

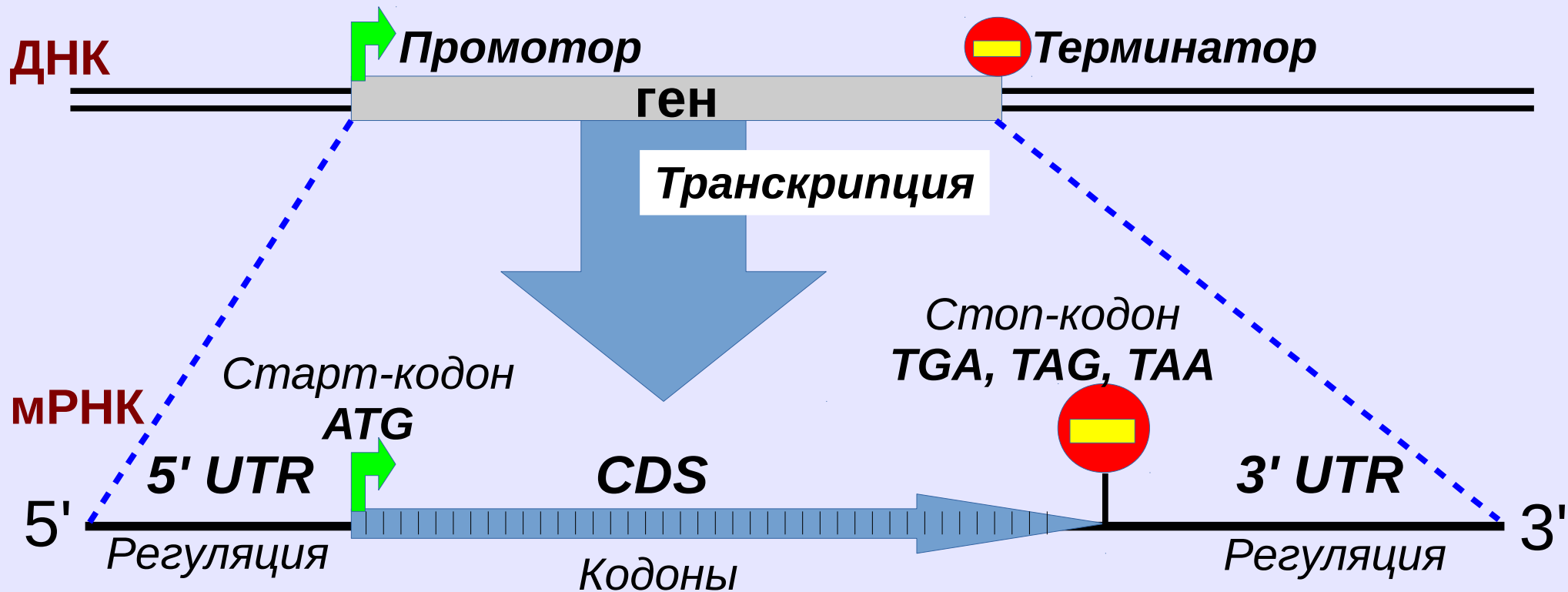
Предсказание генов

Миронов Андрей Александрович
ФББ МГУ

Межфакультетский курс “Биоинформатика”

Предсказание генов

Белок-кодирующий Ген:



Задача: в геноме найти гены (**CDS**)

Видите ген?

gatctaccactgctaaggggaagtcacctaattcttttgttaaatgtaccatccttcca
gatacaagtaggaaaagtcgccagaagacaagagctgtagggaaaaccaccaaccctatc
Ttcaaccacactatggatgagcaagtctgttccagcatttctccaagatgagggtgagtg
cagtgtgatgagtgtttatagtggagactttggcaatctggaagttaaaggaaatattca
gtttgcaattgaatatgtggagtcactgaaggagttgcatgtttttgtggcccagtgtaa
ggacttagcagcagcggatgtaaaaaaacagcgttcagaccatatagtaaaggcctat
gctaccagacaaaggcaaaatgggcaagaagaaaacactcgtagtgaagaaaaccttga
tcctgtgtataacgaaatactgcggtataaaattgaaaaacaaatcttaaagacacagaa
attgaacctgtccatttggcatcgggatacatttaagcgcfaatagtttcttaggggaggt
ggaacttgatttggaaacatgggactgggataacaacagataaacaattgagatggtac
cctctgaagcgggaagacagcaccagttgcccttgaagcagaaaacagaggtgaaatgaaa
ctagctctccagtatgtcccagagccagtccttggtaaaaagcttcttacaactggagaa
gtgcacatctgggtgaaggaatgcctttgatgtatgatgggttcaggcctgaagatctga
tggaagcctgtgtagagcttactgtctgggaccattacaaattaaccaaccaatTTTTGG
gaggtcttcgtattggcctttggaacaggtaaaaagttatgggactga

... а он здесь есть!

gatctaccactgctaaggggaagtcacataaattcttttgttaaatgtaccatccttcca
gatacaagtaggaaaagtcgccagaagacaagagctgtagggaaaaccaccaaccctatc
Ttcaaccacactatgg**atgagcaagtctgttccagcatttctccaagatgaggtgagtgg**
cagtgtgatgagtgtttatagtggagactttggcaatctggaagttaaaggaaatattca
gtttgcaattgaatatgtggagtcactgaaggagttgcatgtttttgtggcccagtgtaa
ggacttagcagcagcggatgtaaaaaaacagcgttcagaccatatagtaaaggcctat
gctaccagacaaaggcaaaatgggcaagaagaaaacactcgtagtgaagaaaaccttgaa
tcctgtgtataacgaaatactgcggtataaaattgaaaaacaaatcttaaagacacagaa
attgaacctgtccatttggcatcgggatacatttaagcgcfaatagtttcctaggggaggt
ggaacttgatttggaaacatgggactgggataacaaacagataaacaattgagatggtac
cctctgaagcgggaagacagcaccagttgcccttgaagcagaaaacagaggtgaaatgaaa
ctagctctccagtatgtcccagagccagtccttggtaaaaagcttcctacaactggagaa
gtgcacatctgggtgaaggaatgcctttgatgtatgatgggttcaggcctgaagatctga
tggaagcctgtgtagagcttactgtctgggaccattacaaattaaccaaccaatTTTTGG
gaggtcttcgtattggctttggaacaggtaaaagttatgggactga

Открытая рамка считывания (ORF)

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

... atggtattatatggaca ...

... 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.

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

Проблемы max ORF

- Не любой ATG есть старт-кодон. Не любой старт-кодон начинается с 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 $P([a,b]=\text{gene})$ 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

$$\text{prob}(x[a,b]|M)$$

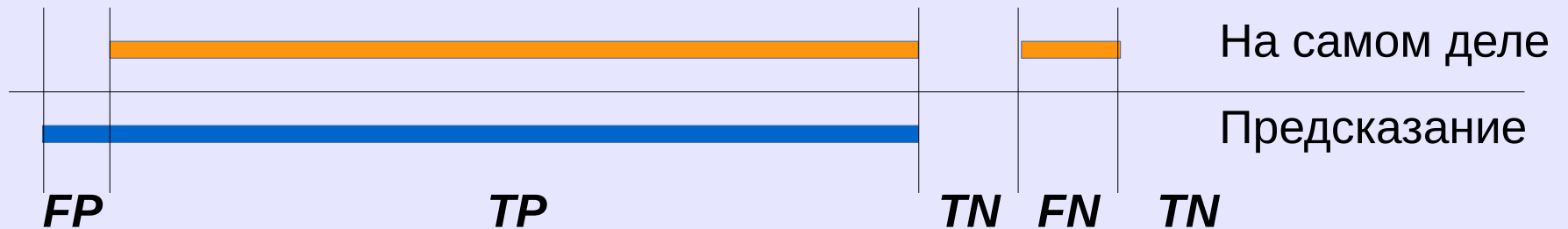
- Вероятность, что последовательность порождена случайной моделью

$$\text{prob}(x[a,b]|R)$$

- Отношение правдоподобия

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

Качество предсказания



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

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

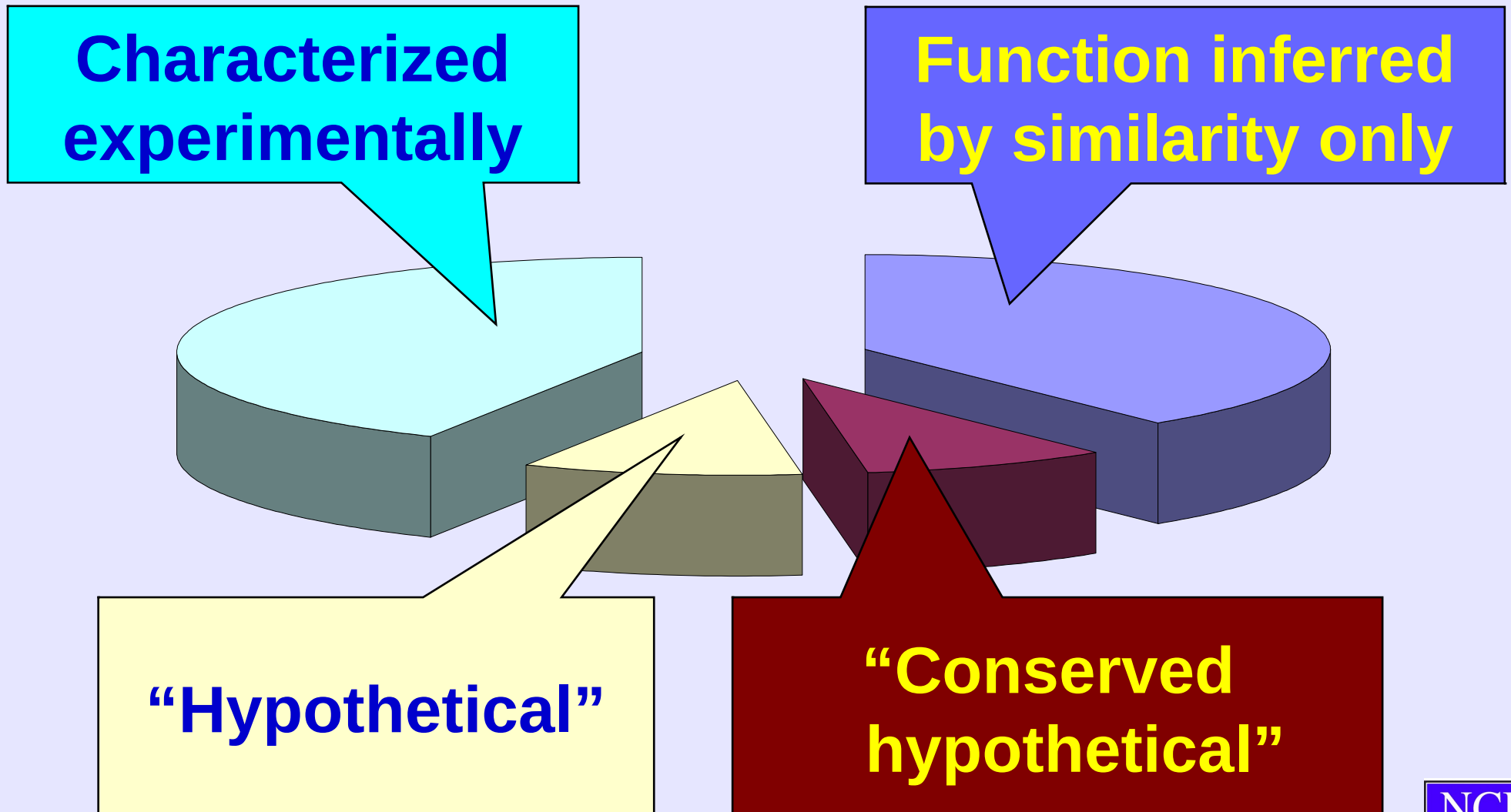
$$J = \frac{TP}{TP + FP + FN}$$

А что есть истина

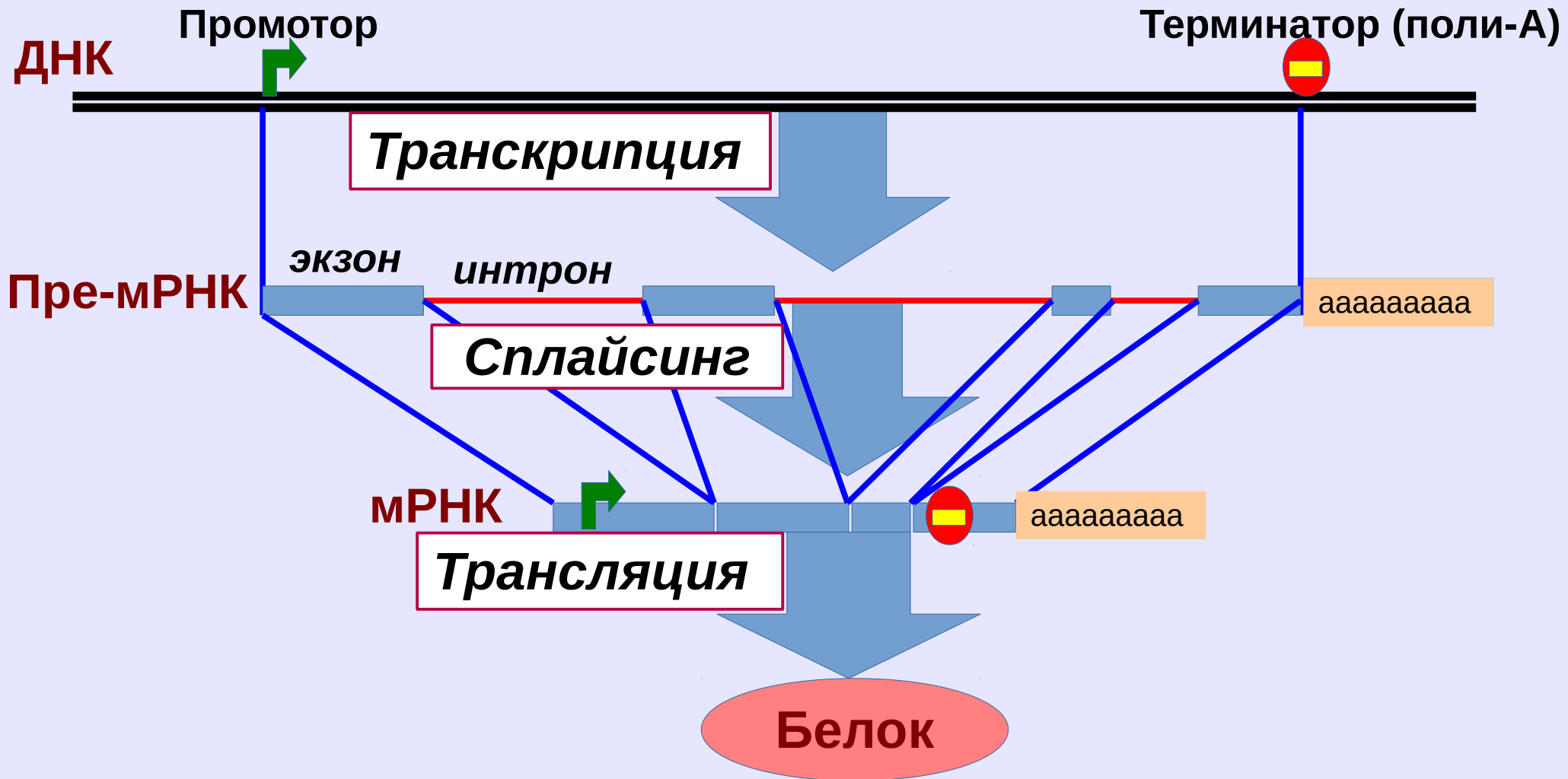
- Надо читать белок (N-конец) (N-конец — Начало; С-конец — концеЦ)
 - Разрезание + хроматография
 - Масс-спектрометрия
- Процессинг белка — N-конец зрелого белка не есть его истинное начало!
- Рибосомное профилирование. Там, где есть рибосомы — там кодирующая область.
- Экспериментальных данных мало

Сравнительный анализ ?

Что мы знаем о протеоме *Escherichia coli* ?



Эукариоты: Сплайсинг



Предсказание генов в эукариотах

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



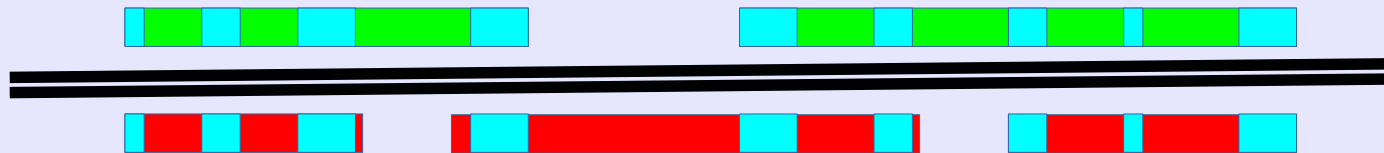
Предсказание генов в эукариотах

- 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 сливается и разделяет гены

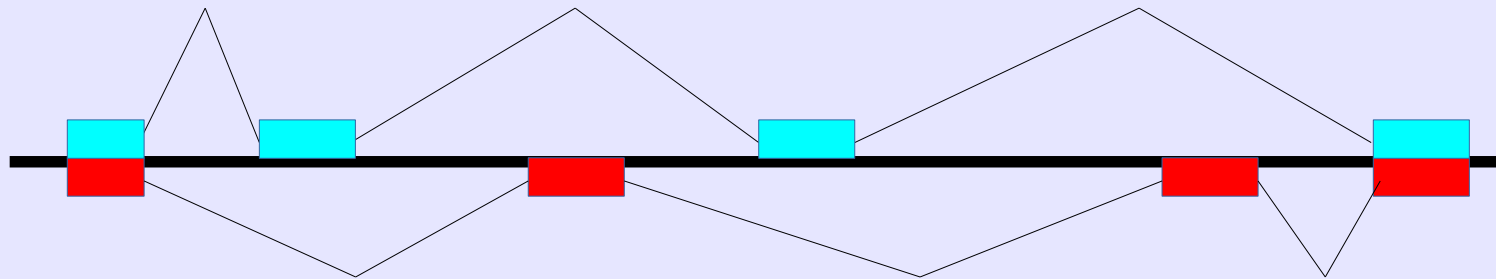
На самом деле



предсказание

Предсказание генов в эукариотах

- Плохие новости –
 - Сплайсинг *альтернативен*



- *Редактирование РНК*

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

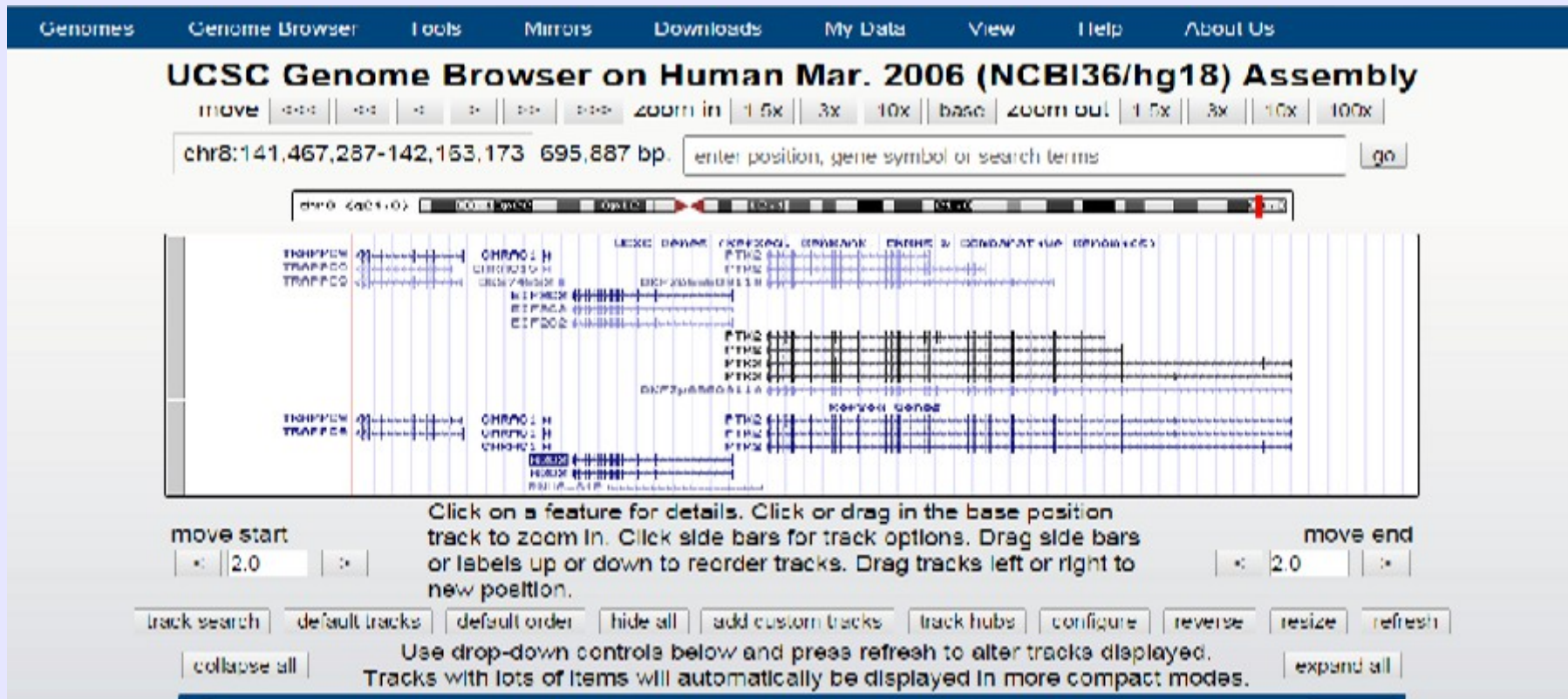
gtttg**T**ccta**I**gctgc

U ~ T
I ~ G

сааиугс гсаиис саассг
гуааасг**АА**сгуааг**А**гуиугс

сааиугс**UU**гсаиис**U**саассг

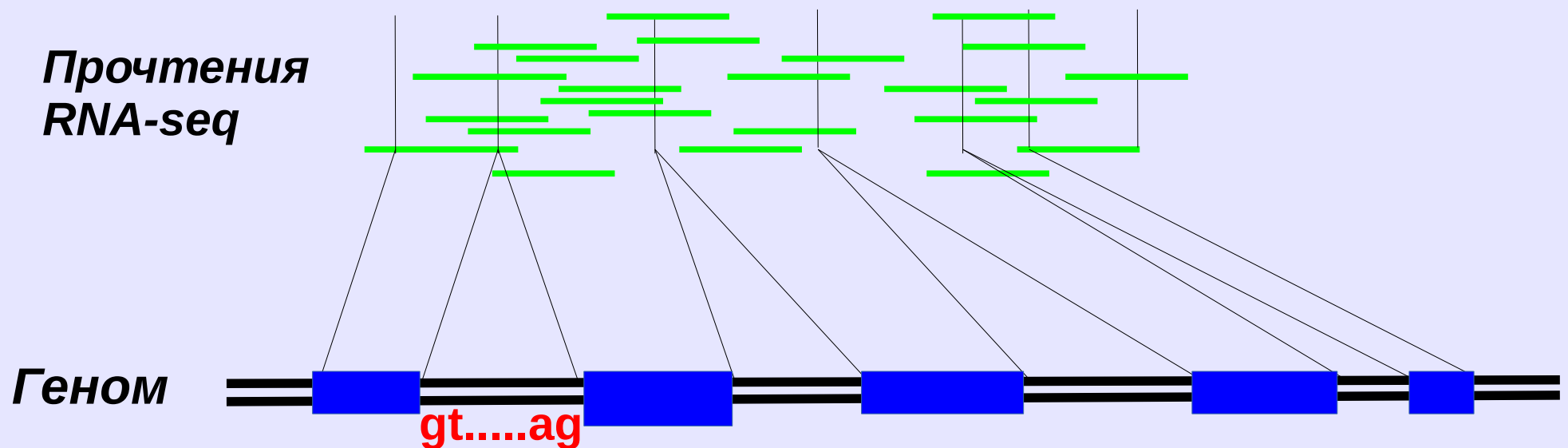
Как это выглядит



The screenshot displays the UCSC Genome Browser interface for the Human Mar. 2006 (NCBI36/hg18) Assembly. The browser is set to chromosome 8 at the genomic coordinates 141,467,287 to 142,163,173 base pairs, which is a region of 695,887 bp. The interface includes a navigation bar at the top with links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, View, Help, and About Us. Below the navigation bar, there are controls for moving and zooming the view, including buttons for 'move', 'zoom in', 'base', and 'zoom out' at various magnifications (1.5x, 3x, 10x, 100x). A search bar is provided for entering genomic positions, gene symbols, or search terms. The main display area shows multiple tracks, including the RefSeq gene track with labels such as TRAPPC2, TRAPPC1, TRAPPC3, TRAPPC4, TRAPPC5, TRAPPC6, TRAPPC7, TRAPPC8, TRAPPC9, TRAPPC10, TRAPPC11, TRAPPC12, TRAPPC13, TRAPPC14, TRAPPC15, TRAPPC16, TRAPPC17, TRAPPC18, TRAPPC19, TRAPPC20, TRAPPC21, TRAPPC22, TRAPPC23, TRAPPC24, TRAPPC25, TRAPPC26, TRAPPC27, TRAPPC28, TRAPPC29, TRAPPC30, TRAPPC31, TRAPPC32, TRAPPC33, TRAPPC34, TRAPPC35, TRAPPC36, TRAPPC37, TRAPPC38, TRAPPC39, TRAPPC40, TRAPPC41, TRAPPC42, TRAPPC43, TRAPPC44, TRAPPC45, TRAPPC46, TRAPPC47, TRAPPC48, TRAPPC49, TRAPPC50, TRAPPC51, TRAPPC52, TRAPPC53, TRAPPC54, TRAPPC55, TRAPPC56, TRAPPC57, TRAPPC58, TRAPPC59, TRAPPC60, TRAPPC61, TRAPPC62, TRAPPC63, TRAPPC64, TRAPPC65, TRAPPC66, TRAPPC67, TRAPPC68, TRAPPC69, TRAPPC70, TRAPPC71, TRAPPC72, TRAPPC73, TRAPPC74, TRAPPC75, TRAPPC76, TRAPPC77, TRAPPC78, TRAPPC79, TRAPPC80, TRAPPC81, TRAPPC82, TRAPPC83, TRAPPC84, TRAPPC85, TRAPPC86, TRAPPC87, TRAPPC88, TRAPPC89, TRAPPC90, TRAPPC91, TRAPPC92, TRAPPC93, TRAPPC94, TRAPPC95, TRAPPC96, TRAPPC97, TRAPPC98, TRAPPC99, TRAPPC100. Below the tracks, there are instructions for using the browser: 'Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.' At the bottom, there are buttons for 'track search', 'default tracks', 'default order', 'hide all', 'add custom tracks', 'track hubs', 'configure', 'reverse', 'resize', 'refresh', 'collapse all', and 'expand all'. The text 'Tracks with lots of items will automatically be displayed in more compact modes.' is also visible.

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)