



Поиск BLAST для нуклеотидных последовательностей

Дарья Владимировна Диброва

к.б.н.,

с.н.с. НИИ ФХБ имени А.Н. Белозерского МГУ имени М.В. Ломоносова,
доцент ФББ МГУ имени М.В. Ломоносова

20 октября 2026 г.

План сегодняшнего занятия

1. «**Кодирующие нуклеотидные последовательности строят козни**»: особенности BLAST для поиска гомологов кодирующих нуклеотидных последовательностей
2. «**Разновидности BLAST** на все случаи жизни»
3. «**NCBI Taxonomy**» и инструменты для визуализации этой базы данных
4. «**NCBI Organelle**»: выборка геномов эукариотических органелл

1. Кодирующие нуклеотидные последовательности строят козны

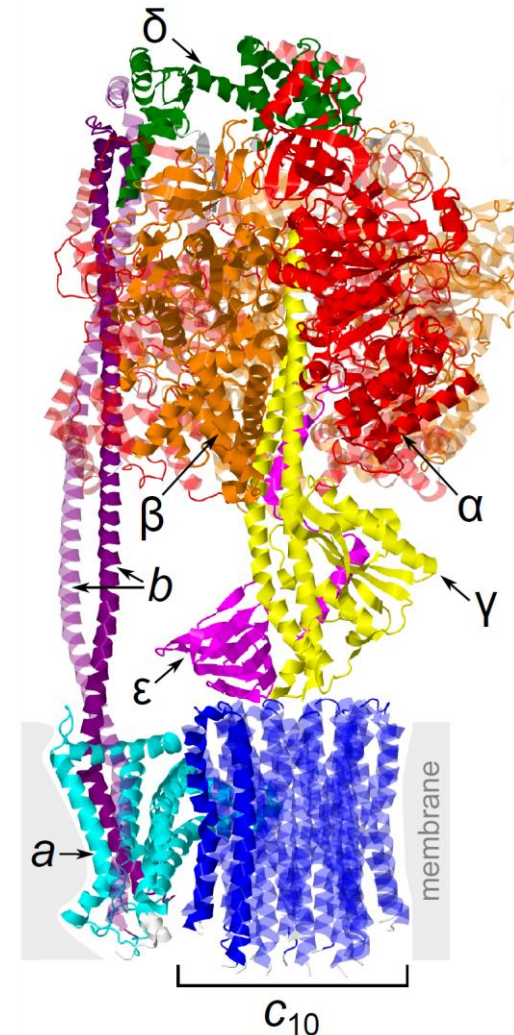
Что может пойти не так при поиске гена белка?..

Дано: *ген* каталитической β -субъединицы роторной мембранной АТФ-синтазы бактерии

Нужно найти: гомологичные последовательности у архей класса Methanomicrobia (по БД RefSeq Protein)

При этом:

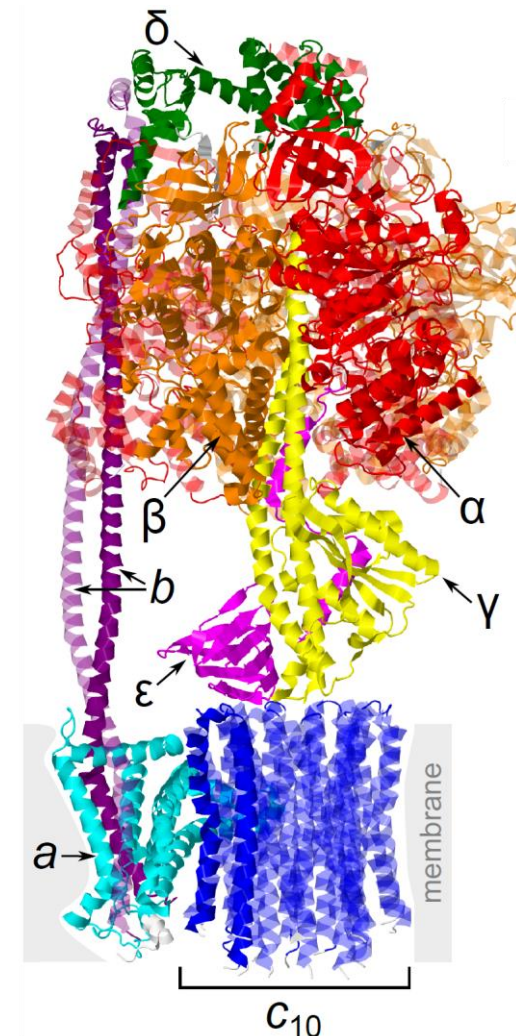
- 1) АТФ-синтаза является универсальным ферментом, она есть почти у всех организмов;
- 2) Разумный поиск **blastp** по последовательности белка (ATPB_ECOLI) дает **638** находок в том же классе.



Печальные результаты поиска

Дано: *ген* каталитической β -субъединицы роторной мембранной АТФ-синтазы бактерии
Нужно найти: гомологичные последовательности у архей класса Methanomicrobia (по БД RefSeq Protein)

При поиске *по гену белка* алгоритмом **blastn** находится **39** (!) находок.
Почему?..

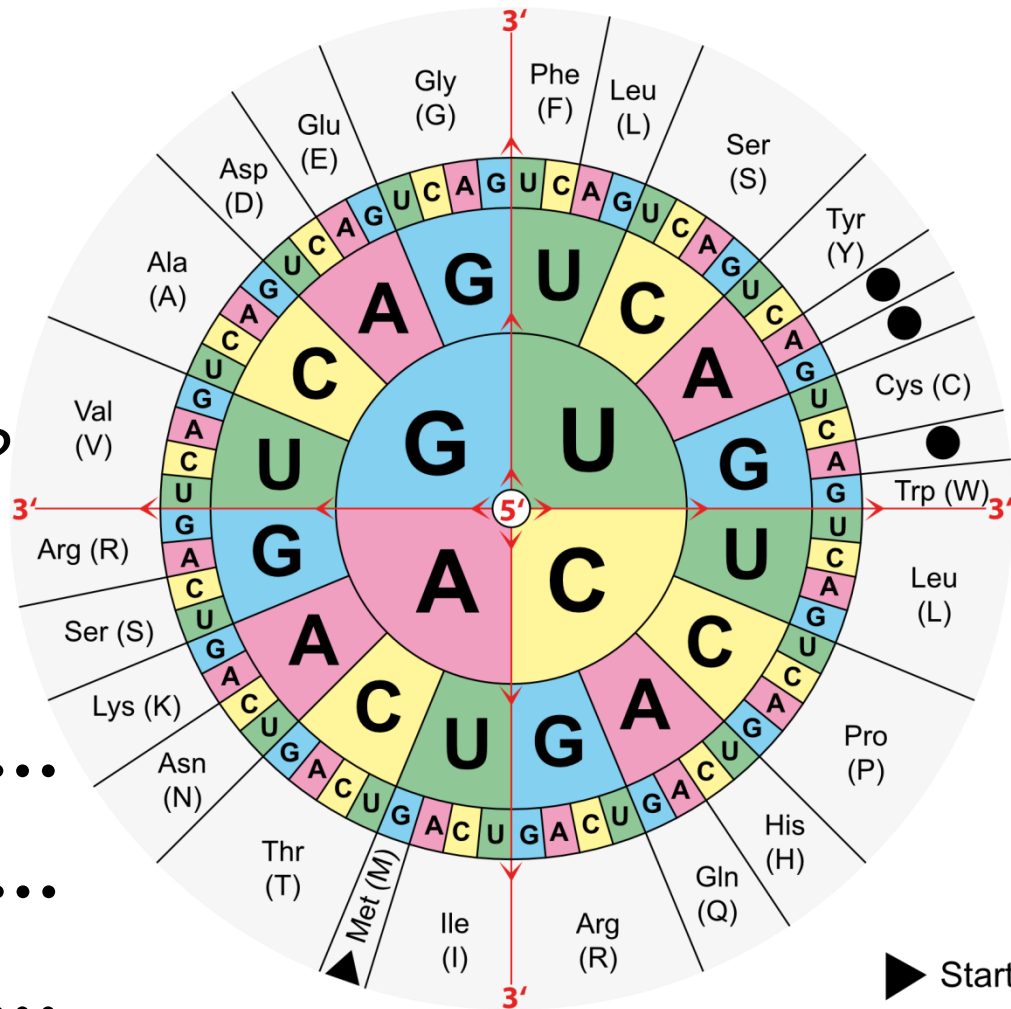


Кодирующие vs. не кодирующие

В чем отличие
кодирующих
нуклеотидных
последовательностей от
не кодирующих —
например, тех же белков?

N ***** **R**

...AATTTGAAGAG...
...AA**C**TGAAGAG...
...AA**C**TGAAG**G**G...



► Start
● Stop

Следствие избыточности* генетического кода

Кодирующие последовательности могут быть переменными в (*почти*) каждой третьей позиции, и это никак не скажется на белке!

Синонимичная замена:

замена, которая **не приводит** к изменению последовательности белка

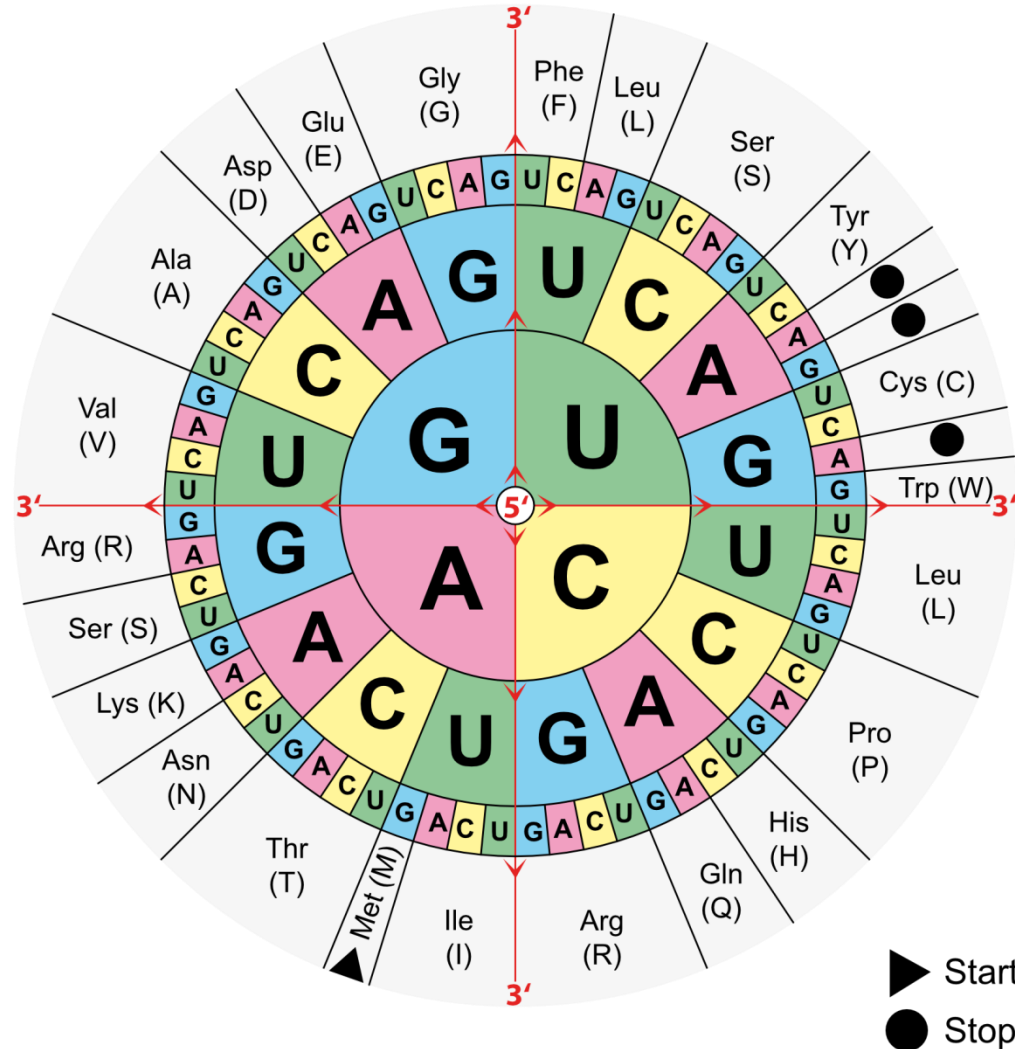
? Синонимичны замены только в третьем нуклеотиде кодона?

* Еще говорят: «вырожденность»

Синонимичные замены бывают самые разнообразные

R

...AATTGAAGA...
...AATTGAAG**G**...
...AATTGA**C**GA...
...AATTGA**CG****G**...
...AATTGA**CG****T**...
...AATTGA**CG****C**...



► Start
● Stop

Но почему BLAST не нашел гомологов?!

Алгоритм BLAST — **эвристический**.

Для ускорения поиска алгоритм BLAST:

- **Подготавливает базу данных**, по которой будет искать (*индексирует* ее), записывая в ней «координаты» всех найденных «слов» определенной длины (*делается 1 раз*)
- **Разбивает на слова определенной длины** последовательность запроса (query) (параметр «word size»)
- При поиске выравнивания **сопоставляет не отдельные буквы, а эти «слова»**, и так находит затравки для будущего выравнивания

Какие ограничения накладывает такой алгоритм?

Пусть длина слова (word size) = **N**

! Если между последовательностью-запросом (query) и последовательностью в базе данных (subject) нет совпадающего участка длиной как минимум **N**, то «зацепки» не будет!

... для нуклеотидного **blastn** размер слова по умолчанию 11, можно сделать (при поиске на сайте) 7 или 15.

Но этого все равно мало, чтобы «покрыть» то, что порождает избыточность генетического кода!

me

Такой хит найдется только с N=7

Methanosarcina mazei SarPi chromosome, complete genome

Sequence ID: [NZ_CP009511.1](#) Length: 4066551 Number of Matches: 1

Range 1: 2234256 to 2234567 [GenBank](#) [Graphics](#)

▼ [Next Match](#) ▲ [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
87.8 bits(96)	6e-15	209/313(67%)	5/313(1%)	Plus/Plus
Query 612	CCTGGTGTATGGCCAGATGAACGAGCCGCCGGGAAACCGTCTGCGCGTTGCTCTGACCGG	671		
Sbjct 2234256	CCTGGTGTGGTCAGATGAATGAACCCCGAGGAGCCGTTTCCGGGTGGGACATACAGC	2234315		
Query 672	TCTGACCATGGCTGAGAAATTCCGTGACGAAG--GTCGTGACGTTCTGCTGTTCTGTTGA	728		
Sbjct 2234316	CCTTACAATGGCAGAGTATTTTCGGGACGAAGAGCATCGAGATGTCTTGCTCCTCATCGA	2234375		
Query 729	CAACATCTATCGTTACACCCTGGCCGGTACGGAAGTATCCGCACTGCTGGGCCGTATGCC	788		
Sbjct 2234376	TAATATTTTCCGGTTCATCCAGGCCGGCTCGGAAGTTTCCGGTCTGATGGGACAGATGCC	2234435		
Query 789	TTCAGCG-GTAGGTTATCAGCCGACCCTGGCGGAAGAGATGGGCGTTCTGCAGGAACGTA	847		
Sbjct 2234436	TTC-GCGCCTGGGCTATCAGCCAACAATGGGTACGGAGTTGTCCGGGTGGAAGAGCGAA	2234494		
Query 848	TCACCTCCACCAAACTGGTTCTATCACCTCCGTACAGGCAGTATACGTACCTGCGGATG	907		
Sbjct 2234495	TTGCCAATACCGATACCGGAGCCATTATGTCAATCCAGGCTGTATACGTGCCGGCTGACG	2234554		
Query 908	ACTTGACTGACCC	920		
Sbjct 2234555	ATTTACGGACCC	2234567		

Поможет ли уменьшение **word_size** до 7?

Параметры для нуклеотидного поиска алгоритмом **blastn**:

- запрос – ген белка **АТРВ_ECOLI** из сборки NC_00913.3;
- база данных **refseq_genomes**;
- фильтр по таксону **Methanomicrobia**;
- максимальное число показываемых находок **5000**;
- размер слова (word_size) **7**.

Ваши ставки: сколько будет находок?

Поможет ли уменьшение **word_size** до 7?

Параметры для нуклеотидного поиска алгоритмом **blastn**:

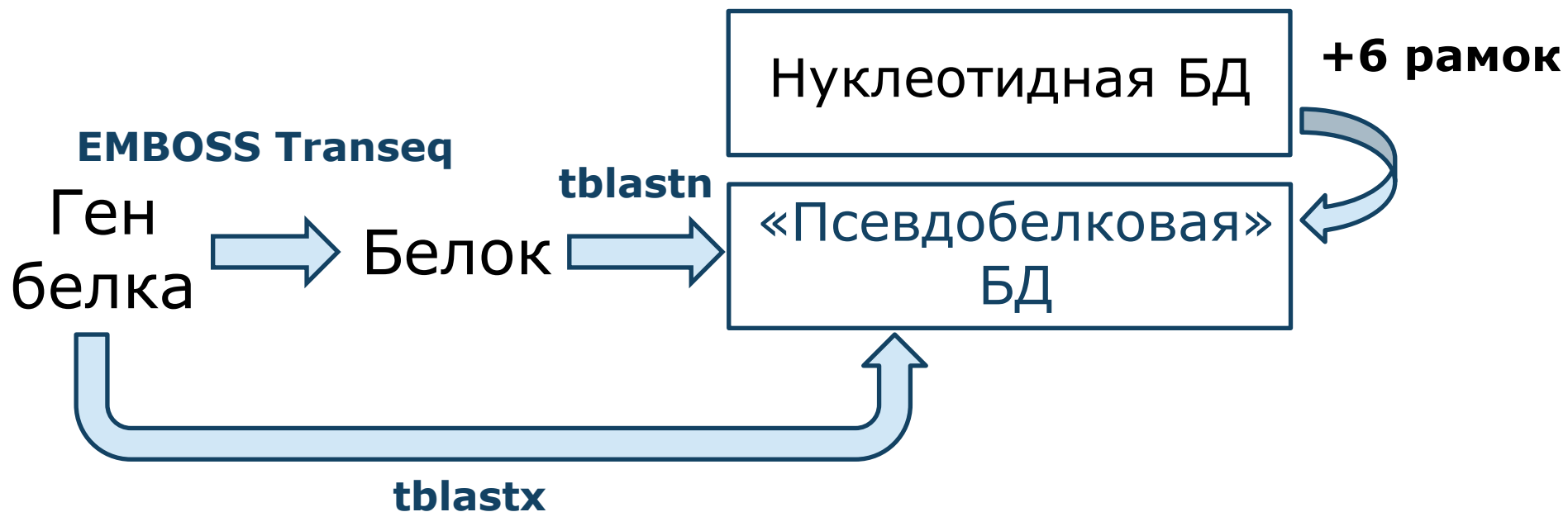
- запрос – ген белка **ATPB_ECOLI** из сборки NC_00913.3;
- база данных **refseq_genomes**;
- фильтр по таксону **Methanomicrobia**;
- максимальное число показываемых находок **5000**;
- размер слова (word_size) **7**.

Ваши ставки: сколько будет находок?

47 vs. 39 ранее

Вывод по поиску кодирующих нуклеотидных последовательностей

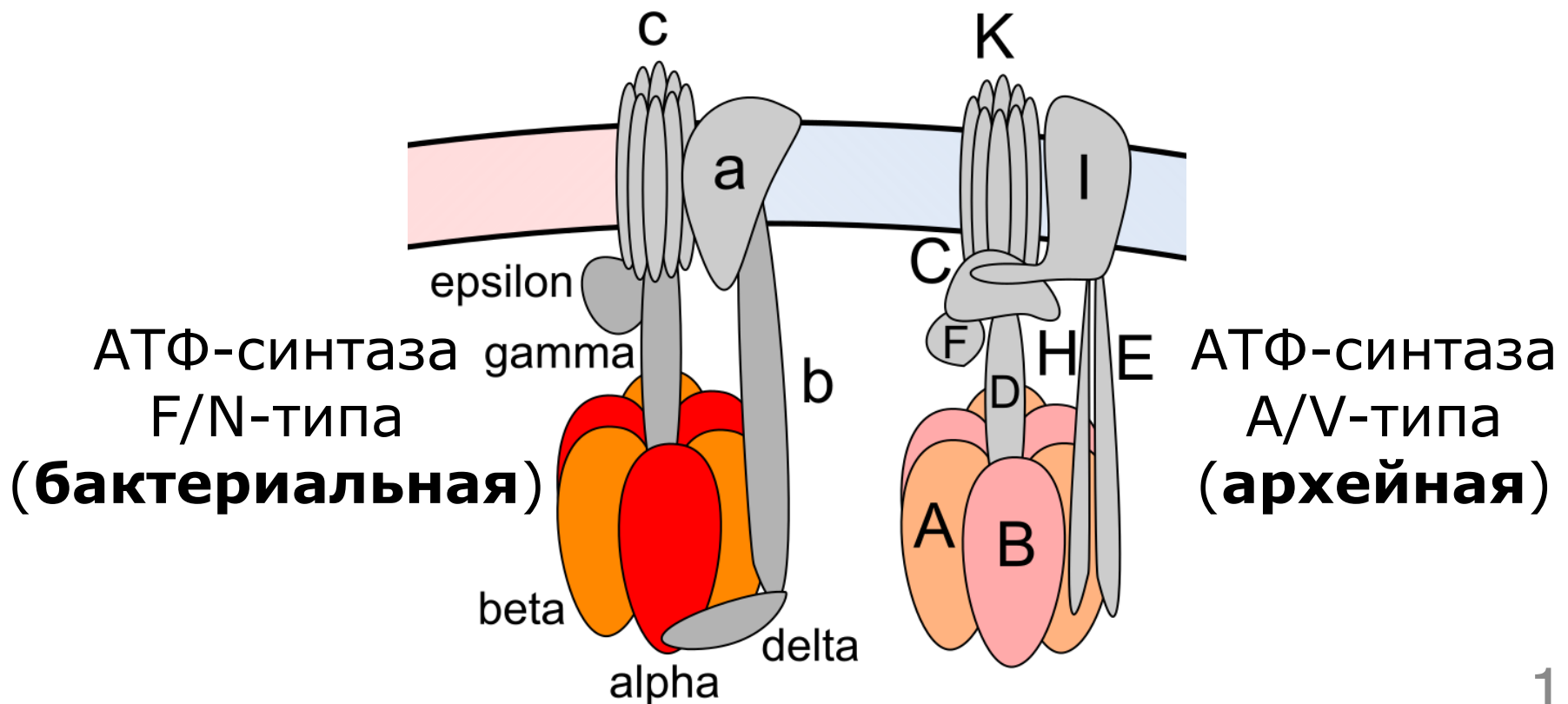
! Для поиска гомологичных кодирующих последовательностей нуклеотидов нужно сперва **транслировать** их



... и все равно все не так просто!

Даже с использованием tblastn таким образом
получается найти только **437** находок
(ср. с **638** находками при поиске по белкам)

ПОЧЕМУ?



Что является находкой при поиске tblastn?

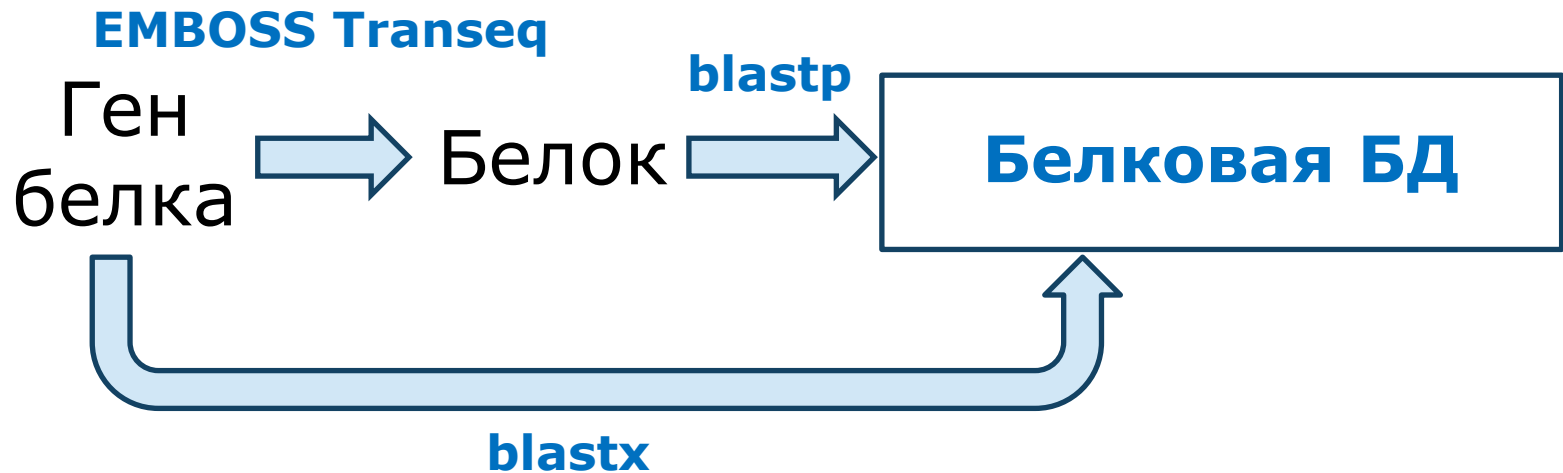
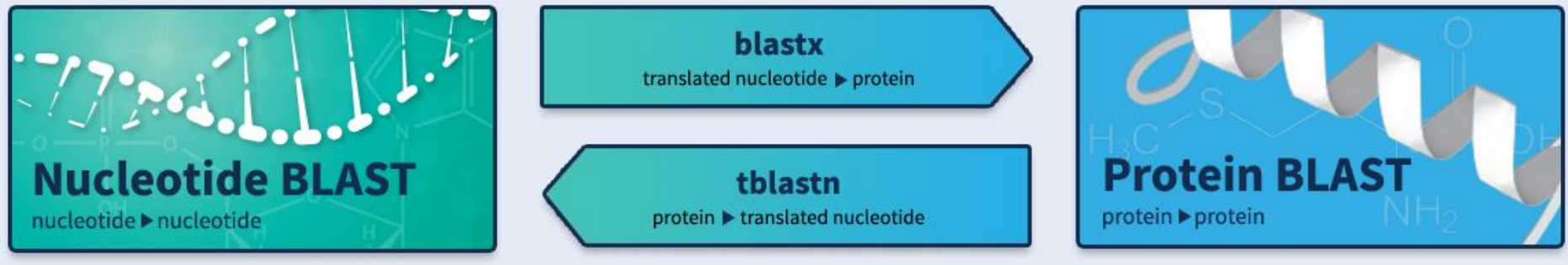
Sequences producing significant alignments				Download	Select columns	Show	5000		
☒ select all 437 sequences selected				GenBank Graphics					
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☒	Candidatus Methanocrinis alkalitolerans strain M04Ac NODE_47...	Candidatu...	442	555	97%	8e-138	49.45%	14355	NZ_JARFPL010000047.1
☒	Methanotherix sp. isolate SO_2017_LW1 bin 36 NODE_2276 leng...	Methanot...	436	538	98%	6e-136	47.80%	18658	NZ_JALNZN010000035.1
☒	Methanosarcina mazei strain 3.H.A.2.4 scaffold38_1_size35990-r...	Methanos...	435	536	99%	1e-135	47.82%	35910	NZ_JJPP010000069.1
☒	Methanosarcina mazei SarPi chromosome, complete genome	Methanos...	435	726	99%	1e-135	47.82%	4066551	NZ_CP009511.1
☒	Methanosarcina barkeri 3 chromosome, complete genome	Methanos...	430	614	98%	1e-133	47.03%	4560446	NZ_CP009517.1
☒	Candidatus Methanoperedens sp. BLZ2 CMPBLZ2contig_254, w...	Candidatu...	428	535	97%	4e-133	46.78%	52908	NZ_NTMG010000069.1
☒	Methanosarcina sp. UBA5 UBA5_contig_1012, whole genome sh...	Methanos...	427	537	98%	9e-133	46.70%	68525	NZ_DAVG010000019.1
☒	Methanosarcina sp. Z-7115 59, whole genome shotgun sequence	Methanos...	426	539	98%	2e-132	46.59%	21280	NZ_JAVKPK010000059.1
☒	Methanotherix sp. isolate AMR_MDS_2986 NODE_3025 length_2...	Methanot...	426	426	97%	4e-132	48.12%	20344	NZ_DAOTJZ010000035.1
☒	Methanotherix sp. isolate Seed_metabat.86 k141_1036892, whole...	Methanot...	426	544	97%	4e-132	48.12%	32858	NZ_JARKUH010000105.1
☒	Methanosarcina sp. isolate I5m-25 NODE_336, whole genome sh...	Methanos...	425	531	98%	4e-132	47.03%	71528	NZ_JAWQJL010000027.1
☒	Methanosarcina sp. isolate I3m-35 NODE_158, whole genome sh...	Methanos...	425	531	98%	4e-132	47.03%	71540	NZ_JAWQJK010000026.1

Находки **tblastn** будут нуклеотидными записями, многие из которых могут содержать несколько совпадений с запросом!

... но есть еще кое-что – длина слова важна и для **blastp**!

- Длина слова по умолчанию в **tblastn** – 5 (как и для **blastp**).
- Но ее можно уменьшить до 2, что увеличит время поиска, но позволит «вычерпать» больше находок.
- В локальной версии **blastp** есть еще влияющие на чувствительность (хорошо) и производительность (плохо) поиска опции:
 - `-window_size`
 - `-threshold`
- Если уменьшить длину слова **tblastn** до 2, при нашем запросе найдется **469** нуклеотидных записей (+32 записи, почти +10%)

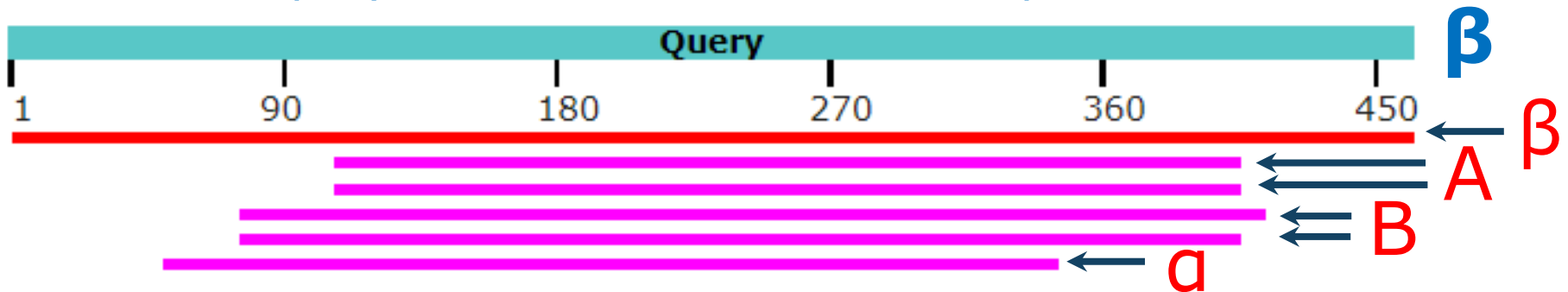
Сайт BLAST на заглавной странице дает подсказки



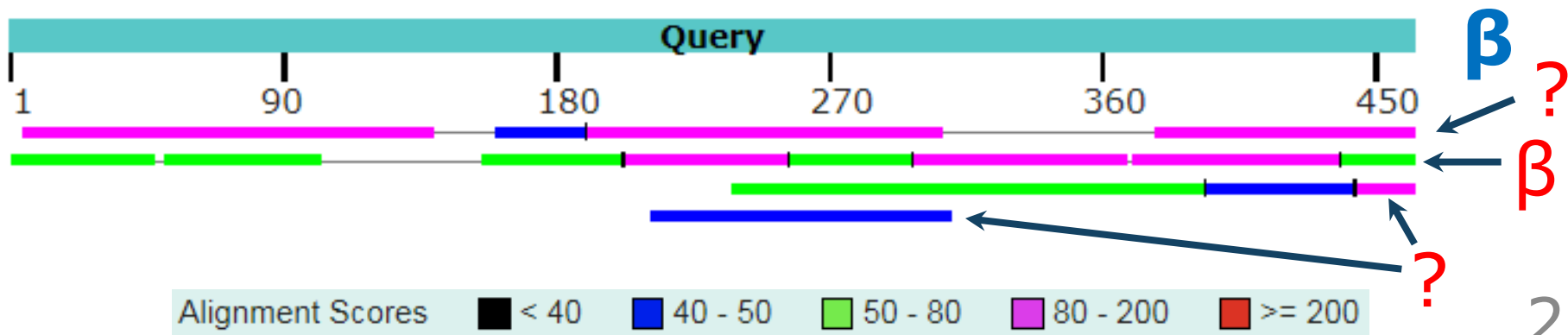
Поиск по прокариотическому белку в геномах эукариот через tblastn

? В чем особенность эукариотических генов белков по сравнению с прокариотическими?

Поиск по β -субъединице АТФ-синтазы среди белков мыши:



Поиск по β -субъединице АТФ-синтазы по геному мыши:



Вывод по поиску кодирующих нуклеотидных последовательностей

! Для поиска гомологичных кодирующих последовательностей нуклеотидов нужно сперва **транслировать** их...

... и без нужды не искать по нуклеотидной базе данных у экариот, когда можно поискать по белковой

2. Разновидности BLAST на все случаи жизни

Классификация алгоритмов BLAST

- ① **Nucleotide - nucleotide**
blastn
megablast (быстрый поиск почти идентичных последовательностей)
- ② **Protein - protein**
blastx *
blastp
- ③ **Protein – translated nucleotide**
tblastx *
tblastn

* На вход подается нуклеотидная последовательность и параметры для трансляции

Задачи и алгоритмы BLAST для их решения

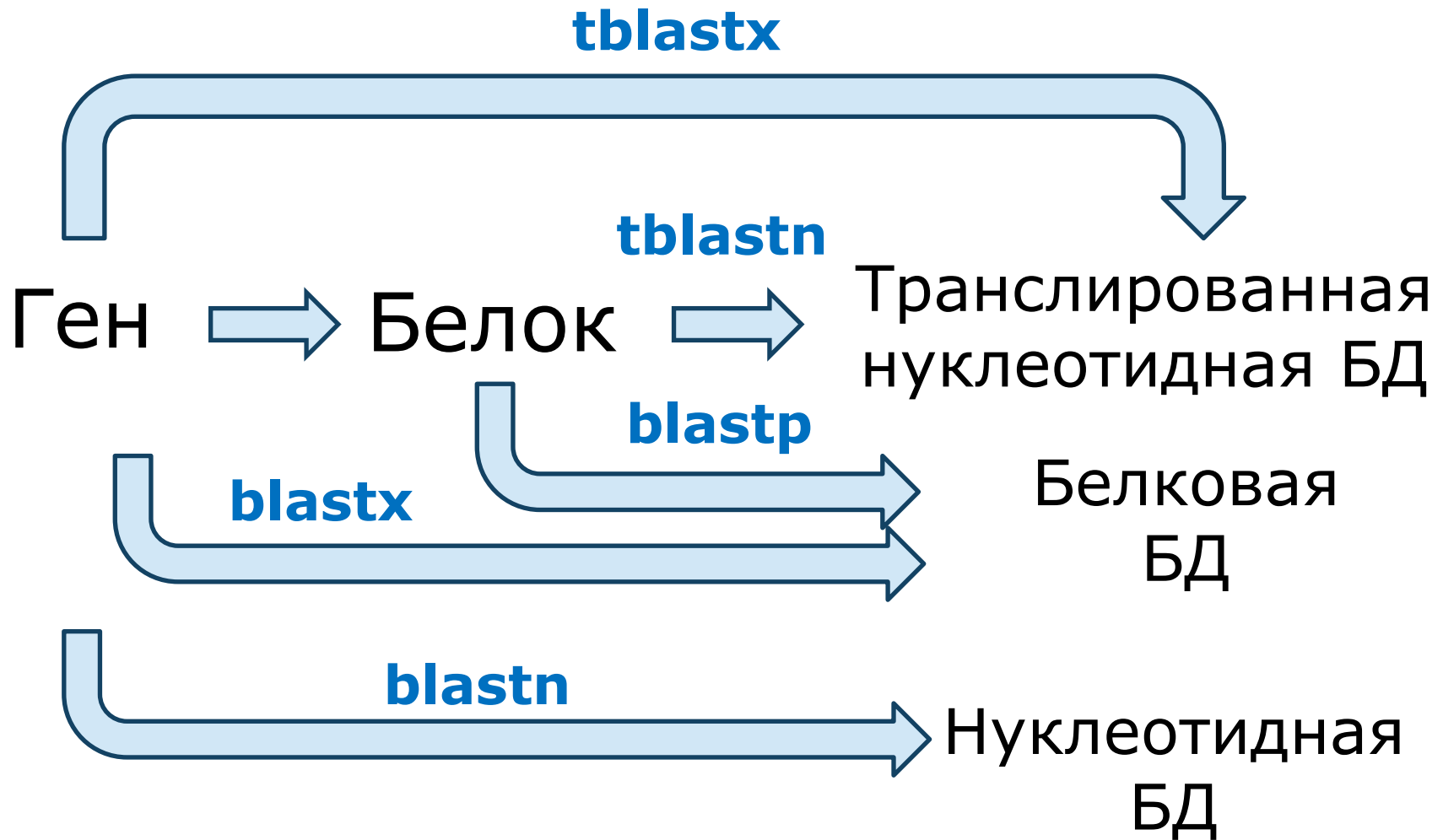
blastn
megablast

blastx
blastp

tblastx
tblastn

1. К какому виду относится 16S рРНК, которую мы секвенировали «из капли воды»?  **blastn**
2. Геном организма известен, нужно картировать на него фрагмент  **megablast**
3. Есть ген белка, нужно найти какие мутации бывают в некоторой позиции  **blastx**
4. Есть белок, а есть ли гомолог в сборке генома, где не предсказаны гены?  **tblastn**
5. Есть белок, есть сборка с предсказанными белками, как найти гомологи?  **blastp**

Типы BLAST еще раз в виде схемы



Часто нужен локальный запуск BLAST «под себя»

Вы можете запускать BLAST локально, на своем компьютере (или на kodomo)

См. инструкцию по работе с разными версиями BLAST:

<https://kodomo.fbb.msu.ru/wiki/Main/blast>

1) Справка с подробным комментарием

`blastn -help`

`tblastn -help`

`blastx -help`

2) Важные опции:

`-outfmt`: формат выдачи, подробный vs. таблица

`-max_target_seqs`: максимальное число находок (!)

НЕ ОБЯЗАТЕЛЬНО ЛУЧШИХ! См.:

<https://academic.oup.com/bioinformatics/article/35/9/1613/5106166>

`-word_size`: размер слова (см. слайды 8–11)

Выдача blastn в табличном виде

```
# BLASTN 2.12.0+
# Query: <...>
# Database: GCF_937001465.1_mOrcOrc1.1_genomic.fna
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens,
q. start, q. end, s. start, s. end, evalue, bit score
# 11 hits found
```

query	NC_064572.1	73.684	114	27	2	887	998	2356223	2356335	1.65e-07	65.3
query	NC_064572.1	73.684	114	27	2	887	998	2395418	2395530	1.65e-07	65.3
query	NC_064572.1	73.684	114	27	2	887	998	3426231	3426119	1.65e-07	65.3
query	NC_064572.1	73.684	114	27	2	887	998	3457752	3457640	1.65e-07	65.3
query	NC_064572.1	73.684	114	27	2	887	998	3489302	3489190	1.65e-07	65.3
query	NC_064572.1	88.372	43	5	0	1494	1536	2356881	2356923	8.55e-05	56.3
query	NC_064572.1	88.372	43	5	0	1494	1536	2396076	2396118	8.55e-05	56.3
query	NC_064572.1	88.372	43	5	0	1494	1536	3425573	3425531	8.55e-05	56.3
query	NC_064572.1	88.372	43	5	0	1494	1536	3457094	3457052	8.55e-05	56.3
query	NC_064572.1	88.372	43	5	0	1494	1536	3488644	3488602	8.55e-05	56.3
query	NC_064565.1	72.807	114	28	2	887	998	8267821	8267709	2.01e-06	60.8

query subject
(запрос) (где
нашлось)

%id

alignment
length

координаты
в query
(начало и
конец)

координаты
в subject
(начало и
конец)

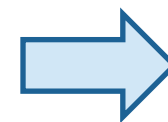
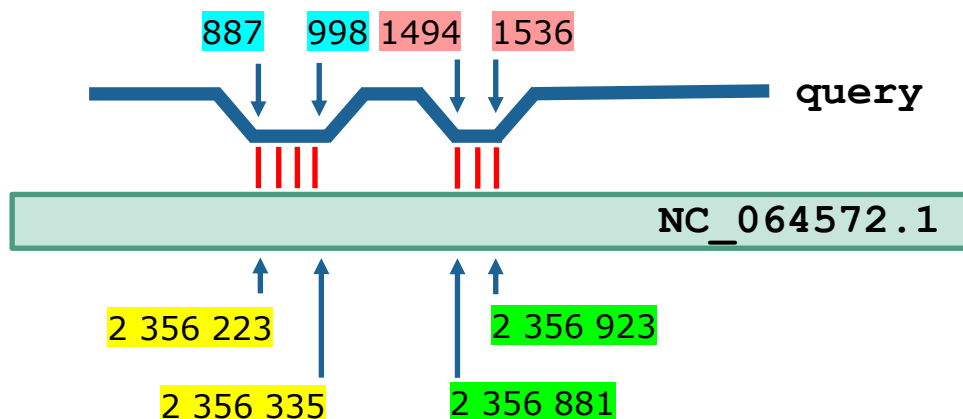
evalue,
score

Интерпретация выдачи blastn

```
# BLASTN 2.12.0+
# Query: <...>
# Database: GCF_937001465.1_mOrcOrc1.1_genomic.fna
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens,
# q. start, q. end, s. start, s. end, evalue, bit score
# 11 hits found
```

query	NC_064572.1	73.684	114	27	2	887	998	2356223	2356335	1.65e-07	65.3
<...>											
query	NC_064572.1	88.372	43	5	0	1494	1536	2356881	2356923	8.55e-05	56.3
<...>											

query (запрос)	subject (где нашлось)	%id	alignment length	координаты в query (начало и конец)	координаты в subject (начало и конец)	evalue, score
-------------------	-----------------------------	-----	---------------------	--	--	------------------



В данном случае
две находки
blastn
соответствуют
одному гомологу в
геноме

3. NCBI Taxonomy

NCBI Taxonomy: старый интерфейс меняется

**National Library of Medicine**
National Center for Biotechnology Information

Log in

 **Taxonomy moving to Datasets**

Visit the new [NCBI Taxonomy homepage](#) and explore the redesigned Taxonomy Browser, Common Tree, and more. The new homepage is planned to replace this legacy page in October 2026.

Taxonomy

Taxonomy

Search

Limits Advanced

Help



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

[Quick Start Guide](#)

[FAQ](#)

[Handbook](#)

[Taxonomy FTP](#)

[Important Update: Phyla Changing](#)

[Important Update: New Flu species Names](#)

[Important Update: New ranks replaced superkingdom](#)

Taxonomy Tools

[Browser](#)

[Common Tree](#)

[Statistics](#)

[Name/ID Status](#)

[Genetic Codes](#)

[Linking to Taxonomy](#)

[Extinct Organisms](#)

Other Resources

[GenBank](#)

[LinkOut](#)

[E-Utilities](#)

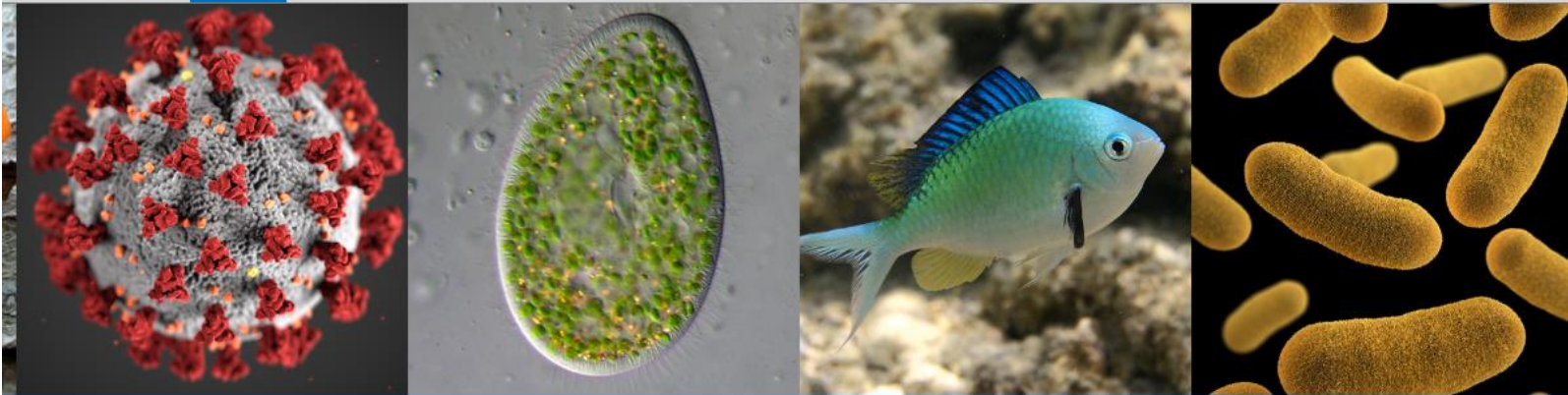
[Batch Entrez](#)

[INSDC](#)

Disclaimer: The NCBI taxonomy database is not an authoritative source for nomenclature or classification
- please consult the relevant scientific literature for the most reliable information.

NCBI Taxonomy: старый интерфейс меняется

NCBI Datasets Taxonomy Genome Gene Command-line tools Documentation



Taxonomy

NCBI Taxonomy is a curated classification system that organizes data in the public sequence databases and provides standardized organism names and lineage information.

Explore our taxonomy resources



Taxonomy browser

[Explore the taxonomic hierarchy](#)



Taxonomy status check

[Verify the current status of a taxon](#)



Common tree

[View a tree from selected taxa](#)



BioCollections

[Explore data on natural history collections](#)

Интерфейс Taxonomy check status

Taxonomy name / ID status

Enter taxonomic names and/or TaxIDs to check their status, one per line (maximum of 1,000). [Upload a list](#).

Example

Escherichia coli
Escherichia coli
Bacillus subtilis

Check status



Taxonomy Name / ID Status

← Back

Download

☐

Query

Current scientific name

TaxID

Rank

Group name

☐

Escherichia coli

☐

Bacillus subtilis

[Bacillus subtilis](#)

1423

species

firmicutes

☐

Escherichia coli

[Escherichia coli](#)

562

species

enterobacteria

По названию вида (или любого таксона) можно получить **выписку** из NCBI

Тахоному, содержащую в том числе:

1. Идентификатор таксономии (**taxid**);
2. Список **taxid** и названий для всех вышестоящих таксонов.

Аналогично, поиск можно делать по **taxid**.

Изменения таксономии

- Таксономия подвержена регулярному пересмотру:
 - Регулярно изменяются и уточняются названия конкретных видов;
 - Бывает что меняются даже названия более «глобальных» групп: из последнего – разделение и переименование группы протеобактерий.
- Нужно проверять периодически, не поменялась ли общепринятая таксономия!

Proteobacteria:

- α -proteobacteria
- β -proteobacteria
- γ -proteobacteria
- ...
- δ -proteobacteria
- ϵ -proteobacteria



Pseudomonadota:

- α -proteobacteria
- β -proteobacteria
- γ -proteobacteria
- ...



Mycococcota

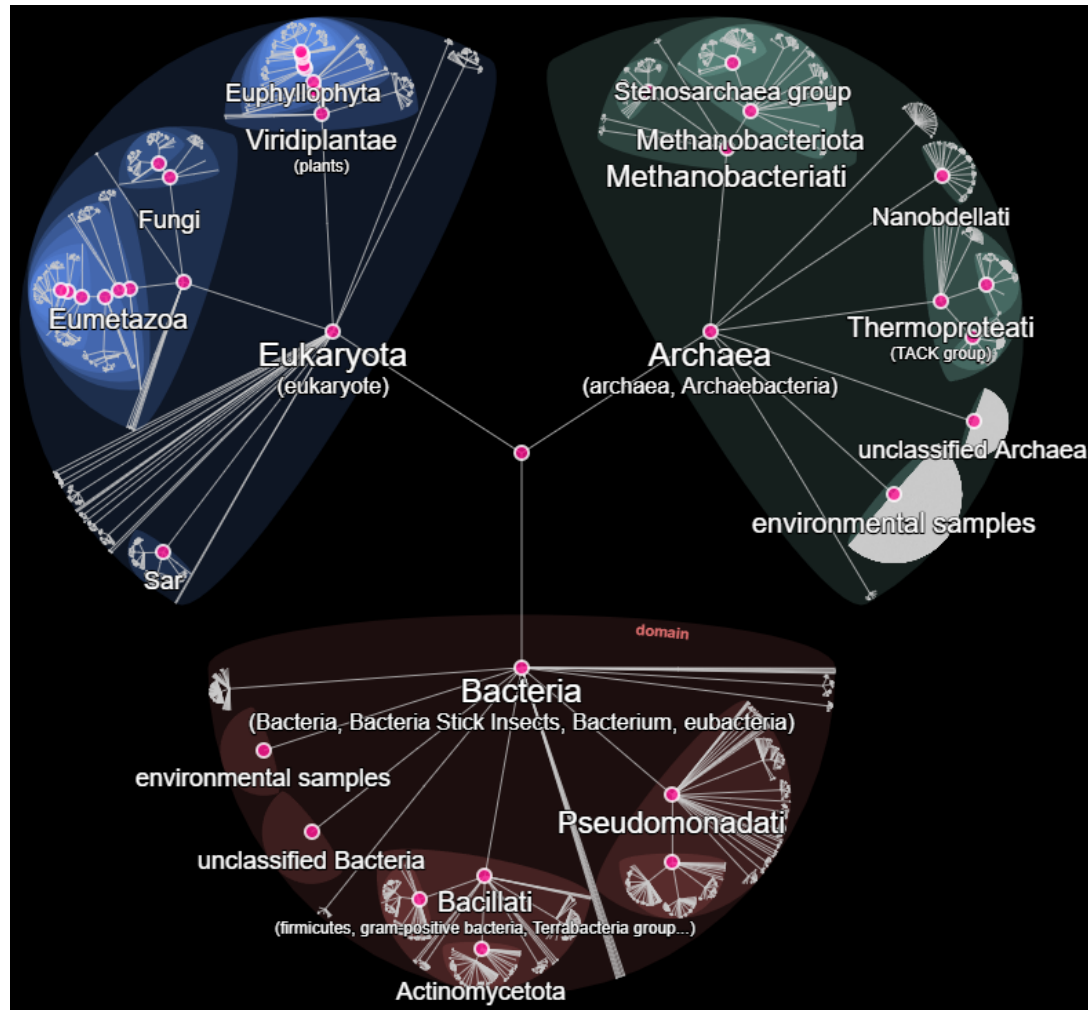
...



Campylobacterota

Lifemap: удобная интерактивная визуализация NCBI Taxonomy

<https://lifemap.cnrs.fr/tree?efficiency-mode=false>

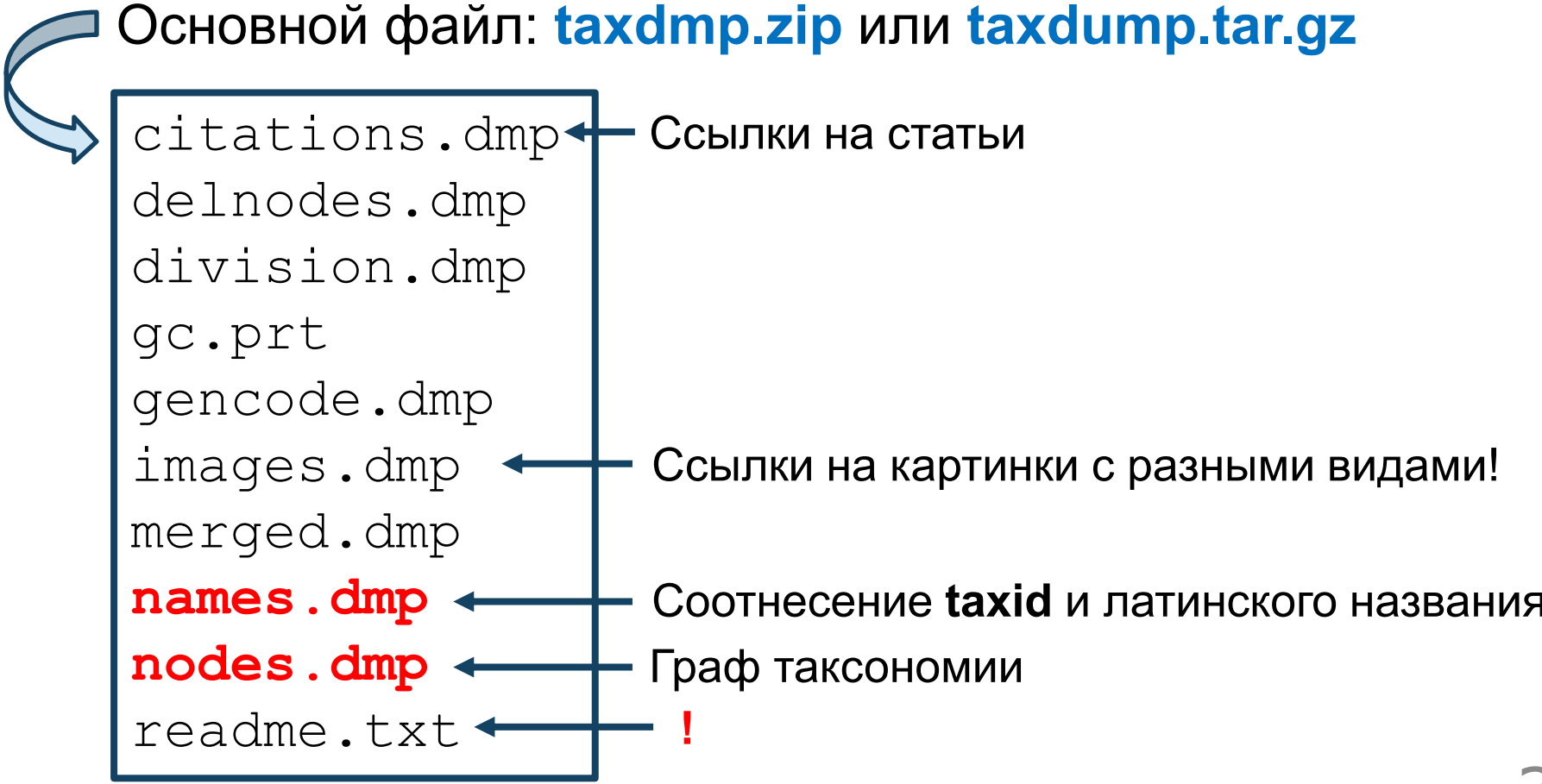


NCBI Taxonomy: доступ через FTP

К этим данным можно получить доступ через FTP:

<https://ftp.ncbi.nih.gov/pub/taxonomy/>

Основной файл: **taxdmp.zip** или **taxdump.tar.gz**



<code>citations.dmp</code>	← Ссылки на статьи
<code>delnodes.dmp</code>	
<code>division.dmp</code>	
<code>gc.prt</code>	
<code>gencode.dmp</code>	
<code>images.dmp</code>	← Ссылки на картинки с разными видами!
<code>merged.dmp</code>	
<code>names.dmp</code>	← Соотнесение taxid и латинского названия
<code>nodes.dmp</code>	← Граф таксономии
<code>readme.txt</code>	← !

4. NCBI Organelle

NCBI Organelle: сборки органелл

<https://www.ncbi.nlm.nih.gov/datasets/organelle/>

Organelle

Selected taxa

Filters

TYPES

☐ Mitochondrion

☐ Chloroplast

☐ Plastid

☐ Apicoplast

☐ Cyanelle

☐ Chromatophore

☐ Kinetoplast

YEAR RELEASED

1980

2026

Число аннотированных CDS в пластидах и хлоропластах (2025 г.)

Данные взяты из:
нуклеотидные
сборки (**всего 15 219**
шт.) из БД NCBI
Organelle
Изображения:
<https://www.wikipedia.org/>

