



ELSEVIER

Physica D 146 (2000) 388–396

PHYSICA D

www.elsevier.com/locate/physd

Correlations in DNA sequences across the three domains of life

Sabyasachi Guharay^{a,*}, Brian R. Hunt^b, James A. Yorke^b, Owen R. White^c

^a Princeton University, Princeton, NJ 08544, USA

^b Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA

^c The Institute for Genomic Research, Rockville, MD 20850, USA

Received 1 June 1999; received in revised form 21 July 2000; accepted 21 July 2000

Communicated by J.D. Meiss

Abstract

We report statistical studies of correlation properties of ~ 7500 gene sequences, covering coding (exon) and non-coding (intron) sequences for DNA and primary amino acid sequences for proteins, across all three domains of life, namely Eukaryotes (cells with nuclei), Prokaryotes (bacteria) and Archaea (archaeobacteria). Mutual information function, power spectrum and Hölder exponent analyses show exons with somewhat greater correlation content than the introns studied. These results are further confirmed with hypothesis testing. While $\sim 30\%$ of the Eukaryote coding sequences show distinct correlations above noise threshold, this is true for only $\sim 10\%$ of the Prokaryote and Archaea coding sequences. For protein sequences, we observe correlation lengths similar to that of “random” sequences. © 2000 Elsevier Science B.V. All rights reserved.

PACS: 87.10.+e; 89.70.+c; 89.90.+n; 87.15.-v

Keywords: DNA correlations; Protein sequences; Statistical genetics; Mathematical biology; Introns versus exons; Three domains of life; Long-range correlations

1. Introduction

Statistical analysis of gene sequences has been a subject of considerable recent interest. In particular, efforts to compare statistical properties of the coding (exons) and the non-coding (introns) regions of DNA are being pursued. It is well known that the coding regions are translated into proteins. However, the origin of the non-coding regions of DNA is still unclear. A basic question arises whether they developed from the coding regions. Whether the introns and exons have similar or different statistical properties is a

critical data point for any theory to explain the origin of introns.

Different techniques, such as formulae derived from Shannon's entropy [1–7], spectral analysis [1,2,8–10], detrended fluctuation analysis [11,12], wavelet transformations [13–15], entropic segmentation and fractal structures [16,17] and random walk analysis [18–21], have been used to determine the information content in gene sequences. Long-range correlations were reported to exist in introns, while exons were indistinguishable from random sequences [1–2,11–15,18]. Azbel [21] claimed that no long-range correlations exist in both coding and non-coding DNA sequences. In the effort to examine any distinction in correlations between introns and intronless DNA, the values of the exponent of a power law, as used in [8,18], were

* Corresponding author. Address: 8453 Bauer Road #32, Springfield, VA 22152, USA

E-mail address: sguharay@princeton.edu (S. Guharay).

evaluated for both types of sequences, and the fits were done on a wide range of window sizes of the nucleotides. They concluded that no consistent difference in the value of the exponent could be found for the intron-containing versus the intronless sequences [22–23]. Therefore, further work is warranted to clarify the apparent contradictions in the above results.

The goal of this work is to investigate the similarities or differences in the statistical properties of introns and exons using large data sets from various specimens and to obtain a more comprehensive picture of the problem. Our paper examines the correlation content of biological sequences derived from high-throughput genome sequences across all the three domains of life, namely, Eukaryotes, Prokaryotes and Archaea. We determine the correlation content using several different mathematical techniques, namely, mutual information function, symbolic power spectrum, and Hölder exponent. This approach minimizes the bias in any particular method. We have performed a statistical test, known as hypothesis testing, on the results obtained from the above mathematical methods to strengthen their validity. In addition to comparing the correlation content in coding and non-coding DNA sequences in the Eukaryote domain, the following questions are addressed in this paper: (1) How does the correlation content vary, if at all, between the coding regions of the Archaea and Prokaryote domains and the coding regions of the Eukaryote domain? (2) Is there any distinction in the correlation content of coding DNA sequences, composed of the four nucleotides, namely, Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), with respect to their corresponding protein sequences consisting of the 20 amino acids, namely, Alanine (A), Leucine (L), etc.?

Section 2 describes the gene data sets we studied. Section 3 explains the mathematical methods used to measure the degree of correlation in the gene sequences. In Section 4, we describe the hypothesis testing. The results and discussion are presented in Section 5. Finally, the conclusions are narrated in Section 6.

2. Gene data sets studied

In order to achieve high confidence in the statistical results, we restrict our investigation to “long exons”, “long introns” and “long proteins”, namely those of length ≥ 500 nucleotides for DNA based sequences and ≥ 500 amino acids for protein sequences. Non-redundant genes are selected from GenBank. Overall, ~ 7500 intron sequences as well as exons and their corresponding protein sequences are studied for the following specimens: (a) non-coding sequences for chicken, fly, fungus, human, mice, primate (non-human) and worms, (b) coding sequences for Archaea, Prokaryotes, chicken, fly, fungus, human, mice, primate (non-human), and worms.

3. Mathematical measures of correlation content

3.1. Mutual information function

The mutual information function (MIF) is a statistical measure of the correlation in a symbolic sequence. This method is derived following Shannon’s definition of entropy [3]. The correlation length calculated from MIF measures how the auto-correlation function of a symbolic sequence changes with distance. MIF is defined as

$$M(d) = \sum_{a,b} P_{a,b}(d) \log \left(\frac{P_{a,b}(d)}{P_a P_b} \right), \quad (1)$$

where P_a and P_b are the probability of occurrence of the symbols a and b , respectively, and $P_{a,b}(d)$ measures the joint probability of occurrence of a and b at a separation of d .

First, “random” sequences, with the same length and same percentage of constituent symbols as the gene sequence to be studied, are generated. We compute $M(d)$ for both “random” and gene sequences following Eq. (1), and these MIF results are compared with each other. The smallest value of d for which $M_{\text{Gene}}(d) \leq M_{\text{Random}}(d)$ gives the correlation length d_c . In order to increase the confidence level of the correlation estimates, five different “random” sequences, with varying initial seeds, are generated for each gene

sequence. Therefore, we obtain five values of d_c . We define the correlation length of a gene sequence to be the average of the five values of d_c .

In order to determine a meaningful value of d_c for a gene sequence, we determine the approximate average correlation length of random sequences. For two random sequences, r_1 and r_2 , the probability that $M_{r_1}(1) \leq M_{r_2}(1)$ is 0.5. Assuming statistical independence, the probability that both $M_{r_1}(2) \leq M_{r_2}(2)$ and $M_{r_1}(1) \geq M_{r_2}(1)$ is 0.25. Therefore, it follows that the probability that $d_c = k$ is approximately 0.5^k . The expected value of d_c for two random sequences is then approximately

$$0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + \dots + k(0.5)^k + \dots, \quad (2)$$

which converges to 2. Thus, random sequences have an average correlation length approximately equal to 2. One can show in this fashion that if we average over five random sequences, there is less than 5% probability that the average value of d_c is greater than 4. Thus, from the viewpoint of correlation length the criterion for non-randomness as $d_c \geq 5$ will have a statistical confidence level of greater than 95%. Therefore, we have used $d_c = 4$ as the noise threshold for MIF studies. Numerical results in Section 5.1 also support this claim.

3.2. Symbolic power spectrum

To calculate the power spectrum for gene sequences it is necessary to first transform them into a numerical sequence. We symmetrically assign a vector to each of the four bases A, C, G, and T. The four vectors are equidistant from each other, and their sum is zero. We can now construct a regular tetrahedron using these vectors. While a 3D space is the lowest possible dimension to describe a tetrahedron, it is much simpler, in the present context, to express the coordinates of a regular tetrahedron in 4D space. For example, the following four vectors are equidistant from each other: (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1). To make these vectors sum to the zero vector (0, 0, 0, 0), we subtract 0.25 from the value of each coordinate of the above four vectors. Thus, the fol-

lowing coordinates are obtained for the nucleotides A, C, G and T: A = (0.75, -0.25, -0.25, -0.25); C = (-0.25, 0.75, -0.25, -0.25); G = (-0.25, -0.25, 0.75, -0.25); and T = (-0.25, -0.25, -0.25, 0.75). Using these vectors, the power spectrum is calculated as follows:

$$P(f) = \sum_{c=1}^4 \frac{1}{N} \left| \sum_{j=0}^{N/2} \chi_c(j) e^{[-i2\pi(f/N)j]} \right|^2, \quad (3)$$

$$f = 0, \dots, \frac{1}{2}N.$$

Here, N is the sequence length and $\chi_c(j)$ the c th coordinate of the vector corresponding to the j th symbol. The unit of frequency f is 1 cycle per 512 symbols or nucleotides; note that similar definition was used in [2]. The above method for calculation of the power spectrum is independent of the choice of coordinates for a regular tetrahedron in 4D space.

3.3. Hölder exponent

We now address the Hölder exponent [24] calculations. Consider a function $f(x)$, where x is a real number, but $f(x)$ may be a number or a vector. Then, $f(x)$ is Hölder continuous at x_0 with exponent α if there exists a constant $c > 0$ such that $|f(x) - f(x_0)| \leq c|x - x_0|^\alpha$ for all x sufficiently close to x_0 . This leads to the following definition of the average Hölder exponent α for a continuous function $f(x)$ on a bounded interval as the limit (provided it exists):

$$\alpha = \lim_{h \rightarrow 0} \frac{\log \langle |f(x+h) - f(x)| \rangle}{\log h}, \quad (4)$$

where angle brackets $\langle \rangle$ denote average over x in the given interval. To calculate a meaningful Hölder exponent for DNA sequences, we determine the cumulative sum of the above vectors assigned to A, C, G, and T. Analogously, summing the vectors corresponding to a random sequence yields Hölder exponent of 0.5. On the other hand, the Hölder exponent of a smooth function, such as $\cos(x)$ is 1. The function $f(x)$, whose average Hölder exponent is to be determined, is defined as the following:

$$f(x) = \sum_{j \leq x} \mathbf{X}(j), \quad (5)$$

where $X(j) = (0.75, -0.25, -0.25, -0.25)$ if the j th symbol is A. For C, G, and T we use their corresponding vectors defined in Section 3.2.

Returning to Eq. (4), a least-squares fit is made to the plot of the numerator, which we call $H(h)$, versus the denominator, $\log h$, and the average Hölder exponent is calculated for $1 \leq h \leq n - 1$, where n is the sequence length.

4. Hypothesis testing using large sample theory

In order to strengthen our analysis of the results on Eukaryote introns and exons obtained from MIF, symbolic power spectrum and Hölder exponent, we perform hypothesis testing as described below.

For each of the above three methods, we define a characteristic value x for exons and y for introns. For example, in the case of MIF, x and y correspond to the correlation lengths for exons and introns, respectively. Let μ_x and μ_y correspond to the true mean characteristic value for exons and introns, respectively. We want to test the null hypothesis, $H_0 : \mu_x = \mu_y$, against the both-sided alternative hypothesis $H_1 : \mu_x \neq \mu_y$. To examine this hypothesis we use large sample theory [25]. According to this theory the test statistic value z is defined as

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_1 + s_y^2/n_2}}. \quad (6)$$

Here \bar{x} is the average of the characteristic exon data, \bar{y} the average of the characteristic intron data, s_x^2 the variance of the characteristic exon data set, s_y^2 the variance of the characteristic intron data set, n_1 the number of exons and n_2 the number of introns. When comparing H_0 versus H_1 , one must reject H_0 if $|z| > z_{\alpha/2}$. We choose here $\alpha = 0.05$ (95% confidence level) so that $z_{\alpha/2} = 1.96$.

5. Results and discussion

5.1. Mutual information function results

Sample MIF results for both coding and a non-coding DNA sequence in the Eukaryote domain

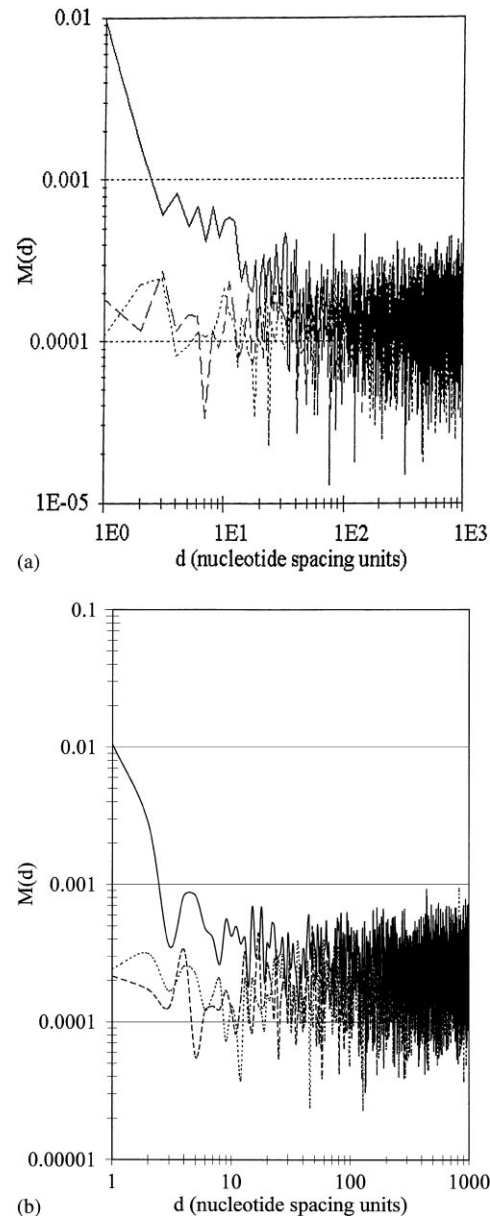


Fig. 1. Mutual information function ($M(d)$) versus d for sample Eukaryote DNA sequences. (a) The solid line represents the sample Eukaryote intron sequence. The long dashed line represents a “random” sequence of identical length and percentage of nucleotides as the sample intron. The dotted line represents a different “random” sequence with identical length and percentage of nucleotides as the sample intron. (b) The solid line represents the sample Eukaryote exon sequence. The long dashed line represents a “random” sequence of identical length and percentage of nucleotides as the sample exon. The dotted line represents a different “random” sequence with identical length and percentage of nucleotides as the sample exon.

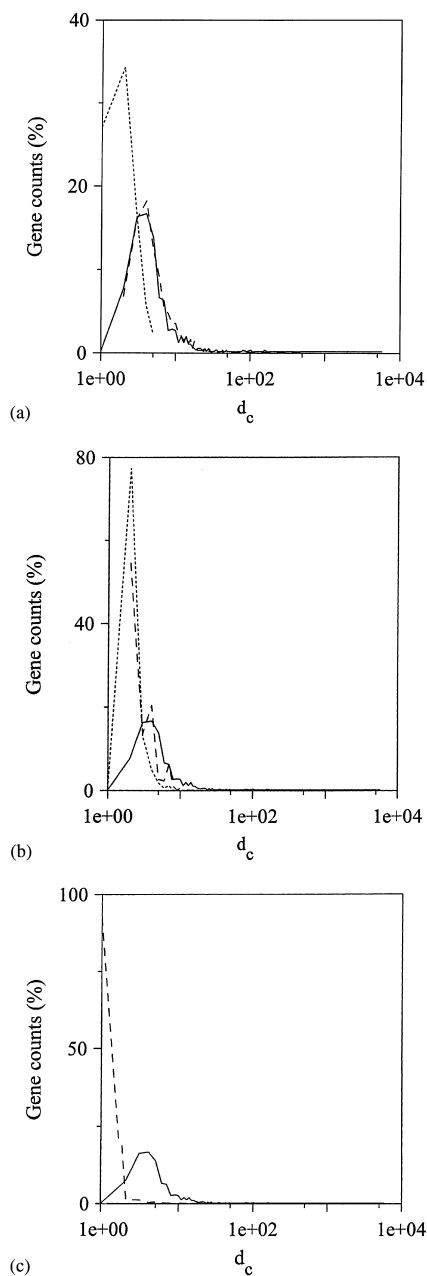


Fig. 2. Distribution of correlation lengths for DNA and protein sequences from mutual information function calculations. (a) The dotted line represents the random cases, and the long dashed line and the solid lines represent, respectively, sample introns and exons in the Eukaryote domain. (b) The dotted line represents the sample Prokaryote coding cases, the long dashed line represents the sample Archaea coding cases and the solid line represents sample Eukaryote exons. (c) The long dashed line represents the sample protein sequences and the solid line represents sample Eukaryote exons.

along with the corresponding MIF results for two “random” sequences are shown in Fig. 1(a) and (b). Notice how MIF of the “random” sequences in both figures behave almost identically to each other, in that, the two random sequences in each case intersect at values of $d < 4$. However, until $d \sim 20$, the MIF of the sample Eukaryote non-coding sequence behaves distinctly from the MIF of the two “random” sequences. Also, in the case of the coding sequence, until $d \sim 20$, the MIF of the coding sequence behaves distinctly from the two “random” sequences. This shows the non-random feature existing in the structural layout of the DNA sequences.

Some characteristic MIF results are displayed in Fig. 2 to show the distribution of correlation lengths in gene sequences. Fig. 2(a) shows the distribution of correlation lengths for specimens in the Eukaryote domain. Several characteristic cases are shown here, for example, fungal exons and human introns in contrast to “random” sequences. More than 95% randomly generated sequences have correlation lengths less than five symbols, i.e., $d_c < 5$. This agrees with the mathematical argument regarding the mean correlation length presented in Section 3.1. The distribution of correlation lengths for sample introns and exons overlap, by and large, with each other, and they are distinct from the random curve. For both sample intron and exon sequences, significant correlations ($d_c \geq 5$) is observed in $\sim 60\%$ of the cases. Longer correlation lengths, with maximum d_c ranging from ~ 10 to ≥ 500 , are noticeable in $\sim 24\%$ of the cases. This suggests that a high percentage of fungal exons and human introns have correlations well above the noise threshold. For the other exon and intron specimens studied, $^1 \geq 30\text{--}60\%$ of the cases shows correlation lengths of $d_c \geq 5$.

Fig. 2(b) shows the distribution of correlation lengths in the coding regions of Archaea, namely, *Methanococcus jannaschii*, and the coding regions of Prokaryotes, namely, *Haemophilus influenzae*. For

¹ Previous studies [4,5] of some specific exon sequences, primarily from yeast chromosomes, reported a period-three oscillation. Our results for these specific sequences also show some periodic oscillations. However in this paper, we have studied larger sample size with the goal to examine the overall statistical properties of both exons and introns.

Table 1
Mutual information function results of exons and introns and hypothesis testing

Case	Percentage with $d_c \geq 5$	Mean correlation length	Test statistic (z -value)	Result
Eukaryote exons	43	33	3.3	$ z > z_{\alpha/2}$; reject H_0
Eukaryote introns	42	10		

comparing the results among the three domains of life we also include the sample Eukaryote exons which were displayed earlier in Fig. 2(a). The distribution plots for both the Prokaryote and Archaea show a peak at $d_c \sim 2$, and then decreases until $d_c = 7$. This indicates that the structural layout in the coding regions of Archaea and Prokaryotes are, for the most part, similar to that of noise, while the sample Eukaryote exons have distinct properties. Therefore, a distinction in the correlation structures is observed between the Eukaryotic domain and the Archaea and Prokaryotes.

Fig. 2(c) shows the distribution of correlation lengths in sample Eukaryote proteins, i.e., sequences of the 20 amino acids, and their corresponding exon sequences. Roughly 94% of the cases for the above proteins have an average correlation length of ~ 1 . We have found that other protein sequences, namely, human, mice, etc., have nearly 100% of the cases with $d_c \leq 2$. This type of low value of correlation length has been observed in the “randomly” generated sequences (sequences composed by randomly generating 20 different symbols). So, most protein sequences exhibit “noisy” ($d_c < 5$) correlation structures and their correlation lengths are much smaller than their corresponding exon sequences.

One of the main goals of this study is to investigate if there are significant overall differences in correlations for coding versus non-coding DNA sequences. Since the Prokaryote and Archaea domains do not have introns, we make this comparison in the Eukaryote domain only. To obtain the largest possible sample we grouped all the Eukaryote exon specimens into a single specimen and called it the Eukaryote exons. Likewise all the introns are grouped into a single specimen, Eukaryote introns. We then calculated the mean correlation length for each group and the percentage of Eukaryote exons and Eukaryote introns that have a value of $d_c \geq 5$. These results are presented in Table 1. We obtain a larger mean value of d_c

for the exons compared to the introns and a slightly greater percentage of $d_c \geq 5$ for Eukaryote exons than introns. In order to further examine the statistical meaningfulness of these results, we perform the hypothesis test described in Section 4. The z -value from Eq. (6) was greater than 1.96, so we rejected the null hypothesis at the 95% confidence level. Since the test suggests that the greater percentage of exons having $d_c \geq 5$ is statistically significant and the mean correlation calculated is much larger for exons than introns, we conclude that the true mean correlation length for the exons is greater than that of the introns.

It is interesting to notice that there is a wide separation in the mean correlation length for the exons compared to the introns while the percentages of exons and introns with $d_c \geq 5$ are very close, within 1%. The difference in the mean values of the correlation lengths is primarily due to a greater number of exons with very large correlation lengths ($d_c \geq 500$).

5.2. Symbolic power spectrum results

Next, we show the symbolic power spectrum results. We determined the scaling of the power P versus frequency f over a spectral range of 6–30 units; we defined the frequency unit in Section 3.2. We obtained the least-squared value of the spectral index²

² It has been argued in Ref. [11] that the meaning of a measured value of the spectral index, β , may be dubious because of the dependence of the scaling range on the parameters, in particular, the window size for computing the power spectrum. However, in our paper, we have fitted the $1/f^\beta$ power law in the same low-frequency region for both introns and exons. We then use β only to make relative comparisons of correlation content between exons and introns. Though our range for scaling estimates is relatively short, we consistently measure a value of β in this range across a large number of sequences. The values of β , as obtained here, will not occur in a sample of random sequences. The power spectrum results reveal non-random correlation structures in gene sequences.

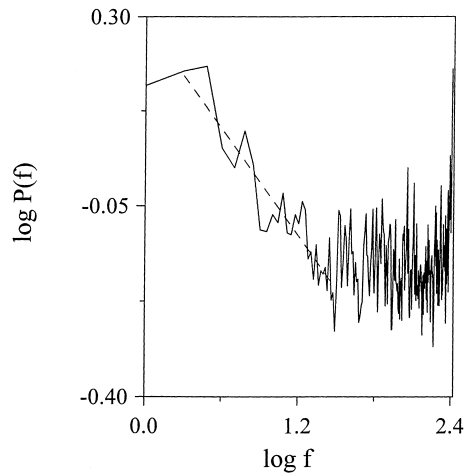


Fig. 3. Symbolic power spectrum for a sample Eukaryote intron (fitted using a $1/f^\beta$ power law). The dashed line represents least-squares fit from 2 to 30 units of frequency.

β for which the power spectrum scaled like $1/f^\beta$. In Fig. 3, a sample symbolic power spectrum result for a human intron is shown, and the spectral index is $\beta = 0.33 \pm 0.03$.

As in the case of the MIF, we want to compare the introns versus the exons. We determined the mean spectral index $\bar{\beta}$ for the Eukaryote exons and introns. The results in Table 2 show a slightly larger $\bar{\beta}$, by 0.01, for Eukaryote exons than introns. Like MIF, we also performed a large sample hypothesis test on the results. The z -value here was greater than 1.96, and the null hypothesis H_0 was rejected hypothesis at the 95% confidence level. As in the case of the MIF, since the

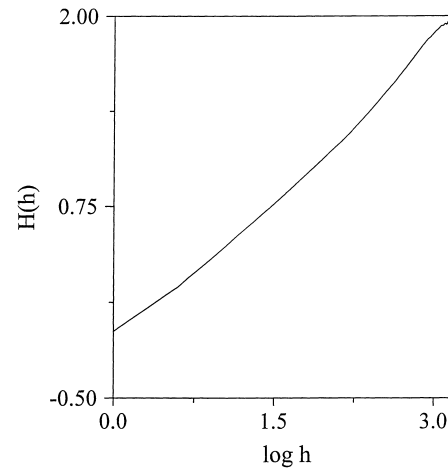


Fig. 4. Hölder exponent calculation for a sample Eukaryote intron.

calculated mean spectral index was slightly larger for the case of the exons rather than introns, this suggests that the true mean spectral index is slightly greater for the exons than the introns.

5.3. Hölder exponent results

Along with the above MIF and power spectrum studies, we calculate the Hölder exponent for Eukaryote exons and introns. Fig. 4 shows a sample Hölder exponent plot for a human intron. The least-squares fit reveals that the average Hölder exponent α for this case is 0.75 ± 0.03 . We determined the mean Hölder exponent, $\bar{\alpha}$ for the Eukaryote exons and introns. The results are shown in Table 3. The value of

Table 2
Symbolic power spectrum results of exons and introns and hypothesis testing

Case	$\bar{\beta}$	Test statistic (z -value)	Result
Eukaryote exons	0.13	8.7	$ z > z_{\alpha/2}$; reject H_0
Eukaryote introns	0.12		

Table 3
Hölder exponent results of exons and introns and hypothesis testing

Case	$\bar{\alpha}$	Test statistic (z -value)	Result
Eukaryote exons	0.77	3.1	$ z > z_{\alpha/2}$; reject H_0
Eukaryote introns	0.75		

$\bar{\alpha}$ for exons was slightly larger, by 0.02, than $\bar{\alpha}$ for introns. From the hypothesis testing we obtain a z -value greater than 1.96 and thus reject the null hypothesis at the 95% confidence level. As in the case of MIF and power spectrum analysis, since our data shows the calculated average Hölder exponent of exons to be slightly larger than introns, this suggests that the true mean Hölder exponent is slightly larger for exons than introns.

6. Conclusions

A key point of this article is comparison of the correlation structures of a large number (~ 7500) of gene sequences. Exons and introns in the Eukaryote domain, in general, exhibit distinct correlation content from “random” sequences. While the mean correlation length computed from MIF is quite different for exons and introns, the percentage of cases with $d_c \geq 5$ is very close, within 1%. The symbolic power spectrum analyses show that the mean spectral indices for exons and introns are very close, namely, 0.13 versus 0.12. The mean Hölder exponents for exons and introns are also similar, namely, 0.77 versus 0.75. The trends of results from the three independent mathematical methods are similar to what is noted by the tests using the large sample theory.

In the Archaea and Prokaryote domains, a large percentage of exons ($\sim 90\%$) show correlation lengths shorter than the exon sequences in Eukaryotes. This possibly indicates that when life diverged from a common ancestor, Eukaryotes may have experienced evolutionary selection that resulted in greater degrees of correlations in their gene sequences. Over the course of evolution, a majority of the Eukaryotes developed structures with high correlation lengths. This may reflect a reduced randomization of their sequences compared to that of the Archaea and Prokaryotes.

Proteins are formed from compressing the DNA sequences into triplet codons, and indeed we observe that they have significantly smaller correlation lengths than their corresponding coding sequences. The cor-

relation lengths observed in the proteins are indistinguishable from noise.

This study augments our understanding of the information contained in a large number (~ 7500) of gene sequences across all the three domains of life. Roughly 100 times the current amount of gene sequence data will be generated in the next 5 years, and the methods presented in this study will be useful for analyzing new larger volume of data.

Acknowledgements

The first author thankfully acknowledges helpful discussions with Professor Wentian Li of Rockefeller University and with Professor Bimal Sinha of University of Maryland, Baltimore County, MD.

References

- [1] W. Li, *Int. J. Bifurc. Chaos* 2 (1992) 137.
- [2] W. Li, K. Kaneko, *Europhys. Lett.* 17 (1992) 655.
- [3] W. Li, *J. Statist. Phys.* 60 (1990) 823.
- [4] W. Li, T. Marr, K. Kaneko, *Physica D* 75 (1995) 217.
- [5] H. Herzel, I. Große, *Physica A* 216 (1995) 518.
- [6] H. Herzel, I. Große, *Phys. Rev. E* 55 (1997) 800.
- [7] H. Herzel, E.N. Trifonov, O. Weiss, I. Grosse, *Physica A* 294 (1998) 449.
- [8] R.F. Voss, *Phys. Rev. Lett.* 68 (1992) 3805.
- [9] E. Coward, *J. Math. Biol.* 36 (1997) 64.
- [10] B. Borstnik, D. Pumpernik, D. Lukman, *Europhys. Lett.* 23 (1993) 389.
- [11] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley, *Phys. Rev. E* 51 (1995) 5084.
- [12] S.V. Buldyrev, N.V. Dokholyan, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, G.M. Viswanathan, *Physica A* 249 (1998) 430.
- [13] A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, *Phys. Rev. Lett.* 74 (1995) 3293.
- [14] A. Arneodo, Y. D'Aubenton-Carafa, E. Bacry, P.V. Graves, J.F. Muzy, C. Thermes, *Physica D* 96 (1996) 291.
- [15] A. Arneodo, Y. D'Aubenton-Carafa, B. Audit, E. Bacry, J.F. Muzy, C. Thermes, *Physica A* 249 (1998) 439.
- [16] P. Bernaola-Galvan, R. Roman-Roldan, J.L. Oliver, *Phys. Rev. E* 53 (1996) 5181.
- [17] R. Roman-Roldan, P. Bernaola-Galvan, J.L. Oliver, *Phys. Rev. Lett.* 80 (1998) 1344.
- [18] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, *Nature* 356 (1992) 168.
- [19] A.C. Dreismann, D. Larhammer, *Nature* 361 (1993) 212.

- [20] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, Z.D. Goldberger, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.-K. Peng, M. Simons, *Physica A* 205 (1994) 214.
- [21] M.Y. Azbel, *Phys. Rev. Lett.* 75 (1995) 168.
- [22] V.V. Prabhu, J.M. Claverie, *Nature* 359 (1992) 782.
- [23] S. Nee, *Nature* 357 (1992) 450.
- [24] F. Kenneth, *Fractal Geometry: Mathematical Foundations and Applications*, Wiley, New York, 1990.
- [25] C.R. Rao, *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.