



# Поиск в Nucleotide

# NCBI

National Center for Biotechnology Information создан в 1988 году как подразделение National Library of Medicine (NLM) в National Institutes of Health (NIH).

All Resources:

- ▶ Databases (> 50)
- ▶ Downloads
- ▶ Submissions
- ▶ Tools
- ▶ How to

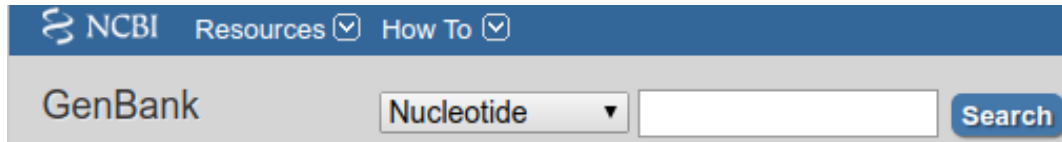
# Nucleotide Database

Сводная база данных нуклеотидных последовательностей и их аннотаций из различных источников:

- ▶ GenBank — аннотированная коллекция всех опубликованных последовательностей ДНК
- ▶ RefSeq — невырожденный набор геномных ДНК, а также транскриптов РНК
- ▶ the Third Party Annotation (TPA) — последовательности, полученные на основании данных GenBank
- ▶ PDB (Protein Data Bank)

# База данных Nucleotide

Перейдите на страницу [Nucleotide](#)



The image shows the top navigation bar of the NCBI website. It includes the NCBI logo, a 'Resources' dropdown menu, and a 'How To' dropdown menu. Below this is a search bar with the 'GenBank' label, a dropdown menu set to 'Nucleotide', an empty text input field, and a blue 'Search' button.

Возможности для поиска

- ▶ Традиционный поиск по ключевым словам
- ▶ Автоматизированный с помощью NCBI e-utilities
- ▶ Поиск конкретных последовательностей (BLAST)
- ▶ Поиск непосредственно на FTP сервере

# Обзор результатов поиска

Сделайте поиск по слову *spike*

The screenshot displays the NCBI Gene database search results for the query 'spike'. The interface is organized into several sections, each highlighted with a numbered purple dashed box:

- Box 1:** The main results list. It shows three entries for 'spike' genes, each with a checkbox, a link to the gene name (e.g., [CCV-C54 spike gene](#)), and details such as '4,435 bp linear DNA', accession number 'A22886.1', and GI number '1249584'. Links for 'Protein', 'Taxonomy', 'GenBank', 'FASTA', and 'Graphics' are provided for each entry.
- Box 2:** The top navigation and filtering area. It includes dropdown menus for 'Summary', '20 per page', and 'Sort by Default order'. A 'Send to:' dropdown and a 'Filters: Manage Filters' link are also present.
- Box 3:** The left sidebar containing filters for 'Species' (Animals, Plants, Fungi, etc.), 'Molecule types' (genomic DNA/RNA, mRNA, etc.), 'Source databases' (INSDC, RefSeq, etc.), 'Sequence Type' (Nucleotide, EST, GSS, etc.), and 'Genetic compartments' (Chloroplast, Mitochondrion, etc.).
- Box 4:** The 'Results by taxon' section. It lists 'Top Organisms' with links to taxonomic trees, including *Triticum aestivum*, *Hordeum vulgare*, *Setaria italica*, *Avian coronavirus*, *Rabies lyssavirus*, and 'All other taxa'.
- Box 5:** The 'Search details' section at the bottom right. It shows the search query 'spike[All Fields]' and a 'Search' button, along with a 'See more...' link.

1. Список последовательностей
2. Настройки отображения
3. Фильтры
4. Экспорт последовательностей
5. Поисковый запрос

# Настройки отображения

1. Summary
2. GenBank и GenBank (full)
3. FASTA и FASTA (text)
4. ASN.1 (Abstract Syntax Notation One)
5. Revision History
6. Accession List
7. GI List

ACCESSION AF000001

VERSION AF000001.5 GI: 7274584

# Разнообразие форматов FASTA

**Расширения:** .fa, .fas, .fasta, .fna, .ffn, .faa, .frn, .afa, .mfa, но не fastq.

**Варианты заголовков:**

```
>id ...  
>lcl|id| ... # local  
>gb|id| ... # GenBank
```

**Точки и дефисы**

```
>sequence1  
AGATACACA  
>sequence2  
.C...-.T-
```

# Символы ИЮПАК для обозначения нуклеотидов

	Обозначение	Расшифровка
A	A	Adenine
T	T	Thymine
G	G	Guanine
C	C	Cytosine
U	U	Uracil
Y	C или T	pYrimidine
R	A или G	puRine
W	A или T	Weak
S	G или C	Strong
K	T или G	Keto
M	A или C	aMino

	Обозначение	Расшифровка
D	не C	следующая буква
V	не T	следующая "свободная" буква
H	не G	следующая буква
B	не A	следующая буква
N	любой	Nucleotide
X	неизвестный	



# Аминокислоты и их обозначения

Alanine	Ala	A
αRginine	Arg	R
asparagiNe	Asp	N
aspartic acid	Asp	D
Cysteine	Cys	C
glutamine	Gln	Q
glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I

Leucine	Leu	L
lysine	Lys	K
Methionine	Met	M
phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
selenocysteine	Se-Cys	U
tryptophan	Trp	W
tYrosine	Tyr	Y

Иногда Asp/Asn обозначают Asx (B), а Glu или Gln - Glx (Z).

# Формат файлов GenBank

## Полное описание формата GenBank

```
LOCUS      A22886                      4435 bp    DNA      linear    PAT 23-JUN-1995
DEFINITION CCV-C54 spike gene.
ACCESSION  A22886
VERSION    A22886.1
KEYWORDS   .
SOURCE     Canine coronavirus
ORGANISM   Canine coronavirus
Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
Alphacoronavirus; Tegacovirus.
REFERENCE  1  (bases 1 to 4435)
AUTHORS    Brown,T.D.K. and Horsburgh,B.C.
TITLE      Canine coronavirus subunit vaccine
JOURNAL    Patent: EP 0510773-A1 5 28-OCT-1992;
AKZO N.V.; Akzo Nobel N.V
```

# Features файлов GenBank

```
FEATURES             Location/Qualifiers
source               1..4435
                    /organism="Canine coronavirus"
                    /mol_type="unassigned DNA"
                    /strain="CCV-V54"
                    /db_xref="taxon:11153"
gene                 60..4421
                    /gene="spike"
CDS                  60..4421
                    /gene="spike"
                    /codon_start=1
                    /protein_id="CAA01637.1"
                    /translation="MIVLTLCLLLFSYNSVICTSNNDCVQVNVTLQPGNENIIKDFLF
...
ORIGIN
1  ttgctcatta gaaacaatgg aaaactacta aacttcggta atcacttggt taatgtgcc
61 tgattgtgct tacattgtgc cttctcttgt tttcatacaa tagtgtgatt tgtacatcaa
...
```

# Примеры features файлов GenBank

- ▶ Источник — `source`
- ▶ Кодировующие последовательности — `CDS`
- ▶ мРНК — `mRNA`
- ▶ Экзоны и интроны — `exon` и `intron`
- ▶ Ген — `gene`
- ▶ Сайты связывания — `misc_binding`
- ▶ Полный список в [Appendix II](#) описания формата GenBank

**Задание 1.** Какой участок НК принято обозначать `stem_loop` в файлах GenBank?

# Обозначение границ участков

- ▶ 1 — первая позиция в последовательности
- ▶ 1..99 — непрерывный участок с 1 по 99 позицию
- ▶ `join(1..99, 101..112)` — несколько участков, образующих единую последовательность
- ▶ `complement(1..99)` — участок в комплементарной цепи
- ▶ `<1..99` — начало участка расположено до указанной позиции, но точная граница не известна
- ▶ `101..>112` — аналогично предыдущему, но не известен конец участка
- ▶ `J00194.1:100..202` — ссылка на участок с 100 по 202 позицию как на отдельную последовательность в базе данных

# Скачивание файлов

Для скачивания файла нужно воспользоваться опцией «Send to:», ее параметры:

- ▶ Complete Record / Coding sequence
- ▶ File / Clipboard / Collection / Analysis Tool
- ▶ Format (при выборе Send to: File)

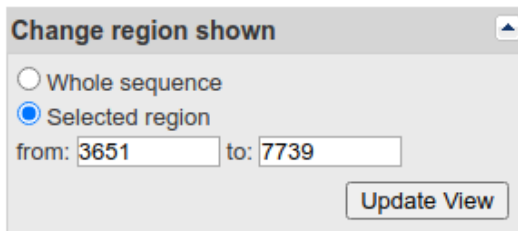
**Задание 2.** Найдите последовательность с идентификатором HE616736.1. Скачайте ее в форматах fasta и GenBank, какой размер у этих файлов? Какому организму принадлежит эта последовательность? Укажите границы кодирующей последовательности.

# Извлечение файлов GenBank частями

**Задание 3.** Сколько генов есть в последовательности с идентификатором AY771998.1?

# Извлечение файлов GenBank частями

**Задание 3.** Сколько генов есть в последовательности с идентификатором AY771998.1?



Change region shown

☐ Whole sequence

☒ Selected region

from: 3651 to: 7739

Update View

В результате будет отображён только нужный участок.

```
LOCUS      AY771998      4089 bp      RNA      linear      VRL 22-FEB-2005
...
ACCESSION  AY771998 REGION: 3651..7739
...
```



# Фильтры

Набор по умолчанию:

- ▶ Species
- ▶ Molecular types
- ▶ Source databases
- ▶ Sequence Type
- ▶ Genetic compartments
- ▶ Sequence length
- ▶ Release date
- ▶ Revision date
- ▶ Search fields

Опция [Show additional filters](#) позволяет добавлять или удалять фильтры.

Опция [Customize](#) — настраивать фильтры

## Species

Animals (1,410)

Plants (642,155)

Fungi (204)

Protists (74)

Bacteria (6,972)

Archaea (210)

Viruses (36,715)

[Customize ...](#)

## Molecule types

genomic DNA/RNA (44,881)

mRNA (634,488)

[Customize ...](#)

## Source databases

INSDC (GenBank) (688,386)

RefSeq (1,276)

[Customize ...](#)

## Sequence Type

Nucleotide (439,585)

EST (250,085)

GSS (2)

## Genetic

compartments

Chloroplast (11)

Mitochondrion (93)

Plasmid (11)

Plastid (14)

## Sequence length

[Custom range...](#)

## Release date

[Custom range...](#)

## Revision date

[Custom range...](#)

## Search fields

[Choose ...](#)

[Clear all](#)

[Show additional filters](#)

# Использование фильтров

**Задание 4.** Сколько последовательностей останется в списке, полученном по запросу *spike*, если выставить следующие фильтры:

1. вирусные последовательности;
2. геномная ДНК или РНК;
3. база данных GenBank.

# Использование фильтров

**Задание 4.** Сколько последовательностей останется в списке, полученном по запросу *spike*, если выставить следующие фильтры:

1. вирусные последовательности;
2. геномная ДНК или РНК;
3. база данных GenBank.

Вместе с настройкой фильтров меняется поисковый запрос (Search details):


```
spike[All Fields] AND (viruses[filter]  
AND biomol_genomic[PROP]  
AND ddbj_embl_genbank[filter])
```

# Расширенный поиск

Nucleotide ▼ spike × Search

[Create alert](#) [Advanced](#)

## Nucleotide Advanced Search Builder

 Filters activated: Viruses, genomic DNA/RNA, INSDC (GenBank), Sequence length from 1000 to 2000. [Clear all](#)

Use the builder below to create your search

[Edit](#)

[Clear](#)

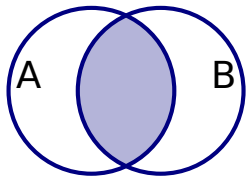
### Builder

All Fields ▼  − [Show index list](#)

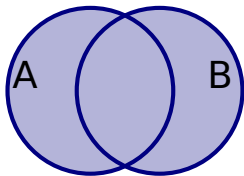
AND ▼ All Fields ▼  − + [Show index list](#)

Search or [Add to history](#)

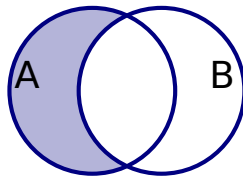
# Логические операторы



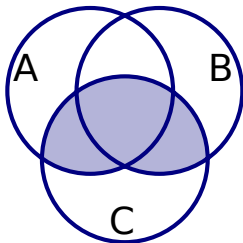
$A \text{ AND } B$



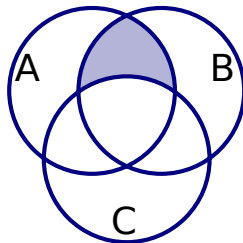
$A \text{ OR } B$



$A \text{ NOT } B$

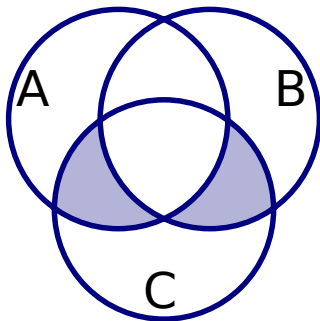


$(A \text{ OR } B) \text{ AND } C$



$(A \text{ AND } B) \text{ NOT } C$

# Логические операторы



**Задание 5.** Напишите выражение, которое бы соответствовало рисунку.

# Поля для поиска

Ссылка на [полный список](#)

- ▶ [Feature Key], [FKEY] — по типу feature
- ▶ [Gene Name], [GENE] — название гена
- ▶ [Protein Name], [PROT] — название белка
- ▶ [Organism], [ORGN] — название организма
- ▶ [Accession], [ACCN] — Accession ID
- ▶ [Publication Date], [PDAT] — дата публикации

**Задание 6.** Задайте поиск слова *spike* только по названиям генов. Сколько последовательностей удовлетворяют этому условию?



**Задание 7.** Задайте поиск слова *spike* либо по названиям генов, либо белков. Сколько последовательностей удовлетворяют этому условию?

**Задание 8.** Самостоятельно найти и скачать последовательность гена (fasta), который носит название *spike* или *s*, или же так назван соответствующий белок, и при этом получена из изолята RaTG13. Какой организм был носителем вируса из этого изолята? Какой размер у полученного файла?

# eUtils

Путь для обращения к eUtils:

`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/`

Синтаксис запроса:

`.../<eutil>.fcgi?db=<database>&term=<query>`

▶ `<eutil>` — **утилита**:

- ▶ `esearch` — поиск по базе данных
- ▶ `efetch` — скачивание из базы данных
- ▶ `esummary` — обзор информации по ID

▶ `<db>` — **база данных**:

- ▶ `pubmed`
- ▶ `Nucleotide, nuccore`

▶ `term=<query>` — запрос (`spike[GENE]`), может меняться в зависимости от утилиты, этот пример для `esearch`

# Пример поискового запроса в eUtils

Путь для обращения к eUtils:

`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/`

Синтаксис запроса:

`.../<eutil>.fcgi?db=<database>&term=<query>`

Пример запроса:

`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/  
esearch.fcgi?db=nuccore&  
term=spike[GENE]`

# Особенности запросов в eUtils

- ▶ Нельзя использовать пробелы, вместо них +  
`spike[GENE]+AND+spike[PROT]`
- ▶ Нельзя использовать кавычки, вместо них %22  
`%22spike+gene%22[GENE]`

**Задание 9.** Составьте запрос, чтобы получить общее количество последовательностей белков с названием "spike glycoprotein" в Nucleotide. В качестве ответа загрузите запрос и количество найденных записей.

# Дополнительные параметры

- ▶ `retstart` — номер первого запроса для вывода
- ▶ `retmax` — максимально количество отображаемых результатов поиска
- ▶ `retmode` — формат вывода (xml, json)

```
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/  
esearch.fcgi?db=nuccore&  
term=s[GENE]&  
retstart=100&  
retmax=100&  
retmode=json
```

Подробное описание параметров

# eFetch. Скачивание файлов

Ссылка на [подробные инструкции](#)

Синтаксис запроса:

```
.../efetch.fcgi?db=nuccore&id=<id>&  
rettype=<format>
```

- ▶ id — идентификатор(ы) последовательностей, в качестве разделителя используют запятую
- ▶ rettype — формат, в котором будут скачены последовательности (fasta, gb)

```
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/  
efetch.fcgi?db=nuccore&  
id=9858048&  
rettype=fasta
```

**Задание 10.** Найдите кодирующие последовательности (cds) генов, которые называются *spike* или *s*, или же так назван соответствующий белок, и при этом принадлежит вирусу (организму) PEDV. С помощью eFetch скачайте любые три последовательности в формате fasta. В ответе загрузите запрос для поиска и обращения к efetch.