Name:

NetID:

Discussion Section:

Course Heading:

Keep your answers concise and to the point.
Long responses won't earn extra credit.

1.(20 pts) Genomics

(a) Which of the following sequencing methods cannot directly provide chromatin accessibility information? (5 pts)

A. ATAC-seq
B. FAIRE-seq
C. DNase-seq
D. Bisulfite-seq

D

(b) Name two sources of sequencing bias in high-throughput sequencing (5 pts)
GC content;
Adapter ligation bias
PCR bias/ Amplification bias
Strand specific bias
Or reasonable answer
(2.5 pt each)

(c) Which of the following is a key advantage of long-read sequencing (e.g., PacBio, ONT) compared to short-read sequencing technologies? (5 pts)

A. Higher throughput per run
B. Lower sequencing cost per base
C. Improved ability to resolve structural variants and repetitive regions
D. Consistently higher per-base accuracy across all sequencing platforms

C

(d) Which of the following best distinguishes somatic mutations from germline mutations? (5 pts)

A. Somatic mutations always occur after birth, whereas germline mutations only occur during embryonic development.
B. Germline mutations affect only a single organ system, while somatic mutations impact the entire body.
C. Somatic mutations arise in non-reproductive cells and cannot be inherited, while germline mutations occur in reproductive cells and can be passed to offspring.
D. Germline mutations are more common than somatic mutations in multicellular organisms.

C

2.(15 pts) Proteomics

(a) Compared to sequencing of DNA, why is proteomics analysis more dependent on sample abundance? (5 pts)

Proteomics lack amplification step
Proteomics depend on sample abundance

(b) Name one major advantage of **mass spectrometry-based proteomics** over traditional antibody-based protein detection methods. (5 pts)

MS doesn't require specific antibodies

(c) What is the primary advantage of Cryo-Electron Microscopy (Cryo-EM) over X-ray crystallography for protein structure determination? (5 pts)

Cryo-EM doesn't require crystallization

3. (10 pts) Variant calling
(a) Which of the following best differentiates a single nucleotide variant (SNV) from a structural variant (SV)? (5 pts)

A. SNVs result from point mutations at a single nucleotide position, whereas SVs involve larger-scale genomic alterations such as inversions, translocations, or copy number changes.
B. SNVs always have functional consequences, while SVs are typically silent mutations.
C. SNVs can only be detected using short-read sequencing, whereas SVs can only be detected using long-read sequencing.
D. SNVs are the primary cause of genetic diseases, while SVs rarely contribute to disease.

A

(b) Name two major sources of false positives in variant calling. (5 pts)

Sequencing errors
Alignment errors
Or reasonable answer
(2.5 pt each)

4. (5 pts) Multi-omics

Which statement about Hi-C is correct?

A. Hi-C is primarily used to detect DNA-binding protein motifs directly.
B. Hi-C measures long-range chromatin interactions by capturing spatial proximity in the nucleus.
C. Hi-C can only be used to detect short DNA fragments of approximately 50 bp.
D. Hi-C cannot provide any information on three-dimensional genome organization.

B

5.Database (10 pt)

(a). Consider the following table (5pt):

Patient_ID	Patient_Name	Zip_Code	City	State
1	John Smith	6511	New Haven	CT
2	Mary Doe	6511	New Haven	CT
3	Adam Brown	6611	Trumbull	CT

What **normal form violation** is present in this table, and how can it be corrected?
A. 1NF violation; split multi-valued fields into separate rows.
B. 2NF violation; remove partial dependencies by creating separate tables for Patient and Address.
C. 3NF violation; eliminate transitive dependencies by separating City and State into another table.
D. No violation; the table is normalized properly.

C

(b). Consider the following **unnormalized** table storing patient appointment data (8pt):

Appointment_ID	Patient_Name	Doctor_Name	Doctor_Specialty	Clinic_Address	Appointment_Date
A001	John Smith	Dr. Lee	Cardiology	123 Main St	1/15/24
A002	John Smith	Dr. Lee	Cardiology	123 Main St	2/10/24
A003	Mary Doe	Dr. Patel	Neurology	456 Elm St	3/5/24

Question:

a) Identify two types of data redundancy present in this table. (3 points)

Repeated **Doctor information** (Dr. Lee, Cardiology, 123 Main St appears multiple times).

Repeated **Patient information** (John Smith listed multiple times).

Transitive dependency

b) Explain how normalization can address these redundancies, and **briefly describe** the structure of the normalized tables. (5 points)

Step 1: Create a Patient table: Patient_ID, Patient_Name

Step 2: Create a Doctor table: Doctor_ID, Doctor_Name, Doctor_Specialty, Clinic_Address

Step 3: Create an Appointment table: Appointment_ID, Patient_ID, Doctor_ID, Appointment_Date

6. (10 pts) Personal genome

(a) (5 pts) What is the typical number of single nucleotide polymorphisms (SNPs) in one person's typical genome with respect to the human reference genome?

A. ~4,000

B. ~40,000

C. ~400,000

D. ~4,000,000

E. ~40,000,000

D

(b) (5 pts) Polygenic Risk Scores (PRS) are increasingly used in personal genomics to predict an individual's susceptibility to complex diseases. Which of the following statements about PRS is **TRUE**?

A. PRS can predict the exact likelihood of developing a specific disease with 100% accuracy.
B. PRS are calculated based on the effects of a single gene variant associated with the disease.
C. PRS aggregate the effects of multiple genetic variants, each contributing a small effect to disease risk.
D. PRS are universally accurate across all populations, regardless of genetic ancestry differences.

C

7. (15 pts) Sequence Alignment

(a) (5pt) List major changes needed to convert the global alignment algorithm into a local alignment algorithm? List at least two to receive full credit.

Allow Negative Scores to Reset to Zero

Traceback Starts from the Highest Score

No Penalty for Gaps at Sequence Ends

(b) (10 pts) Align the following two sequences using the Needleman-Wunsch global alignment or Smith-Waterman local alignment algorithm. Show the complete dynamic programming matrix, and write out the final aligned sequences.

Sequence 1: **GATTACA** Sequence 2: **GCATGCU**

Scoring Scheme: **Match: +2** **Mismatch: -1** **Gap penalty: -2**

Instructions:

1. Construct the dynamic programming matrix and fill in the matrix using the scoring scheme provided.
2. Identify the optimal alignment by performing a traceback and write out the final aligned sequences (use dashes - for gaps).

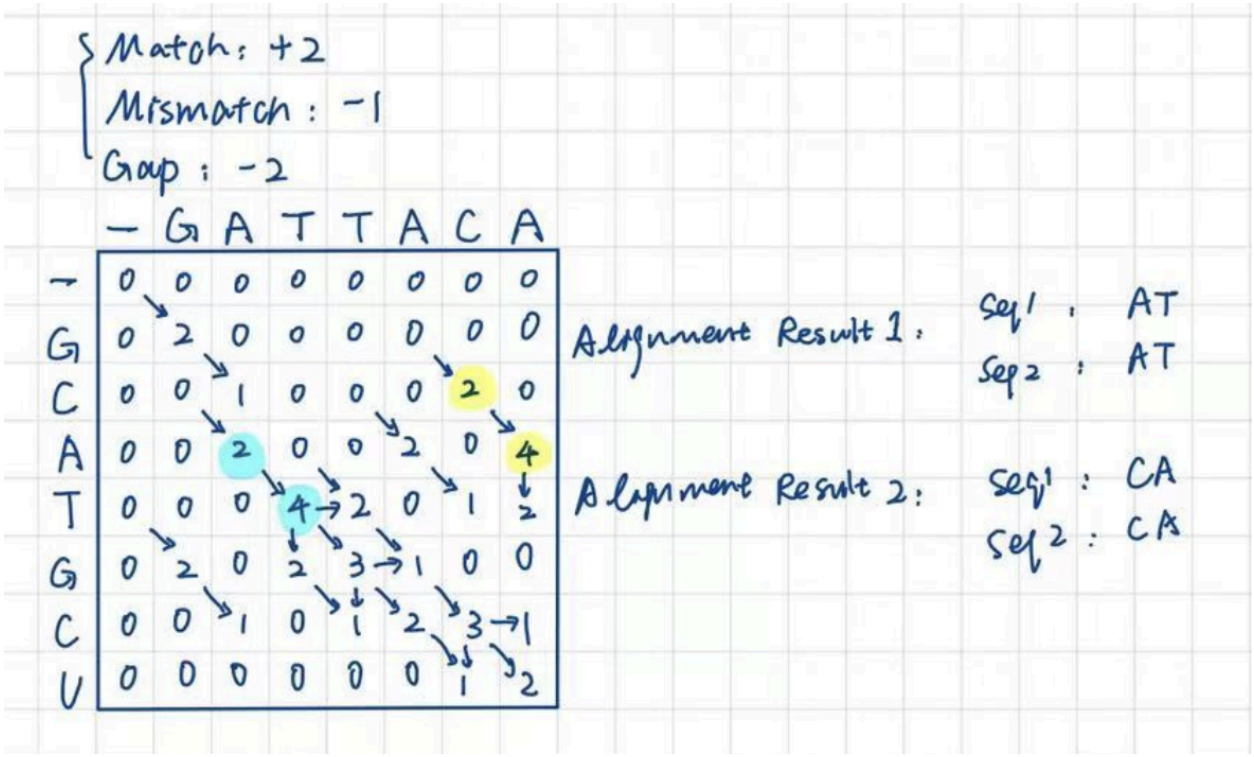
Global Alignment:

	-	G	A	T	T	A	C	A
-	0	-2	-4	-6	-8	-10	-12	-14
G	-2	2	0	-2	-4	-6	-8	-10
C	-4	0	1	-1	-3	-5	-4	-6
A	-6	-2	2	0	-2	0	-2	0
T	-8	-4	0	4	2	0	-2	-2
G	-10	-6	-2	2	3	1	-1	-3
C	-12	-8	-4	0	1	2	3	1
U	-14	-10	-6	-2	-1	-1	1	2

G - A T T A C A

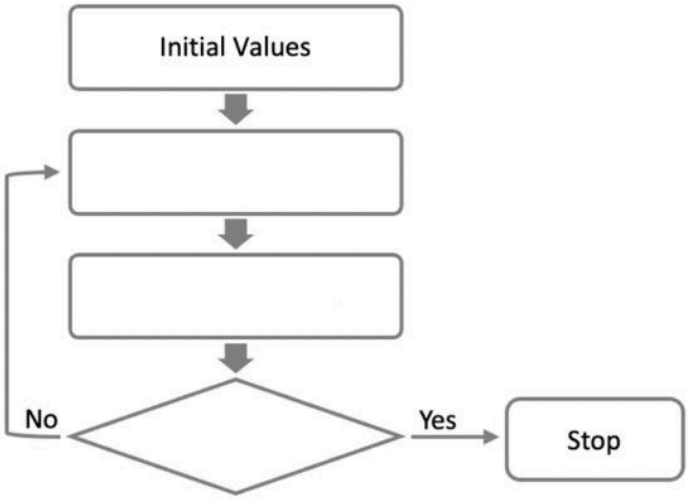
G C A T G - C U

Local Alignment:

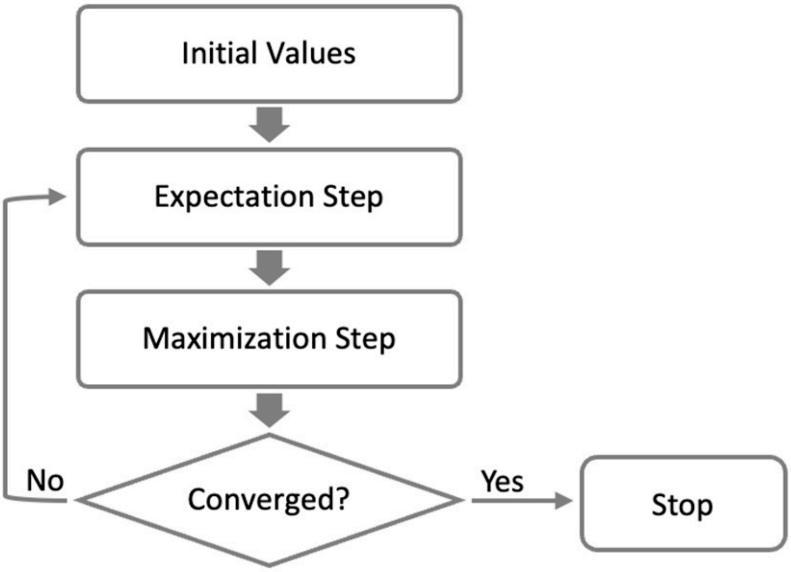


8. Multiple Alignment

(a) Briefly describe the main steps of the EM algorithm. (3 pts)



Answer:



9. FAST Alignment

(a) List the three main steps of the FASTA algorithm used for sequence alignment. (5pt)

(b) Hashing the query sequence into short words (k-mers).

(c) Scanning the database to find exact matches for these k-mers.

(d) Extending matches along diagonals and refining using local alignment.

Keywords: Hashing, short words/k-mers, exact matches, extend match, diagonals, local alignment

(e) Rank the following sequence alignment algorithms from fastest (1) to slowest (3) (5 pt):

- Smith-Waterman
- BLAST
- HMM

(1) BLAST
(2) Smith-Waterman
(3) HMM