Indraprastha Institute of Information Technology Delhi (IIITD)
Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

Rubrics_Quiz-3 (October 21, 2025)

Total time: 50 mins, Total marks: 40

Name: _____ Roll No. _____

Choose the correct answer for the following questions (**1-12**):

1. In RNA-seq, biological replicates are important because **(1 mark)**
   a. Technical variability is higher than biological variability
   b. They increase statistical power by capturing biological variability
   c. They increase sequencing read length
   d. They remove GC bias

2. Loop modelling is difficult because loops **(1 mark)**
   a. Are always missing from templates
   b. Have high conformational variability and depend on long-range packing
   c. Contain no hydrogen bonds
   d. Break templates

3. A statistically significant p-value (<0.05) in Differential Expression analysis means **(1 mark)**
   a. The null hypothesis is definitely false
   b. There is less than 5% probability that the observed difference occurred by chance
   c. The alternative hypothesis is false
   d. The experiment must always be repeated

4. Adapter contamination appears in the quality check as: **(1 mark)**
   a. Uniform peaks across cycles
   b. Sudden drop in sequence complexity
   c. Overrepresented short sequences
   d. Increased GC-bias

5. Why do *ab initio* models frequently become trapped in incorrect conformations? **(1 mark)**
   a. No phi (φ) and psi (ψ) angles are allowed
   b. Energy landscapes have many local minima
   c. Templates are missing
   d. Hydrogen atoms are ignored

6. During DE analysis, log fold change is preferred over raw fold change because **(1 mark)**
   a. RNA-seq cannot c
   b. ompute raw fold change
   c. Log scaling stabilises variance and is symmetric for up/down-regulated genes
   d. It removes GC-bias
   e. It increases counts

7. Which of the following statements about the Ramachandran plot is INCORRECT? **(1 mark)**

a. It plots phi (φ) and psi (ψ) dihedral angles of amino acid residues
b. Glycine has the most restricted conformational space due to its side chain
c. Sterically allowed regions indicate no atomic clashes
d. Proline has limited phi angle rotation due to its cyclic structure

8. In RNA-seq analysis, what is the PRIMARY reason for performing multiple testing corrections?

a. To increase statistical power **(1 mark)**
b. To reduce false positives when testing thousands of genes simultaneously
c. To normalise for sequencing depth
d. To account for biological variability

9. A Phred quality score of 30 indicates a base call accuracy of? **(1 mark)**

a. 90%
b. 99%
c. 99.9%
d. 99.99%

10. Which statement about microarrays vs RNA-seq is CORRECT? **(1 mark)**

a. Microarrays have better dynamic range than RNA-seq
b. RNA-seq requires prior knowledge of the transcriptome
c. RNA-seq can detect novel transcripts and gene fusions
d. Microarrays are better for detecting single-nucleotide polymorphisms

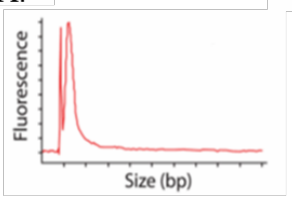
11. Which of the following is true with respect to RNA-seq differential gene expression analysis?

1. The chances of finding false positives increase with the increasing number of genes.
2. For 30000 genes and a p-value of 0.05, one would expect 1500 genes to have a p-value<0.05 by chance.
3. p-value is adjusted using methods like false discovery rate. **(2 marks)**

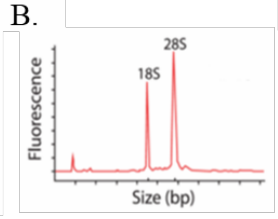
a. 1 and 2
b. 1 and 3
c. 2 and 3
d. All of these
e. None of these

12. Which of the following indicates the best and the worst quality RNA? **(2 marks)**

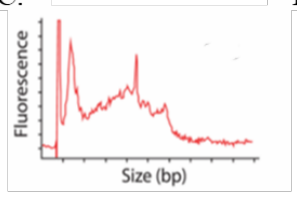
A.



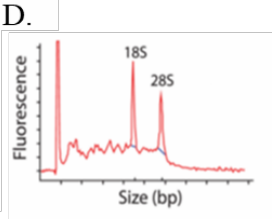
B.



C.



D.



a. Best- A, Worst- D
b. Best- B, Worst- A
c. Best- B, Worst- C
d. Best- A, Worst- D

Answer the following question:

13. A researcher sequences two RNA samples with the following results: Sample A has 10 million reads, Sample B has 50 million reads. Both have 5000 reads mapping to Gene X (length 2 kb). Calculate the RPKM for Gene X in both samples and explain if there's differential expression. **(5 marks)**

Ans: Sample A total reads N_A = 10,000,000.
Sample B total reads N_B = 50,000,000.
Reads mapping to Gene X C=5,000.
Gene length L = 2 kb = 2,000 bp.

RPKM formula: RPKM =(10^9×C)/(N×L).

Calculate Sample A step-by-step:
Numerator: 10^9 × 5,000 = 5 × 10^12.
Denominator: 10,000,000 × 2000 = 10^7 × 2 × 10^3 = 2 × 10^10.
RPKM (A) = 5 × 10^12 ÷ 2 × 10^10 = (5/2) × 10^2 = 2.5 × 10^2 = 250.

Calculate Sample B:
Numerator: 5 × 10^12(same).
Denominator: 50,000,000 × 2000 = 5 × 10^7 × 2 × 10^3 = 1 × 10^11.
RPKM (B) = 5 × 10^12 ÷ 1 × 10^11 = 50.

RPKM for Gene X — Sample A = 250, Sample B = 50.
Interpretation: Gene X appears 5-fold higher in Sample A than in Sample B (250/50 = 5; log fold-change ≈ 2.32).

14. Which of the following is not true about the template Selection Step? Justify your answer.

a. The first step in protein structural modelling is to select appropriate structural templates
b. This forms the foundation for the rest of the modelling process.
c. There is no use of heuristic alignment search programs.
d. The template selection involves searching the non-redundant protein sequence database for homologous proteins to be used as templates. **(5 marks)**

Ans: **C. NOT TRUE (Incorrect Statement)**
Template selection **heavily depends** on heuristic alignment search programs such as:

- BLAST
- PSI-BLAST
- HMMER / HHsearch / HHblits

These programs are essential for:

- Rapid similarity searches
- Identifying homologous proteins
- Detecting distant evolutionary relationships

Therefore, the statement *"There is no use of heuristic alignment search programs"* is **false**.

D. **NOT TRUE (Incorrect Statement)**
Template selection **does not typically use the non-redundant (NR) protein sequence database**.
Instead, it uses **structural databases** such as:

- **PDB (Protein Data Bank)**
- **SCOP / CATH**

- **PDB70 / PDB100 chains** (non-redundant structural sets)
  NR is a *sequence* database, not a *structural* one.
  Template selection requires **structurally solved proteins**, not just sequences.

15. Use the information given in the table to find out the following value: **(8 marks)**

	β-Sheet	Helix	Others	Total
G	200	150	30	380
V	100	400	180	680
All residues	600	2100	500	3200

a. P (SS = β-Sheet | aa = G)

Ans: P (SS = β-Sheet | aa = G) = 200 / 380
Simplify: divide numerator & denominator by 20 → 10 / 19
Decimal: 10 ÷ 19 = 0.526315 (6 d.p.)
Per cent: 52.6316% (4 d.p.)

b. P (SS = Helix | aa = V)

Ans: P (SS = Helix | aa = V) = 400 / 680
Simplify: divide by 20 → 20 / 34 = divide by 2 → 10 / 17
Decimal: 10 ÷ 17 = 0.588235 (6 d.p.)
Per cent: 58.8235% (4 d.p.)

c. P (SS = ~ β-Sheet | aa = G)

Ans: P (SS ≠ β-Sheet | aa = G) = 1 – P (β | G) = (380 − 200) / 380 = 180 / 380
Simplify: divide by 20 → 9 / 19
Decimal: 9 ÷ 19 = 0.473684 (6 d.p.)
Per cent: 47.3684% (4 d.p.)

d. P (SS = ~ Helix | aa = V)

Ans: P (SS ≠ Helix | aa = V) = 1 – P (Helix | V) = (680 − 400) / 680 = 280 / 680
Simplify: divide by 20 → 14 / 34 = divide by 2 → 7 / 17
Decimal: 7 ÷ 17 = 0.411765 (6 d.p.)
Per cent: 41.1765% (4 d.p.)

e. P (SS = β-Sheet)

Ans: P (SS = β-Sheet) = β_total / grand total = 600 / 3200
Simplify: divide by 200 → 3 / 16
Decimal: 3 ÷ 16 = 0.187500 (6 d.p.)
Per cent: 18.7500% (4 d.p.)

f. P (SS = ~β-Sheet)

Ans: P (SS ≠ β-Sheet) = 1 − P(β) = (3200 − 600) / 3200 = 2600 / 3200
Simplify: divide by 200 → 13 / 16
Decimal: 13 ÷ 16 = 0.812500 (6 d.p.)
Per cent: 81.2500% (4 d.p.)

g. P (SS = Helix)

Ans: P (SS = Helix) = Helix_total / grand total = 2100 / 3200
Simplify: divide by 100 → 21 / 32
Decimal: 21 ÷ 32 = 0.656250 (6 d.p.)
Per cent: 65.6250% (4 d.p.)

h. P (SS = ~Helix)

Ans: P (SS ≠ Helix) = 1 − P(Helix) = (3200 − 2100) / 3200 = 1100 / 3200
Simplify: divide by 100 → 11 / 32
Decimal: 11 ÷ 32 = 0.343750 (6 d.p.)
Per cent: 34.3750% (4 d.p.)

16. State whether the following statements are correct or incorrect? In case of an incorrect statement, justify your answer. **(5 marks)**

a. Conformational search algorithm in *ab initio* protein structure modelling explores the potential energy surface and locates the local minimum.

Ans: In**correct.**
Justification: Ab-initio methods (e.g., Monte Carlo, simulated annealing, fragment assembly) explore the protein's potential energy landscape to find low-energy conformations; these searches typically locate local minima (ideally the global minimum, but often local minima).

b. Logs put positive and negative values of fold changes on a symmetric scale.

Ans: **Correct.**
Justification: Log transformation (commonly log) converts fold changes so up- and down-regulation are symmetric (e.g., FC=4 → +2; FC=0.25 → −2), which stabilises variance and simplifies interpretation.

c. The correct sequence of steps involved in RNA-seq analysis includes library preparation, sequencing, quantification, read mapping, and differential expression analysis.

Ans: **Incorrect.**
Justification: The correct order is **library preparation → sequencing → read mapping → quantification → differential expression analysis**. Quantification (counting reads per feature) requires mapped reads, so it comes after read mapping.

d. Technical replicates generally increase statistical power more than biological replicates.

Ans: **Incorrect.**
Justification: Biological replicates capture true biological variability and therefore increase statistical power for detecting real differences. Technical replicates only reduce measurement noise and have much less impact on power for biological hypotheses.

e. Total RNA extracted from the cells can be directly used for sequencing.

Ans: **Incorrect.**
Justification: Total RNA must be processed into a sequencing library (e.g., rRNA depletion or

poly(A) selection, fragmentation, cDNA synthesis, adapter ligation). Raw total RNA cannot be directly loaded onto sequencers.

17. Differentiate between diagnostic and prognostic biomarkers with the help of an appropriate example. **(3 marks)**

Ans: **Diagnostic Biomarkers**

- These biomarkers are used to detect or confirm the presence of a disease.
- They help in early identification or accurate diagnosis.
  Example: Elevated troponin levels are a diagnostic biomarker for myocardial infarction (heart attack).

Prognostic Biomarkers

- These biomarkers provide information about the likely course, outcome, or progression of a disease.
- They predict disease severity, recurrence, or survival.
  Example: HER2 overexpression in breast cancer is a prognostic biomarker indicating aggressive tumour behaviour and poorer prognosis.