Сигналы и мотивы -2

De novo поиск сигналов в последовательностях

Коллоквиум на 5м занятии (пр.10)

- Зачтённые задания снимают соответствующие вопросы коллоквиума
  - PWM
  - IC
  - MEME и FIMO (технология поиска сигнала de novo)
  - И далее ….

Содержание

I. Повторение: PWM, отношение правдоподобия = ln(observed/expected), псевдоотсчёты, информационное содержание (IC), сила сигнала

II. Алгоритмы поиска мотивов de novo

   1) MEME

   2) Gibs sampler

   3) ChiPMunk и HOCOMOCO

III. Поиск сигналов с помощью PWM (FIMO)

IV. Примеры сигналов для поиска de novo в задании

Содержание

- IC повторение
- Алгоритмы поиска мотивов в последовательностях
  - Постановка задачи
  - Пакет MEME, входные параметры
  - Ограничения MEME
  - Идея Gibbs Sampling
  - Другие программы
  - Chip-seq и обработка его результатов
  - Словарик
  - Задания
- Инициация транскрипции у прокариот (сайт посадки сигма субъединицы -35 и -10)
- Инициация трансляции у прокариот.
- Сигнал разрывной транскрипции у коронавирусов.

I. Вес = Логарифм отношения правдоподобия

```
1234567890123456
ACGCAAACGTTTTCTT
TCGCAAACGTTTGCTT
ACGCAAACGTTTTCGT
ACGCAAACGGTTTCGT
ACGCAACCGTTTTCCT
ACGCAAACGTGTGCGT
ACGCAATCGGTTACCT
GCGCAAACGTTTTCGT
AGGAAAACGATTGGCT
AAGCAAACGGTGATTT
ATGCAATCGGTTACGC
AGGCAAACGTTTACCT
GAGCAAACGTTTCCAC
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	**10**	2	0	1	**13**	**13**	**10**	0	0	1	0	0	4	0	1	0
G	2	2	**13**	0	0	0	0	0	**13**	4	1	1	3	1	**5**	0
T	1	1	0	0	0	0	**2**	0	0	**8**	**12**	**12**	**5**	1	3	**11**
C	0	**8**	0	**12**	0	0	**1**	**13**	0	0	0	0	1	**11**	4	2
Всего	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13

Наблюдаемая частота G в позиции 15 равна 0.38

Если GC состав генома равен 0.7, то частота G в геноме равна 0.35. Значит, ожидаемая частота G в колонке 15, как и в любой другой в предположении выравнивания случайных посл-й из генома равна 0.35.

Отношение правдоподобия = (наблюдаемая частота G в позиции 15): (ожидаемая частота G =0.38/0.35

Вес за букву G в позиции 15 этого сигнала заданного последовательностью длины 16 равен

$w(G,15) = \ln(0.38/0.35) = 0.1$

I. Информационное содержание выравнивания последовательностей сигнала. LOGO. «Сила» сигнала

Повторение Л.1.

Информационное содержание IC сигнала, заданного выравниванием

```
1234567890123456
ACGCAAACGTTTTCTT
TCGCAAACGTTTGCTT
ACGCAAACGTTTTCGT
ACGCAAACGGTTTCGT
ACGCAACCGTTTTCCT
ACGCAAACGTGTGCGT
ACGCAATCGGTTACCT
GCGCAAACGTTTTCGT
AGGAAAACGATTGGCT
AAGCAAACGGTGATTT
ATGCAATCGGTTACGC
AGGCAAACGTTTACCT
GAGCAAACGTTTCCAC
```

- Измеряет насколько сигнал отличается от случайной последовательности такой же длины
- Чем дальше от случайного – тем больше в нем информации и меньше его энтропия
- IC = Hbefore - Hafter

ШАГ 1. Подсчёт числа букв N(b,j)

```
1234567890123456
ACGCAAACGTTTTCTT
TCGCAAACGTTTGCTT
ACGCAAACGTTTTCGT
ACGCAAACGGTTTCGT
ACGCAACCGTTTTCCT
ACGCAAACGTGTGCGT
ACGCAATCGGTTACCT
GCGCAAACGTTTTCGT
AGGAAAACGATTGGCT
AAGCAAACGGTGATTT
ATGCAATCGGTTACGC
AGGCAAACGTTTACCT
GAGCAAACGTTTCCAC
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	**10**	2	0	1	**13**	**13**	**10**	0	0	1	0	0	4	0	1	0
G	2	2	**13**	0	0	0	0	0	**13**	4	1	1	3	1	**5**	0
T	1	1	0	0	0	0	**2**	0	0	**8**	**12**	**12**	**5**	1	3	**11**
C	0	**8**	0	**12**	0	0	**1**	**13**	0	0	0	0	1	**11**	4	2
Всего	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13

G C C T A C C C C A T T A T T T...

ШАГ 2. Частоты букв f(b,j)

f(b,j) = N(b,j)/N в примере N=13

Частоты	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.77	0.15	0.00	0.08	1.00	1.00	0.77	0.00	0.00	0.08	0.00	0.00	0.31	0.00	0.0
G	0.15	0.15	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.31	0.08	0.08	0.23	0.08	0.3
T	0.08	0.08	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.62	0.92	0.92	0.38	0.08	0.2
C	0.00	0.62	0.00	0.92	0.00	0.00	0.08	1.00	0.00	0.00	0.00	0.00	0.08	0.85	0.3
Всего	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

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

Величина IC для буквы b в позиции j выравнивания

$$IC(b,j) = f(b,j)*\log_2[f(b,j)/p(b)] = f(b,j)*w(b,j)$$

$\log_2[f(b,j)/p(b)] = \lambda w(b,j)$ – вес из матрицы PWM **без псевдоотсчётов,** где λ - константа перехода от двоичных логарифмов к натуральным $\lambda = \ln 2$

IC(b,j) **положительное число ⇔ f(b,j) > p(b)**

(как вычислять при f(b,j) = 0 ?)

Если f(b,j) = 0, то IC(b,j) = 0 (теорема)

Также IC(b,j) = 0 если частота f(b,j) = p(b)

Максимум IC(b,j) = $\log_2[1/p(b)]$ для минимальной p(b)

Величина IC(j) для колонки j

$$IC(j) = \sum_b f(b,j)*w(b,j))$$

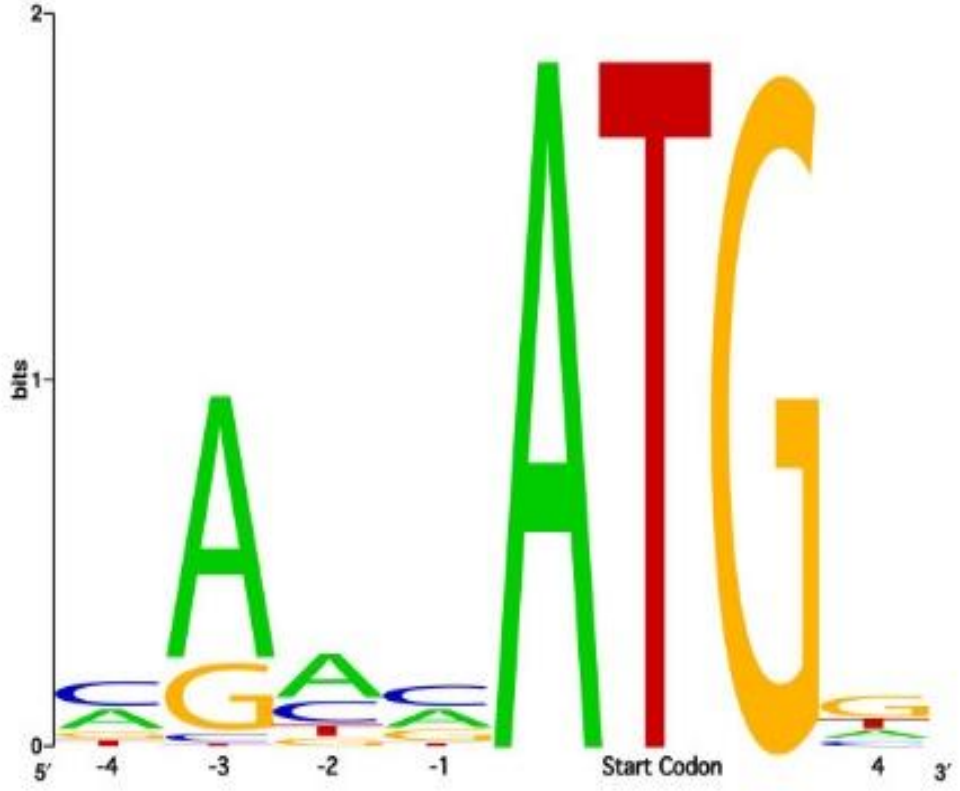
Из формулы следует, что IC(j) – матожидание - веса в колонке при распределении вероятностей букв b заданного частотами букв в колонке

Теорема. $0 \le IC(j) \le$ (?) $\max(\log_2 1/p(b))$ При $p(b) = ¼$ имеем 2

Чем больше IC(j), тем больше частоты букв в колонке отличаются от ожидаемых, тем больше информации в колонке

Информационное содержание IC выравнивания равно

$$IC = \sum_j IC(j)$$



В LOGO сигнала буквы имеют высоту, равную информационному содержанию букв

webLOGO.

$$Rseq = Smax - Sobs = \log_2 N - (- \Sigma f(b) \log_2 f(b))$$

S – энтропия колонки.
N = 4 для ДНК, т.к. 4е буквы, $\log_2 N = 2$

$$IC(\text{колонки}) = \Sigma f(b) \log_2 f(b) - \Sigma f(b) \log_2 p(b)$$

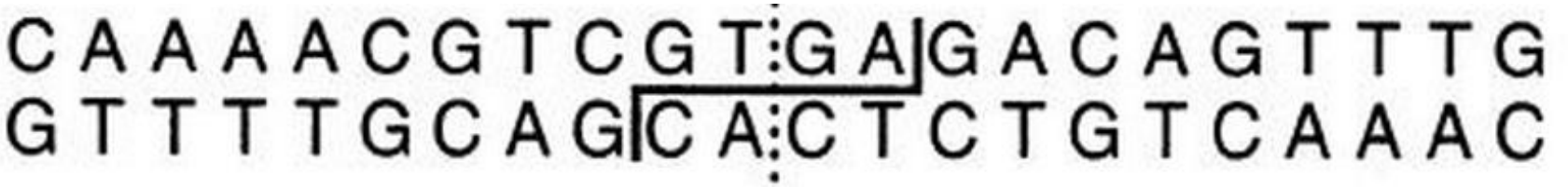
При p(b) = ¼ для всех b получаем

$$IC(\text{колонки}) = \Sigma f(b) \log_2 f(b) + 2 * \Sigma f(b)$$

Совпадает с Rseq

Примеры

- Слабый сигнал:
  - Гомеодомен - консервативный ДНК-узнающий домен многих важных транскрипционных факторов эукариот
  - Узнаёт короткую последовательность ДНК
  - На основании наложения структр гомеодоменов найден единственный общий контакт домена с сайтом ДНК: Asn51 две водородных связи с аденином (!)
  - Сигнал NNANN слабый)))
- Сильный сигнал:
  - Эндонуклеаза I-CreI семейства LAGLIDADG узнает такую последовательность. Вероятность обнаружить в геноме такую последовательность случайно близка к 0



Информационное содержание как мера силы сигнала

- В грубом приближении два выравнивания с одинаковым информационным содержанием дадут одинаковое число «случайных» находок в «случайном» банке

- Информационное содержание «выравнивания» из одной последовательности из n букв равно, 2n (по формуле)
- Сколько раз случайно встретится слово длины n в геноме длины N? В грубом приближении

$$N/(4^n) \text{ раз}$$

Значит если информационное содержание выравнивания равно 10, то случайных находок в геноме размера N будет

$N/(4^5)$ - примерно, 1 на 1000 п.н.

Надо понимать, что такая оценка грубая, но грубые оценки полезны!

ИС измеряет отклонение частот от случайного

II. Алгоритмы поиска мотивов в последовательностях

* MEME: Multiple Expectation Maximization for Motif Elicitation

* gibbs sampling for motif finding

Задача поиска мотивов

Сигнал - последовательность (напр. нуклдеотидов), адресованная одному белку или комплексу белков, и вызывающая одну реакцию. Предполагается, что последовательности одного сигнала похожи (в редких случаях полностью совпадают)

Мотив – описание сигнала: PWM, паттерн, др. правило

Примеры: *от слушателей*

Дано: набор последовательностей, в которых предполагается наличие сигнала

Результат: один или несколько достоверных мотивов. Каждый мотив – предполагаемый сигнал.
Для каждого сигнала **в ответе:** координаты сигнала; выравнивание всех последовательностей, PWM, *информационное содержание и LOGO*

1) Пакет MEME

- Входные параметры позволяют ввести ограничения на искомый сигнал:
  - Число разных сигналов, которые выдает программа
  - Длина последовательности сигнала
  - Ограничения на число находок сигнала в одной последовательности
  - Искать ли на комплементарной цепи
  - Вариант выбора базовой модели для вычисления базовых частот букв

Алгоритм MEME

1. Последовательно берем фрагмент заданной длины в каждой последовательности, ищем похожие фрагменты в других последовательностях, строим выравнивание. Берем базовые частоты букв из дополнения.
2. Для каждого выравнивания получаем PWM с максимальным весом, используя алгоритм EM (Expectation maximization)
3. Выбираем заданное число PWM с лучшим весом
4. Если задан поиск мотивов разной длины, то все заказанные длины перебираются

Алгоритм EM (Expectation maximization)

- На входе выравнивание и PWM
- По очереди удаляем фрагмент из выравнивания, и заменяем его на лучший по PWM фрагмент в соответствующей последовательности
- Повторяем пока процесс не сходится
- Находим максимальный вес, записываем PWM с максимальным весом

E-value мотива, найденного с помощью MEME

- MEME улучшалась несколько раз
- В классическом варианте
  - Нужно одно число на выравнивание (аналог веса для BLAST). Это число – информационное содержание
  - E-value должно показывать мат.ожидание числа мотивов с тем же или большим IC, получаемых поиском MEME в случайном банке того же размера и состава
  - Хорошей математической теории, позволяющей быстро вычислить E-value нет.
  - Используют эвристические алгоритмы

Ограничения MEME

1. Предположение о независимости позиций выравнивания
2. Находит только мотивы без гэпов
3. Последовательности должны быть как можно короче и содержать минимум шума
4. После 40 последовательностей, включение дополнительных последовательностей не улучшает работу алгоритма

2) Gibbs Sampling

- Первый шаг такой же, как в MEME: выбор выравнивания A из случайных фрагментов
- Шаг состоит в удалении одного фрагмента и замене его случайным фрагментом из той же последовательности => новое выравнивание B
- Если I(B) > I(A), то берем B
- Если I(B) < I(A), то с вероятностью

$$\mathbf{P = exp\,[\,(I(B) - I(A))\,/\,T\,]}$$

берем B, иначе оставляем A

- В начале “температура” T большая => почти все замены на худшее выравнивание B принимаются; с каждым шагом температура понижается, так что все более жесткие условия на то, чтобы взять B.
- “Тепловой отжиг” (Как в ПЦР☺)

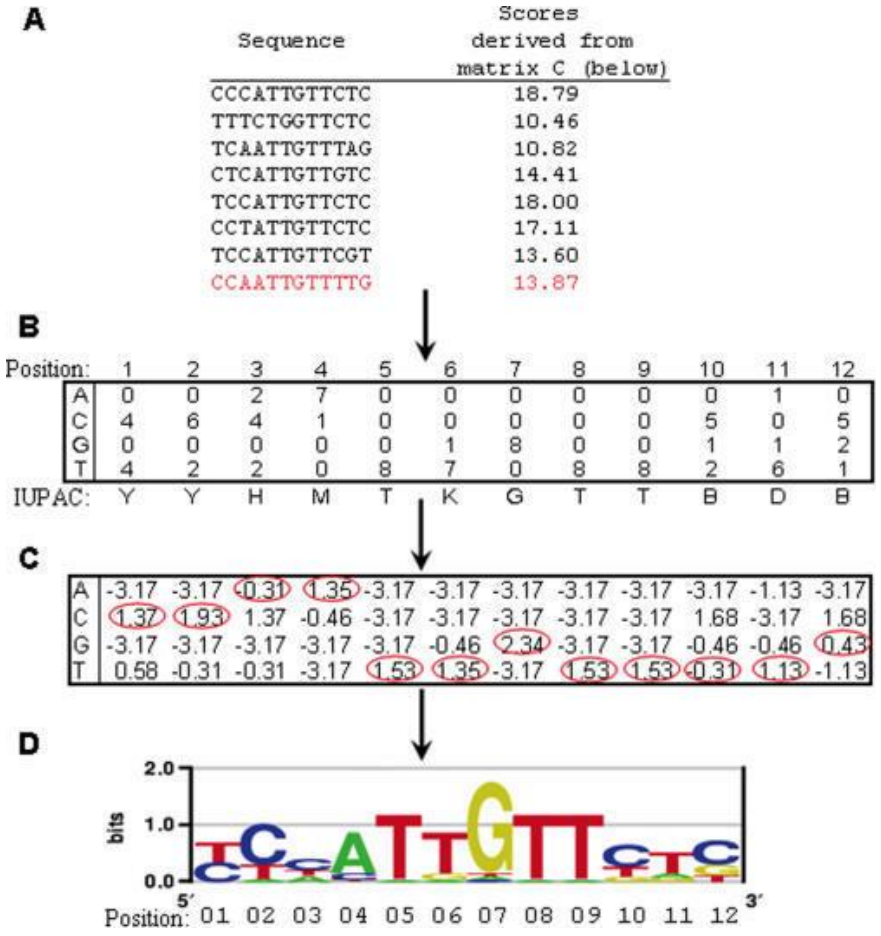
3) Как-то упустил что наши люди – коллеги - тоже сделали детектор мотивов Chipmunk (https://opera.autosome.ru/chipmunk/discovery)

Можете попробовать в своей задаче

III. Find Individual Motif Occurrences (FIMO)

1. FIMO ищет встречи каждого из входных мотивов по очереди, независимо друг от друга
2. Использует алгоритм динамического программирования
3. Вычисляет p-value для каждой находки.
4. Из-за проблемы множественного тестирования, p-value неправильно считать единственным показателем хорошей находки
5. FIMO instead reports for each *P*-value a corresponding *q*-value, which is defined as the minimal FDR threshold at which the *P*-value is deemed significant

Поиск мотива с использованием позиционно-весовой матрицы

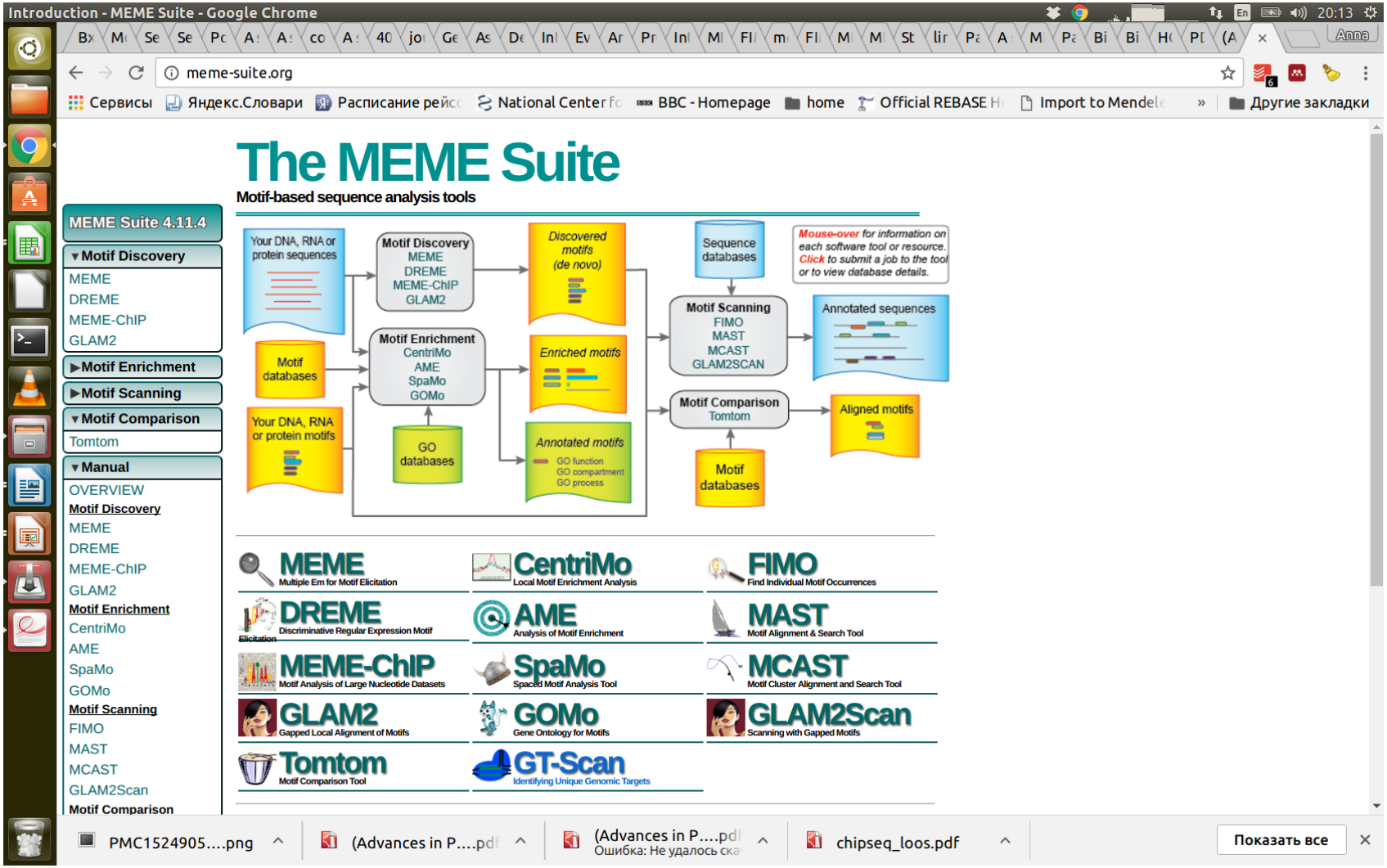


Вес (I(bj) основания b в данной позиции j
I(bj)=f(bj)*log f(bj) — p(b)*log p(b), где f(bj) — частота основания и в позиции j выравнивания, p(b) — фоновая частота основания b

Вес позиции — сумма по столбцу, вес мотива — сумма весов позиций

GuhaThakurta D. Computational identification of transcriptional regulatory elements in DNA sequence. Nucleic Acids Res. 2006 Jul 19;34(12):3585-98.

Набор программ для работы с мотивами



MAST – другая программа из пакета MEME для поиска новых сигналов по нескольким PWM в большом наборе последовательностей

IV Примеры сигналов

Для заданий практикума 7

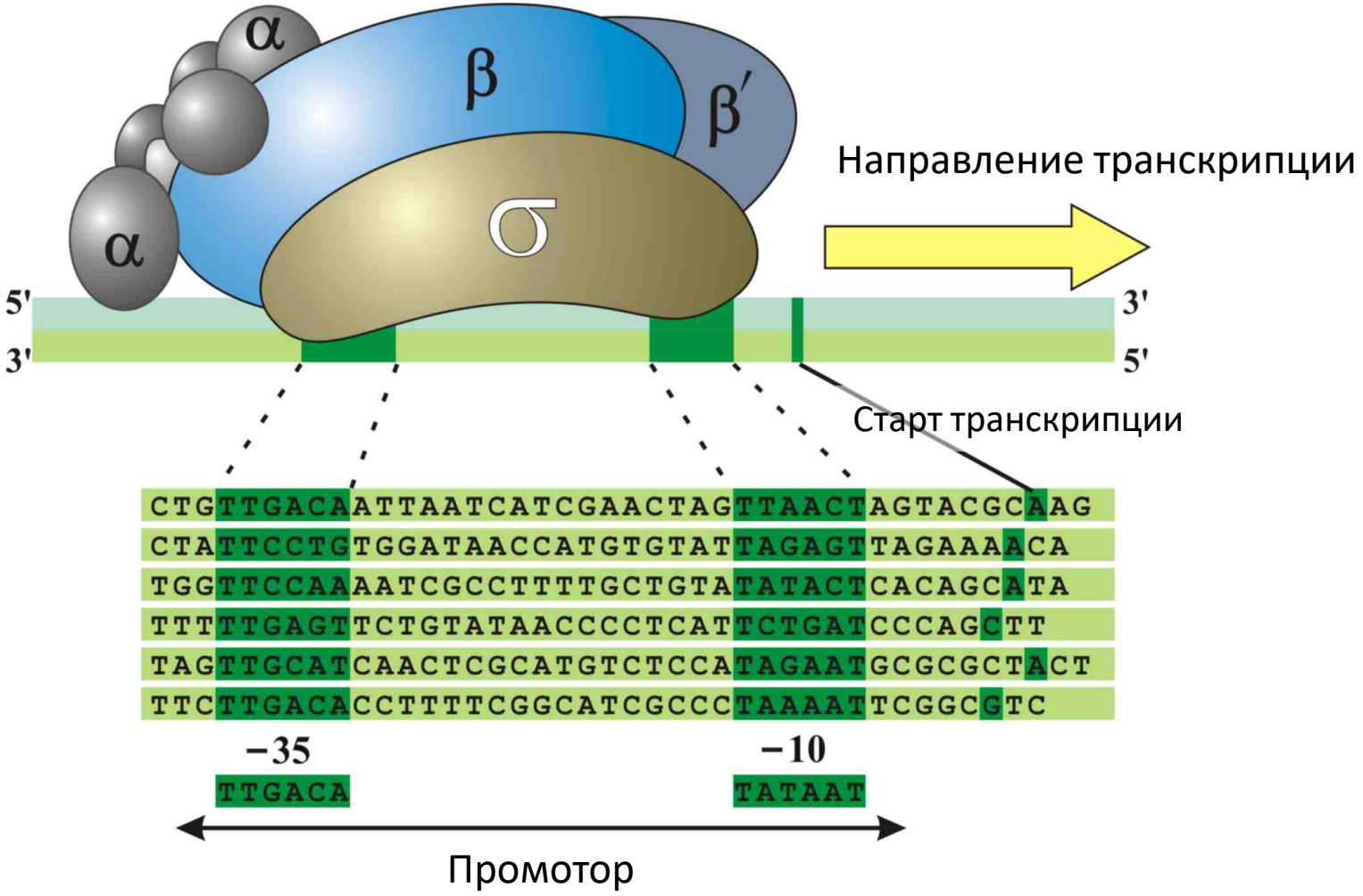
- Промотеры прокариот (инициация транскрипции)
- Сайты посадки рибосомы у прокариот (Shine-Dalgarno = SD последовательности)
- Сигналы разрывной транскрипции у коронавирусов

а. Промотор: последовательность ДНК, узнаваемая белками для инициации транскрипции

- Прокариоты
  - Схема с ДНК и белками
  - Выравнивание для E.coli
- Эукариоты - сложнее
  - Схема инициаторного комплекса TFIID
  - Выравнивание TATA-боксов

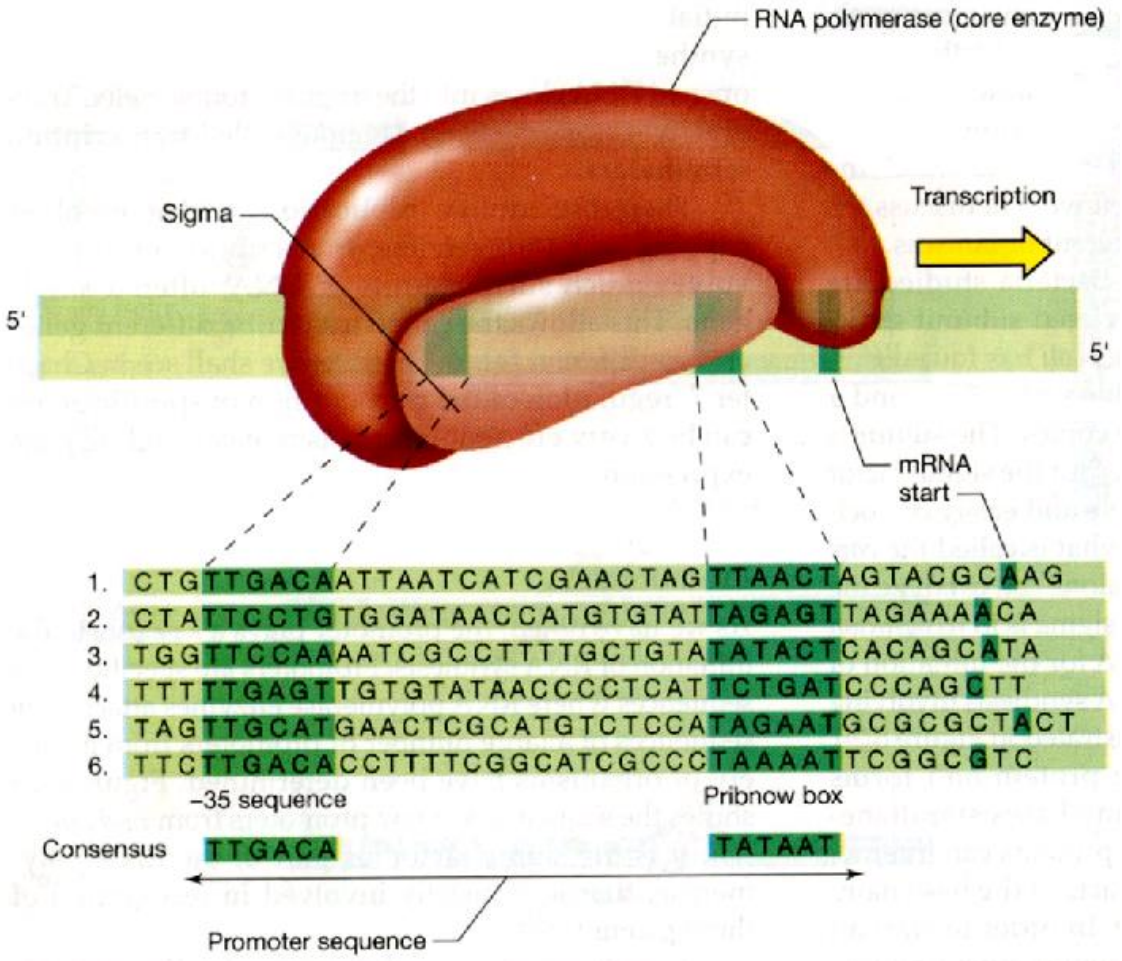
Сигналы промоторов это
- короткие последовательности ДНК, узнаваемые белком;
- расположены перед стартом транскрипции;
- похожие, но не идентичные

Схема инициации транскрипции у прокариот



Источник: РГМ

Initiation of transcription (bacteria)



Источник: **МГ**

	UP-element	-35		-10
TM0373	TTACAAATTCTCATACGACCCC	TTGACA	< 18 bp >	TATAAT
TM1016	TAAAAATTTCATGAAAAATTTC	TTGAAT	< 16 bp >	TTTAAT
TM1272	TTCACATTTTGCATTATACACC	TTGACA	< 17 bp >	TTTAAT
TM1429	CATTGTGATTTTTGTAACTATA	TTGACA	< 17 bp >	TATAAT
TM1667	CAAGTATATCCTAAAAAAATAT	TTGAAA	< 18 bp >	TATAAT
TM1780	GAAAATAACAGTGAAAAAACAC	TTCATA	< 20 bp >	TATAAT
TMt11	AAAAGGGTTATCAGGAAATATC	TTGAAT	< 17 bp >	TAAAAT
TM0032	ATATTAGAATTTGAACTATAAT	TCGAAA	< 18 bp >	CATAAT
TM0477	ACAAAAAAACTTTAGAAAACTC	TTGAAT	< 18 bp >	TATAAT
TM1067	GATTATTTTATACTGAAAGCCC	TTGACC	< 18 bp >	TATTAT
TM1271	GTGATATTTCAACATTAAAATC	TTGACA	< 18 bp >	TATAAT
TMt45	AAGAAGGAAGAAAAATGAAAAC	TTGAAC	< 17 bp >	TATAAT
TM1490	TGAAAATATGCCCAGGAAACGT	TTGACT	< 17 bp >	TAAAAT
		T T		--

Промоторы генов *Termatoga maritima*

Источник: РГМ

РНК-полимераза может использовать разные
sigma-субъединицы.

У E.coli – 7 sigma-субъединиц

Промоторы разных sigma-субъединиц
имеют разные последовательности, но структура:
-35 -10 – одинакова

Экспрессия генов регулируется экспрессией
сигма-факторов (это один из факторов регуляции транскрипции)

Выделяется σ-фактор "домашнего хозяйства", он обслуживает большинство генов, постоянно необходимых бактерии, т.н. генов "домашнего хозяйства".

Вариант а. задания 7 состоит в построении PWM для сигнала посадки превалирующего сигма фактора в геноме бактерии и применении её для поиска промоторов

- Следует набрать несколько десятков промоторных участков, перед стартом транскрипции мРНК (оперона). Например, длиной 100 нукл на кодирующей цепи ДНК.
- С помощью MEME найти подходящие мотивы. Если несколько – выбрать наиболее подходящий с вашей точки зрения.
- Выполнить поиск в других промоторных областях с помощью FIMO; можно попробовать поискать во всем геноме. Описать результат.

b. Сайт посадки рибосомы (прокариоты)

Называется «последовательность Шайн-Далгарно»

Задание 2b: в геноме одной археи или бактерии найти сигнал сайта посадки рибосомы (SD)

Shine-Dalgarno motifs have the consensus sequence GGAGG and can base pair with as many as nine nt in the 3’ terminal sequence of 16S rRNA (ACCUCCUUA in E. coli) referred to as the anti-Shine Dalgarno or ASD (Shine and Dalgarno, 1974).

Saito et al., 2020, eLife

Начала генов *Bacillus subtilis*

dnaN	ACATTATCCGTTAGGAGGATAAAA**ATG**
gyrA	GTGATACTTCAGGGAGGTTTTTTA**ATG**
serS	TCAATAAAAAAAGGAGTGTTTCGC**ATG**
bofA	CAAGCGAAGGAGATGAGAAGATTC**ATG**
csfB	GCTAACTGTACGGAGGTGGAGAAG**ATG**
xpaC	ATAGACACAGGAGTCGATTATCTC**ATG**
metS	ACATTCTGATTAGGAGGTTTCAAG**ATG**
gcaD	AAAAGGGATATTGGAGGCCAATAA**ATG**
spoVC	TATGTGACTAAGGGAGGATTCGCC**ATG**
ftsH	GCTTACTGTGGGAGGAGGTAAGGA**ATG**
pabB	AAAGAAAATAGAGGAATGATACAA**ATG**
rplJ	CAAGAATCTACAGGAGGTGTAACC**ATG**
tufA	AAAGCTCTTAAGGAGGATTTTAGA**ATG**
rpsJ	TGTAGGCGAAAAGGAGGGAAAATA**ATG**
rpoA	CGTTTTGAAGGAGGGTTTTAAGTA**ATG**
rplM	AGATCATTTAGGAGGGGAAATTCA**ATG**

Источник: РГМ

dnaN	ACATTATCCGTTAGGAGGATAAAA**ATG**
gyrA	GTGATACTTCAGGGAGGTTTTTTA**ATG**
serS	TCAATAAAAAAAGGAGTGTTTCGC**ATG**
bofA	CAAGCGAAGGAGATGAGAAGATTC**ATG**
csfB	GCTAACTGTACGGAGGTGGAGAAG**ATG**
xpaC	ATAGACACAGGAGTCGATTATCTC**ATG**
metS	ACATTCTGATTAGGAGGTTTCAAG**ATG**
gcaD	AAAAGGGATATTGGAGGCCAATAA**ATG**
spoVC	TATGTGACTAAGGGAGGATTCGCC**ATG**
ftsH	GCTTACTGTGGGAGGAGGTAAGGA**ATG**
pabB	AAAGAAAATAGAGGAATGATACAA**ATG**
rplJ	CAAGAATCTACAGGAGGTGTAACC**ATG**
tufA	AAAGCTCTTAAGGAGGATTTTAGA**ATG**
rpsJ	TGTAGGCGAAAAGGAGGGAAAATA**ATG**
rpoA	CGTTTTGAAGGAGGGTTTTAAGTA**ATG**
rplM	AGATCATTTAGGAGGGGAAATTCA**ATG**

consensus	a	a	a	g	t	a	t	a	t	a	a	g	**g**	**g**	**a**	**g**	**g**	g	t	t	a	a	t	A	T	G
number	7	6	10	6	6	6	6	5	8	9	6	7	**12**	**12**	**18**	**11**	**10**	6	8	8	8	6	5	9	16	16	16

12 12 18 11 10

Источник: РГМ

Источник: РГМ

Вариант b. задания 7 состоит в построении PWM для сигнала Шайн-Далгарно и применении её для поиска этих сигналов перед другими генами в том же геноме

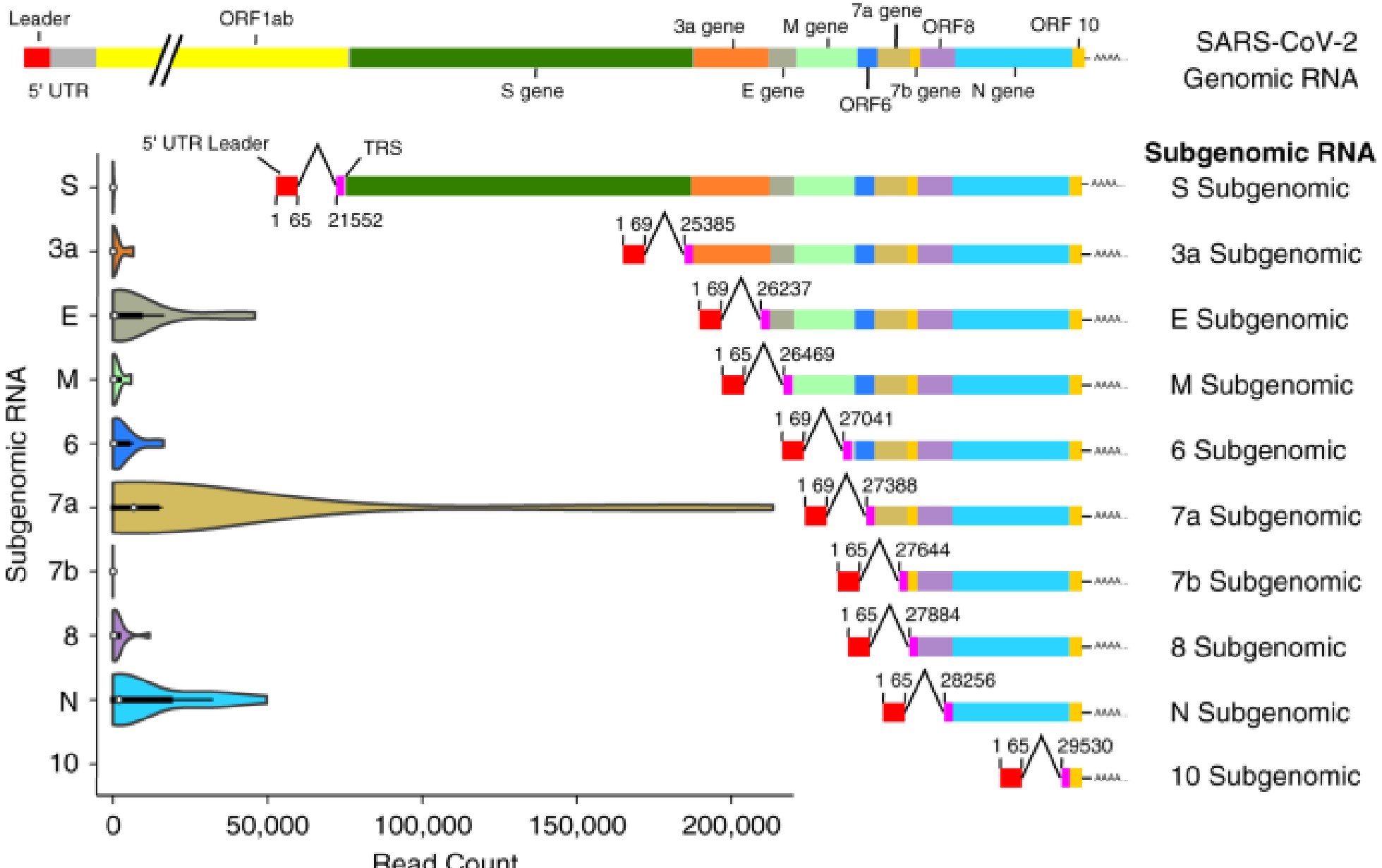
- Следует набрать несколько десятков участков перед стартом первых кодонов генов. Например, длиной 20-30 нукл на кодирующей цепи ДНК.
- С помощью MEME найти подходящие мотивы. Если несколько – выбрать наиболее подходящий с вашей точки зрения.
- Выполнить поиск в других участках перед кодирующими последовательностями с помощью FIMO; можно попробовать поискать во всем геноме. Описать результат.

Проблема с. Трансляция поздних генов коронавирусов

- С РНК вируса транскрибируются мРНК поздних генов. Одна мРНК для одного позднего гена.
- мРНК каждого позднего гена устроена так:
  - Кэпированный 5’ концевой участок мРНК (кончается до ATG кодонов) соединенный с 3’ концевым участком, начинающимся перед ATG кодоном этого позднего гена и до конца
- Эти мРНК называются субгеномными мРНК (sgRNA)

- См. след. слайд

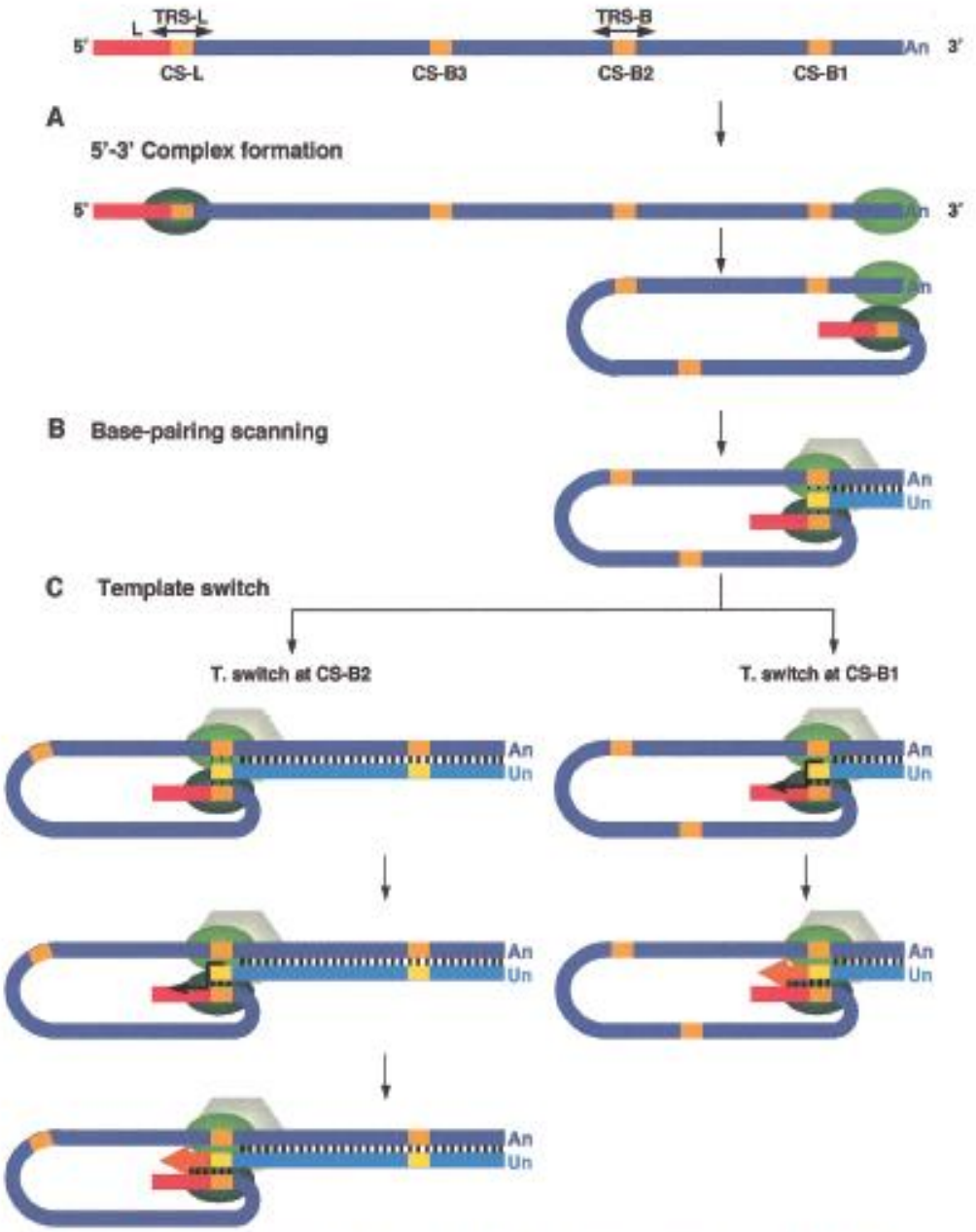
Fig. 1: SARS-CoV-2 genomic and subgenomic RNA structure showing genes and open reading frames (ORF) together with violin plots showing the number of reads per total of 5 million reads in the diagnostic samples mapped to the leader-containing subgenomic RNAs in the fasta file used for mapping.



Транскрипция вирусной РНК

- Вирусная РНК-зависимая РНК-полимераза (RdRP) закодирована в полипротеине (nsp11).
- RdRP по РНК матрице делает комплементарную копию. Из вирусной +RNA получается -RNA; из -RNA получается -(-RNA) = +RNA
- Сигналы разрывной транскрипции направляют перескок RdRP при синтезе -RNA, в результате которого синтезируются -sgRNA.
- -sgRNA является матрицей для RdRP; продукт – субгеномная мРНК (+sgRNA)
- Сигналы разрывной транскрипции называются так: TRS-L в лидере, TRS-B перед каждым поздним геном (TRS=transcription-regulatory sequences)

TRS-L и TRS-B



Лидер –красная полоска

Сигналы TRS – желтые прямоугольники. В них есть общее слово из шести букв (CS)

Мутации в CS влияют на синтез sgmRNA ожидаемым образом

Рисунок - гипотеза, косвенно подтвержденная

Zuniga et al., Journal of Virology, 2004

Сигналы разрывной транскрипции TRS-L, TRS-B; CS

Сигналы TRS-L и все TRS-B имеют высокосходные последовательности. Наиболее похожие их части, часто полностью совпадающие, называются CS (core sequences)

Принято считать, что длина CS – шесть нуклеотидов, TRS включает2-3 нуклеотида с 5’ и 3’ концов CS.

Как все в биологии значения длин не являются мировыми константами

Вариант с. задания 7 состоит в построении PWM для сигнала разрывной транскрипции поздних генов выбранного коронавируса и применении её для поиска этих сигналов в геноме коронавируса

- Выберите коронавирус. Лучше не берите SARS-CoV-2 – надоел.
- Соберите участки перед первым кодоном всех поздних генов и в лидере – перед первым кодоном полипротеина ORFab. Например, длиной 20-30 нукл. . Или лучше так – от предыдущего кодона ATG в любой рамке, до ATG кодона данного позднего гена. Понятно, почему?
- С помощью MEME найдите подходящий мотив. Если несколько – выбрать наиболее подходящий с вашей точки зрения.
- Выполнить поиск в о всем геноме коронавируса с помощью FIMO. Описать результат.

Задание с.: в геноме одного коронавируса найти сигналы TRS (CS)

- У вируса SARS-CoV-2 CS ACGAAC, встречается перед 7-ю из 10-и поздних генов.

КОНЕЦ ПРЕЗЕНТАЦИИ