**Задание**

**Зачет ведется по семи группам.** Дается по два вопроса из каждой группы. Базовые вопросы:

1. Методы вычисления ожидаемого числа встреч слова.

2. Формула для ИС и умение ее применить  
3. Формула для расчета PWM. Умение вычислить вес выравнивания и PWM

4. График покрытия прочтениями Chip-seq на примере TATA-бокса (задание 4 из блока 3; засчитывается, если выполнено на коллоквиуме без представления на веб-странице)

5. Интерпретация дерева домена (задание 2 из блока 4, должно быть выполнено)

6. Вес выравнивания профиля HMM и последовательности. Профиль и выравнивание даны преп.

7. Построение ROC-кривой по результатам поиска по профилю ( задание 3 блока 4; засчитывается построение ROC-кривой на коллоквиуме по данным от преподавателя)

**1. "Точные" сигналы**

1. Как проверить является ли слово TA исключительным по числу встреч в геноме E.coli?   
2. Почему динуклеотид CG встречается в геномах млекопитающих реже, чем ожидается?

3. Как вычислить ожидаемое число встреч слова ATG в геноме E.coli методом, основанным на марковской модели? (формула в применении к слову)

4. Как вычислить ожидаемое число встреч слова ATG в геноме E.coli методом, предложенным С.Карлиным? (формула в применении к слову)

5. В геноме человека динклеотид CG встречается в несколько раз реже, чем ожидается. Оценим ожидаемое число слов TCG исходя из произведения частот нуклеотидов f(T) = 0.3, f(С) = f(G) = 0.2. Как вы думаете, на самом деле число слов TCG в геноме меньше или больше так вычисленного ожидаемого при предположении, что именно это слово не находится под отбором? Почему?

6. Зависит ли контраст слова от метода вычисления? Вычислите с помощью сервиса контраст слова GATC в геноме E.coli по марковской модели, по Карлину и по частотам букв, сравните и объясните результаты.   
7. Обнаружено, что контраст некоторого слова в геноме много меньше единицы. Какой биологический смысл может стоять за этим явлением? (Фантазирование на тему и известные примеры)

**2. Информационное содержание мотива, заданного выравниванием**

8. Приведите формулу для информационного содержания колонки нуклеотидного выравнивания. Примените ее для двух колонок с данными частотами нуклеотидов.

9. Постройте и исследуйте график зависимости информационного содержания от частот букв в случае двухбуквенного алфавита W (A или T) и S (G Или C). Считайте, что базовые частоты W и S равны.

10. Напишите и объясните формулу для вычисления информационного содержания мотива, представленного выравниванием.

11. Найдите максимум и минимум информационного содержания выравнивания длиной 10. Базовые частоты всех нуклеотидов равны.

12. Информационное содержание сигнала равно 10. Грубо оцените число случайных встреч сигнала в геноме бактерии; в совокупности промоторных областей генов. Размер генома бактерии, число генов и размер промоторной области выберите самостоятельно, по возможности, близкими к реальности.

13. Что такое сила сигнала? Как оценить силу сигнала при наличии выравнивания?

14. Объясните, как строится LOGO для выравнивания

**3. Позиционная весовая матрица**

15. Какие исходные данные нужны для рассчета позиционной весовой матрицы выравнивания (PWM)? Напишите формулу для вычисления элемента PWM. Базовые частоты нуклеотидов считать равными.  
16. Что такое псевдоотсчеты? Как и зачем они используются в позиционных весовых матрицах (PWM)?

17. Объясните, что такое отношение правдоподобия на примере PWM

18. Объясните, какая задача решается с помощью позиционной весовой матрицы. Исходные данные, этапы, результат. Какие ограничения у этого метода?

19. Напишите выражение для веса выравнивания последовательности и PWM. Можно - на примере.

20.(\*) Какие исходные данные нужны для рассчета позиционной весовой матрицы выравнивания (PWM)? Напишите формулу для вычисления элемента PWM по Миронову и Co: частоты нуклеотидов в колонке сравниваются с их средним геометрическим

**4. Поиск сигналов в данных Chip-seq**

21. Что такое Chip-seq? Основные этапы эксперимента Chip-seq

22. Какие данные получаются в результате Chip-seq эксперимента (вход для биоинформатического анализа)?

23. В чем состоит контрольный эксперимент (один пример) и зачем он нужен?

24. Какие длины прочтений (ридов) предпочтительней для Chip-seq: 150 п.н. или 35 п.н.? Почему?

25. Что такое парные прочтения? Нужны ли они в Chip-seq экспериментах?

26. Что такое покрытие прочтениями (ридами) при секвенировании? Как зависит ли число пиков сигналов Chip-seq от покрытия прочтениями?

27. Что такое график покрытия по результатам Chip-seq? Как он строится? Показать на примере TATA-боксов в геноме человека (по результатам выполненного задания 4 из блока 3; при отсутствии выполнить его на месте и объяснить).

28. Что такое «pick calling»? Вход, выход. Объяснить на примере TATA-боксов в геноме человека (по результатам выполненного задания 4 из блока 3; при отсутствии выполнить его на месте и объяснить)

28. Этапы биоинформатического анализа Chip-seq

29. Как найти сигналы в данных Chip-seq?

30.(\*) Почему наблюдается сдвиг пиков от разных цепочек ДНК?

31. Найдите сигнал TATA-box в геноме человека с помощью геномного браузера (по результатам выполненного задания 4 из блока 3; при отсутствии выполнить его на месте и объяснить)

**5. Интерпретация филогенетического дерева**

31.Объясните результаты задания № .... .   
Покажите выравнивание, по которому строилось филогенетическое дерево.   
Была ли предобработка выравнивания и если да, то в чем состояла?   
Метод построения дерева.

Как укоренено дерево? Если не укоренено, то как это можно было сделать?

Была ли оценка качества ветвей и если нет, то как ее можно было выполнить?

Объясните и обоснуйте вероятный ход эволюции в соответствии с деревом.

Какие причины, технические и объективные, могли привести к тому, что дерево неправильно отражает эволюцию?

**6. Профиль**

32.В чем различие требований ко входному выравниванию для построения HMM профиля и PWM?

Что такое профиль HMM?

33. Дано выравнивание HMM профиля и последовательности (как в презентации). Вычислите вес этого выравнивания.

34. Что такое калибровка профиля и зачем она нужна?

35. Выбор порога веса для профиля

**7. ROC-кривая**

36. Таблица сопряженности признаков 2х2: предсказание против факта (на примере профиля). Что такое чувствительности и специфичность предсказания?

37. Какие данные нужны для построения ROC-кривой?

38. Объясните алгоритм построения ROC кривой на примере результатов задания №; если не выполнено, то постройте ROC кривую по предложенным преподавателем данным.

39. Какие задачи решаются с помощью ROC-кривой? На примере результатов задания №; если не выполнено, то постройте ROC кривую по предложенным преподавателем данным.