Київський національний університет імені Тараса Шевченка

Факультет радіофізики, електроніки та комп’ютерних систем

Кафедра радіотехніки та радіоелектронних систем

**Звіт**

**З НАУКОВО-ВИРОБНИЧОЇ ПРАКТИКИ**

|  |
| --- |
| Виконавець:  студент групи ІБТКСМ (2М курс)  Гаврилюк Вадим Володимирович  Керівник практики  від кафедри:  Фесенко Сергій Олександрович |
|  |
|  |

Київ-2023

1. МЕТА ПРОХОДЖЕННЯ НАУКОВО-ВИРОБНИЧОЇ ПРАКТИКИ

Моя тема дипломної роботи «Автоматизована індетифікація багатоатомних молекул за методами машинного розпізнавання інфрачервоних спектрів». Мою дипломну роботу можна поділити на два основних етапи. На першому етапі відбувається отримання послідовності молекулярних смуг з експериментальних спектрів, а на другому етапі проводиться аналіз отриманої послідовності для індетифікації молекул. Отримані послідовності молекулярних смуг з інфрачервоних спектрів буду аналізувати за допомогою мови програмування R у програмному середовищі RStudio. Для того, щоб мати розуміння як працювати з послідовностями даних, отримував послідовність нуклеотидів некодуючої ДНК. Використовуючи програмування переводив послідовність у кодуючу ДНК, яка може кодувати синтез білків. З кодуючої послідовності нуклеотидів отримував послідовність амінокислот. Задача була виконана з метою засвоєння мови програмування R та засоби її застосування, для розуміння того, як можна приймати послідовність і записувати іншу послідовність або потрібну інформацію на основі прийнятої. Нескладне завдання було виконане для розуміння основних принципів перед тим як переходити до більш серйозного аналізу даних.

2. ДНК і РНК

Дезоксирибонуклеїнова кислота (ДНК) — один із двох типів природних нуклеїнових кислот, яка забезпечує зберігання, передавання від покоління до покоління і впровадження генетичної програми розвитку й діяльності живих організмів. Основне призначення ДНК в клітинах — довготривале зберігання відомостей про структуру РНК і білків.

З хімічної точки зору ДНК — це довга полімерна молекула, що складається з послідовності блоків — нуклеотидів. Кожний нуклеотид складається з азотистої основи, цукру (дезоксирибози) і фосфатної групи. Зв'язки між нуклеотидами в ланцюгу, утворюються дезоксирибозою й фосфатною групою. У переважній більшості випадків (окрім деяких вірусів, що містять одноланцюгові ДНК) макромолекула ДНК складається з двох ланцюгів, орієнтованих азотистими основами один проти одного. Ця дволанцюгова молекула утворює спіраль. У цілому структура молекули ДНК отримала назву «подвійної спіралі».

Рибонуклеїнова кислота, РНК — клас нуклеїнових кислот, лінійних полімерів нуклеотидів, до складу яких входять залишок фосфорної кислоти, рибоза (на відміну від ДНК, що містить дезоксирибозу) і азотисті основи — аденін, цитозин, гуанін і урацил (на відміну від ДНК, що замість урацила містить тимін).

Рибонуклеїнова кислота — одна з трьох основних макромолекул (дві інші ДНК та білки), яка грає важливу роль в кодуванні, зчитуванні, регулюванні та вираженні генів. Так само як ДНК, РНК складається з ланцюжка нуклеотидів. Кожен нуклеотид складається з азотистої основи, цукру рибози та фосфатної групи. Послідовність нуклеотидів дозволяє РНК кодувати генетичну інформацію. Всі клітинні організми (еукаріоти) використовують РНК (мРНК) для програмування синтезу білків.

Клітинні РНК утворюються під час процесу, який називається транскрипцією, тобто синтезу РНК на матриці ДНК за допомогою спеціальних ферментів — РНК-полімерази. Після цього матричні РНК (мРНК) беруть участь в процесі, який називається трансляцією. Трансляція — це синтез білку на матриці мРНК за участі рибосом. Інші РНК після транскрипції піддаються хімічним модифікаціям, і, після утворення вторинних і третинних структур, виконують функції залежно від типу РНК.

Нуклеїнові кислоти є біополімерами, мономерними ланками яких є нуклеотиди. Нуклеотиди складаються з трьох компонентів: азотиста основа, вуглевод пентоза (рибоза для РНК і дезоксирибоза для ДНК), залишок фосфорної кислоти. Азотисті основи в нуклеїнових кислотах здебільшого присутні у вигляді двох пуринових похідних: аденіну (А) і гуаніну (Г) і трьох піримідинових – цитозину (Ц), тиміну (Т) і урацилу (У). До складу одного нуклеотиду входить тільки одна азотиста основа, яка і визначає назву цього нуклеотиду. До складу ДНК входить чотири такі нуклеотиди: аденіловий (А), гуаніловий (Г), цитидиловий (Ц), тимідиловий (Т). РНК складається з одинарного полінуклеотидного ланцюга, до складу якого входять чотири такі нуклеотиди: аденіловий (А), гуаніловий (Г), цитидиловий (Ц), урациловий (У).

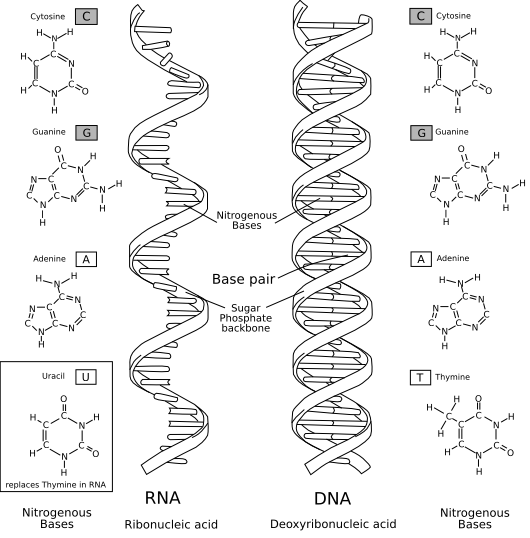


Рис. 2.1

3. ПЕРЕХІД ВІД ПОСЛІДОВНОСТЕЙ КОДУЮЧОЇ ДНК ДО ПОСЛІДОВНОСТІ АМІНОКИСЛОТ

З фрагменту першого ланцюга ДНК за принципом комплементарності можна визначити послідовність нуклеїнових кислот, закодованого фрагментом другого ланцюга ДНК.

У разі нуклеїнових кислот — як оліго- так і полінуклеотидів — азотисті основи нуклеотидів здатні унаслідок утворення водневих зв’язків формувати парні комплекси аденін — тимін (або урацил в РНК) і гуанін—цитозин при взаємодії ланцюжків нуклеїнових кислот (тобто А-Т і Г-Ц або A-T і G-C). Така взаємодія грає ключову роль у ряді фундаментальних процесів зберігання і передачі генетичній інформації: реплікації ДНК, що забезпечує передачу генетичній інформації при клітинному поділі, транскрипції ДНК в РНК при синтезі білків, кодуванні ДНК генів, зберіганні генетичної інформації в дволанцюжковій ДНК і процесах репарації (ремонту) ДНК при її пошкодженні.

Транскрипція – це процес синтезу і – РНК на певній ділянці ДНК (один чи декілька генів) за принципом комплементарності. Тому синтезована і – РНК є генетичною копією цієї ділянки ДНК, має той самий нуклеотидний склад, що ДНК, і, отже, несе тотожну спадкову інформацію про первинну будову поліпептиду. Відбувається цей процес у ядрі клітини за участю ферменту РНК-полімерази. Більшість структурних генів еукаріот (ділянки ДНК) внутрішньо неоднородні. Вони складаються з екзонних (інформативних) та інтронних (безінформаційних) фрагментів. Під час транскрипції в ядрі спочатку синтезується про-іРНК (незріла), яка має в собі як екзони, так і інтрони. Далі за допомогою комплексу ферментів безінформаційні ділянки вирізаються та руйнуються, а інформативні з’єднуються в новий полінуклеотидний ланцюг – зрілу іРНК. Механізм дозрівання іРНК в ядрі називається процесингом. В процесі транскрипції аденін(A) переходить в урацил (U), тимін (T) в аденін (A), гуанін (G) в цитозин (C), цитозин (C) в гуанін (G)

Трансляція – це процес трансформації нуклеотидної послідовності і – РНК в амінокислотну послідовність поліпептиду. Здійснюється цей процес теж за принципом комплементарності, що забезпечує утворення білкової молекули відповідно до спадкової інформації (нуклеотидної послідовності) певної ділянки ДНК. Процес трансляції відбувається позаядерно в рибосомах (полі рибосомах) з участю певних ферментних систем. Під час трансляції послідовність нуклеотидів групуються по три нуклеотиди, і відповідно поділяються на кодони.

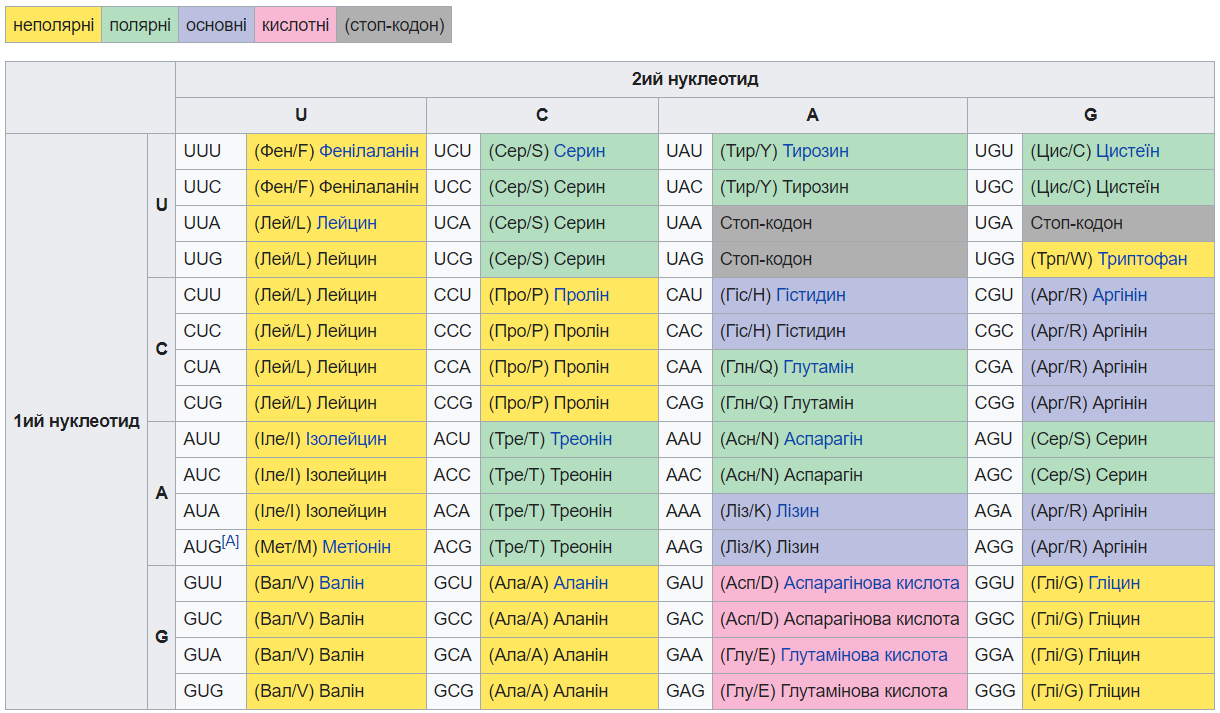


Рис. 3.1

Кожній амінокислоті білка відповідає послідовність з трьох розташованих один за одним нуклеотидів ДНК — триплет.

Кожен триплет нуклеотидів кодує певну амінокислоту, яка буде вбудована у поліпептидний ланцюг. На рисунку 3.1 вказано саме який триплет яку амінокислоту кодує.

Як відомо, до складу ДНК можуть входити чотири нітрогеновмісних основи: аденін (А), гуанін (Г), тимін (Т) і цитозин (Ц). Число сполучень з 4 по 3 складає 64, тобто ДНК може кодувати 64 амінокислоти. Проте кодується тільки 20 амінокислот.

Виявилося, що багатьом амінокислотам відповідає не один, а декілька кодонів. Передбачається, що така властивість генетичного коду (виродженність) підвищує надійність зберігання і передачі генетичної інформації під час ділення клітин.

Для науково-виробничої практики було сформоване завдання. На першому кроці отримуємо будь-який фрагмент послідовностей нуклеотидів у некодуючому ланцюзі молекули ДНК та за принципом комплементарності отримуємо фрагмент послідовності нуклеотидів у фрагменті ланцюга ДНК, що кодує білок. З фрагменту кодуючого ланцюга ДНК переносимо генетичну інформацію в РНК. Потім переводимо послідовність нуклеотидів РНК у поліпептидну послідовність амінокислот. Для виконання такого процесу буде використовуватись мова програмування в R у середовищі програмування RStudio.

4. МОВА ПРОГРАМУВАННЯ R ТА ЇЇ ОСОБЛИВОСТІ

R – це мова програмування і програмне середовище для статистичних обчислень, аналізу та зображення даних в графічному вигляді. Розробка R відбувалась під впливом двох наявних мов програмування: мови програмування S з семантикою успадкованою від Scheme. R названа за першою літерою імен її засновників Роса Іхаки (Ross Ihaka) та Роберта Джентлмена (Robert Gentleman) працівників Оклендського Університету в Новій Зеландії. R має велику кількість розширень у вигляді пакетів, які розробляються спільнотою користувачів. Ці пакети розширюють функціональність R, дозволяючи вам використовувати нові методи та алгоритми для аналізу даних. R є мовою, що означає, що ви можете використовувати її на різних операційних системах, таких як Windows, macOS та Linux. Має активну та велику спільноту користувачів, що допомагає вирішувати проблеми, надає поради та розробляє нові пакети.

R є мовою програмування з орієнтацією на векторизацію. Це означає, що ви можете виконувати операції над цілими векторами даних без необхідності використання явних циклів. Це робить код більш ефективним та зручним для аналізу даних. R зазвичай використовується в інтерактивному режимі, що дозволяє користувачам вводити команди та одразу бачити результати. Це робить вивчення та розробку коду більш ефективним та інтуїтивно зрозумілим. R підтримує об'єктно-орієнтоване програмування, що дозволяє структурувати код у вигляді об'єктів. Це корисно для створення повторно використовуваних та організованих програм.

5. ВИКОНАННЯ ЗАДАЧІ МОВОЮ ПРОГРАМУВАННЯ R В ПРОГРАМНОМУ СЕРЕДОВИЩІ RSTUDIO

Для того, щоб отримати фрагмент послідовності кодуючої ланки ДНК з некодуючої, була записана функція getMDNAfromDNA(DNA). Функція на вхід приймає дані у формі одного вектора. Кожен елемент вектора є нуклеотидом, що має одне із чотирьох типів азотистої основи. Типами азотистих основ є аденін(A), тимін(T), цитозин(C), гуанін(G) і елементи відповідно мають одне із чотирьох значень A, T, C, G. Функція виконує заміну елементів A на T, T на A, C на G, G на C і повертає новий вектор. Для того щоб перевірити, що функція працює правильно, здійснював її виклик, записуючи вектор з невеликою кількістю елементів, сукупність яких є фрагмент послідовності ДНК.

Для того, щоб отримати вектор, що являє собою послідовність РНК записав функцію getRNAfromMDNA(MDNA). Функція приймає послідовність кодуючої ланки ДНК у вигляді вектора, щоб був отриманий з функції getMDNAfromDNA(DNA). Функція getRNAfromMDNA(MDNA) виконує заміну елементів A на U(урацил), T на A, C на G, G на C. Вона повертає фрагмент РНК у вигляді вектора з елементами. Кожен елемент вектора може мати одне із чотирьох значень A, U, C, G. Для перевірки того, що функція працює правильно, був записаний невеликий фрагмент ДНК у вигляді вектора, і був здійснений виклик.

Була записана функція getAminoAcidsFromRNA(RNA) для того, щоб з послідовності РНК отримати послідовність нуклеотидів. Функція приймає послідовності РНК, а повертає послідовності з амінокислот. Функцію можна поділити на три основних етапи дії. На першому етапі функція перевіряє дані, що отримує. Для успішного виконання вона має приймати одну зміну яка є вектором і має довжину (кількість елементів) кратну трьом. Кожний елемент може мати тільки одне із чотирьох значень: A, U, C, G. Якщо вектор відповідає таким умовам, функція продовжує свою роботу. Далі відбувається групування елементів, їх розподіл на кодони. Кожний кодон складається із трьох нуклеотидів і відповідає одній із 20 можливих амінокислот. На цьому етапі здійснюється перетворення послідовності нуклеотидів РНК у послідовність кодонів. Вектор з елементами, що відповідають нуклеотидам, переходить у вектор з елементами, які відповідають кодонам, тобто кожен елемент є строкою з трьох літер. На останньому етапі відбувається заміна елементів вектора з послідовністю кодонів. Кожний елемент вектора замінюється на назву амінокислоти. В результаті отримуємо послідовність з амінокислот.

Перелічені вище функції працюють успішно і виконують бажаний результат. Разом взяті функції і правильна послідовність їх виконують отримання послідовності амінокислот, маючи тільки послідовність нуклеотидів некодуючої ланки ДНК.

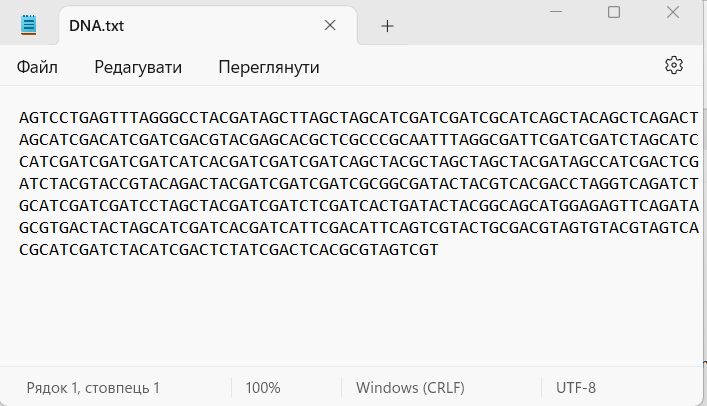


Рис. 5.1

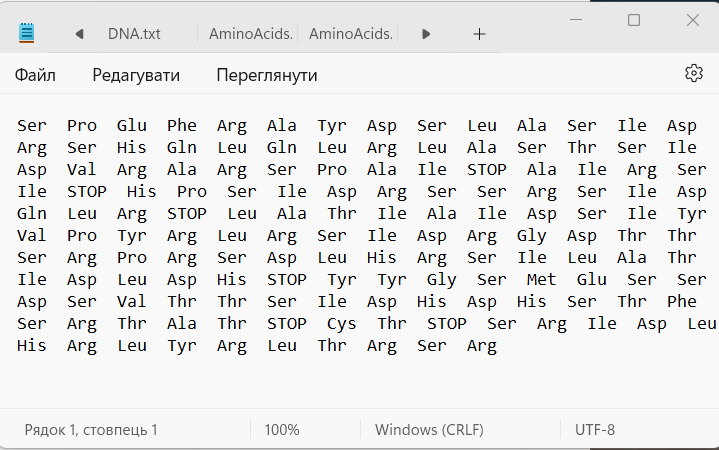


Рис. 5.1

Для розуміння того, як можна за допомогою мови програмування R та засобів її застосування зчитувати дані з файлів та записувати дані у файли, була проведена робота з текстовими файлами. Використовуючи команду read.table() можна записати у зміну дані. Команда write.table() створює файл з даними, які зазвичай означають результат завершеної роботи. Отже, програмний код приймає файл з послідовностями ДНК і записує послідовність амінокислот у новий файл. Фрагмент програмного коду за виключенням об’явлення самих функцій зображений на рисунку 5.3. Файл, з якого зчитується інформація, і файл, у який записуються бажані дані, зображені на рисунках 5.1 і 5.2 відповідно.

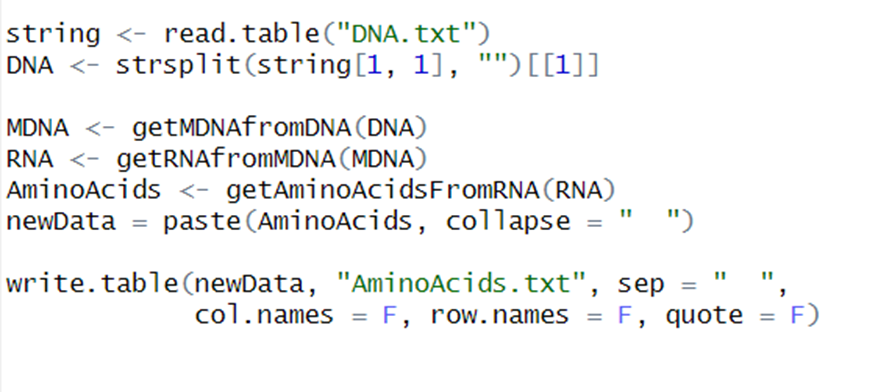


Рис. 5.3

ВИСНОВКИ

Під час виконання науково-виробничої дізнався про ДНК, РНК та амінокислоти. Цікавився такими процесами як транскрипція, трансляція та біосинтез білків. Вивчав основи по мові програмування R для аналізу даних. Вивчав способи застосування для програмування мовою R. Написав програму у програмному середовищі RStudio, яка отримує текстовий файл з послідовностями ДНК і записує текстовий файл з послідовностями амінокислот.