

Парчевська Д. Д., магістр, Грецький І. О., доц., Волошина І. М., доц.

Київський національний університет технологій та дизайну

БІОІНФОРМАЦІЙНІ МЕТОДИ, ЩО ВИКОРИСТОВУЮТЬСЯ В БІОТЕХНОЛОГІЇ

Анотація. У статті розглянуто основні біоінформаційні методи, що використовуються в сучасних біотехнологічних дослідженнях – вирівнювання послідовностей, філогенетичний аналіз, граф де Бройна. Окремо проаналізовано методи передбачення структури та функцій білків та розробки ліків. Охарактеризовано біоінформаційні системи «EkoIS» та «Spark».

Ключові слова: біоінформатика, біотехнології, філогенетика, вирівнювання послідовностей.

Parchevska D. D., Hretskyi I. O., Voloshina I. M.

Kyiv National University of Technologies and Design

BIOINFORMATIVE METHODS USED IN BIOTECHNOLOGY

Abstract. The article discusses the main bioinformatics methods used in modern biotechnological research – sequence alignment, phylogenetic analysis, de Bruijn sequence. The methods of predicting the structure and functions of proteins and developing drugs are analyzed separately. Bioinformation systems «EkoIS» and «Spark» are characterized.

Keywords: bioinformatics, biotechnology, phylogenetics, sequence alignment.

Вступ. За останні роки в галузі біотехнологій, і зокрема в біоінформатиці, відбувся значний прогрес. Це пояснюється появою все нових і нових технологій і великої кількості різних типів даних, які потрібно інтегрувати в перспективу персоналізованої медицини. Через такі великі обсяги нових даних і швидкі темпи прогресу постають виклики в методології досліджень, адже вченим постійно необхідно добирати методи з усього переліку вже наявних, або ж створювати нові.

На сьогодні існують навчально-методичні посібники та практикуми з оглядами баз даних і технологій біоінформаційних досліджень [1]. Однак учені також зосереджуються й на окремих аспектах цієї науки. Так, Л.Ю. Бабінцева розглядає нові напрями біомедичної інформатики в стратегії змінення практичної медицини та біомедичної освіти, О.В. Кленіна вивчає використання баз даних хемоінформатики та біоінформатики у процесах комп'ютерного конструювання ліків, О.М. Ключко із командою досліджує біоінформаційну систему «EkoIS» з базами даних та інкорпорованими біосенсорами [8], а О.С. Ястребова цікавиться моделюванням біологічних процесів на основі даних молекулярної біотехнології та біоінформатики. Серед іноземних учених також популярні загальні огляди біоінформаційних методів і технологій. Проте, є й ті хто зосереджується на конкретних елементах цієї галузі. Наприклад, С. Belser досліджує біоінформаційні підходи до аналізу геномів і біорізноманітності, F. Dufour з їх допомогою вивчає дерегуляцію фізіологічних систем у процесі старіння, D. Lévy та його колеги – електронну кріо-мікроскопію для нового погляду на клітину та її складові [7].

Постановка завдання. Метою нашого дослідження є аналіз біоінформаційних методів, що застосовуються в біотехнології для оптимізації біологічних процесів, генетичних досліджень, розробки нових терапевтичних препаратів і вдосконалення інструментів для обробки великих обсягів біологічних даних. Дослідження спрямоване на систематизацію та розкриття ролі цих методів у сучасній науці.

Результати досліджень. Одним із ключових методів біоінформаційних досліджень є вирівнювання послідовностей – порівняння нуклеотидних (ДНК, РНК) або

амінокислотних (білки) послідовностей для виявлення схожих ділянок, що свідчать про функціональні, структурні або еволюційні зв'язки. З його допомогою ми можемо відповісти на питання «чи схожі дві послідовності ДНК?» та «наскільки вони схожі?» Порівняння нових послідовностей з відомими допомагає встановити їх спорідненість, виявити консервативні та варіабельні ділянки, а також дослідити еволюційні взаємозв'язки. [1, с. 72–73]. Вирівнювання – це алгоритмічний процес порівняння послідовностей, нуклеотид за нуклеотидом, для виявлення подібностей і відмінностей. У випадку подібних послідовностей, відмінності відповідають або природнім мутаціям, або помилкам секвенування: заміні одного нуклеотиду іншим, вставці додаткового нуклеотиду або видаленню нуклеотиду. Щоб визначити оптимальне вирівнювання, яке мінімізує кількість відмінностей між двома послідовностями, необхідно вивчити всі можливі комбінації вставок або видалень. Однак у довжині послідовності є експоненціальне число, тому просто розраховувати всі вирівнювання на практиці – неможливо. Вирівнювання вирішується за допомогою динамічного програмування, ідея якого полягає в тому, щоб розділити проблему на менші підпроблеми й таким чином розрахувати оптимальне вирівнювання [2, с. 4].

Вирівнювання послідовностей буває парним (між двома послідовностями) та множинним (між кількома). Незалежно від типу, можна використовувати глобальне або локальне вирівнювання. Глобальне вирівнювання охоплює всю довжину послідовностей і підходить для гомологічних по всій довжині. Тобто, воно більш ефективне для схожих послідовностей [1, с. 74–75]. Однак для більш віддалених послідовностей краще використовувати локальне вирівнювання, яке знаходить схожі ділянки всередині послідовностей. Найвідоміші методи – програма FastA, яка швидко знаходить схожі області, і BLAST (Basic Local Alignment Search Tool), що порівнює послідовності на основі підпослідовностей, упорядковуючи результати за рівнем подібності [3, с. 31–32].

BLAST – це популярний інструмент для пошуку схожих ділянок між послідовністю запиту та великою базою даних послідовностей, наприклад, Genbank. Його завдання – виявити локальні подібності, тобто максимально схожі області між запитом та послідовністю в базі даних, допускаючи можливість вставок і видалень. Хоча ідеальне розв'язання цієї задачі є обчислювально складним, BLAST використовує спеціально розроблену евристику, яка забезпечує дуже швидкий пошук (іноді за лічені секунди). Для кожного порівняння BLAST надає показник відповідності та оцінку очікуваної кількості збігів з таким самим або вищим балом, що могли б виникнути випадково. Якщо очікуване значення є дуже низьким, то послідовність з бази даних вважається «збігом» і ймовірним гомологом запитуваної послідовності [4].

BLAST визначає схожі фрагменти між різними послідовностями. Програма зіставляє нуклеотидні або білкові послідовності з даними в базах і розраховує статистичну значущість знайдених збігів. Цей інструмент дозволяє робити висновки про функціональні та еволюційні зв'язки між послідовностями, а також допомагає ідентифікувати членів генних сімейств [5].

Для виконання множинного вирівнювання було використано програмне забезпечення Clustal, яке є популярним інструментом у молекулярній біології для вирівнювання нуклеотидних і білкових послідовностей, а також для побудови філогенетичних дерев. Популярність Clustal обумовлена не лише високою точністю результатів, але й його надійністю, зручністю та мобільністю. Останні функції включають підтримку форматів NEXUS і FASTA, виведення нумерації діапазонів та покращену швидкість побудови дерев. Спочатку розроблений для локального використання, Clustal зараз доступний також на багатьох веб-серверах, наприклад, у Європейському інституті біоінформатики [6].

Вирівнювання білкових послідовностей є ключовим елементом сучасної біоінформатики. Завдяки множинному вирівнюванню дослідники можуть аналізувати консервативні еволюційні особливості послідовностей та родинні зв'язки між організмами. Існують два основних типи вирівнювання: глобальне, яке охоплює всю послідовність, і локальне, що охоплює лише окремі ділянки. Найчастіше для глобального вирівнювання використовують програми серії Clustal. Першу версію Clustal розробив Дес Хіггінс у 1988 році з урахуванням обмеженої потужності комп'ютерів того часу. Він поєднав динамічне програмування зі стратегією прогресивного вирівнювання, де вирівнювання послідовностей проводиться поетапно, відповідно до гілок дерева. Для початкових порівнянь застосовувався швидкий алгоритм на основі слів, а дерево створювалось методом UPGMA [6].

У 1992 році вийшла версія ClustalV з новими можливостями вирівнювання профілів та побудовою дерев методом з'єднання сусідніх. У 1994 році з'явилась ClustalW, яка додала удосконалення: вагові коефіцієнти для послідовностей і автоматичний підбір матриць для кожного етапу вирівнювання. Крім того, приблизний пошук слів замінили на більш точний алгоритм динамічного програмування, а UPGMA замінили на NJ для побудови дендограм [6].

Clustal був розроблений для забезпечення надійного і точного інструменту, що легко використовувати, а тісна співпраця біологів і програмістів стала однією з причин його успіху. ClustalW відкрив шлях для нових розробок, зокрема ClustalX, що забезпечує інтерактивне редагування вирівнювань, дозволяючи перевирівнювати окремі ділянки чи послідовності [6].

Обидві програми, ClustalW і ClustalX, продовжують підтримуватися та оновлюватися. Останні вдосконалення включають збереження результатів у форматі NEXUS для сумісності з іншими програмами. Серед нових функцій: збереження результатів у форматі FASTA, генерація матриці ідентичності у відсотках, збереження обраних діапазонів залишків, що корисно при виділенні доменів з багатодомених білків, та пришвидшення алгоритму NJ для побудови дерев [6].

Філогенетичний аналіз – невід'ємна складова багатьох біоінформаційних досліджень. Він вивчає еволюційні зв'язки між організмами та допомагає зрозуміти їхню історію. Ці зв'язки відображаються у вигляді філогенетичних дерев – діаграм, які показують еволюційний родовід. Філогенетичні дерева бувають укорінені (з позначеним загальним предком) і неукорінені (без кореня, лише зв'язки між видами). Укорінене дерево ілюструє розвиток видів від спільного предка, тоді як неукорінене показує зв'язки без зазначення напрямку еволюції [1, с. 88–90].

На увагу заслуговує найпоширеніший метод реконструкції геному за даними секвенування, що базується на «графі де Бройна». Назва походить від імені математика Ніколаса де Бройна, який представив цю структуру даних як комбінаторний об'єкт ще в 1940-х роках. Однак граф де Бройна увійшов у біоінформатику 60 років потому, з появою високопродуктивного секвенування. Метод реконструкції геному на основі графа де Бройна використовує розбиття зчитувань секвенування на короткі фрагменти, k-мери. Граф будується так, що кожен унікальний k-мер є вершиною, а перекриття між ними – ребрами. Геном відновлюється шляхом пошуку ейлерового шляху, який проходить кожне ребро один раз. Метод потребує адаптації для обробки помилок секвенування та повторюваних або непокритих ділянок геному [2, с. 5–6].

Методи передбачення структури та функцій білків – це окрема категорія біоінформаційної методології. Найпоширенішими методами передбачення вторинної структури білків є метод Чоу-Фасмена (базується на частоті появи амінокислот в різних структурах (α-спіралі, β-листи, β-згини) і використовує правила для передбачення),

моделювання нейронних мереж (програми розпізнають амінокислотні комбінації у відомих структурах і відрізняють їх від інших) і метод «найближчого сусіда»: амінокислоти розміщуються у відповідній конформації вторинної структури [1, с. 112–113].

Методи моделювання білків поділяються на два типи: на основі шаблону (гомологічне моделювання та метод протягування) і без шаблону (*de novo* та *ab initio*). Гомологічне моделювання використовує схожість амінокислотної послідовності з білком-шаблоном для створення тривимірної моделі. Його точність залежить від ступеня схожості послідовностей. Метод протягування передбачає «протягування» білка-мішені через бібліотеку відомих білкових укладок. Моделювання *de novo* поєднує фрагментацію та фізичну оптимізацію для побудови моделі. Моделювання *ab initio* базується на фізичних принципах і моделює білок за допомогою молекулярних симуляцій, але потребує великих ресурсів [1, с. 117–124].

Для пошуку та розробки нових лікарських засобів біоінформатика також пропонує свої методи. Виявлення мішеней для ліків можливе через експерименти або комп'ютерне моделювання. Біоінформатика дозволяє проводити віртуальний скринінг і прогнозувати властивості речовин, що зменшує витрати часу і грошей [1, с. 152].

У розробці ліків у нагоді стає і кріоелектронна мікроскопія–метод візуалізації живих організмів, який зараз займає важливе місце в структурній і клітинній біології, біоінформатиці, біомедицині та клітинній фізиці. Вона дозволяє вивчати білкові структури та частинки вірусів через електронний мікроскоп, у якому зображуються заморожені зразки. Цей метод активно застосовується для створення ліків, зокрема для дослідження трансмембранних білків і нейродегенеративних захворювань. Також він використовувався для вивчення структури білків вірусу SARS-CoV-2, що допомогло у розробці терапевтичних засобів [7, с. 379].

Систему «ЕкоІС» розроблено для моніторингу в різних часових інтервалах – від мілісекунд після впливу хімічних речовин на клітини до змін у популяціях організмів через місяці чи роки. Система базується на комп'ютерних технологіях і використовує біосенсиори з електро-, хемочутливими мембранами нейронів мозку. Це дозволяє виявляти речовини у мінімальних кількостях (до 1–2 мілімолей на літр). Завдяки експертній системі, «ЕкоІС» розпізнає хімічні речовини та надає висновки щодо їх безпечності або небезпеки. Доукомплектована версія може також виявляти пожежну небезпеку [8, с. 71–72].

Система «Spark» використовується для прискорення аналізу великих біологічних даних у різних сферах: від аналізу послідовностей до епігенетики, філогенетики, розробки ліків, одноклітинного секвенування РНК та генетичних досліджень населення. SparkSeq значно покращує швидкість обробки NGS-даних, а CloudPhylo вирішує завдання філогенетичної реконструкції. У дослідженні CpG-островів і відкритті ліків застосування система «Spark» дозволила суттєво зменшити час аналізу та підвищити точність. Такі інструменти як VariantSpark і SEQSpark прискорюють генетичні дослідження, а Biospark спрощує обробку великих числових даних. «Spark» забезпечує високу продуктивність і масштабованість для біоінформатики [9].

Сучасні молекулярно-генетичні та біоінформатичні методи дослідження геномів мікроорганізмів відкривають можливості для їх ідентифікації та вивчення мікроеволюційних процесів через аналіз консервативних ділянок геному. Швидкий прогрес молекулярно-біологічних технологій сприяє створенню великих баз даних із результатами досліджень у цій галузі. Однак це породжує нові труднощі – обробка та аналіз значних обсягів інформації стають усе складнішими, особливо для баз даних нуклеотидних послідовностей, таких як GenBank.

Інтернет-сервіси надають доступ до анотованих списків нуклеотидних послідовностей на запити, які містять ключові слова, як-от назва гена, організму чи реєстраційний номер послідовності. Водночас можливості для обробки та аналізу цих даних є досить обмеженими. Коли список містить лише кілька записів, їх можна переглянути вручну, але якщо записів сотні чи тисячі, це стає непродуктивним і практично неможливим завданням.

Коли постає необхідність визначити кількість нуклеотидних послідовностей для кожного виду певного роду і їхній зв'язок з певними генами, ручний перегляд стає надто трудомістким. Серія запитів з різними видами або генами також не спрощує аналіз. Крім того, наявність помилок у базах даних може призводити до втрати інформації, що спотворює результати дослідження.

Висновки. Отже, біоінформатика стала незамінним інструментом у сучасній біотехнології. Її застосування варіюється від розкриття таємниць людського геному до розробки нових лікарських засобів. У роботі було розглянуто як базові, так і нові біоінформаційні технології та системи, що сприяють оптимізації біотехнологічних досліджень, підвищенню їх швидкості та точності. Перспективи подальших досліджень вбачаємо в аналізі біоінформаційних методів для розробки вакцин проти вірусу грипу.

Список використаної літератури

1. Кеца О. В. Основи біоінформатики: навч.-метод. посібник. Чернівці: Чернівецький нац. ун-т ім. Ю. Федьковича, 2018. 192 с. URL: <https://archer.chnu.edu.ua/xmlui/handle/123456789/3149>.
2. Lemaitre, C., Salson, M., Touzet, H. (2022). Comment la bioinformatique a résolu le puzzle du génome du SARS-CoV-2. HAL. 6 p. URL: <https://inria.hal.science/hal-03896532/file/Article1%20-%20decouverte.pdf>.
3. Sidi, M. L. A. (2022). Nouvelles approches informatiques et mathématiques pour la résolution de problèmes biologiques. Doctoral dissertation. Université de Tours-LIFAT. URL: <https://hal.science/tel-03807522/document>.
4. About BLAST. URL: https://www.megasoftware.net/web_help_7/hc_about_blast.htm.
5. BLAST: Basic Local Alignment Search Tool. URL: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.
6. Multiple sequence alignment with the Clustal series of programs. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC168907/>
7. Lévy, D., Di Cicco, A., Bertin, A., Dezi, M. (2021). La cryo-microscopie électronique révèle une nouvelle vision de la cellule et de ses composants. *Médecine/sciences*, No. 37(4), P. 379–385. URL: <https://www.medecine-sciences.org/articles/medsci/pdf/2021/04/msc200425.pdf>.
8. Ключко О. М., Пітерцев О. А., Миронов Д. О., Тишук Б. В. Біоінформаційна система «ЕкоІС» з базами даних та інкорпорованими біосенсорами. Редакційна колегія, 2020. 70 с. URL: https://www.researchgate.net/profile/Olena-Klyuchko/publication/361349651_Bioinformacijna_sistema_EkoIS_z_bazami_danih_ta_inkorporovanimi_biosensorami/links/62acbcc5e1193368baa4c32b/Bioinformacijna-sistema-EkoIS-z-bazami-danih-ta-inkorporovanimi-biosensorami.pdf.
9. Guo, R., Zhao, Y., Zou, Q., Fang, X., Peng, S. (2018). Bioinformatics applications on Apache Spark. *GigaScience*, No. 7(8). URL: <https://doi.org/10.1093/gigascience/giy098>.