INTERNATIONAL
BIOLOGY
OLYMPIAD e. V. IBO

33rd International Biology Olympiad

July 10 - 18, 2022

Yerevan, Armenia

IBO
33rd 2022
ARMENIA

Practical Examination 4

BIOINFORMATICS

Total points: 100

Duration: 90 minutes

Student Code	
Student Password	

Bioinformatics Practical

IBO
33rd 2022
ARMENIA

English (Official)

General Information

For this task, you can get up to **100 points.** The task consists of two independent parts:

Part 1. Study of chemokine signalling pathway perturbations using Pathway Signal Flow Calculator (PSFC,) tool **(72 points)**

Part 2. Sequence alignment **(28 points)**

Materials and Equipment

Name	Quantity
Computer with Application running	1 device
Envelope with materials	1 piece
Calculator	1 piece
Pen	1 piece

Carefully read the instructions before entering the Application.

Write your student code and student password from the envelope in the corresponding cells on the cover page table.

Log into the system by inserting your **Student Code** and **Student Password**, then press the **"Log in"** button. After logging in you will have 87 minutes to complete the task. The remaining time is shown in the left upper corner of the Application. **If you are using a translated version of the task you must answer questions in the corresponding English version directly in the "Exam" tab of the Application. All answers will be recorded and evaluated through the Application.**

Stop working on the Application immediately when you hear the bell ring at the end of the exam.

After 87 minutes your answers will be recorded and the Application will be automatically closed.

You are not allowed to take any material and equipment with you when leaving your table.

The screen of the computer is being recorded and any attempt to use external tools and information with the computer will lead to **disqualification from the exam.**

Use "." as a decimal separator NOT ",", e.g. 1.2.

If you leave all the answers blank you will get 0 points from a given task.

For all multiple choice questions with 4 options the following scheme applies:

- **all 4 correct = 100% of total score of that question**
- **only 3 correct = 60% of total score of that question**
- **only 2 correct = 20% of total score of that question**
- **1 or 0 correct = no points**

NOTE: any statement that is not selected (empty box) will be considered as a false statement.

Part 1 - Study of chemokine signalling pathway perturbations using Pathway Signal Flow Calculator (PSFC) tool (72 points)

Introduction

Genetic studies have moved to a new level where one experiment allows us to perform thousands of measurements for one biological sample at once. This type of experiment allows us to evaluate multiple features of samples comprehensively. As a result, large amounts of data are generated which can be analysed and interpreted by modern computational methods.

In this task, you will have an opportunity to work with different types of large biological data analysis methods and use them for solving these biological tasks.

A **biological pathway** is a chain of interactions of biological molecules that result in a new product or the change of cell state. Pathways can initiate the synthesis of new molecules, “turn on” and “turn off” genes, initiate cell movement, etc.

Generally, biological pathways regulate the cell response to external stimuli. A **signalling pathway** is a special type of biological pathway through which a chemical or physical signal is transmitted through a cell as a series of molecular events, which ultimately result in a cellular response. For instance, insulin’s interaction with its receptor is an external stimulus that initiates activation of the insulin-signalling pathway, a chain of interactions ultimately resulting in the activation of several biological processes such as glycolysis and lipogenesis.

Certain factors can disturb the normal activity of biological pathways. For instance, some mutations can affect protein functions. The presence of certain biologically active substances in the environment can also cause pathway perturbations. These factors may disrupt the regular cell response to the stimulus, resulting in the development of different conditions.

In this assignment, you have to discover the cause of chemokine signalling pathway perturbations and suggest a way to restore the activity of the pathway. Part 1 consists of 4 problems that you need to solve ***sequentially*** from start to finish.

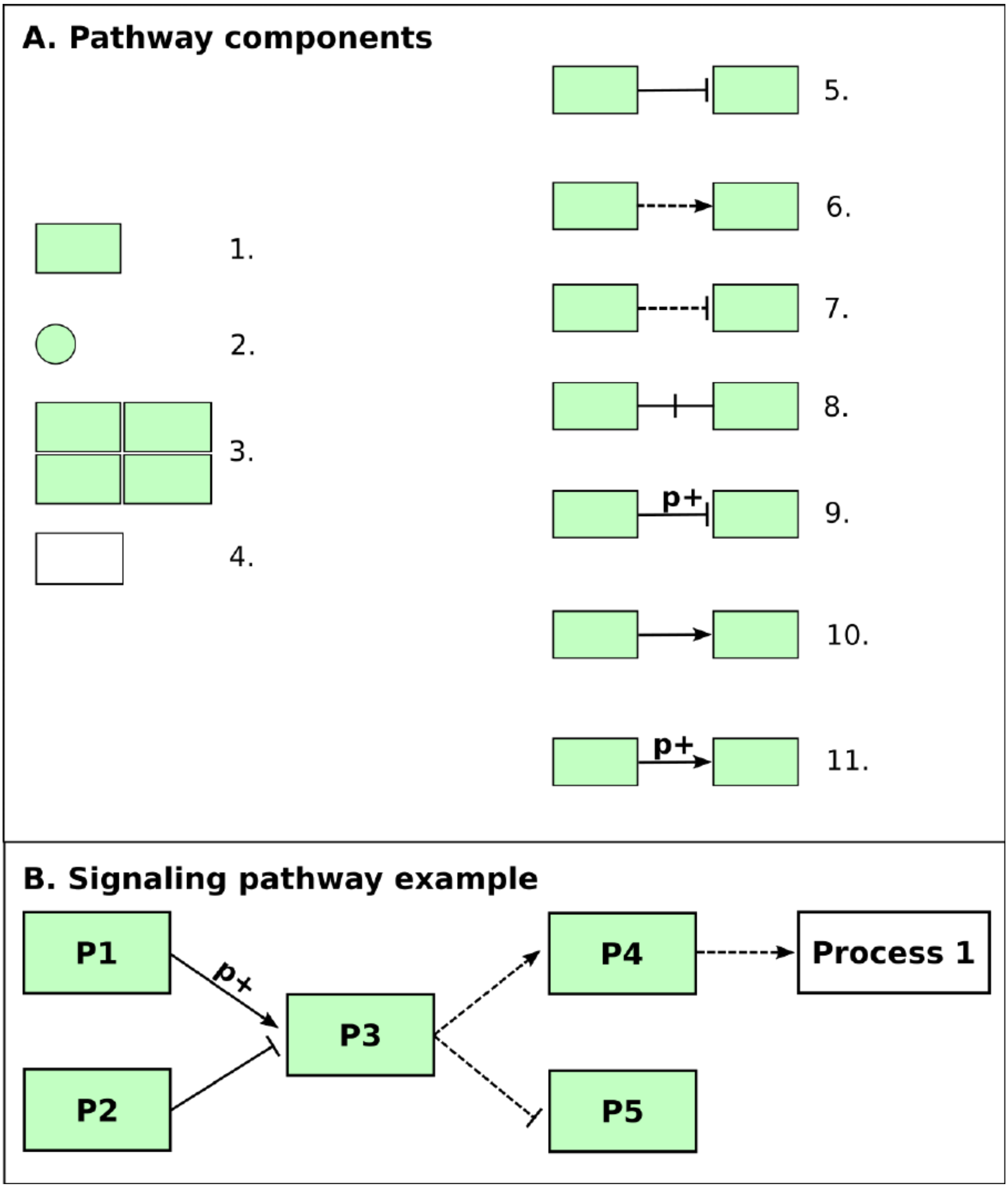
Problem 1.1. Chemokine signalling pathway construction

Chemokines are small cytokines (cell-to-cell signalling proteins), which mediate chemotaxis in cells. Chemokines also regulate processes in hematopoietic cells, such as cell differentiation and activation of cell survival mechanisms.

The chemokine signalling pathway represents the sequence of interactions, which starts with a chemokine receptor and ends with a certain biological process.

In this problem, you will construct one branch of the chemokine signalling pathway, based on the information provided in the table of **Appendix 1**. The table contains the sentences, which describe the components of the pathway and the interactions between them. The names of the pathway components are formatted in **BOLD**.

Have a look at an example of signalling pathways provided in **Figure 1.1.** In section **A**, the components of the pathway are shown. Section **B** shows an example of a pathway. The pathway consists of nodes (proteins, small molecules, DNA, biological processes) and **interactions** (edges) between **nodes** (activation, inhibition or dissociation; direct or indirect). The names of the protein nodes correspond to the names of encoding genes. Usually, the signalling pathways are presented in the following direction: from left to the right. This pathway starts with proteins **P1** and **P2**. Protein **P1** directly activates protein **P3** (the activation is performed by phosphorylation, which is shown by the **p+** character on top of the interaction arrow). **P2** protein directly inhibits **P3**. **P3** indirectly activates **P4** and indirectly inhibits **P5**. Protein **P4** indirectly activates **Process 1**. During direct interaction, the components (nodes) physically interact with each other. On the contrary, during an indirect interaction, there is no physical contact between the components (**nodes**), and the presence of intermediate proteins or chemicals is assumed.



A. Pathway components: 1. Protein; 2. Small molecule or DNA; 3. Protein Complex; 4. Biological process; 5. Inhibition of the protein: direct interaction; 6. Activation of the protein: indirect interaction; 7. Inhibition of the protein: indirect interaction; 8. Dissociation; 9. Inhibition of the protein with phosphorylation: direct interaction; 10. Activation of the protein: direct interaction; 11. Activation of the protein with phosphorylation: direct interaction.

B. Signaling pathway example

Question 1.1.1 (22 points)
Study the sentences represented in the Table of Appendix 1. Based on the sentences, you need to add the interactions between pathway nodes and construct one part of the chemokine signalling pathway. To construct the pathway, use **Instruction 1** provided below (each correct interaction is **1** point, each wrong interaction **-1** point, if you choose wrong attributes you will get 0 points, The lowest you can score for this question is 0).

- Answers will be automatically saved by the Application.

NOTE: You have to submit the constructed network to be able to proceed to the next questions of this section (EVEN IF IT IS JUST A BLANK NETWORK)
Once you submit the network (even a blank one) you CAN NOT go back and change it.

Instruction 1

1. Click on the **«Network construction»** tab.
2. All the nodes of the pathway are already uploaded in the Application according to the pathway components in **Figure 1.1**. The nodes are placed randomly. You can move the nodes across the window by clicking the left mouse button and moving the node without releasing the mouse button. All names of the nodes are matched with the **BOLD** names in the Table of **Appendix 1.** Nodes are shaped and coloured depending on their type (see **Figure 1.1 A** or the right part of the opened tab of the Application).
3. Based on the sentences in the Table of **Appendix 1,** determine all interactions between the nodes (**NOTE 1**: interactions between all node types are possible: protein-protein, protein-small molecule, small molecule-small molecule, protein-biological process, etc. **NOTE 2:** If there is a step-by-step interaction that connects two nodes, there shouldn't be an additional edge that connects them). Similarly, determine the interaction **types** between the nodes, based on the sentences and information from **Figure 1.1**. Add all interactions between the nodes by the following steps:
   - To add the interaction between 2 nodes, click on the **"Add edge"** button at the top of the window. Then click on the first interacting node and without releasing the mouse move to the second interacting node (**NOTE:** the direction should go from the source to the target). Next, a small window will appear on the screen. On the window select the **Interaction type** (activation, inhibition or dissociation) and the **Interaction state** (direct or indirect). Then click on the "**Save**" button to add the interaction.
   - To delete an existing interaction, click on the edge between the nodes and then click the "**Delete selected**" button at the top of the window. To confirm the action, click the **OK** button.
4. When you add all the interactions, click the "**Submit network**" button at the top of the window. Application will ask to confirm the action. **NOTE**, after you click the "**Submit**" button you will **NOT** be able to edit your constructed network anymore.

Problem 1.2. Analysing the perturbations in chemokine signalling pathway based on the gene expression data (27 points)

After submitting the network, click on the "**KEGG**" tab at the top of the window in the Application. Now you will see a full (and correct) chemokine signalling pathway. This (correct) pathway includes only the most important interactions, hence some minor interactions may be omitted. Study the presented pathway. Then go back to the "**Exam**" tab to answer the questions.

Question 1.2.1 (8 points)
Select the true statement(s), based on the pathway structure and information provided in **Figure 1.1 A**

1. Phosphorylation is a protein modification, which always leads to protein activation.
2. The interaction between **chemokine** and its receptors (**chemokineR**) is crucial for the normal functioning of the whole chemokine signalling pathway.
3. The indirect activation of **Raf** by **PI3K IA** involves at least 2 intermediate molecules (proteins, chemicals).
4. **IP3** is a chemical that binds to its receptor on the endoplasmic reticulum, which results in increased **Ca2+** flow into the cytoplasm.

Question 1.2.2 (8 points)
The flow of the chemokine signalling pathway starts at the cell membrane, transfers to the nucleus and ultimately activates the final processes (presented in the right part of the pathway). However, in some pathological cases, the signal flow can be disrupted. Select the true statement(s), based on the pathway structure.

1. A substitution mutation in the DNA sequence, which encodes the cytoplasmic domain of chemokine receptor (**chemokineR**), prevents the interaction between chemokine and its receptor.
2. Inactivation mutations in the **Gαi** gene affect all the final processes in the chemokine signalling pathway.
3. An inactivation mutation in the **ROCK** gene fully suppresses the **regulation of** the **actin cytoskeleton** by the chemokine pathway in the cell.
4. A mutation in the **Ras** gene, which prevents the interaction between **PI3K IA** and **Ras**, does not inhibit any final process in the chemokine signalling pathway.

To evaluate the signal activity, you need to have the expression values of each gene included in that pathway (quantity of messenger RNA (mRNA) in the cell). Depending on the cell type, gene expression can vary. For instance, the expression of insulin is considerably decreased in Type 1 diabetes, which leads to changes in signal activity in the Insulin signalling pathway.

You will study gene expression in the Chemokine signalling pathway in control and case (affected) samples of the same tissue

To see the expression values of control samples, go to the “**KEGG**” tab and click on the “**Map control expression**” button on the right side of the window. The nodes which correspond to the proteins will be coloured based on their gene expression values (from white to green, from low to high, respectively). To see the gene expression values, place the mouse cursor on the corresponding node. Then go back to the “**Exam**” tab to answer the questions.

Question 1.2.3 (8 points)
Select true statement(s) based on the expression values in **Control** samples and Chemokine signalling pathway structure.

1. The low expression of some genes, included in the Chemokine signalling pathway, is always a result of a mutation in the corresponding gene.
2. The expression of the **Tiam1** gene is a limiting factor for the **Tiam1-Par-PRKCZ** complex formation, given that the ratio of proteins forming the protein complex is 1:1:1.
3. Change in **Gαi** activity directly changes the expression value of **Src** in the pathway.
4. If **Akt** gene expression is decreased, **Cytokine production** will increase in the cell.

Question 1.2.4 (3 points)
Look at the gene list provided in the first column of **Table 1** below (in the Application).
Go to the "**KEGG**" tab and determine the gene expression values for each gene from Table **1** by hovering the mouse cursor on the corresponding nodes. Enter the values in the "**Control**" column, rounded to 1 decimal place (use "." as a decimal separator, e.g. 1.2). e.g. 1.55 rounds up to 1.6.
Now you need to determine the gene expression values in case (affected) samples. To map the values on the signalling pathway, click on the "***Map case expression***" button in the "**KEGG**" tab. Determine gene expression values for the genes provided in the Table **1 below** . Write these values under the "**Case**" column, rounded to 1 decimal place (use "." as a decimal separator, e.g. 1.2). e.g. 1.55 rounds up to 1.6. Now compare the expression values in control and case samples. Select "**+**", if there is a difference between expression values, otherwise select "**-**" in the "**Difference**" column of **Table 1** (double click on the Difference column cells to select "+" or "-" from the dropdown menu.).

Problem 1.3. Differential gene expression analysis of Chemokine signalling pathway (3 points)

In bioinformatics, the **Fold change (FC)** measure is used to describe the quantitative change between gene expression of control and case (affected) samples. The **FC** of the gene is defined as a ratio between the expression values of the given gene in affected (Case exp) and control (Control exp) samples.

FC = Case exp/Control exp

Question 1.3.1 (1 point)
In the Application go to "**KEGG**" tab and click on the "**FC table**" button. You will see a table with the gene list.
Based on the gene expression values, obtained in **Question 1.2.4**, calculate the **FC** values for those 5 genes. Write the calculated **FC** values in the "**FC table**" in the "**KEGG**" tab, rounded to 2 decimal places (use "." as a decimal separator, e.g. 1.22). e.g. 1.555 rounds up to 1.56. To enter the **FC** values, double click on the corresponding cell of the "**FC table**". When you finish, click on the "**Submit FC values**" button. **NOTE**: after you click the "**Submit FC values**" button, you will **NOT** be able to edit the "**FC table**" anymore. You must submit the **FC table** to proceed to the next questions of part 1.

Now click on the "**Map FC values**'' button in the "**KEGG**" tab to map all the **FC** values on the signalling pathway. The nodes will be coloured based on the **FC** values of coding genes, where:

- Genes with **FC** = 0-1, will be coloured from blue to white, respectively;
- Genes with **FC** = 1, will be coloured white (gene expression is the same in control and case (affected) samples);
- Genes with **FC** > 1, will be coloured from white to red, respectively;

To see the gene **FC** values, place the mouse cursor on the corresponding node

Question 1.3.2 (2 points)
Based on the **FC** values of the following genes, define which one/s have at least two-times higher expression values in **case** (affected) samples.

1. **p47phox**
2. **STAT**
3. **Itk**
4. **Rac**

Problem 1.4. Analysing the pathologies in Chemokine signalling pathway with Pathway Signal Flow Calculator (PSFC) (20 points)

In affected cells, the pathway signal performance can deviate from the norm. There are many methods to evaluate this deviation. One of them is **the Pathway Signal Flow Calculator (PSFC). PSFC** uses **FC** values to evaluate the pathway signal activity.

PSFC calculates the **Signal value** for each node (gene). **The Signal value** of a node is calculated based on its **FC** value and **FC** values of the nodes, which activate and/or inhibit the given node. For example, in **Figure 1.2, P1** activates **P2**, hence increasing the **P2's Signal value**. On the other hand, **P1** inhibits **P3**, hence, decreases the **P3 Signal value**. The nodes of the pathway are coloured based on each gene's **Signal value**. The higher the node's **Signal value**, the higher is its activity:

Signal value = 0-1 (coloured from blue to white, respectively) shows activity below normal

Signal value = 1 (coloured white) shows normal activity

Signal value > 1 (coloured from white to red, respectively) shows activity above normal.

The **Signal value** of the node, which is the closest one to a final biological process of the pathway, describes the activity of that process. For example, in **Figure 1.2**, the **Signal value** of **P5** (3.4) describes the activity of the final process. Since the **Signal value** of **P5** is higher than 1, the activity of the process can be considered as above normal.

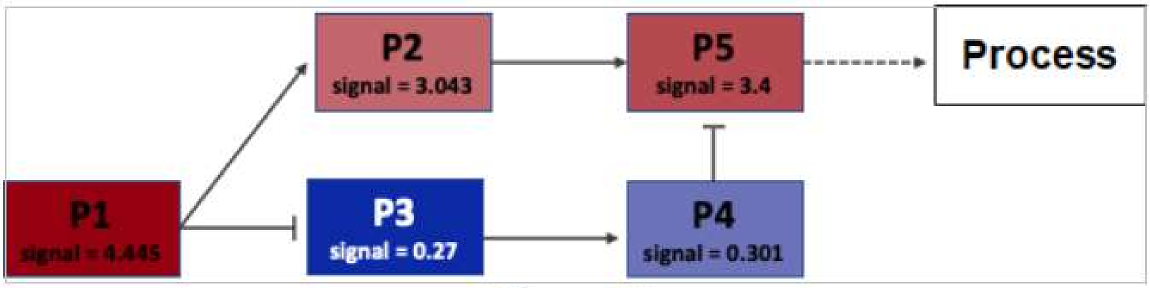


Figure 1.2

In the Application, go to the "**KEGG**" tab and click on the "**Calculate PSF**" button. The Application will calculate the signal values for all genes, included in the pathway. The nodes will be coloured based on their **Signal values**, as described above. To see the Signal values, place the mouse cursor on the corresponding node.

Question 1.4.1 (4 points)
Define the process/es, the activity of which is suppressed (activity below normal) in the case (affected) cells.

1. **Regulation of actin cytoskeleton**
2. **Ubiquitin mediated proteolysis**
3. **ROS production**
4. **Degranulation**

You can manually change the **FC** values of the genes. It will simulate the effect of the mutations or biologically active compounds on the affected pathway. To change the **FC** value of a gene, double click on the gene node, then click the right button of the mouse. A small window will appear on the screen. Click on the "**Change FC value**" button of that window. On the opened window enter the changed **FC** value for the selected node and press "**Change FC and calc PSF**". The **PSF** will be automatically recalculated for the whole pathway and the gene nodes will be coloured according to their updated **Signal values**. To cancel the changes and return to the original **Signal values**, click on the "**Recalc default PSF**" button.

Question 1.4.2 (4 points)
Change **FC** values of 4 genes, according to the values in **Table 2**.
Change the **FC** values for each gene (only one gene at a time). After each step, study how the signalling pathway has been changed. When you change the **FC** value for the next gene, all previous changes will be cancelled.
Choose the gene, whose mutation (changed FC) will have the least consequence on signalling pathway activity.

1. **PI3K IA**
2. **Src**
3. **Cdc42**
4. **Gβ**

Gene name	Changed FC value
PI3K IA	5.1
Src	6.2
Cdc42	4.6
Gβ	4.5

Table 2

Compound A is a biologically active substance, which binds to the active site of **Akt** with high affinity. It was determined that 10 nM **Compound A** binds to 80% of the **Akt** gene product in the affected cells.

Question 1.4.3 (4 points)
Simulate the effect of 10 nM **Compound A** on the Chemokine signalling pathway. You need to manually change the **FC** values of the corresponding node in the pathway. Based on your simulation, select the true statement(s). As a result of adding Compound A:

1. **Ubiquitin mediated proteolysis** is suppressed in the affected cells.
2. The **FC** value of the **FOXO** gene is increased at least three times in the affected cells.
3. The phosphorylation of **BAD** is decreased in the affected cells.
4. None of the processes have their activity changed in the affected cells.

Biologically active compounds can directly affect protein functionality or indirectly change (increase or decrease) the expression of the encoding gene. As a result, the activity of the whole signalling pathway or one of its branches can change. You have 4 different biologically active compounds, which affect the activity of specific proteins in the chemokine signalling pathway. The descriptions of these 4 compounds are shown in **Table 3.**

Biologically active compound	**Description**
Compound 1	Increases **Itk** gene expression, which leads to change of **Itk FC** value to 9.
Compound 2	Increases **Src** gene expression, which leads to change of **Src FC** value to 8.5.
Compound 3	Interacts with the protein encoded by the **PLCβ** gene and changes the **FC** value of the **PLCβ** gene to 0.001.
Compound 4	Interacts with the protein encoded by the **Gβ** gene and changes the **FC** value of the **Gβ** gene to 0.1.

Table 3

Question 1.4.4 (8 points)
Based on your answer to **Q 1.4.1** and the information from **Table 3**, indicate which compound can be used on the affected cells to activate the suppressed process(es), with a minimal side effect on the other processes. Choose only one answer.

1. Compound 1
2. Compound 2
3. Compound 3
4. Compound 4

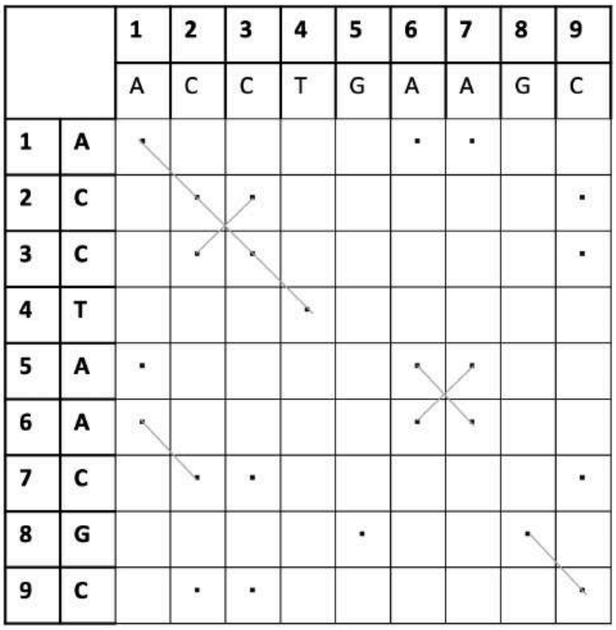
End of Part 1

Part 2 - Sequence alignment (28 points)

The alignment of biological sequences is one of the key steps in bioinformatics analysis, comparing the residues of biological sequences (DNA, RNA, proteins) and identifying similar regions. These imply structural and functional similarities of the sequences, and perhaps an evolutionary relationship between them. In this assignment, you will use different alignment methods to identify and describe similarities between the sequences using dot plot analysis, dynamic programming and heuristic methods.

Problem 2.1 Identification of an unknown sequence by dot plots (6 points)

There are many sequence alignment algorithms. The simplest one is the **dot plot** method, which allows a graphical comparison between two sequences on a coordinate plane or a table. One of the sequences is written on the horizontal axis of the plane, while the other one is written on the vertical axis. Then each letter of the first sequence is compared to all the letters of the second sequence. When two letters coincide, we draw a **point** in the corresponding position of the plane, otherwise, the position remains **empty**. If a pair of sequences has a significant number of coinciding letters, many points align into a continuous diagonal representing the alignment. Figure 2.1 shows the dot plots of **ACCTGAAGC** and **ACCTAACGC** sequences. We can see that the first four letters of each sequence coincide and form a continuous diagonal.



		1	2	3	4	5	6	7	8	9
		A	C	C	T	G	A	A	G	C
1	A	•					•	•		
2	C		•	•						•
3	C		•	•						•
4	T				•					
5	A	•					•	•		
6	A	•					•	•		
7	C		•	•						•
8	G					•			•	
9	C		•	•						•

Figure 2.1

In this problem, you have to identify the unknown sequence, using the dot plot method. The unknown sequence is a spliced gene.

Q 2.1.1 (2 points)
Perform sequence alignment of the unknown sequence using the dot plot method with the genome sequence of each of the following 4 genes: **APOE**, **GABAA1, ADGRG2** and **ORC2**. To do that, click on the **"Alignment"** tab of the Application. Then sequentially choose each of the given genes (**APOE, GABAA1, ADGRG2, ORC2**) from the top window and click on the **"Generate dot plot"** button to generate the dot plot of the unknown sequence and the chosen gene sequence. Observe all the dot plots and identify the unknown gene:

1. APOE
2. GABAA1
3. ADGRG2
4. ORC2

Q 2.1.2 (4 points)
Based on the given information and the obtained dot plots, select the true statement(s).

1. The gene identified in the **Q 2.1** has at least 5 exons.
2. Dot plots allow us to identify insertion/deletion mutations but not substitutions.
3. Dot plots allow us to identify inversions.
4. A dot-gap in the dot plot can be a result of either a point mutation or a DNA sequencing error.

Problem 2.2 The optimal alignment of a pair of sequences by the method of dynamic programming (13 points)

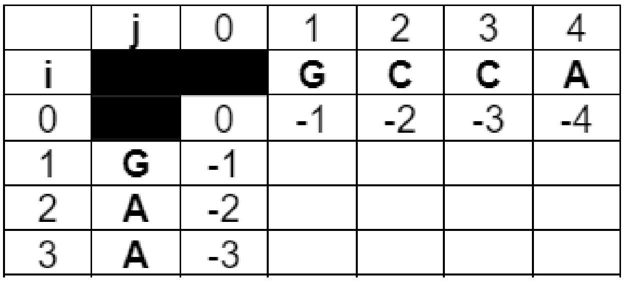
Dynamic programming is another method of sequence alignment, which allows finding the optimal alignment of a pair of sequences. This method is similar to the dot plot method because it is also based on a **two-dimensional alignment table.** However, instead of gaps and points, each pair of letters is assigned a certain numerical value in this method. Afterwards, the optimal alignment can be determined by selecting the corresponding group of cells with the highest value.

The alignment can be local or global. In **global alignment**, we compare two sequences throughout their entire lengths, while in **local alignment**, we search for similar regions in the sequences.

In the dynamic programming method, match-mismatch scores and gap penalties are used to evaluate sequence alignment. Insertion, deletion and substitution within the sequences are assigned gap penalties (usually, a negative score) and substitution penalties (usually, a negative score). On the contrary, a match of the letters is assigned a positive score. Let us assume that each match **M = 1 score**, gap **G = -1 penalty** and substitution **S = 0 penalty.**

Let's look at an example of the global alignment of **GCCA** and **GAA** sequences with the dynamic programming method. This is the tutorial to demonstrate dynamic programming using an example of global alignment of GCCA and GAA sequences.

1. First, we construct an n+1 x m+1 two-dimensional alignment table (see Figure below), where **n** and **m** are the lengths of alignment sequences (in this example **n** = 3 and **m** = 4). Afterwards, the matrix needs to be filled as follows: the cells of the first row from the left to the right and the cells of the first column from top to the bottom are filled with the numbers 0, -1, -2, ...



	j	0	1	2	3	4
i			G	C	C	A
0		0	-1	-2	-3	-4
1	G	-1				
2	A	-2				
3	A	-3				

Step 1.

2. Let's define **C(i, j)** to be the value of the cell with **(i, j)** coordinates. Moving row by row from the left to the right, each **(i, j)** cell needs to be filled with the largest of the following values:
   a. **C(i, j-1) + G** (if this direction is chosen, we assume that there is gap/deletion in vertical sequence), where **C(i, j-1)** is the value of the left neighbouring cell, and **G** is -1.
   b. **C(i-1, j) + G** (if this direction is chosen, we assume that there is gap/deletion in horizontal sequence), where **C(i-1, j)** is the value of the upper neighboring cell, and **G** is -1.
   c. **C(i-1, j-1) + M** (if there is a match in this position), where **C(i-1, j-1)** is the value of the diagonal left upper cell, and **M** is 1.

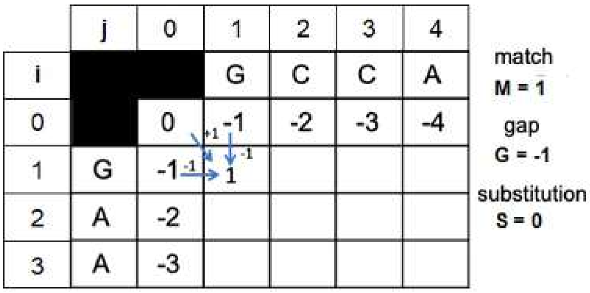
OR

C(i-1, j-1) [value of diagonal left upper cell] + **S** [0] (if there is a substitution in this position)

So, in our example, the cell with coordinates (1, 1) needs to be filled with the largest of the following values:

- C(1,1) = C(0,1) + G = -1-1= -2,
- C(1,1) = C(1,0) + G = -1-1 = -2,
- C(1,1) = C(0,0) + M = 0+1 = 1

Thus, it will be 1.



	j	0	1	2	3	4
i			G	C	C	A
0		0	-1	-2	-3	-4
1	G	-1	1			
2	A	-2				
3	A	-3				

match
M = 1
gap
G = -1
substitution
S = 0

Step 2

match	
gap	
substitution	

3. Let us fill the table in the same way row by row from left to right. **The value of the bottom-right cell is the alignment score.**

	j	0	1	2	3	4
i			G	C	C	A
0		0	-1	-2	-3	-4
1	G	-1	1	0	-1	-2
2	A	-2	0	1	0	0
3	A	-3	-1	0	1	**1**

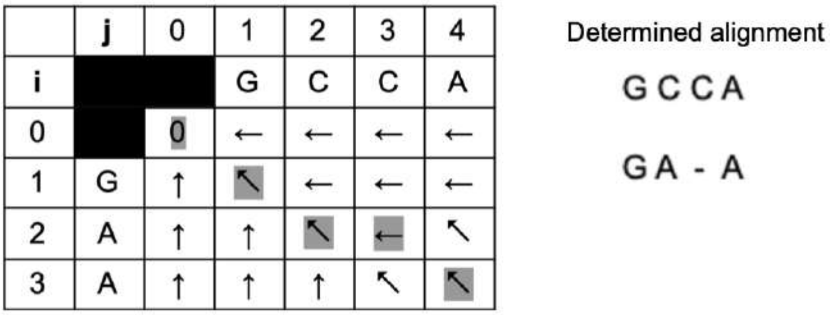
Step 3

4. Let us replace the value of each cell filled in Step 2-3 with an arrow directed to the left, up or upper left diagonally, if accordingly, the **a**, **b** or **c** options introduced in **Step 2** were chosen for that cell. If any two or all of the values from **a**, **b**, **c** are equal for a given cell, then one of the corresponding arrows should be selected randomly.

	j	0	1	2	3	4
i			G	C	C	A
0		0	←	←	←	←
1	G	↑	↖	←	←	←
2	A	↑	↑	↖	←	↖
3	A	↑	↑	↑	↖	↖

Step 4

5. Starting from the bottom-right cell, let us follow the arrow' directions and find the path. Then let us determine the alignment from the path. Take into consideration that «↖» corresponds to a match of the letters, «←» corresponds to a gap within the vertical sequence and «↑» corresponds to a gap within the horizontal sequence.



	j	0	1	2	3	4
i			G	C	C	A
0		0	←	←	←	←
1	G	↑	↖	←	←	←
2	A	↑	↑	↖	←	↖
3	A	↑	↑	↑	↖	↖

Determined alignment

GCCA

GA - A

Step 5

determined aligment	

Q 2.2.1 (8 points, + 0.2 for each correctly filled cell).
Based on the algorithm presented above, perform the sequence alignment using the method of dynamic programming for the following pair of sequences: **AGTAC** and **GCAC**. Fill the left table with numeric values and fill the right table with arrows (double click on the empty cells to select desired arrow type from dropdown menu.).

Q 2.2.2 (3 points)
Enter the calculated alignment score below.

Q 2.2.3 (2 points).
Determine the alignment of the sequences and enter it in the table.
Fill with A, T, C, G or "-" from the dropdown menu.

Problem 2.3 Determining the existence of Taq polymerase gene in the genomes of E. coli and T. thermophilus (9 points)

In this task, you need to search for the Taq polymerase gene of *Thermus aquaticus* bacteria in the genomes of *E. coli* and *T. thermophilus*. For this purpose, you need to perform the alignment of Taq polymerase's DNA sequence with the whole genomes of *E. coli* and *T. thermophilus.*

Q 2.3.1 (1 point)
Choose which type of alignment needs to be performed:

1. Local
2. Global

In the case of alignments where one sequence is significantly longer than the other one (for example, where one sequence is the whole genome), the dynamic programming method is very time-consuming. In these cases, heuristic methods are used instead. The goal of a heuristic method is to find several relatively good alignments in substantially less time instead of the best alignment. Statistical significance score is then calculated for each of the resulting alignments.

Q 2.3.2 (8 points)

Perform an alignment of the Taq polymerase's DNA sequence with the whole genomes of *E. coli* and *T. thermophilus* using the heuristic method. To do that, click on the **"Alignment"** tab of the Application. Then in the **bottom part of the window**, choose the **"Nucleotide"** option in the **"Select alignment type"** menu. Afterwards, choose *E. coli* in the **"Select organism"** menu and click on the **"Perform alignment"** button. Do the same for *T. thermophilus*. Explore the statistical significance of the obtained alignments based on the E-values **("eVal"** column). E-value is calculated to estimate the probability that the result is obtained by chance: the smaller the E-value, the more significant the result. Assume that an E-value **≤ 0.001** shows a statistically significant alignment of two sequences.

It is common to perform an amino-acid sequence alignment instead of a nucleotide sequence alignment. In this case, the nucleotide sequence needs to be translated into the amino-acid alphabet beforehand.

Now perform an alignment of Taq polymerase's protein sequence with the whole genomes of *E. coli* and *T. thermophilus* using the heuristic method. To do that, choose the **"Amino acid"** option in the **"Select alignment type"** menu. Perform the alignment for both *E. coli* and *T. thermophilus*. Explore the statistical significance of the obtained alignments based on the E-values (**"eVal"** column). Assume that an E-value **≤ 0.001** shows a statistically significant alignment of two sequences.

Select true statement(s) based on the results and the information provided above.

1. From an evolutionary point of view, an amino acid sequence is usually more conserved than the nucleotide sequence.
2. The results indicate that the Taq polymerase gene or its homologue has been discovered in the genome of *T. thermophilus*.
3. The majority of amino acids are encoded by more than one nucleotide triplet. Therefore, nucleotide alignment is preferable to the amino acid alignment in phylogenetic studies comparing distant species using coding sequences to avoid information loss.
4. The results indicate that the Taq polymerase gene or its homologue on nucleotide and amino acid level has NOT been discovered in the genome of E. coli.

Appendix 1

Note for translators: Please retain the **terms in bold in the original English font**, in addition to your translation, as these English terms will be used by the students to construct the network of the chemokine signalling pathway.

Chemokine and chemokine receptor **(chemokineR)** interaction plays a crucial role in inflammation and anti-tumour immunity.

Janus tyrosine kinases' family members **(JAK2/3)** appear to be associated with the cytoplasmic domain of many cytokine receptors **(chemokineR)**, but remain catalytically inactive until the binding of the **chemokine** to the receptor **(chemokineR)**.

Phosphorylation of **STAT** at tyrosine 701 by the Janus family of tyrosine kinases **(JAK2/3)** leads to **STAT** dimerization. This leads to **STAT** exposure to a dimer-specific nuclear localization signal, and subsequent nuclear translocation, where **STAT** acts as a transcription activator.

The G-protein subunits (**Gαi** and **Gβ**) are inactive when they bind with guanosine diphosphate (GDP). After the binding of the **chemokine** to a chemokine receptor **(chemokineR)**, G-proteins bind to the chemokine receptor, allowing the exchange of GDP for guanosine triphosphate (GTP). This causes the dissociation of the G protein subunits from each other (**Gαi** and **Gβ**), which leads to their activation.

The cellular level of **cAMP** is decreased as a result of **Gαi**'s direct inhibition of an intermediate molecule.

AC catalyzes the formation of the signaling molecule **cAMP** in response to G-protein signaling.

The ultimate effect of G subunit alpha **(Gαi)** is the inhibition of the cAMP-dependent protein kinase **(PKA)**.

The released activated G-protein subunit **Gβ** directly causes the subsequent activation of an enzyme, called phospholipase C **(PLCβ)** that is associated with the cell membrane.

PLCβ cleaves phosphatidylinositol (4,5)-bisphosphate (PIP2) into two messenger molecules: Inositol triphosphate **(IP3)** and diacylglycerol **(DAG)**.

IP3 induces calcium **(Ca2+)** influx.

Calcium **(Ca2+)** and **DAG** directly activate an enzyme, protein kinase C **(PKC)**.

Pyk2 encodes a tyrosine kinase that regulates cell **migration**. Calcium-activated, phospholipid- and diacylglycerol **(DAG)**-dependent protein kinase C **(PKC)** is indirectly involved in the activation of **Pyk2**.

High glucose levels stimulate **reactive oxygen species (ROS) production** via a PKC-dependent **(PKC)** indirect activation of NAD(P)H oxidase **(p47phox)**.

Activation of protein kinase C **(PKC)** is an essential signal for **degranulation**.

PKC (Protein kinase C) inhibitors, bisindolylmaleimide I (0.1–1 μM) and staurosporine (20-100 nM), effectively block propofol-induced eNOS (endothelial nitric oxide synthase) activation and **NO induction**.