УДК 519.237.8:004.4:616-079

МЕТОДИ КЛАСТЕРИЗАЦІ В ПРОГРАМІ MICROARRAYTOOL ДЛЯ АНАЛІЗУ ДАНИХ ДНК-МІКРОАРРЕ В

С.С. Івахно[1], О.І. Корнелюк[1], О.П. Мінцер[2]

[1]Інститут молекулярної біології і генетики НАН України
[2]Національна медична академія післядипломної освіти імені П.Л.Шупика МОЗ України

Резюме. Мікроаррей-технології або ДНК-чіпи дозволяють проводити кількісний аналіз експресі десятків тисяч генів. В даній роботі описана нова програма Microarraytool для аналізу ДНК мікроаррей-даних, яка дозволяє проводити трансформацію та нормалізацію даних, виконувати кластерний аналіз та порівнювати різні експерименти за допомогою статистичного аналізу. Імплементовано такі методи кластерного аналізу: ієрархічний кластерний аналіз, метод кластеризаці k-середніх, карти ознак, що самоорганізуються (SOM) та SOTA-кластеризація. Проведено тестування алгоритмів для кластерного аналізу для мікроаррей-даних Стенфордсько бази даних з експресі первинних фібробластів людини для 8613 індивідуальних генів на різних часових проміжках після стимуляці . Аналіз даних показав коректне виконання алгоритмів, імплементованих в програмі Microarraytool.

Ключові слова: ДНК-мікроарре , біоінформатика, кластерний аналіз.

МЕТОДЫ КЛАСТЕРИЗАЦИИ В ПРОГРАММЕ MICROARRAYTOOL ДЛЯ АНАЛИЗА ДАННЫХ ДНК-МИКРОАРРЕЕВ

С.С. Ивахно[1], А.И. Корнелюк[1], О.П. Минцер[2]

[1]Институт молекулярной биологии и генетики НАН Украины
[2]Национальная медицинская академия последипломного образования имени П.Л. Шупика МЗ Украины

Резюме. Микроаррей-технологии или ДНК-чипы позволяют проводить количественный анализ экспрессии десятков тысяч генов. В данной работе описана новая программа Microarraytool для анализа ДНК микроаррей-данных, которая позволяет проводить трансформацию и нормализацию данных, выполнять кластерный анализ и сравнивать разные эксперименты с помощью статистического анализа. В программе имплементированы такие методы кластерного анализа: иерархический кластерний анализ, метод кластеризации k-средних, карты самоорганизующихся признаков (SOM) и SOTA-кластеризация. Проведено тестирование алгоритмов кластерного анализа для микроаррей-данных Стэнфордской базы данных по экспрессии первичных фибробластов человека для 8613 индивидуальных генов на разных временных промежутках после стимуляции. Анализ данных показал корректное выполнение алгоритмов, имплементированных в программе Microarraytool.

Ключевые слова: ДНК-микроарреи, биоинформатика, кластерный анализ.

CLUSTERING METHODS IMPLEMENTED INTO MICROARRAYTOOL PROGRAM FOR ANALYSIS OF DNA MICROARRAY DATA

S.S. Ivakhno[1], O.I. Kornelyuk[1], O.P. Mintser[2]

[1]Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine
[2]National Medical Academy of Post-Graduate Education named after P.L. Shupyk of Ministry of Public Health of Ukraine

Summary. Microarray technologies (DNA chips) allow to perform a quantitative analysis of expression of ten thousands genes. In this work a novel Microarraytool program was developed which allows to perform the cluster analysis and to compare the different experiments data by statistical analysis. Several clustering algorithms have been implemented into Microarraytool program: hierarchical clustering, k-means clustering, self-organizing maps (SOM) algorithm and self-organizing tree maps (SOTA) algorithm. The testing of these algorithms was performed using the Stanford Microarray Database for expression of 8613 individual genes in human fibroblasts after stimulation. The testing procedure revealed a correct performance of these algorithms implemented into Microarraytool program.

Key words: DNA microarrays, bioinformatics, clustering analysis.

ВСТУП. Мікроарреї (microarrays) або ДНК-чіпи є новим напрямком молекулярно-біологічних досліджень, який орієнтований на інтегральне вивчення експресії геному. Ця новітня технологія дозволяє проводити детекцію, ідентифікацію та кількісний аналіз експресії десятків тисяч генів [1]. В експерименті досліджується гібридизація мічених флуоресцентними барвниками молекул РНК, виділених з клітин, до ДНК-мішеней на поверхні ДНК-чіпа (рис. 1). На наступному етапі проводиться детекція флуоресценції з кожної точки (окремої ДНК-мішені), причому рівні інтенсивностей залежать від комплементарності молекул ДНК до РНК-проб. Оскільки кожна точка на поверхні ДНК-чіпа має генну анотацію, інтенсивність флуоресценції після зв'язування з відповідною РНК-пробою інтерпретується як рівень експресії специфічної мРНК. Мікроаррей-технології знайшли широке використання в аналізі диференційної генної експресії [2], детекції однонуклеотидного поліморфізму [3], генотипуванні [6], вивченні екзон-інтронної будови генів [7], філогенетичному аналізі [8], ідентифікації маркерів пухлин [9] та інших застосуваннях [8-10].

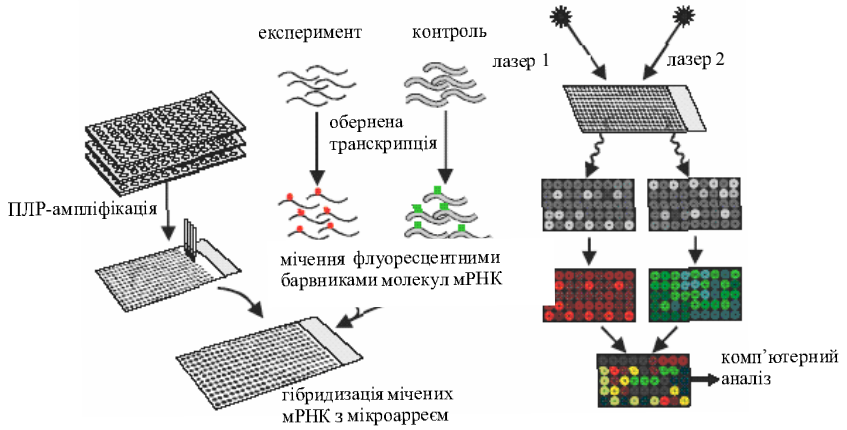

Рис 1. Схема кДНК мікроаррей-експерименту. кДНК наносяться на мікроскопічний слайд, потім кДНК гібридизується з пулом молекул мРНК, мічених флуоресцентними барвниками. Після гібридизації слайд сканується двома лазерами, отримані знімки використовуються в подальшому аналізі.

На даний момент розроблено кілька програм для аналізу мікроаррей-даних, які використовують різноманітні методи нормалізації та кластеризації. Більшість алгоритмів для аналізу мікроарреїв були запозичені з математичних розробок в інших галузях знань, передусім, економіки та інженерії [11]. Ці алгоритми, як правило, розроблені для аналізу даних із невеликою кількістю елементів (не більше 1 000), хоча стандартний мікроаррей містить десятки тисяч генів.

Метою даної роботи є розробка нової програми Microarraytool для аналізу мікроаррей-даних, що використовує адаптовані варіанти алгоритмів нормалізації та кластеризації. Серія адаптацій програмного коду алгоритмів дозволяє прискорити їх виконання і, таким чином, робить можливим аналіз з великою кількістю елементів.

МАТЕРІАЛИ І МЕТОДИ. **Трансформація мікроаррей-даних.** В програмі Microarraytool використані кілька методів трансформації мікроаррей-даних, які можуть бути застосовані перед статистичною обробкою експериментальних даних. Використовується логарифмічна трансформація, оскільки результати експериментів представляються у вигляді пропорцій флуоресцентних барвників. Крім того,

використовуються центрування за середнім чи медіаною, поділ на середнє квадратичне чи дисперсію, дискретизація даних.

Методи обчислення відстаней для алгоритмів кластеризації. Всі кластерні алгоритми використовують методи обчислення відстаней між генними векторами для порівняння рівнів генної експресії. В програмі Microarraytool застосовані наступні методи: кореляційний коефіцієнт Персона, нецентризований коефіцієнт Персона, ступеневий кореляційний коефіцієнт Персона, косинусний кореляційний коефіцієнт, кодисперсія, евклідова відстань, манхеттенська відстань та непараметрична кореляція Спермана.

Кластерний аналіз. Кластерний аналіз може бути визначений як процес розподілу набору об'єктів в окремі підмножини на основі їхньої подібності [12]. Метою аналізу є отримання кластерів, які, з одного боку, максимізують міжкластерну варіабельність, а з іншого – мінімізують внутрішньокластерну відстань. Отже, метою є знаходження набору генів, що мають схожі між собою рівні експресії та максимально відрізняються від рівнів експресії інших генів. В програмі Microarraytool імплементовано кілька різних методи кластерного аналізу.

1. Ієрархічний кластерний аналіз. Цей метод кластеризації трансформує матрицю відстаней в ієрархію вкладених підмножин. Ієрархія може бути представлена деревоподібною дендрограмою, в якій кожний кластер вкладений в інший кластер. В Microarraytool було імплементовано агломеративний кластерний аналіз. Агломеративний кластерний аналіз виконується в наступному порядку : 1). Спершу обчислюється відстань між усіма об'єктами і будується матриця відстаней. Кожний об'єкт представляє кластер, що містить лише його самого. 2). На наступному етапі знаходяться два кластери з мінімальною відстанню один від одного, які потім об'єднуються і замінюються одним кластером. Ці процедури повторюються, доки загальне число кластерів не буде дорівнювати одиниці.

2. Метод кластеризації k-середніх. Метод кластеризації k-середніх – це ітеративне обчислення двох параметрів: параметра централізованості для кожного кластера та параметра розподілу всіх вхідних векторів в кластери. Алгоритм k-середніх складається з двох ітеративних фаз: переміщення вхідних векторів до кластерів, які є найближчими до них; та обчислення нових центрів кластерів відповідно до параметра централізованості. Основою методу кластеризації k-середніх є мінімізація функції оптимізації для всіх кластерів. Алгоритм складається з наступних кроків: включення кожного вектора x_i з X в один з k кластерів, обчислення середнього для кожного з k кластерів, обчислення відстаней між кожним вектором і середнім значенням кожного кластера та переміщення вектора до кластера, що має найближче до нього середнє значення.

Алгоритм кластеризації k-середніх є дуже ефективним для аналізу великого набору даних. Інша перевага алгоритму k-середніх – незначні вимоги до комп'ютерної пам'яті, оскільки алгоритм не виковистовує матрицю відстаней. Кластери, що будуються алгоритмом k-середніх, мають, як правило, колоподібну форму.

3. Карти ознак, що самоорганізуються (Self-Organizing Maps). Однією з найбільш уживаних моделей нейронних сіток для кластерного аналізу є карти ознак, що самоорганізуються (SOM, Self-Organizing Maps) [13]. SOM є методом зменшення вимірності простору, який направляє вхідні дані з багатовимірного простору в вихідні дані меншої вимірності [14]. Як і інші нейроні сітки, SOM складається з нейронів, що розміщені в регулярному, звичайно одно- чи двовимірному просторі [15]. Кожний нейрон i сітки SOM репрезентується n-вимірним референтним вектором, де n дорівнює числу вимірів вхідних векторів. Нейрони сітки приєднані до інших нейронів сусідніми зв'язками, які визначають організацію сітки. SOM виконується в наступному порядку: на першому етапі проводиться ініціалізація. Далі проводиться фаза тренування, коли на кожному кроці тренування один вхідний вектор x вибирається випадково з вхідних даних і обчислюється відстань між ним та всіма референтними векторами, що формують сітку SOM. Параметр навчання – це функція часу, яка змінюється в процесі виконання SOM-алгоритму. Найчастіше параметр навчання лінійно зменшується з часом t. Фаза тренування закінчується після визначеного числа ітерацій. Важливою перевагою SOM є те, що вона може

використовуватися навіть при відсутності значень в елементах векторів генної експресії. На наступному етапі проводиться кластеризація. Алгоритм SOM розподіляє вхідні дані у випуклі регіони Вороного, кожний з яких відповідає одній одиниці (вектору) сітки SOM. Регіон Вороного V_i визначається як множина усіх векторів x, до яких референтний вектор x є найближчим:

$$V_i = \{ \mathbf{x} \mid \|\mathbf{m}_i - \mathbf{x}\| < \|\mathbf{m}_j - \mathbf{x}\|, i \neq j \}.$$

Це означає, що вхідні вектори можна кластеризувати відповідно до їхніх регіонів Вороного. Кожний регіон Вороного, таким чином, відповідає одному кластеру, що складається з подібних між собою вхідних векторів. Отже, число кластерів має бути визначене перед початком роботи алгоритму, оскільки воно залежить лише від кількості нейронів у сітці SOM. Можна використовувати декілька експериментальних ітерацій алгоритму SOM для встановлення оптимального набору параметрів, а потім застосувати їх на кінцевій ітерації алгоритму. Оскільки число циклів у фазі тренування вибирається користувачем, для отримання надійних результатів необхідно як мінімум 100 ітерацій.

4. **SOTA-кластеризація.** Алгоритм SOTA (Self-Organizing Tree Maps) – це нейронна сітка, яка в процесі росту приймає топологію бінарного дерева [16-18]. Алгоритм SOTA базується на сітках SOM, але має декілька інновацій, які роблять його привабливішим для кластерного аналізу. Стандартний алгоритм SOM присвоює вхідні дані складного багатовимірного простору вихідним даним, що складаються з невеликої кількості кластерів. У випадку SOM вихідні дані представлені сіткою у двовимірному просторі, у випадку SOTA вихідні дані складають бінарне дерево варіабельної структури, що змінюється залежно від типу вхідних даних. Як і SOM, референтні вектори в SOTA складаються з n елементів, число яких дорівнює числу елементів у вхідних векторах. Фаза тренування, однак, починається лише з одного вектора, який розподіляється на два, якщо гетерогенність його елементів перевищує встановлений максимальний параметр гетерогенності. Тренування призупиняється, коли зміни показника між двома послідовними ітераціями стають меншими встановленого скаляра. Перевагою SOTA перед SOM є динамічна репрезентація вихідних векторів, що дозволяє більш точно кластеризувати вхідні дані генної експресії.

Інструменти програмування. Java (Sun Microsystems) був вибраний як мова програмування. Для розробки програми Microarraytool використовувався інтегрований простір для розробки програм Jbuilder (Borland Inc.). Технологія Java є одночасно мовою програмування і платформою. В Java програма спочатку компілюється в перехідну мову Java-байткодів, яка є платформно незалежною і легко інтерпретується Java-платформою. Java-платформа є програмною платформою, яка виконується на інших комп'ютерних платформах (Windows, Linux, iMac та інші). Платформа Java складається з двох компонентів – віртуальної машини Java (Java Virtual Machine, Java VM) та Java-інтерфейсу розробки програм (Java Application Programming Interface, Java API).

РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ. Розроблена нова програма Microarraytool для аналізу мікроаррей-даних, яка використовує адаптовані варіанти алгоритмів нормалізації та кластеризації, що дозволяє прискорити їх виконання і робить можливим аналіз з великою кількістю елементів. Ключовим етапом розробки програмного забезпечення є тестування програми та верифікація програмного коду. Для тестування програми Microarraytool ми використали мікроаррей-дані з експресії первинних фібробластів людини, які були стимульовані шляхом вилучення поживного середовища. Експеримент був проведений на кДНК-мікроарреях, що складалися з 9 800 кДНК (8613 індивідуальних генів). Мікроаррей-дані були завантажені з бази даних Стенфордського університету Stanford Microarray Database (США). Дані представлено матрицею генної експресії, що показує відносний рівень експресії генів у часових проміжках 0, 15, 30 хвилин, 1, 2, 3, 4, 8, 12, 16, 20 годин після вилучення поживного середовища. Референтним вектором у цьому експерименті виступав абсолютний рівень експресії генів у часовому проміжку 0 годин після вилучення поживного середовища. Матриця генної експресії була завантажена в структуру даних програми Microarray. Завантажені мікроаррей-дані були проаналізовані з використанням ієрархічного

агломеративного кластерного аналізу, методу кластеризації k-середніх та методу кластерного аналізу карт ознак, що самоорганізуються (SOM). Отримані результати було порівняно для визначення коректності виконання кластерних алгоритмів.

Ієрархічний кластерний аналіз є найбільш поширеним методом для аналізу мікроаррей-даних. Наочна візуалізація даних, отриманих після кластеризації, є перевагою цього методу. В програму Microarraytool було імплементовано такі методи об'єднання кластерів, як метод найближчого сусіда, метод найвіддаленішого сусіда та метод середнього. Перевагою програми є можливість одночасної кластеризації генів та експериментів. Мікроаррей-експеримент зі стимуляції фібробластів людини аналізувався ієрархічним агломеративним кластерним аналізом з використанням наступних параметрів: 1). Кластеризація генів; 2). Метод середнього для об'єднання кластерів. Алгоритм ідентифікував 10 кластерів з різним рівнем відмежованості один від одного: кластери 5, 6 виявилися найбільш відмежованими один від одного; кластери 2, 3 – найбільш подібними (рис. 2). В той же час, аналіз центроїдного графіка цих кластерів показав доречність їхнього розмежування в окремі підмножини. Загалом результат кластеризації підтверджує правильність виконання алгоритму.

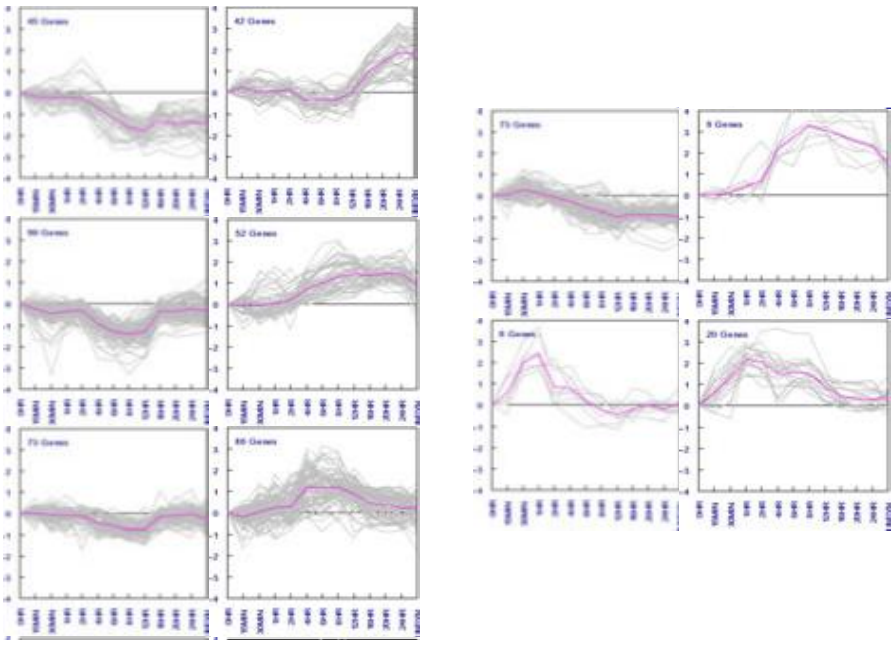

Рис. 2. Результати ієрархічного агломеративного кластерного аналізу мікроаррей-даних із стимуляції фібробластів людини. Центроїдні графіки показують розподіл кластерів.

Мікроаррей-експеримент зі стимуляції фібробластів людини аналізувався методом k-середніх з використанням наступних параметрів: кількість кластерів – 10; число ітерацій – 50. Кількість кластерів була встановлена з урахуванням попередніх результатів ієрархічного кластерного аналізу. Кількість ітерацій алгоритмів визначалось емпірично шляхом підбору оптимального числа ітерацій після повторного виконання алгоритму. Як правило, оптимальність досягалася після 50 ітерацій. Було отримано 10 кластерів, серед яких

кластери 5 та 10 виявилися найбільш подібними (рис. 3). В той же час аналіз центроїдного графіка показав більш знижену експресію для кластера 5, починаючи з 12 години після вилучення поживного середовища. Це спостереження вказує на правильне розмежування кластерів методом k-середніх Microarraytool. Застосовані нами ієрархічний кластерний аналіз та метод k-середніх показали майже однаковий розподіл експресійних профілів по різних кластерах, хоча кількість елементів у кластерах дещо різнилась між двома методами.

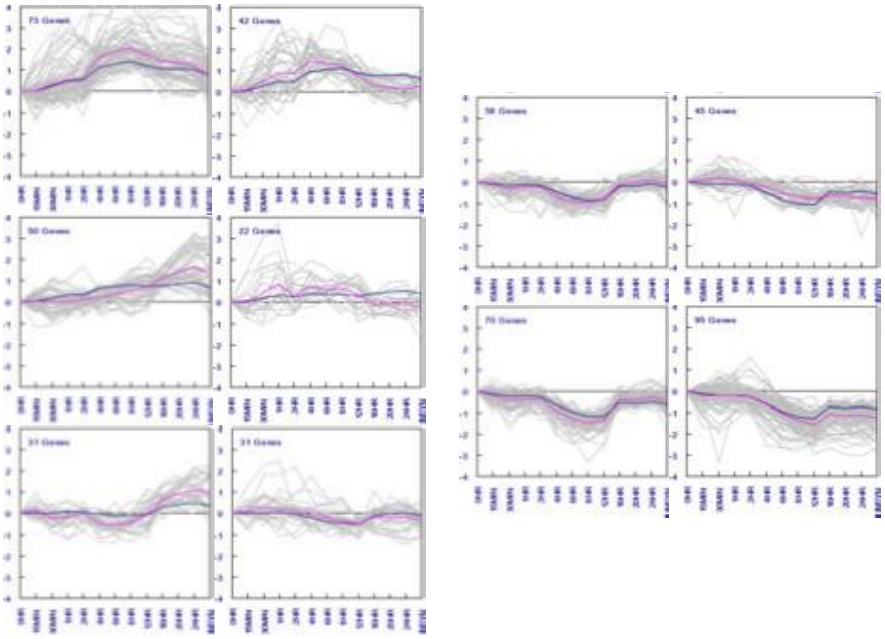

Рис. 3. Результати кластерного аналізу мікроаррей-даних із стимуляції фібробластів людини методом k-середніх.

Алгоритм кластеризації SOM вимагав більше параметрів для встановлення. Були використані такі параметри: число нейронів на осі x – 5; число нейронів на осі y – 2; кількість ітерацій алгоритму – 2 000; показник тренування α – 0,05; сусідський радіус – 3; функція сусідів – функція Гауса (нормальний розподіл); топологія сітки SOM – гексагональна; ініціалізація алгоритму – використання випадкових генних величин. Аналіз генної експресії мікроаррей-даних із стимуляції фібробластів алгоритмом SOM показав кластеризацію, подібну до кластеризації методом k-середніх (рис. 4). Центроїди кластерів майже однакові для обох алгоритмів. В той же час, кластери мають неоднакову кількість генів: кластер 4 в методі k-середніх має 35 генів, а кластер 1 алгоритму SOM – 73 гени. Останнє є результатом нерівномірного розподілу генів між кластерами. Оскільки алгоритми k-середніх та SOM ініціалізуються випадковими величинами, кожне їх виконання буде давати дещо змінений розподіл кластерів. В той же час, як показує порівняння кластера 2 ієрархічного кластерного аналізу методу k-середніх та кластера 5 алгоритму SOM, вони повністю подібні й мають майже однакову кількість генів – 99, 94 та 70 генів відповідно. Це дозволяє нам зробити висновок, що виконання імплементованих у Microarraytool кластерних алгоритмів є правильним і здатним до відтворення результатів.

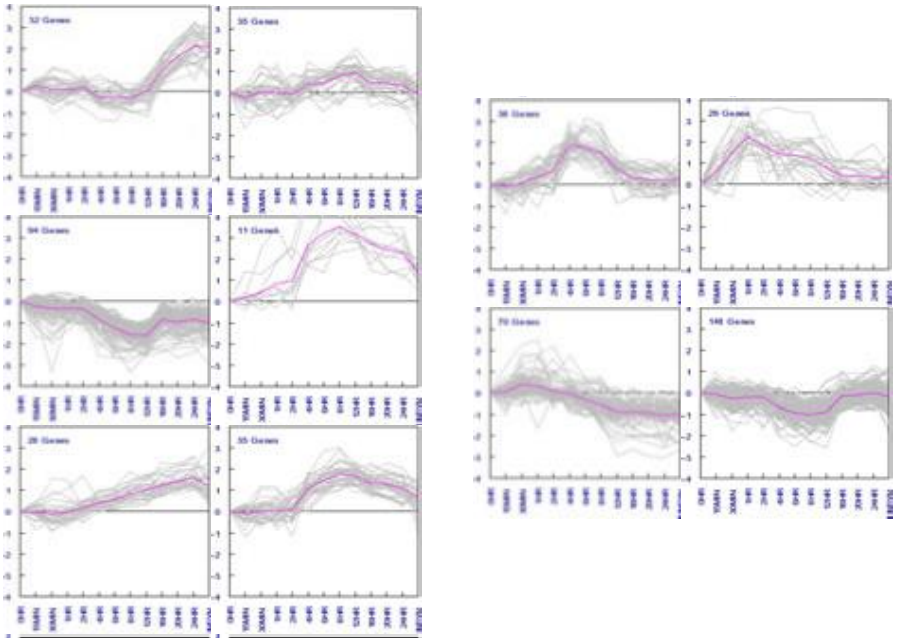

Рис. 4. Результати кластерного аналізу мікроаррей-даних із стимуляції фібробластів людини алгоритмом карт ознак, що самоорганізуються (SOM).

ВИСНОВОК. В даній роботі розроблена програма Microarraytool для аналізу ДНК мікроаррей-даних. Програма дозволяє проводити трансформацію та нормалізацію даних, виконувати кластерний аналіз та порівнювати різні експерименти за допомогою статистичного аналізу. Імплементовано такі методи кластерного аналізу: ієрархічний кластерний аналіз, метод кластеризації k-середніх, карти ознак, що самоорганізуються (SOM) та SOTA-кластеризація.

Проведено тестування алгоритмів для кластерного аналізу для мікроаррей-даних з експресії первинних фібробластів людини, які показували рівень експресії для 8613 індивідуальних генів на різних часових проміжках після стимуляції фібробластів. Аналіз даних показав коректне виконання алгоритмів, імплементованих в програмі Microarraytool.

Ë³òåðàòóðà

1. Baldi P., Hatfield G.W. DNA Mscroarrays and gene expression : From experiments to data analysis modeling. – Cambridge University Press, 2002.
2. Soares M.B. Identification and cloning of differentially expressed genes // Curr. Opin. Biotechnol. – 1997. – V. 8. – ¹ 5. – P. 542-546.
3. Campbell A.M., Heyer L.J. Discovering genomics, proteomics, and bioinformatics. – CSHL Press, 2003.
4. Zhang N., Tan H., Yeung E.S. Automated and integrated system for highthroughput DNA genotyping directly from blood // Anal. Chem. – 1999. – V. 71. – P. 1138- 1145.
5. Raitio M., Lindroos K., Laukkanen M. et al. Y-chromosomal SNPs in Finno-Ugric-speaking populations analyzed by minisequencing on microarrays // Genome Res. - 2001. – V. 11. – ¹ 3. – P. 471-482.
6. Behr M.A., Wilson M.A. Comparative genomics of BCG vacines by whole-genome DNA microarray // Science. – 1999. – V. 284. – P. 1520-1523.
7. Khan J., Saal L.H., Bittner M.L. et al. 1999. Expression profiling in cancer using cDNA microarrays // Electrophoresis. – 1999. – V. 20. – ¹ 2. – P. 223- 229.
8. ²âàõîí Ñ., Êîðíåëþê Î. Ì³êðîàðð¿: îãëÿä òåõí-

ëîã³é òà àíàë³ç äàíèõ // Óêð. á³îõ³ì. æóðí. –2004. – Ò. 76, ¹ 2.– Ñ. 5-19.
9. Heller M. J. DNA Microarray Technology**:** Devices, Systems, and Applications // Annu. Rev. Biomed. Eng. – 2002. – V. 4. – P. 129-153.
10. Zanders E.D. Gene expression analysis as an aid to the identification of drug targets // Pharmacogenomics. - 2000. – V. 1. – ¹ 4. – P. 375-384.
11. Jain A.K., Murty M.N., Flynn P.J.. Data clustering: A review // ACM Computing Surveys. – 1999. – V. 31. – ¹ 3. – P. 264 -323.
12. Jain A. K., Dubes R.C. Algorithms for Clustering Data. Prentice Hall Advanced Reference Series. Prentice Hall, New Jersey, 1988.
13. Toronen P., Kolehmainen M., Wong G. et al. Analysis of gene expression data using self- organizing maps // FEBS Letters. – 1999. – V. 451. – P. 142-146.
14. Dopazo J., Carazo J.M. Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree // J. Mol. Evol. – 1997. – V. 44. – P. 226-233.
15. Fritzke B. Growing cell structures – a self-organizing network for unsupervised and supervised learning // Neural Networks. – 1994. – V. 7. – P. 1141-1160.
16. Dopazo J., Carazo J.M. Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree // J. Mol. Evol. – 1997. – V. 44. – P. 226-233.
17. Fritzke B. Growing cell structures – a self-organizing network for unsupervised and supervised learning // Neural Networks. – 1994. – V. 7. – P. 1141-1160.
18. Herrero J., Valencia A., Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns // Bioinformatics. – 2001. – V. 17. – ¹ 2. – P. 126-136.