УДК 004.048

СУЧАСНИЙ СТАН МЕТОДІВ РЕКОНСТРУКЦІЇ ГЕННИХ РЕГУЛЯТОРНИХ МЕРЕЖ

І. М. Лях

Ужгородський національний університет,
пл. Народна, 3, Ужгород, 88000, Україна

Проаналізовано сучасні методи реконструкції генних регуляторних мереж (ГРМ) на основі даних експресій генів, виокремлено їхні переваги та недоліки й окреслено шляхи подальшого удосконалення. Генна регуляторна мережа подана у вигляді орієнтованого або неорієнтованого графа, при цьому вага дуги визначає силу відповідного зв'язку. Розглянуто наступні методи реконструкції ГРМ: на основі аналізу кореляцій; на основі аналізу взаємної інформації між генами та/або транскрипційними факторами; на основі мережі Байєса; на основі диференційних рівнянь; на основі регресійного аналізу. Аналіз відповідних методів дозволив зробити висновок, що сучасний стан розвитку даного напрямку визначається гібридизацією існуючих моделей, методів та алгоритмів. Застосування ансамблів методів реконструкції ГРМ сприяє підвищенню адекватності реконструйованої ГРМ біологічним генним мережам, що також сприяє кращому розумінню характеру взаємодії генів та/або транскрипційних факторів у мережі.

Ключові слова: *генна регуляторна мережа, алгоритм реконструкції генної регуляторної мережі, транскрипційний фактор, активуючий зв'язок, інгібуючий зв'язок, орієнтований граф, неорієнтований граф.*

Постановка проблеми. Сучасні системи обробки інформації у більшості випадків ґрунтуються на використанні аналогій функціонування біологічних механізмів і процесів, що протікають у живих організмах. До таких процесів потрібно зарахувати функціонування природної нейронної мережі, імунні процеси, генну мережу тощо. Особливістю таких систем є децентралізована паралельна обробка інформації, великий рівень складності, здатність навчатися, розпізнавати інформацію та формувати рішення. Створення штучних моделей сучасних біологічних систем та дослідження їхньої поведінки можливо на основі системного підходу, який передбачає комплексне використання методів молекулярної біології, математики, інформатики, законів фізики і створює умови для розуміння, які чинники визначають характер функціонування біологічної системи з метою корегування цього процесу.

Реконструкція та моделювання генної регуляторної мережі (ГРМ) формує основу для дослідження та аналізу характеру взаємодій генів і впливів цих взаємодій на функціональні можливості біологічного організму. Визначення структури та характеру функціонування мережі пов'язано з великими експериментальними

та теоретичними проблемами. Параметри моделей зазвичай не є очевидними, оскільки входи та виходи елементів молекулярних систем однозначно визначити неможливо. Крім того, принципи, що лежать в основі міжмолекулярних взаємодій є дуже складними або невідомими. Проблема реконструкції генної мережі передбачає процес відновлення регулюючої взаємодії елементів системи на основі заданих експериментальних даних, якими є профілі експресій генів. До особливостей експериментальних даних потрібно зарахувати велику розмірність простору ознак і наявність складної шумової компоненти, яка виникає внаслідок протікання біологічних та технологічних процесів на етапі формування даних. У статті представлений аналіз сучасних методів реконструкції генних регуляторних мереж із виділенням їх переваг та недоліків.

Аналіз останніх досліджень та публікацій. Вирішенню проблеми реконструкції генних регуляторних мереж (ГРМ) на сьогодні присвячено велику кількість наукових праць [1–3]. Одним із ключових елементів ГРМ є транскрипційний фактор (ТФ). Оскільки клітини біологічного організму диференціюються на різні типи відповідно до функцій, які вони виконують, ТФ регулюють клітинну ідентичність, враховуючи тип відповідної клітини. До того ж ТФ контролюють зміни в експресії генів як реакцію на сигнали навколишнього середовища (організму). ГРМ ідентифікує та фіксує механізми взаємодій між ТФ та генами, створюючи умови для проведення досліджень щодо складу та функцій тканин організму у контексті здоров'я та захворювань [4]. На рис. 1 зображений механізм взаємодії елементів у ГРМ [1]. Як бачимо, ГРМ моделює регуляторний вплив ТФ на рівень експресії цільових генів. Кожен вузол в ГРМ являє собою ТФ або ген, кожен край ребра якого відповідає регуляторному зв'язку між ТФ і цільовим геном. ТФ (жовтий вузол) зв'язується з послідовностями ДНК у промоторній ділянці цільового гена (синій вузол) і має активуючий або інгібуючий регуляторний вплив на його транскрипцію (рис. 1а). Ребро спрямоване від ТФ до цільового гена (рис. 1б). Край може мати знак, що вказує, чи є дана взаємодія активуючою (+) або інгібуючою (-).

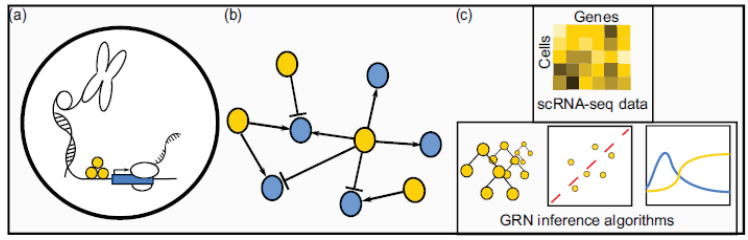

Рис. 1. Модель процесу реконструкції ГРМ на основі даних ланцюгів одноклітинних молекул РНК [1]

Здебільшого експериментальною основою для реконструкції ГРМ є масив експресій генів, отриманий шляхом ДНК-мікрочіпових експериментів або внаслідок застосування більш сучасного і точного методу, заснованого на секвенуванні молекул РНК [5]. Якщо дані, отримані шляхом застосування ДНК мікрочіпових експериментів, дозволяють оцінити середні значення експресії генів у гетерогенній популяції

типів клітин без урахування біологічних сигналів у профілях експресії генів окремих клітин, то застосування методу одноклітинного секвенування молекул РНК (scRNA-seq) [6] дозволяє оцінити експресію генів в окремих клітинах без необхідності виділення кожного типу клітин. Однак слід зазначити, що обробка даних, отриманих шляхом секвенування одноклітинних молекул РНК, і реконструкція на їх основі ГРМ потребують великих обчислювальних ресурсів, оскільки необхідно врахувати стохастичні варіації експресії генів від клітини до клітини, зміни в експресії генів, що виникають на різних стадіях клітинного циклу, високу розрідженість через недостатню чутливість у секвенуванні транскриптів у окремих клітинах для генів з низькою експресією. Однак із різким зростанням обчислювальних можливостей метод реконструкції ГРМ на основі даних експресій генів, отриманих шляхом секвенування молекул РНК, стає більш популярним і привабливим.

На сьогодні існує декілька підходів, що застосовуються для реконструкції ГРМ на основі даних експресій генів. У будь-якому випадку вхідні дані подаються як матриця, яка кількісно відображає рівень експресії кожного гена в клітинах, що досліджуються (рис. 1c). Результатом алгоритму реконструкції ГРМ є мережа прогнозованих зв'язків генів і ТФ (рис. 1b). Кожне ребро може мати вагу, що являє собою силу відповідного зв'язку. У [7] авторами наведено огляд існуючих методів реконструкції ГРМ на основі даних експресій генів. Беззаперечною перевагою методів реконструкції ГРМ на основі даних scRNA-seq є концепція псевдочасу, яка часто відрізняє сучасні алгоритми від попередніх реалізацій, заснованих на даних, отриманих шляхом застосування ДНК-мікрочіпових експериментів. Клітини у зразку можуть перебувати в різних станах і переходити від одного стану до іншого. Таким чином, можна розмістити клітини в псевдочасовому порядку на основі відмінностей у значеннях експресії відповідних генів [8]. Методи реконструкції ГРМ, розроблені спеціально для даних scRNA-seq, можуть використовувати це упорядкування як часові ряди для прогнозування регуляторних взаємодій.

Метою статті є аналіз сучасного стану методів реконструкції генних регуляторних мереж на основі даних експресій генів з виділенням їхніх переваг та недоліків.

Виклад основного матеріалу дослідження. *Метод реконструкції ГРМ на основі аналізу кореляцій* передбачає розрахунок кореляцій Пірсона між усіма парами генів та ТФ для оцінки сили зв'язку між двома змінними [9, 10]. У такий спосіб оцінюється рівень коекспресії ТФ і цільових генів у наборах даних RNA-seq і scRNA-seq. Оскільки отримана кореляційна матриця є симетричною у своїх аргументах, цей вид мережі не передбачає спрямованості при оцінці відповідних взаємодій, можна оцінити тільки силу відповідного зв'язку. При використанні повної матриці коефіцієнтів парної кореляції генна мережа є повнозв'язною завдяки наявності зв'язків між усіма вузлами мережі. У цьому випадку вага дуги дорівнює коефіцієнту парної кореляції між відповідними генами та/або ТФ. Оскільки повнозв'язна мережа не відображає реального характеру зв'язків у генній мережі, виникає необхідність формування топології мережі шляхом видалення зв'язків з низькою вагою. Даний крок реалізується шляхом ініціалізації трешолдінгового

коефіцієнту τ, що визначає порогове значення наявності зв'язку між парою відповідних генів та/або ТФ. Ваговий коефіцієнт дуги, що з'єднує гени g_a і g_b, у цьому випадку визначається наступним чином:

$$\omega(g_a, g_b) = \begin{cases} 0, & якщо r(g_a, g_b) < \tau \\ r(g_a, g_b) & якщо r(g_a, g_b) \geq \tau \end{cases} \quad (1)$$

До переваг даного типу алгоритму слід віднести простоту його реалізації. Великий відсоток суб'єктивізму є одним із головних недоліків, оскільки порогове значення трешолдінгу, що визначає топологію мережі, встановлюється емпіричним шляхом у процесі моделювання. Другим суттєвим недоліком даного типу алгоритму є той факт, що основу мережі складає неорієнтований граф, тобто даний тип мережі не враховує спрямованість взаємодій та їхній характер (активація (+) чи інгібування (-). Але слід зазначити, що даний тип алгоритму може бути застосований на попередньому етапі реконструкції ГРМ для аналізу топології мережі як доповнення до інших алгоритмів реконструкції ГРМ, які враховують спрямованість взаємодій та їхній характер.

Метод реконструкції ГРМ на основі аналізу взаємної інформації між генами та ТФ заснований на розрахунку взаємної інформації між усіма парами генів та/або ТФ із застосуванням відповідного методу оцінки ентропії Шеннона. Взаємна інформація у цьому випадку може бути виражена наступним чином [11]:

$$I(e_s, e_p) = H(e_s) + H(e_p) - H(e_s, e_p), \quad (2)$$

де $H(e_s), H(e_p)$ *і* $H(e_s, e_p)$ – ентропії профілів експресій генів e_s, e_p і взаємна ентропія даних профілів відповідно. Ентропія у цьому випадку визначається як міра невизначеності відповідного стану системи і розраховується за формулою Шеннона [12, 13]:

$$H(e) = -\sum_{i=1}^{m} p_i(e) \log_2 p_i(e), \quad (3)$$

де m – довжина профілю експресій відповідного гена; $p_i(e)$ – ймовірність реалізації відповідного дискретного значення експресії i-ї змінної. У цьому випадку взаємна ентропія Шеннона розраховується за формулою:

$$H(e_s, e_p) = -\sum_{i_s=1}^{m} \sum_{i_p=1}^{m} p_i(e_s, e_p) \log_2 p_i(e_s, e_p), \quad (4)$$

де $p_i(e_s, e_p)$ – ймовірність появи i-го значення у профілях експресій генів e_s і e_p.

Найбільш поширеним алгоритмом реконструкції ГРМ на основі взаємної інформації є алгоритм ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [14]. При застосуванні цього алгоритму зв'язки між генами та/або ТФ формуються на основі аналізу статистичних гіпотез про наявність або відсутність відповідного зв'язку. Як наслідок, на етапі передобробки даних при реконструкції ГРМ для кожного вузла мережі отримується вектор імовірностей, кожний елемент якого визначає наявність та силу зв'язку з іншими вузлами мережі (взаємна інформація):

$$P(\{g_i\}) = \frac{1}{Z}\exp[-\sum_{i=1}^{N}\phi_i(g_i) - \sum_{i,j=1}^{N}\phi_{i,j}(g_i,g_j) - \sum_{i,j,k=1}^{N}\phi_{i,j,k}(g_i,g_j,g_k)], \quad (5)$$

де N – кількість генів; Z – нормалізуючий фактор; ϕ – потенціали, що визначають силу зв'язку відповідної групи генів. Застосування цього методу передбачає, що множина генів взаємодіють між собою, якщо відповідний їм потенціал більший за заздалегідь встановлене емпіричним шляхом значення трешолдінгу τ. У протилежному випадку зв'язок між даною групою генів відсутній. При коректному виборі порогового значення трешолдінгу реконструюється ГРМ з суттєво меншою кількістю зв'язків порівняно з мережею, отриманою при застосуванні алгоритму на основі кореляційного аналізу. Оцінка ступеня взаємозв'язку між парою генів g_i, g_j при наявності M варіантів з'єднання здійснюється із застосуванням Гаусової ядерної оцінки на основі відповідного методу розрахунку ентропії Шеннона:

$$I(\{g_i\},\{g_j\}) = \frac{1}{M}\sum_{k=1}^{M}\log\frac{H_k(g_i,g_j)}{H_k(g_i)H_k(g_j)}. \quad (6)$$

Основна ідея алгоритму ARACNE полягає у тому, що при наявності різних шляхів з'єднання генів у мережі, кожний з яких характеризується ступенем відповідного взаємозв'язку $I(g_i,g_j)$, вибирається зв'язок, який задовольняє умові:

$$I(g_i,g_j) < \min[I(g_i,g_s), I(g_s,g_p),\ldots, I(g_h,g_j)], \quad (7)$$

де $g_s, g_p, \ldots, g_h$ – проміжні гени, через які зв'язуються гени g_i та g_j. Як результат, отримується мережа взаємодіючих елементів (гени та/або ТФ), вага зв'язку між якими визначається ступенем зв'язку між даними елементами.

Але слід зазначити, що, як і у випадку застосування кореляційного аналізу для формування топології ГРМ, застосування методу розрахунку взаємної інформації також потребує введення порогового значення трешолдінгового коефіцієнта, що, безумовно, вносить елемент суб'єктивізму у процес формування топології мережі. До того ж основу ГРМ у цьому випадку також складає неорієнтований граф, тобто напрямок та характер впливу вузлів мережі не враховуються.

Інший алгоритм реконструкції ГРМ на основі взаємної інформації Scribe [15] використовує поняття псевдочасу для обчислення умовної обмеженої спрямованої інформації. Ця величина вимірює взаємну інформацію між минулим і поточним рівнем експресії ТФ та цільового гена, тобто формується залежність між відповідними вузлами мережі у псевдочасовому порядку. Оскільки взаємна інформація між минулим і поточним значеннями експресії може бути несиметричною для ТФ і цільового гена, алгоритм Scribe здатний сформувати ГРМ на основі спрямованого графа.

Метод реконструкції ГРМ на основі регресійного аналізу заснований на припущенні, що рівень експресії цільового гена можна передбачити за рівнями ТФ, які його експресують. Дане припущення створює умови для розробки методу реконструкції ГРМ на основі регресійного аналізу із застосуванням моделі множинної регресії, у якій вихідний параметр представлений як експресія цільового гена, а ТФ, що визначають експресію даного гена, є відповідними регресорами. Регресійна модель для i-го цільового гена у цьому випадку може бути подана наступним рівнянням:

$$g_i = \beta_0 + \beta_1 \cdot TF_{i1} + \beta_2 \cdot TF_{i2} + \cdots + \beta_k \cdot TF_{ik} + \varepsilon_i, \quad (8)$$

де g_i – експресія i-го цільового гена; $TF = \overline{TF_1, TF_k}$ – вектор значень транскрипційних факторів, що мають відповідний вплив на даний цільовий ген; $\beta = \overline{\beta_0, \beta_k}$ – вектор значень коефіцієнтів регресійної моделі, ε_i – похибка моделі, що визначається на етапах моделювання та валідації шляхом застосування відповідних статистичних методів.

Реалізація даної концепції передбачає наступні кроки:

1. Формування підмножини цільових генів і транскрипційних факторів, які визначають експресії відповідних цільових генів, при цьому в ідеальному випадку кореляція між усіма парами ТФ повинна бути мінімальною, а між ТФ і відповідним цільовим геном – максимальною.

2. Визначення характеру залежності між відповідним ТФ і цільовим геном для формування відповідної функції, що буде застосована для даного фактора у межах регресійної моделі.

3. Створення регресійної моделі, оцінка її адекватності шляхом аналізу відповідних критеріїв якості регресійної моделі (індекс детермінації, Акаіке та Байєсівський критерії, Q-Q діаграма тощо).

4. Оцінка ступеня впливу відповідних ТФ на експресію цільового гена шляхом аналізу значень коефіцієнтів регресійної моделі. Коригування моделі з урахуванням вагових значень відповідних коефіцієнтів.

5. Валідація отриманої моделі із застосуванням методів класифікації шляхом ідентифікації значення експресії цільового гена при різних комбінаціях ТФ, що відповідають умовам проведення експериментів, на підставі яких формувалася база експериментальних даних експресій генів.

Як бачимо, реалізація концепції реконструкції ГРМ на основі регресійного аналізу являє собою процес покрокового розв'язання ланцюга задач, що належать до галузей інтелектуального аналізу даних і машинного навчання. На сьогодні ця концепція реалізована у межах програмного забезпечення GENIE3 [16], яке містить алгоритм реконструкції ГРМ, розроблений як для об'ємних вимірювань RNA-seq, так і для даних scRNA-seq. Валідація реконструйованої на основі регресійного аналізу моделі ГРМ у межах запропонованого алгоритму здійснюється шляхом застосування методу випадкового лісу (Random Forest (RF)), який заснований на ансамблі дерев рішень (регресій). Вага відповідної дуги, що з'єднує відповідний ТФ і цільовий ген, визначається як ступінь значущості даного ТФ для прогнозування експресії цільового гена та як усереднене значення ваг даного ТФ для усього дерева регресії, отриманих шляхом застосування алгоритму випадкового лісу.

Програмне забезпечення (ПЗ) GRNBoost2 [17] є логічним продовженням GENIE3. У цьому ПЗ покращено масштабованість з точки зору підвищення ефективності обробки великих одноклітинних наборів даних. Алгоритм SINCERITIES [18] у межах концепції реконструкції ГРМ реалізує регресійну модель, засновану на різниці між розподілом значень експресій профілів кожного гена в послідовних часових або псевдочасових інтервалах часу.

Основним недоліком методів реконструкції ГРМ на основі регресійних моделей є обмежена кількість ТФ, які можуть бути застосовані у рамках моделі. Збільшення кількості ТФ ускладнює як створення адекватної моделі, так і її валідацію. До переваг слід віднести високу точність прогнозування у разі побудови адекватної моделі, що створює умови для аналізу та розуміння характеру взаємодії відповідних молекулярних елементів у генній мережі.

Метод реконструкції ГРМ на основі мережі Байєса. Байєсова мережа (БМ) являє собою ймовірнісну модель, яка приймає як вхідні параметри групу випадкових величин (експресії генів та/або ТФ) і на виході генерує орієнтований ациклічний граф на основі оцінки взаємозв'язків між усіма вхідними параметрами з урахуванням умовних залежностей між ними. У випадку аналізу причинно-наслідкового взаємозв'язку $TF \rightarrow gene$ формулу умовної ймовірності можна подати наступним чином:

$$p(g_i \mid TF_{ik}) = \frac{p(g_i \cap TF_{ik})}{p(TF_{ik})}, k = \overline{1,n}, \tag{9}$$

де $p(g_i \mid TF_{ik})$ означає умовну ймовірність, тобто ймовірність експресії гена g_i за умови наявності ймовірності експресії відповідного даному гену k-го ТФ (TF_{ik}); n – кількість ТФ, що визначають експресію i-го гена.

Математичною основою для створення байєсової мережі є формула Байєса, яка дозволяє визначити ймовірність експресії k-го ТФ за умови наявності певного значення експресії у цільового гена:

$$p(TF_{ik} \mid g_i) = \frac{p(g_i \mid TF_{ik}) \cdot p(TF_{ik})}{\sum_{j=1}^{n} p(g_i \mid TF_{ik}) \cdot p(TF_{ik})}, \tag{10}$$

де TF_{ik} означає значення експресії k-го ТФ, який певним чином впливає на експресію i-го гена, ймовірність $p(g_i \mid TF_{ik})$ визначається шляхом аналізу значень експресій i-го гена за умови наявності значень експресій відповідних ТФ. Дані умовні ймовірності можна розглядати у контексті відповіді на запитання: «Яка буде ймовірність експресії i-го гена, якщо відоме значення експресії k-го ТФ?» Імовірності $p(TF_{ik})$ є апріорними, оскільки вони означають початкові ймовірності відповідних значень експресій для усіх ТФ. Ефективність байєсового методу полягає у тому, що апріорні ймовірності експресій ТФ можна уточнювати або оновлювати у процесі перебору інформації і внаслідок цього уточнювати ймовірність експресії відповідного цільового гена. Знаменник у формулі (10) є нормуючим коефіцієнтом, застосування якого обмежує інтервал варіювання значень умовної ймовірності $p(TF_{ik} \mid g_i)$ від 0 до 1.

На сьогодні є декілька алгоритмів реконструкції ГРМ на основі мережі Байєса. Так, алгоритм GRNVBEM [19] являє собою вектор значень експресій гена та/або ТФ як псевдочас і розбиває діапазон зміни значень псевдочасу на інтервали. Потім він моделює кратну зміну експресії гена або ТФ між послідовними інтервалами як лінійну комбінацію експресії у попередньому інтервалі часу для оцінки відповідних ймовірностей, на підставі яких формується мережа Байєса. До основних

недоліків даного алгоритму можна віднести той факт, що адекватність мережі у цьому випадку визначається кількістю інтервалів, які встановлюються емпіричним шляхом у процесі моделювання. Метод HBFM [20] заснований на визначенні коекспресії гена і відповідних ТФ з використанням розрідженої ієрархічної байєсової факторної моделі для зменшення впливу високої мінливості при переході від однієї клітини до другої та шуму в одноклітинних наборах даних у реконструйованій мережі.

Метод реконструкції ГРМ на основі мережі Байєса із застосуванням даних експресій генів передбачає наявність двох підходів. Перший підхід заснований на ініціалізації та подальшому уточненні (навчанні мережі) значень параметрів, що визначають розподіли граничних та умовних ймовірностей для всіх вузлів мережі. Другий підхід заснований на дослідженні і формуванні оптимальної топології мережі, яка відповідає найбільшій загальній ймовірності розподілу експресій генів та ТФ. Цей підхід заснований на застосуванні певних критеріїв, за якими оптимізується топологія мережі. При навчанні мережі (налаштуванні параметрів) граничні розподіли можуть бути задані експертами заздалегідь або отримані шляхом застосування різних методів, заснованих на ентропійному або ймовірнісному методах [21]. При використанні другого підходу розподіли умовних ймовірностей визначаються із застосуванням таких методів, як оцінка максимальної ймовірності, максимізація очікування тощо [22].

Визначення оптимальної структури мережі Байєса є досить складною проблемою. Зазвичай вирішення даної проблеми можливе шляхом перебору усіх структур із розрахунком топологічних критеріїв якості. При цьому процес визначення оптимальної топології мережі є досить складним через великі часові витрати. Більшість існуючих алгоритмів пошуку оптимальних структур є евристичними та прогресивними. Наприклад, методи ланцюга Монте-Карло Маркова [23] використовують приблизні підходи до вибірки при оцінці і формуванні топології мережі. Алгоритм жадібного пошуку еквівалентності [24] передбачає локальне оптимальне обмеження поетапної схеми пошуку. Для підрахунку балів зазвичай використовується байєсовий інформаційний критерій (BIC) або інформаційний критерій Акаіке (AIC) [25]. Порівняно з методами реконструкції ГРМ на основі оцінки коекспресії, ГРМ на основі мереж Байєса мають властивість спрямованості, тобто кожний зв'язок має певну спрямованість. У такий спосіб виявляються причинно-наслідкові зв'язки між усіма генами та/або ТФ, що значно підвищує інформативність ГРМ.

Результати моделювання показали, що у випадку застосування невеликої кількості генів при наявності шуму у даних ГРМ на основі мереж Байєса можна отримати вищу точність при високій стійкості порівняно з іншими методами реконструкції ГРМ [26]. При збільшенні кількості вузлів (генів та/або ТФ), час формування мережі зростає суперекспоненціально, що створює велику проблему при реконструкції ГРМ на основі повного генома вищих організмів [26]. До того ж для одного набору даних може бути більше однієї оптимальної структури мережі Байєса, яка може мати еквівалентні загальні ймовірності. Дані моделі (структури) називаються класами еквівалентності [24], при цьому моделі, що входять у класи

еквівалентності, неможливо ймовірнісно розрізнити. Наприклад, якщо розглянути два ланцюги мережі Байєса: $A \to B \to C$ і $C \to B \to A$, то вони є ймовірнісно еквівалентні, оскільки їхні загальні ймовірності *P(A)P(B|A)P(C|B) і P(C)P(B|C)P(A|B)* є рівними. Причинно-наслідкова мережа (ПНМ) належить до більш строгої форми мереж Байєса. Це означає, що при реконструкції даних мереж необхідно дотримуватися не тільки умовних залежностей, але й умов Маркова стосовно змінних, що забезпечує чіткий причинно-наслідковий зв'язок між змінними. Отже, мережі на основі застосування спрямованого ациклічного графа не є ймовірнісно еквівалентними через різні умови Маркова: А не залежить від С при заданому В проти С не залежить від А при заданому В. У першому випадку А є причиною С, у другому С є причиною А. Алгоритм жадібного пошуку еквівалентності, згаданий вище [24], є алгоритмом пошуку, розробленим для моделей ГРМ на основі причинно-наслідкових мереж.

До суттєвих недоліків методу реконструкції ГРМ на основі мереж Байєса слід віднести також факт, що причинно-наслідкові мережі не передбачають наявності петель зворотного зв'язку, існування яких було доведено в біологічних мережах. Але слід зазначити, що цю проблему можна вирішити шляхом застосування динамічних мереж Байєса [27, 28]. Другим недоліком є той факт, що причинно-наслідкові мережі не передбачають можливості спостерігати за типом регулювання (активація/інгібування) у ГРМ, створеним на основі мережі Байєса.

Метод реконструкції ГРМ на основі диференційних рівнянь. Наявність псевдочасової інформації в одноклітинних наборах даних також дозволяє моделювати експресію генів за допомогою звичайних диференціальних рівнянь, у яких швидкість зміни експресії цільового гена у часі є функцією експресії його ТФ [29, 30]:

$$\frac{dg_i}{dt} = w_{i1}TF_{i1} + w_{i2}TF_{i2} + \cdots + w_{ik}TF_{ik}, \; i = \overline{1,n}, \tag{11}$$

де $dg_i = g_i(t+dt) - g_i(t)$ – зміна значення експресії *i*-го гена за час *dt*; $w_{i1}, \ldots, w_{ik}$ – вагові коефіцієнти, які визначають силу впливу відповідного ТФ на експресію *i*-го гена.

Розв'язуючи дану систему рівнянь, можна визначити регуляторні відносини на основі ваги кожного ТФ у відповідній функції, що описує зміну експресії гена. Алгоритм SCODE, який реалізує концепцію реконструкції ГРМ на основі методу диференціальних рівнянь [31], містить спрощене припущення, що зміни в експресії генів можна визначити як лінійну комбінацію обмеженої кількості ТФ, яка дозволяє отримати адекватний розв'язок менш складної системи диференційних рівнянь із застосуванням методу лінійної регресії. Аналогічний за призначенням алгоритм GRISLI [32] формує ГРМ шляхом оцінки швидкості зміни експресії кожного гена, враховуючи характер перебігу динамічного процесу зміни експресії кожного гена. Запропонований метод спрощує систему рівнянь, засновану на припущенні, що реконструйована ГРМ містить мінімальну кількість регулюючих зв'язків порівняно із кількістю генів та/або ТФ, створюючи тим самим умову для розв'язання задачі реконструкції ГРМ на основі застосування розрідженої регресії. До беззаперечних переваг алгоритму GRISLI слід віднести те, що цей метод передбачає оцінку експресії цільових генів на основі аналізу декількох траєкторій

диференціювання. Алгоритм dynGENIE3 [33] є логічним продовженням алгоритму реконструкції ГРМ GENIE3. Для розв'язання системи диференційних рівнянь даний алгоритм застосовує метод випадкового лісу алгоритму GENIE3, при цьому зміна експресії одного гена визначається як потенційно нелінійна комбінація експресії інших генів або ТФ.

Слід зазначити, що застосування методу реконструкції ГРМ на основі диференційних рівнянь передбачає застосування невеликої кількості генів і ТФ. Тому у цьому випадку високу значущість отримує проблема формування інформативної підмножини профілів експресії генів стосовно цільових генів. Збільшення кількості профілів експресій генів різко ускладнює модель ГРМ, що беззаперечно відобразиться на її якості. Сучасні алгоритми реконструкції ГРМ здебільшого передбачають етап формування даних експресій генів, висококорельованих щодо цільових генів. Так, алгоритм SCENIC [34] використовує базу даних мотивів зв'язування відповідних ТФ для фільтрації передбачених регуляторних взаємодій, визначених шляхом застосування алгоритму GENIE3, що містить лише ті взаємодії, при яких мотиви для ТФ збагачуються у промоторній ділянці цільового гена. Остання реалізація pySCENIC [35] використовує процес розпаралелювання обробки інформації для підвищення ефективності алгоритму SCENIC.

Висновки. Подано аналіз сучасних методів реконструкції генних регуляторних мереж, розглянуті базові основи реконструкції ГРМ на основі даних експресій генів. Проаналізовано наступні методи реконструкції ГРМ: на основі кореляційного аналізу; на основі оцінки взаємної інформації між генами та/або транскрипційними факторами; на основі мережі Байєса; на основі регресійного аналізу та диференційних рівнянь. У межах кожного методу розглянуті алгоритми, що дозволяють практично реалізувати відповідний метод. Виділені переваги, недоліки та обмеження існуючих методів реконструкції ГРМ, визначені шляхи підвищення їхньої ефективності, що полягають у гібридизації методів обробки даних експресій генів, застосуванні ансамблів методів та більш ретельному виборі критеріїв якості для оцінки топології реконструйованої генної регуляторної мережі порівняно з біологічною генною мережею.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Akers K., Murali T. M. Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*. 2021. Vol. 26. Pp. 87–97.
2. Liu E., Li L., Cheng L. Gene regulatory network revive. *Encyclopedia of Bioinformatics and Computational Biology*. 2019. Vol. 2. Pp. 155–164.
3. Van den Broeck L., Gordon M., Inzé D., Williams C., Sozzani R. Gene Regulatory Network Inference: Connecting Plant Biology and Mathematical Modeling. *Frontiers in Genetics*. 2020. Vol. 11, art no 457.
4. Wagner A., Regev A., Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Natural Biotechnology*. 2016. Vol. 34. Pp. 1145–1160.
5. Marbach D., Costello J., Küffner R. et al. Wisdom of crowds for robust gene network inference. *Natural Methods*. 2012. Vol. 9. Pp. 796–804.

6. Tang F., Barbacioru C., Wang Y. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Natural Methods*. 2009. Vol. 6. Pp. 377–382.
7. Akers K., Murali T. M. Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*. 2021. Vol. 26. Pp. 87–97.
8. Trapnell C., Cacchiarelli D., Grimsby J. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Natural Biotechnology*. 2014. Vol. 32. Pp. 381–386.
9. Kim S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communication for the Statistical Applications and Methods*. 2015. Vol. 30. Pp. 665–674.
10. Specht A. T., Jun Li. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics*. 2017. Vol. 33 (5). Pp. 764–766.
11. Thomas M. C., Joy A. T. Elements of Information Theory. Wiley, 2nd Edition, 2006. 792 p.
12. Shannon C. E. A mathematical theory of communication. *Bell System Technical Journal*. 1948. Vol. 27. Pp. 379–423, 623–656.
13. Чумак О. В. Энтропии и фракталы в анализе данных. Москва-Ижевск : НИЦ «Регулярная и хаотическая динамика», 2011. 164 с.
14. Margolin A. A., Nemenman I., Basso K. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. 2006. Vol. 7. Pp. 1–15.
15. Qiu X., Rahimzamani A., Wang L. et al. Inferring causal gene regulatory networks from coupled single cell expression dynamics using Scribe. *Cell Systems*. 2020. Vol. 10. Pp. 265–274.
16. Huynh-Thu V. A., Irrthum A., Wehenkel L., Geurts P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*. 2010. Vol. 5 (9), art no e12776.
17. Moerman T., Aibar Santos S., Bravo González-Blas C. et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*. 2019. Vol. 35 (12). Pp. 2159–2161.
18. Papili Gao N., Minhaz Ud-Dean S. M., Gandrillon O., Gunawan R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*. 2018. Vol. 34 (2). Pp. 258–266.
19. Sanchez-Castillo M., Blanco D., Tienda-Luna I. M. et al. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*. 2018. Vol. 34 (6). Pp. 964–970.
20. Sekula M., Gaskins J., Datta S. A sparse Bayesian factor model for the construction of gene co-expression networks from single-cell RNA sequencing count data. *BMC Bioinformatics*. 2020. Vol. 21, art no 361.
21. Clarke B. Information optimality and Bayesian modelling. *Journal of Econometrics*. 2007. Vol. 138 (2). Pp. 405–429.
22. Friedman N., Linial M., Nachman I., Pe'er D. Journal of Computational Biology. 2000. Vol. 7 (3–4). Pp. 601–620.
23. Mukherjee S., Speed T. P. Network inference using informative priors. *Proceedings of the National Academy of Sciences.* 2008. Vol. 105 (38). Pp. 14313–14318.
24. Chickering D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*. 2002. Vol. 3. Pp. 507–554.

25. Liu Z., Malone B., Yuan C. Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics*. 2012. Vol. 13, art no S14.
26. Hill S. M., Heiser L. M., Cokelaer T. et al. Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nat. Methods*. 2016. Vol. 13 (4), art no 310.
27. Yu L., Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*. 2004. Vol. 5. Pp. 1205–1224.
28. Murphy K. P., Russell S. Dynamic Bayesian networks: Representation, inference and learning. Thesis of PhD dissertation, 2002.
29. Chen T., He H. L., Church G. M. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*. 1999. Vol. 4. Pp. 29–40.
30. D'Haeseleer P., Wen X., Fuhrman S., Somogyi R. Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing*. 1999. Vol. 4. Pp. 41–52.
31. Matsumoto H., Kiryu H., Furusawa C. et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*. 2017. Vol. 33. Pp. 2314–2321.
32. Aubin-Frankowski P. C., Vert J. P. Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics*. 2020. Vol. 36. Pp. 4774–4780.
33. Huynh-Thu V. A., Geurts P. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Science Report*. 2018. Vol. 8, art no 3384.
34. Aibar S., González-Blas C. B., Moerman T. et al. SCENIC: single-cell regulatory network inference and clustering. *Natural Methods*. 2017. Vol. 14. Pp. 1083–1086.
35. Van de Sande B., Flerin C., Davie K. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Natural Protocols*. 2020. Vol. 15. Pp. 2247–2276.

REFERENCES

1. Akers, K., & Murali, T. M. (2021). Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*, 26, 87–97 (in English).
2. Liu, E., Li, L., & Cheng, L. (2019). Gene regulatory network revive. *Encyclopedia of Bioinformatics and Computational Biology*, 2, 155–164 (in English).
3. Van den Broeck, L., Gordon, M., Inzé, D., Williams, C., & Sozzani, R. (2020). Gene Regulatory Network Inference: Connecting Plant Biology and Mathematical Modeling. *Frontiers in Genetics*, 11, art no 457 (in English).
4. Wagner, A., Regev, A., & Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Natural Biotechnology*, 34, 1145–1160 (in English).
5. Marbach, D., Costello, J., & Küffner, R. et al. (2012). Wisdom of crowds for robust gene network inference. *Natural Methods*, 9, 796–804 (in English).
6. Tang, F., Barbacioru, C., & Wang, Y. et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Natural Methods*, 6, 377–382 (in English).
7. Akers, K., & Murali, T. M. (2021). Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*, 26, 87–97 (in English).

8. Trapnell, C., Cacchiarelli, D., & Grimsby, J. et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Natural Biotechnology*, 32, 381–386 (in English).
9. Kim, S. (2015). ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communication for the Statistical Applications and Methods*, 30, 665–674 (in English).
10. Specht, A. T., & Jun, Li. (2017). LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics*, 33 (5), 764–766 (in English).
11. Thomas, M. C., & Joy, A. T. (2006). Elements of Information Theory. Wiley, 2nd Edition (in English).
12. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656 (in English).
13. Chumak, O. V. (2011). Jentropii i fraktaky v analize dannyh. Moskva-Izhevsk, NIC «Reguljarnaja i haoticheskaja dinamika» (in Russian).
14. Margolin, A. A., Nemenman, I., & Basso, K. et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7, 1–15 (in English).
15. Qiu, X., Rahimzamani, A., & Wang, L. et al. (2020). Inferring causal gene regulatory networks from coupled single cell expression dynamics using Scribe. *Cell Systems*, 10, 265–274 (in English).
16. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, 5 (9), art no e12776 (in English).
17. Moerman, T., Aibar Santos, S., & Bravo González-Blas, C. et al. (2019). GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35 (12), 2159–2161 (in English).
18. Papili Gao, N., Minhaz Ud-Dean, S. M., Gandrillon, O., & Gunawan, R. (2018). SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34 (2), 258–266 (in English).
19. Sanchez-Castillo, M., Blanco, D., & Tienda-Luna, I. M. et al. (2018). A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*, 34 (6), 964–970 (in English).
20. Sekula, M., Gaskins, J., & Datta, S. (2020). A sparse Bayesian factor model for the construction of gene co-expression networks from single-cell RNA sequencing count data. *BMC Bioinformatics*, 21, art no 361 (in English).
21. Clarke, B. (2007). Information optimality and Bayesian modelling. *Journal of Econometrics*, 138 (2), 405–429 (in English).
22. Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Journal of Computational Biology, 7 (3–4), 601–620 (in English).
23. Mukherjee, S., & Speed, T. P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105 (38), 14313–14318 (in English).
24. Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3, 507–554 (in English).

25. Liu, Z., Malone, B., & Yuan, C. (2012). Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics*, 13, art no S14 (in English).
26. Hill, S. M., Heiser, L. M., & Cokelaer, T. et al. (2016). Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nat. Methods*, 13 (4), art no 310 (in English).
27. Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5, 1205–1224 (in English).
28. Murphy, K. P., & Russell, S. (2002). Dynamic Bayesian networks: Representation, inference and learning. Thesis of PhD dissertation (in English).
29. Chen, T., He, H. L., & Church, G. M. (1999). Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*, 4, 29–40 (in English).
30. D'Haeseleer, P., Wen, X., Fuhrman, S., & Somogyi, R. (1999). Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing*, 4, 41–52 (in English).
31. Matsumoto, H., Kiryu, H., & Furusawa, C. et al. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, 33, 2314–2321 (in English).
32. Aubin-Frankowski, P. C., & Vert, J. P. (2020). Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics*, 36, 4774–4780 (in English).
33. Huynh-Thu, V. A., & Geurts, P. (2018). dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Science Report*, 8, art no 3384 (in English).
34. Aibar, S., González-Blas, C. B., & Moerman, T. et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Natural Methods*, 14, 1083–1086 (in English).
35. Van de Sande, B., Flerin, C., & Davie, K. et al. (2020). A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Natural Protocols*, 15, 2247–2276 (in English).

doi: 10.32403/1998-6912-2021-2-63-97-111

THE CURRENT STATE OF METHODS FOR GENE REGULATORY NETWORKS RECONSTRUCTION

I. M. Liakh

Uzhhorod National University,
3, Narodna Square, Uzhhorod, 88000, Ukraine
ihor.lyah@uzhnu.edu.ua

The paper presents the analysis of current methods of gene regulatory networks (GRN) reconstruction on the basis of gene expressions data with the allocation of their advantages, disadvantages and ways of further improvement. The gene regulatory network is represented as both an oriented and non-oriented graph, where the weight of the arc determines the strength of the corresponding connection. Within the framework

of this review, the following methods of gene regulatory networks reconstruction are considered: based on the analysis of correlations; based on the analysis of mutual information between genes and/or transcription factors; based on Bayesian networks; based on differential equations and based on regression analysis. The analysis of the relevant methods has allowed one to conclude that the current state of development of this subject area is determined by the hybridization of existing models, methods and algorithms. Currently, there is no effective information technology for gene regulatory network reconstruction, which takes into account the optimization of the network topology by estimating the distribution of topological parameters in synthetic and reconstructed networks, comparative analysis of different methods for gene regulatory networks reconstruction to form the optimal network topology. The use of ensembles of methods to reconstruct the gene regulatory networks helps to increase the adequacy of the reconstructed gene regulatory network relative to biological gene networks, which, in turn, contributes to a better understanding of the nature of the genes and/or transcription factors interaction in the network.

***Keywords**: gene regulatory network, an algorithm of gene regulatory network reconstruction, transcription factor, activating connection, inhibitory connection, oriented graph, undirected graph.*

Стаття надійшла до редакції 04.10.2021.
Received 04.10.2021.