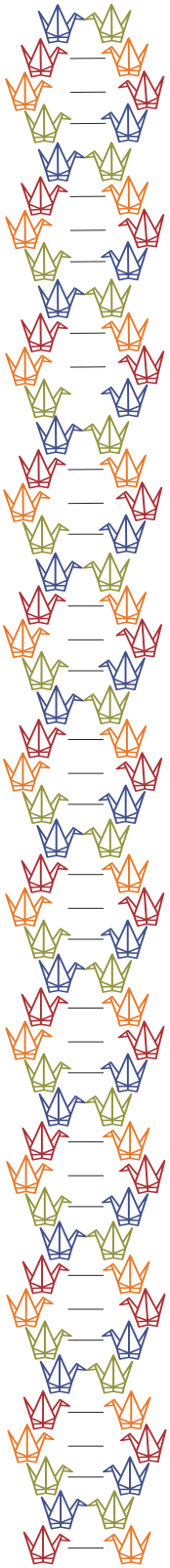
INTERNATIONAL BIOLOGY OLYMPIAD e. V. IBO

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.



Practical Exam 2

Bioinformatics

signature

2020.8.11.

IBO Challenge 2020

A Substitute for The 31st IBO 2020 Nagasaki, JAPAN

Practical Exam 2, Bioinformatics

Use the IBO Challenge 2020 Bioinformatics (IBOC) application to address the following questions. The IBOC application may be accessed using any Web Browser.

Find the URL of your country's server at:
URL: **https://bit.ly/IBO2020file**

Or refer to the URL list at the end of this exam file/booklet.

Then, enter the server using the following username and password

Username: **ibo2020**
Password: **ibo2020binagasaki**

If you have a connection problem, try the following alternative servers;
Competitors in Asia, Oceania, or America: 18.181.30.53/ec2-user/ibo2020bi
Competitors in Europe: 18.184.254.217:3838/ec2-user/ibo2020bi/

IBOC applications should not be accessed except during exam hours. The above URL should not be shared with others even after taking the test.

If you experience any problems using the IBOC application, reload the application using the reload button on of your Web Browser.

The IBOC application consists of 14 tabs. After starting the test, check first to see if you can open all tabs. Immediately after the application is loaded for the first time, it may take about 1 minute for a tab to be loaded.

This exam consists of three parts: Part 1: 13 marks, Part 2: 66 marks, and Part 3: 21 marks. The total score is **100 marks**. You will have **90 minutes** to answer all questions. These questions are designed to be solved in order, from beginning to end.

Overview of the topics covered in the bioinformatics questions

Bioinformatics is a discipline that aims to elucidate how information is transmitted and interpreted within living organisms. Research scientists use bioinformatics tools to convert observations of natural phenomena into digital format, and subsequently to visualize, analyze, and interpret this information using computers.

Digitization has benefited scientific progress in ways that are well beyond simply making the data easier to handle. First, the development of more powerful computers enabled the simultaneous analysis of amounts of data so large that the task would be impossible to perform manually. In addition, online sharing of the data has made it possible for researchers around the world to share their observations on various life phenomena more efficiently. In particular, research on genomic DNA sequences and on the three-dimensional structure of proteins benefited very early on from the advent of computing technology. Nowadays, the benefits of computers in the life sciences are widespread.

Use the IBOC application to address the following questions. IBOC application may be accessed using any Web Browser.
Note: These application tools may return an error string written in red letters like the following: " Error: An error has occurred. Check your logs or contact the app author for clarification". You will see this if the query data is formatted incorrectly.

Part 1: Genome databases

Read the following information about amino acid and nucleotide sequence databases, and answer the questions below.

Genbank is known as one of the first DNA sequence databases and is maintained by NCBI (National Center for Biotechnology Information, USA). In Genbank, information for each gene is displayed in a format called the GenBank Format, as shown in ***Genes 1*** tab. Read the following section and answer the questions.

The ***Genes 1*** tab includes information about the human *HoxA5* gene as registered in the RefSeq database with the accession number NM_019102, and displayed in GenBank Format. In GenBank Format, each gene entry line begins with a **LOCUS** tag, and ends with two slashes (//). Most of the information is described as a combination of tags and data corresponding to the tags. A tag is a word shown at the beginning of a line, such as **LOCUS** or **DEFINITION**. For example, the data corresponding to the **LOCUS** tag is "NM_019102 1670 bp mRNA linear PRI 31-DEC-2019".

SOURCE tag shows the species name (Homo sapiens (human)). **FEATURES** tag corresponds to several sub-sections; **source**, **gene**, **exon**, and **CDS**. "/chromosome=7" in the source sub-section

shows the chromosome number of which the gene of the entry is located. "/gene=HOXA5 in the gene sub-section shows the name of the gene of the entry. The **CDS** sub-section indicates that the region for amino acids starts at the 75th base and ends at the 887th base. The "/translation=" part of the **CDS** sub-section contains the amino acid sequence translated from the coding sequence (CDS) of this gene. There are two **exon** sub-sections in the **FEATURES** tag, described as "exon 1..636" and "exon 637..1670". This means that the gene has two exons, and that the boundary between the two exons is located in between the 636th and 637th bases, for a final transcript length of 1670 bp. Finally, The **ORIGIN** tag contains the DNA sequence of the genomic region that is transcribed as messenger (mRNA), written from the 5'- to the 3'-end.

Question 1. According to the amino acid sequence shown in the /translation= part of the above GenBank entry, the protein encoded by this gene starts with an M (methionine) at the first position, followed by an S (serine) at the second position, and an another S (serine) at the third position. Assuming the nucleotide sequence is exactly the same as the data shown in the above GenBank entry, select the correct nucleotide sequences coding for the second and third position serines. [3 marks] **[No. 1]**

	2nd position serine,	3rd position serine
1.	AAT,	TGT
2.	AAT,	AAT
3.	GAC,	GCA
4.	GAC,	GAC
5.	TCT,	TCT
6.	AGC,	AGC
7.	AGC,	TCT
8.	GCC,	CGG

In the GenBank entry, not only the amino acid sequence encoded by the target mRNA can be found in /translation=, this information is linked to another database (GenPept) as /protein_id="NP_061975.2".
NP_061975.2 is the accession ID of the GenPept database, also operated by NCBI. The GenPept data for NP_061975.2 is shown ***Proteins*** tab.

Proteins consist of several regions with different characteristics functions (protein domains), which confer both structure and function. In the above entry, the protein encoded by the *HOXA5* gene has a total length of 270 amino acids, and the Region sub-section indicates that amino acids 176 to 181 constitute a protein domain called the "Antp-type hexapeptide". Another protein domain called "Homeobox domain" spans amino acids 199 to 251. Similarly, you can find an entry for the *HOXA6* gene in the ***Proteins*** tab.

One of the tools for discovering functional domains in amino acid sequences is hmmscan in Hmmer3 (Mistry *et al.*, 2013). It enables the discovery of common domains (eg.

“Homeodomain”) in a given family of genes. In the ***HMMSCAN*** tab of the IBOC application, you can use hmmscan to examine the functional domains contained in the amino acid sequence of your protein of interest. This requires an existing protein domain database, and IBOC applications are designed to search the Pfam-A database. Let's check the position of the Homeobox domain of *HOXA5* examined in Question 1 using hmmscan.

In the ***Protein Sequences 1*** tab, several amino acid sequences are shown in **FASTA format** which is one of the most commonly used formats to describe nucleotide and amino acid sequences. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning.

From the ***Protein sequences 1*** tab, copy the amino acid sequence of HOXA5 protein in FASTA format and paste it in the "Input sequence" box of the ***HMMSCAN*** tab. An E-value is calculated to estimate the probability of the result happening by chance, and the recommended threshold is 1e-5 (=0.00001). Clicking on "Exec hmmscan" will display the regions containing functional domains for HOXA5 (if present). If done correctly, the following should be displayed (Figure 1.):

target name:	target accession:	tlen	query name:	qlen:	E-value:	score:	#	of	from (hmm coord):	to (hmm coord):	from (ali coord):	to (ali coord):	description of target:
Homeodomain	PF00046.30	57	NP_061975.2homeoboxproteinHox-A5[Homosapiens]	270	4.5e-22	77.7	1	1	1	57	196	252	Homeodomain

Showing 1 to 1 of 1 entries Previous 1 Next

Figure 1. An example of Hmmscan search result. This figure shows the output of the hmmscan.

The search results display the domain name and Accession ID in the "target name:" and "target accession:" columns, respectively. This information depends on the database provided, and in the case of the IBOC application, the domain name and Accession ID from the Pfam-A database are displayed. "tlen" is the total length of that particular domain in amino acids. Here, the homeodomain registered in Pfam-A is 57 amino acids.

The name of the query amino acid sequence (the sequence you copied in, eg. NP_061975.2 homeobox protein Hox-A5 [Homo sapiens] in Figure 1.)" is displayed in the "query name:" column, and its length in the "qlen:" column. If the same domain is found more than once, the total number and serial number are displayed in the "of:" and "#:" columns, respectively. The "from (hmm coord):" column and the "to (hmm coord):" column indicate which part of the domain matched to the database. The "from (ali coord):" and "to (ali coord):" columns show the start and end positions of the domain of interest inside the query sequence. This time, the results show amino acids 1 to 57 of the PF00046 Homeodomain, which is exactly 57 amino acids long, meaning that the entire length of this domain was included in the query sequence.

Question 2. Find a Homeodomain (PF00046.30) in HOXA6 (NP_076919.1 homeobox protein Hox-A6 [Homo sapiens]). Use E-value: 1e-5 as threshold. [4 marks]

The homeodomain (PF00046.30) region in *HOXA6* gene starts ("from (ali coord):") at **[No.2][No.3][No.4]**, and ends ("to (ali coord):") at **[No.5][No.6][No.7]**.

eg) If you would like to answer "starts at **25**" and "ends at **143**", the example of how to answer is as follows;

No. 2 : 0
No. 3 : 2
No. 4 : 5

No. 5 : 1
No. 6 : 4
No. 7 : 3

Next, let's find where the two exons of *HOXA5* are encoded in the DNA genome. It is convenient to use the NCBI-BLAST+ tool developed by NCBI to search for sequence similarity. To do this, you can choose between the following three programs. When **blastp** is selected, an amino acid sequence query is searched for in an amino acid sequence database. If **blastn** is selected, a nucleotide sequence query is searched against a nucleotide sequence database. Finally, when **tblastn** is selected, an amino acid query may be searched against a nucleotide sequence database.

The first three tabs of the IBOC application show the following nucleotide sequences.
The ***"Human Genome DNA 1"*** tab shows the nucleotide sequence of human chromosome 7 from positions 27,140,701 to 27,150,700.
The ***"Human Genome DNA 2"*** tab shows the nucleotide sequence of human chromosome 7 from positions 26,188,001 to 26,218,000.
The ***"B. burgdorferi B31 Genome DNA 1"*** tab is the sequence of the entire cyclic genomic DNA of a bacterium (*Borrelia burgdorferi* strain B31).

In the following questions we will refer to base positions within the analyzed sequence, not positions along the chromosome. For example, we will refer to the first base of the Human Genome DNA 1 sequence as 1, not 27,140,701. The genomic DNA of *B. burgdorferi* strain B31 is circular. Therefore, the first base was arbitrarily defined.

After copying the mRNA sequence of *HOXA5* (NM_019102.4) from the ***Predicted mRNA Sequences*** tab, paste it into the "Input sequence:" window of the ***BLAST*** tab. (Note: By convention, RNA sequences are shown as a complementary DNA sequence. This means a coding DNA strand and the transcribed RNA would be shown as identical sequences). Next, by selecting "Human Genome DNA 1" from the "Reference Sequence:" options, a similarity search is

performed against “Human Genome DNA 1”. This step requires **blastn**. Clicking "Exec" gives the predicted locus of *HOXA5*. In the search results, "sseqid" displays the IDs of the hit sequence.

Question 3. Choose the location which is including the first exon of HOXA5 from the following options. Use “E-value:1e-5” as threshold for blastn. **Hint**: Use ***Predicted mRNA sequences*** tab, and ***BLAST*** tab. [2 marks] **[No.8]**

1. 1 bp ~ 1,000 bp
2. 1001 bp ~ 2000 bp
3. 2001 bp ~ 3000 bp
4. 3001 bp ~ 4000 bp
5. 4001 bp ~ 5000 bp
6. 5001 bp ~ 6000 bp
7. 6001 bp ~ 7000 bp
8. 7001 bp ~ 8000 bp
9. 8001 bp ~ 9000 bp
0. 9001 bp ~ 10,000 bp

Question 4. Next, find proteins with a similar amino acid sequence to HOXA6. Using HOXA6 as a query, perform a blastp search on a human protein dataset (“human proteins” in the Reference sequence field). Use “E-value:1e-5” as threshold for **blastp**. Choose the entry with the highest sequence similarity from the following options, but not HOXA6 itself. Use “E-value:1e-5” as threshold for blastp. **Hint**: Use ***Protein sequences 1*** tab, and ***BLAST*** tab. [4 marks] **[No.9]**

1. 1_06639
2. 1_05172
3. 1_07981
4. 1_09188
5. 1_12205
6. 1_12968
7. 1_15255
8. 1_15496
9. 1_17260
0. 1_18575

In order to obtain the protein sequence, you should go to the ***Entry Database*** tab within the IBOC application (you would need in questions 18 and 19). The amino acid sequence of the protein will be displayed upon entering the protein’s ID (eg. 1_05121) in the **Entry_ID** input field and choose an organism name from the **Sequence** selection. **Note: This feature may be useful for answering Question 18 and 19.**

Part 2. Analysis of sequence features and motif discovery

The question below introduces methods to investigate the GC composition of nucleic acid sequences as a general property of genomic DNA. Read the following information about chromosomes and the GC composition of DNA, and answer the questions below.

As their name suggests, chromosomes can be dyed using one of several staining methods. Among these methods, Giemsa staining results in a striped pattern of bands along the chromosome. The staining is usually performed during the **[No. 10]** of cell division, when the chromosomes are visible because of chromosome condensation. The coloration is darker in regions that have a **[No. 11]** AT content, and lighter and brighter in regions that have a **[No. 12]** GC content. Furthermore, regions with a **[No. 12]** GC content were thought to have more genes. Such a correlation between GC content and gene density was confirmed when the nearly complete sequence of human nuclear DNA (draft genome) was determined by the Human Genome Project in 2000. In addition, DNA tend to undergo denaturation (separation of the double helix into two single strands) at **[No. 13]** temperatures. GC pairs are denatured at a **[No. 14]** temperature than AT pairs. As described above, the GC content is a value that is both easy to obtain and very useful in various aspects of bioscience and biotechnology research.

Question 5. Pick the answer most appropriate for **[No. 10]**. [4 marks]

1. G0 phase
2. S phase
3. Metaphase

Question 6. Pick the answer most appropriate for **[No. 11], [No. 12], [No. 13],** and **[No. 14]**. [8 marks]

1. higher
2. lower

From here, let us focus on the GC content of DNA sequences. The GC content (GC%) is represented by the following formula.

$$GC\% = 100 \times (([G] + [C]) / ([A] + [T] + [G] + [C]))$$

In this formula, [A] represents the number of adenosines (A), and similarly, [T], [G], and [C] each stand for the number of thymines (T), guanosines (G), and cytosines (C), respectively.

Question 7. For the DNA sequence shown in ***"Human Genome DNA 1"*** tab, calculate the GC content in the ten bases from the 1st to the 10th position included. [2 marks]

[No. 15][No. 16][No. 17] %

eg) If you would like to answer "32%", the example of how to answer is as following,

No. 15 : 0
No. 16 : 3
No. 17 : 2

The usage of the ***"Count nucleotide"*** tab and the ***"Window search"*** tab is shown below.

DNA sequences from the region of your choice can be extracted from "Human Genome DNA 1", "Human Genome DNA 2", and "B. burgdorferi B31 Genome DNA 1" by using the **Get Subsequence** function on the left side of the ***"Count Nucleotide" tab***.
Select **Human Genome DNA 1** from the **Target Sequence** selection list, then enter **1** and **10** in the "**Start**" and "**End**" fields, respectively. This will display the first 10 bases from the first to the tenth base of "Human Genome DNA 1". **Note: This feature may be useful for answering Question 15 and 16.**

Window search is a method used to examine the distribution of DNA features over the entire nucleotide sequence by shifting a window of fixed length by a fixed unit. The length of the window is called Window Size. The length of the window shift is called Step Size.

Next, let's examine the sequence characteristics of "Human Genome DNA 1" and "Human Genome DNA 2" using the "***Window Search***" tab of the IBOC application.

Question 8. Let's check the GC content for the first 10,000 nucleotides of "Human Genome DNA 1" using window search.

1. Open the ***"Window Search"*** tab and select "Human Genome DNA 1" from the **Reference Sequence** list.
2. To check the total number of G's and C's combined, select "G+C" from the **Nuc.** selection list.
3. Enter 1 in the **Start** field and enter 10,000 in the **End** field.
4. Set the "Window Size" to 100 bp, the "Step Size" to 100 bp, and the "Bin Size" to 10 %.
5. With the above settings, the GC content will be calculated for every 100 bp window from the 1st base to the 10,000th base of "Human Genome DNA 1".
6. Finally, clicking on "**Show Chart 1**", displays the result as a histogram representing the frequency of GC content values in intervals of 10 % (corresponding to the histogram's **Bin Size**) in "**Histogram for the selected nucleotide(s)**".
7. Additionally, clicking on "**Show Chart 2**", displays the result as a plot representing the frequency of GC content values along the region above in "**Frequency for the selected nucleotide(s)**".

Note: X-axis in Chart-2: nucleotide position (bp). Y-axis in Chart-2: frequency for the selected nucleotide(s).
Note: In Chart 2, Chart 3 and Chart4, the x-axis position may be displayed as “1e-4” instead of “10,000”.
Note: If you have changed the parameters, click on "Show Chart 1(or 2, 3, 4)" again to reflect the change.

Select the most appropriate item for [No. 18] in the following sentence. [4 marks]

Using the above settings, 100 bp windows with a GC content of **[No. 18]** % appear most frequently in "Human Genome DNA 1”.

1. 10-20
2. 20-30
3. 30-40
4. 40-50
5. 50-60
6. 60-70
7. 70-80
8. 80-90

Question 9. As we saw in **Question 7. ~ 8.** the GC% of DNA varies from region to region. Find out where the 200 bp window with the highest GC content is located on "*B. burgdorferi B31* Genome DNA", and choose the corresponding location from the following options. Note that the genome DNA of this organism is circular and 910.724 bp long. Note that this **calculation may take some time** (about one minute per calculation) in IBOC applications. [5 marks] **[No. 19]**

1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

In the following section, we will investigate bacterial DNA replication using the methods for analysis such as **Window Search**, which we learned from the previous question.

Question 10. The replication of circular DNA starts at a single position and propagates in both directions (Figure 2).

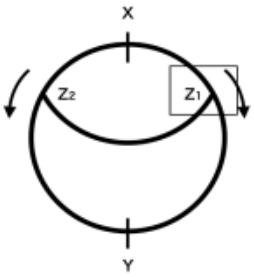


Figure 2. Schematic diagram of DNA replication in a bacterial circular genome. X is the start point of replication (Ori) and Y is the end point of replication. There are two replication forks, shown as Z1 and Z2.

If the whole genome sequence of the bacteria is available, to estimate the start and end points of replication, it is useful to examine the GC-skew. GC-skew is an indication of the bias in the balance between G's and C's on one strand of the DNA. It is expressed as the difference between the number of C's and the number of G's divided by the total number of G's and C's over a given sequence length.

GC-skew = ([C]-[G]) / ([C]+[G])

Using the **Window Search**, look up the GC-skew of the *B. burgdorferi* B31 Genome DNA. In the "***Window Search***" tab, select "*B. burgdorferi* B31 Genome DNA" in the "Reference Sequence", fill the first base and the last base in the **Start** and **End** fields, respectively, and then choose ([C]-[G])/([C]+ [G]) for **Skew** field. Finally, press **Show Chart 3**, the GC-skew plot appears in the Chart 3 area. Note that **this calculation may take some time** (about one minute per calculation) in IBOC applications.

There are two switching points between high GC-skew and low GC-skew in the DNA. Choose the corresponding locations among the following options. [8 marks]

1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

The two switching points are located in **[No. 20]** and in **[No. 21].**

Question 11. It is known that the replication start point, OriC, is in the vicinity of the region encoding the DnaA protein. Examine the region encoding the DnaA protein on the "*B. burgdorferi* B31 Genome DNA" and select the region that contains it.
Think about which tools you need to use to answer this question. You can use the **DnaA** protein sequence from the ***Protein Sequences 2*** tab. [2 marks] **[No. 22]**

1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

Question 12. It is known that GC-skew values change significantly at both the replication start and end points of replication. Considering the answers to the previous questions, select the region that is most likely to contain the replication start point (OriC) for this organism from the following. [4 marks] **[No. 23]**

1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

From here, we will examine the DNA motifs involved in the regulation of gene expression in eukaryotes.
Gene expression can be regulated by the chemical addition of methyl groups to specific bases on DNA (Figure 3).

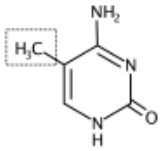


Figure 3. Structural formula of 5-methylcytosine, a methylated cytosine. The dotted area is the added methyl group.

In vertebrates, the cytosines of CpG dinucleotides are often subject to DNA methylation. A CpG dinucleotide, as the name indicates, consists of a consecutive C and G in the 5’ to the 3’ direction (Figure 4). 'p' between C and G, which stands for the phosphodiester bond so that CpG is distinguished from CG which is a nucleotide pair in a double-stranded DNA.

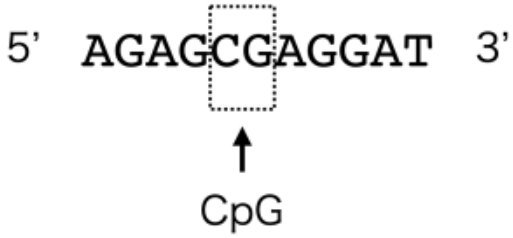


Figure 4. A CpG dinucleotide consists of a consecutive C and G in the 5’ to the 3’ direction. The dotted area is an example of a CpG dinucleotide.

A DNA region that is rich in CpG dinucleotides is known as a CpG island. CpG islands are regularly found in the genomic DNA of vertebrates, and may be located either around the exons of protein-coding genes, or a few hundred bp upstream of their transcription start sites. Not all CpG islands are methylated, but these methylation events are correlated with inactivation of the downstream gene.
Several methods have been proposed to look for CpG islands. Among them, the **CpG-score** is expressed by the following formula, and can be calculated over a shifting window as demonstrated previously

CpG-score = ([CpG] / ([C] x [G])) x Window-Size

In this formula, [CpG] represents the number of CpG in the window, and similarly, [G], and [C] each stand for the number of guanosines (G), and cytosines (C) in the window, respectively.
To obtain the CpG score, "Window-Size = 100" and "Step-Size = 1" are generally the standard parameters used. Subsequently, areas with a CpG score of 0.6 or higher are often considered as candidate CpG islands.

Question 13. As mentioned above, "Human Genome DNA 1" is the sequence from positions 27,140,701 to 27,150,700 on human chromosome 7, for a total of 10,000 bases.
In the "***Window Search***" tab, select "Human Genome DNA 1" from the **Reference Sequence** field, and examine the distribution of the CpG-score. Enter the appropriate values for the **Start** and **End** fields. Use "Window-Size = 800" and "Step-Size = 1" as the parameters.
Select "**CpG-score**" from the **CpG-score** field shown in the lower right corner of the screen, and click **Show Chart 4**. This will display the distribution of CpG-scores for the target sequence region. Estimate the length of the longest candidate CpG island region found in this sequence, and choose the closest value from the following options. In this question, regions with a CpG score greater than 0.6 are considered to be candidate CpG islands. Note that this **calculation may take some time** (about one minute per calculation) in IBOC applications. [6 marks] **[No. 24]**

1. 200 bp
2. 800 bp
3. 1,400 bp
4. 2,000 bp
5. 2,600 bp
6. 3,200 bp
7. 3,800 bp
8. 4,400 bp
9. 5,000 bp

Question 14. Then, predict the transcription start site of human ***HoxA6*** gene in the Human genome DNA 1. Consider the most 5'-end-most position among the search results (also called "hits") of this full-length sequence as the transcription start site, and also consider the most 3'-end position among the hits of this full-length sequence as the end of transcript. Choose the closest answer from the following options. For the *HoxA6* sequence, use the sequence shown in the "***Predicted mRNA sequences***" tab.

The transcription start site of *HoxA6* is nearest to nucleotide position **[No. 25]**. [4 marks]

1. 1,000
2. 2,000
3. 3,000
4. 4,000
5. 5,000

6. 6,000
7. 7,000
8. 8,000
9. 9,000

Question 15. Referring to **Questions** 13 - 14 above, choose the **incorrect** statement from the following options. [6 marks] **[No. 26]**

1. The CpG island, which overlaps with the first exon of *HoxA5*, is longer than another CpG island which overlaps with the first exon of *HoxA6*.
2. The intron of *HoxA6* is roughly consistent with the region containing the fewest CpG dinucleotides.
3. In this sequence, the region with the lowest CpG-score is roughly consistent with the region with the highest AT content.
4. Both *HoxA5* and *HoxA6* have a GC content of >60% in the first exon.
5. The intergenic region between *HoxA5* and *HoxA6* genes is approximately 1.7 kb in length.

In the following question, we will examine the characteristics of functional sequence motifs found on genomic DNA.

DNA contains a variety of functional sequence motifs. A famous example is the polyadenylation signal (AAUAAA sequence) encoded by the transcribed RNA in its 3' untranslated region (UTR). On mRNA precursors, a group of protein complexes recognizes this polyadenylation signal and truncates the 3'-end of RNA before adding a poly(A) tail. In other words, we can expect to find a polyadenylation signal in the last exon at the 3'-end of a coding sequence.

If the location of the motif is known, an analysis method called "sequence logo" enables statistical evaluation of the consensus sequence. As an example, let's use the sequence logo program to visualize the polyadenylation signal motif. The 3'-end of genes are shown in FASTA format below. The sequence of the polyadenylation motif is shown in capitals. Ten bases either side are shown in lowercase. (Same sequences are shown on the bottom of the ***Sequence Logo*** tab).

```
>HoxA5_1638bp_1663bp
tggaacaaaaAATAAActttctattg
>CBX3_2027bp_2052bp
acagttgggaAATAAAagtttcatgt
>HNRNPA2B1_3591bp_3617bp
ggctgtccccAATAAAtgctgttcat
>Stk31_3238bp_3263bp
ggttgtgaaaAATAAAgatgtttggc
>Tra2a_1752bp_1777bp
agtagtctcaAATAAAagctaatttc
```

The figure below is a representation of the sequence alignment as a sequence logo (Figure 5).

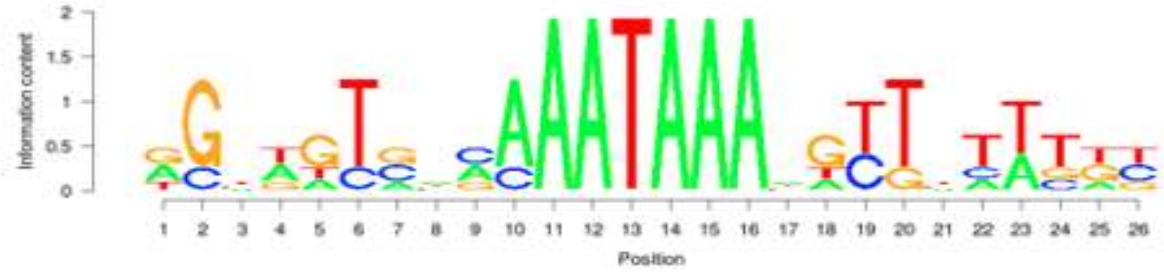


Figure 5. An example of a result of the sequence logo.

Using the ***Sequence Logo*** tab, make sure that you actually get the same diagram as above. (Copy the above sequences in the ***Sequence Logo*** tab and paste those into the "Input Sequence" window in the Sequence Logo tab, and then click on the “Exec Sequence Logo”).

Thus, the AATAAA sequence, a common polyadenylation-signal motif in all five sequences, is highlighted on the sequence log as you can see. In this example, only five gene sequences were used, but in the actual discovery of motifs, many more sequences need to be analyzed. Otherwise, you run the risk of highlighting sequences that just happen to be coincidental matches.

Question 16. In eukaryotes, protein-coding genes consist of several exonic and intronic sequences on a genomic DNA. It is known that the boundary between exons and introns contains a distinctive sequence that allows the spliceosome (the enzyme that cuts out the intron) to recognize the intron sequences to be spliced. In addition to their commonality within eukaryotes, they also have species-specific and taxonomic specificity. The *HNRNPA2B1* gene has a typical exon-intron boundary sequence motif in the human genome. Determine the correct motif of exon-intron boundary and intron-exon boundary. You can use the information of the *HNRNPA2B1* gene in the GenBank format in the ***Genes 1*** tab, mRNA sequence in the ***Predicted mRNA sequences*** tab. Then choose the most appropriate answer from the following options. [13 marks]

Note that these represent the 4 bp motif (2 bp at the 3'-end of the exon and 2 bp at the 5'-end of the intron) that make up the main exon-to-intron boundary in human ([No. 27 – No. 30]), and the 3 bp motif (2 bp at the 3'-end of intron and 1 bp at the 5'-end of exon) that makes up the major intron-to-exon boundary in human ([No. 31 – No. 33]).

[Exon-side bases] - [Intron-side bases]
[[No. 27], [No. 28]] - [[No. 29], [No.30]]

[Intron-side bases] - [Exon-side base]
[[No. 31], [No. 32]] - [[No. 33]]

1. a
2. t
3. g
4. c

eg) If you would like to answer " **…exon… at** - gg …intron... ca - **t … exon …**", the answer should be as follows:

No. 27 : 1
No. 28 : 2
No. 29 : 3
No. 30 : 3
No. 31 : 4
No. 32 : 1
No. 33 : 2

Part 3: Task

Answer the following questions using the IBOC application.

Using the ***BLAST*** tab and the ***Entry Database*** tab, a sequence-similarity search can be performed against the protein data set of the following 7 organisms; 5 animals (human, mouse, fruit fly, octopus, and starlet sea anemone) and 2 unicellular organisms (choanoflagellate. and fission yeast). By examining some of the cadherin family genes in these organisms, answer the following questions.

Question 17. *E-Cadherin* and *N-Cadherin* are known to be representative of the cadherin family of genes. Information on these two genes are shown in the ***Genes*** tab, ***Proteins*** tab, ***Predicted mRNA sequences*** tab, and ***Protein sequences 1*** tab. Many of the cadherin family genes in the above animals have a tandemly duplicated (=repeatedly lined up) domain. Both E-Cadherin and N-Cadherin repeat this domain at least five times or more. The domain is given by the Accession ID of Pfam-A: (i)[PF *****]. Identify the domain, use the e-value of less than 1e-5 for the hmmscan threshold. (Note: Ignore the numbers after the dot in Pfam Accession ID.) [4 marks]

(i) PF **[No. 34][No. 35][No. 36][No. 37][No. 38]**

eg) If you would like to answer "PF00001.12", the answer should be as follows:

No. 34 : 0
No. 35 : 0
No. 36 : 0

No. 37 : 0
No. 38 : 1

Question 18. A comparison of the domain structure of cadherin proteins in animals and non-animals shows that, (ii)[PF ***** . *] is added to the (iii) **[No.44]**-terminus next to the repeat of the (i) domain in animals. (Note: Ignore the numbers after the dot in Pfam Accession ID.)
[10 marks]

(ii) PF **[No. 39][No. 40][No. 41][No. 42][No. 43]**

(iii) **[No.44]**

1. N
2. C

Question 19. From the set of protein entries for the choanoflagellate, find one gene that has multiple repeats of the above domain (i) in Question 17. To identify the domain (i), use the e-value of less than 1e-5 for the hmmscan threshold. [7 marks]
(The more repeats of domain (i) that are in the gene that you find, the higher score you will get.)

Hint: You may use following functions in ***HMMSCAN-tab*** as needed." '**Show [number] entries**' function can be used to change the number of displayed lines. '**Search: [word]**' can be used to perform a string [word] search. The total number of rows in the resulting table is displayed at the bottom left of the screen.

(iv) **[No.45] _ [No. 46][No. 47][No. 48][No. 49][No. 50]**

// End of Practical Exam 2.

URL list

#	Participants	ID	URL for Practical Exam 2
1	Iran	11	13.127.129.209/ec2-user/ibo2020bi
2	Hungary	12	3.122.239.215/ec2-user/ibo2020bi
3	Japan	13	54.250.182.124/ec2-user/ibo2020bi
4	Armenia	15	13.127.129.209/ec2-user/ibo2020bi
5	Russia	16	3.122.239.215/ec2-user/ibo2020bi
6	Kazakhstan	17	13.127.129.209/ec2-user/ibo2020bi
7	Philippines	18	52.79.248.78/ec2-user/ibo2020bi
8	Indonesia	19	54.255.220.196/ec2-user/ibo2020bi
9	South Korea	20	52.79.248.78/ec2-user/ibo2020bi
10	Nepal	21	13.235.100.64/ec2-user/ibo2020bi
11	Sri Lanka	23	13.235.100.64/ec2-user/ibo2020bi
12	Bangladesh	24	13.234.37.5/ec2-user/ibo2020bi
13	Pakistan	25	13.234.37.5/ec2-user/ibo2020bi
14	Thailand	26	52.79.146.192/ec2-user/ibo2020bi
15	Vietnam	27	54.255.220.196/ec2-user/ibo2020bi
16	Singapore	29	54.255.232.154/ec2-user/ibo2020bi
17	China	30	52.79.146.192/ec2-user/ibo2020bi
18	Chinese Taipei	31	52.79.248.78/ec2-user/ibo2020bi
19	Hong Kong, China	32	54.255.232.154/ec2-user/ibo2020bi
20	Syria	34	13.127.123.61/ec2-user/ibo2020bi
21	Saudi Arabia	36	13.127.123.61/ec2-user/ibo2020bi
22	Finland	44	35.156.199.58/ec2-user/ibo2020bi
23	Denmark	47	35.156.199.58/ec2-user/ibo2020bi
24	Iceland	48	18.184.71.168/ec2-user/ibo2020bi
25	Estonia	49	18.184.71.168/ec2-user/ibo2020bi
26	Latvia	50	3.124.195.33/ec2-user/ibo2020bi
27	Lithuania	51	3.124.195.33/ec2-user/ibo2020bi
28	Kyrgyzstan	53	15.236.134.99/ec2-user/ibo2020bi
29	Tajikistan	54	15.236.134.99/ec2-user/ibo2020bi
30	Uzbekistan	56	15.188.75.25/ec2-user/ibo2020bi
31	Azerbaijan	59	15.188.75.25/ec2-user/ibo2020bi
32	Georgia	60	3.124.195.33/ec2-user/ibo2020bi

33	Czech Republic	61	3.127.214.51/ec2-user/ibo2020bi
34	Poland	63	3.127.214.51/ec2-user/ibo2020bi
35	Bulgaria	64	35.180.21.57/ec2-user/ibo2020bi
36	Slovenia	65	35.180.21.57/ec2-user/ibo2020bi
37	North Macedonia	69	15.236.239.75/ec2-user/ibo2020bi
38	Turkey	72	15.236.239.75/ec2-user/ibo2020bi
39	Netherlands	73	15.236.131.4/ec2-user/ibo2020bi
40	Belgium	74	15.236.131.4/ec2-user/ibo2020bi
41	Germany	75	18.185.106.176/ec2-user/ibo2020bi
42	Switzerland	77	18.185.106.176/ec2-user/ibo2020bi
43	Luxembourg	78	3.126.249.168/ec2-user/ibo2020bi
44	United Kingdom	82	35.178.172.104/ec2-user/ibo2020bi
45	United States of America	83	3.128.202.209/ec2-user/ibo2020bi
45	Australia	84	3.104.47.102/ec2-user/ibo2020bi
46	Turkmenistan	89	3.126.249.168/ec2-user/ibo2020bi
47	Croatia	66	35.180.21.57/ec2-user/ibo2020bi
48	Canada	81	3.128.202.209/ec2-user/ibo2020bi
49	France	93	15.236.131.4/ec2-user/ibo2020bi
50	Afghanistan	95	15.236.239.75/ec2-user/ibo2020bi
51	Norway	45	35.178.172.104/ec2-user/ibo2020bi
52	El Salvador / Ibero-America	92	3.128.202.209/ec2-user/ibo2020bi