Предсказание генов

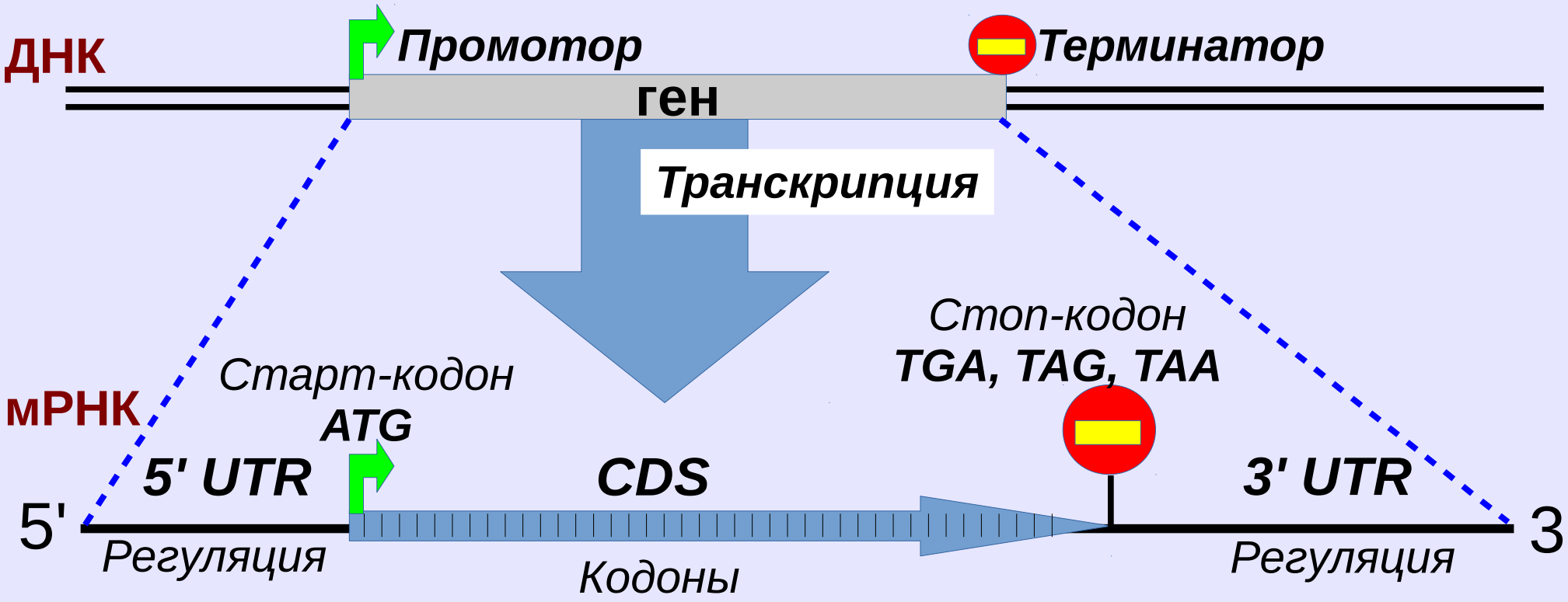
Миронов Андрей Александрович

ФББ МГУ

Межфакультетский курс “Биоинформатика”

Предсказание генов

Белок-кодирующий Ген:



Задача: в геноме найти гены (**CDS**)

Видите ген?

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

… а он здесь есть!

gatctaccactgctaaggggaagtcatctaaattcttttgttaaatgtaccatccttcca
gatacaagtaggaaaagtcgccagaagacaagagctgtagggaaaaccaccaaccctatc
Ttcaaccacactatgg**atgagcaagtctgttccagcatttctccaagatgaggtgagtgg**
cagtgtgatgagtgtttatagtggagactttggcaatctggaagttaaaggaaatattca
gtttgcaattgaatatgtggagtcactgaaggagttgcatgttttgtggcccagtgtaa
ggacttagcagcagcggatgtaaaaaaacagcgttcagacccatatgtaaaggcctattt
gctaccagacaaaggcaaaatgggcaagaagaaaacactcgtagtgaagaaaaccttgaa
tcctgtgtataacgaaatactgcggtataaaattgaaaaacaaatcttaaagacacagaa
attgaacctgtccatttggcatcgggatacatttaagcgcaatagtttcctaggggaggt
ggaacttgatttggaaacatgggactgggataacaaacagataaacaattgagatggtac
cctctgaagcggaagacagcaccagttgcccttgaagcagaaaacagaggtgaaatgaaa
ctagctctccagtatgtcccagagccagtccctggtaaaaagcttcctacaactggagaa
gtgcacatctgggtgaaggaatgcctttgatgtatgatgggttcaggcctgaagatctga
tggaagcctgtgtagagcttactgtctgggaccattacaaattaaccaaccaatttttgg
gaggtcttcgtattggctttggaacaggtaaaagttatgggactga

Открытая рамка считывания (ORF)

- ***Рамка считывания*** – последовательность кодонов, начиная от старт-кодона до стоп-кодона

```
… 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  …
```

+ комплементарная цепь (3 рамки) → итого 6 возможных рамок считывания

Максимальная ORF

Соображения:

1. Начинается со старт-кодона
2. Кончается стоп-кодоном
3. Нет стоп-кодонов в рамке.

Максимальная открытая рамка считывания.
Если длиннее 100 кодонов, то белок.

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

Проблемы max ORF

- Не любой ATG есть старт-кодон. Не любой старт-кодон начинаетсяс ATG.

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

- Где на самом деле находится старт белка?
- Какая рамка считывания (ORF) на самом деле есть ген?
- Порог на длину: либо теряем короткие пептиды, либо набираем много мусора.

Дополнительные соображения

- 1. Не все кодоны используются с одинаковой частотой.
- 2. Во многих прокариотах есть дополнительный сигнал — последовательность Шайна-Дальгарно (SD-последовательность). SD-последовательность вырожденная, и не все гены его используют.

SD спейсер **старт**

aggaggt tgttacgt atg gcc

← 10±3 →

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.
  - Чего-нибудь еще.

Вероятностная модель

- Вероятность, что фрагмент порожден моделью CDS

$$prob(x[a,b]|M)$$

- Вероятность, что последовательность порождена случайной моделью

$$prob(x[a,b]|R)$$

- ***Отношение правдоподобия***

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

Качество предсказания

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

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

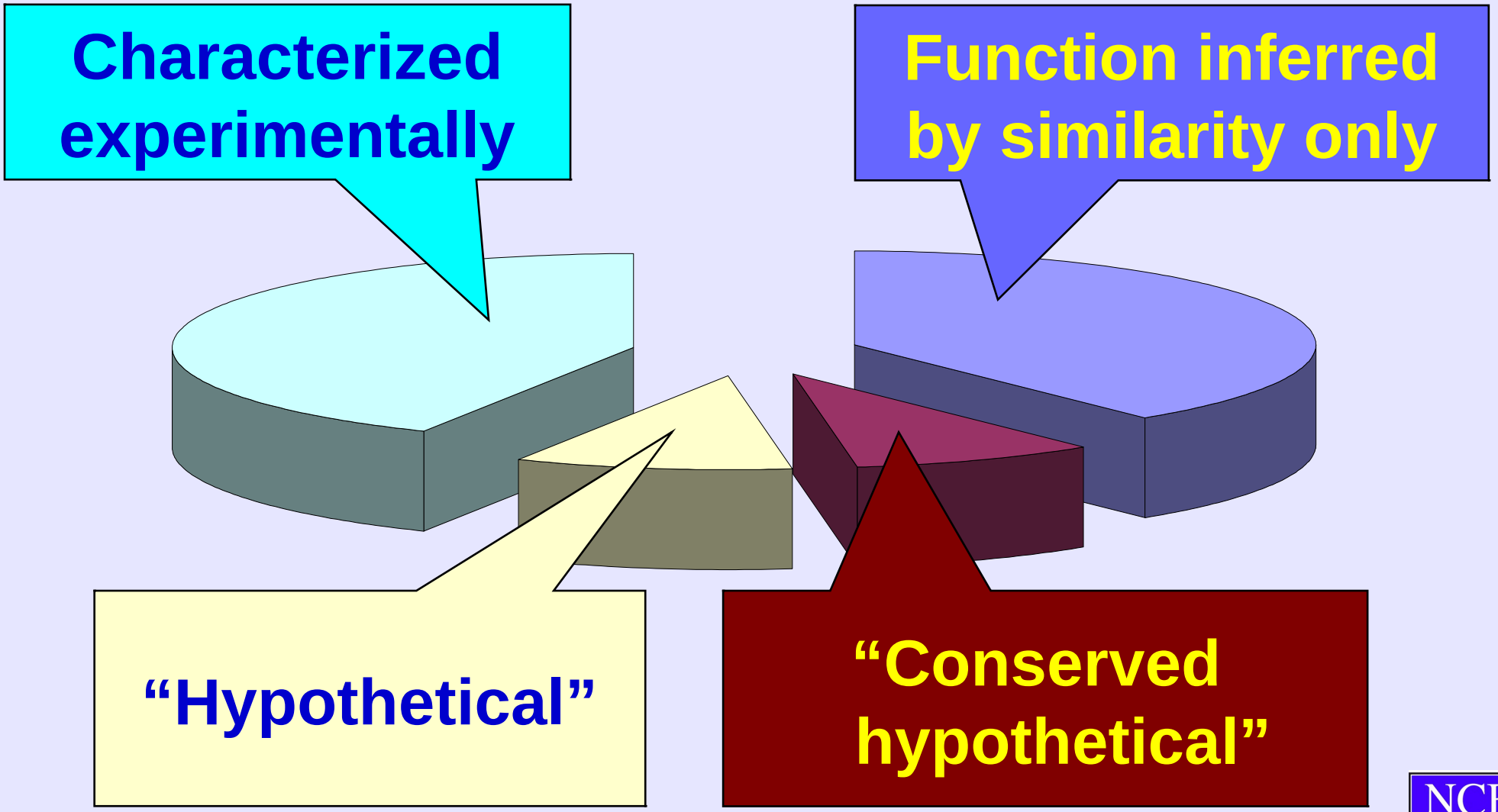
$$J = \frac{TP}{TP + FP + FN}$$

А что есть истина

- Надо читать белок (N-конец) (N-конец — **Н**ачало; С-конец — коне**Ц**)
  - Разрезание + хромотография
  - Масс-спектрометрия
- Процессинг белка — N-конец зрелого белка не есть его истинное начало!
- Рибосомное профилирование. Там, где есть рибосомы — там кодирующая область.
- Экспериментальных данных мало

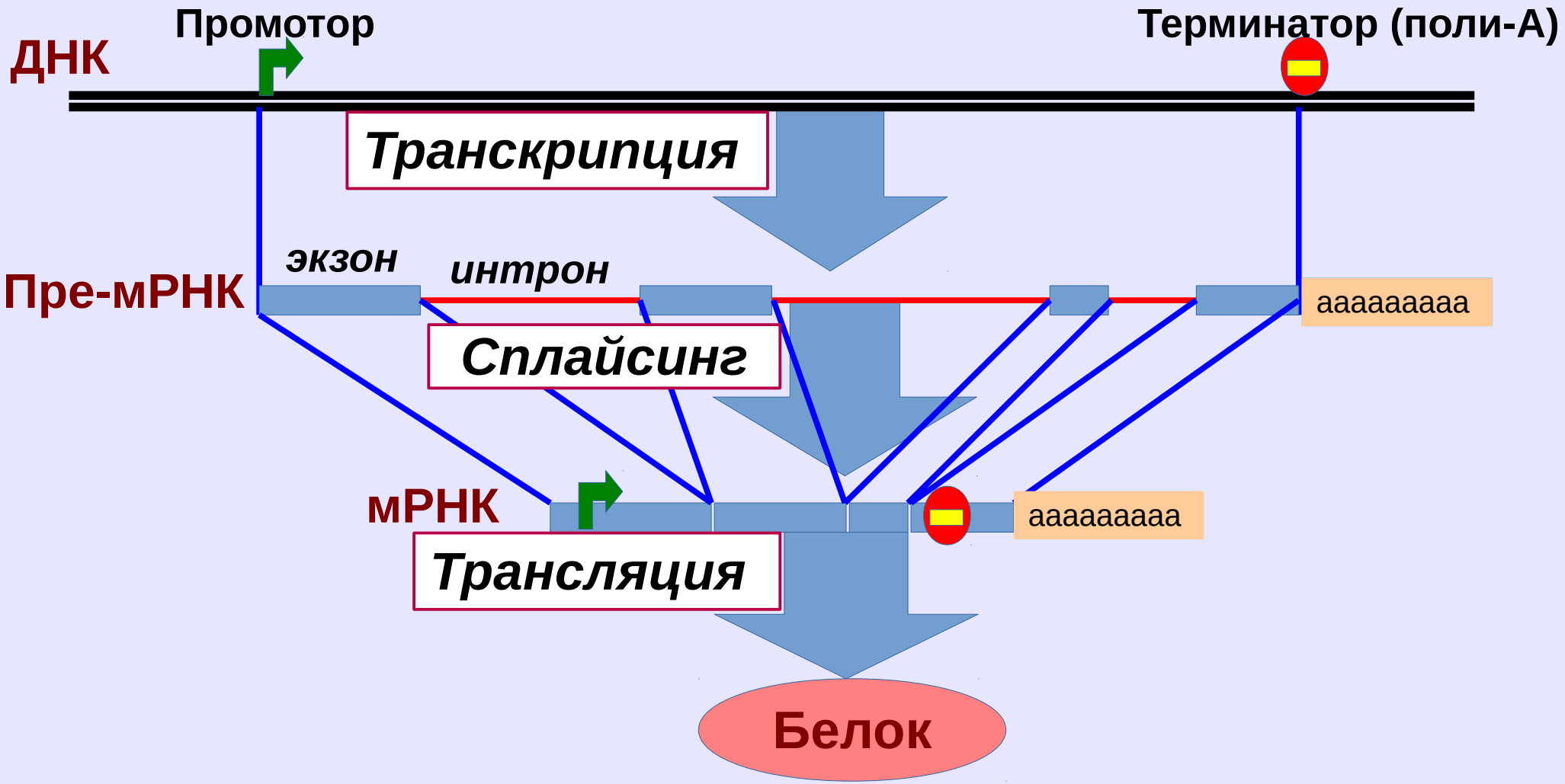
Сравнительный анализ ?

Что мы знаем о протеоме *Escherichia coli* ?



NCBI

Эукариоты: *Сплайсинг*



Предсказание генов в эукариотах

- Плохая новость – сплайсинг
- Хорошая новость — сплайсинг определяется сайтами сплайсинга

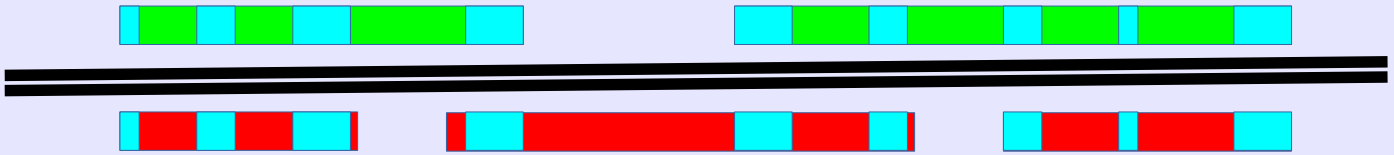
Предсказание генов в эукариотах

- 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

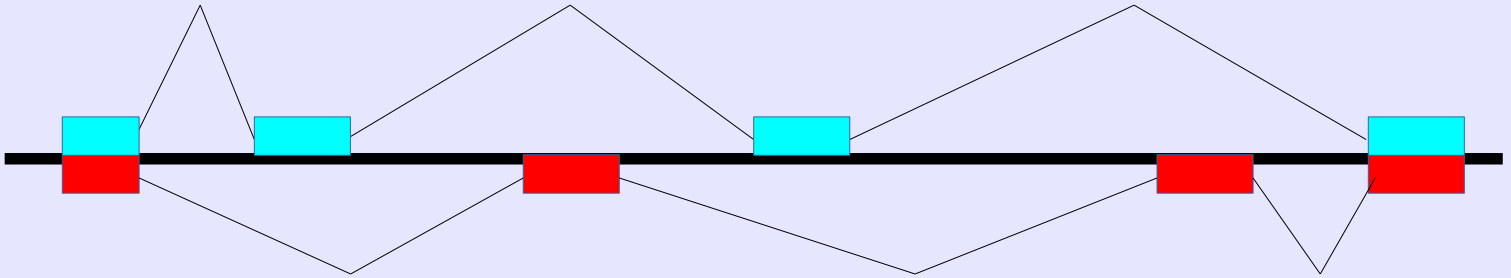
На самом деле



предсказание

Предсказание генов в эукариотах

- Плохие новости –
  - Сплайсинг ***альтернативен***



  - ***Редактирование*** РНК

gtttg**C**ccta**A**gctgc

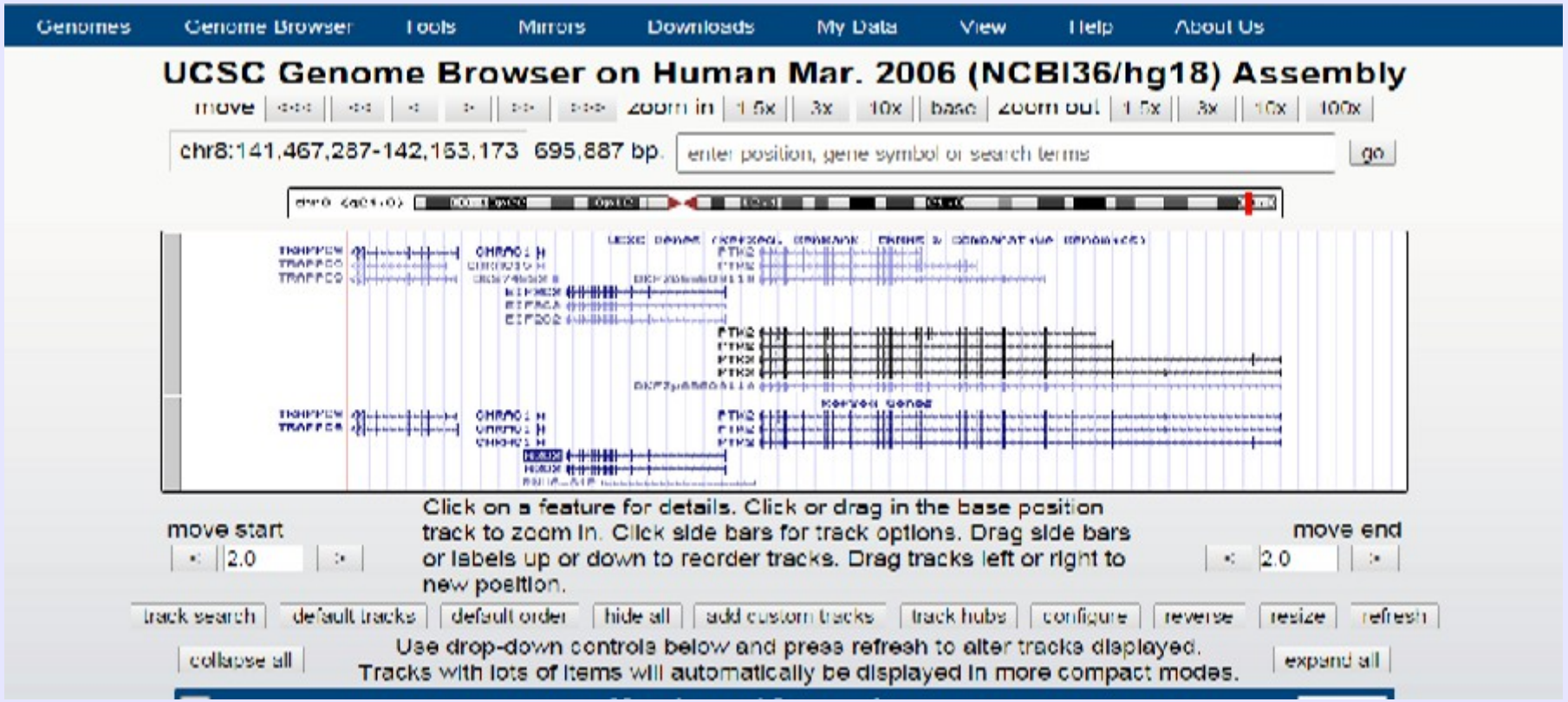
gtttg**T**ccta**I**gctgc

U ~ T
I ~ G

caauugc gcauuc caaccg
guuaacg**AA**cguaag**A**guuggc

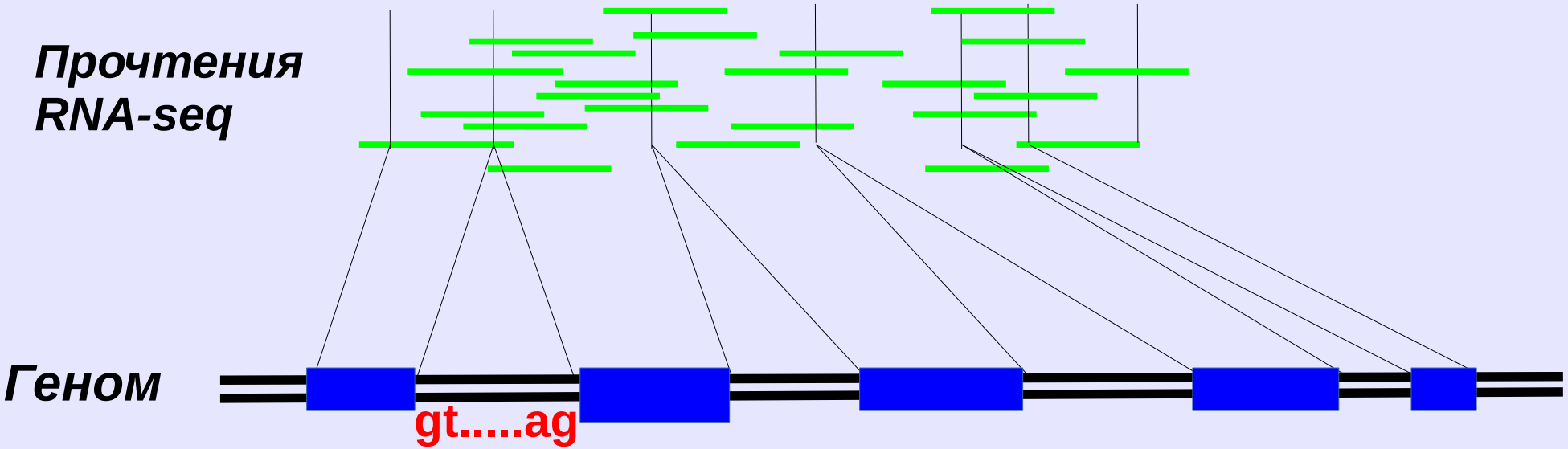
caauugc**UU**gcauuc**U**caaccg

Как это выглядит



Предсказание генов в эукариотах

- Хорошая новость – можно секвенировать зрелую мРНК, положить ее на геном и узнать, где экзоны.



**Не все гены экспрессируются;
Экспрессируются не только гены**

Гибридные подходы

- Учесть сайты сплайсинга
- Учесть частоты кодонов
- Принять во внимание возможные гомологи в других геномах
- Использовать информацию по RNA-SEQ (если она есть)