First Internal Examination (2019-20)
B.Sc. (BT) III Year
Fundamentals of Bioinformatics & Nanotechnology

SET-A

Time: 1:30 Hours **Max. Marks-30**

---

1. **Short Answers Questions (1 Marks Each; 1x8= 8 Marks)**

i. What is Bioinformatics?

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

ii. When was HGP (Human genome project) started?

ANS… The Human Genome Project (HGP) was an international scientific research project with the goal of determining the sequence of nucleotide base pairs that make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and a functional standpoint.
Start date: 1990
Completed: April 2003
Period: 1990 – 2003

iii. What is medical-informatics?

ANS… Medical informatics is the intersection of information science, computer science, and health care. This field deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine.

iv. What is the main role of a bioinformatician in present biological research and development area?

ANS… Biological research is the most fundamental research to understand complete mechanism of living system. In recent decades, bioinformatics methods, with the use of knowledge accumulated in public databases such as NCBI, Pubmed and other databases, make it possible to systematically dissect large gene lists in an attempt to assemble a summary of the most enriched result in current biology, which is able to provide a huge

contribution to biological research. In the same contrast, bioinformatics scanning approaches are emerging as alternative technologies that allow investigators to simultaneously measure the changes and regulation of genome-wide genes under certain biological conditions. These high-throughput technologies usually generate large gene of biological research interest lists as their final outputs.

v. What is data mining?

ANS… Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

vi. Define term database?

ANS… A database is a collection of information that is organized so that it can be easily accessed, managed and updated. ... Databases process workloads to create and update themselves, querying the data they contain and running applications against it.

vii. What do you understand by the term-conserved sequence?

ANS… In evolutionary biology, conserved sequences are similar or identical sequences in nucleic acids or proteins across species or within a genome. Conservation indicates that a sequence has been maintained by natural selection.

viii. Define search engine?

ANS… Search engine is a service that allows Internet users to search for content via the World Wide Web (WWW). A user enters keywords or key phrases into a search engine and receives a list of Web content results in the form of websites, images, videos or other online data.

2. Medium Answer Questions (4 Marks each 2x4= 8 Marks)

i. Which are the main sub-disciplines of bioinformatics?

ANS… 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. Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. There are three important sub-disciplines within bioinformatics:

- the development of new algorithms and statistics with which to assess relationships among members of large data sets.
- the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures

- the development and implementation of tools that enable efficient access and management of different types of information.

ii. What is SWISS-PROT?

ANS… SWISS-PROT (1) is an annotated protein sequence database established in 1986 and maintained collaboratively, since 1987, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library (now the EMBL Outstation-The European Bioinformatics Institute; 2). The SWISS-PROT protein sequence data bank consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardization purposes the format of SWISS-PROT (3) follows as closely as possible that of the EMBL nucleotide sequence database. The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria.

Annotation

In SWISS-PROT annotation is mainly found in the comment lines (CC), in the feature table (FT) and in the keyword lines (KW). Most comments are classified by 'topics', an approach which permits easy retrieval of specific categories of data from the database.

Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. In SWISS-PROT as much as possible data is merged, so as to minimize the redundancy of the database. If conflicts exist between various sequencing reports they are indicated in the feature table of the corresponding entry.

Integration with other databases

It is important to provide the users of biomolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures), as well as with specialized data collections. SWISS-PROT is currently cross-referenced with 24 different databases. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT.

3 Long Answer Questions (7 Marks each 2x7= 14 Marks)

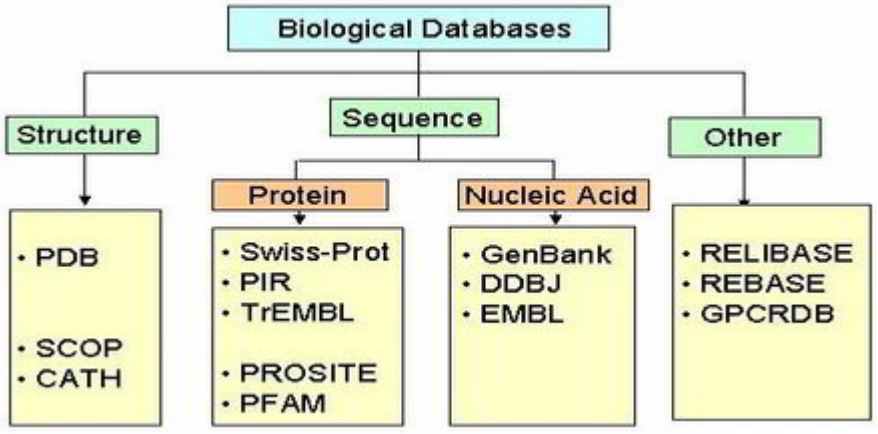
i. What are the types of biological databases, explain?

ANS… Biological databases emerged as a response to the huge data generated by low-cost DNA sequencing technologies. One of the first databases to emerge was GenBank, which is a collection of all available protein and DNA sequences. It is maintained by the National Institutes of Health (NIH) and the National Center for Biotechnology Information (NCBI). GenBank paved the way for the Human Genome Project (HGP). The HGP allowed complete sequencing and reading of the genetic blueprint. The data stored in biological databases is

organized for optimal analysis and consists of two types: raw and curated (or annotated). Biological databases are complex, heterogeneous, dynamic, and yet inconsistent. The inconsistency is due to the lack of standards at the ontological level.

Different types of Biological Databases

Data type	Explanation	Example
Bibliographic DB	Contains article and research papers of different journals	MEDLINE, Pubmed
Genome DB	Contains whole genome sequences of viruses, eukaryotes or prokaryotes	Genome Information Broker (GIB), Entrez genome of NCBI
Sequence DB	Contains protein and nucleotide sequence	DDBJ, EMBL, SWISS prot
Structure DB	Contains 3D structure of proteins and nucleic acids	Nucleotide Database (NDB), Protein Data Bank (PDB)
Metabolic DB	Contains data about various biological pathways	Kyoto Encyclopedia of genes and genomes (KEGG)
Enzyme DB	Contains data about structure, function and pathways of various enzymes	ExPasy, REBASE
Disease DB	Disease related information	OMIM
Chemical DB	Data about biological activity of several small molecules	PubChem
Microarray DB	Mainly Data obtained from microarray experiments	Gene Expression Omnibus (GEO), Human Gene Expression Index (HUGE)



ii. Explain about different types of search engines with examples?

ANS… The purpose of a search engine is to extract requested information from the huge database of resources available on the internet. Search engines become an important day to

day tool for finding the required information without knowing where exactly it is stored. There are different types of search engines to get the information you are looking for.

Search engines are classified into the following three categories based on how it works.

1. Crawler based search engines
2. Human powered directories
3. Hybrid search engines
4. Other special search engines

1. Crawler Based Search Engines

All crawler based search engines use a crawler or bot or spider for crawling and indexing new content to the search database. There are four basic steps, every crawler based search engines follow before displaying any sites in the search results.

- Crawling
- Indexing
- Calculating Relevancy
- Retrieving the Result

1.1. Crawling

Search engines **crawl** the whole web to fetch the web pages available. A piece of software called *crawler* or *bot* or *spider,* performs the crawling of the entire web. The crawling frequency depends on the search engine and it may take few days between crawls. This is the reason sometimes you can see your old or deleted page content is showing in the search results. The search results will show the new updated content, once the search engines crawl your site again.

1.2. Indexing

Indexing is next step after crawling which is a process of identifying the words and expressions that best describe the page. The identified words are referred as keywords and the page is assigned to the identified keywords. Sometimes when the crawler does not understand the meaning of your page, your site may rank lower on the search results. Here you need to optimize your pages for search engine crawlers to make sure the content is easily understandable. Once the crawlers pickup correct keywords your page will be assigned to those keywords and rank high on search results.

1.3. Calculating Relevancy

Search engine compares the search string in the search request with the indexed pages from the database. Since it is likely that more than one page contains the search string, search engine starts **calculating the relevancy** of each of the pages in its index with the search string.

There are various algorithms to calculate relevancy. Each of these algorithms has different relative weights for common factors like keyword density, links, or meta tags. That is why different search engines give different search results pages for the same search string. It is a known fact that all major search engines periodically change their algorithms. If you want to keep your site at the top, you also need to adapt your pages to the latest changes. This is one reason to devote permanent efforts to SEO, if you like to be at the top.

1.4. Retrieving Results

The last step in search engines' activity is **retrieving** the results. Basically, it is simply displaying them in the browser in an order. Search engines sort the endless pages of search results in the order of most relevant to the least relevant sites.

Examples of Crawler Based Search Engines

Most of the popular search engines are crawler based search engines and use the above technology to display search results. Example of crawler based search engines:

- Google
- Bing
- Yahoo!
- Baidu
- Yandex

Besides these popular search engines there are many other crawler based search engines available like DuckDuckGo, AOL and Ask.

2. Human Powered Directories

Human powered directories also referred as open directory system depends on human based activities for listings. Below is how the indexing in human powered directories work:

- Site owner submits a short description of the site to the directory along with category it is to be listed.
- Submitted site is then manually reviewed and added in the appropriate category or rejected for listing.
- Keywords entered in a search box will be matched with the description of the sites. This means the changes made to the content of a web pages are not taken into consideration as it is only the description that matters.
- A good site with good content is more likely to be reviewed for free compared to a site with poor content.

Yahoo! Directory and DMOZ were perfect examples of human powered directories. Unfortunately, automated search engines like Google, wiped out all those human powered directory style search engines out of the web.

3. Hybrid Search Engines

Hybrid Search Engines use both crawler based and manual indexing for listing the sites in search results. Most of the crawler based search engines like Google basically uses crawlers as a primary mechanism and human powered directories as secondary mechanism. For example, Google may take the description of a webpage from human powered directories and show in the search results. As human powered directories are disappearing, hybrid types are becoming more and more crawler based search engines.

But still there are manual filtering of search result happens to remove the copied and spammy sites. When a site is being identified for spammy activities, the website owner needs to take corrective action and resubmit the site to search engines. The experts do manual review of the submitted site before including it again in the search results. In this manner though the crawlers control the processes, the control is manual to monitor and show the search results naturally.

4. Other Types of Search Engines

Besides the above three major types, search engines can be classified into many other categories depending upon the usage. Below are some of the examples:

- Search engines have different types of bots for exclusively displaying images, videos, news, products and local listings. For example, Google News page can be used to search only news from different newspapers.
- Some of the search engines like Dogpile collects meta information of the pages from other search engines and directories to display in the search results. This type of search engines are called metasearch engines.
- Semantic search engines like Swoogle provide accurate search results on specific area by understanding the contextual meaning of the search queries.

First Internal Examination (2019-20)
B.Sc. (BT) III Year
Fundamentals of Bioinformatics & Nanotechnology

SET-B

Time: 1:30 Hours **Max. Marks-30**

---

1. **Short Answers Questions (1 Marks Each; 1x8= 8 Marks)**

i. What is consensus sequence?

ANS… In molecular biology and bioinformatics, the *consensus sequence* (or canonical *sequence*) is the calculated order of most frequent residues, either nucleotide or amino acid, found at each position in a *sequence* alignment.

ii. What is BLAST?

ANS… 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 and/or RNA sequences.

iii. What are biological databases?

ANS… Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis.

iv. The term bioinformatics was coined by?

ANS… The **term bioinformatics** did not mean what it means today. Paulien Hogeweg and Ben Hesper **coined** it in 1970 to refer to the study of information processes in biotic systems.

v. The first published completed gene sequence of was which organism?

ANS… The first organism to have its entire genome sequenced was *Haemophilus* influenzae in 1995.

vi. NCBI was established in which year?

ANS… 4 November 1988

The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper.

vii. Submission to GenBank is made using what tools?

ANS… BankIt is the submission tool of GenBank for sequence submitting to NCBI.

viii. What is DDBJ?

ANS… 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.

2. Medium Answer Questions (4 Marks each 2x4= 8 Marks)

i. How bioinformatics is a multidisciplinary branch?

ANS… Bioinformatics is a multidisciplinary field that applies knowledge of computer science, mathematics and physics to solve problems from biological research fields such as molecular biology and biochemistry.

Coined by Paulien Hogeweg and Ben Hesper in 1978, the term "bioinformatics" was originally used in studies of ecosystems biology and ecology. However, with the development and improvement of technologies for analyzing DNA, RNA and proteins, a greater amount of data began to be generated, requiring the use of computational tools to organize, analyze and understand the growing volume of data.

Through bioinformatics, several methods have been developed to better understand the molecular processes that occur in living organisms. Today, researchers have databases to store a volume of information never imagined before and algorithms to handle large-scale data produced by techniques such as next-generation sequencing and mass spectrometry. Moreover, bioinformatics uses statistical and computational techniques to help undertastand the large volume of data and provide greater explanatory power of biological processes .

Applications of Bioinformatics go through various areas of knowledge, such as genome assembly, comparative genomics, gene expression analysis, gene regulation networks, the study of metabolism, analysis of the structure of macromolecules, drug design and evolutionary biology. Moreover, it is closely related to two new fields of research of great interest in our laboratory: the systems biology and synthetic biology.

As a multidisciplinary field, the research groups within Bioinformatics are composed of researchers familiar with the concepts and methods of biology, as well as able to understand the essential statistical methods for data analysis and knowledge of computer tools necessary to address the biological information.

ii. Describe about search directory with 2 examples?

ANS… A **search directory** is an online index of websites organized by category and alphabetically, similar to a phone book's yellow pages. **Search directories** are generally populated through submission and approval; as such they differ from **search** engines, which depend on web crawlers to index results. Unlike a search engine, which uses bots to sort its information and relies soley on Internet technology to operate, a search directory is human-organized, meaning that real people actually edit the links and classify them into the directory. The result of this distinction is easily apparent when doing an online search: instead of a search engine spit-out of hundreds of thousands of results (many of which are irrelevant), a directory comes back with links that are more accurate and better organized. If you don't know the actual URL of a Web site you want to see, try searching Yahoo! first; if you still can't find any interesting, related information, then go to a search engine (and take time to learn how to use the "advanced search" feature).
Historical perspective: Yahoo! is a well-known search directory, created by David Filo and Jerry Yang of the Department of Computer Science at Stanford University, it's been said that "Yahoo" stands for "Yet Another Hierarchical Officious Oracle," but the founders insist they selected the name because they liked the general definition of a yahoo: rude, unsophisticated, uncouth. The word was invented by Jonathan Swift and used in his book Gulliver's Travels.

3. Long Answer Questions (7 Marks each 2x7= 14 Marks)

i. Comments on following:

a. EMBL 4

ANS… The **European Molecular Biology Laboratory** (**EMBL**) is a molecular biology research institution supported by 25 member states, four prospect and two associate member states. EMBL was created in 1974 and is an intergovernmental organisation funded by public research money from its member states. Research at EMBL is conducted by approximately 85 independent groups covering the spectrum of molecular biology. The Laboratory operates from five sites: the main laboratory in Heidelberg, and outstations in Hinxton (the European Bioinformatics Institute (EBI), in England), Grenoble (France), Hamburg (Germany), Monterotondo (near Rome) and Barcelona (Spain). EMBL groups and laboratories perform basic research in molecular biology and molecular medicine as well as training for scientists, students and visitors. The organization aids in the development of services, new instruments and methods, and technology in its member states.

b. OWL database 3

ANS… OWL is a non-redundant composite of 4 publicly-available primary sources: SWISS-PROT, PIR (1-3), GenBank (translation) and NRL-3D . SWISS-PROT is the highest priority source, all others being compared against it to eliminate identical and trivially-different sequences. The strict redundancy criteria render OWL relatively "small" and hence efficient in similarity searches. The database, OWL, is an amalgam of data from six publicly-available primary sources, and is generated using strict redundancy criteria. The database is updated monthly and its size has increased almost eight-fold in the last six years: the current version contains > 76,000 entries. For added flexibility, OWL is distributed with a tailor-made query language, together with a number of programs for database exploration, information retrieval

and sequence analysis, which together form an integrated database and software resource for protein sequences.

ii. What are protein sequence databases, explain?

ANS… In biology, a protein structure database is a database that is modeled around the various experimentally determined protein structures. The aim of most protein structure databases is to organize and annotate the protein structures, providing the biological community access to the experimental data in a useful way. A variety of protein sequence databases exist, ranging from simple sequence repositories, which store data with little or no manual intervention in the creation of the records, to expertly curated universal databases that cover all species and in which the original sequence data are enhanced by the manual addition of further information in each sequence record. As the focus of researchers moves from the genome to the proteins encoded by it, these databases will play an even more important role as central comprehensive resources of protein information. Several the leading protein sequence databases are discussed here, with special emphasis on the databases now provided by the Universal Protein Knowledgebase (UniProt) consortium.

The two protein sequence databases SWISS-PROT and PIR are different from the nucleotide databases in that they are both **curated**. This means that groups of designated curators (scientists) prepare the entries from literature and/or contacts with external experts.

SWISS-PROT, TrEMBL

SWISS-PROT is a protein sequence database which strives to provide a **high level of annotations** (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.

It was started in 1986 by Amos Bairoch in the Department of Medical Biochemistry at the University of Geneva. This database is generally considered one of the best protein sequence databases in terms of the quality of the annotation. Its size is given in the table below.

TrEMBL is a **computer-annotated supplement** of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT. The procedure that is used to produce it was developed by Rolf Apweiler. The annotation of an entry in TrEMBL has not (yet) reached the standards required for inclusion into SWISS-PROT proper. Its size is given in the table below.

	SWISS-PROT		**TrEMBL**	
Date	**Release**	**# entries**	**Release**	**# entries**
24 Oct 2001	40.1	101,737	18.0	484,388
2 Oct 2000	39.7	88,757	14.17	300,152

SWISS-PROT and TrEMBL are developed by the SWISS-PROT groups at Swiss Institute of Bioinformatics (SIB) and at EBI. The databases can be accessed and searched through the the SRS system at ExPASy, or one can download the entire database as one single flat file. An example of what an entry looks like is given for the human raf oncogene protein, ID KRAF_HUMAN.

The SWISS-PROT database has **some legal restrictions**: the entries themselves are copyrighted, but freely accessible and usable by academic researchers. Commercial companies must buy a license fee from SIB.

PIR

The Protein Information Resource (PIR) is a division of the National Biomedical Research Foundation (NBRF) in the US. It is involved in a collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japanese International Protein Sequence Database (JIPID). The PIR-PSD (Protein Sequence Database) release 70.01 (22 Oct 2000) contains 254,293 entries.

PIR grew out of Margaret Dayhoff's work in the middle of the 1960s. It strives to be **comprehensive**, well-organized, accurate, and consistently annotated. However, it is generally believed that it does not reach the level of completeness in the entry annotation as does SWISS-PROT. Although SWISS-PROT and PIR overlap extensively, there are still many sequences which can be found in only one of them.