i Instructions

Academic responsible: Lars Martin Jakt

Grading deadline: 10.12.2021

Grading scale: A-F

Allowed aids: Pen, ruler, simple calculator and up to two bilingual dictionaries.

Grounds: Students may demand grounds for the assessment of their examination performance. The deadline is one week after the announcement of the original result.

Cheating: Cheating or an attempt to cheat may lead to annulment and suspension, cf. regulations relating to studies and examination section 9.1.

This exam contains a total of 38 questions with a total of 179 marks available.

You are not meant to answer all of the questions but you should choose at least two from every section; after this you may choose freely from the remaining questions until the combined maximum score exceeds 100. This means that you could in theory choose a set of questions whose marks add up to a little bit above 100. Any questions answered after this will not be marked.

If you start to answer a question but decide that you would prefer to answer a different question, then please simply delete your answer (the question will probably still be flagged as answered, but we will ignore any empty answers), or state clearly that you do not want the answer to be marked.

You should spend some time to choose the questions you wish to answer. The easiest way to do this is to have a look at the table of contents and then going to each section individually and finding at least two questions that you want to answer. These can be bookmarked so that you can easily get to them from the table of contents. If you fail to provide two questions from a section then one or two questions (with the lowest amount of score available) will be considered as answered and the marks from those will contribute to the maximum number of marks available to you (i.e. you cannot make up for those marks by answering questions from other sections).

The maximum number of marks available for each question is my estimate for the time that you may need to answer the question. This includes the time to read and think about how to answer the question; some questions will require more thinking time than others, so the marks available from the questions are not a direct function of the amount I expect you to write. You have approximately 1.8 minutes per mark.

1.1 DNA and RNA

What are DNA and RNA and how do they differ? In your answer you should consider the following:

1. The basic structure of the DNA and RNA molecule.
2. The primary difference in the structure and property of RNA and DNA.
3. How the difference in the property of RNA and DNA relate to their most common functions.
4. How the structure of DNA facilitates its most common function.

Please divide your answers into the 4 parts specified.

Fill in your answer here

Maximum marks: 7

1.2 DNA direction

The following is the sequence of short piece of DNA:
5' ACTGATAGA 3'

1. Why do we sometimes include the 5' and 3' when giving a nucleotide sequence.
2. Give the reverse complement of the above sequence. Include the 5' and 3' specifiers correctly.
3. What is the usual relationship between the forward and reverse complement of a sequence (i.e. where will you find forward and reverse complementary sequences).
4. Give the sequence of the RNA molecule that would be created by transcription by a promoter lying to the left of the sequence.

Fill in your answer here

Maximum marks: 6

1.3 DNA strandedness

DNA is usually double stranded. Consider:

1. How double-strandedness might be useful.
2. A potential problem of single stranded DNA.

Fill in your answer here

Maximum marks: 4

1.4 Translation frames

Given the following genetic code:

```
  AAs   = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M------**--*----M---------------M----------------------------
Base1   = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2   = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3   = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Provide all the amino acid sequences that can be encoded by the following (double stranded) DNA sequence:

5' ATGGCACGGTGA 3'

Your answer should show your working and logic.

Fill in your answer here

Maximum marks: 6

1.5 Mammalian genome content

What constitutes the majority of mammalian genomes? What does this tell us about how genome size has evolved in the vertebrates?

Fill in your answer here

Maximum marks: 4

1.6 Genome components

What do genomes contain?

Consider:

1. Structural components.
2. Functional components.
3. Sequence motifs that may or may not have function or activity.

How does the content of these different components vary between large and small genomes. For example, the human, fruit fly (Drosophila) and bacterial genomes.

Fill in your answer here

Maximum marks: 6

1.7 CpG islands

1. What are CpG islands?
2. How do they arise and what are they associated with?

Fill in your answer here

Maximum marks: 5

2.1 Basic alignment

Align the following two sequences using either a global or local alignment.

1. ATGAATTCGGA
2. AGGACCGATCA

Describe how you can assess the quality of the alignment (i.e. provide a reasonable scoring system) and give the scores for both of your alignments.

Note that more than one reasonable scoring system can be used; you need only provide one here.

Fill in your answer here

Maximum marks: 7

2.2 Trace an alignment

Consider the following matrix:

		C	G	A	T	A	C	G	T	G	A
	0	0	0	0	0	0	0	0	0	0	0
			↖					↖		↖	
G	0	0	4	0	0	0	0	4	0	4	0
					↖				↖		
T	0	0	0	0	4	0	0	0	8	0	0
					↖				↖	↖	
T	0	0	0	0	4	0	0	0	4	4	0
				↖		↖					↖
A	0	0	0	4	0	8	0	0	0	0	8
		↖					↖				
C	0	4	0	0	0	0	12 ←	4 ←	3 ←	2 ←	1
					↖		↑	↖	↖		
T	0	0	0	0	4	0	4	8	8	0	0
			↖				↑	↖	↖	↖	
G	0	0	4	0	0	0	3	8	4	12 ←	4
				↖		↖	↑		↖	↑	↖
A	0	0	0	8	0	4	2	0	4	4	16
		↖			↖		↖			↑	↑
C	0	4	0	0	4	0	8	0	0	3	8

1. What algorithm produces a matrix like this?
2. What is that algorithm used for?
3. Extract the optimal alignment from this matrix.

Fill in your answer here

Maximum marks: 6

2.3 Alignment score matrix

Consider the following table:

		C	G	A	T	A	C	G	T	G	A
	0	← -8	← -9	← -10	← -11	← -12	← -13	← -14	← -15	← -16	← -17
G	↑ -8	↖ -4	↖ -4	← -12	← -13	← -14	← -15	↖ -9	← -17	↖ -11	← -19
T	↑ -9	↖ -12	↖ -8	↖ -8	↖ -8	← -16	← -17	↑ -17	↖ -5	← -13	← -14
T	↑ -10	↖ -13	↖ -16	↖ -12	↖ -4	↖ -12	↖ -20	↑ -18	↖ -13	↖ -9	↖ -17
A	↑ -11	↖ -14	↖ -17	↖ -12	↑ -12						
C	↑ -12										
T	↑ -13										
G	↑ -14										
A	↑ -15										

Match 4	Mismatch -4
Gap open -8	Gap ext. -1

1. What is shown in the figure?
2. What kind of alignment can be derived from the completed table?
3. Provide the scores and arrows for the next three cells in row 5. You may give the arrows as U (up), L (left) and D (diagonal) (eg. -4L, would indicate a score of -4 and a leftwards pointing arrow).

Fill in your answer here

Maximum marks: 6

2.4 Scoring gaps

Consider these two alignments:

-
  ```
  ATTACTTAGGATTATAGA
  ||   | | || | |  |
  AT---T-A-GA-T-T--A
  ```

-
  ```
  ATTACTTAGGATTATAGA
  ||||     |||||
  ATTA-----GATTA----
  ```

Which of these is better and why? Describe how you can design a scoring system that gives a higher score to the better alignment. Why would such a scoring system make sense from an evolutionary perspective?

Fill in your answer here

Maximum marks: 6

2.5 Evolutionary homology

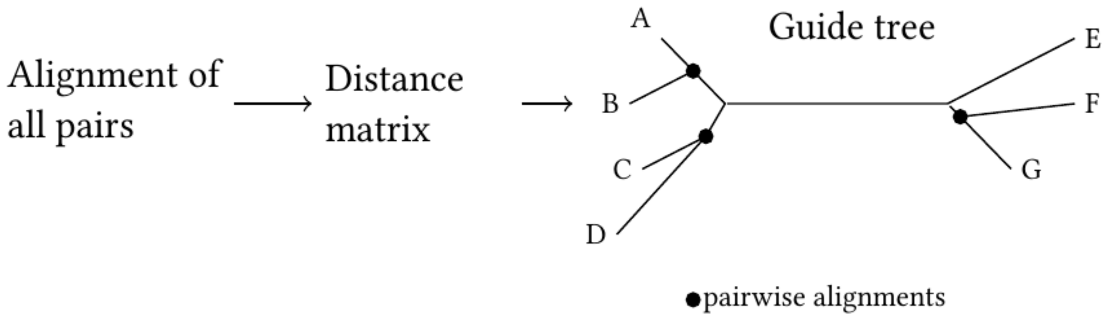
What are homologues, orthologues and paralogues. Describe how they arise during evolution.

Fill in your answer here

Maximum marks: 5

2.6 Clustal MSA

The following figure summarises the way in which clustal performs multiple sequence alignment.



Describe what is done by clustal at each step and the reasons for performing each step.

Fill in your answer here

Maximum marks: 6

2.7 Why align sequences

Why is it useful to align multiple sequences to each other? Give several examples where multiple sequence alignments are used.

Fill in your answer here

Maximum marks: 5

3.1 R logical subset

Consider the following code:

```
v1 <- 1:10
b <- v > 3 & v < 6
v2 <- v[b]
```

Describe what each line does and give the values of the resulting variables.

Fill in your answer here

Maximum marks: 3

3.2 R matrix

Consider the following code:

```
m1 <- matrix(1:12, nrow=4)
m2 <- rbind( c(1:4), c(5:8), c(9:12) )
m3 <- t(m1)
```

1. Show the contents (the values) of m1, m2, and m3.
2. How would you access the 3rd element of the second row of m1?
3. How would you get the sum of the first column of m2?

Fill in your answer here

Maximum marks: 3

3.3 R function

Consider the following code:

```
f1 <- function(x){
     s <- 0
     for(i in 1:length(x)){
         s <- s + x[i]
     }
     s / length(x)
}

v <- c(4, 6, 5)
f1(v)
```

Describe in words what the code does. What is the result of the last operation?

Fill in your answer here

Maximum marks: 4

3.4 R data types

Consider the following code in R:

```
a <- c('one'=3, 'two'=7, 'three'=15)
b <- list('a'=20, 'b'='hello', 'c'=FALSE)
```

Give:

1. What types of data are a and b and what is the fundamental difference between the data types?
2. Two ways in which you can access the second element of a
3. Three ways in which you can access the second element of b

Fill in your answer here

Maximum marks: 3

3.5 32 bit integers

R can only handle integral values as 32 bit signed integers. What (approximately) are the minimal and maximum values that a 32 bit signed integer can hold. Give an example of where this may cause a problem.

Fill in your answer here

Maximum marks: 4

3.6 A model of a gene

Given that a gene is composed in it's simplest form by the positions and order of its exons and its sequence, suggest how you could represent a gene in a computer program.

Fill in your answer here

Maximum marks: 5

3.7 High and low level computing

What is meant by high and low level computing? As a beginner which are you more likely to learn?

Fill in your answer here

Maximum marks: 4

3.8 Numerical values

Describe the two main ways in which numerical values are represented in computer programs. How are these main categories divided into further subtypes?

Fill in your answer here

Maximum marks: 4

3.9 Scripts vs programs

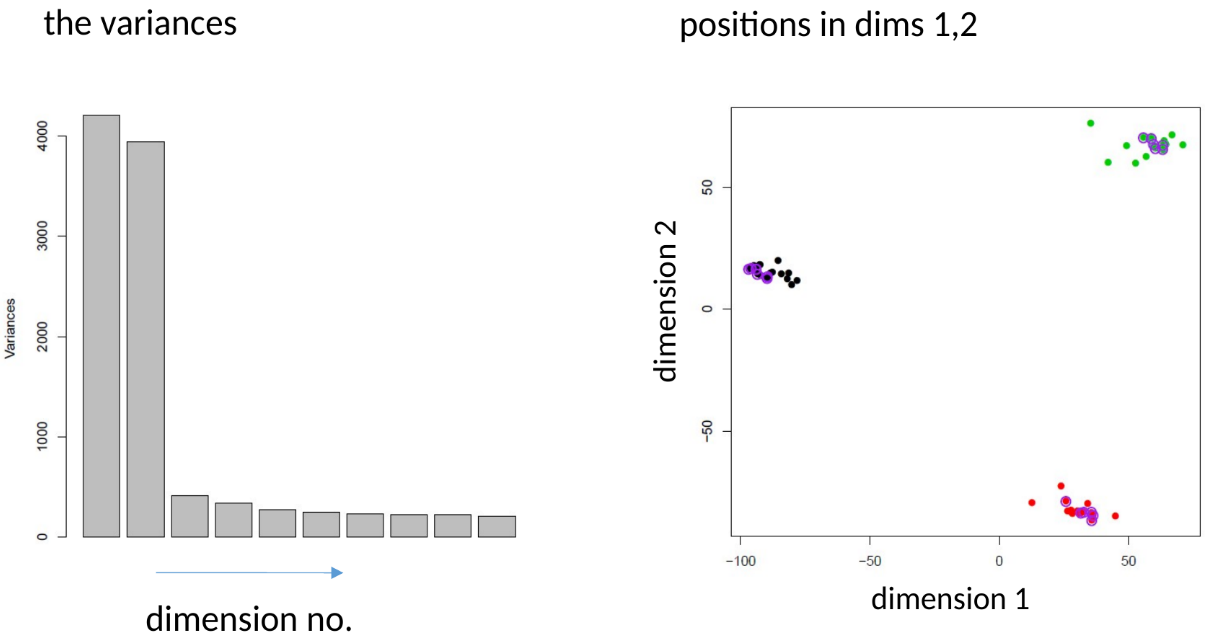
Although there is no clear separating line between what constitutes a script and a program describe some (3-4) general properties that are often used to classify a program as a program or a script. Consider how the code is meant to be used as well as how it is written (eg. choice of language).

Fill in your answer here

Maximum marks: 4

4.1 PCA

The following figure shows the result of a Principal Components Analysis (PCA) on gene expression levels from 54 samples representing 3 different tissue types that have been exposed to 2 different drugs or control conditions. The left-hand plot shows the variances of the first 10 dimensions; the right hand panel shows the samples plotted by the two first positions. The colour of points indicate the tissue type with purple circles indicating samples not exposed to drug.



1. What can you infer about the effect of tissue type and drug treatment on gene expression?
2. Why are there exactly two dimensions that have much larger variance than the others?
3. How many dimensions are required to describe the data completely?

Fill in your answer here

Maximum marks: 6

4.2 Incidental factors

Consider the following scenario:

An experiment was performed to determine the difference in gene expression in the liver of two different strains of a small experimental fish. The strains can be distinguished by the patterning on their skins, and the researchers decided to keep the fish in the same tank to minimise any tank effect. Three fish from each strain were isolated and the liver dissected out followed by RNA extraction. One fish was processed at a time with fish from strain one obtained first, followed by fish from strain two. The extraction procedure was reasonably time consuming and the researcher had lunch at some point during the procedure.

The RNA was converted to cDNA and high-throughput sequencing performed to estimate global gene expression levels. The analysis found differential expression in about 1500 genes.

Assuming that the statistical analyses were done correctly (with p-values correctly adjusted):

1. Would it be reasonable to conclude that the differential gene expression observed is due to the difference in strain?
2. Discuss what other factors may be responsible for the differential gene expression observed.
3. How could the experiment be made more robust?
4. Why is the effect of unknown factors a bigger problem in genomics than in experiments where smaller number of measurements are made per sample?

Fill in your answer here

Maximum marks: 8

4.3 Multiple testing

You have analysed the expression levels of 20,000 genes in two different sample types. After performing a robust statistical test you find that 1,500 genes have a p-value that is below 0.05. How many of these genes do you think are really differentially expressed?

You should explain your reasoning and state the name of the statistical procedure you have followed if you remember it. Note that for this exercise you may consider that a p-value of 0.05 is meaningful for a single statistical test.

Fill in your answer here

Maximum marks: 6

4.4 Simple equations

Given a series of n values x, consider the following series of equations:

1. $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$
2. $\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
3. $\sigma = \sqrt{\sigma^2}$

Please:

1. Describe what the equations calculate and the usual names of the results
2. Describe one use of these equations
3. Describe how the quantities obtained relate to normal distributions
4. These equations are commonly used; however there are circumstances under which they are not suitable. Please provide an example of when not to use the above equations.

Fill in your answer here

Maximum marks: 5

4.5 Object descriptors

Given two objects (can be samples, individuals, sequence sets, etc.) described by a set of n measures given in the vectors A and B, consider the following equation :

$$d = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

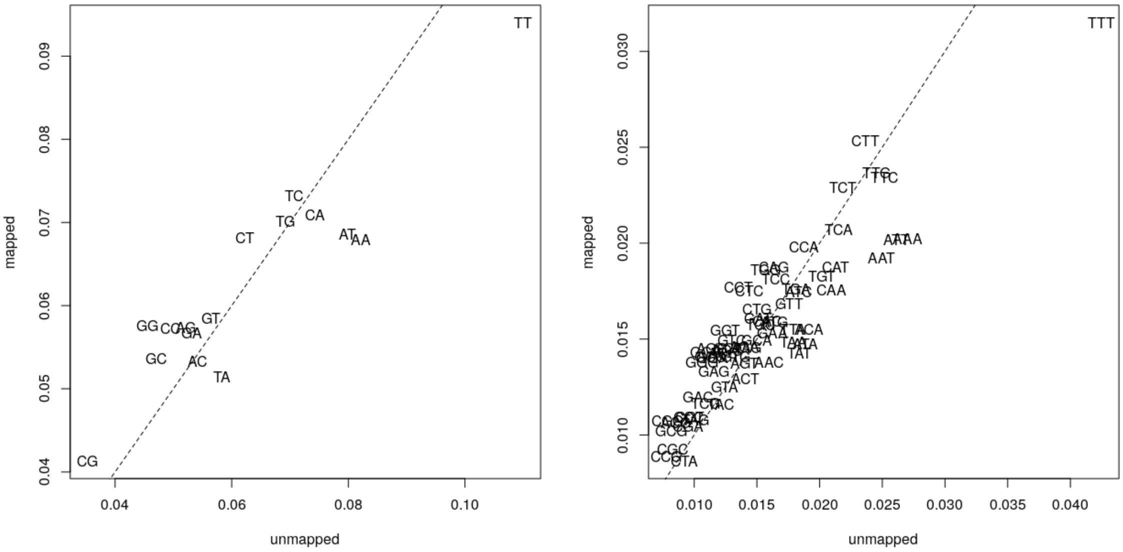
1. Describe in words what is calculated by the equation.
2. What does the quantity d represent?
3. Given a larger set of objects $A, B, C, \ldots$, describe how you might use this equation to tell you something about the relationships between objects.

Fill in your answer here

Maximum marks: 5

4.6 kmer analysis

An experiment has been done where RNA molecules were obtained from an environmental sample and subjected to sequencing. BLAST was used to search for sequences in genbank that could be aligned to 1 million of the sequences obtained. About half of the sequences mapped correctly to sequences in genbank; however no alignments were obtained from the remaining half. The sequences were then divided into two sets (mapped: those that aligned, and unmapped; those that did not) and the frequency of dimers and trimers calculated for the whole set. Plotting unmapped vs mapped provided the following two graphs:



These plots suggest a number of things; give three examples and explain why you think that data suggest these. Include additional analyses that could be done to provide further evidence for one of your inferences.

Fill in your answer here

Maximum marks: 6

4.7 The p-value

1. What is a p-value?
2. What distribution of values should you obtain if you calculate a large number of p-values when the null hypothesis is correct?
3. Explain why your answer to part (2) is reasonable.

Fill in your answer here

Maximum marks: 4

5.1 Database interfaces

Remote databases (eg. Ensembl, NCBI) can often be accessed either through web browsers or public application programming interfaces (APIs). Describe the advantages and dis-advantages of the two methods.

Fill in your answer here

Maximum marks: 4

5.2 Blast e-value

The program BLAST can be used to search a database for sequences that can be aligned to one or more query sequences. In addition to returning the alignments (optional) BLAST also reports a score and an e-value for each alignment it identifies. Describe what these values represent.

Fill in your answer here

Maximum marks: 3

5.3 Novel sequence

Describe what you would do if you were given a single nucleotide sequence and tasked with determining what it is. If you are able to identify the sequence as a vertebrate orthologue where would you go to:

1. Identify related sequences
2. determine the likely biological function of your sequence

What could you do if you have not identified any clearly homologous sequences but you suspect that the sequence may encode a protein?

Note that there are many different ways to answer this question and to some extent what you should do in subsequent steps depends on what you find in earlier steps, and it is fine for you to say, 'if step 1 gives this, then I would do this, followed by...', as you only have time to describe a distinct scenario. In your answer you should state how long the sequence is that you are analysing as this makes a differences as to the steps you might take.

Fill in your answer here

Maximum marks: 5

5.4 Blast types

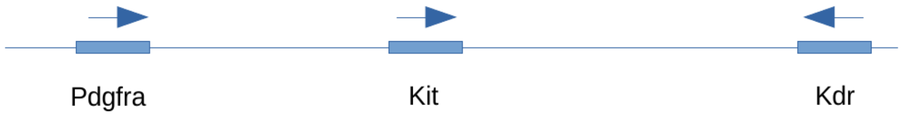
There are four main types of Blast searches. Describe how they differ and give their names if you remember.

Fill in your answer here

Maximum marks: 4

5.5 kit_synteny

Whilst investigating the function of the Kit gene you have found that its neighbouring genes are usually the same and are arranges as follows:



- tyrosine kinase receptor molecules
- syntenic in many (all?) vertebrates

Both its neighbours have similar molecular functions (transmembrane receptor tyrosine kinases).

1. What does this suggest about the evolution of these three genes?
2. How would you go about to find evidence for or against this suggestion?
3. Is there anything else implied by these observations?

Note that although this question is in the database section it is actually a much more general question and your answer should ideally contain elements from several topics covered in the course.

Fill in your answer here

Maximum marks: 10