Two Mark question and Answers

1. **Define Bioinformatics**

   Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. 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; and the development and implementation of tools that enable efficient access and management of different types of information

2. **Swiss-Prot**

   Swiss-Prot is a manually curated biological database of protein sequences. Swiss-Prot was created in 1986 by Amos Bairoch during his PhD and developed by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. Swiss-Prot strives to provide reliable protein sequences associated with 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

3. **BLAST**

   In bioinformatics, **B**asic **L**ocal **A**lignment **S**earch **T**ool, or **BLAST**, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different 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. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence. The BLAST program was designed by Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman and Webb Miller at the NIH and was published in J. Mol. Biol. in 1990.

4. **EMBL**

   European Molecular Biology Laboratory is a molecular biology research institution supported by 20 European countries and Australia as associate member state. The EMBL was created in 1974 and is a non-profit organisation funded by public research money from its member states. The cornerstones of EMBL's mission are: to perform basic research in molecular biology and molecular medicine, to train scientists, students and visitors at all levels, to offer vital services to scientists in the member states, to develop new instruments and methods in the life sciences, and to actively engage in technology transfer.

5. 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 at National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration, or INSDC. GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank continues to grow at an exponential rate. GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

6. PDB

The **Protein Data Bank** (**PDB**) is a repository for the 3-D structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world, can be accessed at no charge on the internet. The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

7. FASTA

In bioinformatics, **FASTA format** (a.k.a. Pearson format) is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences.

E.g.

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

8. ClustalW

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

8. TrEMBL

TrEMBL, ***Tr***anslated ***EMBL*** is a very large protein database in Swiss-Prot format generated by computer translation of the genetic information from the EMBL Nucleotide Sequence Database database.

Computer translation is not entirely perfect, so proteins predicted by the TrEMBL database can be hypothetical, and many TrEMBL entries are poorly annotated. In contrast to Swiss-Prot which contains only proteins actually found in the wild, and PIR which is entirely unchecked.

TrEMBL is currently being combined with the above two databases in the UniProt Consortium.

9. PIR

The **Protein Information Resource** (PIR), located at Georgetown University Medical Center (GUMC), is an integrated public bioinformatics resource to support genomic and proteomic research, and scientific studies.

10. Genome annotation

Genome annotation is the process of attaching biological information to sequences. It consists of two main steps:

1. identifying elements on the genome, a process called Gene Finding, and
2. attaching biological information to these elements.

11. Genomics

Genomics is the study of the genomes of organisms. The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping efforts. The field also includes studies of intragenomic phenomena such as heterosis, epistasis, pleiotropy and other interactions between loci and alleles within the genome.

12. Proteomics

Proteomics is the large-scale study of proteins, particularly their structures and functions. Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. The term "proteomics" was first coined in 1997 to make an analogy with genomics, the study of the genes. The word "proteome" is a blend of "**prote**in" and "gen**ome**", and was coined by Marc Wilkins in 1994. The proteome is the entire complement of proteins, including the modifications made to a particular set of proteins, produced by an organism or system. This will vary with time and distinct requirements, or stresses, that a cell or organism undergoes.

13. Multiple Sequence Alignment

A **multiple sequence alignment (MSA)** is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor.

14. Phylogenetic Analysis

Phylogenetic methods can be used for many purposes, including analysis of morphological and several kinds of molecular data. We concentrate here on the analysis of DNA and protein sequences.

Comparisons of more than two sequences

Analysis of gene families, including functional predictions

Estimation of evolutionary relationships among organisms

15. Genome Assembly

Genome assembly refers to the process of taking a large number of short DNA sequences, all of which were generated by a shotgun sequencing project, and putting them back together to create a representation of the original chromosomes from which the DNA originated. In a shotgun sequencing project, all the DNA from a source (usually a single organism, anything from a bacterium to a mammal) is first fractured into millions of small pieces. These pieces are then "read" by automated sequencing machines, which can read up to 900 nucleotides or bases at a time. (The four bases are adenine, guanine, cytosine, and thymine, represented as AGCT.) A genome assembly algorithm works by taking all the pieces and aligning them to one another, and detecting all places where two of the short sequences, or *reads*, overlap. These overlapping reads can be merged together, and the process continues.