Міністерство розвитку економіки, торгівлі та сільського господарства України

УКРАЇНСЬКИЙ ІНСТИТУТ ЕКСПЕРТИЗИ СОРТІВ РОСЛИН

Стариченко Є. М., Присяжнюк Л. М., Гринів С. М., Лещук Н. В., Ткачик С. О., Києнко З. Б., Мельник С. І.

ЗАСТОСУВАННЯ ТЕСТУ МАНТЕЛЯ ДЛЯ ОЦІНКИ КОРЕЛЯЦІЙНИХ ЗВ'ЯЗКІВ МІЖ ДНК ТА МОРФОЛОГІЧНИМИ МАРКЕРАМИ СОРТІВ РОСЛИН

Науково-методичні рекомендації

Вінниця
ТОВ «ТВОРИ»
2020

УДК 311.16.577.21.32. 633.1

Стариченко Є. М., Присяжнюк Л. М., Гринів С. М., Лещук Н. В., Ткачик С. О., Києнко З. Б., Мельник С. І. Застосування тесту Мантеля для оцінки кореляційних зв'язків між ДНК та морфологічними маркерами сортів рослин. Вінниця: ТОВ «ТВОРИ», 2020. 27 с.

Рецензенти:

Мокрієв М. В., кандидат економічних наук, доцент кафедри інформаційних систем і технологій, Національний університет біоресурсів і природокористування України

Чернуський В. В., кандидат сільськогосподарських наук, провідний науковий співробітник відділу селекції цукрових буряків, Інститут біоенергетичних культур і цукрових буряків НААН

Науково-методичні рекомендації розроблено в Українському інституті експертизи сортів рослин. У науково-методичних рекомендаціях представлено результати досліджень, виконаних в Українському інституті експертизи сортів рослин. Авторами запропоновано застосування надбудови Microsoft Excel XLSTAT та мови програмування для статистичної обробки R для оцінки генетичних дистанцій між сортами рослин та визначення кореляційних зв'язків між матрицями дистанцій за SSR маркерами та морфологічними характеристиками.

Рекомендації призначені для селекціонерів, спеціалістів наукових установ, викладачів університетів та аспірантів, які вивчають теоретичні та методологічні основи сучасної селекції, експертизи сортів рослин та проводять роботу по застосуванню сучасних методів статистичної обробки даних.

Методичні рекомендації розглянуті, схвалені та рекомендовані до видання Вченою радою Українського інституту експертизи сортів рослин, протокол № 12 від «24» грудня 2020 року.

Зміст

ВСТУП

Диференціація близькоспоріднених генотипів, до яких належать сорти рослин, відбувається на основі визначених генетичних дистанцій за морфологічними ознаками та ДНК (дезоксирибонуклеїнова кислота) маркерами. Молекулярно-генетичний аналіз сортів рослин за SSR (Simple Sequence Repeat – прості мікросателітні повтори ДНК) маркерами передбачає проведення ПЛР (полімеразна ланцюгова реакція) ДНК досліджуваних зразків та розділення продуктів ампліфікації за допомогою електрофорезу. Першим етапом аналізу є визначення генетичних дистанцій між досліджуваними сортами за ДНК маркерами та морфологічними ознаками. На сьогодні існує достатньо програмних продуктів, які дозволяють провести аналіз даних та розрахувати статистичні показники. В науково-методичних рекомендаціях представлено приклад аналізу даних молекулярно-генетичних досліджень з використанням надбудови для Microsoft Excel XLSTAT та методу з використанням мови програмування для статистичної обробки R.

Для користувачів передбачено безкоштовну демонстраційну версію XLSTAT на 14 днів або замовлення надбудови з різним набором інструментів для статистичного аналізу даних на рік. Альтернативним засобом розрахунку генетичних дистанцій між сортами на основі ДНК маркерів та морфологічних ознак є мови програмування для статистичної обробки R. Основною перевагою застосування R є можливість роботи з такою мовою на безоплатній основі та написання програм, що реалізують нові статистичні методи.

У науково-методичних рекомендаціях представлено результати аналізу 100 ліній кукурудзи за 8 SSR маркерами. З метою визначення кореляційних зв'язків між ДНК маркерами та морфологічними ознаками застосовується тест Мантеля. Тест Мантеля дозволяє оцінити кореляційні зв'язки між двома матрицями генетичних дистанцій. Одна з переваг тесту Мантеля полягає в тому, що, оскільки він здійснюється за матрицями відстаней (відмінності), він може застосовуватися до змінних різного логічного типу (категоріальні, рангові або інтервальні дані).

ПІДГОТОВКА ФАЙЛІВ ДАНИХ

Унаслідок електрофоретичного розділення продуктів ПЛР отримують певний набір ампліконів, на основі якого складають матрицю, в якій позначають наявність (1) або відсутність (0) амплікону для кожного маркера.

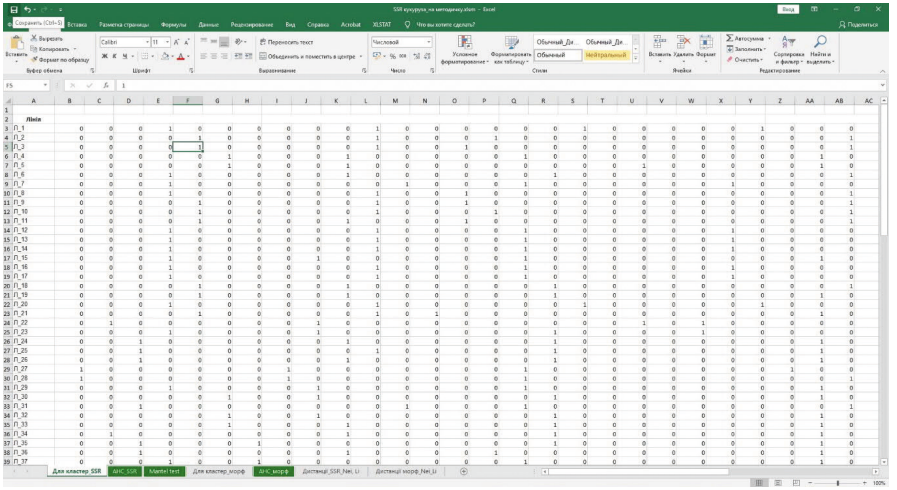


Рис. 1.1. Робоче вікно програми Microsoft Excel з бінарною матрицею результатів молекулярно-генетичного аналізу 100 ліній кукурудзи

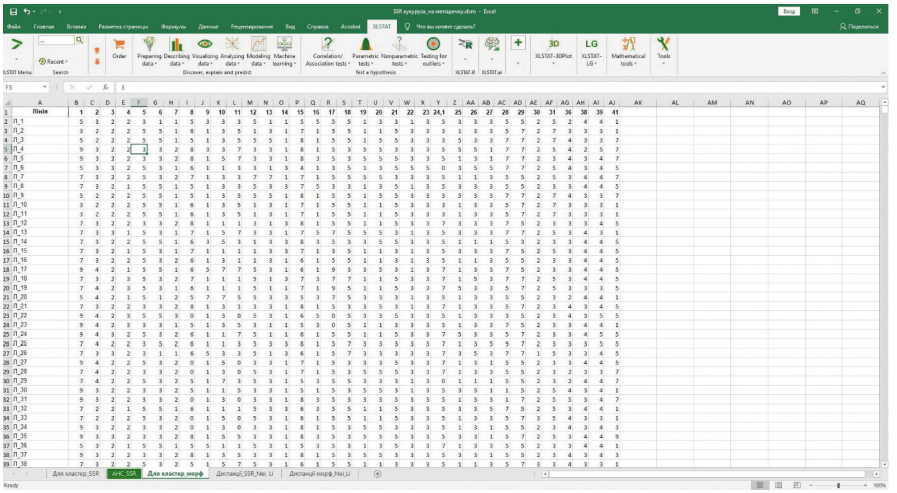


Рис. 1.2. Робоче вікно програми Microsoft Excel з кодами прояву морфологічних ознак 100 ліній кукурудзи: 1-41 – номер ознаки відповідно до Методики проведення експертизи сортів кукурудзи звичайної (*Zea mays* L.) на відмінність, однорідність і стабільність

Для підготовки даних морфологічних ознак для кожного сорту або лінії записують коди прояву ознаки відповідно до методики проведення кваліфікаційної експертизи сортів рослин на відмінність, однорідність та стабільність (ВОС).

Бінарні коди отриманих ампліконів та коди прояву морфологічних ознак для оцінки кореляційних зв'язків за Мантелем з використанням надбудови XLSTAT записують в електронні таблиці Microsoft Excel. Для кожного типу даних (бінарна матриця та коди прояву морфологічних ознак) зручніше використовувати окремі аркуші таблиць Microsoft Excel. Приклад оформлення даних молекулярно-генетичного аналізу та опису морфологічних ознак представлено на рисунках 1.1 та 1.2. Для визначення дистанцій та оцінки кореляційних зв'язків за допомогою мови програмування R формування вихідних даних подібне до поданого вище. Єдиною відмінністю є вимога до назви файлу та аркушів латиницею.

ВИЗНАЧЕННЯ ГЕНЕТИЧНИХ ДИСТАНЦІЙ ТА КОРЕЛЯЦІЙНИХ ЗВ'ЯЗКІВ ЗА МАНТЕЛЕМ ЗА ДОПОМОГОЮ НАДБУДОВИ XLSTAT

1. Визначення генетичних дистанцій між сортами за SSR маркерами та морфологічними ознаками

1.1. Відкрити файл із підготовленими даними. Завантажити XLSTAT натиснувши на піктограму XLSTAT (рис. 1.3) в Microsoft Excel. Після чого відкривається набір доступних інструментів для статистичного аналізу даних (рис. 1.4).

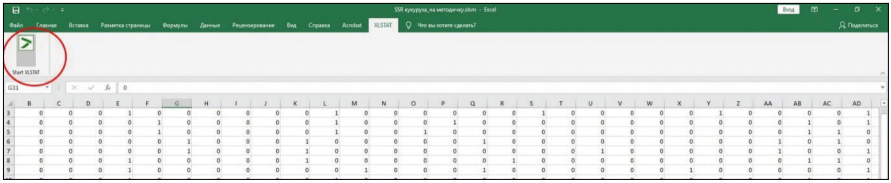


Рис. 1.3. Робоче вікно програми Microsoft Excel із встановленою надбудовою XLSTAT

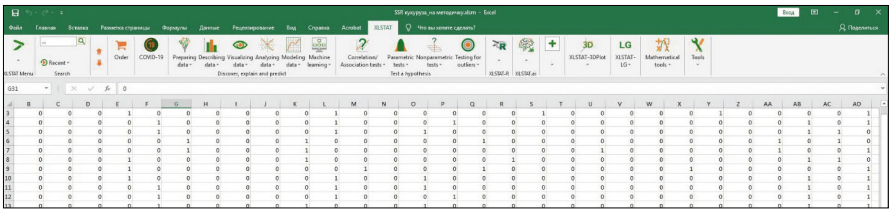


Рис. 1.4. Робоче вікно програми Microsoft Excel із набором інструментів XLSTAT

1.2. Для визначення генетичних дистанцій за SSR маркерами та кодами прояву морфологічних ознак з використанням кластерного аналізу обрати інструмент **Agglomerative hierarchical Clustering (AHC)** (агломераційна ієрархічна кластеризація) в меню **Analyzing data** (Аналіз даних) (рис. 1.5). Цей інструмент дозволяє згрупувати сорти відповідно до їхньої відмінності за досліджуваними ознаками.

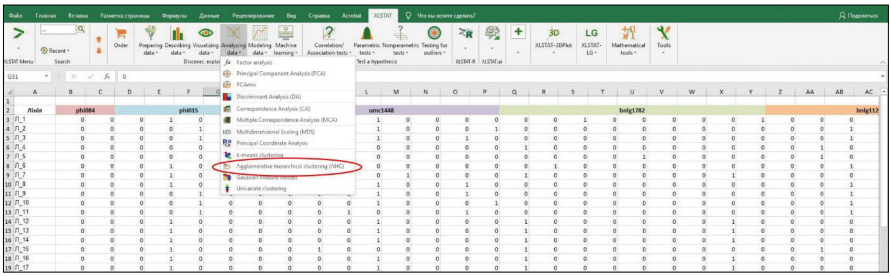


Рис. 1.5. Робоче вікно програми Microsoft Excel: вибір інструменту для визначення генетичних дистанцій в XLSTAT

1.3. У вікні, що відкрилось на вкладці **General** (Загальні параметри) вибрати формат даних **(Data format)** - діапазон значень **(Observations/variables table).** У відповідному полі вибрати діапазон даних, за якими буде проведено кластерний аналіз (бінарна матриця або коди прояву морфологічних ознак) (рис. 1.6).

1.4. Вибрати тип близькості **(Proximity type)** – відмінність **(Dissimilarities)** та тип дистанцій – Евклідові відстані **(Euclidean distance).** Вибрати метод об'єднання в кластери в полі **Agglomeration method** – **Unweighted pair group average** (незважений попарно груповий метод).

1.5. В полі підпис рядків **(Row labels)** вибрати діапазон з назвами сортів.

1.6. На вкладці виведення результатів аналізу **(Outputs)** з усіх запропонованих варіантів вибрати матрицю близькості **(Proximity matrix)** (рис. 1.7).

1.7. Після завершення налаштування параметрів кластерного аналізу натиснути ОК.

1.8. У наступному вікні **(List of selection)** з'явиться інформація про кількість обраних рядків (сортів) та кількість стовпчиків (кількість отриманих ампліконів – для SSR аналізу, або морфологічних ознак). Також буде відображено кількість рядків та стовпчиків, які включають підписи сортів (рис. 1.8).

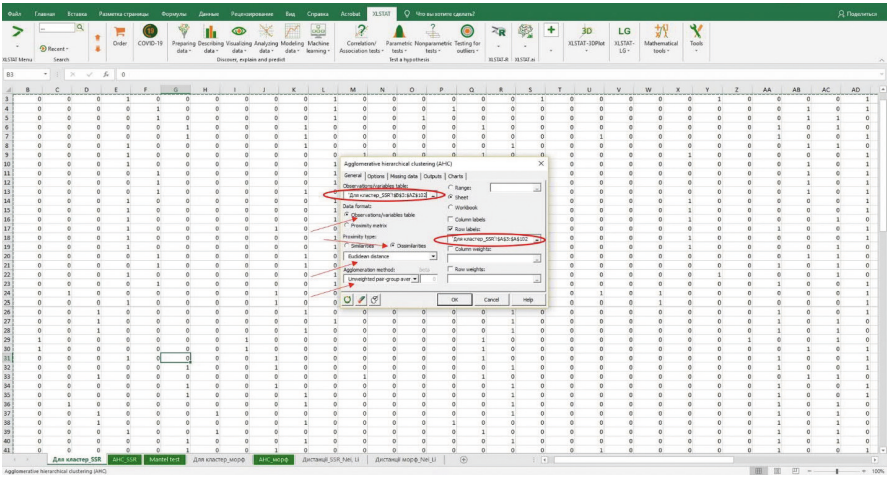


Рис. 1.6. Робоче вікно програми Microsoft Excel: введення параметрів кластерного аналізу (вкладка **General**) з використанням **AHC** в XLSTAT

1.9. Результати кластерного аналізу з використанням **AHC** виводяться на окремий аркуш Microsoft Excel з назвою **AHC** в поточному документі. Відповідно до заданих параметрів результати представлені у вигляді матриці близькості між досліджуваними лініями за

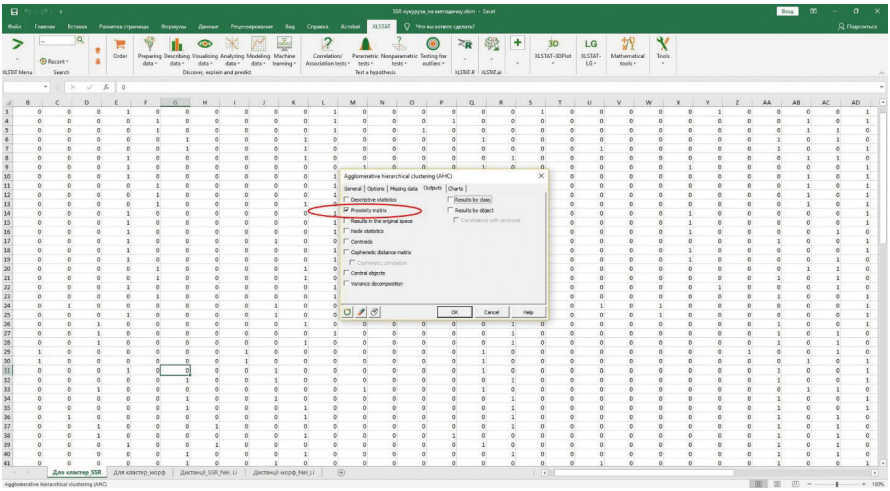


Рис. 1.7. Робоче вікно програми Microsoft Excel: введення параметрів кластерного аналізу (вкладка **Outputs**) з використанням **AHC** в XLSTAT

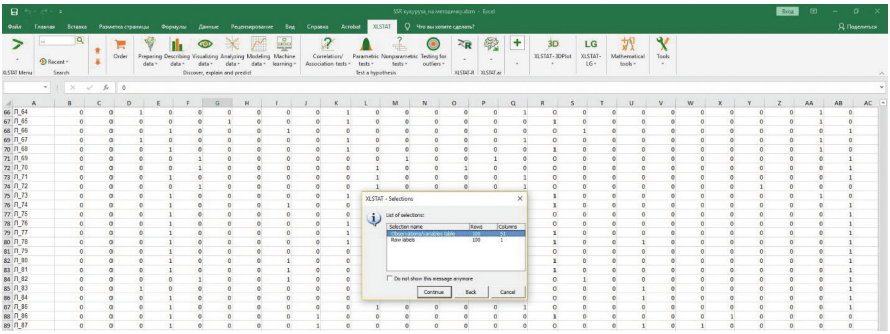


Рис. 1.8. Робоче вікно програми Microsoft Excel: список обраних діапазонів даних для кластерного аналізу з використанням **AHC** в XLSTAT

SSR маркерами (рис. 1.9) та морфологічними ознаками (рис. 1.10). Цифрові значення відображають генетичні дистанції між досліджуваними лініями.

1.10. Обраний тип представлення результатів кластерного аналізу дозволяє отримати дендрограму, що відображає розподіл досліджуваних ліній на кластери відповідно до генетичних дистанцій та дендрограму, яку сформовано за групами отриманих кластерів досліджуваних ліній (рис. 1.11).

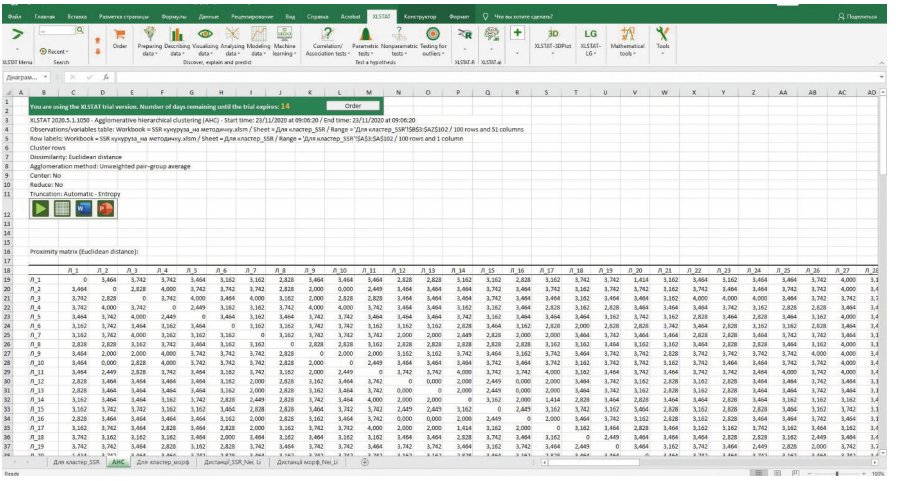


Рис. 1.9. Робоче вікно програми Microsoft Excel: результати кластерного аналізу ліній кукурудзи за SSR маркерами з використанням **AHC** в XLSTAT

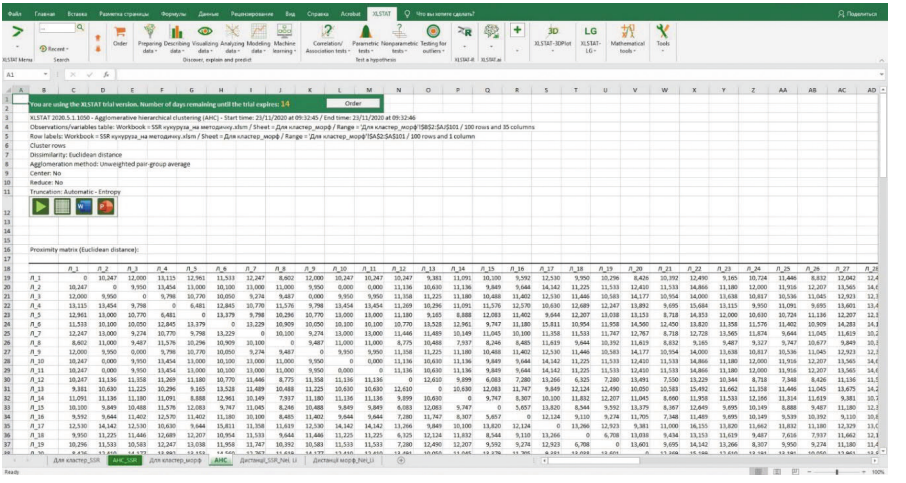


Рис. 1.10. Робоче вікно програми Microsoft Excel: результати кластерного аналізу ліній кукурудзи за морфологічними ознаками з використанням **AHC** в XLSTAT

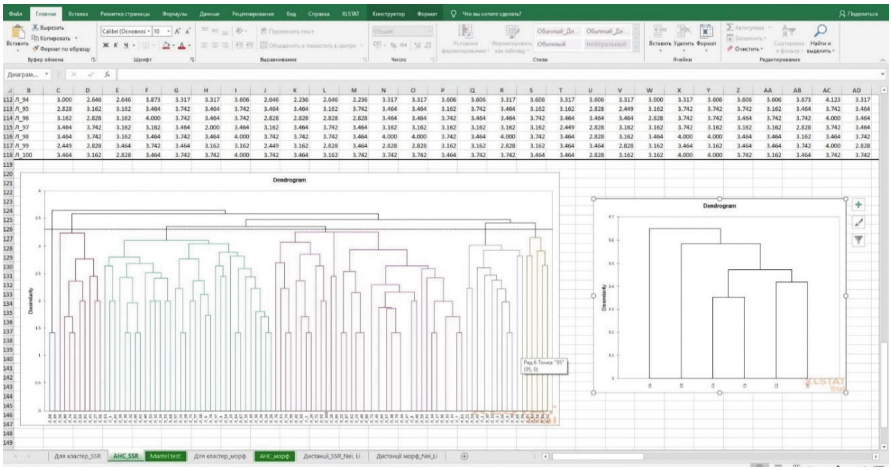


Рис. 1.11. Робоче вікно програми Microsoft Excel: дендрограми, які відображають розподіл досліджуваних ліній на кластери за допомогою **АНС** в XLSTAT

2. Визначення кореляційних зв'язків на основі генетичних дистанцій за SSR маркерами та морфологічними ознаками з використанням тесту Мантеля

2.1. Вихідними даними для проведення тесту за Мантелем є матриці близькості між досліджуваними сортами за ДНК маркерами та морфологічними ознаками. Такі матриці можуть бути розміщені як на одному, так і на окремих аркушах Microsoft Excel.

2.2. Завантажити XLSTAT як описано в п. 1.1 для доступу до набору інструментів статистичного аналізу даних.

2.3. Для початку аналізу обрати інструмент **Correlation/Association test** (Кореляційний/Асоціаційний тест). Із випадаючого списку вибрати **Mantel test** (рис. 2.1).

2.4. У вікні, що відкриється на вкладці **General** (Загальні параметри), зняти позначку **Labels included** (Включити підписи зразків) та виділити діапазони даних: **Matrix A** – матриця близькості між досліджуваними сортами за ДНК маркерами; **Matrix B** – матриця близькості між досліджуваними сортами за кодами прояву морфологічних ознак без підписів сортів (рис. 2.2).

2.5. На вкладці **Options** (Параметри) є можливість обрати тип альтернативної гіпотези **(Alternative hypotheses)**: спрямована (правостороння **r > 0**, лівостороння **r < 0)** та неспрямована **(r = 0)**; рівень значущості **(Significant level)**; метод обчислення ймовірності помилки першого роду **(p-values computation)** та тип кореляції **(Type of correlation)**. Для визначення кореляції за Мантелем між лініями кукурудзи за SSR маркерами та морфологічними ознаками обрані наступні параметри: правостороння альтернативна гіпотеза, рівень значущості – 5%, метод обчислення ймовірності помил-

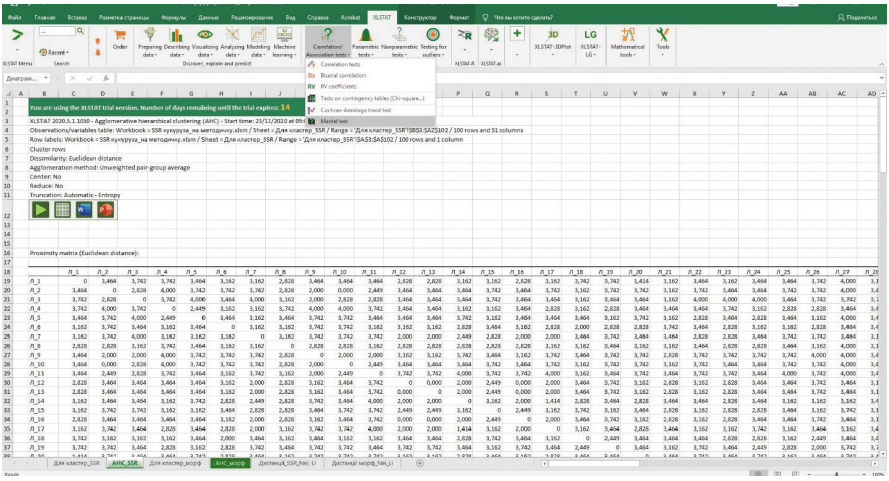


Рис. 2.1. Робоче вікно програми Microsoft Excel: вибір інструменту для проведення тесту за Мантелем

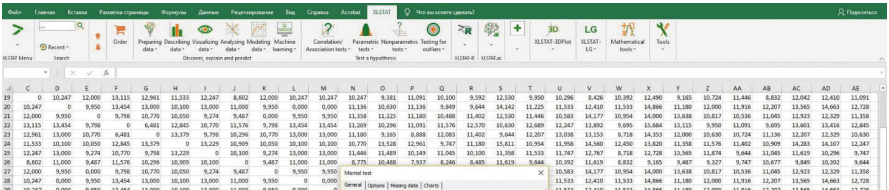


Рис. 2.2. Робоче вікно програми Microsoft Excel: введення параметрів тесту Мантела (вкладка **General**) з використанням **Mantel test** в XLSTAT

ки першого роду – **Monte Carlo method** та кореляція за Пірсоном **(Pearson)** (рис. 2.3).

2.6. На вкладці **Charts** (Діаграми) позначити два типи діаграм, які будуються відповідно до визначених взаємодій за ДНК маркерами та морфологічними ознаками: діаграма розсіяння **(Scatter plot)** та гістограма **(Histogram)** (рис. 2.4).

2.7. Після введення параметрів натиснути ОК. У наступному вікні **(List of selection)** з'явиться інформація про кількість обраних

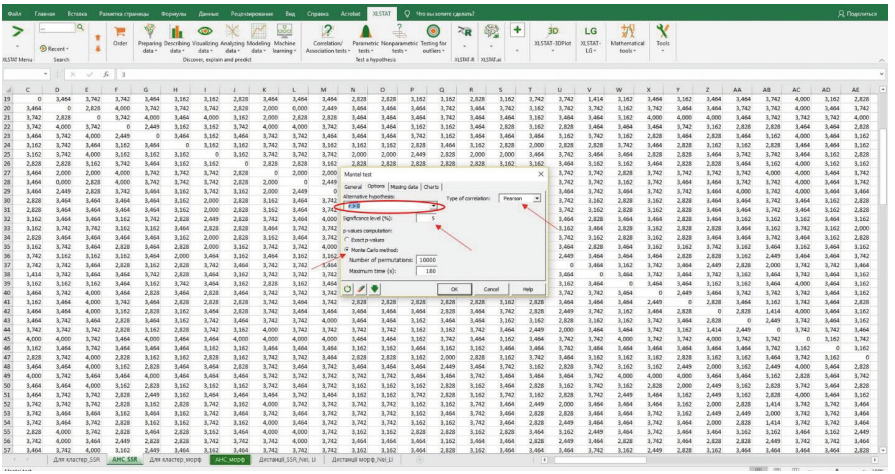


Рис. 2.3. Робоче вікно програми Microsoft Excel: введення параметрів тесту Мантеля (вкладка **Options**) з використанням **Mantel test** в XLSTAT

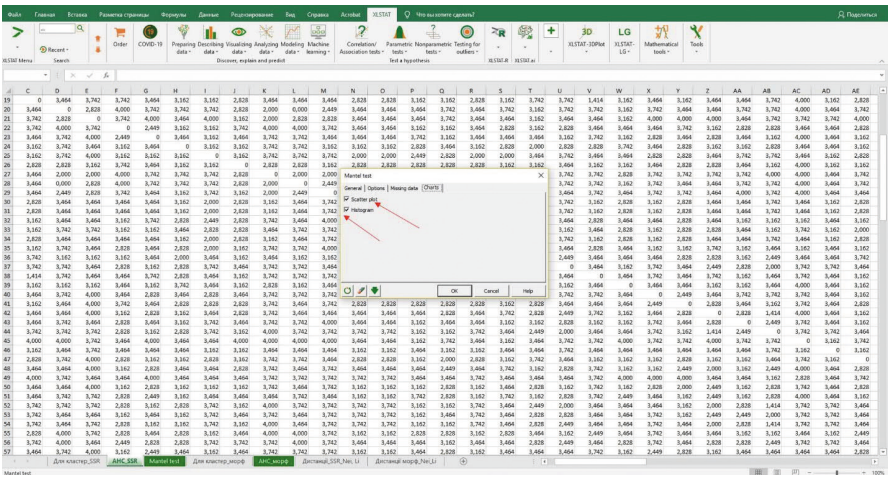


Рис. 2.4. Робоче вікно програми Microsoft Excel: введення параметрів тесту Мантеля (вкладка **Charts**) з використанням **Mantel test** в XLSTAT

рядків та кількість стовпчиків у вибраних діапазонах даних матриць близькості між досліджуваними сортами за ДНК маркерами та за кодами прояву морфологічних ознак (рис. 2.5).

2.8. Результати тесту Мантеля виводяться на окремий аркуш Microsoft Excel з назвою **Mantel test** в поточному документі. Відповідно до заданих параметрів на аркуші результатів наведені діапазони досліджуваних даних, тип альтернативної гіпотези, рівень значущості (α), метод обчислення ймовірності помилки першого

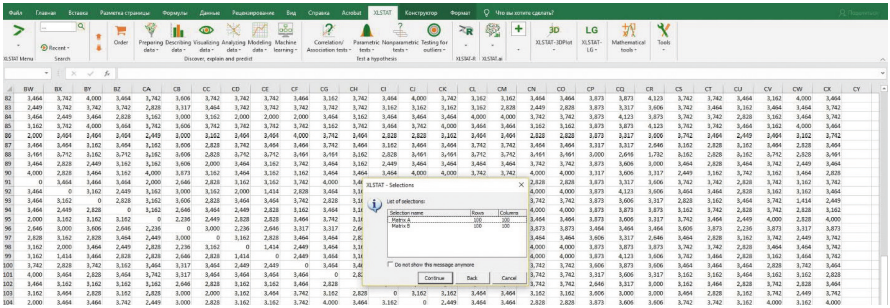
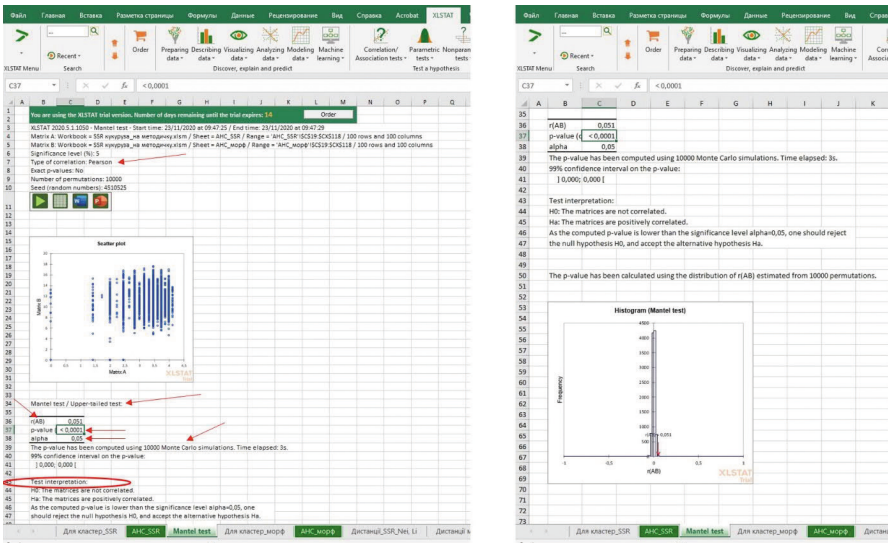


Рис. 2.5. Робоче вікно програми Microsoft Excel: список обраних діапазонів даних для проведення тесту Мантеля в XLSTAT

роду (**p-value**), тип кореляції, коефіцієнт кореляції та пояснення щодо інтерпретації альтернативної гіпотези (рис. 2.6). У результатах також наведені діаграма розсіювання (рис. 2.6 а) та гістограма нормальності розподілу досліджуваних даних (рис. 2.6 б).

2.9. Відповідно до інтерпретації тесту гіпотеза про відсутність кореляції (нульова гіпотеза – H0) приймається за умови $\mathbf{p}>\alpha$. В результаті аналізу 100 ліній кукурудзи за SSR маркерами та морфологічними ознаками з використанням тесту Мантеля визначено, що $\mathbf{p}<\alpha$, тобто слід відхилити нульову гіпотезу та прийняти альтернативну гіпотезу про наявність кореляції (Ha). Коефіцієнт кореляції r становить 0,051.



а **б**

Рис. 2.6. Робоче вікно програми Microsoft Excel: результати тесту Мантеля ліній кукурудзи за SSR маркерами та морфологічними ознаками в XLSTAT

ВИЗНАЧЕННЯ ГЕНЕТИЧНИХ ДИСТАНЦІЙ ТА КОРЕЛЯЦІЙНИХ ЗВ'ЯЗКІВ ЗА МАНТЕЛЕМ ЗА ДОПОМОГОЮ МОВИ ПРОГРАМУВАННЯ ДЛЯ СТАТИСТИЧНОЇ ОБРОБКИ R

3. Визначення генетичних дистанцій між сортами за SSR маркерами та морфологічними ознаками в R

Для роботи з мовою програмування R використано вільне середовище розробки програмного забезпечення RStudio (https://rstudio.com).

3.1. Завантаження вихідних даних до робочого оточення. Для завантаження даних за допомогою RStudio, необхідно натиснути у вкладці інтерфейсу **Environment** кнопку **Import Dataset**, у випадаючому меню якої обрати **From Excel…** (рис. 3.1)

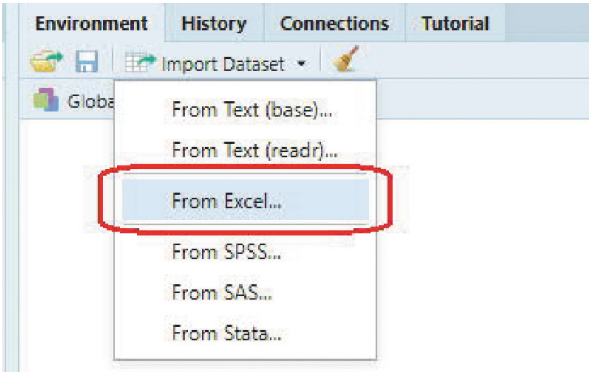


Рис. 3.1. Робоче вікно програмного середовища RStudio: вибір типу завантаження вихідних даних

У вікні, що відкрилося (рис. 3.2), необхідно вибрати файл з вихідними даними (**Browse…**). У пунктах меню **Name** та **Sheet** визначається змінна, в яку записуються вихідні дані (**SSR** для SSR маркерів) та лист Excel, де знаходяться дані (**SSR**) відповідно. В робочому вікні завантаження даних відобразиться попередній перегляд даних для завантаження. Після чого, для завантаження в робоче оточення, необхідно натиснути кнопку **Import** (рис. 3.2).

Інший варіант завантаження даних – використання команд мови R. Для збереження команд R і їхнього повторного використання для інших даних, рекомендовано використовувати файл скриптів R. Для створення файлу скрипта, де зберігатимуться команди, необхідно натиснути на піктограму **New File**, в списку вибору файлів обрати R Script (рис. 3.3).

Для описаного вище завантаження даних SSR маркерів за допомогою команд R, у створеному файлі скрипта необхідно виконати команду:

```
SSR <- read_excel(«Distance/Datas.xlsx», sheet = "SSR")
```

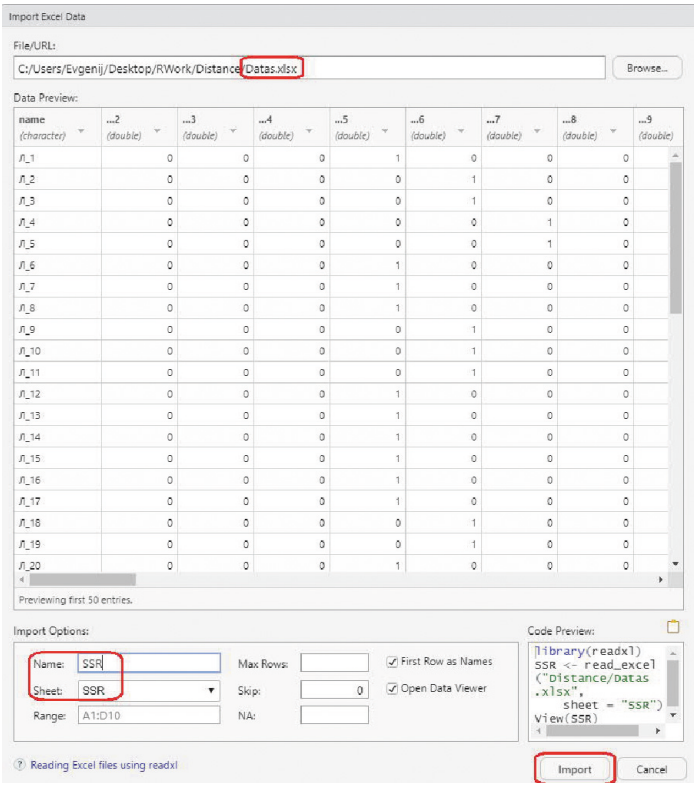


Рис. 3.2. Робоче вікно програмного середовища RStudio: меню завантаження даних з файлів Excel

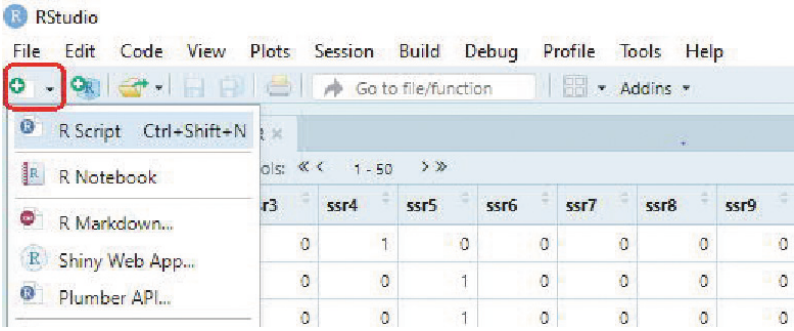


Рис. 3.3. Робоче вікно програмного середовища RStudio: створення файлу із скриптами (командами) R

Аналогічно для завантаження даних морфологічних ознак:
Morf <- **read_excel**(«Distance/Datas.xlsx», sheet = "Morf")

Описані команди імпортують дані та зберігають в змінних **SSR** (SSR маркери) та **Morf** (морфологічні ознаки).

3.2. Встановлення додаткових пакетів та форматування даних. Базові команди мови R дозволяють виконувати великий спектр стандартного статистичного аналізу. Проте для виконання додаткових методів аналізу, побудови складних графіків та трансформації даних необхідне завантаження додаткових пакетів з репозиторію. Так, зокрема для експорту отриманих табличних результатів в Excel необхідний пакет **openxlsx**. Для встановлення пакетів можливостями середовища RStudio у вкладці Packages натиснути кнопку **Install** (рис. 3.4). У діалоговому вікні, що відкриється, в пункті **Packages** ввести назву пакета для встановлення (openxlsx) і натиснути кнопку **Install** (рис. 3.6).

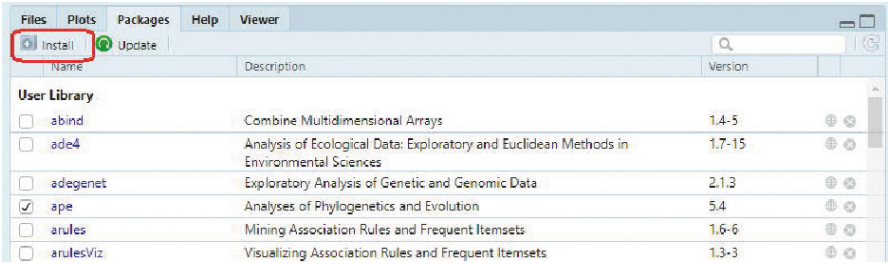


Рис. 3.4. Робоче вікно програмного середовища RStudio: меню встановлення, підключення та оновлення пакетів

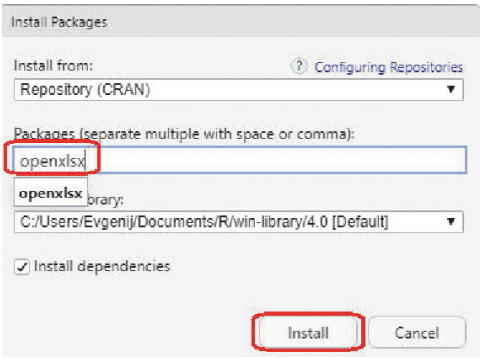


Рис. 3.5. Робоче вікно програмного середовища RStudio: встановлення пакету **openxlsx**

Дані дії, без використання можливостей інтерфейсу, можна виконати за допомогою команди:

```
install.packages("openxlsx ", dependencies = T)
```

Для розрахунку дистанцій та відображення міток на графіках необхідне попереднє форматування даних: додавання назв рядків. Для чого необхідно виконати команди, які перетворюють першу колонку завантажених даних в назви рядків:

```
# Додавання назв рядків для SSR маркерів
SSR <- data.frame(SSR, row.names = 1)
# Додавання назв рядків для морфологічних ознак
Morf <- data.frame(Morf, row.names = 1)
```

3.3. Створення матриці дистанцій. Алгоритм визначення генетичних дистанцій за SSR маркерами та кодами прояву морфологічних ознак з використанням кластерного аналізу в мові програмування R відрізняється від алгоритму в XLSTAT. В R спочатку формують матрицю дистанцій, яку потім використовують для розбиття на кластери. Для створення матриці дистанцій використовують команду **dist**. Формат команди:

dist(x, method = «euclidean», diag = FALSE, upper = FALSE)

де x – таблиця або матриця вихідних даних;

method – міра відстані, яку слід використовувати («евклідові», «максимальні», «манхеттенські», «канберрські», «двійкові» або «мінковські»), за замовчуванням «евклідові»;

diag – логічне значення, яке вказує, чи слід друкувати діагональ матриці;

upper – логічне значення, що вказує, чи слід визначати трикутник матриці відстані.

Побудову матриці евклідових дистанцій для завантажених даних виконують наступними командами:

```
# Створення евклідових дистанцій для SSR маркерів
SSR.dist <- dist(SSR, method='euclidean', diag = T, upper=TRUE)
# Створення евклідових дистанцій для морфологічних ознак
Morf.dist<- dist(Morf, method='euclidean', diag = T, upper=TRUE)
```

Згідно з поданими командами результати зберігаються в змінних SSR.dist (SSR маркери) та Morf.dist (морфологічні ознаки). У параметрах команди вказано Евкладові дистанції: **euclidean**, а також побудова діагоналі матриці та верхнього трикутника: **diag = T** та **upper=TRUE**, відповідно.

Використовуючи команду **write.xlsx** з встановленого раніше пакету **openxlsx** отримані матриці дистанцій можна зберегти в файли формату **.xlsx** (книга Excel):

```
# Підключення бібліотеки (пакету)
library(openxlsx)
# Експорт даних в формат .xlsx
write.xlsx(as.matrix(SSR.dist), file="Distance/DistanceSSR.xlsx",rowNames=TRUE)
write.xlsx(as.matrix(Morf.dist),file="Distance/DistanceMorf.xlsx",rowNames=TRUE)
```

3.4. Проведення кластеризації. Основною командою для ієрархічного кластерного аналізу в мові програмування R є **hclust**. Формат команди:

hclust(d, method = «complete»)

де d – дистанції, створені методом dist;

method – метод агломерації, який буде використовуватися. Може бути одне з «ward.D», «ward.D2», «single», «complete», «average» (=

UPGMA), «mcquitty» (= WPGMA), «median» (= WPGMC) або « centroid" (= UPGMC). За замовчуванням *complete* – метод повних зв'язків.

Для проведення агломераційної ієрархічної кластеризації з використанням незважено попарно групового методу використано вказану команду з указанням аргументу **method = 'average'**:

```
# Побудова кластерів по методу UPGMA для SSR маркерів
hclust_avg <- hclust(SSR.dist, method = 'average')
# Побудова кластерів по методу UPGMA для морфологічних ознак
hclust_avg.Morf <- hclust(Morf.dist, method = 'average')
```

Результати зберігають у відповідні змінні **hclust_avg** та **hclust_avg.Morf**. На рис. 3.6 подано структуру змінної для SSR маркерів.

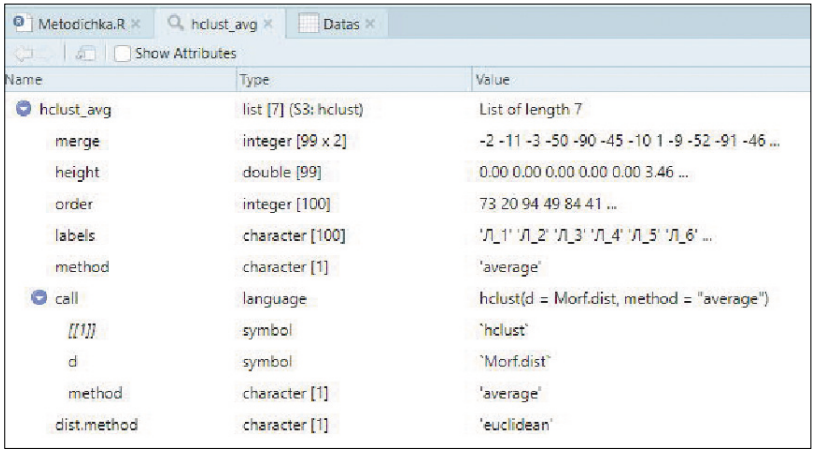


Рис. 3.6. Робоче вікно програмного середовища RStudio: структура змінної результатів кластеризації **hclust_avg**

Дана змінна має формат списку і зберігає в елементах списку відомості про параметри кластеризації:

- merge – матриця *n-1* на *2*. Рядок і об'єднання описує злиття кластерів на *i- му* етапі кластеризації. Якщо елемент *j* у рядку від'ємний, то спостереження *-j* було об'єднано на цьому етапі. Якщо *j* позитивне, тоді злиття було з кластером, сформованим на (попередньому) етапі *j* алгоритму;
- height – набір дійсних значень *n-1*. Висота кластеризації: тобто значення критерію, пов'язаного з методом кластеризації для конкретної агломерації.
- order – вектор, що дає перестановку вихідних спостережень, придатних для побудови графіків;
- label – мітки для кожного з кластеризованих об'єктів;
- method – використаний метод кластеризації;
- dist.method – відстань, яка була використана для створення дистанцій.

Так, використання елемента списку order дозволяє, використовуючи стандартну команду для створення діаграм, вивести дендрограму (рис. 3.7):

```
plot(hclust_avg, cex = 0.7, hang=-1)
```

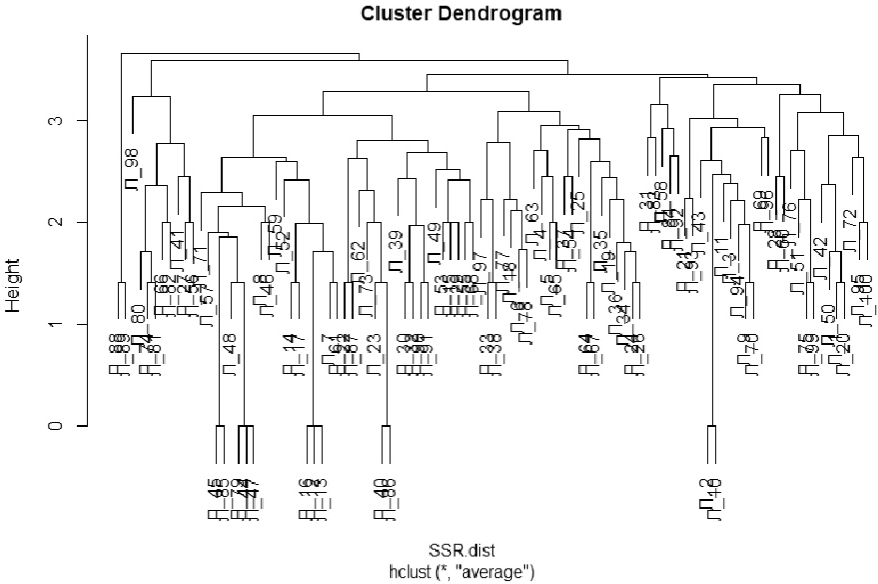


Рис. 3.7. Дендрограма кластеризації SSR маркерів без розбиття

3.5. Побудова дендрограми з розбиттям на кластери. Для побудови дендрограми ієрархічної кластеризації використано можливості пакетів додаткової графіки ggplot2 та dendextend. Для цього виконано наступну послідовність команд:

– підключення бібліотек:

```
library(ggplot2)
library(dendextend)
```

– визначення палітри кольорів:

```
colors6 = c("red", "blue", "green", "black","Gold","Violet")
```

– створення змінної типу **dendrogram** на основі результатів кластеризації SSR маркерів:

```
dendro <- as.dendrogram(hclust_avg)
```

– визначення параметрів дендрограми – кількість кластерів, типів міток, розмірів та створення полотна:

```
dendro.col <- dendro %>% set("branches_k_color", k = 6, value =
colors6) %>% set("branches_lwd", 0.6) %>%  set("labels_colors", value =
c("darkslategray")) %>%  set("labels_cex", 0.5)
ggd1 <- as.ggdend(dendro.col)
```

– створення дендрограми пакетом ggplot2:

```
ggplot(ggd1, theme = theme_minimal()) labs(x = "Observations", y =
"Height", title = "Dendrogram, k = 6")
```

Унаслідок виконання останньої команди отримано дендрограму кластеризації SSR маркерів з розбиттям на 6 кластерів (рис. 3.8).

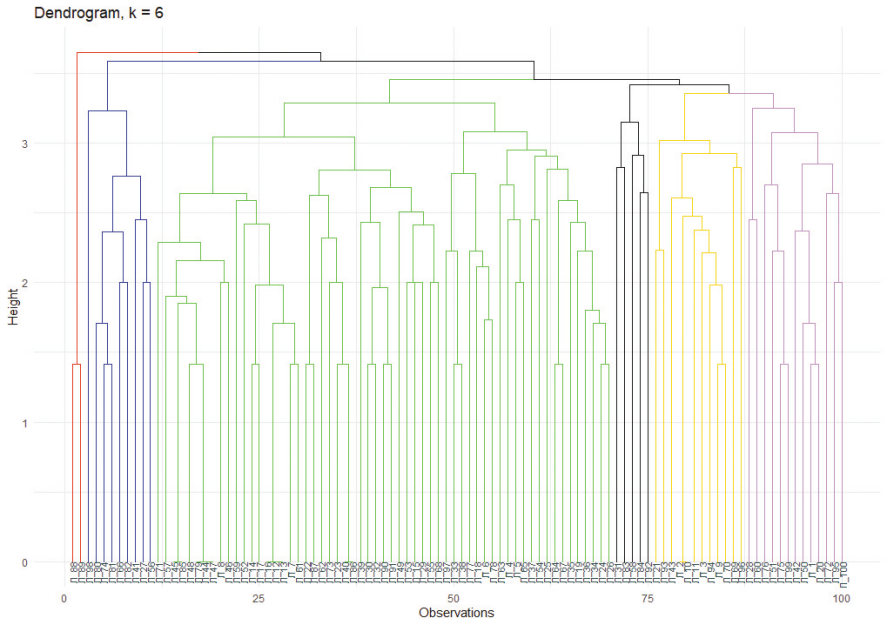


Рис. 3.8. Дендрограма кластеризації SSR маркерів з розбиттям на 6 кластерів

Виконання попередньої команди з вказанням інших параметрів, дозволяє побудувати радіальну дендрограму (рис. 3.9)

```
ggplot(ggd1, labels = T) scale_y_reverse(expand = c(0.2, 0))
coord_polar(theta="x")
```

Повний скрипт всіх команд R подано в лістингу 3.1:

```
# Завантаження даних
library(readxl)
SSR <- read_excel(«Distance/Datas.xlsx», sheet = "SSR")
Morf <- read_excel(«Distance/Datas.xlsx», sheet = "Morf")
# Додавання назв рядків для SSR маркерів
SSR <- data.frame(SSR, row.names = 1)
# Додавання назв рядків для морфологічних ознак
Morf <- data.frame(Morf, row.names = 1)
# Створення евклідових дистанцій для SSR маркерів
SSR.dist <- dist(SSR, method='euclidean', diag = T, upper=TRUE)
# Створення евклідових дистанцій для морфологічних ознак
Morf.dist<- dist(Morf, method='euclidean', diag = T, upper=TRUE)
```

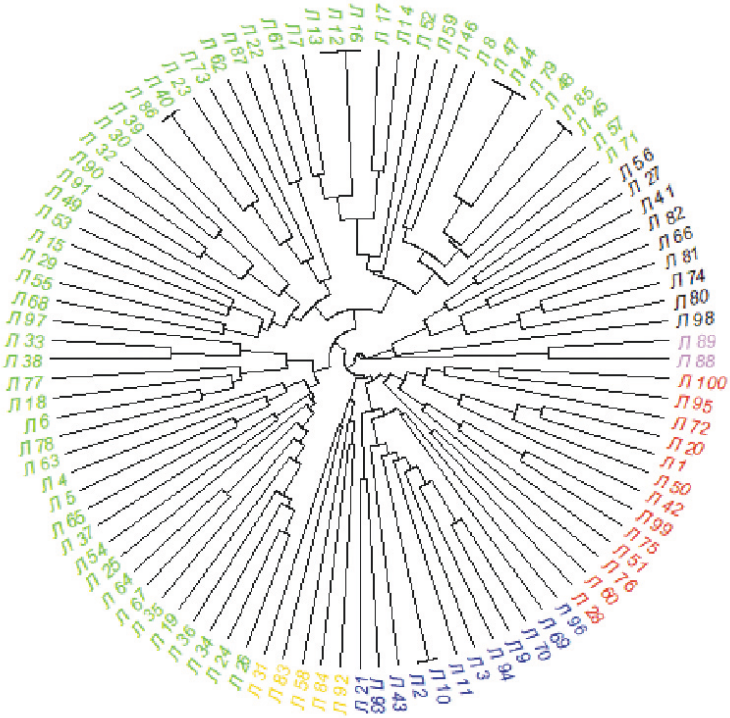


Рис. 3.9. Радіальна дендрограма кластеризації SSR маркерів з розбиттям на 6 кластерів

```
    # Експорт даних в формат .xlsx
library(openxlsx)
    write.xlsx(as.matrix(SSR.dist), file="Distance/DistanceSSR.
xlsx",rowNames=TRUE)
write.xlsx(as.matrix(Morf.dist),file="Distance/DistanceMorf.
xlsx",rowNames=TRUE)
    # Побудова кластерів по методу UPGMA для SSR маркерів
hclust_avg <- hclust(SSR.dist, method = 'average')
# Побудова кластерів по методу UPGMA для морфологічних ознак
hclust_avg.Morf <- hclust(Morf.dist, method = 'average')
    plot(hclust_avg, cex = 0.7, hang=-1)
    #Створення дендрограми з розбиттям
library(ggplot2)
    library(dendextend)
    colors6 = c("red", "blue", "green", "black","Gold","Violet")
    dendro.col <- dendro %>% set("branches_k_color", k = 6, value =
colors6) %>% set("branches_lwd", 0.6) %>%  set("labels_colors", value =
c("darkslategray")) %>%   set("labels_cex", 0.5)
ggd1 <- as.ggdend(dendro.col)
```

```
    ggplot(ggd1, theme = theme_minimal()) labs(x = "Observations", y =
"Height", title = "Dendrogram, k = 6")
    #Створення радіальної дендрограми з розбиттям
ggplot(ggd1, labels = T) scale_y_reverse(expand = c(0.2, 0))
 coord_polar(theta="x")
```

Лістинг 3.1. Скрипт визначення генетичних дистанцій між сортами за SSR маркерами та морфологічними ознаками в R

4. Визначення кореляційних зв'язків на основі генетичних дистанцій за SSR маркерами та морфологічними ознаками з використанням тесту Мантеля методами мови програмування R

Для проведення тесту за Мантелем в мові програмування R існує велика кількість пакетів, що містять необхідні методи, зокрема **ade4, ape, vegan, cultevo, EcoGenetics.** У даній роботі використано останній з перелічених.

EcoGenetics – це пакет R з інфраструктурою та гнучкими методами, розробленими для спрощення та прискорення процедур аналізу даних популяційних екологів та генетиків. Для екогенних об'єктів визначено ряд методів маніпулювання даними та надано функції для дослідницького аналізу просторової автокореляції, для однієї та кількох змінних, з інтерактивним середовищем візуалізації даних.

Для проведення тесту за Мантелем за допомогою пакету **EcoGenetics** використано функцію **eco.mantel**. Формат функції:

eco.mantel(d1, d2, dc = NULL, method = c(«pearson», «spearman», «kendall»), nsim = 99, alternative = c(«auto», «two.sided», «less», «greater»), plotit = T, ...)

де ***d1*** та ***d2*** – матриці відстані, що порівнюються;

method – метод кореляції, що використовується для побудови статистики («пірсона», «спірмана» або «кендалла»);

nsim – кількість моделювань (перестановок) Монте-Карло;

alternative – альтернативна гіпотеза. Варіанти: неспрямований двосторонній – «two.sided», спрямовані правосторонній – «greater» та лівосторонній – «less»;

plotit – побудова гістограми моделювання (за замовчуванням – так).

Результатом виконання є об'єкт класу «eco.gsa» із такими даними:

– Метод, що використовується при аналізі;

– Спостережуване значення кореляції – ***OBS***;

– Очікуване значення кореляції – ***EXP***;

– Значення ймовірності помилки першого роду – ***P-PVAL***;

– Альтернативна гіпотеза – ***ALTER***;

– Кількість моделювань – ***NSIM***.

Для визначення кореляційних зв'язків за Мантель-тестом між матрицями дистанцій в параметрах функції **eco.mantel** вказано матриці відстаней – SSR.dist (SSR маркери) та Morf.dist (морфологічні ознаки); тип кореляції – Пірсона; альтернатива – спрямована правостороння; кількість перестановок – 9999. Послідовність ко-

манд включає підключення пакету **EcoGenetics** та виконання команди **eco.mantel** з вказаними параметрами:

```
   #Підключення бібліотеки:
library(EcoGenetics)
   #Виконання Мантель-тесту
eco.mantel(SSR.dist, Morf.dist, alternative = "greater",
nsim = 9999)
```

Результати виконання подано на рис. 4.1 та лістингу 4.1:

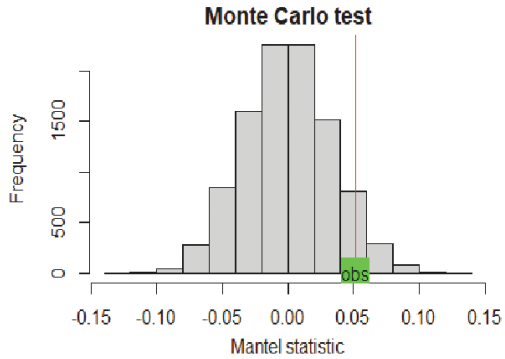


Рис. 4.1. Гістограми моделювання кореляційних зв'язків з використанням тесту Мантеля

```
   ## ############################
##   Mantel test
##   > Correlation coefficient used -->  Pearson
##   > Number of simulations -->  9999
##   > Alternative -->  greater
##   > P-value -->  0.0065
##   > Observed value -->  0.0512
##   > Expected value -->  -2e-04
   ## ############################
```

Лістинг 4.1. Результати тесту Мантеля ліній кукурудзи за SSR маркерами та морфологічними ознаками з використанням функції **eco.mantel** пакету **EcoGenetics**

Отримані в лістингу 4.1 результати відповідають отриманим з використанням надбудови XLSTAT: коефіцієнт кореляції (Observed value) – 0,051; ймовірність отримання помилки першого роду нижче визначеного рівня значимості – 0,0065 (P-value)<0,05.

Даний скрипт використовує збережені в робочому оточенні дані матриць. У разі використання збережених у форматі .xlsx даних матриць дистанцій (приклад збереження подано в попередньому розділі) необхідно виконати подану нижче послідовність дій:

```
   #Завантаження збережених даних з матрицями
SSR.data <- read_excel("Distance/DistanceSSR.xlsx")
Morf.data <- read_excel("Distance/DistanceMorf.xlsx")
```

```
    #Трансформація даних
    SSR.data <- data.frame(SSR.data, row.names = 1)
    Morf.data <- data.frame(Morf.data, row.names = 1)
#Перетворення вихідних даних в тип даних – матриці дистанцій
SSR.dist <- as.dist(SSR.data)
Morf.dist <- as.dist(Morf.data)
    #Проведення Мантель-тесту
eco.mantel(SSR.dist, Morf.dist, alternative = "greater",
nsim = 9999)
```

Отже, запропоноване застосування надбудови для Microsoft Excel XLSTAT та методу з використанням мови програмування для статистичної обробки R дозволяє оцінити генетичні дистанції та кореляційні зв'язки із застосуванням тесту Мантеля за ДНК маркерами та морфологічними ознаками сортів рослин. Можна відзначити, що надбудова для Microsoft Excel XLSTAT дозволяє зосередити увагу не на рутинних обчисленнях за складними формулами, а на безпосередньому аналізі результатів досліджень. В той час як використання мови програмування для статистичної обробки R може задовольнити потреби творчого дослідника, що віддає перевагу самостійному контролю за ходом обчислювального процесу.

Список використаних літературних джерел

1. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. Москва: ДМК Пресс, 2016. 460 с.

2. Мастицкий С. Э., Шитиков В. К. Статистический анализ и визуализация данных с помощью R. М.: ДМК Пресс, 2015. 496 с.

3. Шитиков В. К., Мастицкий С. Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 2017. 351 с.

4. Burstin J., Charcosset A. Relationship between phenotypic and marker distances: theoretical and experimental investigations. Heredity. 1997. Vol. 79, No. 5. P. 477.

5. Darvishzadeh R. Phenotypic and molecular marker distance as a tool for prediction of heterosis and F1 performance in sunflower (Helianthus annuus' L.) under well-watered and water-stressed conditions. *Australian Journal of Crop Science*. 2012. Vol. 6, No. 4. P. 732.

6. Diniz-Filho J. A. F., Soares T. N., Lima J. S. et al. Mantel test in population genetics. *Genetics and molecular biology*. 2013. Vol. *36*, No. 4. C. 475–485.

7. Dutilleul P., Stockwell J. D., Frigon D., Legendre P. The Mantel test versus Pearson's correlation analysis: Assessment of the differences for biological and environmental studies. *Journal of Agricultural, Biological, and Environmental Statistics*. 2000. P. 131–150.

8. Fortin M. J., Dale M. R., Ver Hoef J. M. Spatial analysis in ecology. Wiley StatsRef: Statistics Reference Online. 2002. URL: http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat07766.pub2/full

9. Harmon L. J., Glor R. E. (2010). Poor statistical performance of the Mantel test in phylogenetic comparative analyses. *Evolution: International Journal of Organic Evolution*. 2010. Vol. 64, No. 7. P. 2173–2178. doi:10.1111/j.1558-5646.2010.00973.x

10. Karuri H. W., Ateka E. M., Amata R. et al. Evaluating diversity among Kenyan sweet potato genotypes using morphological and SSR markers. *Int. J. Agric. Biol*. 2010. Vol. 12, No. 1. P. 33–38.

11. Legendre P., Fortin M. J. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular ecology resources*. 2010. Vol. 10, No. 5. P. 831–844. doi: 10.1111/j.1755-0998.2010.02866.x

12. Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer research*. 1967. Vol. 27, No. 2 Part 1. P. 209–220.

13. R в действии. Анализ и визуализация данных в программе R / пер. С англ. Полины А. Волковой. М. ДМК Пресс, 2014. 588 с.

14. Riday H., Brummer E. C., Campbell T. A. et al**.** Comparisons of genetic and morphological distance with heterosis between Medicago sativa subsp. sativa and subsp. falcata. *Euphytica*. 2003. Vol. *131*, No. 1. P. 37–45.

15. Smouse P. E., Long J. C., Sokal R. R. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic zoology*. 1986. Vol. *35*, No. 4. P. 627–632.

Застосування тесту Мантеля для оцінки кореляційних зв'язків між ДНК та морфологічними маркерами сортів рослин

Науково-методичні рекомендації

Рекомендовано до друку Вченою радою Українського інституту експертизи сортів рослин, протокол № 12 від «24» грудня 2020 року.

Науково методичні рекомендації підготували: Стариченко Є. М., Присяжнюк Л. М., Гринів С. М., Лещук Н. В., Ткачик С. О., Києнко З. Б., Мельник С. І.
Український інститут експертизи сортів рослин, 2020.

Комп'ютерна верстка Бойко А. І.

Формат 64Ч84 1/16. Папір крейдовий.
Друк цифровий. Гарнітура Schoolbook
Друк. арк. Умов. друк. арк.
Наклад прим. Зам.
Віддруковано з оригіналів замовника ТОВ «ТВОРИ»
Видавець ТОВ «ТВОРИ»
Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру видавців, виготовлювачів і розповсюджувачів видавничої продукції серія ДК № 6188 від 18.05.2018 р.

21027, м. Вінниця,вул. Келецька, 51а, прим. 143.
Тел.: (0432) 603-000, 69-67-69
e-mail: info@tvoru.com.ua
http://www.tvoru.com.ua