

Семинар 1. Базы данных геномов бактерий. Файлы аннотации геномов генами. Стренды.

ПРОКАРИОТЫ - бактерии и археи

NCBI

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

Genbank

Query of the GenBank database is carried out via the NCBI Entrez system [entrez], which is used to query all

entrez. <http://www.ncbi.nlm.nih.gov/nucleotide>

Homo sapiens[ORGN]) AND 3000:4000[LEN]

EMBL and DDBJ

The European counterpart to GenBank is the ENA [ena], located at the European Bioinformatics Institute (EBI) [ebi]. Another primary nucleotide sequence database, the DDBJ [ddbj], is operated by the National Institute of Genetics (NIG) [nig] in Japan and is the primary nucleotide sequence database for Asia. The three database operators, NCBI, EBI, and NIG, compose the International Nucleotide Sequence Database Collaboration and synchronize their databases every 24 h. A query of all three individual databases is therefore not necessary, nor is it required to enter a new nucleotide sequence into all three databases.

Bacmap An Interactive Atlas for Exploring Bacterial Genomes

<http://wishart.biology.ualberta.ca/BacMap/>

Посмотреть карты геномов разных бактерий и дополнительную информацию о них.

DOWNLOAD MICROBIAL GENOMES

<https://www.ncbi.nlm.nih.gov/genome/microbes/>

Download/FTP Refseq Bacteria genomes


Например, 4ая запись, чтобы долго не грузить -

ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Absiella_dolichum/

ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Absiella_dolichum/latest_assembly_versions/GCF_003474925.1_ASM347492v1

Index of ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Absiella_dolichum/latest_assembly_versions/GCF_003474925.1_ASM347492v1/

 Up to higher level directory

Name	Size	Last Modified
File: GCF_003474925.1_ASM347492v1_assembly_report.txt	20 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_assembly_stats.txt	4 KB	9/8/18 3:00:00 AM GMT+3
 GCF_003474925.1_ASM347492v1_assembly_structure		9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_cds_from_genomic.fna.gz	665 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_feature_count.txt.gz	1 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_feature_table.txt.gz	94 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_genomic.fna.gz	635 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_genomic.gbff.gz	1607 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_genomic.gff.gz	147 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_protein.faa.gz	368 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_protein.gpff.gz	525 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_rna_from_genomic.fna.gz	5 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_translated_cds.faa.gz	478 KB	9/8/18 3:00:00 AM GMT+3
File: GCF_003474925.1_ASM347492v1_wgsmaster.gbff.gz	2 KB	9/8/18 3:00:00 AM GMT+3
README.txt		9/8/18 3:00:00 AM GMT+3
File: annotation_hashes.txt	1 KB	9/8/18 3:00:00 AM GMT+3
File: assembly_status.txt	1 KB	9/9/19 10:57:00 AM GMT+3
File: md5checksums.txt	2 KB	9/8/18 3:00:00 AM GMT+3

Загрузить файлы, разархивировать

>gunzip

Про типы файлов аннотаций бактериальных геномов, доступных на ftp:

<https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/>

fna - полная нуклеотидная последовательность

feature_table - поизучать файл аннотации. Разобраться со стрендами. Увидеть гены, кодирующие белки, и РНК-гены. С помощью Unix-команд посчитать, сколько рибосомальных геномов, сколько tRNA-генов, найти ATPase operon.

cds_from_genomic.fna - FASTA format of the nucleotide sequences corresponding to all CDS features annotated on the assembly, based on the genome sequence.

translated_cds.faa - FASTA sequences of individual CDS features annotated on the genomic records, conceptually translated into protein sequence. The sequence corresponds to the translation of the nucleotide sequence provided in the *_cds_from_genomic.fna.gz file.

protein.faa - FASTA format of the accessioned protein products annotated on the genome assembly.

.gbff - поизучать GenBank format

```
>less GCF_003474925.1_ASM347492v1_cds_from_genomic.fna
```

```
>less GCF_003474925.1_ASM347492v1_cds_from_genomic.fna | wc -l
```

27522

```
>less GCF_003474925.1_ASM347492v1_cds_from_genomic.fna | wc -wc
```

43422 2361983

```
>more GCF_003474925.1_ASM347492v1_feature_table.txt
```

```
>more GCF_003474925.1_ASM347492v1_feature_table.txt | wc -l
```

4503

```
>more GCF_003474925.1_ASM347492v1_feature_table.txt | grep protein_coding | wc -l
```

1931

```
>more GCF_003474925.1_ASM347492v1_feature_table.txt | grep pseudogene | wc -l
```

272

```
> more GCF_003474925.1_ASM347492v1_feature_table.txt | grep rRNA
```

```
>more GCF_003474925.1_ASM347492v1_feature_table.txt | grep "16S rRNA"
```

```
>more GCF_003474925.1_ASM347492v1_feature_table.txt | grep "23S rRNA"
```

```
>more GCF_003474925.1_ASM347492v1_feature_table.txt | grep "tRNA-" | wc -l
```

42

Посмотреть оперонную организацию молекулярной машины ATP synthase, где каждая субъединица закодирована отдельным геном:

```
>more GCF_003474925.1_ASM347492v1_feature_table.txt | grep "ATP synthase"
```

Просмотреть командой more/less все остальные файлы. Изучить формат GenBank