Сигналы и мотивы -2

De novo поиск сигналов в последовательностях

I. КР по Л1

Идёт в зачёт коллоквиума

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.Транскрипционные факторы (TF)

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

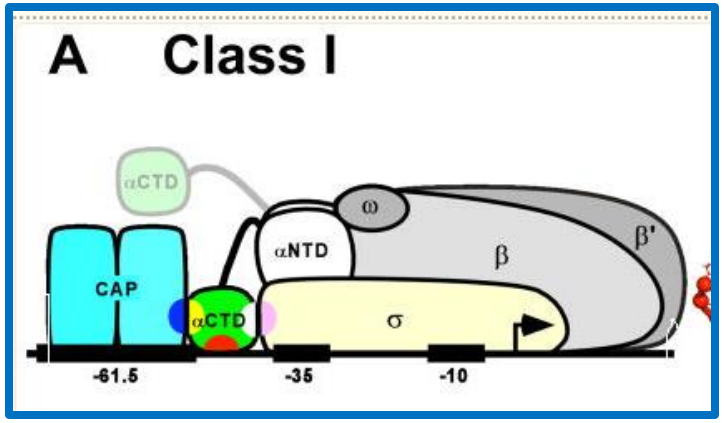
TF прокариот

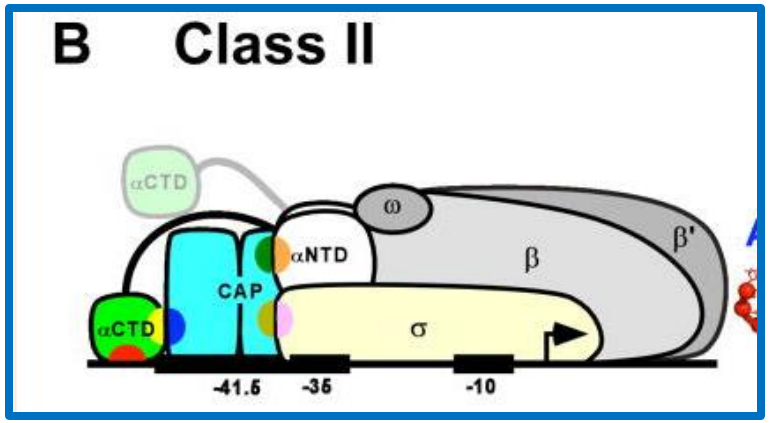
- Репрессоры так или иначе препятствуют сборке комплекса РНК-полимеразы в промоторе.

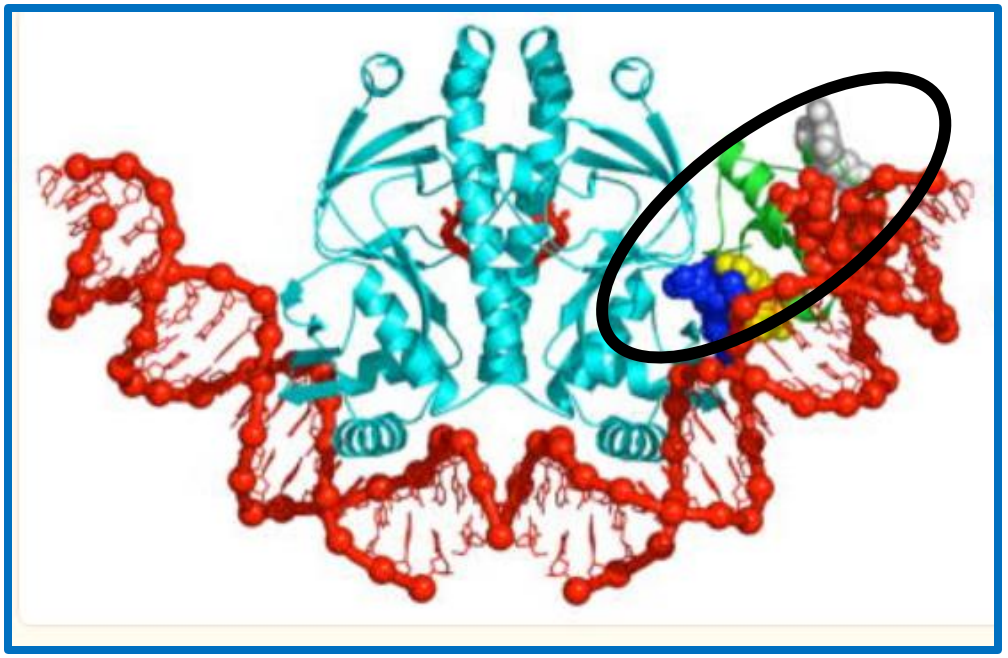
- Активаторы (пример на следю слайде)

Активатор транскрипции CAP

CAP = catabolite activator protein



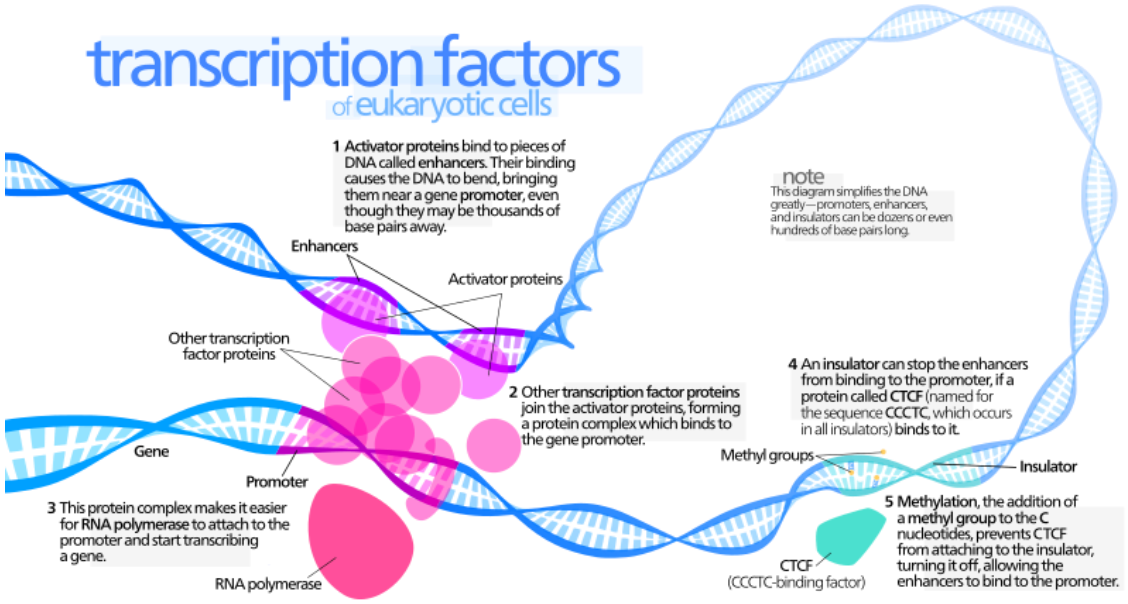


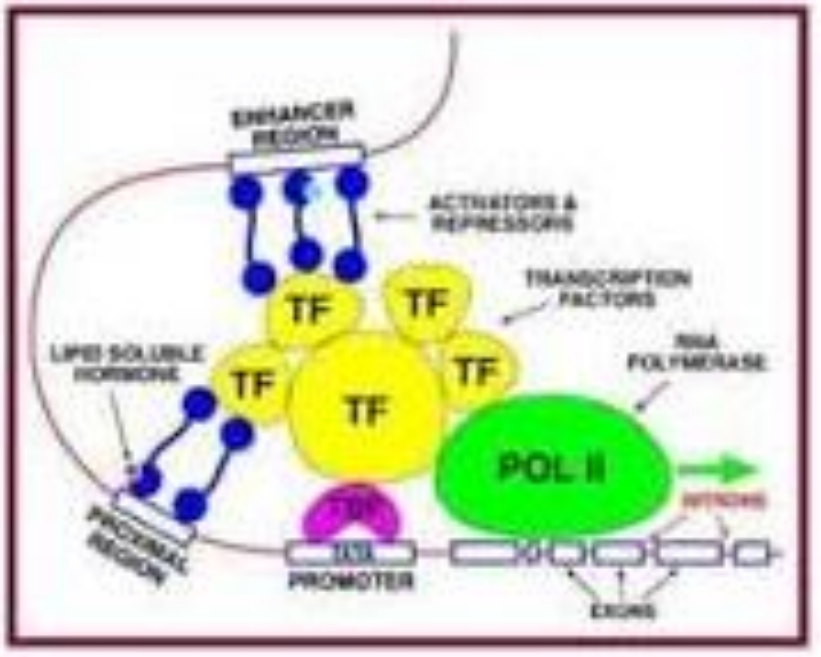


В овале С-концевой домен α-субъединицы полимеразы RNAP

Lawson CL et al., Catabolite activator protein: DNA binding and transcription activation. Curr Opin Struct Biol. 2004

Транскрипционные факторы что бывает





Энхансеры и сайленсеры у эукариот – сигналы для ТФ, расположенные за тысячи и десятки тысяч п.н. от старта транскрипции!

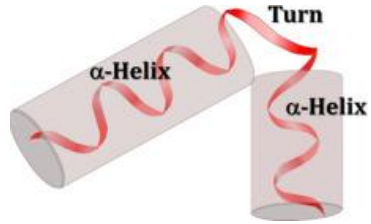
Регуляция трансляции комплексом взаимодействующих ТФ.

Распространённые 3D семейства ТФ

Примеры

Helix-turn-helix TFs

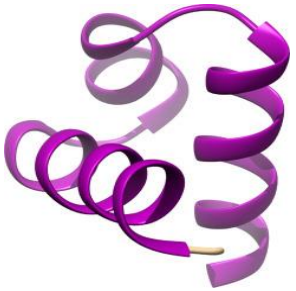
Многие ТФ для узнавания сигнала и связывания с ДНК используют структурный мотив спираль-поворот-спирал (HTH).



HTH структурный мотив

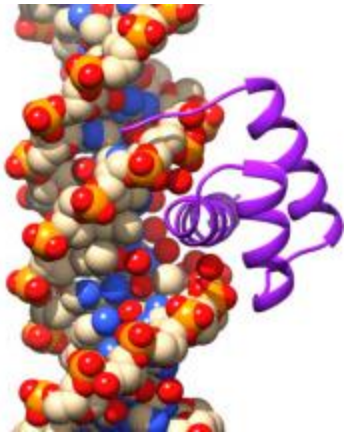
С-концевая спираль – узнающая. Онапомещается в большую бороздку ДНК.
Её а.к. остатки для узнавания сигнала должны образовать несколько (не ковалентных) связей с основаниями ДНК. Так сказать, определить код в большой бороздке!

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



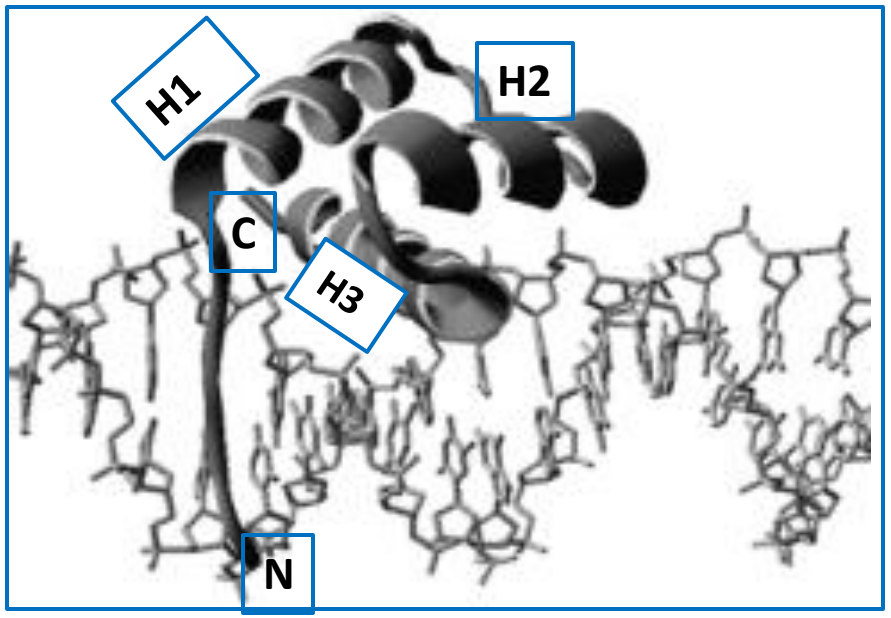
HTH структурные мотивы широко распространены в ДНК узнающих белках прокариот, эукариот и вирусов.

Они делятся на много семейств.

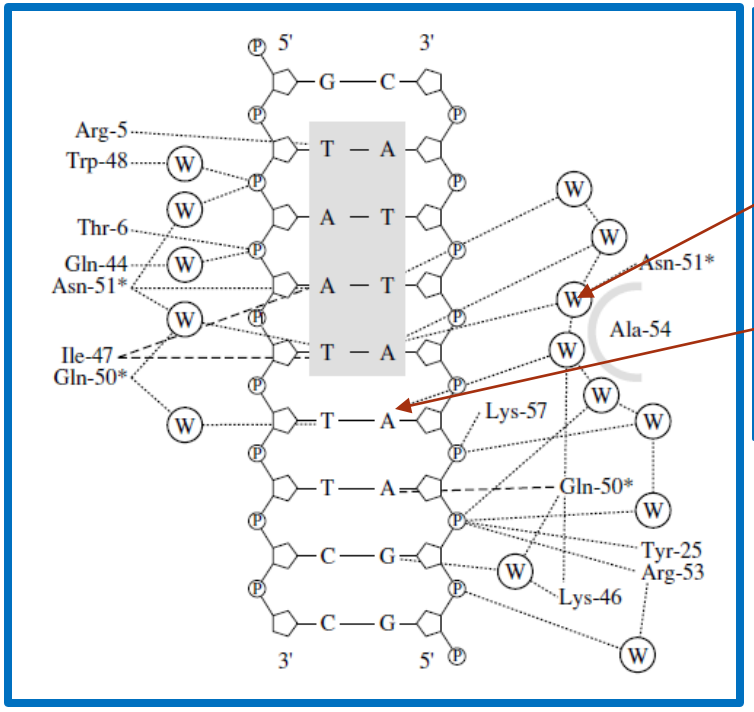
A representation of three-helix bundle class of helix-turn-helix containing proteins

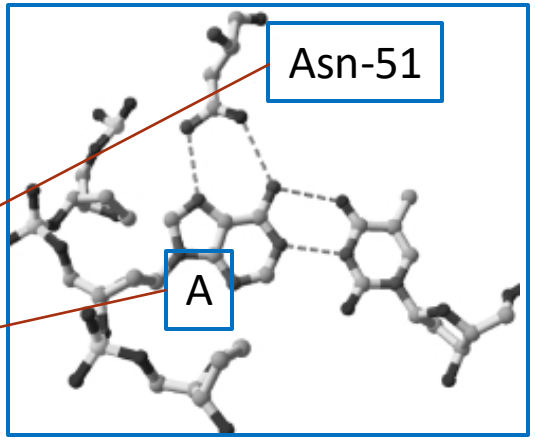
DNA structure | DNA Sequence Recognition by Proteins
K. Rutherford, G.D. Van Duyne, in Encyclopedia of Biological Chemistry (Third Edition), 2013

Гомеодомены – пример HTH Тр.Фактора.



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

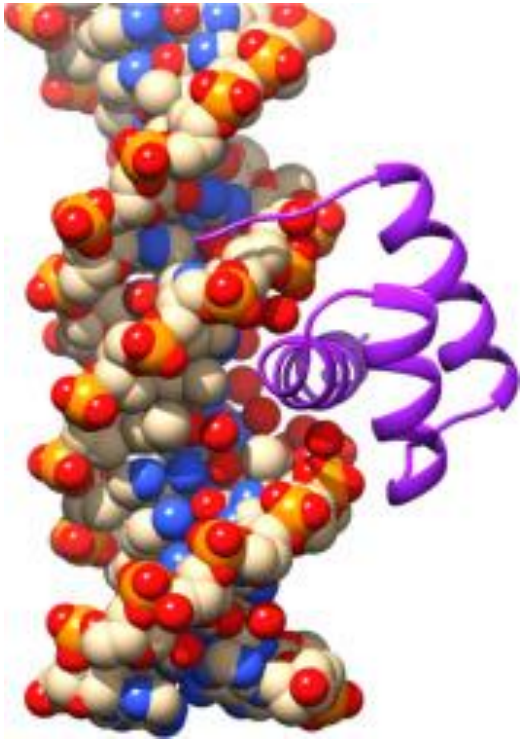


Fig. V.28. The structure of a homeodomain-DNA complex. The image was created from pdb3hdd.

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

Zinc-finger TFs

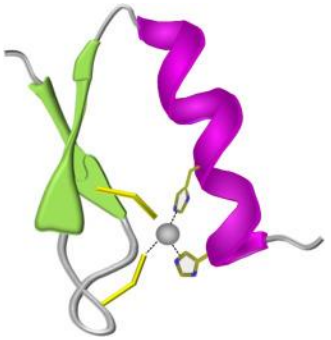
В транскрипционных факторах с цинковыми пальцами (ZF) последовательные узнающие спирали позволяют узнавать длинные сигналы, состоящие из отдельных коротких фрагментов.

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

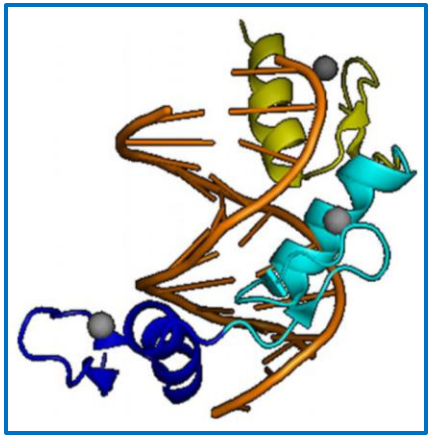
Один ZF имеет малое сродство к сигналу. Поэтому ТФ состоит из каскада ZF-элементов.

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

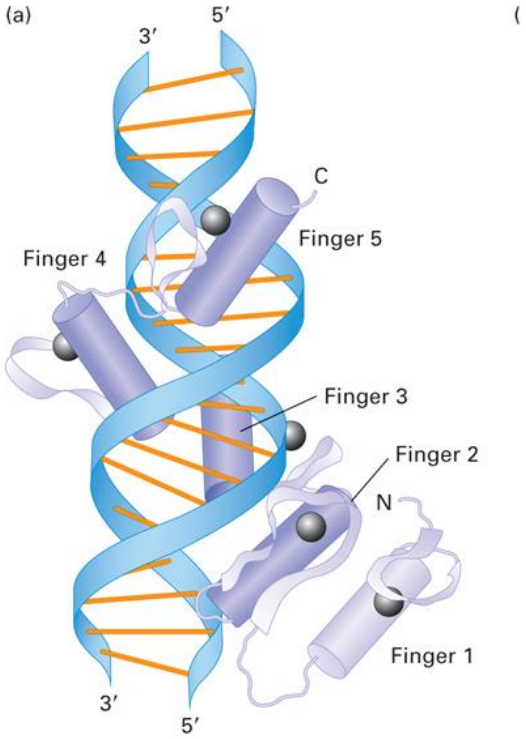


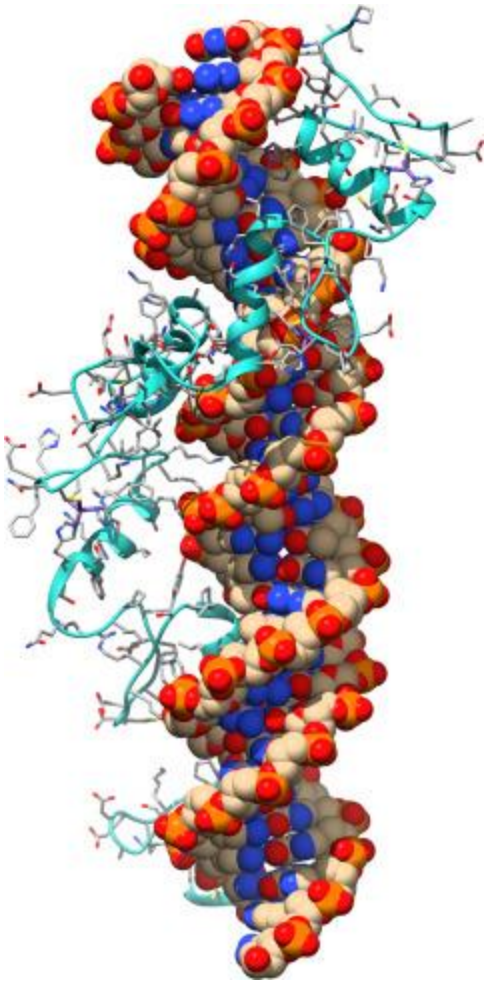
ТФ с тремя ZF

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

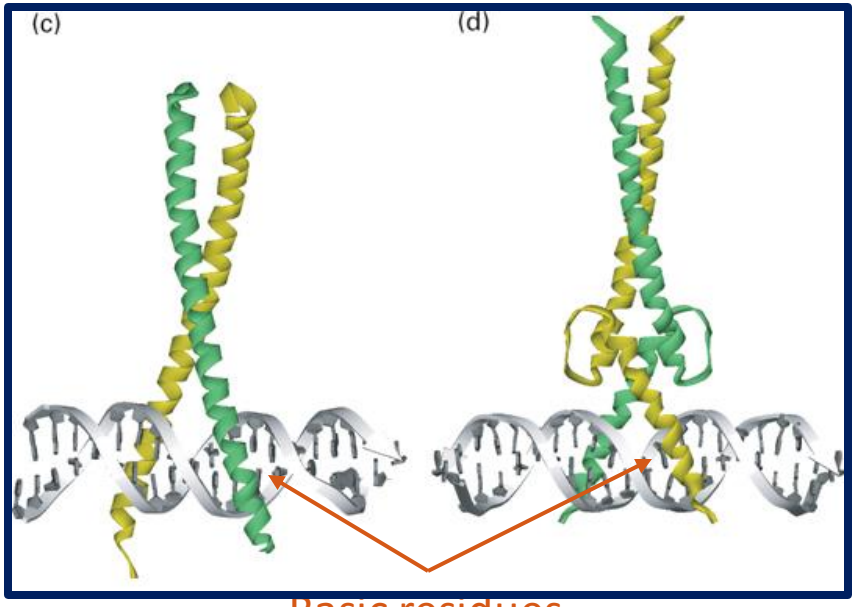
Ещё изображения ZF ТФ





Leucine-zipper TFs

Структурный мотив лейциновая молния содержит две длинные альфа-спирали в которых каждый 7й остаток –ЛЕЙЦИН. Такая периодичность приводит к образованию гидрофобной стороны альфа-спирали. Гидрофобные стороны двух спиралей обращены друг к другу, что энергетически выгодно, так как скрывает гидрофобные остатки от доступа воды. С ДНК взаимодействую N-концевые концы спиралей в большой бороздке, распознавая нужные сигнал.



Базы данных о ТФ

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

БД HOCOMOCO – наши люди

[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

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

III.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.

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
5. Программы пакета MEME строят и используют не PWM, а PFM - частотную матрицу. PWM строится переводом частот в логарифмы (описано в презентации к Л8)

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

MOTIF crp
letter-probability matrix: alength= 4 w= 19
nsites= 17 E= 4.1e-009

0.000000	0.176471	0.000000	0.823529
0.000000	0.058824	0.647059	0.294118
0.000000	0.058824	0.000000	0.941176
0.176471	0.000000	0.764706	0.058824
0.823529	0.058824	0.000000	0.117647
0.294118	0.176471	0.176471	0.352941
0.294118	0.352941	0.235294	0.117647
0.117647	0.235294	0.352941	0.294118
0.529412	0.000000	0.176471	0.294118
0.058824	0.235294	0.588235	0.117647
0.176471	0.235294	0.294118	0.294118
0.000000	0.058824	0.117647	0.823529
0.058824	0.882353	0.000000	0.058824
0.764706	0.000000	0.176471	0.058824
0.058824	0.882353	0.000000	0.058824
0.823529	0.058824	0.058824	0.058824
0.176471	0.411765	0.058824	0.352941
0.411765	0.000000	0.000000	0.588235
0.352941	0.058824	0.000000	0.588235

PFM матрица мотива

MEME version 4

ALPHABET= ACGT

strands: + -

Background letter frequencies
A 0.303 C 0.183 G 0.209 T 0.306

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☺)

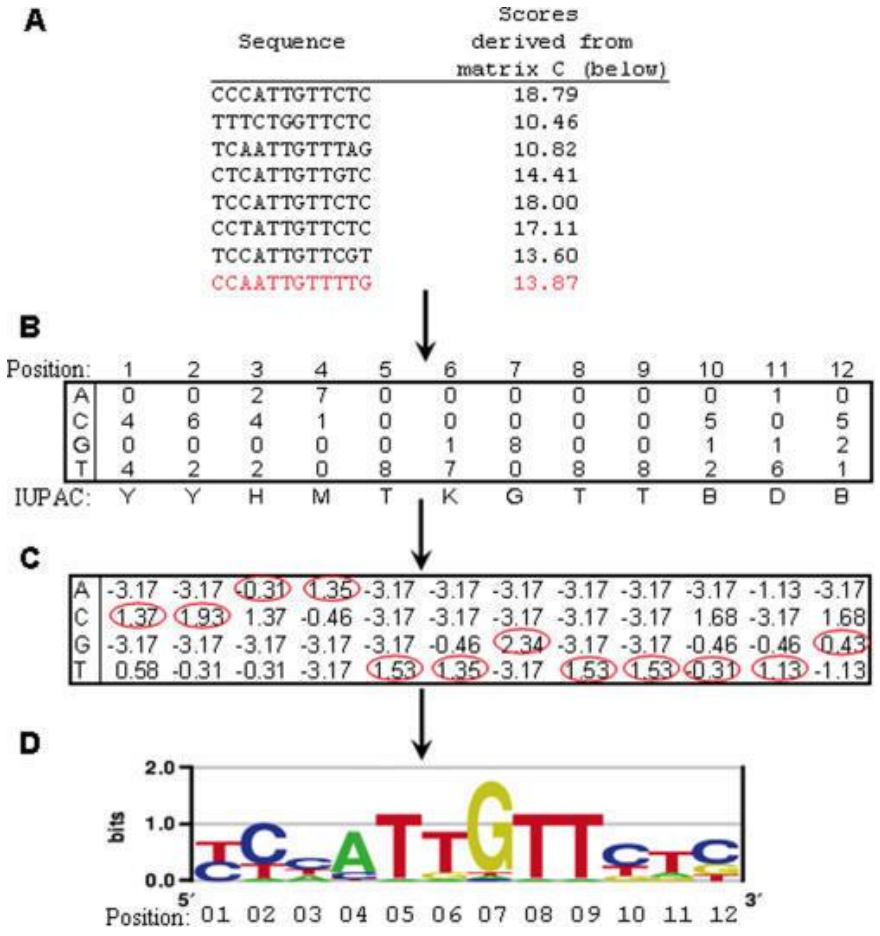
3) Как-то упустил что наши люди – коллеги - тоже сделали детектор мотивов Chipmunk (https://opera.autosome.ru/chipmunk/discovery)

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

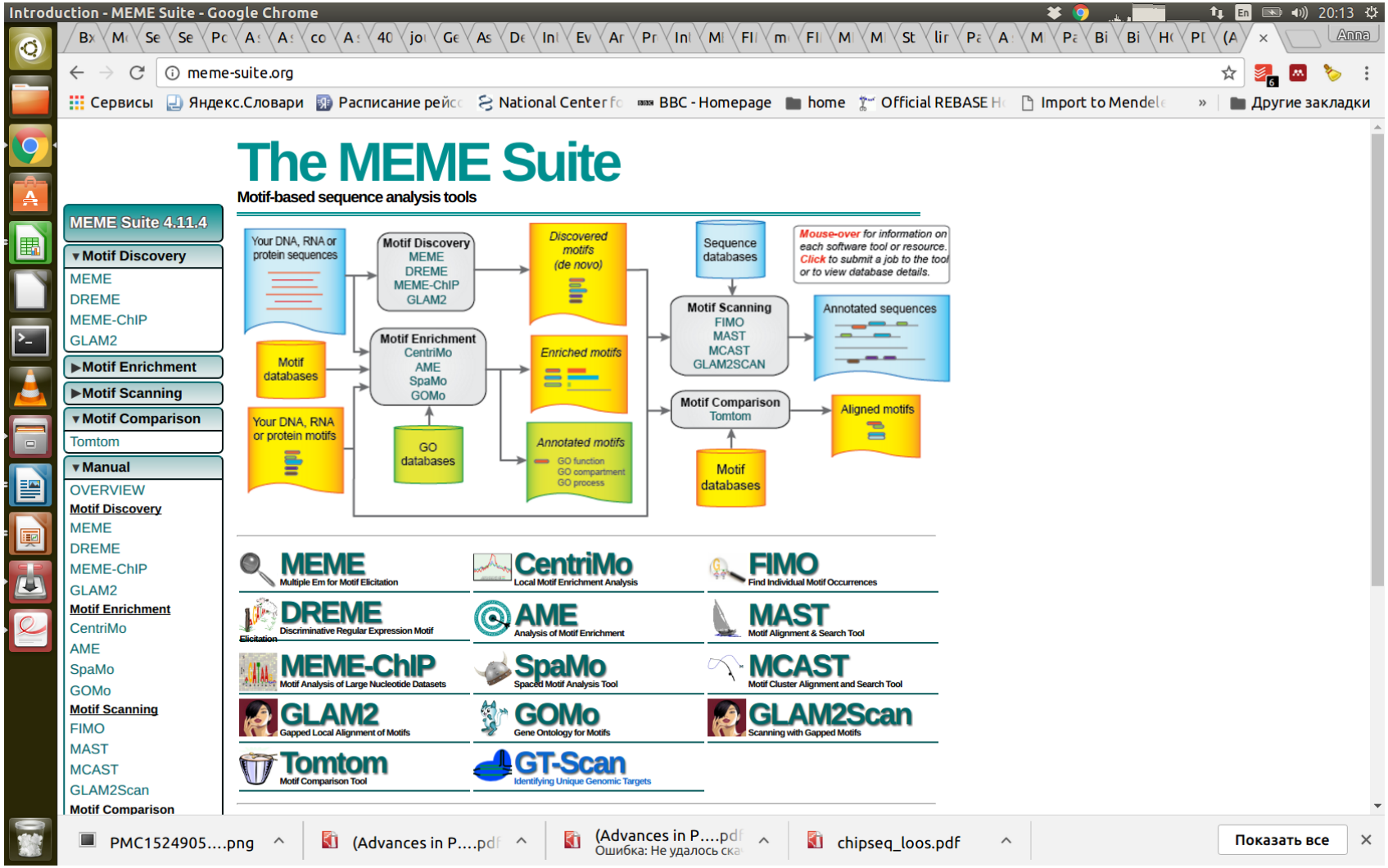


Вес (I(bj) основания b в данной позиции j
I(bj)=f(bj)*log f(bj) — p(b)*log p(b), где f(bj) — частота основания и в позиции j выравнивания, p(b) — фоновая частота основания b

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 сигналов с PWM из базы данных HOCOMOCO в геноме человека

Упражнение

- Выберите мотив TF человека из БД HOCOMOCO https://hocomoco11.autosome.org/
- Сохраните PWM этого мотива
- Найдите с помощью этой PWM несколько наиболее надежных (с маленьким P-value или большим весом) сигналов в геноме человека. Сохраните результат в bed-формате. Используйте сервис ???

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)

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)
9	**itemRgb**	RGB value in the form R,G,B (e.g. 255,0,0) determining the display color of the annotation contained in the BED file
10	**blockCount**	Number of blocks (e.g. exons) on the line of the BED file
11	**blockSizes**	List of values separated by commas corresponding to the size of the blocks (the number of values must correspond to that of the "blockCount")
12	**blockStarts**	List of values separated by commas corresponding to the starting coordinates of the blocks, coordinates calculated relative to those present in the chromStart column (the number of values must correspond to that of the "blockCount")

КОНЕЦ

IV Примеры сигналов

Для заданий практикума 7

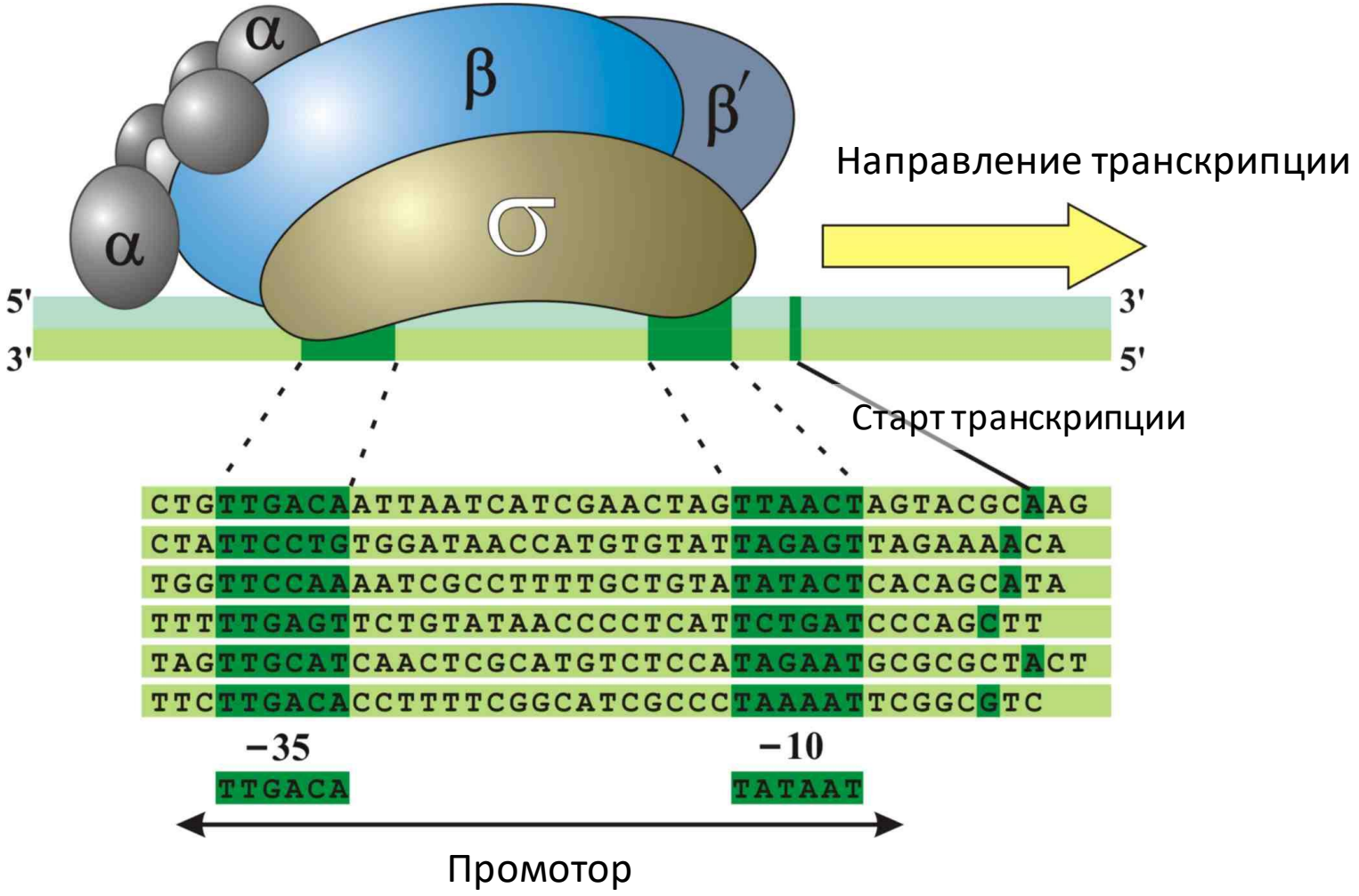
- Промотеры прокариот (инициация транскрипции)
- Сайты посадки рибосомы у прокариот (Shine-Dalgarno = SD последовательности)
- Сигналы разрывной транскрипции у коронавирусов

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

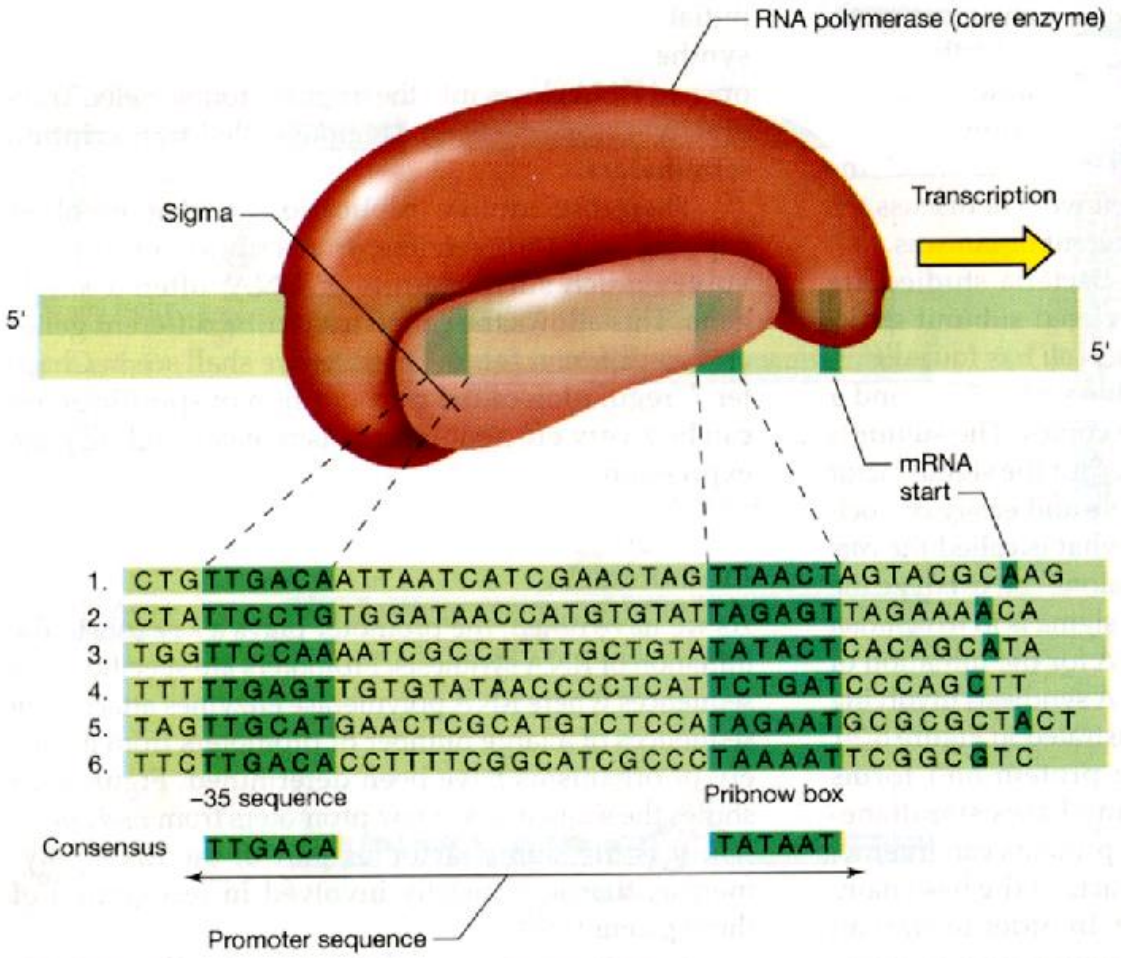
Сигналы промоторов это
- короткие последовательности ДНК, узнаваемые белком;
- расположены перед стартом транскрипции;
- похожие, но не идентичные

Схема инициации транскрипции у прокариот

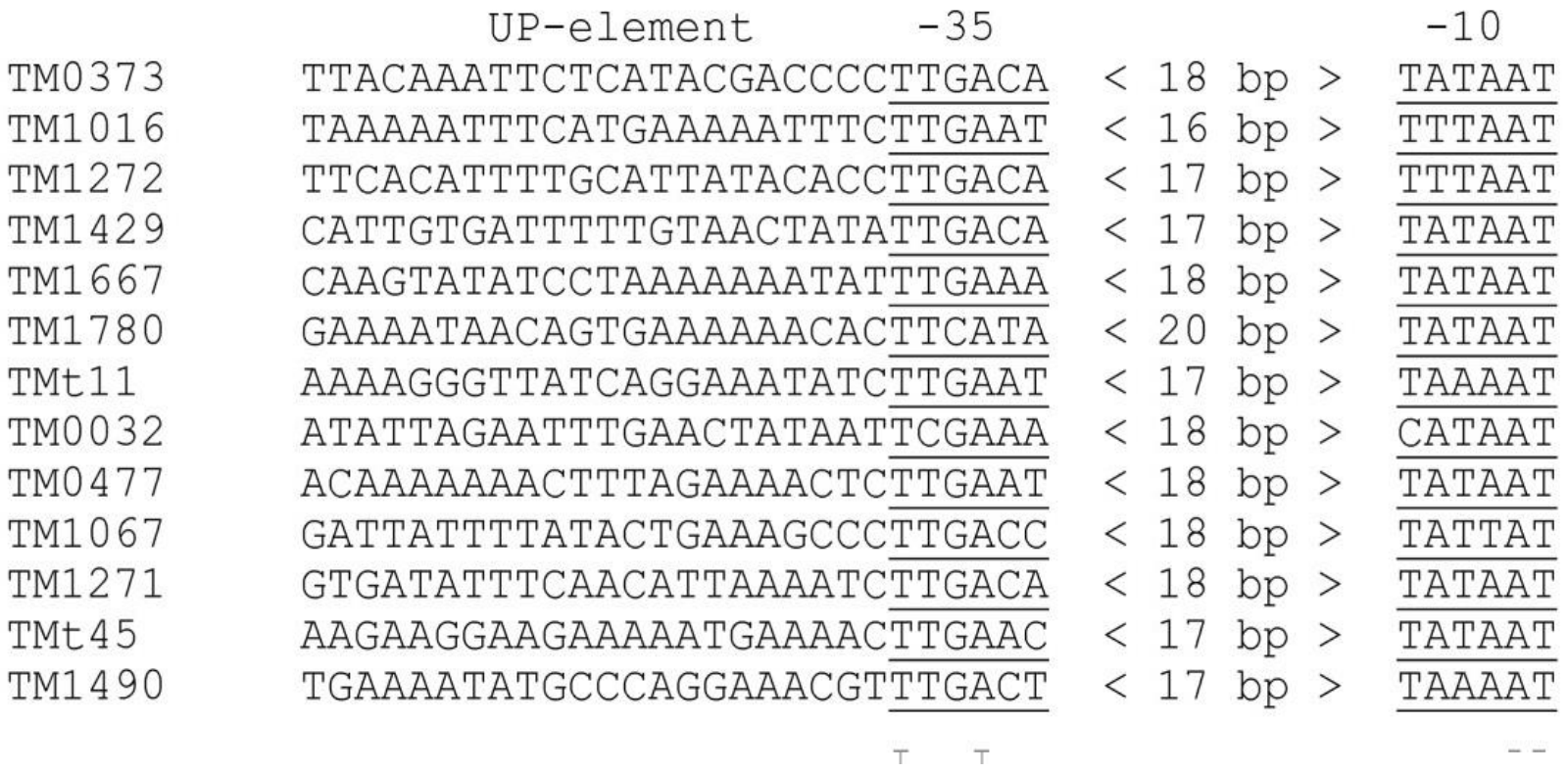


Источник: РГМ

Initiation of transcription (bacteria)



Источник: **МГ**



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)

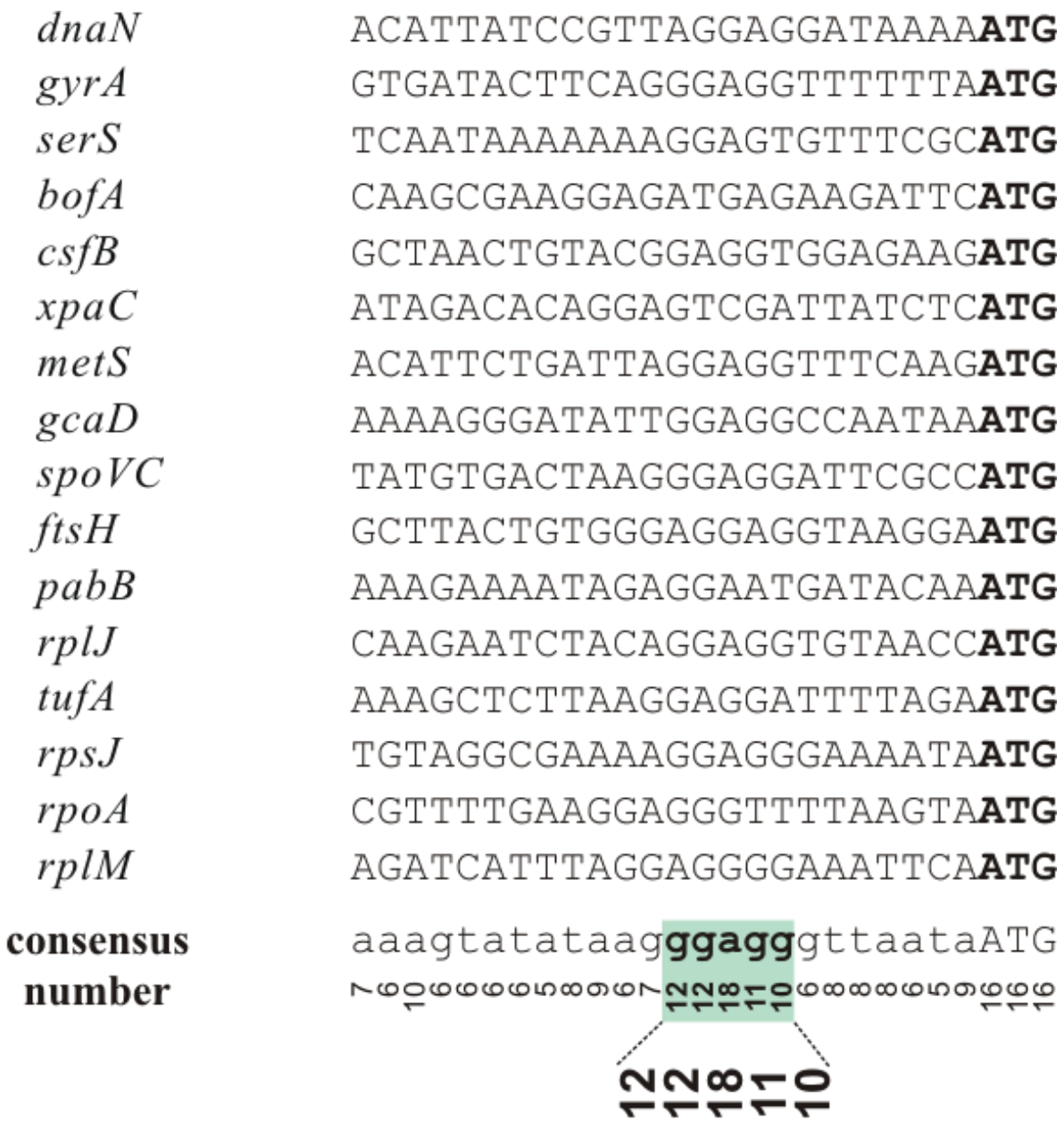
Shine-Dalgarno motifs have the consensus sequence GGAGG and can base pair with as many as nine nt in the 3’ terminal sequence of 16S rRNA (ACCUCCUUA in E. coli) referred to as the anti-Shine Dalgarno or ASD (Shine and Dalgarno, 1974).

Saito et al., 2020, eLife

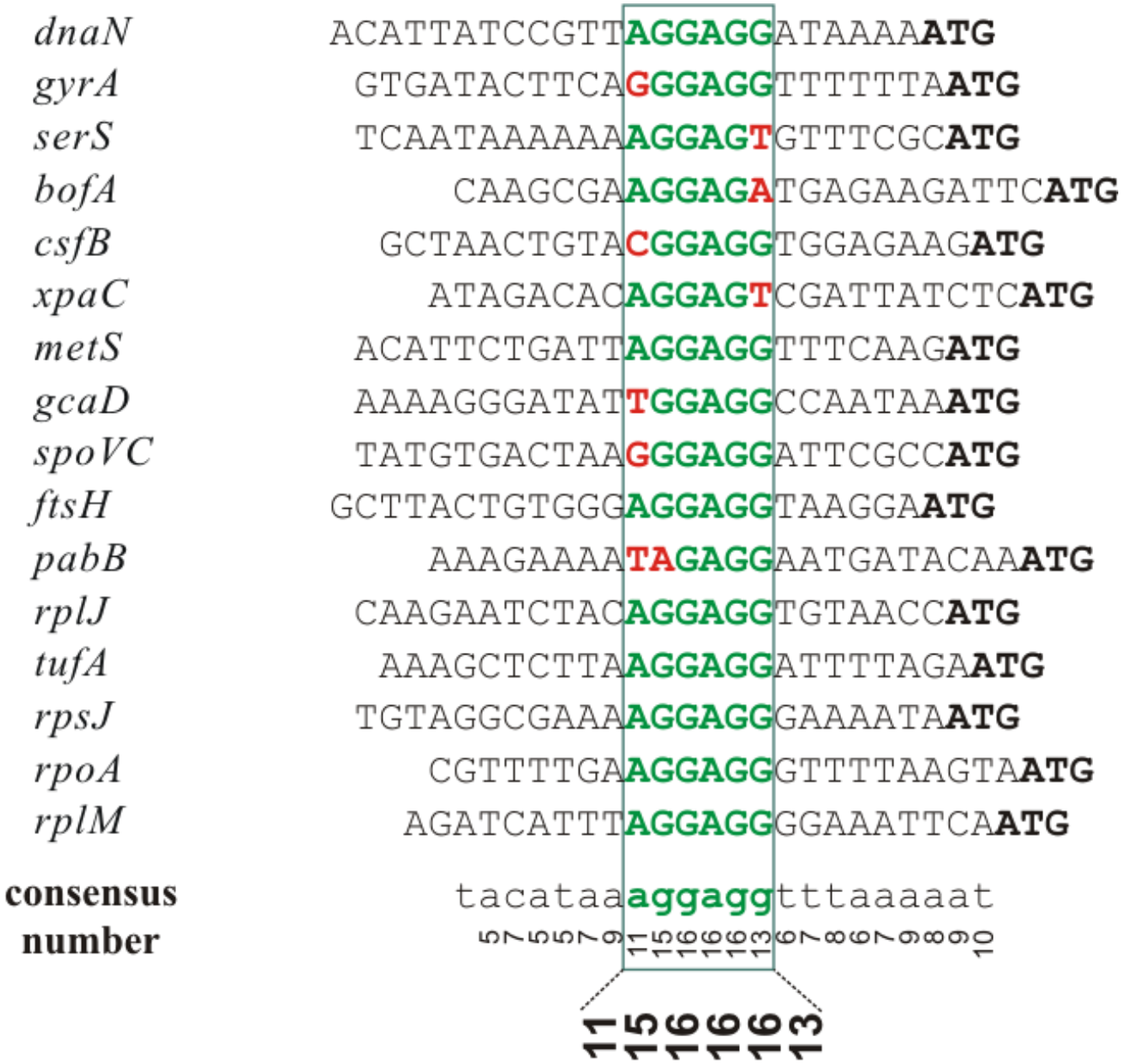
Начала генов *Bacillus subtilis*

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.

Задания

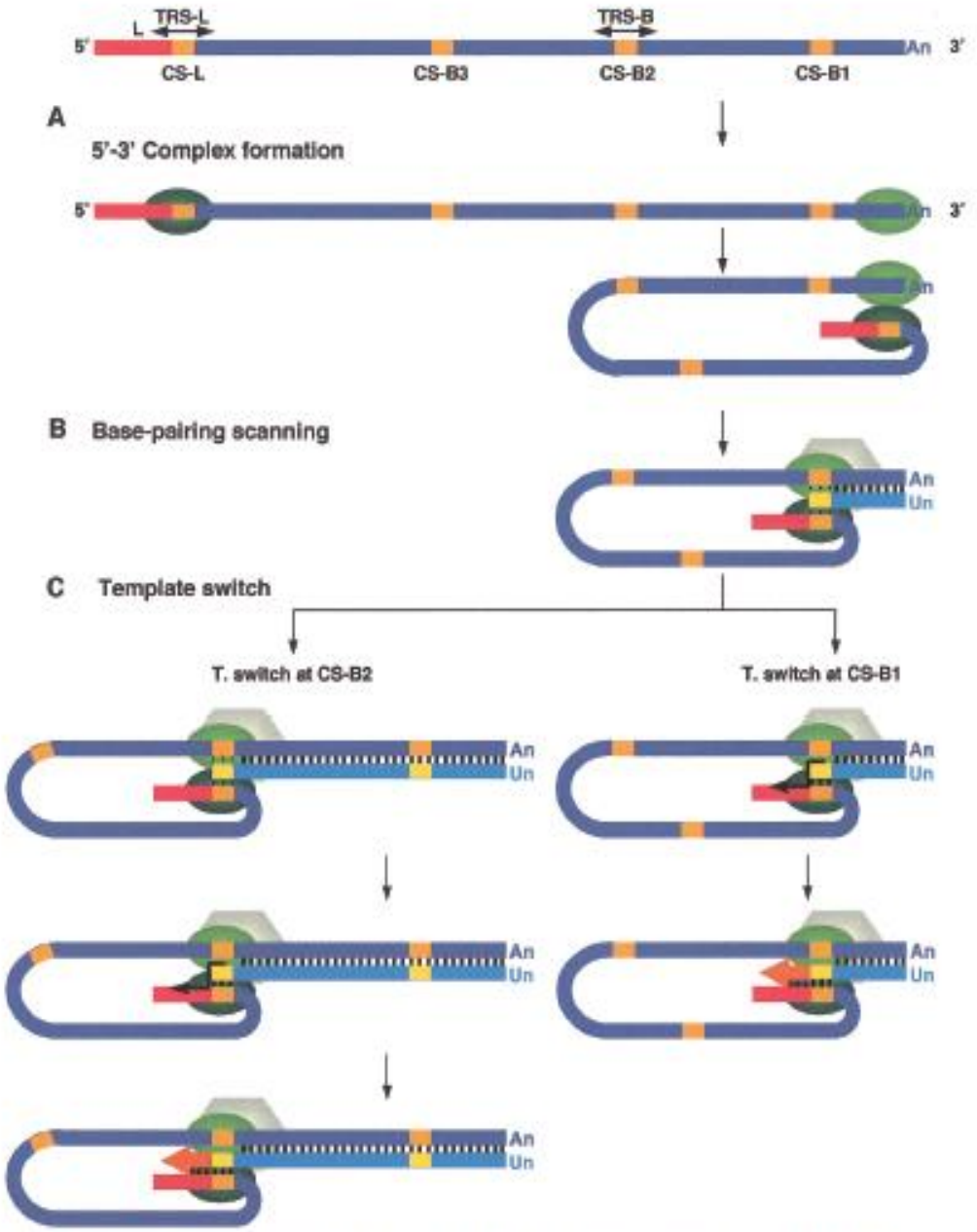
КОНЕЦ

презентации

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

КОНЕЦ

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),