J.Cell.Mol.Med. Vol 6, No 2, 2002 pp. 279-303

Special Article: **Getting DNA to numbers**

Conversion of nucleotides sequences into genomic signals

P. D. Cristea *

Bio-Medical Engineering Center, "Politehnica" University of Bucharest, Romania

Received: March 11, 2002; Accepted: April 29, 2002

Abstract

An original tetrahedral representation of the Genetic Code (GC) that better describes its structure, degeneration and evolution trends is defined. The possibility to reduce the dimension of the representation by projecting the GC tetrahedron on an adequately oriented plane is also analyzed, leading to some equivalent complex representations of the GC. On these bases, optimal symbolic-to-digital mappings of the linear, nucleic acid strands into real or complex genomic signals are derived at nucleotide, codon and amino acid levels. By converting the sequences of nucleotides and polypeptides into digital genomic signals, this approach offers the possibility to use a large variety of signal processing methods for their handling and analysis. It is also shown that some essential features of the nucleotide sequences can be better extracted using this representation. Specifically, the paper reports for the first time the existence of a global helicoidal wrapping of the complex representations of the bases along DNA sequences, a large scale trend of genomic signals. New tools for genomic signal analysis, including the use of phase, aggregated phase, unwrapped phase, sequence path, stem representation of components' relative frequencies, as well as analysis of the transitions are introduced at the nucleotide, codon and amino acid levels, and in a multiresolution approach.

Keywords: genomics - genetic code - genomic signals - complex representation - phase analysis - unwrapped phase - sequence path

Introduction

The almost complete sequencing of the human genome [1, 2], as well as the public access to most of its content [3, 4], offer the opportunity to explore in depth its content and to data mine 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, 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 handling the genomic information to mere pattern matching or statistical procedures. Using a base 4 real representation or an equivalent complex dual binary representation of the nucleotides, allows converting the DNA sequences into digital genomic signals and offers the possibility to apply a wealth of powerful signal processing methods for their analysis. Currently, only about 32000 genes containing the

* Correspondence to: Paul Dan CRISTEA
Spl. Independentei 313, 77206 Bucharest, Romania.
Tel.: +40-1-411 44 37, Fax: +40-1-410 44 14.
E-mail: pcristea@dsp.pub.ro. http://www.dsp.pub.ro

instructions to make proteins, but representing less than 5 percent of the human genome, are considered of interest. The remaining vast majority of the genome is considered "junk DNA" [5, 9]. One of the arguments to sustain this view is that the inter-gene 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. Nevertheless, significant parts of the inter-gene chromosomal DNA play very probable an important role in the control of protein synthesis, in parallel with the recently described molecular chains control mechanism.

The main nucleic genetic material of the cells is represented by the DNA molecules that have a basically simple and well studied structure [10]. The *double helix DNA* molecule comprises two antiparallel intertwined complementary *strands*, each consisting of a linear, one-dimensional and one-directional sequence of *nucleotides*. The repetitive unit, the nucleotide, is made up of three parts: a phosphate group stripped of a certain oxygen atom - from where the prefix "deoxy", a sugar - the "ribose", and a *nucleotide*. The nucleotides differ only by the *nucleotides* they contain. Only four kinds of nucleotides are found in DNA: 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 base pairs T-A and C-G exist. A simple model of the DNA molecule is shown in Fig. 1. The segments of the nucleotide chains that encode a *polypeptide*, *i.e.*, give 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. Triplets of successive nucleotides in the exons form *codons* that, according to the *Genetic Code*, encode the 20 *amino acids* found in the *polypeptide* chains, as well as the *terminator* - marking the end of an encoding segment. *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 unbranched polypeptide chains. The coiling (secondary structure), folding (tertiary structure) and aggregation (quaternary structure) of the polypeptides generate the final very complex spatial

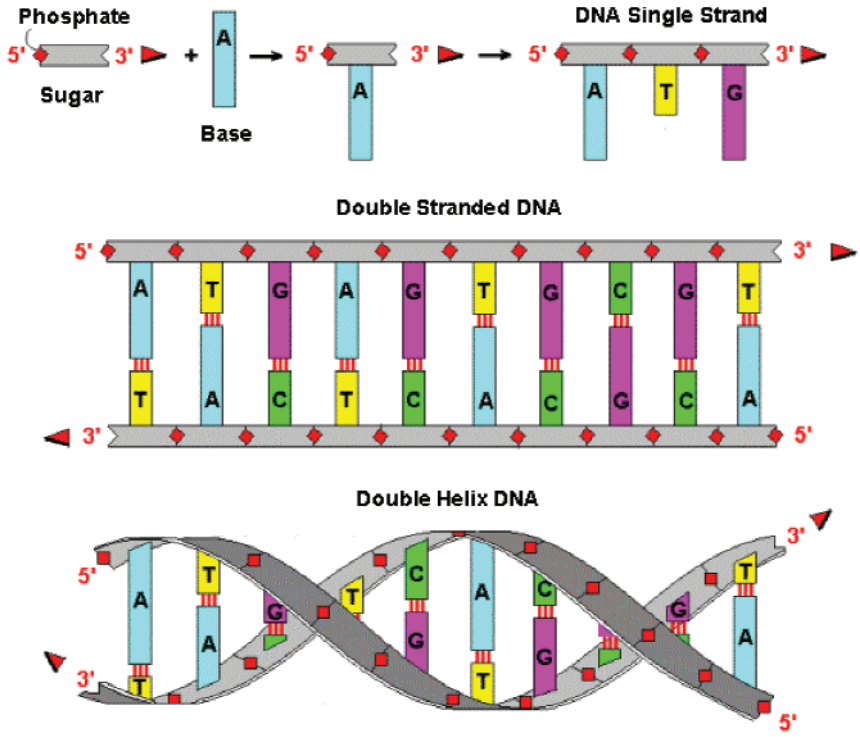


Fig. 1 Schematic model of DNA molecule helicoidal structure.

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. Remarkable 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 nucleotide, but the CG remains otherwise the same. A *polypeptide* chain is synthesized by *ribosomes* that move along the mRNA and *translate* the 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 twenty amino acids. As mentioned, there are only about 32,000 genes in the human genome, but millions of proteins, many of them transitory. Nevertheless, the nucleotide chains and the proteins are the bearers of the essentially the same genetic information.

The paper describes a new tetrahedral modality of representing the GC that better grasps its specific

Table 1. Standard form of the genetic code.

First position in codon	Second position in codon: T			Second position in codon: C			Second position in codon: A			Second position in codon: G			Third position in codon
T	TTT	Phe	[F]	TCT	Ser	[S]	TAT	Tyr	[Y]	TGT	Cys	[C]	T
	TTC	Phe	[F]	TCC	Ser	[S]	TAC	Tyr	[Y]	TGC	Cys	[C]	C
	TTA	Leu	[L]	TCA	Ser	[S]	TAA	Ter	[end]	TGA	Ter	[end]	A
	TTG	Leu	[L]	TCG	Ser	[S]	TAG	Ter	[end]	TGG	Trp	[W]	G
C	CTT	Leu	[L]	CCT	Pro	[P]	CAT	His	[H]	CGT	Arg	[R]	T
	CTC	Leu	[L]	CCC	Pro	[P]	CAC	His	[H]	CGC	Arg	[R]	C
	CTA	Leu	[L]	CCA	Pro	[P]	CAA	Gln	[Q]	CGA	Arg	[R]	A
	CTG	Leu	[L]	CCG	Pro	[P]	CAG	Gln	[Q]	CGG	Arg	[R]	G
A	ATT	Ile	[I]	ACT	Thr	[T]	AAT	Asn	[N]	AGT	Ser	[S]	T
	ATC	Ile	[I]	ACC	Thr	[T]	AAC	Asn	[N]	AGC	Ser	[S]	C
	ATA	Ile	[I]	ACA	Thr	[T]	AAA	Lys	[K]	AGA	Arg	[R]	A
	ATG	Met	[M]	ACG	Thr	[T]	AAG	Lys	[K]	AGG	Arg	[R]	G
G	GTT	Val	[V]	GCT	Ala	[A]	GAT	Asp	[D]	GGT	Gly	[G]	T
	GTC	Val	[V]	GCC	Ala	[A]	GAC	Asp	[D]	GGC	Gly	[G]	C
	GTA	Val	[V]	GCA	Ala	[A]	GAA	Glu	[E]	GGA	Gly	[G]	A
	GTG	Val	[V]	GCG	Ala	[A]	GAG	Glu	[E]	GGG	Gly	[G]	G

structure and features. On this basis, an the symbolic sequences of nucleotides in DNA molecules are converted into digital genomic signals that can be more efficiently processed using existing digital signal processing (DSP) algorithms.

The rest of the paper is structured as follows: Section 2 presents the tetrahedral representation of the GC. Optimal symbolic-to-digital mappings of the genomic information as contained in the linear, one-dimensional and one-directional strands of DNA and RNA molecules, as well as in the primary structure of the corresponding proteins (for genes) are given. "Natural" correspondences between the nucleotides and real or complex digits are discussed, leading to certain "optimal" mathematical properties of the corresponding genomic signals. Section 3 discuses properties of Genetic Signals, focussing on the large scale and global features. New tools for genomic signal analysis, including the use of phase, aggregated phase, unwrapped phase, sequence path, stem representation of components' relative frequencies, as well as the transition analysis are introduced at the nucleotide, codon and amino acid levels, and in a multiresolution approach. It is shown that there are long range regularities of the DNA molecules that contradict Ohno's assertion that mammalian genomes consist of gene oases in an otherwise essentially empty, unstructured deserts [3, 4, 5, 9]. Specifically, it is shown for the first time that the nucleotide complex representations form a counterclockwise helix along the DNA strands that is maintained for ten of millions of base pairs. The results are discussed in Section 4 and the conclusions of the paper are presented in Section 5.

The tetrahedral representation of the genetic code

The Genetic Code (GC) is universal, being used by most known organisms, with only small variations in mitochondria and certain microbes. The GC applies to all known nuclear genetic material, DNA, mRNA and tRNA, and encompasses animals (including humans), plants, fungi, archaea, bacteria, and viruses. A sequence of three nucleotides encodes an amino acid according to the GC in two steps: *transcription* - one strand of DNA is copied into a complementary mRNA (messenger) molecule, and *translation* - in which the language of nucleotides is transformed by ribosomes into the language of amino acids. Only certain limited regions of the genome — the genes — give information to make proteins. Human genes are few and far apart. There are about 12 genes per million bases of human DNA. Genes are divided into exons — sections of the coding sequence, interrupted by introns — non-coding spacers. Human genes have many small exons, some just 19 bases long, separated by introns of an average length of about 3,300 bases, but with a large dispersion. Most introns are only 87 bases long, but some are over 10,000. Statistical analyses of DNA sequences have shown that intra-gene (protein coding) regions are reach in C and G, while inter-gene (non-coding) regions are rich in T and A.

The GC shown in Table 1 [10] gives the mapping between the codons - triplets of nucleotides situated in the exons - and the amino acids. As there are four available bases, there result $4^3 = 64$ distinct codons, out of which 61 encode the 20 amino acids, while the remaining 3 codons encode the terminator that signals the end of an encoding region. Consequently, there is a degeneration of the GC, the amino acids being encoded by one (2 instances), two (9 instances), three (one instance), four (5 instances) or six (3 instances) different codons - triplets of nucleotides. The classic table or Cartesian representations of the GC does not include in their structure its characteristic symmetries and degeneration. We propose a tetrahedral representation of the nucleotides shown in Fig. 2. Each base defines a direction in the representation space, with the four corresponding base vectors symmetrically placed with respect to each other, oriented towards one of the corners of the tetrahedron. In the reference system shown in Fig. 2, the normalized base vectors are:

$$\begin{aligned} \vec{a} &= \qquad\qquad\qquad \vec{k}, \\ \vec{c} &= -\frac{2\sqrt{2}}{3}\vec{i} + \frac{\sqrt{6}}{3}\vec{j} - \frac{1}{3}\vec{k}, \\ \vec{g} &= -\frac{2\sqrt{2}}{3}\vec{i} - \frac{\sqrt{6}}{3}\vec{j} - \frac{1}{3}\vec{k}, \\ \vec{t} &= \frac{2\sqrt{2}}{3}\vec{i} \qquad\qquad - \frac{1}{3}\vec{k}. \end{aligned} \tag{1}$$

The procedure is repeated for each of the three bases in each codon [5], treating each of the three component bases as digits in a three digit number

written in base two: the vector corresponding to the third, *i.e.*, the last base in the codon (the least significant digit) is multiplied with 1, the vector corresponding to the second base in the codon with 2, and the vector corresponding to the first base of the codon (the most significant digit) with 2^2=4. For instance, the vector representing the codon ATG that encodes Methionine is $4\vec{a}+2\vec{t}+\vec{g}$. This results in the tetrahedral representation of the GC [6, 12, 13, 14] shown in Fig. 3. As a consequence, the first base in a codon selects one of the four *first order 16-codon tetrahedrons* that form together the *zero order tetrahedron* of the overall GC, the second base selects one of the *second order 4-codon tetrahedrons* that compose the chosen first order tetrahedron and, finally, the third base identifies one of the vertices. In this way, each of the 64 codons is attached to one of the vertices in the resulting three-level fractal-like tetrahedron structure. Taking into account the codon-to-amino acid mapping imposed by the GC, the amino acids encoded by the codons can also be assigned to one or several of the 64 vertices, according to their degeneration, as shown in Fig. 3. It turns out that the GC tetrahedron representation, as well as the mathematical descriptions based on it, reflect better the metric structure of the GC. Specifically, the codons that correspond to the same amino acid are mapped in neighboring points, so that related codons are clustered. It turns out that degeneration is basically restricted to the second order tetrahedrons and most pairs of interchangeable bases are distributed on the edges parallel to the pyrimidines and purines directions. The tetrahedron representation has also the advantage to naturally determine 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 of the resulting 3D image. To make the base and codon sequences easily readable for an observer, the three axes of the representation space can be assigned to the three basic color components of the RGB - red, green, blue system [11]. Consequently, each point in the representation space – each base in the case of Fig. 2 – corresponds to a distinct, readily identifiable, hue. This approach is useful for the fast exploration of DNA sequences at the nucleotide level and can be readily extended at the codon (Fig. 6) and amino acid levels. The mathematical description of the code can be simplified by

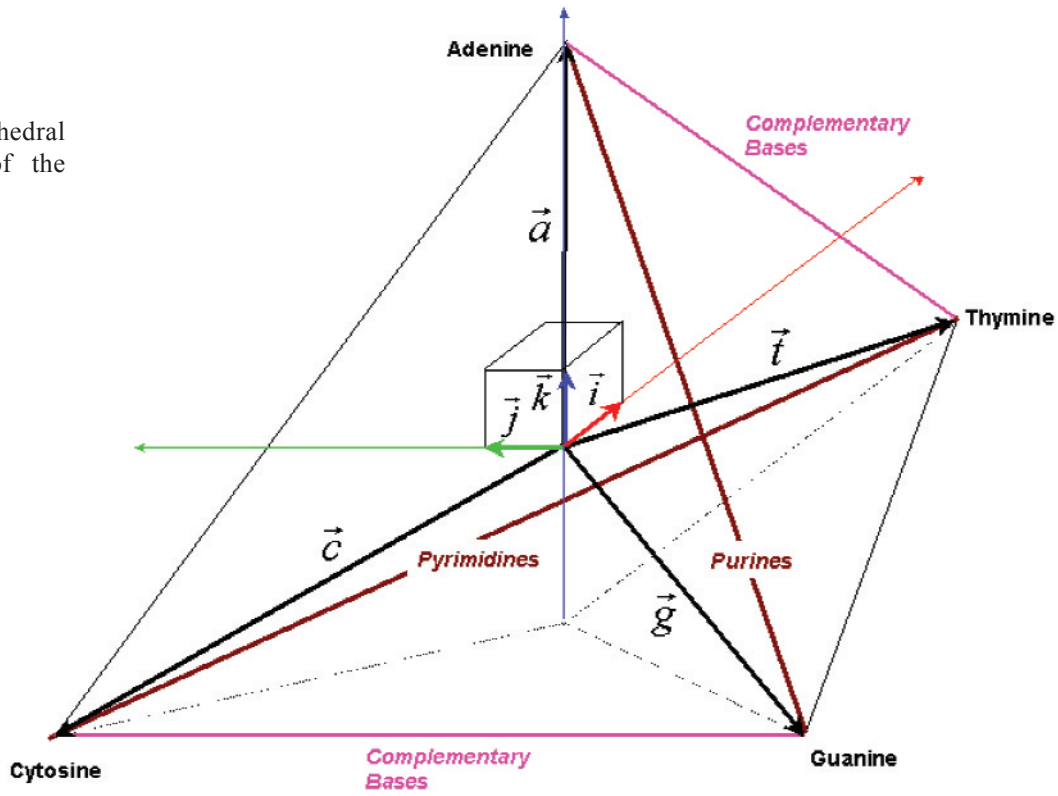


Fig. 2 Tetrahedral representation of the nucleotides.

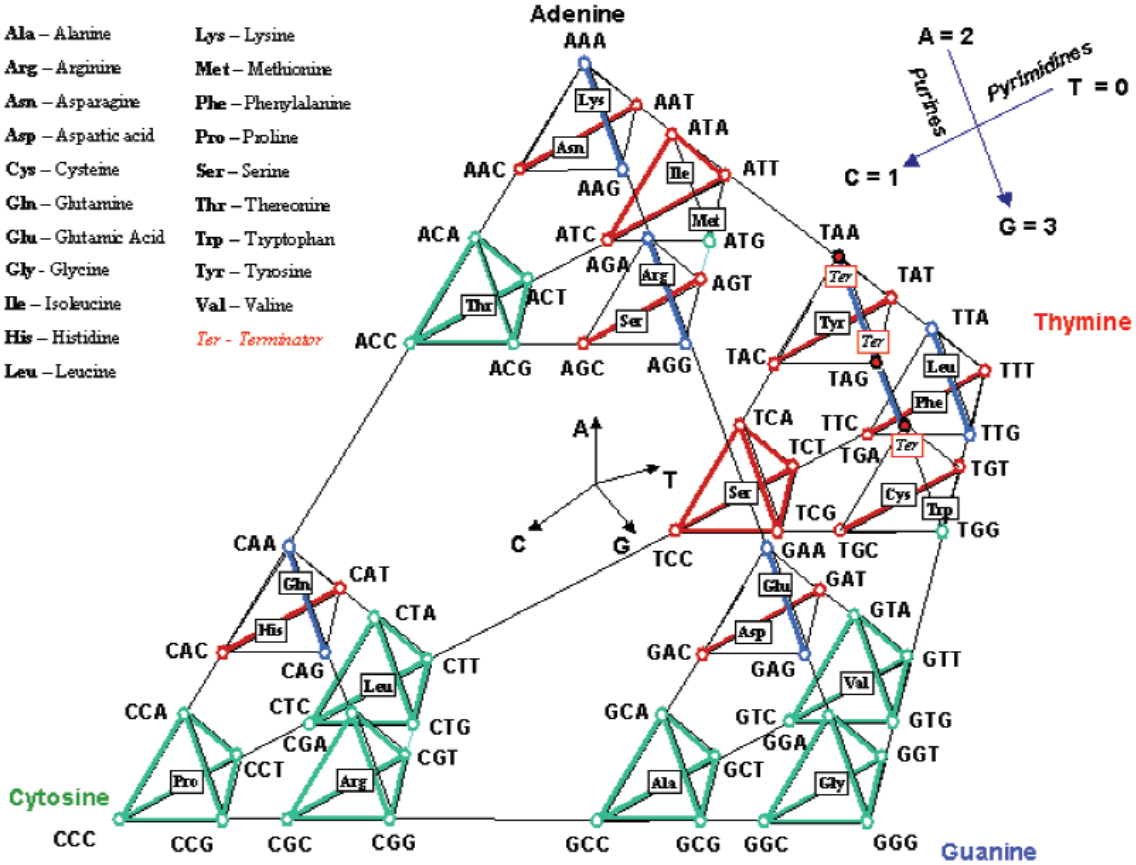


Fig. 3 Tetrahedral representation of the Genetic Code.

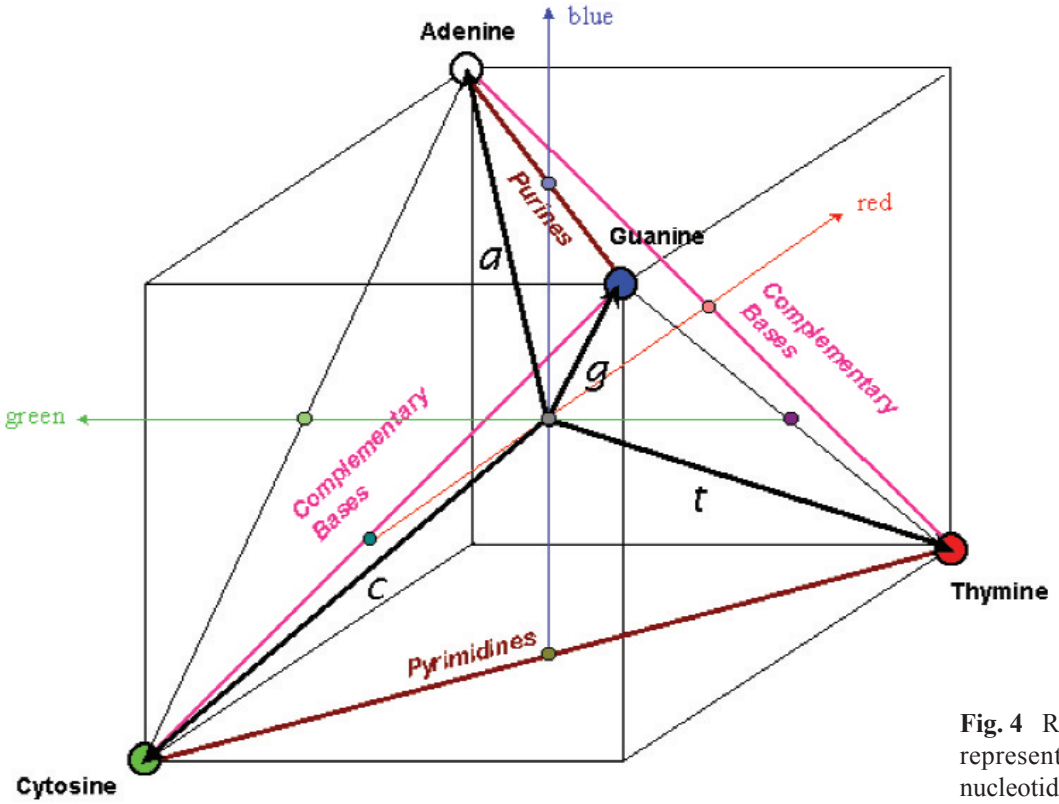


Fig. 4 Rotated tetrahedral representation of the nucleotides.

rotating the reference system as shown in Fig. 4, in which it is emphasized that the vertices of a regular tetrahedron are also a subset of the vertices of a cube. It is also advantageous to give up the Euclidian normalization condition and to choose integer ±1 coordinates for the vertices of this cube, including the points representing the bases, so that the base vectors in (1) take the simpler form:

$$\begin{aligned} \vec{a} &= \vec{i} + \vec{j} + \vec{k} \\ \vec{c} &= -\vec{i} + \vec{j} - \vec{k}, \\ \vec{g} &= -\vec{i} - \vec{j} + \vec{k}, \\ \vec{t} &= \vec{i} - \vec{j} - \vec{k}. \end{aligned} \qquad (2)$$

The dimensionality of the representation can be reduced to two by projecting the basic tetrahedron on a plane. Such planes can be chosen in various ways that conserve the symmetry of the representation and reflect biological properties in corresponding mathematical properties. For instance, the planes can be defined by a pair of the coordinate axes. On the other hand, these planes can be put in correspondence with a complex plane, so that a *complex representation* of the bases is obtained. Choosing the plane *red-blue*, the four bases are placed in a quadrantal symmetry as shown in Fig. 5a and the complex representation of the bases is given by:

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

The complex representation has the advantage of better translating some of the features of the bases into mathematical properties. For instance, in the representation of Fig. 5a, the complementarity of the pairs of bases A-T and G-C, respectively, is expressed by the fact that their representations are complex conjugates, while purines and pyrimidines have the equal imaginary parts and real parts of opposite sign. Relations (3) are equivalent to the use of two mutually orthogonal complex (bipolar)

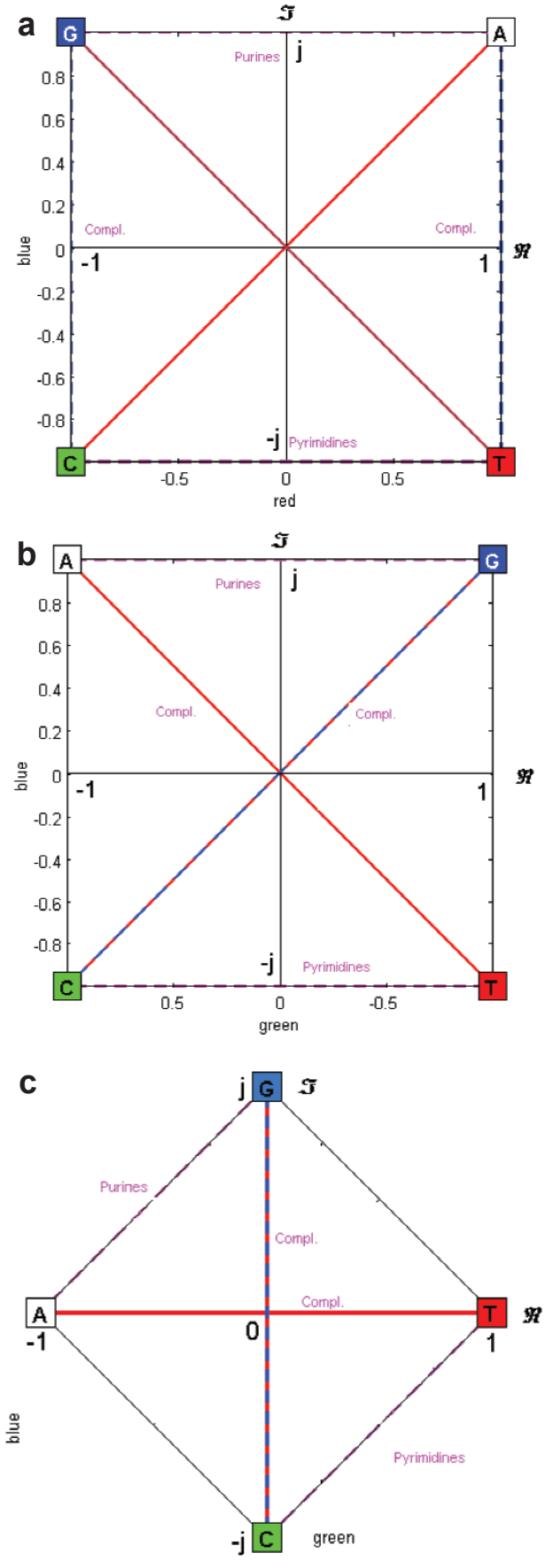


Fig. 5 Projections of the tetrahedral representation of the nucleotides in Fig. 4, on the planes:
(a) red-blue (point-of-view: *az* = 0, *el* = 0, *roll* = 0; *p* = 1, *q* = 2);
(b) green - blue (*az* = -90, *el* = 0, *roll* = 0; *p* = 2, *q* = 2);
(c) green - blue, rotated around the red axis and contracted (*az* = -90, *el* = 0, *roll* = 45; *p* = 1, *q* = 1).

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 attaching odd real and odd imaginary components to the codons. The division by two of the values attached to the bases would make the distance between neighboring numbers equal to one, but also would have the disadvantage of operating with non-integer components.

Permuting the bases A-G, *i.e.*, choosing the blue-green projection plane, the representation shown in Fig. 5b is obtained, for which:

$$\begin{aligned} a &= -1 + j, \\ c &= -1 - j, \\ g &= 1 + j, \\ t &= 1 - j. \end{aligned} \qquad (4)$$

This representation has the advantage that the two complementary strands of a DNA molecule correspond to digital signals of equal absolute values, but opposite signs, so that their sum is always zero.

Rotating the representation in Fig. 5b with 45 degrees counterclockwise, and by contracting the absolute values with $1/\sqrt{2}$, the representation in Fig. 5c is obtained, which conserves the zero sum of complementary strands of DNA and has the advantage of using pure real and pure imaginary complex representations of the bases, with unit absolute value:

$$\begin{aligned} a &= -1, \\ c &= -j, \\ g &= j, \\ t &= 1. \end{aligned} \qquad (5)$$

This representation corresponds to converting base sequences into digital signals expressed in two orthogonal binary systems, one real (for A-T), the

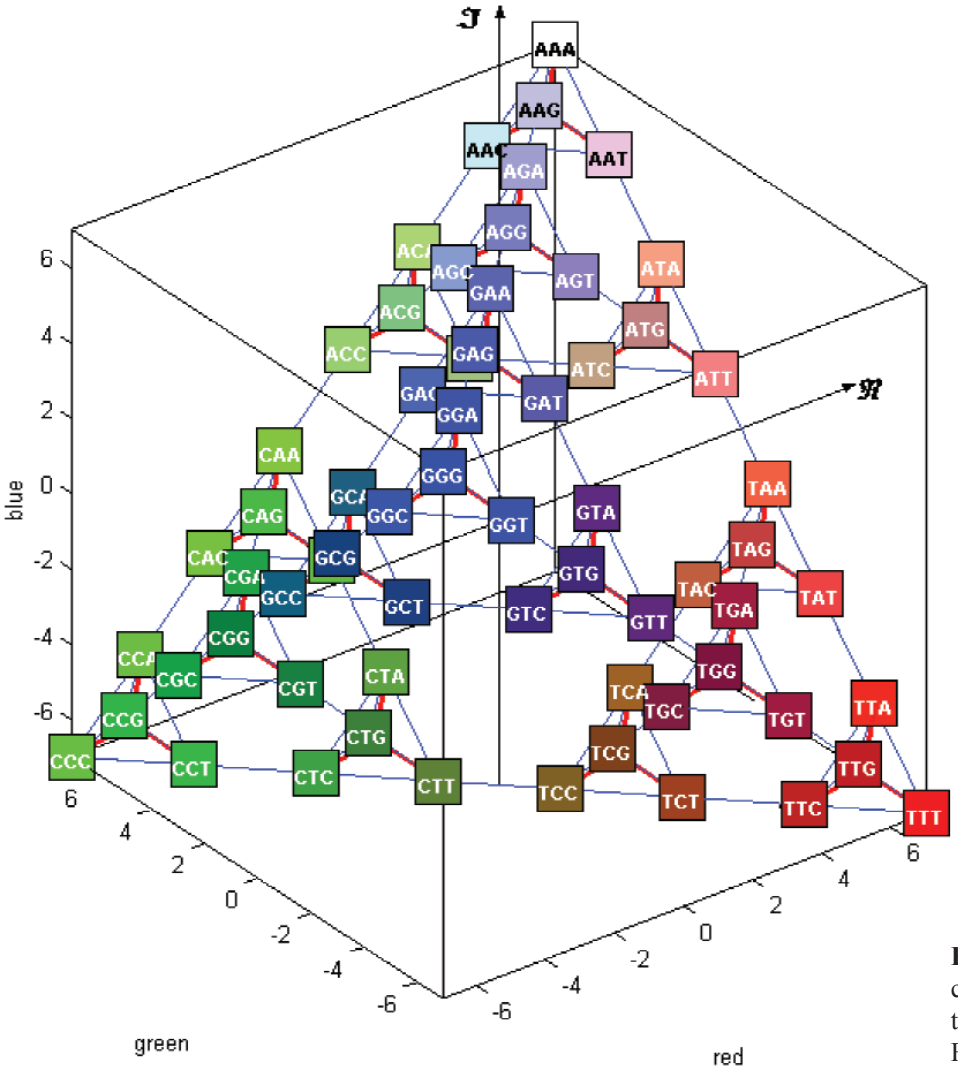


Fig. 6 Color coding of the codon tetrahedral representation in Fig. 3, rotated as in Fig. 4.

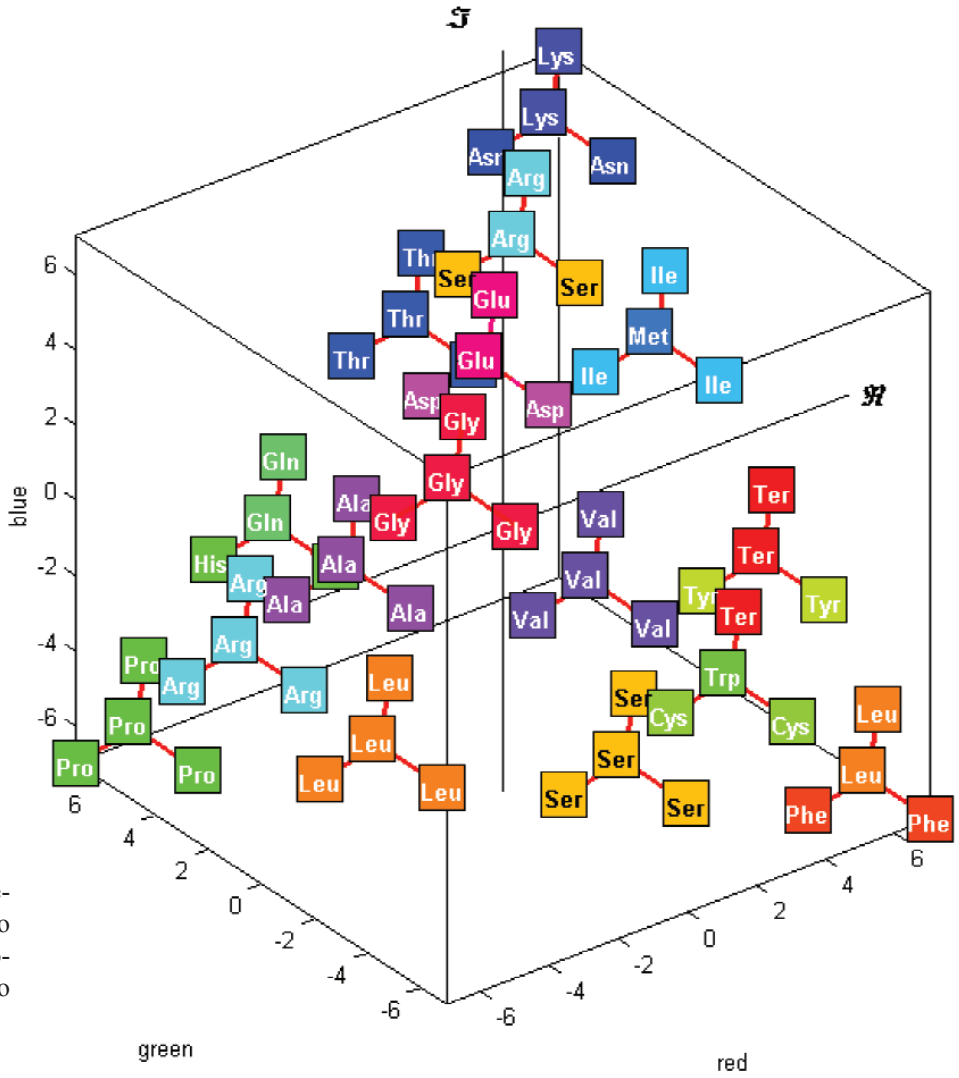


Fig. 7 Tetrahedral representation of the amino acids mapped on the codons in Fig. 6 according to the Genetic Code.

other imaginary (for C-G), each using a bipolar {-1, 1} representation, as opposed to the unipolar {0, 1} representation used in the standard binary system.

Figure 6 shows the tetrahedral representation of the 64 codons in Fig. 3, after using the rotated reference system defined in Fig. 4 and the color convention described above. Correspondingly, Fig. 7 shows the tetrahedral representation of the amino acids mapped on the codons according to the GC. Again, the dimensionality can be reduced by projecting the codon and the amino acid tetrahedrons on adequately chosen planes. The projections of the two tetrahedrons on the *red-blue* plane are shown in Fig. 8a and b, while Fig. 9a and b show the projections on the blue-green plane rotated and contracted as in Fig. 5c. Each of these couples of figures (8a,b and 9a,b) gives a version of the complex representation of the GC. It should be noticed that the complex mapping clusters the amino acids which are agglomerated in contiguous regions of the complex plane, with the exceptions of the three amino acids that have a degeneration of order six. Complex values can be attached in various ways to the amino acids. One modality is to assign to a certain amino acid the average value over the whole area onto which it is mapped. It is possible to compute the average, taking also into account the relative frequencies of occurrence of the different codons that correspond to that amino acid. On the other hand, for specific problems, the assigning of the complex values to the bases and to the amino acids can be adapted for the task. For instance, the optimum values for detecting the exons are different for the optimum values for detecting the reading frames [11]. The choice of the values assigned to nucleotides and/or to amino

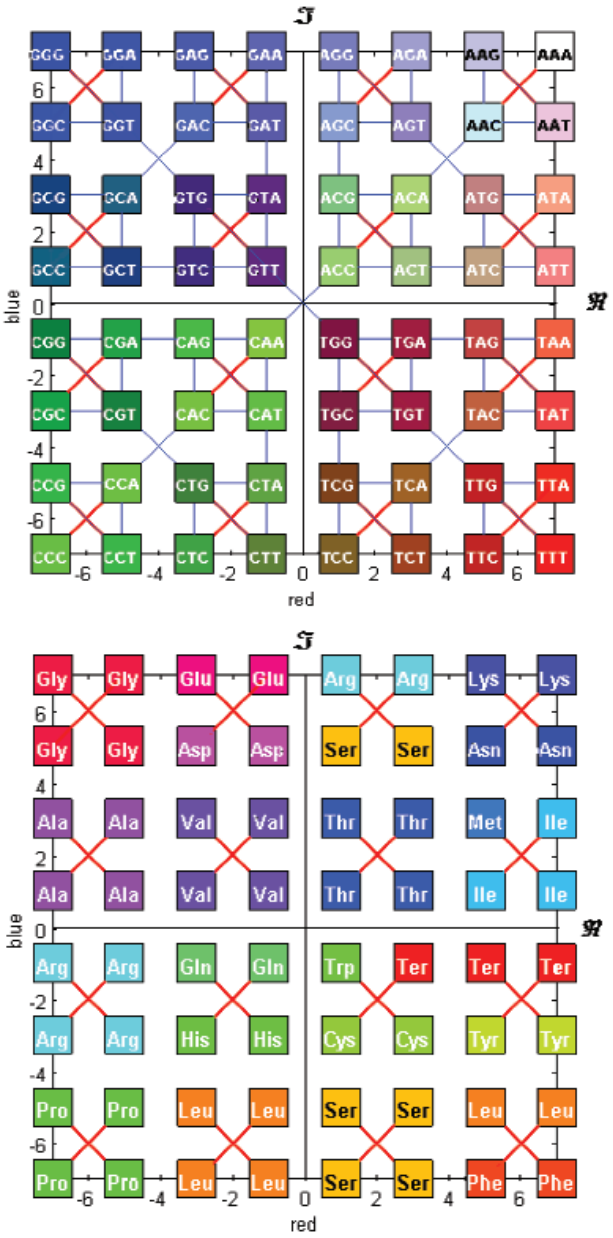


Fig. 8 Projections on the red-blue (complex) plane (point-of-view: *az* = 0, *el* = 0, *roll* = 0; *p* = 1, *q* = 2) of the tetrahedral representation of:
(a) the codons in Fig. 6;
(b) the corresponding aminoacids in Fig. 7.

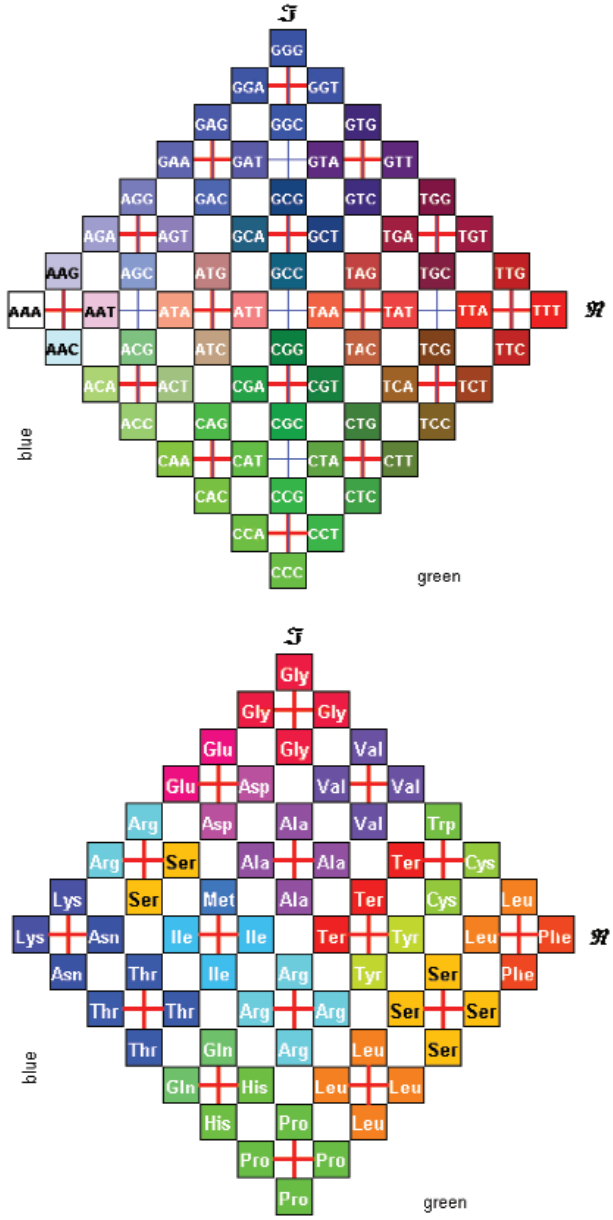


Fig. 9 Projections on the blue-green (complex) plane rotated around the red axis (point-of-view: *az* = -90, *el* = 0, *roll* = 45; *p* = 1, *q* = 1) of the tetrahedral representation of (a) the codons in Fig. 6, (b) the corresponding aminoacids in Fig. 7.

acids alows fitting of the complex representation for the application at hand.

It is also possible to further reduce the dimensionality of the representation of base, codon and amino acid sequences by using a real one-dimensional mapping. The digits {0, 1, 2, 3} are attached to the four nucleotides. 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, , 63}. Actually, a whole DNA sequence can be seen as a huge number written in base four. Nevertheless, it corresponds better 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 2 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 [6,12,13, 14].

Table 2. Mapping of nucleotides to digits in base four.

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

Table 3 gives the mapping of the digital codons to the numerical codes of the amino acids. The numerical codes assigned to the amino acids result from the order of their first reference when gradually incresing 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 dependence has only four reversals of the normal ascending order: for a terminator

Table 3. Optimal mapping of real integer codons to aminoacids.

Digital real codon	Amino acid code	Long name	Short name	Symbol
0,1	1	Phenylalanine	Phe	[F]
2,3,16,17,18,19	2	Leucine	Leu	[L]
4,5,6,7,44,45	3	Serine	Ser	[S]
8,9	4	Tyrosine	Tyr	[Y]
10,11,14	0	Terminator	Ter	[end]
12,13	5	Cysteine	Cys	[C]
15	6	Tryptophan	Trp	[W]
20,21,22,23	7	Proline	Pro	[P]
24,25	8	Histidine	His	[H]
26,27	9	Glutamine	Gln	[Q]
28,29,30,31,46,47	10	Arginine	Arg	[R]
32,33,34	11	Isoleucine	Ile	[I]
35	12	Methionine	Met	[M]
36,37,38,39	13	Thereonine	Thr	[T]
40,41	14	Asparagine	Asn	[N]
42,43	15	Lysine	Lys	[K]
48,49,50,51	16	Valine	Val	[V]
52,53,54,55	17	Alanine	Ala	[A]
56,57	18	Aspartic acid	Asp	[D]
58,59	19	Glutamic acid	Glu	[E]
60,61,62,63	20	Glycine	Gly	[G]

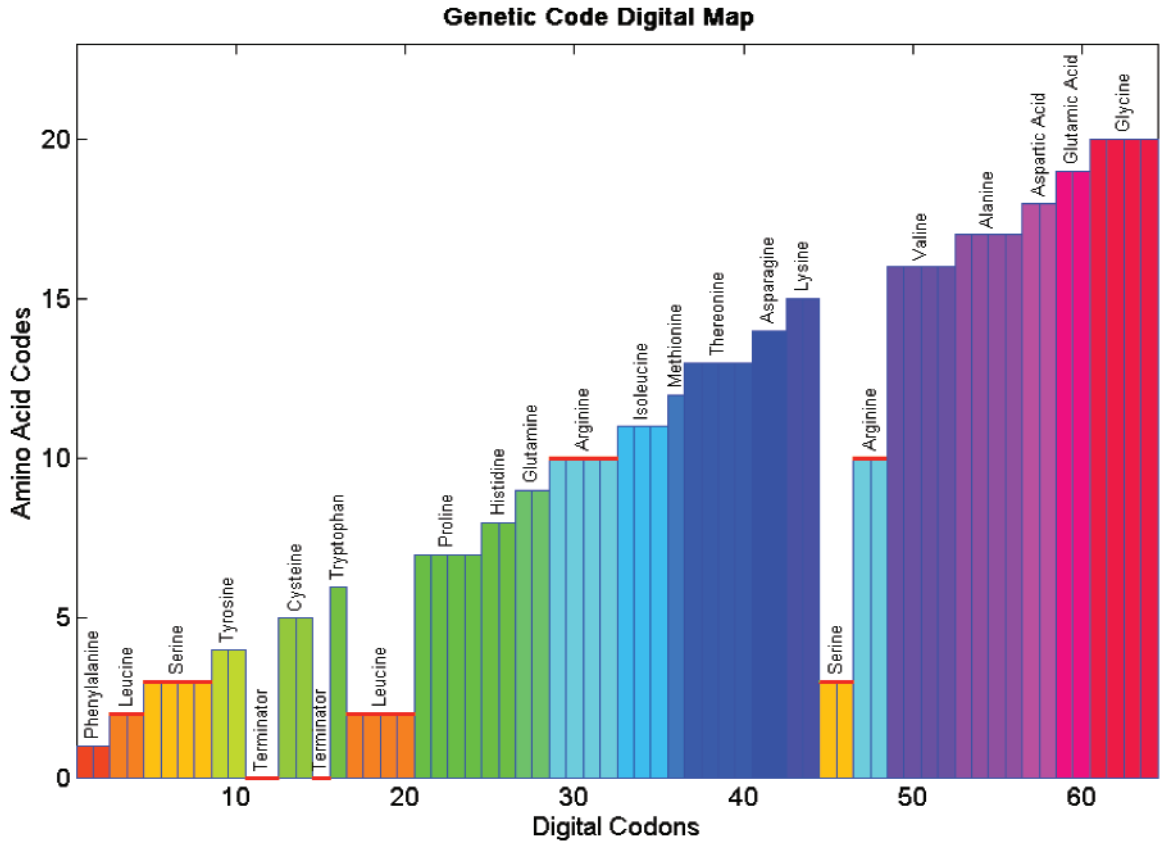


Fig. 10 Optimal mapping of real codons (0 - 63) onto amino acid and terminator real codes (0 - 20), for permutation p = 1. Breakings of the monotonicity occur only for the three amino acids having degenerescency six.

sequence and for the three instances of sextuple degeneration: leucine, serine and arginine. An exhaustive search for all the 24 possible correspondences of the nucleotides 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, as shown in Fig. 10.

The reference to the various real and complex representations of the nucleotides can be simplified by using the pair of indices (p, q) given in Tables 4 and 5. The index p specifies the *base permutations* and takes values from 1 to 24 as given in the table. The index q gives the *representation type* and has the values $q = 0$ for the real representation, $q = 1$ – the representation defined by the mapping of the nucleotides to pure real / pure imaginary numbers, as defined by equation (5) and Fig. 5c (for $p = 1$), and $q = 2$ for the mapping of nucleotides to quadrantally symmetric complex numbers, as defined by equation (3) and Fig, 5a (for $p = 1$).

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 reference eukariotic 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 [3]. In the following, we have chosen a rather different approach: the study of genomic signals mainly at scales of 10^4 - 10^6 bp, to detect general trends of the inter-gene signals, potentially significant in revealing their basic properties and to search for genomic signals with possible control

Table 4. Base order permutations (q = 0).

p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
T	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3
C	1	1	2	2	3	3	0	0	2	2	3	3	0	0	1	1	3	3	0	0	1	1	2	2
A	2	3	1	3	1	2	2	3	0	3	0	2	1	3	0	3	0	1	1	2	0	2	0	1
G	3	2	3	1	2	1	3	2	3	0	2	0	3	1	3	0	1	0	2	1	2	0	1	0

functions. Nucleotide and codon signal analysis are presented below. The amino acid signals are truly useful only for the exon areas.

Nucleotide signal analysis

The human chromosome 11, downloaded from [4] has been arbitrarily divided into segments of 985014 bases each, to search for large range scale feature of the corresponding genomic signals. Albeit, the resolution of the methods and of the software tools that are used, as well as the scope of the study, are not limited to these examples.

Figure 11 shows the *phase*, the *aggregated phase* and the *unwrapped phase* of the nucleotide signals along the segments 1 (a) and 55 (b) of chromosome 11. 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 confines by definition the (principal restriction of the) phase of a complex number to the domain (– π, π] that covers only once all the possible orientations of the associated vector (phasor) in the complex plane. For the sequences analyzed in figures 11 a and b, the complex representation (p = 1, q = 2) defined in Fig. 5a and in equation (3) has been used, for which the *phases* of the base representations can have only the values {–3π/4, –π/4, π/4, 3π/4} radians. The graphs of the phase look just like thicker lines at the scale of figures 11 a and b. A sample of the detailed variation of the phase along a DNA base sequence is shown in Fig. 11 c, which gives the phase for only 201 bases (500 – 700) at the beginning of the segment 1 of chromosome 11.

The *aggregated phase* is the sum of the phases of all the complex base representations starting from the beginning of a segment. As the bases are distributed along the segments with unequal and variable probabilities, the aggregated phase is not zero and drifts between positive and negative values, its 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 (– π, π], which favors π over –π, the aggregated phase can be distorted when using complex representations that include real negative numbers. *E.g.*, this is the case for the representation (p = 1, q = 1), presented in Fig. 5c and in equation (5), for which the *phase* of the base complex representations can have the values {–π/2, 0, π/2, π} radians. This unwanted effect has been avoided by adding uniformly distributed small random complex numbers to each of the complex base representations in the analyzed sequence. This way, phases close to –π are equally probable as the phases close to π and the artificial drift of the aggregated phase towards positive values is eliminated.

Table 5. Types of nucleotide representation (base mappings for p = 1).

	q = 0	q = 1	q =2
T	0	1	1 - j
C	1	-j	-1 - j
A	2	-1	1+ j
G	3	j	-1+ j

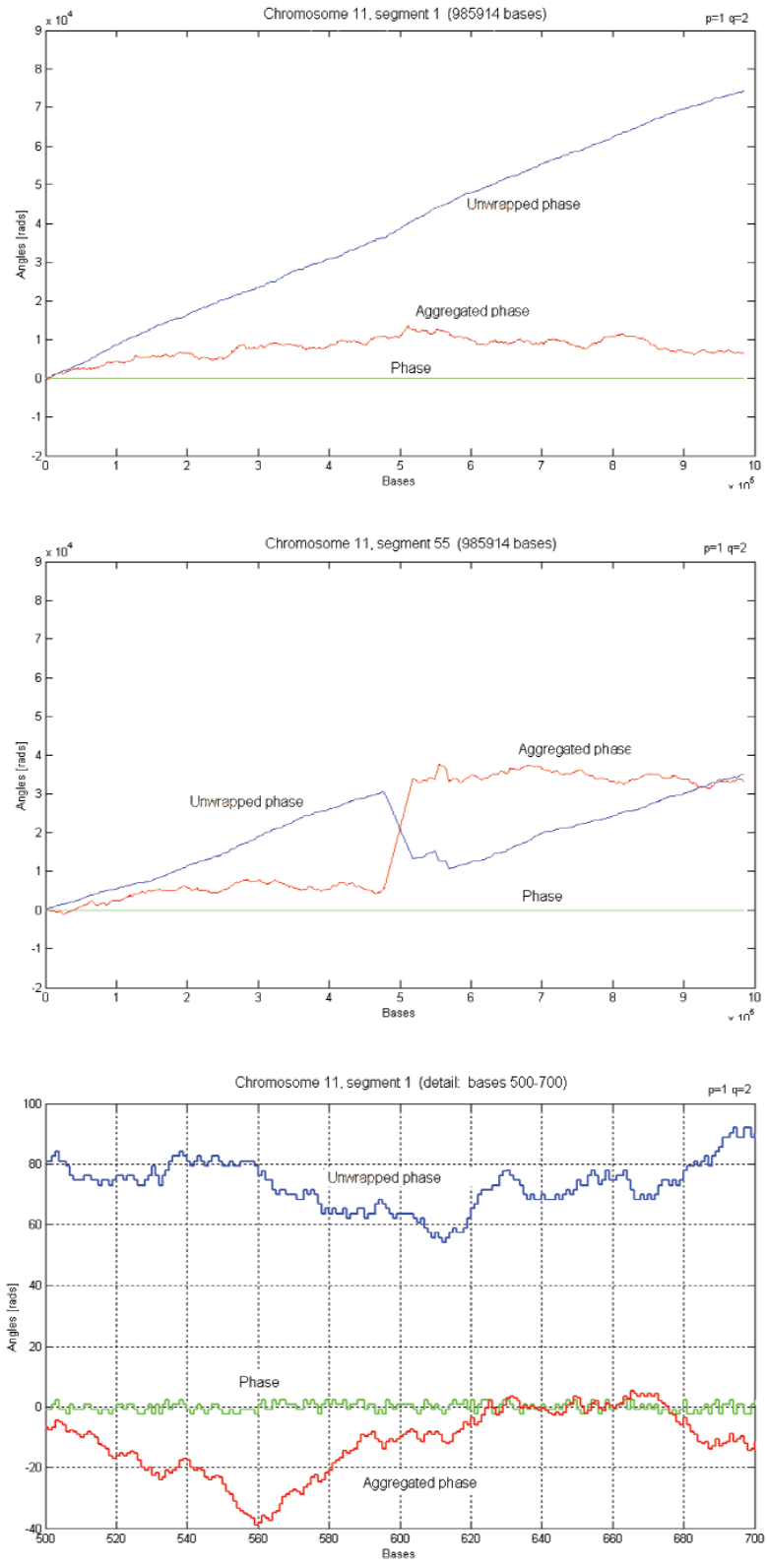


Fig. 11 Phase (green), aggregated phase (red) and unwrapped phase (blue) of nucleotide complex codes along two segments of chromosome 11, for the representation in Fig. 5a (p = 1, q = 2). Segments 1 (a) and 55 (b) comprise each 985014 bases. (c) Details at the beginning of segment 1 (bases 500 - 700).

The *unwrapped phase* is the corrected phase of a sequence of complex numbers, in which the absolute value of the difference between the phase of a new element in the sequence and the phase of the preceding element is kept smaller than π by adding or subtracting an appropriate multiple of 2π to or from the phase of the new element. The unwrapped phase eliminates the phase jumps introduced by the conventional restriction of the phase domain described above and allows observing the true global phase trends along a sequence. The steady and almost linear increase of the unwrapped phase of the nucleotide complex representations along the segment 1 of chromosome 11 in Fig 11a shows an average increase of the phase of about 0.075 rad/bp = 4.3 degree/bp. This behavior is typical for the vast majority of the 80 investigated segments of chr.11, each 985014 bases long, and reveals that the complex representations of the nucleotides in chr.11 form a *counterclockwise helix* with a spinning varying between 0.047 rad/bp = 2.7 degree/bp (for segment 67) and 0.120 rad/bp = 6.9 degree/bp (for the first half of segment 25). A striking exception from this behavior can be found close to the middle of the segment 55 of chr. 11 and is shown in Fig. 11 b. For a total of about 50 kbp, the spinning of the complex base representation helix is oriented clockwise, and the unwrapped phase decreases sharply. For the rest of segment 55, the slope of the unwrapped phase is of about 0.06 rad/bp, similar with the neighboring segments. It is remarkable that the helicoidal wrapping of the complex representations of the

bases is a long range trend along almost all of chromosome 11, and is most probable common to all chromosomes. The trend is maintained over distances of tens of millions of bases and reveals a statistical regularity of the *succession* of bases, not merely of the *distribution* of bases. This contradicts the current oversimplified genomic model that considers the inter-gene areas as domains of randomness, the so called "junk DNA" and recognizes only the meaningful structure of the exons [5, 9].

To check the validity of this large scale feature of the nucleotide sequential structure and to rule out possible artifacts, this result has been verified in several ways. First, as explained above, some low power noise has been added to each sample of the complex representation sequence to eliminate any possible bias related to the conventional restriction of the phase domain that could affect crisp data. Second, other complex representations have been used, changing the orientation and the value of the complex representations of the bases, but keeping unchanged their relative positions and angles. *E.g.*, Fig. 12 a and b show the phase, aggregated phase and unwrapped phase of the nucleotide complex codes for the same segments of chromosome 11 as shown in figures 11a and 11b, but for the representation ($p = 2, q = 1$). This results in a $\pi/4$ radian counterclockwise rotation and a contraction with $\sqrt{2}$ of the representation ($p = 1, q = 2$). As expected, the aggregated phase largely changes, while the unwrapped phase, which is determined by the order and relative angles of the complex numbers in the sequence and not by their individual phases, remains the same. The possible vertical translation of about $\pi/4$ radians resulting from the transformation can not be observed at the scale of the figure. In the third place, an artificial uniform random sequence, comprising the same number of 985014 nucleotides, has been generated and compared with the segments of chr. 11. Again as expected, the phase, aggregated phase and unwrapped phase of the complex codes remain at much lower absolute values and show no systematic variation along the sequence as can be seen in Fig. 13. Finally, it has also been checked directly, using synthesized sequences, that the trend of variation of the unwrapped phase is produced by

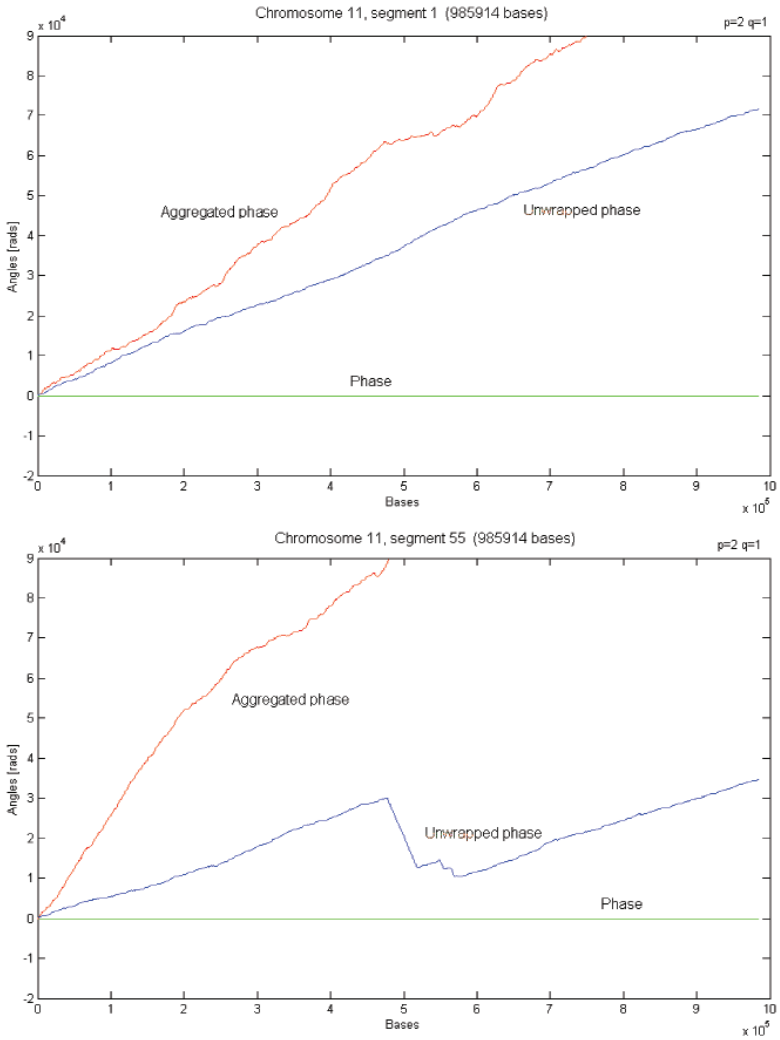


Fig. 12 Phase (green), aggregated phase (red) and unwrapped phase (blue) of the nucleotide complex codes for the same segments of chromosome 11 as shown in Fig. 11a and 11b, but for the representation p = 2, q = 1. This results in a 45 degree counterclockwise rotation and a contraction with of the representation p = 1, q = 2, which changes the aggregated phase. As the relative orientations of the nucleotide codes are the same in the two representations, the unwrapped phase remains essentially unchanged.

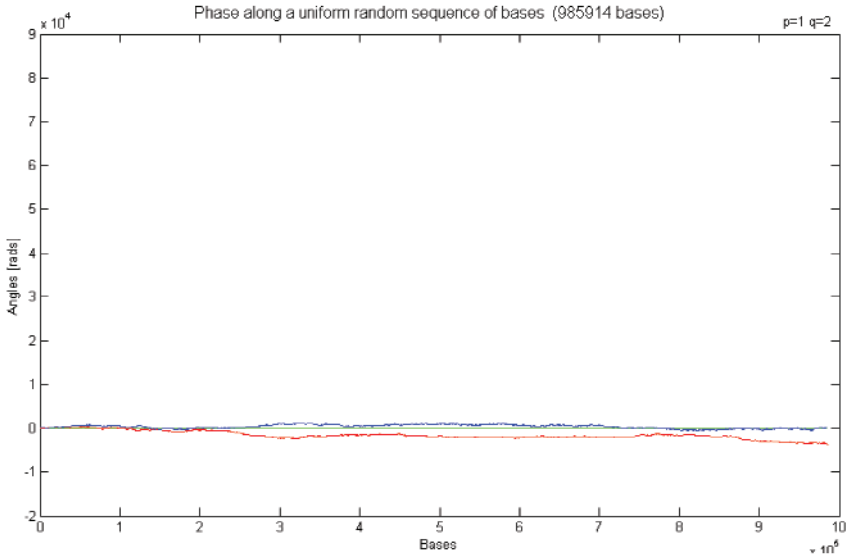
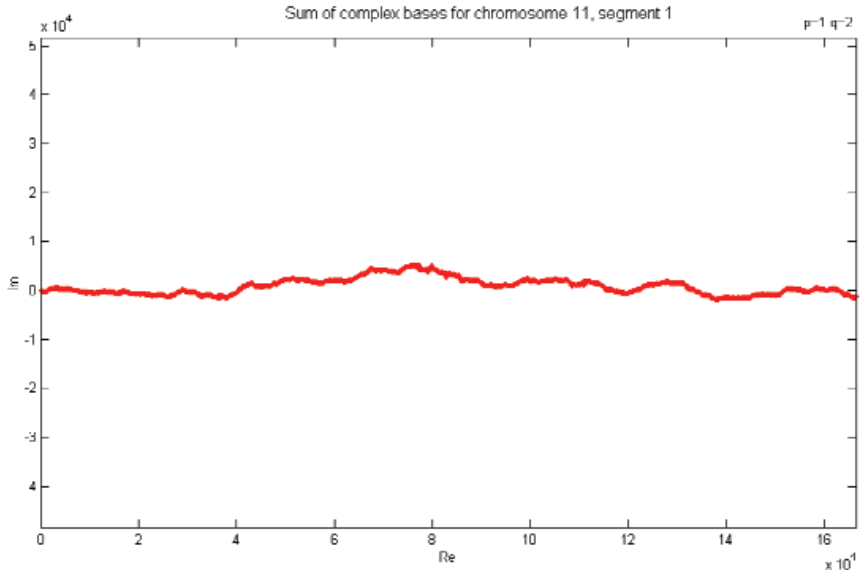


Fig. 13 Phase (green), aggregated phase (red) and unwrapped phase (blue) of the nucleotide complex codes of a uniform random synthetic sequence comprising 985014 nucleotides. The representation p = 1, q = 2 is used as in Fig. 11.



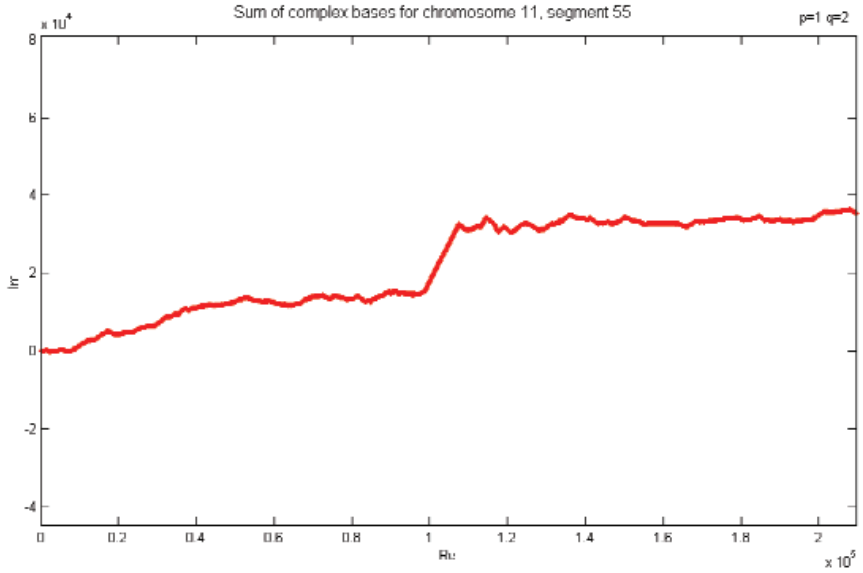


Fig. 14 The *base sequence path*, *i.e.*, the sum of the base complex codes for the segments 1 (a) and 55 (b) of chr.11. Representation $p = 1$, $q = 2$ has been used like in Fig. 11a and b. For all the 80 segments of chr.11 that have been analyzed, each 985014 bases long, the sum of the complex codes displays a marked tendency to progress preferentially from the origin (at left) along the real axis (to the right). Approximately at the middle of the sequence, segment 55 shows a striking exception from this predominant trend that correlates well with the reversal of the unwrapped phase slope in Fig. 11b.

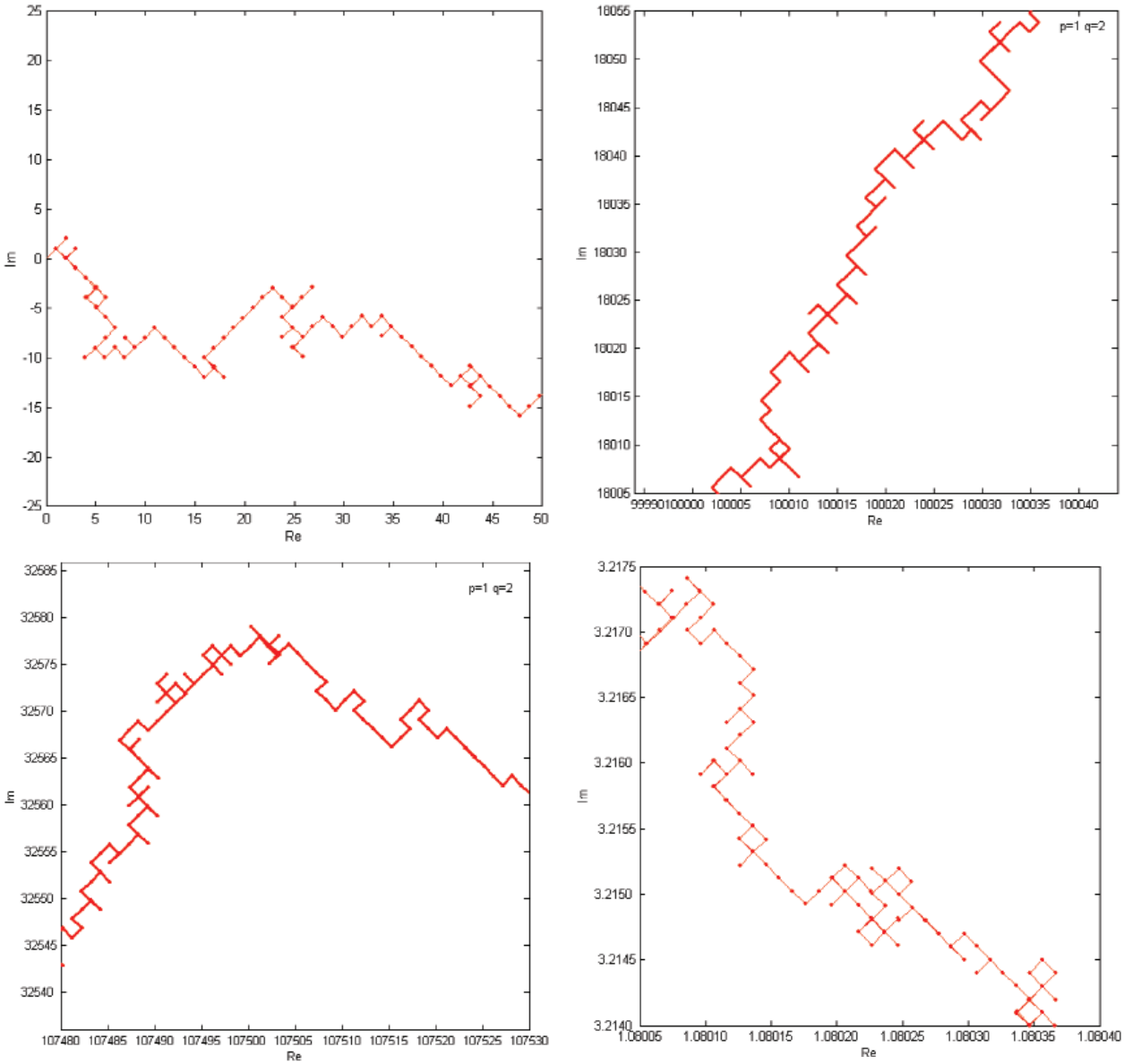


Fig. 15 Details of the base sequence path in the complex plane for the segment 55 of chr.11 shown in Fig. 14b. The four segments, each about 70 nucleotides long, correspond to the (a) initial almost horizontal branch, (b) the rising ramp near the middle of the curve in Fig 14b, (c) the peak at the end of this region, (d) the following descending ramp. The differences in the relative frequency of the four bases determine the shape of each segment. Surprisingly enough, for most of chr11, the trend is maintained over distances of hundreds of thousands of bases.

the predominant succession of the complex elements in the sequence, and disappears when the order of the elements is randomized without changing the number of each of the elements.

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 sum of the elements along the analyzed sequence. For all the 80 segments of the human chromosome 11 that have been analyzed, each of length 985014, the resulting *base sequence path* for the representation ($p = 1$, $q = 2$) 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 probability of the pair A-T over the pair C-G in the nucleotide sequences, but also from the relative balance of the probabilities of purines

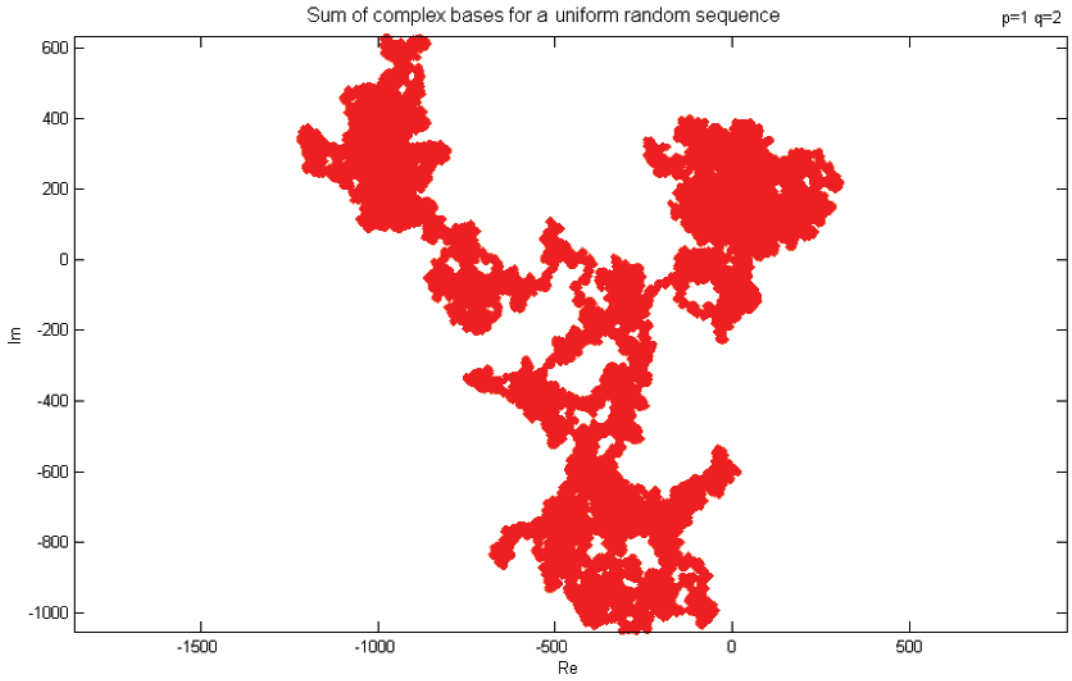


Fig. 16 The base sequence path (sum of the complex codes) of a uniformly distributed random sequence of nucleotides. Note the large difference of scale in comparison to Fig. 14.

and of pyrimidines in each of these pairs. As already mentioned, the A, T bases are more frequent in the inter-gene areas of the genome, while the C, G bases are more frequent in the intra-gene areas, the exons. Fig. 14 a and b show the base sequence path for the segments 1 (a) and 55 (b) of chr.11. As for the aggregated phase and the unwrapped phase shown in Fig. 11 b, approximately at the middle of segment 55, there is an exception from the predominant trend, as the *sequence path* has also a

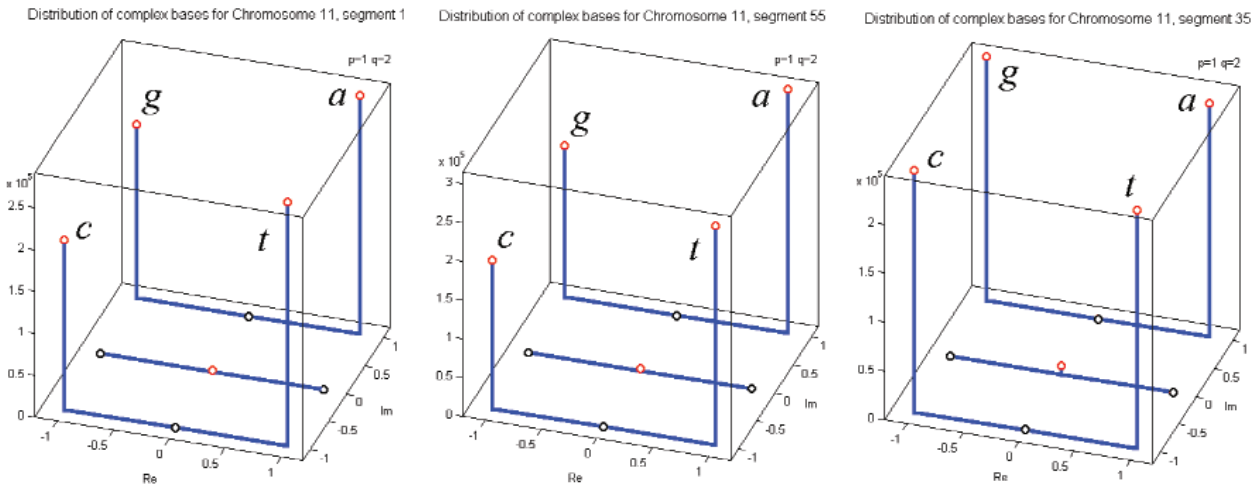


Fig. 17 Stem representation of the total number of the four nucleotides in chromosome 11, segments 1 (a), 55 (b) and 35 (c). In both segments 1 and 55 the number of a-t bases exceeds the number of c-g bases, while for segment 35 the distribution is reversed. Local monitoring of base relative frequencies allows identifying the regions with larger probability to contain exons as the segments with higher c-g density. The small stem in the origin of the complex plane corresponds to not yet identified bases in the sequences.

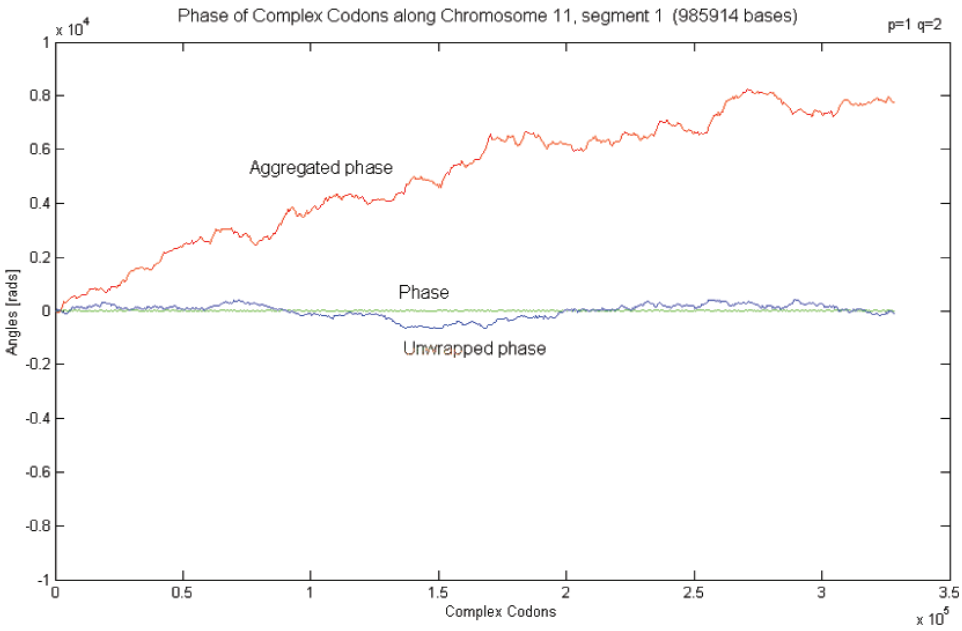


Fig. 18 Phase (green), aggregated phase (red) and unwrapped phase (blue) of complex codons along a DNA sequence. For illustration purposes, the same segment 1 of chromosome 11 in Fig. 11a, and the same representation p = 1, q = 2 are shown.

marked drift along the positive imaginary axis for the same area of about 50 kbp mentioned earlier. It is also obvious that a certain gradual departure from the real axis also occurs in the first half of the segment 55. This shows a higher probability of purines over pyrimidines in these areas. 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. Figure 15 gives details of the base sequence path for the segment 55 of chr.11 shown in Fig. 14b. The four segments, each about 70 nucleotides long, correspond to the (a) initial almost horizontal branch spanning the real axis over the range 1-50, (b) the rising quasi-linear ramp near the middle of the curve in Fig 14b (100000-100040), (c) the peak at the end of this region (107480-107530), (d) the descending quasi-linear ramp (108005-108037). The differences in the relative frequency of the four bases determine the shape of each segment. Surprisingly enough, for most of chr11, the trend is maintained over distances of millions of bases. By comparison, the sequence path of a uniformly distributed random sequence of bases shown in Fig. 16 remains confined in a quite small area around the origin of the complex plane.

Another efficient tool for investigating the relative frequencies of the bases in a certain segment of DNA is the *nucleotide stem representation* used in Fig. 17, again for large scale segments of chr.11. The total number of the four nucleotides in Homo sapiens chromosome 11, segments 1 (a), 55 (b) and 35 (c) are given. In segments 1 and 55 the number of A-T bases exceeds the number of C-G bases, while for segment 35 the distribution is reversed. Local monitoring of base relative frequencies allows identifying the regions with larger probability to contain exons, as the segments with higher C-G density. The small stem in the origin of the complex plane corresponds to not yet identified bases in the publicly available genomic sequences [3, 4] for which the system assigns the conventional default code 0 when a complex representation is used.

Codon signal analysis

DNA sequences converted into real or complex codon sequences by using the various representation presented above have also been investigated. However, from the observed general features of such codon sequences, it results that the codon structure is primarily significant for the exons, *i.e.*, for the intra-gene regions of the genome that directly encode proteins, including the areas that are merely "quotation" of various gene segments scattered along most genomes, including

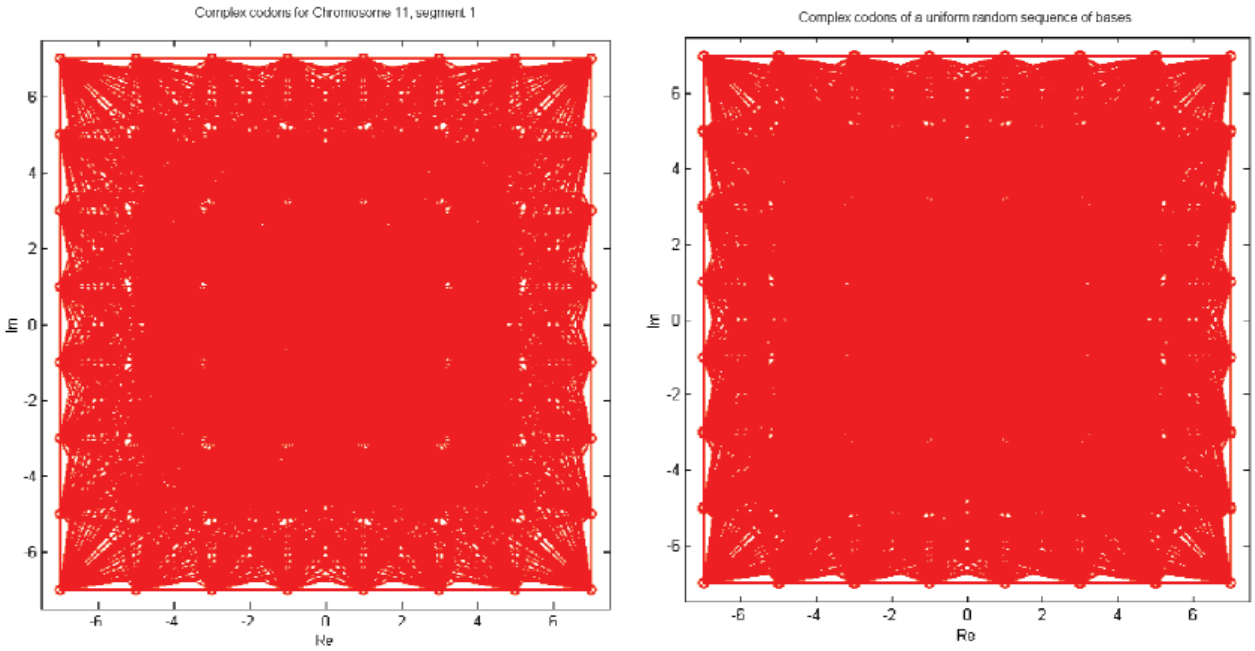


Fig. 19 Transitions between the complex codons along (a) the segment 1 of chromosome 11 (see Figs. 11a, 12, 14a, 17a, and 18), (b) a uniform random sequence of bases (see Figs. 13 and 16).

the homo sapiens genome, and less for the inter-gene regions where other higher order encoding of genomic signals is to be expected. For this reason, the triplets of bases outside the exons and in the absence of a clearly defined reading frame will also be called *quasi-codons*.

Fig. 18 shows the *phase*, *aggregated phase* and *unwrapped phase* of a complex codon sequence. To allow comparison between codon and nucleotide sequences, the same segment 1 of chromosome 11 and the same representation ($p = 1$, $q = 2$) as in Fig. 11a have been used. First, it is to be noticed the conspicuous change of the variation of the *unwrapped phase* along the segment under consideration for the codon sequence with respect to that for the corresponding nucleotide sequence. The steady and almost linear increase of the unwrapped phase that reveals a counterclockwise helix in the case of the nucleotide sequence in Fig. 11a is replaced by a slow quasi-random drift that shows the lack of such a structure for the codon sequence in Fig.18. There is also a significant change of the *aggregated phase* that differs from the one that would correspond to a shift-invariant distribution of the bases along the DNA sequences. This result shows that the distribution of each of the four bases is different for the three possible positions in the codons, *i.e.*, the distribution depend on the reading frames. The detailed analysis of these results are not the object of this paper.

A qualitative *transition diagram* between the quasi-codons to along the segment 1 of chromosome 11 is shown in Fig. 19a. It can be noticed that all the possible transitions do occur for the 328338 quasi-codons of the considered segment. The qualitative inter-codon transition diagrams are quite similar for all the 80 segments of chr.11 that have been studied, and also similar to the transition diagram for the uniformly distributed random sequence of bases having the same length shown in Fig. 19b.

Fig. 20 shows the *codon sequence path* in the complex plan, *i.e.*, the sum of the codon complex representations along the analyzed DNA strand, for (a) the segment 1 of homo sapiens chromosome 11 and (b) a uniformly distributed random sequence of bases. Again, the representation ($p = 1$, $q = 2$) has been used. The codon sequence path in Fig. 20a shows a tendency to progress preferentially in the complex plane from the origin along the real axis similar to that observed for the corresponding base sequence path in Fig. 14a. There is a larger random

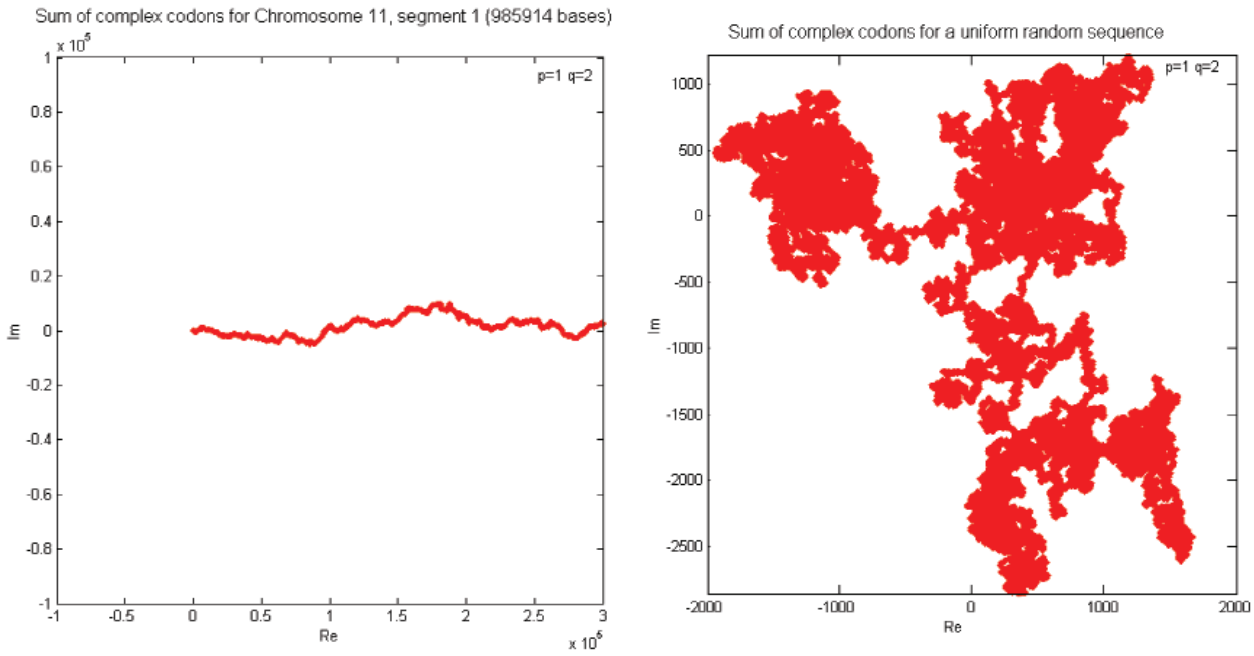


Fig. 20 The codon sequence paths, i.e., the sums of the complex codons along (a) the segment 1 of chromosome 11, (b) a uniformly distributed random sequence of bases. The codon sequence path in Fig 20a displays a similar trend to progress from the origin along the real axis as the corresponding base sequence path in Fig. 14a, while the codon sequence path in Fig 20b remains close to the origin as does the corresponding sum of complex base codes in Fig. 16.

drift in the direction of the imaginary axis. The comparison of the codon and base path advances along the real axis provide information about the distribution of the bases in the three possible codon positions along the analyzed sequences. The empirical ratio of the advances for the codon (Fig. 20a) and for the base (Fig. 14a) sequences is about 1.79. For a shift-invariant distribution of the four

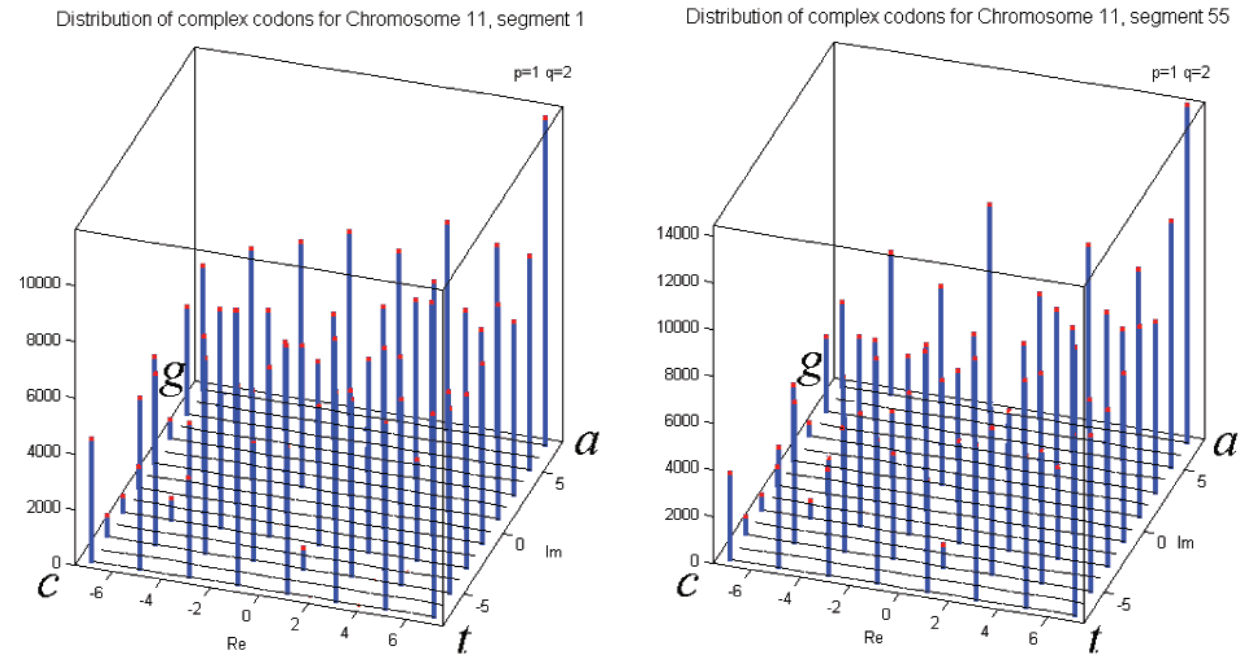


Fig. 21 Stem representation of the distribution of the 64 codons in the segments 1 (a) and 55 (b) of chromosome 11.

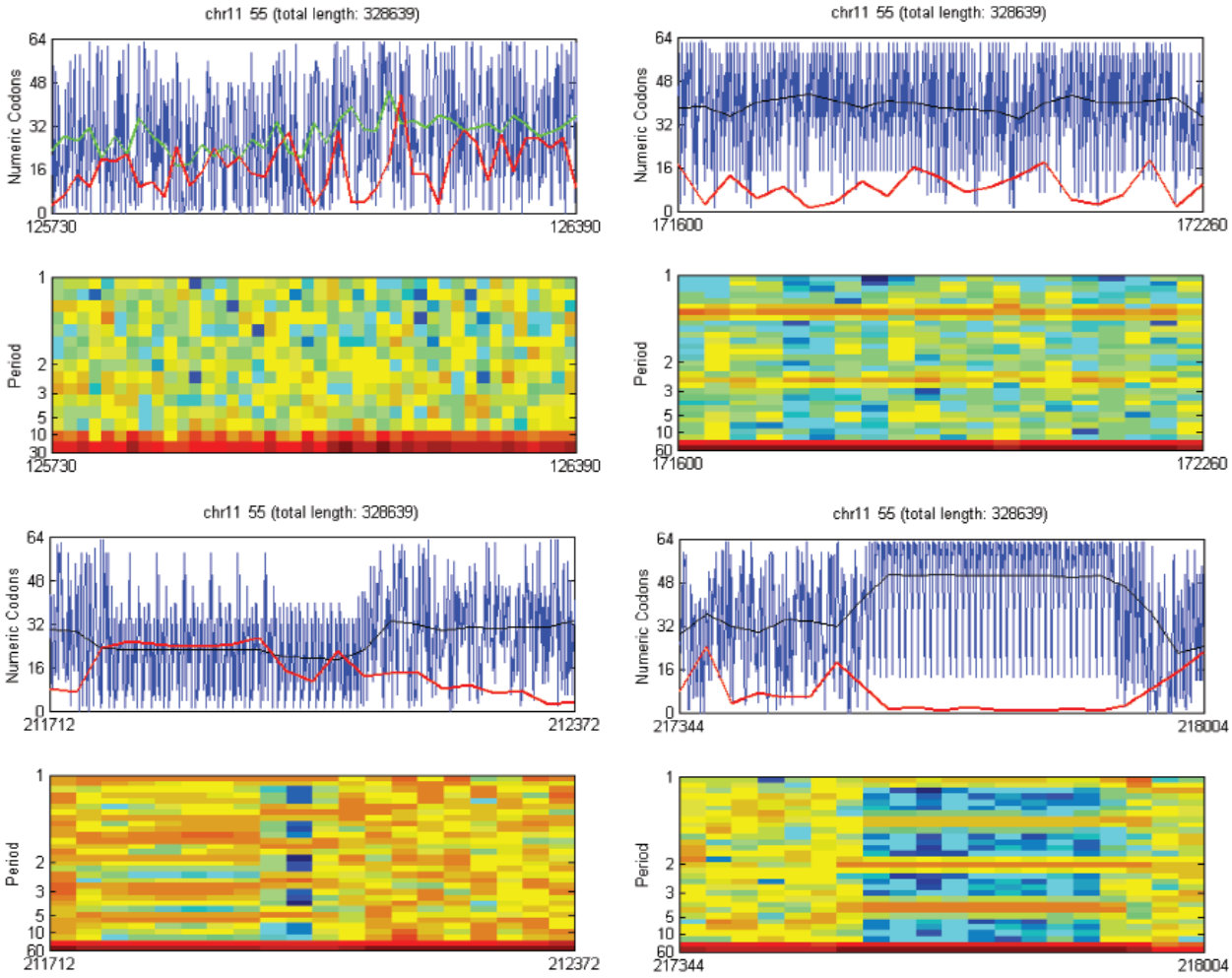


Fig. 22 The upper parts of the graphics show the codon genomic signals (blue) expressed as real sequences in base four, their average (local DC) components (green), and the scaled amplitudes of their local harmonics of period 3 (red). Lower diagrams show the spectrograms of the signals (abscissa: position of along codon sequence, ordinate: periods of harmonic components, color: encodes the amplitude. Figure (a) is typical for a random excerpt of chr 11, while figures (b), (c), and (d) comprise repeats that can be readily identified by the periodicities that show in both the signal and its corresponding spectrogram.

bases along the sequence, *i.e.*, in the case each of the bases would have had the same probability to occupy any of the three positions of the codons, even if these probabilities would have been different from one base to the other, the ratio of the codon path and base path advances along the real axes should have been $7/3 \approx 2.33$. This shows that the pair of bases A-T has globally a slightly lower probability to occupy the most significant position in the quasi-codons, in comparison with the other two, less significant, positions. At the same time, the wandering forth and back in the direction of the imaginary axis shows that the pairs A,G and C, T tend to alternate in being locally more frequent in the most significant position of the quasi-codons, despite the fact that their relative frequencies balance globally as it results from the analysis of the base sequence path in Fig. 14a. On the other hand, the codon sequence path of the random sequence in Fig 20b remains close to the origin as does the corresponding base sequence path in Fig. 16.

The frequency of apparition of each (quasi-) codon in a sequence can be investigated directly. Figure 21 gives the *codon stem representation* of the distributions for the 64 codons in the segments 1 (a) and 55 (b) of chromosome 11. Further work could attach

a meaning to the (quasi-) codon non-uniform distribution by the analysis of shorter DNA sequences.

DNA sequences can be also described as 1D genomic signals using a real representation (*e.g.*, p=1, q=0) at nucleotide, codon, or amino acid levels. Such signals have the advantage of simplicity and allow direct identification of some interesting features by human operators using visual or audio investigation. Figure 22 presents four excerpts of codon digital genomic signals, each 660 samples long. Both the *genomic signal diagram* and its *spectrogram* are given in each of the four figures. The spectrograms display the magnitude of the Fast Fourier Transform (FFT) of the discrete genomic signal computed using a periodic Hanning window of width 60 (or 30) samples, sliding with half that number of samples from one position to the other along the sequence. The magnitude of the FFT for each frequency-position rectangular domain is encoded by its color, with red – maximum, and blue – minimum. Instead of frequency, usually displayed in the spectrograms of time signals, the corresponding length of the period, which has a straightforward meaning in the case of spatial signals, is given on the ordinate axis. Figure 22a is typical for a random excerpt of chr. 11, the genomic signal showing little regularity. Figures 22b, c, and d comprise repeats that can be readily identified by the periodicities that show in both the signal and its corresponding spectrogram. Spectrograms in Fig. 22b, c, d show clearly periodic and (quasi-) periodic components of the genomic signal that correspond to the repeats in the sequence structure.

Discussion

It is to be mentioned that genomic signals corresponding to exons, *i.e.*, to the intra-gene signals, show a rather low auto-correlation, even for neighboring samples. This is a feature usually associated with noise and is consistent with the fact that the functionality of a protein is not given directly by its first order structure, the sequence of amino acids in the polypeptide, but by its higher order spatial structure. Such essentially qualitative features are immeasurable and the concept of quantitative value does not apply. Thus, samples in a sequence can not be correlated, as they do not encode quantities. On the other hand, the inter-gene genomic signals obtained from inter-gene, non-coding, DNA sequences have many features usually associated with "natural" discrete signals in which successive samples in a linear unidirectional sequence encode successive values of a discrete causal or random variable. Such signals are typically piecewise smooth and have a good correlation of close neighbors that decreases rather fast with the distance. The basic hypothesis from which results the need of a careful analysis of inter-gene signals is that this largest part of the genome contains three main components:

- structural elements of the chromosomes with functions in the replication of DNA, but with no direct heredity storage role,
- a repository of repeated excerpts from the genomes of ancestral entities with a possible role in the functioning of the immunitary system and the early selection of viable mutations,
- genomic signals that participate in the control of the gene expression and in the selection of co-transcribed exons

The statistic properties of these types of sequences are significantly different for one another. The control signals are the closest to classic "natural" signals, as the ones currently used in electrical engineering and communication, *i.e.*, signals containing piece-wise smooth regions separated by rather abrupt transitions or discontinuities. The chromosome structural elements include the centromeres and the telomeres that have specialized functions in chromosome duplication and cell reproduction. The centromeres have evolved as points of attachment of the spindle fibers in the metaphase and anaphase to pull apart the sister chromatides during mitosis. Centromeres have similar DNA sequences for any organism and a structure specialized for binding the kinetochore - an aggregate of proteins which further binds the spindle fibers. In most eukaryotes, satellite DNA having a simple and repetitive structure is found in the pericentromeric regions. Telomeres are specialized structures at each end of linear chromosomes and consists of long multiples of 4-6 base-pairs with sequences characteristic for each species. They prevent chromosomes from joining end-to-end and assure their complete replication.

Telomeres are made separately by an enzyme, distinct of DNA polymerase, and are added to the linear DNA molecule after its replication to compensate for the incompleteness of the replication of the lagging strand of DNA. Again, the subtelomeric regions of chromosomes are filled with large recent duplications of DNA segments [5]. Centromeres and telomeres produce quasi-periodic signals which are not significant from the genetic information point of view. As mentioned, the intra-gene DNA making-up the exons that joined together form the genes has, rather surprisingly, statistical features with a low autocorrelation, despite their essential role in bearing the genetic information. This is a consequence of the fact that the genetic information is only indirectly expressed in the linear sequence of nucleotides along DNA. The complex folding of polypeptide chains to generate the final 3D structure of the proteins hides the link between the structural features of DNA and functional features of the resulting proteins. As expected, the archive of ancestral genome excerpts contains sequences with the characteristics of both the intra-gene and the inter-gene materials.

Conclusions and further work

The paper proposes the Tetrahedron Representation [5, 11, 12, 14] of the GC that reflects better its structure and degeneration. Optimal symbolic-to-digital mappings for nucleotides and amino acids are proposed on this basis for various applicative purposes [11]. Some features of the resulting genomic signals are presented. Specifically, the paper reports for the first time the existence of a global helicoidal wrapping of the complex representations of the bases along DNA sequences, a large scale trend of genomic signals. New tools for genomic signal analysis, including the use of phase, aggregated phase, unwrapped phase, sequence path, stem representation of components' relative frequencies, as well as the transition analysis have been introduced at the nucleotide, codon and amino acid levels, and in a multiresolution approach. It is remarkable that, using the phase analysis, the existence of the helicoidal coiling of the complex representations of the bases along DNA sequences has been proved. This is a long range feature, that is maintained over distances of tens of millions of bases and reveals a statistical regularity of the *succession* of bases, not merely of the *distribution* of bases. This contradicts the current oversimplified genomic model that considers the inter-gene areas as domains of randomness, and recognizes only the meaningful structure of the exons [5, 9]. The use of the Projection Pursuit approach, specifically the Independent Component Analysis (ICA), on the genomic signals derived from extra-gene DNA sequences that do not encode proteins, could reveal signals that contribute in the control of the functioning of the genes, *i.e.*, the synthesis of the proteins.

References

1. **Venter J.C., *et al.***, Draft analysis of the human genome by celera genomics, *Science*, **291**: 1304-1351, 2001, *http://www.sciencemag.org*
2. **International Human Genome Sequencing Consortium**, Initial sequencing and analysis of the human genome, *Nature*, **409**: 860-911, 2001
3. **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*.
4. **National center for Biotechnology Information**, NLM, NIH, *ftp://ncbi.nlm.nih.gov/genoms/H.sapiens/*
5. **Gee H.**, Junk Science, Draft of a journey into the genome: What's there, *Nature*, *http://www.nature. com*.
6. **Cristea P.**, Independent component analysis for genetic signals, Short Course, San Jose, U.S.A., *SPIE Conference BiOS 2001 – International Biomedical Optics Symposium*, SC316: 20-26 January 2001, 21 January 2002
7. **Myers E.W., *et al.***, A Whole-Genome Assembly of Drosophila, *Science*, **287**: 2196-2204, 2000
8. **Doolittle W.F.**, Phylogenetic classification and the universal tree, *Science*, **284:** 2124-2128, 1999
9. **Andersson J.O., Nesbø C.L.**, Are there bugs in our genome?, *Science Express*: 2001
10. **Davis R.H., Weller S.G.**, *The Gist of Genetics*, Jones & Bartlett Publishers, 1996, 1998
11. **Anastassiou D.**, Frequency-domain analysis of biomolecular sequences, *Bioinformatics*, **16**: 1073-1081, 2000

12. **Cristea P.**, Genetic signals: An emerging concept, *Proceedings of IWSSIP 2001* pp. 17-22, 2001
13. **Cristea P.**, Genetic signal analysis, *Proceedings of ISSPA 2001 – The Sixth International Symposium on Signal Processing and its Applications*, Invited Paper, Kuala Lumpur, Malaysia, August 13 – 16, 2001, pp. 703–706
14. **Cristea P.**, Genetic signal representation and analysis, *SPIE Conference BiOS 2002 – International Biomedical Optics Symposium*, *Functional Monitoring and Drug-Tissue Interaction,* San Jose, USA, 19-25 January 2002, Conference 4623