

ВІДГУК

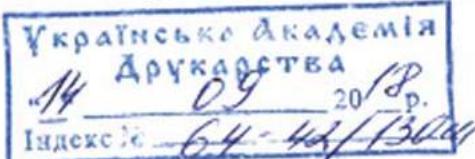
офіційного опонента

на дисертаційну роботу Бабічева Сергія Анатолійовича «Теоретичні та практичні засади інформаційної технології обробки профілів експресій генів для реконструкції генних мереж», яку подано на здобуття наукового ступеня доктора технічних наук за спеціальністю

05.13.06 - інформаційні технології

Актуальність теми дослідження

Зростання обсягів і складності інформаційних потоків, нестационарність та нелінійність даних, що обробляються, вимагають розробки нових та удосконалення існуючих підходів до вирішення проблеми інтелектуального аналізу та обробки даних складної природи. Особливу актуальність дана проблема набуває у галузі генної реверсної інженерії у сучасну епоху розшифровки геному людини. Якісно реконструйована генна регуляторна мережа дозволяє в процесі подальшого моделювання краще зрозуміти характер взаємодії генів у мережі, що в свою чергу створює умови для розробки нових методів діагностики та лікування складних хвороб. Визначення структури та характеру функціонування генної регуляторної мережі пов'язано з великими експериментальними та теоретичними складнощами. 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. До особливостей експериментальних даних слід віднести велику розмірність простору ознак. У випадку застосування технології ДНК-мікрочіпів дані також містять складну шумову компоненту, яка виникає внаслідок протікання біологічних та технологічних процесів на етапі формування даних. Тому особливу актуальність набувають задачі передобробки таких даних, які містять такі етапи, як фільтрація та редукція профілів експресій генів, кластеризація та бікластеризація взаємно-корельованих генів і умов визначення експресій відповідних генів, які використовуються для подальшої реконструкції генних регуляторних мереж. До окремих типів задач у даній предметній галузі слід віднести задачу валідації реконструйованих генних мереж, оскільки розв'язання даної задачі потребує визначення параметрів моделі, за якими



оцінюється адекватність реконструйованих генних мереж. Ефективне розв'язання таких задач можливе на основі комплексного підходу, який ґрунтуються на методах обчислюваного інтелекту, проте їх реалізація потребує розробки та впровадження відповідного інформаційно-технологічного забезпечення предметної області, що досліджується. Слід відзначити, що особливого значення це набуває у медичній галузі, де в умовах накопичування великих обсягів неоднорідних даних у режимі реального часу необхідно приймати рішення про стан біологічного організму, що досліджується. Добре розуміння характеру взаємодії генів у мережі, реконструйованої на основі даних профілів експресій генів, сприяє прийняттю своєчасних рішень по корегуванню характеру розвитку біологічного організму, що досліджується. Даний факт свідчить про актуальність теми дослідження.

Отже, дисертаційна робота присвячена вирішенню актуальної науково-прикладної проблеми розробки теоретичних та практичних зasad інформаційної технології обробки профілів експресій генів для реконструкції генних регуляторних мереж, розробки технології реконструкції та валідації моделей генних мереж із застосуванням кількох критеріїв оцінки топології мережі та створення ефективних методів інтелектуального аналізу й опрацювання складних даних.

Ступінь обґрунтованості наукових положень, висновків і рекомендацій

Основні наукові результати дисертаційного дослідження мають теоретичне обґрунтування та підтверджуються практичною реалізацією. Структура дисертації відображає основні теоретичні та практичні результати, які отримані в процесі дослідження. Дисертація складається із анотації, вступу, шести розділів, висновків та додатків. Характерною ознакою роботи є використання міждисциплінарного підходу, як підґрунтя для розробки гібридних моделей та методів обробки складних даних. Дисертантом проведено критичний аналіз наукових праць за такими напрямками досліджень, як: базові основи реконструкції генних мереж та методи створення експериментальних даних (59 джерел); методи фільтрації, редукції, кластеризації та бікластеризації профілів експресій генів та умов визначення експресій відповідних генів (82 джерела); сучасні методи реконструкції та моделювання генних регуляторних мереж (34 джерела). На основі зробленого аналізу дисертантом виділено ряд

невирішених проблем у галузі дослідження, сформульовано мету дослідження та завдання, розв'язування яких сприяє досягненню поставленої мети, а саме:

- розробка технології визначення оптимальної комбінації методів оцінки експресій генів об'єктів, отриманих при різних умовах проведення експерименту з використанням кількісних ентропійних критеріїв оцінки інформативності даних, що досліджуються;
- розробка технології 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;
- розробка індуктивної технології об'єктивної кластеризації профілів експресій генів на основі комплексного використання індуктивних методів аналізу складних систем та 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, отриманих із використанням рівнопотужних підмножин даних;
- розробка та практична реалізація гібридної моделі кластер-бікластерного аналізу на основі ієрархічних та щільнісних алгоритмів кластеризації у рамках індуктивної технології об'єктивної кластеризації й існуючих алгоритмів бікластеризації для отримання кластерів профілів генів з метою подальшої реконструкції та моделювання генної регуляторної мережі;
- розробка технології реконструкції та валідації моделі генної регуляторної мережі на основі статистичних методів аналізу інформації, проведення порівняльного аналізу різних методів моделювання ГРМ з метою визначення оптимального методу з точки зору критеріїв оцінки якості топології мережі;
- розробка нової інформаційної технології обробки профілів експресій генів для реконструкції генних регуляторних мереж на основі комплексного використання методів індуктивного моделювання складних систем та кількісних критеріїв оцінки якості інформації на кожному етапі її обробки;
- практична реалізація розроблених моделей, методів і алгоритмів у системах обробки профілів експресій генів з метою реконструкції та валідації моделей генних регуляторних мереж.

Центральне місце у математичному апараті, який використовує здобувач, займає підхід на основі обчислюваного інтелекту, що ґрунтуються на індуктивних методах аналізу складних систем та методах багатокритеріальної оптимізації. Комплексне застосування даних методів реалізовано у вигляді індуктивної технології об'єктивної кластеризації, реалізація якої сприяє зменшенню похибки відтворюваності, яка притаманна більшості існуючим алгоритмам кластерного аналізу. Як критерії оцінки якості кластеризації даних здобувачем запропоновано комплексний внутрішній критерій, який враховує як характер розподілу об'єктів у кластерах, так і характер розподілу кластерів у простору ознак, зовнішній критерій, який дозволяє оцінити різницю у кластеризаціях, отриманих на рівнопотужних підмножинах даних, та комплексний критерій балансу, який враховує можливі протиріччя між внутрішніми та зовнішнім критеріями. Даний підхід безумовно сприяє підвищенню об'єктивності кластеризації даних, що досліджується.

Методи багатокритеріальної оптимізації дисертантом також використовуються у запропонованій технології реконструкції генних регуляторних мереж. При цьому розраховуються топологічні параметри, що визначають топологію мережі, і вибір оптимальної топології мережі здійснюється на основі максимального значення комплексного топологічного параметра, який розраховується на основі приватних топологічних параметрів відповідної мережі на основі функції бажаності Харрінгтона. Валідація реконструйованих моделей генних мереж виконувалася шляхом порівняння характеру зв'язків у базовій мережі та у мережах на основі даних отриманих бікластерів із розрахунком похибок першого та другого роду та відносного комплексного критерія валідації відповідної моделі. Більше значення даного критерія свідчило про більший рівень адекватності мережі з точки зору аналізу характеру зв'язків між генами у мережі.

Таким чином, можна стверджувати, що розв'язування завдань дослідженъ дисертантом досягається шляхом розробки математичних моделей, методів та алгоритмів, які є основою розробленої інформаційної технології, ефективність якої підтверджується результатами імітаційного моделювання та фізичних експериментів.

Робота виконувалася у Херсонському національному технічному університеті у рамках досліджень відповідно до науково-дослідних

держбюджетних тем МОН України: «Розробка гіbridних нейро-фазі-імунних алгоритмів інформаційних систем та технологій для розв'язання задач в біоінформатиці та обчислювальній біології», «Синтез гіbridних еволюційних алгоритмів та методів для моделювання генних регуляторних мереж», «Розробка теоретичних зasad побудови інтелектуальних систем класифікації на основі онтології», у яких здобувач був виконавцем окремих етапів.

Достовірність отриманих результатів

Достовірність та обґрунтування отриманих у роботі результатів забезпечується коректним використанням: методів системного аналізу, методів ймовірнісно-статистичного моделювання, вейвлет-аналізу і методів нечіткого моделювання для аналізу та обробки профілів експресій генів, отриманих шляхом ДНК-мікрочіпових експериментів; методів математичної статистики, індуктивного моделювання складних систем, теорії оптимізації, кластерного аналізу для розробки індуктивної технології об'єктивної кластеризації та гібридної моделі кластер-бікластерного аналізу; теорії графів, теорії оптимізації, методів багатокритеріального аналізу та прийняття рішень для розробки технології реконструкції генних мереж; методів, засобів і технологій сучасного прикладного програмування для побудови практичних реалізацій.

Достовірність отриманих результатів також підтверджується публікаціями і тезами міжнародних конференцій, що індексуються у наукометричних базах Scopus і WoS, та результатами впровадження у навчальний процес вищих навчальних закладів, та результатами впровадження у Херсонському обласному онкологічному диспансері при розробці системи ранньої діагностики онкологічних захворювань та у товаристві з додатковою відповідальністю «Херсонський маслозавод» для діагностики якості молочної продукції на основі лабораторних досліджень, що підтверджується відповідними актами впровадження.

Основні наукові результати досліджень і новизна дисертації

Наукова новизна отриманих результатів полягає у розробленні теоретичних та практичних зasad інформаційної технології обробки профілів експресій генів для розв'язання задач реконструкції та валідації моделей генних регуляторних мереж, що дозволило отримати такі основні наукові результати.

1. Вперше розроблено і практично реалізовано гібридну модель кластер-бікластерного аналізу профілів експресій генів для групування генів та зразків з метою подальшої реконструкції генних мереж, що ґрунтуються на комплексному використанні щільнісного алгоритму кластеризації DBSCAN і самоорганізуючого алгоритму кластеризації SOTA у рамках індуктивної технології об'єктивної кластеризації та методу бікластеризації «ensemble», що дозволило підвищити якість обробки інформації за рахунок розпаралелювання процесу її обробки та застосування комплексних кількісних критеріїв якості обробки інформації на кожному етапі групування даних.

2. Вперше розроблено технологію фільтрації та редукції високорозмірних складних даних з метою видалення шумової компоненти і неінформативних ознак за статистичними та ентропійними критеріями, яка ґрунтуються на комплексному використанні вейвлет-аналізу і теорії нечіткого моделювання, що дозволило підвищити інформативність досліджуваних даних за рахунок зменшення рівня шуму та видалення ознак, що не є інформативними за групою критеріїв, які використовуються.

3. Вперше розроблено модель вейвлет-фільтрації профілів експресій генів для видалення фонового шуму, відмінною рисою якої є паралельна оцінка інформативності фільтрованих даних і видаленої шумової компоненти, яка дозволяє оптимізувати визначення параметрів вейвлет фільтру на основі комплексного аналізу фільтрованого сигналу та видаленої шумової компоненти.

4. Вперше розроблено технологію реконструкції та валідації моделей генних мереж, яка ґрунтуються на комплексному використанні топологічних параметрів мережі, індексі бажаності Харрінгтона та ROC-аналізі, що дозволило оптимізувати топологію мережі шляхом об'єктивного визначення параметрів алгоритму реконструкції генної регуляторної мережі.

5. Набула подального розвитку методологія індуктивного моделювання складних систем за рахунок застосування її основних положень до індуктивної технології об'єктивної кластеризації, що підвищує об'єктивність групування об'єктів шляхом використання внутрішніх і зовнішніх критеріїв якості кластеризації та комплексного критерію балансу.

6. Набули подального розвитку методи багатокритеріальної оптимізації в системах прийняття рішень за рахунок комплексного використання внутрішніх

і зовнішніх критеріїв якості обробки інформації та узагальненого індексу бажаності Харрінгтона.

7. Набули подальшого розвитку методи оптимізації визначення параметрів алгоритмів кластерного та бікластерного аналізів за рахунок використання комплексних кількісних критеріїв оцінки якості обробки інформації.

8. Удосконалено індуктивну технологію об'єктивної кластеризації для розв'язання задач групування об'єктів складної природи, яка ґрунтується на основних принципах індуктивного моделювання складних систем, методах багатокритеріальної оптимізації та системного аналізу, що дозволило підвищити об'єктивність кластеризації об'єктів шляхом зменшення похибки відтворюваності в процесі аналізу та обробки інформації.

9. Удосконалено технологію обробки даних мікрочіпових експериментів за рахунок системного підходу до обробки інформації і застосування критерію ентропія Шеннона, що розрахована за методом Джеймса та Стейна, для оцінки якості обробки даних на кожному етапі реалізації даного процесу.

Практичне значення отриманих результатів

Результати, що отримані дисертантом, становлять наукову основу для розробки й удосконалення методів обробки складних даних у різних практичних галузях наукових досліджень. Практична цінність отриманих наукових результатів полягає в тому, що розроблені моделі, методи та алгоритми забезпечують:

- підвищення ефективності фільтрації складних даних за рахунок коректного визначення параметрів вейвлет-фільтру, що відповідають максимальному значенню відношення ентропій Шеннона для виділеної шумової компоненти та корисного фільтрованого сигналу;
- підвищення інформативності даних за рахунок обґрунтовано проведеного процесу редукції даних за статистичними та ентропійними критеріями із застосуванням системи нечіткого логічного виводу;
- підвищення об'єктивності кластеризації складних даних за рахунок коректного використання основних принципів індуктивного моделювання складних систем у рамках індуктивної технології об'єктивної кластеризації;

- підвищення достовірності бікластерного аналізу профілів експресій генів за рахунок використання кількісних внутрішніх критеріїв оцінки якості бікластеризації даних, що досліджуються;
- підвищення якості реконструкції та валідації моделей генних мереж за рахунок коректно проведених процесів передобробки та кластеризації профілів експресій генів і застосування комплексної критеріальної оцінки топології мережі, що отримується.

Результати роботи апробовано і впроваджено у Херсонському обласному онкологічному диспансері в системах ранньої діагностики онкологічних захворювань, у товаристві з додатковою відповіальністю «Херсонський маслозавод» для діагностики якості молочної продукції на основі лабораторних досліджень, у 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», 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».

Оформлення дисертації та автореферату

Дисертаційна робота та автореферат написані на достатньо високому науково-технічному рівні. Опис досліджень має логічну структуру, висновки та рекомендації є доступними для сприйняття. Дисертаційна робота складається з анотацій на 23 сторінках, вступу, шести розділів, висновків, списку використаних літературних джерел на 30 сторінках, що включає 314 найменувань, шести додатків на 19 сторінках, 29 таблиць та 190 рисунків. Загальний обсяг роботи становить 404 сторінки, обсяг основного тексту – 309 сторінок. Роботу оформлено відповідно до вимог МОН України, що висуваються до дисертаційних робіт на здобуття наукового ступеня д.т.н.

Публікації та апробація результатів дисертації

За темою дисертаційної роботи опубліковано 46 наукових праць, з яких 2 публікації у колективних англомовних монографіях, що включені у міжнародну наукометричну базу Scopus, 23 статті у фахових наукових виданнях України та за кордоном з технічних наук, 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 одноосібні.

Відповідність змісту автореферату основним положенням дисертації

Автореферат та опубліковані за темою дисертації наукові публікації здобувача з достатньою повнотою відображають зміст дисертації. Автореферат за структурою, змістом і оформленням цілком відповідає вимогам МОН України до оформлення матеріалів дисертаційних досліджень.

Використання у докторській дисертації результатів наукових досліджень, за якими була захищена кандидатська дисертація

Дисертація на здобуття наукового ступеня к.т.н. дисертантом була захищена у 2003 році. Тема дисертації: «Автоматична система технічної діагностики міцнісних характеристик металів на основі гібридних нейронних мереж». Здобувачем не винесено на захист наукових положень, за якими була захищена його кандидатська дисертація.

Зауваження щодо змісту та оформлення дисертаційної роботи

1. У першому розділі під час аналітичного аналізу сучасного стану методів обробки профілів експресій генів для реконструкції генних мереж велика увага приділяється методам формування експериментальних даних та метода реконструкції та моделювання генних регуляторних мереж. Огляд методів фільтрації, редукції та кластеризації високорозмірних даних є неповним. Хотілося б побачити більш детальний аналіз робіт у галузі аналізу методів передобробки та кластеризації складних даних.

2. Одним із основних критеріїв оцінки інформативності профілю експресій генів у роботі використовувалася ентропія Шеннона, що розраховувалася за методом Джеймса та Стейна. У чому особливість даного методу і чому саме він використовувався для оцінки інформативності даних, що досліджувалися?

3. Яким чином визначалося граничне значення розділення генів на інформативні та неінформативні в системі редукції профілів експресій генів?

4. Існує велика кількість внутрішніх критеріїв оцінки якості кластеризації об'єктів. Чим визначався вибір критеріїв, що були досліджені у дисертаційній роботі? Яка їх перевага у порівнянні з іншими критеріями якості групування об'єктів?

5. Практична реалізація гібридної моделі об'єктивної кластеризації на основі алгоритму DBSCAN передбачає визначення двох параметрів, що визначають результат роботи алгоритму: EPS і MinPts. У дисертаційній роботі запропонована поетапна технологія визначення даних параметрів, при цьому інтервал зміни параметру EPS визначався на основі k-dist графу. Але форма k-dist графу, і як наслідок, і інтервал, може змінюватися в залежності від параметру MinPts. Як цей факт враховується в моделі?

6. У розділі 3 представлена архітектура індуктивної технології об'єктивної кластеризації, практична реалізація якої передбачає використання будь-якого алгоритму кластеризації в рамках даної технології. Чи проводилася автором оцінка ефективності даної технології при використанні інших алгоритмів кластеризації, окрім SOTA і DBSCAN? Якщо так, які результати були отримані?

7. Як видно з роботи, результатом бікластерного аналізу є велика кількість бікластерів різних розмірів. Дисертантом недостатньо обґрунтовано процедуру вибору кількості бікластерів для подальшої реконструкції генних регуляторних мереж.

8. Програмне середовище Cytoscape передбачає можливість застосування різних алгоритмів для реконструкції генних регуляторних мереж. Чим обумовлений вибір алгоритмів кореляційного виводу та ARACNE?

9. Значення відносного критерію валідації реконструйованих моделей генних регуляторних мереж на основі алгоритму кореляційного виводу змінюється від 21 до 1746. Статистична незначима мережа відповідає значенню даного критерію 1. Чи існує умовний поріг відносного критерія, нижче якого генну мережу не можна вважати адекватною для подального моделювання?

Слід зазначити, що вказані зауваження в цілому не змінюють загального позитивного враження та не впливають на оцінку роботи, яка виконана на високому науковому рівні.

Висновок

Дисертаційна робота С.А. Бабічева є завершеним науковим дослідженням, у якому вирішено актуальну науково-практичну проблему розробки теоретичних та практичних зasad обробки профілів експресій генів для реконструкції генних мереж. Отримані в роботі теоретичні та експериментальні результати забезпечують вирішення поставленої проблеми та мають наукове і практичне значення для розвитку інформаційних технологій. Вважаю, що дисертація відповідає паспорту спеціальності 05.13.06 – інформаційні технології та чинним вимогам МОН України «Порядку присудження наукових ступенів» щодо докторських дисертацій, а її здобувач **Бабічев Сергій Анатолійович** заслуговує на присудження наукового ступеня доктора технічних наук за спеціальністю 05.13.06 – інформаційні технології.

Офіційний опонент:

Директор інституту комп'ютерних
систем Одеського національного
політехнічного університету,
д.т.н., професор

С.Г. Антощук

Підпис д.т.н., проф. Антощук С.Г. засвідчує:

Вчений секретар



В. І. Швачук

