MUHAMMAD IMRAN

BIF501 - Bioinformatics II

Current paper solved question mid term December 2019

1. Define Pseudogenes: 2 marks

These are non-functional genes (sometimes there are mutations in the genes and if those mutations are present in some important regions then the genes' functions gets knocked out – known as psudogenes). There is one category of them and is known as processed pseudogenes that lack introns and promoters.

A DNA sequence that resembles a gene but has been mutated into an inactive form over the course of evolution. It often lacks introns and other essential DNA sequences necessary for function. Though genetically similar to the original functional gene, pseudogenes do not result in functional proteins, although some may have regulatory effects.

2. Mitochondrial DNA and Chloroplasts DNA .2

Mitochondrial DNA (both in animals and plants), **Chloroplasts DNA** (in plants), these are membrane-bound organelles and they may be present in hundreds to thousands of copies (so there is also multiple copies of these genomes). Mitochondrion is the site for respiration whereas chloroplast is the site for photosynthesis. Their DNA's can be labeled as mtDNA or cpDNA respectively.
Plasmids, yeast, Transposons, Viral genomes and retroviruses are other examples of organellar DNA.

3. Write detail on sequence LOGOS?

A sequence logo is a graphical representation of the sequence conservation of nucleotides (in a strand of DNA/RNA) or amino acids (in protein sequences) . A sequence logo is created from a collection of aligned sequences and depicts the consensus sequence and diversity of the sequences. **(Internet)**

Sequence LOGOS:
─ A visual representation of a position-specific distribution
─ Easy for nucleotides, but we need colour to depict up to 20 amino acid proportions.

Overall height at position is proportional to the information content
• Proportions of each nucleotide/amino acid are in relation to their observed frequency, with most frequent on top, next most frequent below

4. Biological Databases: 3

Biological databases in general store biological data and their main goals are
1. Data storage
2. Information retrieval
3. Knowledge discovery

Classification:
Biological databases can be classified as
➢ **Primary databases (**that stores the Primary Sequences**)**
➢ **Secondary databases (**the primary sequences are annotated and kept in Secondary Databases**)**
➢ **Specialized Databases (**they are dedicated towards some specific organism or can have some disease data**)**

Biological databases can also be classified on the bases of types of data which they contain, such as:
➢ Nucleotide databases
➢ Protein databases
➢ RNA databases
➢ Genome databases
➢ **Expression databases (**Gene Expression Databases**)**

5. KMP Matcher?

Knuth Morris Pratt (KMP) is an algorithm, which checks the characters from left to right. When a pattern has a subpattern appears more than one in the sub-pattern, it uses that property to improve the time complexity, also for in the worst case. The time complexity of KMP is O(n).J **(Internet)**

Drug Development, Molecular Medicine, Crop Improvement, Personalized Medicine, Microbial Genome. Preventive Medicine, Gene Therapy. Biotechnology, Comparative Study.

Introduction:
A linear time algorithm for string matching
• Does not involve backtracking on string ***s*** i.e., repetitive comparison of nucleotide residues

Components:
• The Prefix Function

• The KMP Matcher

The Prefix Function II:
Encapsulates knowledge about how the pattern matches against shifts of itself
• This information can be used to avoid useless shifts of the pattern ***'p'***

• This enables avoiding backtracking on the string ***'S'***

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

6. Tandem repeats interspersed repeats

Tandem repeats occur in DNA when a pattern of one or more nucleotides is repeated and the repetitions are directly adjacent to each other. Several protein domains also form tandem repeats within their amino acid primary structure, such as armadillo repeats.
Interspersed repeats are **repeated** DNA sequences located at dispersed regions in a genome. They are also known as mobile elements or transposable elements. As discussed in Chapter 8 Section D, a stretch of DNA sequence may be copied to a different location through DNA recombination
they can be *Tandem repeats* (array of repeats together), *interspersed repeats* (those repeats which are separated by normal genome followed by repeats i.e. normal genome is between those repeats).

Short Tandem repeats
Analyzing STRs is more accurate than the RFLP technique because their small size makes them easier to separate. If you want to create a fingerprint, you might look at 20 different STRs at different places in order to create a profile.

7. Genome and Genome informatics? 2

Genome: Genome sequencing provides the sequences of all the genes of an organism • A major application of Bioinformatics is analysis of full genomes that have been sequenced. Challenge is to identify those genes that are predicte d to have a **particular biological function**

Genome Information:

Genome informatics is the field in which computational and statistical techniques are applied to derive biological information from genome sequences. Genome informatics includes methods to analyze DNA sequence information and to predict protein sequence and structure.

8. Generalized algorithm for pattern finding?

Goal: Find all ocuracne of a pattern in 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 T.

9. MS2?

Infects E.Coli and Enterobacteriaceae • Single stranded sense RNA • Encodes MS2 coat protein. MS2 sequencing
• The entire nucleotide sequence was established by nuclease digestion and characterization

10. What is Donor site and Acceptor Splice Junctions:

Donor site
• Coding region | GT (introns starts)

Acceptor
• (introns ends)YAG | coding region -Y can be any pyrimidine.

11. Partial Digest Problem:

Given all pairwise distances between points on a line, reconstruct the positions of those points. Input: The multiset of pairwise distances L, containing (n) integers.

2

Output: A set X, of n integers, such that $\Delta X = L$

ΔA is equal to $\Delta(A \oplus \{v\})$, where $A \oplus \{v\}$ is defined to be

$\{a + v : a \in A\}$,

Also $\Delta A = \Delta(-A)$,where $-A = \{-a : a \in A\}$

$A = \{0, 2, 4, 7, 10\}$, $\Delta(A \oplus \{100\}) =$

$\{100, 102, 104, 107, 110\}$, and $-A = \{-10,-7,-4,-2, 0\}$

$\{0, 1, 3, 8, 9, 11, 12, 13, 15\}$ and $\{0, 1, 3, 4, 5, 7, 12, 13, 15\}$

12. Difference local alignment input and output? 5

Local Alignment Problem : Find the best local alignment between two strings

Input : Strings v and w and a scoring matrix δ

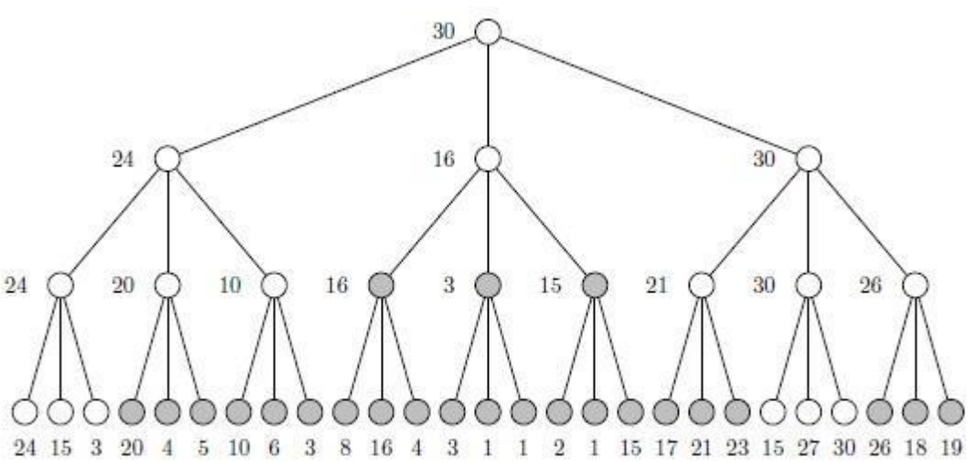
Output: Substring of v and w whose global alignment, as defined by δ, is maximal amoung all global alignment of all substrings of v and w.

13. Four cases of pattern finding?

Pattern recognition is the automated recognition of patterns and regularities in data. Pattern recognition is closely related to artificial intelligence and machine learning,[1] together with applications such as data mining and knowledge discovery in databases (KDD), and is often used interchangeably with these terms. However, these are distinguished: machine learning is one approach to pattern recognition, while other approaches include hand-crafted (not learned) rules or heuristics; and pattern recognition is one approach to artificial intelligence, while other approaches include symbolic artificial intelligence

14. Bypass Algorithm?

Branch and Bound Algorithm
NEXTVERTEX Algorithm
BYPASS Algorithm
Skip the subtree rooted at vertex (a, i) Increment a_i (unless $a_i = k$



A tree that has uninteresting subtrees. The numbers next to a leaf represent the “score” for that L-mer. Scores at internal vertices represent the maximum score in the subtree rooted at that vertex. To improve the brute force algorithm, we can “prune” subtrees that do not contain highscoring leaves. For example, since the score

of the very first leaf is 24, it does not make sense to analyze the 4th, 5th, or 6th leaves whose scores are 20, 4, and 5, respectively. Therefore, the subtree containing these vertices can be ignored.

```
BYPASS(a, i,L, k)
1.  for j ⟵ i to 1
2.      if a_j < k
3.          a_j ⟵ a_j + 1
4.          return (a, j)
5.  return (a, 0)
```

15.gene prediction

In computational biology, **gene prediction** or **gene finding** refers to the process of identifying the regions of genomic DNA that encode **genes**. This includes protein-coding **genes** as well as RNA **genes**, but may also include **prediction** of other functional elements such as regulatory regions.

16. What Kind of Gene Expression OMNIBUS? GEO

A public repository for the archiving and distribution of gene expression data submitted by the scientific community. MIAME compliant data – Minimum Information About a Microarray Experiment. Convenient for deposition of gene expression data, as required by funding agencies and journals • Curated resource for gene expression data • Browsing, querying, analysis and retrieval.

GEO has four kinds of records • **Sample (GSM)** preparation and description of the samples • **Platform (GPL)** technology used and the features detected – microarray or RNASeq. **Series (GSE)** defines a set of samples and how they are related • **Datasets (GDS)** sample data collections assembled by GEO

17. What are suffix tree ? 2

It is a compressed tree containing all the suffixes and allows many problems on strings to be solved quickly. • Exact searching or pattern matching methods • Approximate searching or pattern matching methods • Position weight matrices. • Suffix trees.

18. How mutation in DNA?

“polymerase chain reaction”(PCR) to permit rapid analysis of any known **mutation** in genomic **DNA**

19. XML Format:

It is a modern practice in which we try to put those sequences in kind of a machine language. So, XML stands for Extensible Markup Language. The format is similar to HTML (language for Web programming). The good part is that this language is in between machine and man readable so it’s kind of easy to code over this. And it is becoming standard data format for transferring genome data.

Example:

<xsd:annotation>
<xsd:documentation>
XML Schema for SBOL core data model compatible with RDF/XML serialization.
<dc:date>2012-01-19</dc:date>
<dc:creator>EvrenSirin</dc:creator>
<dc:contributor>Michal Galdzicki</dc:contributor>
</xsd:documentation>
</xsd:annotation>

20. NF-kB binding sites

NF-kB binding sites (nuclear factor kappa-light-chain-enhancer of activated B cells)
Regulatory motifs, transcription factors
Set of upstream regions in genes in the genome, each region containing one NF-kB sites Suppose we do not know either location or sequence of NF-kB sites

21. Difference B/W EMBL and DDJB? 5

EMBL and DDBJ:
• European Molecular biology Lab (**EMBL** established1980) was established in Europe.

DNA databank of Japan **(DDBJ)** established in Mishima japan (1984).

Genebank, DDBJ and EMBL joined together in **I**nternational **N**ucleotide **S**equence **D**atabase **C**ollaboration (INSDC)
You can see in this diagram, the **NCBI** (**N**ational **C**enter for **B**iotechnology **I**nformation), **DDBJ**(**D**NA **D**atabank of **J**apan) and **EBI** (**E**uropean**B**ioinformatics**I**nstitute) /**ENA** (**E**uropean **N**ucleotide **A**rchive)forms an International collaboration known as **INSDC** (**I**nternational **N**ucleotide **S**equence **D**atabase **C**ollaboration).
Where **EMBL**established**EBI**, to deal with Bioinformatics kind of stuff and within them they have established **ENA** to maintain the DNA sequence datasets

22. Lines and sine, chromatin and heterochromatin? 2

SINES (Short Interspersed Nuclear Elements – 80-300 bp – Alu repeats

• **LINES** (Long Interspersed Nuclear Elements) – 6-8 KB long.

Chromatin is a mass of genetic material composed of DNA and proteins that condense to form chromosomes during eukaryotic cell division. Chromatin is locate d in the nucleus of our cells.

Heterochromatin is a tightly packed form of DNA or condensed DNA, which comes in multiple varieties. These varieties lie on a continuum between the two extremes of constitutive heterochromatin and facultative heterochromatin. Both play a role in the expression of genes.

23. Brute Force Algorithm

Largest distance in "L"
Outermost points of "X"
Remaining distance "δ"
$L = \{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$
Size of L is $(n2) = n(n2-1) = 10$ where is "n" is number of points in the solutions. Here "n" is 5 and positions of 'X' as $x_1 = 0$, x_2, x_3, x_4 and x_5.

Applications of Bioinformatics

1. Drug Development
2. Crop Improvement
3. Microbial Genome
4. Gene Therapy
5. Biotechnology
6. Comparative Study
7. Evolutionary Studies
8. Veterinary Science
9. Molecular Medicine
10. Personalized Medicine
11. Preventive Medicine
12. Waste Cleanup
13. Antibiotic Resistance
14. Alternate Energy Science
15. Insect Resistance
16. Climate Change Studies
17. Nutritional Quality

24. Why Sequencing are submit in database? Write precaution of sequencing?

Sequences are submitted to the databases in order to share them with the scientific community Generally sequences are submitted at the time of publication and are reviewed by peers.

Caution: It is important to ensure that sequence files do not contain any special characters • Control characters in addition to standard ASCII characters must be removed else they might mess up the analysis or data transfer.

25. Translation of protein?

Protein synthesis is accomplished through a process called translation. After DNA is transcribed into a messenger RNA (mRNA) molecule during transcription, the mRNA must be translated to produce a protein. In translation, mRNA along with transfer RNA (tRNA) and ribosomes work together to produce proteins.

26. What are MITES?

MITES (Miniature Inverted repeat TEs) – Features of both Class I and II – 400 bp. **(PPTS)**

27. Genequiz and MAGPIE?

MAGPIE (Multipurpose Automated Genome Project Investigation Environment) It's an automated genome analysis tool that is used for structural annotation.

GeneQuiz Focuses on deriving a predicted protein function based upon available evidence, including evaluation of similarity to the closest homologue in database.

28. Restriction Enzyme?

A restriction enzyme is a protein that recognizes a specific, short nucleotide sequence and cuts the DNA only at that specific site, which is known as restriction site or target sequence. More than 400 restriction enzymes have been isolated from the bacteria that manufacture them. **(Intenet)**

29. Exon Changing?

The weight of the path is defined as optimal alignment score between concatenated labels (blocks) and target sequence. ... Exon Chaining assumes the positions and weights of exons are pre-defined. **(Internet)**

Five weighted intervals, (2, 3, 3), (4, 8, 6), (9, 10, 1), (11, 15, 7), and (16, 18, 4), shown by bold edges, form an optimal solution to the Exon Chaining problem. The array at the bottom shows the values s1 , s2 , . . . , s2n generated by the EXONCHANING algorithm. **(PPTS)**

30. Three Steps of Genome annotation?

Genome annotation is an important first step in **genome analysis**, based on; • Finding significant alignment to sequences of known function in databases. There are two steps in genome annotation • Structural annotation • Functional annotation.

31. Structural Annotation

Identification of gene features like; • Promoters • Terminators • Shine-Dalgarno sites • DNA motifs • Co-transcription units • Operons in microbes.

32. Note on ORF?

Open reading frames (ORFs) Fenome of length n as a sequence of n/3 codons The three stopcodons, (TAA, TAG, and TGA) The subsegments of these that start from a start codon,

ATG, are ORFs. ORFs within a single genomic sequence may overlap since there are six possible readiŶg frawes:three o Ŷ one strand starting at positions 1, 2, and 3, and three on the reverse strand. (In molecular genetics, an open reading frame (ORF) is the part of a reading frame

that has the potential to be translated. An ORF is a continuous stretch of codons that contain a start codon (usually AUG) and a stop codon (usually UAA, UAG or UGA).[1] An ATG codon within the ORF (not necessarily the first) may indicate where translation starts. The transcription termination site is located after the ORF, beyond the translation stop codon. If transcription were to cease before the stop codon, an incomplete protein would be made during translation.[2] In eukaryotic genes with multiple exons, ORFs span intron/exon regions, which may be spliced together after transcription of the ORF to yield the final mRNA for protein translation.)

33. Define Structure of annotation? Structural annotation? What we can do by using this technique?

Structural Annotation Identification of gene features like; • Promoters • Terminators • ShineDalgarno sites • DNA motifs • Co -transcription units • Operons in microbes. MAGPIE (Multipurpose Automated Genome

Project Investigation Environment) It's an automated genome analysis tool that is used for structural annotation. **(PPTS)**

Structural annotations are physical regions of a genome that encode a genomic feature. Examples of such annotations are genes, mRNA, transcript, repeat sequences, **(Internet)**

34. Brute runtime algorithm?

In actual cases, the performance (Runtime) of an algorithm depends on n, that is the size of the input or the number of operations is required for each input item. Runtime grows quicker than previous all based on n. Runtime grows even faster than polynomial algorithm based on n.

35. Collinear nucleotide?

Charles Yanofsky proved that a gene and its protein product are collinear. Yanofsky's experiment was so influential that nobody even questioned about codons and for almost 2 decades biologists believed that a protein was encoded by a long string of adjacent triplets.

36. Pre Order Algorithm?

PREORDER(v)

1. output v

2. if v has children

3. PREORDER(left child of v)

4. PREORDER(right child of v)

37. Swiss prot fasta sequence and genebank formate

of fragments. MS2 genome • 49 different codons in the genetic code specify the sequence of the 129 aminoacids long coat polypeptide

SWISS-PROT: 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.

FASTA Sequence Format • DNA sequences are stored in specified formats • Different softwares need sequences in different formats. FASTA is the most frequently used format to present DNA and Protein sequences • It is recommended that all lines of text be shorter than 80 characters in leng th

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

GATGGGATTGGGGTTTTCCCCTCCCATGTGCT CATCTAGAGCCACCGTCCAGGGAGCAGGTAGC TGCTGGGCTCTCCACGACGGTGACACGC--------

Genebank Format • A sequence file in GenBank format can contain several sequences • One sequence starts with a line containing the word LOCUS and a number of annotation lines. The start of the sequence is marked by a line containing "ORIGIN" and the end of the sequence is marked by two slashes ("//")

Example: ORIGIN 1 aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc 61 tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg 121 ccccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc 181 tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc //

Brute force approach-find all local similarities Each substring from the genomic sequence that exhibits sufficient similarity to the target protein could be considered a putative exon Exon-flanking dinucleotides AG and GT Overlapping.

KMP algorithm) searches for occurrences of a "word" W within a main "text string" S by employing the observation that when a mismatch occurs, the word itself embodies sufficient information to determine where the next match could begin

European Molecular biology Lab (EMBL established1980): The European Molecular Biology Laboratory 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.

DNA databank of Japan (DDBJ) established in Mishima japan (1984): The DNA Data Bank of Japan is a biological database that collects DNA sequences. It is located at the National Institute of Genetics in the Shizuoka prefecture of Japan. It is also a member of the International Nucleotide Sequence Database Collaboration or INSDC.

The columns of the alignment containing one space are called **indels**, with the columns containing a space in the top row called **insertions** and the columns with a space in the bottom row **deletions**.

A **splice site mutation** is a genetic mutation that inserts, deletes or changes a number of nucleotides in the specific site at which splicing takes place during the processing of precursor messenger RNA into mature messenger RNA. Splice site consensus sequences that drive exon recognition are located at the very termini of introns. The deletion of the splicing site results in one or more introns remaining in mature mRNA and may lead to the production of abnormal proteins.

38. What is personalized medicine?

Personalized medicine, precision medicine, or theranostics is a medical model that separates people into different groups—with medical decisions, practices, interventions and/or products being tailored to the individual patient based on their predicted response or risk of disease.

39. Describe region of DNA?

DNA. The vast majority of organisms encode their genes in long strands of DNA (deoxyribonucleic acid). DNA consists of a chain made from four types of nucleotide subunits, each composed of: a five-carbon sugar (2-deoxyribose), a phosphate group, and one of the four bases adenine, cytosine, guanine, and thymine.

40. Difference B/W Gene Bank and Fasta format?

GenBank to **FASTA** accepts a **GenBank file** as input and returns the entire DNA sequence in **FASTA format**. Use this program when you wish to quickly remove all of the non-DNA sequence information from a **GenBank file**. Paste the contents of one or more **GenBank files** into the text area below.

41. Write note on Swiss Port? 2

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Recent developments of the database include format and content enhancements, cross-references to additional databases, new documentation files and improvements to TrEMBL, a computer-annotated supplement to SWISS-PROT. TrEMBL consists of entries in SWISSPROT-like format derived from the translation of all coding sequences (CDSs) in the EMBL Nucleotide Sequence Database, except the CDSs already included in SWISS-PROT. We also describe the Human Proteomics Initiative (HPI), a major project to annotate all known human sequences according to the quality standards of SWISS-PROT. SWISSPROT is available at: http://www.expasy.ch/sprot/ and http://www.ebi.ac.uk/swissprot/

42. Difference B/W mouse synteny and human X chromosome? 3

Synteny between species means not only that orthologous (functionality and ancestry identical). Genes are resent but that they are present in the same order on the genome, thus indicating common 4pc uence of the mouse genome and its comparison with the previously sequenced human genome reveal that 90.2% of the human genome and 93.3% of the mouse genome lie in the conserved egment. So not only we can find the homologous sequence between the two genomes; the gene Ily lie on the chromosome in the same order. This is powerful evidence for common ancestry.

43. What are primary medical database? Enlist four medical database?

NIVEL Primary Care Database collects data that is routinely recorded in the health care provider's electronic health record system. This includes data on health problems and treatment. **(Database are not found)**

44. What is long common sequence?

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.

45. Write note on Intron, Exon and Splicing process?

Introns: intron is a long stretch of noncoding DNA found between exons (or coding regions) in a gene. Genes that contain introns are known as discontinuous or split genes as the coding regions are not continuous. Introns are found only in eukaryotic organisms.

Exons: An exon is any part of a gene that will encode a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing. The term exon refers to both the DNA sequence within a gene and to the corresponding sequence in RNA transcripts.

Splicing Process: RNA splicing is a process that removes introns and joins exons in a primary transcript. An intron usually contains a clear signal for splicing (e.g., the beta globin gene). In some cases (e.g., the Tau gene), a splicing signal may be masked by a regulatory protein, resulting in alternative splicing.

46. What is alignment how we calculate in give example?

Alignment and Alignment Score. (Here "alignment score" is "matching score" used in the definition of the sequence identity in problemE.html.) Remark: The program scoreout.c is designed for computing the alignment score of a given pair of amino acid sequences.

47. Type of complexity? 2

Three types of complexity could be considered when analyzing algorithm performance. These are worst -case complexity, best-case complexity, and average-case complexity. ... Average-Case running time for a function, f(n) such that where a,b and c are constants can be described as an average between best-case and worst-case.

48. Exon Predication?

GeneAlign is a coding exon prediction tool for predicting protein coding genes by measuring the homologies between a sequence of a genome and related sequences, which have been annotated, of other genomes. ... The rates of missing exons and wrong exons are smaller than 1%.

49. Synteny its Important and Example?

Additionally, exceptional conservation of synteny can reflect important functional relationships between genes. ... Synteny is widely used in studying complex genomes, as comparative genomics allows the presence and possibly functio n of genes in a simpler, model organism to infer those in a more complex one.

For example, many of the genes of humans are syntenic with those of other mammals—not only apes but also cows, mice, and so on. Study of synteny can show how the genome is cut and pasted in the course of evolution. **(Internet)**

50. List of method finding pattern of genome?

Extac Searching Method, Approximate Searching Method, Position Weight Matrices, Suffix Tree.

51. Difference B/W DNA and RNA?

DNA is a double- stranded molecule while RNA is a single stranded molecule. DNA is stable under alkaline conditions while RNA is not stable. ... DNA and RNA base pairing is slightly different since DNA uses the bases adenine, thymine, cytosine, and guanine; RNA uses adenine, uracil, cytosine, and guanine.

52. Edit Distance and Example?

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, and substitution 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

53. Difference B/W Clustering by subgroup and Cluster by linkage?

Clustering by subgraph • Each sequence is a vertex • Significant alignment score is an edge • Trimming by removing weak edges (High P/E)

Clustering by Linkage • Each sequence is a vertex • Significant alignment score is an edge • Trimming by removing weak edges (High P/E) • Or remove > e-6 • Remaining subgraph should share 2/3rd of edges.

54. Similarities approach of gene prediction?

A similarity-based approachPreviously sequenced genes Unknown genes in newly sequenced DNA fragments Combinatorial puzzle-find a set of substrings (candidate exons) whose splicing best fits the target. Brute force approach-find all local similarities Each substring from the genomic sequence that exhibits sufficient similarity to the target protein could be considered a putative exon Exon-flanking dinucleotides AG and GT Overlapping. Model a putative exon with a weighted interval in the genomic sequence, parameters (l, r, w) l is the left-hand position, r is the right-hand position, and w is the weight of the putative exon $\overline{w}$-local alignment score Likelihood that this interval is an exon A chain is any set of non overlapping weighted intervals. Total weight of a chain A maximum chain Weights of all intervals are positive (w > 0). Model a putative exon with a weighted interval in the genomic sequence

55. Write note on Waardenburg Syndrom?

Hearing loss Two Different colored eyes Gene present on chromosome 2 Splotch gene in mice, Human genomemouse genome Cut into 300 genomic fragments-Synteny blocks Chromosome 2 in humansmouse chromosomes 1, 2, 3, 5, 6, 7, 10, 11, 12, 14, and 17, Genome rearrangement results in a change of gene ordering Analysis of human and mouse genomes-250 genomic rearrangements. **(PPTS)**

Waardenburg syndrome is a group of genetic conditions that can cause hearing loss and changes in coloring (pigmentation) of the hair, skin, and eyes. Although most people with Waardenburg syndrome have normal hearing, moderate to profound hearing loss can occur in one or both ears. **(Internet)**

56. Human Genome Project?

Pilot project, the Human Genome Initiative begun by Department Of Energy (DOE) in 1986. National Human Genome Research Initiative (NHGRI), a federally funded organization (NIH) started in 1988 (Francis Collin) • Celera, a commercial vendor (Craig Venter) joined the venture in 1998. • Both simultaneously announced the completion of the project in 2000 • Total 3.4 billion bases sequenced at a cost of $1/base. **(PPTS)**

The Human Genome Project 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. **(Internet)**

57. Edit Graph?

The grid that is achieved after alignment is similar to the Manhattan grid where each entry in the grid looks like a city block The graph called Edit Graph The main difference between the Manhattan and Edit Graph is that we can move along the diagonals in the Edit Graph.

58. Enzyme commission numbers with example

Enzyme Commission (EC) numbers:

It is another scheme which was put forward by the **E**nzyme **C**ommission (EC) that was working under the **IUBMB (International Union for Biochemistry and Molecular Biology).** They say that the enzymes are classified on the basis of the reactions they catalyze and have a 4-digit scheme which is actually the enzyme commission number: **EC a.b.c.d 'a' (**first digit**)** informs that it is from one of the 6 classes of biochemical reactions (enzyme might be coming from one of these classes). **'b' (**second digit**)** informs that is from the group of substrate (the thing on which the enzyme attacks)**. 'c' (**third digit**)** informs us that it is anaccepter molecule. **'d' (**fourth digit**)** gives thedetails of biochemical reaction

For example,

tripeptideaminopeptidases

EC 3.4.11.4

Where **3** – tells us that it is a Hydrolase (use water to break substrate).

This **3.4-** tell us thatit is aHydrolase that acts on the peptide bonds.

The **3.4.11-** tells us that it is a Hydrolase that cleaves the amino terminal amino acids of polypeptide.

While putting everything together**, EC 3.4.11.4-** it tells us that it is a Hydrolase that cleaves the amino terminal amino acids of a tri-peptide

59. General algorithm 3

Hierarchical agglomerative general algorithm

• Find the 2 closest objects and merge them into a cluster

• Find and merge the next two closest points, where a point is either an individual object or a cluster of objects

• If more than one cluster remains, return to step 2

60. Six model of organisms ?marks 3

• Model organisms:

Most of the times, while we are doing those genome sequencing projects, our objective is to find the cure of some disease, or improving some variety of the crop for enhancing its production, or looking into some drugs against different organisms so it's a good idea to have some model organisms that can be used for studying various processes in labs and there are is a range of model organisms which includes:

1. E. coli – bacteria

2. S. cerevisiae – yeast

3. C. elegans – worm

4. D.melanogaster – fly

5. Daniorerio – zebrafish

6. Musmusculus - mouse

7. Homo sapiens – you and me

8. Arabidopsis - plant

61. self comparison ?

➢ ***Self-comparison of proteome*** (sometimes we are interested in finding genes which are kind of duplicated within the same organism, so in order to do that this self-comparison of proteome is made, where proteome is the collection of the proteins which are derived from those genomes. Therefore, the whole collection of one organism's proteins can be termed as proteome and we can compare it with itself and can find about those sequences which are being duplicated in it).

BEST OF LUCK