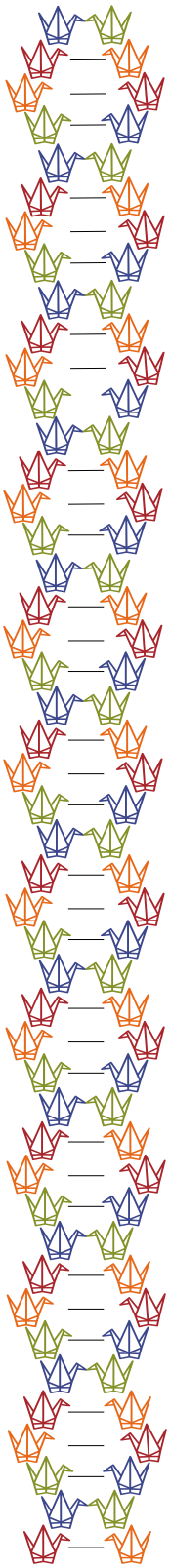


Practical Exam 2

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

اطلاعات زیر را راجع به پایگاه های داده توالی های آمینواسیدی و نوکلئوتیدی بخوانید و به سوالات پیش رو پاسخ دهید.

Genbank یکی از اولین پایگاه های داده توالی DNA بوده و توسط NCBI مدیریت و پشتیبانی می شود. همان طور که در زبانه ***Genes 1*** نشان داده شده است، در Genbank اطلاعات هر ژن در قالبی که GenBank Format نامیده می شود نمایش داده می شود. بخش زیر را بخوانید و به سوالات پاسخ دهید.

زبانه ***Genes 1*** حاوی اطلاعاتی درباره ژن *HoxA5* انسانی که با شماره تخصیص NM_019102 در پایگاه داده RefSeq ثبت شده است می باشد که در قالب GenBank Format نمایش داده شده است. در این قالب اطلاعات هر ژن با یک تگ **LOCUS** آغاز شده و با دو اسلش (//) پایان می یابد. بیشتر اطلاعات به صورت تگ ها و اطلاعات مربوط به هر تگ توصیف شده اند.

یک تگ کلمه ایست که در ابتدای یک خط نمایش داده شده است، مانند **LOCUS** یا **DEFINITION**. برای مثال اطلاعات مربوط به تگ **LOCUS** برابر با "NM_019102 1670 bp mRNA linear PRI 31-DEC-2019" می باشد.

تگ **SOURCE** نام گونه را نمایش می دهد. تگ **FEATURES** حاوی زیر بخش های **source**، **gene**، **exon** و **CDS** می باشد. "/chromosome=7" در زیر بخش **source** نشان دهنده شماره کروموزومی است که ژن مورد نظر در آن قرار دارد. "/gene=HOXA5" در زیر بخش **gene** نشان دهنده نام ژن مورد نظر است. زیر بخش **CDS** مشخص می کند که آمینو اسید ها از 75 امین باز آغاز شده و در 887 امین باز پایان می یابند. قسمت "/translation=" در زیر بخش **CDC** حاوی توالی آمینو اسیدی ترجمه شده از توالی کد کننده (CDC) این ژن می باشد.

دو زیر بخش **exon** ("exon 1..636" و"exon 637..1670") در تگ **FEATURES** وجود دارند. این بدین معنیست که ژن مورد نظر دو اگزون دارد که طول نهایی ترجمه شده 1670 باز را حاصل می کنند و مرز میان دو اگزون میان باز 636 ام و 637 ام قرار دارد. تگ **ORIGIN** حاوی توالی DNA ناحیه ژنومی ای است که به mRNA رونویسی می شود (توالی از سمت ’5 به سمت ’3 نوشته شده است).

سوال 1. با توجه به توالی آمینو اسیدی نمایش داده شده در قسمت =translation/ ژن بالا، پروتئین کد شده توسط این ژن با یک M (متیونین) آغاز شده و با یک S (سرین) در جایگاه دوم و یک S (سرین) دیگر در جایگاه سوم ادامه می یابد.

با فرض اینکه توالی نوکلئوتیدی این ژن دقیقا همان چیزی است که در اطلاعات GenBank این ژن نمایش داده شده است، توالی نوکلئوتیدی کد کننده برای سرین های دومین و سومین جایگاه را انتخاب کنید. **[No. 1] 3 نمره**

	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.

پروتئین ها از چندین بخش با خصوصیات عملکردی متفاوت (protein domains) تشکیل شده اند. در اطلاعات بالا، پروتئین کد شده توسط ژن *HOXA5* طولی معادل 270 آمینواسید داشته و زیر بخش Region مشخص می کند که آمینواسید های 176 تا 181 یک دامین پروتئینی به نام "Antp-type hexapeptide" تشکیل می دهند.

دامین پروتئینی دیگری به نام "Homeobox domain" در طول آمینواسید های 199 تا 251 گسترده شده است.

مشابها، شما می توانید اطلاعات ژن *HOXA6* را در زبانه ***Proteins*** پیدا کنید.

یکی از ابزار های تشخیص دمین های عملکردی در توالی های آمینواسیدی، hmmscan در Hmmer3 است، که کشف دمین های متداول (برای مثال "Homeodomain") را در یک خانواده ژنی ممکن می سازد. در زبانه ***HMMSCAN*** شما می توانید از hmmscan برای بررسی کردن دمین های عملکردی موجود در توالی آمینو اسیدی پروتئین مورد نظرتان استفاده کنید. این کار نیاز به موجود بودن یک پایگاه داده ی دمین های پروتئینی دارد و اپلیکیشن IBOC برای جستجو کرن در پایگاه داده ی Pfam-A طراحی شده است. بگذارید با استفاده از hmmscan جایگاه دمین Homeobox را در HOXA5 که در سوال ۱ مورد تحلیل قرار گرفت، بررسی کنیم.

در زبانه ***Protein Sequences 1***، چندین توالی آمینواسیدی با فرمت **FASTA** نشان داده شده اند که یکی از رایج ترین فرمت ها برای شرح دادن توالی های نوکلئوتیدی و آمینواسیدی است. یک توالی در فرمت FASTA با یک خط توضیحات شروع می شود ودر ادامه چندین خط اطلاعات توالی وجود دارد. خط توضیحات (defline) توسط یک علامت بزرگ تر ("<") از اطلاعات توالی تمایز داده می شود.

از زبانه ی ***Protein sequences 1***، توالی آمینواسیدی پروتئین HOXA5 را در فرمت FASTA کپی کنید و آن را در بخش "Input sequence" زبانه ی ***HMMSCAN*** پیست کنید. E-value برای تخمین احتمال این که نتایج با شانس به وجود آمده باشند، محاسبه می شود و آستانه ی پیشنهادی آن 1e-5 (=0.00001) است.با کلیک کردن روی "Exec hmmscan" ناحیه هایی که دمین های عملکردی HOXA5 را شامل می شوند نمایان می شود. (اگر وجود داشته باشند) اگر این کار را درست انجام دهید، مانند شکل ۱ برای شما نمایش داده می شود:

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

شکل ۱. مثالی از نتایج جستجوی Hmmscan. این شکل خروجی hmmscan را نشان می دهد.

نتایج جستجو اسم دمین و Accession ID را به ترتیب در ستون های "target name:" و "target accession:" نمایش می دهد. این اطلاعات به پایگاه داده موجود بستگی دارد و برای اپلیکیشن IBOC اسم دمین و Accession ID از پایگاه داده Pfam-A نمایش داده می شود. "tlen"طول کامل یک دمین خاص را برحسب آمینواسید ها مشخص می کند.

در اینجا، homeodomain موجود در Pfam-A ، 57 آمینواسید دارد.

اسم توالی آمینواسیدی query (توالی ای که شما کپی کردید، برای مثال NP_061975.2 homeobox protein Hox-A5 [Homo sapiens] در شکل ۱) در ستون "query name:" نشان داده شده است و طول آن در ستون "qlen:" است. اگر یک دمین بیشتر از یکبار پیدا شود، تعداد کل و serial number به ترتیب در ستون های "of:" و "#:" هستند.

ستون "from (hmm coord):" و "to (hmm coord):" مشخص می کند کدام بخش از دمین با پایگاه داده تطبیق یافته است. ستون های "from (ali coord):" و "to (ali coord):" جایگاه شروع و پایان دمین مورد نظر را داخل توالی query نشان می دهد. نتیجه آمینواسید ۱ تا ۵۷ PF00046 Homeodomain را نشان می دهد، که دقیقا 57 آمینواسید طول دارد، به این معنی که تمام طول دمین در توالی query یافت می شود.

سوال ۲. یک Homeodomain (PF00046.30) را در HOXA6 (NP_076919.1 homeobox protein Hox-A6 [Homo sapiens]) پیدا کنید. از E-value : 1e-5 به عنوان آستانه استفاده کنید. **4 نمره**

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]**.

برای مثال) اگر می خواهید پاسخ دهید که از 25 شروع می شود و به ۱۴۳ ختم می شود، باید این گونه پاسخ را وارد کنید:

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

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

حال ، بیاید جایگاهی که دو اگزون *HOXA5* را رمزگذاری می کنند در DNA ژنوم پیدا کنیم. مناسب است که از ابزار NCBI-BLAST+ که توسط NCBI برای جستجوی شباهت توالی ها طراحی شده استفاده کنیم. برای این کار، شما می توانید بین سه برنامه ی زیر انتخاب کنید. وقتی **blastp** را انتخاب می کنیم که توالی آمینواسید **query** را بین پایگاه داده ی توالی آمینواسیدی جستجو می کنیم. اگر **blastn** را انتخاب کنیم، یک توالی نوکلئوتیدی **query** را بین پایگاه داده ی توالی های نوکلئوتیدی جستجو می کنیم. در نهایت، وقتی از **tblastn** استفاده می کنیم که توالی آمینواسیدی **query** را در پایگاه داده ی نوکلئوتیدی جستجو می کنیم.

سه زبانه ی نخست اپلیکیشن IBOC توالی های نوکلئوتیدی ذیل را نشان می دهد.

زبانه ی ***"Human Genome DNA 1"*** توالی نوکلئوتیدی کروموزوم ***7*** انسانی را از جایگاه 27,140,701 تا جایگاه 27,150,700 نشان می دهد.

زبانه ی ***"Human Genome DNA 2"*** توالی نوکلئوتیدی کروموزوم ***7*** انسانی را از جایگاه 26,188,001 تا جایگاه 26,218,000 نشان می دهد.

زبانه ی ***"B. burgdorferi B31 Genome DNA 1"*** توالی کل ***DNA*** ژنوم حلقوی یک باکتری است.

(*Borrelia burgdorferi* strain B31)

در سوالات زیر ما به جایگاه باز در توالی آنالیز شده رجوع می کنیم، نه جایگاه روی کروموزوم. برای مثال، ما به اولین باز توالی Human Genome DNA 1 عدد 1 را منسوب می کنیم، نه 27,140,701 . DNA ژنومی *B. burgdorferi* سویه ی *B31* حلقوی است. به همین دلیل اولین باز بطور قراردادی مشخص شده است.

بعد از کپی کردن توالی mRNA (NM_019102.4) *HOXA5* از زبانه ی ***Predicted mRNA Sequences***، آن را در پنجره ی "Input sequence:" زبانه ی ***BLAST*** پیست کنید. (نکته: طبق قرارداد ، توالی های ***RNA*** به عنوان یک توالی ***DNA*** مکمل نشان داده می شود. این به این معنی است که یک رشته ی کدکننده ی ***DNA*** و ***RNA*** رونویسی شده ی آن کاملا مشابه نشان داده می شوند.) در ادامه، با انتخاب "Human Genome DNA 1" از بین گزینه های "Reference Sequence:" یک جستجوی تشابه بر علیه "Human Genome DNA 1" انجام می شود. این مرحله نیاز به **blastn** دارد. با کلیک کردن روی "Exec" لوکوس پیش بینی شده ی *HOXA5* *داده می شود. در نتایج جستجو،* "sseqid" نشان دهنده ی ID ی توالی های hit است.

سوال ۳. ناحیه ای که اولین اگزون HOXA5 را شامل می شود را از بین گزینه های زیر انتخاب کنید.

از "E-value:1e-5" برای آستانه ی blastn استفاده کنید.

راهنمایی: از زبانه ی ***Predicted mRNA sequences*** و ***BLAST*** استفاده کنید. **[No.8] 2 نمره**

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

سوال ۴. در ادامه ، پروتئین هایی با توالی آمینواسیدی یکسانی با HOXA6 پیدا کنید. با استفاده از HOXA6 به عنوان query، یک جستجوی blastp روی پایگاه داده ی پروتئین های انسانی انجام دهید ("human proteins" در بخش Reference sequence). از "E-value:1e-5" برای آستانه ی **blastp** استفاده کنید. از گزینه های پیش رو، **entry** ای که بیشترین similarity را دارد انتخاب کنید ولی خود HOXA6 را انتخاب نکنید.

راهنمایی: از زبانه ***Protein sequences1*** و ***BLAST*** استفاده کنید. **[No.9] 4 نمره**

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

برای رسیدن به توالی پروتئین ، شما باید به زبانه ی ***Entry Database*** داخل اپلیکیشن ***IBOC*** بروید (شما در سوال ۱۸ و ۱۹ نیاز دارید این کار را انجام دهید.) توالی آمینواسیدی پروتئین با وارد کردن ID پروتئین (برای مثال 1_05121) در بخش **Entry_ID** و انتخاب نام ارگانیسمی از **Sequence** selection، نمایش داده می شود. **توجه: این ویژگی ممکن است در جواب دادن به سوالات *18* و *19* کمک کننده باشد.**

Part 2. Analysis of sequence features and motif discovery

سوال پایین روش های برای بررسی ترکیب CG توالی نوکلئیک اسید ها را به عنوان یکی از خصوصیات DNA ژنومی معرفی می کند. اطلاعات زیر راجع به کروموزوم ها و ترکیب DNA CG بخوانید و سوالات زیر را پاسخ دهید.

کروموزوم ها، همان طور از اسم شان بر می آید، می توانند توسط روش های مختلف رنگ آمیزی رنگ شوند. در میان این روش ها، رنگ آمیزی گیمسا (Giemsa) یک الگوی راه راه از باند ها را در طول کروموزوم حاصل می کند. رنگ آمیزی معمولا در طول در مرحله **[No. 10]** تقسیم سلولی انجام می شود، هنگامی که کروموزوم ها بدلیل متراکم شدن قابل رویت هستند. رنگ گیری در نواحی با محتوای **AT [No. 11]** تیره تر و در نواحی با محتوای **CG [No. 12]** روشن تر و سبک تر است.

همچنین تصور می شود نواحی با محتوای **CG [No. 12]** ژن های بیشتری دارند. این همبستگی میان محتوای **CG** و تراکم ژن ها در سال 2000 با مشخص شدن توالی تقریبا تمام **DNA** هسته ای انسان (human nuclear DNA) در پروژه ژنوم انسان تائید شد.

همچنین DNA در دماهای **[No. 13]** دناتوره می شود به گونه ای که مارپیچ دوتایی آن به دو رشته تکی تبدیل می شود. جفت باز های **CG** در دمای **[No. 14]** ای نسبت به جفت باز های **AT** دناتوره می شوند. همان طور که در بالا شرح داده شد محتوای **CG** مقداری است که علاوه بر اینکه به راحتی بدست می آید، در جنبه های بسیاری از تحقیقات زیست فناوری کاربردی می باشد.

سوال 5. مناسب ترین پاسخ را برای **[No. 10]** انتخاب کنید. **4 نمره**

1. **G0**

2. **S**

3. متافاز

سوال 6. مناسب ترین پاسخ را برای **[No. 11]**، **[No. 12]**، **[No. 13]** و **[No. 14]** انتخاب کنید. **8 نمره**

1. بالاتر (بیشتر)

2. پایین تر (کمتر)

هم اکنون بر روی محتوای CG توالی های DNA تمرکز می کنیم. محتوای CG (GC%) توسط فرمول زیر تعریف می شود :

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

در این فرمول [A] نشان دهنده تعداد آدنوزین ها (A) و همین طور [T]، [G]، و [C] هر کدام نشان دهنده تعداد تیامین ها (T)، گوانوزین ها (G) و سیتوزین ها (C) می باشند.

سوال 7. برای توالی DNA نشان داده شده در زبانه ***"Human Genome DNA 1"***، محتوای CG را در 10 باز از اولین تا دهمین جایگاه محاسبه کنید. **2 نمره**

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

مثال) اگر جواب مدنظر شما 32 درصد است نحوه پاسخ دهی شما به شکل زیر می باشد :

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

نحوه استفاده از زبانه ***"Count nucleotide"*** و ***"Window search"*** به شرح زیر است :

توالی های DNA ناحیه مدنظر شما از "Human Genome DNA 1"، "Human Genome DNA 2" و "B. burgdorferi B31 Genome DNA 1" می توانند توسط قابلیت **Get Subsequence** (سمت چپ زبانه ***"Count Nucleotide"***) استخراج شوند.

با انتخاب کردن **Human Genome DNA 1** از لیست انتخابی **Target Sequence** و وارد کردن 1 و 10 به عنوان مقادیر "**Start**" و "**End**"، 10 باز اول از ابتدا با دهمین باز از "Human Genome DNA 1" نمایش داده می شوند.

توجه : این قابلیت می تواند برای پاسخ گویی به سوالا 15 و 16 مفید باشد.

Window search روشی برای پیدا کردن ویژگی های **DNA** (DNA features) در تمام طول توالی نوکلئوتیدی به وسیله تغییر مکان دادن یک پنجره با طول ثابت به اندازه یک واحد معیین و ثابت در هر مرحله می باشد. طول پنجره Window Size و اندازه تغییر مکان پنجره در هر بار تغییر مکان Step Size نامیده می شود.

هم اکنون ویژگی های توالی را در "Human Genome DNA 1" و "Human Genome DNA 2" به وسیله زبانه

"***Window Search***" بررسی می کنیم.

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".

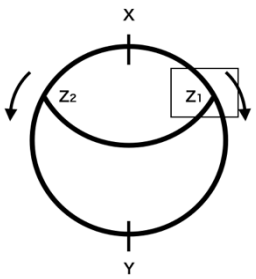
1. 10-20
2. 20-30
3. 30-40
4. 40-50
5. 50-60
6. 60-70
7. 70-80
8. 80-90

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

در قسمت بعدی با استفاده از روش های آنالیزی همانند **Window Search** که در سوال قبل آموختیم به بررسی همانند سازی **DNA** باکتریایی می پردازیم.

سوال 10. همانند سازی DNA حلقوی از یک نقطه شروع شده و در دو جهت پیش می رود (عکس 2).



عکس 2. شکل مفهومی همانند سازی DNA در ژنوم حلقوی باکتریایی

X نقطه شروع همانندسازی (Ori) و Y نقطه پایان همانندسازی می باشند. دو چنگال همانندسازی توسط Z1 و Z2 نمایش داده شده اند.

اگر توالی تمام ژنوم باکتری در دسترس باشد، برای تخمین نقاط شروع و پایان همانندسازی می توان از بررسی GC-skew استفاده کرد. GC-skew یک ملاک از میزان بایاس در تعادل میان تعداد G ها و C ها در یک رشته DNA است. (یعنی ببینیم در جفت باز های G-C آیا بایاسی در اینکه G ها در یک رشته DNA باشن و C ها در رشته مقابل وجود داره یا نه)

این ملاک توسط تفاوت در تعداد C ها و G ها تقسیم بر مجموع تعداد G ها و C ها در یک طول مشخص از توالی مشخص می شود :

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

با استفاده از **Window Search**، DNA GC-skew ژنوم *B. burgdorferi* B31 را بررسی کنید.

در زبانه "***Window Search***"، "*B. burgdorferi* B31 Genome DNA" را در بخش "Reference Sequence" انتخاب کنید. بخش **Start** و **End** را با اولین و آخرین نوکلئوتید پر کنید و سپس ([C]-[G])/([C]+ [G]) را در بخش **Skew** انتخاب کنید. در نهایت دکمه **Show Chart 3** را بفشارید و نمودار GC-skew در ناحیه Chart 3 نمایان می شود.

توجه کنید که این محاسبه ممکن است مقداری طول بکشد (حدود یک دقیقه به ازای هر محاسبه).

دو نقطه تبدیل بین GC-skew بالا و GC-skew پایین در این DNA وجود دارد. نواحی مربوط به آنها را از میان گزینه های زیر انتخاب کنید. **8 نمره**

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

دو نقطه تبدیلی در نواحی **[No. 20]** و **[No. 21]** قرار دارند.

سوال 11. می دانیم که نقطه شروع همانندسازی یا OriC در مجاورت ناحیه کدکننده پروتئین DnaA قرار دارد. ناحیه کدکننده پروتئین DnaA را در "*B. burgdorferi* B31 Genome DNA" پیدا کنید و ناحیه ای که حاوی آن است را انتخاب کنید.

راجع به ابزاری برای پاسخ به این سوال به آن نیاز دارید فکر کنید.

می توانید از توالی پروتئین **DnaA** در زبانه ***Protein Sequences 2*** استفاده کنید. **[No. 22] 2 نمره**

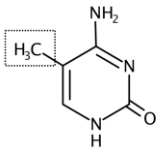
1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

سوال 12. می دانیم که مقادیر GC-skew در نقاط شروع و پایان همانندسازی به شدت تغییر می کنند. با در نظر گرفتن پاسخ خود به سوالات قبلی، محتمل ترین ناحیه برای در بر داشتن نقطه شروع همانندسازی (OriC) را در این جاندار از میان گزینه های زیر انتخاب کنید. **[No. 23] 4 نمره**

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

هم اکنون موتیف های DNA ای که در تنظیم بیان ژن در یوکاریوت ها نقش دارند را بررسی می کنیم.

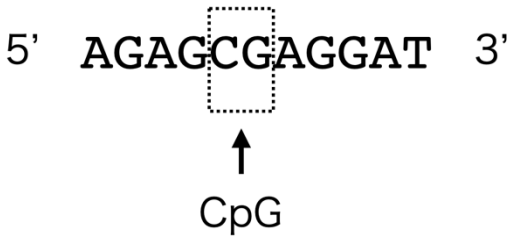
تنظیم بیان ژن می تواند توسط اضافه کردن شیمیایی گروه های متیل به باز های خاصی بر روی DNA رخ دهد.



عکس3. فرمول ساختاری 5-methylcytosine، یک سیتوزین متیله شده.

ناحیه نقطه چین نشان دهنده گروه متیل اضافه شده می باشد.

در مهره داران، سیتوزین CpG dinucleotide ها اغلب متیله می شود. یک CpG dinucleotide همانطور که از اسمش بر می آید از یک C و G پیاپی در جهت ’5 به ’3 تشکیل شده است (عکس 4). p میان C و G نماینده پیوند فسفودی استر می باشد تا CpG از CG که یک جفت نوکلئوتیدی در یک DNA دو رشته ای است تفکیک شود.



عکس 4. یک **CpG dinucleotide** از یک C و G پیاپی در جهت ’5 به ’3 تشکیل شده است.

ناحیه ای DNA که تعداد زیادی CpG dinucleotides دارد CpG island نامیده می شود. CpG islands ها معمولا در DNA ژنومی مهره داران یافت می شوند و ممکن است در اطراف اگزون های ژن های کد کننده پروتئین و یا چند صد جفت باز بالادست (upstream) ناحیه شروع رونویسی آنها قرار گرفته باشند. لزوما همه CpG islands ها متیله شده نیستند ولی این رویداد های متیله شدن با با غیر فعال شدن ژن پایین دست همبستگی دارند.

روش های مختلفی برای پیدا کردن CpG islands ها پیشنهاد شده اند. در میان آنها **CpG-score** توسط فرمول زیر تعریف می شود و می توان آن را در یک پنجره جا به جا شونده (همان طور که قبلا شرح داده شد) محاسبه کرد :

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

در این فرمول [CpG] نشان دهنده تعداد CpG ها در پنجره، و همین طور [G] و [C] نشان دهنده تعداد گوانوزین ها (G) و تعداد سیتوزین ها (C) در پنجره می باشند.

برای بدست آوردن CpG score، "Window-Size = 100" و "Step-Size = 1" معمولا پارامتر های استاندار استفاده شده می باشند. نواحی با CpG score 0.6 یا بالاتر اغلب به عنوان کاندیدای CpG islands در نظر گرفته می شوند.

سوال 13. همانطور که در بالا اشاره شد، "Human Genome DNA 1" توالی جایگاه های 27,140,701 تا 27,150,700 (مجموعا 10,000 باز) بر روی کروموزوم 7 انسان می باشد.

در زبانه "***Window Search***"، "Human Genome DNA 1" را در بخش **Reference Sequence** انتخاب کنید و پراکنش CpG-score را بررسی کنید.

مقادیر مناسب را در بخش **Start** و **End** وارد کرده و از پارامتر های "Window-Size = 800" و "Step-Size = 1" استفاده کنید. "**CpG-score**" را از بخش **CpG-score** در گوشه چپ پایین صفحه انتخاب کرده و **Show Chart 4** را فشار دهید. هم اکنون پراکنش CpG-score ها در توالی هدف نمایش داده می شود. طول ناحیه طولانی ترین کاندید CpG island را در این توالی تخمین زده و نزدیک ترین پاسخ را از میان گزینه های زیر انتخاب کنید.

در این سوال نواحی با CpG score 0.6 یا بالاتر به عنوان کاندیدای CpG islands در نظر گرفته می شوند.

توجه کنید که این محاسبه ممکن است مقداری طول بکشد (حدود یک دقیقه به ازای هر محاسبه). [No. 24] 6 نمره

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

سوال 14. سپس، ناحیه شروع رونویسی ژن ***HoxA6*** را در Human genome DNA 1 پیش بینی کنید.

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.

نزدیک ترین پاسخ را از میان گزینه های زیر انتخاب کنید.

برای توالی *HoxA6* از توالی نمایش داده شده در زبانه "***Predicted Mrna sequences***" استفاده کنید.

ناحیه شروع رونویسی *HoxA6* به جایگاه نوکلئوتید **[No. 25]** نزدیک ترین است. **4 نمره**

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

سوال 15. با توجه به سوالات 13 و 14، عبارت **نادرست** را از میان گزینه های زیر انتخاب کنید. **[No. 26] 6 نمره**

1. CpG island ای که با اولین اگزون *HoxA5* هم پوشانی دارد، از CpG island دیگری که با اولین اگزون *HoxA6* هم پوشانی دارد بلند تر است.

2. اینترون *HoxA6* با ناحیه ای که کمترین CpG dinucleotides ها را دارد تقریبا منطبق است.

3. در این توالی، ناحیه ای که کمترین CpG-score را دارد با ناحیه ای که بیشترین محتوای AT را دارد تقریبا منطبق است.

4. هر دوی *HoxA5* و *HoxA6* محتوای CG بالاتر از 60 درصد در اولین اگزون شان دارند.

5. ناحیه بین ژنی میان *HoxA5* و *HoxA6* طولی در حدود 1.7 kb دارد.

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).

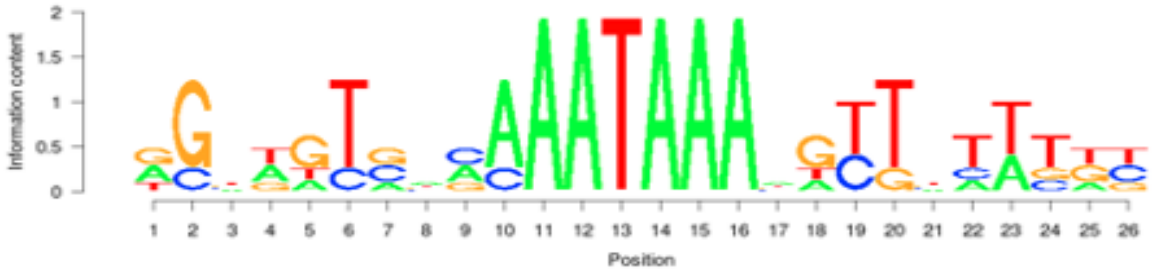


Figure 5. An example of a result of the sequence logo

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

با استفاده از نرم افزار IBOC به سوالات زیر پاسخ دهید.

با استفاده از زبانه های ***BLAST*** و ***Entry Database***، جست و جوی تشابه توالی در میان داده های پروتئین 7 جاندار زیر امکان پذیر است : 5 جانور (انسان، موش، مگس میوه، اختاپوس و starlet sea anemone) و 2 جاندار تک سلولی (choanoflagellate و fission yeast).

با بررسی بعضی از ژن های خانواده کادهرین در این جانداران، به سوالات زیر پاسخ دهید.

سوال 17. *E-Cadherin* و *N-Cadherin* نماینده ژن های خانواده کادهرین می باشند. اطلاعات مربوط این دو ژن در زبانه های ***Genes*، *Proteins*، *Predicted mRNA sequences*** و ***Protein sequences 1*** نمایش داده شده است.

بسیاری از ژن های خانواده کادهرین در جانوران فوق یک دومین پشت سر هم تکرار شونده دارند. این دومین هم در E-Cadherin و هم در N-Cadherin حداقل 5 بار یا بیشتر تکرار می شود. به این دومین شماره شناسایی (Accession ID)، Pfam-A: (i)[PF *****] داده شده است.

دومین را شناسایی کنید، از e-value کمتر از 1e-5 به عنوان آستانه hmmscan استفاده کنید.

نکته : اعداد بعد از نقطه را در شماره شناسایی Pfam نادیده بگیرید. **4 نمره**

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

مثال) اگر جواب مدنظر شما "PF00001.12" است، نحوه پاسخ دهی شما به شکل زیر می باشد :

No. 34 : 0
No. 35 : 0
No. 36 : 0
No. 37 : 0
No. 38 : 1

سوال 18. مقایسه دومین ساختاری پروتئین های خانواده کادهرین در جانوران و غیر جانوران نشان می دهد که در جانوران (ii)[PF ***** . *] به انتهای (iii) **[No.44]** در نزدیکی تکرار دومین (i) اضافه می شود.

(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).

نکته : اعداد بعد از نقطه را در شماره شناسایی Pfam نادیده بگیرید. **10 نمره**

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

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

1. N
2. C

سوال 19. از میان مجموعه اطلاعات پروتئین های choanoflagellate، ژنی را پیدا کنید که تکرارهای متعددی از دومین (i) (در سوال 17) را داشته باشد. برای شناسایی دومین (i) از e-value کمتر از 1e-5 به عنوان آستانه hmmscan استفاده کنید.

7 نمره

(هر چه تعداد تکرار بیشتری از دومین (i) در ژنی که پیدا کردید وجود داشته باشد نمره بیشتری خواهید گرفت.)

راهنمایی: شما می توانید در صورت نیاز از از قابلیت های زیر در زبانه ***HMMSCAN*** استفاده کنید.

قابلیت **Show [number] entries** می تواند برای تغییر تعداد ردیف های نمایش داده شده مورد استفاده قرار گیرد.

قابلیت **'Search: [word]** می تواند برای جستجوی یک رشته متنی مورد استفاده قرار گیرد.

مجموع تعداد ردیف ها در جدول نتایج، در گوشه پایین سمت چپ صفحه نمایش داده می شود.

(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