УДК 004.9

Коренчук О.В., Мазурець О.В.

Хмельницький національний університет, Україна

СУЧАСНІ ТЕНДЕНЦІЇ РОЗВИТКУ БІОІНФОРМАТИКИ

У статті досліджено розвиток біоінформатики на сучасному етапі. Визначено актуальні напрямки застосування інформаційних технологій у біоінформатиці. Запропоновано використання технології паралельних обчислень для підвищення продуктивності обрахунків при рішенні ряду характерних задач.

This article studies the development of bioinformatics in the modern period. It defines current trends in the application of IT in bioinformatics and suggests the use of parallel computing to improve calculation performance for a number of specific tasks.

Біоінформатика є розділом комп'ютерних наук, що швидко розвивається і завдяки якому біологія у 21 столітті перейшла з розділу наук про життя у обчислювальні науки. Під біоінформатикою зазвичай розуміють використання комп'ютерних технологій для вирішення біологічних задач, таких як вивчення специфічних алгоритмів та методи аналізу даних великого об'єму, працюючи переважно з геномними та білковими послідовностями. Біоінформатика набуває нових знань шляхом комп'ютерного аналізу біологічних даних. Біологічні дані можуть містити інформацію, яка міститься в генетичному коді, також експериментальні дані з різних джерел, медичної статистики чи наукової літератури. Досліди в галузі біоінформатики включають в себе розробку методів для зберігання, обробки та аналізу даних. Біоінформатика є мультидисципліною, складові якої швидко розвиваються, про що свідчать численні публікації [1-3]. Цей напрямок використовує методи та поняття інформатики, статистики, математики, хімії, біохімії, фізики, лінгвістики.

Швидкий розвиток обчислювальної техніки сприяє інтенсифікації досліджень в біоінформатиці. Відповідно, виявлення найбільш актуальних напрямків застосування інформаційних технологій в даній галузі дозволить спрогнозувати характер її розвитку в майбутньому. Відповідно, *метою статті* визначено аналіз особливостей розвитку біоінформатики на сучасному етапі.

Загалом, біоінформатика ділиться на основні дослідні області:

- вирівнювання послідовностей геномів;
- пошук генів;
- збірку геномів;
- вирівнювання структур білків;
- передбачення структури білків;
- передбачення експресії генів та білок-білкової взаємодії;
- реконструювання процесу еволюції.

Особливо велику нішу досліджень біоінформатики займає отримання високоякісних послідовностей геномів з фрагментів послідовностей, отриманих за допомогою традиційних методів секвенування ДНК та конструювання сигнальних мереж за даними ДНК-мікрочіпів [4].

У випадку з геномом, коли є багато фрагментів інформації і її потрібно скласти в одне ціле (рис. 1), то вручну це майже неможливо зробити по двом причинам:

- людський фактор – завжди є вірогідність помилки через втому людини чи її психологічний стан;
- час та гроші – збір великої кількості дрібних фрагментів інформації та відновлення геному в умовах відсутності деяких частин – довготривалий та затратний процес.

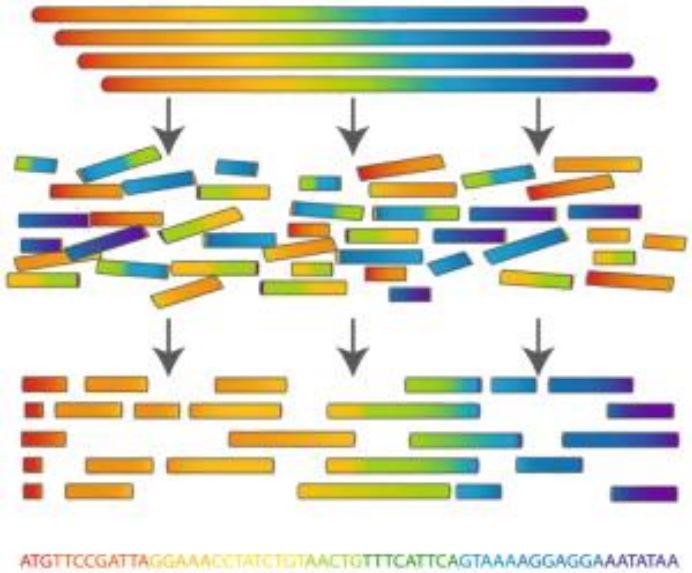

Рис. 1. **Умовне зображення процесу збору генома**

У 21 столітті на допомогу усім біологам світу приходять інформаційні технології – так і зароджується біоінформатика. Перші технології дозволяли зчитувати шматочки генома довжиною до декількох тисяч символів. Ці технології були неймовірно дорогі – на збір першого людського генома було витрачено декілька мільярдів доларів та декілька років кропіткої праці декількох сотень працівників лабораторій по всьому світу. Експериментальне визначення функції тільки одного гена потребує інтенсивної роботи однієї лабораторії як мінімум на протязі декількох місяців. Сучасні ж технології дозволяють з відомим степенем точності охарактеризувати декілька тисяч генів силами невеликої групи дослідників приблизно за тиждень. Звичайно, комп'ютерний аналіз не виключає й експериментальну перевірку, однак в цьому випадку експериментальна робота суттєво спрощується.

Комп'ютерні технології дозволяють читати більш короткі фрагменти, але на порядок дешевше та в більшій кількості. При зборі генома автоматично обробляються гігабайти вхідних даних. Для цього розроблено програми, що називаються геномними збірниками, або частіше – асемблерами (від англ. assemble – збирати), до прикладу, SPAdes, Velvet [5]. В силу деяких особливостей вхідних геномів (наприклад, регіони, що повторюються), а також великого числа помилок у вхідних даних, результатом роботи асемблера є не цілий ген, а лише достатньо довгі його ділянки. Чим довші ділянки отримано, чим більше вони схожі на результат, тим якіснішим вважається результат.

Звичайно, іноді існуючих програм недостатньо для вирішення поставлених задач, або вони мають недостатню точність, чи з'явився новий тип даних [6]. В цьому випадку починається розробка нових програм та алгоритмів. Сучасна біоінформатика при розробці нових алгоритмів широко використовує досягнення теорії ймовірностей, математичної статистики, інформатики. Для створення нової програми чи реалізації алгоритму не потрібні особливі ресурси. Більшість задач біоінформатики вирішується на таких мовах програмування як C++, Perl, Python, Haskell.

Не останню роль у біоінформатиці відіграє візуалізація результатів дослідження. Вчені в цій області працюють з великими об'ємами інформації, яку потрібно зрозуміти та уявити візуально. Хорошими прикладами засобів візуалізації являються браузери геномів (genome browser) (рис. 2).

Браузер генома – це така одномірна карта, яка відображає яку-небудь нуклеотидну послідовність (хромосому чи окремий ген) з супутньою інформацією [7]. Інформація зазвичай структурується в блоки, які називаються треками (tracs). Існує велика кількість таких браузерів. Багато з них спеціалізовані під певний організм чи тип даних. З найпопулярніших слід відмітити Integrated Genome та Browser Integrative Genomic Viewer. Такі браузери мають ряд недоліків. Зокрема – це швидкість візуалізації. При великому надходженні даних на візуалізацію затрачується багато часу. Також дані можуть бути погано структуровані, що також негативно впливає на час роботи браузера генома.

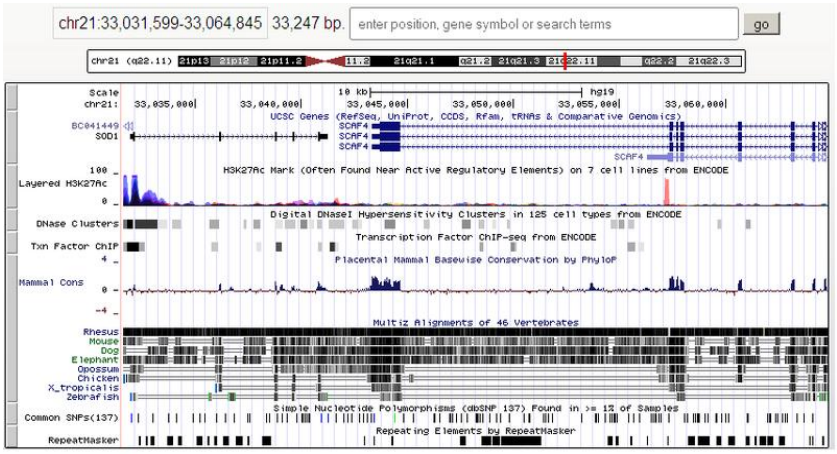

Рис. 2. **Приклад браузера геномів UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly**

Слід зазначити, що біоінформатика активно використовує можливості розподілених обчислень. Так як задачі біоінформатики досить об'ємні та ресурсоємні, виникає потреба у потужних системах, що вирішували б їх. Близько двадцяти проектів біоінформатики були і є учасниками глобальної спілки користувачів World Community Grid (WCG), які надають вільні ресурси своїх комп'ютерів для вирішення складних завдань. Проект для розподілених обчислень запроваджений у 2004 році компанією IBM [8]. Пропонується великий вибір досліджень з боротьби проти раку, СНІДу, грипу та інших захворювань, з яких учасник може зробити вибір. Проект обчислюється не лише добровольцями (у число яких може вступити кожен), але і партнерськими організаціями з багатьох країн. На квітень 2013 року нараховується 614 984 зареєстрованих користувачів, процесорних обчислень більш, ніж на 700 тисяч років. Кожен охочий може приєднатися до спілки на сайті проекту www.worldcommunitygrid.org, скачавши на свій комп'ютер потрібне програмне забезпечення.

Відкритим залишається питання зберігання інформації. Адже геном – це послідовність з тисяч символів. Значна кількість біологічної інформації надходить у різні банки даних. Ці банки даних містять часто первинну інформацію. Далі ця інформація переопрацьовується, в тому числі з залученням наукової літератури. В результаті виникають вторинні банки даних. Інформація в них, як правило, заслуговує великої довіри. В результаті рутинної, добре автоматизованої роботи вже отримана велика кількість генетичних текстів. Так, в базі даних EMBL-EBI на 17 грудня 2012 року міститься 86215682 записів з описом нуклеотидних послідовностей, які містять в собі близько 160000000000 символів (нуклеотидів). Знайти потрібний ген в EMBL-EBI без допомоги комп'ютера просто неможливо. А число даних експонтенціально росте (рис. 3).

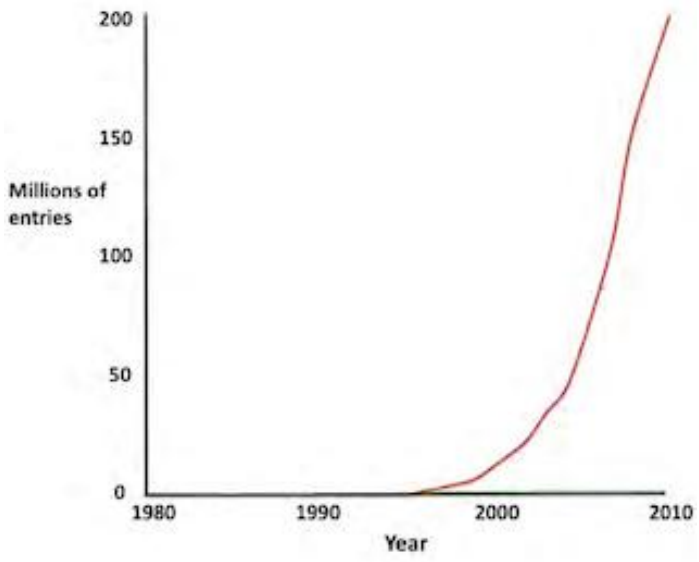

Рис. 3. **Динаміка росту біологічної бази даних EMBL-EBI**

Найбільші сховища первинних структур ДНК та нуклеотидних послідовностей (такі як EMBL, GenBank, DDBJ, SWISS-PROT, Ensembl та інші) поповнюються анотованими послідовностями безпосередньо дослідниками, які їх розшифрували, за допомогою автоматизованої системи поповнення баз даних по мережі Інтернет. Іншим основним джерелом інформації у всіх базах даних є спеціальна наукова

література. Багато баз даних, які працюють над колекціонуванням однорідної інформації, координують свої зусилля шляхом міжнародного розділення праці, наприклад співпраця трьох всесвітніх колекцій послідовностей нуклеотидів EMBL-EBI (Європа), GenBank (США), DDBJ (Японія). Інформація з цих баз є відкритою для як для науковців, так і для пересічних користувачів.

Для збору видових назв, описів, ареалу розповсюдження і генетичної інформації використовуються відповідні бази даних [9]. Спеціалізоване програмне забезпечення застосовується для пошуку, візуалізації й аналізу інформації, і, що важливіше, її доступності іншим людям. Комп'ютерні симулятори моделюють такі речі, як популяційна динаміка, або обчислюють загальне генетичне здоров'я культури в агрономії. Один з найважливіших потенціалів цієї області полягає в аналізі послідовностей ДНК організмів або повних геномів цілих вимираючих видів, дозволяючи запам'ятати результати генетичного експерименту природи в комп'ютері і можливо використовувати знову в майбутньому, навіть якщо ці види повністю вимруть. При чому біоінформатика дозволяє не тільки відновлювати та зберігати генну інформацію, а й змінювати її.

Отже, біоінформатика є галуззю обчислюваної біології, що є потужним інструментом в руках біологів. Завдяки біоінформатиці стало можливим швидке та надійне розшифрування геномів організмів, що дає можливість краще зрозуміти природу усього, що оточує людину. Головне досягнення біоінформатики – розшифрування геному людини. Завдяки досягненням біоінформатиків, які змогли створити штучну бактерію, можливість вилікувати деякі види раку стала реальністю.

Згідно проведеного аналізу, найбільш перспективними напрямками розвитку біоінформатики є відновлення генотипів різноманітних організмів, дослідження та їх аналіз з метою прослідковування еволюції людини, а також появи нових відкриттів, що допомогли б виліковувати досі невиліковні хвороби. Зважаючи на специфіку операцій обробки даних в біоінформатиці та їх значний обсяг, автори вважають ефективним рішенням широке застосування технології паралельних обчислень для суттєвого підвищення продуктивності при рішення багатьох задач, зокрема вирівнювання геномів, їх пошук та збірку, збереження, а також можливість візуалізації результатів дослідження.

Література

1. С.В. Горобець, О.Ю. Горобець, Д.О.Дереча. Біоінформатика як основний інструмент нанобіотехнології та наномедицини. Клиническая информатика и телемедицина, 2008, т.4., Вып.5. с.41- 49.
2. Гельфанд М.С., Миронов А.А. Вычислительная биология на рубеже десятилетий. Молекулярная биология. 1999. Т. 33. С. 969-984.
3. Что такое биоинформатика? [Електронний ресурс]. – Режим доступу: http://www.fbb.msu.ru/res/DOC125/pragrammi_kursov/%C1%E8%EE%E8%ED%F4%EE%F0%EC%E0%F2%E8%EA%E0.pdf
4. Биоинформатика: взгляд изнутри [Електронний ресурс]. – Режим доступу: http://habrahabr.ru/post/143115/
5. Биоинформатика, программирование и анализ данных [Електронний ресурс]. – Режим доступу: http://bioinformatics.ru/
6. Біоінформатика [Електронний ресурс]. – Режим доступу: http://uk.wikipedia.org/wiki/Біоінформатика
7. Браузеры генома [Електронний ресурс]. – Режим доступу: http://habrahabr.ru/post/170429/
8. World Community Grid [Електронний ресурс]. – Режим доступу: http://uk.wikipedia.org/wiki/World_Community_Grid
9. Базы данных в биоинформатике [Електронний ресурс]. – Режим доступу: http://www.microarray.ru/?p=23