UDK 004.048

S.A. BABICHEV
Jan Evangelista Purkině University in Ústi nad Labem, Czech Republic
I.A. LURIE, M.A. VORONENKO
Kherson National Technical Univesity, Ukraine

# JAMES-STEIN SHRINKAGE ESTIMATOR OF SHANNON ENTROPY IN WAVELET-FILTRATION SYSTEMS OF COMPLEX DATA

*The paper presents the wavelet-filtering technology of complex data, where Shannon entropy, which was calculated based on James-Stein shrinkage estimator is used as a criterion to evaluate the quality of the information processing. The gene expression sequence, which was obtained by microchip experiments, was used as experimental data. It have been developed the algorithm of the studied data wavelet filtering, where the level of wavelet decomposition and the type of the wavelet is determined based on Shannon entropy maximum value of the deleted from data noise component and the thresholding coefficient value is determined based on the entropy minimum value of the filtering data.*

*Keywords: Shannon entropy, James-Stein shrinkage estimator, wavelet filtering, gene expression sequence, filtering, thresholding*

С.А. БАБІЧЕВ
Університет Яна Евангелиста Пуркінє в Усті на Лабі, Чехія
І.А. ЛУР'Є, М.О. ВОРОНЕНКО
Херсонський національний технічний університет, Україна

## ОЦІНКА ЕНТРОПІЇ ШЕННОНА НА ОСНОВІ МЕТОДУ СТИСНЕННЯ ДЖЕЙМСА ТА СТЕЙНА У СИСТЕМАХ ВЕЙВЛЕТ-ФІЛЬТРАЦІЇ СКЛАДНИХ ДАНИХ

*У статті представлено технологію вейвлет-фільтрації даних складної природи, у якій як критерій оцінки якості обробки інформації використовується ентропія Шеннона, що розрахована на основі методу стиснення Джеймса та Стейна. Як експериментальні дані було використано послідовність експресій генів, що отримана за допомогою мікрочіпових експериментів. Розроблено алгоритм вейвлет-фільтрації даних що досліджуються, у якому рівень вейвлет декомпозиції і тип вейвлету визначається на основі максимуму ентропії виділеної з сигналу шумової компоненти, а значення трешолдінгового коефіцієнту визначається на основі мінімуму ентропії фільтрованих даних.*

*Ключові слова: ентропія Шеннона, метод стиснення Джеймса-Стейна, вейвлет-фільтрація, послідовність експресій генів, трешолдінг*

С.А. БАБИЧЕВ
Университет Яна Евангелиста Пуркине в Усти на Лабе, Чехия
И.А. ЛУРЬЕ, М.А. ВОРОНЕНКО
Херсонский национальный технический университет, Украина

## ОЦЕНКА ЭНТРОПИИ ШЕННОНА НА ОСНОВЕ МЕТОДА СЖАТИЯ ДЖЕЙМСА-СТЕЙНА В СИСТЕМАХ ВЕЙВЛЕТ-ФИЛЬТРАЦИИ СЛОЖНЫХ ДАННЫХ

*В статье представлена технология вейвлет-фильтрации данных сложной природы, в которой в качестве критерия оценки качества обработки информации используется энтропия Шеннона, рассчитанная на основе метода сжатия Джеймса и Стейна. В качестве экспериментальных данных использовалась последовательность экспрессий генов, полученная посредством микрочиповых экспериментов. Разработан алгоритм вейвлет-фильтрации исследуемых данных, в котором уровень вейвлет-декомпозиции и тип используемого вейвлета определяется на основе максимума энтропии, выделенной с сигнала шумовой компоненты, а значение трешолдингового коэффициента определяется на основе минимума энтропии фильтрованных данных.*

*Ключевые слова: энтропия Шеннона, метод сжатия Джеймса-Стейна, вейвлет-фильтрация, последовательность экспрессий генов, трешолдинг*

### Problem statement

Creation of the models of gene regulatory networks based on the gene expression sequences, which are obtained by DNA microchip experiments or by RNA sequencing methods is one of the actual direction of modern bioinformatics. Accuracy of the obtained model work is determined by the quality of the experimental data

preprocessing, one of the steps of which is the filtration of the gene expression sequences, which are obtained by DNA microchip experiments. Scanning process of the DNA microchip data is accompanied by background noise. Partial correction of the noise component is performed by background correction at the stage of gene expression estimation. However, nowadays, it is not possible to remove completely this noise component during gene expression array creation. Owing to the above, there is a necessity to develop the technology of complex high dimensional data filtration based on the modern computing methods of information processing and estimation.

**Analysis of recent research and publications**

The papers [1,2] are devoted to the questions of visualisation, estimation and preprocessing of high dimensional data. The authors review the common techniques to explore and visualize high dimensional data on examples of gene expression sequences and mass spectrometry protein data. The technology to reduce the uninformativity features in the gene expression array based on the use of the statistical criteria of the studied data estimation is presented in [3]. Implementation of this technology allows reducing of the feature space dimension in the range from 5 to 10%, but it does not solve the problem of noisiness of the remained data. The [4,5] discuss the problem of complex data filtration based on wavelet analyses. The authors proposed the hybrid technology based on complex using the wavelets and Winner filter. However, it should be noted that the problem of the wavelet filter parameters optimisation based on the quantity criteria to estimate the quality of the information processing has not final decision nowadays.

**Unsolved parts of the general problem** are the absence of the efficient methods of complex high dimension data filtration based on the complex use of modern methods of information processing and quality of the obtained results estimation.

**The aim of the paper** is development of the technique of high dimensional complex data wavelet filtration, where the estimation of the data processing quality is performed based on Shannon entropy criterion using James-Stein shrinkage estimator.

**The presentation of the basis material**

Gene expression sequence is a vector the components of which are the expressions of genes, which determine the character of functioning of the appropriate cells of biological organism. Three technologies are actual to determine the gene expression sequences nowadays. These technologies are presented in Fig. 1.

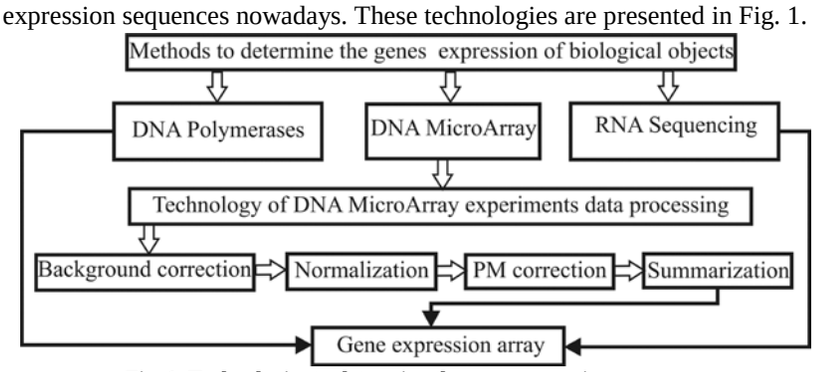


**Fig. 1. Technologies to determine the gene expression sequences**

DNA polymerises and RNA sequencing technologies are more exact in comparison with DNA microarray technology. The gene expression sequences, which are obtained by these technologies, have significantly lower level of noise component, but these technologies are very expensive. DNA microarray technology allows estimating of the gene expression of tens of thousands genes concurrently. This technology is cheaper but the data of genes expressions include the complexity noise component, which is determined by the processes of microchip creation and reading information from it. An example of gene expression sequence of one of the studied objects, which is obtained by DNA microchip method, is shown in Fig. 2.

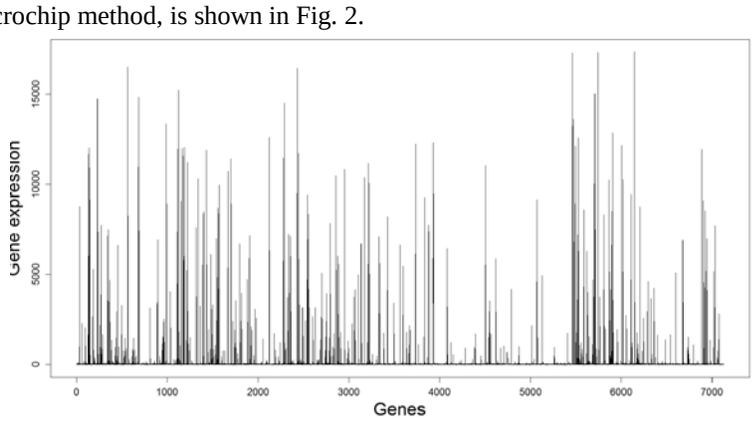


**Fig. 2. An example of gene expression sequence**

The statistical characteristics of the studied vector are presented in Table 1.

Table 1

Statistical characteristics of the studied gene expression sequences

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

As it can be seen, the gene expression sequence includes about seven thousand genes, expression of which is changed in the range from -36,19 to 17360. At the same time, the most of the gene expression have low values. Analysis of Fig.2 and Table 1 allows us to conclude also that the range of change of the gene expression of noise component is significantly lower in comparison with range of change of the gene expression sequence. Moreover, the frequency of the noise component most probably is more in comparison with frequency of the useful component of the studied vector. This fact allows us to use the wavelet analysis to solve the problem of studied data filtration. Wavelets are the families of functions $\Psi_{a,b}(t)$, which are generated from the basis of mother wavelets by choosing the parameters a (scale parameter) and b (shift parameters) [6,7]:

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

The process of wavelet processing for purpose of data filtration is presented in Fig. 3.

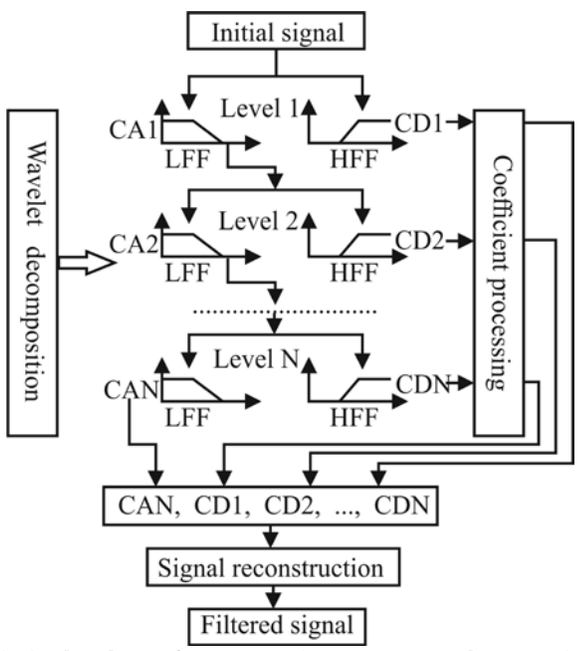


**Fig. 3. The scheme of gene expression sequence wavelet processing**

Approximation coefficients on N level and detail coefficients on levels from 1 to N are calculated during the wavelet decomposition process. The noise component in the most cases are in the detail coefficients since these coefficients have higher frequency, therefore the detail coefficients are processed at the next step using certain value of the thresholding coefficient. The soft thresholding was used for detail coefficients processing in case of gene expression sequence. If $\tau$ – is the thresholding coefficient value and d – is the detail coefficient value, the processing of the detail coessicient in case of the soft thresholding is performed by formula:

$$\begin{cases} d = 0, \ \ if \ \ d \le \tau, \\ d = d - \tau, \ \ if \ \ d > \tau. \end{cases} \tag{2}$$

The signal reconstruction is performed based on the approximation coefficient on N level and processed detail coefficient on levels from 1 to N. The analysis of Fig. 3 allows us to conclude that the wavelet prosessing of the studied data involves the following:

- choise of the mother wavelet;
- determination of the wavelet decomposition level;
- choise of the type of the wavelet from the basis of the mother wavelet;
- determination of the thresholding coefficient value.

Each of these steps involves estimation of the processing quality in order to determine the optimal parameters to process the studied data. To estimate the data processing quality Shannon entropy criterion using James-Stein shrinkage estimator was used. If $k$ – is the quantity of cells with probablities $p_1, p_2, ..., p_k$, where $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$, then Shannon entropy is defined as quantitative measure of the uncertainty of the system state and it is calculated as follow [8]:

$$H = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{3}$$

Two diferent models are the basis of James-Stein shrinkage estimator: a high-dimensional model with low bias and high variance and low-dimensional model with larger bias and lower variance [9]. The probablity of James-Stein shrinkage estimator in $i$ cell is calculated by formula:

$$p_i^{Shrink} = \lambda t_i + (1-\lambda) p_i^{ML}, \tag{4}$$

where $t_i = \frac{1}{n_i}$ – is the target probabbility in $i$ cell or probablity in case what all features in $i$ cell are different; $p_i^{ML}$ – is the probability in $i$ cell calculated by maximum-likelihood method:

$$p_i^{ML} = \frac{n_j}{n_i}, \tag{5}$$

where $n_i$ – is the quantity of features in $i$ cell, $n_j, j = 1,...,k$ – is the quantity of the j-th feature in $i$ cell. $\lambda$ – is the shrinkage intetsity that takes the values from 0 (no shrinkage) to 1 (full shrinkage) and it is calculated by the formula:

$$\lambda = \frac{1 - \sum_{i=1}^{k} \left(p_i^{ML}\right)^2}{(n-1)\sum_{i=1}^{k} \left(t_i - p_i^{ML}\right)^2}, \tag{6}$$

where $n$ – is the quantity of the studied vector features. Taking into account the hereinbefore, the formula to calculate Shannon entropy using James-Stein shrinkage estimator can be presemted as follow:

$$H^{Shrink} = -\sum_{i=1}^{k} p_i^{Shrink} \log_2 p_i^{Shrink}. \tag{7}$$

Obviously, higher value of Shannon entropy corresponds to lower quantity of useful information in the studied vector. Maximum value of Shannon entropy corresponds to white noise component. Thus, lower value of Shannon entropy of the studied vector or higher value of Shannon entropy of the removed noise component corresponds to better quality of the studied vector processing. The structural block diagram of the wavelet filtration process using James-Stein shrinkage Shannon entropy estimator is presented in Fig. 4.

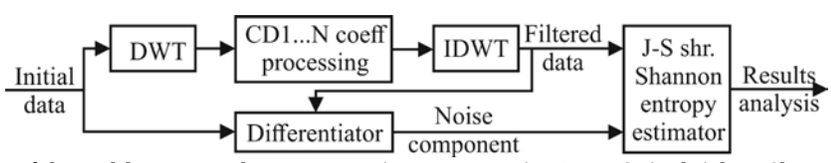


**Fig. 4. Block diagram of the model to process the gene expression sequence using James-Stein shrinkage Shannon entropy estimator**

Implementation of the hereinbefore process involves the following steps:

1. Choise of the mother wavelet from the list of the avialable ones.
2. Determination of the optimal level of wavelet decomposition based on Shannon entropy maximum value for the selected noise component. At this step we use randomly chosen type of wavelet from the family of the mother wavelet.
3. Determination of type of wavelet from the mother wavelet family based on the maxomum value of Shannon entropy for the selected noise component.

4. Determination of the thresholding coefficient value based on the minimum value of Shannon entropy for filtered data.

The practical implementation of the presented technology was performed based on the family of Daubechy wavelets. Wavelets db1 (or haar)… db45 were used during simulation process. Determination of the thresholding coefficient value was carried out in two ways. The first way involves step by step removing of the noise component using low constant value of the thresholding coefficient. Estimation of Shannon entropy of the filtered data is performed at each step of the filtration process. The duration of the filtration process is limited by quantity of the noise removing steps. The second way involves step by step increase of the threshilding coefficient value from $\tau_{\min}$ to $\tau_{\max}$ with step $d\tau$. The estimation of the filtered data Shannon entropy at each step is performed concurently. The final decision about the optimal step of data processing in both cases is taken out based on the minimum value of Shannon entropy for filtered data. The results of the simulation process for gene expression sequence filtering are shown in Fig. 5. Fig. 6 shows the filtered sequence and the removed from sequence noise component.

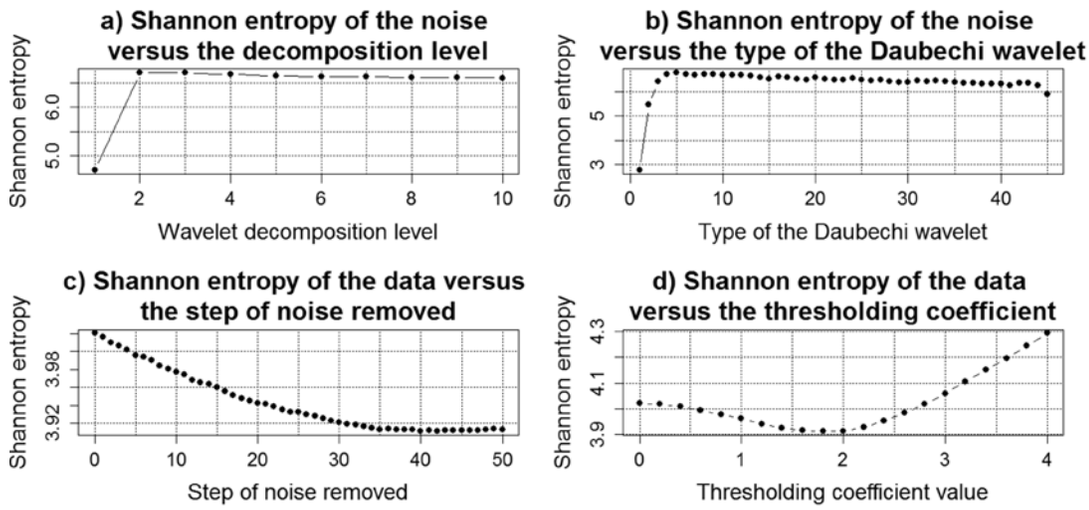


**Fig. 5. Charts of James-Stein shrinkage Shannon entropy versus the: a) level of wavelet decomposition; b) type of the Daubechi wavelet; c) step of noise component remove; d) thresholding coefficient value**

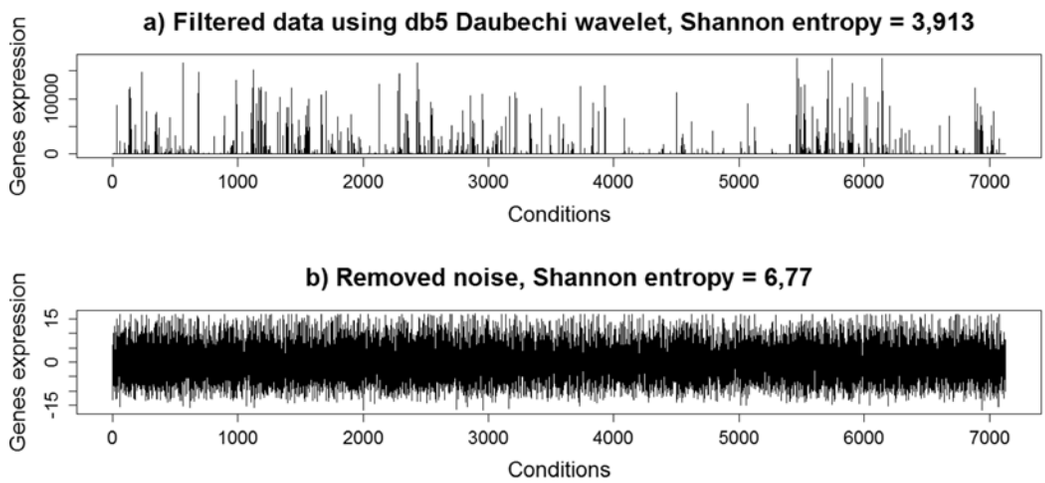


**Fig. 6. Results of the simulation: a) filtered gene expression sequence; b) removed noise component**

The analysis of the obtained results allows us to conclude that an optimal in terms of Shannon entropy criterion is the gene expression processing using wavelet db5 for two level of wavelet decomposition and thresholding coefficient value 1,8. The comparison of the charts in Fig. 5c and 5d allows us also to conclude that the step by step increase of the thresholding coefficient value with little step of its change is more effective in comparison with step by step removing noise component with constant value of the thresholding coefficient. The local minimum of Shannon entropy value in Fig. 5d is more evident in comparison with local minimum in Fig. 5c.

**Conclusion**

The paper presents the technology of filtering the gene expression sequence based on the complex use of the wavelet analysis and James-Stein shrinkage estimator. Implementation of this technology allows us to determine the optimal parameters of the wavelet filter in terms of quantitative criterion of the data processing quality estimation. The family of Daubechies wavelet was used during simulation process. The use of James-Stein shrinkage estimator to calculate Shannon entropy is determined by complex character of this method. James-Stein shrinkage estimator takes into account two very diferent models: a high-dimensional model with low bias and high variance and low-dimensional model with larger bias and lower variance. This fact allows us to obtain higher objectivity during estimation of the data processing quality. The wavelet filtering process of the gene expression sequence includes three stages. The first stage involves the level of wavelet decomposition determination. At the second step the choise of Daubechi wavelet type was performed. In these cases the solution is made based on the maximum Shannon entropy value for the removed noise component. The third stage involves the thresholding coefficient value determination based on the minimum Shannon entropy value for the filtered data. Determination of the thresholding coefficient value was carried out in two ways. The first way involves step by step removing of the noise component using low constant value of the thresholding coefficient. Estimation of Shannon entropy of the filtered data is performed at each step of the filtration process. The duration of the filtration process is limited by quantity of the noise removing steps. The second way involves step by step increase of the threshilding coefficient value from $\tau_{\min}$ to $\tau_{\max}$ with step $d\tau$. The results of the simulation show that the step by step increase of the thresholding coefficient value with little step of its change is more effective in comparison with step by step removing of the noise component with constant value of the thresholding coefficient. Moreover, an optimal in terms of Shannon entropy criterion is the gene expression processing using wavelet db5 for two level of the wavelet decomposition and thresholding coefficient value 1,8. The perspective of the autor's research is the creation of the complex technology of the gene expression sequences preprocessing, where the data filtering will be one of the stages of the studied data processing.

**References**

1. Wu Z. Exploration, visualization, and preprocessing of high-dimensional data / Z. Wu // Methods Molecular Biology, 2010.– Vol. 620. – P. 267-284.
2. Ozsolak F. RNA sequencing: advances, challenges and opportunities / F. Ozsolak, P.M. Milos // Nature Reviews Genetics, 2011**. –** Vol**.**12**. –** P.87-98.
3. Babichev S. Filtration of DNA nucleotide gene expression profiles in the systems of biological objects clustering / S.Babichev, M.A.Taiff, V. Lytvynenko // International Frontier Science Letters. – 2016. – Vol. 8.– P.1-8.
4. Joshi A. Analysis of Adaptive Wavelet Wiener Filtering for ECG Signals: Review / A. Joshi, H.S. Aravind // International Journal of Advanced Research in Electronics and Communication Engineering. – 2014. – Vol. 3. – Issue 4.– P. 395-398.
5. Chandu R. ECG Signal Filtering using an Improved Wavelet Wiener Filtering / R. Chandu. M. Venkateswarlu // International Journal of Advanced Technology and Innovative Research. – 2015. – Vol. 7.– Issue 7.– P. 1242-1247.
6. Daubechies I. The wavelet transform, time-frequency localization and signal analysis / I. Daubechies // IEEE Trans. Inform. Theory. – 1990. – Vol. 36. – P. 961-1005.
7. Coifman R.R. Wavelet Analysis and Signal Processing / R.R. Coifman, Y. Meyer, M.V. Wickerhauser // Wavelets and Their Applications. – Boston Jones and Bartlett, 1992.– P. 153-178.
8. Shannon C.E. A mathematical theory of communication / C. E. Shannon // Bell System Technical Journal, 1948. – V. 27. – P. 379–423, 623–656.
9. Hausser J. Entropy inference and James-Stein estimator, with application to nonlinear gene association networks / J. Hausser, K. Strimmer // Journal of Machine Learning Research. – 2009. – Vol. 10. – P. 1469-1484.