Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

DNA sequence representation by trianders and determinative degree of nucleotides

DUPLIJ Diana[1], DUPLIJ Steven[†2]

([1]Institute of Molecular Biology and Genetics, Kiev 03143, Ukraine)

([2]Theory Group, Nuclear Physics Laboratory, Kharkov National University, Kharkov 61077, Ukraine)

[†]E-mail: Steven.A.Duplij@univer.kharkov.ua

Received Apr. 26, 2005; revision accepted May 30, 2005

Abstract: A new version of DNA walks, where nucleotides are regarded unequal in their contribution to a walk is introduced, which allows us to study thoroughly the "fine structure" of nucleotide sequences. The approach is based on the assumption that nucleotides have an inner abstract characteristic, the determinative degree, which reflects genetic code phenomenological properties and is adjusted to nucleotides physical properties. We consider each codon position independently, which gives three separate walks characterized by different angles and lengths, and that such an object is called triander which reflects the "strength" of branch. A general method for identifying DNA sequence "by triander" which can be treated as a unique "genogram" (or "gene passport") is proposed. The two- and three-dimensional trianders are considered. The difference of sequences fine structure in genes and the intergenic space is shown. A clear triplet signal in coding sequences was found which is absent in the intergenic space and is independent from the sequence length. This paper presents the topological classification of trianders which can allow us to provide a detailed working out signatures of functionally different genomic regions.

Key words: DNA walk, Triander, Determinative degree, Analysis DNA sequences, Dystrophin, Nucleotide
doi:10.1631/jzus.2005.B0743 **Document code:** A **CLC number:** Q34

INTRODUCTION

Genomic DNA sequence analysis using wide range of statistical methods (Torney *et al*., 1999; Bulmer, 1987; Luo *et al*., 1998; Nieselt-Struwe, 1997; Fickett *et al*., 1992; Buldyrev *et al*., 1998; Azbel, 1995) and various symmetry investigations (Findley *et al*., 1982; Hornos and Hornos, 1993; Bashford *et al*., 1997; Bhry *et al*., 1998; Forger and Sachse, 1998; Frappat *et al*., 1998) is an extremely important tool for extracting hidden information on the dynamic process of evolution, especially after the availability of fully sequenced genomes (Nakamura *et al*., 2000). One of the most promising approaches is the DNA walks method (Hamori, 1985; Gates, 1985; Berthelsen *et al*., 1992) first introduced by Azbel (1973; 1995) or genomic landscapes (Lobry, 1996) based on mapping of a sequence into one-, two-, or multidimensional metric space according to various specific rules. In brief, while drawing a DNA walk, the corresponding mappings assign a direction/unit vector to each nucleotide, to dinucleotide or to purine (pyrimidine). The resulting broken lines endow visual presentation to a formal sequence of 4 symbols, where inhomogeneous regions, fluctuations, patches etc. (Bernardi *et al*., 1985) are immediately seen. A modification of the DNA walks method deals with each position in codons independently, which gives three separate broken lines characterized by different angles and lengths (Cebrat and Dudek, 1998), where also addition and subtraction of DNA walks are also considered (Kowalczuk *et al*., 2001a).

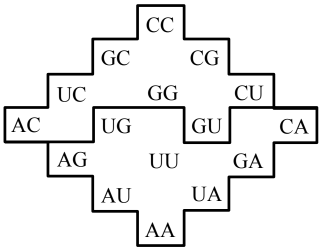
Here we introduce a new version of DNA walks, where all 4 nucleotides are regarded unequal in the sense that their contribution to a walk differs not only by direction, but also by module. It follows from the assumption in (Duplij and Duplij, 2000) that nucleotides have an inner abstract characteristic−the determinative degree (Duplij *et al.*, 2000) reflecting phenomenological properties of genetic code and is adjusted to nucleotides physical properties.

GENETIC CODE REDUNDANCY, DOUBLET MATRIX INNER STRUCTURE AND DETERMINATIVE DEGREE

As well-known, the genetic code is a highly organized system (Yčac, 1969) and has several general properties: triplet character, uniqueness, non-overlapping, comma less, redundancy (degeneracy), which means that most amino acids can be specified by more than one codon (Lewin, 1983; Stent and Kalindar, 1981).

From 64 possible codons one can extract 16 families each defined by the first two nucleotides. Let we denote a triplet (5′-1-2-3-3′) by XYZ. Then the codon sense can be fully determined by the first two nucleotides X and Y independent of the third Z. There are 8 unmixed families (all 4 codons of which encode the same amino acid), and 8 mixed families for which several patterns of assignment exist, in 6 of the latter the pyrimidine codons (Z=C, T) determine one amino acid, and the purine codons (Z=A, G) determine other amino acids or termination signals (in one family). It was found that two thirds of all DNA bases are identical for all organisms in the first two nucleotides in a triplet, and that the variability of DNA composition is given by the third base (Singer and Berg, 1991; Lewin, 1983).

All 16 doublets XY can be presented as the canonical matrix (Rumer, 1968)



This ordering is called the rhombic code (Karasev and Sorokin, 1997; Karasev, 1976). They are grouped together in 2 octets distinguished by the ability of amino acid determination: 8 doublets CC, AC, GC, CU, GU, UC, CG, GG determine amino acid independently of third base (upper part in the rhombic code), and so they can be called strong, and other 8 doublets AA, AU, UU, CA, GA, UG, AG, UA (lower part in the rhombic code) for which third base determines content of codons can be called weak ones (Rumer, 1968; Ratner, 1985). The strong set of doublets has the following (Ratner, 1985) relative content:

$$n(\mathrm{C}):n(\mathrm{G}):n(\mathrm{U}):n(\mathrm{A})=7:5:3:1$$

while the weak set has the reverse content:

$$n(\mathrm{C}):n(\mathrm{G}):n(\mathrm{U}):n(\mathrm{A})=1:3:5:7$$

Note that there is only one A in the strong octet, and one C in weak octet, and that all 4 doublets with Y=C completely determine amino acid, but only 2 doublets with Y=G completely determine it, while doublets with Y=A never determine amino acid. Thus, 4 nucleotides can be arranged in descending order C, G, U, A by their determinative ability (strength) (Rumer, 1969; 2000).

We introduce a numerical characterization of the empirical strength−determinative degree of nucleotide d_{x} in the following way

Pyrimidine	Purine	Pyrimidine	Purine	
C	G	T/U	A	
$d_{\mathrm{C}}=4$	$d_{\mathrm{G}}=3$	$d_{\mathrm{T/U}}=2$	$d_{\mathrm{A}}=1$	(1)
Very “strong”	“Strong”	“Weak”	Very “weak”	
Completely	In 2 cases	In 2 cases	Never	

which allows us change qualitative to quantitative description of genetic code structure (Duplij and Duplij, 2000; 2001)

We use the notation T/U, because genetic code is read from mRNA, and so we will not differentiate their determinative ability (strength) in what follows.

Let us present four bases Eq.(1) as the vector-column

$$\boldsymbol{V}=\begin{pmatrix}\boldsymbol{V}_1\\ \boldsymbol{V}_2\\ \boldsymbol{V}_3\\ \boldsymbol{V}_4\end{pmatrix}=\begin{pmatrix}\mathrm{C}^{(4)}\\ \mathrm{G}^{(3)}\\ \mathrm{T}^{(2)}\\ \mathrm{A}^{(1)}\end{pmatrix} \qquad (2)$$

and the corresponding vector-row

$$\boldsymbol{V}^{\mathrm{T}}=(\mathrm{C}^{(4)}\ \mathrm{G}^{(3)}\ \mathrm{T}^{(2)}\ \mathrm{A}^{(1)}) \qquad (3)$$

where the upper index for nucleotide denotes determinative degree.

We make the exterior product of vector-column Eq.(2) and vector-row Eq.(3) as follows (Duplij and Duplij, 2000; Duplij *et al*., 2000):

$$\boldsymbol{M}=\boldsymbol{V}\times\boldsymbol{V}^{\mathrm{T}}=\begin{pmatrix} \mathrm{C}^{(4)}\mathrm{C}^{(4)} & \mathrm{C}^{(4)}\mathrm{G}^{(3)} & \mathrm{C}^{(4)}\mathrm{T}^{(2)} & \mathrm{C}^{(4)}\mathrm{A}^{(1)} \\ \mathrm{G}^{(3)}\mathrm{C}^{(4)} & \mathrm{G}^{(3)}\mathrm{G}^{(3)} & \mathrm{G}^{(3)}\mathrm{T}^{(2)} & \mathrm{G}^{(3)}\mathrm{A}^{(1)} \\ \mathrm{T}^{(2)}\mathrm{C}^{(4)} & \mathrm{T}^{(2)}\mathrm{G}^{(3)} & \mathrm{T}^{(2)}\mathrm{T}^{(2)} & \mathrm{T}^{(2)}\mathrm{A}^{(1)} \\ \mathrm{A}^{(1)}\mathrm{C}^{(4)} & \mathrm{A}^{(1)}\mathrm{G}^{(3)} & \mathrm{A}^{(1)}\mathrm{T}^{(2)} & \mathrm{A}^{(1)}\mathrm{A}^{(1)} \end{pmatrix} \qquad (4)$$

It is remarkable that the matrix ***M*** in Eq.(4) fully coincides with the canonical matrix of doublets in the rhombic code, if and only if the vector ***V*** has the determinative degree order C, G, U, A in Eq.(1). Although there are 4!=24 possibilities to place 4 bases in row, but all others except one presented in Eq.(1) do not reflect the phenomenological properties of the genetic code. It follows that the intuitive rhombic code and genetic vocabulary (Rumer, 1968; Karasev and Sorokin, 1997; Karasev, 1976) have their own inner abstract structure uniquely defined by the exterior product of special vectors in Eq.(4). This ordering is also adjusted to the schemes (Sukhodolec, 1985; Maslov, 1981), and also (partially) adjusted to the half time of nucleotide substitution under mutational pressure (Kowalczuk *et al*., 2001b) and the nucleotides information weights (Dudek *et al*., 2002). Indeed these facts allows us to introduce the determinative degree, as an abstract variable being a numerical measure of nucleotide difference in ability to determine sense of codon (Duplij and Duplij, 2000; Duplij *et al*., 2000)

An analogous model for the triplet genetic code can be constructed using triple exterior product in the same way (Duplij and Duplij, 2000). We dispose the doublet matrix ***M*** on the XY plane and multiply it on the vector-column ***V*** Eq.(2) disposed along Z axis, i.e. we construct the triple exterior product

$$\boldsymbol{K}=\boldsymbol{V}\times\boldsymbol{M} \qquad (5)$$

Thus we obtain three-dimensional matrix over the set of all triplets, and, since each codon (except three terminal ones) corresponds to an amino acid, that can be treated as a cubic matrix model of the genetic code (Duplij and Duplij, 2000).

DETERMINATIVE DEGREE AND NUCLEOTIDE PROPERTIES

The connection bulk DNA structure and various properties of nucleotides were studied in (Zheltovsky *et al*., 1989; Govorun *et al*., 1992). It is well-known that by chemical structure the 4 nitrous bases can be divided into:

1. purine (A, G) and pyrimidine (C, T);
2. having amino (A, C) group and (G, T) keto group;
3. making 3 (strong) hydrogen bonds (C, G) and 2 (weak) hydrogen bonds (A, T).

They give rise to 3 symmetry transformations:

1. Purine-pyrimidine symmetry

$$\begin{pmatrix} \mathrm{T}^{(2)} \\ \mathrm{A}^{(1)} \\ \mathrm{C}^{(4)} \\ \mathrm{G}^{(3)} \end{pmatrix}=\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} \mathrm{C}^{(4)} \\ \mathrm{G}^{(3)} \\ \mathrm{T}^{(2)} \\ \mathrm{A}^{(1)} \end{pmatrix}=\boldsymbol{R}_{\mathrm{pur}}\boldsymbol{V} \qquad (6)$$

2. Amino-keto symmetry

$$\begin{pmatrix} \mathrm{A}^{(1)} \\ \mathrm{T}^{(2)} \\ \mathrm{G}^{(3)} \\ \mathrm{C}^{(4)} \end{pmatrix}=\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \mathrm{C}^{(4)} \\ \mathrm{G}^{(3)} \\ \mathrm{T}^{(2)} \\ \mathrm{A}^{(1)} \end{pmatrix}=\boldsymbol{R}_{\mathrm{amino}}\boldsymbol{V} \qquad (7)$$

3. Complementary symmetry (leaving invariant the double helix)

$$\begin{pmatrix} \mathrm{G}^{(3)} \\ \mathrm{C}^{(4)} \\ \mathrm{A}^{(1)} \\ \mathrm{T}^{(2)} \end{pmatrix}=\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} \mathrm{C}^{(4)} \\ \mathrm{G}^{(3)} \\ \mathrm{T}^{(2)} \\ \mathrm{A}^{(1)} \end{pmatrix}=\boldsymbol{R}_{\mathrm{compl}}\boldsymbol{V} \qquad (8)$$

where the even (because determinant is +1) permutation matrices $\boldsymbol{R}_{\mathrm{pur}}$, $\boldsymbol{R}_{\mathrm{amino}}$, $\boldsymbol{R}_{\mathrm{compl}}$ satisfy

$$\boldsymbol{R}_{\mathrm{pur}}\,\boldsymbol{R}_{\mathrm{amino}}\,\boldsymbol{R}_{\mathrm{compl}}=\boldsymbol{I}$$

and two of them, e.g. $\boldsymbol{R}_{\mathrm{pur}}$, $\boldsymbol{R}_{\mathrm{compl}}$, together with the identity matrix $\boldsymbol{I}$ form the dihedral group D_2 which is the symmetry group of the dihedron, or regular double-pyramid, with vertices on the unit-sphere (Ziegler, 1995). Another representation of this group by 3×3 rotational matrices is called a DNA group (Zhang, 1997).

The difference in the number of hydrogen bonds causes different interaction with its complementary nucleotide: each strong nucleotide (C and G) has 3 bonds and the energy of C-G interaction is −10.05 kJ/mol, and each weak nucleotide (T and A) has only 2 bonds with the energy of A-T interaction being −5.02 kJ/mol (Lewin, 1983). Therefore each base has its own properties and so dividing them into only 2 groups is not sufficient.

We then can search whether the ordering Eq.(1) is adjusted to some physical properties of nucleotides.

First we observe that the dipole moment of bases is proportional to the determinative degree as shown in Fig.1.

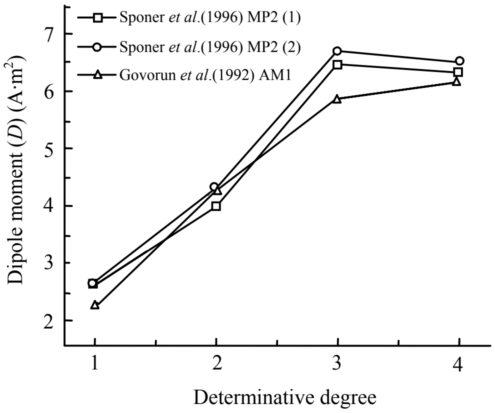


Fig.1 Dipole moment of DNA bases calculated by methods AM1 (Austin Model 1) (Govorun *et al.*, 1992) (triangles) and two modifications of MP2 (second-order Moller-Plesset perturbational method) (Sponer *et al.*, 1996) (squares and circles). The corresponding linear fits are: D_{AM1}=1.21+1.37d_{x} (R=0.96); $D_{\mathrm{MP1(1)}}$=1.45+1.36d_{x} (R=0.93); $D_{\mathrm{MP2(2)}}$=1.5+1.41d_{x} (R=0.93)

Then we see that the weight of hydration sites for bases is also proportional to the determinative degree Fig.2.

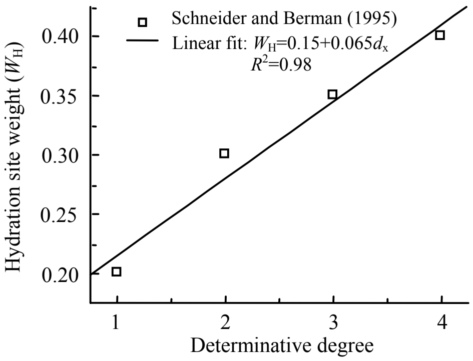


Fig.2 Weights of hydration site (Schneider and Berman, 1995)

We can conclude that the determinative degree reflects not only redundancy of genetic code in the third position, but also connected with some energy properties of the bases themselves.

TRIANDERS AND THEIR CHARACTERISTICS

It can be assumed that the phenomenological properties of genetic code and inequality of bases (reflected in Eq.(1)) will become apparent in real nucleotide sequences. Here we use the introduced determinative degree to build a new kind of sequence analysis based on some special modification of DNA walks method (Berthelsen *et al.*, 1992; Azbel, 1973; Lobry, 1996; Cebrat and Dudek, 1998).

TRIANDER CONSTRUCTION

We embed a nucleotide sequence into the two-dimensional determinative degree space (DD plane) in the following way. The axis assignment corresponds to the value of nucleotide determinative degree as

Axis x: {A}=(−1, 0); {T}=(+2, 0)
Axis y: {G}=(0, −3); {C}=(0, +4)

Moving along a sequence produces a walk in the determinative degree space which we call a determinative degree walk. In general, a current point on DD plane after i steps is determined by the coordinates

$$x_i^{\mathrm{DD}}=d_{\mathrm{T}}n_{\mathrm{T}}(i)-d_{\mathrm{A}}n_{\mathrm{A}}(i)$$

$$y_i^{\mathrm{DD}} = d_{\mathrm{C}} n_{\mathrm{C}}(i) - d_{\mathrm{G}} n_{\mathrm{G}}(i) \tag{9}$$

where $n_{\mathrm{X}}(i)$ is cumulative quantity of nucleotide X after ***I*** steps and d_{X} is the determinative degree of nucleotide X. The standard DNA walks (Cebrat and Dudek, 1998) (genome landscapes (Azbel, 1973)) have all $d_{\mathrm{X}}=1$ in Eq.(9), i.e.

$$\begin{aligned} x_i^{\text{standard}} &= n_{\mathrm{T}}(i) - n_{\mathrm{A}}(i) \\ y_i^{\text{standard}} &= n_{\mathrm{C}}(i) - n_{\mathrm{G}}(i) \end{aligned} \tag{10}$$

The one-dimensional (purine/pyrimidine) (PP) DNA walks are defined by only one coordinate, while x is chosen as position, i.e.

$$\begin{aligned} x_i^{\mathrm{pp}} &= i \\ y_i^{\mathrm{pp}} &= n_{\mathrm{C}}(i) + n_{\mathrm{T}}(i) - n_{\mathrm{A}}(i) - n_{\mathrm{G}}(i) \end{aligned} \tag{11}$$

Therefore, while purine/pyrimidine DNA walks manifestly show the purine/pyrimidine imbalance, the standard DNA walks Eq.(10) applied for one strand show DNA asymmetry (Wu, 1991; Francino and Ochman, 1997) (violation of the second parity rule (Sueoka, 1995)), the determinative degree walk Eq.(9) visually shows strength imbalance in one strand.

Then we build 3 independent determinative degree walks beginning from 1st nucleotide with step 3 (due to the triplet structure of genetic code). In this way we obtain 3 broken lines (branches) starting from the point of origin, and each of them presents the determinative degree walk through the following nucleotide numbers:

1st branch goes through 1,4,7,10,13,... positions;

2nd branch goes through 2,5,8,11,14,... positions;

3rd branch goes through 3,6,9,12,15,... positions. These 3 branches on the determinative degree plane are called triander.

If 1st letter corresponds to the first start codon nucleotide, then the triander branches represent nucleotide sets in three codon positions independently.

As distinct from previous versions of DNA walks in which all 4 nucleotides are regarded equivalent in the sense that they give equal by module shifts, in our approach each nucleotide gives contribution different by module (which is taken equal to its determinative degree). So, although we obtain at first sight isomorphic to (Cebrat and Dudek, 1998) plot, the trianders show not only quantitative composition and pure statistical laws of symbol strings, but also reflect the connection between nucleotide sequences and inner phenomenological properties of genetic code and physicochemical properties of bases.

As an example of triander we will take the dystrophin gene which is the largest gene found in nature, measuring 2.4 millions bases pairs, responsible for Duchenne (DMD) and Becker (BMD) muscular myodystrophy (Yagi *et al.*, 2003). The dystrophin RNA is differentially spliced, producing a range of different transcripts, encoding a large set of protein isoforms. Dystrophin is a large, rod-like cytoskeletal protein found at the inner surface of muscle fibers. The triander for the dystrophin gene is shown in Fig.3.

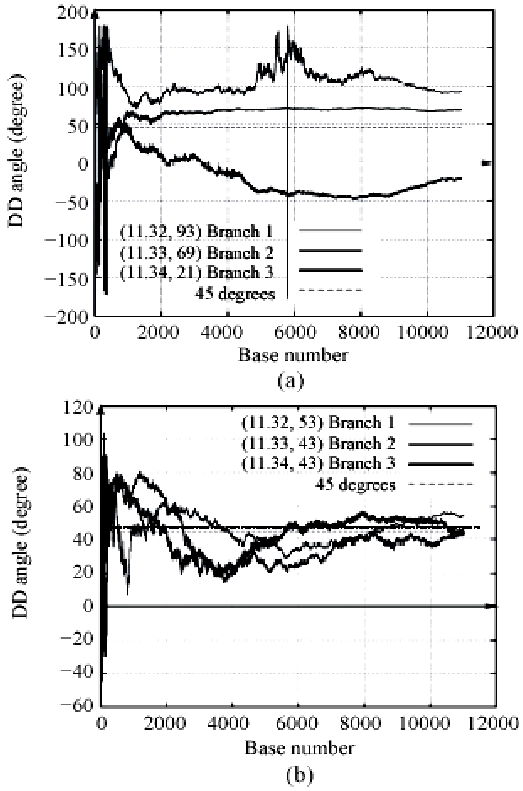


Fig.3 Current DD angle for triander of the Homo Sapiens dystrophin gene (a) and a shuffled sequence of the same nucleotide composition (b)

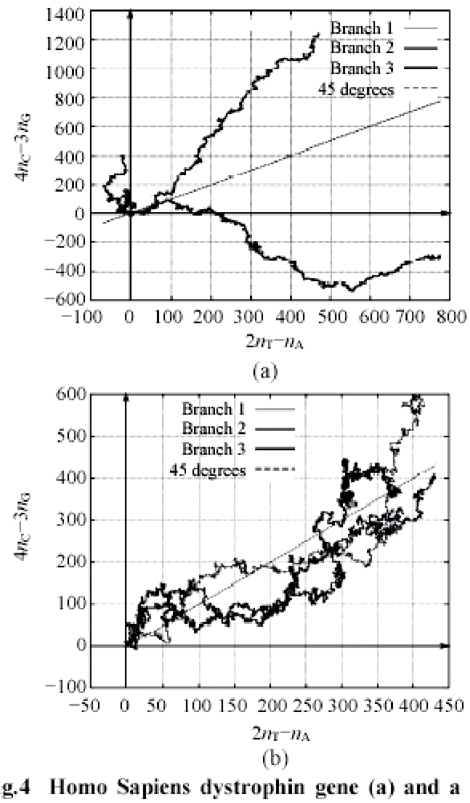
For comparison we also show the triander for a shuffled sequence of the same nucleotide composition. Obviously the ideal triander for uniformly random sequence consists of 3 flowing together lines from the origin having 45 degrees slope. This line also corresponds to the symmetric sequence satisfying the second parity rule (Sueoka, 1995): N_C=N_G, N_T=N_A. Such lines are presented on all triander plots below for normalization.

DETERMINATIVE DEGREE ANGLE

An important visual characteristic of a triander is the slope of its branches, we call it determinative degree (DD) angle (α), which for a current point (i) can be calculated by

$$\tan\alpha(i)=\frac{4n_C(i)-3n_G(i)}{2n_T(i)-n_A(i)} \qquad (12)$$

Evidently, for uniformly random sequence or a symmetric sequence satisfying the parity rule 2 (Sueoka, 1995) the angle will be 45 degree (horizontal dashed line of the plots below), and so the difference from this value will say be about nontrivial ordering. The plots of current values of α for the dystrophin gene and for a shuffled sequence of the same nucleotide composition are presented in Fig.4.



(a)

(b)

Fig.4 Homo Sapiens dystrophin gene (a) and a shuffled sequence of the same nucleotide composition (b)

We stress that trianders show not only quantitative composition, but allow us to find local motives in a more clear way, because different modules for nucleotides lead to less number of superposition and selfintersections. Also trianders more accurately reflect the tendency of the sequence as a whole similarly to DNA walks. Thus triander can be treated as a picture, genome passport or genogram of a given sequence.

If we remember that third base in codon has maximal redundancy, then 3rd branch of a triander gets a definite physical sense. Let us assume that the determinative degree is an additive variable (which can be made as first approximation at least (Duplij and Duplij, 2000)), then 3rd branch can show the current strength of the sequence, that is the bulk ability to determine sense of codon. In this scheme the other two branches can be treated as 3rd branch with shifted ORF (Open Reading Frame).

EUCLIDEAN AND MANHATTAN DISTANCES

As the measure of the sequence strength we can choose length of the radius-vector from the origin to the current point of the triander, i.e. the Euclidean distance

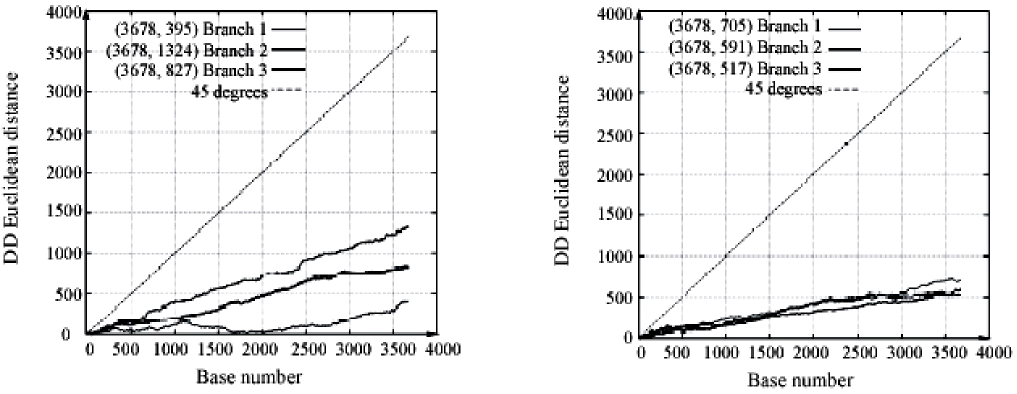
$$D_E(i)=\sqrt{(4n_C(i)-3n_G(i))^2+(2n_T(i)-n_A(i))^2} \qquad (13)$$

We can also use the Manhattan distance (also known as rectilinear distance, and it can be treated as the distance that would be travelled to get from one data point to the other if a grid-like path is followed (a car driving in a city laid out in square blocks, like Manhattan))

$$D_M(i)=|4n_C(i)-3n_G(i)|+|2n_T(i)-n_A(i)| \qquad (14)$$

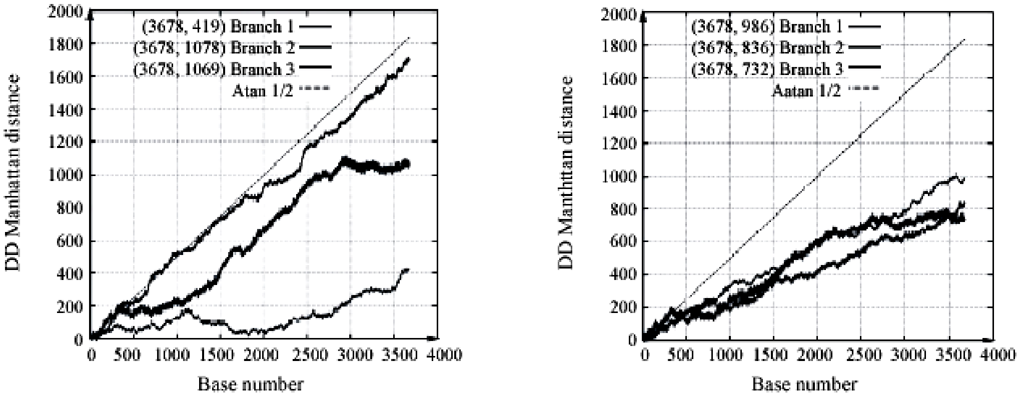
which is the distance between two points measured along axes at right angles (Skiena, 1990).

In case of symmetric sequence (equal number of all nucleotides) at the step i the Euclidean and Manhattan distances are $D_E(i)=i/\sqrt{2}$ and $D_M(i)$=i/2 (which is shown by dashed lines in Fig.5 and Fig.6.



(a) (b)

Fig.5 Current DD Euclidean distance for triander of the Homo Sapiens dystrophin gene (a) and a shuffled sequence of the same nucleotide composition (b)



(a) (b)

Fig.6 Current DD Manhattan distance for triander of the Homo Sapiens dystrophin gene (a) and a shuffled sequence of the same nucleotide composition (b)

VISUALIZATION OF THE GENETIC CODE TRIPLET NATURE

Now we make sure that the triplet character of the genetic code can be seen directly from sequences representation by trianders. As an example we take gene of Homo Sapiens Che-1 mRNA. We consider additionally analogs of trianders with different phases=4,5,7. The result is presented in Fig.7 showing that only the case phase=3 provide nontrivial ordering leading to definite branches, that is we have clear visual presentation of the strong triplet signal.

In such a way one could search for higher phase statistical correlations and possible structures, if any, in nucleotide sequences.

TRANSFORMATIONS OF TRIANDERS

Here we illustrate how the symmetry transformations influence the triander. As an example we take the Homo Sapiens dystrophin in Fig.3, and the result of various symmetry transformations Eq.(6)~Eq.(8) and reversing the sequence is shown on Fig.8. The original non-transformed triander is shown in Fig.3. We observe that the reverse triander is very similar to the original one in Fig.3.

THREE-DIMENSIONAL TRIANDERS

The previously constructed two-dimensional

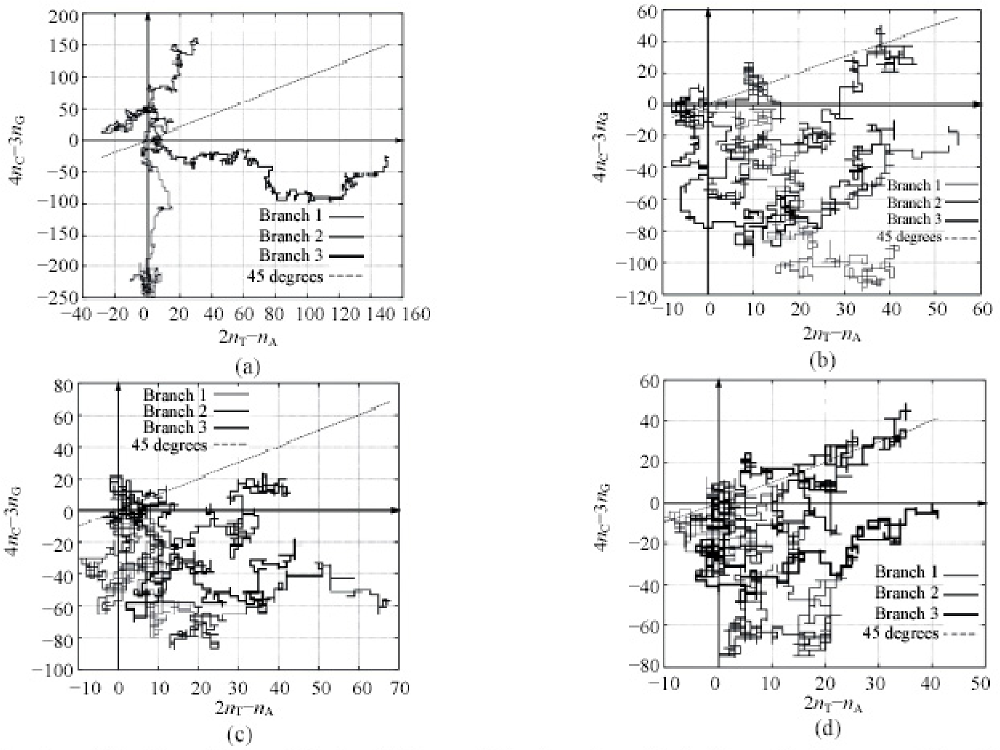


Fig.7 Trianders of the Homo Sapiens Che-1 mRNA gene (a) and analogs of trianders with phases equalling 4 (b), 5 (c),7 (d). The strong triplet signal is clearly seen

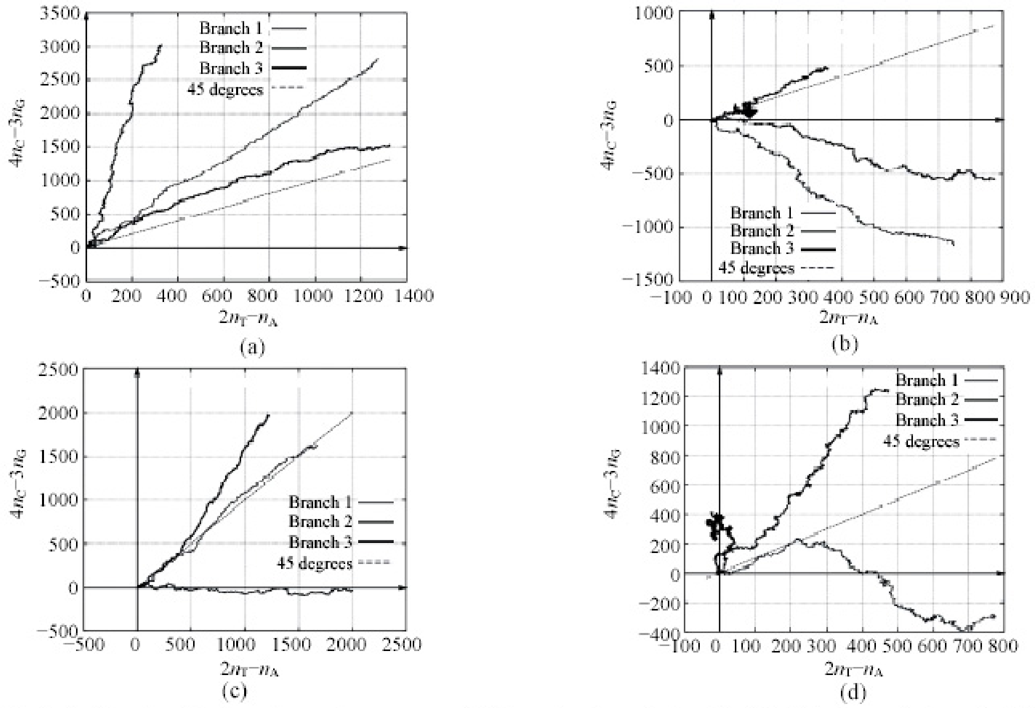


Fig.8 The trianders for the transformed sequences of Homo Sapiens dystrophin (DMD), transcript variant D140ab, mRNA in case of the amino-keto symmetry (a), purine-pyrimidine symmetry (b), complementarity symmetry (c) and reverse sequence reading (d)

trianders have disadvantage in form, because it is not clear, where in the sequence a given point is. To improve this we introduce three-dimensional trianders which are defined by the formula

$$\begin{aligned} x_i^{3\mathrm{D}} &= d_\mathrm{T} n_\mathrm{T}(i) - d_\mathrm{A} n_\mathrm{A}(i) \\ y_i^{3\mathrm{D}} &= d_\mathrm{C} n_\mathrm{C}(i) - d_\mathrm{G} n_\mathrm{G}(i) \\ z_i^{3\mathrm{D}} &= i \end{aligned} \tag{15}$$

which can be treated as mixing of one-dimensional and two-dimensional cases with taking into account the determinative degree. Then, any on the DD space structure can be definitely visually localized using vertical axis. In the Fig.9 we show examples of chaotic and ordered trianders.

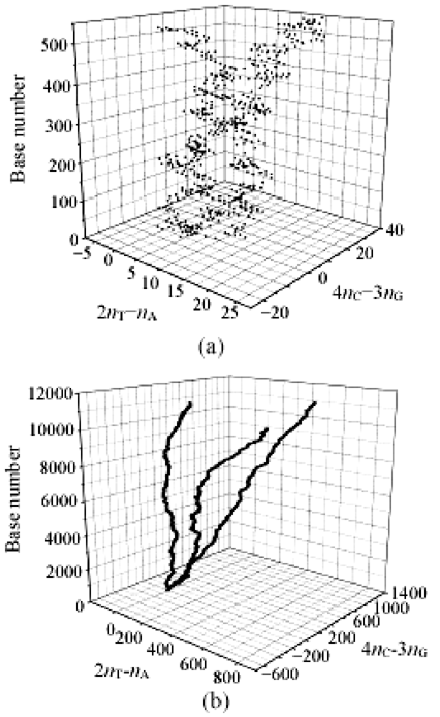


Fig.9 Three-dimensional trianders of the Dengue virus, complete genome, AF226686 (a) and Homo Sapiens dystrophin mRNA (b)

All the graphs start from one point, the origin, and have different length (which can be simply calculated from Eq.(15)), characterizing them as a whole.

In Fig.10 we show the three-dimensional triander of the Homo Sapiens zinc finger and its shuffled version.

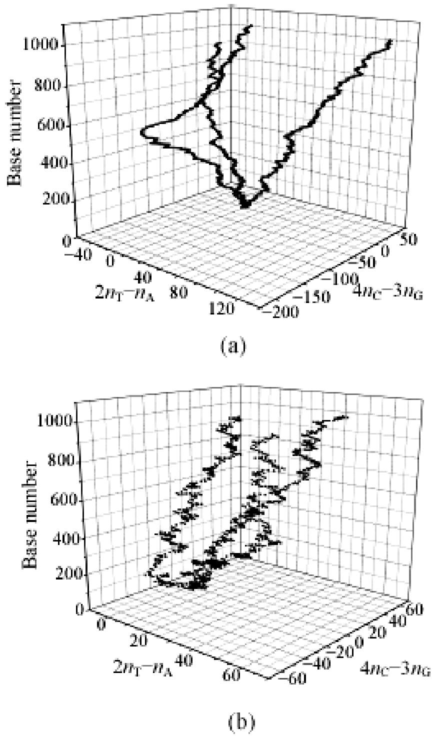


Fig.10 Three-dimensional trianders of the Homo Sapiens zinc finger protein 265 (ZNF265), mRNA (a) and its shuffled version (b)

TOPOLOGICAL CLASSIFICATION OF TRIANDERS

Here we propose the topological classification of trianders by their branches belonging to different quadrants on the DD plane and to the number of intersections and return point. This makes possible studying the fine structure of any length sequences with exactly established functions (genes, intergenic space, repeat regions, etc.) and comparing various loci, as well as searching for homological regions, which can allow us to work out mathematically strong genomic signature formalism (Wu, 2002; Bergmann *et al*., 2002).

We note that there exist many types of trianders. A triander corresponding to a gene we call a genogram, and a triander corresponding to intergenic space we call a gapgram. If some branches intersect each other we call them intersecting trianders, if a branch intersects itself producing knots, we call it knot triander. Branches with (multiple) return points, we call returned trianders. Thus, the determinative degree walk topologizing in our sense means that we identify trianders having definite structure topological features (knot, intersection, return point) and place them into a special class.

So we may hope that such topological classification of trianders can actually help in visually solving the inverse problem: for a given sequence to predict its possible function.

Let $n_X(i)$ be cumulative quantity of nucleotide X after i steps, then the DD plane quadrants are defined by Eq.(9), and therefore

I: $2n_T(i)-n_A(i)>0$; $4n_C(i)-3n_G(i)>0$;
II: $2n_T(i)-n_A(i)<0$; $4n_C(i)-3n_G(i)>0$;
III: $2n_T(i)-n_A(i)<0$; $4n_C(i)-3n_G(i)<0$;
IV: $2n_T(i)-n_A(i)>0$; $4n_C(i)-3n_G(i)<0$.

After examination of around 2000 eukaryotic and prokaryotic sequences we found that all trianders can be distinguished into several types. The first type is a chaotic triander with no definite branch structure, other types can be called ordered trianders. To work out the general classification of ordered trianders and description of branches we introduce the notion:

$$\text{Type A-B-C}_{\text{F}}^{(x,\,y)}\ \text{(E)},$$

where A is quadrant where the 1st branch lies, B is quadrant where the 2nd branch lies, C is quadrant where the 3rd branch lies; E is comprised of triander characteristics as a whole; indices F and (x, y) describe properties of corresponding separate branches (also for A and B) which will be explained below.

In general, there are 4^3=64 possible ordered triander types classified by quadrants only. We will identify trianders which differ by permutation, because it corresponds to ORF shift, thus decreasing to 24 types. Nevertheless, observation showed that there exist only 7 triander types: I-I-I, I-I-II, I-I-III, I-I-IV, I-II-III, I-II-IV, I-III-IV. For example, the Type I-I-II includes the Types I-II-I and Types II-I-I, if we shift ORF to 1 and 2, but on figures we present only first of them.

If e.g. a branch crosses from I quadrant to II quadrant, we denote that by fraction I/II. For instance, the triander of Homo Sapiens dystrophin gene Fig.3 is of Type II-I-I/IV.

The additional qualitative features of triander as a whole observed from sequence examination are

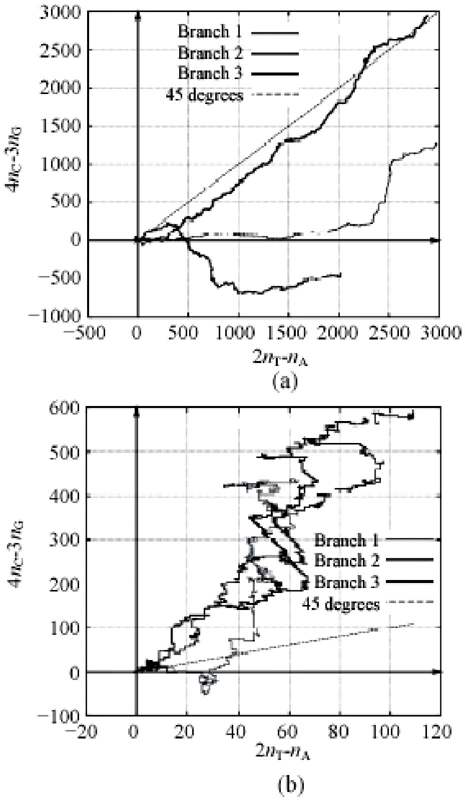
E=sharp, flat, parallel

For branch properties we have x, y denotes axis to which a branch is parallel, F=blurry, loop, smooth, oscillative (horizontal, vertical).

Separately we can describe interaction of branches as:

1. Single intersection of A and B is denoted by sign A#B, which gives intersecting triander;

2. Multiple intersection A and B is called braiding and denoted A&B, which gives braiding triander (Fig.11).



(a)

(b)

Fig.11 Intersecting triander, Chaetosphaeridium globosum chloroplast, complete genome (a) and braiding triander Homo Sapiens cytochrome P450 2f1 (CYP2F1P) (b)

We thoroughly analyzed 150 sequences different by function and evolution level, and for each sequence there were also constructed 100 shuffled sequences having the same nucleotide composition, but not coinciding with the examined one. For every class we show a typical triander of Fig.12, where the following real sequences are presented:

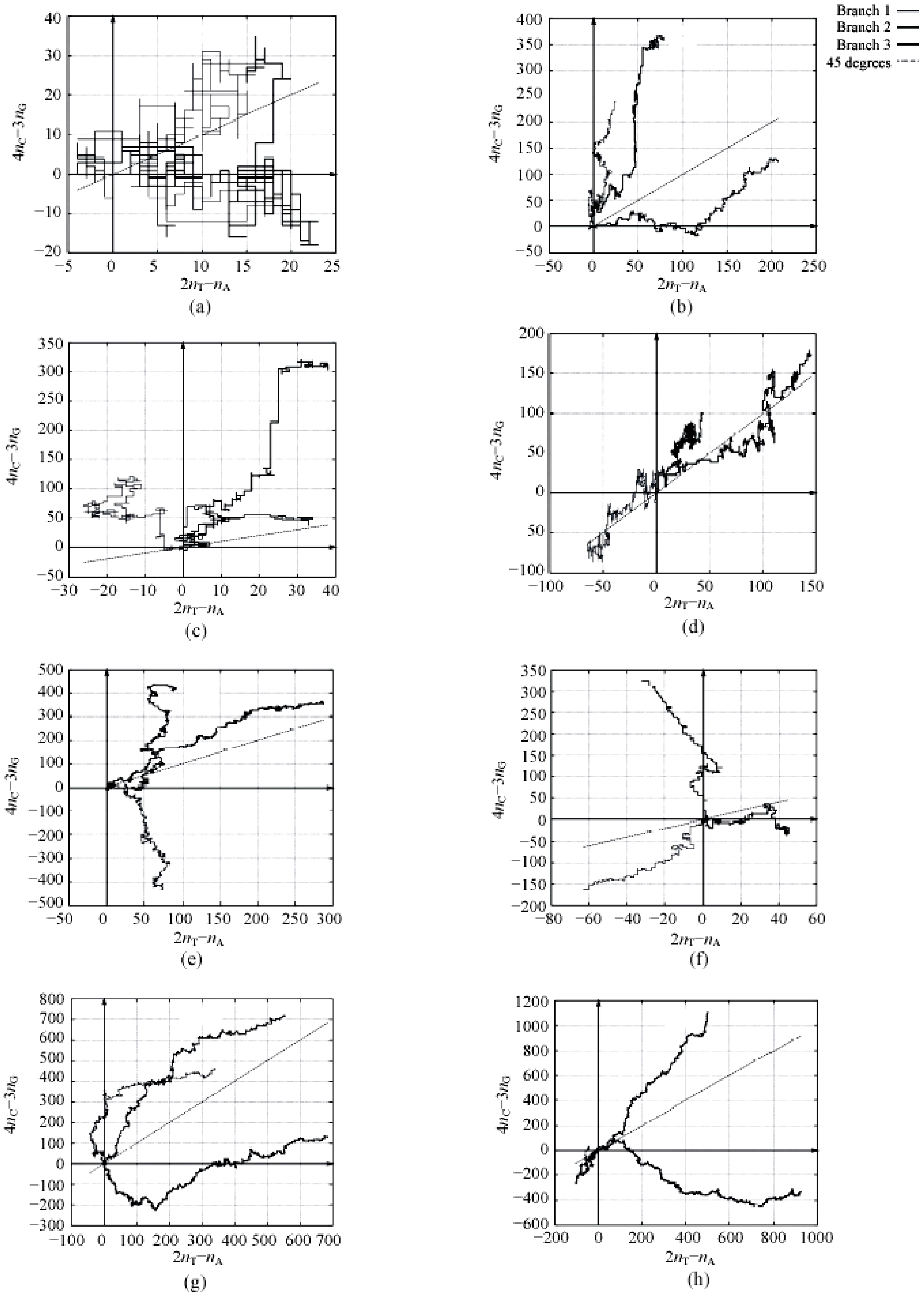


Fig.12 Classification triander (a) Chaotic triander; (b) Type I-I-I^{y}; (c) Type II$^{\text{blury}}$-I^{x}-I^{y} (flat); (d) Type III$^{\text{oscill}}$-I$^{\text{blury}}$-I$^{\text{oscill}}$; (e) Type IV-I#I; (f) Type III-II-I^{x} (sharp); (g) Type II#I-IV; (h) Type III$^{\text{loop}}$-I-IV

1. Chaotic triander (Fig.12a). Dengue virus type 1 strain FGA/NA d1d, intergenic space: AF226686;

2. Type I-I-I^{y} (Fig.12b). Homo Sapiens cytochrome P450, family 2, subfamily F, polypeptide 1 (CYP2F1), mRNA: NM000774;

3. Type II$^{\text{blury}}$-I^{x}-I^{y} (flat) (Fig.12c). Homo Sapiens Cbp/p300-interacting transactivator, mRNA: NM006079;

4. Type III$^{\text{oscill}}$-I$^{\text{blury}}$-I$^{\text{oscill}}$ (Fig.12d). Homo Sapiens collagen, type IX, alpha 2 (COL9A2), mRNA: NM001852;

5. Type IV-I#I (Fig.12e). Caenorhabditis elegans immunoglobulin domain-containing protein family member (106400), mRNA: NM171617;

6. Type III-II-I^{x} (sharp) (Fig.12f). Homo Sapiens H1 histone family, member 5 (H1F5), mRNA: NM005322;

7. Type II#I-IV (Fig.12g). Homo Sapiens dystrophin (muscular dystrophy, Duchenne and Becker types) (DMD), transcript variant D140ab, mRNA: NM004022;

8. Type III$^{\text{loop}}$-I-IV (Fig.12h). Homo sapiens utrophin (homologous to dystrophin) (UTRN), mRNA: NM007124.

Further more careful topological classification and analysis of two- and three-dimensional trianders can be made using some of the topological curves methods (Petrovskiy, 1938; Rokhlin, 1974; Arnold and Oleinik, 1979) or the knot theory (Turaev, 1994; Kauffman, 1991).

CONCLUSION

We can conclude that the introduced determinative degree DNA walk method confirms the mosaic structure of genome, shows parts with different nucleotide content and strength, and so allows us to find the fine structure of nucleotide sequences.

We propose a general method for identification of DNA sequence by triander, which can be treated as a unique genogram, gene passport, etc. The two- and three-dimensional trianders are introduced and their features are studied.

The difference of the nucleotide sequences fine structure in genes and the intergenic space is shown. Also there is a clear triplet signal in coding loci which is absent in the intergenic space and is independent of the sequence length, but depends on composition only. All plots were compared with corresponding shuffled sequences of the same nucleotide composition, which allows us to extract real ordering effect from composition influence. We have constructed the classification of trianders, on the ground that a detailed working out of signatures of functionally different genomic regions can be made.

ACKNOWLEDGMENTS

We sincerely thank S. Cebrat and M.R. Dudek for their kind hospitality at SmORFland (Inst. Microbiology, Wroclaw) and for sharing their useful DNA walk experience with us. Also, we are grateful to A.Y. Berezhnoy, N.A. Chashchin, A.V. Elskaya, A.A. Gusakov, V. Knecht, A.I. Kornelyuk, S.S. Maliuta, G.C. Kourinnoy, P. Mackiewicz, B.V. Novikov, L.A. Livshits, S. Sachse, Y.G. Shckorbatov, O.A. Tretyakov for their very fruitful discussions, and J. Bashford, G. Findley, H. Grubmüller, D.M. Hovorun, P. Jarvis, E. Korotkov, N. Tidjani, D.C. Torney, C. Zhang for kind sending us their interesting papers and useful comments during correspondence. We are thankful to S.M. Donets for language checking and V. Kalashnikov for creation of original special programs for sequence analysis and indispensable assistance in the related GNUPLOT/ PERL/C programming.

References

Arnold, V.I., Oleinik, O.A., 1979. Topology of real algebraic manifolds. *Vestnik Mosk. Univ. Ser. I Mat. I Mekh.*, **A249**:7-17.

Azbel, M.Y., 1973. Random two-component one-dimensional Ising model for heteropolymer melting. *Phys. Rev. Lett.*, **31**:589-592.

Azbel, M.Y., 1995. Universality of DNA statistical structure. *Phys. Rev. Lett.*, **75**:168-171.

Bashford, J.D., Tsohantjis, I., Jarvis, P.D., 1997. Codon and nucleotide assignments in a supersymmetric model of the genetic code. *Phys. Lett.*, **A233**:481-488.

Bergmann, S., Ihmels, J., Barkai, N., 2002. Self-similarity Limits of Genomic Signatures. Weizmann Inst. Science Preprint, Cond-mat/0210038, Rehovot, p.12.

Bernardi, G., Olofsson, B., Filipski, J., 1985. The mosaic genome of warm-blooded vertebtates. *Science*, **228**:953-958.

Berthelsen, C.L., Glazier, J.A., Skolnick, M.H., 1992. Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Phys. Rev.*, **A45**:8902-8913.

Bhry, T., Cziryk, A., Vicsek, T., Major, B., 1998. Application of vector space techniques to DNA. *Fractals*, **6**:205-210.

Buldyrev, S.V., Dokholyan, N.V., Goldberger, A.L., Havlin, S., Peng, C.K., Stanley, H.E., Viswanathan, G.M., 1998. Analysis of DNA sequences using methods of statistical physics. *Physica*, **A249**:430-438.

Bulmer, M., 1987. A statistical analysis of nucleotide sequences of introns and exons in human genes. *Mol. Biol. Evol.*, **4**:395-405.

Cebrat, S., Dudek, M.R., 1998. The effect of DNA phase structure on DNA walks. *Eur. Phys. J.*, **3**:271-276.

Dudek, M., Cebrat, S., Kowalczuk, M., Mackiewicz, P., Nowicka, A., Mackiewicz, D., Dudkiewicz, M., 2002. Information Weights of Nucleotides in DNA Sequences. Inst. Microbiology Preprint, Cond-mat/0301371, Wroclaw, p.8.

Duplij, D., Duplij, S., 2000. Symmetry analysis of genetic code and determinative degree. *Biophysical Bull. Kharkov Univ.*, **488**:60-70.

Duplij, D., Duplij, S., 2001. Determinative degree and nucleotide content of DNA strands. *Biophysical Bull. Kharkov Univ.*, **525**:86-92.

Duplij, D., Duplij, S., Chashchin, N., 2000. Symmetric properties of genetic code. *Biopolymers and Cell*, **16**:449-454.

Fickett, J.W., Torney, D.C., Wolf, D.R., 1992. Base compositional structure of genomes. *Genomics*, **13**:1056-1064.

Findley, G.L., Findley, A.M., McGlynn, S.P., 1982. Symmetry characteristics of the genetic code. *Proc. Natl. Acad. Sci. USA*, **79**:7061-7065.

Forger, M., Sachse, S., 1998. Lie Superalgebras and the Multiplet Structure of the Genetic Code I: Codon Representations. Inst. de Mat. e Estat, Preprint, Math-ph/9808001, Sao Paulo, p.23.

Francino, M.P., Ochman, H., 1997. Strand asymmetries in DNA evolution. *Trends Genet.*, **13**:240-245.

Frappat, L., Sciarrino, A., Sorba, P., 1998. A crystal base for the genetic code. *Phys. Lett.*, **A250**:214-221.

Gates, M.A., 1985. Simpler DNA sequence representations. *Nature*, **316**:219.

Govorun, D.N., Danchuk, V.D., Mishchuk, Y.R., Kondratyuk, I.V., Radomsky, N.F., Zheltovsky, N.V., 1992. AM1 calculation of the nucleic acid bases structure and vibrational spectra. *J. Mol. Structure*, **267**:99-103.

Hamori, E., 1985. Novel DNA sequence representations. *Nature*, **314**:585-586.

Hornos, J.E.M., Hornos, Y.M.M., 1993. Model for the evolution of the genetic code. *Phys. Rev. Lett.*, **71**:4401-4404.

Karasev, V.A., 1976. Rhombic version of genetic vocabulary based on complementary of encoding nucleotides. *Vest. Leningr. Univ.*, **1**:93-97.

Karasev, V.A., Sorokin, S.G., 1997. Topological structure of the genetic code. *Genetika*, **33**:744-751.

Kauffman, L.H., 1991. Knots and Physics. World Sci., Singapore.

Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., 2001a. DNA asymmetry and replicational mutational pressure. *J. Appl. Genet.*, **42**:553-577.

Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., Cebrat, S., 2001b. High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. *BMC evolutionary biology*, **17**:1-13.

Lewin, B., 1983. Genes. Wiley and Sons, New York.

Lobry, J.R., 1996. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**:323-326.

Luo, L., Lee, W., Jia, L., Ji, F., Tsai, L., 1998. Statistical correlation of nucleotides in a DNA sequence. *Phys. Rev.*, **E58**:861-871.

Maslov, S.Y., 1981. On the nature of biological code and its possible evolution. *Biophysics* (Moscow), **26**:632-635.

Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: Status for the year 2000. *Nucl. Acds. Res.*, **28**:292.

Nieselt-Struwe, K., 1997. Graphs in sequence spaces: A review of statistical geometry. *Biophys. Chem.*, **66**:111-131.

Petrovskiy, I.G., 1938. On the topology of real plane algebraic curves. *Ann. Math.*, **39**:189-209.

Ratner, V.A., 1985. Structure and evolution of the genetic code. *Itogi Nauki i Tekhniki. Ser. Mol. Biol.*, **21**:158-197.

Rokhlin, V.A., 1974. Complex orientation of real algebraic curves. *Func. Anal. Appl.*, **8**:71-75.

Rumer, U.D., 1968. Sistematics of codons in the genetic code. *DAN SSSR*, **183**:225-226.

Rumer, U.D., 1969. On codon sistematics in the genetic code. *DAN SSSR*, **187**:937-938.

Rumer, U.D., 2000. Genetic code as a system. *Soros Educational J.*, **6**:15-22.

Schneider, B., Berman, H.B., 1995. Hydration of DNA bases is local. *Biophysical J.*, **69**:2661-2669.

Singer, M., Berg, P., 1991. Genes and Genomes. University Science Books, Mill Valley.

Skiena, S., 1990. Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. Addison-Wesley, Reading.

Sponer, J., Leszczynski, J., Vetterl, V., Hobza, P., 1996. Base stacking and hydrogen bonding in protonated cytosine dimer: The role of molecular ion-dipole and induction interactions. *J. Biomolecular Structure and Dynamics*, **13**:695-705.

Stent, G., Kalindar, R., 1981. Molecular Genetics. Mir, Moscow, p.487.

Sueoka, N., 1995. Intrastrand parity rules of dna base composition and usage biases in synonymous codons. *J. Mol. Evol.*, **40**:318-325.

Sukhodolec, V.V., 1985. A sence of the genetic code: Reconstruction of the prebiologocal evolutin stage. *Genetika*, **21**:1589-1599.

Torney, D.C., Whittaker, C.C., Xie, G., 1999. The statistical properties of human coding sequences. *J. Mol. Biol.*, **286:**1461-1469.

Turaev, V.G., 1994. Quantum Invariants of Knots and 3-Manifolds. W. de Greuter, Berlin.

Wu, C., 1991. DNA strand asymmetry. *Nature*, **352**:114.

Wu, Z.B., 2002. Self-similarity limits of genomic signatures. Inst. Mechanics Preprint, Cond-mat/0212091, Beijing, p.12.

Yagi, M., Takeshima, Y., Wada, H., Nakamura, H., Matsuo, M., 2003. Two alternative exons can result from activation of the cryptic splice acceptor site deep within intron 2 of the dystrophin gene in a patient with as yet asymptomatic dystrophinopathy. *Hum. Genet.*, **267**:164-170.

Yčac, M., 1969. The Biological Code. North-Holland, Amsterdam.

Zhang, C.T., 1997. A symmetrical theory of DNA sequences and its applications. *J. Theor. Biol.*, **187**:297-306.

Zheltovsky, N.V., Samoilenko, S.A., Govorun, D.N., 1989. In Spectroscopy of Biological Molecules. Societa Editrice Esculapio, Bologna, p.159-172.

Ziegler, G.M., 1995. Lectures on Polytopes. Springer-Verlag, Berlin.