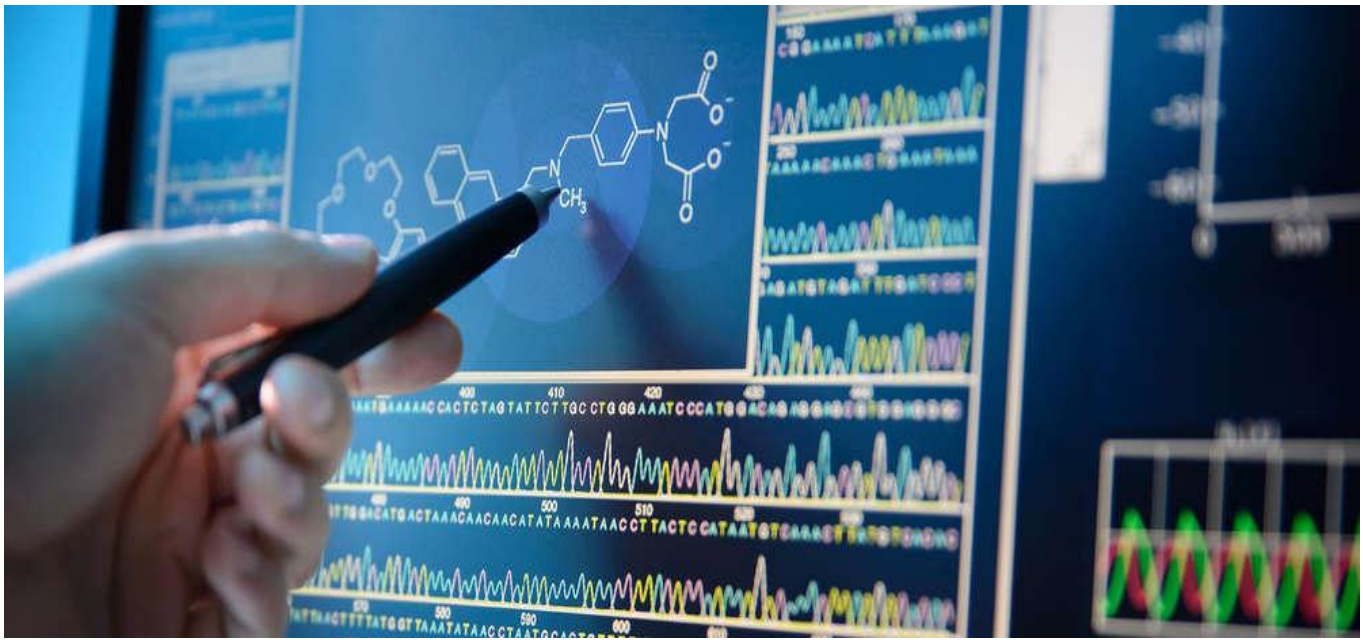
КУЧМЕНКО О. Б., ПЕРЕХОДЬКО К. М.

БІОІНФОРМАТИКА

Навчально-методичний посібник



Ніжинський державний університет імені Миколи Гоголя
Навчально-науковий інститут природничо-математичних,
медико-біологічних наук та інформаційних технологій
Кафедра біології

КУЧМЕНКО О. Б., ПЕРЕХОДЬКО К. М.

БІОІНФОРМАТИКА

Навчально-методичний посібник

Ніжин – 2023

УДК 575.112(075.8)
К95

Рекомендовано Вченою радою
Ніжинського державного університету імені Миколи Гоголя
(НДУ ім. М. Гоголя)
Протокол № 11 від 01.06.2023 р.

Рецензенти:

І. В. Калінін – доктор біологічних наук, професор кафедри біохімії і фізіології тварин ім. акад. М.Ф. Гулого Національного університету біоресурсів і природокористування України;

В. І. Шейко – доктор біологічних наук, професор кафедри біології Ніжинського державного університету імені Миколи Гоголя

Кучменко О. Б., Переходько К. М.

К95 Біоінформатика: навчально-методичний посібник. Ніжин: НДУ ім. М. Гоголя, 2023. 60 с.

ISBN 978-617-527-284-8

Навчально-методичний посібник включає в себе лабораторні роботи з біоінформатики для студентів та аспірантів спеціальності 091 Біологія та біохімія вищих навчальних закладів.

ISBN 978-617-527-284-8

ЗМІСТ

ВСТУП

Навчальна дисципліна «Біоінформатика» є однією із важливих біологічних складових навчального процесу, що відіграє значну роль у підготовці фахівців-біологів. В рамах навчальної дисципліни заплановано лабораторний практикум, який складається з 7 робіт.

Мета навчальної дисципліни «Біоінформатика»: навчити студентів орієнтуватися в сучасних концепціях біоінформатики, дати цілісне уявлення про структуру та методи аналізу біологічних послідовностей, структуру та методи аналізу просторових структур біологічних молекул, структуру та методи комп'ютерного аналізу геномів, сформувати у студентів цілісний і системний погляд на організацію біологічної інформації на молекулярному рівні.

Предметом навчальної дисципліни «Біоінформатика» є загальні закономірності організації і аналізу інформації, що міститься в біомолекулярних системах.

Факторами, що стимулювали розвиток біоінформатики, стала необхідність ефективного опрацювання величезних масивів інформації, отриманих експериментальними методами молекулярної біології і генетики, та розвитку інформаційних технологій, в тому числі ефективних методів розшифровки нуклеотидних послідовностей ДНК. Швидкий розвиток молекулярної біології та генетики став можливим завдяки досягненням у сумісних областях науки: розвитку інформаційних технологій, розробці нових підходів в хімії, появі потужних фізичних засобів дослідження внутрішньої будови речовини (рентгеноструктурного аналізу, електронної та атомної силової мікроскопії, методів, які базуються на використанні явища ядерного магнітного резонансу).

Датою виділення біоінформатики в окрему наукову область вважається 1980 рік, відколи перший номер журналу Nucleic Acids Research був повністю присвячений опису біоінформаційних баз даних та комп'ютерним методам аналізу нуклеотидних і амінокислотних послідовностей.

Нині, вже частково, чи повністю просеквеновано еукаріотичні геноми біля 2000 видів живих істот, серед них геноми людини, шимпанзе, миші, пацюка, кішки, собаки, курки, риби, дрозофіли, малярійного комара, черв'яків і дріжджів, сотні геномів рослин та інших організмів. Окрім того, просеквеновано тисячі бактеріальних геномів та сотні геномів архей. Встановлено, що величина генома людини сягає більше ніж 3 мільярдів пар нуклеотидів (понад 30 тисяч генів), що з урахуванням інформації, яка одержана при його розшифруванні, становить більше десяти терробайт даних.

Біоінформатика – галузь науки, що розробляє і застосовує технології інформатики для аналізу, систематизації молекулярно-біологічних даних, використовує фізико-математичні методи для моделювання процесів, що відбуваються на молекулярному рівні з метою виявлення структур, функцій та взаємодії макромолекул (ДНК, РНК, білків) з подальшим використанням цих знань при створенні нових лікарських препаратів та нановиробів для діагностики і лікування, а також отримання організмів з наперед заданими властивостями.

В багатьох випадках біоінформаційний аналіз геномних даних дозволяє отримати нові, нетривіальні висновки, тобто нові знання, які потім можуть бути перевірені експериментально. Тому біоінформатика є новим інструментом в біології, що стоїть в одному ряду з фізичними (рентгеноструктурний аналіз, електронна та скануюча мікроскопії, мас-спектрометрія, ядерний магнітний резонанс та ін.) та біохімічними методами досліджень.

Основним принципом біоінформатики можна вважати те, що молекули нуклеїнових кислот та білків описуються за допомогою послідовності алфавітних знаків, причому число цих знаків обмежене (4 – для нуклеїнових кислот і 20 для опису білків). З точки зору лише інформатики, геном або білок – це довгий текст, що містить, наприклад, мільярд букв. Біоінформатика ж відрізняється тим, що її задачею є аналіз змісту цього тексту.

Найголовніше призначення біоінформатики – це розширення і поглиблення розуміння людиною біологічних процесів. Традиційна біологія з її

підходами продовжують залишатися важливими, але для більш глибокого розуміння біологічних процесів необхідно посилити кількісну сторону молекулярної біології з акцентуванням уваги на поведінці різних систем організму і особливостей взаємовідношень між ними, що дозволить зрозуміти основи функціонування всього організму.

Вимоги до знань та вмінь.

Знати: основні концепції аналізу біологічних текстів, основні концепції відтворення і аналізу просторової організації біомолекул, основи організації цілих геномів та методи їх порівняльного аналізу.

Вміти: аналізувати та порівнювати біологічні тексти, працювати з банками даних біологічних послідовностей і просторових структур, здійснювати парне та множинне вирівнювання послідовностей, реконструювати просторову структуру, розраховувати поведінку і аналізувати особливості просторової структури білків, визначати білок-кодуючи ділянки в нуклеотидних послідовностях, проводити філогенетичний аналіз, вміти цілісно і системно мислити.

ЛАБОРАТОРНА РОБОТА № 1

Вирівнювання амінокислотних послідовностей (BLAST)

Мета роботи – здійснити порівняння нуклеотидних та білкових послідовностей з наявними в базах даних NCBI, провести парні вирівнювання послідовностей та інтерпретувати отримані результати.

Теоретичні відомості

Вирівнювання амінокислотних послідовностей – біоінформатичний метод, заснований на розміщенні двох або більше послідовностей один під одним так, щоб поєднати подібні ділянки в цих послідовностях. При цьому між амінокислотами послідовностей, що вирівнюються, допустима вставка символу "геп", що дозволяє розташувати гомологічні елементи один під одним.

BLAST (Basic Local Alignment Search Tool) дозволяє швидко порівняти послідовності запиту з базами даних послідовностей. Є фундаментальним для розуміння кревності будь-якої запитуваної послідовності та інших відомих білків або ДНК послідовностей. (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Алгоритм BLAST швидкий, точний та web-доступний.

Програми серії BLAST включають:

1. Нуклеотидні – порівняння нуклеотидної послідовності запиту з базою даних секвенованих нуклеїнових кислот та їх ділянок.
2. Білкові – порівняння амінокислотної послідовності запиту з базою даних білків та їх ділянок.
3. Транслюючі – здатні транслювати нуклеотидні послідовності в амінокислотні.
4. Геномні – призначені для порівняння нуклеотидної послідовності, що вивчається, з базою даних секвенованого геному будь-якого організму.
5. Спеціальні – прикладні програми, що використовують BLAST.

Застосування:

* Визначення ортологів і паралогів

* Виявлення нових генів або білків

* Виявлення варіантів генів або протеїнів

* дослідження expressed sequence tags (ESTs)

* аналіз структури та функції білків.

Хід роботи

1. Вибір послідовності (запиту). Послідовність може бути введена у форматі FASTA або як унікальний номер.

2. Вибір програми BLAST.

* вибрати організм для пошуку

* вибрати фільтрацію on/off

* змінити матрицю замін

* змінити expect(e) value

* змінити word size

* змінити формат виведення даних

3. Вибір бази даних для пошуку, клацнувши на вікно Database

nr = non-redundant (most general database)

dbest = database of expressed sequence tags

dbsts = database of sequence tag sites

gss = genomic survey sequences

4. Вибір додаткових параметрів (додаткові параметри пошуку можна змінити, натиснувши на Algorithm parameters, коментарі по кожній з опцій на http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#wordsize).

Для прикладу, наш вихідний варіант буде білок, виділений з Brachyopsis rostratus (це кістка риба).

Варіанти вихідних даних пошуку BLAST: графічний вигляд, табличний вигляд, вирівнювання, таксономія (об’єднує види із збігами). Поруч із таксономією можна вибрати варіант результатів у вигляді дерева, причому вивести дерево можна в різних видах (rectangle, slanted, radial, force). Клікнувши в Other reports: Search Summary можна отримати таблицю з параметрами пошуку та змінними формулами, що описує статистику.

Програми серії BLAST виробляють локальні вирівнювання, що пов’язано з наявністю в різних білках подібних доменів та патернів. Крім цього, локальне вирівнювання дозволяє порівняти іРНК з геномною ДНК. У разі глобального вирівнювання виявляється менша схожість послідовностей, особливо їх доменів та патернів.

Алгоритм програми BLAST заснований на припущенні про те, що вирівнювання з високим рахунком, ймовірно, містять короткі відрізки ідентичних або майже ідентичних знаків. Ці короткі відрізки називаються словами. «Центральна ідея алгоритму BLAST-обмежити увагу сегментами пар, які містять пару слів завдовжки w з оцінкою, принаймні T.» (Altschul et al. (1990)).

Парне вирівнювання – це вирівнювання двох амінокислотних послідовностей один щодо одного. Воно будується з тією самою метою, як і множинне вирівнювання – тобто з парному вирівнюванню намагаються оцінити гомологію даних послідовностей. Однак робота з парним вирівнюванням завжди передбачає орієнтацію на множинне вирівнювання білків цього сімейства, це потрібно для більш точної оцінки консервативності (рис. 1.1).

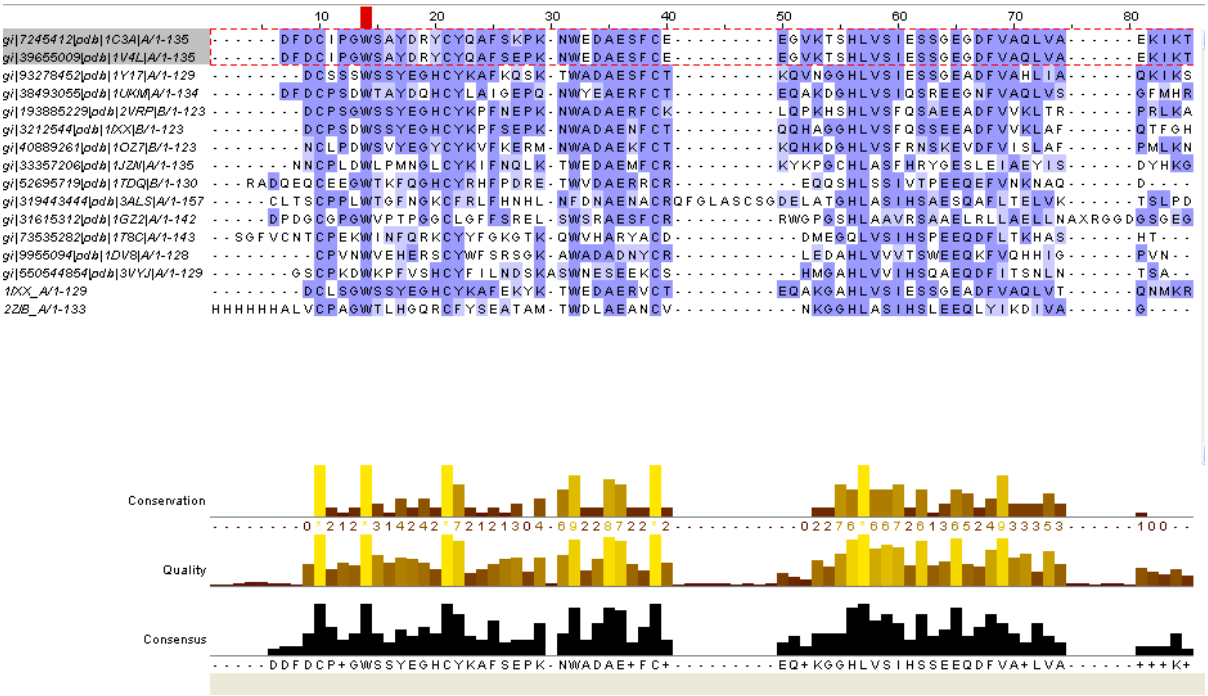


Рис. 1.1. ***Множинне вирівнювання білків***

Для того, щоб наочніше побачити родинні зв’язки між білками, будемо самостійно будувати дерево множинного вирівнювання (рис. 1.2). По ньому видно, що послідовності, які вирівнюються, не дуже гомологічні один одному (принаймні порівняно з рештою). 1IXX – це білок з отрути змії хабу, який пов’язується з факторами згортання крові IX/X. 2ZIB – це Ca-незалежний білок-антифриз, виділений з Brachyopsis rostratus (це кісткова риба).

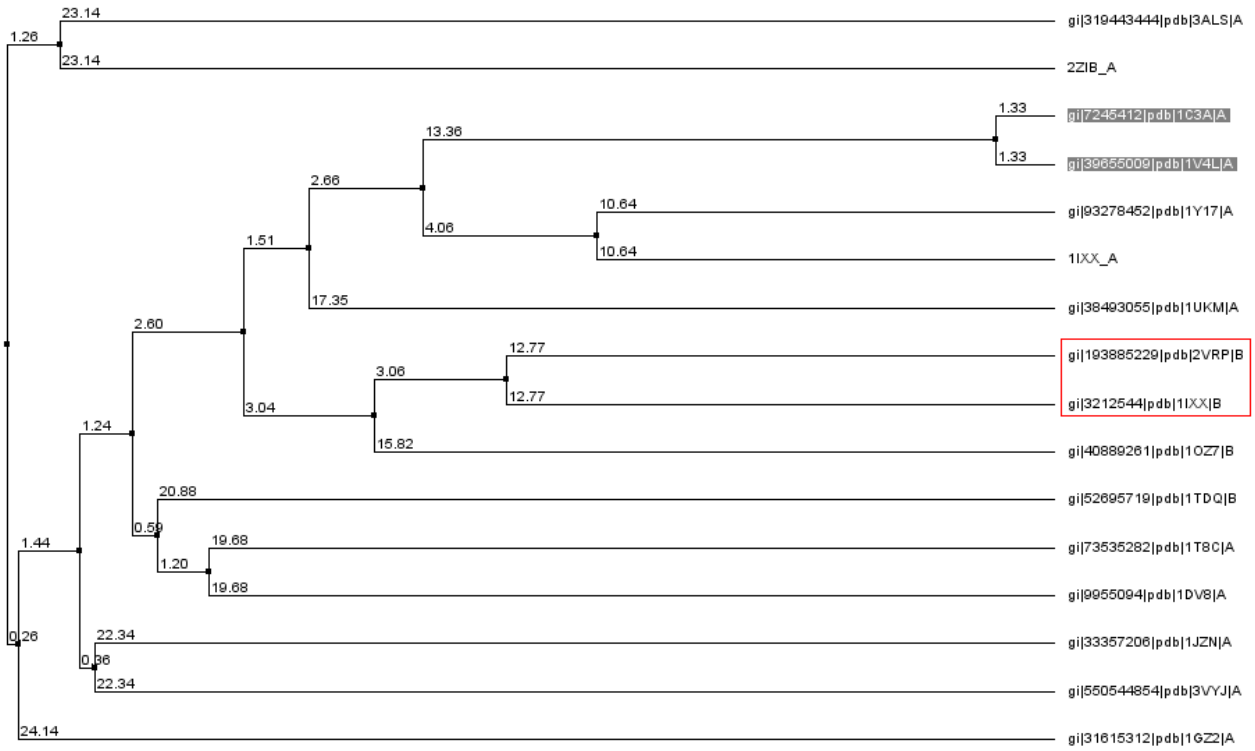


Рис. 1.2. ***Дерево множини вирівнювання***

Вибрані для парного вирівнювання послідовності обведені червоним прямокутником.

За завданням необхідно побудувати вручну вирівнювання двох останніх послідовностей із множинного вирівнювання щодо один одного. Результат дивимось на спектрограммі-рисунку:

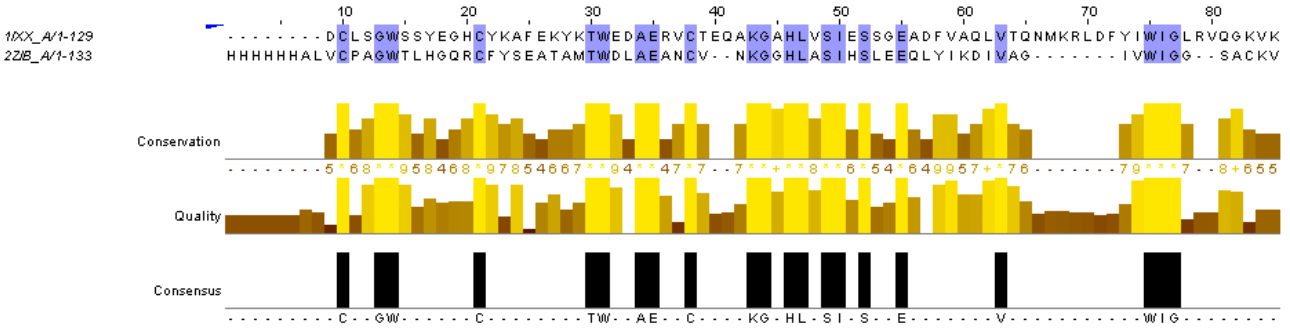


Рис. 1.3. ***Парне вирівнювання послідовностей 1IXX_A та 2ZIB_A, побудоване за допомогою засобів Jalview***

Для виконання наступного завдання скористаємося командою needle на kodomo. Програма needle використовує алгоритм Нідлмана-Вунша для побудови глобального вирівнювання двох послідовностей. Глобальне вирівнювання має на увазі гомологію послідовностей по всій довжині, туди включаються обидві послідовності цілком. Спочатку отримаємо парне вирівнювання двох останніх послідовностей у форматі за замовчуванням, його можна завантажити https://kodomo.fbb.msu.ru/~partyhard/term2/pr10/1ixx_a_1-129.needle

Потім, дописавши опцію -aformat3 fasta, отримуємо output файл у форматі fasta, з ним можна ознайомитися https://kodomo.fbb.msu.ru/~partyhard/term2/-pr10/alignment2.fasta

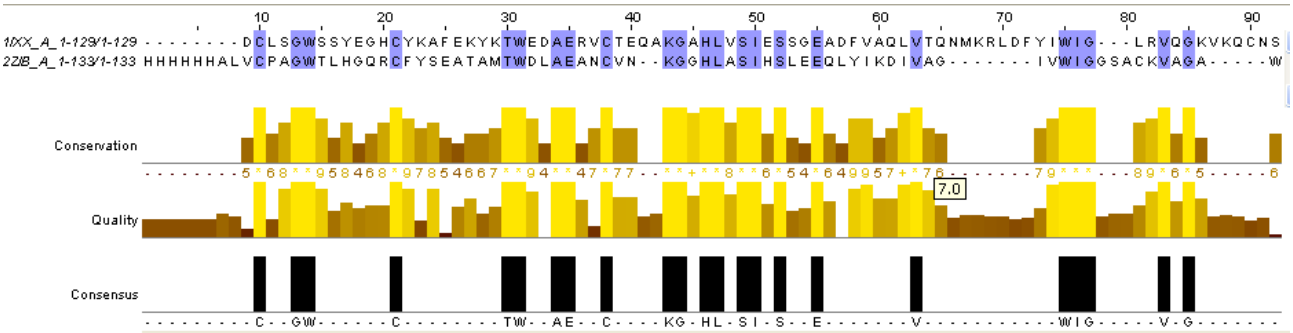


Рис. 1.4. ***Вирівнювання у форматі fasta я візуалізувала у JalView, розмальовка BLOSUM62***

Парне вирівнювання двох останніх послідовностей (1IXX_A і 2ZIB_A) з множинного вирівнювання, побудоване за допомогою програми needle на kodomo. Розмальовка BLOSUM62.

Далі спробуємо побудувати найкраще локальне вирівнювання двох останніх послідовностей (1IXX_A і 2ZIB_B) у форматі fasta, використовуючи команду water на kodomo. Локальне вирівнювання будується, якщо у послідовностях є негомологічні ділянки; тоді вони виключаються і вирівнювання йде між гомологічними ділянками. Для отримання локального вирівнювання використовують алгоритм Сміт-Ватермана. Результат у форматі fast можна побачити https://kodomo.fbb.msu.ru/~partyhard/term2/pr10/-local_align.fasta

У редакторі Jalview це вирівнювання виглядає так:

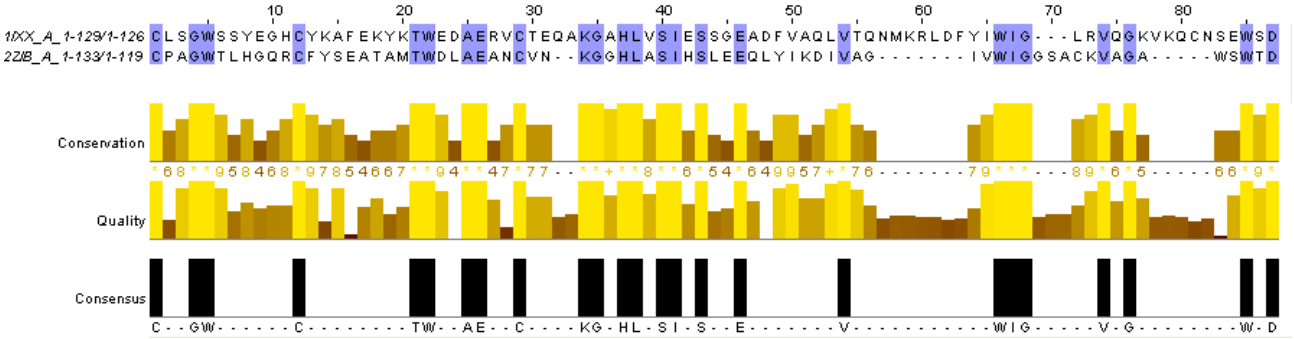


Рис. 1.5. ***Найкраще локальне вирівнювання двох послідовностей (1IXX_A та 2ZIB_B), відкрите в редакторі Jalview. Забарвлення BLOSUM62***

Для виконання фінального завдання потрібно взяти дві послідовності: 1IXX_A (з нею ж велася робота в завданнях вище) і 3GL3_A. Послідовності свідомо негомологічні: 1IXX_A – це зміїна отрута, а 3GL3_A – це прокаріотичний білок дисульфідного обміну.

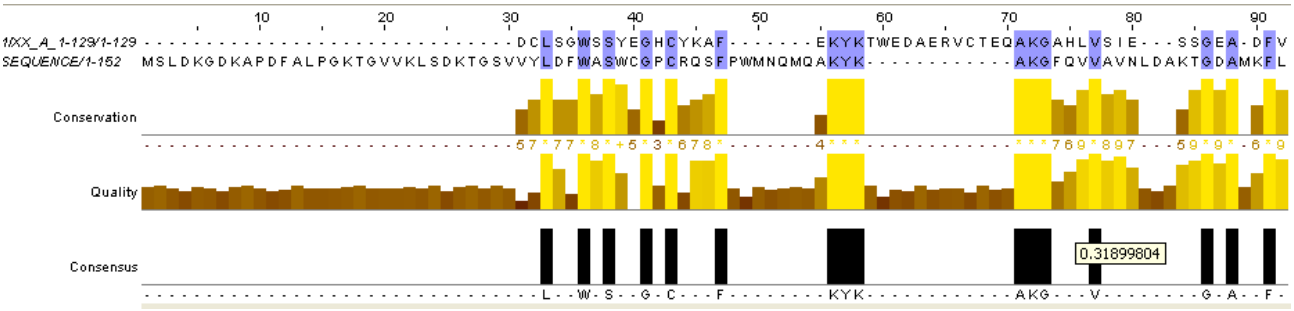


Рис. 1.5.1. ***Результати глобального та локального вирівнювань***

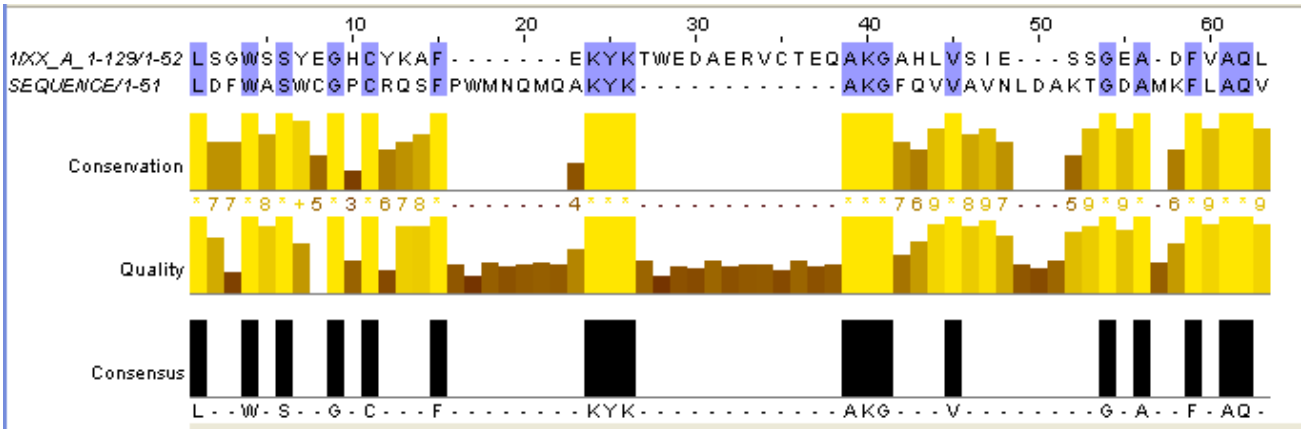


Рис. 1.6. ***Вирівнювання свідомо негомологічних послідовностей білків з ідентифікаторами pdb 3GL3 і 1IXX***

На верхньому рисунку – глобальне вирівнювання, на нижньому – локальне. Як видно, вирівнювання вручну дає менш точні результати, ніж глобальне і локальне вирівнювання з допомогою алгоритмів. Порівнювати ж локальне і глобальне вирівнювання не можна, але слід зазначити, що у разі гомологічних послідовностей довжина глобального і локального вирівнювань відрізняються не набагато, тоді як для свідомо негомологічних послідовностей довжина локального вирівнювання дуже маленька. У вирівнюваннях свідомо негомологічних послідовностей спостерігається помітно більше гепів, за допомогою них програма намагалася знайти хоч якусь схожість. Число і відсоток збігів у світових вирівнюваннях гомологічних і негомологічних послідовностей різняться не сильно за рахунок того, що програма вставила багато гепів у вирівнювання негомологічних послідовностей. У локальному вирівнюванні негомологічних послідовностей спостерігається найбільший

відсоток подібних залишків, але з цього не можна дійти висновку про гомологію, так як це подібність досягається рахунок великого відсотка гепів і вона спостерігається на короткій ділянці проти глобального вирівнювання.

Виконуючи завдання парного вирівнювання, потрібно додати і глобальне вирівнювання, отримане за допомогою needle, до вирівнювання, яке вже отримали вручну. Відразу виявляється висока подібність на початковій ділянці вирівнювань, там нічого не довелося змінювати (рис. 1.7).

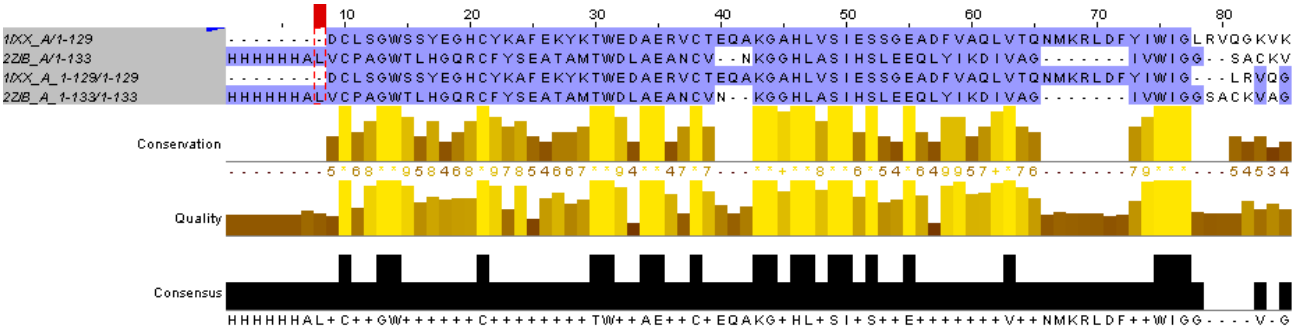


Рис. 1.7. ***Висока подібність двох парних вирівнювань тих самих послідовностей на початковій ділянці***

Однак на кінцевій ділянці вирівнювання поєднуватися не так добре. Щоб отримати суміщення, встановлюємо 3 гепа в обидві послідовності першого вирівнювання так, щоб це не вплинуло на вирівнювання. В результаті маємо добре поєднання (до – рис. 1.8, після – рис. 1.9).

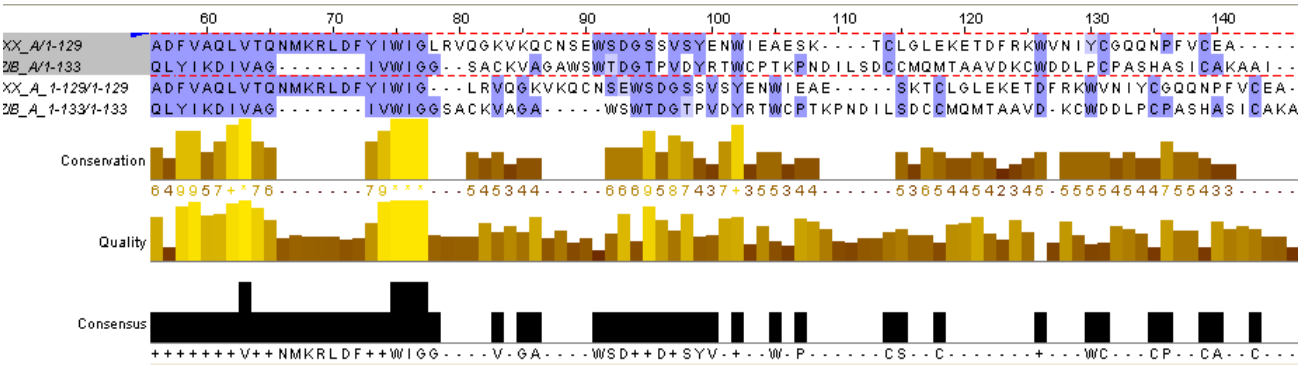


Рис. 1.8. ***Кінцева ділянка вирівнювань, на якій автоматичне вирівнювання дуже відрізняється від отриманого вручну***

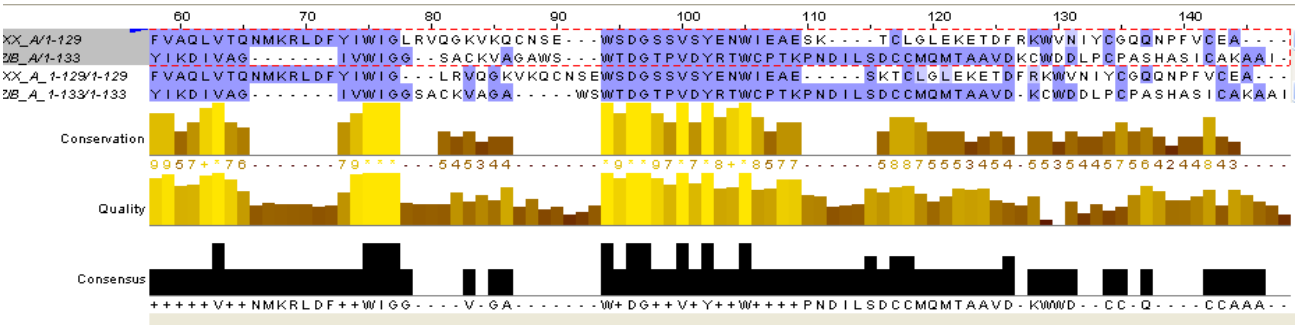


Рис. 1.9. Поєднання двох вирівнювань на кінцевих ділянках шляхом додавання гепів у перше вирівнювання

Перевірка на правильність вирівнювання відбувається за допомогою програми SupCheck, яка поєднує просторові структури двох білків. Це може допомогти знайти помилки у вирівнюванні (зазвичай вони є) та встановити гомологію між ділянками білка. Вікно Rasmol із поєднанням двох структур показано на рис. 1.10 – вирівнювання, отримане за допомогою needle, в цілому більше відповідає просторовому поєднанню, ніж вирівнювання за допомогою видалення гепів.

Рис. 1.10. ***Поєднання двох просторових структур білків 1IXX_A та 2ZIB_A у Rasmol***

Для кожного вирівнювання окремо переглядається його відповідність із поєднанням. Кількість помилок у вирівнюванні оцінюється вручну: 40 помилок I роду (залишки, розташовані в одній колонці вирівнювання, на структурі не

суміщені), 4 помилки II роду (залишки, що добре поєднуються в просторі, не розташовані в одній колонці вирівнювання). Кількість помилок при вирівнюванні needle: 26 помилок I роду, 3 помилки II роду. Насправді обидві ланцюга утворюють подібні в просторі структури на ділянці, виділеній червоним в Jalview, але ці структури розташовані під різним кутом, тому формально не поєднуються. Однак і на цій ділянці можна говорити про гомологію через дві причини: збігаються досить рідкісні амінокислоти і при накладенні незалежно від інших частин білків ці ділянки збіглися б. Всі ці дані надалі необхідні для роботи на платформі BLAST.

Контрольні запитання:

1. Що таке Алгоритм BLAST?
2. Що таке локальні вирівнювання в BLAST?
3. Що таке множинні вирівнювання?
4. Що таке парне вирівнювання?
5. Що таке глобальне вирівнювання? Різниця між локальним вирівнюванням?

ЛАБОРАТОРНА РОБОТА № 2

Пошук гомологів у білків

Мета роботи – засвоїти метод пошуку гомологів у різноманітних білків використовуючи програму PSI-BLAST.

Теоретичні відомості

Емпіричні дослідження амінокислотних замін дозволили з'ясувати, що в процесі еволюції амінокислотні заміни відбувалися не рівномірно. Амінокислоти частіше заміняються на подібні їм по фізико-хімічним властивостям, а саме: розміру, гідрофобності/гідрофільності, заряду, полярності та ін. Так, наприклад, такі амінокислоти, як гліцин, цистеїн та триптофан заміняються рідко. Тому, якщо відбулася амінокислотна заміна, це може майже ніяк не вплинути на структуру та функцію білку, а може значно їх змінити. Так, якщо лізин заміниться на лейцин, який суттєво відрізняється від лізину, то просторова структура білка, а відповідно, і його функція може суттєво змінитися. А заміна лізину на аргінін може не вплинути на просторову структуру та функцію білка.

Характер амінокислотних замін визначається ступенем їх консервативності або радикальності. Консервативною заміною амінокислоти називається мутаційна заміна, яка не призводить до суттєвих змін структури та функції білка. В процесі еволюції консервативні заміни амінокислот відбуваються частіше, ніж радикальні. Ці заміни переважно зустрічаються в функціонально важливих ділянках білкової молекули (наприклад, сайтах зв'язування лігандів). Радикальні заміни амінокислот, напроти, суттєво змінюють структуру та функції білка. Ці всі явища призводять до появи гомологічних білків.

Білки-гомологи – група білків з одного та/або різних організмів, гени яких з великим ступенем вірогідності мають загальне еволюційне походження.

Генетичні причини появи білків-гомологів можуть бути різними: дивергенція організмів (вертикальний перенос), дуплікація генів, горизонтальний перенос.

Сімейством білків-гомологів 15-20 років тому могла бути названа вся сукупність гомологічних між собою білків. Однак підвищення чутливості методів порівняння амінокислотних послідовностей і швидке накопичення даних про більш консервативні тривимірні структури білків, виявило еволюційну спорідненість між багатьма раніше відомими родинами. Термін "сімейство" став більш розмитим, і різні автори можуть його неоднаково трактувати. Часто приналежність білка до конкретного сімейства передбачає відому або передбачувану наявність у нього певної біологічної функції, по якій і дається назва сімейству. Принципово важливим є те, що білки одного сімейства утворюють монофілетичну групу, а рівень подібності їх амінокислотних послідовностей є достатнім для побудови глобального множинного вирівнювання.

Ще 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 один еволюційний домен. Наприклад, такими парними структурними доменами володіють глікозил-гідролази сімейств GH27 і GH32 (рис. 2.1).

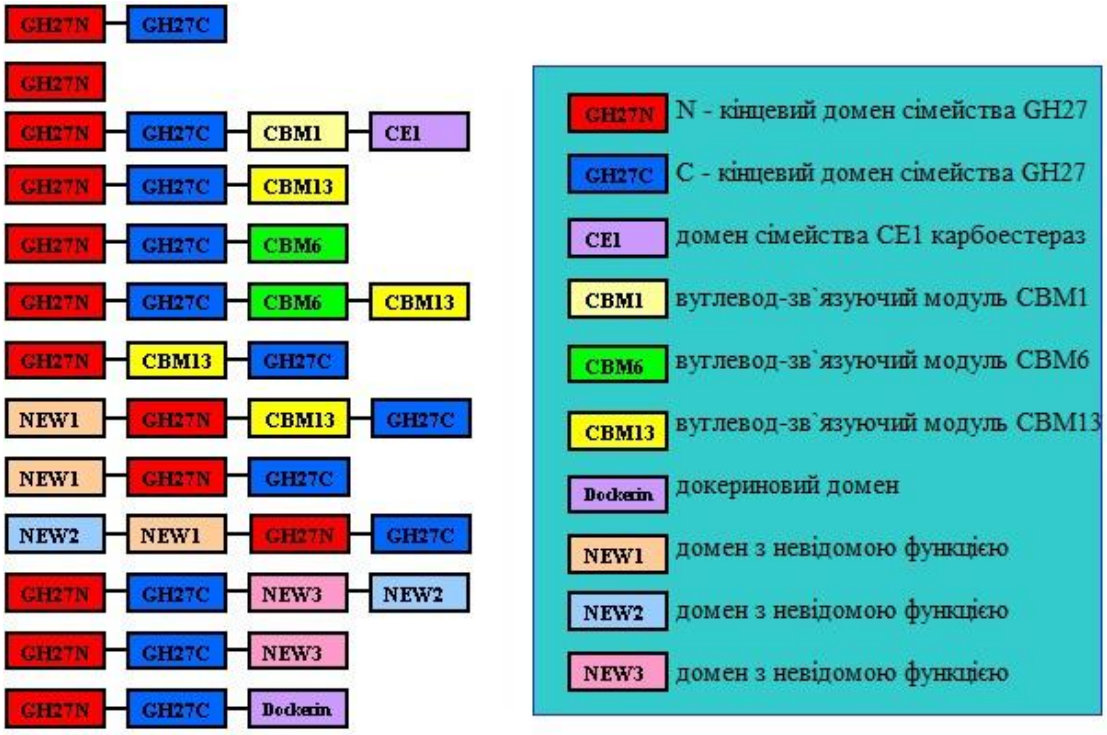


Рис. 2.1. ***Доменна структура білків сімейства GH27 глікозил-гідролаз***

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.

BLAST (basic local alignment search tool – основний (програмний) інструмент пошуку локальних вирівнювань), який майже завжди працює так, як цього вимагає "золотий стандарт". BLAST – сімейство комп’ютерних програм для пошуку гомологів білків або нуклеїнових кислот, для яких відома первинна структура (послідовність) або її фрагмент.

Мета роботи полягає в застосуванні програми BLAST для того, щоб **підібрати з бази даних (БД) кандидатів для порівняння** послідовностей.

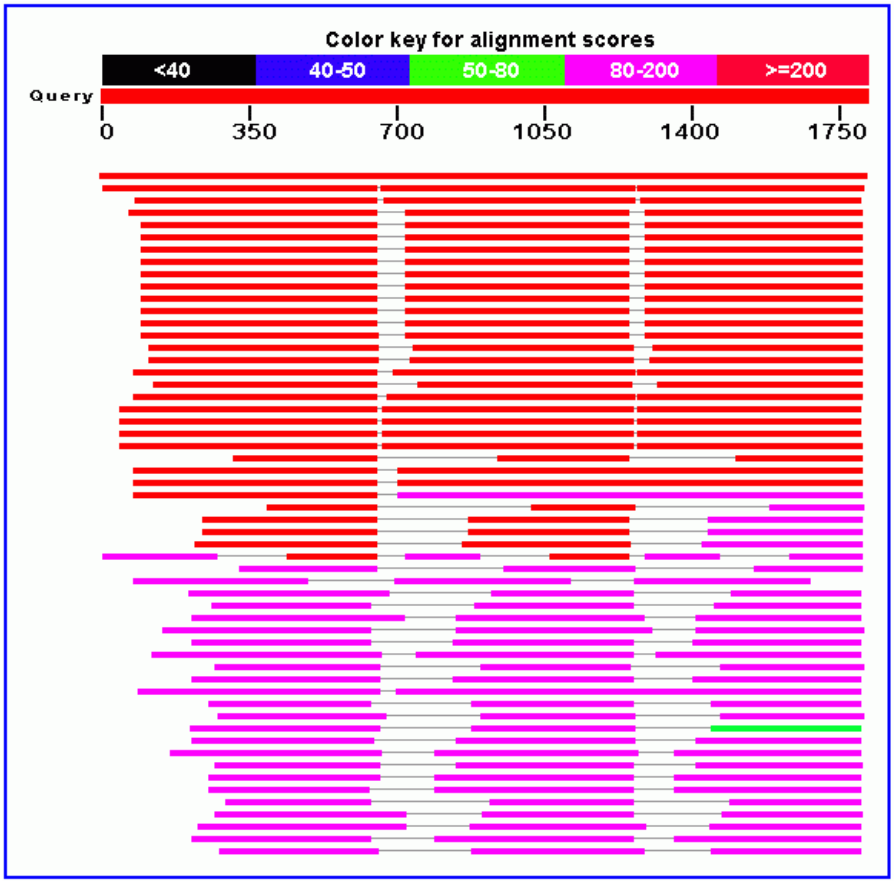


Рис. 2.2. ***Схема, що показує результат пошуку гомологів за допомогою програми PSI-BLAST. В якості запиту був обраний білок, що складається з трьох гомологічних між собою доменів***

Програма **BLAST** була розроблена вченими Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, і David J. Lipman в системі Національного інституту здоров'я США і була опублікована в журналі *Journal of Molecular Biology* в 1990 році.

Сімейство програм серії BLAST ділиться на наступні групи:

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.

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; *tblastx* – переводить досліджувану нуклеотидну послідовність у амінокислотну, а потім порівнює її з трансльованими послідовностями; *tblastn* – амінокислотна послідовність, що вивчається, порівнюється з трансльованими послідовностями з бази даних.

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

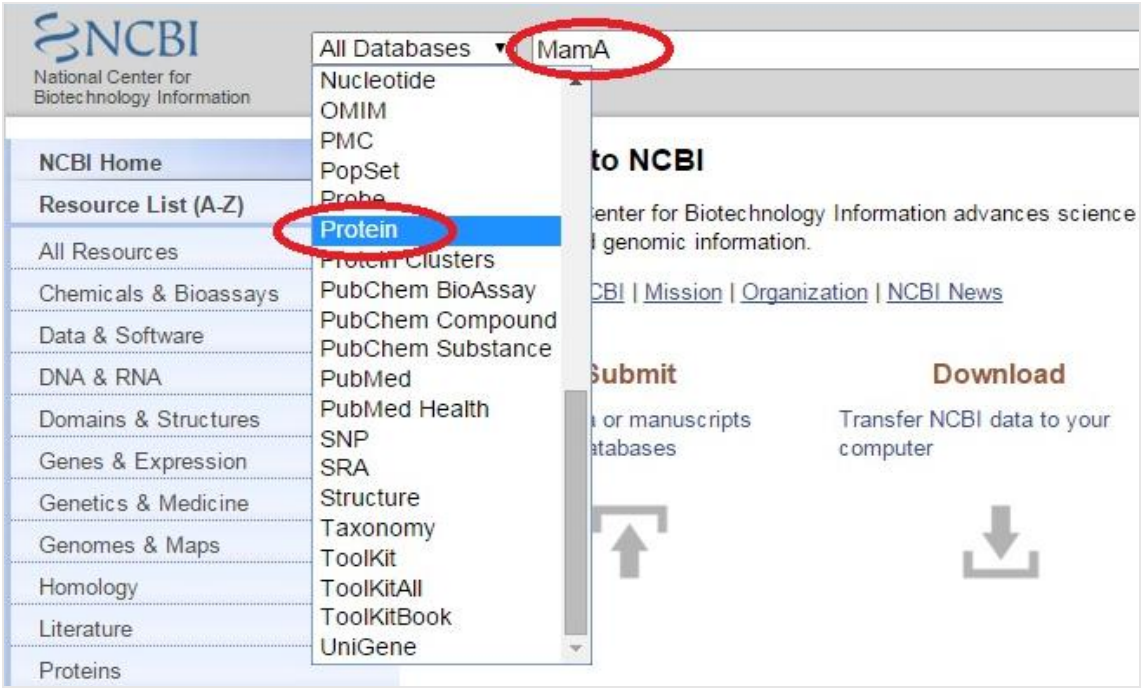
Білкові: призначені для порівняння досліджуваної амінокислотної послідовності білка з наявною базою даних білків і їх ділянок: *BlastP* – повільне порівняння з метою пошуку всіх подібних послідовностей; *PSI-BLAST (Position-Specific Iterated BLAST)* – порівняння білків з метою пошуку дальніх гомологів; *PHI-BLAST (Pattern Hit Initiated BLAST)* – пошук білків, що містять певні паттерни; *DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)* – побудова PSSM з використанням результатів бази даних консервативних доменів.

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.

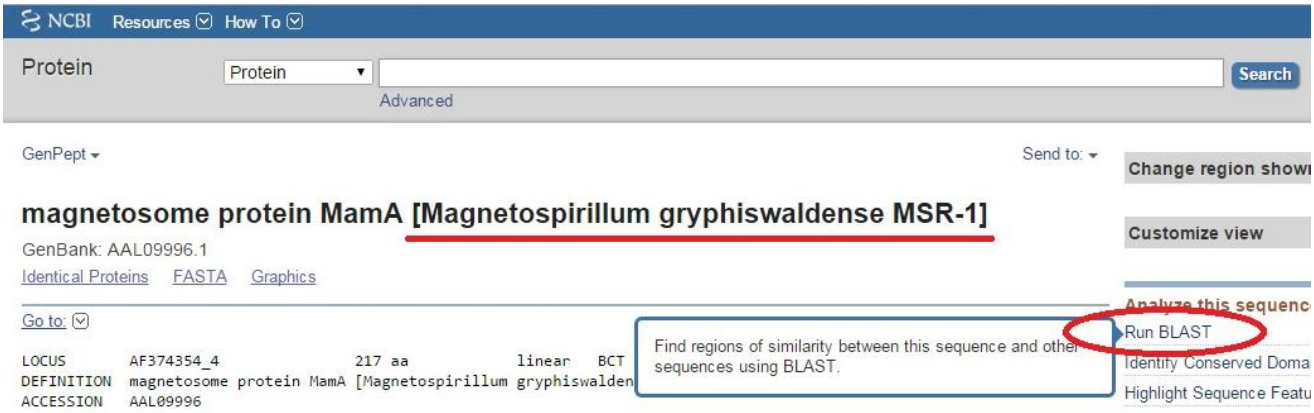
Хід роботи

1. Зайти на сайт **NCBI** (National Center for Biotechnology Information).
2. В верхньому лівому кутку вибрати **Protein**, ввести в поле пошуку необхідний білок (MamA, MamB, MamM, MamE, MamO, MamK).



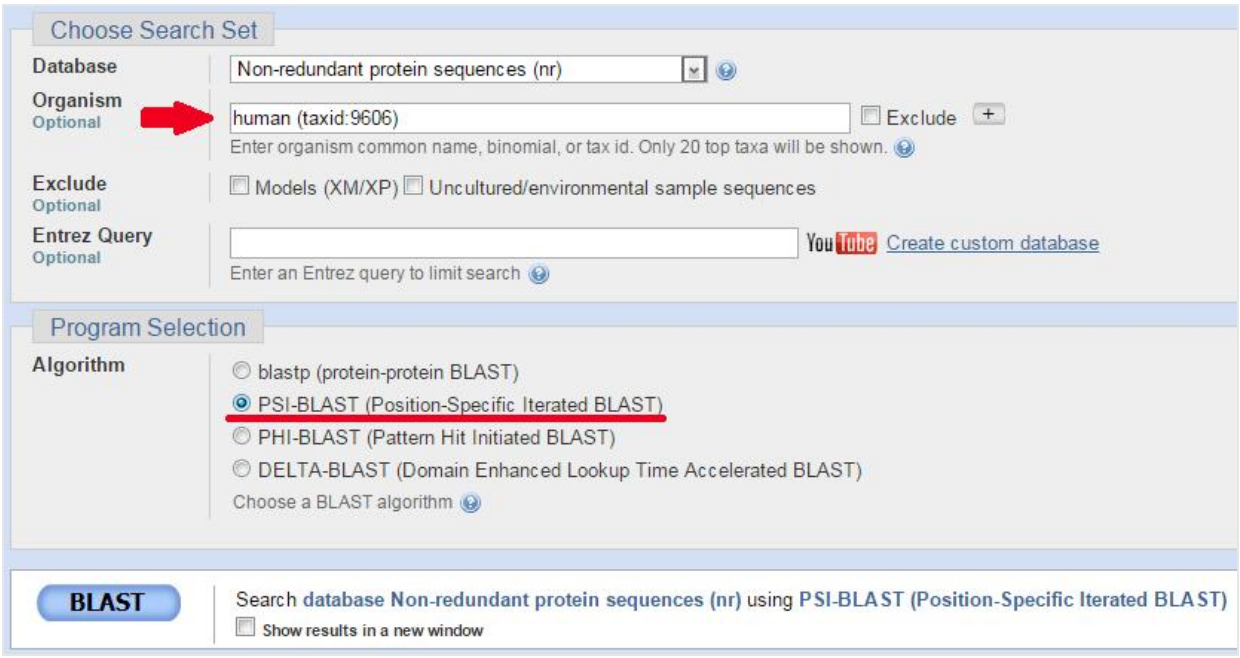
3. Пошук необхідного білку по якому буде проводитися дослідження.

Обравши шуканий білок у мікроорганізма *Magnetospirillum gryphiswaldense* MSR-1 і натиснувши справа **Run Blast**, вирівняти його амінокислотну послідовність з амінокислотними послідовностями білків необхідного організму.

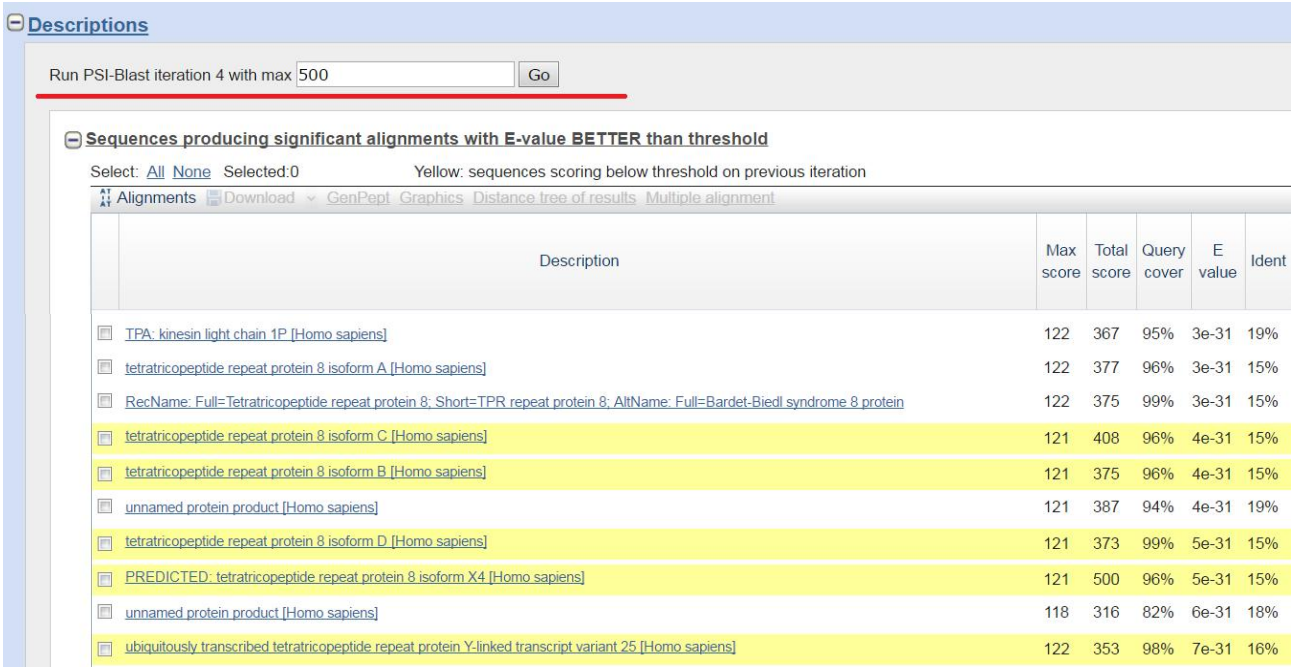


4. Запуск програми **PSI-BLAST**.

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



5. Проведення декількох ітерацій.



Description	Max score	Total score	Query cover	E value	Ident
TPA: kinesin light chain 1P [Homo sapiens]	122	367	95%	3e-31	19%
tetratricopeptide repeat protein 8 isoform A [Homo sapiens]	122	377	96%	3e-31	15%
RecName: Full=Tetratricopeptide repeat protein 8; Short=TPR repeat protein 8; AltName: Full=Bardet-Biedl syndrome 8 protein	122	375	99%	3e-31	15%
tetratricopeptide repeat protein 8 isoform C [Homo sapiens]	121	408	96%	4e-31	15%
tetratricopeptide repeat protein 8 isoform B [Homo sapiens]	121	375	96%	4e-31	15%
unnamed protein product [Homo sapiens]	121	387	94%	4e-31	19%
tetratricopeptide repeat protein 8 isoform D [Homo sapiens]	121	373	99%	5e-31	15%
PREDICTED: tetratricopeptide repeat protein 8 isoform X4 [Homo sapiens]	121	500	96%	5e-31	15%
unnamed protein product [Homo sapiens]	118	316	82%	6e-31	18%
ubiquitously transcribed tetratricopeptide repeat protein Y-linked transcript variant 25 [Homo sapiens]	122	353	98%	7e-31	16%

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

6. Аналіз отриманих білків гомологів на спільну будову і функції.

7. Оформити результати у вигляді таблиці, написати висновки.

Таблиця 2.1

Результати пошуку білків-гомологів

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

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?

ЛАБОРАТОРНА РОБОТА № 3

Оцінка значимості вирівнювань

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

Теоретичні відомості

Припустимо, що вирівнювання показує подібність двох послідовностей. Необхідно з'ясувати, чи має ця подібність біологічний зміст, чи співпадіння двох послідовностей є випадковим.

Основні аспекти для з'ясування подібності послідовностей наступні:

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

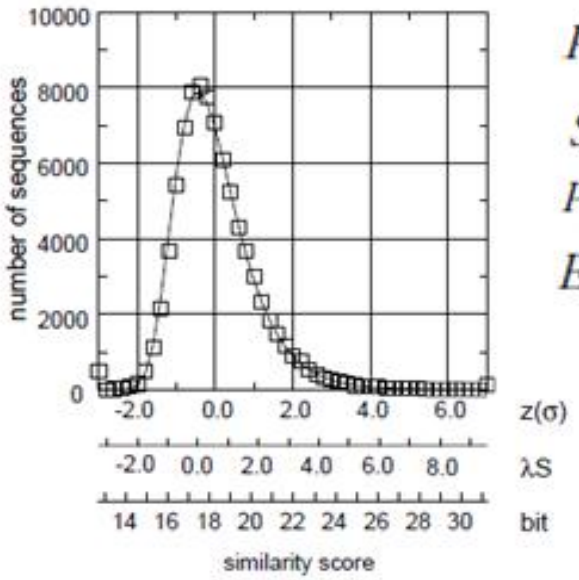
Основний підхід до визначення значимості вирівнювань – це розрахунок статистичної значимості ваги вирівнювань, який базується на: моделях Бернуллі або моделях Маркова та їх модифікаціях; матрицях PAM, BLOSUM та ін.

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 (рис.3.1).



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

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

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

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

Рис. 3.1. ***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 (рис. 3.2).

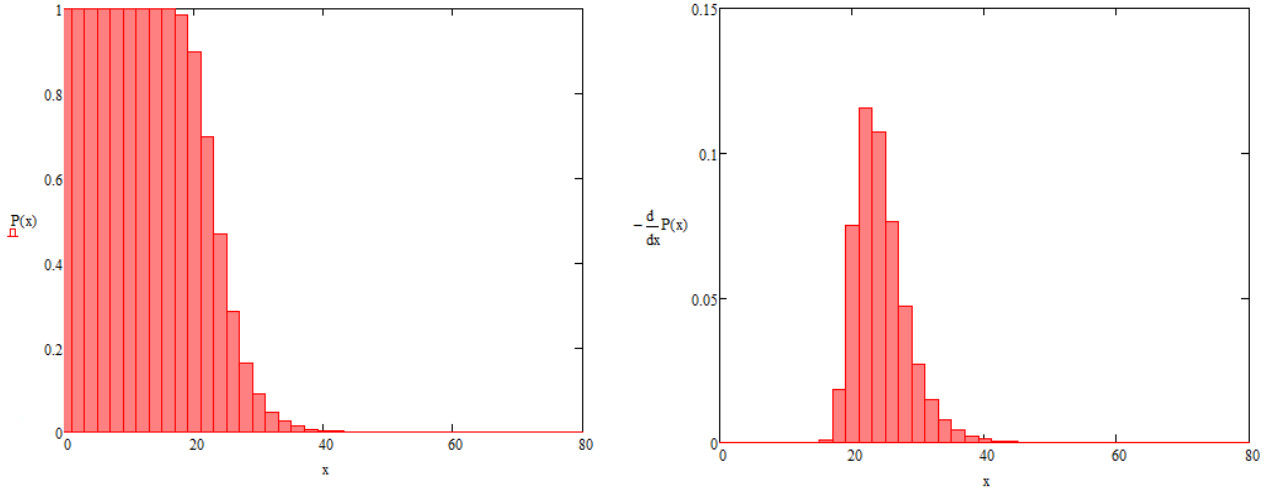


Рис. 3.2. ***Функція розподілу ваг вирівнювань або розподілу ймовірностей ваг представляє собою граничний випадок побудованої гістограми при S, що прямує до нуля***

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

$$P(S' \geq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \langle s \rangle)^2}{2\sigma^2} \right) \tag{3.1}$$

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 (an extreme-value distribution),** 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:

$$P(S' \geq x) = 1 - \exp\left(Kmne^{-\lambda x} \right) \tag{3.2}$$

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 (Poisson distribution).

Оцінка статистичної значимості вирівнювання

Крім ***P*-чисел** (*P-value*) та ***E*-чисел** (*E-value*), розрахунок яких було представлено вище – формули (3.2), (3.3) в біоінформатиці для характеристики статистичної значимості вирівнювання використовуються **Z-числа** (Z-score).

Таким чином, в біоінформатиці використовуються три основні величини 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:

$$Z = \frac{x - \langle s \rangle}{\sigma_s} \tag{3.3}$$

де *x*– вага вирівнювання послідовності запиту з послідовністю з БД;

<s> – середня вага вирівнювань послідовності запиту з випадковими послідовностями, яка дорівнює:

$$\langle s \rangle = \frac{\sum_{i=1}^{n_{\max}} x_i n_i}{n} \tag{3.4}$$

$$\sigma_s = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \langle s \rangle \right)^2 n_i}{n_{\max}(n_{\max} - 1)}} \tag{3.5}$$

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 далекоспоріднені 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 E-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, *E*-чисел та Z-чисел для порівняння послідовностей розраховують відсоток ідентичних залишків. Якщо два білки містять більше 45% ідентичних залишків в їх оптимальному вирівнюванні, то ці білки мають схожі структури і, швидше за все, загальну або, принаймні, схожу функцію.

Якщо вони містять від 25% до 45% ідентичних залишків, вони, ймовірно, мають схожий фолдинг. Низька міра подібності послідовностей не може унеможливити гомології. Область 18%-25%-ої подібності послідовностей – це «область двозначності», для якої можлива гомологія, але таке припущення може бути невірним. Вирівнювання, які знаходяться нижче за цю область (18%) потребують додаткових досліджень для встановлення гомології.

Хід роботи

1. Відкрити **NCBI**, обирати **Protein**, знайти потрібний білок (MamA, MamB, MamM, MamE, MamO, MamK) й обирати будь-який із знайдених.

2. Запустити **Run Blast**.

3. Натиснути на **Search Summary.**

4. З таблиці 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

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

	Description	Max score	Total score	Query cover	E value	Ident
☐	Pregnancy zone protein [Mus musculus]	3104	3104	100%	0.0	100%
☐	pregnancy zone protein, isoform CRA_b [Mus musculus]	3103	3103	100%	0.0	99%
☐	alpha-2-macroglobulin precursor [Mus musculus]	3101	3101	100%	0.0	99%

6. Відкрити файл *Karlin-Altschul statistics* в MathCad, ввести отримані значення X, K, Lambda. Значення n і m дорівнюватимуть 100.

7. Побудувати графіки.

8. Оформити 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.

ЛАБОРАТОРНА РОБОТА № 4

Еволюційні дослідження гомологів білків

Мета роботи – дослідити еволюційну часову шкалу гомологів білків.

Теоретичні відомості

Під час пошуку гомологічних білків намагаємося визначити ті з них, які мали спільного попередника в минулому. На рисунку 4.1 показано загальне еволюційне дерево, корені якого сягають назад до початку історії Землі. Пошук білків-гомологів полягає у порівнянні амінокислотних послідовностей досліджуваного білка з амінокислотними послідовностями білків, взятих з бази даних. 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 зберіглася 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.

Під час дослідження послідовностей білка або ДНК майже завжди вивчаються 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, є приклади послідовностей, які виявлено тільки в хребетних, або тільки в тварин або рослин, але не в обох з них. Такі послідовності менш давні, ніж ті, що наявні і у ссавців, і у бактерій.

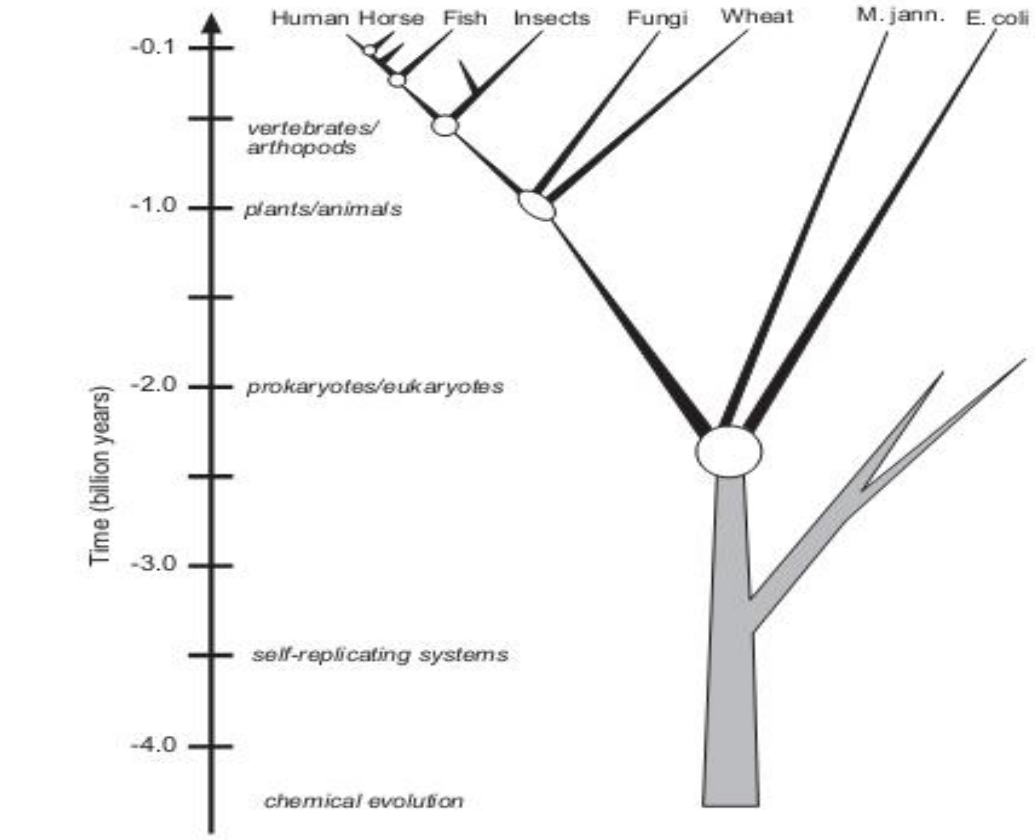


Рис. 4.1. ***Еволюційне дерево, корені якого сягають назад до початку історії Землі***

Для організмів, які дивергували протягом останніх 600 млн. років, інформація про дивергенцію для сучасних організмів взята 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 табл. 4.1 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 4.1

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:

- ортологічні послідовності, які відрізняються, оскільки вони знаходяться у різних видів;

- паралогічні послідовності – послідовності, які відрізняються в результаті події дуплікації генів.

Рис. 4.2 ілюструє еволюційне дерево для послідовностей ортологів цитохрому *c*. Розгалужений візерунок, який відображає відмінності між послідовностями цитохромів *c*, 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 відбулися 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.

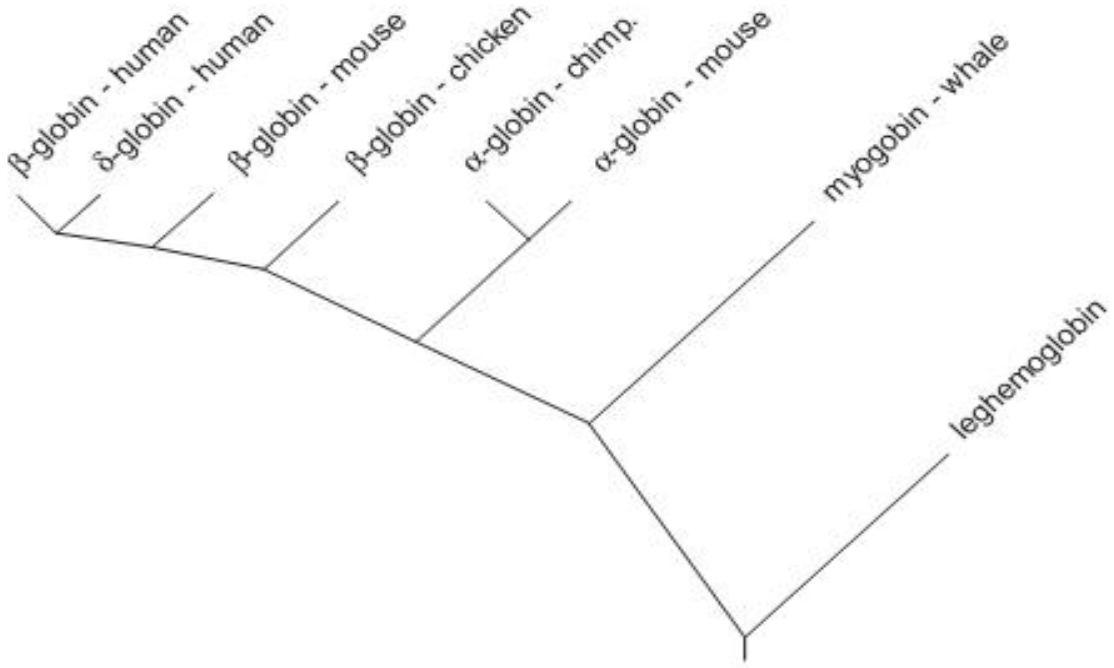
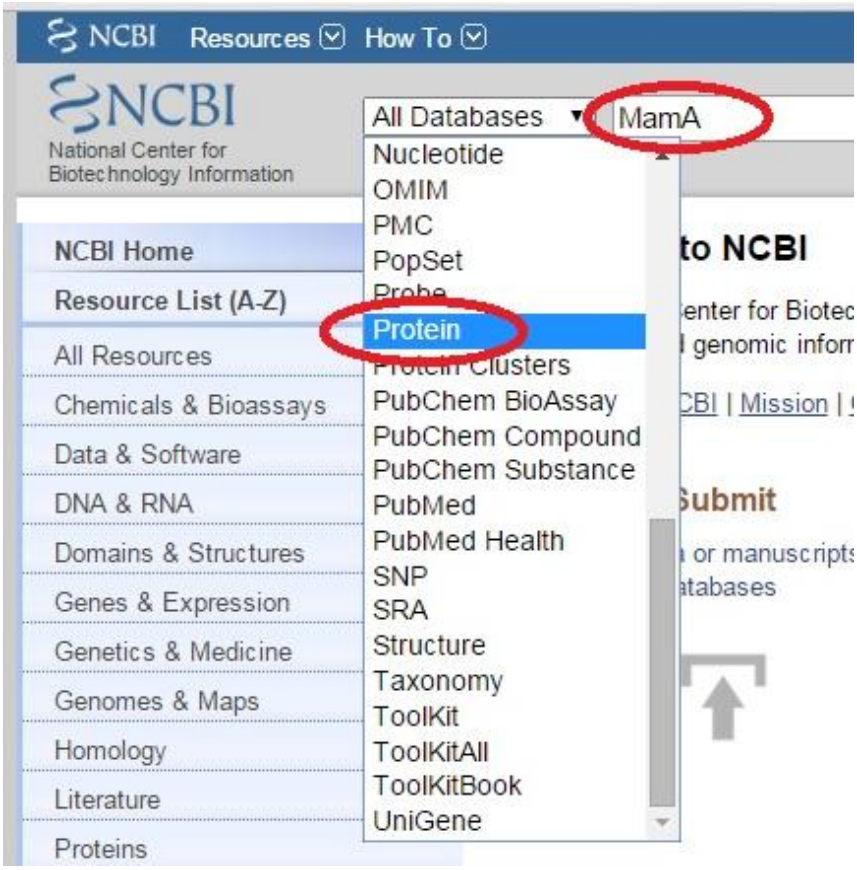


Рис. 4.2. ***Ортологи і паралоги – родина глобінів***

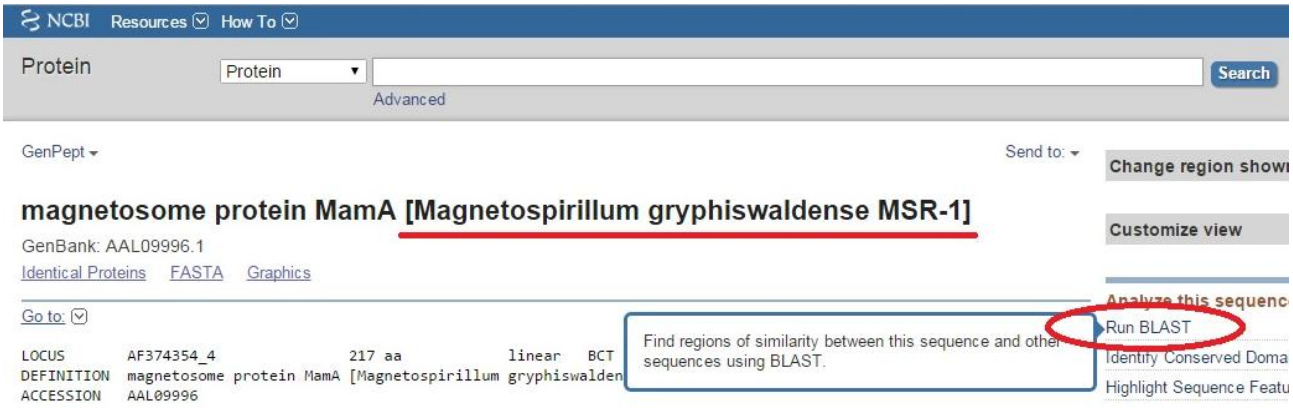
Це справедливо для паралогічних послідовностей, оскільки послідовність дублюється, вона може змінитися дуже швидко, перш ніж селективний тиск на її нову функцію уповільнить швидкість її зміни.

Хід роботи

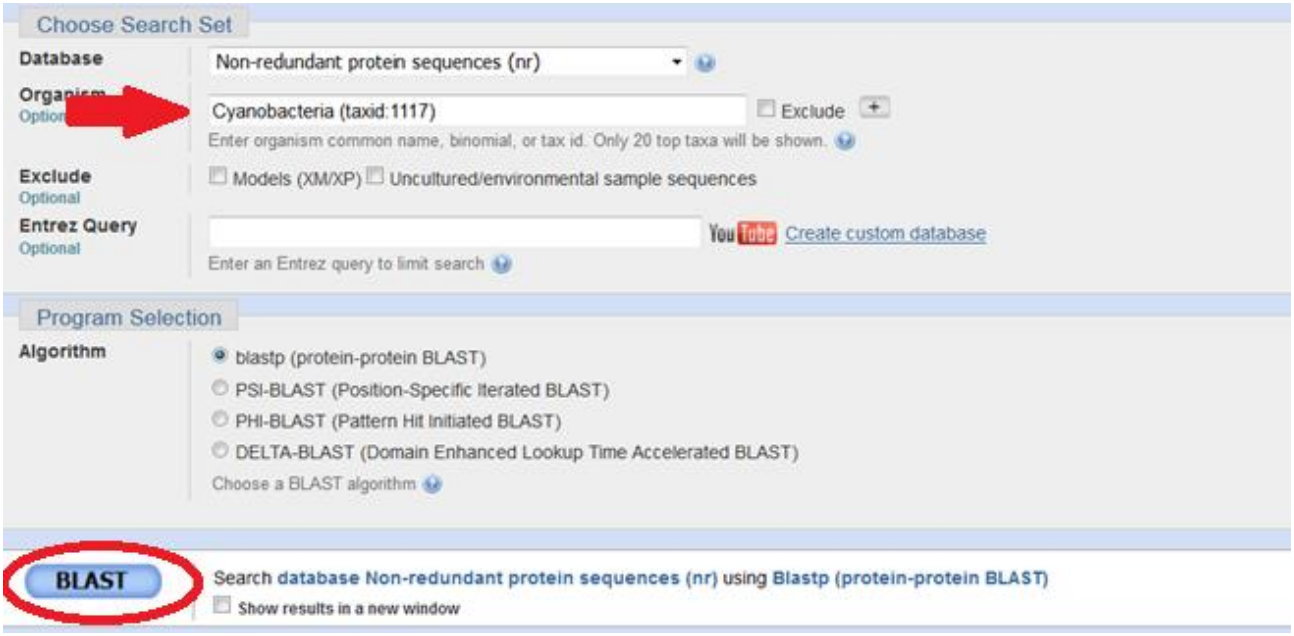
1. Запустити ресурс **NCBI** «http://www.ncbi.nlm.nih.gov/».
2. В верхньому лівому кутку вибрати **Protein**. Ввести в поле пошуку білок (**Mam-білок**) по якому буде проводитися дослідження. Натиснути в верхньому правому кутку **Search**.



3. Обрати організм ***Magnetospirillum gryphiswaldense MSR-1***. 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*.

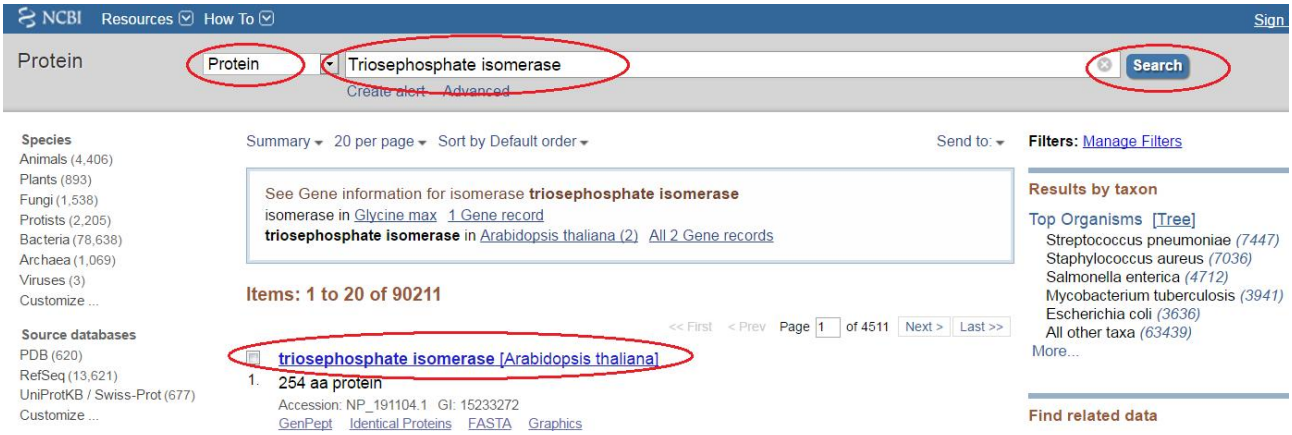
Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident
TPR repeat-containing protein [Cyanothece sp. PCC 7424]	77.8	762	86%	4e-15	27%
hypothetical protein AUI36_12715 [Cyanobacteria bacterium 13_1_40CM_2_61_4]	77.4	219	77%	4e-15	25%
hypothetical protein [Crocosphaera watsonii]	73.6	166	98%	1e-13	25%
hypothetical protein [Oscillatoria sp. PCC 10802]	72.8	497	95%	2e-13	24%
hypothetical protein [Crinalium epipsammum]	72.4	563	94%	3e-13	27%
hypothetical protein [Phormidesmis priestleyi]	71.6	149	76%	4e-13	27%
hypothetical protein [Oscillatoriales cyanobacterium MTP1]	70.9	209	75%	8e-13	29%
Demethylrebeccamycin-D-glucose O-methyltransferase [Prochlorococcus marinus str. MIT 1312]	70.9	148	92%	9e-13	23%
hypothetical protein [Trichodesmium erythraeum]	70.5	396	95%	1e-12	26%
glycosyl transferase, family 2 [Trichodesmium erythraeum IMS101]	70.5	561	96%	2e-12	24%
hypothetical protein [Trichodesmium erythraeum]	70.5	561	96%	2e-12	24%

6. Визначити для обраних білків-гомологів значення параметрів **E-value** та **Ident**. Отримані результати занести до таблиці.

7. У новому вікні запустити ресурс **NCBI**.

8. В верхньому лівому кутку вибрати **Protein**. Ввести в поле пошуку давній 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 давнього 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 давніх 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 давніх білків ціанобактерій. Зробити висновки.

Контрольні запитання:

1. Що таке еволюційне дерево?

2. Яка основна властивість гомологічних послідовностей?

3. На які групи можна поділити гомологічні послідовності? Їх характеристика.

ЛАБОРАТОРНА РОБОТА № 5

Пошук серед мікроорганізмів, що викликають захворювання серця і мозку, потенційних продуцентів біогенних магнітних наночастинок (БМН)

Мета роботи – ознайомитись з процедурою вирівнювання амінокислотних послідовностей білків, а також знайти серед мікроорганізмів, що викликають захворювання серця та мозку, потенційних продуцентів БМН.

Теоретичні відомості

БМН є в багатьох органах організму людини. Серед них: серце, селезінка, печінка, надниркові залози, а також головний мозок. Тому дуже важливо розуміти як магнітні частинки, що накопичуються в наслідок протікання бактеріальних захворювань, збудники яких можуть бути продуцентами БМН, взаємодіють з тими, що вже присутні в організмі.

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 порівняно із здоровими особами, аналогічна ситуація і з іншими хворобами.

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. Причиною запалення можуть бути бактерії, віруси, а іноді і 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.

Здійснити пошук потенційних продуцентів БМН серед вище зазначених мікроорганізмів можна за допомогою програми BLAST.

BLAST – сімейство комп'ютерних програм, що використовуються для пошуку гомологів білків або нуклеїнових кислот, для яких відома первинна структура (послідовність) або її фрагмент. Використовуючи дану програму, можна порівняти наявну в нього послідовність з послідовностями з бази даних і знайти послідовності передбачуваних гомологів. 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 E-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 табл. 5.1, треба обирати ті гомологи, значення E-числа у яких $E \leq 0{,}02$.

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

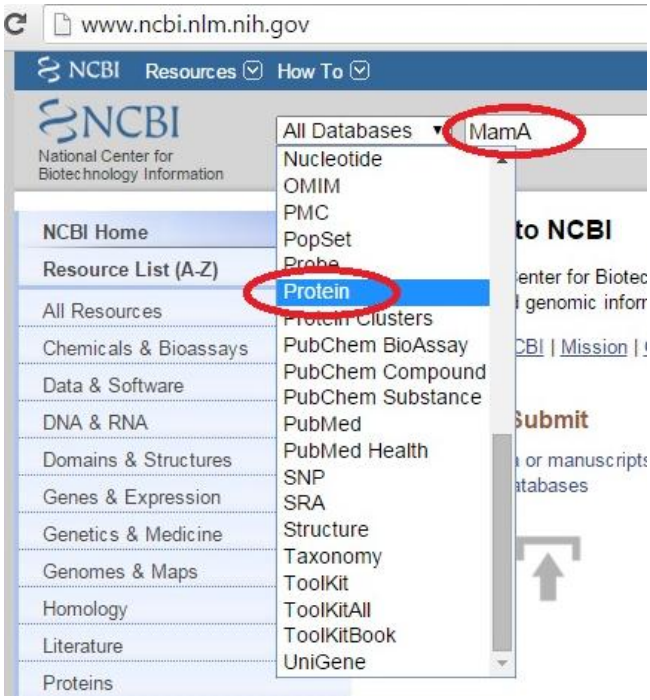
При вирівнюванні виконується пошук гомологів білків магнітосомного острівця магнітотаксисних бактерій *Magnetospirillum gryphiswaldense* MSR-1. MamA, MamB, MamM, MamE, MamO, MamN – ті білки, без яких неможливий процес біомінералізації магнетиту. Для формування кристалічних БМН в досліджуваних мікроорганізмах достатньо наявності гомологів білків: MamA,

MamB, MamE, MamO, MamM, а для формування аморфних БМН необхідна наявність гомологів білків: MamB, MamE, MamO, MamM.

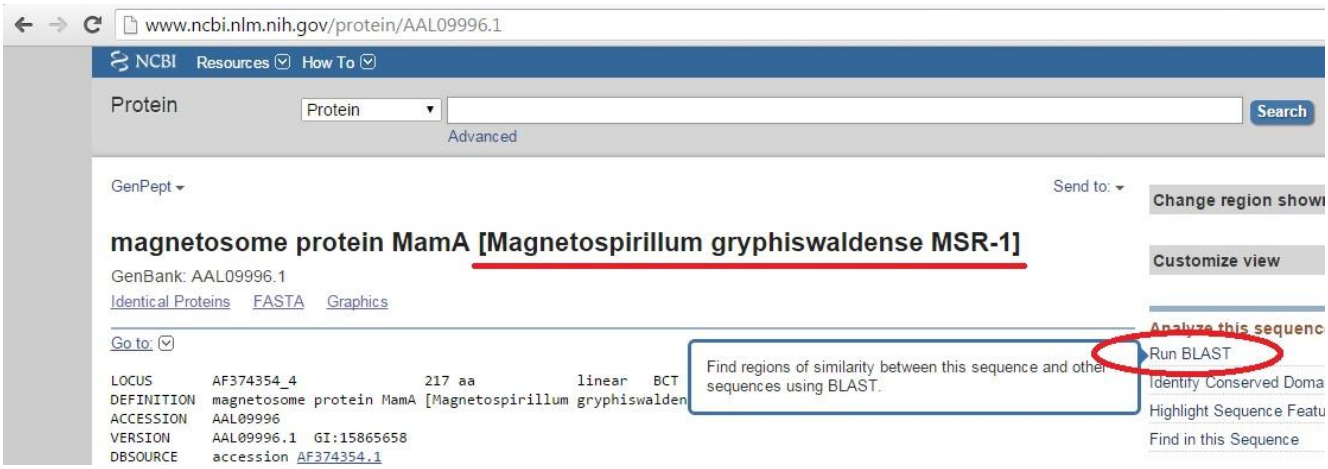
Хід роботи

1. Зайти на сайт **NCBI** (National Center for Biotechnology Information).

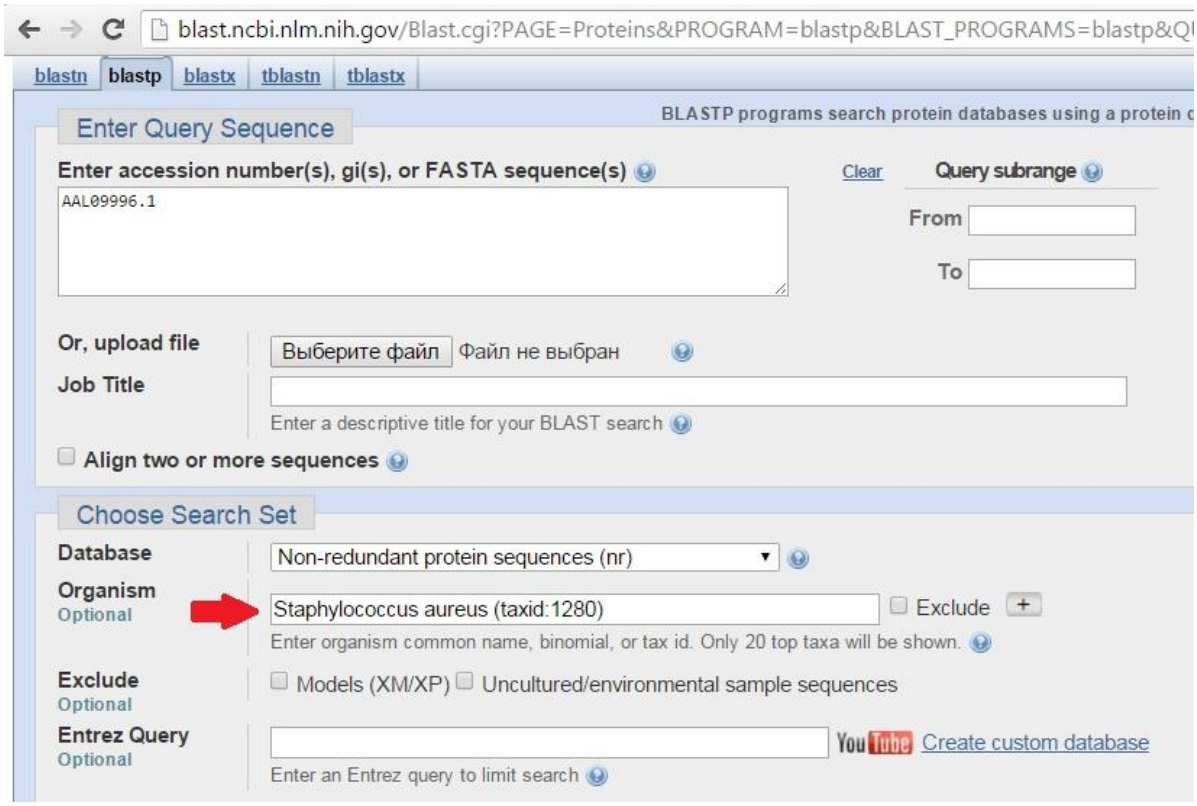
В верхньому лівому кутку вибрати **Protein,** ввести в поле пошуку необхідний білок MamA, MamB, MamM тощо.



2. Обрати досліджуваний білок у мікроорганізма ***Magnetospirillum gryphiswaldense*** **MSR-1** і натиснути справа **Run Blast**.



3. У полі «organism» обрати мікроорганізм, з протеомом якого буде порівнюватись протеом *Magnetospirillum gryphiswaldense* MSR-1.



4. За значенням числа E-value (менше ніж 0,02) та Ident (більше 15%) обрати найкращі гомологи.

5. Результати оформити у вигляді таблиці:

Бактерія-збудник захворювання	Білок Mam	Білок-гомолог	E-value	Ident	Тип БМН (кристалічні/аморфні)

Контрольні запитання:

1. Порогові значення параметрів вирівнювання. Е-число та Ident.

2. Які білки необхідні для формування кристалічних БМН, а які для формування аморфних?

3. Характеристика збудників захворювання серця та мозку

ЛАБОРАТОРНА РОБОТА № 6

Підбір праймерів для полімеразної ланцюгової реакції (ПЛР) біоінформаційними методами

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

Теоретичні відомості

Підбір праймерів здійснюється методом множинного вирівнювання. Множинне вирівнювання – ключовий метод в сучасній молекулярній біології.

Полімеразна ланцюгова реакція (ПЛР) – експериментальний метод молекулярної біології, дозволяє отримати значне збільшення малих концентрацій певних фрагментів нуклеїнової кислоти (ДНК / РНК) в біологічному матеріалі (пробі). Фактично – це метод *in vitro*, за допомогою якого можна примножити (ампліфікувати) специфічну ділянку ДНК.

Реакційна суміш для ПЛР складається з наступних компонентів: праймери, Taq-полімераза, суміш дезоксинуклеотидтрифосфатів (дНТФ), буфер, зразок.

Якщо в аналізованому зразку присутня шукана ДНК, то в процесі реакції ампліфікації з нею відбувається ряд подій, які забезпечуються певними температурними циклами.

Кожен цикл ампліфікації складається з трьох етапів:

1. Денатурація – це перехід ДНК з дволанцюгової форми в одноланцюгову при розриві водневих зв’язків між комплементарними парами основ під впливом високих температур.

2. Відпал – це приєднання праймерів до одноланцюгової ДНК-мішені. Праймери підбирають так, що вони обмежують досліджуваний фрагмент і комплементарні протилежним ланцюгах ДНК. Відпал відбувається відповідно до правил комплементарності Чаргаффа. Якщо ця умова не дотримана, то відпал праймерів не відбувається.

3. Елонгація (синтез). Після відпалу праймерів Taq-полімераза починає добудовувати другий ланцюг ДНК з 3’-кінця праймера.

Праймери – штучно синтезовані олігонуклеотиди, що мають, як правило, розмір від 15 до 30 нуклеотидів, ідентичні відповідним ділянкам ДНК-мішені. Вони відіграють ключову роль в утворенні продуктів реакції ампліфікації. Правильно підібрані праймери забезпечують специфічність і чутливість тест-системи і повинні відповідати ряду критеріїв:

1. праймери повинні бути специфічними. Особливу увагу приділяють 3’-кінцям праймерів, так як саме з них Taq-полімераза починає добудовувати комплементарний ланцюг ДНК. Якщо їх специфічність недостатня, то висока ймовірність, що в пробірці з реакційною сумішшю відбуватимуться процеси неспецифічного зв’язування і синтезу фрагментів різної довжини, відмінних від шуканих. Частина праймерів і дНТФ витрачається на синтез неспецифічної ДНК, що призводить до значної втрати чутливості;

2. працмери не повинні утворювати димери і петлі, тобто не повинно утворюватися стійких подвійних ланцюгів в результаті відпалу (комплементарного приєднання) праймерів самих на себе або одного з одним;

3. область відпалу праймерів повинна знаходитися поза зонами мутацій, делецій або інсерцій в межах видової чи іншої специфічності, взятої в якості критерію при виборі праймерів. При попаданні на таку зону відпал праймерів не відбувається, і, як наслідок, виникають помилково негативні результати.

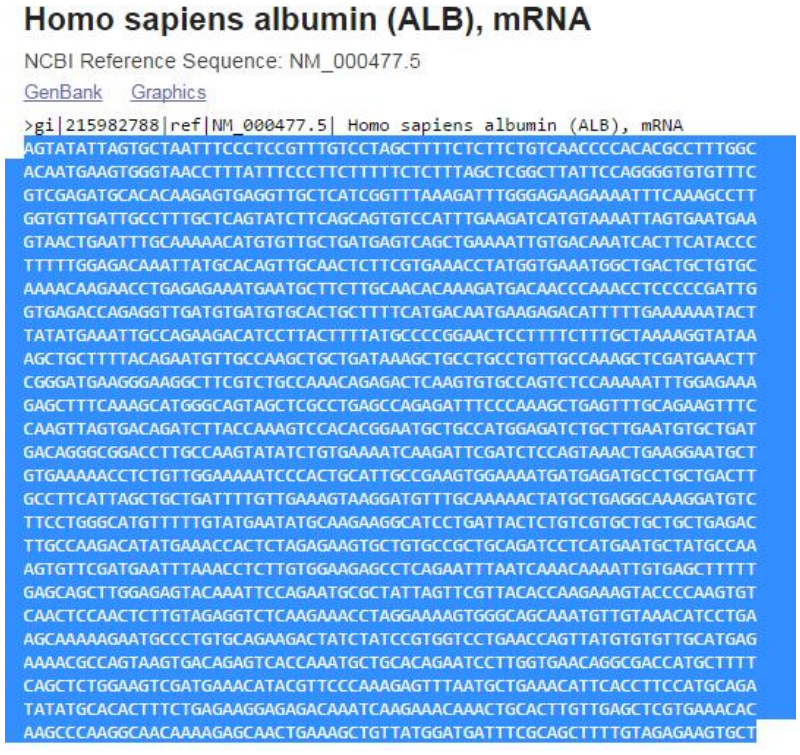
Хід роботи

1. Запуск ресурсу **NCBI.**

2. В верхньому лівому кутку вибрати **Nucleotide,** ввести в поле пошуку білок, по якому буде проводитися дослідження. Обрати організм.

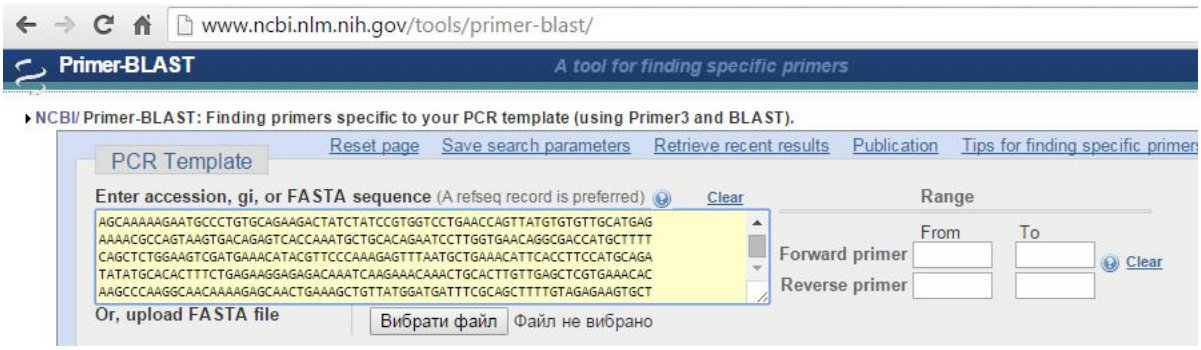
3. Обрати програмний пакет (формат) **FASTA**, який містить нуклеотидну послідовність.

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 послідовність, обрати параметри пошуку й натиснути **GetPrimers.**



7. Оформити результати документом MSWORDу вигляді скріншотів. Зробити висновки.

Контрольні запитання:

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

ЛАБОРАТОРНА РОБОТА № 7

Побудова філогенетичних дерев

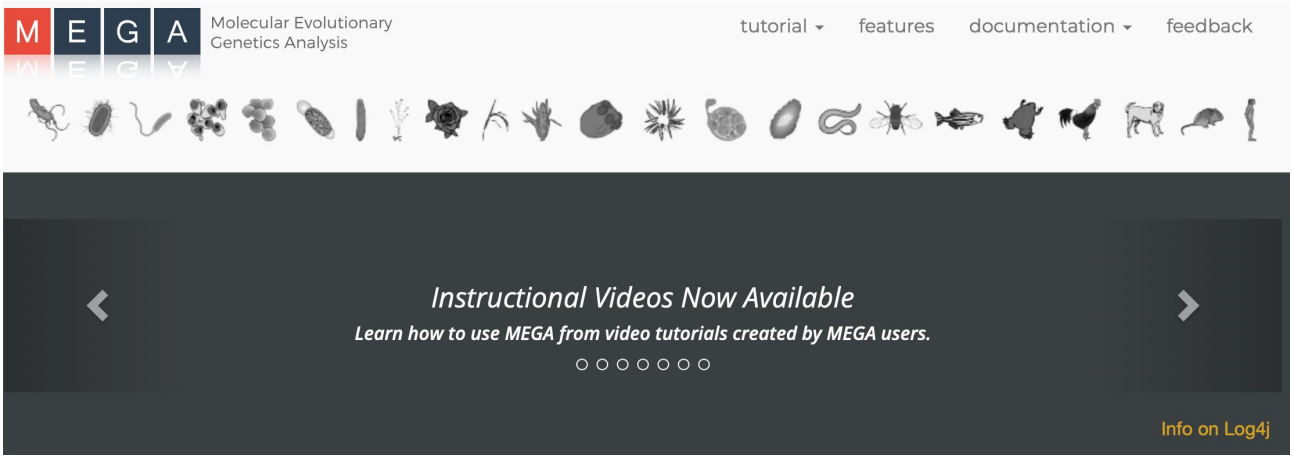
Мета роботи – освоїти метод побудови філогенетичного дерева на платформі MEGA.

Теоретичні відомості

Конструювання філогенетичних дерев на основі множинного вирівнювання.

Філогенія – це опис стосунків між біологічними послідовностями (організмами), що зазвичай зображується у вигляді дерева. Зазначені подібності та відмінності між послідовностями (організмами) використовують для відновлення філогенії.

Філогенетичний аналіз у систематиці визначає взаємовідносини серед таксонів і покликаний допомогти зрозуміти історію еволюційних відносин між живими організмами. Еволюційну історію, відновлену в результаті філогенетичного аналізу, зазвичай зображують у вигляді розгалужених, деревоподібних діаграм, які представляють передбачуваний родовід спадкових відносин між молекулами, організмами, або тим та іншим. Основою розуміння цих процесів є кількісні відносини еволюційних подій, що у кожного окремого організму з його відділення від загального пращура. Наразі розглядаємо лише непрямі показники фактичних подій. У філогенетики найбільш зручний шлях візуального представлення еволюційних взаємин серед груп організмів здійснюється за допомогою графіків, які називаються філогенетичними деревами. Для виконання подібних філогенетичних реконструкцій зручно використовувати (як чорнові макети) on-line програми, які рекомендовано навіть off-line загрузити. Для виконання завдання будемо використовувати виключно доступний ресурс – це MEGA: www.megasoftware.net.

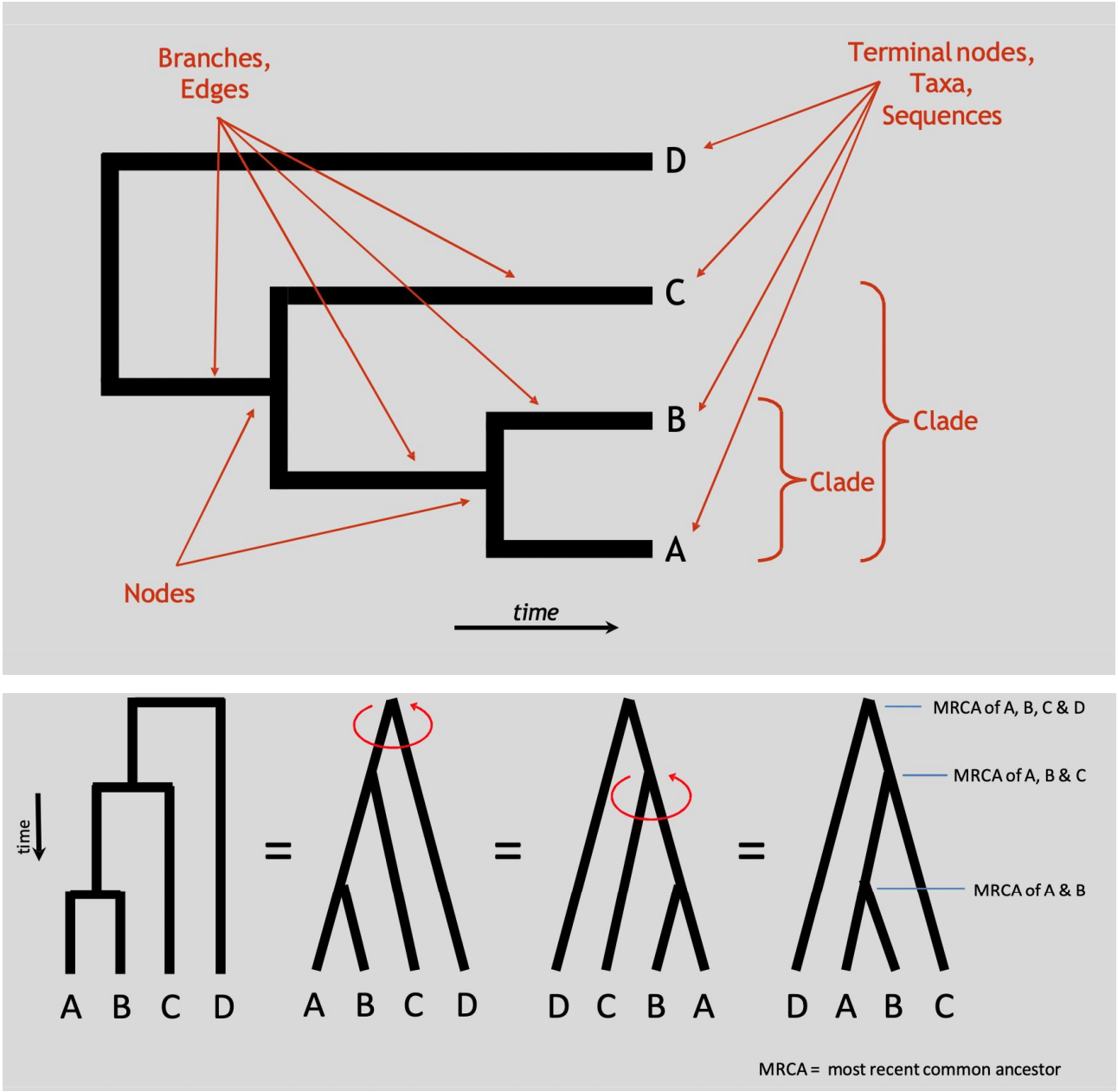


www.megasoftware.net – це безкоштовний, простий у використанні веб-сервіс, присвячений реконструкції та аналізу філогенетичних зв'язків між біологічними послідовностями. В основі роботи сервера MEGA лежить поєднання різних біоінформаційних програм з метою реконструкції надійного дерева філогенезу з набору послідовностей.

У цій роботі будемо використовувати вирівнювання множинних послідовностей (MSA). В результаті філогенетичного аналізу генеруються розгалужені діаграми, які можуть чітко ілюструвати відносини між послідовностями, які не очевидні з BLAST або MSA. Філогенетичні дерева, очевидно, корисні для еволюційних та порівняльних досліджень, орієнтованих на з'ясування еволюційних взаємин та моделей дивергенції, але вони також стають все більш важливими для генерації гіпотез щодо функції гена чи білка для молекулярних та біохімічних досліджень.

Існує дві основні категорії філогенетичних інструментів: методи, що розглядають відстані між послідовностями (distance-based methods) та методи, що розглядають ознаки (character-based methods). В роботі будемо використовувати обидва ці підходи. Часто різні підходи дають різні висновки; отже, потрібно працювати з кількома методами та значеннями параметрів, і в кінцевому підсумку використовувати біологічну інтуїцію, щоб генерувати найкраще дерево та зробити якісний добрий аналіз.

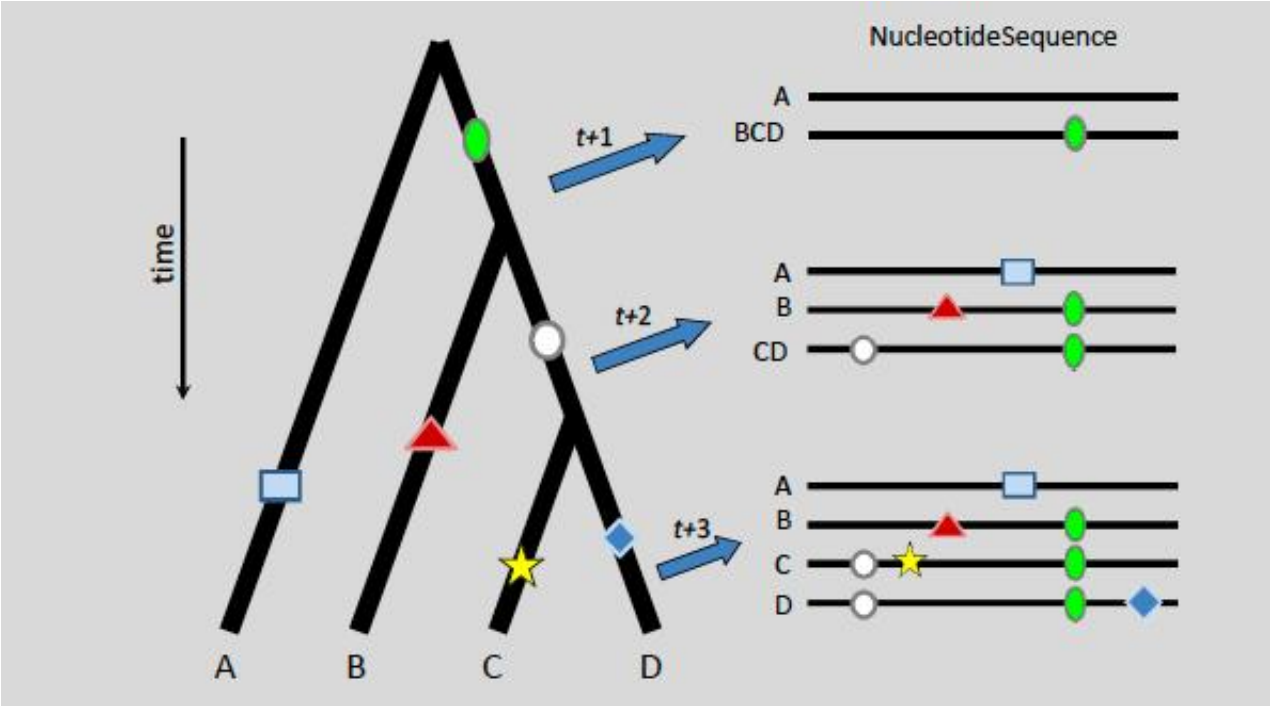
Філогенетичне дерево



Філогенетичні дерева можуть бути представлені в різних формах та орієнтаціях. Важливо, що єдиний спосіб визначити еволюційну відстань між двома послідовностями – це визначити, як далеко назад у часі потрібно піти, перш ніж знайти спільного пращура. Так, на дереві справа, послідовність A фізично розташовується ближче до D порівняно з послідовністю C, хоча насправді послідовність A більш тісно пов’язана з послідовністю C, оскільки вони мають більше спільних пращурів. Еволюційні відносини між A-D також можуть бути представлені з використанням формату Newick наступним чином

(((A, B), C), D): вкладеність у круглі дужки відповідає розподілу на деревах вище.

Зростання філогенетичного дерева



У результаті еволюції організми змінюються, накопичують мутації (позначені кольором). Ці мутації передаватимуться у потомстві всім дочірнім лініям. Мутація, яка відбувається дуже рано в історії групи, наприклад, зелений овал, яка відбулася у пращура послідовностей B, C&D, буде знайдена у всіх трьох лініях нащадків. Мутацію, що відбувається пізніше (наприклад, синій квадрат), можливо, можна знайти в меншій кількості ліній. Положення мутацій на нуклеотидних послідовностях цілком довільне і призначене тільки, щоб показати, як багато унікальних послідовностей є в кожний момент часу, розподіл мутації серед цих послідовностей.

ЕТАПИ ВИКОНАННЯ ФІЛОГЕНЕТИЧНОГО АНАЛІЗУ У ПАКЕТІ MEGA

1. Спочатку необхідно мати файл із нуклеотидними або амінокислотними послідовностями. Використовуємо вже наявний у файл із послідовностями, який використовували для множинного вирівнювання із FASTA форматами.

2. Починаємо з побудови дерева за допомогою neighbour-joining методу в програмі MEGA. MEGA (Molecular Evolutionary Genetic Analysis) досить простий у використанні, хоча і дуже потужний додаток для проведення філогенетичного аналізу. Метод Neighbour-joining – швидкий і досить надійний, тому в ньому роблять більшість стартових дерев для побудови гіпотез про загальне дерево топології / відстані, перш ніж перейти до складніших програм.

2.1. Відкриваємо програму MEGA 5 і конвертуємо вирівняні та збережені на лабораторному занятті № 2 полінуклеотидні послідовності DNA_aligned.fas з FASTA формату у формат MEGA.

2.2. Натискаємо на File/Convert File Format to MEGA…

2.3. Для того, щоб знайти файл, необхідно натиснути на значок дуже невеликої папки з правого боку Data file to convert боксу.

2.4. Знаходимо файл з вирівняними нуклеотидними послідовностями у форматі FASTA (aligned nucleotide sequence file). Вибераємо Data Format (FASTA).

Передбачаючи, що файл перетворено правильно, зберігаємо його. При цьому до імені файлу додається розширення .meg. MEGA та FASTA формати дуже схожі для цих простих файлів, основна відмінність, що MEGA зберігає більше інформації у різних полях.

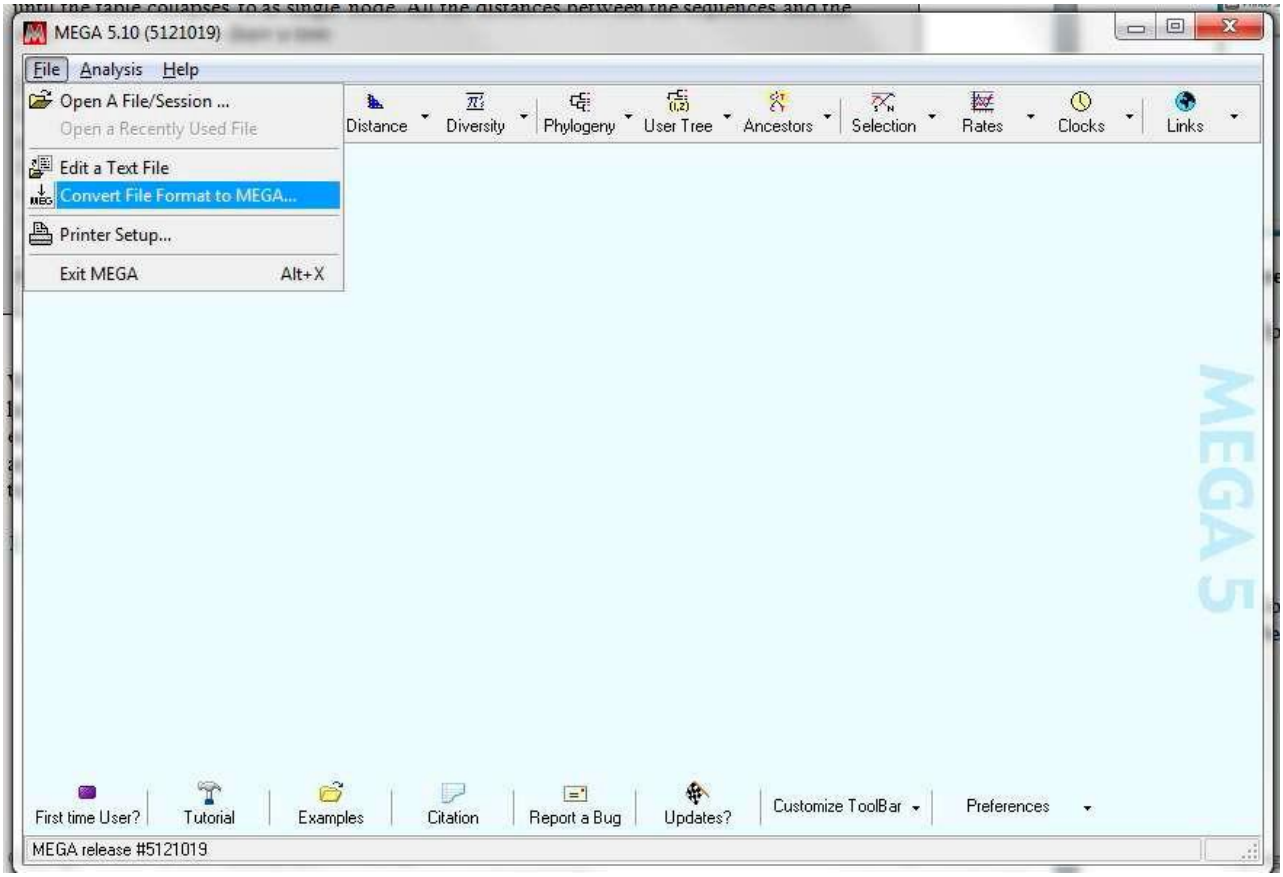
2.5. MEGA формат має особливості, а саме:

2.5.1. вхідні файли не можуть мати імена послідовностей з пробілами, або будь-якого з наступних символів; : ’ “! ? > < [ ] ~ @ # ^ &

2.5.2. фігурні дужки можна використовувати лише тоді, коли вони в парі;

2.5.2. перший рядок завжди: #MEGA;

2.5.3. другий рядок завжди: !Title: xxx , де xxx – будь-що.



3. Відкриваємо новий файл.

3.1. Натискаємо File > Open a File/Session.

3.2. Знаходимо та відкриваємо новий файл із розширенням .meg.

3.3. Позначаємо, що цей файл містить Nucleotide Sequences, та будь-яку іншу необхідну інформацію (protein coding sequence = Y, select genetic code = standard).

3.4. Натискаємо на іконку «ТА» та відкриваємо Data Explorer. Це корисно для візуалізації та вибору даних та областей для аналізу.

Примітка: якщо є неприпустимі символи у файлі, МЕГА покаже вам, на яких лініях вони знаходяться.

4. Повертаємось до головного вікна та входимо до Analysis > Phylogeny > Construct/Test Neighbor- Joining Tree. Дотримуємось параметрів за замовчуванням цього разу, але тільки не забуваємо вибрати "метод Bootstrap" під Phylogeny Test/Test of Phylogeny. Натискаємо Compute.

5. Результатом має бути красиво відформатоване дерево філогенезу в новому вікні. Звертаємо увагу на значення, вказані над кожним вузлом чи гілкою дерева. Вони називаються завантажувальними значеннями та є мірою статистичної достовірності кожного вузла. Будь-яка завантажувальна оцінка>70, як правило, розглядається як досить надійна.

5.1. У вікні TreeExplorer вибераємо View > Options > Branch.

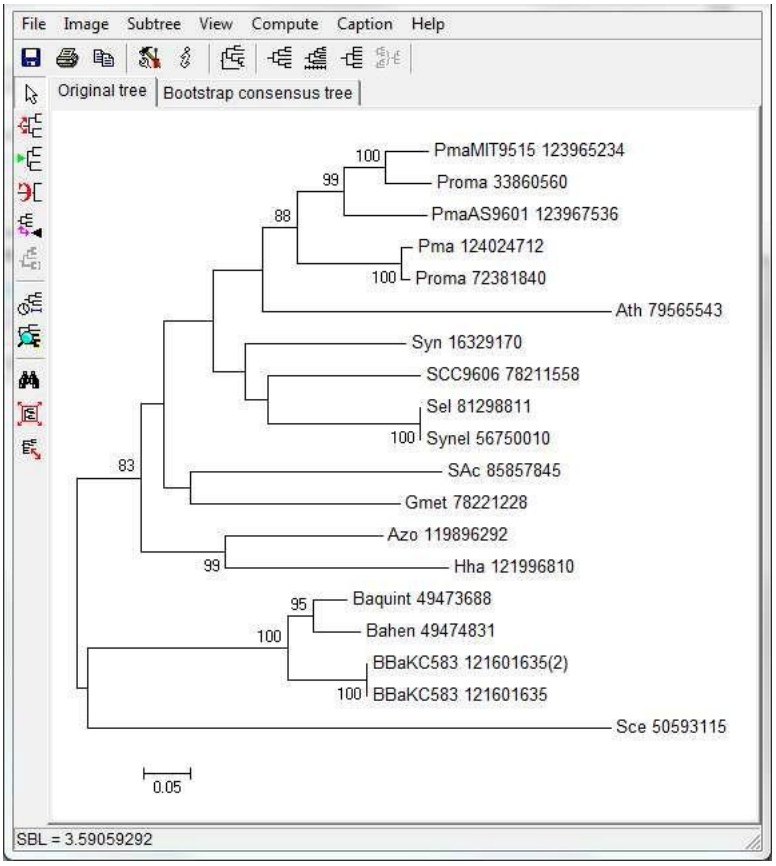
5.2. Вибераємо Hide values lower than та ставимо 70%.

a. Скільки надійних вузлів у дереві?

b. Більш надійні вузли знайдені ближче до основи чи на термінальних кінцях дерева?

c. Чи можете ви назвати можливі причини цього?

d. Чи є у вас велика довіра до цього дерева?



Дерево, побудоване методом найближчого сусіда з бутстрап-підтримкою (Bootstrapped Neighbour-Joining Tree), показано лише бутстрап-рахунки ≥70%.

6. Дуже просто можна змінити в MEGA спосіб подання дерева:

6.1. у вікні TreeExplorer переходимо до View > Tree Branch Style та вибераємо один із форматів: circular, radial, traditional straight…;

6.2. перевіряємо відносний порядок розгалуження із послідовностей з цими різними форматами.

Чи змінилися взаємозв’язки послідовностей?

7. MEGA TreeExplorer дуже потужний та дає можливість маніпулювати побудованим деревом нескінченно за допомогою Options (View > Options) та меню Subtree.

Повертаємо дерево в традиційний формат (traditional/rectangular) і пробуємо змінити різні опції.

8. Наступним етапом є робота з блоком MEGA, пов’язаним з філогенією, в ручному режимі, маючи готову послідовність.

Використаємо вже готову вирівняну послідовність Streptomyces sp. 27 пологів. Для кожного роду Streptomyces sp. через FASTA взяли 2 гомологи з кодом 16s РНК. Щоб з'ясувати їхню «спорідненість», слід врахувати, що для правильної і коректної побудови філогенетичного дерева необхідно перед нуклеотидною послідовністю вказувати найменування роду в такому вигляді: >Streptomyces sp. Для початку збережений файл копіюємо в текстовий редактор .txt, після чого відкриваємо та копіюємо наші 27 пологів Актиноміцетів.

Маємо такий вигляд:

>Streptomyces sp. Lv 1-44

CCTTAACCATGCAAGTCGAACGATGAAGCCCTTCGGGGTGGATTAGTGGC
GAACGGGTGAGTAACACGTGGGCAATCTGCCCTGCACTCTGGGACAAGC
CCTGGAAACGGGGTCTAATACCGGATAACACCTTCTCTCGCATGGGAGGG
GGTTCAAAGCTCCGGCGGTGCAGGATGAGCCCGCGGCCTATCAGCTTGTT
GGTGAGGTAGTGGCTCACCAAGGCGACGACGGGTAGCCGGCCTGAGAGG
GCGACCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGG
CAGCAGTGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCGACGC
CGCGTGAGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCAGCAGGGAAG
AAGCGAAAGTGACGGTACCTGCAGAAGAAGCGCCGGCTAACTACGTGCC
AGCAGCCGCGGTAATACGTAGGGCGCAAGCGTTGTCCGGAATTATTGGGC
GTAAAGAGCTCGTAGGCGGCTTGTCACGTCGGTTGTGAAAGCCCGGGGCT
TAACCCCGGGTCTGCAGTCGATACGGGCAGGCTAGAGTTCGGTAGGGGA
GATCGGAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAAC
ACCGGTGGCGAAGGCGGATCTCTGGGCCGATACTGACGCTGAGGAGCGA
AAGCGTGGGGAGCGAACAGGATTAGATACCCTGGTAGTCCACGCCGTAA
ACGGTGGGCACTAGGTGTGGGCAACATTCCACGTTGTCCGTGCCGCAGCT
AACGCATTAAGTGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTC
AAAGGAATTGACGGGGGCCCGCACAAGCGGCGGAGCATGTGGCTTAATT
CGACGCAACGCGAAGAACCTTACCAAGGCTTGACATACACCGGAAAGCA
TCAGAGATGGTGCCCCCCTTGTGGTCGGTGTACAGGTGGTGCATGGCTGT
CGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAAC

CCTTGTCCCGTGTTGCCAGCAAGCCCTTCGGGGTGTTGGGGACTCACGGG
AGACCGCCGGGGTCAACTCGGAGGAAGGTGGGGACGACGTCAAGTCATC
ATGCCCCTTATGTCTTGGGCTGCACACGTGCTACAATGGCCGGTACAATG
AGCTGCGATACCGCGAGGTGGAGCGAATCTCAAAAAGCCGGTCTCAGTTC
GGATTGGGGTCTGCAACTCGACCCCATGAAGTCGGAGTCGCTAGTAATCG
CAGATCAGCATTGCTGCGGTGAATACGTTCCCGGGCCTTGTACACACCGC
CCGTCACGTCACGAAAGTCGGTAACACCCGAAGCCGGTGGCCCAACCCCT
TGTGGGAGGGAGCTGTCGAAGGTGG

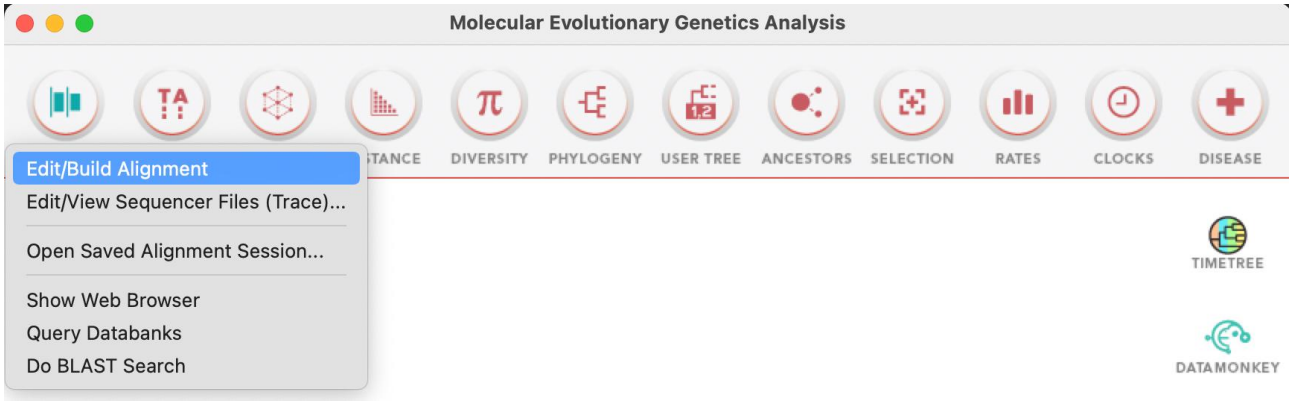
>Streptomyces sp. Lv 1-399

ATGCAAGTCGAACGATGAAGCCCATTCGGGGTGGATTAGTGGCGAACGG
GTGAGTAACACGTGGGCAATCTGCCCTGCACTCTGGGACAAGCCCTGGAA
ACGGGGTCTAATACCGGATGATATCCCCTCTCGCATGGGAGGGGGTTGAA
AGCTCCGGCGGTGCAGGATGAGCCCGCGGCCTATCAGCTAGTTGGTGAGG
TAGAAGCTCACCAAGGCGACGACGGGTAGCCGGCCTGAGAGGGCGACCG
GCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGT
GGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCGACGCCGCGTGA
GGGATGACGGCCTTCGGGTTGTAAACCTCTTTCAGCAGGGAAGAAGCGA
AAGTGACGGTACCTGCAGAAGAAGCGCCGGCTAACTACGTGCCAGCAGC
CGCGGTAATACGTAGGGCGCAAGCGTTGTCCGGAATTATTGGGCGTAAAG
AGCTCGTAGGCGGCTTGTCACGTCGGTTGTGAAAGCCCGGGGCTTAACCC
CGGGTCTGCAGTCGATACGGGCAGGCTAGAGTGTGGTAGGGGAGATCGG
AATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAACACCGGT
GGCGAAGGCGGATCTCTGGGCCATTACTGACGCTGAGGAGCGAAAGCGT
GGGGAGCGAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGGTG
GGAACTAGGTGTTGGCGACATTCCACGTCGTCGGTGCCGCAGCTAACGCA
TTAAGTTCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGGA
ATTGACGGGGGCCCGCACAAGCAGCGGAGCATGTGGCTTAATTCGACGC
AACGCGAAGAACCTTACCAAGGCTTGACATATACCGGAAAGCATCAGAG
ATGGTGCCCCCCTTGTGGTCGGTATACAGGTGGTGCATGGCTGTCGTCAG
CTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGT

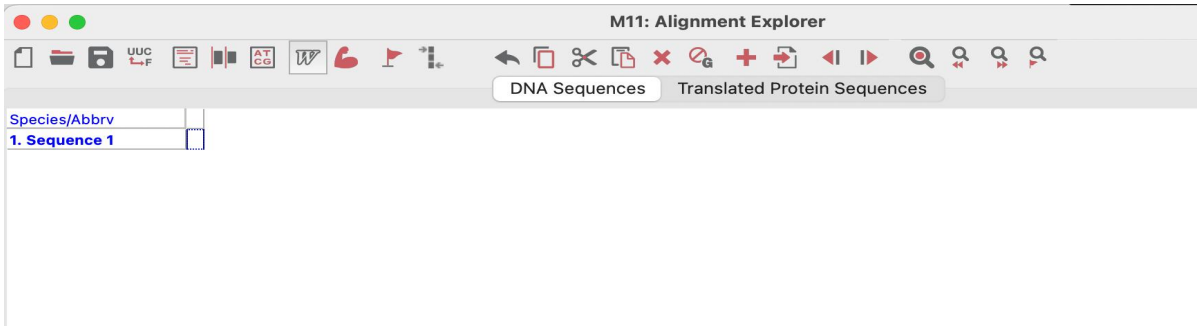
TCTGTGTTGCCAGCATGCCCTTCGGGGTGATGGGGACTCACAGGAGACTG
CCGGGGTCAACTCGGAGGAAGGTGGGGACGACGTCAAGTCATCATGCCC
CTTATGTCTTGGGCTGCACACGTGCTACAATGGCAGGTACAATGAGCTGC
GAAGCCGTAAGGCGGAGCGAATCTCAAAAAGCCTGTCTCAGTTCGGATTG
GGGTCTGCAACTCGACCCCATGAAGTCGGAGTTGCTAGTAATCGCAGATC
AGCATTGCTGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCA
CGTCACGAAAGTCGGTAACACCCGAAGCCGGTGGCCCAACCCCTTGTGGA
GGGACGCTCT

(Ця коротка послідовність із 2 гомологів).

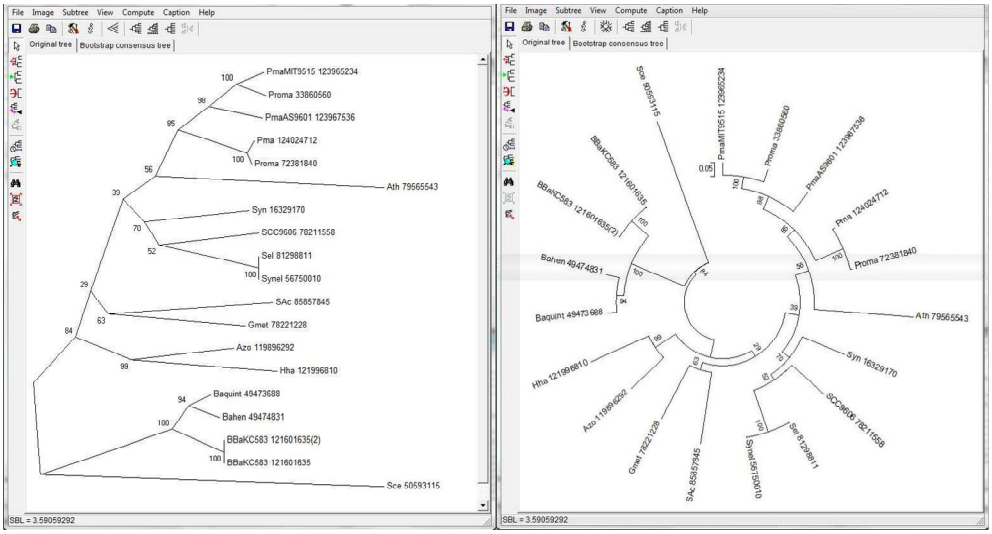
Тепер вставляємо скопійовану послідовність у програму MEGA:



У розділі EDIT. поміщаємо як нову послідовність Create, вибираємо DNA та вставляємо всі 27 пологів. Процес довгий. У вікні Sequence 1, що відкрилося, вставляємо наші послідовності:

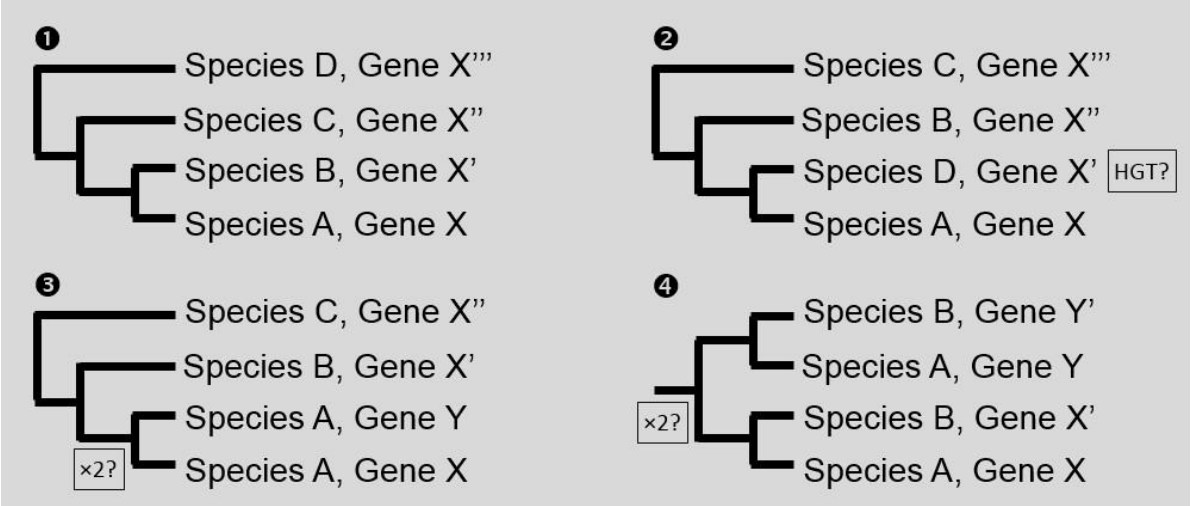


І вже за 10-15 хвилин можемо отримати алгоритм філогенетичного дерева. Далі вибираємо опцію Fylogeny, і через певний час готова візуалізаційна модель філогенетичного дерева.



Інтерпретація результатів філогенетичного аналізу.

Філогенетичний аналіз дуже потужний для визначення еволюційної історії гена, який нас цікавить. Розглянемо наступні чотири сценарії, і припустимо, що всі вузли мають хороші (високі рівні) бутстрап-підтримки. Гени X, X’, X’, X’’’ є ортологами, гени Y та Y’ є паралогічними до цих генів.



Припустимо, що види A, B, C і D будуть більш віддалено пов’язані один з одним, ніж знаходяться по алфавіту. У сценарії 1 ген випадає в очікуваній позиції, тобто дерево для гена аналогічно дереву видів. У сценарії 2 ген не випадає, де очікується, виходячи з дерева видів. Цей ген може бути надбаний при горизонтальному перенесенні (HGT – horizontal gene transfer) від видів (або близькоспоріднених видів), з якими він групується. У сценарії 3 існує паралог у вигляді A, але немає жодних паралогів у інших видів. Припускаючи, що геноми видів були відсеквеновані та поріг E-value не був занадто жорстким, це може

свідчити про часткову дуплікацію або дуплікації всього геному, позначається «×2?». У сценарії 4 є паралоги гена також у гомологів інших видів. Знову ж таки, з тими ж застереженнями гарного покриття геному і належного відсічення E-value (E-value cutoff), це може означати, що подія дуплікації відбулася у пращурів обох видів, у точці, позначеній «×2?».

Набуті вміння та навички:

1. знання термінології, побудова дендрограм і можливість ідентифікувати останнього загального пращура будь-яких двох термінальних вузлів (таксонів) на дереві;
2. знання основних елементів та термінології філогенії та можливі еволюційні шляхи до даного отриманого стану (як можуть виникати подібні форми (homoplasy?);
3. можливість ідентифікувати корінь дерева та знати різницю між укоріненими та некореневими деревами;
4. мати уявлення про методи філогенетичного аналізу, розуміти, як вони працюють, плюси та мінуси кожного з них;
5. познайомитися з різними моделями замін;
6. вміння будувати дерева методами об’єднання найближчих сусідів та максимальної правдоподібності, використовуючи MEGA.

Контрольні питання:

1. Опишіть, як зміна параметрів впливає реконструкцію філогенії.

2. Чи можна зробити обґрунтоване припущення, чому можна (або неможна) включити або виключити певні сайти, такі як в 3-й позиції в кодонах або некодуючих ділянках?

3. В Analysis > Phylogeny > Construct/TestMaximumLikelihood Tree встановити в Phylogeny Test / Test of Phylogeny to Bootstrap – 500 реплік, решту залиште за замовчуванням (проте збільшення кількості потоків може прискорити аналіз). Натисніть Compute, щоб запустити аналіз. Дайте результати.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Біоінформатика: Методичні вказівки до лабораторних робіт для студентів спеціальності 162 – Біотехнології та біоінженерія, спеціалізації «Молекулярна біотехнологія» / Уклад.: С. В. Горобець, О. Ю. Горобець, Є. А. Дарменко, М. О. Булаєвська. Київ: КПІ ім. Ігоря Сікорського, 2017. 52 с.
2. Chapter 7 «Recovering Evolutionary History» in Understanding Bioinformatics by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. P. 223–264.
3. Chapter 8 «Building Phylogenetic Trees» in Understanding Bioinformatics by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. P. 267–311.
4. Tamura K., Dudley J., Nei M., Kumar S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24(8):1596-9.
5. Doolittle W.F. (1999) Phylogenetic classification and the universal tree. Science 284: 2124-2128.
6. Page R.D.M., Charleston M.A. (1997) From gene to organismal phylogeny: reconcidle trees and the gene / species tree problem. Mol. Phylogenet. Evol. 7:231-240.
7. Saitou N., Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.
8. Guidon S., Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696-704.
9. Sitnikova T. (1996) Bootstrap method of interior-branch test for phylogenetic trees. Mol. Biol. Evol. 13:605-611.
10. http://www.megasoftware.net/
11. http://evolution.genetics.washington.edu/phylip.html
12. http://bar.utoronto.ca/webphylip/
13. http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::fastdnaml

Навчальне видання

КУЧМЕНКО О. Б., ПЕРЕХОДЬКО К. М.

БІОІНФОРМАТИКА

Навчально-методичний посібник

Технічний редактор – І. П. Борис
Верстка, макетування – О. В. Борщ

Книга друкується в авторському редагуванні.

Підписано до друку 03.07.23 р.	Формат 60x84/16	Папір офсетний
Гарнітура Times	Обл.-вид. арк. 1,7	Електронне вид-ня
Замовлення № 806	Ум. друк. арк. 3,55	

ВИДАВНИЦТВО НДУ ім.М.В.ГОГОЛЯ

Ніжинський державний університет
імені Миколи Гоголя.
м. Ніжин, вул. Воздвиженська, 3[А]
(04631) 7–19–72
E-mail: vidavn_ndu@ukr.net
www.ndu.edu.ua

Свідоцтво суб'єкта видавничої справи
ДК № 2137 від 29.03.05 р.