Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 2

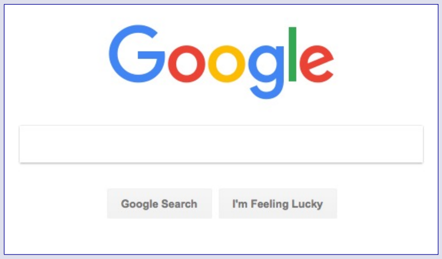
1

TODAY'S TOPICS

- ➢ Introduction to Internet Resources (NCBI, databases)
- ➢ Assignment #1

2

General Web Search

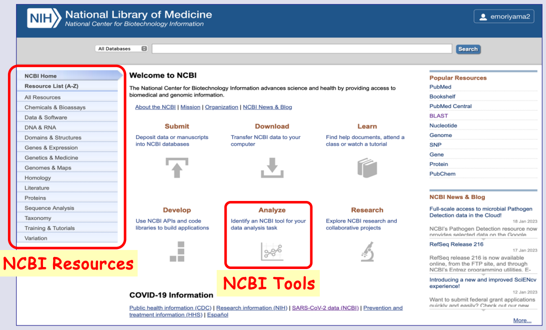


DO NOT blindly believe what you find on internet (*e.g.*, Google, Wikipedia)!
Lots of information are incorrect. Misinformation propagates rapidly!
Always double-check with the original information source (journal articles).

3

National Center for Biotechnology Information (NCBI)

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


NIH
National Library of Medicine
National Center for Biotechnology Information
emoriyama2
All Databases
Search

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

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog

Full-scale access to microbial Pathogen Detection data in the Cloud!
18 Jan 2023
NCBI's Pathogen Detection resource now provides selected data on the Google

RefSeq Release 216
17 Jan 2023
RefSeq release 216 is now available online, from the FTP site, and through NCBI's Entrez programming utilities. E-

Introducing a new and improved SciENcv experience!
12 Jan 2023
Want to submit federal grant applications quickly and easily? Check out our new

More...

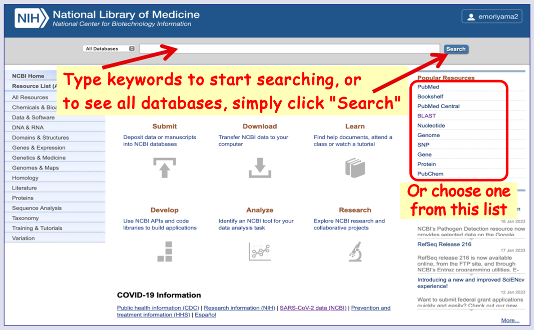
NCBI Resources

NCBI Tools

4

National Center for Biotechnology Information (NCBI)

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


NIH
National Library of Medicine
National Center for Biotechnology Information
emoriyama2
All Databases
Search

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

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

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

18 Jan 2023
NCBI's Pathogen Detection resource now provides selected data on the Google

RefSeq Release 216
17 Jan 2023
RefSeq release 216 is now available online, from the FTP site, and through NCBI's Entrez programming utilities. E-

Introducing a new and improved SciENcv experience!
12 Jan 2023
Want to submit federal grant applications quickly and easily? Check out our new

More...

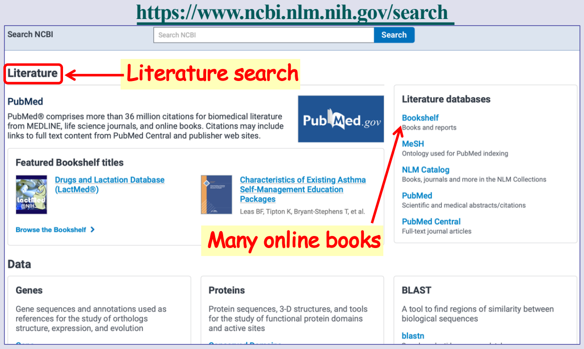
Type keywords to start searching, or to see all databases, simply click "Search"

Or choose one from this list

5

NCBI Databases

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



Literature

Literature search

PubMed

PubMed® comprises more than 36 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

Featured Bookshelf titles

Drugs and Lactation Database (LactMed®)

Characteristics of Existing Asthma Self-Management Education Packages
Leas BF, Tipton K, Bryant-Stephens T, et al.

Browse the Bookshelf ›

Literature databases

Bookshelf
Books and reports

MeSH
Ontology used for PubMed indexing

NLM Catalog
Books, journals and more in the NLM Collections

PubMed
Scientific and medical abstracts/citations

PubMed Central
Full-text journal articles

Many online books

Data

Genes
Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

Proteins
Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

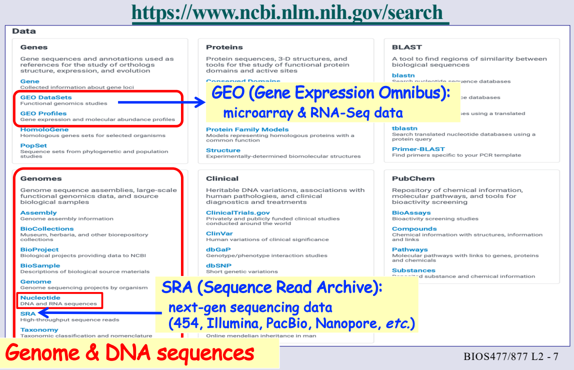
BLAST
A tool to find regions of similarity between biological sequences
blastn

Scroll down to find more databases

6

NCBI Databases

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



Data

Genes
Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

- Gene – Collected information about gene loci
- GEO DataSets – Functional genomics studies
- GEO Profiles – Gene expression and molecular abundance profiles
- HomoloGene – Homologous genes sets for selected organisms
- PopSet – Sequence sets from phylogenetic and population studies

Proteins
Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

- Conserved Domains
- Protein Family Models – Models representing homologous proteins with a common function
- Structure – Experimentally-determined biomolecular structures

BLAST
A tool to find regions of similarity between biological sequences

- blastn – Search nucleotide sequence databases
- tblastn – Search translated nucleotide databases using a protein query
- Primer-BLAST – Find primers specific to your PCR template

Genomes
Genome sequence assemblies, large-scale functional genomics data, and source biological samples

- Assembly – Genome assembly information
- BioCollections – Museum, herbaria, and other biorepository collections
- BioProject – Biological projects providing data to NCBI
- BioSample – Descriptions of biological source materials
- Genome – Genome sequencing projects by organism
- Nucleotide – DNA and RNA sequences
- SRA – High-throughput sequence reads
- Taxonomy – Taxonomic classification and nomenclature

Clinical
Heritable DNA variations, associations with human pathologies, and clinical diagnostics and treatments

- ClinicalTrials.gov – Privately and publicly funded clinical studies conducted around the world
- ClinVar – Human variations of clinical significance
- dbGaP – Genotype/phenotype interaction studies
- dbSNP – Short genetic variations
- Online mendelian inheritance in man

PubChem
Repository of chemical information, molecular pathways, and tools for bioactivity screening

- BioAssays – Bioactivity screening studies
- Compounds – Chemical information with structures, information and links
- Pathways – Molecular pathways with links to genes, proteins and chemicals
- Substances

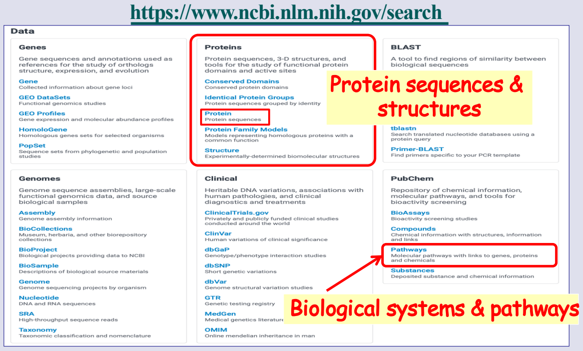
GEO (Gene Expression Omnibus): microarray & RNA-Seq data

SRA (Sequence Read Archive): next-gen sequencing data (454, Illumina, PacBio, Nanopore, *etc.*)

Genome & DNA sequences

NCBI Databases

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



Data

Genes
Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

- Gene – Collected information about gene loci
- GEO DataSets – Functional genomics studies
- GEO Profiles – Gene expression and molecular abundance profiles
- HomoloGene – Homologous genes sets for selected organisms
- PopSet – Sequence sets from phylogenetic and population studies

Proteins
Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

- Conserved Domains – Conserved protein domains
- Identical Protein Groups – Protein sequences grouped by identity
- Protein – Protein sequences
- Protein Family Models – Models representing homologous proteins with a common function
- Structure – Experimentally-determined biomolecular structures

BLAST
A tool to find regions of similarity between biological sequences

- tblastn – Search translated nucleotide databases using a protein query
- Primer-BLAST – Find primers specific to your PCR template

Genomes
Genome sequence assemblies, large-scale functional genomics data, and source biological samples

- Assembly – Genome assembly information
- BioCollections – Museum, herbaria, and other biorepository collections
- BioProject – Biological projects providing data to NCBI
- BioSample – Descriptions of biological source materials
- Genome – Genome sequencing projects by organism
- Nucleotide – DNA and RNA sequences
- SRA – High-throughput sequence reads
- Taxonomy – Taxonomic classification and nomenclature

Clinical
Heritable DNA variations, associations with human pathologies, and clinical diagnostics and treatments

- ClinicalTrials.gov – Privately and publicly funded clinical studies conducted around the world
- ClinVar – Human variations of clinical significance
- dbGaP – Genotype/phenotype interaction studies
- dbSNP – Short genetic variations
- dbVar – Genome structural variation studies
- GTR – Genetic testing registry
- MedGen – Medical genetics literature
- OMIM – Online mendelian inheritance in man

PubChem
Repository of chemical information, molecular pathways, and tools for bioactivity screening

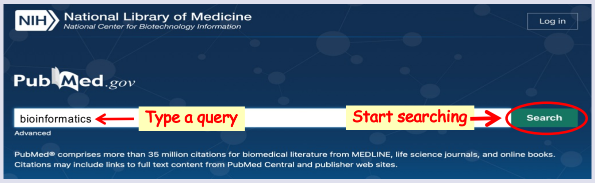
- BioAssays – Bioactivity screening studies
- Compounds – Chemical information with structures, information and links
- Pathways – Molecular pathways with links to genes, proteins and chemicals
- Substances – Deposited substance and chemical information

Protein sequences & structures

Biological systems & pathways

PubMed: Literature Search

https://pubmed.ncbi.nlm.nih.gov



NIH National Library of Medicine
National Center for Biotechnology Information
Log in
PubMed.gov
Search
Advanced

bioinformatics

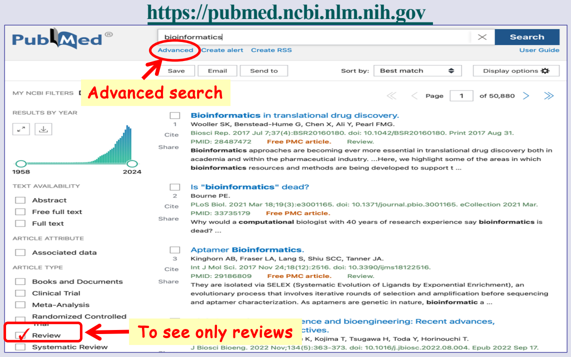
PubMed® comprises more than 35 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

Type a query

Start searching

PubMed: Literature Search

https://pubmed.ncbi.nlm.nih.gov



bioinformatics

Sort by: Best match

MY NCBI FILTERS

Page 1 of 50,880

RESULTS BY YEAR

TEXT AVAILABILITY
- □ Abstract
- □ Free full text
- □ Full text

ARTICLE ATTRIBUTE
- □ Associated data

ARTICLE TYPE
- □ Books and Documents
- □ Clinical Trial
- □ Meta-Analysis
- □ Randomized Controlled Trial
- ☑ Review
- □ Systematic Review

1 **Bioinformatics** in translational drug discovery.
Wooller SK, Benstead-Hume G, Chen X, Ali Y, Pearl FMG.
Biosci Rep. 2017 Jul 7;37(4):BSR20160180. doi: 10.1042/BSR20160180. Print 2017 Aug 31.
PMID: 28487472 Free PMC article. Review.
Bioinformatics approaches are becoming ever more essential in translational drug discovery both in academia and within the pharmaceutical industry. ...Here, we highlight some of the areas in which **bioinformatics** resources and methods are being developed to support t ...

2 Is "**bioinformatics**" dead?
Bourne PE.
PLoS Biol. 2021 Mar 18;19(3):e3001165. doi: 10.1371/journal.pbio.3001165. eCollection 2021 Mar.
PMID: 33735179 Free PMC article.
Why would a **computational** biologist with 40 years of research experience say **bioinformatics** is dead? ...

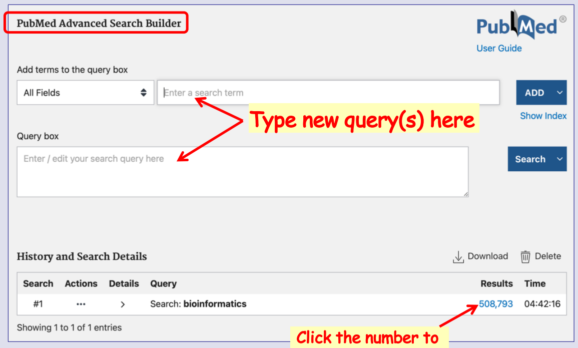
3 Aptamer **Bioinformatics**.
Kinghorn AB, Fraser LA, Lang S, Shiu SCC, Tanner JA.
Int J Mol Sci. 2017 Nov 24;18(12):2516. doi: 10.3390/ijms18122516.
PMID: 29186809 Free PMC article. Review.
They are isolated via SELEX (Systematic Evolution of Ligands by Exponential Enrichment), an evolutionary process that involves iterative rounds of selection and amplification before sequencing and aptamer characterization. As aptamers are genetic in nature, **bioinformatic** a ...

ence and bioengineering: Recent advances, ctives.
K, Kojima T, Tsugawa H, Toda Y, Horinouchi T.
J Biosci Bioeng. 2022 Nov;134(5):363-373. doi: 10.1016/j.jbiosc.2022.08.004. Epub 2022 Sep 17.

Advanced search

To see only reviews

PubMed: Literature Search



PubMed Advanced Search Builder

Add terms to the query box

All Fields | Enter a search term

Query box

Enter / edit your search query here

History and Search Details

Search	Actions	Details	Query	Results	Time
#1	•••	>	Search: **bioinformatics**	508,793	04:42:16

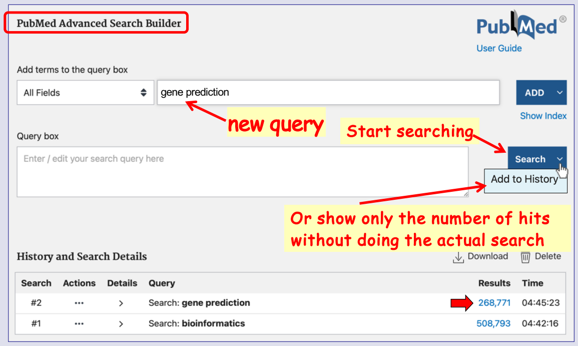
Showing 1 to 1 of 1 entries

Type new query(s) here

Click the number to see the result

https://pubmed.ncbi.nlm.nih.gov/

PubMed: Literature Search



PubMed Advanced Search Builder

Add terms to the query box

All Fields | gene prediction

Query box

Enter / edit your search query here

History and Search Details

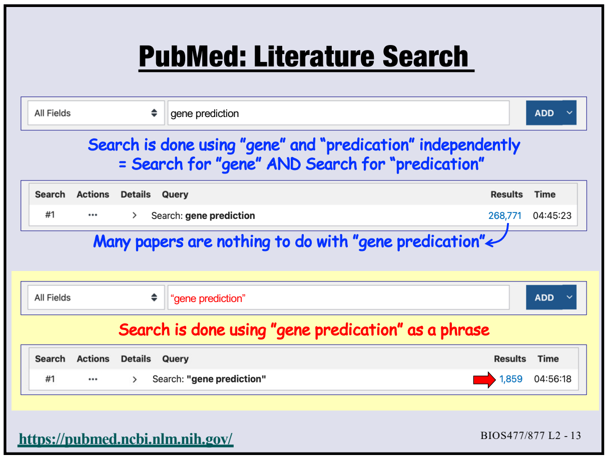
Search	Actions	Details	Query	Results	Time
#2	•••	>	Search: **gene prediction**	268,771	04:45:23
#1	•••	>	Search: **bioinformatics**	508,793	04:42:16

new query

Start searching

Or show only the number of hits without doing the actual search

https://pubmed.ncbi.nlm.nih.gov/



PubMed: Literature Search

All Fields | gene prediction | ADD

Search is done using "gene" and "predication" independently
= Search for "gene" AND Search for "predication"

Search	Actions	Details	Query	Results	Time
#1	•••	>	Search: **gene prediction**	268,771	04:45:23

Many papers are nothing to do with "gene predication"

All Fields | "gene prediction" | ADD

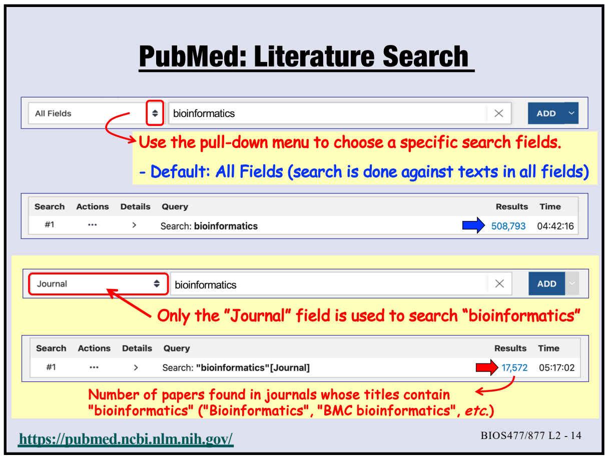
Search is done using "gene predication" as a phrase

Search	Actions	Details	Query	Results	Time
#1	•••	>	Search: **"gene prediction"**	1,859	04:56:18

https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 13

13



PubMed: Literature Search

All Fields | bioinformatics | ADD

Use the pull-down menu to choose a specific search fields.

- Default: All Fields (search is done against texts in all fields)

Search	Actions	Details	Query	Results	Time
#1	•••	>	Search: **bioinformatics**	508,793	04:42:16

Journal | bioinformatics | ADD

Only the "Journal" field is used to search "bioinformatics"

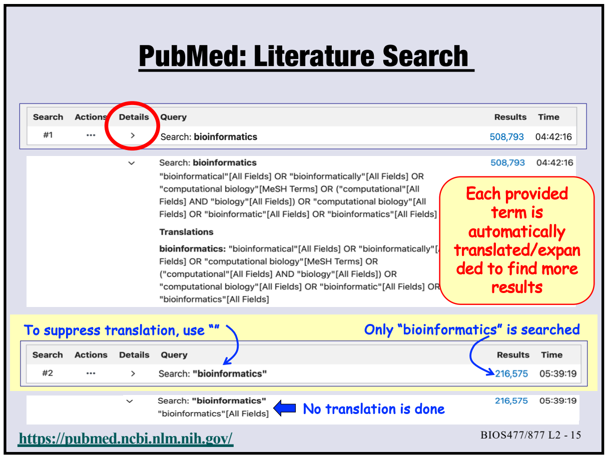
Search	Actions	Details	Query	Results	Time
#1	•••	>	Search: **"bioinformatics"[Journal]**	17,572	05:17:02

Number of papers found in journals whose titles contain
"bioinformatics" ("Bioinformatics", "BMC bioinformatics", *etc.*)

https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 14

14



PubMed: Literature Search

Search	Actions	Details	Query	Results	Time
#1	•••	>	Search: **bioinformatics**	508,793	04:42:16

Search: **bioinformatics** 508,793 04:42:16

"bioinformatical"[All Fields] OR "bioinformatically"[All Fields] OR "computational biology"[MeSH Terms] OR ("computational"[All Fields] AND "biology"[All Fields]) OR "computational biology"[All Fields] OR "bioinformatic"[All Fields] OR "bioinformatics"[All Fields]

Translations

bioinformatics: "bioinformatical"[All Fields] OR "bioinformatically"[All Fields] OR "computational biology"[MeSH Terms] OR ("computational"[All Fields] AND "biology"[All Fields]) OR "computational biology"[All Fields] OR "bioinformatic"[All Fields] OR "bioinformatics"[All Fields]

Each provided term is automatically translated/expanded to find more results

To suppress translation, use ""

Only "bioinformatics" is searched

Search	Actions	Details	Query	Results	Time
#2	•••	>	Search: **"bioinformatics"**	216,575	05:39:19

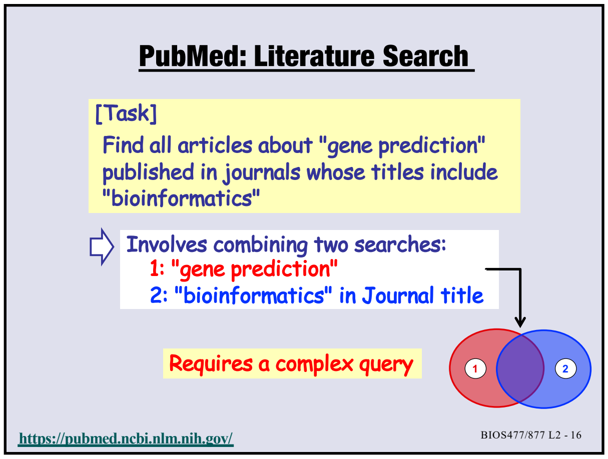
Search: **"bioinformatics"** 216,575 05:39:19
"bioinformatics"[All Fields]

No translation is done

https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 15

15



PubMed: Literature Search

[Task]
Find all articles about "gene prediction" published in journals whose titles include "bioinformatics"

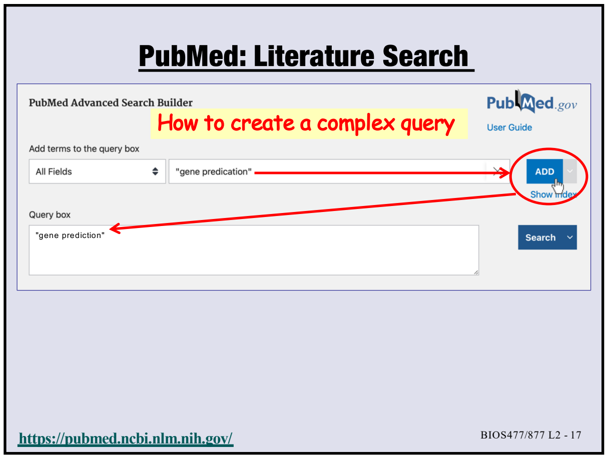
Involves combining two searches:
1: "gene prediction"
2: "bioinformatics" in Journal title

Requires a complex query

https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 16

16



PubMed: Literature Search

PubMed Advanced Search Builder

How to create a complex query

Add terms to the query box

All Fields | "gene predication" | ADD

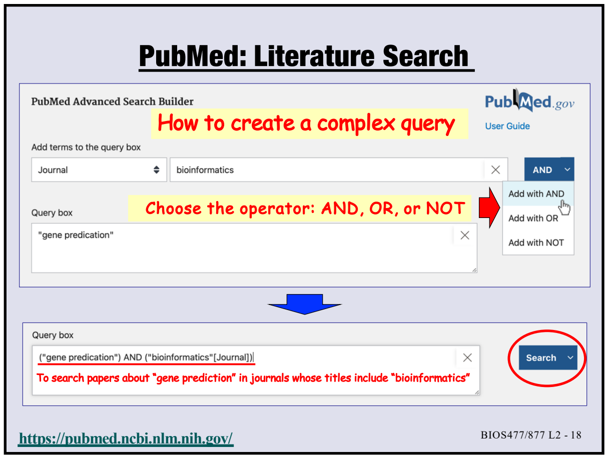
Query box

"gene prediction"

https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 17

17



PubMed: Literature Search

PubMed Advanced Search Builder

How to create a complex query

Add terms to the query box

Journal | bioinformatics | AND

Add with AND
Add with OR
Add with NOT

Choose the operator: AND, OR, or NOT

Query box

"gene predication"

Query box

("gene predication") AND ("bioinformatics"[Journal])

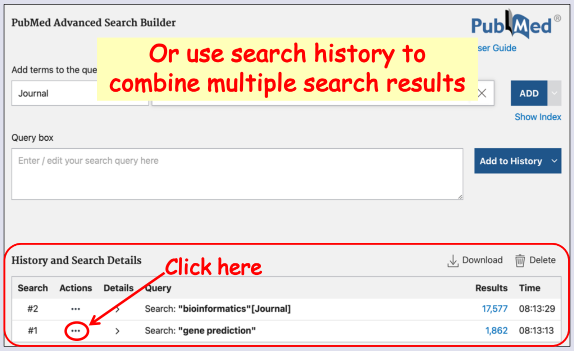
To search papers about "gene prediction" in journals whose titles include "bioinformatics"

https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 18

18

PubMed: Literature Search



PubMed Advanced Search Builder

Or use search history to combine multiple search results

Add terms to the que

Journal

Query box

Enter / edit your search query here

History and Search Details

Click here

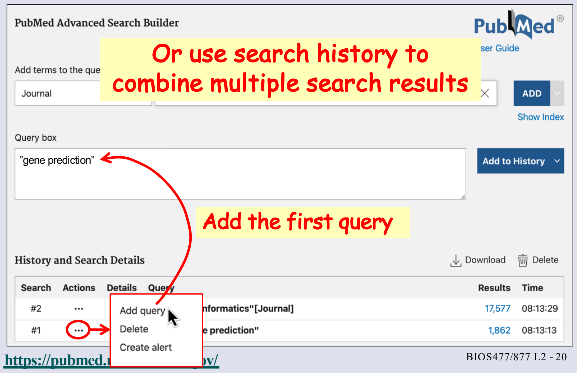
Search	Actions	Details	Query	Results	Time
#2	•••	›	Search: "bioinformatics"[Journal]	17,577	08:13:29
#1	•••	›	Search: "gene prediction"	1,862	08:13:13

https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 19

19

PubMed: Literature Search



PubMed Advanced Search Builder

Or use search history to combine multiple search results

Add terms to the que

Journal

Query box

"gene prediction"

Add the first query

History and Search Details

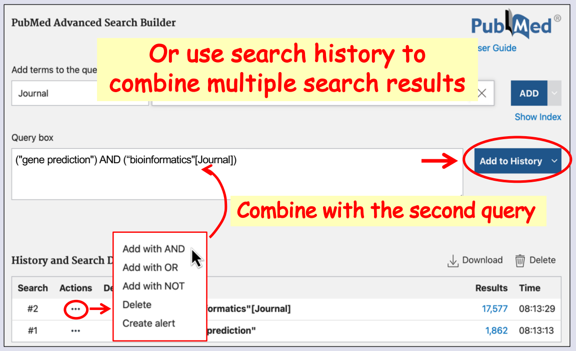
Search	Actions	Details	Query	Results	Time
#2	•••		nformatics"[Journal]	17,577	08:13:29
#1	•••		e prediction"	1,862	08:13:13

https://pubmed.n /

BIOS477/877 L2 - 20

20

PubMed: Literature Search



PubMed Advanced Search Builder

Or use search history to combine multiple search results

Add terms to the que

Journal

Query box

("gene prediction") AND ("bioinformatics"[Journal])

Combine with the second query

History and Search D

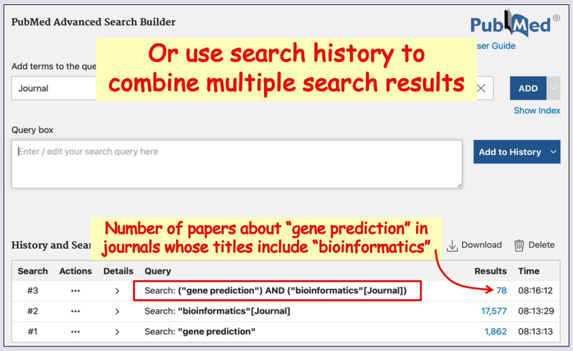
Search	Actions	De		Results	Time
#2	•••		ormatics"[Journal]	17,577	08:13:29
#1	•••		prediction"	1,862	08:13:13

https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 21

21

PubMed: Literature Search



PubMed Advanced Search Builder

Or use search history to combine multiple search results

Add terms to the que

Journal

Query box

Enter / edit your search query here

History and Sear

Number of papers about "gene prediction" in journals whose titles include "bioinformatics"

Search	Actions	Details	Query	Results	Time
#3	•••	›	Search: ("gene prediction") AND ("bioinformatics"[Journal])	78	08:16:12
#2	•••	›	Search: "bioinformatics"[Journal]	17,577	08:13:29
#1	•••	›	Search: "gene prediction"	1,862	08:13:13

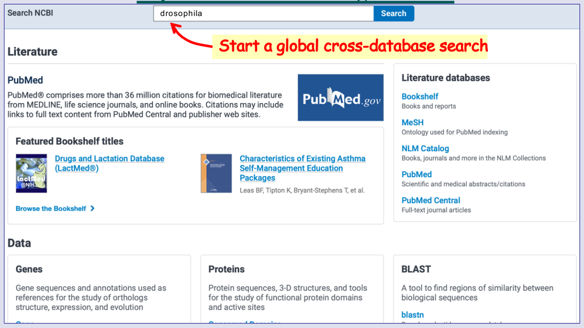
https://pubmed.ncbi.nlm.nih.gov/

BIOS477/877 L2 - 22

22

NCBI Database Search

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



Search NCBI

drosophila

Start a global cross-database search

Literature

PubMed

PubMed® comprises more than 36 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

Featured Bookshelf titles

Drugs and Lactation Database (LactMed®)

Characteristics of Existing Asthma Self-Management Education Packages

Leas BF, Tipton K, Bryant-Stephens T, et al.

Browse the Bookshelf ›

Literature databases

Bookshelf
Books and reports

MeSH
Ontology used for PubMed indexing

NLM Catalog
Books, journals and more in the NLM Collections

PubMed
Scientific and medical abstracts/citations

PubMed Central
Full-text journal articles

Data

Genes

Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

Proteins

Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

BLAST

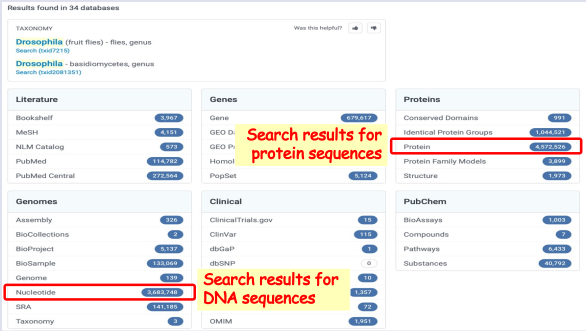
A tool to find regions of similarity between biological sequences

blastn

BIOS477/877 L2 - 23

23

NCBI Entrez Global Query



Results found in 34 databases

TAXONOMY

Drosophila (fruit flies) - flies, genus
Search (txid7215)

Drosophila - basidiomycetes, genus
Search (txid2081351)

Literature	
Bookshelf	3,967
MeSH	4,151
NLM Catalog	573
PubMed	114,782
PubMed Central	272,564

Genes	
Gene	679,617
GEO D	
GEO P	
Homol	
PopSet	5,124

Proteins	
Conserved Domains	991
Identical Protein Groups	1,044,521
Protein	4,572,526
Protein Family Models	3,899
Structure	1,973

Genomes	
Assembly	326
BioCollections	2
BioProject	5,137
BioSample	133,069
Genome	139
Nucleotide	3,683,748
SRA	141,185
Taxonomy	3

Clinical	
ClinicalTrials.gov	15
ClinVar	115
dbGaP	1
dbSNP	0
	10
	1,357
	72
OMIM	1,951

PubChem	
BioAssays	1,003
Compounds	7
Pathways	6,433
Substances	40,792

Search results for protein sequences

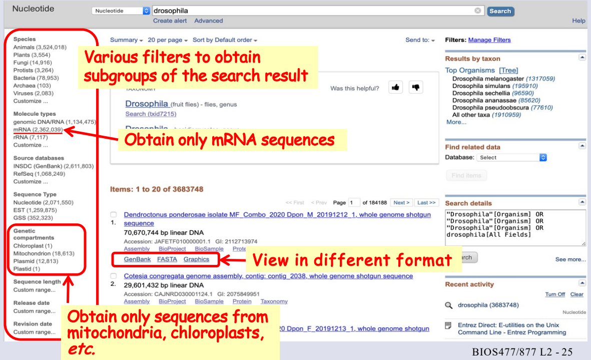
Search results for DNA sequences

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

BIOS477/877 L2 - 24

24

Nucleotide Database Search



Nucleotide | Nucleotide | drosophila | Search
Create alert Advanced Help

Species
Animals (3,524,018)
Plants (3,554)
Fungi (14,916)
Protists (3,264)
Bacteria (78,953)
Archaea (103)
Viruses (2,083)
Customize ...

Molecule types
genomic DNA/RNA (1,134,475)
mRNA (2,362,039)
rRNA (7,117)
Customize ...

Source databases
INSDC (GenBank) (2,611,803)
RefSeq (1,068,249)
Customize ...

Sequence Type
Nucleotide (2,071,550)
EST (1,259,875)
GSS (352,323)

Genetic compartments
Chloroplast (1)
Mitochondrion (18,613)
Plasmid (12,813)
Plastid (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Summary 20 per page Sort by Default order Send to:

Was this helpful?

Drosophila (fruit flies) - flies, genus
Search (txid7215)

Items: 1 to 20 of 3683748

<< First < Prev Page 1 of 184188 Next > Last >>

1. Dendroctonus ponderosae isolate MF_Combo_2020 Dpon_M_20191212_1, whole genome shotgun sequence
70,670,744 bp linear DNA
Accession: JAFETF010000001.1 GI: 2112713974
Assembly BioProject BioSample Prote
GenBank FASTA Graphics

2. Cotesia congregata genome assembly, contig: contig_2038, whole genome shotgun sequence
29,601,432 bp linear DNA
Accession: CAJNRD030001124.1 GI: 2075849951
Assembly BioProject BioSample Protein Taxonomy

0 Dpon_F_20191213_1, whole genome shotgun

Filters: Manage Filters

Results by taxon
Top Organisms [Tree]
Drosophila melanogaster (1317059)
Drosophila simulans (195910)
Drosophila sechellia (96590)
Drosophila ananassae (85620)
Drosophila pseudoobscura (77610)
All other taxa (1910959)
More...

Find related data
Database: Select
Find items

Search details

```
"Drosophila"[Organism] OR
"Drosophila"[Organism] OR
"Drosophila"[Organism] OR
drosophila[All Fields]
```

See more...

Recent activity
Turn Off Clear
drosophila (3683748) Nucleotide
Entrez Direct: E-utilities on the Unix Command Line - Entrez Programming

Various filters to obtain subgroups of the search result

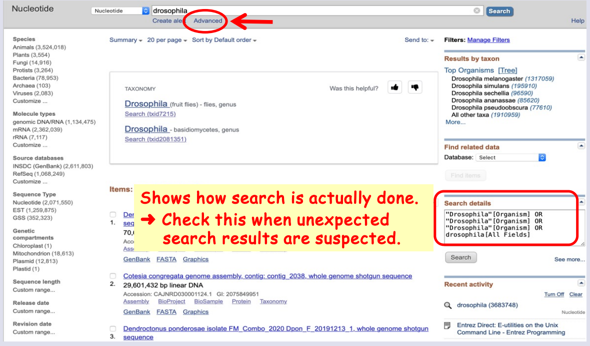
Obtain only mRNA sequences

View in different format

Obtain only sequences from mitochondria, chloroplasts, etc.

25

Nucleotide Database Search



Nucleotide | Nucleotide | drosophila | Search
Create alert Advanced Help

Species
Animals (3,524,018)
Plants (3,554)
Fungi (14,916)
Protists (3,264)
Bacteria (78,953)
Archaea (103)
Viruses (2,083)
Customize ...

Molecule types
genomic DNA/RNA (1,134,475)
mRNA (2,362,039)
rRNA (7,117)
Customize ...

Source databases
INSDC (GenBank) (2,611,803)
RefSeq (1,068,249)
Customize ...

Sequence Type
Nucleotide (2,071,550)
EST (1,259,875)
GSS (352,323)

Genetic compartments
Chloroplast (1)
Mitochondrion (18,613)
Plasmid (12,813)
Plastid (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Summary 20 per page Sort by Default order Send to:

TAXONOMY Was this helpful?

Drosophila (fruit flies) - flies, genus
Search (txid7215)

Drosophila - basidiomycetes, genus
Search (txid2081351)

Items:

1. Der seq 70, Acc Ass
GenBank FASTA Graphics

2. Cotesia congregata genome assembly, contig: contig_2038, whole genome shotgun sequence
29,601,432 bp linear DNA
Accession: CAJNRD030001124.1 GI: 2075849951
Assembly BioProject BioSample Protein Taxonomy
GenBank FASTA Graphics

3. Dendroctonus ponderosae isolate FM_Combo_2020 Dpon_F_20191213_1, whole genome shotgun sequence

Filters: Manage Filters

Results by taxon
Top Organisms [Tree]
Drosophila melanogaster (1317059)
Drosophila simulans (195910)
Drosophila sechellia (96590)
Drosophila ananassae (85620)
Drosophila pseudoobscura (77610)
All other taxa (1910959)
More...

Find related data
Database: Select
Find items

Search details

```
"Drosophila"[Organism] OR
"Drosophila"[Organism] OR
"Drosophila"[Organism] OR
drosophila[All Fields]
```

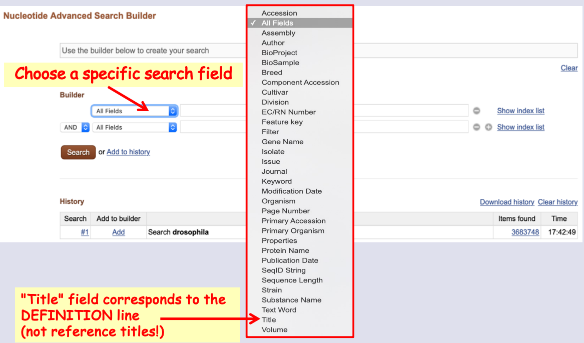
Search See more...

Recent activity
Turn Off Clear
drosophila (3683748) Nucleotide
Entrez Direct: E-utilities on the Unix Command Line - Entrez Programming

Shows how search is actually done.
➔ Check this when unexpected search results are suspected.

26

Nucleotide Database Search



Nucleotide Advanced Search Builder

Use the builder below to create your search

Clear

Builder

All Fields | Show index list
AND | All Fields | Show index list

Search or Add to history

History Download history Clear history

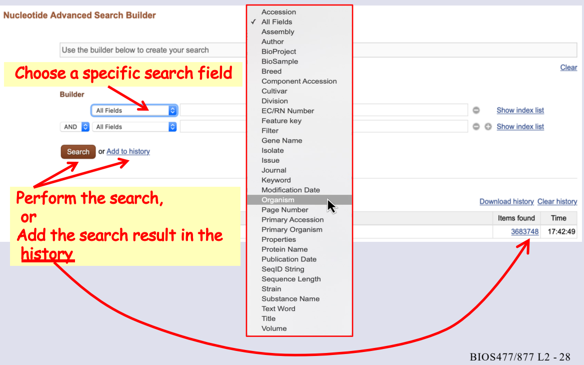
Search	Add to builder		Items found	Time
#1	Add	Search **drosophila**	3683748	17:42:49

Choose a specific search field

"Title" field corresponds to the DEFINITION line (not reference titles!)

27

Nucleotide Database Search



Nucleotide Advanced Search Builder

Use the builder below to create your search

Clear

Builder

All Fields | Show index list
AND | All Fields | Show index list

Search or Add to history

Download history Clear history

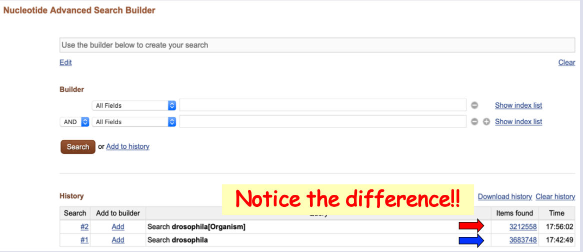
Items found	Time
3683748	17:42:49

Choose a specific search field

Perform the search,
or
Add the search result in the history

28

Nucleotide Database Search



Nucleotide Advanced Search Builder

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

History Download history Clear history

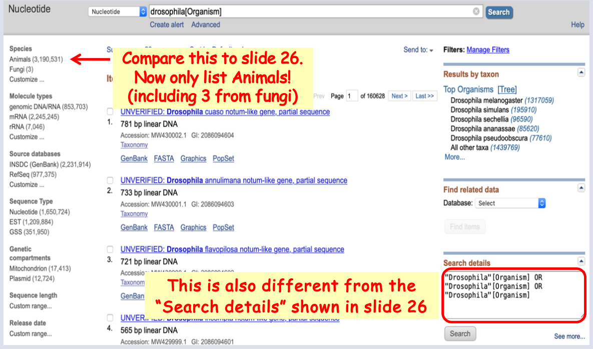
Search	Add to builder	Query	Items found	Time
#2	Add	Search **drosophila[Organism]**	3212558	17:56:02
#1	Add	Search **drosophila**	3683748	17:42:49

Notice the difference!!

#2: Searched [Organism] field
#1: Searched [All Field]

29

Nucleotide Database Search



Nucleotide | Nucleotide | drosophila[Organism] | Search
Create alert Advanced Help

Species
Animals (3,190,531)
Fungi (3)
Customize ...

Molecule types
genomic DNA/RNA (853,703)
mRNA (2,245,245)
rRNA (7,046)
Customize ...

Source databases
INSDC (GenBank) (2,231,914)
RefSeq (977,375)
Customize ...

Sequence Type
Nucleotide (1,650,724)
EST (1,209,884)
GSS (351,950)

Genetic compartments
Mitochondrion (17,413)
Plasmid (12,724)

Sequence length
Custom range...

Release date
Custom range...

Send to:

Prev Page 1 of 160628 Next > Last >>

1. UNVERIFIED: **Drosophila** cuaso notum-like gene, partial sequence
781 bp linear DNA
Accession: MW430002.1 GI: 2086094604
Taxonomy
GenBank FASTA Graphics PopSet

2. UNVERIFIED: **Drosophila** annulimana notum-like gene, partial sequence
733 bp linear DNA
Accession: MW430001.1 GI: 2086094603
Taxonomy
GenBank FASTA Graphics PopSet

3. UNVERIFIED: **Drosophila** flavopilosa notum-like gene, partial sequence
721 bp linear DNA
Accessio
Taxonom
GenBan

4. UNVER
565 bp linear DNA
Accession: MW429999.1 GI: 2086094601

Filters: Manage Filters

Results by taxon
Top Organisms [Tree]
Drosophila melanogaster (1317059)
Drosophila simulans (195910)
Drosophila sechellia (96590)
Drosophila ananassae (85620)
Drosophila pseudoobscura (77610)
All other taxa (1439769)
More...

Find related data
Database: Select
Find items

Search details

```
"Drosophila"[Organism] OR
"Drosophila"[Organism] OR
"Drosophila"[Organism]
```

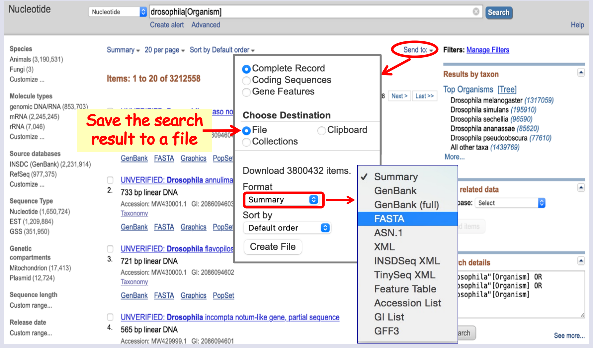
Search See more...

Compare this to slide 26.
Now only list Animals!
(including 3 from fungi)

This is also different from the "Search details" shown in slide 26

30

Nucleotide Database Search



Nucleotide | Nucleotide | drosophila[Organism] | Search
Create alert Advanced Help

Species
Animals (3,190,531)
Fungi (3)
Customize ...

Molecule types
genomic DNA/RNA (853,703)
mRNA (2,245,245)
rRNA (7,046)
Customize ...

Source databases
INSDC (GenBank) (2,231,914)
RefSeq (977,375)
Customize ...

Sequence Type
Nucleotide (1,650,724)
EST (1,209,884)
GSS (351,950)

Genetic compartments
Mitochondrion (17,413)
Plasmid (12,724)

Sequence length
Custom range...

Release date
Custom range...

Summary ▾ 20 per page ▾ Sort by Default order ▾

Items: 1 to 20 of 3212558

2. UNVERIFIED: Drosophila annulima...
733 bp linear DNA
Accession: MW430001.1 GI: 2086094603
Taxonomy
GenBank FASTA Graphics PopSet

3. UNVERIFIED: Drosophila flavopilos...
721 bp linear DNA
Accession: MW430000.1 GI: 2086094602
Taxonomy
GenBank FASTA Graphics PopSet

4. UNVERIFIED: Drosophila incompta notum-like gene, partial sequence
565 bp linear DNA
Accession: MW429999.1 GI: 2086094601

Send to: ▾

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination
● File ○ Clipboard
○ Collections

Download 3800432 items.
Format: Summary
Sort by: Default order
Create File

Filters: Manage Filters

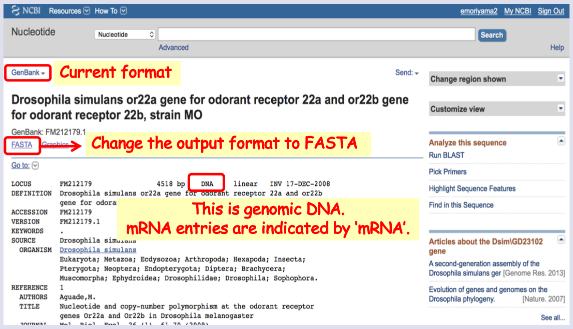
Results by taxon
Top Organisms [Tree]
Drosophila melanogaster (1317059)
Drosophila simulans (195910)
Drosophila sechellia (96590)
Drosophila ananassae (85620)
Drosophila pseudoobscura (77610)
All other taxa (1439769)
More...

Save the search result to a file

BIOS477/877 L2 - 31

31

GenBank Entry



NCBI Resources How To emoriyama2 My NCBI Sign Out
Nucleotide | Nucleotide | Search
Advanced Help

GenBank ▾ — Current format

Send: ▾

Drosophila simulans or22a gene for odorant receptor 22a and or22b gene for odorant receptor 22b, strain MO

GenBank: FM212179.1
FASTA Graphics — Change the output format to FASTA

Go to:

```
LOCUS       FM212179                4518 bp    DNA     linear   INV 17-DEC-2008
DEFINITION  Drosophila simulans or22a gene for odorant receptor 22a and or22b
            gene for odora
ACCESSION   FM212179
VERSION     FM212179.1
KEYWORDS    .
SOURCE      Drosophila simulans
  ORGANISM  Drosophila simulans
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1
  AUTHORS   Aguade,M.
  TITLE     Nucleotide and copy-number polymorphism at the odorant receptor
            genes Or22a and Or22b in Drosophila melanogaster
```

This is genomic DNA.
mRNA entries are indicated by 'mRNA'.

Change region shown
Customize view

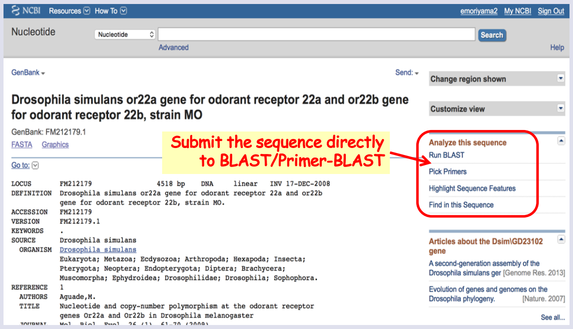
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

Articles about the Dsim\GD23102 gene
A second-generation assembly of the Drosophila simulans ger [Genome Res. 2013]
Evolution of genes and genomes on the Drosophila phylogeny. [Nature. 2007]
See all...

BIOS477/877 L2 - 32

32

GenBank Entry



NCBI Resources How To emoriyama2 My NCBI Sign Out
Nucleotide | Nucleotide | Search
Advanced Help

GenBank ▾ Send: ▾

Drosophila simulans or22a gene for odorant receptor 22a and or22b gene for odorant receptor 22b, strain MO

GenBank: FM212179.1
FASTA Graphics

Go to:

```
LOCUS       FM212179                4518 bp    DNA     linear   INV 17-DEC-2008
DEFINITION  Drosophila simulans or22a gene for odorant receptor 22a and or22b
            gene for odorant receptor 22b, strain MO.
ACCESSION   FM212179
VERSION     FM212179.1
KEYWORDS    .
SOURCE      Drosophila simulans
  ORGANISM  Drosophila simulans
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1
  AUTHORS   Aguade,M.
  TITLE     Nucleotide and copy-number polymorphism at the odorant receptor
            genes Or22a and Or22b in Drosophila melanogaster
```

Submit the sequence directly to BLAST/Primer-BLAST

Change region shown
Customize view

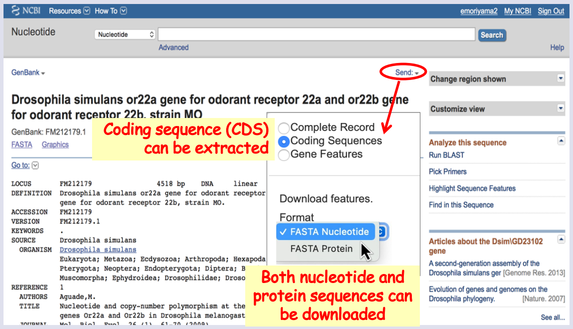
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

Articles about the Dsim\GD23102 gene
A second-generation assembly of the Drosophila simulans ger [Genome Res. 2013]
Evolution of genes and genomes on the Drosophila phylogeny. [Nature. 2007]
See all...

BIOS477/877 L2 - 33

33

GenBank Entry



NCBI Resources How To emoriyama2 My NCBI Sign Out
Nucleotide | Nucleotide | Search
Advanced Help

GenBank ▾ Send: ▾

Drosophila simulans or22a gene for odorant receptor 22a and or22b gene for odorant receptor 22b, strain MO

GenBank: FM212179.1
FASTA Graphics

Go to:

```
LOCUS       FM212179                4518 bp    DNA     linear
DEFINITION  Drosophila simulans or22a gene for odorant receptor
            gene for odorant receptor 22b, strain MO.
ACCESSION   FM212179
VERSION     FM212179.1
KEYWORDS    .
SOURCE      Drosophila simulans
  ORGANISM  Drosophila simulans
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda
            Pterygota; Neoptera; Endopterygota; Diptera; B
            Muscomorpha; Ephydroidea; Drosophilidae; Droso
REFERENCE   1
  AUTHORS   Aguade,M.
  TITLE     Nucleotide and copy-number polymorphism at the
            genes Or22a and Or22b in Drosophila melanogast
```

○ Complete Record
● Coding Sequences
○ Gene Features

Download features.
Format

Coding sequence (CDS) can be extracted

Both nucleotide and protein sequences can be downloaded

Change region shown
Customize view

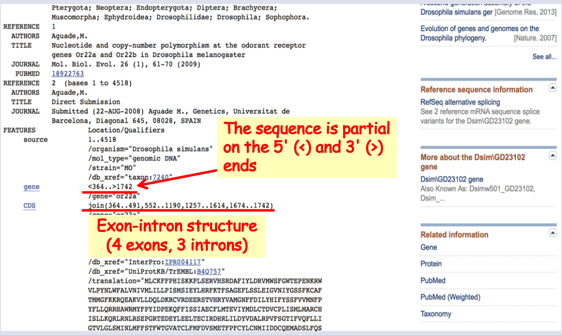
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

Articles about the Dsim\GD23102 gene
A second-generation assembly of the Drosophila simulans ger [Genome Res. 2013]
Evolution of genes and genomes on the Drosophila phylogeny. [Nature. 2007]
See all...

BIOS477/877 L2 - 34

34

GenBank Entry



```
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1
  AUTHORS   Aguade,M.
  TITLE     Nucleotide and copy-number polymorphism at the odorant receptor
            genes Or22a and Or22b in Drosophila melanogaster
  JOURNAL   Mol. Biol. Evol. 26 (1), 61-70 (2009)
   PUBMED   18922763
REFERENCE   2  (bases 1 to 4518)
  AUTHORS   Aguade,M.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-AUG-2008) Aguade M., Genetics, Universitat de
            Barcelona, Diagonal 645, 08028, SPAIN
FEATURES             Location/Qualifiers
     source          1..4518
                     /organism="Drosophila simulans"
                     /mol_type="genomic DNA"
                     /strain="MO"
                     /db_xref="taxon:7240"
     gene            <364..>1742
                     /gene="or22a"
     CDS             join(364..491,552..1190,1257..1614,1674..1742)

                     /db_xref="InterPro:IPR004117"
                     /db_xref="UniProtKB/TrEMBL:B4Q757"
                     /translation="MLCKFFPHISKKPLSERVHSRDAFIYLDRVMWSFGWTEPENKRW
                     VLPYNLWFALVNIVMLILLPISMSIEYLHRFKTFSAGEFLSSLEIGVNIYGSSFKCAF
                     TMMGFKKRQEAKVLLDQLDKRCVRDEERSTVHRYVAMGNFFDILYHIFYSSFVVMNFP
                     YFLLQRRHAWRMYFPYIDPEKQFFISSIAECFLMTEVIYMDLCTDVCPLISMLMARCH
                     ISLLKQRLRNLRSEPGRTEDEYLEELTECIRDHRLILDYVDALRPVFSGTIFVQFLLI
                     GTVLGLSMINLMFFSTFWTGVATCLFMFDVSMETFPFCYLCNMIIDDCQEMADSLFQS
```

The sequence is partial on the 5' (<) and 3' (>) ends

Exon-intron structure (4 exons, 3 introns)

Drosophila simulans ger [Genome Res. 2013]
Evolution of genes and genomes on the Drosophila phylogeny. [Nature. 2007]
See all...

Reference sequence information
RefSeq alternative splicing
See 2 reference mRNA sequence splice variants for the Dsim\GD23102 gene.

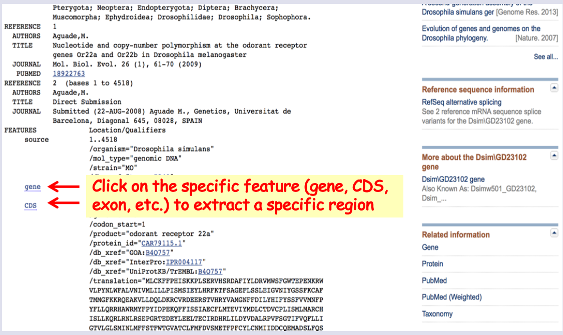
More about the Dsim\GD23102 gene
Dsim\GD23102 gene
Also Known As: Dsimw501_GD23102, Dsim_...

Related information
Gene
Protein
PubMed
PubMed (Weighted)
Taxonomy

BIOS477/877 L2 - 35

35

GenBank Entry



```
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1
  AUTHORS   Aguade,M.
  TITLE     Nucleotide and copy-number polymorphism at the odorant receptor
            genes Or22a and Or22b in Drosophila melanogaster
  JOURNAL   Mol. Biol. Evol. 26 (1), 61-70 (2009)
   PUBMED   18922763
REFERENCE   2  (bases 1 to 4518)
  AUTHORS   Aguade,M.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-AUG-2008) Aguade M., Genetics, Universitat de
            Barcelona, Diagonal 645, 08028, SPAIN
FEATURES             Location/Qualifiers
     source          1..4518
                     /organism="Drosophila simulans"
                     /mol_type="genomic DNA"
                     /strain="MO"
     gene
     CDS
                     /codon_start=1
                     /product="odorant receptor 22a"
                     /protein_id="CAR79115.1"
                     /db_xref="GOA:B4Q757"
                     /db_xref="InterPro:IPR004117"
                     /db_xref="UniProtKB/TrEMBL:B4Q757"
                     /translation="MLCKFFPHISKKPLSERVHSRDAFIYLDRVMWSFGWTEPENKRW
                     VLPYNLWFALVNIVMLILLPISMSIEYLHRFKTFSAGEFLSSLEIGVNIYGSSFKCAF
                     TMMGFKKRQEAKVLLDQLDKRCVRDEERSTVHRYVAMGNFFDILYHIFYSSFVVMNFP
                     YFLLQRRHAWRMYFPYIDPEKQFFISSIAECFLMTEVIYMDLCTDVCPLISMLMARCH
                     ISLLKQRLRNLRSEPGRTEDEYLEELTECIRDHRLILDYVDALRPVFSGTIFVQFLLI
                     GTVLGLSMINLMFFSTFWTGVATCLFMFDVSMETFPFCYLCNMIIDDCQEMADSLFQS
```

Click on the specific feature (gene, CDS, exon, etc.) to extract a specific region

Drosophila simulans ger [Genome Res. 2013]
Evolution of genes and genomes on the Drosophila phylogeny. [Nature. 2007]
See all...

Reference sequence information
RefSeq alternative splicing
See 2 reference mRNA sequence splice variants for the Dsim\GD23102 gene.

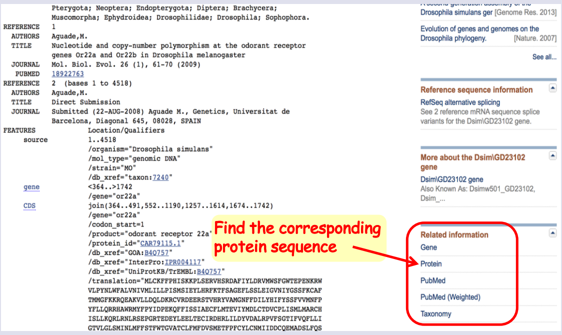
More about the Dsim\GD23102 gene
Dsim\GD23102 gene
Also Known As: Dsimw501_GD23102, Dsim_...

Related information
Gene
Protein
PubMed
PubMed (Weighted)
Taxonomy

BIOS477/877 L2 - 36

36

GenBank Entry

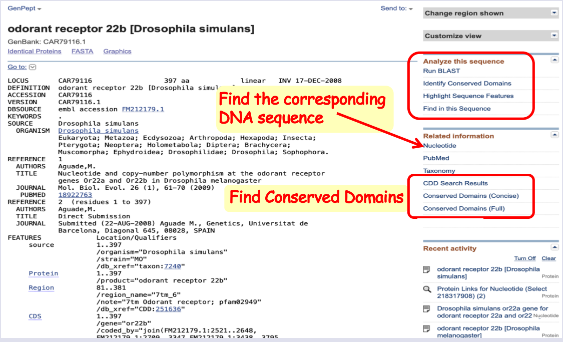


```
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1
  AUTHORS   Aguade,M.
  TITLE     Nucleotide and copy-number polymorphism at the odorant receptor
            genes Or22a and Or22b in Drosophila melanogaster
  JOURNAL   Mol. Biol. Evol. 26 (1), 61-70 (2009)
   PUBMED   18922763
REFERENCE   2  (bases 1 to 4518)
  AUTHORS   Aguade,M.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-AUG-2008) Aguade M., Genetics, Universitat de
            Barcelona, Diagonal 645, 08028, SPAIN
FEATURES             Location/Qualifiers
     source          1..4518
                     /organism="Drosophila simulans"
                     /mol_type="genomic DNA"
                     /strain="MO"
                     /db_xref="taxon:7240"
     gene            <364..>1742
                     /gene="or22a"
     CDS             join(364..491,552..1190,1257..1614,1674..1742)
                     /gene="or22a"
                     /codon_start=1
                     /product="odorant receptor 22a"
                     /protein_id="CAR79115.1"
                     /db_xref="GOA:B4Q757"
                     /db_xref="InterPro:IPR004117"
                     /db_xref="UniProtKB/TrEMBL:B4Q757"
                     /translation="MLCKFFPHISKKPLSERVHSRDAFIYLDRVMWSFGWTEPENKRW
                     VLPYNLWFALVNIVMLILLPISMSIEYLHRFKTFSAGEFLSSLEIGVNIYGSSFKCAF
                     TMMGFKKRQEAKVLLDQLDKRCVRDEERSTVHRYVAMGNFFDILYHIFYSSFVVMNFP
                     YFLLQRRHAWRMYFPYIDPEKQFFISSIAECFLMTEVIYMDLCTDVCPLISMLMARCH
                     ISLLKQRLRNLRSEPGRTEDEYLEELTECIRDHRLILDYVDALRPVFSGTIFVQFLLI
                     GTVLGLSMINLMFFSTFWTGVATCLFMFDVSMETFPFCYLCNMIIDDCQEMADSLFQS
```

A second generation assembly of the
Drosophila simulans ger [Genome Res. 2013]
Evolution of genes and genomes on the Drosophila phylogeny. [Nature. 2007]
See all...
Reference sequence information
RefSeq alternative splicing
See 2 reference mRNA sequence splice variants for the Dsim\GD23102 gene.
More about the Dsim\GD23102 gene
Dsim\GD23102 gene
Also Known As: Dsimw501_GD23102, Dsim_...
Related information
Gene
Protein
PubMed
PubMed (Weighted)
Taxonomy

Find the corresponding protein sequence

GenPept Entry (tranlasted GenBank)



odorant receptor 22b [Drosophila simulans]

GenBank: CAR79116.1

```
LOCUS       CAR79116                 397 aa            linear   INV 17-DEC-2008
DEFINITION  odorant receptor 22b [Drosophila simul
ACCESSION   CAR79116
VERSION     CAR79116.1
DBSOURCE    embl accession FM212179.1
KEYWORDS    .
SOURCE      Drosophila simulans
  ORGANISM  Drosophila simulans
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Holometabola; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1
  AUTHORS   Aguade,M.
  TITLE     Nucleotide and copy-number polymorphism at the odorant receptor
            genes Or22a and Or22b in Drosophila melanogaster
  JOURNAL   Mol. Biol. Evol. 26 (1), 61-70 (2009)
   PUBMED   18922763
REFERENCE   2  (residues 1 to 397)
  AUTHORS   Aguade,M.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-AUG-2008) Aguade M., Genetics, Universitat de
            Barcelona, Diagonal 645, 08028, SPAIN
FEATURES             Location/Qualifiers
     source          1..397
                     /organism="Drosophila simulans"
                     /strain="MO"
                     /db_xref="taxon:7240"
     Protein         1..397
                     /product="odorant receptor 22b"
     Region          81..381
                     /region_name="7tm_6"
                     /note="7tm Odorant receptor; pfam02949"
                     /db_xref="CDD:251636"
     CDS             1..397
                     /gene="or22b"
                     /coded_by="join(FM212179.1:2521..2648,
                     FM212179.1:2709..3347,FM212179.1:3438..3795
```

Analyze this sequence
Run BLAST
Identify Conserved Domains
Highlight Sequence Features
Find in this Sequence
Related information
Nucleotide
PubMed
Taxonomy
CDD Search Results
Conserved Domains (Concise)
Conserved Domains (Full)
Recent activity
Turn Off
Clear
odorant receptor 22b [Drosophila simulans] Protein
Protein Links for Nucleotide (Select 218317908) (2) Protein
Drosophila simulans or22a gene for odorant receptor 22a and or22 Nucleotide
odorant receptor 22b [Drosophila melanogaster] Protein

Find the corresponding DNA sequence

Find Conserved Domains

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

Sequence Formats

➢ **FASTA**: a standard format used in many programs

```
>AAK58030.1 glycoprotein coupled receptor [Human betaherpesvirus 5]
MTPTTTTAELTTEFDYDEDATPCVFTDVLNQSKPVTLFLYGVVFLFGSIGNFLVIFTITWRRRIQCSGDV
YFINLAAADLLFVCTLPLWMQYLLDHNSLASVPCTLLTACFYVAMFASLCFITEIALDRYYAIVYMRYRP
VKQACLFSIFWWIFAVIIAIPHFMVVTKKDNQCMTDYDYLEVSYPIILNVELMLGAFVIPLSVISYCYYR
ISRIVAVSQSRHKGRIVRVLIAVVLVFIIFWLPYHLTLFVDTLKLLKWISSSCEFERSLKRALILTESLA
FCHCCLNPLLYVFVGTKFRQELHCLLAEFRQRLFSRDVSWYHSMSFSRRSSPSRRETSSDTLSDEVCRVS
QIIP
```

- A sequence begins with a single-line description, followed by lines of sequence data
- Up to the first space is usually considered as the sequence id
- The description line is distinguished from the sequence data by a greater-than (>) symbol

FASTA Formats

[Single sequence FASTA file]

```
>P29627
FGNLSSAQAIMGNPRIRAHGKKVLTSLGLAVQNMDNL
KETFAHLSELHCDKLHVDPENFKLLGNVLVIVLSTHF
AKEFTPEVQAAWQKLVAGVANALSHKYH
```

[Multiple sequence FASTA file]

```
>HBB_HORSE
VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWT
QRFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVH
HLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVL
ARHFGKDFTPELQASYQKVVAGVANALAHKYH
>HBA_HUMAN
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPT
TKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDM
PNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLP
AEFTPAVHASLDKFLASVSTVLTSKYR
>HBA_HORSE
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPT
TKTYFPHFDLSHGSAQVKAHGKKVGDALTLAVGHLDDL
PGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLP
NDFTPAVHASLDKFLSSVSTVLTSKYR
>GLB5_PETMA
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDIL
VKFFTSTPAAQEFFPKFKGLTTADQLKKSADVRWHAER
IINAVNDAVASMDDTEKMSMKLRDLSGKHAKSFQVDPQ
YFKVLAAVIADTVAAGDAGFEKLMSMICILLRSAY
```

[Alignment FASTA file]

```
>HBB_HORSE
--------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLV
VYPWTQRFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVH
HLDN-----LKGTFAALSELHCDKLHVDPENFRLLGNVLVVVL
ARHFGKDFTPELQASYQKVVAGVANALAHKYH------
>HBA_HUMAN
---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFL
SFPTTKTYFPHF-DLS-----HGSAQVKGHGKKVADALTNAVA
HVDD-----MPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTL
AAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
>HBA_HORSE
---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFL
GFPTTKTYFPHF-DLS-----HGSAQVKAHGKKVGDALTLAVG
HLDD-----LPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTL
AVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
>GLB5_PETMA
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFT
STPAAQEFFPKFKGLTTADQLKKSADVRWHAERIINAVNDAVA
SMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKVLAAVIADTV
AAG---------DAGFEKLMSMICILLRSAY-------
```

Sequence Databases: DNA

➢ INSDC: International Nucleotide Sequence Database Collaboration

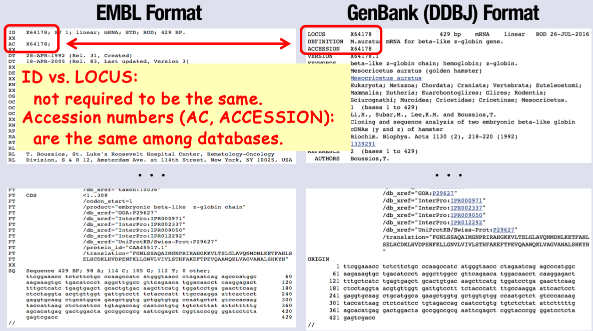
- **GenBank**: NIH genetic sequence database
  https://www.ncbi.nlm.nih.gov/genbank/
- **ENA**: European Nucleotide Archive
  https://www.ebi.ac.uk/ena/browser/home
- **DDBJ**: DNA Data Bank of Japan
  https://www.ddbj.nig.ac.jp/index-e.html

➜ Data are shared among the three databases
➜ They share the accession numbers!
➜ They use different formats

Sequence Databases: DNA

EMBL Format

GenBank (DDBJ) Format



```
ID   X64178; SV 1; linear; mRNA; STD; ROD; 429 BP.
XX
AC   X64178;
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   18-APR-2005 (Rel. 83, Last updated, Version 3)
XX
DE
XX
KW
XX
OS
OC
OC
OC
XX
RN
RP
RA
RT
RL
RL   T. Boussios, St. Luke's Roosevelt Hospital Center, Hematology-Oncology
RL   Division, S & R 12, Amsterdam Ave. at 114th Street, New York, NY 10025, USA
```

. . .

```
FT                   /db_xref="taxon:10036"
FT   CDS             <1..309
FT                   /codon_start=1
FT                   /product="embryonic beta-like  z-globin chain"
FT                   /db_xref="GOA:P29627"
FT                   /db_xref="InterPro:IPR000971"
FT                   /db_xref="InterPro:IPR002337"
FT                   /db_xref="InterPro:IPR009050"
FT                   /db_xref="InterPro:IPR012292"
FT                   /db_xref="UniProtKB/Swiss-Prot:P29627"
FT                   /protein_id="CAA45517.1"
FT                   /translation="FGNLSSAQAIMGNPRIRAHGKKVLTSLGLAVQNMDNLKETFAHLS
FT                   ELHCDKLHVDPENFKLLGNVLVIVLSTHFAKEFTPEVQAAWQKLVAGVANALSHKYH"
XX
SQ   Sequence 429 BP; 98 A; 114 C; 105 G; 112 T; 0 other;
     ttcggaaacc tctcttctgc ccaagccatc atgggtaacc ctagaatcag agcccatggc        60
     aagaaagtgc tgacatccct aggcttggcc gttcagaaca tggacaacct caaggagact       120
     tttgctcatc tgagtgagct gcactgtgac aagcttcatg tggatcctga gaacttcaag       180
     ctcctaggta acgtgttggt gattgtcctt tctacccatt ttgccaagga attcactcct       240
     gaggtgcaag ctgcatggca gaagctggtg gctggtgtgg ccaatgctct gtcccacaag       300
     taccattaag ctctcattcc tgtagaccag caatcctgtg tgtctcttat attctttttg       360
     agcacatgag gactggacta gccggccgcg aattcgagct cggtacccgg ggatcctcta       420
     gagtcgacc                                                               429
//
```

```
LOCUS       X64178                   429 bp    mRNA    linear   ROD 26-JUL-2016
DEFINITION  M.auratus mRNA for beta-like z-globin gene.
ACCESSION   X64178
VERSION     X64178.1
KEYWORDS    beta-like z-globin chain; hemoglobin; z-globin.
            Mesocricetus auratus (golden hamster)
            Mesocricetus auratus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Cricetidae; Cricetinae; Mesocricetus.
            1  (bases 1 to 429)
            Li,H., Subar,M., Lee,K.M. and Boussios,T.
            Cloning and sequence analysis of two embryonic beta-like globin
            cDNAs (y and z) of hamster
            Biochim. Biophys. Acta 1130 (2), 218-220 (1992)
            1339291
REFERENCE   2  (bases 1 to 429)
  AUTHORS   Boussios,T.
```

. . .

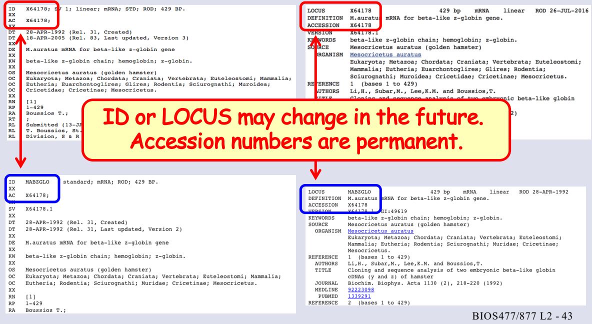
```
                     /db_xref="GOA:P29627"
                     /db_xref="InterPro:IPR000971"
                     /db_xref="InterPro:IPR002337"
                     /db_xref="InterPro:IPR009050"
                     /db_xref="InterPro:IPR012292"
                     /db_xref="UniProtKB/Swiss-Prot:P29627"
                     /translation="FGNLSSAQAIMGNPRIRAHGKKVLTSLGLAVQNMDNLKETFAHL
                     SELHCDKLHVDPENFKLLGNVLVIVLSTHFAKEFTPEVQAAWQKLVAGVANALSHKYH
                     "
ORIGIN
        1 ttcggaaacc tctcttctgc ccaagccatc atgggtaacc ctagaatcag agcccatggc
       61 aagaaagtgc tgacatccct aggcttggcc gttcagaaca tggacaacct caaggagact
      121 tttgctcatc tgagtgagct gcactgtgac aagcttcatg tggatcctga gaacttcaag
      181 ctcctaggta acgtgttggt gattgtcctt tctacccatt ttgccaagga attcactcct
      241 gaggtgcaag ctgcatggca gaagctggtg gctggtgtgg ccaatgctct gtcccacaag
      301 taccattaag ctctcattcc tgtagaccag caatcctgtg tgtctcttat attctttttg
      361 agcacatgag gactggacta gccggccgcg aattcgagct cggtacccgg ggatcctcta
      421 gagtcgacc
//
```

ID vs. LOCUS:
not required to be the same.
Accession numbers (AC, ACCESSION):
are the same among databases.

Sequence Databases: DNA

EMBL Format

GenBank (DDBJ) Format



```
ID   X64178; SV 1; linear; mRNA; STD; ROD; 429 BP.
XX
AC   X64178;
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   18-APR-2005 (Rel. 83, Last updated, Version 3)
XX
DE   M.auratus mRNA for beta-like z-globin gene
XX
KW   beta-like z-globin chain; hemoglobin; z-globin.
XX
OS   Mesocricetus auratus (golden hamster)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC   Cricetidae; Cricetinae; Mesocricetus.
XX
RN   [1]
RP   1-429
RA   Boussios T.;
RT   ;
RL   Submitted (13-JA
RL   T. Boussios, St.
RL   Division, S & R
```

```
LOCUS       X64178                   429 bp    mRNA    linear   ROD 26-JUL-2016
DEFINITION  M.auratus mRNA for beta-like z-globin gene.
ACCESSION   X64178
VERSION     X64178.1
KEYWORDS    beta-like z-globin chain; hemoglobin; z-globin.
SOURCE      Mesocricetus auratus (golden hamster)
  ORGANISM  Mesocricetus auratus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Cricetidae; Cricetinae; Mesocricetus.
REFERENCE   1  (bases 1 to 429)
  AUTHORS   Li,H., Subar,M., Lee,K.M. and Boussios,T.
  TITLE     Cloning and sequence analysis of two embryonic beta-like globin
                                                                  2)
```

ID or LOCUS may change in the future.
Accession numbers are permanent.

```
ID   MABZGLO    standard; mRNA; ROD; 429 BP.
XX
AC   X64178;
XX
SV   X64178.1
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   28-APR-1992 (Rel. 31, Last updated, Version 2)
XX
DE   M.auratus mRNA for beta-like z-globin gene
XX
KW   beta-like z-globin chain; hemoglobin; z-globin.
XX
OS   Mesocricetus auratus (golden hamster)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Cricetinae; Mesocricetus.
XX
RN   [1]
RP   1-429
RA   Boussios T.;
```

```
LOCUS       MABZGLO                  429 bp    mRNA    linear   ROD 28-APR-1992
DEFINITION  M.auratus mRNA for beta-like z-globin gene.
ACCESSION   X64178
VERSION     X64178.1  GI:49619
KEYWORDS    beta-like z-globin chain; hemoglobin; z-globin.
SOURCE      Mesocricetus auratus (golden hamster)
  ORGANISM  Mesocricetus auratus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Cricetinae;
            Mesocricetus.
REFERENCE   1  (bases 1 to 429)
  AUTHORS   Li,H., Subar,M., Lee,K.M. and Boussios,T.
  TITLE     Cloning and sequence analysis of two embryonic beta-like globin
            cDNAs (y and z) of hamster
  JOURNAL   Biochim. Biophys. Acta 1130 (2), 218-220 (1992)
  MEDLINE   92223098
   PUBMED   1339291
REFERENCE   2  (bases 1 to 429)
```

Sequence Databases: Protein

- Protein Databases
  - UniProt the Universal Protein Resource

    https://www.uniprot.org/

    ➜ Collaboration between European Bioinformatics Institute (EMBL-EBI), Swiss Institute of Bioinformatics (SIB), and Protein information Resource (PIR, https://proteininformationresource.org/)

    ➜ UniProtKB Protein Knowledgebase

    Two sections:

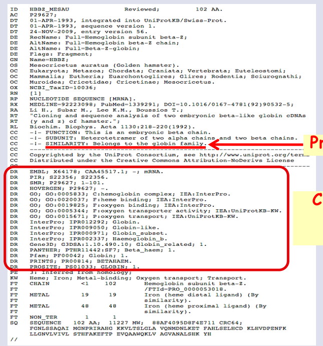
    ◆ Swiss-Prot (manually annotated and reviewed)

    ◆ TrEMBL (automatically annotated and NOT reviewed)

Protein databases have more information
(functions, domains, superfamily, cross-references, etc.)

Sequence Databases: Protein

UniProt Format (similar to EMBL format; used by all protein databases)

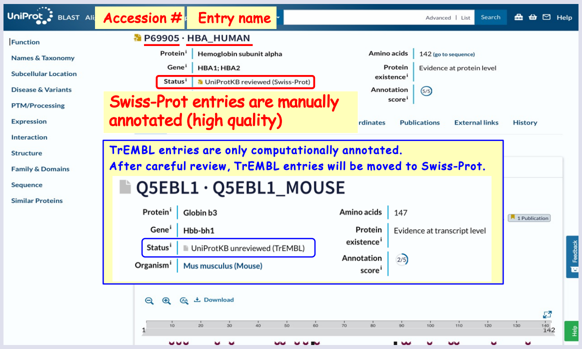


```
ID   HBBZ_MESAU              Reviewed;         102 AA.
AC   P29627;
DT   01-APR-1993, integrated into UniProtKB/Swiss-Prot.
DT   01-APR-1993, sequence version 1.
DT   24-NOV-2009, entry version 56.
DE   RecName: Full=Hemoglobin subunit beta-Z;
DE   AltName: Full=Hemoglobin beta-Z chain;
DE   AltName: Full=Beta-Z-globin;
DE   Flags: Fragment;
GN   Name=HBBZ;
OS   Mesocricetus auratus (Golden hamster).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi;
OC   Muroidea; Cricetidae; Cricetinae; Mesocricetus.
OX   NCBI_TaxID=10036;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RX   MEDLINE=92223098; PubMed=1339291; DOI=10.1016/0167-4781(92)90532-5;
RA   Li H., Subar M., Lee K.M., Boussios T.;
RT   "Cloning and sequence analysis of two embryonic beta-like globin cDNAs
RT   (y and z) of hamster.";
RL   Biochim. Biophys. Acta 1130:218-220(1992).
CC   -!- FUNCTION: This is an embryonic beta chain.
CC   -!- SUBUNIT: Heterotetramer of two alpha chains and two beta chains.
CC   -!- SIMILARITY: Belongs to the globin family.
CC   -----------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/term
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   -----------------------------------------------------------------------
DR   EMBL; X64178; CAA45517.1; -; mRNA.
DR   PIR; S22356; S22356.
DR   SMR; P29627; 1-101.
DR   HOVERGEN; P29627; -.
DR   GO; GO:0005833; C:hemoglobin complex; IEA:InterPro.
DR   GO; GO:0020037; F:heme binding; IEA:InterPro.
DR   GO; GO:0019825; F:oxygen binding; IEA:InterPro.
DR   GO; GO:0005344; F:oxygen transporter activity; IEA:UniProtKB-KW.
DR   GO; GO:0015671; P:oxygen transport; IEA:UniProtKB-KW.
DR   InterPro; IPR012292; Globin.
DR   InterPro; IPR009050; Globin-like.
DR   InterPro; IPR000971; Globin_subset.
DR   InterPro; IPR002337; Haemoglobin_b.
DR   Gene3D; G3DSA:1.10.490.10; Globin_related; 1.
DR   PANTHER; PTHR11442:SF7; Beta_haem; 1.
DR   Pfam; PF00042; Globin; 1.
DR   PRINTS; PR00814; BETAHAEM.
DR   PROSITE; PS01033; GLOBIN; 1.
PE   3: Inferred from homology;
KW   Heme; Iron; Metal-binding; Oxygen transport; Transport.
FT   CHAIN        <1    102       Hemoglobin subunit beta-Z.
FT                                /FTId=PRO_0000053018.
FT   METAL        19     19       Iron (heme distal ligand) (By
FT                                similarity).
FT   METAL        48     48       Iron (heme proximal ligand) (By
FT                                similarity).
FT   NON_TER       1      1
SQ   SEQUENCE   102 AA;  11227 MW;  88AF4095D8F4E711 CRC64;
     FGNLSSAQAI MGNPRIRAHG KKVLTSLGLA VQNMDNLKET FAHLSELHCD KLHVDPENFK
     LLGNVLVIVL STHFAKEFTP EVQAAWQKLV AGVANALSHK YH
//
```

Protein family information

Cross-references to other databases
(DNA, domain, etc.)

UniProt graphic view



Accession # Entry name

P69905 · HBA_HUMAN

Protein: Hemoglobin subunit alpha
Gene: HBA1; HBA2
Status: UniProtKB reviewed (Swiss-Prot)
Amino acids: 142 (go to sequence)
Protein existence: Evidence at protein level
Annotation score: 5/5

Swiss-Prot entries are manually annotated (high quality)

TrEMBL entries are only computationally annotated.
After careful review, TrEMBL entries will be moved to Swiss-Prot.

Q5EBL1 · Q5EBL1_MOUSE

Protein: Globin b3
Gene: Hbb-bh1
Status: UniProtKB unreviewed (TrEMBL)
Organism: Mus musculus (Mouse)
Amino acids: 147
Protein existence: Evidence at transcript level
Annotation score: 2/5

https://www.uniprot.org/uniprotkb/P69905/entry

READSEQ: Sequence Format Converter

- Readseq @ phylogeny.fr

  http://www.phylogeny.fr/one_task.cgi?task_type=readseq
- Sequence Format Conversion @ EBI

  https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/
- Readseq @ mafft website

  https://mafft.cbrc.jp/alignment/server/cgi-bin/readseq.txt
- Format Converter @ lanl.gov

  https://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html

[More on the course website, Links page]

Sequence Analysis Tools on the Web

- NCBI Sequence Analysis Tools (BLAST, *etc.*)

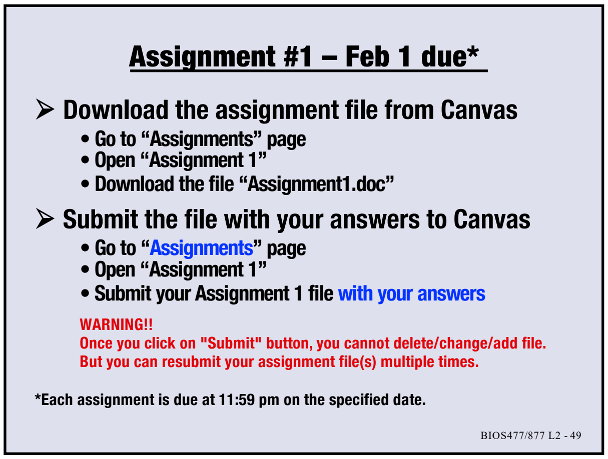
  https://www.ncbi.nlm.nih.gov/guide/sequence-analysis/
- EBI Bioinformatics services

  https://www.ebi.ac.uk/services
- ExPASy: SIB Bioinformatics Resource Portal

  https://www.expasy.org
- Max-Planck Institute Bioinformatics Toolkit

  http://toolkit.tuebingen.mpg.de/

- Many more links are available on the Course website, Links page



Assignment #1 – Feb 1 due*

- Download the assignment file from Canvas
  - Go to “Assignments” page
  - Open “Assignment 1”
  - Download the file “Assignment1.doc”
- Submit the file with your answers to Canvas
  - Go to “Assignments” page
  - Open “Assignment 1”
  - Submit your Assignment 1 file with your answers

WARNING!!
Once you click on "Submit" button, you cannot delete/change/add file.
But you can resubmit your assignment file(s) multiple times.

*Each assignment is due at 11:59 pm on the specified date.

BIOS477/877 L2 - 49

49