Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 353 (2005) 365–387
PHYSICA A
www.elsevier.com/locate/physa

Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences

Alexander Gorban[a,*], Tatyana Popova[b], Andrey Zinovyev[c]

[a]*Department of Mathematics, University of Leicester, Leicester, University Road, Leicester LE1 7RH, UK*
[b]*Institute of Computational Modelling, SB RAS, Krasnoyarsk, Russia*
[c]*Institut des Hautes Études Scientifiques, Bures-sur-Yvette and Bioinformatics Service of Institut Curie, Paris, France*

Received 6 December 2004; received in revised form 7 January 2005
Available online 25 February 2005

Abstract

Three results are presented. First, we prove the existence of a universal 7-cluster structure in all 143 completely sequenced bacterial genomes available in Genbank in August 2004, and explained its properties. The 7-cluster structure is responsible for the main part of sequence heterogeneity in bacterial genomes. In this sense, our 7 clusters is the basic model of bacterial genome sequence. We demonstrated that there are four basic "pure" types of this model, observed in nature: "parallel triangles", "perpendicular triangles", degenerated case and the flower-like type.

Second, we answered the question: how big are the position-specific information and the contribution connected with correlations between nucleotide. The accuracy of the mean-field (context-free) approximation is estimated for bacterial genomes.

We show that codon usage of bacterial genomes is a multi-linear function of their genomic G + C-content with high accuracy (more precisely, by two similar functions, one for eubacterial genomes and the other one for archaea). Description of these two codon-usage trajectories is the third result.

All 143 cluster animated 3D-scatters are collected in a database and is made available on our web-site: http://www.ihes.fr/~zinovyev/7clusters.

Keywords: Genome; Cluster; Codon usage; Correlations; Entropy; Mean field

*Corresponding author.
E-mail addresses: ag153@leicestes.ac.uk, ag153@le.ac.uk (A. Gorban), tanya@icm.krasn.ru (T. Popova), zinovyev@ihes.fr (A. Zinovyev).

0378-4371/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2005.01.043

1. Introduction

The bacterial genomes are compact genomes: most of the sequence contains coding information. Hence any statistical study of bacterial genomic sequence will detect coding information as the main source of heterogeneity (non-randomness). This is confirmed by mining sequences "from scratch", without use of any biological information, using entropic or Hidden Markov Modeling (HMM) statistical approaches (for examples, see [1–4]). All these methods can be seen as specific clustering of relatively short genomic fragments of length in the range 200–400bp comparable to the average length of a coding information piece.

Surprisingly, not much is known about the properties of the cluster structure itself, independently on the gene recognition problems, while it is implicitly used since long time ago (see, for example, early paper [5] about famous GENMARK gene-predictor, or Ref. [6] about GLIMMER approach). Only recently the structure was described explicitly. In Refs. [7–10] the structure was visualized in the 64-dimensional space of non-overlapping triplet distributions for several genomes. Also the same dataset was visualized in Refs. [11,12] using non-linear principal manifolds. In Ref. [13] several particular cases of this structure were observed in the context of the Z-curve methodology in the nine-dimensional space of Z-coordinates: it was claimed that the structure has interesting flower-like pattern but can be observed only for GC-rich genomes. This is somehow in contradiction with the results of Ref. [9], published before, where the same flower-like picture was demonstrated for AT-rich genome of *Helicobacter pylori*. This fact shows that this simple and basic structure is far from being completely understood and described.

The problem can be stated in the following way: there is a set of genomic fragments of length 100–1000 bp representing a genome almost uniformly. There are various ways to produce this set, for example, by sliding window with a given step of sliding (in this case sequence assembly is not generally needed), or it might be a full set of ORFs (in this case one needs to know the assembled sequence). We construct a distribution of points in a multidimensional space of statistics calculated on the fragments and study the cluster structure of this distribution. The following questions arise: what is the number of clusters? What is the character of their mutual locations? Is there a "thin structure" in the clusters? How the structure depends on the properties of genomic sequence, can we predict it?

Every fragment can be characterized by a "frequency dictionary" of short words (see examples Refs. [14–17]). For our purposes we use frequencies of non-overlapping triplets, counted from the first basepair of a fragment. Thus every fragment is a point in 64-dimensional space of triplet frequencies. This choice is not unique, moreover, we use dimension reduction techniques to simplify this description and take the essential features. The cluster structure we are going to describe is universal in the sense that it is observed in any bacterial genome and with any type of statistics which takes into account coding phaseshifts. The structure is basic in the sense that it is revealed in the analysis in the first place, serving as the principal source of sequence non-randomness.

In the series of papers [7–10,13] it was shown that even simple clustering methods like K-Means or Fuzzy K-Means give good results in application of the structure to gene-finding.

Recently, the idea of clustering was developed further by using the Kohonen self-organizing maps (SOM) [18] as a tool for presentation of the space of relative triplet frequencies. That paper aims to show how SOM can be used to automatically identify the major trends in oligonucleotide variation in a genome, and in doing so provide multiple gene models for use in gene prediction. Roughly speaking, one can detect the trends of cluster structure along genome.

The simplest 7-clusters based predictor uses one feature: which cluster the vector of triplet frequency belongs for the tested window. It is based on the simple statistics of triplets. The comparison of this predictor with GLIMMER [7] demonstrated that this one simple feature gives compatible results. The existence of the 7-cluster structure is, perhaps, the main reason, why the content-analysis gene finders work (at least, for bacterial genomes).

There is one essential difference between self-learning clustering approach and GLIMMER: in the clustering approach there is not necessary to deal with ORFs, one can still detect with unassembled genomic sequences or with sequences having a number of gaps.

The gene finding for bacteria is not so hard problem, as it is for human genome, for example, therefore it is desirable to achieve the high accuracy of predictions. Improving the accuracy of prediction of gene starts is one of a few remaining open problems in computer prediction of prokaryotic genes. Its difficulty is caused, in particular, by the absence of relatively strong sequence patterns identifying true translation initiation sites [19]. There are various ways to improve the situation, and one of them is development of the content-analysis self-training methods which can be used even for unassembled genomes.

One example of the observed 7-cluster structure is shown in Fig. 1. In short, this is a PCA plot of the point distribution. In Fig. 1 and further in this paper $F0$ stands for the (spatial) center of the group of "coding" fragments of F-type in which non-overlapping triplets have been red in the correct frame. The center is calculated as a simple mean point and it is a 64-dimensional vector. F1 and F2 correspond to the fragments where the triplets have been red with a frameshift (on one or two positions). Analogously, the B0, B1 and B2 labels stand for the centers of B-type fragment groups, where the triplets have been red with one of three possible frameshifts, respectively.

Every point on this plot presents a fragment of the genetic text, characterized by 64-dimensional vector. Principal component analysis allows to represent the 64-dimensional point distribution on a 2D-plane and, thus, visualize its cluster structure. Let us describe the basic properties of the structure. First, it consists of seven clusters.This fact is rather natural. Indeed, we clip fragments only from the forward strand and every fragment can contain (1) piece of coding region from the forward strand, with three possible shifts relatively to the first fragment position; (2) coding information from the backward strand, with three possible frameshifts; (3) non-coding region; (4) mix of coding and non-coding

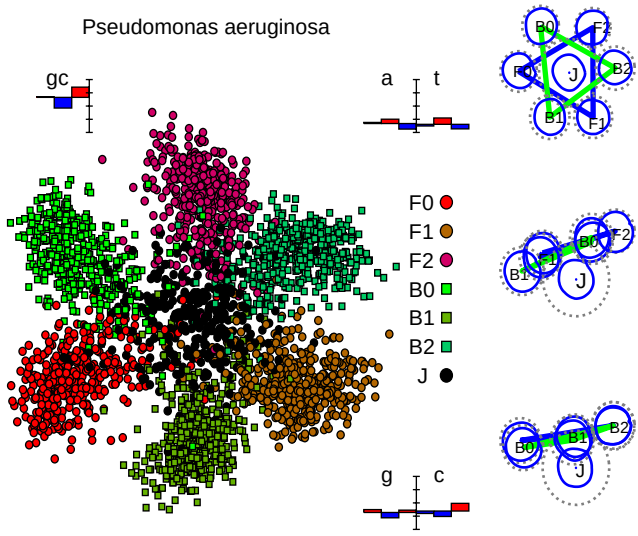

Fig. 1. Seven-cluster structure of *Pseudomonas aeruginosa* genomic sequence (G + C-content 67%). On the left pane the PCA plot of data distribution is shown. The colors specify a frameshift, black circles correspond to non-coding regions. On the right pane the structure is presented in a schematic way, in three projections (first and second principal components on the top, first and third in the middle, second and third in the bottom), with "radii" of the clusters schematically visualized. The diagrams show the codon position-specific nucleotide frequencies (right top and right bottom) as deviations from the average nucleotide frequency and codon position-specific G + C-content (left top).

information: these fragments introduce noise in our distribution, but their relative concentration is not high.

Second, the structure is well pronounced, the clusters are separated from each other with visible gaps. This means that most of learning (and even self-learning) techniques aiming at separation of the clusters from each other will work very well, which is the case for bacterial gene-finders that have performance more than 90% in most cases (for recent overview, see Ref. [20]).

Third, the structure is well represented by a 3D-plot (in this case it is even almost flat, i.e. 2D). Forth, it is indeed has symmetric and appealing flower-like pattern, hinting at there should be a symmetry in our statistics governing the pattern formation.

It is possible to guess the 7-cluster structure before data analysis. First of all, there are six possible frames + intergenic case. The selection control of mutations with respect to frameshift is strong in the coding regions, and they remain non-symmetric with respect to the frame shifts. But a small number of such mutations can symmetrize the codon usage with respect to one-position shifts in any non-coding fragment. Hence, they should form a symmetric cluster. But this guess cannot prove the existence of these clusters, it is necessary to find them in genomes, in unsupervising learning.

The 7-cluster structure includes two triangles of clusters for the coding part of genome: one triangle for the forward strand, another for the backward one. The

geometrical investigation of these triangles mutual positions in the codon usage space gives us a new statement of the classical *DNA asymmetry* problem studied in Refs. [21–25].

In this paper we show how the structure depends on very general properties of genomic sequence and show that it almost uniquely depends on a single parameter: the genomic G+C content. Also, based on analysis of 143 completely sequenced genomes, available in Genbank in August 2004, we describe four "pure" types of the structure observed in bacterial genomic sequences.

The outline of the paper is the following: first we introduce phaseshift and complementary reverse operators, helping to describe the structure. We show that in Nature the set of triplet distributions is almost one-dimensional (a line) for real eubacterial genomes, as well, as for archaeal genomes (another line). Then we analyze the position-specific information and the contribution connected with correlations between nucleotides. The accuracy of the mean-field (context-free) approximation is estimated for bacterial genomes. After that, we explain the properties of the 7-cluster structure and describe four "pure" types of the structures, observed for bacterial genomes. The paper is finalized by the description of the methods utilized and conclusion.

2. Phase-shifts in triplet distributions

Let us denote frequencies of non-overlapping triplets for a given fragment as f_{ijk}, where $i,j,k \in \{A,C,G,T\}$, such as f_{ACT}, for example, is a relative (normalized) frequency of ACT triplet.

One can introduce such natural operations over frequency distribution as *phase shifts* $\mathrm{P}^{(1)}$, $\mathrm{P}^{(2)}$ and *complementary reversion* C^{R}:

$$P^{(1)}f_{ijk} \equiv \sum_{l,m,n} f_{lij} f_{kmn}, \quad P^{(2)}f_{ijk} \equiv \sum_{l,m,n} f_{lmi} f_{jkn}\,,$$
$$\hat{f}_{ijk} \equiv C^R f_{ijk} \equiv f_{\hat{k}\hat{j}\hat{i}}\,, \tag{1}$$

where $\hat{i}$ is complementary to i, i.e. $\hat{A}=T, \hat{C}=G$, etc.

The phase-shift operator $P^{(n)}$ calculates a new triplet distribution, counted with a frame-shift on n positions, in the hypothesis that no correlations exist in the order of triplets in the initial phase. Complementary reversion constructs the distribution of codons from a coding region in the complementary strand, counted from the forward strand ("shadow" frequency distribution).

Phase-shift operators approximate the shifted triplet frequency as superposition of a phase-specific nucleotide frequency and a dinucleotide frequency. This can be better understood if we rewrite definitions (1) in the following way:

$$P^{(1)}f_{ijk} \equiv \sum_{l} f_{lij} \sum_{m,n} f_{kmn} \equiv d^{(right)}_{ij} p^{(1)}_{k}\,,$$
$$P^{(2)}f_{ijk} \equiv \sum_{l,m} f_{lmi} \sum_{n} f_{jkn} \equiv p^{(3)}_{i} d^{(left)}_{jk}\,. \tag{2}$$

We introduce the notion of *mean-field* (or *context free*) approximation of the triplet distributions in the following way:

$$(mf)_{ijk} = m_{ijk} = p_i^{(1)} p_j^{(2)} p_k^{(3)} ,$$
$$p_i^{(1)} = \sum_{jk} f_{ijk}, \quad p_j^{(2)} = \sum_{ik} f_{ijk}, \quad p_k^{(3)} = \sum_{ij} f_{ijk} , \tag{3}$$

i.e. the mean-field approximation is the distribution constructed from the initial triplet distribution neglecting all possible correlations in the order of nucleotides. The $p_i^{(k)}$ are the frequencies of the ith nucleotide ($i \in \{A, C, G, T\}$) at the kth position of a codon ($k = 1 \ldots 3$). In this way we model the 64 frequencies of the triplet distribution using only 12 frequencies of the three position-specific nucleotide distributions. This approximation is widely used in the literature (see, for example, Ref. [3]). All triplet distributions that can be represented in the form (3) belong to a 12-dimensional curved manifold $\mathbf{M}$, parametrized by 12 frequencies $p_i^{(k)}$. The manifold is embedded in the 64-dimensional space of all possible triplet distributions $\mathbf{T}$.[1]

It is easy to understand that any phase-shift for m_{ijk} only rotates the upper (position) indexes:

$$P^{(1)} m_{ijk} = p_i^{(2)} p_j^{(3)} p_k^{(1)} = m_{kij} ,$$
$$P^{(2)} m_{ijk} = (P^{(1)})^2 m_{ijk} = p_i^{(3)} p_j^{(1)} p_k^{(2)} = m_{jki} . \tag{4}$$

Also it is worth noticing that applying the $P^{(1)}$ (or $P^{(2)}$) operator several times to the initial triplet distribution we get the ($p_i^{(1)} p_j^{(2)} p_k^{(3)}$, $p_i^{(2)} p_j^{(3)} p_k^{(1)}$, $p_i^{(3)} p_j^{(1)} p_k^{(2)}$) triangle:

$$(P^{(1)})^3 f_{ijk} = m_{ijk} . \tag{5}$$

Operator $(P^{(1)})^3$ acts as a projector from the full 64-dimensional distribution space $\mathbf{T}$ onto the 12-dimensional manifold $\mathbf{M}$:

$$(P^{(1)})^3 : \mathbf{T} \to \mathbf{M} . \tag{6}$$

On the manifold $\mathbf{M}$ of all possible m_{ijk} we have $P^{(2)} = (P^{(1)})^2$, therefore, there are only two operators: phaseshift $P : Pm_{ijk} = m_{jki}$ and reversion $C : Cm_{ijk} = m_{\hat{k}\hat{j}\hat{i}}$. There are following basic equalities:

$$P^3 = 1, \quad C^2 = 1, \quad PCP = C . \tag{7}$$

Let us consider a point m on $\mathbf{M}$. It corresponds to a set of 12 phase-specific nucleotide frequencies $p_i^{(1)}$, $p_i^{(2)}$ and $p_i^{(3)}$, $i \in \{A, C, G, T\}$. Applying operators P and C in all possible combinations we obtain *an orbit* on $\mathbf{M}$, consisting of 6 points: m, Pm, P^2m, Cm, PCm, P^2Cm. Theoretically, some points can coincide, but only in such a way that the resulting orbit is to consist of 1 (fully degenerated case), 3 (partially

[1]The normalization equality $\sum_{ijk} f_{ijk} = 1$ makes all distributions to form a standard 63-dimensional simplex in R^{64}. For $\mathbf{M}$ one has three independent normalizations: $\sum_i p_i^{(k)} = 1$, $k = 1 \ldots 3$, these equalities distinguish a nine-dimensional set (image of the product of three three-dimensional standard simplexes) in $\mathbf{M}$, where all normalized distributions are located.

degenerated case) or 6 (non-degenerated case) points. The fully degenerated case corresponds to the triplet distribution with the highest possible entropy among all distributions with the same nucleotide composition:

$$f_{ijk} = p_i p_j p_k, \quad p_i = \sum_{mn} \frac{(f_{imn} + f_{mni} + f_{nim})}{3} . \tag{8}$$

This distribution ("completely random") is described by four nucleotide frequencies, with any information about position in the triplet lost. In our **T** space it is a three-dimensional (due to normalization equality) simplex on **M**. For bacterial genomes this distribution can serve as an approximate (zero-order accuracy) model for triplet composition in non-coding regions.

Let us use the introduced notions for the statement of the compositional strand asymmetry problem.[2] The operator C is a linear involution in the codon usage space (because $C^2 = 1$). It has a subspace of fixed points (symmetric distributions m_{ijk} that satisfy the identity $m_{ijk} = m_{\hat{k}\hat{j}\hat{i}}$), it is the eigenspace for eigennumber 1, and the eigensubspace for eigennumber -1 ($m_{ijk} = -m_{\hat{k}\hat{j}\hat{i}}$). Each vector of codon usage m_{ijk} can be decomposed into two parts: symmetric and antisymmetric: $m_{ijk} = m^S_{ijk} + m^A_{ijk}$, $m^S_{ijk} = \frac{1}{2}(m_{ijk} + m_{\hat{k}\hat{j}\hat{i}})$, $m^A_{ijk} = \frac{1}{2}(m_{ijk} - m_{\hat{k}\hat{j}\hat{i}})$. We numerate clusters and their centers F_i, B_i in the reading order of the forward strand. Hence, $F_0 \approx CB_0$, $F_1 \approx CB_2$, $F_2 \approx CB_1$. If forward and backward DNA strands have the same codon composition (in the coding regions), then these equalities become exact. If these compositions are different, then distances between symmetric parts, $\Delta^S = \|F^S_0 - B^S_0\|$, and sums of antisymmetric parts, $\Delta^A = \|F^A_0 + B^A_0\|$, give us the information about the internal structure of asymmetry $\Delta = \|F_0 - B_0\|$. The next step of asymmetry investigation can include analysis of all coordinates of the differences $F_0 - CB_0$. So, the mutual positions of the triangles F_i ($i = 0, 1, 2$) and CB_i ($i = 0, 1, 2$) give rich information for studying DNA asymmetry. In the examples (Fig. 2) we can see the positions of triangles F_i ($i = 0, 1, 2$) and CB_i ($i = 0, 1, 2$). We can mention that in these examples the distances between symmetric parts, Δ^S are close to the sums of antisymmetric parts Δ^A.

3. Information content in the triplet distributions

In this section, we study the information content of the triplet distributions in the coding part of genetic texts, and all the frequencies are computed for the coding part. The questions are: what are the contributions to the total amount of information of the triplet distribution, how significant are: (i) the position-specific information, (ii) the contribution connected with correlations between nucleotides and so on? For

[2]Here we discuss only the problem of finding and presentation of DNA asymmetry. The biological problem is wider, and its consideration includes analysis of the mutation process and of the natural selection pressure, see Refs. [21–25]. The local and systematic deviations from the $C = G$ rule were discussed in Ref. [26]. The question is whether these deviations are a consequence of an underlying bias in mutation or selection (or both).

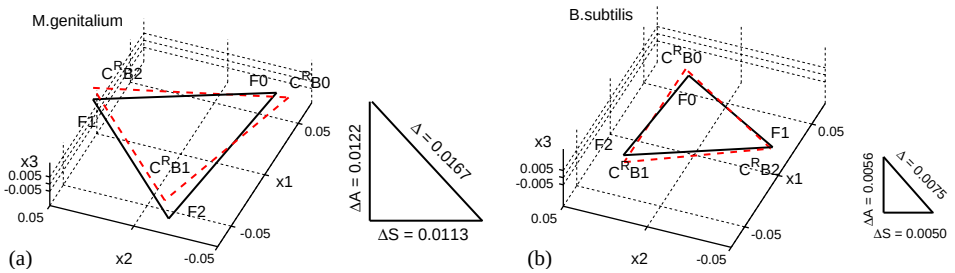

Fig. 2. MDS (multidimensional scaling) plots visualizing mutual positions of $F_0 - F_1 - F_2$ triangle and the triangle obtained after applying complementary reversion operation to the $B_0 - B_1 - B_2$ triangle. Note, that if the two DNA strands would be identical in the nucleotide composition, the triangles would coincide. The difference is less in *B. subtilis* genome and indeed it is known to be symmetric (see PR2-plots in Ref. [27]). The $\Delta - \Delta^S - \Delta^A$ diagram shows the contribution of symmetric (Δ^S) and antisymmetric (Δ^A) parts in the $\|F_0 - C^R B_0\|$ distance. In both cases the contributions are approximately equal.

this purpose we study the mean-field (or context free) approximation of the triplet distributions (3).

Coding information from windows in the forward strand has one of three possible phase shifts. Since this phase shift is not known in advance, approximately one-third of the windows fall into the vicinity of the point that corresponds to the f_{ijk} (0-shift), one-third are close to the $f^{(1)}_{ijk}$ (1-shift), and the last are close to the $f^{(2)}_{ijk}$ (2-shift). This is also true for the backward strand, but with the centers corresponding to complementary distributions.

Let us consider also the averaged three-phase distribution:

$$f^{av}_{ijk} = \frac{1}{3}(f_{ijk} + f^{(1)}_{ijk} + f^{(2)}_{ijk}) .$$

In the f^{av}_{ijk} distribution all position-specific information is eliminated but it still contains some information about the correlations in the order of nucleotides.

One can measure the distance between two distributions g_{ijk} and h_{ijk} as the relative information of the distribution g_{ijk} with respect to h_{ijk} using the Kullback distance [28]:

$$D(g_{ijk}, h_{ijk}) = \sum_{ijk} g_{ijk} \ln \frac{g_{ijk}}{h_{ijk}} .$$

For our purposes we will use a symmetrized version of the Kullback distance

$$D^{SYM}(g_{ijk}, h_{ijk}) = \tfrac{1}{2}(D(g_{ijk}, h_{ijk}) + D(h_{ijk}, g_{ijk})) .$$

To visualize the structure of pair-wise distances between different distributions, we use classical metric multidimensional scaling (MDS) (for reference, see, for example, Ref. [29]). The idea of the MDS method is to put the points onto the 2D plane in such a way that to preserve the structure of the pair-wise distances between the points, given by a distance matrix. The resulting pictures are given in Fig. 3. The axes of the MDS plot correspond to fictive "principal" coordinates that are assigned to the points to preserve the distances between them. Since shift and rotation of the scatters do not change the distances, we use such a shift that the m point (the smallest entropy) is in

the (0,0) point of the plot and the rotation angle such that the f_{ijk} (the f point on the plot) is on the negative side of the x-axis.

We connect points f_{ijk}, $f^{(1)}_{ijk}$, $f^{(2)}_{ijk}$ by solid line. It is the "three-phases" triangle, corresponding to the real triplet distributions in the correct, first and second phases respectively. The second, dashed triangle connects the points of the mean-field approximation $(mf_{ijk}, P^{(1)}mf_{ijk}, P^{(2)}mf_{ijk})$.

Let us discuss general features of the pictures. Qualitatively, the information content (relative entropy) of a point on the plots in Fig. 3 is proportional to the

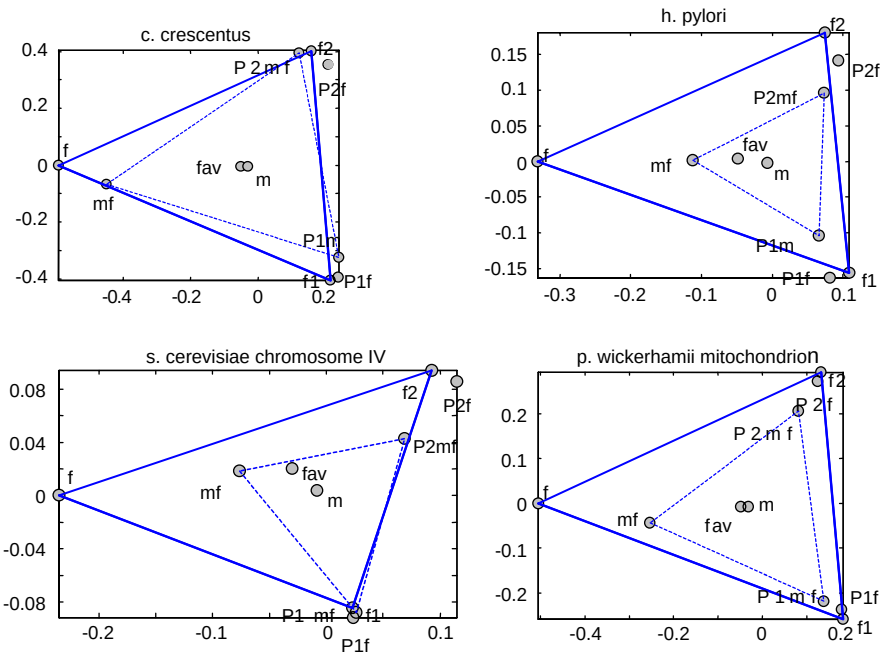

Fig. 3. MDS plots representing the structure of Kullback distances between different distributions. The solid triangle is the "three-phase" triangle, calculated from the real gene sequences. Dashed triangle is the corresponding "mean-field" (context free) approximation.

$f = f_{ijk}$—real triplet distribution in the correct phase;

$f1 = f^{(1)}_{ijk}$—real triplet distribution in the first phase;

$f2 = f^{(2)}_{ijk}$—real triplet distribution in the second phase;

$P1f = P^{(1)}f_{ijk}$—calculated distribution in the first phase;

$P2f = P^{(2)}f_{ijk}$—calculated distribution in the second phase;

$fav = \frac{1}{3}(f_{ijk} + f^{(1)}_{ijk} + f^{(2)}_{ijk})$— average distribution of triplets;

$mf = p^{(1)}_i p^{(2)}_j p^{(3)}_k$—the mean-field (context free) approximation of the codon usage;

$p^{(k)}_i$—are the frequencies of the ith nucleotide ($i \in \{A, C, G, T\}$) at the kth position of codon ($k = 1 \ldots 3$);

$P1mf = P^{(1)}mf = p^{(2)}_i p^{(3)}_j p^{(1)}_k$— mean-field approximation in the first (shifted) phase;

$P2mf = P^{(2)}mf = p^{(3)}_i p^{(1)}_j p^{(2)}_k$— mean-field approximation in the second (shifted) phase;

$m = p_i p_j p_k$—randomized distribution (the highest entropy).

distance from the center of plot (0,0). The maximum of information is contained, of course, in the f_{ijk} distribution (the f point), which is the most distant point on the plots. For example, in the case of *H.pylori*, the relative information of the triplet distribution equals 0.29. The value is higher in the case of *C.crescentus* (0.39) and less in the case of *S.cerevisiae* genome (0.18). In fact, high information content of the triplet distribution in the correct phase gives more contrast cluster structure and better quality of unsupervised gene recognition.

The distances $D^{SYM}(f_{ijk},f^{(1)}_{ijk})$ and $D^{SYM}(f_{ijk},f^{(2)}_{ijk})$ are approximately equal (0.46 and 0.44, for *H.pylori*) and bigger than the distance between $f^{(1)}_{ijk}$ and $f^{(2)}_{ijk}$ (0.32 for *H.pylori*). This can be explained if the correlations in the order of codons in the coding sequences are small (our study shows that this is the case for, at least, bacterial and yeast genomes). In this case, the distributions in the first and second phases can be reconstructed from the f_{ijk} using only position-specific frequencies of nucleotides and di-nucleotides. Indeed, the information contents of $f^{(1)}_{ijk}$ and $f^{(2)}_{ijk}$ are less than in f_{ijk}. (0.19 and 0.19 against 0.29, for *H.pylori*).

Shifted distributions are reconstructed from the initial distribution, applying phase-shifts operators $P^{(1)}$ and $P^{(2)}$. In all cases these reconstructions (points $P1f$ and $P2f$), calculated using assumption about smallness of correlations in the order of codons, are very close to the real distributions in the first and second phases (points $f1$ and $f2$ on the plots).

The ''mean-field approximation'' triangle is isosceles with its center approximately in the m_{ijk} point. The difference in sizes of the ''three-phases'' triangle and the ''mean-field approximation'' triangle reflects presence of correlations in the order of nucleotides. In Fig. 3, this difference is small in the case of *C.crescentus* and considerable in other three genomes. For example, in the case of *H.pylori,* the average length of the ''three-phases'' triangle side is 0.41 while the same value for the ''mean-field approximation'' triangle is only 0.16. The loss of information after neglecting all correlations in the order of nucleotides (the distance from f to mf points) is 0.21 in the case of *H.pylori* and 0.15 in the case *C. crescentus*.

Existence of the universal cluster structure does not depend on the specific codon usage. For the coding part it depends on *nontriviality* of codon usage (in coding regions): the triplet distribution f_{ijk} in the coding phase should be sufficiently far from the randomized distribution $m = p_i p_j p_k$ (see Fig. 3). The lengths of the cluster triangle side for bacterial genomes is proportional to the distance between these two distributions. The separability of these clusters from the non-coding one is determined by this distance, as well as by genome compactness: separation becomes difficult for genomes with large non-coding part.

4. One-dimensional model of codon usage

Let us consider triplet distributions corresponding to the codon usage of bacterial genomes, i.e. the subset of *naturally occurred triplet distributions*. It was found that they can be approximated by their mean-field distributions, i.e. they are located close to **M**. Moreover, in this section we show that in nature, for 143 completely sequenced

bacterial genomes, the m_{ijk} distributions are tightly located along one-dimensional curve on **M**. The curve can be parametrized by the genomic G + C-content.

Twelve dependencies $p_i^{(1)}(GC)$, $p_i^{(2)}(GC)$, and $p_i^{(3)}(GC)$, $i \in \{A, C, G, T\}$, where GC is genomic G + C-content, are presented in Fig. 4(a–d) for 143 fully sequenced bacterial genomes available in Genbank in August, 2004. These dependencies are almost linear. This fact, despite its simplicity, was not explicitly demonstrated before. The numerous results on the structure of codon usage described previously in literature (see, for example, Refs. [30–32]) are in agreement with this picture.

For our purposes, it is important to notice that for the genomes with G + C-content higher than ~60% there is the same well-defined structure in their codon usage: the G + C-content in the first codon position is close to the average of three values, the second is lower than the average and the third is essentially higher than the average. This pattern can be denoted in the form of simple *GC-signature*: $0 - +$. In the next section we develop more complicated signature to classify 7-cluster structures, corresponding to the orbits, generated by P and C operators in a set of genomic fragments of 300–400bp length.

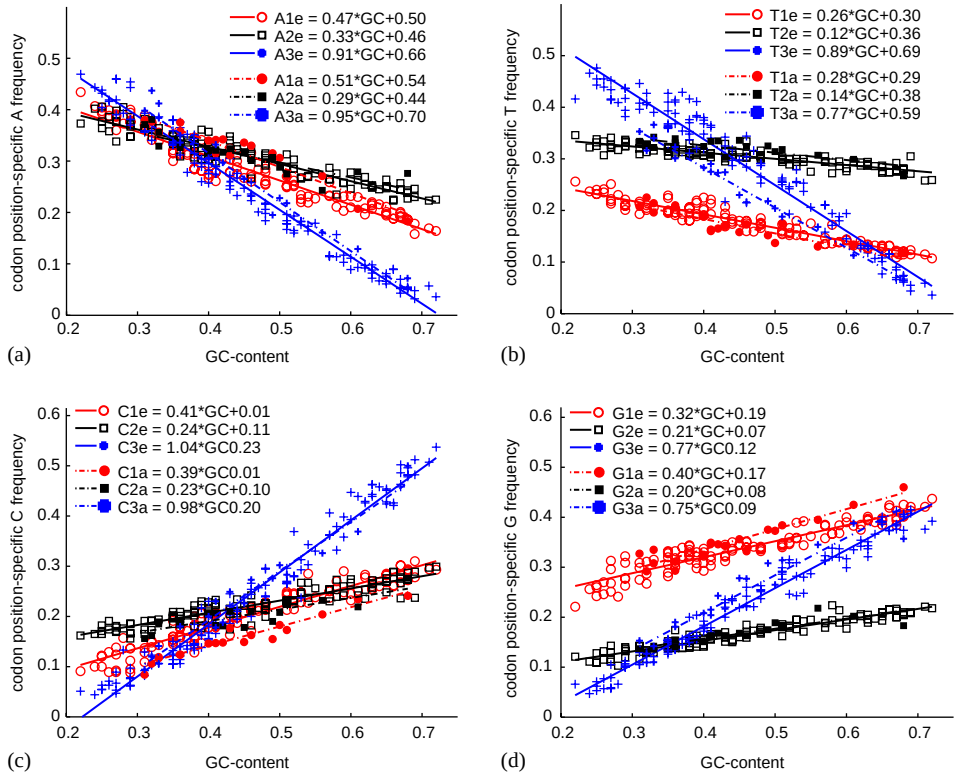

Fig. 4. Codon position-specific nucleotide frequencies (a–d) and codon position-specific GC-content (e). Solid line and empty points correspond to 124 completed eubacterial genomes, broken line and filled points correspond to 19 completed archaeal genomes.

The linear functions, describing codon usage, are slightly different for eubacteria and archaea genomes. Significant differences are observed for $p_A^{(1)}$, $p_C^{(1)}$, $p_G^{(3)}$, $p_G^{(1)}$, $p_T^{(3)}$ functions. For the others, the dependencies are statistically indistinguishable.

Fig. 5 demonstrates that the codon position-specific G+C-content is a linear function of genomic G+C-content, in each position. In Ref. [13] analogous results were shown for 33 completed genomes. For these dependencies the differences between eubacteria and archaea genomes are nonsignificant.

One important conclusion follows from Fig. 4: if we take the set of triplet frequencies, occurred in nature and corresponding to the codon usage of bacterial genomes, then in the 12-dimensional space of codon position-specific nucleotide frequencies this set appears almost as a straight line (more precisely, two close lines, one for eubacteria and the other for archaea). If we look at this picture from the 64-dimensional space of triplet frequencies **T**, then one sees that the distributions are located close to the curved **M** manifold of the mean-field approximations, embedded in the space. As a result, when analyzing the structure of the distribution of bacterial codon usage, one detects that the points are located along two curves. These curves are closer at their AT-rich ends and diverge at GC-rich ends. Moving along these curves one meets all bacterial genomes. Genomes with close G+C-content generally have close codon usage. Many evidences for this structure were reported in studies on multivariate analysis of bacterial codon usage (for example, see Fig. 6 from Ref. [33]), but here the general structure is presented first time in explicit and formal way.

This conclusion is consistent with previous studies: many properties of the codon usage are correlated with genomic G+C-content. For example, the strong correlation between the amino-acid composition and genomic G+C content was proved in Ref. [34] for 59 bacterial species.

The codon usage bias has been widely reported to correlate with G+C composition, and recently the quantitative regression between codon usage bias

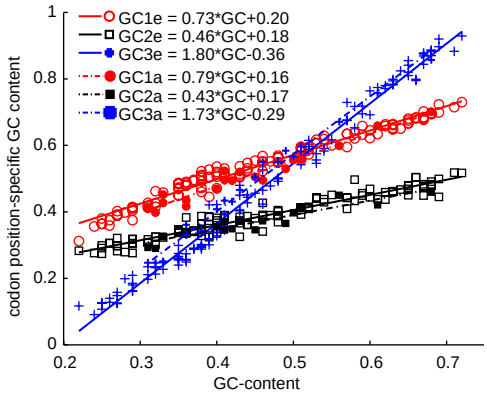

Fig. 5. Codon position-specific GC-content. Solid line and empty points correspond to 124 completed eubacterial genomes, broken line and filled points correspond to 19 completed archaeal genomes.

and GC3 (G + C-content in the third position) was published [35]. The regression equations are based on 70 eubacterial and 16 archaeal genomes.

Some fragments of observed correlations can be found in the Sueoka's neutrality plots [36,37]. A theory of directional mutation pressure was proposed in 1962 [38]. It explained the wide variation of DNA base composition observed among different bacteria and its small heterogeneity within individual bacterial species. The theory was based on the assumption that the effect of mutation on a genome has a directionality toward higher or lower G + C content of DNA, and this pressure generates directional changes more in neutral parts of the genome than in functionally significant parts.

For analysis of codon bias evolution a population genetic model is developed taking into account population size and selective differences between synonymous codons [39].

Following the Sueoka theory, in regression analysis one mostly uses GC3 (the G + C content in the third position), and not the overall genomic G + C content. The reason is that GC1 and GC2 are under strong selection control, while for GC3 and G + C content of intergenic regions this control is much weaker, and the overall genomic G + C content is the linear combination of these quantities. Sometimes, the difference between the GC3 and overall genomic G + C content as reference variable might be significant, but in our case, as it is presented on Fig. 5, the correlation between GC3 and genomic G + C content is strong, one of them is practically a linear function of the other, and both of them can serve as reference variables with the same success.

It is necessary to mention that now we do not know any theory that give a solid explanation of the observed accuracy and linearity of the dependencies, presented on Figs. 4 and 5. Why the bacterial genomes form a straight line in the nine-dimensional space of codon-position specific nucleotide frequencies? We do not know.

It seems natural to apply "mutation + selection" arguments and models [38,36,21,39]. Such models are in good agreement with some data of codon usage [37,24,25,40] (and some quantitative discrepancies [39] and doubts are also reported, even for the problem of genetic code optimality [41]). The problem is that we need to prove the models on another material, it is desirable to verify them independently, by direct measurements. It should be proven that the mutation + selection processes have kinetic constants that could provide such an accuracy, and that the whole process keeps all genomes near the observed straight line. And now we can only ask again, for the new data [42]: Codon usage: mutational bias, translational selection, or both? (Or something else?)

5. Properties and types of the 7-cluster structure

In the paper [13] the authors claim that the codon position-specific nucleotide frequencies (represented as Z-coordinates) in GC-rich genomes show flower-like cluster structure, and the phenomenon is not observed in other genomes. Here we explain the phenomenon and demonstrate other types of structures observed in

genomes. The type of the structure is related to the pattern of symmetric properties of codon usage.

First of all, we point out to the fact that the space used in Ref. [13] is a specific projection from 64-dimensional space of triplet frequencies. The phenomenon can also be observed in 64-dimensional space and in 12-dimensional space of codon position-specific nucleotide frequencies.

Let us consider the context free approximation of codon usage introduced above:

$$m_{ijk} = p_i^{(1)} p_j^{(2)} p_k^{(3)} \tag{9}$$

and consider 3D space with the following coordinates:

$$x = p_G^{(1)} + p_C^{(1)} - f_{GC}, \quad y = p_G^{(2)} + p_C^{(2)} - f_{GC}, \quad z = p_G^{(3)} + p_C^{(3)} - f_{GC}\,. \tag{10}$$

In fact, x, y and z are deviations of GC-content in the first, second and the third position from average GC-content f_{GC} of coding regions. In all GC-rich genomes (starting from $f_{GC} > 60\%$) their codon usage context-free approximation has the following structure (see Fig. 42e): $x \approx 0, y < 0, z > 0$. We can denote this pattern as $0-+$. Applying phaseshift and reverse operators defined above (notice that C operator does not change G + C-content, it only reverses the signature), we obtain the following orbit: $\{0-+, -+0, +0-\}$ and $\{+-0, -0+, 0+-\}$. If we consider now a 3D grid consisting of 27 nodes as shown in Fig. 6, corresponding to all possible patterns (GC-signatures), then it is easy to understand that the orbit corresponds to the points where the grid is cross-sectioned by a plane, coming through the 000 point perpendicular to the $\{---, +++\}$ diagonal. It is well known fact that in this situation the form of the intersection is a regular hexagon. The 000 point in our picture corresponds to the center of the non-coding cluster (this is the fully degenerated distribution described above), where all phases have been

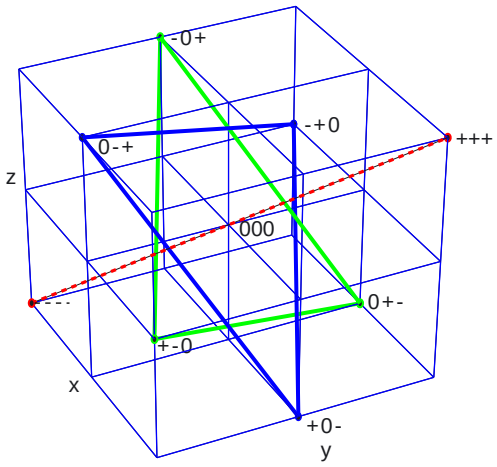

Fig. 6. Model of the flower-like cluster structure. Broken line corresponds to the direction of the fastest G + C-content increase.

mixed. The $\{---,+++\}$ diagonal corresponds to the direction of the fastest G+C-content increase. Hence, this model explains the following features of the flower-like structure observed in GC-rich ($G+C>60\%$) genomes:

(1) In the 64-dimensional space the centers of clusters are situated close to a distinguished 2D-plane, forming regular hexagonal structure.

(2) The third principal component (perpendicular to the cluster plane) is the direction of G+C-content increase (i.e. the gradient of G+C-content linear function, defined in the 64-dimensional triplet space).

In most flower-like structures the cluster that corresponds to the non-coding regions is slightly displaced in the direction perpendicular to the main cluster plane. This happens because G+C-content of non-coding regions is generally slightly lower than of coding regions.

Now let us consider general case of genome with any given genomic G+C-content. The type of the 7-cluster structure depends on values of 12 functions $p_i^{(1)}, p_i^{(2)}, p_i^{(3)}$, $i \in \{A, C, G, T\}$. Applying phaseshift and reverse operators, one obtains an orbit which serves as a skeleton of the cluster structure. The orbit structure reflects symmetries in the set of values of these 12 functions with respect to the P and C operators.

We describe these symmetries in the following simplified manner. Let us order the 12 values in the form of 6×2 table:

$$s_{ij} = \begin{matrix} p_A^{(1)} p_A^{(2)} p_A^{(3)} p_T^{(1)} p_T^{(2)} p_T^{(3)} \\ p_G^{(1)} p_G^{(2)} p_G^{(3)} p_C^{(1)} p_C^{(2)} p_C^{(3)} \end{matrix}. \tag{11}$$

Then the reverse operator C simply reads the table from right to the left:

$$Cs_{ij} = \begin{matrix} p_T^{(3)} p_T^{(2)} p_T^{(1)} p_A^{(3)} p_A^{(2)} p_A^{(1)} \\ p_C^{(3)} p_C^{(2)} p_C^{(1)} p_G^{(3)} p_G^{(2)} p_G^{(1)} \end{matrix}. \tag{12}$$

The phaseshift operator P rotates the values in the table by threes, for every letter:

$$Ps_{ij} = \begin{matrix} p_A^{(2)} p_A^{(3)} p_A^{(1)} p_T^{(2)} p_T^{(3)} p_T^{(1)} \\ p_G^{(2)} p_G^{(3)} p_G^{(1)} p_C^{(2)} p_C^{(3)} p_C^{(1)} \end{matrix}. \tag{13}$$

We reduce the description of s in the following way: every entry in the table is substituted by "+", "−" and "0", if the corresponding value is bigger then the average over the same letter frequencies, smaller or in the $[average - 0.01; average + 0.01]$ interval respectively. For example, for a set of frequencies $p_A^{(1)} = 0.3, p_A^{(2)} = 0.5$, $p_A^{(3)} = 0.401$, we substitute $p_A^{(1)} \to -, p_A^{(2)} \to +, p_A^{(3)} \to 0$. We call "signature" the new table $\hat{s}_{ij}$ with reduced description.

Using linear formulas for eubacteria from Fig. 4(a–d) and calculating the $\hat{s}_{ij}$ tables for the range $[0.2; 0.8]$ of G+C-content, we obtain 19 possible signatures in the intervals of genomic G+C-content, listed in Table 1.

There are 67 different signatures observed for really occurred $p_i^{(k)}$-values for 143 genomes considered in this work (see our web-site [43] with supplementary

materials). Most of them differ from the signature in Table 1 with corresponding G + C value only by changing one of the "+" or "−" for "0" or vise versa.

From Table 1 one can see that the only conserved positions, independent on genomic G + C-content for the interval [0.20; 0.80] are $p_T^{(1)}$ (always "−"), $p_G^{(1)}$ (always "+"), $p_G^{(2)}$ (always "−"). This holds true also for all really observed signatures. This observation confirms already known "invariants" of codon usage described in Refs. [30–32].

Let us look at several typical examples. All genomes with genomic G + C-content higher then 60% have the following genomic signature:

$$\hat{s}_{ij}(GC > 60\%) = \begin{matrix} ++{-}{-}+{-} \\ +{-}+{-}{-}+ \end{matrix}. \tag{14}$$

This signature uniformly reflects the previously mentioned *GC*-signature ("0 − +"): pairs $p_A^{(1)}$, $p_T^{(1)}$ and $p_G^{(1)}$, $p_C^{(1)}$ compensate the signs of each other to give "0" in the first position of *GC*-signature, while in the second position we have "+" for A and T and "−" for G and C, and vice versa for the third position. As a result, we obtain the flower-like structure. In Fig. 7 we visualize the orbit for *Streptomyces coelicolor*,

Table 1
Nineteen possible signatures for one-dimensional codon usage model

$\begin{matrix}--+{-}{-}+\\+{-}{-}++{-}\end{matrix}$ [0.200; 0.255)	$\begin{matrix}000{-}++\\+{-}{-}0+{-}\end{matrix}$ [0.331; 0.373)	$\begin{matrix}0+{-}{-}++\\+{-}{-}{-}0+\end{matrix}$ [0.434; 0.482)
$\begin{matrix}--+{-}{-}+\\+{-}{-}0+{-}\end{matrix}$ [0.255; 0.265)	$\begin{matrix}0+0{-}++\\+{-}{-}0+{-}\end{matrix}$ [0.373; 0.385)	$\begin{matrix}0+{-}{-}++\\+{-}0{-}0+\end{matrix}$ [0.482; 0.487)
$\begin{matrix}--+{-}0+\\+{-}{-}0+{-}\end{matrix}$ [0.265; 0.289)	$\begin{matrix}0+{-}{-}++\\+{-}{-}0+{-}\end{matrix}$ [0.385; 0.388)	$\begin{matrix}0+{-}{-}++\\+{-}0{-}{-}+\end{matrix}$ [0.487; 0.502)
$\begin{matrix}-0+{-}0+\\+{-}{-}0+{-}\end{matrix}$ [0.289; 0.316)	$\begin{matrix}0+{-}{-}++\\+{-}{-}{-}+{-}\end{matrix}$ [0.388; 0.391)	$\begin{matrix}0+{-}{-}+0\\+{-}0{-}{-}+\end{matrix}$ [0.502; 0.515)
$\begin{matrix}00+{-}0+\\+{-}{-}0+{-}\end{matrix}$ [0.316; 0.326)	$\begin{matrix}0+{-}{-}++\\+{-}{-}{-}+0\end{matrix}$ [0.391; 0.424)	$\begin{matrix}++{-}{-}+0\\+{-}0{-}{-}+\end{matrix}$ [0.515; 0.542)
$\begin{matrix}000{-}0+\\+{-}{-}0+{-}\end{matrix}$ [0.326; 0.331)	$\begin{matrix}0+{-}{-}++\\+{-}{-}{-}00\end{matrix}$ [0.424; 0.434)	$\begin{matrix}++{-}{-}+0\\+{-}+{-}{-}+\end{matrix}$ [0.542; 0.545)
		$\begin{matrix}++{-}{-}+{-}\\+{-}+{-}{-}+\end{matrix}$ [0.545; 0.800)

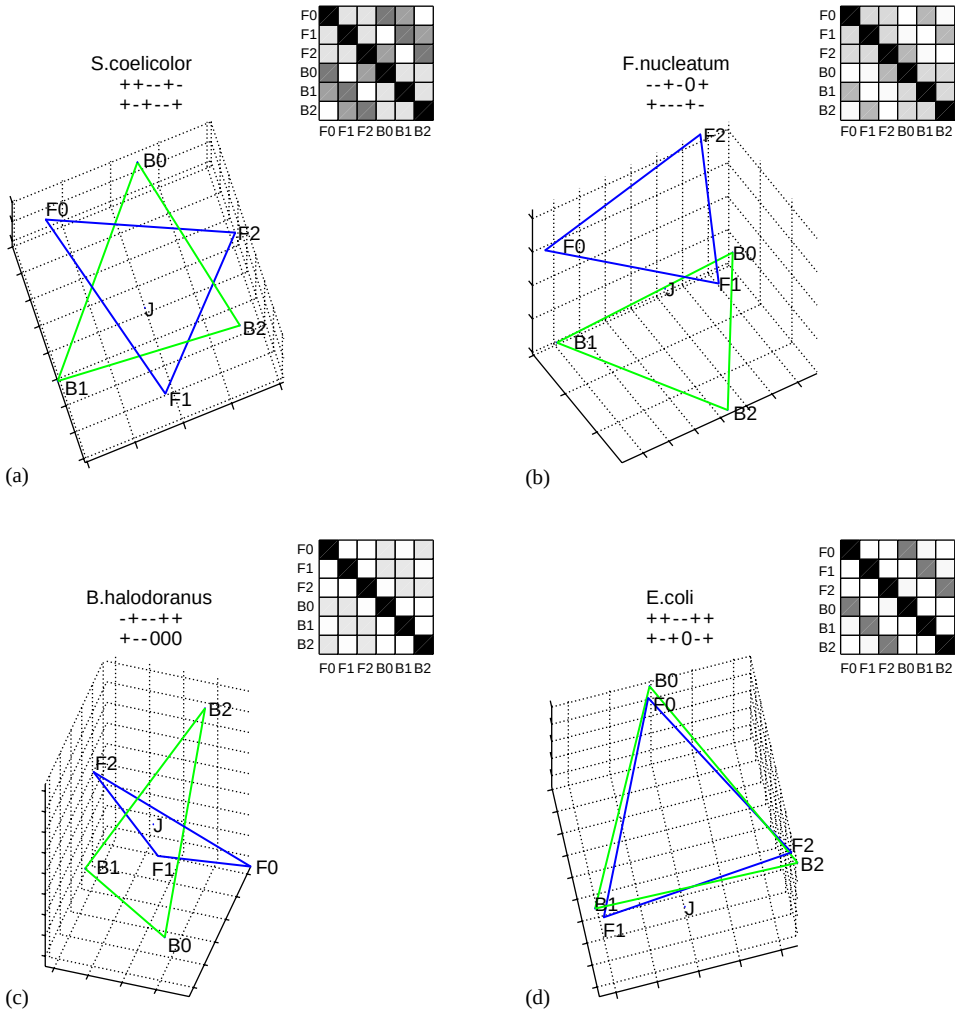

Fig. 7. Four typical examples of the 7-cluster structure: (a) genome of *S. coelicolor* (GC = 72%), flower-like structure; (b) genome of *F. nucleatum* (GC = 27%), "parallel triangles"; (c) *B. halodurans* (GC = 44%), "perpendicular triangles"; (d) *E. coli* (GC = 51%), degenerated case.

genome with high G + C-content: 72%. Together with the orbit we visualize the distance matrix, where the distances are calculated in the full 64-dimensional triplet frequency space **T**. Black color on the plot corresponds to zero distance, white corresponds to the biggest value in the matrix. The most informative 3×3 block of the matrix is in the left bottom corner (or top right, by symmetry): it describes mutual distances between the vertices of two triangles. The left top and right bottom 3×3 blocks contain equal values, since the sides of the triangles have the same length.

Our second example is genome of *Fusobacterium nucleatum* (AT-rich genome, G + C-content is 27%), Fig. 7b. The signature is

$$\hat{s}_{ij}(F.nucleatum) = \begin{matrix} --+-0+ \\ +---+- \end{matrix}. \tag{15}$$

This pattern, commonly observed in AT-rich genomes, can be called "parallel triangles". Notice that two parallel triangles are rotated with respect to their corresponding phase labels: the F0 vertex is located in front of the B1 vertex.

The third example is genome of *Bacillus halodurans* (G + C-content is 44%):

$$\hat{s}_{ij}(B.halodurans) = \begin{matrix} -+--++ \\ +--000 \end{matrix}. \tag{16}$$

We refer to this pattern as "perpendicular triangles". Another example of the pattern is genome of *Bacillus subtilis*. All non-diagonal distances in the distance matrix have in this case approximately the same value. This structure can be easily recognised from its signature: the second row has three zeros while the first one is almost *palindromic*. As we will see in the next example, palindromic rows in the signature (or such that can be made palindromic applying the phase-shift P operator) make zero contribution to the diagonal of the "inter-triangle" part of the distance matrix. This is easy to understand, because the reverse operator C reads the signature from right to the left. The rows with three zeros in different phase positions (when, for example, the phase specific nucleotide frequencies for one letter are equal to their average, as happened in this case) give approximately equal contribution to every value in the "inter-triangle" part of the distance matrix. The resulting matrix corresponds to the "perpendicular triangles" pattern. We should notice that the distance matrix showed on Fig. 7c cannot be effectively represented as a distribution of six points in 3D. Thus the "perpendicular triangles" structure shown on Fig. 7c is only an approximate picture, the real configuration at least four-dimensional, due to the distance matrix symmetry.

In the region of G + C-content about 51% we observe a group of genomes with almost palindromic signatures. One typical example is the genome of *Escherichia coli*:

$$\hat{s}_{ij}(E.coli) = \begin{matrix} ++--++ \\ +-+0-+ \end{matrix}. \tag{17}$$

The resulting pattern is a degenerated case: two triangles are co-located, without phase label rotation (F0 is approximately in the same point as B0). The distance matrix consists of 4 almost identical 3×3 blocks. As a result, we have situation, when 7-cluster structure effectively consists of only four clusters, one for every pair F0-B0, F1-B1, F2-B2 and a non-coding cluster.

The same degenerated case but with rotation of labels (F0-B1,F1-B2,F2-B0) is observed for some AT-rich genomes. For example, for the genome of *Wigglesworthia*

brevipalpis (G + C-content equals 22%) the signature

$$\hat{s}_{ij}(W.brevipalpis) = \begin{matrix} 0-+-0+ \\ +---+- \end{matrix} \tag{18}$$

becomes a perfect palindrom after applying the phaseshift operator:

$$P\hat{s}_{ij}(W.brevipalpis) = \begin{matrix} -+00+- \\ --++-- \end{matrix}. \tag{19}$$

One possible biological consequence (and even explanation) of this degeneracy is existence of overlapping genes: in this case the same codons can be used to encode proteins simultaneously in the forward and backward strands on a regular basis (without frameshift for G + C-content around 50% and with a frameshift for AT-rich genomes), with the same codon usage.

The four patterns are typical for triplet distributions of bacterial genomes observed in nature. The other ones combine features from these four "pure" types. In general, going along the G + C-content scale, we meet first "parallel triangles" which will transform gradually to "perpendicular triangles". On this way one can even meet structures resembling flower-like type in one of the 2D-projections, like for the genome of *Helicobacter pylori* (see our web-site [43] and in [9] for the illustration). Then the pattern goes to the degenerated case with genomic G + C-content around 50% and signatures close to palindromic. After the degeneracy disappears, the pairs F0-B0, F1-B1, F2-B2 diverge in the same 2D-plane and after 55% threshold in G + C-content we almost exclusively have the flower-like structures. It is possible to browse the animated scatters of 7-cluster structures observed for every of 143 genomes on our web-site [43].

6. Web-site on cluster structures in genomic word frequency distributions

To make the images and graphs of 143 genomes 7-cluster structures available for wide public, we established a web-site [43] for cluster structures in genomic word frequency distributions. For the methods see also [44].

For the moment our database contains 143 completely sequenced bacterial genomes and two types of cluster structures: the 7-cluster structure and the gene codon usage cluster structure. When browsing the database, a user can look at animated 3D-representations of these multidimensional cluster structures. For the description of the structures and the methods we refer the reader to the "intro" and "methods" pages of the web-site.

Another possibility which is provided on our web-site is browsing large-scale "maps" of various spaces where all 143 genomes can be embedded simultaneously. One example is the codon usage map: one point on the map is a genome, and close points correspond to the genomes with close codon usage. In fact, this is the same 64-dimensional triplet frequency space, used for construction of the 7-cluster structure. This gives the following hierarchy of maps: general map

of codon usage in 143 genomes, then the 7-cluster structure of in-phase triplet distributions, then the "thin structure" of every coding cluster: gene codon usage map. Clicking on a genome at the first map, the user "zooms" to its more detailed representations.

We strongly believe that the information in the database will help to advance existing tools for bacterial genomes analysis. Also it can serve as rich illustrative material for those who study sequence bioinformatics.

7. Discussion

In this paper we prove the universal 7-cluster structure in triplet distributions of bacterial genomes. Some hints at this structure appeared long time ago, but only recently it was explicitly demonstrated and studied.

Observability of the universal cluster structure depends mainly on two parameters. The triplet distribution f_{ijk} in the coding phase should be sufficiently far from the randomized distribution $m = p_i p_j p_k$ (see Fig. 3). The lengths of the cluster triangle side for bacterial genomes is proportional to the distance between these two distributions. The separability of these clusters from the non-coding one is determined by this distance, as well as by genome compactness.

The non-randomness of DNA sequence is known. For example, the coding DNA sequences were compared with the four-dimensional directed random walk and difference was reported in Ref. [45]. The 7-cluster structure is the main source of sequence heterogeneity (non-randomness) in the bacterial genomes. In this sense, our seven clusters is the basic model of bacterial genome sequence. We demonstrated that there are four basic "pure" types of this model, observed in nature: "parallel triangles", "perpendicular triangles", degenerated case and the flower-like type (see Fig. 7).

To explain the properties and types of the structure, which occur in natural bacterial genomic sequences, we studied 143 bacterial genomes available in Genbank in August, 2004. We showed that the codon usage of the genomes can be approximated by two multi-linear functions of their genomic G + C-content: one function for eubacterial genomes and the other one for archaea.

This observation is consistent with previous studies (Sueoka's neutrality plots, etc. [34–36]), nevertheless, the accuracy of the linear approximations (Figs. 4, 5) seems to be surprising. The difference between these linear dependencies for eubacterial and archaeal genomes is not explained yet (it is not a difference between two or several genomes, it is the difference between two straight lines which model the codon–position specific nucleotide usage with high accuracy). Available archaeal genomes are biased towards thermophilic species and they are known to have their own specific synonymous and non-synonymous codon usage [33]. The results of [46] show that synonymous codon usage is affected by two major factors: (i) the overall G + C content of the genome and (ii) growth at high temperature. It is natural to look for the source of the observed differences in these properties of thermophilic bacteria.

In the 64-dimensional space of all possible triplet distributions the bacterial codon distributions are close to two curves, that are close at their AT-rich ends and diverge at their GC-rich ends. When moving along these curves we meet all naturally occurred 7-cluster structures in the following order: "parallel triangles" for the AT-rich genomes (G + C-content is around 25%), then "perpendicular triangles" for G + C-content is around 35%, switching gradually to the degenerated case in the regions of GC = 50% and, finally, the degeneracy is resolved in one plane leading to the flower-like symmetric pattern (starting from GC = 60%). All these events can be illustrated using the material from the web-site we established [43].

The properties of the 7-cluster structure have natural interpretations in the language of Hidden Markov Models. Locations of clusters in multidimensional space correspond to in-state transition probabilities, the way how clusters touch each other reflects inter-state transition probabilities. Our clustering approach is independent on the Hidden Markov Modeling, though can serve as a source of information to adjust the learning parameters.

The question about mutual position of cluster triangles is an extension of the classical question about symmetry (or asymmetry) between forward and backward DNA strands [21,23–25].

In our paper we analyzed only triplet distributions. It is easy to generalize our approach for longer (or shorter) words. In-phase hexamers, for example, are characterized by the same 7-cluster structure. However, our experience shows that the most of information is contained in triplets: the correlations in the order of codons are small and the formulas (1) work reasonably well. Other papers confirm this observation (see, for example, Refs. [1,7]).

The subject of the paper has a lot of possible continuations. There are several basic questions: how one can explain the one-dimensional model of codon usage or why the signatures in the middle of G + C-content scale have palindromic structures? There are questions about how our model is connected with codon bias in translationally biased genomes: the corresponding cluster structure is the second hierarchical level or the "thin structure" in every cluster of the 7-cluster structure (see, for example, [47]). Also the following question is important: is it possible to detect and use the universal 7-cluster structure for higher eukaryotic genomes, where this structure also exists (see [9]), but is hidden by the huge non-coding cluster?

References

[1] S. Audic, J.M. Claverie, Self-identification of protein-coding regions in microbial genomes, Proc. Natl. Acad. Sci. USA 95 (17) (1998) 10026–10031.

[2] P. Baldi, On the convergence of a clustering algorithm for protein-coding regions in microbial genomes, Bioinformatics 16 (4) (2000) 367–371.

[3] P. Bernaola-Galvan, I. Grosse, P. Carpena, J.L. Oliver, R. Roman-Roldan, H.E. Stanley, Finding borders between coding and noncoding DNA regions by an entropic segmentation method, Phys. Rev. Lett. 85 (6) (2000) 1342–1345.

[4] P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S.D. Ehrlich, B. Prum, P. Bessieres, Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models, Nucleic Acids Res. 30 (6) (2002) 1418–1426.

[5] M. Borodovsky, J. McIninch, GENMARK: parallel gene recognition for both DNA strands, Comput. Chem. 17 (1993) 123–133.
[6] S.L. Salzberg, A.L. Delcher, S. Kasif, O. White, Microbial gene identification using interpolated Markov Models, Nucleic Acids Res. 26 (2) (1998) 544–548.
[7] A. Gorban, A. Zinovyev, T. Popova, Seven clusters in genomic triplet distributions, Silico Biol. 3 (2003), 0039. (E-print: http://arxiv.org/abs/cond-mat/0305681 and http://cogprints.ecs.soton.ac.uk/archive/00003077/)
[8] A.N. Gorban, A.Yu. Zinovyev, T.G. Popova, Statistical approaches to the automated gene identification without teacher, Institut des Hautes Etudes Scientiques. - IHES Preprint, France, 2001, - M/01/34. Available at http://www.ihes.fr web-site. (See also e-print: http://arxiv.org/abs/physics/0108016).
[9] A. Zinovyev, Visualizing the spatial structure of triplet distributions in genetic texts. - IHES Preprint, France, 2002, - M/02/28. Available at http://www.ihes.fr web-site
[10] A.Yu. Zinovyev, A.N. Gorban, T.G. Popova, Self-Organizing Approach for Automated Gene Identification, Open Systems Inform. Dyn. 10 (4) (2003) 321–333.
[11] A.N. Gorban, A.Yu. Zinovyev, Visualization of data by method of elastic maps and its applications in genomics, economics and sociology, Institut des Hautes Etudes Scientiques. - IHES Preprint, France, 2001, - M/01/36. Available at http://www.ihes.fr web-site
[12] A.N. Gorban, A.Y. Zinovyev, D.C. Wunsch, Application of The Method of Elastic Maps In Analysis of Genetic Texts, in: Proceedings of International Joint Conference on Neural Networks (IJCNN), 2003, Portland, Oregon, July 20–24.
[13] H.Y. Ou, F.B. Guo, C.T. Zhang, Analysis of nucleotide distribution in the genome of Streptomyces coelicolor A3(2) using the Z curve method, FEBS Lett. 540 (1–3) (2003) 188–194.
[14] A.N. Gorban, E.M. Mirkes, T.G. Popova, M.G. Sadovsky, A new approach to the investigations of statistical properties of genetic texts, Biofizika 38 (5) (1993) 762–767.
[15] N.N. Bugaenko, A.N. Gorban, M.G. Sadovskii, Maximum entropy method in analysis of genetic text and measurement of its information content, Open Systems and Inform. Dyn. 5 (1998) 265–278.
[16] A.N. Gorban, T.G. Popova, M.G. Sadovsky, Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy, Open System Inform. Dyn. 7 (2000) 1–17.
[17] S. Karlin, Global dinucleotide signatures and analysis of genomic heterogeneity, Curr. Opinion Microbiol. 1 (5) (1998) 598–610.
[18] S. Mahony, J.O. McInerney, T.J. Smith, A. Golden, Gene prediction using the self-organizing map: automatic generation of multiple gene models, BMC Bioinform. (2004) 5(1):23. Online: http://www.biomedcentral.com/1471-2105/5/23
[19] J. Besemer, A. Lomsadze, M. Borodovsky, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes Implications for finding sequence motifs in regulatory regions, Nucleic Acids Res 29 (12) (2001) 2607–2618.
[20] C. Mathe, M.F. Sagot, T. Schiex, P. Rouze, Current methods of gene prediction, their strengths and weaknesses, Nucleic Acids Res. 30 (19) (2002) 4103–4117.
[21] J.R. Lobry, Properties of a general model of DNA evolution under no-strand-bias conditions, J. Mol. Evol. 40 (3) (1995) 326–330.
[22] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (5) (1996) 660–665.
[23] J. Mrazek, S. Karlin, Strand compositional asymmetry in bacterial and large viral genomes, Proc. Natl. Acad. Sci. USA 95 (1998) 3720–3725.
[24] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, DNA asymmetry and the replicational mutational pressure, J. Appl. Genet. 42 (4) (2001) 553–577.
[25] J.R. Lobry, N. Sueoka, Asymmetric directional mutation pressures in bacteria, Genome Biol. 3 (10) (2002) RESEARCH0058.
[26] A.C. Frank, J.R. Lobry, Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, Gene 238 (1) (1999) 65–77.

[27] N. Sueoka, Two Aspects of DNA base composition: $G+C$ content and translation-coupled deviation from intra-strand rule of $A=T$ and $G=C$, J. Mol. Evol. 49 (1999) 49–62.
[28] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
[29] W.S. Torgerson, Theory and Methods of Scaling, John Wiley, New York, 1958.
[30] C.T. Zhang, R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic technique, Nucleic Acids Res. 19 (1991) 6313–6317.
[31] C.T. Zhang, K.C. Chou, A graphic approach to analyzing codon usage in 1562 Escherichia coli protein coding sequences, J. Mol. Biol. 238 (1994) 1–8.
[32] E.N. Trifonov, Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences, J. Mol. Biol. 194 (1987) 643–652.
[33] J.R. Lobry, D. Chessel, Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria, J. Appl. Genet. 44 (2) (2003) 235–261.
[34] J. Lobry, Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species, Gene 205 (1–2) (1997) 309–316.
[35] X.F. Wan, D. Xu, A. Kleinhofs, J. Zhou, Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes, BMC Evol. Biol. 4 (1) (2004) 19.
[36] N. Sueoka, Directional mutation pressure and neutral molecular evolution, Proc. Natl. Acad. Sci. USA 85 (8) (1988) 2653–2657.
[37] N. Sueoka, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, J. Mol. Evol. 40 (3) (1995) 318–325.
[38] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Natl. Acad. Sci. USA 48 (1962) 582–592.
[39] M. Bulmer, The selection-mutation-drift theory of synonymous codon usage, Genetics 129 (3) (1991) 897–907.
[40] S.L. Chen, W. Lee, A.K. Hottes, L. Shapiro, H.H. McAdams, Codon usage between genomes is constrained by genome-wide mutational processes, Proc. Natl. Acad. Sci. USA 101 (10) (2004) 3480–3485.
[41] M. Archetti, Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code, J. Mol. Evol. 59 (2) (2004) 258–266.
[42] P.M. Sharp, M. Stenico, J.F. Peden, A.T. Lloyd, Codon usage: mutational bias translational selection or both?, Biochem. Soc. Trans. 21 (4) (1993) 835–841.
[43] Cluster structures in genomic word frequency distributions. Web-site with supplementary materials. http://www.ihes.fr/~zinovyev/7clusters/index.htm
[44] A.N. Gorban, A.Yu. Zinovyev, T.G. Popova, Four basic symmetry types in the universal 7-cluster structure of 143 complete bacterial genomic sequences. E-print: http://arxiv.org/abs/q-bio.GN/0410033
[45] A. Som, S. Sahoo, J. Chakrabarti, Coding DNA sequences: statistical distributions, Math. Biosci. 183 (1) (2003) 49–61.
[46] D.J. Lynn, A.C. Gregory, G.A.C. Singer, D.A. Hickey, Synonymous codon usage is subject to selection in thermophilic bacteria, Nucleic Acids Res. 30 (19) (2002) 4272–4277.
[47] A. Carbone, A. Zinovyev, F. Kepes, Codon Adaptation Index as a measure of dominating codon bias, Bioinformatics 19 (13) (2003) 2005–2015.