1. Сигналы и мотивы

Поиск сигналов в последовательностях

План

- Геном и информация.
  - Способы кодирования
  - Способы считывания сигнала в ДНК
- Примеры сигналов
- Способы перекодировки сигналов в ДНК для людей – мотивы
  - Последовательность
  - Паттерн
  - PWM – позиционная весовая матрица
- Примеры без подробностей
- Информационное содержание сигнала

Геном и Информация

- Носитель генома – совокупность ДНК клетки.
  У вирусов, хотя вирусы – не клетки, тоже есть геном, ДНК или РНК
- В геноме закодирована информация.
- Что такое информация? Нашел такое определение:
  ИНФОРМАЦИЯ — сведения независимо от формы их представления))) [1)]
- Эту информацию в перекодированном виде и изучает биоИНФОРМАТИКА.
  Значит, для биоинформатики требуется перекодировка молекулярно-биологической информации – какой и как?
- ……

Теория информации основанная Шенноном – математическая теория передачи данных[2)] – используется в биоинформатике, но слишком формализована и проста для объяснения живого:)

[1)] Wiki со ссылкой на Когаловского Р.М. специалиста по информационным систем.
[2)] C.Shannon, “The Mathematical Theory of Communications” , 1948

Геном и информация

- Так какая информация закодирована в геноме?
  - Гены белков
  - Гены РНК
  - Что ещё ……
- Кто, как и зачем в клетке использует информацию из ДНК?
- Это и составляет загадку жизни:)

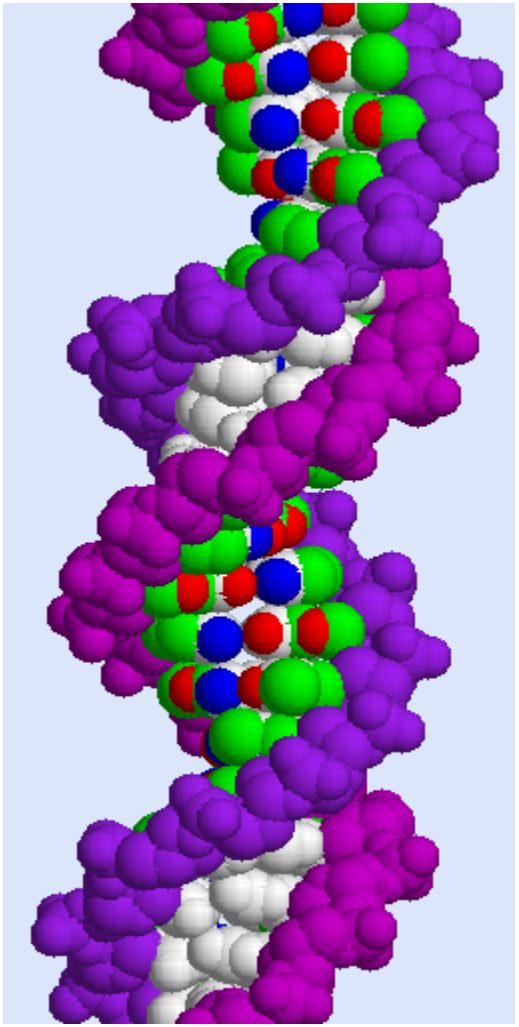
Способы кодирования сигнала в геноме

- Закодированы
  - последовательностью нуклеотидов,
  - для белков – последовательностью аминокислотных остатков
- Закодированы структурой НК или белка
  - G-квадруплекс
  - IRES – особым образом уложенная шпилька РНК
- Закодированы химической модификацией НК или белка
  - Cap
  - Метилирование ДНК

В ДНК закодированы сигналы, необходимые для управления клеточными механизмами

- Сигналы закодированные последовательностью нуклеотидов
  - «Чтение» последовательности с помощью комплементарности РНК-ДНК
  - «Чтение» ДНК без её расплетения
- Сигналы, закодированные химическими модификациями ДНК или (у эукариот) гистонов.
- Сигналы, закодированные вторичной или третичной структурой ДНК, РНК, белков

Так белки читают последовательность ДНК не расплетая её



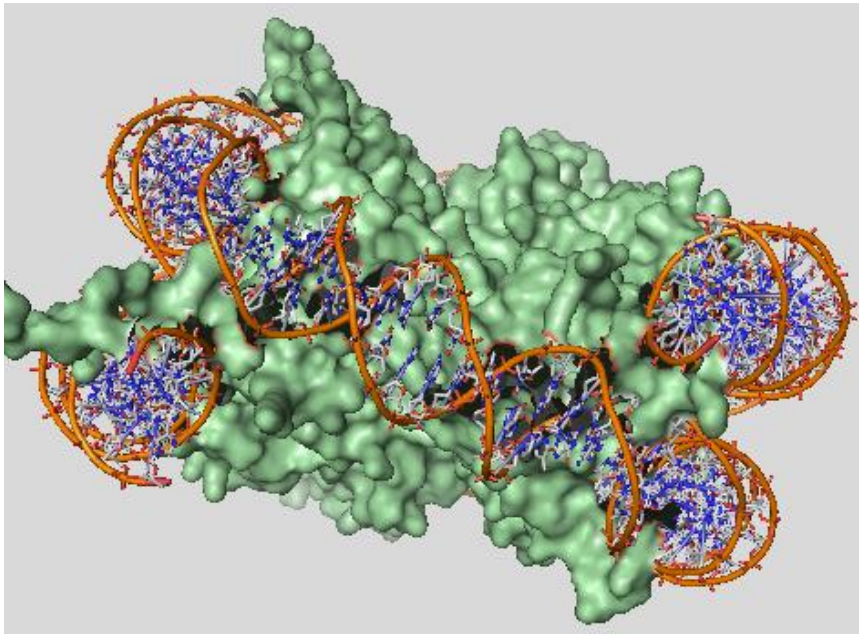
Двойная спираль ДНК.

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

Для эукариот дело 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.

Пример 1й

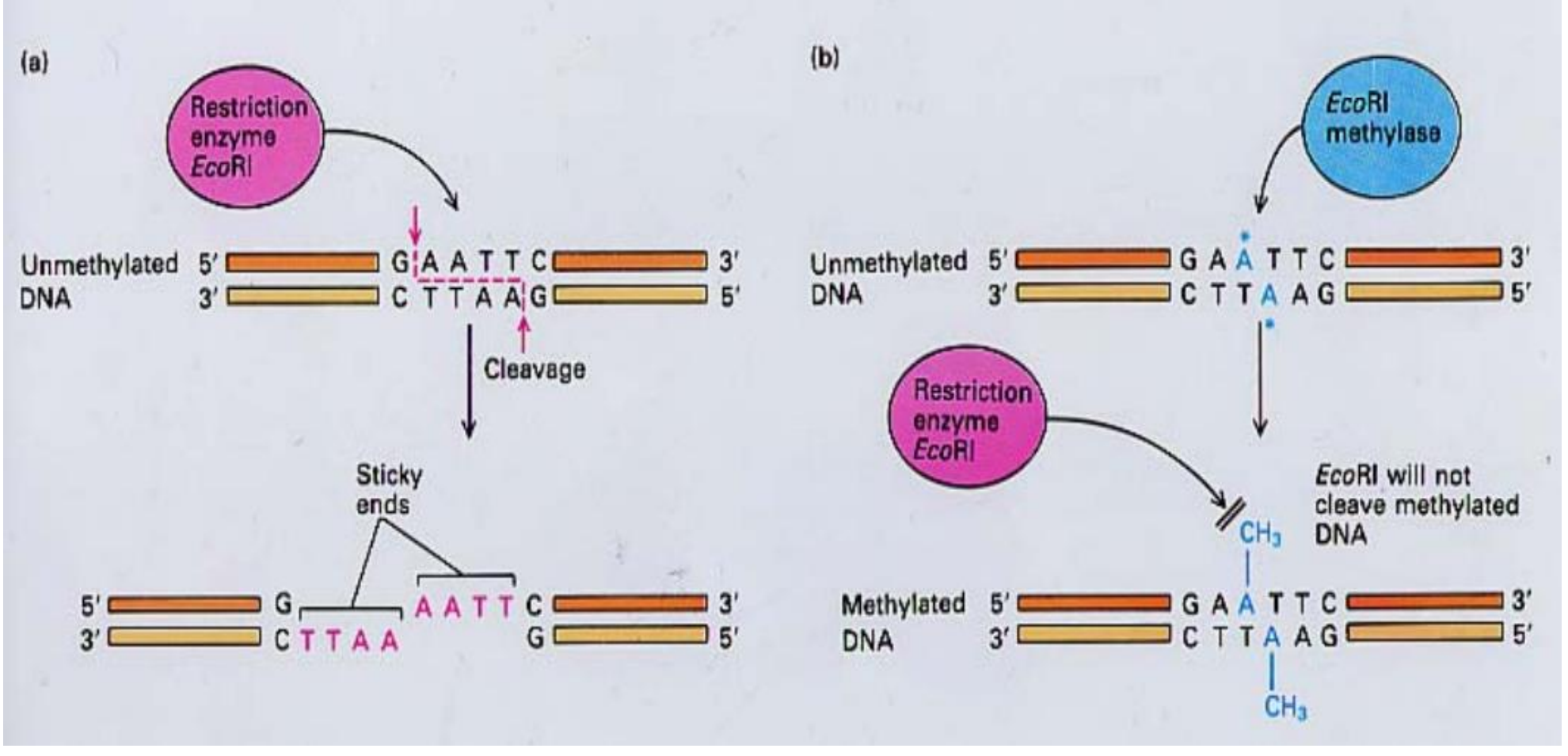
Сигнал «рестрикции» - расщепления ДНК эндонуклеазой рестрикции у прокариот

Сигнал: **GAATTC** в геноме E.coli

Адресован системе рестрикции-модификации EcoRI (белки R и M)

Три состояния:
- не метилирован
- Полуметилирован – метилирован по одной цепочке
- метилирован но двум цепочкам

Предназначен для отличения и расщепления чужой ДНК, и не расщепления своей



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

Для людей всё однозначно. Найти встречи этого сигнал в нуклеотидной последовательности (геноме) легко, используя программу _____ из пакета EMBOSS

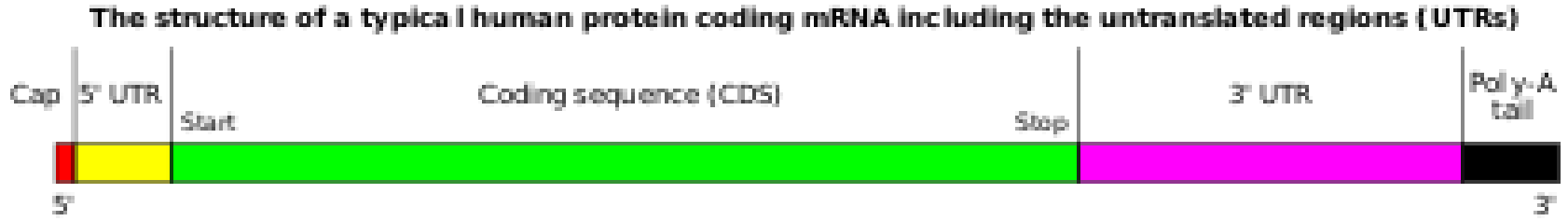
В терминах теории информации для человека **Информационное содержание** (IC) этого сигнала – максимально возможное: есть сигнал ⇔ есть ответ, ДНК будет расщеплена в этом месте эндонуклеазой EcoRI
Энтропия (H), т.е. степень неопределенности сигнала, - нулевая

Для эндонуклеазы – так же, или почти так же, если она иногда ошибается – расщепляет не этот, а похожий сайт (не знаю, но исключить без результатов экспериментов не могу)

Пример 2й

Старт трансляции у эукариот
Он же – задание 1

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

Проблема: старт трансляции со второго ATG кодона

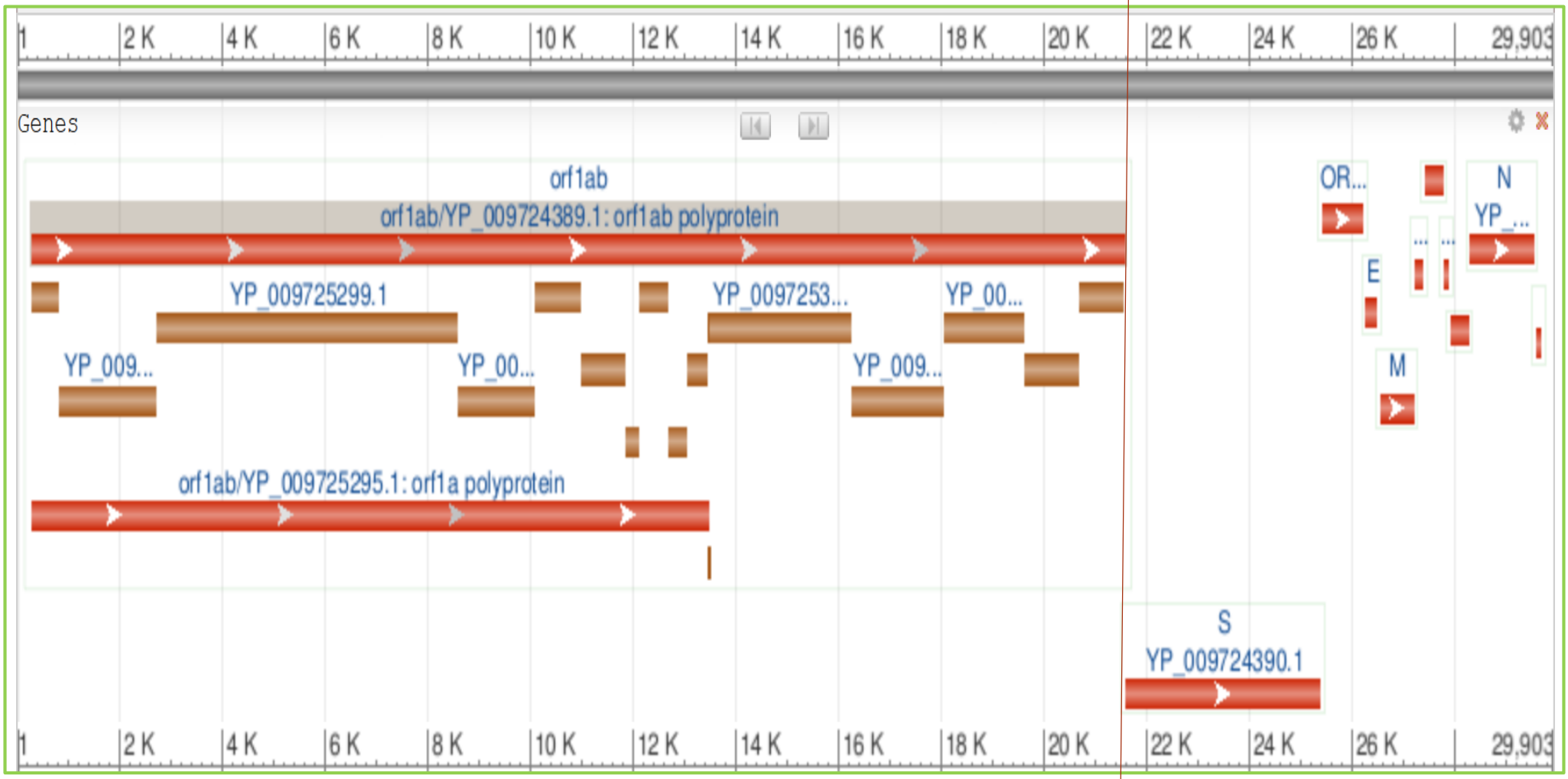
- Первый ген CoV orf1ab начинается с 266 пн (самая длинная красная полоска)
- У SARS-CoV-2 такие ATG до 269-й пн.:
    - 107 – ATG
    - 266 – ATG
- Просто ATG недостаточно для старта трансляции?
- М.Ко́зак в 1986 году проанализировала известные инициаторные кодоны ATG и нашла более длинный *слабый* сигнал
- Сигнал начала трансляции (у эукариот) называется последовательностью Козак. В разных таксонах - отличия

ГЕНЫ БЕЛКОВ SARS-CoV-2

красные и коричневые полоски

По оси X нуклеотиды РНК

1 10 000 20 000 29 903



Вопросы есть?

Кэп-зависимая инициация трансляции

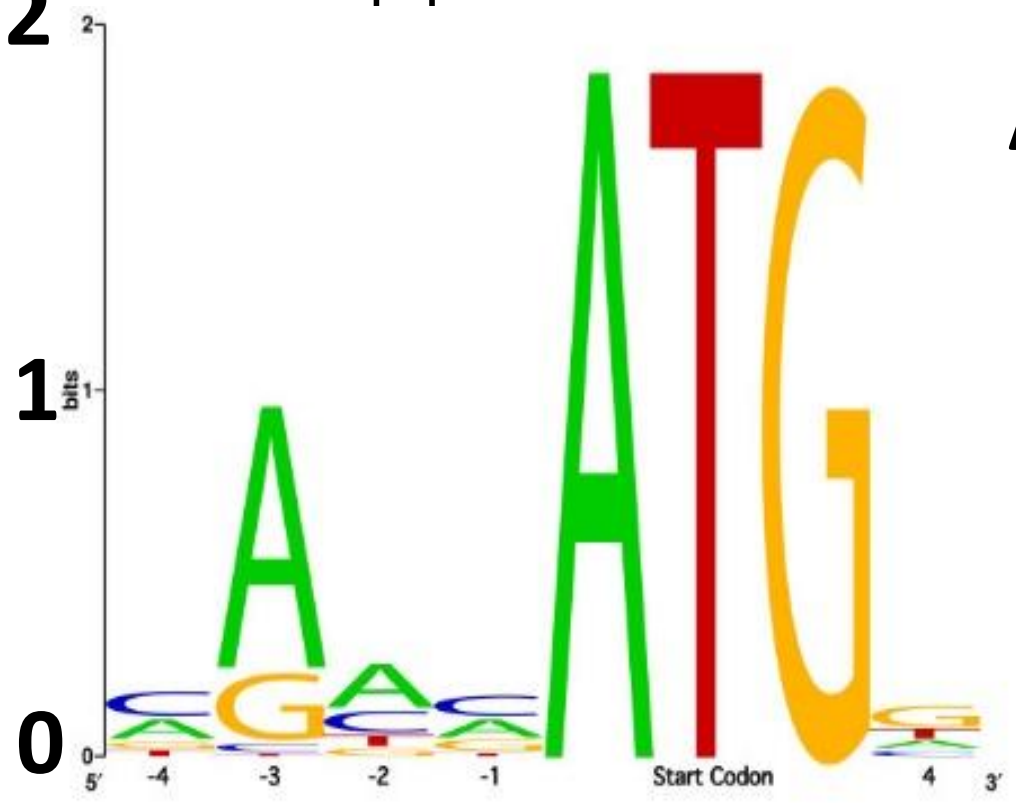
При сканирующем механизме малая субъединица рибосомы садится на 5'-конец мРНК в области кэпа и двигается вдоль молекулы мРНК, «сканирует» кодоны в поисках инициаторного AUG.

- 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)

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

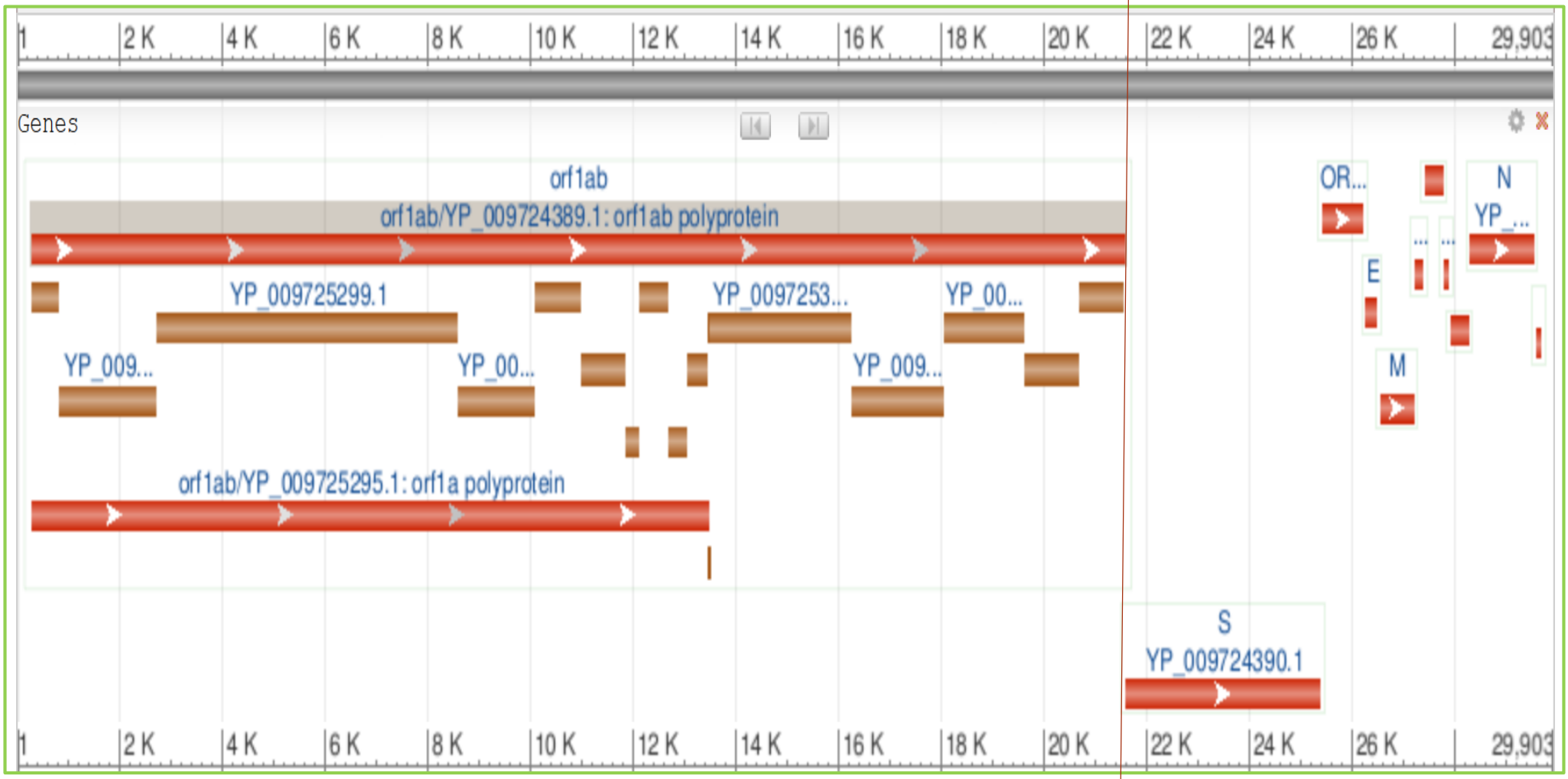
Kozak M, Cell, 1986.

ГЕНЫ БЕЛКОВ SARS-CoV-2

красные и коричневые полоски

По оси X нуклеотиды РНК

1 10 000 20 000 29 903



Вопросы есть?

РНК коронавируса содержит оба сигнала

- ПолиА на 3’-конце [сигнал в последовательности]
- КЭП – 7-метилгуанозин - на 5’ конце [химический сигнал]

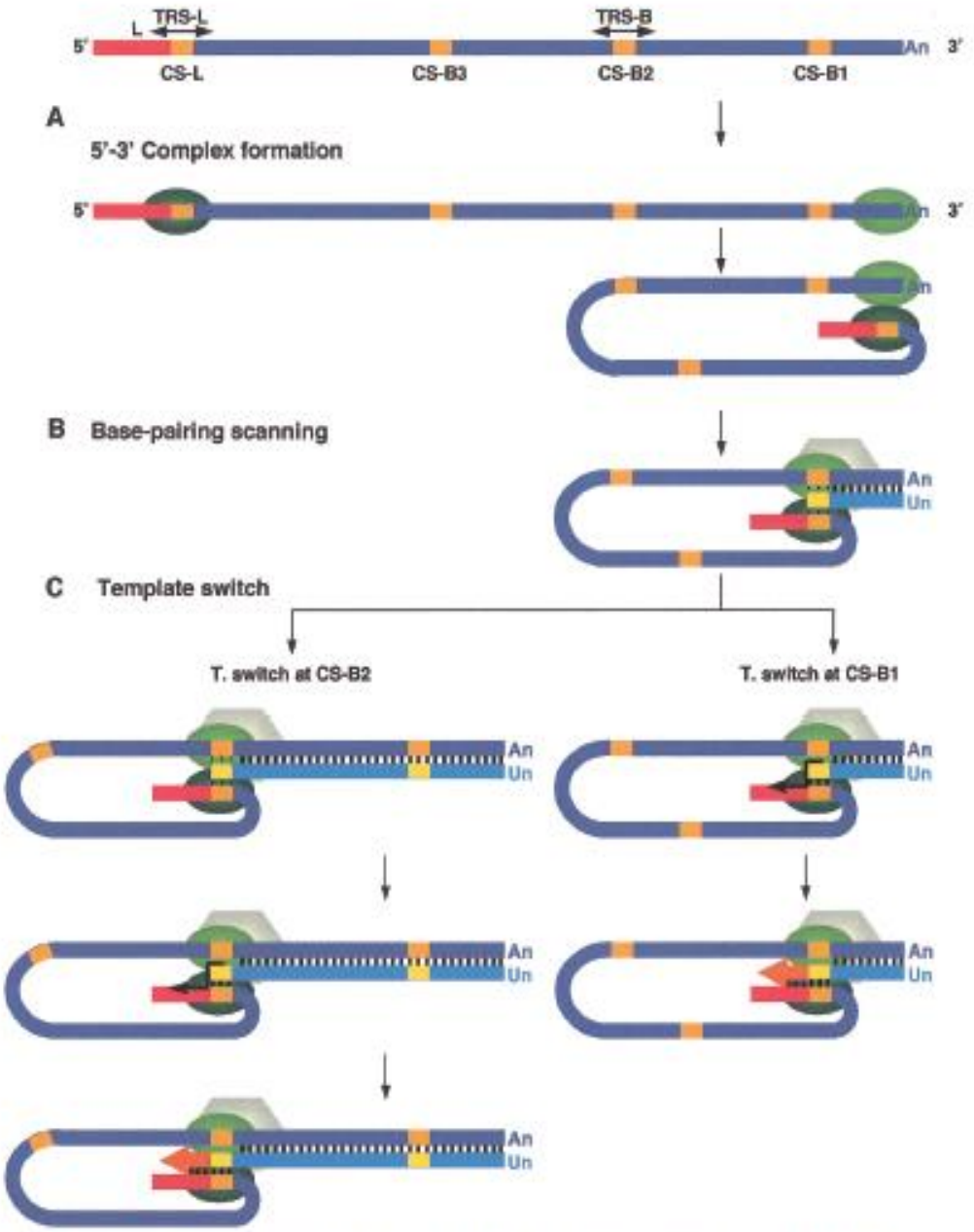
……AAAATTAATTTTAGTAGTGCTATCCCC
ATGTGATTTTAATAGCTTCTTAGGAGAAT
GAC**AAAAAAAAAAAAAAAAAAAAAAAAAAAA**
AAAAAAA – 29903

У человека одна мРНК – один белок!

Транскрипция вирусной РНК

- Вирусная РНК-зависимая РНК-полимераза (RdRP) закодирована в полипротеине (nsp11).
- RdRP по РНК матрице делает комплементарную копию. Из вирусной +RNA получается -RNA; из -RNA получается -(-RNA) = +RNA
- Сигналы разрывной транскрипции направляют перескок RdRP при синтезе -RNA, в результате которого синтезируются -sgRNA.
- -sgRNA является матрицей для RdRP; продукт – субгеномная мРНК (+sgRNA)
- Сигналы разрывной транскрипции называются так: TRS-L в лидере, TRS-B перед каждым поздним геном (TRS=transcription-regulatory sequences)

TRS-L и TRS-B



Лидер –красная полоска

Сигналы TRS – желтые прямоугольники. В них есть общее слово из шести букв (CS)

Мутации в CS влияют на синтез sgmRNA ожидаемым образом

Рисунок - гипотеза, косвенно подтвержденная

Zuniga et al., Journal of Virology, 2004

Сигналы разрывной транскрипции TRS-L, TRS-B; CS

Сигналы TRS-L и все TRS-B имеют высокосходные последовательности. Наиболее похожие их части, часто полностью совпадающие, называются CS (core sequences)

Принято считать, что длина CS – шесть нуклеотидов, TRS включает2-3 нуклеотида с 5’ и 3’ концов CS.

Как все в биологии значения длин не являются мировыми константами

Пример 3й

Сигнал посадки рибосомы у прокариот – последовательность Shine-Dalgarno (SD)

Одно из заданий (по выбору) занятия 7

В геноме одной археи или бактерии найти сигнал сайта посадки рибосомы (SD)

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

Saito et al., 2020, eLife

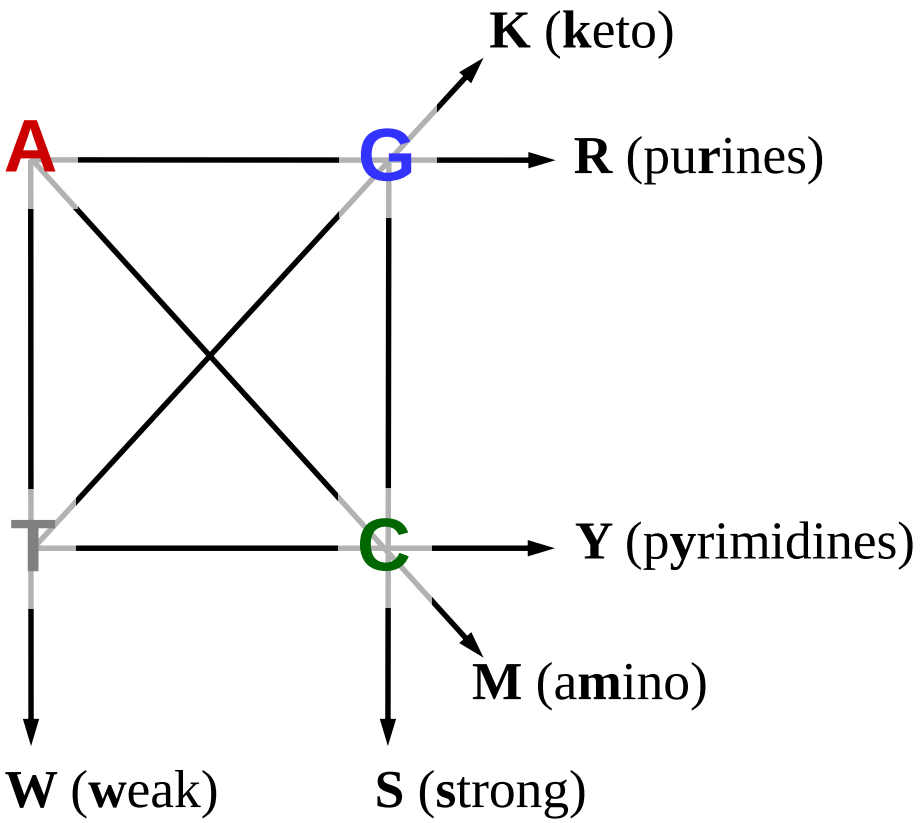
Способы описания сигнала

в последовательности

Мотив – описание последовательностей одного сигнала

- Точная последовательность
  - AAAAAAAAAAAAAAAAAAAAA (от десятков до сотен букв)
  - CpG
  - GATC
- Паттерн
  - CCWGG, и др. примеры из миниКР
- 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

Позиционная весовая матрица (PWM)

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

(Задание 2 этого практикума)

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
```

Идея: вес буквы зависит от позиции в выравнивании

Новая .... `CCTAACCATTATTTTT`...

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
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.
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.
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.
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	-inf	-0.9	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-0.3	-inf	-0.

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 $\varepsilon(b) = \varepsilon/4$

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	0.100	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	0.100	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	0.100	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	0.100	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

W(b,j)	баз. Частоты	e(b)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
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	-(
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	(
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	(
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	-(
	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	-(
			G	**C**	**C**	**T**	**A**	**C**	**C**	**C**	**C**	**A**	**T**	**T**	**A**	**T**	

Выравнивание сайтов связывания PurR *E. coli*

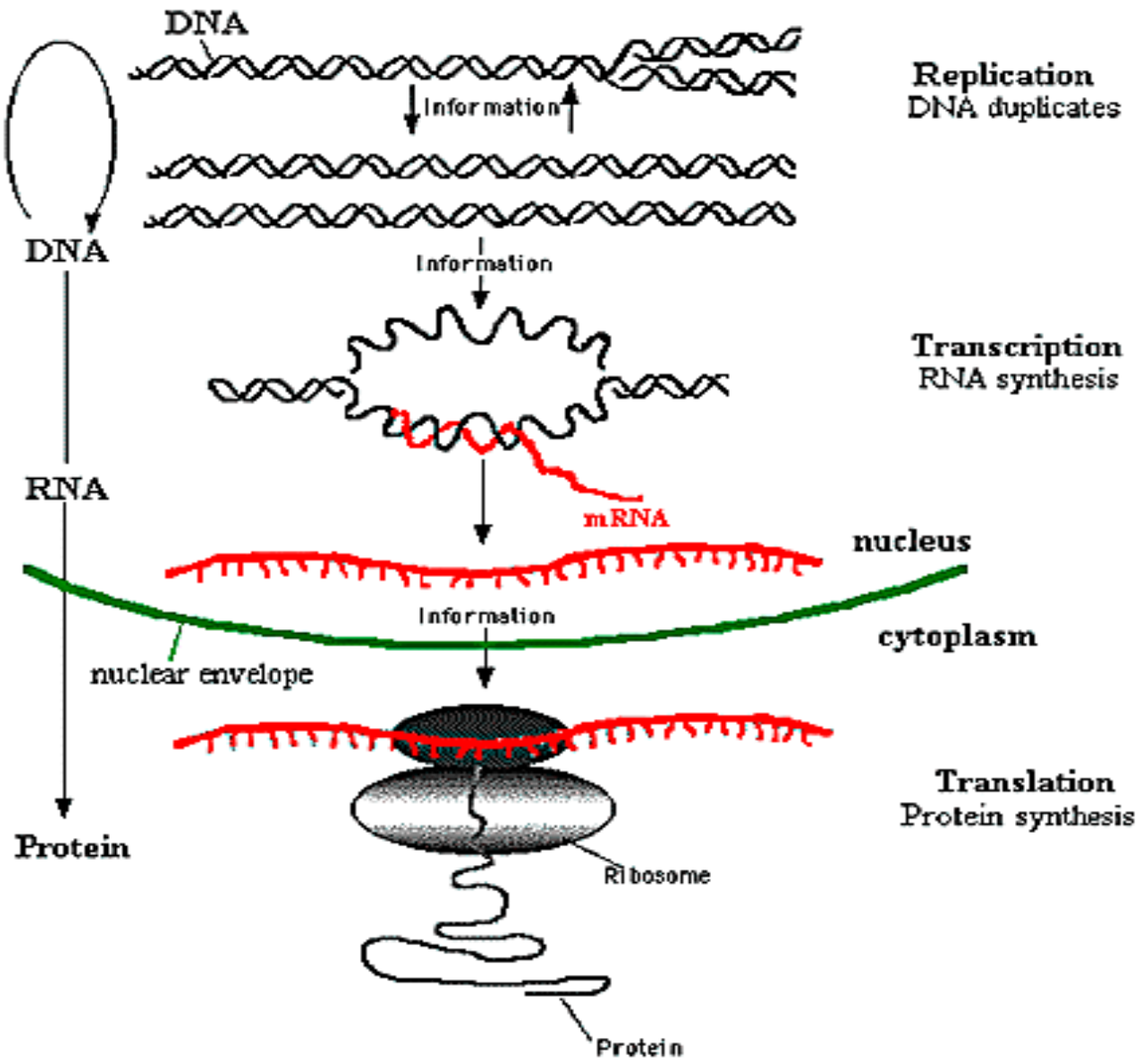
cvpA	**CCTACGCAAACGTTTTCTTTTT**
purM	**GTCTCGCAAACGTTTGCTTTCC**
purT	**CACACGCAAACGTTTTCGTTTA**
purL	**TCCACGCAAACGGTTTCGTCAG**
purE	**GCCACGCAACCGTTTTCCTTGC**
purC	**GATACGCAAACGTGTGCGTCTG**
purB	**CCGACGCAATCGGTTACCTTGA**
purH	**GTTGCGCAAACGTTTTCGTTAC**
$purA_1$	**TTGAGGAAAACGATTGGCTGAA**
$purA_2$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

Источник: РГМ

III. Разнообразные сигналы

Быстро!

В клетке вот что происходит



The Central Dogma of Molecular Biology

Designer Chang

Для человека важно назвать объекты

- Репликация
  Репликацию ДНК осуществляет сложный комплекс, состоящий из 15—20 различных белков-ферментов, называемый реплисомой (wiki)
  - Где начать
    Место начала репликации (англ. origin of replication)
    - *Сайт связывания белка DnaA*
    - *Область первичного раскручивания спирали ДНК*
    - *Сайты метилирования (GATC dam MTase)*
      [wiki на примере E.coli]
  - когда начинать (?)
- Транскрипция
  ДНК зависимая РНК полимераза, комплекс белков.
  - Инициация
    место начала – промотор
  - Терминация
    прокариоты – Rho-зависимая и Rho – независимая
    эукариоты у мРНК сигнал полиаденилирования (поли-А и сигнал кэппирование мРНК)
- (eu) Сплайсинг
- Трансляция мРНК
  - место начала (инициация)
  - место окончания
  - Программируемый сдвиг рамки считывания

Для человека важно назвать объекты

- Трансляция мРНК
  - место начала (инициация)

  эукариоты - ATG в хорошем контексте (последовательность Ко́зак)
  прокариоты – последовательность Шайн-Далгарно
  - место окончания

- Трансляция
- *Рибосома.*
  - Инициация
    место начала – промотор
  - Терминация
    прокариоты – Rho-зависимая и Rho – независимая
    эукариоты у мРНК сигнал полиаденилирования (поли-А и сигнал кэппирование мРНК)
- Регуляция

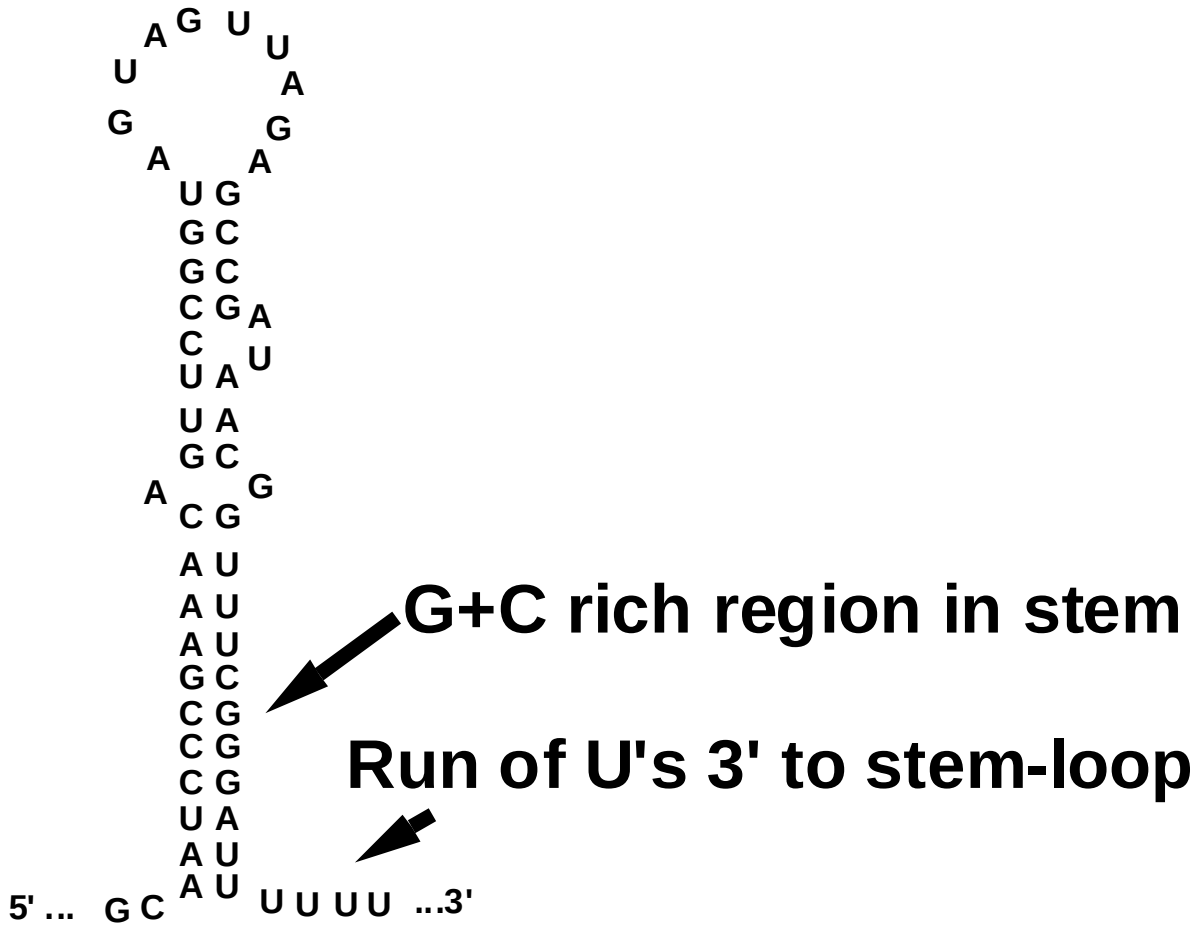
Сигналы, закодированные последовательностью НК или белка

Мотивы в последовательности (что может понять биоинформатик или мол. биолог)

Примеры

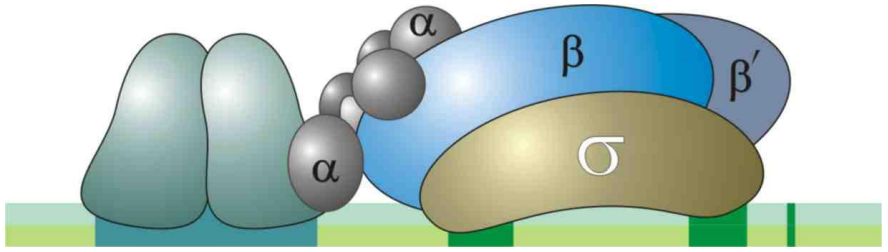
- ПолиА в мРНК, но в ДНК гена белка это:
  - Сигнал полиаденилирования (AAUAAA на 3’ в мРНК)
    - Его читает …. => полиаденилат-полимеразой
    - Важность этой последовательности можно увидеть на примере мутации в гене человеческого 2-глобина, которая изменяет AAUAAA на AAUAAG, что приводит к недостаточному количеству глобина в организме
- Промотор и старт транскрипции
- Конец транскрипта
  - Эукариоты – тот же сигнал полиаденилирования
  - Прокариоты:
    - Rho independent
    - Rho dependent
    - ??????? (Ju, Li, Lui, Nat Microbiol. 2019; о результатах Send-seq технологии)

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

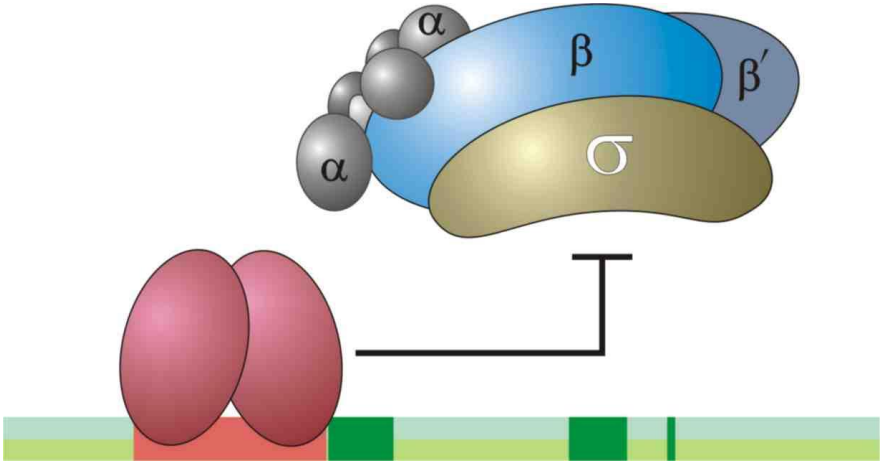


Транскрипция в прокариотах:
Регуляция транскрипции

Активация

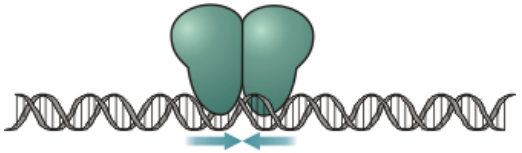
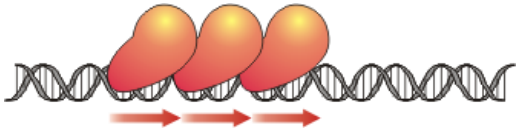
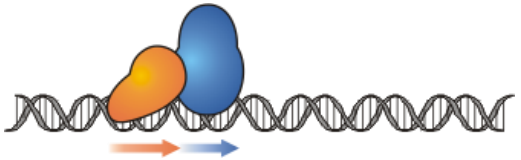


Репрессия



Источник: РГМ

Использование свойств сигнала

- ДНК-связывающие белки и их сигналы
  - Кооперативные однородные
    - Палиндромы 
    - Прямые повторы 
  - Кооперативные неоднородные
    - Кассеты 
  - Другие
- РНК-сигналы

Источник: РГМ

I. Информационное содержание выравнивания последовательностей сигнала. LOGO. «Сила» сигнала

«Классное» задание в конце занятия - вычислить информационное содержание выравнивания

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

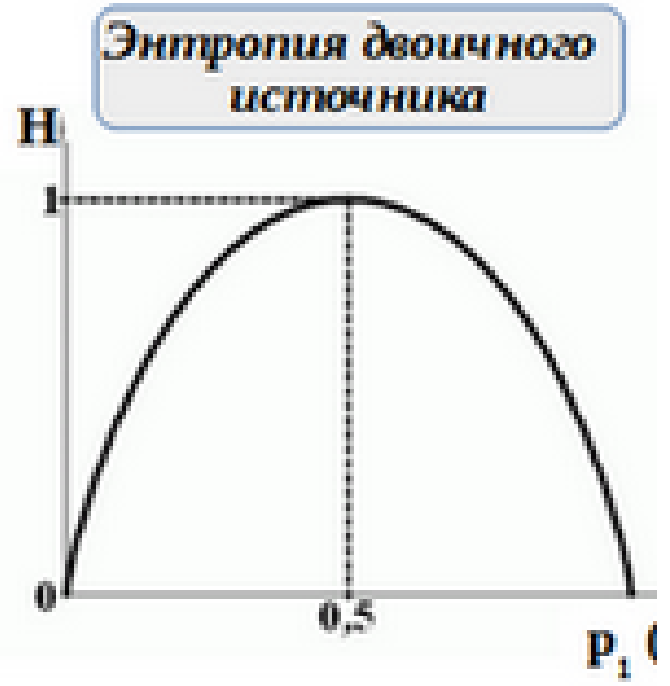
- Изучаем сигнал, который есть последовательность букв. В нашем случае – задан выравниванием представителей сигнала.
- Энтропия H – мера неопределенности, т.е. невозможности предсказать сигнал. Аксиомы:
  - H положительна
  - H = 0 если сигнал однозначно предсказуем (……)
  - Чем менее предсказуем сигнал, тем больше энтропия сигнала (звонок с неизвестного номера). Максимум достигается когда все слова, составляющие сигнал, равновероятны.
  - H аддитивна: энтропия сигнала из одной буквы равна сумме энтропий каждой из них ; энтропия сигнала состоящего из нескольких независимых сигналов 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 равновероятны, то H = 2
- H(сигнала длины N) = N•H в силу аддитивности

Сигнал двоичный:
комплементарная пара W=(A или T) или
S = (G или C).
p = GC состав в долях единицы
(1-p) = частота пары A-T

График $H(p) = p \log_2 p + (1-p) \log_2 (1-p)$
зависимости энтропии от GC состава



I. Информационное содержание выравнивания последовательностей сигнала. LOGO. «Сила» сигнала

«Классное» задание в конце занятия - вычислить информационное содержание выравнивания

КОНЕЦ

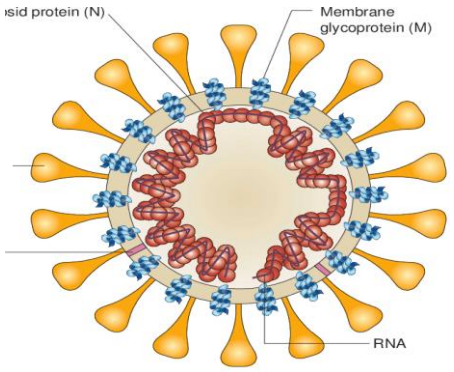
Официальных слайдов

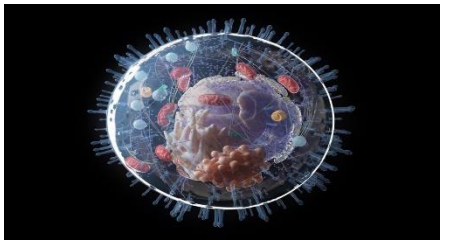
КОНЕЦ ПРЕЗЕНТАЦИИ

I. Сигналы у коронавируса SARS-CoV-2

Вирус для размножения использует механизмы хозяйской клетки; значить, вынужден использовать сигналы, понимаемые клеткой

ОТ КОГО





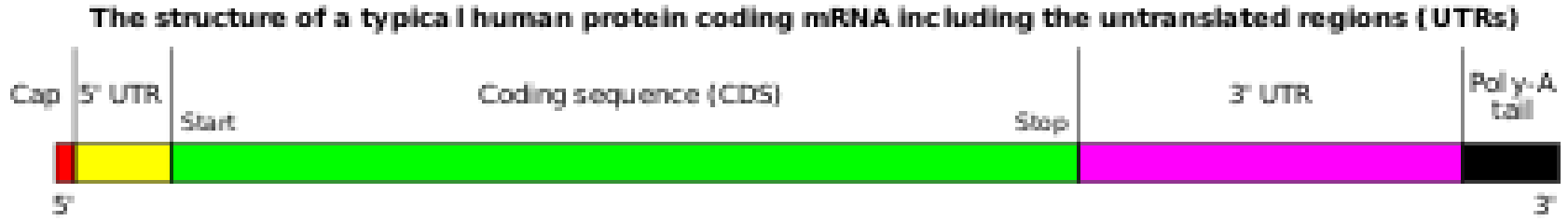
КОМУ

+РНК коронавируса (29 903 пн)

“+” значит, что она кодирует белки вируса

Сигнал рибосомам «Я мРНК»

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

РНК коронавируса содержит оба сигнала

- ПолиА на 3’-конце [сигнал в последовательности]
- КЭП – 7-метилгуанозин - на 5’ конце [химический сигнал]

……AAAATTAATTTTAGTAGTGCTATCCCC
ATGTGATTTTAATAGCTTCTTAGGAGAAT
GAC**AAAAAAAAAAAAAAAAAAAAAAAAAA**
AAAAAAA – 29903

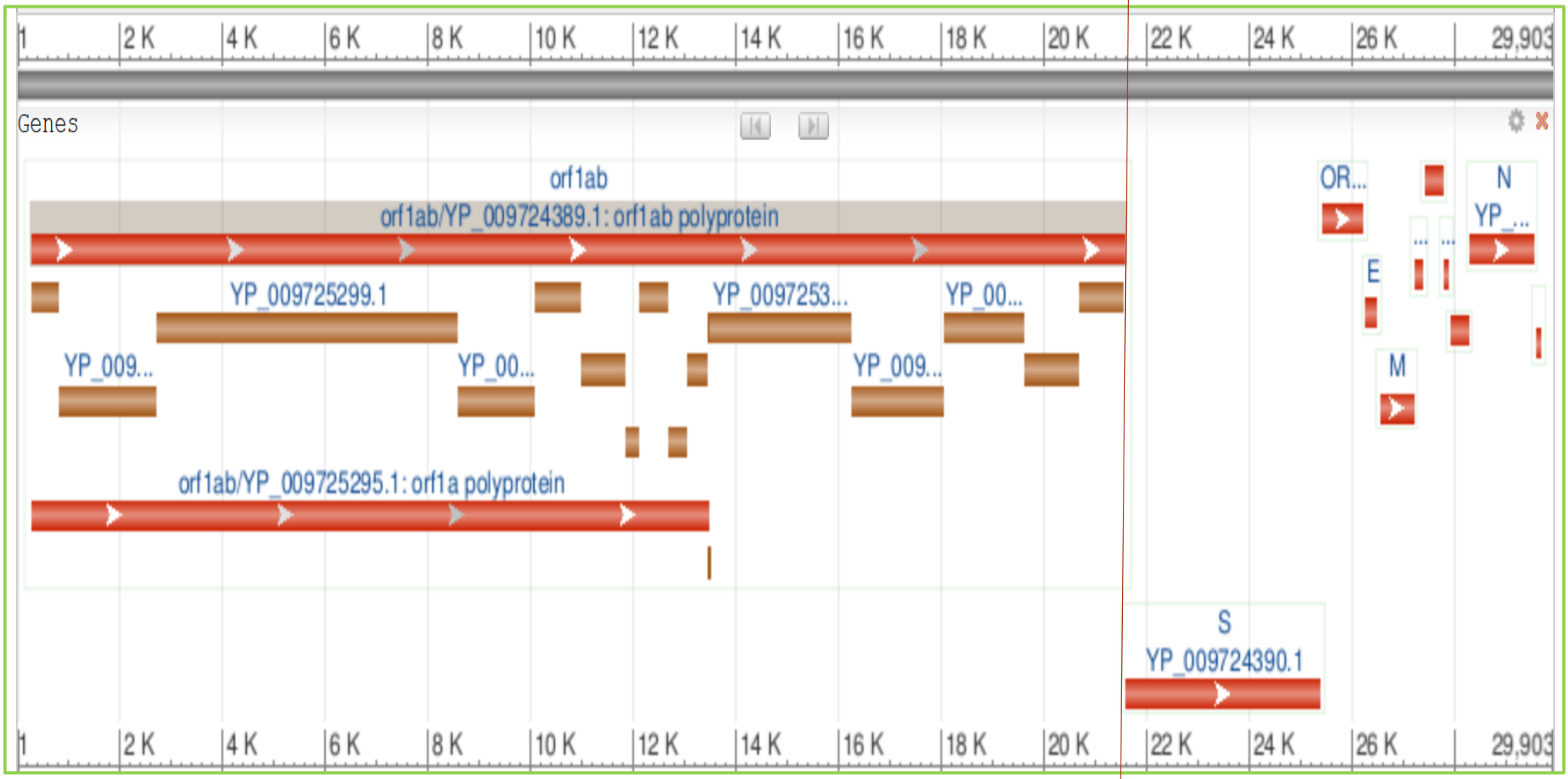
У человека одна мРНК – один белок!

ГЕНЫ БЕЛКОВ SARS-CoV-2

красные и коричневые полоски

По оси X нуклеотиды РНК

1 10 000 20 000 29 903

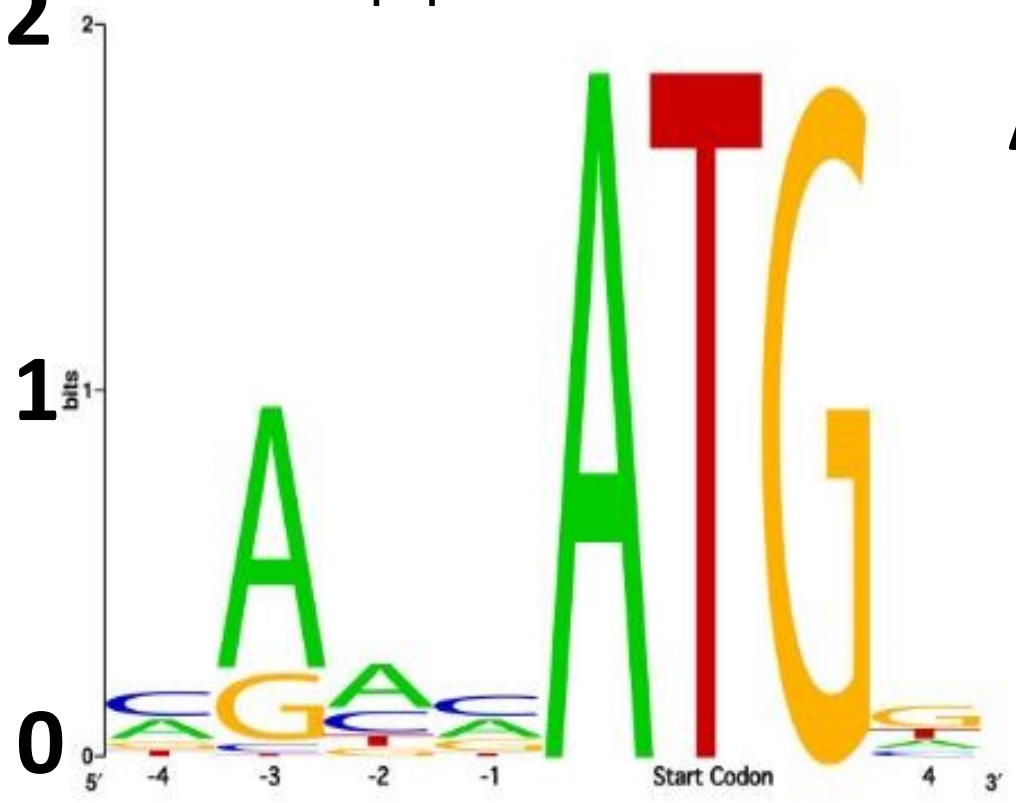


Вопросы есть?

Проблема 1: старт трансляции со второго ATG кодона

- Первый ген CoV orf1ab начинается с 266 пн (самая длинная красная полоска)
- У SARS-CoV-2 такие ATG до 269-й пн.:
  - 107 – ATG
  - 266 – ATG
- Просто ATG недостаточно для старта трансляции?
- 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

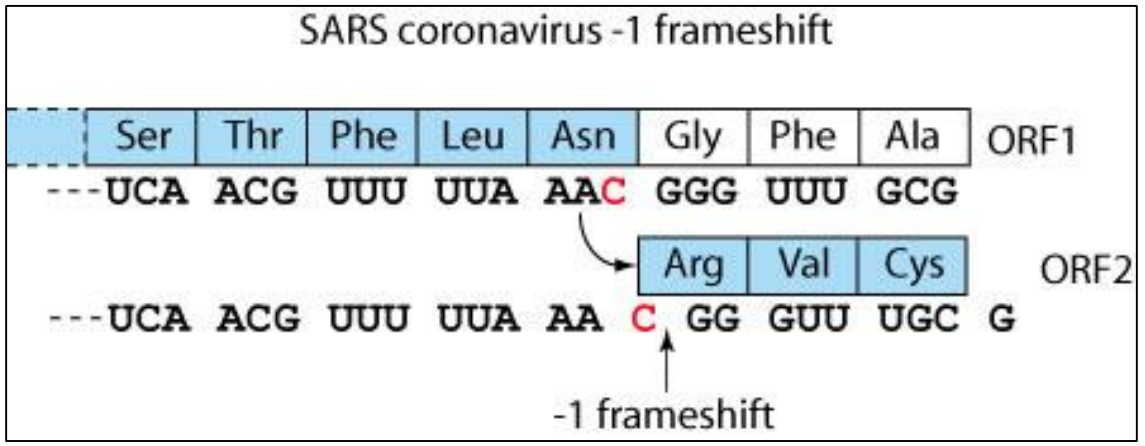
Kozak M, Cell, 1986.

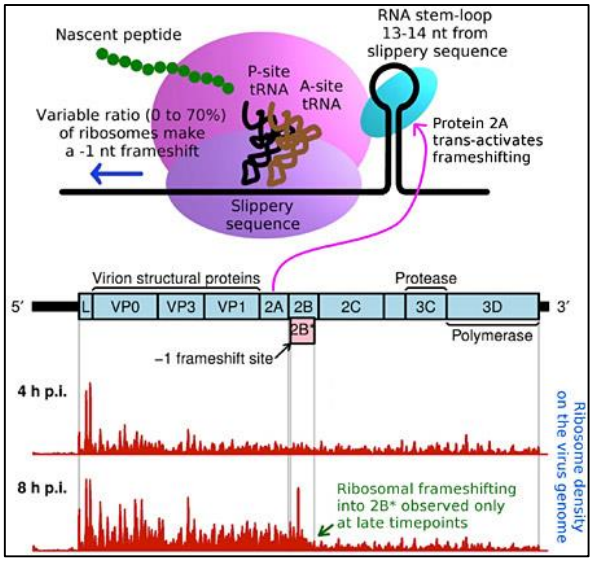
Проблема 2: два гена с одинаковым стартом, но разной длины

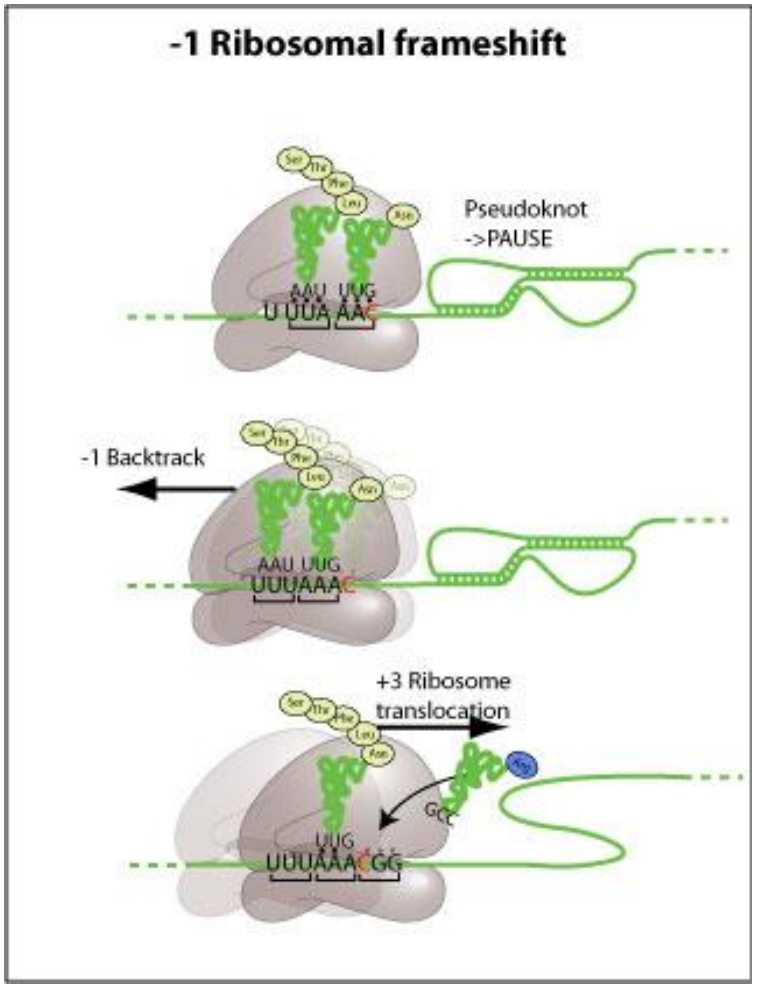
- Короткий ген orf1a завершает стоп кодон
- Получается, что иногда рибосома проскакивает канонический стоп-кодон
- Сигнал программируемого сдвига рамки считывания адресован рибосомам хозяйской клетки

Программируемый рибосомный сдвиг.

Рибосома останавливается из-за шпильки на РНК
и slippery sequence. Отскакивает на ОДИН нуклеотид(букву). И продолжает синтез белка







Ketteler, 2012, Front Genet. 2012; 3: 242.

-1 frameshift

The slippery sequence
is generally of the type X XXY YYZ, where X denotes any nucleotide, Y denotes A or U, and Z is A, U, or C.

Pseudoknots are secondary RNA substructures that contain two or more stem-loop motifs with intercalated stems.

The pseudoknot or stem-loop structure in the mRNA is thought to result in pausing of the ribosome, resulting in eventual frameshifting (Namy et al., 2006)

Сигнал программированного сдвига рамки

- Состоит из двух частей
  - Последовательность
  - Шпилька (псевдоузел) РНК
- Не является сильным, иногда работает иногда нет

Проблема 3. Из одного белка получается много зрелых белков

- Сигналы протеолиза полипротеина, вирусными протеазами
- Сигнал - короткая последовательность аминокислот и их конформация в 3D
- Протеазы – домены полипротеина.

Продукты генов ORF1ab и ORF1a полипротеины

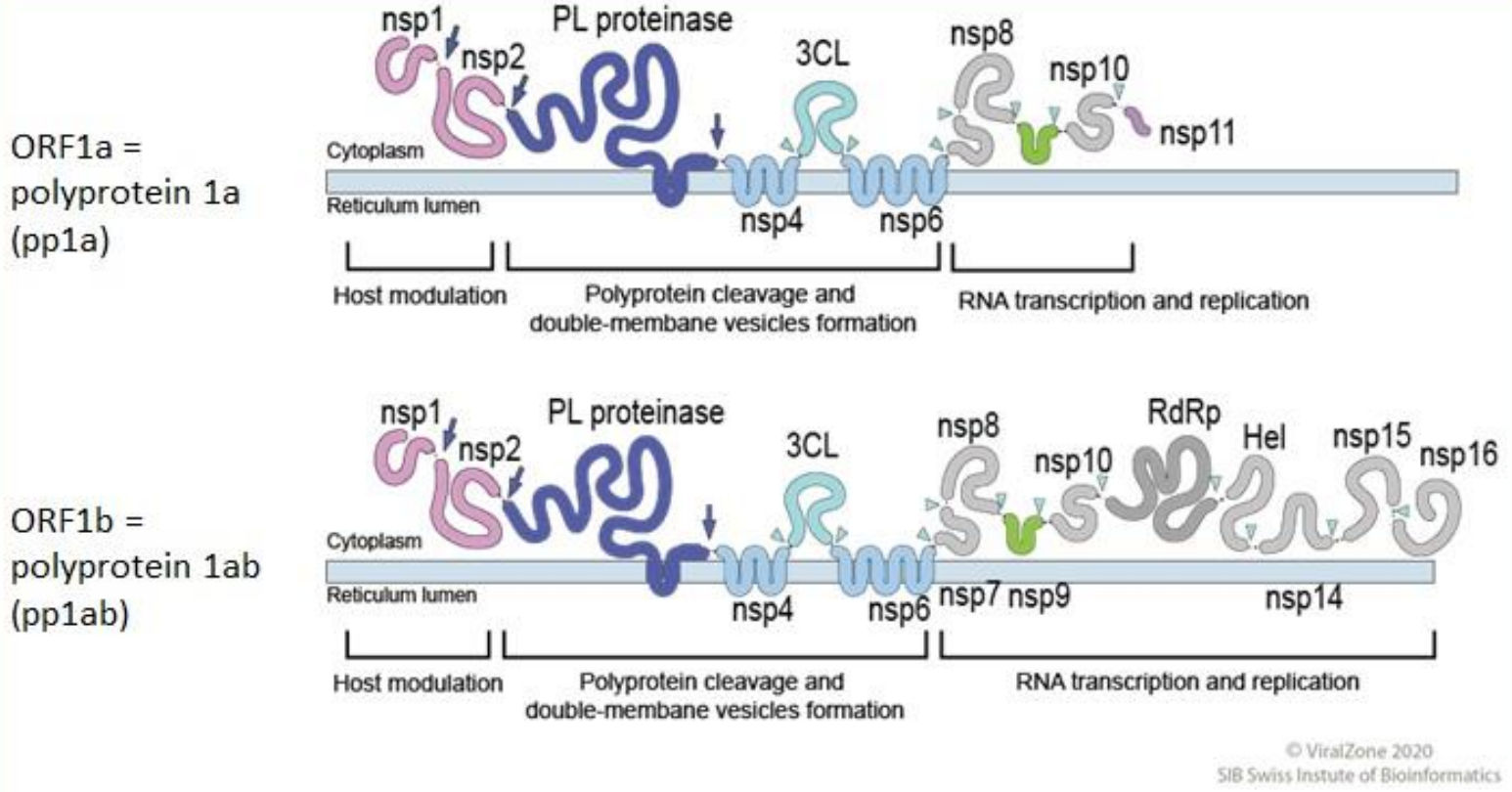


Figure 2: SARS-Cov-2 polyproteins

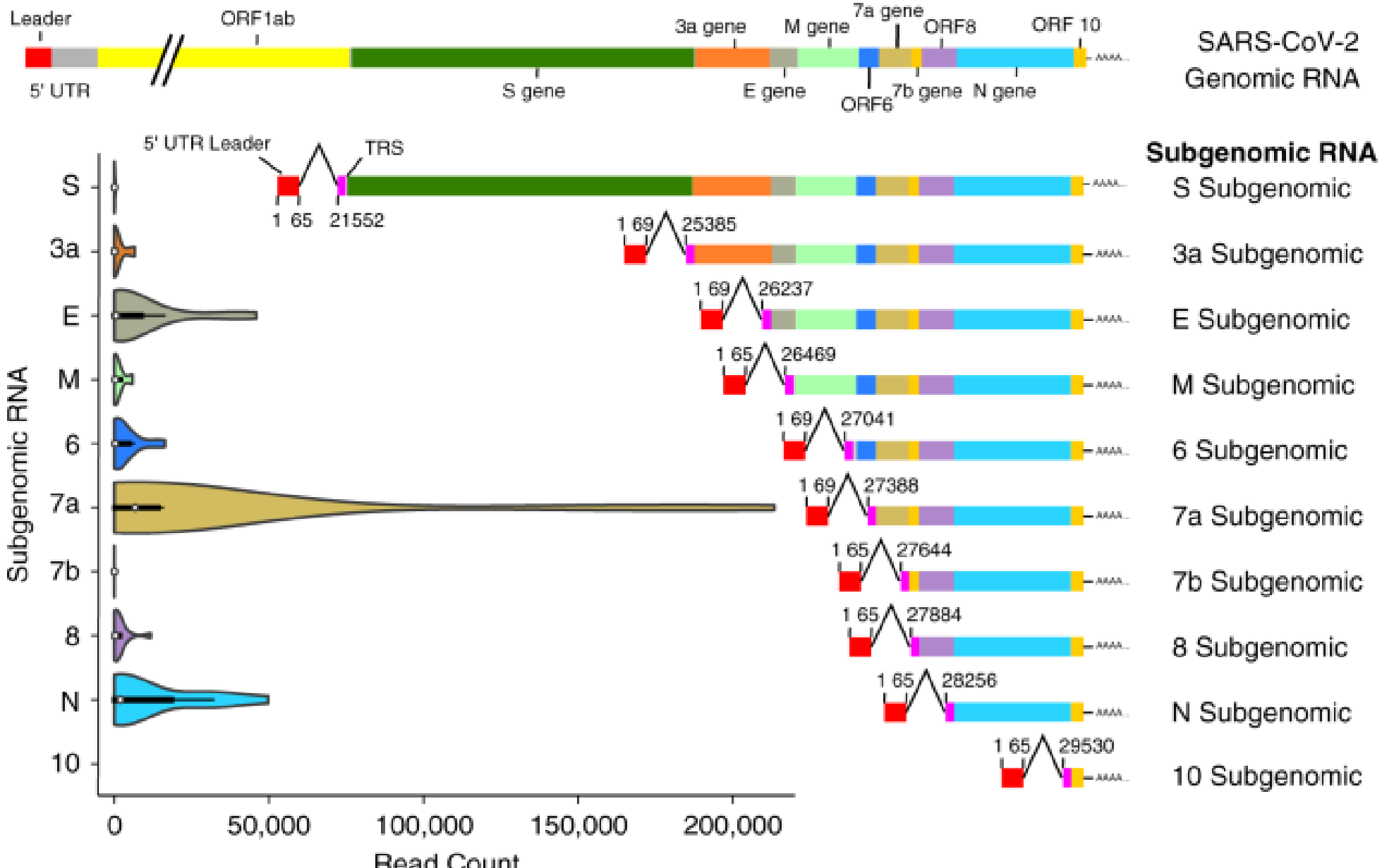
There are two major, partially overlapping, ORFs in the 5' two-thirds of the viral genome. ORF1a is frameshifted to ORF1b, and these encode the replicase polyproteins pp1a and pp1ab respectively. Polyproteins pp1a and pp1ab are proteolytically cleaved into 16 putative non-structural proteins (nsps).

Проблема 4. Трансляция поздних генов

- С РНК вируса транскрибируются мРНК поздних генов. Одна мРНК для одного позднего гена.
- мРНК каждого позднего гена устроена так:
  - Кэпированный 5’ концевой участок мРНК (кончается до ATG кодонов) соединенный с 3’ концевым участком, начинающимся перед ATG кодоном этого позднего гена и до конца
- Эти мРНК называются субгеномными мРНК (sgRNA)

- См. след. слайд

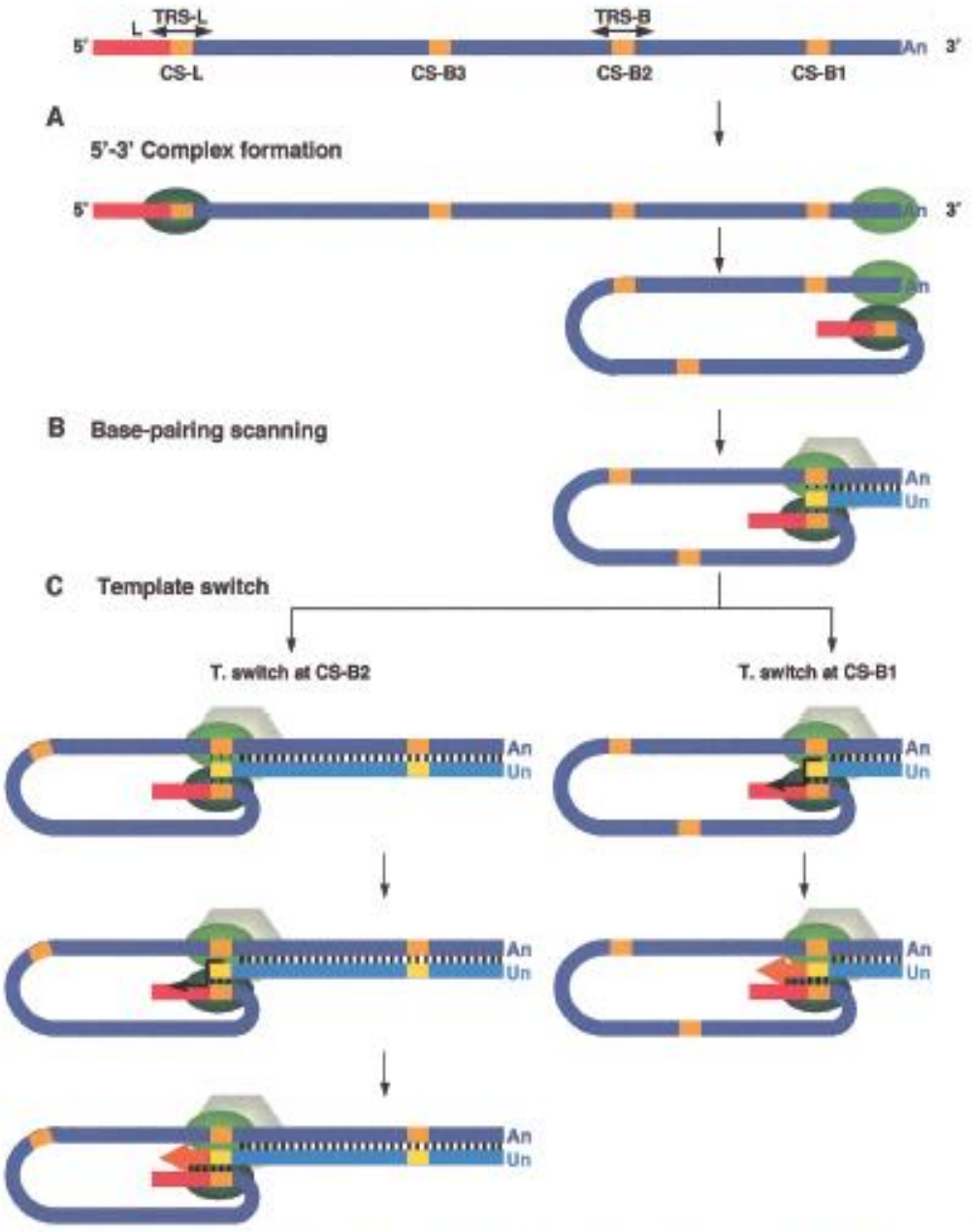
Fig. 1: SARS-CoV-2 genomic and subgenomic RNA structure showing genes and open reading frames (ORF) together with violin plots showing the number of reads per total of 5 million reads in the diagnostic samples mapped to the leader-containing subgenomic RNAs in the fasta file used for mapping.



Транскрипция вирусной РНК

- Вирусная РНК-зависимая РНК-полимераза (RdRP) закодирована в полипротеине (nsp11).
- RdRP по РНК матрице делает комплементарную копию. Из вирусной +RNA получается -RNA; из -RNA получается -(-RNA) = +RNA
- Сигналы разрывной транскрипции направляют перескок RdRP при синтезе -RNA, в результате которого синтезируются -sgRNA.
- -sgRNA является матрицей для RdRP; продукт – субгеномная мРНК (+sgRNA)
- Сигналы разрывной транскрипции называются так: TRS-L в лидере, TRS-B перед каждым поздним геном (TRS=transcription-regulatory sequences)

TRS-L и TRS-B



Лидер –красная полоска

Сигналы TRS – желтые прямоугольники. В них есть общее слово из шести букв (CS)

Мутации в CS влияют на синтез sgmRNA ожидаемым образом

Рисунок - гипотеза, косвенно подтвержденная

Zuniga et al., Journal of Virology, 2004

Сигналы разрывной транскрипции TRS-L, TRS-B; CS

Сигналы TRS-L и все TRS-B имеют высокосходные последовательности. Наиболее похожие их части, часто полностью совпадающие, называются CS (core sequences)

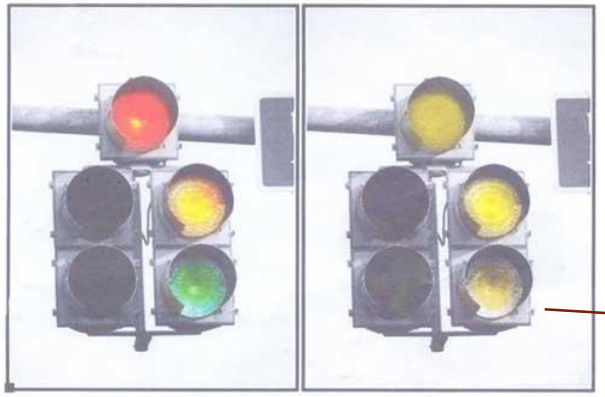
Принято считать, что длина CS – шесть нуклеотидов, TRS включает2-3 нуклеотида с 5’ и 3’ концов CS.

Как все в биологии значения длин не являются мировыми константами

Задание 2а: в геноме одного коронавируса найти сигналы TRS (CS)

- У вируса SARS-CoV-2 CS ACGAAC, встречается перед 7-ю из 10-и поздних генов.

Пример сигнала



Сигнал несет информацию

- **Адресован** пешеходам
- **Два состояния**: красный или зеленый (и не работает)
- **Предназначение** красного сигнала– пешеход не переходит улицу, даже если ему нужно

  Сигнал изменяет поведение получателя
  - Чем чаще сигнал приводит к и реакции, тем больше в нем информации (сильнее сигнал)
- **От кого сигнал** не важно и не всегда ясно

Так видит собака

Не предусмотренные получатели сигнала – бродячие собаки (https://pikabu.ru/story/sobaka_i_svetofor_5931508)

Много сигналов одновременно влияют на действия пешехода и бродячей собаки

- Красный цвет
- Отсутствие машин
- Реакция других пешеходов