Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра інформатики факультету інформатики

РОЗРОБКА ПРОГРАМНОГО ЗАСТОСУНКУ ІМПОРТУ ТА АНАЛІЗУ ДАНИХ ЛАНЦЮЖКІВ МІТОХОНДРІАЛЬНОЇ ДНК

Текстова частина до курсової роботи
за спеціальністю „Комп'ютерні науки" 122

Керівник курсової роботи
к.ф.-м.н., доц. Гулаєва Н. М.
_____
(підпис)
"____" _____ 2020 р.

Виконала:
студентка 3 р.н.
БП «Комп'ютерні науки»
Самовол М. С.
"____" _____ 2020 р.

Київ 2020

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ
Зав.кафедри інформатики,
к.ф.-м.н., доц. Гороховський С.С.
_____
(підпис)
„____” _____2020 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
на курсову роботу

студентці 3-го курсу, факультету інформатики
Самовол Марії Сергіївні

ТЕМА Розробка програмного застосунку імпорту та аналізу даних ланцюжків мітохондріальної ДНК

Розробити програмний застосунок та, за його допомогою, провести аналіз даних ланцюжків мітохондріальної ДНК
Зміст текстової частини до курсової роботи:

Зміст
Анотація
Вступ
1 Аналіз ланцюжків мітохондріальної ДНК людей з різних географічних регіонів
2 Розробка структури бази даних для збереження поліморфізмів нуклеотидів та підготовка даних поліморфізмів нуклеотидів для подальшого аналізу
3 Аналіз імортованих поліморфізмів нуклеотидів
Висновки
Список використаної літератури

Дата видачі „___” листопада 2019 р.
Керівник (підпис)
Завдання отримав (підпис)

Тема: Розробка програмного застосунку імпорту та аналізу даних ланцюжків мітохондріальної ДНК

Календарний план виконання роботи:

№ п/п	Назва етапу курсової роботи	Термін виконання етапу	Примітка
1.	Отримання завдання на курсову роботу	09.10.2019	
2.	Огляд літератури за темою роботи	28.11.2019	
3.	Програмна реалізація парсування даних	14.01.2020	
4.	Розробка бази даних, збереження та аналіз поліморфізмів	13.02.2020	
5.	Розробка бази даних для збереження та аналізу послідовностей носіїв з різних регіонів	10.03.2020	
6.	Програмна реалізація аналізу	31.03.2020	
7.	Написання пояснювальної роботи	08.04.2020	
8.	Створення презентації для доповіді	23.04.2020	
9.	Аналіз отриманих результатів з науковим керівником	30.04.2020	
11.	Коригування роботи за результатами аналізу та остаточне оформлення	07.05.2020	
12.	Захист курсової роботи	18.05.2020	

Студентка Самовол М. С._____

Керівник Гулаєва Н. М. “_____”_____

Зміст

Анотація

Метою даної курсової роботи є аналіз існуючих поліморфізмів нуклеотидів та послідовностей ланцюжків мітохондріальної ДНК людей з різних географічних регіонів.

Розроблено базу даних для збереження послідовностей ланцюжків мітохондріальної ДНК людей що походять з визначених географічних регіонів. Визначені заздалегідь дані послідовностей було імпортовано в створену базу даних. Виконано аналіз послідовностей відносно різних географічних регіонів.

Отримані результати презентовано у вигляді таблиць у форматі .xlsx, що суттєво підвищує зручність їх використання та обробки.

Розроблено базу даних для збереження допустимих варіантів поліморфізмів нуклеотидів мітохондріальної ДНК. Попередньо підготовлені та провалідовані дані було імпортовано у розроблену базу даних. Проаналізовано можливі поліморфізми нуклеотидів (нейтральні та патогенні мутації).

Ключові слова: аналіз ланцюжків мітохондріальної ДНК, поліморфізми нуклеотидів, нейтральні та патогенні мутації, база даних.

Вступ

На сьогоднішній день, багато різних захворювань людського організму викликані мутаціями в послідовностях нуклеотидів ДНК мітохондрій. Дефекти мітохондріальних функцій вкрай рідко діагностуються своєчасно та викликають багато серйозних хвороб, у тому числі цукровий діабет, хворобу Паркінсона, хвороби серця, хворобу Альцгеймера і рак ([1]).

Серед основних особливостей мітохондріальної ДНК людини виділяють відсутність комбінаторної мінливості. Так як мітохондріальна ДНК успадковується у більшості випадків по материнській лінії, то рекомбінаційні події відсутні. Таким чином, послідовність нуклеотидів змінюється з наступним поколінням лише за рахунок послідовного накопичення мутацій ([1],[2]).

Сучасній медицині майже не відомо методів виправлення або компенсування цих мутацій та передбачення патологічних захворювань.

В той же час, високі темпи розвитку генетичної науки наближають перспективи нових відкриттів. Зростає кількість інформації, що стає доступною завдяки генетикам-науковцям. Вона потребує зручних методів обробки, які зможуть дозволити систематизувати, дослідити та використовувати отриману інформацію.

Результати даної роботи будуть корисними в процесі досліджень мутацій мітохондріальних ДНК, дозволять спростити доступ до структурованої інформації та суттєво зменшити час, витрачений на її обробку.

Враховуючи наведену вище інформацію, за *мету* даної роботи було поставлено аналіз допустимих мутацій в структурі послідовностей нуклеотидів мітохондріальної ДНК людського організму та порівняння ланцюжків послідовностей за різними географічними регіонами.

Актуальність теми:

Обрана проблематика відображає гострі проблеми сучасної генетики, яка розвиваючись швидкими темпами надає людству все більше інформації про структуру генома, його функціональність та можливі мутації. Для обробки таких великих інформаційних пластів необхідні сучасні зручні методи, які на сьогодні може запропонувати лише галузь інформаційних технологій, а саме розробка баз даних та аналіз структурованої інформації.

Відповідно до мети, були поставлені наступні *завдання*:

1) розробити структуру бази даних для збереження та аналізу ланцюжків мітохондріальної ДНК, імпортувати надані дані послідовностей ланцюжків у розроблену базу даних;
2) виконати детальний аналіз послідовностей носіїв з різних географічних регіонів;
3) розробити базу даних для зберігання допустимих поліморфізмів нуклеотидів, виконати парсування, валідацію та імпорт даних у розроблену базу даних;
4) провести аналіз допустимих поліморфізмів: порахувати кількість патогенних мутацій та кількість нейтральних мутацій у послідовностях нуклеотидів, визначити кількість можливих поліморфізмів.

В процесі розробки використовувати дані з загальнодоступної бази даних MitoMap ([3], [4], [5]) та дані послідовностей мітохондріальної ДНК носіїв з п’яти регіонів місцевості Сардинія та місцевості Середнього Сходу, що попередньо були надані науковим керівником.

Структура роботи:

Робота складається з трьох розділів.

У першому розділі розробляється база даних для збереження та аналізу ланцюжків мітохондріальних ДНК по різним географічних регіонах (Сардинії та Середнього Сходу). Надається опис структури бази даних та

розробки програмної реалізації детального аналізу послідовностей за різними регіонами. Послідовності нуклеотидів мітохондріальної ДНК порівнюються попарно між собою, визначається дикий тип серед заданого регіону, знаходяться відстані послідовностей до базової послідовності (мітохондріальної Єви ([6], [7], [8])). Результати аналізу наданих даних подаються у вигляді таблиць формату .xlsx та гістограм розподілів відстаней між послідовностями.

У другому розділі описуються вимоги до даних, що повинні необхідні для подальшого аналізу допустимих поліморфізмів нуклеотидів. На основі цих даних розробляється база даних. Надається опис отримання необхідних даних: виконання розбору HTML-сторінок (парсування). Описуються обрані технології. Проводиться підготовка даних до аналізу: валідація та імпорт отриманих даних після парсування.

У третьому розділі виконується аналіз допустимих поліморфізмів нуклеотидів у послідовності мітохондріальної ДНК.

РОЗДІЛ 1: Аналіз ланцюжків мітохондріальної ДНК людей з різних географічних регіонів

1.1. Аналіз предметної області

Мітохондрії являють собою критично значущу складову органел клітини людського організму: вони виконують важливу функцію метаболічного процесу, регулюючи клітинний ріст та окисно-відновлюваний стан клітин ([1], [9]).

Вирішальну роль мітохондрії грають у виробництві енергії за допомогою процесу перетворення молекул органічних речовин на клітинну енергію у вигляді аденозинтрифосорної кислоти (АТФ). Результатом окиснювання основних поживних продуктів (вуглеводів, білків та жирів) є АТФ як головне універсальне джерело енергії для прямого використання клітинами ([9]). В подальшому енергія трансформується у механічну (живлення для м’язових клітин) та біоелектричну енергію (живлення для нервових клітинах).

Кілька тисяч мітохондрій присутні в кожній клітині людського тіла. Клітини серцевих та скелетних м’язів і підшлункової залози містять найбільшу кількість мітохондрій, бо саме цим м’язам необхідна підвищена кількість енергії ([9]).

Мітохондріальна ДНК має 37 генів та існує незалежно від розташованої у ядрі клітини ДНК.

ДНК мітохондрії складається з кільцевої молекули, що містить у собі 16569 пар нуклеотидів ([9]).

Патологічні мутації мітохондріальної ДНК призводять до появи захворювань центральної нервової системи, скелетних та серцевих м'язів, нирки, печінки, ендокринних залоз ([1]).

Кожний із відомих синдромів, спричинених порушенням функціонування мітохондрій, визначається певною мутацією таких типів: заміни, видалення або вставки нуклеотидів у послідовності ([10]).

Існуючі чотири види нуклеотидів позначаються наступним чином:

1) Аденін (позначають літерою A);

2) Цитозін (позначають літерою C);

3) Гуанін (позначають літерою G);

4) Тимін (позначають літерою T).

Отже, у записі послідовності нуклеотидів мітохондріальної ДНК людини зустрічаються лише ці чотири літери, що позначають окремі нуклеотиди та їх комбінації.

Диким типом певного регіону називають послідовність нуклеотидів, значення яких найбільш розповсюджене серед усіх послідовностей мітохондріальної ДНК носіїв що походять з цього регіону. Тобто це значення, що зустрічаються найчастіше ([10]).

Популярна генетична наука пропонує сучасникам наступну версію походження мітохондріальної ДНК. За цією версією, мітохондріальну ДНК люди успадкували від єдиної прародички Мітохондріальної Єви, що мешкала близько двохсот тисяч років тому на території Африки. Хоча Єва була не єдиною пращуркою людей, інші мітохондріальні ДНК не збереглись ([6], [7]). Послідовність нуклеотидів мітохондріальної ДНК Єви прийнято називати базовою послідовністю.

1.2 Розробка структури бази даних

1.2.1 Вимоги до даних

Для дослідження та аналізу мітохондріальних ДНК носіїв з різних географічних регіонів було обрано частину всього ланцюжка нуклеотидів, що відрізняється найбільшою варіацією.

Ланцюжки послідовностей нуклеотидів для подальшої роботи було надано науковим керівником. Також була надана інформація щодо кількості носіїв кожної послідовності та регіонів й місцевостей походження носіїв (додаток А).

В процесі роботи було визначено, які саме дані мають зберігатись у базі даних:

1) джерело послідовності (версія послідовності): шестизначний запис, що однозначно позначає послідовність;

2) FASTA (формат запису ланцюжку послідовності нуклеотидів, що використовується у біоінформатиці ([11]));

3) розташування послідовності в молекулі;

4) довжина цієї послідовності;

5) кількість носіїв послідовності;

6) походження носіїв (регіони та місцевості);

7) дані про базову послідовність та дикий тип;

8) результати аналізу наданих даних (розподіли відносно базової послідовності та дикого типу, попарні розподіли, характеристики до розподілів).

Отримані послідовності належали особам, що походили з п'яти різних регіонів місцевості Сардинії та місцевості Середнього Сходу (розподіл по регіонам не надано). Походження носіїв однозначно визначається двома характеристиками: значенням регіону та значенням місцевості.

Складність задачі полягала в тому, що для деяких послідовностей не було достовірної інформації про походження їх носіїв. Наприклад, про послідовність з версією «J01415» відомо, що вона належала п’ятнадцяти особам, також відомо, що ці особи походили з Сардинії, але з саме якого регіону - невідомо (можливі варіанти регіонів: A, B, C, D).

Тому при подальшому аналізі, дані про носіїв таких послідовностей мають бути віднесені як до кожного з зазначених регіонів, так і до всієї місцевості Сардинія разом.

Аналіз даних проводиться як для кожного з регіонів окремо, так і до усіх регіонів та місцевостей разом.

1.2.2 Опис реляційної моделі

Відповідно до вимог до даних, була спроектована ER модель бази даних для збереження та аналізу ланцюжків нуклеотидів носіїв з різних регіонів (додаток Б).

ER модель містить 7 сутностей: Відомості, Послідовність, Особа, Регіон, Місцевість, Задача, Розподіли.

На основі створеної ER моделі була побудована реляційна модель до бази даних (додаток В).

Декілька зв'язків типу «багато-до-багатьох» зумовили створення додаткових реляцій зі зв’язком типу «багато-до-одного» з початковими реляціями у реляційній моделі. Всього реляційна модель містить 10 реляцій.

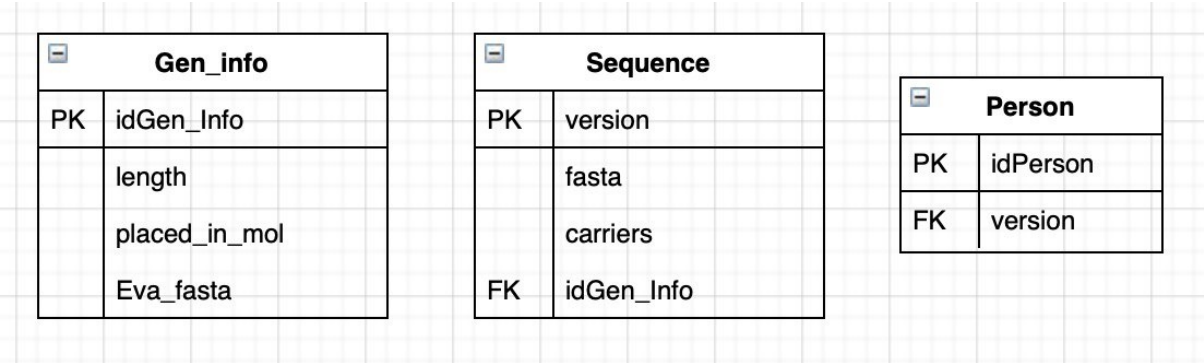


Рисунок 1.1 – Реляції «Gen_Info», «Sequence», «Person»

1) Реляція «Відомості» (Gen_Info)
2) Реляція «Послідовність» (Sequence)
3) Реляція «Особа» (Person)

У реляції «Відомості» наявні атрибути «id відомості» (idGen_Info), «довжина послідовності» (length), «розміщення в молекулі» (placed_in_mol), «базова послідовність» (Eva_fasta). Усі атрибути є обов'язковими. Ключем у реляції «Відомості» є атрибут «id відомості». У атрибуті «базова послідовність» зберігаються дані ланцюжка послідовності нуклеотидів мітохондріальної Єви.

У реляції «Послідовність» містяться атрибути «версія» (version), «FASTA», «кількість носіїв» (carriers). Усі атрибути у даній реляції є обов'язковими. Ключем у реляції «Послідовність» виступає атрибут «версія». Атрибут «id відомості» у реляції «Послідовність» має роль зовнішнього ключа та позначає зв'язок з реляцією «Відомості». Одна послідовність може містити лише одні відомості, але до одних й тих самих значень відомостей можуть відноситись декілька послідовностей. Тому таблиці «Відомості» та «Послідовність» пов'язані між собою зв'язком «один до багатьох».

Атрибут «id носія» (idPerson) з реляції «Особа» є ключем цієї реляції. Ця реляція пов'язана з реляцією «Послідовність» зв'язком «один до

багатьох», тому що у послідовності може бути декілька носіїв, але у одного носія можлива наявність лише однієї послідовності.

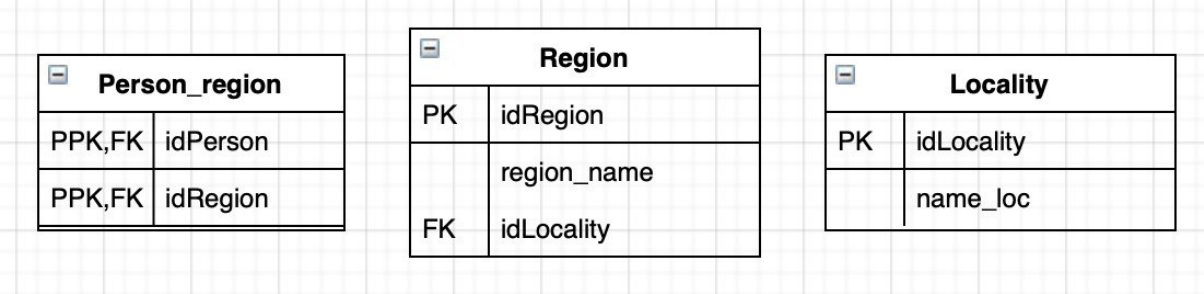


Рисунок 1.2 – Реляції «Person_region», «Region», «Locality»

4) Реляція «Регіон» (Region)

5) Реляція «Місцевість» (Locality)

6) Реляція «Особа_регіон» (Person_region)

У реляції «Регіон» ключем виступає атрибут «id регіону» (idRegion). Також наявний атрибут «назва регіону» (region_name).

У реляції «Місцевість» ключем виступає атрибут «id місцевості» (idLocality). Також наявний атрибут «назва місцевості» (name_loc).

Згідно вимогам до даних, у місцевості знаходяться декілька регіонів. Дивлячись на те, що один регіон можна віднести лише до однієї місцевості, зв’язок між реляціями «Місцевість» та «Регіон» визначається як «один до багатьох».

Реляція «Особа_регіон» з’являється після перетворення ER моделі в реляційну модель. Реляції «Регіон» та «Особа» пов’язані зв’язком «багато до багатьох». З одного регіону може походити декілька носіїв послідовності та одного носія ми можемо одночасно віднести до декількох регіонів (згідно вимогам до даних). В утвореній реляції наявні два атрибута, кожен з яких становить частину ключа реляції та виступає в ролі зовнішнього ключа: «id носія» (idPerson) показує зв’язок з реляцією «Особа», а «id регіону» (idRegion) - з реляцією «Регіон».

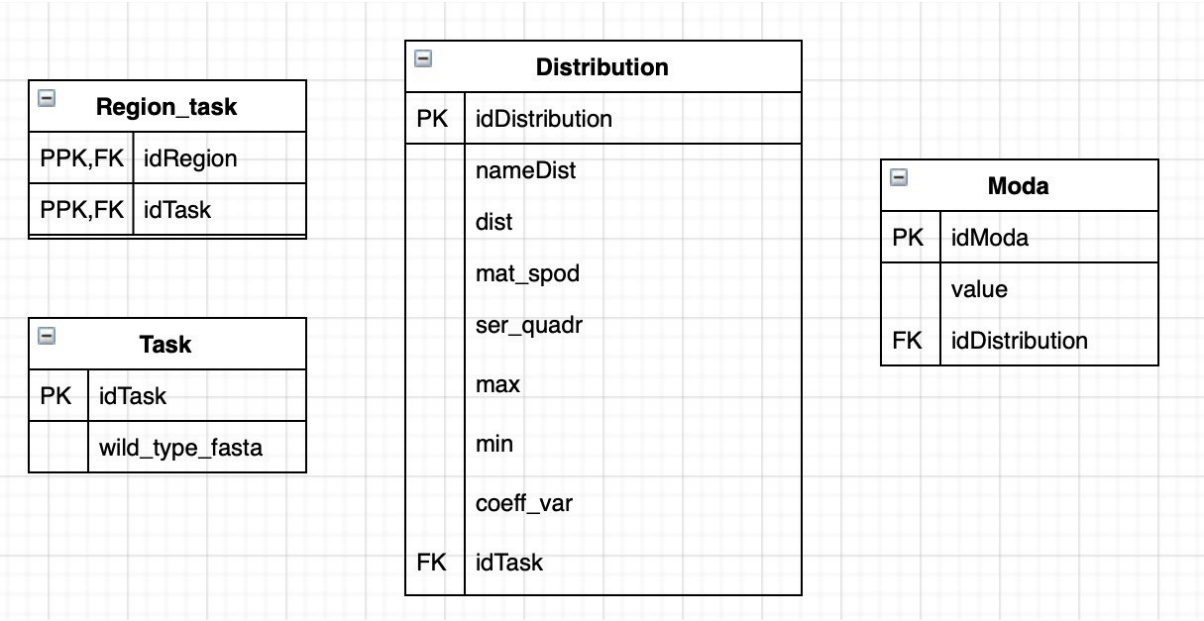


Рисунок 1.3 – Реляції «Region_task», «Task», «Distribution», «Moda»

7) Реляція «Задача» (Task)

8) Реляція «Розподіли» (Distribution)

9) Реляція «Мода» (Moda)

10) Реляція «Регіон_задача» (Region_Task)

Реляція «Задача» містить дані про дикий тип для поставленої задачі (атрибут «дикий тип» (wild_type_fasta)). Ключем цієї реляції виступає атрибут «id Задачі» (idTask).

Реляції «Регіон» та «Задача» пов’язані між собою зв’язком «багато до багатьох», так як одна задача може бути поставлена до сполучення декількох регіонів одночасно, та в одного регіону може бути декілька задач (окрема, та в сукупності з іншими регіонами). Саме це при перетворенні ER моделі у реляційну модель спричинило створення нової реляції «Регіон_задача». В утвореній реляції «Регіон_задача» присутні два атрибута, кожен з яких становить частину ключа реляції та виступає в ролі зовнішнього ключа: «id задачі» (idTask) показує зв’язок з реляцією «Задача», а «id регіону» (idRegion) - з реляцією «Регіон».

У реляції «Розподіли» міститься загальна інформація про кожен з можливих розподілів послідовностей та різні характеристики цих розподілів, що позначаються атрибутами «математичне сподівання» (mat_spod), «середньо квадратичне відхилення» (ser_quadr), «мінімальне значення» (min), «максимальне значення» (max), «коєфіцієнт варіації» (coeff_var). Ключем даної реляції виступає атрибут «id розподілу» (idDistribution). Також присутні атрибути «назва розподілу» (nameDist) та «значення розподілу» (dist). Дана реляція пов’язана з реляцією «Задача» зв’язком типу «один до багатьох», так як один розподіл можна віднести лише до однієї задачі, але у задачі, згідно зазначеним вимогам до даних, може бути декілька різних розподілів. Цей зв’язок позначає зовнішній ключ «id задачі» (idTask).

Реляція «Мода» утворена після перетворення ER моделі на реляційну модель з багатозначного атрибуту «мода» реляції «Розподіли». У одного розподілу може існувати декілька мод (найпоширеніших значень розподілу), тому зв’язок типу «один до багатьох» з реляцією «Розподіли» зумовив появу зовнішнього ключа «id розподілу» (idDistribution). Крім того, у реляції присутні атрибути «значення» (value) та «id моди» (idModa), що також виступає ключем даної реляції.

1.2.3 Вибір цільової СКБД

Для збереження необхідних даних було використано систему керування базами даних (СКБД) MySQL. Основні переваги мого вибору СКБД:

1) MySQL достатньо проста та зручна у використанні;
2) існує досить багато доступних ресурсів для опанування;
3) наявна детальна документація;
4) вільний доступ відкритий для некомерційного використання.

1.3 Програмний застосунок

1.3.1 Використані технології

Імпорт та аналіз даних у базу даних було реалізовано за допомогою мови програмування Java. Вибір було обумовлено перевагами цієї мови.

Основні переваги Java над іншими мовами програмування:

1) велика кількість користувачів, а тому наявні якісні документації та багато матеріалів щодо роботи з даною мовою програмування;

2) зручність у підключенні до бази даних;

3) сучасний та зрозумілий синтакис мови, що дозволяє не витрачати час на пошуки шляхів програмної реалізації, а безпосередньо виконувати поставлені завдання.

Для імпорту отриманих результатів аналізу було використано бібліотеку Apache POI**.** Ця бібліотека дозволяє створювати, змінювати та відображати файли Microsoft Office за допомогою програм Java ([12]). Оперуючи бібліотекою Apache POI, були експортовані попередньо згенеровані дані у файли з розширенням xlsx.

1.3.2 Структура програми

Архітектура програмного застосунку складається з наступних класів:

1) Main;
2) DataBase;
3) Sequence;
4) DistSeq;
5) Pairs;
6) Tasks;
7) toExcel.

У класі Main, що є вхідною точкою програми, відбувається імпортування наданих даних, у файлі з розширенням csv, у необхідну для подальшої роботи структуру даних Sequence. Також у цьому класі знаходяться виклики функцій допоміжних класів (додавання інформації до бази даних, передачу згенерованих розподілів у клас toExcel для подальшого їх аналізу та виведення у файл).

Клас DataBase відповідає за взаємодію з базою даних. Методи доступні в даному класі дозволяють під’єднатися до бази даних, заносити туди необхідні дані, та також обирати інформацію з бази даних щодо різних регіонів та базової послідовності.

У класі Sequence реалізовано структуру даних, необхідну для зручності імпортування даних з наданого файлу, в форматі csv, у розроблену базу даних. Цей клас містить значення версії та FASTA ([11]).

Клас DistSeq допомагає у визначенні відстаней, а саме: за допомогою цієї структури даних реалізовано отримання даних для подальшого знаходження відстаней до базової послідовності та послідовності дикого типу. У цьому класі містяться значення версії, послідовності, кількість носіїв та значення відстані до базової послідовності або дикого типу.

У класі Pairs реалізовано структуру даних для знаходження попарної відстані між послідовностями нуклеотидів та подальшої роботи над ними. Цей клас містить інформацію про значення версії, FASTA, кількості носіїв та дані про відстані до інших послідовностей.

Клас Task містить методи для знаходження відстаней послідовностей певного регіону до базової послідовності, дикого типу та попарних відстаней. На основі цих даних реалізовано метод для визначення розподілів щодо відстаней та визначення їх характеристик. У даному класі також реалізовано метод розрахування дикого типу на основі наданих послідовностей з певного регіону. Також наявні методи для розрахування характеристичних значень визначених розподілів: математичних сподівань,

середньоквадратичних відхилень, максимальних та мінімальних значень розподілу, можливих мод та коефіцієнтів варіації значень даного розподілу.

У класі toExcel реалізовано методи необхідні для експортування отриманих після аналізу даних у файл з розширенням xlsx. Методи наявні у цьому класі дозволяють створити таблиці у форматі xlsx зі значеннями попарних відстаней між послідовностями, зі значеннями відстаней заданих послідовностей до дикого типу та базової послідовності, та з інформацією щодо характеристик розподілів.

1.3.3 Аналітичні методи

Попередньо отримані дані було імпортовано до розробленої бази даних. Для подальшого аналізу необхідно було отримати дані за допомогою запитів SQL з бази даних за певними географічними регіонами.

```
SELECT  Region_GenBank.Sequence.version, fasta, COUNT(DISTINCT Person.idPerson)
FROM ((Region_GenBank.Sequence INNER JOIN Region_GenBank.Person ON Sequence.version = Person.version)
INNER JOIN Region_GenBank.Person_Region ON Person.idPerson = Person_region.idPerson)
INNER JOIN Region_GenBank.Region ON Person_region.idRegion = Region.idRegion
WHERE region_name = "MiddleEast"
GROUP BY Region_GenBank.Sequence.version;
```

Рисунок 1.4 – Запит мовою SQL для отримання інформації по регіону “MiddleEast” (Середній Схід)

На основі отриманих послідовностей визначається дикий тип, властивий заданому регіону: для кожної позиції в послідовності вираховується значення нуклеотиду, що зустрічається у вибірці регіону найчастіше.

Відстані до базової послідовності та дикого типу вираховуються наступним чином: для кожної позиції в послідовності порівнюється

значення нуклеотиду зі значенням нуклеотиду на тотожній позиції у базовій послідовності або дикому типі.

```
public ArrayList<DistSeq> distWild(ArrayList<DistSeq> sequence, String w) {

    ArrayList<Part2.DistSeq> dist = new ArrayList<Part2.DistSeq>();
     char [] wild = w.toCharArray();

    int counter = 0;
    for (DistSeq entry : sequence) {
        for (int i = 0; i < wild.length; i++) {
            if (wild[i] != entry.fasta[i]) {
                counter++;
            }
        }
        DistSeq thisone = createDistSeq(entry.version, entry.fasta, entry.carriers, counter);
        dist.add(thisone);
        counter = 0;
    }

    return dist;
}
```

Рисунок 1.5 – Метод для обчислення відстаней послідовностей до дикого типу

Попарні відстані знаходяться за допомогою парного порівняння нуклеотидів у кожній послідовності з заданої вибірки.

Розподіли відносно базової послідовності, дикого типу та парного порівняння розраховуються на основі отриманих відповідних значень відстаней. Для кожного значення відстані розраховується кількість його отримання. Базуючись на отриманих результатах розподілів відносно відстаней, розраховуються розподіли ймовірностей.

```
public ArrayList<Integer> rozpodilWild(ArrayList<DistSeq> sequence){

    ArrayList<Integer> rozp = new ArrayList<Integer>();

    int sizeRegion = 1;
    for(DistSeq entry: sequence){
        for(int k=0; k<entry.carriers;k++) {
            sizeRegion++;
        }
    }
    int counter =0;
    for(int i = 0; i<sizeRegion; i++){
        for(DistSeq entry: sequence){
            if(i == entry.distance){
                for(int k=0; k<entry.carriers;k++){
                    counter++;
                }
            }
        }
        rozp.add(counter);
        counter = 0;
    }
    return rozp;
}
```

Рисунок 1.6 – Метод для обчислення розподілу відносно відстаней послідовностей до дикого типу

Отримані дані після розрахунку розподілів використовуються для подальшого визначення характеристик цих розподілів:

1) Обчислення значення математичного сподівання (середнього значення) за наступною формулою:

$$m_i = \sum_{i=1}^{l} i * p_i \,, \qquad (1.1)$$

де i – значення змінної розподілу;

p_i – ймовірність змінної розподілу.

Відповідно до формули, було створено програмний код застосування методу вираховування математичного сподівання (рис. 1.7).

```
public double matSpodivW(ArrayList<Double> chastka){

    double matspod =0.0;

    for(int i =0; i<chastka.size();i++){
        matspod += (double) i *chastka.get(i);
    }

    return matspod;
}
```

Рисунок 1.7 – Метод для обчислення математичного сподівання для заданого розподілу

2) Обчислення значення середньоквадратичного відхилення за допомогою наступної формули:

$$\sigma_i = \sqrt{\sum_{i=0}^{l} (i - m_i)^2 * p_{i,}} \qquad (1.2)$$

Де i – значення змінної розподілу;

m_i – значення математичного сподівання;

p_i – ймовірність змінної розподілу.

Відповідно до формули, було створено програмний код застосування методу вираховування середньоквадратичного відхилення (рис. 1.8).

```
public double serQuadraticW(double matspodiv, ArrayList<Double> chastka){

    double quadratic =0.0;
    double helpq =0.0;

    for(int i =0; i<chastka.size();i++){
        helpq += ((double) i - matspodiv) * ((double) i - matspodiv) * chastka.get(i);
    }

    quadratic = Math.sqrt(helpq);
    return quadratic;

}
```

Рисунок 1.8 – Метод для обчислення середньоквадратичного відхилення для заданого розподілу

3) Обчислення коефіцієнту варіації за наступною формулою:

$$\sigma_i / m_i, \tag{1.3}$$

де m_i – значення математичного сподівання розподілу;

σ_i – значення середньоквадратичного відхилення розподілу.

4) Знаходження максимального (рис 1.9) та мінімального (рис 1.10) значень.

```
public int maxW(ArrayList<Integer> rozp){

    int high = rozp.get(0);
    for (Integer integer : rozp) {
        if (high < integer) {
            high = integer;
        }
    }
    return high;
}
```

Рисунок 1.9 – Метод для обчислення максимального значення для заданого розподілу

```
public int minW(ArrayList<Integer> rozp){

    int low = rozp.get(0);
    for (Integer integer : rozp) {
        if (low > integer) {
            low = integer;
        }
    }
    return low;
}
```

Рисунок 1.10 – Метод для обчислення мінімального значення для заданого розподілу

5) Обчислення мод розподілу за допомогою наступної формули:

$$m = \{i: \max p_i\}, \tag{1.4}$$

де i – значення змінної розподілу;

p_i – ймовірність змінної розподілу.

Відповідно до формули, було створено програмний код застосування методу вираховування можливих мод (рис.1.11).

```
public ArrayList<Integer> modaW(int max, ArrayList<Integer> rozp){

    ArrayList<Integer> moda = new ArrayList<Integer>();

    for(int i =0; i<rozp.size(); i++){
        if(max == rozp.get(i)){
            moda.add(i);
        }
    }
    return moda;

}
```

Рисунок 1.11 – Метод для обчислення мод для заданого розподілу

1.3.4 Представлення результатів реалізованих методів

За допомогою бібліотеки Apache POI було реалізовано методи для експортування отриманих даних в процесі аналізу у таблиці формату xlsx.

Нижче наведено приклади створених таблиць для вибірки послідовностей з місцевості Середнього Сходу.

Version	Distance
M58059	4
M58064	1
M58103	3
M58104	6
M58104	6
M58105	5
M58106	6
M58107	3
M58108	5
M58109	4
M58110	5
M58110	5
M58111	10

Рисунок 1.12 – Приклад результату знаходження відстаней послідовностей носіїв з місцевості Середнього Сходу до дикого типу

У зразку таблиці, наведеної на рисунку 1.12, зазначено версії послідовності ланцюжків мітохондріальних ДНК людини та відстань цієї послідовності до послідовності дикого типу місцевості Середнього Сходу.

	M58059	M58064	M58103	M58104	M58104	M58105	M58106	M58107	M58108	M58109	M58110	M58110	M58111	M58112	M58113
M58059	0	5	7	8	8	9	8	7	9	8	9	9	12	9	6
M58064	5	0	4	7	7	6	7	4	6	5	4	4	9	6	3
M58103	7	4	0	7	7	6	7	6	8	5	8	8	11	8	5
M58104	8	7	7	0	0	9	4	7	7	10	11	11	16	9	6
M58104	8	7	7	0	0	9	4	7	7	10	11	11	16	9	6
M58105	9	6	6	9	9	0	9	6	8	7	10	10	13	6	5
M58106	8	7	7	4	4	9	0	5	9	10	11	11	16	9	6
M58107	7	4	6	7	7	6	5	0	6	7	8	8	13	6	3
M58108	9	6	8	7	7	8	9	6	0	9	10	10	15	8	5
M58109	8	5	5	10	10	7	10	7	9	0	7	7	12	9	6
M58110	9	4	8	11	11	10	11	8	10	7	0	0	11	10	5
M58110	9	4	8	11	11	10	11	8	10	7	0	0	11	10	5
M58111	12	9	11	16	16	13	16	13	15	12	11	11	0	15	12
M58112	9	6	8	9	9	6	9	6	8	9	10	10	15	0	5
M58113	6	3	5	6	6	5	6	3	5	6	5	5	12	5	0

Рисунок 1.13 – Результат знаходження попарних відстаней послідовностей носіїв з місцевості Середнього Сходу

У зразку таблиці, наведеної на рисунку 1.13, зазначено версії послідовності ланцюжків мітохондріальних ДНК та попарні відстані цих послідовностей, що належать особам з Середнього Сходу.

Відстань	0	1	2	3	4	5	6	7	8	9	10
Розподіл відносно базової	0	4	4	14	6	9	5	3	1	0	1
Розподіл відносно базової (частка	0.0	0.085	0.085	0.298	0.128	0.191	0.106	0.064	0.021	0.0	0.021
	Мат спод	Серед ква відхиленн	Моди	Мін.	Макс.	Коеф вар					
	4.085	1.922	3	0	14	0.471					
Розподіл відносно дикого_т	0	4	4	14	6	9	5	3	1	0	1
Розподіл відносно дикого_т (частка	0.0	0.085	0.085	0.298	0.128	0.191	0.106	0.064	0.021	0.0	0.021
	Мат спод	Серед кв відхиленн	Моди	Мін.	Макс.	Коеф вар					
	4.085	1.922	3	0	14	0.471					
Розподіл відносно попарних	2	9	28	41	102	115	177	154	164	115	67
Розподіл відносно попарних (частка	0.002	0.008	0.026	0.038	0.094	0.106	0.164	0.143	0.152	0.106	0.062
	Мат спод	Серед ква відхиленн	Моди	Мін.	Макс.	Коеф вар					
	7.057	2.664	6	0	177	0.377					
	Послідов дикого_т	ttctttcatggggaagcagatttgggtaccacccaagtattgactcacccatcaacaaccgctatgtatttcgtacattactgccagccaccatgaatatt									

Рисунок 1.14 – Приклад результату аналізу послідовностей носіїв з місцевості Середнього Сходу

Таблиця на рисунку 1.14 відображає приклад результату аналізу, отриманого в наслідок знаходження відстаней та розподілів, характеристик до них.

За допомогою зазначених методів було проведено аналіз послідовностей носіїв, що походять з різних географічних регіонів. Проаналізовано вибірки наступних регіонів:

1) усі наявні послідовності;

2) регіон A місцевості Сардинія;

3) регіон B місцевості Сардинія;

4) регіон C місцевості Сардинія;

5) регіон D місцевості Сардинія;

6) регіон E місцевості Сардинія;

7) місцевості Сардинія (включаючи усі її регіони);

8) місцевості Середній Схід.

Після отримання документів з розширенням xlsx для кожного регіону за результатами проаналізованих методів, було створено гістограми, що характеризують розподіли відстаней послідовностей кожної вибірки: щодо розподілу відстаней послідовностей до базової послідовності, щодо розподілу відстаней послідовностей до дикого типу, щодо розподілу попарних відстаней послідовностей.

Отримані в результаті аналізу гістограми представлено в додатку Г.

1.4 Висновки за розділом 1

У даному розділі було проаналізовано предметну область та розглянуто основні поняття, що надалі використовуються в роботі.

Також були сформульовані основні вимоги до даних, які мають зберігатись у базі даних для розробки та аналізу послідовностей мітохондріальних ДНК осіб, що походять з різних географічних регіонів.

Відповідно до сформульованих вимог, розроблено структуру бази даних. Надано опис обраних технологій для розробки програмного застосунку та наведено архітектуру програмного застосунку. Крім того, наведено опис аналітичних методів, представлених у програмній частині роботи.

Результати аналізу послідовностей ланцюжків мітохондріальних ДНК осіб, що походять з різних географічних регіонів представлено у вигляді гістограм.

РОЗДІЛ 2: Розробка структури бази даних для збереження поліморфізмів нуклеотидів та їх підготовка для подальшого аналізу

2.1 Розробка бази даних для допустимих поліморфізмів мітохондріальної ДНК

2.1.1. Специфікація вимог до даних

Деякі мутації нуклеотидів можуть мати кілька варіантів допустимих замін, що науково визначаються як поліморфізми.

Точкові заміни нуклеотидів являють собою заміну одного нуклеотиду іншим. Нуклеотиди з різних позицій можуть мати одну, дві чи три варіанти точкових замін ([10]).

У базі даних має зберігатись наступна інформація про нуклеотиди:

а) локус (позиція нуклеотиду);

б) всі можливі варіанти точкових замін нуклеотидів у послідовності;

в) інформація чи викликає дана точкова заміна патології, що спричиняють хвороби.

Дані для аналізу залучено з загальнодоступного онлайн ресурсу MitoMap ([3], [4], [5]). Інформацію про допустимі мутації надано у вигляді трьох таблиць на сторінках сайту: дві таблиці з нейтральними мутаціями (загалом 16 тисяч рядків даних) та одна таблиця з даними про патогенні мутації (майже 500 рядків). Таблиці розміщено на веб сайті, отже всі дані представлено у форматі HTML, який потребує подальшої обробки та підготовки даних до імпортування в базу даних.

У кожній таблиці міститься від шести до одинадцяти колонок з різними даними, але важливою інформацією для аналізу є дані лише двох колонок: Position та Nucleotide Change. Position містить значення позиції локусу нуклеотиду, а у Nucleotide Change представлено зафіксовані варіанти замін цього нуклеотиду.

Для подальшого збереження у базі даних та аналізу поліморфізмів, важливими є лише точкові заміни нуклеотидів, тому мутації видалення та вставок не приймаються до уваги. Для однієї позиції нуклеотиду допускається можливість декількох (до трьох) варіантів точкових замін.

2.1.2 Проектування бази даних

Згідно вимогам до даних, була спроектована ER модель бази даних для збереження допустимих поліморфізмів нуклеотидів (рис 2.1).

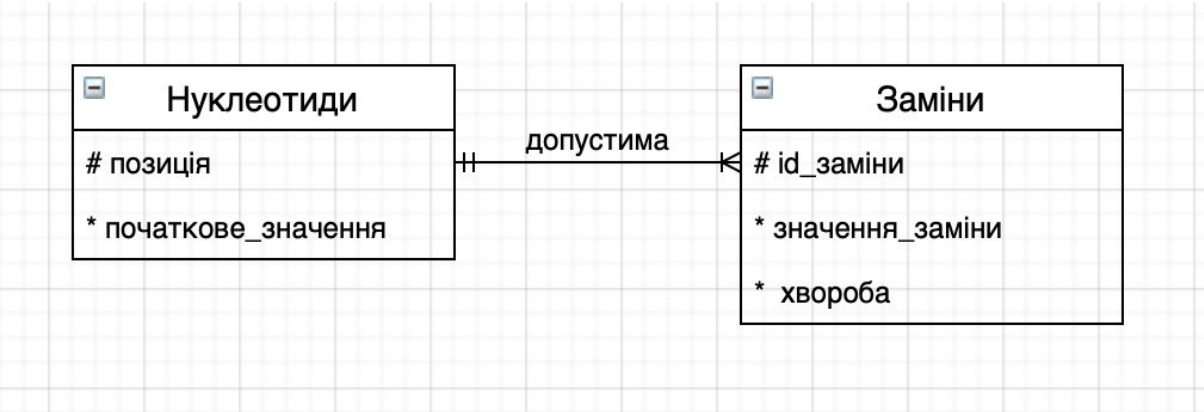


Рисунок 2.1 – ER модель бази даних для збереження допустимих поліморфізмів

На основі створеної ER моделі була побудована реляційна модель до бази даних (рис 2.2).

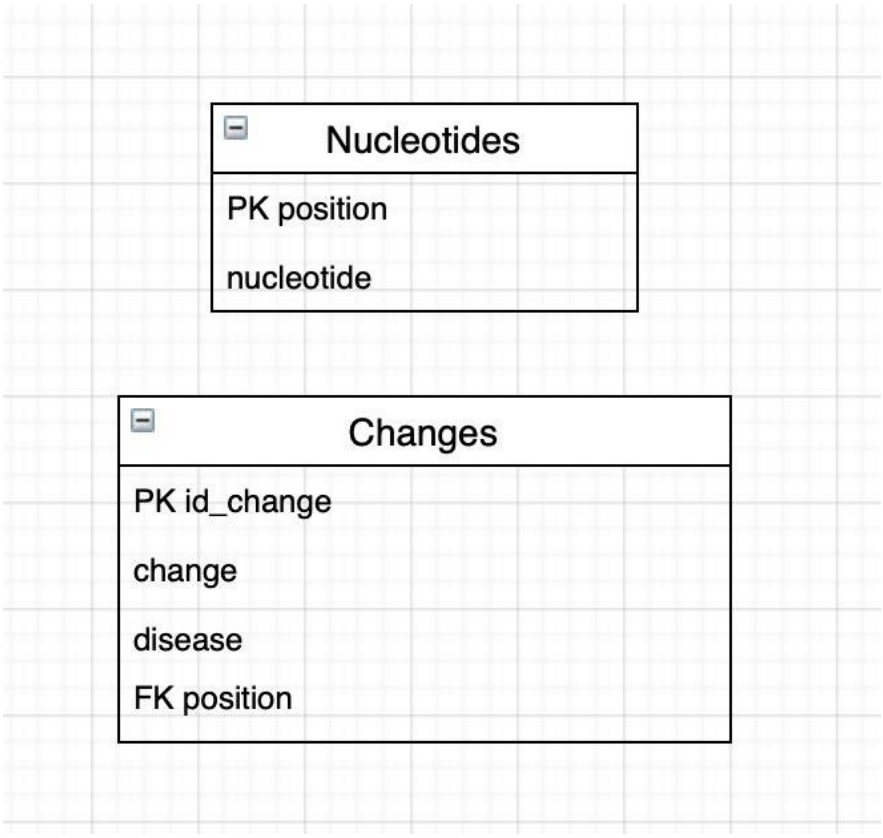


Рисунок 2.2 – Реляційна модель бази даних для збереження допустимих поліморфізмів

Реляційна модель містить 2 реляції:

1) реляція «Нуклеотиди» (Nucleotides);
2) реляція «Заміни» (Changes).

У реляції «Нуклеотиди» наявні атрибути «позиція» (position) та «початкове значення» (nucleotide). Обидва атрибути є обов'язковими. Ключем у реляції «Нуклеотиди» є атрибут «позиція».

У реляції «Заміни» містяться атрибути «id заміни» (id_change), «позиція» (position), «значення заміни» (change) та «хвороба» (disease). Атрибут «хвороба» позначає наявність патологій, пов'язаних з даною заміною нуклеотиду. Усі атрибути у реляції є обов'язковими. Ключем у реляції «Заміни» є атрибут «id заміни».

У одного нуклеотиду може бути до трьох різних варіантів точкових замін, тому таблиці «Нуклеотиди» та «Заміни» пов'язані між собою

зв’язком типу «один до багатьох». Атрибут позиція у реляції «Заміни» виступає в ролі зовнішнього ключа.

Для збереження необхідних даних було використано систему керування базами даних MySQL.

2.2 Розробка програмного застосунку для парсування, валідації та імпорту даних

2.2.1 Технологічні засоби для розробки програмного застосунку

Під терміном «парсування» мається на увазі розбір HTML документу за наявними у ньому тегами.

Парсування та імпорт у базу даних було реалізовано за допомогою мови програмування Java та з використанням бібліотеки JSoup.

Бібліотека JSoup призначена для розбору HTML-сторінок та дозволяє розпарсувати необхідні дані з формату HTML для подальшої маніпуляції з ними ([13]).

Підтримуючи основні специфікації HTML5, бібліотека JSoup орієнтована на гнучкість і простоту використання.

2.2.2 Архітектура програмного застосунку

Архітектура програмного застосунку складається з наступних класів:

1) Main;
2) Parser;
3) Nucleotide;
4) DBConnect.

Клас Main є вхідною точкою програми. В цьому класі відбуваються виклики функцій допоміжних класів, зчитування файлу в форматі HTML та його передачу в клас Parser.

Клас Parser відповідає за розбір наданого HTML документу та валідацію даних у ньому перед занесенням інформації до бази даних.

Клас Nucleotide містить значення позиції та значення заміни нуклеотиду.

Клас DBConnect відповідає за взаємодію з базою даних. Методи доступні в даному класі дозволяють під'єднуватись до бази даних, заносити дані про значення позицій та замін нуклеотидів, позначати, які саме заміни викликають захворювання.

2.2.3 Опис процесу парсування

Дані про мутації нуклеотидів, що були надані у таблицях на трьох HTML-сторінках, потребували підготовки, парсування та валідації для подальшого імпорту в розроблену базу даних.

Щоб зменшити обсяги даних та оптимізувати швидкість роботи, було прийнято рішення відкинути інформацію з інших колонок після збереження HTML коду повних таблиць ще до парсування.

Під час процесу парсування дані з першої колонки, що містили значення позицій нуклеотидів у послідовності, та дані з другої отриманої колонки, що містили початкове значення нуклеотиду та значення його заміни, заносились до створеної структури даних Nucleotide.

```
int position = Integer.parseInt(row.getElementsByTag( tagName: "td").first().text());
String chan = row.getElementsByTag( tagName: "td").next().text();
```

Рисунок 2.3 – Приклад функції парсування та занесення даних до значень структури даних

2.2.4 Опис процесу валідації та імпорту даних

За визначеними вимогами, до уваги не брались видалення та вставки, тому після парсування, дані потребували проходження перевірки.

Перешкодою стала невелика потужність комп’ютера в оброблені великої кількості даних. Тому весь обсяг отриманих після відкидання даних було поділено на кілька частин, щоб поступово імпортувати дані, та таким чином зменшити навантаження на оперативну пам’ять та процесор комп‘ютера.

До розробленої бази даних заносились тільки дані про мутації, значення яких замінялось на значення одного нуклеотиду (точкові заміни), тобто дані, що пройшли валідацію.

```
if (chan.contains("-")){
    String validatingChan = "-".split(chan)[1];
    String validatingChan2 = "-".split(chan)[0];
    if (!validatingChan.equals("del") && validatingChan.length()<=1 && validatingChan2.length()<=1) {
        nucls.add(new Nucleotid(position,chan));
    }
}
```

Рисунок 2.4 – Метод валідації даних для подальшого імпорту в базу даних

```
//заносимо зміни та позиції
public void insCh(LinkedHashSet<Nucleotid> nucleotidsSet) {

    try {
        for (Nucleotid currentNucl : nucleotidsSet) {
            stmt1 = con.createStatement();

            String query2 = " INSERT into Changes (pos, chan)"
                    + " values (?, ?)";

            PreparedStatement prepStmt = con.prepareStatement(query2);
            prepStmt.setInt( parameterIndex: 1, currentNucl.pos);
            prepStmt.setString( parameterIndex: 2, "-".split(currentNucl.chan)[1]);

            prepStmt.execute();
        }
    } catch (SQLException sqlEx) {
        sqlEx.printStackTrace();
    } finally {
        try {
            con.close();
        } catch (SQLException ignored) { }
    }
}
```

Рисунок 2.5 – Метод для імпорту провалідованих даних до бази даних

На рисунку 2.5 зображено приклад методу імпортування даних до бази даних (занесення значення позицій та можливих замін).

2.3 Висновки до розділу 2

У даному розділі сформульовано основні вимоги до даних, що мають зберігатись у базі даних для подальшого аналізу можливих поліморфізмів.

Також наведено опис структури бази даних та програмного застосунку. Крім того, представлено опис процесів парсування, валідації та імпорту позицій та значень точкових замін у розроблену базу даних.

РОЗДІЛ 3. Аналіз допустимих поліморфізмів нуклеотидів

3.1 Результати аналізу поліморфізмів

Після занесення усієї необхідної інформації до бази даних, було підраховано кількість допустимих нейтральних точкових замін за допомогою запитів мовою SQL (рис. 3.1).

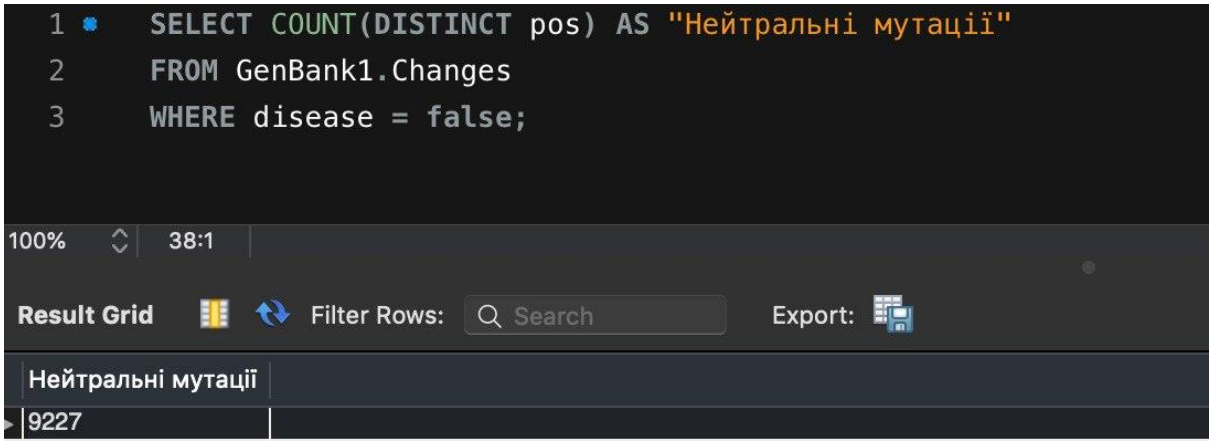


Рисунок 3.1 – Запит SQL для знаходження кількості можливих нейтральних мутацій, спричинених точковими замінами

Також, за допомогою запитів мовою SQL було отримано кількість патогенних мутації, що були спричинені точковими замінами (рис.3.2).

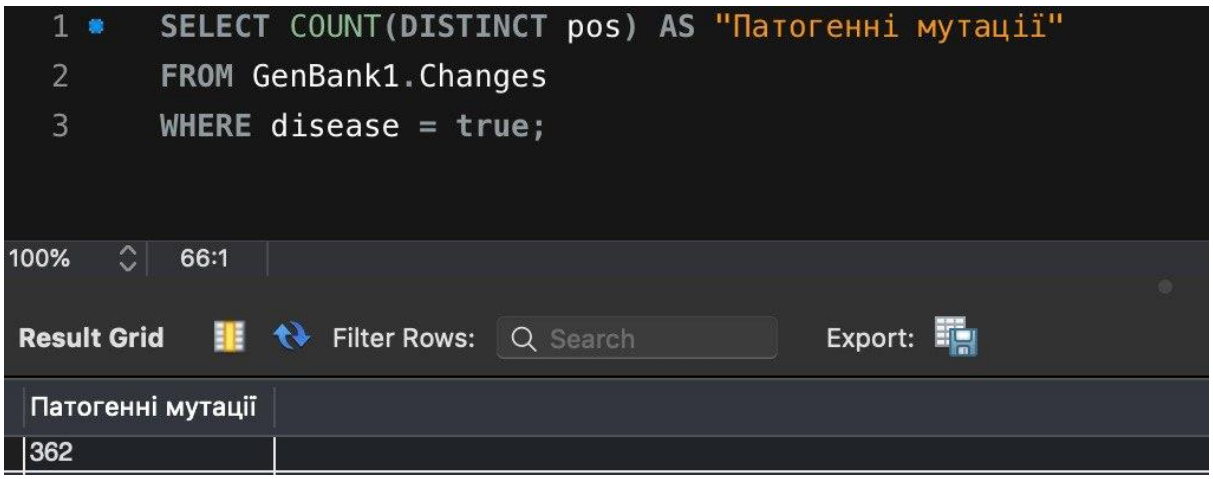


Рисунок 3.2 – Запит SQL для знаходження кількості можливих патогенних мутацій, спричинених точковими замінами

Отримано наступні результати:

а) загальна кількість нейтральних точкових замін становить 9227 різних варіантів;

б) кількість патологічних точкових замін становить 362 варіанта.

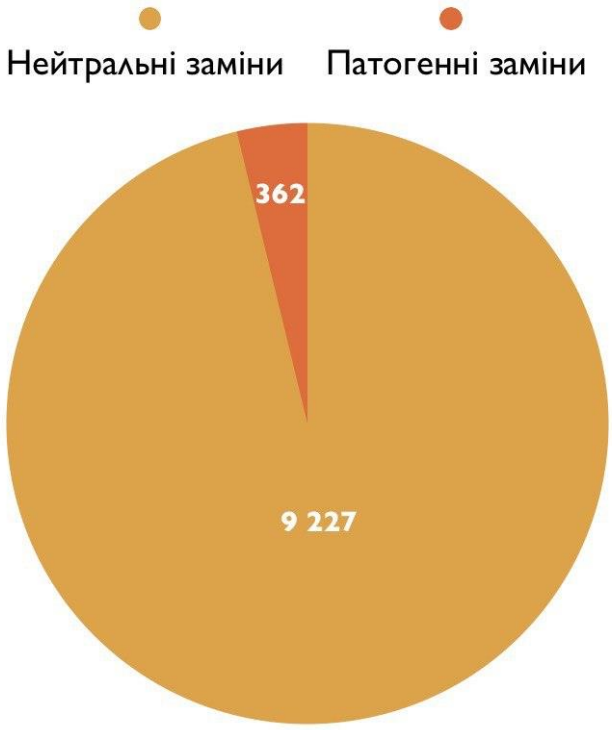


Рисунок 3.2 – Діаграма відношення кількості нейтральних замін до кількості патогенних замін

Також двома різними запитами мовою SQL було обрано інформацію про допустимі поліморфізми нейтральних та патологічних мутацій.

Результати запитів було збережено у файлах з розширенням csv. Приклади отриманих результатів наведено на рисунках 3.3 та 3.4.

Neutral_mutations

Позиція	Кількість поліморфізмів
3	1
5	2
6	3
7	1
8	3
9	2
10	2
11	2
12	2
13	2
14	2
16	3

Рисунок 3.3 – Приклад отриманого файлу з інформацією про кількість допустимих нейтральних точкових замін відносно кожної позиції послідовності

Patological_mutations

Position	Number_of_polymorphisms
114	1
146	1
150	1
195	1
499	1
547	1
3308	2
3310	1
3316	1
3335	1
3336	1
3337	1

Рисунок 3.4 – Приклад отриманого файлу з інформацією про кількість допустимих патологічних точкових замін відносно кожної позиції послідовності

Аналіз допустимих поліморфізмів серед зафіксованих точкових замін показав наступні результати:

Серед нейтральних та патогенних мутацій переважна кількість замін нуклеотидів мала лише один варіант: 6663 точкові заміни серед нейтральних, що становить 72,21% від загальної кількості зафіксованих нейтральних мутацій, та 352 точкові заміни серед патогенних, що становить 97,24% від загальної кількості мутацій з патогенами.

Поліморфізми з двома варіантами замін склали 1948 випадків нейтральних мутацій та 10 випадків патогенних мутацій, що становить 21,11% та 2,76% від загальної кількості відповідно.

Серед патогенних мутацій не було виявлено поліморфізмів з трьома варіантами точкових замін. В той час як серед нейтральних було зафіксовано 616 варіантів таких мутацій що складає 6,68% від загальної кількості точкових замін, що спричинили нейтральні мутації.

Отримані результати візуально представлені на рисунках 3.3 та 3.4.

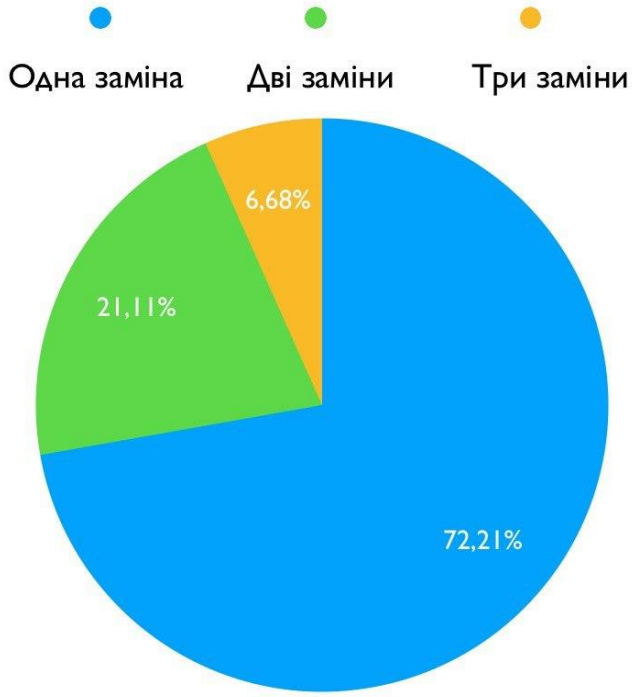


Рисунок 3.5 – Діаграма відношення варіантів поліморфізмів серед нейтральних мутацій

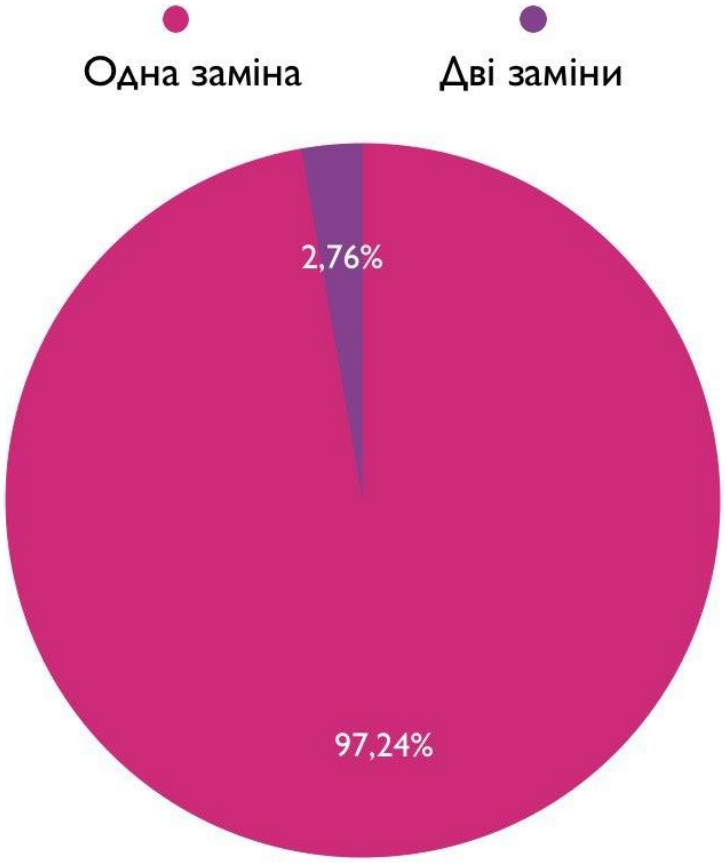


Рисунок 3.6 – Діаграма відношення варіантів поліморфізмів серед патогенних мутацій

3.2 Висновки до розділу 3

У даному розділі було підраховано кількість можливих точкових замін, що спричинили нейтральні та патогенні мутації, та кількість поліморфізмів точкових замін (також у відсотковому співвідношенні).

Надано діаграми відношення варіантів поліморфізмів серед нейтральних та патогенних мутацій для наглядної візуалізації.

З інформації отриманої в результаті аналізу, ми можемо спостерігати, що кількість нейтральних мутацій значно перевищує кількість патогенних мутацій.

Крім того, важливо зазначити, що не зважаючи на відсутність поліморфізмів (заміни лише з одним можливим варіантом) у більшості точкових замін, поліморфізми з двома та трьома допустимими варіантами замін присутні.

Висновки

Високі темпи розвитку сучасної науки в цілому, зокрема генетики і біоінформатики, потребують зручних методів опрацювання інформації, що дозволять заощадити час та інші ресурси.

Дана робота пропонує результати аналізу допустимих мутацій у структурі послідовностей нуклеотидів мітохондріальної ДНК людського організму та порівняння ланцюжків послідовностей за різними географічними регіонами.

Такі результати стануть корисними в процесі досліджень мутацій мітохондріальних ДНК для науковців-генетиків та біоінформатиків.

По закінченню роботи було виконано усі поставлені вище завдання, а саме розроблено структуру бази даних для збереження та аналізу ланцюжків мітохондріальної ДНК, імпортовано надані дані послідовностей ланцюжків у розроблену базу даних та проведено аналіз цих послідовностей відносно різних географічних регіонів.

Також було розроблено базу даних для зберігання допустимих поліморфізмів нуклеотидів, виконано парсування, валідацію та імпорт даних у створену базу даних і проаналізовано їх.

Результати аналізу послідовностей ланцюжків мітохондріальних ДНК осіб, що походять з різних географічних регіонів, представлено у вигляді гістограм для кращого візуального сприйняття інформації.

Під час роботи виникали певні технічні труднощі, на кшталт лімітованої потужності комп’ютера при роботі з великою кількістю даних. Проте розбиття даних на декілька частин та відкидання несуттєвої інформації допомогли вирішити цю проблему.

Даний проект та його тематична актуальність мають перспективи подальшого розвитку. Надалі планується реалізувати додаткові аналітичні методи, додати та проаналізувати дані послідовностей мітохондріальної ДНК осіб з інших географічних регіонів, включаючи Україну.

Список літератури

1. James Holt I. Genetics of Mitochondrial Diseases / Ian James Holt – Oxford University Press, 2003.
2. Yin S. Why Do We Inherit Mitochondrial DNA Only From Our Mothers? [Електронний ресурс] / Steph Yin // The New York Times. – 2016. – Режим доступу до ресурсу: https://www.nytimes.com/2016/06/24/science/mitochondrial-dna-mothers.html
3. MITOMAP: mtDNA Control Region Sequence Variants [Електронний ресурс] – Режим доступу до ресурсу: https://www.mitomap.org/foswiki/bin/view/MITOMAP/PolymorphismsControl
4. MITOMAP: mtDNA Coding Region & RNA Sequence Variants [Електронний ресурс] – Режим доступу до ресурсу: https://www.mitomap.org/foswiki/bin/view/MITOMAP/PolymorphismsCoding
5. MITOMAP: Reported Mitochondrial DNA Base Substitution Diseases: Coding and Control Region Point Mutations [Електронний ресурс] – Режим доступу до ресурсу: https://www.mitomap.org/foswiki/bin/view/MITOMAP/MutationsCodingControl
6. Quenqua D. New Studies Suggest an ‘Adam’ and ‘Eve’ Link [Електронний ресурс] / Douglas Quenqua // The New York Times – 2013. – Режим доступу до ресурсу: https://www.nytimes.com/2013/08/13/science/new-studies-suggest-an-adam-and-eve-link.html
7. Веллз С. Подорож людини: генетична одіссея / Спенсер Веллз. – Книжковий клуб "Клуб Сімейного Дозвілля", 2019.

8. Homo sapiens mitochondrion, complete genome [Електронний ресурс] – Режим доступу до ресурсу: https://www.ncbi.nlm.nih.gov/nuccore/J01415.2
9. Taanman J.W. The mitochondrial genome: structure, transcription, translation and replication [Електронний ресурс] / Taanman J.W. – 1999. – Режим доступу до ресурсу: https://www.sciencedirect.com/science/article/pii/S0005272898001613
10. Гречаніна Ю.Б. Вивчення впливу поліморфізмів мтДНК та поліморфних варіантів генів C677T MTHFR, A66G MTTR на клінічні прояви мітохондріальних дисфункції: дис. на здобуття наукового ступеня доктора медичних наук: 03.00.15/ Гречаніна Юлія Борисівна. – Одеса, 2012.
11. What is FASTA format? [Електронний ресурс] – Режим доступу до ресурсу: https://zhanglab.ccmb.med.umich.edu/FASTA/
12. Apache POI - the Java API for Microsoft Documents [Електронний ресурс] – Режим доступу до ресурсу: https://poi.apache.org/
13. jsoup: Java HTML parser that makes sense of real-world HTML soup [Електронний ресурс] – Режим доступу до ресурсу: https://jsoup.org/apidocs/

Додаток А

(обов'язковий)

Надані дані щодо кількості носіїв кожної послідовності та їх розподіл за регіонами походження

№ в послідовності	Регіон	Місцевість	Кількість осіб-носіїв (1 – якщо не вказано)
J01415	A, B, C, D	Сардінія	15
M58058	E	Сардінія	
M58059	E	Сардінія, Середній Схід	1, 1
M58060	D	Сардінія	
M58061	E	Сардінія	
M58062	E	Сардінія	
M58063	E	Сардінія	
M58064	B, C, E	Сардінія, Середній Схід	3, 1
M58065	E	Сардінія	
M58066	E	Сардінія	
M58067	E	Сардінія	
M58068	C	Сардінія	
M58069	C, E	Сардінія	3
M58070	C	Сардінія	
M58071	D, E	Сардінія	2
M58072	D	Сардінія	
M58073	B, D, E	Сардінія	3
M58074	A	Сардінія	
M58075	E	Сардінія	
M58076	A, C	Сардінія	2
M58077	D	Сардінія	
M58078	D	Сардінія	

M58079	C	Сардінія	
M58080	D	Сардінія	
M58081	D	Сардінія	
M58082	E	Сардінія	
M58083	B	Сардінія	
M58084	C	Сардінія	
M58085	D	Сардінія	
M58086	B	Сардінія	
M58087	C	Сардінія	
M58088	B	Сардінія	
M58089	B, D	Сардінія	2
M58090	B	Сардінія	
M58091	B	Сардінія	
M58092	B	Сардінія	
M58093	B	Сардінія	
M58094	A	Сардінія	
M58095	A	Сардінія	
M58096	A	Сардінія	
M58097	A	Сардінія	
M58098	A	Сардінія	
M58099	A	Сардінія	
M58100	A	Сардінія	
M58101	A	Сардінія	
M58102	A	Сардінія	
M58103		Середній Схід	
M58104		Середній Схід	2
M58105		Середній Схід	
M58106		Середній Схід	
M58107		Середній Схід	
M58108		Середній Схід	

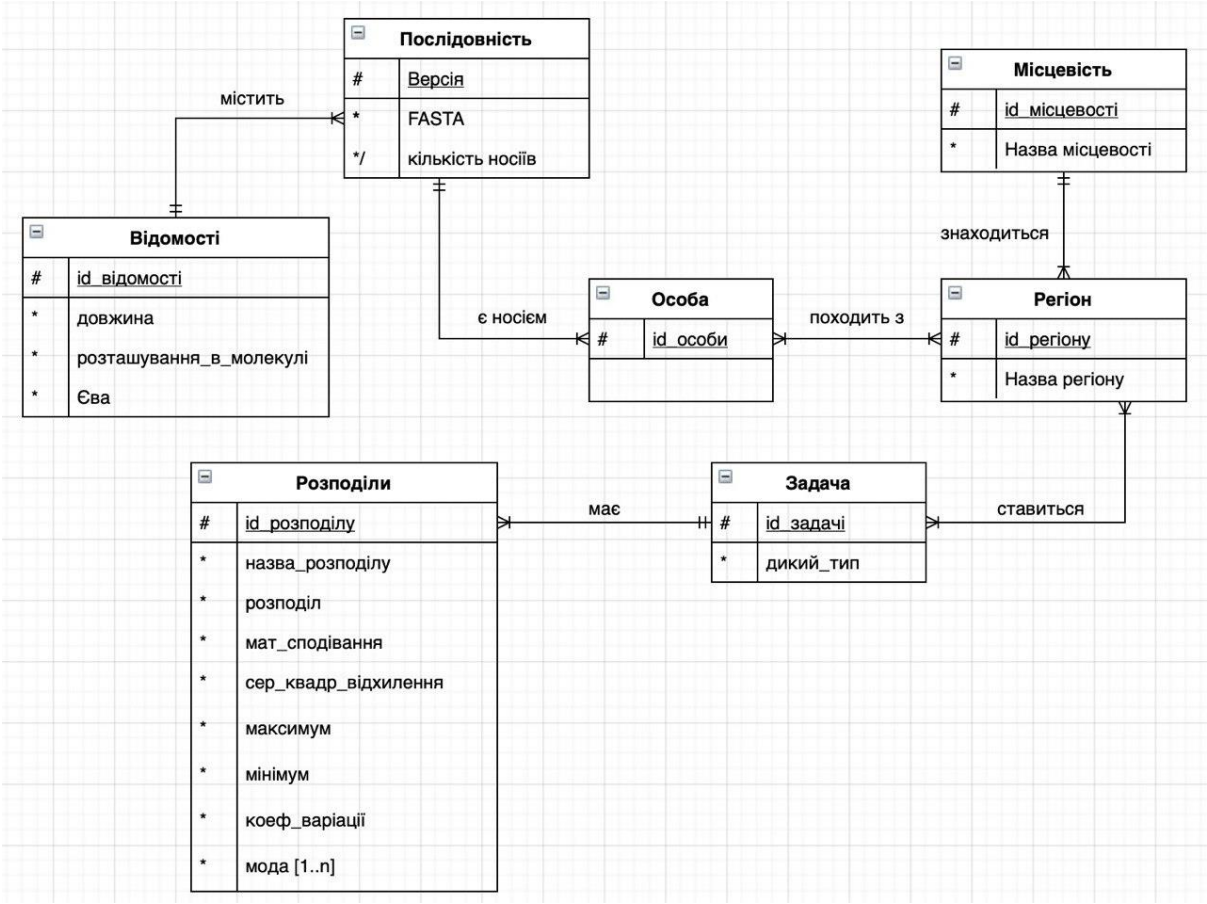
M58109		Середній Схід	
M58110		Середній Схід	2
M58111		Середній Схід	
M58112		Середній Схід	
M58113		Середній Схід	
M58114		Середній Схід	
M58115		Середній Схід	
M58116		Середній Схід	
M58117		Середній Схід	
M58118		Середній Схід	
M58119		Середній Схід	
M58120		Середній Схід	
M58121		Середній Схід	
M58122		Середній Схід	2
M58123		Середній Схід	
M58124		Середній Схід	
M58125		Середній Схід	
M58126		Середній Схід	
M58127		Середній Схід	
M58128		Середній Схід	
M58129		Середній Схід	
M58130		Середній Схід	
M58131		Середній Схід	
M58132		Середній Схід	
M58133		Середній Схід	
M58134		Середній Схід	
M58135		Середній Схід	
M58136		Середній Схід	
M58137		Середній Схід	
M58138		Середній Схід	

M58139		Середній Схід	
M58140		Середній Схід	
M58141		Середній Схід	
M58142		Середній Схід	
M58143		Середній Схід	
M58144		Середній Схід	

Додаток Б

(обов'язковий)

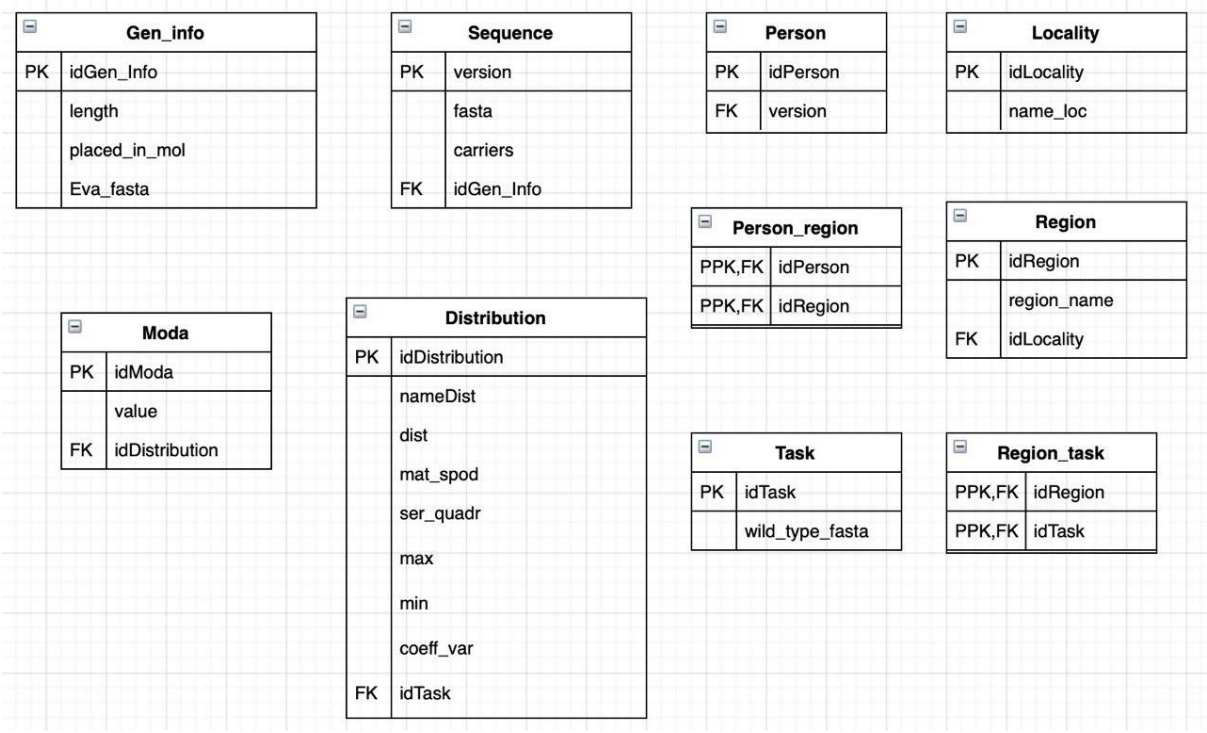
ER модель бази даних для аналізу ланцюжків мітохондріальної ДНК носіїв з різних регіонів



Додаток В

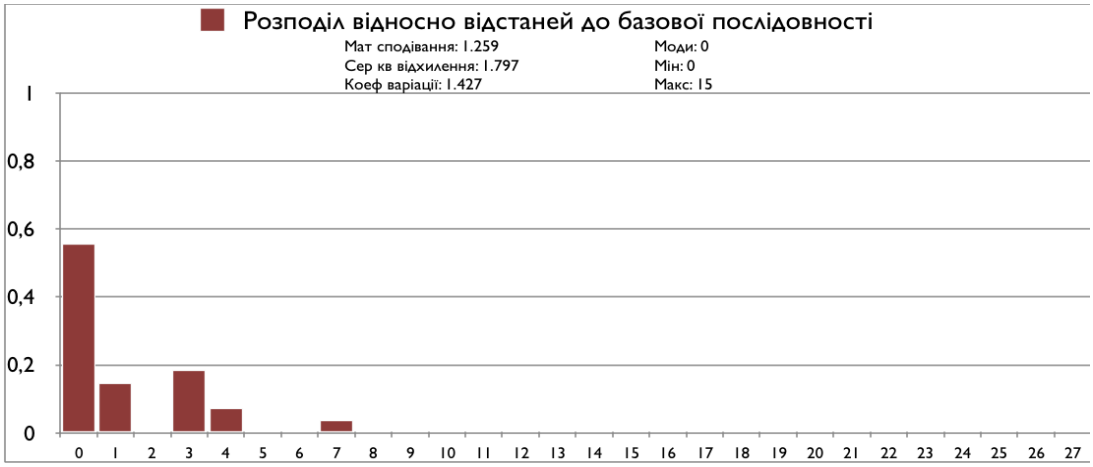
(обов'язковий)

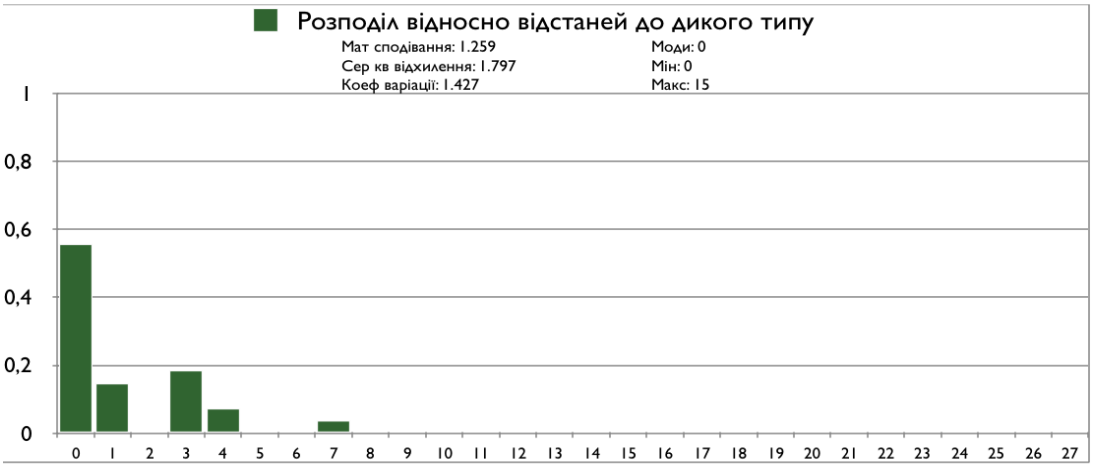
Реляційна модель бази даних для аналізу ланцюжків мітохондріальної ДНК носіїв з різних регіонів

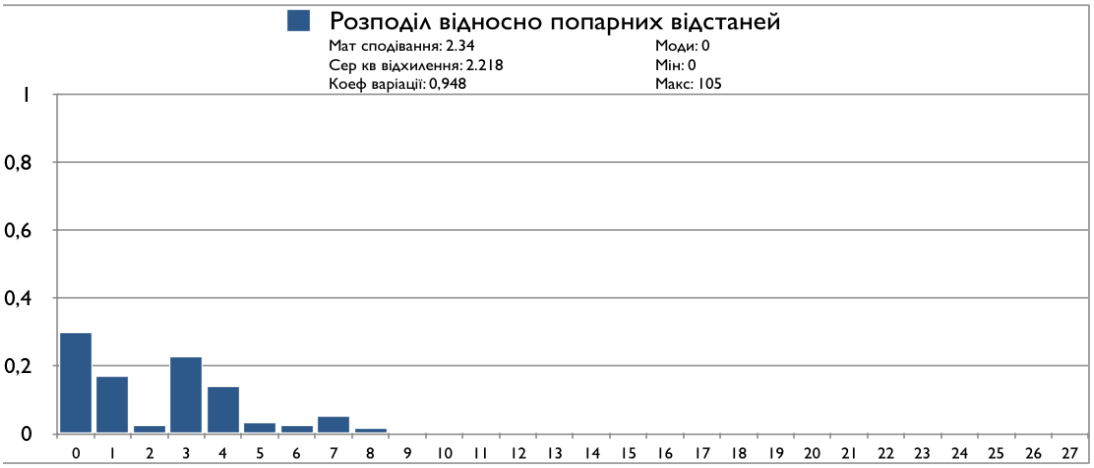


Додаток Г
(обов'язковий)
Результати аналізу ланцюжків мітохондріальної ДНК носіїв за різними географічними регіонами

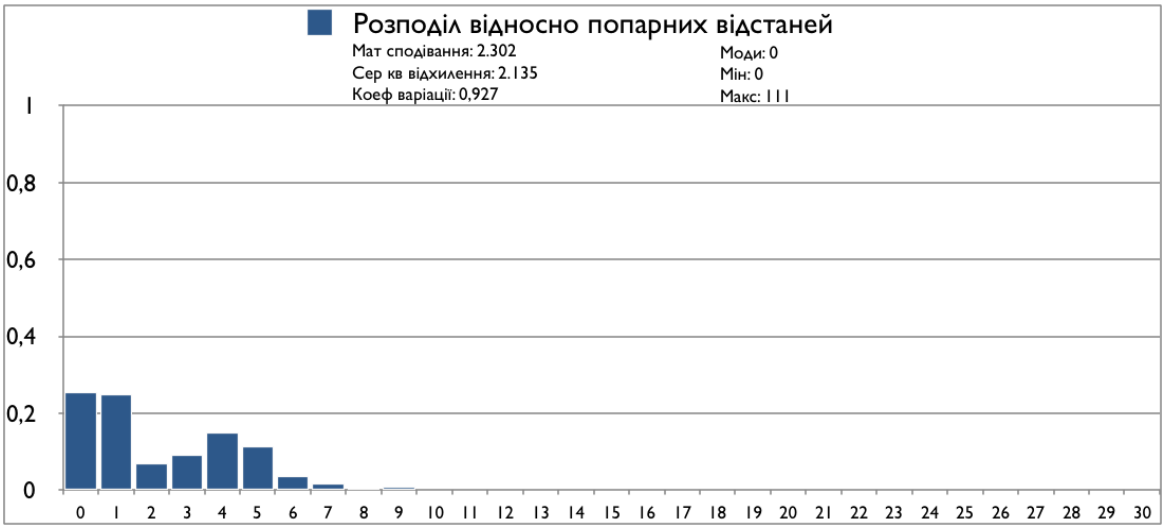
Регіон А місцевості Сардинія:



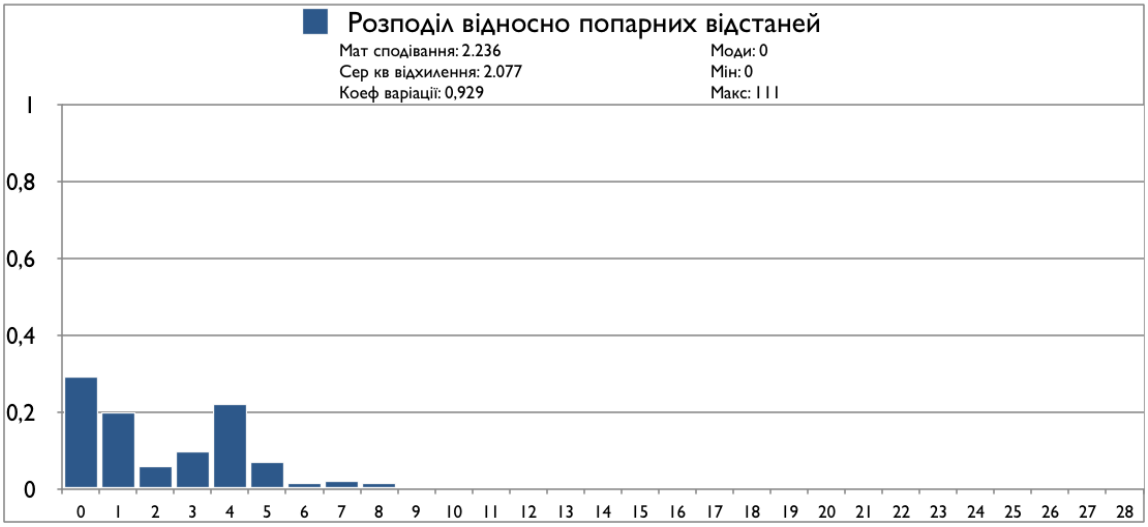




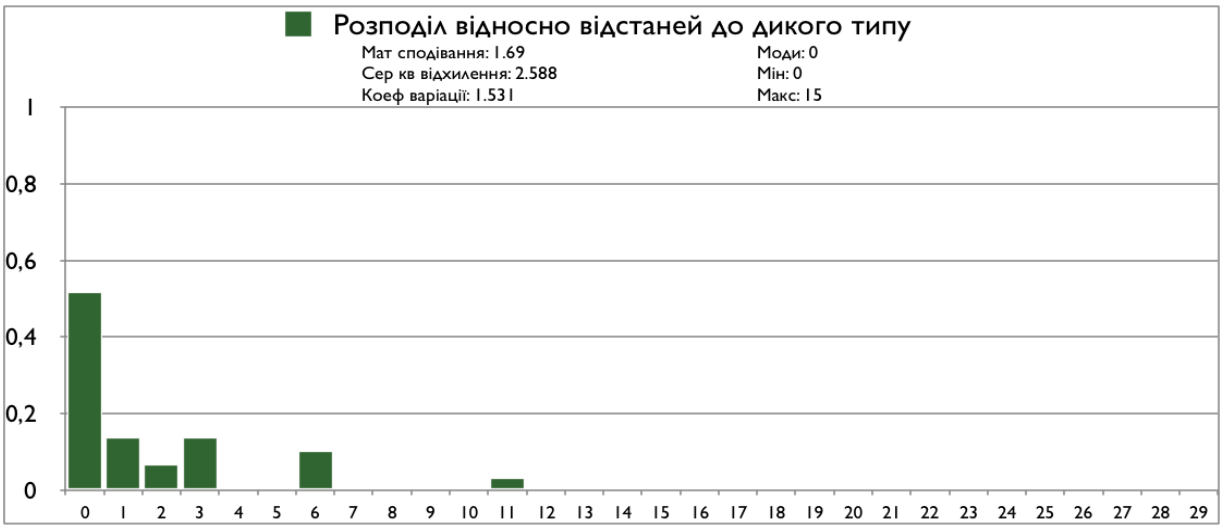
Регіон В місцевості Сардинія:

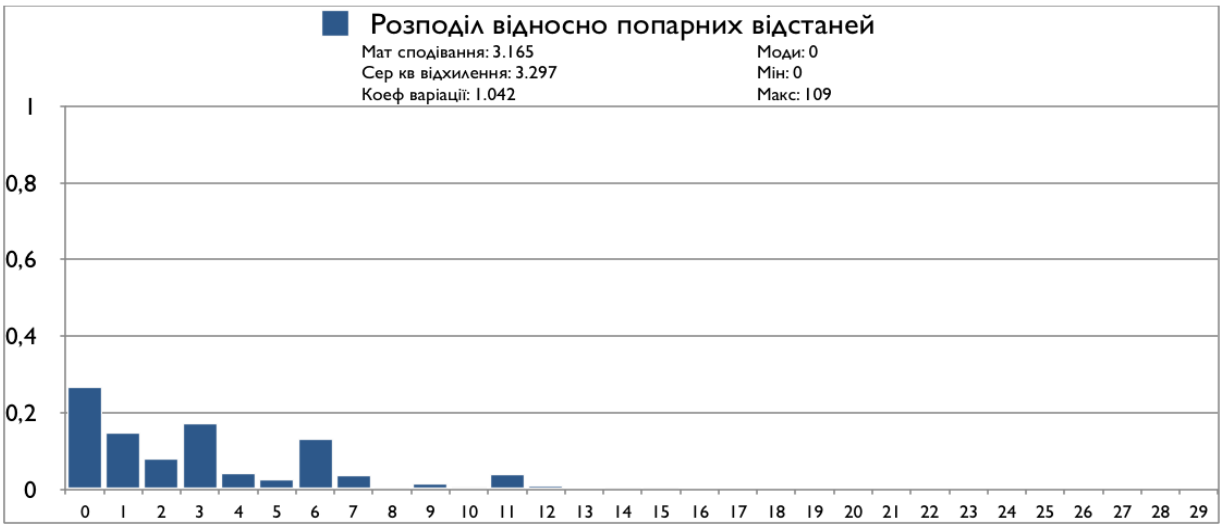


Регіон С місцевості Сардинія:

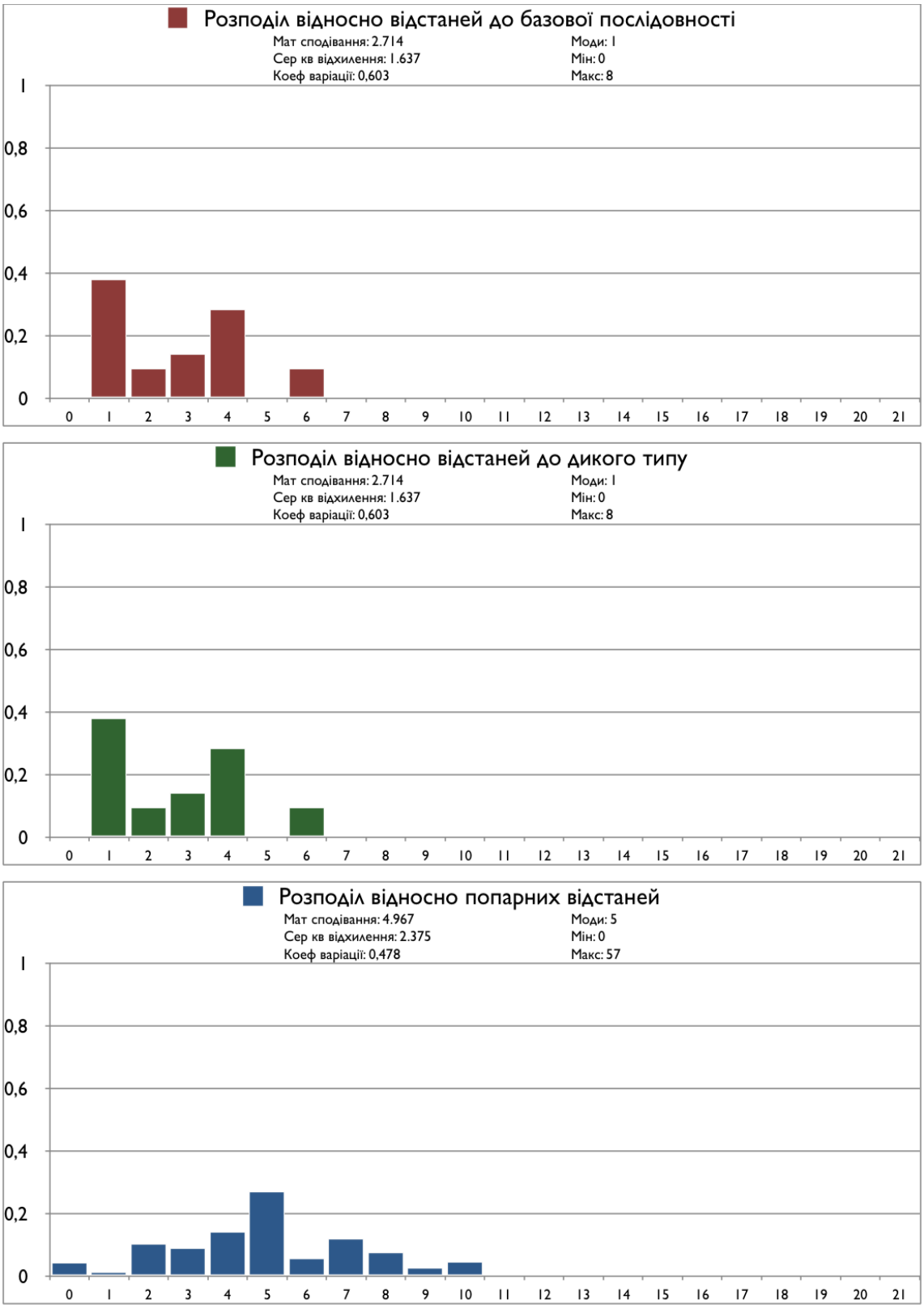


Регіон D місцевості Сардинія:

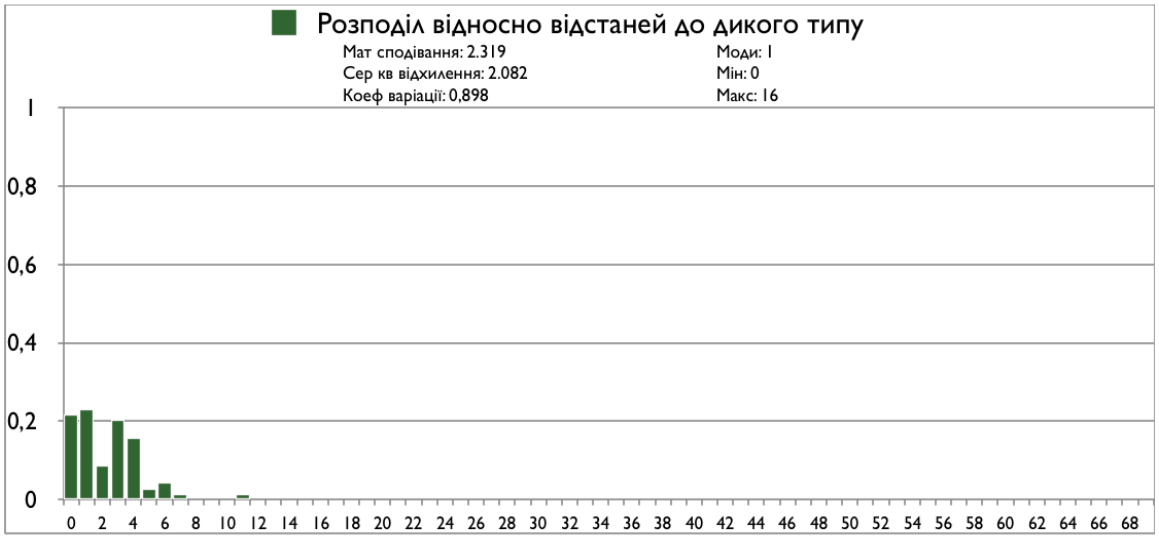


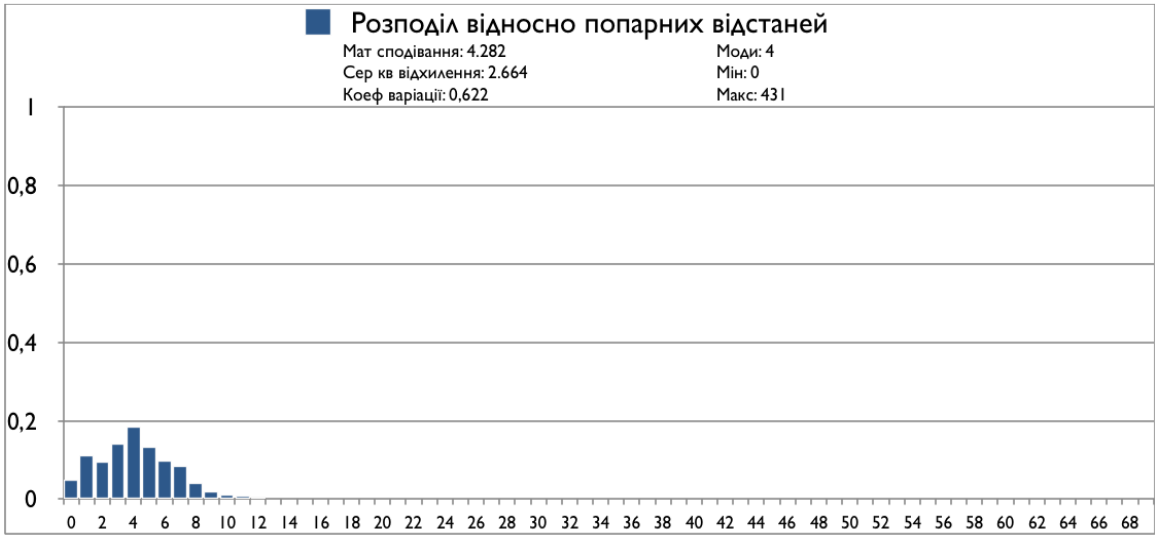


Регіон E місцевості Сардинія:

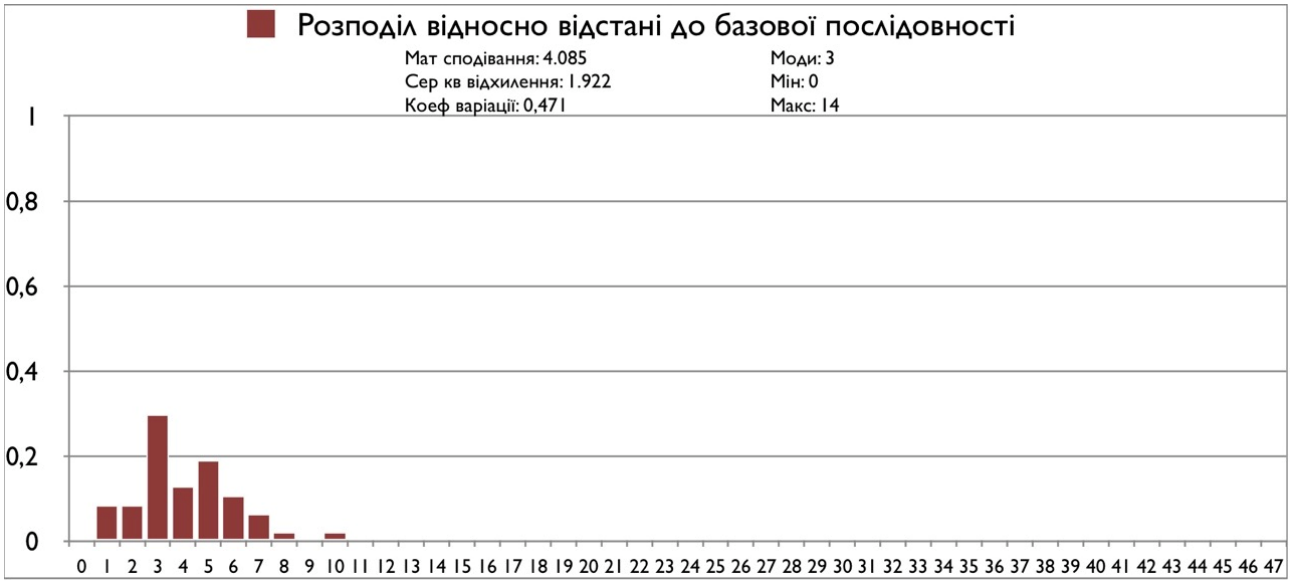


Місцевість Сардинія:





Місцевість Середній Схід:



Усі наявні місцевості: