



# INTRODUCTION TO BIOINFORMATICS

مقدمة في علم المعلومات الحياتية

**W1. Overview On Sequence Databases.**

كلية التقنية الطبية / قسم الأدلة الجنائية

**Dr. Ragheed Hussam Yousif**

**Ragheed.Hussam@alfarahidiuc.edu.iq**



# **OUTLINE**

- 1. The National Center for Biotechnology Information (NCBI).**
- 2. The Main Tools Under NCBI**

# Introduction

- Sequence databases are great tools because they offer a unique window on the past. They make it possible to answer today's biological questions by enabling us to analyze sequences that may have been determined as many as 25 years ago.
- when the whole technology emerged. By doing this, they connect past and present molecular biology.

# Introduction

- The first databases were in fact created as some sort of sequence museum, where sequences could be preserved for all eternity in perfect form, just as they were determined, interpreted, and published by their original authors.
- This historical (time capsule!) perspective pretty much remains in GenBank, the leading nucleotide sequence repository maintained as a consortium between the U.S.
- National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ).

## **The National Center for Biotechnology Information (NCBI)**

- (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). It is approved and funded by the government of the United States. The NCBI in 1988.
- The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for biomedical literature.

# The National Center for Biotechnology Information (NCBI)



- Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine. NCBI was directed by David Lipman, one of the original authors of the BLAST sequence alignment program and a widely respected figure in bioinformatics.

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

**COVID-19 Information**

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

**UNITE**  
A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.

**Ending Structural Racism**  
nih.gov/ending-structural-racism

**LEARN MORE**

**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**

Three outdated browsers (1000 Genomes, dbGaP Data, and Get-EM) to retire in April 2022. Data available in GDV 28 Oct 2021

The Genome Data Viewer (GDV) is now 27 Oct 2021

NCBI will assign 64-bit numeric GIs by November 15th. Update affected software! 25 Oct 2021

As announced last month, NCRI will 25 Oct 2021

Nov 3 Webinar: dbGaP submission improvements and GaPTools 25 Oct 2021

Attention dbGaP submitters! Join us on

## **The Main Tools Under NCBI**

1. GenBank
2. PubMed database
3. NCBI Bookshelf
4. Basic Local Alignment Search Tool (BLAST)
5. Entrez
6. Gene
7. Protein
8. PubChem database



## GenBank

- The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. It is produced and maintained by the National Center for Biotechnology Information (NCBI; a part of the National Institutes of Health in the United States) as part of the International Nucleotide Sequence Database Collaboration (INSDC).

## GenBank

- GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. The database started in 1982 by Walter Goad and Los Alamos National Laboratory. GenBank has become an important database for research in biological fields and has grown in recent years at an exponential rate by doubling roughly every 18 months.
- Release 242.0, produced in February 2021, contained over 12 trillion nucleotide bases in more than 2 billion sequences. GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

## **PubMed database**

- PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintain the database as part of the Entrez system of information retrieval.
- From 1971 to 1997, online access to the MEDLINE database had been primarily through institutional facilities, such as university libraries.[2] PubMed, first released in January 1996, ushered in the era of private, free, home- and office-based MEDLINE searching. The PubMed system was offered free to the public starting in June 1997.

PubMed.gov

Advanced

PubMed® comprises more than 33 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.



**Learn**

- About PubMed
- FAQs & User Guide
- Finding Full Text



**Find**

- Advanced Search
- Clinical Queries
- Single Citation Matcher



**Download**

- E-utilities API
- FTP
- Batch Citation Matcher



**Explore**

- MeSH Database
- Journals

## **NCBI Bookshelf**

- The NCBI Bookshelf is a collection of freely accessible, downloadable, on-line versions of selected biomedical books. The Bookshelf covers a wide range of topics including molecular biology, biochemistry, cell biology, genetics, microbiology, disease states from a molecular and cellular point of view, research methods, and virology. Some of the books are online versions of previously published books, while others, such as Coffee Break, are written and edited by NCBI staff.

NCBI Resources How To Sign in to NCBI

Bookshelf Books Search Browse Titles Advanced Help

**COVID-19 Information** ✕  
[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)


**Bookshelf**  
 Bookshelf provides free online access to books and documents in life science and healthcare. Search, read, and discover.


**Using Bookshelf**  
[Quick Start Guide](#)  
[FAQ](#)  
[Tutorials](#)  
[Copyright and Permissions](#)  
[Follow @ncbibooks](#)


**Read**  
[Browse Titles](#)  
[New Releases](#)  
[PubReader](#)

**Participate**  
[Authors and Publishers](#)  
[How to Apply](#)  
[Participation Agreement](#)  
[File Submission Specifications](#)


**New & Updated**


 **Malnutrition in Hospitalized Adults: A Systematic Review [Internet].** Comparative Effectiveness Review, No. 249. Uhl S, Siddique SM, McKeever L, et al. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021 Oct.


 **Developing and Piloting a Tool To Create Dot Plots To Summarize Pooled Data for Multiple Outcomes in Systematic Reviews [Internet].** Yu Y, Fu R, Wagner J, et al. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021 Oct.

 **Outcome Measure Harmonization and Data Infrastructure for Patient-Centered Outcomes Research in Depression: Data Use and Governance Toolkit [Internet].** Friedler HS, Leavy MB, Bickelman E, et al. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021 Oct.

**Featured Titles**

 **Combating Antimicrobial Resistance and Protecting the Miracle of Modern Medicine [Prepublication Draft].** Buckley GJ, Palmer GH; Committee on the Long-Term Health and Economic Effects of Antimicrobial Resistance in the United States; Board on Population Health and Public Health Practice; Health and Medicine Division; National Academies of Sciences, Engineering, and Medicine. Washington (DC): National Academies Press (US); 2021.

 **Drugs and Lactation Database (LactMed) [Internet].** Bethesda (MD): National Library of Medicine (US); 2006-.

 **Medical Genetics Summaries [Internet].** Pratt VM, Scott SA, Pirmohamed M, et al., editors. Bethesda (MD): National Center for Biotechnology Information (US); 2012-.

**More Information**  
[NLM Literature Archive](#)  
[Open Access Subset](#)  
[Librarians](#)

## **Basic Local Alignment Search Tool (BLAST)**

- BLAST is an algorithm used for calculating sequence similarity between biological sequences such as nucleotide sequences of DNA and amino acid sequences of proteins.
- BLAST is a powerful tool for finding sequences similar to the query sequence within the same organism or in different organisms. It searches the query sequence on NCBI databases and servers and posts the results back to the person's browser in the chosen format.
- Input sequences to the BLAST are mostly in FASTA or Genbank format while output could be delivered in a variety of formats such as HTML, XML formatting, and plain text.

## **Basic Local Alignment Search Tool (BLAST)**

- HTML is the default output format for NCBI's web-page. Results for NCBI-BLAST are presented in graphical format with all the hits found, a table with sequence identifiers for the hits having scoring related data, along with the alignments for the sequence of interest and the hits received with analogous BLAST scores for these.





## Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**NEWS**

**BLAST+ 2.12.0 is here!**  
We have made some improvements to how BLAST multi-threads and the amount of memory required by makeblastdb.

Tue, 13 Jul 2021 12:00:00 EST [More BLAST news...](#)

## Web BLAST



**Nucleotide BLAST**  
nucleotide ▶ nucleotide

**blastx**  
translated nucleotide ▶ protein

**tblastn**  
protein ▶ translated nucleotide



**Protein BLAST**  
protein ▶ protein

## BLAST Genomes

Enter organism common name, scientific name, or tax id

Human   Mouse   Rat   Microbes

## Standalone and API BLAST

 **Download BLAST**  
Get BLAST software and databases

 **Use BLAST API**  
Call BLAST from your application

 **Use BLAST in the cloud**  
Get BLAST on the cloud service

## Entrez

- The Entrez Global Query Cross-Database Search System is used at NCBI for all the major databases such as Nucleotide and Protein Sequences, Protein Structures, PubMed, Taxonomy, Complete Genomes, OMIM, and several others. Entrez is both an indexing and retrieval system having data from various sources for biomedical research.
- NCBI distributed the first version of Entrez in 1991, composed of nucleotide sequences from PDB and GenBank, protein sequences from SWISS-PROT, translated GenBank, PIR, PRF, PDB, and associated abstracts and citations from PubMed.

## Entrez

- Entrez is specially designed to integrate the data from several different sources, databases, and formats into a uniform information model and retrieval system which can efficiently recover that relevant references, sequences and structures.

## Gene

- Gene has been implemented at NCBI to characterize and organize the information about genes. It serves as a major node in the nexus of the genomic map, expression, sequence, protein function, structure, and homology data. A unique GeneID is assigned to each gene record that can be followed through revision cycles.
- Gene records for known or predicted genes are established here and are demarcated by map positions or nucleotide sequences. Gene has several advantages over its predecessor, LocusLink, including, better integration with other databases in NCBI, broader taxonomic scope, and enhanced options for query and retrieval provided by the Entrez system.

## Protein

- Protein database maintains the text record for individual protein sequences, derived from many different resources such as NCBI Reference Sequence (RefSeq) project, GenBank, PDB, and UniProtKB/SWISS-Prot.
- Protein records are present in different formats including FASTA and XML and are linked to other NCBI resources.
- Protein provides the relevant data to the users such as genes, DNA/RNA sequences, biological pathways, expression and variation data, and literature.

## Protein

- It also provides the pre-determined sets of similar and identical proteins for each sequence as computed by the BLAST.
- The Structure database of NCBI contains 3D coordinate sets for experimentally-determined structures in PDB that are imported by NCBI.
- The Conserved Domain database (CDD) of protein contains sequence profiles that characterize highly conserved domains within protein sequences. It also has records from external resources like SMART and Pfam.

## Protein

- There is another database in a protein known as Protein Clusters database which contains sets of proteins sequences that are clustered according to the maximum alignments between the individual sequences as calculated by BLAST.

## Pubchem database

- PubChem is a database of chemical molecules and their activities against biological assays. The system is maintained by the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine, which is part of the United States National Institutes of Health (NIH).
- PubChem can be accessed for free through a web user interface. Millions of compound structures and descriptive datasets can be freely downloaded via FTP. PubChem contains multiple substance descriptions and small molecules with fewer than 100 atoms and 1000 bonds. More than 80 database vendors contribute to the growing PubChem database.





### COVID-19 Information

[Public health information \(CDC\)](#) [Research information \(NIH\)](#) [SARS-CoV-2 data \(NCBI\)](#) [Prevention and treatment information \(HHS\)](#) [Español](#)



National Library of Medicine  
National Center for Biotechnology Information



[About](#) [Blog](#) [Submit](#) [Contact](#)

# Explore Chemistry

Quickly find chemical information from authoritative sources



Try [covid-19](#) [aspirin](#) [EGFR](#) [C9H8O4](#) [57-27-2](#) [C1=CC=C\(C=C1\)C=O](#) [InChI=1S/C3H6O/c1-3\(2\)4/h1-2H3](#)

Use Entrez  Compounds  Substances  BioAssays



Draw Structure



Upload ID List



Browse Data



Periodic Table

111M Compounds 275M Substances 293M Bioactivities 33M Literature 29M Patents

*Thank  
You!*

