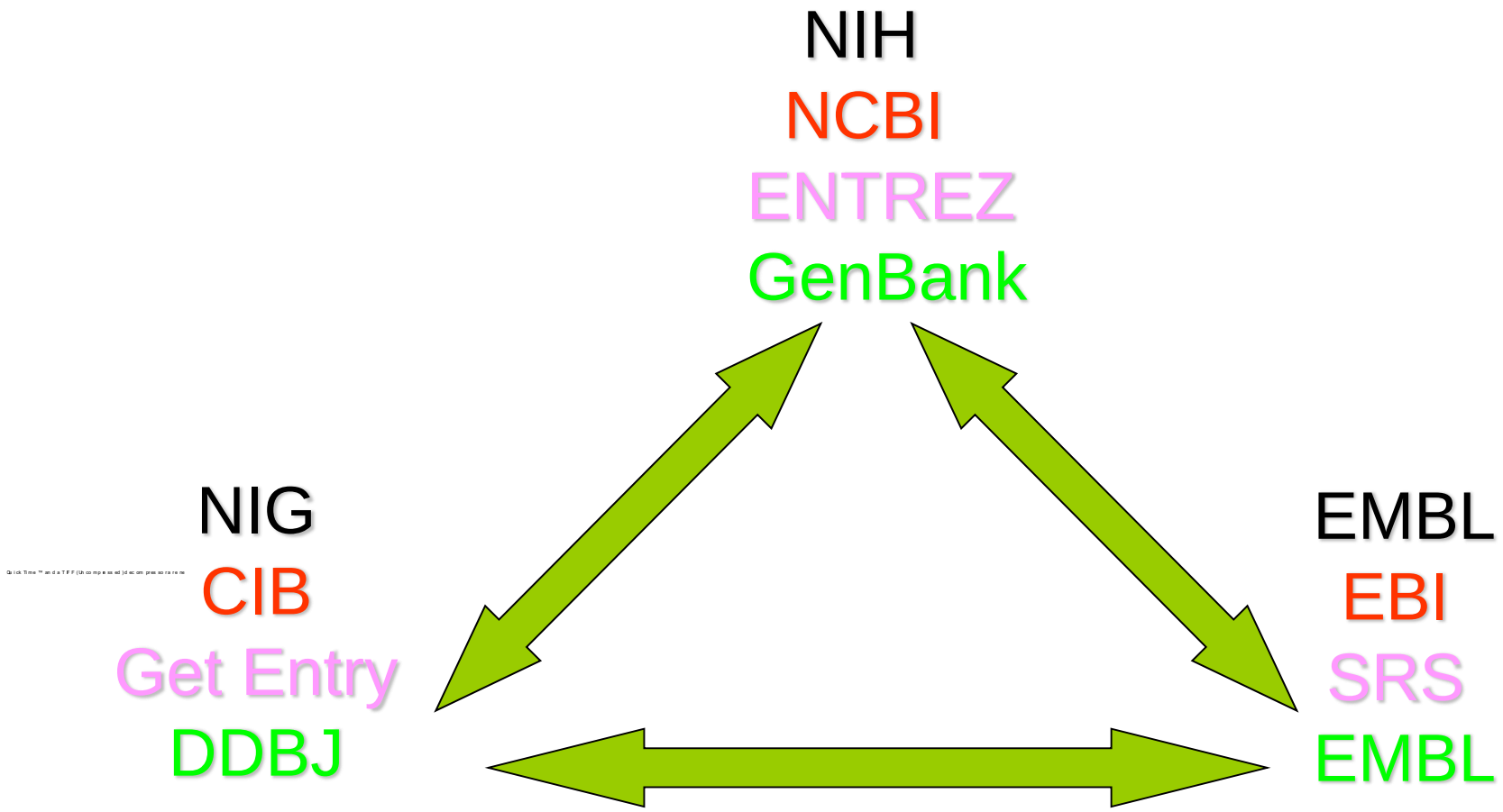
Zaslání sekvence DNA do primární databáze GenBank/EMBL/DDBJ

Nejdůležitější databáze sekvencí nukleových kyselin a proteinů

- V každém ze tří hlavních bioinformatických center je spravována **genomová databáze** sekvencí nukleových kyselin a odpovídajících, z nich přeložených proteinů.
  - **EMBL Nucleotide Sequence Database** (v rámci institutu EBI) – 1980
  - **GenBank** (v rámci institutu NCBI) – 1982
  - **DDBJ** (The DNA Data Bank of Japan) - 1984
- Tři samostatné báze vznikly v důsledku potřeby rychlé dostupnosti databáze sekvencí na jednotlivých kontinentech v době, kdy ještě nebyly rozvinuté vysokorychlostní komunikační sítě.

Mezinárodní spolupráce sekvenčních databází

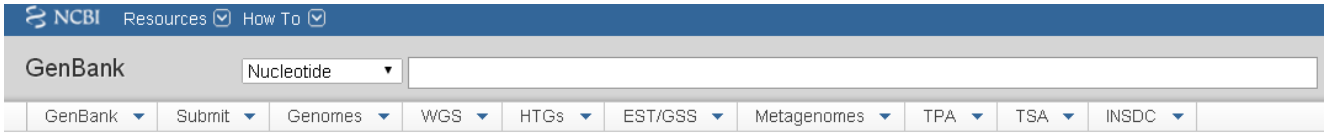
- Databáze sdílejí stejná data



Divize GenBank

https://www.ncbi.nlm.nih.gov/genbank/htgs/divisions/

ftp://ftp.ncbi.nlm.nih.gov/genbank



GenBank Database Divisions

GenBank divisions are divided into two general categories and were described in an (Genome Research (1997) 7(10)) article by Ouellette and Boguski; the full-text article is available (Database Divisions and Homology Search Files: A Guide for the Perplexed). The "Organismal" category includes databases pertaining to sequences derived from specific organisms and the "Functional" databases pertain to different types of sequence data being collected. Sequence records exist only in one GenBank division. For example, the HTG division includes unfinished sequences (phases 0, 1, and 2) being generated from several different organisms. As a sequence is updated to phase 3, it is moved into the appropriate organismal division. For instance, human phase 3 (finished) HTG sequences are located in the PRI division. The GenBank divisions listed here represent the location of the annotated sequence records; for homology search purposes the records are reformatted and stored in the BLAST databases. The different database divisions currently available, as well as the related BLAST database, are listed below. An example of a submission (one accession number) that has progressed through phase 1, phase 2, and phase 3 is available (Examples).

HTGs Resources

About HTGs
Submitting HTGs
Processing HTGs
HTGs FAQ

Organismal Divisions:

Database	Division	BLAST	Example
BCT	Bacterial sequences	nr, month	
PRI	Primate sequences	nr, month	Human Phase 3
ROD	Rodent sequences	nr, month	
MAM	Other mammalian sequences	nr, month	
VRT	Other vertebrate sequences	nr, month	
INV	Invertebrate sequences	nr, month	Drosophila, C. elegans Phase 3
PLN	Plant and Fungal sequences	nr, month	Arabidopsis Phase 3
VRL	Viral sequences	nr, month	
PHG	Phage sequences	nr, month	
RNA	Structural RNA sequences	nr, month	
SYN	Synthetic and chimeric sequences	nr, month	
UNA	Unannotated sequences	nr, month	

Functional Divisions:

Database	Division	BLAST	Example
EST	Expressed Sequence Tags	dbest, month	
STS	Sequence Tagged Sites	dbsts, month	
GSS	Genome Survey Sequences	dbgss, month	
HTG	High Throughput Genomic sequences	htgs, month	All Organisms: Phase 0, 1, and 2

Identifikace záznamu v primárních sekvenčních databázích

- GenBank
- EMBL-Bank (European Nucleotide Archive, ENA)
- DDBJ
- **Přístupový kód (Accession Number)**
- **číslo GI (GenBank Identifier)**

```
LOCUS       AY870395                 553 bp    DNA     linear   BCT 30-JAN-2005
DEFINITION  Micrococcus brunensis strain CCM 4811 60 kDa chaperonin (cpn60)
            gene, partial cds.
ACCESSION   AY870395
VERSION     AY870395.1  GI:58119461
```

Tradiční záznam GenBank

```
LOCUS       AY182241                1931 bp    mRNA    linear   PLN 04-MAY-2004
DEFINITION  Malus x domestica (E,E)-alpha-farnesene synthase (AFS1) mRNA,
            complete cds.
ACCESSION   AY182241
VERSION     AY182241.2  GI:32265057
KEYWORDS    .
SOURCE      Malus x domestica (cultivated apple)
  ORGANISM  Malus x domestica
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; eurosids I; Rosales; Rosaceae; Maloideae; Malus.
REFERENCE   1  (bases 1 to 1931)
  AUTHORS   Pechous,S.W. and Whitaker,B.D.
  TITLE     Cloning and functional expression of an (E,E)-alpha-farnesene
            synthase cDNA from peel tissue of apple fruit
  JOURNAL   Planta 219, 84-94 (2004)
REFERENCE   2  (bases 1 to 1931)
  AUTHORS   Pechous,S.W. and Whitaker,B.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-NOV-2002) PSI-Produce Quality and Safety Lab,
            USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD
            20705, USA
REFERENCE   3  (bases 1 to 1931)
  AUTHORS   Pechous,S.W. and Whitaker,B.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-JUN-2003) PSI-Produce Quality and Safety Lab,
            USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD
            20705, USA
  REMARK    Sequence update by submitter
COMMENT     On Jun 26, 2003 this sequence version replaced gi:27804758.
FEATURES             Location/Qualifiers
     source          1..1931
                     /organism="Malus x domestica"
                     /mol_type="mRNA"
                     /cultivar="'Law Rome'"
                     /db_xref="taxon:3750"
                     /tissue_type="peel"
     gene            1..1931
                     /gene="AFS1"
     CDS             54..1784
                     /gene="AFS1"
                     /note="terpene synthase"
                     /codon_start=1
                     /product="(E,E)-alpha-farnesene synthase"
                     /protein_id="AAO22848.2"
                     /db_xref="GI:32265058"
                     /translation="MEFRVHLQADNEQKIFQNQMKPEPEASYLINQRRSANYKPNIWK
                     NDFLDQSLISKYDGDEYRKLSEKLIEEVKIYISAETMDLVAKLELIDSVRKLGLANLF
                     EKEIKEALDSIAAIESDNLGTRDDLYGTALHFKILRQHGYKVSQDIFGRFMDEKGTLE
                     DFLHKNEDLLYNISLIVRLNNDLGTSAAEQERGDSPSSIVCYMREVNASEETARKNIK
                     GMIDNAWKKVNGKCFTTNQVPFLSSFMNNATNMARVAHSLYKDGDGFGDQEKGPRTHI
                     LSLLFQPLVN"
ORIGIN
        1 ttcttgtatc ccaaacatct cgagcttctt gtacaccaaa ttaggtattc actatggaat
       61 tcagagttca cttgcaagct gataatgagc agaaaatttt tcaaaaccag atgaaacccg
      121 aacctgaagc ctcttacttg attaatcaaa gacggtctgc aaattacaag ccaaatattt
      181 ggaagaacga tttcctagat caatctctta tcagcaaata cgatggagat gagtatcgga
      241 agctgtctga gaagttaata gaagaagtta agatttatat atctgctgaa acaatggatt
//
```

Header

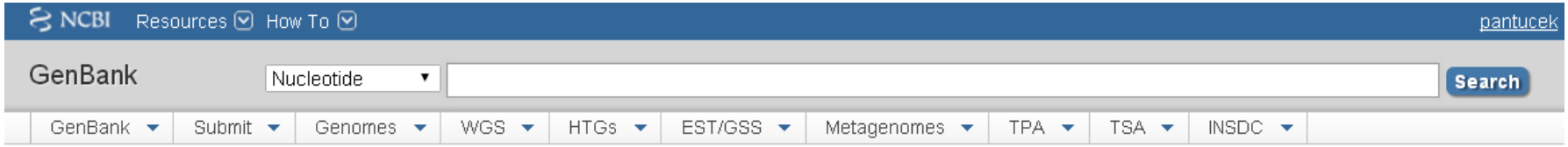
Feature Table

Sequence

Jak se data dostanou do databází?

- Předání dat prostřednictvím WWW portálu
  - BankIt (GenBank)
    - Submission Portal (https://www.ncbi.nlm.nih.gov/WebSub/)
  - WebIn (EMBL/European Nucleotide Archive)
    - http://www.ebi.ac.uk/ena/submit
  - Sakura (DDBJ)
    - http://www.ddbj.nig.ac.jp/sub/websub-e.html
- Samostatná aplikace pro PC
  - Sequin
    - http://www.ncbi.nlm.nih.gov/Sequin/download/seq_download.html
  - pro delší sekvence manuálně anotované
  - fylogenetické, populační nebo mutační studie obsahující sekvenční přiložení
- Tbl2asn – batch submissin
  - command-line program for MAC a Unix
  - automatizuje vytvoření záznamu sekvence
  - určený pro celé genomy, EST, STS a zaslání velkých dávek sekvencí

https://www.ncbi.nlm.nih.gov/genbank/submit/



How to submit data to GenBank

The most important source of new data for GenBank® is direct submissions from scientists. GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit.

Receiving an Accession Number for your Manuscript

Most journals require DNA and amino acid sequences that are cited in articles be submitted to a public sequence repository (DDBJ/EMBL/Genbank - INSDC) as part of the publication process. Data exchange between DDBJ, EMBL and GenBank occurs daily so it is only necessary to submit the sequence to one database, whichever one is most convenient, without regard for where the sequence may be published. Sequence data submitted in advance of publication can be kept confidential if requested. GenBank will provide accession numbers for submitted sequences, usually within two working days. This accession number serves as an identifier for your submitted your data, and allows the community to retrieve the sequence upon reading the journal article. The accession number should be included in your manuscript, preferably in a footnote on the first page of the article, or as required by individual journal procedures.

Submissions to GenBank

There are several options for submitting data to GenBank:

- BankIt, a WWW-based submission tool with wizards to guide the submission process
- Sequin, NCBI's stand-alone submission tool with wizards to guide the submission process is available by FTP for use on for MAC, PC, and UNIX platforms.
- tbl2asn, a command-line program, automates the creation of sequence records for submission to GenBank using many of the same functions as Sequin. It is used primarily for submission of complete genomes and large batches of sequences and is available by FTP for use on MAC, PC and Unix platforms.
- Submission Portal, a unified system for multiple submission types. Currently only ribosomal RNA (rRNA) or rRNA-ITS sequences can be submitted with the GenBank component of this tool. This will be expanded in the future to include other types of GenBank submissions. Genome and Transcriptome Assemblies can be submitted through the Genomes and TSA portals, respectively.
- Barcode Submission Tool, a WWW-based tool for the submission of sequences and trace read data for Barcode of Life projects based on the COI gene.

BankIt, Submission Portal and Barcode Submission Tool entries are automatically submitted to GenBank. Submissions made with Sequin or tbl2asn must be mailed to gb-sub@ncbi.nlm.nih.gov. Large files which may be truncated during mailing with conventional mail tools should be submitted directly using Sequin MacroSend.

You can subscribe to be notified of updates to the submission tools.

There are specialized, streamlined procedures for batch submissions of sequences, such as EST and GSS sequences.

GenBank Resources

- GenBank Home
- Submission Types
- Submission Tools
- Search GenBank
- Update GenBank Records

Genome submissoin portal

https://www.ncbi.nlm.nih.gov/genbank/genomesubmit/

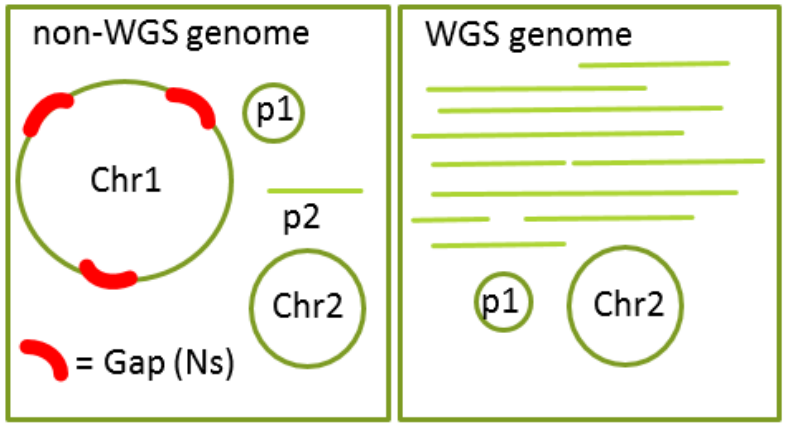
NCBI Resources How To pantucek My NCBI Sign O

GenBank Nucleotide Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

Prokaryotic and Eukaryotic Genomes Submission Guide

Both WGS and non-WGS genomes, including gapless complete bacterial chromosomes, can be submitted via the Submission Portal. You will be asked to choose whether the genome being submitted is considered WGS or not. The differences for GenBank purposes are:



non-WGS

- Each chromosome is in a single sequence and there are no extra sequences
- Each sequence in the genome must be assigned to a chromosome or plasmid or organelle
- Plasmids and organelles can still be in multiple pieces.

WGS

Genome Resources

About WGS
WGS Browser
Genome Submission Guide
Genome Submission Portal
Update Genome Records
FAQ
tbl2asn
Create Submission Template
Eukaryotic Annotation Guide
Prokaryotic Annotation Guide
Annotation Example Files
Discrepancy Report
NCBI Prokaryotic Genome Annotation Pipeline
AGP Format
Complex Assembly Submission Guide
Metagenome Submission Guide
BioProject

http://www.ebi.ac.uk/ena/submit

ENA > Submit and update

Submitting and updating data

We offer a number of services through which data (including updates) can be submitted to the European Nucleotide Archive (ENA). These technologies provide options appropriate for the scale and frequency of submission, the expertise and capacity of the submitter and the nature of the data to be transferred. The choices below lead users most directly to the appropriate submission route.

Submit
read data

Submit
assembled sequence and/or annotation
(No partial or complete assemblies)

Submit
genome assemblies
(contigs/scaffolds/chromosomes)

Email

ENA helpdesk

Protokoly pro zaslání do nukleotidové databáze

- Standard
- ESTs (expressed sequence tags) a GSSs (genome survey sequences)
- Complete Microbial or Eukaryotic Genomes
- Whole Genome Shotgun (WGS)
- High-Throughput Genomic Sequences (HTGs)
- Transcriptome Shotgun Assembly (TSA)
- Third Party Annotation (TPA)
  - záznamy, které upřesňují existující sekvence uložené do databází jinými autory
  - striktní požadavek na přímý experimentální důkaz

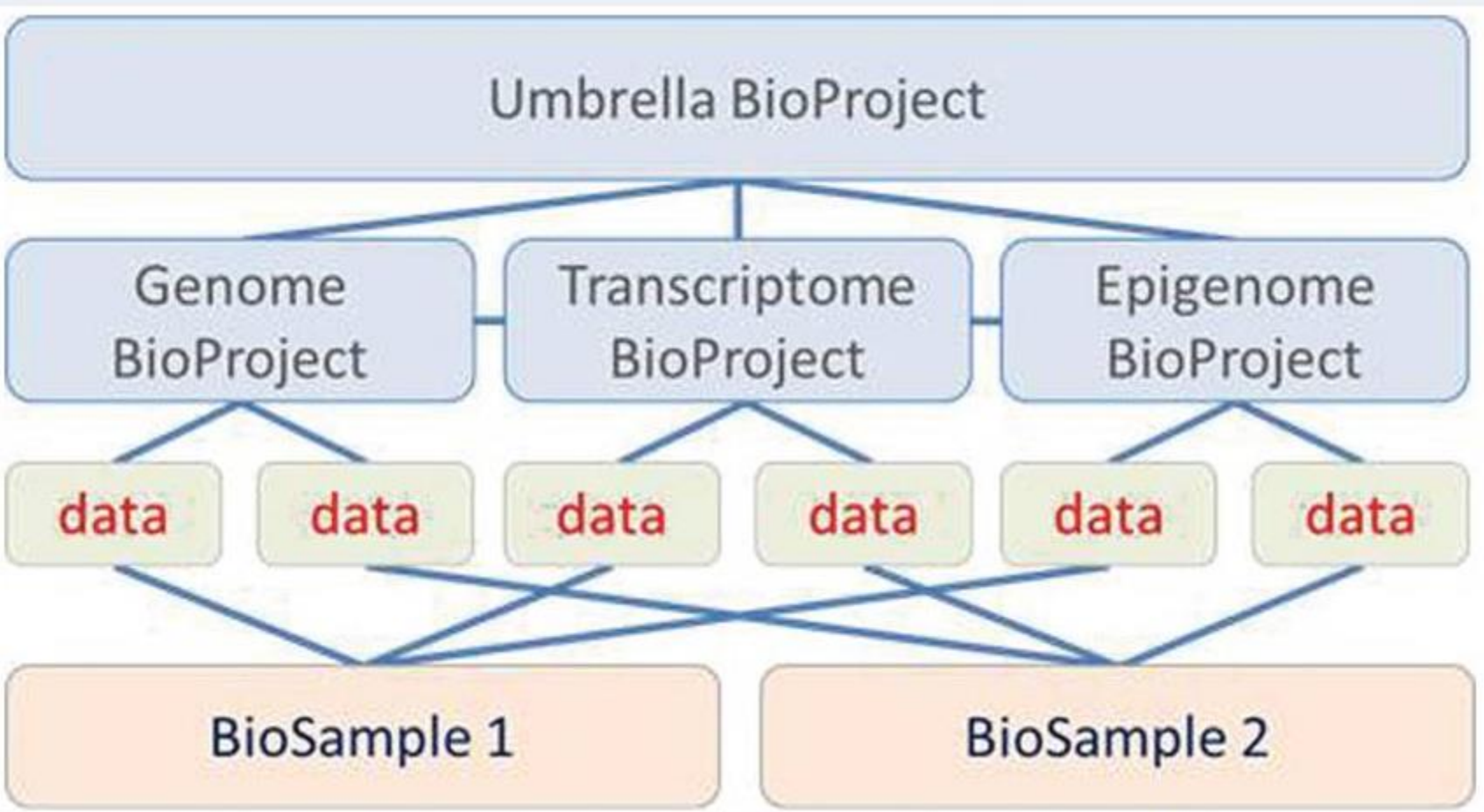
Sekvence, které nejsou akceptovány v primárních databázích

- sekvence bez fyzického (biologického) protějšku – např. konsenzní sekvence
- genomové sekvence více exonů bez údajů o sekvencích intronů
- sekvence <200 bp (vyjma patentových)
- sekvence primerů (mohou být zaslány do NCBI's Probe database)
- pouze sekvence proteinů (mohou být zaslány do UniProt/SwissProt)
- sekvence složené z genomové sekvence a mRNA reprezentované jako jedna sekvence

Typy **standardních** anotovaných sekvencí (nucleotide sequence database)

- prokaryotické geny a genomy
- eukaryotické geny a genomy
- mRNA sekvence
- rRNA a nebo ITS
- virové sekvence
- transpozony a inzerční sekvence
- mikrosatelity
- pseudogeny
- klonovací vektroy
- fylogenetické nebo populační studie (alignmenty)
- nekódující RNA

Celogenomové sekvence
BioSample & BioProject



Whole Genome Shotgun (WGS)

- WGS sekvenační projekty jsou celé genomy nebo chromozomy sekvenované strategií celogenomového shotgun sekvenování
- DDBJ/EMBL/GenBank akceptují jak kompletní, tak nekompletní genomy
- WGS projekty mohou být anotovány, může být zvolena automatická anotace s NCBI pipeline
- Části WGS projektu jsou kontigy, které nesmí obsahovat mezery
- Volitelně - soubor AGP ukazuje, jak jsou kontigy oddělené mezerami uspořádány na chromozomu
- Volitelně lze nahrát BAM nebo FASTQ do SRA (**Sequence Read Archive**)

Sequence Read Archive (SRA)

NCBI Resources How To Sign in to NCBI

SRA

SRA Search

Advanced Help

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Getting Started

- How to Submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- Submission Portal
- Trace Archive
- dbGaP Home
- BioProject
- BioSample

You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA) Support Center

GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
Submit Data

RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology

POPULAR
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene

FEATURED
Genetic Testing Registry
GenBank
Reference Sequences
Gene Expression Omnibus
Genome Data Viewer
Human Genome
Mouse Genome
Influenza Virus

NCBI INFORMATION
About NCBI
Research at NCBI
NCBI News & Blog
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube
Privacy Policy

Automatická anotace

- **NCBI Prokaryotic Genome Annotation Pipeline (PGAP)**
  - https://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/
- **NCBI Eukaryotic Genome Annotation Pipeline**
  - https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/
- **Jiné servery pro automatickou anotaci RAST**
  - http://rast.nmpdr.org/

NCBI Prokaryotic Genome Annotation Pipeline

NCBI Prokaryotic Genome Annotation Pipeline(PGAP) is designed to annotate bacterial and archaeal genomes (chromosomes and plasmids).

Genome annotation is a multi-level process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements.

NCBI has developed an automatic prokaryotic genome annotation pipeline that combines *ab initio* gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Pipeline was developed in 2001 and is regularly upgraded to improve structural and functional annotation quality (Haft DH et al 2018, Tatusova T et al 2016). Recent improvements utilize curated protein profile hidden Markov models (HMMs), including TIGRFAMS and new HMMs for antimicrobial resistance proteins, and curated complex domain architectures for functional annotation of proteins. NCBI's annotation pipeline depends on several internal databases and is not currently available for download or use outside of the NCBI environment.

Related documentation:

- Annotation process
- Annotation standards
- Assemblies excluded from RefSeq
- Release notes

GenBank

The NCBI prokaryotic annotation pipeline is available as a service for GenBank submitters. The pipeline is capable of annotating both complete genomes and draft WGS genomes consisting of multiple contigs. You can request PGAP annotation when you submit your genome to GenBank.

Both WGS and non-WGS genomes, including gapless complete bacterial chromosomes, can be submitted via the Submission Portal. You will be asked to choose whether the genome being submitted is considered WGS or not. The differences for GenBank purposes are:

non-WGS:

- Each chromosome is in a single sequence and there are no extra sequences
- Each sequence in the genome must be assigned to a chromosome or plasmid or organelle
- Plasmids and organelles can still be in multiple pieces.

WGS:

- One or more chromosomes are in multiple pieces and/or some sequences are not assembled into chromosomes

The NCBI Eukaryotic Genome Annotation Pipeline

The NCBI Eukaryotic Genome Annotation Pipeline provides content for various NCBI resources including Nucleotide, Protein, BLAST, Gene and the Genome Data Viewer genome browser.

This page provides an overview of the annotation process. Please refer to the Eukaryotic Genome Annotation chapter of the NCBI Handbook for algorithmic details.

The pipeline uses a modular framework for the execution of all annotation tasks from the fetching of raw and curated data from public repositories (sequence and Assembly databases) to the alignment of sequences and the prediction of genes, to the submission of the accessioned annotation products to public databases. Core components of the pipeline are alignment programs (Splign and ProSplign) and an HMM-based gene prediction program (Gnomon) developed at NCBI.

Important features of the pipeline include:

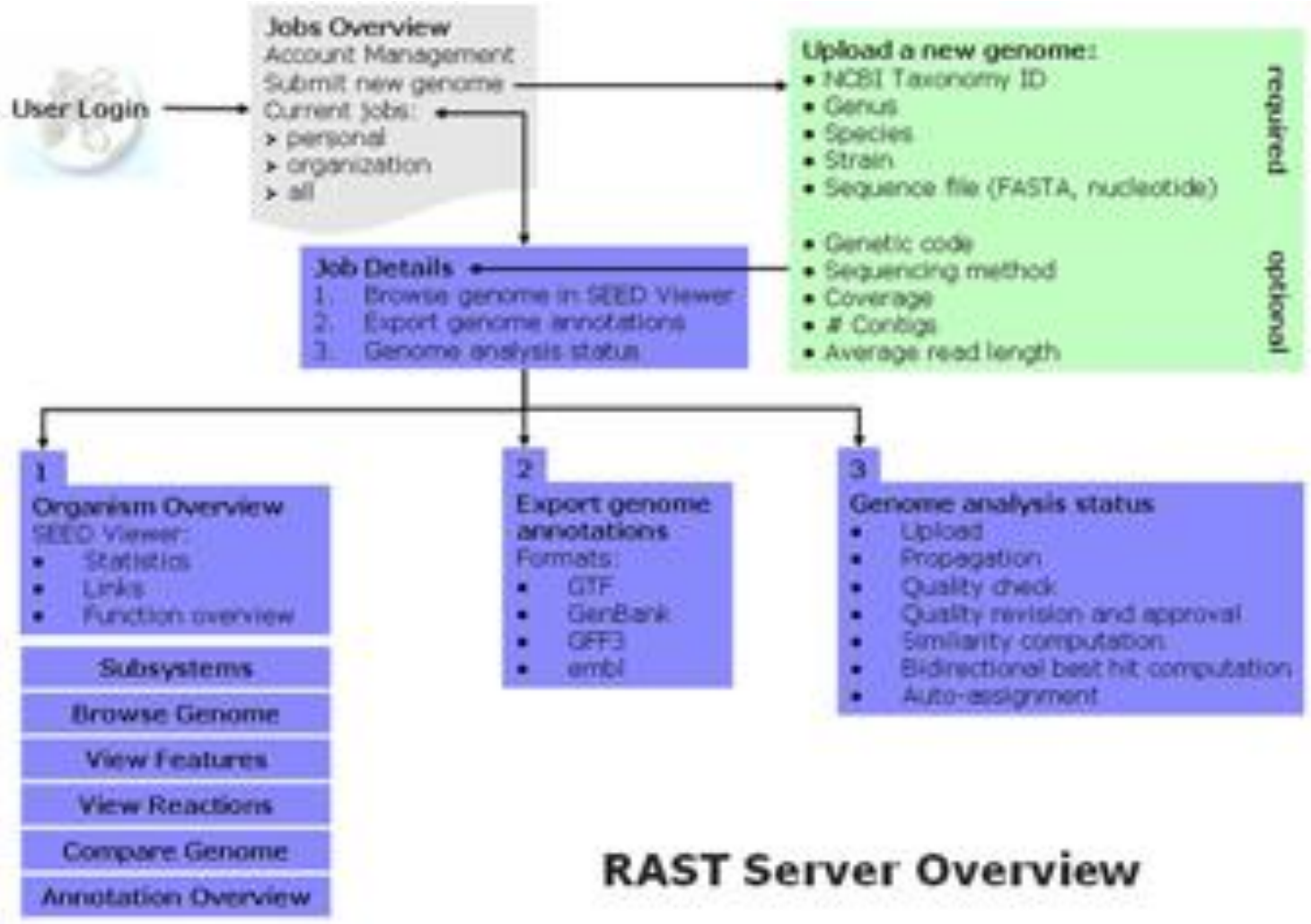
- flexibility and speed
- higher weight given to curated evidence than non-curated evidence
- utilization of RNA-Seq for gene prediction
- production of models that compensate for assembly issues
- tracking of gene loci from one annotation to the next
- ability to co-annotate multiple assemblies for the same organism

The products of an annotation run (chromosome, scaffolds and model transcripts and proteins) are labeled with an Annotation Release number. The Annotation Release name is the combination of the organism name and Annotation Release number (e.g. NCBI *Pongo abelii* Annotation Release 103) and is used throughout NCBI as a way to uniquely identify annotation products originating from the same annotation run.

Contents

RAST (Rapid Annotation using Subsystem Technology) Server

http://rast.nmpdr.org/



Metagenomy

- Metagenomika je genomová analýza společenstev mikroorganismů nezávislá na kultivaci
- Nejrozmanitější skupinou organismů na planetě jsou nekultivovatelné organismy
- Sekvenační metody nezávislé na kultivaci jsou důležité pro pochopení
  - genetické diversity
  - struktury populací
  - ekologické úlohy
  - metabolických funkcí
  - stanovení kompletních genomů nekultivovatelných organismů
  - izolaci nových mikroorganismů z prostředí
- **Sekvence jsou vzájemně propojené v rámci BioProject ID**
- Metagenomové projekty se skládají z neanotovaných sekvencí
  - shromážděné z určitých ekologických zdrojů nebo organismů
  - sestavené do kontigů
  - často obsahují částečné genomy z taxonomicky různých skupin
  - mohou obsanovat převahu informačních sekvencí jako je 16S rRNA

High-Throughput Genomic Sequences (HTGS)

- HTGS je divize nukleotidové databáze vytvořená pro uložení nekompletních genomových sekvencí stanovených ve velkých genomových centrech
- Cílem je zajistit dostupnost sekvencí pro vědeckou veřejnost, zejména prostřednictvím analýzy homologie s BLAST
- Nedokončené sekvence HTG jsou delší než 2 kb a splňují požadavky na kvalitu stanovení
- Jsou získané z jednotlivých klonů (kosmidy, BAC, YAC nebo P1)
- Kolekce klonů má přiřazený přístupový kód
- Může obsahovat chyby

Nezpracovaná data z genomových projektů

- BioSample & BioProject mohou obsahovat různé typy archivů
  - Trace Archive
    - sekvence získaní Sangerovou technikou sekvenování
    - struktura složek se *.scf nebo *.abi soubory

```
TOP_DIRECTORY/
TOP_DIRECTORY/TRACEINFO.txt
TOP_DIRECTORY/MD5
TOP_DIRECTORY/README
TOP_DIRECTORY/traces
TOP_DIRECTORY/traces/HBBA/
TOP_DIRECTORY/traces/HBBA/HBBAA1U0001.scf
TOP_DIRECTORY/traces/HBBA/HBBAA1U0002.scf
TOP_DIRECTORY/traces/HBBA/HBBAA1U0003.scf
```

  - Sequence Read Archive (SRA)
    - archiv obsahující alignment sekvencí získaných při 454, IonTorrent, Illumina, SOLiD, Helicos, PacBio nebo Complete Genomics
  - The database of Genotypes and Phenotypes (dbGaP)
    - interakce genotypu a fenotypu člověka

Formát dat a minimální požadavky pro SRA

- Doporučený formát dat je **BAM** (aligned)
- Minimální požadavek je: primární sekvence (báze) a kvalita = **FASTQ**
- Další akceptovatelné formáty dat jsou
  - SRF
  - General Fastq
  - SOLiD Fastq
  - Illumina Fastq
  - 454 SFF
  - Ion Torrent SFF
  - PacBio HDF5
  - CompleteGenomics Data Package

BAM formát

- Kompletní data z jednotlivých čtení NGS
- Bez přiložení / s přiložením
- Informace o kvalitě
- Mapování k referenční sekvenci
- Konsenzní sekvence
- Variace
- Definice např. zde:
- http://genome.sph.umich.edu/wiki/SAM#What_is_SAM

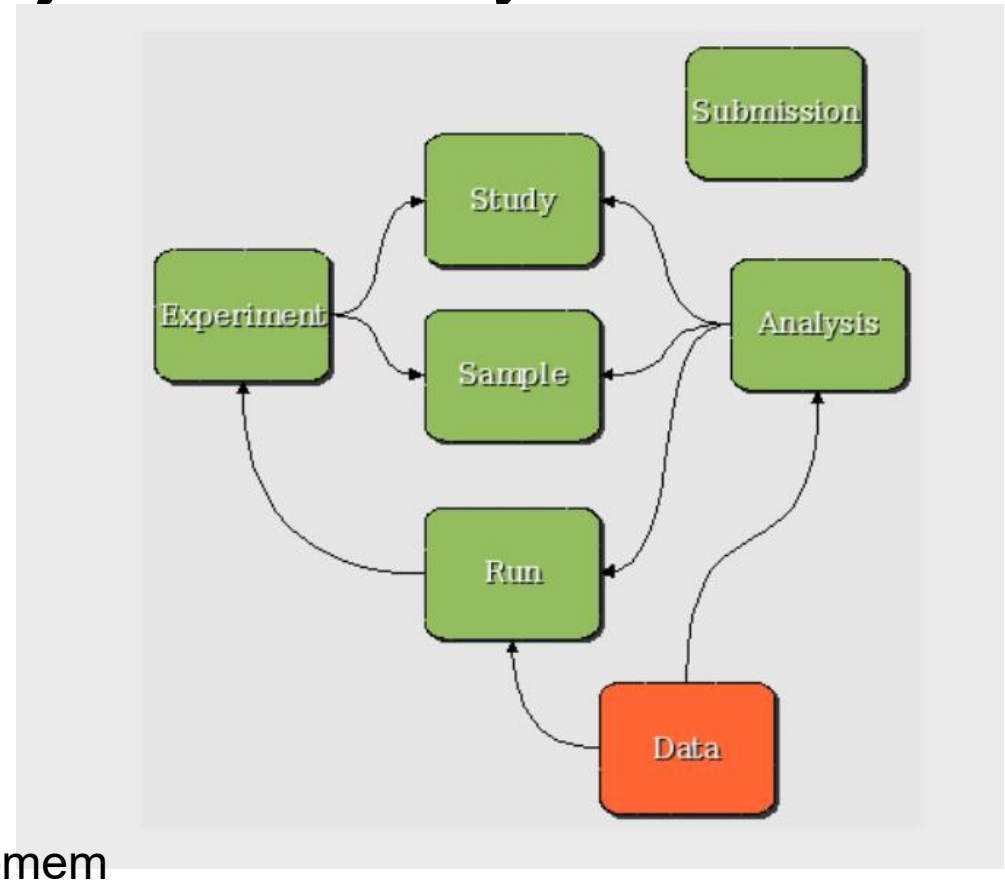
FASTQ formát

- Řádek 1 začíná hlavičkou '@'ID + popis sekvence
- Řádek 2 obsahuje primární sekvenci
- Řádek 3 začíná '+' a může následovat stejné ID a popis
- Řádek 4 obsahuje zakódované hodnoty o kvalitě sekvence a musí obsahovat stejný počet znaků jako řádek 2
- Příklad FASTQ souboru:
- **@SEQ_ID**
  GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAA
  +
  !''*(((*+))%%%++)(%%%%).1***-+*''))**55CCF>>>>A**
- **Kódování kvality, !=nejnižší kvalita, ~= nejvyšší kvalita:**

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Metadata v SRA

- Datové soubory jsou zasílány s metadaty
  - Studie
  - Experiment
  - Vzorek
  - Běh
  - Analýza
  - eticky citlivá data (EGA)

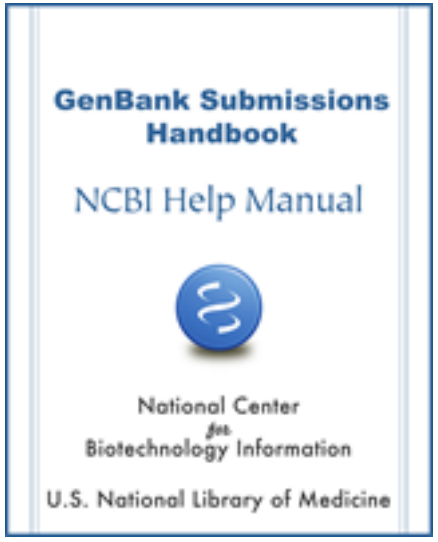


Příklad SRA s mikrobiálním genomem
https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR9600155

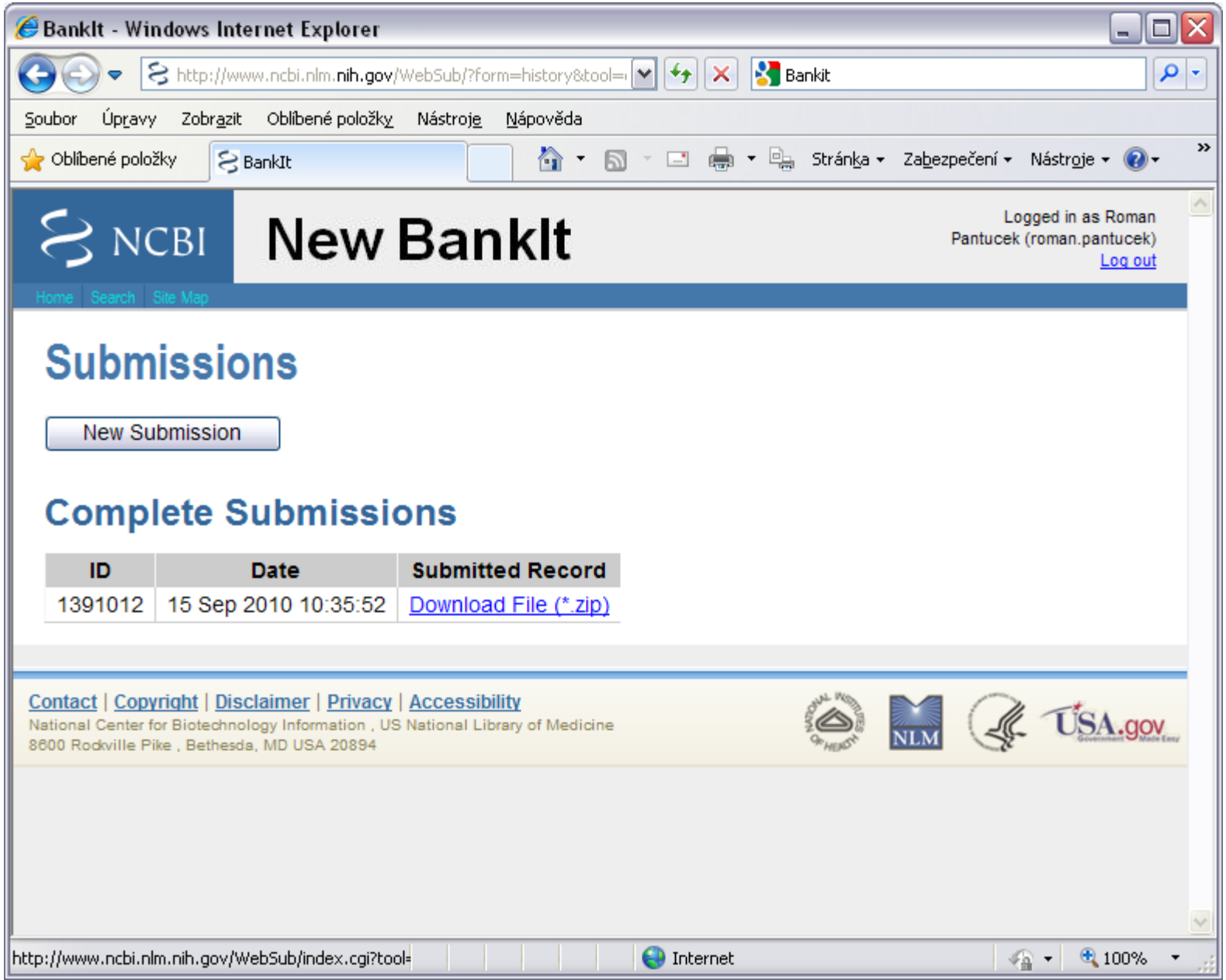
Postup zaslání GenBank Standardního typu

- http://www.ncbi.nlm.nih.gov/books/NBK51157/

The GenBank Submissions Handbook

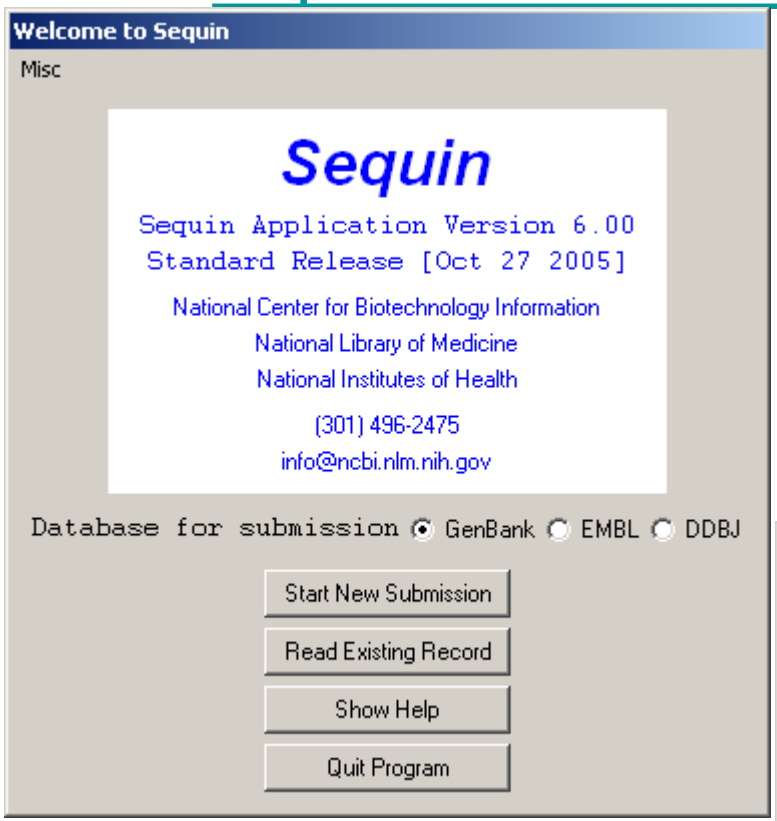


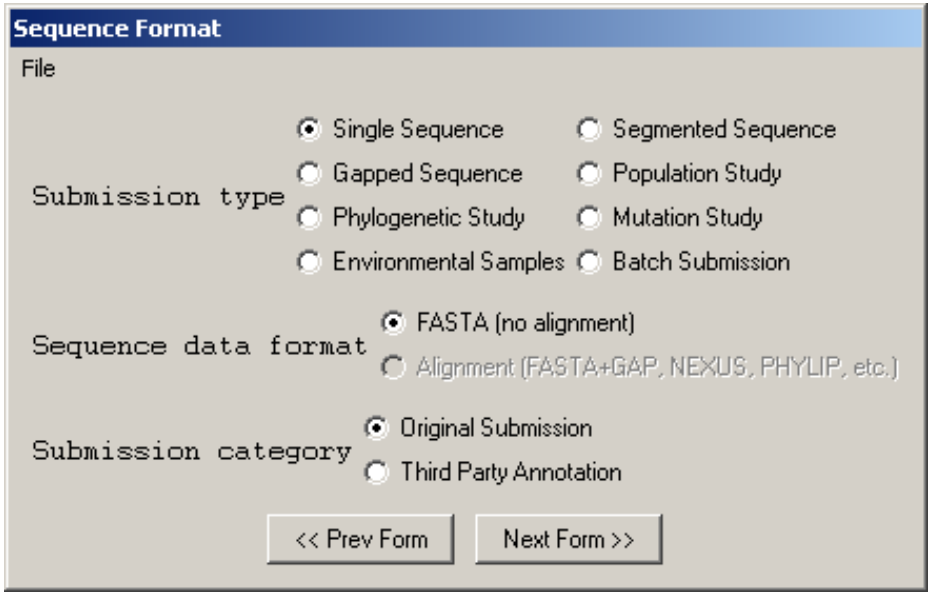
BankIt


BankIt - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/WebSub/?form=history&tool=
Bankit
Soubor Úpravy Zobrazit Oblíbené položky Nástroje Nápověda
Oblíbené položky
BankIt
Stránka Zabezpečení Nástroje
NCBI
New BankIt
Logged in as Roman Pantucek (roman.pantucek)
Log out
Home Search Site Map
Submissions
New Submission
Complete Submissions
ID
Date
Submitted Record
1391012
15 Sep 2010 10:35:52
Download File (*.zip)
Contact | Copyright | Disclaimer | Privacy | Accessibility
National Center for Biotechnology Information , US National Library of Medicine
8600 Rockville Pike , Bethesda, MD USA 20894
NLM
USA.gov
http://www.ncbi.nlm.nih.gov/WebSub/index.cgi?tool=
Internet
100%

Sequin – příprava zaslání sekvence

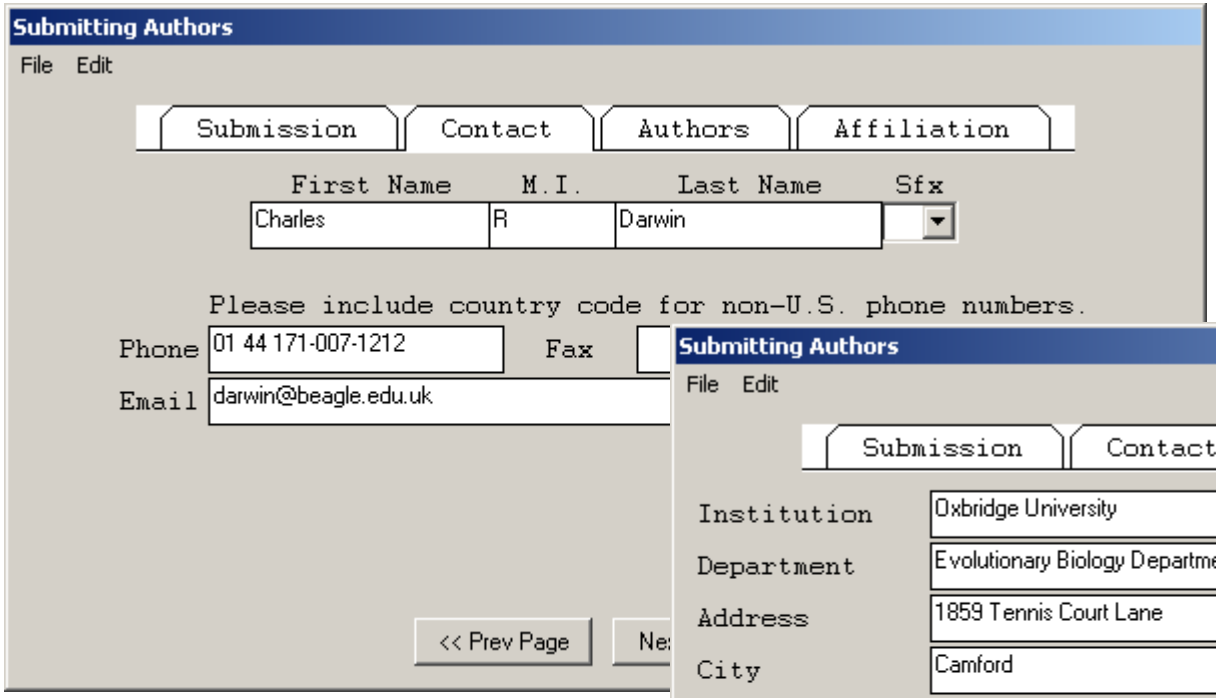
https://www.ncbi.nlm.nih.gov/Sequin/

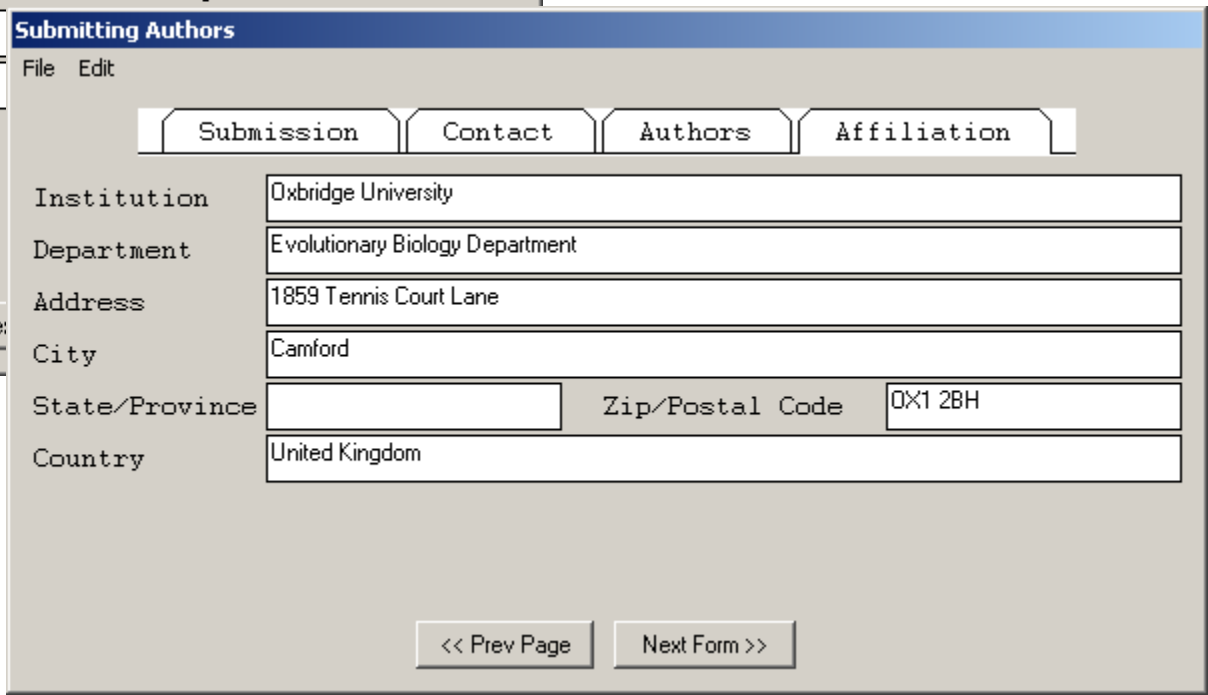




Požadavky na každé zaslání sekvence

- kontaktní informace





Další požadavky na zaslání sekvence

- Informace o datu zveřejnění
- Informace o relevantních publikacích
- Popis zdroje sekvence
- Vlastní sekvence
  - typ a tvar molekuly
  - anotace vlastností sekvence

Popis zdroje sekvence 1

- **organism**
  nezkrácené vědecké jméno
  Příklad: [organism=Drosophila melanogaster]
- **lineage**
  taxonomické zařazení organismu (dle NCBI taxonomy database)
  http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root
- **molecule**
  ve tvaru "DNA" nebo "RNA".
  Příklad : [molecule=DNA]
- **moltype**
  může nabývat následujících hodnot
  Příklad : [moltype=Genomic DNA]
  - Genomic DNA
  - Genomic RNA
  - Precursor RNA
  - mRNA [cDNA]
  - Ribosomal RNA
  - Transfer RNA
  - Small nuclear RNA
  - Small cytoplasmic RNA
  - Other-Genetic
  - cRNA
  - Small nucleolar RNA
- **topology**

Popis zdroje sekvence 2

- **location**
  může nabývat následujících hodnot
  Příklad: [location=mitochondrion]
  - genomic
  - chloroplast
  - kinetoplast
  - mitochondrion
  - plastid
  - macronuclear
  - extrachromosomal
  - plasmid
  - cyanelle
  - proviral
  - virion
  - nucleomorph
  - apicoplast
  - leucoplast
  - proplastid
  - endogenous-virus
  - hydrogenosome
- **Genetic code**
  (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c)

Popis zdroje sekvence 3

Další popisovače ke zdroji sekvence

- **acronym**
- **anamorph**
- **authority**
- **biotype**
- **biovar**
- **breed**
- **cell-line**
- **cell-type**
- **chemovar**
- **chromosome**
- **clone**
- **clone-lib**
- **collected-by**
- **common**
- **country**
- **cultivar**
- **dev-stage**
- **ecotype**
- **endogenous-virus-name**
- **forma**
- **forma-specialis**
- **fwd-pcr-primer-name**
- **fwd-pcr-primer-seq**
- **genotype**
- **group**
- **haplotype**
- **identified-by**
- **isolate**
- **isolation-source**
- **lab-host**
- **lat-lon**
- **map**
- **note**
- **pathovar**
- **plasmid-name**
- **plastid-name**
- **pop-variant**
- **rev-pcr-primer-name**
- **rev-pcr-primer-seq**
- **segment**
- **serogroup**
- **serotype**
- **serovar**
- **sex**
- **specific-host**
- **specimen-voucher**
- **strain**
- **sub-species**
- **subclone**
- **subgroup**
- **substrain**
- **subtype**
- **synonym**
- **teleomorph**
- **tissue-lib**
- **tissue-type**
- **type**
- **variety**

Formát sekvence

- Sekvence nukleové kyseliny a kódovaných proteinů připravené ve formátu FASTA

```
Nucleotide Sequence:

>ABC-1 [organism=Saccharomyces cerevisiae][strain=ABC][clone=1]
ATTGCGTTATGGAAATTCGAAACTGCCAAATACTATGTCACCATCATTGA
TGCACCTGGACACAGAGATTTCATCAAGAACATGATCACTGGTACTT

Protein Sequences:

>4E-I [gene=eIF4E] [protein=eukaryotic initiation factor 4E-I]
MQSDFHRMKNFANPKSMFKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPAGN ...
>4E-II [gene=eIF4E] [protein=eukaryotic initiation factor 4E-II]
MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPAGNTATTTAPAGDD ...
```

Přsrušená sekvence

```
>m_gagei [organism=Mansonia gagei] Mansonia gagei NADH dehydrogenase ...
ATGGAGCATACATATCAATATTCATGGATCATACCGTTTGTGCCACTTCCAATTCCTATTTTAATAGGAA
TTGGACTCCTACTTTTTCCGACGGCAACAAAAAATCTTCGTCGTATGTGGGCTCTTCCCAATATTTTATT
GTTAAGTATAGTTATGATTTTTTCGGTCGATCTGTCCATTCAGCAAATAAATAAAAGTTCTATCTATCAA
TATGTATGGTCTTGGACCATCAATAATGATTTTTCTTTCGAGTTTGGCTACTTTATTGATTCGCTTACCT
>?200 <-- Délka přerušení
GGTATAATAACAGTATTATTAGGGGCTACTTTAGCTCTTGC
TCAAAAAGATATTAAGAGGGGTTTAGCCTATTCTACAATGTCCCAACTGGGTTATATGATGTTAGCTCTA
GGTATGGGGTCTTATCGAGCCGCTTTATTTCATTTGATTACTCATGCTTATTCGAAGGCATTGTTGTTTT
TAGGATCCGGATCCGTTATTCATTCCATGGAAGCTATTGTTGGATATTCTCCAGATAAAAGCCAGAATAT
GGTTTTTATGGGCGGTTTAAGAAAGCATGTGCCAATTACACAAATTGCTTTTTTAGTGGGTACACTTTCT
CTTTGTGGTATTCCACCCCTTGCTTGTTTTTGGTCCAAAGATGAAATTCTTAGTGACAGCTGGTTGT
>?unk100 <-- Přerušení neznámé délky
TCAATAAAACTATGGGGTAAAGAAGAACAAAAAATAATTAACAGAAATTTTCGTTTATCTCCTTTATTAA
TATTAACGATGAATAATAATGAGAAGCCATATAGAATTGGTGATAATGTAAAAAAAGGGGCTCTTATTAC
TATTACGAGTTTTGGCTACAAGAAGGCTTTTTCTTATCCTCATGAATCGGATAATACTATGCTATTTCCT
ATGCTTATATTGGCTCTATTTACTTTTTTTTGTTGGAGCCATAGCAATTCCTTTTAATCAAGAAGGACTAC
ATTTGGATATATTATCCAAATTATTAACTCCATCTATAAATCTTTTACATCAAAATTCAAATGATTTTGA
GGATTGGTATCAATTTTTAACAAATGCAACTCTTTCAGTGAGTATAGCCTGTTTCGGAATATTTACAGCA
TTCCTTTTATATAAGCCTTTTTATTCATCTTTACAAAATTTGAACTTACTAAATTTATTTTCGAAAGGGG
GTCCTAAAAGAATTTTTTTGGATAAAATAATATACTTGATATACGATTGGTCATATAATCGTGGTTACAT
```

Sekvenční přiložení

- Fasta+GAP

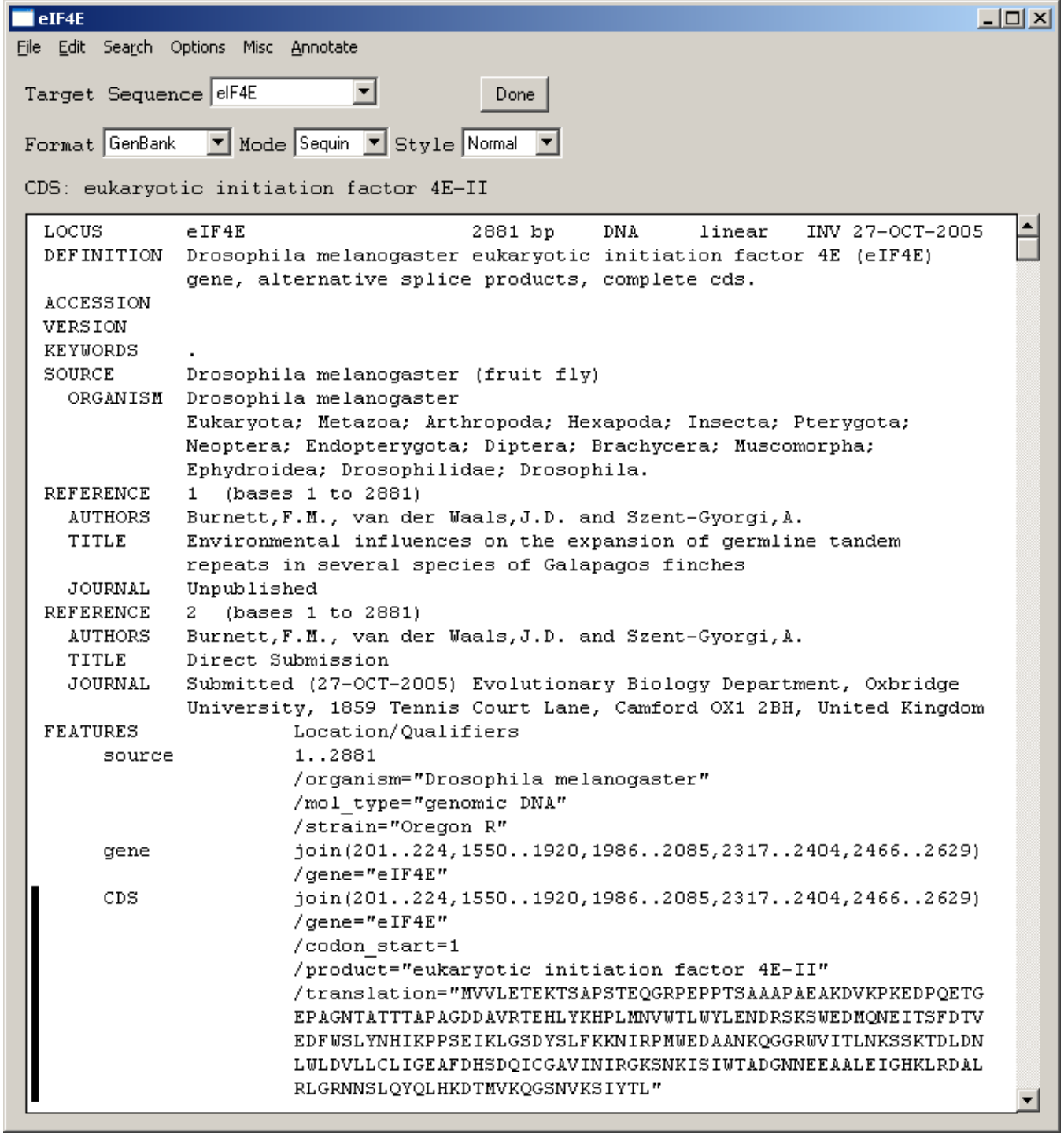
```
>ABC-1 [organism=Saccharomyces cerevisiae][strain=ABC][clone=1]
---ATTGCGTTATGGAAATTCGAAACTGCCAAATACTATGTCACCATCAT
TGATGCACCTGGACACAGAGATTTCATCAAGAACATGATCACTGGTACTT
>ABC-2 [organism=Saccharomyces cerevisiae][strain=ABC][clone=2]
GATATTGCTTTATGGAAATTCGAAACTGCCAAATACTATGTCACCATCAT
TGATGCACCTGGACACAGAAATTTCATCAAGAACATGATCACTGGTACTT
>ABC-3 [organism=Saccharomyces cerevisiae][strain=ABC][clone=3]
---ATTGCTTTATGGAAATTCGAAACTGCCAAATACTATGTTA-------
TGATGCACCTGGACACAGAGATTTCATCAAAAACATGATCACTGGTACTT
```

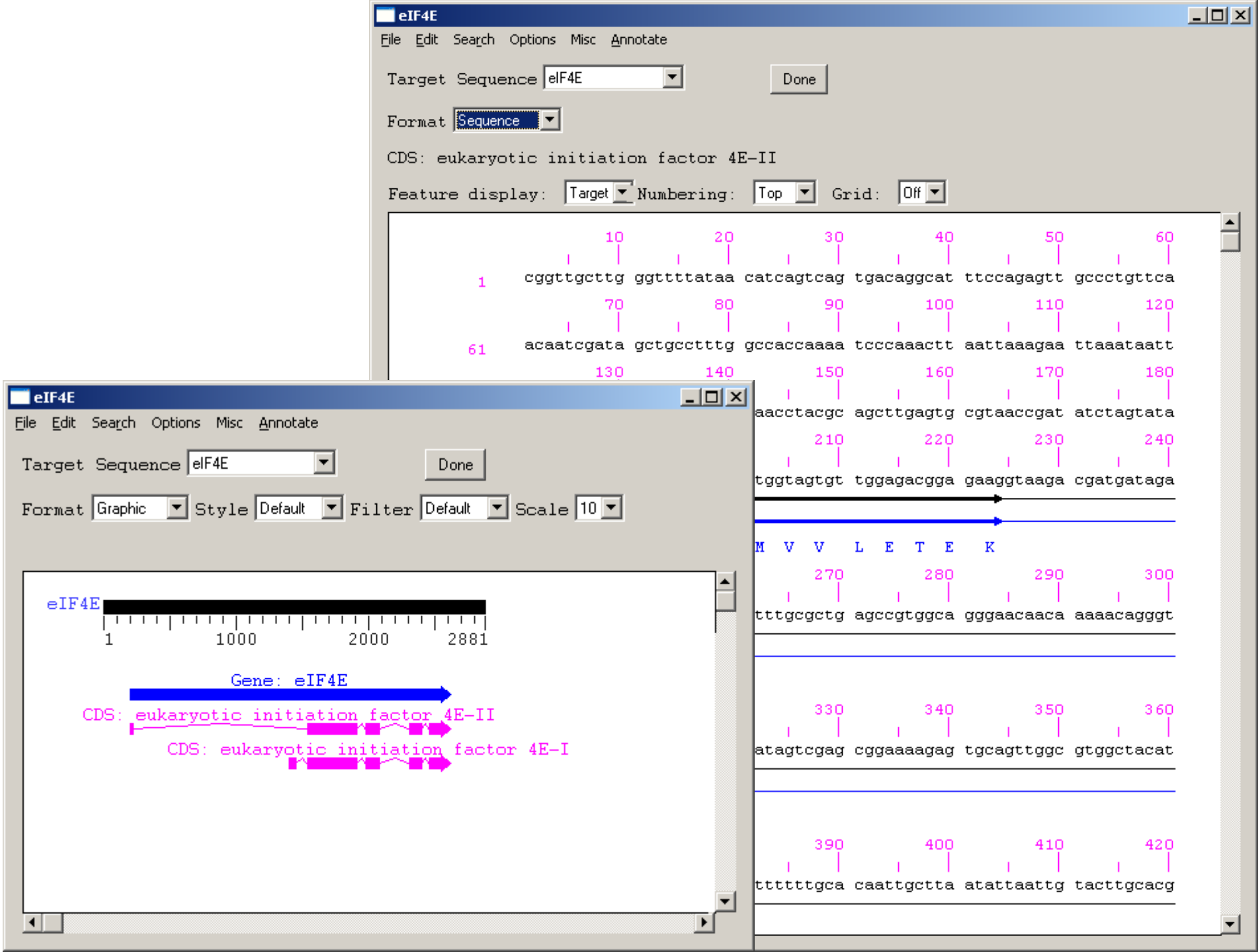
- PHYLIP

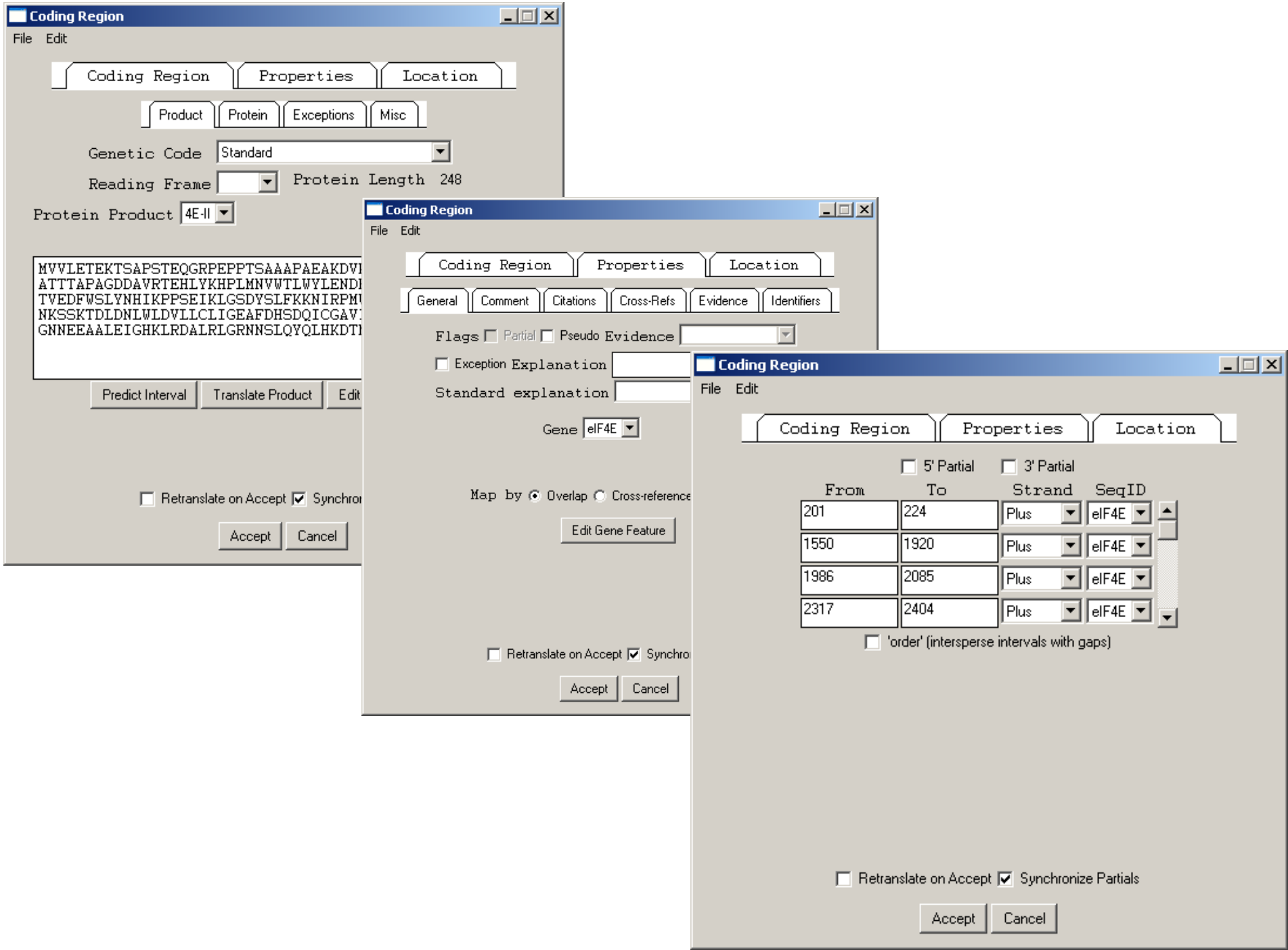
```
 3  100
ABC-1     ---ATTGCGT TATGGAAATT CGAAACTGCC AAATACTATG TCACCATCAT
ABC-2     GATATTGCTT TATGGAAATT CGAAACTGCC AAATACTATG TCACCATCAT
ABC-3     ---ATTGCTT TATGGAAATT CGAAACTGCC AAATACTATG TTA-------

          TGATGCACCT GGACACAGAG ATTTCATCAA GAACATGATC ACTGGTACTT
          TGATGCACCT GGACACAGAA ATTTCATCAA GAACATGATC ACTGGTACTT
          TGATGCACCT GGACACAGAG ATTTCATCAA AAACATGATC ACTGGTACTT

>[organism=Saccharomyces cerevisiae][strain=ABC][clone=1]
>[organism=Saccharomyces cerevisiae][strain=ABC][clone=2]
>[organism=Saccharomyces cerevisiae][strain=ABC][clone=3]
```


eIF4E
File Edit Search Options Misc Annotate
Target Sequence eIF4E
Done
Format GenBank Mode Sequin Style Normal
CDS: eukaryotic initiation factor 4E-II
LOCUS eIF4E 2881 bp DNA linear INV 27-OCT-2005
DEFINITION Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E)
gene, alternative splice products, complete cds.
ACCESSION
VERSION
KEYWORDS .
SOURCE Drosophila melanogaster (fruit fly)
ORGANISM Drosophila melanogaster
Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
Ephydroidea; Drosophilidae; Drosophila.
REFERENCE 1 (bases 1 to 2881)
AUTHORS Burnett,F.M., van der Waals,J.D. and Szent-Gyorgi,A.
TITLE Environmental influences on the expansion of germline tandem
repeats in several species of Galapagos finches
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 2881)
AUTHORS Burnett,F.M., van der Waals,J.D. and Szent-Gyorgi,A.
TITLE Direct Submission
JOURNAL Submitted (27-OCT-2005) Evolutionary Biology Department, Oxbridge
University, 1859 Tennis Court Lane, Camford OX1 2BH, United Kingdom
FEATURES Location/Qualifiers
source 1..2881
/organism="Drosophila melanogaster"
/mol_type="genomic DNA"
/strain="Oregon R"
gene join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
/gene="eIF4E"
CDS join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
/gene="eIF4E"
/codon_start=1
/product="eukaryotic initiation factor 4E-II"
/translation="MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETG
EPAGNTATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQNEITSFDTV
EDFWSLYNHIKPPSEIKLGSDYSLFKKNIRPMWEDAANKQGGRWVITLNKSSKTDLDN
LWLDVLLCLIGEAFDHSDQICGAVINIRGKSNKISIWTADGNNEEAALEIGHKLRDAL
RLGRNNSLQYQLHKDTMVKQGSNVKSIYTL"


eIF4E
File Edit Search Options Misc Annotate
Target Sequence eIF4E
Done
Format Sequence
CDS: eukaryotic initiation factor 4E-II
Feature display: Target Numbering: Top Grid: Off
10 20 30 40 50 60
1 cggttgcttg ggttttataa catcagtcag tgacaggcat ttccagagtt gccctgttca
70 80 90 100 110 120
61 acaatcgata gctgcctttg gccaccaaaa tcccaaactt aattaaagaa ttaaataatt
130 140 150 160 170 180
aacctacgc agcttgagtg cgtaaccgat atctagtata
210 220 230 240
tggtagtgt tggagacgga gaaggtaaga cgatgataga
M V V L E T E K
270 280 290 300
tttgcgctg agccgtggca gggaacaaca aaaacagggt
330 340 350 360
atagtcgag cggaaaagag tgcagttggc gtggctacat
390 400 410 420
ttttttgca caattgctta atattaattg tacttgcacg
eIF4E
File Edit Search Options Misc Annotate
Target Sequence eIF4E
Done
Format Graphic Style Default Filter Default Scale 10
eIF4E
1 1000 2000 2881
Gene: eIF4E
CDS: eukaryotic initiation factor 4E-II
CDS: eukaryotic initiation factor 4E-I


Coding Region
File Edit
Coding Region
Properties
Location
Product
Protein
Exceptions
Misc
Genetic Code
Standard
Reading Frame
Protein Length 248
Protein Product
4E-II
MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVI
ATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDI
TVEDFWSLYNHIKPPSEIKLGSDYSLFKKNIRPMV
NKSSKTDLDNLWLDVLLCLIGEAFDHSDQICGAVI
GNNEEAALEIGHKLRDALRLGRNNSLQYQLHKDTI
Predict Interval
Translate Product
Edit
Retranslate on Accept
Synchro
Accept
Cancel
Coding Region
File Edit
Coding Region
Properties
Location
General
Comment
Citations
Cross-Refs
Evidence
Identifiers
Flags
Partial
Pseudo
Evidence
Exception
Explanation
Standard explanation
Gene
elF4E
Map by
Overlap
Cross-reference
Edit Gene Feature
Retranslate on Accept
Synchro
Accept
Cancel
Coding Region
File Edit
Coding Region
Properties
Location
5' Partial
3' Partial
From
To
Strand
SeqID
201
224
Plus
elF4E
1550
1920
Plus
elF4E
1986
2085
Plus
elF4E
2317
2404
Plus
elF4E
'order' (intersperse intervals with gaps)
Retranslate on Accept
Synchronize Partials
Accept
Cancel

Anotace vlastní sekvence

- Kódované proteiny
  - CDS
    interval
    nekompletnost na N- nebo C- konci
  - gene
    interval odpovídající CDS u experimentálně prokázaných genů
  - mRNA
    interval obsahující 5‘-UTR a 3‘-UTR
- Kódované strukturní RNA

Příklady sekvencí

Sekvence mRNA nebo cDNA

- Kódující oblasti včetně iniciačního a terminačního kodonu
- Název proteinu
- Název genu
- Sekvence proteinu

```
Homo sapiens prolidase (PEPD) mRNA, complete cds.

FEATURES             Location/Qualifiers
     source          1..1888
                     /organism="Homo sapiens"
                     /chromosome="19"
                     /map="19q12-q13.2"
                     /cell_type="fibroblasts"
     mRNA            1..1888
                     /gene="PEPD"
     gene            1..1888
                     /gene="PEPD"
     CDS             17..1498
                     /gene="PEPD"
                     /EC_number="3.4.13.9"
                     /note="imidodipeptidase"
                     /product="prolidase"
```

Sekvence prokaryotického genu

- Kódující intervaly
- Název proteinu
- Název genu, je-li známý
- Aminokyselinová sekvence

```
Escherichia coli RecA protein (recA) gene, complete cds.

FEATURES             Location/Qualifiers

     source          1..3300
                     /organism="Escherichia coli"
                     /strain="K-12"
     gene            783..1961
                     /gene="recA"
     CDS             783..1961
                     /gene="recA"
                     /function="DNA repair protein"
                     /product="RecA protein"
```

Sekvence eukaryotického genu

- Intervaly kódujících oblastí včetně start- a stop-kodonů a intervaly všech intronů
- Název proteinu
- Název genu, je-li známý
- Aminokyselinová sekvence

```
Caenorhabditis elegans tyrosine kinase PTK-2 (ptk-2) gene,
complete cds.

FEATURES             Location/Qualifiers
     source          1..3180
                     /organism="Caenorhabditis elegans"
     gene            211..3011
                     /gene="ptk-2"
     mRNA            join(211..288,533..703,763..890,940..1024,
                     1084..1380,1838..1962,2018..2099,2301..3011)
                     /gene="ptk-2"
                     /product="protein kinase PTK-2"
     CDS             join(250..288,533..703,763..890,940..1024,
                     1084..1380,1838..1962,2018..2099,2301..2456)
                     /gene="ptk-2"
                     /product="protein kinase PTK-2"
```

Ribosomální RNA a vnitřní přepisované mezerníky

- Názvy jakékoli strukturní RNA (např. tRNA-Ile, 16S ribosomal RNA)
- Názvy mezerníkových oblastí (např., internal transcribed spacer 1, 16S/23S intergenic spacer)
- Nukleotidové pozice

```
Saccharomyces cerevisiae 18S ribosomal RNA gene, partial sequence; internal
   transcribed spacer 1, 5.8S ribosomal RNA gene and internal transcribed
   spacer 2, complete sequence; and 28S ribosomal RNA gene, partial
   sequence.

FEATURES             Location/Qualifiers
     source          1..540
                     /organism="Saccharomyces cerevisiae"
                     /strain="UMD 334"
     rRNA            <1..5
                     /product="18S ribosomal RNA"
     misc_RNA        6..178
                     /product="internal transcribed spacer 1 "
     rRNA            179..377
                     /product="5.8S ribosomal RNA"
     misc_RNA        378..519
                     /product="internal transcribed spacer 2"
     rRNA            520..>540
                     /product="28S ribosomal RNA"
```

Oblast promotoru

- Název proteinu nebo genu, ke kterému patří promotor a jeho 5‘ a 3‘ obklopující sekvence
- Intervaly přepisovaných a kódujících sekvencí, pokud jsou přítomné

```
Homo sapiens enhancer-binding protein 2 (EBP2) gene, promoter region and partial
    cds.

FEATURES             Location/Qualifiers

     source          1..3061
                     /organism="Homo sapiens"
                     /chromosome="15"
                     /map="15q13"
                     /cell_line="H441"
                     /tissue_type="lung"
     gene            1..>3061
                     /gene="EBP2"
     promoter        1..2947
                     /gene="EBP2"
     TATA_signal     2918..2923
                     /gene="EBP2"
     mRNA            2948..>3061
                     /gene="EBP2"
                     /product="enhancer-binding protein 2"
     5'UTR           2948..3010
                     /gene="EBP2"
     CDS             3011..>3061
                     /gene="EBP2"
                     /product="enhancer-binding protein 2"
```

Transpozon nebo inzerční sekvence

Specifické jméno elementu

- Nukleotidoné pozice
- Jména a intervaly kódovaných genových produktů, pokud jsou přítomny (např., transposase)
- Pozice a intervaly dalších vlastností (např. LTRs, repeat regions)

```
Bacillus subtilis transposon BLT transposase (tnpA) gene,
complete cds

FEATURES             Location/Qualifiers
     source          1..1221
                     /organism="Bacillus subtilis"
                     /strain="RS2"
     source          21..1127
                     /organism="Bacillus subtilis"
                     /strain="RS2"
                     /transposon="BLT"
     repeat_region   21..61
                     /rpt_type=inverted
     gene            128..1034
                     /gene="tnpA"
     CDS             128..1034
                     /gene="tnpA"
                     /product="transposase"
     repeat_region   1085..1127
                     /rpt_type=inverted
```

Oblasti repeticí

- Intervaly repetitivních sekvencí
- Rodina repeticí (např., Alu, Mer)
- Typ repetice (tandem, inverted, flanking, terminal, direct, dispersed, or other)
- Jednotka repetice (repeat unit) popis intervalů, jestliže sekvence obsahuje více než jednu repetici

```
Homo sapiens repeat regions

FEATURES             Location/Qualifiers
     source          1..2050
                     /organism="Homo sapiens"
                     /chromosome="6"
                     /map="6q25"
     repeat_region   8..126
                     /rpt_type=dispersed
                     /rpt_family="B2"
     repeat_region   197..344
                     /rpt_type="direct"
                     /rpt_unit="197..220"
     repeat_region   389..673
                     /rpt_family="AluSx"
                     /rpt_type=dispersed
     repeat_region   847..876
                     /note="microsatellite BT21"
                     /rpt_type="tandem"
                     /rpt_unit="ca"
     repeat_region    1000..2000
                     /rpt_family="human endogeneous retrovirus K-10
```

Klonovací vektor

- Jedinečné jméno vektoru
- Kódující intervaly, jména genů a proteinů

```
Cloning vector pRB223, complete sequence

FEATURES             Location/Qualifiers
     source          1..4361
                     /organism="Cloning vector pRB223"
     gene            86..1276
                     /gene="tet"
     CDS             86..1276
                     /gene="tet"
                     /product="tetracycline resistance protein"
     RBS             1905..1909
                     /note="Shine-Dalgarno sequence"
     rep_origin      2535
     gene            complement(3293..4194)
                     /gene="bla"
     CDS             complement(3293..4153)
                     /gene="bla"
                     /product="beta-lactamase"
     misc_feature    4069..4125
                     /note="multiple cloning site"
     RBS             complement(4161..4165)
                     /gene="bla"
                     /note="Shine-Dalgarno sequence"
     promoter        complement(4188..4194)
                     /gene="bla"
```

```
Bacteriophage lysis module; endolysin and HNH endonuclease genes, complete CDS

FEATURES             Location/Qualifiers
     source          1..3165
                     /organism="Staphylococcus bacteriophage 812"
                     /virion
                     /mol_type="genomic DNA"
                     /strain="phi812"
                     /lab_host="Staphylococcus aureus CCM 4028"
                     /type="wild type"
     gene            654..3017
                     /gene="lyt812"
     CDS             join(654..1449,2329..3017)
                     /gene="lyt812"
                     /experiment="peptide sequencing"
                     /note="Lyt812"
                     /codon_start=1
                     /transl_table=11
                     /product="endolysin"
                     /translation="MAKTQAEI…… "
     misc_feature    join(1239..1449,2329..2576)
                     /gene="lyt812"
                     /note="SM00644; Ami_2; This family includes zinc amidases
                     that have N-acetylmuramoyl-L-alanine amidase activity;
                     Region: Ami_2"
     intron          1450..2328
                     /gene="lyt812"
                     /standard_name="lyt812-I1"
                     /experiment="cDNA synthesis and sequencing"
     CDS             1617..2117
                     /gene="lyt812"
                     /note="ORFI-812III"
                     /codon_start=1
                     /transl_table=11
                     /product="putative HNH endonuclease
```

Příklady některých dalších modifikací deskriptorů

- Title
  - Informace vyskytující se v databázi v DEFINITION LINE
- Comment
  - Poznámka k různým vlastnostem
- Technique
  - Umožňuje výběr techniky použité pro vytvoření nebo experimentální evidenci vlastností sekvence

Přehled deskriptorů pro popis vlastností sekvence

(http://www.ncbi.nlm.nih.gov/BankIt/help.html)

- attenuator
- C-region
- CAAT_signal
- CDS
- conflict
- D-loop
- D-segment
- enhancer
- exon
- gap
- GC_signal
- gene
- iDNA
- intron
- J_segment
- LTR
- mat_peptide
- misc_binding
- misc_difference
- misc_feature
- misc_recomb
- misc_RNA
- misc_signal
- misc_structure
- modified_base
- mRNA
- N_region
- old_sequence
- operon
- oriT
- polyA_signal
- polyA_site
- precursor_RNA
- prim_transcript
- primer_bind
- promoter
- protein_bind
- RBS
- repeat_region
- repeat_unit
- rep_origin
- rRNA
- S_region
- satellite
- scRNA
- sig_peptide
- snRNA
- snoRNA
- source
- stem_loop
- STS
- TATA_signal
- terminator
- transit_peptide
- tRNA
- unsure
- V_region
- V_segment
- variation
- 3‘clip
- 3‘UTR
- 5‘clip
- 5‘UTR