



УДК 004.9

BIOCONDUCTOR IN SYSTEMS BIOLOGY: ANALYSIS OF BIOLOGICAL DATA USING R

BIOCONDUCTOR У СИСТЕМНІЙ БІОЛОГІЇ: АНАЛІЗ БІОЛОГІЧНИХ ДАНИХ ЗА ДОПОМОГОЮ R

Doroshenko I.V. / Дорошенко І.В.*s. p.-m.s., as.prof. / к. ф.-м.н., доц.*

ORCID: 0000-0001-8729-1768

*Chernivtsi National University, Chernivtsi, Kotsyubynskoho 2, 58012**Чернівецький національний університет, Чернівці, вул.Коцюбинського 2, 58012*

Анотація. Ця стаття присвячена вивченню можливостей використання мови програмування R та біоінформатичних пакетів у біоінформатиці для розвитку системної біології. Завдання дослідження полягають у аналізі сучасних методів обробки біологічних даних у середовищі R, вивченні функціональності біоінформатичних пакетів та їх застосуванні для вирішення конкретних біологічних проблем.

Ключові слова: біоінформатика, системна біологія, аналіз даних, молекулярна біологія, геноміка.

Вступ.

У сучасному світі біоінформатика є важливим компонентом біологічних наук, що сприяє розробці нових методів для аналізу та інтерпретації великих обсягів біологічних даних. Зростаюча популярність серед біоінформатиків середовища R та біоінформатичних пакетів обумовлена їх потужністю, гнучкістю та здатністю автоматизувати процеси аналізу даних.

1. Bioconductor

Проект Bioconductor включає кілька сотень програмних пакетів, що охоплюють широкий спектр завдань у біоінформатиці та обчислювальній біології. Крім того, існує велика кількість пакетів, що надають анотаційну інформацію для різних експериментів з мікрочипами та цілими організмами.

Bioconductor є популярним інструментом для аналізу даних експресії генів. Наприклад, пакет *arrayQualityMetrics* забезпечує комплексну оцінку якості для будь-якого типу масивів. Процес нормалізації, спрямований на усунення небіологічних відмінностей між масивами перед визначенням диференціально експресованих генів, найчастіше реалізується за допомогою пакетів *affy*, *vsn* та



limma. Для оцінки диференціальної експресії часто використовується *limma*, який застосовує метод байєсівського стиснення.

Гіпергеометричний аналіз збагачення генних наборів є популярним методом для тестування заздалегідь визначених груп генів. Для цього широко використовуються пакети *GOstats*, *GSEAbase*, *GSEAlm* та *Category*.

Багато аналітичних процедур залежать від геномної анотації, яка може включати інформацію про функції генів, їхні послідовності, залученість у біологічні шляхи тощо. Bioconductor надає набір незалежних від платформи анотаційних пакетів на основі *SQLite*. Багато з цих пакетів підтримують стандартні API та інтегруються з інструментами для аналізу мікрочипових та інших даних у поєднанні з анотаційною інформацією.

Для візуалізації та представлення результатів пакет *annaffy* пропонує інструменти для створення HTML-звітів. Водночас пакет *annotate* містить інструменти для розробки власних пов'язаних результатів аналізу.

Пакети *graph*, *RBGL* та *Rgraphviz* надають засоби для роботи з графами та мережами, включаючи їх представлення, обчислення та візуалізацію. Вони використовуються для моделювання біологічних шляхів, ієрархічних онтологій та системної біології.

Високопродуктивне генотипування за допомогою чипів однонуклеотидного поліморфізму (SNP) виконується на кількох рівнях. Пакет *oligo* використовується для обробки необроблених SNP-даних з масивів *Affymetrix*, а *beadarraySNP* — для *Illumina*. Організацію та аналіз великих SNP-наборів спрощує *snpMatrix*, тоді як *GGtools* забезпечує інтегрований аналіз експресії генів та даних генотипування, зокрема підтримуючи повне генотипування другої фази *HarMap* (4 мільйони SNP) для різних популяцій.

Для високопродуктивного скринінгу молекулярних сполук або РНК-інтерференції (RNAi) пакет *cellHTS2* пропонує інструменти для аналізу клітинних досліджень. Аналіз проточної цитофлуориметрії підтримується пакетами *flowCore* та *flowViz*, які забезпечують управління даними, візуалізацію, автоматизоване секвенування та статистичний аналіз.



Скринінг з високим вмістом даних за допомогою автоматизованої мікроскопії підтримується пакетом *EImage*, який надає інструменти для введення, виведення та аналізу зображень у R, а також розширені функції їх обробки. Для роботи з даними високопродуктивного секвенування пакет *ShortRead* забезпечує доступ до форматів коротких послідовностей, тоді як *Biostrings* містить інструменти для вирівнювання та порівняння послідовностей із цільовими геномами.

2. Експериментальні дані

З самого початку проект Bioconductor підкреслював важливість самоописових структур даних, які включають експериментальні дані та інші відомості, пов'язані з експериментом, такі як анотації на рівні функцій, протокол експерименту та опис експериментального дизайну. Існують важливі загальні риси, які можна найкраще врахувати за допомогою об'єктно-орієнтованого підходу до програмування. Це передбачає створення невеликої кількості класів, що надають будівельні блоки для більш конкретного використання. Головна мета цих класів - зберігати дані, які вже були попередньо оброблені та готові до аналізу.

Пакет Biobase надає основну інфраструктуру та визначення базових класів для роботи з біологічними даними. Головним класом є *eSet*, який містить простори для зберігання даних аналізу, фенотипової інформації, характеристик вимірних ознак, а також деталей експерименту. Цей клас є гнучким і може бути розширений для підтримки різних типів вхідних даних або додавання нових слотів, що відповідають специфіці певного експерименту. Для представлення даних експресії створено клас *ExpressionSet*, який є універсальним і може використовуватися різних типів експресійних досліджень. Подібні структури застосовуються для представлення даних мас-спектрометрії білків і проточної цитометрії.

Експериментальні набори даних відіграють ключову роль у розробці, тестуванні та демонстрації статистичних методів. На платформі Bioconductor доступні різні джерела даних, включаючи репозиторій Bioconductor та базу



даних GEO. У репозиторії Bioconductor міститься велика колекція пакетів R, що включають реальні експериментальні дані. Для роботи з базою GEO використовується пакет *GEOquery*, який забезпечує зручний інтерфейс для доступу до експериментальних наборів даних (<http://www.ncbi.nlm.nih.gov/geo>). Він дозволяє завантажувати ці дані та конвертувати їх у формат ExpressionSet.

Мікроматриці широко застосовуються для аналізу різних геномних характеристик, зокрема окремих SNP, варіацій кількості копій та поширеності мікроРНК, що є лише частиною можливих напрямків їх використання. Великий інтерес викликає інтеграція даних із цих різних джерел. Для аналізу певних типів даних існують спеціалізовані пакети, наприклад, *aCGH* та *DNAcopy*.

Окремий напрямок досліджень — генетична геноміка, де експресія генів розглядається як кількісна ознака, а її зміни аналізуються у зв'язку з варіаціями SNP. Для таких досліджень пакет *GGtools* надає програмні інструменти, що дозволяють проводити аналізи на основі наборів даних SNP та відповідних даних про експресію генів.

Дані метаболоміки

Метаболоміка — це наука, що вивчає сукупність метаболітів у біологічному зразку. Метаболіти — це малі органічні молекули, зокрема цукри, амінокислоти та ліпіди, які беруть участь у біохімічних процесах під дією ферментів (білків). Вони виконують різні функції: збереження енергії, передача сигналів та формування будівельних блоків для білків і ДНК.

У Bioconductor пакет *xcms* забезпечує потужні інструменти для аналізу даних GC-MS та LC-MS у метаболоміці. Він підтримує імпорт даних у форматах NetCDF або mzXML, які є стандартами у мас-спектрометрії. Пакет дозволяє проводити ключові етапи попередньої обробки, зокрема:

- пошук піків у хроматографічних профілях;
- вирівнювання зразків за часом;
- групування мас-спектральних сигнатур для однакових метаболітів;
- кількісний аналіз метаболітів.



Результати можуть бути легко представлені у форматі ExpressionSet, що дозволяє їх подальший аналіз за допомогою інших інструментів R та Bioconductor, таких як *limma*.

Для ідентифікації метаболітів пакет *Rdisop* інтегрується з *xcms* і дозволяє генерувати передбачувані молекулярні формули на основі мас-спектральних сигнатур, що значно спрощує інтерпретацію отриманих даних.

3. Анотації

З точки зору системної біології, глибоке розуміння біологічної системи є необхідною умовою для ефективного аналізу та інтерпретації експериментальних даних. Отримання цих знань включає анотацію біологічних зразків та молекулярних видів, однак цей процес може бути складним через розрізненість даних та різноманітність форматів їх зберігання. Bioconductor пропонує інструменти для створення та доступу до анотаційних ресурсів. Інформацію про молекулярні види можна отримати як із пакетів метаданих Bioconductor, так і з онлайн-баз даних.

Розділ анотацій у репозиторії Bioconductor містить широкий і постійно оновлюваний набір R-пакетів, які містять ресурси для анотацій. Кожен пакет може охоплювати дані про певну мікрочип-платформу або геном організму. Основна мета цих пакетів — інтеграція різних баз даних, що дозволяє пов'язувати інформацію про послідовності генів та їхні функції через ідентифікатори.

Хоча більшість анотаційних пакетів у Bioconductor зосереджені на певних мікрочипах від різних виробників, також доступні анотації повних геномів для більшості модельних організмів. Назви таких пакетів мають формат *org.Hs.eg.db*, де:

- *org* означає, що пакет містить анотацію для всього організму,
- *Hs* вказує на *Homo sapiens*,
- *eg* позначає, що первинними ключами є ідентифікатори генів Entrez,
- *db* свідчить про використання баз даних SQLite.



У сучасній обчислювальній біології можливість доступу до анотаційних даних через Інтернет є важливою функцією програмного забезпечення. Bioconductor пропонує низку інструментів для інтерактивного отримання даних, серед яких найбільш універсальним є пакет *biomaRt*. Він забезпечує зручний інтерфейс до баз даних BioMart, які використовуються для доступу до інформації про модельні організми, а також до великих геномних ресурсів.

Геномні анотації

Багато анотацій, які надають пакети метаданих Bioconductor, є геномними, тобто вони прив'язані до конкретних позицій у геномі. У базах даних геномних анотацій накопичено величезний обсяг інформації, і нові анотації можуть бути отримані шляхом різних типів аналізу. Bioconductor надає інфраструктуру для представлення та візуалізації геномних анотацій у R, що значно полегшує аналіз експериментальних даних.

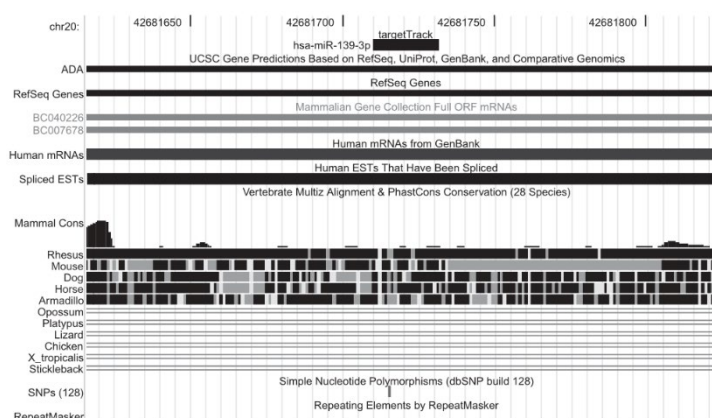


Рисунок 1 - Геномна візуалізація на прикладі rtracklayer

Інтеграція геномних анотацій на рівні послідовностей є важливим аспектом аналізу, оскільки просте зіставлення експериментальних ознак вже не є достатнім. Пакет *rtracklayer* у Bioconductor забезпечує імпорт геномних анотацій, також відомих як треки, у середовище R. Треки завантажуються як екземпляри класу *RangedData*, що використовується в пакеті *IRanges* для зберігання ранжованих ознак. Геномні ознаки визначаються за:

- хромосоною,
- початковою та кінцевою позиціями,



- ланцюгом (strand).

За допомогою *rtracklayer* можна імпортувати RangedData із файлів у форматах General Feature Format (GFF), Browser Extended Display (BED) та Wiggle (WIG), що забезпечує зручну роботу з геномними даними.

4. Представлення моделей біологічних систем

Системна біологія моделює біологічні системи, включаючи різні біологічні об'єкти, такі як гени, білки та метаболіти. Із розширенням обсягів даних у майбутньому очікується проведення ще складніших експериментів для їх збору.

Графи та мережі є ефективними інструментами для формалізації взаємодій і процесів у біологічних системах, зокрема метаболічних та сигнальних шляхів. Вони дозволяють відображати як статичні зв'язки між об'єктами, так і динаміку життєвих процесів, використовуючи системи диференціальних рівнянь.

У Bioconductor доступні три основні пакети для роботи з графами та мережами:

- *graph* – для представлення графів,
- *RBGL* – для маніпулювання ними,
- *Rgraphviz* – для їхньої візуалізації.

RBGL надає інтерфейс до бібліотеки Boost Graph Library (BGL), що забезпечує широкий набір інструментів для роботи з графами. Rgraphviz взаємодіє з Graphviz, дозволяючи створювати графічні зображення графів. Для передачі мереж у Cytoscape, потужну платформу для аналізу та візуалізації мереж, можна використовувати пакет *gaggle*.

Для роботи з мовою розмітки системної біології (SBML) у Bioconductor доступні два пакети: SBMLR та *rsbml*. Обидва дозволяють завантажувати, змінювати, моделювати та зберігати дані у форматі SBML, проте відрізняються за підходом до їх обробки:

- *rsbml* використовує зовнішню бібліотеку *libsbml*, що забезпечує глибоку інтеграцію з SBML-файлами та розширені можливості обробки.

- *SBMLR* аналізує SBML-дані безпосередньо в R, працюючи як із SBML-файлами, так і з текстовими векторами, що містять ці дані.



Висновки.

Було розглянуто низку інструментів Bioconductor, призначених для інтегрованого аналізу та візуалізації експериментальних даних, анотацій і мереж у контексті біологічних систем. Bioconductor забезпечує структури даних, процедури попередньої обробки, методи аналізу та іншу інфраструктуру для роботи з різними типами експериментальних даних, зокрема з експресією генів, а також метаболомікою, протеомікою, проточною цитометрією тощо.

Однією з головних переваг R і Bioconductor є їхні гнучкі структури даних, потужні статистичні методи, високоякісна графіка та розвинена інфраструктура для отримання та імпорту даних. Завдяки цим особливостям R і Bioconductor є зручною та ефективною платформою для дослідницького аналізу даних і розробки прототипів аналітичних методів.

Література:

- [1] Jae K. Lee. Statistical Bioinformatics: For Biomedical and Life Science Researchers. - Wiley-Blackwell. 2014. – 368 p.
- [2] Sanges, R., Cordero, F., and Calogero, R. A. (2007). oneChannelGUI: A graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language. Bioinformatics

Abstract. *This article is dedicated to exploring the potential of using the R programming language and bioinformatics packages in bioinformatics for the advancement of systems biology. The research objectives include analyzing current methods for processing biological data in the R environment, examining the functionality of bioinformatics packages, and their application in solving specific biological problems.*

Keywords: *bioinformatics, systems biology, data analysis, molecular biology, genomics.*

Стаття відправлена: 13.02.2025 р.

© Дорошенко І.В.