Сигналы и мотивы -2

De novo поиск сигналов в последовательностях

КР по Л1

Идёт в зачёт коллоквиума

План

- IC и энтропия – повторение
- ТФ
- Технология поиска de novo MEME
  - НК
  - Белки (?)
- Поиск по PWM – Fimo
- Упражнение
- Задание

PWM, энтропия и информационное содержание

повторение

PWM

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

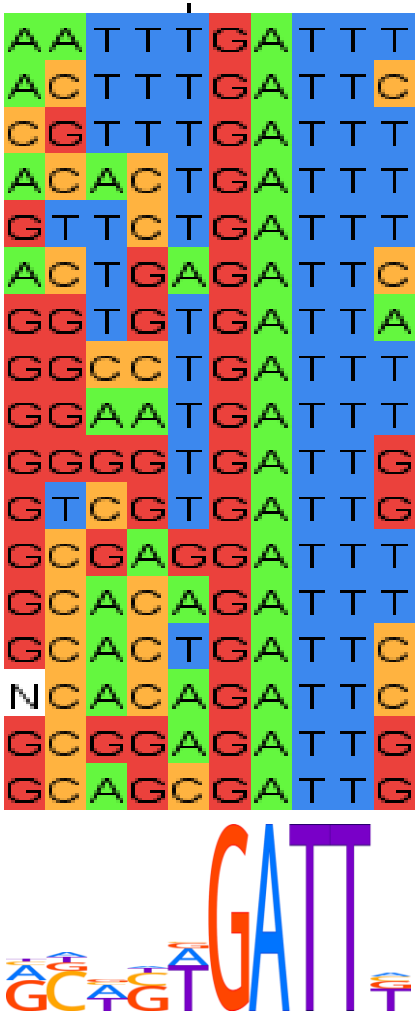
Вес за букву G в позиции 15 этого сигнала равен
$w(G,15) = \ln(0.38/0.35) = 0.1$

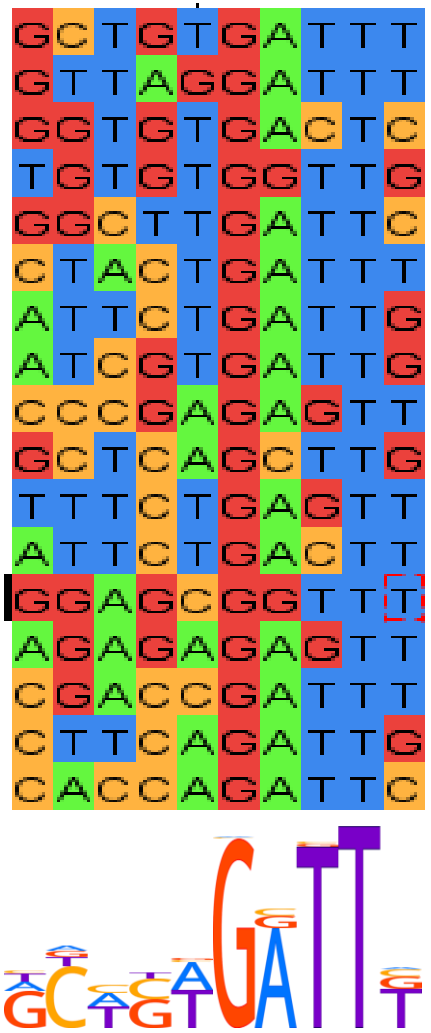
Информационное содержание и энтропия

Повторение

Какой из двух наборов представителей одного и того же сигнала взять для построения PWM?

Сигналы подобраны для не идентичных, но бизкородственных белков – Транскрипционных факторов (TF): GFI1_HUMAN и GFI1B_HUMAN [1]





Что на интересует:
Поиск по какой из матриц **PWM1 или PWM1b** даст меньше случайных находок?

Философский ответ:
По PWM, построенной по выравниванию, в котором больше информации

В обеих сигналах по 17 посл. И по 10 колонок.

Источник примеров БД HOCOMOCO [1] адаптировано мной.

[1] Kulakovskiy et al., HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018

Как измерять информацию в сигнале?

Уменьшением энтропии по сравнению с энтропией случайного сигнала – выравнивания, составленного из случайных букв

Случайное выравнивание

- то, в котором в каждой колонке j выбираем буквы случайно с базовыми вероятностями p(b), одинаковыми для всех колонок (например, частотами буквы b в геноме)

- Его энтропия H вычисляется по формуле
  $H = \sum_j H_j$

  Энтропия колонки H_j вычисляется по формуле

  $H_j = - \sum_b p_j(b) \log_2 p_j(b)$

j – номер колонки, b – буква A, T, G или C

Информационное содержание сигнала, заданного выравниванием

- Вычисляется по формуле

$$IC = \sum_j IC_j$$

Информ. содержание IC_j вычисляется по формуле

$$IC_j = \sum_b f_i(b) \log_2 f_i(b)/p(b)$$

- В выравнивании в колонке j частоты букв $f_j(b)$.

Если $f_j(b) >> p(b)$, то значит в сигнале буква b в этой позиции предпочитаема.

Если $f_j(b) \approx p(b)$, то буква b не даёт новой информации – безразлична или даже избегаема

j – номер колонки, b – буква A, T, G или C

Итоговая формула для информационного содержания сигнала, заданного выравниванием:

$$IC = \sum_i IC_j$$

$$IC_j = \sum_b f_i(b) \log_2 f_i(b)/p(b)$$

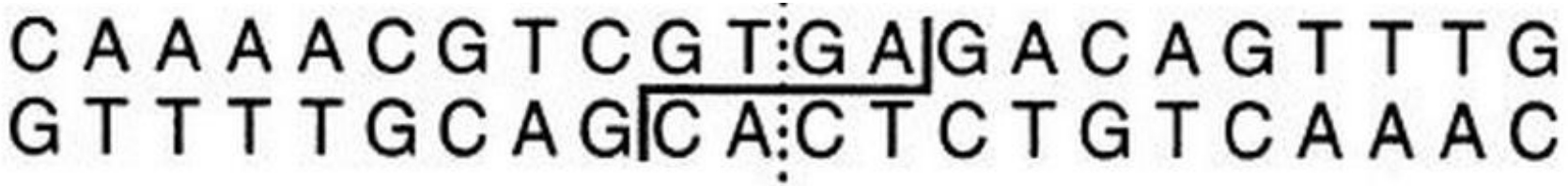
IC_j - информационное содержание колонки j выравнивания

Часто для простоты, предполагают, что p(A) = p(T) = p(C) = p(G) = 1/4

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)))
  - по формуле **IC = 2** при базовых частотах ¼

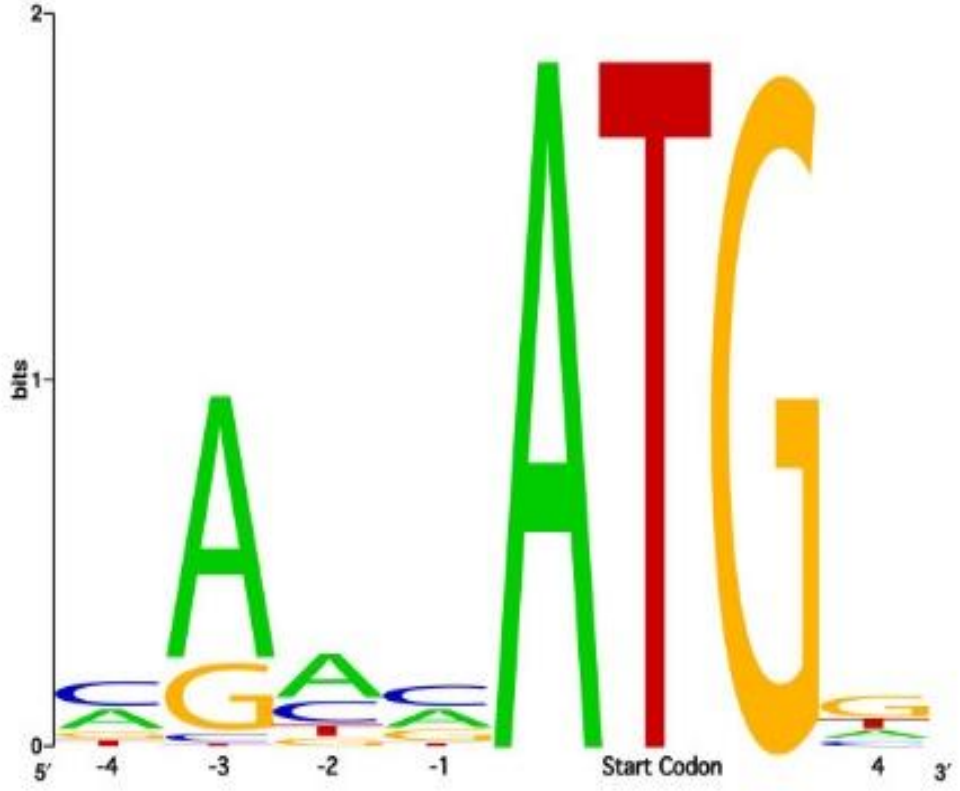
- Сильный сигнал:
  - Эндонуклеаза I-CreI семейства LAGLIDADG узнает такую последовательность. Вероятность обнаружить в геноме такую последовательность случайно близка к 0



по формуле **IC = 22 x 2 = 44** при базовых частотах ¼

Информационное содержание IC выравнивания равно

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



В LOGO сигнала буквы имеют высоту, равную информационному содержанию буквы в предположении, что базовые частоты всех 4х букв равны ¼ [2]

Поэтому max(IC(j)) = 2

[2] Schneider, Stephens , Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990

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

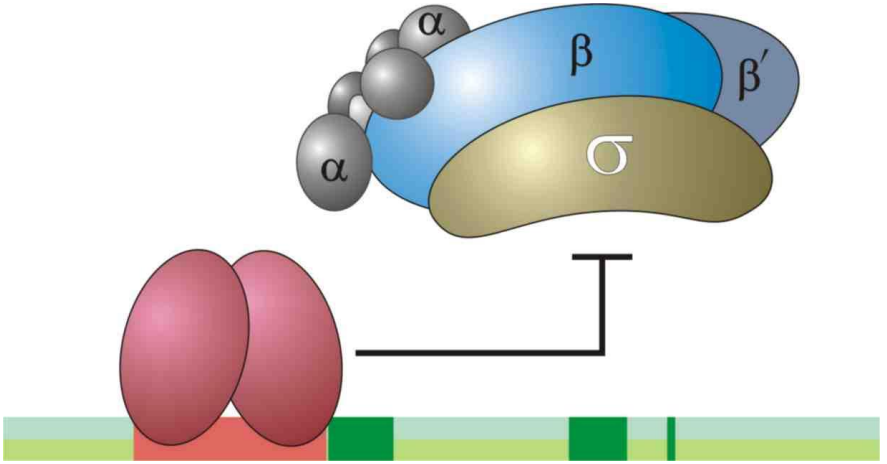
Транскрипционные факторы (TF)

Белки, которые регулируют транскрипцию определенных генов, связываясь со специфическими сайтами в промоторах генов;

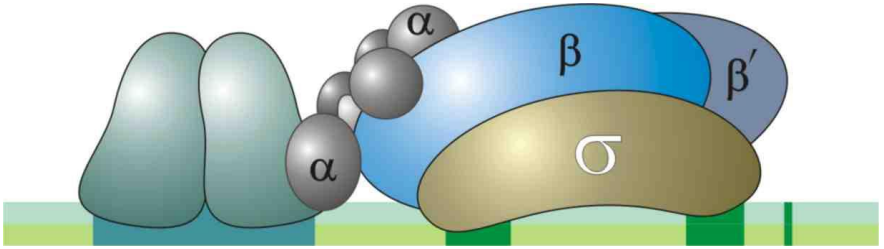
у эукариот еще в энхансерах и сайленсерах

TF прокариот

- Репрессия



- Активация



3D структуры распространенных семейств TF

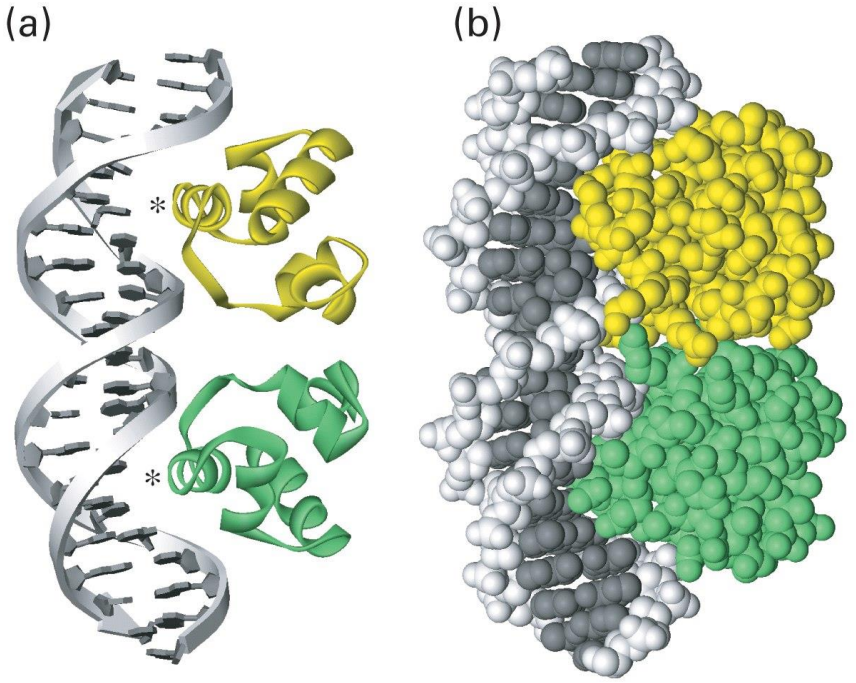
Слайды заимствованы из презентаций из The Ohio State University. Tice lectures in Biochemistry !

The author is not clear for me. Either Carlos O. Miller himself or one of the authors of lecture series in honor of Miller

Запрос «miller lecture on proteins»
Использовал Chap. 7 Transcriptional Control of Gene Expression (Part B)

Helix-turn-helix TFs

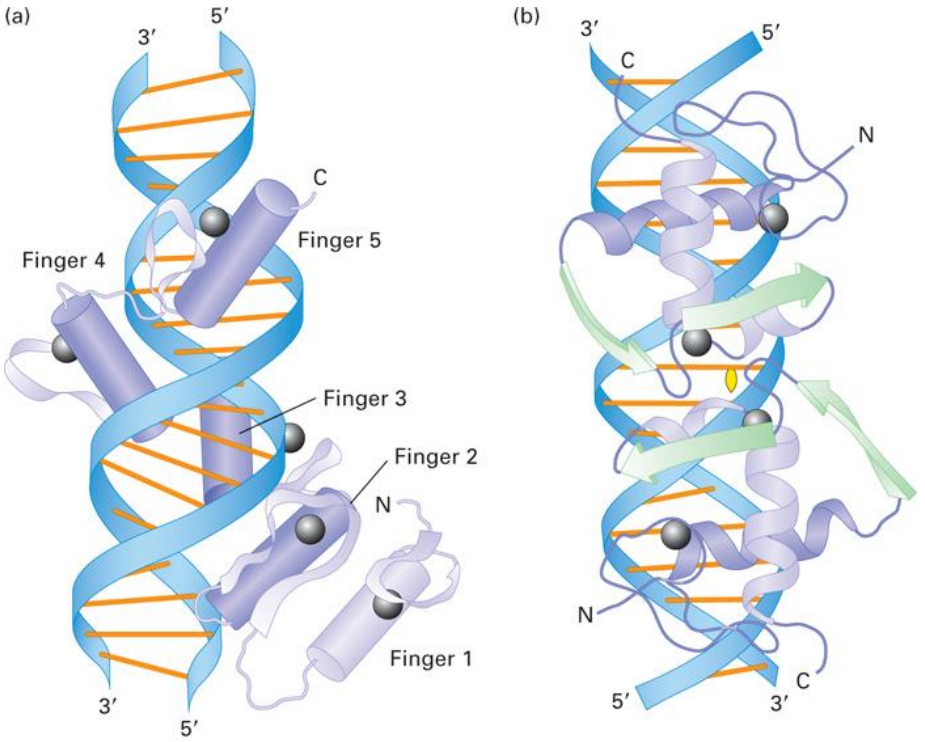
DNA-binding proteins bind specifically to DNA via non-covalent interactions. α-helices are one of the most common types of DNA-binding sequences (Fig. 7.28). The side-chains of residues within the α-helix often bind to the surfaces of bases exposed in the major groove of double-helical DNA. Binding to phosphates and bases in the minor groove typically is less important. One of the most common DNA-binding structure motifs is the helix-turn-helix.The second helix in this motif (the DNA recognition helix) typically



binds to a specific sequence of bases in DNA. The recognition helices in the dimeric bacteriophage 434 repressor are indicated with asterisks in Fig. 7.28a. Helix-turn-helix TFs are common in bacteria.

Zinc-finger TFs

The most common DNA-binding motif in human and multicellular animal TFs is the zinc finger. Two types of zinc finger TFs are discussed here--C_2H_2 zinc finger TFs (Fig. 7.29a) and C_4 zinc finger TFs (Fig. 7.29b). Most TFs that contain C_2H_2 zinc fingers are monomeric. Its 2 cysteine and 2 histidine residues bind to zinc ions (Zn^{2+}) (Fig. 7.29a), and the α-helix containing the 2 histidines binds to bases in the major groove. Much less

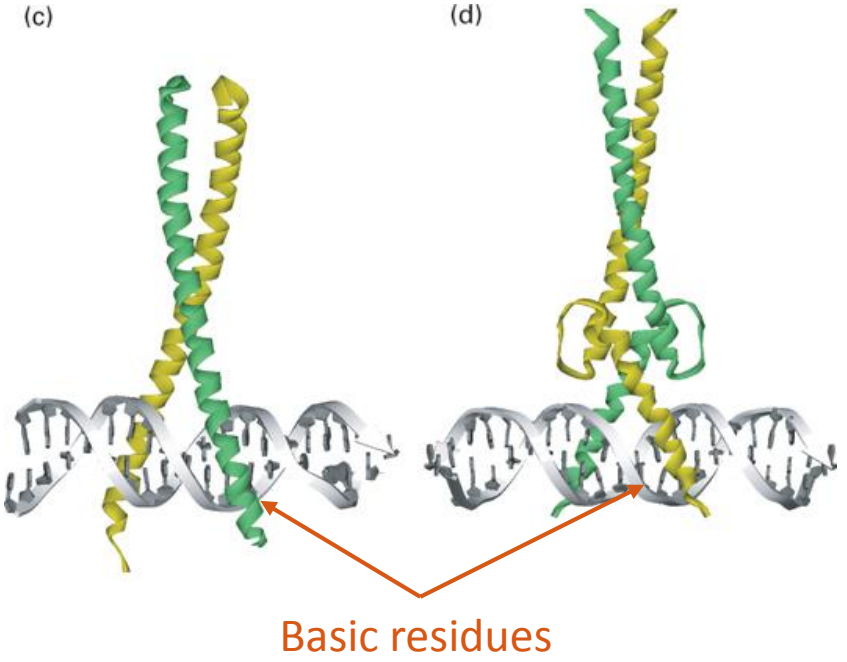


common are TFs containing C_4 zinc fingers. Most TFs containing this motif are dimeric. Nuclear receptors, which bind steroid hormones and other compounds, contain this motif. The glucocorticoid receptor is shown in Fig. 7.29b. Zinc ions are bound to the DNA recognition helix of this motif, which contacts bases in the major groove.

Leucine-zipper TFs

Leucine-zipper TFs contain extended α-helices wherein every 7th amino acid is leucine. This periodicity creates a nonpolar face on one side of the helix that is ideal for dimerization with another such protein via a coiled-coil motif (Fig. 7.29c). So-called basic zipper (bZip) TFs have a similar structure except that some leucines are replaced by other nonpolar amino acids. The N-terminal ends of both leucine-zipper and bZip proteins contain basic amino acids that interact with bases in the major groove (Fig. 7.29c). Leucine zipper proteins are now considered to be a subclass of bZip proteins.

Another class of TF, the basic helix-loop-helix (bHLH) proteins are similar to bZip proteins, but contain a loop between the DNA recognition helix and the coiled-coil region (Fig. 7.29d). bZip and bHLH proteins commonly form heterodimeric TFs.



См. также базу TF человека HOCOMOCO [1]

https://hocomoco11.autosome.org/

Jaspar https://jaspar.uio.no/

[1] Kulakovskiy et al., HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018

Проблемы

- Известен TF, как определить его сайт узнавания. Эксперимент. Коллекции HOCOMOCO(human) [1], Jaspar [3](7 таксонов эукариот)
- Если известно несколько сайтов одного TF, то найти все гены, транскрипцию которых регулирует этот TF.
- Найти консервативный сигнал, встречающийся в промоторах нескольких генов. Если IC сигнала большое, то это не случайно. Значит, можно искать объяснение, а именно, TF или иной белок, который связывается с этим сайтом. [4]

[3] Castro-Mondragon et al., JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles, NAR, 2022

[4] Baumgarten et al. Improved linking of motifs to their TFs using domain information. *Bioinformatics*. 2020

Поиск сайтов *de novo*

Пакет MEME

MEME siute

- On line
- На kodomo
  >meme --help
- Параметры командной строка с примерами лучше смотреть на MEME siut
  https://meme-suite.org/meme/doc/meme.html#examples

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.

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

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☺)

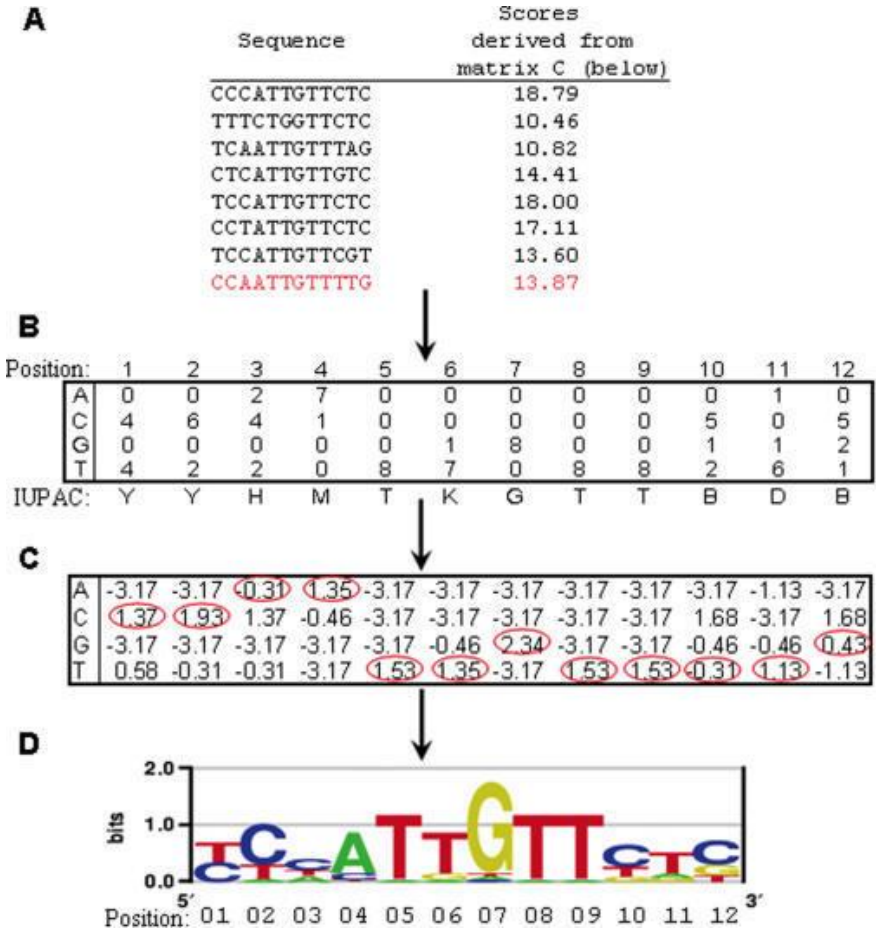
3) Как-то упустил что наши люди – коллеги - тоже сделали детектор мотивов Chipmunk (https://opera.autosome.ru/chipmunk/discovery)

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

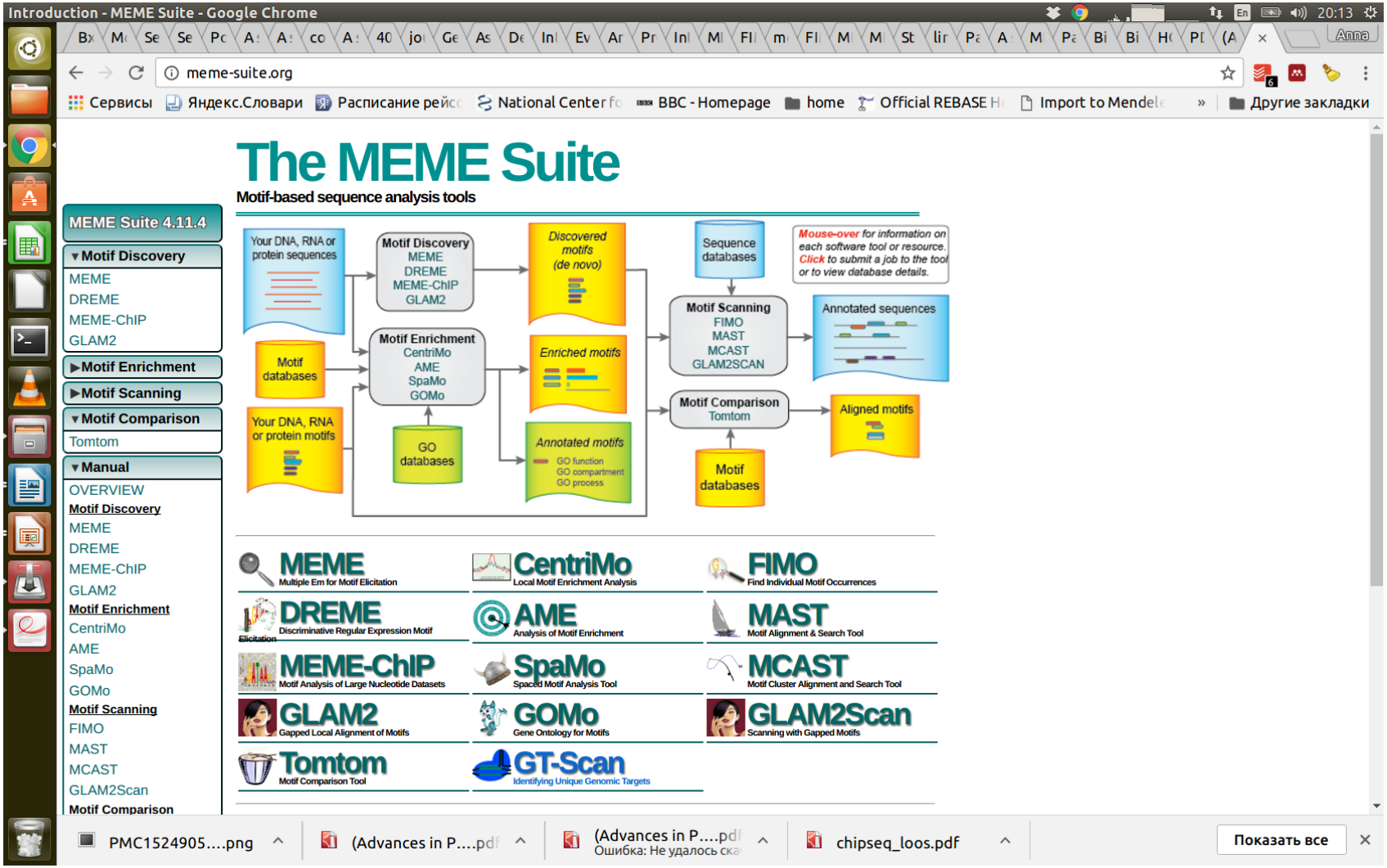


Вес (I(bj) основания b в данной позиции j
I(bj)=f(bj)*log f(bj) — p(b)*log p(b), где f(bj) — частота основания и в позиции j выравнивания, p(b) — фоновая частота основания b

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

УПРАЖНЕНИЕ

Найти встречи сигнала с PWM из базы данных HOCOMOCO в геноме человека

Упражнение

1. Выберите мотив TF человека из БД HOCOMOCO
   https://hocomoco11.autosome.org/
   На основании LOGO предпочитайте мотив с большим IC
2. **Сохраните PWM и LOGO этого мотива**
3. Найдите с помощью этой PWM сигналы в геноме человека.
   Используйте сервис PWMscan пакета PWMTools
   https://ccg.epfl.ch/pwmtools/pwmscan.php
   Берите версию генома hg38. Выберите Matrix format: Real PWM
   Используйте максимально строгий порог (P-value, Score или Percentage), чтобы находок было поменьше.
4. Сохраните результат
   1. **в bed-формате** (хромосома, координаты сигнала, вес и P-value)
   2. **последовательности сигнала с окрестностями по 3 нукл.** с каждой стороны т.е. From = -3 а To = длина сигнала +3

Задание пр. 7

Найти мотивы с помощью MEME в промоторах двух десятков генов из генома бактерии.

Для одного мотива найти представителей того же сигнала в промоторах других генов.

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

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.

Задания

КОНЕЦ

презентации

Дополнительные слайды

- Bed формат

Bed формат

olumn number	Title	Definition
1	**chrom**	Chromosome (e.g. chr3, chrY, chr2_random) or *scaffold* (e.g. scaffold10671) name
2	**chromStart**	Start coordinate on the chromosome or scaffold for the sequence considered (the first base on the chromosome is numbered 0)
3	**chromEnd**	End coordinate on the chromosome or scaffold for the sequence considered. This position is non-inclusive, unlike chromStart.
4	**name**	Name of the line in the BED file
5	**score**	Score between 0 and 1000
6	**strand**	DNA strand orientation (positive ["+"] or negative ["-"] or "." if no strand)
7	**thickStart**	Starting coordinate from which the annotation is displayed in a thicker way on a graphical representation (e.g.: the start codon of a gene)
8	**thickEnd**	End coordinates from which the annotation is no longer displayed in a thicker way on a graphical representation (e.g.: the stop codon of a gene)
		RGB value in the form R,G,B (e.g. 255,0,0)

КОНЕЦ ПРЕЗЕНТАЦИИ