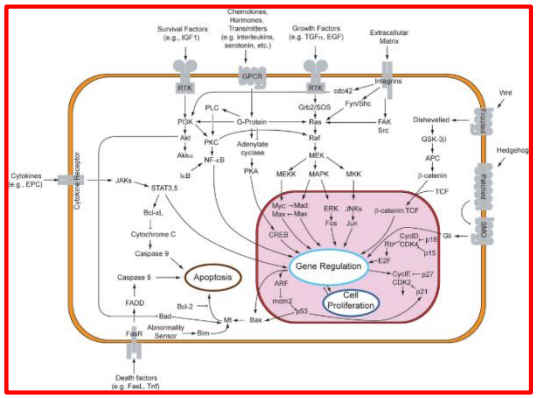
Гомологичные ДНК (с почти одинаковыми последовательностями) считать за одну.
Так считается в БД Reference genomes NCBI

№	Чей геном	вид	Число молекул ДНК	Длина генома в п.н.	Число генов белков
1	млекопитающие	человек			
2[1])	бактерия				
3[2])					

Понятно, до некоторой степени☺

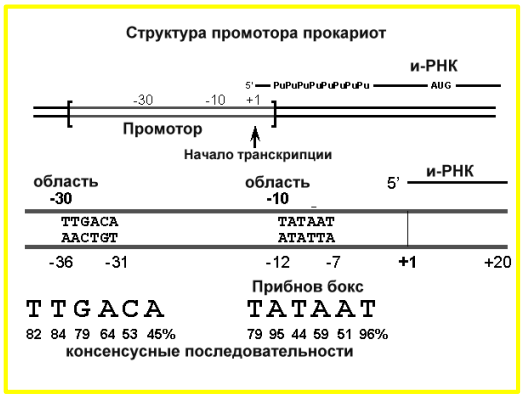
https://vrnikc.ru/wp-content/uploads/2021/04/neverbalnye-1024x640.jpg

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

Сигнал несет информацию

Носитель сигнала – светофор

Какие значения –

красный, зелёный, не работает

Инфа. – можно ли переходить улицу

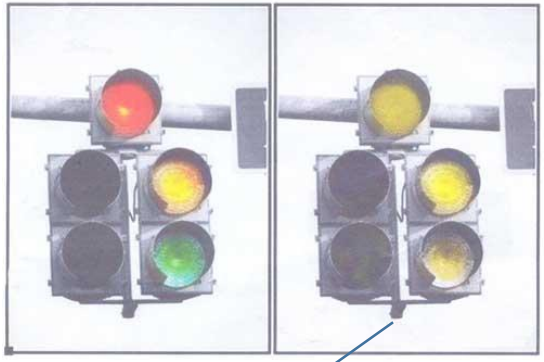
Кому – пешеходу

и любому, кто реагирует на сигнал

Правильная реакция –

“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)

Много сигналов одновременно влияют на действия пешехода и бродячей собаки

- Красный цвет
- Отсутствие машин
- Реакция других пешеходов
- Другие – вспомните)))
  Про барбоса на другой стороне шоссе

Так и в клетке, и в геноме.
Поэтому на каждое правило есть исключения

Сила сигнала –

одна из важных характеристик 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

Геном – носитель всей информации, передаваемой от предка к потомку.

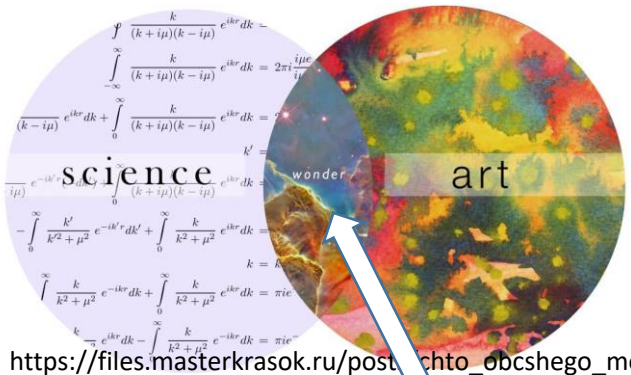
ВСЕ СОГЛАСНЫ!?

Всей ли???

А при делении клетки?

ИНФОРМАЦИЯ НУЖНА ДЛЯ АДРЕСАТОВ:
в клетке – белков (яйцеклетки)

Наука про геном точная, но на грани с искусством. Всегда есть исключения



https://files.masterkrasok.ru/post[illegible]chto_obcshego_mezhdu_naukoj_i_iskusstvom.jpeg

https://www.surgpu.ru/media/uploads/2022/03/30/untitled_8.jpg

Как много наука знает об информации, закодированной в геноме!!!

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

- **Функции белков и метаболические пути**

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

Как мало наука знает об информации, закодированной в геноме!

Как закодирована в геноме информация о

* фенотипе будущего взрослого человека – в геноме зародыша!
(черты лица, пропорции тела и конечностей, характера; часто дети похожи на родителей, т.е. это записано в геноме, как?)

* составах протеомов в клетках разных тканей

* поведении пчёл и пчелиного роя

Клетки разных тканей человека

- Имеют одинаковый геном (модификации не в счёт) С оговорками
- Протеомы клеток из разных тканей существенно различаются Меряют по транскриптомам, mass spectrometry и др.
- Протеом – совокупность разных белков и их процентное соотношение Белки – разные, если транслируются с разных зрелых мРНК, т.е. имеют разные последовательности а.к.о.

Где и как записаны в геноме зародыша составы протеомов в клетках разных тканей

Данные о протеомах разных тканей собираются в проекте HUPO Human Proteome Project

Современные технологии позволяют определять протеомы клеток отдельных тканей человека.

Фантазии. Может, когда-нибудь, мутируя методами генной инженерии геном мышиного зародыша, можно будет направлено изменять состав протеома гепатоцитов??? Это было бы торжество науки!!!

I-1. Omenn GS et al., The 2022 Report on the Human Proteome from the HUPO Human Proteome Project. J Proteome Res. 2023

I-2 Adhikari S, A high-stringency blueprint of the human proteome. Nat Commun. 2020

Кто и как решает какие белки и в каком количестве экспрессировать в клетках данной ткани?

Факт “на лице”:

Наука знает, что это закодировано в геноме.
Но КАК?

Фантазии: сложно устроенные взаимодействующие регуляторные каскады. Мутируем в зародыше регулятор транскрипции (TF) A, он репрессирует TF B, который активирует экспрессию TF C, и ещё пара сотен таких шагов и в результате вырастает мышка с костями удвоенной длины!!!

Современный уровень. Экспериментально показаны способы регуляции экспрессии отдельных генов и генов определённых метаболических путей.

Важно знать ВСЕ сигналы в геноме и всё о каждом из них

Характеристики сигнала в геноме

Носитель – ДНК (по определению)

Вид сигнала – мотив: характерная последовательность участка ДНК (уточнения в примерах)

Адресат – чаще всего, определенный белок или комплекс белков; РНК тоже может участвовать в узнавании мотива

Реакция – белок узнаёт мотив и связывается с ДНК в этом месте

Результат - зависит от белка, что он делает

Название сигнала – его важно знать

(см. дом. задание 1)

II. Основные сигналы

Следует выучить к коллоквиуму

1. Репликация у бактерий

1. 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 [1-2]

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

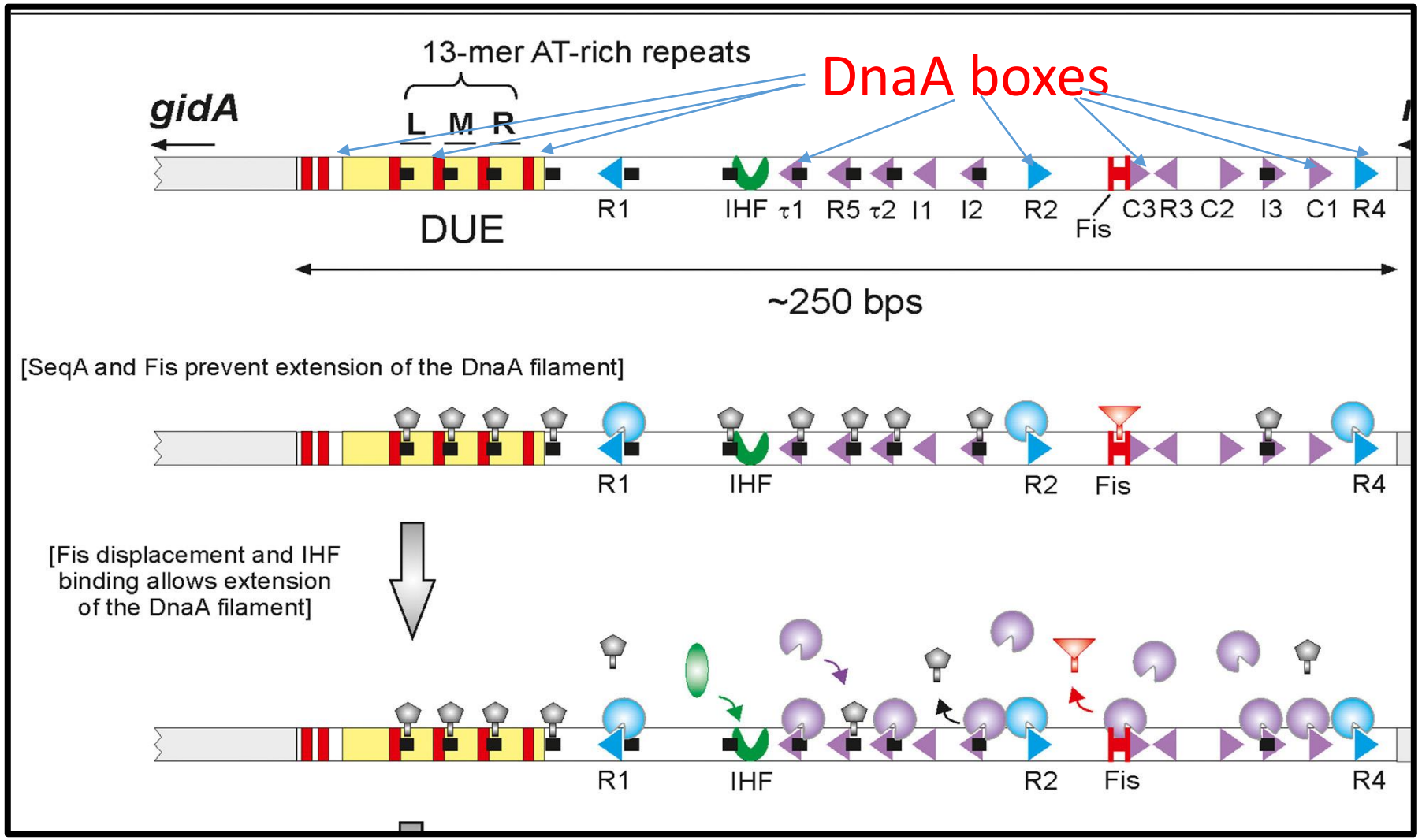
Область oriC вариабельна у бактерий.
Общее у всех – три функциональных участка

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

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

(3)Последовательности, узнаваемые другими регуляторными белками

[1-2] Wolanski et al., oriC-encoded instructions for the initiation of bacterial chromosome replication, 2015



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

Литература

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

[1-2] 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.

[1-3] Wegrzyn KE, Gross M, Uciechowska U, Konieczny I. Replisome Assembly at Bacterial Chromosomes and Iteron Plasmids. Front Mol Biosci. 2016

[1-4] Wegrzyn K, Konieczny I. Toward an understanding of the DNA replication initiation in bacteria. Front Microbiol. 2024

2. 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.

Сигналы сборки RNAP

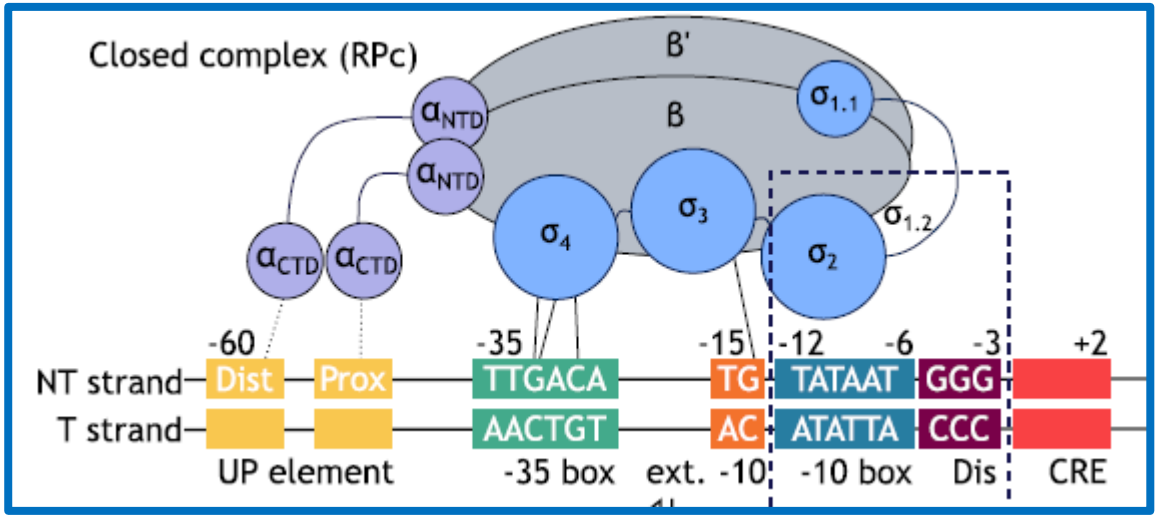
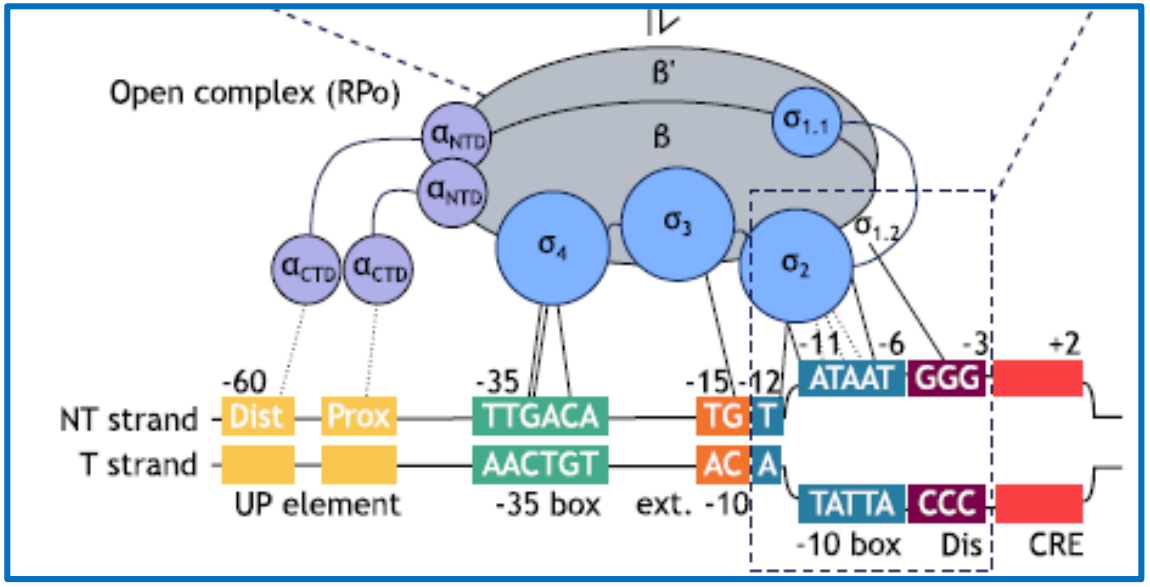


Рис. Кружками одного цвета изображены домены одного и того же белка



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

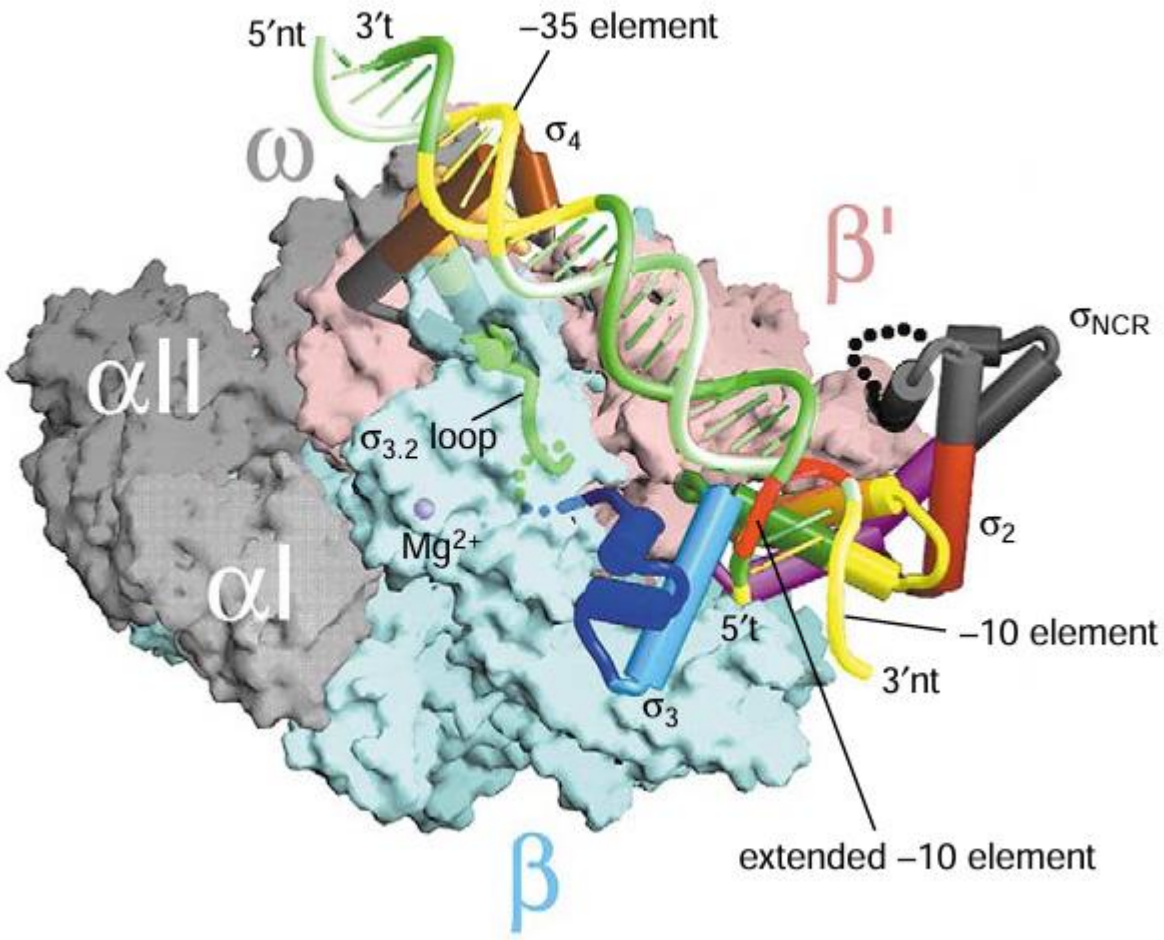
2)Потом собирается закрытый комплекс RNAP

3) Потом расплавляется ДНК и образуется раскрытый комплекс

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

[2-1] Deal C et al., Towards a rational approach to promoter engineering: understanding the complexity of transcription initiation in prokaryotes. FEMS Microbiol Rev. 2024

Вы хочите ω?
Их есть у меня!



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

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

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

[2-1] Deal C et al., Towards a rational approach to promoter engineering: understanding the complexity of transcription initiation in prokaryotes. FEMS Microbiol Rev. 2024

[2-2] 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

[2-3] 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

3.Терминация транскрипции у прокариот

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

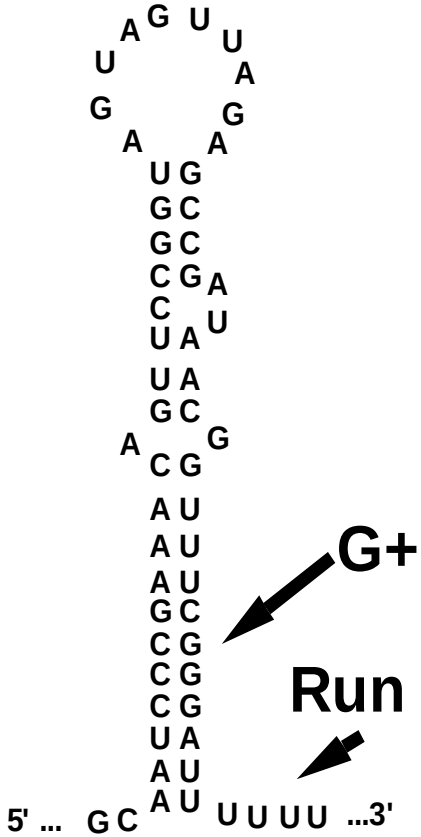
2. Rho – независимая. Сложная шпилька мРНК [3-1]
   web service: http://rssf.i2bc.paris-saclay.fr/toolbox/arnold/

[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



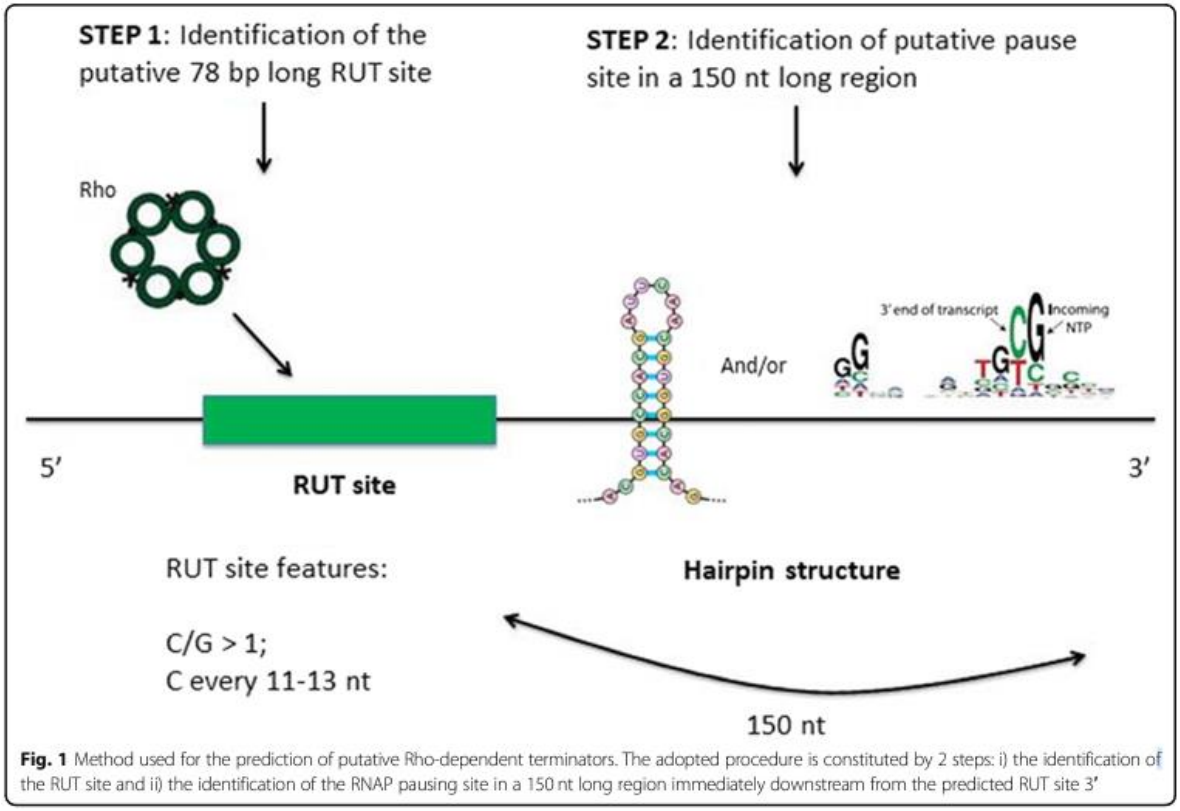


Fig. 1 Method used for the prediction of putative Rho-dependent terminators. The adopted procedure is constituted by 2 steps: i) the identification of the RUT site and ii) the identification of the RNAP pausing site in a 150 nt long region immediately downstream from the predicted RUT site 3'

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

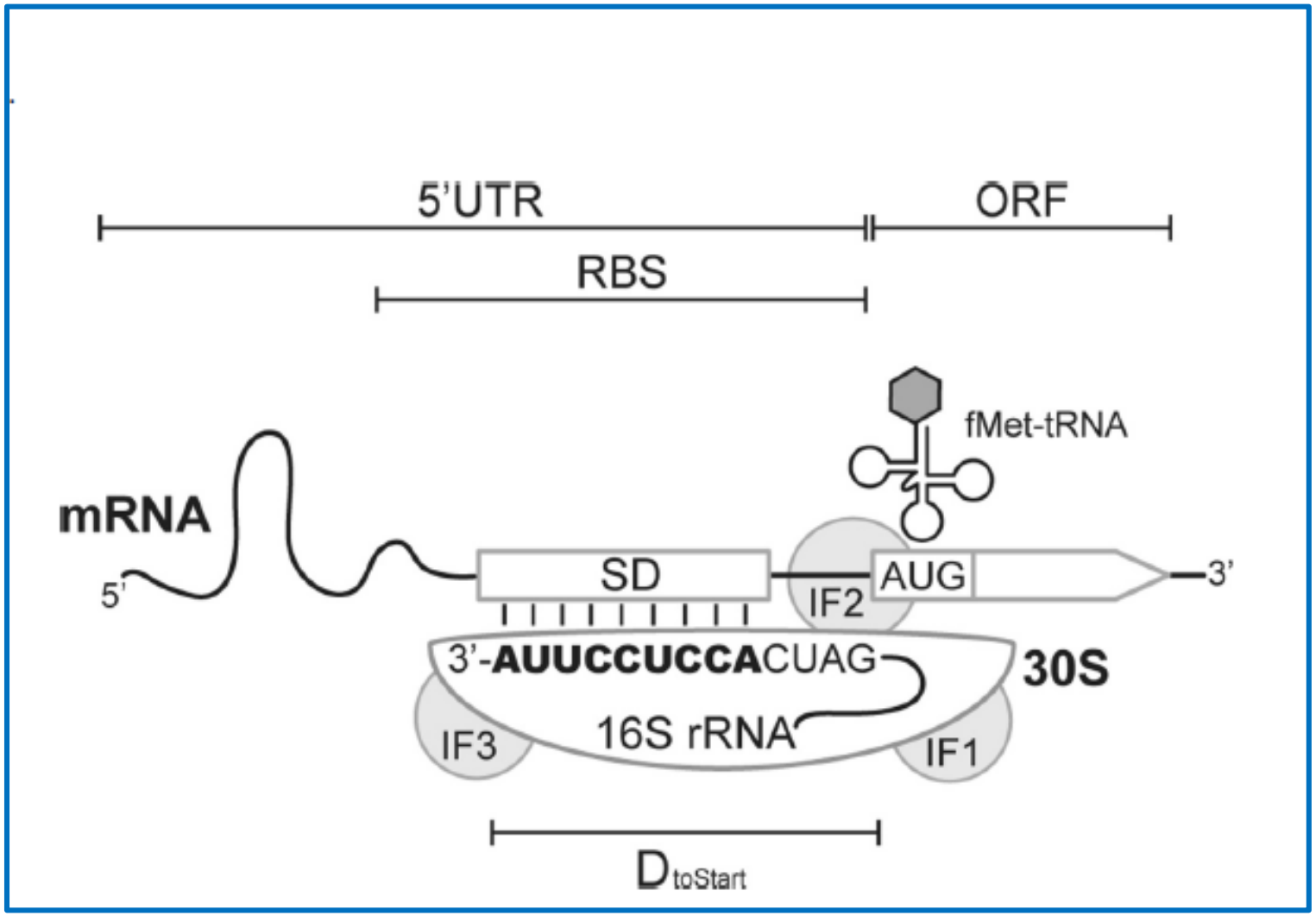
4.Сигнал старта трансляции у прокариот - последовательность Shine-Dalgarno (SD)

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

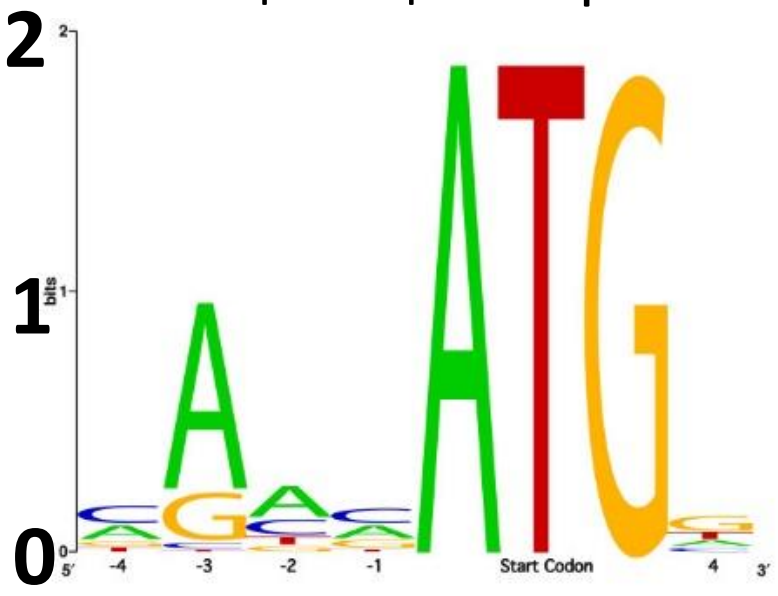
Было обнаружено, что комплементарность SD-ASD способствует эффективности сигнала SD но не является строго обязательной [4-3, 4-2]

[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.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.

Про последовательность Козак в геноме SARS-Cov-2 – пропустим. Надоел

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

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

III. Технологии поиска и описания сигналов в геноме

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

Способы описания мотива

Мотив – биологически значимый сигнал для адресата, чаще всего, белка. Мотивы бывают разные

(i) Известна точная последовательность мотива

(ii) Паттерн – способ описать последовательности, известных мотивов в случае если они похожи, но есть различия (*так, чтобы программа могла находить предполагаемые мотивы в других ДНК*)

(iii) Позиционная весовая матрица (PWM) для описания выравнивания сигналов с более различающимися последовательностями

(iv) Другие (3D структура РНК или ДНК, повторы, ….)

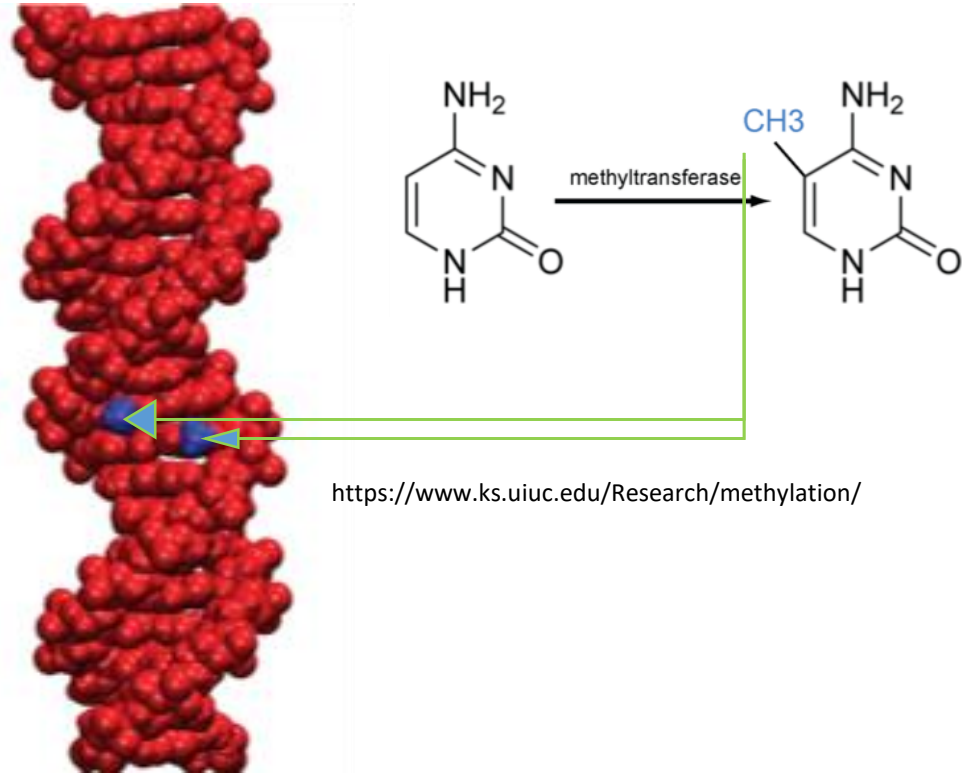
1. Точная последовательность. Поли-А

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



https://www.ks.uiuc.edu/Research/methylation/

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 (реплиация и т.п.)

Метилированные цитозины являются важными сигналами для разных процессов

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

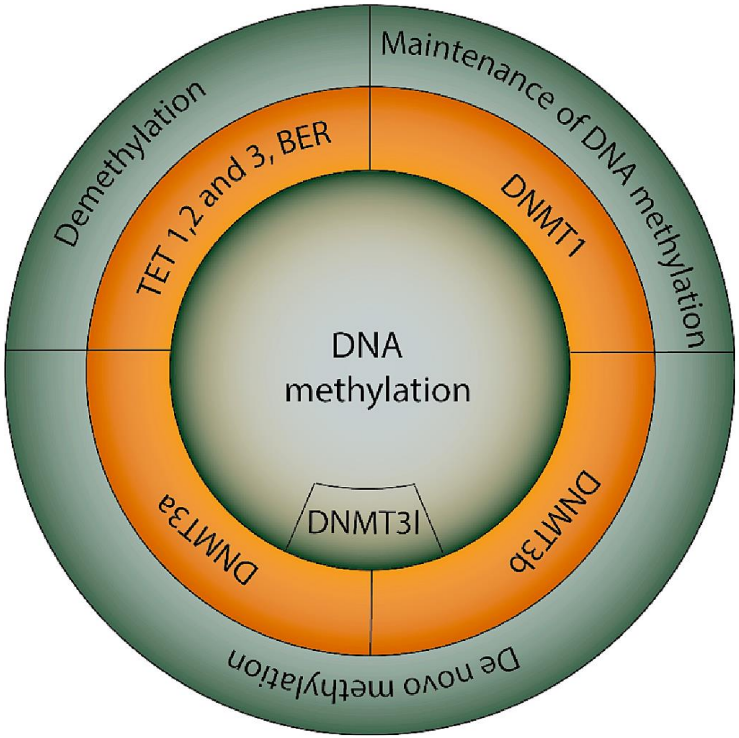
Фрагмент паттерна метилирования

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.

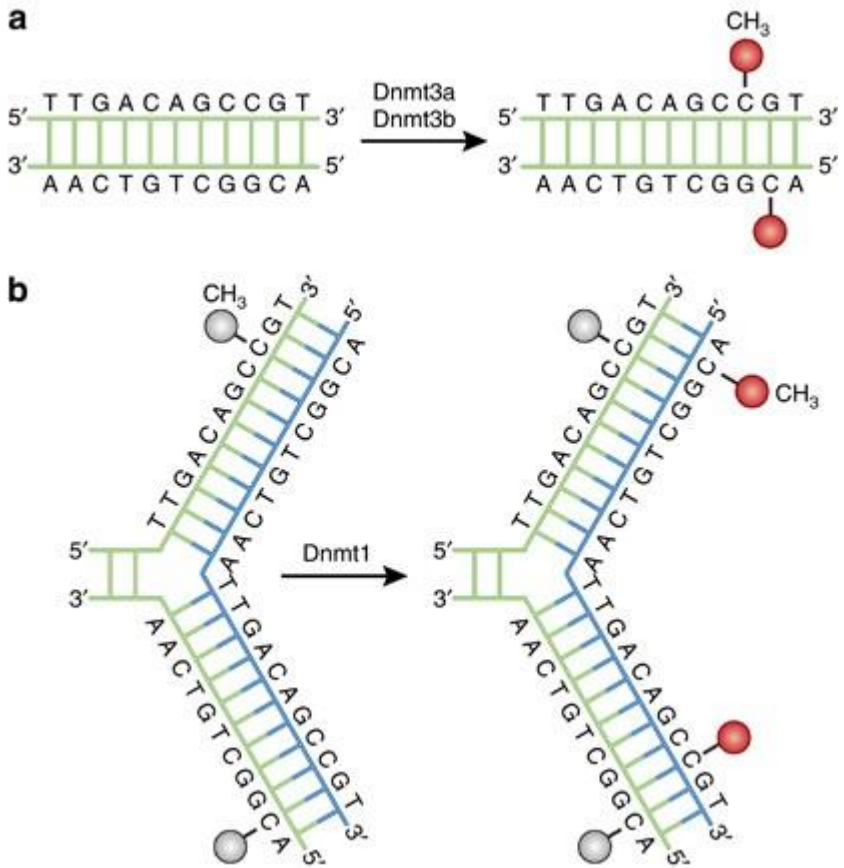
Эти паттерны имеют важное значение в нормальном развитии человека, старении, онкогенезе и других заболеваниях.

Три процесса связанные с CpG-метилированием



1. Метилирование CpG сайтов и паттернов CpG *de novo* для регуляции экспрессии генов (белки DNMT3a, DNMT3b)
2. Поддержка паттерна метилирования при делении клеток (DNMT1)
3. Деметилирование CpG – участвую несколько белков (BER, TET1, TET2)

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
[2-4]

Литература

[2-1] Cain JA et al., Intragenic CpG Islands and Their Impact on Gene Regulation. Front Cell Dev Biol. 2022

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

[2-3] Sergeeva A, et al. Mechanisms of human DNA methylation, alteration of methylation patterns in physiological processes and oncology. Gene. 2023

[2-4] Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013

Сайты метилирования у прокариот

Метилирование ДНК у прокариот гораздо разнообразнее, чем у эукариот.

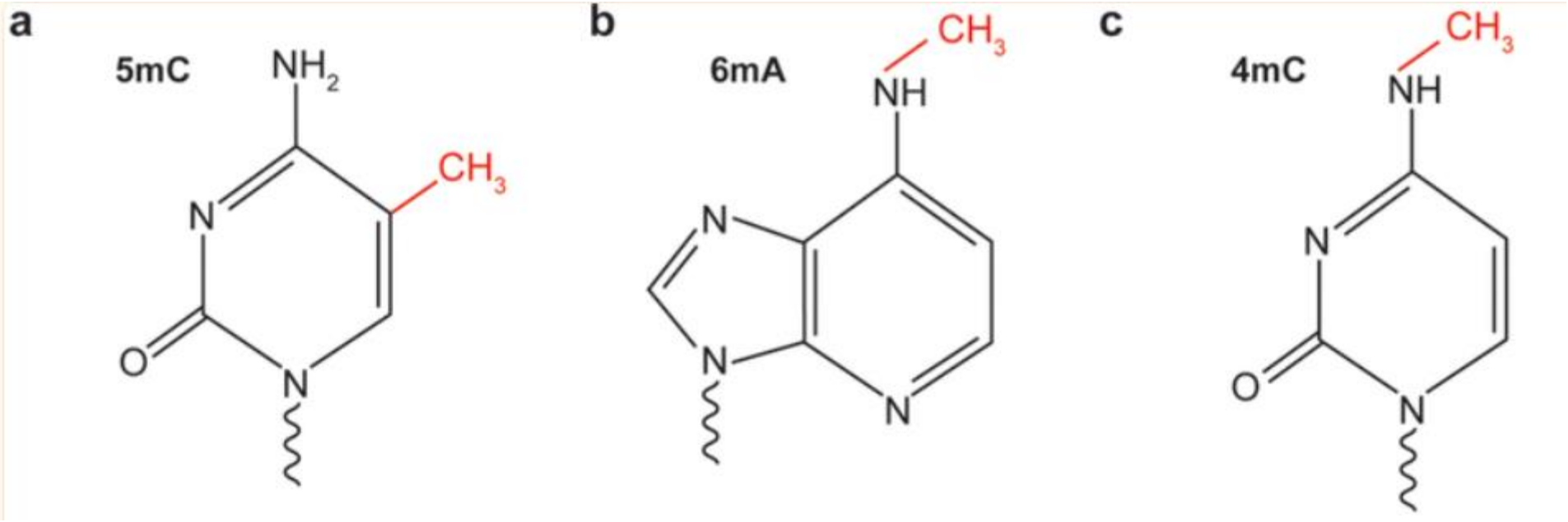
- Кроме метилирования типа m5C – как у эукариот в CpG – есть метилирования ДНК типов m4C, m6A (*)
- Сайты узнавания более разнообразны

Функции метилирования разнообразны

- Выделяется метилирование ДНК в системах рестрикции-модификации (Р-М) как сигнал для отличения своей ДНК от чужеродной – ДНК бактериофагов.
- Регуляция транскрипции отдельных генов (менее распространена, чем у эукариот)
- Пометка материнской цепочки при делении (ниже)
- Другие роли

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

Типы метилирования ДНК



3. Метилирование GATC dam метилтрансферазой - сигнал системе репарации неправильно спаренных оснований ДНК (MMR – MisMatches Repair). GATC есть также в oriC

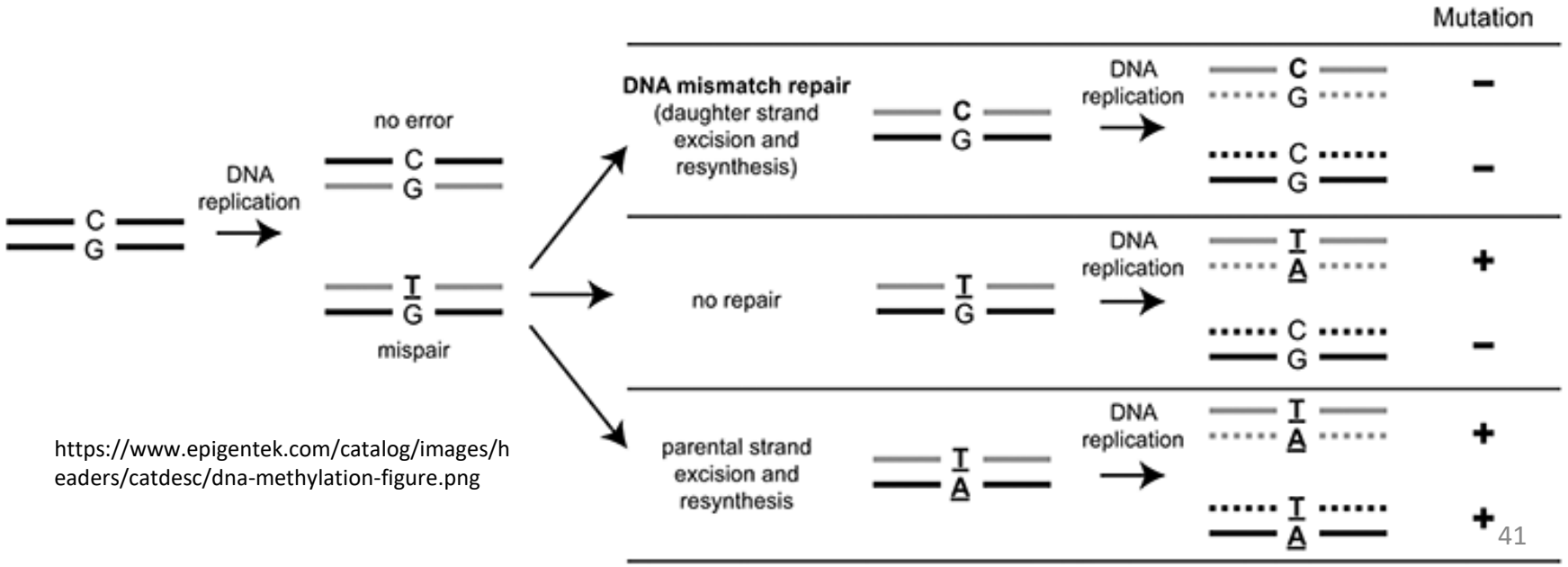
Самая частая мутация в ДНК C => T путем спонтанного дезаминирования. Из пары

C T

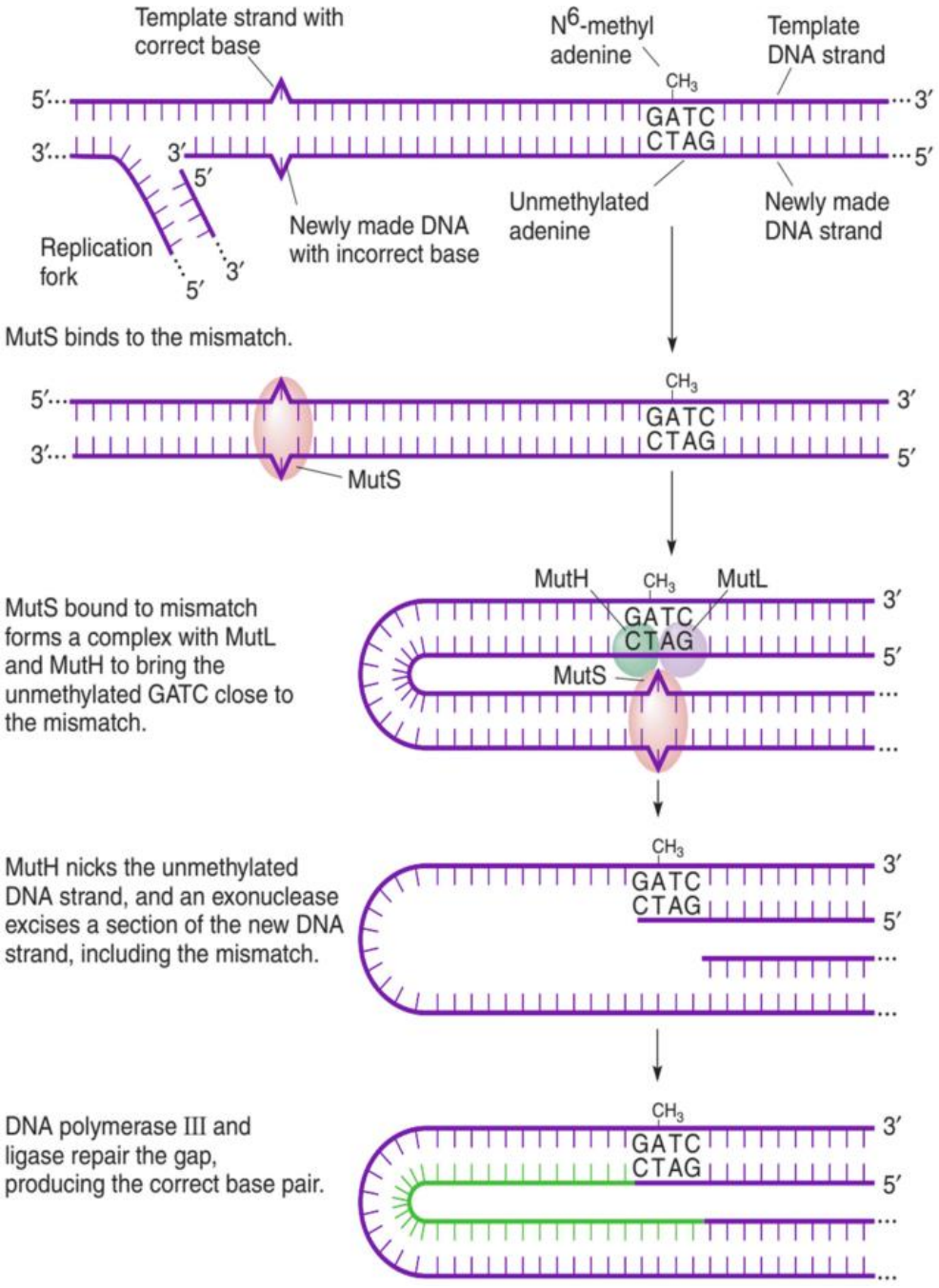
| получается пара | Как известно, G с T образуют в дцДНК одну водородную связь

G G

Такие неправильно спаренные основания часто возникают в процессе репликации Мизмтч пары G-T репарируются. Как решить репарировать в A-T или G-C??? [3-1]



https://www.epigentek.com/catalog/images/headers/catdesc/dna-methylation-figure.png



https://media.springernature.com/lw685/springer-static/image/chp%3A10.1007%2F978-981-16-7041-1_9/MediaObjects/505724_1_En_9_Fig22_HTML.png

Все GATC в материнской ДНК метилированы по Аденину в обеих цепочках. Значит при репликации GATC в материнской цепочке метилирована, а в новой – НЕТ!

Белок MutS из MMR узнаёт мизматч

Привлекает MutL и MutH. MutH скользит по ДНК, изгибая ДНК и сохраняя связь с MutS и MutL до встречи GATC в ДНК.

MutH выбирает цепочку с неметилированным GATC и скушивает её до мизматча

Привлекается ДНК полимераза III, которая достраивает вторую цепочку по комплементарности

[3-2] Not Free. Ray, A. (2022). DNA Mutation, Repair, and Recombination. In: Kar, D., Sarkar, S. (eds) Genetics Fundamentals Notes. Springer, Singapore. https://doi.org/10.1007/978-981-16-7041-1_9

Литература

3-1. Putnam CD. Strand discrimination in DNA mismatch repair. DNA Repair (Amst). 2021

3-2. Not Free. Ray, A. (2022). DNA Mutation, Repair, and Recombination. In: Kar, D., Sarkar, S. (eds) Genetics Fundamentals Notes. Springer, Singapore. https://doi.org/10.1007/978-981-16-7041-1_9

4. 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

ЧТО ЗНАЧАТ БУКВЫ, отличные от A, T, G, C?
См. след слайд

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

Ambiguous nucleotide codes

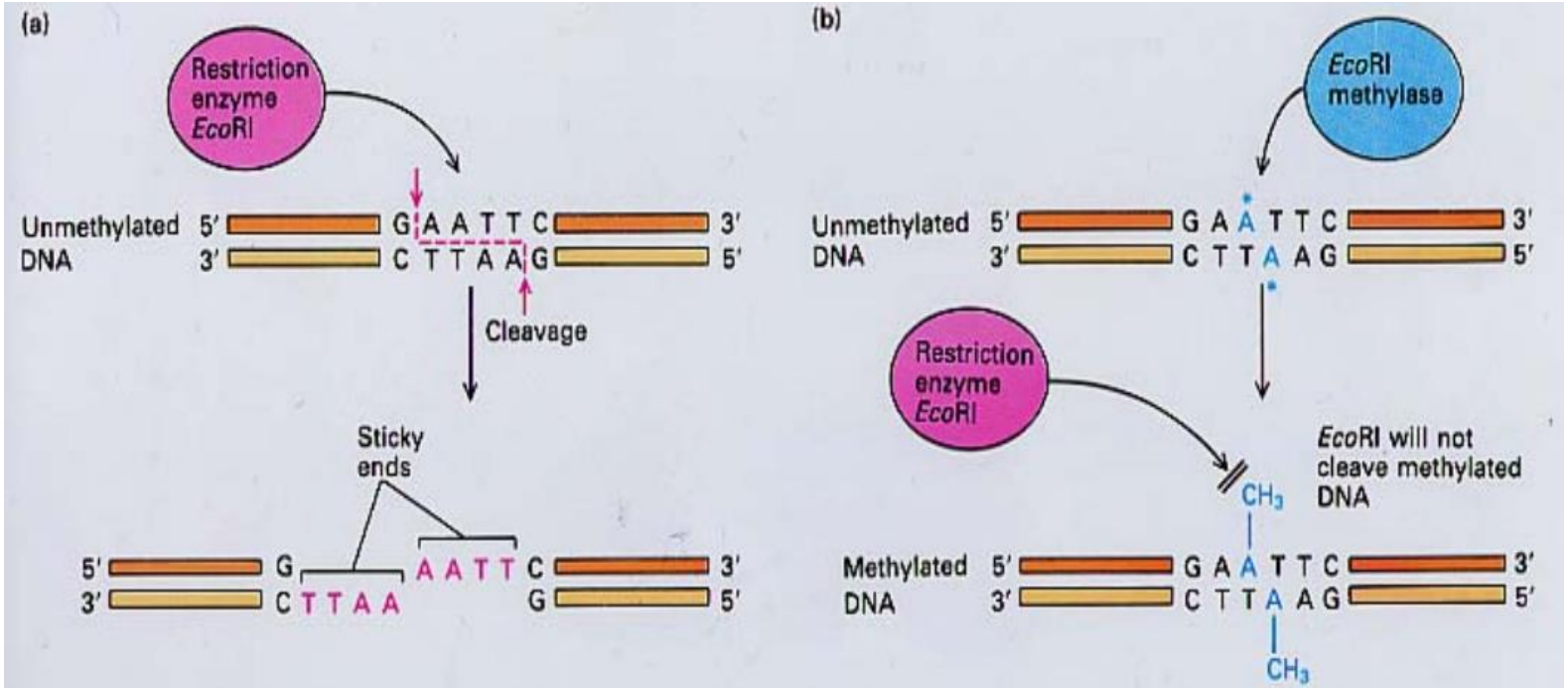
Symbol	Meaning	Description Origin
G	G	**G**uanine
A	A	**A**denine
T	T	**Thymine**
C	C	**C**ytosine
R	G or A	pu**R**ine
Y	T or C	p**Y**rimidine
M	A or C	a**M**ino
K	G or T	**K**etone
S	G or C	**S**trong interaction
W	A or T	**W**eak interaction
H	A or C or T	**H** follows G in alphabet
B	G or T or C	**B** follows A in alphabet
V	G or C or A	**V** follows U in alphabet
D	G or A or T	**D** follows C in alphabet
N	G or A or T or C	a**N**y

Упражнение.

На предыдущем слайде найдите палиндромы (последовательность комплементарной цепочки совпадает с последовательностью прямой цепочки)

Работа системы рестрикции-модификации на примере EcoRI

- Эта система в геноме штаммов E.coli (не всех?)
- В системе два гена:
  R = эндонукдеаза рестрикци EcoRI
  M =ДНК метилтрансфераза M.EcoRI
- Служит для защиты от посторонней – фаговой ДНК



5. Позиционная весовая матрица (PWM)

1. Известны несколько последовательностей одного и того же сигнала одинаковой длины

2. Получаем выравнивание БЕЗ ГЭПОВ!!!

3. Строим матрицу PWM (далее)

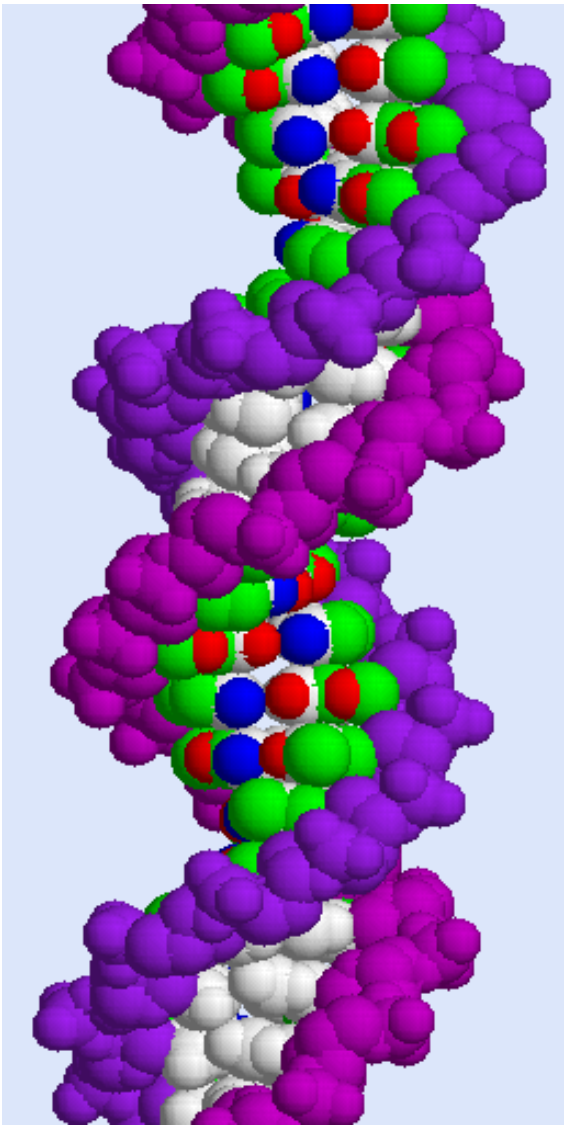
4. С помощью PWM оцениваем сходство любой последовательности с сигналом.

Выбрав порог можем найти все сигналы в геноме.

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

Почему без гэпов. Белки опознают последовательность ДНК по большой бороздке не расплетая ДНК.

Атомы на поверхности малой бороздки (белая) почти не зависят от последовательности



Двойная спираль ДНК

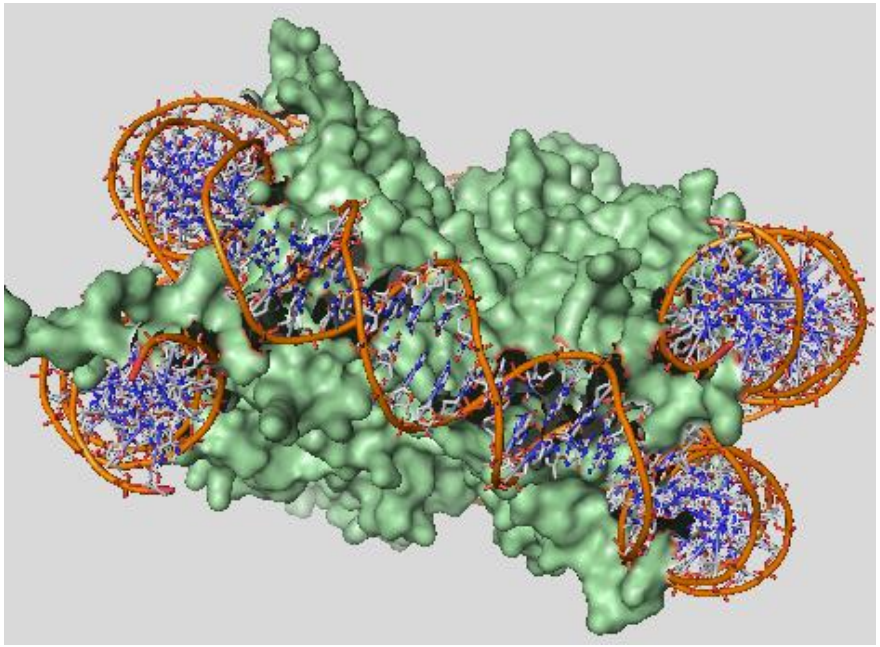
Остов одинаков

Разница оснований закодирована расположением атомов
в большой бороздке:
Синий – азот
красный – кислород
Зелёный – углерод

Белок узнаёт сине-красный код в 3D, а не нуклеотиды!!! Конформация остова ДНК немножко зависит от последовательности оснований.

Делеция пары оснований - очевидно разрушает этот пространственный код. Поэтому **в коротком сигнале ДНК не может быть делеций!**

Для эукариот дело усложняется доступностью ДНК для белков



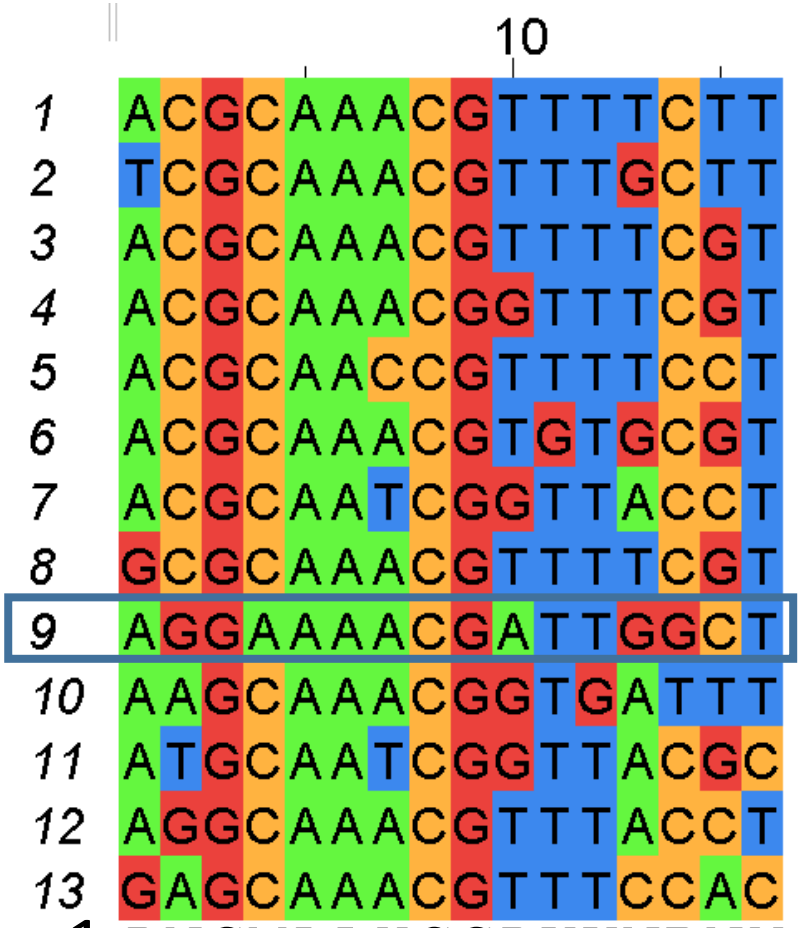
Нуклеосома:
ДНК человека на “катушке” из гистонов: вид сбоку (гистоны – такие белки)

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

Ещё сложнее с доступностью на более высоких уровнях организации хроматина. Даже у прокариот начали изучать.

Чем плохи паттерны-1 и -2?

```
123456789012345 6
ACGCAAACGTTTTCTT
TCGCAAACGTTTGCTT
ACGCAAACGTTTTCGT
ACGCAAACGGTTTCGT
ACGCAACCGTTTTCCT
ACGCAAACGTGTGCGT
ACGCAATCGGTTACCT
GCGCAAACGTTTTCGT
AGGAAAACGATTGGCT
AAGCAAACGGTGATTT
ATGCAATCGGTTACGC
AGGCAAACGTTTACCT
GAGCAAACGTTTCCAC
```



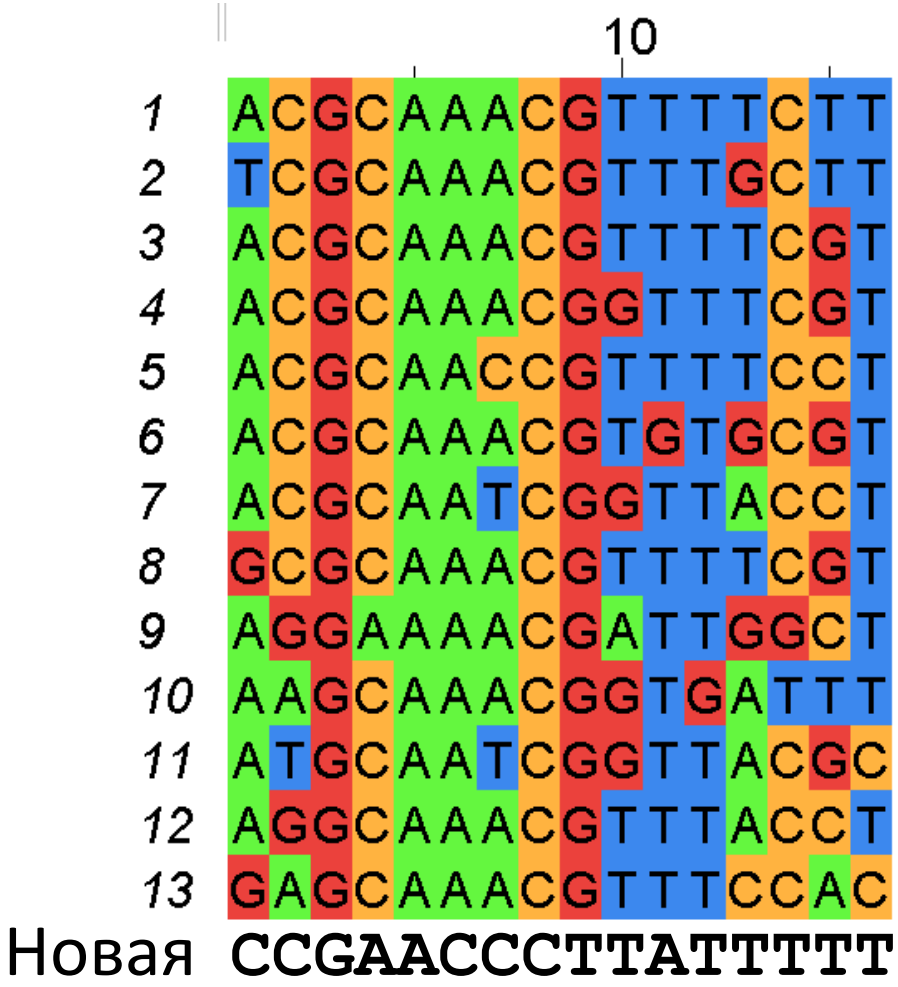
Паттерн-1 DNGMAAHCGDKKNBNY

Паттерн-2 . . **G.AA.CG**. . . . . .

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

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

ИДЕЯ PWM: оценить вес сходства новой последовательности по каждой позиции выравнивания и сложить



Подсчёт числа букв N(b,j) в колонке

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

b\j	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
Все го	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)

Вес буквы = LOG(Отношения правдоподобия)

Вес буквы b в колонке j должен зависеть от частоты этой буквы

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

где N – число последовательностей, в примере N = 13

Большая частота – больший вес!

Не совсем так. Играет значение не сама частота, а превышение частоты над её базовой частотой – частотой в геноме!

Отношение правдоподобия: наблюдаемое/ожидаемое

Для выравнивания $f(b,j)/p(b)$ где p(b) базовая частота буквы.

Значит, $W(b,j) = \ln f(b,j)/p(b)$

Логарифм берется, чтобы веса суммировались
(частоты как вероятности перемножаются)

Матрица весов PWM. Беда с f(b,j) = 0

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																

ЕСЛИ БУКВА **C** не встретилась в колонке **1** ни разу, то её частота f(**C,1**) равна 0. Значит, отношение правдоподобия равно 0.

Значит, вес W(C,1) = минус бесконечность

Псевдоотсчёты: борьба с -inf и не только … Pseudocounts

Идея в том, чтобы немножко изменить ЧАСТОТЫ букв.

(1) Избавляемся 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

Частоты с псевдоотсчётами

b/j	баз. час-тоты	$\varepsilon(b)$	1	2	3	4	5	6	7	8	9	10	11	12	13	1
A	0.15	0.1	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	0.0
G	0.35	0.1	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	0.0
T	0.15	0.1	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	0.0
C	0.35	0.1	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	0.8
Σ	1	0.4	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	1.0

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**

6. Информация и энтропия сигнала

Энтропия – мера неупорядоченности (обозначается H)

Информация противоположна энтропии (обозначается IC)

Чем больше энтропия, тем меньше информации.

Чем больше информации, тем меньше энтропия

Идеалистически, IC = Hbefor – Hafter.

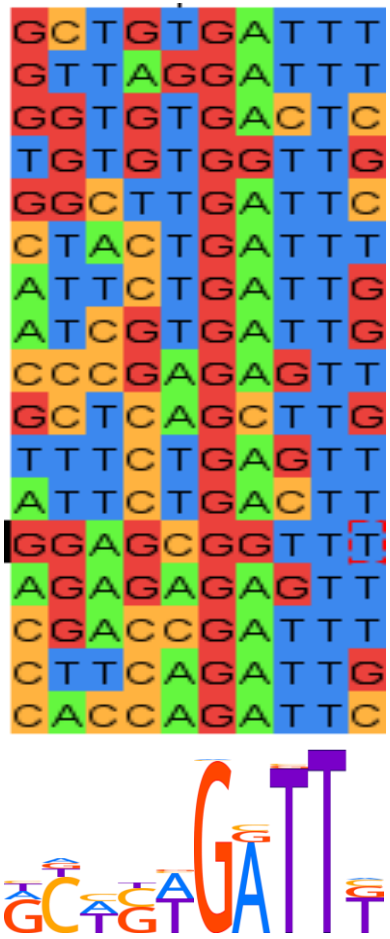
Мера содержания информации в сигнале равна тому, насколько уменьшилась неопределённость (т.е. энтропия) после появления сигнала.

Чем больше IC, тем сильнее сигнал.

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

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

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



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

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

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

Энтропия H по Шеннону и IC

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

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

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

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

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

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

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

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

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

$$IC = \sum_i IC_j$$

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

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

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

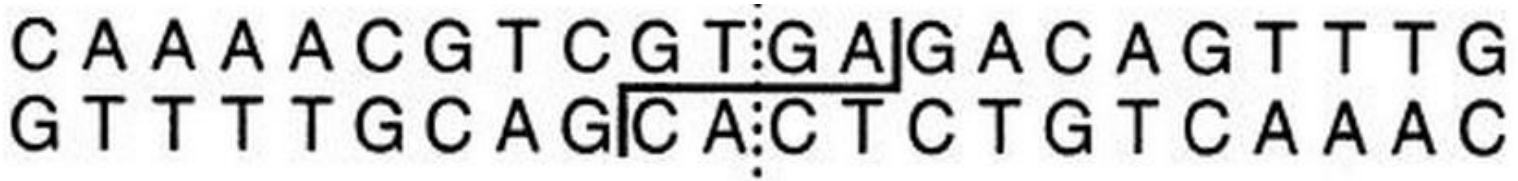
$f_j(b)$ - частота буквы в колонке, p(b) – базовая частота буквы b

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

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

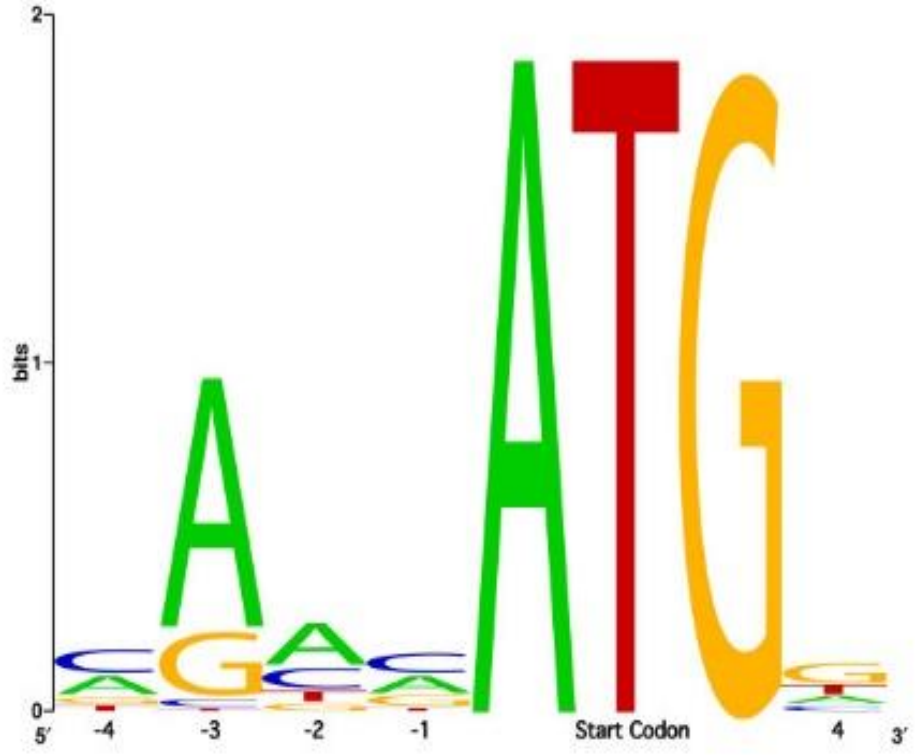
IC слабого и сильного сигналов

- Слабый сигнал:
  - Гомеодомен - консервативный ДНК-узнающий домен многих важных транскрипционных факторов эукариот
  - Узнаёт короткую последовательность ДНК
  - На основании наложения структр гомеодоменов найден единственный общий контакт домена с сайтом ДНК:
    Asn51 имеет две водородных связи с аденином (!)
  - Сигнал NNANN очень слабый , слабее не придумаешь)))
  - по формуле **IC = 2** при базовых частотах ¼
- Сильный сигнал:
  - Эндонуклеаза I-CreI семейства LAGLIDADG узнает такую последовательность. Вероятность обнаружить в геноме такую последовательность случайно близка к 0



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

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



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

Поэтому max(IC(j)) = 2

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

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

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

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

webLOGO

ПРО ЗАДАНИЯ

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

   Поиск по Pubmed.

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

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

ПРО ЗАДАНИЯ

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

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

КОНЕЦ

IV.Сложный пример

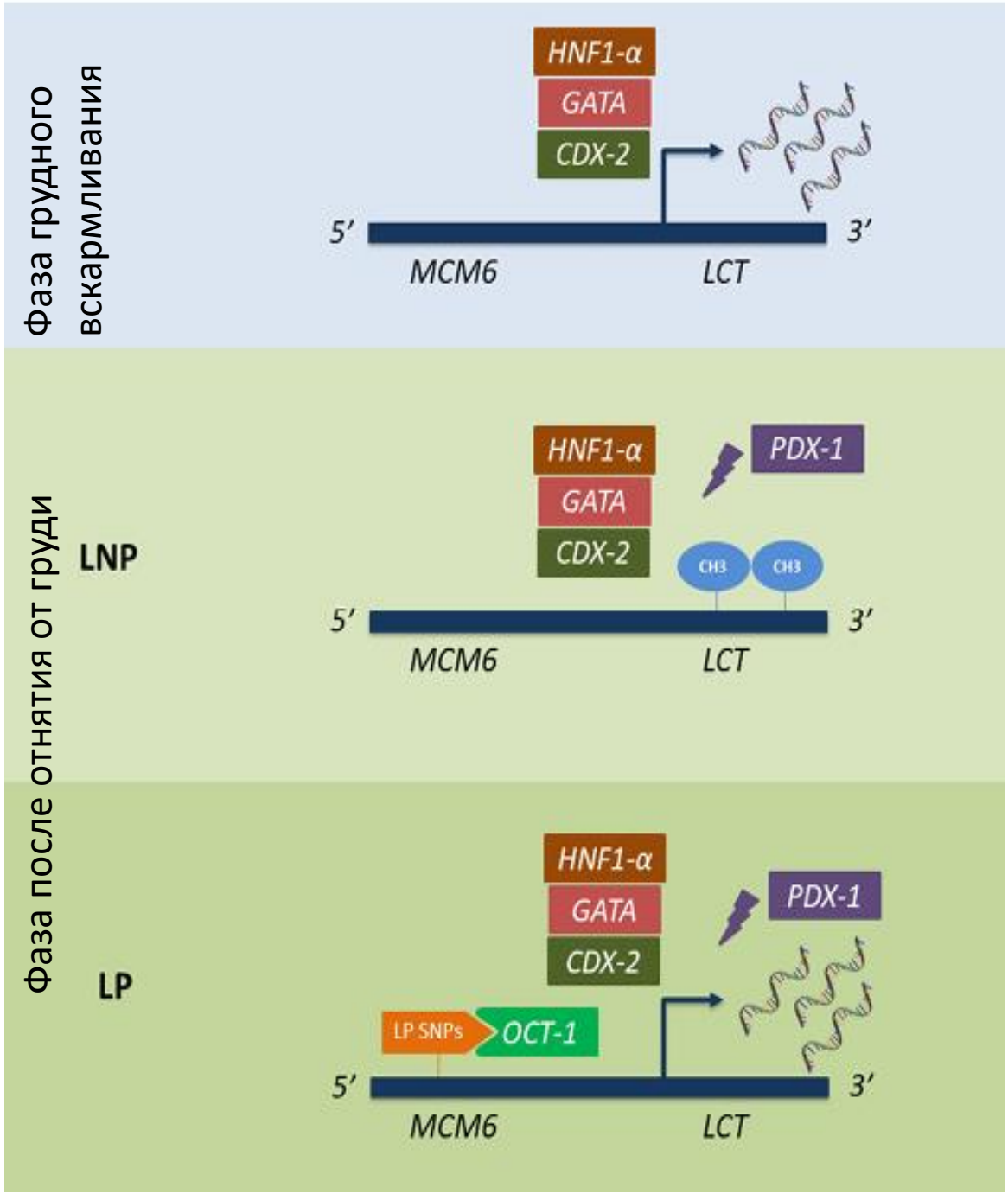
Непереносимость молока (лактозы) во взрослом возрасте (Лактазная недостаточность)

Распространённость лактазной недостаточности по миру по официальной медицинской статистике стран

Чем краснее кружок, тем выше частота непереносимости. Чем синее – тем выше частота переносимости.

Половина человечества имеет лактазную недостаточность

- Лактаза - ген LCT белок LCH_HUMAN
  Лактаза необходима для гидролиза лактозы из молока.
  Лактаза работает у грудничков, и лет до пяти.
  Если не экспресируется у взрослых, то серьёзные проблемы с потреблением молока.
- полифоризмом [2], но статьи с объяснением того через кого сигнал передаётся LCT не обнаружил.

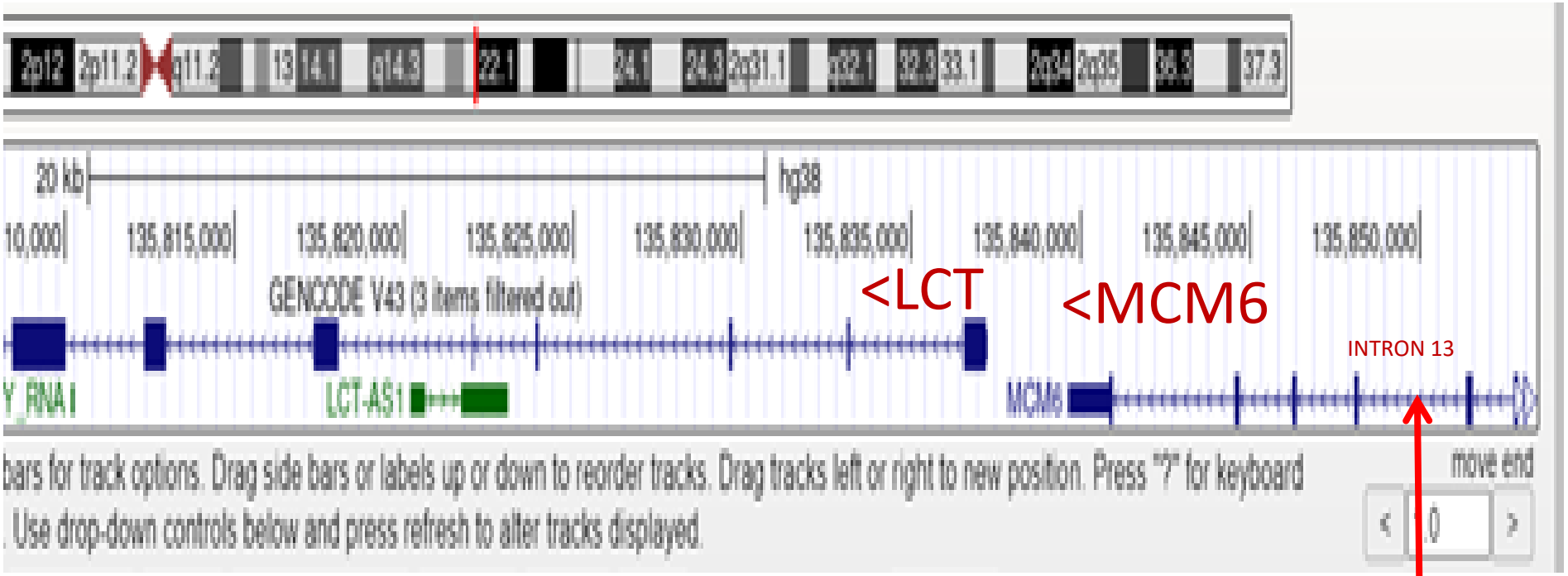


LP – переносимость лактозы
LNP – непереносимость лактозы
ТРАНСКРИПЦИОННЫЕ ФАКТОРЫ
АКТИВАТОРЫ - способствуют экспрессии LCT
HNF1-α
GATA
CDX2
OCT-1
CDX-2: caudal type homeobox 2;
HNF1-α: hepatocyte nuclear factor 1α;
OCT-1: octamer-binding protein 1;
PDX-1: pancreatic and duodenal homeobox 1.
РЕПРЕССОРЫ – предотвращают экспрессию гена PDX-1
Голубые кружочки CH3 – метилированные ди-нуклеотиды CpG

Одно-нуклеотидные полиморфизмы (SNP)
Некоторые из полиморфизмов на оранжевом участке способствуют LP потому, что образуется сайт связывания активатора OCT-1.

Нетривиальность в том, что этот сайт находится за 14 тыс. п.н. до LCT.
Такое называется ЭНХАНСЕРОМ. Влияние ТФ на экспрессию связано с образованием петли ДНК

Anguita-Ruiz A et al., Genetics of Lactose Intolerance: An Updated Review and Online Interactive World Maps of Phenotype and Genotype Frequencies. Nutrients. 2020

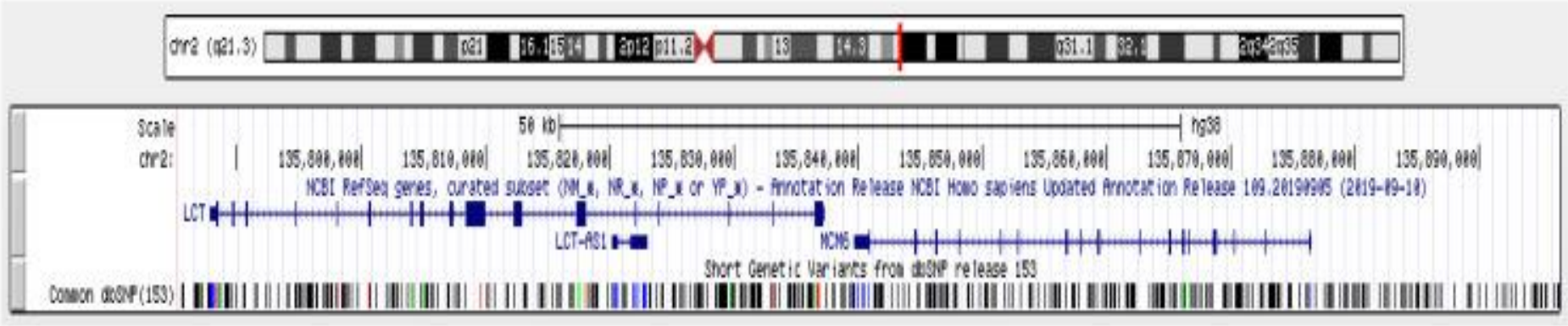


Сайленсеры и энхансеры – сайты связывания факторов транскрипции, удалённые на значительное расстояние от старта транскрипции

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)



[1] Anguita-Ruiz A, Aguilera CM, Gil Á. Genetics of Lactose Intolerance: An Updated Review and Online Interactive World Maps of Phenotype and Genotype Frequencies. Nutrients. 2020

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