INTERNATIONAL BIOLOGY OLYMPIAD e. V. IBO





35th International Biology Olympiad

7th - 14th July 2024 Astana, Kazakhstan

Practical Exam: Bioinformatics

Table of Content

General Instructions	G0-3
Important Information	G0-4
Equipment and Materials	G0-5
Appendix 1. Data format types	Q1-1
Appendix 2. Bioinformatics tools available in the application	Q1-4
Appendix 3. Standard genetic code table	Q1-6
Appendix 4. Multiple cloning sites of the pGLO-GFP-3UAG	Q1-7
Appendix 5. Definition of flanking sequence	Q1-8
Appendix 6. Methods for Statistical Ecology	Q1-9
Task I. Public Health	Q2-1
Task II. Gene Design	Q3-1
Task III. Statistical Ecology	Q4-1

Bioinformatics

IBO 2024 Kazakhstan

G0-2

English (Official)

General Instructions

- You have 90 minutes to complete the THREE tasks in this practical exam.
- You can do the tasks in any order.
- Task I: **Public Health (18 points)**
- Task II: **Gene Design (35 points)**
- Task III: **Statistical Ecology (25 points)**

Bioinformatics

IBO 2024 Kazakhstan

G0-3

English (Official)

Important Information

- No answers on the exam paper will be graded.
- You will use a dedicated browser application to perform the bioinformatics analyses and enter your responses. This application includes four different types of screens, namely question screen, “Notepad”, “Input”, and “Tools”.
- **You do not need to submit the exam. It will be submitted automatically once the exam is over. There will be a timer showing the time remaining until the end of the exam.**
- You **MUST PRESS “Save”** after completing EACH question or else your answers will not be registered. After the end of the exam all unsaved responses will not be submitted.
- You **MUST NOT** quit the "Safe Exam Browser".
- Questions that are fully or at least partially answered and saved are highlighted in green in the side panel. You can flag questions, and these will be highlighted in orange in the side panel. The current question is highlighted in blue.
- In questions, where you need to select more than 1 option a **NEGATIVE MARKING** is applied for each incorrect option. The lowest mark you can receive in such questions is 0.
- The “Reset” button will erase answers provided for the currently selected question. It will never erase answers provided for other questions. Using this button, you can resubmit your answers as many times as you want.
- The right click of the mousepad is disabled during the exam.
- You can search for specific text strings using “Ctrl” + “F”, but you must first click on the text field (the window within the window) to execute the search against the sequence.
- You can open multiple instances of the same tool or different tools and have several windows side-by-side (for instance, to compare the results of different analyses).
- You can copy the inputs and outputs by pressing "Ctrl" + "C". You can paste the copied data by pressing "Ctrl" + "V".
- You can select all the text inside the text box by pressing "Ctrl" + "A".
- You can switch between windows using “Alt” + “Tab”.
- If you minimize the electronic copy of the questions, you can quickly return to it by clicking the icon on the bottom left of the screen.
- To ensure proper formatting of the obtained analysis output you can maximize the window.
- In questions where you need to enter a number, you **MUST** use period "." to separate integer part from decimal part.
- You do not need to modify the obtained analysis output after you have pasted it into the answer box.
- If you face any technical issues with your computer, raise your card.
- If you get an error message after pressing “Submit” for any tool, your input is incorrect, so the assistants will not deal with such queries.

- Use the following cards to ask for drinking water/ restroom/ technical help/ other queries.

	Drinking water
	Restroom
	Technical issues
	Other queries

- No paper, materials, or equipment should be taken out of the venue.
- Stop answering as soon as you hear the stop signal at the end of the exam

Equipment and Materials

Please check that the following items are provided for you otherwise raise your hand immediately:

Name	Quantity
Laptop	1 piece
Mouse	1 piece
Blue card	1 piece
Yellow card	1 piece
Purple card	1 piece
Red card	1 piece
Watch	1 piece
Lamp	1 piece

Appendix 1. Data format types

"FASTA"

The "fasta" format is a format used to store DNA/RNA/protein sequences.

Each sequence has a label (header) that starts with a ">" sign. The actual sequence starts on the next line and goes on until the next label or the end of the file. An example is shown below.

>Examp1

AGTCGATCGACTAGCATCAGC

CACTACGTCAGCAT

>Examp2

AGTCGATGCACTAGCATCAGCCACTA

Note: Usually only four letters are used to represent both RNA and DNA sequences: A, C, G, and T, as 'T' stands for both thymine and uracil. For amino acid sequences, single-letter codes are used (see Appendix 3).

"CLUSTAL"

Below is an example of a sequence alignment using the "clustal" format (Figure 1).

```
CLUSTAL X (1.81) multiple sequence alignment


Examp3                          ----------------------CCACTCGGTCTCATGCCAGTGAAAAATC
Examp1                          --------AGAATCCAAACTGACGACTTAGTATGACTC--------AGGC
Examp2                          GATTGGGCTGCGGCTAGTTGAACCATTCAGTCTTAGTTCATTG---AATC
                                          * * *  ** * *          *  *

Examp3                          CTCCGCATCATGCACACTGATCGATGCCAGTCACCTTCGTCGGTATA---
Examp1                          TTTAGGAAGGAGTGCTCGGCT----------AGCTCCGCCGATGGCGGA
Examp2                          CGTCGTATGACGCA----------------AGCCGCGTC---------
                                   * *    *                 * *  ** *
```

Figure 1. "CLUSTAL X (1.81)" multiple sequence alignment

The first line is a header. It specifies the alignment format. The header is followed by sequence blocks (in the example above there are two blocks). Each block has the sequence labels followed by the corresponding sequences. In this example, there are three sequences labeled "Examp1", "Examp2" and "Examp3".

The dashes "-" in the sequence indicate alignment gaps resulting from insertions or deletions during the evolution of the analyzed sequences. Beneath the last sequence, there is a line with "*" symbols indicating positions conserved across all analyzed sequences

"NEWICK"

Below is an example of a tree expressed from the "newick" format, followed by a graphical representation of the tree it describes (Figure 2):

Input "((Examp2:0.30000,Examp1:0.30000):0.02250,Examp3:0.32250):0.00000";

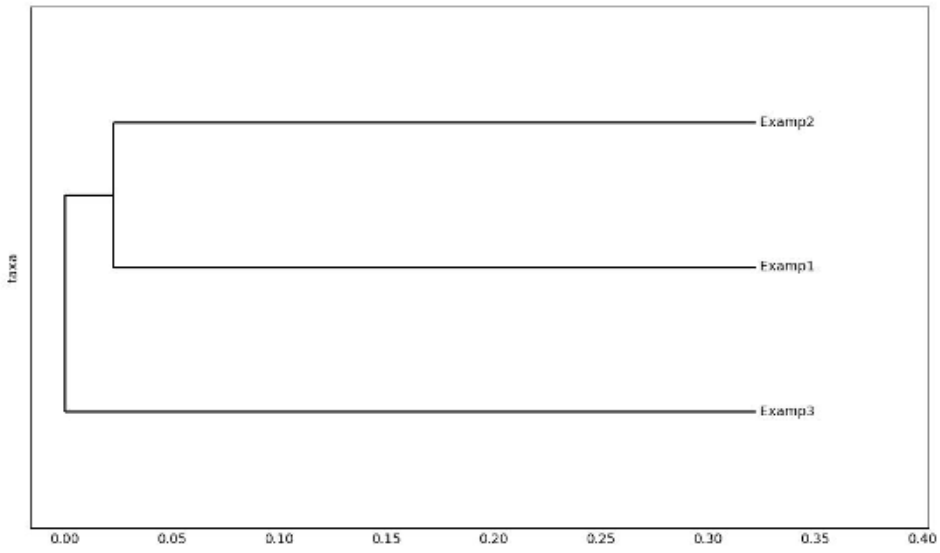


Figure 2. Taxonomical tree representation from the "newick" input.

In the "newick" format, characters to the left of the colon ":" symbol are node or leaf labels; numbers to the right of the colon ":" symbol indicate the length of the corresponding branch and brackets "()" are used to group leaves and nodes into branches; the tree ends with a semi-colon symbol ";".

"CSV"

Comma-separated values "CSV" is a text file format that uses commas to separate values, and newlines to separate records. Below is an example of a "CSV" file format

examp1, examp2, examp3

1,2,3

4,5,6

7,8,9

The first row represents column names. Each new row represents a new row of data, separated by a comma. This way, the examp1 column has values 1, 4, and 7, the examp2 column has values 2, 5, and 8, and the examp3 column has 3, 6, and 9

"PNG"

Portable Network Graphics "PNG" is a raster graphics file format that supports lossless data compression. Below is an example of a "PNG" file (Figure 3.)

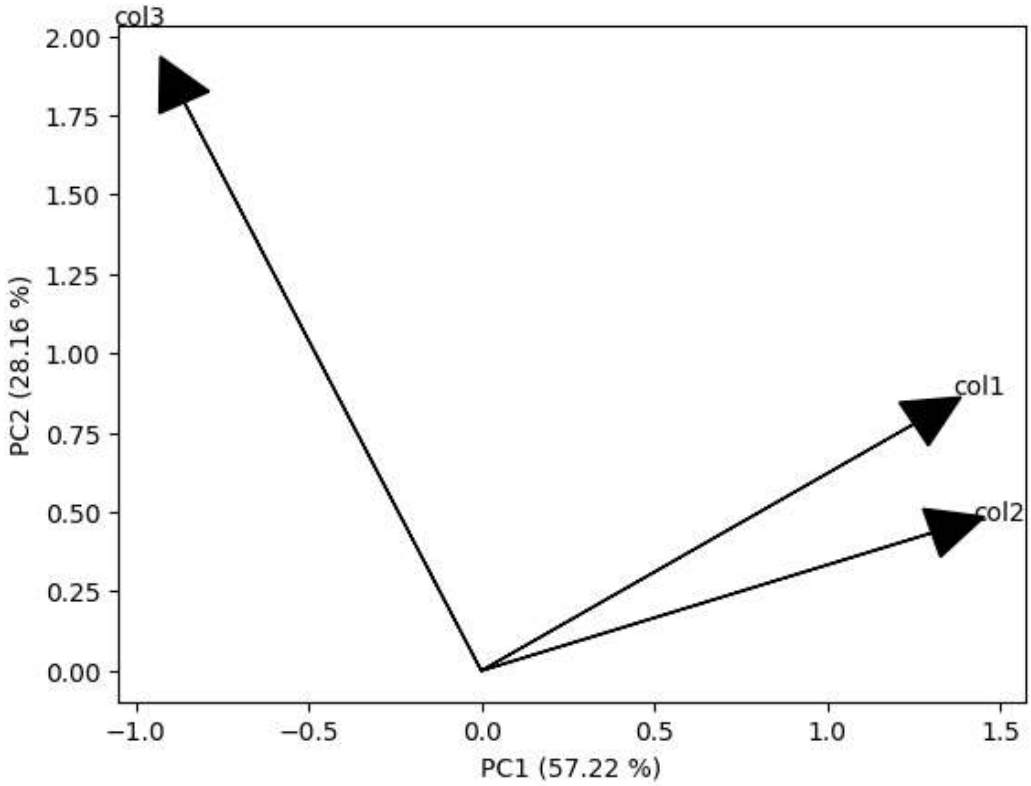


Figure 3. An example of a "PNG" file.

The axis labels, meaning, and legend are subject to change depending on the type of test and input provided.

Appendix 2. Bioinformatics tools available in the application

Tool	Description	Input	Output
"1. Codon Alignment"	Align two or more nucleotide sequences by codons	1. A protein alignment ("fasta") 2. Corresponding nucleotide codons sequences ("fasta")	Nucleotide sequences with only aligned codons shown
"2. DNA to protein"	Translate DNA nucleotide sequence to protein amino acid sequence	Nucleotide sequence (text)	Encoded protein amino acid sequence
"4. Restriction mapper"	Find restriction sites in a nucleotide sequence	Nucleotide sequence (text)	Number of cut sites for the corresponding restriction enzymes
"5. Reverse Complement"	Returns the reverse complement of a nucleotide sequence	Nucleotide sequence (text)	Reverse complement nucleotide sequence
"6. Sequence Editor"	Returns the length, start and end position of a selected subsequence	Nucleotide sequence (text)	Length, start and end position of a selected subsequence
"7. Sequence Alignment"	Performs alignment of two or more nucleotide / protein sequences. IMPORTANT NOTE: Do not paste the result of this tool into notepad, and use this tool in maximized window only	A set of at least two nucleotide / protein sequences ("fasta")	Aligned sequences ("Clustal")
"8. Tm Calculator"	Calculates melting temperature for a nucleotide sequence	Nucleotide sequence (text)	Melting temperature in degrees Celsius

"9. Tree Builder"	Builds a phylogenetic tree based on a nucleotide / protein sequence alignment	Nucleotide/protein alignment ("Clustal")	Phylogenetic tree ("newick"+image)
"10. Chi-square test"	Performs chi-square test on the dataset	Frequency values separated by comma	Chi-square value, degrees of freedom, p-value
"11. Reverse translation"	Performs reverse translation on the amino acid sequence. For the sake of simplicity, one amino acid is mapped to only one codon.	Amino acid sequence (text)	Nucleotide sequence (text)
"12. Codon optimization"	Changes codons in the sequence provided	Nucleotide sequence (text)	Optimized nucleotide sequence (text)
"13. Alpha diversity"	Performs alpha diversity analysis on dataset. Available methods are: "observed_otus", "shannon", "simpson", "simpson_e"	"CSV" with integers, **including column names and IDs**	"PNG" plot (x-axis IDs, y-axis corresponding index)
"14. NMDS"	Performs NMDS analysis on dataset	"CSV" with integers, **including column names and IDs**	"PNG" plot (x-axis "NMDS1", y-axis "NMDS2", numbers represent IDs of input)
"15. PCA"	Performs PCA analysis on dataset	"CSV" with integers, **including column names and IDs**	"PNG" plot (x-axis "PC1" with percentage, y-axis "PC2" with percentage)
"16. PCoA"	Performs PCoA analysis on dataset	"CSV" with integers, **including column names and IDs**	"PNG" plot (x-axis "PC1" with percentage, y-axis "PC2" with percentage, numbers represent IDs of input)

Appendix 3. Standard Genetic Code table

Below is the Standard Genetic Code table.

	U	**C**	**A**	**G**	
U	UUU - Phe	UCU - Ser	UAU - Tyr	UGU - Cys	**U**
	UUC - Phe	UCC - Ser	UAC - Tyr	UGC - Cys	**C**
	UUA - Leu	UCA - Ser	UAA - Stop	UGA - Stop	**A**
	UUG - Leu	UCG - Ser	UAG - Stop	UGG - Trp	**G**
C	CUU - Leu	CCU - Pro	CAU -His	CGU -Arg	**U**
	CUC - Leu	CCC - Pro	CAC -His	CGC -Arg	**C**
	CUA - Leu	CCA -Pro	CAA -Gln	CGA -Arg	**A**
	CUG - Leu	CCG -Pro	CAG -Gln	CGG -Arg	**G**
A	AUU - Ile	ACU - Thr	AAU - Asn	AGU - Ser	**U**
	AUC - Ile	ACC - Thr	AAC - Asn	AGC - Ser	**C**
	AUA - Ile	ACA - Thr	AAA - Lys	AGA - Arg	**A**
	AUG - Met	ACG - Thr	AAG - Lys	AGG - Arg	**G**
G	GUU - Val	GCU - Ala	GAU - Asp	GGU - Gly	**U**
	GUC - Val	GCC - Ala	GAC - Asp	GGC - Gly	**C**
	GUA - Val	GCA - Ala	GAA - Glu	GGA - Gly	**A**
	GUG - Val	GCG - Ala	GAG - Glu	GGG - Gly	**G**

Below is the table of 3- and 1-letter amino acid codes.

Amino acid	3-letter code	1-letter code	Amino acid	3-letter code	1-letter code
Glycine	Gly	G	Proline	Pro	P
Alanine	Ala	A	Valine	Val	V
Leucine	Leu	L	Isoleucine	Ile	I
Methionine	Met	M	Cysteine	Cys	C
Phenylalanine	Phe	F	Tyrosine	Tyr	Y
Tryptophan	Trp	W	Histidine	His	H
Lysine	Lys	K	Arginine	Arg	R
Glutamine	Gln	Q	Asparagine	Asn	N
Glutamate	Glu	E	Aspartate	Asp	D
Serine	Ser	S	Threonine	Thr	T

Appendix 4. Multiple cloning site of the pGLO-GFP-3UAG.

The figure below presents the Multiple Cloning Site (MCS) for the plasmid pGLO-GFP-3UAG (Figure 4). The number on the right represents the nucleotide number. Above the sequence are the names of the restriction sites.



Figure 4. Multiple cloning site of pGLO-GFP-3UAG (GFP - Green Fluorescent Protein; MCS - Multiple Cloning Site).

Appendix 5. Definition of flanking sequence

The flanking region is a region of DNA that is adjacent to either 5' or 3’ end of the gene. This sequence can vary in length. Figure 5 shows the graphical representation of flanking regions in the DNA sequence. As a general rule 6 base pairs should be added on either side of the recognition site to cleave efficiently. The extra bases should satisfy 2 rules:

1. No primer dimers
2. No palindrome sequences

In this exam please use the following flanking sequence: AAACCC

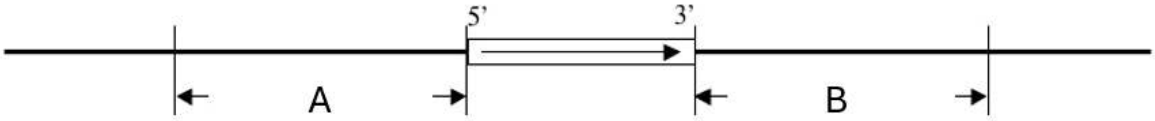


Figure 5. Graphical representations of flanking regions (A - 5’ flanking region; B - 3’ flanking region)

Appendix 6. Methods for Statistical Ecology

Method:	Definition
Observed Diversity Index • method name in the system: "observed_otus"	The Observed Diversity Index, also known as species richness, is a measure of biodiversity that simply counts the number of different species present in a given area or sample. It is one of the simplest measures of diversity and does not take into account the abundance of each species
Shannon Diversity Index (Modified) • method name in the system: "shannon"	The Shannon Diversity Index, named after Claude Shannon, is a measure of biodiversity that takes into account both the richness (number of different species) and the evenness (relative abundance) of the species present in a sample or area. It provides a more comprehensive measure of diversity compared to species richness alone. Increase of Shannon diversity index usually reflects higher species richness and evenness.

Method:	Definition
Simpson Diversity Index • method name in the system: "simpson"	The Simpson Diversity Index, named after Edward H. Simpson, is another measure of biodiversity that considers both richness and evenness. However, unlike the Shannon index, Simpson's index gives more weight to the dominant species in a community. A higher Simpson index usually indicates lower diversity.
Inversed Simpson Diversity Index • method name in the system: "simpson_e"	The Inversed Simpson Diversity Index is the reciprocal of Simpson's Diversity Index. It provides a measure of the probability that two individuals randomly selected from a sample will belong to different species.

Method:	Definition
Non-metric Multidimensional Scaling (NMDS)	NMDS is a method used in ecology to visualize the similarity or dissimilarity of samples based on their species composition. It is a dimensionality reduction technique that transforms high-dimensional data (e.g., species abundances across different samples) into a lower-dimensional space while while trying to preserve original distances using a heuristic method. NMDS is particularly useful for exploring and visualizing complex ecological datasets. The greater the distance between points, the higher the dissimilarity between them.

Method:	Definition
Principal Component Analysis (PCA)	PCA is a statistical method commonly used for dimensionality reduction and exploratory data analysis of continuous data. It transforms the dataset into principal components (PC), often followed by the percent values. Percent values indicate the contribution of each PC to variation. PCA can be used to identify patterns and relationships in ecological datasets, such as gradients or clusters of species or ecological factors
Principal Coordinates Analysis (PCoA)	PCoA, also known as metric multidimensional scaling (MDS), is a technique similar to PCA but specifically designed to analyze dissimilarity or distance matrices of discrete (categorical) data. Like PCA, PCoA identifies patterns and relationships in ecological data but focuses on the similarities or dissimilarities between samples rather than the original variables themselves. It is often used to visualize and explore patterns of species composition or community structure across different sites or treatments

Task I. Public Health

Nowadays, bioinformatics is one of the integral research fields that is capable of advancing biomedical research, medicine, healthcare, and public health. Bioinformatics has a variety of applications in the public health sector. Among them, statistical analysis of patients nationwide is one of the most general usages.

In this task, you are provided with a fictional dataset of patients from a small village in Kazakhstan. Your main goal for this task is to find if there is any dependency between the sugar level in the blood of the patients and myocardial infarction.

The dataset for this task can be found by clicking on the “Input” dropdown menu and selecting the input labeled as “patients.csv”. This file contains information about 68 patients and involves fields like “ID, blood sugar level, blood pressure, height, weight, and chronic condition”. In public health research, one of the purposes is to identify the potential risk group. For this reason, for each patient in your dataset, a small fictional motif sequence of geneA is provided for you in the “geneA.fasta”. To see the sequence, you can copy-paste the provided data into the "Notepad" page of the application and edit it appropriately.

In order to correctly adjust your raw dataset and select the right statistical method, you need to identify your null and alternative hypotheses correctly.

Q.1.1.1 Select the sentence that is most appropriate to describe the null hypothesis. 1.0pt

A) There is no association between blood sugar level and myocardial infarction
B) There is association between blood sugar level and myocardial infarction

Q.1.1.2 Select the sentence that is most appropriate to describe the alternative hypothesis. 1.0pt

A) There is no association between blood sugar level and myocardial infarction
B) There is association between blood sugar level and myocardial infarction

Now copy your raw dataset to "Notepad" and use the "Notepad" page to adjust your raw dataset.

Sometimes in public health datasets, some of the patients do not provide all of the data asked by a doctor; consequently, some of the fields in the dataset are labeled as “NA” (which means not available). Remove the patients with “NA” values from the dataset.

Q.1.2 How many patients are retained in the dataset? 1.0pt

Since you are interested in studying the dependency between the sugar level in the blood of the patients and myocardial infarction, you need to find two groups: 1) people suffering from myocardial infarction; and 2) a control group. In this task, control group are healthy patients.

Q.1.3.1 How many patients are suffering from myocardial infarction? 1.0pt

Q.1.3.2 How many patients are in the control group? 1.0pt

The high sugar level in the blood is considered to be above 100 mg/dL. Using the information above count the number of people in each group.

	Q.1.4.1 How many patients have normal blood sugar levels among people with myocardial infarction?	1.0pt

	Q.1.4.2 How many patients have high blood sugar levels among people with myocardial infarction?	1.0pt

	Q.1.4.3 How many patients have normal blood sugar levels in the control group?	1.0pt

	Q.1.4.4 How many patients have high blood sugar levels in the control group?	1.0pt

You are working as a group with your colleagues. After you combine your previous results with theirs, you ended up with having new data:

- group 1 (people with myocardial infarction with normal blood sugar level) = 80
- group 2 (people with myocardial infarction with high blood sugar levels) = 90
- group 3 (control group with normal blood sugar level) = 50
- group 4 (control group with high blood sugar levels) = 30

Now, you need to check the dependency between the blood sugar level and myocardial infarction. For this, you need to perform a chi-square test. Adjust the dataset and enter it into the input fields of the "#10. Chi-square" tool. Into the first row, enter the values for the case group, into the second row enter the control group. Values should be separated by a comma.

	Q.1.5.1 What is your chi-square value? (Round to two decimal places)	1.0pt

	Q.1.5.2 How many degrees of freedom does this chi-square test has?	1.0pt

	Q.1.5.3 What is your p-value for this dataset? (Round to two decimal places)	1.0pt

Now, you can decide whether you reject or fail to reject your null hypothesis.

	Q.1.6.1 [CANCELED]	1.0pt

Q.1.6.2 Make the correct interpretation of your chi-square results at the 5% significance value. Fill out the following text blanks: 3.0pt

Since _____(Q1.6.2.1) of the test is _____(Q1.6.2.2) than _____(Q1.6.2.3) , we _____(Q1.6.2.4). This means we _____(Q1.6.2.5) have sufficient evidence to say that there is an association between _____(Q1.6.2.6) and myocardial infarction.

Q1.6.2.1:

A) the chi-square value
B) the p-value
C) the df value

Q1.6.2.2:

A) less
B) more

Q1.6.2.3:

A) 0.05
B) 0.1
C) 0.25
D) 0.5
E) 1

Q1.6.2.4:

A) reject the null hypothesis
B) fail to reject the null hypothesis

Q1.6.2.5:

A) do
B) do not

Q1.6.2.6:

A) blood sugar level
B) blood pressure
C) height
D) weight

In our research, let's consider that geneA together with increased levels of sugar definitely contributes to the risk of myocardial infarction. We want to do a prophylaxis of myocardial infarction in the control group. Assuming a dependency has been found between high blood sugar levels and myocardial infarction, we can align the DNA sequences of geneA of people suffering from myocardial infarction with high blood sugar level using the "#7. Sequence Alignment" tool. The data is provided in the "geneA.fasta" input file, where the number at the end of the sequence name is the ID of the patient. For example, Sequence_1 is the geneA of patient 1. It will return an alignment in the "clustal" format (for more about it see Appendix 1).

Q.1.7 Enter the longest continuous segment of geneA that is the same in all people suffering from myocardial infarction with high blood sugar level. 2.0pt

Now, using the obtained sequence segment, identify which person/people in the control group, who has/have high sugar levels in their blood, is/are at risk of developing myocardial infarction. You can perform additional analyses of these sequences using the available tools if needed.

Q.1.8 Select the ID(s) of a person/people at the risk of developing myocardial infarction if any exists. 1.0pt

Task II. Gene Design

UniProt is a comprehensive and freely accessible database of protein sequences and functional information. It stands for Universal Protein Resource and serves as a centralized repository that integrates data from various sources. This database contains the protein sequence for human insulin, which is labeled as P01308.

In this task, you are provided with the proinsulin sequence. Your main goal for this task is to obtain two separate genetic vectors (for insulin A and insulin B) that could be transformed in *Escherichia coli* cells, and further used for the manufacture of human insulin.

The insulin sequence can be found by clicking on the “Input” dropdown menu and selecting the input labeled as “proinsulin.fasta”.

Q.2.1 What is the length of the proinsulin molecule? (unit: amino acids)	1.0pt

Only functional insulin conformation is used for the bacterial transformation process. This conformation is a product of two consecutive enzymatic reactions. The initial reaction hydrolyzes the peptide bond at the following region (-AAA | FVN-). At the end of this reaction, the heavier part remains, while the lighter one is degraded.

Q.2.2 Enter the remaining proinsulin sequence after the end of the first reaction	1.0pt

The second reaction involves two enzymatic cleavages of the remaining peptide from Q2.2: one between the amino acids 30 and 31, and the other between the amino acids 55 and 56. This way the middle part of the sequence is removed. From the remaining parts, the lighter sequence is the A chain, while the heavier one is the B chain of the insulin.

Q.2.3.1.1 Enter the sequence of the insulin A chain.	1.0pt
Q.2.3.1.2 Enter the sequence of the insulin B chain.	1.0pt
Q.2.3.2.1 What is the length of the insulin A chain? (unit: amino acids)	1.0pt
Q.2.3.2.2 What is the length of the insulin B chain? (unit: amino acids)	1.0pt
Q.2.3.3 How many cysteine residues are there in chain A of insulin? (see Appendix 3)	1.0pt
Q.2.3.4 How many cysteine residues are there in chain B of insulin?	1.0pt
Q.2.3.5 "What's the maximum number of disulfide bonds an insulin molecule may form?	1.0pt
Q.2.3.6 What's the maximum number of disulfide bonds that may connect chains A and B of an insulin molecule?	1.0pt

Bioinformatics

Q2-2

English (Official)

For research purposes you had decided to use mouse model instead of humans. Run the "#11. Reverse Translation" tool on the "mouse_insulin_A.fasta" and "mouse_insulin_B.fasta". Use these inputs for the questions from Q.2.4.1.1 to Q.2.6.2

Q.2.4.1.1 Enter the DNA sequence of the mouse insulin chain A. 1.0pt

Q.2.4.1.2 Enter the DNA sequence of the mouse insulin chain B. 1.0pt

Q.2.4.2.1 What is the length of DNA sequence of mouse insulin A? (unit: nucleotides) 1.0pt

Q.2.4.2.2 What is the length of DNA sequence of mouse insulin B? (unit: nucleotides) 1.0pt

The DNA sequence we got is from mouse and is not very well suited for expression in another organism. First, we need to do codon optimization for a smooth expression in *E. coli*. Copy the DNA sequences of the insulin A and B chains and paste them in the "#12. Codon Optimizer" tool. Using this tool perform codon optimization from the *Mus musculus* sequence to the *E.coli* sequence.

Q.2.5.1 Enter the optimized DNA sequence of the insulin A chain 1.0pt

Q.2.5.2 Enter the optimized DNA sequence of the insulin B chain 1.0pt

You can perform additional analyses of these sequences using the available tools if needed

Q.2.6.1 Compared the DNA sequence encoding the mouse insulin chain A, the *E. coli* codon-optimized DNA sequence of insulin chain A (More than 1 option could be correct): 1.0pt

A) Remained unchanged
B) Changed at positions 1-20
C) Changed at positions 21-40
D) Changed at positions 41-till the end

Q.2.6.2 Compared to the DNA sequence encoding the mouse insulin chain B, the *E. coli* codon-optimized DNA sequence of insulin chain B DNA sequence (More than 1 option could be correct): 1.0pt

A) Remained unchanged
B) Changed at positions 1-20
C) Changed at positions 21-40
D) Changed at positions 41-60
E) Changed at positions 61-80
F) Changed at positions 81-till the end

Linear DNA is unlikely to be adopted by a target organism. In order to perform the transformation of *E. coli* cells, we need to insert the codon-optimized linear DNA sequence into a plasmid. In our case, we will use pGLO-GFP-3UAG, which is provided for you. The pGLO-GFP-3UAG sequence can be found by clicking on the “Input” dropdown menu and selecting the input labeled as “plasmid.fasta”.

Q.2.7 What is the length of the MCS unit in pGLO-GFP-3UAG? (unit: nucleotides) (see Appendix 4)	1.0pt

We will insert our insulin sequences into the Multiple Cloning Site (MCS). For that, we will use the Restriction Ligation cloning technique. While the most basic, it still requires some preparations and modifications to our insulin sequences. We will use 2 restriction enzymes to cut our plasmid: enzyme 1 and enzyme 2. Enzymes 1 and 2 recognition sites start at the positions 2078 and 2093 respectively.

Q.2.8.1 What is the name of enzyme 1?	1.0pt

Q.2.8.2 What is the name of enzyme 2?	1.0pt

Fortunately, one of your colleagues was able to optimize the obtained insulin sequences even further. They are stored in the "optimized_insulin_A.fasta" and "optimized_insulin_B.fasta". From question Q.2.9.1 to the end of the exam use these inputs.

Now we need to design primers that will contain restriction sites of our optimized sequences. The restriction site sequences of enzymes 1 and 2 can be found by clicking on the “Input” dropdown menu and selecting the input labeled as “restriction.fasta”. The length of the primer which is complementary to the protein coding sequence, should be 20 nucleotides long. Additionally, for further optimization of the cutting efficiency of our sequence, we need to add a flanking sequence to the 5' end (see Appendix 5) before the restriction site. The flanking sequence to be used is AAACCC. Because we would need to express our protein, we also need to add a Ribosome Binding Site (AGGAGG) and a start codon (ATG) before the protein coding sequence to the forward primer.

Design the forward and reverse primers for chains A and B. All of the answers should be provided from 5' to 3'. Notes about the primers:

- The melting temperatures of the primers do not need to match.
- The length of the primer can exceed the recommended 24 nucleotides.
- Normally, RBS and start codon are separated by a number of nucleotides, however for this task we are not doing that.

Q.2.9.1 Enter the sequence of forward primer for insulin A chain	2.0pt

Q.2.9.2 Enter the sequence of reverse primer for insulin A chain	2.0pt

Q.2.9.3 Enter the sequence of forward primer for insulin B chain	2.0pt

Q.2.9.4 Enter the sequence of reverse primer for insulin B chain	2.0pt

Now, identify the resulting sequences of pGLO-GFP-3UAG after the insertion of the target sequences

	Q.2.10.1 Enter the sequence of pGLO-GFP-3UAG after the insertion of DNA of insulin A chain (The answer should not contain the plasmid index).	3.0pt

	Q.2.10.2 Enter the sequence of pGLO-GFP-3UAG after the insertion of DNA of insulin B chain (The answer should not contain the plasmid index).	3.0pt

Task III. Statistical Ecology

You are an environmental researcher tasked to study the effect of climate change on the diversity of trees in Kazakhstan. For this, you counted the number of trees in different parts of Kazakhstan, recorded the conditions at which they are growing as well as general information about the climate of the region. You need to identify and compare the species distribution across all regions of Kazakhstan. The description of methods used for statistical ecology is provided in Appendix 6.

The fictional datasets for this task can be found by clicking on the "Input" dropdown menu and selecting the input labeled as "Kaz_trees_2023.csv", "Kaz_region_names_2023.csv" and "Kaz_factors_regions_2023.csv". ID values of "Kaz_trees_2023.csv" and "Kaz_factors_regions_2023.csv" indicate the same regions of Kazakhstan. Assume that the number inside the file name is the year of data collection.

Answer the following questions using the appropriate statistical tools provided.

Q.3.1 What is the most important environmental factor explaining the environmental differences between the regions of Kazakhstan based on the given data?	1.0pt
Q.3.2 Based only on the species richness, what is / are the location(s) with the highest diversity? (Label all that are applicable)	2.0pt
Q.3.3 Based on species richness and evenness, which are the four Kazakhstan regions with the highest diversity?	2.0pt
Q.3.4 Based on species richness and evenness, which are the two Kazakhstan regions with the lowest diversity?	2.0pt
Q.3.5 Based on species dominance, which region of Kazakhstan has the lowest diversity?	1.0pt
Q.3.6 Based on species dominance, which are the two Kazakhstan regions with the highest diversity?	2.0pt
Q.3.7 Which region(s) of Kazakhstan is the most distinct from the others based on its species composition?	3.0pt
Q.3.8 Mark as True or False: Based on the obtained results, Zhetysu, West-Kazakhstan, North-Kazakhstan, and Ulytau regions have the least difference in diversity.	1.0pt
Q.3.9.1 Which 3 regions of Kazakhstan have the closest diversity to the Akmola region?	2.0pt
Q.3.9.2 Which 2 regions of Kazakhstan have the closest diversity to the Aktobe region?	2.0pt

Bioinformatics

Q3-2

English (Official)

You receive a government request to study the diversity of trees in the Karagandy region. You are given the dataset containing the data about the count of trees in the area dated to 1971. However, in 2021 Karagandy region was split into two regions, Karagandy region and Ulytau region.

The fictional datasets for this task can be found by clicking on the "Input" dropdown menu and selecting one of the inputs labeled as:

- "karagandy_trees_1971.csv",
- "karagandy_factors_1971.csv",
- "karagandy_factors_2023.csv",
- "karagandy_factors_2023_names.csv".

You can perform additional analyses of this dataset using the available tools if needed.

Q. 3.10.1 Considering the species richness, how has the diversity of the old Karagandy region territory changed over the last 50 years? 1.0pt

A) Increased
B) Decreased
C) Unchanged

Q.3.10.2 Considering the species dominance, how has the diversity of the old Karagandy region territory changed over the last 50 years? 1.0pt

A) Increased
B) Decreased
C) Unchanged

Q.3.10.3 Considering the species richness and evenness, how has the diversity of the old Karagandy region territory changed over the last 50 years? 1.0pt

A) Increased
B) Decreased
C) Unchanged

Q. 3.10.4 Based on the given data, what is the most important environmental factor explaining the differences between the studied sites in Ulytau in 2023? 1.0pt

Q.3.10.5 Based on the given data, what is the most important environmental factor explaining the differences between the studied sites in Karagandy in 2023? 1.0pt

Q.3.10.6 Based on the given data, what is the most important environmental factor explaining the differences between the studied sites in Karagandy in 1971? 1.0pt

Q.3.10.7 Mark as True or False: The same environmental factor is the most important predictor of the environment in all regions studied in Q3.10.1 - Q3.10.6 both in 2023 and in 1971. 1.0pt

Task1.

questions:	**max points:**	**question type:**	**Marking and correct answers:**	**tool:**
Q1.1.1	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: A	- Knowledge-based question - No tools needed
Q1.1.2	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: B	- Knowledge-based question - No tools needed
Q1.2	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 57	- Notepad - visual analysis - “Ctrl + F” to remove all “NA” values
Q1.3.1	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 25	- Notepad - Visual analysis - “Ctrl + F” to find all “myocardial infarction” values
Q1.3.2	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 18	- Notepad - Visual analysis - “Ctrl + F” to find all “No” values
Q1.4.1	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 9	- Notepad - Visual analysis - Count the number of people with normal blood sugar levels within people with myocardial infarction

Q1.4.2	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 16	- Notepad - Visual analysis - Count number of people with high blood sugar levels within people with myocardial infarction
Q1.4.3	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 13	- Notepad - Visual analysis - Count number of people with normal blood sugar levels within control group
Q1.4.4	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 5	- Notepad - Visual analysis - Count number of people with high blood sugar levels within control group
Q1.5.1	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 4.59-4.60	- enter the corresponding numbers of all 4 groups to “Chi-square test” - chi-square value rounded
Q1.5.2	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 1	- “Chi-square test” - find df value
Q1.5.3	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 0.03	- “Chi-square test” - find p-value rounded
Q1.6.1	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: A	- Analysis of results from the “Chi-square” tool - select correct answer - No extra tool needed

Q1.6.2.1	0.5	multiple choice with 1 answer	0.5 point for correct answer 0 points for incorrect correct answer: B	- Analysis of results from the “Chi-square” tool - select correct answer - No extra tool needed
Q1.6.2.2	0.5	multiple choice with 1 answer	0.5 point for correct answer 0 points for incorrect correct answer: A	
Q1.6.2.3	0.5	multiple choice with 1 answer	0.5 point for correct answer 0 points for incorrect correct answer: A	
Q1.6.2.4	0.5	multiple choice with 1 answer	0.5 point for correct answer 0 points for incorrect correct answer: A	
Q1.6.2.5	0.5	multiple choice with 1 answer	0.5 point for correct answer 0 points for incorrect correct answer: A	
Q1.6.2.6	0.5	multiple choice with 1 answer	0.5 point for correct answer 0 points for incorrect correct answer: A	
Q1.7	2	essay type	2 points for correct answer 0 points for incorrect correct answer: TATAGGCACGGGGAAGTA	- Sequence alignment - Input:people with myocardial infarction and high blood sugar level - select the aligned region

Q1.8	1	multiple choice with multiple answers	1 point for 2 correct IDs 0.5 points for 1 correct ID negative 0.7 points for each incorrect ID the lowest mark is 0 correct answer: ID23, ID50	- Sequence alignment - Input:people from control group with high blood sugar level and aligned region (answer from the previous question) - find the IDs of people with that region in control group

Task2.

questions:	**max points:**	**question type:**	**Marking and correct answers:**	**tool:**
Q2.1	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 100	- “Sequence editor” - find the length of proinsulin
Q2.2	1	essay type	1 point for correct answer 0 points for incorrect correct answer: FVNQHLCGSHLVEALYLVCGERGFFYTPKT RREAEDGPGAGSLQPLALEGSLQKRGIVEQ CCTSICSLYQLENYCN	- Notepad or sequence editor - remove the corresponding region - enter the remaining sequence
Q2.3.1.1	1	essay type	1 point for correct answer 0 points for incorrect correct answer: GIVEQCCTSICSLYQLENYCN	- Notepad or sequence editor - remove the corresponding region - enter the remaining sequence
Q2.3.1.2	1	essay type	1 point for correct answer	- Notepad or sequence editor

			0 points for incorrect correct answer: FVNQHLCGSHLVEALYLVCGERGFFYTPKT	- remove the corresponding region - enter the remaining sequence
Q2.3.2.1	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 21	- “Sequence editor” - find the length of insulin A chain
Q2.3.2.2	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 30	- “Sequence editor” - find the length of insulin B chain
Q2.3.3	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 4	- Notepad - “Ctrl + F” to find cysteine - use Appendix 3 - count cysteine residues in insulin A chain
Q2.3.4	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 2	- Notepad - “Ctrl + F” to find cysteine - use Appendix 3 - count cysteine residues in insulin B chain
Q2.3.5	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 3	- might use Notepad - Knowledge-based - No tool needed
Q2.3.6	1	Numerical	1 point for the exact number 0 points for any other number	- might use Notepad - Knowledge-based - No tool needed

			correct answer: 2	
Q2.4.1.1	1	essay type	1 point for correct answer 0 points for incorrect correct answer: GCACTTGTAGAGCAATGTTGTACAAGCCTT TGTAGCATCTATCAAATCGAGAACTATTGTA AC	- Reverse translation - input the protein sequence of insulin A chain - select the DNA sequence from the tool
Q2.4.1.2	1	essay type	1 point for correct answer 0 points for incorrect correct answer: TTTGTAAACCAACATATCTGTGCAAGCCAT ATCGTAGAGGGTATCTATATCGTATGTGCA GAGAGGGCATTTTTTTATACACCTAAGACA	- Reverse translation - input the protein sequence of insulin B chain - select the DNA sequence from the tool
Q2.4.2.1	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 63	- Sequence editor - count the length of the insulin A chain
Q2.4.2.2	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 90	- Sequence editor - count the length of the insulin B chain
Q2.5.1	1	essay type	1 point for correct answer 0 points for incorrect correct answer: GCCCTGGTTGAACAGTGTTGTACGAGTTT GTGCTCCATCTATCAGATCGAAAACTACTG TAACTGA	- Codon optimization - Input: answer from reverse translation - Enter the optimized sequence of insulin A chain

Q2.5.2	1	essay type	1 point for correct answer 0 points for incorrect correct answer: TTTGTGAATCAGCATATTTGTGCAAGTCATA TTGTTGAAGGCATCTACATTGTGTGCGCC GAACGCGCGTTTTTCTACACCCCGAAAAC CTAA	- Codon optimization - Input: answer from reverse translation - Enter the optimized sequence of insulin B chain
Q2.6.1	1	multiple choice with multiple answers	1 point for 3 correct answers 0.333 points for each correct answer negative 1 point for the incorrect answer the lowest mark is 0 correct answer: B, C, D	- Sequence alignment - input: initial and optimized insulin A chains - fing regions with optimized nucleotides
Q2.6.2	1	multiple choice with multiple answers	1 point for 5 correct answers 0.2 points for each correct answer negative 1 point for the incorrect answer the lowest mark is 0 correct answer: B, C, D, E, F	- Sequence alignment - input: initial and optimized insulin B chains - fing regions with optimized nucleotides
Q2.7	1	Numerical	1 point for the exact number 0 points for any other number correct answer: 57	- Sequence editor or visual analysis - use Appendix 4
Q2.8.1	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: Acc65I	- use Appendix 4 - no tool needed
Q2.8.2	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect	- use Appendix 4 - no tool needed

			correct answer: XbaI	
Q2.9.1	2	essay type	2 point for correct answer 0 points for incorrect correct answer: AAACCCGGTACCAGGAGGATGGCACTGGTTGAACAATGTTG	- Notepad or Sequence editor - make the corresponding primer based on the description + flanking region - use Appendix 5
Q2.9.2	2	essay type	2 point for correct answer 0 points for incorrect correct answer: AAACCCTCTAGATCAATTGCAGTAGTTCTCGA	
Q2.9.3	2	essay type	2 point for correct answer 0 points for incorrect correct answer: AAACCCGGTACCAGGAGGATGTTCGTGAACCAACATATTTG	
Q2.9.4	2	essay type	2 point for correct answer 0 points for incorrect correct answer: AAACCCTCTAGATTAAGTTTTCGGCGTGTAGA	
Q2.10.1	3	essay type	3 point for correct answer 0 points for incorrect correct answer:	- Notepad or Sequence editor - cut the region between Acc651 and XbaI - enter the corresponding DNA chain (A

			at the end of the document	or B)
Q2.10.2	3	essay type	3 point for correct answer 0 points for incorrect correct answer: at the end of the document	

Task3.

questions:	**max points:**	**question type:**	**Marking and correct answers:**	**Tool**
Q3.1	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: C	- PCA - input: Kaz_factors_regions.csv - analyze graph using Appendix 6 - select the factor
Q3.2	2	multiple choice with multiple answers	2 point for 4 correct IDs 0.5 points for 1 correct ID negative 1 points for each incorrect ID the lowest mark is 0 Correct answers: ID 5, ID 6, ID 13, ID14	- Alpha diversity, observed otus - input: Kaz_trees.csv - analyze graph using Appendix 6
Q 3.3	2	multiple choice with multiple answers	2 point for 4 correct IDs 0.5 points for 1 correct ID negative 1 points for each incorrect ID the lowest mark is 0 Correct answers: ID 5, ID 6, ID 13, ID14	- alpha diversity, shannon - input: Kaz_trees.csv - analyze graph using Appendix 6
Q 3.4	2	multiple choice with multiple answers	2 points for 2 correct IDs 1 points for 1 correct ID negative 2 points for each incorrect ID	- alpha diversity, shannon - input: Kaz_trees.csv - analyze graph using Appendix 6

			the lowest mark is 0 Correct answers: ID 2, ID 17	
Q 3.5	1	multiple choice with multiple answers	1 points for correct ID 1 points for 1 correct ID negative 2 points for each incorrect ID the lowest mark is 0 Correct answers: ID 14	- alpha diversity, simpson - input: Kaz_trees.csv - analyze graph using Appendix 6
Q 3.6	2	multiple choice with multiple answers	2 points for 2 correct IDs 1 points for 1 correct ID negative 2 points for each incorrect ID the lowest mark is 0 Correct answers: ID 2, ID 17	- alpha diversity, simpson - input: Kaz_trees.csv - analyze graph using Appendix 6
Q 3.7	3	multiple choice with multiple answers	3 points for 1 correct IDs negative 2 points for each incorrect ID the lowest mark is 0 Correct answers: ID 13	- NMDS - input: Kaz_trees.csv - analyze graph using Appendix 6
Q 3.8	1	True/False	1 point for correct answer 0 points for incorrect correct answer: False	- Analysis of the result from the previous question - Knowledge-based - analyze graph using Appendix 6
Q 3.9.1	2	multiple choice with multiple answers	0.66 points for 1 correct IDs negative 1.32 points for each incorrect ID the lowest mark is 0 Correct answers: ID 1, ID 4, ID 16	- NMDS - input: Kaz_trees.csv - analyze graph using Appendix 6
Q 3.9.2	2	multiple choice with multiple answers	1 points for 1 correct ID negative 2 points for each incorrect ID the lowest mark is 0 Correct answers: ID 6, ID 7	- NMDS - input: Kaz_trees.csv - analyze graph using Appendix 6

Q3.10.1	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: A	- alpha diversity, observed otus - Input: 1) sum of tree diversity in Kazagandy and Ulitay in Kaz_trees.csv as one, 2)karagandy_trees_1971.csv - compare 2 graphs and decide whether it decreased/increased / unchanged - use Appendix 6
Q3.10.2	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: A	- alpha diversity, simpsons - Input: 1) sum of tree diversity in Kazagandy and Ulitay in Kaz_trees.csv as one, 2)karagandy_trees_1971.csv - compare 2 graphs and decide whether it decreased/increased / unchanged - use Appendix 6
Q3.10.3	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: B	- alpha diversity, shannon - Input: 1) sum of tree diversity in Kazagandy and Ulitay in Kaz_trees.csv as one, 2)karagandy_trees_1971.csv - compare 2 graphs and decide whether it decreased/increased / unchanged - use Appendix 6
Q3.10.4	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: averageYearlyTemp	- PCA - input: first 7 rows of karagandy_factors_2021.csv - analyze graph using Appendix 6 - select the factor

Q3.10.5	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: averageYearlyTemp	- PCA - input: 7-14 rows of karagandy_factors_2021.csv - analyze graph using Appendix 6 - select the factor
Q3.10.6	1	multiple choice with 1 answer	1 point for correct answer 0 points for incorrect correct answer: averageYearlyTemp	- PCA - input: karagandy_factors_1971.csv - analyze graph using Appendix 6 - select the factor
Q 3.10.7	1	True/False	1 point for correct answer 0 points for incorrect correct answer: True	- no specific tool needed - use Appendix 6 - choose the correct answer based on the results of questions 3.10.4 / 3.10.5 / 3.10.6

Q2.10.1

atcgatgcataatgtgcctgtcaaatggacgaagcagggattctgcaaaccctatgctactccgtcaagccgtcaattgtctgattcgttaccaattatgacaacttgacggctacatcattcacttttttcttcacaa
ccggcacggaactcgctcgggctggccccggtgcatttttttaaataccgcgagaaatagagttgatcgtcaaaaccaacattgcgaccgacggtggcgataggcatccgggtggtgctcaaaagcagct
tcgcctggctgatacgttggtcctcgcgccagcttaagacgctaatccctaactgctggcggaaaagatgtgacagacgcgacggcgacaagcaaacatgctgtgcgacgctggcgatatcaaaattgct
gtctgccaggtgatcgctgatgtactgacaagcctcgcgtacccgattatccatcggtggatggagcgactcgttaatcgcttccatgcgccgcagtaacaattgctcaagcagatttatcgccagcagctccg
aatagcgcccttcccccttgcccggcgttaatgatttgcccaaacaggtcgctgaaatgcggctggtgcgcttcatccgggcgaaagaaccccgtattggcaaatattgacggccagttaagccattcatgcca
gtaggcgcgcggacgaaagtaaacccactggtgataccattcgcgagcctccggatgacgaccgtagtgatgaatctctcctggcgggaacagcaaaatatcacccggtcggcaaacaaattctcgtcc
ctgatttttcaccaccccctgaccgcgaatggtgagattgagaatataacctttcattcccagcggtcggtcgataaaaaaatcgagataaccgttggcctcaatcggcgttaaacccgccaccagatgggc
attaaacgagtatcccggcagcaggggatcattttgcgcttcagccatacttttcatactcccgccattcagagaagaaaccaattgtccatattgcatcagacattgccgtcactgcgtcttttactggctcttctc
gctaaccaaaccggtaaccccgcttattaaaagcattctgtaacaaagcgggaccaaagccatgacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacgg
cgtcacactttgctatgccatagcatttttatccataagattagcggatcctacctgacgctttttatcgcaactctctactgtttctccatacccgtttttttgggctagaaataatttttgtttaactttaagaaggagatat
acatatggcttagagcaaaggagaagaacttttcactggagttgtcccaattcttgttgaattagatggtgatgttaatgggcacaaattttctgtcagtggagagggtgaaggtgatgctacatacggaaagct
taccccttaaatttatttgcactactggaaaactacctgttccatggccaacacttgtcactactttctcttatggtgttcaatgcttttcccgttatccggatcatatgaaacggcatgacttttcaagagtgccatgccc
gaaggttatgtacaggaacgcactatatctttcaaagatgacgggaactacaagacgcgtgctgaagtcaagtttgaaggtgatacccttgttaatcgtatcgagttaaaaggtattgattttaaagaagatgg
aaacattctcggacacaaactcgagtacaactataactcacactaggtatagatcacggcagacaaacaaaagaatggaatcaaagctaacttcaaaattcgccacaacattgaagatggatccgttca
actagcagaccattatcaacaaaatactccaattggcgatggccctgtccttttaccagacaaccattacctgtcgacacaatctgcccttttcgaaagatcccaacgaaaagcgtgaccacatggtccttcttg
agtttgtaactgctgctgggattacacatggcatggatgagctctacaaataatgaattcgagctcggtaccAGGAGGATGGCACTGGTTGAACAATGTTGTACAAGCTTGTGTT

CCATCTATCAGATCGAGAACTACTGCAATTGATtctagagtcgacctgcaggcatgcaagcttggctgttttggcggatgagagaagattttcagcctgatacagattaaatcagaacgc
agaagcggtctgataaaacagaatttgcctggcggcagtagcgcggtggtcccacctgaccccatgccgaactcagaagtgaaacgccgtagcgccgatggtagtgtggggtcccccatgcgagagta
gggaactgccaggcatcaaataaaacgaaaggctcagtgcaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctcctgagtaggacaaatccgccgggagcggatttgaacgttgcgaa
gcaacggcccggagggtggcgggcaggacgcccgccataaactgccaggcatcaaattaagcagaaggccatcctgacggatggccttttgcgtttctacaaactctttgtttatttttctaaatacattcaa
atatgtatccgctcatgagacaataaccctgataaatgcttcaataatattgaaaaaggaagagtatgagtattcaacatttccgtgtcgcccttattccctttttgcggcattttgccttcctgtttttgctcacccag
aaacgctggtgaaagtaaaagatgctgaagatcagttgggtgcacgagtgggttacatcgaactggatctcaacagcggtaagatccttgagagttttcgccccgaagaacgttttccaatgatgagcacttt
taaagttctgctatgtggcgcggtattatcccgtgttgacgccgggcaagagcaactcggtcgccgcatacactattctcagaatgacttggttgagtactcaccagtcacagaaaagcatcttacggatggca
tgacagtaagagaattatgcagtgctgccataaccatgagtgataacactgcggccaacttacttctgacaacgatcggaggaccgaaggagctaaccgcttttttgcacaacatgggggatcatgtaactc
gccttgatcgttgggaaccggagctgaatgaagccataccaaacgacgagcgtgacaccacgatgcctgcagcaatggcaacaacgttgcgcaaactattaactggcgaactacttactctagcttcccg
gcaacaattaatagactggatggaggcggataaagttgcaggaccacttctgcgctcggcccttccggctggctggtttattgctgataaatctggagccggtgagcgtgggtctcgcggtatcattgcagca
ctggggccagatggtaagccctcccgtatcgtagttatctacacgacggggagtcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgattaagcattggtaactgtcaga
ccaagtttactcatatatactttagattgatttacgcgccctgtagcggcgcattaagcgcggcgggtgtggtggttacgcgcagcgtgaccgctacacttgccagcgccctagcgcccgctcctttcgctttcttc
ccttcctttctcgccacgttcgccggctttccccgtcaagctctaaatcgggggctccctttagggttccgatttagtgctttacggcacctcgaccccaaaaaacttgatttgggtgatggttcacgtagtgggcca
tcgccctgatagacggtttttcgccctttgacgttggagtccacgttctttaatagtggactcttgttccaaactggaacaacactcaaccctatctcgggctattcttttgatttataagggattttgccgatttcggcct
attggttaaaaaatgagctgatttaacaaaaatttaacgcgaattttaacaaaatattaacgtttacaatttaaaaggatctaggtgaagatccttttgataatctcatgaccaaaatcccttaacgtgagttttcgtt
ccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaaaaaaccaccgctaccagcggtggtttgtttgccggatcaagagctac
caactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgtccttctagtgtagccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgtt
accagtggctgctgccagtggcgataagtcgtgtcttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacgggggttcgtgcacacagcccagcttggagcgaac
gacctacaccgaactgagatacctacagcgtgagcattgagaaagcgccacgcttcccgaagggagaaaggcggacaggtatccggtaagcggcagggtcggaacaggagagcgcacgaggga
gcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtgatgctcgtcaggggggcggagcctatggaaaaacgccagcaacgcggccttttacg
gttcctggccttttgctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataaccgtattaccgcctttgagtgagctgataccgctcgccgcagccgaacgaccgagcgcagcgagtcagt
gagcgaggaagcggaagagcgcctgatgcggtattttctccttacgcatctgtgcggtatttcacaccgcatatggtgcactctcagtacaatctgctctgatgccgcatagttaagccagtatacactccgcta
tcgctacgtgactgggtcatggctgcgccccgacacccgccaacacccgctgacgcgccctgacgggcttgtctgctcccggcatccgcttacagacaagctgtgaccgtctccgggagctgcatgtgtca
gaggttttcaccgtcatcaccgaaacgcgcgaggcagcaaggagatggcgcccaacagtcccccggccacggggcctgccaccatacccacgccgaaacaagcgctcatgagcccgaagtggcga
gcccgatcttccccatcggtgatgtcggcgatataggcgccagcaaccgcacctgtggcgccggtgatgccggccacgatgcgtccggcgtagaggatctaattctcatgtttgacagcttatc

Q2.10.2

atcgatgcataatgtgcctgtcaaatggacgaagcagggattctgcaaaccctatgctactccgtcaagccgtcaattgtctgattcgttaccaattatgacaacttgacggctacatcattcacttttttcttcacaa
ccggcacggaactcgctcgggctggccccggtgcatttttttaaataccgcgagaaatagagttgatcgtcaaaaccaacattgcgaccgacggtggcgataggcatccgggtggtgctcaaaagcagct
tcgcctggctgatacgttggtcctcgcgccagcttaagacgctaatccctaactgctggcggaaaagatgtgacagacgcgacggcgacaagcaaacatgctgtgcgacgctggcgatatcaaaattgct
gtctgccaggtgatcgctgatgtactgacaagcctcgcgtacccgattatccatcggtggatggagcgactcgttaatcgcttccatgcgccgcagtaacaattgctcaagcagatttatcgccagcagctccg
aatagcgcccttcccccttgcccggcgttaatgatttgcccaaacaggtcgctgaaatgcggctggtgcgcttcatccgggcgaaagaaccccgtattggcaaatattgacggccagttaagccattcatgcca

gtaggcgcgcggacgaaagtaaacccactggtgataccattcgcgagcctccggatgacgaccgtagtgatgaatctctcctggcgggaacagcaaaatatcacccggtcggcaaacaaattctcgtcc
ctgatttttcaccacccccctgaccgcgaatggtgagattgagaatataacctttcattcccagcggtcggtcgataaaaaaatcgagataaccgttggcctcaatcggcgttaaacccgccaccagatgggc
attaaacgagtatcccggcagcaggggatcattttgcgcttcagccatacttttcatactcccgccattcagagaagaaaccaattgtccatattgcatcagacattgccgtcactgcgtcttttactggctcttctc
gctaaccaaaccggtaaccccgcttattaaaagcattctgtaacaaagcgggaccaaagccatgacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacgg
cgtcacactttgctatgccatagcatttttatccataagattagcggatcctacctgacgcttttatcgcaactctctactgtttctccatacccgttttttgggctagaaataatttttgtttaactttaagaaggagatat
acatatggcttagagcaaaggagaagaacttttcactggagttgtcccaattcttgttgaattagatggtgatgttaatgggcacaaattttctgtcagtggagagggtgaaggtgatgctacatacggaaagct
taccccttaaatttatttgcactactggaaaactacctgttccatggccaacacttgtcactactttctcttatggtgttcaatgcttttcccgttatccggatcatatgaaacggcatgacttttcaagagtgccatgccc
gaaggttatgtacaggaacgcactatatctttcaaagatgacgggaactacaagacgcgtgctgaagtcaagtttgaaggtgatacccttgttaatcgtatcgagttaaaaggtattgattttaaagaagatgg
aaacattctcggacacaaactcgagtacaactataactcacactaggtatagatcacggcagacaaacaaaagaatggaatcaaagctaacttcaaaattcgccacaacattgaagatggatccgttca
actagcagaccattatcaacaaaatactccaattggcgatggccctgtccttttaccagacaaccattacctgtcgacacaatctgccctttcgaaagatcccaacgaaaagcgtgaccacatggtccttcttg
agtttgtaactgctgctgggattacacatggcatggatgagctctacaaataatgaattcgagctcggtaccAGGAGGATGTTCGTGAACCAACATATTTGTGCAAGTCATATTGT
TGAGGGCATCTACATTGTTTGCGCCGAACGGGCGTTTTTCTACACGCCGAAAACTTAAtctagagtcgacctgcaggcatgcaagcttggctgttttggcggatgaga
gaagattttcagcctgatacagattaaatcagaacgcagaagcggtctgataaaacagaatttgcctggcggcagtagcgcggtggtcccacctgaccccatgccgaactcagaagtgaaacgccgtag
cgccgatggtagtgtggggtcccccatgcgagagtagggaactgccaggcatcaaataaaacgaaaggctcagtgcaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctcctgagtagg
acaaatccgccgggagcggatttgaacgttgcgaagcaacggcccggagggtggcgggcaggacgcccgccataaactgccaggcatcaaattaagcagaaggccatcctgacggatggccttttg
cgtttctacaaactctttgtttatttttctaaatacattcaaatatgtatccgctcatgagacaataaccctgataaatgcttcaataatattgaaaaaggaagagtatgagtattcaacatttccgtgtcgcccttattcc
ctttttgcggcattttgccttcctgtttttgctcacccagaaacgctggtgaaagtaaaagatgctgaagatcagttgggtgcacgagtgggttacatcgaactggatctcaacagcggtaagatccttgagagtt
ttcgccccgaagaacgttttccaatgatgagcacttttaaagttctgctatgtggcgcggtattatcccgtgttgacgccgggcaagagcaactcggtcgccgcatacactattctcagaatgacttggttgagta
ctcaccagtcacagaaaagcatcttacggatggcatgacagtaagagaattatgcagtgctgccataaccatgagtgataacactgcggccaacttacttctgacaacgatcggaggaccgaaggagct
aaccgcttttttgcacaacatgggggatcatgtaactcgccttgatcgttgggaaccggagctgaatgaagccataccaaacgacgagcgtgacaccacgatgcctgcagcaatggcaacaacgttgcgc
aaactattaactggcgaactacttactctagcttcccggcaacaattaatagactggatggaggcggataaagttgcaggaccacttctgcgctcggcccttccggctggctggtttattgctgataaatctgga
gccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccctcccgtatcgtagttatctacacgacggggagtcaggcaactatggatgaacgaaatagacagatcgctgaga
taggtgcctcactgattaagcattggtaactgtcagaccaagtttactcatatatactttagattgatttacgcgccctgtagcggcgcattaagcgcggcgggtgtggtggttacgcgcagcgtgaccgctaca
cttgccagcgccctagcgcccgctcctttcgctttcttcccttcctttctcgccacgttcgccggctttccccgtcaagctctaaatcgggggctccctttagggttccgatttagtgctttacggcacctcgaccccaa
aaaacttgatttgggtgatggttcacgtagtgggccatcgccctgatagacggtttttcgccctttgacgttggagtccacgttctttaatagtggactcttgttccaaactggaacaacactcaaccctatctcggg
ctattcttttgatttataagggattttgccgatttcggcctattggttaaaaaatgagctgatttaacaaaaatttaacgcgaattttaacaaaatattaacgtttacaatttaaaaggatctaggtgaagatccttttga
taatctcatgaccaaaatcccttaacgtgagttttcgttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaaaaaaccaccg
ctaccagcggtggtttgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgtccttctagtgtagccgtagttaggccaccacttcaagaactct
gtagcaccgcctacatacctcgctctgctaatcctgttaccagtggctgctgccagtggcgataagtcgtgtcttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacg
gggggttcgtgcacacagcccagcttggagcgaacgacctacaccgaactgagatacctacagcgtgagcattgagaaagcgccacgcttcccgaagggagaaaggcggacaggtatccggtaagc
ggcagggtcggaacaggagagcgcacgagggagcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtgatgctcgtcaggggggcggagc
ctatggaaaaacgccagcaacgcggccttttacggttcctggccttttgctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataaccgtattaccgcctttgagtgagctgataccgctcgc
cgcagccgaacgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcctgatgcggtattttctccttacgcatctgtgcggtatttcacaccgcatatggtgcactctcagtacaatctgctct
gatgccgcatagttaagccagtatacactccgctatcgctacgtgactgggtcatggctgcgccccgacacccgccaacacccgctgacgcgccctgacgggcttgtctgctcccggcatccgcttacaga
caagctgtgaccgtctccgggagctgcatgtgtcagaggttttcaccgtcatcaccgaaacgcgcgaggcagcaaggagatggcgcccaacagtcccccggccacggggcctgccaccatacccacg

ccgaaacaagcgctcatgagcccgaagtggcgagcccgatcttccccatcggtgatgtcggcgatataggcgccagcaaccgcacctgtggcgccggtgatgccggccacgatgcgtccggcgtaga
ggatctaattctcatgtttgacagcttatc