GUDLAVALLERU ENINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

Seshadri Rao Knowledge Village, Gudlavalleru – 521 356.

Department of Computer Science and Engineering

HANDOUT

on

BIO-INFORMATICS

Vision

To be a Centre of Excellence in computer science and engineering education and training to meet the challenging needs of the industry and society.

Mission

- To impart quality education through well-designed curriculum in tune with the growing software needs of the industry.
- To serve our students by inculcating in them problem solving,leadership,teamwork skills and the value of commitment to quality,ethical behavior & respect for others.
- To foster industry-academia relationship for mutual benefit and growth.

Program Educational Objectives

- Identify, analyze, formulate and solve Computer Science and Engineering problems both independently and in a team environment by using the appropriate modern tools.
- Manage software projects with significant technical, legal, ethical, social, environmental and economic considerations
- Demonstrate commitment and progress in lifelong learning, professional development, leadership and Communicate effectively with professional clients and the public.

HANDOUT ON BIO-INFORMATICS

Class& Sem. : IV B.Tech – I Semester Year :2017-18

Branch : CSE Credits : 3

==========

1. Brief History and Scope of the Subject

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data.

2. Pre-Requisites

- Familiar with the fundamental concepts of biological science and computers.

[

3. Course Objectives:

- To know the importance of Bioinformatics for computational learning.
- To understand basic biological databases, algorithms for proteomics and genomics analysis
- To learn the bioinformatics packages to solve the biological problems

4. Course Outcomes:

CO1: the differences between genomics and proteomics
CO2 to solve the biological problems using computational approach
CO3: to perform data sequence search

5. Program Outcomes:

Graduates of the Computer Science and Engineering Program will have

a. Apply knowledge of computing, mathematics, science and engineering fundamentals to solve complex engineering problems.

b. Formulate and analyze a problem, and define the computing requirements appropriate to its solution using basic principles of mathematics, science and computer engineering.
c. Design, implement, and evaluate a computer based system, process, component, or software to meet the desired needs.
d. Design and conduct experiments, perform analysis and interpretation of data and provide valid conclusions.
e. Use current techniques, skills, and tools necessary for computing practice.
f. Understand legal, health, security and social issues in Professional Engineering practice.
g. Understand the impact of professional engineering solutions on environmental context and the need for sustainable development.
h. Understand the professional and ethical responsibilities of an engineer.
i. Function effectively as an individual, and as a team member / leader in accomplishing a common goal.
j. Communicate effectively, make effective presentations and write and comprehend technical reports and publications.
k. Learn and adopt new technologies, and use them effectively towards continued professional development throughout the life.
l. Understand engineering and management principles and their application to manage projects in the software industry.

6. Mapping of Course Outcomes with Program Outcomes:

	a	b	c	d	e	f	g	h	i	j	k	l
CO1							M					
CO2	H											
CO3				H								

7. Prescribed Text Books

a. S.P.T.K Attwood & D J Parry-Smith,Introduction to Bioinformatics, Pearson Education Publications.
b. M.L.R Dane Krane, Wright State University, Fundamental concepts of Bioinformatics

8. Reference Text Books

a. C.N Jean-Michel Claveriw,Bioinformatics-A Beginners guide,WILEY Dream Tech-2003
b. S.M.D. Leon,Sequence Analysis in a Nutshell,1st ed,O'REILLY-2003

9. URLs and Other E-Learning Resources

a. https://www.atdbio.com/content/14/Transcription-Translation-and-Replication

10. Digital Learning Materials:

- https://onlinecourses.nptel.ac.in

11. Lecture Schedule / Lesson Plan

Topic	No. of Periods	
	Theory	Tutorial
UNIT –1: Introduction and DNA Sequence analysis		
Introduction to Bioinformatics-history of bioinformatics	1	1
Role of bioinformatics in biological sciences	1	
Scope of bioinformatics	1	
The Central dogma	2	1
DNA and Protein	2	
	7	**2**
UNIT – 2: Applications		
Genetic code	2	1
Sequencing, biological sequence/structure	2	
Genome projects	2	1
Pattern recognition an prediction	2	
Folding problem	1	
Sequence analysis,homology and anology	1	

	10	**2**
UNIT – 3: Databases in Bioinformatics		
Protein information resources- biological databases	1	1
Primary sequence databases	1	
Protein sequence databases	1	
Secondary databases	1	1
Protein pattern databases	2	
structural classification databases		1
	6	**3**
UNIT – 4: genome information resources		
DNA sequence databases	3	1
Specialized genome resources	3	1
	6	**2**
UNIT – 5: Alignment Techniques		
Pair-wise alignment techniques- database searching, alphabets and complexity	2	1
Alogirthms and programs,comparing two sequences, sub-sequences	3	
Identity and similarity, the Dotpot,Local and global similarity	2	
Different alignment techniques,dynamic programming	2	1
Pair-wise database searching	2	
	11	**2**
UNIT – 6: Database Searching and Analysis Packages		
Secondary database searching-Importance and need of secondary database searches	2	1
Secondary database structure and building a sequence search protocol	2	
Analysis packages- analysis package structure	3	1
Commercial databases	2	
Commercial software	1	
	9	**2**
Total No.of Periods:	**49**	**13**

12. Seminar Topics

- Measures of similarity in amino acids sequences
- Cancer informatics ecosystem
- Proteomics and bioinformatics
- Bioinformatics tools for heath care
- Parallel algorithms for bioinformatics applications
- Computational intelligence in bioinformatics

UNIT –I

Objective:

- To know the importance of Bioinformatics for computational learning.

Syllabus:

Introduction and DNA Sequence analysis

Introduction to Bioinformatics-history of bioinformatics, Role of bioinformatics in biological sciences, Scope of bioinformatics, The Central dogma, DNA and Protein

Learning Outcomes:

The student will be able to

- describe how to experimentally obtain and evaluate DNA sequence information

Learning Material

- Bioinformatics is a highly interdisciplinary field of biology, relying on basic principles from computer science, biology, physics, chemistry and mathematics.
- It is the subject that involves the use of techniques from all these subjects to deal with problems of biology understand biological processes, and find methods and solutions with information technology to solve biological problem.

What is Bioinformatics (Broader or Loose definition)

- From Introduction to Bioinformatics by Attwood and Parry Smith – "The term Bioinformatics is used to encompass all computer applications in biological sciences".
- as per Cynthia Gibas, Bioinformatics is the intersection of information technology and biology

Bioinformatics definition: (Narrow or Tight definition)

- Fredij Tekaia at the Institute Pasteur offers the definition of bioinformatics as " The mathematical statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information"

- Bio informatics is the application of information technology to the study of living things, usually at the molecular level
- Bioinformatics involves the use of computers to collect organize and use Biological information to answer questions in field like evolutionary biology

Introduction to Bioinformatics

- One of the important properties of living things is their ability to reproduce
- Living things are able to make copies of themselves, not perfect copies, but close enough
- The process of copying anything always involve a transfer of information
- Transformation of information from one form to another always has some error associated with it, just as the transformation of energy from one form to another will always have some waste
- Energy and information are connected through these errors and the concept of entropy in Thermo dynamics ties the together.
- Life is dependent on processes that convert information into chemical energy and eventually into the chemistry of macro molecules. This was unknown until the discovery of deoxyribonucleic acid (DNA)

Bioinformatics – When and Why

- It was in the 17th century that biologist started dealing with problems of information management
- By the middle of the 17th century, John ray introduced the concept of distinct species of animals and plants developed guidelines based on anatomical features for distinguishing conclusively between species.
- In 1730, Carolus Linnaeus established the basis for the modern taxonomic naming system of kingdoms, classes, geneera, and species.
- Taxonomy was the first informatics problem in biology.
- The university of Arizona's tree of Life project and NCIB'S taxonomy database are two example of quline taxonomy projects

- Bioinformatics has, fact been in existence for more than 30 years and is now middle – aged.
- Collecting and cataloguing information about individual genes in human DNA determining the sequence of three billion chemical bases that made up the human DNA became the second informatics in biology,
- In 1980, human genome project was initiated as a prominent bioinformatics solution to the problem and this labeled the 21st century as the era of genomes.

Role of Bioinformatics in biological sciences

With adequate data and right tools, it is possible to explore a number of new areas in biology.

- Sequence Analysis: For sequence analysis, there are many powerful tools and computers which perform the duty of analyzing the genome of various organisms. These computers and tools also see the DNA mutations in an organism and also detect and identify those sequences which are related.

- Prediction of Protein Structure: It is easy to determine the primary structure of proteins in the form of amino acids which are present on the DNA molecule but it is difficult to determine the secondary, tertiary or quaternary structures of proteins. For this purpose either the method of crystallography is used or tools of bioinformatics can also be used to determine the complex protein structures.

- Comparative Genomics: Comparative genomics is the branch of bioinformatics which determines the genomic structure and function relation between different biological species.For this purpose, intergenomic maps are constructed which enable the scientists to trace the processes of evolution that occur in genomes of different species. These maps contain the information about the point mutations as well as the information about the duplication of large chromosomal segments.

- Health and Drug discovery: The tools of bioinformatics are also helpful in drug discovery, diagnosis and disease management.

Complete sequencing of human genes has enabled the scientists to make medicines and drugs which can target more than 500 genes. Different computational tools and drug targets has made the drug delivery easy and specific because now only those cells can be targeted which are diseased or mutated. It is also easy to know the molecular basis of a disease.

- DNA forensics: DNA profile of an individual, called DNA fingerprints can help in identifying criminals, establishing family relationships, protecting rare wildlife species, and matching organ donors

Scope of Bioinformatics

- Current biological and medical labs use methods that produce extremely large date sets, which cannot be analyzed by hand – for instance sequencing human genomes
- Modern Biological and medical research and development cannot be done without bio informatics
- Future applications in biology, chemistry, pharmaceuticals, medicine and agriculture
- In addition, bioinformatics plays an important role in bio medical research
- Research work in the area of genetic diseases and medical genomics is rapidly increasing and the future of personalized medicine depends on Bioinformatics approaches

The central dogma

- DNA (deoxyribonucleic acid) is the genetic material
- it represents the answer to questions that have pondered by philosophers and scientists for thousand of year
  1. What is the basis of inheritance?
  2. What allows living things to be different from non living things?

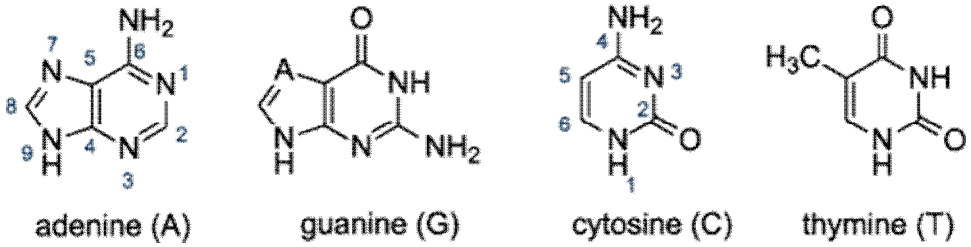
- It is the information stored in DNA that allows the organization of inanimate molecules into functioning, living cells and organisms that are able to regulate their internal chemical components growth and reproduction.

Nucleotides

- Genes themselves contain their information as a specific sequence of nucleotides that are found in DNA molecules.
- Only four different bases are used in DNA
  (1) Guanine (2) Adenine (3) Thymine (4) Cytosine

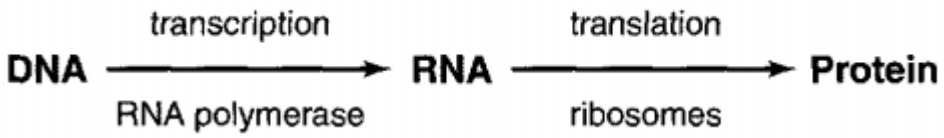
Two purines called **adenine** (**A**) and **guanine** (**G**)

two **pyrimidines**, called **thymine** (**T**) and **cytosine** (**C**)



- Each base in attached to phosphate group and a deoxyribose sugar to form a nucleotide
- The only thing that makes one nucleotide different from another is which nitrogeneous base it contains

Central Dogma



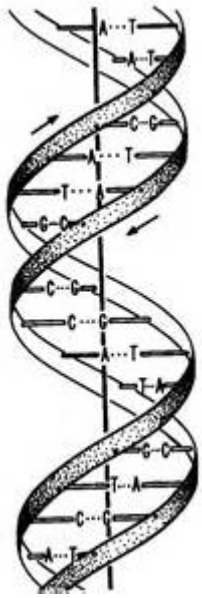
- The "central dogma "is the process by which the instructions in DNA are converted into a functional product. It was first proposed in 1958 by Francis crick, discover of the structure of DNA
- The central dogma suggests that DNA contains the information needed to make all of use proteins, and that RNA is a messenger that carries this information to the ribosome's
- The ribosome's serve as factories in the cell when the information is translated from a code into a functional product
- The process by which the DNA instructions are converted into the functional product is called gene expression
- Gene expression has two key stages
  (1) Transcription
  (2) Translation
- In transcription, the information in the DNA of every cell is converted into small, portable RNA messages

- During translation, these messages travel from where the DNA is in the cell nucleus to the ribosome's when they are read to make specific proteins.
- The central dogma states that the pattern of information that occurs most frequently in one cells is
  (1) From existing DNA to make new DNA (DNA replication)
  (2) From DNA to make new RNA (transcription)
  (3) From RNA to make new proteins (translation)

DNA

- DNA is a single, large molecule, sometimes up to 2- meters long
- It consists of two long stings of smaller molecules called nucleotide, wound up against each other in a structure, now famous called the double helix.
- The two stings of the double helix are connected to each other, periodically, by a set of four molecules or bases Adenine, Guanine, Cytosine and Thymine
  - o two **purines**, called **adenine** (**A**) and **guanine** (**G**)
  - o two **pyrimidines**, called **thymine** (**T**) and **cytosine** (**C**)

- The structure of DNA is like a ladder, twisted around to form a helix.



- The steps of this ladder are formed of pairs of molecules Adenine with Thymine, or Guanine with Cytosine. AT or GC
- The DNA molecules exist inside every cell nucleus in a living Organism
- The sequence of base pairs in DNA are the " raw materials" as well as "instructions" for building and maintaining the Organism
- Sequences of base pairs after millions of alphabets in numbers are grouped together into genes.
- When a gene inside a DNA molecule needs to be activated, The DNA molecule in a cell nucleus uncoils and unfurls to just the right intent to expose that gene and at this time, RNA molecule is formed.
- Some organisms, for example retroviruses, use ribonucleic acid (RNA) instead of DNA as their store of genetic information.

- Reverse transcription is the transfer of information from RNA to make new DNA, this occurs in the case of retroviruses, such as HIV. It is the process by which the genetic information from RNA is assembled into new DNA

DNA and protein

- DNA or otherwise called deoxyribonucleic acid is the building block of the life
- It contains the information the cell requires to synthesize protein and to replicate itself, to be short it is the storage repository for the information that is required for any cell to function
- Watson – crick has discovered the current structure in 1953
- The famous double – helix structure of DNA has its own significance
- The DNA sequence looks like this "ATTGCTGAAGGTGCGA"
- DNA Is measured according to the number of base pairs it consists of usually in kbp or mbp
- Each base has its complementary
  A has T as complementary.
  G has C as complementary
- The DNA base of the human genome were typed as A, C, T and G , the 3 billion letters would fill 4000 books of 500 pages
- The DNA is broken down into bits that wound into coils which are called chromosomes
- Human beings have 23 pairs of chromosomes

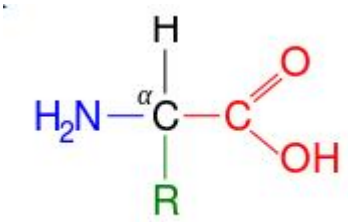
- Each chromosome is further broken into pieces called genes.
- The 23 pairs of chromosomes consists of about 70,000 genes and every gene has its own function
- Determining the gene's functionality and position of the gene in the chromosome is called gene mapping
- An intermediate language, encoded in the sequence of Ribonucleic Acid (RNA) translates a gene's message into a protein's amino acid sequence
- RNA is somewhat similar to DNA, they both are which acids of nitrogen – containing bases joined by sugar – phosphate back bone

RNA	DNA
(1)Single stranded	(1)Double stranded
(2) Has uracil as base	(2) Has thymine as base
(3) Ribose as the sugar	(3) deoxyribose as sugar
(4) Use protein encoding information	(4) Maintains protein encoding information

- In the synthesis of protein these are three types of RNA
  (1) Messenger RNA (mRNA) --- carries the genetic information from DNA and is used for a template for protein synthesis
  (2) Ribosomal RNA (rRNA) ---- major consistent of the cellular particles called ribosome's on which protein synthesis actually takes place.
  (3) A set of transfer RNA (tRNA), each of which incorporates a particular amino acid submit into the growing protein when it recognizes a specific group of three adjacent bases in the m RNA .
- DNA maintains genetic information in the nucleus
- RNA takes that information into the cytoplasm
- Bacteria have at least three distinct DNA polymerases: Pol I, Pol II and Pol III; it is Pol III that is largely involved in chain elongation.

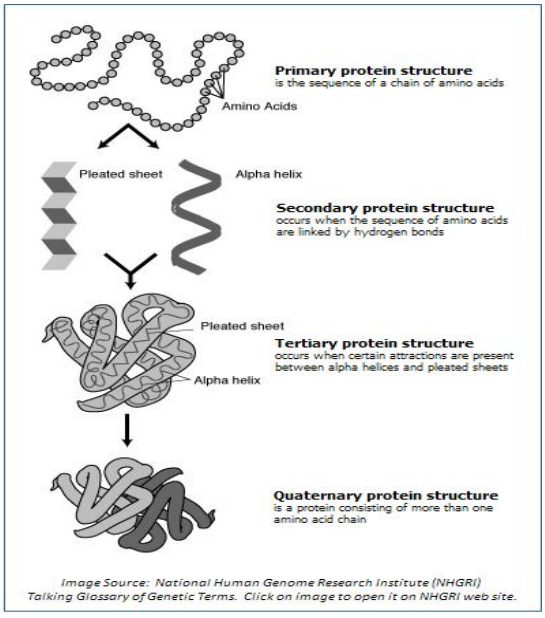
Protein—

- Proteins are molecules that are called polymers that are built using independent blocks linked together in repeating units
- They are building blocks of proteins
- For every amino acid there are amino group, □- carbon, cooboxyl group and variable group
- The only difference is variable group that varies between any amino acid and it is the variable group that shows the function of that amino acid.



- Multiple amino acids are linked together to form polypeptide or protein. They can be linked together through a reaction called condensation reaction. This reaction moves a molecule of water to create a bond
- It is the bond that links together two amino acids
- When you have more than 3 amino acids they we consider it is polypeptide through peptide bond.

Protein structure---



(1)Primary structure ---

It is simply the order of amino acids that make up the polypeptide chain. It is the sequence that they are linked together.

(2)Secondary structure ---

It is particular shape that polypeptide takes place. Then structures formed by regular intermolecular hydrogen bonding

These are two types

(1) α- helix

(2) β pleated sheet _____

- Often the secondry structures are the first portions of the portion to fold after translation
- α - helix an characterized by phi and pri angles of roughly – 60 degrees and exhibit a spring like helical shape with 3.6 amino acids per complete 360 degrees turn
- B sheet an characterized by regions by extended back bone confirmation with

(3)Tertiary structure ---

The regions of secondary structure in a protein pack together and combine with other less structured regions of the protein backbone to farm an overall three dimensional shape which is called tertiary structure.

(4) Quaternary structure ---

An action enzyme is compound of two or more protein chains that come together into single large complex. When this occurs, the overall structure farmed by the interacting proteins is commonly referred as quaternary structure

UNIT-I

Assignment-Cum-Tutorial Questions

SECTION-A

I. Objective Questions

1. For his breakthrough in rapid sequencing techniques, ------- earned a second Nobel Prize for Chemistry in 1980. [ ]

A) Sanger B) Walter Gilbert C) Paul Berg D) Har Gobind Khorana

2. _____discovered the basic rules of heredity of garden pea that an individual organism has two alternative heredity units for a given trait (dominant trait Vs. recessive trait). [ ]

A) Gregor Mendel B) Walter Gilbert

C) Sanger D) Har Gobind Khorana

3. Using the data from Franklin and Wilkins, _____were able to determine the double-stranded structure of DNA. [ ]

A) Watson and Crick B) Walter Gilbert

C) Sanger D) Har Gobind Khorana

4. _____is an enzyme that is responsible for copying a DNA sequence into an RNA sequence, during the process of transcription. [ ]

A. RNA polymerase B. Purine C. Cytosine D. ALL

5. A macromolecule, usually a protein, that catalyzes biochemical reactions, lowering the activation energy and increasing the rate of reaction is ___ . [ ]

A. Enzyme B. Thymine C. Guanine D. None of the above

6. The study of protein structure, function, and interactions produced by a particular cell, tissue, or organism is called as _____. [ ]

A) Recombinant s B)Proteomics C) Genomics D) prokaryotics

7. The combination of a nucleobase and a pentose is called a _____[ ]

A) nucleoside B) prokaryotic C) Recombinant D) sugar

SECTION-B

II. Descriptive Questions

1. Explain why the nucleic acids have two distinctive ends: the 5′ (5-prime) and 3′ (3-prime) ends?
2. Differentiate between purines and pyrimidines
3. Differentiate between prokaryotic and eukaryotic cells.
4. The sequence of bases of one strand of DNA is given as CGACCCCAG. Give the base sequence that will be produced as a result of transcription.

UNIT –II

Objective:

- To know the importance of Bioinformatics for computational learning.

Syllabus:

Applications

Genetic code, Sequencing, biological sequence/structure, Genome projects, Pattern recognition an prediction, Folding problem, Sequence analysis, homology and anology

Learning Outcomes:

The student will be able to

- describe the contents and properties of the most important bioinformatics databases
- understand the basics of Bioinformatics resources.

http://fig.cox.miami.edu/~cmallery/150/gene/genome.size.htm

Learning Material

Genetic code

- The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA) is translated into proteins by living cells.
- Specifically, the code defines a mapping between tri – nucleotide sequences called codons and amino acids, every triplet of nucleotides in a nucleic acid sequence specifies a single amino acid.
- Because the vast majority of genes encoded with exactly the same code, this particular code is often referred to as the canonical are standard genetic code, are simply the genetic code, though in fact there are many variant codes thus the canonical genetic code is not universal.
- The genome of an organism is inscribed in DNA, or in some viruses in RNA
- The portion of the genome that codes for a protein or an RNA is referred to as a gene
- Those genes that code for proteins or composed of tri – nucleotide units called codons, each coding for a single amino acid.

- Each nucleotide sub unit consists of a phosphate, deoxyribose sugar and one of the 4 nitrogenous nucleotide bases.
- The purine bases adenine (A) and guanine (G) are larger and consist of two aromatic rings.
- The pyrimidine bases cytosine (C) and thymine (T) are smaller and consists of only one aromatic ring.
- In the double – helix configuration, two strands of DNA are joined to each other by hydrogen bonds in an arrangement known as base pairing
- These bonds almost always form between an adenine base on one strand and a thymine on the other strand and between a cutosine base on one strand and a guanine base on the other
- This mean that the number of A and T residues will be the same in a given double helix as will the number of G and C residues
- In RNA, thymine (T) is replaced by uracil (U) and the deoxyribore is substituted by ribose.

DNA sequencing

- DNA sequencing is the process of determining the sequence of nucleotides (As, Ts, Cs, and Gs) in a piece of DNA
- Sequencing an entire genome remains a complex task
- It requires breaking the DNA of the genome into many smaller pieces, sequencing the pieces, and assembling the sequences into a single long "Consensus"
- With latest methods, genome sequencing is now much faster and less expensive than it was during the Human genome project

Methods of sequencing

(1) Sanger di-deoxyl (primer extension / chain – termination) method

2.) Maxam – Gilbert chemical cleavage method:-

Sanger sequencing

(1) Sanger sequencing developed by Fred Sanger et al in the mid 1970's

(2) Uses dideoxynucleotides for "chain termination", generating fragments of different lengths ending in ddATP, ddGTP, ddCTP, ddTTP

Requirements of Sanger method

- DNA to be sequenced must be in single strand form
- Primer
- DNA polymerase
- di deoxy nucleotide
- 4 dideoxynucleotidephosphates (dd NTPS)

Sanger technique principle

- The sanger technique uses dideoxynucleotides (di deoxyadenine, etc) these are molecules that resemble normal nucleotides but lack the normal –OH group
- Because they lack the –OH (which allows nucleotides to join a grouping DNA standard), replication stops

Sanger Method:

Procedure:-

1. Denaturation

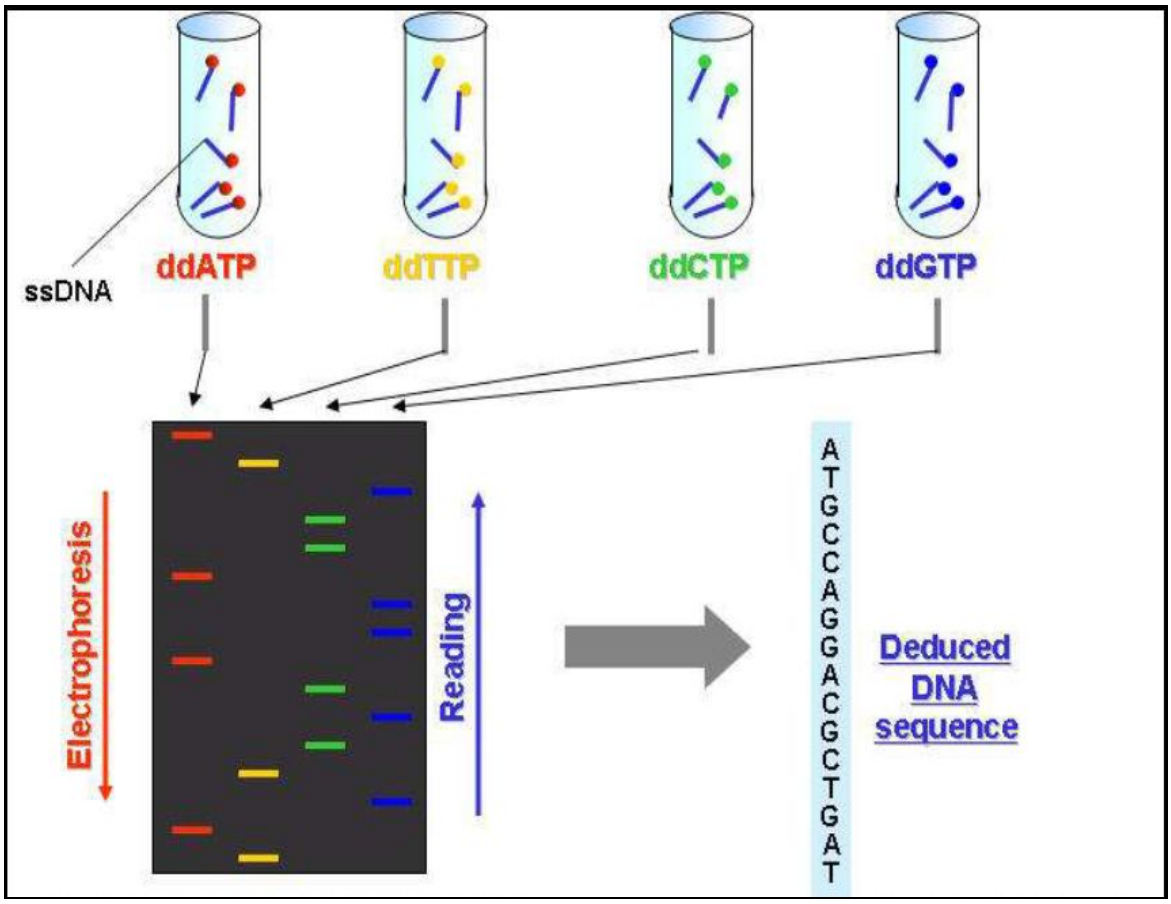
2. Primer attachment and extension of bases

3. Termination

4. Gel electrophoresis

- The DNA template is treated with heat so that it becomes single stranded
- A short, single-stranded primer which is radioactively labelled is added to the end of the DNA template
- Add template DNA and primer in 4 Tubes.
- Now add ddNTPs In tubes in the way that single tube contain one type of ddNTP.
- Extension is start and band formed of various sizes.

- The fragments of DNA are separated by electrophoresis.
- Overlap these sequences to find out sequence of Target DNA.



Sequencing gel (Slab or capillary gel)

Maxam – Gibert sequencing:-

Principle :-

- purification of the DNA fragment that to be sequenced and labeled with radioactive material.
- Chemical treatment generates breaks at a specific nitrogenous bases and thus a series of labelled fragments is generated.
- The fragments in the four reactions are arranged side by side in gel electrophoresis for size separation.

- The fragments visualize in X-ray for autoradiography.
- To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radio labelled DNA fragment, from which the sequence may be inferred.

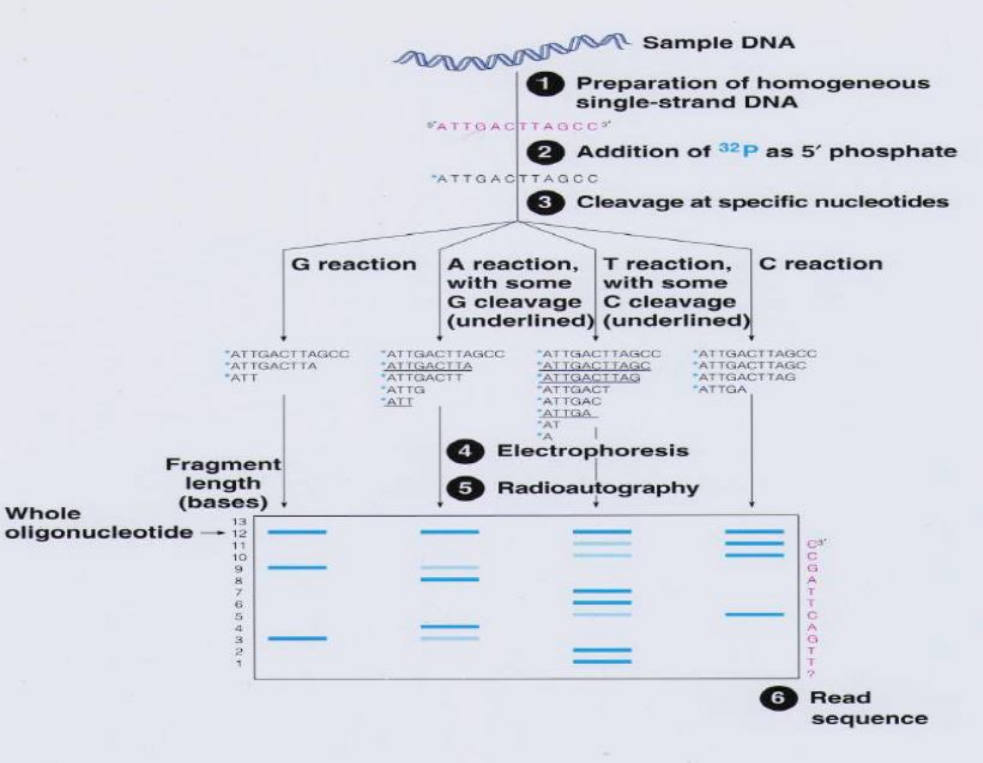


Fig:-Maxam – Gibert sequencing:-

Procedure:-

Maxam–Gilbert sequencing requires radioactive labeling at one 5′ end of the DNA fragment to be sequenced (gamma-32P).

Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T).

1. the purines (A+G) by using formic acid,
2. the guanines (and to some extent the adenines) by dimethyl sulfate,
3. the pyrimidines (C+T) by using hydrazine.
4. NaCl add to hydrazine for Cytosine.

Add each chemical in separate tube.

Thus a series of labeled fragments is generated.

The fragments in the four reactions are electrophoresed side by side for size separation.

To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each showing the location of identical radio labeled DNA molecules.

Advantages:-

- **Forensics**:-

  DNA sequencing has been applied in forensics science to identify particular individual because every individual has unique sequenceof his/her DNA. It is particularly ed to identify the criminals byfinding some proof from the crime scene in the form of hair, nail, skinor blood samples.

- **Agriculture**:-

  DNA sequencing has played vital role in the field of agriculture.The mapping and sequencing of the whole genome ofmicroorganisms has allowed the agriculturists to make them useful for the crops and food plants.

- **Medicine**:-

  In medical research, DNA sequencing can be used to detect thegenes which are associated with some heredity or acquireddiseases. Scientists use different techniques of genetic engineering like gene therapy to identify the defected genes andreplace them with the healthy ones.

- More and more old crimes are beingsolved by resubmitting evidence for enhanced DNA testing.
- Another major advantage of DNAanalysis is the ability to screen for certaingenetic diseases or risk factors.
- Women involved in certain fertilitytreatments can also get information aboutan embryo before it is implanted.
- The chance of a DNA match between two persons who aren't twins is from 1/7000 to
- 1/1,000,000,000, depending on the frequency of the patterns being compared.
- This is a much more specific test than other methods such as blood type, and DNA is

- present in any of kind of body tissue, so it is more likely to be found at a crime scene than blood.
- DNA testing is also more reliable than eyewitness testimony

Dis advantages:-

- ✓ One key disadvantage of DNA analysis is the potential for invasion of individual privacy;

- ✓ Because a person's DNA reveals so much information about their physical state, it is sensitive information that must be carefully guarded;

- ✓ Information about an individual's ethnic background and parentage could become cause for discrimination;

- ✓ Disadvantages include incomplete coverage, which can lead to false normal results, and the ability to test only for unbalanced rearrangements (duplications and deletions), and not balanced translocations or inversions

Genome projects

- ➢ In the mid – 1980's the United States Department of Energy (DOE) initiated number of projects to construct detailed genetic and physical maps of the human genome to determine its complete nucleotide sequence
- ➢ The human genome project (HGP) was the international collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings. All our genes together are known as our "genome"
- ➢ The hereditary material of all multi – cellular organisms is the famous double helix of deoxybonuclic acid (DNA) which contains all of our genes
- ➢ HGP researches have deciphered the human genome in three major ways
  (1) Determining the order of sequence of all the bases in our genome's DNA
  (2) Making maps that show the location of genes for major selections of all of chromosomes

(3) Producing what are called linkage maps, complex versions of the type originated in early drosophila research, through which inherited traits can be tracked over generations

- Similar research efforts were also launched to map and sequence the genomes of variety of organisms used extensively in research laboratories as mode systems
- Although the sequencing projects of only a small number of relatively small genomes had been completed including the human genome is the results of such projects

Pattern recognition and prediction

- Pattern recognition methods as the name suggests are built on the assumption that some underlying characteristic of a protein sequence, or of a protein structure, can be used to identify similar traits in related proteins
- If part of a sequence or structure is preserved or conserved this characteristic may be used to diagnose new family member
- Searches of sequence pattern databases, and of fold template databases, are now routinely used to diagnose family relationship
- By definition, both sequence and structure based pattern recognition methods demand that a particular sequenced or structure has been ‘ seen ‘ before, and that some characteristics of it can be launched in a reference database
- Prediction, the holy grail of bioinformatics, is still not possible and is unlikely to be so for decodes to come
- Prediction stems from the idea that a functional site, or indeed a complete structure, need not have been ‘seen’ before , but can be deducted directly from the amino acid sequence
- This approach the need to create reference database of functional site or structural templates, but requires inserted the design of sophisticated

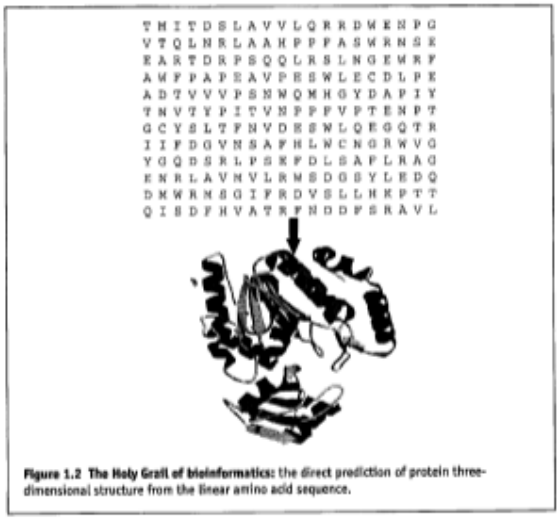


Figure 1.2 The Holy Grail of bioinformatics: the direct prediction of protein three-dimensional structure from the linear amino acid sequence.

The folding problem

- The following problem is a central theme of molecular biology
- In 1961, Anfineen showed that ribonucleare could be denatured and refolded without loss of enzymatic activity
- This experiment suggested that all the information for a protein to adopt its native conformation is encoded in its primary structure.
- In 1998, methods of secondary structure prediction little more than 50-60% reliable
- There are three main approaches to secondary structure prediction

  (1) Empirical statistical methods use parameters derived from known 3d structures

  (2) Methods based on physicochemical criteria

  (3) Prediction algorithms that are known structures to assign secondary structure
- Tertiary structure prediction is still further beyond reach
- In 1998, it is clear that direct prediction of structure sequence remains decoder away

Sequence analysis

- The exact nature of the information encoded in the primary structure unclear, and we still cannot read the language used to describe the final 3d fold of a biologically macro molecule.
- Using sequence analysis techniques, we can attempt to identify similarities between novel query sequences and databases sequences whose structures and functions have been elucidated

The twilight zone

- The twilight zone is a zone of sequence similarity in which alignments may appear plausible to the eye, but are no longer statistically significant
- To penetrate deeper into the twilight zone is the goal of most analytical methods
- Many different approaches have been devised, source of
- These involve database searches with single sequences
- Use characteristic chunks of sequence alignments
- Some weight database searches
- Others are only observed amino acid sequence data
- Each method offers a different perspective, depending on the type of information used in the search

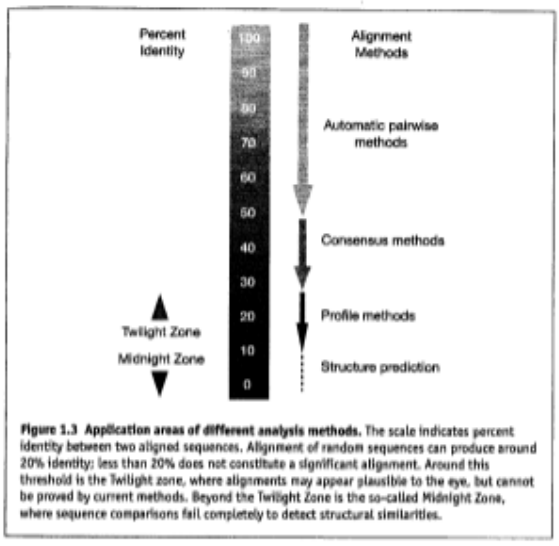


Figure 1.3 Application areas of different analysis methods. The scale indicates percent identity between two aligned sequences. Alignment of random sequences can produce around 20% identity; less than 20% does not constitute a significant alignment. Around this threshold is the Twilight zone, where alignments may appear plausible to the eye, but cannot be proved by current methods. Beyond the Twilight Zone is the so-called Midnight Zone, where sequence comparisons fail completely to detect structural similarities.

- While the algorithms that recognize folds reliably, or that can predict structures, it is important to use the sequence analysis tools we have at our disposal, but in an intelligent way

Homology and Analogy

- The term homology, although easy to understand, is confounded and abused in the literature
- Sequences are said to be homologues if they are related by divergence from a common ancestor.
- For example, in everyday life, people look like one another for different reasons
- Two sisters, for example might look alike because they both inherited brown eyes and black hair from their father
- It works the same way in biology
- Some traits shared by two living things were inherit from their ancestor and some similarities evolved in other ways, these are called homologies and analogies
- Homology among proteins or DNA is typically inferred from their sequence similarity
- Significant similarity is strong evidence that two sequences are related by divergent evolution of a common ancestor
- Alignments of multiple sequences are used to indicate which regions of each sequence are homologous
- In some cases, when sequence and structure are different, we can infer with a degree of confidence that the triads result from convergent evolution
- In the former case, however, where folds are similar but the sequences differ, which such folds are usually considered to be analogues
- It is sometimes difficult to rule out the existence of a common ancestor because structures are more highly conserved than the sequence
- The errence of sequence analysis is the detection of homologues

  Sequence by means of routines database searches, usually with unknown or uncharacterised query searches.

Orthology and paralogy

- Among homologous sequences, it is useful to distinguish between proteins that perform the same function in different species refund to as ORTHOLOGUES. And those that program different but related functions within one oraginsm called PARALOGUES

Genome projects

- In the mid – 1980's the United States Department of Energy (DOE) initiated number of projects to construct detailed genetic and physical maps of the human genome to determine its complete nucleotide sequence
- The human genome project (HGP) was the international collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings. All our genes together are known as our "genome"
- The hereditary material of all multi – cellular organisms is the famous double helix of deoxybonuclic acid (DNA) which contains all of our genes
- HGP researches have deciphered the human genome in three major ways
  (4) Determining the order of sequence of all the bases in our genome's DNA
  (5) Making maps that show the location of genes for major selections of all of chromosomes
  (6) Producing what are called linkage maps, complex versions of the type originated in early drosophila research, through which inherited traits can be tracked over generations
- Similar research efforts were also launched to map and sequence the genomes of variety of organisms used extensively in research laboratories as mode systems
- Although the sequencing projects of only a small number of relatively small genomes had been completed including the human genome is the results of such projects

Pattern recognition and prediction

- Pattern recognition methods as the name suggests are built on the assumption that some underlying characteristic of a protein sequence, or of a protein structure, can be used to identify similar traits in related proteins
- If part of a sequence or structure is preserved or conserved this characteristic may be used to diagnose new family member
- Searches of sequence pattern databases, and of fold template databases, are now routinely used to diagnose family relationship
- By definition, both sequence and structure based pattern recognition methods demand that a particular sequenced or structure has been ' seen ' before, and that some characteristics of it can be launched in a reference database

- Prediction, the holy grail of bioinformatics, is still not possible and is unlikely to be so for decodes to come
- Prediction stems from the idea that a functional site, or indeed a complete structure, need not have been 'seen' before , but can be deducted directly from the amino acid sequence
- This approach the need to create reference database of functional site or structural templates, but requires inserted the design of sophisticated

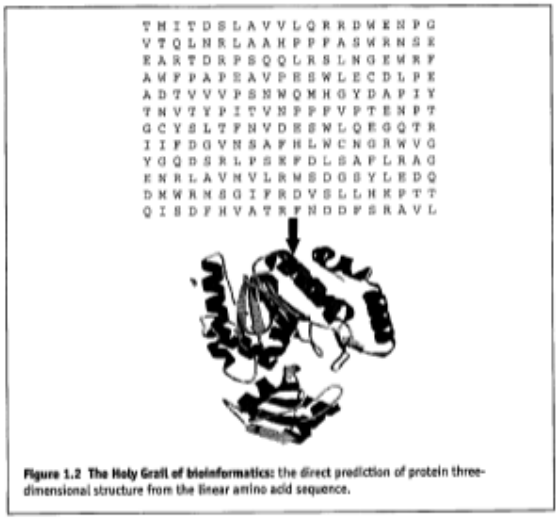


Figure 1.2 The Holy Grail of bioinformatics: the direct prediction of protein three-dimensional structure from the linear amino acid sequence.

The folding problem

- The following problem is a central theme of molecular biology
- In 1961, Anfineen showed that ribonucleare could be denatured and refolded without loss of enzymatic activity
- This experiment suggested that all the information for a protein to adopt its native conformation is encoded in its primary structure.
- In 1998, methods of secondary structure prediction little more than 50-60% reliable
- There are three main approaches to secondary structure prediction
  (4) Empirical statistical methods use parameters derived from known 3d structures
  (5) Methods based on physicochemical criteria
  (6) Prediction algorithms that are known structures to assign secondary structure
- Tertiary structure prediction is still further beyond reach
- In 1998, it is clear that direct prediction of structure sequence remains decoder away

Sequence analysis

- The exact nature of the information encoded in the primary structure unclear, and we still cannot read the language used to describe the final 3d fold of a biologically macro molecule.

- Using sequence analysis techniques, we can attempt to identify similarities between novel query sequences and databases sequences whose structures and functions have been elucidated

The twilight zone

- The twilight zone is a zone of sequence similarity in which alignments may appear plausible to the eye, but are no longer statistically significant
- To penetrate deeper into the twilight zone is the goal of most analytical methods
- Many different approaches have been devised, source of
- These involve database searches with single sequences
- Use characteristic chunks of sequence alignments
- Some weight database searches
- Others are only observed amino acid sequence data
- Each method offers a different perspective, depending on the type of information used in the search

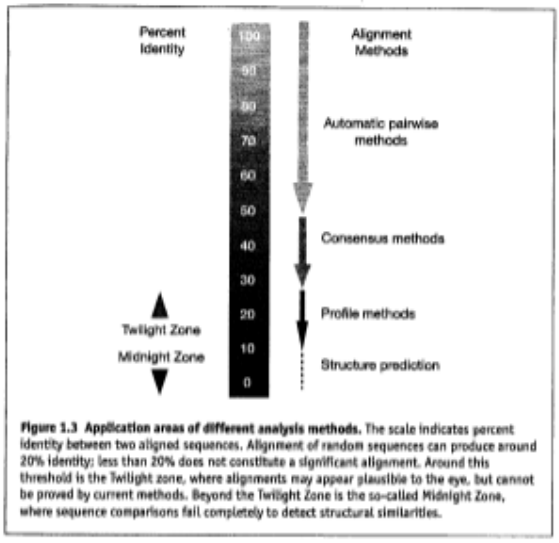


Figure 1.3 Application areas of different analysis methods. The scale indicates percent identity between two aligned sequences. Alignment of random sequences can produce around 20% identity; less than 20% does not constitute a significant alignment. Around this threshold is the Twilight zone, where alignments may appear plausible to the eye, but cannot be proved by current methods. Beyond the Twilight Zone is the so-called Midnight Zone, where sequence comparisons fail completely to detect structural similarities.

While the algorithms that recognize folds reliably, or that can predict structures, it is important to use the sequence analysis tools we have at our disposal, but in an intelligent way

Homology and Analogy

- The term homology, although easy to understand, is confounded and abused in the literature
- Sequences are said to be homologues if they are related by divergence from a common ancestor.
- For example, in everyday life, people look like one another for different reasons
- Two sisters, for example might look alike because they both inherited brown eyes and black hair from their father
- It works the same way in biology
- Some traits shared by two living things were inherit from their ancestor and some similarities evolved in other ways, these are called homologies and analogies
- Homology among proteins or DNA is typically inferred from their sequence similarity

- Significant similarity is strong evidence that two sequences are related by divergent evolution of a common ancestor
- Alignments of multiple sequences are used to indicate which regions of each sequence are homologous
- In some cases, when sequence and structure are different, we can infer with a degree of confidence that the triads result from convergent evolution
- In the former case, however, where folds are similar but the sequences differ, which such folds are usually considered to be analogues
- It is sometimes difficult to rule out the existence of a common ancestor because structures are more highly conserved than the sequence
- The errence of sequence analysis is the detection of homologues
  Sequence by means of routines database searches, usually with unknown or uncharacterised query searches

Orthology and paralogy

- Among homologous sequences, it is useful to distinguish between proteins that perform the same function in different species refund to as ORTHOLOGUES. And those that program different but related functions within one oraginsm called PARALOGUES

UNIT-II
Assignment-Cum-Tutorial Questions
SECTION-A

Objective Questions

1) _____is the set of rules by which information encoded in genetic material is translated into proteins by living cells.
2) The purine bases adenine (A) and guanine (G) are larger and consist of _____aromatic rings.
3) In RNA, thymine (T) is replaced by _____ and the deoxyribose is substituted by _____
4) If part of a sequence or structure is preserved or conserved this characteristic may be used to _____
5) Similar structures that evolved independently are called _____.
6) In _____ perspective sequence based prediction of DNA binding protein has much more speciality than DNA sequence classification.
7) _____ are homologous genes where a gene diverges after a speciation event, but the gene and its main function are conserved.
8) The human genome is: [ ]

A. All of our genes
B. All of our DNA
C. All of DNA and RNA in our cells
D. Responsible for all our physical characterstics

9) Genes are made up of [ ]

A. DNA B. RNA C. Proteins D. Enzymes

10) How many chromosomes do human have? [ ]

A. 46 B. 48 C. 54 D. 56

11) In the Sanger method of DNA sequencing, DNA synthesis ____ when a deoxy base is encountered Pattern recognition [ ]

A. Commences B. Continues C. Stops D. increases

12) Scientists now think humans have how many protein-encoding genes

A. 20-25,000 B. 30-40,000 C. 65-75,000 D. More than 100,000

13) What role does messenger RNA play in the synthesis of proteins?

A. Transported around the body to make proteins [ ]

B. Used a blueprint to assemble protein it codes for
C. Passed on from parents to children
D. Replicated within the cell

14) DNA and RNA are each made up of four chemical bases joined to a sugar and phosphate, called nucleotides, that match up with each other. In DNA, the adenine base forms special bonds to make a base pair with: [ ]

A. Guanine B. Uracil C. Cytosine D. Thymine

15) The aim of HGP [ ]

A. To identify and map 20,000-25,000 genes of humans
B. To determine chemical base pairs of DNA of humans
C. Nucleotides contained in a human haploid reference genome
D. All of these

16) The rules used to determine the link between the nucleotide sequence of a gene and the amino acid sequence of the protein specified by that gene is referred to as _____ [ ]

A. Secondary structure guidelines C. R group rules
B. Codon assignment rules D. The genetic code

17) How many different codons are possible? [ ]

A. 3 B. 20 C. 64 D. An infinite number

18) Codons that specify the amino acids often differs in the [ ]

A. First base B. Second base C. Third base D. None of these

19) What term is used to describe the process by which proteins are synthesised from a genetic code? [ ]

A. Reproduction B. Replication C. Translation D. Transcription

20) On which of the following molecules would you find a codon?[ ]

A. Messenger RNA C. Ribosomal RNA
B. Transfer RNA D. Small nuclear RNA

SECTION-B

SUBJECTIVE QUESTIONS:

1) Briefly explain the formation of genetic code.
2) Give an outline of Human Genome Project and also the research done on genome.
3) Briefly explain the folding problem
4) What are the two pattern recognition methods? Explain in detail
5) What is twilight zone? Explain in detail.
6) Explain homology and anology.
7) Is the Genome just a way of finding out where in the human genome a particular gene is located?

8) DNA sequencing and analysis has been the bread and butter of bioinformatics and computational biology for genomics and evolutionary research. Give your opinion.
9) Illustrate dideoxynucleotides have in Sanger DNA sequencing?
10) Justify the statement “Homology is not a synonym for similarity”.
11) Compare and contrast the terms Homology, Analogy, Orthology and paralogy in relation to bioinformatics.
12) What does the subject of bioinformatics deal with? State the role of bioinformatics in biological sequencing.
13) Illustrate the differences and similarities between homologous and analogous structures in bioinformatics.
14) If only a part of genome consists of gene, why sequencing the whole DNA?

SECTION-C

QUESTIONS AT THE LEVEL OF GATE

1) The lagging strand of a DNA molecule undergoing replication reads 3’-CGCATGTAGCGA-5’. What is the code of the DNA that is the template for the complementary leading strand of this segment?
2) Explain about Genetic code.
3) Explain Sequencing and biological sequence/structure
4) Describe Genome projects and Pattern recognition and prediction,
5) Analyse Folding problem and Sequence analysis using homology and anology.

UNIT –III

Objective:

- To understand basic biological databases, algorithms for proteomics and genomics analysis.

Syllabus:

Databases in Bioinformatics

Protein Information Resources- Biological databases, Primary sequence databases, Protein Sequence databases, Secondary databases, Protein pattern databases, and Structure Classification databases.

Learning Outcomes:

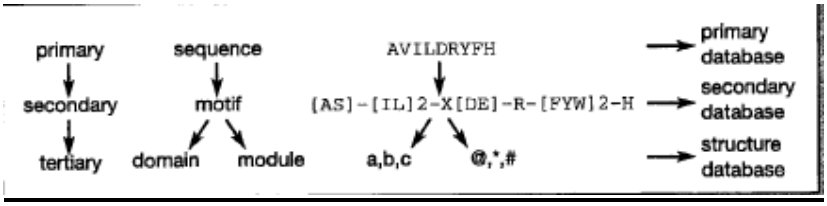
The student will be able to

- describe the contents and properties of the most important bioinformatics databases
- existing software effectively to extract information from large databases and to use this information in computer modelling.

Learning Material

Biological databases

- The first in, then in,analysing sequence information is to assemble it into central, shareable resources i.e databases.
- Databases are effectively electronic filling cabinets, a convenient and efficient method of storing vast amount of information.
- There are many different databases types, depending on both nature of the information being stored and on the manner of data storage.
- In the context of protein sequence analysis, primary and secondary databases are encountered
- Such resources different levels of information in totally different formats.
- Primary and secondary databases are used to address different aspects of sequence analysis, because they store different levels of protein sequence information.



Primary sequence databases

- In the early 1980s, sequence information started to become more abundant in the scientific literature.
- Several laboratories saw that there might be advantages to harvesting and storing these sequences in central repositories
- Several primary database projects began to evolve in different parts of the world.

Nucleic acid	*Protein*
EMBL	PIR
GenBank	MIPS
DDBJ	SWISS-PROT
	TrEMBL
	NRL-3D

- The above table shows some of the most important nucleic acid and protein sequence databases that arose with such intiatves

Nucleic acid sequence databases

- The principal DNA sequence databases are GenBank(USA), EMBL(Europe) and DDBJ(Japan), which exchange data on a daily basis to ensure comprehensive coverage at each of the sites.

Protein sequence databases

- The Protein Sequence Database was developed at the National Biomedical Research Foundation(NBRF) in the early 1960s by Maraget Dayhoff as a collection of sequences for investigating evolutionary relationships among proteins.
- Since 1988, the Protein Sequence Database has been maintained by PIR-International, an association of macromolecular sequence data collection centres; the consortium includes the Protein Information Resource(PIR) at the NBRF.
- The database is split into four distinct sections, designated PIR1-PIR4, which differ in terms of the quality of data and level of annotation provided: PIR1 contains fully classified and annotated entries; PIR2 includes preliminary

entries, which have not been thoroughly reviewed and PIR3 contains unverified entries; which have not been verified.

- PIR4 entries fall into one of four categories
    a) Conceptual transactions of artefactual sequences
    b) Conceptual translations of sequences that are not transcribed or translated
    c) Protein sequences or conceptual translations that are extensively genetically engineered
    d) Sequences that are not genetically encodedand not produced on ribosomes

MIPS

The Martinsried Institute for Protein Sequences collects and processes sequence data for the tripartite PIR-International Protein Database project.

- The Munich Information Center for Protein Sequences (MIPS-GSF, Neuherberg, Germany) continues to provide genome-related information in a systematic way. MIPS supports both national and European sequencing and functional analysis projects, develops and maintains automatically generated and manually annotated genome-specific databases, develops systematic classification schemes for the functional annotation of protein sequences, and provides tools for the comprehensive analysis of protein sequences.

SWIPS-PROT

- SWISS-PROT is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the European Molecular Biology Laboratory (EMBL), since 1987. SWISS-PROT is now an equal partnership between the EMBL and the Swiss Institute of Bioinformatics (SIB). The EMBL activities are carried out by its Hinxton Outstation, the European Bioinformatics Institute (EBI)
- SWISS-PROT strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.
- Recent developments of the database include format and content enhancements, cross-references to additional databases, new documentation files and improvements to TrEMBL, a computer-annotated supplement to SWISS-PROT.

- TrEMBL consists of entries in SWISS-PROT-like format derived from the translation of all coding sequences (CDSs) in the EMBL Nucleotide Sequence Database, except the CDSs already included in SWISS-PROT.
- We also describe the Human Proteomics Initiative (HPI), a major project to annotate all known human sequences according to the quality standards of SWISS-PROT.
- The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria: (i) annotations, (ii) minimal redundancy and (iii) integration with other databases.

Structure of SWISS-PROT entries

- Each sequence entry is composed of lines. Different types of lines, each with their own format, are used to record the various data which make up the entry.
- Each line begins with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order in which they appear in an entry, are shown below:

```
ID      - Identification.
AC      - Accession number(s).
DT      - Date.
DE      - Description.
GN      - Gene name(s).
OS      - Organism species.
OG      - Organelle.
OC      - Organism classification.
RN      - Reference number.
RP      - Reference position.
RC      - Reference comments.
RX      - Reference cross-references.
RA      - Reference authors.
RL      - Reference location.
CC      - Comments or notes.
DR      - Database cross-references.
KW      - Keywords.
FT      - Feature table data.
SQ      - Sequence header.
        - (blanks) sequence data.
        - Termination line.
```

- The program ignores all the description lines and uses only these line types: 'ID', 'DE', 'OS', 'SQ' and '//'.
- The program uses the 'ENTRY_NAME' which is the first field of the ID line as the first line of the title
- The data of the 'DE' and 'OS' lines are collected by the program and are used as the remaining lines of the title

- The 'SQ' line is used to identify the beginning of the sequence. The program collect all the following lines until the teminalion line is found or end is reached

TrEMBL

- TrEMBL is a computer-annotated protein sequence database supplementing the Swiss-Prot Protein Sequence Data Bank.
- TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database not yet integrated in Swiss-Prot. TrEMBL can be considered as a preliminary section of Swiss-Prot.
- For all TrEMBL entries which should finally be upgraded to the standard Swiss-Prot quality, Swiss-Prot accession numbers have been assigned.
- TrEMBL is split in two main sections: SP-TrEMBL and REM-TrEMBL:
- SP-TrEMBL (Swiss-Prot TrEMBL) contains the entries (300'192) which should be eventually incorporated into Swiss-Prot. Swiss-Prot accession numbers have been assigned for all SP-TrEMBL entries.

NRL-3D

- The NRL-3D is produced by PIR from sequences extracted from the Brookhaven Protein Databank(PDB)
- NRL-3D is valuable resource, as it makes the sequence information in the PDB available both for keyword interrogation and for similarity searches
- The database may be searched using ATLAS retrieval system, a multi-database information retrieval program specifically designed to access macromolecular sequence databases.

Secondary databases

- By contrast, **secondary databases** comprise data derived from the results of analysing primary data.
- They are often referred to as curated databases but this is a bit of a misnomer because primary databases are also curated to ensure that the data in them is consistent and accurate.
- Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary), controlled vocabularies (see later section) and the scientific literature.
- They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science.

Essential aspects of primary and secondary databases.

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

- SWISS-PROT has emerged as the most popular primary sources and many secondary databases now use it as their basis

Secondary database	*Primary source*	*Stored information*
PROSITE	SWISS-PROT	Regular expressions (patterns)
Profiles	SWISS-PROT	Weighted matrices (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)

*SWISS-PROT is OWL's highest priority source.

PROSITE

- The first secondary database have been created is PROSITE.
- PROSITE is a protein database.
- It consists of entries describing the protein families, domains and functional sites as well as amino acid patterns and profiles in them.

- These are manually curated by a team of the Swiss Institute of Bioinformatics and tightly integrated into Swiss-Prot protein annotation. PROSITE was created in 1988 by Amos Bairoch, who directed the group for more than 20 years. Since July 2009, the director of the PROSITE, Swiss-Prot and Vital-IT groups is Ioannis Xenarios.
- PROSITE's uses include identifying possible functions of newly discovered proteins and analysis of known proteins for previously undetermined activity. Properties from well-studied genes can be propagated to biologically related organisms, and for different or poorly known genes biochemical functions can be predicted from similarities.
- PROSITE offers tools for protein sequence analysisand motif detection (see sequence motif, PROSITE patterns). It is part of the ExPASy proteomics analysis servers.
- The database ProRule builds on the domain descriptions of PROSITE.
- It provides additional information about functionally or structurally critical amino acids.
- The rules contain information about biologically meaningful residues, like active sites, substrate- orco-factor-binding sites, posttranslational modification sites or disulfide bonds, to help function determination.
- These can automatically generate annotation based on PROSITE motifs.
- PROSITE consists of documentation entries describing protein domains, families and functional sites, as well as associated patterns and profiles to identify them.
- It is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of these profiles and patterns by providing additional information about functionally and/or structurally critical amino acids.
- The PROSITE database uses two kinds of signatures or descriptors to identify conserved regions, i.e. patterns and generalized profiles, both having their own strengths and weaknesses defining their area of optimum application.
- Each PROSITE signature is linked to an annotation document where the user can find information on the protein family or domain detected by the signature, such as the origin of its name, taxonomic occurrence, domain architecture, function, 3D structure, main characteristics of the sequence, domain size and literature references.

PRINTS

- In molecular biology, the PRINTS database is a collection of so-called "fingerprints".
- It provides both a detailed annotation resource for protein families, and a diagnostic tool for newly determined sequences.
- A fingerprint is a group of conserved motifs taken from a multiple sequence alignment - together, the motifs form a characteristic signature for the aligned protein family.

- The motifs themselves are not necessarily contiguous in sequence, but may come together in 3D space to define molecular binding sites or interaction surfaces.
- The particular diagnostic strength of fingerprints lies in their ability to distinguish sequence differences at the clan, superfamily, family and subfamily levels.
- This allows fine-grained functional diagnoses of uncharacterised sequences, allowing, for example, discrimination between family members on the basis of the ligands they bind or the proteins with which they interact, and highlighting potential oligomerisation or allosteric sites.
- PRINTS is a founding partner of the integrated resource, InterPro, a widely used database of protein families, domains and functional
- The fingerprinting method arose from the need for a reliable technique for detecting members of large, highly divergent protein super-families.
- The idea was to exploit the most conserved regions within sequence alignments to build diagnostic signatures of family membership.
- In a database search, there would then be a greater chance of identifying a distant relative, whether or not all parts of a signature were matched (providing the motifs were found in the correct order and the distances between them were consistent with those expected of true neighbouring motifs).
- The ability to tolerate mismatches, both at the level of residues within individual motifs, and at the level of motifs within entire fingerprints, rendered fingerprinting a powerful diagnostic approach.
- PRINTS was formerly built as a single ASCII (text) file. With the continued growth of the database, however, maintenance was becoming inefficient and error-prone.
- We have therefore designed an object-relational schema, which places existing database fields (e.g., relating to motifs, sequence data, true and false assignments, etc.) into separate but related tables.

BLOCKS

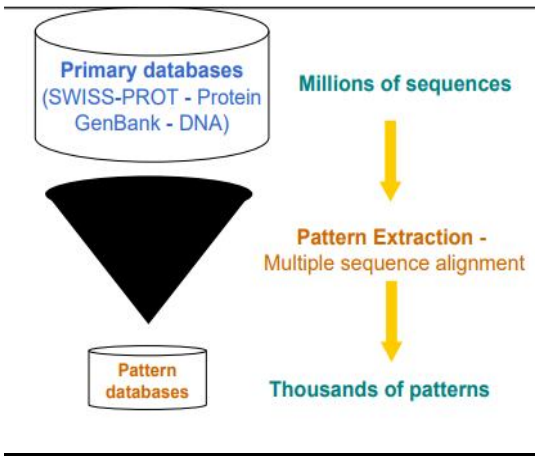
- Many known proteins can be grouped into families according to functional and sequence similarities.
- The similarity of the proteins across the sequences in each family is far from uniform.
- While some regions are clearly conserved, others display little sequence similarity
- The description of a protein family by its conserved regions focuses on the family's characteristic and distinctive sequence features, thus reducing noise.
- Databases of conserved features of protein families can be utilized to classify sequences from proteins, cDNAs and genomic DNA.
- An example is the Blocks Database (3), which consists of ungapped multiple alignments of short regions, called 'blocks'.
- The database was constructed from sequences of protein families using a fully automated method.
- Searching the Blocks database with a sequence query allows detection of one or more blocks representing a family.

- The BLIMPS (Blocks IMProved Searcher) program searches the Blocks Database.
- BLIMPS transforms each block into a position specific scoring matrix (PSSM), sometimes called a profile.
- Each PSSM column corresponds to a block position and contains values based on the amino acid frequencies in each position.
- To prevent domination of the PSSM by a large subgroup of related sequences, each sequence segment in a block is weighted using position-based sequence weights.
- To reduce the effect of small sequence samples, the amino acid frequencies in each PSSM position (observed counts) are supplemented with artificial 'pseudo-counts'.
- Currently we model pseudo-counts on amino acid substitution probabilities (13; SH and JGH, unpublished results).
- BLIMPS compares a query sequence with a block by sliding the PSSM over the sequence (nucleotide sequences are translated in all the frames into six amino acid sequences).
- For every alignment, each sequence position receives the value of its amino acid in the aligned PSSM column.
- These scores are summed to obtain the score of the sequence segment. This is repeated with all blocks in the database, and the top scores are saved.
- In addition to searching a sequence against a database of blocks, BLIMPS can search a block against a database of sequences.

Protein pattern databases

Definition

Secondary databases derived from conserved obtained from multiple sequence alignment of primary databases such as GenBank, EMBL,DDBJ, SP/TrEMBL,PIR,etc



- Composite database render sequence searching much more efficient, because they obviate the need to interrogate multiple resource.
- A composite resource will be non-identical if it eliminates only identical sequence copies during the amalgamation process.

- The choice of different sources and the application of different redundancy criteria have led to the emergence of different composites.
- These databases are mainly created by analytical methods from the primary database.
- At the heart of the analysis methods that underpin pattern databases is the multiple sequence alignment method and different SID Protein Databases techniques have evolved to exploit the fact of classifying proteins in different pattern databases.
- The different analytical methods to create pattern databases are given

Applications:

- Function prediction of protein/ nucleotide sequences even when sequence similarity is low (< 25%).
- Useful for classification of protein sequences into families.
- It takes less time to search the pattern than the primary database.
- Since "patterns" is the compact representation of features of many sequences

- The different analytical methods to create pattern databases are given below:
- **Single motif methods.**
  - The idea is that a particular protein family can be characterized by the single most conserved region,
- **Multiple motif methods**.
  - This method finds several motifs that characterize the aligned family within a sequence alignment in a database search,
  - Therefore a greater chance of identifying a distant relative exists
- **Profile methods**
  - This method uses the variable regions between conserved motifs. In this method, the complete conserved position of the alignment (including gaps), becomes a discriminator or a profile, this profile defines which residues are allowed at the given positions, which residues are allowed at conserved and which degenerate.
  - Profiles are sometimes related weight matrices and they provide a sensitive mean of detecting.
  - Distant sequence relationships, where only a few residues are well conserved can be found using this method
- The different methods mentioned above have given rise to different pattern databases, despite the data source; these pattern databases are emerged from the conserved motifs shared in homologous sequences, which is thought to be crucial to the structure and function of proteins.
- Therefore searching the pattern database theoretically offers an insight to the sequence biological function.
- Since these pattern databases are derived from the multiple sequence alignment methods, one can identify a distant relationship between the sequences

STRUCTURE CLASSIFICATION DATABASES

- These databases provide structural comparisons for the proteins currently in the Brookhaven PDB and access to the sequences of these proteins.
- Each database offers different family coverage and different levels of annotation.
- This is vital for a user, who not only wants to discover whether a sequence has matched a predefined motif, but also needs to understand its biological significance

- Many proteins share structural similarities, reflecting, in some cases, common evolutionary origins
- The evolutionary process involves substitutions, insertions and deletions in amino acid sequences

Two well-known classification schemes

SCOPE

- The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domainsbased on similarities of their structures and amino acid sequences.
- A motivation for this classification is to determine the evolutionary relationship between proteins.
- Proteins with the same shapes but having little sequence or functional similarity are placed in different super families, and are assumed to have only a very distant common ancestor.
- Proteins having the same shape and some similarity of sequence and/or function are placed in "families", and are assumed to have a closer common ancestor.
- The SCOP database is freely accessible on the internet.
- SCOP was created in 1994 in the Centre for Protein Engineeringand the Laboratory of Molecular Biology.
- It was maintained by Alexey G. Murzin and his colleagues in the Centre for Protein Engineering until its closure in 2010 and subsequently at the Laboratory of Molecular Biology in Cambridge, England.

SCOP Classification:

- The source of protein structures is the Protein Data Bank.
- The unit of classification of structure in SCOP is the protein domain.
- The shapes of domains are called "folds" in SCOP.
- Domains belonging to the same fold have the same major secondary structures in the same arrangement with the same topological connections.

The levels of SCOP are as follows.

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

CATH

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

The four main levels of the CATH hierarchy:

#	Level	Description
1	**C**lass	the overall secondary-structure content of the domain. (Equivalent to the SCOP Class)
2	**A**rchitecture	high structural similarity but no evidence of homology. (Equivalent to the 'fold' level in SCOP)
3	**T**opology/fold	a large-scale grouping of topologies which share particular structural features
4	**H**omologous superfamily	indicative of a demonstrable evolutionary relationship. (Equivalent to SCOP superfamily)

- The domains are then classified within the CATH structural hierarchy: at the Class (C) level, domains are assigned according to their secondary structure content, i.e. all alpha, all beta, a mixture of alpha and beta, or little secondary structure.
- At the Architecture (A) level, information on the secondary structure arrangement in three-dimensional space is used for assignment.
- At the Topology/fold (T) level, information on how the secondary structure elements are connected and arranged is used.
- Assignments are made to the Homologous superfamily (H) level if there is good evidence that the domains are related by evolution i.e. they are homologous.

PDBsum

- PDBsum is a web-based database providing a largely pictorial summary of the key information on each macromolecular structure deposited at the Protein Data Bank (PDB).
- It includes images of the structure, annotated plots of each protein chain's secondary structure, detailed structural analyses generated by the PROMOTIF program, summary PROCHECK results and schematic diagrams of protein–ligand and protein–DNA interactions.
- To date, the 3D structures of over 13 000 biological macromolecules have been determined experimentally, principally by X-ray crystallography and NMR spectroscopy.
- The majority of these are protein structures, including protein–DNA and protein–ligand complexes.
- Together with sequence, physicochemical and functional annotations they provide a wealth of information crucial for the understanding of biological processes.
- Each new structure is deposited in the Protein Data Bank (PDB), which is currently run by the Research Collaboratory in Structural Biology (RCSB).
- The structures can be downloaded from the RCSB's PDB web server, which also provides additional information about each one.

UNIT-III
Assignment-Cum-Tutorial Questions
SECTION-A

Objective Questions

1) Margaret Dayhoff developed the first protein sequence [ ]
a) SWISS PROT b) PDB
c) Atlas of protein structure d) Protein sequence databank
2) Translated EMBL was created in 1996 as a computer annotated supplement to_____.
3) Which of the following ia protein sequence database [ ]
a) DDBJ b) EMBL c) GenBank d) PIR
4) The first secondary database developed was [ ]
a) PRINTS b) PROSITE c) PDB d) PIR
5) GeneBank and SWISSPORT are example of [ ]
a) primary database c) secondary data base
b) composite database d) none
6) SWISS PROT is related to [ ]
a) Portable data c) Swissbank data
b) Sequence data bank d) Sequence sequence data
7) Which one of the following is not a primary nucleic acid database?
a) GenBank. b) DDBJ. c) EMBL. d) TREMBL [ ]
8) _____is a composite database. [ ]
a) PROSITE. b) DDBJ. c) NRDB. d) EMBL
9) Which one of the following is not a primary nucleic acid database?
a) GenBank. b) DDBJ. c) EMBL d) TREMBL [ ]
10) SWISS PROT protein sequence database began in [ ]
a) 1985 b) 1986 c) 1987 d) 1988

11) A comprehensive database for the study of human genetics and molecular biology [ ]
a) PDB b) STAG c) OMIM d) PSD
12) All th following are protein sequence database except [ ]
a) PIR b) PSD c) SWISS PROT d) EMBL
13) Information of all known nucleotide and protein sequences are available on [ ]
a) EMBL b) DDBT c) NCBI's Gene Bank d) All of these
14) SCOP is [ ]
a) It is primary database
b) It is nucleotide sequence database
c) SCOP database is a hierarchial classification of protein 2D domain structures

d) Structural database, which identity and evolutionary relationships

15) PDB is [ ]

a) Primary database for macromolecules
b) Can be determined by gel electrophesis
c) Composite database
d) Database for three dimensional structure of biological macromolecule

16) PRINTS are software used for [ ]

a) detection of genes from genome sequence
b) detection of tRNA genes
c) prediction of function of a new gene
d) Identification of functional domains/motifs of protein

17) Which one of the following is a complementary DNA database?[ ]

a) SwissProt. b) GenBank c) UniSTS. d) NRDB

18) STAG is maintained by [ ]

a) Brookhaven laboratory
b) DNA database of Japan
c) European Molecular Biology Laboratory
d) National Centre for Biotechnology Information

19) FASTA' was published by [ ]

a) Joseph Sambrook c) Pearson and Lipman
b) Sanger d) Altschul et al

20) CATH shares maximum of resource with _____ [ ]

a) SCOP b) PROSITE c) Pfam d) BLOCKs

SECTION-B

SUBJECTIVE QUESTIONS

1) Explain the primary sequence databases?
2) Describe protein classification on SCOP database.
3) Briefly explain the essential aspects of primary and secondary databases.
4) Illustrate the structure of SWISS PROT entries.
5) Explain the applications protein pattern databases.
6) Describe the features and importance of NCBI.
7) Classify and explain major databases in bioinformatics giving examples of each database.
8) Explain the steps for data mining and knowledge discovery of biological databases.
9) Write the file format of EMBL Nucleotide Sequence Database.
10) How to compute the physical Properties Based on Sequence?

11) Justify the statement "PRINTS is a compendium of protein fingerprints".
12) Why create secondary databases?
13) Highlight the importance of SCOP,CATH and PROSITE databases towards prediction exercises.
14) Write a comparative note on PDB and MMDB.

SECTION-C

QUESTIONS AT THE LEVEL OF GATE

1) Explain the parameters percentage identify, percentage similarity, E-value and Gap penalty scores in the context of pair-wise alighnments**.**

UNIT –IV

Objective:

- To understand basic biological databases, algorithms for proteomics and genomics analysis.

Syllabus:

Genome Information Resources

DNA sequence databases, specialized genomic resources

Learning Outcomes:

The student will be able to

- understand how to find a DNA sequence and save it in the correct format
- interpret sequence analysis results and understand the biological impact of functional regions

Learning Material

DNA sequence databases

EMBL:

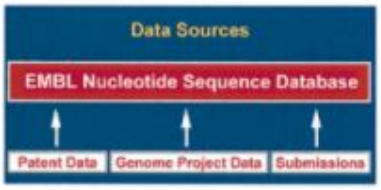
- The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database is maintained at the European Bioinformatics Institute (EBI) in an international collaboration with the DNA Data Bank of Japan (DDBJ) and GenBank (USA).
- The major contributors to the EMBL database are individual authors and genome project groups.
- DNA and RNA sequences are directly submitted from researchers and genome sequencing groups and collected from the scientific literature and patent applications.
- The database is produced in collaboration with DDBJ and GenBank; the participating groups each collect a portion of total sequence data reported worldwide, and all new updated entries are then exchanged between the groups on a daily basis.
- The EMBL Databasecollects, organizes and distributes a database of nucleotide sequence data and related biological information.
- Since 1982 this work has been done in collaboration with GenBank (NCBI, Bethesda, USA) and the DNA Database of Japan (Mishima).
- Each of the three international collaborating databases DDBJ/EMBL/GenBank, collect a portion of the total sequence data reported world-wide.
- All new and updated database entries are exchanged between the International Nucleotide Sequence Collaboration on a daily basis.
- EMBL Database releases are produced quarterly and are distributed on CD-ROM.
- The most up-to-date data collection is available via Internet and World Wide Web interface.

- The explosive growth of the database continues.
- Currently at over 1200 million base pairs, the database almost doubles in size each year, with a new sequence being deposited on average once a minute.
- Sequencing projects like the Human Genome Project and a growing number of other genome sequencing groups produce large quantities of new sequence data.
- The recent increase in data volume is a direct consequence of ongoing collaborations between major sequencing projects and the EMBL Database.

Database divisions

- EMBL Database entries are grouped into divisions.
- The grouping is based mainly on taxonomy with a few exceptions like the new HTG (High Throughput Genome Sequences) and GSS (Genome Survey Sequences) divisions, for which grouping is based on the specific nature of the underlying data.
- Thus, divisions provide subsets of the database which reflect the areas of interest of many of our users.
- The EMBL Database currently consists of 17 divisions with each entry belonging to exactly one division.
- In each entry the according division is indicated using the three letter codes as shown below:

Division	Code
Bacteriophage	PHG
ESTs	EST
Fungi	FUN
High throughput genome	HTG
Genome survey sequences	GSS
Human	HUM
Invertebrates	INV
Organelles	ORG
Other mammals	MAM
Other vertebrates	VRT
Plants	PLN
Prokaryotes	PRO
Rodents	ROD
STSs	STS



Database entry structure

- Database entries are distributed in EMBL flat-file format which is supported by most sequence analysis software packages and also provides a structure usable by human readers.
- A typical database entry contains a brief description for cataloging purposes, the taxonomic description of the source organism, reference information, a sequence and the feature table describing locations of coding regions and other biologically significant sites.
- Different line types are used to record the various types of data which make up the sequence entry, e.g., DE, description; OS, organism species; RT, reference title; FT, feature table data; etc.

DDBJ:

- The DNA Data Bank of Japan (DDBJ) is a biological database that collects DNA sequences.
- It is located at the National Institute of Genetics (NIG) in the Shizuoka prefecture of Japan.
- It is also a member of the International Nucleotide Sequence Database Collaboration or INSDC.
- It exchanges its data with European Molecular Biology Laboratory at the European Bioinformatics Institute and with GenBank at the National Center for Biotechnology Information on a daily basis.
- Thus these three databanks contain the same data at any given time.
- The DNA Data Bank of Japan (DDBJ) provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval and annotation analysis.
- A new archive for quantitative genomics data, the DDBJ Omics aRchive (DOR). The DOR stores quantitative data both from the microarray and high-throughput new sequencing platforms.
- Other improvements include improved content of the DDBJ patent sequence, released a new submission tool of the DDBJ Sequence Read Archive (DRA) which archives massive raw sequencing reads, and enhanced a cloud computing-based analytical system from sequencing reads, the DDBJ Read Annotation Pipeline.
- The objective of DDBJ is to support and promote the sharing and use of biological data as a public resource.

GenBank:

- The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations.
- This database is produced and maintained by the National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration (INSDC).
- The National Center for Biotechnology Information is a part of the National Institutes of Health in the United States.
- GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms.
- The database started in 1982 by Walter Goad and Los Alamos National Laboratory.
- GenBank has become an important database for research in biological fields and has grown in recent years at an exponential rate by doubling roughly every 18 months.
- Release 194, produced in February 2013, contained over 150 billion nucleotide bases in more than 162 million sequence.
- GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

GSDB:

- In 1997 the primary focus of the Genome Sequence DataBase (GSDB) located at the National Center for Genome Resources was to improve data quality and accessibility.
- Efforts to increase the quality of data within the database included two major projects; one to identify and remove all vector contamination from sequences in the database and one to create premier sequence sets (including both alignments and discontiguous sequences).
- Data accessibility was improved during the course of the last year in several ways. First, a graphical database sequence viewer was made available to researchers. Second, an update process was implemented for the web-based query tool, Maestro. Third, a web-based tool, Excerpt, was developed to retrieve selected regions of any sequence in the database. And lastly, a GSDB flatfile that contains annotation unique to GSDB (e.g., sequence analysis and alignment data) was developed.

dBEST:

- dbEST is a division of **GenBank** that contains sequence data and other information on "single-pass" cDNA sequences, or "Expressed Sequence Tags", from a number of organisms.
- EST sequences are included in the EST division of GenBank, available from NCBI by anonymous ftp and through Entrez.
- The nucleotide sequences may be searched using the BLAST electronic mail server. The TBLASTN program which takes an amino acid query sequence and compares it with six-frame translations of dbEST DNA sequences is particularly useful.

Specialized genomic resources

SGD:

- The 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.
- The Saccharomyces Genome Database (SGD) provides Internet access to the complete Saccharomyces cerevisiae genomic DNA sequence, its genes and their products, the phenotypes of its mutants, and the literature supporting these data.
- The amount of information and the number of features provided by SGD have increased greatly following the release of the S. cerevisiae genomic sequence.
- SGD aids researchers by providing not only basic information, but also tools such as sequence similarity searching that lead to detailed information about features of the genome and relationships between genes.
- SGD presents information using a variety of user-friendly, dynamically created graphical displays illustrating physical, genetic and sequence feature maps.
- The biocurators at SGD aim to annotate each gene by identifying function(s) from primary literature and linking to terms using the structured knowledge representation in the Gene Ontology.
- Additionally, functions identified from high throughput experiments as well as computationally predicted function annotations are included from GO Annotation project.

UniGene:

- The UniGene resource, developed at NCBI, clusters ESTs and other mRNA sequences, along with coding sequences (CDSs) annotated on genomic DNA, into subsets of related sequences.
- In most cases, each cluster is made up of sequences produced by a single gene, including alternatively spliced transcripts. However, some genes may be represented by more than one cluster.
- The clusters are organism specific and are currently available for human, mouse, rat, zebrafish, and cattle. They are built in several stages, using an automatic process based on special sequence comparison algorithms.
- First, the nucleotide sequences are searched for contaminants, such as mitochondrial, ribosomal, and vector sequence, repetitive elements, and low-complexity sequences.
- After a sequence is screened, it must contain at least 100 bases to be a candidate for entry into UniGene. mRNA and genomic DNA are clustered first into gene links.
- A second sequence comparison links ESTs to each other and to the gene links. At this stage, all clusters are ‘‘anchored,’’ and contain either a sequence with a polyadenylation site or two ESTs labeled as coming from the 3 end of a clone.

- Clone-based edges are added by linking the 5 and 3 ESTs that derive from the same clone. In some cases, this linking may merge clusters identified at a previous stage.
- Finally, unanchored ESTs and gene clusters of size 1 (which may represent rare transcripts) are compared with other UniGene clusters at lower stringency. The UniGene build is updated weekly, and the sequences that make up a cluster may change.
- Thus, it is not safe to refer to a UniGene cluster by its cluster identifier; instead, one should use the GenBank accession numbers of the sequences in the cluster.
- As of July 2000, the human subset of UniGene contained 1.7 million sequences in 82,000 clusters; 98% of these clustered sequences were ESTs, and the remaining 2% were from mRNAs or CDSs annotated on genomic DNA.
- These human clusters could represent fragments of up to 82,000 unique human genes, implying that many human genes are now represented in a UniGene cluster. (This number is undoubtedly an overestimate of the number of genes in the human genome, as some genes may be represented by more than one cluster.) Only 1.4% of clusters totally lack ESTs, implying that most human genes are represented by at least one EST.
- Conversely, it appears that the majority of human genes have been identified only by ESTs; only 16% of clusters contain either an mRNA or a CDS annotated on a genomic DNA.
- Because fewer ESTs are available for mouse, rat, and zebrafish, the UniGene clusters are not as representative of the unique genes in the genome.
- Mouse UniGene contains 895,000 sequences in 88,000 clusters, and rat UniGene contains 170,000 sequences in 37,000 clusters.
- Homologs are identified as the best match between a UniGene cluster in one organism and a cluster in a second organism.
- When two sequences in different organisms are best matches to one another (a reciprocal best match), the UniGene clusters corresponding to the pair of sequences are considered putative orthologs.
- A special symbol indicates that UniGene clusters in three or more organisms share a mutually consistent ortholog relationship.
- It is important to note that clusters that contain ESTs only (i.e., no mRNAs or annotated CDSs) will be missing some of these fields, such as LocusLink, OMIM, and mRNA/Gene links.
- UniGene titles for such clusters, such as ‘‘EST, weakly similar to ORF2 contains a reverse transcriptase domain [H. sapiens],’’ are derived from the title of a characterized protein with which the translated EST sequence aligns.
- The cluster title might be as simple as ‘‘EST’’ if the ESTs share no significant similarity with characterized proteins.

TDB :

- The TIGR database(TDB) provides a substancial suite of databases containing DNA and protein sequence, gene expression,cellular role and protein family information and taxonomic data for microbes, plants and humans.
- The resources include a microbial database that links to worldwide and TIGR genome sequencing projects

ACeDB:

- ACeDB is 'A C elegans DataBase' arising from the C.elegans genome project.
- The resource includes restriction maps, gene structural information and so on.
- The software designed to organise and browse the information known as ACEDB presents a graphical interface that enables the user to view genomic data at different stages of resolution from the level of a complete chromosome down to the physical level
- The use of ACeDB and ACEDB to refer to both the database and the software can lead to confusion, so users should be aware of the distinction.

UNIT-IV
Assignment-Cum-Tutorial Questions
SECTION-A

Objective Questions

1) The full form of EMBL_____.
2) The EMBL Database currently consists of _____with each entry belonging to exactly one division.
3) DDBJ is located at the _____.
4) The National Center for Biotechnology Information is apart of _____.
5) The GeneBank database started in_____by_____
6) The full form of SGD_____
7) The primary focus of the Genome Sequence DataBase is _____
8) _____ are identified as the best match between a UniGene cluster in one organism and a cluster in a second organism.
9) Which of the following is the first biological database? [ ]

A. GenBank B. DDBJ C. Atlas of Protein Structure D. OMIM

10) GenBank is maintained by [ ]

A. EBI B. NIG C. NCBI D. SIB

11) Approximately what proportion of the human genome is made up of repetitive DNA sequences? [ ]

A. 1% B. 15% C. 50% D. 90%

12) The biological sequence data was first published in [ ]

A. 1962 B. 1963 C. 1964 D. 1965

13) _____ database was found in Maryland [ ]

A. GDB B. GTD C. GSDB D. PIR

14) One of the following is not a major nucleotide database. Which is it?

A. GenBank B. PDB C. EMBL D. DDBJ [ ]

15) This database was started in 1979 and is maintained by NCBI of the United States since 1992. [ ]

A. GenBank B. PDB C. UniProt D. DDBJ

16) The nucleotide database established in Europe, specifically in Heidelberg in 1980, and maintained by EBI-Cambridge since 1994. [ ]

A. GenBank B. Swiss-Prot C.DDBJ D. EMBL

17) The premier database for protein structure [ ]

A. PDV B. EMBL C. ProtBank D. Protein Data Bank

18) Your TA tells you to go to the NCBI Human Genome page. What does she probably want you to do? [ ]

A. Determine what genes are around your protein gene on its chromosome

B. Identity a DNA sequence and see if it came from a human

C. Look up papers about diseases caused by abornormalities in certain protein

D. Look at colourful, rotating 3-D pictures of the tertiary structure of a protein

19) During 2009-10, DDBJ contributed _____ of the entries and ____of the bases added to INSDC [ ]

A. 25.4%, 21.5% B. 21.5%,25.4% C. 25.4%,25.4% D. 21.5%,21.5%

20) The most importance source of new data for GenBank is [ ]

A. Direct submission from scientists C.Indirect submission from scientists

B. Direct and indirect submission from scientists D. No submission

SECTION-B

SUBJECTIVE QUESTIONS

1) Write the file format of EMBL Nucleotide Sequence Database
2) What type of information is present in GDB?
3) Briefly explain the database entries of EMBL.
4) What kinds of data are acceptable at DDBJ?
5) Explain the terms TDB and ACeDB.
6) Explain about the clustring in UniGene.
7) Are there homologues in the databases?
8) Compare and contrast UniGene and GeneBank
9) Which type of databases are used in bioinformatics
10) Give a survey on how specialized genomic resources are used in bioinformatics.
11) Categorize all the databases in DNA sequence databases
12) Categorize the differences of clusters present in different species that are stored in UniGene.
13) Compare and contrast dBESt and GSDB
14) Explain how data can be analyzed in bioinformatics.

SECTION-C

QUESTIONS AT THE LEVEL OF GATE

1) Random clone assembly requires that substantially more primary sequence be determined than does map-based assembly. Why is random clone assembly now used more commonly for the determination of complete genome sequences?

UNIT –V

Objective:

- To understand basic biological databases, algorithms for proteomics and genomics analysis.

Syllabus:

Alignment Techniques

Pair-wise alignment techniques- database searching, alphabets and complexity, Algorithms and programs,comparing two sequences, sub-sequences, Identity and similarity, the Dotplot, Local and global similarity, Different alignment techniques,dynamic programming, Pair-wise database searching

Learning Outcomes:

The student will be able to

- explain the major steps in pairwise and multiple sequence alignment, explain the principle for, and execute pairwise sequence alignment by dynamic programming

Learning Material

Pair-wise alignment techniques:

Sequence alignment:

- Sequence alignment is a fundamental procedure (implicitly or explicitly) conducted in any biological study that compares two or more biological sequences (whether DNA, RNA, or protein).
- It is the procedure by which one attempts to infer which positions (sites) within sequences are homologous, that is, which sites share a common evolutionary history.
- it is important to remember that alignment algorithms produce a hypothesis of homology.
- these alignments may contain more or less error depending on the nature of the data, some of which may have huge downstream effects on other analyses.

Pair-wise alignment:

Pair-wise comparison is a fundamental process in sequence analysis, underpinning as it does, database search algorithms, which seek out relationships based on properties rather than simple interrogation of textual annotation.

Database searching:

- Database interrogation can take the form of text queries or sequence similarity searches.
- Text-based querying has its place in the armoury of the sequence analysis software and should certainly not be overlooked in any worthwhile analysis.
- The purpose of database searching is to describe how the relationships between a query sequence and another sequence can be quantified and their similarity assessed.

- In oder to identify an evolutionary relationship between newly determined sequence and known gene family, we need to assess the extent of shared similarity.
- When the degree of similarity is low, the relationship must remain putative,until evidence has been collected.
- In order to assess effectively the results of database searches, we need to understand the way in which tools work.

Alphabets and complexity:

- A sequence consists of letters from an alphabet.
- The complexity of alphabet can be defined using number of different letters it contains.
- For example, the complexity of English language is 26, the complexity of DNA is 4.
- Sometimes additional characters are used in an alphabet to indicate the degree of ambiguity in the identity of particular residue or base.

Algorithms and programs:

- It is important to note the difference between an algorithm and a program
- The former is the set of steps that define some computational process at an abstract level, the latter is the implementation of an algorithm.
- There are many different implementations of the same algorithm, but these should give the same results, if the algorithm has been clearly defined.

Comparing two sequences:

- Here we need to consider how to develop algorithm for determining the similarity between two sequences, each selected from an alphabet of complexity 20.
- The naive approach is to line up the sequences against each other and insert additional characters to bring the two strings in vertical alignment.

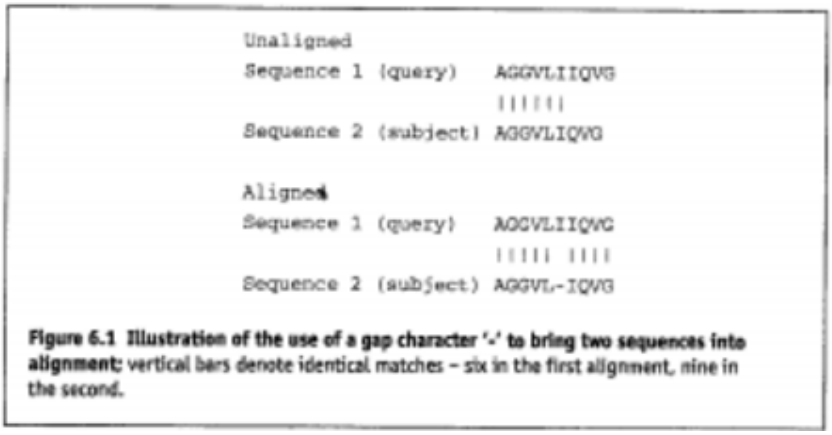


Figure 6.1 Illustration of the use of a gap character '-' to bring two sequences into alignment: vertical bars denote identical matches – six in the first alignment, nine in the second.

- At this point the task is complete
- We could score the alignment by counting how many positions match identically at each position , here the unaligned score is 6, while the aligned score is 9.

- In the above example, we can see that the score increases when more identical residues have been aligned.
- But this is only illustration, the sequences are very short, the sequences are almost the same length and there are nearly identical anyway.
- The process of alignment can be measured in terms of the number of gaps introduced and the number of mismatches remaining in the alignments.
- A metric relating such parameters represents the distance between two sequences so called edit-distance.

Sub-sequences:

- Consider more realisitic pair of sequences.
- Sequence A is 400residues long and B contains 650 residues.
- If sequence A is in its entirety identical to any portion of sequence B, then A is said to be a sub-sequence of B.
- Gaps simply need to be inserted, as required, to bring A into register with B as shown in figure

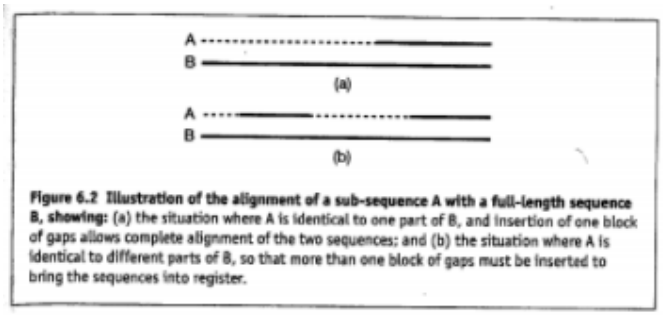


Figure 6.2 Illustration of the alignment of a sub-sequence A with a full-length sequence B, showing: (a) the situation where A is identical to one part of B, and insertion of one block of gaps allows complete alignment of the two sequences; and (b) the situation where A is identical to different parts of B, so that more than one block of gaps must be inserted to bring the sequences into register.

- Now consider that sequence A has two extended regions that show identity to sequence B.
- We would need to identify these regions and then insert gaps into A to bring them into alignment with B
- This algorithm could stop at that point, having found the highest scoring scoring sub-sequences between A and B.

Identity and similarity:

- If sequence comparison is depended only on finding regions of strict identity between two sequences, this method can be implemented into reasonable program.
- Generally alignment is not restricted to sub-sequence matching, but involves comparison of full-length sequences.
- A comprehensive alignment must account fully for the positions of all the residues in both sequences.
- This means that many residues may have to be placed at positions that are not strictly identical.

- In this case, the positioning of gaps in the alignment becomes more complex to compute, this can be done simply by just maximising the number of identical matches by inserting gaps in an unrestricted area.
- Although achieving the optimum score, the result of such a process would be biologically meaningless.
- Instead, scoring penalities are introduced to minimize the gap, that are begun and extension penalities are then incurred when a gap has to be extended.
- In calculating the score for an alignment, we have only considered residue identities.
- Using a unitary matrix, i.e one that weights identical residue matches with a score of 1. Such a matrix is called sparse matrix most of the elements are zero.
- All identical matches carry equal weighting

Table 6.1 Unitary scoring matrices: (a) DNA and (b) protein – the amino acids are grouped according to their physicochemical properties.

(a)

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

(b)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	B	Z	X
C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

The Dayhoff Mutation Data Matrix:

- The most widely used scheme for scoring amino acid pairs is that developed by Dayhoff and co-workers.
- The system arose out of a general model for the evolution of proteins. Dayhoff and co workers examined alignments of closely similar sequences where the the likelihood of a particular mutation (e. A-D) being the result of a set of successive mutations (eg. A-x-y-D) was low.

- A complete picture of the mutation process including those amino acids which did not change was determined by calculating the average ratio of the number of changes a particular amino acid type underwent to the total number of amino acids of that type present in the database.
- This was combined with the point mutation data to give the mutation probability matrix ( )where each element gives the probability of the amino acid in column mutating to the amino acid in row after a particular evolutionary time, for example after 2 PAM (Percentage of Acceptable point Mutations per years).
- The 1978 family of Dayhoff matrices was derived from a comparatively small set of sequences.
- Many of the 190 possible substitutions were not observed at all and so suitable weights were determined indirectly.
- Recently, Jones *et al.* have derived an updated substitution matrix by examining 2,621 families of of sequences in the SWISSPROT database release 15.0.
- The principal differences between the Jones *et al.* matrix (PET91) and the Dayhoff matrix are for substitutions that were poorly represented in the 1978 study. However, the overall character of the matrices is similar.
- Both reflect substitutions that conserve size and hydrophobicity, which are the principle properties of the amino aids

The BLOSUM matrix:

- In bioinformatics, the **BLOSUM** (**BLO**cks **SU**bstitution **M**atrix) matrix is a substitution matrix used for sequence alignment of proteins.
- BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments.
- BLOSUM matrices were first introduced in a paper by Steven Henikoff and Jorja Henikoff. They scanned the BLOCKS database for very conserved regions of protein families (that do not have gaps in the sequence alignment) and then counted the relative frequencies of amino acids and their substitution probabilities.
- Then, they calculated a log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids.
- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins like the PAM Matrices.
- **BLOSUM**: Blocks Substitution Matrix, a substitution matrix used for sequence alignment of proteins.
- Scoring metrics (statistical versus biological): When evaluating a sequence alignment, one would like to know how meaningful it is.
- This requires a scoring matrix, or a table of values that describes the probability of a biologically meaningful amino-acid or nucleotide residue-pair occurring in an alignment.
- Scores for each position are obtained frequencies of substitutions in blocks of local alignments of protein sequences.
- Several sets of BLOSUM matrices exist using different alignment databases, named with numbers.
- BLOSUM matrices with high numbers are designed for comparing closely related sequences, while those with low numbers are designed for comparing distant related sequences.

- For example, BLOSUM80 is used for less divergent alignments, and BLOSUM45 is used for more divergent alignments.
- The matrices were created by merging (clustering) all sequences that were more similar than a given percentage into one single sequence and then comparing those sequences (that were all more divergent than the given percentage value) only; thus reducing the contribution of closely related sequences.
- The percentage used was appended to the name, giving BLOSUM80 for example where sequences that were more than 80% identical were clustered.
- BLOSUM r: the matrix built from blocks with more than r% of similarity – E.g., BLOSUM62 is the matrix built using sequences with more than 62% similarity. – Note: BLOSUM 62 is the default matrix for protein BLAST.
- Experimentation has shown that the BLOSUM-62 matrix is among the best for detecting most weak protein similarities.

The statistical measure of alignment significance

- When performing sequence alignment computationally, a match between the two sequences are created according to mathematical model.
- The model describes the concept of alignment of two sequence strings and fine detail (gap penalities,impact of sequence length differences, effect of alphabet complexity and so on.) is dealt with use of parameters.
- Appropriate choice of parameters is used for minimizing the number of gaps.
- A program produces an alignment between the sequences should not be takenas proof in itself that any relationship exists between them.
- Any standard program will produce some statistical value indicating the level of confidence that should be attached to the alignment
- The statistics quoted in BLAST for pair-wise comparisons are probability (p) or expected frequency (E) values.
- The p-value relates the the score returned for an alignment to the likelihood of its having arisen by chance:
  - Closer the value approaches to zero, the greater the confidence that the match is real.
  - The nearer the value to unity, the greater the chance of the match to be fake.
- The E-value expects the number of hits one can expect to see by chance when searching a database of a particular size.

The DotPlot:

- The most basic method of comparing two sequences is a visual approach known as dotplot.
- One of the simplest methods for evaluating similarity between two sequences is to visualize regions of similarity using dotplot.
- To construct a simple dotplot, the first sequence to be compared is assigned to the horizontal axis of a plot space and the second is then assigned to the vertical axis.
- Kdots are then replaced in the plot space at easch position where both of the sequence elements are identical
- Adjacent regions of identity between the two sequences give rise to diagonal line of dot in the plot.
- Such plots quickly become overly complex and crowded when large, similar sequences are compared.

- The figure illustrated this method for a window size of 10 and also involves a similarity cutoof of 8.
- First,nucleotides 1-10 of the X-axis sequence are compared with nucleotides 1-10 of the sequence of Y-axis.
- If 8 or more of the 10 nucleotides(nt) in the first comparison are identical, a dot is placed in the position(1,1) of the plot space.
- Next window is advanced one nucleotide on the X-axis, so that nucleotides 2-11 of theX-axis sequence are now compared with 1-10 of the sequence of y-axis.
- This procedure is repeated until each 10nt subsequences of the x-axis has been compared to nts 1-10 of the y-axis.
- The y-axis window is advanced by one nucleotide, and the process repeats until all 10nt subsequences of both sequences have been compared.
- Window sizes and cutoff scores can both be varied easily depending on the similarity of the two sequences being compared.
- The ultimate objective is typically to choose criteria that draw attention to regions of significant similarity without allowing noise levels to be distracting.
- The trail and error approcach is often best when first analyzing new data sets.

Local and global similarity:

- Two general models view alignments in rather different ways:
  - The first considers similarity across the full extent of the sequences(global)
  - The second focuses on regions of similarity in parts of the sequences only(local)
- It is very important to understand these distinctions,to appreciate that sequences are uniformly similar, therefore there is no value for performing a global alignment on sequences that have only local similarity.
- Some of the publicly available implementations of pair-wise comparison programs(BLAST and FastA) are fast because they look for local alignment and are made to run more faster by incorporating heuristics.

Pair wise database searching.

FASTA

- FASTA is a DNA and protein sequence alignment software package first described (as FASTP) by David J. Lipman and William R. Pearson in 1985.
- Its legacy is the FASTA format which is now ubiquitous in bioinformatics.
- The original FASTP program was designed for protein sequence similarity searching.
- Because of the exponentially expanding genetic information and the limited speed and memory of computers in the 1980s heuristic methods were introduced aligning a query sequence to entire data-bases.
- FASTA (developed in 1988) added the ability to do DNA:DNA searches, translated protein:DNA searches, and also provided a more sophisticated shuffling program for evaluating statistical significance.
- There are several programs in this package that allow the alignment of protein sequences and DNA sequences.

- Nowadays, increased computer performance makes it possible to perform searches for local alignment detection in a database using the Smith-Waterman algorithm.
- FASTA takes a given nucleotide or amino acid sequence and searches a corresponding sequence database by using local sequence alignment to find matches of similar database sequences.
- The FASTA program follows a largely heuristic method which contributes to the high speed of its execution.
- It initially observes the pattern of word hits, word-to-word matches of a given length, and marks potential matches before performing a more time-consuming optimized search using a Smith-Waterman type of algorithm.

Uses:

- FASTA is pronounced "fast A", and stands for "FAST-All", because it works with any alphabet, an extension of "FAST-P" (protein) and "FAST-N" (nucleotide) alignment.
- The current FASTA package contains programs for protein:protein, DNA:DNA, protein:translated DNA (with frameshifts), and ordered or unordered peptide searches.
- Recent versions of the FASTA package include special translated search algorithms that correctly handle frameshift errors (which six-frame-translated searches do not handle very well) when comparing nucleotide to protein sequence data.
- In addition to rapid heuristic search methods, the FASTA package provides SSEARCH, an implementation of the optimal Smith-Waterman algorithm.
- A major focus of the package is the calculation of accurate similarity statistics, so that biologists can judge whether an alignment is likely to have occurred by chance, or whether it can be used to infer homology.
- The FASTA package is available from fasta.bioch.virginia.edu.
- The web-interface to submit sequences for running a search of the European Bioinformatics Institute (EBI)'s online databases is also available using the FASTA programs.
- The FASTA file format used as input for this software is now largely used by other sequence database search tools (such as BLAST) and sequence alignment programs (Clustal, T-Coffee, etc.).

➢ The FASTA programs find regions of local or global similarity between Protein or DNA sequences, either by searching Protein or DNA databases, or by identifying local duplications within a sequence.

➢ Other programs provide information on the statistical significance of an alignment. Like BLAST, FASTA can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Protein

- Protein–protein FASTA.
- Protein–protein Smith–Waterman (ssearch).
- Global protein–protein (Needleman–Wunsch) (ggsearch)
- Global/local protein–protein (glsearch)

- Protein–protein with unordered peptides (fasts)
- Protein–protein with mixed peptide sequences (fastf)

Nucleotide

- Nucleotide–nucleotide (DNA/RNA fasta)
- Ordered nucleotides vs nucleotide (fastm)
- Unordered nucleotides vs nucleotide (fasts)

Translated

- Translated DNA (with frameshifts, e.g. ESTs) vs proteins (fastx/fasty)
- Protein vs translated DNA (with frameshifts) (tfastx/tfasty)
- Peptides vs translated DNA (tfasts)

Statistical significance

- Protein vs protein shuffle (prss)
- DNA vs DNA shuffle (prss)
- Translated DNA vs protein shuffle (prfx)

Local duplications

- Local protein alignments (lalign)
- Plot protein alignment "dot-plot" (plalign)
- Local DNA alignments (lalign)
- Plot DNA alignment "dot-plot" (plalign)

Basic Local Alignment Search Tool

- In bioinformatics, BLAST for Basic Local Alignment Search Tool is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences.
- A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.
- Different types of BLASTs are available according to the query sequences.

    1. BLASTP which searches a protein database.
    2. BLASTN to search a nucleotide database.
    3. TBLASTN which searches for a protein sequence in a nucleotide database by translating nucleotide sequences in all 6 reading frames.
    4. BLASTX which can search for a nucleotide sequence against a protein database by translating the query via all 6 reading frames.
    5. Gapped-BLAST
    6. psi-BLAST

- BLAST locates patches of regional similarity instead of calculating the best overall alignment using gaps.

- The program then uses a scoring matrix to rank these matches as positive, negative or zero. If the initial match is scored highly, the search is expanded in both directions until the ranking score falls off.
- BLAST is one of the most widely used bioinformatics programs for sequence searching.
- It addresses a fundamental problem in bioinformatics research.
- The heuristic algorithm it uses is much faster than other approaches, such as calculating an optimal alignment.
- This emphasis on speed is vital to making the algorithm practical on the huge genome databases currently available, although subsequent algorithms can be even faster.

USES OF BLAST:

- BLAST can be used for several purposes. These include identifying species, locating domains, establishing phylogeny, DNA mapping, and comparison.
- **Identifying species**:With the use of BLAST, you can possibly correctly identify a species or find homologous species. This can be useful, for example, when you are working with a DNA sequence from an unknown species.
- **Locating domains:**When working with a protein sequence you can input it into BLAST, to locate known domains within the sequence of interest.
- **Establishing phylogeny:**Using the results received through BLAST you can create a phylogenetic tree using the BLAST web-page. Phylogenies based on BLAST alone are less reliable than other purpose-built computational phylogenetic methods, so should only be relied upon for "first pass" phylogenetic analyses.
- **DNA mapping**:When working with a known species, and looking to sequence a gene at an unknown location, BLAST can compare the chromosomal position of the sequence of interest, to relevant sequences in the database(s).
- **Comparison:**When working with genes, BLAST can locate common genes in two related species, and can be used to map annotations from one organism to another.

Comparing BLAST and Smith-Waterman Process:

- While both Smith-Waterman and BLAST are used to find homologous sequences by searching and comparing a query sequence with those in the databases, they do have their differences.
- Due to the fact that BLAST is based on a heuristic algorithm, the results received through BLAST, in terms of the hits found, may not be the best possible results, as it will not provide you with all the hits within the database.
- BLAST misses hard to find matches.
- A better alternative in order to find the best possible results would be to use the Smith-Waterman algorithm.
- This method varies from the BLAST method in two areas, accuracy and speed. The Smith-Waterman option provides better accuracy, in that it finds matches that BLAST cannot, because it does not miss any information.
- Therefore, it is necessary for remote homology. However, when compared to BLAST, it is more time consuming, not to mention that it requires large amounts of computer usage and space.
- However, technologies to speed up the Smith-Waterman process have been found to improve the time necessary to perform a search dramatically. These technologies include FPGA chips and SIMD technology.

- In order to receive better results from BLAST, the settings can be changed from their default settings.
- However, there is no given or set way of changing these settings in order to receive the best results for a given sequence.
- The settings available for change are E-Value, gap costs, filters, word size, and substitution matrix.
- Note, that the algorithm used for BLAST was developed from the algorithm used for Smith-Waterman.
- BLAST employs an alignment which finds "local alignments between sequences by finding short matches and from these initial matches (local) alignments are created".

UNIT-V
Assignment-Cum-Tutorial Questions
SECTION-A

Objective Questions

1) Database interrogation can take the form of _____ or _____.

2) In bioinformatics, a _____ is a graphical method that allows the comparison of two biological sequences and identify regions of close similarity between them.

3) Identity matrices are _____.

4) To generate gapped alignments, _____ algorithm can be used.

5) The full form of BLAST is_____.

6) _____ is the input sequence format in BLAST.

7) Finding good solution using dynamic programming involves the process of _____ and _____.

8) The procedure of aligning two sequences by searching for patterns that is in the same order in the sequences [ ]

a) sequence alignment b) pair wise alignment
c) multiple sequence alignment d) all of these

9) FASTA format starts with _____ [ ]
a) / b) * c) > d) #

10) The procedure of aligning many sequences simultaneously is called[ ]
a) Multiple sequence alignment c) Pair wise alignment
b) Global alignment d) Local alignment

11) _____ compares protein sequence against protein databases.[ ]
a) Blastp b) blastn c) blastx d) tblastx

12) The complexity of alphabet for DNA is [ ]
a) 4 b) 7 c) 20 d) 25

13) Which of the following is the sequence alignment tool? [ ]
a) Chime b) BLAST c) FASTA d) Clustal W

14) All are sequence alignment tools except [ ]
a) Rasmol b) BLAST c) FASTA d) Clustal W

15) Which is the default scoring matrix used in BLAST? [ ]

a) PAM62 b) BLOSUM 62 c) BLOSUM 60 d) BLOSUM 80

16) In Needleman Wunsch algorithm of pairwise alignment of sequences with lengths n and m, thecomputational time is proportional to: [ ]

a) n x m b) (n+1) x (m+1) c) n + m d) n x (m+1

17) You have two distantly related proteins. Which of the following sets is the best for comparingthem? [ ]

a) BLOSUM45 or PAM250 b) BLOSUM45 or PAM1

c) BLOSUM80 or PAM250 d) BLOSUM80 or PAM1

18) Which alignment is used to predict whether two sequences are homologous or not?

[ ]

a) Local b) Global c) Pair-wise d) Multiple

19) BLASTx is used to [ ]

a)search a nucleotide database using a nucleotide query

b) search protein database using a protein query

c) search protein database using a translated nucleotide query

d) search translated nucleotide database using a protein query

20) One PAM means one accepted point mutation per [ ]

a) 10^2 residues

b) 10 residues

c) 10^3 residues

d) 10^4residues

SECTION-B

SUBJECTIVE QUESTIONS

1) Explain following methods of sequence alignment :

a) The BLOSUM matrices

b) The Dayhoff Mutation Data Matrix

2) Explain similarities and differences between BLAST and FASTA tools for sequence alignment.

3) Explain BLAST algorithm. State the major refirements included in gapped BLAST.

4) What is filtering in BLAST?

5) Briefly explain about the dynamic programming.

6) Illustrate the differences between local and global similarity.

7) Discuss how a sequence alignment might be evaluated statistically, illustrating your answer with an example.

8) Discuss how to find matches in a genome sequence efficiently.

9) Illustrate the most basic method of comparing two sequences.

10) When should one use either a global or local sequence alignment?

11) What does sequence comparison measure? Similarity versus homology

12) Justify the statement “"What sounds simple in principle isn't at all simple in practice. Choosing a good alignment by eye is possible, but life is too short to do it more than once or twice”.

13) Why use BLAST?

14) What are the major extensions of BLAST? Discuss the areas of applications of these programs.

SECTION-C

QUESTIONS AT THE LEVEL OF GATE

1) You do protein BLAST searches of the SWISS-PROT and the non-redundant databank using the same sequence as query and you get the same top hits however the E-values for the tophit are different. Why could this happen

UNIT –VI

Objective:

- To understand basic biological databases, algorithms for proteomics and genomics analysis.

Syllabus:

Database Searching and Analysis Packages

Secondary database searching-Importance and need of secondary database searches, Secondary database structure and building a sequence search protocol, Analysis packages- analysis package structure, Commercial databases, Commercial software

Learning Outcomes:

The student will be able to

-

Learning Material

Database Searching and Analysis Packages

Secondary database searching:

- Different secondary databases have evolved as a result of the different analysis methods used in the derivation of family signatures.

Regular expressions:

- the simplest approach to pattern recognition is to characterize a family by means of a single conserved motif and to reduce the sequence data within the motif to a regular expression pattern.
- The expression derived from the motif shown in the table indicated that position 2,4,8,12 and 15 are completely conserved; positions 1,11,13 and 14 allow one of two possible residues;position 3 allows one of the three possible residues; positions 5 to 8 can be anything except proline or glycine.

Alignment	Regular expression
ADLGAVFALCDRYPQ SDVGPRSCPCERFYQ ADLGRTQNRCDRYYQ ADIGQPHSLCERYPQ	[AS]-D-[IVL]-G-x4-{PG}-C-[DE]-R-[FY]2-Q

- In order to reduce the likelihood of the pattern making too many incorrect matches, the software that makes use of regular expressions often does not tolerate similarity and searches are thus limited in scope to the retrieval of identical matches.
- In a sequence, inspite of sequence being 99% identical to the expression,will be nevertheless be rejected as a mismatch. Eventhoough the mismatch is a conservative,biological feasible replacement.

- Alternatively, a sequence matching all positions of the patterns,but with an additional residue inserted in the non-conserved region following the glycine, will again fail to match, because the expression does not cater for sequences with more than 4 linking residues at this point.
- Searching a database in this way thus results in either an exact match or match at all.
- creating a regular expression that performs well in database searches is always a compromise between the tolerance that can be built into it, and the amount of noise it will match; fuzzier the pattern; noiser its results; but greater the hope of finding the results.

Finger prints:

- Within a sequence alignment, it is usual to find not one, but several motifs that characteristic the aligned family.
- it makes sense to use many or all, of the conserved regions to create a signature or fingerprint, so that, in a database search, there is a higher chance of identifying a distant relative, whether or not all the parts of signature are matched.
- Groups of motifs are excised from alignments and the sequence information they contain is converted into matrices populated by other frequencies observed at eacg position
- This type of scoring system is said to be unweighted in the sense that no additional scores.
- Patterns have very limited diagnostic power because they require target sequences to match the expression exactly.
- Thus, as we have seen, a sequence that fits a pattern at all but one position will be rejected as a mismatch.
- To address this problem, it is possible to build more powerful discriminators. Within a sequence alignment, it is usual to find not one, but several motifs that characterize the aligned family.
- Diagnostically, it makes sense to use all the conserved regions to build a family signature, so that in a database search there is a higher chance of identifying a distant relative, whether or not all parts of the signature are matched.

Blocks:

- the constituent motifs of a fingerprint are unweighted (i.e. the scoring system uses only observed residue frequencies), which sometimes leads to relatively poor diagnostic performance.
- However, it is possible to build more powerful representations of motifs by applying different weighting schemes.
- In one approach, sequence segments within a motif are clustered to reduce multiple contributions to residue frequencies from groups of closely related sequences.

- Each cluster is then treated as a single segment, each of which is assigned a score that gives a measure of its relatedness; the most distant segment is given a weight of 100.
- Most protein families will be defined by several such clustered, weighted motifs. For a given query sequence, as with a fingerprint, the more motifs matched, the greater the confidence that the sequence belongs to that family.
- Aligned, ungapped, weighted segments of this type are most frequently termed blocks (synonymous terms are motif, segment, and feature).

Profiles:

- By contrast with the use of patterns, blocks, and fingerprints, an alternative approach is to distil the sequence information within complete alignments into scoring tables.
- Such tables define which residues are allowed at given positions, which positions are highly conserved and which degenerate, and which positions, or regions, can tolerate insertions.
- The scoring system is intricate, and may include evolutionary weights and results from structural studies, as well as data implicit in the alignment.
- In addition, variable penalties are specified to weight against insertions and deletions occurring in core secondary structure elements. Such tables are used to search databases for similar matches, their inherent complexity rendering them very potent discriminators. They are most frequently termed profiles.

Hidden Markov models:

- Hidden Markov models (HMMs) have been extensively used in biological sequence analysis.
- HMMs are well-known for their effectiveness in modeling the correlations between adjacent symbols, domains, or events, and they have been extensively used in various fields, especially in speech recognition and digital communication.
- Considering the remarkable success of HMMs in engineering, it is no surprise that a wide range of problems in biological sequence analysis have also benefited from them.
- For example, HMMs and their variants have been used in gene prediction [2], pairwise and multiple sequence alignment , base-calling , modeling DNA sequencing errors, protein secondary structure prediction , ncRNA identification , RNA structural alignment , acceleration of RNA folding and alignment , fast noncoding RNA annotation, and many others.
- A *hidden Markov model (HMM)* is a statistical model that can be used to describe the evolution of observable events that depend on internal factors, which are not directly observable.
- We call the observed event a `symbol' and the invisible factor underlying the observation a `state'.

- An HMM consists of two stochastic processes, namely, an invisible process of hidden states and a visible process of observable symbols.
- The hidden states form a *Markov chain*, and the probability distribution of the observed symbol depends on the underlying state.
- For this reason, an HMM is also called a doubly-embedded stochastic process

Secondary database structure and building a sequence search protocol

- One of the central goals of the bioinformatics is the prediction of protein function and ultimately of structure, from the linear amino acid sequence
- By searching secondary databases, which house abstractions of functional and structural sites characteristic of particular proteins.
- Similarily by searching fold libraries, which use template of known structures, it is possible to recognize the previously characterized fold.

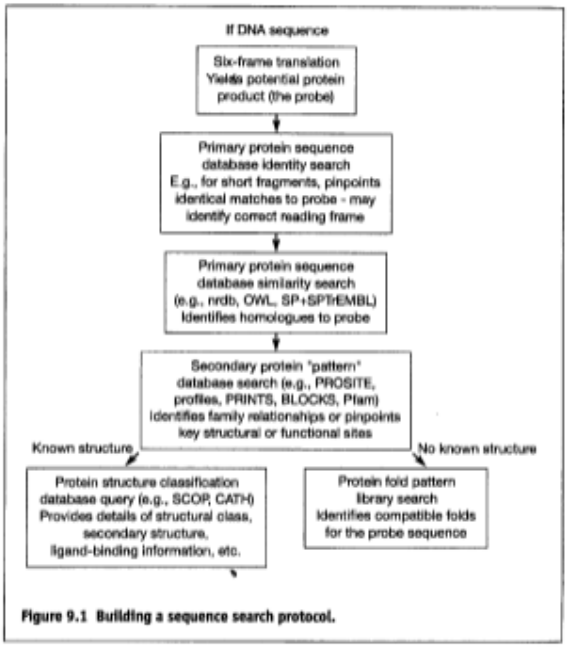


Figure 9.1 Building a sequence search protocol.

- One practical approach presented in the above figure is that, we start by checking for identical matches and then move on to search for closely similar sequences in the primary databases.
- The strategy then involves searching for previously characterized sequence and where possible, foldpatterns in a variety of pattern databases.
- The deciding step is the integration of results from all the searches to build a consistent family/functional/structural diagnosis.

Searching the primary databases:

- If an identity search fails to find a If an identity search fails to find a match, all is not lost all is not lost.
- The next step is to look for similar The next step is to look for similar sequences sequences.
- It is recommended to perform similarity It is recommended to perform similarity searches on peptides that are longer searches on peptides that are longer than ~30 residues than ~30 residues (the shorter the the shorter the peptide, the greater the likelihood of peptide,

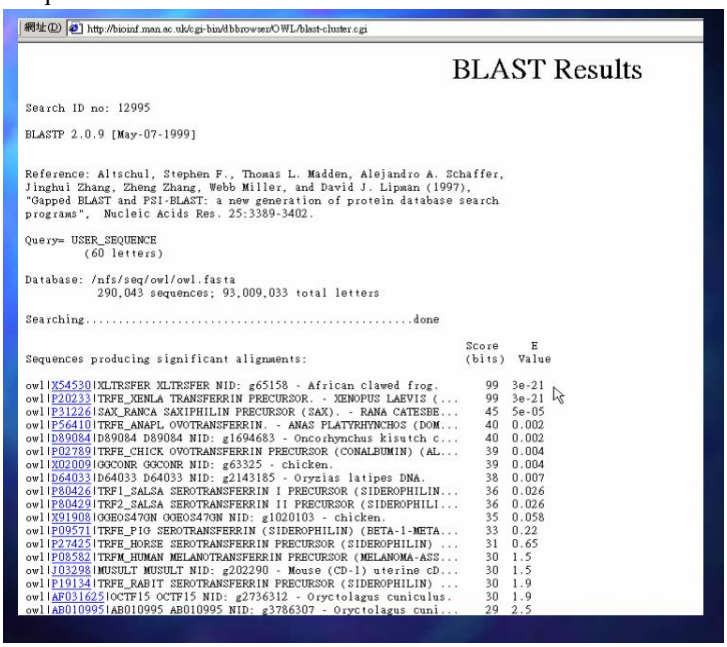
the greater the likelihood of finding chance matches that finding chance matches that have no have no biological relevance biological relevance) „

Similarity searching:

- The most rapid and simple option is use The most rapid and simple option is use BLAST (of FastA) BLAST (of FastA) „
- We can use the sequence retrieved by We can use the sequence retrieved by from the primary database (OWL)

BLAST Output:

- „ There are several important features to note There are several important features to note in BLAST Output.
- In BLAST Output We are looking for matches that have We are looking for matches that have high scores high scores with correspondingly low probability values with correspondingly low probability values.
- Low probability indicates that a match is Low probability indicates that a match is unlikely unlikely to have arisen by chance to have arisen by chance.
- The results show a cluster of high scores (with low The results show a cluster of high scores (with low probabilities) at the top of the list, indicating a probabilities) at the top of the list, indicating a likely relationship between the query and the likely relationship between the query and the family of sequences in the cluster. family of sequences in the cluster



Searching the secondary databases:

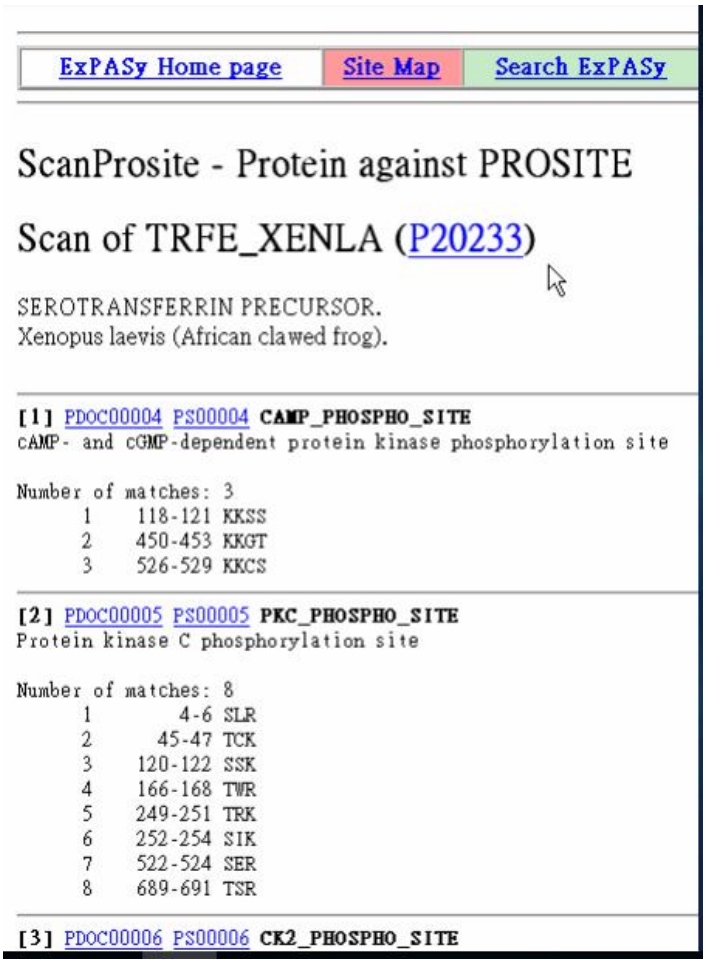
- Although a family of sequences was Although a family of sequences was identified by the BLAST search, we can identified by the BLAST search, we can continue on and search the secondary continue on and search the secondary database in order to discover if our database in order to discover if our query sequence contains any known query sequence contains any known characteristic characteristic conserved motifs conserved motifs. „

- The first secondary database to consider is consider is PROSITE.

Searching PROSITE:

- At this example, the database code (TRFE_XENLA)
- Three regular expression patterns have Three regular expression patterns have been matched been matched Î
  - TRANSFERRIN_1 TRANSFERRIN_1
  - TRANSFERRIN_2 TRANSFERRIN_2
  - TRANSFERRIN_3 TRANSFERRIN_3

  Y-x(0,1)-[VAS]-V-[IVAC]-[IVA]-[IVA]-[RKH]-[RKS]-[GDENSA] Pattern of PS00205



ExPASy Home page | Site Map | Search ExPASy

ScanProsite - Protein against PROSITE

Scan of TRFE_XENLA (P20233)

SEROTRANSFERRIN PRECURSOR.
Xenopus laevis (African clawed frog).

[1] PDOC00004 PS00004 **CAMP_PHOSPHO_SITE**
cAMP- and cGMP-dependent protein kinase phosphorylation site

Number of matches: 3

1	118-121 KKSS
2	450-453 KKGT
3	526-529 KKCS

[2] PDOC00005 PS00005 **PKC_PHOSPHO_SITE**
Protein kinase C phosphorylation site

Number of matches: 8

1	4-6 SLR
2	45-47 TCK
3	120-122 SSK
4	166-168 TWR
5	249-251 TRK
6	252-254 SIK
7	522-524 SER
8	689-691 TSR

[3] PDOC00006 PS00006 **CK2_PHOSPHO_SITE**

Searching Pfam:

- Another important resource to search is Another important resource to search is the Pfam collection of Hidden Markov the Pfam collection of Hidden Markov Models. Models.
- The sequence must be in FastA format.

Searching PRINTS:

- The Output of PRINTS is divided into distinct The Output of PRINTS is divided into distinct sections: sections:

- ❖ The program offers an intelligent The program offers an intelligent ‘guess’ based on based on the occurrence of the highest the occurrence of the highest -scoring complete or scoring complete or partial fingerprint match partial fingerprint match -it then provides an it then provides an expanded table that shows the top 10 best expanded table that shows the top 10 best -scoring scoring matches. matches.
- ❖ If the guess is considered either to be wrong or If the guess is considered either to be wrong or to has missed something, the to has missed something, the remaining sections remaining sections of output provide more of the row data, again of output provide more of the row data, again allowing the user to search for anything that allowing the user to search for anything that might have been missed. might have been missed.

Searching IDENTIFY:

- ❖ The output of IDENTIFY is normally The output of IDENTIFY is normally given at several stringency levels given at several stringency levels

Analysis packages:

- In the course of analysing a protein sequence. Several search methods need to be applied.
- But homology searching is only one aspect of analysis process.
- Other research tools are also available including hydropathy profiles for detection of possible transmembrane domains.
- Sequence alignment and phylogenetic tree tools for charting evolutionary relationships, secondary structure prediction plots for locating α-helices and β-strands
- Because of the need to employ a range of techniques for effective sequence analysis, software packages have been developed to bring a variety of these methods under a single umbrella, remove the need to use different tools with different interfaces, with different input requirements and different output formats.

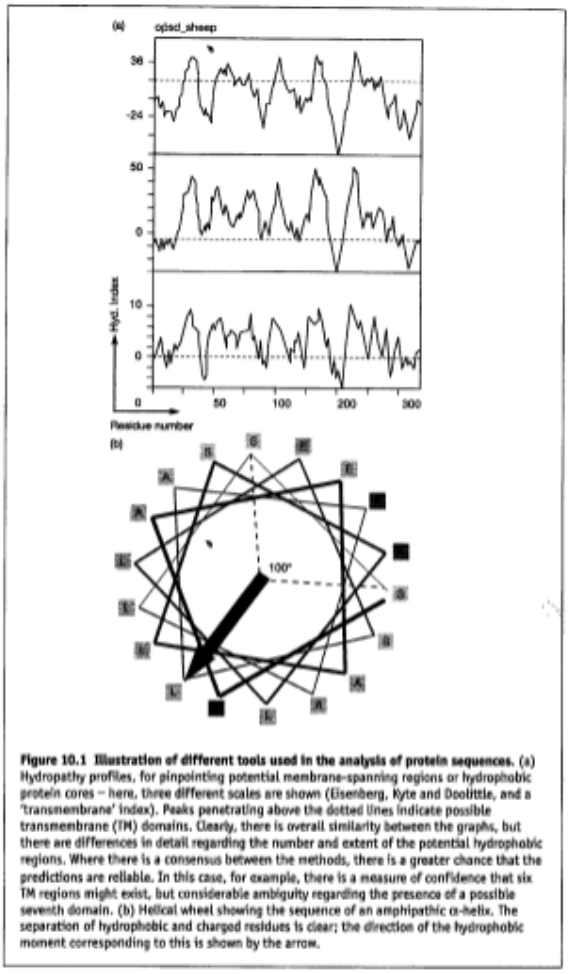


Figure 10.1 Illustration of different tools used in the analysis of protein sequences. (a) Hydropathy profiles, for pinpointing potential membrane-spanning regions or hydrophobic protein cores – here, three different scales are shown (Eisenberg, Kyte and Doolittle, and a 'transmembrane' index). Peaks penetrating above the dotted lines indicate possible transmembrane (TM) domains. Clearly, there is overall similarity between the graphs, but there are differences in detail regarding the number and extent of the potential hydrophobic regions. Where there is a consensus between the methods, there is a greater chance that the predictions are reliable. In this case, for example, there is a measure of confidence that six TM regions might exist, but considerable ambiguity regarding the presence of a possible seventh domain. (b) Helical wheel showing the sequence of an amphipathic α-helix. The separation of hydrophobic and charged residues is clear; the direction of the hydrophobic moment corresponding to this is shown by the arrow.

Commercial databases:

- If the majority of the biological databases are available from publicly accessible resources, why we require commercial databases?
- The answer lies in the industrial approach to information technology i.e the desire to purchase solutions to well-defined problems, rather than the more explanatory academic approach.
- If services exist to develop and maintain databases, can be purchase, finance and man power can be released for the more exciting scientific task.
- Major releases of DNA and protein takes place for every 3 to 4 months.
- In mean time, newly determined sequences are updated daily.
- To keep in-house database up-to-date, synchronized FTP scripts are used.
- If new database evolve, and it is considered advantageous also to bring them in-house, existing scripts must be updated to incorporate the new resources.

- One answer to these problems is to find out a good database service provider, who will either supply up-to-date databases or alternatively offer easily maintainable software for database updating.

Commercial software:

- In an industrial environment, the requirement for software licenses is often simply a matter of legal probity-although commercial releases from academic institutions are often identical to their freely available academic counterparts.
- It is reasonable that software used for commercial purposes should attract fee.
- Commercial organizations usually require some level of support to be assured
- Another significant concern lies with company use of the internet for database searching; the act of performing a database search with a propertiary sequence on a public server is equivalent to publishcation.

UNIT-VI
Assignment-Cum-Tutorial Questions
SECTION-A

Objective Questions

1) _____ removes the burden of learning how to use different tools, with different interfaces.
2) _____ is an integrated set of tools for sequence analysis being developed at the Sanger center.
3) _____ is the composite resource that can be queried directly by means of its query language.
4) BLAST tool is _____ faster than dynamic programming.
5) The database splits the sequence into two domains, which are assigned CATH numbers _____ and _____.
6) The simplest pattern recognition approach characterises families using _____.
7) The first secondary database developed by [ ]
a) PRINTS b) PROSITE c) PDB d) PIR
8) Which tool can be used for the identification of motifs? [ ]
a) COPIA b) Patternhunter c) PROSPECT d) BLAST
9) Which algorithm is used by global alignment [ ]
a) Needleman and Wunsch c) Smith-Waterman
b) BLAST d) PAM
10) In pairwise alignment result, sequences reported as similar due to chance represents _____ result [ ]
a) True positive b) True negative c) False positive d) False negative

11) Sequence alignment helps scientists [ ]
a) To trace out evolutionary relationships
b) To infer the functions of newly synthesised genes
c) To predict new members of gene families
d) All of these
12) Which of the following is the sequence alignment tool [ ]
a) BLAST b) PRINT c) PDB d) PIR
13) Which is data retrieving tool? [ ]
a) ENTREZ b) EMBL c) PHD d) All of these
14) Which of the following is a multiple sequence alignment tool? [ ]
a) Clustal W b) Chime c) Dismol d) PDB
15) Phylogenetic relationship can be shown by [ ]
a) Dendrogram b) Gene bank c) Data retrieving tool d) Nucleic acid sequence analysis tool

16) PRINTS are the software used for [ ]
a) Detection of genes from genome sequence d) Detection of tRNA
b) Prediction of function of a new gene
c) Identification of functional domains/motifs of proteins
17) The process of finding the relative location of genes on a chromosome is called [ ]
a) Gene tracing c) Genome trapping
b) Genome walking d) Chromosome walking
18) BLOSUM matrices are used for [ ]
a) Multiple sequence alignment c) Pairwise sequence alignment
b) Phylogentic analysis d) All of the above
19) GeneBank and SWISS-PROT are the examples of [ ]
a) Primary database c) Secondary database
b) Composite database d) None of the above
20) When you are comparing two or more than two sequences of same or different organisms,what is the type of alignment [ ]
a) Global b) Pairwise sequence c) Local d) Multiple sequence

SECTION-B

SUBJECTIVE QUESTIONS

1) Explain following methods of secondary database searching :
a)Fingerprints b)Blocks C) Profiles
2) Briefly explain the BLAST output.
3) Give an outline on Hidden Markov Models.
4) Briefly explain analysis packages.
5) Explain the following
a) Searching PROSITE c)Searching pfam
b) Searching PRINTS
6) Give a brief note on analysis packages.
7) Illustrate the different tools used in analysis of protein sequences.
8) If the majority of the biological databases are available from publicly accessible resources, why we require commercial databases?
9) Compare and contrast commercial databases and commercial software.
10) Why we require analysis packages?
11) Justify the statement “homology searching is only one aspect of analysis process.”
12) Compare and contrast Searching PROSITE, Searching pfam, Searching PRINTS.
13) Why HMM is called as doubly-embedded stochastic process ?
14) Why different analysis methods used in the derivation of family signatures?