BIF501 Mid Term Solve
Subjective

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.

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.

What are MITES?

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

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.

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.

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)**

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)**

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.

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

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 readiŶg fraŵes: three oŶ 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.)

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)**

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

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.

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)

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

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)**

Application of Bioinformatics? Five application of bioinformatics?

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

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|56881558l: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 //

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 aj < k
3. aj aj
+ 1
4. return (a, j)
5. return (a, 0) **(Proper Answer PPTS mn dykhain mujy ye sahi trha samj nh aaya)**

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.

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

Difference B/W EMBL and DDJB?

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.

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**.

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.

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.

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.

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

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 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/

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.

Difference B/W mouse synteny and human X chromosome?

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 lly lie on the chromosome in the same order. This is powerful evidence for common ancestry.

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)**

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.

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.

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.

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)**

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)**

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%.

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)**

List of method finding pattern of genome?

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

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:

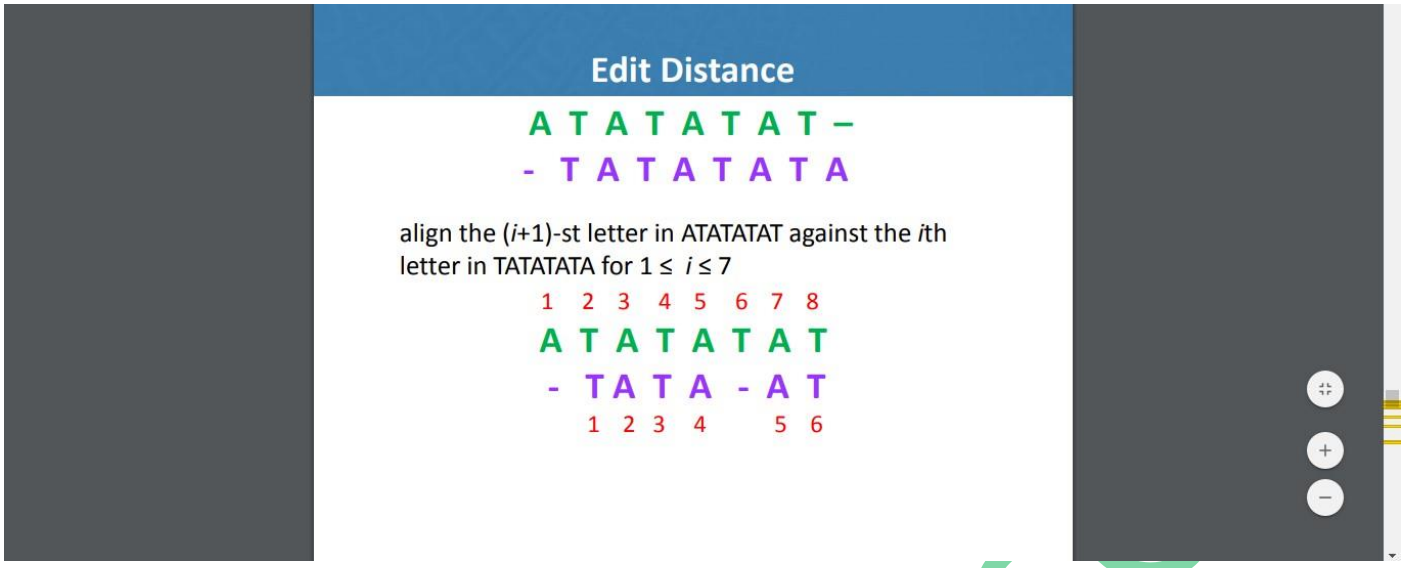
An indication that pattern P exist in T, or it does not exist in T.

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.

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



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.

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 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

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)**

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)**

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.

Write detail note on project initiated for Haemophilus?

The project initiated at The Institute of Genome Research (TIGR) under leadership of Craig Ventor.

What is LCS?

The longest common subsequence (LCS) problem is the problem of finding the longest subsequence common to all sequences in a set of sequences (often just two sequences).

What are solution for exon prediction?

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.

Local alignment input output with example?

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.

Conserved DNA?

conserved sequences are identical or similar sequences in nucleic acids (DNA and RNA) or proteins across species (orthologous sequences), or within a genome (paralogous sequences), or between donor and receptor taxa (xenologous sequences). **(Internet)**

Proteome and Proteomer core?

Core Proteome • All-against-all comparison provides an indication of Gene/Protein families • Unique set of proteins is core proteome.

Proteomer core:

??????????

Describe approximate approach algorithm?

The approximation ratio (or approximation factor) of an algorithm is the ratio between the result obtained by the algorithm and the optimal cost or profit. ... An algorithm with approximation ratio k is called a k-approximation algorithm; both algorithms above would be called 2-approximation algorithms. **(Internet)**

Cluster analysis?

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

Global sequence problem?

The dynamic programming solves the original **problem** by dividing the **problem** into smaller independent sub **problems**. These techniques are used in many different aspects of computer science. The algorithm explains **global sequence** alignment for aligning nucleotide or protein **sequences**.

Transcription of DNA? Which enzyme is involved in this process?

RNA polymerase enzymes

Transcription. Transcription is the process by which DNA is copied (transcribed) to mRNA, which carries the information needed for protein synthesis. Transcription takes place in two broad steps. First, pre-messenger RNA is formed, with the involvement of RNA polymerase enzymes.

What do you mean by gene split?

split genes. genes in which the genomic sequences are interrupted by intervening sequences (introns) that are spliced out of the mRNA before translation.

Definition of match mismatch, deletion, insertion?

Columns that contain the same letter in both rows are called **matches**, while columns containing different letters are called **mismatches**. the columns containing a space in the top row called **insertions** and the columns with a space in the bottom row **deletions**.

Proteome advantage?

is important because proteins represent the actual functional molecules in the cell. When mutations occur in the DNA, it is the proteins that are ultimately affected. Drugs, when they have beneficial effects, do so by interacting with proteins.

Difference between prokaryote and eukaryotes genome?

The prokaryotic genome is almost entirely made up of coding DNA; no introns! ... Genes that serve a related function are often clustered closely together on the genome. Prokaryotes, mostly bacteria, also often have short, usually circular, double-stranded segments of DNA called plasmids.

DNA mutation?

A mutation is a change in DNA, the hereditary material of life. An organism's DNA affects how it looks, how it behaves, and its physiology. So a change in an organism's DNA can cause changes in all aspects of its life. Mutations are essential to evolution; they are the raw material of genetic variation.

Transposible element?

A **transposable element** (TE, **transposon**, or jumping gene) is a DNA sequence that can change its position within a genome, sometimes creating or reversing mutations and altering the cell's genetic identity and genome size. Transposition often results in duplication of the same genetic material.

Array and variable with example?

An **array** is a group (or collection) of same data types. For **example** an int **array** holds the elements of int types while a float **array** holds the elements of float types.

Start and Stop codon?

Genetic code table. Each three-letter sequence of mRNA nucleotides corresponds to a specific amino acid, or to a **stop codon**. UGA, UAA, and UAG are **stop codons**. AUG is the **codon** for methionine, and is also the **start codon**.

Edit distance?

In computational linguistics and computer science,**edit distance** is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other.

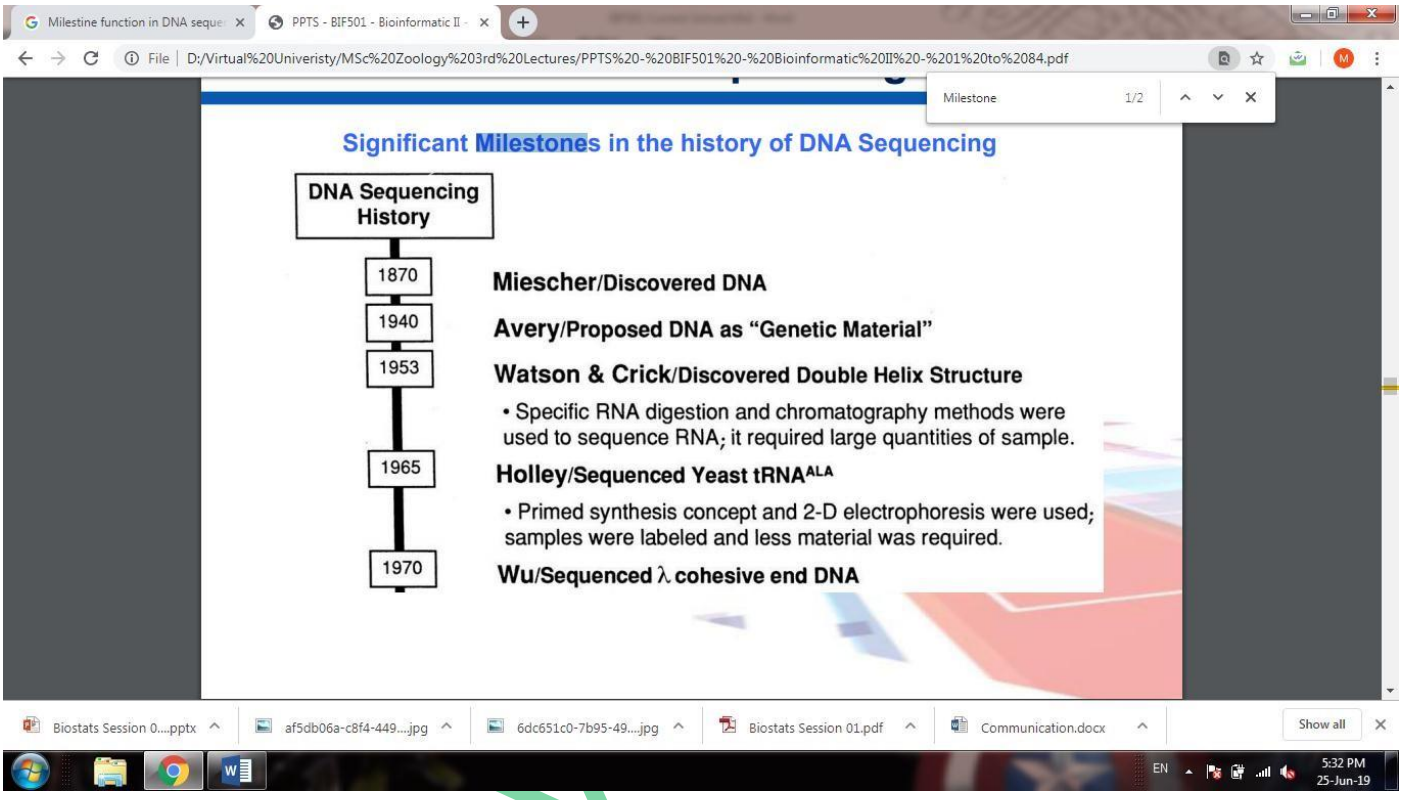
Frame Hexamer Brute force approach?

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

Classes of biological database?

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

Milestone function in DNA sequence?



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.

Euchromatin and heterochromatin differences?

The tightly packed form of DNA in the chromosome is called as **heterochromatin**, while the loosely packed form of DNA in the chromosome is called as **euchromatin**. In**heterochromatin**, the density of DNA is high and are stained dark, whereas in **euchromatin** the density of DNA is little and are lightly stained.Jul 1, 2017

Local alignment problem?

The **Sequence Alignment problem** is one of the fundamental **problems** of Biological Sciences, aimed at finding the similarity of two amino-acid sequences.

Enzyme commission numbers with example?

Every **enzyme** code consists of the letters "EC" followed by four **numbers** separated by periods. ... For **example**, the tripeptide aminopeptidases have the code "EC 3.4.11.4", whose components indicate the following groups of **enzymes**: EC 3 **enzymes** are hydrolases (**enzymes** that use water to break up some other molecule)

Note on splice signal?

The term intron refers to both the DNA sequence within a gene and the corresponding sequence in the unprocessed RNA transcript. ... The **splice** acceptor**site** at the 3' end of the intron terminates the intron with an almost invariant AG sequence.

Shot on KE-KB?

?????????

KMP algorithm?

Why we cannot use beret force to approach motif? Why Brute Force not use for motif finding?

Four cases of pattern finding?

EM BK note?

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

Note on ETS?

What do you know about Enter2?

Fibrositis?

Describe four approaches finding gene expression?

Step of pattern finding fasta?

Collinear arrangement of nucleotides?

Bypass algorithm its advantages?

Five task of Genome analysis?