

1. Сколько белков длины 16 аминокислот можно составить из 20 аминокислот? (1 балл)

20<sup>16</sup>

2. Каков возраст Вселенной и возраст Земли в миллиардах лет. (1 балл)

13.7 млрд лет и 5 млрд лет

3. Что такое коалесценция? (1 балл)

Существование общего предка для общих генов, или сходимость генетических родословных к общему предку. Например, с помощью теории коалесценции было оценено время, когда жил общий предок митохондриальной ДНК или Y-хромосомы.

4. Что такое эволюционное расстояние между последовательностями? (1 балл)

Ожидаемое число замен между двумя последовательностями. ( В моделях эволюции нуклеотидных последовательностей выводится через наблюдаемое число замен)

5. Назовите модели эволюции нуклеотидных последовательностей. В чем их основное отличие друг от друга? (2 балла)

Модель Джукса-Кантора. Двупараметрическая модель Кимуры. Модель Хасегава-Кишино-Яно. Они отличаются вероятностями нуклеотидных замен.

6. Назовите модели эволюции аминокислотных последовательностей. (2 балла)

Эмпирические матрицы замен - PAM, BLOSSUM, JTT

7. В матрице BLOSSUM на пересечении G (глицина) и L (лейцина) стоит цифра -4. Что это значит? (2 балла)

Матрицы BLOSSUM заполняются по следующей формуле,

$$S_{ij} = 2 \log_2 \frac{P_{ij}}{P_i P_j}$$

где  $p_i$ ,  $p_j$  – частоты встречаемости аминокислот в рассматриваемой базе,  $p_{ij}$  - частота реальных замен. Получается, что замена глицина на лейцин происходит реже, чем ожидалось бы при случайном распределении аминокислот в белках.

8. Вес матриц счета является (выберите правильный ответ) (1 балл)

- (A) энтропией Шэннона
- (B) относительной энтропией
- (C) ожидаемой энтропией

9. Две выровненные последовательности идентичны на 80%. Какая из ниже перечисленных матриц является **наиболее** адекватной? (выберите правильный ответ) (1 балл)

- (A) PAM 1
- (B) PAM 8
- (C) PAM 25
- (D) PAM 200

10. Две выровненные последовательности идентичны на 20%. Какая из ниже перечисленных матриц является **наиболее** адекватной? (выберите правильный ответ) (1 балл)
- (A) PAM 1  
 (B) PAM 8  
 (C) PAM 25  
 (D) PAM 200
11. Вам нужно выровнять две последовательности, которые идентичны на 90%. Какая из матриц будет **наиболее** подходящей? (выберите правильный ответ) (1 балл)
- (A) Blosum 35  
 (B) Blosum 65  
 (C) Blosum 90  
 (D) Blosum 80
12. Вы осуществили поиск белковой последовательности с помощью программы BLAST. Вы получили хит с E-value=1.2 e-13. Что это значит? (выберите правильный ответ) (2 балла)
- (A) Вероятность найти исходную последовательность в базе данных равна 1.2 e-13  
 (B) Вероятность полученного веса выравнивания искомой и найденной последовательности в базе данных равна 1.2 e-13  
 (C) Число последовательностей, которые можно найти в базе данных с полученным весом выравнивания, равно 1.2 e-13  
 (D) Число последовательностей, которые можно найти в базе данных с полученным весом выравнивания или большим, равно 1.2 e-13
13. Вы сравнили последовательность A с последовательностью B и получили хорошее выравнивание по всей длине последовательностей (E-value  $<10^{-9}$ ). Последовательность B также получила значимый хит с последовательностью C (E-value  $<10^{-12}$ ). Вы считаете эти значения E-values достаточными для доказательства гомологии. (2 балла)
- (A) Эти факты свидетельствуют, что последовательность A также гомологична последовательности C.  
 (B) Эти факты могут быть использованы для того, чтобы определить наличие гомологии между A и C.  
 (C) Эти факты говорят о возможном существовании гомологии между A и C, но чтобы быть полностью уверенным, нужно знать значение E-value между A и C.  
 (D) Ни один из вышеперечисленных.
14. Для выравнивания двух последовательностей AAAC и AGC составьте матрицу методом динамического программирования и найдите обратный путь. 1 – совпадение, 0 – несовпадение, -1 разрыв. (5 баллов)

		A	A	A	C
	0	-1	-2	-3	-4
A	-1	$\max(1,-2,-2) = 1$	0	-1	-2
G	-2	$\max(-1,-3,0) = 0$	1	0	-1
C	-3	$\max(-3,-4,-1) = -1$	0	1	1

15. Что такое афинная функция штрафов? (1 балл)

Функция, которая отдельно учитывает открытие разрыва и продолжение разрыва  
 $\gamma(n) = -d \cdot (n-1)e$ , где  $d$  - штраф за открытие,  $e$  - штраф за продолжение разрыва

16. Что такое правдоподобие? (2 балла)

Например: Условная вероятность наблюдения данных при том, что использовалась данная модель.

В общем случае – совместное распределение выборки из параметрического распределения, рассматриваемое как функция параметра.

17. Теорема Байеса (2 балла)

$$P(X|Y) = P(Y|X) \cdot P(X) / P(Y)$$

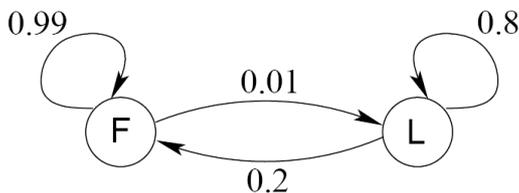
18. Что скрыто в скрытых моделях Маркова? (1 балл)

состояния

19. В каких задачах биоинформатики применяются скрытые цепи Маркова? (2 балла)

Предсказание кодирующих участков, границ экзонов-интронов, CpG-островков, определение принадлежности последовательности к семейству по профилю множественного выравнивания

20. Внизу дана диаграмма цепи Маркова для нечестного казино. F – честная кость (fair), L – нечестная кость (loaded). Для честной кости все вероятности равновероятны. Для нечестной –  $P(6) = 0.5$ , остальные равно 0.1. Распишите вероятность наблюдения данных 656 для какого-нибудь одного возможного пути. (5 баллов)



Например, для пути LFL:  $P(L) \cdot P(6|L) \cdot P(L \rightarrow F) \cdot P(5|F) \cdot P(F \rightarrow L) \cdot P(6|L)$

21. Будет ли вероятность наблюдения данных 626 отличаться от вероятности наблюдения данных 656? (2 балла)

нет

22. Что находит алгоритм Витерби? (3 балла)

Наиболее вероятный путь состояний, которые испустили наблюдаемые данные

23. Что находит алгоритм просмотра вперед? (3 балла)

Полную вероятность данных, сгенерированных данной моделью (скрытой цепью Маркова)

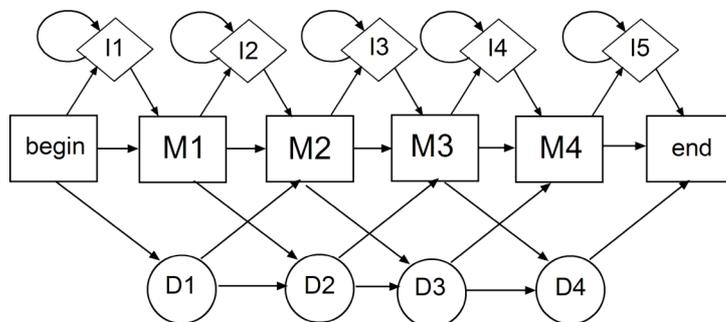
24. Чем отличаются гены прокариот от генов эукариот? (2 балла)

Интроны, экзоны, 5' UTR, 3' UTR

25. Перечислите состояния скрытых цепей Маркова для предсказания генов прокариот. (2 балла)

Старт, стоп, кодирующее, некодирующее

26. Нарисуйте схематически скрытую цепь Маркова для профилей множественного выравнивания (используется в базе данных Pfam) (4 балла)



27. Дан профиль.

AAT-T  
A-A-T  
ATT-T  
AATCT  
MDMIM

Присвойте колонкам состояния M-match (совпадение), I-insertion (вставка), D-deletion (делиция). Распишите вероятность в ФОРМУЛЬНОМ виде того, что

последовательность АААСТ была испущена данной моделью, т.е. через обозначения  $P(M)$ ,  $P(A|M)$ , и так далее. (5 баллов)

$P(M)P(A|M) P(M \rightarrow D)P(A|D) P(D \rightarrow M)P(A|M) P(M \rightarrow I)P(C|I) P(I \rightarrow M)P(T|M)$