

NCBI

Data retrieval using GenBank



**CENTURION
UNIVERSITY**
*Shaping Lives...
Empowering Communities!*

What is NCBI?

On November 4, 1988 that President Ronald Reagan signed the Health Omnibus Extension Act to create The National Center for Biotechnology Information as part of National Library of Medicine at NIH.



- Create automated systems for knowledge about **molecular biology**, biochemistry, and genetics.
- Perform research into advanced methods of analyzing and interpreting **molecular biology** data.
- Enable **biotechnology** researchers and medical care personnel to use the systems and methods developed.

GenBank Sequence Format

- To search GenBank effectively using the text-based method requires an understanding of the GenBank sequence format.
- GenBank is a relational database. However, the search output for sequence files is produced as flat files for easy reading.
- The resulting flat files contain three sections; Header, Features, and Sequence entry.
- There are many fields in the Header and Features sections. Each field has a unique identifier for easy indexing by computer software.
- Understanding the structure of the GenBank files helps in designing effective search strategies.

Header

```

LOCUS       Q9ZGE9             440 aa             linear   BCT 15-JUN-2002
DEFINITION  Light-independent protochlorophyllide reductase subunit N (LI-POR
subunit N) (DPOR subunit N).
ACCESSION   Q9ZGE9
VERSION     Q9ZGE9   GI:18203677
DBSOURCE    swissprot: locus BCHN_HELMO, accession Q9ZGE9;
            class: standard.
            created: Oct 16, 2001.
            sequence updated: Oct 16, 2001.
            annotation updated: Jun 15, 2002.
            xrefs: gi: 3820536, gi: 3820556
            xrefs (non-sequence databases): InterProIPR000510, PfamPF00148
KEYWORDS    Photosynthesis; Bacteriochlorophyll biosynthesis; Oxidoreductase.
SOURCE      Heliobacillus mobilis
ORGANISM    Heliobacillus mobilis
            Bacteria; Firmicutes; Clostridia; Clostridiales; Heliobacteriaceae;
            Heliobacillus.
REFERENCE   1 (residues 1 to 440)
AUTHORS    Xiong,J., Inoue,K. and Bauer,C.E.
TITLE      Tracking molecular evolution of photosynthesis by characterization
of a major photosynthesis gene cluster from Heliobacillus mobilis
JOURNAL    Proc. Natl. Acad. Sci. U.S.A. 95 (25), 14851-14856 (1998)
MEDLINE    29061957
PUBMED     2943979
REMARK     SEQUENCE FROM N.A.
COMMENT

```

```

-----
This SWISS-PROT entry is copyright. It is produced through a
collaboration between the Swiss Institute of Bioinformatics and
the EMBL outstation - the European Bioinformatics Institute.
The original entry is available from http://www.expasy.ch/sprot
and http://www.ebi.ac.uk/sprot
-----
[FUNCTION] Uses Mg-ATP and reduced ferredoxin to reduce ring D of
protochlorophyllide (Pchlde) to form chlorophyllide a (Chlide) (By
similarity). This reaction is light-independent.
[PATHWAY] Light-independent bacteriochlorophyll biosynthesis.
[SUBUNIT] Protochlorophyllide reductase is thought to be composed
of three subunits; bchL, bchN and bchB. Could form a heterotetramer
of two bchB and two bchN subunits.
[SIMILARITY] BELONGS TO THE BCHN / CHLN FAMILY.

```

Features

```

FEATURES             Location/Qualifiers
  source              1..440
                    /organism="Heliobacillus mobilis"
                    /db_xref="taxon:28064"
  gene                1..440
                    /gene="BCHN"
  Protein             1..440
                    /gene="BCHN"
                    /product="Light-independent protochlorophyllide reductase
subunit N"
                    /EC_number="1.16.--"

```

Sequence

```

ORIGIN
1  merverengc fhtfcpiasv awlhrkikds fflivgthtc ahfiqtaldv mvyahsrfgf
61  avleesdlvs aspteelgkv vqyvvdewhp kvifvlstcs vdilkmdley sckdlstrfg
121 fvpvlpastgs idrsftgged avlhallpfv pkeapavepv eekkpwrwfs gkesekekae
181 parrnlvliga vtdstiqglq welkqlgipk vdvfpdgdtr kmpvineqtv vvpilgpylnd
241 tlatirrerz akvlstvfpi gpdgtarfle aiclefgltd srikekeaga wrdleplqi
301 lqgkkmflg dallelplar fltscdvqv v eagtpyihsk dlqelellk erdvrviesp
361 dftkqlqrmq eykpdlvvag lgicnpleam gfttawsief tfaqihgfvn aidliklftk
421 pllkrqalme hgwaeeagwle
//

```

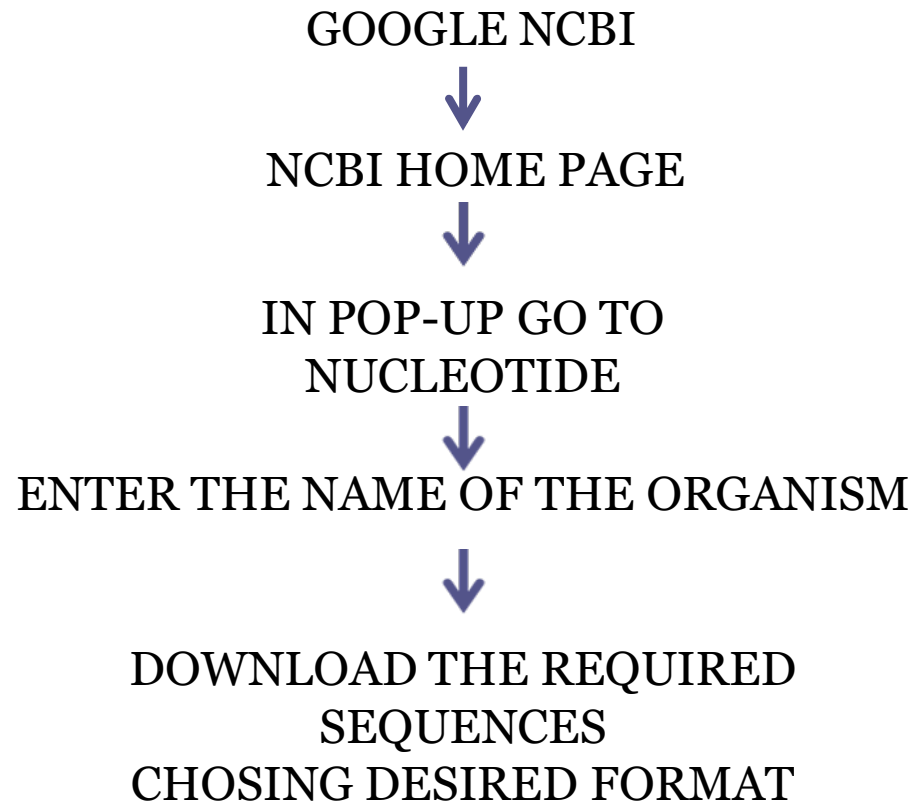
Figure 2.3: NCBI GenBank/GenPept format showing the three major components of a sequence file.

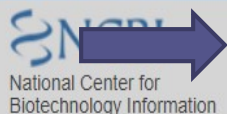


AIM OF THE EXPERIMENT

Data retrieval using GenBank, NCBI.

Procedure





Nucleotide ▾ Arabidopsis thaliana

Search

National Center for Biotechnology Information

COVID-19 is an emerging, rapidly evolving situation.
 Get the latest public health information from CDC: <https://www.coronavirus.gov>.
 Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
 Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

- NCBI Home**
- Resource List (A-Z)**
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts into NCBI databases



Download

Transfer NCBI data to your computer



Learn

Find help documents, attend a class or watch a tutorial



Develop

Use NCBI APIs and code libraries to build applications



Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects



Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

NCBI Virus: Test drive our new SARS-CoV-2 interactive data dashboard! 03 Dec 2020

Are you looking for SARS-CoV-2 sequence data? Look no further! The

December 9 Webinar: Using Docker and on the cloud



Join us on December 9, 2020 to learn about containerized BI A

Read assembly and Annotation Pipeline Tool (RAPT) is available for use and



COVID-19 is an emerging, rapidly evolving situation.
 Get the latest public health information from CDC: <https://www.coronavirus.gov>.
 Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
 Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

Species

- Animals (2,793)
- Plants (2,298,383)
- Fungi (7,881)
- Protists (2,064)
- Bacteria (21,508)
- Archaea (2)
- Viruses (49)
- Customize ...

Molecule types

- genome DNA/RNA (687,938)
- mRNA (33,269)
- rRNA
- Customize ...

Source databases

- DDBJ
- EMBL
- GenBank (2,339,764)
- PDB
- Customize ...

Sequence Type

- Nucleotide (439,967)
- EST (1,338,780)
- GSS (561,017)

Genetic

- compartments
- Chloroplast (2,472)
- Mitochondrion (125)
- Plasmid (8)

Summary 20 per page Sort by Default order

Send to: Filters: Manage Filters

GENOME ASSEMBLY

Was this helpful?



TAIR10.1

Arabidopsis thaliana (thale cress)

The Arabidopsis Information Resource (TAIR) (June 2018)

RefSeq GCF_000001735.4

PubMed (7)

Genome Browser

BLAST

Get data

Assembly statistics

Items: 1 to 20 of 2339764

<< First < Prev Page 1 of 116989 Next > Last >>

Filters activated: GenBank. Clear all

Arabidopsis thaliana chromosome 1 sequence

1. 30,427,671 bp linear DNA

/genome/gdv/browser/?acc=GCF_000001735.4&context=geno... 34.1 GI: 332189094

Results by taxon

Top Organisms [Tree]

- Arabidopsis thaliana (1855876)
- Triticum turgidum (137489)
- Triticum urartu (84575)
- Boechera stricta (42895)
- Pinus taeda (35235)
- All other taxa (183694)

Find related data

Database: Select

Find items

Search details

("Arabidopsis thaliana"[Organism] OR Arabidopsis thaliana[All Fields]) AND genbank[filter]

Search

Recent activity



CENTURION UNIVERSITY

Here in source databases we can filter by choosing GenBank, for retrieving genbank data only.

GSS (13,799)

Genetic compartments

- Chloroplast (6)
- Mitochondrion (4)
- Plasmid (2)
- Plastid (7)

Sequence length

Custom range...

Release date

Custom range...

Revision date

Custom range...

[Clear all](#)

[Show additional filters](#)

Items: 1 to 20 of 494718

Selected: 3

<< First < Prev Page 1 of 24736 Next > Last >>

Filters activated: DDBJ. [Clear all](#)

[Arabidopsis thaliana DNA, chromosome 3, complete sequence](#)

1. 23,403,063 bp linear DNA
Accession: BA000014.8 GI: 55417891
[Assembly](#) [BioProject](#) [Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Arabidopsis thaliana DNA, chromosome 5, complete sequence](#)

2. 23,810,767 bp linear DNA
Accession: BA000015.5 GI: 55417889
[Assembly](#) [BioProject](#) [Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Oryza sativa Japonica Group genomic DNA, chromosome 1, complete sequence](#)

3. 43,342,410 bp linear DNA
Accession: BA000010.8 GI: 55417890
[Assembly](#) [BioProject](#) [Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Arabidopsis thaliana DNA, Partitivirus CP-like sequence 1 \(AtPCLS1\(Ler\)\), ecotype: Ler](#)

4. 472 bp linear DNA
Accession: AB576170.1 GI: 340545477
[PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Arabidopsis thaliana DNA, corresponding sequence to flanking region of AtRE1 in ecotype C24 genome \(Chromosome 2\), ecotype: Nossen](#)

5. 1,414 bp linear DNA
Accession: AB605883.1 GI: 338746608
[Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Arabidopsis thaliana DNA, corresponding sequence to flanking region of AtRE1 in ecotype C24 genome \(Chromosome 3\), ecotype: Nossen](#)

6. 620 bp linear DNA

("Arabidopsis thaliana"[Organism] OR Arabidopsis thaliana[All Fields]) AND ddbj[filter]

Search

[See more...](#)

Recent activity

[Turn Off](#) [Clear](#)

🔍 Arabidopsis thaliana AND (ddbj[filter]) (494718) Nucleotide

🔍 Arabidopsis thaliana (3341235) Nucleotide

🔍 Arabidopsis thaliana AND (genbank[filter]) (2339764) Nucleotide

🔍 Arabidopsis thaliana mitochondrion AND (genbank[filter]) (70) Nucleotide

🔍 arobidopsis thaliana AND (genbank[filter]) (2455258) Nucleotide

[See more...](#)



CENTURION UNIVERSITY
Shaping Lives... Empowering Communities!

Mark the desired sequence which you want to download.

Nucleotide

Nucleotide

Arabidopsis thaliana



Search

Create alert Advanced

Help



COVID-19 is an emerging, rapidly evolving situation.
 Get the latest public health information from CDC: <https://www.coronavirus.gov>.
 Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
 Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/covid19>.



Species

- Animals (110)
- Plants (492,948)
- Fungi (52)
- Protists (1)
- Bacteria (49)
- Viruses (10)
- Customize ...

Molecule types

- genomic DNA/RNA (24,360)
- mRNA (447,896)
- Customize ...

Source databases

clear

- DDBJ (494,718)
- EMBL (0)
- GenBank (0)
- PDB (0)
- Customize ...

Sequence Type

- Nucleotide (45,934)
- EST (434,985)
- GSS (13,799)

Genetic

compartments

- Chloroplast (6)
- Mitochondrion (4)
- Plasmid (2)
- Plastid (7)

Summary 20 per page Sort by Default order

Send to

Filters: Manage Filters

GENOME ASSEMBLY

Was this helpful?



TAIR10.1

Arabidopsis thaliana (thale cress)

The Arabidopsis Information Resource (TAIR) (June 2018)

RefSeq GCF_000001735.4

PubMed (7)

Genome Browser

BLAST

Get data

Assembly statistics



Items: 1 to 20 of 494718

Selected: 3

<< First < Prev Page 1 of 24736 Next > Last >>

Filters activated: DDBJ. Clear all

Arabidopsis thaliana DNA, chromosome 3, complete sequence

1. 23,403,063 bp linear DNA

Results by taxon

Top Organisms [Tree]

- Arabidopsis thaliana (375330)
- Physcomitrium patens (82320)
- Eutrema halophilum (31597)
- Brassica rapa subsp. pekinensis (2166)
- Oryza sativa (1407)
- All other taxa (1898)

More...

Find related data

Database: Select

Find items

Search details

("Arabidopsis thaliana"[Organism] OR Arabidopsis thaliana[All Fields] AND ddbj[filter])

Search



CENTURION UNIVERSITY Shaping Lives... Empowering Communities!

- Species**
- Animals (110)
 - Plants (492,948)
 - Fungi (52)
 - Protists (1)
 - Bacteria (49)
 - Viruses (10)
 - Customize ...
- Molecule types**
- genomic DNA/RNA (24,360)
 - mRNA (447,896)
 - Customize ...
- Source databases** clear
- DDBJ (494,718)
 - EMBL (0)
 - GenBank (0)
 - PDB (0)
 - Customize ...
- Sequence Type**
- Nucleotide (45,934)
 - EST (434,985)
 - GSS (13,799)
- Genetic compartments**
- Chloroplast (6)
 - Mitochondrion (4)
 - Plasmid (2)
 - Plastid (7)

Summary 20 per page Sort by Default order

GENOME ASSEMBLY

[TAIR10.1](#)

[Arabidopsis thaliana \(thale cress\)](#)

The Arabidopsis Information Resource (TAIR) (June 2018)

RefSeq GCF_000001735.4

[PubMed \(7\)](#)

Genome Browser BLAST Get data

Choose Destination

Complete Record
 Coding Sequences
 Gene Features

File Clipboard

Collections

Download 3 items.

Format

- GenBank
- Summary
- GenBank**
- GenBank (full)
- FASTA
- ASN.1
- XML
- INSDSeq XML
- TinySeq XML
- Feature Table
- Accession List
- GI List
- GFF3

Items: 1 to 20 of 494718

Selected: 3

Filters activated: DDBJ. [Clear all](#)

[Arabidopsis thaliana DNA, chromosome 3, complete sequence](#)

- 23,403,063 bp linear DNA

<< First < Prev Page 1 of 24736 Next > Last >>

Search details

("Arabidopsis thaliana"[Organism] OR Arabidopsis thaliana[All Fields]) AND ddbj[filter]

Search

- In choose destination choose **file**
- In the format section chose any format you want to- **here I am choosing genbank for getting all the details.**
- To retrieve only the sequences we can download in **fasta format.**



Results- in GenBank format

sequence (2).gb - Notepad

File Edit Format View Help

```
LOCUS       BA000014               23403063 bp    DNA    linear    CON 19-MAY-2007
DEFINITION  Arabidopsis thaliana DNA, chromosome 3, complete sequence.
ACCESSION   BA000014
VERSION     BA000014.8
DBLINK      BioProject: PRJNA13190
KEYWORDS    .
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
            Camelineae; Arabidopsis.
REFERENCE   1
  AUTHORS   European Union Chromosome 3 Arabidopsis Genome Sequencing
            Consortium, The Institute for Genomic Research and Kazusa DNA
            Research Institute.
  TITLE     Sequence and analysis of chromosome 3 of the plant Arabidopsis
            thaliana
  JOURNAL   Nature 408, 820-822 (2000)
COMMENT     On Nov 5, 2004 this sequence version replaced BA000014.7.
            Constructed (26-Dec-2000) by DDBJ.
FEATURES             Location/Qualifiers
     source           1..23403063
                     /organism="Arabidopsis thaliana"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:3702"
                     /chromosome="3"
                     /ecotype="Columbia"
CONTIG            join(AC067753.4:1..4580,AC008261.5:1145..93735,
AC010676.6:16803..58448,AC010870.6:41653..86324,
AC009325.8:5472..105543,AC010797.4:28249..89154,
AC011664.10:39082..103960,AC009755.7:28826..94369,
AC068900.3:5412..22087,AC021640.7:7198..103904,
AC018363.8:2513..100559,AC012328.6:2096..96540,
AC009895.4:2933..82289,AC009327.8:28249..89934,
AC009540.6:10210..101410,AC011698.6:13672..63654,
AC016829.7:17115..110804,AC022287.9:5541..71616,
AC011437.6:5133..95310,AC009465.7:5248..93234,
AC009177.7:946..140064,AC009606.4:15146..91924,
AC011620.8:14598..100887,AC012393.5:12002..87045,
AC013454.6:24997..62271,AC068073.1:5033..24745,
AC018907.5:5358..93478,AC011623.5:21013..86022,
AC020580.11:634..94704,AC036106.6:16992..51646,
```

❖ Then open the downloaded file on notepad.



CENTURION
UNIVERSITY
*Shaping Lives...
Empowering Communities!*

If you download the file in Fasta format

sequence.fasta - Notepad

File Edit Format View Help

```
>AB576170.1 Arabidopsis thaliana DNA, Partitivirus CP-like sequence 1 (AtPCLS1(Ler)), ecotype: Ler
CCTCGAAAAGATCTCTTCTTTCTCCACATGAAACACAGCATGGCTTGGTTCAATCAAGTAAAAGGAGTT
GCGGATGATGTTGCTGCATCCTTTGAAGGATCCGGCACCCCTTGCTGACTGCTCTCCACATGGGTTGGTTG
CTAACCAAGTCATGGTCGTTCTCTTACCCCGAACATCTCCAGAATCACCAAATTGTATCGTGATAA
GCGTGCAACCTATGAGTTCGGTTACCAGCTCAAAGCACAGTTCGCAACCTCCCTCCCTTGGCGAAGCA
CTTGCCGCTTCTCTCAGACATATATCAGGATGTTTCTAACCATCCCTTCTCGGGACGTTTGGGTCGA
AAACTCTTGATCATGGCCCGTTCTGAAAAATAAGGCCTATTGGCTCCAGCCTTACCGATAACAGCTCCTA
CCTCACTATCCCATCGATTGTCAAGCAAGCATTCAAAGGCCAGCCACGTAG
```

```
>AB605883.1 Arabidopsis thaliana DNA, corresponding sequence to flanking region of AtRE1 in ecotype C24 genome (Chromosome 2), ecotype: Nossen
TTTTAGGTTTTTGTCTTTGTCTTTTCAAGAATAGAGTCAGATCTGGCCGATCCATCCAAGG
CAGAAGCGTTTCTATCCAAATTTTCGCCCTCAGACTTCTGAAATCCACCTGAAGATCCCGATCTCGTTCT
CTCCTTCATGAGGCTCTCACTCCTTTGGGCTCTGGATGATGGTTTTATCTCATCGCCATTTCGTGTTTA
GGTTGAGTTAAGTGAGGCTTCTAGTCTCAGATCTATCAGTTTATCATTGTTTGTAGACTTTGTATATT
TCGTTTGAACCTTTGGATTTTGGGCTCGGGAAGTACTTGATCAGATGTGATGTAATTTGTAGATCGGTT
TTTTTCCCTTTGTTTCATTTAAATATGTTTTAGTTCTGGAAGGTCCCATGTACTCTCCGGCTTTTAG
AAGAGTGTCTTCTGTGACTCTTGTACTCAGATCTTCTATCAATGAAATCTATATTTCCAGTGAAAAA
AAAAATAGTTATAGATAGATGTTGGTCTGAGACTCTCCGACTCATTAGACCATCCTCATCGAATAATTTT
TCATAGATGATTTTATATAAAATAATATAAAGGTATAGATGCTTAAATGTTTTAGGAAGATATAATT
TTCATTTAATGACCCTCATATTTTTTATTTGTTTTACTCTTTTATATTTAAAAATAGCCATCTAAAAAC
ACTCATAAAAAATAACCAATGTGGATACTTAAATCAGTTTGGACATAACCAACAAAAAGGAAAGAATAA
AGAAGTAGGTGGTCGAAAAATCACTAGGATTTGATAGATCAAGCAGAAACTAAAAGGAGCGAGATTTCT
GGGAAATATCAGAGAAGCAGACGAGGTCAAGTCTATAAATCTGATGGAGAGTAATGTATAGAGGCATCA
ATTGTACACTAGAAAATCCCTAATATTACTTACTCATCTAGCTCTAAGAAGTAAGAACATCTCAGCGAAA
TCTTTGTAACTTATTTTCGGTATATAAAAGTCCAGTTTGATTCTTAAATTTGTGATCATGTTTGATTGCA
TGGAGGACTAGTTATGCGATCTTGATCCACAATAGAGTAATATTGACCAGAAAGAAATTACAAACAAG
TTACAATAGTATGACATGAAAAAATACCAATCAAAGGATCATCAACGTTGGAATGAAAAATTTTGAT
TTGTGTAACCTCATAGTCATAGGATGTGTTTCGAGGAGTGAATAATCTCCCATGCTTAATACTAGTATATAT
CTGGATGCGAAATTCGCAATTTTTTATTTTGTAGCGAAATGGTAAAAAGAGATTCAAAACATTCATCC
AGCAATCACTTAATTAATCCGAATGATTGCAAATGAAAAATAGATTCTTTATTGAACACATATGTTTACA
AAACAAGAATAGTA
```

```
>AB605847.1 Arabidopsis thaliana DNA, corresponding sequence to flanking region of AtRE1 in ecotype Columbia genome, ecotype: Nossen
AAGACTTCATATTATCCGTTTGGTCTTTGCTCTTCTTCTATGAGTGCTAAAGAAAAAAGTTCAAATT
GTCTTCTGAGTTCCCTCAAAGCGCAGGACGTCTTTAGGGTGTATGCATCTTCCAACTGGTCACCCTTCTT
CGCCATGGCCATGTAGAACCCGAAACAAAATGCTACATGGCCTGCGTTTTGGATGCACCACTGTTTGAC
CAGGCACCAAGCTTCTCCCTCGTGTTCA
```



CENTURION
UNIVERSITY
*Shaping Lives...
Empowering Communities!*

Conclusion

- GenBank is the most complete collection of annotated nucleic acid sequence data for almost every organism.
- The content includes genomic DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms.
- There is also a GenPept database for protein sequences, the majority of which are conceptual translations from DNA sequences, although a small number of the amino acid sequences are derived using peptide sequencing techniques.

Thank You



**CENTURION
UNIVERSITY**
*Shaping Lives...
Empowering Communities!*