Міністерство освіти і науки України
Харківський національний аграрний університет ім. В.В. Докучаєва
Випробувальна лабораторія ТОВ «АГРОГЕН НОВО»

ОСНОВИ БІОІНФОРМАТИКИ

Навчальний посібник

Харків–2021

УДК [577.21:004](075.8)
О75

Рекомендовано до друку вченою радою ХНАУ ім. В.В. Докучаєва
(протокол № 5 від 25 травня 2021 р.)

Рецензенти: **О.М. Утевська**, доктор біологічних наук, професор кафедри генетики та цитології ХНУ ім. В.Н. Каразіна;
Ю. Є. Колупаєв, доктор біологічних наук, професор кафедри ботаніки та фізіології рослин ХНАУ ім. В.В. Докучаєва

Автори:
В.М. Попов, С.В. Лиманська, Г.Є. Чернишенко, Ю.М. Тереняк

О75 Основи біоінформатики: навч. посіб. / В.М. Попов, С.В. Лиманська, Г.Є. Чернишенко, Ю.М. Тереняк. – Харків: ХНАУ, 2021. – 108 с.

Висвітлено основні методи біоінформатики для дослідження генома рослин – алгоритми вирівнювання послідовностей ДНК, РНК і білків, філогенетичний аналіз, бази даних та комп'ютерні програми.

Призначено для здобувачів другого (магістерського) рівня спеціальності 201 «Агрономія» ОПП «Селекція і генетика сільськогосподарських культур», здобувачів третього (освітньо-наукового) рівня, викладачів, науковців.

УДК [577.21:004](075.8)

ЗМІСТ

ВСТУП

Біоінформатика як наука виникла у 80-х рр. XX ст. Її бурхливий розвиток пов'язаний не тільки з дослідженнями в області молекулярної генетики, але й з появою сучасних комп'ютерних технологій та їх інтеграцією з класичними науками – генетикою, селекцією, біохімією тощо. Сучасна біоінформатика займається системним аналізом нуклеотидних послідовностей ДНК, РНК, а також амінокислотних послідовностей і структурою білків. Важливим етапом розвитку біоінформатики стало секвенування геномів різних організмів – людини, рису, кукурудзи, пшениці, сої, соняшнику тощо.

Нині методів секвенування послідовностей ДНК дуже багато. Усі вони об'єднані під загальною назвою – next generation sequence (NGS), собівартість їх низька, через те нові дані щодо структури геномів з'являються швидко та інтегруються у різні бази даних з біоінформатики. Інформація, яка міститься в цих базах даних, дає новий поштовх до розвитку селекції, медицини, лабораторної діагностики, популяційної генетики та інших галузей, які використовують надбання молекулярної генетики. Водночас накопичені масиви даних щодо нуклеотидної будови окремих генів та цілих геномів живих істот потребують упорядкування, систематизації і статистичного аналізу. Часто виникає необхідність порівняти секвеновані послідовності генів у різних таксонів для встановлення еволюційних подій, зокрема інсерцій, делецій, транзицій і трансверсій, визначити закономірності філогенетичних зв'язків між таксонами, підібрати найбільш інформативні для аналізу окремих послідовностей праймери. Зручними інструментами для таких досліджень є різні комп'ютерні програми – BioEdit, MEGA, AmplifX тощо.

1. ПОНЯТТЯ ПРО БІОІНФОРМАТИКУ ТА ЇЇ ЗНАЧЕННЯ В СЕЛЕКЦІЇ РОСЛИН

Біоінформатика – це наука про моделювання процесів еволюції та оптимізації селекційного процесу з використанням методів прикладної математики й інформатики. Біоінформатика вивчає нуклеотидні послідовності ДНК та РНК, а також амінокислотні послідовності в структурі білків.

Мета біоінформатики:

1) організація масиву даних;
2) розробка комп'ютерних програм та інформаційних ресурсів;
3) аналіз даних та інтерпретація результатів.

Завдання біоінформатики:

1) визначення подібності нуклеотидних або амінокислотних послідовностей;
2) аналіз генома (визначення ділянок ДНК, які кодують білок, різні типи РНК, а також регуляторні ділянки);
3) передбачення вторинної структури РНК та білків;
4) філогенетичний аналіз;
5) створення та підтримання баз даних з геноміки живих організмів.

Накопичення масивів даних із секвенування геномів сільськогосподарських рослин дозволяє застосовувати інструменти біоінформатики для оптимізації селекційного процесу. Тепер селекціонерам доступна повна нуклеотидна послідовність генома пшениці озимої м'якої, кукурудзи, ячменю, гороху, сої, нуту, соняшнику, ріпака тощо. Цю інформацію можна використовувати для пошуку генів або генів-кандидатів агрономічних ознак та їх аналізу. Наприклад, у базах даних представлено нуклеотидну послідовність деяких генів стійкості до бурої іржі (*Lr*) та борошнистої роси (*Pm*), а також генів розвитку (*Vrn*) пшениці. Біоінформаційний аналіз генів якості дозволяє розширити уявлення про біосинтез, який забезпечує накопичення білка, антиоксидантів, різних жирних кислот та ін. У селекції рослин такий підхід дозволяє розробити принципово нову стратегію добору для покращання вихідного матеріалу.

У базах даних міститься інформація за геномними маркерами EST (Expressed Sequence Tag). EST являють собою часткові або

повні нуклеотидні послідовності генів, що експресуються, з відомою або передбаченою функцією. Їх розмір, як правило, 500–700 п.н. Ці послідовності доступні через GenBank або спеціалізовані бази даних. EST часто містять SNP, які потенційно можуть впливати на адаптивні ознаки або бути зчепленими з генами, які відповідають за ці ознаки. Для ідентифікації SNP можливий комп'ютерний пошук (*in silico*) в базах даних.

2. ВИРІВНЮВАННЯ ПОСЛІДОВНОСТЕЙ

Вирівнювання двох або більше послідовностей – розміщення однієї послідовності над іншою для визначення рівня їх ідентичності. Можна вирівнювати нуклеотидні послідовності ДНК, а також амінокислотні послідовності білків. Вирівнювання послідовностей спрямовано на ідентифікацію гомологічних ділянок в аналізованих послідовностях. Під час вирівнювання кодувальних нуклеотидних послідовностей бажано транслювати їх в амінокислотні послідовності, після чого провести вирівнювання на амінокислотному рівні. Під час аналізу послідовностей нуклеїнових кислот та білків користуються кодами IUPAC (табл. 1–2).

1. Коди IUPAC для нуклеїнових кислот

Код	Нуклеотиди	Код	Нуклеотиди
A	**A**denine	R	A або G (pu**R**ine)
C	**C**ytosine	Y	C або T (p**Y**rimidine)
G	**G**uanine	B	C або G або T
T	**T**hymine	D	A або G або T
W	A або T (**W**eak)	H	A або C або T
S	C або G (**S**trong)	V	A або C або G
M	A або C (a**M**ino)	N	будь-який нуклеотид
K	G або T (ke**T**o)	-	розрив (gap)

2. Коди IUPAC для амінокислот

Код	Амінокислота	Код	Амінокислота
A	аланін	N	аспарагін
B	аспартат або аспарагін	P	пролін
C	цистеїн	Q	глютамін
D	аспартат	R	аргінін
E	глутамат	S	серин
F	фенілаланін	T	треонін
G	гліцин	U	селеноцистеїн
H	гістидин	V	валін
I	ізолейцин	W	триптофан
K	лізин	Y	тирозин
L	лейцин	Z	глютамат або глютамін
M	метіонін		
*	стоп-кодон	X	будь-яка

Під час вирівнювання послідовностей ураховують мутаційні зміни (заміна одного нуклеотиду іншим, вставки та делеції нуклеотидів). Вставки та делеції (скорочено англійською називають INDEL від ***in****sertion* та ***del****etion*) при вирівнюванні називають розривами (gaps) та позначають "-". Біологічний смисл вирівнювання полягає в тому щоб записати одну послідовність над іншою таким чином, щоб гомологічні нуклеотиди були розташовані один над одним, а математичний – у тому, щоб знайти спосіб кількісної оцінки якості порівняння послідовностей макромолекул.

Розрізняють кілька типів вирівнювань (табл. 3).

3. Типи вирівнювання послідовностей

Тип вирівнювання	Опис
Парне вирівнювання (pair sequence alignment)	Вирівнюють дві послідовності
Множинне вирівнювання (multiple sequence alignment)	Вирівнюють три або більше послідовностей
Глобальне вирівнювання (global alignment)	Вирівнювання здійснюють за повною довжиною
Локальне вирівнювання (local alignment)	Вирівнювання тільки частини послідовності

Під час вирівнювання послідовностей необхідно отримати оптимальне вирівнювання, критерієм якого є сума всіх оцінок (score), які були введені при процедурі вирівнювання. Якщо нуклеотиди збігатимуться, то оцінки будуть мати високі (позитивні) значення, а за наявності розривів вводитимуть штрафи, які становитимуть менші, а іноді й негативні значення. Штрафи застосовуватимуть у таких випадках: незбіжність нуклеотиду, початок пробілу (gap opening penalty) та продовження пробілу (gap extension penalty). У разі отримання більших значень оцінок можна стверджувати, що певне вирівнювання є оптимальним. На рис. 1 показано дві амінокислотні послідовності до та після проведення вирівнювання.

Вирівнювання послідовностей ДНК та білків дозволяє визначити структуру та функції цих послідовностей. Наприклад, послідовності з геномів різних організмів, які мають високе значення подібності, ймовірно, виконують однакову функцію та мають спільного предка.

1 E E E L T K P R L L W A L Y F N M R D A L S S G
2 V E K P R I L Y A L Y F N M R D S S D E

1 E E E *L* *T* **K P R** L **L** *W* **A L Y F N M R D** A L **S S** *G* -
2 - - - *V* *E* **K P R** I **L** *Y* **A L Y F N M R D** - - **S S** *D* E

Рис. 1. Вирівнювання двох амінокислотних послідовностей 1 та 2

Примітка. Виділення жирним шрифтом – збіжність (match); виділення курсивом – незбіжність (mismatch); "-" – розриви (gap); підкреслені літери – консервативна заміна.

Під час вирівнювання послідовностей використовують такі алгоритми: принцип матриці точок (dot-matrix method) та динамічне програмування.

Принцип матриці точок. Використання цього методу полягає в тому, що одну послідовність записують у стовпчик, а другу – у рядок таблиці. Однакові нуклеотиди або амінокислотні послідовності записують у вигляді точок. Якщо дві послідовності є ідентичними, точки будуть розташовані в кожній комірці по діагоналі таблиці (рис. 2).

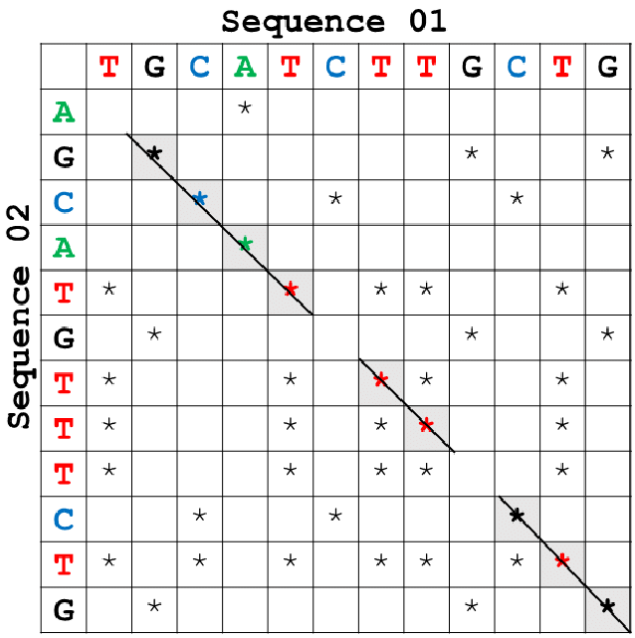


Рис. 2. Матриця точок

У вирівнюваних послідовностях можливі заміни та пробіли. За наявності замін у двох послідовностях точки в комірках будуть

розташовані по діагоналі таблиці, а у випадку наявності пробілів у послідовностях діагональ буде зміщатися праворуч або ліворуч.

Вирівнювання двох послідовностей методом динамічного програмування. Алгоритм вирівнювання двох послідовностей розроблено Нідлменом-Вуншем у 1970 р. Цей алгоритм використовують для вирівнювання й амінокислотних, і нуклеотидних послідовностей. Принцип цього алгоритму полягає в тому, що дві послідовності, які необхідно порівняти між собою, розташовують у таблиці-матриці по вертикалі та горизонталі. Під час вирівнювання двох послідовностей присвоюють штрафні бали, наприклад, при збіжності нуклеотидів – 1, незбіжності – -1, за розрив – також -1. У верхньому лівому куті ставлять нуль, а потім заповнюють поза матрицею колонки біля кожного нуклеотиду значеннями -1, -2, -3 тощо. Аналіз починають з верхнього лівого кута матриці та закінчують у правому нижньому куті цієї матриці.

Необхідно заповнити всю таблицю шляхом визначення для кожної комірки найбільшого можливого значення, використовуючи три типи переміщень: праворуч, униз і по діагоналі. Для цього до початкового числа з вертикальної, горизонтальної або діагональної комірок додають певний штрафний бал, залежно від того, чи збігаються нуклеотиди, і серед трьох отриманих значень обирають найбільше. Саме це значення заносять до таблиці-матриці. Воно може бути позитивним, негативним або дорівнювати нулю. Після заповнення всієї таблиці здійснюють прохід по матриці у зворотному напрямку за максимальними балами. Отриманий маршрут відповідає оптимальному вирівнюванню (рис. 3).

Алгоритм Нідлемана-Вунша використовують при глобальному вирівнюванні, тоді як при локальному застосовують алгоритм Сміта-Уотермена. Останній передбачає, що оптимальний шлях через матрицю може починатися та закінчуватися у будь-якій комірці. При пошуку подібної нуклеотидної послідовності в генетичних базах даних з аналізованою послідовністю широко застосовують саме алгоритм Сміта-Уотермена. Обидва ці алгоритми з математичного погляду спямовані на пошук оптимального вирівнювання при певних заданих параметрах.

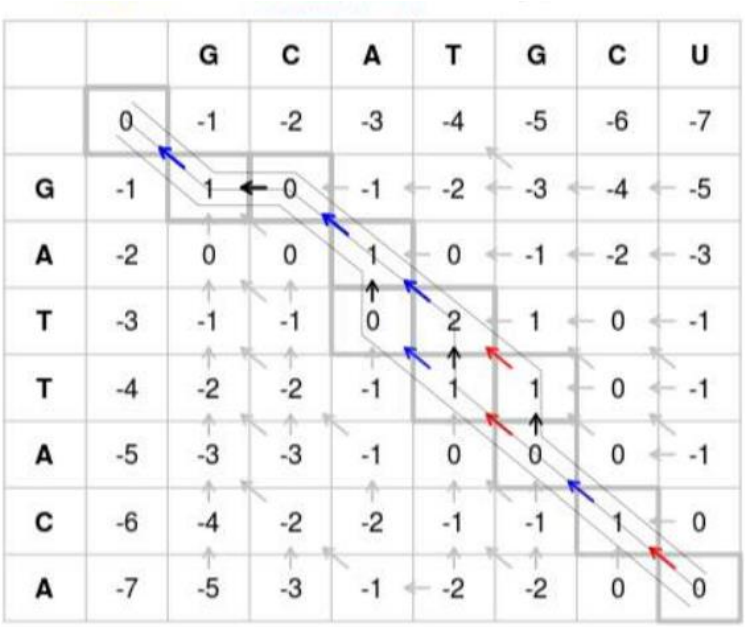


Рис. 3. Матриця для динамічного програмування

При множинному вирівнюванні послідовностей (multiple alignment) використовують алгоритм прогресивного множинного вирівнювання, який складається з трьох етапів: 1) усі послідовності вирівнюють між собою з подальшою ідентифікацією груп схожих між собою послідовностей; 2) проводять вирівнювання в кожній групі; 3) вирівнюють групи між собою.

Еволюційні моделі. Наступним етапом після вирівнювання нуклеотидних або амінокислотних послідовностей є розрахунок еволюційних дистанцій, які враховують такі мутаційні події: заміна нуклеотидів, вставки та делеції. Для розрахунку еволюційних дистанцій використовують такі еволюційні моделі.

1. ***Парні дистанції (****p*-distance***)***. В основу методу покладено розрахунок еволюційної дистанції як частки нуклеотидів, що не збігаються під час попарного порівнювання відповідних позицій у гені. Метод є недостатньо чутливим, оскільки не враховує можливості повторних мутацій в одній позиції і різну імовірність замін різних нуклеотидів.

2. ***Модель Jukes-Cantor*** – однопараметричний метод розрахунку еволюційних дистанцій, який ураховує частку нуклеотидів, що не збігається під час попарного порівнювання. Він базується на припущенні однакової частоти нуклеотидів (25 %) та однакової ймовірності заміщення в будь-якій парі нуклеотидів.

3. ***Модель Tajima-Nei:*** враховує неоднакову частоту чотирьох нуклеотидів (A, T, G, C) у послідовностях, а також різну імовірність замін цих нуклеотидів.

4. ***Модель Kimura 2-parameter*** – метод розрахунку еволюційних відстаней, який передбачає, що різні варіанти замін нуклеотидів мають різну імовірність та розглядає два типи замін: транзиції (A↔G, C↔T) і трансверсії (A, G ↔ C, T). Базується на припущенні, що частоти нуклеотидів дорівнюють 0,25 протягом усього еволюційного процесу.

5. ***Модель Tamura 3-parameter*** – метод розрахунку еволюційних відстаней, який ураховує можливість різних частот нуклеотидів у послідовностях і оцінює максимальну імовірність, яка найбільше відповідає кожному випадку.

6. ***Модель Tamura-Nei*** – модель розрахунку еволюційних відстаней, яка враховує, що рівень транзицій між пуринами (A і G) та піримідинами (T і C) часто відрізняється.

7. ***MCL-метод*** (maximum composite likelihood method, метод максимальної сумарної імовірності) – модель розрахунку еволюційних відстаней, придатна для опрацювання одночасно великої кількості вирівняних послідовностей. Точність методу підвищується зі збільшенням кількості досліджуваних послідовностей.

Основні принципи вибору певної моделі:

1) у випадках, якщо різні моделі дають однакові результати, варто використовувати простішу модель;

2) при дистанціях до 0,05 рекомендовано використовувати модель Jukes-Cantor; також її застосовують при збільшенні дистанції до 0,3 та незначному співвідношенні числа транзицій і трансверсій, наприклад, менше 2.

3) якщо частки транзицій і трансверсій суттєво відрізняються, а сама послідовність є достатньо довгою, то використовують модель Kimura 2-parameter;

4) якщо дистанція знаходиться в межах від 0,3 до 1, а нуклеотидна частка значно відрізняється та немає переваги між транзиціями і трансверсіями, то використовують модель Tajima-Nei, а при збільшенні співвідношення транзиції : трансверсії застосовують модель Tamura-Nei;

5) у випадку збільшення дистанції понад 1 жодна модель не буде давати адекватних результатів.

Розглянуті еволюційні моделі включено до різних комп’ютерних програм, наприклад MEGA.

Контрольні запитання

1. Що називають біоінформатикою?
2. Яка мета і завдання біоінформатики?
3. У чому полягає біологічний смисл вирівнювання послідовностей ДНК та білків?
4. Які існують типи вирівнювання?
5. Що називають кодами IUPAC? Яке їх значення?
6. Опишіть вирівнювання послідовностей методом матриці точок.
7. Дайте визначення динамічного програмування.
8. Які існують еволюційні моделі?

3. ФІЛОГЕНЕТИЧНИЙ АНАЛІЗ

Філогенетика – розділ еволюційної біології, що вивчає еволюційні зв'язки між різними формами життя. Молекулярна філогенетика – це встановлення еволюційних зв'язків між живими організмами на основі даних структури ДНК, РНК та білків. Результатом філогенетичного аналізу є побудова філогенетичного дерева.

Філогенетичний аналіз складається з таких етапів:

1) запис вихідних даних для обробки, що залежить від типу даних (дод. А);
2) розрахунок певної еволюційної моделі (див. розділ 2);
3) побудова філогенетичного дерева.

Філогенетичне дерево складається з внутрішніх та зовнішніх гілок та вузлів, а також кореня (якщо вибрана така опція при будуванні дерева). Особливість гілкування філогенетичного дерева називається його *топологією* (рис. 4). Об'єктами філогенетичного дослідження є гени або їх ділянки, нуклеотидні та амінокислотні послідовності, популяції, індивідууми тощо. Ці об'єкти називаються оперативними таксономічними одиницями – ОТО (OUT, operational taxonomic units).

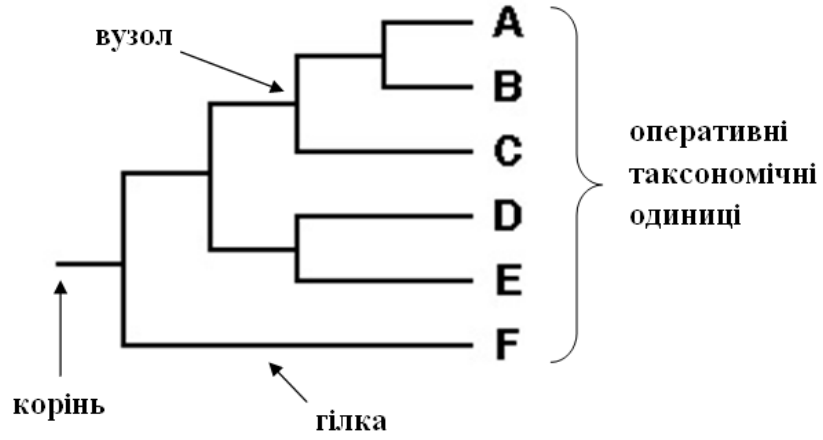


Рис. 4. Загальний вигляд філогенетичного дерева

Виділяють такі особливості філогенетичного дерева:

1. Філогенетичне дерево становить у загальному розумінні дендрограму, яку використовують для відображення еволюційних взаємозв'язків між оперативними таксономічними одиницями, і є гіпотезою, а не остаточним фактом.

2. Оперативні таксономічні одиниці розміщені на кінчиках гілок, а вузли на топології дерева відображають наявність спільних предків.

3. Довжина гілок філогенетичного дерева вказує на еволюційну відстань.

4. Філогенетичні дерева можна побудувати в різних еквівалентних стилях для кращої візуалізації отриманих результатів.

5. У філогенетичних деревах два види споріднені більшою мірою, якщо вони мають недавнього спільного предка, і меншою – якщо мають давнього спільного предка.

Типи філогенетичних дерев наведено в табл. 4.

Існують такі методи будування філогенетичних дерев:

1) дистанційні;

2) метод максимальної економії;

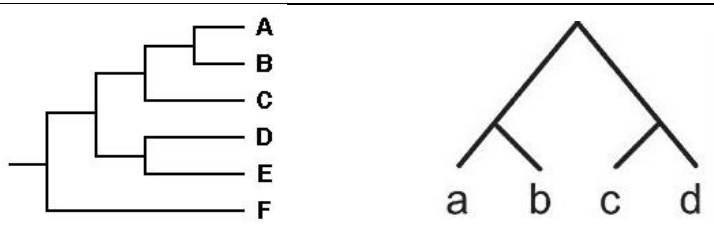
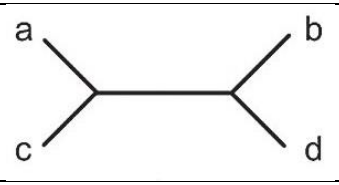
3) метод максимальної правдоподібності.

Дистанційні методи. На першому етапі необхідно встановити еволюційну відстань, використовуючи певну модель (див. розд. 2) і представити у вигляді матриці дистанцій (табл. 5). Дистанційні методи також дозволяють будувати дерева на основі інших даних – морфологічних, біохімічних та молекулярних з розрахунком коефіцієнтів подібності або генетичних відстаней. Ідея генетичної відстані належить L. Sanghvi. У середині XX ст. він запропонував вимірювати дистанції між популяціями математичною величиною. У розрахунку коефіцієнтів подібності існує така закономірність: високі значення свідчать про більшу подібність, а для генетичних відстаней – навпаки. Наразі існує безліч генетичних відстаней, які наведено в дод. Б.

Оцінку цих параметрів також можна використовувати для підбору пар у селекції автогамних та алогамних сільськогосподарських культур. Отримані результати щодо розрахунку генетичних відстаней між вихідним матеріалом рослин свідчили про таку закономірність – чим вище значення розрахованих показників, тим більше вихідний матеріал – сорти, інбредні лінії – віддалені один від одного; відповідно, схрещування генетично віддалених сортів або інбредних ліній може приводити до збільшення генетичного різноманіття.

Одержані значення генетичних відстаней використовують для групування ОТО за допомогою кластерного аналізу, який дозволяє візуалізувати результати у вигляді спеціального графіка – дендрограми.

4. Типи філогенетичних дерев

Назва	Вид	Характеристика
Дерево орієнтоване, ієрархічне, укорінене		На дереві вказано основу та визначено направлення переходів між вершинами. Такі дерева в філогенетиці відображають напрям еволюції
Дерево неорієнтоване, неукорінене		Дерево, на якому не вказано основу. Такі дерева показують тільки родинні зв'язки між аналізованими об'єктами
Дерево узгоджене	Може бути представлене у вигляді укоріненого або неукоріненого дерева	Дерево, отримане після проведення статистичного аналізу для оцінки достовірності топології дерева з використанням бутстреп-аналізу або методу «складного ножа» тощо. Дерево будують на основі одного масиву даних
Супердерево		Дерево будують на основі дерев, які отримано за результатами аналізу різних масивів даних (наприклад, морфологічні, біохімічні та молекулярні дані)

5. **Матриця парних дистанцій для t оперативних таксономічних одиниць (ОТО)**

ОТО	1	2	3	…	t
1	-	d_{1-2}	d_{1-3}	…	d_{1-t}
2	d_{2-1}	-	d_{2-3}	…	d_{2-t}
3	d_{3-1}	d_{3-2}	-	…	d_{3-t}
…	…	…	…	…	…
t	d_{t-1}	d_{t-2}	d_{t-3}	…	-

Кластерний аналіз – це метод багатомірної статистики, який дозволяє упорядкувати об’єкти в кластери або групи подібних об’єктів. Основними методами кластерного аналізу є ієрархічний та К-середніх (дод. В). Наприклад, з погляду генетичних даних кластеризацію можна розуміти як появу в просторі частот генів, ОТО з вищою щільністю, ніж в інших областях цього простору. Кластеризація – універсальний метод, який можна використовувати для будь-яких генетичних та селекційних вихідних даних. Альтернативою використання в генетико-селекційних дослідженнях кластерного аналізу є факторний аналіз, який описано в дод. Г.

Для визначення кластерів використовують різні методи – метод незваженого попарного групування із середнім арифметичним, (UPGMA, unweighted pair group method with arithmetic mean), метод приєднання сусідів (NJ neighbor-joining method) та ін.

Метод незваженого попарного групування із середнім арифметичним. Метод базується на використанні послідовного групування об’єктів. При цьому дерево будують у кілька етапів. На першому етапі визначають об’єкти з максимальною подібністю, які надалі розглядають як один об’єкт. Наступним етапом є приєднання до них об’єкта, який має найбільшу подібність (рис. 5).

Метод приєднання сусідів. Під час використання цього методу класифікації у кластер об’єднують об’єкти, які мають найменшу суму серед усіх гілок дерева, тобто враховують також довжину решти гілок, на відміну від методу UPGMA, де в кластер об’єднують об’єкти, які мають найменшу дистанцію між собою незалежно від довжини інших гілок. Побудова дерева методом NJ починається зі створення дерева із зореподібною топологією. Далі розглядають кожну можливу пару об’єктів. З усіх можливих обирають ту пару об’єктів, яка має найменшу суму довжин гілок. Далі цю пару розглядають як один зразок, розраховують нову матрицю відстаней, обирають нову пару об’єктів, яка дає найменшу суму довжин гілок дерева (рис. 6).

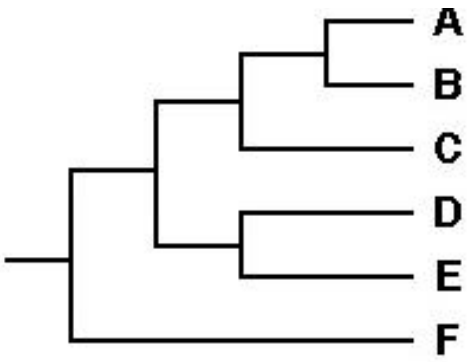


Рис. 5. Укорінене дерево, побудоване методом UPGMA

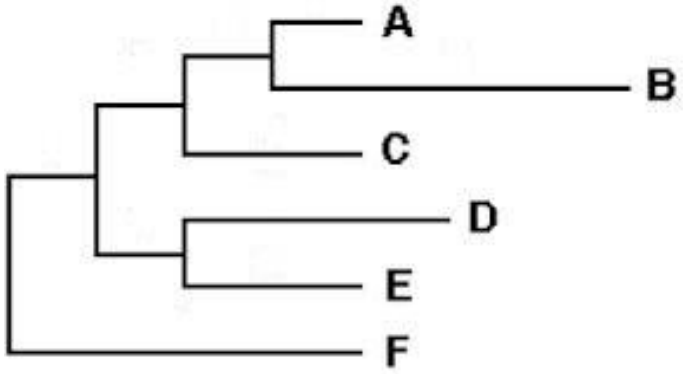


Рис. 6. Неукорінене дерево, побудоване методом NJ

Метод максимальної економії (parsimony). Метод був розроблений Едвардсом та Каваллі-Сфорца для заданих частот генів. Цей метод ураховує безпосередньо значення ознаки, яку ідентифіковано в ОТО, наприклад, різницю між послідовностями ДНК в певній позиції. Але він не враховує еволюційних відстаней. Метод максимальної економії спрямовано на пошук філогенетичного дерева, яке включає найменше число нуклеотидних або амінокислотних замін.

Метод максимальної правдоподібності. Метод був розроблений Едвардсом та Каваллі-Сфорца для заданих частот генів і пізніше був модифікований Фельзенштейном. Цей метод ураховує еволюційні моделі для будування філогенетичного дерева. На перших етапах визначають певну топологію дерева, а потім довжину гілок з подальшим розрахунком величини правдоподібності. Отримані для різних філогенетичних дерев статистичні величини правдоподібності порівнюють між собою та добирають дерево з максимальною правдоподібністю.

Бутстреп-аналіз. Статистичний метод, який дозволяє оцінити достовірність топології філогенетичного дерева за допомогою створення великої кількості випадкових вибірок, у яких значення однієї або декількох ознак розподілені випадково. Зазвичай під час

проведення бутстреп-аналізу створюють від 100 до 1000 випадкових вибірок. Після цього для кожної вибірки будують дерево із використанням тих же методів, що і для аналізованого дерева. Усі отримані дерева порівнюють із побудованим раніше, визначаючи, чи є у цих дерев ті ж самі внутрішні вузли, що й у дерева, яке аналізують. Для кожного внутрішнього вузла вихідного дерева підраховують відсоток випадкових дерев, у яких є той самий вузол. Цей відсоток називають значенням бутстреп-аналізу та записують поряд із вузлом (рис. 7). Як правило, достовірно встановленими вузлами вважають ті, для яких значення бутстреп-аналізу перевищує 70.

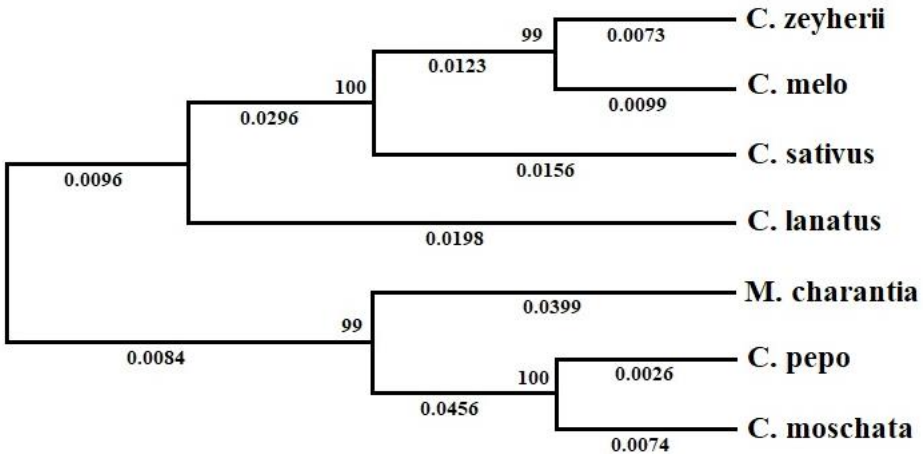


Рис. 7. Філогенетичне дерево, яке побудовано після бутстреп-аналізу

Контрольні запитання

1. Дайте визначення поняття філогенетичний аналіз.
2. Опишіть основні етапи філогенетичного аналізу.
3. Які типи філогенетичних дерев використовують для встановлення взаємовідносин між оперативними еволюційними одиницями?
4. Охарактеризуйте дистанційні методи будування дерев.
5. У чому полягає принцип будування дерев методом максимальної економії та максимальної правдоподібності?
6. Дайте визначення поняття кластерний аналіз.
7. Дайте визначення поняття бутстреп-аналіз.
8. У чому полягає принцип факторного аналізу?

4. ГЕНОМНІ БАЗИ ДАНИХ

База даних (БД) – комп'ютерна система збереження, пошуку та видачі необхідної інформації. БД з геноміки створюють для накопичення, збереження, систематизації та аналізу великих масивів інформації. Основними БД є інформаційні ресурси, де накопичується інформація про всі живі організми (табл. 6).

6. Характеристика основних баз даних

Назва	Опис
GenBank http://www.ncbi.nlm.nih.gov	Містить і підтримує архів нуклеотидних послідовностей, а також надає програми та інформаційні ресурси для інтеграції з іншими біологічними даними. Інтегрована з DDBJ та EMBL
UniProt, The Universal Protein Resource https://www.uniprot.org/	Універсальний ресурс про білки
GrainGenes http://wheat.pw.usda.gov/GG3	БД для родин *Triticeae* та *Avena,* є джерелом фенотипової і молекулярної інформації для пшениці, ячменю, жита, вівса та інших споріднених видів пшениці
Plant Genome Database http://www.plantgdb.org/	БД про геноми різних видів рослин. Містить геномні послідовності і транскрипти, а також дані з порівняльної геноміки
Phytozome http://phytozome.jgi.doe.gov/	БД з порівняльної генетики рослин (пшениця, соняшник, ячмінь, амарант тощо)
MaizeGDB http://www.maizegdb.org/new_genes	Основна база даних з генетики та геноміки кукурудзи. Містить дані з функціональної геноміки, хромосомні карти, інформацію про гени та їх алелі, молекулярні маркери та опис фенотипів
MAS wheat http://maswheat.ucdavis.edu	Містить інформацію про ДНК-маркери до різних генів агрономічних ознак пшениці

До таких БД належать GenBank (розроблена National Centre for Biotechnology Information, NCBI), DDBJ (DNA Data Bank of Japan), ENA (European Nucleotide Archive), які пов'язані між собою.

Найбільшою БД з протеоміки є UniProt, яка проводить анотацію усіх білкових послідовностей і структурно-функціональну організацію білків. Створено і такі БД, які інтегрують інформацію з геноміки різних видів рослин, наприклад, GrainGenes, Plant Genome Database, Phytozome, а також БД для окремих видів MaizeGDB (для кукурудзи) та MASwheat (для пшениці).

Користувач у різних БД може знайти інформацію про: 1) геноміку, протеоміку і транскриптоміку; 2) експресію генів; 3) генетичні, фізичні та хромосомні карти; 4) ДНК-маркери; 5) фенотипічні ознаки; 6) посилання на наукові статті.

БД поділяють на такі типи: 1) архівні (інформацію розміщують безпосередньо вчені, які за неї відповідають, наприклад, до архівних БД відносять GenBank); 2) БД, які мають кураторів (за зміст записів у таких БД відповідають куратори, а інформацію відбирають експерти з архівних БД, наприклад, SwissProt); 3) автоматичні (у таких БД записи моделюють спеціальні комп'ютерні програми, наприклад, БД UniProt); 4) інтегровані (такі БД об'єднують інформацію із різних БД, наприклад, Entrez).

БД контролюються різними організаціями, одним з лідерів яких є Національний центр біотехнологічної інформації (National Centre for Biotechnology Information, NCBI). Його було засновано в 1988 р. у США при Національній медичній бібліотеці. NCBI надає інформацію про різні бази даних: нуклеотидні послідовності ДНК (GenBank) та амінокислотні послідовності білків (Protein), таксономічну інформацію про певні біологічні види (Taxonomy), базу даних статей наукової літератури (PubMed) тощо. Усього NCBI налічує 35 баз даних. Вони доступні через пошукову систему Entrez.

Для більшості традиційних сільськогосподарських культур України, зокрема ячменю, пшениці, кукурудзи, соняшнику, завдяки міжнародним проектам з вивчення їх геномів накопичено численні дані щодо структури та функцій останніх. Основні БД стосовно ячменю наведено в табл. 7.

Згідно з накопиченими даними повтори в хромосомах розподіляються таким чином: дистальна частина характеризується наявністю великої кількості низькокопійних елементів, високим вмістом генів і високою частотою мейотичних рекомбінацій; середня зона має середню щільність генів; проксимальна частина має мінімальну кількість генів, переважно це гени домашнього господарства, які є дуже консервативними.

7. Бази даних ячменю

База даних	Розробник	Посилання
Ensembl Plants	EBI (European Bioinformatics Institute, UK)	http://plants.ensembl.org
BARLEX	IPK (Leibniz Institute of Plant Genetics and Crop Plant Research, Germany)	http://barlex.barleysequence.org
IPK barley web BLAST	IPK (Leibniz Institute of Plant Genetics and Crop Plant Research, Germany)	http://webblast.ipk-gatersleben.de/ barley
Barley GenomeZipper	PGSB (Plant Genome and Systems Biology unit at the Helmholtz Center Munich, Germany)	http://pgsb.helmholtz-muenchen.de/ plant/barley/gz/download/index.jsp
CrowsNest homepage	PGSB (Plant Genome and Systems Biology unit at the Helmholtz Center Munich, Germany)	http://pgsb.helmholtz-muenchen.de/ plant/crowsNest/
RNA-Seq data of barley cultivar Morex	JHI (James Hutton Institute, Scotland)	https://ics.hutton.ac.uk/morexGenes/
TCAP homepage	TCAP (Triticeae Coordinated Agricultural Project, USA)	http://triticeaetoolbox.org
International Database for Barley Genes and Barley Genetic Stocks	NordGen (Nordic Genetic Resource Centre, Nordic Institute, Iceland)	http://www.nordgen.org/bgs/index. php?pg=bgs_tables&m=loc

Для селекціонерів важливі гени стійкості до біотичних та абіотичних факторів середовища, гени продуктивності тощо, розташовані в дистальних частинах, для яких характерна висока частота рекомбінації. Порівняння з бібліотекою повторів, специфічною для *Triticeae,* виявило, що приблизно 80 % генома ячменю належить до мобільних елементів.

У міру зниження собівартості секвенування ДНК зросла кількість ідентифікованих однонуклеотидних поліморфізмів – SNP, які потенційно зчеплені або асоційовані з генами агрономічних ознак. Зокрема, для ячменю розроблено ДНК-чип 50K Infinium iSelect, який дозволяє одночасно проаналізувати 6000 SNP. Використання таких ДНК-чипів дозволяє проводити масовий пошук асоціацій між

фенотипом та генотипом, цей аналіз називається повногеномний пошук асоціацій (GWAS, genome wide association analysis). Наразі завдяки цьому методу виявлено геномні райони, які асоційовані з агрономічними ознаками ячменю – маса 1000, кількість зерен у колосі, висота рослини, стійкість до посухи та засолення, вміст білка, β-глюкану та мікроелементів у зерні тощо.

У геномних БД часто міститься детальна інформація про фенотипові градації морфологічних ознак рослин. Така інформація може бути отримана за унікальним ідентифікатором гена або алеля. Наприклад, у БД MaizeGDB за запитом «purple» будуть знайдені всі можливі аномалії органів кукурудзи зі зміною забарвлення з нормального на фіолетовий. При виборі певної фенотипової градації буде видано список алелей, а для кожного алеля – надано його характеристику та опис мутантного фенотипу з обов'язковим посиланням на джерело інформації (автор, організація або наукова стаття).

База даних GenBank (www.ncbi.nlm.nih.gov/genbank). БД GenBank розроблено National Centre for Biotechnology Information, NCBI у 1982 р. GenBank організовано за біологічним принципом, основним елементом є нуклеотидна послідовність нуклеїнової кислоти або амінокислотна послідовність білка.

Ця БД складається з таких елементів:

1) послідовність і її характеристика – послідовність нуклеїнової кислоти без пропусків. Вона являє собою текст, який складається з літер (*a, c, g, t*). Описуючи РНК, уридин позначають літерою *t*. Невизначеності за нуклеотидами прописано в стандартному коді IUPAC. Також указують довжину, тип нуклеїнової кислоти. Крім того, наводять інформацію про організм, стадію розвитку, тип тканини, номер клону тощо, які було використано при секвенуванні нуклеотидної послідовності;

2) бібліографічна інформація – містить джерело інформації;

3) біологічна анотація – тут розміщують інформацію про продукт, який кодується нуклеїновою кислотою, та вказують тип продукту (наприклад, білок або тРНК), розмір кодуючої області і функціональних сайтів (мутації, сайти рекомбінації, регуляторні послідовності тощо).

Елементи даних організовано в такі записи та поля (табл. 8).

8. Елементи даних БД

Поле	Опис
LOCUS	Ідентифікатор запису, який містить номер, розмір і тип нуклеїнової кислоти, а також дату реєстрації
DEFENITION	Містить визначення нуклеотидної послідовності
ACCESSION	Містить список номерів, за якими здійснюється доступ
VERSION	Версія нуклеотидної або амінокислотної послідовності
KEY WORDS	Ключові слова
SOURCE	Назва організму
ORGANISM	Таксономічна інформація
REFERENCE	Містить інформацію про те, яку послідовність було взято з наукової роботи
AUTHORS	Список авторів
TITLE	Назва літературного джерела
JOURNAL	Містить посилання на журнал або на інше джерело
PUBMED	Містить номер посилання в базі літературних джерел PubMed
REFERENCE	Додаткова інформація
AUTHORS	
TITLE	
JOURNAL	
FEATHURE	Опис функціональних районів нуклеотидної послідовності: CDS (**co**d**i**ng **s**equences, область, що кодується); 3’-UTR (3′-untranslated region, 3’-область, що не транслюється)
ORIGIN	Нуклеотидна послідовність, яку представлено згідно зі стандартом кодування

Приклад запису гена *Pm3 Triticum aestivum* L. наведено на рис. 8.

Формат FASTA. У БД усі послідовності нуклеїнових кислот і білків записано у форматі FASTA (**FAST** – швидкий та **A**ligment – вирівнювання). Це основний формат для читання послідовностей та оформлення результатів.

Рядок починається із символу >, за яким стоїть ідентифікатор запису та опис кислоти. У форматі FASTA використовують однолітерні коди для нуклеотидів та амінокислот відповідно до UPAC. На рис. 9 наведено нуклеотидну послідовність гена *Pm3*, а на рис. 10 показано амінокислотну послідовність білка цього гена у форматі FASTA.

GenBank ▾ Send t

Triticum aestivum powdery mildew resistance protein (Pm3) mRNA, Pm3-d allele, partial cds

GenBank: AY605285.1

FASTA Graphics

Go to:

```
LOCUS       AY605285                759 bp    mRNA    linear   PLN 26-JUL-2016
DEFINITION  Triticum aestivum powdery mildew resistance protein (Pm3) mRNA,
            Pm3-d allele, partial cds.
ACCESSION   AY605285
VERSION     AY605285.1
KEYWORDS    .
SOURCE      Triticum aestivum (bread wheat)
  ORGANISM  Triticum aestivum
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BOP
            clade; Pooideae; Triticodae; Triticeae; Triticinae; Triticum.
REFERENCE   1  (bases 1 to 759)
  AUTHORS   Srichumpa,P., Brunner,S., Keller,B. and Yahiaoui,N.
  TITLE     Allelic series of four powdery mildew resistance genes at the Pm3
            locus in hexaploid bread wheat
  JOURNAL   Plant Physiol. 139 (2), 885-895 (2005)
   PUBMED   16183849
REFERENCE   2  (bases 1 to 759)
  AUTHORS   Srichumpa,P., Yahiaoui,N. and Keller,B.
  TITLE     Direct Submission
  JOURNAL   Submitted (23-APR-2004) Institute of Plant Biology, University of
            Zurich, Zollikerstrasse, 107, Zurich 8008, Switzerland
FEATURES             Location/Qualifiers
     source          1..759
                     /organism="Triticum aestivum"
                     /mol_type="mRNA"
                     /cultivar="Kolibri"
                     /db_xref="taxon:4565"
                     /tissue_type="leaf"
                     /dev_stage="10 days"
     gene            <1..759
                     /gene="Pm3"
                     /allele="d"
     CDS             <1..357
                     /gene="Pm3"
                     /allele="d"
                     /note="Pm3d"
                     /codon_start=1
                     /product="powdery mildew resistance protein"
                     /protein_id="AAT38871.1"
                     /translation="RLPAPLKRLFIMGNSGLTSLECLSGEHPPSLESLWLERCSTLAS
                     LPNEPQVYRSLWSLEIRGCPAIKKLPRCLQQQLGSIKRKWLDARYEVTEFKPLKPKTW
                     KEIPRLVRERRQACRS"
     3'UTR           358..759
                     /gene="Pm3"
                     /allele="d"
ORIGIN
        1 cgtctgcctg cacccctcaa gagactgttc attatgggca acagtgggct gacatcgctg
       61 gagtgtctgt cgggagagca ccccccatcg ctggaatccc tttggcttga aagatgcagt
      121 accctggcat ccctgccgaa tgagccgcaa gtatacaggt ctctctggtc tcttgaaatt
      181 agaggctgcc ctgctataaa gaagctccct agatgcctgc agcagcaact gggcagcatc
      241 aaacgcaaat ggctagatgc ccgttatgaa gtaacggaat tcaaaccatt gaaaccgaag
      301 acatggaagg aaataccgag gctagtccgt gagcggaggc aggcctgccg gagctgaagc
      361 tgaaggtcat tcagcttcta ttcaatgaca ggcctgatcc ctgattgttg cactgctctg
      421 ctctgctctg ctaggggaat ggcctgtgat tgcttctaca gctgaactgt gccgcagtcc
      481 gctctcggtc gagaactgtc ggggcctgcg gtgtagggcg gctgctgctg cttggcgttg
      541 agtcgggcat ctctgaatca acatcctagg cccagaccat ctcatattga tctggtttgc
      601 tgccgcctgc tggcccttgg ggcgtgtaca acattcctgt catgtagccc aagggttgtt
      661 gtatctgaat agcttctttgc accatacaat gtttaatttg tacaccaaaa actaggggaa
      721 ggataaaaaa tggttgtttt aaaaaaaaaa aaaaaaaaa
//
```

Рис. 8. Опис гена *Pm3 Triticum aestivum* L. з бази даних GenBank

```
>AY605285.1 Triticum aestivum powdery mildew resistance protein (Pm3) mRNA,
Pm3-d allele, partial cds
CGTCTGCCTGCACCCCTCAAGAGACTGTTCATTATGGGCAACAGTGGGCTGACATCGCTGGAGTGTCTGT
CGGGAGAGCACCCCCCATCGCTGGAATCCCTTTGGCTTGAAAGATGCAGTACCCTGGCATCCCTGCCGAA
TGAGCCGCAAGTATACAGGTCTCTCTGGTCTCTTGAAATTAGAGGCTGCCCTGCTATAAAGAAGCTCCCT
AGATGCCTGCAGCAGCAACTGGGCAGCATCAAACGCAAATGGCTAGATGCCCGTTATGAAGTAACGGAAT
TCAAACCATTGAAACCGAAGACATGGAAGGAAATACCGAGGCTAGTCCGTGAGCGGAGGCAGGCCTGCCG
GAGCTGAAGCTGAAGGTCATTCAGCTTCTATTCAATGACAGGCCTGATCCCTGATTGTTGCACTGCTCTG
CTCTGCTCTGCTAGGGGAATGGCCTGTGATTGCTTCTACAGCTGAACTGTGCCGCAGTCCGCTCTCGGTC
GAGAACTGTCGGGGCCTGCGGTGTAGGGCGGCTGCTGCTGCTTGGCGTTGAGTCGGGCATCTCTGAATCA
ACATCCTAGGCCCAGACCATCTCATATTGATCTGGTTTGCTGCCGCCTGCTGGCCCTTGGGGCGTGTACA
ACATTCCTGTCATGTAGCCCAAGGGTTGTTGTATCTGAATAGCTTCTTGCACCATACAATGTTTAATTTG
TACACCAAAAACTAGGGGAAGGATAAAAAATGGTTGTTTTAAAAAAAAAAAAAAAAAAA
```

Рис. 9. Нуклеотидна послідовність гена *Pm3* у форматі FASTA

```
>AAT38871.1 powdery mildew resistance protein, partial [Triticum aestivum]
RLPAPLKRLFIMGNSGLTSLECLSGEHPPSLESLWLERCSTLASLPNEPQVYRSLWSLEIRGCPAIKKLP
RCLQQQLGSIKRKWLDARYEVTEFKPLKPKTWKEIPRLVRERRQACRS
```

Рис. 10. Амінокислотна послідовність гена *Pm3* у форматі FASTA

Пошук у базах даних NCBI. Усі бази даних NCBI пов’язані між собою, на них існують посилання. Для пошуку використовують систему Entrez (https://www.ncbi.nlm.nih.gov/search/).

Можна здійснити пошук одразу в усіх базах даних NCBI чи вибрати певну базу даних і зробити пошук за ключовими словами або за ідентифікатором запису. Між базами даних також є посилання.

Для ефективного пошуку необхідної інформації можна використовувати логічні оператори між термінами, які пишуть великими літерами: 1) AND – у випадку, якщо в одному документі необхідно знайти два або більше термінів (цей оператор використовувати не обов’язково, оскільки система автоматично його додає, якщо не вказаний інший оператор); 2) OR – якщо в документі потрібно знайти хоча б один з термінів; 3) NOT – у випадку, якщо тільки один з термінів потрібно знайти, а другий обов’язково має бути відсутнім.

Програми серії BLAST. До пакета програм BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) входять програми для локального вирівнювання між експериментальною послідовністю (query) та послідовністю з бази даних (subject). Цей пакет було розроблено в 1990 р. американським математиком Стефаном Альтшулем. BLAST розміщено на сервері NCBI.

До пакета програм входять такі основні програми:

1. Геномні – для порівняння аналізованої нуклеотидної послідовності з базою даних секвенованого генома певного виду, наприклад *Arabidopsis thaliana*.

2. Нуклеотидні – для порівняння аналізованої нуклеотидної послідовності з базою даних секвенованих послідовностей:

- blastn – повільне порівняння для пошуку усіх схожих послідовностей;
- megablast – швидке порівняння для пошуку дуже схожих послідовностей;
- dmegablast – швидкий пошук схожих, але не ідентичних послідовностей.

3. Білкові – для порівняння амінокислотної послідовності досліджуваного білка з базою даних білків та їх ділянок:

- blastp – повільне порівняння для пошуку усіх схожих послідовностей;
- psi-blast – порівняння для пошуку послідовностей, які мають незначну схожість.

Алгоритм програми BLAST оснований на тому, що експериментальну нуклеотидну або амінокислотну послідовність розбивають на короткі сегменти ідентичних знаків, які називаються словами, – по 11 нуклеотидів для нуклеотидних послідовностей і по три амінокислоти для білкових послідовностей (позначають літерою *W*). Далі ці слова використовують у процесі сканування бази даних інших послідовностей, які містять ідентичні або схожі слова. Кожне збігання або подібність перевіряють із встановленням оцінки, яка відповідає певному пороговому значенню – *T*. Якщо оцінка відповідає пороговому значенню, то послідовність починає продовжуватися в обох напрямах з нарахуванням балів за збіг та штрафів за введення розривів. Подовження послідовності триватиме доти, доки оцінка для цього вирівнювання не опуститься нижче від заздалегідь визначеного порога. Для оцінки значущості вирівнювання програма розраховує такі статистичні критерії – вагу вирівнювання (score) та *E*-value. Чим більша вага вирівнювання, тим вища схожість двох послідовностей. Якщо $E \leq 0{,}02$ – послідовності є гомологічними; $0{,}02 \leq E \leq 1$ – гомологія нечітка; $E > 1$ – випадковий збіг.

База даних UniProt. UniProt (Universal Protein Resource, універсальний білковий ресурс) – вільнодоступний інтернет-ресурс, який містить інформацію щодо послідовностей та функцій білків.

Інтернет-адреса: https://www.uniprot.org/.

UniProt було розроблено консорціумом UniProt, який складається з груп Європейського інституту біоінформатики (European Bioinformatics Institute, EBI), Інформаційного ресурсу білків (Protein Information Resource, PIR) та Швейцарського інституту біоінформатики (Swiss Institute of Bioinformatics, SIB).

За допомогою UniProt можна виконувати такі завдання:

1. Пошук наукової літератури про білки.
2. Аналізування широкого спектра експериментальних даних, які пройшли ручну анотацію.
3. Використання інструментів, таких як BLAST, вирівнювання послідовностей білків, конвертація ідентифікаторів білків тощо.
4. Пошук посилань на інформацію більше ніж 140 біологічних баз даних.
5. Переглядання та завантажування повних наборів протеомів.

Структура UniProt

1. UniProt Knowledgebase (UniProtKB) – центральна база даних інтегрованої інформації про білки. Кожний запис про білок обов'язково містить послідовності амінокислот, назву або опис білка, таксономічні дані та інформацію про цитування. Крім того, зазначають якомога більше додаткової інформації: прийняті біологічні онтології, класифікації, перехресні посилання та чіткі вказівки якості анотації у вигляді посилання на експериментальні та обчислювальні дані.

База даних UniProt складається з двох розділів:

1) *UniProtKB/Swiss-Prot* – розділ, який переглядає та анотує вручну куратор. У UniProtKB/Swiss-Prot записи з інформацією, вилученою з літератури, також перевіряє та аналізує куратор;

2) *UniProtKB/TrEMBL* – розділ з автоматично проаналізованими записами, які очікують повної ручної анотації.

Понад 95 % послідовностей білків, наданих у UniProtKB, отримано з кодувальних послідовностей (CDS), які опубліковано в загальнодоступних базах даних нуклеїнових кислот EMBL-Bank/GenBank/DDBJ databases (INSDC). Усі ці послідовності, а

також відповідні дані, подані авторами, автоматично інтегруються в UniProtKB/TrEMBL.

UniProt має чотири компоненти, які призначено для різних цілей (рис. 11).

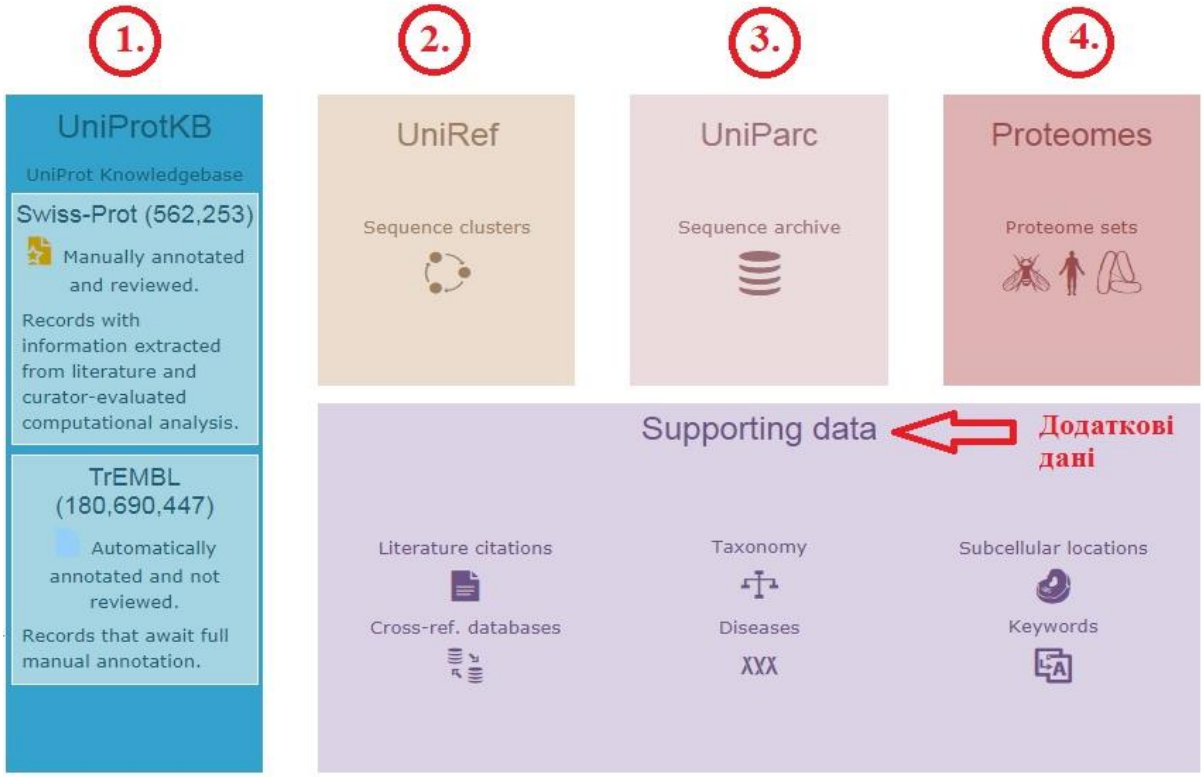


Рис. 11. Структура UniProt

2. The UniProt Archive (UniParc) – архів UniProt, всеосяжна база даних, яка містить більшість загальнодоступних білкових послідовностей у світі. Білки можуть існувати в різних початкових базах даних і в декількох копіях в одній базі даних. UniParc уникнув такої надмірності, зберігаючи кожну унікальну послідовність лише один раз і надаючи їй стабільний і унікальний ідентифікатор (UPI), що дозволяє ідентифікувати той самий білок з різних баз даних. UPI ніколи не знімають і не змінюють. UniParc містить лише послідовності білків. Усю іншу інформацію про білок можна отримати з вихідних баз даних, використовуючи перехресні посилання. UniParc відстежує зміни послідовності у вихідних базах даних та архівує історію всіх змін. UniParc об'єднав багато баз даних в одну на рівні послідовності, а пошук UniParc рівнозначний пошуку багатьох баз даних одночасно. У наш час UniParc містить білкові послідовності з 22 загальнодоступних баз даних.

3. UniProt Reference Clusters (UniRef) – еталонні кластери UniProt, згруповані набори послідовностей з UniProtKB (включаючи ізоформи) та вибраних записів UniParc для отримання повного покриття простору послідовностей у кількох роздільних здатностях. Фрагменти послідовностей білків об'єднують у UniRef таким чином:

1) база даних UniRef100 об'єднує однакові послідовності й субфрагменти з 11 і більше залишками амінокислот від будь-якого організму в єдиний запис UniRef, відображаючи послідовність репрезентативного білка, номери приєднання всіх злитих записів та посилання на відповідні записи UniProtKB й UniParc;

2) UniRef90 побудований шляхом кластеризації послідовностей UniRef100 з 11 або більше залишками за допомогою алгоритму MMseqs2 (Steinegger M. та Soeding J., Nat. Commun. 9 (2018)), так що кожен кластер складається з послідовностей, що мають щонайменше 90 % ідентичності і на 80 % перекриваються з найдовшою послідовністю кластера;

3) UniRef50 будують шляхом кластеризації послідовностей UniRef90, які мають щонайменше 50 % ідентичності та 80 % перекриття з найдовшою послідовністю в кластері.

4. Протеоми (Proteomes results) – це набори білків, які експресуються організмом. UniProt надає протеоми для видів з повністю секвенованими геномами.

За допомогою інструменту «Фільтр» (Filter by) можна обрати референтні протеоми – еталонні протеоми, які охоплюють добре вивчені модельні організми та інші організми для біомедичних та філогенетичних досліджень.

Додаткові дані (Supporting data)

1. Пошук джерел літератури (Literature citations results) – дає змогу шукати публікації, які цитуються в UniProtKB, за ключовими словами, ім'ям автора, назвою журналу, роком видання.

2. Пошук у перехресних базах даних – у цьому розділі відображають посилання на інші бази даних (нуклеотидних послідовностей, модельних видів, геномні та протеомні ресурси) для запису UniProtKB. Один запис може мати перехресні посилання на кілька десятків різних баз даних і кілька сотень індивідуальних посилань.

3. Пошук за таксономічною назвою (Taxonomy) – за результатами пошуку відображають усі записи UniProtKB для

конкретного виду організму. За допомогою фільтрів можна відсортувати протеоми, записи, які анотуються вручну та автоматично.

4. Хвороби людини (Human diseases) – цей розділ надає інформацію про захворювання, пов'язані з генетичними варіаціями цього білка. Інформація надходить з джерел наукової літератури. Хвороби також описано в базі даних OMIM (Online Mendelian Inheritance in Man).

5. Пошук локалізації білка в клітині (Subcellular location) – цей підрозділ надає інформацію про розташування і топологію зрілого білка в клітині.

6. Пошук за ключовим словом (Keywords results) – усі записи UniProtKB позначені ключовими словами, за допомогою яких можна отримати певні підмножини записів.

Існує 10 категорій ключових слів: 1) біологічний процес; 2) клітинний компонент; 3) різноманіття кодувальних ділянок; 4) етап розвитку; 5) захворювання; 6) домен; 7) ліганд; 8) молекулярна функція; 9) посттрансляційна модифікація; 10) технічний термін.

КОРИСТУВАННЯ UNIPROT

Текстовий пошук. Для пошуку в UniProt слід виконати такі дії (рис. 12):

1. Вибрати відповідний розділ UniProt (вибір за замовчуванням – UniProt KB).
2. Ввести свій запит і натиснути опцію пошуку.
3. Також можна використати розширений варіант пошуку, натиснувши опцію «**Advanced**». Інтерфейс розширеного пошуку являє собою набір декількох полів, у які можна ввести текст або вибрати дані з розкривних списків. Після заповнення полів для здійснення пошуку необхідно натиснути опцію «**Search**» (розпочати пошук).

Рис. 12. Текстовий пошук UNIPROT

Сторінка результатів UniProt. Після здійснення пошуку користувач побачить сторінку у вигляді таблиці, де відображено результати (рис. 13).

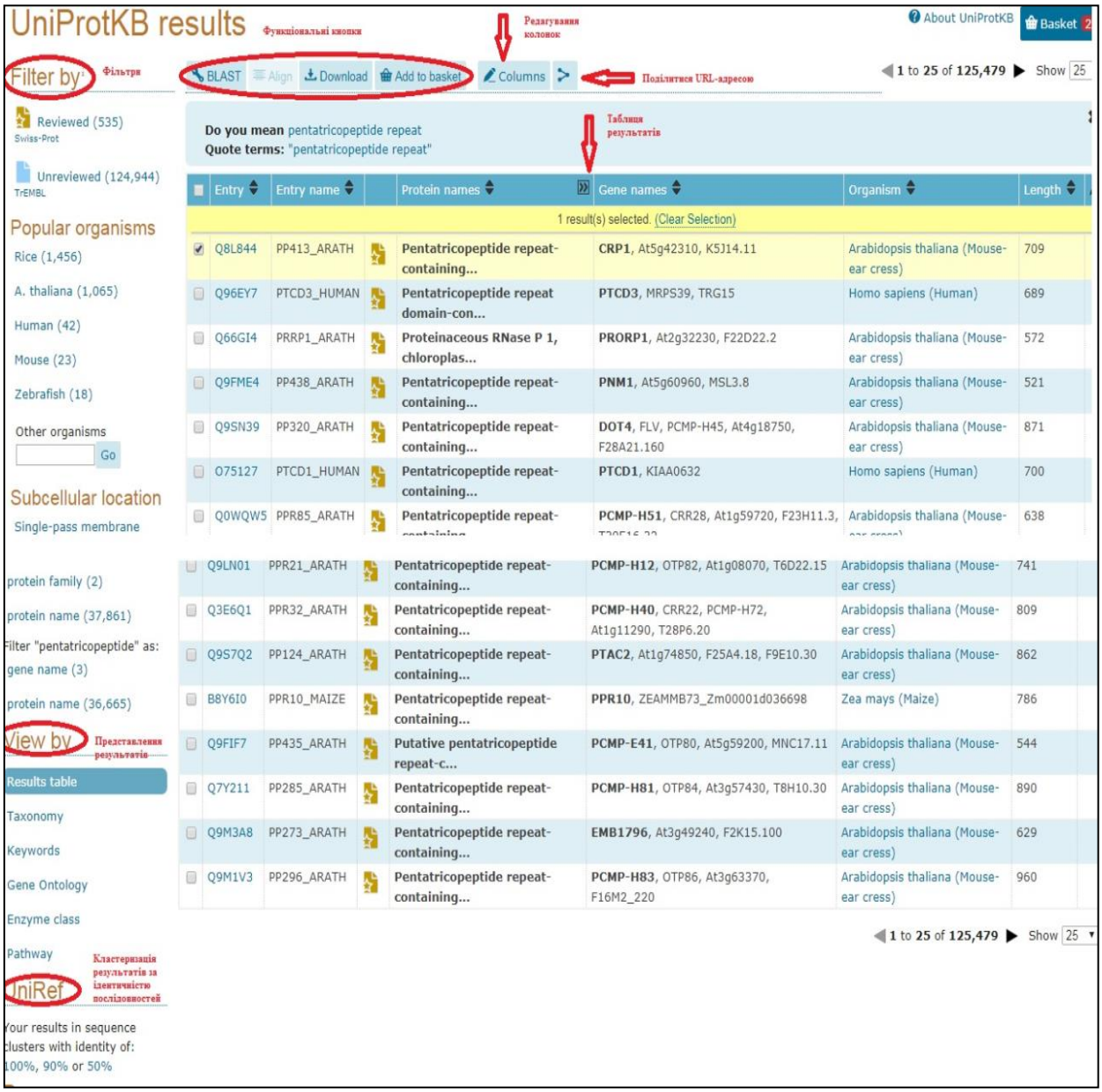


Рис. 13. Таблиця результатів UniProtKB

Стовпці таблиці (Columns). У таблиці результатів відображено таку інформацію: ідентифікатор запису, назва запису, стан анотації (вручну або автоматично переглянутий запис), назви білків, назви генів, організм та довжина послідовності. Можна налаштувати таблицю результатів, відредагувавши стовпці, щоб додати або видалити інформацію.

Фільтри (Filter by) та відображення результатів (View by). Можна відфільтрувати результати ***(Filter by)*** за статусом анотації, організмом та іншими критеріями залежно від завдань пошуку.

За допомогою інструмента ***View by*** можна змінити представлення результатів. Наприклад, "**погляд за систематикою"** (***«Taxonomy»)*** показує дерево таксономії всіх організмів, знайдених у наборі результатів. Кількість результатів про організм відображають у дужках поруч з назвою організму. Натиснувши на номер, ви перейдете до цієї підмножини результатів пошуку у вікні результатів UniProtKB. Натиснувши на назву організму, ви перейдете на сторінку таксономії, що описує організм та його походження.

Інструмент UniRef дозволяє кластеризувати результати за ідентичністю послідовностей білків. Застосовують у випадку, якщо основний пошук було проведено у UniProtKB.

Функціональні клавіші. Зі сторінки результатів можна запустити пошук схожих ділянок послідовностей (BLAST), вирівнювання послідовностей (Align), завантажити записи результатів у різних форматах, поділитися URL-адресою таблиці результатів, зберегти записи для подальшого використання, додавши їх у свій кошик.

Використання кошика UniProt. Для збереження результатів пошуку UniProt з метою подальшого перегляду та аналізу можна використовувати ***"Кошик"*** (***«Basket»)*** (рис. 14). Для додавання у кошик необхідно вибрати записи та опцію ***"Додати у кошик"*** (***«Add to basket»***). Після додавання записів до кошика під ідентифікатором запису з'явиться значок кошика. При цьому число у кошику змінюється, що свідчить про додавання нових білків. При натисканні опції кошика розгортається інформація про збережені білки. Кошик складається з трьох окремих вкладок для записів UniProtKB, UniRef та UniParc. У кошику є функції, які дозволяють вибирати в ньому білки, щоб вирівняти їх, запустити BLAST-пошук або завантажити їх у різних форматах.

Для видалення записів у кошику по черзі необхідно натиснути ***«Видалити***» (***«Delete»)*** у кожному стовпці або скористатися опцією ***«Очистити»*** (***«Clear»***), щоб видалити всі збережені записи. Вміст кошика залишатиметься незмінним, доки не будуть видалені файли cookie з браузера або користувач не очистить кошик самостійно.

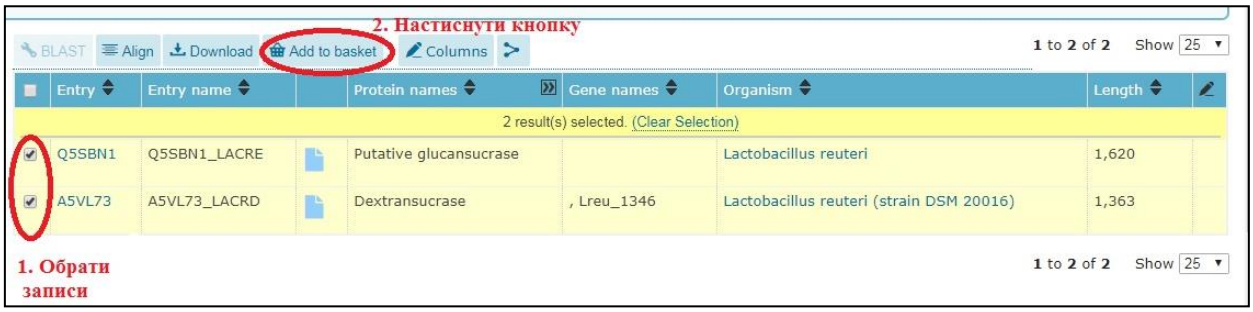


Рис. 14. Використання кошика

Інструменти Uniprot. UniProt надає чотири інструменти для аналізу даних білків. До всіх чотирьох можна отримати доступ із посилань у заголовку веб-сайта (рис. 15).

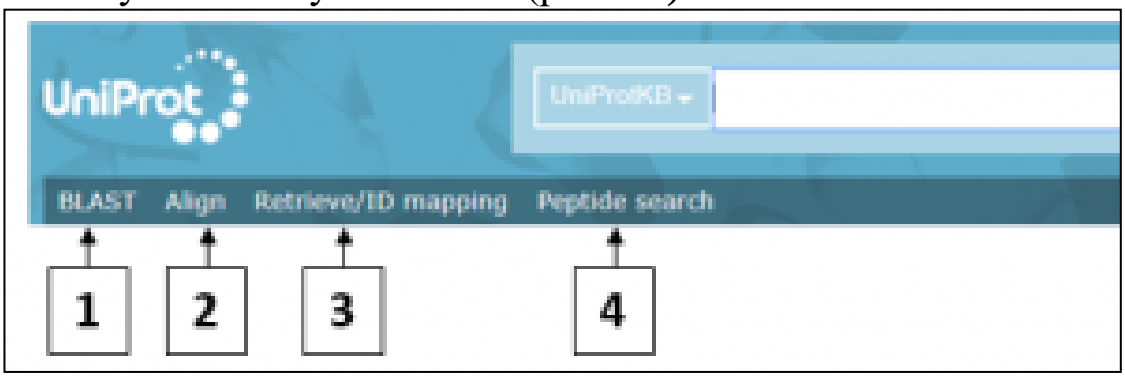


Рис. 15. Інструменти Uniprot:

1 – пошук подібності послідовностей (BLAST); 2 – вирівнювання послідовностей білків (Align); 3 – отримання / конвертація ідентифікаторів білків (Retrieve/ID mapping); 4 – пошук пептидів (Peptide search)

Розглянемо детальніше ці інструменти.

1. Пошук подібності послідовностей (BLAST).

Пошук подібності послідовностей визначає ділянки локальної схожості між послідовностями, які можна використовувати для дослідження функціональних та еволюційних зв'язків між послідовностями, а також для ідентифікації членів родини генів.

Як користуватися

1. Виберіть вкладку BLAST на панелі інструментів, щоб запустити пошук подібності послідовностей з програмою BLAST.
2. Введіть білкову, нуклеотидну послідовність (необроблена послідовність або формат FASTA) чи ідентифікатор UniProt у поле форми.
3. Натисніть опцію «Run BLAST».

Якщо запустити BLAST на панелі інструментів у UniProtKB, UniRef або UniParc, поточна послідовність автоматично заповнюється у формі.

2. Вирівнювання послідовностей білків (Align).

Вирівнювання послідовностей білків у Uniprot здійснюють за допомогою програми Clustal Omega.

Як користуватися

1. Виберіть вкладку «***Вирівняти***» (***«Align»***) на панелі інструментів, щоб вирівняти дві або більше послідовностей білка.

2. Введіть у поле форми або послідовності білків у форматі FASTA, або UniProt. Натисніть опцію ***«Виконати вирівнювання»*** (***Run Align***).

Якщо необхідно використати іншу програму для вирівнювання послідовностей, натисніть клавішу ***«Завантажити» («Download»)***, щоб зберегти інформацію для подальшого використання.

Запустити вирівнювання білків можна також з результатів пошуку у UniProtKB, UniRef або UniParc. Для цього необхідно поставити позначки біля відповідних білків у таблиці результатів та натиснути опцію ***«Вирівняти» («Align»)*** (рис. 16).

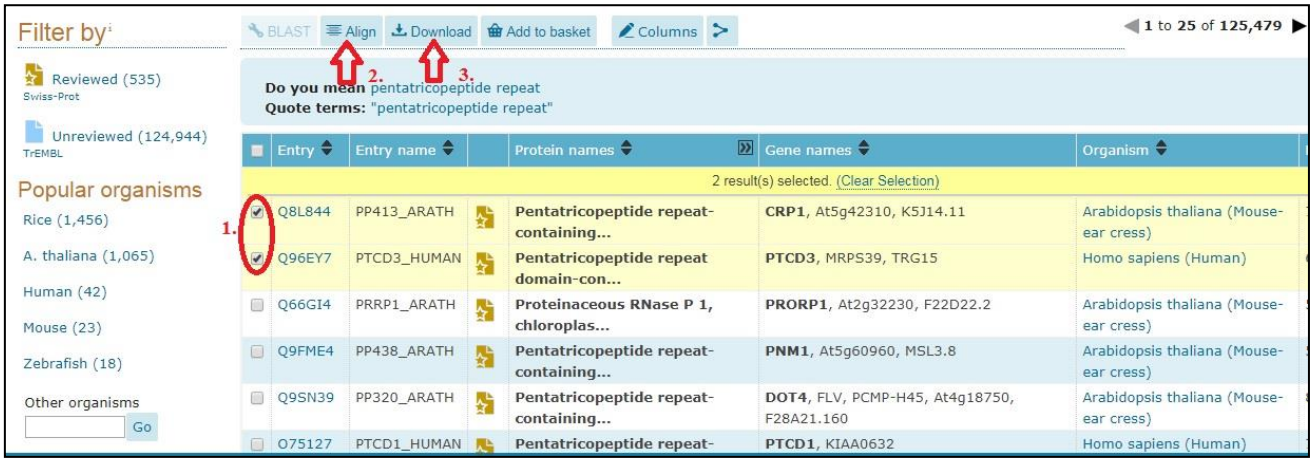


Рис. 16. Запуск вирівнювання білків з таблиці результатів UniProtKB, UniRef або UniParc:

1 – обрати відповідні записи; 2 – натиснути «Align»; 3 – для вирівнювання в інших програмах натиснути «Download»

3. Отримання / конвертація ідентифікаторів білків (Retrieve/ID mapping).

Цей інструмент дозволяє:

Отримати відповідні записи UniProt, щоб завантажити їх або працювати з ними на веб-сайті UniProt.

Здійснити конвертацію ідентифікаторів іншого типу в ідентифікатори UniProt або навпаки та завантажити списки ідентифікаторів.

Як користуватися

1. Введіть ідентифікатори, розділені пробілами або новими рядками, у поле форми, наприклад: P31946 P62258, ALBU_HUMAN, EFTU_ECOLI.
2. Якщо вам потрібно перетворити на інший тип ідентифікатора, виберіть вихідний та цільовий тип ідентифікатора зі спадного меню.
3. Натисніть опцію «***Надіслати***» (***«Submit»***).

4. Пошук пептидів (Peptide search).

Інструмент пошуку пептидів дозволяє подати запит пептидних послідовностей щонайменше з трьох залишків і знайти всі послідовності UniProtKB, які мають точну відповідність послідовності запитів.

Як користуватися

1. Для доступу до інструменту натисніть на посилання «***Пошук пептидів***» (**«*Peptide search*»**) у заголовку, який розміщено вгорі кожної сторінки на веб-сайті UniProt.
2. Введіть пептидну послідовність, що має принаймні три амінокислоти, у поле пошуку. Інструмент дає змогу вибрати обмеження таксономії, а також аналізувати ізолейцин та лейцин як еквіваленти (Treat Isoleucine and Leucine as equivalent).
3. Натисніть «***Запустити пошук пептидів***» («***Run peptide search***»).

Пошук SPARQL

UniProt надає кілька інтерфейсів прикладного програмування (API) для запиту та доступу до бази даних програмно. Один з таких API – пошук SPARQL, посилання на який розміщено в заголовку веб-сайта (рис. 17). SPARQL здійснює пошук для всіх даних UniProt, що зберігаються у форматі Resource Description Framework (RDF).

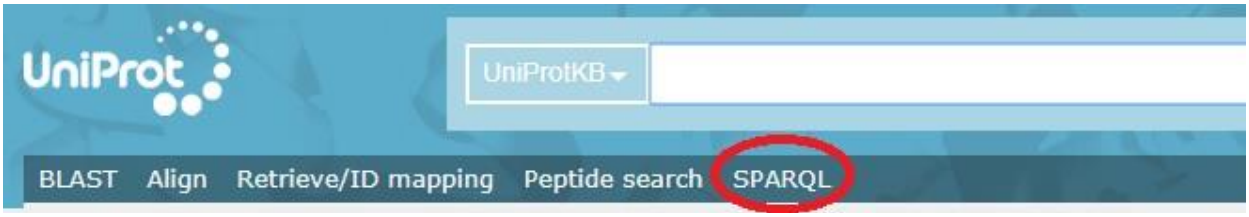


Рис. 17. Пошук SPARQL

Контрольні запитання

1. Що називають базою даних?
2. Опишіть класифікацію баз даних з біоінформатики.
3. Дайте визначення поняття «серія програм BLAST».
4. Дайте визначення поняття «формат FASTA».
5. Охарактеризуйте базу даних GenBank.
6. Опишіть базу даних UniProt.

5. КОМП'ЮТЕРНІ ПРОГРАМИ ДЛЯ РОБОТИ З БІОЛОГІЧНИМИ ПОСЛІДОВНОСТЯМИ

Програма BioEdit

BioEdit – це редактор вирівнювання біологічних послідовностей (ДНК, РНК, поліпептидів).

Програму BioEdit створено Томом Хеллом у 1999 р. Сьогодні це один з найпоширеніших багатоцільових ресурсів, який використовують у молекулярно-біологічних дослідженнях.

BioEdit має інтуїтивно зрозумілий інтерфейс, містить багато функцій для різних режимів вирівнювання послідовностей та легкої маніпуляції ними. BioEdit характеризується автоматичною інтеграцією з іншими програмами, такими як ClustalW та Blast.

У навчальному посібнику розглянуто функції BioEdit для проведення вирівнювання послідовностей і деяких маніпуляцій з ними. Проте можливості BioEdit цим не обмежуються. Програма також дозволяє розраховувати генетичні відстані, будувати філогенетичні дерева з використанням різних методів філогенетичного аналізу (NJ, максимальної подібності, максимальної економії), переводити послідовності нуклеїнових кислот у поліпептиди та білкові послідовності в нуклеотидні, через програму BLAST здійснювати пошук подібних послідовностей у базах даних тощо. Детальнішу інформацію щодо можливостей і алгоритмів роботи у BioEdit можна знайти за посиланням: http://darwin.uvigo.es/dposada/repository/pdfs/BioEdit.help.pdf (матеріали від розробника програми Т. Хелла).

BioEdit підтримує різні формати файлів, які зазвичай використовують в інших біоінформаційних програмах. Це дозволяє обмінюватися файлами даних між BioEdit та іншим програмним забезпеченням. Серед форматів, які приймає BioEdit, такі: розширений текстовий формат *.rtf, файли fasta (*.fas, *.fasta, *.fst, *.fsa), файли генбанку (*.gbk, *.gen, *.gb, *.gnk), файли форматів *.csv, *.txt, *.excel, файл хроматограми формату abi (*.ab1, *.ab), файл послідовності *.seq, плазмідні файли *.pmd, файли проекту bioedit – *.bio та ін. У програмі можна легко запустити новий документ, скопіювати (Ctrl + C) або вставити (Ctrl + V) дані з буфера обміну. До недоліків програми також можна віднести те, що час аналізу суттєво збільшується зі збільшенням розміру і кількості

одночасно досліджуваних послідовностей; для використання деяких функцій потрібен певний досвід.

Установка програми. Програма BioEdit є загальнодоступною. Її можна безкоштовно завантажити з багатьох серверів:

- http://www.mbio.ncsu.edu/BioEdit/bioedit.html;
- http://www.mbio.ncsu.edu/bioedit/page2.html;
- http://en.bio-soft.net/format/BioEdit.html;
- http://www.mybiosoftware.com/alignment/1013;
- http://www.etoology.net/index.php/software/genetics/100-bioedit-709.html.

Після завантаження програму слід установити на жорсткому диску, використовуючи установчий файл. Програма сумісна з більшістю версій Windows.

Робота з програмою. Розглянемо можливості програми BioEdit на прикладі вирівнювання нуклеотидних послідовностей господарсько цінних генів *Wx,* які кодують білок Waxy, що впливає на утворення крохмалю амілопектинового типу у запасаючих тканинах рослин. Ці гени присутні в геномі зернових і, зокрема, злакових культур. Виникає питання: наскільки подібні гени *Wx* у різних видів у межах одного роду, а також у представників різних родів? Вирівнювання секвенованих послідовностей указаних генів дозволить відповісти на поставлене питання.

Вхідні файли. Роботу слід починати з пошуку послідовностей ДНК у літературних джерелах або базах даних, наприклад NCBI, MAS Wheat, GRIN та ін.

Після біоінформаційного пошуку формуємо базу вхідних даних. Такими можуть бути файли у форматі *.txt, а також файли, завантажені з баз даних.

Назви файлів формату *.txt потрібно вводити латинськими літерами. Сам файл повинен містити послідовність, у якій нуклеотиди або амінокислоти також записані без проміжків латинськими великими або малими літерами загальноприйнятими позначеннями: а) для нуклеотидів: A – аденін, G – гуанін, T – тимін, C – цитозин, M – A/C, R – A/G, W – A/T, S – C/G, Y – C/T, K – G/T, N – невідомий нуклеотид; б) для амінокислот: A – аланін, R – аргінін, N – аспарагін, D – аспарагінова кислота, V – валін тощо (детальніше – за посиланням: http://kodomo.cmm.msu.ru/~youthofchemist/projects/Term1/Amin oAcid/index.html).

Щоб завантажити файл із бази даних (розглянемо на прикладі бази NCBI), необхідно в ній відкрити сторінку з бажаною послідовністю, обрати функцію «**Send to**», у графі «**Choose Destination**» обрати призначення «**File**», установити формат «**GenBank**» і зберегти послідовність, натиснувши «**Create File**» (рис. 18).

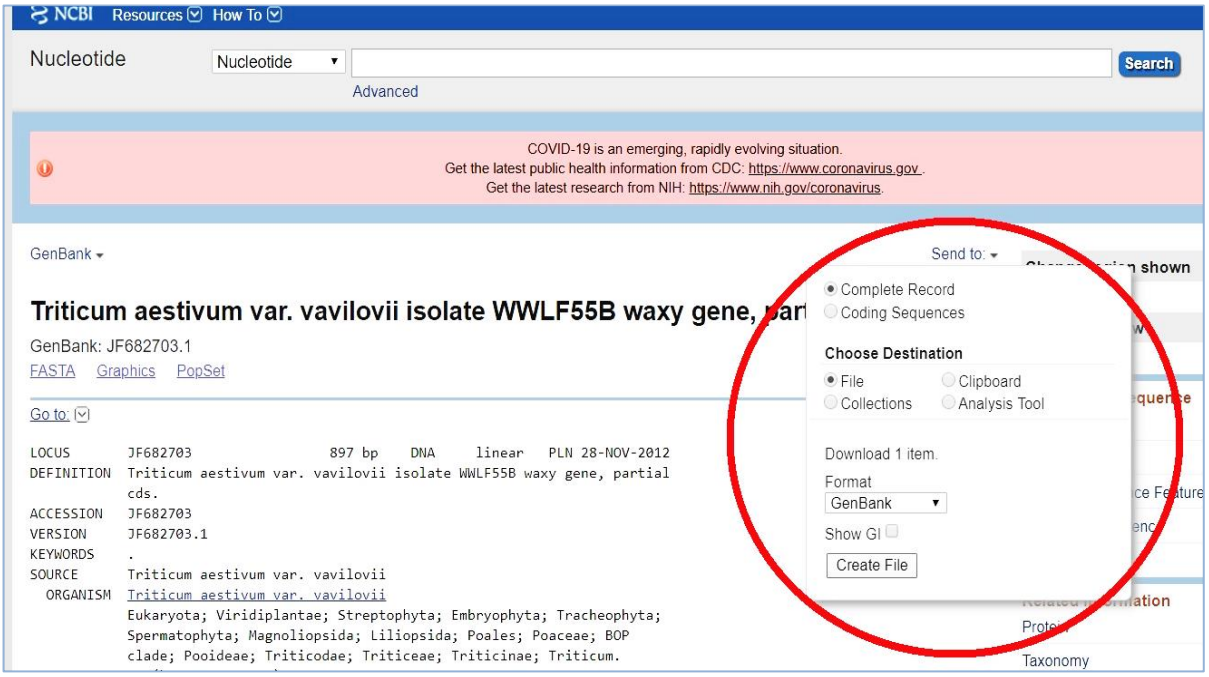


Рис. 18. Збереження послідовності ДНК з бази даних NCBI

Імпортування даних у програму і початок роботи. Щоб розпочати роботу в програмі BioEdit, потрібно відкрити цю програму, двічі натиснувши на значок програми у папці, у яку її було завантажено. Коли з'явиться вікно з програмою, відкрити вкладку **File** на панелі задач, обрати опцію **New Alignment** – якщо розпочинаємо новий проект, або опцію **Open** – якщо потрібно відкрити вже створений документ (рис. 19). У робочій області програми з'явиться вікно з інструментами.

Потім знову натискаємо File на панелі задач, обираємо опцію Import, і далі – Sequence alignment file (рис. 20).

У діалоговому вікні обираємо потрібні файли із вхідними даними. Зазначимо, що для зручності всі файли з досліджуваними послідовностями слід розміщувати в одній папці. За один раз оптимально аналізувати 20–50 послідовностей. Розробник програми Т. Хелл вказує на можливість одночасної обробки 50 документів з 20000 послідовностей необмеженого розміру в

кожному. Але час розрахунків програми за умови збільшення кількості одночасно оброблюваних ланцюжків сильно збільшуватиметься, програма може «зависнути».

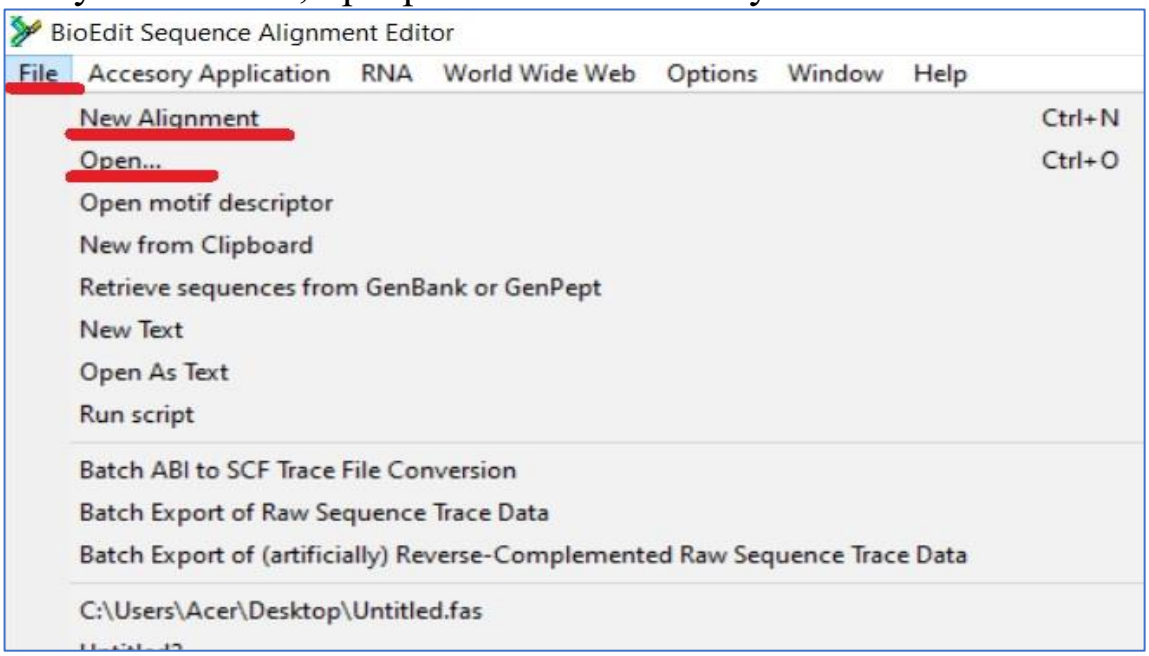


Рис. 19. Створення нового проекту в BioEdit

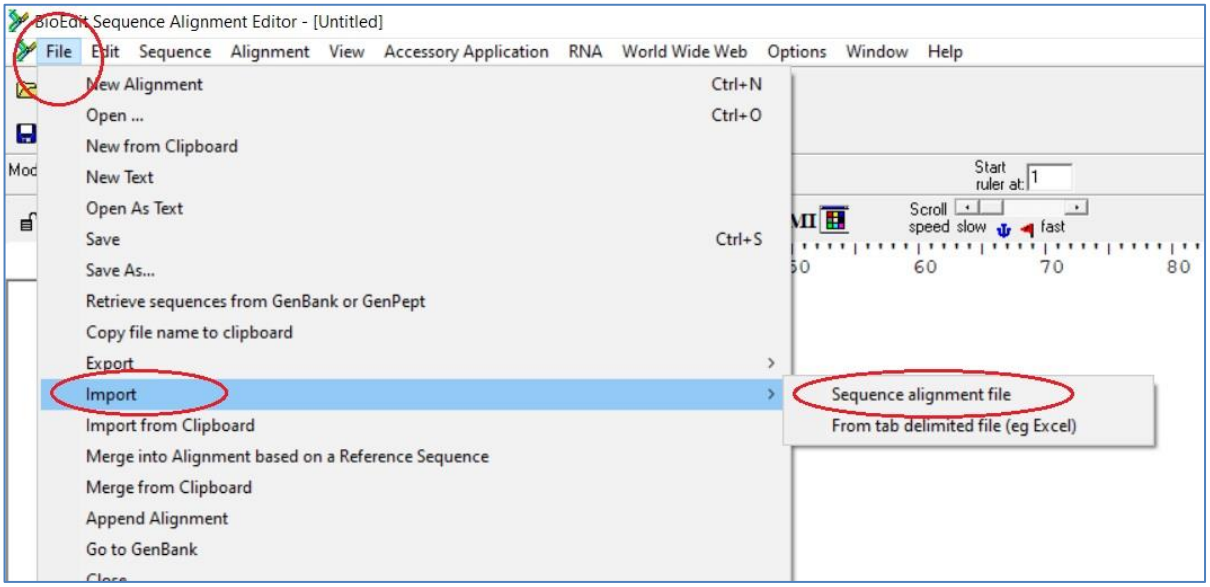


Рис. 20. Завантаження вхідних даних у програму BioEdit

Усі вибрані файли після натискання кнопки «Відкрити» завантажуються в програму. При цьому в лівій частині робочої області відображуються назви завантажених об’єктів, а праворуч – безпосередньо послідовності (рис. 21).

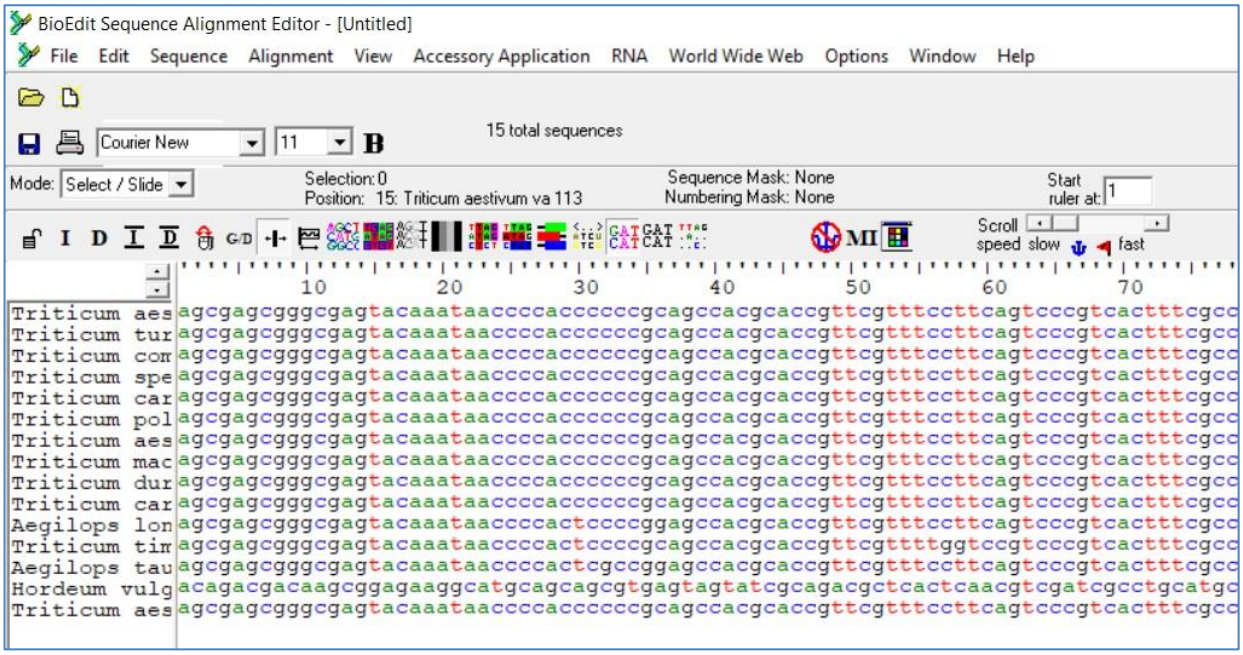


Рис. 21. Вигляд програми BioEdit із завантаженими даними

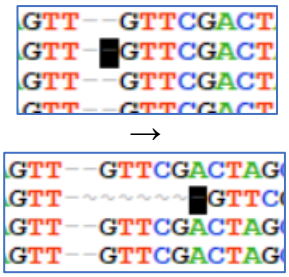
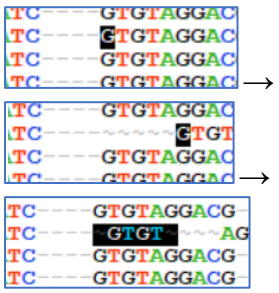
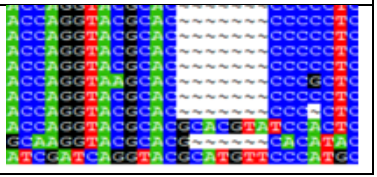
Вирівнювання послідовностей. Для прикладу розглянемо вирівнювання 14 послідовностей ДНК гена *Wx* різних видів пшениці, а також егілопсу і ячменю, які секвеновані. Послідовності було підібрано і завантажено з платформи NCBI з використанням програми BLAST. Далі послідовності імпортували в програму BioEdit.

Процедуру вирівнювання досліджуваних послідовностей можна виконувати вручну, використовуючи опції на панелі інструментів робочого вікна (табл. 9). Проте це займає багато часу і результат буде обумовлений суб'єктивним баченням дослідника.

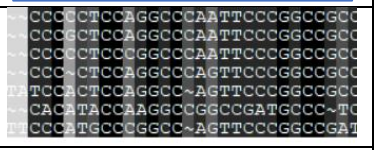
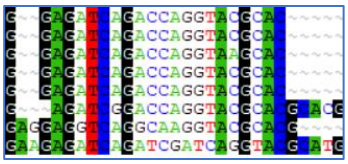
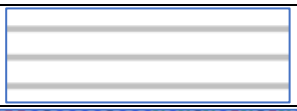
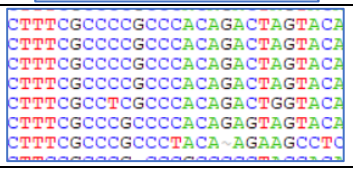
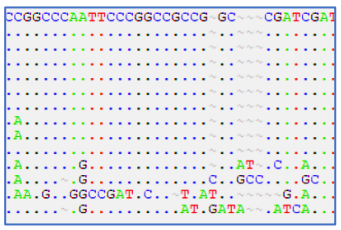
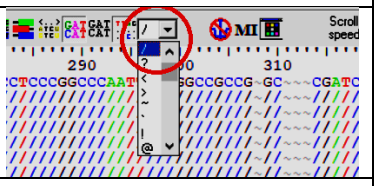
9. Функції BioEdit для маніпуляції послідовностями

Опції на панелі інстру-ментів	Функція	Зовнішній вигляд результату
1	2	3
 	Блокує/розблоковує усі проміжки у вирівнюванні. Зміни в послідовності можна вносити лише за умов «розблокованого» положення	Зовнішній вигляд послідовності не змінюється, але в заблокованому стані програма не сприймає зміни, внесені користувачем

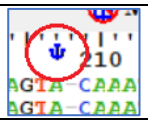
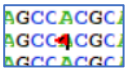
Продовження табл. 9

1	2	3
	Дозволяє вставити окремі пропуски, натиснувши праву клавішу миші	
	Дозволяє видалити пропуски, натиснувши праву клавішу миші	
	Дозволяє вставити проміжки в усіх послідовностях, крім тієї, яку натискали правою клавішею миші	
	Видаляє проміжки в усіх послідовностях, крім тієї, яку натискали правою клавішею миші. Послідовності, які не мають проміжку у вибраному положенні, будуть незмінними	
	Змінює функції за замовчуванням лівої та правої клавішей миші	
	Активує режим «Grab & Drag» («Захопити і перетягнути»)	
	Коли ця функція активована, то при натисканні на якийсь символ у послідовності та горизонтальному «протягуванні» праворуч з'являються додаткові проміжки, а сама послідовність зсувається. Протягування ліворуч дозволяє зменшити кількість пропусків і повернути послідовність у початкове положення без змін. Якщо на цій ділянці пропусків немає, то після «протягування» послідовність не змінюється. Функція діє за умови затискання клавіші **Shift** під час протягування	
	Активація цієї функції при протягуванні затиснутого символу праворуч дає такий самий результат, як попередня функція. А при протягуванні ліворуч дозволяє повернути на місце лише виділений фрагмент. Функція діє за умови затискання клавіші **Shift** під час протягування	
	Затушовує нуклеотиди різними кольорами	

Продовження табл. 9

1	2	3
	Змінюють кольорове забарвлення нуклеотидів на монохромне	
	Забезпечує монохромне забарвлення стовпчиків	
	Затушовує ідентичні і дуже подібні ділянки послідовностей. Ступінь подібності затушованих стовпчиків можна задавати самостійно	
	Забезпечує перегляд послідовностей відповідно до попередньо встановлених функцій	Залежить від функцій, які до цього були застосовані
	Креслить лише функції. Не показує послідовностей	
	Забезпечують кольорове забарвлення нуклеотидів	
	Дозволяє переглянути зображення, позначаючи ідентичні стосовно до стандарту нуклеотиди у вигляді крапки, або іншої позначки. У цьому режимі найзручніше аналізувати наявність/ відсутність SNP у досліджуваних послідовностях	
	Дозволяє змінити вигляд позначок у попередній функції («.», «/», «<», «>», «?», «%» тощо)	
	Вмикає режим «ігнорувати якірні точки». Коли ця функція активна – анкерні стовпці ігноруються, коли вимкнена – анкерні стовпці обмежують діапазон вирівнювання	
	Показує або приховує аналізатор спільної інформації (лише для аналізу RNA)	
	Опція, яка дозволяє самостійно обрати колір нуклеотидів і всіх позначень у послідовностях	

Продовження табл. 9

1	2	3
	Керує швидкістю горизонтальної смуги прокрутки	
	Додає або видаляє точку прив'язки стовпця	
	Додає або видаляє прапор позиційного маркера	

Зручніше використовувати вбудовану в BioEdit програму ClastalW.

ClastalW. Програма, створена J.D. Thompson, D.G. Higgins та T.J. Gibson (1994), призначена для автоматичного вирівнювання безлічі біологічних послідовностей із процедурою прогресивного евристичного вирівнювання. Вона забезпечує отримання найбільш коректних біологічних результатів.

Прогресивне вирівнювання включає етапи будування парних вирівнювань, формування корегуючого дерева і проведення множинного вирівнювання за цим деревом.

Clastal W можна використовувати як самостійне програмне забезпечення, доступне для вільного завантаження з багатьох серверів (http://npsa-pbil.ibcp.fr, http://www.ebi.ac.uk/clastalw/index.html).

У BioEdit пакет Clastal W використовують без змін, а основну онлайн-допомогу для цієї програми надають у вигляді пов'язаної версії оригінальної документації, що поширюється з програмою. Інтерфейс BioEdit для ClustalW є простим, а опції, доступні для цієї програми, описано в документації на Clustal W.

Вирівнювання у BioEdit виконуємо за алгоритмом Сміта–Уотермана. Зокрема, для реалізації процедури вирівнювання виділяємо назви всіх залучених до роботи послідовностей. Далі вибираємо опцію **Accessory Application** на панелі задач. У вкладці вибираємо функцію **ClastalW Multiple Alignment** (рис. 22), натискаємо **Run ClastalW** (рис. 23) і **OK** у діалоговому вікні.

Програма виконує обчислення і видає результат у вигляді вирівняних послідовностей у новому вікні (рис. 24). Різні нуклеотиди підсвічуються різними кольорами. Відсутність нуклеотидів на певній ділянці, позначена значком «–», свідчить про наявність інсерцій/делецій у відповідних локусах послідовностей досліджуваних біологічних об'єктів.

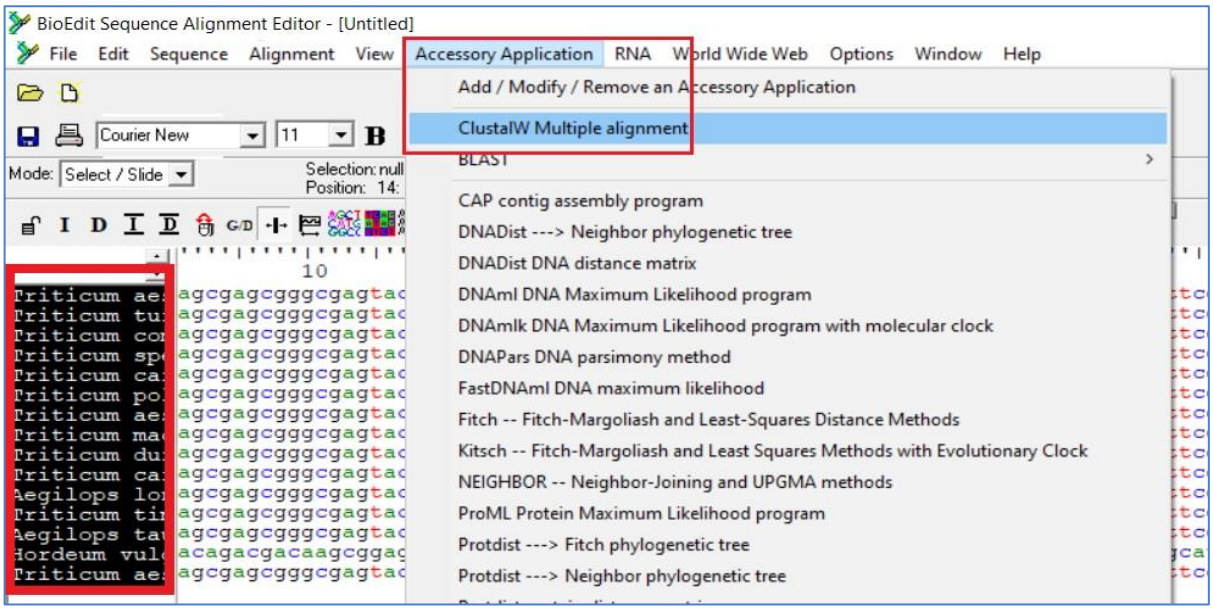


Рис. 22. Опції для вирівнювання послідовностей у програмі BioEdit

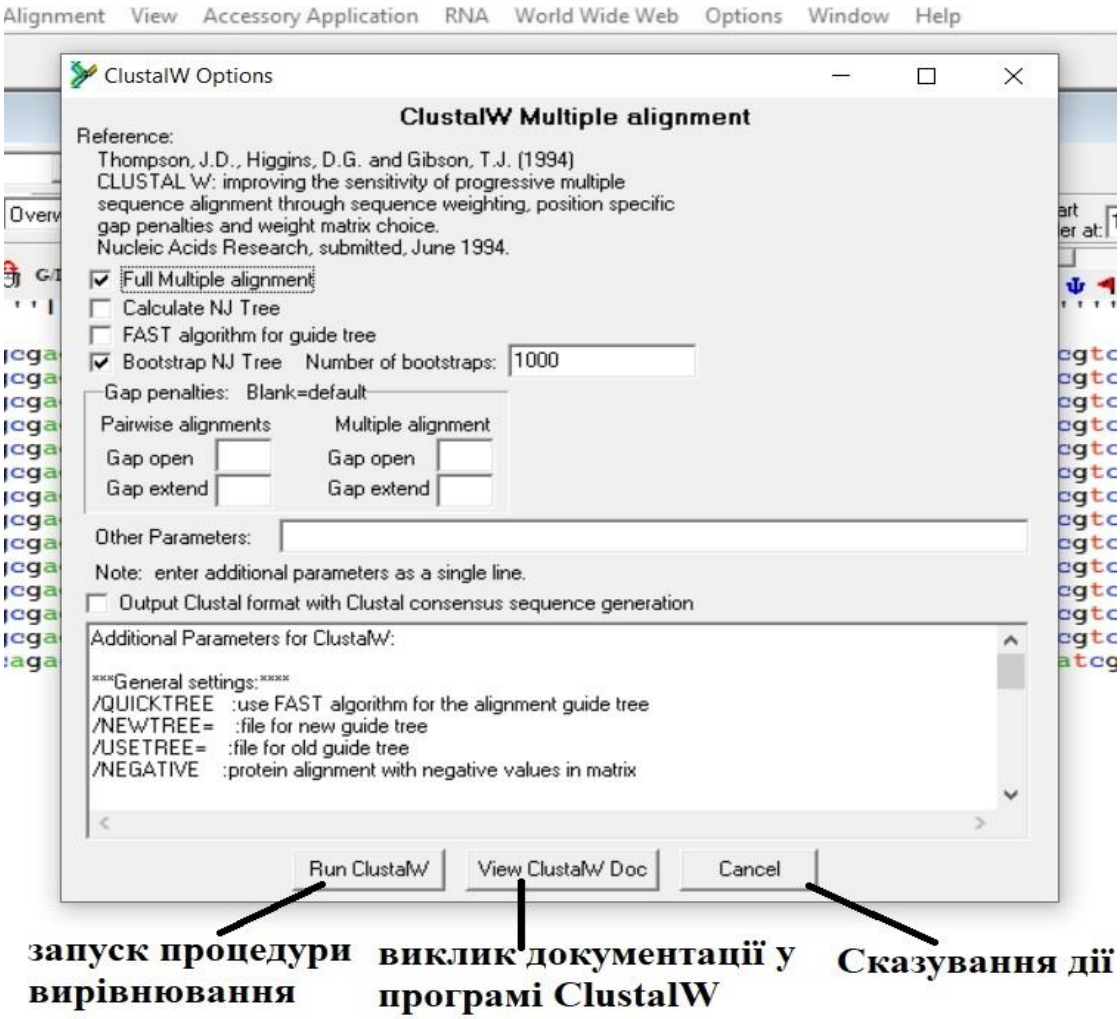


Рис. 23. Вікно програми ClastalW з параметрами вирівнювання

Рис. 24. Вікно BioEdit з результатами вирівнювання послідовностей

Для зручності сприйняття отриманих результатів можна використовувати опції на панелі інструментів, які дозволяють замінити однакові нуклеотиди в колонках крапками або іншими позначками. При цьому першу послідовність приймають за стандарт, а всі інші порівнюють з нею. Однакові нуклеотиди відповідно до стандарту у другій та всіх наступних послідовностях замінюють обраною позначкою. Нуклеотиди, які відрізняються від стандарту, залишаються у вигляді літерних символів (A, T, G, C тощо). У такому режимі найзручніше аналізувати наявність SNP.

За допомогою опцій на панелі інструментів можна змінити колір нуклеотидів, замінити кольорове забарвлення на монохромне, затушувати однаковими кольорами області з однаковими нуклеотидами тощо.

Часто досліджувані послідовності суттєво відрізняються за кількістю нуклеотидів і після вирівнювання можна побачити неінформативні зони перед або після порівнюваних областей (рис. 25). У нашому прикладі неінформативні зони виникли через те, що останній ланцюжок (*H. vulgare*) містив ділянки розміром 191 п.н. на початку сиквенсу і 2763 п.н. – у кінці, аналоги яких були відсутні в усіх інших досліджуваних нами послідовностях.

Наявність таких ділянок заважає зручному сприйманню результатів і може спричиняти неправильне групування зразків під час подальшого аналізу послідовностей, у тому числі з використанням іншого програмного забезпечення (на процес вирівнювання ці зони не впливають).

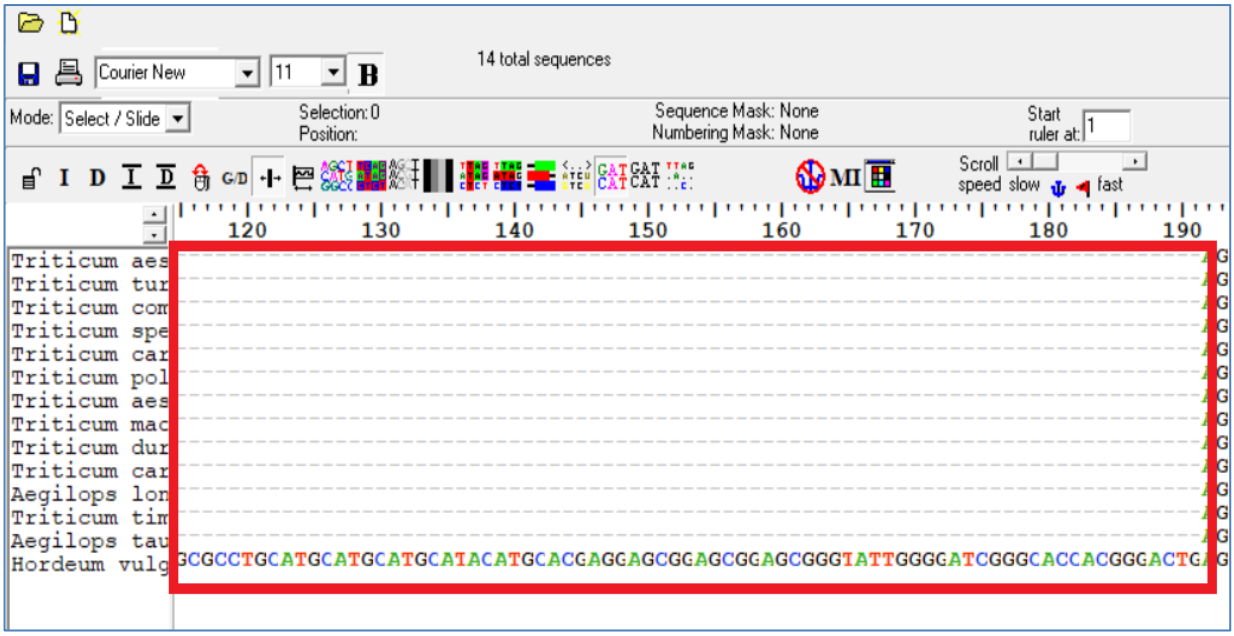


Рис. 25. Неінформативна зона вирівняних послідовностей

Неінформативні ділянки бажано вилучити. Це можна зробити двома шляхами:

1. Установити кількість «непотрібних» нуклеотидів на початку та в кінці послідовностей-вискочок. Зайти у вхідні файли цих послідовностей на ПК і видалити непотрібні фрагменти безпосередньо у вхідних файлах.

У програмі BioEdit до початку вирівнювання виділити назву послідовності, яку потрібно вкоротити. Двічі натиснути на неї лівою клавішею миші. З’явиться вікно, у межах якого можливе маніпулювання відповідною послідовністю (рис. 26).

У вікні, яке відкрилося, можна подивитися інформацію про відповідну послідовність ДНК (таксономія досліджуваного об’єкта, посилання на літературні джерела, загальна довжина та інші особливості послідовності). Виділяємо ту частину ланцюжка, яку слід видалити, і натискаємо клавішу **Delete** або **Backspace**. У цьому ж вікні можна вписати потрібні нуклеотиди. Щоб зберегти зміни, натискаємо кнопку **Apply** або **Apply and Close**.

Збереження результатів. Щоб зберегти результати вирівнювання, потрібно після виконання процедури у вкладці **File** обрати опцію **Save** або **Save As**, задати шлях, у яку саме папку потрібно помістити цей файл, обрати формат, у якому слід його зберігати, і натиснути клавішу «Зберегти». BioEdit дозволяє зберегти результати у форматах *.fas, *.fst, *.fsa (Fasta), *.gb, *.gbk, *.gen, *.gnk (GenBank), *.phy (Phylip), *.nbr, *.pir (NBRF/PIR), *.txt (текстові редактори), *.bio (BioEdit), *.xml (XML files). Формат обирають залежно від способу і програм, у яких планують подальшу обробку результатів, отриманих у BioEdit.

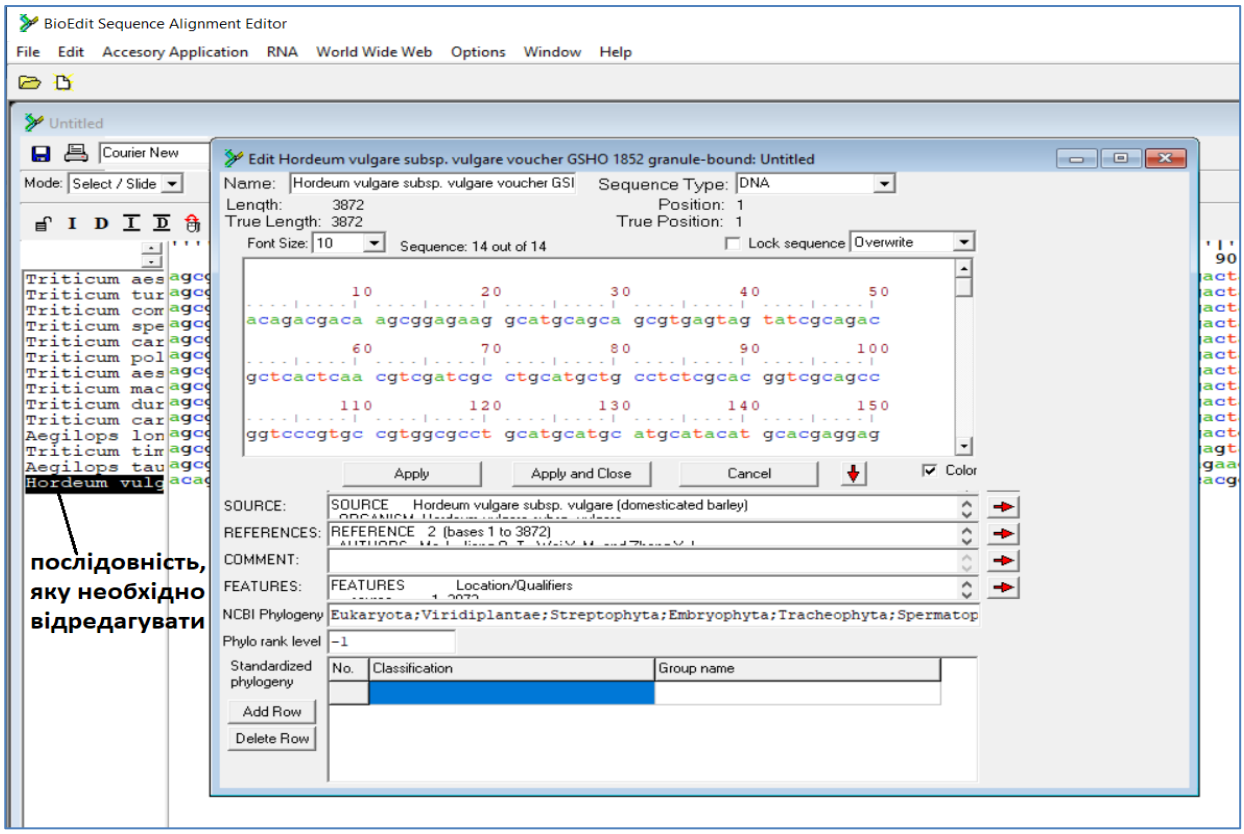


Рис. 26. Вікно з послідовністю ДНК, яку потрібно відредагувати, у програмі BioEdit

Другий спосіб збереження результатів полягає у безпосередньому перенесенні вирівняних послідовностей у документ MS Word. Для цього після процедури вирівнювання потрібно на панелі інструментів натиснути опцію **Print**. У вікні, що відкриється, виділяємо область з вирівняними послідовностями, яку потрібно зберегти. Натискаємо комбінацію клавіш **Ctrl+C** – виділена область завантажиться до буфера обміну (рис. 27).

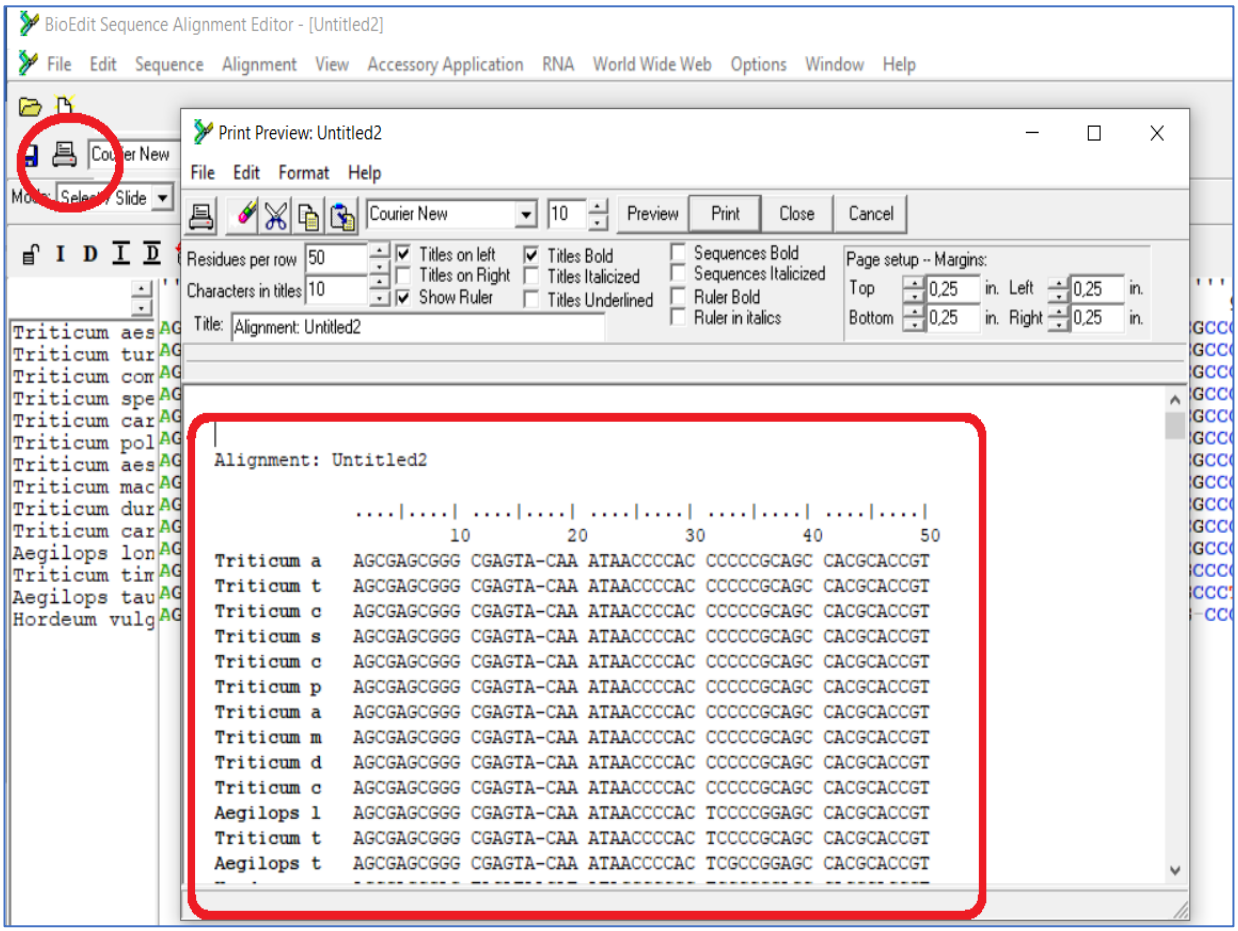


Рис. 27. Копіювання результатів вирівнювання

Далі відкриваємо файл формату *.doc, *.docx, *.rtf. Ставимо курсор у позицію, куди потрібно додати скопійовану послідовність, і натискаємо **Ctrl+V** – отримуємо копію результатів вирівнювання. Нижче наведено фрагмент результатів вирівнювання 14 послідовностей генів *Wx*, секвенованих у різних видів пшениці, а також егілопсу та ячменю.

```
                ....|....| ....|....| ....|....| ....|....| ....|....|
                         10         20         30         40         50
Tr. aestiv      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. turgid      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. compac      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. spelta      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. carthl      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. poloni      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. aestiv      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. macha       AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. durum       AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Tr. carthl      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC CCCCCGCAGC CACGCACCGT
Ae. longis      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC TCCCCGGAGC CACGCACCGT
Tr. timoph      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC TCCCCGCAGC CACGCACCGT
Ae. tausch      AGCGAGCGGG CGAGTA-CAA ATAACCCCAC TCGCCGGAGC CACGCACCGT
H. vulgare      AGCGAGCGAG TACATAACAT ATAGGCCCGC TCCCCGGAGC CACGCACCGT

                ....|....| ....|....| ....|....| ....|....| ....|....|
                         60         70         80         90        100
Tr. aestiv      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. turgid      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. compac      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. spelta      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. carthl      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. poloni      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. aestiv      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. macha       TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. durum       TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. carthl      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Ae. longis      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCTCGCCC ACAGACTGGT
Tr. timoph      TCGTTTTGGT C-----CGTC CCGTCACTTT CGCCCGCCCC ACAGAGTAGT
Ae. tausch      TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCGCCCT ACA-AGAAGC
H. vulgare      TCGTTCGTTT CCTTGGAGTC CCGTCACTTC CGCCCG-CCC GCCCCCTACC

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        110        120        130        140        150
Tr. aestiv      ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. turgid      ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. compac      ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. spelta      ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. carthl      ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. poloni      ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. aestiv      ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. macha       ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. durum       ACAACCAGGA GGCGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGAACGACA
Tr. carthl      ACAACCAGGA GGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGCACGACA
Ae. longis      ACAACCAGGA CGAGGAGGAG AGGAGAAGCC TCTGCCAGTG AAGGAGGACA
Tr. timoph      ACAACCAGGA GCCTCTCCCA GCGAACAACA ACAACAA--C AATAAGGACA
Ae. tausch      CTCTCCCAGT GAAGAAGAAG AAGAA----- ---------G AAGAAGGACA
H. vulgare      ACACACTACA ACCTCTGCCA CTCAA----- ---------C AACAACAACA
```

Слід зазначити, якщо змінити формат скопійованої послідовності, то рядки і стовпчики можуть зміститися, як показано нижче.

```
....|....| ....|....| ....|....| ....|....| ....|....|
        60        70        80        90       100
Tr. aestiv   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. turgid   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. compac   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. spelta   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. carthl   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. poloni   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. aestiv   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. macha    TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. durum    TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Tr. carthl   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCCGCCC ACAGACTAGT
Ae. longis   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCTCGCCC ACAGACTGGT
Tr. timoph   TCGTTTTGGT C-----CGTC CCGTCACTTT CGCCCGCCCC ACAGAGTAGT
Ae. tausch   TCGTTTCCTT C-----AGTC CCGTCACTTT CGCCCGCCCT ACA-AGAAGC
H. vulgare   TCGTTCGTTT CCTTGGAGTC CCGTCACTTC CGCCCG-CCCGCCCCCTACC
```

Сприймати й аналізувати такий запис важко. Отже, щоб уникнути зміщення, варто зберегти скопійовані вирівняні послідовності у форматі *.pdf і тільки потім переносити їх у текстовий документ як рисунок.

Програма MEGA X

MEGA (The Molecular Evolutionary Genetics Analysis, молекулярно-еволюційний генетичний аналіз) – програма для автоматичного та ручного вирівнювання послідовностей, виведення філогенетичних дерев, пошуку даних веб-баз, оцінки швидкості молекулярної еволюції і тестування еволюційних гіпотез. Першу версію програми розробили у 1994 р. S. Kumar, K. Tamura і M. Nei. Програму широко використовують дослідники в усьому світі для статистичного опрацювання філогенетичних та еволюційних проектів.

Робота з програмою. MEGA сумісна з операційними системами Microsoft Windows, Mac OS X та Linux, має простий, інтуїтивно зрозумілий інтерфейс. Програма є безкоштовною. Її можна завантажити з платформи www.megasoftware.net.

Не рекомендовано переносити файли програми з одного ПК на інший, оскільки в цьому випадку програма може працювати некоректно, а окремі функції можуть стати недоступними.

Скачану програму потрібно встановити на ПК шляхом активації інсталяційного файлу. Ярлик програми можна винести на робочий стіл ПК. Після запуску інтерфейс програми виглядає так (рис. 28).

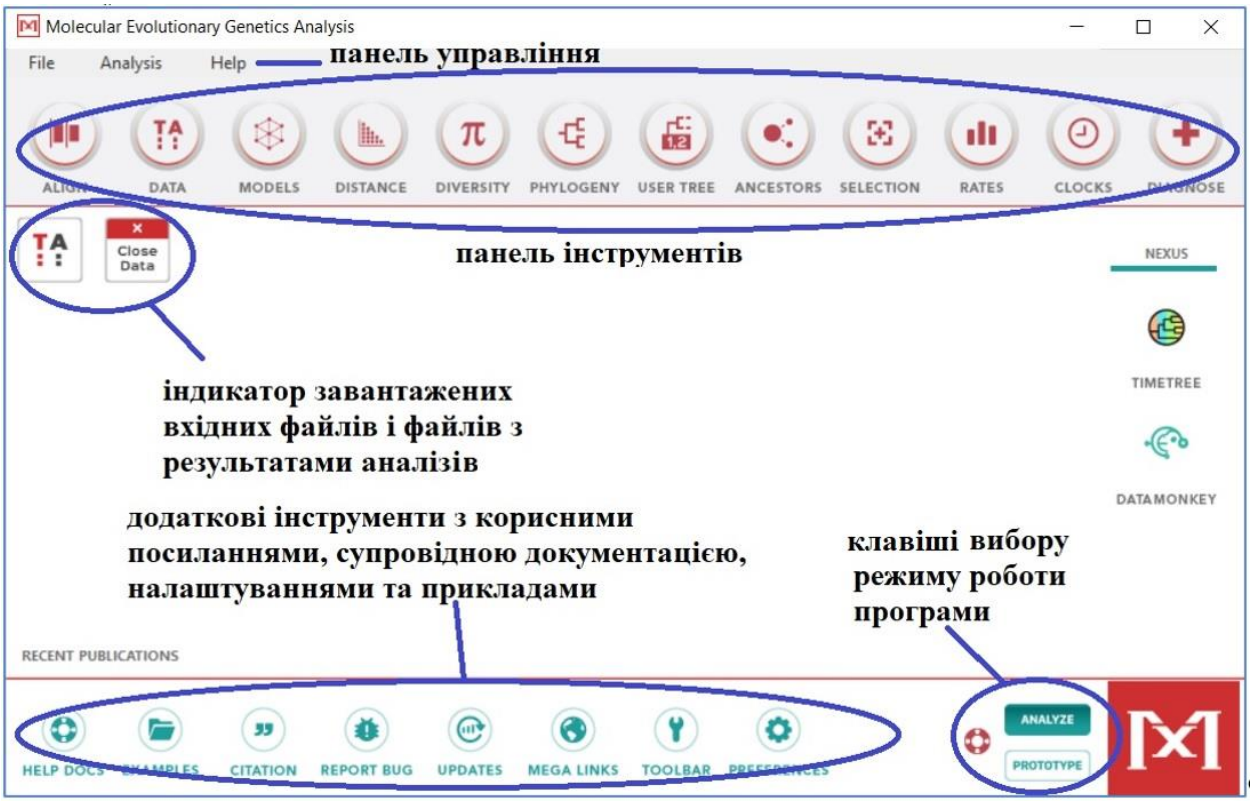


Рис. 28. Інтерфейс програми MEGA X

Програма MEGA X може працювати у двох режимах: Analyze та Prototype (див. рис. 28). Їх потрібно обирати перед початком роботи до завантаження вхідних даних. Режим **Analyze** активує повний графічний інтерфейс, у якому доступні всі візуальні засоби для аналізу даних та дослідження результатів. Цей режим установлено за замовчуванням. Режим **Prototype** дозволяє працювати тільки з файлами, створеними у MEGA (*.mao), з інтерфейсом командного рядка MEGA-CC. У разі активації цього режиму всі інструменти візуалізації даних і результатів вимкнено. Доступними залишаються тільки меню аналізу та діалогові вікна параметрів. Переважну більшість філогенетичних обчислень виконують у режимі Analyze.

Для початку роботи програми необхідно завантажити вхідні файли з послідовностями. Для проведення розрахунків і будування філогенетичних дерев використовують лише заздалегідь вирівняні послідовності однакової довжини. Для проведення аналізів потрібні файли форматів mega (*.meg, *.mao та ін.) і fasta (*.fas, *.fasta, *.fst, *.fsa). Переглянути можна також файли формату *.txt. Проте, щоб здійснити будь-які маніпуляції та обчислення, їх потрібно конвертувати у формат MEGA. MEGA підтримує чимало інших форматів (*.an, *.nexus, *.phylip, *.gcg, *.pir, *.nbrf, *.msf, *.ig, *.xml). Але для використання їх також необхідно конвертувати у формат MEGA, що можна здійснити в межах самої програми.

Щоб завантажити вхідні дані, потрібно на панелі управління обрати меню **File**, у вкладці натиснути ***Open a File/Session*** (відкриває новий файл) або ***Open a Recently Used File*** (пропонує завантажити один з попередніх файлів, з якими вже працювали), у діалоговому вікні обрати і відкрити потрібний файл даних. З'явиться діалогове вікно, у якому пропонується обрати одну з функцій Align або Analyze (рис. 29).

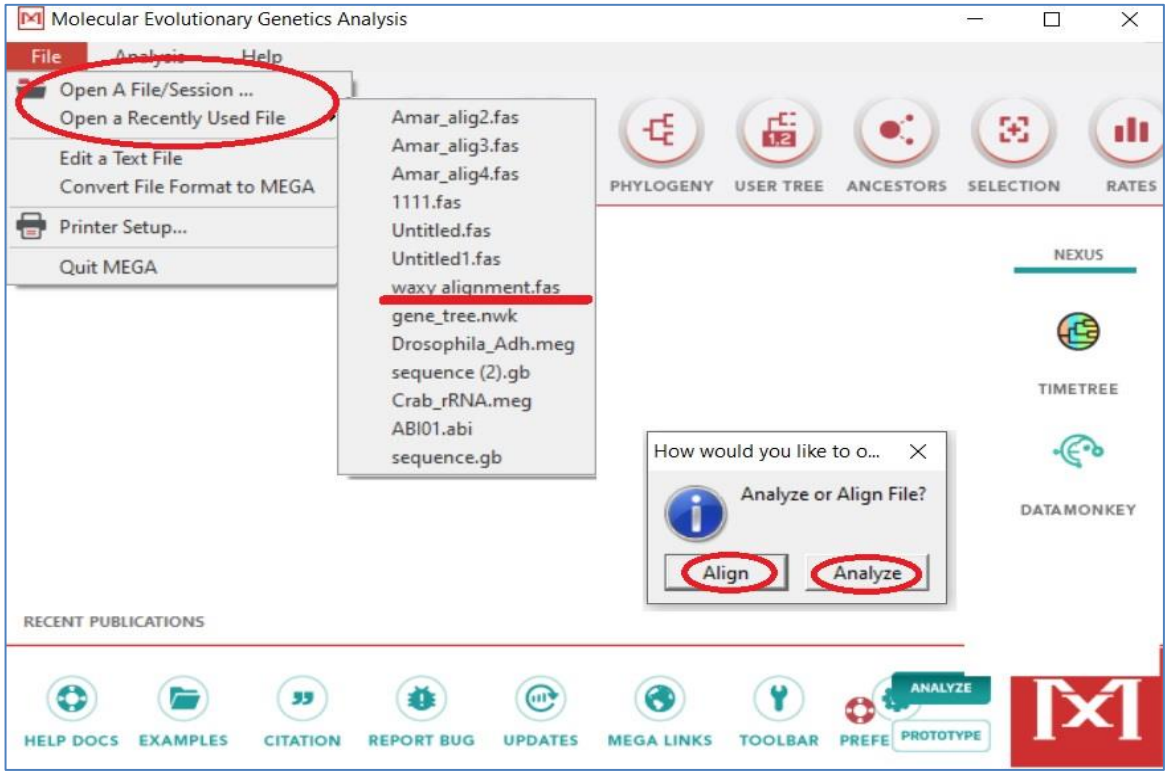


Рис. 29. Завантаження файлу даних у програмі MEGA X

Align відкриє обраний файл даних для здійснення вирівнювання послідовностей власноруч або за допомогою вбудованих у MEGA програм CLUSTAL W та MUSCLE (рис. 30).

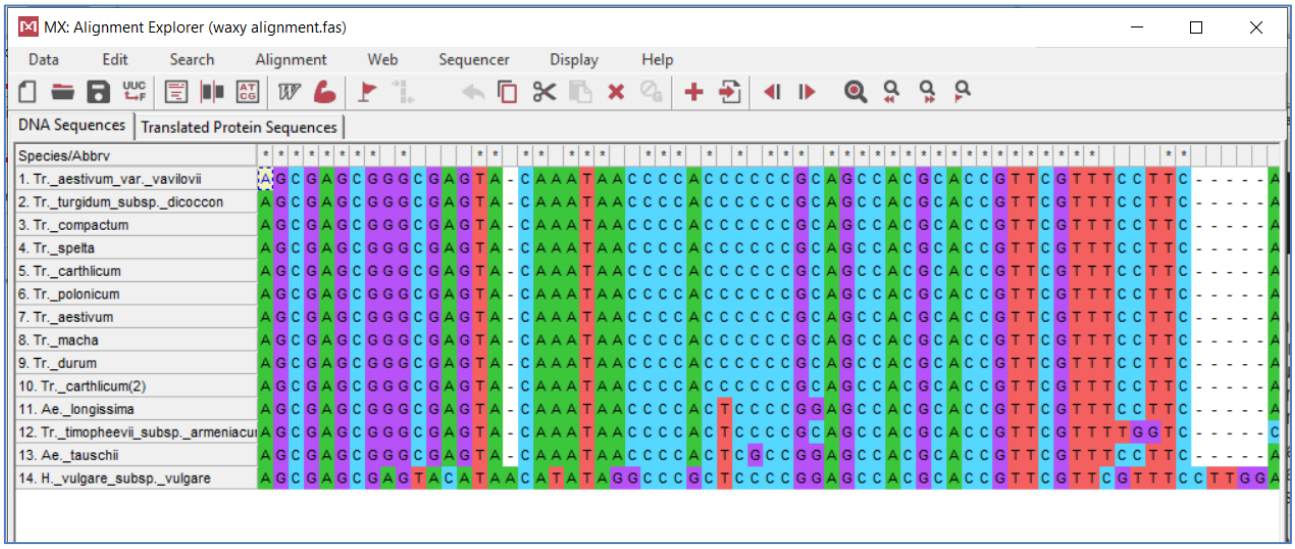


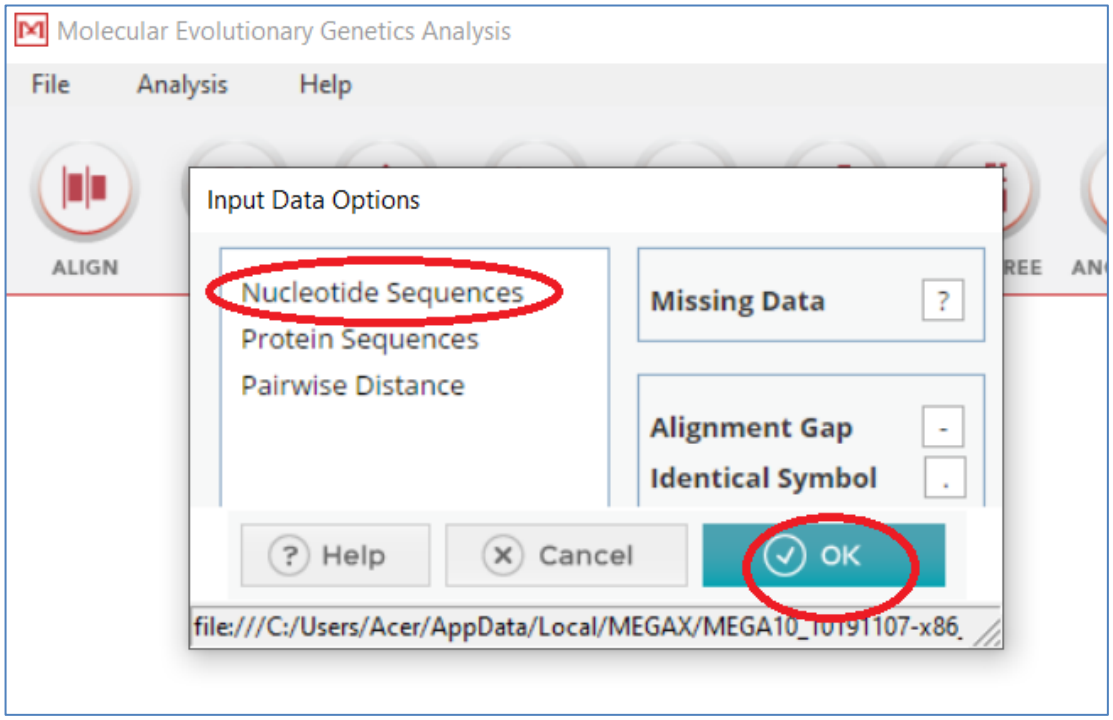
Рис. 30. Завантажений у програму MEGA X файл даних для редагування послідовностей

Для здійснення вирівнювання власноруч потрібно поставити курсор у місце на послідовності, у якому потрібно додати/видалити нуклеотиди або розрив і здійснити потрібну маніпуляцію з використанням літерної клавіатури та клавіш Delete і Beckspace. Для реалізації автоматичного вирівнювання з використанням вбудованого програмного забезпечення потрібно виділити всі послідовності або ті, які потребують вирівнювання, обрати на панелі управління опцію **Alignment** і у вкладці, яка з'явиться, вибрати програму для вирівнювання. Якщо вхідні послідовності різної довжини або містять неоднакові змістові фрагменти, то результати вирівнювання будуть некоректні. Тому краще проводити вирівнювання до початку роботи у MEGA з використанням інших програм.

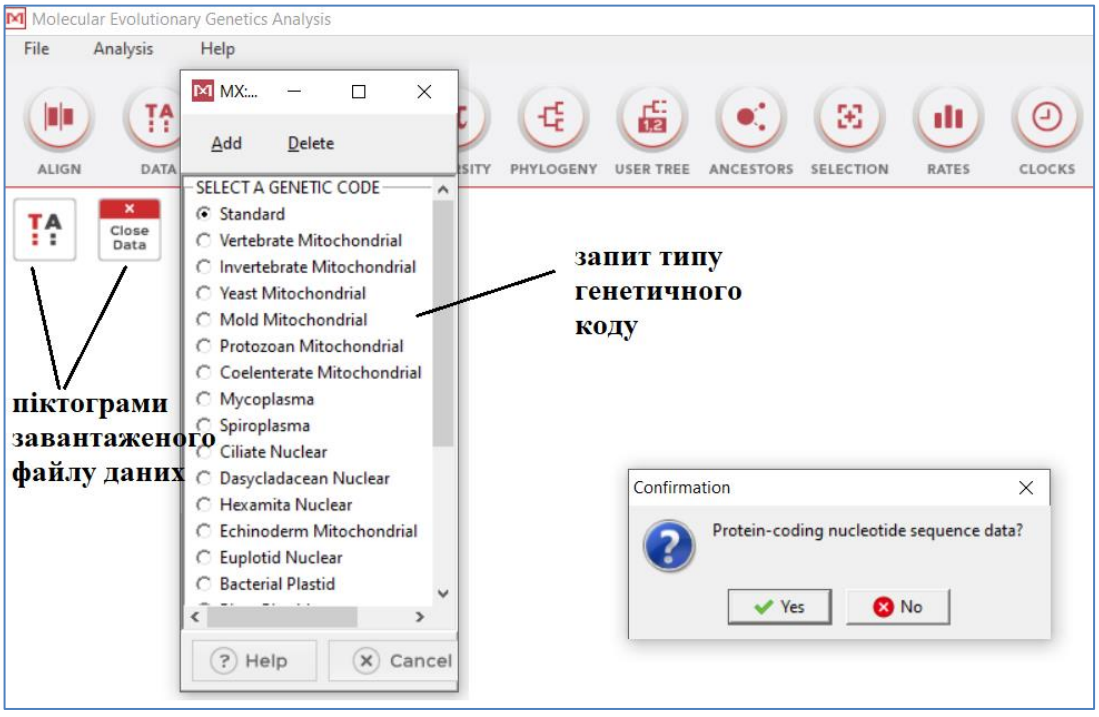
Функція **Analyze** завантажує файл даних для філогенетичного аналізу (рис. 31, а). Після її активації з'являється діалогове вікно, в якому пропонується обрати тип даних (нуклеотидні чи білкові послідовності, попарні дистанції), які аналізуватимуться.

Розглянемо принципи роботи MEGA на прикладі 14 нуклеотидних послідовностей генів ***Wx*** різних видів пшениці, а також егілопсу і ячменю. У діалоговому вікні обираємо тип даних

Nucleotide Sequences і натискаємо ***OK***. Далі з'являється віконце із запитом, чи це є протеїн-кодувальна послідовність, чи ні (рис. 31, б).



а



б

Рис. 31. Завантаження файлу даних у програмі MEGA X для філогенетичного аналізу:

а – вибір типу даних (нуклеотидні, білкові послідовності, парні дистанції), б – завантаження файлу з білок-кодувальними послідовностями

Якщо обрати **NO**, то файл даних завантажується одразу і з'являється в робочій зоні програми MEGA X у вигляді піктограм **TA** і **Close Data**. У разі вибору **YES** разом з піктограмами відкривається вікно, де запропоновано обрати, який саме генетичний код використано у файлі даних (Standard, Vertebrate Mitochondrial, Mold Mitochondrial, Mycoplasma, Euploid Nuclear та інші).

Піктограма **TA** відкриває вікно, у якому можна переглянути завантажені послідовності, з використанням меню на панелі інструментів затушувати консервативні («C»), варіабельні («V»), економічно інформативні («Pi») і SNP («S») сайти, зберегти дані в іншому форматі тощо (рис. 32). Піктограма **Close Data** при натисканні закриває завантажений файл.

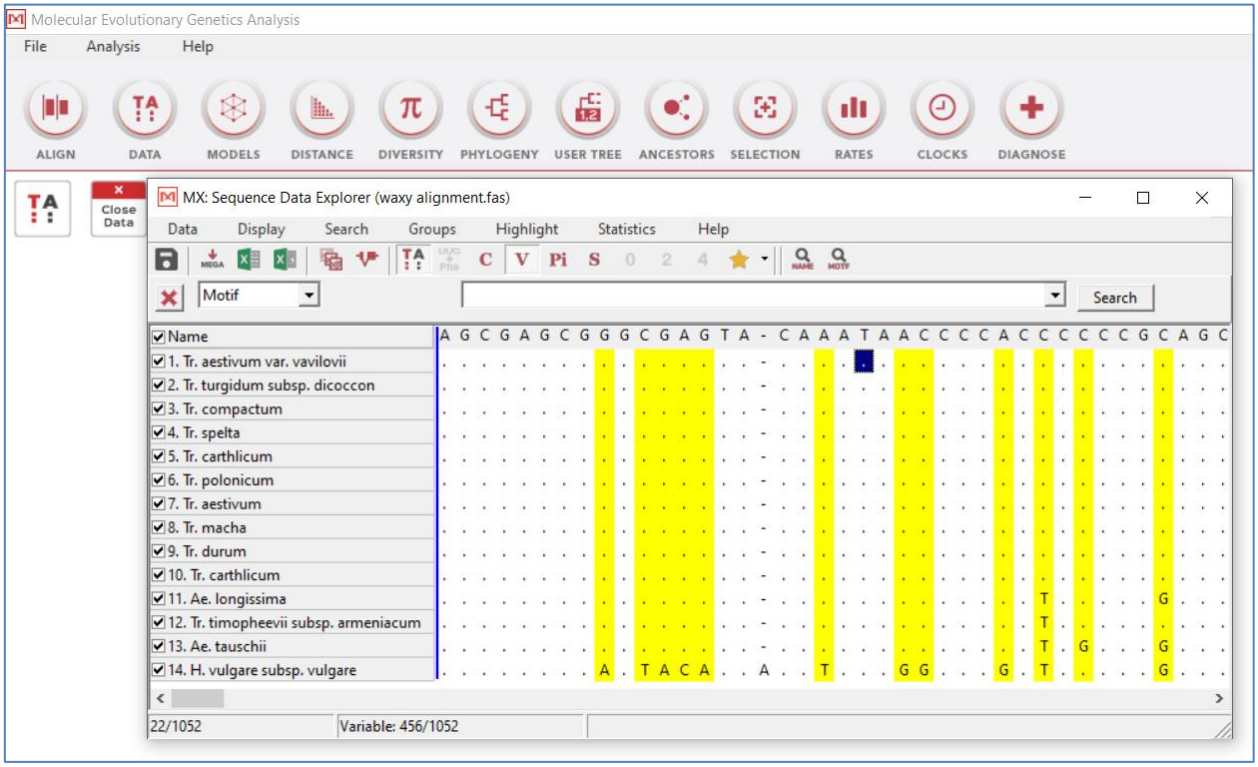


Рис. 32. Вікно із завантаженими послідовностями, активоване піктограмою TA

Примітка. Жовтим затушовано варіабельні сайти.

Аналіз даних у MEGA X. Після завантаження дані готові до подальшого аналізу. Для його здійснення потрібно натиснути функцію **Analysis** на панелі управління і у вкладці вибрати тип аналізу, який необхідно здійснити. Також можна обрати потрібну опцію на панелі інструментів головного вікна програми (див. рис. 28).

Меню **ALIGN** призначене для множинного вирівнювання нуклеотидних і білкових послідовностей. Після вибору цієї опції з'являється вкладка, де можна обрати спосіб завантаження даних для аналізу (рис. 33).

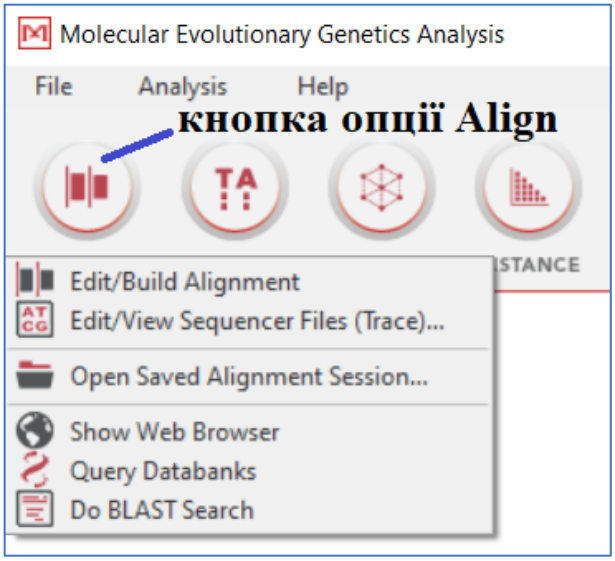


Рис. 33. Доступні способи завантаження даних після активації опції Align у програмі MEGA X

Опція ***Edit/Build Alignment*** дозволяє відкрити файл формату fasta, знайти послідовність з використанням BLAST на платформі NCBI або побудувати послідовність власноруч. При цьому програма запитає, це буде послідовність ДНК або білка.

Опцію ***Edit/View Sequencer File (Trace)*** слід використовувати для завантаження вже створених файлів з послідовностями формату *.abi та *.ab1.

Опція ***Open Saved Alignment Session*** відкриває вже створений файл формату *.mas та *.masx (файли MEGA) із розпочатим раніше, але незавершеним вирівнюванням.

Опції ***Show Web Browser***, ***Query Databanks*** та ***Do BLAST Search*** дозволяють завантажити біологічні послідовності з різних платформ мережі Інтернет.

Меню **DATA (TA)** призначене для маніпуляцій із файлами даних. Воно дозволяє відкривати збережені на ПК файли, поєднувати вирівняні послідовності, використовувати активні дані, завантажувати і конвертувати вирівнювання, з якими нещодавно працювали, у формат MEGA, Nexus, Phylip, Fasta, (рис. 34), перекладати послідовності ДНК↔білок, задавати тип генетичного коду, обирати гени, домени, таксони і групи у вирівнюваннях.

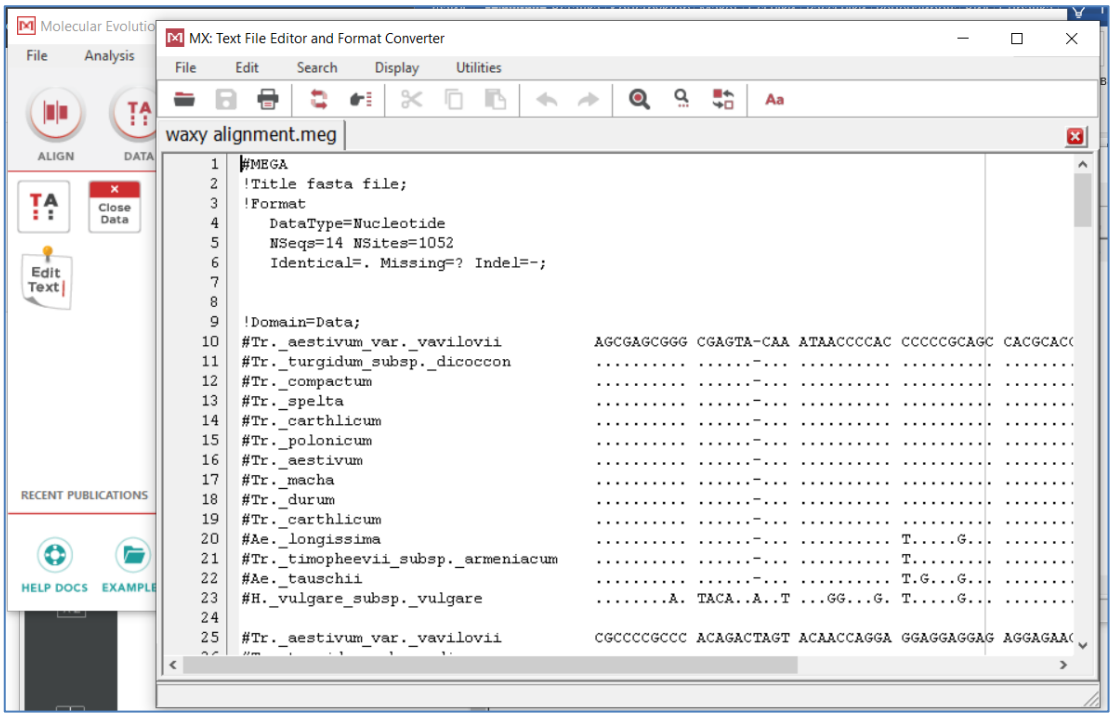


Рис. 34. Вікно з конвертованим у формат *.meg вирівнюванням послідовностей гена *Wx* різних видів пшениці, егілопсу і ячменю

Меню **MODELS** допомагає обрати найефективнішу модель філогенетичної оцінки вирівняних послідовностей (рис. 35).

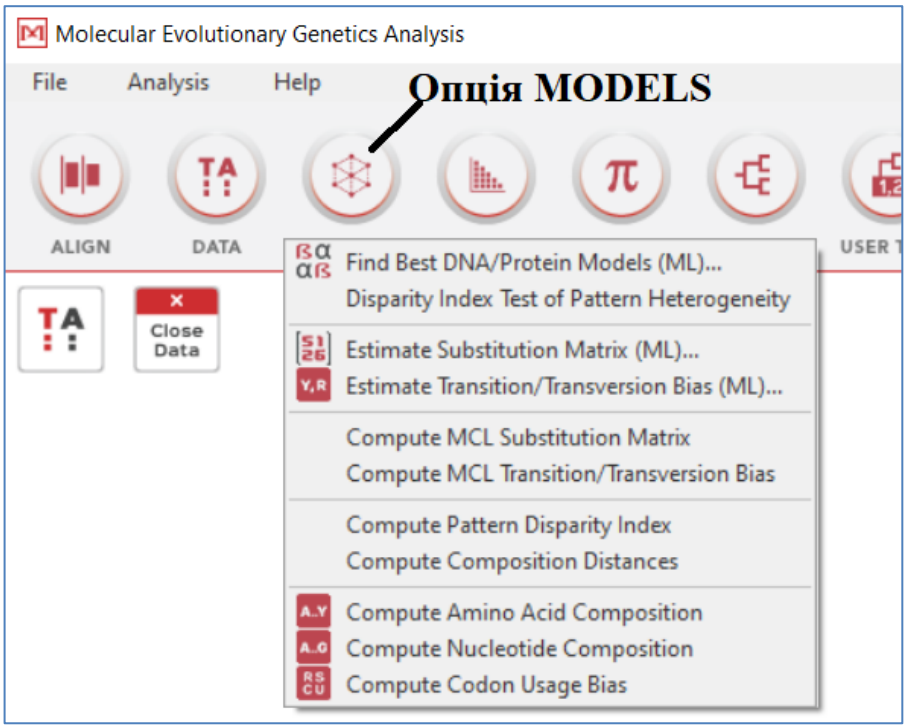


Рис. 35. Доступні функції, активовані опцією MODELS у програмі MEGA X

Параметр ***Find Best DNA/Protein Models (ML)*** тестує файл даних (нуклеотид або амінокислота) на відповідність деяким популярним моделям еволюції та повертає значення кількох критеріїв, за допомогою яких можна вибрати найбільш ефективну еволюційну модель для свого аналізу.

Параметри ***Disparity Index Test of Pattern Heterogeneity*** (тест показника невідповідності гетерогенності шаблону) і ***Compute Pattern Disparity Index*** перевіряють нульову гіпотезу про те, що послідовності еволюціонували за тією ж схемою заміщення, за якою оцінено ступінь відмінностей зміщень базових комбінацій у послідовностях.

Параметр ***Estimate Substitution Matrix (ML)*** оцінює та відображає матрицю швидкості заміщення нуклеотидів, використовуючи метод максимальної ймовірності для поточного набору даних та вибраної еволюційної моделі. Використовують лише для нуклеотидних даних.

Функція ***Estimate Transition/Transversion Bias (ML)*** оцінює параметри κ (модель Kimura), κ1 і κ2 (модель Tamura-Nei) транзицій/трансверсій, використовуючи метод максимальної ймовірності. Доцільна лише для нуклеотидних даних.

Опція ***Compute MCL Substitution Matrix*** оцінює та відображає матрицю швидкості заміщення, оцінену методом максимальної сумарної ймовірності (MCL), для поточного набору даних (лише нуклеотидні дані).

Опція ***Compute MCL Transversion/Transition bias*** оцінює параметри транзицій/трансверсій κ (для пуринів і піримідинів), κ1 (лише пурини) та κ2 (лише для піримідинів) за моделлю максимальної сумарної ймовірності (тільки для нуклеотидних даних).

Опція ***Compute Composition Distance*** аналізує ступінь різниці нуклеотидного (або амінокислотного) складу для заданої пари послідовностей.

Команда ***Amino Acid Composition*** обчислює частоти амінокислот для кожної послідовності, а також загальне середнє значення, яке відображатиметься за доменом. Опція є активною лише для білкових послідовностей, а також нуклеотидних послідовностей, заздалегідь переведених у білкові.

Команда ***Compute Nucleotide Composition*** обчислює частоти азотистих основ для кожної послідовності, а також загальне

середнє значення. Результати подають у таблиці Excel. Опція активна лише для даних з нуклеотидними послідовностями.

Команда ***Compute Codon Usage Bias*** обчислює частоту заміщень кодонів.

Інструмент **DISTANCE** призначений для обчислення попарних та середніх відстаней між послідовностями, стандартних помилок; середніх, внутрішньо- і міжгрупових відстаней у групах; статистику різноманітності послідовностей для багатьох популяцій (рис. 36).

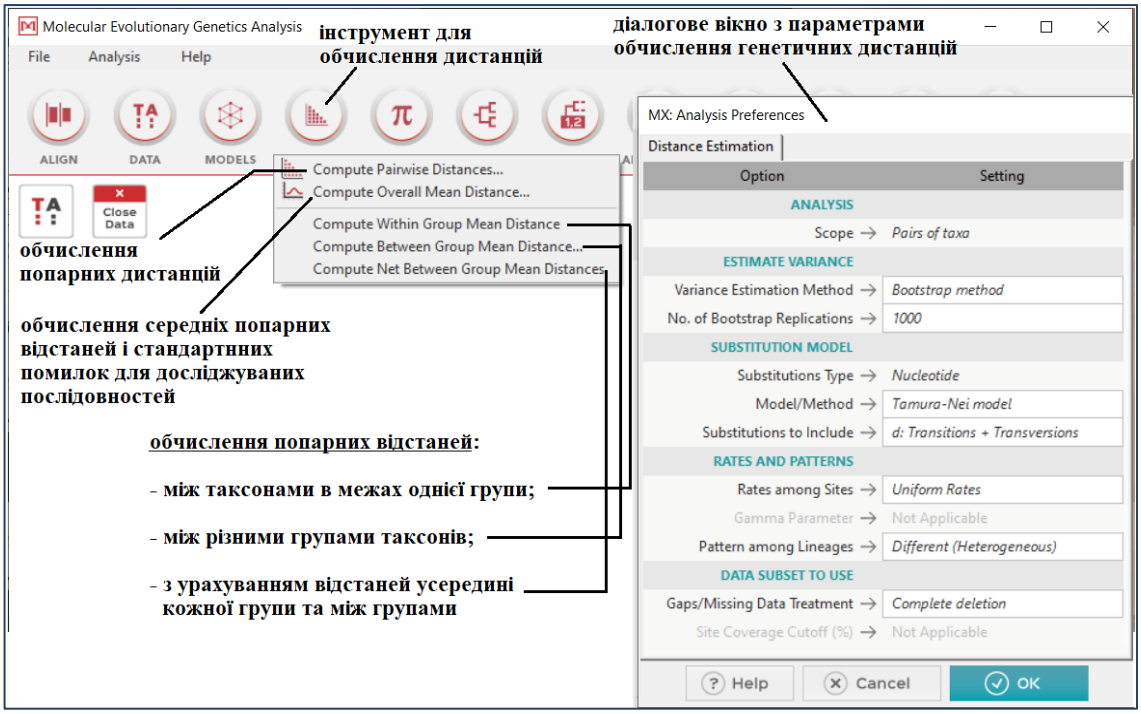
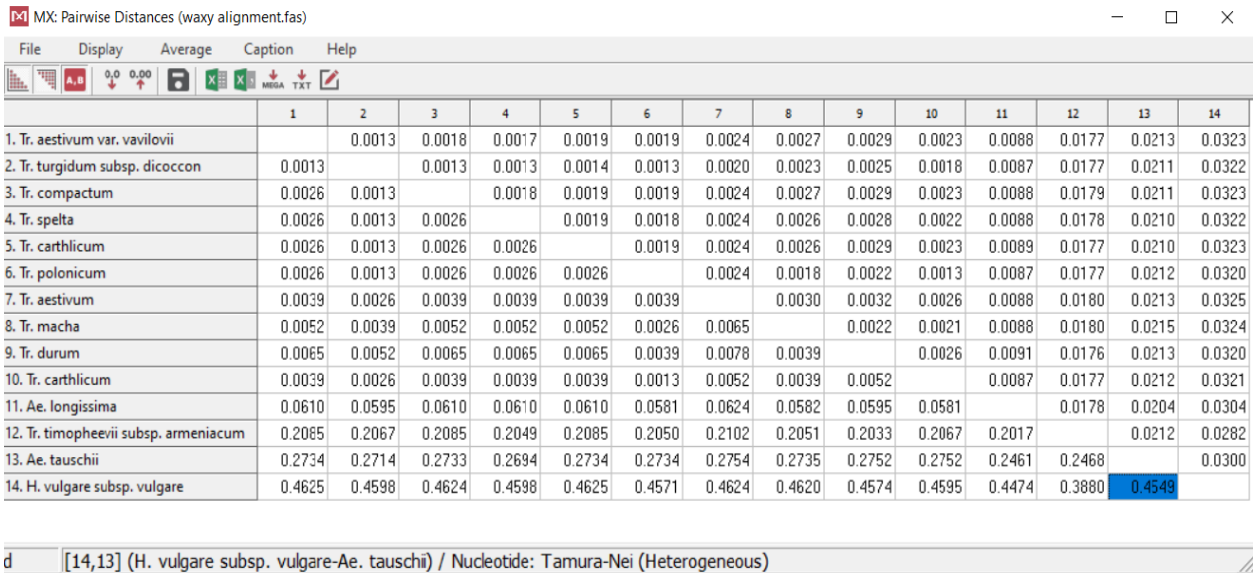


Рис. 36. Меню DISTANCE в програмі MEGA X

Після активації опцій у меню DISTANCE з'являється діалогове вікно, у якому можна задати потрібну модель оцінки відстаней, метод розрахунку помилок (бутстреп або аналітичний), тип замін, які буде враховано (транзиції, трансверсії, їх поєднання, їх відношення), обрати рівень розподілу сайтів (рівномірний або gamma), відмітити гомогенність/гетерогенність послідовностей усередині родин, вибрати спосіб обробки пропущених та відсутніх даних (повне, парне або часткове вилучення).

У MEGA можна розрахувати такі еволюційні моделі: парні дистанції (p-distance), Jukes-Cantor, Kimura 2-parameter, Tajima-Nei, Tamura 3-parameter, Tamura-Nei, MCL-метод (Maximum Composite Likelihood), які розглянуто в розділі 2, а формули наведено в дод. Б.

Результат обчислення генетичних відстаней за однією з обраних моделей у програмі MEGA оформлено як матриця відстаней. Нижче наведено приклад матриці відстаней, розрахованих за моделлю Tamura-Nei (рис. 37, а, б).

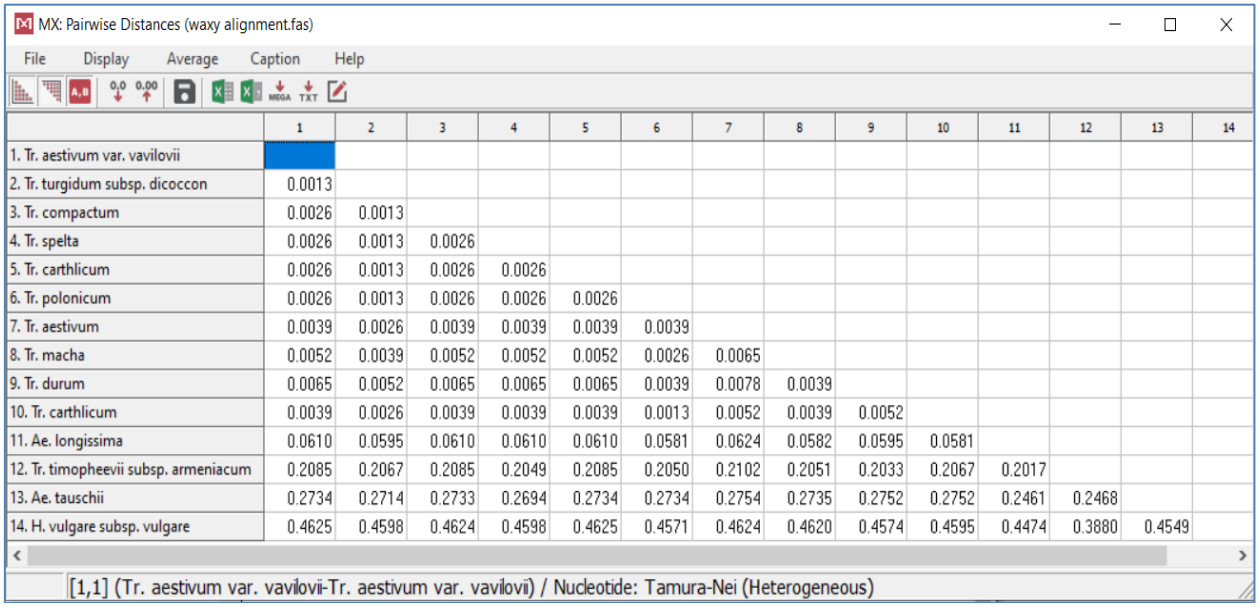

MX: Pairwise Distances (waxy alignment.fas)

File Display Average Caption Help

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Tr. aestivum var. vavilovii		0.0013	0.0018	0.0017	0.0019	0.0019	0.0024	0.0027	0.0029	0.0023	0.0088	0.0177	0.0213	0.0323
2. Tr. turgidum subsp. dicoccon	0.0013		0.0013	0.0013	0.0014	0.0013	0.0020	0.0023	0.0025	0.0018	0.0087	0.0177	0.0211	0.0322
3. Tr. compactum	0.0026	0.0013		0.0018	0.0019	0.0019	0.0024	0.0027	0.0029	0.0023	0.0088	0.0179	0.0211	0.0323
4. Tr. spelta	0.0026	0.0013	0.0026		0.0019	0.0018	0.0024	0.0026	0.0028	0.0022	0.0088	0.0178	0.0210	0.0322
5. Tr. carthlicum	0.0026	0.0013	0.0026	0.0026		0.0019	0.0024	0.0026	0.0029	0.0023	0.0089	0.0177	0.0210	0.0323
6. Tr. polonicum	0.0026	0.0013	0.0026	0.0026	0.0026		0.0024	0.0018	0.0022	0.0013	0.0087	0.0177	0.0212	0.0320
7. Tr. aestivum	0.0039	0.0026	0.0039	0.0039	0.0039	0.0039		0.0030	0.0032	0.0026	0.0088	0.0180	0.0213	0.0325
8. Tr. macha	0.0052	0.0039	0.0052	0.0052	0.0052	0.0026	0.0065		0.0022	0.0021	0.0088	0.0180	0.0215	0.0324
9. Tr. durum	0.0065	0.0052	0.0065	0.0065	0.0065	0.0039	0.0078	0.0039		0.0026	0.0091	0.0176	0.0213	0.0320
10. Tr. carthlicum	0.0039	0.0026	0.0039	0.0039	0.0039	0.0013	0.0052	0.0039	0.0052		0.0087	0.0177	0.0212	0.0321
11. Ae. longissima	0.0610	0.0595	0.0610	0.0610	0.0610	0.0581	0.0624	0.0582	0.0595	0.0581		0.0178	0.0204	0.0304
12. Tr. timopheevii subsp. armeniacum	0.2085	0.2067	0.2085	0.2049	0.2085	0.2050	0.2102	0.2051	0.2033	0.2067	0.2017		0.0212	0.0282
13. Ae. tauschii	0.2734	0.2714	0.2733	0.2694	0.2734	0.2734	0.2754	0.2735	0.2752	0.2752	0.2461	0.2468		0.0300
14. H. vulgare subsp. vulgare	0.4625	0.4598	0.4624	0.4598	0.4625	0.4571	0.4624	0.4620	0.4574	0.4595	0.4474	0.3880	0.4549	

d [14,13] (H. vulgare subsp. vulgare-Ae. tauschii) / Nucleotide: Tamura-Nei (Heterogeneous)

а


MX: Pairwise Distances (waxy alignment.fas)

File Display Average Caption Help

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Tr. aestivum var. vavilovii														
2. Tr. turgidum subsp. dicoccon	0.0013													
3. Tr. compactum	0.0026	0.0013												
4. Tr. spelta	0.0026	0.0013	0.0026											
5. Tr. carthlicum	0.0026	0.0013	0.0026	0.0026										
6. Tr. polonicum	0.0026	0.0013	0.0026	0.0026	0.0026									
7. Tr. aestivum	0.0039	0.0026	0.0039	0.0039	0.0039	0.0039								
8. Tr. macha	0.0052	0.0039	0.0052	0.0052	0.0052	0.0026	0.0065							
9. Tr. durum	0.0065	0.0052	0.0065	0.0065	0.0065	0.0039	0.0078	0.0039						
10. Tr. carthlicum	0.0039	0.0026	0.0039	0.0039	0.0039	0.0013	0.0052	0.0039	0.0052					
11. Ae. longissima	0.0610	0.0595	0.0610	0.0610	0.0610	0.0581	0.0624	0.0582	0.0595	0.0581				
12. Tr. timopheevii subsp. armeniacum	0.2085	0.2067	0.2085	0.2049	0.2085	0.2050	0.2102	0.2051	0.2033	0.2067	0.2017			
13. Ae. tauschii	0.2734	0.2714	0.2733	0.2694	0.2734	0.2734	0.2754	0.2735	0.2752	0.2752	0.2461	0.2468		
14. H. vulgare subsp. vulgare	0.4625	0.4598	0.4624	0.4598	0.4625	0.4571	0.4624	0.4620	0.4574	0.4595	0.4474	0.3880	0.4549	

[1,1] (Tr. aestivum var. vavilovii-Tr. aestivum var. vavilovii) / Nucleotide: Tamura-Nei (Heterogeneous)

б

Рис. 37. Матриця генетичних відстаней, розрахованих за моделлю Tamura-Nei в програмі MEGA X: а – з використанням бутстреп-аналізу, б – без оцінки значущості результатів

Додаткові параметри такі: тип замін – транзиції + трансверсії, рівномірний розподіл сайтів, гетерогенність послідовностей, повне вилучення пропущених та відсутніх даних, оцінка статистичної

достовірності результатів методом бутстреп (див. рис. 37, а) та без оцінки значущості (див. рис. 37, б).

Зміна параметра оцінки достовірності результатів призвела до зміни зовнішнього вигляду матриці.

Меню **DIVERSITY** є корисним для розрахунків показників генетичної різноманітності, які використовують у популяційній генетиці. Це, зокрема, показники субпопуляційної, популяційної та міжпопуляційної різноманітності, коефіцієнт диференціації.

Меню **PHYLOGENY** призначене для побудування філогенетичних дерев. Для кластеризації у MEGA можна обрати такі методи: максимальної подібності, приєднання сусідів (NJ), мінімальної еволюції, незваженого парного групування із середнім арифметичним (UPGMA), максимальної економії (рис. 38).

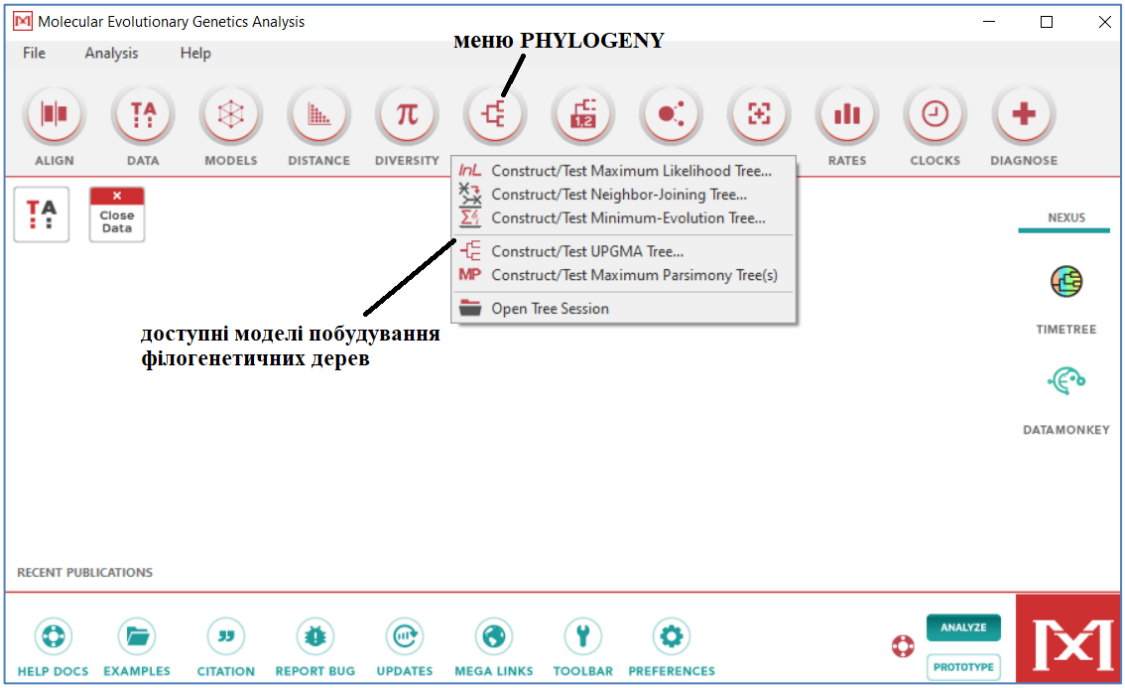


Рис. 38. Активне меню PHYLOGENY в програмі MEGA X

Після вибору методу кластеризації з’являється діалогове вікно, у якому можна задати додаткові параметри. Натискаємо OK – MEGA будує філогенетичне дерево і видає результат у новому вікні. Нижче наведено філогенетичне дерево, побудоване з використанням методу NJ (рис. 39).

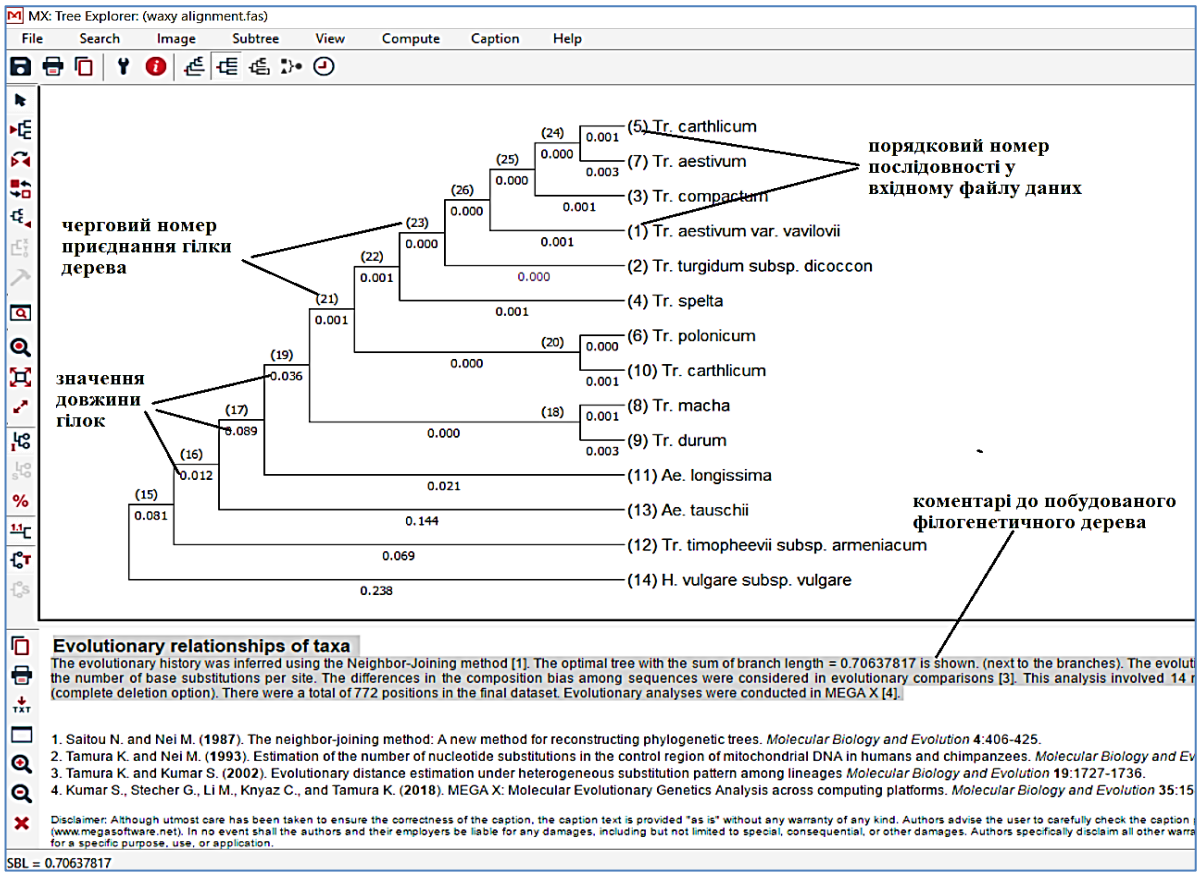


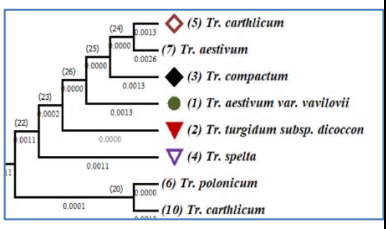
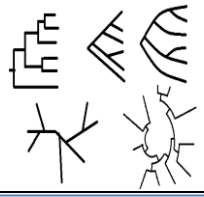
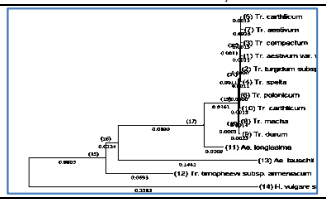
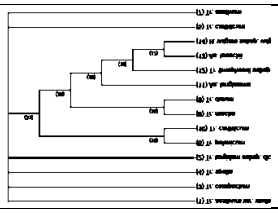
Рис. 39. Філогенетичне дерево, побудоване методом NJ у програмі MEGA X

Вікно з деревом має панель управління, вертикальну і горизонтальну панелі інструментів, за допомогою яких можна зберегти, експортувати або роздрукувати результат; зберегти дерево як рисунок у форматі *.bmp, *.pdf, *.pgn, *.sgn, *.tiff, *.emf; форматувати дерево або окремі його гілки; розфарбувати окремі гілки або кластери різними кольорами; змінити масштаб дерева; додати або видалити підписи тощо. У табл. 10 наведено деякі функції щодо редагування філогенетичного дерева в MEGA.

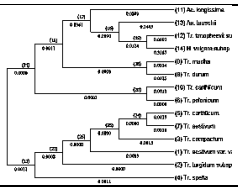
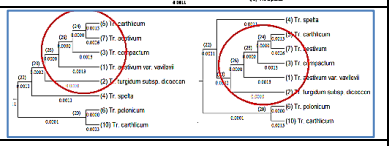
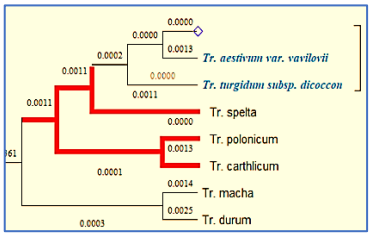
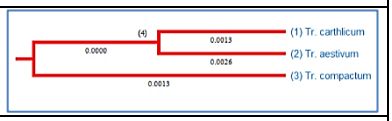
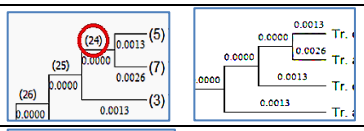
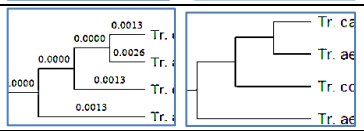
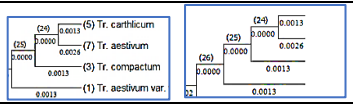
Меню **USER TREE** призначене для роботи з деревами, створеними користувачем. За допомогою цього інструменту можна порівняти дерево користувача з поточними даними послідовностей, редагувати топологію таких дерев, перевірити їх за моделями максимальної подібності, найменшого квадрата, парсимонії.

Меню **ANCESTORS** є опцією роботи з родоводами. Дозволяє на філогенетичному дереві оцінити вузли предків та підвищити їх імовірність, а також побудувати біологічну послідовність предкових форм або запропонувати декілька варіантів.

10. Функції редагування та форматування філогенетичних дерев в програмі MEGA X

Значок функції	Функція	Зовнішній вигляд результату
1	2	3
	Функція налаштування. Дозволяє змінити товщину і довжину гілок, відстань між ними; додати/ прибрати підписи; змінити колір, тип та розмір шрифту підписів; поставити мітку біля окремих зразків	
	Відкриває вікно, яке містить інформацію про побудоване дерево (за якими параметрами будувалося, тип використаних даних, скільки зразків проаналізовано тощо)	
	Змінює тип дерева: традиційний (прямокутний, прямий або вигнутий), радіальний або круговий	
	Відображає дерево у формі зв’язків між таксонами, ігноруючи довжину гілок	
	Коли внутрішні гілки філогенетичного дерева не мають статистично значущої довжини, вибір цієї команди конденсує дерево у топологію, в якій кожна гілка з меншою, ніж бажана, статистичною значимістю згортається	Функція не була активною для нашого прикладу
	Будує консенсусне дерево для дерев, одержаних методом максимальної економії	
	Розраховує час дивергенції на основі довжини гілок дерева	Функція не була активною для нашого прикладу
	Активує інструмент-вказівник	

Продовження табл. 10

1	2	3
	Укорінює дерево за обраною гілкою	
	Міняє місцями виділені вузли	
	Перевертає вибрані вузли на 180°	Ефект такий самий
	Компресує виділені вузли	Ефект такий самий
	Відкриває діалогове вікно, у якому можна задати параметри форматування окремих гілок дерева. Дозволяє змінити колір гілок та підписів, змінити тип та розмір шрифту підписів, додати дужки для виділення кластерів для окремих гілок, а також стискати виділені гілки (на дереві компресовані гілки об'єднуються в одну без зазначення таксонів, які входили у цю гілку). Функція активна, лише коли увімкнений інструмент-вказівник	
	Відкриває виділені вузли у новому вікні	
	Включає режим автомасштабування дерева	
	Додає/видаляє підписи порядкових номерів приєднання гілок	
	Додає/видаляє підписи довжини гілок	
	Додає/видаляє підписи таксонів (зразків) поруч з гілками	

Меню **SELECTION** містить інструменти, призначені для оцінки спрямування еволюції (позитивна, негативна або нейтральна) на основі аналізу кодонів, синонімічних і несинонімічних вставок нуклеотидів з використанням Z-тесту, критерію Фішера або тесту на нейтральність Tajima.

Меню **RATES** призначено для оцінки параметрів окремих сайтів послідовностей. Воно дозволяє встановити швидкість еволюції на кожній ділянці вирівнювання.

Меню **CLOCKS** дає змогу здійснити оцінку гіпотези молекулярного годинника для поточного масиву даних, а також конструювання часових дерев.

Меню **DIAGNOSE** призначено для проведення діагностики мутацій у послідовностях та їх впливу на еволюційний процес.

У MEGA запрограмовано можливість збереження результатів по кожній функції у форматі таблиць та рисунків. Надалі їх можна використовувати як графічний матеріал для візуалізації даних під час оформлення наукових праць.

Відзначимо, що MEGA містить інструкцію з детальним описом команд, функцій та моделей, які використовують для аналізу біологічних послідовностей і маніпуляцій з ними, а також корисні посилання на літературні джерела. Інструкція доступна за вкладкою **Help** на панелі інструментів головного вікна програми, а також у діалогових вікнах з результатами за кожною функцією.

Програма STRUCTURE

Structure – статистична програма кластерного аналізу для вивчення структури популяцій за полілокусними генотиповими даними. Ця програма доступна безкоштовно через Інтернет-ресурс: http://pritchardlab.stanford. edu/structure.html.

Програму створено на основі математичного алгоритму, який називається *байєсівський алгоритм*. У цій програмі розраховано частку належності кожного окремого генотипу до певної популяції (кількісний кластерний аналіз). За допомогою Structure можна чітко виділяти популяції, оцінювати належність індивідів до певної популяції, виявляти зони гібридизації та проміжні індивіди.

Аналіз у Structure можна застосовувати для більшості розповсюджених нині молекулярних маркерів, включаючи SNP та SSR, також можна використовувати фенотипові дані.

Установлення програми. Для встановлення програми необхідно завантажити архів з файлами на сайті програми (http://pritchardlab.stanford.edu/structure_software/release_versions/v2.3.4/html/structure.html), розпакувати його на обраному диску. Далі запустити файл *.exe та виконати інструкції майстра встановлення.

Після цього на робочому столі комп'ютера з'явиться позначка, а на диску, куди встановили програму, папка Structure 2.3.4. У цій папці містяться декілька внутрішніх папок. Більшість з них мають системні файли, необхідні для роботи програми. Проте є дві папки, які можуть знадобитися користувачу:

1) **help files** – містить інструкцію до програми;

2) **samples** – містить приклади файлів з вихідними даними.

Підготовка вихідних даних. Файл з вихідними даними має відповідати таким вимогам:

1) об'єкти повинні бути розташовані в рядках, а локуси – у стовпчиках;

2) дані мають бути представлені тільки цілими числами;

3) кількість рядків даних для кожного об'єкта повинна відповідати типу даних (гаплоїдні дані – один рядок, диплоїдні дані – два рядки);

4) дані, які відсутні, слід позначати числом, яке більше не використовують у матриці даних;

5) файл з вихідними даними потрібно зберігати в текстовому (*.txt) форматі, а не у форматі Exel (*.xls).

Імпортування даних у програму. Для імпортування даних у Structure необхідно відкрити програму подвійним натисканням на ярлик програми. Потім виконати такі кроки: **File**, потім **Open Data File** та обрати папку, де зберігається файл з вихідними даними. Потім необхідно вибрати файл з вихідними даними. Після цього відкриється вихідний файл у вікні Structure (рис. 40).

Початок роботи. Необхідно вибрати команди **File** та **New Project.** Відкриється вікно (рис. 41), у якому слід:

а) дати проекту будь-яку назву у графі **Project Name**;

б) обрати папку, де зберігається файл з вихідними даними (**Select directory**);

в) вибрати файл з вихідними даними (**Choose data file**).

Потім необхідно виконати крок, натиснувши клавішу Next, після чого з'явиться вікно (рис. 42).

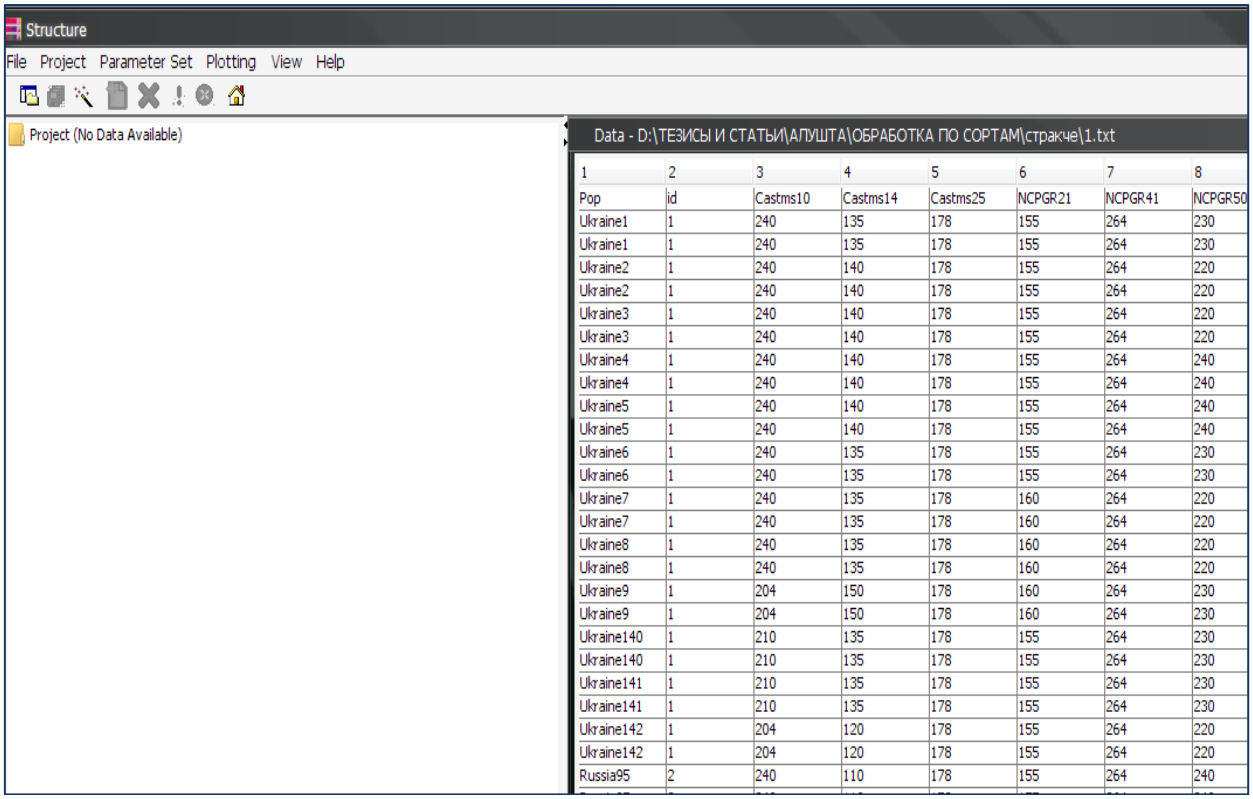
У цьому вікні потрібно зазначити таку інформацію:

а) кількість зразків (Number of Individuals);

б) плоїдність (Ploidy data) – якщо використовуються диплоїдні дані, то файл слід записати у дворядковому форматі і вказати цифру 2; якщо ж гаплоїдні дані, то вихідний файл записують в однорядковому форматі і ставлять цифру 1;

в) кількість вивчених ознак (маркерів) (Number of loci);

г) позначення пропущених даних (Missing data value), наприклад, «-1».



1	2	3	4	5	6	7	8
Pop	id	Castms10	Castms14	Castms25	NCPGR21	NCPGR41	NCPGR50
Ukraine1	1	240	135	178	155	264	230
Ukraine1	1	240	135	178	155	264	230
Ukraine2	1	240	140	178	155	264	220
Ukraine2	1	240	140	178	155	264	220
Ukraine3	1	240	140	178	155	264	220
Ukraine3	1	240	140	178	155	264	220
Ukraine4	1	240	140	178	155	264	240
Ukraine4	1	240	140	178	155	264	240
Ukraine5	1	240	140	178	155	264	240
Ukraine5	1	240	140	178	155	264	240
Ukraine6	1	240	135	178	155	264	230
Ukraine6	1	240	135	178	155	264	230
Ukraine7	1	240	135	178	160	264	220
Ukraine7	1	240	135	178	160	264	220
Ukraine8	1	240	135	178	160	264	220
Ukraine8	1	240	135	178	160	264	220
Ukraine9	1	204	150	178	160	264	230
Ukraine9	1	204	150	178	160	264	230
Ukraine140	1	210	135	178	155	264	230
Ukraine140	1	210	135	178	155	264	230
Ukraine141	1	210	135	178	155	264	230
Ukraine141	1	210	135	178	155	264	230
Ukraine142	1	204	120	178	155	264	220
Ukraine142	1	204	120	178	155	264	220
Russia95	2	240	110	178	155	264	240

Рис. 40. Діалогове вікно файлу з вихідними даними

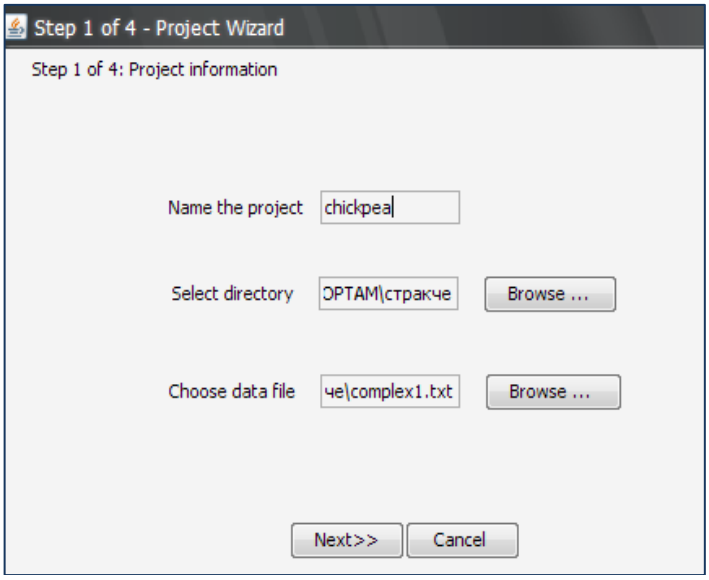


Рис. 41. Діалогове вікно про інформацію проекту

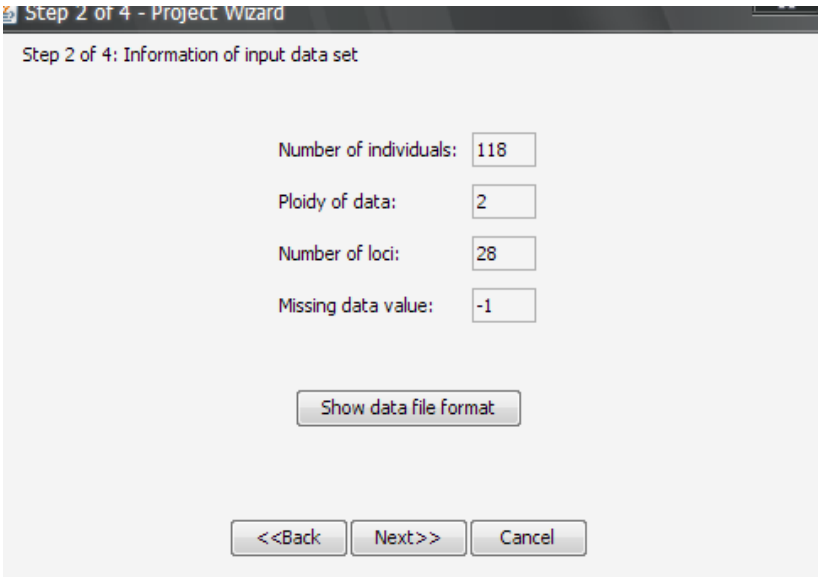


Рис. 42. Вікно вводу інформації про вихідні дані

Після цього необхідно натиснути клавішу **Next**. Також у цьому вікні доступна опція **Повернутись** до попереднього вікна **(Back**), де можна змінити дані, та опція **Відміна (Cancel)**, якщо потрібно відмінити весь проект.

Після натискання клавіші **Next** з'явиться вікно (рис. 43), у якому необхідно відмітити «галочкою», які рядки даних містять файл з вихідними даними:

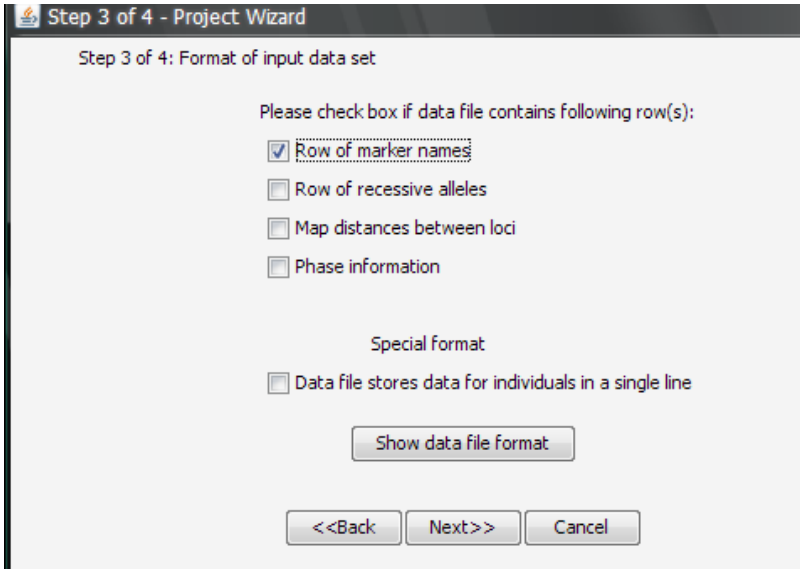


Рис. 43. Вікно для вибору рядків даних, які записані у вихідному файлі

а) row of marker names – рядок з назвами маркерів;

б) row of recessive alleles – рядок з рецесивними алелями;

в) map distances between loci – матриця відстаней між локусами;

г) phase information – інформація про фазу. Використовують тільки для диплоїдних даних при застосуванні моделі зчеплення;

д) special information – стовпчики з додатковою інформацією, які можуть бути необхідними користувачу, але ігноруються програмою.

Потім потрібно знову натиснути клавішу **Next**. Також у цьому вікні доступна клавіша **Повернутись** до попереднього вікна **(Back)**, де можна змінити дані, та клавіша **Відміна (Cancel)**, якщо необхідно відмінити увесь проект.

Після натискання клавіши Next з'явиться вікно (рис. 44), у якому слід відмітити дані, які містить вихідний файл:

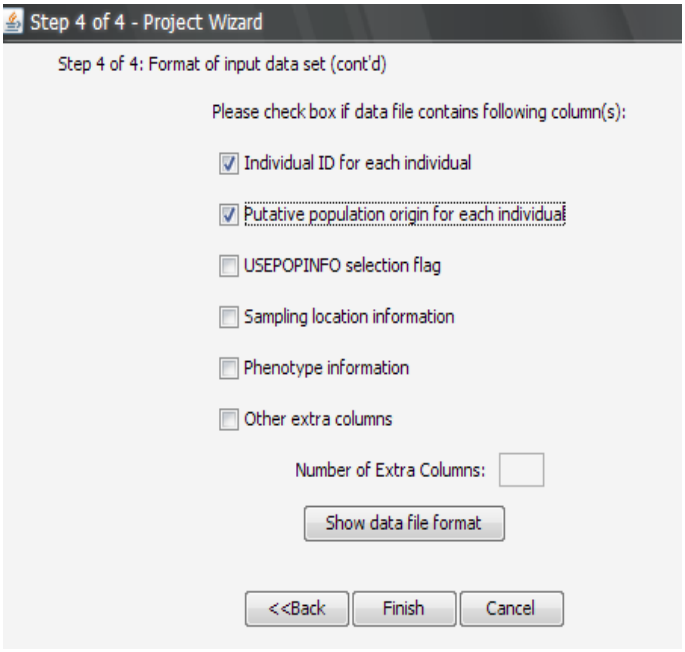


Рис. 44. Вікно для вибору типу даних, які містяться у вихідному файлі

а) individual Id for each population – назва кожного зразка;

б) putative population origin of each individual – передбачуване походження (популяція) кожного зразка;

в) USERPOPINFO – функція, необхідна для вирішення таких завдань:

- визначення походження мігрантів – у випадках, якщо користувачу точно відомо походження більшох кількості зразків (на основі попередніх досліджень, даних систематики тощо) та необхідно встановити походження невеликої групи об'єктів, які, ймовірно, можуть бути мігрантами однієї з відомих популяцій;

- визначення походження невідомих об’єктів на основі наявної інформації про походження інших об’єктів;

г) phenotype information – фенотипові дані;

д) other extra columns – інші додаткові колонки.

Якщо додаткові колонки є, то необхідно відмітити їх кількість у графі Number of extra columns.

Потім необхідно перейти до наступного вікна шляхом натискання кнопки **Next**. Після цього з’явиться вікно з узагальненою інформацією про створений проект (рис. 45).

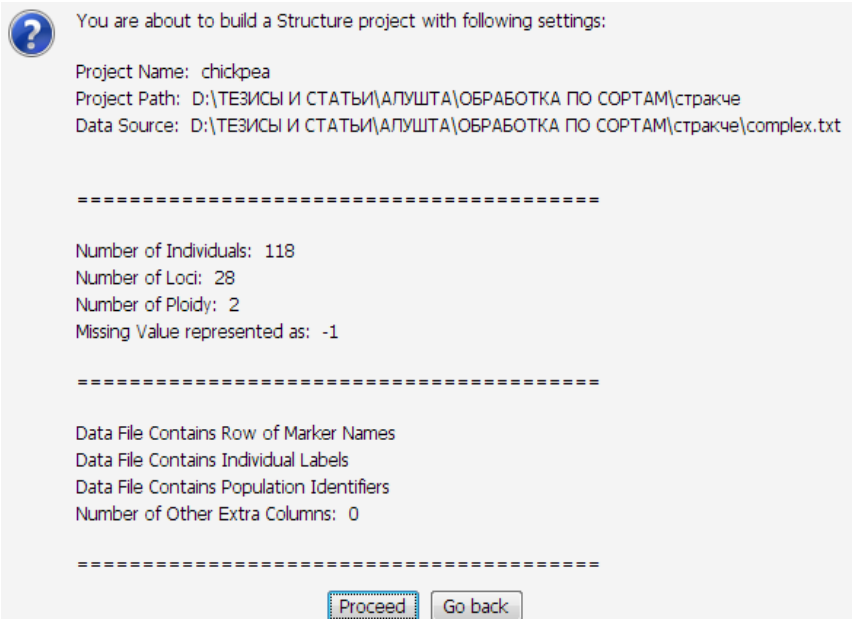


Рис. 45. Узагальнена інформація про створений проект

Далі необхідно активувати опцію **Proceed**. Якщо при створенні проекту були помилки, програма генерує повідомлення – **Bad Format** (неправильний формат). У цьому випадку необхідно натиснути опцію **Go back** (повернутись) та виправити помилки.

Якщо створення проекта успішне, то на екрані програми з’явиться таблиця з вихідними даними (рис. 46).

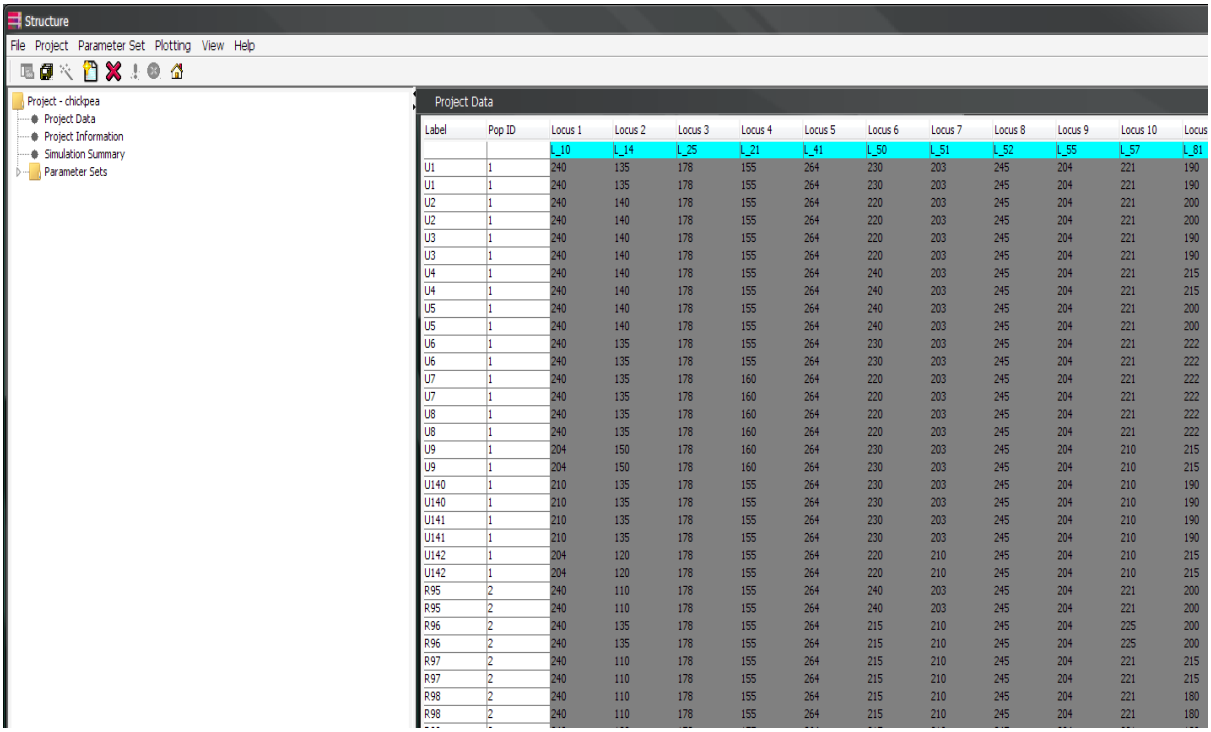
Запис параметрів для розрахунків (Parameter Set). Необхідно виконати такі кроки:

1) зайти у вкладку **Parameter Set – New**. На екрані з’явиться вікно із внутрішніми вкладками (рис. 47);

2) перейти у внутрішню вкладку – **Run Length** і встановити такі параметри:

- **Length of Burnin Period** – час, за який повинен запускатись математичний алгоритм перед початком розрахунків – 5000;

- **Number of MCMC Reps after Burnin** – час, за який повинен працювати математичний алгоритм після закінчення розрахунків для отримання коректних результатів – 50000;



Label	Pop ID	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5	Locus 6	Locus 7	Locus 8	Locus 9	Locus 10	Locus
		L_10	L_14	L_25	L_21	L_41	L_50	L_51	L_52	L_55	L_57	L_81
U1	1	240	135	178	155	264	230	203	245	204	221	190
U1	1	240	135	178	155	264	230	203	245	204	221	190
U2	1	240	140	178	155	264	220	203	245	204	221	200
U2	1	240	140	178	155	264	220	203	245	204	221	200
U3	1	240	140	178	155	264	220	203	245	204	221	190
U3	1	240	140	178	155	264	220	203	245	204	221	190
U4	1	240	140	178	155	264	240	203	245	204	221	215
U4	1	240	140	178	155	264	240	203	245	204	221	215
U5	1	240	140	178	155	264	240	203	245	204	221	200
U5	1	240	140	178	155	264	240	203	245	204	221	200
U6	1	240	135	178	155	264	230	203	245	204	221	222
U6	1	240	135	178	155	264	230	203	245	204	221	222
U7	1	240	135	178	160	264	220	203	245	204	221	222
U7	1	240	135	178	160	264	220	203	245	204	221	222
U8	1	240	135	178	160	264	220	203	245	204	221	222
U8	1	240	135	178	160	264	220	203	245	204	221	222
U9	1	204	150	178	160	264	230	203	245	204	210	215
U9	1	204	150	178	160	264	230	203	245	204	210	215
U140	1	210	135	178	155	264	230	203	245	204	210	190
U140	1	210	135	178	155	264	230	203	245	204	210	190
U141	1	210	135	178	155	264	230	203	245	204	210	190
U141	1	210	135	178	155	264	230	203	245	204	210	190
U142	1	204	120	178	155	264	220	210	245	204	210	215
U142	1	204	120	178	155	264	220	210	245	204	210	215
R95	2	240	110	178	155	264	240	203	245	204	221	200
R95	2	240	110	178	155	264	240	203	245	204	221	200
R96	2	240	135	178	155	264	215	210	245	204	225	200
R96	2	240	135	178	155	264	215	210	245	204	225	200
R97	2	240	110	178	155	264	215	210	245	204	221	215
R97	2	240	110	178	155	264	215	210	245	204	221	215
R98	2	240	110	178	155	264	215	210	245	204	221	180
R98	2	240	110	178	155	264	215	210	245	204	221	180

Рис. 46. Вікно з вихідними даними

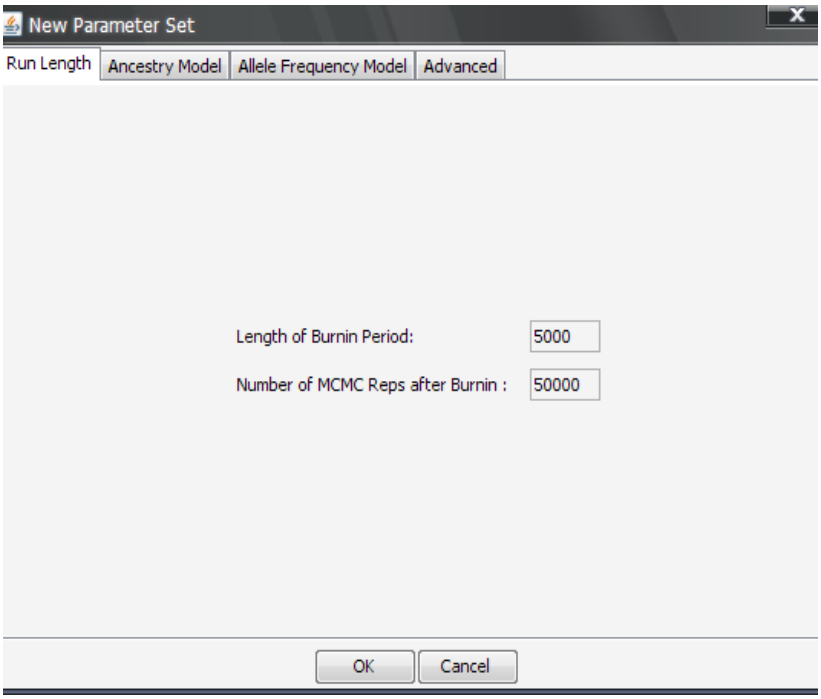


Рис. 47. Вікно запису параметрів роботи математичного алгоритму

3) перейти у внутрішню вкладку **Ansentry model** і вибрати ту модель походження, яка відповідає експериментальним даним (рис. 48):

Use No Admixture model – ймовірно, усі досліджувані об’єкти мають спільне походження;

Use Admixture model – ймовірно, усі досліджувані об’єкти мають різне походження;

Use population information – використовувати в розрахунках інформацію про популяції, яка вказана у файлі з вихідними даними.

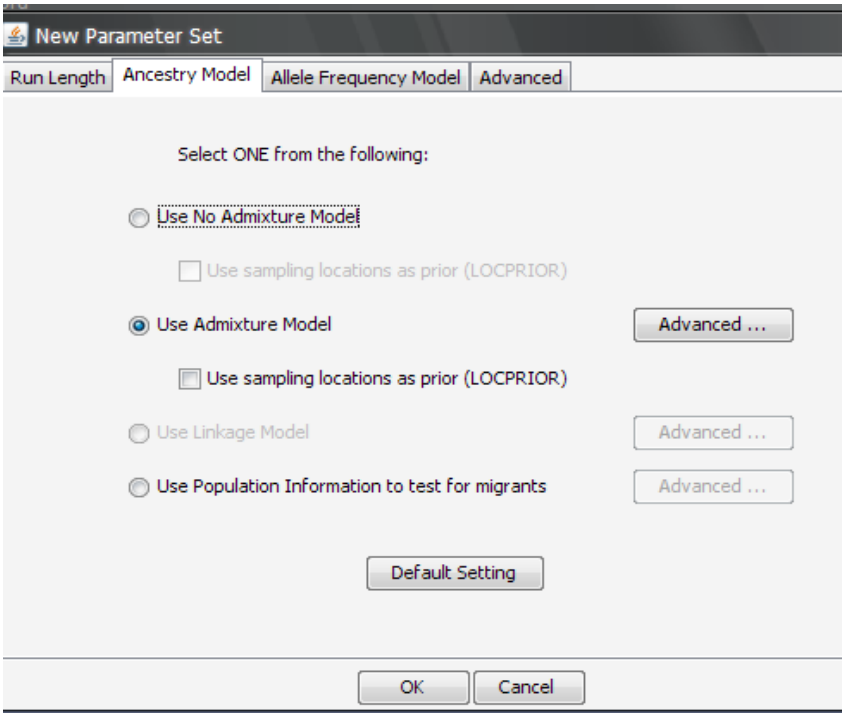


Рис. 48. Вікно для вибору моделі походження для вихідних даних

4) обрати внутрішню вкладку **Allele Frequency Model** та поставити мітку на одному з трьох варіантів (рис. 49):

Allele Frequencies Correlated – модель розподілу взаємопов’язаних частот алелів;

Allele Frequencies Independent – модель розподілу незалежних частот алелів;

Infer Lambda – модель розподілу незалежних частот алелів з використанням коефіцієнта λ, яка запропонована Pritchard et al.

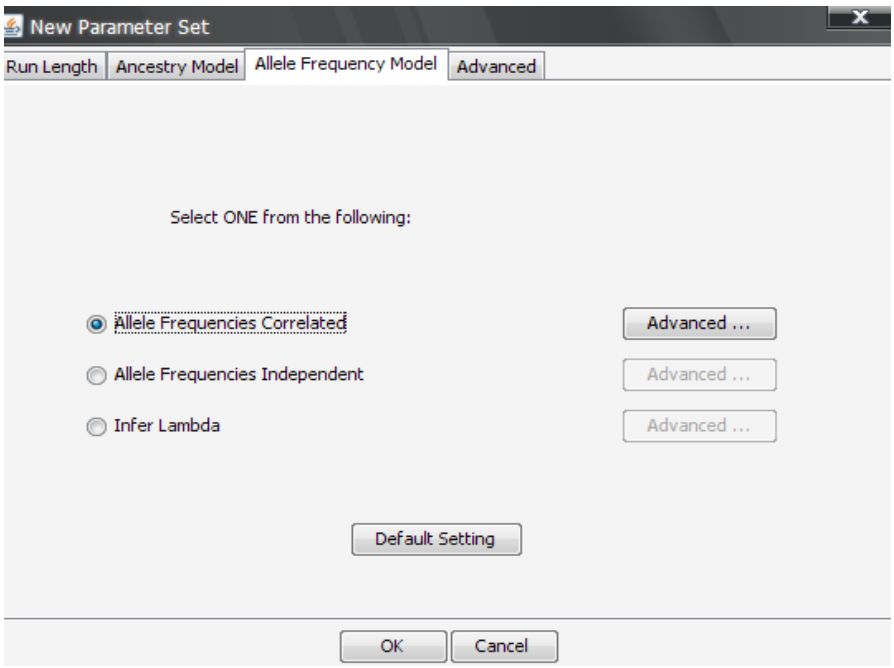


Рис. 49. Вікно вибору моделі типу розподілу алелів

Автори програми рекомендують використовувати для розрахунків моделі **Allele Frequencies Correlated** або **Allele Frequencies Independent**;

5) перейти у внутрішню вкладку **Advanced** та, якщо необхідно, обрати додаткові параметри (рис. 50).

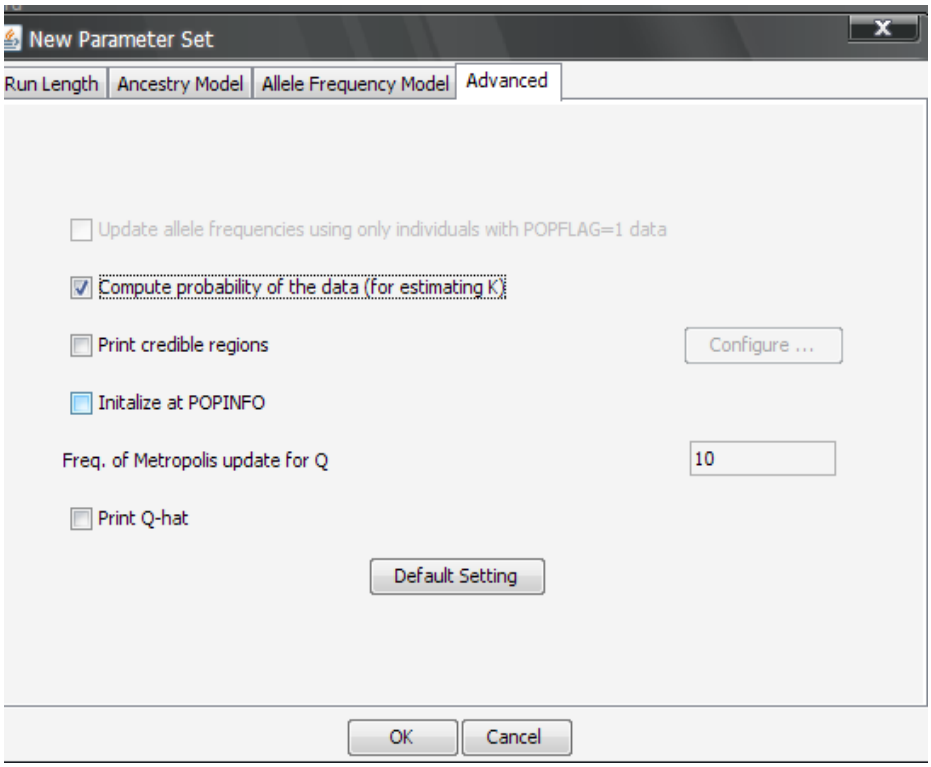


Рис. 50. Вікно для вибору додаткових параметрів

Для скорочення часу розрахунків автори рекомендують відключити функцію **Compute probability of data (for estimating K)** – розрахунок показників вірогідностей. Проте, якщо необхідне визначення статистично значущої кількості кластерів (див. нижче), цю функцію відключати не можна. З іншими статистичними параметрами, їх розрахунком, що доступно у Structure, можна ознайомитися в інструкції до програми, яку розміщено в папці Help files.

Після того, як усі параметри у вкладці **Parameter Set** будуть установлені, необхідно натиснути клавішу OK. З'явиться вікно, у якому слід записати ім'я нових параметрів розрахунку (рис. 51).

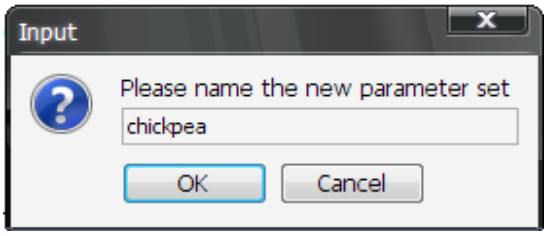


Рис. 51. Вікно для введення імені файлу

У рядку необхідно прописати вибране ім'я параметрів та натиснути OK. На екрані з'явиться вікно, у якому прописано встановлені параметри для розрахунків (рис. 52).

Запуск розрахунків. Після того, як матриця із даними та параметри розрахунків сформовані, можна приступати до запуску розрахунків. Для цього необхідно натиснути **Parameter set** та **Run**. Після цього відкриється вікно, де необхідно задати кількість імовірних кластерів.

Після закінчення розрахунків з'явиться вікно з результатами – **Simulation result**.

Аналіз результатів. Вікно **Simulation result** містить таку інформацію:

1. Частка належності кожної вивченої популяції до певного кластера (**Proportion of membership of each pre-defined population in each of the 3 clusters**), яку оцінено за коефіцієнтом Q;

2. Дивергенція між популяціями, оцінена за частотами алелів (**Allele-freq. divergence among pops (Net nucleotide distance)**);

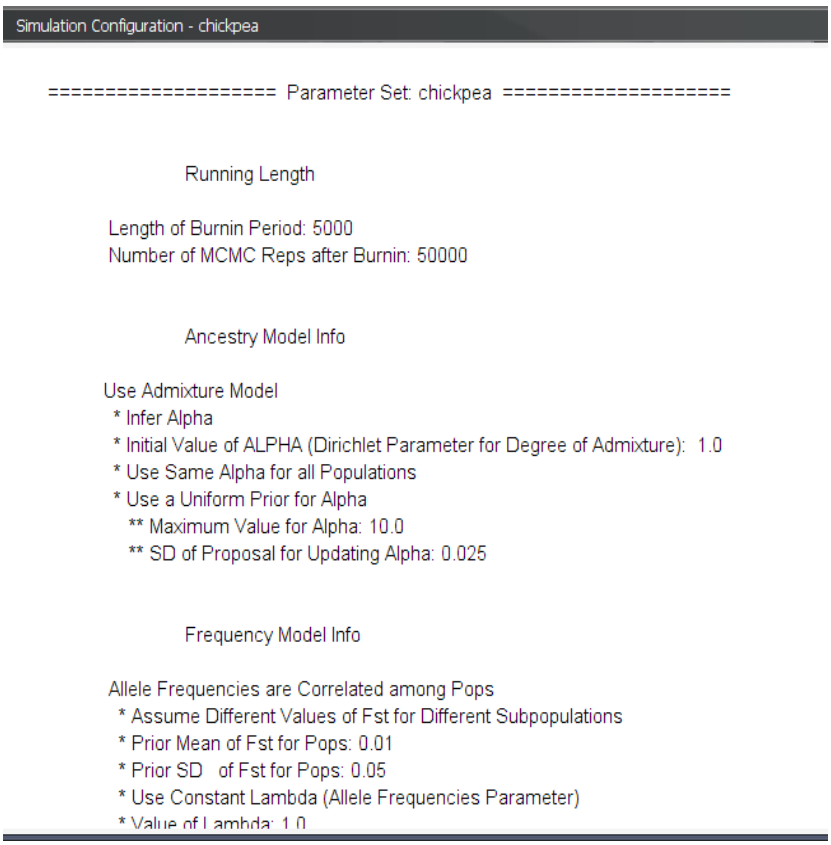


Рис. 52. Інформація про встановлені параметри для розрахунку

3. Середні генетичні відстані між популяціями – очікувана гетерозиготність (**Average distances (expected heterozygosity) between individuals in same cluster**));

4. Проміжні розрахункові дані та коефіцієнти (Alpha – поточне значення розподілу Дірліхе, який показує відсоток змішаності популяції, Corr (коефіцієнт кореляції між локусами), D1, 2 (дивергенція між популяціями), Ln Like, Est Ln P(D) – коефіцієнти вірогідності;

5. Кількісна належність кожного об’єкта до певної популяції (**Inferred ancestry of individuals**), оцінювана за коефіцієнтом Q;

6. Частоти алелів за кожним локусом, а також відсоток пропущених даних (**missing data**).

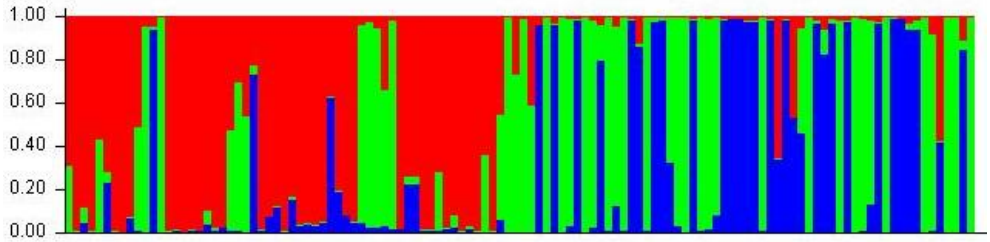
У вікні **Simulation result** є також чотири внутрішні вкладки, за допомогою яких можна графічно візуалізувати результати аналізу.

Візуалізація результатів. **Bar plot** (стовпчикова діаграма, гістограма). Кожного індивіда представлено у вигляді вертикального стовпчика, стовпчики об’єднано в блоки за певним кольором. Кожний колір відповідає одному з кластерів. Об’єкти, які складаються із сегментів різних кольорів, займають проміжне

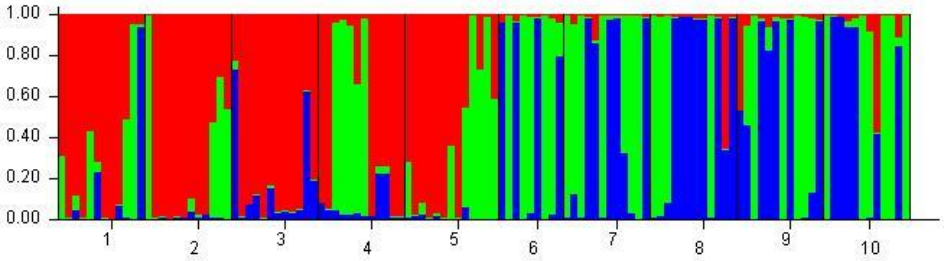
місце між кластерами. Кількісну оцінку належності таких об'єктів до певного кластера можна здійснити за допомогою коефіцієнта Q.

У програмі доступно чотири типи гістограм:

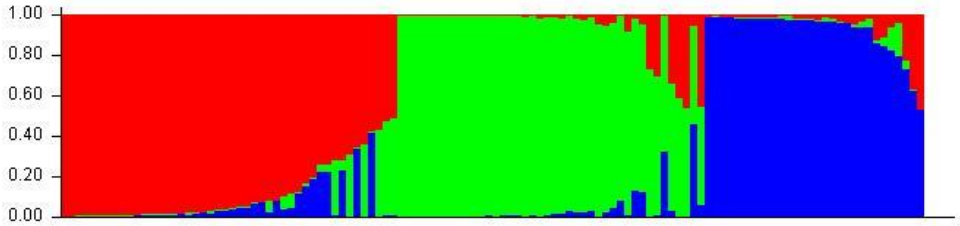
а) оригінальна (**original order**) – представлення кластерів об'єктів за результатами аналізу;



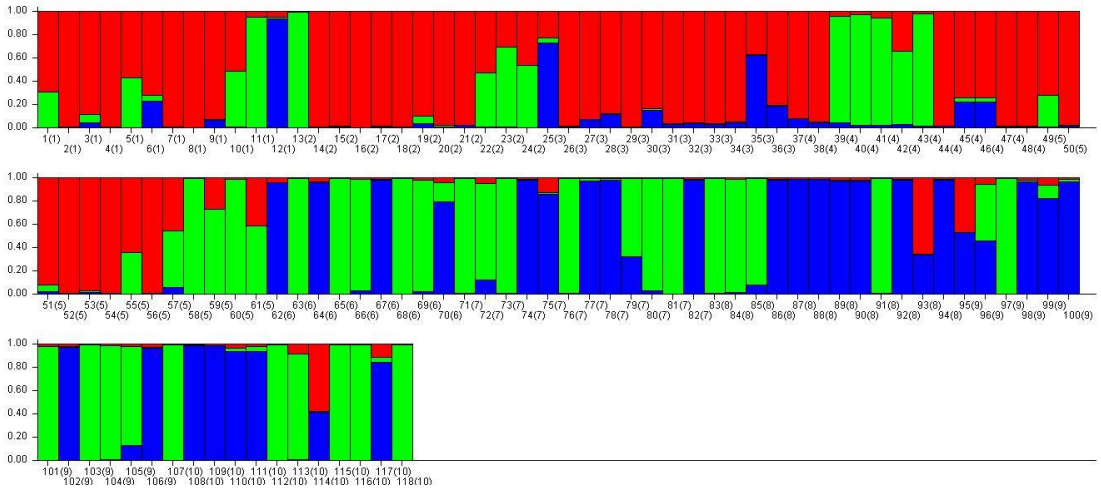
б) об'єднання об'єктів за популяціями (**group by POP id**);



в) об'єднання об'єктів за коефіцієнтом Q (**Sort by Q**);



г) усі типи гістограм може бути представлено однією лінією (**plot in single line**) або за допомогою множинних ліній (**plot in multiple lines**). В останньому випадку гістограма структурується на окремі об'єкти, як показано нижче.



Трикутна діаграма (Triangle Plot). Кожний об'єкт представлено у вигляді кольорової точки. Колір точок відповідає певній популяції. Коли масив даних представлений трьома кластерами (K = 3), тоді представники кожного кластера тяжіють до одного з кутів трикутника. Якщо масив даних класифікований на більшу кількість кластерів (K > 3), то користувачу пропонується обрати два кластери, які будуть відображатись окремо у двох кутах трикутника. Представники решти кластерів у такому випадку будуть відображатись разом у третьому куті трикутника (рис. 53).

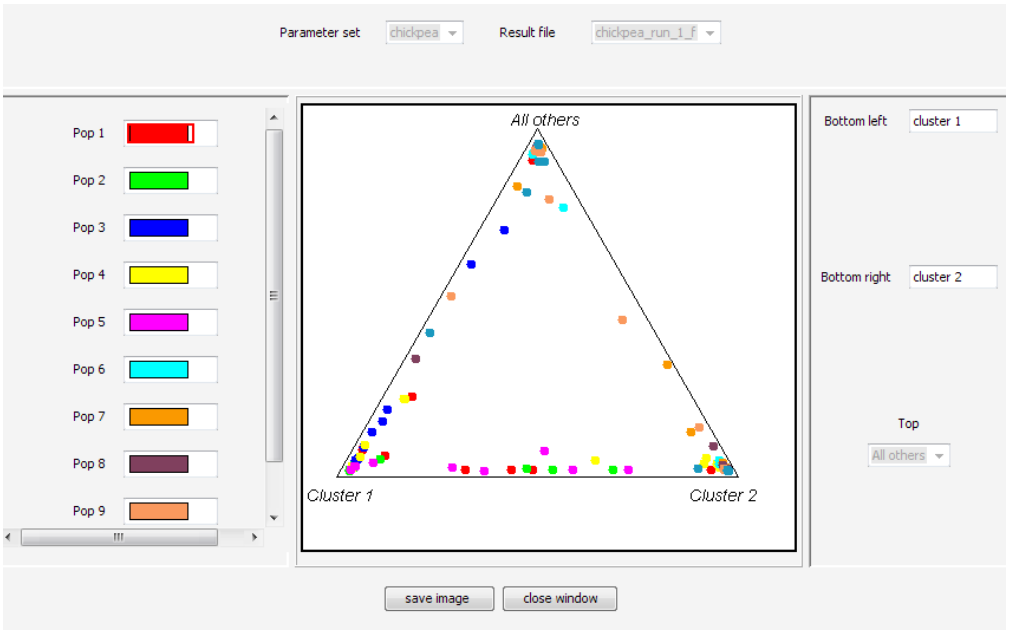


Рис. 53. Вигляд трикутної діаграми

Дерево (Tree plot)

Графік цього типу представляє генетичні відстані між кластерами. Дерево будують методом приєднання найближчих сусідів (Neighbor) (рис. 54). Візуалізація дерева здійснюється за допомогою програми **DRAWTREE** з пакета **PHYLIP**.

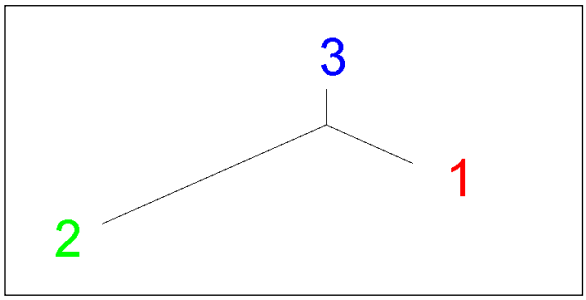


Рис. 54. Загальний вигляд дерева

Визначення статистично значущої кількості кластерів. У програмі Structure при виборі кількості ймовірних кластерів є частка суб'єктивізму. Evano зі співавторами (2005) запропонували математичний алгоритм, який дозволяє визначати статистично значущу кількість кластерів у програмі Structure (рис. 55).

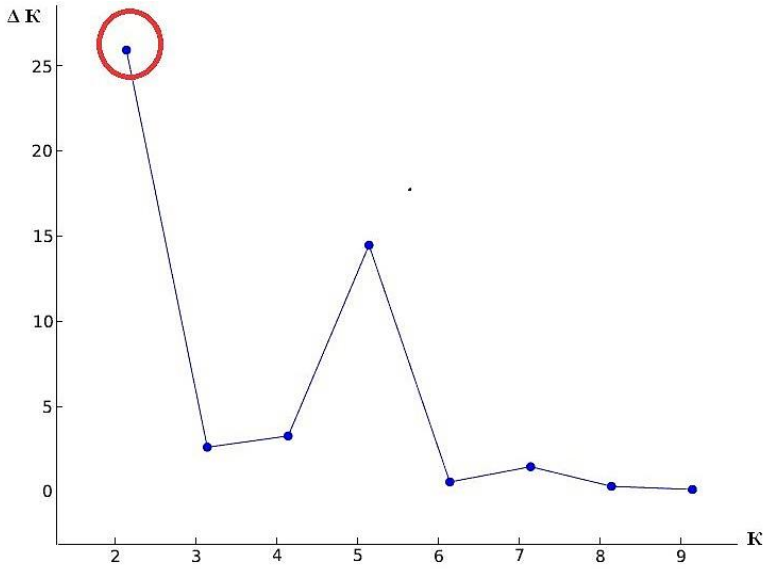


Рис. 55. Залежність кількості кластерів (K) від показника Δ K. Найвищий пік відповідає статистично значущій кількості кластерів

Для цього необхідно виконати такі кроки:

1. Підготувати дані і встановити параметри розрахунків, як описано вище. Потім натиснути **Project-Start a Job**. На екрані з'явиться наведене нижче вікно (рис. 56).

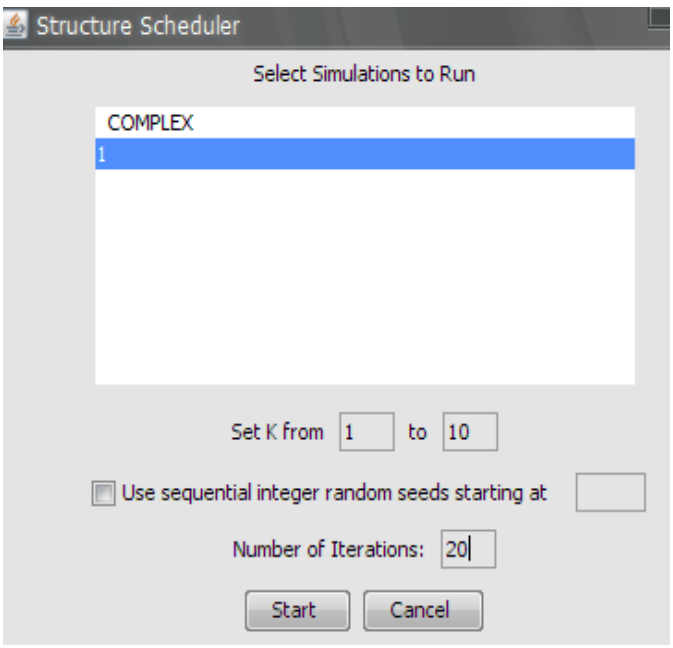


Рис. 56. Вікно для запису параметрів розрахунку

У цьому вікні необхідно задати такі параметри:
- кількість кластерів (**Set K**) – від 1 до 10;
- кількість повторень розрахунків (**Number of Iteration**) – 20;
- параметри для розрахунків.

2. Далі натиснути клавішу **Start**. Почнеться процес розрахунків, який залежно від параметрів комп'ютера може тривати декілька годин. Після закінчення розрахунків з'явиться повідомлення про завершення процесу – **Job is Completed**.

3. На диску, де збережений файл з вихідними даними, з'явиться папка з назвою проекту розрахунків, у цьому випадку Chickpea. У цій папці міститься ще дві вкладені папки – **PlotData** та **Results**. Папку **Results** необхідно додати у **zip**-архів.

Увага! Інші типи архівів недопустимі!!!

4. Потім слід відкрити Інтернет-сторінку http://taylor0.biology.ucla.edu/structureHarvester/ з онлайн-додатком **Structure Harvester**, який дозволяє реалізувати алгоритм Evanno et al (рис. 57).

На сторінці **Structure Harvester** є клавіша «Оберіть файл». Необхідно натиснути її та у відповідній папці на комп'ютері обрати **zip**-архів з результатами аналізу. Після цього на екрані з'явиться серія графіків і таблиць.

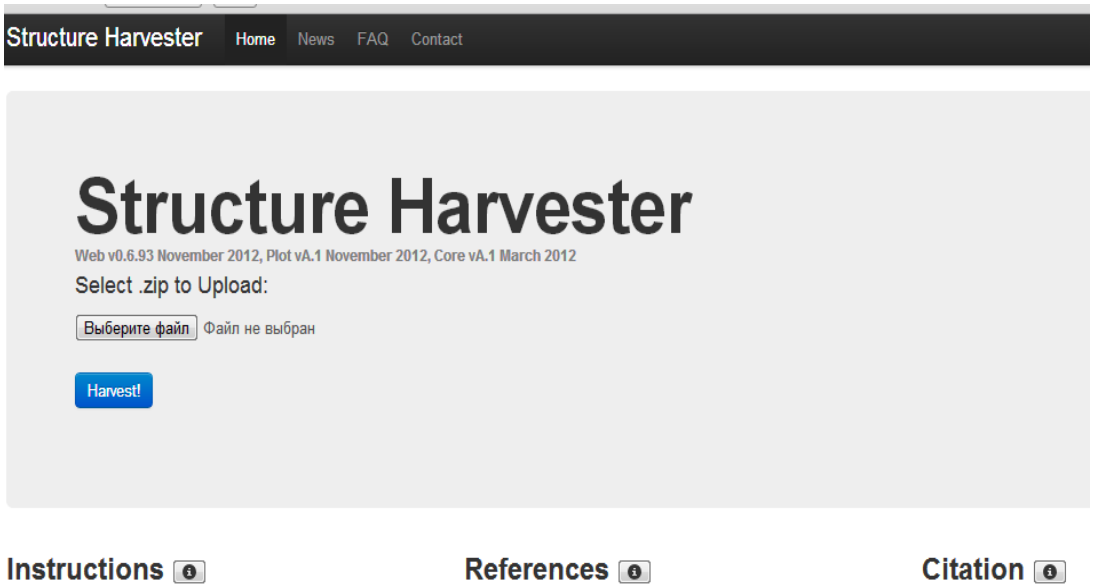


Рис. 57. Інтернет-сторінка з онлайн-додатком

Програма AmplifX

Це програма, розроблена N. Jullien та A.-M. Univ (CNRS, INP, Інститут нейрофізіопатології, Марсель, Франція) для конструювання та перевірки праймерів за показниками – Tm, якість, довжина, формування власних баз даних, тестування праймерів для ПЛР *in silico* (комп'ютерне моделювання, симуляція).

AmplifX сумісна з операційними системами Windows, Mac OS, Linux. Програма є безкоштовною і доступною для вільного завантаження з платформи https://inp.univ-amu.fr/en/amplifx-manage-test-and-design-your-prime rs-for-pcr.

Робота з програмою. Для інсталяції програми на ПК потрібно відкрити установчий файл. Ярлик програми для зручності доступу можна винести на робочий стіл ПК. Після активації програми на робочому столі відкривається головне вікно (рис. 58).

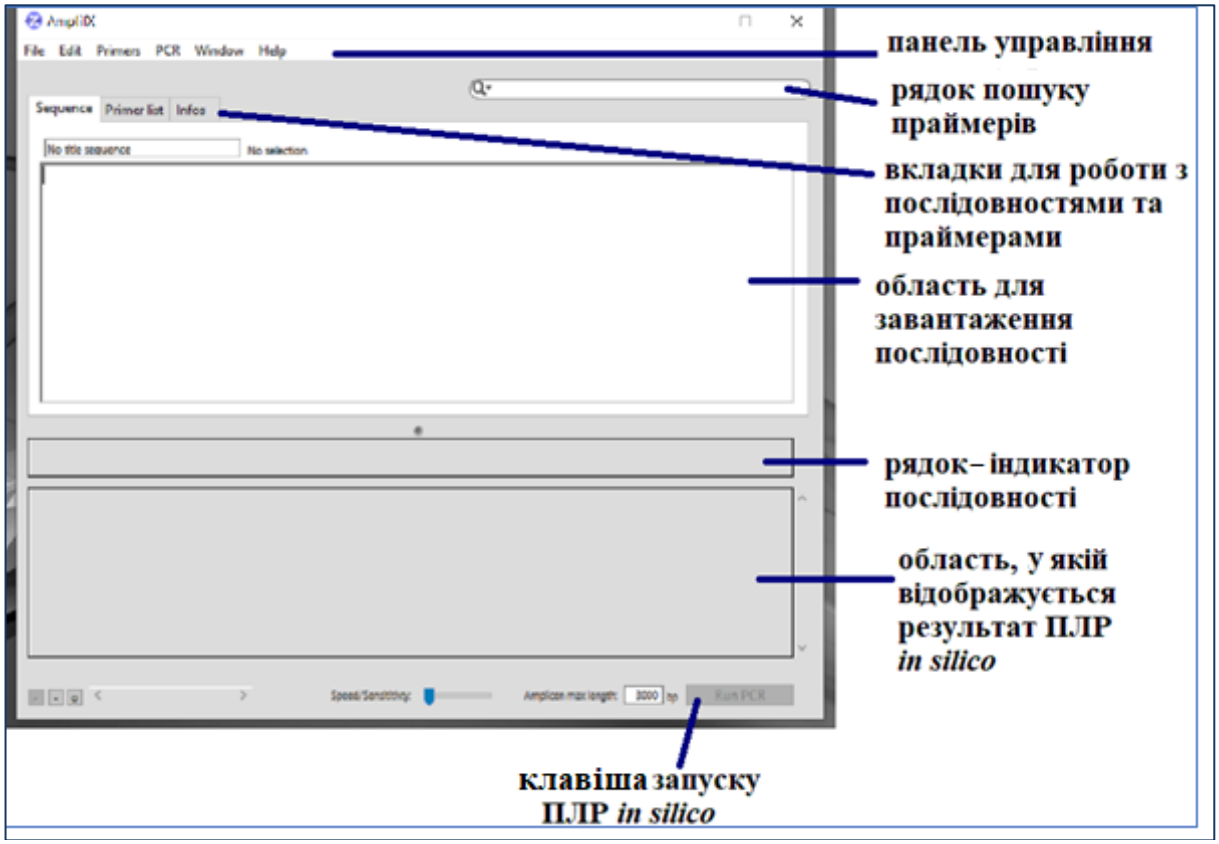


Рис. 58. Головне вікно програми AmplifX

Меню **File** містить команди для створення, завантаження, збереження, експортування, друку списків праймерів.

У меню **Edit** розміщено інструменти для копіювання, вирізання, вставки з буфера обміну послідовностей праймерів і ділянок генів, пошуку файлів з послідовностями та списками праймерів. Опція ***Preferences*** у межах цього меню відкриває діалогове вікно, у якому можна задати параметри за замовчуванням (температура плавлення (Tm), кількість димерів, рівень GC%, допустимий показник виродження та ін.). Результатом унесених змін є селективне завантаження зі списку тільки тих праймерів, які задовольняють установлені вимоги.

Меню **Primers** призначене для маніпуляцій зі списками праймерів. У ньому є інструменти для імпортування списків праймерів, виділення окремих або всіх праймерів у списку, додавання/вилучення праймерів, експорту виділених праймерів у задане місце. Функцію ***Design primers*** у цьому меню використовують для конструювання нових праймерів для цільової послідовності.

Меню **PCR** використовують для проведення симуляції ПЛР *in silico* із застосуванням розроблених для цільової послідовності праймерів, а також для тестування праймерів на утворення димерів.

Меню **Help** відкриває доступ до інформації про програму AmplifX, супровідної документації та інструкції.

Інтерфейс та призначення робочих панелей програми AmplifX. Головне вікно AmplifX розділено на дві частини (див. рис. 58). У верхній частині міститься текстова зона, організована у вигляді трьох панелей: «Sequence», «Primer List» та «Infos». У нижній частині відображується графічна схема розташування праймерів на цільовій послідовності після симуляції ПЛР *in silico*.

Панель ***Sequence*** призначена для завантаження цільової послідовності ДНК, для якої потрібно сконструювати нові праймери або підібрати зі списку розроблених раніше. AmplifX сприймає лише символи A, T, C, G. Усі інші символи та позначки, включаючи вироджені нуклеотиди (IUPAC: N, S, W та ін.), розриви і пропуски, під час вставки усуваються. З одного боку, це дозволяє копіювати послідовності безпосередньо з певного джерела (наприклад, геномної бази даних). З другого – це може призвести до некоректного розміщення праймерів на послідовності під час симуляції ПЛР *in silico,* а також до помилок у розрахунках довжини ампліконів. Усунути цей недолік можна шляхом використання послідовностей, які не містять зазначених елементів.

AmplifX підтримує файли форматів «DNA Strider», «GeneBank», «EMBL», «GCG», «Fasta», а також текстовий формат (*.txt), з яких можна завантажувати послідовності. Також зручно переносити послідовності, відкриті за допомогою BioEdit (див. рис. 26). Для цього потрібно в BioEdit імпортувати цільову послідовність, відкрити її для редагування, скопіювати у буфер обміну (Ctrl+C) та вставити у зону *Sequence* в AmplifX (Ctrl+V) (рис. 59).

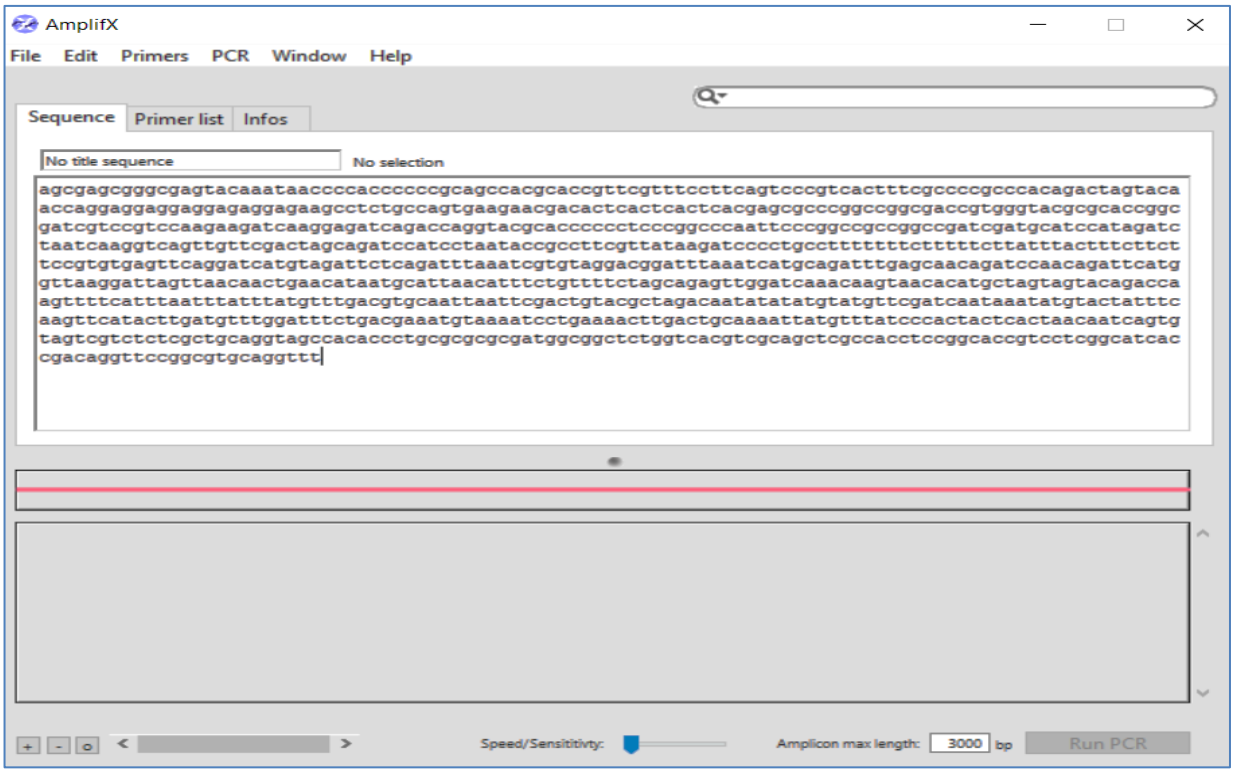


Рис. 59. Панель Sequence із завантаженою послідовністю нуклеотидів гена *Wx* пшениці виду *Tr. aestivum*

Вкладка ***Primer List*** містить список праймерів і пов'язаної з ними інформації (послідовність праймера, ID-номер у каталозі, назва або ідентифікатор, довжина, показник якості, TM) (рис. 60).

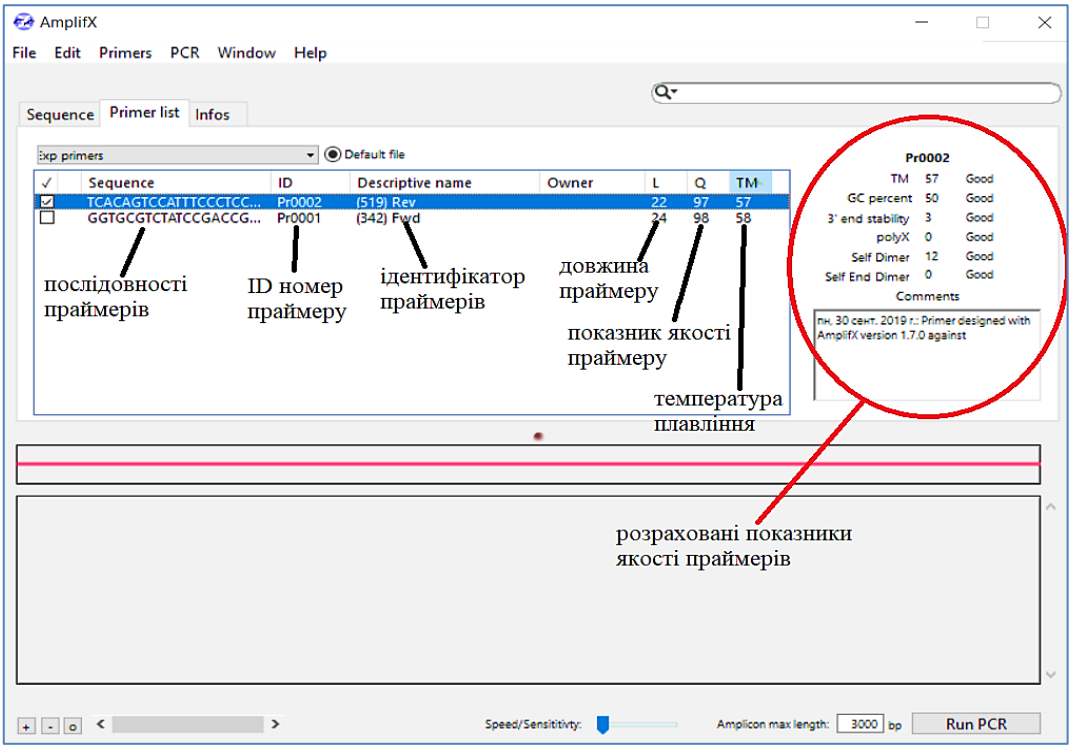


Рис. 60. Вкладка Primer List із завантаженим списком праймерів

Праймери до списку можна завантажувати із файлів, створених в Excel або FileMaker. Також можна ввести послідовності праймерів власноруч. Для цього у полі Primer List потрібно натиснути праву клавішу миші та обрати меню **Add primer** (рис. 61). У списку з'явиться порожній рядок для введення послідовності праймера з клавіатури ПК.

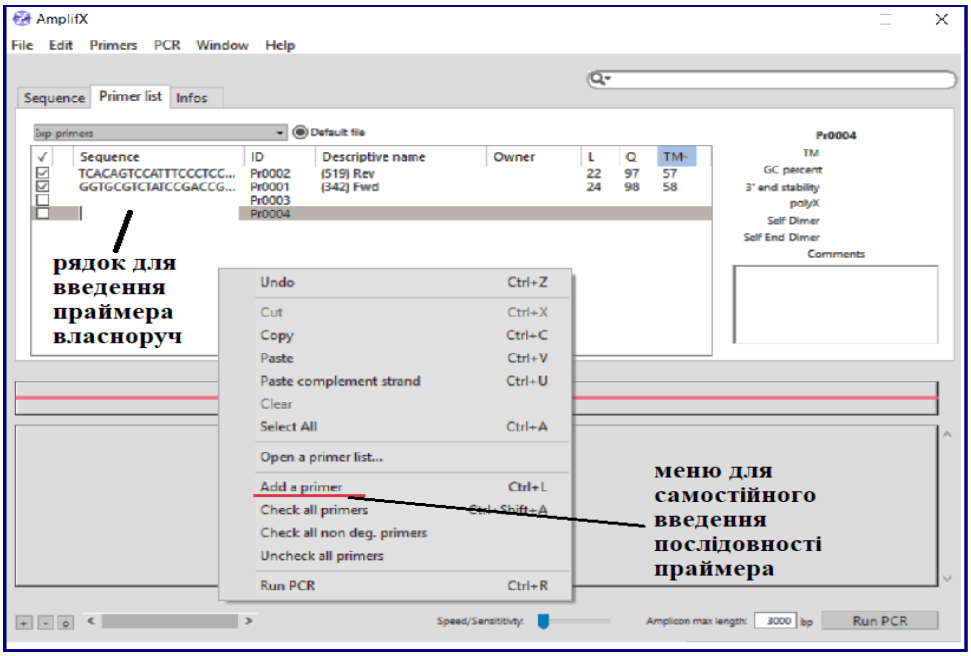


Рис. 61. Самостійне введення олігонуклеотидної послідовності до списку праймерів у програмі AmplifX

Панель ***Infos*** надає інформацію про будь-який графічний елемент, відображений у нижній частині вікна AmplifX. Щоб отримати таку інформацію, достатньо клікнути курсором по графічному елементу.

Дизайн праймерів у програмі AmplifX. Функція дизайну праймерів у AmplifX доступна в меню Primers на панелі управління головного вікна програми. Після активації функції ***Design primers*** з'являється вікно, розділене на дві частини (рис. 62). У верхній потрібно вказати параметри, яким повинні відповідати праймери. Зокрема, можна задати довжину, максимальну різницю температури плавлення прямого та зворотного праймерів, мінімальний рівень якості, діапазон місця розташування прямого і зворотного праймерів, діапазон довжини ампліконів, кількість пар праймерів, яку сконструює і відобразить програма). У нижній частині вікна міститься поле, у якому відображатимуться розроблені праймери після активації процесу дизайну.

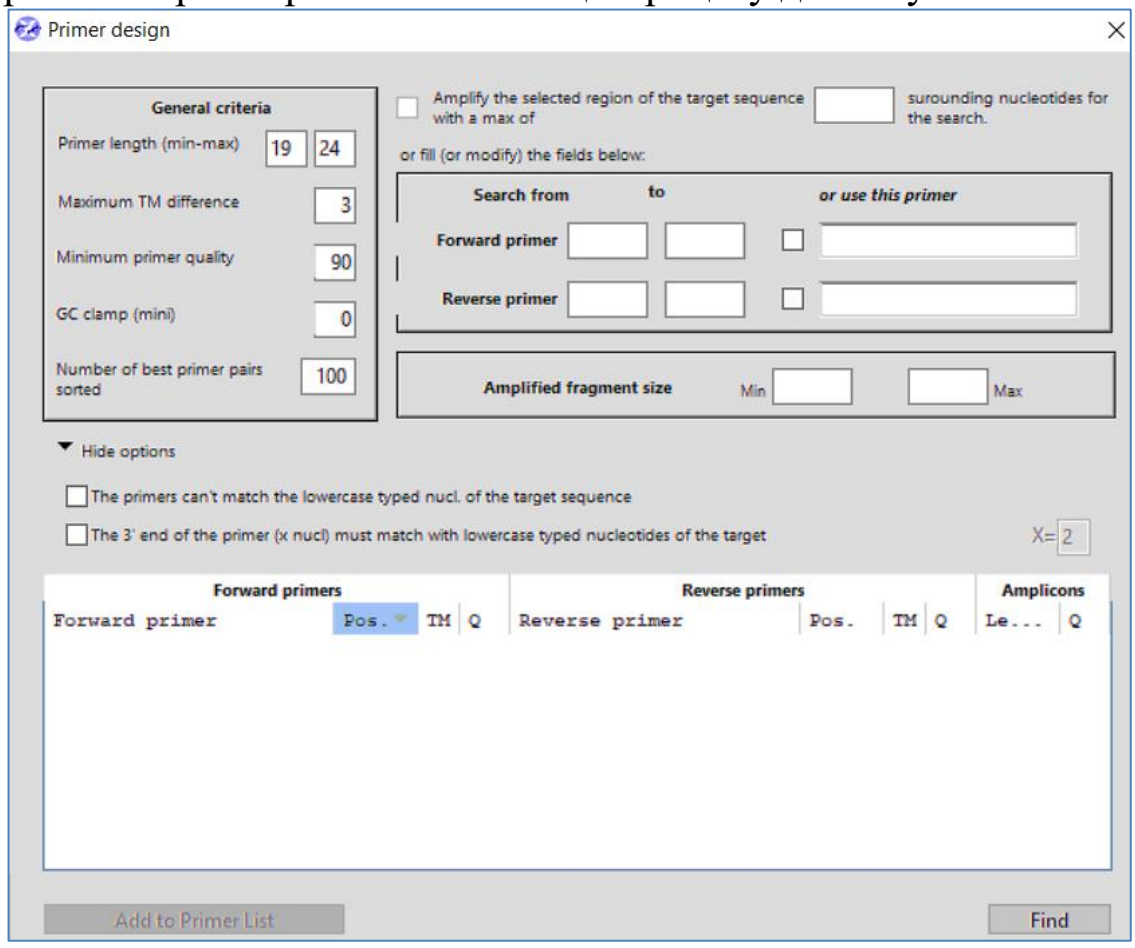


Рис. 62. Інтерфейс вікна дизайну праймерів у програмі AmplifX

Розглянемо процес дизайну праймерів на прикладі послідовності гена *Wx* пшениці м'якої (*Tr. aestivum*; GenBank:

JF682687.1). Пошук послідовності здійснювали в базі даних NCBI. До AmplifX послідовність переносили із програми BioEdit (див. рис. 59). Загальна довжина послідовності становить 897 п.н.

Параметри довжини праймерів (19–24 п.н.), різниці температури плавлення (3 °C), рівень якості (90 %) та кількість пар праймерів (100) введено за замовчуванням. Вони задовольняють вимоги більшості досліджень, тому їх можна не змінювати.

Діапазон місця розташування прямого і зворотного праймерів залежить від довжини досліджуваної послідовності, локалізації рамки зчитування (екзони, CDS), мети дослідження (розробка діагностичних праймерів, вивчення експресії генів тощо) та бажаної довжини ампліконів.

У наведеному нижче прикладі розроблятимемо діагностичні праймери до вказаної нуклеотидної послідовності для проведення класичної ПЛР. Нас задовольнить довільне розташування праймерів незалежно від локалізації CDS-ділянок.

Щодо розміру ампліконів зазначимо, що емпірично нами визначено найбільш оптимальні для детекції продукти розміром 200–500 п.н. (саме ці межі вводитимемо у програму). Зменшення нижньої межі може призвести до збільшення кількості димерів. Збільшення прогнозованого розміру продуктів подовжить час ампліфікації, оскільки подовжиться тривалість етапу елонгації.

Для визначення діапазонів, у яких мають приєднуватися прямий F (forward) і зворотний R (reverse) праймери, ділимо навпіл загальну довжину послідовності (897÷2≈449 п.н.). Від початку та з кінця послідовності відступаємо по 20 п.н., оскільки в цих місцях висока імовірність виникнення помилок під час секвенування. Отримуємо діапазон 21–449 п.н. для прямого праймера і 449–877 п.н. для зворотного праймера.

Після визначення параметрів і занесення їх до програми потрібно натиснути опцію **Find**. У нижній частині вікна дизайну праймерів з’являється ряд праймерів, що відповідають заданим нами параметрам (рис. 63).

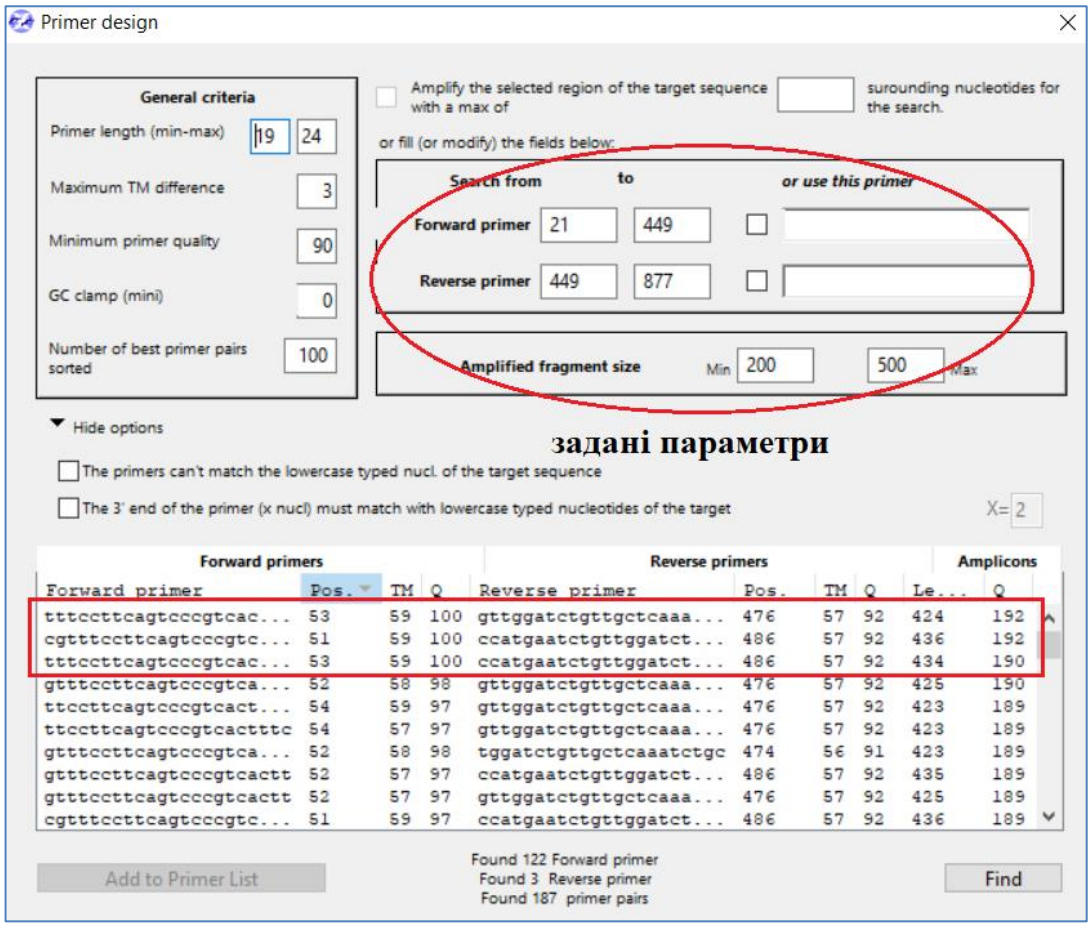


Рис. 63. Розроблені праймери в програмі AmplifX

Серед списку для подальшої роботи обираємо декілька пар праймерів з максимально високою якістю (Q. 97...100). У нашому випадку – це перші три пари (табл. 11).

11. Праймери до гена *Wx* пшениці виду *Tr. aestivum*, розроблені в AmplifX

Прямий праймер	Зворотний праймер
tttccttcagtcccgtcactttcg	gttggatctgttgctcaaatctgc
cgtttccttcagtcccgtcacttt	ccatgaatctgttggatctgttgc
tttccttcagtcccgtcactttcg	ccatgaatctgttggatctgttgc

Щоб занести розроблені праймери до списку AmplifX та протестувати їх, необхідно виділити пару праймерів, потім натиснути опцію **Add to Primer List** у лівому нижньому куті вікна Primers design. Обрані олігонуклеотиди завантажуються на панель Primer List (рис. 64).

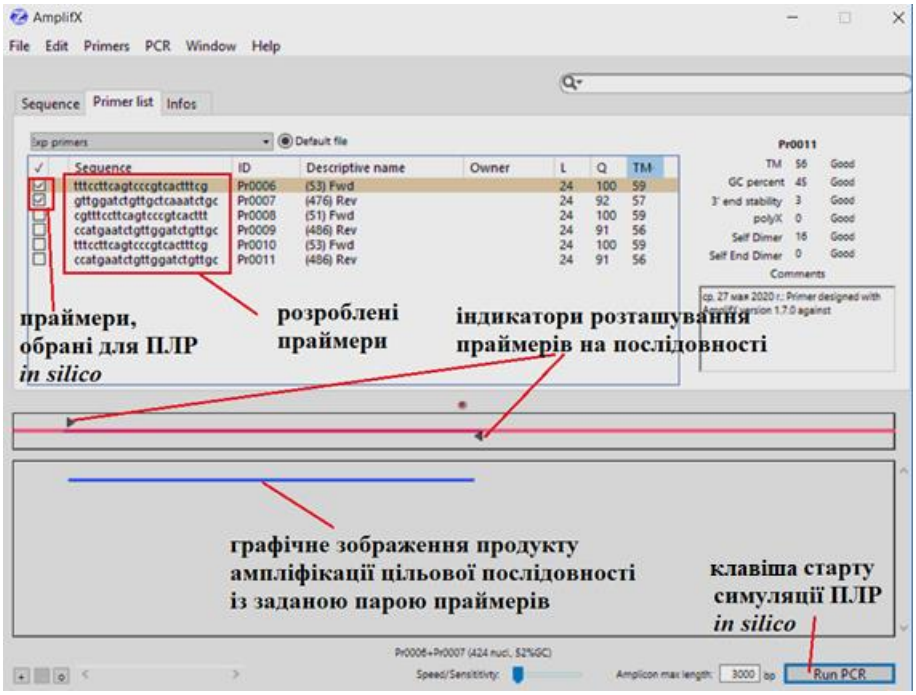


Рис. 64. Панель Primer List із завантаженими праймерами та проведеною симуляцією ПЛР

Для оцінки прогнозу роботи праймерів *in silico* потрібно вибрати пару F-R зі списку та натиснути клавішу ***Run PCR***. У нижній частині вікна з'являться індикатори розташування прямого та зворотного праймерів і графічне зображення амплікону. Можна одночасно тестувати декілька пар праймерів. При цьому в нижній частині вікна графічно відображатимуться продукти ампліфікації з кожною з вибраних пар праймерів, що дозволяє порівняти локалізацію детектованих ділянок у межах однієї послідовності й ефективніше добирати пари праймерів для подальших досліджень. У нашому прикладі усі три пари F-R приєднувалися майже в одному і тому ж місці.

Якщо натиснути на графічне зображення амплікону, автоматично відкривається панель ***Infos*** з інформацією про загальний розмір продукту ампліфікації, відсоток нуклеотидів GC у складі праймерів, температуру плавлення; графічно зображено ділянки приєднання та порядкові номери нуклеотидів, до яких приєднуються праймери, а також димери, якщо вони утворюються (рис. 65).

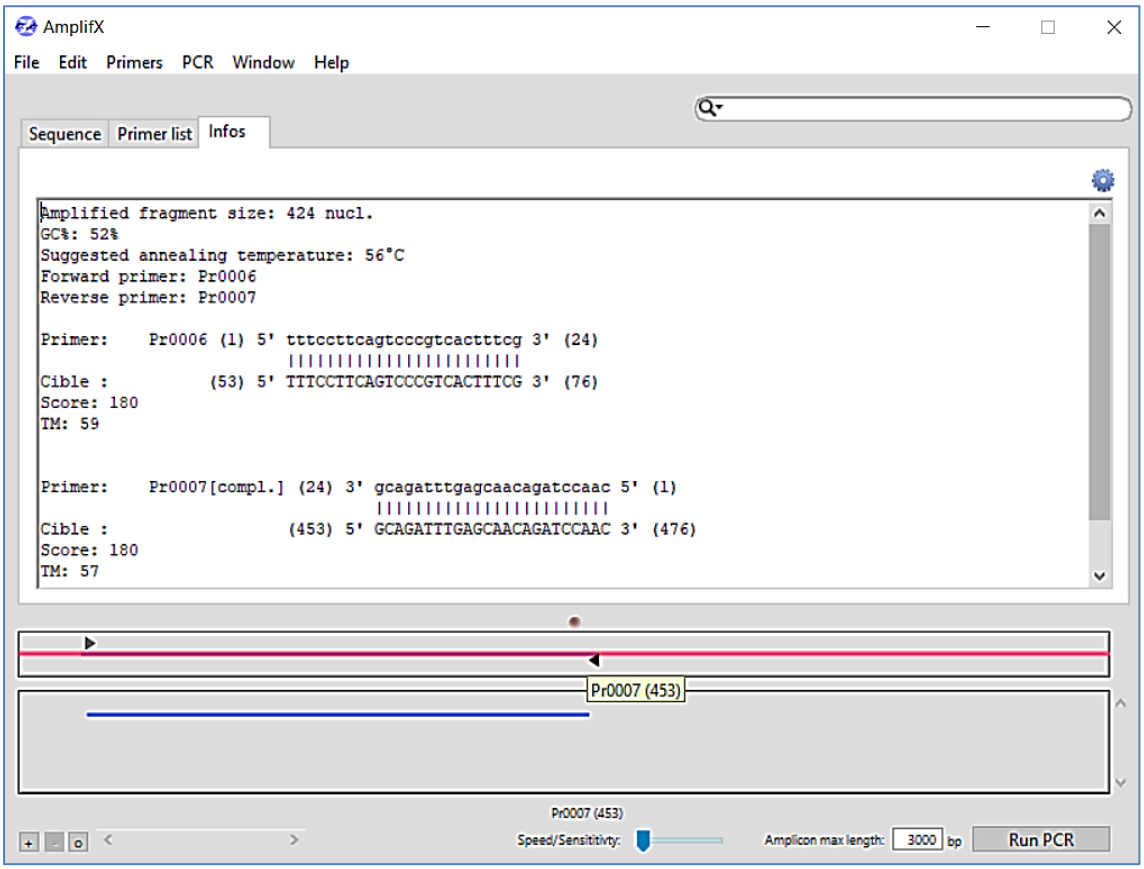


Рис. 65. Панель Infos з інформацією про праймери та амплікони, одержані за результатами ПЛР *in silico*

На рис. 65 представлено результати розрахунків за першою парою праймерів. Forward (вміст GC – 50 %) приєднуються до ДНК-матриці на ділянці з 53-го до 76-го нуклеотиду, reverse (вміст GC – 45 %) – з 453-го до 476-го нуклеотиду. Загальний розмір продукту, який ампліфікуватиметься, – 424 п.н. (вміст GC – 52 %). Через неоднаковий вміст GC у складі прямого та зворотного праймерів вони мають неоднакову температуру плавлення: F – 59 °C, R – 57 °C. Рекомендована AmplifX температура відпалу праймерів становить 56 °C.

Вибір праймерів серед списку розроблених програмою AmplifX є досить суб'єктивним. Обрані пари F-R потребують валідації. Також для кожної пари праймерів варто визначити параметри температури та часу для проведення ПЛР.

Кожний цикл ПЛР включає три основні етапи:

I етап – первинна денатурація: 1–5 хв за температури 92–95 °C;

II етап – 10–40 циклів, протягом яких відбувається безпосередньо накопичення продуктів ампліфікації. Кількість

циклів зумовлена початковою концентрацією ДНК в реакційній суміші. Кожний цикл складається з трьох стадій:

- ***денатурація ДНК*** – відбувається при температурі 94–95 °С за період від 5 с до 1 хв, залежно від передбачуваної довжини амплікону, типу використовуваних пробірок, буфера, типу Taq-полімерази;

- ***гібридизація праймерів***: температура відпалу варіює в межах 40–72 °С, підбирається індивідуально, залежно від довжини, нуклеотидного складу, *Tm* окремих праймерів. Температура відпалу має бути на 2–5 °С нижчою за *Tm*. Тривалість етапу – від 5 с до 2 хв;

- ***елонгація***: температурний режим залежить від типу Taq-полімерази і найчастіше становить 72 °С. Тривалість етапу варіює залежно від Taq-полімерази та передбачуваної довжини ампліконів. У середньому швидкість синтезу компліментарного ланцюга дорівнює 1–2 тис. основ за 1 хв, або 30–60 нуклеотидів за 1 с;

ІІІ етап – фінальна елонгація – не є обов'язковою стадією, використовується для збільшення виходу реакції, для остаточної дуплікації одноланцюгових фрагментів.

Температуру плавлення праймерів зручно розраховувати за допомогою **олігокалькуляторів** – мініпрограм, призначених для розрахунку фізичних (довжина, молекулярна маса, вміст GC%, концентрація мкг/мл), термодинамічних і температурних параметрів для заданої олігонуклеотидної послідовності. Для здійснення розрахунків у таких програмах достатньо знати лише послідовність нуклеотидів, яку потрібно ввести в програму і натиснути «Порахувати». Ми використовували олігокалькулятор за посиланням http://www.basic.northwestern.edu/biotools/OligoCalc.html (рис. 66). Цей калькулятор розраховує *Tm* трьома методами: простим, з коригуванням за концентрацією солей і за алгоритмом «найближчого сусіда».

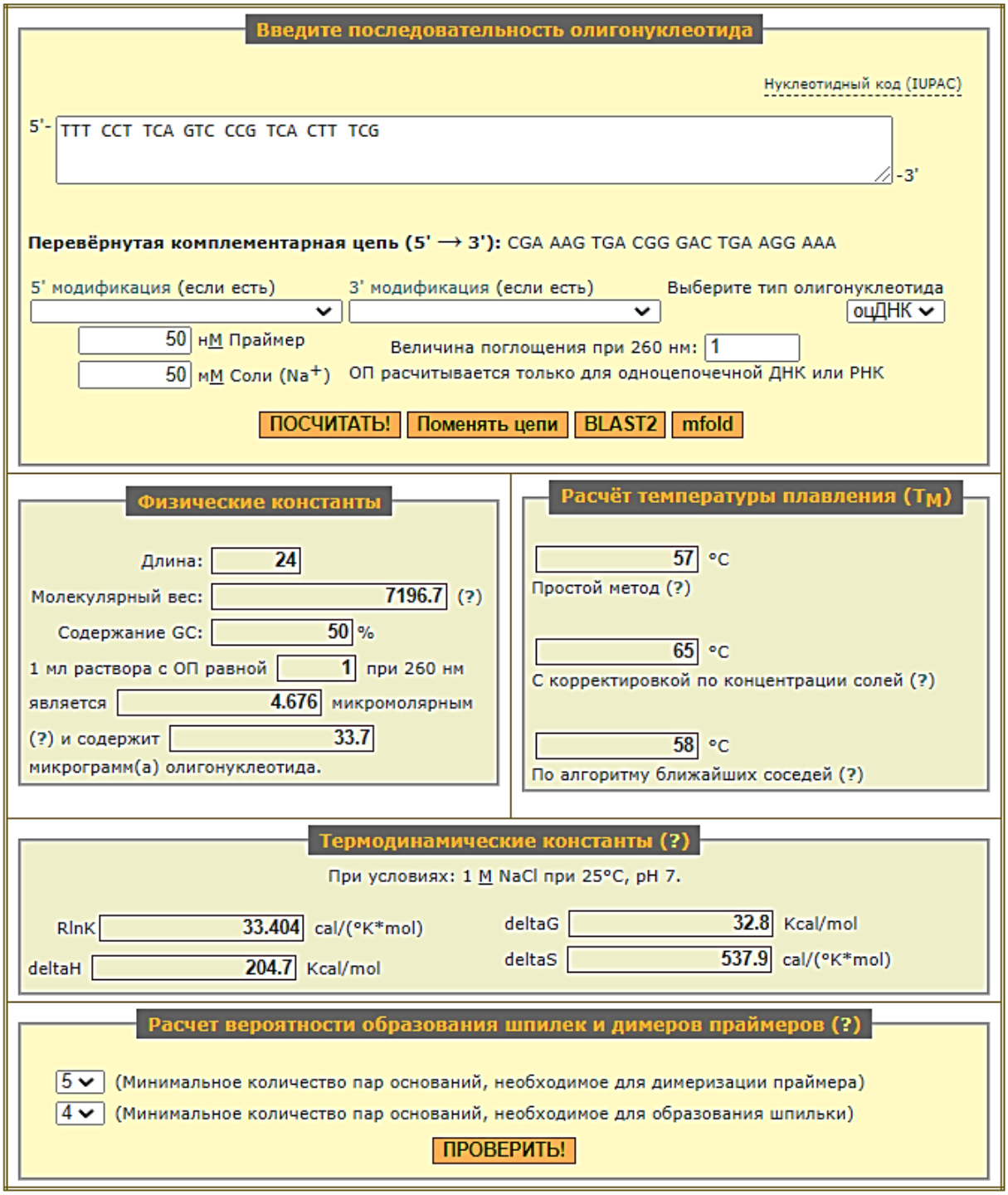


Рис. 66. Інтерфейс олігокалькулятора із сайта БГУ (http://www.basic.northwestern.edu/biotools/OligoCalc.html) з розрахованими параметрами для праймера tttccttcagtcccgtcactttcg (F)

Простий метод, найчастіше використовуваний науковцями, є досить неточним. Обчислення *Tm* виконують за формулами:

1) для послідовностей менше 14 нуклеотидів:

$$Tm = (wA + xT) \cdot 2 + (yG + zC) \cdot 4$$

де *w*, *x*, *y*, *z* – кількість відповідних нуклеотидів (A, T, G, C);

2) для послідовностей більше 13 нуклеотидів:

$$Tm = 64{,}9 + 41 \cdot (yG + zC - 16{,}4)/(wA + xT + yG + zC).$$

Метод з коригуванням за концентрацією солей ураховує поправку на концентрацію іонів натрію, які суттєво впливають на формування сольових містків між ланцюгами ДНК. *Tm* розраховують за формулами:

1) для послідовностей менше 14 нуклеотидів:

$$Tm = (wA + xT) \cdot 2 + (yG + zC) \cdot 4 - 16{,}6 \cdot \ln(0{,}050) + 16{,}6 \cdot \ln([Na+]);$$

2) для послідовностей більше 13 нуклеотидів:

$$Tm = 100{,}5 + (41 \cdot (yG + zC)/(wA + xT + yG + zC)) - (820/(wA + xT + yG + zC)) + 16{,}6 \cdot \ln([Na+]).$$

Це рівняння доцільно використовувати для послідовностей довжиною 18–25 нуклеотидів.

Метод розрахунку Tm за **алгоритмом «найближчого сусіда»** є ефективним для праймерів довжиною 8–40 нуклеотидів, він оснований на термодинамічній залежності між ентропією, ентальпією, вільною енергією і температурою. Розраховують за формулою:

$$T = \frac{\Delta H - 3{,}4 \frac{kcal}{^\circ K\ mole}}{\Delta S + R ln\left(\frac{1}{[primer]}\right)} + 16{,}6 ln([Na^+]),$$

де ΔH – ентальпія, ΔS – ентропія, R – універсальна газова стала (1,987 кал/моль К), [$primer$] – концентрація незв'язаних праймерів, 3,4 – кількість ккал, на яку змінюється енергія під час переходу від одноніткової до В-форми ДНК, $^\circ K$ – температура у кельвінах.

Визначимо умови ПЛР для першої пари праймерів (див. табл. 11), розроблених нами у програмі AmplifX (**F**: tttccttcagtcccgtcactttcg; **R**: gttggatctgttgctcaaatctgc): температура відпалу – 56 °С (розраховано в AmplifX); очікувана довжина амплікону – 424 п.н.; отже, елонгація має тривати приблизно 30 с.

Розрахункові умови для ПЛР із цими праймерами такі:

1 цикл – 95 °С – 5 хв;

30 циклів: 95 °С – 30 с, 56 °С – 30 с, 72 °С – 30 с;

фінальна елонгація – 72 °С – 5 хв.

Розраховані умови ПЛР на практиці не завжди бувають досконалими, а отже, потребують попереднього відпрацювання та іноді внесення змін до протоколу.

Визначення специфічності праймерів *in silico* в базі даних GenBank. Для порівняння досліджуваного праймера з усіма послідовностями ДНК, що зберігаються в базах даних, використовують програму blastn (порівняння нуклеотид-нуклеотид). Для пошуку гомології праймерів необхідно встановити такі параметри:

1) розмір слова (word size) – 7. В програмі BLAST є два алгоритми – megablast та blastn. Megablast – нуклеотидну послідовність буде розбито на короткі субпослідовності, які називаються «розмір слова». Тому високі значення при здійсненні пошуку ідентичності праймера можуть перешкодити його вирівнюванню з випадкової незбіжності;

2) значення припущення (expect value) – 1000. Цей параметр дозволяє встановити очікувану кількість збігань гомологічних послідовностей. Для перевірки специфічності праймерів починати аналіз необхідно зі значення припущення «е»=1000. При аналізі буде спостерігатися така закономірність: чим більша величина параметру припущення, тим більший список збіжностей з невеликою кількістю очок буде отримано. Наприклад, значення припущення, яке буде дорівнювати п’яти, означає, що у базі даних очікується лише п’ять випадкових збіжностей.

3) штраф за відкриття гепу (gap opening penalty) – 12 та штраф за поводження гепу (gap extension penalty) – 8. Під час пошуку послідовностей, гомологічних досліджуваним праймерам, вставка гепів є небажаною. Тому параметри штрафу за відкриття гепів мають бути високими або дуже високими;

4) фільтрування складності (complexity filtering) виключити. Виключення цього параметра гарантуватиме збіжність праймера незалежно від складу його нуклеотидів.

Пошук здійснюють шляхом уведення послідовності досліджуваного праймера у вікно пошуку. Якщо запит менший за 30 п.н., то програма blastn автоматично перевіряє, чи є запит коротким, та налаштовує параметри пошуку. Під час перевірки специфічності праймерів бажано задати пошук відразу обох з них. Для цього між послідовностями праймерів вставляють ланцюг з 20

або більше літер N, після чого проводять пошук. У такому запиті немає необхідності робити комплементарний реверс зворотного праймера перед їх з'єднанням, оскільки програма автоматично здійснює пошук в обох напрямах. Якщо проводиться одночасний запит праймерів, з'єднаних літерами N, автоматичне прилаштування короткого запиту не відбувається, тому встановлення параметрів пошуку має бути здійснене вручну.

Під час оцінювання результатів слід звертати увагу на такі моменти:

1) послідовність нуклеотидів геномної ДНК, з якою відбувається гібридизація праймерів, може містити ортологи або паралоги, через це може не синтезуватися продукт ПЛР необхідного розміру;

2) гібридизація праймерів із цільовою ДНК може відбутися в кількох місцях;

3) можлива незбіжність праймера на 3'-кінці, що, ймовірно, спричинить відсутність формування комплексу праймер-ДНК;

4) не ідентифіковано жодного збігу послідовності праймера із ДНК, через те потрібно змінити параметри пошуку, а праймер може бути занадто коротким.

Контрольні запитання

1. Які завдання з біоінформатики можна вирішити за допомогою програми BioEdit?
2. Опишіть структуру програми BioEdit.
3. Які завдання з біоінформатики можна вирішити за допомогою програми MEGA?
4. За допомогою яких програм здійснюють дизайн праймерів?
5. Як розрахувати оптимальну температуру гібридизації праймерів до матричної ДНК?
6. Як визначити специфічність праймерів *in silico*?
7. Який алгоритм дизайну праймерів у програмі AmplifX?
8. Яке призначення олігокалькуляторів?
9. У чому особливості програми Structure?

РЕКОМЕНДОВАНІ ДЖЕРЕЛА

Основні

1. Леск А. Введение в биоинформатику/ А. Леск. – Москва: Бином, 2009. – 318 с.

2. Игнасимуту С. Основы биоинформатики / С. Игнасимуту. – Москва–Ижевск: НИЦ «Регулярная и хаотическая динамика», 2007. – 320 с.

3. Лукашов В.В. Молекулярная эволюция и филогенетический анализ/ В.В. Лукашов. – Москва: Бином, 2009. – 256 с.

4. Вейр Б. Анализ генетических данных / Б. Вейр. – Москва: Мир, 1998. – 400 с.

5. Акинина Г.Е. Статистичекий анализ данных с использованием компьютерных программ Arlequin, Philyp, Clann, Structure / Г.Е. Акинина, Ю.Н. Дугарь, В.Н. Попов. – Харьков, 2014. – 100 с.

6. Basu C. Methods in Molecular Biology. PCR Primer Design, 2 / C. Basu // Aufl. Humana Press, Springer. – New York, 2015. – V. 1275. – 221 p.

Допоміжні

1. Alzohairy A.M. BioEdit: An important software for molecular biology / A.M. Alzohairy // GERF Bulletin of Biosciences. – 2011. – V.2, I. 1. – P. 60–61.

2. Hall T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT / T.A. Hall // Nucl. Acids. Symp. Ser. – 1999. – V. 41. – P. 95–98.

3. Thompson J.D. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice / J.D. Thompson, D.G. Higgins, T.J. Gibson // Nucleic Acids Research. – 1994. – V. 22, I. 22. – P. 4673–4680.

4. Jullien N. AmplifX 1.7.0. [Electronic resource] / N. Jullien. – 2013. – Access mode: http://crn2m.univ-mrs.fr/pub/amplifx -dist.

5. Kibbe W.A. OligoCalc: An Online Oligonucleotide Properties Calculator / W.A. Kibbe // Nucleic Acids Res. 35(Web Server issue). – 2007. – W43-6. – DOI: 10.1093/nar/gkm234.

6. Kumar S. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets / S. Kumar, G. Stecher, K. Tamura // Mol. Biol. Evol. – 2016. – V. 33, I. 7. – P. 1870–1874. – DOI: 10.1093/molbev/msw054.

Електронні ресурси

https://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional – текстові та відеоінструкції користування Uniprot.

https://www.uniprot.org/help/ – розділ «Допомога» на сайті Uniprot.

ДОДАТКИ

Додаток А

Вихідні дані

Тип даних	**Представлення вихідних даних**		**Рекомендована міра подібності**
Морфологічні	Бінарна матриця складається за принципом «відсутність-наявність» певної градації ознаки. Частка об’єктів з певною градацією. Абсолютні значення для кількісних ознак		Критерій Jaccard
Біохімічні (запасні білки, ізоферменти)	Бінарна матриця. Частота алелів гена ферменту		Стандартна генетична відстань Nei (*D*), Nei & Li
Молекулярні	SSR (залежно від еволюційної моделі)	1) для **моделі нескінченної кількості алелів – IAM (infinite alleles model)** – матриця частот алелів	Генетична відстань Nei; Reynolds, Weir and Cockerham's,
		2) для **покрокової мутаційної моделі – SMM (stepwise mutation model)**; 3) для **двофазної моделі – TPM (two phase model)**, **узагальненої покрокової моделі – GSM (generalized stepwise model)** – матриця, яка містить розміри алелів, число нуклеотидних повторів у них або нуклеотидні послідовності ДНК	Генетична відстань Slatkin, Reynolds
	RAPD, ISSR, AFLP	Бінарна матриця: «1» – наявність фрагменту ПЛР; «0» – відсутність фрагменту ПЛР	Генетична відстань Nei & Li

Додаток Б

Показники подібності і генетичні відстані

Показник	Формула	Значення символів	Опис
1	2	3	4
Показник подібності			
Ідентичність алельних генів у двох популяціях для *j*-локусу (I_j)	$I_j = \sum x_i y_i / \sqrt{\sum x_i^2 \sum y_i^2}$	x_i та y_i – частоти алелів у популяціях X і Y	Генетична подібність може набувати значень від нуля (немає спільних алелів у порівнюваних популяціях) до одиниці (частота усіх алелів однакова в обох популяціях)
Генетична подібність (I) для сукупності локусів	$I = \bar{J}_{xy} / \sqrt{\bar{J}_x \bar{J}_y}$	$\bar{J}_x, \bar{J}_y, \bar{J}_{xy}$ – арифметичні середні за всіма локусами; m – кількість алелів; L – кількість локусів	
	$\bar{J}_x = \frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{m} x_i^2$		
	$\bar{J}_y = \frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{m} y_i^2$		
	$\bar{J}_{xy} = \frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{m} x_i y_i$		
Генетична відстань			
Евклідова відстань для двох та більше змінних (d_{PQ})	$\mathrm{d_{PQ}} = \sqrt{\sum_1^k x_{il} - x_{jl}}^{\,2}$	d_{PQ} – відстань між об'єктами; x_{il}, x_{jl} — значення l-ї ознаки у i-го та j-го об'єкта	Евклідову відстань визначають за допомогою координат двох точок у n-мірному просторі

Продовження дод. Б

1	2	3	4
Махаланобіса (d_{ij})	$d_{ij} = \left(X_i - X_j\right)^T S^{-1}(X_i - X_j)$	X_i та X_j – значення змінних для *i*- та *j*-об'єктів; S – загальна внутрішньогрупова дисперсійно-коваріаційна матриця	Відстань Махаланобіса, яку розраховують за допомогою матриць дисперсій-коваріацій, пов'язана з кореляціями між змінними. Якщо кореляція між змінними дорівнює нулю, ця метрика еквівалентна квадратичній евклідовій відстані
Nei (D)	$D = -lnI$	I – генетична подібність	Генетична відстань Nei варіює від нуля до нескінченності. Її використовують як міру генетичної диференціації популяції. Використання цієї відстані передбачає таке: а) IAM модель еволюції; б) усі локуси мають однакову швидкість мутацій; в) мутації та дрейф генів перебувають у рівновазі; г) стабільний ефективний розмір популяції
Хордова відстань Cavalli-Sforza	$cos\ \theta = \sum_{i=1}^{m} \sqrt{p_i q_i}$ $D_{CH} = 2 \cdot \sqrt{\frac{2}{\pi}} \cdot \sqrt{1 - \cos\theta}$	p_i та q_i – частоти алелів або фенотипів	Ця відстань ураховує тільки дрейф генів. Показник використовують для аналізу не тільки частот алелів, але й частот фенотипів

1	2	3	4
Nei & Li	$D_{ij} = 1 - S_{ij} = \frac{2a}{2a + b + c}$	a – кількість загальних градацій ознаки у двох порівнюваних об'єктах (тип 1,1); b – кількість однієї градації ознаки, яка властива тільки першому об'єкту (тип 1,0); c – кількість однієї градації ознаки, яка властива тільки другому об'єкту (тип 0,1); d – кількість відсутності градацій ознаки в обох об'єктах(тип 0,0)	Арифметичні доповнення до коефіцієнта подібності. Використовують для обробки бінарних даних (RAPD, ISSR, фенотипові бінарні ознаки)
Gower	$D_{ij} = 1 - S_{ij} = 1 - \frac{a}{a + b + c}$		
Sneath & Sokal	$D_{ij} = 1 - S_{ij} = 1 - \frac{a + d}{a + b + c + d}$		
Відстань Reynolds, Weir, and Cockerham's	$\theta_w = \sqrt{\frac{\sum_l \sum_u (X_u - Y_u)^2}{2\sum_l (1 - \sum_u X_u Y_u)}}$	X_u, Y_u – частоти алеля u у локусі l у популяціях X та Y	Використання цієї відстані передбачає таке: а) IAM модель еволюції; б) розроблено для алозимних даних та передбачено, що тільки генетичний дрейф впливає на частоту алелів (не враховує мутації)
Slatkin	$S_w = \frac{1}{d_s} \sum_{j=1}^{d_n} \frac{2}{2n(2n-1)} \sum_{i \le i'} (a_{ij} - a_{i'j})^2$	d_s – кількість субпопуляцій j; n – кількість об'єктів у кожній субпопуляції; a_{ij} – кількість тандемних повторів і-алеля у j-субпопуляції	Під час використання SMM, TPM, GSM мутаційних моделей

Продовження дод. Б

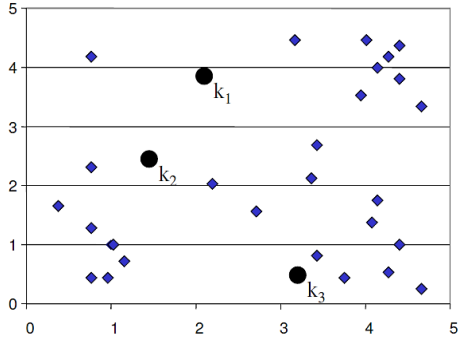
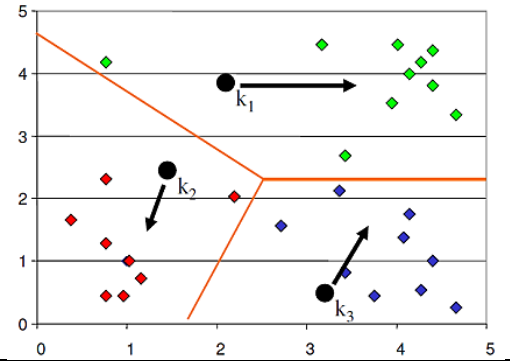
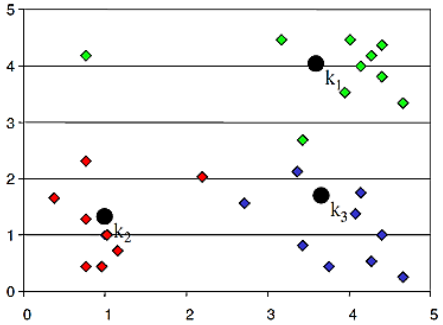
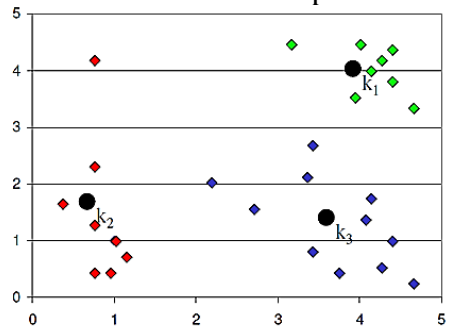
1	2	3	4
Jukes-Cantor	$d_{JC} = -\frac{3}{4} ln\left(1 - \frac{4p}{3}\right)$	p – парна дистанція	Метод ураховує частку нуклеотидів, що не збігається під час попарного порівнювання; базується на припущенні однакової частоти нуклеотидів (25 %) та однакової ймовірності заміщення в будь-якій парі нуклеотидів
Tajima-Nei	$d = -b \ln\left(1 - \frac{p}{b}\right),$ $b = \frac{1}{2}\left(1 - \sum_{i=1}^{4} g_i^2 + \frac{p^2}{c}\right),$ $c = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \frac{x_{ij}^2}{2 g_i g_j}$	x_{ij} – відносні частоти пар нуклеотидів, g_i *і* g_j – відносні частоти i-го і j-го нуклеотидів	Модель ураховує неоднакову частоту чотирьох нуклеотидів (A, T, G, C) у послідовностях, а також різну імовірність замін цих нуклеотидів
Kimura 2-parameter	$d_{K2P} = -\frac{1}{2} \ln(1 - 2P - Q) - \frac{1}{4} \ln(1 - 2Q)$	P – частка транзицій, Q – частка трансверсій	Метод передбачає, що поява транзицій і трансверсій має різну імовірність. Базується на припущенні, що частоти нуклеотидів дорівнюють 0,25 протягом усього еволюційного процесу
Tamura 3-parameter	$d_T = -2\Theta(1-\Theta)\ln\left(1 - \frac{P}{2\Theta(1-\Theta)} - Q\right) -$ $-\frac{1}{2}\left(1 - 2\Theta(1-\Theta)\right)\ln(1-2Q)$	P – частка транзицій, Q – частка трансверсій, Θ – вміст C+G	Враховує можливість різних частот нуклеотидів у послідовностях і оцінює максимальну імовірність, яка найбільше відповідає кожному випадку

Закінчення дод. Б

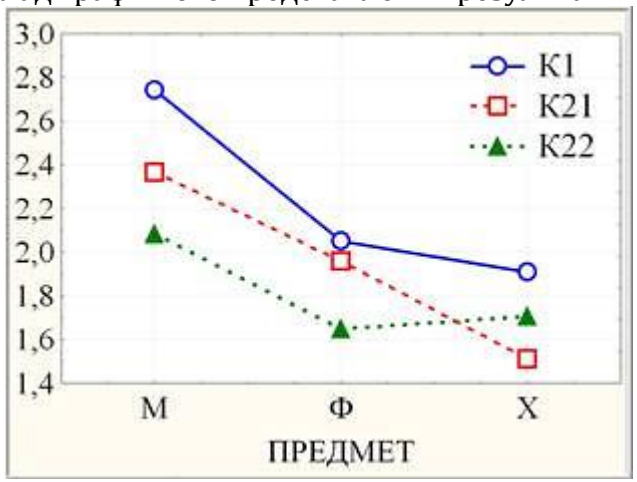
1	2	3	4
Tamura-Nei	$d_{TN} = \frac{2g_A g_G}{g_R} \ln\left(1 - \frac{g_R}{2g_A g_G} P_1 - \frac{1}{2g_R} Q\right) - \frac{2g_T g_C}{g_Y} \ln\left(1 - \frac{g_Y}{2g_T g_C} P_2 - \frac{1}{2g_Y} Q\right) - 2\left(g_R g_Y - \frac{g_A g_G g_Y}{g_R} - \frac{g_T g_C g_R}{g_Y}\right) ln\left(1 - \frac{1}{2g_R g_Y} Q\right)$	P_1 та P_2 – частка транзицій між A і G та між T і C, відповідно, Q – частка трансверсій, g_A, g_G, g_C, g_T, g_R, g_Y – відносні частоти відповідних нуклеотидів, R – вироджений нуклеотид A/G, Y – вироджений нуклеотид C/T	Модель ураховує, що рівень транзицій між пуринами (A і G) та піримідинами (T і C) часто відрізняється
MCL-метод	$d_{ij} = 4(g_A g_G k_1 + g_T g_C k_2 + g_R g_Y) \times b_{ij}$	g_A, g_G, g_C, g_T – частоти нуклеотидів A, G, C і T, $k_1 = (P_{1ij}/g_A g_T)/(Q_{ij}/ g_R g_Y)$, $k_2 = (P_{2ij}/g_T g_C)/(Q_{ij}/ g_R g_Y)$, $b_{ij} = Q_{ij} /(4g_R g_Y)$, де P_{1ij}, P_{2ij} і Q_{ij} – частоти транзицій A↔G, транзицій C↔T і трансверсій відповідно	Точність методу підвищується зі збільшенням кількості досліджуваних послідовностей

Додаток В

Ітеративні методи групування

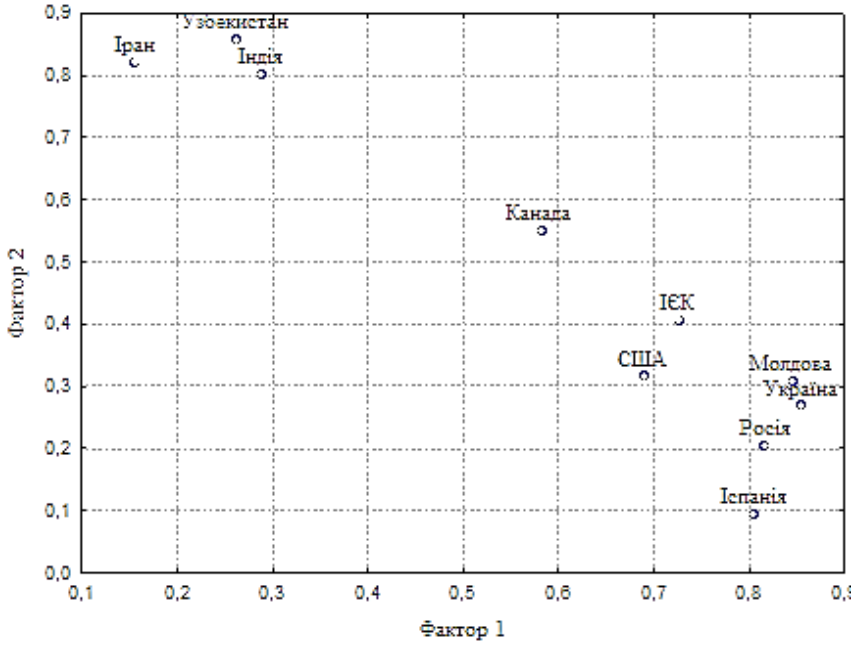
<table>
<tr><td>Методи</td><td colspan="2">Опис методу</td></tr>
<tr><td>1</td><td colspan="2">2</td></tr>
<tr><td rowspan="2">Метод К-середніх</td><td>1) дані розбивають на деяке задане число кластерів. Розраховують центри ваги всіх кластерів
 
</td><td>2) кожну точку даних переносять у кластер з найближчим центром ваги
 
</td></tr>
<tr><td>3) розраховують нові центри ваги
 
</td><td>4) кроки 2 та 3 повторюють до тих пір, поки не перестануть змінюватися кластери
 
</td></tr>
</table>

Продовження дод. В

1	2
	5) приклад графічного представлення результатів аналізу 
Байєсівський підхід (Bayesian Approach)	Класифікація за допомогою байєсівського підходу основана на оцінці коефіцієнта ймовірності відповідно до теореми Байєса. Цей підхід використовують для кластеризації даних у програмі Structure

Додаток Г

Факторні методи

Метод	Опис	Графічне представлення результатів аналізу
Факторний аналіз	Факторний аналіз передбачає виділення гіпотетичних факторів з великої кількості змінних. Він дозволяє отримати просту структуру, яка відображає певні закономірності у масиві даних. Кількість факторів, які виділяють, повинна бути менша від набору вихідних змінних. Факторний аналіз складається з чотирьох етапів: 1) розрахунок кореляційної матриці для всіх змінних; 2) вилучення факторів; 3) перетворення факторів з метою одержання простої структури; 4) інтерпретація факторів. Виділяють декілька видів факторного аналізу: 1) стандартний (R) – факторний аналіз змінних; 2) зворотній (Q) – факторний аналіз об'єктів; 3) часовий (T) – факторний аналіз часових тенденцій у різних об'єктів за однією змінною; 4) зворотній часовий (S) – факторний аналіз об'єктів залежно від часу; 5) одного об'єкта (P) – факторний аналіз змінних одного об'єкта у різний час; 6) часовий в одного об'єкта (O) – факторний аналіз часових тенденцій за різними показниками в одного об'єкта. Для статистичної обробки біологічних даних найчастіше використовують R- та Q-техніки факторного аналізу	 Приклад графічного представлення результатів з використанням Q-техніки факторного аналізу Розташування у двомірному просторі сортів нуту з різних країн, оцінених за поліморфізмом мікросателітних локусів

Навчальне видання

Попов Віталій Миколайович
Лиманська Світлана Василівна
Чернишенко Галина Євгенівна
Тереняк Юлія Миколаївна

ОСНОВИ БІОІНФОРМАТИКИ

Навчальний посібник

Редактор О.В. Васільєва
Коректор І.О. Бутильська
Комп'ютерний набір і верстка – В.М. Попов, С.В. Лиманська, Г.Є. Чернишенко, Ю.М. Тереняк

Підпис. до друку 17.12.2021 р. Формат 60х84/16. Гарнітура Таймс. Друк офсетний. Обсяг: 6,3 ум. друк. арк.; 6,3 обл.-вид. арк. Тираж 100. Замовлення № ___

Виробник – редакційно-видавничий відділ Харківського національного аграрного університету ім. В.В. Докучаєва.
62483, Харківська обл., Харківський р-н, п/в «Докучаєвське–2», навч. містечко ХНАУ, тел. 99-72-70. E-mail: office@knau.kharkov.ua