BIF501-Bioinformatics-II

Solved Questions FINAL TERM (PAST PAPERS)

1. What is the run time of brute force algorithm?

Ans. Brute force algorithm to solve the Motif Finding problem running time of $O(l \cdot nt)$. Cannot runs it on biological samples. It is a Faster greedy techniquenot correct but have good performance. It is an Approximation algorithm , Running time of this algorithm is $O(ln2 + lnt)$, which is vastly better than the $O(lnt)$ of IMPLEMOTIFSEARCH or even the $O(4lnt)$ of BRUTEFORCEMEDIANSTRING

2. What is the difference between protein sequence and DNA sequence?

Ans:

DNA	Protein
A DNA sequence has a chain of nucleotides	A protein sequence has chain of amino acids
The nucleotides are A, C, G and T for the DNA and A, C, G and U for RNA.	There are around 20 amino acids for protein
Nucleotides are joined together through phosphodiester linkage	Amino acids are join together by peptide bonds

3. What are substitution matrices? Also mention its types.

Ans: **Scoring or Substitution Matrices:**

It is use for scoring amino acid substitutions in pairwise alignments. They reflect substitution rates that are originated by evolutionary events. For this we align the sequence and count the mutation at each position.

Types:

Some of the substitution matrices to compute sequence alignments are:

PAM: Point Accepted mutations BLOSUM: BLOCK

Substitution Matrix

4. Define suffix trees?

Ans: **Suffix Trees:**

It is a compressed tree containing all the suffixes and allows many problems on strings to be solved quickly.

5. What is the difference between smith water man algorithm and needle man wunshman algorithm in term of matrix filling?

Ans: There ar two main differences between them,

Needleman-Wunsch Algorithm

1. It performs global alignment on two sequences.
2. The final step in the algorithm is the trace back for the best alignment
   Start at the bottom-right corner. Follow where maximum value comes from

Smith-Waterman Algorithm

1. It is use to finds the best local alignment between two subsequences.

2. Traceback starts with maximum value in the matrix and then go backwards,

6. What is MATE and MATE strategy?

Ans: During fragment assembly, the reads of Sequences whose suffix and prefix are overlapping and they are likely to be consecutives are called as mate fragments or simply mates.

Mates strategy:

Overlap: Finding potentially overlapping reads

Layout: Finding the order of reads along DNA

Consensus: Deriving the DNA sequence from layout Best match between the suffix of one read and the prefix of another. By using this strategy we can sequence the whole fragment into single genome. Large number of reads to ensure that experimental errors are reduced to minor noise.

7. Which algorithm avoid the back tracking?

Ans**: KMP Algorithm:**

A linear time algorithm for string matching which does not involve backtracking of strings Is called KMP algorithm.

8. What would be the lexicographic order of, (TAT,ATG,TGG,GGT,GTG,TGC) ?

Ans: Lexicographic order of that spectrum is , **ATG,GGT,GTG, TAT, TGC, TGG**

9. l=3, TATGGTGC find the spectrum?

Ans: If l = 3 and s = **TATGGTGC**, then its spectrum will be as follow,

Spectrum(s, l) = {**TAT, ATG, TGG, GGT, GTG, TGC**}

and its Lexicographic order will be,

Lexicographic order, like **ATG,GGT,GTG, TAT, TGC, TGG**

10. What is denovo protein sequencing?

Ans: It is used for the protein that are not present in any data base, So its sequence is not known.

Or if it is known then it may be modified by some mutation or by some chemical modification.

11. Explain Primary Tertiary and Quaternary Structure of Protein?

Primary Structure: the **primary structure** of a **protein** is the linear sequence of amino acids. Together, this linear sequence is referred to as a polypeptide chain. The amino acids in the **primary structure** are held together by covalent bonds, which are made during the process of **protein** synthesis (translation).

Tertiary Structure: the tertiary **structure** will have a single polypeptide chain "backbone" with one or more **protein** secondary **structures**, the **protein** domains. Amino acid side chains may interact and bond in a number of ways. The interactions and bonds of side chains within a particular **protein** determine its tertiary **structure**.

Quaternary Structure: Protein quaternary structure is the number and arrangement of multiple folded **protein** subunits in a multi-subunit complex. It includes organisations from simple dimers to large homo oligomers and complexes with defined or variable numbers of subunits.

12. Discus RNA (10) RNA Structure and types and SecondaryStructure and Process? (10)

RNA, abbreviation of **ribonucleic acid**, complex compound of high molecular weight that functions in cellular protein synthesis and replaces DNA (deoxyribonucleic acid) as a carrier of genetic codes in some viruses. RNA consists of ribose nucleotides (nitrogenous bases appended to a ribose sugar) attached by phosphodiester bonds, forming strands of varying lengths. The nitrogenous bases in RNA are adenine, guanine, cytosine, and uracil, which replaces thymine in DNA.

RNA Structure: RNA is typically single stranded and is made of ribonucleotides that are linked by phosphodiester bonds. A ribonucleotide in the **RNA** chain contains ribose (the pentose sugar), one of the four nitrogenous bases (A, U, G, and C), and a phosphate group.

RNA Types and Function:

- mRNA - Messenger RNA: Encodes amino acid sequence of a polypeptide.
- tRNA - Transfer RNA: Brings amino acids to ribosomes during translation.
- rRNA - Ribosomal RNA: With ribosomal proteins, makes up the ribosomes, the organelles that translate the mRNA.
- snRNA - Small nuclear RNA: With proteins, forms complexes that are used in RNA processing in eukaryotes. (Not found in prokaryotes.)

13. Supervised and Unsupervised Machine Learning?

Supervised: Supervised Learning Prediction of future cases Knowledge extraction Compression Outlier detection. **Un Supervised:** Learning "what normally happens" No output.
Clustering: Grouping similar instances.
Applications Customer segmentation in CRM Image compression Bioinformatics.

14. Swiss Port?

Swissport International Ltd. is the world's largest ground handling company, **Swissport's** activities encompass everything an airline needs on the ground to move passengers and cargo efficiently and safely. The company also performs such services as line maintenance, aircraft cleaning, and fueling.

15. Secondary Structure prediction process?

Secondary structure prediction is a set of techniques in bioinformatics that aim to **predict** the local **secondary structures** of proteins based only on knowledge of their amino acid sequence.

16. Mate Pair Sequencing Strategy Sequencing?

Mate-pair **sequencing** can also provide a powerful alternative for whole genome **sequencing** as the longer DNA fragments can be used to bridge together DNA fragments even in the presence of highly repetitive DNA **sequences** which are rife in human and many other complex genomes.

17. Homology, Paralogy and Orthalogy? Homology Modeling?

Homology modeling, also known as comparative **modeling** of **protein**, refers to constructing an atomicresolution **model** of the "target" **protein** from its amino acid sequence and an experimental three-dimensional structure of a related **homologous protein** (the "template").

orthology and **paralogy** relationships between two sequences are derived from such trees based on which evolutionary event (speciation or duplication, respectively) is assigned for their last common ancestor.

18. What method used for sequence?

DNA sequencing methodsFred Sanger and Walter Gilbert Cells make copies of DNA DNA fragments of different lengths- if one base is missing.

19. Limitation of AB intio?

Computationally expensive • Suitable for proteins with less than 100 residues. Conclusion • Ab initio methods rely on computing the energies of folded proteins • The protein structures with the lowest energy are deemed as plausible predictions.

20. Challenge of AB intio?

Very hard to accurately describe energy functions that can reliably discriminate native and non-native structures. Enormous amount of computations.

21. Two target discovery approaches?

Target discovery to clinical application saga • Physiology-based approach • Target-based approach.

Physiology-based approach Is a disease-centric approach in which target is not identified, multiple targets are involved. In vivo screening is done by using drugs, siRNAor antisense oligonucleotides.

Target-centric approach Target based discovery starts with the identification of genes and their protein products. Aim to develop drugs affecting one gene or a molecular mechanism.

22. Pharmacology and Pharmacogenomics?

Pharmacology: Pharmacology (science of drugs) Genomics (the study of genes and their functions) Develop effective, safe medications & doses tailored to a person's genetic makeup.

Pharmacogenomics: Detection of genetic variability of drug effects on the genome level • Agent selection • Analysis of drug reactions and drug toxicity on gene expression • Development of new indications for already approved drugs. Discovery of new drug targets • Identification of (non) responders in clinical trials of phase I-IV • Identification of genotype dependent adverse drug reactions • Identification of individuals at risk for severe adverse drug effects.

23. Homology Folding recognition Steps? Folding recognition process of homology modeling and its important?

Homology / fold recognition predict protein structures without computing fundamental physical/chemical properties of the mechanisms and driving forces in structure formation. Ab initio methods, in contrast, base their predictions on physical models for these mechanisms • Energy released during the folding process is computed for predicting structure.

Importants:

Homology modeling is **the** most accurate computational method to create reliable structural **models** and is commonly used in many biological applications. **Homology modeling** predicts **the** 3D structure of a query protein through **the** sequence alignment of template proteins.

24. What is the simple way of scoring and algorithm?

Scoring: A simple way (but not the best) to score an alignment is to count 1 for each match and 0 for each mismatch,

Algorithm: Algorithms can be written in ordinary language, and that may be all a person needs. In computing, an **algorithm** is a precise list of operations that could be done by a Turing machine. For the purpose of computing, **algorithms** are written in pseudocode, flow charts, or programming languages.

25. Three step of needlman wunch algorithm?

Three steps • Initialization • Matrix filling • Traceback.

Initialization • The cell of first row and first column of the matrix is initially filled with zero • Add gap penalty for each shift to the right.

Matrix Fill • Move through the cells row by row, calculating the score for each cell • Compute three scores: – A match score – Vertical gap score – Horizontal gap score.

Traceback • The final step in the algorithm is the trace back for the best alignment • Start at the bottom-right corner • Follow where maximum value comes from

26. Structure of Protein?

Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule. **Proteins** are polymers – specifically polypeptides – formed from sequences of amino acids, the monomers of the polymer. ... Very large aggregates can be formed from **protein** subunits.

27. Types of RNA?

RNA Types and Function:

- mRNA - Messenger RNA: Encodes amino acid sequence of a polypeptide.
- tRNA - Transfer RNA: Brings amino acids to ribosomes during translation.
- rRNA - Ribosomal RNA: With ribosomal proteins, makes up the ribosomes, the organelles that translate the mRNA.
- snRNA - Small nuclear RNA: With proteins, forms complexes that are used in RNA processing in eukaryotes. (Not found in prokaryotes.)

28. Needleman Algorithm?

The **Needleman**–Wunsch **algorithm** is an **algorithm** used in bioinformatics to align protein or nucleotide sequences. It was one of the first applications of dynamic programming to compare biological sequences. The **algorithm** was developed by Saul B. **Needleman** and Christian D. Wunsch and published in 1970.

29. Homology Modeling?

Homology modeling, also known as comparative **modeling** of protein, refers to constructing an atomicre solution **model** of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related **homologous** protein (the "template").

2 Marks questions:

1. **Differentiate between Pharmacology and pharmacogenomics. (2)?**

   Ans:**pharmacology:**

   It is the branch of science which deals with the uses, effects and mode of action of drugs. Pharmacology is actually the study of drugs.

   pharmacogenomics:

   it is the deal with this aspect that how genes affect persons response to drugs. Pharmacology (science of drugs). Genomics (the study of genes and their functions). It develops effective, safe medications & doses tailored to a person's genetic

   makeup.

2. **Write three applications of pharmacogenomics. (2)**

   Improve drug safety

   Reduce ADRs

   Optimal dosing

3. **What do you know about GENEQUIZ. (2)?**

   Ans: There are two tools which are important worth mentioning here, one is MAGPIE and the other is GENEQUIZ and these are designed to assist with gene the genome annotations.

Join VU BIO A

GENEQUIZ- Focuses on deriving a predicted protein function based upon the available evidence; including evaluation of similarity to the closest homologue in the database (i.e. it is good tool for functional annotation).

4. **Write a note swiss port. 2**

   Swiss-Prot is a Collaboration between the SIB (**S**wiss **I**nstitute of **B**ioinformatics) and EBI (**E**uropean **B**ioinformatics **I**nstitute) and it weekly releases from about 50 servers across the world, the main source being ExPASy in Geneva.

5. **What are the limitations of ab intio? 2**

   Ans: Ab initio methods are the very difficult to design. Because it require energy function it is very difficult to find.

   These methods are slow due to the huge possibilities

   It can only deal with shorter length sequences.

6. **Why sanger award noble prize. 2**

   Ans: Fredrik Sanger got Noble Prize in 1940 for his work to sequence the proteins routinely. He basically work on the sequencing of Insulin.

7. **Write a short note on cystic fibrosis gene? 2**

   Ans: Cyctic fibrosis is a genetic disorder in which produce abnormal secretions and is diagnosed in children. It produces abnormally thick mucus that clogs the lungs.

8. **What do you know about KMP Matcher? 2**

   The KMP Matcher:

   Given: string '***S***', pattern '***p***' and prefix function **'Π'**

   Find: the occurrence of '***p***' in '***S***'

   Return: the number of shifts of '***p***' after which occurrence is found,

9. **Write the online tools for homology modeling. 2**

   Ans: Homology modelling helps predict protein structures by using prior structural information. Several tools are available to perform homology modelling in a programmatic or automated way.

   Modeller is a software for homology modeling.

10. **Write any four amino acids that are mutable. 2**

    Ans: 1) Asparagine 2) Serine 3) Glutamate 4) Aspartate

11. **XML stand for what? 2**

    Ans: XML stands for Extensible Markup Language.

Join VU BIO ANIMALS

12. Enlist online tool for homology modeling.

Modeller is an online tool for homology modeling. It is available at Salli lab. Inputs needed are

13. What is primary biological database? List four databases? (2 marks)

- Swiss-Prot and PIR for protein sequences
- GenBank and DDBJ for genome sequences
- Protein Databank for protein structures

14. **How sanger method solved Edman degradation problem of protein identification? (2)**

Ans: Sanger overcome this problem by using some digestive enzymes for the breakdown of protein. For this purpose he use protease enzyme to digest the insulin protein into peptides and then he sequence all the peptides , he also use same strategy for the fragments that are overlapping each other and come to the sequence of Insulin.

15. Differentiate b/w supervised and unsupervised machine learning. (2 marks)

Ans: Supervised: A type of machine learning in which every task has labels with it is called supervised learning.

Unsupervised: A type of machine learning which don't have any labels and that normally happens is learned is called unsupervised learning.

16. **Hamming distance and how it is calculate? (2)**

The hamming distance between two equal strings of sequences is the number of entries/bases which are different. For example if there are two strings

String 1 - A T C G C T G A T

String 2 - A C C G T T G A T

These two strings has 2 mismatches so the hamming distance between them is 2.

17. What is Non canonical splicing junctions? (2)

Ans: The splice junction other than AG and GT are called as non-canonical splice junctions. For example the percentage of occurrence of canonical junction is 99.24% the rest are non canonical which may be,

GC-AG (0.69%) AT-AC (0.05%) Others (0.02%)

18. When short gun sequencing is used normally? 2

Ans: shotgun is a method of sequencing in which large genomes are fragmented into short reads and then sequenced individually. So when the size of the sequence is large normally shotgun sequence is used.

19. How many base pair triptosin gene has in human T cells? 2

Ans: In human T-cell locus, there are 5 closely related fragments of triptosan gene and the size of this triptosan gene is 4kb. It is a large size fragment and we can't read it because read limit is only 500-700bp. These sequences repeat 1 million times in the human genome.

3 marks questions:

20. **Explain structure annotation and technique? (3)**

Ans: Structural annotations are physical regions of a genome that encode a genomic feature. In structural annotations where we try to identify features of the genes like Promoters , Terminators , Shin-dalgarno site, DNA motifs, Co-transcription units , Operons in microbes. We can use this technique we can identify the regions within genes and their functions.

21. Why BAC is used? (3)

Ans: BAC is the abbreviation of bacterial artificial chromosome. These are specially design vectors in which the large size of the fragments can be ligated, the size of fragments will be almost 150,000 bp. BAC is mostly used for sequencing. Now a days BACs are also called as mini genomes. Using BAC is helpful to simplify computational assembly problems by reducing the numbers of times have to be sequenced.

22. Explain fasta and NBRF formats with example. 3

FASTA Sequence Format

FASTA is the most recognized and well distributed format to present DNA and Protein sequences. The sequence starts with a '*greater than*' sign (>), whereas the actual sequence is always on the next line. It is recommended that all lines of text should be shorter than 80 characters in length (generally we have 60 characters)

Example:

>gi|568815581:c7687550-7668402 Homo sapiens chromosome 17, GRCh38 Primary Assembly
GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCATCTAGAGCCACCGTCCAGGGAGC
AGGTAGCTGCTGGGCTCTCCACGACGGTGACACGC--------

NBRF, **N**ational **B**iomedical **R**esearch **F**oundation.

Join VU BIO ANIMALS

23. Explain spliceosome? 3

1. Spliceosome is a large and complex molecular machine that is basically used in the splicing process.
2. It is approximately contain more than 100 proteins.
3. The spliceosome removes introns from a transcribe pre-mRNA which is a type of primary transcript.

24. What is the function of DELETE(y, L)? 3

Ans: In Brute Force Algorithm we make a List of pairwise distances, L, and uses the function DELETE(y, L). It means the value of "y" will be deleted from the subset "L".

25. What are the Challenges of AB initio? 3

Ab initio methods base their predictions on physical models of folding mechanisms. Stabilization is measured by energy released during the folding process. Start with an energy function. Fold structures in order to obtain the most stable structure. This structure will have the minimum energy.

26. What is Generalize algorithm for pattern finding? 3

Pattern Finding:

– Much of the data processing in bioinforma[cs involves searching and recognizing certainpatterns within DNA,

RNA or protein sequences.

– In Biology it means finding moifs in DNA or proteins while in computa[onal means it is finding a pattern in a string.

Goal: Finding all occurrences of a pattern in a text

Input:

Pattern p = [p1…pn] of length n

Text t = [t1…tm] of length m

Output:

An indication that pattern P exist in T or it does not exist in text T

27. Explain the digital revolution in recent time? 3

Pakistan is a country with limited scientific infrastructure. In our country we don't have high resolution instruments in the labs to perform scientific research. But to our rescue there comes digital revolution. Developed countries like US, Europe, Japan, Australia and Canada are producing very huge amount of digital data which is available online. With the digitalization of this data and its availability through internet we in Pakistan have access to this data. By simply obtaining these data sets and processing it

Join VU BIO ANIMALS

and analyzing it we have the capacity to make novel discoveries. So this is a big opportunity for us to work with that data only with a computer and internet and to buy very expensive instrumentation is not needed any more.

28. Longest common subsequence problem (only single strip) with example (3

Largest Common Subsequences (LCS)

The simplest form of a sequence similarity analysis is the Longest Common Subsequence (LCS) problem, where we eliminate the operation of substitution and allow only insertions and deletions. A subsequence of a string **v** is simply an (ordered) sequence of characters (not necessarily consecutive) from **v.**

Typically there are many common subsequences between two strings **v** and **w**, but how to find the longest one. s(**v**,**w**)-be the length of the longest common subsequence of **v** and **w** edit distance between **v** and **w** under the assumption that only insertions and deletions are allowed.

d(**v**,**w**) = $n + m$ -2s(**v**,**w**)

This corresponds to the minimum number of insertions and deletions needed to transform **v** into **w.**

Longest Common Subsequence Problem:

Find the longest subsequence common to two strings

Input: Two strings, **v** and **w**

Output: The longest common subsequence of **v** and **w**

29. **Write down Dynamic programming and its three steps (3)**

Ans:

Dynamic Programming:

Dynamic programming is used to find an optimal alignment of two sequences and its scores. It is a method by which a larger problem may be solved by first solving smaller, partial versions of the problem. **Three steps in dynamic programming:**

• Initialization

• Matrix filling (scoring)

• Traceback (alignment

1. **Initialization:** Create matrix with M+1 columns and N+1 rows where M and N correspond to the size of sequences to be aligned

Join VU BIO ANIMALS

2. **Matrix filling:**Fill the matrix with highest possible score

3. Trace back: Move from the last corner and follow the arrow

30. What are personalized medicine? 3

The drugs which are made specifically for a patient with respect to its genome, physiology and system biology are called personalized medicine. It's a hot cake today to grow in this field being a bioinformatician.

31. What are cluster linkage and subgroups? 3

Clustering by Linkage :

There is another technique, where the clustering is done by linkage (almost similar to the previous ones except some changes). The method of doing it is as follows:

- Each sequence is a vertex.
- Significant alignment score is an edge.
- Trimming by removing weak edges (High P/E).
- Or remove > e-6 (we remove the ones which are not linked).
- Remaining sub-graph should share 2/3rd of edges.

32. What is Brute Force and Greedy algorithm? 3

Brute force algorithms are exhaustive in nature and search each and every nucleotide. They take much more time.

While greedy algorithms focus only required data by skipping the useless steps and take shorter time to retrieve and give result.

33. What is partial consensus score?

Ans: The difference between calculate score Si of any substring from total score t of superstring is called partial consensus score. Partial consensus score = total score St-Si or simply (t-i)

34. What is infram hexamer count?

Ans: While finding exons in the transcription start sites the frame is shifted from left to right by leaving one nucleotide every time and adding next nucleotide in the frame. Then all the frequencies of codons are counted which are likely to code an amino acid more often. This frequency count is called as inframe hexamer count.

35. What is Mate?

Ans: During fragment assembly, the reads of Sequences whose suffix and prefix are overlapping

Join VU BIO ANIMALS

and they are likely to be consecutives are called as mate fragments or simply mates.

5 Marks questions:

36. **Differentiate between homology, orthology, paraology? (5)**

Homology:

Homology is the combination of orthologus and paralogus. Homology modelling is used to predict structures of proteins having high sequence similarity with other proteins with known structures. OR The duplication of genes results into different homologs during course of evolution. Study of these homologs which have same ancestral origin are called as homology.

Orthology:

If same gene is present in two different species. Orthlogs are those genes which are originated from gene duplication from a single gene and are present in different organisms. Study of orthologs is called as **orthology**. For example alpha-globin genes present in frogs, chicks and mouse are orthologs.

Paralogy:

In this case two different genes are present in the same species encoding two different types of proteins. Homologs which arise from rare gene duplication but are present in same species having different genetic products are called paralogs. Study of these paralogs is called as **paralogy**. For example alpha-globin and beta-globin both are present in mouse.

37. **Explain needle wunsch algorithm. (5)**

Ans: The Needleman-Wunsch algorithm finds best global alignment between two sequences. It performs global alignment on two sequences. The algorithm was developed by Saul B. Needleman and Christian D. Wunsch and published in **1970.**

Basic idea is

• to build up the best alignment by using optimal alignments of smaller subsequences

• It was the first application of dynamic programming to compare biological sequences

Steps:

There are three steps as following,

• Initialization

• Matrix filling

• Traceback

Creating the Matrix:

Initial matrix is created with M+1 columns and N+1 rows

Join VU BIO ANIMALS

• Where M and N correspond to the length of sequences

Initialization:

• The cell of first row and first column of the matrix is initially filled with zero

• Add gap penalty for each shift to the right

Matrix Fill:

• Move through the cells row by row, calculating the score for each cell

• Compute three scores:

– A match score

– Vertical gap score

– Horizontal gap score

• The *match score* is the sum of the diagonal cell score and the score for a match

• The *horizontal gap score* is the sum of the cell to the left and the gap score

• The *vertical gap score* is computed analogously

Traceback:

• The final step in the algorithm is the trace back for the best alignment

• Start at the bottom-right corner

• Follow where maximum value comes from

• Scoring scheme introduced can be user defined

• It contains specific scores for match and mismatch residues as well as gap

Conclusions:

• Needleman and Wunsh Algorithm performs global alignment on two sequences using a dynamic programming approach

38. **Explain protein translation. (5)**

Ans: protein translation has three basic steps.

1. Initiation
2. Elongation
3. Termination

1. Initiation:

As the translation takes place in ribosome. Ribosome divided into two subunits.

1. smaller subunit 2. Larger subunit

Join YU BIO ANIMALS

When these subunits combine with the help of other proteins form a initiation complex afterthis the molecule of mRNA binds with the smaller subunit. The larger subunit has three basicspecific sites.

A site: generally called as the accepter site. Where t RNA has 1st comes and bind at his site.

P site: it is the 2nd site responsible of translation.

E site: it is referred as exit site. From where t RNA exit from ribosome.

When m RNA binds with the smaller subunit of the ribosome it recoganised special three nucleotide codons. 1st codon is AUG present on m RNA and its complimentary codon is on t RNA. From here the initiation will start.

2. Elongation:

Once a t RNA binds to the codon while moves through P site it will be exit from E site, then the next t RNA comes to A site and binds codons in similar way and exit from E site. During this activity each tRNA binds amino acid molecules together which they have with them. There will be a amino acid chain that is generated in the form of polypeptide chain. This chain will be elongated until the desired structure of protein will be formed. So this step is known as elongation.

3. Termination:

Elongation process will stop when the stop codons (UAG, UAA, UGA) will arrived at A site. When such a specific t RNA reach to A site the termination process occur which follows the segregation. In which smaller and larger subunits will separate and protein will be librated. After this some post translational changes occur that make this protein functional.

39. Explain Target discovery. (5)

Target discovery is mainly used for clinical application saga

• Physiology-based approach

• Target-based approach

1. **Physiology-based approach**

Is a disease-centric approach in which target is not identified, multiple targets are involved. In vivo screening is done by using drugs, siRNAor antisense oligonucleotides.

This process relies on disease, not on the target.

2. **Target-centric approach**

Target based discovery starts with the identification of genes and their protein products. Aim to develop drugs affecting one gene or a molecular mechanism. The identification of disease-relevant genes in vitro cellular models has been possible due to several tools. Gene-suppression tool used to linked the genes with disease.

Join VU BIO ANIMALS

Target has two types

• Genetic

• Mechanistic

Genetic targets are represented by genes and genes products. Mechanistic targets include mechanism based targets such as receptors, enzymes or genes, identified on the basis of the disease state.

40. Explain fragment assembly algorithms 5

When 500 – 700 bp fragments are sequence then the next step is to assemble them into a entire genome. This assembly is called fragment assembly. While assemble the fragments there are certain problems in which the major problem is the DNA repeated sequence. These repeats creats problems when we read the sequence bcz reads are only 500 – 700 bp but sometimes repeated sequence much larger then reads. This problem could be trace by increasing the size of reads but it is not yet possible so for this purpose we clone the large fragment of genome into BAC.

James Weber and Gene Myer introduce this strategy. Gene Myers basically work on fragment assembly problem and demonstrated that read size should be increase from 500 to much higher to overcome the problem of this repeated units. They also suggest that the both ends of the fragments should be sequenced. This is also known as Mates strategy.

Most fragment assembly algorithms consist of the following three steps

Overlap: Finding potentially overlapping reads

Layout: Finding the order of reads along DNA

Consensus: Deriving the DNA sequence from layout Best match between the suffix of one read and the prefix of another. By using this strategy we can sequence the whole fragment into single genome. Large number of reads to ensure that experimental errors are reduced to minor noise.

41. Explain Endosymbiont hypothesis. 5

Ans: There is ahypothesis that mitochondria actually evolvedfrom bacteria and is known as endosymbiont hypothesis.

Endosymbiont hypothesis:

According to this hypothesis, these organelles originated as separate prokaryotic organisms that were taken

inside a primordial eukaryotic cell. Such symbiotic relationships in which two species are dependent upon

one another to varying extents served as crucial elements of the evolutionary progression ofeukaryotic cells.

Join VU BIO ANIMALS

This hypothesis was originally proposed in 1883 by ***Andreas Schimper***, butextended by ***Lynn Margulis***in the

1980s. So according to this theory, ***Mitochondria*** and***Chloroplast***are derived from endosymbiotic bacteria

(that got incorporated into the cells).

42. List the stop and start codons and their functions? 5

Gene finding, especially in prokaryotes starts form searching for open reading fram (ORF).An ORF is a sequence of DNA that starts with start codon "ATG" (not always) and ends withany of the three termination codons (TAA, TAG, TGA).Depending on the start point, we can define different ORFs, so if we want to go for those tripletcodons, we can start with any nucleotide, in this way we have three different possibilities for oneof the strands and since we have two strands in the DNA, so in total we can have six ORFs (so 3of them are from 3' to 5' direction whereas the other 3 are from 5' to 3' direction).

43. Edit distance? With examples.

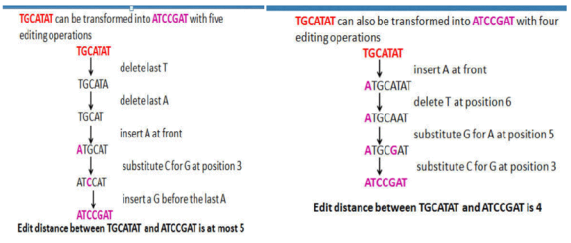
Edit distance between two strings as the minimum number of editing operations needed to transform one string into another, where the edit operations are insertion, deletion, andsubstitution of one symbol for another It is

often the case that the *i*th symbol in one sequence corresponds to a symbol at different position in other. Mutation in DNA-evolutionary process: DNA replication- substitutions, insertions, and deletions of nucleotides,

leads to "edited" DNA

texts. Whether the *i*th symbol in one DNA sequence corresponds to the *i*th symbol in the other.

Example:



44. Write down the Five tasks of genome analysis. 5

Genome Analysis:

In Genome Analysis we mainly perform the following tasks; (you can write any five from following)

- ***Sequencing***
- ***Assembly*** (since the sequencing is done in a way that whole genome is broken down into

JU BIO ANIMALS

short fragments and once those fragments are sequenced, we need to put them together, this step in genome analysis is known as Assembly).

• ***Repeat identification and masking out*** (once we assemble that genome, we try to find out the regions in which we have large number of repeats because assemblies jumble up where we have those repeats so we need to find those regions and it is one of the important task to go and look into those assemblies while keeping in mind those regions in which we have those repeats).

• ***Gene prediction*** (after we have assembled a finished genome, now we can go for the prediction of the genes where we can find the genes by using different patterns or features of those genes).

• ***Genome annotation*** (in which we can find out similar functions performed by different genes)

• ***Expression analysis*** (once we have the idea about the regions of the gene in which we can have the gene expression then we can explore the quantification i.e. how much those genes are being expressed).

• ***Metabolic pathways and regulation studies*** (once those genes are expressed, their products interact with each-other and then they perform different metabolic roles in the shape of different metabolic pathways and networks).

• ***Functional genomics*** (where we are actually looking into the different functions performed by different regions of the genome that are under the control of different genes and what exactly would be the effect of changes in those genes specifically if we want to study about the genes related to diseases).

• ***Gene location/gene map identification*** (map the location of those genes on the chromosomes).

• ***Comparative genomics*** (in which we can take one genome and compare it with another genome, where we can find the comparative features; what is present in the first genome and not in the second one and the intersections between them, etc.).

• ***Evolutionary modeling*** (so the identification of the clusters of functionally related gene can help us in making an evolutionary model)

Join VU BIO ANIMALS

45. Explain two computational problems which are used to identify proteins 5

- Protein identification problem
- Modified protein identification problem

Protein identification problem:

Problem: Find a protein from data base that best matches with experimental spectrum S.

Input A protein database, experimental spectrum S, a set of ion types and a parent mass m.

Output: A protein of mass m from data base that best matches with experimental spectrum S

Modified protein identification problem:

Problem Find a peptide in a protein database with best match to the experimental spectrum with k modifications

Input: Protein database, experimental spectrum S, a set of ion types, a parent mass and parameter k cappingthe modifications

Output:A protein of mass m with best match to experimental spectrum S at most k modifications away from an entry in database

46. Explain gene expression omnibus (5 marks)

GEO is a public repository for archiving and distribution of gene expression data submitted by scientists.

It is under the umbrella of NCBI. It is efficient in submission and retrieval of data by browsing.

The minimum requirement for data submission in GEO can be found from MIAME (Minimum information about microarray experiment).

There are four records of GEO.

- **Sample (GSM)**
  - It is information about samples, their collection, their storage, resource and use.
- **Platform (GPL)**
  - This kind of record is required about technology used and in which category the data is to be submitted like microarrays or RNAseq.
- **Series (GSE)**

Join VU BIO ANIMALS

- If samples are more than one then they are arranged in a series order.

- **Datasets (GDS)**
  - The data which is to be submitted to GEO.

47. Write approaches for prediction of exon. 5

Exons can be predicted within ORF from the information gathered about

• splice junctions

Approach:

For each segment [acceptor, donor], we get three scores (coding potential, donor score, acceptor score)

Various possibilities

all three scores are high – probably true exon.

— all three scores are low – probably not a real exon.

— all in the middle -- ?.

Some score are high and some are low.?

we can get the evidence by the help of that information which we gathered from those splice sites.

Prediction:

• Collect a set of exons and non-exon

• Score them using our scoring schemes

• Plot them as follows

draw a separate line between exon and non-exon.

Linear Discriminate Analysis: (LDA)

— linear discriminate analysis (LDA) finds an optimal plane surface that best separates points that belong to two classes.

— For example, if there are ten true exons and ten introns, and two feature.

— These samples could be represented by 20 points in a two

— LDA would compute a straight line through the space that can best separate the two classes with the minimal classification error.

Join VU BIO ANIMALS

48. Differentiate b/w prokaryote and eukaryote genome. 5

Prokaryotic Genome	Eukaryotic Genome
Prokaryotes genetic material is not enclosed in membrane	Genetic material is enclosed in nuclear membrane
Organelles are not membrane bounded	Organelles are membrane bounded
They are small in size	They are larger than prokaryotic genome
Contains single circular chromosomes	Contains multiple linear chromosomes
They are haploid	They are diploid
DNA is not associated with histone	DNA is associated with histones
They are condensed in nucleoid via DNA supper coiling	They are condensed via histone , Euchromatin and Euchromatin
No introns present	Introns are present between exons

49. Differentiate satellite DNA, micro and mini DNA?

Satellite DNA: Satellite DNA has following features:

- It may be one to several thousand bp long and it can also be present as million bases long.
- They are present near centromere and telomere. It can be classified as mini satellite and micro satellite.

Mini-satellite: Mini-satellite features are as follows:

- They are 15 bases long in array of several hundred to thousands kb.
- They are typically present in euchromatin region.
- Example is VNTR and is used to identify human individuals in forensics.

Microsatellite: Micro-satellite features are as follows:

- They are 2-6 bases long and can be in arrays of 10-100 bases.
- They are inherited to offspring.
- Mainly they are useful markers for genetic analysis and evolutionary studies (just like in previous example)

Join VU BIO ANIMALS

- They are found in telomeres TTAGGG (one example of six nucleotides segment).
- SSRs and STRs (**S**imple **S**egment **R**epeats and **S**hort **T**andem **R**epeats) are typical examples.

50. **Why Edman degradation failed in protein sequencing?**

Edman degradation method was based on removal of terminal amino acid by labeling it. It failed because when the terminal amino acid encountered is modified by acetylation already, then it will stop. Similarly is required more than one pico mole of peptides to start the degradation. Also it could not identify disulphide bridges in the proteins. Due to these reason Edman Degradation failed to sequence proteins.

51. Explain synteny of human X-chromosome.

If human X-chromosome is cut into 5 large blocks of genes, each block will be called as synteny. So in this way human X-chromosome will have 5 synteny. Similarly if we cut mice X-chromosome into 5 blocks or synteny and compare it to the human X-chromosome synteny is similar to the mouse X-chromosome with different locations. If we rearrange mice X-chromosome synteny according human, then it would take only 3 flipping to become exactly human X-chromosome. This similarity in the synteny of human and mice chromosome reflect common ancestry.

52. Explain components of KMP algorithm.

KMP Algorithm:

A linear time algorithm for string matching which does not involve backtracking of strings

Is called KMP algorithm.

Components of KMP algorithm:

There are two components of KMP algorithm

- The prefix function Π
- KMP matcher

Prefix function Π:

The component of KMP which encapsulate the knowledge about how pattern matches against the shifts of itself is called prefix function. The information obtained from shifts can be used to Avoid useless shifts of pattern P. So it enables avoiding backtracking on string S.

KMP matcher:

Join VU BIO ANIMALS

Given: String S, pattern P and prefix function Π.

Find: Occurrence of pattern p in String S.

Return: Number of shifts of p after which occurrence is found.

10 Marks question:

53. What is RNA? Explain types with function and how secondary structure is predicted?

Definition:

Ribonucleic acid (RNA) is a single stranded polymeric macromolecule consisted of four types of nucleotides which are joined together through phosphodiester linkage. Each nucleotide is made up of three molecules namely ribose sugar, phosphate group and nitrogenous base. There are four types of nitrogenous bases in RNA namely cytosine, guanine, uracil so there are four kinds of nucleotides.

RNA types with function:

- **Messenger RNA(mRNA)**
  - Protein translation and coding
- **Transfer RNA (tRNA)**
  - Protein translation
- **Ribosomal RNA(rRNA)**
  - Protein translation
- **Heterogeneous nuclear RNA(hnRNA)**
  - Precursor and intermediate of mRNA and other RNAs
- **Small cytoplasmic RNA (scRNA)**
  - Signal recognition particle and tRNA processing
- **Small nuclear and nucleolar RNAs(snRNA and snoRNA)**
  - mRNA Processing, maturation, poly adenylation and rRNA processing, maturation and methylation
- **Regulatory RNAs (Small interfering RNA and microRNA)**
  - Regulation of transcription and translation and other

Secondary structure of RNA

Non coding RNAs have to stay in the cytoplasm so they have to be stable to avoid degradation. For this purpose they fold it by pairing with complementary bases of strand. During folding to form secondary structure there are several conformations which are created in the secondary structure. Some of them are

Loops:

Join VU BIO ANIMALS

During folding process if few bases don't have their complementary bases and could not pair and remain single stranded in the folding process they form loops mostly commonly hair pin loops. Hair pin loops results into 180 degree bending of strand. If there are four bulge of equal base pairs on either side, it results into junction loop.

Bulges:

If one or two bases left unpaired, in a continuous strand they result into bulges which look like bulge originating out from continuous strand on one side and result in 90 degree bending of structure. If bulges are present interior to the strand on both sides, they are called interior bulges and chain remain continuous without considerable bend.

Stem:

The double strand where no bulges and loops are present is straight in structure and is called as stem.

Pseudoknots:

When two hairpin loops or bulges connect with each other externally, they result in Pseudoknots. Pseudoknots are big challenge for algorithms to consider them.

Secondary structure prediction:

There are different approaches to predict secondary structure of RNAs.

- Energy minimization
- Comparative sequence analysis
- Base pair maximization
- Folding and alignment

Energy minimization:

The algorithm which is followed in this strategy has following charactrisitics

- Thermodynamically stable
- Energies are estimated using experimental techniques
- More stable is more likely structure
- No Pseudoknots
- Uses dynamic programming alignment strategy
- Scores are maximized thermodynamically
- Example include MFOLD and Vienna RNA

Comparative sequence analysis:

In this technique already available data is used to compare and predict the structure which is a best match. It requires multiple sequence alignments with best scores. Also it requires less matching sequences to predict covariance of junctions.

It gives more reliable and accurate results.

Join VU BIO ANIMALS