Межфакультетский курс «Биоинформатика»
Факультет биоинженерии и биоинформатики МГУ
весна 2026

Лекция 10, часть 2

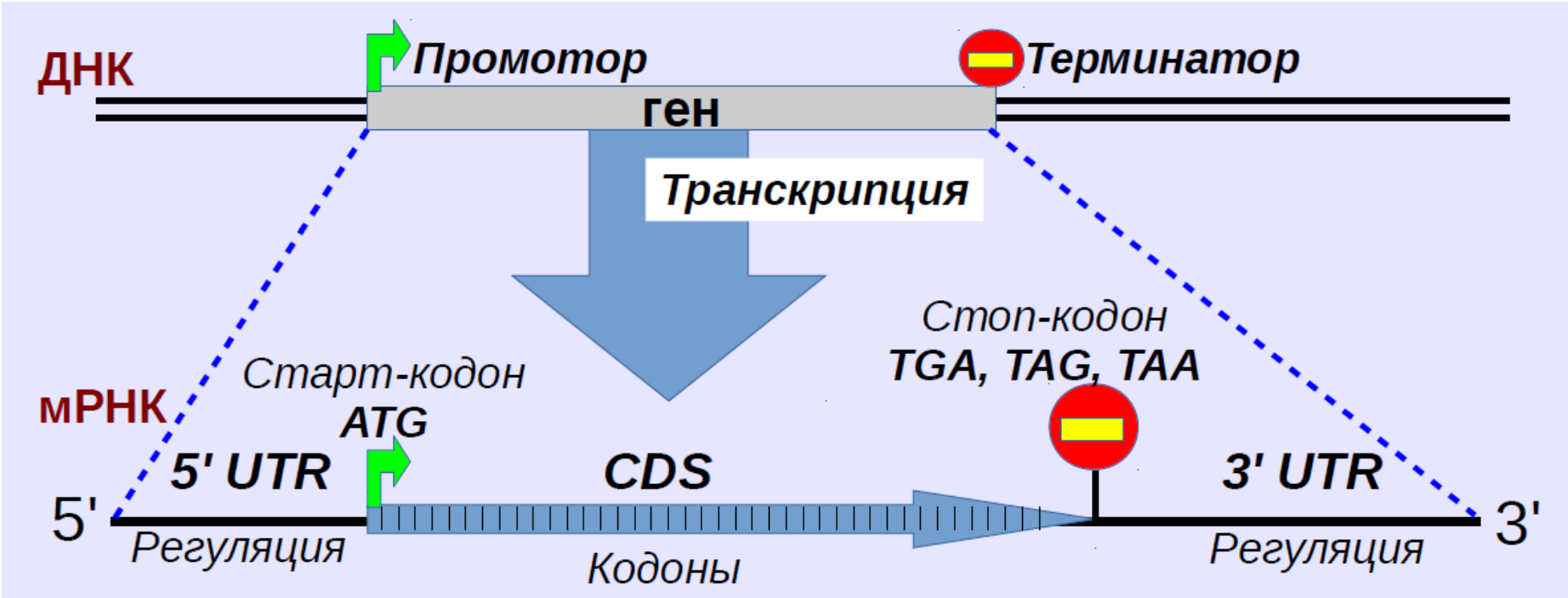
Предсказание генов

С.А.Спирин по слайдам А.А.Миронова

22 апреля 2026

Предсказание генов

Белок-кодирующий ген бактерии:



Задача: в геноме найти гены (или хотя бы CDS)

Видите ген?

gatctaccactgctaaggggaagtcatctaaattcttttgttaaatgtaccatccttcca
gatacaagtaggaaaagtcgccagaagacaagagctgtagggaaaaccaccaaccctatc
Ttcaaccacactatggatgagcaagtctgttccagcatttctccaagatgaggtgagtgg
cagtgtgatgagtgtttatagtggagactttggcaatctggaagttaaaggaaatattca
gtttgcaattgaatatgtggagtcactgaaggagttgcatgttttgtggcccagtgtaa
ggacttagcagcagcggatgtaaaaaaaacagcgttcagacccatatgtaaaggcctattt
gctaccagacaaaggcaaaatgggcaagaagaaaacactcgtagtgaagaaaaccttgaa
tcctgtgtataacgaaatactgcggtataaaattgaaaaacaaatcttaaagacacagaa
attgaacctgtccatttggcatcgggatacatttaagcgcaatagtttcctaggggaggt
ggaacttgatttggaaacatgggactgggataacaaacagataaacaattgagatggtac
cctctgaagcggaagacagcaccagttgcccttgaagcagaaaacagaggtgaaatgaaa
ctagctctccagtatgtcccagagccagtccctggtaaaaagcttcctacaactggagaa
gtgcacatctgggtgaaggaatgcctttgatgtatgatgggttcaggcctgaagatctga
tggaagcctgtgtagagcttactgtctgggaccattacaaattaaccaaccaatttttgg
gaggtcttcgtattggctttggaacaggtaaaagttatgggactga

… а он здесь есть!

gatctaccactgctaaggggaagtcatctaaattcttttgttaaatgtaccatccttcca
gatacaagtaggaaaagtcgccagaagacaagagctgtagggaaaaccaccaaccctatc
Ttcaaccacactatggatgagcaagtctgttccagcatttctccaagatgaggtgagtgg
cagtgtgatgagtgtttatagtggagactttggcaatctggaagttaaaggaaatattca
gtttgcaattgaatatgtggagtcactgaaggagttgcatgttttgtggcccagtgtaa
ggacttagcagcagcggatgtaaaaaaacagcgttcagacccatatgtaaaggcctattt
gctaccagacaaaggcaaaatgggcaagaagaaaacactcgtagtgaagaaaaccttgaa
tcctgtgtataacgaaatactgcggtataaaattgaaaaacaaatcttaaagacacagaa
attgaacctgtccatttggcatcgggatacatttaagcgcaatagtttcctaggggaggt
ggaacttgatttggaaacatgggactgggataacaaacagataaacaattgagatggtac
cctctgaagcggaagacagcaccagttgcccttgaagcagaaaacagaggtgaaatgaaa
ctagctctccagtatgtcccagagccagtccctggtaaaaagcttcctacaactggagaa
gtgcacatctgggtgaaggaatgcctttgatgtatgatgggttcaggcctgaagatctga
tggaagcctgtgtagagcttactgtctgggaccattacaaattaaccaaccaatttttgg
gaggtcttcgtattggctttggaacaggtaaaagttatgggactga

Рамки считывания

Рамкой считывания называется разбиение последовательности ДНК на кодоны. Каждой последовательности ДНК соответствует шесть разных рамок.

```
… atggtattatatggacaa …
… atg gta tta aat gga cat ga  … рамка 1
… Met-Val-Leu-Asn-Gly-His     …

… a tgg tat taa atg gac atg a … рамка 2
…   Trp-Tyr-Stp-Met-Asp-Met   …

… at ggt att ata agg aca tga  … рамка 3
…    Gly-Ile-Ile-Arg-Thr-Stp  …
```

+ три рамки на комплементарной цепи

Открытая рамка считывания (ORF)

Открытая рамка считывания — последовательность кодонов от старт-кодона до стоп-кодона, без стоп-кодонов между ними. Сокращение: ORF = Open Reading Frame

Простейший вариант предсказания гена (у прокариот): максимальная (т.е. нерасширяемая) ORF, если она длиннее заданного порога (скажем, 100 кодонов)

$$P(ORF \geqslant 100) = P(\text{не стоп})^{100} = (1 - 3/64)^{100} = 0.0082$$

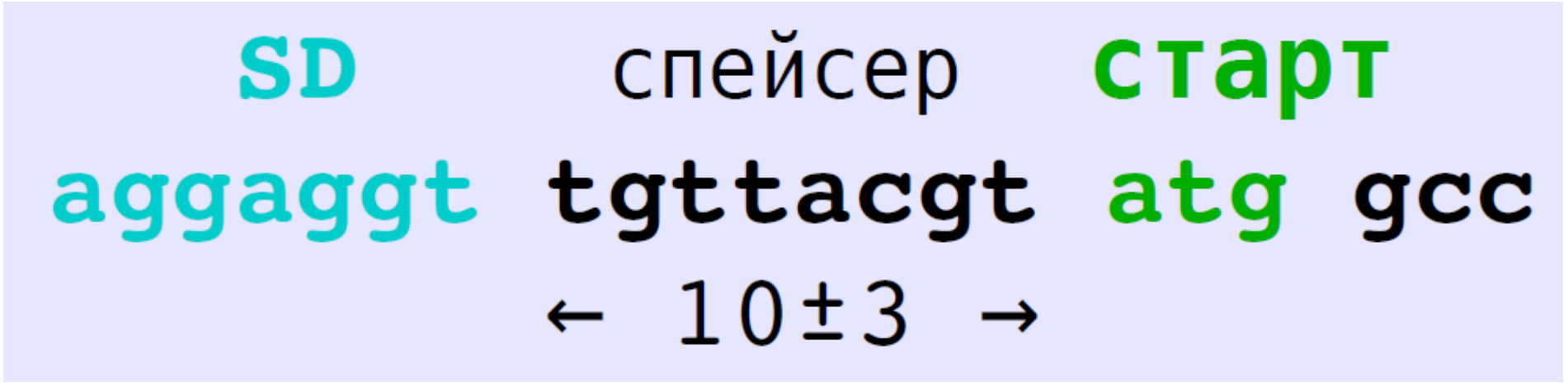
Проблемы с максимальной ORF

- Не любой кодон ATG — старт-кодон. Даже если мы угадали рамку, где настоящее начало CDS?
- Кроме ATG, у бактерий встречаются другие старт-кодоны (чаще всего GTG)
- Порог на длину произволен. При низком пороге набираем много лишнего, при высоком — теряем гены коротких белков.

atg agc aag tct gtt cca atg ttt
ctc caa gat gag gtg agt tgt atg
cat gaa atg tac tca acc taa tga

Дополнительные сообращения

- Кодоны используются с разной частотой
- У многих прокариот в 5’-некодирующей области (т.е. перед старт-кодоном) имеется дополнительный сигнал — последовательность Шайна – Дальгарно (SD-последовательность).
  Но SD-последовательность вырожденная, и не во всех генах она есть.



Вероятностная модель

Каждой ORF припишем вероятность P([a,b]=gene) того, что она кодирует настоящий белок.

Учтем:
- Характер использования кодонов.
- Наличие и качество SD-последовательности.
- Длину рамки.
- Тип стоп-кодона.
- Что-нибудь еще.

Вероятностная модель

- Вероятность, что фрагмент порожден моделью CDS
  prob(x[a,b]|M)
- Вероятность, что фрагмент порожден случайной моделью
  prob(x[a,b]|R)
- *Логарифм* ***отношения правдоподобия:***

$$L(x[a,b]\,is\,CDS) = \log \frac{prob(x[ab]|M)}{prob(x[ab]|R)}$$

Если *L* выше порога, предсказываем CDS

Качество предсказания

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TP}{TP + FP}$$

$$J = \frac{TP}{TP + FP + FN}$$

Откуда брать то, что «на самом деле»?

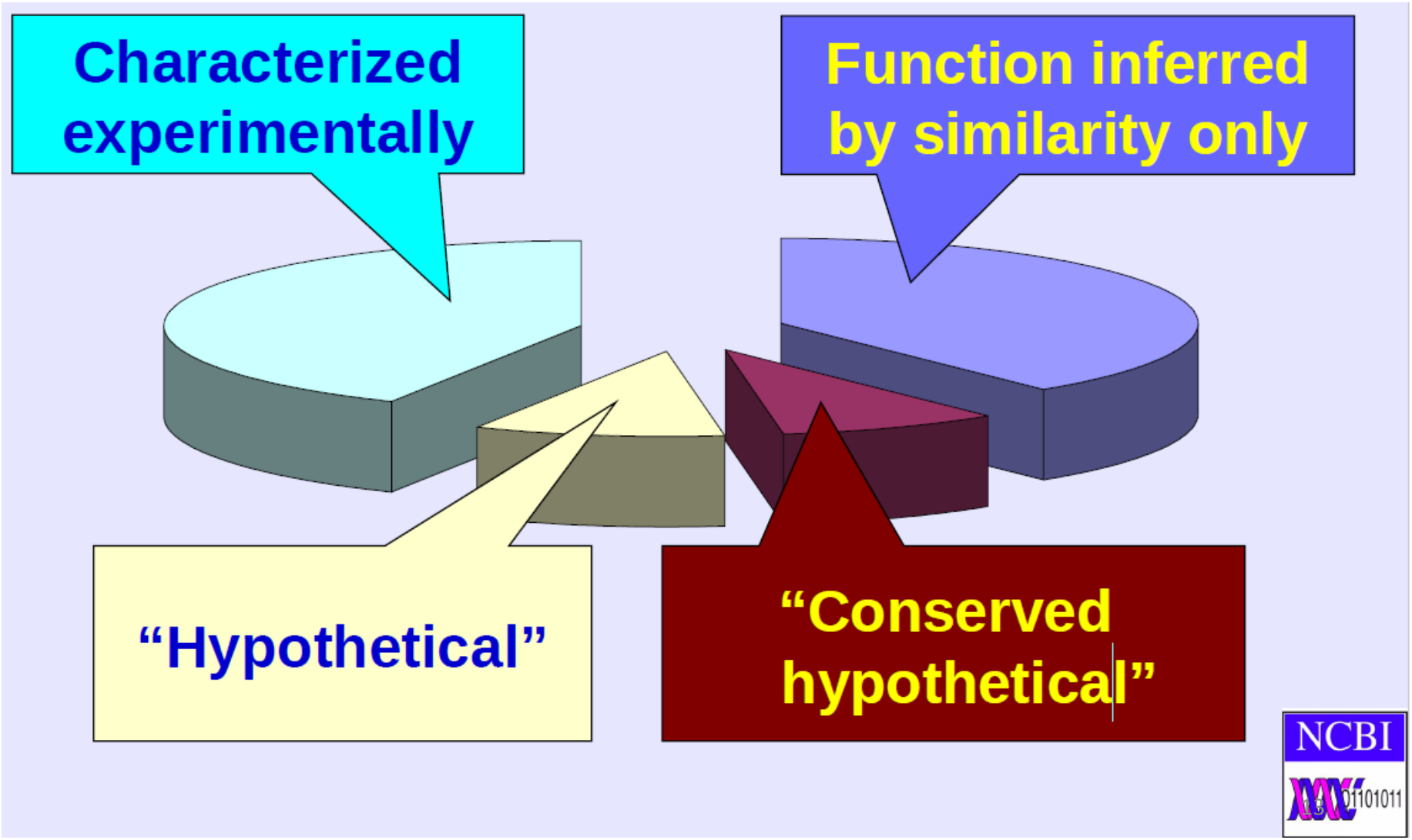
- Надо читать белок (N-конец)
  (N-конец — **Н**ачало; С-конец — коне**Ц**)
  – Разрезание + хроматография
  – Масс-спектрометрия
  Учитываем процессинг белка — N-конец зрелого белка не всегда соответствует первым кодонам CDS!
- Рибосомное профилирование. Там, где есть рибосомы — там кодирующая область.

Экспериментальных данных мало.

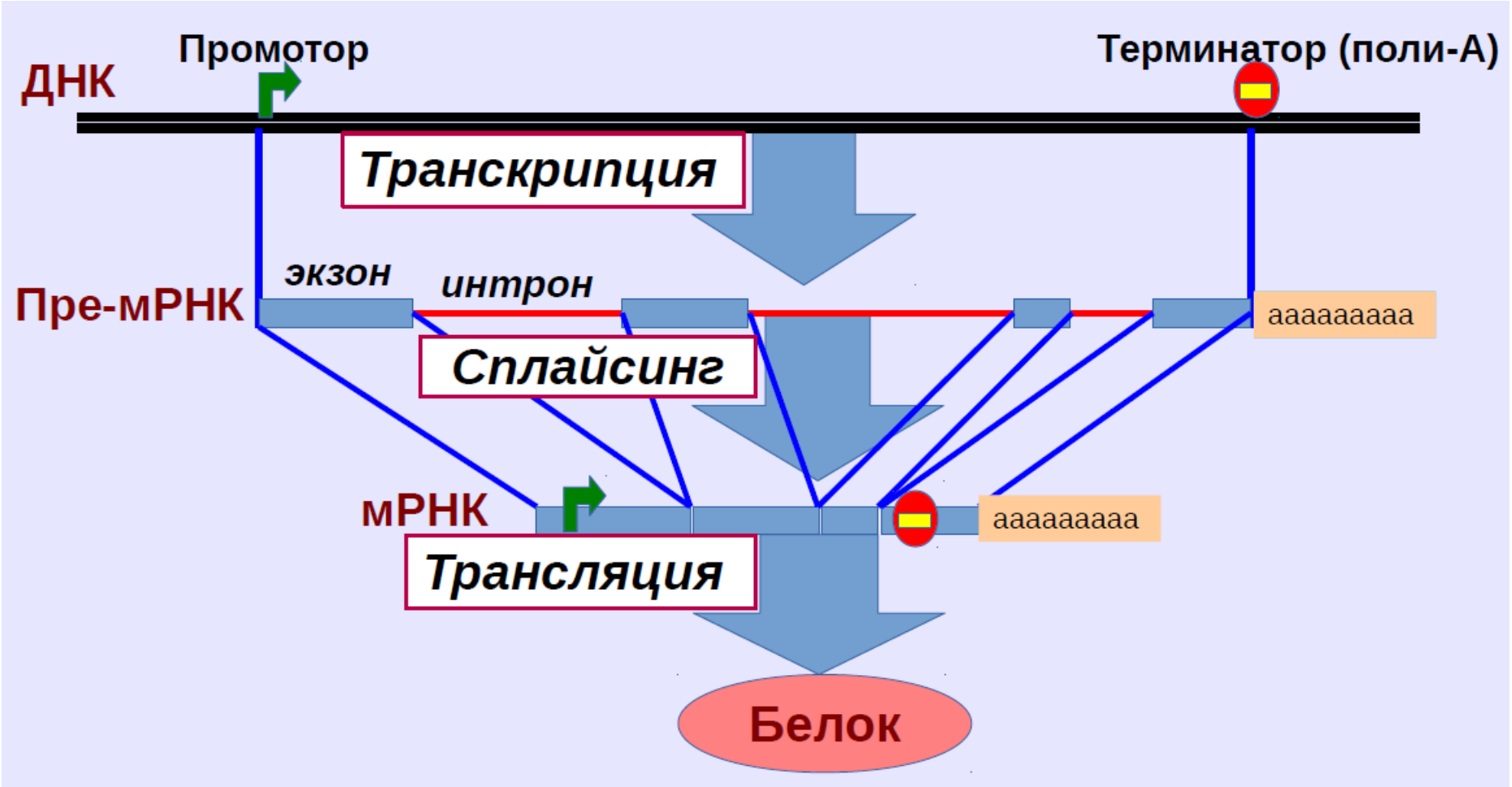
Сравнительный анализ?

Что мы знаем о протеоме кишечной палочки?

Escherichia coli — самая изученная бактерия



Эукариоты: сплайсинг



Предсказание генов в эукариотах

Плохая новость – сплайсинг
Хорошая новость — сплайсинг определяется сайтами сплайсинга

Предсказание генов в эукариотах

Модель учитывает:
– Сайты сплайсинга
– Распределение длин интронов
– Распределение длин экзонов
– Частоту использования кодонов

Обучение модели (подбор параметров):
– Ищем известные гены (BLAST)
– Определяем на них частоты кодонов и распределение длин интронов и экзонов

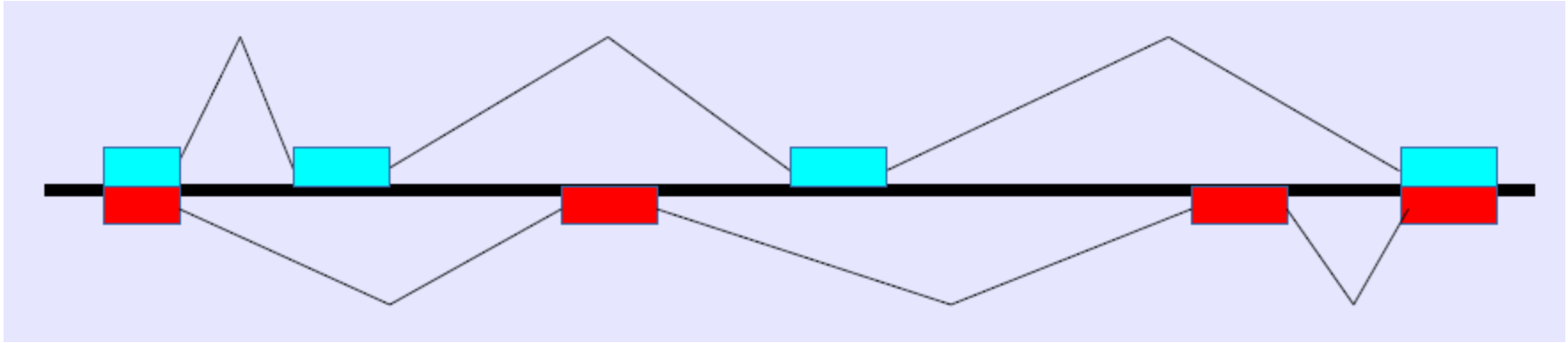
Предсказание генов в эукариотах

Проблема: предсказание сливает и разделяет гены

Предсказание генов в эукариотах

Плохие новости:

– сплайсинг бывает альтернативен



– редактирование РНК

gtttgCcctaAgctgc

gtttgTcctaIgctgc

U ~ T
I ~ G

caauugc gcauuc caaccg

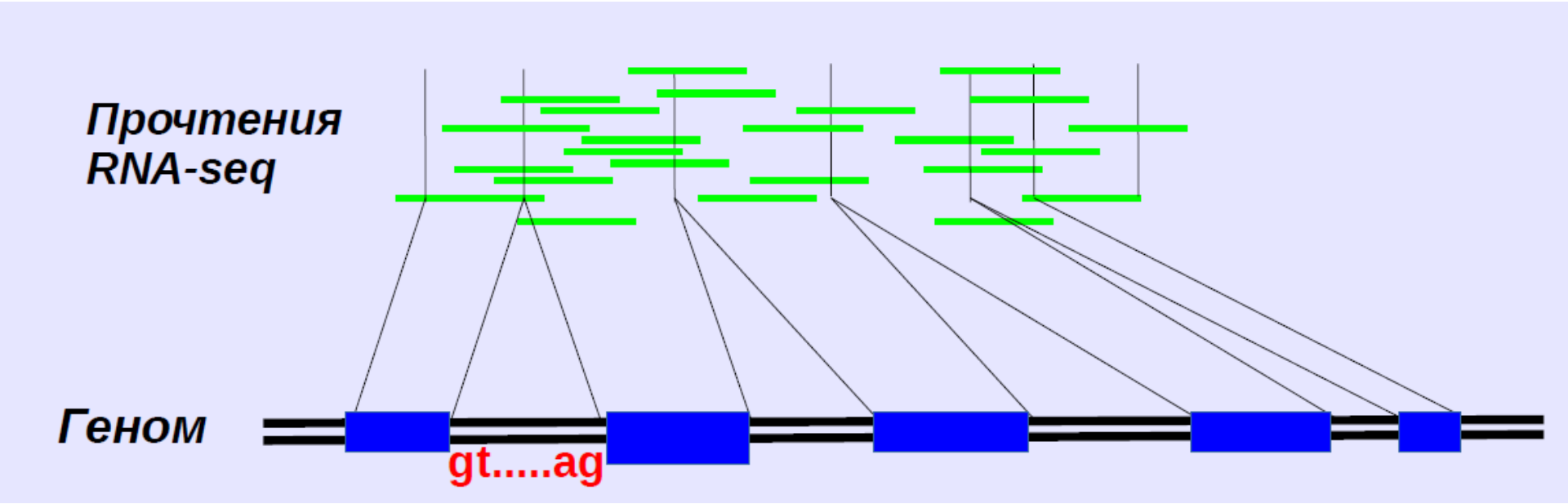
guuaacgAAcguaagAguuggc

caauugcUUgcauucUcaaccg

Предсказание генов в эукариотах

Хорошая новость:

– можно секвенировать зрелую РНК, выровнять её на геном и узнать, где экзоны



Но:

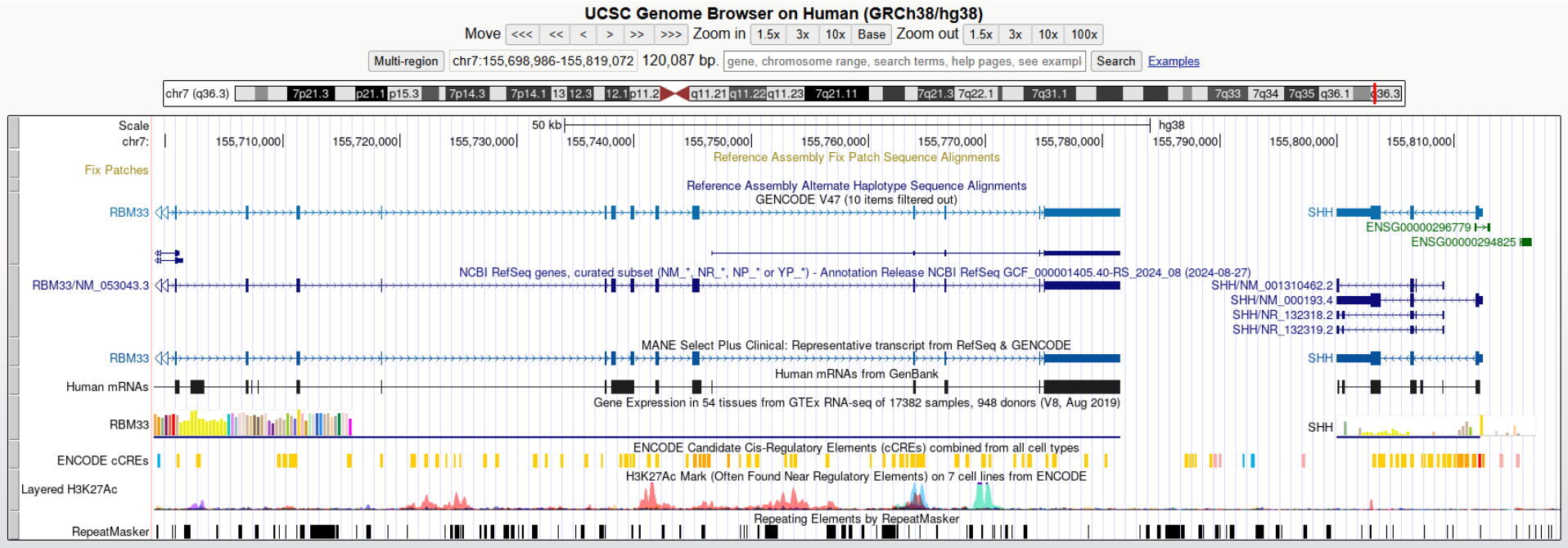
- Многие гены экспрессируются очень редко и слабо
- Изредка экспрессируются случайные фрагменты генома

Предсказание генов в эукариотах

Надо использовать всю информацию:

– учесть сайты сплайсинга
– учесть частоты кодонов
– сравнивать с геномами близких организмов
– использовать результаты RNA-Seq (если они доступны)

Как это выглядит



https://genome-euro.ucsc.edu/
https://www.ensembl.org/