International Journal of Environmental Engineering– IJEE
Volume 2 : Issue 1 *[ISSN : 2374-1724]*
Publication Date: 30 April, 2015

Nonlinear Methods of DNA Coding Regions Identification

[Nataliia V. Kudriavtseva[1], Pavel Chmelar[1], Vyacheslav A. Tykhonov[2]]

***Abstract*—We used the higher orders spectra to improve observability of the coding regions. Smoothened spectrograms of the second and sixth orders were received. Efficiency of the proposed method for protein coding regions was estimated on 8000 samplings of DNA C. elegans F56F11.4. We used signal-to-noise ratio to evaluate the accuracy of the measures in predicting the coding regions.**

***Keywords*—autoregressive models, DNA sequence, spectrogram, power spectrum density**

I. Introduction

The important problem of DNA sequence analysis is coding regions identification. Literature survey shows that there is no prevailing method for identification of protein coding regions.

We propose a new method of DNA parametric power spectrum density (PSD) estimations for identification of protein coding regions in this paper. We used discrete Fourier transform method for the second and the higher orders of analysis. The higher order spectra shows non-Gaussian characteristics of a DNA. The spectra were calculated after preliminary DNA transformation on the basis of codons redundancy during amino acids encoding. Preliminary DNA transformation shows that it is possible to strengthen the nucleotides periodicity in the coding regions.

II. Periodicity of Nucleotides

Periodicity of nucleotides is caused by redundancy of genetic code, preferences in using of specific codons for coding of amino acids, and prevalence of proteins by certain amino acids. There is a hypothesis that triplet periodicity can emerge as a result of a need for control of mutations using shift of a reading frame. In real genome DNA sequences, periodic characteristics appear on a strong random background. Such a randomness in caused by point mutations, insertion segments, deletions, translocations, etc.

Nataliia Kudriavtseva[1], Pavel Chmelar[1]
Department of Electrical Engineering, Faculty of Electrical Engineering and Informatics, University of Pardubice
Pardubice, Czech Republic

Vyacheslav A. Tykhonov[2]
Department of Radioelectronic Systems, Kharkiv National University of Radioelectronics
Kharkiv, Ukraine

The insertion segments and deletions are the reason of period length variations. The 3-base periodicity is a universal property of protein coding regions. There are some explanations about origin and universality of the 3-base periodicity p = 3 [1]. Authors in [2] explain it by evolutionary origin of genetic code and dominance of codons in mRNA on an early stage of a molecular evolution. The most rapid case is expressed by more frequent occurrence of G in the first position of mRNA triplets, and avoidance of G in the second positions (G-non-G-N). The same pattern can be observed in the form (RNY) (R-is A or G, Y is C or U, N is any base).

III. The Spectral Analysis of DNA Coding Regions

Spectral analysis is traditionally used to identify protein coding regions. We calculated PSD estimations on the frequency that corresponds to the 3-base periodicity of a DNA sequence. The Voss mapping was used in this paper. Spectrograms were calculated for a moving window along the DNA sequence using discrete Fourier transform

$$X_m = X_m[k] = \sum_{n=0}^{N-1} x_m[n] \exp(-j2\pi nk / N),$$

where $k = 0, \ldots, N-1$, $m = \{A, C, G, T\}$. If we take the moving window's size N, then in the magnitude of X_m the frequency index is equal to $k = N / 3$.

In the Voss mapping [3] a symbol sequence is divided into four numerical sequences. Coefficient that were calculated in [4], are $a = 0.1 + 0.12j$, $t = -0.3 - 0.2j$, $c = 0$, $g = 0.45 - 0.19$. We can show that parametric PSD also can be expressed by a square of absolute value spectrum sum. We propose a parametrical autoregressive (AR) estimation of spectrum on the base of an additive linear prediction AR model. This spectrum estimation improved frequency resolution of the traditional AR spectrum. The additive AR model can be used for spectral analysis DNA.

AR model can be defined using an operator form

$$\Phi(z^{-1})x[t] = a[t], \quad (1)$$

where $a[t]$ – is a prediction error such as white noise. Equation (1) can be also presented in the following form

$$x[t] = \Phi^{-1}(z^{-1})a[t] = H(z^{-1})a[t]. \quad (2)$$

This expression describes generating linear filter with a feedback loop. The input process of this filter is a white noise *a[t]* and the output process is a correlated process *x[t]*. The PSD is equal by form to the amplitude-frequency characteristic (AFC)

$$|H[f]|^2 = \left|\Phi^{-1}[f]\right|^2 = \left|\sum_{n=0}^{p}\Phi[i]e^{-j2\pi fiT}\right|^{-2},$$

and is depended on AR coefficients. AFC of a generating linear filter has inversely AFC of prediction filter. In some cases it is useful to calculate AR coefficients of generating and inverse prediction filter using spectral characteristics of *x[t]* or AFC. For example, for spectral peak of DNA sequences with period 3 we can find a relation with AR coefficients and PSD characteristics of spectral peaks and broadband at the level of 0.5. The stationarity condition of AR process follows from the robustness condition of generating AR filter with transfer function [5]. Stationarity condition of a stochastic AR process follows from a characteristic equation

$$c^p - \Phi[1]c^{p-1} - \ldots - \Phi[p] = 0. \quad (3)$$

If the roots *c*[i] of the characteristic equation (3) lie inside the unit circle then the process is stationary. AFC of inverse prediction filter is always robust because it does not contain the feedback loop. From (3) the following equality follows

$$c^p - \Phi[1]c^{p-1} - \ldots - \Phi[p] = \prod_{i=1}^{p}(c - c[i]).$$

The roots of the characteristic equation fully describe the AR model. If a root is a real it can be represented using an exponential function

$$c[i] = e^{-\pi\Delta f[i]T},$$

where *Δf[i]* – broadband of i-th spike PSD, *T* – quantization interval.

We can make random series which is a sum of four known components. Such a random process that is described by a linear prediction model we call additive linear prediction process. Generating filter for an additive linear prediction process is composed by four parallel connected generating filters with the same input generating process. Gaussian and non-Gaussian white noise is used as input generating process.

Additive stationary stochastic process of AR is represented by an equation [6, 7] similar to (2). The expression for parametric PSD of additive AR model is the following

$$S(f) = \left| \frac{1}{\sum_{i=0}^{\sigma_1}\Phi_1[i]e^{-i\Delta\pi fT}} + \frac{1}{\sum_{i=0}^{\sigma_2}\Phi_2[i]e^{-i\Delta\pi fT}} + \frac{1}{\sum_{i=0}^{\sigma_3}\Phi_3[i]e^{-i\Delta\pi fT}} + \frac{1}{\sum_{i=0}^{\sigma_4}\Phi_4[i]e^{-i\Delta\pi fT}} \right|^2 D_z. \quad (4)$$

Our investigations [6] show, that this expression gives a more exact estimation of PSD by Yule-Walker and Burg algorithms than traditional parametrical AR estimation of PSD. The resolution capability of proposed estimation (4) is significantly higher because we calculated mutual spectra of the additive processes, not only separate spectra of the processes. Naturally, Fourier's spectra of the additive processes have the same features.

In the work [4] optimal weighting coefficients *a*, *t*, *c*, *g* are described for fixed DNA region. DNA spectrum is presented like sum of spectrums for each nucleotide. Therefore, such spectrum of DNA is described by spectrum of additive linear prediction AR model.

The most of the coding regions studies the prediction Fourier analysis is used. Therefore, we analyzed the efficiency of our methods using the Fourier's spectra.

Efficiency of the proposed spectral analysis method for protein coding regions was estimated on the area of 8000 DNA samples of C. elegans F56F11.4 (www.ncbi.nlm.nih.gov). The spectrogram of the DNA is shown on the Fig. 1. Analysis of the figures shows an improvement of the gene prediction.

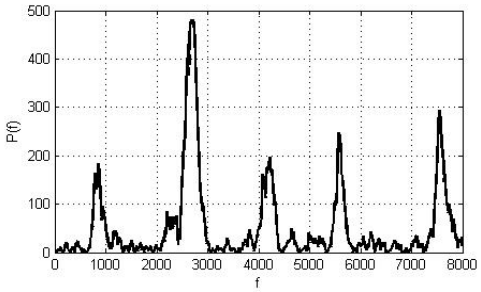

Figure 1. The spectrogram of the second order of the DNA area

IV. The Spectral Analysis of Higher Orders

Spectra of higher orders are necessary to study various properties of non-Gaussian processes [8, 9]. A non-Gussian process is fully described by a set of spectra of all orders, higher orders spectra as well as cumulant functions can contain additional information about non-Gaussian stochastic process. It is known, that DNA sequences possess non-Gaussian properties with nonzero skewness and kurtosis values.

Apart from researching of spectral second order statistics of coding regions, possibilities of spectral higher order statistics using the Fourier method were studied. For analysis of non-Gaussian processes we used Fourier sixth order

statistics of a DNA sequence after replacement of some codons to the synonyms. This characteristics has been received by the product of Fourier transform sequences. The spectra of k-th order of an additive DNA sequence consists of an additive sum of four nucleotides sequences. It is calculated using the expression

$$X_k = \left| aX_A + tX_T + cX_C + gX_G \right|^k . \quad (5)$$

Spectra of higher orders for stochastic processes can be useful in various real-life situations. Spikes in a spectrum of higher order are more powerful then spectra of the second order for some stochastic processes. The reason for this is in the difference of density distribution of a useful signal and a noise of a coding region and in the difference between coding and non-coding regions. A coding sequence can be considered as a mixture of a useful signal and a noise. If statistical characteristics of the white noise are close to the Gaussian distribution than that of the useful signal's characteristics the effect of the white noise on the accuracy of the higher order spectra estimation is less than that of the effect on the spectra of the second order. To receive the most accurate spectrogram it is necessary to select the spectrum's order. Even in a case of a correlated noise that "covers" spectrum of the useful signal it is possible to detect this spectrum using an appropriate spectrum's order. An application of a nonlinear transformation to calculate higher order spectra (for example, for periodogram method involution to k-th degree) intensifies powerful components of the spectrum of coding regions.

Spectra of higher orders are multidimensional and depend on k–1 frequencies. But in many cases instead of multidimensional spectra we can analyze unidimensional spectra of higher orders. Unidimensional spectra are certain sections of multidimensional spectra. In our work we used a section, where $f_2, f_3, \ldots, f_{k-1} = 0$.

Generally, to analyze spectra of higher orders spectra of the third and fourth orders are used. Spectrum of the third order sometimes is called bispectrum [9], and spectrum of the fourth order is called triplespectrum. It is difficult to determine the most effective spectrum's order beforehand. It depends on multidimensional distributions of the coding and non-coding regions. Hence, it is necessary to select an appropriate spectrum's order.

The spectrograms of the sixth order that were calculated using (5) when k = 6 of analyzed a DNA sequence before replacement of codons by synonyms and after replacement of the spectrogram are shown on the Fig. 2. Analysis of the figures shows that spikes of the first, third, fourth and fifth spectrograms of the coding regions are higher. Spectrograms were received after replacement of codons by synonyms. But at the same time the level of spectrogram's non-coding regions became higher.

When the spectrum's order is high the spectrogram's level of non-coding regions decreases. The Fig. 1 and Fig. 2 show that the spectrogram's level of non-coding regions is significantly low. We can observe a significant difference in the spectrogram's level of coding and non-coding regions before and after the replacement of codons to synonyms. The strongest decrease of the spectrogram's level of non-coding regions appears when using spectra of the sixth order. To demonstrate the decrease of the spectrogram's level of the non-coding regions we will limit the spectra's amplitude Fig. 3(b) by level 10^7. The clipped spectrograms are shown in Fig. 4.

Statistical averaging is not used in the methods of the spectrograms calculation for DNA sequences. Therefore, figures of spectrograms have strong fluctuations. The fluctuations can be decreased using smoothening windows and filters. In our paper, we used AR generating filter for smoothing. Filtration using an AR filter is described in (1).The input process in this case is a mapping DNA sequence after the replacement of codons to synonyms, but not white noise. The spike's frequency of the amplitude-frequency characteristic is $N / 3$, where N – length of the moving window, which we used for calculation of the spectrogram.

The bandwidth is selected to receive appropriate smoothing and not to accept shifting of the spectrogram coding regions on the x-coordinate. A narrow-band AR filter amplifies smoothing of the spectrum. But the lower the filter's band the higher the filter's lag. This is the reason for the spectrogram's displacement. Smoothed spectrograms that were received for spectra estimations when $k = 2, 6$ are shown in Fig. 3. Clipped smoothed spectrograms are shown in Fig. 5. Comparison of Fig. 3 and Fig. 5 show, that the shift of the spectrograms of the coding regions after smoothing is small.

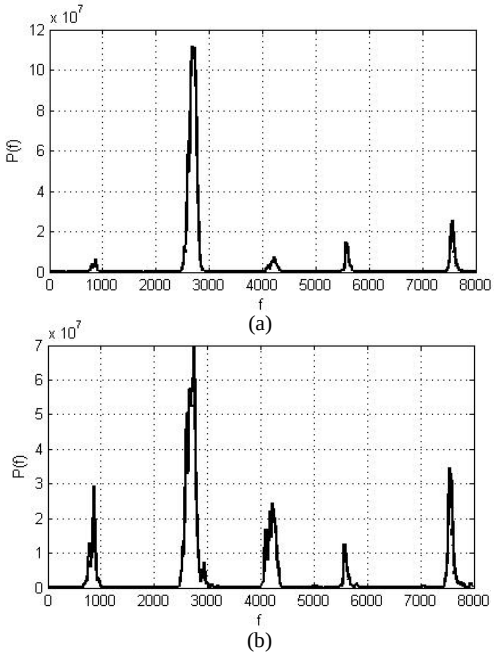

Figure 2. The spectrogram of sixth order of the DNA area: (a) – before the replacement of codons to synonyms; (b) – after the replacement of codons to synonyms

Smoothened spectrograms of the second and sixth orders after the replacement of codons by synonyms are shown in Fig. 3.

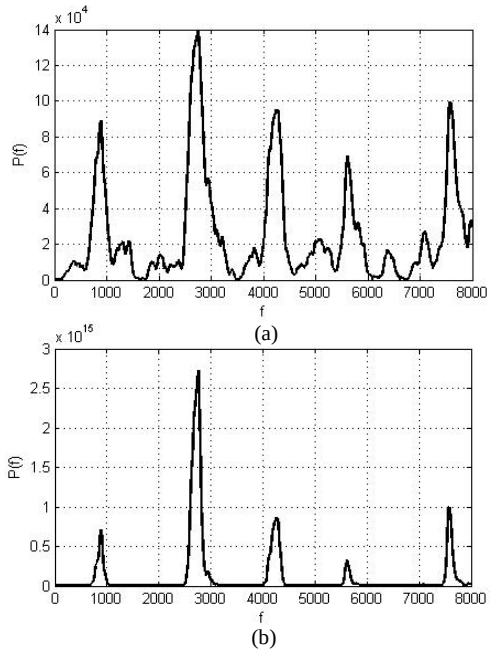

Figure 3. Smoothen spectrograms after the replacement of codons to synonyms of data: (a) – for spectrogram on Fig. 1; (b) – Fig. 2(b)

Clipped spectrograms after the replacement of codons by synonyms of sixth order with thresholds are shown in Fig. 4.

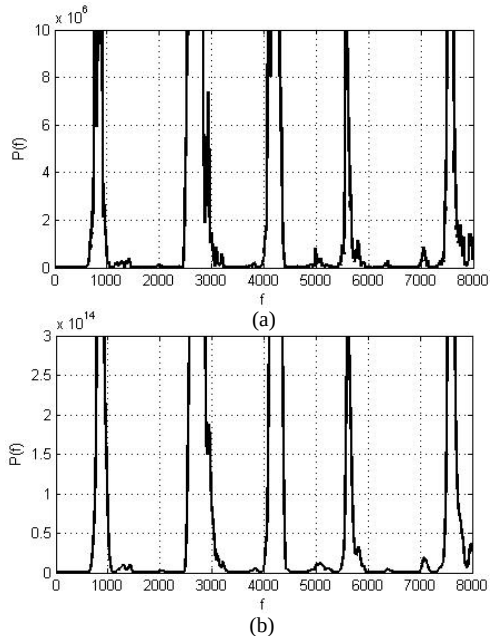

Figure 4. Clipped spectrograms and smoothen clipped spectrograms after the replacement of codons to synonyms of data with thresholds when p = 6: (a) – clipped spectrograms with threshold 10^7 when p = 6; (b) – smoothen clipped spectrograms with threshold $3 \cdot 10^{14}$ when p = 6

As it shown in Fig 2(a), when the spectrogram's order is higher the strong growth of the second spike substantially reduces observability of the other spikes. Improving the observability of all spikes and, therefore, of coding DNA regions, is possible if spectrum is limited by the level.

$$X_1 = \left| aX_A + tX_T + cX_C + gX_G \right|.$$

When spectrum's order is rising, we can see an equal rise of all spectrogram's spikes with respect to the noise level (Fig. 5).

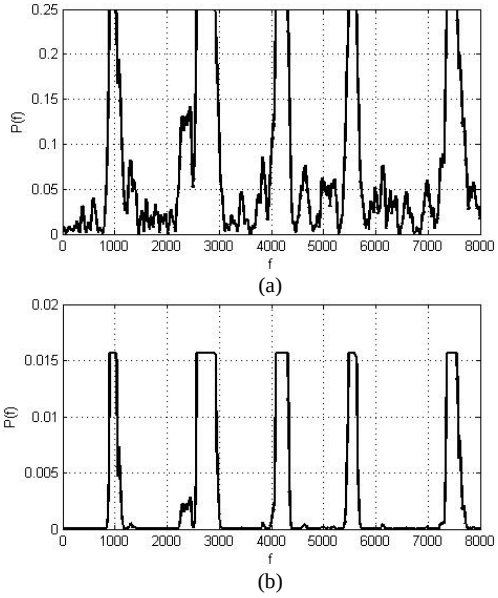

Figure 5. Smoothed clipped spectrograms of data for: (a) – p = 2; (b) – p = 6.

To evaluate the accuracy of the measures in predicting the coding regions employed the Signal-to-Noise Ratio (SNR)

$$\mathrm{SNR} = \frac{\sum \mathrm{P_e}}{\sum \mathrm{P_i}},$$

where ΣP_e – the sum of spectrogram's counts of coding DNA regions, ΣP_i – the sum of spectrogram's counts of non-coding DNA regions.

Using spectrograms clipped by levels does not influence fundamentally on the coding regions identification when the spectrum's order is rising. However, such nonlinear transformation that includes clips and powering significantly increases signal-to-noise ratio. Signal-to-noise ratio depends on the clip's level. SNR grows when clip's level is growing, but when clip's level of normalized by maximum spike of a spectrogram is more than 0.7, the level of second spike is significantly higher than the rest four spikes' levels of the spectrogram. Therefore, the high SNR appear only due to high level of the second spike. This significantly decreases possibility of the coding regions identification. Dependencies

of SNR on the spectrum's order using clips 0.4, 0.6, 0.7 are shown on Fig. 6. According to the Fig. 6, SNR is as high as the *k* degree. The accuracy of coding regions identification does not depend on the clip's level. When the spectrum's order rises the accuracy of the coding regions bounds does not change very much as it is shown on the Fig. 7. Therefore, we can divide the task of the coding regions identification in two stages. The first stage is detection of the coding regions using spectrograms of higher orders. The second stage is estimation of coding regions' limits using spectrograms of the second and higher orders. A similar approach is used in radiolocation, where the task of signal's detection and the task of parameter's estimation of a radar object are solved using different devices and methods.

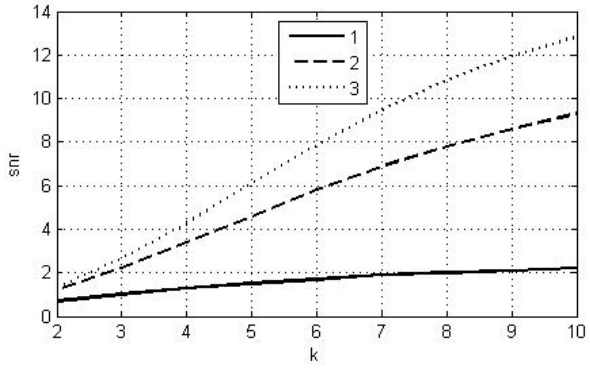

Figure 6. Dependence of SNR from spectrum's order clipped by levels: 1 – 0.4; 2 – 0.6; 3 – 0.7

As it shown in figure, when the model's order is high, the SNR is higher. Smoothing using AR filter does not influence on SNR value.

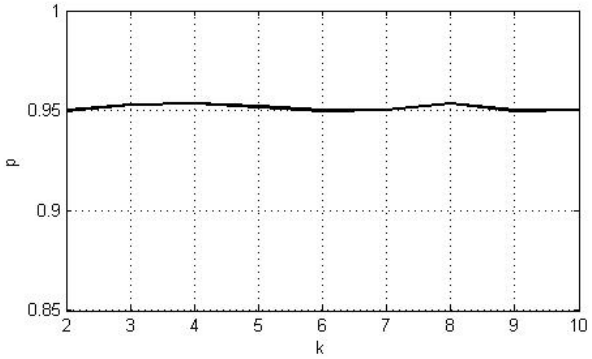

Figure 7. Dependence of parameter p from spectrum's order.

v. Conclusions

Using the higher orders spectra for spectral analysis it is possible to raise spectrograms' level of the coding regions and to reduce spectrograms' level of non-coding regions. In the paper the spectrograms of the sixth order was used. As it is shown in the figures, the most effective spectrograms we received on the base of the sixth order spectra. Using smoothening AR filter allows to decrease fluctuation of the spectrograms and to intensify the spectrograms on *N / 3* frequencies. DNA sequence is an additive process and, therefore, it is more efficient to use expression (4) for additive process's spectral estimation with coefficients from [4]. This expression gives more precise spectral estimations than that of the traditional parametric spectral methods of spectral estimation. Using of smoothed clipped spectrograms of higher orders improves observability of the coding regions.

Acknowledgment

The research was supported by the Internal Grant Agency of University of Pardubice, the project No. SGFEI 02/2014.

References

[1] M. Eigen, W. Gardiner, P. Schuster and R. Winkler-Oswatitsch, "The origin of genetic information", Sci. Am., 1981, pp. 88-96.

[2] E. N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences", Physica A 249, 1998, pp. 511-516.

[3] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", Physical Review Letters, vol. 68, pp. 3805-3808, 1992.

[4] D. Anastassion, "Genomic signal processing", In: IEEE Signal processing magazine, 2001, pp.8-20.

[5] G. E. P. Box, G. M. Jenkins, "Time series analysis", Forecasting and control, San Francisco, Calif: Holden Day, 1976.

[6] V. A. Tykhonov, N.V. Kudriavtseva, "Spectral analysis of adjoined linear prediction model for Non-Gaussian processes", In: Radio Electronics and Informatics Journal, 1 (48), KhNURE, 2010, pp. 35-37.

[7] V. A. Tykhonov, N.V. Kudriavtseva, I.O. Fil, "Interference rejection using filter based on additive linear prediction models," In: Eastern-European Journal of Enterprise Technologies, 1/4 (49), 2011, pp. 22-24.

[8] M. Rosenblatt, J.S. Van Ness, "Estimation of the bispectrum", Ann. Mathematical Statistics, No. 36, pp. 1120-1136, 1965.

[9] J. W. Tukey "An introduction of the measurement of spectra", Probability and statistics, Ed. U. Grenander, 1959, pp. 300-330.

About Authors:

My main scientific interests include the following areas: signal processing and time series analysis, statistical data analysis, linear prediction models, regression analysis, spectral data analysis, Gaussian and non-Gaussian processes, medicine, biostatistics, bioinformatics, Bayesian statistical modeling, statistics in economics.

The main scientific interests are connected to stochastic signals and process digital processing with aim of determination, recognition, and parameters estimation. With this purpose linear prediction models are widely used (AR, MA, ARMA, ARIMA).

I am currently PhD student at the University of Pardubice, Faculty of Electrical Engineering and Informatics, Department of Electrical Engineering. I am interesting in image processing, space mapping and microcontrollers.