Формат экзамена

Экзамен рассчитан на 3 часа и состоит из 21 вопроса и 3 бонусных вопросов. Вопросы распределены по модулям, которые были пройдены на парах. Внутри модулей вопросы расположены в произвольном порядке, независимо от их уровня сложности.

Все файлы для выполнения заданий находятся на сервере по пути /srv/common/midterm/. Вы не можете редактировать эти файлы, но можете их читать. Этого должно быть достаточно для выполнения всех заданий. В крайнем случае вы можете скопировать файл в свою папку.

Записывайте ответы на задания в текстовом редакторе, который позволяет сохранить файл в формате .pdf. По умолчанию, ответы на задания должны содержать ваш код команд и их вывод. Некоторые задания могут требовать ответа в виде картинки или текста.

Удачи! (c) Мичил, Saleem, Layal, Лидия

Форматы файлов

1. Сконвертируйте sam файл в bam файл и создайте для полученного bam файла индекс. Файл находится по пути: /srv/common/midterm/task_adh1b.sam.
2. Определите сборку bam файла полученного вам в первом задании. В ответе приведите сборку.
3. Определите, есть ли у человека непереносимость лактозы с помощью IGV по bam расположенному по пути: /srv/common/midterm/task_lactose.bam. В ответе приведите мутации которые смотрели, их генотип и финальный вывод.
4. Найдите количество всех ридов, выровненных на хромосому 4, в bam файле расположенному по пути: /srv/common/midterm/task_adh1b.bam. В ответе приведите код и количество ридов.
5. Определите пол человека по bam файлу: /srv/common/midterm/task_gender.bam. В ответе приведите код и пол.
6. Сделайте bed файл включающий только регион гена *ADH1B*. Посчитайте среднее покрытие данного региона для файла полученного в первом задании. При помощи bedtools определите, какая доля гена *ADH1B* имеет покрытие x1+. В ответе приведите среднее покрытие, долю гена с покрытием x1+ и код.
7. Подсчитайте количество позиций в файле /srv/common/midterm/chip.vcf.gz. В ответе приведете количество позиций и код.

8. Используя bcftools, извлеките все варианты для интервала 5215000-5233000 хромосомы 21 из VCF файла: /srv/common/midterm/chip.vcf.gz. В ответе приведите количество вариантов.
9. В ответе запишите следующие данные: хромосома, позиция, референсный, альтернативный аллель, генотип для вариантов с гетерозиготой или гомозиготой по альтернативному аллелю у образца NA21135 в том же регионе что в прошлом задании для файла: /srv/common/midterm/chip.vcf.gz
10. Отфильтруйте позиции по колонке INFO/AF. Оставьте варианты с частотой не менее 5%. В ответе запишите количество вариантов.
11. В ответе запишите генотип всех образцов в буквенном формате (пример: 0/1 стало AG) для позиций chr21, pos 5231680, REF:T ALT: C и chr21, pos 5225197, REF: G, ALT: T.

Глобальные и локальные выравнивания

1. Выровняйте следующие последовательности с помощью алгоритма Нидлмана-Вунша:

   ATGCCCGA
   GTCACCC

Используйте следующие параметры для выравнивания: награда за совпадение: +1, штраф за замену: -1, штраф за вставку или удаление: -2.

Постройте таблицу динамического программирования для вычисления оптимального глобального выравнивания, вручную на бумажке. Найдите оптимальное выравнивание для данных последовательностей. Вычислите итоговый выравнивающий счет. В ответе приведите фотографию/скриншот таблицы с оптимальным выравниванием и итоговый выравнивающий счёт.

2. Найдите нуклеотидную последовательность белка эндонуклеазы III (Nth) из бактерии *Escherichia coli*, штамм K-12, субштамм MG1655, в базе данных NCBI (www.ncbi.nlm.nih.gov). Проведите соответствующий последовательности BLAST. Напишите какой организм имеет лучший мэтч с данной последовательностью помимо

Escherichia coli. Выпишите Max Score и E-value. В ответе приведите организм, max score и e-value.

3. Найдите белковую последовательность человеческого (Homo sapiens) гена *BRCA1*, в базе данных белков NCBI (www.ncbi.nlm.nih.gov). Проведите соответствующий последовательности BLAST. Выпишите Max Score и E-value лучшего мэтча с отличным от организма запроса. Выполните парное выравнивание мэтча с исходным запросом и сохраните получившийся dotplot. В ответе приведите. В ответе приведите организм, max score, e-value и dotplot

Множественные выравнивания

1. Возьмите белковую последовательность человеческого гена *BRCA1.* Выполните белковый BLAST с использованием базы данных Reference proteins database (Refseq protein). Из полученных результатов выберите последовательности для 4ех любых видов (Homo sapiens, Gorilla gorilla gorilla и т.д.). Получите последовательности в формате FASTA. Сократите названия чтобы они содержали только название белка и вид (например, BRCA1_Homo_sapiens).

   Подсказка: Если существует несколько изоформ белков, выберите ту, которая имеет наименьший номер. Кроме того, если для одного и того же белка имеется несколько записей, выберите ту из них, которая начинается с «NP_», или, как вариант, с «XP_».

   Выполните множественное выравнивание любым удобным для вас тулом.

   Сохраните результат множественного выравнивания в .fasta формате. Он понадобится вам для построения дереве в следующем задании

   В ответе запишите количество строк в полученном .fasta файле.

Филогенетика

1. Нарисуйте произвольное ультраметрическое укорененное древо с 6-ю листьями. В ответе приведите фотографию/скриншот.

2. Постройте филогенетические древо используя расстояние по Хэммингу и посчитав матрицы расстояний, для следующих последовательностей. Приведите матрицы расстояний и древо в качестве ответа

   AGCTGA
   AGTTGA
   CGCTGA
   AGCTGG
   CGTTGA

3. Постройте дерево записанное в Newick формате (A1:0.1,(A2:0.2,(A3:0.3,A4:0.4):0.2):0.1,A5:0.5);. В ответе приведите древо
4. Постройте дерево для файла Mammals.fasta в программе MEGA, методами UPGMA, Neighbourhood Joining, Maximum Parsimony. Укорените деревья в нужном месте на ваш взгляд. Проанализировав ВСЕ деревья в ответе приведите ответы на следующие вопросы:
   a. Кто ближайший сосед человеку – мышь или собака?
   b. Можем ли мы доказать независимое происхождение ламантинов и китообразных?
   c. Кто ближайший сосед летучим мышам – собака или человек?
5. Визуализируйте любое дерево на ваш вкус из задания 4 в программе iTOL, сделав его круговым и покрасив узел с человеком в зеленый цвет. В ответе приведите скриншот.
6. Используя программу MEGA, постройте древо методом максимального правдоподобия (ultra fast bootstrap with 1000 replicates) для результата множественного выравнивания из задания “Множественные выравнивания”. В ответ приведите дерево.

Бонусные задания на дополнительные баллы

A

У вас есть последовательности ДНК в файлах /srv/common/FASTA1.txt и /srv/common/FASTA2.txt.

1. Определите процент идентичности между этими двумя последовательностями.

2. Сделайте аннотацию к двум последовательностям (укажите ген и вид организма).

Бонус: как биоинформатик, интерпретация и прогнозирование результатов также важны. Объясните результат с эволюционной точки зрения (вы знаете, что они на x% похожи, почему вы так думаете?)

3. В FASTA2.txt, какой длины будет полученный белок в каждом из следующих случаев
   -Инсерция AATAGACCC после позиции 132
   -Делеция TCCC в позиции 623
   -Делеция TG после позиции 165

4. Праймеры - это небольшие участки ДНК, которые позволяют связаться ферменту ДНК-полимеразе с ДНК и амплифицировать последовательность ДНК в методе ПЦР, поскольку они выполняют функцию “границ” последовательности. Праймеры образуют пары, и каждый из праймеров прикрепляется к каждой из нитей ДНК. У вас есть 6 праймеров, и вам нужно найти, какой из них может прикрепиться к какой последовательности в файлах FASTA1.txt и FASTA2.txt (помните, что каждой последовательности нужно присвоить пару)

5'-AACAATAGTGATGCGG-3
5'-GCATGCGCGTACGG-3
5'-ATGAGGGCATTTCC-3
5'-GAAACGGCACCGTAAC-3
5'-ACTTCGACGATCACCA-3
5'-CTGTGCAAAATGTCGG-3

B

По пути есть файл /srv/common/midterm/chip1.vcf.gz:

1. найдите MAF для следующих вариантов:

rsID	POS	REF	ALT
rs1336612648	97431	C	T
rs61157983	49972112	GGT	G

2. найдите количество инделов.
3. Рассмотрим вариант rs61157983, расположенный в гене A и вызывающий миссенс-мутацию, которая приводит к наследственному заболеванию W в случае альтернативной гомозиготности. Если известно, что образец (HG00100) вступает в брак со здоровым человеком, укажите доли потомства, которое будет иметь заболевание (все аллели ALT), носителей (гетерозиготных) и здоровых (имеющих только REF).

C

Мотивы ДНК - это паттерны в последовательности, которые в некоторых позициях оказываются относительно консервативными. Поиск мотивов основан на сравнении нескольких последовательностей в одном и том же локусе. Например:

A|A|A|T|G|G|C|C|C|G|A|A|G|T|A

A|C|A|A|A|G|C|G|G|G|T|A|C|T|A

A|T|A|A|G|G|C|G|C|G|A|A|T|G|A

A|A|A|A|C|G|C|A|C|G|T|A|T|A|T|A

Среди предыдущих последовательностей мотивов (например, от разных видов) мы видим, что некоторые позиции более консервативны, чем другие.

Мы называем мотив **консенсусом**, если он представляет собой наиболее сохранившуюся последовательность в данной позиции (среди видов, образцов и т. д.).

Таким образом, консенсусом предыдущего мотива является:

AAAAGGCGCGWANTA

!W: обозначает слабый, A или T (см. вики: нуклеотиды; сокращения)

N: обозначает любой нуклеотид.

У вас есть список из n сегментов ДНК, каждый из которых имеет одинаковую длину k. Они выравниваются в одном и том же месте, образуя мотив. Найдите консенсус этого мотива, используя свои навыки программирования.

Мотивы:

«TCGGGGGTTTTT», „CCGGTGACTTAC“, „ACGGGGATTTTC“, „TTGGGGGACTTTT“, „AAGGGGACTTCC“, „TTGCGGACTACC“, „TCGCGGATTAAT“, „TCGCGGATTACT“, „TAGCGGAACAAC“, „TCGCGTATAACC“.