Біологічні науки

УДК 575

Ліщинська Руслана Віталіївна

студентка кафедри біомедичної кібернетики

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

Лищинская Руслана Витальевна

студентка кафедры биомедицинской кибернетики

Национального технического университета Украины

«Киевский политехнический институт имени Игоря Сикорского»

Lishchynska Ruslana

Student of the Department of Biomedical Cybernetics of the

National Technical University of Ukraine

"Igor Sikorsky Kyiv Polytechnic Institute"

Кисляк Сергій Володимирович

старший викладач кафедри біомедичної кібернетики

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Кисляк Сергей Владимирович

старший преподаватель кафедры биомедицинской кибернетики

Национальный технический университет Украины

«Киевский политехнический институт имени Игоря Сикорского»

Kyslyak Sergii

Senior Lecturer of the Department of Biomedical Cybernetics

National Technical University of Ukraine

"Igor Sikorsky Kyiv Polytechnic Institute"

ПРОГРАМНИЙ ПРОДУКТ ДЛЯ ПОШУКУ CRISPR-СИСТЕМ
ПРОГРАММНЫЙ ПРОДУКТ ДЛЯ ПОИСКА CRISPR-СИСТЕМ
SOFTWARE TO SEARCH FOR CRISPR-SYSTEMS

Анотація. *Досліджено CRISRP-системи. Проаналізовано найпопулярніші алгоритми для пошуку паліндромних повторів. Реалізовано алгоритм Манакера для пошуку паліндромних повторів в біологічних послідовностях. Розроблено програмний додаток, що знаходить паліндроми і виводить їх кількість в заданій біологічній послідовності.*

Ключові слова: *CRISRP, алгоритми для пошуку паліндромів, алгоритм Манакера, біологічні послідовності, редагування геномів.*

Аннотация. *Исследованы CRISRP-системы. Проанализированы самые популярные алгоритмы для поиска палиндромный повторов. Реализован алгоритм Манакера для поиска палиндромных повторов в биологических последовательностях. Разработан программный продукт, что находит палиндромы и выводит их количество в заданной биологической последовательности.*

Ключевые слова: *CRISRP, алгоритмы для поиска палиндромов, алгоритм Манакера, биологические последовательности, редактирование геномов.*

Summary. *CRISRP-systems are investigated. The most popular algorithms for finding palindromic repetitions are analyzed. The Manaker algorithm is implemented to search for palindromic repeats in biological sequences. The software application, which finds palindromes and displays their quantity in the given biological sequence, is developed.*

Key words: *CRISRP, algorithms for finding palindroms, Manaker's algorithm, biological sequences, genome editing.*

Постановка проблеми. CRISPR / Cas9 - це нова технологія для редагування геномів вищих організмів на основі імунної системи бактерій. Ця система може бути в ділянках бактеріальної ДНК, що містить короткі паліндромні кластерні повтори або CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). Між ідентичними повторами розташовуються відмінні один від одного фрагменти ДНК - спейсери, більшість з яких відповідають фрагментам геномів бактеріофагів. Коли вірус надходить у бактеріальну клітину, його виявляють за допомогою спеціалізованих білків Cas (CRISPR-associated sequence - послідовність, пов'язана з CRISPR), асоційована з РНК CRISPR. Якщо фрагмент вірусу зустрічається в спейсері CRISPR РНК, Cas-білки вирізають вірусну ДНК і знищують її, захищаючи клітину від інфекції. На початку 2013 р. декілька груп вчених показали, що CRISPR / Cas системи можуть працювати не тільки в бактеріальних клітинах, але і в клітинах еукаріотичних організмів, що означає, що CRISPR / Cas системи дозволяють редагувати неправильні послідовності генів, що дозволить лікувати спадкові хвороби людини [1]. Саме тому є надзвичайно актуальним створення програмних продуктів для пошуку CRISPR в біологічних послідовностях.

Аналіз останніх досліджень та публікацій. З 2012 року технології CRISPR було присвячено понад 5000 досліджень [2]. Запропоновано метод вивчення хвороб, створення нових лікарських засобів та методів лікування. В 2016 федеральна комісія з біологічної безпеки та етики США затвердила проведення першого експерименту з редагування генома людини за допомогою системи CRISPR / Cas9. Учені застосували технологію для модифікації лімфоцитів з метою лікування злоякісних захворювань крові [2]. Сьогодні найактивнішою країною в питаннях клінічних випробувань методу на людях став Китай. Зараз CRISPR найбільш активно розглядають як інструмент боротьби з раком. Перше

дослідження на живій людині провели в Сичуанському університеті. Вчені ввели модифіковані Т-лімфоцити хворому на рак легень з метою редагування гену, що контролює експресію білка PD-1, що приймає участь в диференціації імунних клітин [3]. У США перший подібний експеримент розпочався в березні 2018 року. Вчені з Пенсільванського університету проводять дослідження на 18 добровольцях, хворих множинною мієломою, саркомою і меланомою. На кінець року заплановані перші клінічні випробування в Європі. Але там лікарі будуть лікувати пацієнта з спадковим захворюванням крові - бета-таласемією [4].

Виділення не вирішених раніше частин загальної проблеми. Існуючі програмні додатки для пошуку CRISPR користуються недосконалими алгоритмами для пошуку паліндромів, що вимагає вирішення проблем пов'язаних з колізіями, використанням великих об'ємів пам'яті чи часу. Застосований алгоритм Манакера виключає всі ці недоліки та достовірно знаходить всі паліндроми в необхідній послідовності.

Мета статті. Головною метою даної роботи було створення програмного продукту для пошуку CRISPR, тобто застосування алгоритму Манакера для пошуку паліндромних повторів в біологічних послідовностях. Предметом дослідження є алгоритм Манакера та CRISPR- системи.

Виклад основного матеріалу. CRISPR (від англ. Clustered Regularly Interspaced Short Palindromic Repeats, короткі паліндромні повтори, регулярно розташовані групами) - це імунна система прокаріот, що забезпечує захист від чужорідних репліконів, в першу чергу - вірусів і плазмід. CRISPR-система складається з двох принципових компонентів: CRISPR-касет і Cas-білків (від англ. CRISPR-associated proteins). Кожна функціональна касета містить елементи трьох типів: лідерну послідовність, спейсери і повтори (Рис. 1) [5].

Рис. 1. Структура CRISPR-касети [5]

На початку CRISPR-касети розташовується лідерна послідовність, вона задає напрямок транскрипції касети. Після неї розташовані повтори та спейсери. Довжина повторів складає від 24 до 48 пар нуклеотидів. Повтори в межах однієї касети, як правило, ідентичні між собою по послідовності і довжині, наведена кількість повторів може відрізнятись парою кінцевих нуклеотидів. Між повторами розташовані варіабельні участки ДНК - спейсери. Порівняння послідовностей спейсерів з відомими нуклеотидними послідовностями показало, що деякі спейсери збігаються з ділянками вірусних і плазмідних геномів, що дозволило довести імунну роль CRISPR [6].

Для програмної ідентифікації CRISPR-касет часто застосовують алгоритми пошуку паліндромних повторів. Паліндром - число, буквосполучення, слово або текст, що однаково читається в обох напрямках. Іноді паліндромом називають будь-який симетричний щодо своєї середини набір символів. Найпопулярнішими алгоритмами для пошуку паліндромів є: тривіальні алгоритми з асимптотикою $O(N^2)$ та $O(N^3)$, алгоритм з використанням хешів та дерев паліндромів. Проте, в ході дослідження було з'ясовано, що найоптимальнішим рішенням в пошуку CRISPR є застосування алгоритму Манакера, в зв'язку з швидкодією в часі, що дає беззаперечну перевагу при роботі з послідовностями великої довжини. Його суть полягає в наступному - для швидкого обчислення притримуються границь (l, r) найправішого з виявлених підпаліндромів (тобто підпаліндромів з найбільшим значенням

r), $d_1[]$ приймається за масив паліндромів. Нехай необхідно обчислити значення $d_1[i]$ для чергового i, при цьому всі попередні значення $d_1[]$ вже підраховані. Якщо *i* не перебуває у межах поточного підпаліндрома, тобто $i > r$, то просто виконується тривіальний алгоритм. Тобто необхідно послідовно збільшувати значення $d_1[i]$, і перевіряти кожен раз – чи правда поточний підрядок - $[i - d_1[i] ; i + d_1[i]]$ є паліндромом. Коли знаходиться перша розбіжність, або коли досягається границі рядка s - зупиняються: значення $d_1[i]$ вирішено. Для того, щоб витягти частину інформації з уже підрахованих значень $d_1[]$ необхідно відобразити позицію *i* всередині підпаліндрома (*l*, *r*), тобто відображення *j = l + (r - i)*, і розглянути значення $d_1[j]$. Оскільки *j* - позиція, симетрична позиції *i*, то можна привласнити $d_1[i] = d_1[j]$. Ілюстрація цього відображення (паліндром навколо *j* фактично "копіюється" в паліндром навколо *i*) [7]:

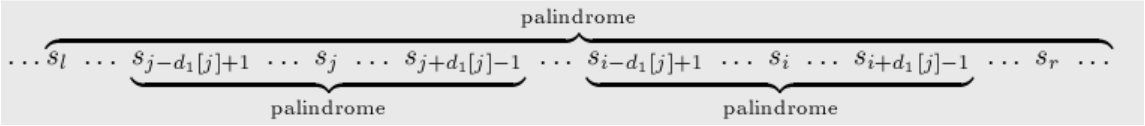

Рис. 2. Ілюстрація відображення

З метою пошуку паліндромів в біологічних послідовностях за допомогою алгоритму Манакера було створено програмний додаток, інтерфейс якого зображений на рис. 3.

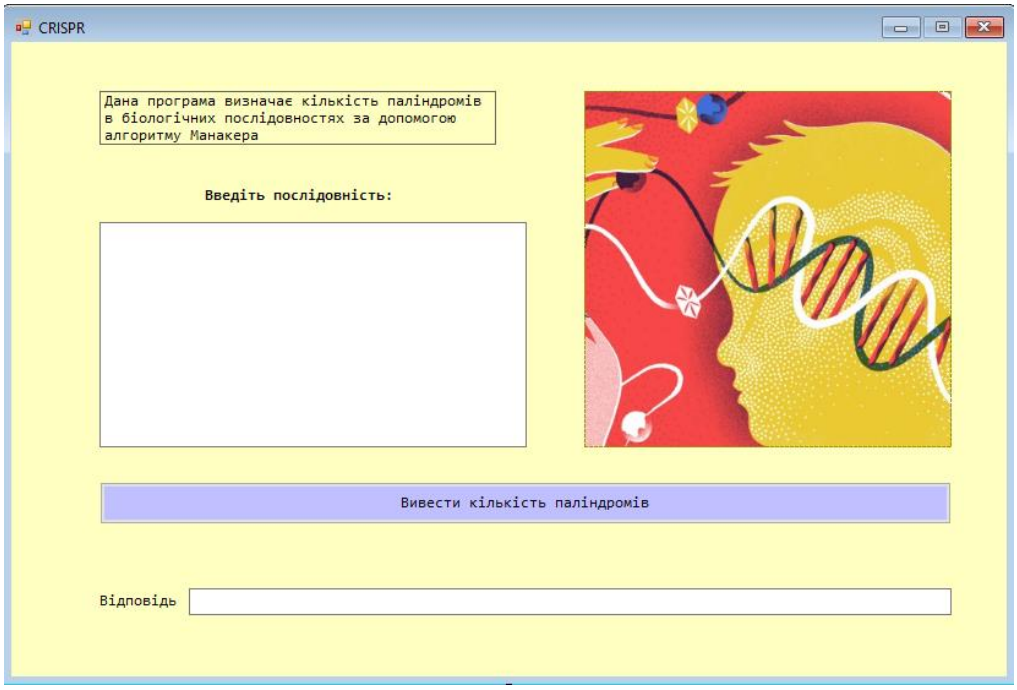

Рис. 3. Інтерфейс програмного додатку

В текстове поле вводиться біологічна послідовність в FASTA форматі. Необхідну для опрацювання біологічну послідовність можна знайти за допомогою сайту https://www.ncbi.nlm.nih.gov/, де знаходиться база біологічних послідовностей. Алгоритм починає опрацювання введеної біологічної послідовності та пошуку паліндромів. Після знаходження та підрахунку кількості всіх паліндромів, їх кількість виводиться в текстове поле внизу вікна.

Приклад розв'язку задачі пошуку. Перед пошуком паліндромів в біологічній послідовності перевіримо коректність виконання програми для рядка «AAAA», так як відомо, що для нього загальна кількість паліндромів дорівнює 6. Вводимо «AAAA» в текстове поле та очікуємо результат.

Рис. 4. Перевірка коректності виконання

Як бачимо, алгоритм працює коректно та виводить достовірний результат, тому переходимо до головного завдання, тобто визначення кількості паліндромів в біологічній послідовності. Необхідну для опрацювання біологічну послідовність знаходимо за допомогою сайту https://www.ncbi.nlm.nih.gov/, де знаходиться база біологічних послідовностей в FASTA форматі. Покажемо на прикладі як працює даний програмний додаток. В даній роботі робота програми буде продемонстрована на фрагменті геному миші.

```
AGACTATTGATGACTGCCTCTATTTCTTTAGGGGAAATGGGACTTTTAGTCCATGAATCTGATCCTGATT
TAGCTTTGGTACCTGGTATCTGTCTAGGAAGTTGTCCATTTCATCCAGGTTTTCCTGGTTTTTTTTTAGT
ATAGCCTTTCATAGTAAAATCTGATGATGTTTTTGATATCCTCATGTTCTGTTGGTATGTCTCCTTTTTC
ATTTCTGATTTTGTTAATTATAGTACAGTCCCTATGCCCTCTAGTTAGTCTGGCTAAGGGTTTATCTATC
TTGTTGACTTTCTCAAAGAACCAGCTACTATTTTGGTTGATTCTTTGAATATTTCTTTTTGTTTCCACTT
```

Рис. 5. Фрагмент геному миші

Рис. 6. Результат для фрагменту геному миші

Для біологічних послідовностей кількість паліндромів надзвичайно велика, навіть для не дуже великих фрагментів. Саме тому було прийнято рішення виводити тільки кількість паліндромів без їх явного представлення, що вимагає використання додаткових ресурсів.

Висновки та перспективи розвитку. Було створено програмний продукт для пошуку паліндромних повторів за допомогою алгоритму Манакера. На даному етапі програмний продукт готовий до використання, проте планується його вдосконалення з реалізацією можливості ідентифікації паліндромів, що є частиною CRISPR систем. Таким чином, в майбутньому ця програма буде доповнена пошуком співпадаючих пар паліндромів та перевіркою участків між ними з базою вірусних геномів за допомогою алгоритму BLAST.

Література

1. Finance.ua [Електронний ресурс]. – Режим доступу: https://news.finance.ua/ru/news/-/417752/na-smenu-crispr-idet-gennoe-redaktirovanie-20
2. Мой геном [Електронний ресурс]. – Режим доступу: http://mygenome.su/news/1050/
3. 112.ua [Електронний ресурс]. – Режим доступу: https://112.ua/statji/crispr-panaceya-ot-smertelnyh-bolezney-ili-prizrachnaya-perspektiva-430350.html
4. E. Mick, A. Stern, and R. Sorek, "Holding a grudge: persisting anti-phage CRISPR immunity in multiple human gut microbiomes.," RNA Biol., vol. 10, no. 5, pp. 900–6, 2013.
5. E. Deltcheva, K. Chylinski, C. M. Sharma, and K. Gonzales, "Europe PMC Funders Group Europe PMC Funders Author Manuscripts CRISPR RNA maturation by trans -encoded small RNA and host factor RNase III," vol. 471, no. 7340, pp. 602–607, 2011.
6. R. Sorek, V. Kunin, and P. Hugenholtz, "CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea.," Nat. Rev. Microbiol., vol. 6, no. 3, pp. 181–6, Mar. 2008.
7. MAXimal [Електронний ресурс]. – Режим доступу: http://e-maxx.ru/algo/palindromes_count