УДК 004.048+004.94

С.А. БАБІЧЕВ
Університет Яна Евангелиста Пуркінє в Усті на Лабі, Чехія;
Херсонський державний університет
О.Р. ЯРЕМА
Львівський національний університет імені Івана Франка

АНАЛІЗ СУЧАСНОГО СТАНУ МЕТОДІВ ФОРМУВАННЯ ПІДМНОЖИН ЗНАЧУЩИХ ТА ВЗАЄМНО ЕКСПРЕСОВАНИХ ДАНИХ ЕКСПРЕСІЇ ГЕНІВ

У статті здійснено детальний аналіз сучасних підходів до формування підмножин значущих і взаємно експресованих профілів експресії генів, отриманих за допомогою технологій ДНК-мікрочипів та секвенування молекул РНК. Це є важливим аспектом, оскільки високовимірні матриці експресії генів, які створюються в ході таких досліджень, вимагають ефективної обробки для виділення генів, які мають критичне значення для розуміння стану біологічних систем. Сучасні методи кластерного і бікластерного аналізу дають змогу зменшити кількість генів для подальшого дослідження, що є важливим для підвищення точності діагностики та аналізу біологічних процесів. Крім того, використання генної онтології (GO) дає змогу структуровано описати функціональні ролі генів у різних біологічних процесах, молекулярних функціях та клітинних компонентах, що сприяє підвищенню якості обробки даних і дає змогу зосередитися на ключових генах, які відіграють важливу роль у патологічних процесах. У роботі також розглянуто різні етапи передобробки даних, включно з видаленням неекспресованих генів, визначенням диференційно експресованих генів за допомогою інструментів, таких як DESeq2 та EdgeR, і застосуванням мета-аналізу для інтеграції результатів різних досліджень. GO-аналіз дає змогу ефективно знаходити збагачені GO терміни, які пов'язані з функціонально значущими генами, та інтерпретувати отримані результати за допомогою візуалізацій у вигляді графів і схем. Однак важливим викликом залишається стандартизація результатів та їх узгодження між різними дослідницькими групами, що є необхідним для інтеграції даних у єдину діагностичну систему. У статті наголошується на важливості подальшого вдосконалення підходів до аналізу експресії генів та інтеграції даних, що дасть змогу підвищити ефективність досліджень у галузі діагностики захворювань та персоналізованої медицини.

Ключові слова: експресія генів, система діагностики, генна онтологія, ентропія Шеннона, статистичні критерії.

S.A. BABICHEV
Jan Evangelista Purkyně University in Ústi nad Labem, Czech Republic;
Kherson State University
O.R. YAREMA
Ivan Franko National University of Lviv

ANALYSIS OF THE CURRENT STATE OF METHODS FOR SUBSET FORMATION OF SIGNIFICANT AND MUTUALLY EXPRESSED GENE EXPRESSION DATA

The article provides a comprehensive analysis of the modern approaches to forming subsets of significant and mutually expressed genes based on gene expression data obtained through DNA microarray and RNA sequencing technologies. This topic is of particular relevance since the high-dimensional gene expression matrices generated in such studies require effective preprocessing to identify genes that are critical for understanding the biological systems' conditions. Modern clustering and biclustering methods play a vital role in reducing the number of genes for further analysis, thereby improving the accuracy of diagnostics and biological process analysis. Moreover, the use of Gene Ontology (GO) facilitates the structured description of the functional roles of genes in various biological processes, molecular functions, and cellular components. This enhances data processing quality and enables researchers to focus on key genes playing significant roles in pathological processes. The article also addresses different stages of data preprocessing, including the removal of non-expressed genes, the identification of differentially expressed genes using tools like DESeq2 and EdgeR, and the application of meta-analysis to integrate results from multiple studies. GO analysis allows researchers to effectively identify enriched GO terms associated with functionally significant genes and interpret the results through visualizations such as graphs and diagrams. However, one of the key challenges remains the standardization of results and ensuring consistency across various research groups, which is essential for integrating data into a unified diagnostic system. The paper highlights the importance of further enhancing approaches to gene expression analysis and data integration, which will significantly improve the efficiency of bioinformatics research in disease diagnostics and personalized medicine.

Key words: gene expression, diagnostic system, gene ontology, Shannon entropy, statistical criteria.

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

Постановка проблеми

Проблема формування підмножин значущих генів з великих наборів даних є надзвичайно актуальною в контексті сучасних підходів до персоналізованої медицини. Процес отримання експериментальних даних за допомогою технології ДНК-мікрочипів та секвенування РНК приводить до створення великої високовимірної матриці, де кожен рядок представляє досліджуваний об'єкт, а стовпці відповідають ідентифікаторам генів, значення експресії яких визначають стан об'єктів. Однак після видалення неекспресованих або слабко експресованих генів у наборі даних зазвичай залишаються десятки тисяч генів (близько 10–20 тис.), що значно ускладнює подальший аналіз і обробку. Цей виклик є критичним для розроблення ефективних діагностичних систем і реконструкції генних регуляторних мереж, особливо в межах персоналізованої медицини.

Персоналізована медицина зосереджується на індивідуалізації підходів до лікування кожного пацієнта, зокрема на основі аналізу їхньої генетичної інформації. Велика кількість генів, що вивчаються, ускладнює аналіз, що створює потребу у методах фільтрації та зменшення даних. Сучасні підходи, такі як кластерний і бікластерний аналіз, дають змогу ідентифікувати підмножини значущих генів, які відіграють ключову роль у визначенні станів досліджуваних об'єктів. Ці підмножини формують основу для моделювання стану пацієнтів, прогнозування прогресування захворювань та вибору відповідних методів лікування.

Формування підмножин значущих генів за допомогою методів, таких як аналіз генних онтологій (GO), дає змогу структуровано описати функції генних продуктів у біологічних процесах. Це надає інструменти для глибшого розуміння складних молекулярних механізмів і взаємодій, які лежать в основі патологій. GO-аналіз пропонує стандартизовану термінологію для опису біологічних функцій, що покращує порівнянність результатів між різними дослідженнями і полегшує інтеграцію даних з різних джерел, що є важливим для розроблення діагностичних систем у персоналізованій медицині. Таким чином, розробка методів формування підмножин значущих генів є важливим кроком для підвищення точності та ефективності діагностики на основі геномних даних. Це також закладає основу для інтеграції таких даних у системи прогнозування стану пацієнтів, що є важливим елементом у розвитку підходів до персоналізованого лікування та терапії.

Аналіз останніх досліджень та публікацій

Сучасні дослідження в галузі формування та обробки даних експресії генів зосереджені на розробці ефективних методів ідентифікації значущих генів, які використовуються для діагностики захворювань [1; 2]. Використання технологій ДНК-мікрочипів та RNA-seq дає змогу отримувати високовимірні дані, однак їх обробка є проблематичною через велику кількість генів, які необхідно відфільтрувати для подальшого аналізу. Кластерні та бікластерні методи стали ключовими інструментами для сегментації даних, що дає змогу ідентифікувати групи генів зі схожими патернами експресії [3; 4]. Генна онтологія (GO) широко використовується для функціональної анотації генів, значно підвищуючи точність ідентифікації генів, які відіграють ключові ролі в біологічних процесах та патологіях [5]. Однак проблема об'єднання результатів з різних джерел та стандартизації критеріїв значущості генів залишається невирішеною, що створює виклики в інтеграції даних для створення універсальних діагностичних систем.

Мета дослідження

Метою цього дослідження є аналіз сучасних методів формування підмножин значущих та взаємно експресованих генів з акцентом на вирішенні проблем стандартизації та інтеграції для підвищення точності діагностики на основі експресії генів.

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

Виклад основного матеріалу дослідження

Формування підмножин значущих генів та їх подальший аналіз є важливим завданням у біоінформатиці та молекулярній біології. Використання генних онтологій (GO) дає змогу структуровано описувати функції продуктів генів, що сприяє глибшому розумінню біологічних процесів, молекулярних функцій та клітинних компонентів. Онтологія, як філософська дисципліна, вивчає природу буття, сутність речей та категорії існування [6]. У контексті інформатики та біоінформатики онтологія описується як формальне представлення знань у певній галузі за допомогою концепцій та взаємозв'язків між ними. Вона забезпечує спільний словник для дослідників та інструменти для інтеграції та аналізу даних. Онтології дають змогу моделювати складні системи та процеси, допомагаючи зрозуміти їхню структуру та функції.

Генна онтологія є одним з найвідоміших та найширше використовуваних інструментів для анотації генів та їхніх продуктів у біології. GO надає структурований словник для опису функцій генів, їхньої участі в біологічних процесах та їхньої локалізації всередині клітини [5; 7]. GO складається з трьох основних компонентів:

– *молекулярна функція*: описує основні дії, які виконують продукти генів на молекулярному рівні, такі як зв'язування або каталіз;

– *біологічний процес*: визначає серію подій або молекулярних функцій, що спільно досягають конкретної біологічної мети, такої як метаболічні процеси або сигнальні шляхи;

– *клітинний компонент*: описує місця всередині клітини, де функціонують гени, такі як органели або макромолекулярні комплекси.

GO була створена у відповідь на необхідність стандартизації термінів, що використовуються для опису продуктів генів та їхніх функцій у різних організмах. Основна мета GO полягає в наданні єдиного словника для опису продуктів генів у будь-якому організмі, що дає змогу покращити порівнянність і аналіз даних у різних дослідженнях. Терміни GO організовані ієрархічно, де кожен посідає чітке місце в деревоподібній структурі, що показує взаємозв'язки між різними термінами (рис. 1) [8]. Це дає змогу не тільки знайти інформацію про конкретний продукт гена, але й зрозуміти, як він взаємодіє з іншими генами та біологічними процесами. Таким чином, GO дає змогу стандартизувати біологічну інформацію та полегшує обмін даними між різними дослідницькими групами, а також інтеграцію даних з різних джерел. Методологія застосування аналізу GO передбачає наявність таких кроків.

1) Збір даних та попередня обробка. Дані можуть бути отримані з RNA-seq або мікрочип-аналізу, які надають інформацію про експресію генів у різних умовах. Нормалізація даних необхідна для усунення технічних варіацій, що можуть вплинути на результати аналізу. Це забезпечує порівняння експресії генів між різними зразками.

2) Визначення диференційно експресованих генів (DEGs). Для визначення диференційно експресованих генів використовуються такі інструменти, як DESeq2 [9] або EdgeR [10]. Вони дають змогу статистично оцінити зміни в експресії генів між контрольними та експериментальними умовами, що дає можливість виявити гени, які показують значущі зміни в експресії під час аналізу різних зразків.

3) Використання мета-аналізу для об'єднання результатів. Мета-аналіз дає змогу об'єднати результати з різних досліджень для підвищення надійності та узагальненості висновків. Для цього можуть бути використані різні методи комбінування p-значень, як-от тести Фішера та Колмогорова-Смірнова, що дають змогу інтегрувати результати з різних досліджень у єдину картину [11].

4) Аналіз генної онтології. GO аналіз включає кілька підходів.

– Over-Representation Analysis (ORA). Цей метод оцінює, чи є певні GO-терміни надмірно представленими у вибраній підмножині генів. Використовуються такі інструменти, як DAVID та Qiagen IPA, які дають змогу оцінити статистичну значущість збагачення GO-термінів у підмножині генів [12].

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

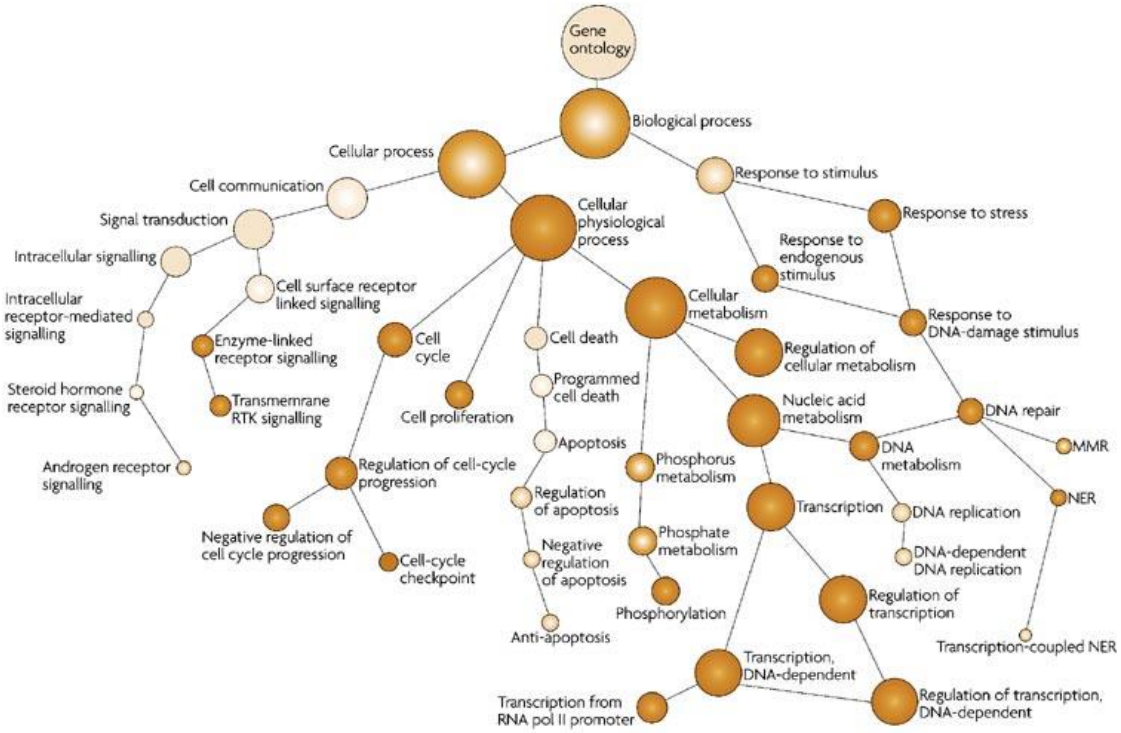

Рис. 1. Ілюстрація графу розподілу GO-термів

– Functional Class Scoring (FCS). Враховує всі гени та використовує методи ранжування, такі як GSEA, для визначення значущих наборів генів. Цей підхід дає змогу виявити, які гени з набору показують найзначущі зміни в експресії, а також як ці зміни впливають на біологічні процеси [11].

– Pathway Topology (PT). Аналізує взаємодії генів у контексті біологічних шляхів. Використовуючи інструменти, такі як Cytoscape та його плагіни, як-от ClueGO, можемо візуалізувати та аналізувати взаємозв'язки між генами та біологічними шляхами [11].

5) Візуалізація та інтерпретація результатів. Результати GO-аналізу часто надаються у вигляді графіків та схем, які показують збагачені GO-терміни та їхні ієрархічні зв'язки. Інструменти, такі як Cytoscape з плагінами ClueGO або EnrichmentMap, допомагають створювати інтерактивні візуалізації, що полегшує інтерпретацію даних та виявлення нових взаємозв'язків між генами [12; 13].

Таким чином, формування підмножин значущих генів з подальшим GO-аналізом є потужним інструментом для дослідження функціональної ролі генів. Використання сучасних методів та інструментів дає змогу глибоко зрозуміти біологічні процеси та виявити нові взаємозв'язки між генами. Цей підхід допомагає не лише у визначенні важливих біологічних механізмів, але й у прогнозуванні нових функцій генів, що може мати значний вплив на подальші дослідження та розроблення терапевтичних стратегій.

На рис. 2 зображено структурну схему покрокової процедури застосування аналізу GO для виділення значущих генів на основі анотації GO.

Як зазначають автори [14], загалом практична реалізація вищенаведеної процедури включає такі етапи.

1) Підготовка даних. На цьому етапі формується список генів, наявних у досліджуваних даних. Далі ці гени анотуються з використанням наявних баз даних, що надають інформацію про їх асоціацію з різними термами GO.

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

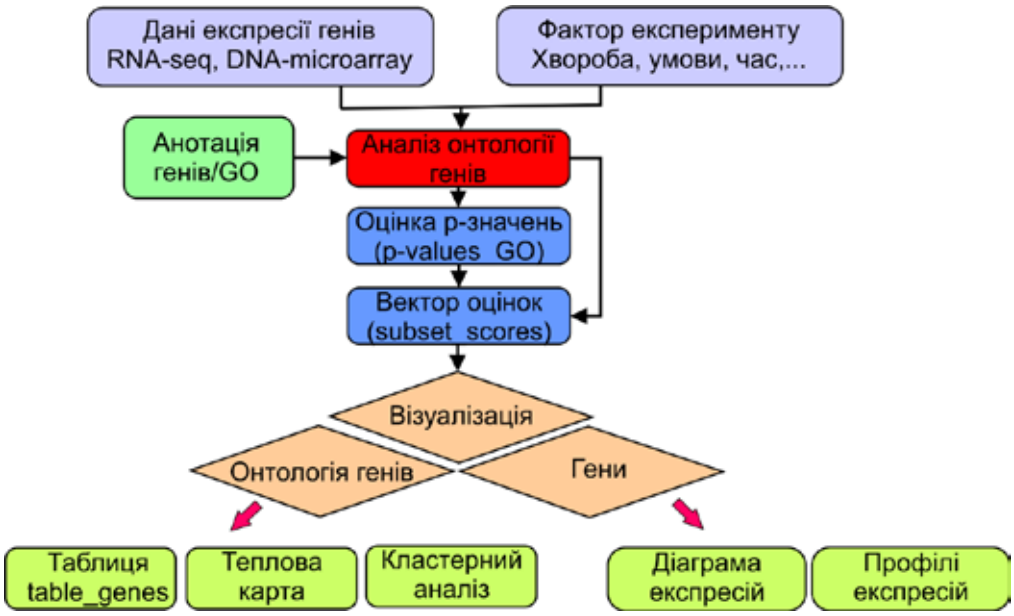

Рис. 2. Структурна схема покрокової процедури застосування аналізу GO для формування підмножини значущих генів

2) Створення об'єкта GO. На цьому етапі створюється об'єкт, який містить інформацію про всі терміни GO та їх взаємозв'язки.

3) Застосування тестової статистики. До даних експресії генів застосовуються статистичні тести для порівняння частоти кожного терму GO в наборі вибраних генів з частотою у фоновому наборі (загальна популяція генів). Зазвичай на цьому етапі використовуються тести ANOVA, Фішера та Колмогорова-Смірнова.

4) Аналіз збагачення термів GO. Оцінюється, чи є певні терми GO надмірно представленими (збагаченими) серед вибраних генів. Також обчислюється p-значення для кожного терму GO, що вказує на ймовірність випадкового отримання такої кількості генів, що відповідають цьому терму.

5) Корекція на множинні порівняння. Оскільки в аналізі GO виконується велика кількість тестів, необхідна корекція для уникнення помилкових позитивних результатів. Зазвичай на цьому етапі застосовувався тест Бенджаміні-Хохберга для корекції p-значень.

6) Інтерпретація та візуалізація результатів. Оцінюються та аналізуються значущі терми GO, які були ідентифіковані як збагачені серед вибраних генів, аналізуються зв'язки між різними термами, створюються мережеві діаграми, що відображають біологічні шляхи або процеси. Візуалізація результатів включає створення мережевих діаграм найбільш збагачених GO термів.

7) Формування списку значущих генів. Створюється підмножина даних експресії генів, що містять значущі за GO гени, для подальшого аналізу та використання у системах діагностики стану об'єктів або реконструкції та моделювання генних регуляторних мереж.

У джерелі [14] автори представили результати моделювання щодо застосування наведеної процедури для формування підмножини значущих генів з подальшою оцінкою ефективності цієї процедури шляхом застосування класифікатору до зразків, що містили як атрибути виділені значущі гени. Процедура моделювання здійснювалася з використанням функцій та модулів пакету *topGO* [15] пакета *Bioconductor* [16] середовища програмування R. Моделювання процесу застосування аналізу GO проводилося з використанням даних експресії генів пацієнтів, хворих на чотири типи ракових захворювань: у 502 пацієнтів було виявлено плоскоклітинну карциному легень (LUSC), у 541 – аденокарциному легень (LUAD), у 542 – нирково-клітинну

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

карциному (KIRC), у 534 – гліому низького ступеня злоякісності мозку (LGG). Дані були отримані за допомогою методу секвенування RNA (RNA-seq) в рамках проєкту "The Cancer Genome Atlas" (TCGA) і доступні на вебсайті проєкту [17]. Початковий набір даних містив 2 119 зразків та 19 947 генів. Після видалення генів, які не експресувалися (мали нульову експресію для всіх зразків), кількість генів скоротилася до 19 043. Анотація генів через порівняння з відповідними ідентифікаторами в базах даних людського геному (з використанням модуля "org.Hs.eg.db") зменшила кількість генів до 18 930, оскільки не анотовані в базі даних гени були видалені. Результат застосування наведеної процедури на основі аналізу GO привів до зменшення кількості генів до 14 488, при цьому для розділення генів на значущі та незначущі було використано p-значення 0,01, тобто гени вважалися значущими як за тестом Фішера, так і за тестом Колмогорова-Смірнова з ймовірністю 99%. Результати класифікації зразків за результатами досліджень, що представлені у джерелі [14], зображені у табл. 1.

Однак варто зазначити, що кількість генів залишається доволі великою. Крім того, прийняття рішень щодо стану об'єкта на основі великої бази даних містить значний елемент суб'єктивності. Підвищити об'єктивність у цьому разі можна шляхом розпаралелювання процесу обробки інформації за допомогою кластерного або бікластерного аналізу. На кожному рівні значущі гени можна визначати з використанням аналізу GO. Окрім того, дослідження підмножин значущих генів, що відповідають окремим термам GO, враховуючи відповідний біологічний процес (хвороба, що досліджується) також може підвищити об'єктивність отримання результату.

Таблиця 1

Результати класифікації даних на основі значущих генів, виділених із застосуванням аналізу GO [14]

Class	Prediction				Precision	Recall	F1	Accuracy
	kirc	lgg	luad	lusc				
kirc	162	1	1	0	0,988	1,000	0,994	97,6%
lgg	0	159	0	0	1,000	0,994	0,997	
luad	0	0	155	7	0,957	0,957	0,957	
lusc	0	0	6	143	0,960	0,953	0,957	

Застосування аналізу генної онтології на першому етапі передобробки даних експресії генів дає змогу виділити значущі гени, враховуючи тип біологічного організму, при цьому рівень значущості може бути визначений із застосуванням різних статистичних тестів (тести Фішера, Колмогорова-Смірнова тощо). Але слід зазначити, що застосування різних тестів може приводити до неузгоджених результатів. Тому у джерелі [14] авторами запропоновано метод формування підмножин значущих генів на основі комплексного застосування тестів Фішера та Колмогорова-Смірнова. Ген вважався значущим, якщо він ідентифікований як значущий за обома тестами. При цьому кількість значущих генів визначається значенням гіперпараметру алгоритму p-value (ймовірність, що ген не є значущим), яке визначається емпіричним шляхом у процесі моделювання. Більш того, за однакового значення p кількість значущих генів суттєво залежить від якості експериментальних даних і може варіюватися у достатньо широкому інтервалі, що вносить певну суб'єктивність у процес фільтрації даних на етапі їх передобробки.

Аналіз даних експресії генів пацієнтів, що досліджувалися на різні типи ракових захворювань, що складають базу даних, сформовану в рамках TCGA проєкту [17], показав, що у початковому стані дані містять приблизно 60 000 генів. Видалення неекспресованих для всіх зразків генів призводить до зменшення кількості генів приблизно до 25 000. Як показано у джерелі [14], застосування аналізу генної онтології дає змогу скоротити кількість генів приблизно до 19 000, але при цьому кількість генів залишається достатньо великою, що ускладнює процес

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

подальшої обробки даних. Підвищити ефективність застосування методу фільтрації генів на основі аналізу генної онтології можна шляхом більш ретельного формування даних на попередньому етапі застосування аналізу ГО.

У роботі [18] авторами представлені результати досліджень щодо формування підмножин значущих генів із застосуванням статистичних та ентропійних критеріїв, при цьому остаточне рішення щодо рівня значущості гена ухвалювалось на основі застосування як функції бажаності Харрінгтона, так і моделі нечіткого логічного виводу. На думку авторів, рівень значущості гена визначається на основі комплексного застосування трьох параметрів: максимального значення експресії профіля експресії гена, значення дисперсії та ентропії Шеннона цього профіля. При цьому передбачалося, що більші максимальне значення експресії та дисперсії та менше значення ентропії Шеннона відповідають більш високому рівню значущості цього гена. Визначення границі, що розділяє гени на значущі та незначущі, здійснювалося на основі граничних значень відповідних параметрів за формулою:

$$\left\{e_{ij}\right\} = \left\{ \begin{array}{c} \max\limits_{i=1,n} e_{ij} \geq e_{lim}, and\ var\left(e_j\right) \geq var_{lim}, \\ and\ entr\left(e_j\right) \leq entr_{lim} \end{array} \right\}, j = \overline{1,m}, \quad (1)$$

де n – кількість зразків, що досліджуються; m – кількість профілів експресії генів; e_{ij} – значення експресії гена, що відповідає i-му зразку та j-му профілю; e_{lim}, var_{lim} і $entr_{lim}$ – граничні значення експресії, дисперсії та ентропії Шеннона відповідно. Граничні значення у цьому разі визначалися емпіричним шляхом у процесі моделювання, враховуючи приблизну кількість генів, які повинні складати підмножину експериментальних даних для подальшого дослідження.

Проте слід зазначити, що запропонована авторами концепція має значний недолік. Високе значення дисперсії певного профілю експресії гена або низьке значення ентропії Шеннона (за цими критеріями профіль вважається значущим) за умов низьких абсолютних значень експресії генів у всіх досліджуваних об'єктів не гарантує, що цей профіль дійсно є значущим. Це пов'язано з тим, що за абсолютними значеннями експресії він не забезпечує високу точність ідентифікації досліджуваних об'єктів. Таким чином, виникає потреба у визначенні пріоритетності виконання відповідних операцій або через встановлення послідовності їхнього застосування, або через ініціалізацію ваг для кожної операції, при цьому важливо обґрунтувати вибір значення відповідної ваги. У роботі [19] автори представили результати досліджень, присвячених формуванню підмножин профілів експресії генів різного рівня значущості з використанням системи нечіткого логічного виведення. Як експериментальні дані, авторами були застосовані дані експресії генів пацієнтів, що досліджувалися на ранній стадію рака легенів. Дані GSE19188 були взяті з доступної бази даних "Gene Expression Omnibus" [20] і містили дані експресії генів 156 пацієнтів, з яких 65 були ідентифіковані за результатами клінічних досліджень як здорові, а у 91 була ідентифікована ракова пухлина у початковій стадії. У початковому стані дані містили 54 675 профілів експресії генів. У цьому дослідженні пріоритетність певних операцій враховується під час створення бази нечітких правил, яка є основою нечіткої моделі. Діапазон зміни значень вхідних параметрів у запропонованій моделі визначається на основі аналізу загальної статистики. Спочатку для кожного профілю обчислюється максимальне значення експресії генів, а потім формується загальна статистика для отриманого вектору максимальних значень, вектору дисперсії профілів експресії генів та ентропії Шеннона. Для створення нечіткої моделі використовувалися міжквантільні інтервали зміни максимальних абсолютних значень, дисперсії та ентропії Шеннона, які були розділені на три інтервали з відповідними термами. Для дисперсії та максимальних значень експресії генів: 0%≤x<25% – «Низьке» (Н); 25%≤x<75% – «Середнє» (С); x≥75% – «Високе» (В). Для ентропії

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

Шеннона: x≥75% – «Високе» (В); 25%≤x<75% – «Середнє» (С); x<25% – «Низьке» (Н). Діапазон варіювання вихідного параметра (значущість профілю) змінювався від 0 до 100 і був поділений на п'ять рівних інтервалів: 0≤y<20 – «Дуже низьке» (ДН); 20≤y<40 – «Низьке» (Н); 40≤y<60 – «Середнє» (С); 60≤y<80 – «Високе» (В); 80≤y≤100 – «Дуже високе» (ДВ). Щодо функцій належності нечітких множин, для вхідних параметрів з термами «Низьке» та «Високе» застосовувалася трапецеїдальна функція належності, а для середнього діапазону (С) – трикутна функція належності. Для всіх підмножин вихідного параметра використовувалися трикутні функції належності. Параметри функцій належності для вхідних критеріїв передбачали коригування під час моделювання з урахуванням розподілу значень експресії генів у досліджуваних експериментальних даних. На рис. 3 зображено покрокову процедуру, що була реалізована в процесі моделювання.

Рис. 3. Структурна схема покрокової процедури формування підмножин профілів експресії генів різного ступеня значущості за статистичними та ентропійними критеріями із застосуванням нечіткої логіки

Результати моделювання показали, що з 54 675 профілів експресії генів тільки 29 були ідентифіковані як «Дуже висока ступінь значущості». З цієї причини групи зі ступенем значущості «Дуже висока» і «Висока» були об'єднані для подальшого моделювання. Як результат, початкова множина профілів експресії генів за ступенем значущості була розділена на чотири підмножини: 16 734 – «Висока (Hg)»; 13 076 – «Середня (Md)»; 13 605 – «Низька (Low)»; 14 240 – «Дуже низька (VLow)». Оцінка адекватності моделі здійснювалася шляхом застосування класифікатору до об'єктів, атрибутами яких були сформовані підмножини профілів експресії генів з подальшою оцінкою результатів класифікації шляхом розрахунку критеріїв якості класифікації (Accuracy, F1-index, Matthews Correlation Coefficient (MCC)) та застосування ROC-аналізу з розрахунком площі під ROC-кривою (AUC). Результати моделювання представлені у табл. 2. Аналіз результатів моделювання дає змогу дійти висновку щодо адекватності запропонованого авторами методу формування підмножин профілів експресії генів на основі системи нечіткого логічного виведення, оскільки значення усіх критеріїв якості класифікації об'єктів позитивно корелюють з рівнем значущості профілів експресії генів.

Таблиця 2

Результати моделювання класифікації об'єктів за даними експресії генів різного ступеня значущості, що отримані з використанням нечіткої логіки

Значущість генів	Критерії якості класифікації об'єктів			
	точність, %	F-індекс	MCC	AUC
Висока	98,4	0,992	0,967	0,999
Середня	93,5	0,967	0,873	0,998
Низька	90,3	0,894	0,801	0,955
Дуже низька	85,5	0,862	0,701	0,950

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

Однак слід зазначити, що запропонований авторами метод має суттєві недоліки. По перше, висока трудомісткість. Базу правил та значення функцій належності необхідно адаптувати до експериментальних даних, що використовуються на поточному етапі моделювання. Другим суттєвим недоліком є високий рівень суб'єктивізму під час формування бази нечітких правил та налаштуванні нечіткої моделі. Це обмежує застосування запропонованого авторами методу на етапі передобробки даних експресії генів.

У роботі [21] авторами реалізовано інший підхід щодо формування узагальненого показника значущості профілів експресії генів із застосуванням статистичних критеріїв та ентропії Шеннона, що базується на функції бажаності Харрінгтона, яка наразі широко застосовується в різних наукових дослідженнях. Цей метод заснований на такому рівнянні:

$$d = \exp(-\exp(-Y)), \tag{2}$$

де Y – безрозмірний параметр, що варіюється в межах від -2 до 5, а d – приватна бажаність, що відповідає кожному з критеріїв, які використовуються для формування узагальненого показника значущості.

Важливо зазначити, що межі, які визначають крайні інтервали значень бажаності – 0,2 (незадовільно – погано) і 0,8 (добре – відмінно), – є умовними та можуть бути скориговані залежно від характеру змін параметрів, що подаються на вхід моделі. Фіксовані межі 0,37 = 1/e (погано – задовільно) і 0,63 = 1 – 1/e (добре – відмінно) відповідають точкам перетину функції бажаності. В рамках запропонованої авторами моделі передбачалося, що залежність параметру *Y* і значень критеріїв, що подаються на вхід моделі, відповідає лінійному закону.

Результати моделювання з розрахунком критеріїв якості класифікації зразків, що представлені у дослідженнях авторів, дають змогу дійти висновку щодо суттєво більш низької ефективності методу на основі функції бажаності Харрінгтона порівняно з моделлю на основі системи нечіткого логічного виведення. За граничного значення узагальненого показника 0,04, що розділяв значущі та незначущі профілі експресії генів, з 54 675 було виділено тільки 9 630 профілів. Точність класифікації зразків дорівнювала 91,9%, що значно менше, ніж у моделі на основі нечіткої логіки. Цей факт можна пояснити нелінійним характером залежності значень відповідних критеріїв від значення показника Y, що, безумовно, впливає на об'єктивність рішення щодо визначення ступеня значущості відповідного профіля експресії гену.

Висновки

У статті представлено огляд сучасних методів формування підмножин значущих та взаємно експресованих генів, зосереджений на покращенні якості даних експресії для подальшого аналізу та їх застосування в системах діагностики. Особливу увагу приділено застосуванню генної онтології (GO) для виділення генів, що беруть участь у ключових біологічних процесах. Використання аналізу GO дає змогу структурувати велику кількість експериментальних даних і підвищити точність у виявленні взаємозв'язків між генами та біологічними процесами.

Розглянуто етапи обробки даних, які включають збір та попередню обробку даних, визначення диференційно експресованих генів, застосування мета-аналізу для об'єднання результатів різних досліджень та візуалізацію результатів. Використання таких методів, як Over-Representation Analysis, Functional Class Scoring, та Pathway Topology, дає змогу покращити точність у відборі значущих генів для подальших досліджень та систем діагностики.

Одним з ключових викликів залишається стандартизація результатів з різних джерел даних, що дасть змогу підвищити точність та узагальненість діагностичних систем. Важливим напрямом майбутніх досліджень є розробка більш точних та надійних методів для формування підмножин значущих генів, які сприяють кращому розумінню біологічних процесів

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

і, відповідно, удосконаленню терапевтичних підходів. Таким чином, стаття підкреслює важливість інтеграції сучасних біоінформаційних методів для досягнення більшої точності у діагностиці на основі експресії генів, що є важливим компонентом персоналізованої медицини.

Список використаної літератури

1. Green H. RNA Processing. *Cold Spring Harbor Perspectives in Biology*. 2017. Vol. 9 (5). Art. no. a032425.
2. Wang Z., Gerstein M., Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2019. Vol. 10 (1). P. 57–63.
3. Bolstad B.M., Irizarry R.A., Astrand M., Speed T.P. A Comparison of Normalization Methods for High-Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*. 2023. Vol. 19. P. 185–193.
4. Qin S., Tang X., Chen Y. et al. mRNA-based therapeutics: powerful and versatile tools to combat diseases. *Signal Trans. and Targeted Therapy*. 2022. Vol. 7 (1). Art. no. 166.
5. Ashburner M., Ball C.A., Blake J.A., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 2000. Vol. 25 (1). P. 25–29.
6. Gruber T.R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*. 1993. Vol. 5 (2). P. 199–220.
7. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*. 2021. Vol. 49 (D1). P. D325–D334.
8. Huntley R.P., Sawford T., et al. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *GigaScience*. 2015. Vol. 4 (1). Art. no. 4.
9. Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014. Vol. 15. Art. No. 550.
10. Chen Y., Chen L., Lun A.T.L., et al. EDGER 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *bioRxiv*. 2024.
11. Pandey D., Perumal P.O. Improved meta-analysis pipeline ameliorates distinctive gene regulators of diabetic vasculopathy in human endothelial cell (hECs) RNA-Seq data. *PLoS ONE*. 2023. Vol. 18 (11). Art. no. e0293939.
12. Suzi A., James B., Seth C., et al. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023. Vol. 224(1). Art. no. iyad031.
13. Shin M.G., Pico A.R. Using published pathway figures in enrichment analysis and machine learning. *BMC Genomics*. 2023. Vol. 24. Art. no. 713.
14. Babichev S., Korobchynskyi M., Rudenko M., Batenko H. Applying biclustering technique and gene ontology analysis for gene expression data processing. *CEUR Workshop Proceedings*. 2024. Vol. 3675. P. 14–28.
15. Ouma W.Z., Pogacar K., Grotewold E. Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. *PLoS Computational Biology*. 2018. Vol. 14 (4). Art. no. e1006098.
16. Bioconductor: Open source software for Bioinformatics. 2024. July, 29. URL: https://www.bioconductor.org.
17. The Cancer Genome Atlas Program (TCGA). National Cancer Institution. Center for Cancer Genomics. 2024. July, 27. URL: https://www.cancer.gov/ccg/research/genome-sequencing/tcga.
18. Babichev S., Škvor J. Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and Classification Methods. *Diagnostics*. 2020. Vol. 10 (8). Art. no. 584.
19. Liakh I., Babichev S., Durnyak B., Gado I. Formation of Subsets of Co-expressed Gene Expression Profiles Based on Joint Use of Fuzzy Inference System, Statistical Criteria and

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

Shannon Entropy. *Lecture Notes on Data Engineering and Communications Technologies*. 2023. Vol. 149. P. 25–41.
20. Hou J., Aerts J., den Hamer B., et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*. 2010. Vol. 5. Art. no. e10312.
21. Yasinska-Damri L., Babichev S., Spivakovsky A., Lemeshchuk O. Formation and Analysis of Gene Expression Data Based on the Joint Use of Data Mining and Machine Learning Techniques. *CEUR Workshop Proceedings*. 2023. Vol. 3373. P. 87–98.

References

1. Green, H. (2017). RNA Processing. *Cold Spring Harbor Perspectives in Biology*. 9 (5). Art. no. a032425 [in English].
2. Wang, Z., Gerstein, M., & Snyder, M. (2019). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 10 (1), 57–63 [in English].
3. Bolstad, B.M., Irizarry, R.A., Astrand, M., & Speed, T. P. (2023). A Comparison of Normalization Methods for High-Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*. 19, 185–193 [in English].
4. Qin, S., Tang, X., & Chen, Y. et al. (2022). MRNA-based therapeutics: powerful and versatile tools to combat diseases. *Signal Trans. and Targeted Therapy*. 7 (1). Art. no. 166 [in English].
5. Ashburner, M., Ball, C.A., & Blake, J.A., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 25 (1). 25–29 [in English].
6. Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5 (2), 199–220 [in English].
7. The Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*. 49 (D1). D325–D334 [in English].
8. Huntley, R.P., & Sawford, T., et al. (2015). Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *GigaScience*. 4 (1). Art. no. 4 [in English].
9. Love, M.I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 15. Art. no. 550 [in English].
10. Chen, Y., Chen, L., & Lun, A.T.L, et al. (2024). EDGER 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *bioRxiv* [in English].
11. Pandey, D., & Perumal, P.O. (2023). Improved meta-analysis pipeline ameliorates distinctive gene regulators of diabetic vasculopathy in human endothelial cell (hECs) RNA-Seq data. *PLoS ONE*. 18 (11). Art. no. e0293939 [in English].
12. Suzi, A., James, B., & Seth, C., et al. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*. 224 (1). Art. no. iyad031 [in English].
13. Shin, M.G., & Pico, A.R. (2023). Using published pathway figures in enrichment analysis and machine learning. *BMC Genomics*. 24. Art. no. 713 [in English].
14. Babichev, S., Korobchynskyi, M., Rudenko, M., & Batenko, H. (2024). Applying biclustering technique and gene ontology analysis for gene expression data processing. *CEUR Workshop Proceedings*. 3675. 14–28 [in English].
15. Ouma, W.Z., Pogacar, K., & Grotewold, E. (2018). Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. *PLoS Computational Biology*. 14 (4). Art. no. e1006098 [in English].
16. Bioconductor: Open source software for Bioinformatics. (2024). July, 29. Retrieved from: https://www.bioconductor.org/ [in English].
17. The Cancer Genome Atlas Program (TCGA). National Cancer Institution. Center for Cancer Genomics. (2024). July, 27. Retrieved from: https://www.cancer.gov/ccg/research/genome-sequencing/tcga [in English].

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2

18. Babichev, S., & Škvor, J. (2020). Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and Classification Methods. *Diagnostics*. 10 (8). Art. no. 584 [in English].
19. Liakh, I., Babichev, S., Durnyak, B., & Gado, I. (2023). Formation of Subsets of Co-expressed Gene Expression Profiles Based on Joint Use of Fuzzy Inference System, Statistical Criteria and Shannon Entropy. *Lecture Notes on Data Engineering and Communications Technologies*, 149, 25–41 [in English].
20. Hou, J., Aerts, J., & den Hamer, B., et al. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*. 5. Art. no. e10312 [in English].
21. Yasinska-Damri, L., Babichev, S., Spivakovsky, A., & Lemeshchuk O. (2023). Formation and Analysis of Gene Expression Data Based on the Joint Use of Data Mining and Machine Learning Techniques. *CEUR Workshop Proceedings*. 3373. 87–98 [in English].

Бабічев Сергій Анатолійович – д.т.н., професор, професор кафедри інформатики Університету Яна Євангеліста Пуркінє в Усті на Лабі, Чехія; професор кафедри фізики Херсонського державного університету. E-mail: sergii.babichev@ujep.cz, sbabichev@ksu.ks.ua, ORCID: 0000-0001-6797-1467.

Ярема Олег Романович – к.т.н., доцент, доцент кафедри цифрової економіки та бізнес-аналітики Львівського національного університету імені Івана Франка. E-mail: oleh.yarema@lnu.edu.ua, ORCID: 0000-0003-3736-4820.

Babichev Sergii Anatoliiovych – Doctor of Technical Sciences, Professor, Professor at the Department of Informatics at Jan Evangelista Purkyně University in Ústí nad Labem, Czech Republic; Professor at the Department of Physics at Kherson State University. E-mail: sergii.babichev@ujep.cz, sbabichev@ksu.ks.ua, ORCID: 0000-0001-6797-1467.

Yarema Oleh Romanovych – Candidate of Technical Sciences, Associate Professor, Associate Professor at the Department of Digital Economy and Business Analytics at Ivan Franko National University of Lviv. E-mail: oleh.yarema@lnu.edu.ua, ORCID: 0000-0003-3736-4820.

https://doi.org/10.32782/mathematical-modelling/2024-7-2-2