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.

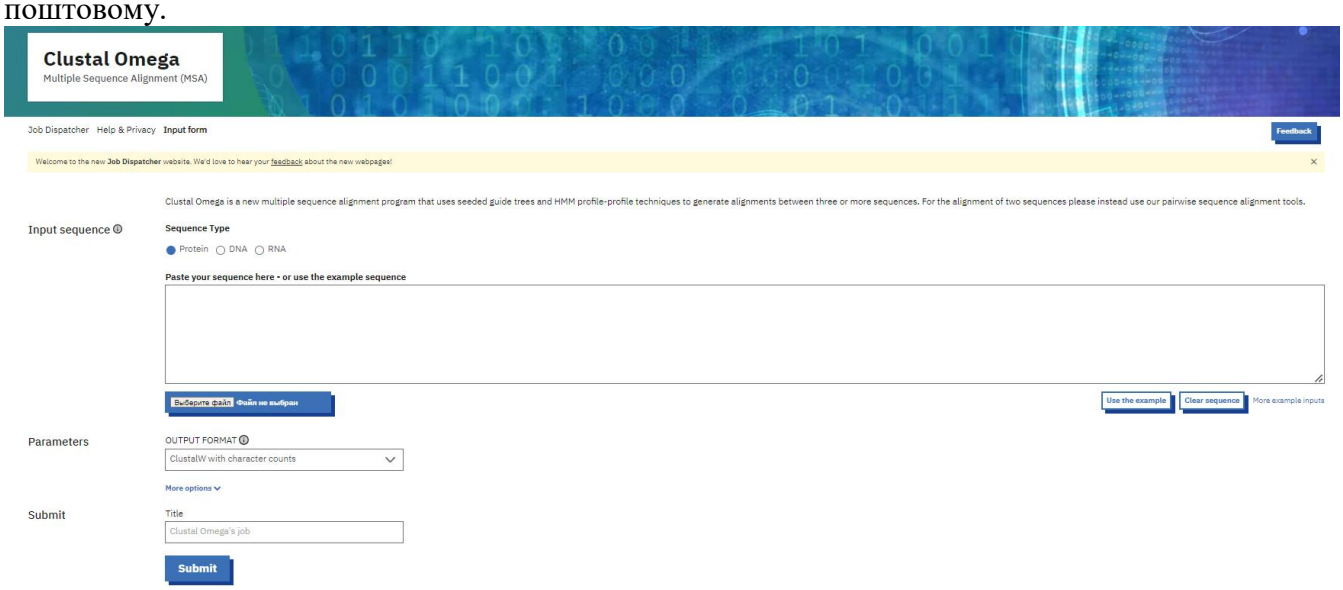


Рисунок 1. Вікно програми Clustal

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

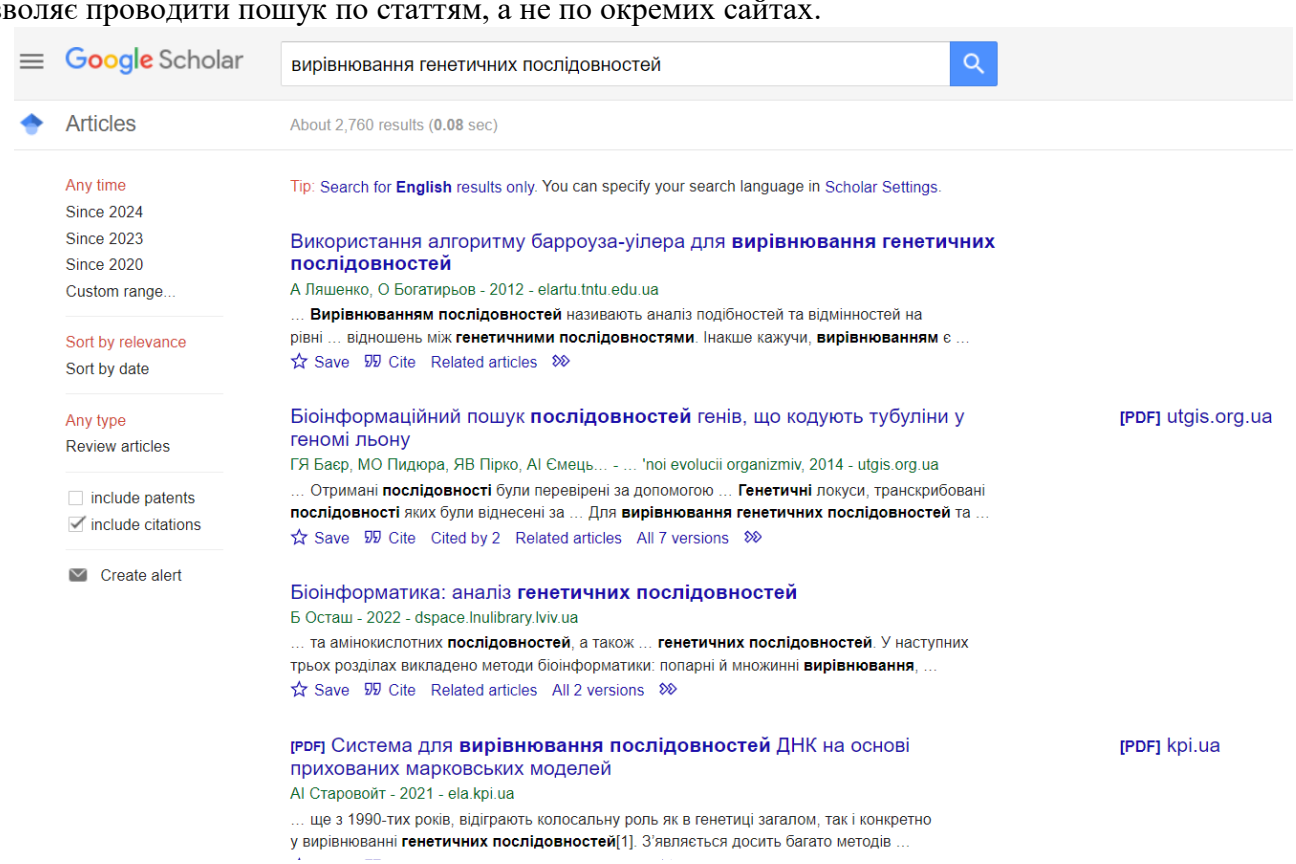


Рисунок 2. Інтерфейс Google Scholar

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.

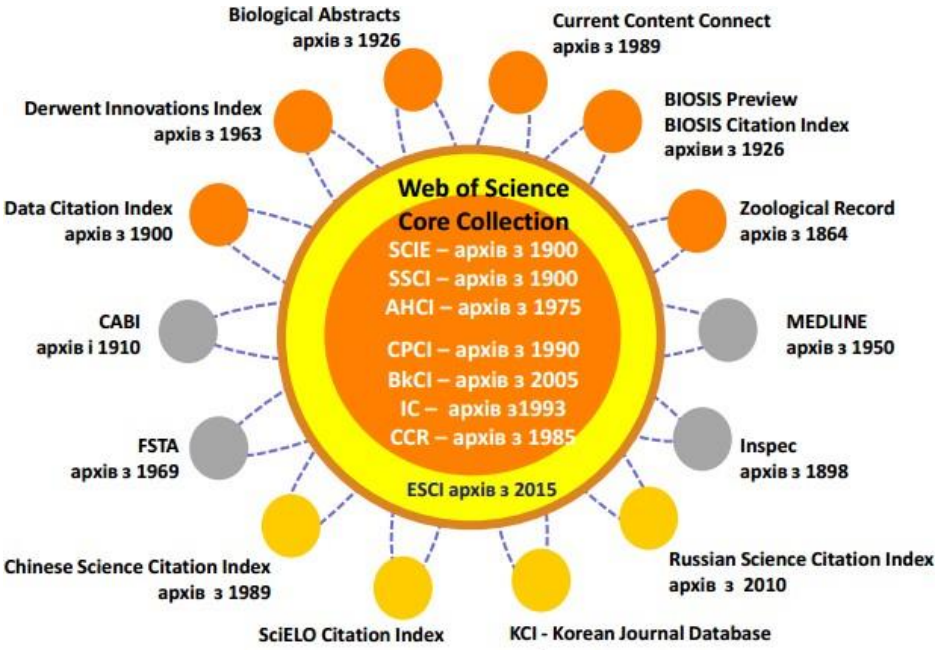


Рисунок 3. Бази даних на платформі Web of Scince

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 імпакт-фактор обраного видання;
- знайти колег у світі, які займаються даною проблемою;
- зберігати знайдені публікації;
- переоформити матеріал для іншого видання.

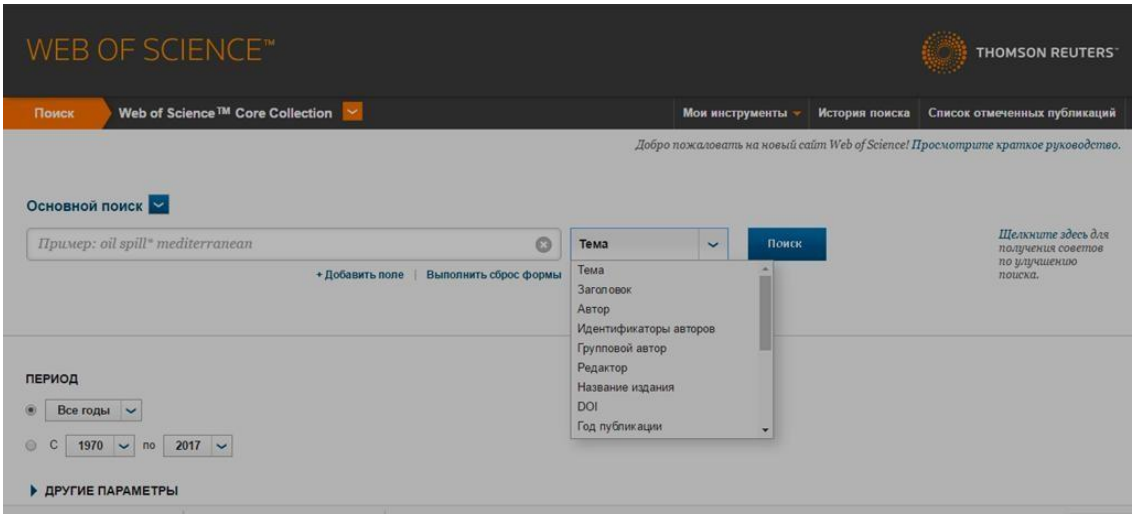


Рисунок 4. Вікно пошуку платформи Web of Science

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» воєдино одну і ту ж статтю, що міститься урізних файлах. Корисною функцією програми є автоматичне створення посилань у тексті та списку використаних джерел.

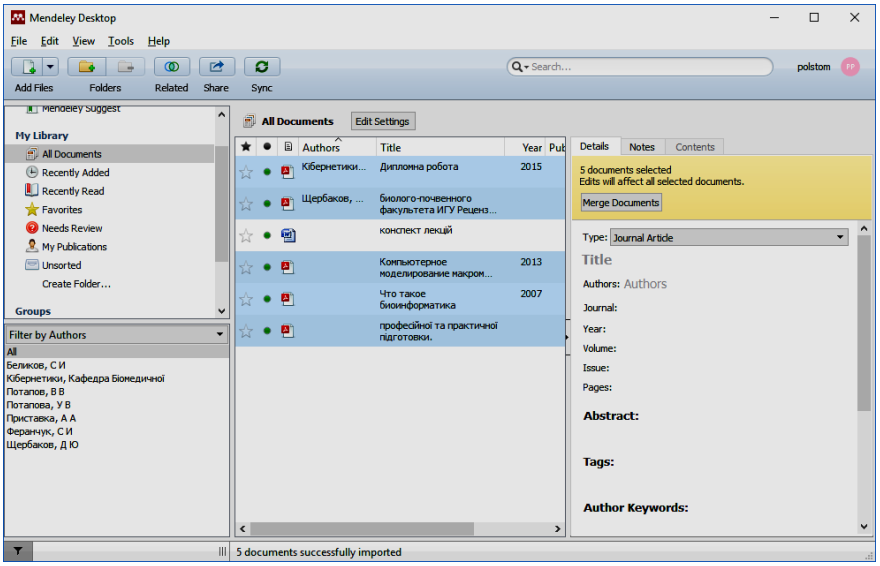


Рисунок 5. Вікно програми Mendeley

Схожою за функціями є програма Red Cube Web Reader.

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

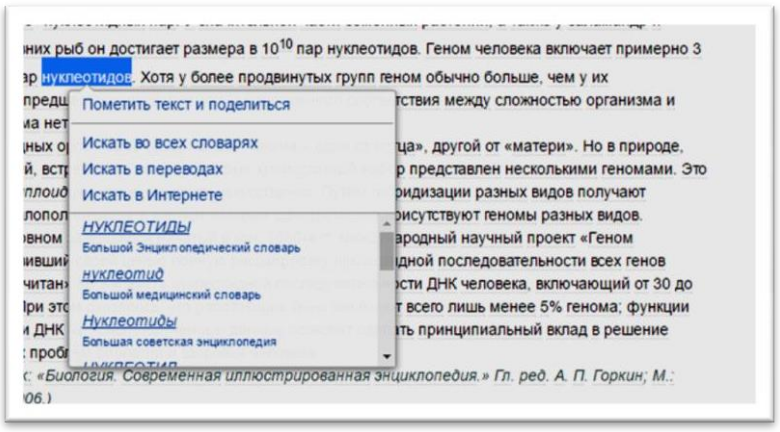


Рисунок 6. Використання плагіну reflect.ws

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.

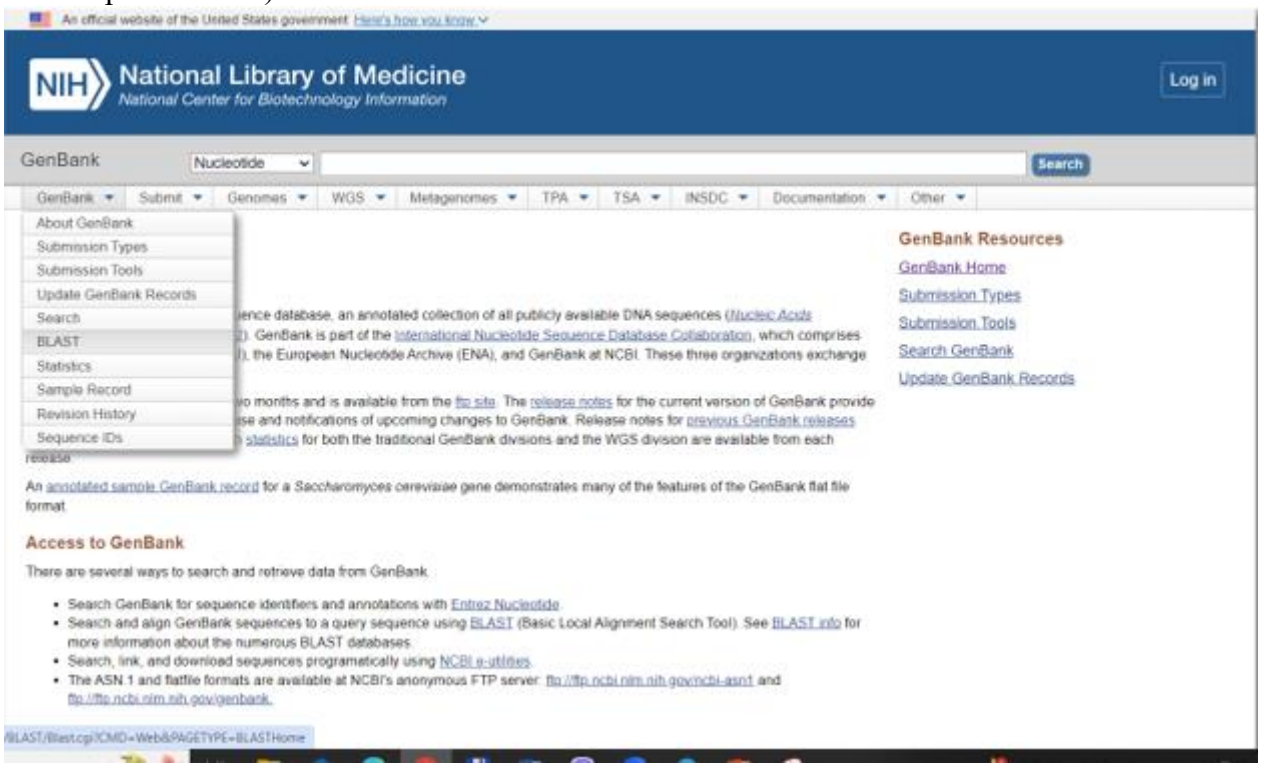
База даних генетичних послідовностей GenBank

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/). 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.

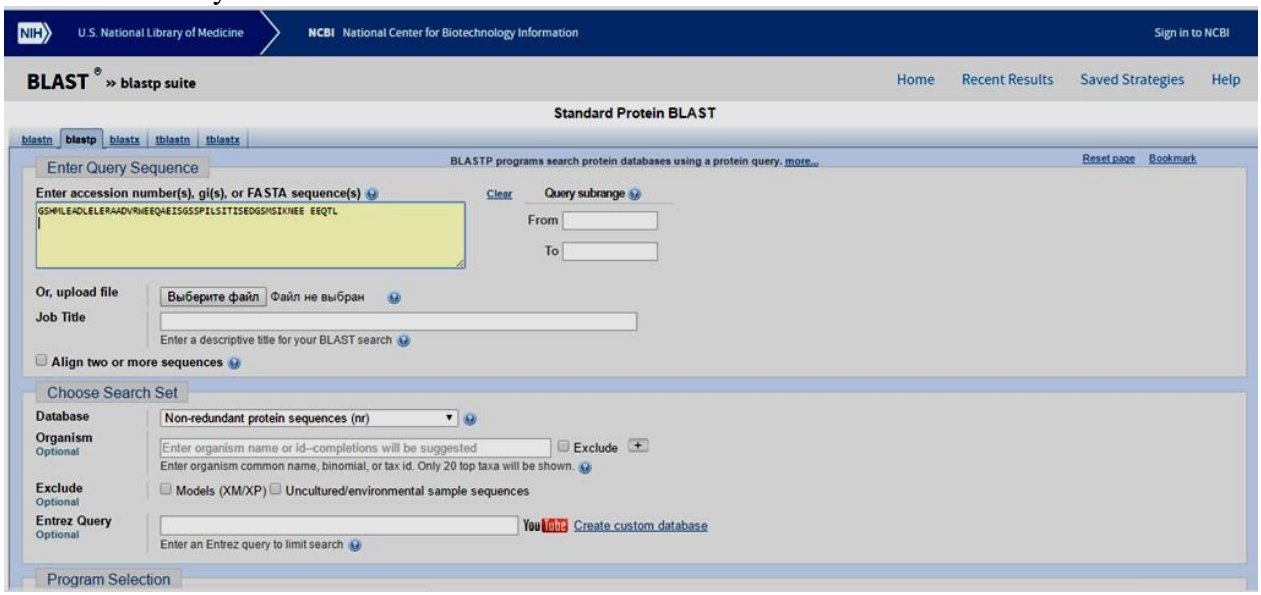


Рисунок 7. Вікно пошуку GenBank

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

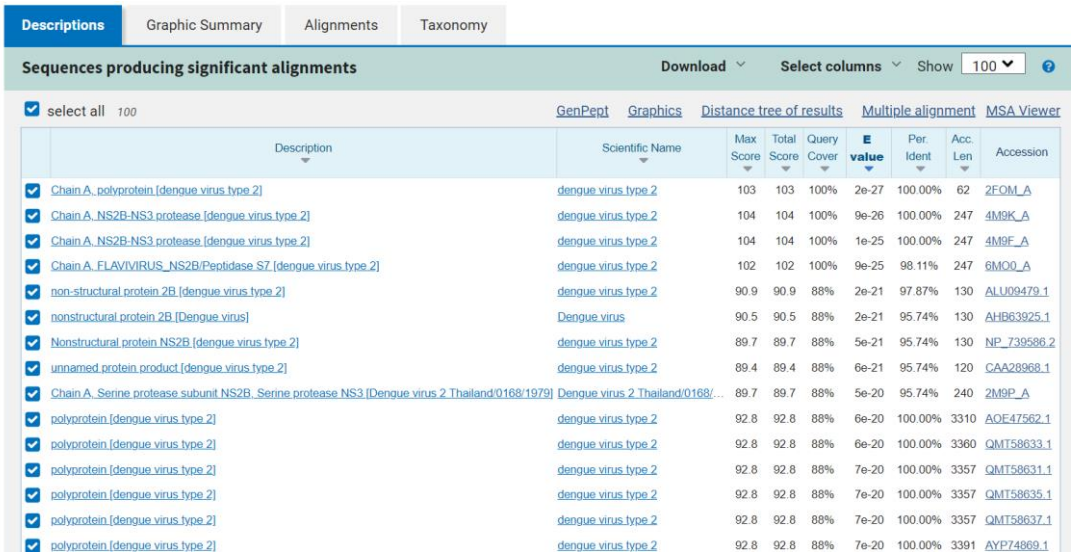


Рисунок 8. Інформація про послідовність у GenBank

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

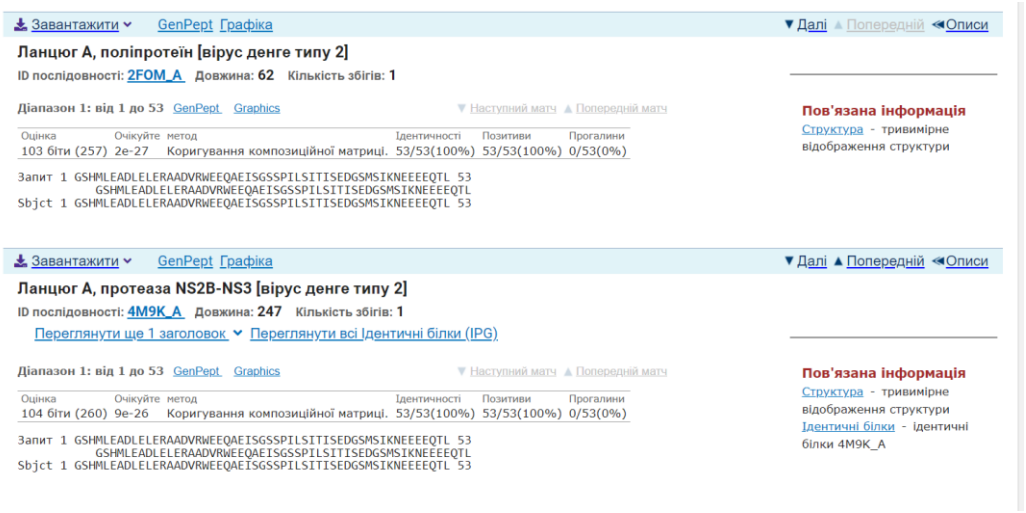


Рисунок 9. Детальна інформація про послідовність у GenBank

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.

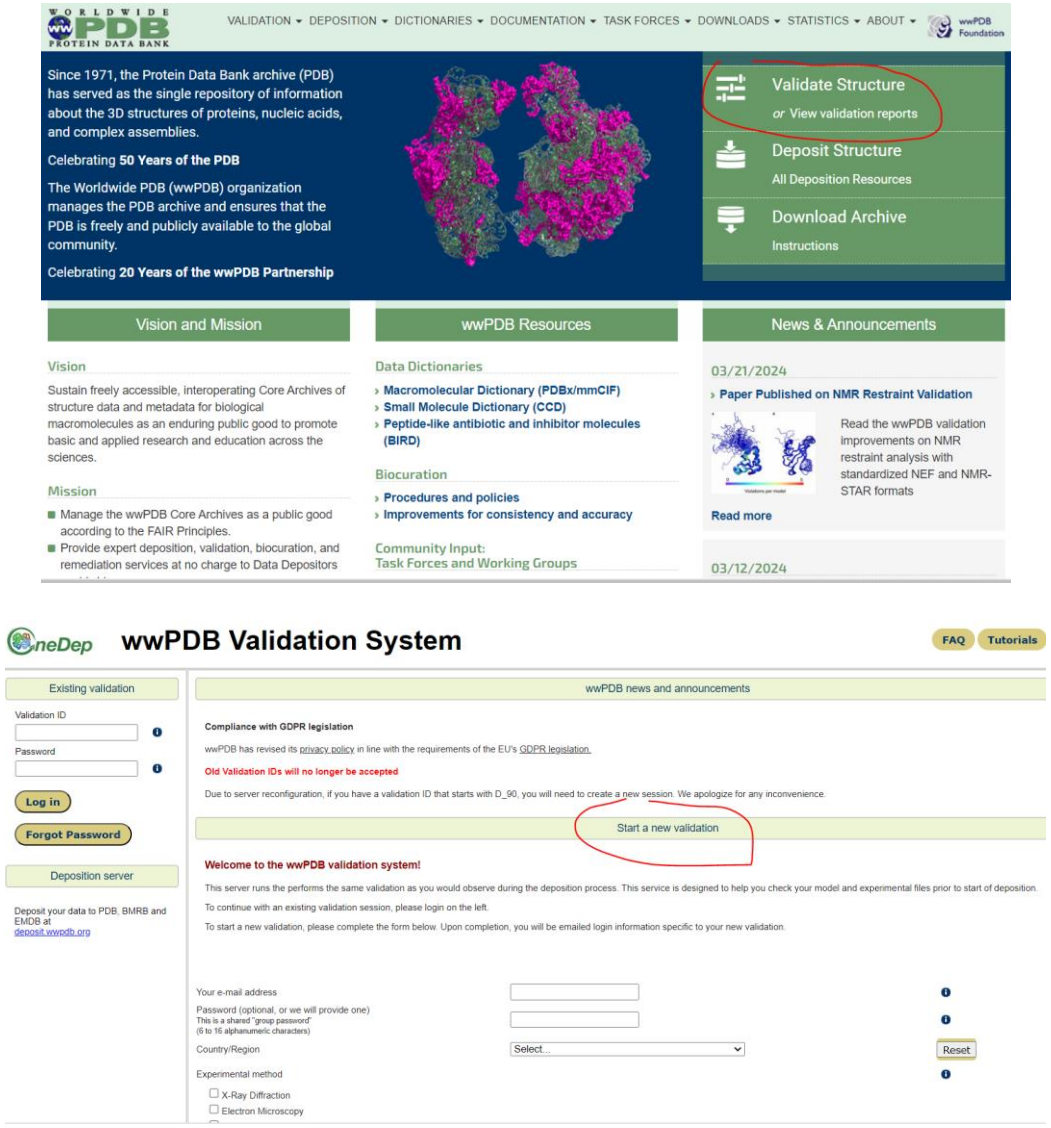


Рисунок 10-11. Вікно Protein Data Bank

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

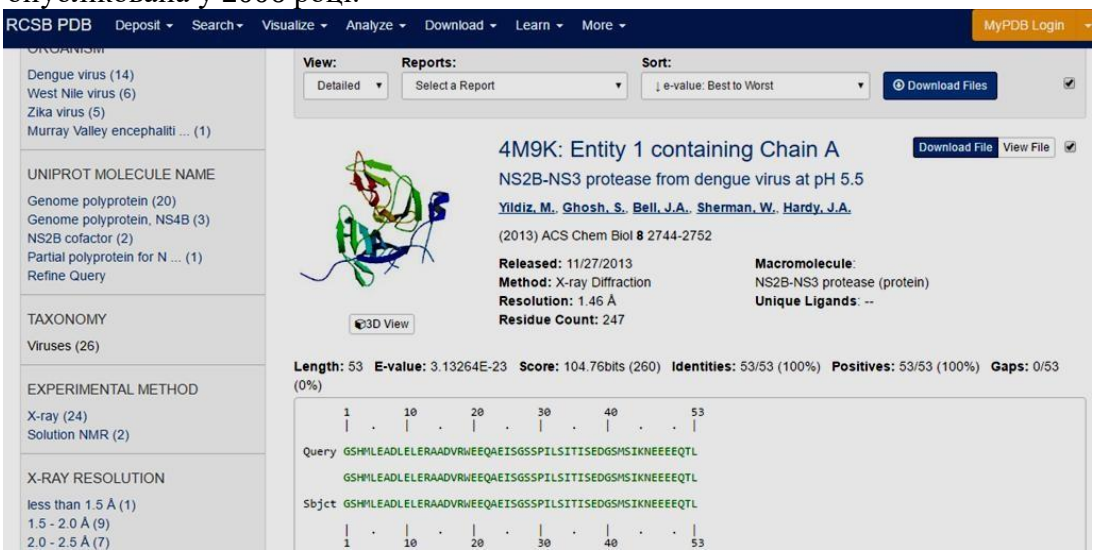


Рисунок 12. Результати дослідження послідовності у Protein Data Bank

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.

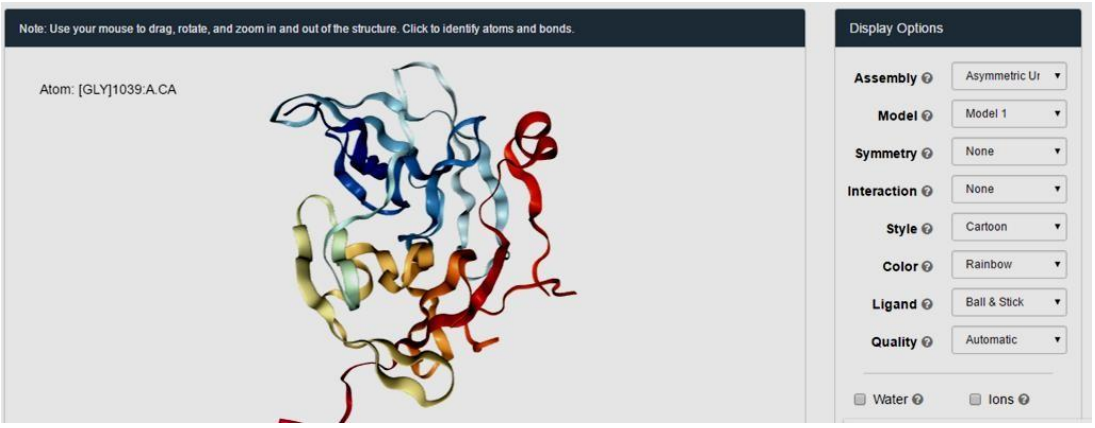


Рисунок 13. Прокрутка тримірної структури моделі у Protein Data Bank

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).

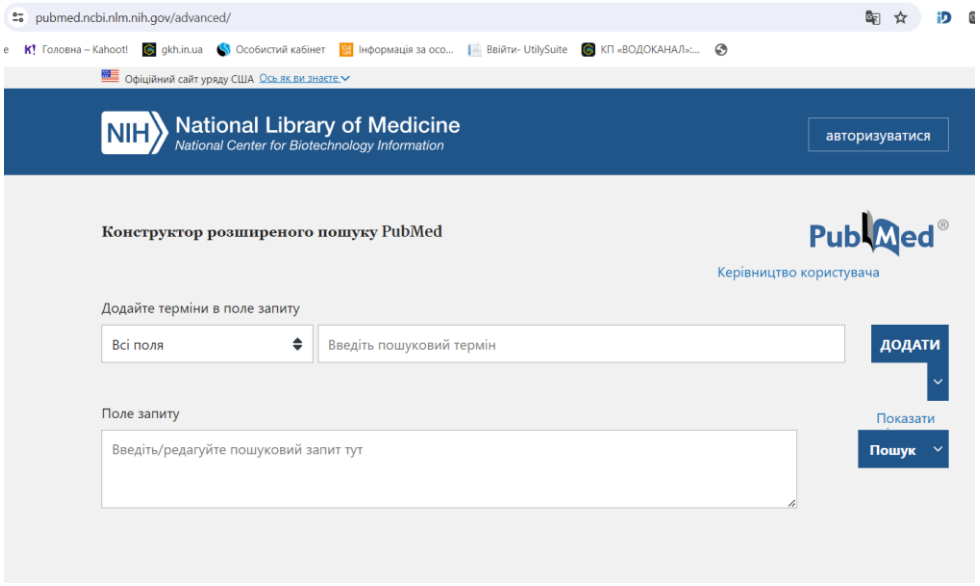


Рисунок 14. Головний пошуковий екран PubMed

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

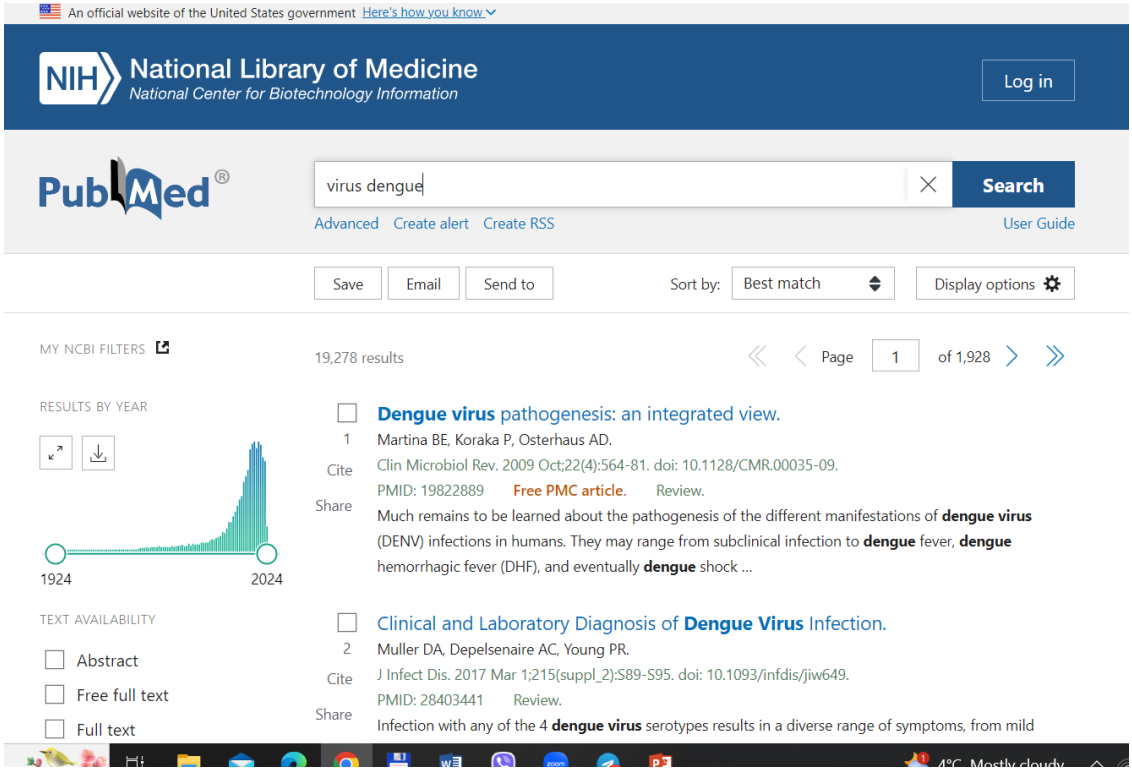


Рисунок 15. Результат пошуку в PubMed

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 місять даний термін у якості текстового слова;
- пошук за покажчиком журналів (Journals Table) – терміни пошуку перевіряються у покажчику журналів. Покажчик включає повні заголовки журналів та абревіатури, 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 незнаходиться, 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 звузити 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 авторівпублікації. Написавши авторам на вказані адреси, можна отримати повнотекстову версію статті від самих авторів.