BIOSC 1540 - Computational Biology

Final Exam

Dec 16, 2024

100 points

Please read the following instructions carefully before beginning your assessment.

- **Time limit:** You have 150 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.

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

Recommendation: More challenging problems are worth fewer points, so answer the easier problems first. Choose the best answer for each problem unless specified otherwise.

Problem 1

Build a De Bruijn graph with k_{edge} = 3 with all the following reads: `GGATT`, `GATTA`, `TACAG`, `AGATT`, `TACCG`. Using your De Bruijn graph, write down the optimal contig (there is only one). Assume that you can only use each edge up to two times.

(4 points)

Problem 2

In RNA-seq analysis, what is the most significant problem caused by adapter contamination?
(2 points)

(A) False splice junctions during transcript mapping.

(B) Reduction in overall sequencing depth.

(C) Bias in GC content calculations.

(D) Alteration of read quality scores.

Problem 3

What key advantage does RPKM (Reads Per Kilobase Million) provide in RNA-seq analysis?
(3 points)

(A) It corrects for sequence-specific biases in PCR amplification.

(B) It adjusts for differences in RNA degradation between samples.

(C) It enables direct comparison of expression levels between different gene lengths.

(D) It normalizes for differences in GC content between genes.

Problem 4

How does increasing sequencing coverage from 10x to 30x most significantly impact variant calling in genomic analysis?
(2 points)

(A) It increases the maximum length of insertions that can be detected.

(B) It improves confidence in identifying heterozygous variants.

(C) It reduces the time required for computational analysis.

(D) It enables detection of more complex structural variants.

Problem 5

In de Bruijn graph construction for genome assembly, what is the relationship between k-mer size and graph complexity?
(1 point)

(A) Larger k-mers result in simpler graphs with fewer branching points but require higher coverage.

(B) Smaller k-mers produce simpler graphs and require less coverage for accurate assembly.

(C) K-mer size has no effect on graph complexity, which is determined solely by genome size.

(D) Larger k-mers always produce more accurate assemblies regardless of coverage.

Problem 6

During genome assembly graph traversal, which strategy would most likely lead to inefficient or incorrect assembly?

(1 point)

(A) Breaking cycles in the graph by identifying repeat regions.

(B) Implementing backtracking when encountering branching points in the graph.

(C) Starting traversal from nodes with high coverage and extending in both directions.

(D) Using a depth-first random walk without considering coverage information.

Problem 7

Which feature of prokaryotic genes most significantly influences the accuracy of computational open reading frame (ORF) detection?

(4 points)

(A) The presence of well-defined ribosomal binding sites upstream of start codons.

(B) The length distribution of intergenic regions.

(C) The presence of transcription termination sequences.

(D) The GC content of coding regions.

Problem 8

In RNA-seq analysis, which scenarios would require a negative binomial model rather than a simpler Poisson model? Select all that apply.

(1 point)

(A) When biological replicates show higher variance than expected from sampling alone.

(B) When samples are sequenced at different depths.

(C) When samples come from different experimental batches.

(D) When analyzing differential expression between conditions.

(E) When analyzing technical replicates from the same sample.

Problem 9

What is the main advantage of Salmon's online phase in transcript quantification?
(2 points)

(A) It performs complete alignment of all reads against the transcriptome.

(B) It eliminates the need for further refinement in abundance estimation.

(C) It identifies all possible splice variants for each gene.

(D) It provides rapid initial abundance estimates.

Problem 10

Why do Sanger sequencing reads typically show higher quality scores at the middle positions compared to both ends of the read?
(3 points)

(A) DNA polymerase has higher accuracy in synthesizing medium-length fragments.

(B) Base-calling algorithms are optimized for the middle of reads.

(C) The middle positions have more balanced representation of fragment lengths in the reaction.

(D) PCR amplification is more efficient for medium-length fragments.

Problem 11

In which scenario would protein threading be more likely to succeed than homology modeling?
(2 points)

(A) When the protein has 20% sequence identity but conserved fold patterns with known structures.

(B) When the protein shares 60% sequence identity with a known structure.

(C) When the protein contains multiple domains with varying levels of conservation.

(D) When crystal structures exist for close homologs of the target protein.

Problem 12

What makes correlated mutations in protein sequences valuable for predicting three-dimensional protein structure?
(3 points)

(A) They determine the evolutionary age of specific protein domains.

(B) They reveal which amino acids are most conserved across species.

(C) They predict the rate of protein folding in different cellular environments.

(D) They identify pairs of residues that maintain physical contact.

Problem 13

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

(4 points)

		G	**C**	**A**	**T**	**A**	**T**	**A**	**C**	**G**	**C**
T											
C											
G											
T											
A											
G											
C											
T											
A											

Problem 14

When developing a new force field for protein simulations, which approach would provide the most reliable parameter validation?

(2 points)

(A) Comparing protein-ligand binding free energies with experimental measurements.

(B) Testing if the force field can fold a protein from a random coil.

(C) Measuring how well the force field reproduces quantum mechanical energies.

(D) Counting how many hydrogen bonds form during a simulation.

Problem 15

Which challenge best explains why multiple types of scoring functions are often used together in molecular docking?

(3 points)

(A) Individual scoring functions are too computationally expensive to use alone.

(B) Different scoring functions capture complementary aspects of protein-ligand binding.

(C) Single scoring functions cannot handle different protein structures.

(D) Using multiple scoring functions increases the speed of virtual screening.

Problem 16

During a protein-ligand binding simulation, a researcher observes that important binding events are too rare to study effectively. Which simulation approach would be most appropriate to address this?

(3 points)

(A) Change to a different force field parameter set.

(B) Increase the simulation box size to include more solvent molecules.

(C) Reduce the simulation temperature to slow molecular motions.

(D) Sample along a collective variable between bound and unbound states.

Problem 17

Why is the minimum image convention essential when implementing periodic boundary conditions in molecular dynamics simulations?

(2 points)

(A) It prevents particles from interacting with multiple copies of themselves.

(B) It reduces the computational cost by eliminating the need for calculating forces for all atoms.

(C) It allows particles to move freely between different simulation boxes.

(D) It increases the accuracy of long-range electrostatic calculations.

Problem 18

Why does X-ray crystallography require multiple protein molecules arranged in a crystal lattice rather than a single protein molecule?

(4 points)

(A) Individual protein molecules move too quickly to be analyzed by X-rays.

(B) The crystal structure prevents radiation damage to the protein.

(C) The regular arrangement of proteins amplifies weak X-ray scattering signals to detectable levels.

(D) Crystallization removes water molecules that interfere with X-ray diffraction.

Problem 19

What typically happens if a molecular dynamics simulation uses a time step that is too large relative to the fastest molecular motions?

(3 points)

(A) During equilibration, the simulation will adjust the time step to maintain stability.

(B) The simulation will have unrealistic atomic movements and energy conservation violations.

(C) The simulation slows down to compensate for the large time step.

(D) The simulation runs more efficiently and sample more conformations.

Problem 20

Why are Fourier series particularly well-suited for modeling dihedral angle potentials in force fields?
(2 points)

(A) They naturally capture the periodic nature of rotation around chemical bonds.

(B) They require fewer computational resources than other mathematical functions.

(C) They allow for direct incorporation of quantum mechanical data.

(D) They automatically adjust to different types of chemical bonds.

Problem 21

In molecular dynamics simulations, why are multiple independent simulation runs considered more statistically robust than a single long simulation?
(2 points)

(A) They allow for parallel processing of different initial molecular configurations, reducing overall computational time.

(B) Multiple runs provide a more comprehensive sampling of the system's conformational space by exploring different initial microstates.

(C) Independent runs enable direct comparison of simulation outcomes to identify systematic biases in the computational method.

(D) They provide redundant data points that can be averaged to reduce statistical noise in the simulation results.

Problem 22

In molecular binding processes, entropy is best described as:
(3 points)

(A) A static property that determines molecular interactions based on molecular size and shape.

(B) A measure of thermal energy transfer between molecules during complex formation.

(C) The quantitative change in molecular degrees of freedom upon binding.

(D) An exclusively enthalpic phenomenon that predicts the stability of molecular complexes.

Problem 23

In computational drug discovery, the primary goal of generating molecular fingerprints through hashing is to:

(1 point)

(A) Simulate molecular interactions through mathematical transformations.

(B) Generate unique identifiers that perfectly capture a molecule's three-dimensional structure.

(C) Compress complex molecular structural information into a computationally manageable format.

(D) Standardize molecular representations for rapid computational comparison.

Problem 24

In molecular binding processes, Gibbs free energy (ΔG_{bind}) fundamentally represents:

(4 points)

(A) The total energy required to initiate molecular interactions under standard conditions.

(B) The maximum work that can be extracted from a molecular binding process at constant temperature and pressure.

(C) A measure of the spontaneity and energetic favorability of molecular association.

(D) The precise mechanical work needed to overcome intermolecular repulsive forces.

Problem 25

In molecular interactions, which type of noncovalent force typically provides the most important energetic contribution to specific molecular recognition?

(1 point)

(A) Van der Waals interactions that depend on temporary electron density fluctuations.

(B) Specialized chemical interactions that form precise, directional patterns.

(C) Charge-based interactions that create long-range attraction between molecular partners.

(D) Quantum mechanical coupling effects between molecular electronic structures.

Problem 26

In computational molecular simulations, the representation of chemical bonds as a mechanical model primarily aims to:

(3 points)

(A) Provide a computationally efficient approximation of molecular bond dynamics.

(B) Capture the exact quantum mechanical behavior of electron interactions.

(C) Simulate the complete breaking and reformation of chemical bonds during interactions.

(D) Replicate the precise vibrational modes of molecular structures.

Problem 27

In structural biology, constructing an accurate protein model from experimental electron density data involves:

(3 points)

(A) Systematically comparing theoretical atomic models with observed experimental data.

(B) Using computational algorithms to predict protein folding based on sequence information.

(C) Manually tracing electron density contours to determine molecular structure.

(D) Generating multiple independent structural models to capture protein variability.

Problem 28

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

(4 points)

Problem 29

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

(4 points)

Problem 30

In transcriptomics, the fragment assignment matrix represents a critical step in:

(2 points)

(A) Precisely determining the genomic origin of sequencing fragments with deterministic mapping.

(B) Probabilistically resolving fragment compatibility across multiple potential transcripts.

(C) Generating a comprehensive catalog of all possible transcript variants.

(D) Calculating absolute fragment count distributions for each transcript.

Problem 31

In transcriptomics, a generative model's primary purpose is to:
(2 points)

(A) Create a predictive framework for experimental design.

(B) Develop a comprehensive mapping of transcript diversity.

(C) Reconstruct the precise molecular pathway of RNA synthesis.

(D) Simulate the probabilistic process of RNA sequencing fragment production.

Problem 32

In molecular biology research, the primary purpose of assessing RNA sample quality involves:
(4 points)

(A) Quantifying the molecular characteristics that impact downstream experimental reliability.

(B) Determining the potential for accurate gene expression measurements.

(C) Establishing the biochemical potential of RNA molecules for experimental use.

(D) Identifying the structural stability of RNA for long-term storage.

Problem 33

In genomic research, functional annotation primarily aims to:
(4 points)

(A) Classify genes based on their evolutionary conservation patterns.

(B) Determine the structural characteristics of genomic regions.

(C) Predict the associated molecular interactions and cellular processes.

(D) Quantify the expression levels of newly identified genes.

Problem 34

In genome assembly algorithms, which parameter would have the least direct impact on determining the most reliable contig path?
(4 points)

(A) Computational resource requirements for path exploration.

(B) Diversity of sequencing reads contributing to the path.

(C) Statistical confidence in path connectivity.

(D) Potential for introducing sequencing artifacts.

Problem 35

Why do ddNTPs cause chain termination in DNA synthesis while dNTPs allow continued elongation?
(4 points)

(A) ddNTPs lack the 3' hydroxyl group necessary for forming the next phosphodiester bond in DNA elongation.

(B) ddNTPs form weaker hydrogen bonds with template DNA, causing the polymerase to release the growing strand.

(C) ddNTPs change the conformation of DNA polymerase, preventing it from adding more nucleotides.

(D) ddNTPs block the binding site for the next incoming nucleotide, physically preventing further additions.

Problem 36

In Illumina sequencing, what happens to DNA fragments that lack properly ligated adapters?
(4 points)

(A) They bind to the flow cell but cannot form clusters due to incomplete bridge formation.

(B) They fail to bind to the flow cell surface and are washed away during bridge amplification.

(C) They form clusters but cannot be sequenced due to missing primer binding sites.

(D) They produce weak signals during sequencing due to reduced fluorescent marker incorporation.