1. Database

a. [5 pts] In relational databases, what is the difference between a primary key and a foreign key?

Primary keys are unique to each table and aim to identify rows; a foreign key is a primary key of another table predominantly used to join tables

b. [5 pts] How can foreign keys be used to join tables?

A foreign key in one table is the primary key of another table; tables are joined by matching the values of the foreign key with the values in the table where the column is a primary key.

2. Genomics

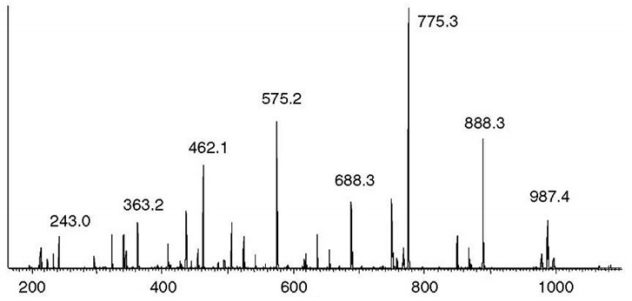
a. [5 pts] In sequencing, what do we mean by “read coverage”?

Number of reads mapped to a region

b. [5 pts] What is the main advantage of Third-generation sequencing technologies (i.e., Nanopore-based or Pacific Bioscience)?

Longer read lengths

3. [10 pts] A sample mass spectrum is shown below. What do X-axis and Y-axis of the mass spectrum denote?



* X = m/z (mass-to-charge ratio)
* Y = Relative abundance or Intensity

4. [10 pts] Circle all methods to identify protein structures

A. X-ray crystallography

B. NMR

C. Mass Spectrometry

D. Cryo-EM

E. SILAC

A, B and D

5. [10 pts] What is the typical number of single nucleotide polymorphism (SNP) 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

6. [10 pts] Why are FASTA and BLAST preferred to dynamic programming approaches to searching sequence databases?

Acceptable answers are:

FASTA and BLAST are faster by utilizing a hash table

(reduces time complexity, search space)

Dynamic programming is impractical to use due to O(n^2) complexity

7. [20 pts] Align the following two sequences using the Needleman-Wunsch global alignment algorithm. Show the complete dynamic programming matrix, and circle one optimal traceback on the matrix.

Sequence 1: ACTGCA

Sequence 2: ACATGA

Use the following scoring scheme in the score matrix:

Match: +2

Mismatch: 0

Gap: 0

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | C | T | G | C | A |
| A |  |  |  |  |  |  |
| C |  |  |  |  |  |  |
| A |  |  |  |  |  |  |
| T |  |  |  |  |  |  |
| G |  |  |  |  |  |  |
| A |  |  |  |  |  |  |

Answer (bottom up):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | C | T | G | C | A |
| A | 10 | 6 | 4 | 4 | 2 | 2 |
| C | 6 | 8 | 4 | 2 | 4 | 0 |
| A | 8 | 6 | 4 | 2 | 2 | 2 |
| T | 4 | 4 | 6 | 2 | 2 | 0 |
| G | 2 | 2 | 2 | 4 | 2 | 0 |
| A | 2 | 0 | 0 | 0 | 0 | 2 |

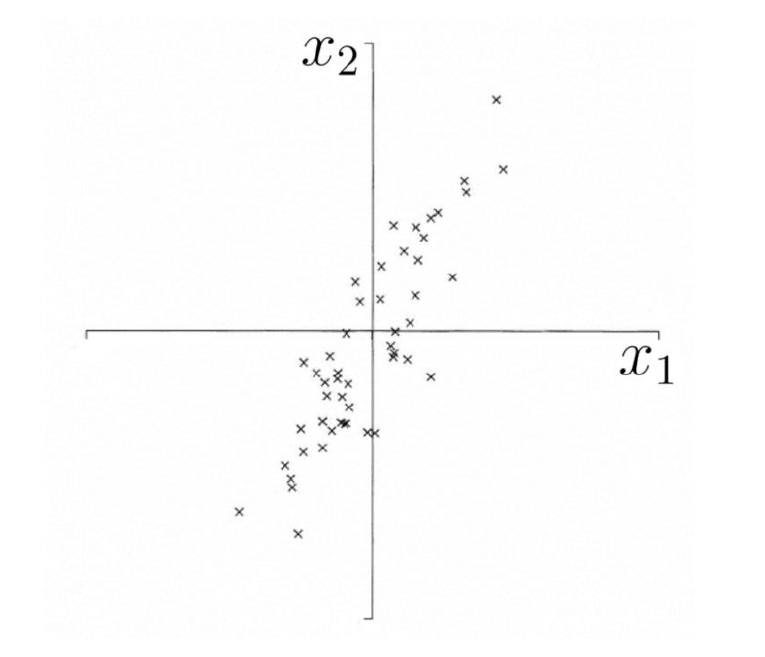
Alternatively (top down):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | C | T | G | C | A |
| A | 2 | 0 | 0 | 0 | 0 | 2 |
| C | 0 | 4 | 2 | 2 | 4 | 2 |
| A | 2 | 2 | 4 | 4 | 4 | 6 |
| T | 0 | 2 | 6 | 4 | 4 | 4 |
| G | 0 | 2 | 4 | 8 | 6 | 6 |
| A | 2 | 2 | 4 | 6 | 8 | 10 |

AC-TGCA

ACATG-A

8. [5 pts] Draw the two principal components (PC1, PC2) on the graph below.



9. [5 pts] In SVD, the data matrix A is decomposed as A = USV’. Suppose A is a 3\*18 matrix. What are the dimensions of U, S, V respectively?

U: 3 x 3

S: 3 x 18

V: 18 x 18

Also acceptable answer (based on lecture note):

U:3 x 18

S: 18 x 18

V: 18 x 18

10. Position Probability Matrix (PPM) is commonly used to represent motifs (patterns) in biological sequences.

a. [5 pts] Given the following DNA sequences, complete the corresponding position probability matrix profile.

DNA 1: GAGGTTGA

DNA 2: TCCGTTCA

DNA 3: CAGGTAGA

DNA 4: ACAGTTTA

DNA 5: TAGGTCAA

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pos 1 | Pos 2 | Pos 3 | Pos 4 | Pos 5 | Pos 6 | Pos 7 | Pos 8 |
| A | 0.2 | 0.6 | 0.2 | 0 | 0 | 0.2 | 0.2 | 1 |
| C | 0.2 | 0.4 | 0.2 | 0 | 0 | 0.2 | 0.2 | 0 |
| G | 0.2 | 0 | 0.6 | 1 | 0 | 0 | 0.4 | 0 |
| T | 0.4 | 0 | 0 | 0 | 1 | 0.6 | 0.2 | 0 |

b. [5 pts] Using the profile above, calculate the probability of the sequence S = GAGGTACA being observed.

0.00288

Or

288 \* 10^-5

Or

0.288 %

Or

9/3125