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

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

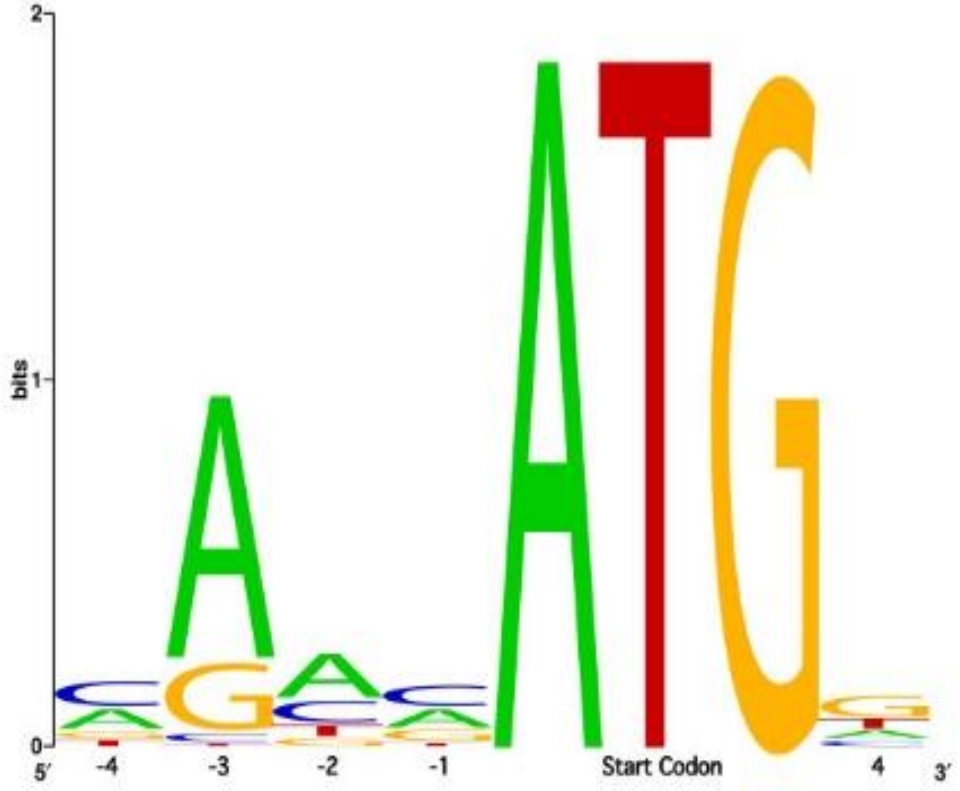
Из формулы следует, что IC(j) – матожидание - веса в колонке при распределении вероятностей букв b заданного частотами букв в колонке

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, $\log_2 N = 2$

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

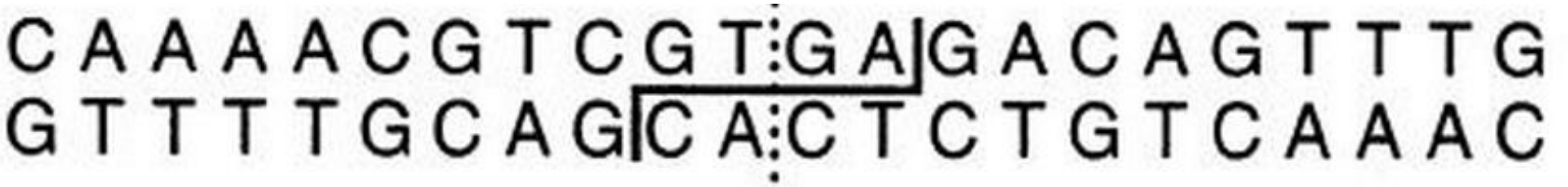
При p(b) = ¼ для всех b получаем

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

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☺)

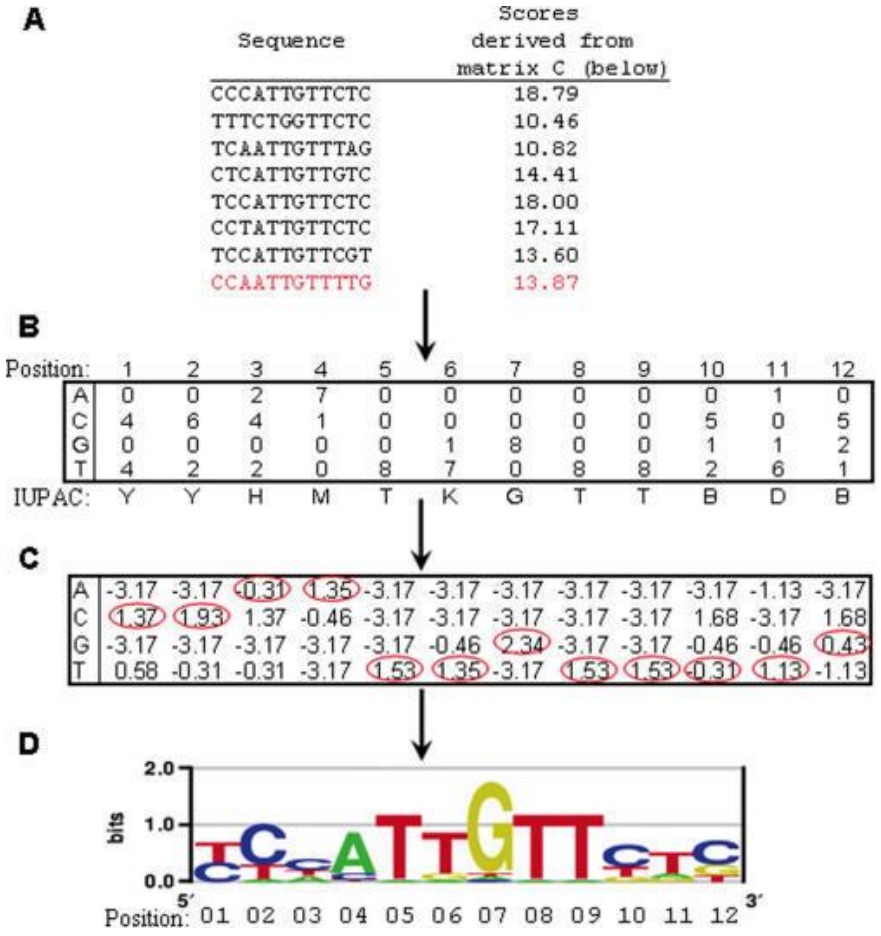
3) Как-то упустил что наши люди – коллеги - тоже сделали детектор мотивов Chipmunk (https://opera.autosome.ru/chipmunk/discovery)

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

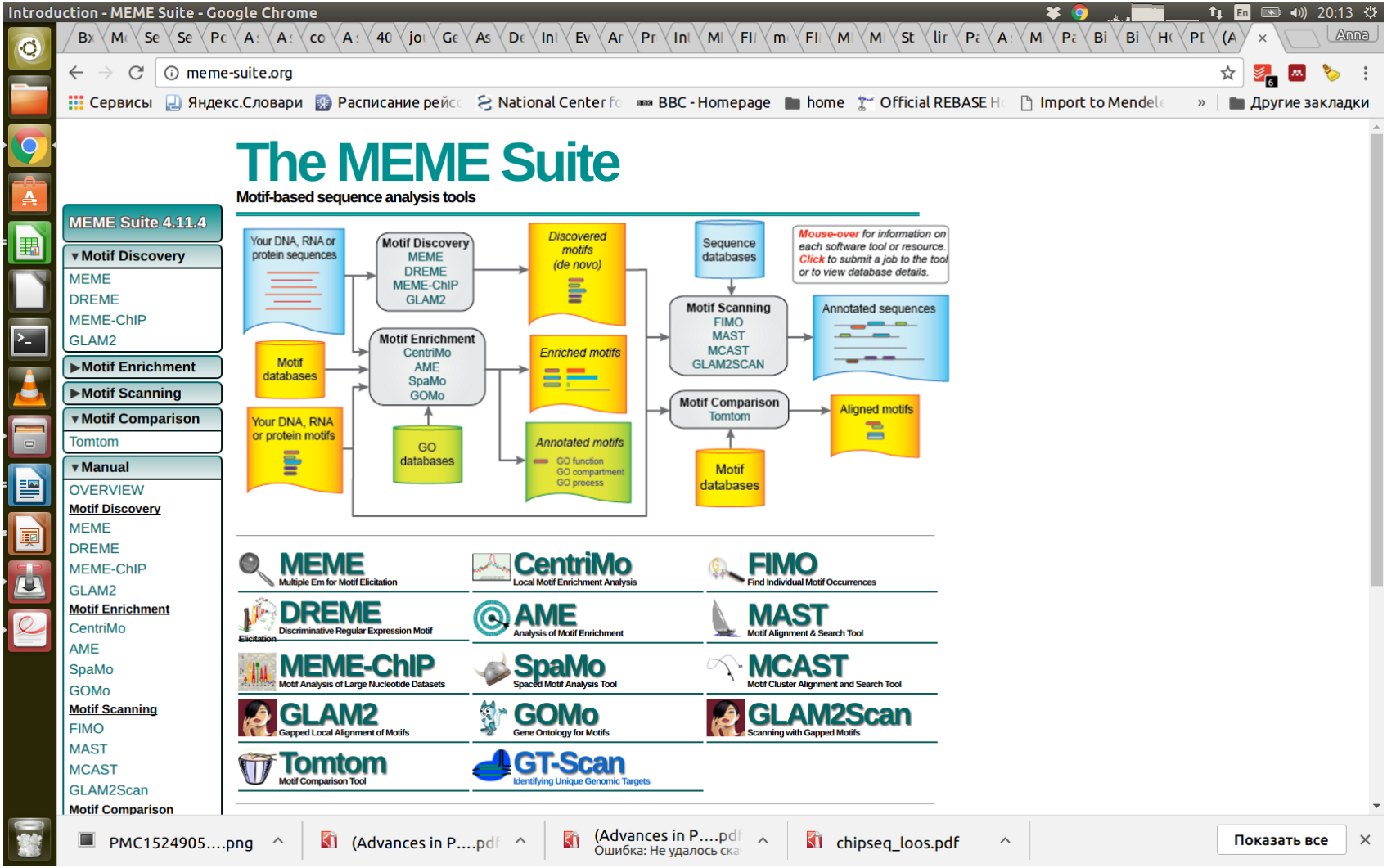


Вес (I(bj) основания b в данной позиции j
I(bj)=f(bj)*log f(bj) — p(b)*log p(b), где f(bj) — частота основания и в позиции j выравнивания, p(b) — фоновая частота основания b

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

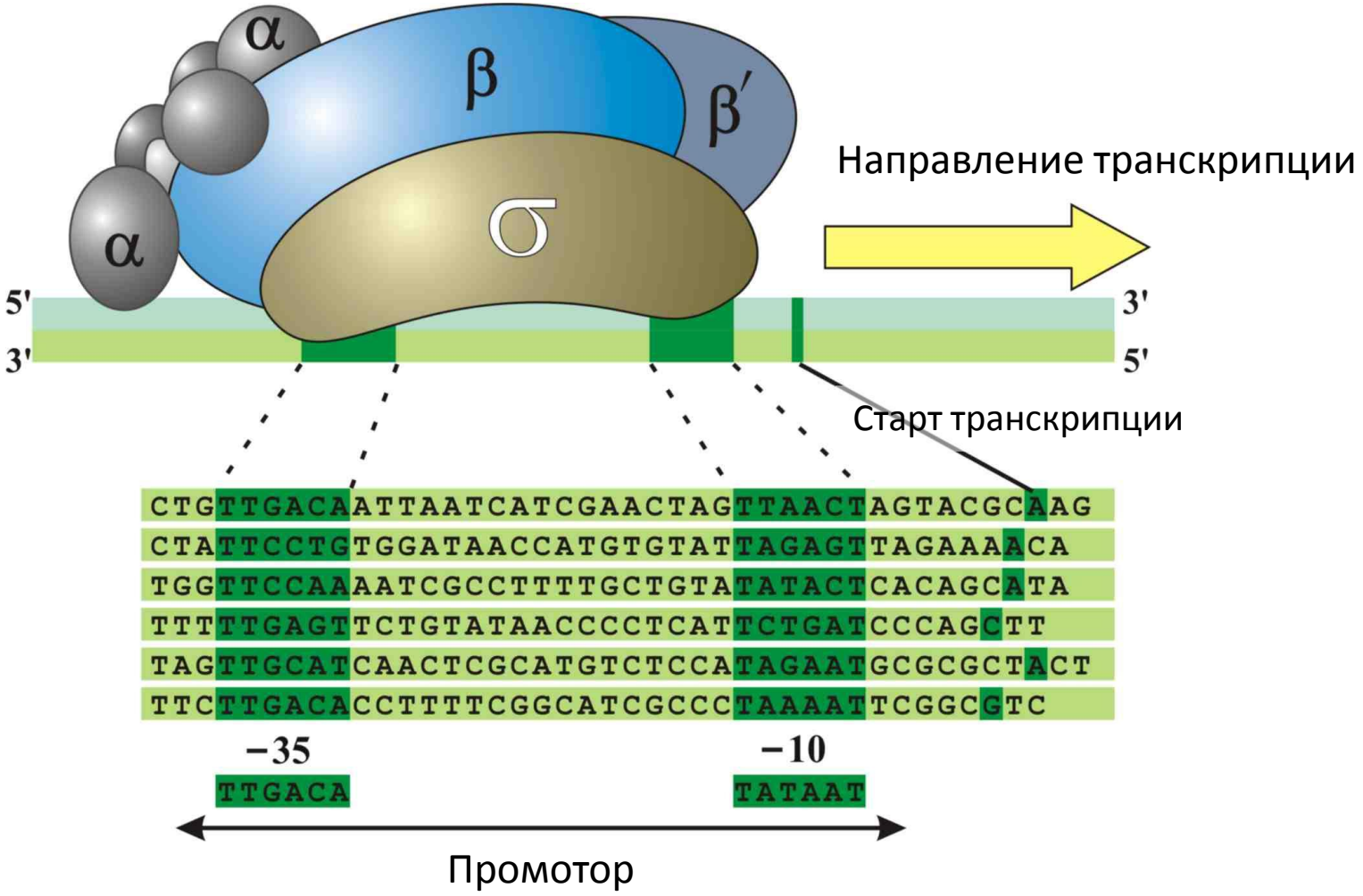
- Промотеры прокариот (инициация транскрипции)
- Сайты посадки рибосомы у прокариот (Shine-Dalgarno = SD последовательности)
- Сигналы разрывной транскрипции у коронавирусов

a. 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

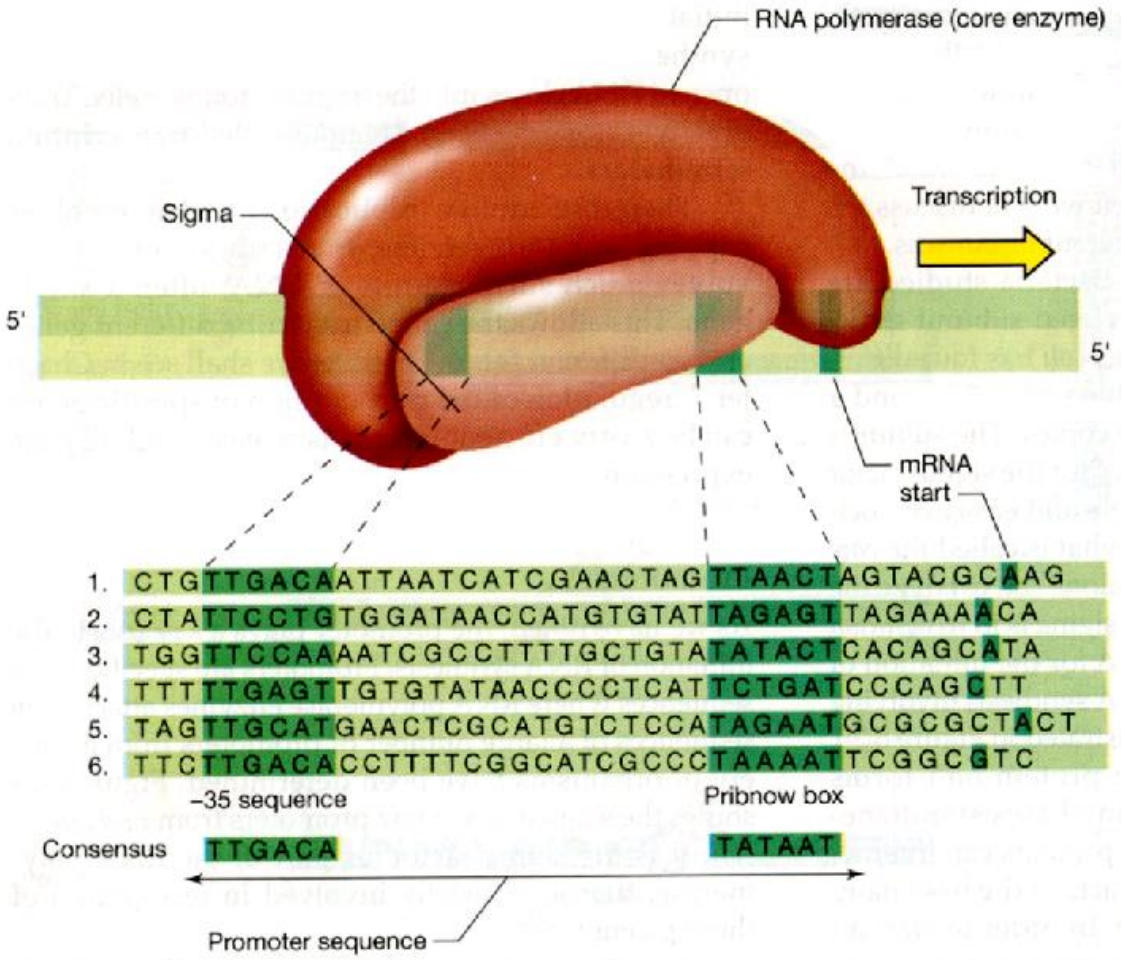
Сигналы промоторов это
- короткие последовательности ДНК, узнаваемые белком;
- расположены перед стартом транскрипции;
- похожие, но не идентичные

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)

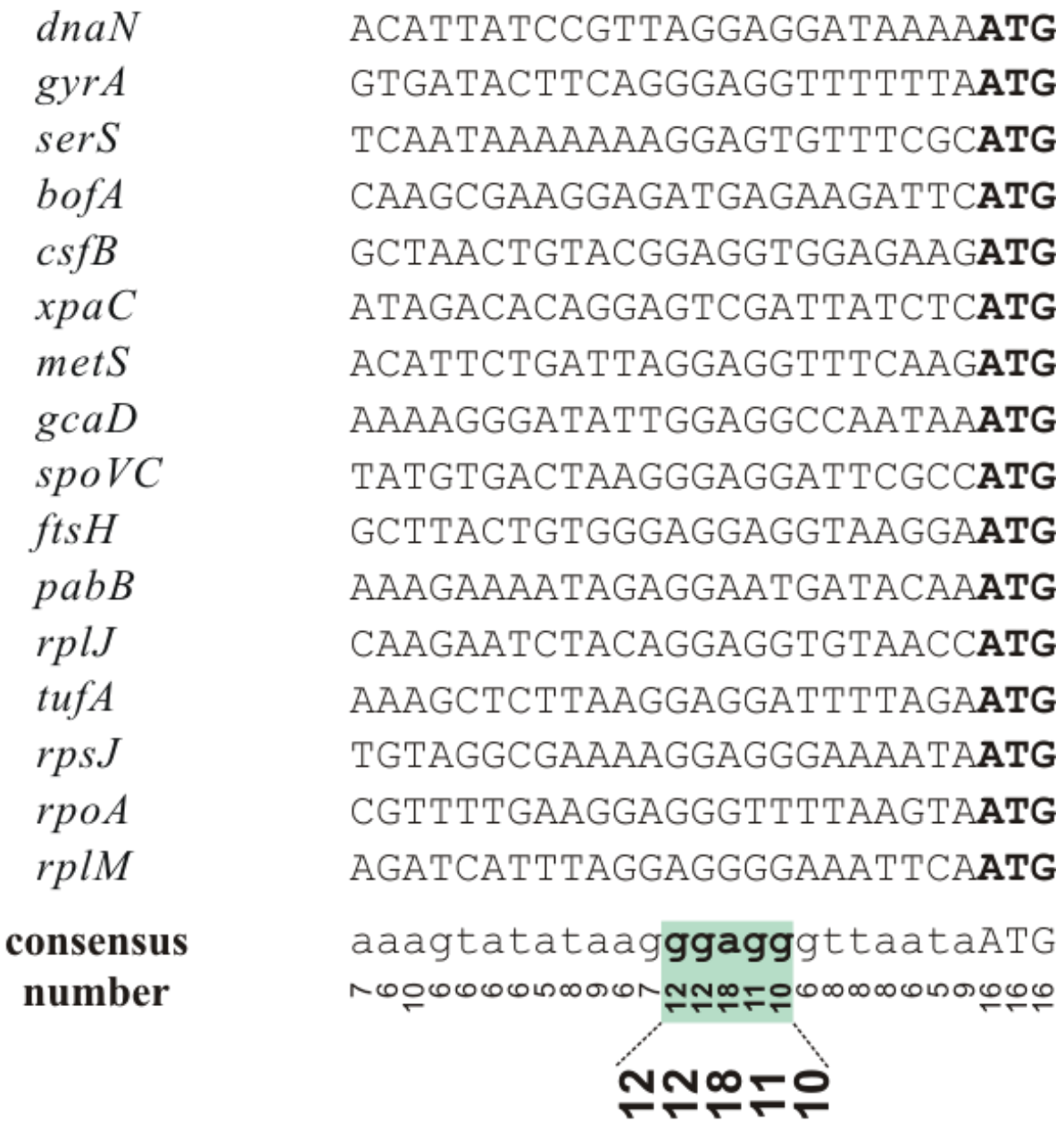
Shine-Dalgarno motifs have the consensus sequence GGAGG and can base pair with as many as nine nt in the 3’ terminal sequence of 16S rRNA (ACCUCCUUA in E. coli) referred to as the anti-Shine Dalgarno or ASD (Shine and Dalgarno, 1974).

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



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

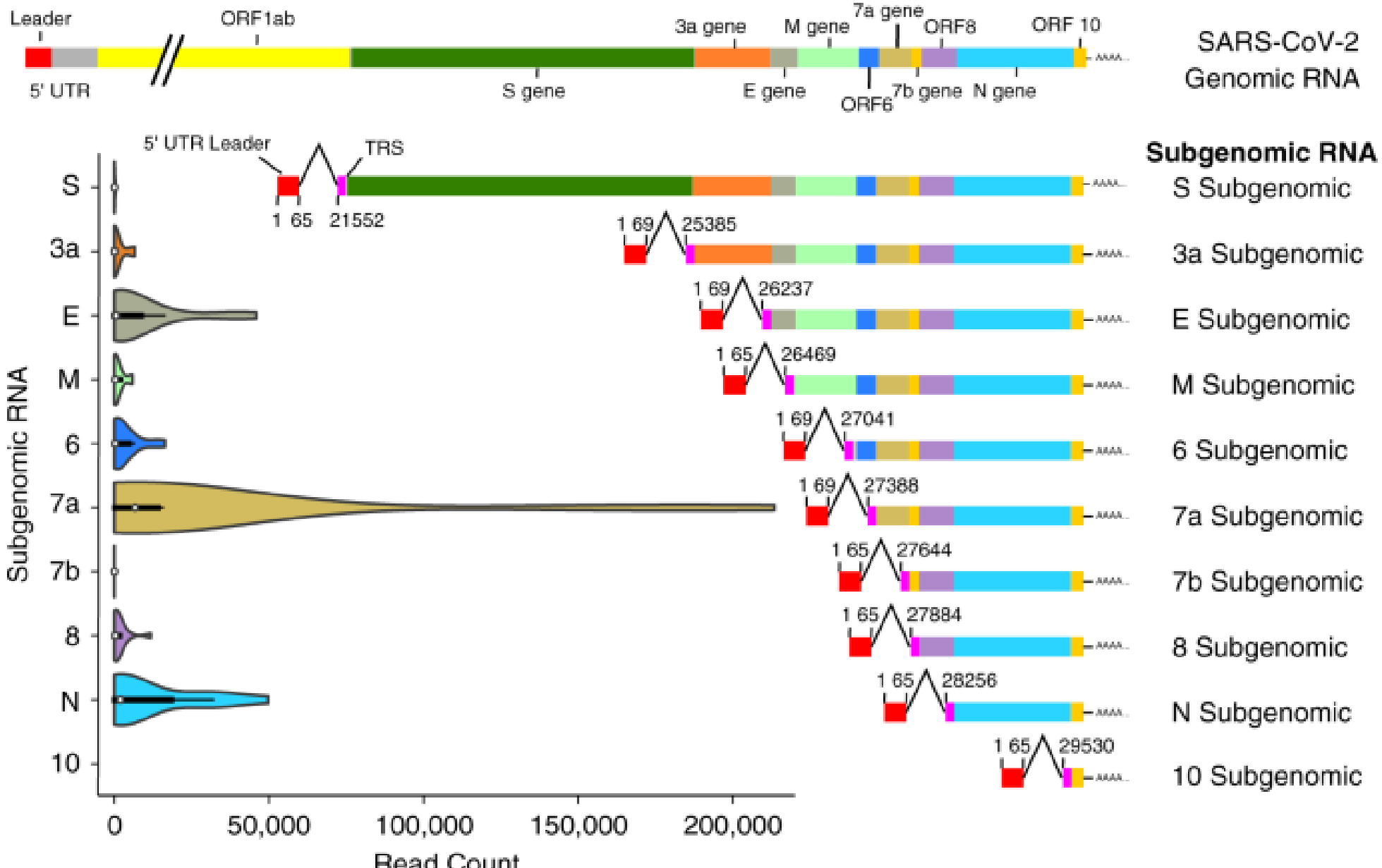
- Следует набрать несколько десятков участков перед стартом первых кодонов генов. Например, длиной 20-30 нукл на кодирующей цепи ДНК.
- С помощью MEME найти подходящие мотивы. Если несколько – выбрать наиболее подходящий с вашей точки зрения.
- Выполнить поиск в других участках перед кодирующими последовательностями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

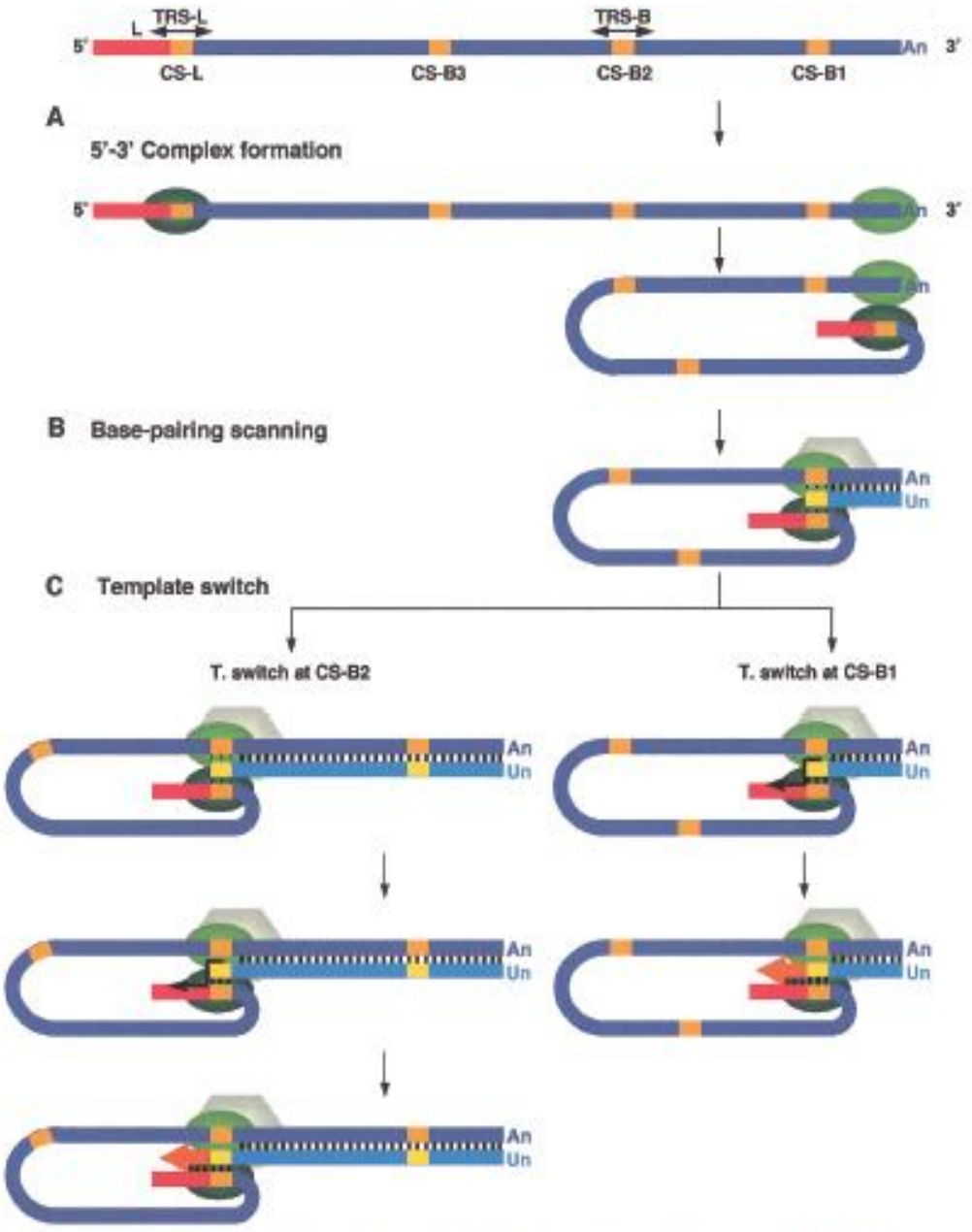
Fig. 1: SARS-CoV-2 genomic and subgenomic RNA structure showing genes and open reading frames (ORF) together with violin plots showing the number of reads per total of 5 million reads in the diagnostic samples mapped to the leader-containing subgenomic RNAs in the fasta file used for mapping.



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

Вопросы о сигнале

1. Как называется
2. В каком процессе используется
3. Конкретный адресат – белок или комплекс белков, реагирующий на сигнал
4. Предназначение, какую реакцию вызывает у адресата
    1. [не запланированные получатели сигнала, если есть]
5. Какой тип сигнала (хим. модификация, вторичная структура, 3D структура, последовательность; составной), описание сигнала
6. Информационное содержание сигнала
7. Эффективность сигнала для адресата
    1. Какая вероятность, что адресат среагирует на сигнал при встрече его

Примеры сильных и слабых сигналов

Сигнал (мотив) читают белки и биологи. И те, и особенно другие, ошибаются

Сигнал эффективный, если он вызывает ожидаемый ответ (какой сигнал, такой ответ).

Сигнал GATC в ДНК адресован эндонуклеазе рестрикции DpnII и двум ДНК метилтрансферазам M1.DpnII и M2.DpnII из стрептококка.

Ответ ДНК метилтрансфераз состоит в навешивании метильной группы на основания А на прямой и обратной цепи.

Ответ DpnII ПРИ ОТСУТСТВИИ метильных групп состоит в расщеплении обеих цепочек ДНК между G и предыдущим основанием.

Сигнал для DpnII эффективный: есть сигнал => есть ответ.

Красный светофор – сильный сигнал там, где машины не едут на красный цвет

Примеры

Сигнал CG (пишут CpG) в геноме человека адресован ДНК метилтрансферазе DNMT3A.

Ответ – метилирование по цитозину в одной цепочке. Однако не все сайты CpG метилированы. Сигнал не эффективный. Наше знание того, как DNMT3A распознаёт CpG которые метилирует недостаточно для понимания происходящего *in vivo.*

Метилирование полуметилированных сайтов – эффективный сигнал для DNTM1. На этом основана эпигенетика. При репликации ДНК DNTM1 восстанавливает метилирование по обеим цепочкам.

КОНЕЦ ПРЕЗЕНТАЦИИ

4. Сужение области поиска мотива

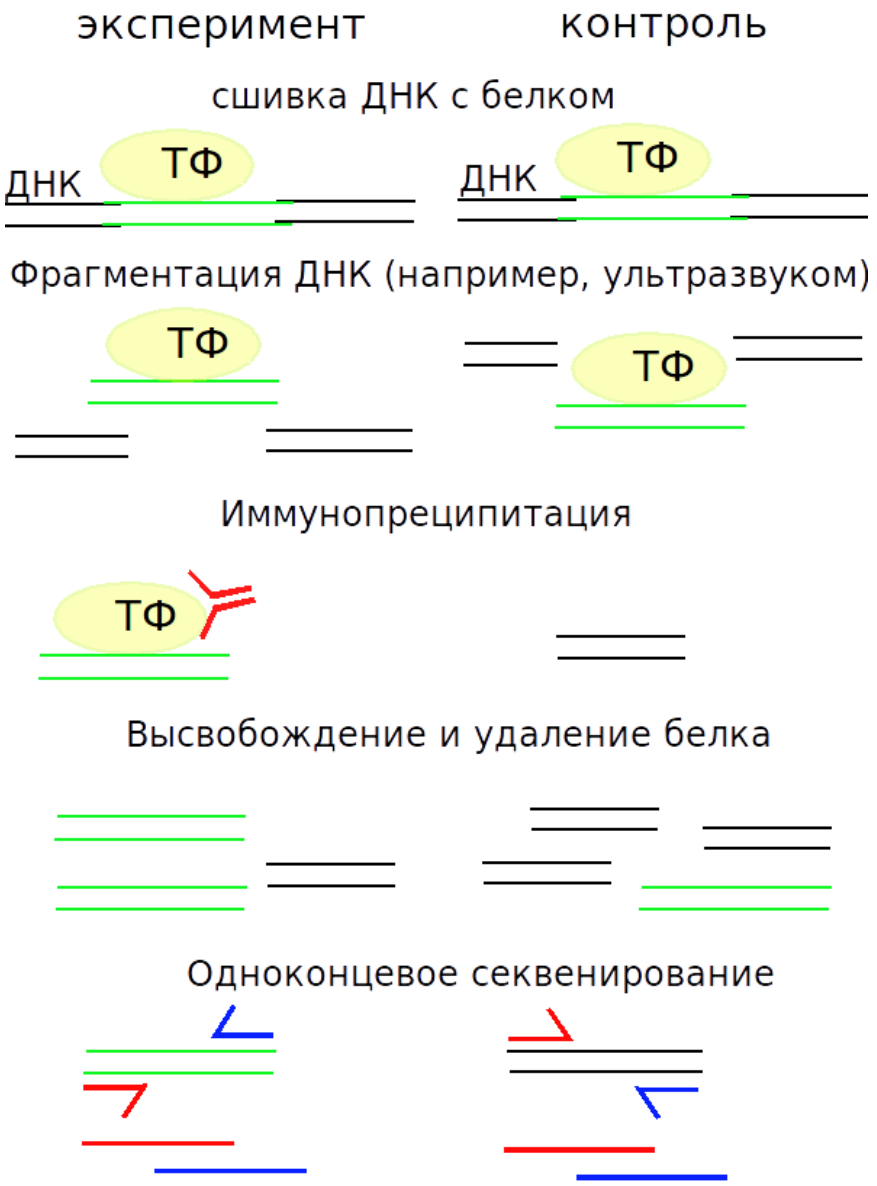
Чем меньше область поиска тем надежнее детектируются мотивы

ChIP-seq

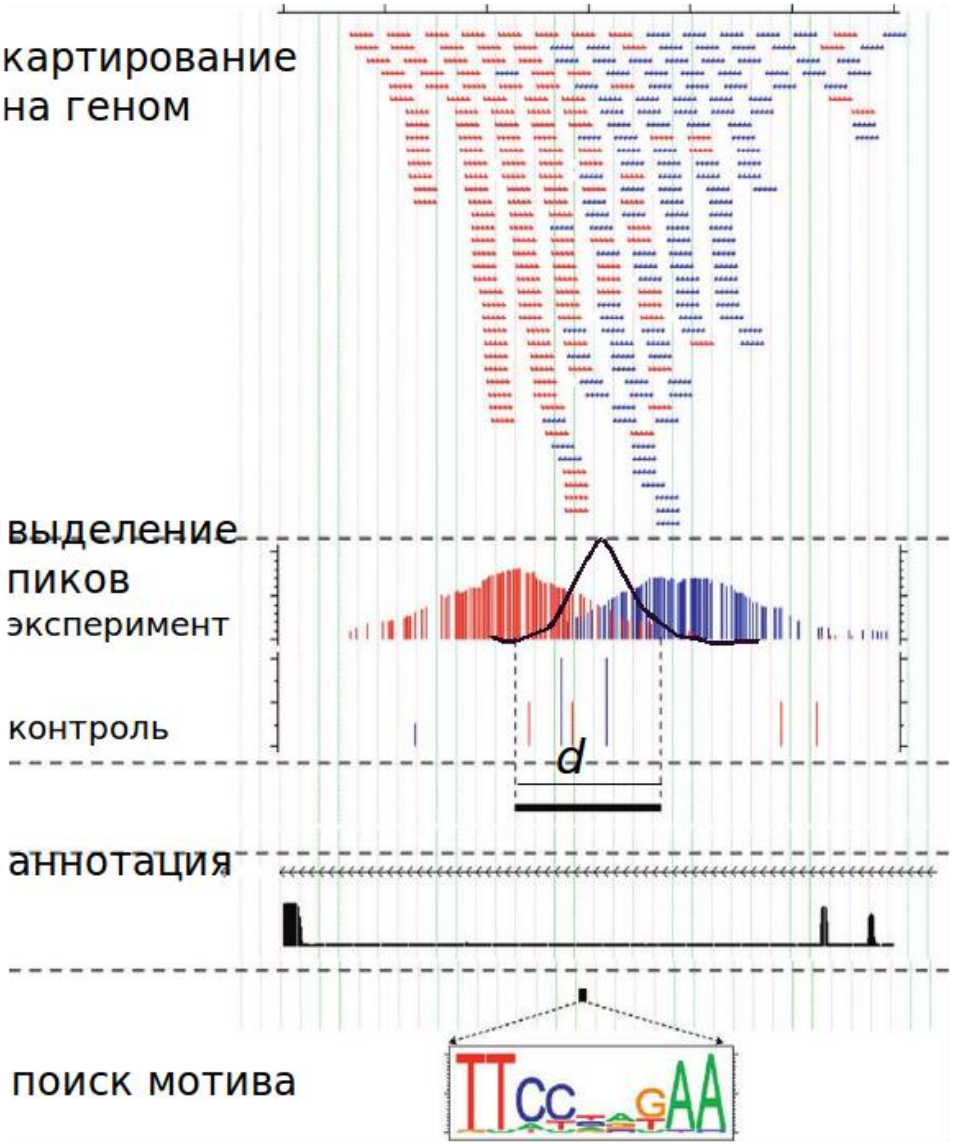
- **Ch**romatin **i**mmuno**p**recipitation (Chip) с последующим высокопроизводительным секвенированием

Эксперимент и анализ данных

Эксперимент



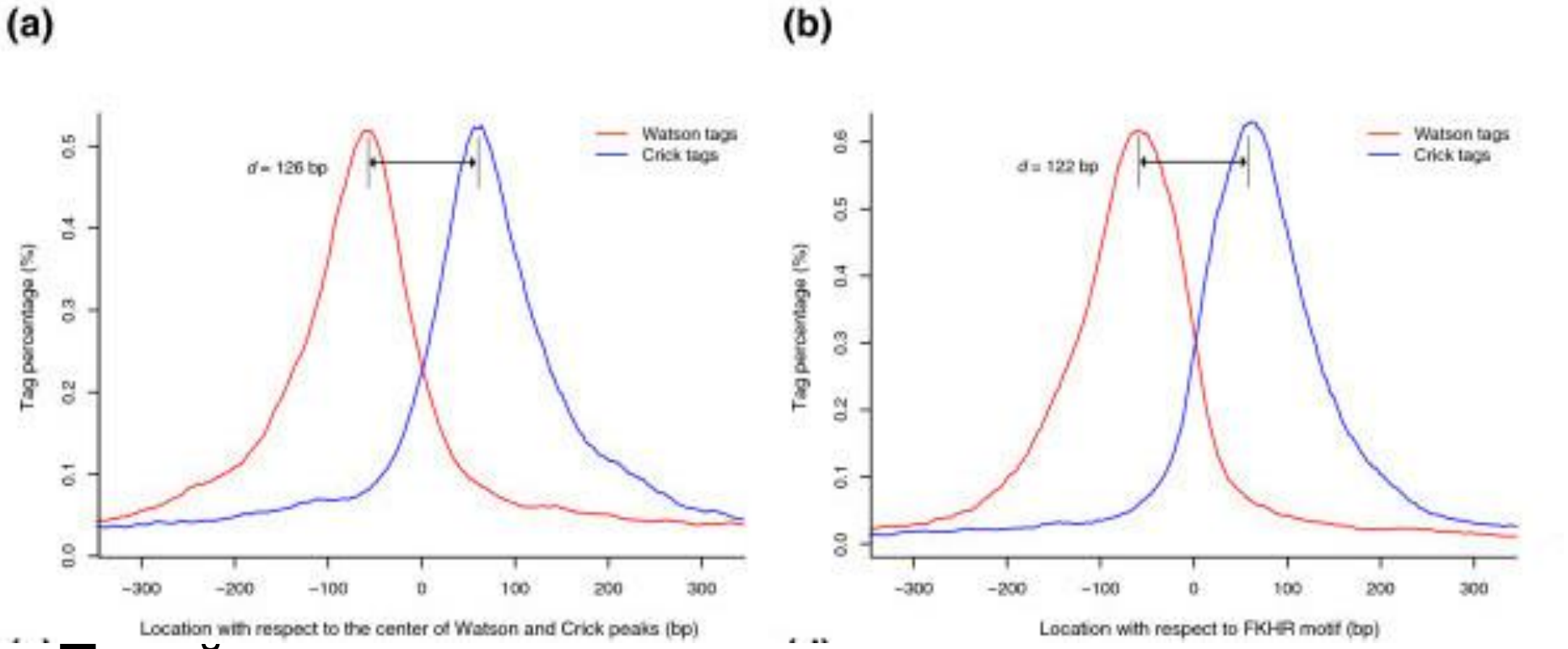
Анализ данных ChIP-seq



Ma W, Wong WH. The analysis of ChIP-Seq data. Methods Enzymol. 2011;497:51-73.

Выбор величины сдвига

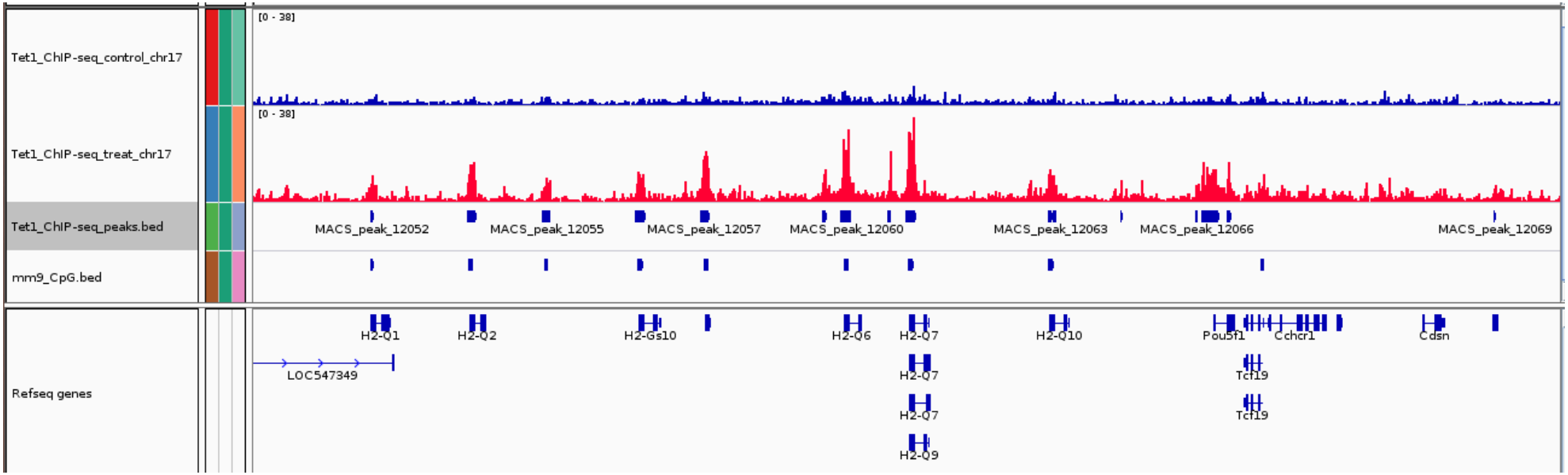
Длина секвенированного фрагмента — 200 п.н.



Такой сдвиг пиков проискодит, если длина анализируемого фрагмента примерно равна длине секвенируемого фрагмента

Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

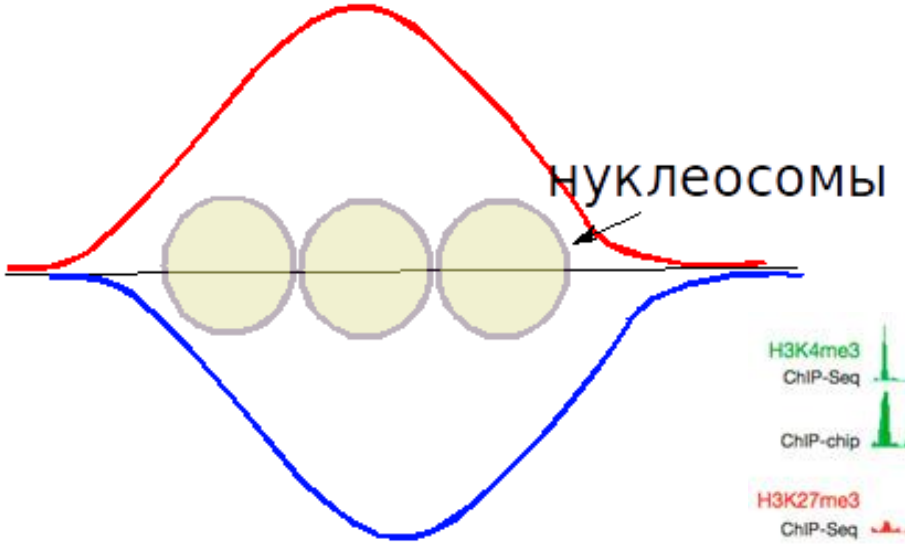
Выбор достоверных пиков

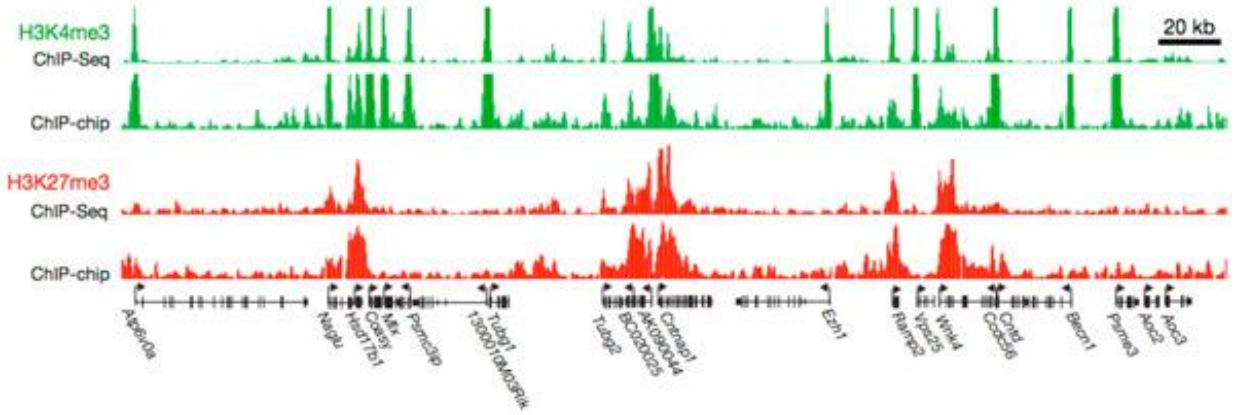


Сравнивают пики в эксперименте и контроле, считают p-value.

http://crazyhottommy.blogspot.ru/2013_12_01_archive.html

Связывание с хроматином

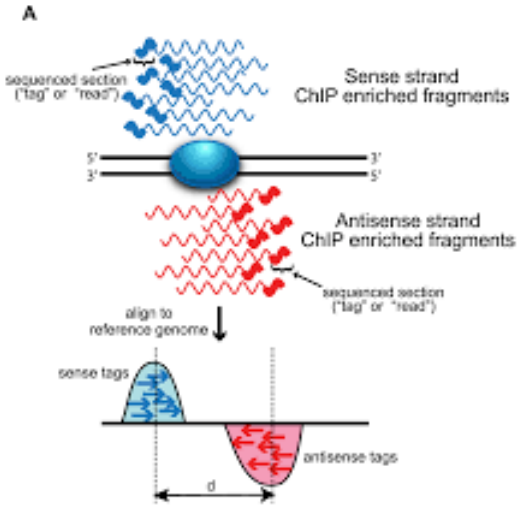


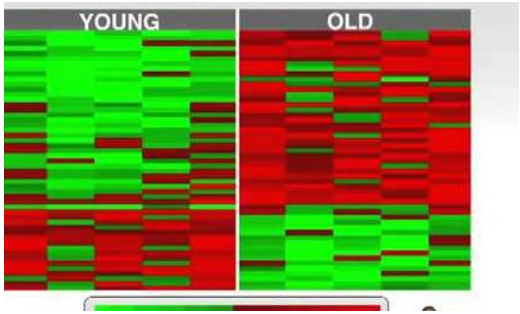


Нет асимметрии пиков

http://compbio.pbworks.com/w/page/16252888/Epigenetic%20Regulation

Примеры



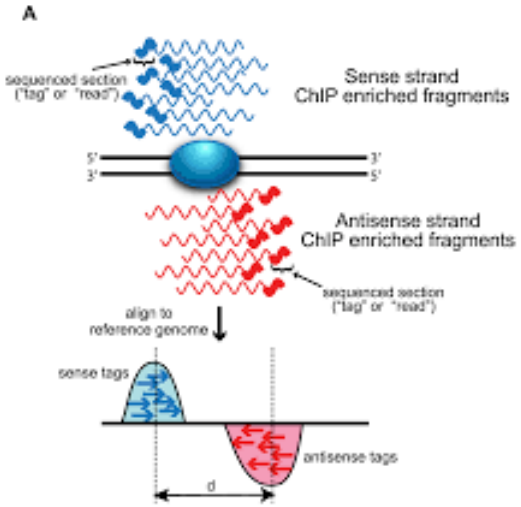


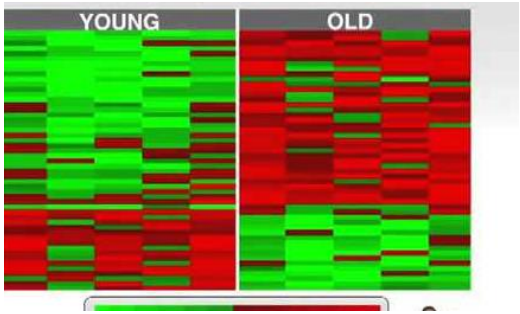
- Данные экспериментов ChiP-Seq
- Upstream области коэкспрессирующихся генов

```
CCTACGCAAACGTTTTCTTTTT
GTCTCGCAAACGTTTGCTTTCC
CACACGCAAACGTTTTCGTTTA
TCCACGCAAACGGTTTCGTCAG
GCCACGCAACCGTTTTCCTTGC
GATACGCAAACGTGTGCGTCTG
CCGACGCAATCGGTTACCTTGA
GTTGCGCAAACGTTTTCGTTAC
```

А это выравнивание –
то, что нужно найти в более длинных последовательностях

Примеры





- Данные экспериментов ChiP-Seq
- Upstream области коэкспрессирующихся генов

```
CCTACGCAAACGTTTTCTTTTT
GTCTCGCAAACGTTTGCTTTCC
CACACGCAAACGTTTTCGTTTA
TCCACGCAAACGGTTTCGTCAG
GCCACGCAACCGTTTTCCTTGC
GATACGCAAACGTGTGCGTCTG
CCGACGCAATCGGTTACCTTGA
GTTGCGCAAACGTTTTCGTTAC
```

А это выравнивание –
то, что нужно найти в более длинных последовательностях

Chip-seq

эксперимент, позволяющий находить сайты связывания конкретного белка с ДНК с помощью NGS

Поиск точных последовательностей или паттернов

Ищем подпоследовательности или паттерны,которые:

- часто встречаются в наборе последовательностей, связывающихся с белком, и
- не встречаются в контрольном наборе
- Недостатки — ищет точное совпадение, в то время как большинство сайтов связывания ТФ устроены более сложно
- Метод применим, например, для сайтов рестрикции систем рестрикции-модификации, которые обычно определены однозначно:

GATC
CCNGG
GGWCC

Теория

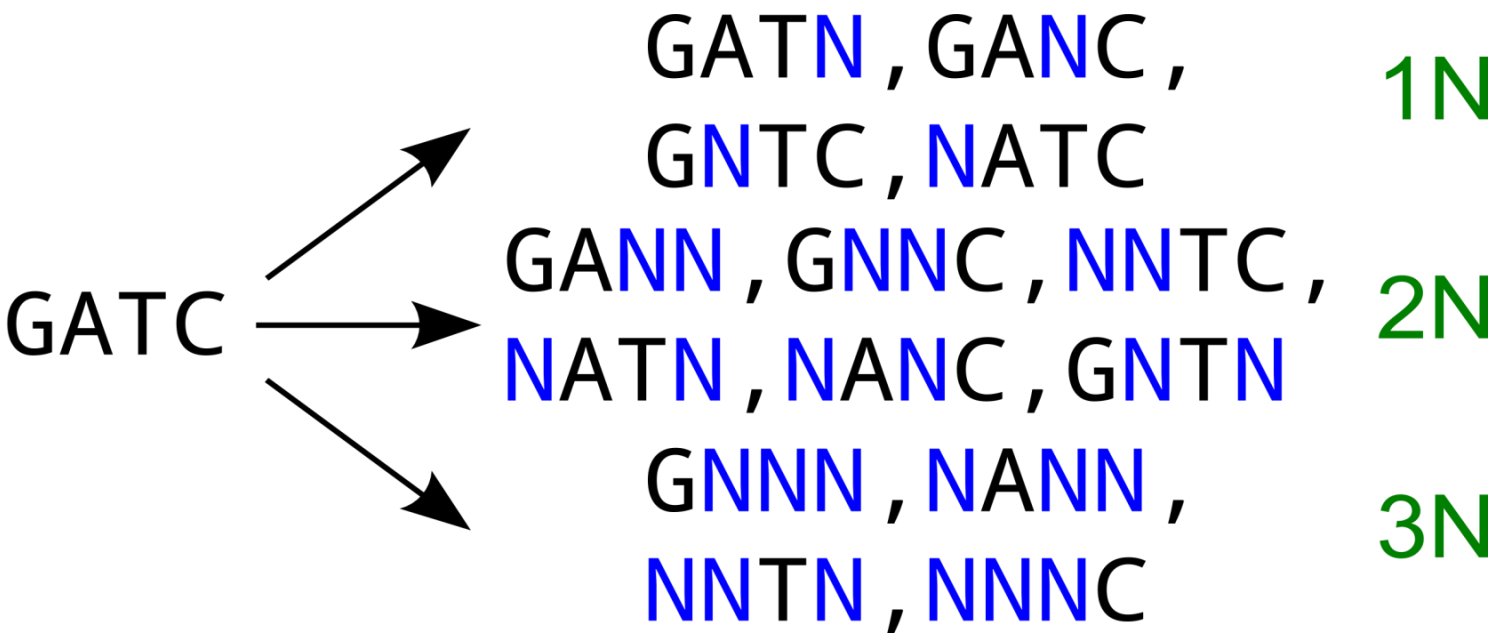
Как учесть зависимость позиций сигнала?

Недостатки PWM и других подходов с весом выравнивания
Предположение о независимости букв в колонках. (Есть работы о том, что часто это близко к реальности)
Учет колонок даже тех, в которых фактически нет значимого сигнала (есть работа, в которой предлагается способ уменьшить их роль) пример
Предложение:
при поиске *de novo* найти слова, в т.ч. вырожденные, которые встречаются чаще, чем ожидалось бы в соответствии со статистической моделью
Если удается найти правильные слова, то придумать правило как их использовать для поиска

Site	A	C	G	U	χ^2	P	A	C	G	U
1	83	30	49	84	10.10	0.0177	0.0525	−0.6332	−0.0260	0.3143
2	103	44	46	53	10.04	0.0182	0.3613	−0.0878	−0.1162	−0.3434
3	121	36	38	51	30.01	0.0000	0.5920	−0.3739	−0.3886	−0.3981
4	122	38	33	53	32.16	0.0000	0.6038	−0.2969	−0.5893	−0.3434
5	81	40	81	44	28.33	0.0000	0.0177	−0.2238	0.6933	−0.6081
6	0	1	245	0	948.34	0.0000	−6.6464	−5.0056	**2.2841**	−6.6469
7	0	9	0	237	582.23	0.0000	−6.6464	−2.3190	−6.6480	**1.8032**
8	239	1	2	4	462.46	0.0000	**1.5693**	−5.0056	−4.3320	−3.8633
9	16	24	1	205	387.81	0.0000	−2.2655	−0.9496	−5.0680	**1.5946**
10	2	0	243	1	928.96	0.0000	−4.8476	−6.6483	**2.2723**	−5.3416
11	9	7	2	228	521.06	0.0000	−3.0427	−2.6612	−4.3320	**1.7475**
12	87	15	34	110	53.66	0.0000	0.1198	−1.6111	−0.5468	0.7006
13	84	49	30	83	11.71	0.0085	0.0696	0.0659	−0.7246	0.2971
14	111	39	33	63	19.09	0.0003	0.4684	−0.2599	−0.5893	−0.0969
15	106	38	31	71	17.24	0.0006	0.4024	−0.2969	−0.6781	0.0738
16	92	30	40	84	13.69	0.0034	0.1997	−0.6332	−0.3155	0.3143
17	80	38	36	92	14.32	0.0025	−0.0001	−0.2969	−0.4655	0.4445

Оценка контраста сайта

$$Kr = \frac{Fo}{Ke} \qquad Ke = \frac{\prod Fo(odd\,N)}{\prod Fo(even\,N)}$$



Сайт посадки рибосомы (прокариоты)

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**

Источник: РГМ

Источник: РГМ

конец

Shannon coined the name "bit" as a unit measure of information in his 1948 paper "The Mathematical Theory of Communication," as a short form of a "binary digit."

Информация и энтропия сигнала

- Изучаем сигналы, которые кодируются последовательностью букв (пример, нуклеотидов)
- Энтропия H – мера неопределенности сигнала
- Информация I состоит в уменьшении неопределенности.
- Неопределенность появления случайно одной буквы
  - H = sum pi*log2 pi, I in
- Пусть появилась буква A
- Ищем сигнал, который есть последовательность букв.
- Очередной нуклеотид
- Белок сканирует РНК. Очередной нуклеотид A
- Начнем с одной , например, слово из 3х нуклеотидов.
- Число таких слов N = 4^3 = 2^6. Значит закодировать все слова можно 6-ю битами (последовательностями из 6 нулей или единиц)
  - Смотрим на первую букву.

The more surprised we are, the more information we gain Similarly, if we receive a message solely consisting of 'A's, as in AAAAAAA..., and nothing more, that would not be extremely surprising either. Alternatively, we could make it a long chain consisting of only ones or only zeros. Either way, there's no surprise reading along either of the messages; Thus, the amount of information in these messages is zero.

Shannon coined the name "bit" as a unit measure of information in his 1948 paper "The Mathematical Theory of Communication," as a short form of a "binary digit."

$$I = H_{Before} - H_{After}$$

Позиционно-весовая матрица выравнивания (без гэпов)

Частоты букв в колонках. Пример.

	a	m	G	A	A	A	a	C	G	k	T	T	w	C	w	T
j																
A	10	2	0	1	13	13	10	0	0	1	0	0	4	0	1	0
C	0	8	0	12	0	0	1	13	0	0	0	0	1	11	4	2
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

Позиционная весовая матрица PWM. (По РГМ)

A	1,6	-0,2	-0,8	0,0	2,5	2,5	1,6	-0,8	-0,8	-0,4	-1,1	-1,1	0,3	-1,3	-0,8	-1,2
C	-1,4	1,0	-0,8	2,1	-0,8	-0,8	-0,3	2,5	-0,8	-1,5	-1,1	-1,1	-0,8	1,8	0,3	0,4
G	0,2	-0,2	2,5	-1,1	-0,8	-0,8	-1,4	-0,8	2,5	0,7	0,0	0,0	0,0	-0,2	0,5	-1,2
T	-0,3	-0,7	-0,8	-1,1	-0,8	-0,8	0,2	-0,8	-0,8	1,3	2,1	2,1	0,5	-0,2	0,0	1,9

Упражнение. Посчитайте вес последовательности

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

относительно этой PWM.

Обоснование PWM

Для каждой буквы b_j из слова S мы можем оценить вероятность ее увидеть в столбце j частотой $F(b_j, j)$

Произведение P(S|выравнивание) = $\Pi_j F(b_j, j)$ по всем позициям оценивает вероятность слова S быть похожим на выравнивание. Чем больше частоты букв каждой позиции, тем выше вероятность слова S

Нужен контроль того, насколько эта вероятность велика. Контролем будем считать вероятность увидеть слово S в геноме. И считать ее будем как произведение частот букв из S в геноме: $P(S) = \Pi_j p_{b_j}$

Решение про сходство S с выравниванием принимается просто: что вероятнее, тому и верим:

Если P(S|выравнивание) >> P(S), то S похоже на выравнивание

Если P(S) << P(S|выравнивание) , то S больше похоже на случайное слово из генома на выравнивание

Мерой сходства служит отношение правдоподобия P(S|выравнивание)/P(S). На практике всегда переходят к логарифмам – чтобы не умножать, а складывать:

W(S, выравнивание) = ln (P(S|выравнивание)/P(S))

Напишите выражение для W как сумму по позициям

РГМ* придумали свою формулу для W(b,j)

$$W(b,j) = \ln [N(b,j)+0{,}5] - 0{,}25 \sum_i \ln [N(i,j)+0{,}5]$$

РГМ вместо базовой вероятности буквы p(b) в геноме использует такую штуку :

p(b,j) = M({F(A, j), F(T,j), F(G,j), F(C, j) })

где M(…) обозначает среднее геометрическое четырех частот

У них базовая частота меняется от колонки к колонки, но не зависит от буквы!

КРУТЫЕ!

Имею документ, подписанный лично А.А.Мироновым, что формула верна ☺ ААл

*Равчеев, Гельфанд, Миронов

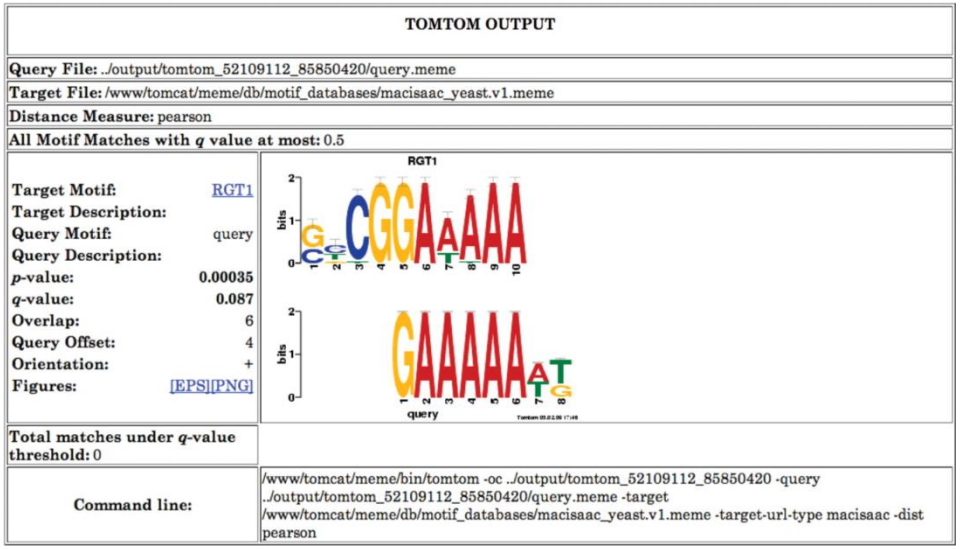
Обозначения, $S = (b_1. b_2,\ldots, b_m)$ – посл., выровненная с выравниванием

N	Число последовательностей	f(b,j)	= N(b,j)/N	Частота буквы b в колонке j
i	i-я последовательность	F(b,j)	= $(N(b,j)+\varepsilon_b)/(N+\varepsilon)$ $\varepsilon = \Sigma_b\ \varepsilon_b$	Частота буквы b в колонке j с учетом псевдокаунтов
m	Число колонок	I(b,j)	= $F(b,j)\cdot\log_2[F(b,j)/p_b]$	Информационное содержание буквы b в колонке j
j	j-я колонка	I(j)	= $\Sigma_b\ I(b,j)$	Информационное содержание колонки j
N(b,j)	Число букв b в колонке j	I	= $\Sigma_j\ I(j)$	Информационное содержание выравнивания сигналов
S	Последовательность длины m	W(b,j)	= ln $F(b,j)/p_b$	Вес буквы b относительно колонки j
b	Буква A,T,G или C	W(S)	= $\Sigma_j\ W(b_j,j)$	Вес последовательности относительно выравнивания

Tomtom output

From: **MEME Suite: tools for motif discovery and searching**

Nucleic Acids Res. 2009;37(suppl_2):W202-W208. doi:10.1093/nar/gkp335



TOMTOM OUTPUT

Query File: ../output/tomtom_52109112_85850420/query.meme

Target File: /www/tomcat/meme/db/motif_databases/macisaac_yeast.v1.meme

Distance Measure: pearson

All Motif Matches with *q* value at most: 0.5

Target Motif:	RGT1
Target Description:	
Query Motif:	query
Query Description:	
***p*-value:**	0.00035
***q*-value:**	0.087
Overlap:	6
Query Offset:	4
Orientation:	+
Figures:	[EPS][PNG]

Total matches under *q*-value threshold: 0

Command line: /www/tomcat/meme/bin/tomtom -oc ../output/tomtom_52109112_85850420 -query ../output/tomtom_52109112_85850420/query.meme -target /www/tomcat/meme/db/motif_databases/macisaac_yeast.v1.meme -target-url-type macisaac -dist pearson

Figure Legend:

The figure shows the Tomtom output from searching a single DNA motif against a collection of yeast transcription factor binding site motifs identified via ChIP-seq (9). Tomtom shows that the query motif closely resembles the binding motif for transcription factor RGT1.

OXFORD UNIVERSITY PRESS

FIMO output

From: **MEME Suite: tools for motif discovery and searching**

Nucleic Acids Res. 2009;37(suppl_2):W202-W208. doi:10.1093/nar/gkp335

Sequence Analysis with fimo

Pattern Name	Sequence Name	Start	Stop	Score	p-value	q-value	Matched Sequence
1	NP_418484.4|yjcB	281	298	21.2367	5.3e-09	0.00758	AATTGTGATATAGTTCAC
1	NP_418485.1|yjcC	149	132	21.2367	5.3e-09	0.00758	AATTGTGATATAGTTCAC
1	NP_418031.1|yiaJ	175	158	19.8034	3.86e-08	0.0173	AAGTGTGCCGTAGTTCAC
1	NP_418032.1|yiaK	26	43	19.8034	3.86e-08	0.0173	AAGTGTGCCGTAGTTCAC
1	NP_418535.1|proP	37	54	19.7078	4.3e-08	0.0173	ATGTGTGAAGTTGATCAC
1	NP_414666.1|gcd	126	143	19.6123	4.85e-08	0.0173	AATTGTGATGACGATCAC
1	NP_414667.4|hpt	80	63	19.6123	4.85e-08	0.0173	AATTGTGATGACGATCAC

Figure Legend:

Полезные ссылки

- Анализ данных ChiPSeq https://books.google.ru/books?hl=ru&lr=&id=YC2K_v1mficC&oi=fnd&pg=PA135&dq=makeev+vsevolod&ots=uSo84sL8A6&sig=xOTJH2RcPWhsjL7cBELQqkCkyfI&redir_esc=y#v=onepage&q=makeev%20vsevolod&f=true
- Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research. 2006;34(Web Server issue):W369-W373. doi:10.1093/nar/gkl198.
- Tran NTL, Huang C-H. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. Biology Direct. 2014;9:4. doi:10.1186/1745-6150-9-4.
- Kulakovskiy IV, Makeev VJ. DNA sequence motif: a jack of all trades for ChIP-Seq data. Adv Protein Chem Struct Biol. 2013;91:135-71.

Сравнение двух PWM

- Как в MEME
- Vorontsov et co-authors https://almob.biomedcentral.com/articles/10.1186/1748-7188-8-23

Некоторые термины

- **Сигнал** – то, что узнает белок. Иногда и он тоже может ошибаться

  .
- **Мотив** – описание сигнала, придуманное человеком;
  - обычно, основано на последовательности
  - позволяет предсказывать в геноме сигналы – сайты, с которыми связывается изучаемый белок
  - Идеал: предсказываются все сигналы и не предсказывается ни одного сигнала в геноме там, где их нет – НЕДОСТИЖИМ ☹

RRvnGAAASKGAAASK GAAAGTGAAA

Виды описаний сигналов (мотивов)

Пусть есть выравнивание последовательностей известных или предполагаемых сигналов

```
CCTACGCAAACGT
GTCTCGCAAACGT
GCCACGCAACCGT
```

Консенсус — способ отображения мотива, в котором в каждой позиции указывается самое частое основание/аминокислота

```
GCCACGCAAACGT
```

- **Паттерн** – указание допустимых букв в каждой позиции

```
SMMWCGCAAMCGT или
[GC][TC]…[AC]CGT
```

- **LOGO** – в каждой позиции каждая буква изображаются прямоугольником высоты, равной *её информационному содержанию* (ждите!)

- **Позиционная весовая матрица PWM** – вес каждой буквы в каждой позиции (ждите)
- **Профиль** – обобщение PWM, позволяющее учитывать делеции и вставки (позже); другая математика