УДК 519.7

Є.О. Терпіловський

МЕТОД *K-MER* У ЗАВДАННЯХ ВИЯВЛЕННЯ ЗАКОНОМІРНИХ ПОСЛІДОВНОСТЕЙ

Терпіловський Єгор Олександрович

Інститут кібернетики імені В.М. Глушкова НАН України, м. Київ,
https://orcid.org/0000-0001-5137-0126

terpilovskiy.egor@gmail.com

У статті порівнюються дві методології попередньої обробки послідовностей ДНК людини для покращення ідентифікації конкретних генетичних захворювань за допомогою методів машинного навчання. Перший підхід забезпечує вибірку слів *k-mer*, тоді як другий використовує Multiple EM for Motif Elicitation (MEME) для розпізнавання мотиву. Метод *k-mer* передбачає розбиття послідовності ДНК на менші фрагменти фіксованої довжини, що дозволяє структурувати та аналізувати великі обсяги даних ефективно. З іншого боку, MEME застосовує алгоритм максимізації сподівань (EM) — Expectation-Maximization для виявлення статистично значущих біологічних мотивів у послідовностях, що дає змогу глибше зрозуміти функціональні області ДНК. Всебічний аналіз передбачає тренування моделі машинного навчання на вибірках даних, оцінку точності та інші метрики продуктивності, а також можливість практичного впровадження обох методів. Дані для дослідження надані центром U.S. National Library of Medicine і репрезентовані у форматі FASTA, який забезпечує стандартизоване представлення нуклеотидних послідовностей. Кожен зразок ДНК належить людям, які дали згоду на використання їхніх генетичних матеріалів у наукових дослідженнях. Для забезпечення всебічного аналізу дані оброблені як за допомогою *k-mer*, так і MEME. Перший метод сумісний з різними алгоритмами машинного навчання та дозволяє ефективно обробляти великі обсяги генетичних даних, а другий є потужним інструментом для розпізнавання мотивів, але потребує значних обчислювальних ресурсів та часу для аналізу. Порівняння цих методів показало, що у контексті ідентифікації генетичних захворювань за геномними послідовностями *k-mer* має переваги у швидкості та ефективності, тому більш придатний для практичного застосування у клінічних умовах. Виявлено, що цей метод забезпечує також високу точність і ефективність, що робить більш доцільним його інтеграцію у клінічні системи для швидшої діагностики. Отримані висновки сприятимуть удосконаленню підходів до генетичної діагностики та розвитку персоналізованої медицини.

Ключові слова: мотив, MEME, *k-mer*, машинне навчання, ДНК-послідовність, геном, нейронна мережа, CNN.

Вступ

Секвенування ДНК революціонізувало багато аспектів геноміки та стало незамінним інструментом у виявленні генетичних захворювань. З появою та розвит-

ком технологій машинного навчання з'явилися нові методи для більш точної та швидкої обробки геномних даних, що дозволяє з великою точністю виявляти потенційні патогенні мутації та варіанти спадковості. Використання таких передових методів, як *k-mer* і MEME (Multiple EM for Motif Elicitation) для розпізнавання мотивів, сприяє значному підвищенню якості аналізу ДНК.

Ці техніки дозволяють не лише ефективно розпізнавати і каталогізувати генетичні послідовності, але й адаптувати алгоритми машинного навчання для передбачення можливих функціональних наслідків генетичних змін. Такий інтегрований підхід відкриває нові горизонти для персоналізованої медицини та дозволяє розробляти індивідуальні стратегії лікування та профілактики генетичних захворювань, що є великим кроком уперед для покращення якості життя пацієнтів.

Постановка мети та задач

Метою статті є дослідження можливостей та ефективності застосування методу *k-mer* у задачах встановлення мотивів, зокрема при аналізі класифікації секвенованих ДНК на виявлення генетичних захворювань.

Досягнення визначеної мети передбачає вирішення наступних завдань.

1. Зробити огляд та аналіз існуючих підходів до задач виявлення мотивів у секвенованих ДНК.

2. Розробити архітектуру нейронної мережі, адаптованої для класифікації послідовностей.

3. Провести навчання та тестування розробленої мережі на різних наборах даних для оцінки її здатності до класифікації.

4. Проаналізувати результати та порівняти ефективність методу *k-mer* у задачах встановлення мотивів, зокрема при аналізі класифікації секвенованих ДНК на виявлення генетичних захворювань.

5. Окреслити перспективи застосування методу *k-mer* у різних областях, де потрібний ефективний пошук послідовностей.

Огляд існуючих підходів та літератури

З розвитком біоінформатики та геноміки методи секвенування та аналізу ДНК, спрямовані на виявлення генетичних захворювань, були значно розширені та удосконалені. Однією з фундаментальних технологій, що виникли на початку 90-х років минулого століття, є метод відбору слів *k-mer* [1], який дозволяє структурувати та аналізувати послідовності ДНК, розбиваючи їх на фрагменти фіксованої довжини. Історично *k-mer* використовувався у перших програмах для вирішення задач секвенування та асемблірування геномів, де швидкість та здатність обробляти великі обсяги даних відігравали вирішальну роль.

На відміну від *k-mer* метод MEME аналізує послідовності для ідентифікації статистично значущих біологічних мотивів, що дозволяє глибше зрозуміти потенційні функціональні області в ДНК. Він є ефективним для детального вивчення мотивів, але потребує значно більше часу та ресурсів для обробки порівняно з *k-mer*.

Сучасні алгоритми машинного навчання, такі як глибоке навчання, вносять революційний вклад у біоінформатику. Зокрема нейронні мережі CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks) і LSTM (Long Short-Term Memory) використовуються для класифікації та прогнозування генетичних варіантів [2]. CNN ефективні при виявленні складних патернів у візуальних даних, що робить їх ідеальними для аналізу медичних зображень та генетичних послідовностей, де важливо розпізнавати тонкі відмінності у патернах. RNN та LSTM особливо корисні для обробки послідовних даних, де зв'язки між елементами можуть мати критичне значення для передбачення, наприклад, у визначенні функціональних доменів або регуляторних послідовностей ДНК.

Ці технології разом з методами *k-mer* та MEME забезпечують більш глибоке розуміння генетичних основ захворювань та сприяють розробці ефективніших стратегій для діагностики та лікування, що є основою персоналізованої медицини.

Комбінація цих інструментів та методологій відкриває нові можливості для наукових досліджень та клінічних застосувань, що дозволяє вченим не лише виявляти, але й глибше розуміти складні генетичні взаємозв'язки.

Теоретичні основи

Алгоритм методу *k-mer*. У біоінформатиці одним з основних підходів для аналізу послідовностей ДНК є метод *k-mer*, який базується на розбиванні довгої послідовності ДНК на менші сегменти фіксованої довжини, відомі як *k*-мери. Цей метод дозволяє перетворити складну структуру ДНК на простіші, кількісно аналізовані дані, які можуть бути використані для подальшої обробки алгоритмами машинного навчання. Покроковий алгоритм можна представити таким чином:

— визначення розміру *k*: вибір довжини *k* впливає на результати аналізу, де занадто маленьке значення *k* може призвести до втрати інформації, тоді як занадто велике — до збільшення шуму;

— генерація *k*-мерів: послідовність ДНК сканується від початку до кінця, з виокремленням усіх можливих послідовностей довжиною *k* (наприклад, для послідовності «*AGTCAGTC*» і $k = 4$ *k*-мери будуть: *AGTC, GTCA, TCAG, CAGT, AGTC*);

— кодування *k*-мерів: кожен *k*-мер може бути закодований числовими індексами для спрощення подальшої обробки, звичайно через методи, які враховують частоту його виникнення в датасеті;

— статистичний аналіз: аналізується розподіл *k*-мерів у різних геномних наборах, що може допомогти ідентифікувати статистично значимі відмінності, пов'язані з певними станами або характеристиками [3].

Алгоритм визначення мотивів методом MEME. Складним алгоритмом для ідентифікації мотивів у біополімерних послідовностях, що широко використовується для визначення нових та відомих біологічних мотивів у геномних, транскриптомних чи протеомних даних, є MEME. Покроковий опис має такий вигляд:

— підготовка даних: визначення набору послідовностей, які будуть аналізуватись на наявність мотивів;

— ініціалізація пошуку: вибір початкової моделі мотиву, який може бути випадково згенерованим або заснованим на попередніх знаннях;

— оцінка EM-алгоритмом: використання цього алгоритму для уточнення моделі мотиву з метою оцінювання ймовірності кожного мотиву в кожній послідовності та вдосконалення моделі до стабілізації параметрів;

— рефайнмент моделі: визначення остаточних параметрів мотиву з урахуванням його структури та варіативності у різних послідовностях;

— аналіз та інтерпретація результатів: вивчення характеристик та біологічного значення ідентифікованих мотивів, що може включати порівняння з відомими мотивами та їхнє непотенційне пов'язування з біологічними функціями [4].

Обидва методи, *k-mer* та MEME, відіграють критичну роль у геномній біоінформатиці, оскільки забезпечують потужні інструменти для розуміння складних генетичних інтеракцій та механізмів, що спричиняють захворювання. Їхнє використання в сучасних дослідженнях сприяє більш точному розумінню генетичних основ захворювань та розробці нових методів діагностики та лікування.

Основні методи і результати дослідження

Дослідження базується на секвенуванні ДНК, отриманих від пацієнтів з підозрою на спадкові захворювання, з використанням технології Illumina для зчитування послідовностей.

Застосування методу відбору слів *k-mer* відбувається таким чином, що послідовності ДНК розбиваються на фрагменти (мери) фіксованої довжини *k*, які використовуються для подальшого статистичного аналізу. Мери обираються за формулою

$$\text{Кількість } k\text{-мерів} = N - k + 1,$$

де N — довжина послідовності, а k — довжина меру [5].

Метод MEME залучається для розпізнавання мотивів, оскільки аналізує послідовності щодо наявності статистично значущих мотивів, які можуть бути пов'язані з біологічними функціями. Формула для оцінки значущості мотивів містить розрахунок *E*-значення, яке показує, чи могла дана послідовність з'явитися випадково.

Алгоритми класифікації, такі як Supervised Machine Learning (Навчання з вчителем) та Random Forest (Випадковий ліс), використовуються для розрізнення послідовностей із специфічними захворюваннями та контрольних зразків.

Розглянемо архітектуру нейронної мережі, адаптовану для класифікації послідовностей ДНК, які були оброблені за допомогою підходу *k-mer*. Дана архітектура базується на згорткових нейронних мережах (CNN), які ефективні при роботі з такими даними завдяки своїй здатності виявляти в них локальні патерни [6].

Відповідно, архітектура мережі будується таким чином:

— Вхідний шар (Input layer):

• розмір вхідних даних: залежно від розміру k у підході *k-mer*. Наприклад, якщо $k = 6$, кожен *k-mer* може бути закодований як вектор з шістьох чисел (одне число на символ, *A*, *C*, *G*, *T*, кожне репрезентоване унікальним числом). У конкретному випадку $k = 6$ підібрано емпірично.

— Перший згортковий шар (Convolutional layer 1):

• фільтри: 64;

• розмір ядра: 2;

• функція активації: ReLU (Rectified Linear Unit);

• подовження (Padding): Same, для збереження розміру вихідних даних;

• крок (Stride): 1.

— Шар пулінгу (MaxPooling):

• розмір пулінгу: 1;

• пулінг допомагає зменшити розміри вихідних даних і зробити мережу стійкішою до малих зміщень у вхідних даних.

— Другий згортковий шар (Convolutional layer 2):

• фільтри: 128;

• розмір ядра: 2;

• функція активації: ReLU;

• подовження (Padding): Same;

• крок (Stride): 1.

— Ще один шар MaxPooling:

• розмір пулінгу: 1.

— Третій згортковий шар (Convolutional layer 3):

• фільтри: 256;

• розмір ядра: 3;

• функція активації: ReLU;

• подовження (Padding): Same;

• крок (Stride): 1.

— Ще один шар MaxPooling:

• розмір пулінгу: 1.

— Четвертий згортковий шар (Convolutional layer 4):

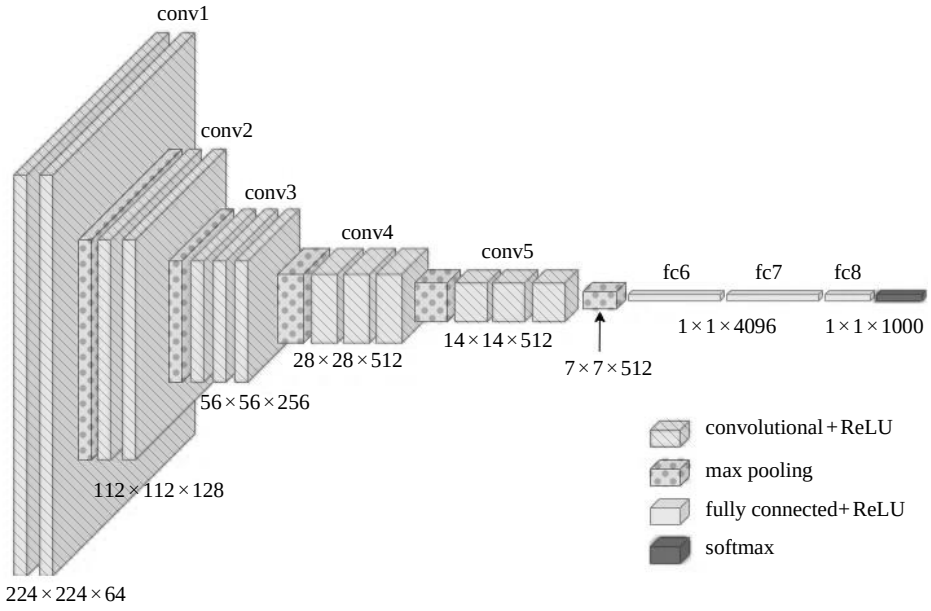
• фільтри: 512;

• розмір ядра: 3;

• функція активації: ReLU;

• подовження (Padding): Same;
• крок (Stride): 1.
— Ще один шар MaxPooling:
• розмір пулінгу: 1.
— П’ятий згортковий шар (Convolutional layer 5):
• фільтри: 512;
• розмір ядра: 3;
• функція активації: ReLU;
• подовження (Padding): Same;
• крок (Stride): 1.
— Ще один шар MaxPooling:
• розмір пулінгу: 1.
— Повнозв’язний шар (Fully connected layer):
• кількість нейронів: 4096;
• функція активації: ReLU.
— Вихідний шар (Output layer):
• кількість нейронів: 2 (для бінарної класифікації: виявлено захворювання чи ні);
• функція активації: Softmax (для бінарної класифікації також можна використовувати Sigmoid).

На рисунку наведена автоматично згенерована ілюстрація архітектури нейронної мережі.



Використано наступний підхід до тренування:

— оптимізатор: Adam (Adaptive Moment Estimation), завдяки його адаптивним властивостям у коригуванні швидкості навчання;

— функція втрат: Cross-entropy, що є стандартом для задач класифікації;

— розмір пакету (Batch size): зазвичай 32 або 64, залежно від обчислювальних ресурсів;

— кількість епох: 20–50, також залежить від доступних даних.

Ця архітектура може бути адаптована або змінена залежно від специфіки задачі та доступних даних.

Усі дані для дослідження надані U.S. National Library of Medicine та застосовуються з дозволу учасників, які погодилися з використанням їх ДНК у наукових дослідженнях [7]. Вибірки ДНК секвеновані та подані у форматі FASTA,

що дозволяє зберігати інформацію у текстовому форматі, використовуючи символи, які репрезентують нуклеотидні послідовності. Кожна з цих послідовностей у файлі починається з описової стрічки та відділяється від неї символом «>». Для їхньої попередньої обробки використовувалися алгоритми *k-mer* та MEME. Перший вживається для розбиття довгих послідовностей ДНК на коротші фрагменти фіксованої довжини (мери) [8], що полегшує подальший аналіз і класифікацію. Водночас для ідентифікації та аналізу повторюваних мотивів або патернів у цих послідовностях, що мають біологічне значення, застосовується другий метод (MEME).

Наведемо приклад послідовності у FASTA-форматі.

>Sample_1

CACATGATAAGAAAATGAAAGAAATGATATGAT…

Цей формат містить ідентифікатор зразка та саму нуклеотидну послідовність, яка репрезентована літерами, що вказують на конкретні нуклеотиди (*A* — аденін, *T* — тимін, *C* — цитозин, *G* — гуанін). Послідовності, оброблені за допомогою *k-mer* і MEME, надають важливу інформацію, яка використовується для тренування нейронних мереж, з метою вдосконалення процесів ідентифікації та класифікації генетичних захворювань.

Методологія передбачає ретельне тренування моделей машинного навчання на наборах даних, підготовлених за допомогою обох методів попередньої обробки, з наступним порівняльним аналізом на основі коректності, точності, похибки та оцінки *F*1 [9].

Згідно з цими параметрами отримано результати методу *k-mer*, де коректність становить 98,6 %, точність — 98,5 %, похибка — 1,4 %, оцінка *F*1 — 0,984.

У разі методу MEME встановлено результати, які мають коректність 89,2 %, точність — 87,5 %, похибку — 10,8 % та оцінку *F*1 — 0,951.

Метод *k-mer* стабільно перевершував підхід MEME за всіма показниками і додатково продемонстрував кращі показники: акуратність (98,6 %) та точність (98,5 %), порівняно з останнім (акуратність — 89,2 % та точність — 87,5 %).

Аналіз підтверджує, що метод *k-mer* є більш ефективним та швидшим у виявленні генетичних захворювань. MEME, хоч і є потужним у розпізнаванні мотивів, вимагає значно більше часу для обробки та має нижчу точність.

Висновок

Метод відбору слів *k-mer* продемонстрував високу точність та ефективність у виявленні специфічних генетичних захворювань, що робить його ідеальним кандидатом для інтеграції у клінічні системи. Його використання може значно прискорити процес діагностики, що дозволить медичним фахівцям швидше виявляти та реагувати на генетичні аномалії, які можуть вплинути на здоров’я пацієнтів. Цей метод надає можливість автоматизувати та оптимізувати обробку великих обсягів геномних даних із забезпеченням високої точності визначення потенційних патогенних варіацій.

Окрім цього, метод відбору слів *k-mer* може сприяти розвитку персоналізованої медицини, оскільки дозволяє точно адаптувати лікувальні підходи до генетичного профілю кожного пацієнта. Інтеграція цього методу у клінічні практики не лише покращить якість лікування, але й сприятиме більш швидкому впровадженню інноваційних терапевтичних методів, базованих на генетичному аналізі.

THE *K-MER* METHOD IN TASKS OF IDENTIFYING REGULAR SEQUENCES

Yehor Terpilovskyi

V.M. Glushkov Institute of Cybernetics of NAS, Kyiv,

terpilovskiy.egor@gmail.com

In this study, we compare two methodologies for preprocessing human DNA sequences to improve the identification of specific genetic diseases using machine learning techniques. The first approach involves *k-mer* word sampling, while the second uses Motif Elicitation (MEME) for motif recognition. The *k-mer* method involves dividing the DNA sequence into smaller fragments of a fixed length, which allows structuring and analyzing large volumes of data efficiently. MEME, on the other hand, applies an expectation maximization (EM) algorithm to detect statistically significant biological motifs in sequences, which allows for a deeper understanding of functional DNA regions. Our comprehensive analysis includes training a machine learning model on data samples, accuracy scores and other performance metrics, as well as considerations for the practical implementation of both methods. Data for this study were provided by the U.S. National Library of Medicine and presented in the FASTA format, which provides a standardized representation of nucleotide sequences. Each DNA sample belongs to people who have given consent for their genetic data to be used in scientific research. The data were processed with both *k-mer* and MEME to ensure a comprehensive analysis. The *k-mer* method allows efficient processing of large volumes of genetic data and is compatible with various machine learning algorithms. On the other hand, MEME is a powerful tool for motif recognition, but requires significant computational resources and time for analysis. A comparison of these methods showed that *k-mer* has advantages in speed and efficiency, which makes it more suitable for practical use in clinical settings. The main goal of our research is to find out the advantages of the *k-mer* method over MEME in the context of identifying genetic diseases by genomic sequences. The results of our study show that the *k-mer* method provides high accuracy and efficiency, which makes it more suitable for integration into clinical systems for faster diagnosis. The conclusions of our research can contribute to the improvement of approaches to genetic diagnostics and the development of personalized medicine.

Keywords: motif, MEME, *k-mer*, machine learning, DNA sequence, genome, neural network, CNN, FASTA.

ПОСИЛАННЯ

1. Jain A.K., Murty M.N., Flynn P.J. Data clustering: a review. *ACM computing surveys (CSUR)*. 1999. Vol. 31, N 3. P. 264–323. DOI: https://doi.org/10.1145/331499.331504
2. Xu R., Wunsch D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*. 2005. Vol. 16, N 3. P. 645–678. DOI: https://doi.org/10.1109/tnn.2005.845141
3. Alshayeji M., Sindhu S.C., Abed S. Viral genome prediction from raw human DNA sequence sampling. *Expert System with Applications*. 2023. Vol. 218, N 3. DOI: https://doi.org/10.1016/j.eswa.2023.119641
4. Zhang Y., Wang P., Yan, M. An entropy-based position projection algorithm for motif discovery. *BioMed Research International*. 2016. P. 1–11. DOI: https://doi.org/10.1155/2016/9127474
5. Dunn J.C. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*. 1974. Vol. 4, N 1. P. 95–104. DOI: https://doi.org/10.1080/01969727408546059
6. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015. Vol. 521 (7553). P. 436–444. DOI: https://doi.org/10.1038/nature14539
7. Goodfellow I., Bengio Y., Courville A. Deep learning. Cambridge : The MIT Press, 2016. 800 p. DOI: https://doi.org/10.1007/s10710-017-9314-z
8. Krizhevsky A., Sutskever I., Hinton G.E. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 2012. Vol. 25. P. 1097–1105. DOI: https://doi.org/10.1145/3065386
9. Hinton G.E., Osindero S., Teh Y.W. A fast learning algorithm for deep belief nets. *Neural Computation*. 2006. Vol. 18 (7). P. 1527–1554. DOI: https://doi.org/10.1162/neco.2006.18.7.1527

Отримано 16.04.2024
Доопрацьовано 09.05.2024