UNIVERSAL SEVEN-CLUSTER STRUCTURE OF GENOME FRAGMENT DISTRIBUTION: BASIC SYMMETRY IN TRIPLET FREQUENCIES

A. Gorban[1], A. Zinovyev[2], T. Popova[3*]
[1] *Centre for Mathematical Modeling, University of Leicester, Leicester, UK;* [2] *Institutes des Hautes Etudes Scientifiques, Bures-sur-Yvette, France;* [3] *Institute of Computational Modeling, Krasnoyarsk, Russia, e-mail: tanya@icm.krasn.ru*
[*] *Corresponding author*

Abstract: We found a universal seven-cluster structure in bacterial genomic sequences and explained its properties. Based on the analysis of 143 completely sequenced bacterial genomes available in GenBank in August 2004, we show that there are four 'pure' types of the seven-cluster structure observed. The type of cluster structure depends on GC content and reflects basic symmetry in triplet frequencies. Animated 3D-scatters of bacterial genomes seven-cluster structure are available on our web site: http://www.ihes.fr/~zinovyev/7clusters.

Key words: triplet frequencies; genome fragments; codons; mean-field approximation; symmetry; visualization

1. INTRODUCTION

Coding information is the main source of statistical inhomogeneity in bacterial genomes. There exist well-known compositional differences between codon positions in coding regions, which we observed using pure data exploration strategy and determined as universal seven-cluster structure. We considered 64D vectors of non-overlapping triplet frequencies in sliding window within the direct strand of bacterial DNA sequence (see details in section 2.1). Visualization of 64D vectors data set in the subspace of the first three principal components shows a clear cluster structure presented in Figure 1 by examples of two genomes and in the form of the general pattern.

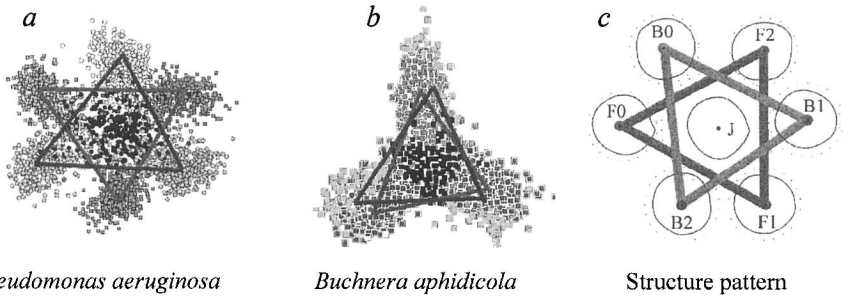


Figure -1. Universal seven-cluster structure: (*a* and *b*) visualization of genome fragment distribution and (*c*) pattern of the cluster structure.

Biologically, there are seven significantly different positions of a sliding window according to coding information: three possible reading frames of coding regions in two complementary strands plus non-coding regions. The obtained cluster structure corresponds to biologically relevant one with a higher than 90 % accuracy at the nucleotide level (Gorban et al., 2003).

This cluster structure is universal in the sense that it is observed in any bacterial genome and with any type of statistic, which takes into account three possible reading frames. The structure is basic in the sense that it is revealed in the analysis in the first place, reflecting the principal source of sequence non-randomness. The structure is well represented by a 3D-plot, while initially we have 64D vectors of frequencies. It has a symmetric and appealing flower-like pattern, hinting that there should be a symmetry in our statistics (triplet frequencies) governing the pattern formation.

The seven-cluster structure was implicitly used since long time ago in gene recognition problem (Borodovsky et al., 1993; Salzberg et al., 1998). Specific clustering of relatively short genome fragments is used in entropic or Hidden Markov Modeling (HMM) statistical approaches (Audic et al., 1998; Baldi, 2000; Bernaola-Galvan et al., 2000; Nicolas et al., 2002), which are effective due to non-randomness in DNA sequence being reflected by the seven-cluster structure. However, the structure itself was described explicitly and visualized for several genomes only recently (Zinovyev et al., 2003). We refer to the structure itself because of simplicity and formality of the presented approach: it is based on the 64 frequencies of non-overlapping triplets in sliding window and data exploration strategy regardless of any model of genome organization.

Several particular cases of flower-like pattern were observed in the 9D space of Z-coordinates (Ou et al., 2003). However, the structure was reported to pertain to GC-rich genomes only, while Zinovyev (2002) and Gorban et al. (2003) demonstrated that AT-rich genome of *H. pylori* had a

flower-like cluster structure. This fact shows this simple and basic structure to be far from being completely understood and described.

In this paper, we show that the seven-cluster structure is determined by a single parameter: the genomic GC content. Based on the analysis of 143 completely sequenced bacterial genomes, available in GenBank in August 2004, we describe four 'pure' types of the structure and basic symmetries in triplet frequencies that they reflect.

2. SEVEN-CLUSTER STRUCTURE FOR COMPACT GENOMES

2.1 Algorithm of data table construction

To visualize the seven-cluster structure for some bacterial genome, a data set is prepared as follows.

1. The complete genome sequence and its annotation are extracted from GenBank. Let N be the length of a given sequence. One defines a step size p (~ 10–100 bp) and a fragment size W (odd number ~ 300–400 bp).
2. For $i = 1\ldots [N/(W + 1 + p)]$, a fragment of the length $W + 1$ centered at position $S_i = ip + W/2$ is clipped from the DNA sequence.
3. According to genome annotation, each clipped fragment is labeled by one of the F0, F1, F2, B0, B1, B2, and J labels with the letter F for S_i being inside the forward strand CDS, the letter B for S_i being inside the complementary strand CDS, and J for S_i of inter-CDS regions. Here, indices 0, 1, or 2 are equal to shift modulo 3 of the first base pair in the clipped fragment relative to the first base pair of the start codon in the corresponding CDS frame.
4. Frequencies of non-overlapping triplets are counted within each clipped fragment forming the table of 64D vectors of frequencies, which are characterized by window position S_i and annotation label.
5. Standard principal component analysis (PCA) is performed, and the first three principal components are calculated. Each vector is projected into the 3D basis of principal components and visualized (Figure 1).

2.2 Overall properties of seven-cluster structure

Typical 3D plots of data distribution obtained for bacterial genomes are shown in Figure 1. Distribution has well-detectable seven-cluster structure: each type of annotated points constitutes their own cluster, and clusters are separated from each other with visible gaps. Automatic clustering by the

method of k-means with Euclidean distance attributes a data point to its annotated cluster with a more than 90 % accuracy (see Gorban et al., 2003 for more details), which corresponds to efficiency of automatic gene identification methods for bacterial genomes (Mathe et al., 2002). The seven-cluster structure reflects well-known differences in three letter word distributions between seven types of fragments. However, we show that these seven types of fragments are extracted primarily without any preliminary knowledge of genome organization, as a main source of sequence heterogeneity. Further, we consider some features of triplet frequencies in bacterial genomes that govern the cluster structure formation and its particular shape.

2.3 Phase triangles

According to annotation labels in the data table, we calculate an arithmetic mean vector of frequencies for each type of genome fragments, denoted here as f, $f^{(1)}$, and $f^{(2)}$ for coding regions of the forward strand (labels F0, F1, and F2) and as $\hat{f}$, $\hat{f}^{(1)}$, and $\hat{f}^{(2)}$ for coding regions of the complementary strand (labels B0, B1, and B2). Referring to the seven-cluster structure, these six vectors should be the centers of corresponding clusters. These centers constitute two *phase triangles* in 64D space of frequencies (Figure 1): forward strand triangle (f, $f^{(1)}$, $f^{(2)}$) and complementary strand triangle ($\hat{f}$, $\hat{f}^{(1)}$, $\hat{f}^{(2)}$).

Having codon frequencies $f = (f_{\mathrm{AAA}}, f_{\mathrm{AAC}}, \ldots, f_{\mathrm{TTG}}, f_{\mathrm{TTT}})$, one can easily calculate estimations $P^{(1)}f$ and $P^{(2)}f$ of the shifted distributions $f^{(1)}$ and $f^{(2)}$ under the assumption that no correlation exists in codon order:

$$P^{(1)}f_{ijk} \equiv \sum_{lmn} f_{lij} f_{kmn}\,,\quad P^{(2)}f_{ijk} \equiv \sum_{lmn} f_{lmi} f_{jkn}\,,\ i, j, k, l, m, n \in \{\mathrm{A,T,G,C}\}. \quad (1)$$

An estimation of complementary strand phase triangle vertex $\hat{f}$ on the basis of f (denoted here as $C^R f$) consists in rearrangement of f coordinates according to reverse reading and complementary translation of the corresponding codons: $C^R f_{ijk} = f_{\hat{k}\hat{j}\hat{i}}$, i, j, $k \in \{\mathrm{A,T,G,C}\}$, where $\hat{i}$ is complementary to ith nucleotide. Estimations of shifted distributions $\hat{f}^{(1)}$ and $\hat{f}^{(2)}$ are calculated as $C^R P^{(1)} f$ and $C^R P^{(2)} f$.

Five calculated distributions $P^{(1)}f$, $P^{(2)}f$, $C^R f$, $C^R P^{(1)} f$, and $C^R P^{(2)} f$ appeared to be very close to the centers of corresponding clusters obtained according to genome annotation (Gorban et al., 2003; see also section 2.5). It means that (1) the *codon frequencies* determine seven-cluster structure and (2) between-codon correlations are, in average, much less than within-codon.

2.4 Mean-field approximation of codon frequencies

In order to reveal the properties of codon frequencies that guarantee the clusters appearance, we are to consider a *mean-field* approximation of ***f***. Mean-field approximation, ***mf***, assumes 64 codon frequencies to be modeled by 12 position-specific nucleotide frequencies as follows:

$$(\boldsymbol{mf})_{ijk} = p_i^1 p_j^2 p_k^3, \quad i,j,k \in \{\mathrm{A, T, G, C}\}, \tag{2}$$

where

$$p_i^1 = \sum_{jk} f_{ijk}\ ,\ p_j^2 = \sum_{ik} f_{ijk}\ ,\ p_k^3 = \sum_{ij} f_{ijk}\ , i,j,k \in \{\mathrm{A,C,G,T}\}.$$

Mean-field approximation models codon frequencies under the hypothesis of independent position-specific nucleotide generation in codon. This approximation is widely used in literature (Bernaola-Galvan et al., 2000).

Two another vertexes of the mean-field phase triangle, $P^{(1)}\boldsymbol{mf}$ and $P^{(2)}\boldsymbol{mf}$, can be easily calculated under the same hypothesis:

$$P^{(1)}(\boldsymbol{mf})_{ijk} = p_i^2 p_j^3 p_k^1,\ P^{(2)}(mf)_{ijk} = p_i^3 p_j^1 p_k^2,\quad i,j,k \in \{\mathrm{A,T,G,C}\}. \tag{3}$$

There exists exactly triangle ($\boldsymbol{mf}$, $P^{(1)}\boldsymbol{mf}$, $P^{(2)}\boldsymbol{mf}$) in 64D space of triplet frequencies because no more than three different triplet distributions can be produced under the accepted model.

Assuming coding regions in the complementary strand to have the same position-specific frequencies, one can easily get complementary strand phase triangle of mean-field approximation: ($C^R\boldsymbol{mf}$, $C^R P^{(1)}\boldsymbol{mf}$, $C^R P^{(2)}\boldsymbol{mf}$).

Thus, the differences in nucleotide frequencies dependent on their position in codon provide existence of six possible 64D vectors of triplet frequencies. They are $\boldsymbol{mf}$, $P^{(1)}\boldsymbol{mf}$, $P^{(2)}\boldsymbol{mf}$, $C^R\boldsymbol{mf}$, $C^R P^{(1)}\boldsymbol{mf}$, and $C^R P^{(2)}\boldsymbol{mf}$ in mean-field approximation notation. Theoretically, some vectors can coincide, but only in such a way that makes the resulting set to consist of six (non-degenerated case), three (partially degenerated case), two, or one (completely degenerated cases) vectors.

Non-degenerated case corresponds to the presence of six 'coding' clusters, which is typical of a number of real genomes. Coincidence of phase triangles of the forward and complementary strands in any combination of their vertexes produces the partially degenerated case, which is an ordinary case too for certain bacterial genomes (see section 3.1).

The completely degenerated cases appear, if true and shifted codon distributions are identical. Referring to Eq. (3), they correspond to position-independent distribution of nucleotides in codons. Denoting this distribution as $\boldsymbol{m}$, one calculates it using four position-independent nucleotide frequencies:

$$m_{ijk} = p_i\, p_j p_k\,, \quad p_i = 1/3(p_i^1 + p_i^2 + p_i^3)\,, \quad i,j,k \in \{\mathrm{A, T, G, C}\}. \qquad (4)$$

It corresponds to the simplest zero order model of coding regions and constitutes approximately the center of phase triangle. Completely randomized distribution $\boldsymbol{m}$ and its complementary reversion $C^R\boldsymbol{m}$ would coincide iff $p_A = p_T$ and $p_C = p_G$. Thus, the number of degenerated clusters depends on the interstrand symmetry in nucleotide frequencies. However, the cases were called 'degenerated' because of unusual for real genomes coincidence of coding and shifted distributions.

2.5 Information content in the triplet distributions

Visual illustration of the information content of some true and modeled triplet distributions is shown in Figure 2 by the examples of two bacterial genomes. Pairwise distance between the two triplet distributions $\boldsymbol{g}$ and $\boldsymbol{h}$ was calculated according to symmetrized Kullback–Leibler distance

$$D^{SYM}(\boldsymbol{g};\boldsymbol{h}) = \frac{1}{2}\left(\sum g_i \ln\frac{g_i}{h_i} + \sum h_i \ln\frac{h_i}{g_i}\right).$$

Metric multidimensional scaling (MDS) technique was applied to visualize the distributions on 2D plane on the basis of the obtained pairwise distances.

Figure 2 shows relative information content of true phase triangle distributions ($\boldsymbol{f}, \boldsymbol{f}^{(1)}$, and $\boldsymbol{f}^{(2)}$) and mean-field approximation ones ($\boldsymbol{mf}$, $P^{(1)}\boldsymbol{mf}$, and $P^{(2)}\boldsymbol{mf}$). The calculated shifted distributions $P^{(1)}\boldsymbol{f}$, $P^{(2)}\boldsymbol{f}$ and the center of true phase triangle $\boldsymbol{f}^{(\mathrm{av})}$ (which is arithmetic mean of $\boldsymbol{f}$, $\boldsymbol{f}^{(1)}$, and $\boldsymbol{f}^{(2)}$) are shown as well with the origin set at the $\boldsymbol{m}$ point. Information content of all shown distributions is proportional to their distance to the origin.

The maximum of information is contained in the codon distribution $\boldsymbol{f}$, which is the most distant point from the origin. Its high information content gives more contrast cluster structure and better quality of unsupervised gene recognition. Calculated shifted distributions $P^{(1)}\boldsymbol{f}$, $P^{(2)}\boldsymbol{f}$ are very close to real shifted distributions $\boldsymbol{f}^{(1)}$, $\boldsymbol{f}^{(2)}$, confirming the fact of small correlation in codon order for bacterial genomes.

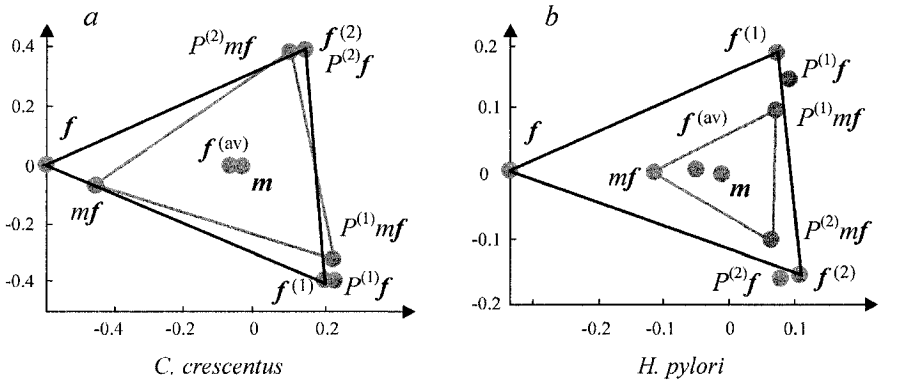


Figure -2. MDS plots representing relative information content in triplet distributions.

The difference in sizes between true phase triangle and mean-field one reflects the presence of correlation in the order of nucleotides. This difference is small for *C. crescentus* genome and considerable for *H. pylori* genome. Among all considered bacterial genomes, *H. pylori* demonstrates the largest difference between *f* and *mf*, while *C. crescentus* is in the very middle. It agrees with the fact that bacterial codon usage is reasonably well approximated by its mean-field distribution (Bernaola-Galvan et al., 2000).

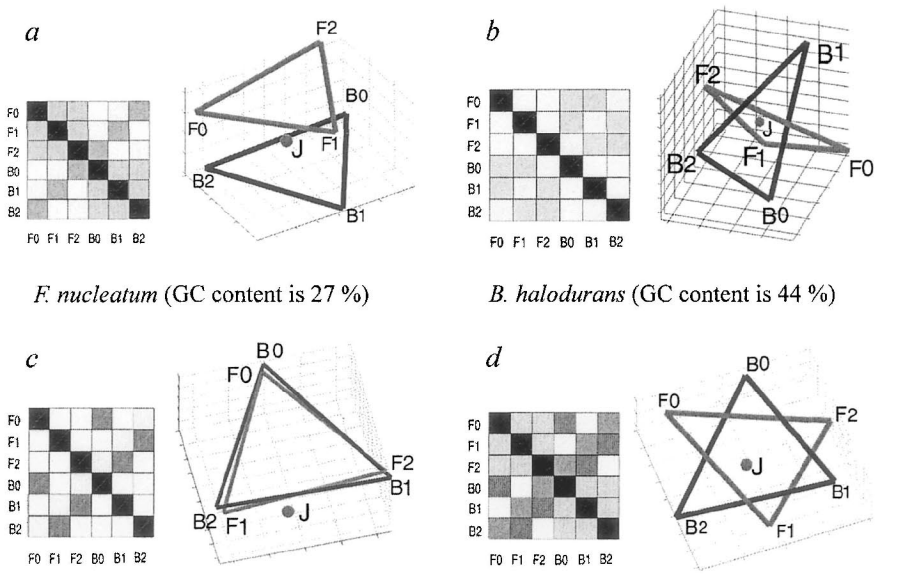
3. TYPES OF SEVEN-CLUSTER STRUCTURE

The skeleton of seven-cluster structure is created by positional relationship of two phase triangles: the forward strand and complementary strand ones. The types of mutual position of cluster triangles refer to the classical problem of symmetry (or asymmetry) between the forward and backward DNA strands (Mrazek et al., 1998; Lobry and Sueoka, 2002) as well as to the pattern of symmetric properties of codon usage.

3.1 Four 'pure' types of seven-cluster structure

Among the seven-cluster structures of all considered bacterial genomes, we picked out four 'pure' types of positional relationship of two phase triangles, which are shown in Figure 3 by the examples of corresponding genomes.

The first pattern—'parallel triangles' (Figure 3*a*)—corresponds to the AT-rich genome of *Fusobacterium nucleatum* (GC content is 27 %). The phase triangles exhibit an opposite rotation of the vertex indices with the F1 vertex meeting the B1 one. This pattern is commonly observed in AT-rich genomes.



F. nucleatum (GC content is 27 %)

B. halodurans (GC content is 44 %)

E. coli (GC content is 51 %)

E. coli (GC content is 51 %)

Figure -3. Four 'pure' types of universal seven-cluster structure and corresponding matrices of pair distances between cluster centers in 64D space (the distances are shown by gray scale intensity: the darker color corresponds to the less distance): (*a*) 'parallel triangles', (*b*) 'perpendicular triangles', (*c*) coinciding triangles, and (*d*) flower-like structure.

The second pattern—'perpendicular triangles' (Figure 3*b*)—belongs to the genome of *Bacillus halodurans* (GC content is 44 %). The 'perpendicular triangles' structure is only an approximate picture; the real configuration is almost 6-dimensional due to the distance matrix symmetry: all non-diagonal elements have similar big value.

The third pattern (Figure 3*c*) represented by the *Escherichia coli* genome (GC content, 51 %) corresponds to partially degenerated case of coinciding phase triangles of the forward and complementary strand with F0–B0, F1–B2, and F2–B1 pairs of coinciding vertexes.

The fourth flower-like pattern (Figure 3*d*) being represented by the GC-rich genome of *Streptomyces coelicolor* (GC content, 72 %) is close to plane regular hexagon with non-coding cluster slightly displaced in the direction perpendicular to the hexagon plane. The displaced J cluster position is connected with the CG content of non-coding regions, which is less than that of coding regions. The same situation was observed for the third pattern of coinciding triangles. The four patterns are typical of triplet distributions of bacterial genomes observed in nature by the moment. The other ones combine features of these four 'pure' types.

3.2 Genomic GC content and type of seven-cluster structure

To represent distribution of seven-cluster structure types over bacterial genomes, we created a data table of 64D vectors of *codon* frequencies (codon usage) for all the 143 bacterial genomes and visualized it in 2D space of the first two principal components. It is a well-known fact that many properties of the codon usage are correlated with genomic GC content (Lobry, 1997; Wan et al., 2004). The first principal component explains near 60 % of the total variance in codon usage, and factor scores reflect GC content: coding, genomic, and position-specific ones are equally highly correlated with factor scores having $r > 0.95$. Figure 4 shows PCA plot of bacterial genomes distribution in the space of their codon usage. Ascribing a bacterial genome to the type of its seven-cluster structure resulted from automatic classification of distance matrices. Locations of all the mentioned genomes are highlighted by big markers and denoted by their source name abbreviation.

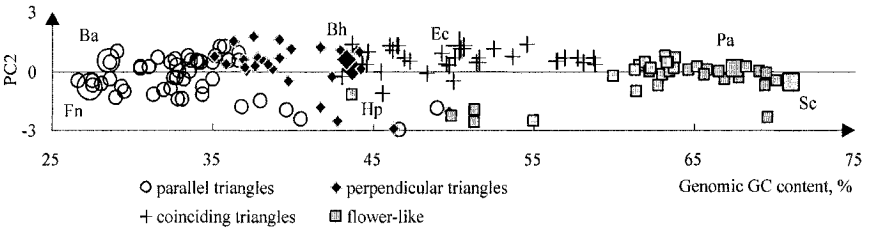


Figure -4. PCA plot of bacterial genomes distribution in the 64D space of codon frequencies with their seven-cluster structure attribution to one of the four pure types. The first principal component scores were replaced by the corresponding genomic GC content ($r > 0.95$); axes were adjusted in length to reflect approximately the explained variance ratio (60 % and 8 %).

The presented PCA plot confirms the fact that bacterial codon usage is determined essentially by the GC content and so do the type of seven-cluster structure. In 64D space of codon frequencies, bacterial codon usage located near one dimensional curve, that is almost straight line reflecting the GC content scale. In general, going along the curve, one meets at first 'parallel triangles', which transform gradually to 'perpendicular triangles'. On this way, one can meet flower-like patterns in one of the 2D projections, like that of *H. pylori* genome (Zinovyev et al., 2003). Then, the pattern goes to the coinciding triangles with genomic GC content around 50 %. Further pairs F0–B0, F1–B2, and F2–B1 diverge in the same 2D plane and after 55 % threshold in GC content, the flower-like structures are present almost exclusively.

3.3 Basic symmetry in triplet frequencies

The types of seven-cluster structure depend strongly on the genomic GC content, because bacterial codon usage is at most determined by it. The shape of mutual position of phase triangles presents a visual illustration of the peculiarities of bacterial codon usage and in particular, it reflects the symmetry (or asymmetry) between the forward and backward DNA strands and other symmetric properties of codon usage. Thus, interstrand symmetry is displayed by coincidence of the centers of phase triangles. Namely, the 'parallel triangles' type shows a strong asymmetry between DNA strands of AT-rich bacterial genomes, whereas other types reflect relative interstrand symmetry of corresponding genomes.

In terms of the mean-field approximation, the structure type reflects symmetries in the set of 12 position-specific frequencies with respect to the phase shift and complementary reverse operations. Since phase triangles exist due to the difference between position-specific frequencies p_i^1, p_i^2, p_i^3 and randomized ones p_i, $i \in \{A,T,G,C\}$, some features of cluster structure could be explained in terms of these differences.

Plane structures like hexagon and coinciding triangles are easy to observe in 3D space of the differences in coding GC content between position-specific GC frequency and the mean one: $\Delta_{GC}^k = p_C^k + p_G^k - (p_C + p_G)$, $k = 1, 2, 3$. GC-rich bacterial genomes perform the special pattern of Δ_{GC}^k: $\Delta_{GC}^1 \approx 0, \Delta_{GC}^2 < 0, \Delta_{GC}^3 > 0$, or $(0 - +)$, if denoting them by triplet of signs. Phase shifts operation rotate the signs, while complementary reverse one only reads them from back to front (GC content is not changed under C^R transformation). The forward strand phase triangle $(0 - +, - + 0, + 0 -)$ and the complementary strand one $(+ - 0,\ 0 + -,\ - 0 +)$ constitute exactly hexagon in 3D space of Δ_{GC}^k, $k = 1, 2, 3$. Note that none of the vertexes coincide.

Coinciding triangles that appear in the vicinity of 50 % of genomic CG-content have a $(+ - +)$ symmetric pattern of Δ_{GC}^k signs. It obviously gives two identical phase triangles. Moreover, the pattern of coinciding vertexes (Figure 3*c*) reflects symmetric features of position-specific nucleotide frequencies with respect to complementary reversion: $p_i^1 \approx p_{\hat{i}}^3$, $p_i^2 \approx p_{\hat{i}}^2$, $i \in \{A,T,G,C\}$.

Similar features can be observed for genomes with 'parallel triangles' structure pattern. 'Parallel' location of triangles with F1–B1 'coincidence' reflects $p_i^3 \approx p_{\hat{i}}^3$ and $p_i^1 > p_{\hat{i}}^2$ $p_i^2 > p_{\hat{i}}^1$, $i \in \{A, G\}$ properties of codon usage in the corresponding genomes.

More complicated interrelations of position-specific frequencies determine the pattern of perpendicular triangles. Complementary reversion

symmetry in A and T frequencies $p_i^1 \approx p_{\hat{i}}^3$, $p_i^2 \approx p_{\hat{i}}^2$, $i \in \{A,T\}$ together with special asymmetry in G and C frequencies $\Delta_C^k = 0$, $\Delta_G^k \neq 0$, $k = 1, 2, 3$ produce distance matrix like that in Figure 3*b*.

A more detailed description of the symmetries in codon usage, which become apparent from the cluster structure, is available in (Gorban et al., 2005).

4. CONCLUSION

In this paper, we prove the universal seven-cluster structure in triplet distributions of bacterial genomes that reflects the main source of sequence heterogeneity. We showed the seven-cluster structure to be determined by a single parameter: genomic GC content. Based on the analysis of 143 completely sequenced bacterial genomes, we describe four 'pure' types of the structure and basic symmetries in triplet frequencies they reflect.