is used, parametric bootstrapping generates more reasonable replicates than random replicates. Thus, this procedure is considered more robust than nonparametric bootstrapping.

Jackknifing

- In addition to bootstrapping, another often used resampling technique is jackknifing.

- In jackknifing, one half of the sites in a dataset are randomly deleted, creating datasets half as long as the original.

- Each new dataset is subjected to phylogenetic tree construction using the same method as the original.

Advantage of jackknifing

- is that sites are not duplicated relative to the original dataset and that computing time is much shortened because of shorter sequences.

Disadvantage of this approach

- is that the size of datasets has been changed into one half and that the datasets are no longer considered replicates. Thus, the results may not be comparable with that from bootstrapping.

PHYLOGENETIC PROGRAMS

There are numerous phylogenetic programs available,

For a list of hundreds of phylogenetic software programs, available in Felsenstein's collection at: http://evolution.genetics.washington.edu/phylip/software.html.

Most of these programsare freely available. Some are comprehensive packages; others are more specialized to perform a single task.

PAUP* (Phylogenetic analysis using parsimony and other methods, by David Swofford,http://paup.csit.fsu.edu/)

- is a commercial phylogenetic package.
- It is probably one of the most widely used phylogenetic programs available from Sinauer Publishers.
- It is a Macintosh program (UNIX version available in the GCG package) with a very user-friendly graphical interface.
- PAUP was originally developed as a parsimony program, but expanded to a comprehensive package that is capable of performing distance, parsimony, and likelihood analyses.
- The distance options include NJ, ME, FM, andUPGMA.
- PAUP is also able to perform nonparametric bootstrapping, jackknifing, KH testing, and SH testing.

Phylip(Phylogenetic inference package; by Joe Felsenstein) at http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html.

- is a free multi platform comprehensive package containing thirty-five subprograms for performing distance, parsimony, and likelihood analysis, as well as bootstrapping for both nucleotide and amino acid sequences.
- It is command-line based, but relatively easy to use for each single program.

PHYML(http://atgc.lirmm.fr/phyml/)

- is a web-based phylogenetic program using the GA.
- It first builds an NJ tree and uses it as a starting tree for subsequent iterative refinement through subtree swapping.
- Branch lengths are simultaneously optimized during this process.
- The tree searching stops when the total ML score no longer increases.
- The main advantage of this program is the ability to build trees from verylarge datasets with hundreds of taxa and to complete tree searching within a relatively short time frame.

Database similarity searching

- A main application of pairwise alignment is retrieving biological sequences in databases based on similarity.
- This process involves submission of a query sequence and performing a pairwise comparison of the query sequence with all individual sequences in a database.
- Thus, database similarity searching is pairwise alignment on a large scale.
- This type of searching is one of the most effective ways to assign putative functions to newly determined sequences.

UNIQUE REQUIREMENTS OF DATABASE SEARCHING

There are unique requirements for implementing algorithms for sequence database searching.

1. ***sensitivity***, which refers to the ability to find as many correct hits as possible.

These correct hits are considered “true positives” in the database searching exercise.

2. ***selectivity***, also called*specificity*, which refers to the ability to exclude incorrect hits.These incorrect hits are unrelated sequences mistakenly identified in database searching

and are considered “false positives.”

3. ***speed***, which is the time it takes to get results from database searches.In database searching, there are two fundamental types of algorithms.

1. ***exhaustive type***, which uses a rigorous algorithm to find the best or exact solution fora particular problem by examining all mathematical combinations. Dynamic programming is an example of the exhaustive method and is computationally very intensive.

2. ***heuristic type***, which is a computational strategy to find an empirical or near optimalsolution by using rules of thumb. Essentially, this type of algorithms take shortcuts by reducing the search space according to some criteria.

HEURISTIC DATABASE SEARCHING

Currently, there are two major heuristic algorithms for performing database searches: **BLAST** and **FASTA**.

- These methods are not guaranteed to find the optimal alignment or true homologs, but are 50–100 times faster than dynamic programming.

- Both BLAST and FASTA use a heuristic *word method* for fast pairwise sequence alignment. This is the third method of pairwise sequence alignment.

- It works by finding short stretches of identical or nearly identical letters in two sequences. - - These short strings of characters are called *words*, which are similar to the windows used in the dot matrix method.

- Once regions of high sequence similarity are found, adjacent high-scoring regions can be joined into a full alignment.

BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST)

- was developed by Stephen Altschul of NCBI in 1990

  - become one of the most popular programs for sequence analysis.
  - BLAST uses heuristics to align a query sequence with all sequences in a database.

The objective is

- to find high-scoring ungapped segments among related sequences.

- helps to discriminate related sequences from unrelated sequences in a database. BLAST performs sequence alignment through the following steps.

The **first step**

Is to create a list of words fromthe query sequence.

Each word is typically three residues for protein sequences and eleven residues for DNA sequences.

The list includes every possible word extracted from the query sequence.

This step is also called *seeding*.

The second step

is to search a sequence database for the occurrence of these words.

This step is to identify database sequences containing the matching words.

The matching of the words is scored by a given substitution matrix.

A word is considered a match if it is above a threshold.

The fourth step

involves pairwise alignment by extending from the words in both directions while counting the alignment score using the same substitution matrix.

The extension continues until the score of the alignment drops below a threshold due to mismatches (the drop threshold is twenty-two for proteins and twenty for DNA).

The resulting contiguous aligned segment pair without gaps is called *high-scoringsegment pair* (HSP).

In the original version of BLAST, the highest scored HSPs are presented as the final report.

They are also called maximum scoring pairs.

A recent improvement in the implementation of

BLAST is the ability to provide gapped alignment.

- In gapped BLAST, the highest scored segment is chosen to be extended in both directions using dynamic programming where gaps may be introduced.

Variants of BLAST

BLAST is a family of programs that includes

BLASTN,

BLASTP,

BLASTX

TBLASTN,

TBLASTX.

BLASTN - queries nucleotide sequences with a nucleotide sequence database.

BLASTP - uses protein sequences as queries to search against a protein sequence database. **BLASTX** - uses nucleotide sequences as queries and translates them in all six reading frames to produce translated protein sequences, which are used to query a protein sequence database. **TBLASTN** - queries protein sequences to a nucleotide sequence database with the sequences translated in all six reading frames.

TBLASTX - uses nucleotide sequences, which are translated in all six frames, to searchagainst a nucleotide sequence database that has all the sequences translated in six frames.

BLASTweb server(www.ncbi.nlm.nih.gov/BLAST/)

- ✓ has been designed in such a way as to simplify the task of program selection.
- ✓ The programs are organized based on the type of query sequences, protein sequences, nucleotide sequences, or nucleotide sequence to be translated.
- ✓ In addition, programs for special purposes are grouped separately;

For example,

bl2seq, immunoglobulinBLAST, and VecScreen,

- ✓ The BLAST programs specially designed for searching individual genome databases are also listed in a separate category.
- ✓ The choice of the type of sequences also influences the sensitivity of the search.

POSSIBLE QUESTIONS

UNIT-I

PART-A (1 MARKS)

(Q.NO 1 TO 20 Online Examination)

PART-B (2 MARKS)

1. Define alignment?
2. Differentiate pairwise and multiple Alignment?
3. Differentiate local and global alignment?
4. Define sequence homology?
5. Define sequence similarity?
6. Name any two methods for local alignment?
7. Name any two tools for alignment?
8. What are scoring matrices?
9. What are the algorithms developed for optimizing MSA?
10. Define phylogeny?
11. What is evolution?
12. Define taxa?
13. Define monophyletic group?
14. Differentiate dichotomy and polytomy?
15. Differentiate rooted tree and unrooted tree?
16. Differentiate cladogram and dendrogram?
17. What are substitution models? What is its application in predicting phylogeny?
18. Define bootstrapping?
19. What is meant by sensitivity of database searching?
20. Define BLAST?

PART-C (8 MARKS)

1. Describe dot matrix method of sequence alignment?
2. Write notes on the role of scoring matrices in sequence alignment?
3. Describe exhaustive algorithm of MSA?

4. Describe heuristic algorithm of MSA?
5. Write notes on ClustalW?
6. Define phylogenetic tree? What are the different ways of representing a tree?
7. Write notes on the unique requirements of database searching?
8. Describe BLAST.

Unit III Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
Example of progresive method of multiple sequence alignment is	Clustal	Iterative approach	Exhaustive alignment	Text based	Clustal
Reptitive refinement of sub-optimal alignment is	Iterative approach	Exhaustive alignment	Text based	Heuristic	Iterative approach
The percentage matches of similar physiochemical characters of amino acids between two aligned sequences is called as	Sequence similarity	Sequence identity	Global alignment	Local alignment	Sequence similarity
If two aligned sequences assumes similar over their entire length is called as	Global alignment	Sequence similarity	Sequence identity	Local alignment	Global alignment
Sequence similarity is a statement of	Quantitative	qualitative	relative	average	Quantitative
Sequence similarity level depends on the sequence	Type and length	Type and average	length	Type	Type and length
Molecular fossils contains the information of	DNA and proteins	DNA and RNA	DNA and Lipids	Lipids and proteins	DNA and proteins
No sampling bias involved in	Molecular fossil	Phylogenetics	Clade	Morphological	Molecular fossil
The branching pattern in a tree is called as	Tree topology	Root	Node	Branches	Tree topology
If all branches bifurcate in phylogenetic tree is referred as	Dichotomy	Root	Node	Branches	Dichotomy
Neighbor joining method builds a tree based on the method of	Stepwise distance matrix	Phylogram	Cladogram	Polytomy	Stepwise distance matrix
The statistical technique that tests the sampling errors of a phylogenetic tree is called as	Bootstrapping	Phylogram	Cladogram	Polytomy	Bootstrapping
Process of aligning two sequence is called as	Pairwise alignment	Protein alignment	Multiple alignment	Process alignment	Pairwise alignment
The products of evolution are	DNA and proteins	DNA and RNA	DNA and Lipids	RNA and Lipids	DNA and proteins
The chance of being identical unrelated nucleotide sequence is	25%	30%	40%	50%	25%
The chance of being identical unrelaed	5%	10%	15%	20%	5%

protein sequence is					
If two aligned sequences assumes similar at local region is called as	Local alignment	Global alignment	Sequence similarity	Sequence identity	Local alignment
Graphical way of comparing two sequences in two dimensional matrix is	Dot matrix	Dot matcher	Dottup	Dot analysis	Dot matrix
Simultaneously analyzing all possible aligned positions is called as	Exhaustive alignment	Iterative approach	Text based	Heuristic	Exhaustive alignment
Clustal W version provides the interface of	Text based	Exhaustive alignment	Iterative approach	Heuristic	Text based
The study of evolutionary history of genes and other macromolecules is called as	Molecular phylogenetics	Phylogenetics	Clade	Branches	Molecular phylogenetics
The lines in the phylogenetic tree are called as	Branches	Node	Root	Clade	Branches
The branch point with more than two descendents is called as	Multifurcating node	Root	Node	Branches	Multifurcating node
The phylogeny with multifurcating branches is called as	Polytomy	Root	Node	Branches	Polytomy
Expansion of UPGMA	Unweighted Pair group method using arithmetic average	Phylogram	Cladogram	Polytomy	Unweighted Pair group method using arithmetic average
UPGMA builds a tree based on the method of	Sequential clustering	Phylogram	Cladogram	Polytomy	Sequential clustering
Two taxa shares a unique common ancestor are called as	Sister taxa	Root	Node	Branches	Sister taxa
Number of taxa shares more than one common ancestors are called as	Paraphyletic	Root	Node	Branches	Paraphyletic
The tree branching pattern of evolutionary divergence is called as	Phyogeny	Molecular fossil	Clade	Morphological	Phyogeny
Fossil records contains the information of	Morphological	Biochemical	Chemical	Molecular	Morphological
The scoring function for multiple sequence alignmnet is based on the	Sum of pairs	Squre of pairs	Multiple of pairs	Division of pairs	Sum of pairs

concept of					
Commonly used algorithms in optimal alignment is	Heuristic	Iterative approach	Exhaustive alignment	Text based	Heuristic
Midnight zone of sequence homology is	< 20 %	< 30 %	< 40 %	< 50 %	< 20 %
The percentage matches of same amino acids between two aligned sequences is called as	Sequence identity	Sequence similarity	Global alignment	Local alignment	Sequence identity
Sequence homology is a statement of	qualitative	quantitative	relative	average	qualitative
If two sequences shares similar physiochemical properties are known as	Sequence similarity	Sequence homology	Sequence difference	Structure homology	Sequence similarity
Degree of sequence conservation in the alignment reveals the evolutionary relatedness of	Different sequence	same sequence	Different segment	Same segment	Different sequence
The basis for structure and function prediction of uncharacterized sequence is	Sequence alignment	Structure alignment	Segment alignment	Sequence analysis	Sequence alignment
Lower similarity depends on the sequence length of	Longer	Shorter	Medium	Average	Longer
Higher similarity depends on the sequence length of	Shorter	Longer	Medium	Average	Shorter
The web-server aligns two sequence in FASTA format is	Dot matcher	Dot matrix	Dottup	Dot analysis	Dot matcher
The web-server aligns two sequence in word format is	Dottup	Dot matcher	Dot matrix	Dot analysis	Dottup
Clustal X version provides the interface of	Graphical	Text based	Exhaustive alignment	Iterative approach	Graphical
The matrix used by clustal for closely related sequences is	BLOSUM62	BLOSUM72	BLOSUM90	BLOSUM250	BLOSUM62
Taxa represents	Present day species	Past day species	Future day species	Average day species	Present day species
The ancestor of extant taxa represented by	Node	Root	Clade	Branches	Node
Unscaled branch length representation of phylogentic tree is called as	Cladogram	Root	Node	Branches	Cladogram

Scaled branch length representation of phylogenetic tree is called as	Phylogram	Root	Node	Branches	Phylogram
Distance based method of phylogentic tree is based on	Distance	Root	Node	Branches	Distance
Character based method of phylogenetic tree is based on	Character	Root	Node	Branches	Character
Common ancestor of all members of the tree represented by	Root	Node	Clade	Branches	Root
Group of taxa descended from a single common ancestor is called as	Clade	Root	Node	Branches	Clade
The matrix used by clustal for divergent sequences is	BLOSUM45	BLOSUM72	BLOSUM90	BLOSUM250	BLOSUM45
The study of the evolutionary history of living organisms using tree like diagrams is called as	Phylogenetics	Molecular fossil	Phyogeny	Clade	Phylogenetics
The handling capacity of SIM scoring matrices is	Tens of Kb	Thousands of Kb	Millions of Kb	Billions of Kb	Tens of Kb
Alignment of multiple related sequence is called as	Multiple sequence alignment	Pair-wise sequence alignment	Multiple structure alignment	Pairwise structure alignment	Multiple sequence alignment
Safe zone of homology between two protein sequence is	30%	40%	50%	60%	30%
Twilight zone of homology between two protein sequence is	20 to 30 %	30 to 40 %	40 to 50 %	40 to 60 %	20 to 30 %
Sequence alignment provides the inference of the two sequence	Relatedness	Difference	Variation	Homology	Relatedness
If two sequence shares high degree of similarity is known as	Sequence homology	Sequence difference	Structure homology	Sequence variance	Sequence homology

Unit IV: Genome organization and analysis

Diversity of Genomes: Viral, prokaryotic & eukaryotic genomes Genome, transcriptome, proteome, 2-D gel electrophoresis, Maldi Tof spectroscopy. Major features of completed genomes: *E.coli*, *S.cerevisiae*, *Arabidopsis*, Human.

Genome organization refers to the sequential, not the structural organization of the genome. Besides the **coding exons**, the non-coding DNA in Eukaryotes may fall in the following classes

- **Introns**. They are DNA sequences inserted between the exons and found in the ORF. They are spliced after the first level of transcription. Most introns are junk inserted within genes.
- **Pseudogenes**. 'Dead', non-functional copies of genes present elsewhere in the genome, but no longer of any use.
- **Retropseudogenes**. Like pseudogenes, but have been processed, i.e. lack introns. Produced by the action of reverse transcriptase (RT) on mRNA, and subsequent incorporation of the cDNA into the genome.
- **Transposons.** Jumping genes, which splice themselves in and out of the genome (in DNA form) randomly, by the action of transposase.
- **Retrotransposons**. Transcribed into an mRNA, which encodes an RT enzyme, which then copies the mRNA back to DNA and incorporates it into the genome.

Infact in humans only 1.5% of the entire genome length corresponds to coding DNA. This 1.5% codes for about 27,000 genes which in turn code for proteins that are responsible for all the cellular processes.

Genome Organisation

Definitions

Genome:

The total amount of genetic material, stored as DNA. The nuclear genome refers to the DNA in the chromosomes contained in the nucleus; in the case of humans the DNA in the 46 chromosomes. It is the nuclear genome that defines a multicellular organism; it will be the same for all (almost) cells of the organism. You can have organelle genomes as well such as the mitochondrial genome. When you want to identify or distinguish one organism from another, such as in forensic testing, you investigate the genome.

Transcriptome:

The total amount of genetic information which has been transcribed by the cell. This information will be stored as RNA. The transcriptome is unique to a cell type and is a measure of the gene expression. Different cells within an organism will have different transcriptomes. Cell types can be identified by their transcriptome.

Proteome:

The cell's complete protein output. This reflects all the mRNA sequences translated by the cell. Cell types have different proteomes and these can be used to identify a particular cell.

The genome organisation:

Base composition

The base composition of DNA in an organism is a fixed value and it is expressed as the % of (G + C) of the total genome. The variation of this value between different prokaryotes is large; this is surprising given that many of the individual proteins

produced by the species have similar amino acid sequences. Prokaryotic genomic DNA can have as little as 25 % (G + C) in Mycoplasma genitalium to as high as ~72 % in Micrococcus lysodeikticus. Eukaryotic genomic DNA does not display the same variation between species. The % (G + C) composition of most plant and animal species falls within a narrow range, averaging at 39% with a variation of only ± 6%.

Base composition within the genome.

In prokaryotes the bases are distributed evenly throughout the genome with a slightly lower (G + C) content in promoter and intergenic regions; these often have A + T rich segments which melt more readily than G+C rich regions. The relatively constant base distribution within a given bacterial genome suggests that although there may be unequal nucleotide pool sizes inside the cell, the system will have evolved over many generations to be like this, and the rate of DNA replication is constant. The distribution of the (G + C) content throughout each genome in eukaryotes, however, varies significantly, unlike prokaryotes. Whereas the mean variation in % (G + C) content throughout the E. coli genome is only 8.6 % in eukaryotes, this variation is over 30 %. Certain regions of eukaryotic genomic DNA are found to be (A + T) rich, with a % (G + C) content as low as 18 %, while other regions have a (G + C) content as high as 70 %.

The genome organisation: repetitive and unique sequences

The C-value paradox:

The C-value is the total number of DNA bases in the genome (per haploid set of chromosomes).

Some organisms seem to have far too much DNA for their complexity e.g. the carp has 52 chromosomes while the alligator has 16. Some flowers have far more genetic material than humans. Clearly the amount of DNA is not proportional to that required to produce all the proteins made by the organism or to their position on the food chain.

Diversity of genome

The genetic information for every organism is written in the universal language of DNAsequences, and the DNA sequence of any given organism can be obtained by

standard biochemical techniques, it is now possible to characterize, catalogue, and compare any set of living organisms with reference to these sequences

The human haploid genome consists of about 3 x 10^9 base pairs of DNA. Genomic DNA exists as single linear pieces of DNA that are associated with a protein called a nucleoprotein complex. The DNA-protein complex is the basis for the formation of chromosomes, virtually all of the genomic DNA is distributed among the 23 chromosomes that reside in the cellular nucleus. A very small fraction of the genome is also found in a 16,000 base pair circular piece of DNA that is found in the mitochondria. The double helical DNA of the chromatin is replicated with the chromatin fiber condensing into discrete bodies, the chromosomes, each consisting of two identical chromatids. The two sister chromatids separate, one moving to each pole of the cell, where they become part of the newly formed nucleus of each daughter cell. The cells that make up most of the body of a multicellular organism, the somatic cells, have two copies of each chromosome and are said to be diploid (2n). Egg and sperm for example, produced by meiosis and having only one copy of each chromosome, are haptoid (n). The DNA of chromatin and chromosomes is bound tightly to a family of positively charged proteins, the histones, which associate strongly with the many negatively charged phosphate groups in DNA. The histones and DNA associate in complexes called nucleosomes in which the DNA strand winds around a core of histone molecules.

Genome sequencing -Saccharomyces cerevisiae

A species of yeast used in winemaking, baking and brewing. This was the first fungi to be sequenced. It is sequenced by International Collaboration for the Yeast Genome Sequencing. *S. cerevisiae* was the first eukaryotic genometo be completely sequenced.

The genome sequence was released to the public domainon April 24, 1996. Since then, regular updates have been maintained at the Saccharomyces GenomeDatabase. This database isa highly annotated and cross-referenced database for yeast

researchers. Another important *S. cerevisiae* database is maintained by the Munich Information Center for Protein Sequences (MIPS).

- The *S.cerevisiae* genome is composed of about 12,156,677 base pairsand 6,275 genes, compactly organized on 16 chromosomes.
- Only about 5,800 of these genes are believed to be functional. It is estimated at least 31% of yeast genes have homologsin the human genome.
- Yeast genes are classified using gene symbols (such as sch9) or systematic names. In the latter case the 16 chromosomes of yeast are represented by the letters A to P, then the gene is further classified by a sequence number on the left or right arm of the chromosome, and a letter showing which of the two DNA strands contains its coding sequence.

Systematic gene names for Baker's yeast

gene name	**YGL118W**
Y	the Y to show this is a yeast gene
G	chromosome on which the gene is located
L	left or right arm of the chromosome
118	sequence number of the gene/ORF on this arm, starting at the centromere
W	whether the coding sequence is on the Watson or Crick strand

Saccharomyces cerevisiae has been widely utilized in the exploration ofbiochemistry, molecular biology, cell biology and systems biology because of the ease with which it can be grown and manipulated, the extensive conservation of its genes and pathways with those of higher organisms, and the powerful genetic techniques that it offers.

- The completion of the *S.cerevisiae* genomic DNA sequence in 1996 provided the sequence of each of its genes and currently represents the only complete sequence of a eukaryotic genome.

ARABIDOPSIS GENOME ANALYSIS:

The most up-to-date version of the *A. thaliana* genome is maintained by the Arabidopsis Information Resource (TAIR). Much work has been done to assign functions to its 27,000 genes and the 35,000 proteins they encode.

Initiation and progress

1983	First genetic map published
1988-1989	Publication of RFLP maps
1990	Multinational Coordinated Arabidopsis thaliana Genome project initiated
1991	First YAC libraries
1995- 1996	Standard BAC and P1 libraries constructed
1996	Arabidopsis Genome Initiative organised and started sequencing
1998	Physical maps of all chromosomes completed
1999	Sequence and analysis of chromosome 2 and 4
2000	Sequence and analysis of chromosome 1, 3 and 5
2000	Completion of whole genome sequencing

Salient Features

1. Small size plant (6–12 inch height).
2. Can be grown in Petri dish.
3. Life cycle (5–6 weeks).
4. Small genome compared to other plant genomes.
5. Large number of seeds per plant (10,000 per plant).
6. Easy to generate transgenic plants.

7. Genome contains much less repetitive DNA. Genome Size: Nucleus – 125 Mb, Plastid – 154 kb, Mitochondria – 367 kb. It has 12.1 million bases and 32 chromosomes

8. Well studied for light sensing and flower development.

9. Largest collections of mutants are available.

10. Translucent nature of the plant parts can be used to take fluorescent images to perform *in situ* analysis.

Genome of Arabidopsis thaliana

Salient Features

Arabidopsis genome sequence creates the potential for direct and efficientaccess to a much deeper understanding of plant development and environmental responses, and permits the structure and dynamics of plant genomes to be assessed and understood.

- One of the reasons *Arabidopsis* was chosen for complete sequencing was its relative lack of repeat sequences compared to other experimentally tractable plants.

The genomes of *C. elegans*, *Drosophila*, and *Arabidopsis* are intermediate in size and complexity between those of yeasts and humans. Distinctive features of each of these organisms make them important models for genome analysis:

Genome Organization

Genome organization refers to the sequential, not the structural organization of the genome. Besides the **coding exons**, the non-coding DNA in Eukaryotes may fall in the following classes

- **Introns**. They are DNA sequences inserted between the exons and found in theORF. They are spliced after the first level of transcription. Most introns are junk inserted within genes.
- **Pseudogenes**. 'Dead', non-functional copies of genes present elsewhere in thegenome, but no longer of any use.
- **Retropseudogenes**. Like pseudogenes, but have been processed, i.e. lackintrons. Produced by the action of reverse transcriptase (RT) on mRNA, and subsequent incorporation of the cDNA into the genome.
- **Transposons.** Jumping genes, which splice themselves in and out of the genome(in DNA form) randomly, by the action of transposase.
- **Retrotransposons**. Transcribed into an mRNA, which encodes an RT enzyme,which then copies the mRNA back to DNA and incorporates it into the genome.

Infact in humans only 1.5% of the entire genome length corresponds to coding DNA. This 1.5% codes for about 27,000 genes which in turn code for proteins that are responsible for all the cellular processes

Human Genome Organisation

Understanding the human genome will help us to improve human health and sequenced by Human Genome Organisation (HUGO) and Celera Genomics. It contains 3.2 billion bases and 23 chromosomes.

Salient Features

- The human genome contains 3164.7 million nucleotide bases.
- 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 30,000–much lower than previous estimates of 80,000 to 1,40,000 genes. Almost all (99.9 per cent) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50 per cent of discovered genes.
- Less than 2 per cent of the genome codes for proteins.
- Repeated sequences make up very large portion of the human genome.
- Repetitive sequences are stretches of DNA sequences that are repeated many times, sometimes hundred to thousand times. They are thought to have no direct coding functions, but they shed light on chromosome structure, dynamics and evolution.
- Chromosome 1 has most genes (2968), and the Y has the fewest (231).
- 99.9% of the nucleotide bases are exactly similar in all human beings. Only 0.1% of human genome with some 3.2 million nucleotides represents the variability observed in human beings.
- Repeated or repetitive sequences make up a large portion of human genome. There are some 30,000 minisatellite loci, each having 11 -60 bp repeated tandemly upto thousand times. These are about 2,00,000 microsatellites, each with upto 10 bp repeated 10-100 times.
- Approximately 1 million copies of short 5-8 base pair repeated sequences are clustered around centromeres and near the ends of chromosomes. They represent junk DNA.
- Scientists have identified about 1.4 million locations where single- base DNA differences (SNPs– single nucleotide polymorphism ,pronounced as ‘snips’) occur in humans. This information promises to revolutionise the processes of finding chromosomal locations for disease-associated sequences and tracing human history.

Applications and Future Challenges:

1. Disorders:

More than 1200 genes are responsible for common human cardiovascular diseases, endocrine diseases (like diabetes), neurological disorders (like Alzheimer's disease), cancers and many more.

2. Cancers:

Efforts are in progress to determine genes that will change cancerous cells to normal.

3. Health Care:

It will indicate prospects for a healthier living, designer drugs, genetically modified diets and finally our genetic identity.

4. Interactions:

It will be possible to study how various genes and proteins work together in an interconnected network.

5. Study of Tissues.

All the genes or transcripts in a particular tissue, organ or tumour can be analysed to know the cause of effect produced in it.

6. Nonhuman Organisms:

Information about natural capabilities of nonhuman organisms can be used in meeting challenges in health care, agriculture, energy production and environmental remediation. For this a number of model organisms have been sequenced, e.g., bacteria,

yeast Coenorhabditis elegans (free living non-pathogenic nematode), Drosophila (fruitfly), Rice, Arabidopsis.

***E.coli* genome organisation**

Physical Characteristics of the *E. coli* genome:

- single chromosome/cell (haploid)
- 4.6×10^6 bp (4600 kilobases)
    - about 4300 potential coding sequences
    - only about 1800 known *E. coli* proteins
- 70% is composed of single (monocistronic) genes
- 6% is polycistronic
- Roughly equal number of genes on each strand

About 30% of the sequenced ORF's (Open Reading Frames, areas that look like they could be the start points of transcription) have unknown function.

Proteomics

2-D gel electrophoresis

Two-dimensional gel electrophoresis (2-D electrophoresis) is a powerful and widely used method for the analysis of complex protein mixtures extracted from cells, tissues, or other biological samples. This technique expands the number of proteins that

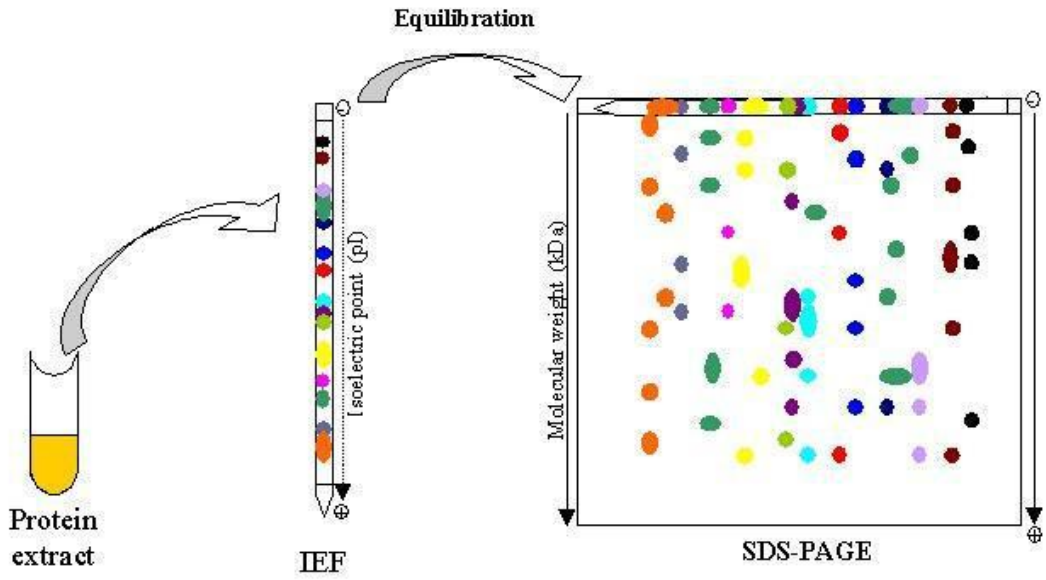
could be identified, provides more efficient data and detailed information for proteomics analysis.

This technique separate proteins in two steps, according to two independent properties: the first-dimension is isoelectric focusing (IEF), which separates proteins according to their isoelectric points (pI); the second-dimension is SDS-polyacrylamide gel electrophoresis (SDS-PAGE), which separates proteins according to their molecular weights (MW). In this way, complex mixtures consisted of thousands of different proteins can be resolved and the relative amount of each protein can be determined.

The procedure involves placing the sample in gel with a pH gradient, and applying a potential difference across it. In the electrical field, the protein migrates a long the pH gradient, until it carries no overall charge. This location of the protein in the gel constitutes the apparent pI of the protein.

There are two alternatives methods to create the pH gradient - carrier ampholites and immobilized pH gradient (IPG) gels. The IEF is the most critical step of the 2-D electrophoresis process. The proteins must be solubilize without charged detergents, usually in high concentrated urea solution, reducing agents and chaotrophs. To obtain high quality data it is essential to achieve low ionic strength conditions before the IEF itself. Since different types of samples differ in their ion content, it is necessary to adjust the IEF buffer and the electrical profile to each type of sample. After 2-DE analysis, a gel with proteins spread out based on their pI and molecular mass is obtained. These proteins can then be detected by a variety of means, but the most commonly used stains are silver and coomassie brilliant blue staining. In the former case, a silver colloid is applied to the gel. The silver binds to cysteine groups within the protein. The silver is darkened by exposure to ultra-violet light. The amount of silver can be related to the darkness, and therefore the amount of protein at a given location on the gel. This measurement can only give approximate amounts, but is adequate for most purposes. Other staining method such as coomassie brilliant blue combined with in-gel digestion is suitable for MS detection afterwards.

The separation in the second dimension by molecular size is performed in slab SDS- PAGE. Twelve parallel gels can be separated in a fixed temperature to minimize the separation variations between individual gels.



In two-dimensional gel electrophoresis, proteins are first separated by their pI in isoelectric focusing and then further separated by molecular weight through SDS-PAGE, thus the sample proteins are distributed across the two-dimensional gel profile.

2-DE is a powerful and widely used method for the analysis of complex protein mixtures extracted from cells, tissues, or other biological samples. This technique sorts proteins according to two independent properties in two discrete steps. Each spot on the resulting two-dimensional array corresponds to a single protein species in the sample. Thus thousands of different proteins can be separated, and information including the protein pI, the apparent molecular weight, and the amount of each protein is obtained.

Mass Spectrometry

Mass spectrometry is an analytical technique in which samples are ionized into charged molecules and ratio of their mass-to-charge (m/z) can be measured. In MALDI-TOF mass spectrometry, the ion source is matrix-assisted laser desorption/ionization (MALDI), and the mass analyzer is time-of-flight (TOF) analyzer.

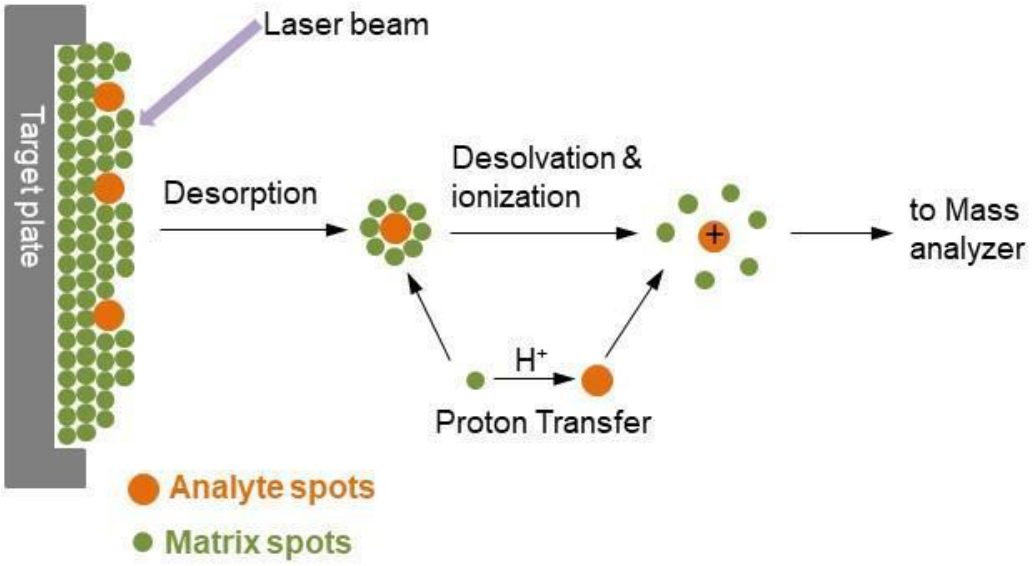
MALDI

MALDI is the abbreviation for "Matrix Assisted Laser Desorption/Ionization."

The sample for MALDI is uniformly mixed in a large quant it y o f matrix.

The matrix absorbs the ultraviolet light (nitrogen laser light, wavelengt h 337 nm) and converts it to heat energy. A small part of the matrix (down to 100 nm from the top outer surface of the Analyte in the diagram) heats rapidly (in several nano seconds) and is vaporized, together wit h the sample.

MALDI is a soft ionization that involves a laser striking a matrix of small molecules to make the analyte molecules into the gas phase without fragmenting or decomposing them. Some biomolecules are too large and can decompose when heated, and traditional techniques will fragment or destroy macromolecules. MALDI is appropriate to analyze biomolecules like peptides, lipids, saccharides, or other organic macromolecules.



The analyte is embedded in a very large excess of a matrix compound deposited on a solid surface called a target, usually made of a conducting metal and having spots for several different samples to be applied. After a very brief laser pulse, the irradiated spot is rapidly heated and becomes vibrationally excited. The matrix molecules energetically ablated from the surface of the sample, absorb the laser energy and carry the analyte molecules into the gas phase as well. During the ablation process, the analyte molecules are usually ionized by being protonated or deprotonated with the nearby matrix molecules. The most common MALDI ionization format is for analyte molecules to carry a single positive charge.

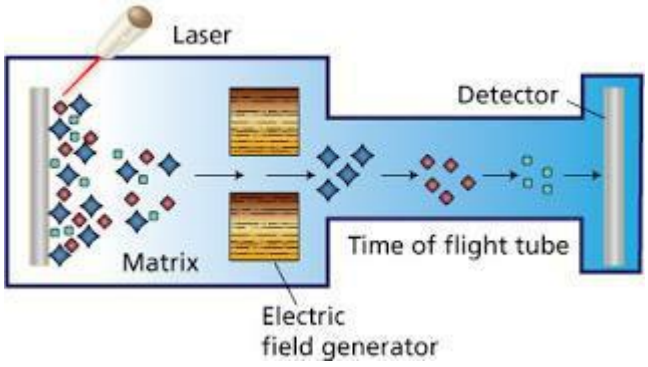
- **Types of laser commonly used in MALDI**

Lasers of both ultraviolet (UV) and infrared (IR) wavelengths are in use, but UV lasers are by far the most important light sources in analytical MALDI. Among these, nitrogen lasers and frequency-tripled or quadrupled Nd: Yag lasers often serve for the majority of applications. IR-MALDI is dominated by Er:Yag lasers while TEA-CO2 lasers are rarely used.

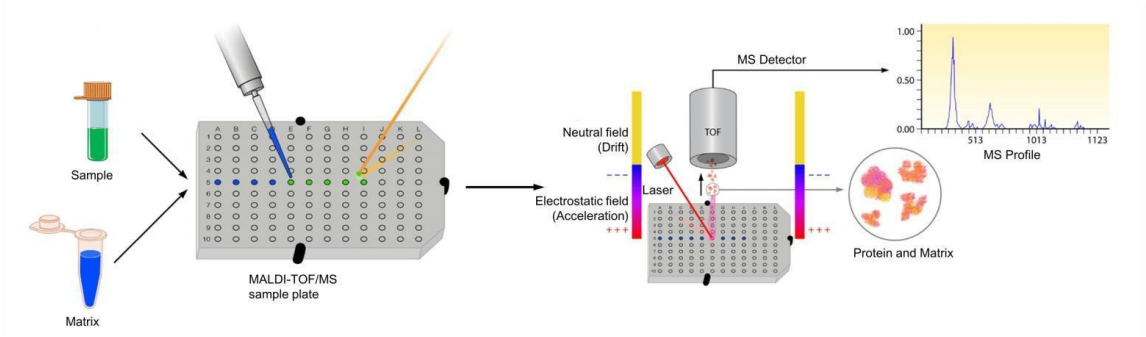
- **Commonly used MALDI matrix substance**

It is believed that the first function of the matrix essentially is to dilute and isolate analyte molecules from each other. This occurs during solvent evaporation and concomitant formation of a solid solution. Then, upon laser irradiation, it functions as a mediator for energy absorption. The choice of the right matrix is key to the success in MALDI. In general, highly polar analytes work better with highly polar matrices, and nonpolar analytes are preferably combined with nonpolar matrices. Currently, the most commonly used matrixes are α-cyano-4-hydroxycinnamic acid, 2,5-dihydroxybenzoic acid, 3,5-dimethoxy-4-hydroxycinnamic acid, and 2,6-dihydroxyacetophenone.

What is TOF MS?



TOF MS is the abbreviation for Time of Flight Mass Spectrometry. Charged ions of various sizes are generated on the sample slide, as shown in the diagram. A potential difference V_0 between the sample slide and ground attracts the ions in the direction shown in the diagram. The velocity of the attracted ions v is determined by the law of conservation of energy. As the potential difference V_0 is constant with respect to all ions, ions with smaller m/z value (lighter ions) and more highly charged ions move faster through the drift space until they reach the detector. Consequently, the time of ion flight differs according to the mass-to-charge ratio (m/z) value of the ion. The method of mass spectrometry that exploits this phenomenon is called Time of Flight Mass Spectrometry.



Application of MALDI-TOF mass spectrometry

- **Intact Mass determination**

The intact mass determination is basic and important for protein characterization, due to the correct molecular weight of a protein can indicate the intact structure. MALDI, a soft ionization technique, is suitable for proteins which tend to be fragile and fragment when ionized by other ionization methods. The performance of MALDI-TOF MS is less affected by buffer components, detergents, and contaminants. In addition, it permits intact protein mass determination with sufficient accuracy (≤ 500 ppm) for sequence validation. After protein digestion, MALDI-TOF MS can be also used to analyze the obtained peptides for further primary sequence confirmation by peptide mass fingerprinting.

- **Peptide mass fingerprinting (PMF)**

MALDI-TOF mass spectrometry has simple operation, good mass accuracy, as well as high resolution and sensitivity. Therefore, it has widespread uses in proteomics to identify proteins from simple mixtures by a method called peptide mass fingerprinting, which are often used with two-dimensional gel electrophoresis (2-DE). In this approach, peptides are generated by digesting proteins of interest with a sequence-specific enzyme like trypsin. And then peptides are analyzed by MALDI-TOF mass spectrometry to get the peptide masses. The experimental masses are compared against a database containing theoretical peptide masses from a given organism with the same sequence-specific protease.

- **Post source decay (PSD) MALDI-TOF analysis**

MALDI-TOF mass spectrometers equipped with reflectrons can analyze fragment ions produced from precursor ions that spontaneously decompose in the flight. Such ions are generally referred to as metastable ions, and the process of decomposition in the field free region between the ion source and the reflectron is commonly referred to as PSD. PSD fragment ions are formed within the field free region before entering the reflectron. PSD fragment ions can be separated, collected, and recorded on the detector by continuously changing the reflector voltage to form a PSD mass spectrum that provides very rich and effective structural information for the primary structure of peptides and proteins. In the proteomics study, some 2DE-separated protein samples cannot be identified by PMF or the results of identification are not clear. The PSD sequencing function can be applied to the identification of these proteins. Using PSD spectroscopy, combined with a database search, proteins can be identified quickly and with high specificity.

- **Oligonucleotides analysis**

With the development of molecular biology techniques and antisense nucleic acid drug technologies, more and more oligonucleotide fragments have been synthesized to be used as primers, probes and antisense drugs. It is entirely necessary to quickly detect these fragments to determine whether the synthesis is complete and whether the synthesized sequence is correct. Mass spectrometry, including MALDI-TOF-MS, is by far the best means of doing this. Oligonucleotide analysis using MALDI-TOF-MS was simple, rapid, accurate, and sensitive, which can be used to determine the complete oligonucleotide sequence.

- **MALDI imaging**

The MALDI-TOF can be used in profiling and imaging proteins directly from thin tissue sections, known as MALDI imaging mass spectrometry (MALDI-IMS). It provides specific information about the local molecular composition, relative abundance

and spatial distribution of peptides and proteins in the analyzed section. MALDI-IMS can analyze multiple unknown compounds in biological tissue sections simultaneously through a single measurement that can obtain molecule imaging of the tissue while maintaining the integrity of cells and molecules in tissues.

MALDI-TOF mass spectrometry can analyze a wide variety of biomolecules, such as peptides, proteins, carbohydrate, oligonucleotide. Due to the fact that formed ions have low internal energy, a great advantage of MALDI-TOF is that the process of soft-ionization enables observation of ionized molecules with little to fragmentation of analytes, allowing the molecular ions of analytes to be identified, even within mixtures. It is easy to use and maintain with fast data acquisition. Choosing the appropriate matrix substance is important for successful MALDI-TOF mass spectrometry.

Review Questions:

Short Answer Questions **(2 Marks)**

1. Define genome?
2. Define transcriptome?
3. Define Proteome?
4. Explain Human genome project?
5. Define mass spectrometry?
6. Difference between viral and prokaryotic genome?
7. Define C-value paradox?
8. Define : MALDITOF
9. Explain base composition of DNA?
10. Define Oligonucleotides
11. What is retrotransposons?
12. What is SDS?
13. Define PAGE?
14. Advantages of genome sequencing?
15. Give an example for sequenced viral and prokaryotic genome?
16. Give an account of *Arabidopsis thalliana*?
17. What is Bakers yeast
18. Life cycle of *A.thalliana*

Essay Answer Questions **(8 Marks)**

1. Describe Human genome project and its advantages
2. Explain MALDI TOF in detail
3. Describe 2-D gel electrophoresis with its principle and applications
4. Describe Eukaryotic genome organisation
5. Explain Viral genome organisation
6. Major features of *E.coli* genome
7. Major features of *S.cervisiae* genome

Unit IV Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
The first proposal of the human genome project was made by	Winfield	David	Churchill	R.Sinsheimer	R.Sinsheimer
NCBI was established in the year	1988	2001	2000	2017	1988
National Institute of health is abbreviated as	NIH	NH	INH	NI	NIH
Craig Venter established	Institute for Genomic Research	Institute for Taxonomic Research	Institute for Mitochindrial Research	Institute for Protein Research	Institute for Genomic Research
Craig Venter invented ……. Technology	EST	TSE	ETS	STE	EST
Haemophilius influenzae genome was published in the year	1992	1993	1994	1995	1995
E.coli genome was published in the year	1997	1995	1996	1999	1997
S.cerevisiae genome was published in the year	1996	1994	1887	2017	1996
Genome of C.elegans was published in the year	2016	2000	1998	1999	1998
Bakers yeast genome was sequenced in the year	1998	1999	1995	2000	1998
Genome of P.aeruginosa was published in the year	2000	2004	2007	2008	2000

Genome of A.thaliana was published in the year	1975	2000	2004	2006	2000
D.melanogaster sequence was published In the year	1888	1987	2000	2008	2000
E.coli genome contains ……protein coding genes	4284	7098	6543	6666	4284
E.coli genome contains ……structural RNA genes	111	122	133	144	122
The first archeal genome to be sequenced was	E.coli	Virus	M.jannaschii	Nematodes	M.jannaschii
M.jannaschii contains …… number of genes	1745	1754	1888	1743	1743
The simplest organism genome is	M.genitalium	C.elegans	HIV virus	Nematodes	M.genitalium
First plant genome sequenced was	Arabidopsis	Rice	Wheat	Sugarcane	Arabidopsis
The scientific endeavours that aim to determine the complete genome sequence of an organism	Genome project	Organelle project	Proteomics	Metabolomics	Genome project
The approximate number of genes in Human genome	10000-15000	15000-20000	20000-25000	25000-50000	25000-50000
Sequence database	Watson	crick	Sanger	Wilbur Lipman	Wilbur Lipman

searching algorithm was developed by					
Expansion for EST	Expressed sequence tag	Expressed short tag	Enriched sequence tag	Enriched short tag	Expressed sequence tag
Rice genome sequences released at	1990	1995	2000	2005	2005
SSS means	Sequence Search Services	Structure Search Services	Similarity Search Services	Smart Search Services	Sequence Search Services
Working draft of human genome was releases in	1995	2000	2005	2010	2000
Complete draft of human genome was released in	1995	2000	2003	2005	2003
Human genome project initiated by	Francis collins	Ari Patrinos	Sanger	Wilbur Lipman	Ari Patrinos
Human genome project directed by	Francis collins	Ari Patrinos	Sanger	Wilbur Lipman	Francis collins
Number of nucleotides in Human haploid reference genome	1 billion	2 billion	3 billion	4 billion	3 billion
EMBL created at	1990	1992	1994	1996	1994
Celera group used the method for human genome sequencing	shot gun	genome walking	cDNA library	ESTs	shot gun
BLAST was introduced during	1980	1990	2000	2010	1990
Atlas of Protein sequences developed by	Margaret Dayhoff's	Owen White	Francis collins	Ari Patrinos	Margaret Dayhoff's
IHGSC means	International	Indian Human	Italian Human	Internal Human	International

	Human Genome sequencing consortium	Genome sequencing consortium	Genome Sequencing consortium	Genome Sequencing Consortium	Human Genome sequencing consortium
The percentage of protein coding regions in human genome is	1.1 to 1.4%	2.0 to 2.5%	3.0 to 4%	5.0 to 6.0%	1.1 to 1.4%
Human genome project was initiated by	NIH and DOE	NIH and DDBJ	NIH	EMBL	NIH and DOE
The largest gene in humans is	Immunoglobin	Dystrophin	Heamoglobin	Antibody	Dystrophin
According to HGP geentic similarity of all humans is	99%	95%	99.99%	98%	99.99%
The first draft of HGP was published in the journal	Science	Cell	PloS Biology	Nature	Nature
The private company involved in Human genome sequencing in parallel with HGP is	IBM	TATA	HCL	Celera	Celera
HGP was also focussed on identifying	SNP	VNTR's	Satellite DNA	Junk DNA	SNP
All are genome sequencing strategies except	shot gun	Edman Degradation method	Directed gene sequencing	Whole genome short gun sequenicng	Edman Degradation Method
The term genomics was coined by	Abraham Christophen	Elizabeth Angel	Winston Churchill	Thomas Roder	Thomas Roder
Chromosomal and linkage maps are	family data	RFLP	positional Cloning	Sequenceing	family data

generated by					
The human genome project began as researchers mapped _____ and sites of cytogenetic abnormalities.	RFLP's	lods	PCR	VNTR's	RFLP's
In the Sanger method of DNA sequencing, DNA synthesis ____ when a dideoxy base is encountered.	Commences	Stops	Continue	Increases	Stops
Approximately what proportion of the human genome is made up of repetitive DNA sequences?	20%	25%	15%	50%	50%
How many protein coding genes do human genome have	10 - 15,000	30 - 40,000	More than 100,000	20 - 25,000	20 - 25,000
How many chromosomes do humans have	48	44	44	46	46
Genes are made up of	DNA	RNA	Proteins	Enzymes	DNA
The human genome is	Responsible for all our physical characteristics	All of our genes	All of the DNA and RNA in our cells	All of our DNA	All of our DNA

Working draft of human genome was releasedin	1999	2000	2001	2004	2000

Unit V – Protein Structure Predictions

Hierarchy of protein structure - primary, secondary and tertiary structures, modeling. Structural Classes, Motifs, Folds and Domains.Protein structure prediction in presence and absence of structure template Energy minimizations and evaluation by Ramachandran plot Protein structure and rational drug design.

Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence — that is, the prediction of its folding and its secondary, tertiary, and quaternary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes).There are three major theoretical methods for predicting the structure of proteins: comparative modelling, fold recognition, and ab initio prediction.

Comparative modelling:

Comparative modelling exploits the fact that evolutionarily related proteins with similar sequences, as measured by the percentage of identical residues at each position based on an optimal structural superposition, have similar structures.

Fold recognition or "threading":

Threading uses a database of known three-dimensional structures to match sequences without known structure with protein folds. This is accomplished by the aid of a scoring function that assesses the fit of a sequence to a given fold.

Ab initio prediction:

The ab initio approach is a mixture of science and engineering. The science is in understanding how the three-dimensional structure of proteins is attained. The engineering portion is in deducing the three-dimensional structure given the sequence.

Protein structural bioinformatics

Protein Structure Basics

Proteins perform most essential biological and chemical functions in a cell.

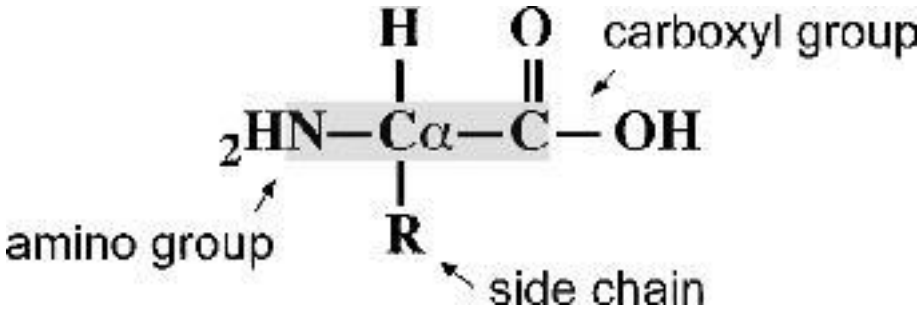
They play important roles in structural, enzymatic, transport, and regulatory functions.

The protein functions are strictly determined by their structures.

Therefore, protein structural bioinformatics is an essential element of bioinformatics.

AMINO ACIDS

The building blocks of proteins are twenty naturally occurring amino acids, small molecules that contain a free amino group (NH2) and a free carboxyl group (COOH). Both of these groups are linked to a central carbon (Cα), which is attached to a hydrogen and a side chain group (R) (Fig. 12.1). Amino acids differ only by the side chain R group.



The chemical reactivities of the R groups determine the specific properties of the amino acids.

Amino acids can be grouped into several categories based on the chemical and physical properties of the side chains, such as size and affinity for water.

According to these properties, the side chain groups can be divided into small, large, hydrophobic, and hydrophilic categories.

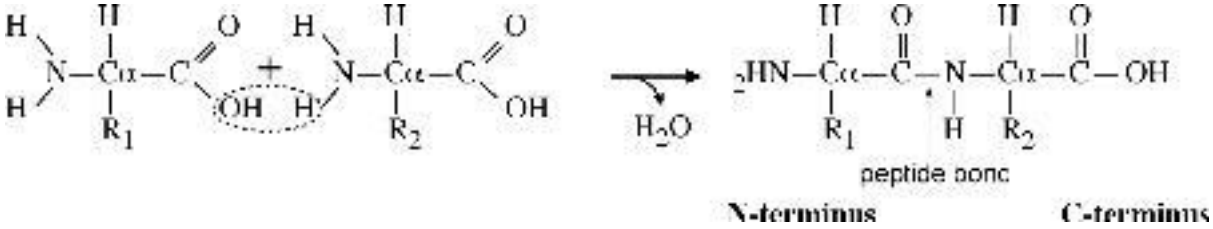
Within the hydrophobic set of amino acids, they can be further divided into aliphatic and aromatic.

Aliphatic side chains are linear hydrocarbon chains and *aromatic side chains* are cyclicrings.

Within the hydrophilic set, amino acids can be subdivided into polar and charged. *Charged amino acids* can be either positively charged (basic) or negatively charged(acidic).

PEPTIDE FORMATION

The peptide formation involves two amino acids covalently joined together between the carboxyl group of one amino acid and the amino group of another (shown in Figure).



This reaction is a condensation reaction involving removal of elements of water from the two molecules. The resulting product is called a *dipeptide.*

The newly formed covalent bond connecting the two amino acids is called a *peptidebond.* Once an amino acid is incorporated into a peptide, it becomes an amino acidresidue. Multiple amino acids can be joined together to forma longer chain of amino acid polymer.

A linear polymer of more than fifty amino acid residues is referred to as a *polypeptide.* A polypeptide, also called a protein, has a well-defined three-dimensional arrangement. On the other hand, a polymer with fewer than fifty residues is usually called a peptide without a well-defined three-dimensional structure.

The residues in a peptide or polypeptide are numbered beginning with the residue containing the amino group, referred to as the *N*-terminus, and ending with the residue containing the carboxyl group, known as the *C*-terminus.

DIHEDRAL ANGLES

A peptide bond is actually a partial double bond owing to shared electrons between O=C–N atoms.

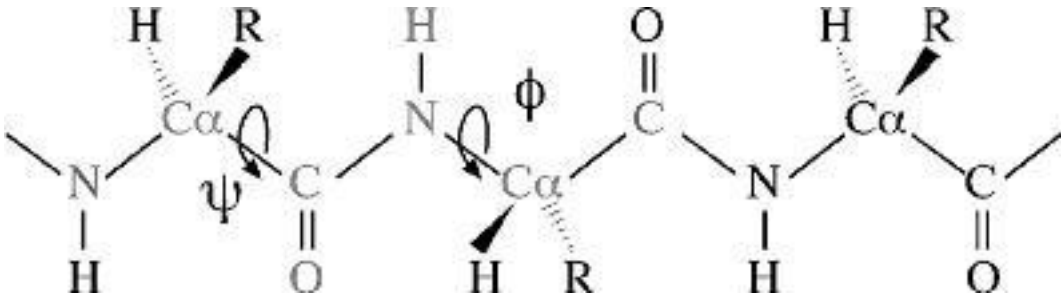
The rigid double bond structure forces atoms associated with the peptide bond to lie in the same plane, called the *peptide plane*.

Because of the planar nature of the peptide bond and the size of the R groups, there are considerable restrictions on the rotational freedom by the two bonded pairs of atoms around the peptide bond.

The angle of rotation about the bond is referred to as the *dihedral angle* (also called the *tortional angle*).

For a peptide unit, the atoms linked to the peptide bond can be moved to a certain extent by the rotation of two bonds flanking the peptide bond.

This is measured by two dihedral angles (as shown in Figure).



One is the dihedral angle along the N–Cα bond, which is defined as phi (φ); and the other is the angle along the Cα–C bond, which is called psi (ψ).

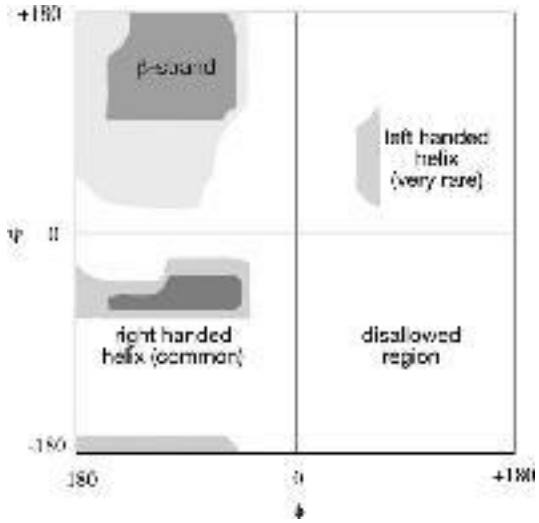
Various combinations of φ and ψ angles allow the proteins to fold in many different ways.

Ramachandran Plot

The rotation of φ and ψ is not completely free because of the planar nature of the peptide bond and the steric hindrance from the side chain R group. Consequently, there is only a limited range of peptide conformation.

When φ and ψ angles of amino acids of a particular protein are plotted against each other, the resulting diagram is called a Ramachandran plot.

This plot maps the entire conformational space of a peptide and shows sterically allowed and disallowed regions (as shown in Figure).



It can be very useful in evaluating the quality of protein models.

Protein structures

can be organized into four levels of hierarchies with increasing complexity.

These levels are

[1] primary structure,

[2] secondary structure,

[3] tertiary structure,

[4] quaternary structure.

(1) Primary structure:

A linear amino acid sequence of a protein is the primary structure.

This is the simplest level with amino acid residues linked together through peptide bonds.

(2) Secondary structure:

defined as the local conformation of a peptide chain.

The secondary structure is characterized by highly regular and repeated arrangement of amino acid residues stabilized by hydrogen bondsbetween main chain atoms of the C=Ogroup and the NH group of different residues.

(3) Tertiary structure:

which is the three dimensional arrangement of various secondary structural elements and connecting regions.

The tertiary structure can be described as the complete three-dimensional assembly of all amino acids of a single polypeptide chain.

(4) Quaternary structure:

which refers to the association of several polypeptide chains into a protein complex, which is maintained by noncovalent interactions.

In such a complex, individual polypeptide chains are called *monomers* or *subunits.* Intermediate between secondary and tertiary structures, a level of supersecondary structure is

often used, which is defined as two or three secondary structural elements forming a unique functional domain, a recurring structural pattern conserved in evolution.

SECONDARY STRUCTURES

As mentioned, local structures of a protein with regular conformations are known as secondary structures.

They are stabilized by hydrogen bonds formed between carbonyl oxygen and amino hydrogen of different amino acids.

Chief elements ofsecondary structures are α-helices and β-sheets.

Useful rules in guiding the prediction of protein secondary structures.

For α-Helices

- ✓ Anα-helix has a main chain backbone conformation that resembles a cork screw.
- ✓ Nearly all known α-helices are right handed, exhibiting a rightward spiral form.

- ✓ In such a helix, there are 3.6 amino acids per helical turn.
- ✓ The structure is stabilized by hydrogen bonds formed between the main chain atoms of residues *i* and *i* + *4*.
- ✓ The hydrogen bonds are nearly parallel with the helical axis.
- ✓ The average φ and ψ angles are 60◦ and 45◦, respectively, and are distributed in a narrowly defined region in the lower left region of a Ramachandran plot.
- ✓ Hydrophobic residues of the helix tend to face inside and hydrophilic residues of the helix face outside. Thus, every third residue along the helix tends to be a hydrophobic residue. Ala, Gln, Leu, and Met are commonly found in anα-helix, but not Pro, Gly, and Tyr.

For β-Sheets

- ✓ A β-sheet is a fully extended configuration built up from several spatially adjacent regions of a polypeptide chain.
- ✓ Each region involved in forming the β-sheet is a β-strand.
- ✓ The β-strand conformation is pleated with main chain backbone zigzagging and side chains positioned alternately on opposite sides of the sheet.
- ✓ β-Strands are stabilizedby hydrogen bonds between residues of adjacent strands.
- ✓ β-strands near the surface of the protein tend to show an alternating pattern ofhydrophobic and hydrophilic regions, whereas strands buried at the core of a protein are nearly all hydrophobic.
- ✓ The β-strands can run in the same direction to form a parallel sheet or can run every other chain in reverse orientation to form an antiparallel sheet, or a mixture of both.
- ✓ The hydrogen bonding patterns are different in each configurations.
- ✓ The φ and ψ angles are also widely distributed in the upper left region in a Ramachandran plot. Because of the long-range nature of residuesinvolved in this type of conformation, it is more difficult to predict β-sheets than α- helices.

Coils and Loops

- ✓ There are also local structures that do not belong to regular secondary structures (α-helices and β-strands).
- ✓ The irregular structures are coils or loops.
- ✓ The loops are often characterized by sharp turns or hairpin-like structures.
- ✓ If the connecting regions are completely irregular, they belong to random coils.
- ✓ Residues in the loop or coil regions tend to be charged and polar and located on the surface of the protein structure.
- ✓ They are often the evolutionarily variable regions where mutations, deletions, and insertions frequently occur.
- ✓ They can be functionally significant because these locations are often the active sites of proteins.

Coiled Coils

- ✓ Coiled coils are a special type of super secondary structure characterized by a bundle of two or more α-helices wrapping around each other.
- ✓ The helices forming coiled coils have a unique pattern of hydrophobicity, which repeats every seven residues (five hydrophobic and two hydrophilic).

Protein Secondary Structure Prediction (page 200 to 212)

- Protein secondary structures are stable local conformations of a polypeptide chain.

- They are critically important in maintaining a protein three-dimensional structure.

- The highly regular and repeated structural elements include α-helices and β-sheets.

- It has been estimated that nearly 50% of residues of a protein fold into either α-helices andβ-strands.

Protein secondary structure prediction

refers to the prediction of the conformational state of each amino acid residue of a protein sequence as one of the three possible states, namely, helices, strands, or coils, denoted as H, E, and C, respectively.

The prediction is based on the fact that secondary structures have a regular arrangement of amino acids, stabilized by hydrogen bonding patterns.

The structural regularity serves the foundation for prediction algorithms.

Applications of predicting protein secondary structures:

- It can be useful for the classification of proteins and for the separation of protein domains and functional motifs.
- Secondary structures are much more conserved than sequences during evolution. As a result, correctly identifying secondary structure elements (SSE) can help to guide sequence alignment or improve existing sequence alignment of distantly related sequences.
- In addition, secondary structure prediction is an intermediate step in tertiary structure prediction as in threading analysis.

SECONDARY STRUCTURE PREDICTION FOR GLOBULAR PROTEINS

The formation of α-helices is determined by short-range interactions, whereas the formation of β-strands is strongly influenced by long-range interactions. Prediction for long-range interactions is theoretically difficult.

After more than three decades of effort, prediction accuracies have only been improved from about 50% to about 75%.

The secondary structure prediction methods can be either

(1) ab initio based- which make use of single sequence information only,

(2) homology based - which make use of multiple sequence alignment information.

Ab initio methods,

which belong to early generation methods, predict secondary structures based on statistical calculations of the residues of a single query sequence.

Homology-based methods

do not rely on statistics of residues of a single sequence, but on common secondary structural patterns conserved among multiple homologous sequences.

Ab Initio–Based Methods

This type of method predicts the secondary structure based on a single query sequence. It measures the relative propensity of each amino acid belonging to a certain secondary structure element.

The propensity scores are derived from known crystal structures.

Examples of ab initio prediction are the

(i) Chou–Fasmanalgorithum

(ii) Garnier, Osguthorpe, Robson (GOR) methods.

The ab initio methods were developed in the 1970s when protein structural data were very limited.

(i) Chou–Fasman algorithm

(http://fasta.bioch.virginia.edu/fasta/chofas.htm)determines the propensity or intrinsic tendency of each residue to be in the helix, strand, and β-turn conformation using observed frequencies found in protein crystal structures (conformational values for coils are not considered).

(ii) The GOR method (http://fasta.bioch.virginia.edu/fasta www/garnier.htm)

is also based on the "propensity" of each residue to be in one of the four conformational states, helix (H), strand(E), turn(T),andcoil (C).

However, instead of using the propensity value from a single residue to predict a conformational state, it takes short-range interactions of neighboring residues into account.

It examines a window of every seventeen residues and sums up propensity scores for all residues for each of the four states resulting in four summed values.

The highest scored state defines the conformational state for the center residue in the window (ninth position).

The GOR method has been shown to be more accurate than Chou–Fasman because it takes the neighboring effect of residues into consideration.

Homology-Based Methods

The third generation of algorithms were developed in the late 1990s by making use of evolutionary information.

This type of method combines the ab initio secondary structure prediction of individual sequences and alignment information from multiple homologous sequences (>35% identity).

The idea behind this approach is that close protein homologs should adopt the same secondary and tertiary structure.

When each individual sequence is predicted for secondary structure using a method similar to the GOR method, errors and variations may occur.

However, evolutionary conservation dictates that there should be no major variations for their secondary structure elements.

Therefore, by aligning multiple sequences, information of positional conservation is revealed. Because residues in the same aligned position are assumed to have the same secondary structure, any inconsistencies or errors in prediction of individual sequences can be corrected using a majority rule.

This homology based method has helped improve the prediction accuracy by another 10% over the second-generation methods.

The following lists several frequently used third generation prediction algorithms available as web servers.

PHD(Profile network from Heidelberg;http://dodo.bioc.columbia.edu/predictprotein/submit def.html)

- is a web-based program that combines neural network with multiple sequence alignment.
- It first performs a BLASTP of the query sequence against a nonredundant protein sequence database to find a set of homologous sequences, which are

aligned with the MAXHOM program (a weighted dynamic programming algorithm performing global alignment).

PSIPRED(http://bioinf.cs.ucl.ac.uk/psiform.html)

- is a web-based program that predicts protein secondary structures using a combination of evolutionary information and neural networks.
- The multiple sequence alignment is derived from a PSI-BLAST database search.
- A profile is extracted from the multiple sequence alignment generated from three rounds of automated PSI-BLAST.

SSpro(http://promoter.ics.uci.edu/BRNN-PRED/)

- is a web-based program thatcombines PSI-BLAST profiles with an advanced neural network, known as *bidirectional recurrent neural networks* (BRNNs).
- is an algorithm that combines PSI-BLAST profiles and a multistaged neural network, similar to that in PHD.
- In addition, it uses a linear discriminant function to discriminate between the three states.

HMMSTR(Hidden Markov model [HMM] for protein STRuctures;www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php)

- uses a branched and cyclic HMM to predict secondary structures.
- It first breaks down the query sequence into many very short segments (three to nine residues, called I-sites) and builds profiles based on a library of known structure motifs.
- It then assembles these local motifs into a super secondary structure.
- It further uses an HMM with a unique topology linking many smaller HMMs into a highly branched multicyclic form
- combines the analysis results from six prediction algorithms, including PHD, PREDATOR, DSC, NNSSP, Jnet, and ZPred.

- The query sequence is first used to search databases with PSI-BLAST for three iterations. Redundant sequence hits are removed.
- The resulting sequence homologs are used to build a multiple alignment from which a profile is extracted.
- The profile information is submitted to the six prediction programs.
- If there is sufficient agreement among the prediction programs, the majority of the prediction is taken as the structure.

- Is another multiple prediction server that uses Jpred, PHD, PROF, and PSIPRED, among others.
- The difference is that the server does not run the individual programs but sends the query to other servers which e-mail the results to the user separately.
- It does not generate a consensus.
- It is up to the user to combine multiple prediction results and derive a consensus.

Protein Tertiary Structure Prediction

Structural prediction is a powerful tool to understand the functions of biological macromolecules at the atomic level.

DNA structure, a double helix, is invariable regardless of sequence variations.

Necessary for protein structure prediction:

- protein structures vary depending on the sequences.
- much slower rate of structure determination by x-ray crystallography or NMR spectroscopy compared to gene sequence generation from genomic studies.
- Consequently, the gap between protein sequence information and protein structural information is increasing rapidly. Protein structure prediction aims to reduce this sequence–structure gap.

- In contrast to sequencing techniques, experimental methods to determine protein structures are time consuming and limited.Currently, it takes 1 to 3 years to solve a protein structure.
- Certain proteins, especially membrane proteins, are extremely difficult to solve by x-ray or NMR techniques.
- There are many important proteins for which the sequence information is available, but their three-dimensional structures remain unknown.
- The full understanding of the biological roles of these proteins requires knowledge of their structures.

Therefore, it is often necessary to obtain approximate protein structures through computer modeling.

Methods of protein three-dimensional structure prediction:

There are three computational approaches to protein three-dimensional structural modeling and prediction.

- Homology modeling - knowledge-based methods
- Threading - knowledge-based methods
- Ab initio prediction.

In Knowledge-based methods - predict protein structures based on knowledge of existing protein structural information in databases.

Homology modeling

builds an atomic model based on an experimentally determined structure that is closely related at the sequence level.

Threading

identifies proteins that are structurally similar, with or without detectable sequence similarities.

Ab initio approach

is simulation based and predicts structures based on physicochemical principles governing protein folding without the use of structural templates.

Homology modelling or comparative modelling

Homology modeling predicts protein structures based on sequence homology withknown structures. It is also known as *comparative modeling*.

Principle

if two proteins share a high enough sequence similarity, they are likely to have very similar three-dimensional structures. If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence.

Homology modeling produces an all-atom model based on alignment with template proteins.

Homology modeling procedure consists of six steps:

First step - is template selection, which involves identification of homologous sequences in the protein structure database to be used as templates for modeling.

Second step - is alignment of the target and template sequences.

Third step - is to build a framework structure for the target protein consisting of main chain atoms.

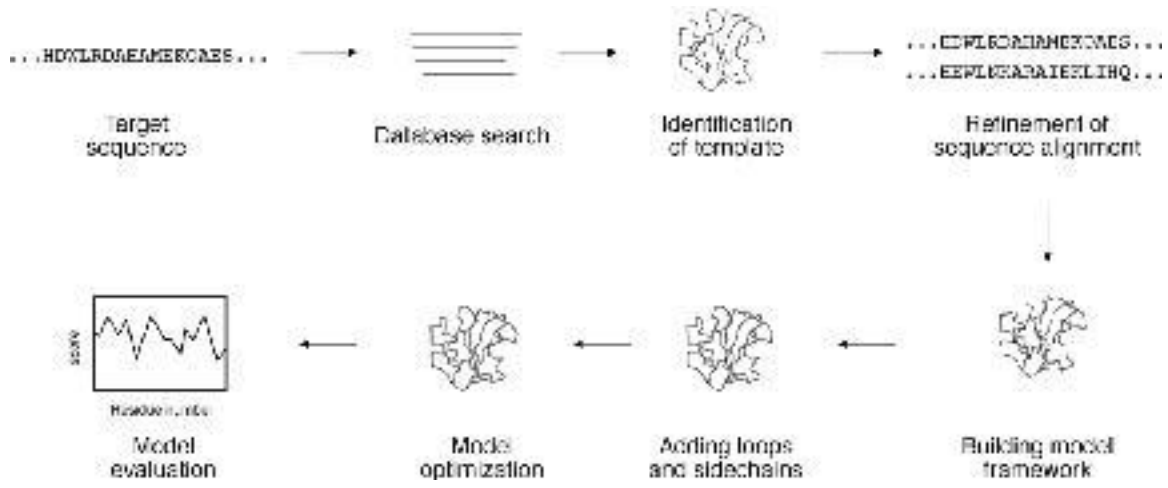
Fourth step - of model building includes the addition and optimization of side chain atoms and loops.

Fifth step - is to refine and optimize the entire model according to energy criteria.

Sixth step - involves evaluating of the overall quality of the model obtained.

If necessary, alignment and model building are repeated until a satisfactory result is obtained.

Flow chart showing steps involved in homology modelling:



Template Selection

- The first step in protein structural modeling is to select appropriate structural templates.
- This forms the foundation for rest of the modeling process.
- The template selection involves searching the Protein Data Bank (PDB) for homologous proteins with determined structures.
- The search can be performed using a heuristic pair-wise alignment search program such as BLAST or FASTA.

As a rule of thumb,

a database protein should have at least 30% sequence identity with the query sequence to be selected as template.

Thus it is recommended that the structure(s) with the highest percentage identity, highest resolution, and the most appropriate cofactors is selected as a template. If, no highly similar sequences can be found in the structure database,

- either a more sensitive profile-based PSI-BLAST method or

- a fold recognition method such threading can be used to identify distant homologs.

Modeling can therefore only be done with the aligned domains of the target protein.

Sequence Alignment

Once the structure with the highest sequence similarity is identified as a template,

- the full-length sequences of the template and target proteins need to be realigned using refined alignment algorithms to obtain optimal alignment.
- This realignment is the most critical step in homology modeling, which directly affects the quality of the final model.
- Errors made in the alignment step cannot be corrected in the following modeling steps.
- Therefore, the best possible multiple alignment algorithms, such as Praline and T-Coffee should be used for this purpose.
- Even alignment using the best alignment program may not be error free and should be visually inspected to ensure that conserved key residues are correctly aligned.
- If necessary, manual refinement of the alignment should be carried out to improve alignment quality.

Backbone Model Building

Once optimal alignment is achieved,

residues in the aligned regions of the target protein can assume a similar structure as the template proteins, meaning that the coordinates of the corresponding residues of the template proteins can be simply copied onto the target protein.

- If the two aligned residues are identical, coordinates of the side chain atoms are copied along with the main chain atoms.
- If the two residues differ, only the backbone atoms can be copied.
- The side chain atoms are rebuilt in a subsequent procedure.
- In backbone modeling, it is simplest to use only one template structure.

- As mentioned, the structure with the best quality and highest resolution is normally chosen if multiple options are available. This structure tends to carry the fewest errors.
- Occasionally, multiple template structures are available for modeling.
- In this situation, the template structures have to be optimally aligned and superimposed before being used as templates in model building.
- One can either choose to use average coordinate values of the templates or the best parts from each of the templates to model.

Loop Modeling

- In the sequence alignment for modeling, there are often regions caused by insertions and deletions producing gaps in sequence alignment.
- The gaps cannot be directly modeled, creating "holes" in the model.
- Closing the gaps requires loop modeling, which is a very difficult problem in homology modeling and is also a major source of error.
- Loop modeling can be considered a mini–protein modeling problem by itself.
- Unfortunately, there are no methods available that can model loops reliably.

Currently, there are **two main techniques** used to approach the problem:

(1) database searching method and (2) *ab initio* method.

(1) **database searching method**

The best loop can be selected based on sequence similarity as well as minimal steric clashes with the neighboring parts of the structure. The conformation of the best matching fragments is then copied onto the anchoring points of the stems.

(2) ***ab initio* method**

The ab initio method generates many random loops and searches for the one that does not clash with nearby side chains and also has reasonably low energy and φ and ψ angles in the allowable regions in the Ramachandran plot.

If the loops are relatively short (three to five residues), reasonably correct models can be built using either of the two methods. If the loops are longer, it is very difficult to achieve a reliable model.

The following are specialized programs for loop modeling:

FREAD (www-cryst.bioc.cam.ac.uk/cgi-bin/coda/fread.cgi) is a web server that modelsloops using the database approach.

PETRA (www-cryst.bioc.cam.ac.uk/cgi-bin/coda/pet.cgi) is a web server that uses theab initio method to model loops.

CODA (www-cryst.bioc.cam.ac.uk/~charlotte/Coda/search coda.html) is a web serverthat uses a consensus method based on the prediction results from FREAD and PETRA. For loops of three to eight residues, it uses consensus conformation of both methods and for nine to thirty residues, it uses FREAD prediction only.

Side Chain Refinement

- Once main chain atoms are built, the positions of side chains that are not modeled must be determined.
- Modeling side chain geometry is very important in evaluating protein–ligand interactions at active sites and protein–protein interactions at the contact interface.
- A side chain can be built by searching every possible conformation at every torsion angle of the side chain to select the one that has the lowest interaction energy with neighboring atoms. Most current side chain prediction programs use the concept of *rotamers*, which are favored side chain torsion angles extracted from known protein crystal structures.
- A collection of preferred side chain conformations is a rotamer library in which the rotamers are ranked by their frequency of occurrence.
- Having a rotamer library reduces the computational time significantly because only a small number of favored torsion angles are examined.

- In prediction of side chain conformation, only the possible rotamers with the lowest interaction energy with nearby atoms are selected.
- In many cases, even applying the rotamer library for every residue can be computationally too expensive.
- To reduce search time further, backbone conformation can be taken into account.
- It has been observed that there is a correlation of backbone conformations with certain rotamers.
- By using such correlations, many possible rotamers can be eliminated and the speed of conformational search can be much improved.
- After adding the most frequently occurring rotamers, the conformations have to be further optimized to minimize steric overlaps with the rest of the model structure.
- Most modeling packages incorporate the side chain refinement function.
- A specialized side chain modeling program that has reasonably good performance is SCWRL (sidechain placement with a rotamer library; www.fccc.edu/research/labs/dunbrack/scwrl/),
- It removes rotamers that have steric clashes with main chain atoms.
- The final, selected set of rotamers has minimal clashes with main chain atoms and other side chains.

Model Evaluation

- The final homology model has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules.
- This involves checking anomalies in φ–ψ angles, bond lengths, close contacts.
- Another way of checking the quality of a protein model is to implicitly take these stereochemical properties into account.
- This is a method that detects errors by compiling statistical profiles of spatial features and interaction energy from experimentally determined structures.

- By comparing the statistical parameters with the constructed model, the method reveals which regions of a sequence appear to be folded normally and which regions do not. If structural irregularities are found, the region is considered to have errors and has to be further refined.

Procheck(www.biochem.ucl.ac.uk/~roman/procheck/procheck.html) is a UNIXprogram that is able to check general physicochemical parameters such as φ–ψ angles, chirality, bond lengths, bond angles.

The parameters of the model are used to compare with those compiled from well-defined, high-resolution structures.

If the program detects unusual features, it highlights the regions that should be checked or refined further.

WHAT IF (www.cmbi.kun.nl:1100/WHATIF) is a comprehensive protein analysisserver that validates a protein model for chemical correctness. It has many functions, including checking of planarity, collisions with symmetry axes (close contacts), proline puckering, anomalous bond angles, and bond lengths. It also allows the generation of Ramachandran plots as an assessment of the quality of the model.

ANOLEA (Atomic Non-Local Environment Assessment;http://protein.bio.puc.cl/cardex/servers/anolea/index.html)is a web server that uses thestatistical evaluation approach. It performs energy calculations for atomic interactions in a protein chain and compares these interaction energy values with those compiled from a database of protein x-ray structures.

Verify3D (www.doe-mbi.ucla.edu/Services/Verify 3D/) is another server using thestatistical approach. It uses a precomputed database containing eighteen environmental profiles based on secondary structures and solvent exposure, compiled from high-resolution protein structures.

Comprehensive Modeling Programs

A number of comprehensive modeling programs are able to perform the complete procedure of homology modeling in an automated fashion.

Some freely available protein modeling programs and servers are listed.

Modeller(http://bioserv.cbs.cnrs.fr/HTML BIO/frame mod.html) is a web server forhomology modeling.

Swiss-Model (www.expasy.ch/swissmod/SWISS-MODEL.html) is an automatedmodeling server that allows a user to submit a sequence and to get back a structure automatically.

3D-JIGSAW (www.bmm.icnet.uk/servers/3djigsaw/) is a modeling server that works ineither the automatic mode or the interactive mode. Its loop modeling relies on the database method.

Homology Model Databases

The availability of automated modeling algorithms has allowed several research groups to use the fully automated procedure to carry out large-scale modeling projects.

ModBase(http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi) is a database of proteinmodels generated by the Modeller program. For most sequences that have been modeled, only partial sequences or domains that share strong similarities with templates are actually modeled.

3Dcrunch (www.expasy.ch/swissmod/SWISS-MODEL.html) is another databasearchiving results of large-scale homology modeling projects. Models of partial sequences from the Swiss-Prot database are derived using the Swiss-Model program.

What determines Fold?

- Anfinsen's experiments in 1957 demonstrated that proteins can fold spontaneously into their native conformations under physiological conditions. This implies that primary structure does indeed determine folding or 3-D stucture.
- Some exceptions exist
  - *Chaperone* proteins assist folding
  - Abnormally folded *Prion* proteins can catalyze misfolding of normal *prion* proteins that then aggregate

THREADING AND FOLD RECOGNITION

*threading*or *structural fold recognition* predicts the structural fold of an unknownprotein sequence by fitting the sequence into a structural database and selecting the best-fitting fold. The comparison emphasizes matching of secondary structures, which are most evolutionarily conserved. Therefore, this approach can identify structurally similar proteins even without detectable sequence similarity.

The algorithms can be classified into **two categories**,

pairwise energy based and profile based.

The pairwise energy–based method was originally referred to as *threading* and the profile-based method was originally defined as *fold recognition*.

Pairwise Energy Method

In the pairwise energy based method, a protein sequence is searched for in a structural fold database to find the best matching structural fold using energy-based criteria.

Profile Method

In the profile-based method, a profile is constructed for a group of related protein structures. The structural profile is generated by superimposition of the structures to expose corresponding residues. Statistical information from these aligned residues is then used to construct a profile.

The profile scores contain information for secondary structural types, the degree of solvent exposure, polarity, and hydrophobicity of the amino acids.

To predict the structural fold of an unknown query sequence, the query sequence is first predicted for its secondary structure, solvent accessibility, and polarity. The predicted information is then used for comparison with propensity profiles of known structural folds to find the fold that best represents the predicted profile.

3D-PSSM (www.bmm.icnet.uk/~3dpssm/) is a web-based program that employs thestructural profile method to identify protein folds.

GenThreader(http://bioinf.cs.ucl.ac.uk/psipred/index.html) is a web-based programthat uses a hybrid of the profile and pairwise energy methods.

Fugue (www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html) is a profile-based foldrecognition server. It has precomputed structural profiles compiled from multiple alignments of homologous structures, which take into account local structural environment such as secondary structure, solvent accessibility, and hydrogen bonding status.

Levels of Description of Structural Complexity

- **Primary Structure (AA sequence)**
- **Secondary Structure**
  - **Spatial arrangement of a polypeptide's backbone atoms withoutregard to side-chain conformations**
    - **α, β, coil, turns**
  - **Super-Secondary Structure**
    - **α, β, α/β, α+β**
- **Tertiary Structure**
  - **3-D structure of an entire polypeptide**
- **Quarternary Structure**
  - **Spatial arrangement of subunits (2 or more polypeptide chains)**

Techniques of Structure Prediction

- **Computer simulation based on energy calculation**
- **Based on physio-chemical principles**
- **Thermodynamic equilibrium with a minimum free energy**
- **Global minimum free energy of protein surface**
- **Knowledge Based approaches**
- **Homology Based Approach**
- **Threading Protein Sequence**
- **Hierarchical Methods**

AB INITIO PROTEIN STRUCTURAL PREDICTION

- Both homology and fold recognition approaches rely on the availability of template structures in the database to achieve predictions.
- If no correct structures exist in the database, the methods fail.
- However, proteins in nature fold on their own without checking what the structures of their homologs are in databases.
- Obviously, there is
- some information in the sequences that provides instruction for the proteins to

"find" their native structures. Early biophysical studies have shown that most proteins fold spontaneously into a stable structure that has near minimum energy. This structural state is called the *native state*. This folding process appears to be non random; however, its mechanism is poorly understood.

- The limited knowledge of protein folding forms the basis of ab initio prediction.
- The ab initio prediction method attempts to produce all-atom protein models based on sequence information alone without the aid of known protein structures.

The following web program is such an example using this approach:

Rosetta (www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php) is a web server that predictsprotein three-dimensional conformations using the ab initio method. This relies on a "mini-threading" method. The method first breaks down the query sequence into many very short segments (three to nine residues) and predicts the secondary structure of the small segments using a hidden Markov model–based program, HMMSTR.

Protein Structure Visualization:

- Once a protein structure has been solved, the structure has to be presented in a three dimensional view on the basis of the solved Cartesian coordinates.
- Before computer visualization software was developed, molecular structures were represented by physical models of metal wires, rods, and spheres.

- With the development of computer hardware and software technology, sophisticated computer graphics programs have been developed for visualizing and manipulating complicated three-dimensional structures.
- The computer graphics help to analyze and compare protein structures to gain insight to functions of the proteins.

- interactivity, which allows users to visually manipulate the structural images through a graphical user interface.

At the touch of a mouse button, a user can move, rotate, and zoom an atomic model on a computer screen in real time, or examine any portion of the structure in great detail, as well as draw it in various formsin different colors.

- Further manipulations can include changing the conformation of a structure by protein modeling or matching a ligand to an enzyme active site through docking exercises.

- The visualization program should also be able to produce molecular structures in different styles, which include wire frames, balls and sticks, space-filling spheres, and ribbons.

Wire-frame diagram

is a line drawing representing bonds between atoms.

The wire frame is the simplest form of model representation and is useful for localizing positions of specific residues in a protein structure, or for displaying a skeletal form of a structure when Cα atoms of each residue are connected.

Balls and sticks

are solid spheres and rods, representing atoms and bonds, respectively.

These diagrams can also be used to represent the backbone of a structure.

Space-filling representation

each atom is described using large solid spheres with radii corresponding to the van der Waals radii of the atoms.

Ribbon diagrams

use cylinders or spiral ribbons to represent α-helices and broad, flat arrows to represent β-strands. This type of representation is very attractive in that it allows easyidentification of secondary structure elements and gives a clear view of the overall topology of the structure.

Some widely used and freely available software programs for molecular graphics:

RasMol(http://rutgers.rcsb.org/pdb/help-graphics.html#rasmol download)

- is a command-line–based viewing program that calculates connectivity of a coordinate file and displays wireframe, cylinder, stick bonds, α-carbon trace, space-filling (CPK) spheres, and ribbons.
- It reads both PDB and mmCIF formats and can display a whole molecule or specific parts of it.
- It is available in multiple platforms: UNIX, Windows, and Mac.
- is a new version of RasMolforWindows with a more enhanced user interface.
- is a structure viewer for multiple platforms.
- It is essentially a Swiss-Army knife for structure visualization and modeling because it incorporates so many functions in a small shareware program.
- It is capable of structure visualization, analysis, and homology modeling.
- It allows display of multiple structures at the same time in different styles, by charge distribution, or by surface accessibility.
- It can measure distances, angles, and even mutate residues.
- In addition, it can calculate molecular surface, electrostatic potential, Ramachandran plot, and so on.
- The homology modeling part includes energy minimization and loop modeling.

Molscript(www.avatar.se/molscript/)

- is a UNIX program capable of generating wire-frame, space-filling, or ball-and-stick styles. In particular, secondary structure elements can be drawn with solid spirals and arrows representing α-helices and β-strands, respectively.

Grasp(http://trantor.bioc.columbia.edu/grasp/)

- is a UNIX program that generates solid molecular surface images and uses a gradated coloring scheme to display electrostatic charges on the surface. **WebMol**(www.cmpharm.ucsf.edu/cgi-bin/webmol.pl)

- is a web-based program built based on a modified RasMol code and thus shares many similarities with RasMol.
- It runs directly on a browser of any type as an applet and is able to display simple line drawing models of protein structures.
- It also has a feature of interactively displaying Ramachandran plots for structure model evaluation.
- Is a plug-in for web browsers; it is not a stand alone program and has to be invoked in a web browser.
- The program is also derived from RasMol and allows interactive display of graphics of protein structures inside a web browser.

Cn3D(www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml)

- is a helper application for web browsers to display structures in the MMDB format from the NCBI's structural database.
- It can be used on- or offline as a stand-alone program.
- It is able to render three-dimensional molecular models and display secondary structure cartoons.

The drawback is that it does not recognize the PDB format

Review Questions:

Short Answer Questions **(2 Marks)**

1. Define gene prediction.

2. Define ORF?
3. Define RBS?
4. Draw the structure of an amino acid?
5. What are the levels of structural organization of proteins?
6. What are the methods of protein tertiary structure prediction?
7. Define threading?
8. What are the tools used for visualizing a protein structure?
9. Define homology modeling?
10. Define propensity?
11. Name any four hydrophobic amino acids?
12. Name any four hydrophilic amino acids?
13. Name four amino acids that forms α-helix?
14. Name four amino acids that form β-sheets?
15. What are coils?
16. Define quarternary structure of a protein?
17. Write any two differences between prokaryotic and eukaryotic gene?
18. Write the three stop codons?
19. How can you identify a splice junction in a eukaryotic gene?
20. How is transcription terminated in a prokaryotic gene?

Essay Answer Questions: **(6 & 8 Marks)**

1. Describe a prokaryotic gene structure with diagram?
2. Describe a eukaryotic gene structure with diagram?
3. Write notes on secondary structure of a protein?
4. Describe protein secondary structure prediction methods.
5. How is tertiary structure of a protein predicted using homology based methods?
6. Describe threading method of tertiary structure prediction?

7. Describe some protein visualization tools

Further Readings:

Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press.

Applications of bioinformatics – en.wikipedia.org/wiki/**Bioinformatics** Attwood TK and Parry- Smith DJ (2006) Introduction to Bioinformatics. Pearson Education Ltd.

Unit V Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
Well-conserved regions in multiple sequence alignments	reflect areas of structural importance	reflect areas of functional importance.	reflect areas of both functional and structural importance	reflect areas of both functional and structural importance	
Which server is used to compare threedimensional protein structures?	DALI	FSSP	SCOP.	CATH.	DALI
Which one of the following tool is used to predict the threedimensional structure of a protein?	AutoDock	Gromacs	ChemSketch.	Modeller.	Modeller.
Which one of the following tool is not used to predict the threedimensional structure of a protein?	GLIDE	SwissPDB Viewer	JACKAL	Modeller.	GLIDE
Which one of the following tool uses comparative modeling method to predict the threedimensional structure of a protein	Rosetta.	Threader	CASP	Modeller	Modeller.
Homology modeling is also called as _____.	comparative modeling	abinitio prediction	threading	surface modeling.	comparative modeling
Which one of the following is a computational method to predict the threedimensional structure of the protein?	Xray crystallography	NMR	UV Spectroscopy	Threading.	Threading.
Which one of the following is an experimental method to determine the three-dimensional structure of the protein?	Threading	Xray crystallography	Homology modeling	Abinitio method.	Xray crystallography

Which method is used for predicting protein tertiary structure in the absence of homology to a known structure	Comparative modeling	Abinitio prediction	Threading	Surface modeling	Abinitio prediction
Who is the father of Genomics?	Altschul.	Gregor Mendel.	. Richard	Craig Venter	Craig Venter
Which of the following is the character based method?	UPGMA	Maximum Parsimony and Maximum Likelihood.	Maximum Likelihood and NeighborJoining	NeighborJoining	Maximum Parsimony and Maximum Likelihood.
A phylogenetic tree that explicitly represents number of character changes through its branch lengths is	dendogram	cladogram	phylogram	. chronogram	phylogram
Which one of the following is a command based offline tool for molecular structural visualization?	SwissPDB Viewer	. RasMol.	QMol.	PyMol.	. RasMol.
Molecular phylogeny can be performed with _____ sequences.	only DNA	only RNA	only Protein	DNA, RNA and Protein	DNA, RNA and Protein
Which one of the following is actually based on MolView?	Raswin.	. QMol.	RasMol	Moldraw.	Moldraw.
Which tool can be used for viewing molecular structures and animating molecular trajectories?	Chimera.	QMol.	Arguslab.	ChemSketch.	QMol.
Energy minimization of a modeled protein can be done using _____	ChemSketch.	Moldraw	RasMol.	SwissPDB Viewer	SwissPDB Viewer
Homology modeling can be done using _____.	SwissPDB Viewer	QMol	Raswin	Babel.	SwissPDB Viewer

Which one of the following tools can be used for both modeling the protein and structure visualization?	SwissPDB Viewer	QMol.	RasMol.	ChemSketch.	SwissPDB Viewer
Which one of the following tool can be used to generate neighbor joining trees with or without bootstrap values?	ClustalX	BLAST	SwissPDB viewer	ChemSketch	ChemSketch
Which one of the following is more weighted mutation?	Transitions	. Transversions	Transitions and transversion	Deletion	Transitions
Single substitution in the nucleotide sequence is called _____.	single substitution	simple substitution	single nucleotide polymorphism	simple nucleotide polymorphism	single nucleotide polymorphism
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.
The study of evolutionary relationships is _____.	Phylogenics	Molecular Evolution.	Cladogenesis.	Cladistics.	Phylogenics
PAUP stands for _____.	. Phylogenetic Analysis Using Parsimony	Phylogenetic Analysis Using Pairwise.	Proteomic Analysis Using Parsimony.	Phylogenetic Analysis Using Protein.	. Phylogenetic Analysis Using Parsimony
Which one of the following is not a characterbased method in tree construction?	Maximum parsimony.	Minimum likelihood	Minimum evolution method.	Neighbor joining.	Neighbor joining.
Which of these methods is a distancebased method in tree construction?	Unweighted pair group method with arithmetic mean	JukesCantor	Minimum evolution.	Maximum parsimony	Unweighted pair group method with arithmetic mean
Which is the only method that permits for the incorporation of alignable gaps as	Maximum parsimony.	Maximum likelihood	Neighbor Joining	Unweighted pair group method with arithmetic	Unweighted pair group method with arithmetic mean

characters?				mean	
Template based protein modeling techniques is called as _____	comparative modeling.	surface modeling	threading	abinitio prediction.	comparative modeling.
Which one of the following helps to calculate a structural similarity measure between pairs of structures of protein chains taken from the PDB?	CATH.	SCOP	FSSP	DALI.	DALI.
DDD stands for _____.	. Dali Domain Dictionary.	Distance Matrix Alignment Server.	Distance Matrix Domain Dictionary	Distance Domain Dictionary	Dali Domain Dictionary.
A _____ is defined in SCOP as a collection of superfamilies	primary structure of protein	secondary structure of protein.	protein fold	mutated protein sequences	protein fold
SCOP stands for _____.	Similar Classification of Proteins.	Structural Classification of Proteins	Similar Characterization of Proteins	Similar Classification of Proteins.	Structural Classification of Proteins
Which one of the following method predicts the protein structure based on fold recognition?	Comparative modeling.	Threading	Abinitio.	Homology modeling.	Threading
Which server is used to deposit the protein structures in PDB?	ClustalW.	. ClustalX.	ExPASy.	ADIT	ADIT
Which experimental structures cannot be deposited in PDB?	Xray crystallography.	NMR.	Mass spectrometry.	Comparative modeling	Comparative modeling
PDBID is a combination of _____ number of letters	1	2	3	4	4
PDBID is a _____representation	. SMILES.	ROSDAL.	WLN.	ALPHANUMERIC.	ALPHANUMERIC.
Which of the following is the distance based method?	PGMA	Maximum parsimony.	Maximum likelihood.	NeighborJoining	NeighborJoining
How many methods are there to predict 3dimensional	1	3	5	7	3

structure of a protein?					
Which is a repository for the 3-dimensional structure data for large biomolecules?	NCBI.	EMBL.	SwissProt.	PDB	PDB
Coordinates for known protein structures are housed in?	CATH	SCOP.	. PDBsum.	PDB	PDB
Hydropathy plots are usually used to predict _____.	beta secondary structure.	transmembrane domains.	alpha secondary structure.	tertiary structure	transmembrane domains.
A term used to classify protein domains according to their secondary structural content and organization is _____.	class	architecture	taxonomy.	homologs.	class
. In pairwise alignment result, sequences reported as similar due to chance represents _____ result	true positive.	. true negative	false positive.	false negative.	false positive.
In pairwise alignment result, sequences reported as related homologous represents _____ result.	true positive	true negative.	false negative.	false positive.	true positive
Which one of the PAM matrix represents amino acid substitutions that occur in distantly related proteins?	PAM1.	PAM250	PAM60	PAM45	PAM250
Which matrix is based upon the alignments of closely related protein sequences?	PAM 1	PAM250	PAM60	PAM 250	PAM 1
Which matrix uses the data on accepted mutations and the	PAM250	PAM 1	PAM 45	PAM 60	PAM 1

probabilities of occurrence of each amino acid to generate a mutation probability?					
Which method of multiple sequence alignment uses genetic recombination?	Progressive	Dynamic Programming.	Genetic Algorithm.	Hidden Markov Model.	Genetic Algorithm.
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
BLAST2 compares _____ number of sequences.	two	three	four	five	two
Single dot in the sequence alignment represents _____.	identity.	semiconserved substitutions.	conserved substitutions.	gaps.	semiconserved substitutions.
The paired dot in the sequence alignment represents _____	conserved substitutions.	semiconserved substitutions.	gaps	identity.	conserved substitutions.
Speciation event results in_	zoologs.	paralogs.	xenologs.	orthologs.	orthologs.
Gene duplication results in _____.	orthologs	paralogs.	xenologs.	zoologs.	paralogs.
PAM matrices are based on _____ of protein evolution.	Needleman and Wunsch	SmithWaterman.	Dayhoff model.	Markov model.	Dayhoff model.

What is the full form of BLOSUM?	Blocks of Amino Acid Substitution Mutation	Basic Amino Acid Substitution Mutation	Blocks of Amino Acid Substitution Matrix.	Basic Amino Acid Substitution Matrix.	Blocks of Amino Acid Substitution Matrix.
Define PAM?	Parallel Align Mutation.	Point Altered Mutation	Point Accepted Mutation.	Point Arranged Mutation	Point Accepted Mutation.
The PRINTS database consists of protein finger prints that define families in the _____ databases	SwissProt/TrEMBL	SwissProt/EMBL.	PIR/TrEMBL.	PIR/EMBL.	SwissProt/TrEMBL

Reg. No. : _____

[17MBU502B]

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Under Section 3 of UGC Act 1956)
COIMBATORE – 641 021
B.Sc. DEGREE EXAMINATION, July 2019
DEPARTMENT OF MICROBIOLOGY

I Internal Test – Fifth Semester

BIOINFORMATICS

Time: 2 hours **Maximum: 50 marks**
Date / Session: 18.12.2018/FN

Part A

Multiple Choice Questions: No 1 to 20 **20x1 = 20 marks**

1. FTP stands for _____.
   A. File transfer protocol B. File transfer process
   C. File transmit program D. File divert protocol
2. Relational databases that stores data in a structured format using _____.
   A. rows and columns B. segments
   C. lines and circles D. none of the above
3. Source and sink modes are the operations of _____.
   A. FTP B. SFTP
   C. SCP D. All the above
4. Cipher text is a _____.
   A. Encrypted data B. Unencrypted data
   C. Transferred data D. Plain data
5. Standard way of accessing data from a relational database is through _____.
   A. SQL B. SLQ C. SQS D. Both A and B
6. Amino acids are _____.
   A. building block of carbohydrates B. building block of vitamins
   C. building block of minerals D. building block of proteins
7. The computerized language used to store and organize data is called as_____.
   A. Database B. Databasic
   C. Data store D. Warehouse
8. PDB is a _____.
   A. structural database B. sequence database
   C. specialized database D. All the above
9. Which one of the following is the specialized database _____.
   A. REBASE B. EMBL C. DDBJ D. NCBI
10. Pick out the primary protein sequence database _____.
   A. PRINTS B. BLOCKS C. PIR D. PROSITE
11. GenBank division for bacterial sequences is_____
   A. BCT B. VRL C. ROD D. MAM
12. The unique identifier for a sequence record is _____
   A. Accession No. B. Keywords C. Reference D. Title
13. dbEST is a _____.
   A. metabolic pathway database B. gene expression database
   C. nucleotide database D. None of the above
14. Microarray technology works for _____.
   A. Gene expression profiling B. Genotyping

C. Mapping of gene D. All the above

15. SMD stands for _____.

A. Stanford Microarray database B. Stanford Molecular database

C. Single Microarray database D. None of the above

16. KEGG is the database for _____.

A. nucleotides B. protein C. metabolic pathways D. gene expression

17. Class Architecture Topology Homology is _____.

A. SCOPE B. CATH C. STRING D. PIR

18. The database GenBank is the collection of _____.

A. Nucleic acid sequences B. Protein sequences

C. protein structures D. All the above

19. EMBL was established during _____.

A. 1978 B. 1974 C. 1981 D. 1982

20. DDBJ belongs to _____.

A. USA B. GERMANY C. INDIA D. JAPAN

Part B

Answer all the questions **3x2 = 6 marks**

21. Define Relational Database.
22. Write about SCP – data transfer protocol.
23. Expand NCBI, PDB, EMBL and ESTs.

Part C

Answer all the questions **3x8 = 24 marks**

24. A. Give a detailed account on Relational Database Management System (RDBMS).

Or

B. Write a detailed note on FTP with their data representation modes.

25. A. Write the advantages of encrypted data transfer.

Or

B. Write in detail about biological databases and its classification.

26. A. Write short notes on primary and secondary protein sequence databases. Give one example of SWISS PROT data structure.

Or

B. Write a detailed note on Gene expression and metabolic pathways databases.

****