Л2.Поиск сигналов de novo

I. Контрольная

По Л1: паттерны, PWM

Как написать контрольную

- Откройте файл CW-1.docx по ссылке
- Один вариант – одна страница. Задание написано на странице варианта.
- Выберите любой вариант. Скопируйте страницу к себе в файл
- После выполнения, ответ вносите в форму на этой странице, открываете на своём сайте и записываетесь в очередь на проверку
- Выполняете задание любым доступным для вас методом – программирование, калькулятор на телефоне, на листочке бумаги, в электронной таблице.
- Укажите какой метод использовали и этапы вычисления

II. Оценки пригодности

паттернов, PWM для поиска сигналов

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

Информационное содержание и энтропия сигнала

Энтропия – мера неупорядоченности (обозначается H)

Информация противоположна энтропии (обозначается IC)

Чем больше энтропия, тем меньше информации.

Чем больше информации, тем меньше энтропия

Идеалистически, IC = Hbefor – Hafter.

Мера содержания информации в сигнале равна тому, насколько уменьшилась неопределённость (т.е. энтропия) после появления сигнала.

Чем больше IC, тем сильнее сигнал.

Другое дело, как это соображения применить для сигнала, заданного выравниванием последовательностей

Шнайдер с соавт. в 1986 году предложили формулу для IC, которая с используется как основная [6-1].

$$IC = \sum_i IC_j$$

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

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

Преимущества. Формула Шнайдера простая. Она правильно отражает интуитивные представления. Она успешно применялась во множестве работ

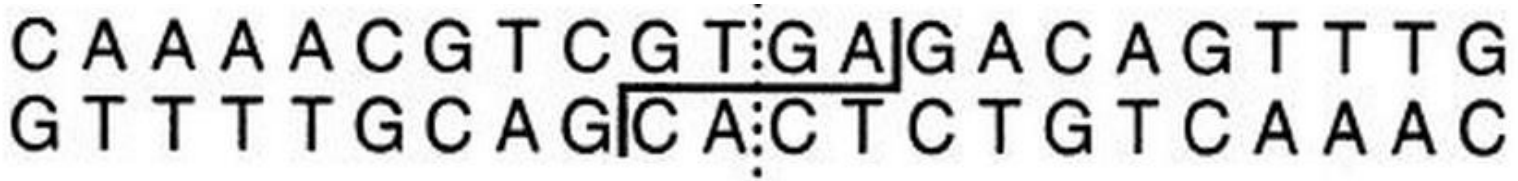
$f_j(b)$ - частота буквы в колонке, p(b) – базовая частота буквы b

Есть $f_j(b) >> p(b)$, то IC_j большое число. Значит, в сигнале буква b в этой позиции предпочитаема.

Если $f_j(b) \approx p(b)$, то $\log_2 f_i(b)/p(b) \approx 0$. Значит, буква b в колонке j не даёт новой информации – безразлична или даже избегаема

IC слабого и сильного сигналов

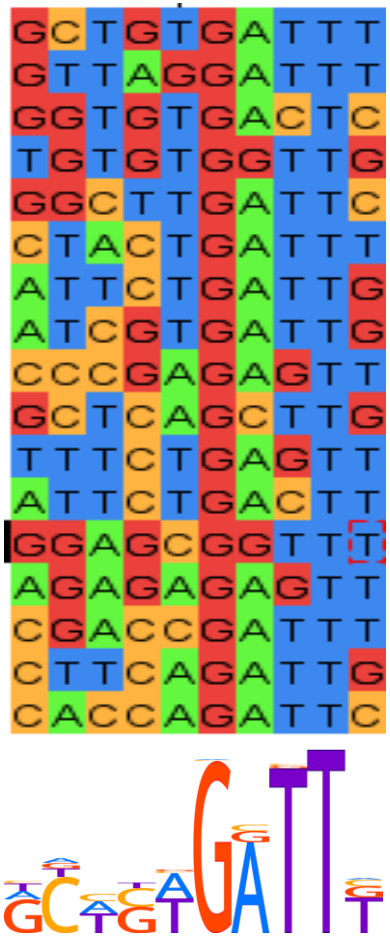
- Слабый сигнал:
  - Гомеодомен - консервативный ДНК-узнающий домен многих важных транскрипционных факторов эукариот
  - Узнаёт короткую последовательность ДНК
  - На основании наложения структр гомеодоменов найден единственный общий контакт домена с сайтом ДНК:
    Asn51 имеет две водородных связи с аденином (!)
  - Сигнал NNANN очень слабый , слабее не придумаешь)))
  - 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

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]



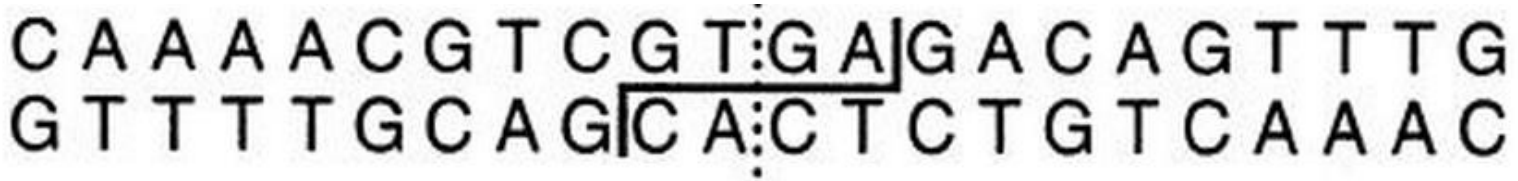
Что на интересует:
Поиск по какой из матриц **PWM1 или PWM1b** даст меньше случайных находок?

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

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

IC слабого и сильного сигналов

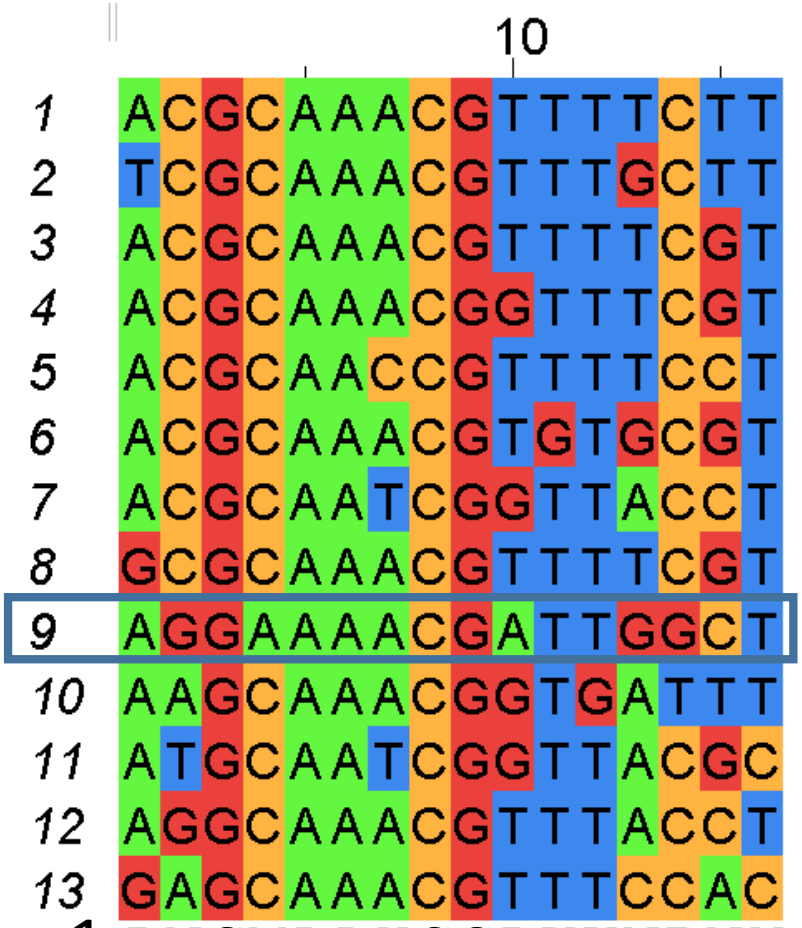
- Слабый сигнал:
  - Гомеодомен - консервативный ДНК-узнающий домен многих важных транскрипционных факторов эукариот
  - Узнаёт короткую последовательность ДНК
  - На основании наложения структр гомеодоменов найден единственный общий контакт домена с сайтом ДНК:
    Asn51 имеет две водородных связи с аденином (!)
  - Сигнал NNANN очень слабый , слабее не придумаешь)))
  - 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

Чем плохи паттерны-1 и -2?

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



Паттерн-1 DNGMAAHCGDKKNBNY

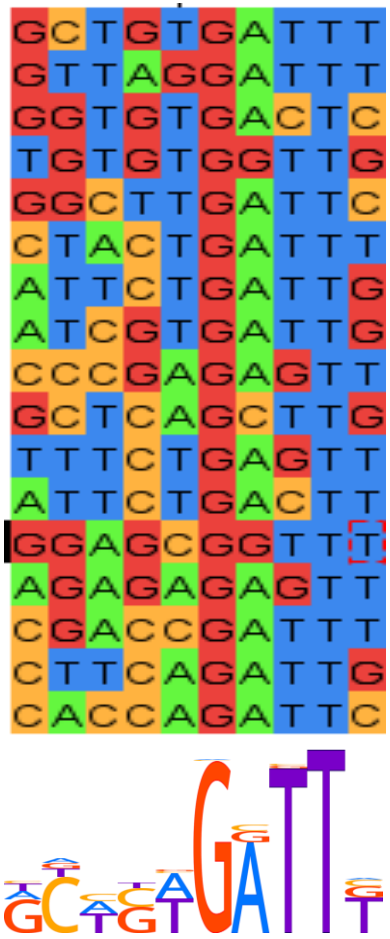
Паттерн-2 . . **G.AA.CG**. . . . . .

Паттерн-1 сильно изменится, если убрать посл. 9. Он описывает очень большое число возможных последовательностей

Паттерн 2 - точки заменяют N. Всего 5 условий. Колонки с одной заменой несут информацию, но не учитываются в паттерне. По случайным причинам встретится примерно один раз на 1000 п.н.

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

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

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

Энтропия H по Шеннону и IC

- **Энтропия Hg для частот букв в геноме**. Четыре буквы - четыре частоты p(A), p(T), p(G), p(C)

  $H_g = -\sum_b p(b) \log_2 p(b)$ b in {A,T,G,C}

- Теорема (Шеннон, 1948). H максимальна если частоты всех букв равны p(A) = p(T) = p(G) = p(C) = ¼ $H_{g\_max} = 2$

- **Энтропия Hj частот букв fj(b) в колонке j выравнивания** равна

  $H_j = -\sum_b f_j(b) \log_2 f_j(b)$ b in {A,T,G,C}

- **Первое определение IC для колонки выравнивания.**
  $IC_j = H_g - H_j$ [иногда используют и такую упрощённую оценку $H_{g\_max} - H_j$]

IC_{aln} выравнивания равна $IC_{aln} = \sum_j IC_j$ в предположении независимости колонок и в силу аксиом энтропии

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

Шнайдер с соавт. в 1986 году предложили формулу для IC, которая с используется как основная [6-1].

$$IC = \sum_i IC_j$$

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

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

Преимущества. Формула Шнайдера простая. Она правильно отражает интуитивные представления. Она успешно применялась во множестве работ

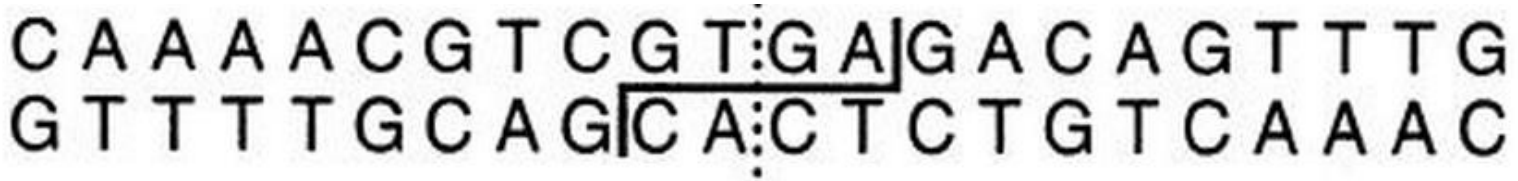
$f_j(b)$ - частота буквы в колонке, p(b) – базовая частота буквы b

Есть $f_j(b) >> p(b)$, то IC_j большое число. Значит, в сигнале буква b в этой позиции предпочитаема.

Если $f_j(b) \approx p(b)$, то $\log_2 f_i(b)/p(b) \approx 0$. Значит, буква b в колонке j не даёт новой информации – безразлична или даже избегаема

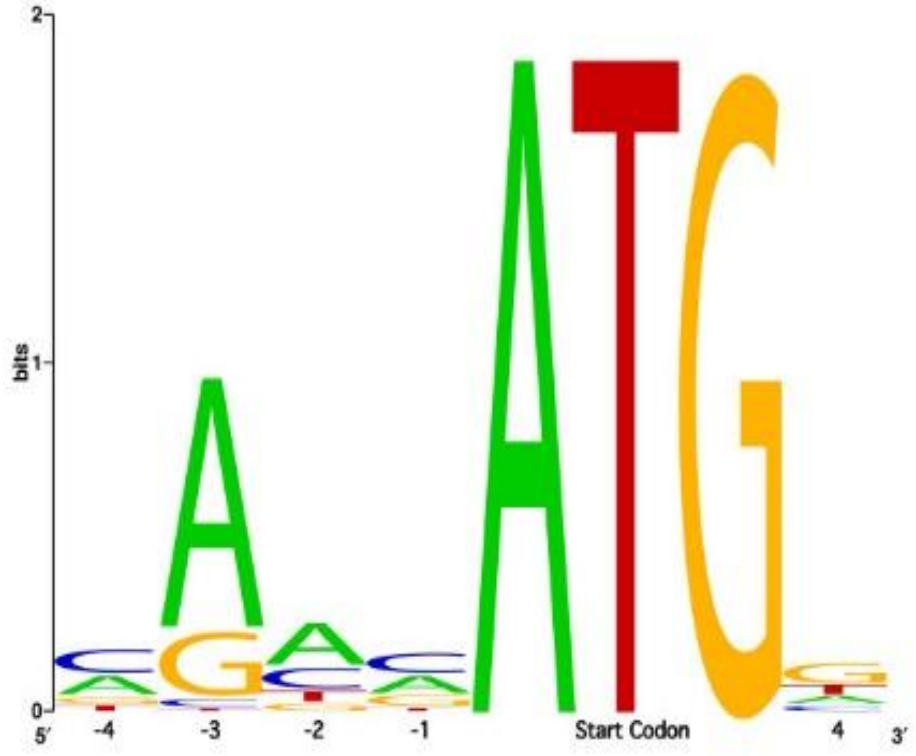
IC слабого и сильного сигналов

- 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 имеет две водородных связи с аденином (!)
  - Сигнал NNANN очень слабый , слабее не придумаешь)))
  - 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

LOGO высота буквы b в позиции j равна $IC_j(b)$ в битах



В LOGO сигнала буквы имеют высоту, равную информационному содержанию буквы в предположении, что базовые частоты всех 4х букв равны ¼ [6-2]

Поэтому max(IC(j)) = 2

Если $IC_j(b) < 0$, то считают $IC_j(b) = 0$

[6-1] Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J Mol Biol. 1986

[6-2] Schneider, Stephens , Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990
webLOGO

Распространённые ДНК связывающие домены

Названия, 3D, Pfam

Двойная спираль ДНК

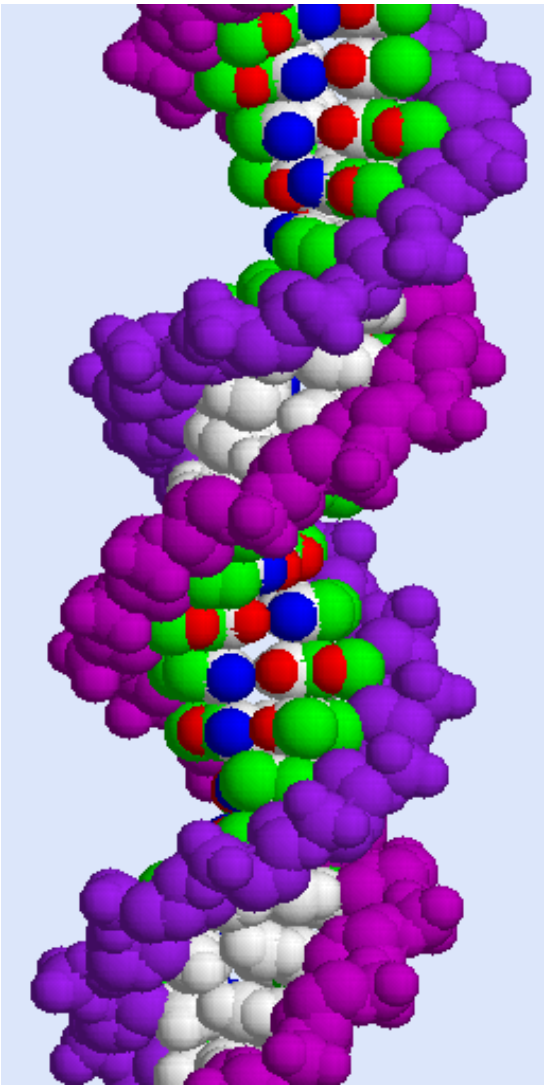


Рис. 3D структура фрагмента ДНК *in vitro*.

Получена с помощью рентгеноструктурного анализа (РСА).

Шарик = атом

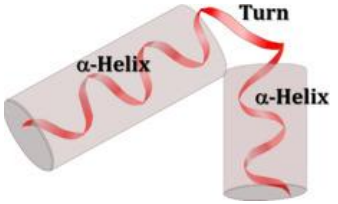
Водороды не видны, т.к. маленькие и РСА их не видит

Атомы остова – фиолетовые двух оттенков чтобы различать цепочки.

В большой бороздке красный – кислород, синий – азот, зеленый – углерод. Малая бороздка не окрашена

БЕЛКИ (некоторые) МОГУТ УЗНАТЬ ПОСЛЕДОВАТЕЛЬНОСТЬ УЧАСТКА ДНК по расположению N, O, C в большой бороздке. СИГНАЛЫ!

HTH (Helix-turn-helix TFs

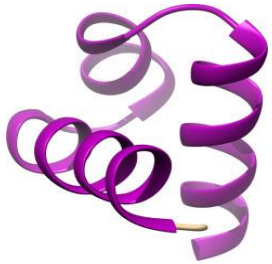


HTH структурный мотив

Многие ТФ для узнавания сигнала и связывания с ДНК используют структурный мотив спираль-поворот-спирал (HTH).

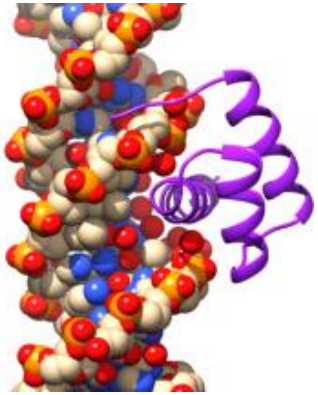
С-концевая спираль – узнающая. Онапомещается в большую бороздку ДНК.
Её а.к. остатки для узнавания сигнала должны образовать несколько (не ковалентных) связей с основаниями ДНК. Так сказать, определить код в большой бороздке!

Дополнительно, поворот и предыдущая спираль образуют связи с остовом и иногда малой бороздкой для правильно стабилизации HTH содержащего домена белка относительно ДНК



Three-helix bundle”

Часто HTH входит состав “three-helix bundle”. Добавляется спираль H1 с N-конца, антипараллельная 2й спирали и перпендикулярная 3й - узнающей



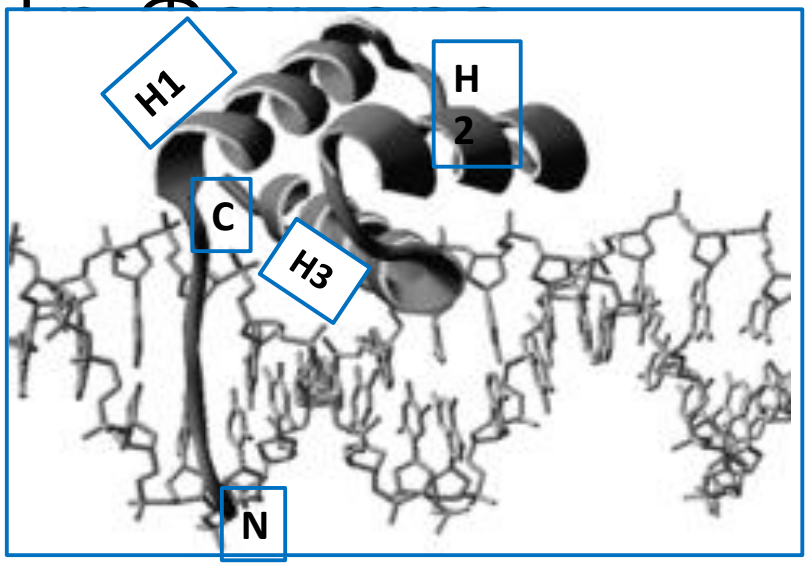
HTH структурные мотивы широко распространены в ДНК узнающих белках прокариот, эукариот и вирусов.

Они делятся на много семейств.

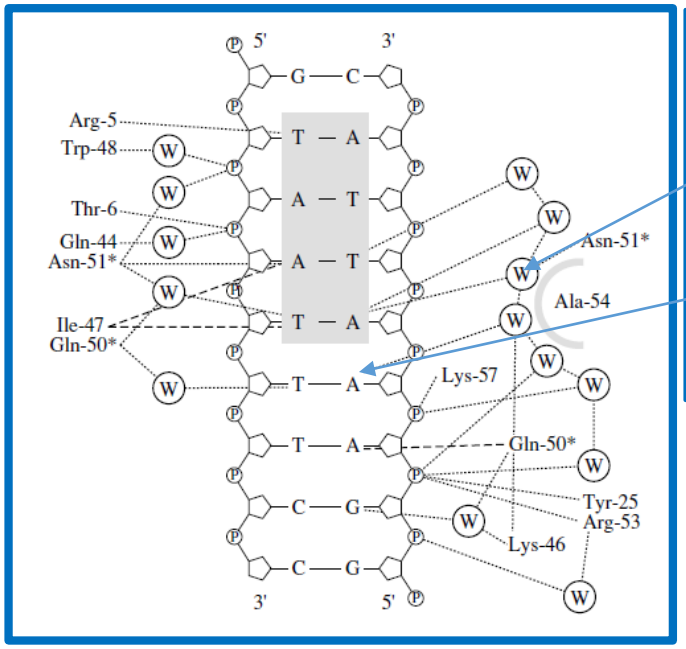
A representation of three-helix bundle class of helix-turn-helix containing proteins

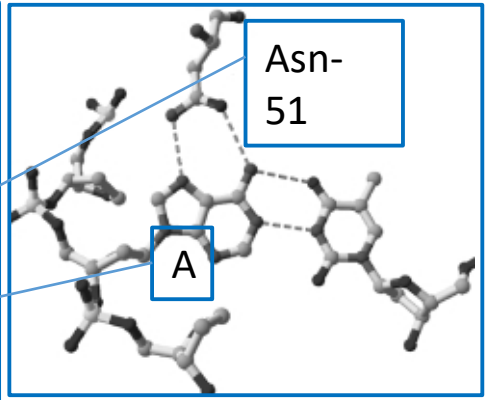
DNA structure | DNA Sequence Recognition by Proteins
K. Rutherford, G.D. Van Duyne, in Encyclopedia of Biological Chemistry (Third Edition), 2013

Гомеодомены – пример HTH



X-ray 3D, PDB **1B72**
Узнающая спираль H3 лежит в большой бороздке, придавленная сверху перпендикулярной спиралью H2.
Спирали H2 и H3 ссоответствуют структурному мотиву (HTH).
Спираль H1 антипараллельна спирали H2.
N-концевая рука подвижна. А.к. остатки руки и спиралей H1, H2 взаимодействуют с остовом, стабилизируя комплекс.





PDB 3HDD. Контакт ASN_51 с аденином в большой бороздке был обнаружен во всех
3D структурах гомеодомен – ДНК.
ASN-51 был обнаружен в 629 последовательностях из 631.
Сегодня PF00046 (homeodomain) seed – 136, full - 244133
Seed подтверждает наше предположение 2001 года.
Проверить на 244тыс.посл. не успел(((Кто поможет?

Ledneva RK, Alekseevskiĭ AV, Vasil'ev SA, Spirin SA, Kariagina AS. Structural aspects of homeodomain interactions with DNA. Mol Biol (Mosk). 2001

Ещё один гомеодомен. Изображен в хорошем ракурсе

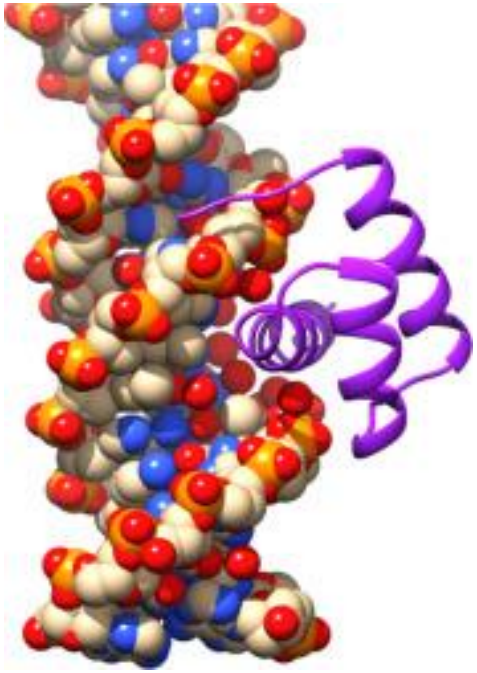


Fig. V.28. The structure of a homeodomain-DNA complex. The image was created from pdb3hdd.

DNA structure | DNA Sequence Recognition by Proteins
K. Rutherford, G.D. Van Duyne, in Encyclopedia of Biological Chemistry (Third Edition), 2013

Zinc-finger TFs

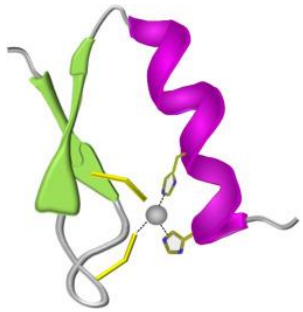
В транскрипционных факторах с цинковыми пальцами (ZF) последовательные узнающие спирали позволяют узнавать длинные сигналы, состоящие из отдельных коротких фрагментов.

Отдельный цинковый палец считают элементом супер-вторичной структуры. Он состоит из антипараллельной бета-шпильки и альфа-спирали, координированными атомом цинка. Чаще всего цинк координирует Cys_2-His_2 мотив

Один ZF имеет малое сродство к сигналу. Поэтому ТФ состоит из каскада ZF-элементов. а Cys_2-His_2

ZF в полипептидной цепи ТФ не идентичны, их узнающие спирали узнают разные короткие сигналы. Этим в ZF ТФ достигается высокая специфичность к длинным последовательностям ДНК

Цинковые пальцы широко распространены среди многоклеточных эукариотических ТФ . Встречаются у прокариот и вирусов (ссылка справа)



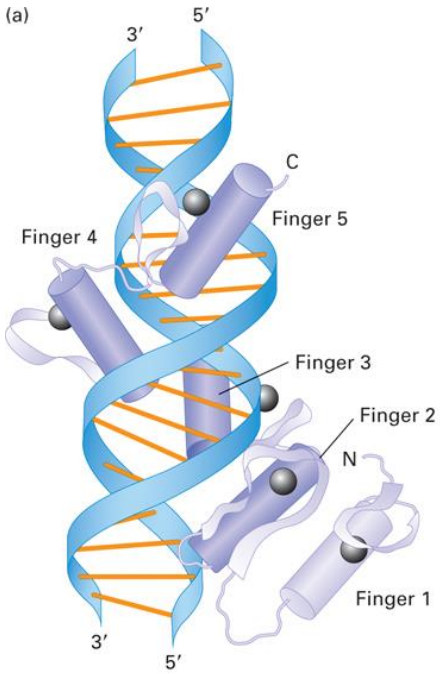
Один изолированный ZF

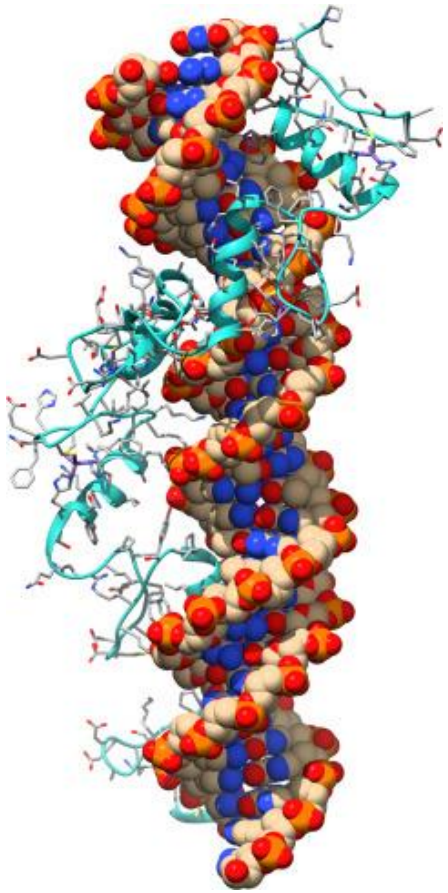
ТФ с тремя ZF

Malgieri G et al., The prokaryotic zinc-finger: structure, function and comparison with the eukaryotic counterpart. FEBS J. 2015

DNA structure | DNA Sequence Recognition by Proteins
K. Rutherford, G.D. Van Duyne, in Encyclopedia of Biological Chemistry (Third Edition), 2013

Ещё изображения ZF

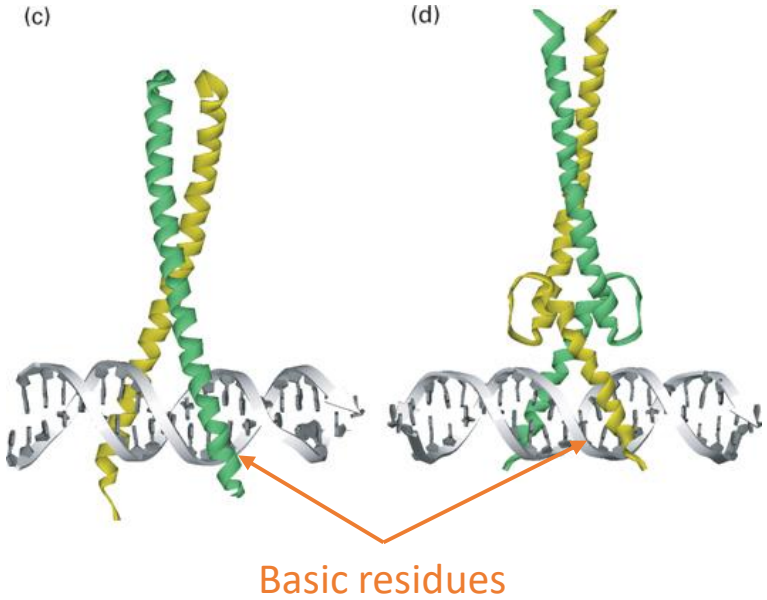




Leucine-zipper TFs

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.



Коллекции TF и их сайтов на 2019 [5]

TRANSFAC eukaryotic TFs, their genomic binding sites, and DNA-binding profiles
JASPAR motifs for multicellular eukaryotes
PROSITE protein domains, families and functional sites in addition to related patterns and profiles to recognize them
YEASTRACT predicted TFs for S. cerevisiae.
SCPD http://rulai.cshl.edu/SCPD/
RegulonDB E. coli both computational as well as experimental data of predicted objects
CisBP a list of >160,000 predicted TFs from >300 species
DBTBS TFs for Bacillus subtilis

[5] Hashimi et al., Review of Different Sequence Motif Finding Algorithms, 2019

Ref.

Free.Tognon M, Giugno R, Pinello L. A survey on algorithms to characterize transcription factor binding sites. Brief Bioinform. 2023

Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. Genome Res. 2015

or example, we found that most of the TFs
belonging to
homeodomain family (88 out of 96 members),
POU family (10 out of 13),
forkhead family (14 out of 16) prefer binding to
regions with low GC content surrounding the
core motif, as opposed to
C2H2 zinc finger (19 out of 41) and
ETS TFs (12 out of 22),

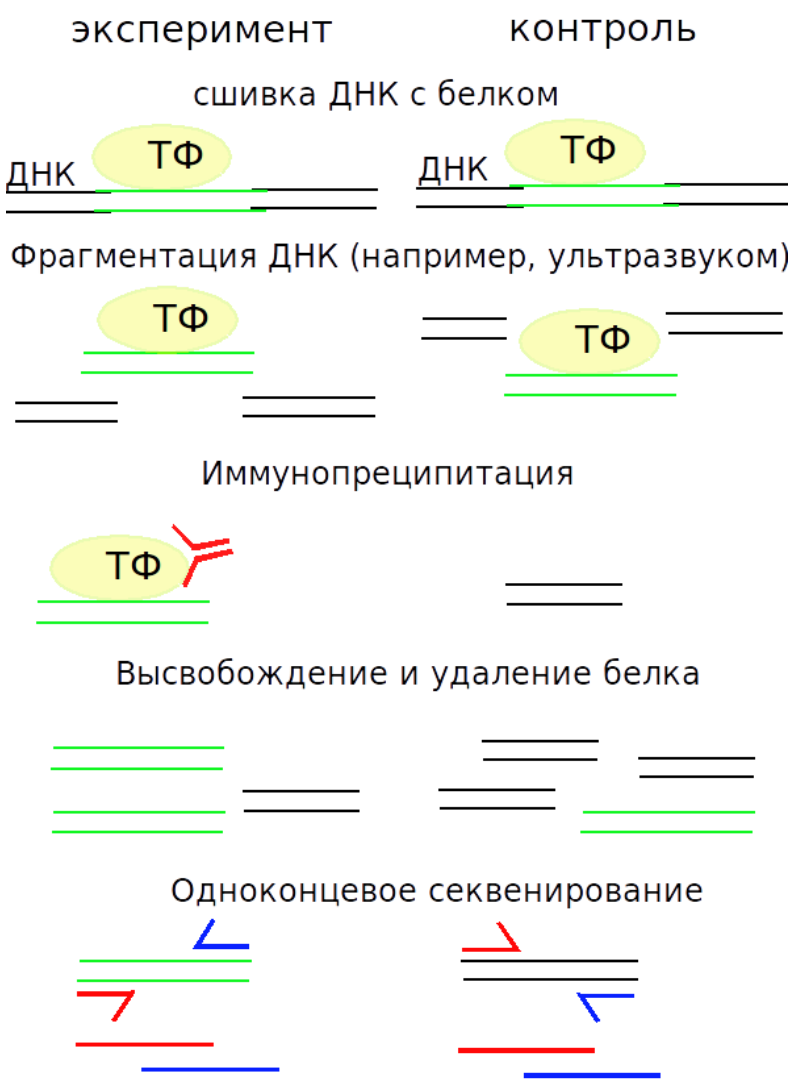
Белки узнают нужные им сигналы лучше, чем самые крутые биоинформатики!

ChIP-seq

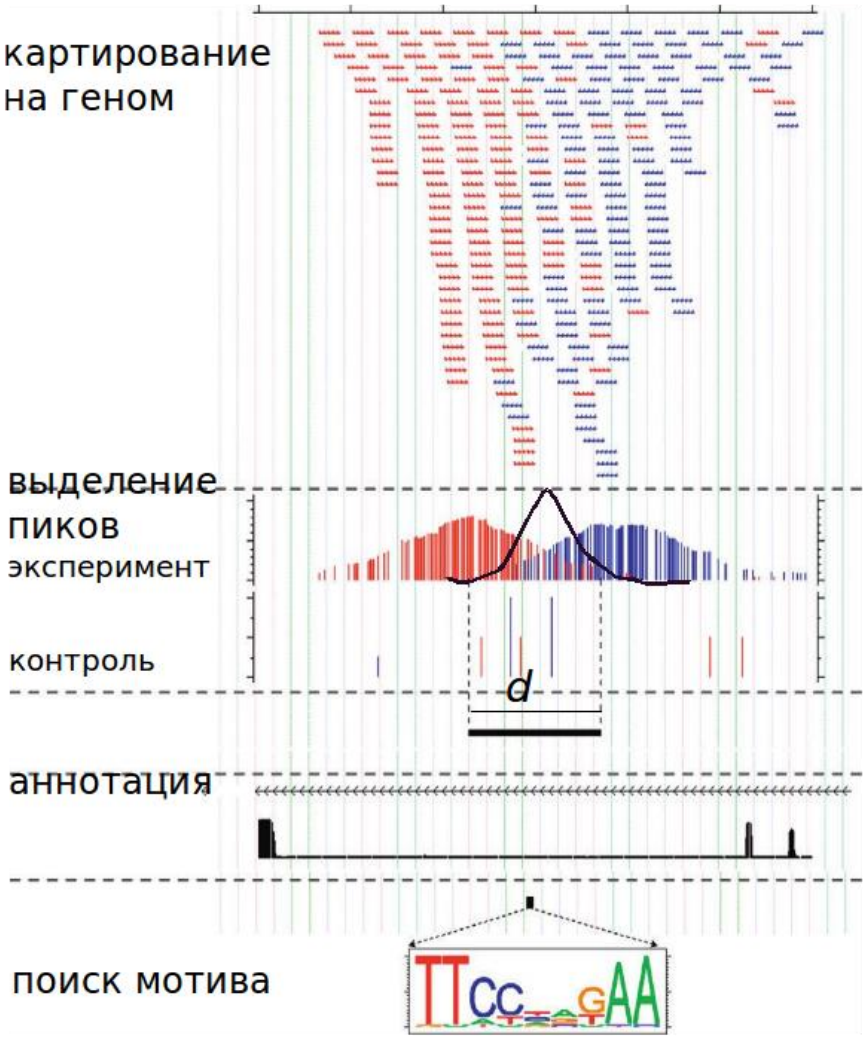
- **Ch**romatin **i**mmuno**p**recipitation (Chip) с последующим высокопроизводительным секвенированием

Эксперимент и анализ данных

Эксперимент



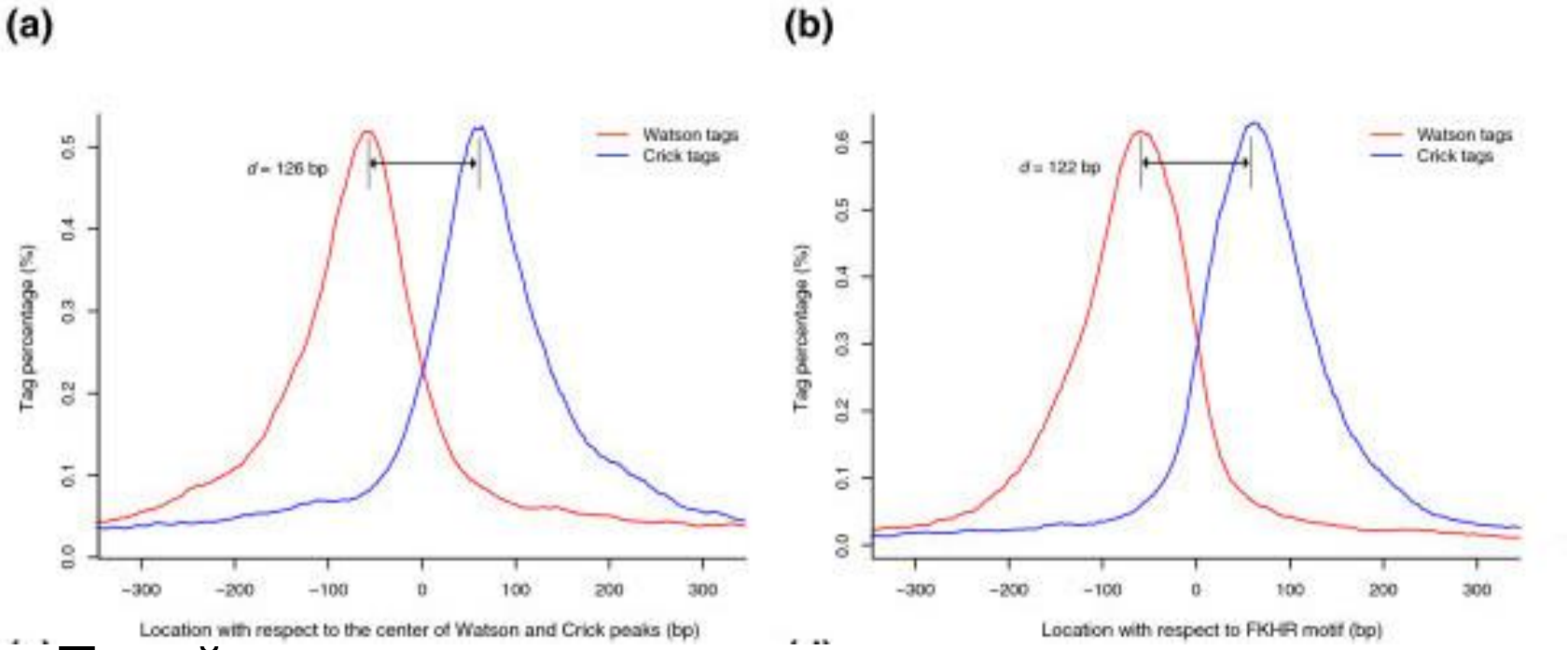
Анализ данных ChIP-seq



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

Выбор величины сдвига

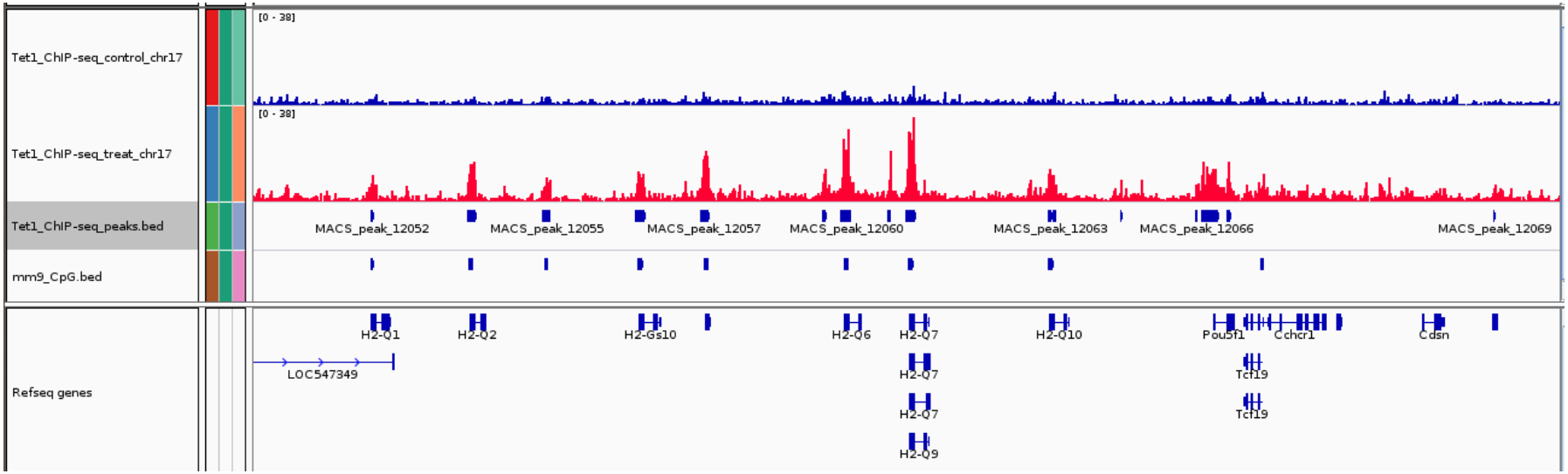
Длина секвенированного фрагмента — 200 п.н.



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.

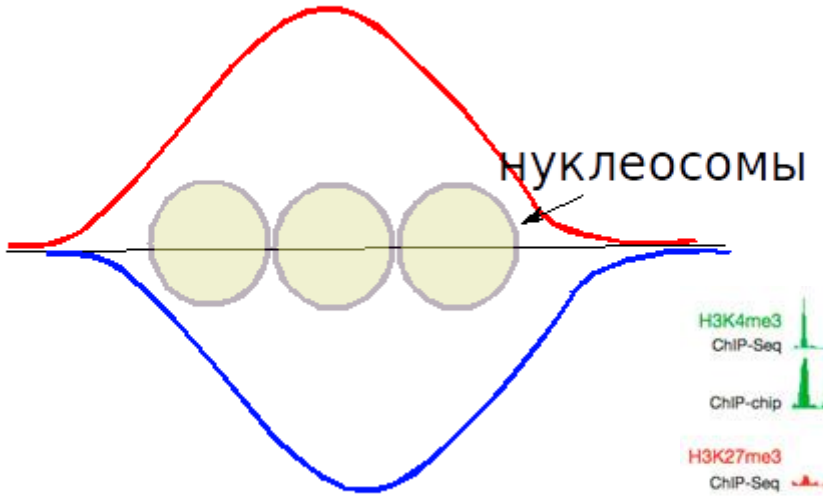
Выбор достоверных пиков

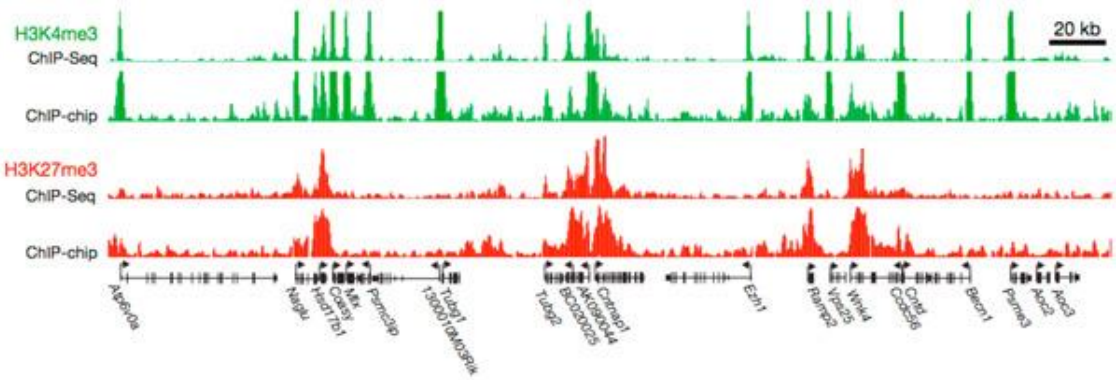


Сравнивают пики в эксперименте и контроле, считают p-value.

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

Связывание с хроматином

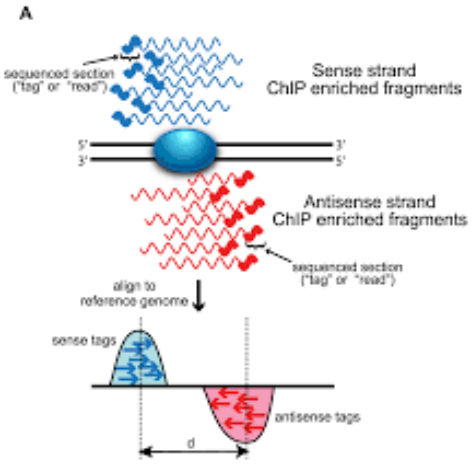




Нет асимметрии пиков

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

Поиск мотивов de novo

1. 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☺)

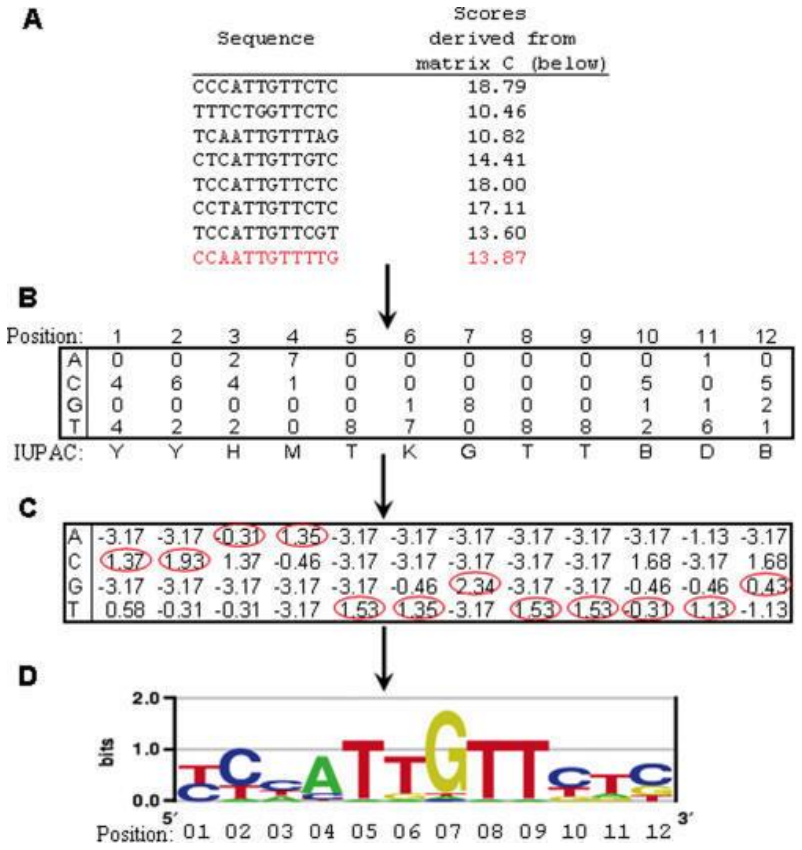
3) Как-то упустил что наши люди – коллеги - тоже сделали детектор мотивов Chipmunk (https://opera.autosome.ru/chipmunk/discovery)

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

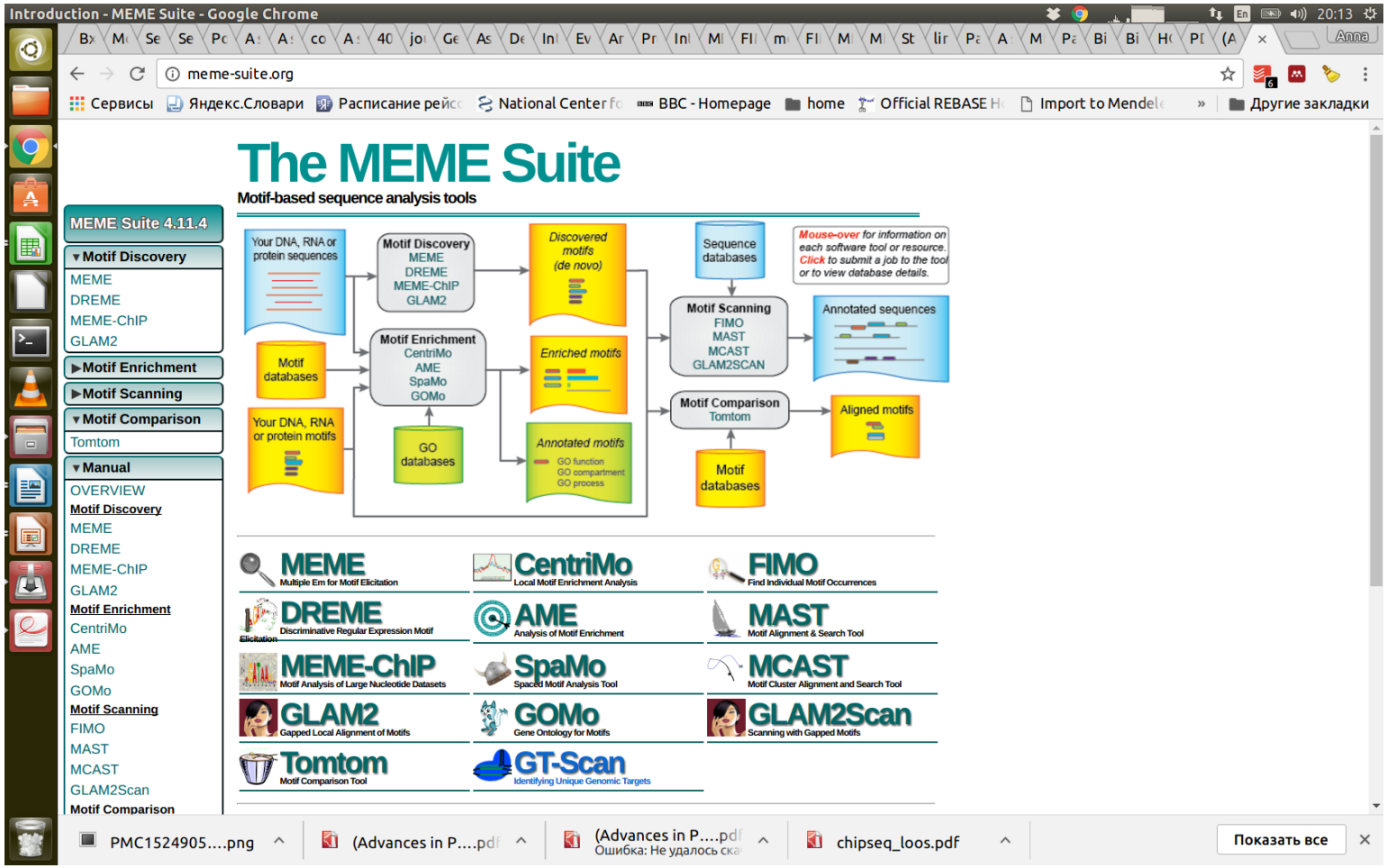


Вес (I(bj) основания b в данной позиции j
I(bj)=f(bj)*log f(bj) — p(b)*log p(b), где f(bj) — частота основания и в позиции j выравнивания, p(b) — фоновая частота основания b

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

Задания на дом

ПРО ЗАДАНИЯ

1. Описать один сигнал.
   Включает литературную составляющую и демонстрацию примеров сигналов

   Поиск по Pubmed.

a. Advanced:
“di Salvo M[au] 2019[dp]” позволяет найти статью этого автора, вышедшую в указанном году.
Можно задать [1au]; 2016:2020[dp];
Pribnow[ti] - название статьи включает Pribnow[-Scheller box] то же, что -10 сигнал для сигма фактора РНК полимеразы.
ORI Finder[ti]
Работают кавычки для выражений "RNA polymerase" не то, что RNA polymerase
Больше можно найти в Advanced

b. Filters:
   полезно использовать
   Free full text
   Review
   PUBLICATION DATE 5 - последние пять лет,
   но некоторые базовые вещи придумали сильно раньше☺

ПРО ЗАДАНИЯ

2. Построить PWM по выборке сигналов. Проверить её работу на независимой выборке. Вычислить информационное содержание выравнивания, по которому строилась PWM И построить LOGO

Пока в указаниях не успел описать сигналы, подходящие для этого задания.
Очевидно, это те, для которых легко набрать более десятка представителей.
Например, таковым является последовательность Козак для человека, другой зверюшки или даже бактерии – интересно же, и статья была по этой теме. Указана в презентации

Как- то так

Энтропия H по Шеннону и IC

- **Энтропия Hg для частот букв в геноме**. Четыре буквы - четыре частоты p(A), p(T), p(G), p(C)

  $H_g = -\sum_b p(b) \log_2 p(b)$ b in {A,T,G,C}

- Теорема (Шеннон, 1948). H максимальна если частоты всех букв равны p(A) = p(T) = p(G) = p(C) = ¼ $H_{g\_max} = 2$

- **Энтропия Hj частот букв fj(b) в колонке j выравнивания** равна

  $H_j = -\sum_b f_j(b) \log_2 f_j(b)$ b in {A,T,G,C}

- **Первое определение IC для колонки выравнивания.**
  $IC_j = H_g - H_j$ [иногда используют и такую упрощённую оценку $H_{g\_max} - H_j$]

IC_{aln} выравнивания равна $IC_{aln} = \sum_j IC_j$ в предположении независимости колонок и в силу аксиом энтропии

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

Шнайдер с соавт. в 1986 году предложили формулу для IC, которая с используется как основная [6-1].

$$IC = \sum_i IC_j$$

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

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

Преимущества. Формула Шнайдера простая. Она правильно отражает интуитивные представления. Она успешно применялась во множестве работ

$f_j(b)$ - частота буквы в колонке, p(b) – базовая частота буквы b

Есть $f_j(b) >> p(b)$, то IC_j большое число. Значит, в сигнале буква b в этой позиции предпочитаема.

Если $f_j(b) \approx p(b)$, то $\log_2 f_i(b)/p(b) \approx 0$. Значит, буква b в колонке j не даёт новой информации – безразлична или даже избегаема