Slide 1

BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

| Spring 2026 | Lecture 2 |

1

Slide 2

Today's topics

➢ **Introduction to Internet Resources (NCBI, databases)**

➢ **Assignment #1**
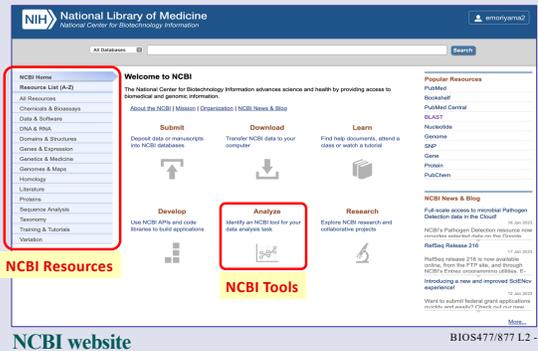
2

Slide 3

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

Slide 4

National Center for Biotechnology Information (NCBI Home)

NCBI Resources

NCBI Tools

NCBI website

4

Slide 5

National Center for Biotechnology Information (How to start database search)

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

Or choose one from this list
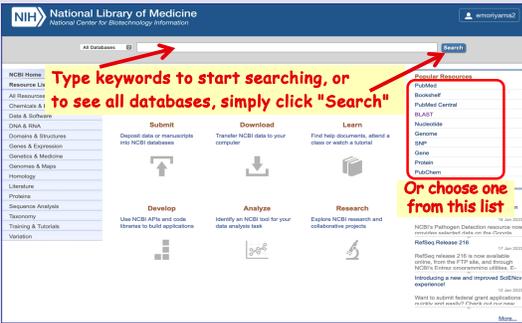
Next slide shows all databases available at NCBI

5

Slide 6

NCBI Databases: Literature

Literature databases are found at the top of the page

PubMed (biomedical literature)

Many online books

Scroll down to find more databases

6

1

NCBI Databases: Genes and Genomes

GenBank, NCBI's nucleotide sequence database, is found in "Genomes"

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

GEO (Gene Expression Omnibus): microarray & RNA-Seq data
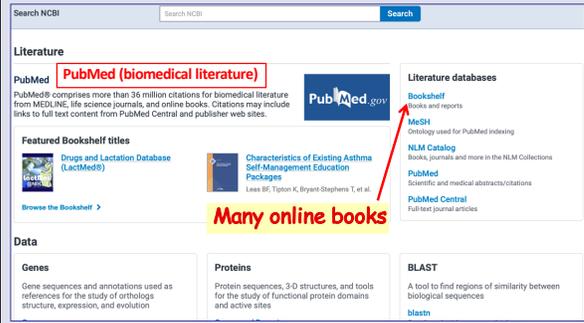
HomoloGene
Homologous gene sets for selected organisms

PopSet
Sequence sets from phylogenetic and population studies

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

Nucleotide
DNA and RNA sequences — **GenBank**

SRA
High-throughput sequence reads

Taxonomy
Taxonomic classification and nomenclature

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

blastx
Search protein databases using a translated nucleotide sequence database query

Protein Family Models
Models representing homologous proteins with a common function

Structure
Experimentally-determined biomolecular structures

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 and links

OMIM
Online mendelian inheritance in man

BLAST
A tool to find regions of similarity between biological sequences

blastn
Search nucleotide sequence databases

tblastn
Search translated nucleotide databases using a translated protein query

Primer-BLAST
Find primers specific to your PCR template

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

SRA (Sequence Read Archive): database for next-gen sequencing data

7

NCBI Databases: Proteins

For proteins, sequences, structures, and domain databases are available

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 gene sets for selected organisms

PopSet
Sequence sets from phylogenetic and population studies

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

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

Protein
Protein sequences

Protein sequences & structures

Protein Family Models
Models representing homologous proteins with a common function

Structure
Experimentally-determined biomolecular structures

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

ClinicalTrials.gov
Privately and publicly funded clinical studies

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

BLAST
A tool to find regions of similarity between biological sequences

blastn
Search nucleotide databases

Primer-BLAST
Find primers specific to your PCR template

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

Biological systems & pathways

8

Literature search using PubMed

Click on PubMed link to start the PubMed search

Search NCBI [Search NCBI] **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.

PubMed.gov

Featured Bookshelf titles

Drugs and Lactation Database (LactMed®)

Characteristics of Existing Asthma Self-Management Education Packages
Lees 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

9

PubMed search: how to start

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed.gov

[bioinformatics] **① Type a query**   **② Start searching →** **Search**

Advanced

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.

Many online tutorials are available @ NCBI
PubMed Online Training

10

PubMed search: result page

PubMed®

[bioinformatics] **Search**

Advanced Create alert Create RSS   User Guide

Save Email Send to   Sort by: Best match   Display options ⚙

Advanced search

MY NCBI FILTERS

RESULTS BY YEAR
< Page 1 of 50,880 >

1958 — 2024

Sort order. To see more recent papers, this needs to be changed to "most recent".

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

☐ 1 **Bioinformatics in transla...**
Cite Wooler SK, Benstead-Hume G,
Share Biosci Rep. 2017 Jul 7;37(4):BSR...
PMID: 28487472 Free PMC...
Bioinformatics approaches are... in academia and within the pharm...
bioinformatics resources and r...

☐ 2 **Is "bioinformatics" dead...**
Cite Bourne PE.
Share 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.**
Cite Kinghorn AB, Fraser LA, Lang S, Shiu SCC, Tanner JA.
Share 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 ...

☐ 4 **Bioinformatics in bioscience and bioengineering: Recent advances, applications, and perspectives.**
Cite Uesaka K, Oka H, Kato R, Kanie 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.

Check "Review" if you want to find only review papers

11

PubMed search: search history

PubMed Advanced Search Builder

PubMed®
User Guide

Add terms to the query box

[All Fields] [Enter a search term] **ADD ▼**
Show Index

Query box

[Enter / edit your search query here]

Search ▼

Click the number to see the result of previous search

History and Search Details   ⬇ Download 🗑 Delete

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

Showing 1 to 1 of 1 entries

12

2

13

14

15

16

17

18

PubMed search: term translation

Click on ">" to see the query detail

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

Actual queries used for the search

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

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 related 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"**
"bioinformatics"[All Fields]   ← No translation is done   216,575   05:39:19

BIOS477/877 L2 - 19

19

PubMed search: complex query

[Task]
Find all articles about "gene prediction" published in journals that have "bioinformatics" in their titles

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

Find 1∩2

This search requires a complex query

① ②

BIOS477/877 L2 - 20

20

How to create a complex query

PubMed Advanced Search Builder   PubMed.gov
User Guide

Add terms to the query box

All Fields ⬍   "gene predication" ————→   ADD
Show Index

Query box
"gene prediction"

First query is added   Search ⌄

BIOS477/877 L2 - 21

21

How to create a complex query (continued)

PubMed Advanced Search Builder   PubMed.gov
User Guide

Add terms to the query box   Second query

Journal ⬍   bioinformatics   ✕   AND ⌄
                                          Add with AND
The second search is limited to "Journal" field   Add with OR
                                          Add with NOT
"gene predication"   ✕

The second query can be added by using AND, OR, or NOT operator (AND is chosen)

Query box
("gene predication") AND ("bioinformatics"[Journal])   ✕   Search ⌄

To search articles about "gene prediction"
published in journals that have "bioinformatics" in their
titles (e.g., "Bioinformatics", "BMC Bioinformatics")

BIOS477/877 L2 - 22

22

How to combine multiple searches

PubMed Advanced Search Builder   PubMed
User Guide

Add terms to the query box

Journal ⬍   Enter a search term   ADD
Show Index

Query box

Enter / edit your   Add to History ⌄

A complex search can be achieved by combining multiple searches using search history

History and Search Details   ⬇ Download   🗑 Delete

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

Click here

BIOS477/877 L2 - 23

23

How to combine multiple searches
(adding the first query)

PubMed Advanced Search Builder   PubMed
User Guide

Add terms to the query box

Journal ⬍   Enter a search term   ✕   ADD
Show Index

Query box
"gene prediction"

Add to History ⌄

Add the first search query

History and Search Details   ⬇ Download   🗑 Delete

Search	Actions	Details	Query	Results	Time
#2	...		Add query	nformatics"[Journal]	17,577	08:13:29
#1	...		Delete	e prediction"	1,862	08:13:13
			Create alert			

BIOS477/877 L2 - 24

24

4

How to combine multiple searches
(adding the second query)

25

How to combine multiple searches
(search result)

26

NCBI global database search
(how to start)

27

NCBI global database search
(search results across all databases)

28

Nucleotide database search
(search result page – #1)

29

Nucleotide database search
(search details)

30

Nucleotide database search (problem?) — slide 31

Nucleotide database search (Advanced search) — slide 32

31

32

Nucleotide database search (Advanced search for a specific field) — slide 33

Nucleotide database search (search result page – #2) — slide 34

33

34

How to download the search result — slide 35

GenBank entry: Web view — slide 36

35

36

GenBank entry: tool list

37

GenBank entry: downloading CDS

38

GenBank entry: sequence features

39

GenBank entry: extracting sequences

40

GenBank entry: related information

41

GenPept (translated GenBank) entry

42

FASTA sequence format

➢ **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 description in a single line**, followed by lines of sequence data

▪ In the description line, 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

43

Variation of FASTA format

[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-------
```

44

DNA sequence databases

➢ **INSDC: International Nucleotide Sequence Database Collaboration**

▪ **GenBank**: NIH genetic sequence database

▪ **ENA**: European Nucleotide Archive

▪ **DDBJ**: DNA Data Bank of Japan

➢ **Data are shared among the three databases**

➢ **They share the accession numbers!**

➢ **They use different formats**

45

DNA sequence databases: format

EMBL format   **GenBank (DDBJ) format**

For the same sequence entry:
- **ID** vs. **LOCUS**:
  → not required to be the same between databases.
- Accession numbers (**AC**, **ACCESSION**):
  → are the same among databases.

46

DNA sequence databases: identifier

EMBL format   **GenBank (DDBJ) format**

- Even within the same database, **ID** or **LOCUS** may change between versions.
- Accession numbers (**AC** or **ACCESSION**) are permanent.

47

Protein sequence databases

➢ **Protein Databases**

▪ **UniProt: The Universal Protein Resource**
  → Collaboration between:
    ○ European Bioinformatics Institute (**EMBL-EBI**)
    ○ Swiss Institute of Bioinformatics (**SIB**), and
    ○ Protein information Resource (**PIR**)

▪ **UniProtKB: UniProt Knowledgebase**
  → Comprehensive resource for protein sequence and annotation data
  → One of UniProt databases
  → Contains two sections:
    ○ **Swiss-Prot** (manually annotated and reviewed)
    ○ **TrEMBL** (automatically annotated and NOT reviewed)

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

48

Protein sequence databases: format

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

Protein family information

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

BIOS477/877 L2 - 49

49

UniProt Web interface

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.

Example entry: P69905

BIOS477/877 L2 - 50

50

Sequence format converters

- **Readseq @ phylogeny.fr**
- **SEQRET: Sequence Format Conversion @ EBI**
- **Readseq @ mafft website**
- **Format Converter @ lanl.gov**

[More on the course website, Links page]

BIOS477/877 L2 - 51

51

Sequence analysis tools on the Web

- **NCBI Sequence Analysis Tools (BLAST, etc.)**
- **EBI Bioinformatics services (InterPro, etc.)**
- **ExPASy: SIB Bioinformatics Resource Portal**
- **Max-Planck Institute Bioinformatics Toolkit**

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

BIOS477/877 L2 - 52

52

Assignment #1 – Jan 22 due*

➢ **Download the assignment file from Canvas**
- Go to "Assignments" page
- Open "Assignment 1"
- Download the file "Assignment1.docx"

➢ **Write your answers in the Assignment 1 file**

➢ **Submit your assignment file to Canvas**
- Go to "Assignments" page
- Open "Assignment 1" and submit your file

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 - 53

53