

УДК 004.048

СУЧАСНИЙ СТАН МЕТОДІВ, МОДЕЛЕЙ ТА АЛГОРИТМІВ ВАЛІДАЦІЇ І МОДЕЛЮВАННЯ ГЕННИХ РЕГУЛЯТОРНИХ МЕРЕЖ

І. М. Лях¹, Б. В. Дурняк², С. А. Бабічев³

¹Ужгородський національний університет,
пл. Народна, 3, Ужгород, 88000, Україна

²Українська академія друкарства,
вул. Під Голоском, 19, Львів, 79020, Україна

³Херсонський державний університет,
вул. Університетська, 27, Херсон, 73000, Україна

Подано аналіз сучасних методів валідації та моделювання генних регуляторних мереж на основі даних експресій генів, виділено їхні переваги та недоліки й сформульовано шляхи подальшого удосконалення. Генна регуляторна мережа подана у вигляді орієнтованого або неорієнтованого графа, при цьому вага дуги визначає силу відповідного зв'язку. Визначено шляхи підвищення ефективності реконструйованої генної мережі внаслідок застосування порівняльного аналізу синтетичних та реконструйованих ГРМ як за топологією, так і наявністю або відсутністю зв'язків між відповідними елементами мережі, при цьому враховано також характер відповідних зв'язків. Запропоновано структурну блок-схему покрокової процедури формування даних експресій генів, виділення підмножини інформативних генів, реконструкції, валідації та моделювання ГРМ, практична реалізація якої передбачає можливість застосування різних методів, що відповідають кожному із етапів. Показано, що підвищити ефективність реалізації відповідного етапу можна шляхом гібридизації моделей, що передбачає комплексне застосування різних методів та алгоритмів для підвищення достовірності прийняття рішення на відповідному етапі.

Ключові слова: генна регуляторна мережа, валідація, моделювання, експресія генів, регулювальна взаємодія генів.

Постановка проблеми. Проблема реконструкції генних регуляторних мереж є однією із актуальних проблем сучасної біоінформатики. Враховуючи широкий спектр підходів до реконструкції ГРМ, виникає проблема вибору алгоритму, який найкраще підходить для відповідних експериментальних даних. У цьому випадку виникає необхідність у стандартизованій оцінці ефективності доступних методів реконструкції ГРМ.

Аналіз останніх досліджень та публікацій. Незалежно опубліковані результати порівняльного аналізу різних підходів показують, що є величезні можливості

для підвищення точності існуючих методів реконструкції ГРМ [1, 2]. Одним із напрямків підвищення ефективності ГРМ є оптимізація її топології на основі комплексного аналізу топологічних параметрів, що дозволяють кількісно оцінити структуру ГРМ.

Ефективність алгоритмів реконструкції ГРМ можна оцінити за багатьма параметрами, включно з точністю прогнозованих регуляторних зв'язків, узгодженістю результатів роботи алгоритму в декількох моделюваннях, а також шляхом оцінки масштабованості методу для обробки більших наборів даних. Відсутність добре налагоджених і загальноузгоджених мереж реальних масштабів, безумовно, є найбільшою проблемою для підвищення надійності при оцінці точності алгоритмів реконструкції ГРМ [3]. Дані, отримані в результаті застосування ДНК-мікрочіпових експериментів або методів секвенування молекул РНК, здебільшого дозволяють реконструювати адекватну ГРМ, але при цьому виникає проблема визначення оптимальної топології мережі. Ця проблема вирішується різними способами. Одна із стратегій полягає в тому, щоб вручну зіставити результати ДНК-мікрочіпових експериментів і дані, отримані шляхом секвенування одноклітинних молекул РНК, враховуючи типи клітин, що застосовувалися у результаті експерименту [3]. Менш поширеним підходом є використання інтегрованих баз даних, які охоплюють декілька типів клітин [3]. Хоча ці набори даних можуть не бути специфічними для будь-якого типу клітин, вони часто створюються на основі аналізу літературних джерел [4] і дозволяють оцінити надійність окремих експериментів [5]. Нарешті, мережі фізичних або функціональних взаємодій між білками також можуть бути використані для оцінки точності алгоритмів реконструкції ГРМ [3,6].

Мета статті – аналіз сучасних методів, моделей та алгоритмів валідації і моделювання реконструйованих генних регуляторних мереж із формулюванням подальших перспектив їхнього удосконалення.

Виклад основного матеріалу дослідження. Через проблеми, пов'язані з отриманням високоякісних наборів адекватних даних для валідації реконструйованих моделей ГРМ, виникає необхідність оцінки адекватності ГРМ шляхом їх порівняння із синтетичними ГРМ відомих структур, отриманих шляхом аналізу сучасних літературних джерел. До того ж синтетичні мережі можуть бути дієвим методом для оцінки ефективності алгоритмів реконструкції ГРМ у більш широкому діапазоні відомих топологій. База даних BEELINE [3] містить набір синтетичних моделей ГРМ та експериментально отриманих ГРМ для оцінки ефективності алгоритмів реконструкції ГРМ.

Алгоритм GeneNetWeaver [7] трансформує модель ГРМ у систему диференціальних рівнянь, у якій експресія генів моделюється як функція ТФ білка, яка, зі свого боку, визначається як функція експресії ТФ відповідного гена. Метод, на якому заснований даний алгоритм, додає випадковий шум до кожного рівняння для моделювання стохастичності експресії генів. Отримана система диференціальних рівнянь дозволяє зробити оцінку експресії генів, відібраних у змодельовані моменти часу. Хоча алгоритм GeneNetWeaver був розроблений для моделювання

масивів даних експресії генів, нещодавно він був використаний для створення синтетичних даних на основі одноклітинних послідовностей молекул РНК [2, 3, 8]. До недоліків даного алгоритму слід віднести той факт, що він поєднує регулятори кожного гена, використовуючи випадково згенеровану логічну функцію. Отже, він може синтезувати дані, які неточно або неадекватно відображають динаміку реконструйованої ГРМ.

Алгоритм BoolODE [3] є логічним продовженням алгоритму GeneNetWeaver. В основі цього алгоритму закладена булева логічна функція, що виражає комбінований регуляторний ефект кількох ТФ у специфікації ГРМ. Клітини у наборах даних, змодельованих алгоритмом BoolODE, можуть бути впорядковані за траєкторіями, які закінчуються кластерами, що узгоджуються зі стійкими станами ГРМ. Крім того, існує безперервна модель експресії генів уздовж кожної траєкторії, яка може бути інтерпретована даною моделлю ГРМ. Як результат, алгоритм BoolODE може бути більш прийнятним, ніж GeneNetWeaver для оцінки методів реконструкції ГРМ, які спираються на дані одноклітинних послідовностей молекул РНК [3].

Складнощі, що виникають при реконструкції якісних ГРМ та подальшому точному моделюванні варіацій експресії генів від клітини до клітини, визначаються великою кількістю генів у вихідних даних. Як наслідок, виникає необхідність редукції даних з урахуванням їхніх особливостей та мети поставленої задачі. До того ж на якість даних впливає наявність технічного шуму, який виникає на етапі реалізації технології секвенування. Складність експериментальних даних також визначається наявністю гетерогенності у популяції типів клітин. Для визначення параметрів розподілу експресії генів, з яких генеруються змодельовані дані, алгоритм Splatter використовує альтернативний підхід аналізу реальних наборів даних scRNA-seq [9]. Симулятор SERGIO [10], що використовується у процесі роботи алгоритму, генерує профілі експресії генів, що являють собою різні клітини, типи та моделі додаткових джерел технічних варіацій.

Слід зазначити, що кожний з цих методів генерує дані експресії генів. Виникнення та розвиток більшої кількості методів реконструкції ГРМ, що інтегрують мультиомічні дані, призведе до необхідності розробки методів моделювання для більш широких наборів даних з однією клітиною і, як наслідок, вирішення проблеми моделювання додаткових стохастичних ефектів, специфічних для різних технологій секвенування. Нещодавно розроблений пакет *dynngen* [11] може моделювати мультимодальні регуляторні події, включно з транскрипцією, сплайсингом, трансляцією та деградацією, і стохастично генерувати відповідну кількість мРНК та білка з одноклітинною гранулярністю.

Найбільш поширені методи валідації моделей реконструйованих ГРМ засновані на аналізі похибок першого та другого роду і передбачають оцінку площі під відповідною кривою (рис. 1 [12]).

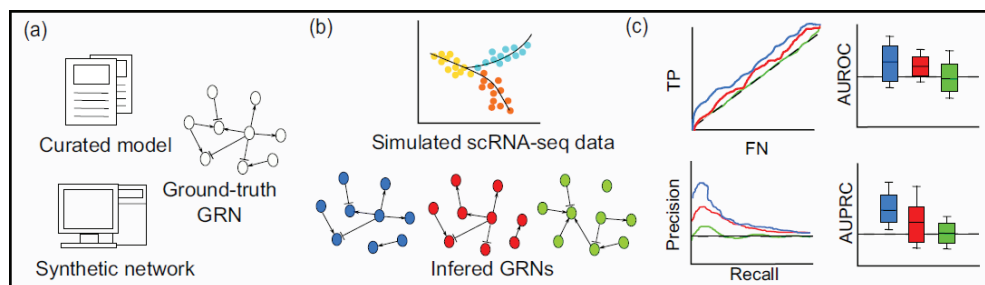


Рис. 1. Процес валідації моделей ГРМ на основі аналізу помилок першого та другого роду [12]

Зазвичай порівнюються синтетичні ГРМ та мережі, реконструйовані на основі даних реальних експериментів стосовно визначення експресій відповідних генів. Синтетичні мережі реконструюються шляхом аналізу літературних джерел із застосуванням різного типу симуляторів, описаних вище. При цьому враховується, що ГРМ мають великий ступінь розрідженості з відносно невеликою кількістю регуляторних взаємодій між усіма можливими парами генів [13]. Вочевидь, даний метод може бути застосований для порівняння ефективності різних алгоритмів реконструкції ГРМ на етапі їхньої валідації.

Враховуючи той факт, що структура реальних ГРМ являє собою орієнтований граф, валідація моделей ГРМ може бути виконана шляхом оцінки біологічної релевантності передбачуваних країв відповідних дуг, включно з ідентифікацією регуляторних відносин між генами шляхом аналізу відповідних літературних джерел [8]. Мотиви зв'язування ТФ, збагачені промоторною ділянкою передбачуваних цільових генів, використовуються для валідації моделей ГРМ [6]. Даний метод передбачає порівняння розподілу ступенів вузлів для ТФ і цільових генів у реконструйованих ГРМ з відповідними розподілами у реальних біологічних ГРМ [1].

Враховуючи високу мінливість експресії генів від клітини до клітини та низьку глибину секвенування, властиву для одноклітинних наборів даних, стабільність прогнозів алгоритму у відповідь на варіації вхідних даних також може бути корисним показником для валідації моделей ГРМ [2]. При цьому необхідно враховувати час роботи алгоритму та об'єм пам'яті обчислювальної машини, що використовується для розрахунків з урахуванням діапазону варіювання вхідних даних.

Аналіз перспективних напрямків у галузі реконструкції моделей ГРМ на основі даних експресій генів. Останнім часом з'явилася низка інноваційних технологій, що дозволяють одночасно оцінити декілька типів омічних даних в одних і тих же окремих клітинах. Наприклад, технології scNMT-seq [14], SNARE-seq [15], sci-CAR [16], scMT-seq [17] і scM&T-seq [18] дають можливість одночасного профілювання епігеноміки та транскриптоміки. Технології RAID [19], CITE-seq [20] і REAP-seq [21] застосовуються для одночасного вимірювання протеому і транскриптома. Для вивчення таких захворювань, як рак, особливо для виявлення варіацій окремих нуклеотидів між соматичними клітинами, існують методи, які одночасно секвенують ДНК і РНК в окремих клітинах, включно з G&T-seq [22], DR-Seq

[23], SDR [24] і TARGET-seq [25]. Метод scNT-seq [26] дає можливість одночасного секвенування щойно транскрибованих і вже існуючих мРНК для профілювання тимчасової динаміки експресії генів. Оскільки дедалі більше наборів даних буде створено з використанням цих останніх досягнень у секвенуванні однієї клітини, виникне потреба у розробці інструментів для обчислювального аналізу цих інтегрованих мультиомічних даних з подальшим розвитком існуючих методів реконструкції ГРМ.

Нейтралізуючий скринінг та одноклітинна мультиоміка. Збільшення або зниження експресії інших генів або ТФ у відповідь на нейтралізацію цільового гена є альтернативним підходом, який можна використовувати для прогнозування регуляторних взаємодій. У цьому контексті інша категорія одноклітинних експериментальних методів поєднує скринінг на нейтралізацію CRISPR-Cas9 з одноклітинним секвенуванням для вимірювання змін у експресії генів окремих клітин у відповідь на генетичні збурення. Дані технології вміщують CRISP-seq [27], Perturb-seq [28] та CROPseq [29]. Метод Perturb-ATAC [30] дозволяє профілювати хроматин у відповідь на CRISPR-індуковані клітинно-специфічні нейтралізації. Активуючий або інгібуючий вплив результату нейтралізації гена на експресію інших генів можна визначити, вимірюючи кратну зміну експресій кожного гена у клітинах відносно середньої експресії кожного гена в інших клітинах [27]. Алгоритм scMAGeCK [31] використовує підхід, заснований на лінійній регресії для моделювання експресії генів як функції нейтралізації інших генів і для обчислення балів для впливу кожної нейтралізації на експресію інших генів.

Останні досягнення в експериментальних техніках (такі, як ECCITE-seq [32]) дозволяють одночасно вимірювати багатоомічні дані у відповідь на нейтралізації. Вимірювання змін експресії генів в окремих клітинах, що інтегроване з асоціаціями між профілями експресії в гетерогенній популяції типів клітин, може бути використане для ідентифікації різних регуляторних впливів на цільовий ген, які відбуваються в контексті інших експресованих у клітині генів. Ця додаткова інформація може покращити прогнози щодо взаємодії багатьох факторів на експресію генів та ідентифікації прямих та непрямих регуляторних зв'язків.

Метод реконструкції ГРМ з “учителем”. У загальному випадку методи реконструкції ГРМ на основі даних scRNA-seq не контролюються в тому сенсі, що вони не використовують відомі знання про взаємовідносини пар ТФ – гени для прогнозування нових з'єднань. Останнім часом у літературі з'явилися методики для прогнозування ГРМ з вчителем [6, 33]. Наприклад, згортоква нейронна мережа для коекспресії (Convolutional Neural Network (CNNC)) обчислює нормалізовану емпіричну функцію розподілу ймовірностей для кожної пари генів на основі їх профілів експресії [6]. На наступному кроці дана нейронна мережа вводить цю функцію як зображення до звичайної згорткової нейронної мережі, щоб передбачити регуляторні відносини. Для тренування згорткової нейронної мережі використовуються пари ТФ – цільові гени, що отримані при застосуванні ChIP-seq експериментів. Сформована множина пар ТФ – цільові гени розбивається на дві підмножини, одна з яких застосовується для навчання нейронної мережі, а друга – для її тестування.

Інший підхід тих самих авторів застосовує графові згорткові нейронні мережі на просторових транскриптомних даних для прогнозування генів, що взаємодіють між клітинами [34]. Хоча цей підхід не дозволяє зробити прямий висновок щодо оптимальної структури ГРМ, але він дає змогу оцінити можливості та ефективність контрольованих методів для вирішення проблеми реконструкції ГРМ.

Аналіз сучасного стану досліджень щодо валідації та моделювання генних регуляторних мереж дозволяє зробити висновок, що на сьогодні не існує ефективної інформаційної технології обробки даних експресій генів, реконструкції, валідації та моделювання генної регуляторної мережі, яка здатна з високим ступенем довіри діагностувати складні захворювання та передбачити характер подальшого розвитку біологічного організму на генному рівні шляхом моделювання характеру зміни експресії цільових генів при їхній взаємодії з ТФ та іншими генами мережі.

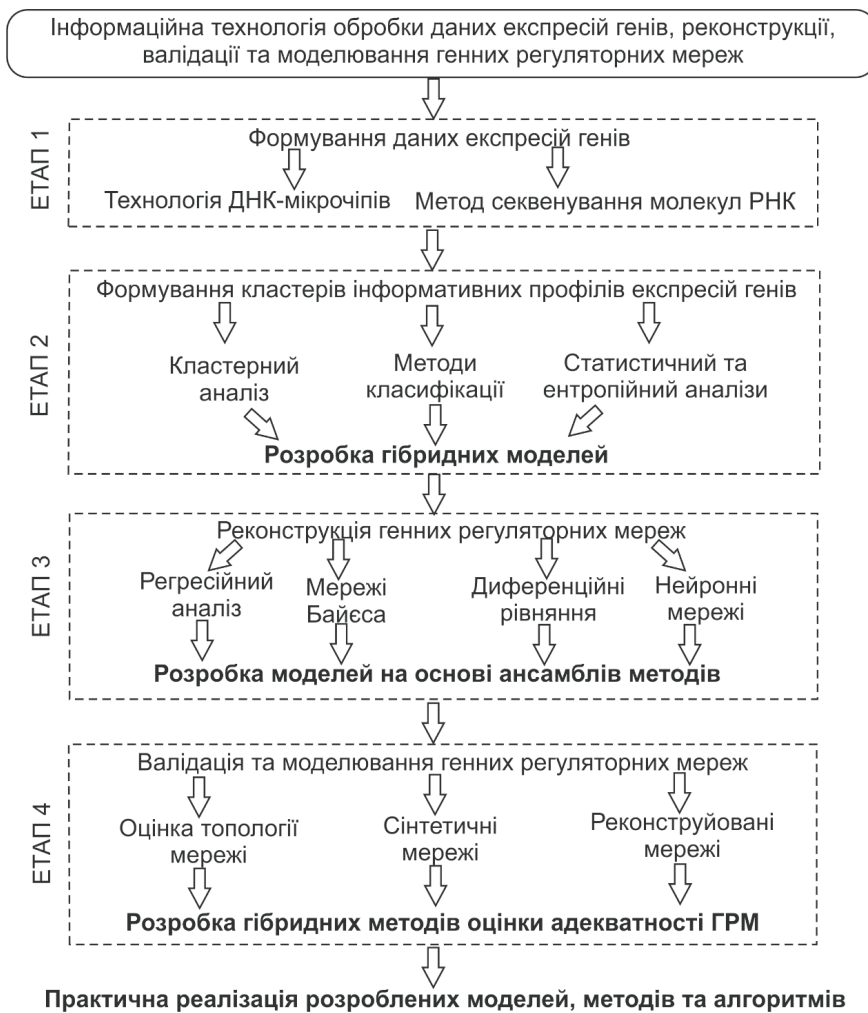


Рис. 2. Структурна блок-схема інформаційної технології обробки даних експресій генів, реконструкції, валідації та моделювання ГРМ

На рис. 2 зображено структурну блок-схему покрокової процедури формування даних експресій генів, виділення підмножини інформативних генів, реконструкції, валідації та моделювання ГРМ. Як бачимо, практична реалізація даної процедури передбачає можливість застосування різних методів, що відповідають кожному з етапів, при цьому слід зазначити, що на сьогодні проблема вибору оптимальної комбінації методів не вирішена. Вочевидь, підвищити ефективність реалізації відповідного етапу можна шляхом гібридизації моделей, що передбачає комплексне застосування різних методів та алгоритмів для підвищення достовірності прийняття рішення на відповідному етапі. Цей крок також вимагає розробки гібридних критеріїв якості для оцінки результату реалізації відповідного етапу.

Іншим шляхом підвищення ефективності інформаційної технології реконструкції, валідації та моделювання ГРМ є застосування ансамблів методів з подальшим порівнянням результату, отриманим кожним із методів, із застосуванням відповідних критеріїв якості реалізації відповідного етапу та подальшим розрахунком комплексного критерію, на підставі якого приймається остаточне рішення щодо структури моделі. Таким чином, можна зробити висновок, що зараз існує ряд невирішених питань, які безпосередньо впливають на ефективність процесу реконструкції, валідації та моделювання ГРМ.

Висновки. У статті представлені результати дослідження щодо аналізу сучасних методів валідації та моделювання генних регуляторних мереж із виділенням переваг та недоліків відповідних методів. Проведене дослідження дало змогу визначити шляхи підвищення ефективності процесу реконструкції генної мережі завдяки ретельнішій валідації в процесі моделювання із застосуванням як синтетичних, так і реконструйованих моделей ГРМ. При цьому враховується топологія реконструйованої мережі і характер зв'язків між молекулярними елементами ГРМ. Проведений аналіз дозволив також виділити невирішену частину загальної проблеми, яка полягає у відсутності ефективної технології валідації ГРМ, що враховує порівняльний аналіз топологій синтетичних та реконструйованих ГРМ за розподілом відповідних топологічних параметрів, значення яких визначають структуру мереж, з одного боку, та відсутності інформаційної технології моделювання ГРМ, яка дозволяє з високим ступенем достовірності передбачити характер взаємодії цільових генів та транскрипційних факторів з метою кращого розуміння закономірностей взаємодії елементів мережі з урахуванням типу хвороби, що досліджується, з іншого боку. Виділені невирішені проблемні питання дозволили окреслити шляхи удосконалення процедури реконструкції, валідації та моделювання генних регуляторних мереж.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Chen S., Mar J. C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*. 2018. Vol. 19. Pp. 232.
2. Pratapa A., Jalihal A. P. et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Natural Methods*. 2020. Vol. 17. Pp. 147–154.

3. Pratapa A., Jaliha A. P. et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Natural Methods*. 2020. Vol. 17. Pp. 147–154.
4. Han H., Cho J. W., Lee S. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*. 2018. Vol. 46. Pp. D380–D386.
5. Garcia-Alonso L., Holland C. H., Ibrahim M. M. et al. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*. 2019. Vol. 29. Pp. 1363–1375.
6. Yuan Y., Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences of the USA*. 2019. Vol. 116. Pp. 27151–27158.
7. Schaffter T., Marbach D., Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*. 2011. Vol. 27. Pp. 2263–2270.
8. Nguyen H., Tran D., Tran B., Pehlivan B., Nguyen T. A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Brief Bioinformatics*. 2020. Vol. 22 (3), art no. bbaa190.
9. Zappia L., Phipson B., Oshlack A. Splatter: simulation of singlecell RNA sequencing data. *Genome Biology*. 2017. Vol. 18, art no. 174.
10. Dibaeinia P., Sinha S. SERGIO: a single-cell expression simulator guided by gene regulatory networks. *Cell Systems*. 2020. Vol. 11. Pp. 252–271.
11. Cannoodt R., Saelens W. et al. Spearheading future omics analyses using *dyngen*, a multi-modal simulator of single cells. *Nature Communication*. 2021. Vol. 12, art no. 3942.
12. Akers K., Murali T. M. Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*. 2021. Vol. 26. Pp. 87–97.
13. Saito T., Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*. 2015. Vol. 10, art no. e0118432.
14. Clark S. J., Argelaguet R., Kapourani C. A. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*. 2018. Vol. 9, art no. 781.
15. Chen S., Lake B. B., Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotech*. 2019. Vol. 37. Pp. 1452–1457.
16. Cao J., Cusanovich D. A., Ramani V. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. 2018. Vol. 361. Pp. 1380–1385.
17. Hu Y., Huang K., An Q. et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biology*. 2016. Vol. 17, art no. 88.
18. Angermueller C., Clark S. J., Lee H. J. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*. 2016. Vol. 13. Pp. 229–232.
19. Gerlach J. P., van Buggenum J. A. G., Tanis S. E. J. et al. Combined quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells. *Science Reports*. 2019. Vol. 9, art no. 1469.
20. Stoeckius M., Hafemeister C., Stephenson W. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*. 2017. Vol. 14. Pp. 865–868.

21. Peterson V. M., Zhang K. X., Kumar N. et al. Multiplexed quantification of proteins and transcripts insingle cells. *Nature Biotechnology*. 2017. Vol. 35. Pp. 936–939.
22. Macaulay I. C., Haerty W., Kumar P. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*. 2015. Vol. 12. Pp. 519–522.
23. Dey S. S., Kester L., Spanjaard B. et al. Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*. 2015. Vol. 33. Pp. 285–289.
24. Han K. Y., Kim K. T. et al. SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Research*. 2018. Vol. 28. Pp. 75–87.
25. Rodriguez-Meira A., Buck G., Clark S. A. et al. Unravelling intratumoral heterogeneity through highsensitivity single-cell mutational analysis and parallel RNA sequencing. *Molecular Cell*. 2019. Vol. 73. Pp. 1292–1305.
26. Qiu Q., Hu P., Qiu X. et al. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods*. 2020. Vol. 17. Pp. 991–1001.
27. Jaitin D. A., Weiner A., Yofe I. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*. 2016. Vol. 167. Pp. 1883–1896.
28. Dixit A., Parnas O., Li B. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*. 2016. Vol. 167. Pp. 1853–1866.
29. Datlinger P., Rendeiro A. F., Schmidl C. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*. 2017. Vol. 14. Pp. 297–301.
30. Rubin A. J., Parker K. R. et al. Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*. 2019. Vol. 176. Pp. 361–376.
31. Yang L., Zhu Y., Yu H. et al. scMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens. *Genome Biology*. 2020. Vol. 21, art no. 19.
32. Mimitou E. P., Cheng A., Montalbano A. et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*. 2019. Vol. 16. Pp. 409–412.
33. Turki T., Taguchi Y. H. SCGRNs: novel supervised inference of single-cell gene regulatory networks of complex diseases. *Computers in Biology and Medicine*. 2020. Vol. 118, art no. 103656.
34. Yuan Y., Bar-Joseph Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology*. 2020. Vol. 21, art no. 300.

REFERENCES

1. Chen, S., & Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19, 232 (in English).
2. Pratapa, A., & Jaliha, A. P. et al. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Natural Methods*, 17, 147–154 (in English).
3. Pratapa, A., & Jaliha, A. P. et al. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Natural Methods*, 17, 147–154 (in English).
4. Han, H., Cho, J. W., & Lee, S. et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46, D380–D386 (in English).

5. Garcia-Alonso, L., Holland, C. H., & Ibrahim, M. M. et al. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29, 1363–1375 (in English).
6. Yuan, Y., & Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences of the USA*, 116, 27151–27158 (in English).
7. Schaffter, T., Marbach, D., & Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27, 2263–2270 (in English).
8. Nguyen, H., Tran, D., Tran, B., Pehlivan, B., & Nguyen, T. (2020). A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Brief Bioinformatics*, 22 (3), art no. bbaa190 (in English).
9. Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: simulation of singlecell RNA sequencing data. *Genome Biology*, 18, art no. 174 (in English).
10. Dibaeinia, P., & Sinha, S. (2020). SERGIO: a single-cell expression simulator guided by gene regulatory networks. *Cell Systems*, 11, 252–271 (in English).
11. Cannoodt, R., & Saelens, W. et al. (2021). Spearheading future omics analyses using *dyngen*, a multi-modal simulator of single cells. *Nature Communication*, 12, art no. 3942 (in English).
12. Akers, K., & Murali, T. M. (2021). Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*, 26, 87–97 (in English).
13. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10, art no. e0118432 (in English).
14. Clark, S. J., Argelaguet, R., & Kapourani, C. A. et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9, art no. 781 (in English).
15. Chen, S., Lake, B. B., & Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotech*, 37, 1452–1457 (in English).
16. Cao, J., Cusanovich, D. A., & Ramani, V. et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361, 1380–1385 (in English).
17. Hu, Y., Huang, K., & An, Q. et al. (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biology*, 17, art no. 88 (in English).
18. Angermueller, C., Clark, S. J., & Lee, H. J. et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13, 229–232 (in English).
19. Gerlach, J. P., van Buggenum, J. A. G., & Tanis, S. E. J. et al. (2019). Combined quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells. *Science Reports*, 9, art no. 1469 (in English).
20. Stoeckius, M., Hafemeister, C., & Stephenson, W. et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14, 865–868 (in English).
21. Peterson, V. M., Zhang, K. X., & Kumar, N. et al. (2017). Multiplexed quantification of proteins and transcripts insingle cells. *Nature Biotechnology*, 35, 936–939 (in English).
22. Macaulay, I. C., Haerty, W., & Kumar, P. et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12, 519–522 (in English).

23. Dey, S. S., Kester, L., & Spanjaard, B. et al. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, 33, 285–289 (in English).
24. Han, K. Y., & Kim, K. T. et al. (2018). SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Research*, 28, 75–87 (in English).
25. Rodriguez-Meira, A., Buck, G., & Clark, S. A. et al. (2019). Unravelling intratumoral heterogeneity through highsensitivity single-cell mutational analysis and parallel RNA sequencing. *Molecular Cell*, 73, 1292–1305 (in English).
26. Qiu, Q., Hu, P., & Qiu, X. et al. (2020). Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods*, 17, 991–1001 (in English).
27. Jaitin, D. A., Weiner, A., & Yofe, I. et al. (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, 167, 1883–1896 (in English).
28. Dixit, A., Parnas, O., & Li, B. et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167, 1853–1866 (in English).
29. Datlinger, P., Rendeiro, A. F., & Schmidl, C. et al. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14, 297–301 (in English).
30. Rubin, A. J., & Parker, K. R. et al. (2019). Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*, 176, 361–376 (in English).
31. Yang, L., Zhu, Y., & Yu, H. et al. (2020). scMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens. *Genome Biology*, 21, art no. 19 (in English).
32. Mimitou, E. P., Cheng, A., & Montalbano, A. et al. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*, 16, 409–412 (in English).
33. Turki, T., & Taguchi, Y. H. (2020). SCGRNs: novel supervised inference of single-cell gene regulatory networks of complex diseases. *Computers in Biology and Medicine*, 118, art no. 103656 (in English).
34. Yuan, Y., & Bar-Joseph, Z. (2020). GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology*, 21, art no. 300 (in English).

doi: 10.32403/0554-4866-2021-2-82-92-103

THE CURRENT STATE OF METHODS, MODELS AND ALGORITHMS FOR GENE REGULATORY NETWORKS VALIDATION AND SIMULATION

I. Liakh¹, B. Durnyak², S. Babichev³

¹*Uzhhorod National University,
3, Narodna Square, Uzhhorod, 88000, Ukraine
ihor.lyah@uzhnu.edu.ua*

²*Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine
durnyak@uad.lviv.ua*

³Kherson State University,
27, University St., Kherson, 73000, Ukraine
sbabichev@ksu.ksu.ua

The paper presents an analysis of current methods of gene regulatory networks validation and simulation on the basis of genes expression data with the allocation of their advantages, shortcoming and ways for the further improvement of the existing techniques. The gene regulatory network is represented as a directed or non-directed graph, in which the weight of the arc determines the strength of the appropriate connection. The way to increase the efficiency of the reconstructed gene network is determined by analyzing synthetic and reconstructed timing systems both in topology and in the presence or absence of connection between the relevant elements of the network, and the nature of the corresponding connections. The conducted analysis has allowed one to allocate the unsolved part of the general problem, which is the absence of effective technology of gene regulatory networks validation, which takes into account the comparative analysis of the topologies of synthetic and reconstructed gene regulatory networks by the distribution of the relevant topological parameters, the values of which determine the structure of the networks on the one hand and, the absence of the information technology of gene regulatory network simulation that allows one to predict with a high degree of reliability the nature of the target genes and transcription factors interacting in order to better understand the patterns of the network elements interconnection, taking into account the type of disease, on the other hand. The structural block chart of the step-by-step procedure of gene expression data generation, selection of a subset of informative genes, reconstruction, validation and simulation of gene regulatory networks is proposed. The practical implementation of the proposed procedure involves the use of different methods taking into account the appropriate stage. It is shown that it is possible to increase the efficiency of the respective stage implementation by hybridization of models with joint application of various methods and algorithms for an increase of reliability of decision-making at the corresponding stage.

Keywords: gene regulatory network, validation, simulation, gene expression, regulatory interaction of genes.

*Стаття надійшла до редакції 01.10.2021.
Received 01.10.2021.*