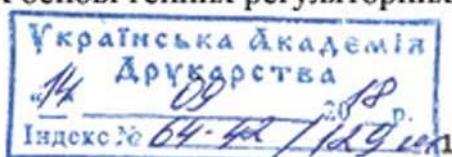


ВІДГУК
офіційного опонента

доктора технічних наук, професора **Михальова Олександра Ілліча**
на дисертаційну роботу **Бабічева Сергія Анатолійовича** «Теоретичні та
практичні засади інформаційної технології обробки профілів експресій генів
для реконструкції генних мереж», яка подана на здобуття наукового ступеня
доктора технічних наук за спеціальністю
05.13.06 - інформаційні технології

Дисертація присвячена розв'язку задачі реконструкції генних регуляторних мереж на основі дослідження профілів експресій генів. Ця проблема нерозривно пов'язана із необхідністю створення нових та удосконалення існуючих методів обробки складних даних, особливістю яких є велика розмірність простору ознак та наявність складної шумової компоненти. Як раз до таких даних і відносяться профілі експресій генів, що отримані із застосуванням технологій ДНК мікрочіпів або методом секвенування молекул РНК. Підвищення якості обробки профілів експресій генів дасть можливість підвищити адекватність реконструйованих генних регуляторних мереж, що, у свою чергу, створить умови для розробки нових систем ранньої діагностики складних системних явищ, а саме, тяжких захворювань, виготовлення нових ефективних ліків і, в цілому - створення сучасних методів лікування складних хвороб. Дані факти свідчать про несумлінну *актуальність* теми дослідження.

Сучасні інформаційні технології реконструкції генних регуляторних мереж на основі масивів профілів експресій генів, як звісно, ґрунтуються на використанні бікластерного аналізу, практична реалізація якого дозволяє отримати групи взаємно корельованих генів та умов проведення експериментів, які використовуються для реконструкції генної мережі. Проте кількість та розмір бікластерів варіюється у дуже великому діапазоні, що ускладнює процес реконструкції якісних генних мереж. Визначення структури та характеру функціонування такої мережі пов'язано з великими експериментальними та теоретичними проблемами. При цьому більшість параметрів 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 зовсім невідомими. Тому існуючі моделі та методи не є достатньо ефективними для обробки профілів експресій генів для реконструкції генних регуляторних мереж. Насамперед це пов'язано із відсутністю відповідного математичного забезпечення, а також комплексного підходу до вирішення завдання обробки високорозмірних складних даних із метою обґрунтованого виділення взаємокорельованих рядків і стовпців для подальшої реконструкції на їх основі генних регуляторних мереж.



Основним напрямом ефективного вирішення цієї проблеми є розробка нових та удосконалення існуючих інформаційних технологій передобробки профілів експресій генів із застосуванням кількісних критеріїв оцінки якості обробки інформації на кожному етапі. В умовах глобалізації методів та засобів інтелектуального аналізу даних все більш впевнені позиції отримує напрям інтелектуалізації інформаційних технологій, як альтернативного підходу для забезпечення вирішення завдання обробки даних складної природи та реконструкції на їх основі генних регуляторних мереж. Таким чином, розвиток інтелектуальних технологій обробки профілів експресій генів для реконструкції генних регуляторних мереж сприяє укріпленню фундаменту методів обробки складних даних та є потужним інструментом вирішення широкого спектру задач на основі математичного апарату обчислювального інтелекту.

Автором виконано змістовний аналіз достатньо великого обсягу наукових публікацій за напрямом досліджень у різних предметних галузях, який показав відсутність сучасних інформаційних технологій обробки профілів експресій генів для реконструкції генних регуляторних мереж та необхідність розробки моделей, методів та алгоритмів для обробки профілів експресій генів для реконструкції генних регуляторних мереж на основі даних ДНК-мікрочіпових експериментів або методів секвентного аналізу молекул РНК.

Актуальність виконаної роботи

Дисертація С.А. Бабічева присвячена виріщенню науково-прикладної проблеми розробки теоретичних та практичних зasad інформаційної технології обробки профілів експресій генів для реконструкції генних регуляторних мереж, створення ефективних методів інтелектуального аналізу й опрацювання складних даних. Проте розв'язок цієї проблеми засновано на глибинному дослідженні теоретичних основ обробки високорозмірних даних складної природи, що, у свою чергу, обумовлено непридатністю існуючих моделей та методів для аналізу вищезазначених даних з метою обґрунтованого виділення векторів генів та зразків з високим ступенем їх взаємної кореляції, що використовуються як експериментальна основа для реконструкції генних регуляторних мереж.

На основі проведених досліджень здобувачем було виявлено недостатню ефективність існуючих підходів до обробки складних високорозмірних даних та підтверджено необхідність розробки інформаційного забезпечення у вигляді:

- інформаційної технології покрокового процесу обробки профілів експресій генів для реконструкції та валідації моделей генних регуляторних мереж;
- гібридної моделі фільтрації та редукції профілів експресій генів;
- індуктивної технології об'єктивної кластеризації на основі індуктивних методів аналізу складних систем;

- гібридної моделі кластер-бікластерного аналізу профілів експресій генів та умов проведення експерименту по визначеню експресії відповідного гену;
- технології реконструкції та валідації моделей генних регуляторних мереж;
- технічного забезпечення у вигляді програм і алгоритмів, реалізованих у програмному середовищі R.

Дисертаційна робота виконувалась відповідно до науково-дослідних держбюджетних тем МОН України: «Розробка гібридних нейро-фазі-імунних алгоритмів інформаційних систем та технологій для розв'язання задач в біоінформатиці та обчислювальній біології» (ДР № 0116U002841), «Синтез гібридних еволюційних алгоритмів та методів для моделювання генних регуляторних мереж» (ДР № 0116U002840), «Розробка теоретичних зasad побудови інтелектуальних систем класифікації на основі онтології» (ДР № 0113U007833), у яких здобувач був виконавцем окремих етапів.

Особливої уваги заслуговує той факт, що тематика розглянутих у роботі завдань лежить у площині загальнодержавних науково-технічних програм, що сформульовані в Законах України “Про наукову і науково-технічну діяльність”, “Про національну програму інформатизації”, а також відповідають планам найважливіших науково-технічних програм Міністерства освіти та науки України, зокрема: 6 – Інформатика, автоматизація та приладобудування; 6.2.1 – Інтелектуалізація процесів прийняття рішень; 6.2.2 – Перспективні інформаційні технології і системи.

Предметом дослідження є інформаційні технології обробки даних складної біологічної природи в задачах реконструкції генних мереж, методи й алгоритми їх виконання. При цьому автором зроблено аналіз літературних джерел щодо базових основ реконструкції генних регуляторних мереж та існуючих методів створення експериментальних даних для реконструкції генних мереж (59 джерел), методів фільтрації, редукції, кластеризації та бікластеризації складних даних (82 джерела), методів реконструкції та моделювання генних регуляторних мереж (34 джерела). На основі зробленого аналізу сформульовано мету дослідження та задачі, розв'язання яких сприяє досягненню поставленої мети.

Аналіз рівня попередньо досягнутих результатів в області, що розглядається у роботі, дозволив автору виявити недоліки та протиріччя існуючих моделей та методів, які пов’язані з їх непристосованістю для аналізу високорозмірних складних даних. На основі системного підходу було сформульовано науково-прикладну проблему та завдання, які необхідно розв’язати для її вирішення, що полягають у розробці нової інформаційної технології обробки профілів експресій генів для реконструкції генних регуляторних мереж, відмінною рисою якої є висока об’єктивність отриманих результатів за рахунок комплексного використання методів індуктивного моделювання складних систем, вейвлет аналізу, нечіткого логічного виводу,

теорії графів і кількісних критеріїв оцінки якості інформації на кожному етапі її обробки.

Таким чином, актуальність дисертаційної роботи полягає у необхідності створення нової інформаційної технології обробки профілів експресій генів для реконструкції генних регуляторних мереж, практична реалізація якої сприяє кращому розумінню характеру взаємодії генів у мережі, що створює умови для розробки нових сучасних методів діагностування технічних та біомедицинських складних систем, лікування тяжких хвороб.

Наукова новизна дисертаційних досліджень

В дисертаційній роботі отримано теоретичні та практичні результати щодо розробки наукових і методологічних основ створення та застосування інформаційних технологій та інформаційних систем на основі удосконалення та створення нових методів аналізу і оцінювання інформації, моделювання процесів і їх класифікації для побудови ефективних прикладних інформаційних технологій. Це дозволяє стверджувати, що тема і зміст дисертації повністю відповідають паспорту спеціальності 05.13.06 – інформаційні технології.

В дисертаційній роботі поставлено і вирішено такі основні завдання наукового дослідження:

1. Розроблено технологію визначення оптимальної комбінації методів оцінки експресій генів об'єктів, що отримані при різних умовах проведення експерименту з використанням кількісних ентропійних критеріїв оцінки інформативності даних, які досліджуються. Автором проаналізовано різні методи оцінки ентропії Шеннона для аналізу одновимірних модельних даних та профіля експресій генів при різних рівнях шумової компоненти. На основі результатів моделювання запропоновано покроковий алгоритм визначення оптимальної за критерієм ентропія Шеннона комбінації методів обробки даних ДНК-мікрочіпових експериментів для формування масиву експресій генів (п. 2.2.1 дисертації).

2. Створено технологію аналізу, фільтрації та редукції профілів експресій генів із використанням вейвлет аналізу, методів нечіткої логіки та кількісних статистичних й ентропійних критеріїв оцінки інформативності даних, що досліджуються. При цьому визначення оптимальних параметрів вейвлет-фільтру проводилося шляхом паралельної оцінки інформативності фільтрованих даних та виділеної шумової компоненти. На основі проведених досліджень дисертантом запропоновано структуру інформаційної технології визначення оптимальних параметрів вейвлет-фільтру (п. 2.2.2 дисертації). Проте редукція неінформативних профілів експресій генів за статистичними критеріями та ентропією Шеннона проводилася із застосуванням системи нечіткого логічного виводу. Автором запропоновано модель системи нечіткого логічного виводу, практична реалізація якої дозволила визначити порогові

значення відповідних критеріїв, що розділяють гени на інформативні та неінформативні. При цьому ген видається з подальшого дослідження, якщо він визнаний неінформативним за усіма критеріями, що використовуються (п. 2.2.3 дисертації).

Використання такого підходу дійсно дає змогу підвищити інформативність експериментальних даних за рахунок зменшення шумової компоненти та видалення генів, що визнані неінформативними за групою критеріїв.

3. Розроблено інформаційну технологію об'єктивної кластеризації профілів експресій генів на основі комплексного використання індуктивних методів аналізу складних систем, внутрішніх і зовнішніх критеріїв оцінки якості групування даних і комплексного критерію балансу, який враховує як характер розподілу даних в окремих кластеризаціях, так і різницю у результатах кластеризацій, що отримано із використанням рівнопотужних підмножин даних. Дисертантом у процесі моделювання визначено оптимальну функцію афінності для оцінки відстані між відповідними профілями, а також внутрішні і зовнішні критерії оцінки якості кластеризації та комплексний критерій балансу, який містить як компоненти відповідні внутрішні та зовнішній критерії. Як результат, запропоновано архітектуру індуктивної технології об'єктивної кластеризації та покрокову процедуру її реалізації (п. 3.5.2 дисертації), практична реалізація якої можлива на основі будь-якого алгоритму кластерного аналізу. При цьому слід відзначити, що реалізація запропонованої автором технології дозволяє зменшити похибку відтворюваності, яка властива більшості існуючим алгоритмам кластерного аналізу.

4. Розроблено та практично реалізовано гібридну модель кластер-бікластерного аналізу на основі ієрархічних та щільнісних алгоритмів кластеризації в класі індуктивної технології об'єктивної кластеризації й існуючих алгоритмів бікластеризації для отримання кластерів профілів генів з метою подальшої реконструкції генної регуляторної мережі. Дисертантом розроблено гібридні моделі індуктивної технології об'єктивної кластеризації на основі щільнісного алгоритму DBSCAN та самоорганізуючого алгоритму SOTA (п.п. 4.1 і 4.2 дисертації). Практична реалізація запропонованих моделей дозволяє визначити оптимальні параметри відповідного алгоритму кластеризації за критеріями, що використовуються (п. 4.3 дисертації). У рамках даного завдання автором також розроблено технологію бікластерного аналізу профілів експресій генів на основі методу бікластеризації «ensemble». Практична реалізація даної технології дозволяє визначити оптимальні параметри даного методу за внутрішнім критерієм якості бікластеризації, ефективність якого доведено в процесі моделювання бікластеризації на модельних даних із різним рівнем зашумленості (п. 4.4 дисертації). Як

результат досліджень, дисертантом запропоновано гібридну модель кластер-бікластерного аналізу у вигляді структурної блок-схеми покрокової процедури групування даних із застосуванням алгоритмів кластеризації і рамках індуктивної технології об'єктивної кластеризації та методу бікластеризації «ensemble». Результатом практичної реалізації даної технології є отримання бікластерів для подальшої реконструкції генних регуляторних мереж (п. 4.5 дисертації).

5. Запропоновано технологію реконструкції та валідації моделей генних регуляторних мереж на основі статистичних методів аналізу інформації. На підставі проведеного автором дослідження статистичних методів реконструкції генних мереж із розрахунком топологічних параметрів відповідної мережі створено технологію оптимізації топології генної мережі (п. 5.2 і 5.3 дисертації). Для валідації отриманих моделей генних мереж дисертантом запропоновано технологію валідації, реалізація якої передбачає порівнювальний аналіз характеру зв'язків між генами у базовій мережі та у мережах, що реконструйовані на основі отриманих бікластерів із розрахунком похибок першого та другого роду (п. 5.5 дисертації). Автором проведено порівнювальний аналіз методів реконструкції і валідації моделей генних мереж на основі алгоритмів кореляційного виводу і ARACNE, який показав більшу ефективність алгоритму кореляційного виводу за запропонованими критеріями (п. 5.5.1 і 5.5.2).

6. Створена нова інформаційна технологія обробки профілів експресій генів для реконструкції генних регуляторних мереж на основі комплексного використання методів індуктивного моделювання складних систем, методів нечіткого моделювання та кількісних критеріїв оцінки якості інформації на кожному етапі її обробки. Дано технологія представлена у вигляді структурно-логічної схеми покрокової процедури реалізації запропонованих методів і моделей обробки профілів експресій генів для реконструкції і валідації моделей генних регуляторних мереж (п. 5.6 дисертації).

7. Перевірено адекватність розроблених моделей шляхом практичної реалізації запропонованої технології обробки профілів експресій генів для реконструкції генних мереж із застосуванням даних ДНК-мікрочіпових експериментів пацієнтів, що досліджувалися на хвороби Альцгеймера, Паркінсона і різні типи раку. Реалізація проводилася у програмному середовищі R із застосуванням технології Cytoscape для реконструкції генних мереж (розділ 6 дисертації). Отримані результати показали високий рівень адекватності отриманих моделей генних мереж за критеріями, що використовувалися.

Достовірність наукових положень і результатів

Достовірність та обґрутування отриманих у роботі результатів забезпечується коректним використанням методів математичного та комп’ютерного моделювання, адекватність яких підтверджується експериментами із використанням фізичних моделей.

Під час теоретичних досліджень використано теорію системного аналізу, теорію індуктивного моделювання складних систем, теорію оптимізації, теорію нечіткого моделювання, теорію графів, нечіткого логічного виводу, тощо.

Ефективність застосування розроблених інформаційних технологій підтверджується результатами їх впровадження у Херсонському обласному онкологічному диспансері при розробці системи ранньої діагностики онкологічних захворювань, у товаристві з додатковою відповіальністю «Херсонський маслозавод» для діагностики якості молочної продукції на основі лабораторних досліджень, та у ряді навчальних закладів, що підтверджується відповідними актами впровадження.

Ступінь обґрутованості наукових положень, висновків і рекомендацій, що сформульовано в дисертації

Грунтовний аналіз основних методів розв’язку поставлених задач на різних етапах обробки профілів експресій генів дозволив автору розробити низьку методів і моделей обробки профілів експресій генів для реконструкції і валідації моделей генних мереж. Коректне застосування методів системного аналізу, методів ймовірнісно-статистичного моделювання, вейвлет-аналізу і методів нечіткого моделювання дозволило розробити технологію передобробки даних ДНК-мікрочіпових експериментів для формування масиву експресій генів. Розробка індуктивної технології об’єктивної кластеризації заснована на коректному використанні методів математичної статистики, методів індуктивного моделювання складних систем, теорії оптимізації, методів кластерного та бікластерного аналізів. Коректне застосування теорії графів, теорії оптимізації та методів багатокритеріального аналізу дозволило розробити технологію реконструкції та валідації моделей генних мереж. Отримані результати та висновки є логічно та математично аргументованими.

Дисерант вирішив поставлену проблему, що пов’язана з розробкою теоретичних та практичних зasad інформаційної технології обробки профілів експресій генів, отриманих шляхом ДНК-мікрочіпових експериментів або методом секвенування молекул РНК для реконструкції генних мереж.

Підтверджую, що у докторській дисертації Бабічева Сергія Анатолійовича не використано результатів наукових досліджень із його кандидатської дисертації, яка була присвячена питанням розробки автоматизованої системи технічної діагностики міцнісних характеристик металів на основі гібридних нейронних мереж та захищено у 2003 році.

Повнота освітлення результатів дисертації

Результати роботи висвітлені у 46 наукових публікаціях, з яких 2 публікації у колективних англомовних монографіях, що включені до міжнародної наукометричної бази Scopus, 23 статті у фахових наукових виданнях з технічних наук України та за кордоном, 3 статті у наукових виданнях України, які індексуються у наукометричній базі Scopus, 18 публікацій у збірниках матеріалів міжнародних і національних конференцій, з яких 4 індексуються у наукометричних базах Scopus і Web of Science. 11 публікацій є одноосібними. Серед них статті, які опубліковано в авторитетних журналах: «Управляющие системы и машины», «Системні технології», «Biopolymers and Cell», «Communications in Computer and Information Science», «Advances in Intelligent Systems and Computing», «Індуктивне моделювання складних систем», «Штучний інтелект».

Автореферат у повній мірі відображує зміст дисертації.

Практична цінність результатів роботи

Практична цінність отриманих наукових результатів полягає в тому, що розроблені технології, методи та алгоритми становлять наукову основу для розробки й удосконалення методів обробки профілів експресій генів для реконструкції генних мереж, що створює умови для розробки нових методів ранньої технічної діагностики та лікування складних хвороб на генному рівні. Результати дисертаційної роботи впроваджені у Херсонському обласному онкологічному диспансері в системах ранньої діагностики онкологічних захворювань, у товаристві з додатковою відповідальністю «Херсонський маслозавод» для діагностики якості молочної продукції на основі лабораторних досліджень, Українській академії друкарства на факультеті видавничо-поліграфічної, інформаційної технології у процесі проведення лекційних занять і лабораторних робіт із курсів «Організація баз даних і баз знань», «Аналіз даних» та «Моделювання інформаційних систем і процесів», у Херсонському національному технічному університеті на кафедрі інформатики і комп’ютерних наук у процесі проведення лекційних занять і лабораторних робіт із курсів «Інтелектуальний аналіз даних і знань» та «Комп’ютерні інформаційні технології» та в університеті імені Яна Євангеліста Пуркіне в Усті над Лабем, Чехія на кафедрі інформатики в процесі проведення лекційних і семінарських занять із курсів «Data Mining», «Аналіз та візуалізація даних». Результати впровадження підтверджено відповідними актами.

Зауваження щодо змісту та оформлення дисертації

За змістом дисертації можна висловити такі зауваження:

1. У першому розділі роботи під час аналітичного огляду методів передобробки профілів експресій генів аналіз методів фільтрації та редукції профілів

експресій генів є неповним. Хотілося б побачити більш детальний аналіз існуючих методів обробки одновимірних даних.

2. Вейвлет-фільтрація одновимірних та двовимірних даних є досить розповсюдженою у даний час. Яка особливість запропонованої технології вейвлет-фільтрації профілів експресій генів у порівнянні з методами, що використовуються у даний час? Наприклад, з методом Гільберта-Хуанга. У чому новизна запропонованої моделі?
3. З підрозділу 2.2.3 не зовсім зрозуміло, чим визначався вибір функцій належності нечітких множин для вхідних та вихідних параметрів при налаштуванні системи нечіткого логічного виводу?
4. Об'єктивна кластеризація профілів експресій генів передбачає розрахунок внутрішніх та зовнішніх критеріїв якості кластеризації. Чи врахувалася в процесі розрахунку комплексного критерія вага відповідних критеріїв?
5. Практична реалізація індуктивної технології об'єктивної кластеризації виконувалася на основі алгоритмів кластеризації DBSCAN і SOTA. Чим обумовлений даний вибір? Чи проводилися дослідження по створенню гіbridних моделей об'єктивної кластеризації на основі інших алгоритмів кластерного аналізу?
6. Алгоритм кластеризації DBSCAN, який практично реалізований у програмному середовищі R, не передбачає використання кореляційної метрики, доцільність застосування якої для кластеризації профілів експресій генів показана здобувачем. Як даний факт враховувався у запропонованій моделі?
7. Гіbridна модель кластер-бікластерного аналізу передбачає виділення великої кількості бікластерів. Як у даному випадку обиралися бікластери, що використовувалися для реконструкції генних мереж?
8. Реконструйовані моделі генних регуляторних мереж містять інформацію про силу зв'язку між відповідними генами через ваговий коефіцієнт. Чи враховує запропонована технологія характер зв'язків між генами (спрямованість впливу)?
9. Частина рисунків містить позначення на 2-ох мовах одночасно: англійською та українською, потрібно було б узгодити термінологію на рисунках (наприклад: рис. 1.3, 1.14, 4.3, 5.16 і т.п.).
10. Відповідно до рис. 2.2, 2.3, 2.5 автор пише, що кращі результати були отримані методом James-Stein shrinkage estimator. Однак в позначеннях на рисунках цей метод отримав назву shrink (без авторських роз'яснень), що є не досить вдалим, бо інші використані методі скорочено до назв імен їх розробників.

Проте вказані зауваження не знижують загальну позитивну оцінку роботи, її наукову і практичну цінність.

Загальна оцінка дисертаційної роботи

На підставі вивчення змісту дисертаційної роботи, автореферату та наукових публікацій здобувача можна зробити такі висновки: дисертаційна робота С.А. Бабічева є завершеним науковим дослідженням, де виявлено та вирішено важливу наукову-прикладну проблему розробка теоретичних та практичних засад інформаційної технології обробки профілів експресій генів для реконструкції генних мереж.

В роботі отримано нові результати, що мають наукове та практичне значення для розвитку науки і техніки. Наведені у роботі наукові результати знайшли впровадження у різних галузях, що підтверджує їх практичну значимість, а їх використання сприятиме науково-технічному прогресу в областях технічної та медичної діагностики при розв'язанні завдань ранньої діагностики стану складних об'єктів, виготовлення ефективних ліків та створення нових ефективних методів лікування складних хвороб.

Дисертаційна робота С.А. Бабічева повністю відповідає чинним вимогам п.п. 9, 10, 12-14 «Порядку присудження наукових ступенів і присвоєння вченого звання старшого наукового співробітника» (Постанова Кабінету Міністрів України №567 від 24.07.2013 р.) щодо докторських дисертацій, а її автор, Бабічев Сергій Анатолійович, заслуговує на присудження наукового ступеня доктора технічних наук за спеціальністю 05.13.06 – інформаційні технології.

Офіційний опонент:

завідувач кафедри інформаційних технологій і систем,
Національної металургійної академії України,
Лауреат Державної премії України в галузі науки і техніки,
доктор технічних наук, професор

О.І. Михальов

Підпис офіційного опонента,
доктора технічних наук, професора Михальова О.І. засвідчує:
Вчений секретар Національної металургійної академії України,
професор



О.Ю. Потап

