INTERNATIONAL BIOLOGY OLYMPIAD e. V. IBO

All IBO examination questions are published under the following Creative Commons license:

CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) -
https://creativecommons.org/licenses/by-nc-sa/4.0/

The exam papers can be used freely for educational purposes as long as IBO is credited and new creations are licensed under identical terms. No commercial use is allowed.

Bioinformatics

34th International Biology Olympiad
3-10 July 2023, United Arab Emirates University

Practical Exam

Total points: 98
Duration: 90 minutes

General Instructions:
You have 90 minutes to complete **TWO tasks in this practical exam**.
You can do the tasks in any order.

Task 1: Molecular phylogenetics (61 points)
Task 2: Genome editing (37 points)

Bioinformatics

34th International Biology Olympiad
United Arab Emirates 2023

Q1-2

English (Official)

Important Information:

No answers on the exam paper will be graded.

You will use a dedicated browser application both to perform the bioinformatics analyses and to enter your responses. This application includes four different types of screens, namely question screen, "Notepad", "Input Sequences" 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.

You **MUST NOT** quit the Safe Exam Browser.

Questions which 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.

"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.

You can search 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 results of different analyses).

You can switch between Windows using "Alt" + "Tab".

If you minimise the electronic copy of the questions, you can quickly return to it by clicking the icon on the bottom left of the screen.

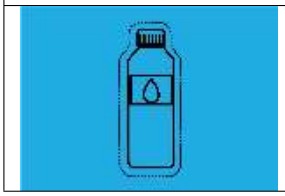
To ensure proper formatting of the obtained analysis output you can maximize the window.

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, this is because of your input being incorrect, so such queries will not be dealt with by the assistants.

Use the following cards to ask for water/washroom/help.

Drinking water	Washroom	Other queries
		

No paper, materials or equipment should be taken out of the laboratory.

Good luck!

TASK 1. MOLECULAR PHYLOGENY

One of the great milestones in molecular phylogenetics was the work by Carl Woese (1977). He compared the sequence of the small ribosomal subunit rRNA (16S/18S rRNA), between different species, by digesting it with T1 RNase followed by hybridization of conservative loops.

In this task, you will also use 16S/18S rRNA sequences, of 10 different organisms, to build a phylogenetic tree. To do that, you will first need to make an alignment of all corresponding sequences, and then use that alignment as input for the tree-building algorithm.

The input data for this task can be found by clicking on the "Input Sequences" dropdown menu and choosing the input labeled as "rRNA.fasta". This file contains the sequences of both the 16S/18S and the 23S/28S rRNA for each species. The labels have either "16S" or "23S" ending, respectively. Hence, you should prepare a "*fasta*" formatted input which contains only the sequences of 16S/18S rRNA. For this you can copy-paste the provided data into the Notepad page of the Application and edit it appropriately.

Notes: a) each sequence should have a header starting with ">"; b) you can enter empty lines between sequences in the "*fasta*" format to keep them visually separated. For more about the "*fasta*" format see Appendix 1.

Now align the 16S/18S sequences of each species using the "7. Sequence Alignment" tool. It will return an alignment in the "*clustal*" format (for more about it see Appendix 1).

Q1.1 True or False?

Q1.1.1	The alignment provides evidence for insertions or deletions happening during the evolution of 16S/18S rRNA genes.	1.0pt
Q1.1.2	*C. paramecium* 16S/18S rRNA is the shortest sequence in the alignment.	1.0pt
Q1.2	Copy your resulting alignment and paste it into the input field of question Q1.2 for the results validation.	5.0pt

Copy your resulting alignment and paste it into the input field of the "9. Tree Builder" tool. Press "Submit" for tree building.

When the analysis is finished, you will get the resulting tree in the output field in *"newick"* format (for more about the *"newick"* format see Appendix 1), as well as an image.

Note: the tree is unrooted!

Q1.3	Copy the "*newick*" tree you have obtained from the output of the "9. Tree Builder" tool and paste it into the input field of question Q1.3.	5.0pt

Based on a tree similar to the one you just obtained, Woese proposed that all cellular life is divided into three domains: Bacteria, Archaea and Eukaryotes.

Q1.4	For each of the species in Q1.4, select the domain of life it belongs to, using letters **"A**" for Archaea, "**B"** for Bacteria, "**E"** for Eukaryotes.	6.0pt

Q1.5 What can you conclude about the *Z. mays* (corn, *Zea mays*) sequence you used based on its position in the tree? For each statement indicate if it is True or False.

Q1.5.1	The sequence could be the sequence of the nuclear 18S rRNA gene.	1.0pt
Q1.5.2	The sequence could be the sequence of the chloroplast 16S rRNA gene.	1.0pt
Q1.5.3	The sequence could be the sequence of the mitochondrial 16S rRNA gene.	1.0pt
Q1.5.4	The position of *Z. mays* sequence must be incorrect / a tree-building artifact.	1.0pt

Q1.6 True or False?

Q1.6.1	The tree demonstrates that Archaea is evolutionarily closer to Eukaryotes than to Bacteria.	1.0pt
Q1.6.2	The tree can be rooted (the position of the Last Universal Common Ancestor of all cellular life on Earth can be identified) by including a viral sequence into the analysis.	1.0pt
Q1.6.3	The Last Universal Common Ancestor of all cellular life on Earth likely had a homolog of the 16S/18S rRNA.	1.0pt
Q1.6.4	The T1 enzyme used by Woese cleaves single-stranded RNA.	1.0pt
Q1.6.5	Based on the tree, *P. syntrophicum* is closer to *M. formicicum* than *P. syntrophicum* is to *T. peptonophilus.*	1.0pt
Q1.6.6	A taxon that includes *H. sapiens* and *S. cerevisiae* but excludes *C. paramecium* is monophyletic.	1.0pt

Molecular phylogenetics can also compare species on a finer scale. In this task you will explore the relationships between 5 bacteria species from the *Streptococcus* genus.

For this, you will use amino acid sequences of two proteins: DnaJ and Cas9.

- DnaJ is a molecular chaperone involved in protein folding.
- Cas9 is an endonuclease which is a part of the CRISPR-Cas9 system.

You have two "*fasta*" inputs, for all five *Streptococcus* species:

- Amino acid sequences of the DnaJ protein labeled “DnaJ-protein.fasta”
- DNA sequences of the Cas9 locus labeled “Cas9-locus.fasta”.

For the Cas9 locus, you need to find the Cas9 open reading frames (ORF; the part of a sequence which can be translated) and the corresponding amino acid sequences using the “3. ORF Finder” Tool. To do this, copy the "*fasta*" sequence of a single Cas9 locus, together with the species name label from the "Input Sequences" menu, and paste it into “3. ORF Finder”. Specify the following parameters for an ORF search by entering specific numbers into the corresponding fields:

- **Parameter 1:** The minimum ORF length (in amino acid residues). Note that Cas9 proteins in all given species are longer than 1000 amino acid residues.
- **Parameter 2:** The index of genetic code type which you would like to use to translate RNA to protein. The table below gives the available indices. Note that option 4 is not included deliberately.

Genetic code type	Genetic code index
The Standard Code	1
The Vertebrate Mitochondrial Code	2
The Yeast Mitochondrial Code	3
The Invertebrate Mitochondrial Code	5

The output of “3. ORF Finder” will be a table with 3 columns, showing ORF length, coding nucleotide sequence and encoded amino acid sequences for each Cas9 ORF.

You will have to run “3. ORF Finder” five times; once for each species.

For each input species, you should copy the DNA and the protein sequences of the correct ORF and paste it into the "Notepad". Each sequence should have a "*fasta*" header identical to the header in the original input (“>S.anginosus-Cas9” and so on; the corresponding DNA and protein sequences should have the same name).

Next, rearrange the sequences in the Notepad so that you have grouped the five DNA sequences first, and then the five protein sequences. Keep both the amino acid and the coding DNA sequences in the "Notepad" – you will need them later.

Q1.7 What type of genetic code did you use? Choose the corresponding index number in Q1.7. 1.0pt

Q1.8 Paste the five coding DNA sequences into the input field of Q1.8 in the "*fasta*" format. The stop codon should NOT be included. 5.0pt

Q1.9 When searching for ORFs, “3. ORF Finder” checks all possible reading frames (possible ways to divide the provided DNA sequence into codons) on both strands. How many reading frames in total does it check? Enter your answer as a number into the input field of Q1.9. 1.0pt

Follow the same workflow as for the rRNA sequences, to build trees based on the amino acid sequences of DnaJ (provided for you in the "Input Sequences" menu) and Cas9 (you have obtained these in the previous task).

Q1.10	Paste the DnaJ protein alignment in *"clustal"* format into the input field of Q1.10.	1.0pt

Q1.11	Paste the Cas9 protein alignment in *"clustal"* format into the input field of Q1.11.	1.0pt

Also save the Cas9 protein alignment in your Notepad as you will need it later on.

Q1.12	Look at the two alignments. Which of the two is expected to have fewer errors (non-homologous amino acid positions identified as homologous)? Choose DnaJ or Cas9 in Q1.12.	1.0pt

Q1.13	Paste the DnaJ tree in *"newick"* format into the input field of Q1.13.	2.5pt

Q1.14	Paste the Cas9 tree in *"newick"* format into the input field of Q1.14.	2.5pt

Q1.15 True or False?

Q1.15.1	The Cas9 protein is more evolutionarily conserved, compared to DnaJ.	1.0pt

Q1.15.2	The Cas9 tree is more likely to show the genome-average phylogenetic relationships between these species, compared to the DnaJ tree.	1.0pt

Q1.15.3	The difference between the trees for the two genes could be a result of horizontal gene transfer.	1.0pt

Q1.15.4	Both Cas9 and DnaJ trees include a branch with *S. anginosus* and *S. mitis* as closest neighbors.	1.0pt

Q1.15.5	If we did the analysis using some other gene from the same five species, we could get a tree with a different topology.	1.0pt

Using bioinformatics tools, we can also explore selective pressure on DNA and protein sequences. One method is the dN/dS test (also known as the Ka/Ks test). Here, negative selection is defined as amino acid substitutions being on average deleterious, while positive selection is a scenario of amino acid substitutions being on average beneficial.

The test compares the observed rates of non-synonymous (N; leading to a change in amino acid) and synonymous (S; not leading to a change in amino acid) nucleotide changes in a protein-coding DNA sequence, calculating their ratio (dN/dS). The rate of each type of substitutions (dN or dS) is defined as the number of observed substitutions of this type normalized (divided) by the total number of possi-

ble substitutions of this type. The table below shows the number of possible non-synonymous (N) and synonymous (S) substitutions for some codons.

Codon	Number of possible non-synonymous (N) substitutions	Number of possible synonymous (S) substitutions
ATG	9	0
AAA	8	1
AGA	7	2
ACA	6	3

The dN/dS test can be applied to either a pair of sequences, or to a tree (in the latter case, dN/dS ratio can be calculated for each branch).

Q1.16 True or False?

Note: the standard genetic code table can be found in Appendix 5.

Q1.16.1	The maximum number of possible single nucleotide synonymous changes for a codon in the standard genetic code is 4.	1.0pt
Q1.16.2	The described normalization is needed because the relative probability of N and S substitutions depend on the original sequence.	1.0pt
Q1.16.3	This test assumes that synonymous substitutions are selectively neutral.	1.0pt
Q1.16.4	One can directly compare dN/dS values between genes with moderately different total mutation rates.	1.0pt
Q1.17	What is the expected dN/dS value for a sequence that evolves completely neutrally (even non-synonymous substitutions are neither deleterious nor beneficial)? Enter your answer as a whole number into the answer field of Q1.17.	1.0pt

Q1.18 For each DNA sequence described below, choose if the result of the dN/dS test is expected to be equal, smaller or bigger than X, where X is the dN/dS value for a sequence evolving completely neutrally.

Q1.18.1	Coding sequence of histone H4.	1.0pt

Q1.18.2	Coding sequence of a viral surface protein that can be recognized by the host's immune defense.	1.0pt

Q1.18.3	Coding sequence of a pseudogene (a gene which no longer produces a functional product).	1.0pt

Now you should apply the dN/dS test to explore evolution of the Cas9 gene in the *Streptococcus* genus. The first step is to align the coding sequences by codons using the "1. Codon Alignment" tool.

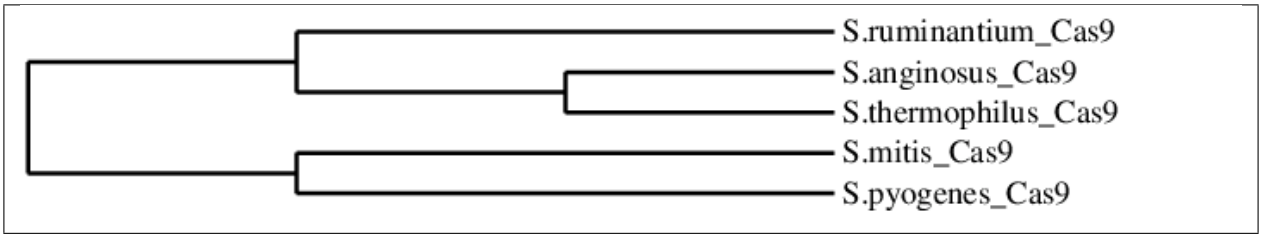
Copy the Cas9 protein alignment for the 5 species from your "Notepad" and paste it into the first field labeled "Format: *clustal*" .

Next, copy the Cas9 coding DNA sequences from your Notepad (all five together including the *"fasta"* headers) and paste them into the second field labeled "Format: *fasta*" , and press "Submit".

Q1.19	Paste the resulting Cas9 codon alignment data into the input field of Q1.19	1.0pt

Now, the codon alignment can be used as an input for the dN/dS test. However, because running it for each branch of the tree would take too long, the results are already provided to you below.

The cladogram below shows the clustering of the the five Cas9 sequences. Branch length conveys no information here. The table under the cladogram tabulates the dN/dS ratios for each branch of the cladogram.



Branch leading to	dN/dS ratio
S.ruminantium-Cas9	0.002
S.thermophilus-Cas9	0.028
S.anginosus-Cas9	0.011
S.pyogenes-Cas9	0.002
S.mitis-Cas9	0.022
(*S.pyogenes*-Cas9,*S.mitis*-Cas9)	0.112
(*S.thermophilus*-Cas9,*S.anginosus*-Cas9)	0.009
((*S.thermophilus*-Cas9,*S.anginosus*-Cas9), *S.ruminantium*-Cas9)	79.086

Q1.20

Analyze the results of the dN/dS test of the Cas9 gene in the *Streptococcus* species.

True or False?

Q1.20.1	It can be concluded that Cas9 was under constant negative selection during its evolution in *Streptococcus*.	1.0pt
Q1.20.2	During the evolution of Cas9, since the common ancestor of *S.thermophilus* and *S.anginosus* to modern-day *S.thermophilus* and *S.anginosus*, most of the non-synonymous substitutions were deleterious.	1.0pt
Q1.20.3	The divergence in Cas9 amino acid sequence between ((*S.thermophilus, S.anginosus*), *S.ruminantium*) on the one hand, and (*S.mitis*, *S.pyogenes*) on the other, could be explained by positive selection.	1.0pt
Q1.20.4	The difference in Cas9 amino acid sequence between ((*S.thermophilus*, *S.anginosus*),*S.ruminantium*) on one hand, and (*S.mitis*, *S.pyogenes*) on the other, could be explained by a high error (non-homologous amino acid positions identified as homologous) rate in the alignment.	1.0pt

END OF TASK 1

TASK 2. GENOME EDITING

Genome editing became a widespread technology thanks to the *Streptococcus pyogenes* Cas9 enzyme. In this approach, mammalian cells are transfected with a vector containing the sequence encoding Cas9 and a sequence encoding gRNA (guide RNA). The mechanism behind Cas9 application is described in Appendix 3.

Natural Cas9 uses two different RNAs: crRNA for DNA target sequence recognition and tracrRNA for proper enzyme assembly and crRNA binding. Both these functions can be combined in an artificial RNA, called gRNA.

The guide sequence consists of the target sequence (usually 20 nucleotides) and should be followed by the PAM sequence (protospacer adjacent motif; 5'-NGG-3', where N is any base). Therefore, only sequences with a PAM sequence at their 3' end can be targeted by the gRNA.

In this task your ultimate goal is to design an experiment to create a mouse model of the human sickle-cell anemia disease, by editing the *Hbb* gene in mice. Specifically, you need to replace a part of the first exon of the mouse *Hbb* gene with the corresponding human sequence carrying the disease-causing mutation.

Initially, you need to design primers to amplify and clone the complete Cas9 coding sequence into a mammalian expression vector.

The primers should meet the following requirements:

- Allow to amplify the complete *S. pyogenes* Cas9 coding sequence.
- The part of the primer annealing to the Cas9 sequence should be 18-25 bases long.
- The annealing part of both primers should have a melting temperature between 47-54°C.

- Have a G or a C on the 3’ end.
- Have non-annealing 5’ overhangs containing restriction sites for the restriction enzymes you plan to use for cloning. To ensure proper orientation, the forward and the reverse primer should be cut by different enzymes. **Note:** The multiple cloning site (MCS) of the expression sequence is described in Appendix 4.
- Restriction sites included in the overhangs should be located right next to the annealing part of the primers, with no additional nucleotides in between.
- Contain 5 additional nucleotides on the 5’ end to improve activity of the restriction enzymes by ensuring proper binding to the DNA.
- Allow cloning of the Cas9-coding sequence in the same frame as the FLAG-coding sequence present in the vector (i.e. a fusion protein of Cas9 and FLAG will be translated).

You should start with designing the annealing parts of the primers.

To do this for the forward primer you need to:

1. Open the “S.pyogenes-Cas9-cds” input which contains the sequence of the *S. pyogenes* Cas9 coding sequence and copy it.
2. Open the “6. Sequence Editor” tool and paste the sequence into the input field of the tool.
3. Select a sub-sequence from the 5’ end of the Cas9 sequence. You can see the length of the selected subsequence in the “6. Sequence Editor” tool. Copy the selected sequence if it meets the requirements.
4. Check its melting temperature using the “8. Tm Calculator” tool.
5. If all criteria for the annealing part are met, you can paste it into your Notepad.

The general procedure for the reverse primer is similar. However, using additional tools (see Appendix 2) might be needed. Note that both forward and reverse primer sequences should be written in 5' – 3' orientation.

Q.2.1

Enter the sequences of the annealing parts of the primers:

Q2.1.1 Forward Primer annealing part, in 5' to 3' orientation	4.0pt

Q2.1.2 Reverse Primer annealing part, in 5' to 3' orientation	4.0pt

The next step is designing the overhangs. For this, you first need to choose the restriction enzymes you want to use.

Use Appendix 4 to find which enzymes can cut the ‘multiple cloning site’ of the vector. **Note**: the same sequence can be found in the "Input Sequences" menu labeled "MCS".

You should also investigate which enzymes may cut the Cas9-coding sequence. Copy the corresponding sequence and paste it into the input field of the “4. Restriction Mapper” tool. It will return the number of cut sites for commonly used restriction enzymes.

Q.2.2

Based on the results you obtained choose the restriction enzymes that you will use for cloning.

Q2.2.1 Forward Primer	1.0pt

Q2.2.2 Reverse Primer	1.0pt

Q2.3

Determine the reason why the following restriction enzyme **pairs** are **inappropriate** for cloning the Cas9-coding sequence into the mammalian expression vector.

Q2.3.1 PmeI and XhoI	1.0pt

Q2.3.2 BamHI and HindIII	1.0pt

Reasons (you may pick more than one for each enzyme pair):

A. At least one of the enzymes will cut *Cas9* gene inside the coding sequence.

B. At least one of the enzymes will cut the MCS twice.

C. The FLAG-coding sequence will be lost.

D. The FLAG-coding sequence will not be in-frame with the Cas9-coding sequence.

Now you have to design the primer overhangs to meet all the requirements listed previously.

Q2.4

Enter the primer overhang sequences in the fields below:

Q2.4.1 Forward Primer overhang, in 5' to 3' orientation	4.0pt

Q2.4.2 Reverse Primer overhang, in 5' to 3' orientation	5.0pt

Q2.5

Indicate if the following statements regarding the cloning procedure and mammalian expression vector is **True or False**.

Q2.5.1 Cloning the Cas9 gene ($>$4000 bp) requires using Pfu polymerase (error rate 1.3x10^{-6}) instead of Taq polymerase (error rate 1.8x10^{-4}).	1.0pt

Q2.5.2	Codon optimization (replacing naturally rare codons with more common codons encoding the same amino acids) may increase the expression level of Cas9.	1.0pt

Q2.5.3	The vector should contain an antibiotic-resistance gene for the selection of bacterial colonies.	1.0pt

Q2.5.4	Bacterial RNA polymerase can start transcription from the transcription start site adjacent to the multiple cloning site.	1.0pt

In humans, sickle-cell anemia is caused by a single nucleotide mutation in one of the beta-globin genes, which results in the substitution of glutamate for valine at 6th amino acid position (E6V substitution; GAG codon changing to GTG).

In this task you need to design two nucleotide sequences:

1. A gRNA which will guide the Cas9 enzyme towards the proper mouse *Hbb* gene.
2. A Homology Directed Repair (HDR) template (see Appendix 3) which will allow introducing the appropriate mutation.

The table below contains annotations for the GenBank file (accession number: X14061.1) (labeled "*M.musculus*-Hbb-complex" in the "Input Sequences" menu), containing the **mouse** genomic DNA sequence of the beta-globin complex. This sequence is 55856 bp, and includes several genes and pseudogenes (genes which do not produce functional products).

Note: the coordinates of the exons are inclusive on both ends (a region defined as 1..2 spans 2 bp).

The wild-type coding sequence of this **human** beta-globin is available in the "Input Sequences" menu labeled "*H.sapiens*-Hbb-cds".

Type	Name	Expression	Coordinates for exons
Gene	*Hbb-y*	High expression level in early embryo	12781..12872, 13200..13422, 14274..14402
Gene	*Hbb-bh0*	Low expression level in early embryo	12781..12872, 13200..13422, 14274..14402
Gene	*Hbb-bh1*	High expression level in late embryo	21102..21194, 21301..21522, 22331..22459
Pseudogene	*Hbb-bh2*	None	23786..23882, 23963..24182, 24905..25030
Pseudogene	*Hbb-bh3*	None	30435..30525, 30617..30707, 31409..31534
Gene	*Hbb-b1*	High expression level in pups and adult	38339..38430, 38547..38769, 39424..39552
Gene	*Hbb-b2*	Low expression level in pups and adult	53548..53652, 53757..53978, 54607..54735

Q2.6 Select the mouse gene to target with CRISPR-Cas9 which would probably create the best model of human sickle-cell disease. 1.0pt

Q2.7

True or False?

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

Q2.7.1 The N-terminal methionine residue is present in the mature human beta-globin chain. 1.0pt

Q2.7.2 The coding sequence of the target mouse beta-globin chain (chosen in Q2.6) has the same length as the coding sequence of the human beta-globin chain. 1.0pt

Q2.7.3 The sequences of the last 10 amino acid residues in the target mouse beta-globin chain and human beta-globin chain are identical. 1.0pt

Q2.7.4 Both mouse beta-globin pseudogenes result from reverse transcriptase activity. 1.0pt

Q2.7.5 The wild-type sequence of the target mouse beta-globin chain contains a glutamate in the 6th position (the target position) 1.0pt

Q2.8	Choose the appropriate sequence for part of the gRNA hybridizing to the target sequence (excluding PAM; shown in dark blue in Appendix 3) from those provided in Q2.8.	1.0pt

Q.2.9	Choose the N nucleotide within the NGG sequence (PAM sequence) for the appropriate gRNA sequence.	1.0pt

Analyze the following HDR template sequence. Shaded fragments are flanking sequences which are homologous to the target sequence.

```
5'-CACAGCATCCAGGGAGAAAT-[160 nucleotides]-CAACCCCAGAAACAGACATC
          ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCC
CTGTGGGGAAAGGTGAACTC-[160 nucleotides]-ACCCTTGGACCCAGCGGTAC-3'
```

This sequence is also available in the "Input Sequences" menu labeled "HDR-template". There, each of the three parts of the sequence are separated onto different lines.

Q.2.10

True or False?

Q2.10.1	The central sequence (non-shaded) in the HDR template should be divisible by 3.	1.0pt

Q2.10.2	Using this HDR template will change the total number of amino acid residues in the target mouse beta-globin chain.	1.0pt

Q2.10.3	Non-Homologous-End-Joining repair may induce frame shifting in the Hbb gene.	1.0pt

Q2.10.4	Using this HDR template is likely to affect the amino acid sequence of the mouse embryonic beta-globin chain, encoded by the *Hbb-bh1* gene.	1.0pt

END OF TASK 2

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 5).

CLUSTAL

Below is an example of a sequence alignment using the *"clustal"* format

CLUSTAL X (1.81) multiple sequence alignment

```
Examp1   GAGAGGGAGCCTGAGAGATGGCTACCACATCCAAGGAAGGCAGCAGGCGC
Examp2   CACAGGGGCACTGAGACACGGGCCCCACTCCTACGGGAGGCAGCAGTTAG
Examp3   GAGATGGAACCTGAGACAAGGTTCCAGGCCCTACGGGGCGCAGCAGGCGC
          * * **    ****** * **    *      * * **    *******

Examp1   GCAAATTACCCAAT---------CCTGATTCAGGGAGGTAGCGACAGAAA
Examp2   GAATCTTCCGCAATGGGCGCAAGCCTGACGGAGCGACGCCGCTTGGAGGA
Examp3   GAAACCTCCGCAATGCACGAAAGTGCGACGGGGGAAACCCAAGTGCCAC-
         * *    * * ****             **     *   *
```

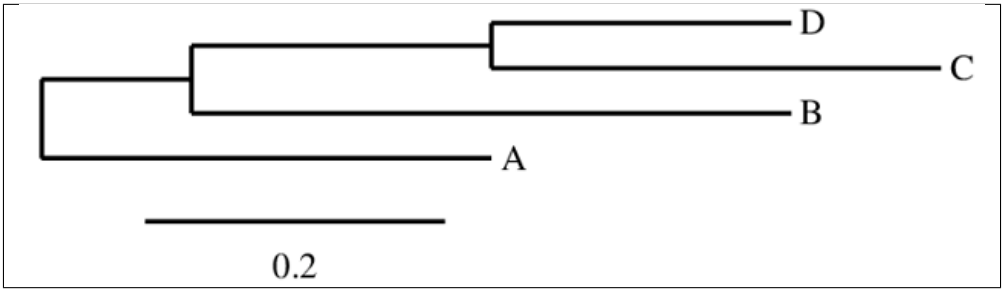
The first line is a header. It specifies the alignment format. The header is followed by sequence blocks (in the example above there are two blocks). Each block has the sequence labels followed by the corresponding sequences. In this example there are three sequences labeled 'Examp1', 'Examp2' and 'Examp3'.

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 in the *"newick"* format, followed by a graphical representation of the tree it describes:

(A:0.3,(B:0.4,(C:0.3,D:0.2):0.2):0.1);



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 ";".

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 ("*clustal*") 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
3 ORF Finder	Find open reading frames in nucleotide sequence	1. Minimum ORF length (numeric) 2. Genetic code type (numeric) 3. Nucleotide sequence ("*fasta*")	ORF size (in amino acid residues), DNA coding sequence and encoded protein 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	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*)*

Appendix 3. Cas9: function and application

To undertake genome editing of mammalian cells, scientists transform them with a vector to synthesise Cas9 enzyme and gRNA. Cas9 then cleaves DNA at the locus targeted by gRNA.

This can trigger the homology directed repair (HDR) pathway, which repairs the the double-strand break using complementary sequences (either sister chromosomes, or an artificially introduced repair template). The repair template can contain a desired mutation, or even an additional DNA fragment.

An alternative mechanism used by cells to repair double-strand breaks is Non-Homologous-End-Joining (NHEJ). NHEJ typically deletes or inserts several random nucleotides whilst fusing the cut DNA back together.

An overview of the CRISPR-Cas9-mediated genome editing is shown in the figure below.

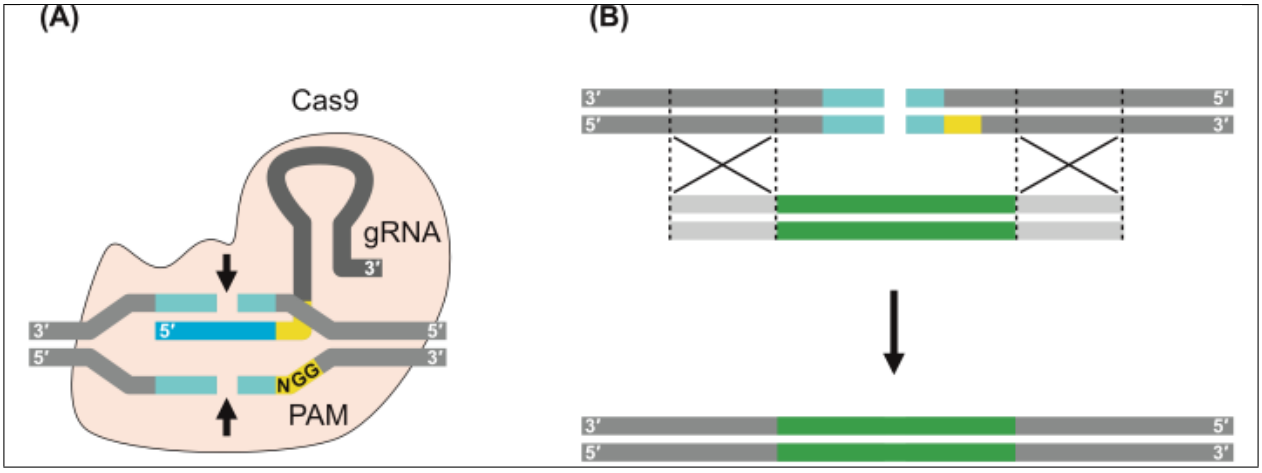


Figure **A.** shows how the Cas9 enzyme recognizes the cleavage site and introduces a double strand break into the genomic DNA. The target sequence is indicated in blue, while the PAM sequence (see main exam text) is shown in yellow. **B.** shows repair by HDR. Homologous recombination, which occurs between genomic DNA and flanking sequences in the HDR template, is indicated by crossed lines

Appendix 4. Multiple cloning sites of the vector

The figure below presents the sequence, with some features labelled, of the 'multiple cloning site' of the mammalian expression vector that you plan to use in your cloning.

First, the vector will be transformed into *Escherichia coli* for amplification. It is then isolated from bacterial cells, sequenced, and transfected into mammalian cells to produce recombinant Cas9-FLAG enzyme.

FLAG is a commonly used artificial tag recognised by tool antibodies for immunodetection of recombinant proteins. In this case, western blotting may be performed on a crude lysate of transfected mammalian cells to confirm proper expression of the recombinant Cas9 enzyme.

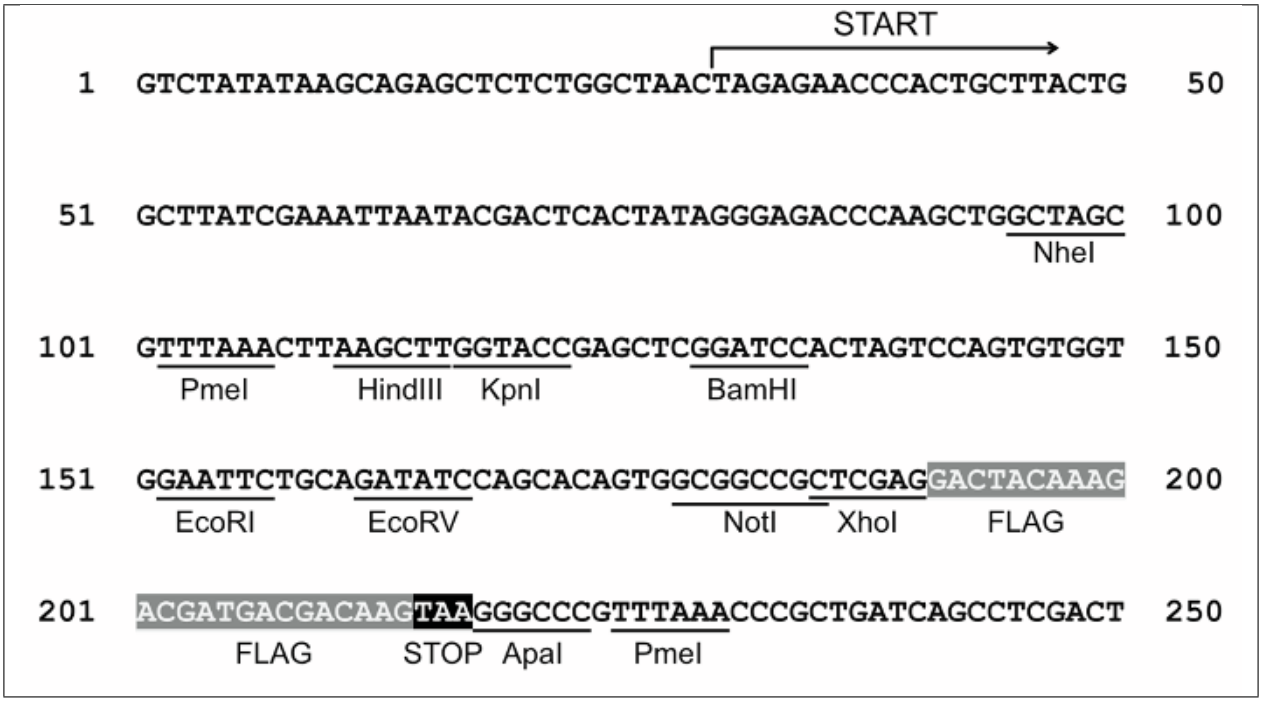


Figure: Multiple cloning site (MCS) and promoter region of the vector.

Legend: START – transcription start site FLAG – coding sequence for the FLAG epitope STOP – stop codon for FLAG-coding sequence Underlined sequence is recognized by the enzyme indicated under the sequence. Numbers either side of each line of sequence indicate the nucleotide position, within the vector sequence, at the start and end of that line, respectively

Appendix 5. Standard Genetic Code table

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

Amino acids encoding with 3- and 1-letter 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	T
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

Bioinformatics answers key and marking

Maximum score: 98

Q1.1 (2 points)

	T	F
Q1.1.1	X	
Q1.1.2		X

Q1.1.1 There are gaps in the alignment (‘-’) suggesting insertions/deletions happening during the evolution of this sequence.

Q1.1.2 *C. paramecium* sequence is the longest. It can be seen from several blocks like the one below

```
H.sapiens_16S          TCTCGGCGCCCCCTCGATGC------------------------------

S.cerevisiae_16S       TTCTGGCTAACCTT--GAGT------------------------------

C.paramecium_16S       GTATCGAATACAATCGGAGTGAAGGGAATATGCTTCTTTTTTTTTTTTAA

T.thermophilus_16S     ----------------GCGT------------------------------

E.coli_16S             ----------------AAGT------------------------------

S.pneumoniae_16S       ----------------AAGT------------------------------

Z.mays_16S             ----------------AAGT------------------------------

P.syntrophicum_16S     ----------------AAGT------------------------------

M.formicicum_16S       ----------------AAGT------------------------------

T.peptonophilus_16S    ----------------AAGT------------------------------
```

where C. paramecium has extra sequence parts not present in any other organism compared. Also smart ones can take the sequence and paste it into the sequence editing tool that will return the length of the sequence.

Q1.2 (5 points)

Here we evaluate that the input is in clustal format and there are only 10 sequences with correct labels and length.

For each sequence with the correct label and length one would get 0.5 point so 5 points is the maximum for this task.

If there are more than 10 sequences in the submitted result, each extra sequence should result in a penalty of -0.5 points. This means that having either nine 16S sequences or ten 16S sequences and one 23S will both result in 4.5 points. But the total for the question can't be negative.

Key:

H.sapiens-16S:1869
S.cerevisiae-16S:1800
C.paramecium-16S:1970
M.formicicum-16S:1476
P.syntrophicum-16S:1495
T.peptonophilus-16S:1489
E.coli-16S:1542
S.pneumoniae-16S:1546
Z.mays-16S:1491
T.thermophilus-16S:1519

Q1.3 (5 points)

The tree will be graded by checking each of the 10 terminal branches for the correct label and correct length of the branch.

0.5 points for each correct label and terminal branch length combination.

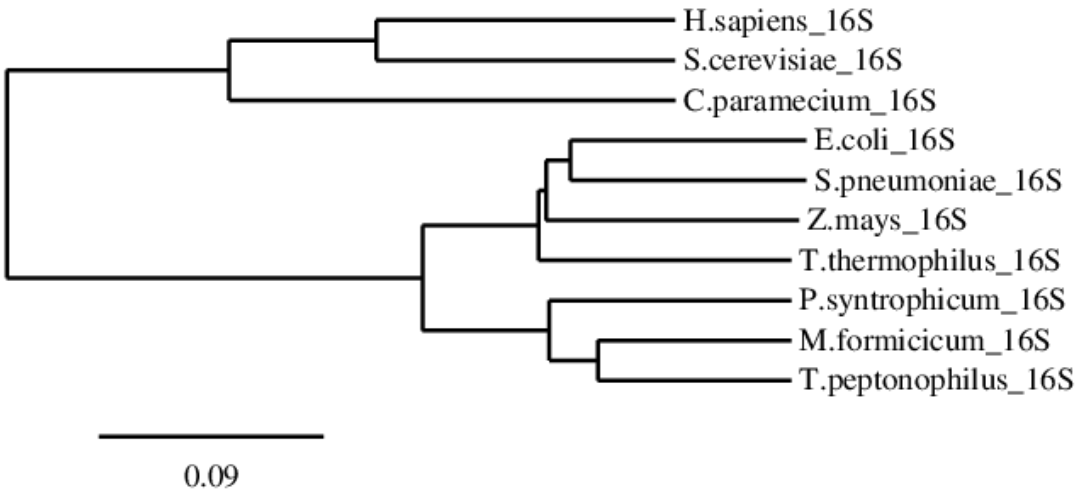
Key:

T.peptonophilus-16S:0.07977
M.formicicum-16S:0.07977
P.syntrophicum-16S:0.09667
T.thermophilus-16S:0.10347
Z.mays-16S:0. 10347
S.pneumoniae-16S:0.09409
E.coli-16S:0.09409
C.paramecium-16S:0.17539
S.cerevisiae-16S:0.11732
H.sapiens-16S:0.11732

Correct newick:

((((T.peptonophilus-16S:0.07977,M.formicicum-16S:0.07977):0.01689,P.syntrophicum-16S:0.09667):0.05146,(T.thermophilus-16S:0.10347,(Z.mays-16S:0.10347,(S.pneumoniae-16S:0.09409,E.coli-16S:0.09409):0.01027):0.00088):0.04465):0.16452,(C.paramecium-16S:0.17539,(S.cerevisiae-16S:0.11732,H.sapiens-16S:0.11732):0.05807):0.09261):0.00000;

Correct tree (just to visualize it for the jury; students don't have to submut the image)



Q1.4 (6 points)

	A	B	E
C. paramecium			X
M. formicicum	X		
T. peptonophilus	X		
T. thermophilus		X	
S. cerevisiae			X
S. pneumoniae		X	

Q1.5 (4 points)

	T	F
Q1.5.1		X
Q1.5.2	X	
Q1.5.3	X	
Q1.5.4		X

Z. mays sequence is in the bacteria branch which can be explained by it being either the mitochondrial or the chloroplast 16S rRNA sequence and it is not possible to know which one from the data.

Q1.6 (6 points)

	T	F
Q1.6.1		X
Q1.6.2		X
Q1.6.3	X	
Q1.6.4	X	
Q1.6.5		X
Q1.6.6	X	

Q1.6.1 – False. This particular tree doesn't show it. It is unrooted and hence can't tell you much about evolutionary relatedness between domains of life. Even if we rely on the root being on the longest branch, it is between Eukaryotes on one side and Bacteria + Archaea on the other.

Q1.6.2 – False. No known viruses have sequences homologous to the rRNA genes + even if any are found in the future these can't be considered as an outgroup by default as it can result from a recent acquisition.

Q1.6.3 – True. This is the most parsimonious explanation for all cellular forms having it.

Q1.6.4 – True. As rRNA have a secondary structure with both single and double-stranded regions, molecules with different sequences would have different secondary structure and hence different digestion profile.

Q1.6.5 – False. *M.formicicum* and *T.peptonophilus* are equally close to *P.syntrophicum*.

Q1.6.6 – True. These two species are each other's closest relatives.

Q1.7 (1 point)

1

The standard genetic code can be applied here.

Q1.8 (5 points)

Checked for correct sequence labels and correct sequence length. One point per sequence.

Key:

S.anginosus-Cas9:3378
S.mitis-Cas9:4176
S.pyogenes-Cas9:4104
S.ruminantium-Cas9:3372
S.thermophilus-Cas9:3384

Q1.9 (1 point)

6
+1,+2 +3 frames on both + and – strands

Q1.10 (1 point)

Here we evaluate that the input is in clustal format and there are only 5 sequences with correct labels and length.

For each sequence with the correct label and length one would get 0.2 point so 1 point is the maximum for this task.

If there are more than 5 sequences in the submitted result, each extra sequence should result in a penalty of -0.2 points.

Key:

S.anginosus-DnaJ:378
S.mitis-DnaJ:378
S.pyogenes-DnaJ:378
S.ruminantium-DnaJ:378
S.thermophilus-DnaJ:377

Q1.11 (1 point)

Here we evaluate that the input is in clustal format and there are only 5 sequences with correct labels and length.

For each sequence with the correct label and length one would get 0.2 point so 1 point is the maximum for this task.

If there are more than 5 sequences in the submitted result, each extra sequence should result in a penalty of -0.2 points. But the total can’t be negative.

Key:

S.anginosus-Cas9:1126
S.mitis-Cas9:1392
S.pyogenes-Cas9:1368
S.ruminantium-Cas9:1124
S.thermophilus-Cas9:1128

Q1.12 (1 point)

DnaJ

DnaJ alignment has fewer gaps and more conservative positions. This means DnaJ sequences are less divergent and hence are easier to align. Cas9 sequences, on the other hand, are more divergent and hence alignment errors are more likely.

Q1.13 (2.5 point)

The tree will be graded by checking each of the 10 terminal branches for the correct label and correct length of the branch.

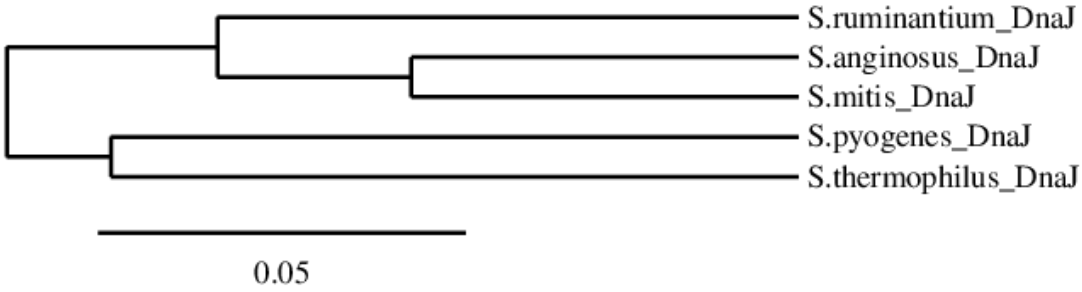
0.5 points for each correct label and terminal branch length combination.

Key:

S.thermophilus-DnaJ:0.09342
S.pyogenes-DnaJ:0.09342
S.mitis-DnaJ:0.05263
S.anginosus-DnaJ:0.05263
S.ruminantium-DnaJ:0.07895

Newick:
((S.thermophilus-DnaJ:0.09342,S.pyogenes-DnaJ:0.09342):0.01414,((S.mitis-DnaJ:0.05263,S.anginosus-DnaJ:0.05263):0.02632,S.ruminantium-DnaJ:0.07895):0.02862):0.00000;

Correct tree (just to visualize it for the jury; students don't have to submut the image)



Q1.14 (2.5 point)

The tree will be graded by checking each of the 10 terminal branches for the correct label and correct length of the branch.

0.5 points for each correct label and terminal branch length combination.

Key:

S.pyogenes-Cas9:0.20096

S.mitis-Cas9:0.20096

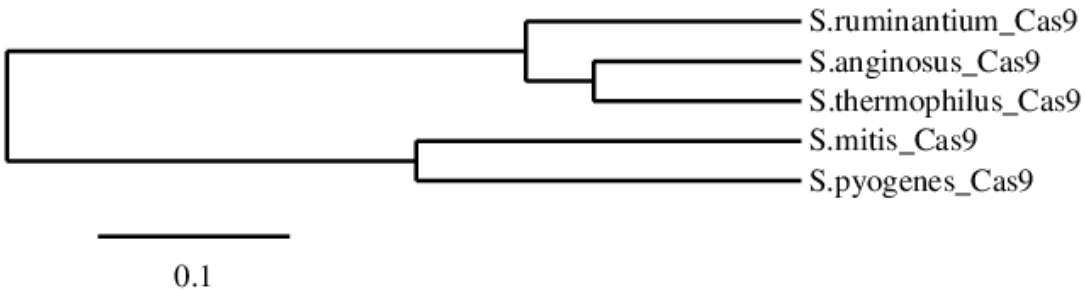
S.thermophilus-Cas9:0.10837

S.anginosus-Cas9:0.10837

S.ruminantium-Cas9:0.14369

Newick:
((S.pyogenes-Cas9:0.20096,S.mitis-Cas9:0.20096):0.21378,((S.thermophilus-Cas9:0.10837,S.anginosus-Cas9:0.10837):0.03532,S.ruminantium-Cas9:0.14369):0.27105):0.00000;

Correct tree (just to visualize it for the jury; students don't have to submut the image)



Q1.15 (5 points)

	T	F
Q1.15.1		X
Q1.15.2		X
Q1.15.3	X	
Q1.15.4		X
Q1.15.5	X	

Q1.15.1 False. DnaJ is more conserved as can be seen from shorter branches of the tree (mind the scale bar).

Q1.15.2 False. As the two trees are not congruent, at least one of them is not reflecting the true phylogeny of the species. Cas9 is likely to be showing a topology different from the species one because of being less evolutionary conserved.

Q1.15.3 True. It is not the only explanation but definitely a possible one: there could be an exchange (not necessarily directly) between S. mitis and S. thermophilus.

Q1.15.4 False.

Q1.15.5 True. Different genes might show different phylogenies because of HGT as well as due to analysis artifacts.

Q1.16 (4 points)

	T	F
Q1.16.1	X	
Q1.16.2	X	
Q1.16.3	X	
Q1.16.4	X	

Q1.16.1 – True. In codons CGA and CGG coding for Arginine and CUA and CUG coding for Leucine any of the three changes of the 3rd nucleotide and one change of the first nucleotide are synonymous.

Q1.16.2 – True. We can't use just the ratio between absolute numbers of nonsynonymous and synonymous substitutions because this ratio under neutrality depends on the sequence and hence will differ from gene to gene.

Q1.16.3 – True. Yes, they are used as a neutral control. Basically, the test can be re-written as dN=dS; dS is used as a neutral expectation and then we compare dN to it. However, this assumption is not always true.

Q1.16.4 – True. By using dS as an "internal" control for a given gene and a given lineage one can compare dN/dS results between different genes or different branches of the tree even if they differ in mutation rate.

Q1.17 (1 points)

1

We are testing if the rate of fixation is the same for the two substitution classes, e.g. dN=dS

Q1.18 (3 points)

	dN/dS = X	dN/dS < X	dN/dS > X
Q1.18.1		X	
Q1.18.2			X
Q1.18.3	X		

Q1.18.1. dN/dS < 1. H4 is very conserved = under strong negative selection = most non-synonymous substitutions are selected against and hence are not observed as differences between lineages. So dN < dS and dN/dS is < 1.

Q1.18.2 dN/dS > X Proteins evolving under an arms race like the proteins involved in pathogen recognition by the host's immune defense are likely to have an accelerated mode of evolution – on average changes to the amino acid sequence are beneficial as they help avoid the immune system.

Q1.18.3 dN/dS = X. A pseudogene is not functional and hence evolves neutrality.

Q1.19 (1 point)

1 point if the sequence labels are correct, there are 5 sequences and the alignment is 3138 nucleotides long. Zero points otherwise.

Q1.20 (4 points)

	T	F
Q1.20.1		X
Q1.20.2	X	
Q1.20.3	X	
Q1.20.4	X	

Q1.20.1 False. Ths would have been true if all the branches had a dN/dS value below 1.

Q1.20.2 True. Both species' terminal branches have dN/dS value much below 1.

Q1.20.3 True. The dN/dS value on this branch is much higher than 1 which can be explained by positive selection.

Q1.20.4 True. If the protein alignment contains a lot of errors (non-homologous codon identified as homologous) this will distort the results of the test. Specifically, misaligning different amino acids will increase the observed dN value.

Q2.1 (8 points)

There is actualy only one pair meeting all the requirements

Q2.1.1

ATGGATAAGAAATACTCAATAGGC

Q2.1.2

GTCACCTCCTAGCTGACTC

Common criteria for both Q.2.1.1 and Q2.1.2 (1 point for each criteria for each sequence, 6 in total)

1. The sequence is within the specified length range (inclusive on both ends). 18 – 25.

2. The melting temperature, as calculated by our tool, is between 47-54 degrees.

3. The last letter in the submitted sequence is "C" or "G".

Q2.1.1 specific criteria (1 point)

The input is a substring of "ATGGATAAGAAATACTCAATAGGCTTAGATATCGGCACAAATAGCGTCGG" starting from the first character and is at least 14 nucleotides long.

Q2.1.2 specific criteria (1 point)

Just as above but the string is different:

"GTCACCTCCTAGCTGACTCAAATCAATGCGTGTTTCATAAAGACCAGTAA"

Q2.2 (2 points)

Q2.2.1 KpnI

Q2.2.2 XhoI

Q2.3 (2 points)

Q2.3.1 PmeI and XhoI B, C

Q2.3.2 BamHI and HindIII A,D

Q2.4 (9 points)

1. Does the sequence contain any of the following strings (restriction sites in the MCS): GCTAGC, TTTAAA, AAGCTT, GGTACC, GGATCC, GAATTC. GATATC, GCGGCCGC, CTCGAG, GGGCCC? One point for each of the two sequences, hence the maximum is 2.
2. Does the sequence contain more than one restriction site from Step 1? If "YES", stop evaluating the sequence keeping the points scored so far. The two sequences are checked independently so the evaluation might be stopped for one but not the other.
3. If both sequences passed Step 2, do they contain the same restriction site? If "YES", penalize the pair with -1.
4. Does the sequence contain any character on the right of the restriction site? Evaluate each of the two sequences, if "NO", +1 point for each sequence, hence the maximum is 2.

5. Does the sequence contain 5 nucleotides on the left of the restriction site? Evaluate each of the two sequences, if "YES", +1 point for each sequence, hence the maximum is 2.
6. Is the restriction site in Q2.4.1 GGTACC (KpnI)? If "YES" +1 point for this sequence.
7. Is the restriction site in Q2.4.2 any of the three following: GCTAGC, GAATTC, CTCGAG (these are in frame with FLAG)? If "YES" +1 point for this sequence.
8. Is the restriction site in Q2.4.2 CTCGAG (XhoI)? If "YES" +1 point for this sequence.

Q2.5 (4 points)

	T	F
Q2.5.1	X	
Q2.5.2	X	
Q2.5.3	X	
Q2.5.4		X

Q2.5.1 - True. WIth such a length error rate becomes important so it is necessary to use a high-fidelity polymerase.

Q2.5.2 - True. Yes, codon usage might differ between bacteria where the gene comes from and eukaryotes where we want it to be expressed.

Q2.5.3 - True. We first want to amplify the construct in bacteria.

Q2.5.4 - False. This is a eukaryotic expression vector so it should not be recognized by bacterial transcription machinery.

Q2.6 (1 points)

Hbb-b1

To recreate the sickle-cell anemia in mice we want the major adult beta-globin chain to be affected.

Q2.7 (5 points)

	T	F
Q2.7.1		X
Q2.7.2	X	

Q2.7.3		X
Q2.7.4		X
Q2.7.5		X

Q2.7.1 - False. The GAG codon is 7th in the provided cds but the text tells us that glutamate is the 6th amino acid residue. Hence, the starting methionine is removed.

Q2.7.2 - True. Both are 146 (without the starting methionine) amino acid residues long.

Q2.7.3 - False.

`h.sapience VANALAHKYH`

`m.musculus VATALAHKYH`

Q2.7.4 - False. Those pseudogenes contain introns and hence don't originate from mature mRNAs.

Q2.5.5 - False. The mouse protein has alanine (A) in this position.

Q2.8 (1 point)

E

Only sequences D and E are present in the mouse beta-globin complex with the following coordinates:

D: 47938-47957

E: a) 38344-38363 b) 53553-53572

But only E recognizes the correct gene Hbb-b1. Note that 38344-38363 is followed by AGG while 53553-53572 is followed by AGT so the latter will not be cut.

Q2.9 (1 point)

A

See question Q2.8.

Q2.10 (4 points)

	T	F
Q2.10.1		X
Q2.10.2		X
Q2.10.3	X	

Q2.10.4		X

Q2.10.1 - False. You don’t need to replace the whole coding sequence, you can replace just a part of it which doesn’t have to have a whole number of codons.

Q2.10.2 - False. The template is 442 bp long and it will replace a sequence 443 bp long hence causing a single bp deletion. However, this deletion happens just upstream of the first ATG codon and hence won’t affect the length of the protein.

Q2.10.3 - True. If instead of homologous recombination the double-strand break is repaired through Non-Homologous-End-Joining some nucleotides can be deleted or inserted. As the cut is introduced within the coding sequence, this might lead to frameshifts.

Q2.10.4 - False. The guide RNA is not homologous to it.