

Лекція 6

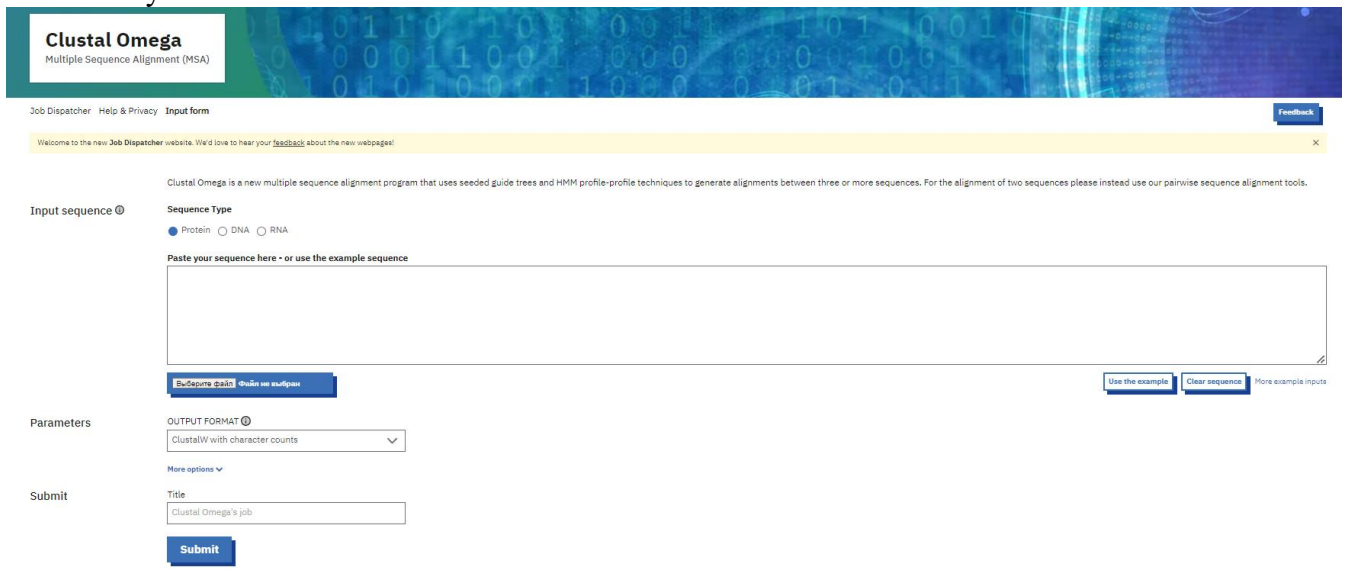
Тема: Інструменти біоінформатики: інструменти роботи з біологічними даними, світові бази даних

План:

1. Виконання множинного вирівнювання у програмі ClustalW.
2. Робота з публікаціями
3. Робота з базами даних

1. Виконання множинного вирівнювання у програмі ClustalW

Множинне вирівнювання послідовностей сьогодні стало можливим за допомогою програм, що доступні у on-line режимі. Одна з них – ClustalW. Серія даних програм з'явилася у 1994 році. Програма дозволяє обирати матриці порівняння амінокислот і нуклеотидів. Результати вирівнювання представляються у вигляді формату FASTA, що забезпечує високу сумісність програми з іншими пакетами програм. Програма Clustal доступна на багатьох серверах, зокрема, <http://npsa-pbil.ibcp.fr>, <http://www.ebi.ac.uk/services>, у двох варіантах: інтерактивному та поштовому.



The screenshot shows the Clustal Omega web interface. At the top, it says "Clustal Omega Multiple Sequence Alignment (MSA)". Below this, there are navigation links: "Job Dispatcher", "Help & Privacy", "Input form", and a "Feedback" button. A yellow banner contains a welcome message: "Welcome to the new Job Dispatcher website. We'd love to hear your feedback about the new webpage!". The main content area is titled "Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools." Under "Input sequence", there is a "Sequence Type" section with radio buttons for "Protein" (selected), "DNA", and "RNA". Below this is a text input field for "Paste your sequence here - or use the example sequence". There are buttons for "Выберите файл" (Choose file), "Файл не выбран" (File not selected), "Use the example", "Clear sequence", and "More example inputs". The "Parameters" section has an "OUTPUT FORMAT" dropdown menu set to "ClustalW with character counts" and a "More options" link. The "Submit" section has a "Title" input field with "Clustal Omega's job" and a "Submit" button.

Рисунок 1. Вікно програми Clustal

Перший дозволяє виконувати вирівнювання невеликої кількості послідовностей (<100), а другий – по електронній пошті (застосовується за умови великої кількості послідовностей).

До основних можливостей програми можна віднести наступні:

- виконання множинного вирівнювання;
- підрахунок еволюційних дистанцій між послідовностями;
- визначення характеру і типу амінокислотних замін.

До використання методів секвенування дослідження у молекулярній біології проводилися з допомогою амінокислотних послідовностей. Визначення амінокислотних послідовностей займало багато часу і часто було помилковим. Наразі визначити послідовність ДНК легше, ніж послідовність білка, тому амінокислотні послідовності отримують з нуклеотидних послідовностей, використовуючи таблиці генетичного коду. Джерелом генетичних баз даних є наукові статті.

2. Робота з публікаціями

Наукові статті мають індекс цитування - скільки разів статтю процитували в інших статтях. Вважають, що чим більше разів цитують публікацію, тим достовірніша інформація в ній,

вищий рейтинг журналу та автора статті.

Індекси цитування неможливо підрахувати вручну, тому наукові праці відображаються у базах даних.

Наукова діяльність передбачає, у першу чергу, роботу з публікаціями. Робота з великими масивами наукової інформації: статтями, монографіями, патентами та ін. ускладнена сьогодні, насамперед, великою її кількістю, а також великою кількістю неперевіреної та нерецензованої інформації у мережі інтернет. За останніми даними сучасний вчений читає за рік не більше 200 статей, що складає приблизно 0,4 % від усього масиву наявних наукових журналів. На допомогу науковцю приходять різні інструменти роботи з науковими публікаціями.

Робота з Google Scholar

Робота з Google Scholar – це зручний ресурс для пошуку наукових статей, що знаходиться за посиланням <https://scholar.google.com.ua/scholar>. Браузер Google «знайшов» та проіндексував всю інформацію, яку можна вважати науковими публікаціями і розмістив її у межах одного ресурсу, що дозволяє проводити пошук по статтям, а не по окремих сайтах.

The screenshot shows the Google Scholar search interface. The search bar contains the text 'virivnyuvannya genetychnykh poslidovnostey'. Below the search bar, it indicates 'Articles' and 'About 2,760 results (0.08 sec)'. On the left side, there are filters for 'Any time' (with options: Since 2024, Since 2023, Since 2020, Custom range...), 'Sort by relevance' (with 'Sort by date' selected), 'Any type' (with 'Review articles' selected), and checkboxes for 'include patents', 'include citations', and 'Create alert'. The main content area displays search results. The first result is titled 'Використання алгоритму барроуза-уїлера для вирівнювання генетичних послідовностей' by А Ляшенко, О Богатирьов - 2012 - elartu.tntu.edu.ua. The second result is 'Біоінформаційний пошук послідовностей генів, що кодують тубуліни у геномі льону' by ГЯ Баєр, МО Пидюра, ЯВ Пірко, АІ Ємець... - 'noi evolucii organizmiv, 2014 - utgis.org.ua'. The third result is 'Біоінформатика: аналіз генетичних послідовностей' by Б Осташ - 2022 - dspace.ln.ulb.lviv.ua. The fourth result is '[PDF] Система для вирівнювання послідовностей ДНК на основі прихованих марковських моделей' by АІ Старовойт - 2021 - ela.kpi.ua. Each result includes a brief abstract and options to 'Save', 'Cite', and view 'Related articles'.

Рисунок 2. Інтерфейс Google Scholar

Ресурс дозволяє отримати інформацію про:

- цитування статті;
- схожі статті;
- версії однієї статті в різних журналах.

Також можна сортувати отриманий список за часом та налаштувати оповіщення на електронну пошту за ключовими словами та авторами.

Робота з платформою Web of Science

Web of Science – це міжнародна мультидисциплінарна реферативна платформа, що включає в себе велику кількість предметних баз (рис. 3).

Платформа індексує більше 17500 найвпливовіших світових журналів та містить інформацію про матеріали, на які посилається автор конкретної статті, а також на матеріали, які цитують дану

статтю.

Платформа Web of Science дозволяє:

- обирати мову інтерфейсу;
- обирати критерії пошуку за автором, ключовими словами та ін..(рис. 4);
- знайти контактні дані авторів статті;
- створити власну бібліотеку публікацій;
- мати можливість аналізувати отримані результати.

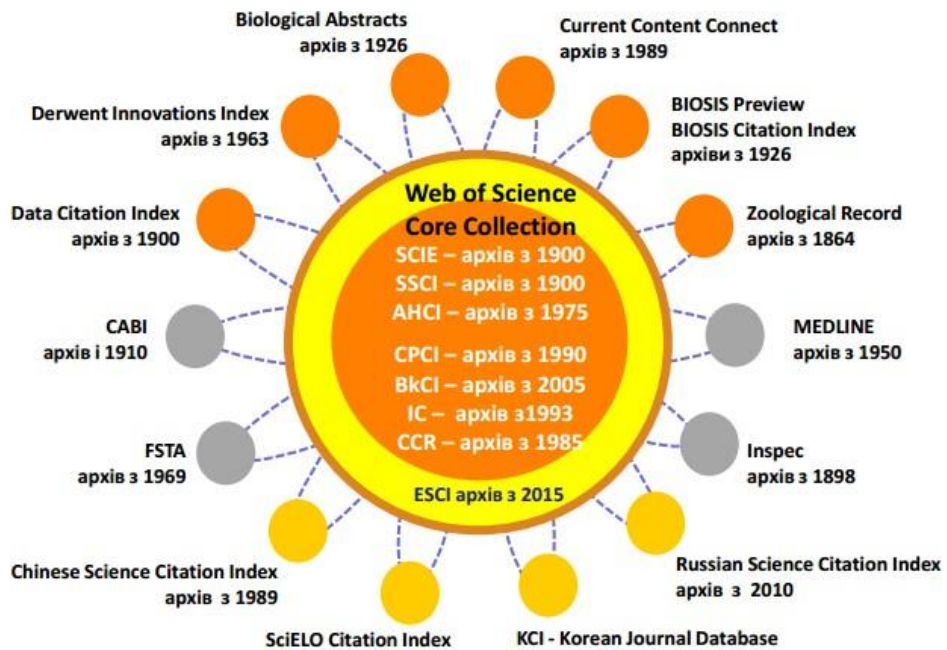


Рисунок 3. Базы данных на платформе Web of Science

Платформа Web of Science має інструментарій що дозволяє:

- мати картотеку статей за обраною темою;
- оформити статтю за вимогами обраного видання;
- автоматизувати процес оформлення списку літератури;
- дізнатися імпаکت-фактор обраного видання;
- знайти колег у світі, які займаються даною проблемою;
- зберігати знайдені публікації;
- переоформити матеріал для іншого видання.

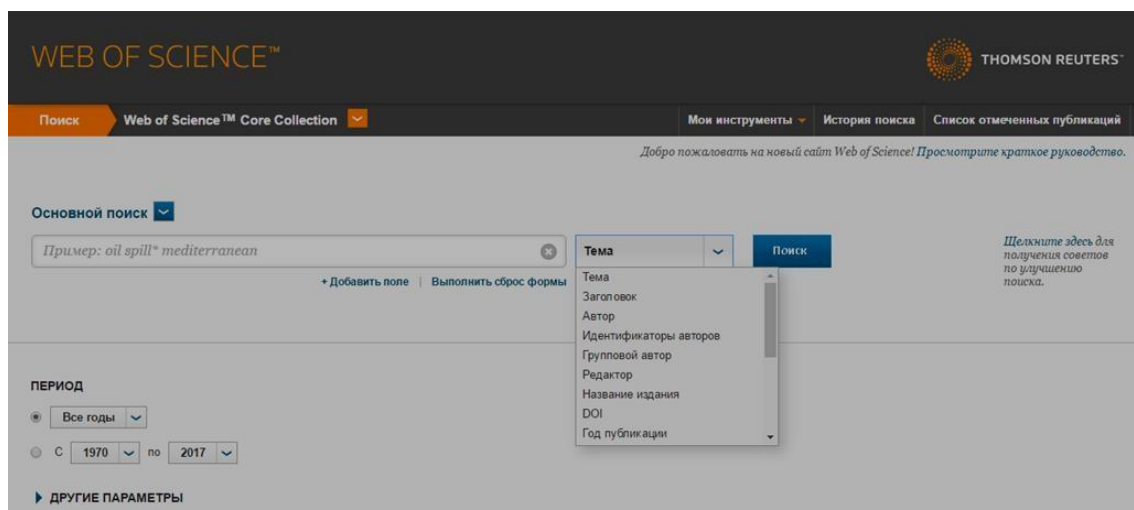


Рисунок 4. Вікно пошуку платформи Web of Science

Робота з програмою Mendeley

Читати та аналізувати великі масиви статей досить складно та незручно. Для ефективної роботи з великою кількістю інформації використовують програму Mendeley (рис. 5). Програма у файлах з розширенням pdf сама знаходить назву журналу, автора статті, а також дозволяє сортувати статті за прізвищем автора, «зшивати» воедино одну і ту ж статтю, що міститься у різних файлах. Корисною функцією програми є автоматичне створення посилань у тексті та списку використаних джерел.

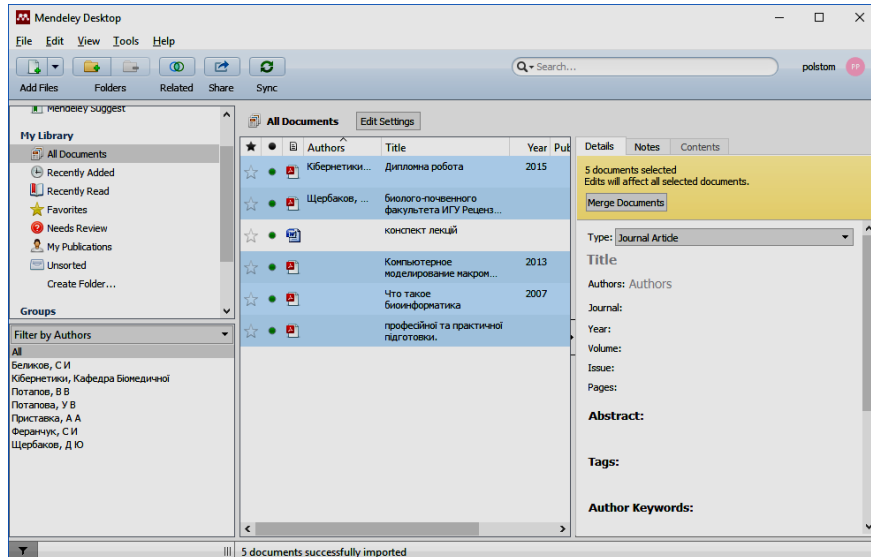


Рисунок 5. Вікно програми Mendeley

Схожою за функціями є програма Red Cube Web Reader.

Корисним може бути плагін reflect.ws. Він встановлюється у браузер і дозволяє прямо у самій статті знаходити визначення певних термінів (рис. 6)

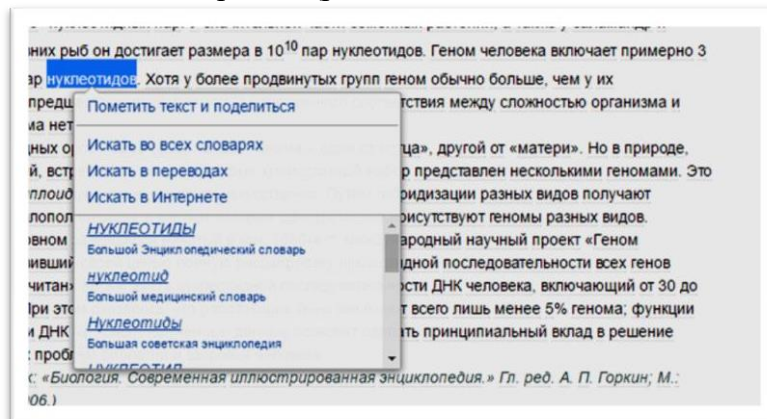


Рисунок 6. Використання плагіну reflect.ws

3. Робота з базами даних

Бази даних, що містять інформацію про біологічні послідовності, мають логічно організовану структуру даних. Кожному запису бази відповідає ідентифікатор. Ці ідентифікатори, зазвичай, різняться для різних баз даних. Інформація у базах даних часто буває неповною і, навіть, містить помилки, але наукова діяльність без їхнього використання неможлива.

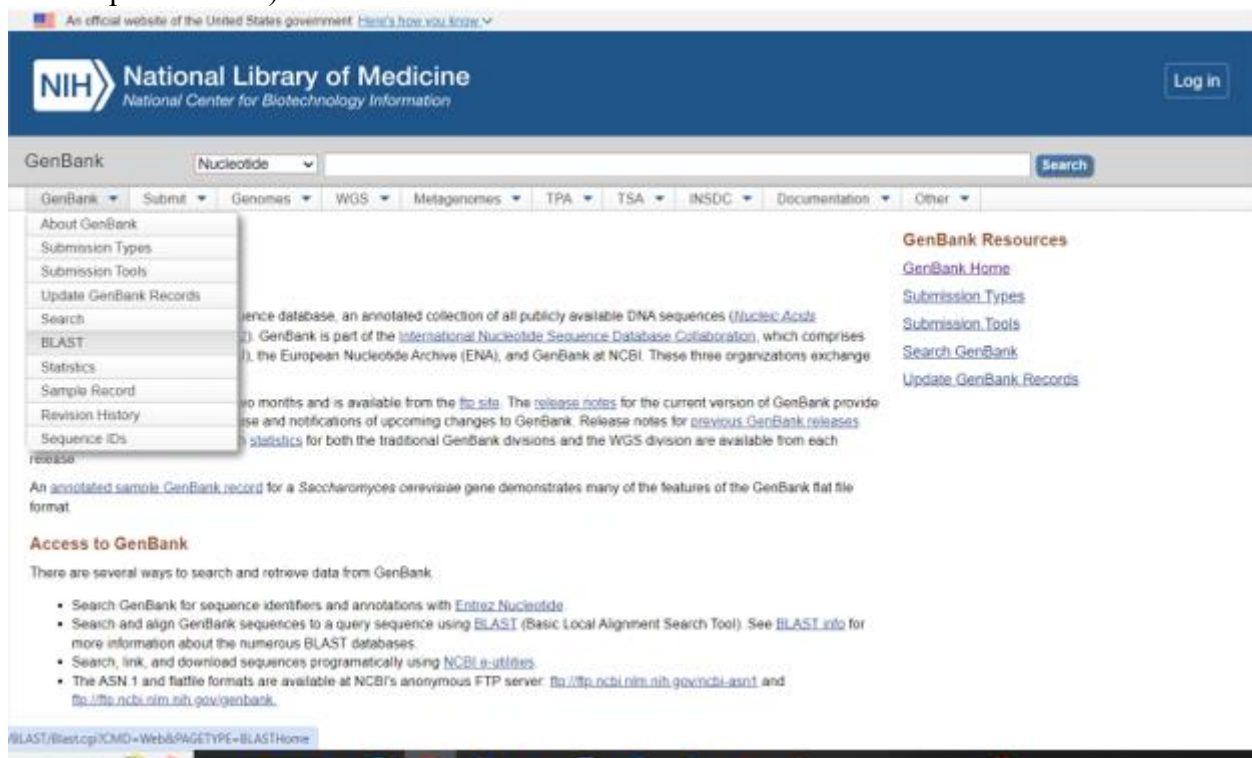
База даних генетичних послідовностей GenBank

GenBank – світовий архів послідовностей нуклеїнових кислот, що виник завдяки партнерству Національного центру біотехнологічної інформації США, Бібліотеки даних Європейського інституту біоінформатики і Банку даних ДНК Японського національного інституту генетики. Ці

бази щоденно обмінюються інформацією. У бази надходять дані про послідовності ДНК і РНК, що отримані із проектів дослідження геномів, наукових публікацій та заявок на патенти. Наукові журнали при прийомі до друку статті вимагають занесення нових послідовностей до вищезгаданих баз.

Системи пошуку GenBank допомагають здійснити пошук інформації про біологічну послідовність, що в ньому міститься.

Для пошуку GenBank необхідно знайти сайт NCBI ([http:// www.ncbi.nlm.nih.gov/genbank/](http://www.ncbi.nlm.nih.gov/genbank/)). На цьому ресурсі потрібно знайти BLAST (Basic Local Alignment Search Tool – ресурс для пошуку локального вирівнювання).



Далі знаходять у білковій базі (protein blast) амінокислотну послідовність із літер, копіюють і натискають кнопку BLAST.

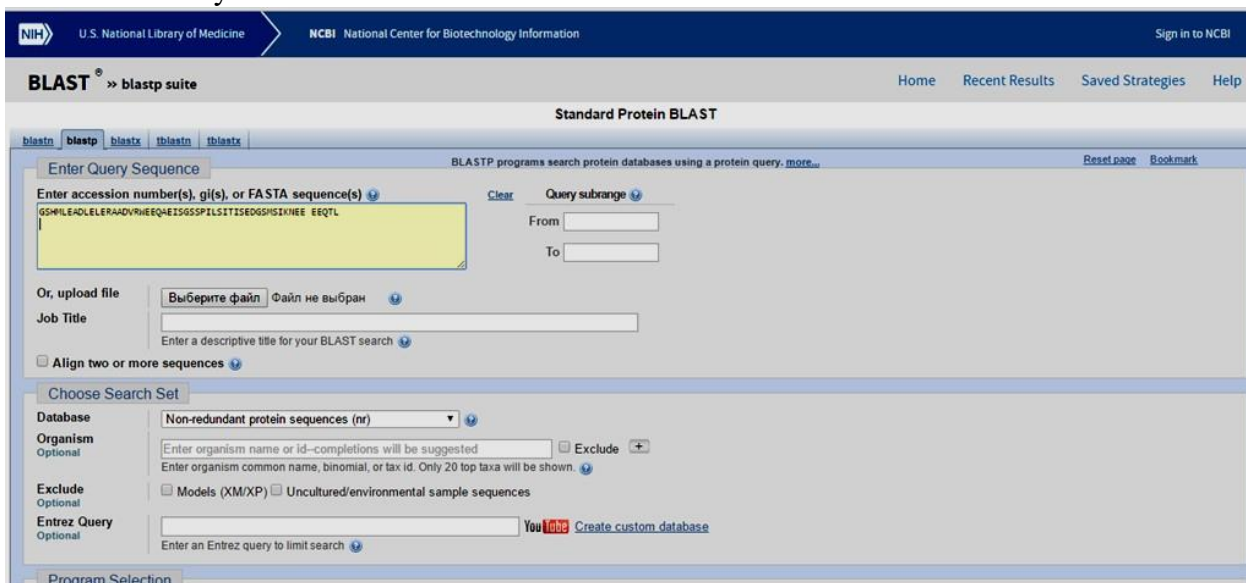


Рисунок 7. Вікно пошуку GenBank

Наприклад, послідовність GSHMLEADLELERAADVRWEEQAEISGSSPIL SITISEDGSM SIKNEEEEQTL за результатом більше всього схожа на послідовність 2FOM Chain A, Dengue Virus Ns2bNS3 PROTEASE (рис. 8).

Descriptions		Graphic Summary	Alignments	Taxonomy				
Sequences producing significant alignments								
Download Select columns Show 100								
select all 100								
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per Ident	Acc. Len	Accession
Chain A, polyprotein [dengue virus type 2]	dengue virus type 2	103	103	100%	2e-27	100.00%	62	2FOM_A
Chain A, NS2B-NS3 protease [dengue virus type 2]	dengue virus type 2	104	104	100%	9e-26	100.00%	247	4M9K_A
Chain A, NS2B-NS3 protease [dengue virus type 2]	dengue virus type 2	104	104	100%	1e-25	100.00%	247	4M9F_A
Chain A, FLAVIVIRUS_NS2B/Peplidase S7 [dengue virus type 2]	dengue virus type 2	102	102	100%	9e-25	98.11%	247	6M00_A
non-structural protein 2B [dengue virus type 2]	dengue virus type 2	90.9	90.9	88%	2e-21	97.87%	130	ALU09479.1
nonstructural protein 2B [Dengue virus]	Dengue virus	90.5	90.5	88%	2e-21	95.74%	130	AHB63925.1
Nonstructural protein NS2B [dengue virus type 2]	dengue virus type 2	89.7	89.7	88%	5e-21	95.74%	130	NP_739586.2
unnamed protein product [dengue virus type 2]	dengue virus type 2	89.4	89.4	88%	6e-21	95.74%	120	CAA28968.1
Chain A, Serine protease subunit NS2B_Serine protease NS3 [Dengue virus 2 Thailand/0168/1979]	Dengue virus 2 Thailand/0168/...	89.7	89.7	88%	5e-20	95.74%	240	2M9P_A
polyprotein [dengue virus type 2]	dengue virus type 2	92.8	92.8	88%	6e-20	100.00%	3310	AOE47562.1
polyprotein [dengue virus type 2]	dengue virus type 2	92.8	92.8	88%	6e-20	100.00%	3360	QMT58633.1
polyprotein [dengue virus type 2]	dengue virus type 2	92.8	92.8	88%	7e-20	100.00%	3357	QMT58631.1
polyprotein [dengue virus type 2]	dengue virus type 2	92.8	92.8	88%	7e-20	100.00%	3357	QMT58635.1
polyprotein [dengue virus type 2]	dengue virus type 2	92.8	92.8	88%	7e-20	100.00%	3357	QMT58637.1
polyprotein [dengue virus type 2]	dengue virus type 2	92.8	92.8	88%	7e-20	100.00%	3391	AYP74869.1

Рисунок 8. Інформація про послідовність у GenBank

Перейшовши за посиланням, можна отримати більш детальну інформацію про послідовність (рис. 9).

Завантажити GenPept Графіка Дати Попередній Описи

Ланцюг А, поліпротеїн [вірус денге типу 2]
 ID послідовності: [2FOM_A](#) Довжина: 62 Кількість збігів: 1

Діапазон 1: від 1 до 53 GenPept Graphics Наступний матч Попередній матч

Оцінка	Очікуйте метод	Ідентичності	Позитиви	Прогалини
103 бйти (257)	2e-27	Коригування композиційної матриці. 53/53(100%)	53/53(100%)	0/53(0%)

Запит 1 GSHMLEADLELERAADVRIEEQAEISGSSPILSITISEDGSMSIKNEEEQTL 53
 GSHMLEADLELERAADVRIEEQAEISGSSPILSITISEDGSMSIKNEEEQTL 53
 Sbjct 1 GSHMLEADLELERAADVRIEEQAEISGSSPILSITISEDGSMSIKNEEEQTL 53

Пов'язана інформація
[Структура](#) - тривимірне відображення структури

Завантажити GenPept Графіка Дати Попередній Описи

Ланцюг А, протеаза NS2B-NS3 [вірус денге типу 2]
 ID послідовності: [4M9K_A](#) Довжина: 247 Кількість збігів: 1

[Переглянути ще 1 заголовок](#) [Переглянути всі ідентичні білки \(IPG\)](#)

Діапазон 1: від 1 до 53 GenPept Graphics Наступний матч Попередній матч

Оцінка	Очікуйте метод	Ідентичності	Позитиви	Прогалини
104 бйти (260)	9e-26	Коригування композиційної матриці. 53/53(100%)	53/53(100%)	0/53(0%)

Запит 1 GSHMLEADLELERAADVRIEEQAEISGSSPILSITISEDGSMSIKNEEEQTL 53
 GSHMLEADLELERAADVRIEEQAEISGSSPILSITISEDGSMSIKNEEEQTL 53
 Sbjct 1 GSHMLEADLELERAADVRIEEQAEISGSSPILSITISEDGSMSIKNEEEQTL 53

Пов'язана інформація
[Структура](#) - тривимірне відображення структури
[Ідентичні білки](#) - ідентичні білки 4M9K_A

Рисунок 9. Детальна інформація про послідовність у GenBank

Protein Data Bank

Protein Data Bank заснований у 1971 році Уолтером Гамільтоном у Національній лабораторії Брукхавена. Він поповнюється даними про просторову структуру білків і нуклеїнових кислот, що отримані за допомогою рентгенівської кристалографії. Якщо послідовність можливо знайти у Protein Data Bank, то це значить, що для цієї послідовності можна отримати тримірну структуру.

Зайти у базу даних Protein Data Bank можна за посиланням www.pdb.org. Відкриємо сайт бази (рис. 10, 11) і спробуємо знайти вищезгадану послідовність. Ввівши ідентифікатор у рядку пошуку, отримаємо інформацію про послідовність.

WORLDWIDE PDB PROTEIN DATA BANK

VALIDATION • DEPOSITION • DICTIONARIES • DOCUMENTATION • TASK FORCES • DOWNLOADS • STATISTICS • ABOUT

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

Celebrating 50 Years of the PDB

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 20 Years of the wwPDB Partnership

Validate Structure
or View validation reports

Deposit Structure
All Deposition Resources

Download Archive
Instructions

Vision and Mission

Vision

Sustain freely accessible, interoperating Core Archives of structure data and metadata for biological macromolecules as an enduring public good to promote basic and applied research and education across the sciences.

Mission

- Manage the wwPDB Core Archives as a public good according to the FAIR Principles.
- Provide expert deposition, validation, biocuration, and remediation services at no charge to Data Depositors

wwPDB Resources

Data Dictionaries

- Macromolecular Dictionary (PDBx/mmCIF)
- Small Molecule Dictionary (CCD)
- Peptide-like antibiotic and inhibitor molecules (BIRD)

Biocuration

- Procedures and policies
- Improvements for consistency and accuracy

Community Input:
Task Forces and Working Groups

News & Announcements

03/21/2024

Paper Published on NMR Restraint Validation

Read the wwPDB validation improvements on NMR restraint analysis with standardized NEF and NMR-STAR formats

Read more

03/12/2024

oneDep **wwPDB Validation System** FAQ Tutorials

Existing validation

Validation ID

Password

Log in

Forgot Password

Deposition server

Deposit your data to PDB, BMRB and EMBD at deposit.wwpdb.org

wwPDB news and announcements

Compliance with GDPR legislation

wwPDB has revised its [privacy policy](#) in line with the requirements of the EU's GDPR legislation.

Old Validation IDs will no longer be accepted

Due to server reconfiguration, if you have a validation ID that starts with D_90, you will need to create a new session. We apologize for any inconvenience.

Start a new validation

Welcome to the wwPDB validation system!

This server runs the same validation as you would observe during the deposition process. This service is designed to help you check your model and experimental files prior to start of deposition. To continue with an existing validation session, please login on the left.

To start a new validation, please complete the form below. Upon completion, you will be emailed login information specific to your new validation.

Your e-mail address

Password (optional, or we will provide one)
This is a shared "group password"
(5 to 16 alphanumeric characters)

Country/Region

Experimental method

X-Ray Diffraction

Electron Microscopy

Reset

Рисунок 10-11. Вікно Protein Data Bank

Послідовність, що досліджується, належить протеазі NS2b/NS3 вірусу Денге (рис. 12), модель якої опублікована у 2006 році.

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

View: Detailed **Reports:** Select a Report **Sort:** e-value: Best to Worst **Download Files**

4M9K: Entity 1 containing Chain A **Download File** **View File**

NS2B-NS3 protease from dengue virus at pH 5.5

[Yildiz, M., Ghosh, S., Bell, J.A., Sherman, W., Hardy, J.A.](#)

(2013) ACS Chem Biol **8** 2744-2752

Released: 11/27/2013 **Macromolecule:** NS2B-NS3 protease (protein)

Method: X-ray Diffraction **Unique Ligands:** --

Resolution: 1.46 Å **Residue Count:** 247

Length: 53 **E-value:** 3.13264E-23 **Score:** 104.76bits (260) **Identities:** 53/53 (100%) **Positives:** 53/53 (100%) **Gaps:** 0/53 (0%)

Query GSH*LEADLELEERAADVRIEEQAEISGSSPILSITISEDGSHSIKNEEEQTL

GS*H*LEADLELEERAADVRIEEQAEISGSSPILSITISEDGSHSIKNEEEQTL

Sbjct GSH*LEADLELEERAADVRIEEQAEISGSSPILSITISEDGSHSIKNEEEQTL

Рисунок 12. Результати дослідження послідовності у Protein Data Bank

Існує можливість прокрутити тримірну структуру моделі у браузері, скориставшись

функцією 3D View (рис. 13), підібрати моделі, схожі за амінокислотною послідовністю чи за структурою і т.д.



Рисунок 13. Прокрутка тримірної структури моделі у Protein Data Bank

Робота з PubMed

PubMed – це джерело біологічних публікацій (рис. 14). Ознайомитися з ним можна на сайті NCBI (Національного Центру Біоінформатики США (National Center for Biotechnology Information)).

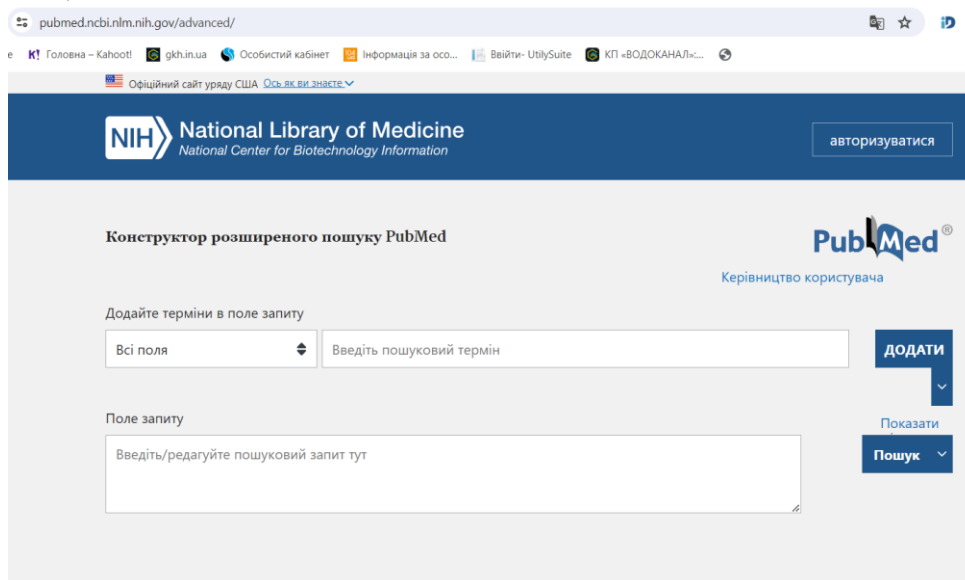


Рисунок 14. Головний пошуковий екран PubMed

Раніше PubMed містив лише публікації медичного спрямування, а зараз – містить всі біологічні статті. Ресурс має зручний інструмент пошуку (рис. 15).

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed®

virus dengue

Advanced Create alert Create RSS Search User Guide

Save Email Send to Sort by: Best match Display options

MY NCBI FILTERS 19,278 results Page 1 of 1,928

RESULTS BY YEAR

1924 2024

TEXT AVAILABILITY

Abstract

Free full text

Full text

Dengue virus pathogenesis: an integrated view.

1 Martina BE, Koraka P, Osterhaus AD.
Cite Clin Microbiol Rev. 2009 Oct;22(4):564-81. doi: 10.1128/CMR.00035-09.
PMID: 19822889 [Free PMC article](#). [Review](#).
Share Much remains to be learned about the pathogenesis of the different manifestations of **dengue virus** (DENV) infections in humans. They may range from subclinical infection to **dengue fever**, **dengue hemorrhagic fever** (DHF), and eventually **dengue shock** ...

Clinical and Laboratory Diagnosis of Dengue Virus Infection.

2 Muller DA, Depelseinaire AC, Young PR.
Cite J Infect Dis. 2017 Mar 1;215(suppl_2):S89-S95. doi: 10.1093/infdis/jiw649.
PMID: 28403441 [Review](#).
Share Infection with any of the 4 **dengue virus** serotypes results in a diverse range of symptoms, from mild

Рисунок 15. Результат пошуку в PubMed

PubMed – це електронно-пошукова система, що включає:

- MEDLINE;
- PreMEDLINE;
- Видавничі описи.

MEDLINE – база даних медичної інформації, що містить бібліографічні описи більше 4800 медичних періодичних видань від початку 1960 р.р. База знаходиться у відкритому доступі і включає описи статей із медичних журналів на 30 мовах. Більше 70 % описів мають реферати. За тематикою MEDLINE включає широкий спектр галузей, що відносяться до біології та медицини:

- наукові дослідження та їх методологія;
- клінічна практика;
- медичні аспекти біології;
- стоматологія;
- фармакологія.

У PubMed передбачено декілька варіантів пошуку:

- пошук за ключовими словами (можна проводити як за ключовими словами, так і за термінами);
- пошук за покажчиком медичних предметних рубрик (MeSH Translation Table) – система відбирає всі документи, що включені у предметну рубрику і документи, що містять даний термін у якості текстового слова;
- пошук за покажчиком журналів (Journals Table) – терміни пошуку перевіряються у покажчику журналів. Покажчик включає повні заголовки журналів та аббревіатури, міжнародні серійні номери (ISSN);
- пошук за покажчиком фраз (Phrase List) – одна і та ж фраза буде шукатися у всіх пошукових полях системи; якщо така фраза не знаходиться, то система здійснить пошук за окремими словами фрази;
- пошук за авторським покажчиком (Author Index) – терміни перевіряються у авторському покажчику, якщо заданий параметр пошуку буде містити ініціали;
- пошук терміна за коренем слова (Truncation) – використовується для пошуку різних варіантів одного і того ж слова; для пошуку необхідно замінити закінчення слова у рядку пошуку на

зірочку (*);

- пошук з допомогою логічних операторів (Boolean Operators) – можливе використання логічних операторів між термінами, щоб знайти їх комбінації в одному документі (AND – якщо обидва слова повинні бути знайдені в документі; OR – якщо хоча б одне зі слів повинне знайтися у документі; NOT – якщо лише один зі термінів повинен знаходитися у документі за умови, що іншого там не буде);
- пошук за обмеженнями (Limits) – функція, що доступна на різних етапах пошуку і застосовується щоб звужити коло результатів.

PubMed має низку можливостей щодо:

- сортування та збереження результатів пошуку,
- замовлення повнотекстових версій у авторів,
- отриманню результатів пошуку електронною поштою.
- крім цього можна використовувати суміжні із PubMed пошукові ресурси.

Не всі знайдені статті можуть бути представлені у відкритому доступі. Частина результатів пошуку може бути представлена лише у вигляді анотацій (повнотекстові бази даних дають доступ до повних версій статей, на відміну від реферативних). Це є особливістю бази даних, за допомогою якої проводився пошук інформації. Наукові журнали можуть працювати за принципом трьох бізнес-моделей:

- традиційної – за перегляд статті платять самі читачі чи бібліотеки;
- модель відкритого доступу – за прийом статті до редакції та її відкритий доступ читачам платять самі автори;
- гібридна модель – автори самі вирішують у якому доступі буде знаходитися стаття, у разі відкритого – сплачують самі.

У разі традиційної моделі постає проблема пошуку повнотекстової версії статті. Більшість баз даних мають посилання на електронну адресу авторів/публікації. Написавши авторам на вказані адреси, можна отримати повнотекстову версію статті від самих авторів.