BIF101 – INTRODUCTION TO BIOINFORMATICS
MID TERM IMPORTANT QUESTIONS

1. Define database.
   A shared collection of logically related data design to meet the information needs of multiple users in an organization the term database is often in erroneously referred to as a synonym for a ‘’database management system (DBMS)’’ .

   Database is not only being used in the commercial applications rather today many of the scientific engineering applications are also using databases less or more. Databases are concern of the respectively ladder form of appliations are more commercial applications.

2. Differentiate b/w database & algorithm?
   A database is a named location that can be used to store and organize data. And, an algorithm is a collection of steps to solve a particular problem. Learning data structures and algorithms allow us to write efficient and optimized computer programs.

3. Explain the modifications in the life cycle of protein?
   Modifications during Protein Cycle:
   Modifications those occur early in the life of the protein
   • Carboxylation of glutamate residues
   • Removal of the N-terminal methionine
   • Glycosylation
   • Addition of Prosthetic groups
   • Formation of multisubunit complexes
   • Prenylation of cysteine residues assists anchoring of proteins in or on membranes.

4. What is difference between the Modellers and I-Tasser?
   Modeller is a software for homology modelling. This link provides “MODELLER” for protein modling i.e. salilab.org/modeller. It takes input as python script file, sequence alignment and as a template (PDB).

   ITASSER stands for “Iterative Threading ASSEmbly Refinement (I-TASSER) server. It’s a software for automated protein structure and for functional prediction which is based on the sequence-to-structure-to-function.

5. Write the definitions of primary key, forms and queries.
   Primary key:
   A primary key is the column or columns that contain values that uniquely identify each row in a table. A database table must have a primary key for Optim to insert, update, restore, or delete data from a database table. Optim uses primary keys that are defined to the database.

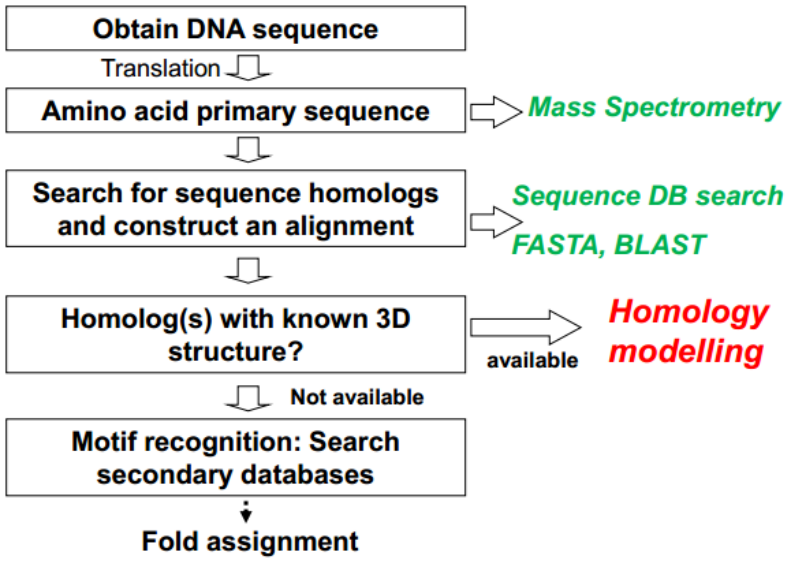
   Forms:
   A form is a database object that you can use to enter, edit, or display data from a table or a query. You can use forms to control access to data, such as which fields of data are displayed. For example, certain users may not need to see all of the fields in a table.

Queries:
A query can either be a request for data results from your database or for action on the data, or for both. A query can give you an answer to a simple question, perform calculations, combine data from different tables, add, change, or delete data from a database.

6. What is Medical Databases?
   Medical databases store and provide medical information. The premier database for biomedical literature is the National Library of Medicine (NLM)'s MEDLINE, which is accessible through PubMed.
   There are other databases where we can have medical information in addition to MEDLINE and are as follows:
   - AcademicOneFile
   - CINAHL (Cumulated Index of Nursing and Allied Health Literature)
   - PsycINFO
   - Web of Knowledge

7. What is expression proteomics ?
   EXPRESSION PROTEOMICS:
   Expression proteomics is used to study the qualitative and quantitative expression of total proteins under two different conditions.
   i. Normal and diseased state. E.g.: tumor or normal cell.
   ii. It studied that protein is over expressed or under expressed. E.g.: 2-D electrophoresis.

8. Explain work flow of Structural modeling.



9. Write applications of database management and its two operation.
   The main purpose of database applications is to provide a way for data to be consumed either by end users (via UI) or other higher-level applications (via APIs). A database application can be used for storing or retrieving data, processing transactions, or various machine learning calculations.

   Configuring authentication and authorization. Easily configure user accounts, define access policies, modify restrictions, and access scopes. These operations allow administrators to

limit access to underlying data, control user actions, and manage users in databases. Providing data backups and snapshots.

10. Define tertiary structure of proteins.
    The overall three-dimensional arrangement of all atoms in a protein is referred to as the protein's tertiary structure.

    - It includes longer-range aspects of amino acid sequence.
    - Amino acids that are far apart in the polypeptide chain may interact within the completely folded structure of a protein.
    - Interacting segments of polypeptide chains are held in their characteristic tertiary positions by different kinds of weak interactions (and sometimes by covalent bonds) between the segments.
    - Large polypeptide chains usually fold into two or more globular clusters known as domains, which often give these proteins a bi- or multilobal appearance.

11. Define Ab Initio modeling.
    Introduction:
    Ab initio method is based on Anfinsen's dogma (or thermodynamic hypothesis). This method helps us to determine the structure with minimum free energy.

    NEED FOR AB INITIO MODELLING
    Ab Initio method is applicable for all sequences. But biologically it's not very accurate. Its accuracy and applicability are limited based on our requirements.

    LIMITATION
    Ab Initio method is computationally expensive and it is only suitable for those proteins who have less than 100 residues.

12. What is Global alignment?
    A global alignment is a sequence alignment over the entire length of two or more nucleic acid or protein sequences. In a global alignment, the sequences are assumed to be homologous along their entire length.

1. BACKGROUND
In pairwise sequence alignment, we align two sequences of nucleotides or amino acids. Matches are shown in colored form whereas mismatches are denoted by ".".

2. SALIENT POINTS
Sometimes sequences are slides on each other to maximizing the matches and mis matches whereas gaps are used for deletions and insertions. Gaps are reducing overall score of alignment.

13. What are the technique of protein sequencing?
    TECHNIQUES OF PROTEIN SEQUENCING
    There are many types of techniques for protein sequencing. But we will discuss only two of them.

i. Edman Degradation: It is basically two step method. First is, labeling of amino terminal residues. Second is, removing the labeled residues.
ii. Mass Spectrometry: Mainly it consists of two steps which are MS1 and MS2. And further it is dividing on many steps. It generates spectrum against proteins or peptides. It is more reliable method than Edman degradation.

14. What are types of RNA and their function?

There are mainly three types of Ribonucleic acids (RNAs) present in the cells of living organisms.
Messenger RNA (mRNA)
Transfer RNA (tRNA)
Ribosomal RNA (rRNA)

Messenger RNA (mRNA)

- It is the type of RNA that carries genetic information from DNA to the protein biosynthetic machinery of the ribosome.
- It provides the templates that specify amino acid sequences in polypeptide chains.
- The process of forming mRNA on a DNA template is known as transcription.
- It may be monocistronic or polycistronic.
- The length of mRNA molecules is variable and it depends on the length of gene.

Transfer RNA (tRNA)

- Transfer RNAs serve as adapter molecules in the process of protein synthesis.
- They are covalently linked to an amino acid at one end.
- They pair with the mRNA in such a way that amino acids are joined to a growing polypeptide in the correct sequence.

Ribosomal RNA (rRNA)

- Ribosomal RNAs are components of ribosomes.
- rRNA is a predominant material in the ribosomes constituting about 60% of its weight.
- It has a number of functions to perform in the ribosomes.

15. What are applications of Computational genomics?

i. proposing cellular signalling networks
ii. proposing mechanisms of genome evolution
iii. predict precise locations of all human genes using comparative genomics techniques with several mammalian and vertebrate species
iv. predict conserved genomic regions that are related to early embryonic development
v. discover potential links between repeated sequence motifs and tissue-specific gene expression
vi. measure regions of genomes that have undergone unusually rapid evolution

16. What is the chemical composition of protein?

i. Proteins are polymers of amino acids
ii. They range in size from small to very large
iii. All the proteins are made up of twenty different types of amino acids. So these amino acids are called standard amino acids

iv. In a protein molecule, each amino acid residue is joined to its neighbour by a specific type of covalent bond which is called Peptide Bond
v. Amino acids can successively join to form dipeptides, tripeptides, tetrapeptides, oligopeptides and polypeptides

17. Write differnce homology amd ab intio modelling.
To make it simple in few words, Homology modeling is used to predict the structure of a protein based on the structure of its close relative (This relative is found by comparing its sequence with a sequence database) and ab initio method is used when we don't find any close relative.

18. What is Entrez?
It is an integrated search engine that works behind NCBI, so you can do lot of researches and can look for variety of data using it (It can be accessed from the site www.ncbi.nlm.nih.gov/Entrez/). It integrates PubMed and 39 other scientific literatures, nucleotide and protein databases. For example, it can be protein domain data, population studies, expression data, pathways, genome details and taxonomic information.
It is an integrated system which operates between different databases, so you can simply search for whatever you are looking after and ENTREZ will search it for you.

19. Write main difference between Similarity and identity.
The key difference between these two terms is that similarity is the resemblance between two sequences in comparison whilst identity is the number of characters that match exactly between two different sequences.

20. Define BIGDATA.
Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields offer greater statistical power, while data with higher complexity may lead to a higher false discovery rate.

21. What is fasts35?
fasts35: Compare unordered peptides to a protein sequence database

22. What are Prokaryotic cells?
No membrane enclosed internal compartments.
Plasma membrane regulates traffic (barrier).
Nucleoid region contains DNA.
Most have cell wall.

23. What are Protein database ?
Protein databases store protein data which may include the following:
Protein sequences
Motif (patterns of amino acids)
Structure
Structure alignments (aligned structures)

24. What are types of proteomics ?

STRUCTURAL PROTEOMICS:

• Structural proteomics helps to understand three dimensional shape and structural complexities of functional proteins.
• It determine either by amino acid sequence in protein or from a gene this process is known as homology modeling.
• It identify all the protein present in complex system or protein-protein interaction.
• Mass spectroscopy is used for structure determination.

FUNCTIONAL PROTEOMICS:

• Functional proteomics explains understanding the protein functions as well as unrevealing molecular mechanisms within the cell that depend on the identification of the interacting protein partners. So that detailed description of the cellular signaling pathways might greatly benefit from the elucidation of protein- protein interactions

25. Explain Human genome project.

Human Genome Project:

Human Genome Project started initially as a Pilot Project which begun by Department OfEnergy (DOE) of USA in 1986. Two organizations, one is National Human Genome Research Initiative (NHGRI), which was federally funded organization through NIH (National Institute of Health) that started in 1988 by Francis Collin which was joined to Commercial organization named *Celera (Celera Genomics)*in 1998, a commercial under the leadership of Craig Venter.

According to human genome project, the number of genes in each cell is approximately 20,000. This microscopic cell has an ultramicroscopic commanding center called nucleus in which DNA is packaged. The number of nucleotides is ~3*10^9. That is much enormous data in a cell.

26. Explain basic amino acid structure.

amino acid, any of a group of organic molecules that consist of a basic amino group (—NH2), an acidic carboxyl group (—COOH), and an organic *R* group (or side chain) that is unique to each amino acid. The term amino acid is short for α-amino [alpha-amino] carboxylic acid. Each molecule contains a central carbon (C) atom, called the α-carbon, to which both an amino and a carboxyl group are attached. The remaining two bonds of the α-carbon atom are generally satisfied by a hydrogen (H) atom and the *R* group.

27. What is Optimal energy function?

This function is easy to design. And we should remember that native structure of protein not always found at the global minimum. So, we have not clear way to generate alternative structure.

28. Expolain cell as a unit of life.

A cell is the smallest unit of a living thing. A living thing, whether made of one cell (like bacteria) or many cells (like a human), is called an organism. Thus, cells are the basic building blocks of all organisms. Several cells of one kind that interconnect with each other and perform a shared function form tissues; several tissues combine to form an organ (your

stomach, heart, or brain); and several organs make up an organ system (such as the digestive system, circulatory system, or nervous system). Several systems that function together form an organism (like a human being). There are many types of cells all grouped into one of two broad categories: prokaryotic and eukaryotic. For example, both animal and plant cells are classified as eukaryotic cells, whereas bacterial cells are classified as prokaryotic.

29. What is Coding?

Coding, sometimes called computer programming, is **how we communicate with computers**. Code tells a computer what actions to take, and writing code is like creating a set of instructions. By learning to write code, you can tell computers what to do or how to behave in a much faster way.

30. What are types of integrity?

There are mainly four types of Data Integrity:

- Domain Integrity.
- Entity Integrity.
- Referential Integrity.
- User-Defined Integrity.

31. Define scientific Application aur Local applications.

Scientific applications involve more computations, that is, different type of calculations that vary from simple to very complex. Todaysuch applications exist, like in the fields of space, nuclear, medicine that take hours or days of computations on even computers of the modern age. On the other hand, the applications that are termed as commercial or business applications do not involve much computations, rather minor computation but mainly they perform the input/output operations. That is, these applications mainly store the data in the computer storage, then access and present it to the users in different formats (also termed as data processing) for example, banks, shopping, production, utilities billing.

32. Write interactions in alpha helix structure.

1. The simplest arrangement which a polypeptide chain could assume with its rigid peptide bonds is a helical structure, which Pauling and Corey called the **α-helix**.
2. The helical twist of the α-helix found in all proteins is right-handed.
3. The repeating unit is a single turn of the helix, which extends about 5.4 Å (includes 3.6 amino acid residues) along the long axis
4. The amino acid residues in an helix have conformations with psi = –45 to –50 and phi = –60.
5. An helix makes optimal use of internal hydrogen bonds.
6. About one-fourth of all amino acid residues in polypeptides are found in α-helices while in some proteins it is the predominant structure

33. What are yypes of NCBI?

NCBI resources include Entrez, the Entrez Programming Utilities, MyNCBI, PubMed, PubMed Central, Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link (BLink), Primer-BLAST, COBALT, Splign, RefSeq, UniGene, HomoloGene, ProtEST, dbMHC,

dbSNP, dbVar, Epigenomics, the Genetic Testing Registry, Genome and related tools, the Map Viewer, Model Maker, Evidence Viewer, Trace Archive, Sequence Read Archive, BioProject, BioSample, Retroviral Genotyping Tools, HIV-1/Human Protein Interaction Database, Gene Expression Omnibus, Probe, Online Mendelian Inheritance in Animals, the Molecular Modeling Database, the Conserved Domain Database, the Conserved Domain Architecture Retrieval Tool, Biosystems, Protein Clusters and the PubChem suite of small molecule databases. Augmenting many of the web applications are custom implementations of the BLAST program optimized to search specialized data sets. All of these resources can be accessed through the NCBI home page.

34. Explain DNA and it's structure.

The chemical structure of DNA

Deoxyribonucleotides is an organic chemical that give instructions and genetic information about the synthesis of protein.

DNA is a polymer of Deoxyribonucleotides. It is composed of three components.

Deoxyribose, Nitrogenous Base, Phosphoric acid.

DNA has three parts such as; a phosphate group, a sugar group and one to four types of nitrogen bases. DNA is also composed of chemical building blocks that is called Nucleotides.

In DNA strand the nucleotides are linked into chain.

There are four bases in DNA such as

Adenine

Guanine

Cytosine

Thymine

These bases forms pairs (Adenine A with thymine T) and (Guanine G with Cytosine C)

The chemical DNA was first discovered in 1869, but its role in genetic inheritance was not demonstrated until 1943. In 1953James Watson and Francis Crick determined that the structure of DNA is a double-helix polymer, a spiral consisting of two DNA strands wound around each other. Each strand is composed of a long chain of monomer nucleotides. The nucleotide of DNA consists of a deoxyribose sugar molecule which is attached a phosphate group and one of four nitrogenous

bases: two purines (adenine and guanine) and two pyrimidines (cytosine and thymine). The nucleotides are joined together by covalent bonds between the phosphate of one nucleotide and the sugar of the next, forming a phosphate-sugar backbone from which the nitrogenous bases protrude. One strand is held to another by hydrogen bonds between the bases; the sequencing of this bonding is specific—i.e., adenine bonds only with thymine, and cytosine only with guanine.

35. What is the structure of RNA strand.

mRNA is always single stranded when it is formed from DNA.

But this single strand assumes a double helical conformation soon after its formation.

This confirmation is achieved mainly due to base stacking interactions**.**

Self-complementary sequences may occur in the RNA molecules which produce more complex structures.

So RNA can base-pair with complementary regions of either RNA or DNA.

RNA has no any regular secondary structure that serves as a reference point. The three-dimensional structures of many RNAs are complex and unique.

Breaks in the helix caused by mismatched or unmatched bases in one or both strands are common and result in bulges or internal loops.

36. Why do we perform Ab initio modelling?

Unlike the template-based modelling, successful ab initio modelling procedure could **help answer the basic questions on how and why a protein adopts the specific structure out of many possibilities**. Typically, ab initio modelling conducts a conformational search under the guid- ance of a designed energy function.

37. Why is it called central dogma in biology?

It begins with the sequence of amino acids that make up the protein. Instructions for making proteins with the correct sequence of amino acids are encoded in DNA. **Discovering this sequence of events was a major milestone in molecular biology**. It is called the central dogma of molecular biology.

38. What is simulation of folded proteins.

Simulation of the folding process

Step 1: Build an accurate initial model (including energy and forces)

Step 2: Accurately simulate the dynamics of the protein folding process

Step 3: The native structure will steadily emerge

39. Define Bioinformatics .

Bioinformatics is **an integrated field, which combines computational, mathematical and statical methods to manage and analyze the biological data or information**. Or. Bioinformatics is the management and analysis of biological information's contained in biological databases.

40. What are data records of geo?

GEO has four kinds of data records

– Platform (GPL) = the technology used and the features detected.
– Sample (GSM) = preparation and description of the sample.
– Series (GSE) defines a set of samples and how they are related.
– Datasets (GDS) sample data collections assembled by geo staff.
– Submitters may provide raw data
– Original microarray scans
– Raw quantification data

41. Steps of homology modeling

After determination of template sequence, we flow next steps. In homology modeling, there are 7 steps for prediction of proteins.

Step 1: Template recognition and initial alignment
Step 2: Alignment correction
Step 3: Backbone generation
Step 4: Loop modeling
Step 5: Side-chain modeling

Step 6: Model optimization
Step 7: Model validation

42. Why protein fold?
Proteins fold itself in 3-D structure to acquire more stability. And these proteins formed cells with the collaboration of other biomolecules.

43. Write modeller for homolgy modeling.
BACKGROUND
Homology modelling operates in 7 steps.
Step 1: Template recognition and initial alignment
Step 2: Alignment correction
Step 3: Backbone generation
Step 4: Loop modeling
Step 5: Side-chain modeling
Step 6: Model optimization
Step 7: Model validation

INTRODUCTION
Modeller is a software for homology modelling. This link provides "MODELLER" for protein modling i.e. salilab.org/modeller. It takes input as python script file, sequence alignment and as a template (PDB).

44. Define Swissprot.
SwissProt Format:
SwissProt protein sequence format is similar to EMBL format but there is considerably more information about physical and biochemical properties of a protein (as you can see below there is more description).
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.
// - Termination line.

45. What is Computational proteomics?
For advanced level of study in Proteomics, you may take "Computational Proteomics" course. You will study given below topics:
- Protein Sequencing
- PTM search
- Structure Modelling
- PPI Studies

46. What are subunit of ribosomes?
Ribosomes consist of two major components: the small and large ribosomal subunits. Each subunit consists of one or more ribosomal RNA (rRNA) molecules and many ribosomal proteins (RPs or r-proteins). The ribosomes and associated molecules are also known as the translational apparatus.

47. What is sequence analysis?
HOW DO WE SEQUENCE?
Generally, we classify sequences into two ways.
Genomes: In this type of sequence, we sequenced at genetic level in other meanings, in the form of nucleotides. E.g. AAAACCCGGGTTT etc.
Proteomes: In this type of sequence, we sequenced at protein level in other meanings, in the form of amino acids e.g. AEPEVLEGGI etc.

2. HOW DO WE COMPARE SEQUENCE?
We compared sequences into two ways.
Pair-wise Sequence Alignment: In this type of comparison, we compared two sequences with each other.
Multiple Sequence Alignment: In this type of comparison, we compared more than two sequences with each other.

TYPES OF ALIGNMENT
Mainly there are two types of alignment.
Global Alignment: In this type of alignment, we align whole sequences. In this alignment, we usually used Needle-Wunsch Algorithm.
Local Alignment: In this type of alignment, we align some specific region/part of sequence. In this alignment, we usually used Smith-Waterman Algorithm.

4. ADVANCED TOOLS
There are many types of tools for alignment. But two tools of alignments are most popular.
Fast Alignment (FASTA): In this type of tool, we align the sequences. But it is not guaranteed that FASTA can find best alignment between query and alignment because it prefers speed.
Basic Local Alignment Search Tool (BLAST): In this type of tool, we align the sequences. BLAST can search sequence databases and identify unknown sequences by

comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history. Updated version of FASTA is BLAST.

5. ONLINE DATABASES

There are many online databases. But two databases are most popular.
GenBank: In this type of database, we have genetic sequences which is regularly updated Internationally.
UniProt: In this type of database, we have protein sequences which is regularly updated Internationally.

6. ONLINE PORTALS

There are many online portals. But here we discuss only three portals.
Ensemble: It is genome search engine which is used to search the genome of every recorded species. And it is regularly updated.
ExPASy: It provides access to a variety of online databases and tools. Depending upon your requirement, you can find sequence information from ExPASy.
UniprotKB: It is the central hub for the collection of functional information's on protein. And this is a part of UniProt.

48. Define structural proteomics.

STRUCTURAL PROTEOMICS:

• Structural proteomics helps to understand three dimensional shape and structural complexities of functional proteins.
• It determines either by amino acid sequence in protein or from a gene this process is known as homology modeling.
• It identifies all the protein present in complex system or protein-protein interaction.
• Mass spectroscopy is used for structure determination.

49. What is system biology?

A biological system, sometimes simply referred to as a system, is a group of entities or organs that work together to carry out a particular task. It is a biological unit of the body or of an organism.

50. Differentiate between Fasta and Blast.

The main difference between BLAST and FASTA is that BLAST is mostly involved in finding of ungapped, locally optimal sequence alignments whereas FASTA is involved in finding similarities between less similar sequences.

51. What is Reverse and forward genomics?

Forward Genomics is a framework to associate phenotypic differences between species to differences in their genomes [1]. This framework focuses on phenotypes that were present in the common ancestor of a group of species and repeatedly lost in independent descendant species.

Reverse genetics is an experimental molecular genetics technique that enables researchers to elucidate gene function by examining changes to phenotypes (of cells or organisms) caused by genetically engineering specific nucleic acid sequences (within DNA or RNA).

52. Explain Quternary structur of protein.
    1. Some proteins contain two or more separate polypeptide chains or subunits, which may be identical or different.
    2. The spatial arrangement of these subunits is known as a protein's quaternary structure
    3. A multi-subunit protein is also referred to as a multimer
    4. A multimer with just a few subunits is called as oligomer and a single subunit or a group of subunits, is called a protomer
    5. Identical subunits of multimeric proteins are generally arranged in a symmetric .patterns
    6. Oligomers can have either rotational symmetry or helical symmetry
    7. There are several forms of rotational symmetry. The simplest is cyclic symmetry, involving rotation about a single axis
    8. A somewhat more complicated rotational symmetry is dihedral symmetry, in which a twofold rotational axis is present
    9. More complex rotational symmetries include icosahedral symmetry
    10. An icosahedron is a regular 12-cornered polyhedron having 20 triangular faces
    11. The other major type of symmetry found in oligomers is helical symmetry

53. Evolutionary system of bioinformatics
    Bioinformatics has aided research in the evolutionary process by allowing comparison of DNA sequences, sharing of data, prediction of future evolution, and classification of complex evolutionary processes.