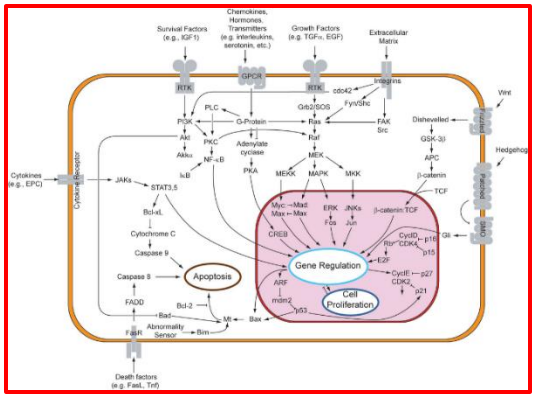
СИГНАЛЫ В ГЕНОМЕ

Поиск известных сигналов

Понятно, до некоторой степени☺

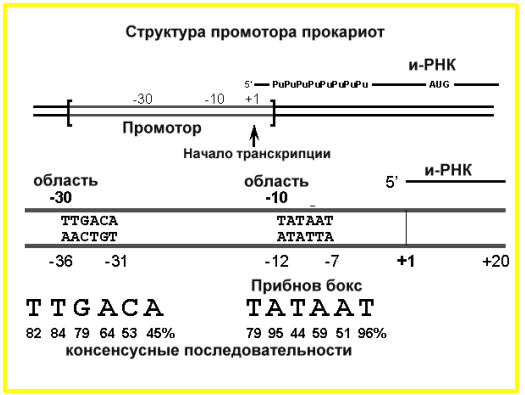
https://vrnikc.ru/wp-content/uploads/2021/04/neverbalnye-1024x640.jpg

В клетке – чёрт ногу сломит! Показана примитивная схема



https://upload.wikimedia.org/wikipedia/commons/thumb/f/fb/Signal_transduction_pathways.png/1200px-Signal_transduction_pathways.png

В геноме – кажется, что понятно, но...



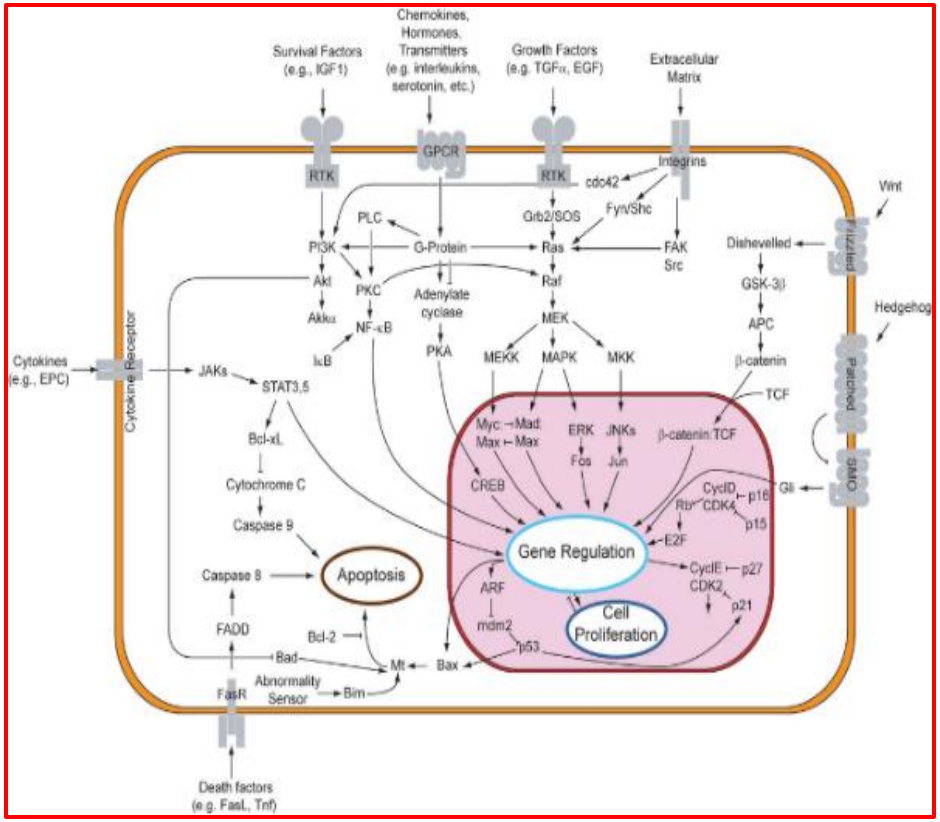
https://studfile.net/html/2706/365/html_TMQTMVH3gQ.IAMF/htmlconvd-eRHp66_html_c67aaedb6bd877a8.png

наша тема

I.Сигналы

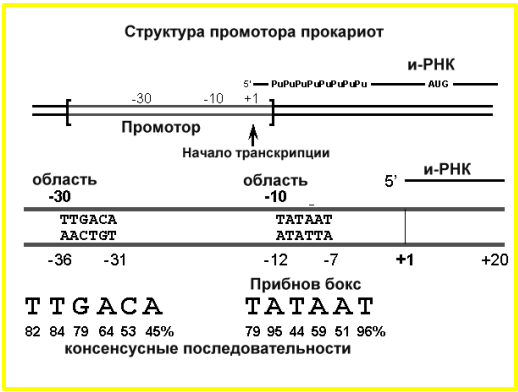
Их роль в жизни и в геноме

В клетке – чёрт ногу сломит!
Показана примитивная схема



https://upload.wikimedia.org/wikipedia/commons/thumb/f/fb/Signal_transduction_pathways.png/1200px-Signal_transduction_pathways.png

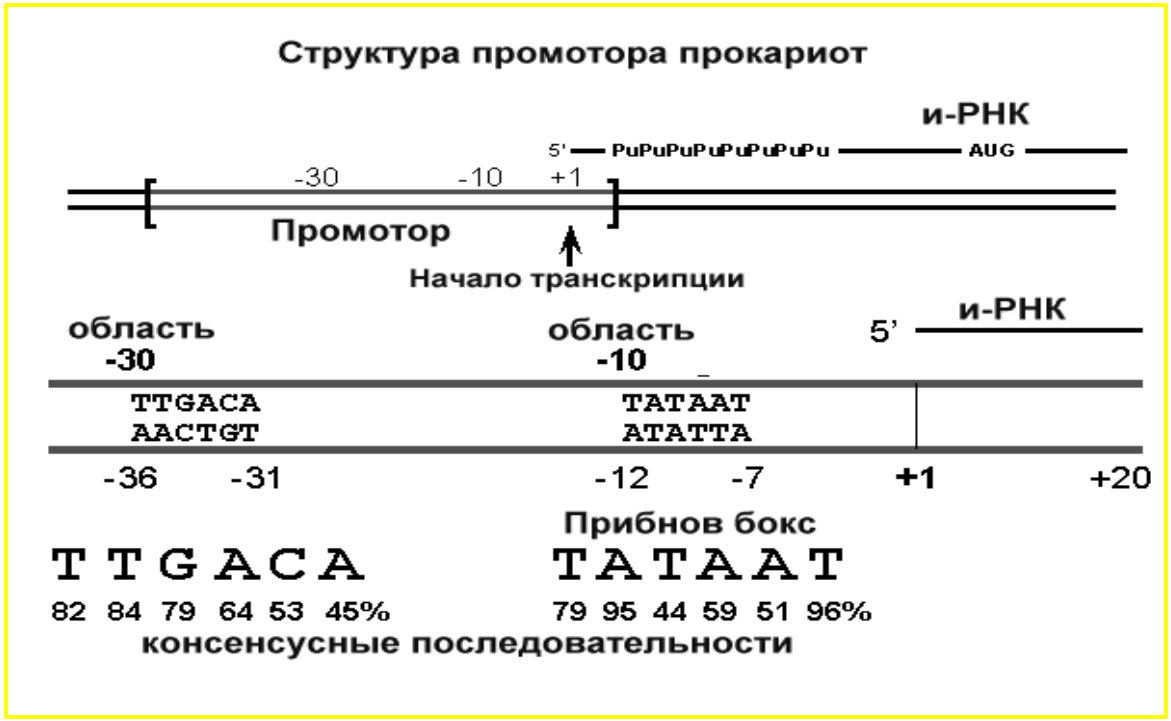
В геноме –
кажется, что понятно, но...



https://studfile.net/html/2706/365/html_TMQTMVH3gQ.IAMF/htmlconvd-eRHp66_html_c67aaedb6bd877a8.png

наша тема

В геноме –
кажется, что понятно, но...



https://studfile.net/html/2706/365/html_TMQTMVH3gQ.IAMF/htmlconvd-eRHp66_html_c67aaedb6bd877a8.png

наша тема

Характеристики; сигнал несет информацию что такое информация И.М.

Носитель сигнала – светофор

Какие значения –

красный, зелёный, не работает

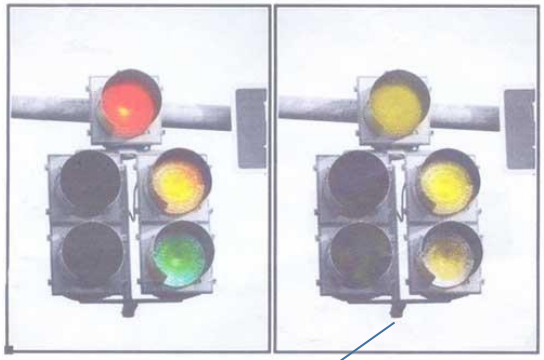
Кому – пешеходу

и любому, кто реагирует на сигнал

Действие – переход улицы:

“красный” стой, “зелёный” иди,

“не работает” - ???



Так видит собака

Бродячие собаки тоже переходят улицу. “Нет колбочек, воспринимающих красный цвет: собаки не отличают красный цвет от зеленого и могут спутать оба этих цвета с желтым или оранжевым.”

https://vetsas.by/base/stati/zrenie-sobak-kak-vidyat-sobaki

Мужественность сигналов, влияющих на пешехода и бродячую собаку

- Светофор
- Отсутствие машин
- Действия других пешеходов
- Другие – вспомните)))

Так и в геноме.

Повторение для коллоквиума

- Носитель сигнала
- Какие значения принимает
- Кому адресован
- Какие действия вызывает
- Сила сигнала - чем чаще правильная реакция на сигнал, тем сигнал сильнее

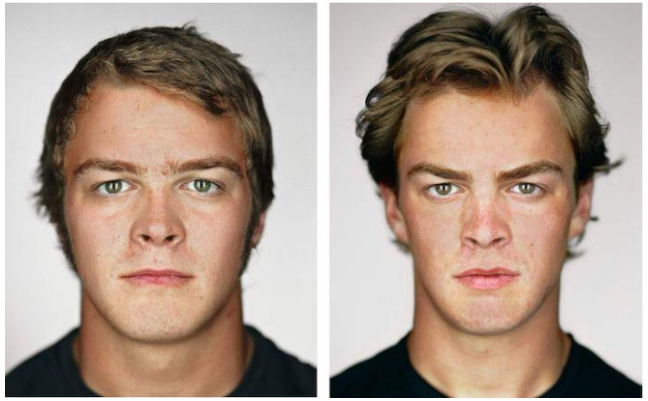
Светофор

- Для водителя – сильный сигнал (штраф большой)
- Для пешехода – несколько слабее
- Для бродячей собаки – слабый сигнал

II Роль сигналов в геноме.

От генома зиготы до фенотипа организма и психотипа индивидуума

Однояйцовые близнецы. Геномы одинаковы. Сходство закодировано в зиготе.



Цвет глаз определен, в основном двумя генами OCA2 и HERC2 на хромосомах родителей, и то бывают варианты.

Brancato et al. 2023 Forensic DNA Phenotyping: Genes and Genetic Variants for Eye Color prediction Genes (Basel).

А как закодирована в геноме **форма носа**:

- прямой
- с горбинкой
- курносый
- крючком (как у бабы Яги)
- широкий
- узкий

По другому. Можно ли секвенировав геном человека сказать какой у него нос?

Не пустой вопрос: *ВАЖНО для поиска преступника*☺

Известны и успешно изучаются

- Гены, последовательности и структуры белков и РНК
- Функции белков и РНК
- комплексы белков и белков с РНК
- метаболические пути
- Сигналы многих процессов в клетках
- И многое другое

- Как закодирован в геноме зиготы
  - фенотип будущего взрослого человека? *(черты лица, пропорции тела и конечностей, характера)*;
  - составах протеомов в клетках разных тканей? только кое-что – пример

Протеом – совокупность разных белков и их процентное соотношение в клетке.
Белки – разные, если транслируются с разных зрелых мРНК, т.е. имеют разные последовательности а.к.о

- Как закодировано в геноме пчёл поведение пчелиного роя?

III. Сигналы в геноме

ДНК и РНК

Носители сигналов

1. Последовательность ДНК, РНК
(-35) TTGACA (-10) TATAAT (слайд 4)

2. Множественные сигналы
ориджин у бактерий

3. Вторичная и пространственная структура РНК

Терминация транскрипции у прокариот

8. Особые структуры ДНК

G-квадруплекс

7. Модификации ДНК, РНК
Метилирование цитозина в динуклеотиде CpG

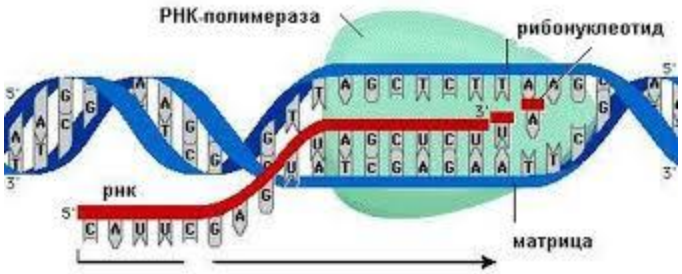
6. Доступность сигнала ДНК для получателя – белков (и РНК)

Кому адресованы

- Белкам
- РНК
- Комплексам белков
- Комплексам белков и РНК

1. Старт транскрипции у прокариот

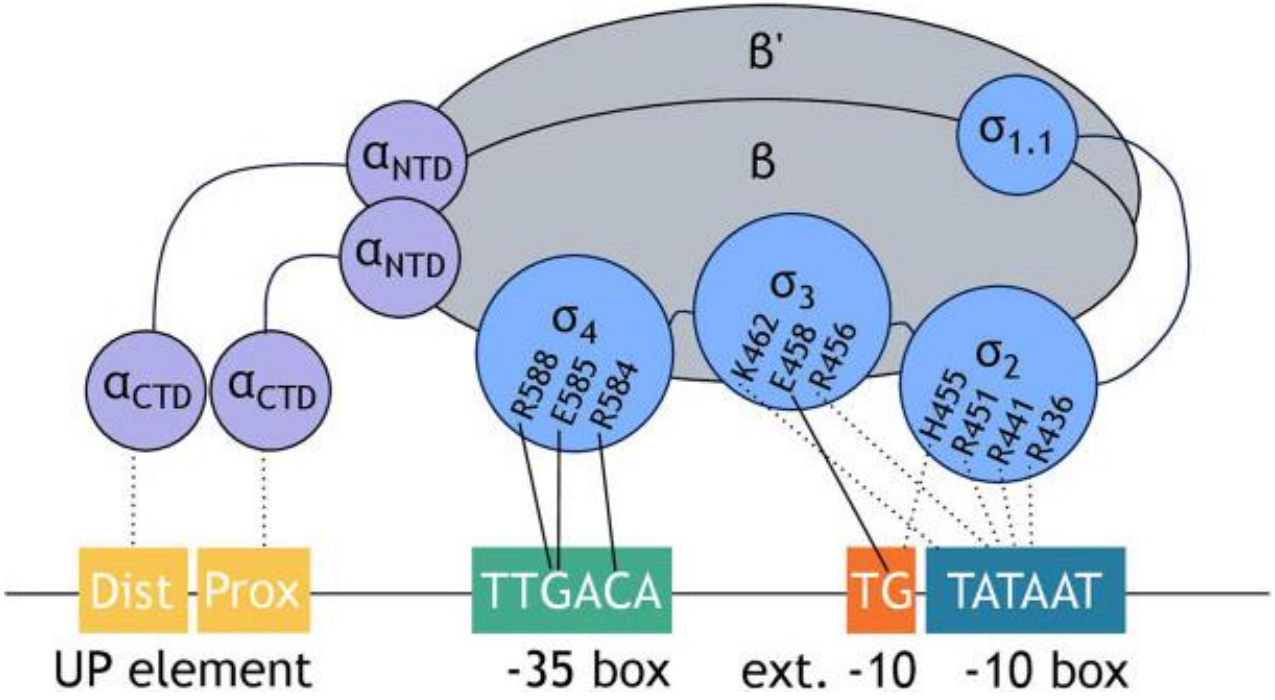
Последовательность как сигнал



Старт транскрипции

1. **Промотор** – участок перед стартом транскрипции (TSS) содержащий много сигналов, регуляции транскрипции (длина примерно 200 п.н.)
2. Для начала транскрипции на промоторе должен собраться **комплекс – РНК полимераза** – состоящая из нескольких субъединиц. **RNAP** холоэнзим состоит из субъединиц ααββ’ωσ
3. Первой с промотором связывается **σ-субъединица** и инициирует сборку RNAP
4. Бывают разные sigma факторы, сигма-70 самый распространенный у бактерий.
5. В одном геноме в промоторах мРНК транскрибируемых с одним сигма фактором последовательности -10 и -35 консервативны.

   Как всегда в эволюции, те же сигналы в геномах близкородственных бактерий имеют больше шансов быть похожими.


β'
β
σ1.1
αNTD
αNTD
αCTD
αCTD
σ4
R588
E585
R584
σ3
K462
E458
R456
σ2
H455
R451
R441
R436
Dist
Prox
TTGACA
TG
TATAAT
UP element
-35 box
ext. -10
-10 box

Сигналы сборки RNAP

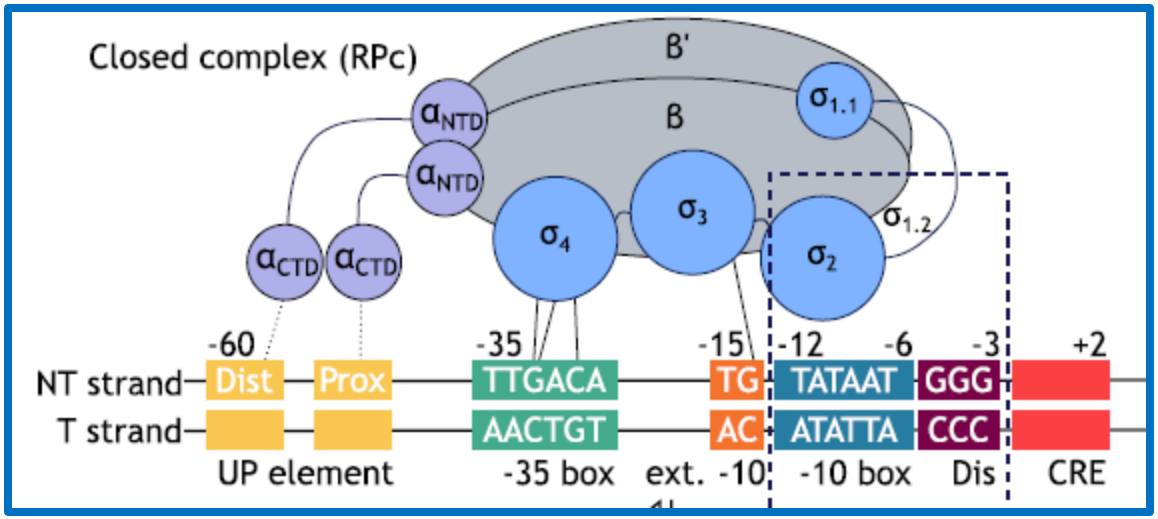
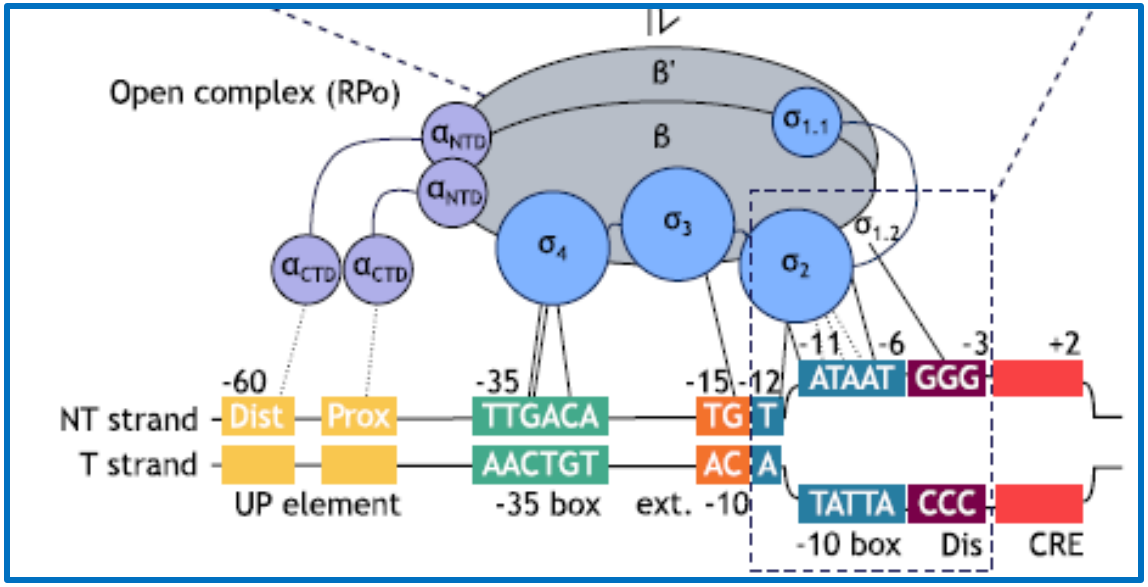


Рис. Кружками одного цвета изображены домены одного и того же белка



1) Сначала σ-фактор сзывается специфически с -10 и -35 боксами

2)Потом собирается закрытый комплекс RNAP

3) Потом расплавляется ДНК и образуется раскрытый комплекс

Самыми консервативными являются сайты
-10 = Pribnow box
-35 box

[1-1] Deal C et al., Towards a rational approach to promoter engineering: understanding the complexity of transcription initiation in prokaryotes. FEMS Microbiol Rev. 2024

Последовательности -10 и -35 для сигма-фактора SigB отличны от таковых для sigma-70 [1-2]

gene	-35	spacer(bp)	-10	Species	referen ce
ctc General stress protein,	GTTTAA	14	GGGTAT	B.Subtilis	Reder et al., 2012a
gspA General stress protein,	GTTT	14	GGGTAT	B.Subtilis	Reder et al., 2012a
trxA Thioredoxin	GTTT	16	GGGCAT	B.Subtilis	Reder et al., 2012a
usfx SigF anti-sigma factor	GTTTC	15	GGGTAT	M.tuberculosis	Williams et al., 2007
phoY1 transcriptional regulatory	GGATTG	16	GGGTAT	M.tuberculosis	Williams et al., 2007
Rv2884 transcriptional regulatory	AGTTGG	18	GGGTAC	M.tuberculosis	Williams et al., 2007

SigB используется транскрипции >150 генов, важных для ответов на стрессы и выживания

Литература

[1-1] Deal C et al., Towards a rational approach to promoter engineering: understanding the complexity of transcription initiation in prokaryotes. FEMS Microbiol Rev. 2024

[1-2] Rodriguez et al. The Stress-Responsive Alternative Sigma Factor SigB of *Bacillus subtilis* and Its Relatives: An Old Friend With New Functions. Front Microbiol. 2020

[1-3] Jensen and Galburta, The Context-Dependent Influence of Promoter Sequence Motifs on Transcription Initiation Kinetics and Regulation, 2021

[2-4] Murakami KS, Darst SA. Bacterial RNA polymerases: the wholo story. Curr Opin Struct Biol. 2003

2. Участок инициация репликации у бактерий **oriC**

Множественный сигнал

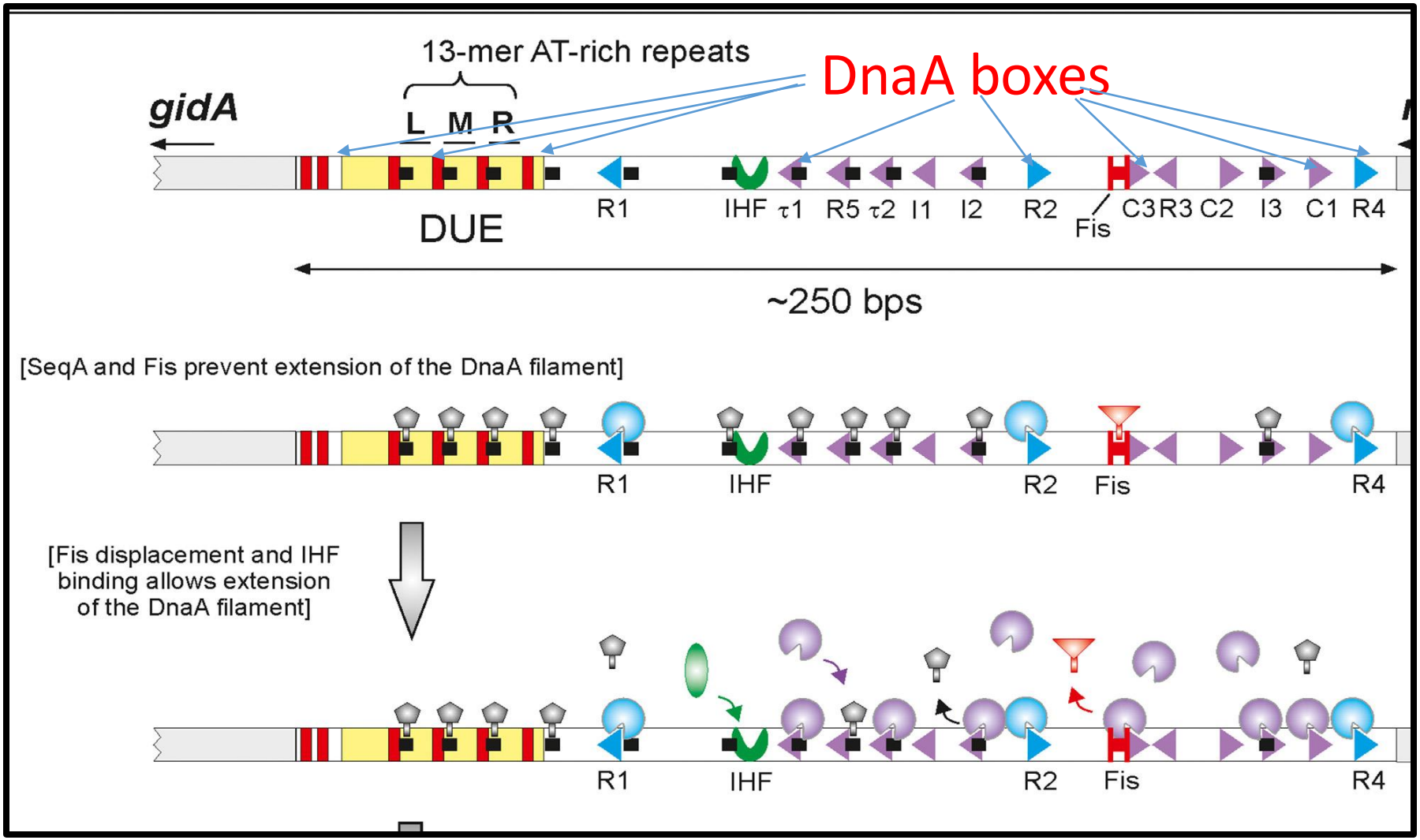
Origin of replication (**oriC**).

Участок ДНК ≈250 п.н. со многими сайтами определенной последовательности. **Белки DnaA – первыми связываются со своими сайтами (DnaA boxes) и инициируют репликацию**

Replisoma - комплекс, состоящий из 15—20 различных белков.

Вопрос о вариабельности сайта связывания DnaA обсуждается в [2-2]

Как решается вопрос когда пора делиться? (сигналов не знаю ААл)



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

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

(1) Кластер сайтов связывания белка DnaA (DnaA boxes) The default DnaA box motif is the standard motif (TTATCCACA) of *E.coli* [2-1]
(2) Участок DUE (DNA unwinding element) A-T богатый

(1) Последовательности, узнаваемые другими регуляторными белками

[2-2] Wolanski et al., oriC-encoded instructions for the initiation of bacterial chromosome replication, 2015

Литература

[2-1] Ori-Finder 2022
A Comprehensive Web Server for Prediction and Analysis of Bacterial Replication Origins
https://tubic.org/Ori-Finder2022/public/index.php

[2-2] Wolański M, Donczew R, Zawilak-Pawlik A, Zakrzewska-Czerwińska J. oriC-encoded instructions for the initiation of bacterial chromosome replication. Front Microbiol. 2015 Jan 6;5:735. doi: 10.3389/fmicb.2014.00735. PMID: 25610430; PMCID: PMC4285127.

[2-3] Wegrzyn KE, Gross M, Uciechowska U, Konieczny I. Replisome Assembly at Bacterial Chromosomes and Iteron Plasmids. Front Mol Biosci. 2016

[2-4] Wegrzyn K, Konieczny I. Toward an understanding of the DNA replication initiation in bacteria. Front Microbiol. 2024

3.Терминация транскрипции у прокариот

Вторичная и 3D структура РНК

1. Rho-зависимая терминация. Rho – белок, узнаёт rut-сайт в mRNA (длинный 78нукл. с C>G, потом шпилька RNA [3-2])

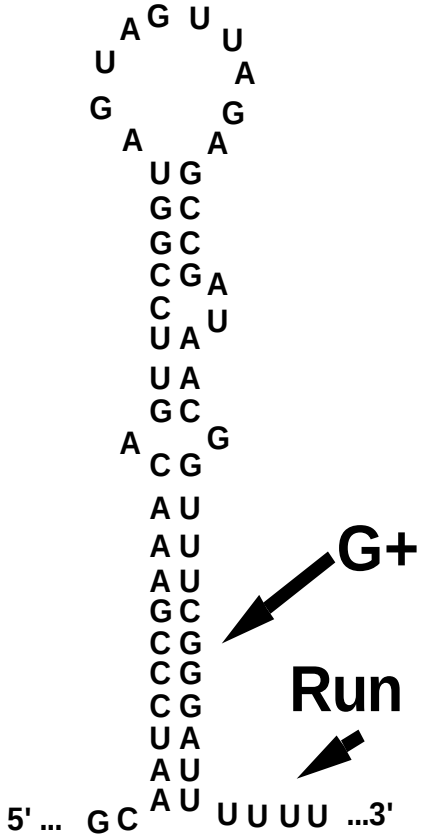
2. Rho – независимая. Сложная шпилька мРНК [3-1]
   web service: http://rssf.i2bc.paris-saclay.fr/toolbox/arnold/

[3-1] Naville M et al., ARNold: a web tool for the prediction of Rho-independent transcription terminators. RNA Biol. 2011

[3-2]Di Salvo M et al., **RhoTermPredict: an algorithm** for predicting Rho-dependent transcription terminators based on Escherichia coli, Bacillus subtilis and Salmonella enterica databases. BMC Bioinformatics. 2019

[3-3] Неплохой текст в википедии про это
https://en.wikipedia.org/wiki/Terminator_(genetics)

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



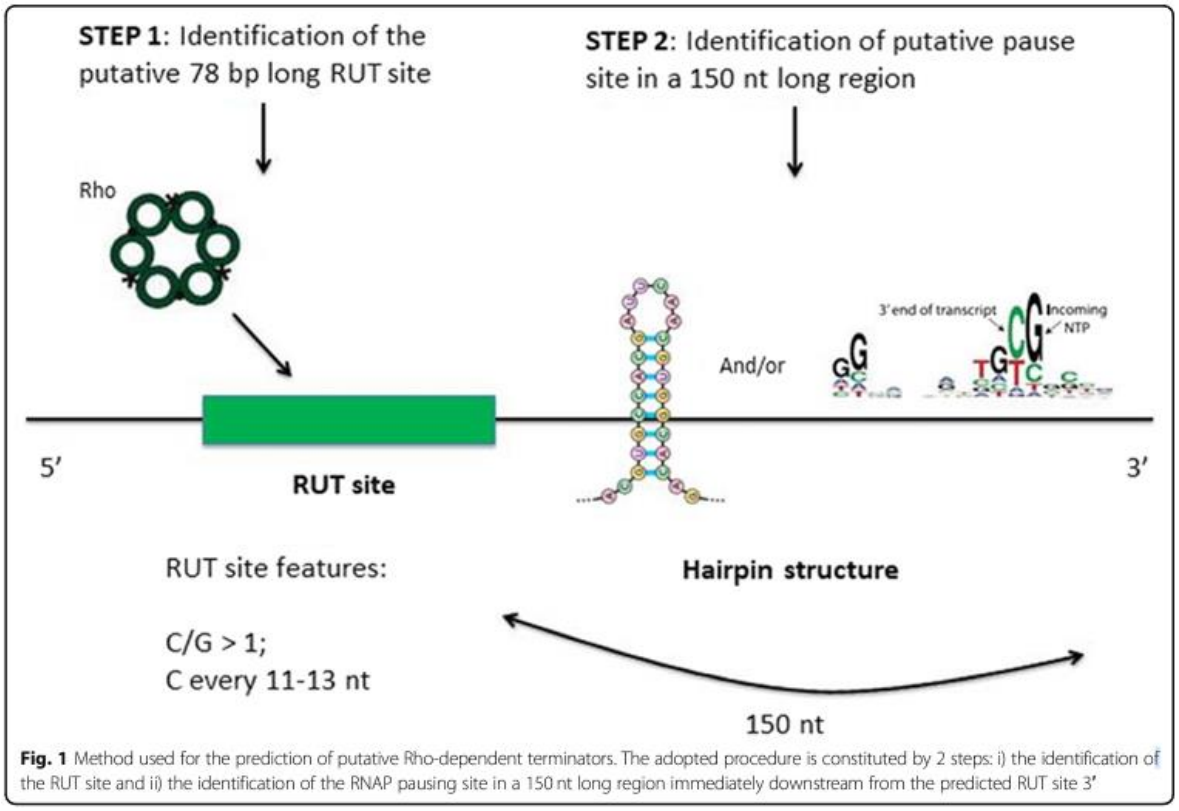


Fig. 1 Method used for the prediction of putative Rho-dependent terminators. The adopted procedure is constituted by 2 steps: i) the identification of the RUT site and ii) the identification of the RNAP pausing site in a 150 nt long region immediately downstream from the predicted RUT site 3'

Termination of transcription Rho-**dependent.**
Rut-site (C/G>1 и шпилька. Алгоритм [3-1]

8. G-квадруплекс (G4)

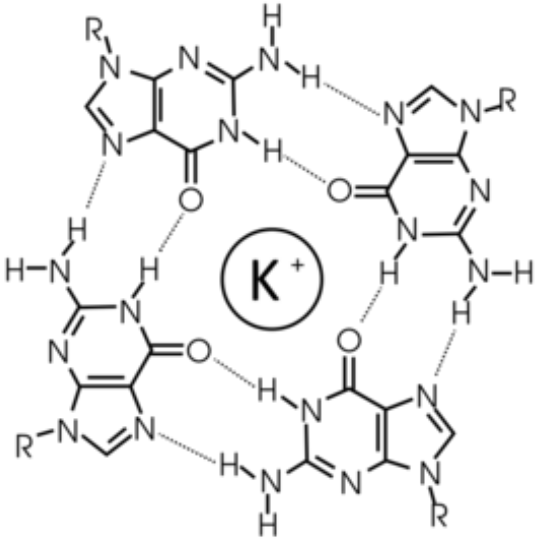
Паттерн.
Особая структура ДНК, сигнал для разных процессов. Например, для регуляции транскрипции гена

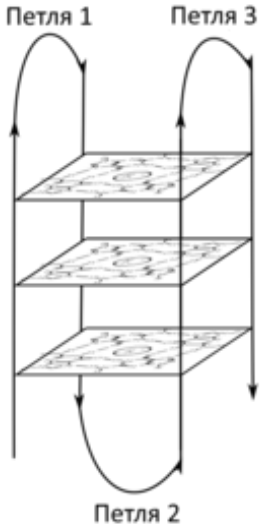
В углах этажерки расположены гуанины. Поэтому такая структура ДНК может образоваться только при наличии четырех троек гуанинов в последовательности ДНК, связанных петлями неопределенной длины. У этажерки может быть и больше полочек.

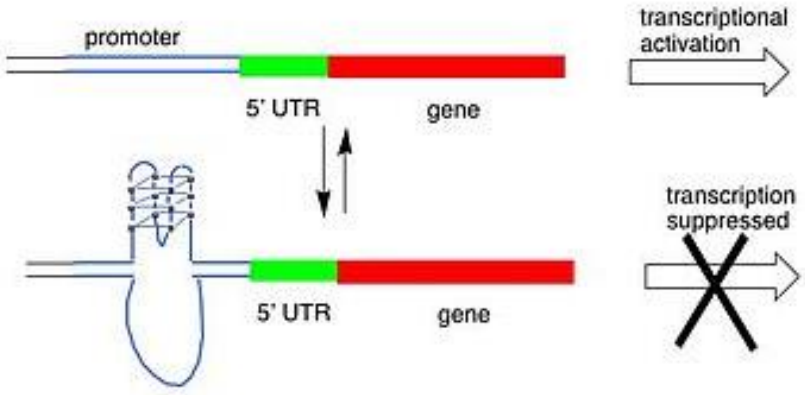
Потенциальную возможность того, что участок ДНК может образовать G4 принято определять паттерном

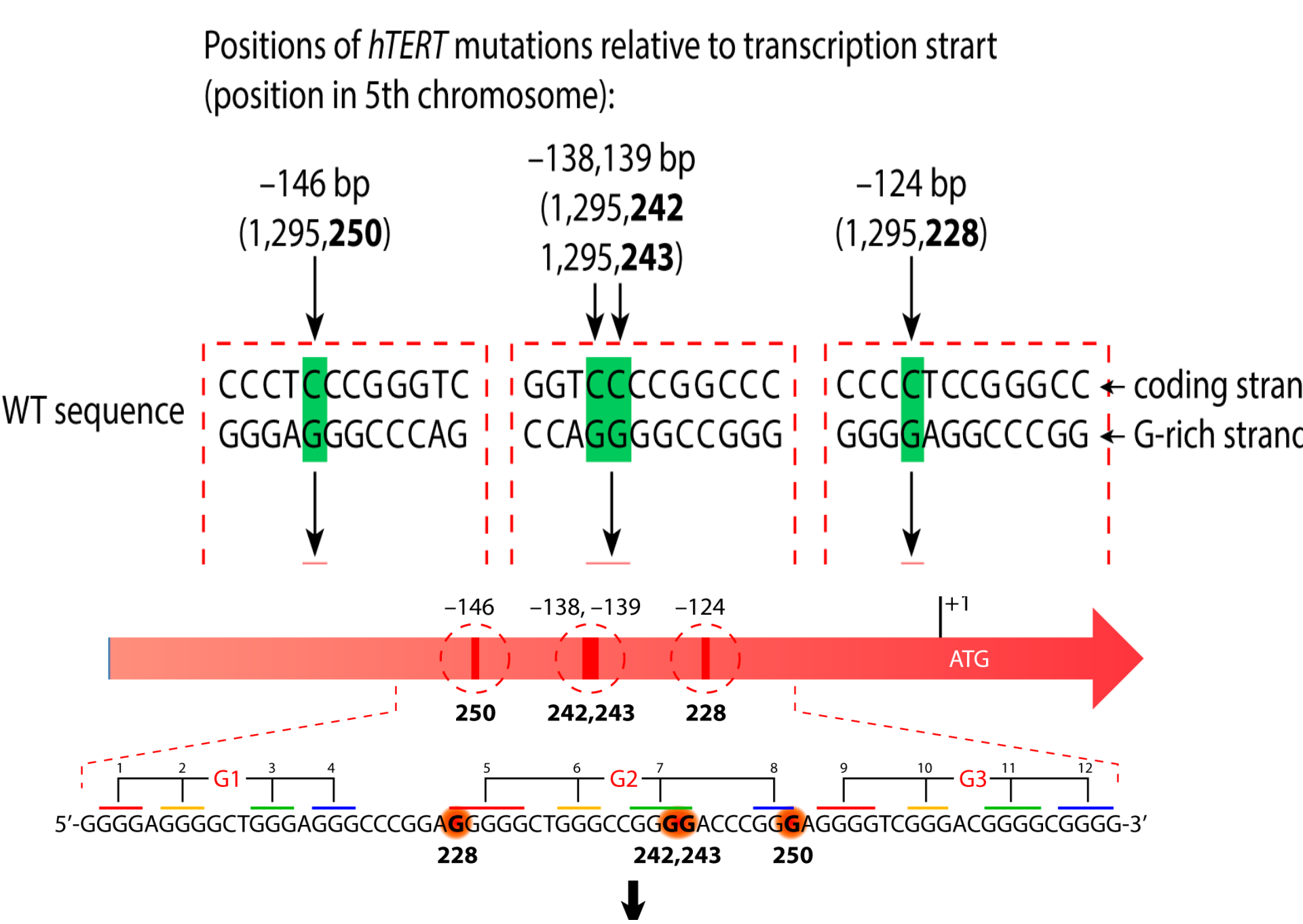
$G_{3+}\ L_{1-7}\ G_{3+}\ L_{1-7}\ G_{3+}\ L_{1-7}\ G_{3+}$ (объяснить)

Такая последовательность может образовать G4, может и не образовать. Если это происходит в промоторе перед стартом транскрипции, это влияет на транскрипцию.








Positions of hTERT mutations relative to transcription strart (position in 5th chromosome):
−146 bp (1,295,250)
−138,139 bp (1,295,242 1,295,243)
−124 bp (1,295,228)
WT sequence
CCCTCCCGGGTC
GGGAGGGCCCAG
GGTCCCCGGCCC
CCAGGGGCCGGG
CCCCTCCGGGCC ← coding stran
GGGGAGGCCCGG ← G-rich strand
−146
−138, −139
−124
+1
ATG
250
242,243
228
G1
G2
G3
5′-GGGGAGGGGCTGGGAGGGCCCGGAGGGGGCTGGGCCGGGGACCCGGGAGGGGTCGGGACGGGGCGGGG-3′
228
242,243
250

1.Повышение частоты мутаций в окрестности экспериментально подтвержденных G-квадруплексов в геноме человека

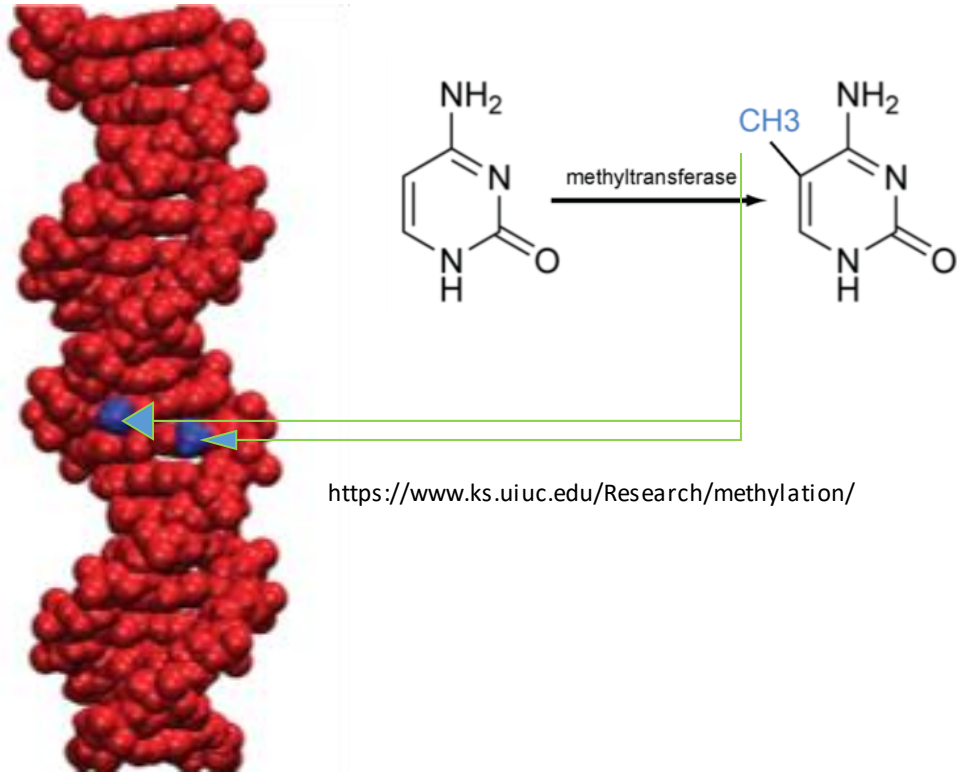
Автор Вера Панова

1. Введение: в промоторе гена hTERT есть три G-квадруплекса (G4).
    1. *In vitro* показано, что MutL из системы репарации связывается с G4 и система MMR неэффективна в окрестности G4 (Pavlova et al., 2022)
    2. a statistically significant higher frequency of nucleotide substitutions in the conserved G4 motifs compared to the surrounding regions was confirmed only for the order Primates. These data support the assumption that G4s can interfere with the DNA repair process (Panova et al., 2023)

7. CpG Метилирование ДНК

Пример сигнала, основанного на модификации ДНК

2. Точная последовательность. CpG[1)]-methylation in mammals.



Метилирование ДНК типа mC5 см. на рисунке.

У млекопитающих 60-80% сайтов CpG метилированы.

Метилированные цитозины смотрят в большую борозду ДНК, не нарушают структуру ДНК, сохраняется возможность основных операций с ДНК (репликация и т.п.)

Метилированные цитозины являются важными сигналами для разных процессов

[1)] CpG – динуклеотид; p значит фосфат, чтобы не путать с комплементарной парой CG

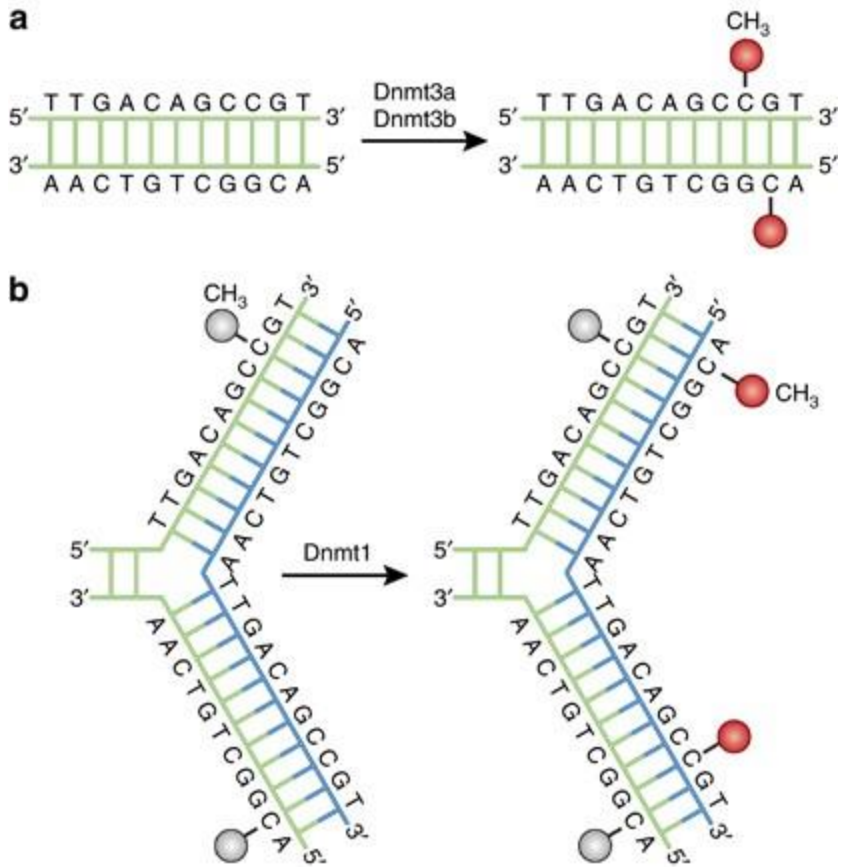
Фрагмент паттерна метилирования

https://www.labclinics.com/wp-content/uploads/2022/12/DNA_Methylation_figure_1-1024x357-1024x357-1.png

Паттерны метилирования CpG в геномах млекопитающих являются т.н. ЭПИГЕНЕТИЧЕСКИМИ сигналами, влияющими на экспрессию многих генов, и наследующимися при делении клеток.

Эти паттерны имеют важное значение в нормальном развитии человека, старении, онкогенезе и других заболеваниях.

Воспроизведение паттерна метилирования при делении клетки



Важно, что в CpG динуклеотиде поддерживается метилирование обоих цитозинов.

Поэтому при репликации в обоих дочерних ДНК материнская цепочка метилирована а новая – нет

ДНК метилтрансфераза DNMT1 ищет такие полуметилированные сайты и метилирует их по второй цепочке рис. b
[7-4]

Литература

[7-1] Cain JA et al., Intragenic CpG Islands and Their Impact on Gene Regulation. Front Cell Dev Biol. 2022

[7-2]Moore LD et al., DNA methylation and its basic function. Neuropsychopharmacology. 2013

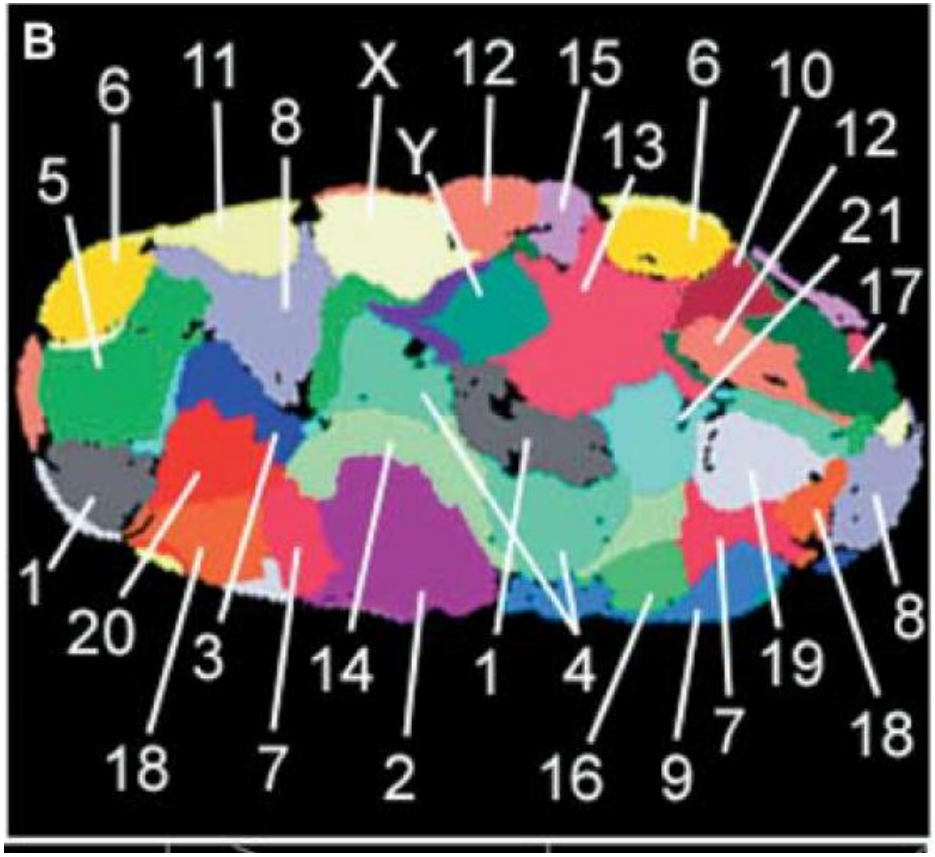
[7-3] Sergeeva A, et al. Mechanisms of human DNA methylation, alteration of methylation patterns in physiological processes and oncology. Gene. 2023

[7-4] Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013

6. Доступность ДНК

Как регулятор экспрессии генов

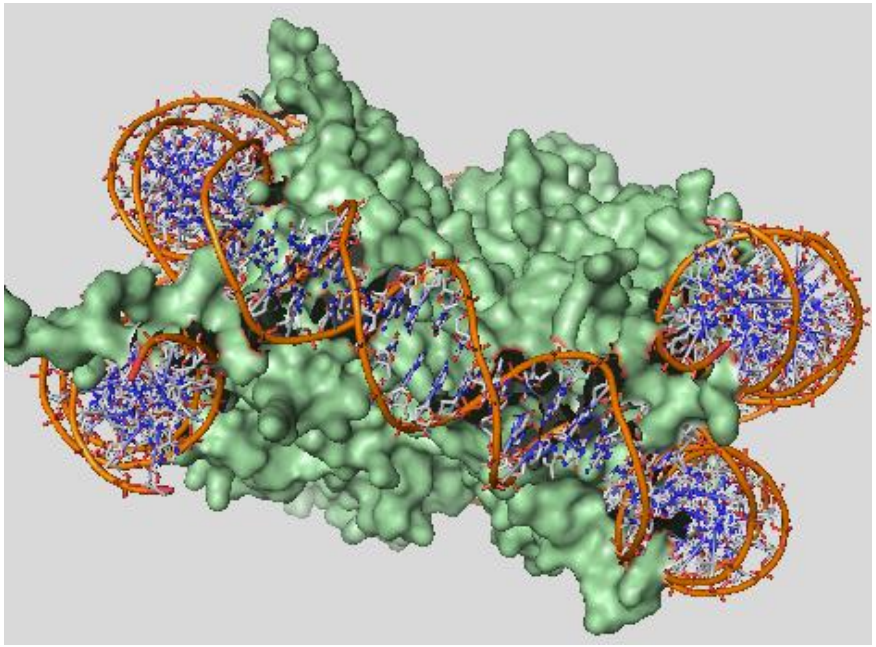
[6-1] Хромосомы плотно упакованы в ядре клетки. Разные ДНК покрашены в разные цвета, одинаковые по последовательности (>99%) – в одинаковые цвета.



Клетка фибробласта человека

Клетка находится перед стадией деления: каждая хромосома состоит из 2х одинаковых ДНК после удвоения (репликации). При делении клетки он разойдутся в разные дочерние клетки

Для эукариот дело усложняется доступностью ДНК для белков



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

PDB код забыл
3LEL тоже X-ray нуклеосомы, 2019 год

Сложнее с доступностью на более высоких уровнях организации хроматина. [6-2]

Даже у прокариот начали изучать.

Литература

[6-1] Bolzer A et al. Three-dimensional maps of all chromosomes in fibroblast nuclei and prometaphase rosettes. PLoS Biol. 2005

[6-2] Selivanovskiy AV, Molodova MN, Khrameeva EE, Ulianov SV, Razin SV. Liquid condensates: a new barrier to loop extrusion? Cell Mol Life Sci. 2025 80.

IV. Технологии поиска и описания сигналов в геноме

Сегодня разбираем случаи, когда для сигнала есть материал обучения – десятки последовательностей сигнала известны.

Способы находки известного сигнала в геноме

- Точная последовательность
- Последовательность с вариациями. Паттерн.
- PSSM по выравниванию последовательностей
- Машинное обучение – не ко мне. Устарел. Хочу понимать а не гадать.
  Возможность построения машинного мышления была предметом главного препирательства Бонгарда и Нюберга (1960 -1970 г.г.)
  https://www.trv-science.ru/2024/12/dialogi-bongarda-i-nyuberga-o-postroenii-mashinnogo-myshleniya/
  ПОЧИТАЙТЕ – ЛЮБОПЫТНО И АКТУАЛЬНО!

Последний проект Бонгарда Михаила Моисеевича (см. wiki) был «животное», Нюберг принципиально отрицал возможность машинного мышления.

Поговаривали, что Мика Бонгард делает гадалку, которая предсказывает следующее случайное число!!

1. Точная последовательность. Поли-А

1) AAAAAAAAAAAAAAAAAAAAA (от десятков до сотен букв)

Сигнал зрелой мРНК у эукариот.

[1-1] wiki https://en.wikipedia.org/wiki/Polyadenylation
[1-2] Rodríguez-Molina JB, Turtola M. Birth of a poly(A) tail: mechanisms and control of mRNA polyadenylation. FEBS Open Bio. 2023

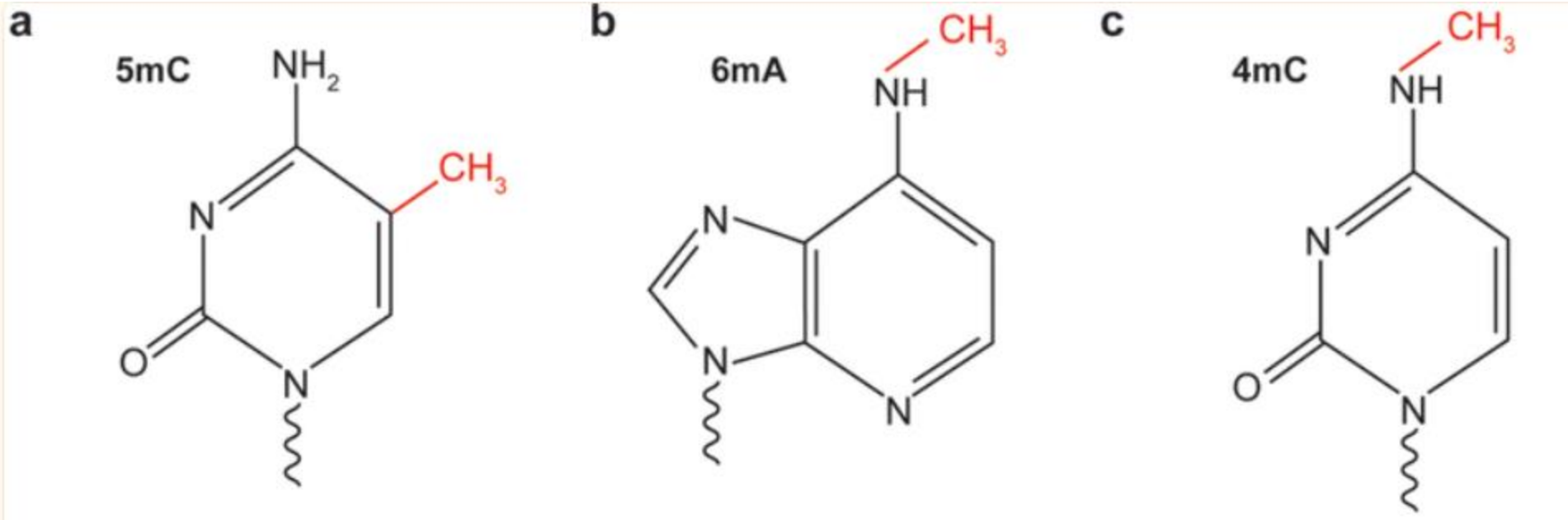
2.Сайты метилирования систем рестрикции-модификации

Сигнал определяется ПАТТЕРНОМ

2. Сайты метилирования у прокариот

Паттерны

Типы метилирования ДНК у прокариот



(*) В последние годы метилирование m6A и m4C обнаружены и у эукариот, но пока они изучены недостаточно и мало распространены

Ambiguous nucleotide codes

Symbol	Meaning	Description Origin	complement
G	G	Guanine	C
C	C	Cytosine	G
A	A	Adenine	T
T	T	Thymine	A
R	G or A	puRine	Y
Y	T or C	pYrimidine	R
M	A or C	aMino	K
K	G or T	Ketone	M
H	A or C or T	H follows G in alfabet	D
D	G or A or T	D follows C in alfabet	H
B	G or T or C	B follows A in alfabet	V
V	G or C or A	V follows U in alfabet	B
S	G or C	Strong pairing	S
W	A or T	Weak pairing	W
N	G or A or T or C	aNy	N

Разнообразие сайтов метилирования в системах рестрикции-модификации (P-M)

сайт узнавания	имя системы P-M
CG	M.SssI
GC	M.CviPI
ATTACY	M.Pmu10382II
GGNAC	M.Chy11610IV
CCWGG	M.Msp42II
GACNNNG	M.Tth111I
CAYGAC	AbaPBA3II
CBACAG	M.Csp423IV
CCTGCAGG	M.SbfI
GCGGCCGC	M.NotI

Условные обозначения.
Метилируемые основания, по прямой и по обратной цепочке, выделены цветом

Цвет	тип метилирования
Фиолетовый	m5C
Синий	m6A
Оранжевый	m4C

Выборка паттернов, с которыми связывается ДНК метилтрансфераза (МТ)

Всего в БД REBASE, хранящей информацию о системах P-M и ДНК метилтрансферазах, более 3 тысяч сайтов МТ, определённых экспериментально

3. Позиционная весовая матрица (PWM)

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

Впервые предложена в работе:

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. **1982**;10(9):2997-3011

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 PWM

w(b,j)	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 ε(b)

В работе Nishida K, Frith MC, Nakai K. Pseudocounts for transcription factor binding sites. Nucleic Acids Res. 2009 Feb;37(3):939-44.
Исследовали вопрос о лучшем выборе псевдоотсчётов для нукл. последовательностей. Заключение авторов:

выбирать ε примерно равным 1, а ε(b) = ε/4
(проверить по статье)

Однако, по прежнему, выбор псевдоотсчётов остаётся на усмотрении авторов и может меняться в зависимости от ситуации

ШАГ 4. Частоты с псевдоотсчётами

F(b,j)	баз. Частоты	e(b)	1	2	3	4	5	6	7	8	9	10	11	12	
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
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
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
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
	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

Шаг 5. Матрица PWM с псевдоотсчётами
Вес последовательности относительно PWM

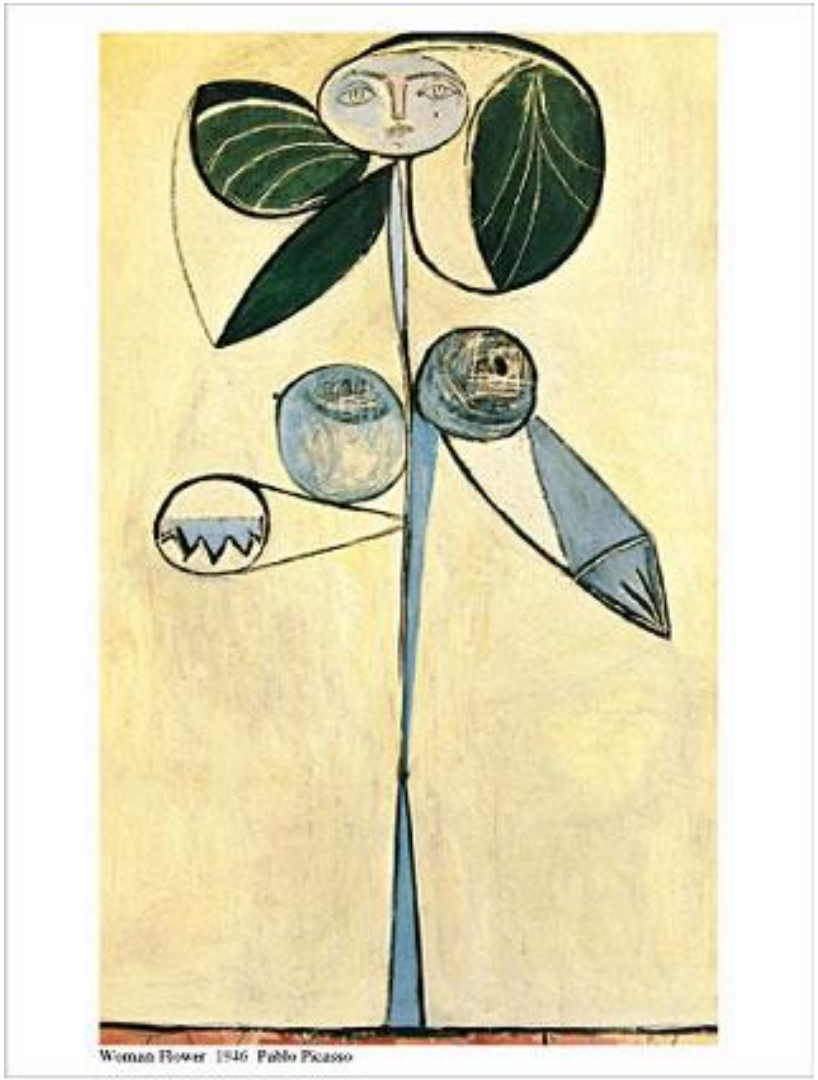
b/j	p(b)	ε(b)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0.15	0.1	1.6	0.0	-3.0	-0.6	1.9	1.9	1.6	-3.0	-3.0	-0.6	-3.0	-3.0	0.7	-3.0	-0.6	-3.0
G	0.35	0.1	-0.8	-0.8	1.0	-3.8	-3.8	-3.8	-3.8	-3.8	1.0	-0.1	-1.5	-1.5	-0.4	-1.5	0.1	-3.8
T	0.15	0.1	-0.6	-0.6	-3.0	-3.0	-3.0	-3.0	0.0	-3.0	-3.0	1.4	1.8	1.8	0.9	-0.6	0.4	1.7
C	0.35	0.1	-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	-0.8

		A	**G**	**G**	**C**	**T**	**A**	**A**	**C**	**G**	**G**	**T**	**T**	**A**	**T**	**T**	**T**
W =	**10.9**	**1.6**	**-0.8**	**1.0**	**0.9**	**-3.0**	**1.9**	**1.6**	**1.0**	**1.0**	**-0.1**	**1.8**	**1.8**	**0.7**	**-0.6**	**0.4**	**1.7**
		G	**C**	**C**	**T**	**A**	**C**	**C**	**C**	**C**	**A**	**T**	**T**	**A**	**T**	**T**	**T**
W=	-13.2	**-0.8**	**0.5**	**-3.8**	**-3.0**	**-3.8**	**-3.8**	**-1.5**	**1.0**	**-3.8**	**-0.6**	**1.8**	**1.8**	**0.9**	**-0.6**	**0.4**	**1.7**

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


Woman Flower 1946 Pablo Picasso

Как- то так

4.Сигнал старта трансляции у прокариот - последовательность Shine-Dalgarno (SD)

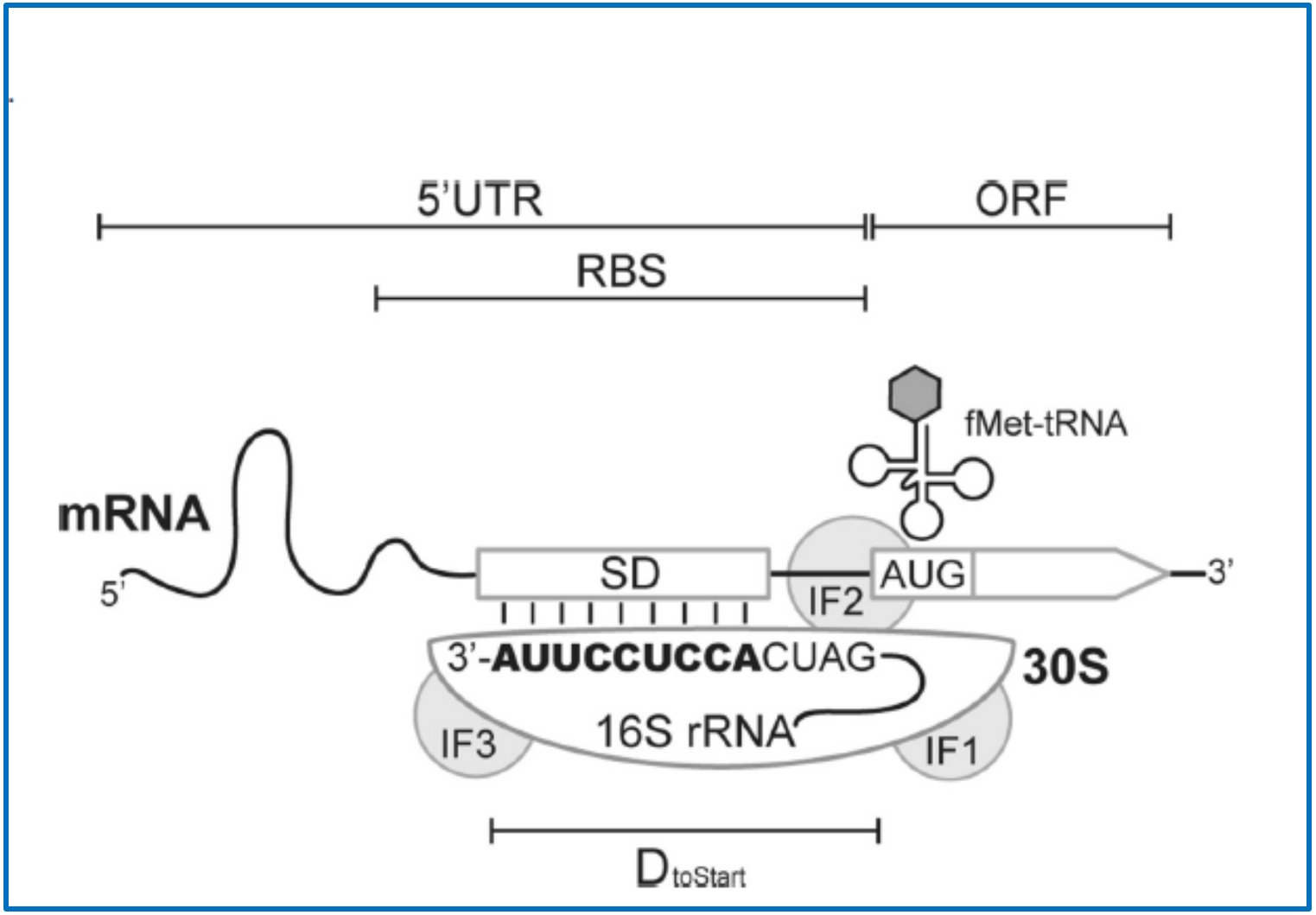
Последовательность РНК узнаётся 16S rRNA

Shine и Dalgarno обнаружили, что у *E.coli* мотив GGAGGU посадки Рибосомы на мРНК комплементарен последовательности 3’ конца 16S rRNA (ACCUCCUUA in E. coli) [4-1].

С тех пор сайт посадки Рибосомы называется SD, а последовательность 3’ конца 16S - anti-Shine Dalgarno or ASD

Было обнаружено, что комплементарность SD-ASD способствует эффективности сигнала SD но не является строго обязательной [4-3, 4-2]

[4-1] Shine J, Dalgarno L. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci U S A. 1974
[4-2] Saito K, Green R, Buskirk AR. Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. Elife. 2020
[4-3] Wen JD, Kuo ST, Chou HD. The diversity of Shine-Dalgarno sequences sheds light on the evolution of translation initiation. RNA Biol. 2021

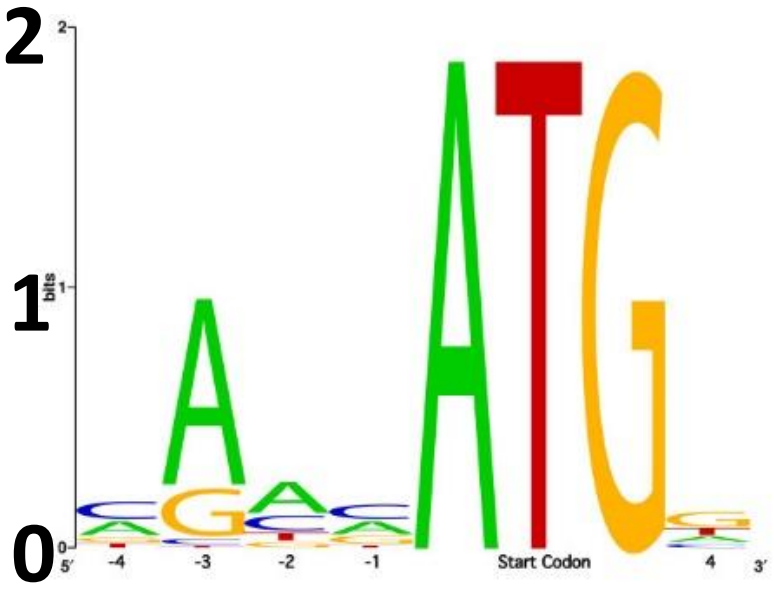


[4-3] Wen JD, Kuo ST, Chou HD. The diversity of Shine-Dalgarno sequences sheds light on the evolution of translation initiation. RNA Biol. 2021

5.Инициации трансляции у эукариот

Сильно вариабльная последовательность РНК

5.Последовательность Козак человека в инициации трансляции у эукариот



Marilyn Kozak в 1986 году обнаружила оптимальное окружение старт кодона ATG для эффективности инициации трансляции у эукариот.
Изображено на рис. слева с помощью LOGO. [5-1]

Неожиданно нашёл недавние ссылки на последовательность Козак в связи с генной инженерией [5-2], иследование последовательностей Козак у млекопитающих [5-3] и даже бактерий [5-4]

Поэтому решил оставить такое задание для выбора.

Литература

[5-1] Kozak M. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. Proc Natl Acad Sci U S A. 1986
[5-2] Kondratov O, Zolotukhin S. Exploring the Comprehensive Kozak Sequence Landscape for AAV Production in Sf9 System. Viruses. 2023 Sep 23;15(10)
[5-3] Noderer WL, Flockhart RJ, Bhaduri A, Diaz de Arce AJ, Zhang J, Khavari PA, Wang CL. Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol Syst Biol. 2014
[5-4] Saito K, Green R, Buskirk AR. Translational initiation in E. coli occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. Elife. 2020