УДК 577.323

# MULTISCALE PROPERTIES OF DNA PRIMARY STRUCTURE: CROSS-SCALE CORRELATIONS*

***M.V.Altaisky, V.V.Ivanov, R.V.Polozov*****

Cross-scale correlations of wavelet coefficients of the DNA coding sequences are calculated and compared to that of the generated random sequence of the same length. The coding sequences are shown to have strong correlation between large and small scale structures, while random sequences have not.

The investigation has been performed at the Laboratory of Information Technologies, JINR.

## Многомасштабные свойства первичной структуры ДНК: межмасштабные корреляции

***М.В.Алтайский, В.В.Иванов, Р.В.Полозов***

Исследованы межмасштабные корреляции вейвлет-коэффициентов, вычисленных для различных кодирующих последовательностей *E.coli*. Произведено сравнение структуры полученных корреляций с корреляциями таких же коэффициентов, вычисленных для случайных последовательностей той же длины. Обнаружено, что у реальных последовательностей имеются выраженные корреляции между вейвлет-коэффициентами различных масштабов, в то время как случайные последовательности таких корреляций не имеют.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.

## 1. INTRODUCTION

To understand the structural organization of genetic sequences is one of the challenging problems in molecular biology. The most nontrivial problem concerns the detailed analysis of primary structures of DNA sequences including the identification of hidden patterns inside these structures and the comparative study of primary structures in connection to their function and the evolutionary origin.

*The work was partially supported by the Commission of the European Community within the framework of the EU-RUSSIA Collaboration under the ESPRIT project CTIAC-21042, and by the Russian Foundation for Basic Research, grant 97-01-01027. The support of the Solvay Institute is appreciated.

**ITEB RAS, Pushchino, Russia

The DNA sequences written in 4-letter alphabet {A,T,C,G} look, at first glimpse, like random. Despite some universal regularities — the triplet code, the excess of CG over TA nucleotides and certain relations between triplets (codons), — it is unknown what exactly is written in $64/20$ redundant triplet code.

There are a lot of indications on the existence of long-range correlations in DNA sequences (see [1–5] and references therein). These correlations may indicate that different parts of nucleotide sequences are causally connected and may have evolved from the same parts of the pre-DNA [6, 7]. To reveal the hidden structures of DNA sequences we need a method which can analyze local structures and, at the same time, relate these structures to the whole sequence. The Fourier analysis, which is essentially nonlocal, does not match these requirements.

If the DNA sequences of the present organisms have really originated from short pre-DNA sequences, it should have been some multiplicative process on each stage of evolution, by which the length is increased to the present length and the new information encoded. At present, we can't see these archaic pre-DNA sequences, but we can look for the hierarchical structure of the hypothetical multiplicative process, that has resulted in present DNA structure.

The wavelet transform, due to its self-similar structure, is capable of revealing the hierarchical (tree-like) fragmentation processes using only the final distribution (the present nucleotide sequence, that is the final result of the evolution, in our case) as an input. The ability of wavelet transform to reveal such structures has been shown many times, for the «devil staircase» measure, for the hydrodynamic turbulence, and also for DNA sequences; scaling in DNA sequences is also known [8–10].

The scaling (self-similarity) itself, if observed, does not tell us whether or not the sequence carries some information or is random. The random walk (Brownian motion) is self-similar. It has *global scaling*, and there are no visual differences between its oblique at zoom window 100, 1000, or 10000 time steps. In other situation, the scaling exponents may be scale-dependent themselves, e.g., $l^{\zeta}(l)$, the multifractal law of hydrodynamic turbulence [8, 14]. Thus, the presence of certain scaling law, just indicates the presence of multiplicative processes of certain class.

However there are principal differences between scaling in chaotic systems, like turbulence, and scaling in DNA sequences. First, in DNA the scaling laws were found to be position dependent [10]. The presence of local scaling means that different fragments of a given DNA sequence may have originated from different starting points of the predecessor, the nucleotide structure present on some previous stage of evolution. The roots of such processes, «forks», are visible on two dimensional wavelet plots calculated for DNA sequences [9]. Second, in dynamic chaotic systems, say in Kolmogorov turbulence, the scaling is important itself: there is an overaging over all possible configurations (phase space integration) at each scale, the *statistical* self-similarity is the main thing. At the same time, the large and small scale structures could be simultaneously observed (at least in principle) for chaotic processes. Thus the correlations between large and small scale structures are physically measurable. For instance, it is possible to measure the velocities of small vortices within a large one, and compare the structure of calculated wavelet coefficients with the velocity field really observed at both small and large scales. In case of the DNA sequence the only object we have to deal with is the *«small scale structures»*, long nucleotide sequences of present DNA, originated by means of some multiplicative process from short pre-DNA sequences (*«large scale structure»*). But we do not have this «large scale structure» at our disposal!

In the present paper we exploit the ability of wavelet analysis to reveal structure properties of the multiplicative process which resulted in given samples, DNA sequences, by studying the correlations of wavelet coefficients of different scales [11]. If a sequence is random, the wavelet coefficient correlation function will coincide with that of random signal, if no, the structure of wavelet coefficient correlation function will be different. In particular, this technique has been applied to different coding sequences taken from the full *E.coli* genome. The coding sequence was found to be different from randomly generated sequence of the same length by the presence of modulated correlations between small and large scales.

## 2. METHOD

Let us start with definitions. The convolution of function $f(t) \in \mathrm{L}^2(\mathrm{R})$ with a certain locally supported function $g(t)$ shifted and dilated is called the wavelet transform (WT) of $f$

$$W_g(a,b)[f] := \int \frac{1}{\sqrt{a}} g\left(\frac{t-b}{a}\right) f(t)dt. \tag{1}$$

Referring the reader to [15] for a general review on wavelets, we just mention that WT is a straightforward generalization of the Fourier transform. While the Fourier transform is a decomposition of a function with respect to the translation group $G : x' = x + b$, the wavelet transform is a decomposition with respect to the affine group $G : x' = ax + b$ [16], where $a$ is a scale parameter. This new parameter provides different window width (see Eq. 1) for different scales and, therefore, provides a local resolution-dependent analysis. Very often the «Mexican hat wavelet»

$$g_2(x) = (1 - x^2)\exp(-x^2/2)$$

is used as a basis in (1).

To analyze the nucleotide sequences we have first to digitize a symbolic sequence written in 4-letter nucleotide alphabet {A,T,C,G}, which stands for adenine, thymine, guanine and cytosine. In present paper we use the DNA walk mapping : A,G $\to 1$; T,C $\to -1$, as in [17], and the alternative one A,T $\to 1$; C,G$\to -1$. The former regards if purine or pyrimidine occures in certain position, the later is its complement with regard to the 4-letter alphabet. Our observation shows that these two mappings are not completely equivalent: the sequence that seems random in the former coding may have correlations in the latter. (More details on this subject are presented in Appendix).

For completely random sequence a Brownian motion type signal is expected.

To illustrate the method we present the path mapping (Fig. 1), the Fourier transform (Fig. 2), and $g_2$ wavelet transform (Fig. 3) of the recA *E.coli* coding sequence, GenBank accession number V00328 [12, 13].

The tree-like structure displayed in Fig. 3 obviously reassembles a branching process, like that of one-third Cantor set construction.

In fact, there is the hypothesis that modern DNA sequences have been originated from short (a few nucleotides in length) primodal sequences [6]. This branching processes may be

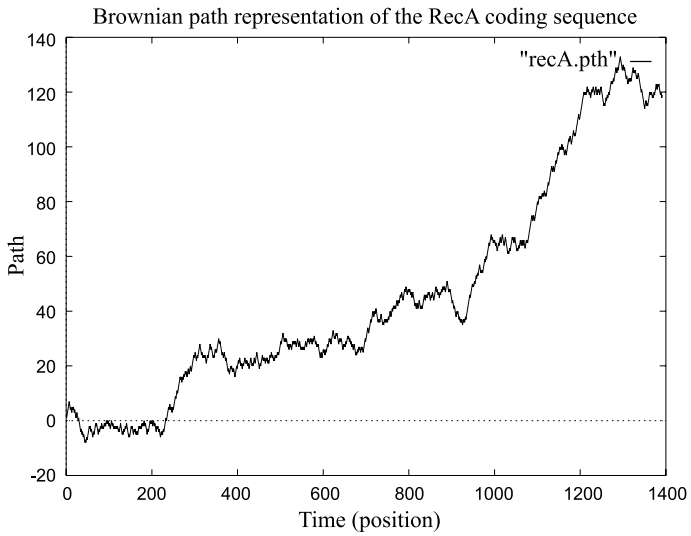


Fig. 1. Path mapping for the recA *E.coli* coding sequence. A,G $\to 1$; T,C $\to -1$ coding is used

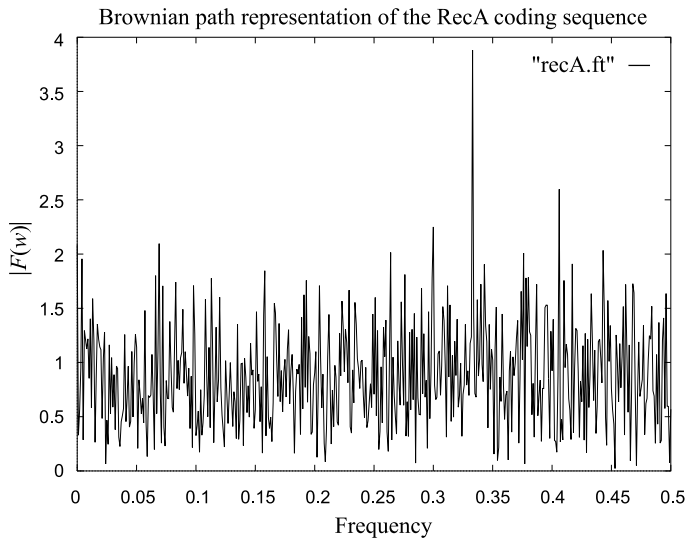


Fig. 2. Modulus of Fourier image for the recA *E.coli* coding sequence fragment of 1000 bp. A,G $\to 1$; T,C $\to -1$ coding is used normalized to frequency (inverse period) $\Delta f = 1/T$. Maximum at $f = 0.3$ corresponding to the triplet code periodicity is observed

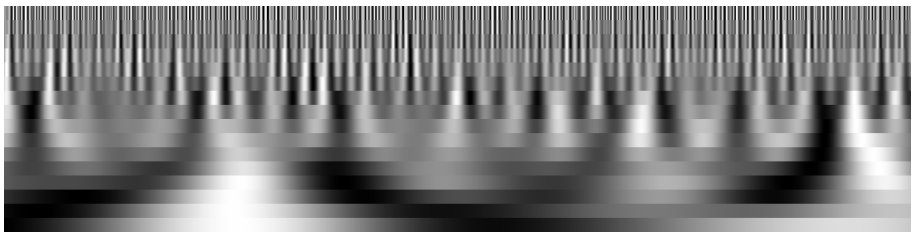

Fig. 3. The shadow plot of the recA *E.coli* coding sequence $g_2$ wavelet coefficients. A,G $\to$ 1; T,C $\to -1$ coding is used

visually traced at the shadow plot in Fig. 3. To study the matter quantitatively we calculate the correlations between wavelet coefficients

$$R(a_1, a_2, b_1 - b_2) = \langle W_g(a_1, b_1) W_g(a_2, b_2) \rangle \tag{2}$$

at different scales. The curly brackets $\langle\rangle$ mean the covariance

$$\mathbf{cov}(W_1, W_2) := \mathrm{E} \frac{(W_1 - \mathrm{E}(W_1))(W_2 - \mathrm{E}(W_2))}{\sqrt{\mathrm{D}W_1 \cdot \mathrm{D}W_2}},$$

where D is the dispersion and E is the mathematical expectation.

## 3. RESULTS

In Figs. 4,5 we present wavelet cross-scale correlations $R(a_1, a, b)$ calculated for the recA *E.coli* coding sequence and in Figs. 6, 7 — for a random sequence of the same length as recA *E.coli* coding sequence (The sequence is random with equal probability of $1/4$ for all «nucleotides» and there is no need to use different mappings to code it). To ensure, that the cross-scale correlations (the nucleotide sequences are seen to display) are not induced by some periodicity of the sequences, we have also simulated the random sequences *with periodic modulation* and we were not able to reproduce the modulation of wavelet coefficient cross-scale correlation function at relatively small scales ($< 100$ bp) as observed for DNA sequences. In the right of Fig. 6 we present the wavelet coefficients cross-correlation plot obtained for a random work with the period of harmonical modulation $t = 20$.

In all pictures the $X$-axis corresponds to $a_1$ scale in Eq. 2 notation; the $a_2$ scale is taken fixed $a_2 = (\sqrt{2})^{14} = 128$. The $Y$-axis is the position lag $b = b_1 - b_2$.

The landscapes of Figs. 4–7 are different. The difference is not very striking, but the wavy form of the left edge of the plots for coding sequences clearly shows the correlation between large and smale scales. While for random sequence plot has no special modulation. The left edges of Figs. 4, 5 are not a plane, and the correlations are clearly traced to the higher scales. The difficulty with getting more clear difference is due to *relatively short length of coding sequences*, of about 1000-2000 bp. We have made simulation for longer random sequences (2000 bp and longer), and for that case the left (small-scale) edge of the wavelet cross-scale correlation plots are quite plain, with no visible modulation.

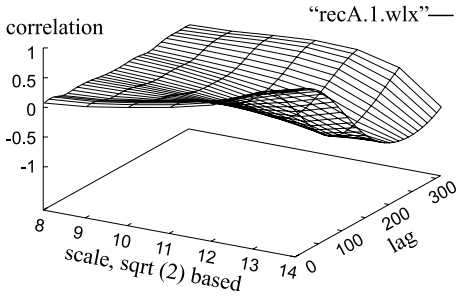


Fig. 4. The plot of $g_2$ wavelet coefficients cross-scale correlations for recA *E.coli* coding sequence. A,G $\to 1$; T,C $\to -1$ coding is used. Calculated for 8-15 layers at the base $2^{1/2}$

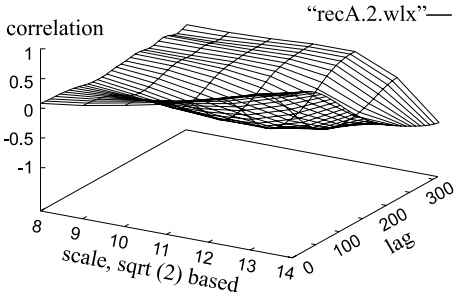


Fig. 5. The plot of $g_2$ wavelet coefficients cross-scale correlations for recA *E.coli* coding sequence. A,T $\to 1$; C,G $\to -1$ coding is used. Calculated for 8-15 layers at the base $2^{1/2}$

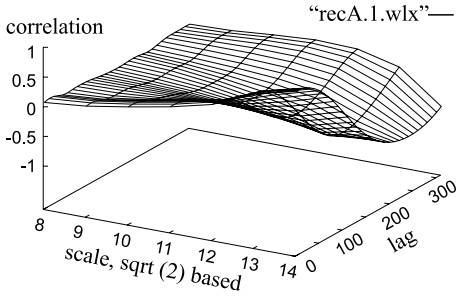


Fig. 6. The plot of $g_2$ wavelet coefficients cross-scale correlations for the randomly generated nucleotide sequence. Calculated for 8-15 layers at the base $2^{1/2}$

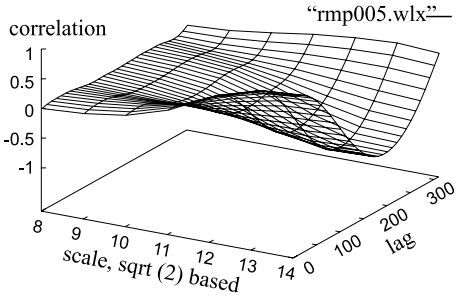


Fig. 7. The plot of $g_2$ wavelet coefficients cross-scale correlations for the randomly generated nucleotide sequence with sinusoidal modulation $\sin(2\pi 0.05n)$. Calculated for 8-15 layers at the base $2^{1/2}$

To some extent, we can say that local distribution of the nucleotides in coding sequences, «knows» which macro-block it lives in. The macro-blocks, revealed by the wavelet analysis, may be considered as the imprintings of prenucleotides at the level of the present DNA structure.

To check the effect, we have also done the same calculations for more than twenty coding sequencies taken from the *E.coli* genome, GenBank accession number U00096 [18]. The typical effect — the waving of the low scale edge of the wavelet coefficient cross-scale correlation plot is observed for most of these sequences. For some sequences it is manifested even more strongly, than for the coding sequence we have used for Figs. 4,5.

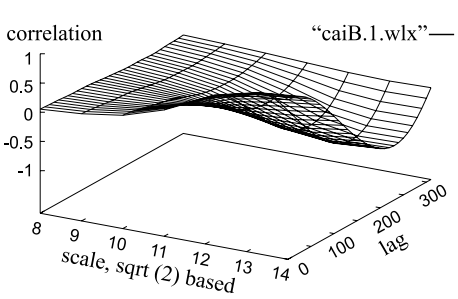


Fig. 8. The plot of $g_2$ wavelet coefficients cross-scale correlations for caiB *E.coli* coding sequence. A,G $\to 1$; T,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

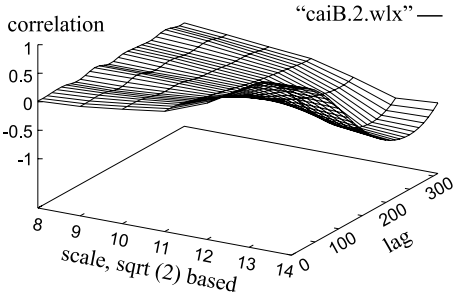


Fig. 9. The plot of $g_2$ wavelet coefficients cross-scale correlations for caiB *E.coli* coding sequence. A,T $\to 1$; C,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

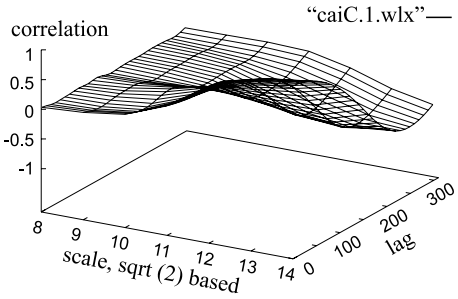


Fig. 10. The plot of $g_2$ wavelet coefficients cross-scale correlations for caiC *E.coli* coding sequence. A,G $\to 1$; T,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

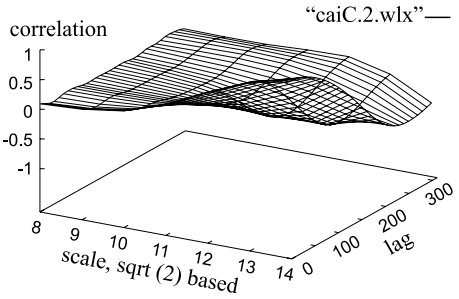


Fig. 11. The plot of $g_2$ wavelet coefficients cross-scale correlations for caiC *E.coli* coding sequence. A,T $\to 1$; C,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

Here below we present 6 plots of the wavelet coefficients cross-scale correlations. To get more complete information we used two alternative mappings

$$1: \quad AG \to 1 \quad TC \to -1$$
$$2: \quad AT \to 1 \quad CG \to -1$$

The plots presented in the second coding seems to display more structural information about cross-scale correlations. The biological relevance of this observation, that the nucleotides in pairs (AG — purines, TC — pyrimidines) are unlikely to duplicate each other in the sense of information content. Therefore the second coding seems more informative (alas, both are «random» unless we know the genetic code exactly).

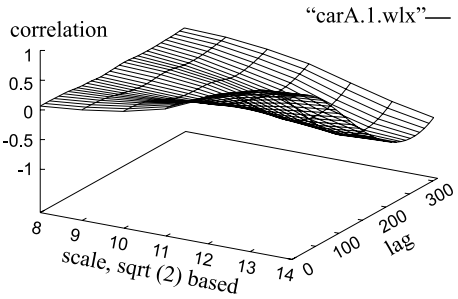


Fig. 12. The plot of $g_2$ wavelet coefficients cross-scale correlations for carA *E.coli* coding sequence. A,G $\to 1$; T,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

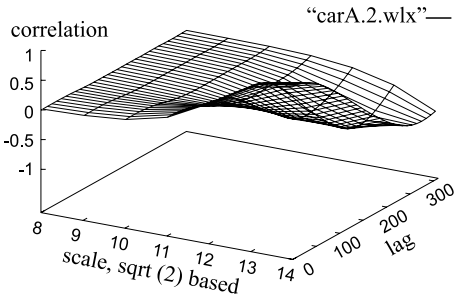


Fig. 13. The plot of $g_2$ wavelet coefficients cross-scale correlations for carA *E.coli* coding sequence. A,T $\to 1$; C,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

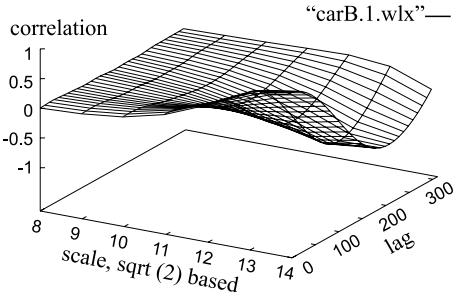


Fig. 14. The plot of $g_2$ wavelet coefficients cross-scale correlations for carB *E.coli* coding sequence. A,G $\to 1$; T,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

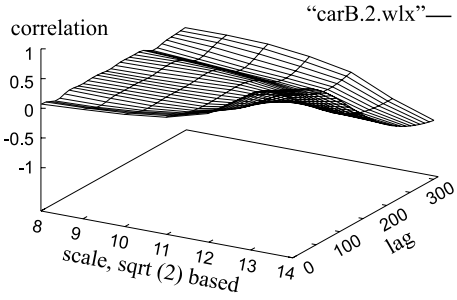


Fig. 15. The plot of $g_2$ wavelet coefficients cross-scale correlations for carB*E.coli* coding sequence. A,T $\to 1$; C,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

**3.1. Plots of Wavelet Coefficients Cross-Scale Correlations.** All pictures in the left column are taken with respect to the first coding $AG \to 1, \quad TC \to -1$; pictures in the right column are taken with respect to the second coding $AT \to 1, \quad CG \to -1$.

All coding sequences were taken from the same *E.coli* genome [18]. It is clearly seen that only the left edge of plot in Figs. 16 and 12 are visually as flat as the corresponding landscape for the random sequence shown in Fig. 6 (left). All other plots display wavy surface at the left edge, which means the lag ($b = b_1 - b_2$) varying correlations between small- ($a_1$) and large- ($a_2$) scale wavelet coefficients. The typical (lag) period of these variations is visually much less than 100 bp, and seems to be about 15–30 bp or so.

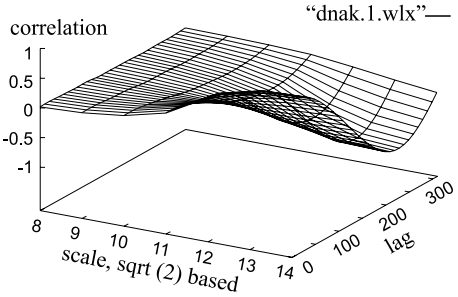


Fig. 16. The plot of $g_2$ wavelet coefficients cross-scale correlations for dnak *E.coli* coding sequence. A,G $\to 1$; T,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

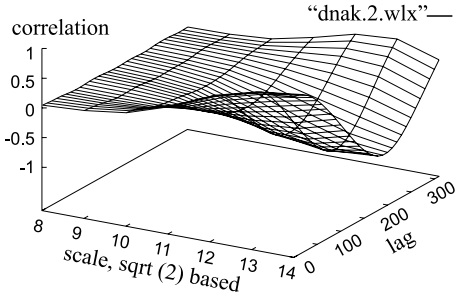


Fig. 17. The plot of $g_2$ wavelet coefficients cross-scale correlations for dnak *E.coli* coding sequence. A,T $\to 1$; C,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

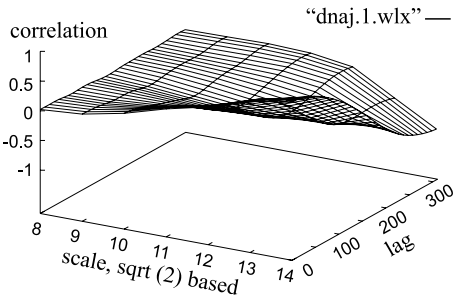


Fig. 18. The plot of $g_2$ wavelet coefficients cross-scale correlations for dnaj *E.coli* coding sequence. A,G $\to 1$; T,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

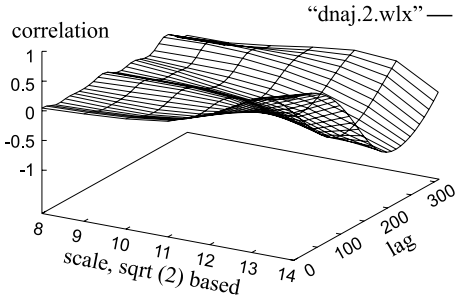


Fig. 19. The plot of $g_2$ wavelet coefficients cross-scale correlations for dnaj*E.coli* coding sequence. A,T $\to 1$; C,G $\to -1$ coding is used. Calculated for 6-14 layers at the base $2^{1/2}$

## 4. CONCLUSION

The method for the nucleotide sequences analysis based on the wavelet transform is proposed. In the present contribution we show that the cross-scale correlations of wavelet coefficients for the DNA coding sequences have strong correlation between large and small scale structures, while random sequence have not. This feature can be used to classify the nucleotide sequences and to study their functional organization.

## References

1. Peng C.-K. et al. — Nature, 1992, v.356, p.168.

2. Viswanathan G.M. et al. — Biophys. J., 1997, v.72, p.866.

3. Voss R.F. — Phys. Rev. Lett., 1992, v.68, p.3805.

4. Li W., Kaneko K. — Europhys. Lett., 1992, v.17, p.655.

5. Lee W.J., Luo L.F. — Phys. Rev. E, 1997, v.56, p.848.

6. Ohno S. — Proc. Natl. Acad. Sci. (USA), 1988, v.85, p.4378.

7. Yomo T., Ohno S. — Proc. Natl. Acad. Sci. (USA), 1989, v.86, p.8452.

8. Arneodo A. et al. — Ondelettes, multifractales et turbulence de I'ADN aux crosissance cristallines, Diderot Editeur, Paris, 1995.

9. Tsonis A.A. et al. — Phys. Rev. E, 1996, v.53, p.1828.

10. Altaisky M., Mornev O., Polozov R. — Genetic Analysis: Techniques and Applications, 1996, v.12, p.165.

11. Houdré C. — In: Wavelets, Mathematics and Applications, ed. by J.J.Benedetto and M.W.Frazier, CRC Press Inc, 1994.

12. Sancar A. et al. — Proc. Natl. Acad. Sci. (USA), 1980, v.77, p.2611.

13. Horii T., Ogawa T., Ogawa H. — Proc. Natl. Acad. Sci. (USA), 1980, v.77, p.313.

14. Frisch U. — Turbulence, Oxford, 1993.

15. Daubechies I. — Ten Lectures on Wavelets, SIAM, Philadelphia, PA, 1992.

16. Grossmann A., Morlet J., Paul T. — Math. Phys., 1985, v.26, p.2473.

17. Stanley H. et al. — Physica A, 1992, v.191, p.1.

18. Blattner F.R. et al. — Science, 1997, v.277:5331, p.1453; GenBank accession number U00096.

Received on April 5, 2000.