EURASIP Journal on Applied Signal Processing 2004:1, 132–137

Genomic Signals of Reoriented ORFs

Paul Dan Cristea

Biomedical Engineering Center, Politehnica University of Bucharest, Splaiul Independentei 313, Bucharest 77206, Romania
Email: pcristea@dsp.pub.ro

Received 14 March 2003; Revised 12 September 2003

Complex representation of nucleotides is used to convert DNA sequences into complex digital genomic signals. The analysis of the cumulated phase and unwrapped phase of DNA genomic signals reveals large-scale features of eukaryote and prokaryote chromosomes that result from statistical regularities of base and base-pair distributions along DNA strands. By reorienting the chromosome coding regions, a "hidden" linear variation of the cumulated phase has been revealed, along with the conspicuous almost linear variation of the unwrapped phase. A model of chromosome longitudinal structure is inferred on these bases.

Keywords and phrases: genomic signals, open reading frames, ORF orientation.

1. INTRODUCTION

The conversion of nucleotide sequences into digital signals offers the opportunity to apply signal processing methods to analyze genomic information. Using the genomic signal approach, long-range features of DNA sequences, maintained over distances of 10^6–10^8 base pairs, that is, at the scale of whole chromosomes, have been found [1, 2, 3, 4, 5, 6, 7]. One of the most conspicuous results is that the unwrapped phase of the complex genomic signal varies almost linearly along all investigated chromosomes for both prokaryotes and eukaryotes. The slope is specific for various taxa and chromosomes. Such a behavior reveals a large-scale regularity in the distribution of the pairs of successive nucleotides—a rule for the statistics of second order: *the difference between the frequency of positive nucleotide-to-nucleotide transitions (*A → G, G → C, C → T, T → A*) and that of negative transitions (the opposite ones) along a strand of nucleic acid tends to be small, constant, and taxon and chromosome specific*. There is a similarity between this rule and Chargaff's rules referring to the frequencies of occurrence of nucleotides, that is, to statistics of the first order [8].

The paper shows that the abrupt changes in nucleotide frequencies along DNA strands of prokaryote chromosomes, as revealed by the piecewise linear variation of the cumulated phase of complex genomic signals [1, 2, 3, 4, 5, 6, 7] or by the skew diagrams [9, 10, 11], are the effect of corresponding abrupt changes in the distribution of direct and inverse open reading frames (ORFs) along the strand. It is also shown that, by reorienting all the negative (inverse) ORFs in the direction of the positive (direct) ones, an almost linear variation of the cumulated phase along the concatenated sequence is obtained, corresponding to almost constant frequencies of nucleotides along the entire chain of concatenated reordered ORFs. This large-scale homogeny of the reordered ORFs, together with the taxon specific large-scale regularities of the actual nucleic DNA strands, suggests that the distribution of direct and inverse coding segments along chromosomes, as reflected in the slope of the cumulated phase, has a functional role, most probably linked to the control of the crossing-over/recombination process, thus playing a role in the separation of species. A similar property probably exists in eukaryote chromosomes too, but the relative extension of the coding regions is much lower than in the case of prokaryotes, so that there is too little information for the reordering of the extremely large number of direct and inverse individual chromosome patches.

The paper also presents a model of chromosome longitudinal structure. The model explains why the frequency of nucleotide-to-nucleotide transitions does not change significantly in the points of abrupt changes of the nucleotide frequencies or as a consequence of ORF reordering. Correspondingly, the model explains the ubiquitous almost linear variation of the unwrapped phase of the genomic signals along all investigated chromosomes.

2. DATA AND METHOD

Complete genomes or complete sets of available contigs for eukaryote and prokaryote taxa have been downloaded from the GenBank [12] database of National Institutes of Health (NIH), converted into genomic signals, and analyzed at the scale of whole chromosomes.

As the detailed methodology of the nucleotide, codon, and amino acid sequence conversion into digital signals has been presented elsewhere [3, 4], we give here only a short summary of the quadrantal complex representation used throughout this paper. The nucleotides (adenine (A), cytosine (C), guanine (G), and thymine (T)) are mapped to four

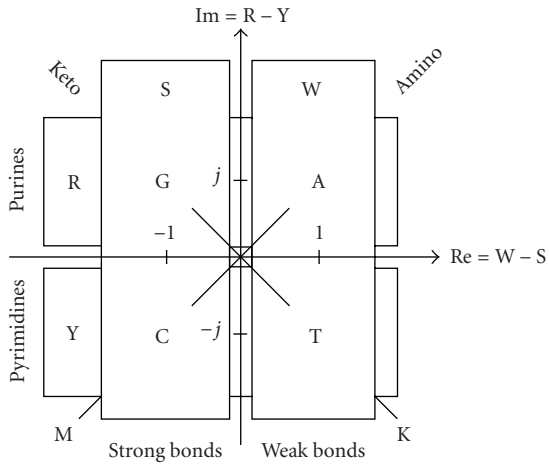

Figure 1: Nucleotide quadrantal complex representation.

complex numbers as shown in Figure 1:

$$a = 1 + j, \quad c = -1 - j, \quad g = -1 + j, \quad t = 1 - j. \tag{1}$$

The representation (1) conserves the main six classes of nucleotides:

(i) strong bonds S = {C, G},
(ii) weak bonds W = {A, T},
(iii) amino M = {A, C},
(iv) keto K = {G, T},
(v) pyrimidines Y = {C, T},
(vi) purines R = {A, G},

and readily expresses the W-S and R-Y dichotomies. This representation allows also the classification of nucleotide pairs in three sets of transitions, in accordance with the change of the unwrapped phase they produce when occurring in a sequence:

(i) the *positive transitions* A → G, G → C, C → T, and T → A that determine a variation with $+\pi/2$ in the trigonometric sense,
(ii) the set of *negative transitions* A → T, T → C, C → G, and G → A—that determines a variation of $-\pi/2$, clockwise,
(iii) the set of *neutral transitions* that correspond to a zero-mean change of the unwrapped phase.

The slopes s_c of the cumulated phase and s_u of the unwrapped phase of a complex genomic signal, obtained by applying the representation (1) to a DNA sequence, are linked to the nucleotide and the nucleotide-to-nucleotide transition frequencies by the following equations [2]:

$$s_c = \frac{\pi}{4}\left[3(f_G - f_C) + (f_A - f_T)\right], \tag{2}$$

$$s_u = \frac{\pi}{2}(f_+ - f_-), \tag{3}$$

where f_A, f_C, f_G, and f_T are the nucleotide frequencies, while f_+ and f_- are the positive and negative transition frequencies.

Thus, the phase analysis of complex genomic signals is able to reveal features of both the nucleotide frequencies and the nucleotide-to-nucleotide transition frequencies along DNA strands.

Relations (1) can be seen as representing the nucleotides in two orthogonal bipolar binary systems with complex bases (units).

3. A MODEL OF DNA LONGITUDINAL STRUCTURE

The chromosomes of both prokaryotes and eukaryotes have a very "patchy" structure comprising many intertwined coding and noncoding segments oriented in a direct and inverse sense. The reversed orientation of DNA segments has been found first for the coding regions, where direct and inverse ORFs have been identified. The analysis of the modalities in which DNA segments can be chained together along the DNA double helix is important for understanding genomic signal large-scale properties [1, 2, 3].

The direction reversal of a DNA segment is always accompanied by the switching of the antiparallel strands of its double helix. This property is a direct result of the requirement that all the nucleotides be linked to each other along the DNA strands only in the 5′ to 3′ sense.

Figure 2 schematically shows the way in which the 5′ to 3′ orientation restriction is satisfied when a segment of a DNA double helix is reversed and/or has its strands switched. In the case in Figure 2a, the two component helices have the chains $(A_0A_1)(A_1A_2)(A_2A_3)$ and $(B_0B_1)(B_1B_2)(B_2B_3)$, respectively, ordered in the 5′ to 3′ sense indicated by the arrows. The reversal of the middle segment, without the corresponding switching of its strands (Figure 2b), would generate the forbidden chains $(A_0A_1)(A_2A_1)(A_2A_3)$ and $(B_0B_1)(B_2B_1)(B_2B_3)$ that violate the 5′ to 3′ alignment condition. Similarly, the switching of the strands of the middle segment, without its reversal, would generate the equally forbidden chains $(A_0A_1)(B_2B_1)(A_2A_3)$ and $(B_0B_1)(A_2A_1)(B_2B_3)$ shown in Figure 2c. Finally, the conjoint reversal of the middle segment and the switching of its strands (Figure 2d) generate the chains $(A_0A_1)(B_1B_2)(A_2A_3)$ and $(B_0B_1)(A_1A_2)(B_2B_3)$, compatible with the 5′ to 3′ orientation condition. As a consequence, there is always a pair of changes (direction reversal and strand switching) produced by an inversed insertion of a DNA segment so that the sense/antisense orientation of individual DNA segments affects the nucleotide frequencies but not the frequencies of the positive and negative transitions. Figure 3 shows the effect of the *segment reversal* and *strand switching* transformations on the positive and negative nucleotide-to-nucleotide transitions for the case of the complex genomic signal representation given by (1). After a pair of segment reversal and strand switching transformations of a DNA segment, the nucleotide transitions do not change their type (positive or negative). As a consequence, the slope of the unwrapped phase does not change as the slope of the cumulated phase. This explains why the cumulated phase and the unwrapped phase

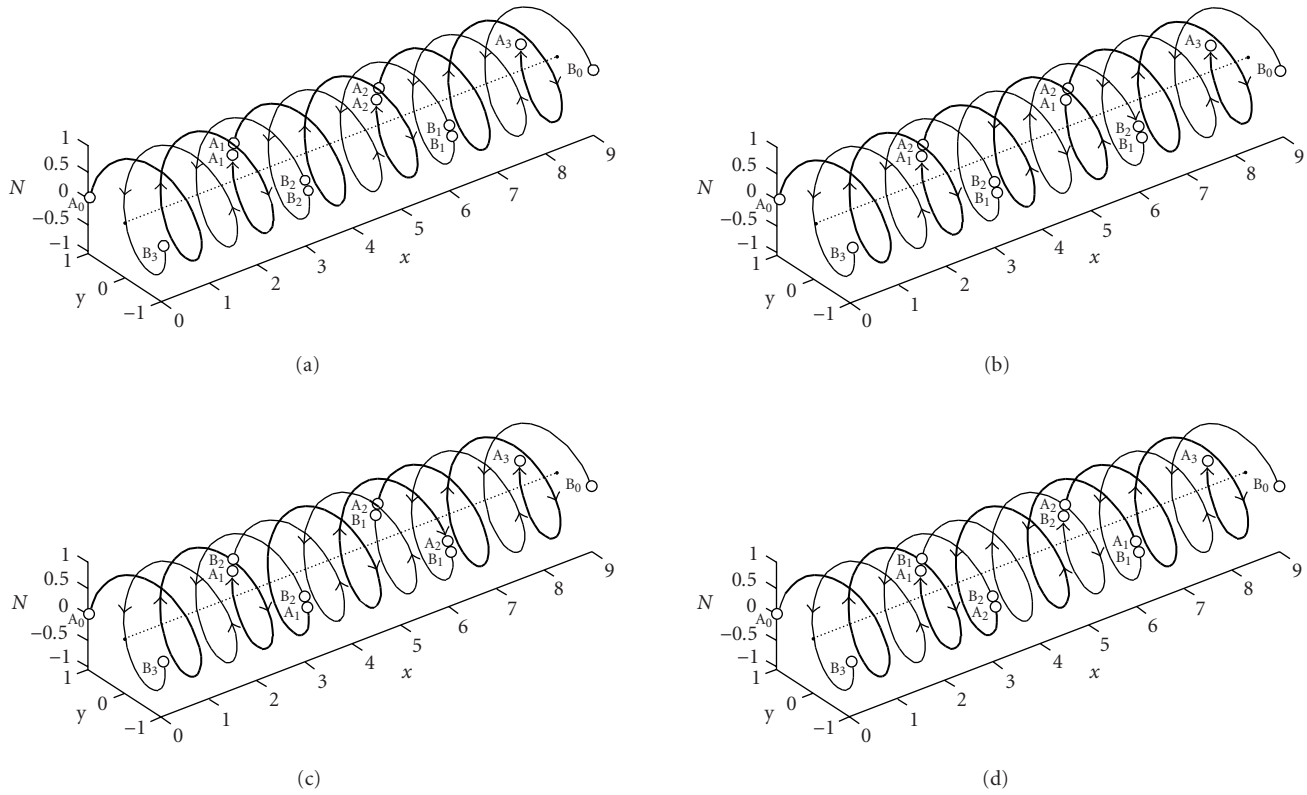

Figure 2: Schematic representations of the direction reversal of a DNA segment. (a) Initial state in which the two antiparallel strands have all the marked segments ordered in the 5′ to 3′ direction, indicated by arrows. (b) Hypothetic reversal of the middle segment without the switching of the strands. (c) Hypothetic switching of strands for the middle segment without its reversal. (d) Direction reversal and strand switching for the middle segment. The 5′ to 3′ alignment condition is violated in cases (b) and (c) but reestablished in (d).

of genetic signals have completely different types of variations along DNA molecules that contain a large number of reversed segments.

4. CUMULATED AND UNWRAPPED PHASE VARIATION ALONG CHROMOSOMES AND CONCATENATED REORIENTED CODING REGIONS

Figure 4 presents the cumulated and the unwrapped phases of the complete circular chromosome of *Salmonella typhi*, the multiple-drug resistant strain CT18 [13] (accession AL5113382 [12]). The locations of the breaking points, where the cumulated phase changes the sign of the slope of its variation along the DNA strand, are given in Figure 4. Even if, locally, the cumulated phase and the unwrapped phase do not have a smooth variation, at the large scale used in Figure 4, the variation is quite smooth and regular. A pixel in the curves of Figure 4 represents 6050 data points, but the absolute value of the difference between the maximum and minimum values of the data in the set of points represented by each pixel is smaller than the vertical pixel dimension expressed in data units. This means that the local data variation falls between the limits of the width of the line used for the plot so that the graphic representation of data by a line is adequate. As found for other prokaryotes [2, 3, 4, 5], the cumulated phase has an approximately piecewise linear variation over two almost equal domains, one of positive slope

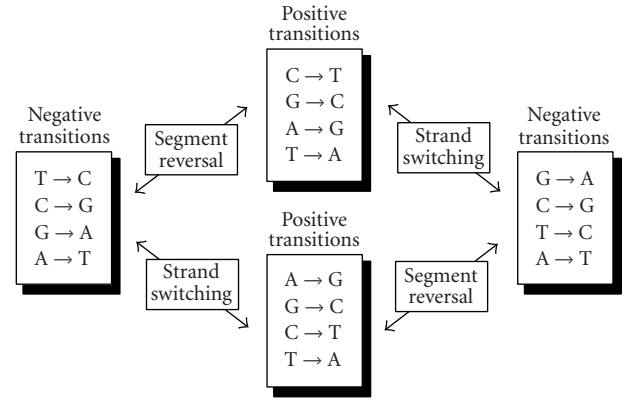

Figure 3: Effect of segment reversal and strand switching on positive and negative nucleotide-to-nucleotide transitions. An even number of transforms do not change the type of the transitions.

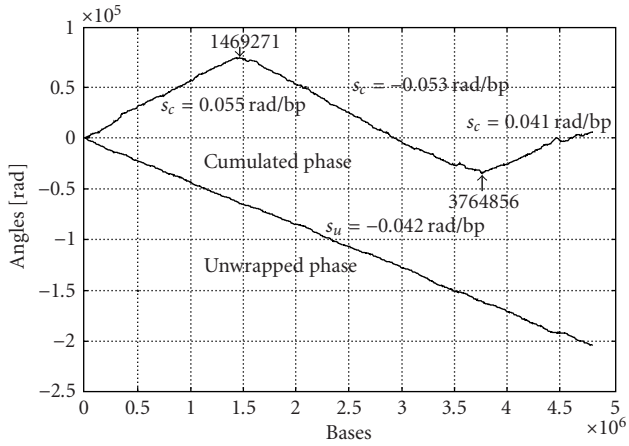

FIGURE 4: Cumulated and unwrapped phases for the genomic signal of the complete chromosome (4809037 bp) of *Salmonella typhi* [13] (accession AL5113382 [12]).

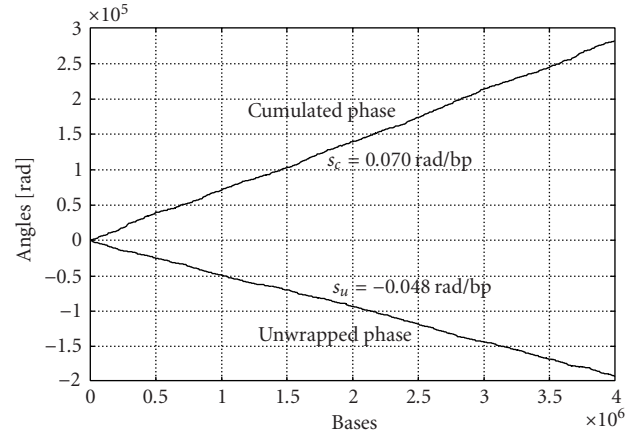

FIGURE 5: Cumulated and unwrapped phases of the genomic signal for the concatenated 4393 reoriented coding regions (3999478 pb) of *Salmonella typhi* genome [13] (accession AL5113382 [12]).

(apparently divided in the intervals 1-1469271 and 3764857-4809037, but actually contiguous on the circular chromosome) and the second of negative slope (1469272-3764856), while the unwrapped phase has an almost linear variation for the entire chromosome, showing little or no change in the breaking points. The breaking points, like the extremes of the integrated skew diagrams, have been put in relation with the origins and termini of chromosome replichores [2, 9, 11]. The slope of the cumulated phase in each domain is related to the nucleotide frequency in that domain by (2). In the breaking points, a macroswitching of the strands, accompanied by a reversal of one of the domain-large segments, occurs. On the other hand, the two domains comprise a large number of much smaller segments, oriented in the direct and the inverse sense. At the junctions of these segments, reversals and switchings of DNA helix segments take place as described in Section 3. The average slope of each large domain is actually determined by the density of direct and inverse small segments along that domain. This model can be verified by using the "**.ffn*" files in the GenBank [12] database that contain the coding regions of the sequenced genomes, together with their orientation. Concatenating the coding regions oriented in the positive direction (positive ORFs) with the reoriented (reversed and complemented) coding regions read in the negative direction (negative ORFs), a nucleotide sequence with all the coding regions (exons and introns) oriented in the same direction is obtained. Because the intergenic regions for which the orientation is not known have to be left out of the reoriented sequence, this new sequence is shorter than the one that contains the entire chromosome or all the available contigs given in the "**.gbk*" files of the GenBank database [12].

Figure 5 shows the cumulated and unwrapped phases of the genomic signal obtained by concatenating the 4393 reoriented coding regions of *Salmonella typhi* genome [13] (accession AL5113382 [12]). Each inverse coding region (inverse ORF) has been reversed and complemented, that is,

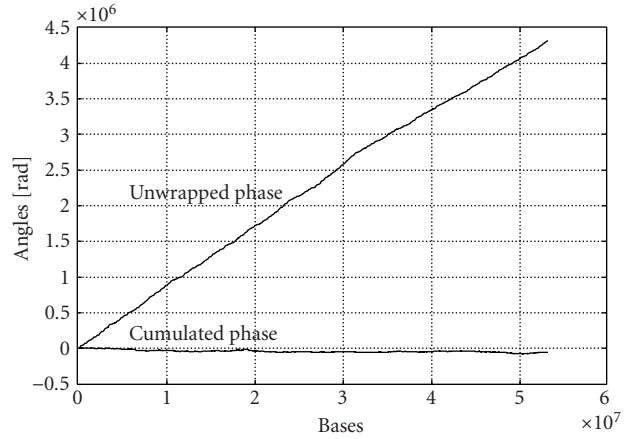

FIGURE 6: Cumulated and unwrapped phases along the complete chromosome 4 of *Mus musculus* [14] (NT019246 53208110 bp [12]).

the nucleotides inside the same W (adenine-thymine) or S (cytosine-guanine) class have been replaced with each other to take into account the switching of the strands that accompanies the segment reversal.

As expected from the model, the breaking points in the cumulated phase disappear and the absolute values of the slopes increase as there is no longer interweaving of direct and inverse ORFs. The average slope s_c of the cumulated phase of a genomic signal for a domain is linked to the average slope $s_c^{(0)}$ of the concatenated reoriented coding regions by the relation

$$s_c = \frac{\sum_{k=1}^{n_+} l_k^{(+)} - \sum_{k=1}^{n_-} l_k^{(-)}}{\sum_{k=1}^{n_+} l_k^{(+)} + \sum_{k=1}^{n_-} l_k^{(-)}} s_c^{(0)}, \tag{4}$$

where $\sum_{k=1}^{n_+} l_k^{(+)}$ and $\sum_{k=1}^{n_-} l_k^{(-)}$ are the total lengths of the n^+ direct and n^- inverse ORFs in the given domain.

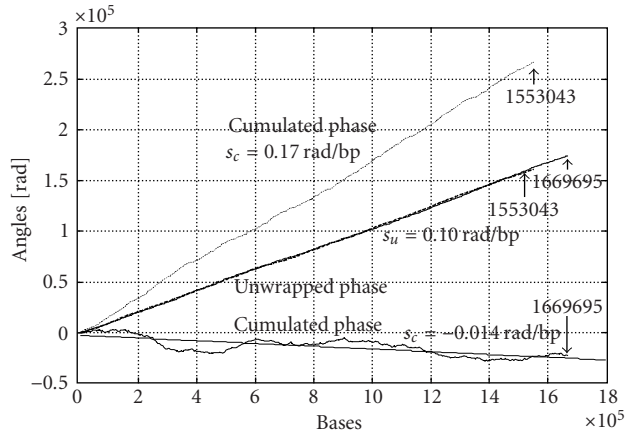

Figure 7: Cumulated and unwrapped phases of the genomic signals for the complete nucleotide sequence and the concatenated reoriented coding regions of *Aeropyrum pernix K* genome [15] (NC000854 [12]) versus all genomes.

The unwrapped phase, which is linked by (3) to the nucleotide positive and negative transition frequencies, shows little or no change when replacing the chromosome nucleotide sequence with the concatenated sequence of reoriented coding regions. As explained, the reorientation of the inverse coding regions consists in their reversal and switching of their strands.

The model also explains the finding that the unwrapped phase, which reveals second-order statistical features, has an almost linear variation even for eukaryote chromosomes [1, 2, 3, 4, 5, 6, 7] despite their very high fragmentation and quasirandom distribution of direct and inverse ORFs, while the cumulated phase, linked to the frequency of nucleotides along the DNA strands, displays only a slight drift close to zero. Figure 6 gives the cumulated phase and the unwrapped phase along the complete chromosome 4 [14] of *Mus musculus* (accession NT019246 [12]). The unwrapped phase increases almost linearly (actually there are two domains of quasilinearity with distinct slopes), while the cumulated phase remains almost zero (at the scale of the plot). Similar results have been obtained for all *Mus musculus* and *Homo sapiens* chromosomes.

The reversal of all inverse segments along the same positive direction, as performed for prokaryotes, would most probably reveal a similar "hidden linear variation" of the cumulated phase. Unfortunately, for eukaryotes, the information about the OFR orientation is not sufficient to perform the reordering, because the extension of the coding regions is only a small fraction from the total length of the chromosome. We illustrate the way the "hidden" linear variation of the cumulated phase could be revealed by DNA segment reorientation, by using again the case of a prokaryote, the aerobic hyperthermophilic crenarchaeon *Aeropyrum pernix K*, for which the genome has been completely sequenced [12, 14]. Figure 7 presents the cumulated and the unwrapped phases of the genomic signal for the entire genome comprising 1669695 base pairs. The unwrapped phase varies almost linearly, like in all the other investigated prokaryote and eukaryote genomes [1, 2, 3, 4, 5, 6, 7], confirming the rule stated in Section 1 and explained in this paper. The cumulated phase decreases irregularly, an untypical behavior for prokaryotes that tend to have a regular piecewise linear variation of the cumulated phase, as shown above. Figure 7 also shows the cumulated and unwrapped phases of the signal that correspond to a sequence obtained by concatenating the 1839 coding regions in the genome after reorienting them all in the same reference direction. The new sequence comprises only the 1553043 base pairs involved in the coding regions for which the sense information is available; the intergenic regions, for which this information is missing, have been left out. As seen in the figure, the cumulated phase changes to a uniform, almost linear, increase while the unwrapped phase remains practically unchanged.

5. CONCLUSION

DNA sequences of complete chromosomes or sequences obtained by concatenating all reoriented coding regions of chromosomes have been converted into genomic signals by using a nucleotide complex representation derived from the nucleotide tetrahedral representation. Some large-scale features of the resulting genomic signals have been analyzed. The cumulated phase and unwrapped phase of genomic signals are correlated with the statistical distribution of bases and base pairs, respectively. The paper presents a model of the longitudinal structure of the chromosomes that explains the almost linear variation of the unwrapped phase of the complex genomic signals for all prokaryotes and eukaryotes [1, 2, 3, 4, 5, 6, 7]. The linearity of the cumulated phase for the reordered ORFs, reflecting a large-scale homogeny of the nucleotide distribution in such sequences, on one hand, and the taxon specific variation of the cumulated phase for the actual nucleic DNA strands, on the other, suggest the hypotheses of a primary ancestral genomic material and of a functional role of the particular orientation of direct and inverse DNA segments that generate specific densities of the first- and second-order repartition of nucleotides along chromosomes. The relevance of these large-scale features of chromosomes in the control of the crossing-over/recombination process, the identification of the interacting regions of chromosomes, and the separation of species, as well as the mechanisms that generate the specific arrangements of direct and inverse ORFs remain to be further investigated.

REFERENCES

[1] P. Cristea, "Genomic signals for whole chromosomes," in *Manipulation and Analysis of Biomolecules, Cells, and Tissues*, vol. 4962 of *Proceedings of SPIE*, pp. 194–205, San Jose, Calif, USA, January 2003.

[2] P. Cristea, "Large scale features in DNA genomic signals," *Signal Processing*, vol. 83, no. 4, pp. 871–888, 2003.

[3] P. Cristea, "Conversion of nucleotides sequences into genomic signals," *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 279–303, 2002.

[4] P. Cristea, "Genetic signal representation and analysis," in *Functional Monitoring and Drug-Tissue Interaction*, vol. 4623 of *Proceedings of SPIE*, pp. 77–84, San Jose, Calif, USA, January 2002.
[5] P. Cristea, "Genetic signal analysis," in *Proc. 6th International Symposium on Signal Processing and Its Applications (ISSPA '01)*, pp. 703–706, Kuala Lumpur, Malaysia, August 2001.
[6] P. Cristea, "Genetic signals," *Rev. Roum. Sci. Techn. Electrotechn. et Energ.*, vol. 46, no. 2, pp. 189–203, 2001.
[7] P. Cristea and R. Tuduce, "Signal processing of genomic information: Mitochondrial genomic signals of hominidae," in *Proc. 4th EURASIP Conference Focused on Video/Image Processing and Multimedia Communications (EC-VIP-MC '03)*, Zagreb, Croatia, July 2003.
[8] E. Chargaff, "Structure and function of nucleic acids as cell constituents," *Federation Proceeding*, vol. 10, pp. 654–659, 1951.
[9] J. M. Freeman, T. N. Plasterer, T. F. Smith, and S. C. Mohr, "Patterns of genome organization in bacteria," *Science*, vol. 279, no. 5358, pp. 1827–1832, 1998.
[10] A. Grigoriev, "Analyzing genomes with cumulative skew diagrams," *Nucleic Acids Research*, vol. 26, no. 10, pp. 2286–2290, 1998.
[11] J. R. Lobry, "Asymmetric substitution patterns in the two DNA strands of bacteria," *Molecular Biology and Evolution*, vol. 13, no. 5, pp. 660–665, 1996.
[12] National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, GenBank, http:// www.ncbi.nlm.nih.gov/genoms/.
[13] J. Parkhill, G. Dougan, K. D. James, et al., "Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18," *Nature*, vol. 413, no. 6858, pp. 848–852, 2001.
[14] J. Kawai, A. Shinagawa, K. Shibata, et al., "Functional annotation of a full-length mouse cDNA collection," *Nature*, vol. 409, no. 6821, pp. 685–690, 2001, RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium.
[15] Y. Kawarabayasi, Y. Hino, H. Horikawa, et al., "Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, Aeropyrum pernix K1," *Journal of DNA Research*, vol. 6, no. 2, pp. 83–101, 1999.

Paul Dan Cristea graduated from the Faculty of Electronics and Telecommunications, Politehnica University of Bucharest (PUB) in 1962, and the Faculty of Physics, PUB, as head of the series. He obtained the Ph.D. degree in technical physics from PUB, in 1970. His research and teaching activities have been in the fields of genomic signals, digital signal and image processing, connectionist and evolutionary systems, intelligent e-learning environments, computerized medical equipment, and special electrical batteries. He is the author or coauthor of more than 125 published papers, 12 patents, and contributed to more than 20 books in these fields. Currently, he is the General Director of the Biomedical Engineering Center of PUB and Director of the Romanian Bioinformatics Society.