PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 5, Number 2/2004, pp.000-000

DIGITAL REPRESENTATION AND ANALYSIS OF GENOMIC DATA

Paul Dan CRISTEA

Bio-Medical Engineering Center
"Politehnica" University of Bucharest, Romania
pcristea@dsp.pub.ro

Complex representation of nucleotides is used to convert sequences of nucleotides into complex digital genomic signals that are analyzed using signal processing methods. Using phase analysis, we establish global and large scale features of prokaryote and eukaryote chromosomes, including all human chromosomes. We show that both the cumulated and unwrapped phases of concatenated re-oriented ORFs have a linear variation along prokaryote DNA molecules. A hypothesis on the existence of a primary ancestral genomic material from which the current DNA molecules have evolved is advanced on these bases.

Keywords: Genomic signals, Complex representation, Phase analysis, Open Reading Frames

INTRODUCTION

The almost complete sequencing of the human [10,15] and mouse genomes [13] and of other eukaryote and prokaryote organisms [9,11,14], as well as the public access to the genomic databases, offer the opportunity to explore in depth this unique information depository in the attempt to extract useful knowledge from this vast volume of raw data.

After the publishing of the first versions of the almost complete sequence of the human genome, it has been formulated the opinion that only the genes, the approximately five percent of the genome that contains the information to synthesize the proteins was of real interest, while the remaining vast majority of the genome was simply "junk DNA". The genes have been considered the "blue-prints" of any organism and the role of other endogenous and exogenous factors in the complex ontogeny, function and dysfunction of the living has been diminished, if not ignored. This reductionist view, reminiscent of the classic concept "*one gene – one trait*", reformulated as "*one gene – one protein*", has also been motivated by the potential importance of genes in pharmacogenomics, in the hope of leading to the synthesis of proteins, potentially useful as targeted medication [12]. Certainly, the importance of proteins for the living can not be underestimated. Proteins are the main contributors to cell structure and, as enzymes, catalyse the chemical reactions specific to the functioning of any cells. Almost everything in the organism is made *of* or *by* proteins. The genes encode the primary structure of proteins, *i.e.*, the composition of the amino acid sequences that build-up the polypeptide chains. A protein can contain several polypeptide chains and its biological functions take place at the level of the very complex spatial structure that results from the coiling (secondary structure), folding (tertiary structure) and aggregation (quaternary structure) of the polypeptide chains. The complexity of the proteome – the set of proteins existing in a cell, exceeds by far the complexity of its genome. The total number of genes in the human genome is of about 30000, while the proteome comprises more than one million proteins. It became gradually clear that the key to organism complexity is not in the gene number, but in the way parts of genes are expressed and combined to build different proteins using alternative splicing. The complexity of the phenotype also results from processes like pleiotropy – one gene affecting several phenotypic characteristics, and polygeny – a group of genes acting together to cumulatively produce a certain trait. It is to be expected that the regulatory mechanisms that control these processes are sensitive to signals from the external environment and from the cells, tissues and organs themselves. Despite the fact that the intergenic part of the human genome contains repetitive, quasi-random sequences and a large amount of transposable elements that bear a close resemblance to the DNA of

independent entities like viruses and bacteria, significant parts of the inter-gene chromosomal DNA most likely play an important role in the control of protein synthesis, conjointly with other gene regulatory systems.

The standard approach of symbolically representing the genomic information by sequences of nitrogenous base symbols in the strands of DNA and RNA molecules (a = adenine, c = cytosine, g = guanine, t = thymine / u = uracil), by symbolic codons (triplets of nucleotides), or by symbolic sequences of amino acids in the corresponding polypeptide chains (for the genes) limits the methodology of processing the genetic information to mere pattern matching and statistical analysis. Converting the DNA sequences into digital signals [5] opens the possibility to apply signal processing methods to the analysis of genomic data [2-8]. The genomic signal approach has already proven its potential in revealing large scale features of DNA sequences maintained over distances of 10^6 - 10^8 base pairs, including both coding and non-coding regions, *i.e.*, at the scale of whole chromosomes [3]. One of the most conspicuous results is that the unwrapped phase of DNA complex genomic signals varies almost linearly along all investigated chromosomes, for both prokaryotes and eukaryotes. The slope is specific for various taxa and chromosomes. Such a behaviour reveals large scale second order statistical rules for the distribution of pairs of successive nucleotides, similar to Chargaff's first order rules for the frequencies of occurrence of nucleotides [1]. The existence of this regularity supports the view that extra–gene DNA sequences, that do not encode proteins, can play important functional roles, most likely in the control of gene expression. This is strongly supported by the recent publications of a high quality draft sequence of the mouse genome [13] that allowed a comparative analysis of the mouse and human genomes. Along with the about 30000 genes, the *homo sapiens* and the *mus musculus* genomes share twice as long other extra–gene DNA sequences. These sequences must have important functions to explain their exact conservation over the 75 million year divergent evolution of human and mouse lineages.

The paper gives a brief overview of our current work on digital genomic signal representation and analysis and reports several new results obtained by using this approach.

2. VECTORIAL AND COMPLEX REPRESENTATION OF NUCLEOTIDES

The conversion of DNA sequences from the symbolic form given in the genomic data bases [12] into digital signals allows using powerful signal processing procedures for handling and analysing genomic data. We have investigated a large variety of mappings and we have selected the tetrahedral (3D) and the complex (2D) representations used throughout our work [2-8] based on several requirements that must be satisfied by an adequate mapping. First of all, the mapping has to be *truthful* and *un-biased*. A truthful mapping expresses the relevant features of the represented objects in corresponding mathematical properties of the digital signal samples. An un-biased mapping does not introduce artefacts, *i.e.*, properties of the resulting digital signal without correspondent in the of the features of the represented symbolic sequence. On the other hand, the mapping should allow a fast and computationally effective conversion and should provide an output easy to read for humans. The last request favours representations with a low dimensionality of the output, preferably 1D or 2D. In this section we briefly present the digital representation of nucleotides starting from the essential features of DNA sequences. A detailed study of the symbolic to digital conversion of genomic sequences can be found in [5].

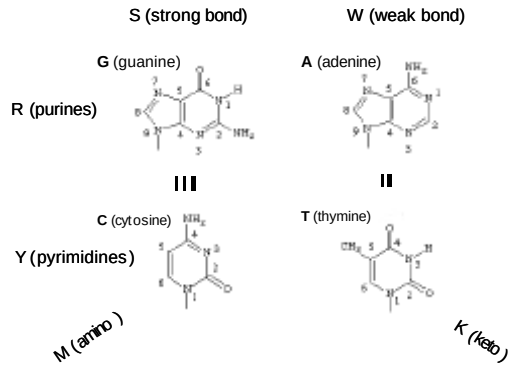

Fig. 1. Dichotomies of nitrogenous bases

The double helix DNA molecules comprises two antiparallel intertwined complementary strands, each a helicoidally coiled heteropolymer [16]. The repetitive units are the nucleotides which consist each of three parts linked by strong covalent bounds: a phosphate group, a sugar – the deoxyribose, and a nitrogenous base. There are four nucleotides in DNA molecules differing by the nitrogenous basis they contain: adenine

(A), cytosine (C), guanine (G) or thymine (T). Along the two strands of the DNA double helix, a basis in one chain always faces its complementary basis in the other chain, and only the base pairs T-A and C-G exist. The hydrogen bonds within these base pairs keep together the two strands. The entities in the nucleotide chains that encode polypeptides, i.e., specify the primary structure of proteins, are called genes. The genes are made up of several exons – coding regions separated by introns – non-coding regions. The protein coding is governed by the Genetic Code (GC) that establishes the mapping of codons – triplets of successive nucleotides in the exons to the 20 amino acids found in the polypeptide chains and to the terminator that marks the end of an encoding segment. There is a large redundancy (degeneration) of the GC as there are 4^3 = 64 codons to specify only 21 distinct outputs. The redundancy is distributed unevenly among the outputs: there are amino acids encoded by one (2 instances), two (9 instances), three (one instance), four (5 instances) or six (3 instances) distinct codons, while the terminator is encoded by three codons. When a gene is expressed, the original DNA strand is first transcribed into a complementary messenger RNA (mRNA) sequence, which is edited by the excision of all introns and the joining of all exons. The number of nucleotides in an exon is not necessarily a multiple of three, *i.e.*, an exon does not necessarily comprise an integer number of codons. In RNA molecules, thymine is replaced by uracil – a related nitrogenous base, but the GC remains otherwise the same. A polypeptide chain is synthesized by ribosomes that translate the codon sequence of mRNA into an amino acid sequence. Each of the 20 amino acids is brought by a specific transfer RNA (tRNA).

As schematically shown in Fig. 1, there are three main dichotomies of the nitrogenous bases biochemical properties that allow arranging them in classes: (1) *molecular structure* – A and G are purines (R), while C and T are pyrimidines (Y); (2) *strength of links* – bases A and T are linked by two hydrogen bonds (W - weak bond), while C and G are liked by three hydrogen bonds (S - strong bond); (3) *radical content* – A and C contain the amino (NH_3) group (M class), while T and G contain the keto (C=O) group (K class). To conserve the symmetry of the nucleotides and to express their classification in the couples shown in Fig. 1, we have proposed the nucleotide tetrahedral representation [5, 8] given in Fig. 2.

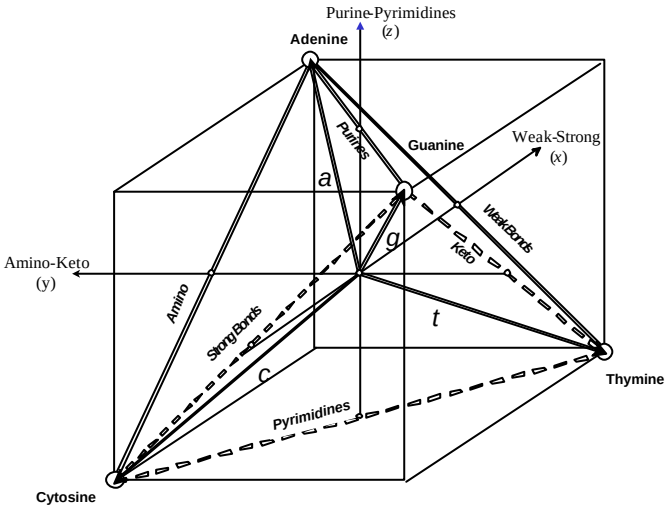

Fig. 2. Nucleotide tetrahedron

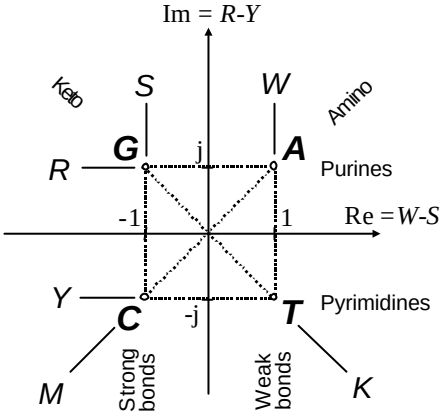

Fig. 3. Nucleotide quadrantal representation

The nucleotides are mapped to four vectors oriented towards the vertices of a regular tetrahedron. Each of the six edges corresponds to a certain class comprising a pair of nucleotides. The representation is three-dimensional and the axes chosen in Fig. 2 correspond to the differences: x = W – S, y = M – K, z = R –Y. By choosing $\{\pm 1\}$ coordinates for the vertices of the embedding cube, the vectors that represent the four nucleotides take the simple form:

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

The dimensionality of the representation can be reduced by projecting the nucleotide tetrahedron on an adequately chosen plane. This plane can be put in correspondence with the complex plane, so that a complex

representation of the nucleotides is obtained. The choice of the projection plane is determined by the features that have to be conserved as being most relevant in the given context. For the representation of DNA sequences, we have found that the projection on the plane xz is best as it expresses the S-W and Y-R dichotomies. The corresponding quadrantal representation is given in Fig. 3, in which the pairs of nucleotides are grouped in the six above mentioned classes, while the corresponding complex representation of the nucleotides is given by the equations:

$$a = 1+j,\ c=-1-j,\ g=-1+j,\ t = 1-j. \tag{2}$$

For a codon consisting of the sequence of three nucleotides $X = B_2B_1B_0,\ B_i \in \{A, C, G, T\};\ i = 0, 1, 2$, the vector and the complex representations are obtained using the basis 2 expansions:

$$\vec{x} = 2^2\vec{b}_2 + 2^1\vec{b}_1 + 2^0\vec{b}_0;\ \vec{b}_i \in \{\vec{a},\vec{c},\vec{g},\vec{t}\};\quad i = 0,1,2, \tag{3}$$

$$x = 2^2 b_2 + 2^1 b_1 + 2^0 b_0;\ b_i \in \{a,c,g,t\};\quad i = 0,1,2, \tag{4}$$

respectively. For example, in the case of methionine, to which corresponds the codon ATG that also starts any gene, the vector representation is $\vec{M} = 5\vec{i} + \vec{j} + 3\vec{k}$, while the complex representation is $M = 5 + j\,3$. Relations (2) and similar ones can be seen as representing the nucleotides in two mutually orthogonal (bipolar) binary systems, each with a complex basis, instead of a system in base four.

Using equation (3) one can build the codon tetrahedron on which the Genetic Code is mapped [5]. It is remarkable that the different representations of an amino acid, resulting from the redundancy of the Genetic Code, are mapped in neighbouring points of the codon tetrahedron, with the exception of the three instances of amino acids degenerated of order six for which none of the investigated mapping can obtain the full contiguity of the representations. Similarly, equation (4) generates a complex representation of the Genetic Code.

3. PHASE ANALYSIS OF GENOMIC SIGNALS

Several analysis tools have been developed for extracting local and large scale features of genetic signals [2,6]. Phase analysis concepts have been presented in detail elsewhere [3], but are briefly reviewed in this section for convenience.

The ***phase*** of a complex number is periodic with period 2π. The standard mathematical convention restricts the phase of a complex number to the domain $(-p, p]$, covering only once all the possible directions in the complex plane. The ***cumulated phase*** is the sum of the phases of the complex samples in a sequence, from the first to the current one. To avoid the bias introduced by the restriction of the phase, which favours π over $-\pi$ for real negative samples, a small uniformly distributed complex noise is added to each sample in the sequence. For the representation (2), the slope s_c of the cumulated phase variation is related to the frequencies of occurrence of the nucleotides along the DNA by the equation [3]:

$$s_c = \frac{p}{4}\left[3\left(f_G - f_C\right) + \left(f_A - f_T\right)\right] \tag{5}$$

The ***unwrapped phase*** is a corrected phase of the elements in a complex sequence, in which the absolute value of the difference between the phase of each element and the phase of its preceding element is kept smaller than π by adding or subtracting an appropriate multiple of 2π to or from the phase of the current element. Again, a small complex noise has been added to avoid the phase bias described above. The unwrapped phase is a measure of the frequencies of the nucleotide pairs (transitions) in a sequence . For the complex representation given in Eq.(2), each *positive transition* A→G, G→C, C→T, T→A determines an increase of the unwrapped phase with $\frac{p}{2}$, each *negative transition* A→T, T→C, C→G, G→A determines a similar decrease, while all other transitions are *neutral* and do not change the unwrapped phase on the average. Correspondingly, the slope s_u of the unwrapped phase along a DNA strand is related to the

difference between the frequency f_+ of the positive transitions and the frequency f_- of the negative ones by the relation:

$$s_u = \frac{p}{2}(f_+ - f_-)\cdot \tag{6}$$

4. LARGE SCALE FEATURES OF GENOMIC SIGNALS

Contigs from several eukaryote and prokaryote genomes have been downloaded from GenBank of NIH [12] and converted to genomic signals using the complex representation given in Eq. (2). Large scale features of the signals have been sought for at the scale of contigs or concatenated contigs.

The cumulated phase and the unwrapped phase have well defined long range trends which are specific for the different eukaryote and prokaryote genomes. Fig. 4 shows the unwrapped phase along the cumulated contigs of all *homo sapiens* chromosomes. The unwrapped phase varies almost linearly or piece-wise linearly, along all chromosomes, with several exceptions. Remarkable enough, when DNA draft sequences are refined, passing to higher quality drafts, the linearity of the unwrapped phase improves. Phase 3 data show best this feature. The average slope of the unwrapped phase for all *homo sapiens* chromosomes are given in Fig. 4 and Table 1.

Table 1. Average slope of the unwrapped phase of *homo sapiens* chromosomes

Chromosome	s_u [rad / bp]	Chromosome	s_u [rad / bp]	Chromosome	s_u [rad / bp]
1	0.072	9	0.067	17	0.078
2	0.064	10	0.067	18	0.060
3	0.063	11	0.069	19	0.084
4	0.056	12	0.069	20	0.073
5	0.060	13	0.057	21	0.056
6	0.062	14	0.066	22	0.091
7	0.066	15	0.072	X	0.057
8	0.062	16	0.075	Y	0.054

The linear increase of the unwrapped phase of a sequence of nucleotide complex representations along a DNA strand shows that the complex representations form on the average a counter clockwise helix that completes a turn over the spatial period:

$$L = \frac{2p}{s_u}, \tag{7}$$

where s_u is the slope of the unwrapped phase. It is remarkable that the trend of variation of the unwrapped phase is maintained over distances of tens of millions of bases, revealing a statistical regularity in the distribution of the succession of bases (base-pairs), not only in the distribution of the bases themselves:

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 & chromosome specific.

Similar bng range behaviour has been found for the *mus musculus* (mouse) contigs. Chromosomes 1 and 4 of mus musculus displays each two distinct linear domains of the unwrapped phase that suggest a composite structure of these chromosomes. For both *homo sapiens* and *mus musculus*, the cumulated phase has locally a less regular variation but, at the scale of representation used for the unwrapped phase, remains close to the horizontal axis showing an overall approximate balance of the purines and pyrimidines in all chromosomes as stated by the well known Chargaff's law.

The behaviour is quite different for prokaryotes. Typically, the cumulated phase varies piece-wise linearly along the DNA strands, as shown in Fig. 5 for the complete genome (3031430 bp) of *Clostridium perfringens* (NC 006633, [12]), an "anaerobic flesh-eater" bacteria [14]. The unwrapped phase presents the same linear variation found for all taxa and all chromosomes. The chromosomes of both prokaryotes and eukaryotes have a very "patchy" structure comprising many intertwined coding and non-coding segments oriented in direct and inverse sense. Figure 5 also shows the effect of re-orienting all the 2660 coding regions

in the genome along the same positive direction. A striking change is displayed by the cumulated phase which becomes approximately linear along the whole sequence of 2530146 bp. The unwrapped phase remains almost unchanged, with the exception of the shortening of the sequence resulting from the removal of the non-coding regions, for which there are no orientation data. As shown in [4], the direction reversal of a DNA segment is always accompanied by the switching of the antiparallel strands of its double helix. This fact explains why the statistics of first order and the cumulated phase change obviously when re-orienting ORFs, while the statistic of second order and the unwrapped phase remain almost unchanged

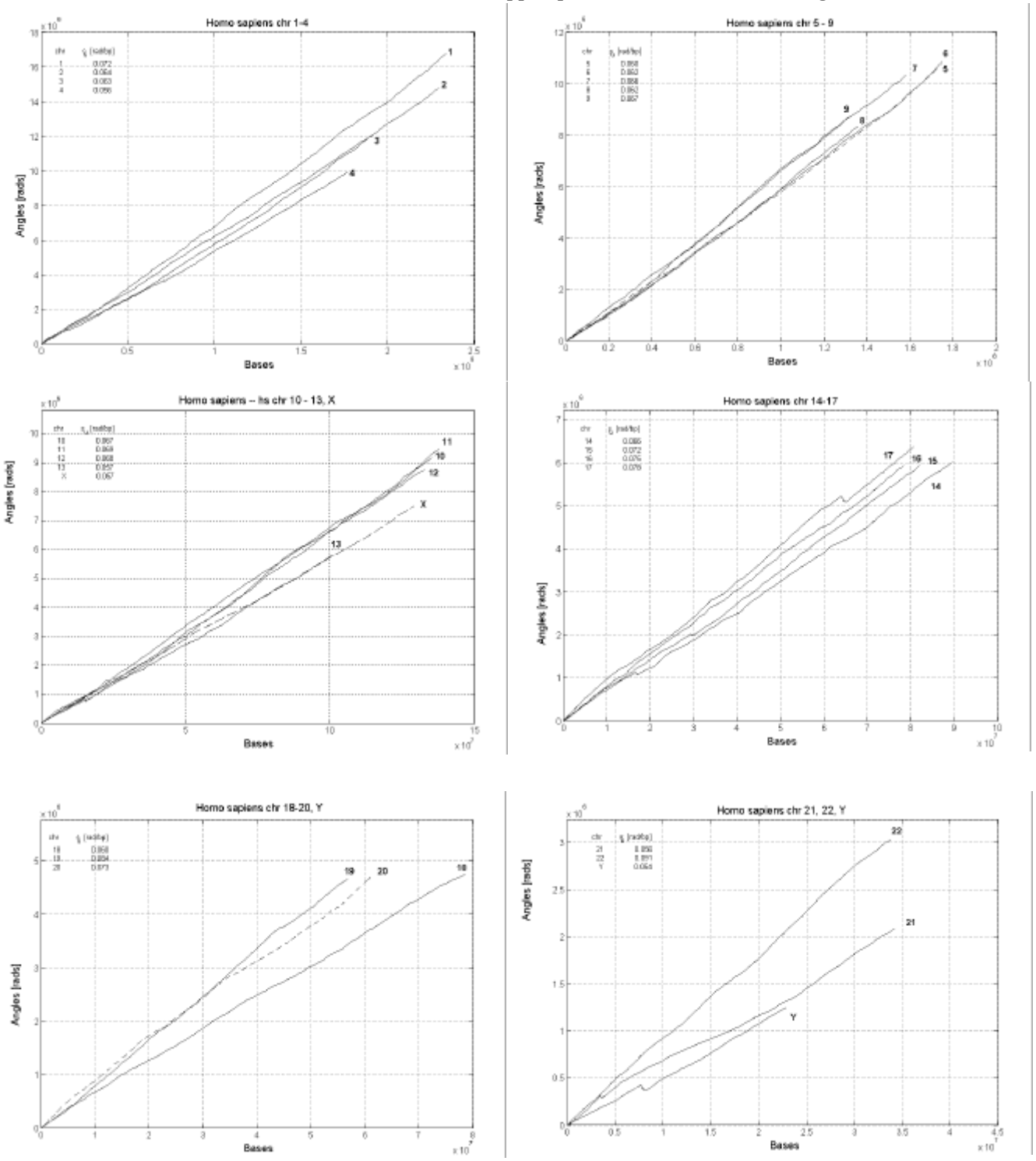

Fig. 4. Unwrapped phase along the concatenated contigs of all *homo sapiens* chromosomes

Figure 6 presents the same analysis for the complete genome of another prokaryote, the hyperthermophilic bacterium *Aquifex aeolicus* [9] (AE000657 [12]). In this case, the cumulated phase has a

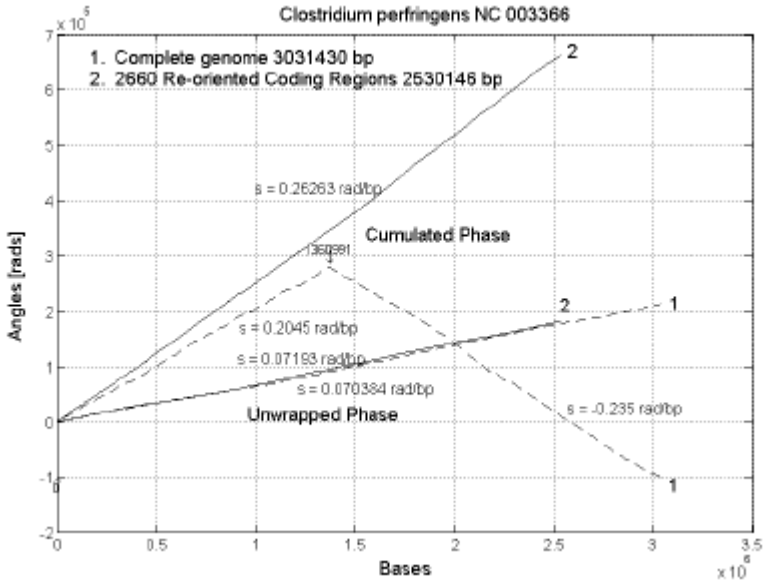

Fig. 5. Cumulated and unwrapped phase of the genomic signals for the complete nucleotide sequence and the concatenated re-oriented 2660 coding regions of *Clostridium perfringens* genome [14] (NC003366 [12]).

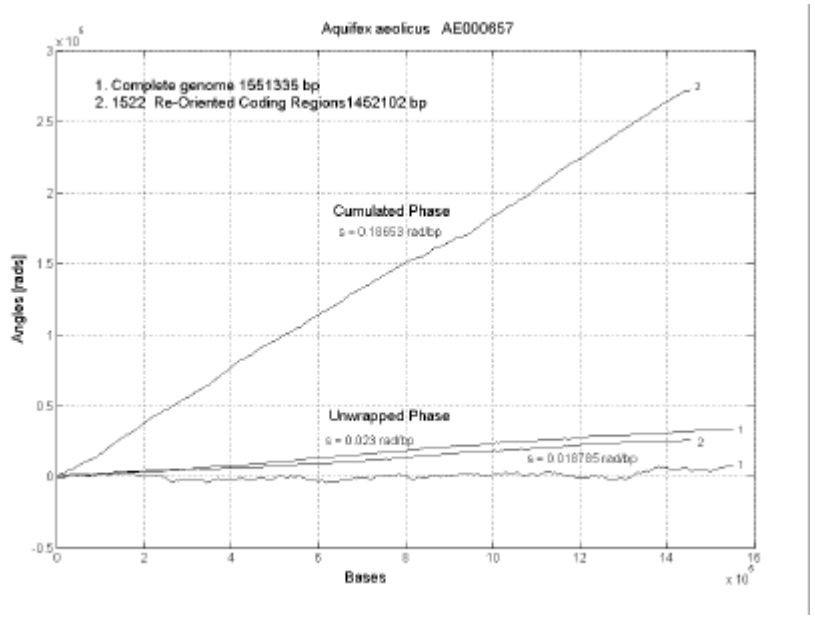

Fig. 6. Cumulated and unwrapped phase of the genomic signals for the complete nucleotide sequence and the concatenated re-oriented 1522 coding regions of *Aquifex aeolicus* genome [9] (AE000657 [12]).

quasi-random close to zero variation for the complete genome, and becomes linear when the 1522 coding regions are re-oriented along the positive sense of the molecule. Again, the unwrapped phase conserves its linear variation before and after the ORF re-orientation.

5. CONCLUSIONS

DNA sequences have been converted into genomic signals by using a complex quadrantal representation of the nucleotides that has been derived from the tetrahedral representation of nucleotides. The

main advantage of genomic signals over symbolic sequences is that they can be analysed by using signal processing methods. The paper presents results of the phase analysis of complex genomic signals for both prokaryotes and eukaryotes. Specifically, results of the phase analysis for all the chromosomes of the human genome and for two bacteria are given. The genomic signal cumulated phase and unwrapped phase are put in correspondence with the statistical distribution of bases and base-pairs, respectively. Large scale regularities, maintained over distances of 10^6 - 10^8 base pairs, *i.e.*, at the scale of whole chromosomes, are reported. This result contradicts the oversimplified genomic model that considers extra-genic regions as domains of randomness, recognizing only the meaningful structure of the exons [10,12,15].

In the case of prokaryotes, it is shown that the cumulated phase displays either a piece-wise linear, or a quasi-random, close to zero, variation along the DNA molecules, while the unwrapped phase has always an almost linear variation. In statistical terms, this result confirms the well known Chargaff's laws for the firs order distribution of nucleotides, but also reveals a similar law for the second order distribution of the nucleotide-to-nucleotide transitions.

The re-ordering in the same (positive) direction of all the coding regions of a chromosome, changes completely the variation of the cumulated phase, but leaves unchanged the unwrapped phase. The linearity of the cumulated phase for the re-ordered ORFs strongly suggests two hypotheses: (1) the existence of a primary ancestral genomic material from which the current DNA molecules have evolved, (2) the functional role of the particular orientation of direct and inverse ORFs, that generates specific densities of the first and second order repartition of nucleotides along chromosomes. The role of these large scale features in the control of crossing-over/recombination processes and the separation of species remain to be further investigated.

REFERENCES

1. CHARGAFF E., *Structure and function of nucleic acids as cell constituents*, Fed. Proc., **10**, 1951, pp. 654-659.
2. CRISTEA, P., *Genomic signals for whole chromosomes*, SPIE Conference, BiOS 2003 – International Biomedical Optics Symposium, Molecular Analysis and Informatics, San Jose, USA, BO 4962-19, January 25-31, 2003.
3. CRISTEA, P., *Large Scale Features in DNA Genomic Signals*, ELSEVIER, Signal Processing, Special Issue on Genomic Signal Processing, **83**, 2003, pp. 871-888.
4. CRISTEA, P., *Genomic Signals of Re-Oriented ORFs*, EURASIP JASP, Special Issue on Genomic Signal Processing, vol. 2004, 1 January 2004, no. 1, pp. 132-137.
5. CRISTEA, P., *Conversion of Nitrogenous Base Sequences into Genomic Signals*, Journal of Cellular and Molecular Medicine, **6**, 2, April – June 2002, pp. 279-303.
6. CRISTEA, P., *Genetic signal representation and analysis*, SPIE Conference, BiOS 2002 – International Biomedical Optics Symposium, Molecular Analysis and Informatics, San Jose, USA, BO 4623-10, January 21-24, 2002.
7. CRISTEA, P., *Genetic Signal Analysis*, ISSPA 2001 – The Sixth International Symposium on Signal Processing and its Applications, Invited Paper, Kuala Lumpur, Malaysia, August 13 – 16, 2001, pp. 703–706.
8. CRISTEA, P., *Genetic Signals*, Rev. Roum. Sci. Techn. Electrotechn. et Energ., **46**, 2, 2001, pp. 189-203.
9. DECKERT, G. et al, *The complete genome of the hyperthermophilic bacterium Aquifex aeolicus*, Nature, **392** (6674), 1998, pp. 353-358.
10. International Human Genome Sequencing Consortium, *Initial sequencing and analysis of the human genome*, Nature, **409**, February 15, 2001, pp. 860-911.
11. MYERS, E.W. et al., *A Whole-Genome Assembly of Drosophila*, Science, **287**, March 24, 2000, pp. 2196-2204.
12. National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, GenBank, *http://www.ncbi.nlm.nih.gov/genoms/.*
13. RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, *Functional annotation of a full-length mouse cDNA collection*, Nature, **409**, February 8, 2001, pp. 685-689.
14. SHIMIZU, T. et al, *Complete genome sequence of Clostridium perfringens, an anaerobic flesh-eater*, Proc. Natl. Acad. Sci. U.S.A., **99**, 2002, pp. 996-1001.
15. VENTER, J.C. et al., *Draft Analysis of the Human Genome by Celera Genomics*, Science, **291**, February 16, 2001, pp. 1304-1351.
16. WATSON, J.D., CRICK, F.H.C., *A Structure for Deoxyribose Nucleic Acid*, Nature, **171**, April 2, 1953, pp. 737.

Received October 27, 2003