УДК 004.056.53

DOI: 10.20535/2523-4455.mea.217265

Моделювання та аналіз сигналів біонанопорового секвенування ДНК для виявлення генетичних мутацій

Євдощенко[f] І. М., ORCID 0000-0003-0049-2159
Іванько К. О., к.т.н. доц., ORCID 0000-0002-3842-2423
Іванушкіна Н. Г., к.т.н. доц., ORCID 0000-0001-8389-7906
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського" ROR 00syn5v21
Київ, Україна

Вішвеш Кулкарні, доц., ORCID https://orcid.org/0000-0002-2285-8652
Інженерна школа, Університет Ворика ROR 01a77tt86
м. Ворик, Сполучене Королівство Великої Британії та Північної Ірландії

***Анотація*—Робота присвячена розвитку методів цифрової обробки геномних сигналів, які представляють собою дані щодо будови ДНК, з метою використання методів обробки сигналів до задачі аналізу геномних даних. За фрагментами послідовностей нуклеотидів змодельовано сигнали іонного струму крізь біологічну нанопору при секвенції ДНК для випадків норми, точкових мутацій, вставки та видалення ділянки ДНК. Модельні сигнали іонного струму у білковій нанопорі отримано на основі реальних послідовностей нуклеотидів з атласів ракового геному. В роботі використано кореляційний аналіз для визначення подібності сигналів нанопорового секвенування ДНК за допомогою функції взаємної кореляції між двома сигналами іонного струму крізь білкову нанопору, зокрема між сигналами у нормі та з наявністю мутації. За розташуванням максимуму взаємної кореляційної функції визначається тип мутації (інсерція або делеція), а також проводиться вирівнювання однакових нуклеотидних послідовностей за допомогою визначеного зсуву сигналу.**

Проаналізовано застосування методів машинного навчання до класифікації геномних сигналів нанопорового секвенування ДНК. Для визначення найкращих моделей класифікації застосовано алгоритми на основі дерев рішень, дискримінантного аналізу, методу опорних векторів, логістичної регресії, методу k-найближчих сусідів та ансамблевого навчання. Для різних методів машинного навчання визначено та порівняно точність класифікації на 4 класи: норма, точкова мутація (місенс або нонсенс), мутація делеції та інсерції декількох нуклеотидів. Показано, що результати застосування методів машинного навчання до проблеми класифікації сигналів нанопорового секвенування ДНК суттєво залежать від рівня шуму у зареєстрованих сигналах іонного струму крізь білкову нанопору та типу мутації. Найкращі результати класифікації отримано для методу опорних векторів. Застосування лінійної, квадратичної та кубічної функцій ядра показало високу точність вірно класифікованих сигналів – від 93 до 100%.

Ключові слова — секвенація ДНК; білкова нанопора; мутації; обробка геномних сигналів; класифікація; машинне навчання.

I. Вступ

Дезоксирибонуклеїнова кислота (ДНК) — це макромолекула, що забезпечує зберігання, передачу і реалізацію генетичної програми розвитку та функціонування живих організмів. Молекула ДНК зберігає біологічну інформацію у вигляді генетичного коду, що складається з послідовності нуклеотидів. Секвенування ДНК — це процес визначення послідовності нуклеотидів в ДНК, що включає в себе методи, які використовуються для визначення порядку слідування у молекулі ДНК чотирьох азотистих основ: аденіну (А), гуаніну (Г), цитозину (Ц) і тиміну (Т) [1-4].

В останні роки набувають розвитку методи обробки геномних сигналів, які перетворюють дані щодо будови ДНК у числові значення з метою використання існуючих методів цифрової обробки сигналів для геномних даних [5-13]. Все більшої важливості набуває застосування методів обробки геномних сигналів до задачі виявлення та прогнозування раку, руйнівної хвороби з величезною захворюваністю та смертністю [8, 9, 12]. Рак – це генетичне захворювання, ключову роль в розвитку якого грає пошкодження ДНК клітин. Метою даної роботи є розробка алгоритмів для моделювання і аналізу сигналів нанопорового секвенування ДНК для дослідження генетичних змін, що призводять до раку. Важливим є визначення початкової стадії переходу між нормальною та раковою структурами ДНК, а також пошук ознак для прогнозування розвитку раку та його виявлення на

DOI: 10.20535/2523-4455.mea.217265

ранній стадії.

Для виявлення мутацій в генах проводяться молекулярно-генетичний аналіз мутацій методом полімеразної ланцюгової реакції (ПЛР) та молекулярно-генетичне дослідження генів методом визначення нуклеотидної послідовності ДНК за допомогою секвенування. Секвенування дозволяє виявити різноманітні зміни в певних ділянках хромосом та визначити наявність та вид мутацій.

Під час нанопорового секвенування відбувається зчитування нуклеотидного складу безпосередньо аналізованої молекули ДНК, що приводить до спрощення процесу, бо секвенування не потребує стадії підготовки проби (ПЦР, мічення зразка і т.д.). Принцип нанопорового секвенування заключається в переміщенні молекули ДНК крізь нанопору і реєстрації змін в оточуючому електричному полі в процесі руху молекули [1-3].

Для класифікації раку та розділення геномних зразків у нормі та за наявності генетичних змін у роботі [6] дослідники використовують методи кластеризації, наприклад, ієрархічну кластеризацію та метод *K*-середніх. Для дослідження пар генів подібність між зразками у [15] визначається на основі порогового коефіцієнта кореляції. У джерелах [7, 14], присвячених методам обробки геномних сигналів, запропоновано застосування методів машинного навчання, таких як *K*-найближчих сусідів, метод опорних векторів, нейронних мереж та ін. Автори дослідження [9] припускають, що закономірності в спектрі власних значень мають перспективу прогнозу раку на ранніх стадіях. У роботах [10-12] до обробки геномних сигналів застосовано частотний, частотно-часовий та вейвлет- аналіз. Отримані дослідниками результати точності розпізнавання ракових зразків відрізняються для різних видів раку (шкіри, легень, печінки, простати і т.д.) і застосованих методів обробки та сягають до 85-98% вірно класифікованих зразків. Однак напрямок обробки геномних сигналів потребує подальших досліджень та вдосконалень.

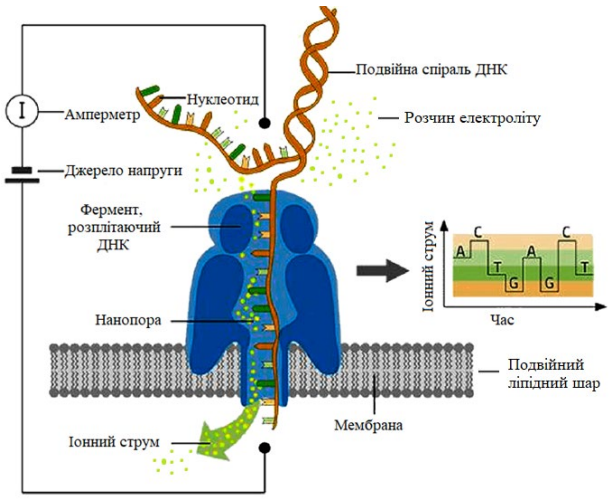


Рис. 1 Принцип нанопорового секвенування ДНК (модифіковано з [4])

II. Моделювання та аналіз сигналів нанопорового секвенування ДНК

Пристрій для біонанопорового секвенування ДНК складається з α-гемолізинового білкового каналу, введеного в подвійний ліпідний шар, який відокремлює дві камери, що містять розчин електроліту KCl [1-4] (Рис.1). В технологіях типу Oxford Nanopore використовуються пороутворюючі білки, що створюють пори в мембранах [1]. Білок α-гемолізин і подібні йому пороутворюючі білки існують в клітинних мембранах, де вони функціонують в якості каналів для переносу іонів і молекул як у клітину, так і в навколишнє середовище. Білок α-гемолізин представляє собою гептамірну білкову пору з внутрішнім діаметром порядка нм одиниць, що близько до розмірів одиничних біологічних молекул, таких як ДНК та РНК.

Напруга прикладається до сторін подвійного ліпідного шару, створюючи іонний струм крізь нанопору, який реєструється для подальшого аналізу. Під дією електрорушійної сили негативно заряджена молекула ДНК дифундує до пори, втягується в канал і проходить крізь канал до позитивно зарядженої сторони подвійного ліпідного шару. Коли молекула ДНК захоплюється і рухається крізь нанопору, провідність каналу G_c зменшується, викликаючи падіння іонного струму, який вимірюється. Ця зміна провідності і, як наслідок, зміна струму та його тривалості використовуються для характеристики захопленої нанопорою молекули ДНК. Секвенування ДНК за допомогою нанопори використовується з метою ідентифікації послідовності нуклеотидів ДНК на основі змін іонного струму крізь нанопору.

Патч-кламп — метод електрофізіології, який дозволяє ізолювати фрагмент клітинної мембрани з іонними каналами, задавати необхідну різницю потенціалів крізь цей фрагмент клітинної мембрани, створювати по обидві сторони мембрани середовище з певним іонним складом і вимірювати в цих контрольованих умовах струм навіть одиничних іонних каналів [2, 16]. В експерименті з постійною керуючою напругою V_c зміна іонного струму I_i свідчить про зміну провідності каналу G_c. $V_c = I_i/G_c$ - таким чином, про зміщення значень провідності G_c сигналізує зміна іонного струму I_i.

Діаметр нанопори складає декілька нанометрів, через що молекула ДНК здатна проходити крізь нанопору тільки в одноланцюговій формі. Під час проходження молекули нуклеїнової кислоти крізь нанопору окремі нуклеотиди затримуються в певних сайтах всередині пори, затримуючи проходження іонів електроліту, в результаті чого відбувається падіння струму. Струм крізь нанопору та провідність каналу пропорційні розміру нуклеотиду, який проходить в даний момент, що викликано різною будовою азотистих основ ДНК — аденіну, гуаніну, тиміну і цитозину.

Аналіз біологічних послідовностей вимагає обробки сигналів нанопорового секвенування ДНК з метою усунення шуму та знаходження однакових сегментів у послідовностях, а також ділянок з мутаціями. Крім того знаходження однакових послі-

DOI: 10.20535/2523-4455.mea.217265

довностей є важливим для відстеження співпадінь геномних фрагментів у базах даних. Важливою задачею для розуміння мутацій є вирівнювання однакових біологічних послідовностей та виявлення ділянок з мутаціями та визначення типів цих мутацій. Вирівнювання послідовностей ДНК — це паралельне розташування первинних послідовностей ДНК або амінокислот різних зразків з метою виявлення областей подібності між ними. Це допомогає виявити зв'язки між послідовностями. Таким чином, вирівнювання послідовностей ДНК зазвичай виконується для того, щоб визначити функції та структуру нових секвенованих генів шляхом порівняння з геном, функція якого відома. Результатом такого порівняння може бути визначення еволюційного зв’язку між послідовностями внаслідок мутацій або локально однакових областей у віддалених послідовностях. Таке порівняння може привести до виявлення точної або близької відповідності між послідовністю секвенованих нуклеотидів і послідовностями, які вже досліджені і зберігаються в базах даних. Точна відповідність ділянок може спостерігатися по всій довжині послідовності нуклеотидів (глобальна відповідність) або для менших сегментів послідовності нуклеотидів (локальна відповідність).

З метою розробки алгоритму знаходження мутацій у ДНК тканин, уражених раком, в даній роботі використано фрагменти послідовностей нуклеотидів з наступних баз даних (Рис.2):

- база даних соматичних мутацій раку COSMIC (catalogue of somatic mutation in cancer) [17];
- база даних Національного центру біологічної інформації NCBI (National Center for Biology Information) [18];
- база даних програми Атлас ракового геному TCGA (Cancer Genome Atlas Program) [19, 20].

Для дослідження можливостей застосування методів обробки геномних сигналів було обрано нуклеотидні послідовності довжиною 45 нуклеотидів в нормі і з наявністю різних форм мутацій, таких як:

1) місенс-мутація (точкова мутація, де відбувається заміна нуклеотиду, що кодує іншу амінокислоту),
2) нонсенс-мутація (точкова мутація, де кодується стоп – білок),
3) мутація інсерції (вставка ділянки ДНК),
4) мутація делеції (видалення ділянки ДНК).

Приклад послідовностей нуклеотидів, які використовувалися для подальшого порівняння, знаходження мутацій та вирівнювання однакових фрагментів у послідовностях:

Норма:
GGAGGGGAGAAACTTGCCACTTTGGCTGAGTTG
GTCCAGTATTAC

Місенс-мутація:
GGAGGGGAGAAA**T**TTGCCACTTTGGCTGAGTTG
GTCCAGTATTAC

Нонсенс-мутація:
GGAGGG**T**AGAAACTTGCCACTTTGGCTGAGTTG
GTCCAGTATTAC

Мутація інсерції (вставлено фрагмент TGC):
GGA**TGC**GGGGAGAAACTTGCCACTTTGGCTGA
GTTGGTCCAGTAT

Мутація делеції (видалено фрагмент GGG):
GGA|видалено фрагмент|GAGAAACTTGCCACTTTGGCTGAGTTGGTC
CAGTATTACGGG

З метою застосування методів аналізу сигналів до задачі знаходження мутацій дані послідовності нуклеотидів було використано для моделювання сигналів, отриманих за допомогою нанопорового секвенування (Рис.3-4). Проходження кожного нуклеотида крізь білкову нанопору змінює значення рівня іонного струму. Негативно заряджена одноланцюгова ДНК проходить крізь нанопору в мембрані, зовнішня поверхня якої має негативний заряд, а внутрішня — позитивний. Як тільки черговий нуклеотид перекриває внутрішній отвір в нанопорі, електропровідність мембрани (струм мембрани) змінюється в залежності від хімічної будови нуклеотиду. Нуклеотиди різного типу, які складають ланцюжок ДНК, відрізняються хімічними та фізичними властивостями. Тому ступінь блокування нанопори кожним нуклеотидом різна. Відповідно до цього, змінюються електрична провідність та струм мембрани. Під час проходження ДНК крізь нанопору іонний струм значно зменшуться.

Рис. 2 Фрагмент послідовності нуклеотидів з ділянкою, на якій виникли мутації (база даних COSMIC [17])

DOI: 10.20535/2523-4455.mea.217265

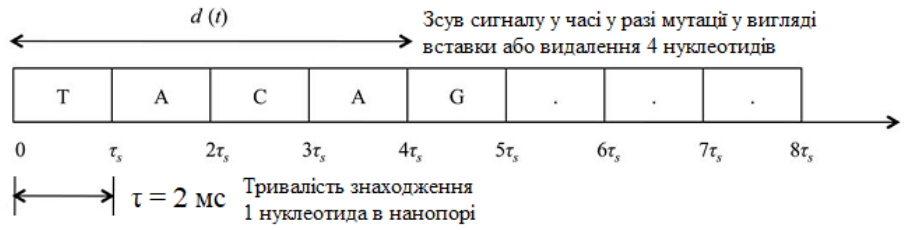
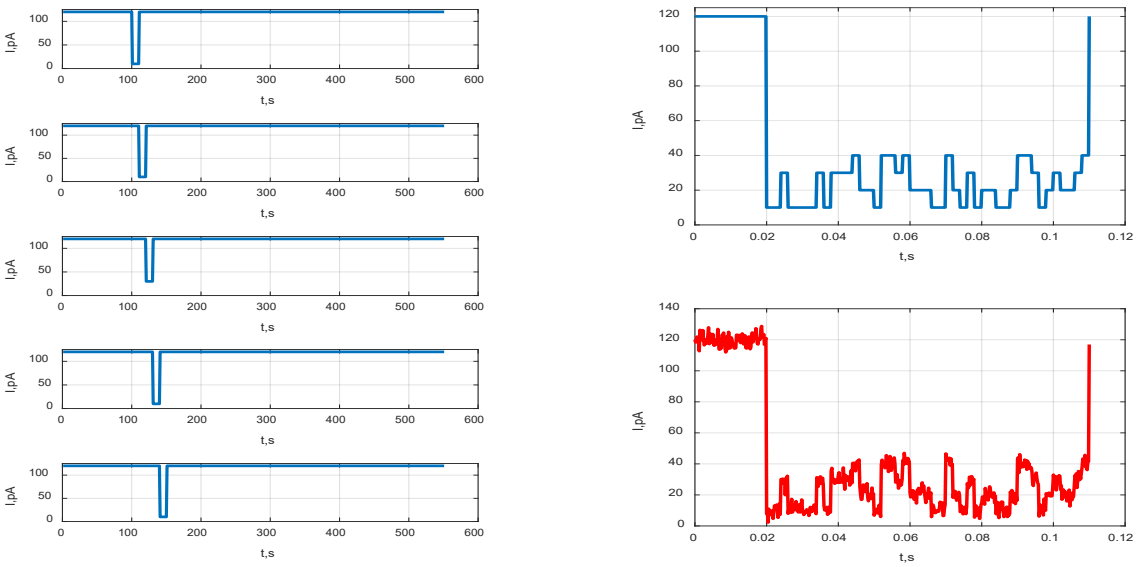


Рис. 3 Пояснення до моделювання сигналу струму крізь нанопору під час протягування одноланцюгової ДНК з мутаціями інсерції або делеції



а) б)

Рис. 4 Моделювання сходинок струму крізь нанопору під час протягування одноланцюгової ДНК: а) послідовне блокування нанопори 5-ти нуклеотидами GGAGG; б) сумарний сигнал під час послідовного блокування нанопори 45 нуклеотидами GGAGGGGAGAAACTTGCCACTTTGGCTGAGTTGGTCCAGTATTAC (зверху – без шума, знизу – з наявністю білого шуму)

При моделюванні сигналу струму крізь нанопору під час протягування одноланцюгової ДНК було використано наступні параметри, які відповідають реальним патч-кламп установкам у режимі фіксації напруги [1-3, 16]:

- частота дискретизації сигналу 5 кГц;
- швидкість протягування ДНК крізь нанопору – 500 нуклеотидів/с;
- у разі прикладеної напруги V_c=120 мВ та провідності каналу G_c=1 нСм, іонний струм крізь відкриту нанопору складає 120 pA;
- середній струм іонів крізь нанопору у випадку перекриття пори гуаніном складає 10 пА, у випадку перекриття тіміном – 20 пА, у випадку перекриття аденіном – 30 пА, у випадку перекриття цитозином – 40 пА;
- тривалість знаходження нуклеотиду в порі складає 2 мс.

Загалом сигнал, при проходженні крізь пору 45 нуклеотидів, за умови Fs=5 кГц, має тривалість 90 мс та кількість відліків N=450. Час дискретизації сигналу дорівнював 0,2 мс. Знаходження певного нуклеотиду в нанопорі моделювалося як сходинка струму іонного каналу тривалістю τ = 2 мс (десять відліків сигналу), що відображає час блокування нанопори 1 нуклеотидом (Рис. 4 а).

Серед параметрів моделі струму крізь нанопору у разі проходження 1 нуклеотиду слід відзначити наступні:

- d_{st} – час входження ланцюга ДНК до каналу нанопори та початку секвенування;
- τ – тривалість знаходження 1 нуклеотида в порі;
- $c_{st}=d_{st}+\tau/2$ – часове розташування центру сходинки іонного струму крізь нанопору;
- I_{fr} – струм іонів крізь вільну нанопору (без захопленої ДНК);
- A – амплітуда прямокутного імпульсу, яка відповідає за зменшення струму крізь нанопору за умови знаходження в ній певного нуклеотиду.

З метою отримання реалістичного сигналу струму крізь нанопору під час протягування одноланцюгової ДНК до отриманого сходинкового сигналу необхідно додати шумові складові (Рис. 4, б). Таким чином,

DOI: 10.20535/2523-4455.mea.217265

було змодельовано низку сигналів в нормі та при наявності мутацій.

Задача вирівнювання послідовностей нуклеотидів полягає у пошуку областей подібності між двома або більше послідовностями. В даній роботі пропонується застосування кореляційного аналізу для визначення подібності за допомогою кореляційної функції між двома сигналами іонного струму у нанопорі (в нормі та з наявністю мутації ДНК).

Взаємна кореляційна функція визначається як

$$C(\tau) = \int_{-\infty}^{\infty} g(t)p^*(t-\tau)dt,$$

де $g(t)$ та $p(t)$ представляють сигнали нанопорового секвенування ДНК, які порівнюються, * - оператор комплексного спряження [15]. Це добуток між двома послідовностями сигналів струму іонного каналу, одна з яких зміщена відносно іншої на τ секунд часу. У випадку дискретного сигналу $\tau=n*F_s$ в діапазоні від негативних до позитивних значень n, кількості відліків зсуву сигналів один відносно іншого. Для визначення кількості змінених у результаті мутації нуклеотидів знаходився пік кореляційної функції та його зсув відносно початку координат (у відліках та за часом). По значенню зсуву, тобто позитивний він чи від'ємний, визначалося, в якому напрямку потрібно зсунути сигнал для накладання однакових послідовностей і який тип мутації виник - інсерція чи делеція. У випадку мутації інсерції спостерігався зсув піку взаємної кореляційної функції вправо, оскільки відбувалася вставка нуклеотидів, для делеції пік зсувався вліво, бо відбулося видалення нуклеотидів (рис. 5 а). Для заданих в даному дослідженні параметрів відстань між відліками за часом складала 0,2 мс. Тоді для мутації інсерції, де присутня вставка 3 нуклеотидів, тобто зсув у 30 відліків або за часом 0,006 с, маємо результат, наведений на рис. 5 б.

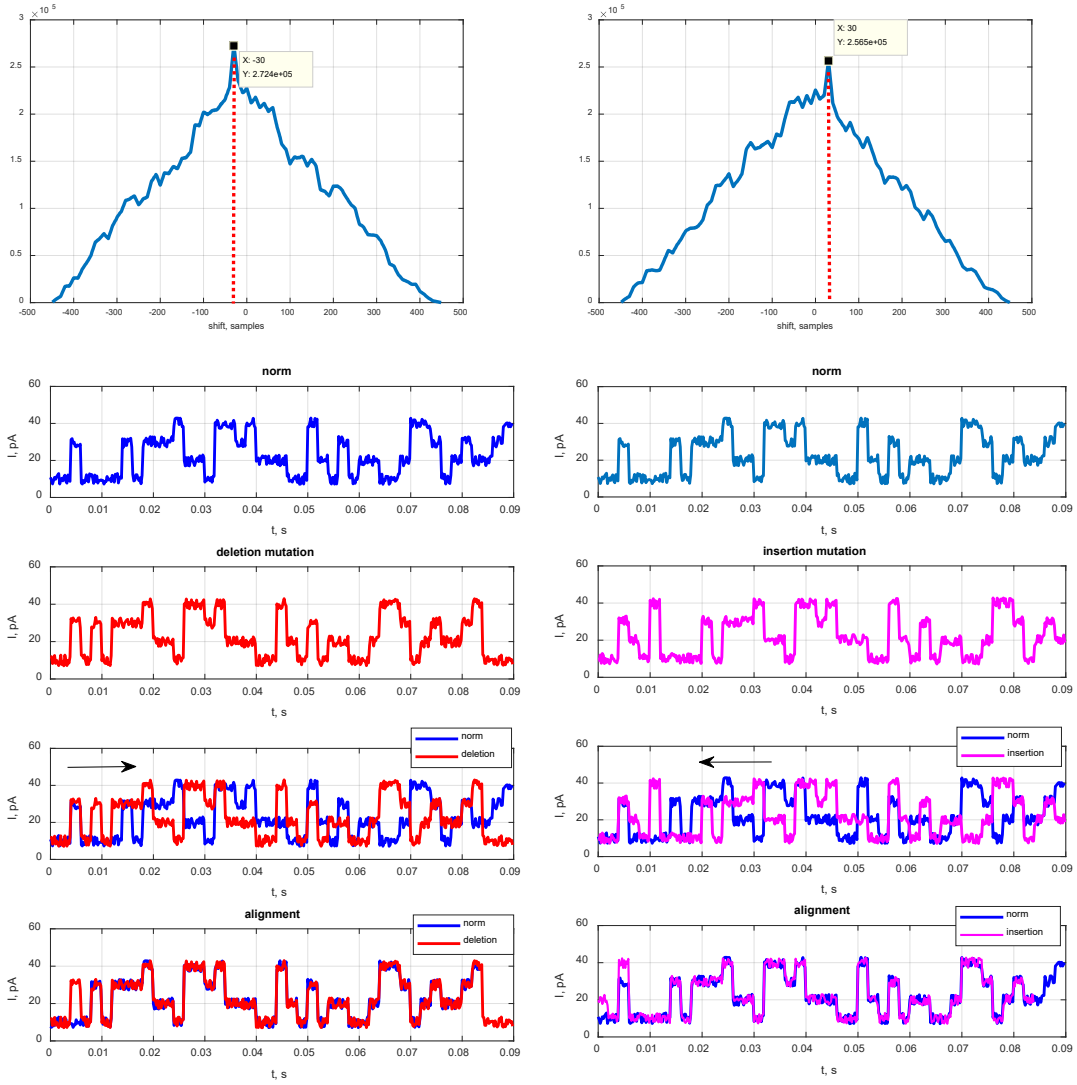


Рис. 5 Результати кореляційного аналізу: а) взаємна кореляційна функція між сигналом іонного струму для послідовності нуклеотидів в нормі та сигналом для послідовності нуклеотидів з мутацією делеції 3х нуклеотидів - спостерігаємо зсув піку на 30 відліків вліво від початку координат; б) відповідні сигнали до рисунку а) та накладання послідовностей однакових нуклеотидів; в) взаємна кореляційна функція між сигналом іонного струму для послідовності нуклеотидів в нормі та сигналом для послідовності нуклеотидів з мутацією інсерції 3х нуклеотидів - спостерігаємо зсув піку на 30 відліків вправо від початку координат; г) відповідні сигнали до рисунку в) та накладання послідовностей однакових нуклеотидів

DOI: 10.20535/2523-4455.mea.217265

III. Застосування методів машинного навчання для класифікації сигналів нанопорового секвенування ДНК

Різниця між онкологічними та нормальними зразками ДНК в області сигналів іонного струму може бути виявлена методами машинного навчання, хоча у багатьох випадках може все ще бракувати засобів для біологічної інтерпретації цих відмінностей.

З метою дослідження можливості застосування алгоритмів машинного навчання до задачі класифікації сигналів нанопорового секвенування було змодельовано по 300 сигналів відрізку послідовності 45 нуклеотидів для 4х класів: 1) норма; 2) місенс-мутація у вигляді зміни 1 нуклеотиду; 3) мутація делеції – видалення фрагменту з трьох нуклеотидів; 4) мутація інсерції – вставка зайвого фрагменту з трьох нуклеотидів. До змодельованих сигналів кожного класу додавався адитивний гаусівський шум.

Використання методу кластеризації *K*-середніх передбачає, що зразки формують *K* кластерів даних, де кількість кластерів *K* є заздалегідь визначеною. Алгоритм ініціалізує *K* центрів, по одному для кожного кластера, і кожен зразок приписується одному з кластерів на основі його схожості з *K* центрами цих кластерів. Потім центри оновлюються як середнє значення зразків у кожному кластері. Процес між присвоєнням окремих зразків кластерам та оновленням кластерних центрів є ітераційним. Етапи зміни центрів кластерів і перепризначення зразків ітераційно повторюються до тих пір, доки границі кластерів і розташування їх центрів не перестануть змінюватися, тобто на кожній ітерації в кожен кластер будуть потрапляти одні і ті самі дані. Таким чином, методи кластеризації здатні групувати гени з подібними послідовностями нуклеотидів.

Для реалізації алгоритмів кластеризації необхідно задати кількість кластерів і це може становити головну проблему. Оптимальну кількість кластерів складно визначити універсально для різних застосувань, оскільки це залежить від різних захворювань, мутацій, які їм відповідають, та від різних наборів досліджуваних генів. У нашому модельному дослідженні кількість кластерів заздалегідь відома і складає 4, що відповідає ділянкам сигналу секвенованої ДНК
у нормі та при наявності 3х видів мутацій (місенс-мутація, інсерція, делеція). При кластеризації даних за методом k-середніх, який мінімізує сумарне квадратичне відхилення точок кластерів від центрів цих кластерів у багатовимірному просторі ознак, сигнали розділилися на кластери, які відповідали нормі, місенс-мутації, мутацій делеції та інсерції трьох нуклеотидів (Рис.6).

Алгоритм *k*-найближчих сусідів (kNN), є одним з найпростіших алгоритмів машинного навчання. Побудова моделі полягає в запам'ятовуванні навчального набору даних. Для того, щоб зробити прогноз для нового зразку (ділянки сигналу секвенованої ДНК), алгоритм *k*-найближчих сусідів знаходить найближчі до неї дані навчального набору, тобто знаходить її «найближчих сусідів». Враховуючи навчальні зразки з відомим класом та немічені зразки для тестування, кожному зразку з тестової вибірки присвоюється мітка класу на основі більшості голосів *k* найбільш близьких навчальних зразків.

У найпростішому варіанті алгоритм *k*-найближчих сусідів розглядає лише одного найближчого сусіда - точку навчального набору, найближче розташовану до точки, для якої потрібно отримати прогноз. Але можна розглянути будь-яку кількість (*k*) сусідів. Коли приймаємо до уваги більше одного сусіда, для присвоєння мітки використовується клас, який найбільш часто зустрічається, тобто обирається клас, який набрав більшість серед *k*-найближчих сусідів. У разі мультикласової класифікації, як у нашому випадку з 4 класами, підраховується кількість сусідів, що належать до кожного класу, і прогнозується клас, який найбільш часто зустрічається. В такому класифікаторі є два важливі параметри: кількість сусідів і міра відстані між зразками даних. У даній роботі порівнювалося використання у якості параметрів кількості сусідів *k*=1 та *k*=10, евклідової відстані та відстані Мінковського (Табл.1).

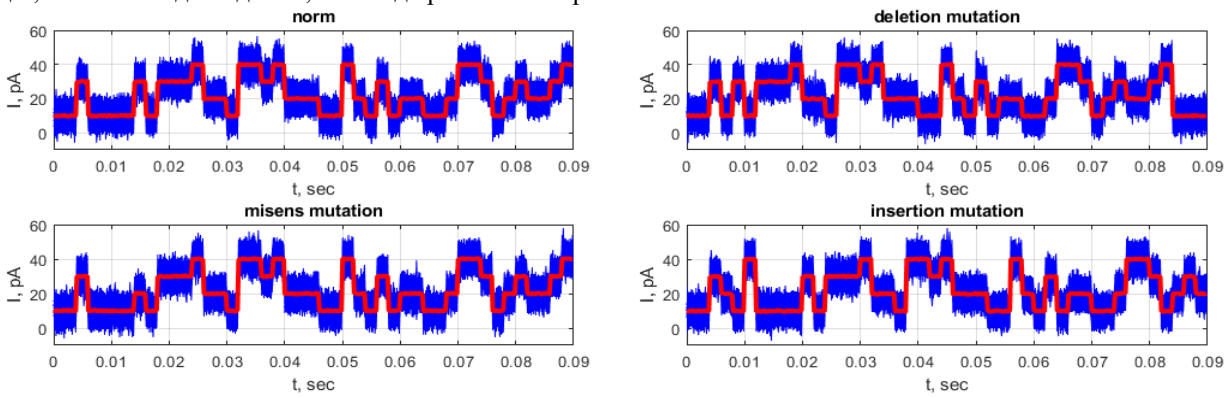


Рис.6. Сигнали нанопорового секвенування, які відповідають 4 кластерам, отриманим за методом k-середніх (норма, місенс-мутація, мутація делеції та інсерція трьох нуклеотидів)

DOI: 10.20535/2523-4455.mea.217265

Таблиця 1 – Точність розпізнавання модельних сигналів нанопорового секвенування ДНК в залежності від відношення сигнал/шум (SNR) та методу машинного навчання (% вірно визначених сигналів при класифікації на 4 класи, у дужках послідовно вказано % вірно визначених випадків для класів “мутація інсерції”, “мутація делеції, “мутація місенс” та “норма”)

SNR	Метод машинного навчання							
	Лінійний дискримінант	***Метод k-найближчих сусідів (10 сусідів, евклідова відстань)***	***Метод k-найближчих сусідів (1 сусід, евклідова відстань)***	***Метод k-найближчих сусідів (10 сусідів, відстань Мінковського)***	***Зважений метод k-най-ближчих сусідів***	***Ансамбль підпросторів k-найближчих сусідів***	***Метод опорних векторів (лінійна, квадратична, кубічна функція ядра)***	***Дерева рішень***
15	100	98,8 (100, 100, 100, 99)	97,8 (100, 100, 96, 96)	98,8 (100, 100, 99, 96)	99,8 (100, 100, 100, 99)	100	100; 100; 100	99,3 (100, 100, 98, 99)
13	100	99,0 (100, 100, 100, 99)	96,6 (100, 100, 96, 90)	97,3 (100, 100, 99, 90)	99,8 (100, 100, 100, 99)	100	100; 100; 100	99,3 (100, 100, 98, 99)
10	99,8 (100, 100, 99, 99)	99,3 100, 100, 99, 97)	93,6 (100, 100, 85, 90)	97,0 (100, 100, 99, 89)	99,2 (100, 100, 99, 98)	99,8 (100, 100, 100, 99)	100; 100; 100	97,2 (99, 99, 95, 94)
8	99,2 (100, 100, 98, 98)	97,4 (100, 100, 98, 91)	90,8 (100, 100, 79, 84)	95,1 (100, 100, 96, 85)	97,8 (100, 100, 96, 95)	99,2 (100, 100, 99, 98)	99,8 (100, 100, 99, 100); 99,9 (100, 100, 99, 100); 99,9 (100, 100, 99, 100)	93,1 (99, 100, 86, 87)
4	91,2 (100, 100, 83, 81)	86,4 (100, 100, 76, 70)	80,3 (100, 100, 54, 68)	85,8 (100, 100, 77, 66)	85,8 (100, 100, 63, 80)	87,7 (100, 100, 69, 82)	93,9 (100, 100, 88, 88); 92,7 (100, 100, 87, 84); 93,3 (100, 100, 88, 85)	83,0 (96, 95, 70, 71)

Метод опорних векторів є потужним методом машинного навчання, який показав гарні результати у багатьох біомедичних застосуваннях [5, 7, 21, 22]. Використовуючи набір навчальних даних (наприклад, профілі експресії генів суб’єктів у нормі та з наявністю онкології), метод опорних векторів знаходить гіперплощину, яка найкраще відокремлює два класи навчальних даних. Такою гіперплощиною є гранична гіперплощина, яка має максимальну відстань від різних класів навчальних даних. Такий класифікатор ідентифікує клас, виходячи з того, на якій стороні гіперплощини знаходяться дані з тестової вибірки.

Оскільки сигнал іонного струму крізь біологічну чи твердотільну нанопору є дуже малим за амплітудою, шум представляє значну проблему, яка суттєво обмежує чутливість нанопори. В роботах [23-25] наведено порівняння рівнів шуму для різних матеріалів нанопор, їх діаметрів та довжин. У [24] показано, що біологічні нанопори мають кращі показники шуму порівняно з твердодільними нанопорами. Відношення сигнал/шум розраховувалося у часовій області як відношення зміни амплітуди сигналу при блокуванні пори нуклеотидами до середньоквадратичного значення іонного струму у порожній порі. Так, в залежності від матеріалу, відношення сигнал/шум у біологічній нанопорі коливалося від 4 до 15 [23].

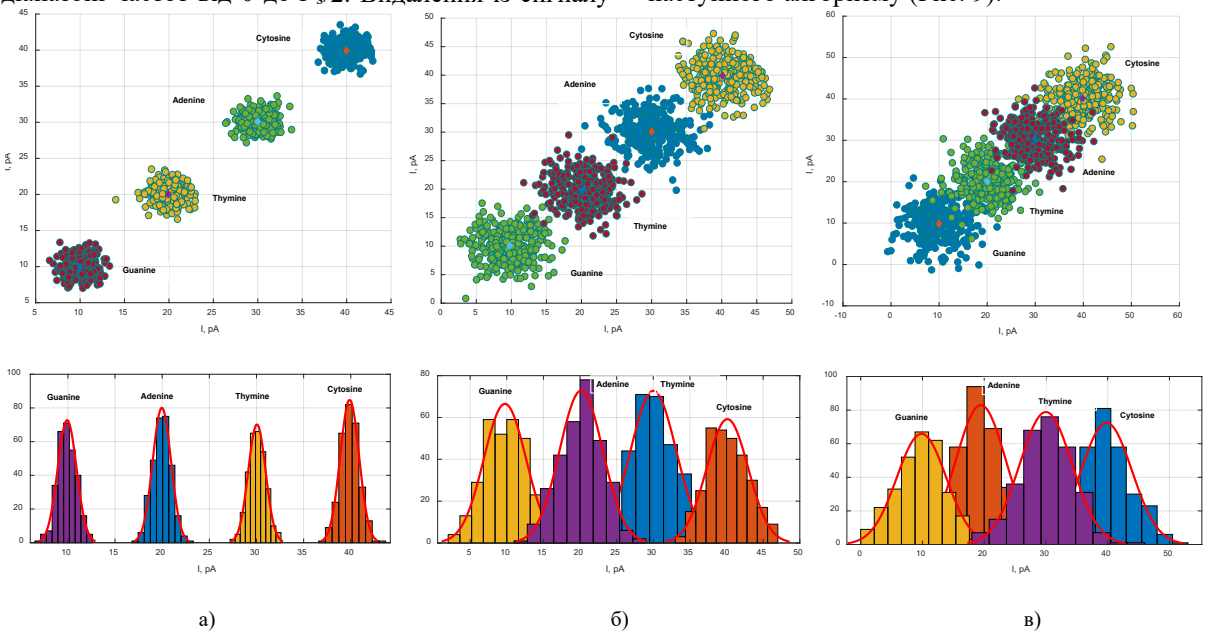
В даній роботі проведено порівняння результатів класифікації в залежності від рівня шуму у сигналі, що відповідав експериментальним даним з джерел [23, 24]. У таблиці 1 наведено точність розпізнавання модельних сигналів нанопорового секвенування ДНК в залежності від відношення сигнал/шум та методу машинного навчання. Наведені значення відповідають відсотку вірно визначених сигналів при класифікації на 4 класи, а також окремо вказано відсоток вірно визначених випадків для класів “мутація інсерції”, “мутація делеції, “мутація місенс” та “норма”.

При відносно низькому рівні шуму, коли значення сходинок струму для різних нуклеотидів не перекриваються (Рис. 6, 7), точність класифікації досягає 100%. Але у разі збільшення амплітуди шумових складових, виникає перекриття класів для норми і однонуклеотидної мутації типу місенс чи нонсенс, особливо якщо рівні сходинок для 2х нуклеотидів в нормі і у разі мутації знаходяться поруч (як для гуаніну та тиміну, тиміну та аденіну, аденіну та цитозину). У цьому випадку виникають помилки класифікації через перекриття рівнів сигналу для цих нуклеотидів (Рис. 8). У випадку мутацій делеції та інсерції (вставки або видалення послідовності з декількох нуклеотидів) спостерігається зсув між сигналом в нормі та у разі мутації. Через цей зсув буде зростати відстань між сигналами. Тому ці класи класифікатор виявляє з нижчим рівнем помилок чи взагалі без помилок.

Сигнали з патч-кламп установки реєструються зі значними шумами [16] і, для подальшого визначення нуклеотидів та виявлення мутацій, їх необхідно очистити від шумів, щоб уникнути помилок розпізнавання. У разі частотної фільтрації із використанням фільтру нижніх частот або у разі застосування

DOI: 10.20535/2523-4455.mea.217265

вейвлет-розкладу для зменшення рівня шуму, отримуємо сигнал зі спотворенням сходинок, які передають іонний струм у нанопорі під час блокування її нуклеотидами. Це пояснюється тим, що спектр прямокутного сигналу має частотні складові на всьому діапазоні частот від 0 до $F_s/2$. Видалення із сигналу високочастотних складових викликає спотворення фронтів прямокутних сходинок, що в свою чергу приводить до некоректного розшифрування послідовності нуклеотидів. Тому у даній роботі для усунення шумових складових пропонується використання наступного алгоритму (Рис. 9):



а) б) в)

Рис. 7 Ілюстрація перекриття значень іонного струму в каналі нанопори для різних азотистих основ ДНК у разі збільшенні рівня шуму

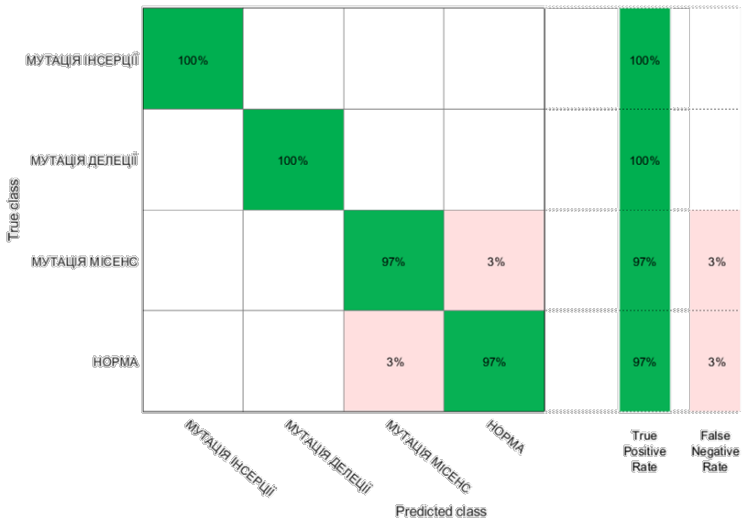


Рис. 8 Матриця похибок класифікації для застосування методу машинного навчання *k*-найближчих сусідів (10 сусідів, відстань Мінковського) до задачі класифікації сигналів нанопорового секвенування ДНК – загальна точність класифікації на 4 класи 98.5%

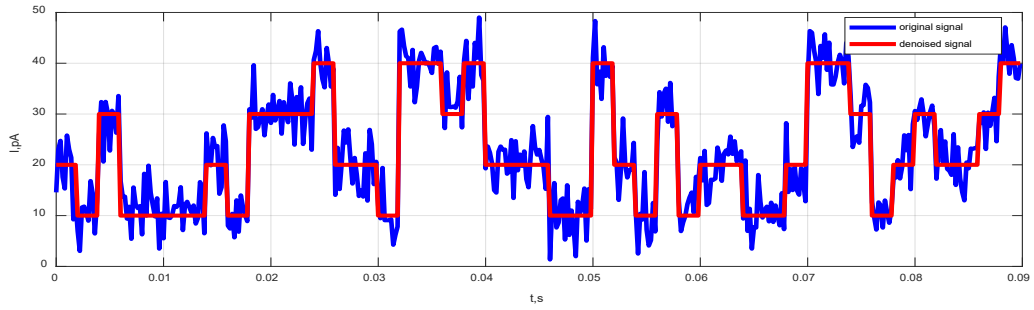


Рис. 9 Застосування знешумлення сигналу іонного струму під час секвенції ДНК

DOI: 10.20535/2523-4455.mea.217265

1) Фіксується початок входження чергового нуклеотида в нанопору. Створюється часове вікно довжиною в тривалість знаходження 1 нуклеотида в нанопорі (τ = 2 мс, що за умови Fs=5 кГц відповідає 10 відлікам сигналу).

2) Визначається середнє значення струму крізь нанопору під час знаходження в ній поточного нуклеотиду. Так як білий шум має нульове середнє значення, то середнє значення струму крізь нанопору під час знаходження в ній поточного нуклеотиду буде наближатися до фіксованих значень, властивих саме цьому нуклеотиду.

3) Визначається модуль різниці між вектором значень струму крізь нанопору, властивим для 4 типів азотистих основ A, T, G, C та отриманим середнім значенням на часовому інтервалі вікна. Далі необхідно визначити, де ця різниця є мінімальною і присвоїти всій ділянці струму у часовому вікні знайдене фіксоване значення.

4) Процедура повторюється для всього сигналу зі зсувом часового вікна вздовж сигналу на крок, що відповідає тривалості знаходження 1 нуклеотиду в нанопорі (τ = 2 мс, що за умови Fs=5 кГц відповідає 10 відлікам сигналу).

5) Отримується очищений від шуму сходинковий сигнал струму крізь білкову нанопору.

6) Згідно зі значеннями сходинок струму повертається код з літер, які відповідають послідовності нуклеотидів секвенованої ділянки одноланцюгової ДНК.

Застосування запропонованого алгоритму знешумлення модельних сигналів іонного струму при секвенції ДНК дозвлило без помилок ідентифікувати всі сигнали розглянутих вище 4 класів.

Висновки

В роботі проаналізовано застосування методів обробки геномних сигналів до задачі моделювання та аналізу сигналів нанопорового секвенування ДНК. На основі послідовностей нуклетидів в нормі та у разі мутацій змодельовано 1200 сигналів, які представляють 4 класи: норма, місенс-мутація, мутація інсерції та мутація делеції. Застосовано кореляційний аналіз для визначення подібності сигналів нанопорового секвенування ДНК за допомогою взаємної кореляційної функції між двома сигналами струму у нанопорі – у нормі та за наявності мутації. По розташуванню піка кореляції визначено тип мутації (інсерція чи делеція), а також проведено вирівнювання однакових послідовностей нуклеотидів за допомогою визначеного зсуву між сигналами.

Результати застосування методів машинного навчання до задачі класифікації сигналів нанопорового секвенування ДНК суттєво залежать від рівня шуму зареєстрованих сигналів іонного струму крізь білкову нанопору та типу мутації. За умови відносно низького рівня шуму, коли значення сходинок струму для різних нуклеотидів не перетинаються, точність класифікації досягає 100%. У разі збільшення середньоквадратичного відхилення шумових складових, виникає перекриття рівнів значень струму у нанопорі під час її блокування нуклеотидами близьких розмірів. Як наслідок, найчастіше виникають помилки у визначенні класів норми і однонуклеотидної мутації (місенс або нонсенс), особливо якщо рівні сходинок струму у нанопорі для двох нуклеотидів в нормі і при мутації знаходяться поруч (наприклад, гуаніну та тиміну, тиміну та аденіну, аденіну та цитозину) і шум маскує їх внесок до зменшення струму у нанопорі. Мутації інсерції та делеції, тобто вставки чи вирізання деякої послідовності нуклеотидів, найчастіше класифікуються без помилок, оскільки для таких мутацій характерний зсув в декілька нуклеотидів між сигналами в нормі і патології, через який зростає відстань між цими сигналами. Серед методів машинного навчання, які продемонстрували високу точність класифікації змодельованих сигналів струму у білковій нанопорі під час секвенування ДНК, слід відзначити метод лінійного дискримінанту, метод k-найближчих сусідів (за умови використання евклідової відстані та достатньої кількості найближчих сусідів – в даній роботі десяти), а також метод опорних векторів. Найкращі результати класифікації отримано для методу опорних векторів. Застосування лінійної, квадратичної та кубічної функцій ядра показало високі відсотки вірно класифікованих сигналів – від 93 до 100%.

Перелік посилань

[1]. Hengyun Lu, Francesca Giordano, Zemin Ning. Oxford Nanopore MinION Sequencing and Genome Assembly, Genomics, Proteomics & Bioinformatics, Vol. 14, Issue 5, 2016, pp. 265-279, **DOI:** 10.1016/j.gpb.2016.05.004.

[2]. D. R. Garalde, C. R. O'Donnell, R. D. Maitra, D. M. Wiberg, G. Wang and W. B. Dunbar, "Modeling the Biological Nanopore Instrument for Biomolecular State Estimation," in *IEEE Transactions on Control Systems Technology*, vol. 21, no. 6, pp. 2038-2051, 2013, **DOI**: 10.1109/TCST.2012.2224349.

[3]. J. Kim, R. Maitra, K. D. Pedrotti and W. B. Dunbar, "A Patch-Clamp ASIC for Nanopore-Based DNA Analysis," in *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 3, pp. 285-295, 2013, **DOI**: 10.1109/TBCAS.2012.2200893.

[4]. Nanoporovoe sekvenuvannya: na porozi tret'oyi henomnoyi revolyutsiyi [Nanoporous sequencing: on the threshold of the third genomic revolution].
URL: https://biomolecula.ru/articles/nanoporovoe-sekvenirovanie-na-poroge-tretei-genomnoi-revoliutsii

[5]. Anastassiou, Dimitris. (2001). Genomic Signal Processing. Signal Processing Magazine, IEEE. 18. 8-20. **DOI**: 10.1109/79.939833.

[6]. Mendizabal-Ruiz G, Román-Godínez I, Torres-Ramos S, Salido-Ruiz RA, Vélez-Pérez H, Morales JA. Genomic signal processing for DNA sequence clustering. PeerJ. 2018 Jan 24;6:e4264. **DOI:** 10.7717/peerj.4264. PMID: 29379686; PMCID: PMC5786891.

[7]. P. Dixit and G. I. Prajapati, "Machine Learning in Bioinformatics: A Novel Approach for DNA Sequencing," *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, Haryana, 2015, pp. 41-47, **DOI:** 10.1109/ACCT.2015.73

[8]. J. Chen and S. T. c. Wang, "Nanotechnology for genomic signal processing in cancer research - A focus on the genomic signal processing hardware design of the nanotools for cancer ressearch," in *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 111-121, Jan. 2007, **DOI:** 10.1109/MSP.2007.273064

[9]. P. Qiu, Z. J. Wang and K. j. R. Liu, "Genomic processing for cancer classification and prediction - Abroad review of the recent advances in model-based genomoric and proteomic signal

DOI: 10.20535/2523-4455.mea.217265

processing for cancer detection," in *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 100-110, Jan. 2007, **DOI:** 10.1109/MSP.2007.273063.

[10]. Ravichandran Lakshminarayan et al.. (2011). Waveform Mapping and Time-Frequency Processing of DNA and Protein Sequences. Signal Processing, IEEE Transactions on. 59. 4210 - 4224. **DOI:** 10.1109/TSP.2011.2157915.

[11]. S. Deng, Z. Chen, G. Ding and Y. Li, "Prediction of protein coding regions by combining Fourier and Wavelet Transform," *2010 3rd International Congress on Image and Signal Processing*, Yantai, 2010, pp. 4113-4117, **DOI:** 10.1109/CISP.2010.5648065.

[12]. T. Meng *et al*., "Wavelet Analysis in Current Cancer Genome Research: A Survey," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 6, pp. 1442-14359, 2013, **DOI:** 10.1109/TCBB.2013.134.

[13]. David Stoddart et al. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore//Proceedings of the National Academy of Sciences, 2009, 106 (19), pp.7702-7707. **DOI:** 10.1073/pnas.0901054106

[14]. Kim, Bong-Hyun & Yu, Kijin & Lee, Peter. (2019). Cancer classification of single cell gene expression data by neural network. Bioinformatics (Oxford, England). 36. **DOI:** 10.1093/bioinformatics/btz772.

[15]. Rockwood AL, Crockett DK, Oliphant JR, Elenitoba-Johnson KS. Sequence alignment by cross-correlation. J Biomol Tech. 2005; 16(4):453-458. **URL:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2291754 **PMID**: 16522868

[16]. Single-channel recording /edited by Bert Sakmann and Erwin Neher. - Springer. - 705 p. **DOI**: 10.1007/978-1-4419-1229-9

[17]. Bindal, N., Forbes, S.A., Beare, D. et al. COSMIC: the catalogue of somatic mutations in cancer. Genome Biol 12, P3 (2011). **DOI**: 10.1186/1465-6906-12-S1-P3

[18]. NCBI. – **URL**: http://www.ncbi.nlm.nih.gov

[19]. The Cancer Genome Atlas Program. **URL**: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[20]. Genomic Data Commons Data Portal. **URL**: https://portal.gdc.cancer.gov/

[21]. İ. B. AYDİLEK, "Examining Effects of the Support Vector Machines Kernel Types on Biomedical Data Classification," *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, 2018, pp. 1-4, **DOI:** 10.1109/IDAP.2018.8620879.

[22]. A. David and B. Lerner, "Pattern classification using a support vector machine for genetic disease diagnosis," *2004 23rd IEEE Convention of Electrical and Electronics Engineers in Israel*, Tel-Aviv, Israel, 2004, pp. 289-292, **DOI:** 10.1109/EEEI.2004.1361148.

[23]. Alessio Fragasso, Sonja Schmid, and Cees Dekker, "Comparing Current Noise in Biological and Solid-State Nanopores," ACS Nano 2020, 14(2), 1338-1349, **DOI:** 10.1021/acsnano.9b09353

[24]. Shengfa Liang, Feibin Xiang, Zifan Tang, Reza Nouri, Xiaodong He, Ming Dong, Weihua Guan, "Noise in nanopore sensors: Sources, models, reduction, and benchmarking," Nanotechnology and Precision Engineering, Volume 3, Issue 1, 2020, Pages 9-17, **DOI:** 10.1016/j.npe.2019.12.008.

[25]. Wen, C., Zeng, S., Zhang, Z., Hjort, K., Scheicher, R. et al. On nanopore DNA sequencing by signal and noise analysis of ionic current. Nanotechnology, 27: 215502, 2016 **DOI:** 10.1088/0957-4484/27/21/215502

Надійшла до редакції 26 листопада 2020 року.

UDC 004.056.53

DOI: 10.20535/2523-4455.mea.217265

Simulation and Analysis of Bionanopore DNA Sequencing Signals for Genetic Mutations Detection

I. M. Ievdoshchenko[f], ORCID 0000-0003-0049-2159
K. O. Ivanko, PhD Assoc.Prof., ORCID 0000-0002-3842-2423
N. H. Ivanushkina, PhD Assoc.Prof., ORCID 0000-0001-8389-7906
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute» ROR 00syn5v21
Kyiv, Ukraine

Vishwesh Kulkarni, Assoc.Prof., ORCID https://orcid.org/0000-0002-2285-8652
School of Engineering, University of Warwick ROR 01a77tt86
Warwick, United Kingdom of Great Britain and Northern Ireland

***Abstract*—The application of genomic signal processing methods to the problem of modeling and analysis of nanoporous DNA sequencing signals is considered in the paper. Based on the nucleotide sequences in the norm and in the case of mutations, 1200 signals are simulated, which represent 4 classes: norm, missense mutation, insertion mutation and deletion mutation. Correlation analysis was used to determine the similarity of nanoporous DNA sequencing signals using a cross-correlation function between two current signals in the protein nanopore, specifically signal in norm and in the presence of mutation. The location of the correlation peak determines the type of mutation (insertion or deletion), as well as the alignment of the same nucleotide sequences using a defined signal shift.**

The results of applying machine learning methods to the problem of classification of nanoporous DNA sequencing signals significantly depend on the noise level of the registered current signals through the protein nanopore and the type of mutation. Given a relatively low noise level, when the values of the ion current through a protein nanopore for different nucleotides do not intersect, the classification accuracy reaches 100%. In the case of increasing the standard deviation of the law of distribution of noise components, there is an overlap of the levels of current values in the nanopore in the case of its blocking by nucleotides of the close size. As a result, errors in the definition of normal and single nucleotide mutations (missense or nonsense) often occur, especially if the levels of current steps in the nanopore for two nucleotides are similar (for example, guanine and thymine, thymine and adenine, adenine and cytosine) and noise masks their contribution to reduction current in the nanopore. Mutations of insertion and deletion of a certain nucleotide sequence are often classified without errors, because these mutations are characterized by a shift of several nucleotides between normal signals and pathology, which increases the distance between these signals. Among the machine learning methods that have demonstrated the high accuracy of classification of the signals of nanopore-based DNA sequencing, the methods of linear discriminant, k-nearest neighbors classifier (with Euclidean distance and the sufficient number of nearest neighbors), as well as the method of reference vectors should be mentioned. The best results were obtained for the classification method of support vector machines. The use of linear, quadratic and cubic kernel functions shows the high accuracy of correctly classified signals — from 93 to 100%.

Keywords — DNA sequencing; single protein molecule nanopore; mutations; genomic signal processing; classification; machine learning.