МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

С.В. Горобець, О.Ю. Горобець, І.В. Дем'яненко

БІОІНФОРМАТИКА
ПРАКТИКУМ

Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для студентів,
які навчаються за спеціальністю 162 «Біотехнологія та біоінженерія»

Київ
КПІ ім. Ігоря Сікорського
2020

Рецензент *Галкін О.Ю.,* професор, д.б.н., завідувач кафедри трансляційної медичної біоінженерії факультету біомедичної інженерії Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського»

Відповідальний редактор *Горго Юрій Павлович*, професор, д.б.н.,

Гриф надано Методичною радою КПІ ім. Ігоря Сікорського (протокол № від р.)
за поданням Вченої ради факультету біотехнології і біотехніки (протокол № від р.)

Електронне мережне навчальне видання

Горобець Світлана Василівна, д-р техн. наук, проф.
Горобець Оксана Юріївна, д-р фіз.-мат.наук, проф.
Дем'яненко Ірина Володимирівна, к-т. техн. наук.

БІОІНФОРМАТИКА
ПРАКТИКУМ

Біоінформатика. Практикум [Електронний ресурс] : навч. посіб. для студ. спеціальності 162 «Біотехнологія та біоінженерії» / КПІ ім. Ігоря Сікорського; автор.: С.В. Горобець, О.Ю.Горобець, І.В. Дем'яненко. – Електронні текстові дані (1 файл: 12,78 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2020. – 87 с.

В навчальний посібник «Біоінформатика. Практикум» входить ряд унікальних розроблених авторами практичних робіт, які дозволять здобувачам вищої освіти освоїти новітні біоінформатичні методи дослідження біологічних агентів (ДНК, РНК, білків). В процесі виконання практичних завдань здобувачі освоюють навички роботи з різними біоінформатичними прикладними програмами, навчаться будувати філогенетичні дерева, шукати промотери для постановки полімеразно-ланцюгової реакції.

Даний посібник рекомендується до використання в освітній діяльності для забезпечення підготовки бакалаврів та магістрів спеціальності 162-Біотехнології та біоінженерія.

Зміст

Вступ

За останні 25-30 років науковцями світу створено колосальний теоретичний та експериментальний матеріал про будову і функціонування біологічних молекул (білків і нуклеїнових кислот). Цей матеріал потребує розвинутих комп'ютерних методів для свого аналізу, саме тому біоінформатика широко застосовується для розв'язання задач генетики, молекулярної біології, молекулярної біотехнології, біохімії та ін. На сьогодні ці дисципліни є лідерами як за обсягом надходжень нової інформації, так і за темпами її впровадження в різні технології. Математикам, наприклад, від обчислення площі кола до створення перших обчислювальних машин знадобилося більше 2000 років, а фізикам від відкриття закону всесвітнього тяжіння до польотів у космос – біля 300 років. Молекулярній же біології від відкриття структури ДНК та принципів кодування генетичної інформації у 1953 році до її широкомасштабного промислового використання знадобилося всього біля 30 років.

У 1951 році – американський хімік і фізик К. Л. Полінг разом із американським біохіміком Р. Б. Корі розробили уявлення про структуру поліпептидного ланцюга у білках, вперше висунули гіпотезу про її спіральну будову та описали альфа-спіраль.

1953 рік став знаковим та ознаменував початок нової епохи для генетиці та молекулярній біології з відкриттям подвійної спіралі ДНК 25-річним американцем Джеймсом Уотсоном та 36-річним англійцем Френсісом Кріком.

У 1962 році відкривачі вторинної структури ДНК стали Нобелівськими лауреатами. Першим розшифрованим білком був бичачий інсулін (1956 р.), що складався із 51 амінокислотного залишку. Приблизно десятиліттям пізніше була розшифрована перша нуклеїнова послідовність тРНК дріжджів із 77 основ.

У 1965 році Маргарет Дейхофф з співробітниками Національного Фонду біомедичних досліджень (Вашингтон) систематизувала усі наявні дані про амінокислотні послідовності і створила першу біоінформаційну базу даних – атлас білкових послідовностей та їх структур, яка містила опис 65 послідовностей. Через декілька років після розшифровки перших амінокислотних послідовностей були створені перші біоінформаційні (у той час – біологічні) бази даних білків і нуклеотидних послідовностей.

А у 1982 році були організовані банки даних нуклеотидних послідовностей – GenBank (база послідовностей ДНК) у США та EMBL (Європейська молекулярно-біологічна бібліотека) у Європі. Провив в молекулярній біології відбувся з відкриттям зворотньої транскрипції ДНК та клонуванням у 70-х 20 століття. Зі зростанням ринкового попиту на секвенування послідовностей та цілих геномі було сконструйовано автоматичний секвенатор (1987 рік, Л. Худ, Т. Хункапиллер). Обробка та аналіз цієї інформації традиційними статистичними методами стали практично неможливими.

З накопиченням інформації сформувався новий науковий напрям – біоінформатика в кінці 80-х років та почали активно розвиватися біологічні, біоінформатичні та ін. бази даних. Знання зростали експотенційно. Науковці

все більше дізнавалися про структуру та функції окремих ділянок геному (генів), пов’язували їх з білками або шукали їх фенотиповий прояв. Наразі біоінформатичні методи можна зустріти в усіх галузях пов’язаних з ДНК, РНК та білками. Саме тому сучасним здобувачам вищої освіти необхідно мати ґрунтовні знання з новітніх методів та алгоритмів біоінформатики. Це і є основною метою даного навчального посібника.

Практична робота 1

Вирівнювання нуклеотидних та амінокислотних послідовностей за допомогою пошукової системи BLAST в NCBI

Мета роботи: Оволодіти навиками проведення пошуку гомологів білків методами порівняльної геноміки

Теоретичні відомості

Національний центр біотехнологічної інформації США (англ. National Center for Biotechnological Information, NCBI) заснований у 1988 році в Бетесда (штат Меріленд, США) як центральний інститут обробки і зберігання даних молекулярної біології. Є частиною Національної медичної бібліотеки США (англ. United States National Library of Medicine, NLM), підрозділу Національного інституту здоров'я (англ. National Institutes of Health, NIH).

Керує центром Девід Ліпман, один з авторів програми пошуку локальних вирівнювань BLAST і широко визнаний професіонал в області біоінформатики. Він також керує науковими програмами центру, включаючи наукові групи Стефана Альтшуля (співавтор програми BLAST), Девіда Ландсмана, Євгенія Куніна.

NCBI – це сукупність великої кількості баз даних з різних галузей знань. Вона надає інформацію про різні аспекти біотехнології, молекулярної біології, медицини та ін. Наприклад, база даних PubMed є електронною бібліотекою наукових публікацій, де частина статей є у вільному доступі (PubMedCenter). Або є одним з найбільших сховищ даних ДНК (GenBank). Також NCBI має набір інструментів та прикладних програм для аналізу даних (BLAST, Cn3D і т.д.). Для пошуку інформації в базі даних використовують систему Entrez.

Завдання NCBI:

- Створення автоматизованих систем для зберігання та аналізу даних з молекулярної біології, біомедицині та генетики.
- Комп'ютерна обробка даних, отриманих у дослідженнях структури і значення біологічно активних молекул і речовин.
- Сприяння широкому використанню баз даних і програмного забезпечення для дослідників у галузі біотехнологій та медичного персоналу.
- Координування зусиль по накопиченню біотехнологічної інформації по всьому світу.

Куратори NCBI приділяють велику увагу освітньому сегменту та організації наукових онлайн конференцій. На інформаційному порталі є розділ з навчальними матеріалами, які описують та навчають роботі з базою даних.

Одним з частіш використовуваних інструментів є BLAST (англ. Basic Local Alignment Search Tool). Це- сімейство комп'ютерних програм для пошуку гомологів білків або нуклеїнових кислот, для яких відома первинна структура (послідовність) або її фрагмент.

Програма BLAST створена групою вчених Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, і David J. Lipman на замовлення Національного

інституту охорони здоров'я США і вперше повідомлення про неї з'явилося в журналі Journal of Molecular Biology (англ.) в 1990 році [1].

До сімейства програм групи BLAST входять 4 підгрупи :

Нуклеотидні: призначені для порівняння досліджуваної нуклеотидної послідовності з базою даних секвенованих нуклеїнових кислот та їх ділянок:

megablast – швидке порівняння з метою пошуку високоподібних послідовностей,

dmegablast – швидке порівняння з метою пошуку дивергованих послідовностей, що мають незначну схожістю,

blastn - повільне порівняння з метою пошуку всіх подібних послідовностей та ін.

Білкові: призначені для порівняння досліджуваної амінокислотної послідовності білка з наявною базою даних білків і їх ділянок.

blastp – повільне порівняння з метою пошуку всіх подібних послідовностей,

cdart – порівняння з метою пошуку гомологічних білків з доменною структурою,

rpsblast – порівняння з базою даних консервативних доменів,

psi-blast – порівняння з метою пошуку послідовностей, що мають незначну схожістю,

phi-blast – пошук білків, що містять певний патерн та ін.

Транслюючі: здатні транслювати нуклеотидні послідовності в амінокислотні:

blastx – переводить досліджувану нуклеотидну послідовність у амінокислоту, а потім порівнює її з наявними в базі даних амінокислотними послідовностями білків;

tblastn – амінокислотна послідовність, що вивчається, порівнюється з трансльованими послідовностями з бази даних;

tblastx – переводить досліджувану нуклеотидну послідовність у амінокислотну, а потім порівнює її з трансльованими послідовностями.

Геномні: призначені для порівняння досліджуваної нуклеотидної послідовності будь-яких організмів (людини, миші та ін.) з базою даних [2].

Принципи роботи BLAST

Є декілька алгоритмів, які використовують для пошуку гомологів серед амінокислотних та нуклеїнових послідовностей. Найчастіше використовують алгоритм глобального (послідовності порівнюються повністю) та локального (порівнюються лише певні ділянки послідовностей) вирівнювання. Власне BLAST засновано на алгоритмі локального вирівнювання, що пов'язано з наявністю у різних білках подібних доменів. Крім цього локальне вирівнювання дозволяє порівняти іРНК з геномною ДНК.

Після введення досліджуваної нуклеотидної або амінокислотної послідовності (запит) на одну з веб-сторінок BLAST, вона разом з іншою вхідною інформацією (база даних, розмір «слова» (ділянки), значення величини

Е-числа та ін.) надходить на сервер. BLAST створює таблицю всіх «слів» (у білку – це ділянка послідовностей, які за замовчуванням складається з трьох амінокислот, а для нуклеїнових кислот з 11 нуклеотидів) з цих послідовностей.

Потім в базі даних проводиться пошук. При встановленні відповідності проводять автоматично збільшення довжини слова (для амінокислот – 4 і більше, для нуклеотидів – 12 і більше), після цього використовують вставлення пробілів (гепів). Коли досягається максимальний розмір слів, програма визначає вирівнювання з найкращими збігами (тимчасово зберігається в SeqAlign) та формує документ виходячи з отриманих даних у формі таблиць та графічного зображення.

Для кожної виявленої в базі даних послідовності програма BLAST визначає, наскільки ця послідовність схожа з досліджуваною послідовністю (запитом) і наскільки значиме це вирівнювання. Гомологію визначають за значенням Е-числа, що розраховується статистичним методом.

При визначенні гомології послідовностей ключовим елементом є матриці замін PAM або BLOSUM, так як вони визначають показники подібності для будь-якої можливої пари нуклеотидів або амінокислот. У більшості програм серії BLAST використовується матриця BLOSUM62 (Blocks Substitution matrix 62 % identity, блокова матриця замін з 62 % ідентичності).

Визначення пари сегментів відбувається за допомогою модифікованих алгоритмів Сміта-Уотермана або Селлерса. Такі пари видовжених «слів» називаються парами сегментів з максимальною схожістю (high-scoring segment pairs, HSP). У разі достатньо великої довжини досліджуваних послідовностей (m) і послідовностей бази даних (n) показники схожості HSP характеризуються двома параметрами K (розмір області пошуку) і P (системи підрахунку) від яких залежать результативні показники подібності досліджуваної послідовності та послідовності бази даних (S).

Для порівняння показників подібності (B) їх перетворюють використовуючи наступну формулу:

$$B = (P \cdot S - \ln K) / l n 2 \qquad (1.1)$$

Величина B показує, наскільки схожі послідовності (чим більше число бітів, тим більше схожість) та відповідає величині E (E-value), визначається за формулою:

$$E = m \cdot n \cdot 2^{-B} \qquad (1.2)$$

Програми BLAST переважно визначають значення E, а не P (імовірності наявності хоча б одного HPS з показником, що перевищує або дорівнює S). Але при $E < 0{,}01$ значення P і E майже ідентичні.

Визначення величини E базується на порівнянні досліджуваної послідовності довжиною m з множиною послідовностей баз даних. Є два положення, які використовують для визначення Е-числа. Перше говорить про те, що всі послідовності однаково схожі з досліджуваною в базі даних. Це

означає, що значення Е для вирівнювання з короткою послідовністю, що міститься в базі даних, слід прирівняти зі значенням Е для вирівнювання з довгою послідовністю. Для обчислення значення Е по базі даних необхідно помножити значення Е, отримане при попарному вирівнюванні, на число послідовностей в ній. Вивчення схожості між короткими послідовностя ми є більш суттєвим ніж з довгими, оскільки вони часто містять різні ділянки (домени). Це друге положення.

Програма BLAST використовує для розрахунку Е-числа ймовірність подібності, що пропорційна довжині послідовності. Якщо попарне значення Е для послідовності бази даних довжиною n треба помножити на N/n, де N - загальна довжина амінокислот або нуклеотидів в базі даних. Якщо описати більш просто, то E-value – очікуване число збігів з такою вагою (тобто такої якості), якщо б у нас наша послідовність і банк були випадковими.

Алгоритм BLAST

1. Йдемо вздовж Query послідовності і формуємо підслова.
2. За допомогою хеш-таблиці знаходимо в банку відповідні послідовності.
3. Будуємо для них вирівнювання.
4. Оцінюємо статичну значимість вирівнювання (E-value).

E-value - це очікуване число подій, може бути більше одиниці. Якщо E-value маленьке, то, значить, збіг значущий, і воно несе велику біологічну інформацію. P-value - це ймовірність зустрічі такої ж послідовності (не може бути більше одиниці). При оцінці E-value, та і взагалі при будь-яких статистичних оцінках, важливо, яка модель лежить в основі оцінки значимості вирівнювань [3].

В даній роботі порівнюємо бактерії виду *Magnetospirillum gryphiswaldense* та бактерії роду *Gloeobacter.*

Послідовність вирівнювання мікроорганізмів за допомогою пошукової системи BLAST

1. Сторінка сайту NCBI має наступний вигляд:

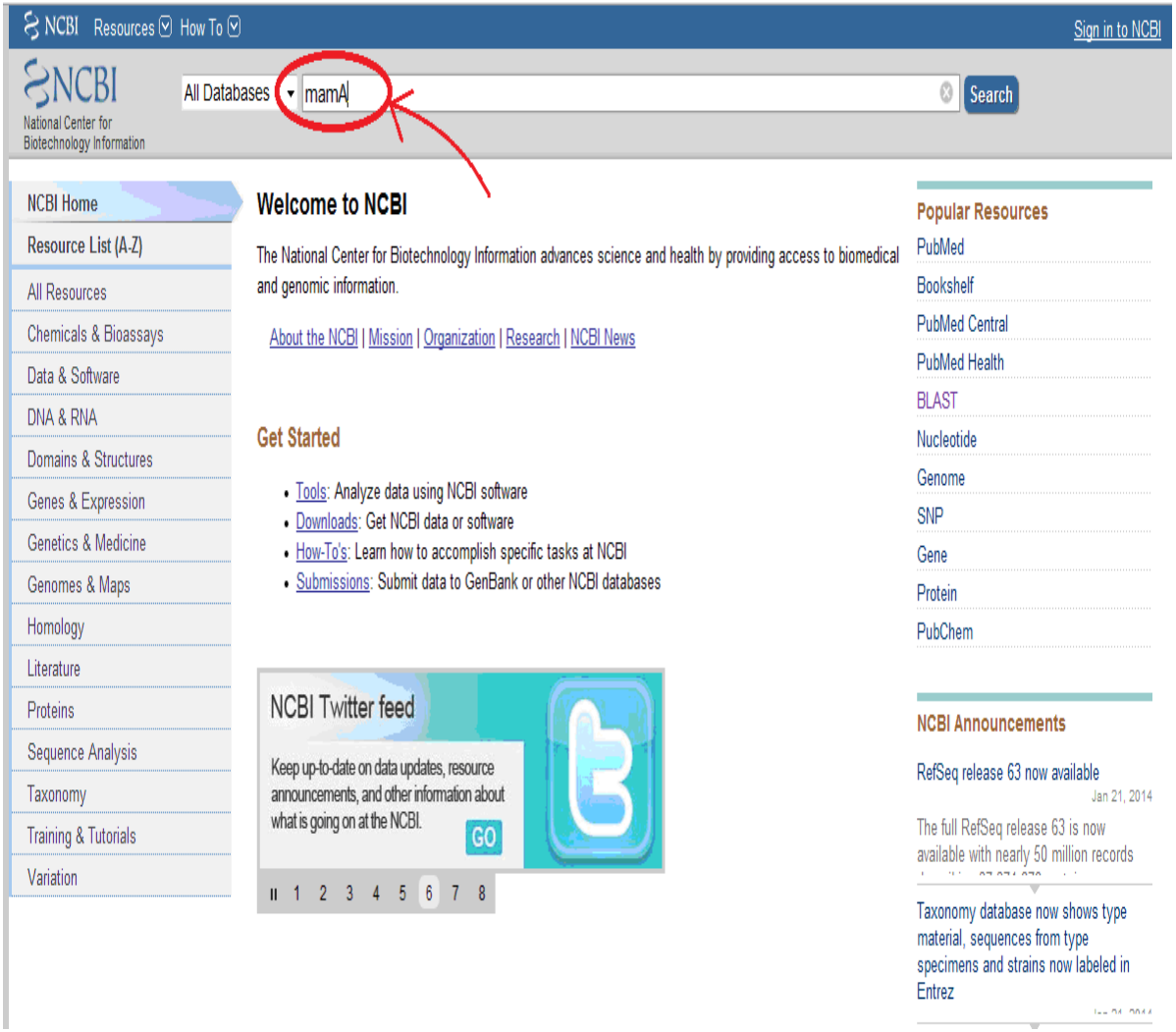


Рисунок 1.1 – Домашня сторінка сайт NCBI

В строчці пошуку вводимо назву білка «mamA» і натискаємо «Search».

2. Інформація, яка виводиться на сторінці, розділена на групи, наприклад, «Genomes», «Genes», «Proteins», «Chemicals». Обираємо розділ «Proteins».

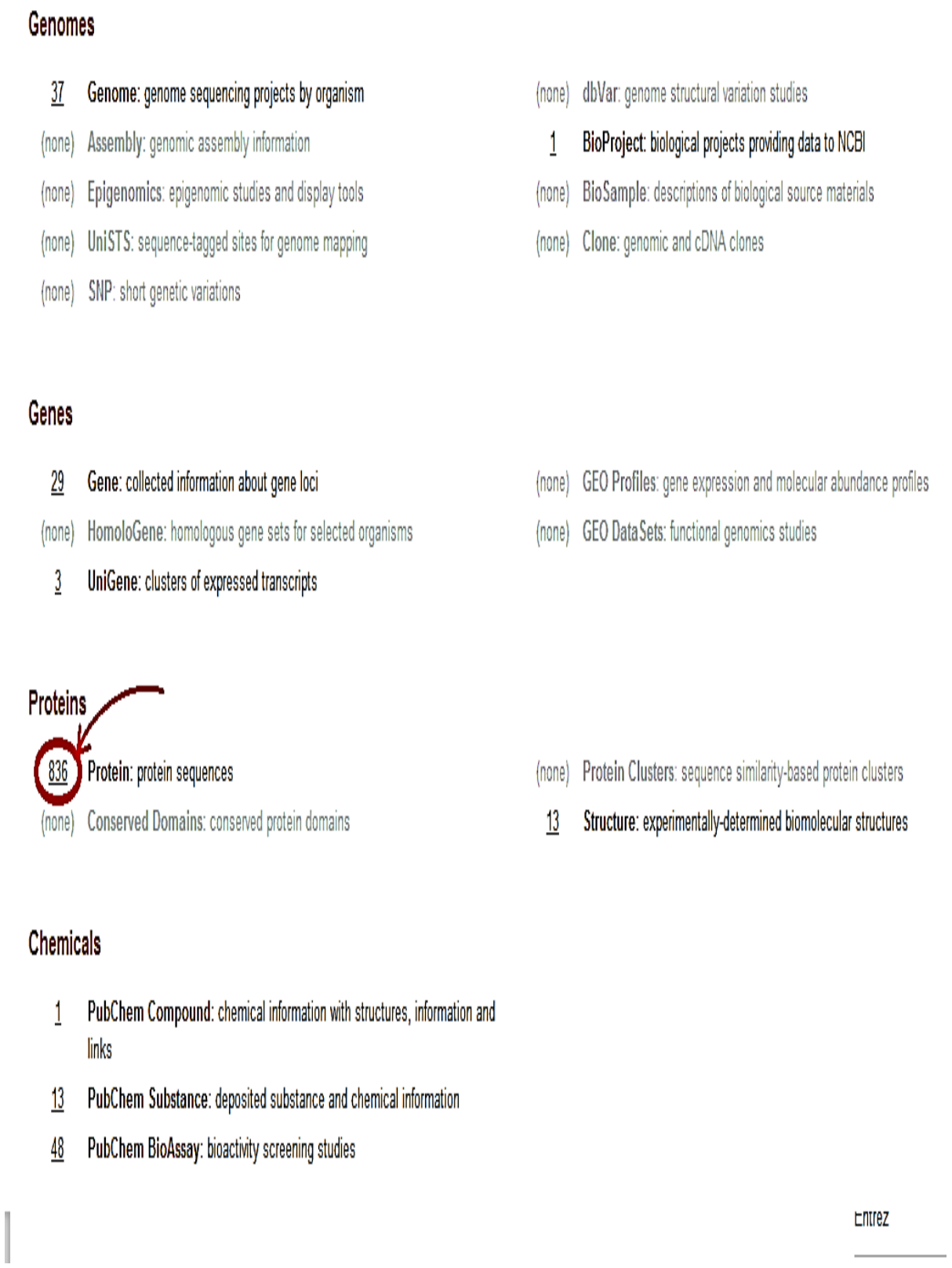


Рисунок 1.2 - Інформація, яка виводиться на сторінці, розділена на групи, наприклад, «Genomes», «Genes», «Proteins», «Chemicals».

3. Отримуємо інформацію про організми, що містять білок mamA. Якщо необхідно знайти бактерії, можна обрати розділ «Bacteria» в лівій частині вікна. Шуканий вид - *Magnetospirillum gryphiswaldense* – знаходиться в списку під номером 5. Переходимо за посиланням, натиснувши на відповідний рядок.

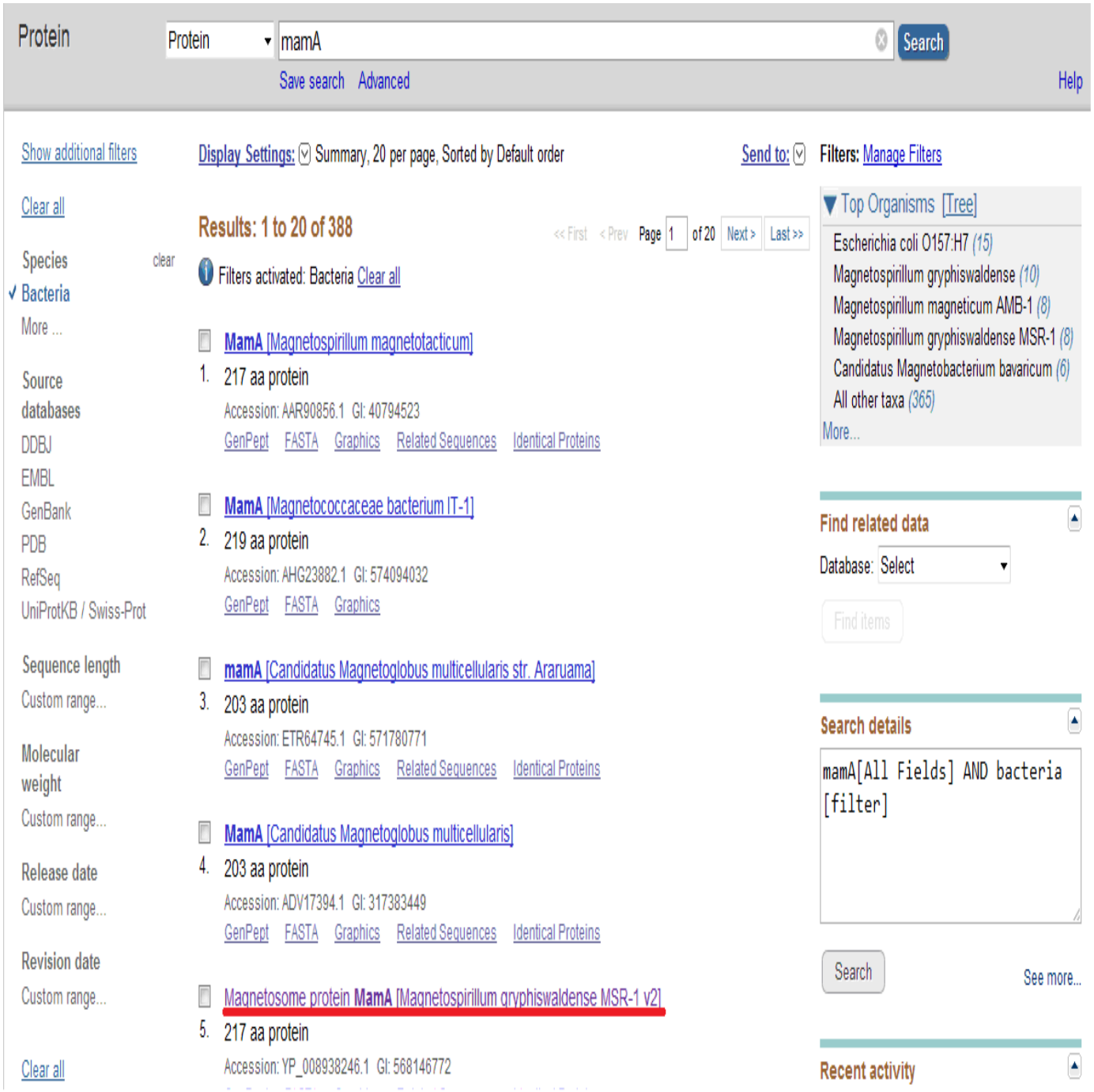


Рисунок 1.3 – Результати пошуку mam A

4. У вікні, що відкрилося, міститься інформація щодо конкретного білка (mamA) даної бактерії *Magnetospirillum gryphiswaldense*.

Для вирівнювання послідовності натискаємо «Run BLAST».

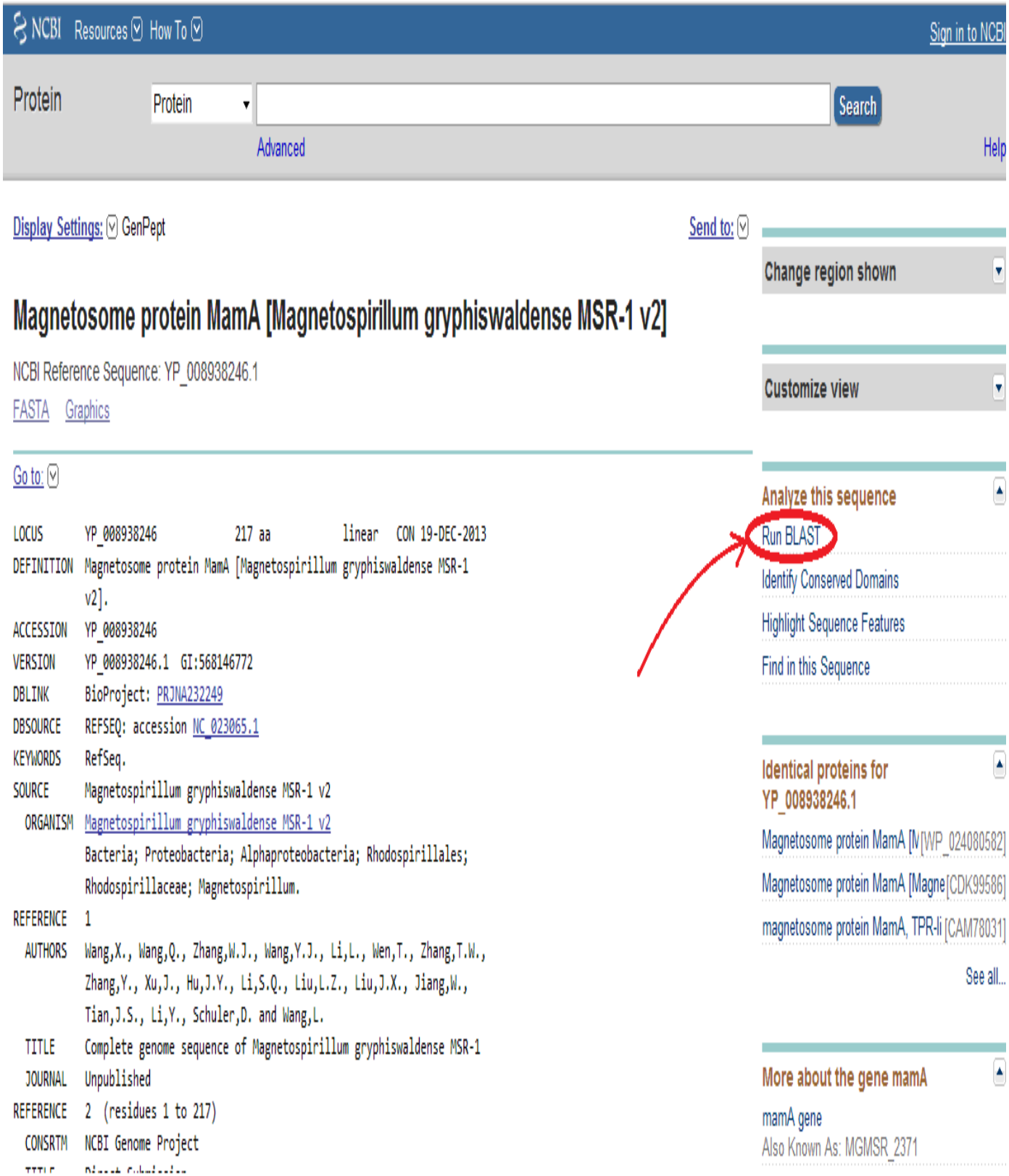


Рисунок 1.4 – Загальний вигляд закладки про білок

5. Навпроти напису «Organism» вводимо назву роду чи виду, з яким необхідно порівняти послідовність білка mamA *Magnetospirillum gryphiswaldense*. В даному випадку, вводимо *Gloeobacter*. Вибираємо алгоритм «blastp (protein-protein BLAST)» та натискаємо «BLAST».

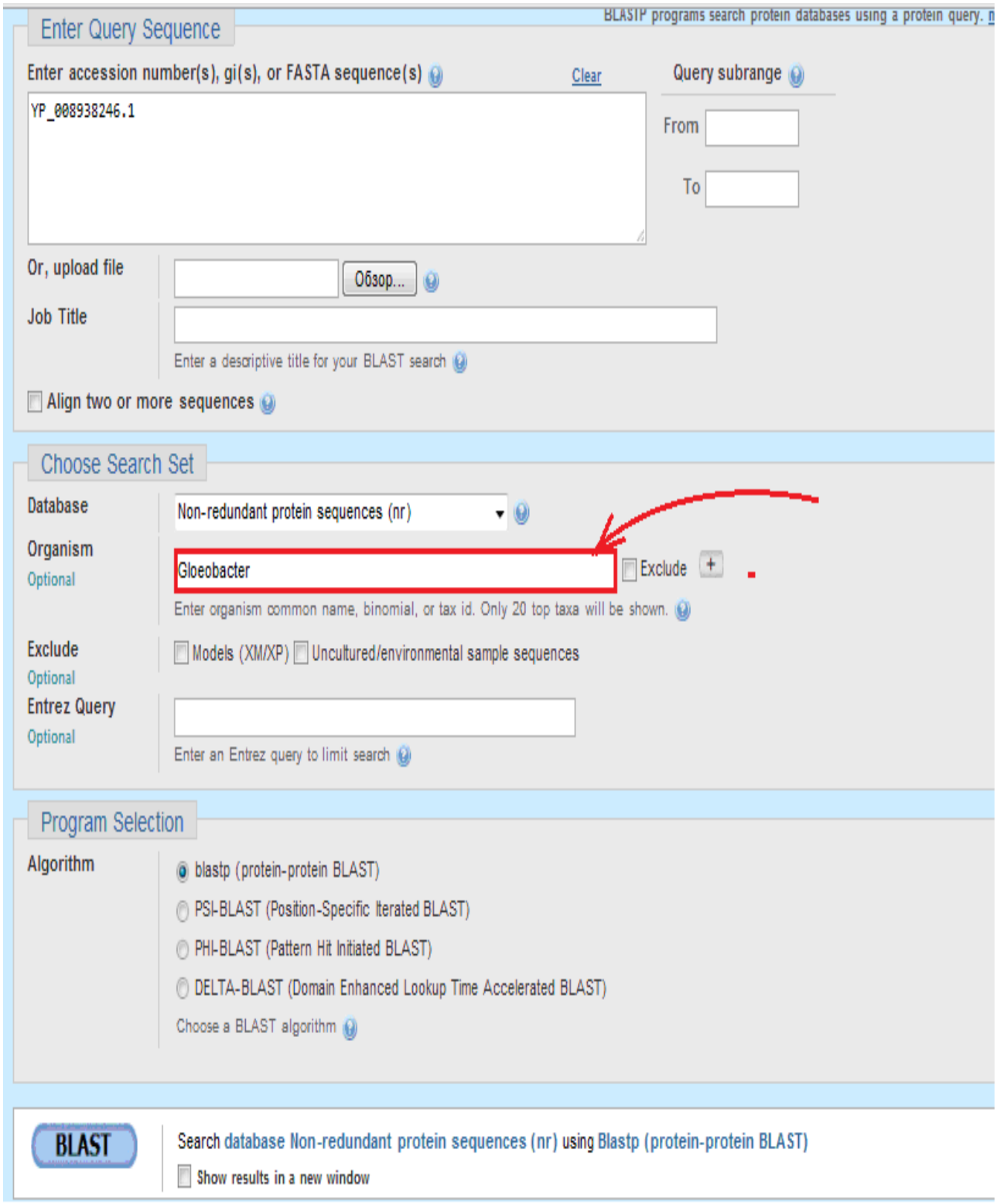


Рисунок 1.5 – Приклад роботи в BLAST

6. Отримуємо результати вирівнювання. З даної таблиці до результуючих таблиць виписуємо Accession, Query cover, E-value, Ident.

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
TPR repeat-containing protein [Gloeobacter kilaueensis JS1] >ref|WP_023171413.1| TPR repeat-containing protein [Gloeobacter kilauee	58.5	250	96%	1e-10	24%	YP_008710120.1
TPR repeat-containing protein [Gloeobacter kilaueensis JS1] >ref|WP_023174309.1| TPR repeat-containing protein [Gloeobacter kilauee	55.5	138	71%	1e-09	31%	YP_008712795.1
hypothetical protein glr1902 [Gloeobacter violaceus PCC 7421] >ref|WP_011141900.1| hypothetical protein [Gloeobacter violaceus] >db	52.4	160	76%	8e-09	23%	NP_924848.1
hypothetical protein gll3576 [Gloeobacter violaceus PCC 7421] >ref|WP_011143565.1| hypothetical protein [Gloeobacter violaceus] >db	51.2	81.6	74%	3e-08	27%	NP_926522.1
TPR repeat-containing protein [Gloeobacter kilaueensis JS1] >ref|WP_023171830.1| TPR repeat-containing protein [Gloeobacter kilauee	51.2	137	73%	3e-08	24%	YP_008710506.1
cellulose synthase subunit BcsC [Gloeobacter kilaueensis JS1] >ref|WP_023173295.1| cellulose synthase subunit BcsC [Gloeobacter	50.4	126	90%	4e-08	28%	YP_008711878.1
TPR repeat-containing protein [Gloeobacter kilaueensis JS1] >ref|WP_023174323.1| TPR repeat-containing protein [Gloeobacter kilauee	49.3	235	74%	1e-07	23%	YP_008712810.1
tetratricopeptide repeat protein [Gloeobacter kilaueensis JS1] >ref|WP_023171773.1| tetratricopeptide repeat protein [Gloeobacter kilau	47.8	181	75%	3e-07	23%	YP_008710452.1
cellulose synthase subunit BcsC [Gloeobacter kilaueensis JS1] >ref|WP_023174657.1| cellulose synthase subunit BcsC [Gloeobacter	46.2	291	91%	1e-06	30%	YP_008713102.1
hypothetical protein glr3240 [Gloeobacter violaceus PCC 7421] >ref|WP_011143230.1| hypothetical protein [Gloeobacter violaceus] >db	45.1	234	78%	3e-06	26%	NP_926186.1

Рисунок 1.6 – Результати роботи програми BLAST, де: **Max score -** високий бал вирівнювання (біт-оцінка) між послідовністю запитів і сегментом послідовності в базі даних.

Total score - сума вирівнювання всіх сегментів з тієї послідовності бази даних, що відповідають запиту (розрахована по всіх сегментах). Total score відрізняється від max score, якщо деякі частини послідовності в базі даних відповідають різним частинам запитуваної послідовності.

Query coverage – відсоток довжини запиту, який включено до вирівняних сегментів. Розраховується по всіх сегментах.

Ident – ступінь відповідності між двома послідовностями (без пробілів між ними). Ідентичність 25% або вище вказує на подібність функції, в той час як ідентичність 18-25% вказує на подібність структури або функції.

Accession – номер доступу, що є унікальним ідентифікатором для запису послідовностей в базі даних GenBank. При пошуку за цим номером можна знайти тільки один файл з послідовністю ДНК.

7. Натискаючи на посилання з найбільшим E-value для даного виду мікроорганізму, отримуємо наступну таблицю.

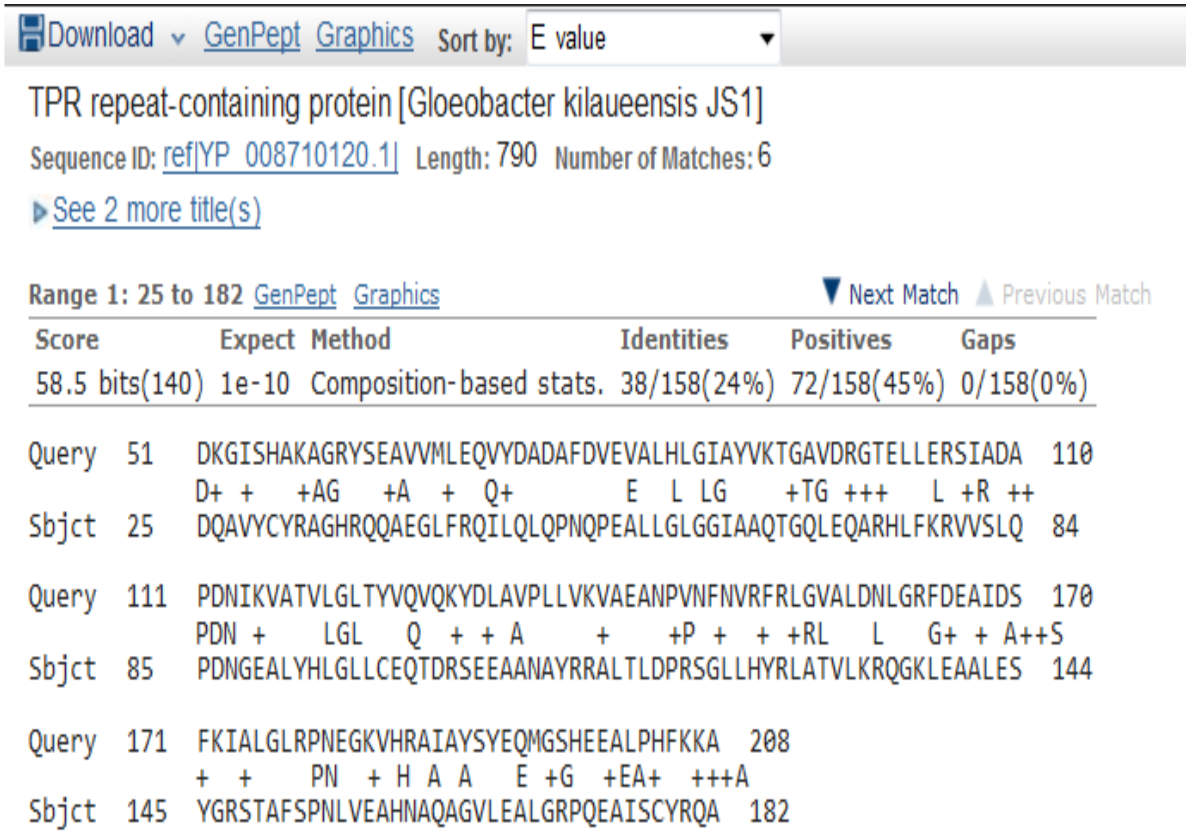


Рисунок 1.7 – Приклад графічного зображення результатів вирівнювання

З отриманої таблиці виписуємо значення Name ID, Positives до результуючої таблиці.

8. Описані вище пункти 1-7 повторюємо для білків mam A, B, E, M, K, O, N, Q, Z.

Послідовність визначення повноти геному

1. На головній сторінці NCBI вводимо назву роду бактерій повноту геному шукаємо і обираємо в лівому віконці «Genome» та натискаємо «Search».

Рисунок 1.8– Приклад пошуку БД Genome

2. Отримуємо інформацію про бактерії роду *Gloeobacter* та обираємо з переліку досліджуваний вид. Шуканий вид - *Gloeobacter violaceus* – знаходиться в списку під номером 2. Переходимо за посиланням, натиснувши на відповідний рядок.

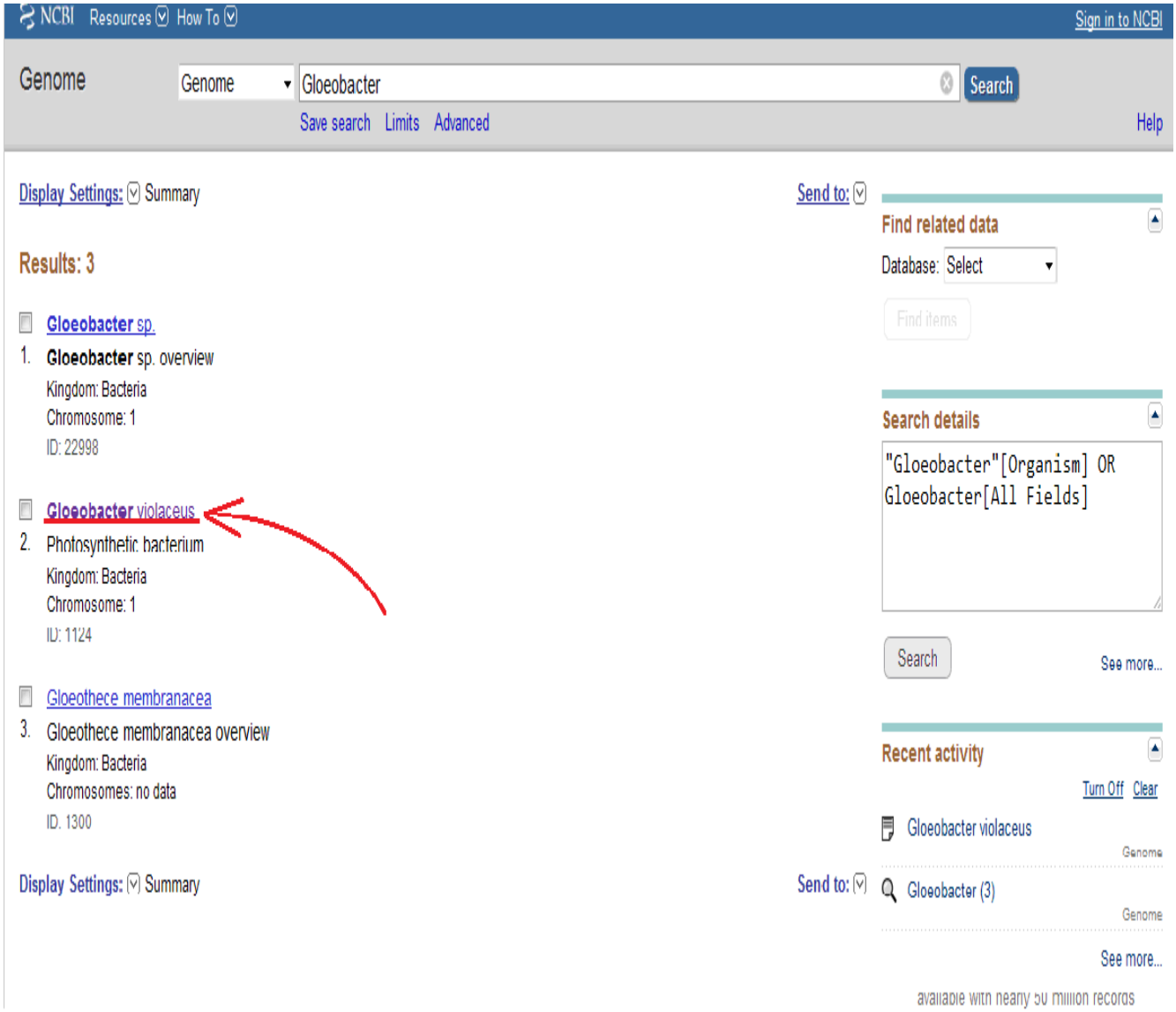


Рисунок 1.9 – Приклад вибору певного організму

3. На сторінці висвітлюється інформація про даний вид бактерій в табличці в столбці «Status» вказується повноту розшифровки геному. Зафарбоване коло означає, що геном повністю розшифрований, напівзафарбоване коло - розшифровано тільки певні ділянки, не зафарбоване коло - нерозшифрований геном.

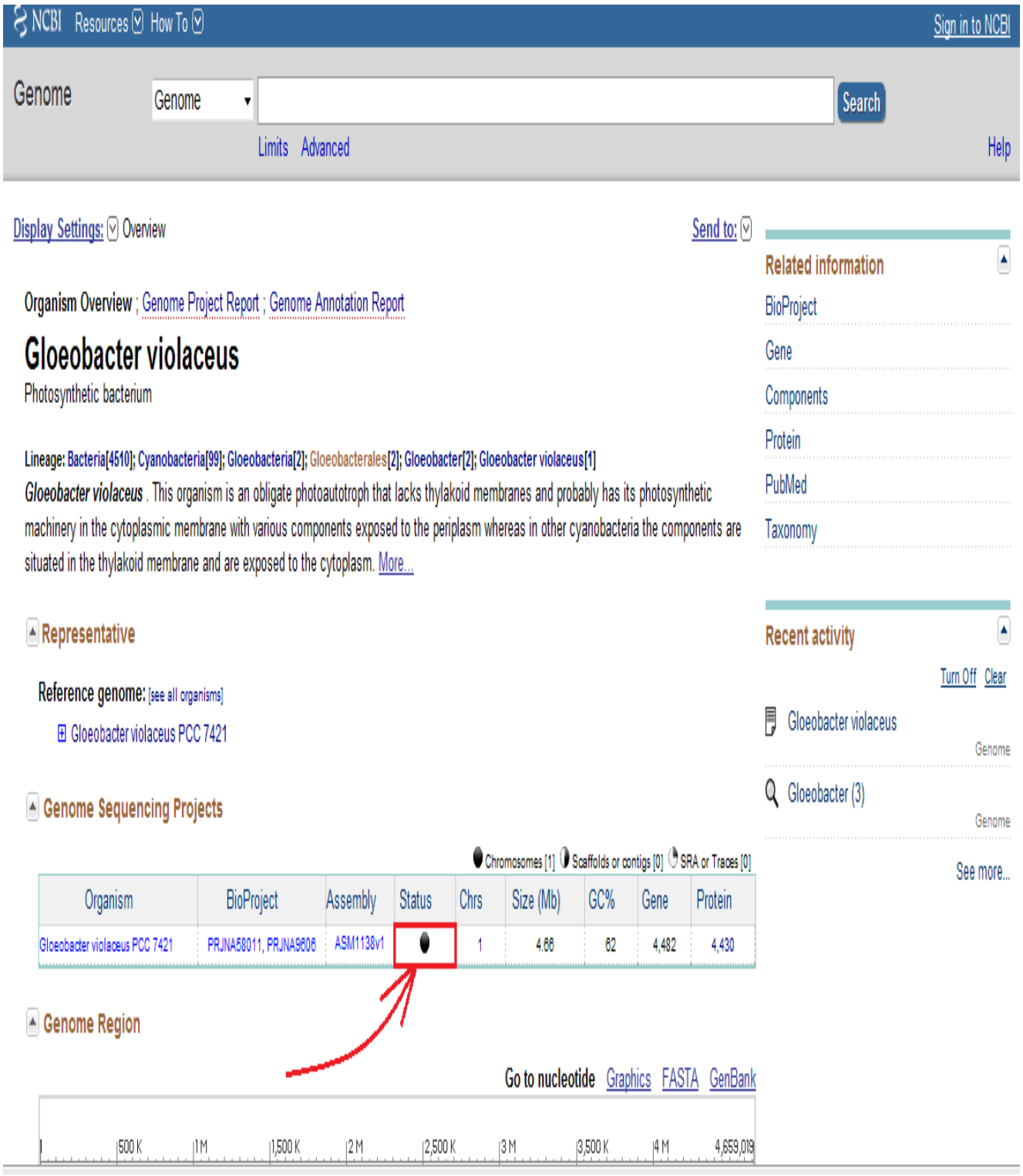


Рисунок 1.10 – Приклад аналізу повноти геному

Завдання для практичної роботи

1. З використанням програми BLAST онлайн-ресурсу NCBI провести попарні вирівнювання амінокислотних послідовностей білків групи Mam та Еукаріот з компонентами протеому представників наведених нижче

таксономічних одиниць. У дослідженні використати амінокислотні послідовності білків Mam, наведених у таблиці 1.1.

Таблиця 1.1. Перелік використаних у дослідженні білків

Білок родини Mam	Код доступу в базі даних NCBI	Білок родини Mam	Код доступу в базі даних NCBI
MamA	AAL09996.1	MamN	CAM78028.1
MamB	AAL09999.1	MamK	CAJ30118.1
MamM	CAM78027.1	MamQ	CAM78032.1
MamE	CAJ30116.1	MamH	CAJ30114.1
MamO	CAJ30122.1	MamZ	YP_008938198.1

2. Перелік таксономічних одиниць, в межах яких необхідно провести пошук організмів, що містять гомологи білків групи Mam та в яких експериментально виявлено біогенні магнітні наночастинки або магніточутливі наноструктури:,,

1) Родина **Comamonadaceae**
2) Родина **Acidithiobacillaceae**
3) Родина **Shewanellaceae**
4) Родина **Chlorobiaceae**
5) Родина **Geobacteraceae**
6) Родина **Rhodocyclaceae**
7) *Родина* **Geobacteraceae**
8) *Родина* **Lactobacillaceae**
9) *Родина* **Lactobacillaceae**
10) *Родина* **Streptococcaceae**
11) *Родина* **Caulobacteraceae**
12) *Родина* **Staphylococcaceae**
13) *Родина* **Pseudomonadaceae**
14) *Родина* **Bacillaceae**
15) *Родина* **Halobacteriaceae**
16) *Родина* **Bradyrhizobiaceae**
17) *Відділ* **Ascomycota** (царство Гриби)

3. За отриманими результатами побудувати таблицю:

Таблиця 1.2 – Результати біоінформатичного аналізу

Назва штаму організ-му*	Назва білка-гомолога та його ID	Е-число	% ідентич-них АК	% співпадінь з урахуванням рівноцінних замін	% перекриття ділянок, де присутні співпадіння	Кількість амінокислот в білку-гомологу
Назва таксономічної одиниці						
Гомологи білка Mam(X)						

*- відмітити наявність послідовностей повного геному для даного штаму організму в базі даних (перевірити за допомогою ресурсу Genome).

4. Для штамів організмів, у протеомі яких виявлено білки-гомологи побудувати наступну таблицю

Таблиця 1.3 – Аналіз організмів, які є теоретичними продуцентами БМН

Назва штаму організму	Середовище існування	Тип живлення	Вимоги до умов культивування

Контрольні питання

1.Принцип роботи Blast.
2. Яка відмінність між ортологами та паралогами?
3. За якими параметрами можна судити про гомологію між білками(генами)?
4. Що таке магнітотаксисні бактерії?
5. Назвіть мам-білки незамінні для біомінералізації БМН

Практичне заняття 2

Оцінка значимості вирівнювань

Мета роботи: оцінити значимість проведених вирівнювань білкових послідовностей та їх статистичні показники.

Теоретичні відомості

Для встановлення біологічного змісту вирівнювання між декількома послідовностями проводять оцінку значимості вирівнювання.

Принцип оцінки залежать від декількох аспектів:

- тип вирівнювання;
- система оцінки якості вирівнювання (системи премій та штрафів, які необхідно використовувати);
- алгоритми, які використовуються для находження оптимальних вирівнювань;
- статистичні методі, які використовуються для оцінки значимості вирівнювання.

Основний підхід до визначення значимості вирівнювань – це розрахунок статистичної значимості ваги вирівнювань, який базується на: моделях Бернуллі або моделях Маркова та їх модифікаціях; матрицях PAM, BLOSUM та ін.

Для знаходження значимості ваги вирівнювань необхідно пройти наступні етапи:

1) Вирівнювання послідовності запиту з рандомізованими (випадковими) послідовностями.

2) Побудова функції розподілу (або розподілу ймовірностей) ваг оптимальних вирівнювань між послідовністю запитом та кожною випадковою послідовністю.

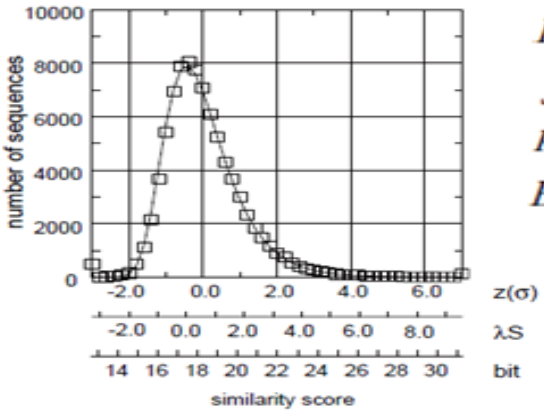
3) Оцінка статистичної значимості вирівнювання.

4) Емпіричні правила оцінки відсотка ідентичних залишків.

5) Аналіз наявності/відсутності спільних функцій.

Вирівнювання послідовності запиту з рандомізованими (випадковими) послідовностями.

Такі вирівнювання виконуються багато разів (на ансамблі 100-500 та більше випадкових послідовностей), знаходяться оптимальні вирівнювання послідовності запиту з кожною рандомізованою послідовністю x_i, дані зводяться у таблицю. Після чого будується розподіл ваг вирівнювань.



$$P(S \geq x) = 1 - exp(-Kmne^{-\lambda x})$$

$$S' = \lambda S - \ln Kmn$$

$$P(S' \geq x) = 1 - exp(-e^{-x})$$

$$E(S' \geq X) = PD$$

Рисунок 2.1 – Графік розподілу ваг вирівнювання

Побудова функції розподілу (або розподілу ймовірностей) ваг оптимальних вирівнювань між послідовністю запитом та кожною випадковою послідовністю.

В результаті вирівнювань послідовності запиту з великою кількістю випадкових послідовностей можна побудувати функцію розподілу ваг оптимальних вирівнювань. На практиці для отримання такого розподілу будують гістограму

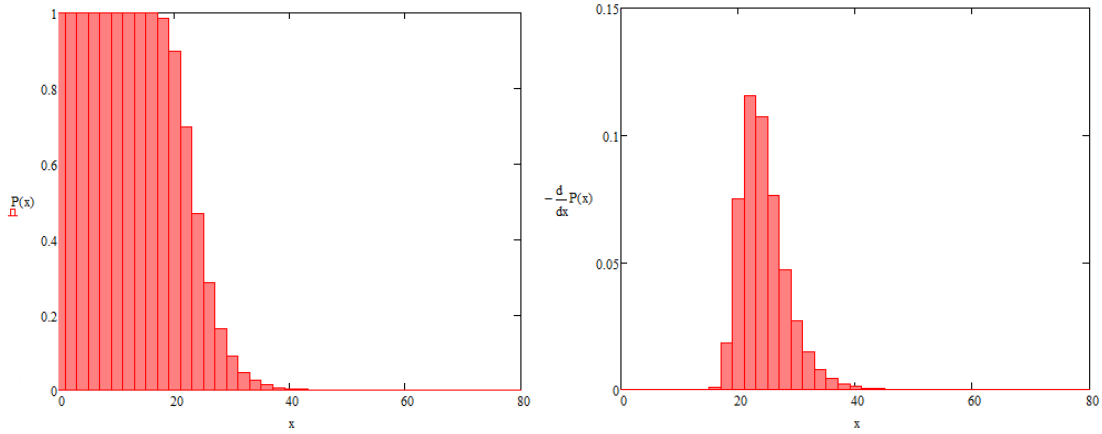


Рисунок 2.2 – Функція розподілу ваг вирівнювань або розподілу ймовірностей ваг представляє собою граничний випадок побудованої гістограми при S, що прямує до нуля.

Права, повільно спадаюча частина графіку означає, що вона не описується нормальним розподілом:

$$P(S' \geq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \langle s \rangle)^2}{2\sigma^2}\right) \qquad (2.1)$$

де σ^2 – дисперсія, x – вага вирівнювання послідовності запиту з i-ою рандомізованою послідовністю, $<s>$– середнє значення ваг оптимальних вирівнювань послідовності запиту з рандомізованими послідовностями.

Тому Стефаном Альтшулем запропоновано зазначену функцію розподілу наблизити **розподілом екстремальних значень (anextreme-**

valuedistribution), якщо вага вирівнювання S' задовольняє умові $S' \geq x_{max}$, де x_{max} визначає положення максимуму густини функції розподілу ваг оптимальних вирівнювань. Це емпірична формула, яка отримана на основі експериментальних даних:

$$P(S' \geq x) = 1 - \exp\left(Kmne^{-\lambda x}\right) \qquad (2.2)$$

де S' – вага вирівнювання, *K і* λ *- параметри, пов'язані з розташуванням максимуму та шириною* функції розподілу ваг оптимальних вирівнювань.

При порівнянні випадкових послідовностей достатньо великих довжин кількість віддалених вирівнювань з вагою щонайменше *x* приблизно описується розподілом Пуасона (Poissondistribution)

Оцінка статистичної значимості вирівнювання

Крім ***P*-чисел** (*P-value*) та ***E*-чисел** (*E-value*), розрахунок яких було представлено вище - формули (2), (3) в біоінформатиці для характеристики статистичної значимості вирівнювання використовуються **Z-числа** (Z-score).

Таким чином, три основні величини використовуються в біоінформатиці для характеристики статистичної значимості вирівнювання:

- **E-число** (E-*value*),
- **P-число** (*P-value*),
- ***Z*-число** (Z-*score*),

Z-число – це міра не випадковості співпадінь при вирівнюванні послідовностей. Розрахунок величини z-числа відбувається за формулою:

$$Z = \frac{x - \langle s \rangle}{\sigma_s} \qquad (2.3)$$

де *x*– вага вирівнювання послідовності запиту з послідовністю з БД;
<s> – середня вага вирівнювань послідовності запиту з випадковими послідовностями, яка дорівнює:

$$\langle s \rangle = \frac{\sum_{i=1}^{n_{\max}} x_i n_i}{n} \qquad (2.4)$$

$$\sigma_s = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \langle s \rangle\right)^2 n_i}{n_{\max}(n_{\max} - 1)}} \qquad (2.5)$$

n– кількість точок на графіку, σ_s– середньоквадратичне відхилення ваг вирівнювань досліджуваної послідовності з випадковими послідовностями, n_i– кількість випадкових послідовностей, що мають вагу x_i, n_{max}– загальна кількість випадкових послідовностей.

***P*-число** – це ймовірність того, що знайдена подібність може бути випадковою, тобто ймовірність того, що досліджуване вирівнювання не краще ніж випадкове.
Орієнтовні *значення P-числа* та їх інтерпретації наступні:
$P \leq 10^{-100}$ – точне співпадіння;
$10^{-100} < P \leq 10^{-50}$ – послідовності майже ідентичні, наприклад, наявні алелі або поліморфізми;
$10^{-50} < P \leq 10^{-10}$ – гомологія очевидна, близькоспоріднені послідовності, близька гомологія;
$10^{-10} < P \leq 10^{-1}$ – скоріш за все дальноспоріднені послідовності, гомологія незначна, дальня гомологія;
$P > 10^{-1}$ – співпадіння не є значущим.

***E*-число** – це очікувана кількість послідовностей в БД, що мають таке саме, або краще значення числа Z, що і досліджуване вирівнювання.
Орієнтовні значення Е-числа та їх інтерпретації наступні:

$E \leq 0.02$ – послідовності ймовірно гомологічні;
$0{,}02 \leq E \leq 1$ – неможливо точно встановити гомологію, гомологія не очевидна;
$E > 1$ – випадкове співпадіння.

Емпіричні правила оцінки відсотка ідентичних залишків

Крім *P*-чисел, і *E*-чисел та z-чисел програми для порівняння послідовностей розраховують відсоток ідентичних залишків.

Для оцінки використовують градацію відсотку ідентичності залишків: 45% - білки мають дуже схожі структури, загальну або схожу функцію; 25%-45% - мають подібний фолдінг; 18%-25% - відносять до області «двозначності» (Р.Ф.Дулітл)**;** 18% - є необхідність додаткових досліджень.

Хід роботи

1. Відкриваємо **NCBI**, обираємо **Protein** , шукаємо потрібний нам білок (MamM, MamB …) й обираємо будь-який із знайдених:

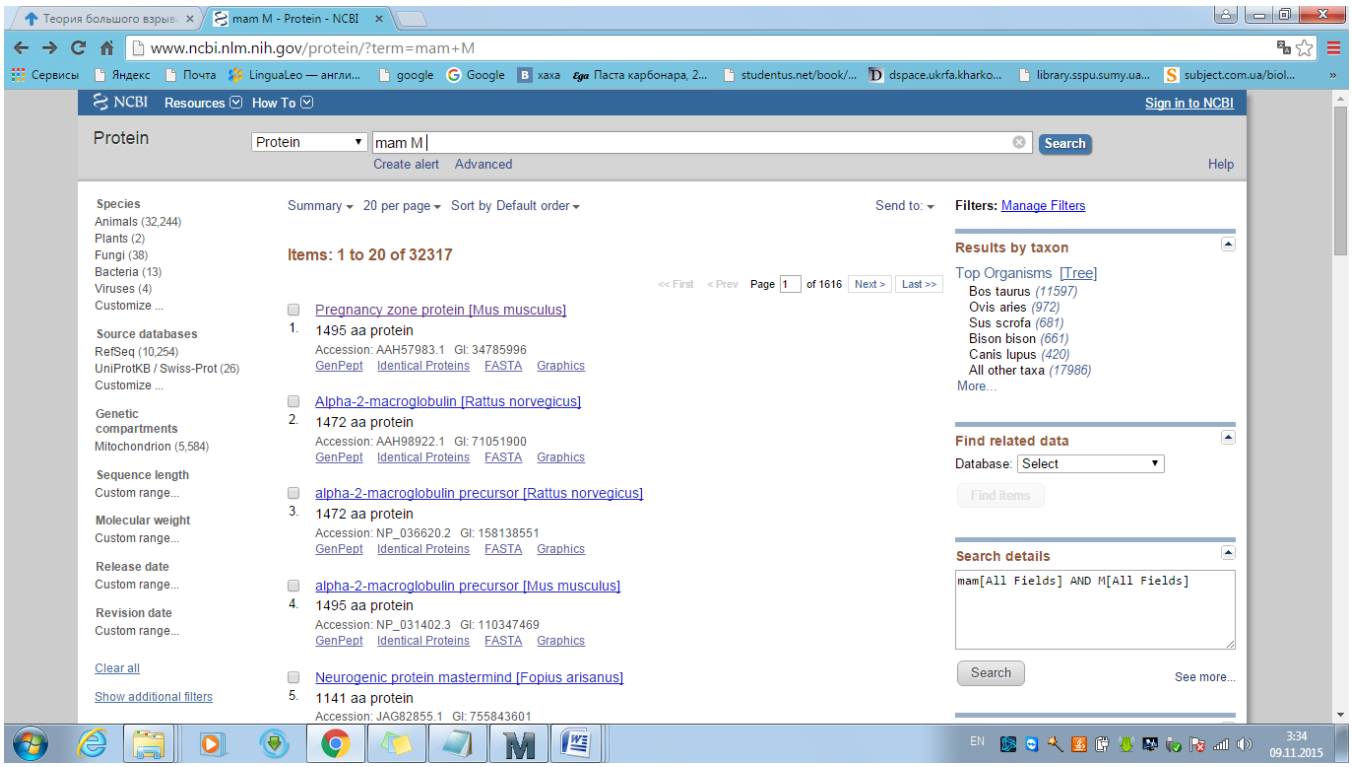


Рисунок 2.1 – Приклад пошуку білку mamB

2. Запускаємо **Run Blast**:

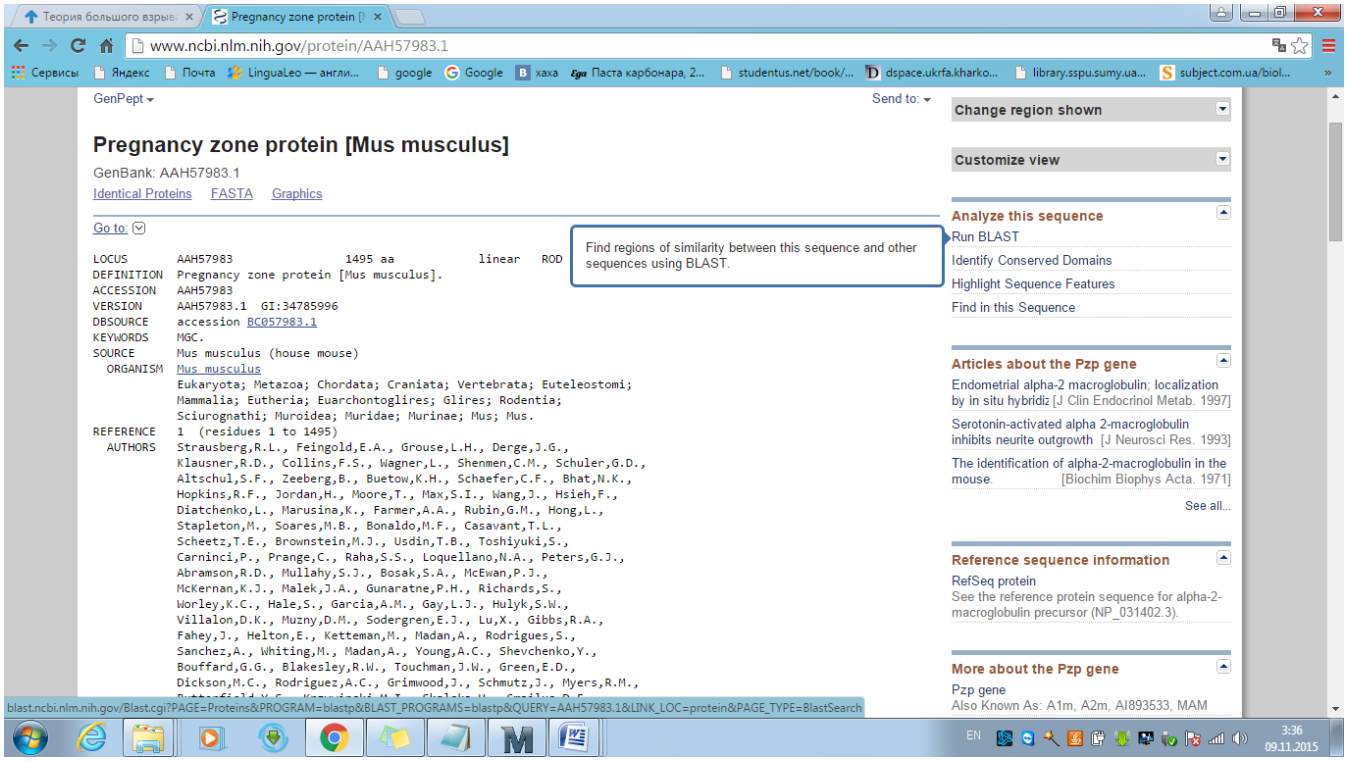


Рисунок 2.2 – Приклад пошуку програми BLAST на сторінці

3. Використовуємо **Blast**

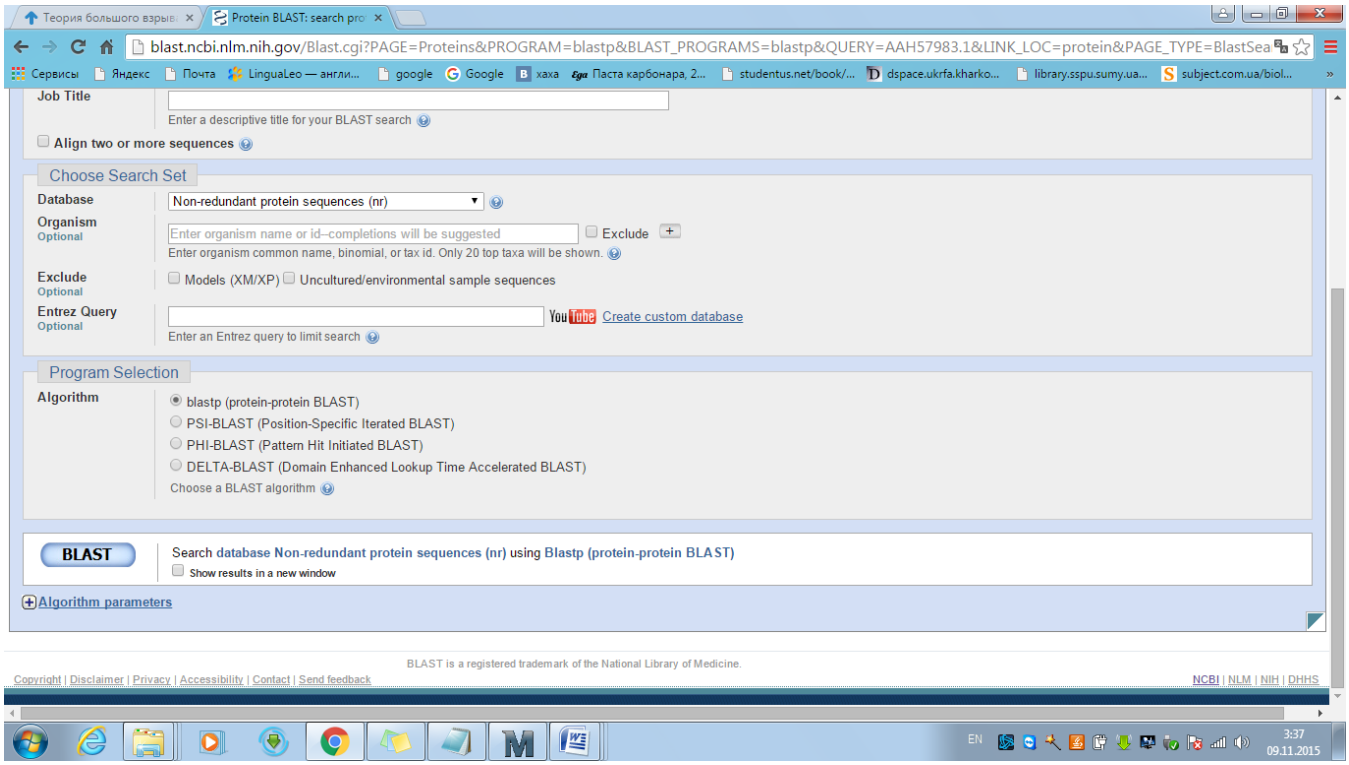


Рисунок 2.3 – Приклад запуску програми BLAST

4. Натискаємо на **Search Summary :**

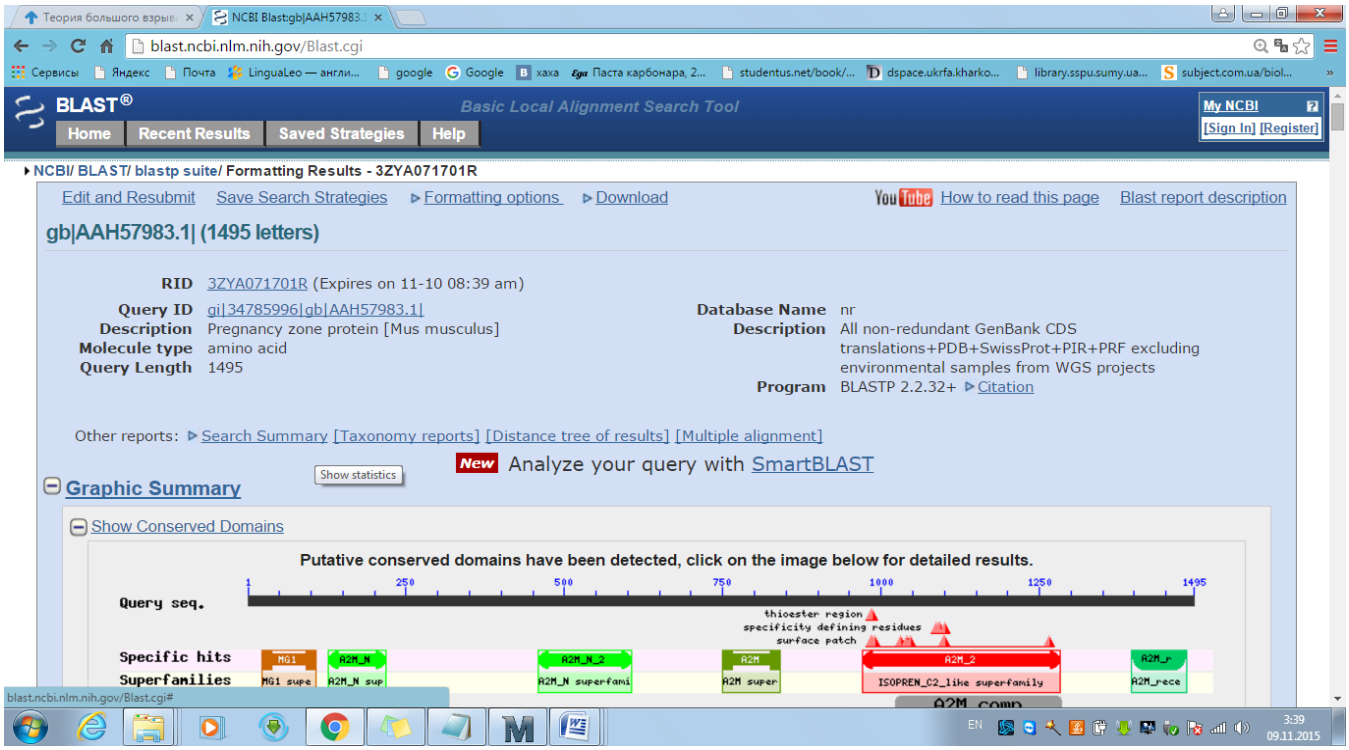
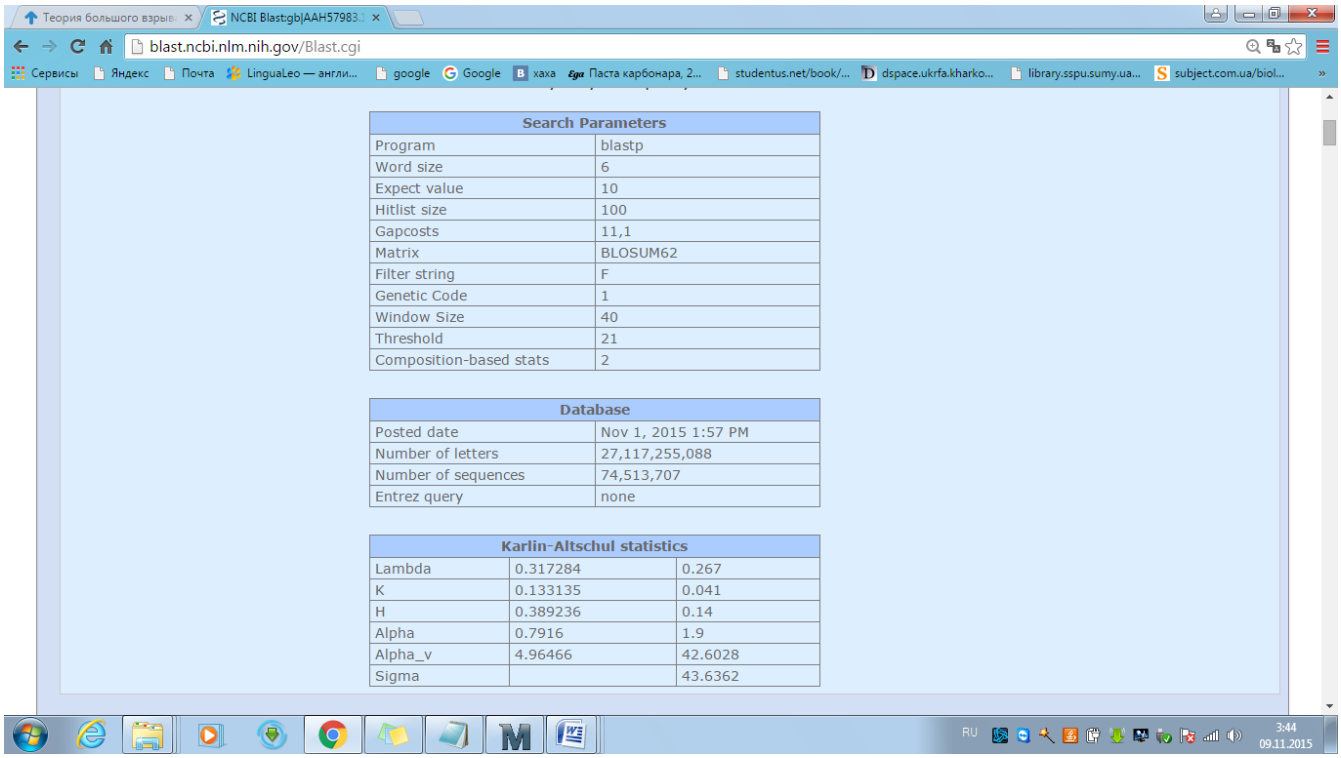


Рисунок 2.4 – Вигляд сторінки сайту з отриманими результатами вирівнювання

5. Отримуємо таблицю, з якої нам необхідно взяти значення **K**й **Lambda**



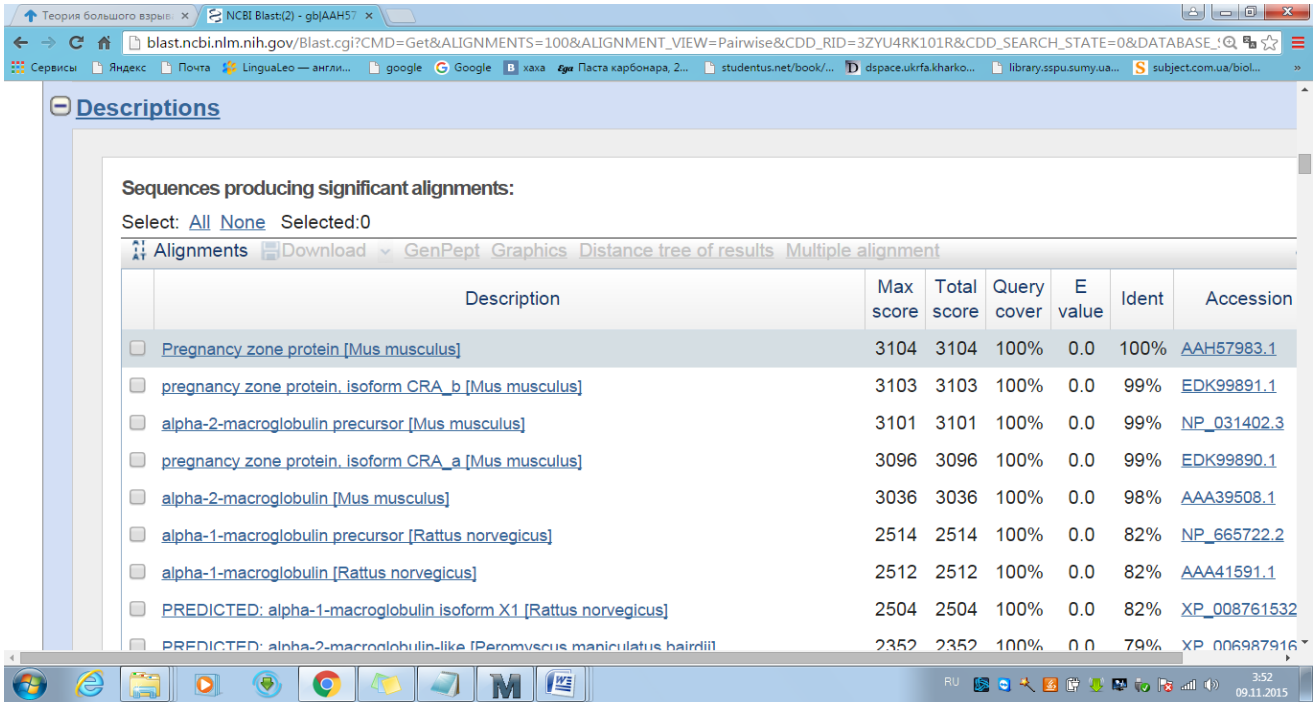
Search Parameters	
Program	blastp
Word size	6
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

Database	
Posted date	Nov 1, 2015 1:57 PM
Number of letters	27,117,255,088
Number of sequences	74,513,707
Entrez query	none

Karlin-Altschul statistics		
Lambda	0.317284	0.267
K	0.133135	0.041
H	0.389236	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Рисунок 2.5 – Параметри пошуку

6. В цьому ж вікні ,але нижче ми знаходимо значення **X** ,яке буде дорівнювати значенню Maxscore, округленому в більший бік.



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Pregnancy zone protein [Mus musculus]	3104	3104	100%	0.0	100%	AAH57983.1
pregnancy zone protein, isoform CRA_b [Mus musculus]	3103	3103	100%	0.0	99%	EDK99891.1
alpha-2-macroglobulin precursor [Mus musculus]	3101	3101	100%	0.0	99%	NP_031402.3
pregnancy zone protein, isoform CRA_a [Mus musculus]	3096	3096	100%	0.0	99%	EDK99890.1
alpha-2-macroglobulin [Mus musculus]	3036	3036	100%	0.0	98%	AAA39508.1
alpha-1-macroglobulin precursor [Rattus norvegicus]	2514	2514	100%	0.0	82%	NP_665722.2
alpha-1-macroglobulin [Rattus norvegicus]	2512	2512	100%	0.0	82%	AAA41591.1
PREDICTED: alpha-1-macroglobulin isoform X1 [Rattus norvegicus]	2504	2504	100%	0.0	82%	XP_008761532
PREDICTED: alpha-2-macroglobulin-like [Peromyscus maniculatus bairdii]	2352	2352	100%	0.0	79%	XP_006987916

Рисунок 2.6 – Результати вирівнювання

7. Відкриваємо файл *11arlin-Altschul statistics* , де й вводимо отримані значення X, K, Lambda. n і m дорівнюватимуть 100.

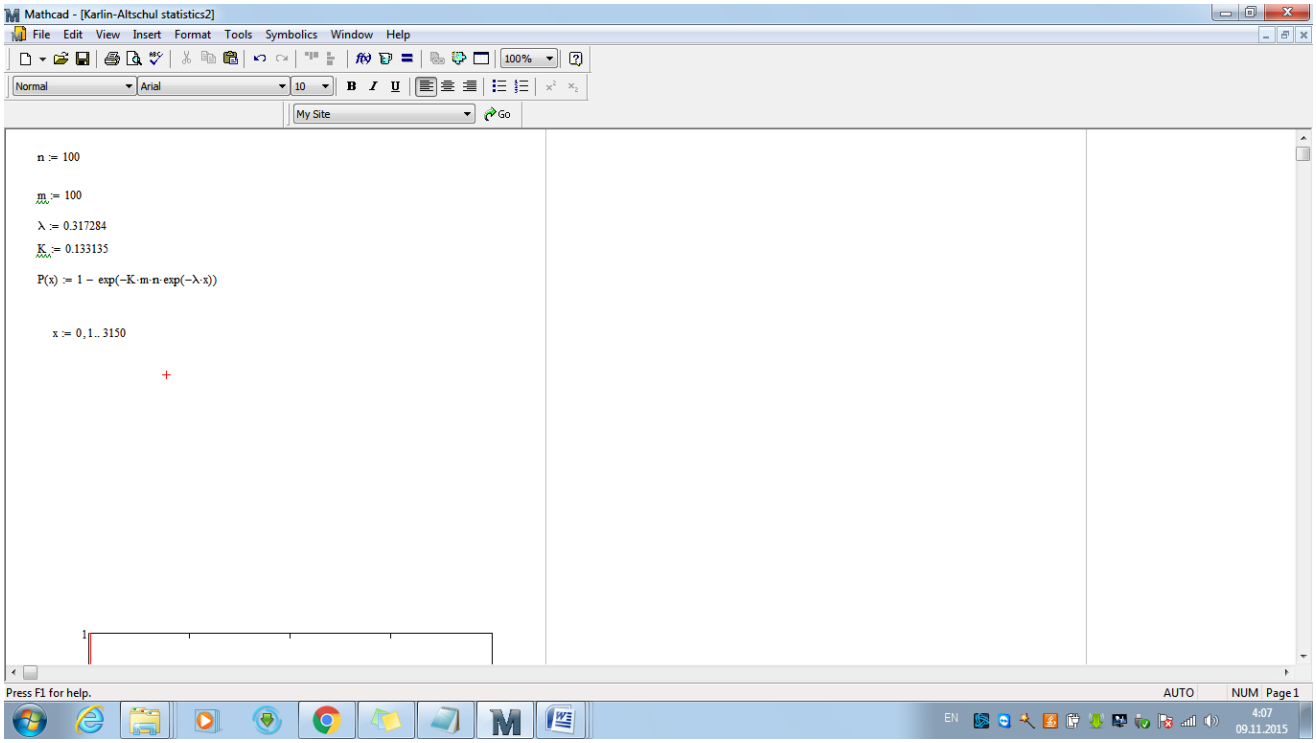


Рисунок 2.7 – Текстове представлення параметрів

Отримуємо 3 графіки які й є результатом цієї роботи

Оформлюємо висновки

Контрольні питання:

1. Які етапи необхідно пройти для знаходження значимості ваги вирівнювань?
2. Які основні аспекти для з'ясування подібності послідовностей?
3. Як будується функцію розподілу ваг оптимальних вирівнювань?
4. Що таке розподіл екстремальних значень?
5. Що таке P-число, як воно розраховується?
6. Що таке Z-число, як воно розраховується?
7. Що таке S-число, як воно розраховується?
8. Емпіричні правила оцінки відсотка ідентичних залишків

Практичне заняття 3

Розрахунок PSSM-матриці

Мета роботи: Освоїти алгоритм розрахунку PSSM-матриці

Теоретичні відомомсті

Позиційно-специфічна матриця ваг Position-Speciic Scoring Matrix (PSSM)

PSSM вперше запропонована в 1987 році [1], отримується з набору послідовностей, що були вирівняні раніше з використанням матриці ваг PAM або BLOSUM.

PSSM розраховується за допомогою програми PSI-BLAST (Position-Speciic Scoring BLAST), яка порівнює профілі PSSM для виявлення пов'язаних гомологічних білків або ДНК.

Параметри позиційно-специфічної матриці ваг Position-Specific Scoring Matrix [2]:

(1)**Позиція (Position):** індекс кожного амінокислотного залишку в послідовності після вирівнювання.

(2) **Набір (Зонд) (Probe):** група типових послідовностей, функціонально споріднених білків вже вирівняних за допомогою алгоритмів вирівнювання послідовностей або за допомогою структурного вирівнювання протеїнів з використанням матриці ваг PAM або BLOSUM.

(3) **Профіль (Profile):** матриця амінокислотних замін, яка складається з 20 стовпчиків, що відповідають 20 амінокислотам.

(4) **Консенсус (Consensus):** послідовність амінокислотних залишків, яка є найбільш близькою до всіх вирівняних послідовностей Набору (Probe) в кожному положенні, консенсусна послідовність генерується, вибираючи найвищий бал в профілі на кожній позиції.

Кожен елемент позиційно-специфічної матриці (PSSM) для даного білку з довжиною N розраховують як:

$$PSSM\,(i,j) = \sum_{k=1}^{20} \omega\,(i,k) \times M\,(j,k),$$
$$i = 1,\ldots,N,\ \ j = 1,\ldots,20, \qquad (3.1)$$

де ω(i,k) є відношення між частотою k-тої амінокислоти в положенні (комірці) і до загальної кількості операцій і M(j,k) це значення мутацій в матриці Дейгхоф між j-тою і k-тою амінокислотами (M(j,k) це матриця замін).

Малі значення PSSM (i, j) вказують на слабо консервативні, а великі значення вказують на сильно консервативні ділянки послідовностей.

Строчка в цьому алгоритмі множиться на строку

PSSM (1,1) = $\sum_{k=1}^{20} \omega(1,k) \cdot M(1,k) = \omega(1,11) \cdot M(1,11) = 3.$

PSSM (2,4) = $\sum_{k=1}^{20} \omega(2,k) \cdot M(4,k) = \frac{2}{3} \cdot 0 + \frac{1}{3} \cdot 7 = 2{,}3$

PSSM (5,6) =
$= \sum_{k=1}^{20} \omega(5,k) \cdot M(6,k) = \frac{1}{3} \cdot 2 + \frac{2}{3} \cdot 8 + \frac{1}{3} \cdot 4 = \frac{2}{3} + \frac{16}{3} + \frac{4}{3} = \frac{22}{3} = 7{,}3$

Ми бачимо, що консервативні стовпчики у множинному вирівнюванні мають більшу вагу і якщо в консервативному стовпчику є мутація, то вона буде мати дуже малу ймовірність, тобто, наприклад, якщо це сайти зв'язування, то мутації в ньому мало ймовірні.

Хід роботи

Побудувати матриці Дейхофф S і M

1. Вирівняні за допомогою глобального вирівнювання послідовності

S

t_1

2. Розрахувати P(a) =

A/K	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P(a)																				

3. f_{ab}

A/K	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A																				
C																				
D																				
E																				
F																				
G																				
H																				
I																				
K																				
L																				
M																				
N																				
P																				
Q																				
R																				
S																				
T																				
V																				
W																				
Y																				

f_a =

f =

m_a =

M_{aa} =

4. M_{ab} =

5. S = /K	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A																				
C																				
D																				
E																				
F																				
G																				
H																				
I																				
K																				
L																				
M																				
N																				
P																				
Q																				
R																				
S																				
T																				
V																				
W																				
Y																				

A/K	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A																				
C																				
D																				
E																				
F																				
G																				
H																				
I																				
K																				
L																				
M																				
N																				
P																				
Q																				
R																				
S																				
T																				
V																				
W																				
Y																				

Варіанти практичного заняття:

1)

S	M	S	S	K	P	S	N	M	L	E	V	W
t	M	M	S	K	R	M	N	S	L	W		

2)

S	F	A	R	R	C	V	W	S	M	K	L	Y
t	F	V	K	R	C	A	M	M	M	Y		

3)

S	K	K	A	S	M	V	V	C	H	I	L	P
t	L	K	A	M	S	A	C	H	L	I		

4)

S	G	G	A	T	D	L	E	H	H	I	F	T
t	G	C	V	T	D	I	E	G	H	L		

5)

S	M	W	I	D	L	L	A	R	E	R	S	D
t	S	W	L	D	I	L	V	R	R	S	D	

6)

S	M	R	K	S	G	C	A	V	C	S	R	S
t	S	R	A	M	G	C	V	C	S	R		

7)

S	M	E	I	G	E	T	M	G	D	Q	P	T
t	M	I	L	G	E	A	S	G	D	P		

8)

S	N	K	I	V	F	C	E	R	S	W	K	A
t	N	R	L	A	F	C	E	R	M	A		

9)

S	P	V	S	I	L	A	F	L	I	L	V	T
t	P	A	M	L	I	A	F	I	L	V		

10)

S	G	A	Y	L	L	D	N	Y	D	E	S	M
t	G	V	Y	I	I	D	N	D	E	M		

11)

S	I	A	A	T	G	P	T	G	K	V	E	G
t	L	A	V	T	G	Q	T	G	K	A		

12)

S	F	G	Y	G	M	Y	L	I	R	P	V	V
t	A	G	Y	G	S	Y	I	L	R	V		

13)

S	G	A	R	L	M	R	V	L	P	V	R	V
t	G	V	R	I	S	R	V	I	P	A		

14)

S	M	A	A	I	A	I	K	S	L	T	T	V
t	M	V	A	L	A	L	K	M	I	T		

15)

S	R	S	A	V	L	G	L	P	M	G	L	F
t	R	M	V	A	L	G	I	P	S	G		

16)

S	S	W	E	Q	T	G	G	P	L	V	A	S
t	M	W	E	Q	T	G	C	P	I	A	V	

Контрольні запитання:

1. Що таке програма BLAST?
2. На які групи поділяється сімейство програм BLAST?
3. Як розраховується PSSM-матриця?
4. Практичне значення PSSM-матриці.
5. Параметри PSSM-матриці.

Практичне заняття 4

Пошук дальніх гомологів у білків, що відповідають за біомінералізацію біогенних магнітних наночастинок

Мета роботи*:* освоїти метод пошуку гомологів у різноманітних білків використовуючи програму PSI-BLAST.

Теоретичні відомості

Однією проблемою при виділенні сімейств є складна доменна структура багатьох білків. Структурні домени білків найкраще виявляються при аналізі їх просторової організації. Наявність експериментальних даних по тривимірних структурах дозволяє визначити число доменів і межі між ними в первинній структурі білка. Різні структурні домени, як правило, виконують різні біологічні функції, будучи тим самим і функціональними доменами. Відсутність інформації про просторову структуру білка істотно ускладнює визначення його доменної структури. Часто різні домени одного білка мають незалежну еволюційну історію.

В таких випадках вони є одночасно і еволюційними доменами. Однак у багатьох випадках два структурні домени майже завжди присутні в білках одночасно, створюючи однин еволюційний домен. Наприклад, такими парними структурними доменами володіють глікозил-гідролази сімейств GH27 і GH32.

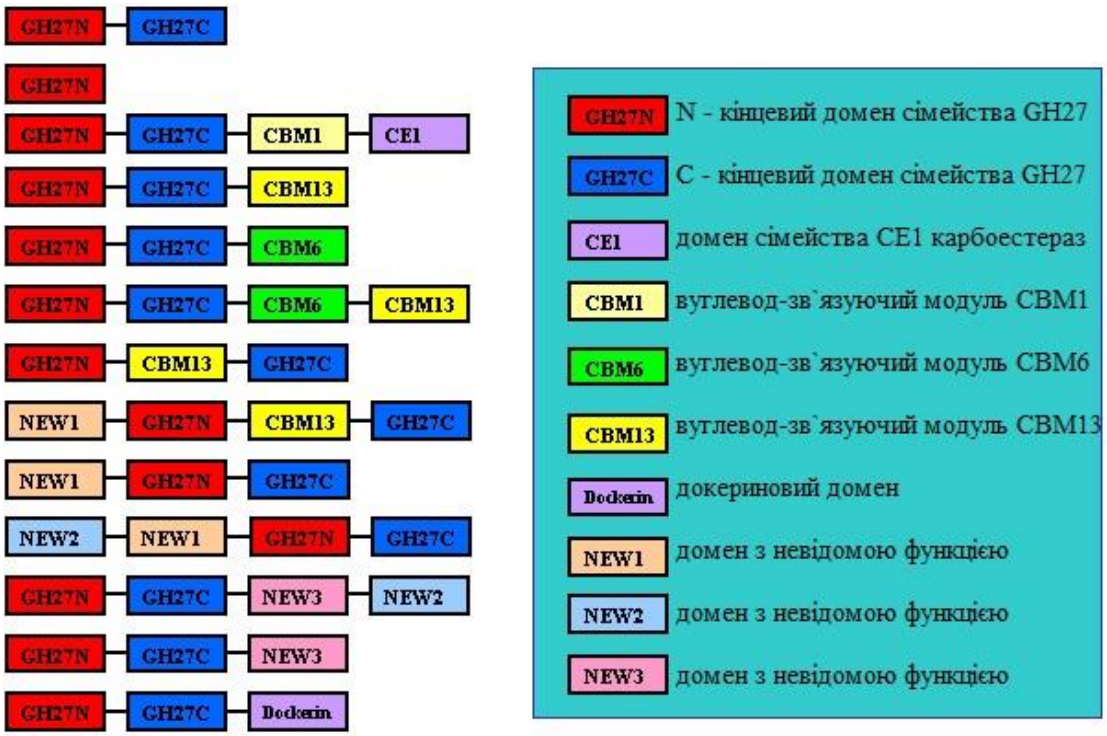


Рисунок 4.1 – Доменна структура білків сімейства GH27 глікозил-гідролаз

Більшість білків цього сімейства складаються з двох доменів: GH27N і GH27C. Лише кілька білків містять тільки каталітичний домен GH27N. Ряд білків також мають додаткові домени декількох типів.

Часто виявляється, що у складі якогось сімейства немає жодного детально дослідженого білка. У такій ситуації певні висновки про структуру та функції білків цього сімейства можна зробити виходячи з інформації про білки з еволюційно споріднених сімейств.

Наприклад, наявність експериментальних даних по третинній структурі якогось білка дозволяє передбачити просторову будову не тільки інших білків того ж сімейства, але і для представників споріднених сімейств.

Мета полягає в застосуванні програми BLAST для того, щоб **підібрати з БД кандидатів для порівняння** послідовностей.

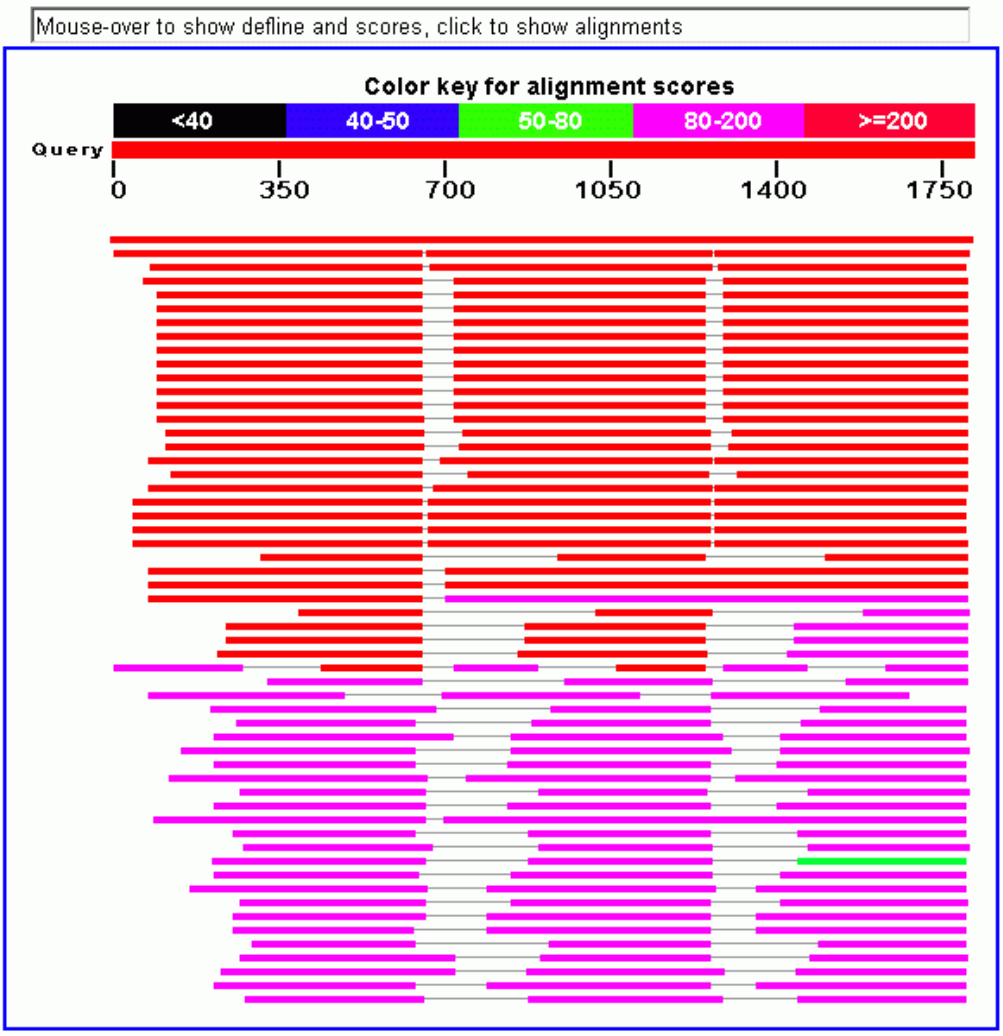


Рисунок 4.2 – Схема, що показує результат пошуку гомологів за допомогою програми PSI-BLAST. В якості запиту був обраний білок, що складається з трьох гомологічних між собою доменів.

Сімейство програм серії BLAST ділиться на наступні групи:

Для пошуку еволюційно споріднених сімейств білків доцільно використовувати програму PSI-BLAST. У результаті своєї першої ітерації вона зазвичай знаходить майже виключно білки даного сімейства, а подальші ітерації виявляють представників споріднених сімейств. В якості порогового значення E-value для включення послідовності в наступну ітерацію має сенс використовувати 0.01 або 0.001. Ітерації варто проводити до припинення появи нових білків із заданим рівнем схожості. Білки, знайдені в кожній з ітерацій, треба досліджувати на приналежність до відомих чи нових сімейств. При цьому слід враховувати той факт, що білки можуть містити більше одного домену, а також можливість появи серед результатів скринінгу бази даних амінокислотних послідовностей і негомологічних білків. Слід очікувати того, що спорідненість двох сімейств білків повинна бути взаємною, тобто якщо використання послідовностей білків одного сімейства дозволяє знайти серед гомологів членів другого сімейства, то і використання представників другого сімейства повинно виявляти білки першого відповідно.

Параметри позиційно-специфічної матриці ваг Position-Specific Scoring Matrix [1]:

1. Позиція (Position): індекс кожного амінокислотного залишку в послідовності після вирівнювання.
2. Набір (Зонд) (Probe): група типових послідовностей, функціонально споріднених білків вже вирівняних за допомогою алгоритмів вирівнювання послідовностей або за допомогою структурного вирівнювання протеїнів з використанням матриці ваг PAM або BLOSUM.
3. Профіль (Profile): матриця амінокислотних замін, яка складається з 20 стовпчиків, що відповідають 20 амінокислотам.
4. Консенсус (Consensus): послідовність амінокислотних залишків, яка є найбільш близькою до всіх вирівняних послідовностей Набору (Probe) в кожному положенні, консенсусна послідовність генерується, вибираючи найвищий бал в профілі на кожній позиції.

Кожен елемент позиційно-специфічна матриці (PSSM) для даного білку з довжиною N розраховується як:

$$PSSM(i,j) = \sum_{k=1}^{20} \omega\,(i,k) \times M(j,k),$$
$$i = 1,\ldots,N,\ \ j = 1,\ldots,20, \qquad (4.1)$$

де ω(i,k) є відношення між частотою k-тої амінокислоти в положенні (комірці) і до загальної кількості операцій і *M*(j,k) це значення мутацій в матриці Дейгхоф між j-тою і k-тою амінокислотами (*M*(j,k) це матриця замін).

Малі значення PSSM (i, j) вказують на слабо консервативні, а великі значення вказують на сильно консервативні ділянки послідовностей.

Хід роботи

1. Запустити програму Blast.

В верхньому лівому кутку вибрати Protein, ввести в поле пошуку необхідний білок (MamA, MamB, MamM, MamE, MamO, MamK).

2. Пошук необхідного білку по якому буде проводитися дослідження.

Обравши шуканий білок у мікроорганізма *Magnetospirillum gryphiswaldense MSR-1* і натиснувши справа **Run Blast**, вирівняти його з необхідним мікроорганізмом.

3. Запуск програми PSI-BLAST

Після вибору мікроорганізму, вибираємо програму за алгоритмом PSI-BLAST (Position-Specific Iterated BLAST), та натискаємо BLAST.

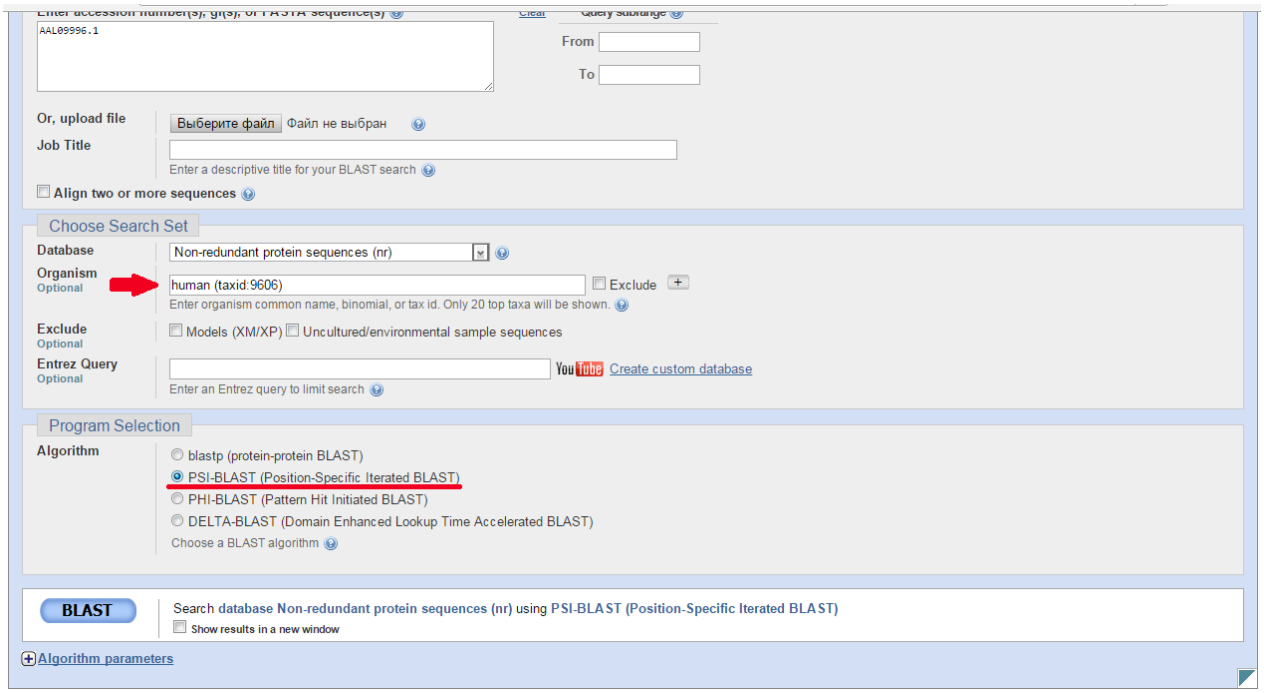


Рисунок 4.3 – Вибір програми Psi-BLAST

4. Проведення декількох етапів ітерацій

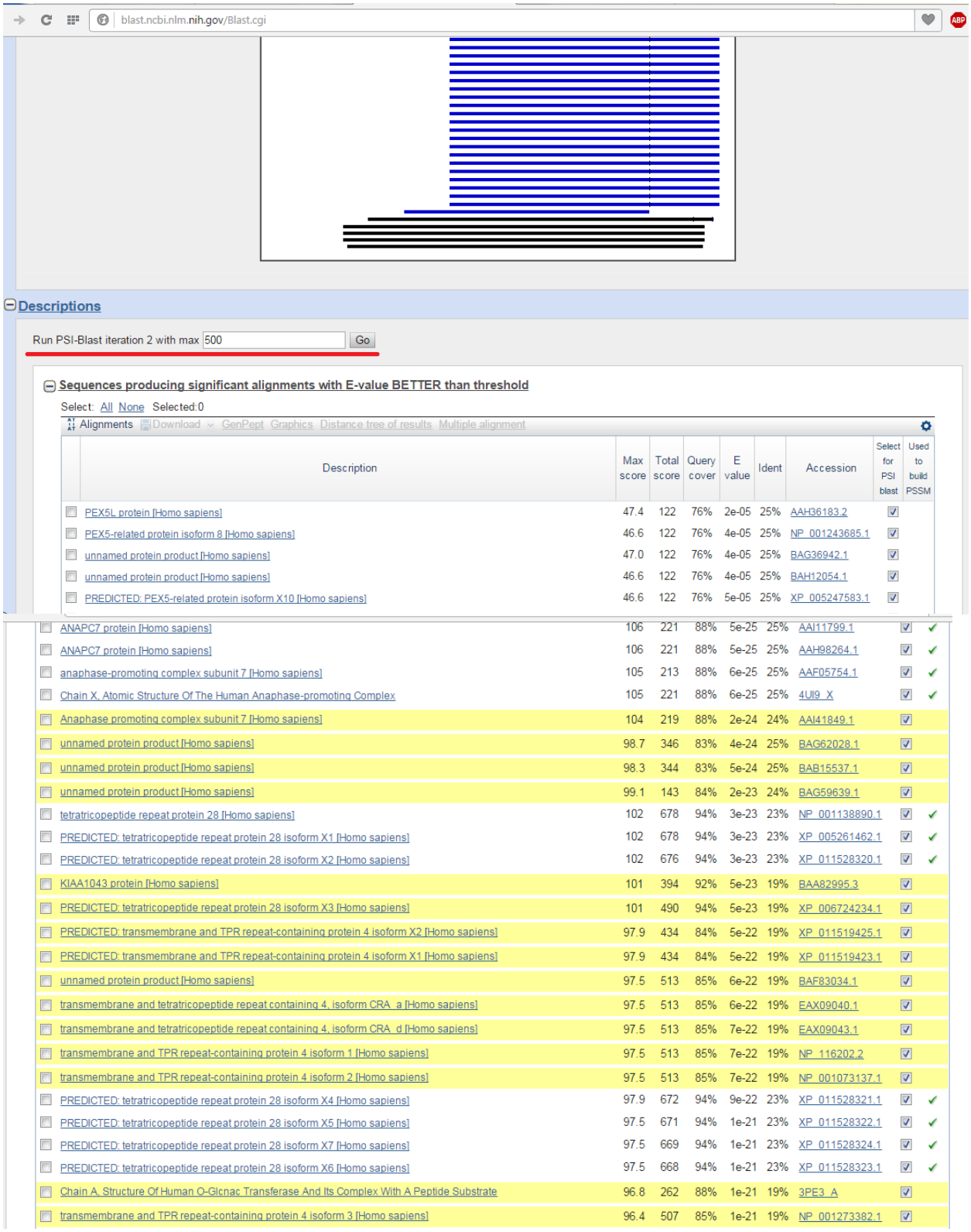


Рисунок 4.4 – Етапи ітерації

Потім знову натискаємо Run PSI-Blast iteration 3 with max - 500 (так проводимо декілька ітерацій підряд).

5. Аналіз отриманих білків гомологів на спільну будову і функції
6. Оформлення результатів у вигляді таблиці, написання висновків.

Таблиця 4.1 Результати роботи

№	Назва гомологічного білка (переклад)	E-value / I (%)	Опис білку(функції)
1			
2			
3			
…			
N			

Контрольні запитання:

1. Що таке амінокислотні заміни, які вони бувають?
2. Що таке білки-гомологи?
3. Який вплив має доменна структура білків на методи порівняння білкових послідовностей?
4. Що таке монофілетична група?
5. Як проводиться пошук дальніх білків-гомологів?

Практична робота 5

Пошук серед мікроорганізмів, що викликають захворювання серця і мозку потенційних продуцентів БМН

Мета роботи*:* визначити мікроорганізми, які можуть мати білки-гомологи білків магнітосомного острівця магнітотаксисних бактерій (МТБ) MamA, MamB, MamM, MamE, MamE і т.д., а також знайти серед них продуцентів біогенних магнітних наночастинок (БМН).

Теоретичні відомості

Захворювання таких надзвичайно важливих органів як серце і мозок завжди будують актуальними серед суспільства. Це ті захворювання які іноді дуже важко точно діагностувати, і передбачити їх протікання. Великий відсоток таких захворювань викликається бактеріальними організмами і тому має сенс їх дослідження.

В тканинах мозку вже давно було знайдено біогенний магнетит, який демонструє його унікальні властивості та функції. Питання про його функції в цих органах на сьогоднішній день залишається відкритим.

Вченими було доведено, що кількість магнетиту в мозку у пацієнтів з захворюванням Альцгеймера набагато більша ніж без нього, аналогічна ситуація і з іншими хворобами.

Якщо серед мікроорганізмів, що викликають захворювання серця і мозку є продуценти біогенних магнітних частинок, це буде величезним проривом сучасної медицини у боротьбі з страшними хворобами і дасть змогу пояснити і передбачити механізми їх протікання.

Серед мікроорганізмів, що викликають захворювання серця можна виділити наступні: *Borrelia burgdorferi, Corynebacterium diphtheria* (викликають інфекційний міокардит); *Staphylococcus aureus* (викликають бактеріальний ендокардит).

Менінгіт – це запалення м'якої мозкової оболонки, що покриває головний мозок людини і спинний мозок. Запалення можуть спричинити віруси, бактерії або інші мікроорганізми, й, у більш рідкісних випадках, деякі лікарські препарати. Головні збудники бактеріального менінгіту - це *Streptococcus agalactiae, Neisseria meningitidis і Streptococcus pneumoniae***.** У всьому світі вони викликають 75-80% випадків цього захворювання, хоча співвідношення між цими трьома збудниками в різних країнах різна.

Під час пошуку гомологів найбільш важливими параметрами є числа Ident та E-value.

E-число – це очікувана кількість послідовностей в базі даних, що мають таке саме, або краще значення числа **z** (міра не випадковості співпадінь при вирівнюванні послідовностей).

Таблиця 5.1. Орієнтовані значення Е-числа

$E \leq 0{,}02$	Послідовності ймовірно гомологічні
$0{,}02 \leq E \leq 1$	Неможливо точно встановити гомологію
$E > 1$	Випадкове співпадіння

Тобто, як видно з таблиці 1, треба обирати ті гомологи, значення Е-числа у яких $E \leq 0{,}02$.

Окрім числа Е, необхідно також врахувати параметр **Ident,** по якому можно судити про структуру та функції вирівнюваних послідовностей. Наприклад, якщо два білки мають Ident більше ніж 45%, то вони мають схожу структуру та функції.

Хід роботи

1. Запустити програму **Blast**. В верхньому лівому кутку вибрати **Protein,** ввести в поле пошуку необхідний білок MamA, MamB, MamM тощо.
2. Обравши шуканий білок у мікроорганізма *Magnetospirillum gryphiswaldense MSR-1* і натиснувши справа **Run Blast**, вирівняти його з необхідним мікроорганізмом – збудником захворювання серця та мозку.
3. За значенням числа E-value (менше ніж 0,02) та Ident (більше 15%) обрати найкращі гомологи.
4. Результати оформити у вигляді таблиці

Бактерія-збудник захворювання	Білок Mam	Білок-гомолог	E-value	Ident

5. Зробити висновки

Контрольні питання

1. Дати визначення Е-числу?
2. Які мікроорганізми є збудниками хвороби серця?
3. Який метод використовували для виявлення гомологів?
4. Яке значення параметра Ident вважається достатнім для встановлення гомології?
5. Яке граничне значення Е-числа для встановлення гомології?

Практичне заняття 6

Еволюційні дослідження білків – гомологів, що відповідають за біомінералізацію БМН

Мета роботи: дослідити еволюційну часову шкалу білків – гомологів, що відповідають за біомінералізацію БМН

Теоретичні відомості

Під час пошуку гомологічних білків, ми намагаємося визначити ті з них, які мали спільного пращура в минулому. На рис. 1 показано загальне еволюційне дерево, корені якого сягають назад до початку історії Землі. Метою порівняння послідовностей білка є взяти послідовність білка, наприклад, кодованого хромосомою людини , і шукати у базі даних білків з метою знайти гомологічні послідовності, часто в дуже віддалених організмах. Таким чином, якщо при пошуку встановлено значний рівень подібності з білком, знайденим в **дріжджах**, то предкова послідовність білка повинна була існувати в організмі не менше **1 мільярду років** тому і послідовність цього організму збереглися в сучасних людині і дріжджах. Аналогічно, якщо послідовність дріжджового білка гомологічна знайденій в паличці ***E. coli***, то ця послідовність повинна була існувати *2 млрд років* тому в первісному організмі, що був предком бактерій і грибів.

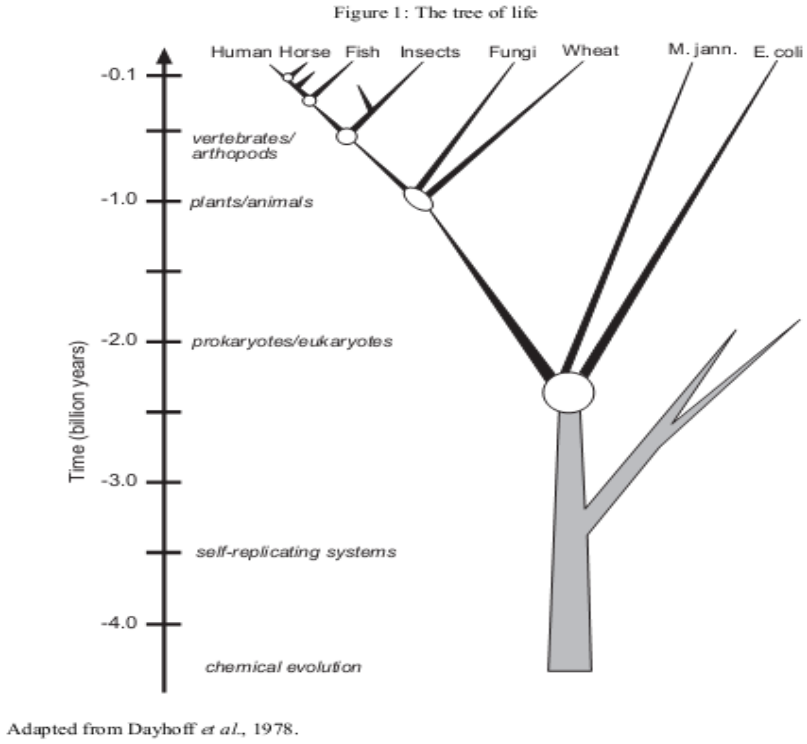


Рисунок 6.1. – Загальне еволюційне дерево

Під час дослідження послідовностей білка або ДНК, ми майже завжди вивчаємо сучасні (на теперішній час) послідовності. Таким чином, не має

ніякого сенсу твердження, що послідовності дріжджів або бактерій є більш примітивними, ніж послідовності у ссавців; всі ці послідовності є сучасними. Однак, як ми побачимо пізніше, є приклади послідовностей, які виявлено тільки в хребетних, або тільки в тварин або рослин, але не в обох з них. Такі послідовності менш древні, ніж ті, що наявні і у ссавців, і у бактерій.

Для організмів, які дивергували протягом останніх 600 млн. років, інформація про дивергенцію для сучасних організмів взято з геологічних даних; більш древні часи дивергенції виводяться з екстраполяції еволюційних «годинників». **Еволюційні годинники засновані як на послідовностях білків, так і рибосомальних РНК**, **які повільно змінюються**; оцінка часу розбіжності вимагає показника швидкості змін, що в середньому є постійною. Найдавнішим скам'янілостям прокаріот в скелях понад 2,5 мільярди років; цей геологічний вік узгоджується з віком виведеним з еволюційних темпів дивергенції.

Теоретичний погляд на дані таблиці 1 свідчить про те, що можна виділити білки, які характеризуються наявністю близько 20% ідентичних послідовностей по всій довжині. Це буде зрозуміло з подальших прикладів, де буде показано, що якщо дві послідовності білка мають 25% ідентичних амінокислотних залишків по всій довжині, то вони гомологічні, а в деяких випадках, переконливим свідченням спільного походження може бути тільки 20% ідентичних амінокислотних залишків. Виявлений еволюційний час може бути підтверджено на практиці, наприклад, з використанням ретельних високоточних алгоритмів порівняння послідовностей можна встановити значну схожість між глобінами рослин і тварин.

Таблиця 6.1. Еволюційні горизонти (PAMs, point accepted mutations – набуті точкові мутації)

Білок	PAMs/100 залишків /10^8 років	Теоретичний Lookback-час, років тому	Горизонт
Псевдогени	400	45 млн.	Примати, Гризуни
Фібринопептиди	90	200 млн.	Поширення ссавців
Лактальбуміни	27	670 млн.	Хребетні
Рибонуклеази	21	850 млн.	Тварини
Гемоглобіни	12	1.5 (х 1000 млн.)	Рослини/Тварини
Кислі протеази	8	2.3 (х 1000 млн.)	Прокаріоти/ Еукаріоти
Трифосфатізомерази	3	6 (х 1000 млн.)	Археї
Глутаматдегідрогенази	1	18 (х 1000 млн.)	

Звичайна дивергенція від спільного пращура

Гомологічні послідовності можна поділити на 2 групи: (1) ортологічні послідовності, які відрізняються, оскільки вони знаходяться у різних видів, і (2) паралогічні послідовності - послідовності, які відрізняються в результаті події дуплікації генів. Рис. 3 показує еволюційне дерево для послідовностей ортологів цитохрому *c*. Розгалужений візерунок, який відображає відмінності між послідовностями цитохромів с, відповідає еволюційним відношенням між видами, у яких експресуються ці білки.

Для багатьох родин білків з різними швидкостями дивергенції, швидкість зміни протягом еволюційного часу є відносно сталою. Ці швидкості можуть бути використані на сьогоднішній день для датування подій дивергенції (наприклад, рослин і тварин), що відбулося понад 600 млн. років і, отже, види не мають скам'янілостей. Проте, різні родини білків дивергують з різною швидкістю, так що, загалом, кількість відмінностей між парою послідовностей не може бути використана для оцінки часу, за який ці дві послідовності дивергували.

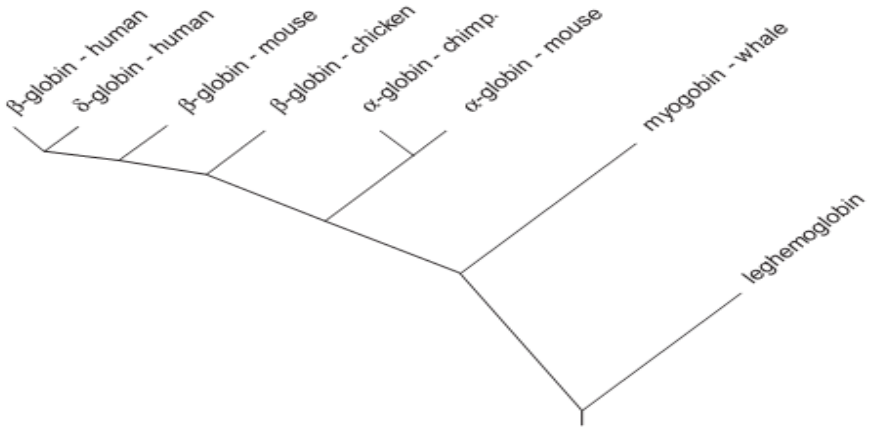


Рисунок 6.2. – Ортологи і паралоги – родина глобінів.

Це особливо вірно для паралогічних послідовностей; оскільки послідовність дублюється, вона може змінитися дуже швидко, перш ніж селективний тиск на її нову функцію уповільнить швидкість її зміни.

Хід роботи

1. Запуск ресурсу **NCBI** «http://www.ncbi.nlm.nih.gov/».
2. В верхньому лівому кутку вибрати **Protein**. Ввести в поле пошуку білок (**Mam-білок**) по якому буде проводитися дослідження. Натиснути в верхньому правому кутку **Search**. Обрати організм ***Magnetospirillum gryphiswaldense MSR-1***.
3. Справа в модулі **Analyze this sequence** натиснути **Run BLAST**.
4. Обрати організм ***Cyanobacteria*** і натиснути внизу зліва **BLAST**.

5. Обрати 3 гомологічні білки до досліджуваного білку магнітосомного острівця *Magnetospirillum gryphiswaldense MSR-1*.

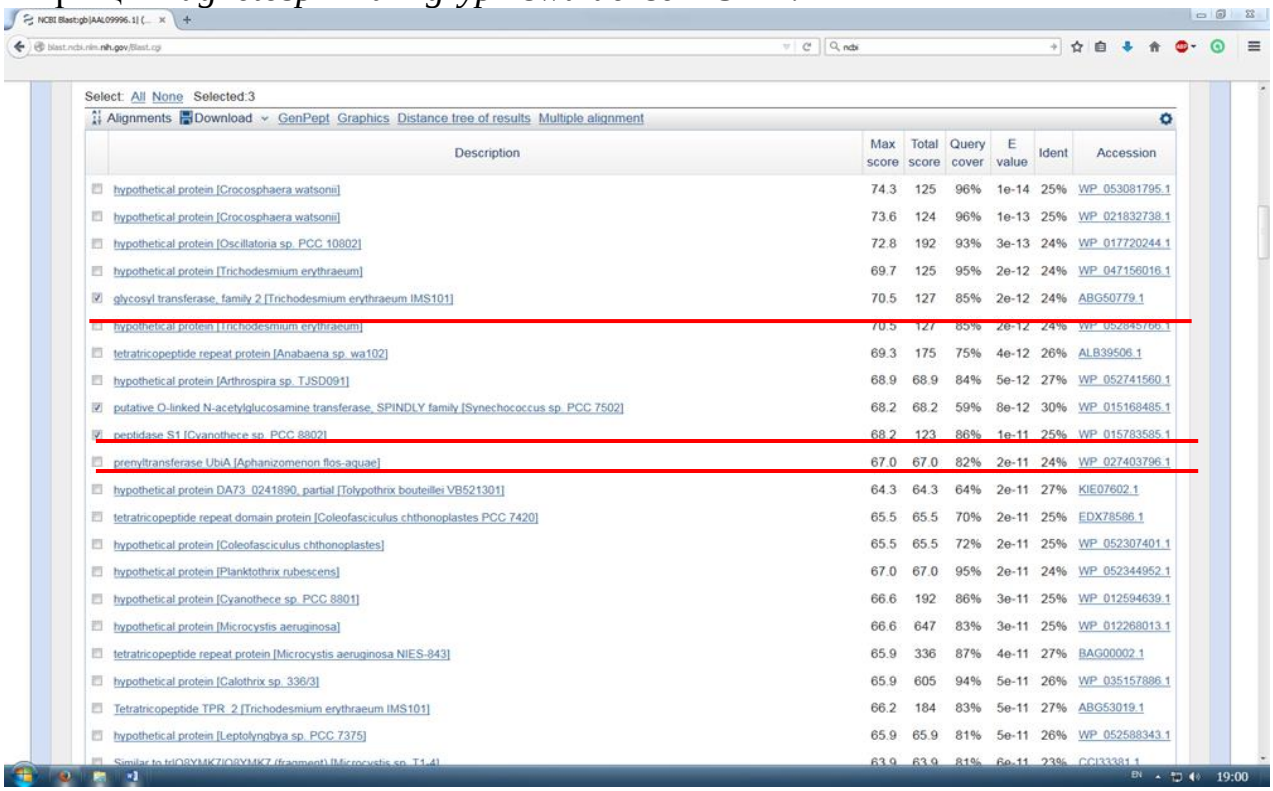


Рисунок 6.3 – Вибір гомологів

6. Визначити для обраних білків-гомологів значення параметрів **E-value** та **Ident**. Отримані результати занести до таблиці.
7. У новому вікні запустити ресурс **NCBI** «http://www.ncbi.nlm.nih.gov/».
8. В верхньому лівому кутку вибрати **Protein**. Ввести в поле пошуку древній білок **Triosephosphate isomerase**. Натиснути в верхньому правому кутку **Search**. Обрати організм ***Arabidopsis thaliana*** (або будь-який інший).
9. Справа в модулі **Analyze this sequence** натиснути **Run BLAST**.
10. Обрати організм ***Cyanobacteria*** і натиснути внизу зліва **BLAST**.
11. Обрати білок-гомолог до досліджуваного древнього білку (бажано обирати білок ціанобактерій, якщо такий є).
12. Визначити для обраного білку-гомологу значення параметрів **E-value** та **Ident**. Отримані результати занести до таблиці.
13. Повторити пункти 7-12 для наступних древніх білків *Glutathione reductase* та *Glutamate dehydrogenase*.
14. Оформити результати в MS WORD у вигляді таблиць. Порівняти отримані значення параметрів **E-value** та **Ident** гомологічних білків магнітосомного острівця магнітотаксисних бактерій та древніх білків ціанобактерій. Зробити висновки.

Контрольні питання

1. На якому принципі заснований еволюційний годин?
2. Принцип пошуку еволюційних відстаней між білками, ДНК та РНК
3. Назвіть типи гомологічних послідовностей
4. Що таке ортологи?
5. Що таке паралоги?

Практичне заняття 7

Множинне вирівнювання за допомогою програми ClustalOmega

Мета роботи: вирівняти декілька послідовностей, виявити гомологію вхідних послідовностей, виділити їх консервативні ділянки, провести філогенетичний аналіз.

Теоретичні відомості

Множинне вирівнювання – алгоритмічно дуже складна задача, тому для її вирішення використовується декілька методів:

- динамічне програмування,
- прогресивні методи,
- ітеративні методи,
- профільний аналіз для знаходження мотивів та ін.

Множинне вирівнювання будується точно так, як і попарне вирівнювання методом д**инамічного програмування,** яке потребує конструювання *k*-вимірного еквівалента матриці ваг префіксів.

- Вирівнювання 2-х послідовностей - двовимірна матриця
- Вирівнювання 3-х послідовностей - тривимірна матриця
- Вирівнювання ***k*** послідовностей – ***k***-вимірна матриця

Припустимо, що відомі *k* послідовностей, довжиною *n*:

$$s^1 = s_1^1 \ldots s_n^1, \quad s^2 = s_1^2 \ldots s_n^2, \quad s^k = s_1^k \ldots s_n^k \tag{7.1}$$

(верхній індекс – номер послідовності, нижній індекс – номер символу у послідовності). У даному випадку для зберігання ваг префіксів знадобиться *k*-вимірний масив розмірністю у кожному вимірі.

Модифікація основного алгоритму для множинного вирівнювання:

1. ***Ініціалізація масиву ваг префіксів a.***

$$a(0,\ldots,0) = 0 \tag{7.2}$$

2. Побудова матриці ваг префіксів вирівнювання k послідовностей

$$a(i_1, i_2 .. i_k) = \max \begin{cases} a(i_1 - 1, \ldots i_k - 1) + \omega_S\left(s_{i_1}^1, \ldots s_{i_k}^k\right), \\ a(i_1, i_2 - 1, \ldots i_k - 1) + \omega_S\left(\_, s_{i_2}^2, \ldots s_{i_k}^k\right), \\ a(i_1 - 1, i_2, i_3, \ldots i_k - 1) + \omega_S\left(s_{i_1}^1, \_, s_{i_3}^3, \ldots s_{i_k}^k\right), \\ \ldots \\ a(i_1 - 1, \ldots i_{k-1} - 1, i_k) + \omega_S\left(s_{i_1}^1, \ldots s_{i_{k-1}}^{k-1}, \_\right), \\ a(i_1, i_2, i_3 - 1, \ldots i_k - 1) + \omega_S\left(\_, \_, s_{i_3}^3, \ldots s_{i_k}^k\right), \\ \ldots \end{cases} \tag{7.3}$$

3. Побудова зворотнього шляху, починаючи з *a(n,n,n).*

Час роботи алгоритму

• Для 3-х послідовностей довжиною *n*, час роботи алгоритму пропорційний величині – $(2^k-1)(n^k) = 7n^3$; $O(n^3)$.
• Для *k* послідовностей – $(2^k-1)(n^k)$; $O(2^k n^k)$.

Вибір послідовностей для множинного вирівнювання

Необхідно вирівнювати білки, а не ДНК, якщо є вибір.
Послідовностей краще багато, але не занадто (~ 10-15).

У вибірці краще уникати :
• занадто схожих послідовностей (>90%);
• занадто різних послідовностей (<30%);
• неповних послідовностей (фрагментів);
• тандемних повторів.

Вивчаючи нову послідовність, необхідно використовувати:
- Вирівнювання на основі BLAST.
- Детально охарактеризовані послідовності – анотації.
- Зовсім неохарактеризовані послідовності (hypothetical proteins) – достатній рівень різноманітності.
- Вирівнювання послідовностей по всій довжині.
- Уникати неповних послідовностей (partial sequences).
- E-value – 10^{-40} – 10^{-5}

Множинне вирівнювання необхідне для вирішення наступних задач:
- Анотації послідовностей.
- Передбачення функції білків (наприклад, виявлення залишків, складових активних сайтів ферментів).
- Моделювання 3D – структури білків.
- Реконструкції еволюційної історії послідовностей (філогенія).
- Виявлення патерну функціональних сімейств і сигналів в ДНК.
- Побудови доменних профілів.
- Дизайну праймерів для аналізу ПЛР

Хід роботи

1. Відкриваємо NCBI, вводимо довільний з mam і вибираємо 4-ри довільні білки:

magnetosome protein MamM [*Magnetospirillum gryphiswaldense MSR-1*],
magnetosome protein MamM [*Magnetovibrio blakemorei*],
cation transporter [*Geobacter sp. OR-1*],
cation diffusion facilitator transporter [*Clostridium algidicarnis*]

Окремо відкриваємо кожен білок , після цього відкриваємо вікно FASTA і копіюємо послідовність кожного гомологу.

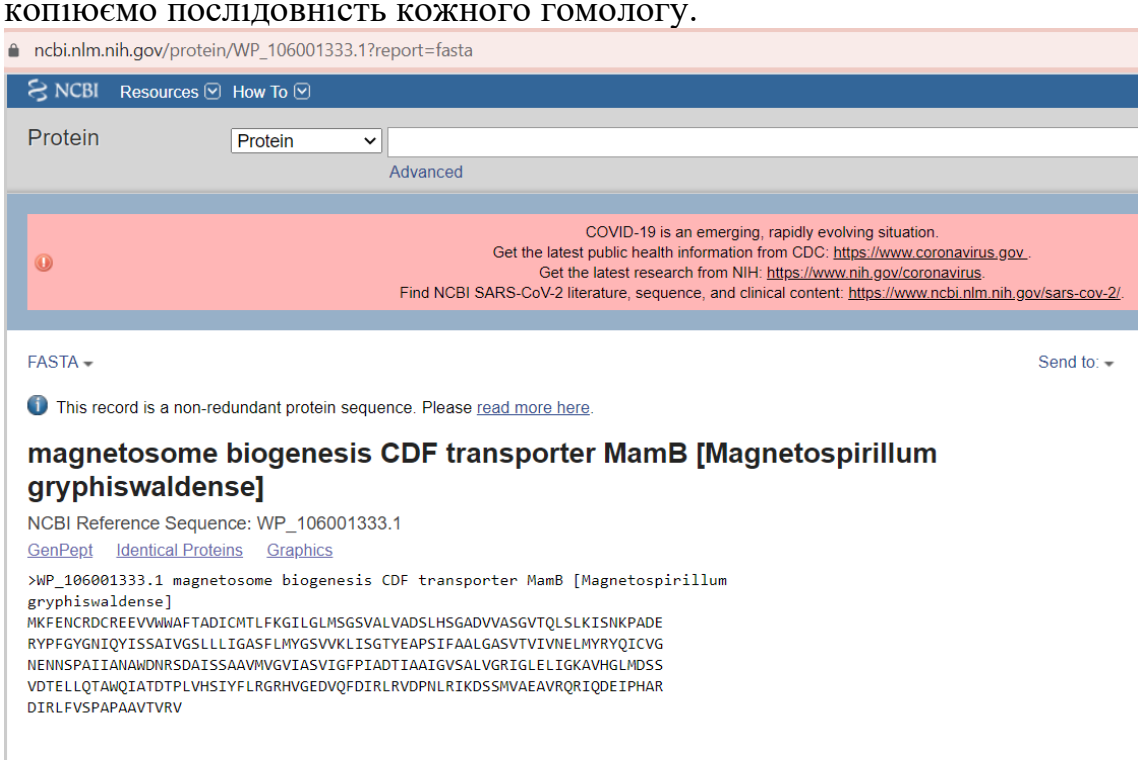


Рисунок 7.1 – Приклад файлу в FASTA форматі

2. Паралельно відкриваємо програму Clustal Omega.
3. У полі *Enter or paste a set of sequences in any supported format*: вставляємо послідовності почергово, видаляючи пробіли.

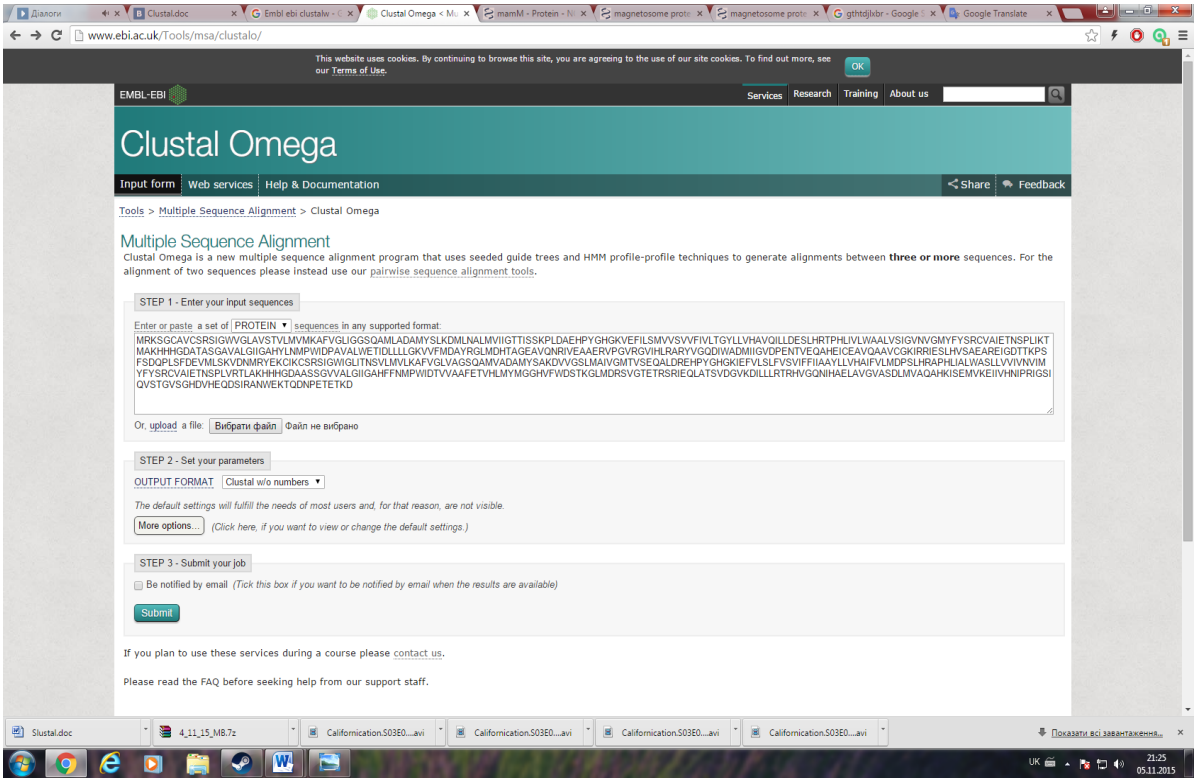


Рисунок 7.2 – Приклад роботи в програмі Clustal Omega

4. Натискаємо кнопку *Submit.*
5. Отримуємо вирівняну послідовність за допомогою множинного вирівнювання:

Рисунок 7.4 – Результат роботи програми Clustal Omega

1: gi|78033502|emb|CAJ30120.1|	100.00	54.66	30.85	33.92
2: gi|238653872|emb|CAV30814.1|	54.66	100.00	26.76	31.47
3: gi|754576520|ref|WP_041969491.1|	30.85	26.76	100.00	31.82
4: gi|657648193|ref|WP_029452641.1|	33.92	31.47	31.82	100.00

6. Зробити відповідні висновки за даними знаючи, що:

- Ідентичні амінокислотні залишки або нуклеотиди відзначаються зірочкою (*****) – єдиний залишок, який зберігся.
- Консервативні заміни – двокрапкою (**:**) – амінокислотні залишки мають схожі властивості.
- Напівконсервативної - точкою (**.**) – амінокислотні залишки слабо аналогічні властивості.
- За кольорами:
  Червоний колір – маленька, гідрофобна амінокислота;
  Синій – амінокислота з кислотними властивостями.
  Рожевий – амінокислота з основними властивостями.
  Зелений – амінокислота має інші радикали (сульфігідрильні групи)

Контрольні запитання:
1. Що таке множинне вирівнювання?
2. Які є види множинного вирівнювання?
3. Алгоритм множинного вирівнювання
4. Практичне значення використання алгоритму множинного вирівнювання

Практичне заняття 8

Підбір праймерів для полімеразної ланцюгової реакції біоінформаціними методами

Мета роботи: освоїти метод біоінформаційного підбору праймерів для полімеразної ланцюгової реакції

Теоретичні відомості

Підбір праймерів здійснюється методом множинного вирівнювання. Крім цього множинне вирівнювання використовують не лише для цього:

- Анотація послідовностей.
- Передбачення функцій консервативних ділянок білків (наприклад, виявлення залишків, складових активних сайтів ферментів, сигналів в ДНК).
- Моделювання вторинної і третинної структури білків.
- Філогенетичний аналіз, реконструкції еволюційної історії послідовностей білків.
- Виявлення характерних фрагментів послідовностей для характеристики білкових сімейств.
- Побудова доменних профілів.
- Дизайн праймерів для аналізу ПЛР.

Множинне вирівнювання – ключовий метод в сучасній молекулярній біології.

Полімеразна ланцюгова реакція - це метод in vitro, використовуваний для того, щоб ферментативно ампліфікувати (помножити) специфічну ділянку ДНК, розташовану між двома ділянками ДНК з відомою послідовністю.

Метод ПЛР заснований на механізмі реплікації ДНК в природних умовах: дволанцюгова ДНК (dsДНК) розкручується до одноланцюгових ДНК (ssДНК), проходить дуплікація і вона знову закручується. Ця методика складається з повторюваних циклів:

• Денатурація ДНК шляхом плавлення при підвищеній температурі для перетворення дволанцюгової ДНК в одноланцюгову ДНК

• Відпал (гібридизація) двох олігонуклеотидів, використовуваних як праймери для цільової ДНК

• Подовження ланцюга ДНК, починаючи від праймерів, шляхом додавання нуклеотидів з використанням ДНК полімерази в якості каталізатора і в присутності іонів Mg^{2+}

Етапи денатурації матриці, відпалу праймерів і подовження праймера складають один «цикл» в методології ампліфікації з використанням ПЛР.

Після кожного циклу, знову синтезований ланцюг ДНК може слугувати матрицею в наступному циклі. Основний продукт експоненційної реакції – це сегмент з dsДНК, кінці якого визначаються 5'-кінцями олігонуклеотидних праймерів, а довжина визначається відстанню між праймерами. Продуктами успішного першого етапу ампліфікації є гетерогенні за розміром молекули

ДНК, довжина яких може перевищувати відстань між сайтами (ділянками) зв’язування двох праймерів. У другому етапі, ці молекули генерують ланцюги ДНК певної довжини, які будуть накопичуватися експоненційно у наступних циклах ампліфікації і утворювати домінуючі продукти реакції. Ефективність ПЛР буде варіювати від матриці до матриці, і у відповідності зі ступенем оптимізації, яка була проведена.

Праймери

В основному, використовують праймери довжиною 16-30 нуклеотидів, що дозволяє застосовувати досить високу температуру відпалу. Слід уникати в структурі праймерів протяжних ділянок послідовності, які складаються з однієї основи (наприклад, poly dG), або повторюваних мотивів - вони можуть гібридизуватися невідповідним чином на матриці, слід уникати інвертованих повторів, щоб не допустити утворення вторинної структури в праймері, яка може заважати гібридизації з матрицею. Слід також уникати послідовностей, комплементарних іншим праймерам, використовуваним в ПЛР, щоб запобігти гібридизації між праймерами, або утворенню димерів праймерів (особливо, це важливо для 3'-кінця праймера). Якщо можливо, 3'-кінець праймера повинен бути багатий G і C основами для підвищення ефективності відпалу цього кінця праймера, який буде зростати. Відстань між праймерами повинна бути менше 10 тисяч основ. Зазвичай, істотне зниження виходу продукту спостерігається, коли праймери віддалені один від одного більше, ніж на 3 тисячі основ. Зазвичай, в ПЛР олігонуклеотиди використовують у концентрації 1 мкМ. Цього достатньо, принаймні, для 30 циклів ампліфікації. Присутність вищих концентрацій олигонуклеотидів може призвести до ампліфікації небажаних, нецільових послідовностей. У противагу цьому, недостатня концентрація праймерів робить метод ПЛР неефективним.

Відпал праймерів

Відпал, або ре-гібридизація ланцюгів ДНК відбувається при температурі (зазвичай 55 - 65 °C). Як тільки температура знижується, дві комплементарні ssДНК починають відтворювати dsДНК молекулу. На цій стадії праймери плавають у реакційному середовищі, і між одноланцюговим праймером і одноланцюговою матрицею постійно утворюються і руйнуються іонні зв'язку. Більш стабільні зв'язки існують трохи довше (праймери, які точно підходять до матричної ДНК), і на цьому маленькому відрізку дволанцюгової ДНК (матриця і праймер), полімераза може приєднатися і почати копіювати матрицю. Як тільки приєднаються кілька основ, іонні сили між матрицею і праймером стають настільки сильними, що не можуть розірватися.

Подовження праймера

На цьому етапі праймери розростаються вздовж цільової послідовності з використанням термостабільної ДНК полімерази (часто - Taq ДНК полімерази) у присутності dNTPs, що призводить до дуплікації вихідного цільового матеріалу. Ідеальна робоча температура для Taq полімерази 72 °C. Коли праймери збільшаться на кілька основ, вони починають володіти більш сильним іонним зв’язком для матриці, що знижує ймовірність зворотного

процесу. Праймери, які не є строго комплементарними, знову вивільняються (внаслідок більш високої температури) і не забезпечують подовження фрагмента. Основи (комплементарні матриці) спаровуються з праймером на 3'-кінці (полімераза додає dNTPs в напрямку з 5' до 3', прочитуючи матрицю від 3' до 5' кінця). Тривалість етапу нарощування праймера може бути збільшена, якщо ділянка ДНК, яка підлягає ампліфікації досить довга.

Хід роботи

1. Запуск ресурсу NCBI«http://www.ncbi.nlm.nih.gov/»
2. В верхньому лівому кутку вибрати **Nucleotide,** ввести в поле пошуку білок по якому буде проводитися дослідження. Обрати організм.

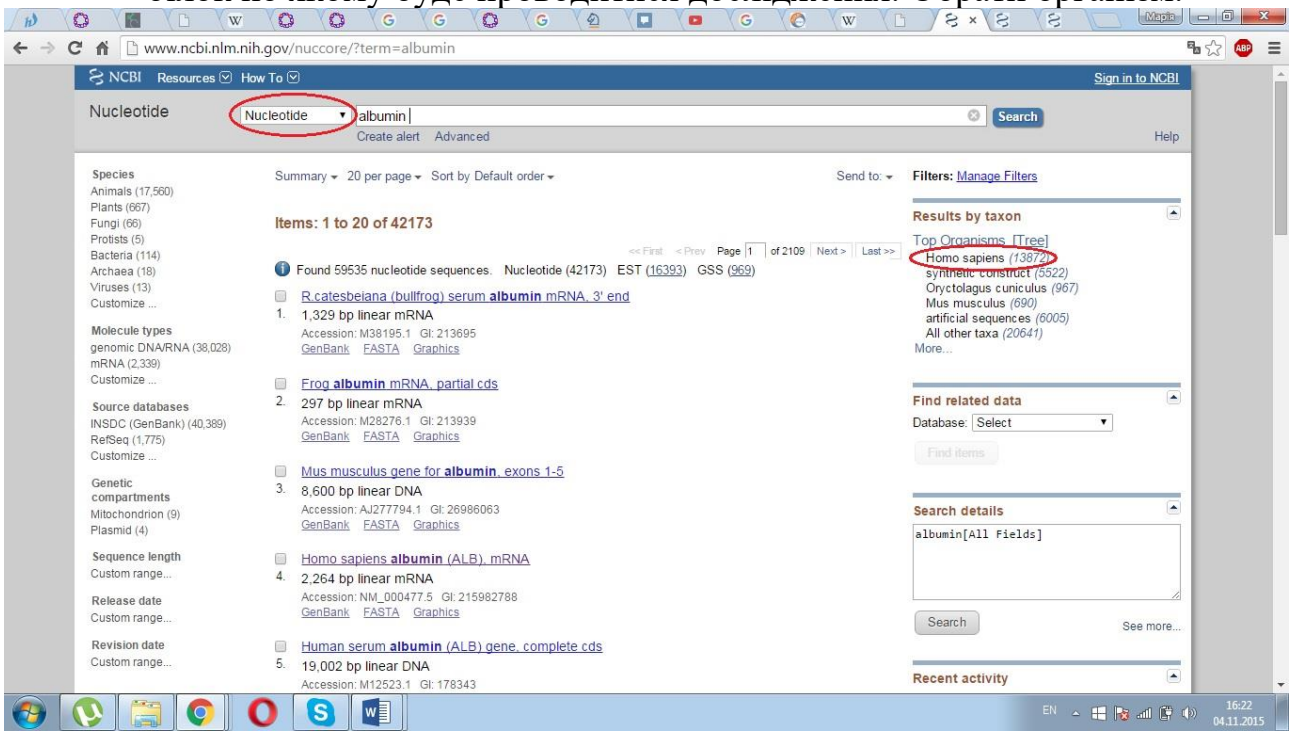


Рисунок 8.1 – Пошук в БД Nucleotide

3. Обрати програмний пакет (формат) **FASTA**, який містить пептидну послідовність.
4. Скопіювати пептидну послідовність.
5. У новому вікні відкрити посилання «http://www.ncbi.nlm.nih.gov/tools/primer-blast/» або у нижньому правому кутку знайти FEATURED і обрати Primer-BLAST.
6. Вставити нуклеотидну послідовність, обрати параметри пошуку й натиснути **GetPrimers.**

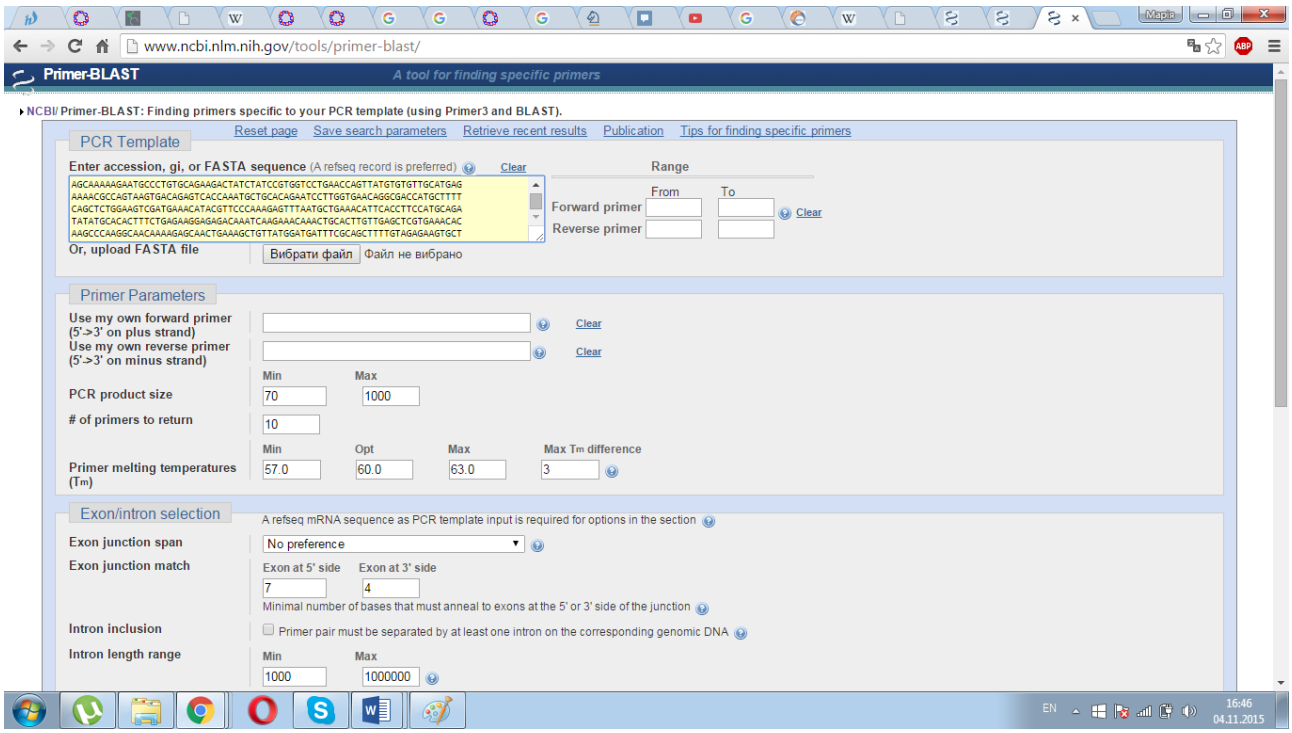


Рисунок 8.2 – Загальний вигляд сторінки програми для пошуку праймерів

7. Оформити результати документом MS WORD у вигляді скріншотів. Зробити висновки.

Контрольні запитання:

1. Що таке праймери, для чого вони потрібні?
2. В яких етапах полімеразної ланцюгової реакції беруть участь праймери?
3. За допомогою якого алгоритму динамічного програмування можна проводити підбір праймерів?

Практична робота 9

Візуалізація 3D структури білків

Мета роботи: оволодіти навиками роботи з 3D структурами білкків

Теоретичні відомості

На сьогоднішній день, відомо про офіційне існування первинної, вторинної, третинної (глобули) і четвертинної структурах білків. Однак четвертинну структуру можна віднести до взаємодій між різними глобулами, тому вона буде цікавити нас у меншій мірі.

Відомо 20 видів амінокислот, з яких будуються всі білки живих організмів.

У більшості літератури можна зустріти трибуквенне позначення назв амінокислот, але в біоінформатиці прийнято використовувати однобуквену номенклатуру (FASTA - формат) для роботи з найбільш перспективними програмами.

Дослідження третинної (просторової) структури білка - одна з найбільш важливих завдань біоінформатики. З часів виникнення біохімії, вченим вдалося встановити безліч закономірностей, завдяки яким стало можливим обчислювати будова і структуру білків. Однак хімічні та біологічні властивості білків вивчаються емпірично (шляхом проведення лабораторних експериментів). Досі, чітка кореляція між будовою білкових молекул та їх властивостями не виявлена. 3D-моделі в майбутньому допоможуть ефективно зіставляти теоретичні та емпіричні дані, щоб створювати штучні білки з необхідним набором властивостей.

Хід роботи

1. Візуалізацію 3D структури білків проведемо на прикладі білків магнітосомного острівця МТБ та їх гомологів
2. Заходимо в базу даних NCBI. В ресурсах знаходимо домени і структури.
3. В Datebases переходимо в Structure (Molecular Modeling Datebases).

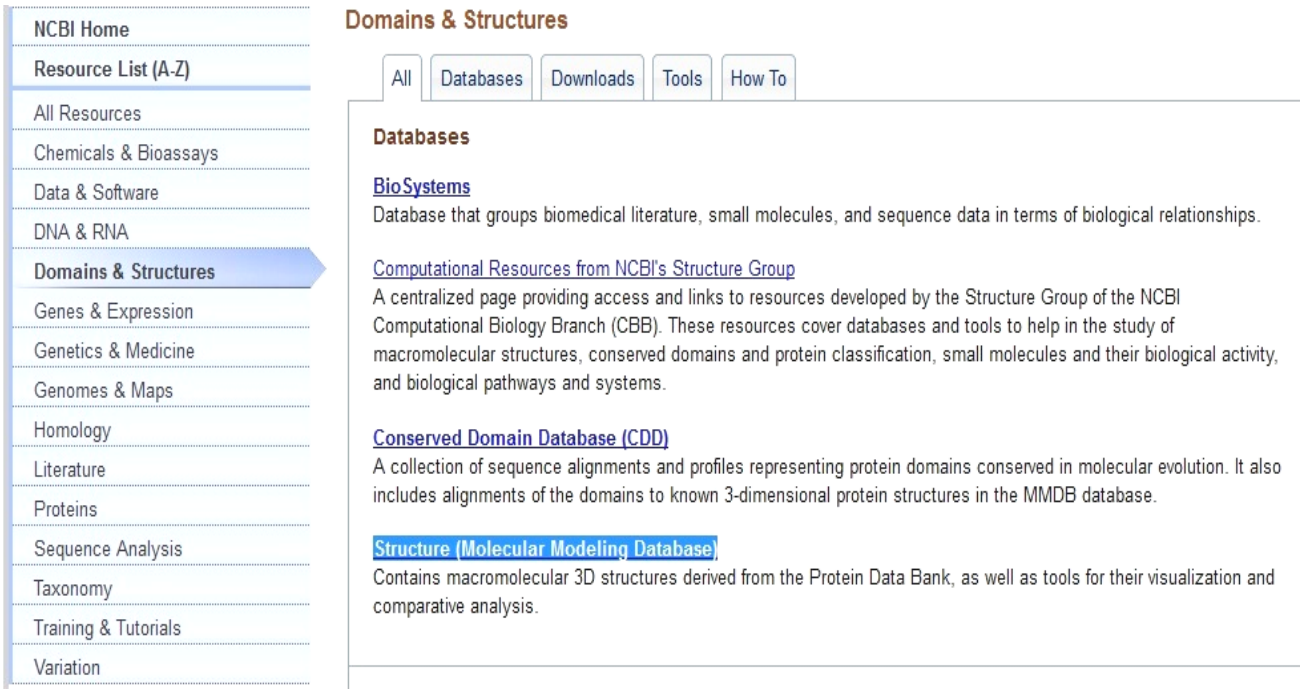


Рисунок 9.1 – Розташування вкладки Domains and Structure

4. В Structure Tools Macromolecular знаходимо Resources Overview.
5. Із ресурсів обираємо один із запропонованих інструментів для перегляду та дослідження 3D структури білків магнітосомного острівця МТБ та їх гомологів. Для цього потрібно знати певні особливості кожного з них.

MMDB: Інструмент, що дозволяє експериментально розв’язати структури білків, РНК та ДНК, отримані з білків банку даних (PDB), з високою доданою вартістю функцій, таких як явних хімічних графів, обчислювально виявлених 3D доменів (компактних підструктур), які використовуються для виявлення схожих 3D структур, а також посилання на літературу, подібних послідовностей, інформації про хімічні речовини, пов'язаної зі структурами, знайти 3D структури для гомологів послідовності шуканого білка, зв’язки в структурі, активні центри, журнальні статті та багато іншого.
Шукають потрібну інформацію за допомогою *MMDB ID* чи *PDB ID* досліджуваного білка.
CBLAST: Інструмент, який порівнює послідовність шуканого білка з всіма білковими послідовностями з дозволених 3D структур за допомогою білка BLAST проти набору даних PDB. Мета полягає в тому, щоб знайти представництва 3D структур для запиту та / або його гомологів, а наявні. Ви також можете ввести послідовність білка безпосередньо в сторінці пошуку CBLAST, щоб знайти схожі послідовності 3D структури. Результати пошуку можна переглянути в Cn3D (звідси і назва "CBLAST").
Шукають потрібну інформацію за допомогою GI досліджуваного білка.
Cn3D: Інструмент для візуалізації тривимірних структур з акцентом на інтерактивній експертизі послідовності структурних зв’язків і суперпозиції

геометрично подібних структур. Може бути використаний для відображення MMDB структури, суперпозиції VAST подібних структур.

VAST: Комп'ютерний алгоритм розроблений в NCBI і використовується для ідентифікації подібних білкових 3-мірні структури за допомогою чисто геометричних критеріїв, а також визначити віддалені гомологи, які не можуть бути визнані порівняння послідовностей.

VAST+: Інструмент, призначений для ідентифікації макромолекули, які мають аналогічні 3-мірні структури, з акцентом на пошуку тих, з подібними біологічними одиницями. Подібність розраховують з використанням чисто геометричних критеріїв, і, отже, можна визначити віддалені гомологи.

6. Вводимо відповідно до інструментів бази даних NCBI MMDB ID чи PDB ID, у випадку використання MMDB, або GI досліджуваного білка при використанні CBLAST.

Наприклад, для magnetosome protein MamA [*Magnetospirillum gryphiswaldense MSR-1*] GI:15865658.

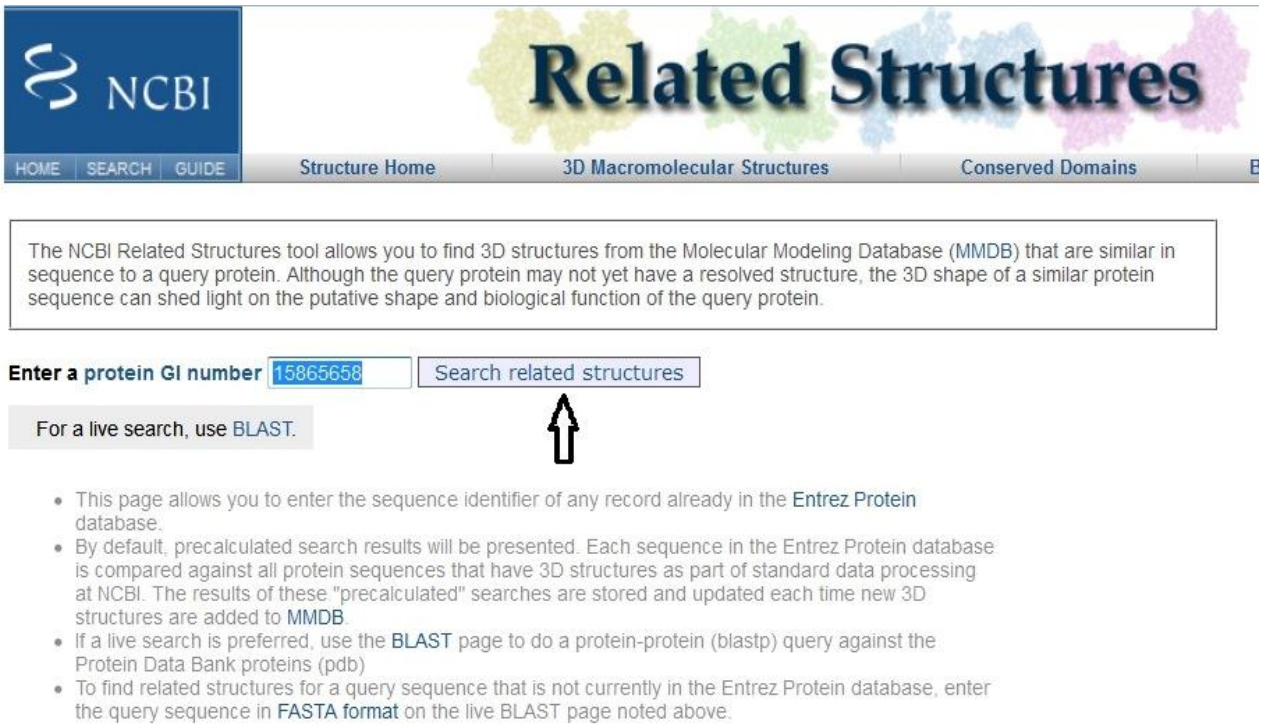


Рисунок 9.2 – Приклад використання поля для пошуку

7. Серед знайдених гомологів обирається певний білок з хорошим *E-value.* Наприклад, 2PL2_A.

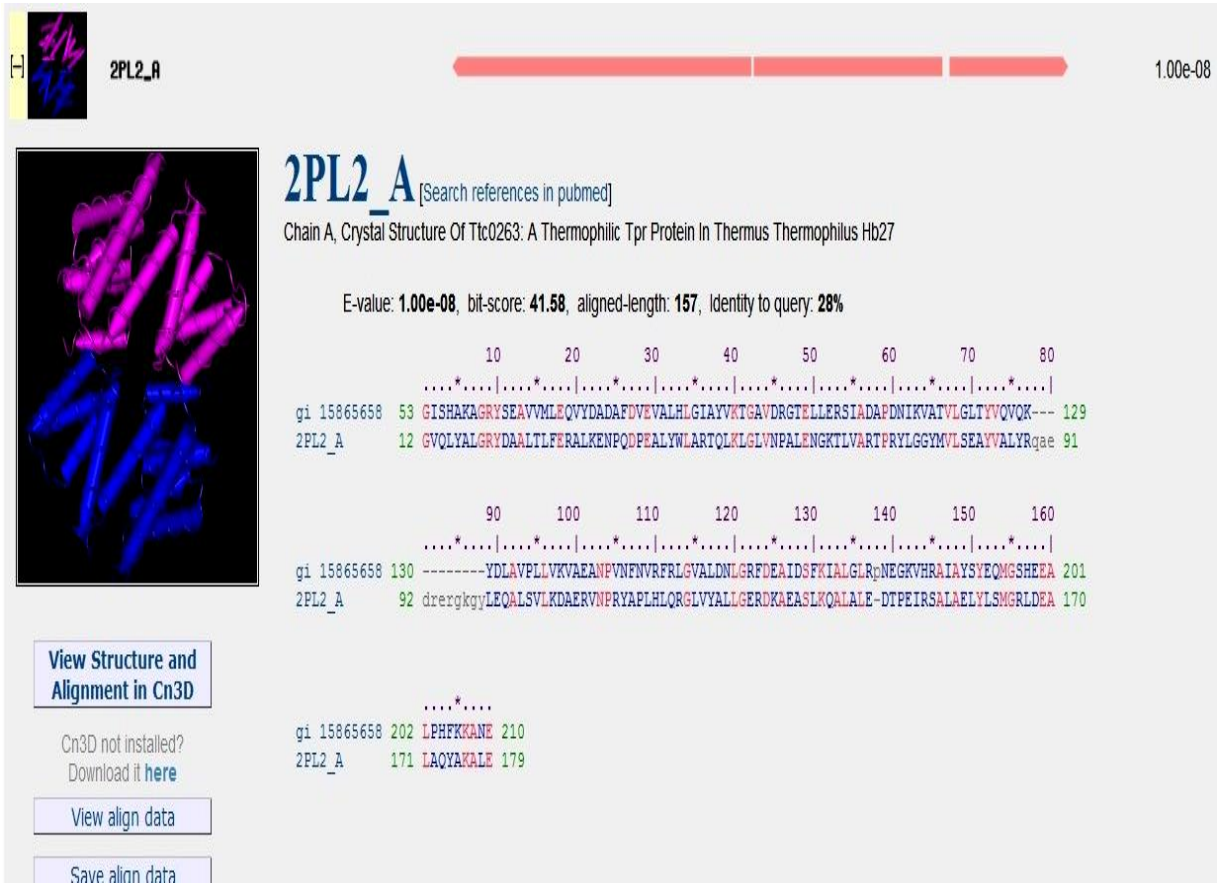
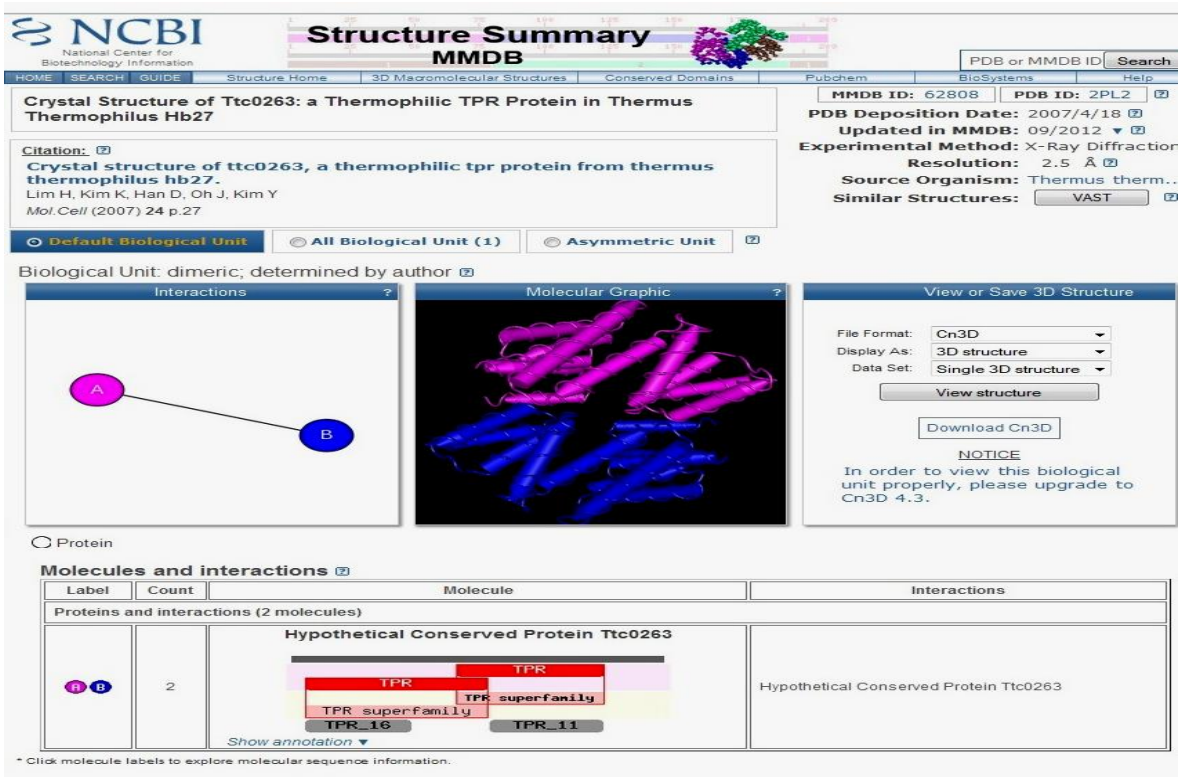


Рисунок 9.3 – Структурні та конформаційні особливості обраного білка



8 Використовуючи програму *ArgusLab* ми можемо бачити 3D структуру обраного білку та працювати з нею, використовуючи *View*.

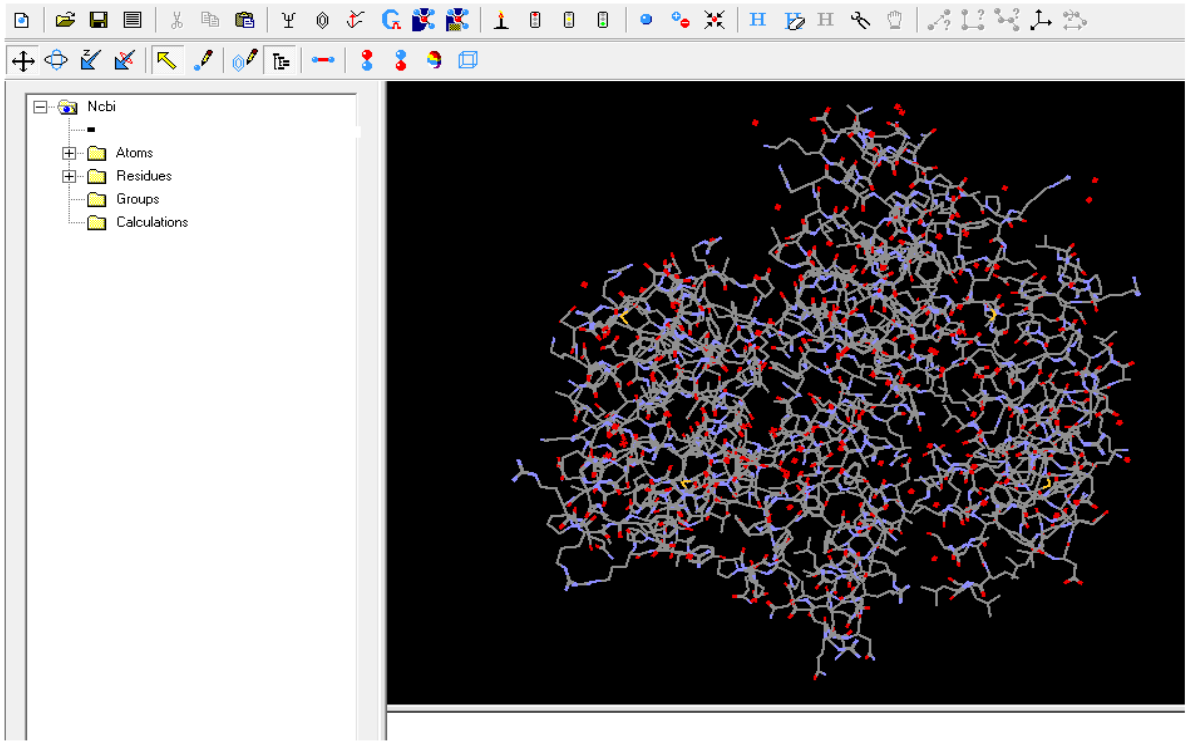


Рисунок 9.4 – 3D структура обраного білку

Контрольні питання

1. За допомогою яких програмних засобів можна вивчати 3д структури білків.
2. Опишіть основні можливості Cn3D.
3. Опишіть основні можливості Vast+.
4. Опишіть основні можливості CBLAST.

Практична робота 10

Побудова філогенетичних дерев

Мета роботи: оволодіти навиками побудови філогенетичних дерев за допомогою біоінформаційних методів

Теоретичні відомості

Філогенетичне дерево (еволюційне дерево, дерево життя) — дерево, що відображає еволюційні взаємозв'язки між різними видами, іншими таксонами, генами або іншими об'єктами, що мають загального пращура.

Вершини філогенетичного дерева діляться на три класи: листя, вузли і (максимум один) корінь. Листя — це кінцеві вершини, тобто ті, в які входять рівно по одному ребру; кожен лист відображає деякий вид живих організмів (або інший об'єкт, схильний до еволюції, наприклад білковий домен). Кожен вузол представляє еволюційну подію: розділення предкового виду на два або більше, які надалі еволюціонували незалежно. Корінь представляє загального предка всіх даних об'єктів. Ребра дерева філогенезу прийнято називати «гілками».

Відомо 20 видів амінокислот, з яких будуються всі білки живих організмів. У більшості літературних джерел можна зустріти трибуквенне позначення назв амінокислот, але в біоінформатиці прийнято використовувати однобуквену номенклатуру (FASTA - формат) для роботи з найбільш перспективними програмами [1].

Основним параметром, що характеризує еволюцію амінокислотних послідовностей, є еволюційна дистанція між ними. Еволюційні дистанції використовуються для побудови філогенетичних дерев (дендрограм) та визначення часу дивергенції, що можливо завдяки існуванню однозначної відповідності між часом дивергенції та ступенем амінокислотних відмінностей при умові постійності швидкостей еволюції білків в різних філогенетичних лініях.

Більш точним методом визначення кількості відмінностей між послідовностями є обчислення долі різних амінокислот в цих послідовностях (позначається p, Pd, p-distance від англ. part of differences). Використовуючи цю характеристику, можна оцінити кількість відмінностей між послідовностями навіть, якщо їх довжини значно відрізняються. p-дистанція визначається за формулою:

$$p_d = n_d / n. \qquad (10.1)$$

Якщо заміни у всіх амінокислотних сайтах відбуваються з рівною ймовірністю однієї заміни на сайт p_1, то ймовірність k замін на сайт розподілена за біноміальним розподілом:

$$P(k) = C^k_n \, p^k_1 (1- p_1)^{n-k} , \qquad (10.2)$$

де $C_n^k = \frac{n!}{k! \cdot (n-k)!}$, k - кількість замін.

Дисперсія величини P(k) визначається за формулою:

$$D(p) = np(1 – p). \quad (10.3)$$

Слід відмітити, що p-дистанція не є строго пропорційною часу дивергенції таксономічних груп організмів t і тому описаний метод використовується для отримання приблизної оцінки еволюційних відмінностей амінокислотних послідовностей.

Еволюційна дистанція Кімури d_K є одним з варіантів корекції p-дистанції. Її величина обчислюється по емпіричній формулі

$$d_K = -\ln(1 - p - 1/5p^2) \quad (10.4)$$

Кімура з'ясував, що ця формула справедлива при значеннях p, які не перевищують 0,75.

Еволюційна відстань між двома послідовностями розраховується у таких моделях:

1) Kimura: не може бути обчислена для фракції неузгоджених основ, більших за 0,75.

2) Grishin: приблизно така ж модель, як Кімура, але може бути обчислена для фракції неузгоджених амінокислот, більших за 0,75.

3) Grishin General: більш загальна еволюційна модель: норми заміщення варіюються як для амінокислот, так і для сайтів.

Для побудови філогенетичного дерева за допомогою ресурсів NCBI спочатку треба обрати білки, які будуть об'єктом аналізу. В даній лабораторній роботі будемо порівнювати білки магнітосомного острівця магнітотаксисних бактерій з білками тварин. Для цього заходимо в базу даних NCBI та вводимо в пошукову строку білок mamA. Обираємо пошук по базі даних білків. Результатом пошуку є перелік різних гомологів білку mamA в різних організмах.

Обираємо для порівняння перший варіант пошуку – білок mamA в бактерії *Magnetospirillum magnetotacticum*. Переходимо за посиланням на даний білок й далі праворуч в меню переходимо в програму BLAST.

В новому вікні відкривається система BLAST для порівняння амінокислотних послідовностей. У полі «Enter accession number(s), gi(s), or FASTA sequence(s)» бачимо номер доступу (accession number) обраного білка. Для порівняння даного білка з іншими організмами, що є в базах даних NCBI, в полі «Organism» вводимо назву однієї з таксономічних груп травин, наприклад, Aves (птахи). В полі «Algorithm» обираємо алгоритм порівняння послідовностей, в даному випадку це порівняння білок-білкових послідовностей (blastp (protein-protein BLAST)). Далі натискаємо на кнопку «BLAST».

В новому вікні бачимо результати порівняння послідовності обраного білку з білками птахів.

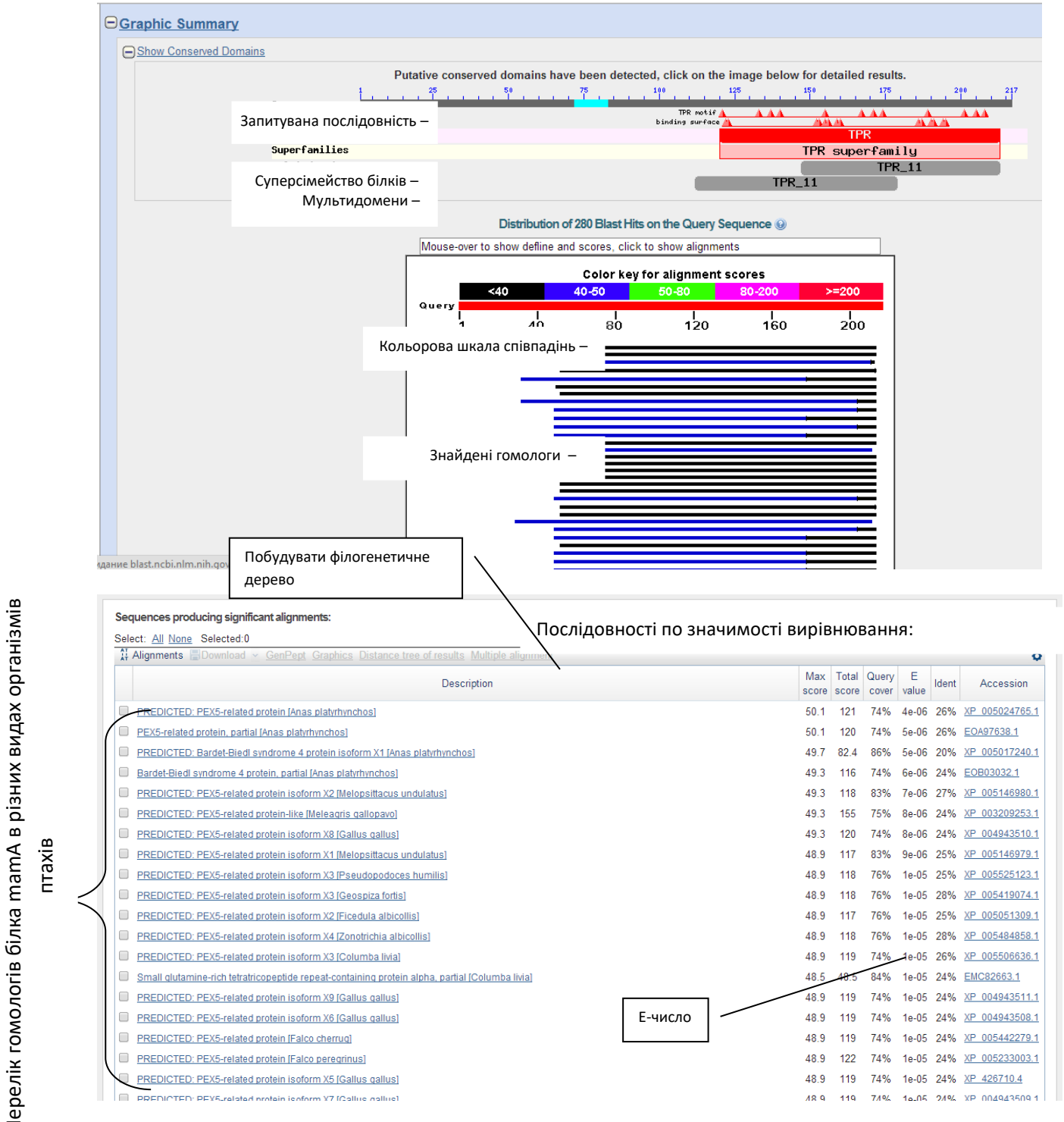


Рисунок 10.1 – Критерії вибору вирівнювань для побудови філогенетичного дерева

Для побудови філогенетичного дерева ставимо «галочку» навпроти тих гомологів, які нас цікавлять (наприклад, обираємо лише перший результат вирівнювання). Далі натискаємо на кнопку «Distance tree of results». В новому вікні відкривається побудоване філогенетичне дерево.

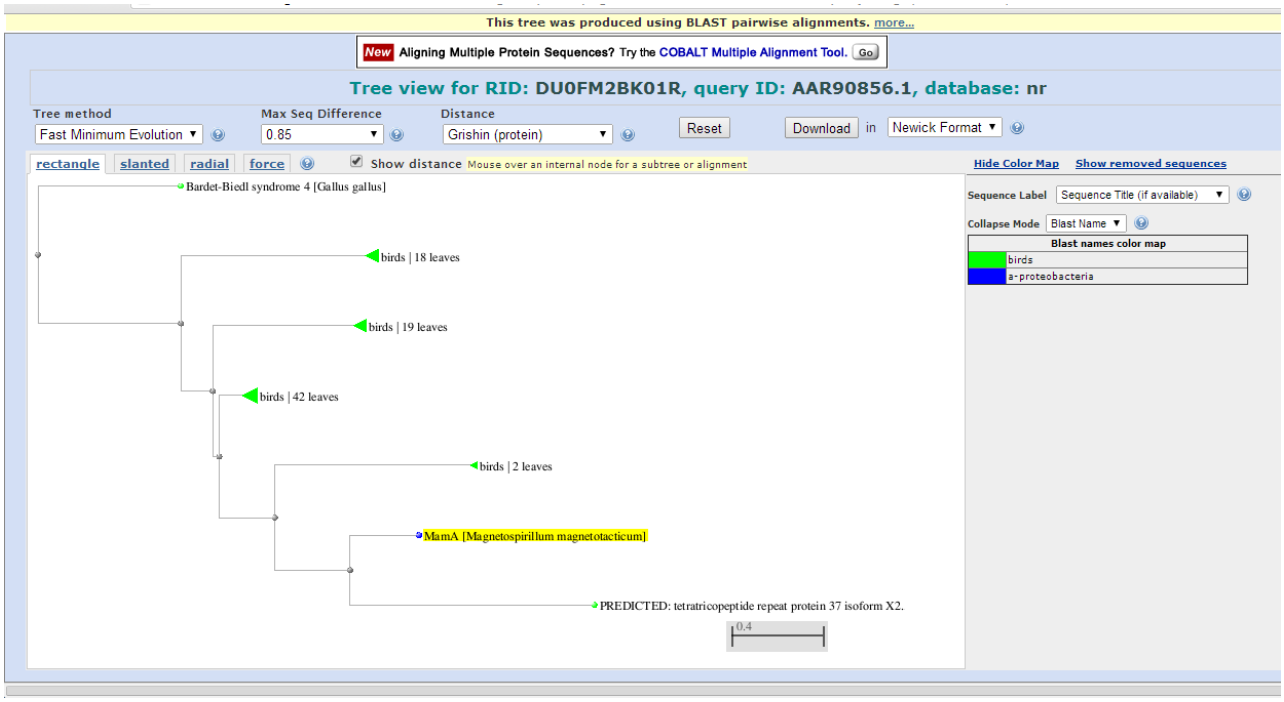


Рисунок 10.2 – Приклад побудовано філогенетичного дерева

На рисунку бачимо відрізок, що означає довжину дистанції Грішина. Для отримання точних даних необхідно завантажити файл з усіми дистанціями, натиснувши на кнопку «Download».

Хід роботи:

1. За допомогою інтернет-ресурсу NCBI побудувати філогенетичне дерево для білків у тварин, які є гомологами білків магнітосомного острівця магнітотаксисних бактерій. (*Magnetospirillum magnetotacticum*).
2. На відомому філогенетичному дереві для тварин позначити набір гомологів, згідно BLAST-результату.
3. Позначити середню відстань Грішина між гомологом тварини і білком МО МТБ. Написати визначення відстані Грішина.
4. При виборі гомолога порівняти E-числа визначених вирівнювань. E-число має бути не більше 10^{-5}.
5. Для аналізу обрані гени магнітосомного острівця: mamZ, mamE, mamA, mamN, mamB, mamH, mamN, mamM, mamQ.

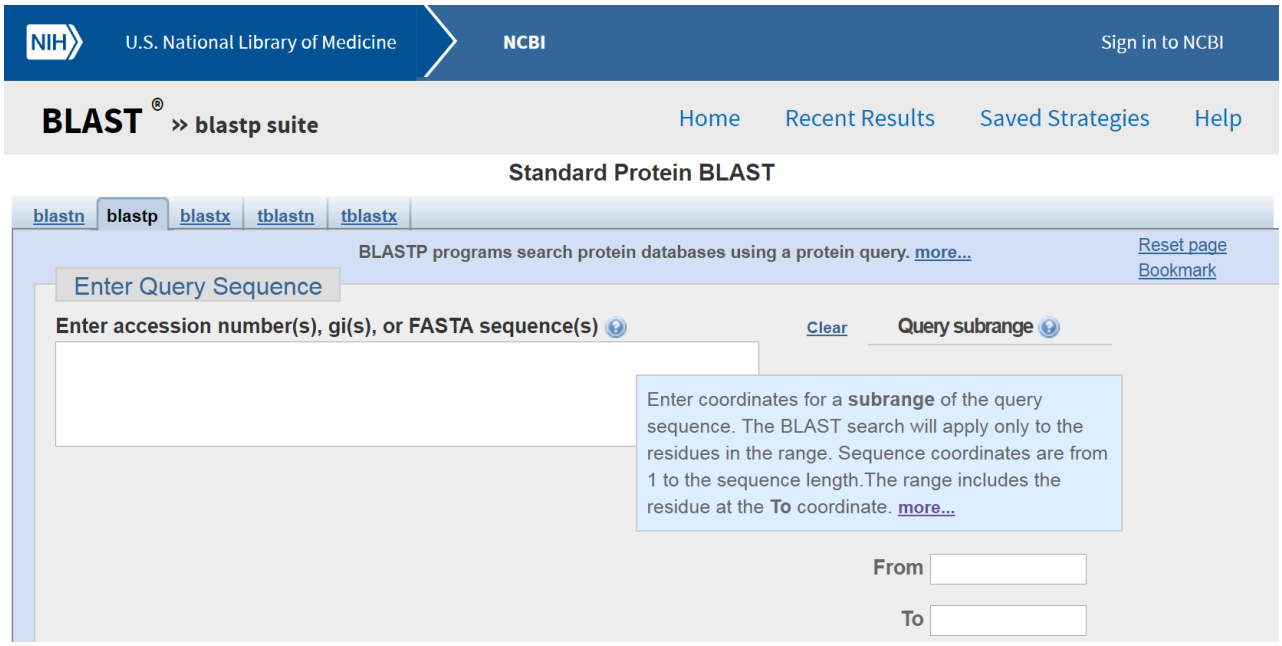
Контрольні питання

1. Що таке філогенетичне дерево?
2. Які параметри використовують для побудови філогенетичного дерева?
3. Будова філогенетичного дерева.
4. Що таке відстань Грішина?

Додаток

Опис основних параметрів BLASTP та BLASTN

BLASTP

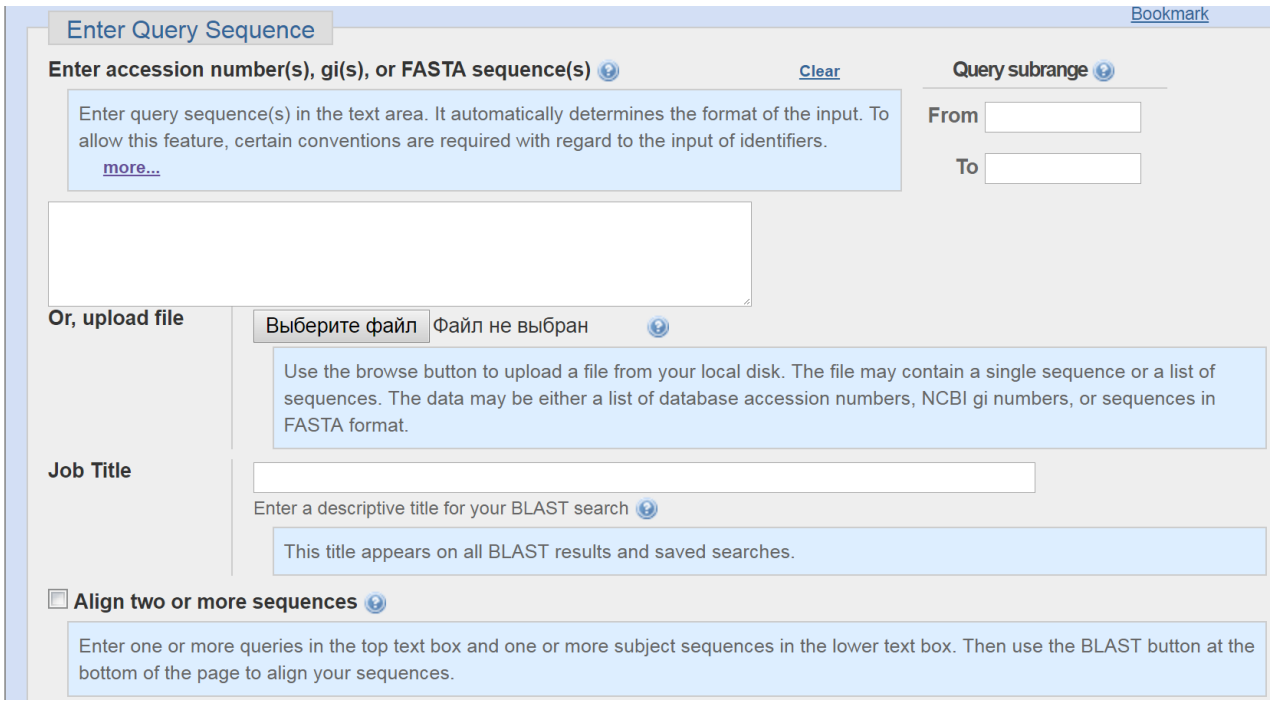


Query subrange

Піддіапазон запиту

Enter coordinates for a subrange of the query sequence. The BLAST search will apply only to the residues in the range. Sequence coordinates are from 1 to the sequence length. The range (межі) includes the residue (залишок) at the **To** coordinate.

Ввести координати послідовності запиту. Координати послідовності від 1 до довжини послідовності. BLAST працює тільки з вказаним діапазоном.



Enter Query Sequence

Введення послідовності запиту

Enter accession number(s), gi(s), or FASTA sequence(s). Enter query sequence(s) in the text area. It automatically determines the format of the input. To allow this feature, certain conventions are required with regard to the input of identifiers.

Ввести номер доступу, унікальний номер послідовності в БД або послідовність у FASTA форматі. Ввести послідовність запрос в поле для тексту. Формат введених даних визначається автоматично. Щоб функція виконувалась, точні правила вимагаються відносно ідентифікаторів.

Or, upload file

Або завантажте файл

Use the browse button to upload a file from your local disk. The file may contain a single sequence or a list of sequences. The data may be either a list of database accession numbers, NCBI gi numbers, or sequences in FASTA format.

Використовуйте кнопку перегляду для завантаження файлу з Вашого локального диску. Файл може містити одну послідовність або ж список

послідовностей. Дані можуть також являти собою список номерів доступу у базі даних, унікальний номер послідовності в БД NCBI, чи послідовность у FASTA форматі.

Job Title

Назва роботи

Enter a descriptive title for your BLAST search.

This title appears on all BLAST results and saved searches.

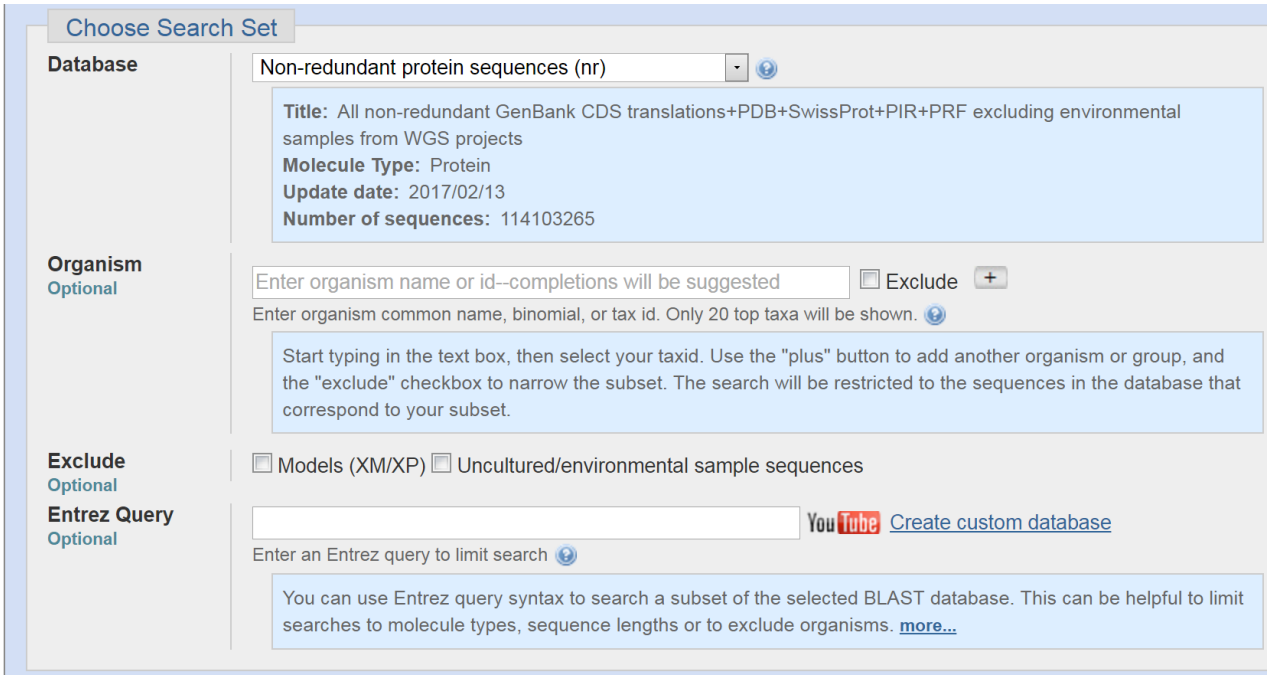
Введіть відповідну назву для Вашого пошуку у BLAST. Ця назва з’явиться на всіх результатах BLAST і збережених пошуках.

Align two or more sequences

Вирівнювання двох або більше послідовностей

Enter one or more queries in the top text box and one or more subject sequences in the lower text box. Then use the BLAST button at the bottom of the page to align your sequences.

Введіть одну чи більше послідовностей запитів в верхню текстову комірку і одну чи більше послідовність з БД в нижню текстову комірку. Потім використайте кнопку BLAST внизу сторінки для вирівнювання Ваших послідовностей.



Choose Search Set

Виберіть параметри пошуку

Organism

Optional

Організм (не обов'язково)

Start typing in the text box, then select your taxid. Use the "plus" button to add another organism or group, and the "exclude" checkbox to narrow the subset. The search will be restricted to the sequences in the database that correspond to your subset.

Почніть вводити текст в текстовому полі, а потім оберіть taxid. За допомогою кнопки "плюс", щоб додати ще один організм або групу, а також кнопку - прапорець "виключити", щоб обмежити підмножини. Пошук буде обмежений послідовностями в базі даних, які відповідають Вашій підгрупі.

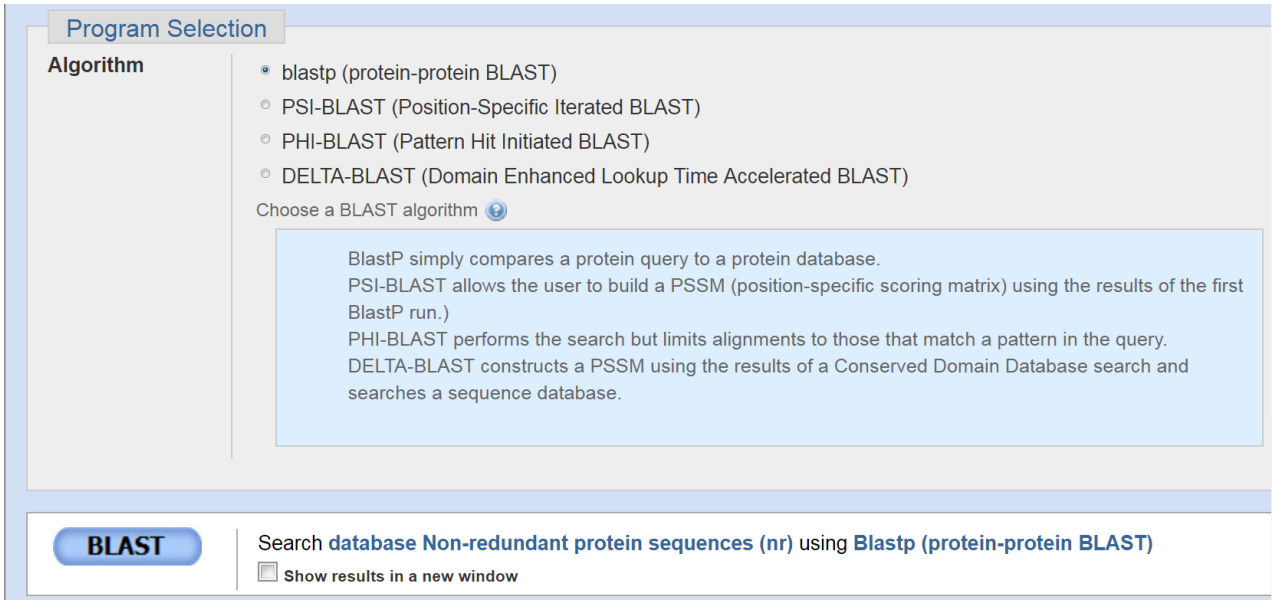
Entrez Query

Optional

Послідовність запит (не обов'язково)

You can use Entrez query syntax to search a subset of the selected BLAST database. This can be helpful to limit searches to molecule types, sequence lengths or to exclude organisms.

Ви можете використовувати **Entrez** (пошукова система) для пошуку підмножини вибраних в БД BLAST-ом. Це може бути корисно для обмеження пошуків до молекулярних типів, довжин послідовностей, або щоб виключити організми.



Program Selection

Вибір програми

Algorithm

Вибір алгоритму

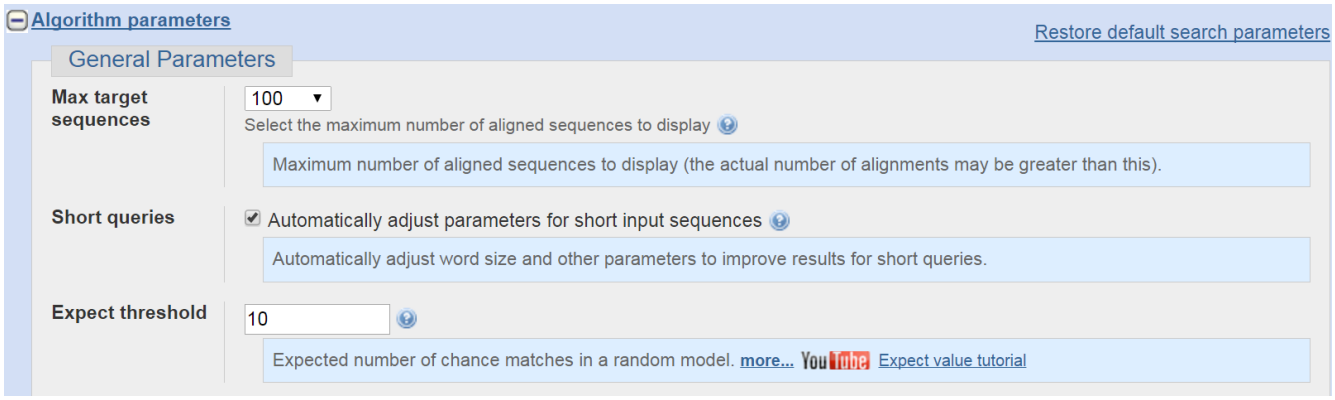
Blastp simply compares a protein query to a protein database.

PSI-BLAST allows the user to build a PSSM (position-specific scoring matrix) using the results of the first Blastp run.)

PHI-BLAST performs the search but limits alignments to those that match a pattern in the query.

DELTA-BLAST constructs a PSSM using the results of a Conserved Domain Database search and searches a sequence database.

Blastp – просто порівнює амінокислотну послідовність з білками в базі даних.
PSI-BLAST – дозволяє побудувати специфічну матрицю ваг (PSSM), використовуючи результати першого запуску Blastp.
PHI-BLAST – виконує пошук, але обмежує вирівнювання паттернів в запиті.
DELTA-BLAST – конструює PSSM матрицю, використовуючи результати пошуку в БД консервативних доменів і пошук послідовностей в базі даних.



Max target sequences

Максимальна кількість цільових послідовностей

Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Максимальна кількість вирівняних послідовностей для відображення (фактична кількість вирівнювань може бути більше, ніж ця).

Short queries

Короткі запити

Automatically adjust word size and other parameters to improve results for short queries.

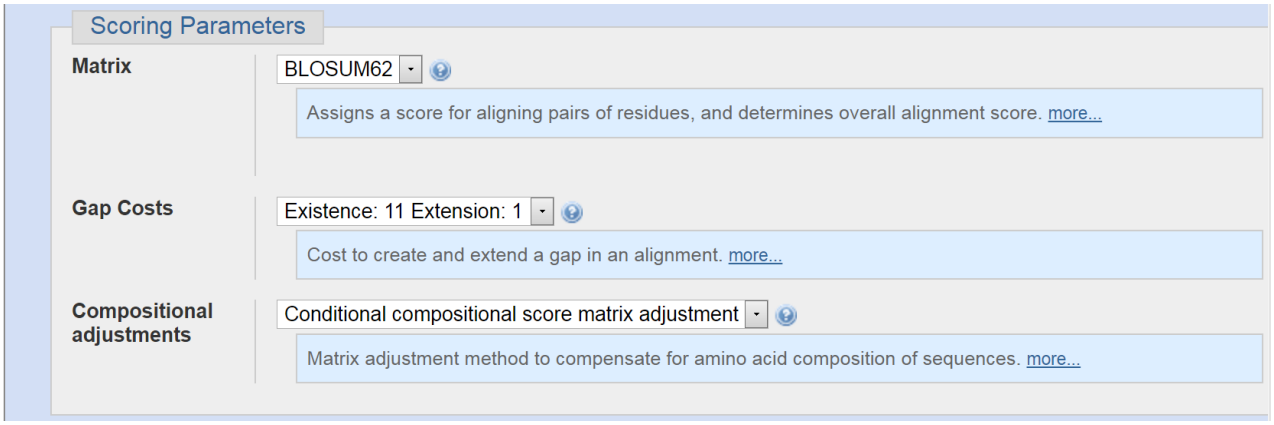
Автоматично встановлюється розмір слова та інші параметри для поліпшення результатів для коротких запитів.

Expect threshold

Поріг для Е-числа

Expected number of chance matches in a random model

Ймовірнісна кількість збігів у випадковій моделі.



Scoring Parameters

Вагові параметри

Matrix

Матриця

Assigns a score for aligning pairs of residues, and determines overall alignment score.

Призначає вагову матрицю для вирівнювання пар залишків, і визначає загальну вагу вирівнювання.

Gap Costs

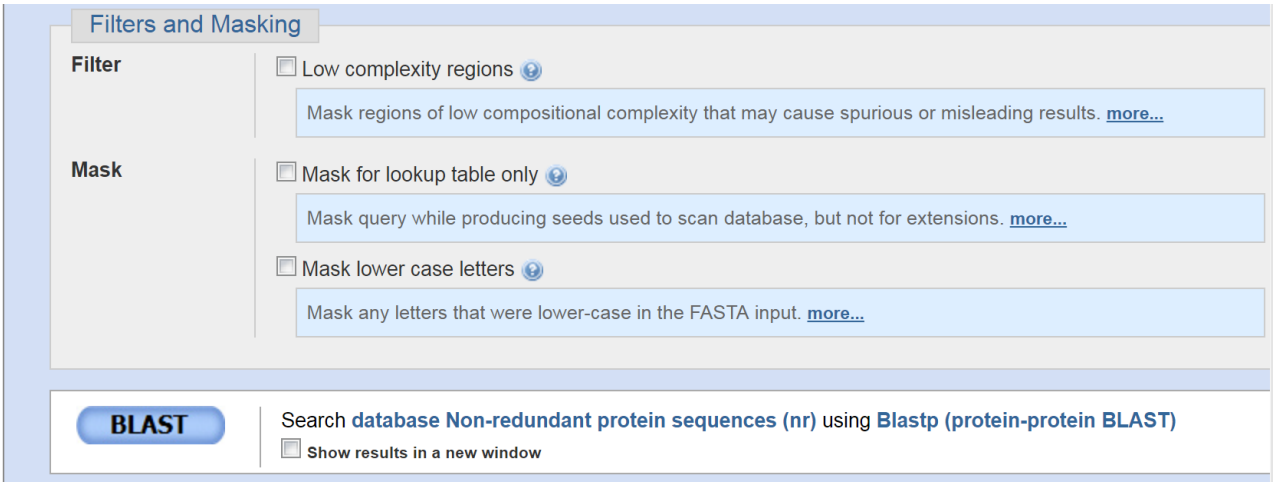
Ціна пробілів

Compositional adjustments

Композиційні регулювання

Matrix adjustment method to compensate for amino acid composition of sequences

Матричний метод регулювання для побудови амінокислотного складу послідовностей



Filter

Фільтри

Mask regions of low compositional complexity that may cause spurious or misleading results.

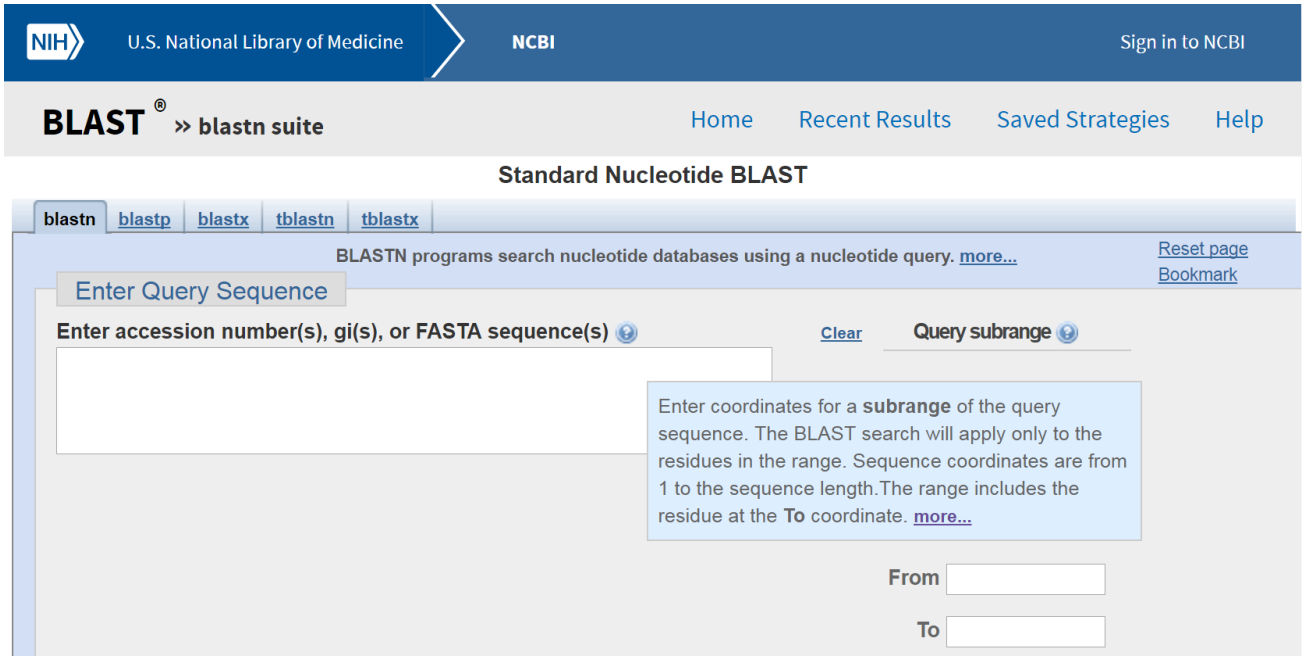
Маска області низької композиційної складності, які можуть викликати помилкові результати.

Mask

Mask any letters that were lower-case in the FASTA input.

Маскування будь-яких літер, які були в нижньому регістрі на вході FASTA.

BLASTN

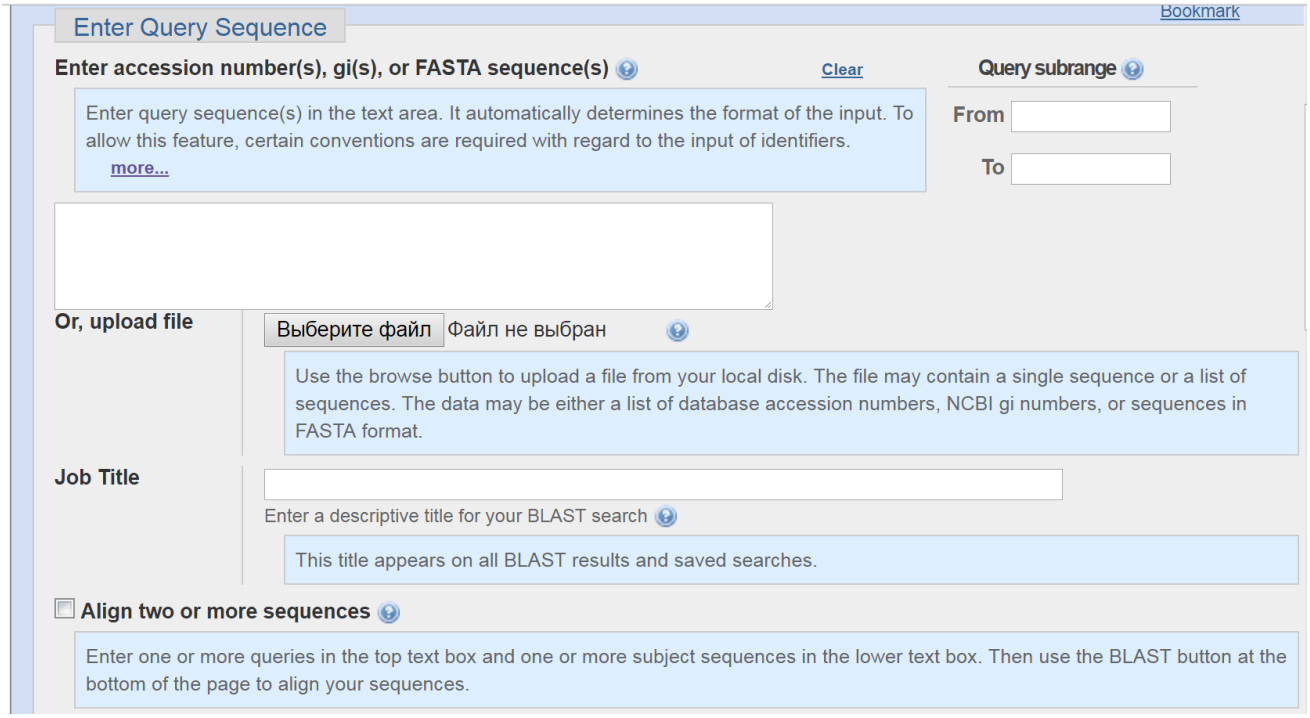


Query subrange

Піддіапазон запиту

Enter coordinates for a subrange of the query sequence. The BLAST search will apply only to the residues in the range. Sequence coordinates are from 1 to the sequence length. The range includes the residue at the **To** coordinate.

Ввести координати послідовності запиту. Координати послідовності від 1 до довжини послідовності. BLAST працює тільки з вказаним діапазоном.



Enter Query Sequence

Введення послідовності запиту

Enter accession number(s), gi(s), or FASTA sequence(s)

Enter query sequence(s) in the text area. It automatically determines the format of the input. To allow this feature, certain conventions are required with regard to the input of identifiers.

Ввести номер доступу, унікальний номер послідовності в БД або послідовність у FASTA форматі. Ввести послідовність запрос в поле для тексту. Формат введених даних визначається автоматично. Щоб функція виконувалась, точні правила вимагаються відносно ідентифікаторів.

Or, upload file

Або завантажте файл

Use the browse button to upload a file from your local disk. The file may contain a single sequence or a list of sequences. The data may be either a list of database accession numbers, NCBI gi numbers, or sequences in FASTA format.

Використовуйте кнопку перегляду для завантаження файлу з Вашого локального диску. Файл може містити одну послідовність або ж список послідовностей. Дані можуть також являти собою список номерів доступу у базі даних, унікальний номер послідовності в БД NCBI, чи послідовності у FASTA форматі.

Job Title

Назва роботи

Enter a descriptive title for your BLAST search

This title appears on all BLAST results and saved searches.

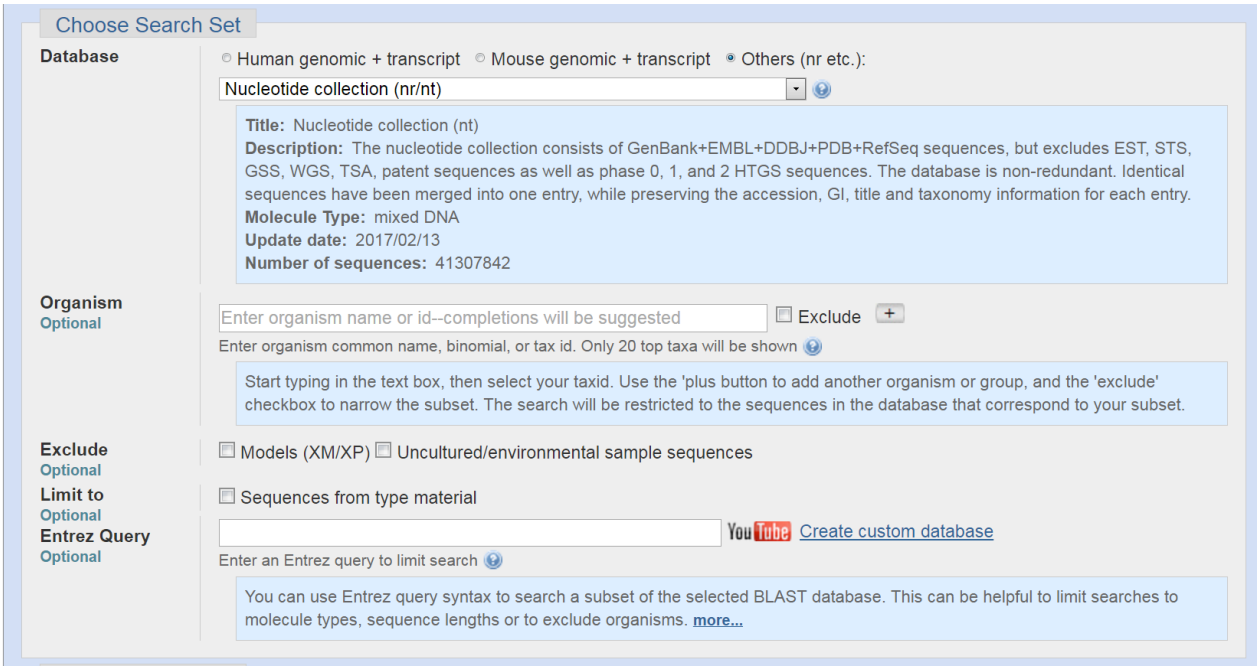
Введіть відповідну назву для Вашого пошуку у BLAST. Ця назва з’явиться на всіх результатах BLAST і збережених пошуках.

Align two or more sequences

Вирівнювання двох або більше послідовностей

Enter one or more queries in the top text box and one or more subject sequences in the lower text box. Then use the BLAST button at the bottom of the page to align your sequences.

Введіть одну чи більше послідовностей запитів в верхню текстову комірку і одну чи більше послідовність з БД в нижню текстову комірку. Потім використайте кнопку BLAST внизу сторінки для вирівнювання Ваших послідовностей.



Choose Search Set

Виберіть параметри пошуку

Organism (optional)

Організм (не обов’язково)

Start typing in the text box, then select your taxid. Use the "plus" button to add another organism or group, and the "exclude" checkbox to narrow the subset. The search will be restricted to the sequences in the database that correspond to your subset.

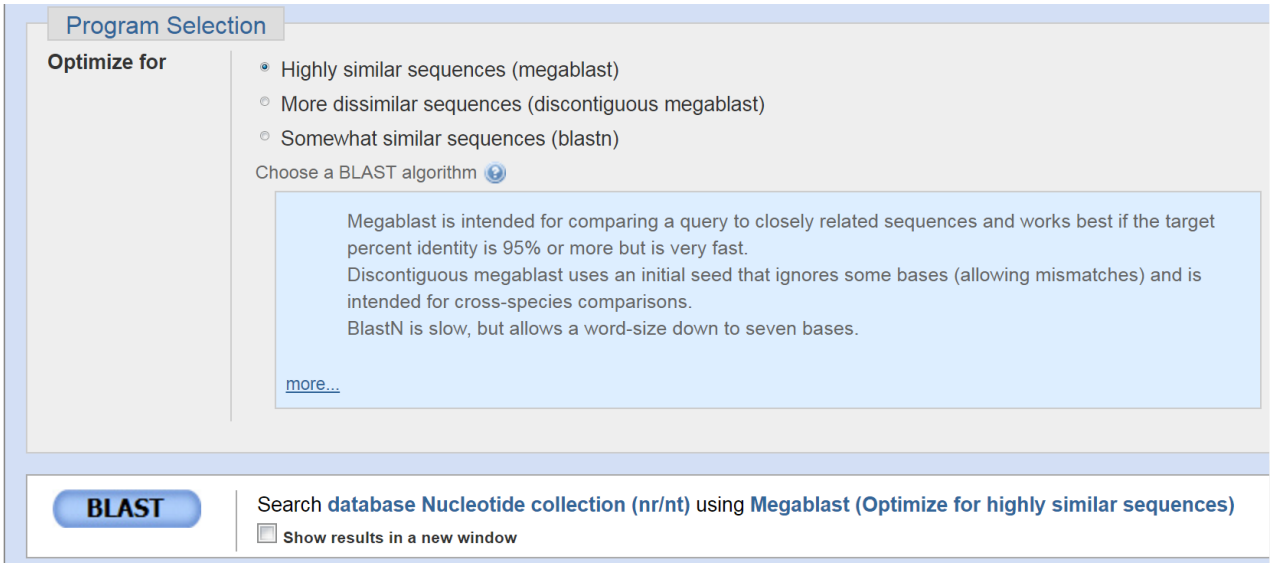
Почніть вводити текст в текстовому полі, а потім виберіть taxid. За допомогою кнопки "плюс", щоб додати ще один організм або групу, а також кнопку-прапорець "виключити", щоб обмежити підмножини. Пошук буде обмежений послідовностями в базі даних, які відповідають вашій підгрупі.

Entrez Query (optional)

Послідовність запит (не обов’язково)

You can use Entrez query syntax to search a subset of the selected BLAST database. This can be helpful to limit searches to molecule types, sequence lengths or to exclude organisms.

Ви можете використовувати **Entrez** (пошукова система) для пошуку підмножини вибраних в БД BLAST-ом. Це може бути корисно для обмеження пошуків до молекулярних типів, довжин послідовностей, або щоб виключити організми.



Program Selection

Вибір програми

Optimize for
Оптимізація

Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast.

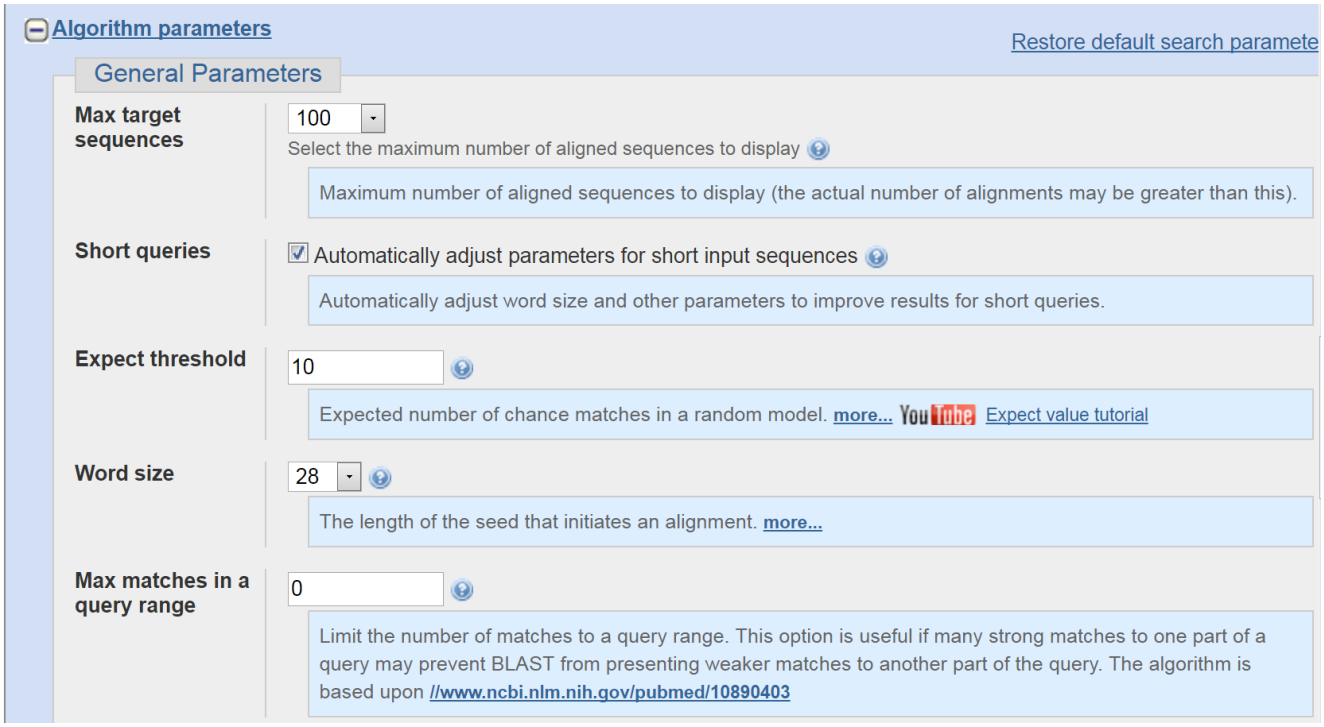
Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.

BlastN is slow, but allows a word-size down to seven bases.

Megablast призначений для порівняння послідовності та тісно пов'язаних послідовностей і працює краще, якщо відсоток ідентичності становить 95% або більше.

При неспівпадіннях, Megablast використовує початкове значення, яке ігнорує деякі основи (з урахуванням неспівпадінь) і призначений для міжвидових порівнянь.

Не дивлячись на те, що BLASTN повільний він дозволяє використовувати довжину слова до семи.



Algorithm parameters

Параметри алгоритму

Max target sequences

Максимальна кількість цільових послідовностей

Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Максимальна кількість вирівняних послідовностей для відображення (фактична кількість вирівнювань може бути більше, ніж ця).

Short queries

Короткі запити

Automatically adjust word size and other parameters to improve results for short queries.

Автоматично встановлюється розмір слова та інші параметри для поліпшення результатів для коротких запитів.

Expect threshold

Очікуваний поріг

Expected number of chance matches in a random model.

Ймовірнісна кількість збігів у випадковій моделі.

Word size

Розмір слова

The length of the seed that initiates an alignment.

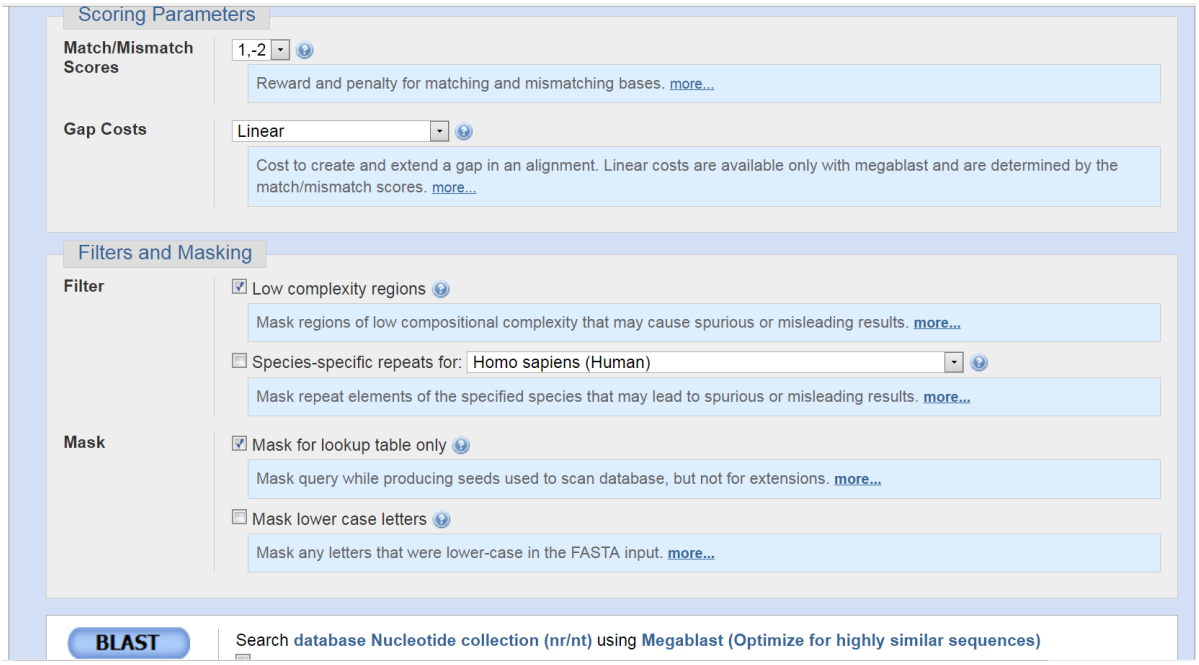
Довжина слова, яке ініціює вирівнювання.

Max matches in a query range

Максимальна кількість співпадінь в діапазоні запиту

Limit the number of matches to a query range. This option is useful if many strong matches to one part of a query may prevent BLAST from presenting weaker matches to another part of the query.

Обмежте кількість співпадінь до ряду запитів.



Scoring Parameters

Вагові параметри

Match/Mismatch Scores

Вага співпадіння / неспівпадіння

Reward and penalty for matching and mismatching bases.

Премія і штраф за співпадіння і неспівпадіння.

Gap Costs

Вага пробілів

Cost to create and extend a gap in an alignment. Linear costs are available only with megablast and are determined by the match/mismatch scores.

Вага появи і подовження

пробілів в вирівнюванні. Лінійна вага доступна тільки з megablast і визначається оцінками за співпадіння / неспівпадіння.

Filters and Masking

Фільтри і Маскування

Filter

Фільтри

Mask regions of low compositional complexity that may cause spurious or misleading results.

Маска області низької композиційної складності, які можуть викликати помилкові результати.

Mask repeat elements of the specified species that may lead to spurious or misleading results.

Маска повторюваних елементів зазначених видів, які можуть привести до помилкових результатів.

Mask

Mask any letters that were lower-case in the FASTA input.

Маскування будь-яких літер, які були в нижньому регістрі на вході FASTA.

Список використаної літератури

1. Електронний ресурс: http://www.ncbi.nlm.nih.gov/
2. Bioinformatics: Sequence, structure and database – Oxford University Press, 2001.
3. Азимов А. Генетический код. От теории эволюции до расшифровки ДНК. — М.: Центрполиграф, 2006. — 208 с — ISBN 5-9524-2230-6.
4. Ратнер В. А. Генетический код как система — Соросовский образовательный журнал, 2000, 6, № 3, с.17-22.
5. S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. Journal of Molecular Biology, 215(3):403–410, October 5, 1990.
6. Chang, S. R. and J. L. Kirschvink. «Magnetofossils, the magnetization of sediments, and the evolution of magnetite biomineralization», 1989. Annual Review of Earth and Planetary Sciences 17: 169—195.
7. Dobson J.P., Fuller M., Moser S., Wieser H.G., Dunn J.R. and Zoeger J. vocation of epileptiform activity by weak D.C. magnetic fields and iron biomineralization in the human brain. In: Biomagnetism: Fundamental Research and Applications, eds. C Baumgartner, L Deecke, G Stroink, SJ Williamson. Elsevier, Amsterdam: – 1995. – P. 16–19.
8. Dobson J.P. and Grassi P. Magnetic Properties of Human Hippocampal Tissue - Evaluation of Artefact and Contamination Sources. Brain Res. Bull., – 1996. – 39. – P. 255–259.
9. Joseph L Kirschvink*, Michael M Walker† and Carol E Diebe. Magnetite-based magnetoreception // Sensory systems. - Current Opinion in Neurobiology 2001, 11:462–467
10. Bazylinski, Dennis, Controlled biomineralisation of magnetic minerals by magnetotactic bacteria // Chemical Geology Elsevier. – 1995.
11. Zhou K, Zhang W.Y., Yu-Zhang K., Pan H.M., Zhang S.D., Zhang W.J., Yue H.D., Li Y, Xiao T, Wu L.F. A novel genus of multicellular magnetotactic prokaryotes from the Yellow Sea // Environ Microbiol. 2012 Feb;14(2):405-13
12. Cat Faber, Living Lodestones: Magnetotactic bacteria, Strange Horizons, 2001
13. Ulysses Lins, Marcos Farina. Phosphorus-rich granules in uncultured magnetotactic bacteria // FEMS Microbiology Letters. Volume 172, Issue 1, 1 March 1999, Pages 23–28
14. Schüler, Dirk, «The biomineralization of magnetosomes in Magnetospirillum gryphiswaldense», 2002 Int. Microbiology
15. Bazylinski, Dennis, «Controlled biomineralization of magnetic minerals by magnetotactic bacteria», 1995 Chemical Geology Elsevier
16. Горобець, С. В. Основи біоінформатики [Електронний ресурс]: підручник для студентів напряму підготовки 6.051401 «Промислова біотехнологія» факультету біотехнології і біотехніки / С. В. Горобець, О. Ю. Горобець, Т. А. Хоменко ; НТУУ «КПІ». - Електронні текстові дані (1 файл: 2,72 Мбайт). – Київ : НТУУ «КПІ», 2010

17.Горобець С.В., Горобець О.Ю., Булаєвської М.О. «Біоінформатичні бази даних» електронний навчальний посібник . – Київ : НТУУ «КПІ», 2020р.