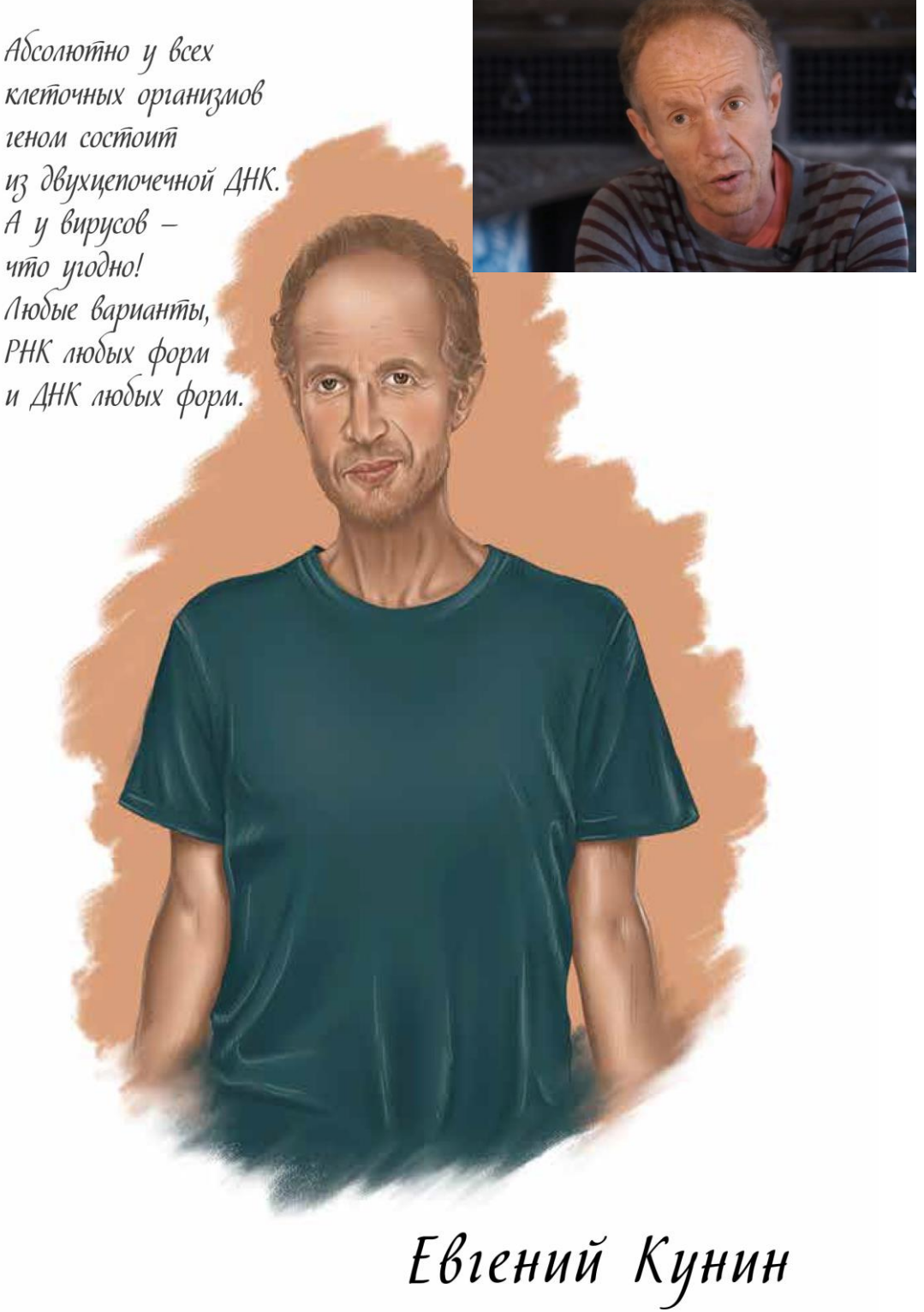
1. Введение

Геном и информация

Геном

- совокупность наследственной информации <u>о живом</u>
- Живое умеет:
  - размножаться (это его жизненная установка)
  - копировать свой геном потомкам
  - с ошибками => эволюция этим отличаемся от компьютерных вирусов)))
  - Стареть и умирать



Размножение

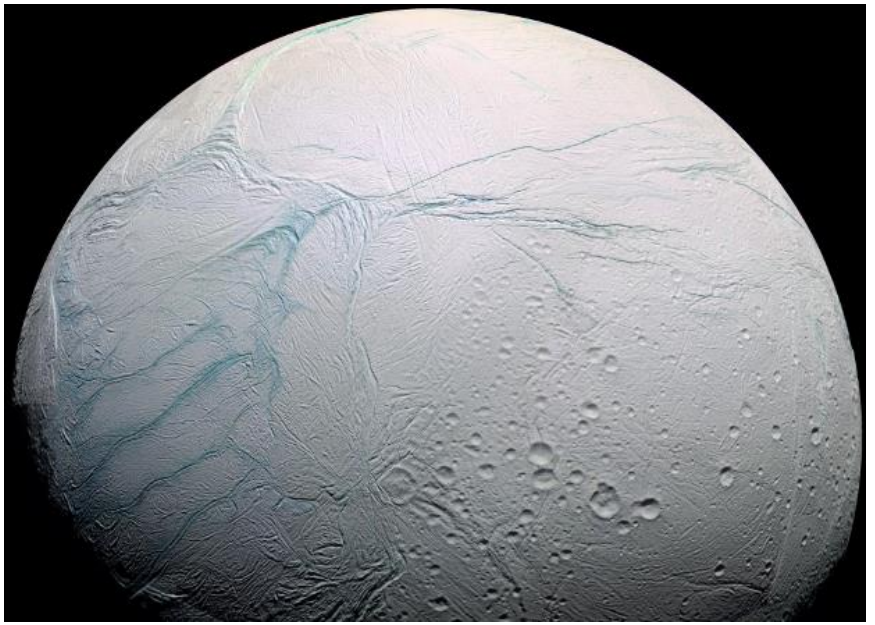
- *эу***кариоты**
  многоклеточные эу это мы: папа + мама => ребёнок
  одноклеточные эу: удвоение ядер и деление,
  несколько способов.

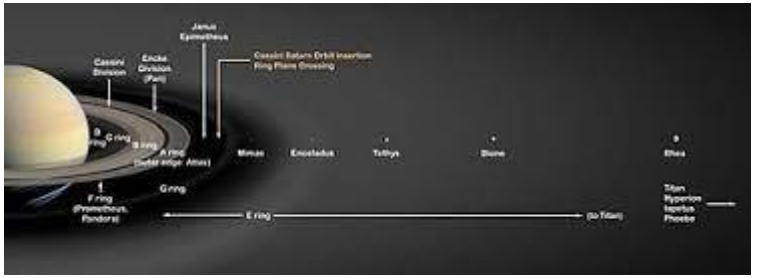
- *про***кариоты**
  бактерии
  археи
  р*епликация*: клетка делится на 2, удвоение генома

- Клетка многоклеточных эукариот – живая!
  (кроме тех, которые не делятся) размножается как прокариоты. *Еще и про митохондрии вспомним*

- Вирусы – вообще, анекдот)))
  и сами не живут, и другим не дают жить здоровыми

К вопросу об археях





[1] Taubner at al., Biological methane production under putative Enceladus-like conditions. Nat Commun. 2018

Информация в геноме

Информация в геноме

- **Гены белков** – участки ДНК, кодирующие химическую формулу без модификаций (=аминокислотную последовательность) белка.

Информация в геноме

- **Гены белков** – участки ДНК, кодирующие химическую формулу без модификаций (=аминокислотную последовательность) белка.

По Ф. Энгельсу:
«Жизнь есть способ существования белковых тел»

Информация в геноме

- **Гены белков** – участки ДНК, кодирующие химическую формулу без модификаций (=аминокислотную последовательность) белка.

По Ф. Энгельсу:
«Жизнь есть способ существования белковых тел»

- **Гены молекул РНК**
  - мРНК рРНК тмРНК тРНК
  - Рибозимы, некодирующие РНК, в т.ч. малые РНК

Информация в геноме

- **Гены белков** – участки ДНК, кодирующие химическую формулу без модификаций (=аминокислотную последовательность) белка.

По Ф. Энгельсу:
«Жизнь есть способ существования белковых тел»

- **Гены молекул РНК**
  - мРНК рРНК тмРНК тРНК
  - Рибозимы, некодирующие РНК, в т.ч. малые РНК
- **Как закодирована информация о**:
  - Поведении пчелиного роя (сколько мёда достаточно, чтобы пережить зиму, как вести себя при морозе и др.)
  - Том, когда белке начинать строить гнездо
  - Половом влечении в определенном периоде жизни
  - Высиживание яиц скопой – мамой и папой

2. Сигналы в ДНК и РНК

Закодированные последовательностями нуклеотидов

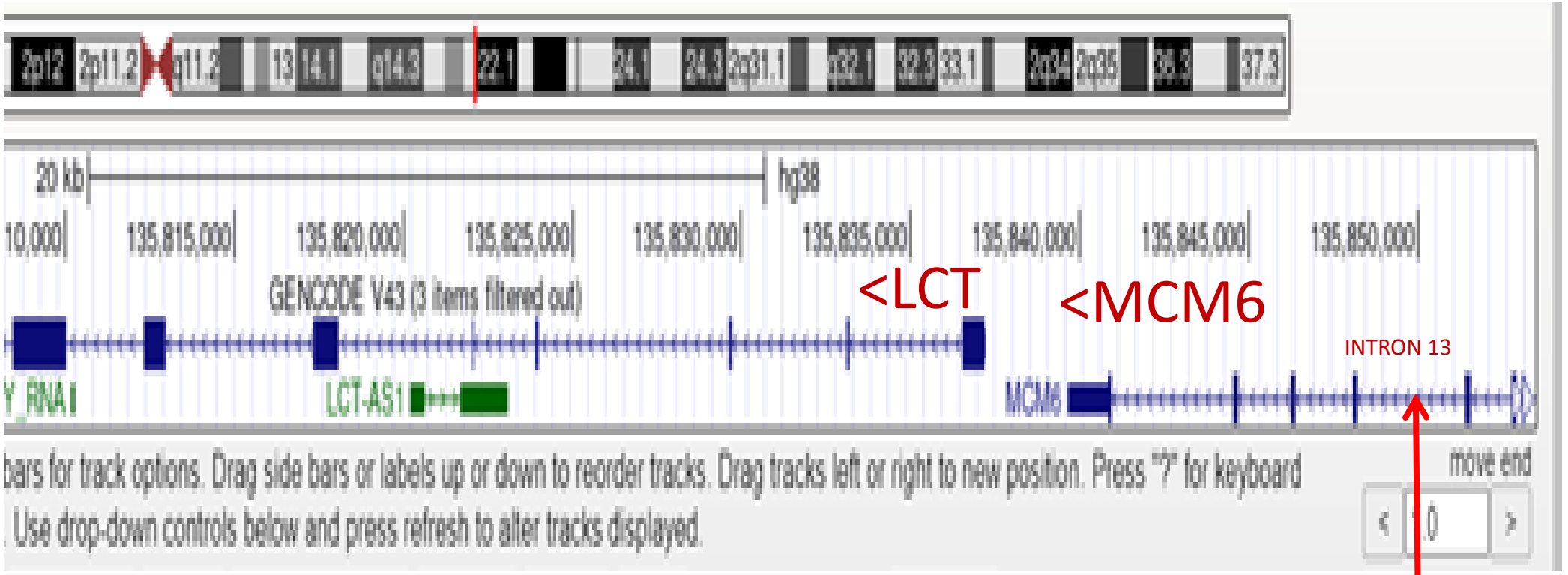
Сложный пример

Один вариант лактазной недостаточности у людей

- Лактаза - ген LCT белок LCH_HUMAN
  Лактаза необходима для гидролиза лактозы из молока.
  Лактаза работает у грудничков.
  Если не экспресируется у взрослых, то серьёзные проблемы с потреблением молока.
- Распространенный вариант заболевания связан с мутацией в гене MCM6
- Я ожидал, что всё просто – белок MCM6_HUMAN регулирует экспрессию LCT у взрослых. И расскажу, что в промоторе LCT есть сигнал для связывания белка MCM6.
  Оказалось – всё не так просто!!!
- Заболевание определяется точечным полиморфизмом в ИНТРОНЕ!!! гена MCM6, находящегося перед геном LCT. Значит, белок MCM6 тут не при чем.
  Полиморфизм расположен за 13 910 bp перед стартом транскрипции LCT
- Значит, сигнал влияющий на уровень транскрипции LCT гена во взрослом состоянии находится в некодирующей последовательности в предшествующем участке ДНК

- Всё доказано экспериментально – на клетках человека с тем или иным полифоризмом [2], но статьи с объяснением того через кого сигнал передаётся LCT не обнаружил.

[2] Olds and Sibley, Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element, 2003



Сайленсеры и энхансеры – сайты связывания факторов транскрипции, удалённые на значительное расстояние от старта транскрипции

Позиция 13910 п.н. до старта транскрипции гена лактозы LCT. В сайленсере (гипотеза) гена LCT

T – переносимость лактозы у взрослого

C – непереносимость лактозы у взрослого (гомозигота)

Вопрос. Как проверить гипотезу с помощью компьютера?

Предположение.
Если полиморфизм в сайте, связываемом белком – транскрипционным фактором, то он будет похожим в геномах родственников

Трудно, но попробовать можно.

Простой пример

Сигнал системе рестрикции-модификации EcoRI:

GAATTC в геноме E.coli

Сигнал адресован системе рестрикции-модификации EcoRI (белки R и M)

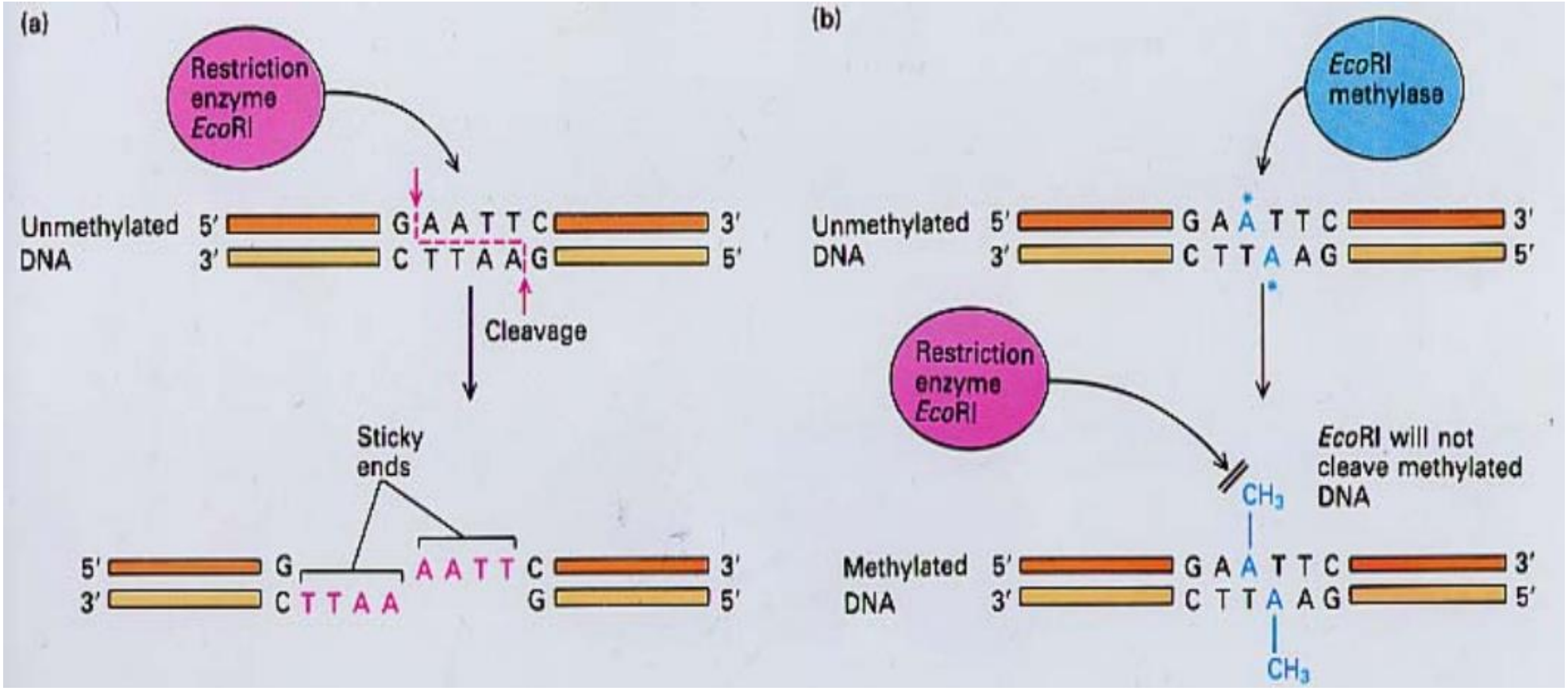
Сигнал GAATTC бывает в 2х состояниях

(1) не метилирован

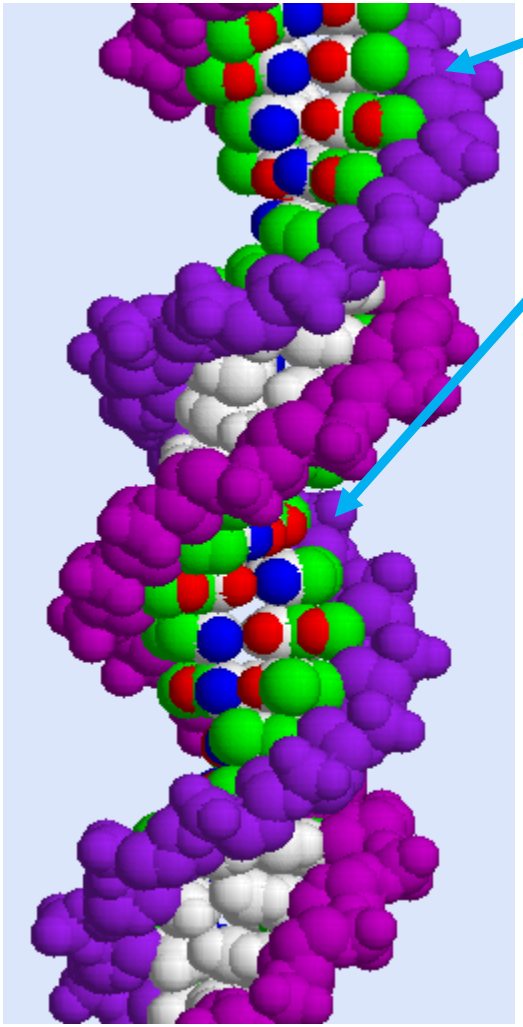
(2) метилирован по двум цепочкам ДНК метилтрасферазой M

При репликации кратковременно полуметилирован.

Предназначен для отличения и расщепления чужой ДНК, и не расщепления своей. Гидролизует обе цепочки ДНК эндонуклеаза R



Белки читают последовательность ДНК по большой бороздке, не расплетая цепочки ДНК



Двойная спираль ДНК.

Раскраска моя ААл ☺

Глядя на рисунок легко представить себе почему в последовательностях сайтов ДНК, связываемых одним белком (и его близкими гомологами) не может быть делеций!

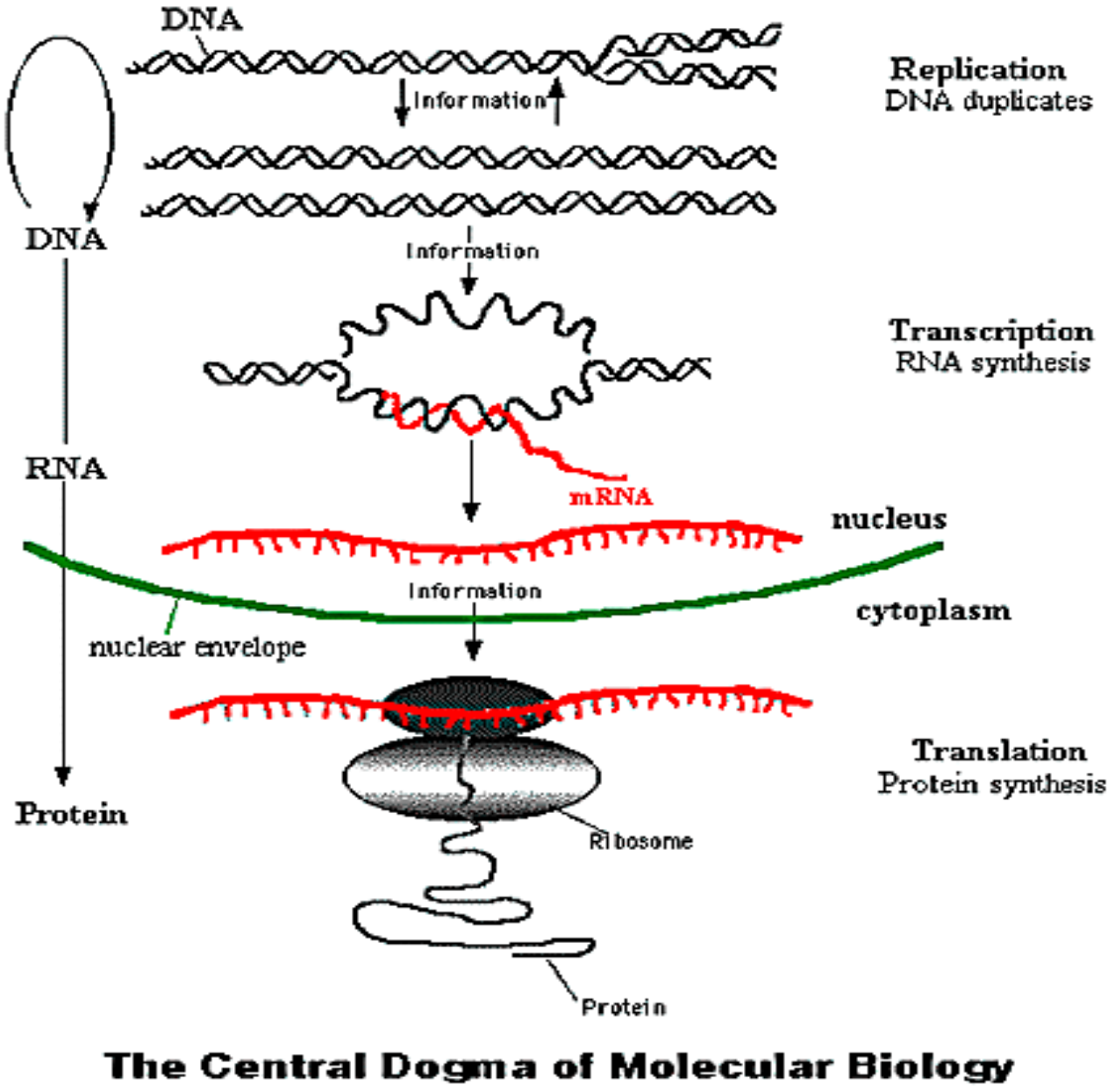
Сигнал, по существу, трехмерный. К тому же, известно, что конформация остова ДНК немножко зависит от последовательности оснований

Этот сигнал определен экспериментально в 1971 году в Phd thesis автор Yoshimori R.M. В университете Сан-Франциско

Найти все встречи этого сигнал в нуклеотидной последовательности (геноме) сумеет любой студент 2го курса ФББ. Так?

Надо бы задание дать: посчитать число сайтов GAATTC в полном геноме одного штамма E.coli и определить насколько и в какую сторону оно отличается от ожидаемого по статистике, насколько отличается и достоверно ли отличается.

Сигналы процессов передачи генетической информации



The Central Dogma of Molecular Biology

Какие сигналы нужны?

Designer Chang

Что надо узнать про сигнал.

1. Сигнал чего?
2. Носитель сигнала, что из себя представляет сигнал
3. Название сигнала
4. Сигнал кому, кто воспринимает сигнал и реагирует на него?

Какие сигналы нужны для:

- Репликации

  ……

  ……
- Транскрипции

  ……

  ……
- Сплайсинга у эукариот

  ……

  ……
- Трансляции мРНК

  ……

  ……

Репликация у бактерий

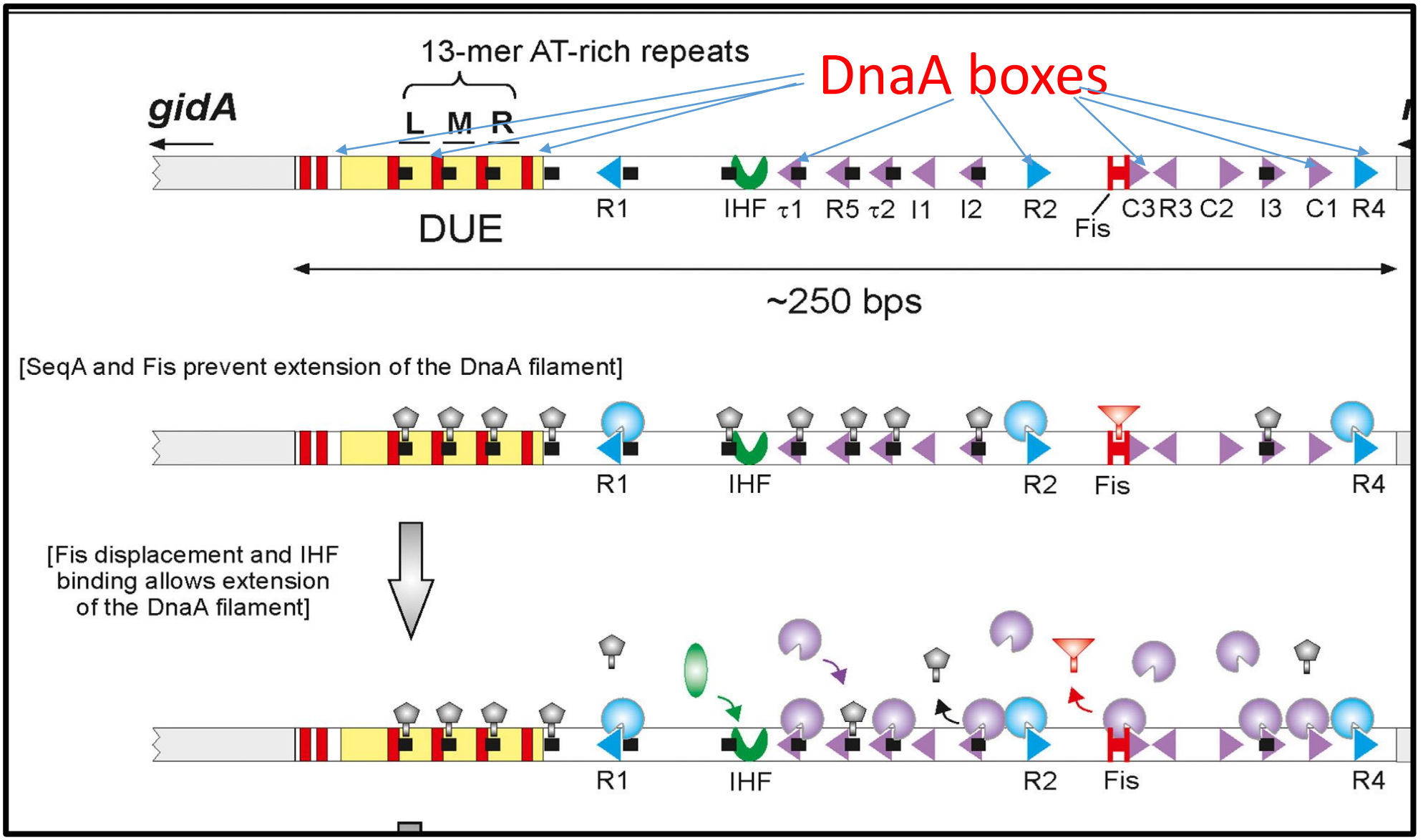
1. **Место начала репликации - ориджин**
   Origin of replication

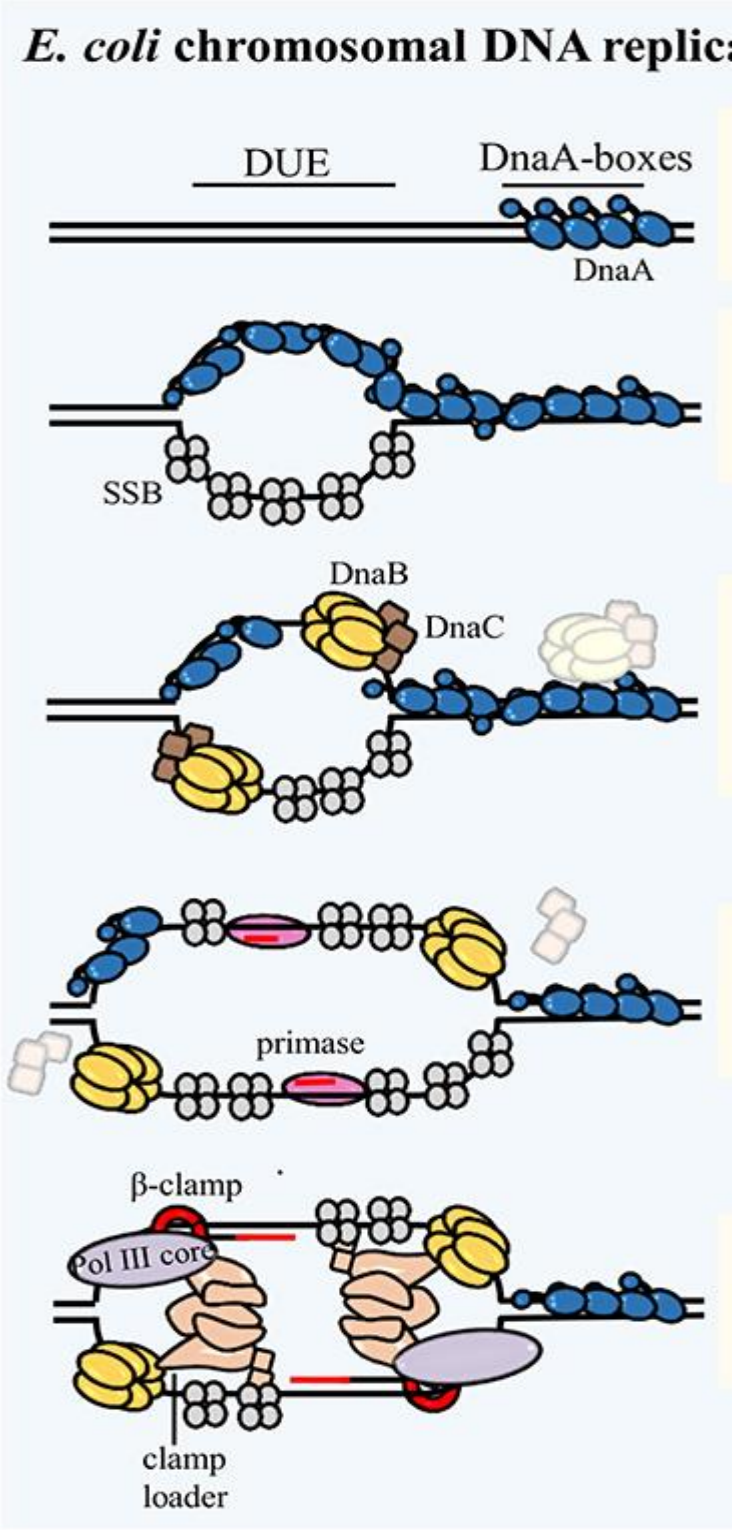
   Участок ДНК с несколькими сайтами определенной последовательности.
   oriC region ≈250 п.н. См. СЛЕДУЮЩИЙ СЛАЙД

   Белки DnaA – первыми связываются со своими сайтами (DnaA boxes)

   Replisoma - комплекс, состоящий из 15—20 различных белков.
   Инициируется белками DnaA, связанными с ДНК.

Как решается вопрос когда пора делиться? (не знаю)


13-mer AT-rich repeats
DnaA boxes
gidA
L M R
DUE
R1
IHF
τ1
R5
τ2
I1
I2
R2
Fis
C3
R3
C2
I3
C1
R4
~250 bps
[SeqA and Fis prevent extension of the DnaA filament]
[Fis displacement and IHF binding allows extension of the DnaA filament]
protein binding sites
high- / low- affinity DnaA boxes
Fis binding site
SeqA binding site
IHF binding site
6-mer ssATP-DnaA boxes
oriC binding proteins
Fis
IHF
SeqA
DnaA
+ATP or -ATP
+ATP



Из
[3] Wegrzyn et al.,
Replisome Assembly at Bacterial Chromosomes and Iteron Plasmids. Front Mol Biosci. 2016

Область oriC вариабельна у бактерий.
Общее у всех – три функциональных участка

(1)Кластер сайтов связывания белка DnaA (DnaA boxes)

(2)Участок DUE (DNA unwinding element) A-T богатый

(3)Последовательности, узнаваемые другими регуляторными белками

[4] Wolanski et al., oriC-encoded instructions for the initiation of bacterial chromosome replication, 2015

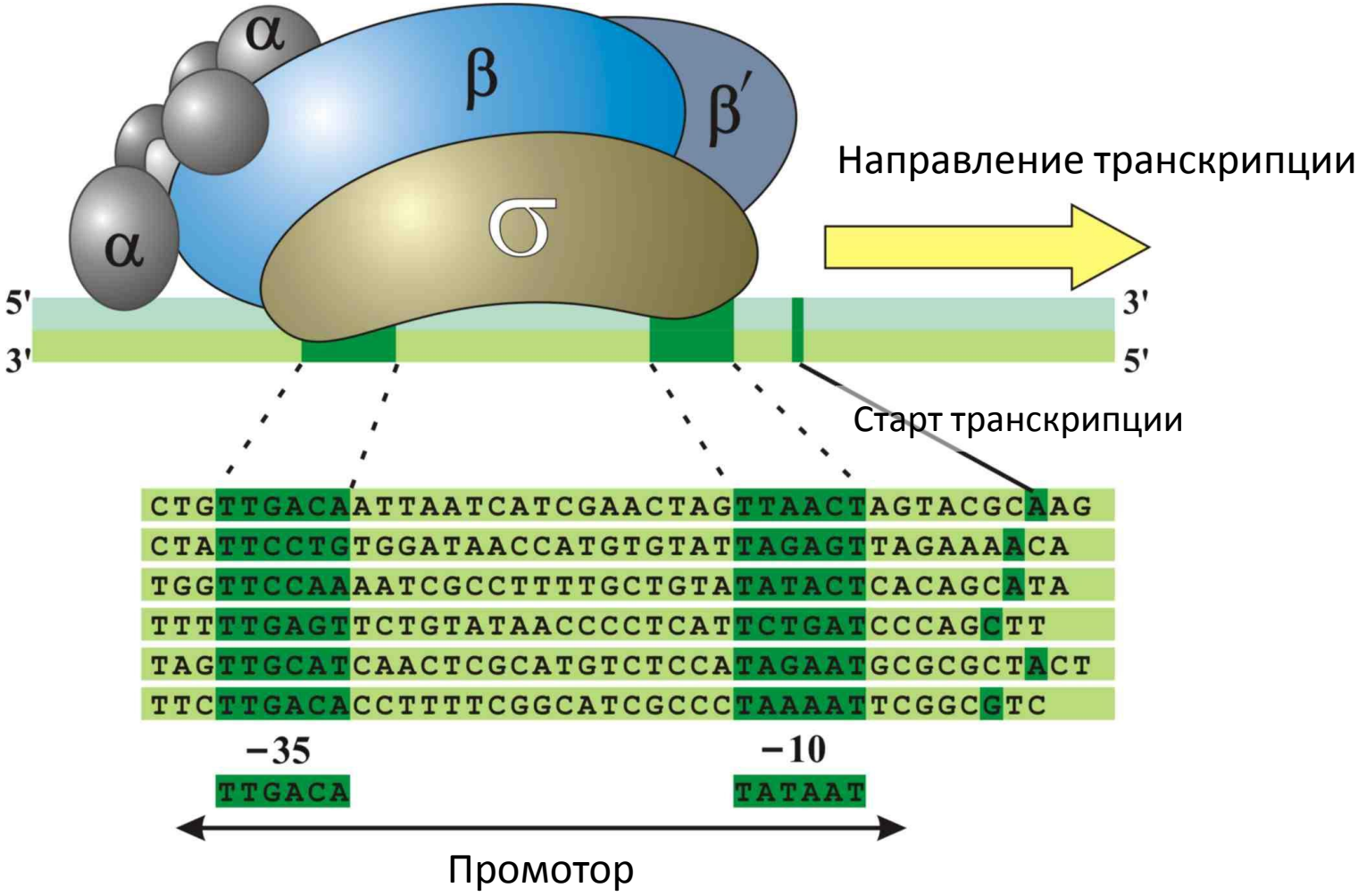
Вопрос: последовательности сайтов DnaA

Одинаковы ли у штаммов одного вида (напр.E.coli)

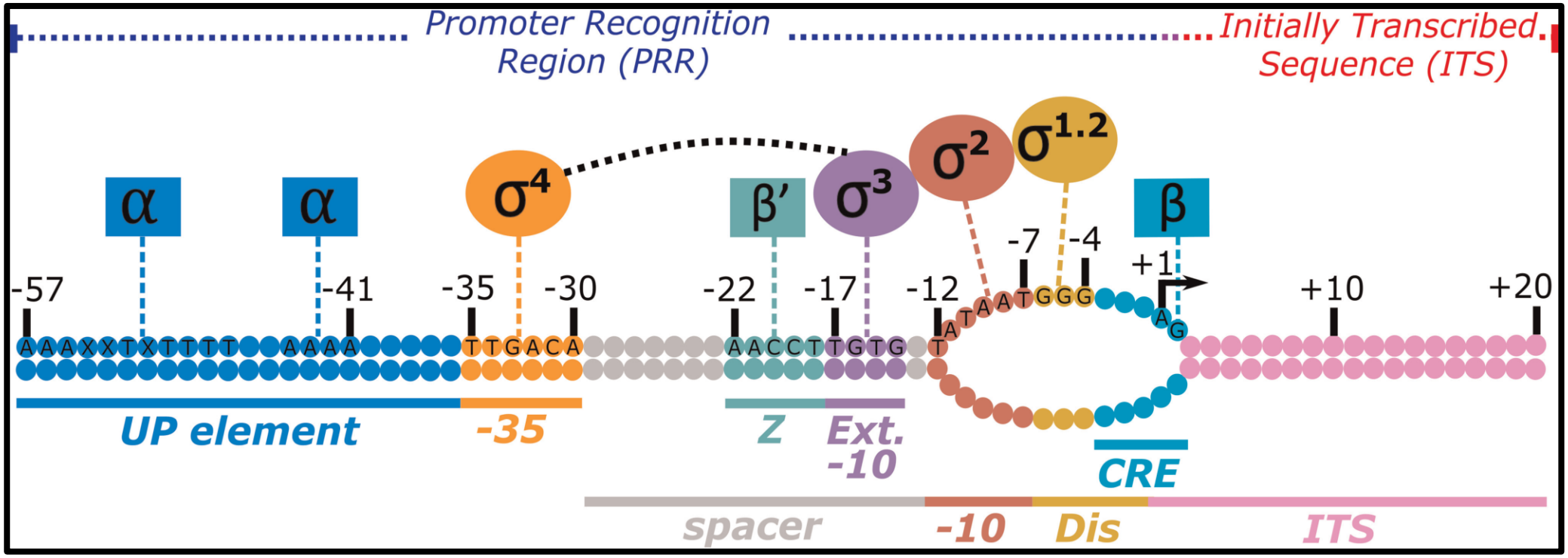
Старт транскрипции

- Промотор: последовательность ДНК, узнаваемая белками для инициации транскрипции
- Для начала транскрипции на промоторе должен собраться комплекс – РНК полимераза – состоящая из нескольких субъединиц
- Первой с промотором связывается σ-субъединица RNAP

Схема инициации транскрипции у прокариот



Источник: РГМ



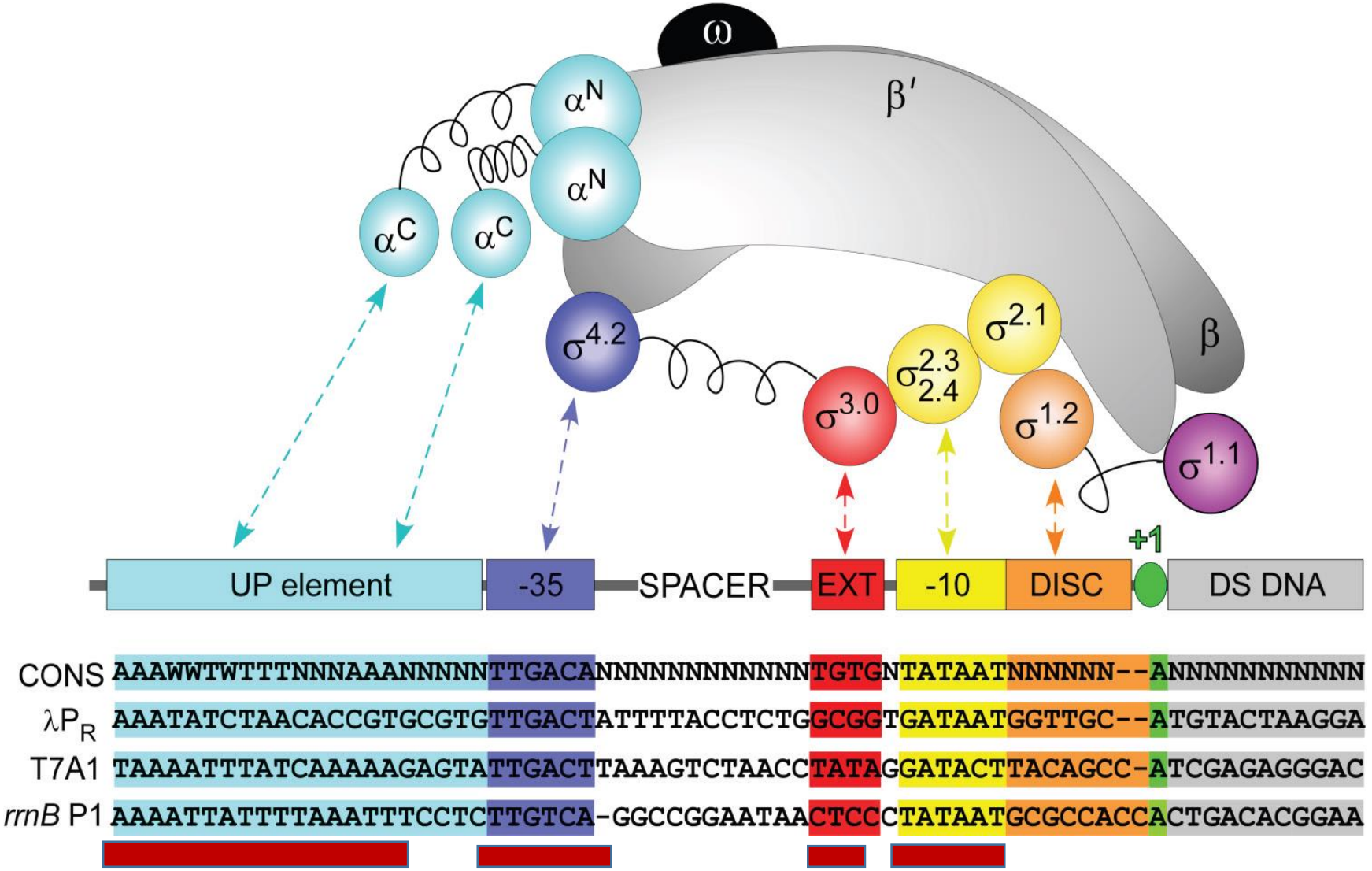
Обозначения $\sigma^{1.2}, \sigma^{2}, \sigma^{3}, \sigma^{4}$ - домены σ-субъединицы RNAP

α и β, β' - соответствующие субъединицы RNAP.

upstream (UP) element

[5 Jensen and Galburta, The Context-Dependent Influence of Promoter Sequence Motifs on Transcription Initiation Kinetics and Regulation, 2021

Консенсус для промоторов 3х генов

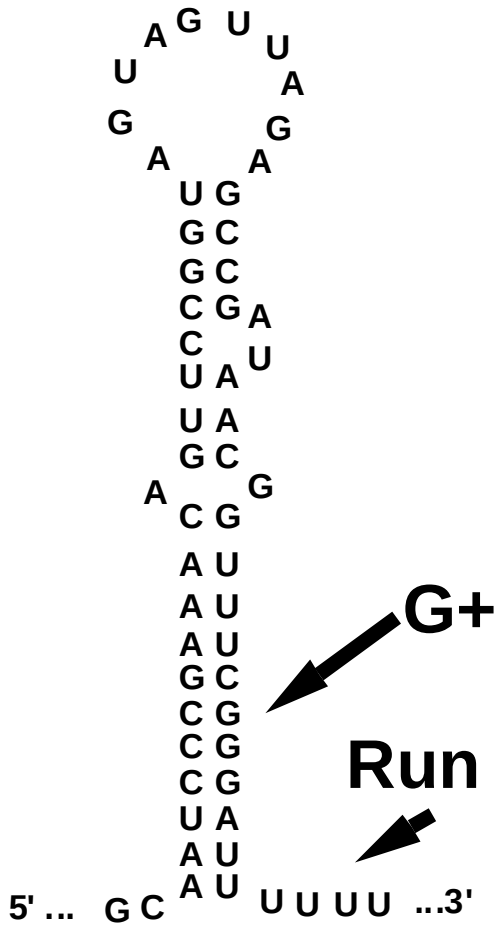


Терминация транскрипции

- Прокариоты
  - Rho-зависимая терминация
  - Rho – независимая
- эукариоты

Неплохой текст в википедии про это
https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D1%80%D0%BC%D0%B8%D0%BD%D0%B0%D1%82%D0%BE%D1%80_(%D0%BC%D0%BE%D0%BB%D0%B5%D0%BA%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%B0%D1%8F_%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F)

Termination of transcription in *E. coli*: Rho-**independent** site



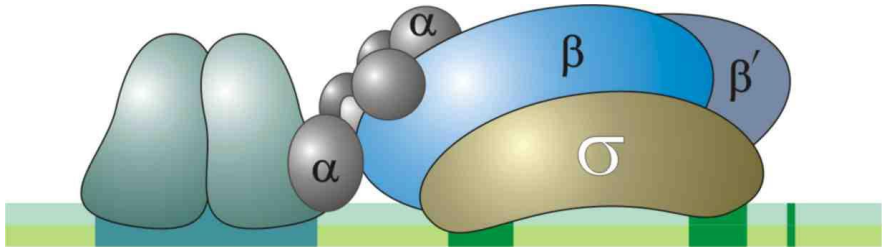
Сигналом является вторичная структура РНК – шпилька, а не последовательность

Регуляция транскрипции

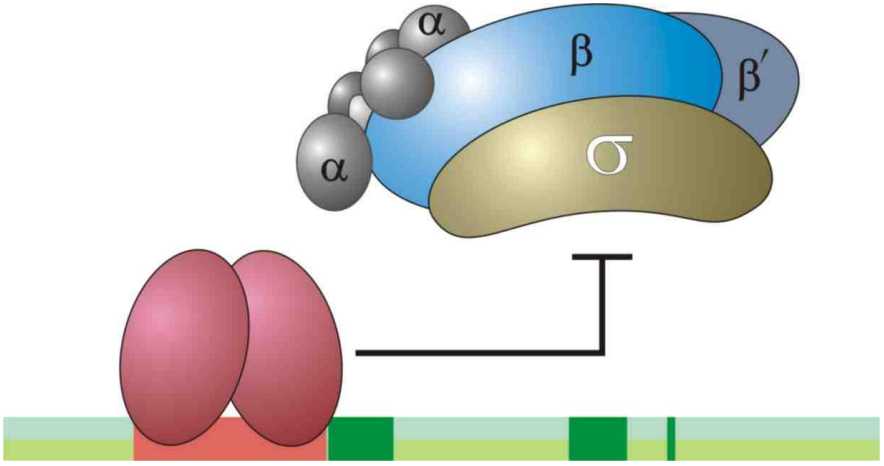
- Прокариоты
- Эукариоты

Транскрипция в прокариотах:
Регуляция транскрипции

Активация

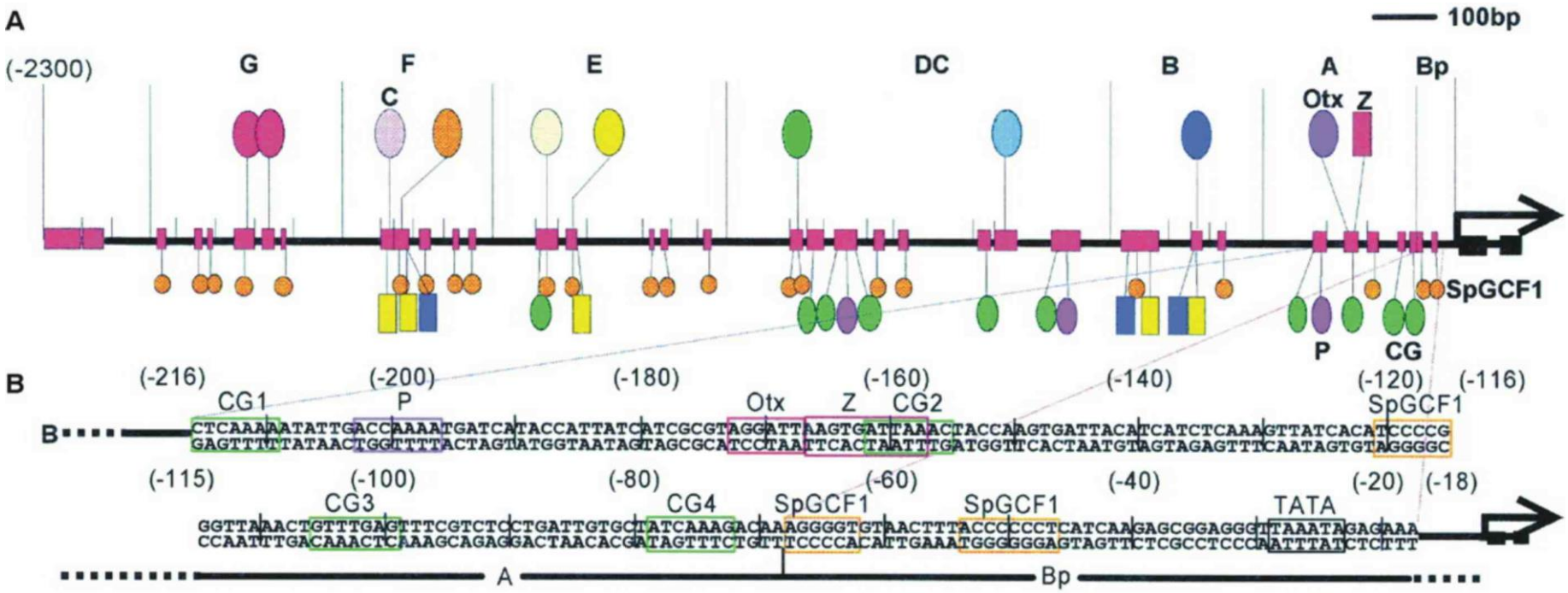


Репрессия



Источник: РГМ

Регуляция транскрипции у эукариот



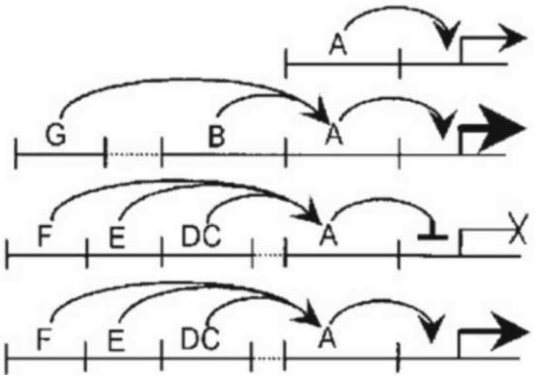
C Функции модуля A:

Экспрессия в вегетативных бляшках на ранних стадиях

Синергизм с модулями B и G - усиление экспрессии в энтодерме на поздних стадиях

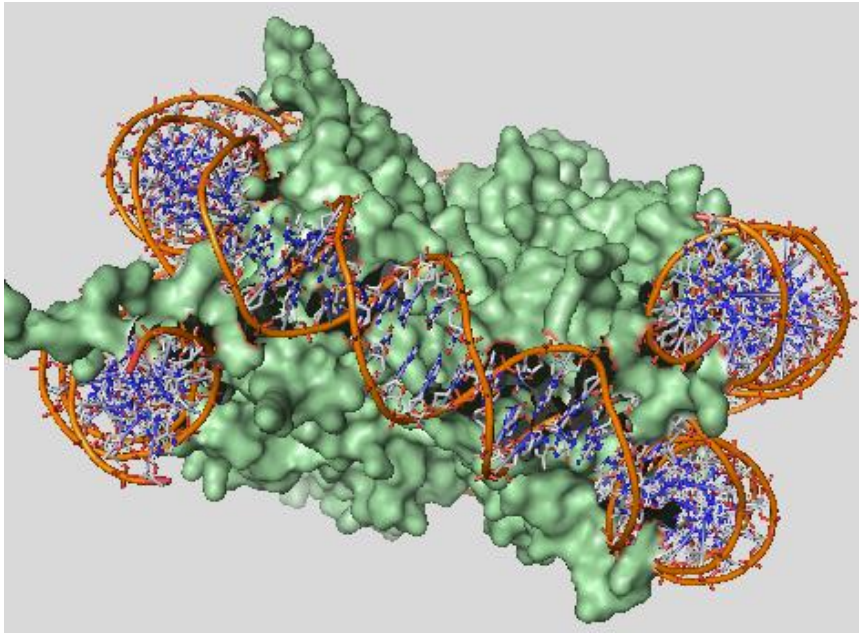
Репрессия в энтодерме (модули E и F) и скелетогенной мезенхиме (модуль DC)

Модули E, F и DC при обработке LiCl



Источник: РГМ

Для эукариот чтение без расплетения цепочек ДНК усложняется доступностью ДНК для белков



Нуклеосома:
ДНК человека на “катушке” из гистонов: вид сбоку (гистоны – такие белки)

Ещё сложнее на более высоких уровнях организации хроматина.

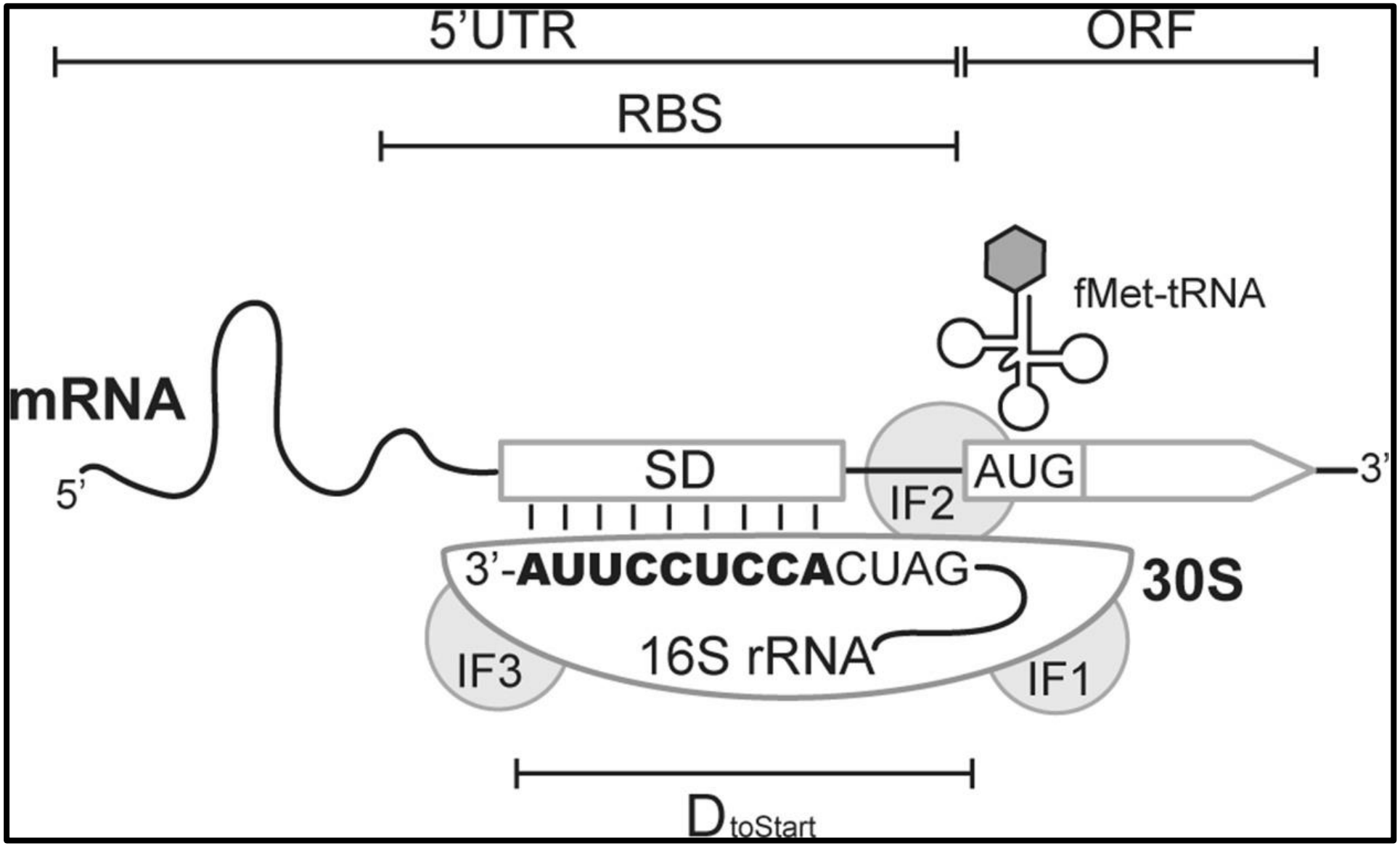
Старт трансляции у прокариот

Задание 2

Старт трансляции у прокариот

- Трансляция инициируется связыванием 30S субъединицы рибосомы с последовательностью Шайна — Дальгарно (SD).
- SD комплементарна 3’ концу 16S рРНК, входящей в состав рибосомы.
  aSD – анти Шайн – Дальгарно последовательность

Architecture of prokaryotic ribosome binding sites



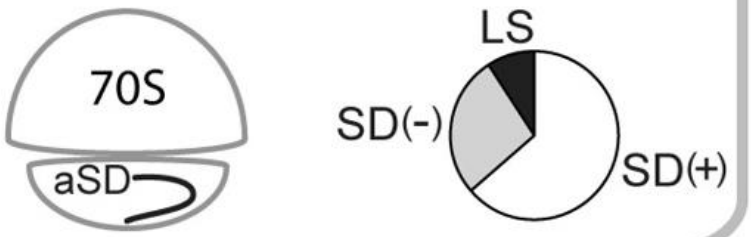
IF, initiation factors, 5'UTR, 5' untranslated region; RBS, ribosome binding site; $D_{toStart}$, the distance between the start codon and the 3' end of 16S rRNA

[6] Wen et al., The diversity of Shine-Dalgarno sequences sheds light on the evolution of translation initiation. RNA Biol. 2021

У бактерий SD(+) чаще других

Bacteria

SD(+) mRNA is dominant in major bacterial species. Yet the composition of mRNA pools varies greatly within and between taxonomic groups.

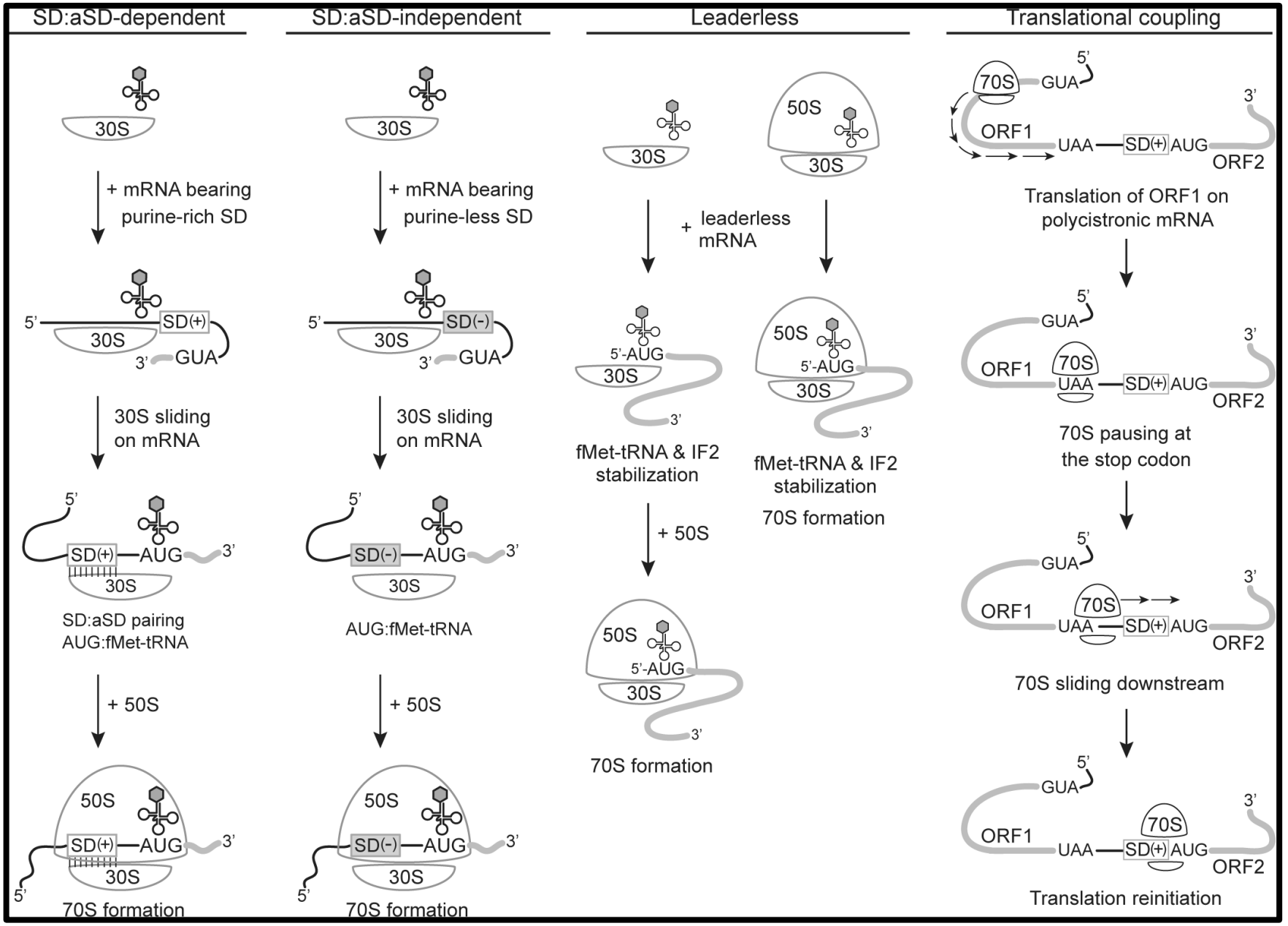


Archaea

Promoters evolved to locate immediate upstream of the open reading frames. Leaderless mRNA became dominant in the majority of archaeal species.

Бывают и SD(-) способы инициации трансляции [6]

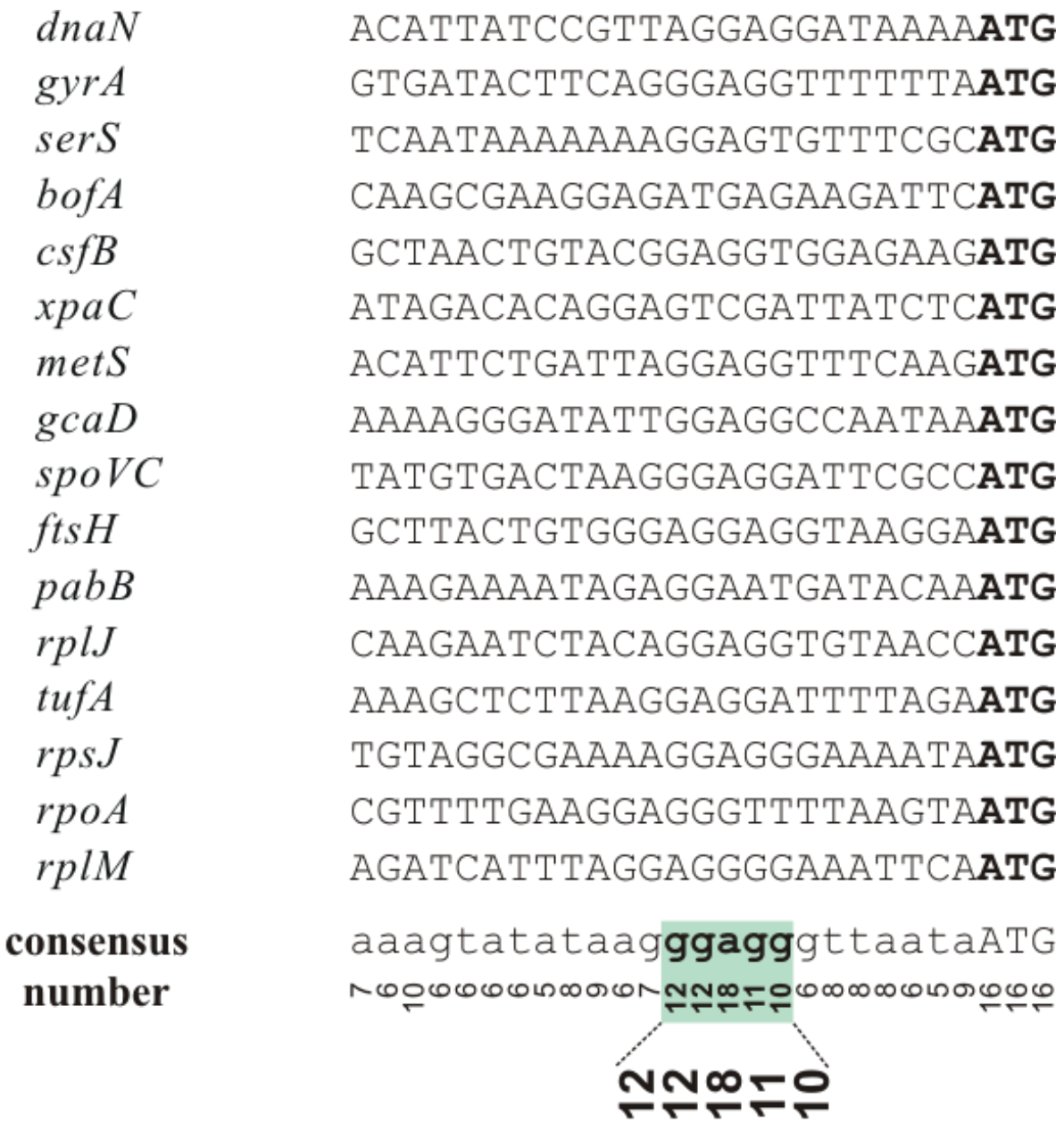
SD(+) SD(-) LS – leader less Полицистронная мРНК



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

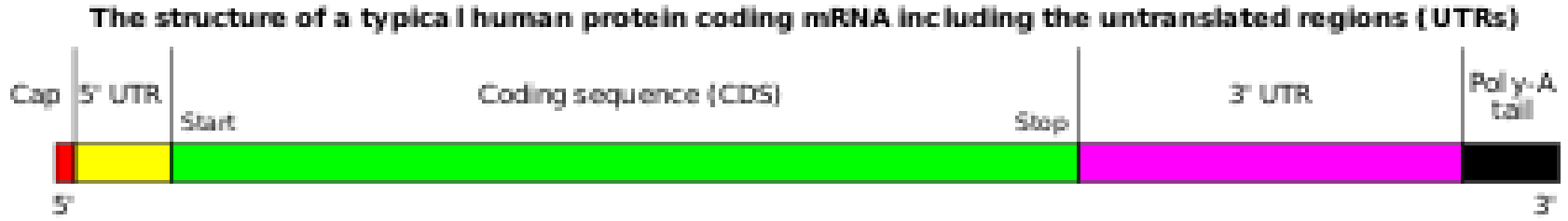


Источник: РГМ

Источник: РГМ

Старт трансляции у эукариот

Сигналы, позволяющие рибосоме отличить мРНК человека (эук.) от остальных РНК



<u>мРНК эукариот</u> содержит такие сигналы рибосоме:

- 5’: **КЭП (cap)** - 7-метилгуанозин
  - присоединяет кэп связывающий комплекс (CBC)
- 3’: **ПолиА** - много-много-много А (аденинов)
  - Присоединяет поли(А)-полимераза при наличии сигнала полиаденилирования в 3’ концевой части транскрипта

Инициация, элонгация, терминация

в объёме одного слайда

- Фактор инициации трансляции узнаёт кэп и связывается с ним. Белки PABP связываются с полиА и они же связываются с инициаторным комплексом, стабилизируя его
- Малая субъединица рибосомы садится на 5’ конец мРНК и сканирует её до старта инициации трансляции, ATG (кодон метионина)
- Привлекается большая субъединица рибосомы и начинается трансляция
- Терминация – на ближайшем стоп-кодоне в рамке

У человека одна мРНК – один белок

SARS CoV2 имеет +РНК геном

РНК коронавируса содержит оба сигнала

- ПолиА на 3’-конце
- КЭП – 7-метилгуанозин - на 5’ конце

- Первый ген CoV orf1ab начинается с 266 пн
- У SARS-CoV-2 такие ATG до 269-й пн.:
    - 107 – ATG
    - 266 – ATG

ПРОТИВОРЕЧИЕ с пред. слайдом

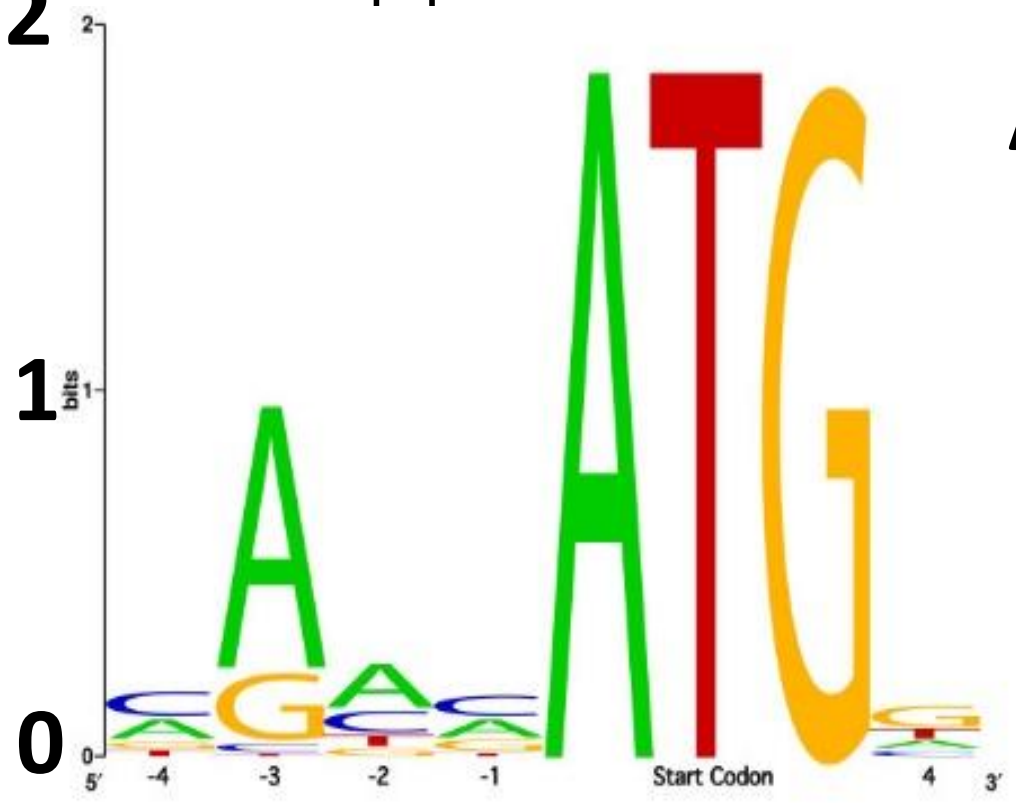
Последовательность Козак

Задание 2

Проблема: старт трансляции со второго ATG кодона

- М.Ко́зак в 1986 году проанализировала известные инициаторные кодоны ATG и нашла более длинный *слабый* сигнал
- Сигнал начала трансляции (у эукариот) называется последовательностью Козак. В разных таксонах - отличия

Последовательность Козак человека



ATG между 1 и 269 в геноме SARS-CoV-2:

104-TGC **ATG** C -110

263-AAG **ATG** G -269

Контекст (окружение) ATG в позиции 266 более похож на последовательностью Козак

Kozak M, Cell, 1986.

Кэп-зависимая инициация трансляции

При сканирующем механизме малая субъединица рибосомы садится на 5'-конец мРНК в области кэпа и двигается вдоль молекулы мРНК, «сканирует» кодоны в поисках инициаторного AUG.

• Консенсусная последовательность Ко́зак, играющая важную роль в инициации трансляции у эукариот, включает четыре-шесть нуклеотидов, предшествующих старт-кодону, и один-два нуклеотида непосредственно после старт-кодона.
• Оптимальный нуклеотидный контекст AUG кодона, коррелирует с высоким уровнем синтеза белка с соответствующей мРНК in vivo и является характеристикой так называемой "сильной" (эффективно инициирующей трансляцию) последовательности Козак
• Последовательность Козак не является сайтом связывания рибосомы (англ. ribosomal binding site, RBS), в отличие от прокариотической последовательности ШайнаДальгарно.

из презентации М.Скоблова

Как ещё может инициироваться трансляция у эукариот? _____
(И.Н. Шацкий и команда)

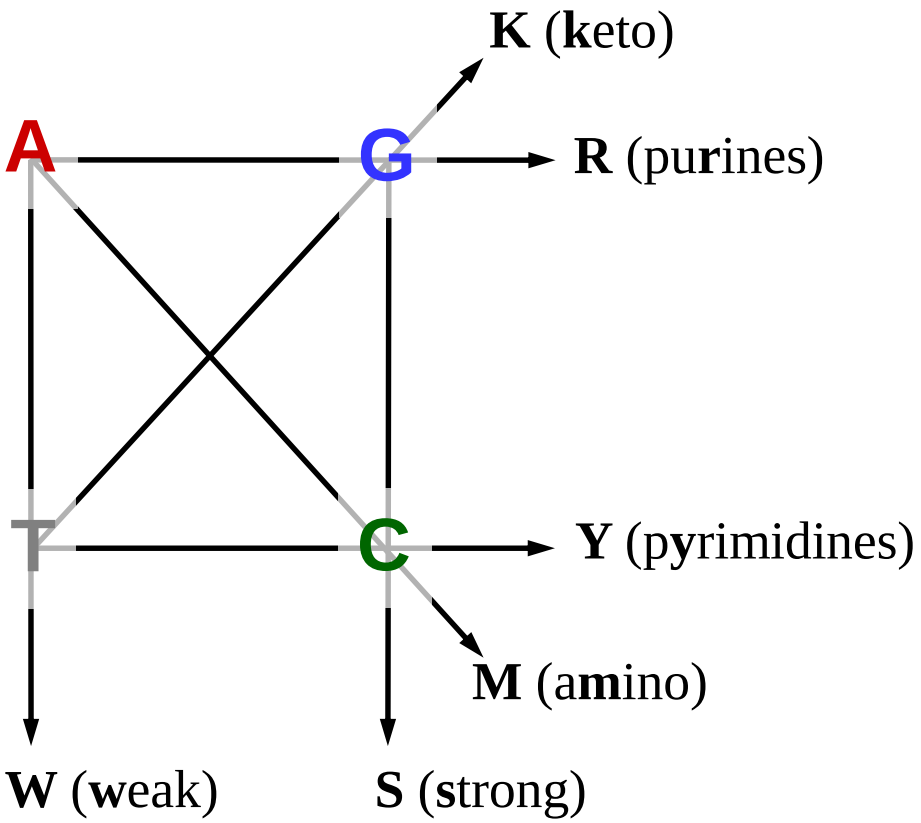
Способы описания сигнала для поиска

Для поиска в последовательности новых представителей

Мотив – описание последовательностей одного сигнала

- Точная последовательность
  - AAAAAAAAAAAAAAAAAAAAA (от десятков до сотен букв)
  - CpG
  - GATC
- Паттерн
  - CCWGG, и др. примеры
- Выравнивание
  - Консенсус
  - LOGO
  - Позиционная весовая матрица (PWM)
  - Профиль (чаще для белков)

Для справки: Ambiguity codes



C / G / T (“не A”) → **B**
A / G / T (“не C”) → **D**
A / C / T (“не G”) → **H**
A / C / G (“не T”) → **V**

A / C / G / T → **N** (**n**ucleotide)

Источник: РГМ

55

Выравнивание сайтов связывания PurR *E. coli*

cvpA	**CCTACGCAAACGTTTTCTTTTT**
purM	**GTCTCGCAAACGTTTGCTTTCC**
purT	**CACACGCAAACGTTTTCGTTTA**
purL	**TCCACGCAAACGGTTTCGTCAG**
purE	**GCCACGCAACCGTTTTCCTTGC**
purC	**GATACGCAAACGTGTGCGTCTG**
purB	**CCGACGCAATCGGTTACCTTGA**
purH	**GTTGCGCAAACGTTTTCGTTAC**
$purA_1$	**TTGAGGAAAACGATTGGCTGAA**
$purA_2$	**TTTAAGCAAACGGTGATTTTGA**
guaB	**TAGATGCAATCGGTTACGCTCT**
$purR_1$	**TAAAGGCAAACGTTTACCTTGC**
$purR_2$	**AACGAGCAAACGTTTCCACTAC**
consensus	**AcGCAAACGtTTtCgT**
pattern	**dnGMAAhCGdKKnbnY**

Источник: РГМ

Позиционная весовая матрица (PWM)

Для поиска сигналов в последовательностях, если известны последовательности ряда сигналов.

Задание 3

PWM Известно выравнивание (без гэпов) последовательностей сигнала

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

Задача: найти все сигналы в геноме

Похожа ли новая последовательность на выравнивание?

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

Новая .... CCTAACCATTATTTTT ....

Идея: вес буквы зависит от позиции в выравнивании

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
	A	**C**	**G**	**C**	**A**	**A**	**A**	**C**	**G**	**T**	**T**	**T**	**t**	**C**	**g**	**T**
	G	**C**	**C**	**T**	**A**	**C**	**C**	**C**	**C**	**A**	**T**	**T**	**A**	**T**	**T**	**T**

Проверяемая последовательность

Самая частая буква в колонке (консенсус)

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	16
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	0.08	0.00
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	0.38	0.00
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	0.23	0.85
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	0.31	0.15
Всего	1	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**	**T**

Повышенная частота буквы может объясняться её повышенной частотой в геноме!!!

Частота G в позиции 15 равна 0.38

Значит ли это что-нибудь, если GC состав генома равен 0.7,
Т.е. частота G в геноме равна 0.35?

ЛОГАРИФМ Отношения правдоподобия W как вес различия наблюдаемой частоты и ожидаемой:

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

ШАГ 4. Матрица весов w(b,j) - PWM

	Баз. частоты	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0.15	1.6	0.0	-inf	-0.7	1.9	1.9	1.6	-inf	-inf	-0.7	-inf	-inf	0.7	-inf	-0.7	-inf
G	0.35	-0.8	-0.8	1.0	-inf	-inf	-inf	-inf	-inf	1.0	-0.1	-1.5	-1.5	-0.4	-1.5	0.1	-inf
T	0.15	-0.7	-0.7	-inf	-inf	-inf	-inf	0.0	-inf	-inf	1.4	1.8	1.8	0.9	-0.7	0.4	1.7
C	0.35	-inf	0.6	-inf	1.0	-inf	-inf	-1.5	1.0	-inf	-inf	-inf	-inf	-1.5	0.9	-0.1	-0.8
	1	-inf	-0.9	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-0.3	-inf	-0.3	-inf

Шаг 5. Псевдоотсчёты: борьба с –inf и не только... Pseudocounts

Идея в том, чтобы немножко изменить ЧАСТОТЫ букв.

(1) Избавляется от возможности нулевой частоты буквы

(2) Если частота A равна единицы, то разрешим другим буквам появляться с малой частотой, вдруг у нас просто мало последовательностей, чтобы все буквы появились

$F(b,j) = [N(b,j) + \varepsilon(b)] \ /(N + \varepsilon)$ вместо

$f(b,j) = N(b,j)/N$

Здесь $\varepsilon = \varepsilon(A) + \varepsilon(G) + \varepsilon(T) + \varepsilon(C)$

Все $\varepsilon(b)$ маленькие в сравнении с N

Подбираются опытным путем

Выбор $\varepsilon(b)$

ШАГ 4. Частоты с псевдоотсчётами

F(b,j)	баз. Частоты	e(b)	1	2	3	4	5	6	7	8	9	10	11	12	13
A	0.15	0.10	0.75	0.16	0.01	0.08	0.98	0.98	0.75	0.01	0.01	0.08	0.01	0.01	0.31
G	0.35	0.10	0.16	0.16	0.98	0.01	0.01	0.01	0.01	0.01	0.98	0.31	0.08	0.08	0.23
T	0.15	0.10	0.08	0.08	0.01	0.01	0.01	0.01	0.16	0.01	0.01	0.60	0.90	0.90	0.38
C	0.35	0.10	0.01	0.60	0.01	0.90	0.01	0.01	0.08	0.98	0.01	0.01	0.01	0.01	0.08
	1	0.40	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

ШАГ 5. Матрица PWM с псевдоотсчётами Вес последовательности

	баз. Частоты	e(b)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.15	0.10	1.6	0.0	-3.0	-0.6	1.9	1.9	1.6	-3.0	-3.0	-0.6	-3.0	-3.0	0.7	-3.0	-0.6
G	0.35	0.10	-0.8	-0.8	1.0	-3.8	-3.8	-3.8	-3.8	-3.8	1.0	-0.1	-1.5	-1.5	-0.4	-1.5	0.1
T	0.15	0.10	-0.6	-0.6	-3.0	-3.0	-3.0	-3.0	0.0	-3.0	-3.0	1.4	1.8	1.8	0.9	-0.6	0.4
C	0.35	0.10	-3.8	0.5	-3.8	0.9	-3.8	-3.8	-1.5	1.0	-3.8	-3.8	-3.8	-3.8	-1.5	0.9	-0.1
	1	0.40	-3.6	-0.8	-8.8	-6.5	-8.8	-8.8	-3.6	-8.8	-8.8	-3.2	-6.5	-6.5	-0.2	-4.2	-0.2
			G	C	C	T	A	C	C	C	C	A	T	T	A	T	T

КОНЕЦ

Больше не успеем

Информационное содержание выравнивания последовательностей сигнала

LOGO

«Сила» сигнала

Информация и энтропия сигнала

Информация противоположна энтропии.
Энтропия – мера неупорядоченности.
Чем больше энтропия, тем меньше порядка.
Чем больше информации, тем меньше энтропия

- Информационная ёмкость - потенциально возможное количество информации в сигнале (матем.)
- Информационное «содержание» – насколько сигнал отличается от случайного (статист.)
- Содержательность - чем чаще сигнал приводит к реакции, тем более содержательна информация в сигнале

Энтропия

- Изучаем сигнал, который есть последовательность букв. В нашем случае – задан выравниванием представителей сигнала.
- Энтропия H сигнала, заданного выравниванием, – число характеризующее неопределенность сигнала. Чем ближе сигнал к набору случайных посл-й, тем больше энтропия H
  Аксиомы:
  - H положительна
  - H = 0 если сигнал однозначно предсказуем (задан точной последовательностью)
  - Чем менее предсказуем сигнал по выравниванию, тем больше энтропия сигнала. Максимум достигается когда все слова, составляющие сигнал, равновероятны.
  - H аддитивна: энтропия сигнала длиной в одну букву равна сумме энтропий каждой из букв ; энтропия сигнала состоящего из нескольких независимых сигналов (колонок выравнивания) равна сумме энтропий
  - H можно вычислить в два шага через группировку.
    Пример группировки: W= {A или T}, S = {G или C}.
    Энтропию сигнала в алфавите (A,T,G,C) можно вычислить через энтропию в алфавите (W, S) и энтропии W в алфавите (A,T) и S в алфавите (G,C)

Теорема Шеннона: существует единственная функция H, удовлетворяющая аксиомам

На примере сигналов из нуклеотидов ДНК

- Энтропия сигнала из одного нуклеотида
  $H = -\Sigma_b\, p(b) \log_2 p(b)$ b пробегает A, T, G, C. Если буквы равновероятны, то H = 2
- H(сигнала из N равновероятных букв) = N•H в силу аддитивности

Если сигнал двоичный. Например, последовательность комплементарных пар W=(A или T) и S = (G или C).
Пусть p = «GC состав в долях единицы»
Тогда (1-p) = «частота пар A-T»
Обозначим через H(p) энтропию однобуквенного сигнала при данном p

Если W и S – равновероятны, то сигнал из одной буквы ничего не значит. H максимальна.
Если GC состав очень близок к 0 (напр. $p=10^{-10}$), то сигнал предсказуем: почти всегда будет W. Появление S невероятно, и потому может что-то значить. H близка к 0
Аналогично с p близким к 1.

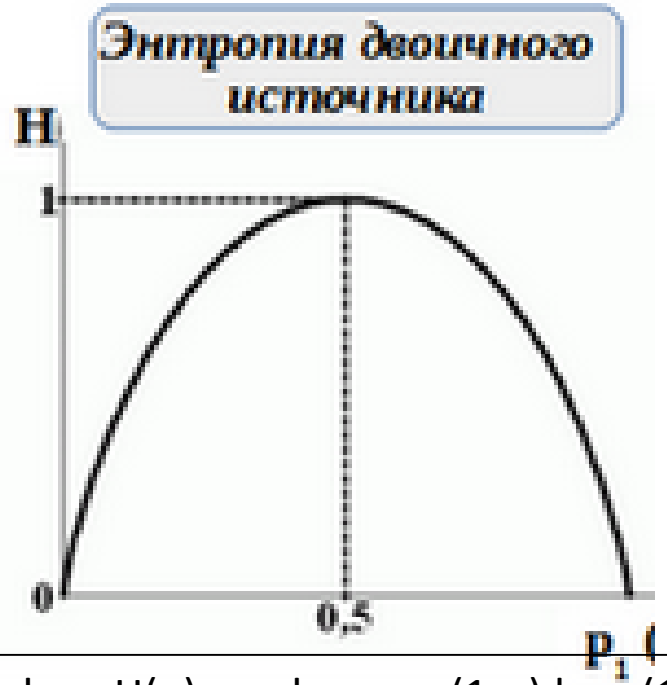


График $H(p) = p \log_2 p + (1-p) \log_2 (1-p)$ зависимости энтропии 1-буквенного сигнала от p = «GC состав»

Содержание информации IC в сигнале

- Информация IC измеряется тем, насколько уменьшилась неопределенность после получения информации о сигнале.
- $IC(\text{сигнала}) = H_{before} - H_{after}$

В нашем случае H_{before} - энтропия полностью случайного выравнивания фрагментов той же длины и того же нуклеотидного состава

H_{after} - энтропия выравнивания известных сигналов.

Использование терминологии мат.теории передачи данных Шеннона – некоторая историческая условность. Впрочем, разумная и полезная для целей анализа сигналов, заданных выравниваниями. См. сайт ниже:

"The meaning of information has nothing to do with Shannon's amount of information.

For example, the word "AND" contains the same amount of information even we spell it backward to "DNA."

Similarly, 010101 carries the same amount of information 101010 as well as a DNA codon ATG and GTA or GAT all carry the same amount of information only the meaning is different."

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/introduction_to_information_theory_page3.html

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 – тем больше в нем информации и меньше его энтропия
- IC = Hbefore - Hafter

Вычисление IC мотива M, заданного выравниванием

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

Сигнал NNN….NN длины n с независимым появлением букв

$$H_{before} = - \sum_i \sum_b p(b) \log_2 p(b)$$

i номер буквы в сигнале; все буквы равноправны,

p(b) - априорная вероятность появления буквы b в изучаемых последовательностях (напр., частота в геноме)

$$H_{after} = - \sum_i \sum_b f_i(b) \log_2 f_i(b)$$

Это энтропия того знания о мотиве M, которое мы получаем глядя на выравнивание: $f_i(b)$– частота буквы b в i-м столбце

$$H'_{before} = - \sum_i \sum_b f_i(b) \log_2 p(b)$$ Обоснование?

$$IC = H'_{before} - H_{after} = - \sum_i \sum_b f_i(b) \log_2 p(b) + \sum_i \sum_b f_i(b) \log_2 f_i(b) =$$

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

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

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

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

$\log_2[f(b,j)/p(b)] = w(b,j)$ – вес из матрицы PWM **без псевдоотсчётов.**

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

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

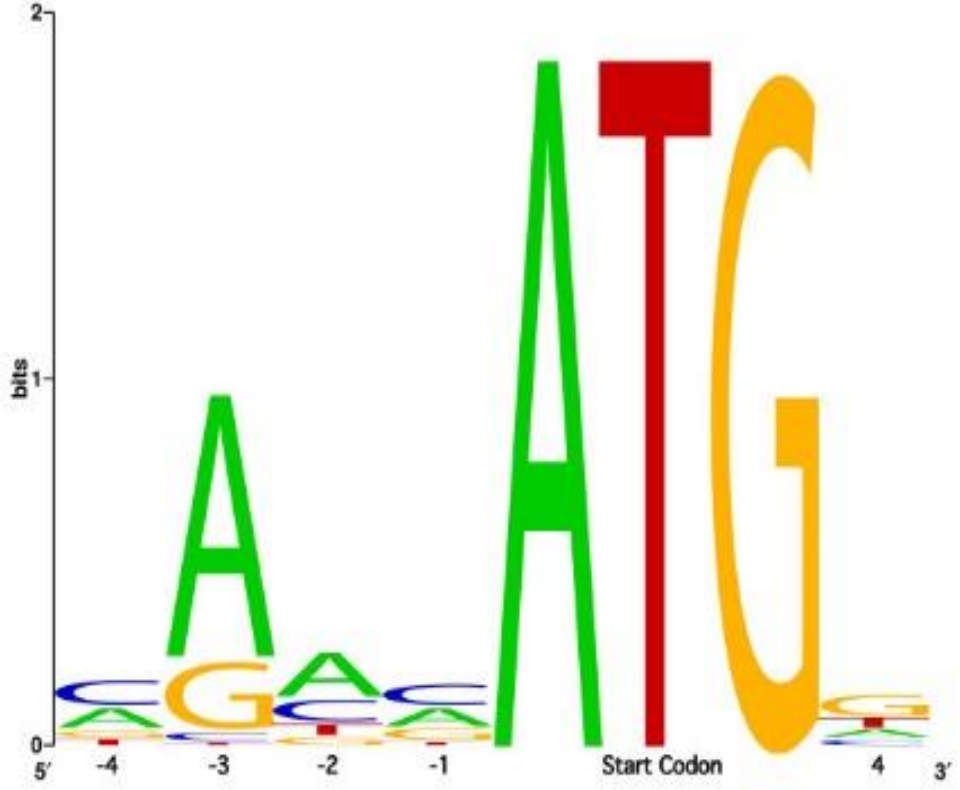
Из формулы следует, что IC(j) – матожидание - веса в колонке при распределении вероятностей букв b заданного частотами букв в колонке

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 сигнал от буквы b в позиции j имеют высоту, равную информационному содержанию IC(b,j)

webLOGO.

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

N = 4 для ДНК, т.к. 4е буквы, $\log_2 N = 2$

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

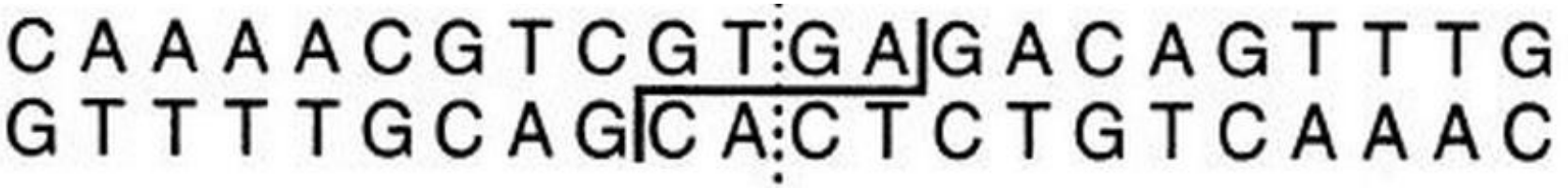
При p(b) = ¼ для всех b получаем

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

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

КОНЕЦ ПРЕЗЕНТАЦИИ