

1- 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.

2- Delete the one, two and three nucleotide from strand with example? ??????

3- 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.

4- What are MITES?

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

5- 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.

6- Genome and Genome informatics?

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 predicted 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.

7- 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)**

8- 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 $s_1, s_2, \dots, s_{2n}$ generated by the EXONCHANNING algorithm. **(PPTS)**

9- 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.

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

11- Note on ORF?

Open reading frames (ORFs) Fenome of length n as a sequence of $n/3$ codons The three stop codons, (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 reading frames; three on 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.)

12- Define Structure of annotation? Structural annotation? What we can do by using this

technique? Structural Annotation Identification of gene features like; • Promoters • Terminators • Shine-Dalgarno 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)**

13- What Kind of Gene Expression OMNIBUS? Kind of Records 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

14- EM BK note?

?????????

15- 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.

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.

16- 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

17- 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)

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

19- KMP Matcher?

Knuth Morris Pratt (KMP) is an algorithm, which checks the characters from left to right. When a pattern has a sub-pattern 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 $O(n)$. **(Internet)**

20- Edit the Manhattan Graph?

21- Application of Bioinformatics? Five application of bioinformatics?

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

SOUHLA-FATIMA

22- Swiss Port, Fasta, Gene Bank, ORF?

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 length

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 ccgagtcgg gtccttggg cccaacctcc catcgtgtc 61 tattgtaccc tgggtctcg
gcgggccgc cgcttgcgg ccgcccggg ggcgccttg 121 cccccgggc ccgtgccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc 181
tgagttgatt gaatgcaatc agttaaact ttcaacaatg gatctcttgg ttccggc //

23- Bypass Algorithm?

NEXTVERTEX Algorithm BYPASS Algorithm Skip the subtree rooted at vertex (a, i) Increment ai (unless ai = k BYPASS(a, i, L, k)

1. for j i to l

2. if $a_j < k$

3. a_j

a_j

+ l

4. return (a, j)

5. return (a, 0) (**Proper Answer PPTS mn dykhain mujy ye sahi trha samj nh aaya**)

24- Brute Force?

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.

25- KMP algorithm?

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

26- Why we cannot use brute force to approach motif? Why Brute Force not use for motif finding?
?????????

27- Six model of organism?

E.Coli- Bacteria S.Cerevisiae- Yeast

C.elegans- Worm,

D-Melangaster- Fly

Daniorerio- Zebrafish Musmusculus-Mouse

Homo sapien- Human

Arabidopsis- Plant

28- Difference B/W EMBL and DDJB?

European Molecular biology Lab (EMBL established 1980): 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.

29- Difference Indel. Insertion and Deletion in an alignment?

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**.

30- Note on splice signal?

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](#).

SOUHLA-FATIMA

31- Lines and sine, chromatin and heterochromatin?

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 located 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.

32- 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.

33- 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.

34- Difference local alignment input and output?

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 among all global alignment of all substrings of v and w .

35- 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.

36- Five task of Genome analysis?

37- How mutation in DNA?

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

38- Write note on Swiss Port?

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 SWISS-PROT-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. SWISS-PROT is available at: <http://www.expasy.ch/sprot/> and <http://www.ebi.ac.uk/swissprot/>

39- What are suffix?

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.

40- Write note on transposable elements?

41- Difference B/W mouse synten and human X chromosome?

Synten between species means not only that orthologous (functionality and ancestry identical). Genes are present 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 segment. So not only we can find the homologous sequence between the two genomes; the gene lly lie on the chromosome in the same order. This is powerful evidence for common ancestry.

42- 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)**

43- What is long common sequence?

SOUHLA-FATIMA

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 .

44- 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.

45- Shot on KE-KB?

?????????

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- 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)**

48- Type of complexity?

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. **(Internet)**

49- Note on ETS?

????????/

50- 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%.

51- 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 function 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)**

52- List of method finding pattern of genome?

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

53- Generalized algorithm for pattern finding?

Goal: Find all occurace of a pattern in text.

Input:

Pattern $p=[p_1 \dots p_n]$ of length n

Text $t=[t_1 \dots t_m]$ of length m

Output:

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

54- 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.

55- 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

Edit Distance

A T A T A T A T -
- T A T A T A T A

align the $(i+1)$ -st letter in ATATATAT against the i th letter in TATATATA for $1 \leq i \leq 7$

1 2 3 4 5 6 7 8
A T A T A T A T
- T A T A - A T
1 2 3 4 5 6

56- 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.

57- What are hammering score and how calculate it?

???????/

58- Similarities approach of gene prediction?

A similarity-based approach Previously 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 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

59- Write note on Waardenburg Syndrom?

Hearing loss Two Different colored eyes Gene present on chromosome 2 Splotch gene in mice, Human genome-mouse genome Cut into 300 genomic fragments-Syteny 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)**

60- 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)**

61- 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.

SOUHLA-FATIMA