

VU Learning HUB

BIF501 Quiz File

Chapter-wise Solved MCQs

vulearninghub.xyz

Lesson 1: Applications of Bioinformatics & Biological Databases

1. Which organism is cited in the handout as an alternative energy source due to its capacity for generating energy from light?

- A) Enterococcus faecalis
- B) Chlorobium tepidum
- C) Bacillus thuringiensis
- D) Saccharomyces cerevisiae

2. What are the three primary goals of biological databases?

- A) Data encryption, data compression, and data deletion
- B) Data storage, information retrieval, and knowledge discovery
- C) Sequence alignment, structure rendering, and gene editing
- D) Sample collection, lab automation, and clinical trials

3. According to the source, how frequently does the number of bases in GenBank double?

- A) Every 6 months
- B) Every 12 months
- C) Every 18 months
- D) Every 24 months

4. The International Nucleotide Sequence Database Collaboration (INSDC) is a joint venture of which three repositories?

- A) WHO, CDC, and NIH
- B) NCBI, EMBL, and DDBJ
- C) PIR, EBI, and SIB
- D) TAIR, SGD, and AceDB

Answer Key

1: B — Chlorobium tepidum is highlighted in the handout for its capacity to generate energy from light as an alternative energy source

2: B — Biological databases store biological data with the main goals of data storage, information retrieval, and knowledge discovery

3: C — The growth of bases in GenBank is exponential, doubling every 18 months

4: B — INSDC is an international joint venture formed by NCBI, EMBL, and DDBJ

Lesson 2: Sequence Submission, Retrieval, & Formats

1. UniProt is an international partnership formed between which three organizations?

- A) NCBI, EMBL, and DDBJ
- B) PIR, EBI, and SIB
- C) MIT, Berkeley, and MRC
- D) WHO, FAO, and UNESCO

2. Which delimiter line in GenBank format specifies the termination/end of a record?

- A) LOCUS
- B) ORIGIN
- C) //
- D) END

3. Which software program interconverts DNA or protein sequence files between different formats?

- A) BLAST
- B) READSEQ
- C) MAGPIE
- D) TopHat

4. Which network protocol is highlighted as the standard protocol for transferring biological data files over the internet?

- A) HTTP
- B) FTP
- C) SMTP
- D) SSH

Answer Key

- 1: B — UniProt is maintained as an international partnership between PIR, EBI, and SIB
- 2: C — In GenBank format entries, // represents the termination line marking the end of a sequence entry
- 3: B — READSEQ recognizes DNA or protein sequence files and interconverts them between different formats
- 4: B — File Transfer Protocol (FTP) is the standard network protocol used to transfer files via command line or client applications

Lesson 3: Genome Informatics & Transposable Elements

1. What are Insertion Sequences (IS elements) defined as?

- A) Complex eukaryotic retrotransposons containing long terminal repeats
- B) The simplest transposable elements in prokaryotes coding only for transposition
- C) Long non-coding RNA molecules regulating mRNA splicing
- D) Structural annotations that identify ribosomal binding sites

2. Which elements belong to the Non-LTR category of Class I transposable elements?

- A) IS elements and plasmids
- B) LINEs and SINEs
- C) Exons and introns
- D) Promoters and terminators

3. In prokaryotic structural gene annotation, what is the biological function of the Shine-Dalgarno site?

- A) Termination of RNA transcription
- B) Ribosomal binding site during protein synthesis
- C) Splicing donor recognition site
- D) Reverse transcription initiation site

4. DNA sequencing began in the 1970s with the development of which two original techniques?

- A) Maxam-Gilbert method and Sanger sequencing
- B) Illumina HiSeq and PacBio sequencing
- C) Edman degradation and Mass Spectrometry
- D) Chou-Fasman algorithm and BLAST

Answer Key

- 1: B — Insertion Sequences (IS elements) are the simplest transposable elements found in prokaryotes and code strictly for transposition
- 2: B — Within Class I retrotransposons, Non-LTR retrotransposons are divided into LINEs and SINEs
- 3: B — Shine-Dalgarno sites function as ribosomal binding sites during protein synthesis in structural annotation
- 4: A — DNA sequencing originated in the 1970s using Maxam-Gilbert's chemical method and Sanger's dideoxy sequencing

Lesson 4: Molecular Biology Fundamentals & Central Dogma

1. Which of the following set of codons functions as standard stop codons?

- A) ATG, ACG, AGG
- B) TAA, TAG, TGA
- C) AAC, AAG, AAT
- D) TAT, TAC, TCG

2. How many total reading frames must be evaluated when searching for Open Reading Frames (ORFs) on double-stranded DNA?

- A) 2
- B) 3
- C) 4
- D) 6

3. Which segments of a gene are transcribed into mRNA but are NOT translated into protein?

- A) Exons 1 and 2
- B) 5'-UTR and 3'-UTR
- C) Promoter and Enhancer
- D) ORF and Coding Region

4. Which of the following base pairs represents a non-canonical "wobble" base pair in RNA?

- A) A-U
- B) G-C
- C) G-U
- D) C-G

Answer Key

- 1: B — The three standard stop codons in the genetic code are TAA, TAG, and TGA
- 2: D — DNA has six possible reading frames: three on the forward strand (+1, +2, +3) and three on the reverse strand (-1, -2, -3)
- 3: B — Both 5'-UTR and 3'-UTR regions are transcribed into mRNA but remain untranslated during protein synthesis
- 4: C — In RNA secondary structure, G-U represents a wobble base pair, whereas A-U and G-C are canonical base pairs

Lesson 5: Dynamic Programming & Restriction Mapping

1. In the Block Game dynamic programming problem (FASTBLOCK), what is the output when both grid dimensions n and m are even?

- A) Win (W)
- B) Loss (L)
- C) Draw (D)
- D) Undefined (U)

2. If a restriction map X contains n cut sites, how many pairwise distances are present in its distance multiset?

- A) $n-1$
- B) $2n$
- C) $n(n-1)/2$
- D) n squared

3. In the Partial Digest Algorithm (PARTIALDIGEST), what value is initially assigned to width?

- A) Minimum element in L
- B) Average of elements in L
- C) Maximum element in L
- D) Total count of elements in L

4. What search strategy is implemented by the PLACE(L, X) subroutine in the Partial Digest Algorithm?

- A) Greedy search
- B) Recursive backtracking
- C) Genetic algorithm
- D) Machine learning classification

Answer Key

- 1: B — The FASTBLOCK algorithm returns Loss (L) if n and m are both even; otherwise, it returns Win (W)
- 2: C — A set of n points yields $n(n-1)/2$ pairwise distances in the multiset
- 3: C — PARTIALDIGEST initializes width as the maximum element present in the distance multiset L
- 4: B — PLACE uses a recursive backtracking search to test candidate points y or width-y and backtracks if constraints fail

Lesson 6: DNA Motifs, Profiles, & Search Trees

1. How is Hamming distance $dH(v,w)$ defined between two l-mers v and w?

- A) Total number of identical nucleotides between v and w
- B) Number of positions at which the characters in v and w differ
- C) The ratio of GC content between two sequences
- D) The length of the longest common substring

2. What is the mathematical relationship between total Hamming distance and consensus score $\text{Score}(s, \text{DNA})$ for t sequences of length l?

- A) $dH = l * t + \text{Score}(s, \text{DNA})$
- B) $dH = l * t - \text{Score}(s, \text{DNA})$
- C) $dH = l * t / \text{Score}(s, \text{DNA})$
- D) $dH = (l-t) * \text{Score}(s, \text{DNA})$

3. In Branch and Bound motif search algorithms, what action is taken when the optimistic score for a branch is strictly LESS than bestScore?

- A) Explore all leaf nodes of the subtree
- B) Terminate the program execution
- C) Call the BYPASS algorithm to skip the subtree
- D) Reset bestScore to zero

4. NF-kB binding sites (TCGGGGATTCC) are cited as an example of which biological feature?

- A) Splice donor junctions
- B) Transcription factor regulatory motifs in DNA
- C) Transposable insertion sequences
- D) Secondary protein alpha-helices

Answer Key

- 1: B — Hamming distance counts the number of positions with differing characters between two equal-length strings
- 2: B — Total Hamming distance equals $l \cdot t - \text{Score}(s, \text{DNA})$, making consensus score maximization equivalent to distance minimization
- 3: C — If the optimistic score is lower than bestScore, BYPASS is called to skip the entire unpromising subtree
- 4: B — NF-kB binding sites are specific regulatory motifs located upstream of genes where transcription factors bind

Lesson 7: Genome Rearrangements & Sequence Alignment

1. Comparative analysis of human and mouse genomes reveals approximately how many major genomic rearrangements?

- A) 25
- B) 100
- C) 250
- D) 1000

2. What is Edit Distance between two strings defined as?

- A) The number of identical characters in matching positions
- B) The minimum number of editing operations (insertions, deletions, substitutions) needed to transform one string into another
- C) The total length of the combined sequences
- D) The number of GC pairs in the alignment

3. Which editing operations are permitted in the Longest Common Subsequence (LCS) alignment problem?

- A) Substitutions only
- B) Insertions and deletions only
- C) Transpositions and inversions only
- D) Substitutions, insertions, and deletions

4. Which algorithm modified global alignment to discover highly conserved local regions (such as homeodomains) by setting negative path scores to zero?

- A) Needleman-Wunsch algorithm
- B) Smith-Waterman algorithm
- C) Chou-Fasman algorithm
- D) Maxam-Gilbert method

Answer Key

- 1: C — Genome comparison indicates approximately 250 genomic rearrangements between human and mouse genomes
- 2: B — Edit distance measures the minimum number of insertions, deletions, and substitutions required to convert one string to another
- 3: B — LCS eliminates substitution and permits ONLY insertion and deletion operations
- 4: B — The Smith-Waterman algorithm (Local Alignment Problem) modified global alignment by resetting negative matrix values to zero

Lesson 8: Gene Prediction, ORFs, & Genome Annotation

1. What percentage of eukaryotic splice junctions follow the canonical GT-AG rule?

- A) 50.0%
- B) 75.5%
- C) 90.0%
- D) 99.24%

2. In Sequence LOGO visual representations, what does the overall height of symbols at a given position represent?

- A) Sequence read depth
- B) Information content at that position
- C) Exon length in base pairs
- D) Molecular weight of the base

3. Predicting exons for a segment between an acceptor and a donor site relies on evaluating which three scores?

- A) GC content, AT ratio, and melting temperature
- B) Coding potential score, donor score, and acceptor score
- C) E-value, Bit score, and Alignment length
- D) Hydrophobicity, charge, and mass

4. Which software platform is used during genome annotation to search predicted protein models against domain databases such as Pfam and Prosite?

- A) READSEQ
- B) InterProScan
- C) AceDB
- D) FASTBLOCK

Answer Key

- 1: D — Eukaryotic splice junctions adhere to the canonical GT-AG rule in 99.24% of instances
- 2: B — The overall height of letters at a position in a Sequence LOGO is proportional to the information content at that site
- 3: B — Exon prediction evaluates coding potential, donor site score, and acceptor site score for candidate segments
- 4: B — InterProScan annotates predicted gene models against protein domain databases including Pfam, Prosite, and PANTHER

Lesson 9: Pattern Finding Algorithms & Substitution Matrices

1. In the Naive Brute Force exact pattern matching algorithm, how far is the window shifted upon encountering a mismatch?

- A) By the length of the pattern (m)
- B) Exactly 1 position to the right
- C) To the next matching anchor
- D) 2 positions to the left

2. BLOSUM matrices were introduced by Henikoff and Henikoff in 1992 based on which dataset?

- A) Single nucleotide polymorphism (SNP) databases
- B) Aligned ungapped regions in the BLOCKS database
- C) Full-length ribosomal RNA alignments
- D) Synthetic random DNA sequences

3. Which substitution matrix is noted as being particularly suited for ungapped sequence alignments?

- A) PAM1
- B) BLOSUM62
- C) BLOSUM50
- D) PAM250

4. Which classic dynamic programming algorithm determines the optimal GLOBAL alignment spanning the entire length of two sequences?

- A) Smith-Waterman algorithm
- B) Needleman-Wunsch algorithm
- C) Chou-Fasman algorithm
- D) Partial Digest algorithm

Answer Key

- 1: B — Naive Brute Force shifts the search window by exactly 1 position to the right whenever a character mismatch occurs
- 2: B — BLOSUM matrices are derived from ungapped blocks of aligned sequence segments in the BLOCKS database
- 3: B — BLOSUM62 is optimal for ungapped alignments, whereas BLOSUM50 is suited for gapped alignments
- 4: B — The Needleman-Wunsch algorithm computes the optimal global alignment covering the full length of both sequences

Lesson 10: High-Throughput Sequencing, DNA Arrays, & Proteomics

1. What is the 3-mer composition Spectrum(s,3) for the sequence s=TATGGTGC?

- A) {TAT, ATG, TGG, GGT, GTG, TGC}
- B) {TAT, TGG, TGC}
- C) {ATG, GTG, TGG}
- D) {TAA, TGG, GCC}

2. A Universal DNA array designed to detect target l-mers contains how many total probe combinations?

- A) 2^l
- B) 4^l
- C) 4^l
- D) l^4

3. In Tandem Mass Spectrometry (MS/MS) of peptides, what designation is given to the most frequent N-terminal fragment ions?

- A) a-ions
- B) b-ions
- C) x-ions
- D) y-ions

4. Which scientist received the Nobel Prize for early protein sequencing work, including digesting insulin with proteases?

- A) Francis Crick
- B) Frederick Sanger
- C) Margaret Dayhoff
- D) Steve Fodor

Answer Key

- 1: A — The 3-mer spectrum of TATGGTGC consists of all overlapping 3-base sub-strings: TAT, ATG, TGG, GGT, GTG, and TGC
- 2: C — A universal DNA array incorporates all 4^l possible oligonucleotide probes of length l
- 3: B — In MS/MS, b-ions correspond to N-terminal partial peptide fragments, while y-ions represent C-terminal fragments
- 4: B — Frederick Sanger received the Nobel Prize for pioneering protein sequencing using insulin digestion

Lesson 11: Protein Structure Prediction & Modeling

1. In the Chou-Fasman algorithm, how many contiguous amino acid residues with $P(\alpha) > 1.0$ are required to initiate an Alpha Helix?

- A) 2
- B) 4
- C) 6
- D) 8

2. Which structure prediction strategy is recommended when target sequence identity falls into the "Twilight Zone" (<30% identity)?

- A) Homology Modeling
- B) Fold Recognition / Threading
- C) Naive Brute Force
- D) Exon Chaining

3. Ab Initio protein structure modeling relies centrally on which thermodynamic concept?

- A) Dayhoff's evolutionary mutation rate
- B) Anfinsen's thermodynamic hypothesis (minimizing free energy)
- C) Chou's propensity product
- D) Hamming's minimum distance

4. Approximately how many stable protein folds are estimated to exist in nature?

- A) 500
- B) 1,000
- C) 5,000
- D) 50,000

Answer Key

- 1: B — Chou-Fasman requires 4 contiguous residues with $P(\alpha) > 1.0$ to form an alpha-helical nucleation center
- 2: B — Fold Recognition (Threading) is used for low alignment identity (<30%) in the "Twilight Zone" where homology modeling fails
- 3: B — Ab Initio modeling is centered on Anfinsen's hypothesis, seeking the native conformation with minimum free energy
- 4: C — Structural biology indicates there are approximately 5,000 stable protein folds in nature

Lesson 12: RNA Structure & Machine Learning in Bioinformatics**1. In RNA-Seq transcript quantification, what does RPKM normalize for?**

- A) GC content and exon junction counts
- B) Transcript length (in kilobases) and total mapped reads (in millions)
- C) Polymerase speed and ribosome density
- D) Mutation rate and edit distance

2. Although humans share 99.9% of their DNA sequence, what percentage of variable sequence forms the basis for forensic DNA profiling?

- A) 0.01%
- B) 0.1%
- C) 1.0%
- D) 5.0%

3. In machine learning applications for bioinformatics, which category does K-means or Hierarchical Clustering belong to?

- A) Supervised Learning
- B) Unsupervised Learning
- C) Reinforcement Learning
- D) Semi-supervised Regression

4. In the drug discovery pipeline, an approach that begins without a pre-identified gene target and screens physiological effects in animal/organ models is termed:

- A) Target-centric approach
- B) Physiology-based (disease-centric) approach
- C) Virtual docking approach
- D) Structure-based de novo approach

Answer Key

- 1: B — RPKM normalizes raw read counts by transcript length in kilobases and total mapped reads in millions
- 2: A — Forensic DNA profiling utilizes the 0.01% of unique sequence variation that differentiates human individuals

- 3: B — Clustering is an unsupervised learning technique that groups objects by similarity without pre-labeled training outputs
- 4: B — Physiology-based discovery is disease-centric, screening drugs or oligonucleotides for physiological effects in organ/animal models

End of Quiz

BIF501 — 48 solved MCQs
Prepared by VU Learning HUB
vulearninghub.xyz