УДК 004.048

С.А. Бабічев

Технологія вейвлет-фільтрації профілів експресій генів з метою видалення фонового шуму

Представлена технология вейвлет-фильтрации профилей экспрессий генов для удаления фонового «белого» шума на основе критерия энтропия Шеннона, который рассчитан посредством использования метода оценки Джеймса и Стейна. Предложена структурная блок-схема процесса определения параметров вейвлет-фильтра, предполагающая расчет энтропии Шеннона как для фильтрованного сигнала, так и для удаленной шумовой компоненты.

Ключевые слова: профили экспрессий генов, вейвлеты, трешолдинг, фильтрация.

Подано технологію вейвлет-фільтрації профілів експресій генів для видалення фонового «білого» шуму на основі критерію ентропія Шеннона, розрахованого з використанням методу оцінки Джеймса та Стейна. Запропоновано структурну блок-схему процесу визначення параметрів вейвлет-фільтру, яка передбачає розрахунок ентропії Шеннона як для фільтрованого сигналу, так і для видаленої шумової компоненти.

Ключові слова: профілі експресій генів, вейвлети, трешолдінг, фільтрація.

Вступ. Детальний аналіз процесу формування матриці експресій генів, отриманих шляхом мікрочіпових експериментів [1, 2] показує, що етап сканування зображення мікрочіпу ДНК супроводжується появою фонового шуму. Часткова фонова корекція проводиться при формуванні матриці експресій генів, але слід зазначити, що повне видалення шумової складової на даному етапі є проблематичним. Тому виникає необхідність подальшої фільтрації отриманих профілів експресій генів з використанням кількісних критеріїв оцінки інформативності профілів, що обробляються, або шумової складової, що видаляється з відповідних генів. У даній статті задача розв'язується з використанням вейвлет-аналізу [3–5], який отримав широке розповсюдження у різних галузях для обробки складних нестаціонарних сигналів та зображень [6–10].

Аналіз сучасних досягнень та публікацій

Питанням створення матриці експресій генів, отриманих шляхом ДНК-мікрочіпових експериментів, присвячено роботи [11–14], де детально розглянуто етапи обробки експериментальних даних, які містять фонову корекцію, нормалізацію, PM-корекцію та сумарізацію. На кожному етапі передбачено використання різних методів, кожен з яких має і переваги, і недоліки. У [15] представлено порівняльний аналіз методів обробки профілів експресій генів, хворих на рак легенів, отриманих шляхом ДНК-мікрочіпових експериментів, визначено оптимальну комбінацію методів з використанням критерію ентропія Шеннона. Питанням редукції неінформативних профілів експресій генів з огляду на статистичні та ентропійні критерії присвячено роботу [16]. Представлена технологія покрокової обробки профілів експресій генів дозволяє скоротити кількість ознак на 6–10 відсотків, що підвищує інформативність даних для подальших досліджень. Однак слід зазначити, що на етапі фонової корекції проблемі видалення фонового білого шуму, який виникає на етапі сканування зображення мікрочіпу, приділяється недостатня увага. У [17] запропоновано методику фільтрації хроматограмм з використанням вейвлет-аналізу та критерію ентропія Шеннона. Але для профілів експресій генів дана задача на даний час однозначного розв'язку не має.

Метою статті є розробка технології вейвлет-фільтрації профілів експресій генів із застосуванням критерію ентропія Шеннона, що розра-

хована на основі методу Джеймса та Стейна, з метою видалення фонового білого шуму.

Основний матеріал

До вейвлетів належать функції, що будуються на основі одного материнського вейвлету $\psi(t)$ шляхом операцій зсуву за аргументом τ та масштабною зміною за параметром a [4]:

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-\tau}{a}\right), \qquad (1)$$

де $(a,\tau) \in R$, $\psi(t) \in L^2(R)$, а множник $1/\sqrt{|a|}$ забезпечує незалежність норми функції від масштабного числа a. Материнському вейвлету $\psi(t)$ притаманні наступні властивості:

• *обмеженість*, тобто вейвлет-функції $\psi(t)$ повинні мати кінцеву енергію:

$$\|\psi(t)\| = \int_{-\infty}^{+\infty} |\psi(t)|^2 \, dt < \infty; \qquad (2)$$

• *локалізація*, тобто вони мають бути визначені на кінцевому інтервалі як у часовій, так і у частотній областях. Для цього достатньо, щоб виконувалися при C = const та $\varepsilon > 0$ наступні умови:

$$|\psi(t)| \le C(1+|t|)^{-1-\varepsilon} \text{ та } |\psi(\omega)| \le C(1+|\omega|)^{-1-\varepsilon}, \qquad (3)$$

де ω – середня частота вейвлету;

• *нульове* середнє значення, тобто виконання наступної умови для нульового моменту:

$$\int_{-\infty}^{+\infty} \psi(t)\, dt = 0. \qquad (4)$$

У випадку безперервного вейвлет-перетворення сигналу $s(t) \in L^2(R)$ параметри a та τ набувають будь-яких значень у межах області їх визначення і вейвлетний масштабно-часовий спектр $C(a,b)$ розраховується у відповідності з формулою

$$C(a,b) = \int_{-\infty}^{+\infty} s(t) a^{-1/2} \psi\left(\frac{t-b}{a}\right) dt. \qquad (5)$$

Вейвлет-реконструкція сигналу $s(t)$ за умови використання такого ж базису функцій виконується так:

$$s(t) = \frac{1}{C_\psi} \int_{R^+} \int_R C(a,\tau) a^{-1/2} \psi\left(\frac{t-\tau}{a}\right) \frac{da \cdot d\tau}{a^2}, \qquad (6)$$

де C_ψ – константа, що визначається функцією ψ:

$$C_\psi = 2\pi \int_{-\infty}^{\infty} |\hat{\psi}(\omega)|^2 |\omega| \, d\omega, \qquad (7)$$

де $\hat{\psi}$ – є фур'є образ функції ψ. Для кількісного аналізу сигналів, який передбачає його декомпозицію та реконструкцію можливе використання будь-яких локалізованих функцій $\psi(t)$, якщо для них існують такі функції-двійники $\psi'(t)$, що сімейства $\{\psi_{a,\tau}(t)\}$ та $\{\psi'_{a,\tau}(t)\}$ утворюють парні базиси функціонального простору $L^2(R)$. Якщо вейвлет $\psi(t)$ ортогональний, то $\psi(t) \equiv \psi'(t)$ і вейвлетний базис також має властивість ортогональності. У випадку неортогональності вейвлету $\psi(t)$ наявність двійника $\psi'(t)$ дає можливість сформувати сімейства $\{\psi_{m,k}(t)\}$ та $\{\psi'_{g,p}(t)\}$, що задовольняють умову біортогональності на множині цілих чисел Z:

$$\langle \psi_{m,k}(t), \psi'_{g,p}(t) \rangle = \delta_{m,k}\delta_{g,p}, \; m,k,g,p \in Z. \qquad (8)$$

У цьому випадку можлива декомпозиція будь-якого сигналу та його подальша реконструкція. При дискретній вейвлет-декомпозиції сигналу параметри a та τ набувають дискретних значень на множині:

$$a = 2^j \text{ та } \tau = k2^j, \qquad (9)$$

де j та k – є цілі числа, коефіцієнт j є параметром масштабу, а k визначає рівень вейвлет-декомпозиції сигналу. За дискретних значень a і τ вейвлет-функція набуває вигляду

$$\psi_{j,k}(t) = a_0^{-j/2} \psi\left(a_0^{-j} t - k\right), \qquad (10)$$

а пряме дискретне вейвлет-перетворення зводиться до обчислення деталізуючих коефіцієнтів:

$$d_{j,k} = \int_{-\infty}^{+\infty} a_0^{-j/2} \psi\left(a_0^{-j} t - k\right) s(t) dt. \qquad (11)$$

Обернене дискретне вейвлет-перетворення задається за допомогою того ж базису, як і пряме:

$$s(t) = \frac{1}{C_\psi} \sum_{j \in Z} \sum_{k \in Z} d_{j,k} a_0^{-j/2} \psi\left(a_0^{-j} t - k\right). \quad (12)$$

Структурну схему процесу дискретного вейвлет-перетворення даних, що досліджуються, подано на рис. 1. Процес вейвлет-декомпозиції одновимірного сигналу передбачає розрахунок вектора коефіцієнтів, які апроксимують на *N*-рівні вейвлет-декомпозиції та вектори коефіцієнтів, які деталізують на рівнях від одиниці до *N*:

$$s = CA_0 \to \{CA_1, CD_1\} \to \{CA_2, CD_2, CD_1\} \to \\ \to ... \to \{CA_N, CD_N, CD_{N-1}, ..., CD_1\}. \quad (13)$$

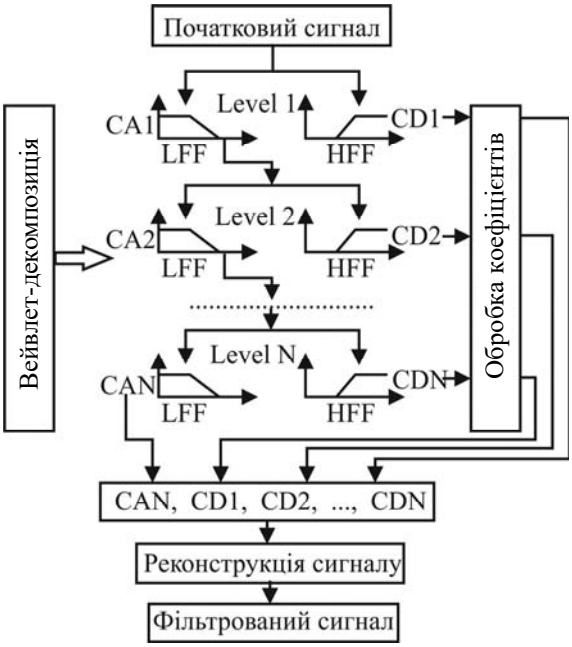


Рис. 1

Вектори коефіцієнтів, які апроксимують та деталізують, розраховуються шляхом використання для вектора відповідних даних фільтра низьких частот *LFF* при апроксимації та фільтра високих частот *HFF* для деталізації. Апроксимуючі коефіцієнти містять інформацію про грубу складову сигналу; інформація про деталі сигналу та високочастотна шумова складова містяться у більшості деталізуючих коефіцієнтів. Видалення шумової складової з профілів експресій генів у межах запропонованої моделі проведено з використанням м'якого трешолдингу:

$$\begin{cases} d = 0, \ \text{якщо } d \le \tau, \\ d = d - \tau, \ \text{якщо } d > \tau, \end{cases} \quad (14)$$

де τ – значення трешолдингового коефіцієнта, d – значення деталізуючих коефіцієнтів на усіх рівнях досліджуваних вейвлет-декомпозицій вектора даних.

Реконструкція даних відбувається з використанням апроксимуючих коефіцієнтів на *N*-рівні вейвлет-декомпозиції та оброблених деталізуючих коефіцієнтів на рівнях від одиниці до *N*.

Аналіз рис. 1 дозволяє визначити шляхи оптимізації процесу вейвлет-фільтрації профілів експресій генів. Процес передбачає наступні кроки:

- вибір материнського вейвлету;
- визначення оптимального рівня вейвлет-декомпозиції вектора досліджуваних даних;
- вибір типу вейвлету з сімейства материнського вейвлету;
- визначення оптимального значення трешолдингового коефіцієнта.

Оцінку якості обробки інформації на кожному кроці проведено з використанням ентропії Шеннона на основі оцінки Джеймса та Стейна [18]. Вочевидь, що мінімальне значення критерію ентропія Шеннона, розраховане для профілю експресій гена, відповідає найбільш високій якості обробки інформації. З іншого боку, максимальне значення ентропії Шеннона, розраховане для видаленої шумової компоненти, відповідає максимальному наближенню до безупорядкованого білого шуму, який має бути видалений з досліджуваних даних. Структурну блок-схему системи вейвлет-фільтрації профілів експресій генів на основі критерію ентропія Шеннона з використанням ймовірнісної оцінки Джеймса та Стейна подано на рис. 2.

Рис. 2

Реалізація даного процесу передбачає наступні кроки:

1. *Вибір* материнського вейвлету зі списку доступних для типу досліджуваних даних.

2. *Визначення* оптимального рівня вейвлет-декомпозиції сигналу на основі максимального

значення ентропії Шеннона, що розраховується для видаленої шумової компоненти. На цьому етапі вибір типу вейвлету з сімейства материнського вейвлету та значення трешолдингового коефіцієнта встановлюються випадково з інтервалу допустимих значень.

3. *Визначення* типу вейвлета з сімейства материнського вейвлету на основі максимального значення ентропії Шеннона для видаленої шумової компоненти.

4. *Визначення* оптимального значення трешолдингового коефіцієнта на основі мінімального значення ентропії Шеннона, що розраховується для фільтрованого сигналу.

Отже, процес оптимізації параметрів вейвлет-фільтру профілів експресій генів передбачає паралельну оцінку ентропії Шеннона для фільтрованих даних та для видаленої шумової компоненти.

Експерименти та результати

Моделювання процесу вейвлет-фільтрації було проведено з використанням профілів експресій генів хворого на рак легенів, отриманих шляхом ДНК-мікрочіпових експериментів [19] та передобробленого за методикою [16]. Діаграма розподілу експресій генів досліджуваного об'єкта представлено на рис. 3.

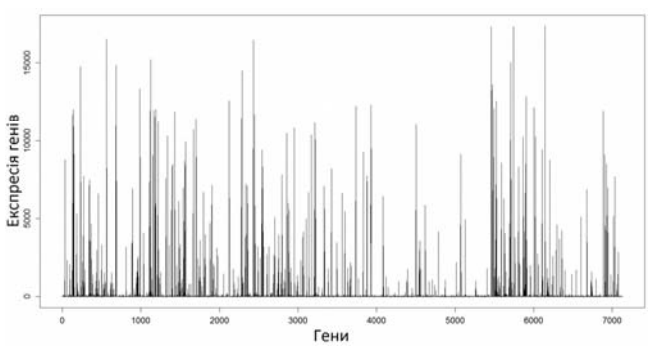


Рис. 3

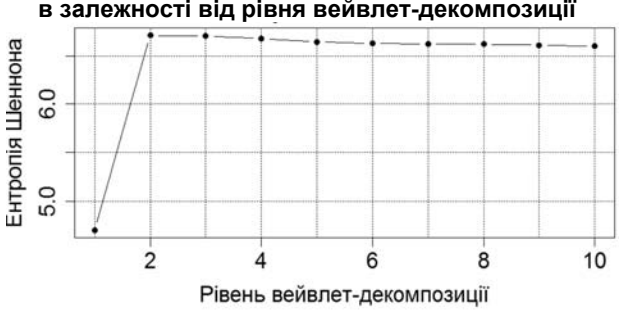
Статистичні характеристики вектора профілю експресій досліджуваних генів представлено у таблиці.

Minimum	1 *Quantile*	*Median*	*Mean*	3 *Quantile*	*Maximum*
–36,19	2,18	11,10	236,91	22,63	17360,00

Аналіз даних таблиці та рис. 3 дозволяє дійти висновку, що вектор експресій генів, використаних для моделювання процесу вейвлет-фільтрації, містить 7129 генів, експресія яких змінюється від –36,19 до 17360, при цьому слід зазначити, що більшість генів мають низьке значення експресії. Більш того, з великою ймовірністю можна стверджувати, що шумова компонента міститься у високочастотній частині спектра, що обґрунтовує використання вейвлет-аналізу для очистки даних від шуму.

В процесі моделювання досліджено ортогональні вейвлети Добеші (*db*1, *db*2,…, *db*45), симплети (*sym*2, *sym*3,…, *sym*30), койфлети (*coif*1, *coif*2,…, *coif*5), біортогональні вейвлети (*bior*1.1, *bior*1.3, *bior*1.5, *bior*2.2, *bior*2.4, *bior*2.6, *bior*2.8, *bior*3.1, *bior*3.3, *bior*3.5, *bior*3.7, *bior*4.4, *bior*5.5, *bior*6.8), та обернені біортогональні вейвлети (*rbio*1.1, *rbio*1.3, *rbio*1.5, *rbio*2.2, *rbio*2.4, *rbio*2.6, *rbio*2.8, *rbio*3.1, r*bio*3.3, *rbio*3.5, *rbio*3.7, *rbio*4.4, *rbio*5.5, *rbio*6.8). Експериментальне визначення порогового значення трешолдингового коефіцієнта проведено двома способами. У першому випадку проводилася покрокова обробка деталізуючих коефіцієнтів у відповідності з формулою (14), коли значення трешолдингового коефіцієнту було досить малим (0,2) і не змінювалося у процесі моделювання. Тривалість експерименту обмежувалася кількістю кроків обробки деталізуючих коефіцієнтів. Другий випадок передбачав покрокове збільшення значення трешолдингового коефіцієнту від $\tau_{\min}$ до $\tau_{\max}$ з кроком $d\tau$. На рис. 4 показано результати моделювання з використанням вейвлетів Добеші. У відповідності зі схемою (див. рис. 2), вибір типу вейвлету та визначення рівня вейвлет-декомпозиції проводилися на основі максимального значення ентропії Шеннона, розрахованої для видаленої шумової компоненти.

а) Ентропія Шеннона шумової компоненти в залежності від рівня вейвлет-декомпозиції



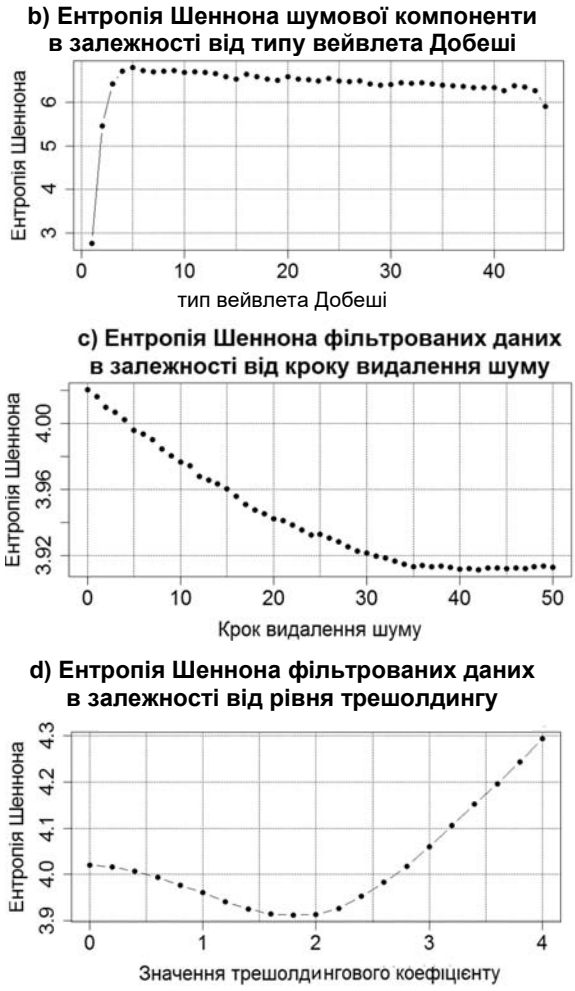


Рис. 4

Визначення трешолдингового коефіцієнта проведено на основі мінімального значення ентропії Шеннона для фільтрованих даних.

Аналіз отриманих діаграм дозволяє зробити висновок, що при визначенні максимального значення ентропії Шеннона виділеної шумової компоненти оптимальним є використання вейвлета *db*5 (див. рис. 4, *b*), при другому рівні вейвлет-декомпозиції даних (див. рис. 4, *a*). Методика покрокового видалення шумової компоненти при сталому значенні трешолдингового коефіцієнта (див. рис. 4, *c*) не є ефективною, оскільки вона не дає можливості однозначного визначення кроку зупинки роботи алгоритму. Для визначення оптимального трешолдингу є ефективною методика покрокового збільшення значення коефіцієнта трешолдингу, оскільки у цьому випадку спостерігається яскраво виражений мінімум значення ентропії Шеннона фільтрованих даних, який відповідає значенню коефіцієнта трешолдингу $\tau = 1{,}8$ (див. рис. 4, *d*). Фільтровані дані експресій генів та виділена шумова компонента при використанні вейвлету Добеші *db*5, другому рівні вейвлет-декомпозиції та значенні коефіцієнта трешолдингу $\tau = 1{,}8$ подано на рис. 5.

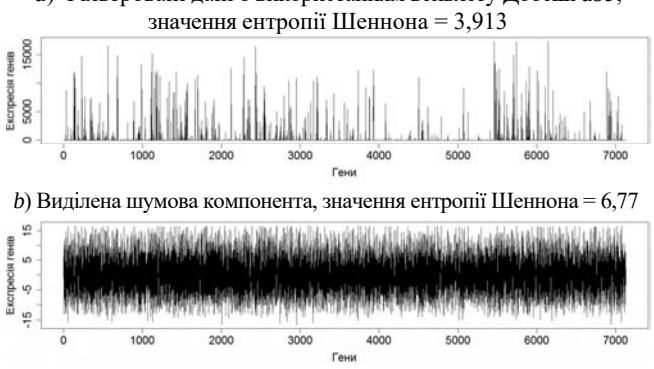
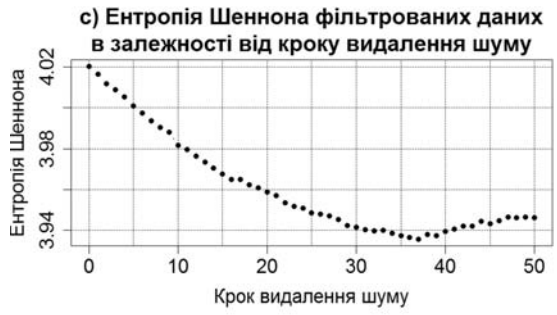


Рис. 5

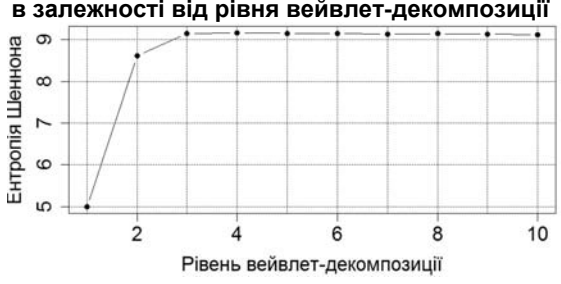
Аналогічні результати при використанні койфлетів, симплетів, біортогональних та обернених біортогональних вейвлетів показано на рис. 6–9. Аналіз результатів дозволяє зробити висновок, що оптимальними за критерієм ентропія Шеннона є використання наступних параметрів вейвлет-фільтру: вейвлет Добеші *db*5 на другому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 1,8; койфлет *coif*4 на другому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 1,6; симплет *sym*5 на третьому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 1,8; біортогональний вейвлет *bior*1.5 на третьому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 2,2 і обернений біортогональний вейвлет *rbio*1.5 на четвертому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 1,8.

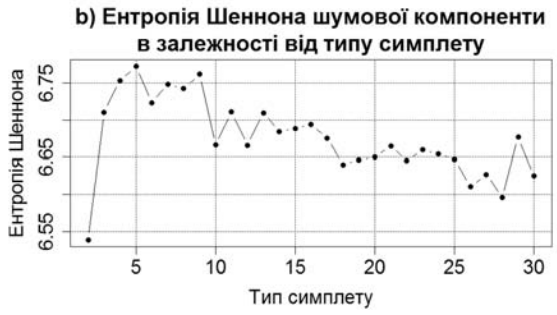

b) Ентропія Шеннона шумової компоненти
в залежності від типу койфлету
Ентропія Шеннона
6.3
6.5
6.7
1
2
3
4
5
Тип койфлету

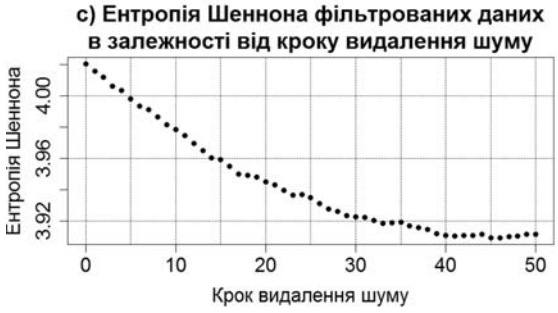

c) Ентропія Шеннона фільтрованих даних
в залежності від кроку видалення шуму
Ентропія Шеннона
3.94
3.98
4.02
0
10
20
30
40
50
Крок видалення шуму


d) Ентропія Шеннона фільтрованих даних
в залежності від рівня трешолдингу
Ентропія Шеннона
4.0
4.2
4.4
0
1
2
3
4
Значення трешолдингового коефіцієнту

Рис. 6


a) Ентропія Шеннона шумової компоненти
в залежності від рівня вейвлет-декомпозиції
Ентропія Шеннона
5
6
7
8
9
2
4
6
8
10
Рівень вейвлет-декомпозиції


b) Ентропія Шеннона шумової компоненти
в залежності від типу симплету
Ентропія Шеннона
6.55
6.65
6.75
5
10
15
20
25
30
Тип симплету


c) Ентропія Шеннона фільтрованих даних
в залежності від кроку видалення шуму
Ентропія Шеннона
3.92
3.96
4.00
0
10
20
30
40
50
Крок видалення шуму


d) Ентропія Шеннона фільтрованих даних
в залежності від рівня трешолдингу
Ентропія Шеннона
3.9
4.0
4.1
4.2
4.3
0
1
2
3
4
Значення трешолдингового коефіцієнту

Рис. 7


a) Ентропія Шеннона шумової компоненти
в залежності від рівня вейвлет-декомпозиції
Ентропія Шеннона
5.0
5.5
6.0
6.5
2
4
6
8
10
Рівень вейвлет-декомпозиції


b) Ентропія Шеннона шумової компоненти в
залежності від типу біортогонального вейвлету
Ентропія Шеннона
4.0
5.0
6.0
2
4
6
8
10
12
14
Тип біортогонального вейвлету

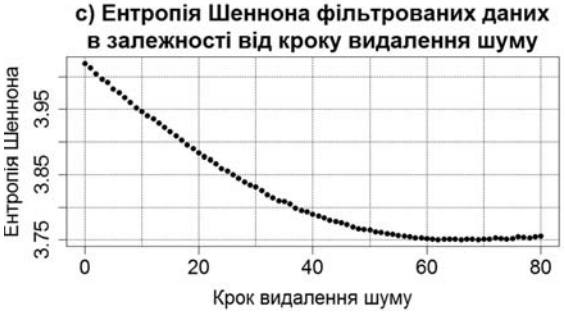

c) Ентропія Шеннона фільтрованих даних
в залежності від кроку видалення шуму
Ентропія Шеннона
3.75
3.85
3.95
0
20
40
60
80
Крок видалення шуму

Рис. 8

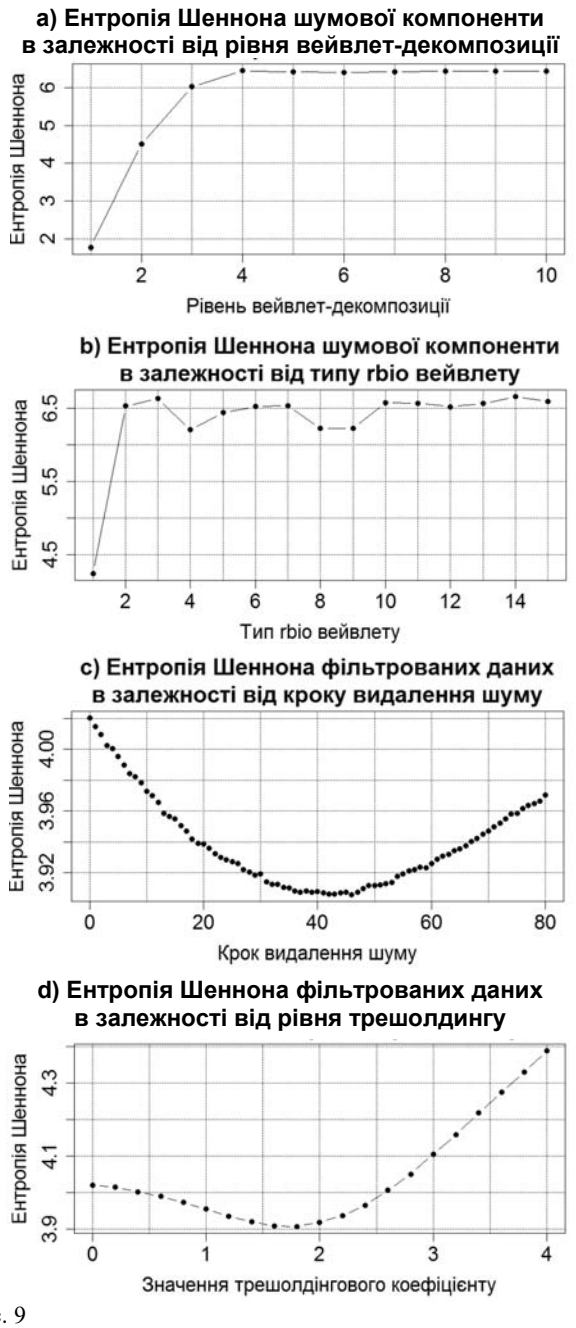


Рис. 9

На рис. 10 представлено діаграму відношення ентропій фільтрованих даних та виділеного шуму в залежності від використаного вейвлету. Аналіз рис. 10 дозволяє зробити висновок, що вибір типу материнського вейвлету з сім'ї ортогональних та біортогональних вейвлетів у випадку фільтрації профілів експресій генів не є визначальним. З урахуванням відношення ентропій Шеннона для фільтрованих даних та виділеної шумової компоненти кращі результаті щодо вейвлет-фільтрації отримуються з використанням біортогонального вейвлету *bior*1.5. Але різниця між результатами, отриманими з використанням інших вейвлетів досить мала. Визначальними у даному випадку є вибір типу вейвлету з сім'ї використаного материнського вейвлету, вибір рівня вейвлет-декомпозиції та визначення оптимального значення коефіцієнта трешолдингу для обробки деталізуючих коефіцієнтів.

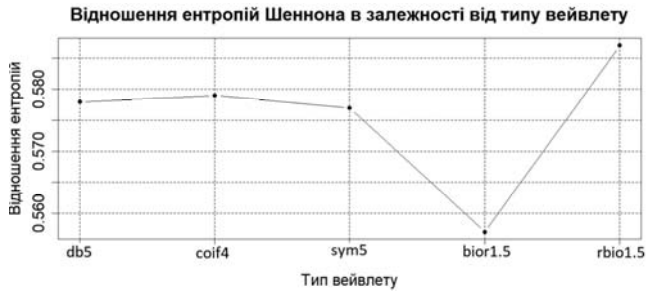


Рис. 10

Дослідження дозволяють розробити технологію визначення оптимальних параметрів вейвлет-фільтру для обробки профілів експресій генів у вигляді структурної блок-схеми покрокової обробки інформації. Архітектуру даної технології показано на рис. 11. Практична реалізація представленої технології передбачає наступні етапи:

***Eman* I**. Ініціалізація вихідних параметрів вейвлет-фільтру.

1. *Формування вектора материнських вейвлетів* та векторів вейвлетів у межах виділених материнських вейвлетів:

$$wv = \{wv_i\}, \ i = 1,...,k, \qquad wv_i = \{wv_i^j\}, \ j = 1,...,p, \tag{15}$$

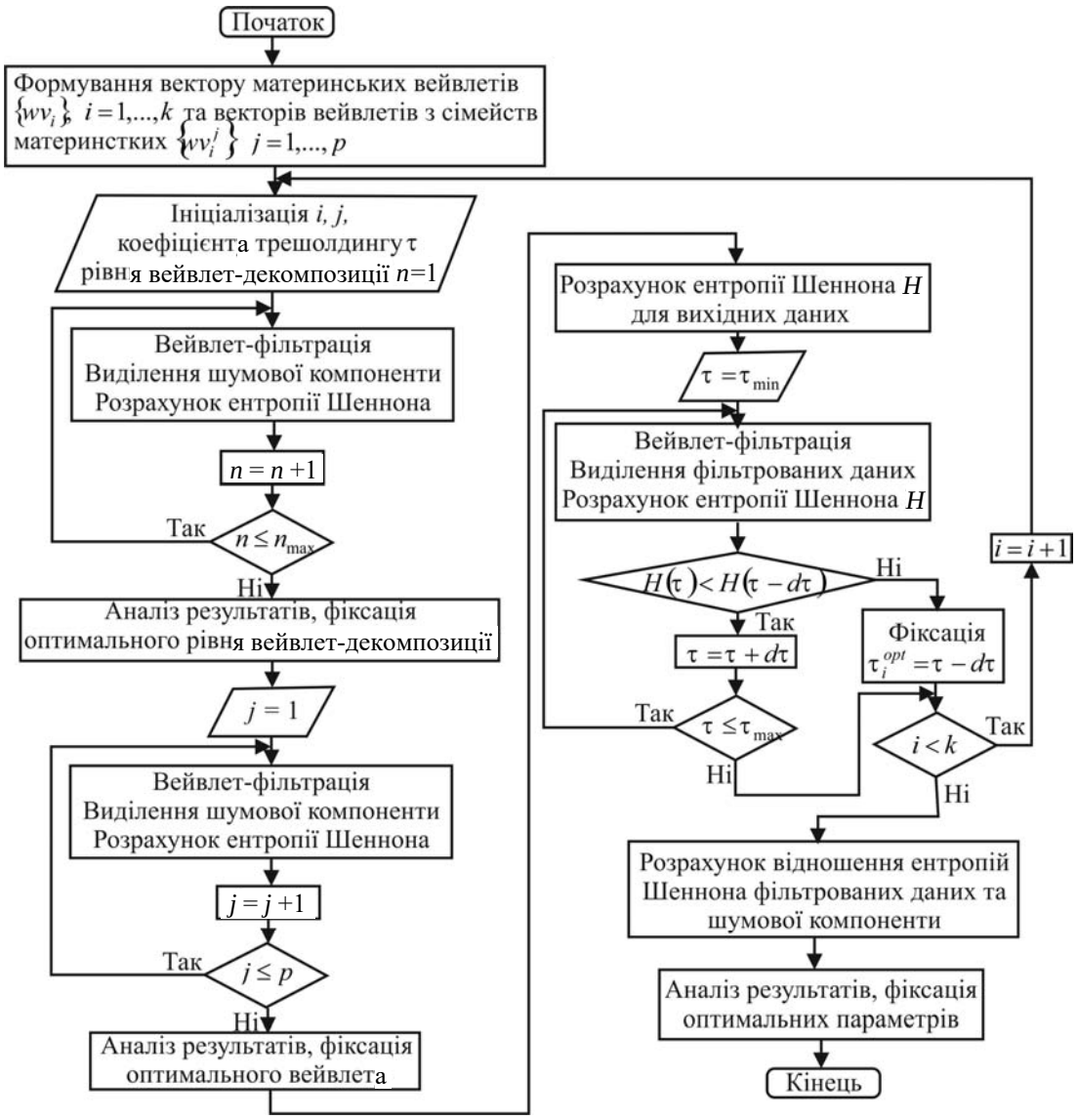


Рис. 11

де k – кількість материнських вейвлетів, що досліджуються у процесі моделювання, p – кількість типів вейвлетів материнського вейвлету i. Визначення інтервалу та кроку зміни коефіцієнта трешолдингу $\tau_{min}, \tau_{max}, d\tau = \tau_{min}$, вибір максимального рівня вейвлет-декомпозиції.

2. *Вибір материнського вейвлету*, що відповідає першому порядковому номеру вектора материнських вейвлетів ($i = 1$), довільний вибір типу вейвлету даного материнського вейвлету. Встановлення коефіцієнта трешолдингу $\tau = \tau_{min}$ та рівня вейвлет-декомпозиції $n = 1$.

***Етап* II**. Визначення оптимального рівня вейвлет декомпозиції.

3. *Вейвлет-фільтрація вектора* профілів експресій генів відповідно до схеми, зображеної на рис. 2 у межах встановленого інтервалу зміни рівня вейвлет-декомпозиції даних, виділення шумової компоненти на кожному її рівні та розрахунок ентропії Шеннона шумової компоненти на кожному кроці.

4. *Аналіз отриманих результатів*. Фіксація оптимального рівня вейвлет-декомпозиції для даного материнського вейвлету n_i^{opt}, який відповідає максимальному значенню критерію ентропія Шеннона.

***Етап* III**. Визначення типу вейвлета з сім'ї відповідного материнського вейвлета.

5. *Вейвлет-фільтрація вектора* профілів експресій генів для усіх p типів даного материнського вейвлета $\left\{wv_i^j\right\}$, $j=1,...,p$. Виділення шумової компоненти, відповідної кожному типу вейвлета та розрахунок ентропії Шеннона шумової компоненти на кожному кроці.

6. *Аналіз результатів*. Фіксація оптимального такого типу вейвлета материнського вейвлета, який відповідає максимальному значенню критерія ентропія Шеннона.

***Етап* IV**. Визначення оптимального коефіцієнта трешолдінгу для обробки деталізуючих коефіцієнтів.

7. *Розрахунок* ентропії Шеннона для вектора вихідних даних $H(\tau - d\tau)$.

8. *Вейвлет-фільтрація* вектора профілів експресій генів при використанні коефіцієнта трешолдінгу τ. Виділення вектора фільтрованих даних.

9. *Розрахунок ентропії* Шеннона на даному кроці обробки інформації $H(\tau)$.

10. Якщо $H(\tau) < H(\tau - d\tau)$, збільшення коефіцієнта трешолдінгу на $d\tau$ та перехід на крок 8 даного алгоритму. У протилежному випадку – фіксація оптимального коефіцієнта трешолдінгу для i-го материнського вейвлету: $\tau_i^{\text{opt}} = \tau - d\tau$.

11. *Операція інкремента* параметра i ($i = i+1$) та за умови $i \le k$ повторення етапів II–IV даного алгоритму.

***Етап* V**. Формування остаточного рішення щодо вибору параметрів вейвлет-фільтру.

12. *Розрахунок відношень* ентропій Шеннона для фільтрованих даних та виділеної шумової компоненти для кожного материнського вейвлету з використанням оптимальних параметрів вейвлет-фільтру.

13. *Фіксація оптимальних параметрів* вейвлет-фільтра, відповідних глобальному мінімуму критерію відношення ентропій Шеннона фільтрованих даних та виділеної шумової компоненти.

Висновки. Представлено технологію вейвлет фільтрації профілів експресій генів з метою видалення фонового білого шуму. Як основний критерій оцінки інформативності досліджуваного сигналу використано ентропію Шеннона на основі методу оцінки ймовірності Джеймса та Стейна, заснованого на комплексному використанні двох моделей даних: високорозмірної з малим зміщенням та високою дисперсією розподілу даних та низькорозмірної з високим зміщенням та низькою дисперсією. Структурна блок-схема системи вейвлет-фільтрації у запропонованій моделі передбачає оцінку ентропії як фільтрованого сигналу, так і видаленої шумової компоненти, коли остаточне рішення щодо вибору параметрів вейвлет-фільтра приймається на основі відносного критерію, розрахованого як відношення ентропій Шеннона фільтрованого сигналу та видаленої шумової компоненти. У процесі моделювання досліджено сім'ї ортогональних вейвлетів Добеші, симплети, койфлети, біортогональні та обернені біортогональні вейвлети. Проведено дослідження з оптимізації процесу вибору типу вейвлета, рівня вейвлет-декомпозиції та значення трешолдінгового коефіцієнта. Експериментальне визначення порогового значення трешолдингового коефіцієнта проведено у два способи. У першому випадку проведено покрокову обробку деталізуючих коефіцієнтів, коли значення трешолдингового коефіцієнта досить мале (0,2) і не змінювалося у процесі моделювання. Тривалість експерименту обмежувалася кількістю кроків обробки деталізуючих коефіцієнтів. Другий випадок передбачав покрокове збільшення значення трешолдингового коефіцієнта від $\tau_{\min}$ до $\tau_{\max}$ з кроком $d\tau$. Результати моделювання подано у вигляді графіків залежності ентропії Шеннона від відповідного параметра, максимальне або мінімальне значення яких дозволяє прийняти об'єктивне рішення у виборі відповідного параметра.

Аналіз результатів показав, що оптимальними за критерієм ентропії Шеннона є використання наступних параметрів вейвлет-фільтра: вейвлет Добеші *db*5 на другому рівні вейвлет-

декомпозиції та коефіцієнті трешолдингу 1,8; койфлет *coif*4 на другому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 1,6; симплет *sym*5 на третьому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 1,8; біортогональний вейвлет *bior*1.5 на третьому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 2,2 та обернений біортогональний вейвлет *rbio*1.5 на четвертому рівні вейвлет-декомпозиції та коефіцієнті трешолдингу 1,8. Аналіз діаграми залежності відносної ентропії від типу вейвлету дозволяє зробити висновок, що вибір типу материнського вейвлету сім'ї ортогональних та біортогональних вейвлетів у випадку фільтрації профілів експресій генів не є визначальним. З урахуванням відношення ентропій Шеннона для фільтрованих даних та виділеної шумової компоненти кращі результаті щодо вейвлет-фільтрації отримано за використання біортогонального вейвлету *bior*1.5. Але різниця між результатами, отриманими за використання інших вейвлетів, досить мала. Визначальними у даному випадку є вибір типу вейвлету з використаної сім'ї материнського вейвлету, вибір рівня вейвлет-декомпозиції та визначення оптимального значення коефіцієнта трешолдингу для обробки деталізуючих коефіцієнтів. На основі досліджень запропоновано технологію визначення оптимальних параметрів вейвлет-фільтра для обробки профілів експресій генів у вигляді структурної блок-схеми покрокової обробки інформації.

Перспективами досліджень автора є практична реалізація запропонованої технології у межах гібридної моделі передобробки профілів експресій генів з метою подальшої реконструкції генної регуляторної мережі.

1. *Schermer M.J.* DNA microarrays: a practical approach // Oxford University Press. – 1999. – P. 17–42.
2. *Microarray* biochip technology / T. Basarsky, D. Verdnik, J.Y. Zhai et al. – Eaton Publ. – 2000. – P. 265–284.
3. *Daubechies I.* The wavelet transform, time-frequency localization and signal analysis // IEEE Trans. Inform. Theory, 1990. – **36**. – P. 961–1005.
4. *Daubechies I.* Ten lectures on wavelets // CBMS–NSF conf. series in applied math. SLAM Ed., 1992. – 343 p.
5. *Coifman R.R., Meyer Y., Wickerhauser M.V.* Wavelet Analysis and Signal Processing // Wavelets and Their Applications. – Boston Jones and Bartlett, 1992. – P. 153–178.
6. *Antoshchuk S.G.* Realizaciya vejvletnogo preobrazovaniya pri strukturnom analize izobrazhenij. Elektromashinobuduvannya ta elektroobladnannya, 2004. – **62**. – P. 153–157. (In Russian).
7. *Benedetto J.J.* Wavelets: Mathematics and Application // CRC Press, Series: Studies in Advanced Mathematics, 1993. – 592 p.
8. *Samsul A., Karim A., Mohd T.I.* Compression of Chemical Signal Using Wavelet Transform // European J. of Scientific Research, 2009. – **36**(4). – P. 513–520.
9. *Joshi A., Aravind H.S.* Analysis of Adaptive Wavelet Wiener Filtering for ECG Signals: Review // Int. J. of Advanced Research in Electronics and Communication Engineering, 2014. – **3**. – Issue 4. – P. 395–398.
10. *Chandu R., Venkateswarlu M.* ECG Signal Filtering using an Improved Wavelet Wiener Filtering // Int. J. of Advanced Technology and Innovative Research, 2015. – **7**. – Issue 7. – P. 1242–1247.
11. *Baldi P., Hatfield G.W.* DNA Microarrays and gene expression: From experiments to data analysis modeling // Cambridge University Press. – 2002. – P. 22–23.
12. *A Comparison* of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias / B.M. Bolstad, R.A. Irizarry, M. Astrand et al. // Bioinformatics, 2003. – **19**. – P. 185–193.
13. *Exploration*, normalization, and summaries of high density oligonucleotide array probe level data / R.A. Irizarry, B. Hobbs, F. Collin et al. // Biostatistics. – 2003. – **4**, № 2. – P. 249–264.
14. *Chen Z., McGee M., Liu Q.* Distribution-Free Convolution Model for background correction of oligonucleotide microarray data // BMC genomics, 2009. – **10**. – P. 1–19.
15. *Computational* analysis of gene expression profiles of lung cancer / S. Babichev, A. Kornelyuk, V. Lytvynenko et al. // Biopolymers and Cells. – 2016. – **32**(1). – P. 70–79.
16. *Babichev S., Taif M.A., Lytvynenko V.* Filtration of DNA nucleotide gene expression profiles in the systems of biological objects clustering // Int. Frontier Science Letters. – 2016. – **8**. – P. 1–8.
17. *Fil'traciya* hromatogramm s pomoshch'yu vejvlet-analiza s ispol'zovaniem kriteriya ehntropii / S.A. Babichev, A.A. Didyk, V.I. Litvinenko et al. // System technologies. – 2010. – N **6**(71). – P. 117–131. (In Russian).
18. *Hausser J., Strimmer K.* Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks // J. of Machine Learning Research. – 2009. – **10**. – P. 1469–1484.

19. *Gene-expression* profiles predict survival of patients with lung adenocarcinoma / D.G. Beer, S.L. Kardia, C. Huang at al. // Nature Medicine. – 2002. – **8**(8). – P. 816–824.

Поступила 19.09.2017
E-mail: *sergii.babichev@ujep.cz*

С.А. Бабичев

Технология вейвлет-фильтрации профилей экспрессий генов с целью удаления фонового шума

Введение. Детальный анализ процесса формирования матрицы экспрессий генов, полученных посредством микрочиповых экспериментов [1, 2] показывает, что этап сканирования изображения микрочипа ДНК сопровождается возникновением фонового шума. Частичная фоновая коррекция проводится на этапе формирования матрицы экспрессии генов, но следует отметить, что полное удаление фоновой шумовой составляющей на данном этапе проблематично. Вследствие этого возникает необходимость дальнейшей фильтрации полученных профилей экспрессии генов с использованием количественных критериев оценки информативности обрабатываемых профилей и шумовой составляющей, выделяемой из соответствующих профилей экспрессий генов. В данной статье задача решается с использованием вейвлет-анализа [3–5], получившим широкое распространение в различных областях современной науки и техники [6–10].

Анализ современных достижений и публикаций

Вопросам создания матрицы экспрессий генов, полученных посредством ДНК-микрочиповых экспериментов, посвящены работы [11–14], где подробно рассматриваются этапы обработки экспериментальных данных, предполагающих фоновую коррекцию, нормализацию, PM-коррекцию и суммаризацию. На каждом этапе предусмотрена возможность использования различных методов со своими преимуществами и недостатками. В [15] представлен сравнительный анализ методов обработки профилей экспрессий генов больных раком легких, полученных посредством ДНК-микрочиповых экспериментов, определена оптимальная комбинация методов с использованием критерия энтропия Шеннона. Вопросам редукции неинформативных профилей экспрессий генов с учетом статистических и энтропийных критериев посвящена работа [16]. Представленная технология пошаговой обработки профилей экспрессии генов позволяет сократить количество признаков на 6–10 процентов, что повышает информативность данных для дальнейших исследований. Однако следует отметить, что на этапе фоновой коррекции проблеме удаления фонового белого шума, возникающего на этапе сканирования изображения микрочипа, уделено недостаточное внимание. Методика фильтрации хроматограмм с использованием вейвлет-анализа и критерия энтропия Шеннона предложена в [17]. Но применительно к профилям экспрессий генов данная задача в настоящее время однозначного решения не имеет.

Цель статьи – разработка технологии вейвлет-фильтрации профилей экспрессий генов для удаления фонового белого шума, оценка качества обработки данных в которой осуществляется с применением критерия энтропия Шеннона, рассчитанного на основе метода расширения Джеймса и Стейна.

Основной материал

К вейвлетам относятся функции, которые строятся на основе одного материнского вейвлета $\psi(t)$ посредством операций сдвига по аргументу τ и масштабирования по параметру a [4]:

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-\tau}{a}\right), \qquad (1)$$

где $(a,\tau) \in R$, $\psi(t) \in L^2(R)$, а множитель $\frac{1}{\sqrt{|a|}}$ обеспечивает независимость нормы функции от масштабного числа a. Материнский вейвлет $\psi(t)$ обладает следующими свойствами:

• *ограниченность*, т.е. вейвлет функции $\psi(t)$ должны иметь конечную энергию:

$$\|\psi(t)\| = \int_{-\infty}^{+\infty} |\psi(t)|^2 \, dt < \infty; \qquad (2)$$

• *локализация*, т.е. они должны быть определены на конечном интервале как по времени, так и по частоте. Для этого достаточно, чтобы при C = const и $\varepsilon > 0$ выполнялись следующие условия:

$$|\psi(t)| \le C\left(1+|t|\right)^{-1-\varepsilon} \text{ и } |\psi(\omega)| \le C\left(1+|\omega|\right)^{-1-\varepsilon}, \qquad (3)$$

где ω – средняя частота вейвлета;

• *нулевое* среднее значение, т.е. выполнение следующего условия для нулевого момента:

$$\int_{-\infty}^{+\infty} \psi(t)\, dt = 0. \qquad (4)$$

В случае непрерывного вейвлет-преобразования сигнала $s(t) \in L^2(R)$ параметры a и τ принимают любые значения в пределах области их определения и вейвлетный масштабно-временной спектр $C(a,b)$ рассчитывается в соответствии с формулой

$$C(a,b) = \int_{-\infty}^{+\infty} s(t) a^{-1/2} \psi\left(\frac{t-b}{a}\right) dt. \qquad (5)$$

Вейвлет-реконструкция сигнала $s(t)$ при условии использования такого же базиса функций выполняется так:

$$s(t) = \frac{1}{C_\psi} \int_{R^+} \int_R C(a,\tau) a^{-1/2} \psi\left(\frac{t-\tau}{a}\right) \frac{da \cdot d\tau}{a^2}, \qquad (6)$$

где C_ψ – константа, определяемая функцией ψ:

$$C_\psi = 2\pi \int_{-\infty}^{\infty} \left|\hat{\psi}(\omega)\right|^2 |\omega| d\omega, \qquad (7)$$

где $\hat{\psi}$ – фурье-образ функции ψ. Для количественного анализа сигналов, предусматривающего его декомпозицию и реконструкцию, возможно использование любых локализованных функций $\psi(t)$, если для них существуют такие функции-двойники $\psi'(t)$, что семейства $\{\psi_{a,\tau}(t)\}$ и $\{\psi'_{a,\tau}(t)\}$ образуют парные базисы функционального пространства $L^2(R)$. Если вейвлет $\psi(t)$ ортогональный, то $\psi(t) \equiv \psi'(t)$ и вейвлетный базис также имеет свойства ортогональности. В случае неортогональности вейвлета $\psi(t)$ наличие двойника $\psi'(t)$ дает возможность сформировать семейства $\{\psi_{m,k}(t)\}$ и $\{\psi'_{g,p}(t)\}$, удовлетворяющие условию биортогональности на множестве целых чисел Z:

$$\left\langle \psi_{m,k}(t), \psi'_{g,p}(t) \right\rangle = \delta_{m,k}\delta_{g,p}, \; m,k,g,p \in Z. \qquad (8)$$

В этом случае возможна декомпозиция и реконструкция любого сигнала. При дискретной вейвлет-декомпозиции сигнала параметры a и τ принимают дискретные значения на множестве

$$a = 2^j \text{ и } \tau = k2^j, \qquad (9)$$

где j и k – целые числа, коэффициент j – параметр масштаба, а k определяет уровень вейвлет-декомпозиции сигнала. При дискретных значениях a и τ вейвлет-функция приобретает вид

$$\psi_{j,k}(t) = a_0^{-j/2} \psi\left(a_0^{-j} t - k\right), \qquad (10)$$

а прямое дискретное вейвлет-преобразование сводится к вычислению детализирующих коэффициентов

$$d_{j,k} = \int_{-\infty}^{+\infty} a_0^{-j/2} \psi\left(a_0^{-j} t - k\right) s(t) dt. \qquad (11)$$

Обратное дискретное вейвлет-преобразование задается тем же базисом, что и прямое:

$$s(t) = \frac{1}{C_\psi} \sum_{j \in Z} \sum_{k \in Z} d_{j,k} a_0^{-j/2} \psi\left(a_0^{-j} t - k\right). \qquad (12)$$

Структурная схема процесса вейвлет-фильтрации исследуемых данных представлена на рис. 1.

Процесс вейвлет-декомпозиции одномерного сигнала предполагает расчет вектора аппроксимирующих коэффициентов на N-уровне вейвлет-декомпозиции и векторов детализирующих коэффициентов на уровнях от единицы до N:

$$s = CA_0 \to \{CA_1, CD_1\} \to \{CA_2, CD_2, CD_1\} \to \ldots \\ \ldots \to \{CA_N, CD_N, CD_{N-1}, \ldots, CD_1\}. \qquad (13)$$

Векторы аппроксимирующих и детализирующих коэффициентов рассчитываются путем использования для вектора соответствующих данных фильтра низких частот *LFF* для аппроксимации и фильтра высоких частот *HFF* для детализации. Аппроксимирующие коэффициенты содержат информацию о грубой составляющей сигнала. Информация о деталях сигнала и высокочастотная шумовая составляющая содержатся в большинстве случаев в детализирующих коэффициентах. Удаление шумовой составляющей из профилей экспрессии генов в рамках предложенной модели проводилось с использованием мягкого трешолдинга так:

$$\begin{cases} d = 0, \text{ если } d \le \tau, \\ d = d - \tau, \text{ если } d > \tau, \end{cases} \qquad (14)$$

где τ – значение трешолдингового коэффициента, d – значение детализирующих коэффициентов на всех уровнях вейвлет-декомпозиции вектора исследуемых данных. Реконструкция сигнала проведена на основе аппроксимирующих коэффициентов на N-уровне вейвлет-декомпозиции и обработанных детализирующих коэффициентов на уровнях от единицы до N.

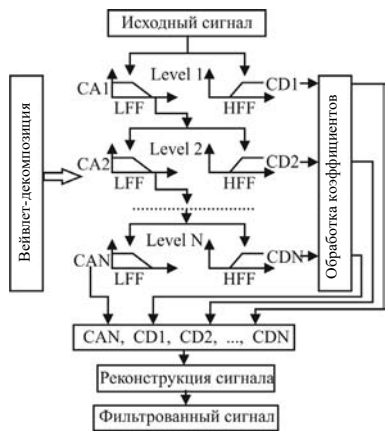


Рис. 1

Детальный анализ рис. 1 позволяет определить пути оптимизации процесса вейвлет-фильтрации профилей экспрессии генов. Данный процесс предусматривает следующие шаги:

- *выбор* материнского вейвлета;
- *определение* оптимального уровня вейвлет-декомпозиции вектора исследуемых данных;
- *выбор* типа вейвлета из семейства материнского вейвлета;
- *определение* оптимального значения трешолдингового коэффициента.

Оценка качества обработки информации на каждом шагу проводилась с использованием энтропии Шеннона на основе оценки Джеймса и Стейна [18]. Очевидно, что минимальное значение критерия энтропия Шеннона, рассчитанное для профиля экспрессии гена, соответствует наибо-

лее высокому качеству обработки информации. С другой стороны, максимальное значение энтропии Шеннона, рассчитанное для выделенной шумовой компоненты, соответствует максимальному приближению к хаотическому белому шуму, который должен быть удален из исследуемых данных. Структурная блок-схема системы вейвлет-фильтрации профилей экспрессии генов на основе критерия энтропия Шеннона с использованием вероятностной оценки Джеймса и Стейна представлена на рис. 2.

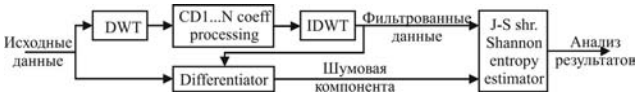


Рис. 2

Реализация данного процесса предусматривает следующие шаги:

1. *Выбор* материнского вейвлета из списка доступных для данного типа исследуемых данных.

2. *Определение* оптимального уровня вейвлет-декомпозиции сигнала на основе максимального значения энтропии Шеннона, рассчитываемой для выделенной шумовой компоненты. На этом этапе выбор типа вейвлета из семейства материнского вейвлета и значение трешолдингового коэффициента устанавливаются произвольно из интервала допустимых значений.

3. *Определение* типа вейвлета из семейства материнского вейвлета на основе максимального значения энтропии Шеннона для выделенной шумовой компоненты.

4. *Определение* оптимального значения трешолдингового коэффициента на основе минимального значения энтропии Шеннона, рассчитываемой для фильтрованного сигнала.

Таким образом, процесс оптимизации параметров вейвлет-фильтра профилей экспрессии генов предусматривает параллельную оценку энтропии Шеннона для фильтруемых данных и для выделенной шумовой компоненты.

Эксперименты и результаты

Моделирование процесса вейвлет-фильтрации было проведено с использованием профиля экспрессии генов больного раком легких, полученного посредством ДНК-микрочиповых экспериментов [19] и предобработанного по методике [16]. Диаграмма распределения экспрессий генов исследуемого объекта представлена на рис. 3.

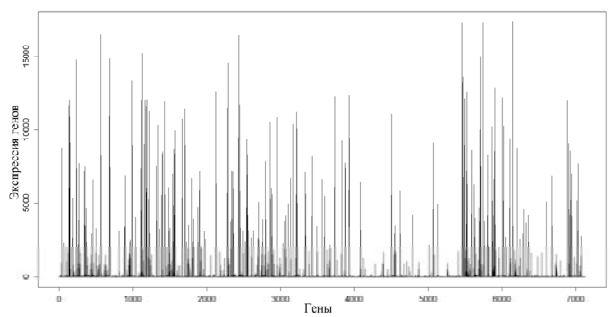


Рис. 3

Статистические характеристики вектора экспрессий исследуемых генов представлены в таблице.

Minimum	25% *Quartile*	*Median*	*Mean*	75% *Quartile*	*Maximum*
–36,19	2,18	11,10	236,91	22,63	17360,00

Анализ данных таблицы и рис. 3 позволяет сделать вывод, что вектор экспрессий генов, использованный в процессе моделирования, содержит 7129 генов, экспрессия которых меняется от –36,19 до 17360, при этом следует отметить, что большинство генов имеют низкое значение экспрессии. Более того, с большой вероятностью можно утверждать, что шумовая компонента содержится в высокочастотной части спектра. Данный факт обосновывает использование вейвлет-анализа для очистки данных от шума.

В процессе моделирования исследовались: ортогональные вейвлеты Добеши (*db*1, *db*2,…, *db*45), симплеты (*sym*2, *sym*3, …, *sym*30), койфлеты (*coif*1, *coif*2, …, *coif*5), биортогональные вейвлеты (*bior*1.1, *bior*1.3, *bior*1.5, *bior*2.2, *bior*2.4, *bior*2.6, *bior*2.8, *bior*3.1, *bior*3.3, *bior*3.5, *bior*3.7, *bior*4.4, *bior*5.5, *bior*6.8), и обратные биортогональные вейвлеты (*rbio*1.1, *rbio*1.3, *rbio*1.5, *rbio*2.2, *rbio*2.4, *rbio*2.6, *rbio*2.8, *rbio*3.1, *rbio*3.3, *rbio*3.5, *rbio*3.7, *rbio*4.4, *rbio*5.5, *rbio*6.8). Экспериментальное определение порогового значения трешолдингового коэффициента было проведено двумя способами. В первом случае проводилась пошаговая обработка детализирующих коэффициентов в соответствии с (14), при этом значение трешолдингового коэффициента было достаточно малым (0,2) и не менялось в процессе моделирования. Продолжительность эксперимента ограничивалась количеством шагов обработки детализирующих коэффициентов. Второй случай предусматривал пошаговое увеличение значения трешолдингового коэффициента от $\tau_{\min}$ до $\tau_{\max}$ с шагом $d\tau$. На рис. 4 представлены результаты моделирования при использовании вейвлетов Добеши. В соответствии со схемой, изображенной на рис. 2, выбор типа вейвлета и определение уровня вейвлет-декомпозиции проводилось на основе максимального значения энтропии Шеннона, рассчитанной для выделенной шумовой компоненты. Определение трешолдингового коэффициента проводилось на основе минимального значения энтропии Шеннона для фильтрованных данных. Анализ полученных диаграмм позволяет сделать вывод, что с учетом максимального значения энтропии Шеннона выделенной шумовой компоненты оптимально использование вейвлета *db*5 (см. рис. 4, *b*) при втором уровне вейвлет-декомпозиции данных (см. рис. 4, *a*). Методика пошагового удаления шумовой компоненты при постоянном значении трешолдингового коэффициента (см. рис. 4, *c*) не эффективна, поскольку не дает возможности однозначного определения шага остановки работы алгоритма. Для определения оптимального трешолдинга эффективна методика пошагового увеличения значения трешолдингового коэффициен-

та, поскольку в этом случае наблюдается ярко выраженный минимум значения энтропии Шеннона фильтрованных данных, соответствующий значению коэффициента трешолдинга τ = 1,8 (см. рис. 4, *d*). Вектор фильтрованных экспрессий генов и выделенная шумовая компонента при использовании вейвлета Добеши *db*5, втором уровне вейвлет-декомпозиции и значении коэффициента трешолдинга τ = 1,8 представлены на рис. 5.

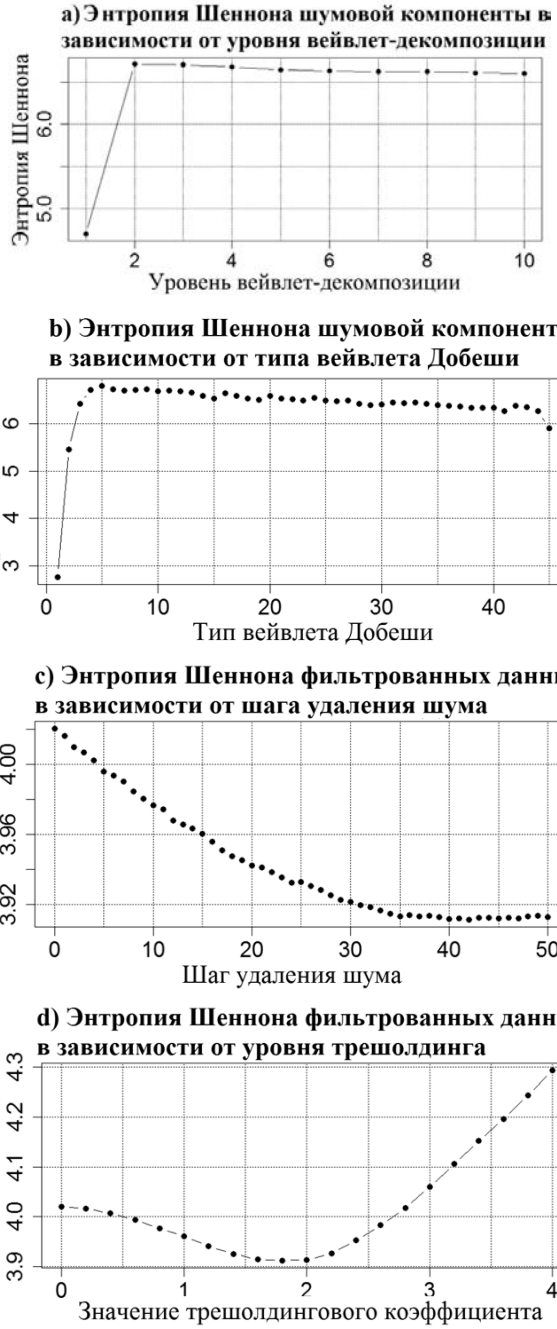


Рис. 4

Аналогичные результаты при использовании койфлетов, симплетов, биортогональных и обратных биортогональных вейвлетов представлены на рис. 6–9.

a) Фильтрованные данные с использованием вейвлета Добеши *db*5, значение энтропии Шеннона = 3,913

b) Выделенная шумовая компонента, значение энтропии Шеннона = 6,77

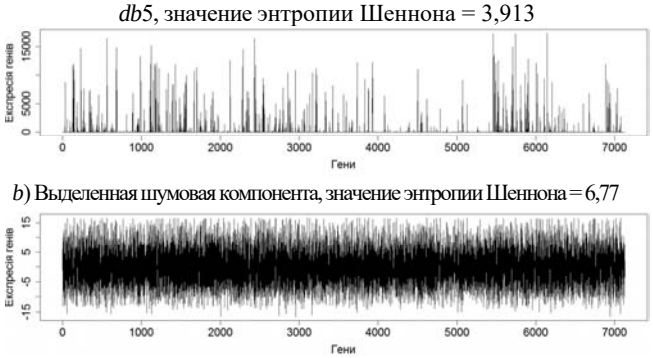


Рис. 5

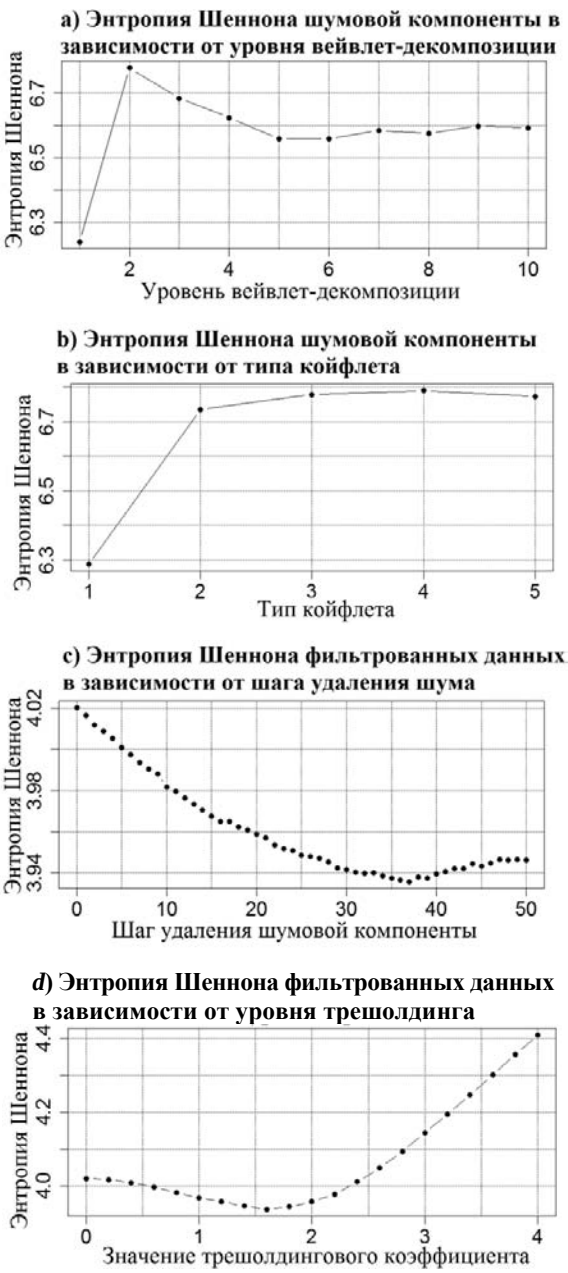
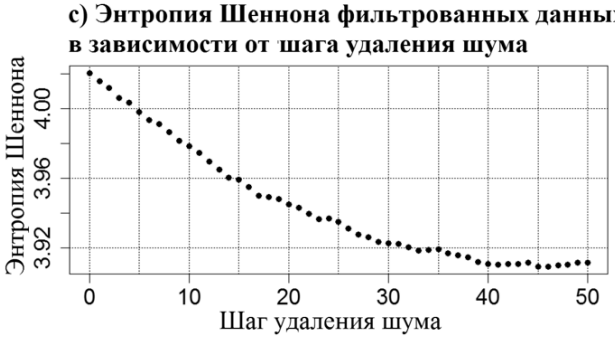


Рис. 6.


a) Энтропия Шеннона шумовой компоненты в зависимости от уровня вейвлет-декомпозиции
Энтропия Шеннона
Уровень вейвлет-декомпозиции


b) Энтропия Шеннона шумовой компоненты в зависимости от типа симплета
Энтропия Шеннона
Тип симплета


c) Энтропия Шеннона фильтрованных данных в зависимости от шага удаления шума
Энтропия Шеннона
Шаг удаления шума


d) Энтропия Шеннона фильтрованных данных в зависимости от уровня трешолдинга
Энтропия Шеннона
Значение трешолдингового коэффициента

Рис. 7


a) Энтропия Шеннона шумовой компоненты в зависимости от уровня вейвлет-декомпозиции
Энтропия Шеннона
Уровень вейвлет-декомпозиции

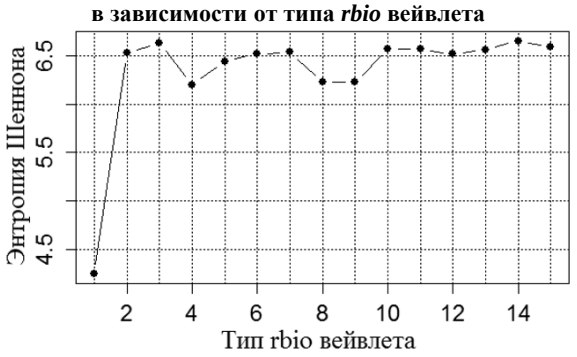

b) Энтропия Шеннона шумовой компоненты в зависимости от типа биортогонального вейвлета
Энтропия Шеннона
Тип биортогонального вейвлета


c) Энтропия Шеннона фильтрованных данных в зависимости от шага удаления шума
Энтропия Шеннона
Шаг удаления шумовой компоненты

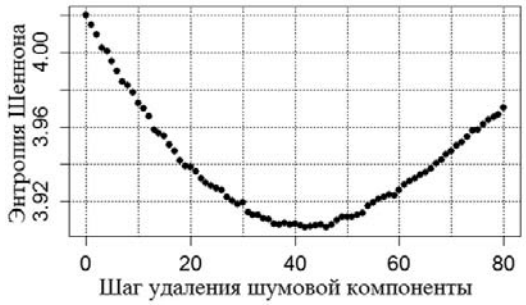

d) Энтропия Шеннона фильтрованных данных в зависимости от уровня трешолдинга
Энтропия Шеннона
Значение трешолдингового коэффициента

Рис. 8


a) Энтропия Шеннона шумовой компоненты в зависимости от уровня вейвлет-декомпозиции
Энтропия Шеннона
Уровень вейвлет-декомпозиции


b) Энтропия Шеннона шумовой компоненты в зависимости от типа *rbio* вейвлета
Энтропия Шеннона
Тип rbio вейвлета

***c*) Энтропия Шеннона фильтрованных данных в зависимости от шага удаления шума**



***d*) Энтропия Шеннона фильтрованных данных в зависимости от уровня трешолдинга**

Рис. 9

Анализ результатов позволяет сделать вывод, что оптимальными по критерию энтропия Шеннона является использование следующих параметров вейвлет-фильтра: вейвлет Добеши *db*5 при втором уровне вейвлет-декомпозиции и коэффициенте трешолдинга 1,8; койфлет *coif*4 при втором уровне вейвлет-декомпозиции и коэффициенте трешолдинга 1,6; симплет *sym*5 при третьем уровне вейвлет-декомпозиции и коэффициенте трешолдинга 1,8; биортогональный вейвлет *bior*1.5 при третьем уровне вейвлет-декомпозиции и коэффициенте трешолдинга 2,2 и обратный биортогональный вейвлет *rbio*1.5 при четвертом уровне вейвлет-декомпозиции и коэффициенте трешолдинга 1,8.

На рис. 10 представлена диаграмма отношения энтропий фильтрованных данных и выделенного шума в зависимости от типа используемого вейвлета при оптимальных параметрах вейвлет-фильтра. Анализ рис. 10 позволяет сделать вывод, что выбор типа материнского вейвлета из семейства ортогональных и биортогональных вейвлетов в случае фильтрации профилей экспрессии генов не является определяющим. С учетом отношения энтропии Шеннона для фильтруемых данных и выделенной шумовой компоненты лучшие результаты вейвлет-фильтрации получаются с использованием биортогонального вейвлета *bior*1.5. Но разница между результатами, полученными с использованием других вейвлетов достаточно мала. Определяющими в данном случае являются выбор типа вейвлета из семейства материнского вейвлета, выбор уровня вейвлет-декомпозиции и определение оптимального значения коэффициента трешолдинга для обработки детализирующих коэффициентов.

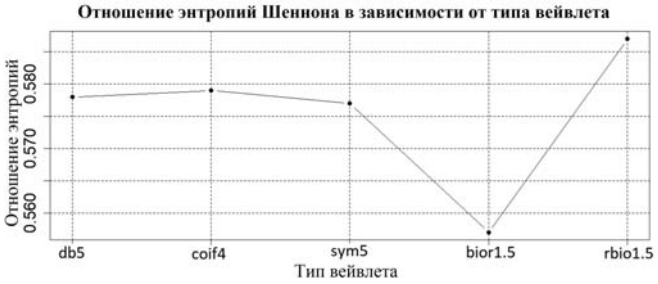


Рис. 10

Исследо вания позволяют разработать технологию определения оптимальных параметров вейвлет-фильтра для обработки профилей экспрессии генов в виде структурной блок-схемы пошаговой обработки информации. Архитектура данной технологии представлена на рис. 11. Практическая реализация представленной технологии предполагает наличие следующих этапов:

***Этап* I**. Инициализация исходных параметров вейвлет-фильтра.

1. *Формирование вектора материнских вейвлетов* и векторов вейвлетов из семейств выделенных материнских вейвлетов:

$$wv = \{wv_i\}\ ,\ i = 1,...,k,$$
$$wv_i = \{wv_i^j\}\ ,\ j = 1,...,p, \qquad (15)$$

где k – количество материнских вейвлетов, исследуемых в процессе моделирования, p –количество типов вейвлетов из семейства материнского вейвлета i. Определение интервала и шага изменения трешолдингового коэффициента: $\tau_{\min}, \tau_{\max}, d\tau = \tau_{\min}$, задание максимального уровня вейвлет-декомпозиции.

2. *Выбор материнского вейвлета*, соответствующего первому порядковому номеру вектора материнских вейвлетов ($i = 1$), произвольный выбор типа вейвлета данного материнского вейвлета, установление коэффициента трешолдинга $\tau = \tau_{\min}$ и уровня вейвлет-декомпозиции $n = 1$.

***Этап* II**. Определение оптимального уровня вейвлет-декомпозиции.

3. *Вейвлет-фильтрация вектора профилей* экспрессии генов в соответствии со схемой, изображенной на рис. 2, в пределах установленного интервала изменения уровня вейвлет-декомпозиции данных, выделение шумовой компоненты на каждом уровне вейвлет-декомпозиции и расчет энтропии Шеннона шумовой компоненты на каждом шаге.

4. *Анализ результатов*. Фиксация оптимального уровня вейвлет-декомпозиции для данного материнского вейвлета n_i^{opt}, соответствующего максимальному значению критерия энтропия Шеннона.

***Этап* III**. Определение типа вейвлета из семейства данного материнского вейвлета.

5. *Вейвлет-фильтрация* вектора профилей экспрессии генов для всех p типов данного материнского вейвлета $\left\{wv_i^j\right\}$, $j = 1,..., p$. Выделение шумовой компоненты, соответствующей каждому типу вейвлета и расчет энтропии Шеннона шумовой компоненты на каждом шагу работы алгоритма.

6. *Анализ результатов.* Фиксация оптимального типа вейвлета данного материнского вейвлета, соответствующего максимальному значению критерия энтропия Шеннона.

***Этап* IV**. Определение оптимального коэффициента трешолдинга для обработки детализирующих коэффициентов.

7. *Расчет энтропии* Шеннона для вектора исходных данных $H(\tau - d\tau)$.

8. *Вейвлет-фильтрация* вектора профилей экспрессии генов при использовании коэффициента трешолдинга τ. Выделение вектора фильтрованных данных.

9. *Расчет энтропии* Шеннона на данном шаге обработки информации $H(\tau)$.

10. Если $H(\tau) < H(\tau - d\tau)$, увеличение коэффициента трешолдинга на $d\tau$ и переход на шаг 8 данного алгоритма. В противном случае фиксация оптимального коэффициента трешолдинга для i-го материнского вейвлета: $\tau_i^{\text{opt}} = \tau - d\tau$.

11. *Операция инкремента* параметра i ($i = i + 1$) и, при условии $i \le k$, повторение этапов II–IV данного алгоритма.

***Этап* V**. Формирование окончательного решения по выбору параметров вейвлет-фильтра.

12. *Расчет отношений энтропий* Шеннона для фильтрованных данных и выделенной шумовой компоненты для каждого материнского вейвлета с использованием оптимальных параметров вейвлет-фильтра.

13. *Фиксация оптимальных параметров* вейвлет-фильтра, соответствующих глобальному минимуму критерия отношение энтропий Шеннона фильтрованных данных и выделенной шумовой компоненты.

Заключение. Представлена технология вейвлет-фильтрации профилей экспрессии генов с целью удаления фонового белого шума. В качестве основного критерия оценки информативности исследуемого сигнала использована энтропия Шеннона, рассчитанная по методу оценки вероятности Джеймса и Стейна, основанного на комплексном использовании двух моделей данных: высокоразмерной модели с малым смещением и высокой дисперсией распределения данных и низкоразмерной – с высоким смещением и низкой дисперсией. Реализация предложенной структурной блок-схемы системы вейвлет-фильтрации предполагает оценку энтропии как фильтрованного сигнала, так и удаленной шумовой компоненты, при этом окончательное решение по выбору параметров вейвлет-фильтра принимается на основе относительного критерия, который рассчитывается как отношение энтропий Шеннона фильтрованного сигнала и выделенной шумовой компоненты. В процессе моделирования исследованы семейства ортогональных вейвлетов Добеши, симплеты, койфлеты, биортогональные и обратные биортогональные вейвлеты. Проведено исследование по оптимизации процесса выбора типа вейвлета, уровня вейвлет-декомпозиции и значения трешолдингового коэффициента. Экспериментальное определение порогового значения трешолдингового коэффициента было проведено двумя способами. В первом случае проводилась пошаговая обработка детализирующих коэффициентов, когда значение трешолдингового коэффициента было достаточно малым (0,2) и не менялось в процессе моделирования. Продолжительность эксперимента ограничивалась количеством шагов обработки детализирующих коэффициентов. Второй случай пред-

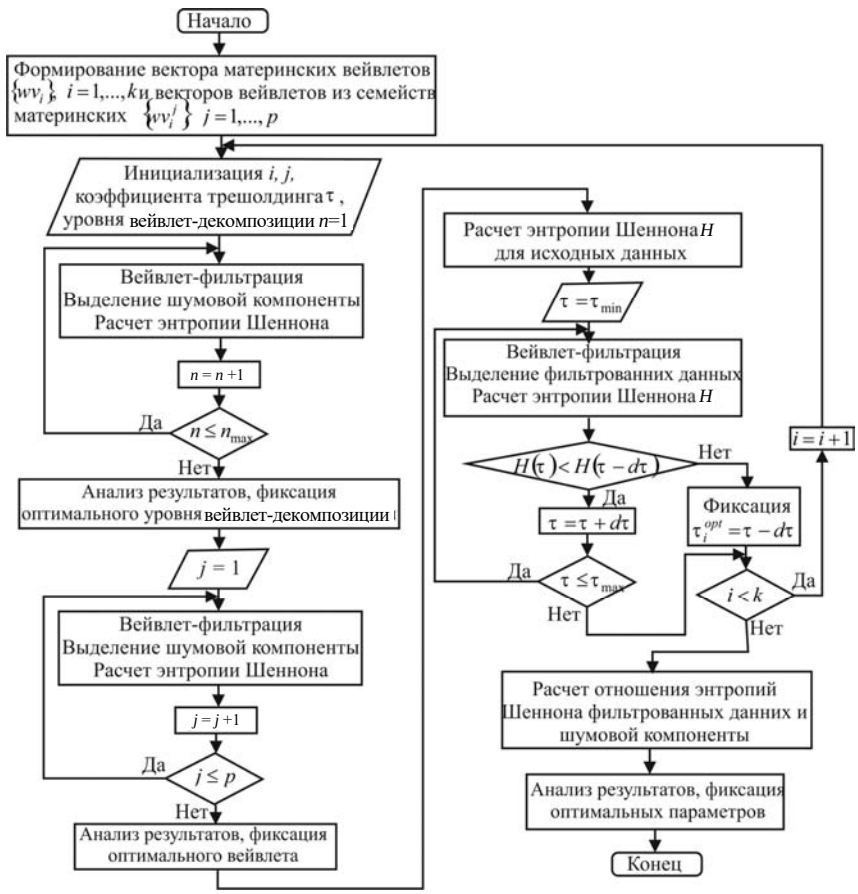


Рис. 11

полагал пошаговое увеличение значения трешолдингового коэффициента от τ_{min} до τ_{max} с шагом $d\tau$. Результаты моделирования представлены в виде графиков зависимости энтропий Шеннона от соответствующего параметра, максимальное или минимальное значение которых позволяет принять объективное решение по выбору соответствующего параметра.

Анализ результатов показал, что по критерию энтропия Шеннона оптимальны использование следующих параметров вейвлет-фильтра: вейвлет Добеши *db*5 при втором уровне вейвлет-декомпозиции и коэффициенте трешолдинга 1,8; койфлет *coif*4 при втором уровне вейвлет-декомпозиции и коэффициенте трешолдинга 1,6; симплет *sym*5 при третьем уровне вейвлет-декомпозиции и коэффициенте трешолдинга 1,8; биортогональный вейвлет *bior*1.5 при третьем уровне вейвлет-декомпозиции и коэффициенте трешолдинга 2,2 и обратный биортогональный вейвлет *rbio*1.5 при четвертом уровне вейвлет-декомпозиции и коэффициенте трешолдинга 1,8. Анализ диаграммы зависимости относительной энтропии от типа вейвлета позволяет сделать вывод, что выбор типа материнского вейвлета из семейства ортогональных и биортогональных вейвлетов в случае фильтрации профилей экспрессии генов не является определяющим. С учетом отношения энтропий Шеннона для фильтруемых данных и выделенной шумовой компоненты лучшие результаты по вейвлет-фильтрации получаются с использованием биортогонального вейвлета *bior*1.5. Но разница между результатами, полученными с использованием других вейвлетов достаточно мала. Определяющими в данном случае является выбор типа вейвлета из семейства используемого материнского вейвлета, выбор уровня вейвлет-декомпозиции и определение оптимального значения коэффициента трешолдинга для обработки детализирующих коэффициентов. На основе проведенных исследований предложена технология определения оптимальных параметров вейвлет-фильтра для обработки профилей экспрессии генов в виде структурной блок-схемы пошаговой обработки информации.

Перспективой дальнейших исследований является практическая реализация предложенной технологии в рамках гибридной модели предобработки профилей экспрессии генов с целью реконструкции генной регуляторной сети.

UDC 004.048

S.A. Babichev

PhD., associate professor, associate professor of department of informatics, Jan Evangelista Purkyne University in Usti nad Labem, Czech Republic, 8, Ceske mladeze Str., Usti nad Labem, Czech Republic, 400 96.

Technology of Wavelet-Filtration of the Gene Expression Profiles in Order to Remove the Background Noise

Keywords: gene expression profiles, wavelets, tresholding, filtration.

Introduction. The solved task is focused on increasing the gene expression profiles quality, which are used to reconstruct the gene regulatory networks. The filtration process is one of the stages of data preprocessing, implementation of which corresponds to the increasing data quality by removing the background “white” noise component.

The aim of the paper is development of the wavelet filtration technology of gene expression profiles based on the Shannon entropy criterion, which calculated by James-Stein shrinkage estimator using.

Methods. During the research, the methods of the computer simulation, wavelet analysis, and entropy methods to estimate the studied data comprehension are used.

Results. The results of the simulation prove that the choice of the mother wavelet type from orthogonal and biorthogonal wavelets in case of the gene expression profiles filtration is not determinative. In terms of the relative criterion calculated as the Shannon entropy ratio of the filtered gene expression profiles and the extracted noise component, the best results are obtained using the biorthogonal wavelet bior1.5, however the difference obtained using other types of wavelets is insignificant. The choice of the type of the wavelet from the family of the mother’s wavelets, the choice of the level of the wavelet decomposition, and the choice of the value of the thresholding coefficient are determining in this case.

Conclusions. The wavelet filtration technology of gene expression profiles based on complex use of the methods to estimate the filtered signal and extracted noise comprehension component is proposed based on the performed simulation. The implementation of this technology allows us to optimise the wavelet filtration process of complex signals in order to remove the “white” noise component.