Candidate Number _____ Fraction _____

2023 Republic of China Biology Olympiad National Team Selection Camp Practical Test Questions

Third practical test

Bioinformatics Practice

Project	quantity
laptop (MEGA-11 program installed) Laptop	1 unit
desktop - Folder for storing bioinformatics practice data (containing...) mtCDNA SEQ.meg file) Laptop	1
desktop (folder containing student exam numbers)	1

Please note:

1. Please check if the exam number on the exam paper matches your exam number.
2. This exam paper (including cover and question paper) consists of 6 pages and must be returned in its entirety upon submission.
3. The total time allotted for answering the questions is 90 minutes.
4. Please answer the questions on this paper. You may write your answers on the back of the paper, but please indicate the question number.

Question 1: In

bioinformatics research, sequence alignment is a method of arranging DNA, RNA, or protein sequences.

The method of arranging columns to identify similar regions is one of the most basic tasks, and subsequent work on these region sequences...

Analysis of energy, structure, or evolutionary relationships. Methods for aligning nucleotide or amino acid bases have been developed, with

dot-matrix methods being a rapid and simple approach. Dot-matrix methods can be used to assess the similarity between two sequences.

Using a dot plot method, taking DNA sequences as an example, one sequence is arranged on the X-axis of the plot, and another sequence is arranged on the Y-axis.

One sequence is used to compare two sequences. When the bases of the two sequences match at the same positions in the diagram, the corresponding positions are plotted.

Set a point. The starting points of the two sequences on the two axes must be at the same point. If the sequence on the X-axis corresponds to the sequence on the Y-axis...

If the bases are the same, a point is plotted in this grid, as shown in the example below. Sequence 1 and Sequence CTTGCC are listed on the X-axis.

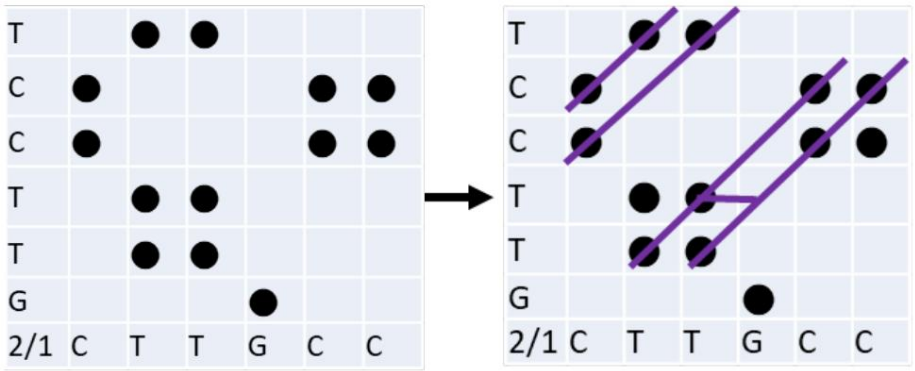
2. The GTTCCT sequence is plotted on the Y-axis, and a point is drawn at the corresponding base position on both the X and Y axes, forming the following structure.

fruit.

Once the points are drawn, connect neighboring points with 45-degree lines. The proximity of similar sequences will determine the diagonal and display.

The proximity of the connecting lines in a direct relationship is influenced by the differences between the two sequences, such as shifts.

The position is displayed by adding spaces when arranging the sequence.



The methods for permuting two sequences can be summarized into the following 5 possibilities:

```
1 CTTGCC- 1 ---CTTGCC 1----CTTGCC 1 CTTGCC 1 CTTGCC-
    || ||         | |          ||         || |      || ||
2 GTT-CCT 2 GTTCCT--- 2 GTTCCT---- 2 GTTCCT 2 GTT-CCT
```

The optimal arrangement is determined using a calculation method, as shown in the following formula:

- $D = s + wg$

s represents the number of base substitutions, g represents the total number of spaces, and w is the space length weighted score. Therefore, the following four permutations...

The calculations for the formulas are as follows:

```
1 CTTGCC- 1 ---CTTGCC 1----CTTGCC 1 CTTGCC 1 CTTGCC-
    || ||         | |          ||         || |      || ||
2 GTT-CCT 2 GTTCCT--- 2 GTTCCT---- 2 GTTCCT 2 GTT-CCT
```

D =1+1*2=2 D =1+3*6=19 D =0+4*8=32 D =3+0*0=3 D =1+1*2=3

Based on the results above, the optimal permutation of these two sequences is as follows:

```
1 CTTGCC-
    || ||
```

2 GTT-CCT because of calculation D The value is 2, which is the minimum value.

Given two sequences:

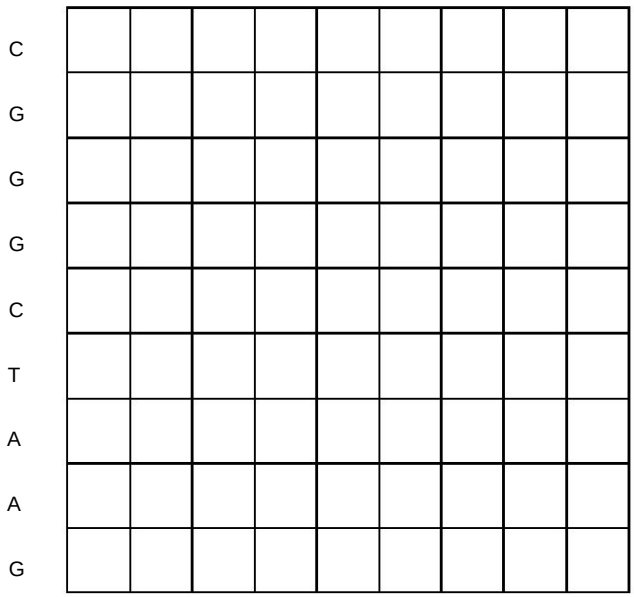
Sequence A: ATTCGGGCA

and Sequence B: GAATCGGGC,

align and arrange these two sequences to answer the following questions:

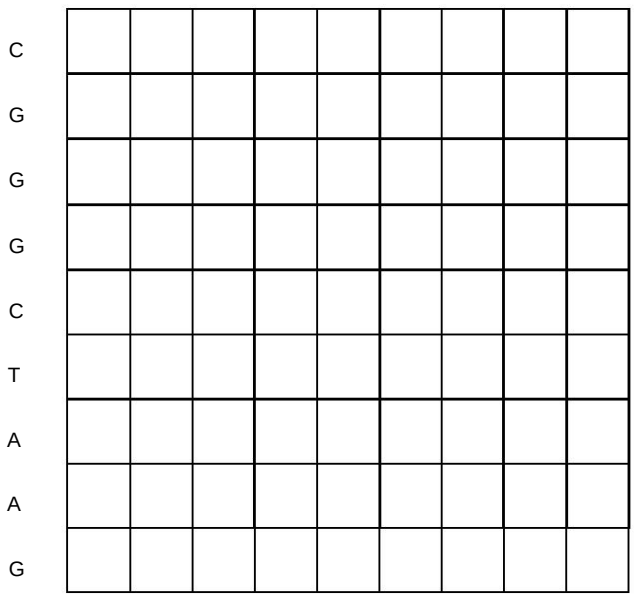
Question 1:

Using the raster chart method, draw points in the following image: (5 points)



B/A ATTCGGGCA

Question 2: Using the raster diagram from Question 1, connect neighboring points with 45-degree lines to draw all possible sequence arrangements: (5 points)



B/A ATTCGGGCA

Question 3

Using question 2, draw up the possible permutations of the sequences, as described in the question above, and organize the possible permutations of the two sequences.

Write down five possibilities: (10 points)

Question 4

Using the $D = s + wg$ The calculation method involves using this formula to calculate the possible permutations of the two sequences derived from Problem 3.

algorithm, s represents the number of base substitutions, g represents the total number of spaces, and w is the weighted score based on space length, for each possible permutation...

Each example needs to clearly show the calculation formula and the calculation process: (10 points)

Question 5

Based on the results above, write down the optimal permutation of these two sequences and the reasons why. (4 points)

Question 2:

DNA sequence analysis is a common method for exploring biological evolutionary history. The mtCDNA SEQ.meg file contains sequence fragments from seven primates: humans, chimpanzees, bonobos, gorillas, orangutans, Sumatran gibbons, and gibbons. Analyze this sequence data. Among the methods for constructing a phylogenetic tree, the unweighted pairing-group method using the arithmetic mean is used.

Arithmetic Means (UPGMA) is a commonly used and simple method for studying the similarity between different samples. It constructs a tree-like relationship graph by clustering the samples. The UPGMA method first clusters the two samples with the smallest distance together to form a new node, with its branch point located at half the distance between the two samples. Then, it calculates the average distance between this branch point and other samples, and finds the two smallest distances to cluster together again. This process is repeated until all samples are clustered together to form a complete tree-like relationship graph. However, the first step in constructing UPGMA requires building a genetic distance matrix. Therefore, please use the MEGA-11 program to perform the following analysis based on the given problem description.

Question 6:

The sequences of these 7 species total 3331 nucleotides in length. After testing, errors were found in the first 120 and last 131 nucleotides of the sequence, which need to be removed. After removing the erroneous segments, what percentage of the original total length is the length of the nucleotide chain of the erroneous segments? What percentage are conserved nucleotide sites? What percentage are variable nucleotide sites? What percentage are parsimony information nucleotide sites? What percentage are singleton nucleotide sites? (Hint: Save the new sequences to the folder with your exam number on your desktop, named "exam number-mtCDNA SEQ-1.MEG" and "exam number-mtCDNA SEQ-1.FAS".) (2 points each, 10 points total)

project	Answer
Example: Percentage of nucleotide chain length relative to the original total length	
Conserved (C) nucleotide site percentage example:	
Variable (V) percentage of nucleotide sites:	
Parsimony information (P) - Percentage of nucleotide sites (e.g.):	
Singleton (S) nucleotide site percentage example:	

Question 7

Obtain new sequence data for these 7 species, construct a genetic distance matrix using MEGA 11, set up a p-distance replacement model, and set the rate between loci to Gamma Distributed. Calculate the genetic distances between each pair of species and fill them in the table below. (1 point each, 21 points in total)

	Homo chimpanzee bonobo gorilla				orangutan Sumatran gibbon		
homo							
chimpanzee							
bonobo							
gorilla							
orangutan							
sumatran							
gibbon							

Question 8

The genetic distance matrix of these 7 species was obtained, and a dendritic relationship was constructed by cluster analysis using the arithmetic mean unweighted pairing method.

Based on the information in the previous question, identify which two species have the closest genetic distance and fill in the table below. (2 points each, 4 points in total)

Question 9

The clustering analysis and reconstruction of the genetic distance matrix are based on the arithmetic mean unweighted pairing method, continuing from the previous topic.

Find the two species with the closest genetic distance, merge them, and reconstruct a new genetic distance matrix. Enter the calculation results below.

The matrix contains species information and genetic distance information (the diagonally lined cells do not require answers). (1 point each, 21 points total)

point)

	1:	2:	3:	4:	5:	6:
1:						
2:						
3:						
4:						
5:						
6:						

Question 10

Based on the new DNA sequence data of these seven species, dendrograms were constructed using the MEGA 11 program's neighbor-joining method (NJ) and maximum likelihood method (ML), respectively.

In the analysis, the adjacent linkage method uses a p-distance substitution model, and the rate between sites is set to Gamma Distributed. The maximum likelihood rule requires selecting the most suitable nucleotide substitution model based on the BIC (Bayesian Information Criterion) score obtained from the nucleotide substitution model analysis. The analysis is repeated using the Bootstrap method, and the NJ method is performed 1000 times.

The ML method is run 100 times to obtain the bootstrap value of each branch, constructing NJ and ML tree diagrams, and then...

Bootstrap provides hand-drawn tree diagrams of NJ and ML below, along with a comparison and explanation of the differences between the two diagrams. (NJ tree 3 points, ML tree 3 points)

Tree (3 points, comparison and explanation (4 points))