BIOSC 1540 - Computational Biology

Bioinformatics Exam

Oct 3, 2024

100 points

Please read the following instructions carefully before beginning your assessment.

- **Time limit:** You have 75 minutes to complete and turn in this assessment.
- **Open note:** You may use notes, but with the following restrictions:
  - Notes must be hand-written on either (1) paper or (2) a tablet with a stylus, then printed.
  - You may use a maximum of one sheet of 8.5 $\times$ 11 in. paper for notes (front and back allowed).
  - Notes must be your own work. Sharing or copying notes from others is strictly prohibited.
  - Your name must be clearly written on your notes.
- **No digital devices:** The use of digital devices, including calculators, is not allowed.
- **Submission requirements:** You must submit both your completed assessment and all notes used.

I agree to follow the above instructions. I affirm that all work on this assessment will be my own and that I will not give or receive any unauthorized assistance. To have your assessment graded, you must write your name, sign, and provide your student ID below.

_____ _____

Name Signature

_____

Student ID

Sequencing

Problem 01

During Sanger sequencing, it is commonly observed that base call quality deteriorates toward the terminal regions of sequencing reads. What is the main technical factor contributing to this decline in data quality at the read termini?

(3 points)

(A) Low concentration of chain-terminating nucleotides, leading to incomplete termination events.

(B) Increased likelihood of dNTP misincorporation near the end of the sequence.

(C) Variability in fragment mass and electrophoretic mobility affecting resolution.

(D) Reduced signal intensity due to the lower quantity of fragments.

Problem 02

Select all advantages of Sanger sequencing.

(2 points)

(A) Enhanced scalability and throughput suitable for high-volume sequencing projects.

(B) Superior accuracy in homopolymeric tracts due to lower indel error rates.

(C) Ability to generate the longest read lengths among sequencing technologies.

(D) Greater cost efficiency when performing large-scale sequencing.

Problem 03

What is the primary functional role of Illumina adapters within the sequencing process?

(2 points)

(A) They incorporate fluorescent markers essential for the detection of base incorporations during sequencing.

(B) They facilitate the binding of DNA fragments to complementary oligonucleotides on the flow cell.

(C) They protect DNA fragments from degradation during the sequencing process by providing a stable binding interface.

(D) They allow for the simultaneous sequencing of multiple DNA fragments by providing unique barcode sequences.

Problem 04

Which issue is generally not associated with standard base call accuracy concerns for Illumina sequencing?

(3 points)

(A) Cross-talk between adjacent detection channels.

(B) Delayed or incomplete polymerase activity.

(C) Variability in the removal of fluorescent terminator molecules.

(D) Temperature fluctuations affecting sequencing cycle rates.

Problem 05

What is sequencing coverage, and how does it affect downstream genomic analyses?

(1 point)

(A) The proportion of the total genome size that has at least one read aligned, affecting assembly completeness.

(B) The mean number of sequencing reads encompassing each nucleotide position, influencing confidence in base accuracy.

(C) The evenness of read distribution across the genome, impacting variant detection reliability.

(D) The overall fidelity of nucleotide identification during sequencing, affecting error rates.

(E) The maximum read length achieved during sequencing runs, affecting the ability to span large genomic features.

Problem 06

In high-throughput sequencing data, which of the following provides essential per-base error probability metrics for quality control?

(1 point)

(A) The N50 contiguity statistic.

(B) The sequence in FASTQ file format.

(C) Phred scores embedded in FASTQ file format.

(D) Coverage depth information within assembled contigs.

(E) Per-sequence GC content from the sequencing summary file.

Problem 07

What is the immediate biochemical consequence when a ddNTP is incorporated by DNA polymerase?
(2 points)

(A) It terminates DNA strand synthesis, preventing further elongation.

(B) It allows the DNA strand to continue synthesizing until the next dNTP is encountered.

(C) It enhances sequencing accuracy by preventing incorporation of incorrect nucleotides.

(D) It labels the DNA strand with a fluorescent marker that signals successful base pairing.

Problem 08

What is the main purpose of performing bridge amplification during high-throughput sequencing?
(2 points)

(A) To minimize the occurrence of sequencing errors by proofreading base incorporation through amplification cycles.

(B) To generate localized clonal clusters of DNA fragments for robust signal detection.

(C) To enable high-throughput sequencing by amplifying template DNA fragments, ensuring sufficient quantities for downstream enzymatic reactions.

(D) To selectively amplify longer DNA fragments, allowing improved sequencing coverage.

(E) To enhance base calling accuracy by reducing background noise during sequencing.

Genome assembly

Problem 09

Which of the following best defines a contig in the context of genome assembly?
(2 points)

(A) A contiguous region within the genome that exhibits high sequencing coverage.

(B) A sequence of nucleotides constructed by overlapping sequencing reads.

(C) A segment of the genome prone to errors due to consistently low base quality scores.

(D) A repetitive genomic region that complicates the assembly process.

Problem 10

What is the most accurate definition of a “k-mer”, and how does it contribute to the construction of de Bruijn graphs in genome assembly algorithms?

(3 points)

(A) A substring of fixed length derived from reads and form de Bruijn graph nodes.

(B) A short sequence of nucleotides from a reference genome and represent de Bruijn graph edges.

(C) A DNA fragment of variable length for constructing nodes and edges between reads in de Bruijn graphs.

(D) Sequences with lengths less than five derived from reads to build a de Bruijn graph.

Problem 11

How is the N50 value defined, and what does it convey about the quality of an assembly?

(1 point)

(A) The sum of all contig lengths divided by the number of contigs, reflecting the average contig size across the assembly.

(B) The length of the smallest contig in a sorted list of contigs that reaches 50% of the assembly length.

(C) The contig length at which 50% of the total reads have been mapped during the assembly process, reflecting sequencing coverage uniformity.

(D) The number of contigs whose cumulative length adds up to 50% of the total genome size, reflecting contig distribution in the assembly.

Problem 12

Which of the following genomic characteristics is most likely to increase the complexity of de novo genome assembly efforts?

(2 points)

(A) Genomic regions with high GC content.

(B) Genomes with extensive repetitive elements and high repeat content.

(C) Sequencing data with minimal depth of coverage across the genome.

(D) Use of sequencing platforms that generate only short read lengths.

Problem 13

What fundamental principle is the basis for building a de Bruijn graph from sequencing reads?
(3 points)

(A) Aligning sequencing reads to an existing reference genome to identify overlaps and reconstruct the target sequence.

(B) Identifying and utilizing overlapping k-mers to construct nodes and edges within the graph

(C) Aggregating short sequencing reads to form longer contigs by prioritizing overlaps between reads.

(D) Incorporating base quality scores into the assembly algorithm to weigh k-mers based on read accuracy.

Problem 14

In graph-based genome assembly approaches, which of the following actions is typically not incorporated into the graph traversal algorithms?
(1 point)

(A) Initiating traversal from nodes with optimal coverage levels to avoid erroneous low-coverage paths.

(B) Extending the current path in the graph until a termination node, such as a dead-end or circular path, is encountered.

(C) Employing a linear search algorithm to examine all nodes in islands for potential paths systematically.

(D) Implementing backtracking strategies to resolve ambiguous branches or repetitive regions in the sequencing graph.

Problem 15

Within De Bruijn graph-based genome assembly frameworks, selecting shorter k-mer lengths facilitates which of the following aspects?
(1 point)

(A) Enhanced resolution in repetitive genomic regions by reducing the number of ambiguous nodes.

(B) Increased number of overlaps due to the higher frequency of shorter k-mers, which can complicate the identification of unique sequences.

(C) Improved management of regions with sparse sequencing coverage, as shorter k-mers provide more reliable read alignment.

(D) Increased connectivity in the assembly graph, allowing for better handling of sequencing errors and structural variations.

Problem 16

When determining the most favorable pathway for contig extraction in genome assembly, which of the following factors is considered least critical?
(1 point)

(A) Length of the traversed sequences.

(B) Uniformity of read coverage across the path.

(C) Frequency of branching points within the graph.

(D) Existence of unique, linear sequencing paths.

Problem 17

What is the primary rationale for employing paired-end sequencing reads to assemble genomes?
(2 points)

(A) To decrease computational processing time during assembly.

(B) To supply information regarding inter-contig spacing.

(C) To rectify low-quality nucleotide bases at read termini.

(D) To prevent misassembly of repetitive elements.

Problem 18

What is the main repercussion of retaining untrimmed adapter sequences?
(2 points)

(A) Causing misalignment of reads, resulting in the generation of shorter contigs.

(B) Leading to erroneous gene annotations and the omission of open reading frames.

(C) Inflating estimates of the total genome size.

(D) Introducing inaccuracies in assembly due to incorrect indels.

Problem 19

Select all statements regarding the SPAdes genome assembly tool that are incorrect.
(2 points)

(A) SPAdes utilizes De Bruijn graph structures for genome assembly.

(B) SPAdes is capable of processing both short and long sequencing reads during assembly.

(C) SPAdes constructs a singular k-mer size graph to streamline the assembly procedure.

(D) SPAdes effectively assembles bacterial genomes even in regions with low sequencing coverage.

Problem 20

In the context of prokaryotic genome assembly, what do bulges in the assembly graph typically represent?

(1 point)

(A) Errors or misassemblies that cause small divergences in sequence paths.

(B) Gaps between reads that could not be closed by paired-end reads.

(C) Repeats in the genome that are unresolved during assembly.

(D) Regions of high GC content that are difficult to sequence accurately.

Gene annotation

Problem 21

Which methodological approach does Prokka employ to detect open reading frames (ORFs) within prokaryotic genomes?

(2 points)

(A) Utilization of Hidden Markov Models (HMMs).

(B) Precise alignment of ribosomal RNA gene sequences.

(C) Identification based on learned codon motifs.

(D) Alignment with gene databases.

Problem 22

What does “functional annotation” primarily entail within the scope of gene annotation?

(3 points)

(A) Inferring the biological roles of genes through sequence similarity analyses.

(B) Detecting open reading frames (ORFs) within genomes.

(C) Constructing contigs from cleaned sequencing data.

(D) Identifying gene start and stop codons using probabilistic modeling techniques.

Problem 23

Assemble a genome sequence by using the greedy algorithm and the following reads: `ATTAGACCTG`, `CCTGCCGGAA`, `AGACCTGCCG`, `GCCGGAATAC`

(5 points)

Problem 24

Build a De Bruijn graph with k_{edge} of 5 with the following reads: `GATTAC`, `TACAGATT`, `AGATTAC`, `TACCGG`, `GGATTA` Then, using your De Bruijn graph, determine the optimal contig.

(5 points)

Alignment

Problem 25

Select all aspects that distinguish the Needleman-Wunsch from the Smith-Waterman algorithm.
(3 points)

- (A) Needleman-Wunsch performs global alignment, ensuring the entire length of both sequences is aligned.
- (B) Needleman-Wunsch is inherently parallelizable, unlike Smith-Waterman.
- (C) Smith-Waterman is optimized for aligning sequences of significantly different lengths.
- (D) Smith-Waterman allows for multiple optimal local alignments to be identified simultaneously.
- (E) Needleman-Wunsch uses affine gap penalties, whereas Smith-Waterman employs linear gap penalties.
- (F) Needleman-Wunsch initializes the scoring matrix with gap penalties along the first row and column.

Problem 26

How are affine gap penalties characterized within the sequence alignment algorithms?
(1 point)

- (A) Applying a uniform penalty to each gap irrespective of its length.
- (B) Assigning a higher penalty for initiating a gap and a lower penalty for each gap extension.
- (C) Exempting terminal gaps from penalty assignments.
- (D) Imposing progressively larger penalties as the length of the gap increases.

Problem 27

What is the principal objective of performing a multiple sequence alignment (MSA)?
(2 points)

- (A) Identifying conserved functional domains across diverse species.
- (B) Executing fast local alignments between pairwise sequences.
- (C) Detecting point mutations within individual DNA sequences.
- (D) Comparing tertiary structures of proteins within a singular organism.

Problem 28

Perform a Needleman-Wunch and Smith-Waterman alignment with the following sequences: `AATCG` and `AACG`. Use a scoring scheme of 1 for match, −1 for mismatch, and −2 for gap for both. Show all possible tracebacks and their respective alignments.

(5 points)

Needleman-Wunch

		A	A	T	C	G
A						
A						
C						
G						

Smith-Waterman

		A	A	T	C	G
A						
A						
C						
G						

Transcriptomics

Problem 29

Which of the following best describes the main difference between transcriptomics and genomics?
(3 points)

(A) Genomics studies the DNA sequence while transcriptomics analyzes RNA transcripts.

(B) Genomics focuses on protein function while transcriptomics deals with gene expression.

(C) Transcriptomics can predict protein structures while genomics predicts gene sequences.

(D) Transcriptomics focuses on the non-coding regions of DNA while genomics focuses on exons.

Problem 30

Which of the following statements best describes the primary function of the RNA Integrity Number (RIN) in evaluating RNA samples for applications like next-generation sequencing and microarray analysis?
(2 points)

(A) It quantifies the absolute amount of mRNA in a sample.

(B) It assesses the extent of RNA degradation by analyzing fragmentation patterns.

(C) It measures RNA purity by detecting contaminants like proteins and DNA.

(D) It evaluates the efficiency of reverse transcription in cDNA synthesis.

Read mapping

Problem 31

What is the key benefit of using the Burrows-Wheeler Transform (BWT) in sequence alignment algorithms?
(3 points)

(A) It enables fast and memory-efficient searching of large genomes.

(B) It helps detect novel splice sites during RNA-seq analysis.

(C) It eliminates the need for gap penalties in sequence alignment.

(D) It provides an efficient way to construct phylogenetic trees.

Problem 32

Which of the following describes the search strategy in hash-based alignment?
(2 point)

(A) Finding a short exact match and then attempting to extend it in both directions.

(B) Using suffix arrays to find starting indices of k-mers.

(C) Constructing a sequence alignment by iterating through each possible alignment.

(D) Immediately calculating the optimal alignment score.

Problem 33

How do suffix trees improve sequence alignment in comparison to hash-based methods?
(1 point)

(A) Suffix trees allow for faster exact pattern matching.

(B) Suffix trees can align sequences in parallel, speeding up the process.

(C) Suffix trees handle larger genomes with lower memory requirements.

(D) Suffix trees remove the need for gap penalties in alignment algorithms.

Problem 34

Given the initial string `banana`, what are the zero-based starting indices of the suffix array?
(1 point)

(A) [6, 5, 3, 1, 0, 4, 2]

(B) [6, 5, 3, 1, 0, 2, 4]

(C) [6, 3, 5, 1, 4, 2]

(D) [1, 3, 0, 5, 2, 4]

Problem 35

Perform the Burrows-Wheeler Transform (BWT) of the string `CANADA`. Show all intermediate steps.

(5 points)

Problem 36

Given the Burrows-Wheeler Transform (BWT) `GT$ATCTGCGA`, determine the original string. Show all of your work.

(5 points)

Quantification

Problem 37

What is the key distinction between pseudoalignment and full alignment?

(2 points)

(A) Pseudoalignment is slower but more accurate than full alignment.

(B) Pseudoalignment only determines transcript compatibility.

(C) Full alignment does not account for sequencing errors.

(D) Pseudoalignment requires fewer transcripts in the reference.

Problem 38

Which of the following best explains the purpose of Salmon's generative model in RNA-seq data analysis?

(3 points)

(A) It identifies the precise position of each read on the transcript.

(B) It generates new reads from known transcripts for quality control.

(C) It models how sequencing reads are generated from a population of transcripts.

(D) It adjusts the effective length of each transcript to correct for biases.

Problem 39

What is the main goal of the Expectation-Maximization (EM) algorithm used in Salmon's inference process?

(1 point)

(A) To minimize the variance in transcript abundance estimates.

(B) To optimize the likelihood of observing the RNA-seq reads.

(C) To identify the most compatible transcripts for each read.

(D) To estimate the total number of fragments generated by the sequencing experiment.

Problem 40

How does Salmon's two-phase inference process achieve both speed and accuracy?
(2 points)

(A) By running the Expectation-Maximization algorithm in both the online and offline phases.

(B) By performing a quasi-mapping in the online phase and refining results with full alignment in the offline phase.

(C) By using quasi-mapping for initial abundances and refining these estimates using more accurate methods.

(D) By combining read mapping and assembly-based quantification in both phases.

Problem 41

What is the significance of the transcript-fragment assignment matrix, Z, in Salmon's model?
(2 points)

(A) It records the exact positions of all fragments within each transcript.

(B) It is used to map each fragment to all possible compatible transcripts.

(C) It stores the probabilities of fragments originating from each transcript

(D) It provides the nucleotide sequences of all transcripts.

Differential gene expression

Problem 42

Why is the Negative Binomial distribution commonly used in differential gene expression analysis of RNA-seq data?
(2 points)

(A) It accounts for the large number of zeros often observed in RNA-seq data.

(B) It models the overdispersion in the data where the variance exceeds the mean.

(C) It ensures that only the most highly expressed genes are analyzed.

(D) It simplifies the analysis by assuming equal variance across all genes.

Problem 43

In RNA-Seq data analysis, what does a very low p-value suggest when comparing gene expression between two conditions?

(2 points)

(A) The observed differences in gene expression are likely due to random chance.

(B) The observed differences in gene expression are statistically significant.

(C) The gene expression is identical between the two conditions.

(D) The statistical model is not suitable for the data.

Problem 44

Which of the following best describes the TPM normalization method in RNA-Seq?

(1 point)

(A) It adjusts for differences in gene length and sequencing depth.

(B) It normalizes based on the total number of mapped reads, ignoring gene length.

(C) It uses logarithmic scaling to adjust for gene expression levels.

(D) It is identical to the FPKM method but includes a correction for GC content.

Problem 45

Select all factors that could contribute to overdispersion in normalized RNA-seq count data.

(1 point)

(A) Variability in gene length across the transcriptome.

(B) Presence of unmodeled confounding covariates such as batch effects.

(C) Heterogeneous RNA integrity (RIN) scores among samples.

(D) Differential isoform expression is not accounted for in the analysis.

(E) Technical variability in sequencing depth among samples.

(F) Low detection rate of highly expressed genes.

(G) Uniform gene expression levels across all genes and samples.

(H) Nonuniform coverage across inserts.