Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT•

Signal Processing 83 (2003) 871–888

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Large scale features in DNA genomic signals

Paul Dan Cristea*

*Bio-Medical Engineering Center, Politehnica, University of Bucharest, Spl. Independentei 313,
Bucharest 77206, Romania*

Received 16 June 2002; received in revised form 20 August 2002

Abstract

Complex representations of the nucleotides, codons and amino acids derived from the projection of the Genetic Code Tetrahedron on adequately oriented planes are presented. By converting the sequences of nucleotides and polypeptides into digital genomic signals, this approach offers the possibility of using signal processing methods for the analysis of genomic information. New tools for genomic signal analysis are introduced at the nucleotide, codon and amino acid levels, in a multiresolution approach. It is shown that some important features of nucleotide sequences can be revealed using these signal representations. The paper reports the existence of large scale and global trends of DNA genomic signals in both eukaryotes and prokaryotes, reflecting an almost constant second order nucleotide statistics along DNA strands even at the points where the first order nucleotide statistics show marked changes, as it is the case in prokaryotes.

Keywords: Genomics; Genetic code; Genomic signals; Complex representation; Phase analysis; Unwrapped phase; Sequence path

1. Introduction

The sequencing of the human genome [17,23], as well as the public access to most of its content and to several other complete genomes [15,21,22], offer the opportunity to data mine and to explore in depth this unique information depository. The standard approach of representing the genomic information by sequences of nucleotide symbols in the strands of DNA and RNA molecules, sequences of codon symbols—triplets of nucleotides with the adequate reading frame (starting point) in the exons, or sequences of amino acid symbols in the corresponding polypeptide chains—limits the methodology of handling the genomic information to mere pattern matching or statistical procedures. The use of interdependent mappings of the nucleotide, codon and amino acid symbols to real or complex numbers allows converting the genomic sequences into digital genomic signals and offers the possibility to apply a wealth of powerful signal processing methods for their analysis [1,6–11]. The current Special Issue of Signal Processing on Genomic Signal Processing witnesses the increasing interest in the use of GSP as an efficient approach to handling, processing and knowledge eliciting from genomic data.

Currently, only the about 50,000 genes containing instructions to make proteins—but representing less than 5% of the human genome—are considered of interest. The remaining vast majority of the genome is considered “junk DNA” [2,14]. One of the arguments to sustain this view is that the intergenic part of the human genome contains repetitive, quasi-random sequences and a large amount of transposable elements

* Tel.: +40-1-411-44-37; fax: +40-1-410-44-14.
E-mail address: pcristea@dsp.pub.ro (P.D. Cristea).

0165-1684/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0165-1684(02)00477-2

that bear a close resemblance to the DNA of independent entities like viruses and bacteria. Nevertheless, significant parts of the inter-gene chromosomal DNA most likely play an important role in the control of protein synthesis, conjointly with the recently identified gene regulatory networks.

The main genetic material of eukaryote cells is represented by the nucleic DNA molecules that have a beautiful and well studied structure, first described in the famous Watson–Crick paper [24]. The double helix DNA molecule comprises two antiparallel intertwined complementary strands, each consisting of a helicoidally coiled linear, one-dimensional and one-directional sequence of nucleotides. The repetitive unit, the nucleotide, is made up of three parts: a phosphate group, the deoxyribose—a sugar stripped of a certain oxygen atom, and a nitrogenous base linked by strong covalent bounds. There are four kinds of nucleotides that differ by their nitrogenous bases: thymine (T) and cytosine (C)—which are pyrimidines, adenine (A) and guanine (G)—which are purines. Along the two strands of the DNA double helix a pyrimidine in one chain always faces a purine in the other and only the complementary base pairs T–A and C–G exist. The weak 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 a protein, are called genes. The genes are made up of several exons—coding regions, that are separated along the DNA strand 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) different codons, while the terminator is encoded by three codons.

Proteins are the main contributors to the cell structure and, as enzymes, catalyze the chemical reactions specific to the functioning of the cells. Almost everything in the organism is made of or by proteins. The primary structure of a protein is given by the polypeptide chains formed of amino acid sequences. A protein contains one or several such polypeptide chains. The coiling (secondary structure), folding (tertiary structure) and aggregation (quaternary structure) of the polypeptides generate the final very complex spatial structure of the protein, essential for its biological functions. Any gene starts with the codon ATG that also encodes Methionine. 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. Remarkably enough, 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, 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 mRNA codon sequence into an amino acid sequence. Each of the 20 amino acids is brought by a specific transfer RNA (tRNA).

There is a sharp contrast between the deceivingly simple structure of DNA nucleotide chains—unbranched linear code written in a four letters alphabet—and the overwhelming complexity of the protein 3D structure built of 20 amino acids. As mentioned, there are only about 50,000 genes in the human genome, but millions of proteins, many of them transitory. Nevertheless, the exons from the nucleotide chains and the primary structure of the proteins are the bearers of essentially the same genetic information.

The rest of the paper is organized as follows: Section 2 presents the tetrahedral representation of the nucleotides and codons [6] that leads to a natural ordering of the GC expressing its specific structure and features. By projecting the tetrahedral representation on adequately chosen plans, complex representations of nucleotides, codons and amino acids are obtained. On this basis, the symbolic sequences of nucleotides in DNA molecules are converted into complex digital genomic signals. Section 3 presents large scale and global features of Genomic Signals for both eukaryote and prokaryote genomes. New tools for genomic signal analysis [10,11] are presented, including the use of phase, cumulated phase, unwrapped phase, sequence path, stem representation of sequence components' relative frequencies, as well as the transition analysis at the nucleotide, codon and amino acid levels in a multiresolution approach. It is shown that there are long range regularities of the DNA molecules that contradict Ohno's assertion that genomes consist of scarce gene oases in an otherwise essentially empty, unstructured desert [2,14,15,22]. An almost linear variation of the average unwrapped phase over distances of tens of millions of bases, i.e., at the scale of whole chromosomes, has been found. This long range feature, which is equivalent to an average helicoidal coiling of the nucleotide complex representations along DNA sequences, holds for all the investigated genomes, being shared by both prokaryotes and eukaryotes. The magnitude and sense of the helicoidal wrapping is specific for various taxa and chromosomes. Such a behavior proves a rule similar to Chargaff's rule [4] for the *distribution* of nucleotides—a first order statistics, but reveals a statistical regularity of the *succession* of the nucleotides—a second order statistics. The compliance of a certain contig with the large scale regularities of the chromosome to which it belongs can be used to spot out exceptions and errors. The conclusions of the paper are presented in Section 4.

2. Tetrahedral representation

2.1. Nucleotide tetrahedron

We proposed [6–10] a tetrahedral representation of the nucleotides: four equal length vectors, symmetrically placed with respect to each other, i.e., oriented towards the corners of a tetrahedron, are put in correspondence with the nucleotides as shown in Fig. 1, in which it is emphasized that the vertices of a regular tetrahedron are a subset of the vertices of a cube.

The mathematical description of the resulting code can be simplified by choosing integer $\{\pm 1\}$ coordinates for the vertices of this cube, including the points representing the bases, without imposing a normalization condition, so that the base vectors take the simple form:

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

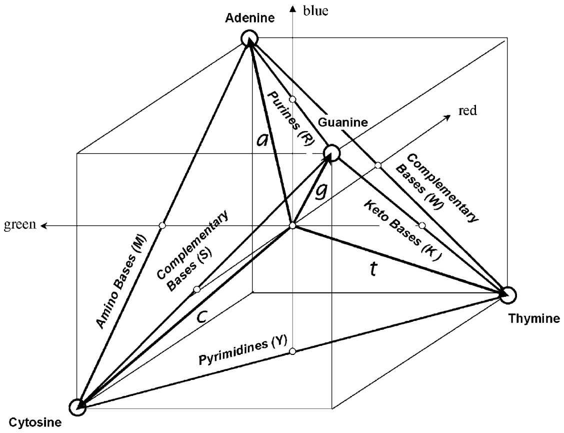

Fig. 1. Nucleotide tetrahedron.

To make the genomic sequences easier to read for a human observer, the three color components of the RGB system can be assigned to the three axes of the representation space [1,10]. Consequently, each point in the representation space, each nucleotide in the case of Fig. 1, corresponds to a distinct hue. This approach is useful for the fast exploration of DNA sequences at the nucleotide level and can be readily extended at the codon and amino acid levels.

The dimensionality of the representation can be reduced to two by projecting the basic tetrahedron on a plane. The projection planes can be chosen in various ways that conserve the symmetry of the representation and reflect some selected features of the nucleotides in corresponding mathematical properties. A simple way is to choose planes defined by pairs of coordinate axes. On the other hand, the projection planes can be put in correspondence with the complex plane, so that a *complex representation* of the nucleotides is obtained.

Choosing the *red–blue* (xz) plane, the four bases are placed in quadrantal symmetry as shown in Fig. 2 and the complex representation of the bases is

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

In this representation the complementarity of the pairs of bases A–T and C–G, respectively, is expressed by the symmetry with respect to the real axis

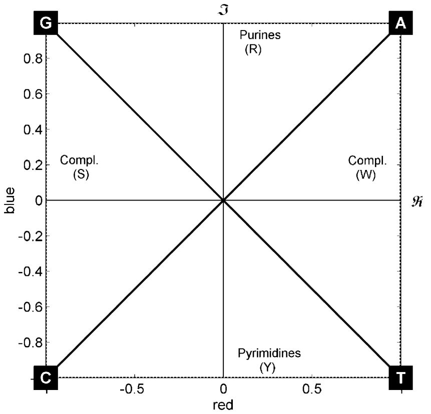

Fig. 2. Nucleotide complex representation obtained by projecting the nucleotide tetrahedron in Fig. 1 on the *red–blue* (xz) plane.

(the representations are complex conjugates: $t = a^*$, $g = c^*$), while the chemical similarity of the pairs A–G (purines) and C–T (pyrimidines) is expressed by the symmetry with respect to the imaginary axis (equality of imaginary parts and equality of absolute values, but opposite signs of real parts). Relations (2) and similar ones are equivalent to representing the nucleotides in two mutually orthogonal complex (bipolar) binary systems, instead of one real system of base four. The use of bipolar systems with the distance 2 between first-order neighbor numbers results in assigning odd real and odd imaginary components to the codons. The division by two of the values attached to the bases makes the distance between neighboring numbers equal to one, but has the disadvantage of operating with non-integer components.

Various other complex representations can be used [8–11]. For instance, the A–G bases can be permuted by projecting the tetrahedron in Fig. 1 on the *reversed_green–blue* $(-yz)$ plane, or equivalently, by mirroring the tetrahedron with respect to the *pyrimidine-blue* diagonal plane and projecting it on the *red–blue* (xz) plane:

$$\begin{aligned} a &= -1 + j, \\ c &= -1 - j, \\ g &= 1 + j, \\ t &= 1 - j. \end{aligned} \tag{3}$$

In this case, the A–T and C–G complementary is expressed in opposite signs $(t = -a,\ g = -a)$, so that the two strands of a DNA molecule correspond to digital signals of equal absolute values, but opposite sign and ordering. i.e., if the sequence of samples in the signal for one of the strands is $\{b_1, b_2, \ldots, b_n\}$, than for the antiparallel strand the signal is $\{-b_n, -b_{n-1}, \ldots, -b_1\}$. The amino–keto relationship is expressed by the symmetry with respect to the real axis, while the purine-pyrimidine relationship by the symmetry with respect to the imaginary axis.

2.2. Codon and amino acid tetrahedrons

2.2.1. The tetrahedral representation of the genetic code

The vector and complex mappings of the nucleotides described above can be extended for the representation of the codons and amino acids [10]. The nucleotides forming a codon can be considered the digits of a three digit number written in base two: the vector corresponding to the first nucleotide of the codon, i.e., the most significant digit, is multiplied with $2^2 = 4$, the vector corresponding to the second nucleotide in the codon with 2, and the vector corresponding to the last nucleotide in the codon, i.e., the least significant digit, is left unchanged. For instance, the vector representing the codon ATG that encodes *Methionine* and also signals the start of a gene is $4\vec{a} + 2\vec{t} + \vec{g}$. In this way, each of the 64 codons is attached to one of the vertices of a resulting three-level fractal-like tetrahedron structure—the *codon tetrahedron*. By mapping the 20 amino acids and the terminator to the codons in accordance to the GC, one obtains the *amino acid tetrahedron* [6–11]. Again, the dimensionality can be reduced by projecting the codon and the amino acid tetrahedrons on adequately chosen planes that can be put in correspondence with the complex plane to get various complex representations of the Genetic Code. It turns out that the GC tetrahedral representation, as well as the complex representations based on it, reflect adequately the metric structure of the GC. This can readily be noticed from the fact that codons corresponding to the same amino acid are mapped to neighboring points of the tetrahedron representation, i.e., related codons are clustered [10]. Correspondingly, the complex mappings cluster the multiple representations of the same amino acid in contiguous regions of the complex plane, with the exception of the three amino acids that have a degeneration of order six. On the other hand, the degeneration is basically restricted to the second order tetrahedrons and most pairs of interchangeable bases are distributed on the edges parallel to the pyrimidine or purine directions. A distance between the amino acids can be defined based on the probability. The tetrahedron representation of the genetic code has also the advantage to naturally indicate putative ancestral coding sequences by the simple passage to a lower level tetrahedron. Thus, the tetrahedron representation grasps some essential features of the GC which appear as symmetries and regularities in the resulting 3D image.

Complex values can be assigned in various ways to the amino acids either as domains, or as punctual values. In the later case, to a certain amino acid can be attached the average value over the whole area onto which it is mapped. It is also possible to take into account the relative frequencies of occurrence of the different codons that correspond to the amino acid. On the other hand, the assigning of the complex values to the bases and to the amino acids can be adapted to the specific problem.

2.3. *Real representations*

It is possible to further reduce the dimensionality of the representation of base, codon and amino acid sequences by using a real one-dimensional mapping [6–11]. The digits $\{0,1,2,3\}$ are attached to the four nitrogenous bases. The three-base-codons are interpreted as three-digit-numbers written in base four, i.e., the codons along the DNA strands are mapped to the numbers $\{0,1,2,\ldots,63\}$. Actually, a whole DNA sequence can be seen as a huge number written in base four. Nevertheless, it better corresponds to the biological reality to interpret each codon as a distinct sample of a digital genomic signal distributed along the DNA strand. There are $4! = 24$ choices for attaching the digits 0–3 to the bases A, C, G, T. The optimal choice given in Table 1 results from the condition to obtain the most monotonous mapping of the codons 0–63 to the amino acids plus the terminator 0–20, that leads to best auto-correlated inter-gene genomic signals [6–11].

The numerical codes assigned to the amino acids result from the order of their first reference when gradually increasing the codons from 0 to 63. By convention, the code zero is assigned to the terminator. As mentioned above, there are only two *one codon—one amino acid* (non-degenerated) mappings—for tryptophan and methionine, but nine double, one triple, five quadruple, and three sextuple degeneration, plus the three codons corresponding to the terminator. The minimum non-monotonic dependency has only four reversals of the normal ascending order: for a terminator sequence and for the three instances of sextuple degeneration: leucine, serine and arginine. An exhaustive search for all the 24 possible correspondences of the nitrogenous bases to the digits 0–3 has shown that there does not exist a more monotonic mapping. The proposed mapping gives a piece-wise constant function, with only the three mentioned reversals of the order [6–11].

Table 1
Mapping of nucleotides to digits in base four

Pyrimidines	Purines
Thymine = T = 0	Adenine = A = 2
Cytosine = C = 1	Guanine = G = 3

3. Large scale features of genomic signals

Prompted by the significance of genes for practical purposes, primarily for the pharmaceutical industry, important efforts have been made not only to finish the sequencing of the human genome and other eukariote and prokariote genomes, but also to analyze and annotate these sequences at the resolution of one base or one codon, and to identify exons in order to enable the synthesis of potentially useful proteins [5,12,15,22]. In this paper we have chosen a rather different approach: the study of genomic signals mainly at scales of 10^4–10^8 bp, to detect general trends of the genomic signals, potentially significant in revealing their basic properties and to search for specific genomic signals with possible control functions. Contigs from several eukaryote and prokaryote genomes have been downloaded from GenBank of NIH [22] and converted to complex genomic signals using the mapping given in Eq. (2) and Fig. 2 at nucleotide level. Several tools have been developed for their analysis and the results are presented in the following sections.

3.1. Phase, cumulated phase and unwrapped phase

The *phase* of a complex number is a periodic multi-valued magnitude: adding or subtracting any multiple of 2π to or from the phase of the complex number does not change the number itself. To remove the ambiguity, the standard mathematical convention restricts the phase of a complex number to the domain $(-\pi, \pi]$ that covers only once all the possible orientations of the associated vector (phasor) in the complex plane. For the complex representations given in Eq. (1) and Fig. 1, the *phases* of the complex vectors attached to the four nucleotides have the values $\{-3\pi/4, -\pi/4, \pi/4, 3\pi/4\}$ radians.

The *cumulated phase* is the sum of the phases of the complex numbers in a sequence from the first element in the sequence, up to the current element. For the complex representations of DNA nucleotide sequences, the cumulated phase drifts between positive and negative values, its current value giving an indication on the relative frequencies of the purines (A, G) vs. pyrimidines (C, T) in the segment under consideration. Because of the bias introduced by the conventional restriction of the phase to the domain $(-\pi, \pi]$, which favors π over $-\pi$, the cumulated phase can be distorted when using complex representations that include pure real negative numbers; e.g., this is the case of the representation [10]:

$$\begin{aligned} a &= -1, \\ c &= -j, \\ g &= j, \\ t &= 1, \end{aligned} \tag{4}$$

for which the *phase* of the nucleotide complex representations takes the values $\{-\pi/2, 0, \pi/2, \pi\}$ radians. Consequently, the average phase of a uniform distribution of nucleotides would not be zero, but $\pi/4$. This unwanted effect, which would bias the phase analysis, has been avoided by introducing a uniform complex noise, i.e., by adding uniformly distributed small random complex numbers to each of the nucleotide representations in the sequence. For the noisy complex sequence, the phases close to $-\pi$ are equally probable with the phases close to $+\pi$ and there is no spurious drift of the cumulated phase towards positive values.

The *unwrapped phase* is the corrected phase of the elements in a sequence of complex numbers, in which the absolute value of the difference between the phase of each element in the sequence 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. The unwrapped phase eliminates the phase jumps introduced by the standard conventional restriction of the phase domain described above and allows observing the true global phase trends along a sequence. The value of the unwrapped phase gives an indication on the relative frequencies of the transitions between the nucleotides. For the complex representation given in Eq. (2), the *positive transitions* A → G, G → C, C → T, T → A determine an increase of the unwrapped phase, corresponding to a rotation in the trigonometric sense, the *negative transitions* A → T, T → C, C → G, G → A determine a decrease, corresponding to a clockwise rotation, while all other transitions are *neutral*. Care must be taken to avoid the bias described in the previous paragraph referring to the cumulated phase.

The cumulated phase and the unwrapped phase have well defined long range trends which are specific for the different eukaryote and for prokaryote genomes. As an example, Fig. 7 presents the cumulated phase and the unwrapped phase of the nucleotide complex representations for the *Homo sapiens* (hs) chromosome (chr) 11 contig with accession number NT 030106 [22] that contains 9,991,932 bp. The unwrapped phase shows a steady and almost linear increase along the contig, with an approximate average slope $s_u = 0.09$ rad/bp $= 5.2°$/bp.

The linear increase of the unwrapped phase in a sequence of nucleotide complex representations along the corresponding DNA strand shows that the complex representations form in the average a *counterclockwise helix* that completes a turn over the spatial period:

$$L = \frac{2\pi}{s_u}, \tag{5}$$

where s_u is the slope of the unwrapped phase. For the example in Fig. 3, it results $L = 70$ bp.

It is remarkable that this counterclockwise helicoidal wrapping of the complex representations of nucleotides along the DNA strands is a common long range feature for all the chromosomes of *Homo sapiens*. The genomic complex signals for all the other investigated genomes also long range helices with

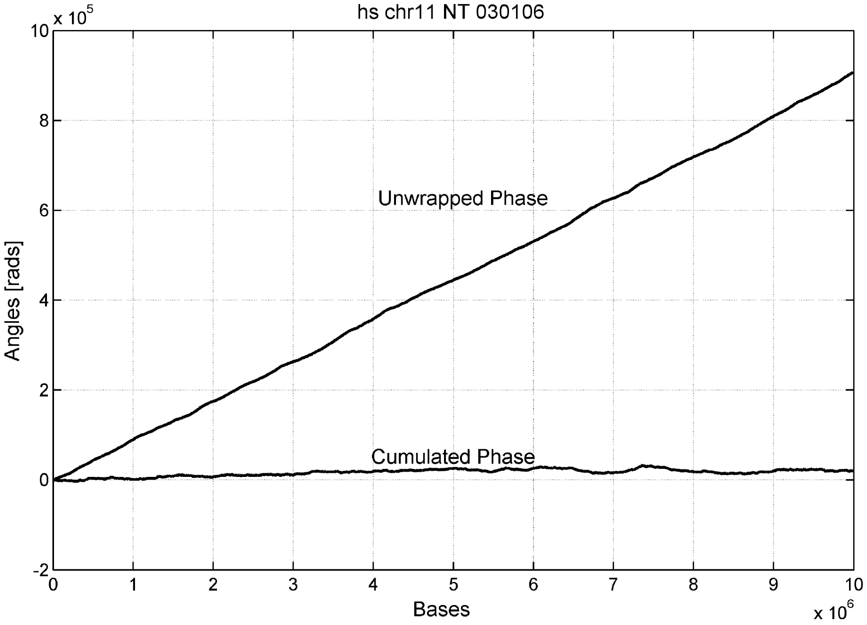

Fig. 3. Cumulated and unwrapped phase along the contig NT_030106 [22] of *Homo sapiens* chr 11.

specific parameters that depend on the species and the section of the genome under consideration. The trend is maintained over distances of tens of millions of bases and reveals a statistical regularity in the *distribution of the succession* of bases (transitions from one base to the other), not only in the *distribution of the bases themselves*. This result contradicts the current oversimplified genomic model that considers the exons separated by domains of almost complete randomness, the so called "junk DNA" [2,14].

For the complex representation given by (2), the slope s_u of the variation of the unwrapped phase along the DNA strand, thus the spinning of the complex representation helix, is related to the difference Δf_{pm} between the frequency f_+ of positive transitions and the frequency f_t: of the negative ones by the relation:

$$s_u = \frac{\pi}{2}(f_+ - f_-) = \frac{\pi}{2}\Delta f_{pm}. \qquad (6)$$

Similarly, the slope $s_c$of the variation of the cumulated phase along a DNA strand is related to the frequencies of occurrence of the bases by the equation

$$s_c = \frac{\pi}{4}[3(f_G - f_C) + (f_A - f_T)], \qquad (7)$$

that can also be expressed as

$$s_c = \frac{\pi}{2}[(f_R - f_Y) + (f_K - f_M)/2] = \frac{\pi}{2}\Delta f_{RhK}, (8)$$

where f_R is the frequency of occurrence of purines $R = A + G$, f_Y of pyrimidines $Y = C + T$, f_K of keto bases $K = G + T$ and f_M of amino bases $M = A + C$. For the example in Fig. 3, the average slope of the unwrapped phase $s_u = 0.09$ rad/bp corresponds to a plus–minus difference of transition frequencies $\Delta f_{pm} = 5.7\%$/bp, while the average slope of the cumulated phase $s_c = 2$ mrad/bp corresponds to a purine-half-keto difference of occurrence frequencies $\Delta f_{RhK} = 0.12\%$/bp.

The almost linear variation of the unwrapped phase of the nucleotide complex representation along the DNA strands shows that the difference between the frequencies of the two classes of transitions remains approximately constant along the sequences under consideration. Fig. 4 shows the cumulated phase and unwrapped phase along the ensemble of all concatenated contigs of *hs* chr 11. The average slope of the unwrapped phase is $s_u = 0.0667$ rad/bp, while the various contigs have slopes in the range between 0.047 rad/bp = 2.7°/bp and 0.120 rad/bp = 6.9°/bp.

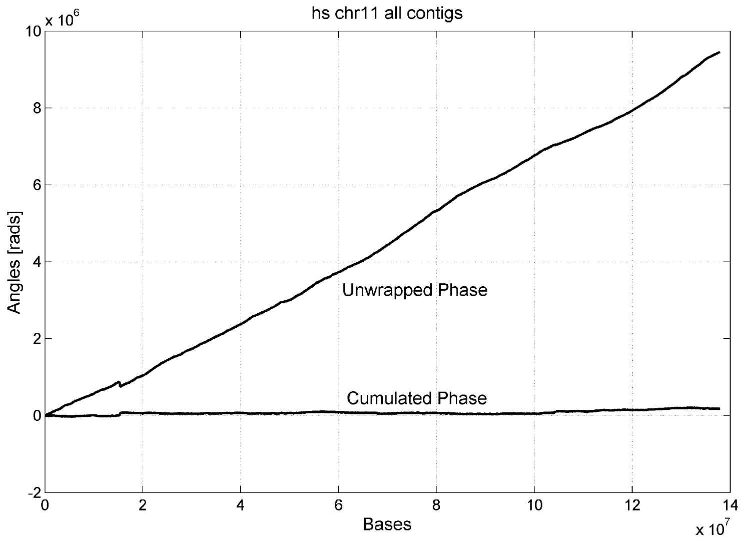

Fig. 4. Cumulated and unwrapped phase along the concatenated contigs [22] of *hs* chr 11.

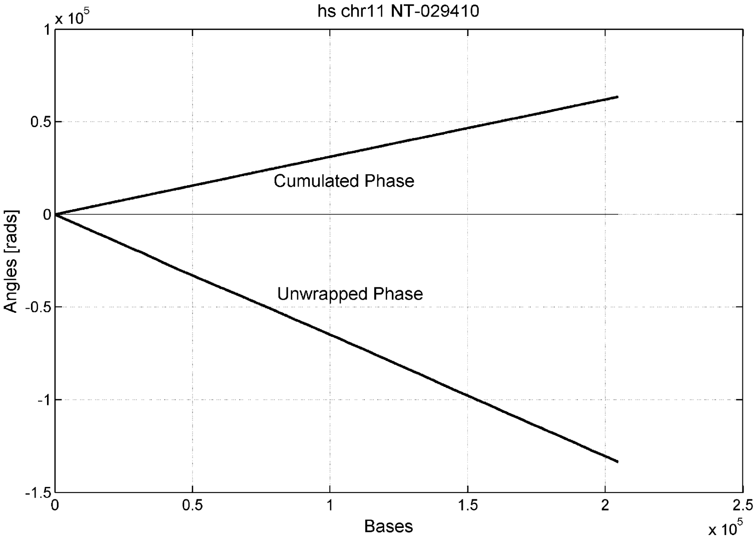

Fig. 5. Cumulated and unwrapped phase along contig NT_029410 [22] of *hs* chr 11.

A striking exception is found in the interval ~ 15.17–15.38 Mbp of the concatenated string of contigs in Fig. 4 and corresponds to the contig of accession NT 029410 [22] of *Homo sapiens* chr 11 for which the nucleotide complex representation phases are shown in Fig. 5. On a length of about 210 kbp, the unwrapped phase decreases linearly with a sharp average slope $s_u = -0.65$ rad/bp $= -37.2°$/bp, which corresponds to a large negative difference in the frequencies of positive and negative transitions $\Delta f_{pm} = f_+ - f_- = -39.4\%$/bp and to a nucleotide average helix oriented clockwise, completing a turn for about every 9.7 bp. At the same time, the cumulated phase increases linearly with a slope

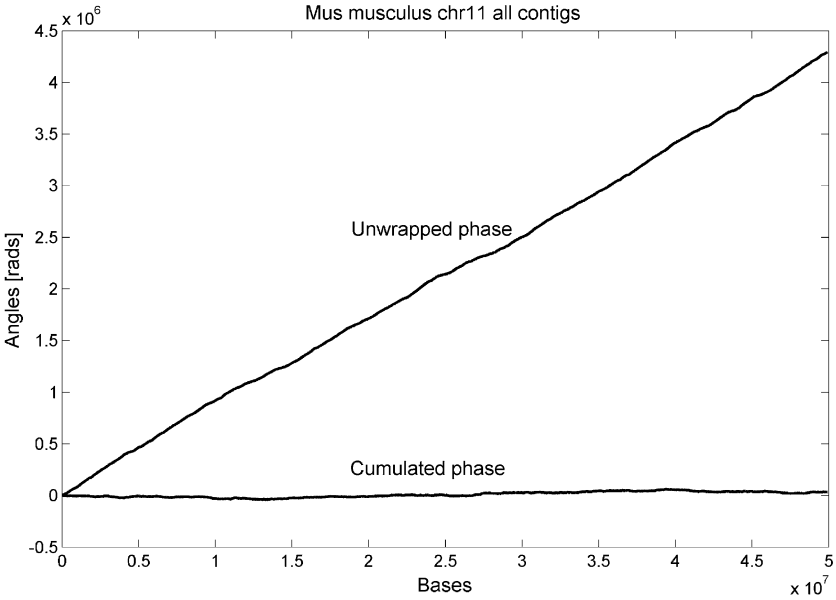

Fig. 6. Cumulated and unwrapped phase along concatenated contigs of *mus musculus* chr 11 [22].

$s_c = 0.325$ rad/bp $= 18.6°$/bp, ($s_c \approx^{|s_u|}$ /2), which corresponds to $\Delta_{\mathrm{RhK}} = 19.7\%$/bp.

Similar large scale properties can be found in other eukaryote genomes. As seen in Fig. 6 that shows the phase diagram for all concatenated contigs of *mus musculus* chr 11, the unwrapped phase increases also almost linearly with an average slope $s_u = 0.086$ rad/bp $= 4.93°$/bp, while the cumulated phase varies very little.

On the contrary, markedly different large scale features are found for the genomic signals of most prokaryote organisms. For the reference direction and the origin of the DNA strand chosen along the circular chromosome of *Escherichia coli* in the GenBank sequence NC-000913 [22], the unwrapped phase of the DNA complex nucleotide signal decreases on the average with a negative slope $s_u = -0.041$ rad/bp $= -2.35°$/bp (Fig. 7), which corresponds to a clockwise turn of the complex signal helix for about $L = 153$ bp and, in accordance to Eq. (6), to a difference in the frequencies of the positive and negative transitions $\Delta f_{\mathrm{pm}} = -2.61\%$/bp. Even more conspicuous, the long range cumulated phase also varies piece-wise linearly along two domains of the circular DNA having almost equal length: a region of positive slope $s_{c+} = 0.0393$ rad/bp $= 2.25°$/bp of length $l_+ = 2,266,409$ bp (split into two sub domains $1\ldots1,550,413$ bp and $3,923,226\ldots4,639,221$ bp by the choice of the origin of the sequence) and a region of negative slope $s_{c-} = -0.0375$ rad/bp $= -2.15°$/bp of length $l_- = 2,372,812$ bp. Each extreme has been attributed arbitrarily to the domain on its left. According to Eq. (8), the switch of the sign of the slope of the cumulated phase when passing from one region to the other shows a switch of the nucleotide type distribution along the analyzed strand from purine and keto excess $\Delta f_{\mathrm{RhK}+} = 2.50\%$/bp to pyrimidine and amino excess $\Delta f_{\mathrm{RhK}-} = -2.39\%$/bp (see also Figs. 18a and b) As it can be seen from Fig. 7, there is no corresponding change of the unwrapped phase slope in the extreme points and, according to Eq. (6), there is no switch of the frequencies of positive and negative transitions between the nucleotides (see also Figs. 17a and b). In a simple model, a circular chromosome is composed from two linear sections connected in a ring, one of the sections being reversed and having the strands switched. Consequently, the two inversions between clockwise (negative) and counterclockwise (positive) orientations of the nucleotide-to-nucleotide transitions along the GenBank strand induced by the

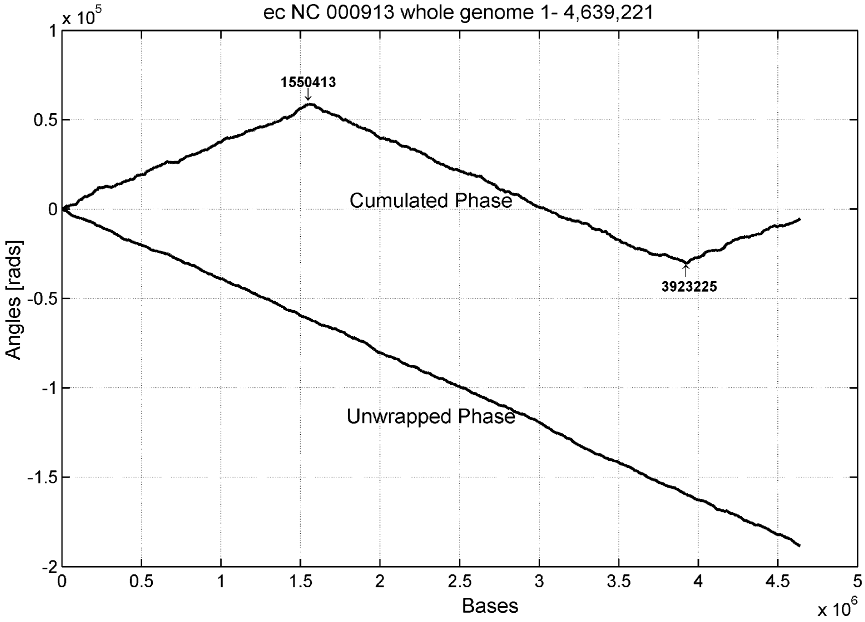

Fig. 7. Cumulated and unwrapped phase along the chromosome of *Escherichia coli*, NC-000913 [22].

reversal and by the switching cancelled out and there is no change in the unwrapped phase diagram.

We have found similar features from the phase analysis of the genomic signal for the circular DNA of *Bacillus subtilis* (bs) with the accession number NC-000964 [22] (Fig. 8). The following parameters have been calculated from the unwrapped phase curve: $s_u = -0.057$ rad/bp $= -3.27°$/bp, $L = 110$ bp, $\Delta f_{\mathrm{pm}} = -3.64\%$/bp, and from the cumulated phase curve: $s_{c+} = 0.106$ rad/bp $= 6.07°$/bp, $l_+ = 1,941,643$ bp, $\Delta f_{\mathrm{RhK}_+} = -6.75\%$/bp—for the region of positive slope, and $s_{c-} = -0.0965$ rad/bp $= -5.53°$/bp, $l_- = 2,273,121$ bp, $\Delta f_{\mathrm{RhK}_-} = -6.14\%$/bp— for the region of negative slope.

Other approaches to the whole-genome analysis of prokaryotes, based on skew [19,20] and integrated skew [13,16] plots, have been presented in literature. The integrated skew diagrams look quite similar to the cumulated phase curves in the case of prokaryote genomes, but the cumulated phase given by (7) and (8) for the complex representation (2) is a different function than the skew functions (G–C)/(G + C), (A − T)/(A + T) or the similar ones. The cumulated phase conserves all the information about the corresponding DNA strand, so that a symbolic sequence can be re-constructed from its cumulated phase diagram. This is not the case for the skew diagrams, that extract each a partial information, emphasizing certain specific properties of DNA strands. The same holds true for the purine-excess or the keto-excess functions. On the other hand, the skew plots are constructed using a sliding window of a certain (arbitrary) width that limits the resolution of the analysis, while the cumulated phase has a one-base resolution. The extremes of the integrated skew diagram have been put in relation with the origins and termini of chromosome replication. The differences in base occurrence frequencies have been explained by the differences in mutation probabilities resulting from the asymmetry of replication mechanisms for the leading and lagging strands. In the case of *Escherichia coli*, the origin of replication has been found experimentally at 3,923,500 bp, and the terminus at 1,588,800 bp [3], while the GC skew minimum has been found at $84.5\% \times 4,639,222 = 3,896,946$ bp ± 0.1%, and the maximum at $33.4\% = 1,549,500 \pm 0.1\%$ [13,16]. The corresponding extremes of the cumulated phase diagrams for *E. coli* are found at 3,923,225 bp and

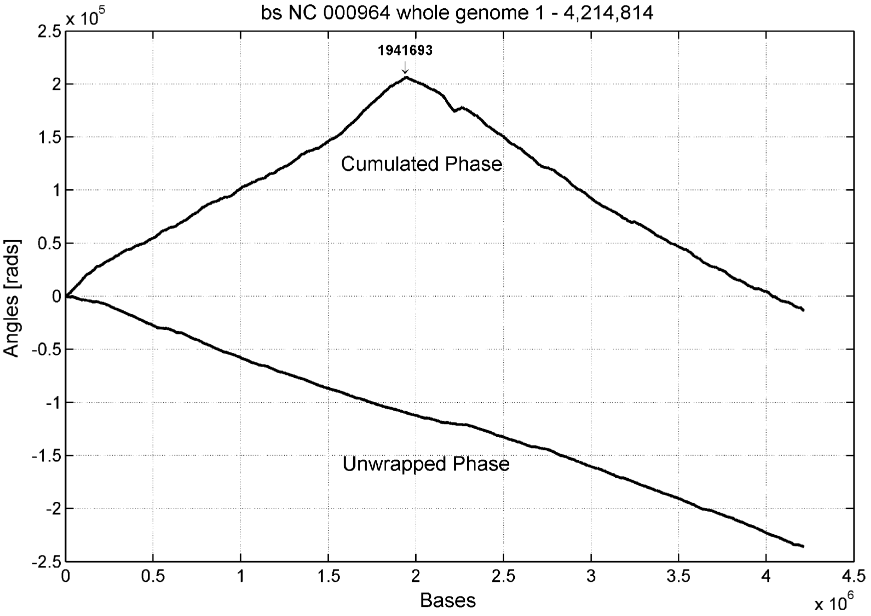

Fig. 8. Cumulated and unwrapped phase along the chromosome of *Bacillus subtilis*, NC-000964 [22].

$1,550,413$ bp, respectively, closer to the experimental data than the GC, AT or purine-excess prediction. Similarly, for *Bacillus subtilis*, the experimental origin and terminus of chromosome replication are in positions 1 bp and 2,017,000 bp [18], respectively, again a little closer to the extremes 1 bp and $1,941,693$ bp of the cumulated phase in Fig. 8, than from the points 1 bp and $46.0\% \times 4,214,814 = 1,939,814$ bp $\pm 0.1\%$, resulted from skew diagrams [13,16].

To ensure the reliability and robustness of the unwrapped phase diagram as a large scale feature detector for nucleotide sequences and to rule out possible artifacts, the above results have been verified in several ways and additional precautions have been taken. Firstly, as explained above, some low power complex noise has been added to each sample in the complex representation sequence to eliminate any possible bias related to the conventional restriction of the phase domain that could affect in this instance the difference of phase when using crisp data. Secondly, various complex representations have been used, changing the orientation and the value of the complex representations of the bases, but keeping their relative positions and differences of phase unchanged. E.g., Figs. 9a and b show the cumulated phase and the unwrapped phase of the nucleotide complex signals for the same contig NT-011005 [22] (1-3355638 bp) of *Homo sapiens* chr 18. The complex representation in Eq. (2) has been used in Fig. 9a, while the complex representation in Eq. (4), after permuting of *a* with *g*, has been used in Fig. 9b.

This way, the second representation corresponds to a counterclockwise rotation with $\pi/4$ rad and a contraction with $\sqrt{2}$ of the previous representation. As expected, the cumulated phase largely changes because it corresponds to other base occurrence frequencies in the two instances: $\Delta f_{\text{a}} = \frac{1}{2}\,[3(f_{\text{G}} - f_{\text{C}}) + (f_{\text{A}} - f_{\text{T}})]$ and $\Delta f_{\text{b}} = f_{\text{A}} - f_{\text{C}}$, respectively. On the other hand, the unwrapped phase remains the same, as it is determined only by the relative phases of the nucleotide complex representations in the sequence and not by the individual phases,. The vertical translation of about $\pi/4$ rad resulting from the change of mapping (2) with mapping (4) cannot be observed at the scale of Fig. 9.

In the third place, artificial uniform random sequences have been synthesized and their features have been compared to those of natural genomic signals of similar length. Again as expected, the cumulated phase and the unwrapped phase of the artificial complex

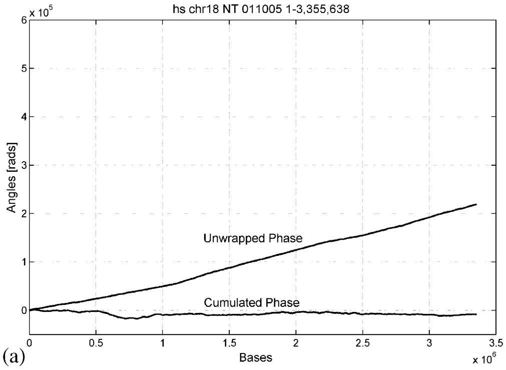

Fig. 9. Cumulated and unwrapped phase of the contig NT-011005 [22], *hs* chr 18, for the complex representations in (a) Eq. (2) and (b) Eq. (4) with switched a and g. The representation (b) differs from (a) by a 45° counterclockwise rotation and a $\sqrt{2}$ contraction. The cumulated phase—depending on individual base representations—changes significantly, while the unwrapped phase—which depends only on the differences of phase—remains essentially unchanged.

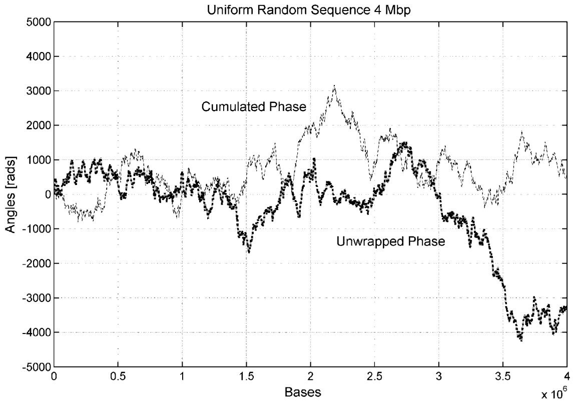

Fig. 10. Cumulated and unwrapped phase of the complex signal of a synthetic uniformly random sequence comprising 4 Mbp. The complex representation in Eq. (2) has been used.

signals remain at much lower absolute values and show no systematic variation along the sequence—as can be seen in Fig. 10, where the ordinate scale is two orders of magnitude smaller than in Fig. 9.

Finally, it has also been verified on synthetic sequences that the long range linear variation of the unwrapped phase can be controlled by the frequency of the transitions between the complex elements in the sequence and that this trend disappears when the order of the elements is randomized, even when the number of elements of each type is not changed.

3.2. Sequence path

Another approach to explore large scale trends, but also local features of complex sequences, is computing the *sequence path* in the complex plan, i.e., the cumulative sum of the elements along the analyzed sequence. Fig. 11 shows the nucleotide sequence path for the *Homo sapiens* chr 11 contig with accession number NT-030106 [22] comprising 9, 991, 932 bp for which the phase diagrams is shown in Fig. 4. Like for all the other contigs of the human genome that have been analyzed, the resulting *nucleotide sequence path* for the representation in Eq. (2) displays a marked tendency to progress preferentially in the complex plane from the origin in the positive direction of the real axis. This tendency results from the known global higher occurrence frequency of the complementary pair W = A + T over the complementary pair S = C + G in the extra-genic regions of the *hs* genome, which are predominant, but also from the detailed balance of A and T, on one hand, and C and G, on the other—the well known Chargaff's rules [4], that result in the global approximate balance between purines (R) and pyrimidines (Y) along the strands of a DNA molecule.

To indicate the advancement along the sequence path, equal length intervals (1 Mbp) have been marked on the curve. The non-uniformity of the advancement shows the changes in the nucleotide distribution along the sequence. As revealed in the phase analysis, the large scale trend in the distribution of bases is

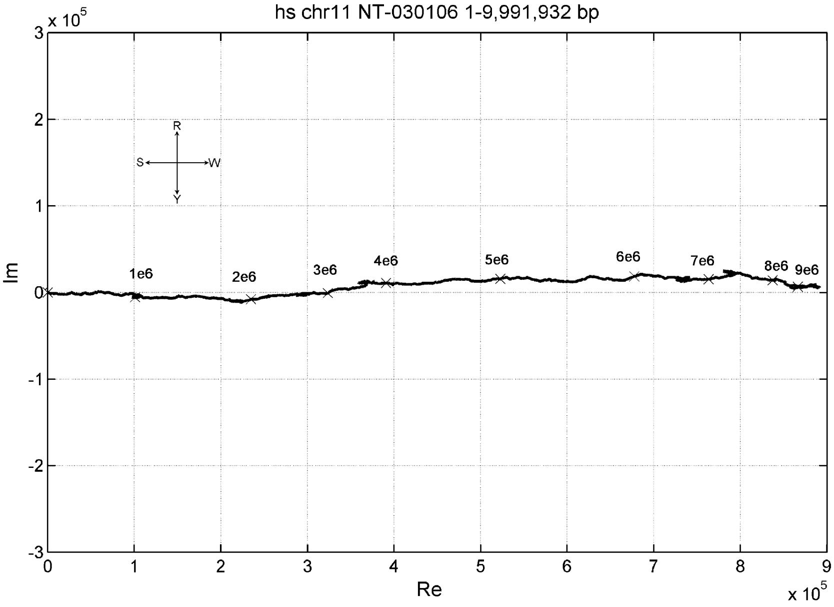

Fig. 11. Nucleotide Path for the *hs* chr 11 contig with accession number NT-030106 [22] comprising 9, 991, 932 bp. Each marked interval is 1 Mb long.

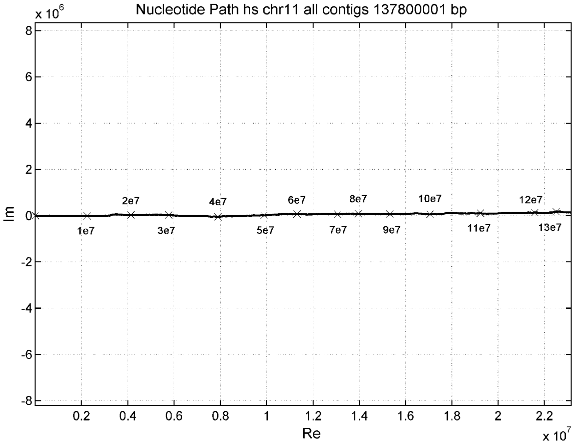

Fig. 12. Nucleotide Path for all concatenated contigs of *hs* chr 11 [22].

maintained over distances larger than 100 Mb. This defines a basic reference frame for the encoding segments, in which the local nucleotide distribution varies around the average value that gives the general trend.

Fig. 12 shows the nucleotide path for all concatenated contigs of *Homo sapiens* chromosome 11 [22] comprising about 137.8 Mb. At the scale of the (almost) whole chromosome, the preferential advancement along the real axis of the complex plane becomes even more noticeable, with a relatively small drift along the imaginary axis.

As already shown by phase analysis, there are contigs with largely different behavior. Fig. 13 presents

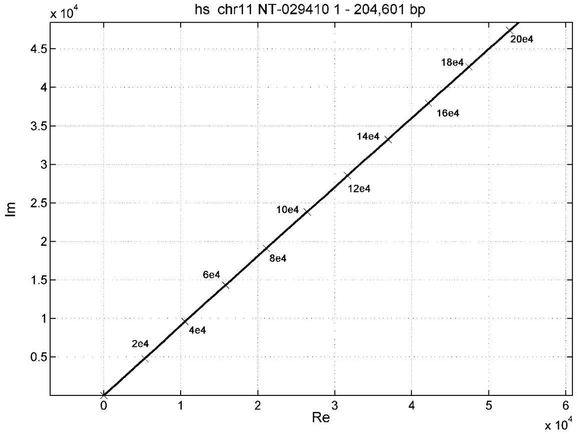

Fig. 13. Nucleotide Path for the *hs* chr 11 contig with accession number NT-029410 [22] comprising 204, 601 bp. Each marked interval is 10^4 bp long.

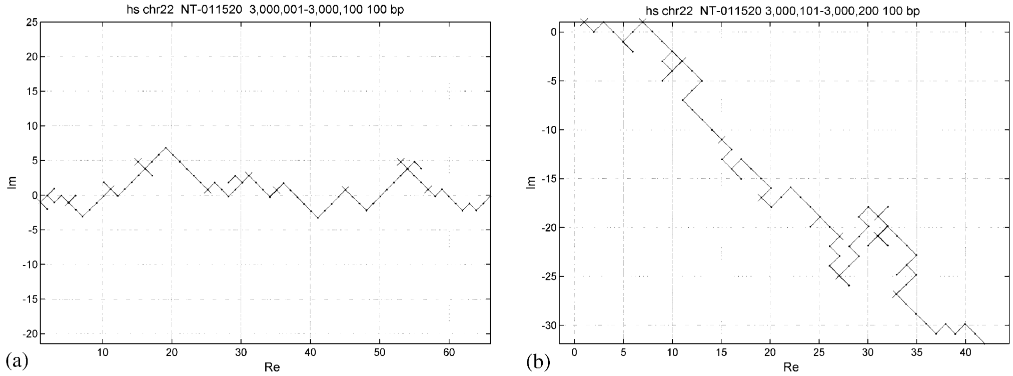

Fig. 14. Nucleotide Path for segments of the *hs* chr 22 contig with accession number NT-011520 [22] comprising each 100 bp. Marked intervals are 1 kbp long each.

the nucleotide path for *Homo sapiens* chr 11 contig with the accession number NT-029410 [22] comprising 204601 bp, for which the phase diagrams are represented in Fig. 5. The path is still almost linear, but advances along both the real and imaginary axes, forming with the real positive half-axis an angle of tangent 0.9125. This shows a higher occurrence frequency of adenine over cytosine and a quite balanced distributions of guanine and thymine in this area.

The sequence path in the complex plane is a useful instrument for investigating the statistical distribution of the elements in a sequence at various resolutions. Fig. 14 gives details of the nucleotide path for two segments of *hs* chromosome 22, each about 100 nucleotides long. Some local features can be recognized by direct visual inspection of such images.

In contrast, the sequence path of a uniformly distributed random sequence of bases remains confined in a quite small area around the origin of the complex plane.

3.3. Stem representations

Another convenient tool for investigating the nucleotide occurrence frequency (first order statistics) and the nucleotide-to-nucleotide transition frequency (second order statistics) along a certain segment of a DNA sequence is the *stem representation*. Such a study is useful to compare the results of the phase analysis of genomic signals with the statistics of the corresponding symbolic sequences. Fig. 15 shows the stem representation of the nucleotide-to-nucleotide transitions along the first 1 Mbp segment of the *hs* chr 11 contig NT_030106 (a), and along the *hs* chr 11 contig NT_029410 (b), which have the phase diagrams given in Figs. 3 and 5, respectively. The values of the difference $\Delta f_{\mathrm{pm}} = f_{+} - f_{-}$ between the positive and negative transition frequencies given in Figs. 15a and b from the statistical analysis are close to those calculated in Section 3.1 with Eq. (6) from the slope of the unwrapped phase. The small difference comes from the estimation of the unwrapped phase average slope by linear regression on the whole contig. The rather unusual features of contig NT-029410 are expressed in the markedly non-uniform distribution of nucleotide-to-nucleotide transition frequencies readily noticeable in Fig. 15b.

The stem representations of the nucleotide occurrence statistics for the same two DNA segments of *hs* chr 11 are shown in Figs. 16a and b. Again, there are significant changes in the nucleotide statistics, the most conspicuous being the reversal of the small pyrimidine excess for the first 1 Mbp segment of the contig NT_030106 (Fig. 16a) into a very large purine excess for contig NT_029410 that comprises only 204,601bp (Fig. 16b). The small stem in the origin of the complex plane corresponds to not yet identified bases in the publicly available genomic sequences [15,22] for which the system assigns the conventional default code 0 when a complex representation is used.

Figs. 17a and b present the nucleotide-to-nucleotide transition statistics for two 1 Mbp DNA segments belonging to the two replichores of the *Bacillus subtilis* circular chromosome with the accession number

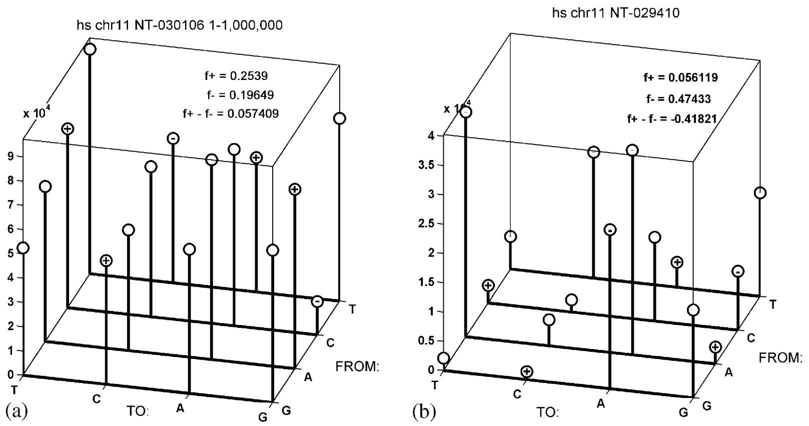

Fig. 15. Stem representation of the nucleotide-to-nucleotide transition (second order) statistics for the first 1 Mbp segment of *hs* chr 11 contig NT_030106 (a), and for *hs* chr 11 contig NT_029410 comprising 204, 601 bp (b).

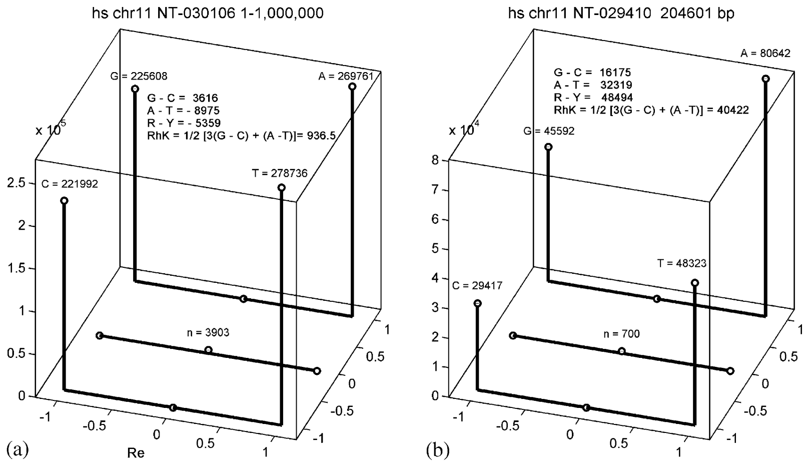

Fig. 16. Stem representation of the nucleotide occurrence statistics for the two DNA segments of *hs* chr 11 shown in Fig. 15.

NC-000964. Figs. 18a and b present the nucleotide occurrence statistics for the same segments. Despite the switch from a purine excess (R–Y = 56, 814) to a pyrimidine excess (R–Y = −66, 868), obvious in the nucleotide statistics of Figs. 18a and b, accompanied by the switch of *RhK* from +64669 to −69757, there is no corresponding change in the transition statistics that remain almost the same, with $\Delta f_{\mathrm{pm}} \approx -3.7\%$ for Fig. 17a and $\Delta f_{\mathrm{pm}} \approx -3.1\%$ for Fig. 17b. The transition statistics stem plots in Figs. 17a and b are quite similar, while the nucleotide statistics stem plot in Figs. 18a and b look almost like mirrored with respect to the middle vertical plane separating the complementary pairs AT and GC. The stem representations have the advantage of making clear the difference in the first and second order statistics properties.

The stem representation can be coupled with a window sliding along the nucleotide strand, allowing a direct visual investigation of the number of nucleotides of each type in the segment comprised in the window. The size of the window can be chosen according to the purpose of the scan. Local monitoring of base relative frequencies allows identifying the regions with specific local features.

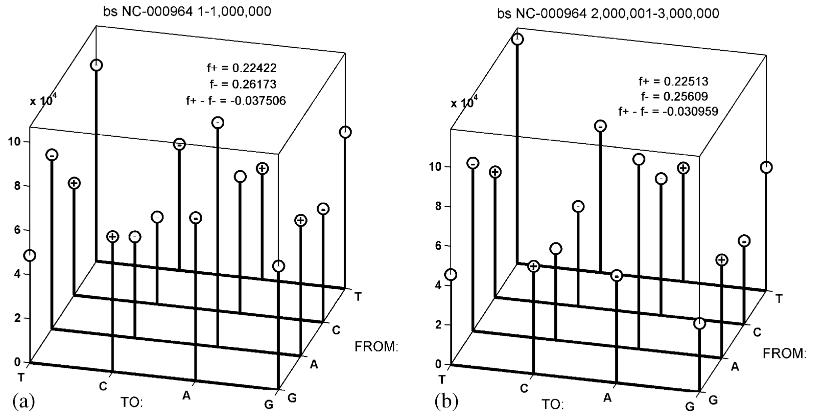

Fig. 17. Stem representation of the nucleotide-to-nucleotide transition statistics for two 1 Mbp segments belonging to the two replichores of the *Bacillus subtilis* chromosome NC-000964.

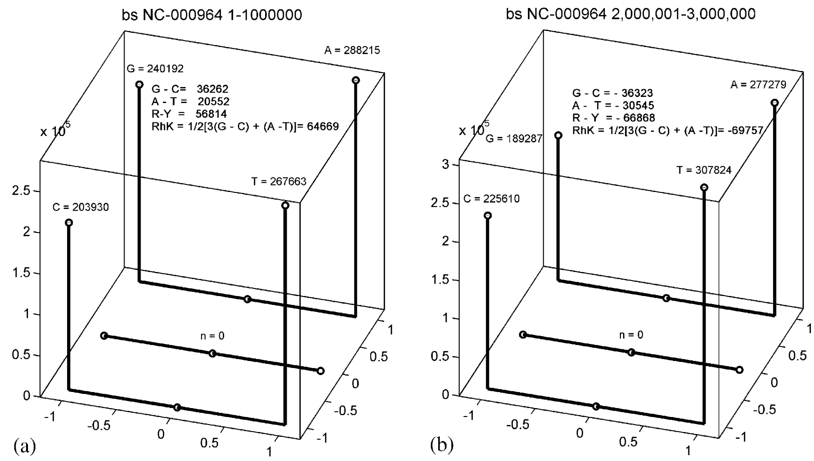

Fig. 18. Stem representation of the nucleotide occurrence statistics for the two 1 Mbp segments of the *Bacillus subtilis* chromosome NC-000964 shown in Fig. 17.

As shown above, stem representations of order one statistics (base distribution) are correlated with the cumulated phase diagrams, while second order statistics stem representations (nucleotide-to-nucleotide transition distributions) are correlated with the unwrapped phase diagrams.

4. Discussions and conclusions

The paper presents some new results in the representation and analysis of genomic information: (1) tetrahedral representation based methods to map genomic symbolic sequences to real or complex signals, (2) new analysis tools for the study of genomic signals, mainly at scales of 10^4–10^8 bp to detect large scale features, (3) results of the large scale analysis for several eukaryote and prokaryote genomes.

The Tetrahedron Representation expresses the symmetry and the degeneration (redundancy) of the genetic code and generates various mappings of the symbolic nucleotide, codon or amino acids sequences to genomic signals that also translate selected sequence properties in corresponding signal properties.

The new tools for genomic signal analysis comprise the phase analysis—phase, cumulated phase and unwrapped phase, the sequence path, and the stem representation of occurrence and transition statistics of DNA sequences, in a multiresolution approach.

An almost linear variation of the unwrapped phase over distances of tens of millions of bases has been found. This is a long range feature that is maintained at the scale of whole chromosomes and shows the existence of a average uniform helicoidal coiling of the complex representations of the bases along DNA sequences. The property holds for all the investigated genomes, being shared by both prokaryotes and eukaryotes, while the magnitude and sense of the helicoidal wrapping is specific for various taxa and chromosomes. Such a behavior proves a rule similar to Chargaff's rule [4], but reveals a statistical regularity of the *succession* of the nucleotides—a second order statistics, not merely of the *distribution* of nucleotides—a first order statistics. In statistical terms, this means that the difference between the frequencies of positive and negative nucleotide-to-nucleotide transitions along the strand of a chromosome tends to be small and constant.

At the same time, interesting behavior of the cumulated phase has been found for most sequenced prokaryote genomes. The circular chromosomes are divided in two almost equal domains in which the slope of the variation of the cumulated phase along the strand has opposite signs and which can be put in correspondence with the replichores of the chromosome, in a way similar to the approach based on skew diagrams. Moreover, the almost linear variation of the unwrapped phase along the chromosomes of prokaryotes continues over the points where the cumulated phase switches from positive to negative slopes or the reverse. The comparison between the behavior of the cumulated phase and of the unwrapped phase across the putative origins and termini of the replichores suggests an interesting model for the equivalence of circular and linear chromosomes.

Both the cumulated phase and the unwrapped phase conserve each all the information about the corresponding DNA strand, so that a symbolic sequence can be re-constructed from either of them, which have a one-base resolution.

To fully reveal the biological significance of the newly found large scale features of the genomic signals additional co-operative work is necessary. Nevertheless, even at this point, the degree of compliance of a certain contig with the large scale regularities of the chromosome it belongs to can be used to spot out contigs that comprise either special information, or possible errors.

The use of the Projection Pursuit approach, specifically of the Independent Component Analysis (ICA), for the genomic signals derived from extra-gene DNA sequences that do not encode proteins, could reveal signals that play an important role in the control of gene expression, i.e., in protein synthesis, working conjointly with the recently described molecular chains control mechanism.

References

[1] D. Anastassiou, Frequency-domain analysis of biomolecular sequences, Bioinformatics 16 (12) (2000) 1073–1081.

[2] J.O. Andersson, W.F. Doolittle, C.L. Nesbø, Are there bugs in our genome? Science 292 (2001) 1848–1850.

[3] F.R. Blattner, et al., The complete genome sequence of *Escherichia coli* K-12, Science 277 (5331) (September 1997) 1453–1474.

[4] E. Chargaff, Structure and function of nucleic acids as cell constituents, Fed. Proc. 10 (1951) 654–659.

[5] J.M. Claverie, Computational methods for the identification of genes in vertebrate genomic sequences, Human Mol. Genet. 6 (10) (1997) 1735–1744.

[6] P. Cristea, Independent component analysis for genetic signals, SPIE Conference BiOS 2001, International Biomedical Optics Symposium, SC316, Short Course, San Jose, USA, 20–26 January 2001, 21–24 January 2002.

[7] P. Cristea, Genetic signal analysis, Proceedings of ISSPA 2001, The Sixth International Symposium on Signal Processing and its Applications, Kuala Lumpur, Malaysia, 13–16 August 2001, Invited Paper, pp. 703–706.

[8] P. Cristea, Genetic signal representation and analysis, SPIE Conference BiOS 2002, International Biomedical Optics Symposium, Conference 4623, Functional Monitoring and Drug–Tissue Interaction, San Jose, USA, 19–25 January 2002.

[9] P. Cristea, Genetic signals, Rev. Roum. Sci. Techn. Electrotechn. Energ. 46 (2) (April–June 2001) 189–203.

[10] P. Cristea, Conversion of nitrogenous base sequences into genomic signals, J. Cell. and Mol. Med. 6 (2) (April–June 2002) 279–303.

[11] P. Cristea, Real and complex genomic signals, Proceedings of the DSP 2002, 14th International Conference on Digital Signal Processing, Santorini, Greece, 1–3 July 2002, pp. 543–546.

[12] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis, Cambridge University Press, Cambridge, 1998.

[13] J.M. Freeman, et al., Patterns of genome organization in bacteria, Science 279 (1998) 1827–1832.

[14] H. Gee, A journey into the genome: what's there, Nature, 12 February 2001, http://www.nature.com/nsu/010215/010215-3.html.

[15] The Genome Data Base, http://gdbwww.gdb.org/, Genome Browser, http://genome.ucsc.edu, European Informatics Institute, http://www.ebl.ac.uk, Ensembl, http://www.ensembl.org.

[16] A. Grigoriev, Analyzing genomes with cumulative skew diagrams, Nucleic Acids Res. 26 (10) (1998) 2286–2290.

[17] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature 409 (February 2001) 860–911.

[18] F. Kunst, et al., The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*, Nature 390 (1997) 249–256.

[19] J.R. Lobry, Origin of replication of *Mycoplasma genitalium*, Grams. Nucleic Acids Res. 26 (1996) 2286–2290; Science 272(3) (May 1996), 745–746.

[20] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (5) (May 1996) 660–665.

[21] E.W. Myers, et al., A whole-genome assembly of Drosophila, Science 287 (2) (March 2000) 2196–2204.

[22] National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/, ftp://ftp.ncbi.nlm.nih.gov/genoms/, GenBank, http://www.ncbi.nlm.nih.gov/Genbank/index.html.

[23] J.C. Venter, 272 others, The sequence of the human genome, Science 291 (2001) 1304–1351.

[24] J.D. Watson, F.H.C. Crick, A structure for deoxyribose nucleic acid, Nature 171 (2) (April 1953) 737–738.