INFORMATICA, 2019, Vol. 30, No. 3, 553–571

DOI: http://dx.doi.org/10.15388/Informatica.2019.218

Local Symmetry of Non-Coding Genetic Sequences

Marijus RADAVIČIUS[1]*, Tomas REKAŠIUS[2],
Jurgita ŽIDANAVIČIŪTĖ[2]

[1] *Institute of Data Science and Digital Technologies, Vilnius University,*
Akademijos st. 4, LT-04812 Vilnius, Lithuania
[2] *Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania*
e-mail: marijus.radavicius@mii.vu.lt, tomas.rekasius@vgtu.lt, jurgita.zidanaviciute@vgtu.lt

Received: November 2018; accepted: May 2019

Abstract. The simplest hypothesis of DNA strand symmetry states that proportions of nucleotides of the same base pair are approximately equal within single DNA strands. Results of extensive empirical studies using asymmetry measures and various visualization tools show that for long DNA sequences (approximate) strand symmetry generally holds with rather rare exceptions. In the paper, a formal definition of DNA strand *local symmetry* is presented, characterized in terms of generalized logits and tested for the longest non-coding sequences of bacterial genomes. Validity of a special regression-type probabilistic structure of the data is supposed. This structure is compatible with probability distribution of random nucleotide sequences at a steady state of a context-dependent reversible Markov evolutionary process. The null hypothesis of strand local symmetry is rejected in majority of bacterial genomes suggesting that even neutral mutations are skewed with respect to leading and lagging strands.

Key words: generalized logit, DNA strand symmetry, Markov random field, characterization, hypothesis testing.

> *Due to symmetry, the nature is perfect.*
> *Spices of asymmetry make it beautiful.*

1. Introduction

Genetically (or biologically) informative sequences can be defined as those which are either close to a known genetically important sequence or are far from sequences known to be noninformative. The first criterion seems to be more practical, however it is limited since it tries to reproduce what is already known. The second principle is more fundamental and more convenient for mathematical formalization and statistical inference. When employing this principle, the problem is how to define the noninformative genetic sequence (we call it the *genetic noise*), i.e. the sequence which has no genetically or biologically important information.

[*] Corresponding author.

A model of the genetic noise is also crucial for statistical hypotheses testing, the phylogenetic tree reconstruction, simulations of the (neutral) evolutions, and in assessing the variability and uncertainty.

Genome regions whose evolution is not subjected to natural selection pressure and hence evolve with a neutral mutation rate can be viewed as the genetic noise. Those regions could be parts of non-coding regions of genoms of primitive species.

A generic formulation of empirical findings is sometimes called a *stylized fact*. The definition of the genetic noise should be consistent with the stylized facts about non-coding DNA sequences as well as with a probabilistic model of their evolution. Thus, the general aim of our investigation is to specify and to test statistically the basic properties of non-coding DNA sequences implied by a model of DNA evolution (Markov property, homogeneity, long-range dependence, reverse-complement symmetry, CpG content, etc.). In this work we focus on *symmetry/asymmetry* properties of two complementary DNA strands.

Chargaff's second parity rule. The simplest hypothesis of DNA strand symmetry (sometimes referred to as *Chargaff's second parity rule*) states that proportions of nucleotides of the same base pair are approximately equal within single DNA strands (Rudner *et al.*, 1968), i.e. %A $\approx$ %T and %C $\approx$ %G. Since the lagging strand is read in the reverse order, an extension of this first-order symmetry to higher-orders is called *reverse-complement symmetry*, or *intra-strand parity* (ISP) (Powdel *et al.*, 2009), or simply *strand symmetry* (Baisnée *et al.*, 2002; Zhang and Huang, 2008). Although rather natural, this universal phenomenon of strand symmetry in the chromosomes needs explicit description and explanation. Actually, it may be the effect of a wide range of mechanisms operating at multiple orders and length scales (Baisnée *et al.*, 2002).

Thus far the issue about strand symmetry, its origins and biological significance is controversial. On the one hand, results of empirical studies using various asymmetry measures and visualization tools show that for long DNA sequences (approximate) strand symmetry generally holds with rather rare exceptions. The fact that the strand symmetry should hold at the equilibrium state is also derived theoretically (Sueoka, 1995; Lobry, 1995). Baisnée *et al.* (2002) defined strand symmetry indices through relative L_1 distance between the observed frequencies of respective reverse-complementary oligonucleotides and compare them with critical values calculated for completely random sequences. In Kong *et al.* (2009), various symmetry indices (reverse, complement and inverse symmetry indices, global as well as segmental) based on L_2 distance have been calculated for 786 complete chromosomes. The authors have found that reverse-complement symmetry (inverse-complement plus reverse-symmetry in terms of the authors) is prevalent in complex patterns in most chromosomes. Rosandić *et al.* (2016) considered 20 symbolic quadruplets of trinucleotides obtained via interstrand mirror symmetry mappings (direct, reverse complement, complement, and reverse) and demonstrated quadruplet's symmetries in chromosomes of wide range of organisms, from Escherichia coli to human genomes. Powdel *et al.* (2009) have noticed another strand symmetry manifestation, intra-strand frequency distribution parity (ISFDP), which represents closeness of frequency distributions between the complementary mono/oligonucleotides. This general feature (with rare exceptions) was observed in chromosomes of bacteria, archaea and eukaryotes. It has been also noticed that the frequency of an genomic word is more similar to the frequency of its reversed complement than to the frequencies of other words of equivalent composition. This phenomenon is called exceptional symmetry. Afreixo *et al.* (2017) proposed a new measure to evaluate the exceptional symmetry effect based on discrepancy between frequency of symmetric word pair and frequencies of word pairs of equivalent composition. They identified words that show high symmetry effect across the 31 species, and across the 9 animal species studied. Fractal-like symmetry structures are considered in Petoukhov *et al.* (2018). Sobottka and Hart (2011) proposed a model based on a hidden Markov process for approximating the distributions of primitive DNA sequences. The model provides an alternative interpretation of strand symmetry and describes new symmetries in bacterial genomes. Cristadoro *et al.* (2018) introduced flexible statistical measures of symmetry and used them to define an extended Chargaff symmetry. The definition actually coincides with global strand symmetry of genoms defined and studied in Simons *et al.* (2005). Domain models introduced in Cristadoro *et al.* (2018) alow to explain simultaneously symmetries as well as non-random structures in genetic sequences and unravel previously unknown symmetries, which are organized hierarchically through different scales.

On the other hand, statistical analyzes of the genomic sequences (Shporer *et al.*, 2016; Tavares *et al.*, 2018), especially those based on Markov-type models (Hart and Martínez, 2011; Hart *et al.*, 2012), have demonstrated significant deviations from the second Chargaff's parity rule and its extensions. A statistical IS-Poisson model introduced in Shporer *et al.* (2016) assumes that frequencies of oligonucleotides (DNA k-mers) follow the Poisson distribution. The model allows to conclude that for k-mers with low k (even for nucleotides, $k = 1$) violations of symmetry, although extremely small, are significant. In Tavares *et al.* (2018), both the distance distributions and the frequencies of symmetric words in the human DNA have been compared. The results obtained suggest that some asymmetries in the human genome go far beyond Chargaff's rules.

One of the explanations of strand asymmetry (skew), i.e. violation of symmetry, is mutation bias. When investigating asymmetries in mutation patterns, phylogenetic estimation based on maximum likelihood can be applied. Usually independent evolution models completely determined by nucleotide substitution rates are employed, see, e.g. Faith and Pollock (2003), Marin and Xia (2008). Note that mathematical models for evolutionary inference considered in Parks (2015) also assume independent evolution. However, Siepel and Haussler (2004) presented extensions of standard phylogenetic models with context-dependent substitution and showed that the new models improve goodness of fit substantially for both coding and non-coding data. Moreover, considering context dependence leads to much larger improvements than does using a richer substitution model or allowing for rate variation across sites, under the assumption of site independence. We refer to Bérard and Guéguen (2012) for a more recent application of context-dependent substitution models in a phylogenetic context.

In this paper, *DNA strand local symmetry* introduced in Židanavičiūtė (2010) is tested for the longest non-coding (in the both leading and lagging strands) sequences of bacterial genomes taken from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Validity of a special regression-type probabilistic structure of the data is supposed. This structure is compatible with probability distribution of random nucleotide sequences at a steady state of a context-dependent reversible Markov evolutionary process (Jensen, 2005), see also Arndt *et al.* (2003), Lunter and Hein (2004). The null hypothesis of strand local symmetry is rejected in majority of bacterial genomes suggesting that even neutral mutations are skewed with respect to leading and lagging strands.

The rest of the paper is organized as follows. In the next section the definition of *strand local symmetry* is presented and characterization of this property in terms of generalized logits is given. Results of statistical analysis are discussed in Section 3. We end with some concluding remarks.

2. Local Symmetry

In this section we present the formal definition of local symmetry (Židanavičiūtė, 2010) and recall necessary notions and facts about discrete Markov random fields and loglinear modelling.

2.1. *Complementary Transformation*

Nucleotide sequences $x = x_{[n]}$ are sequences of elements $(x_i, i \in [n])$ with values from the alphabet $\mathcal{A} := \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Here $[n] := [1, n] = \{1, \dots, n\}$ is an interval of (positive) integers.

If $x = (x_1, \dots, x_n)$ is the leading strand of a DNA sequence, then the complementary one (the lagging strand read in the opposite direction) is denoted by $x^* = (x_1^*, \dots, x_n^*)$, where x_i^* is the complementary nucleotide to x_i in the ith base pair, and $x_* = x_{*[n]} := (x_{*1}, \dots, x_{*n}) = (x_n^*, \dots, x_1^*)$. This determines the *complementary transformation*. For instance,

$$x = (x_1, \dots, x_n): \quad \overrightarrow{\mathtt{\dots CGGATTTAGCTA\dots}}, \tag{1}$$

$$x^* = (x_1^*, \dots, x_n^*): \quad \overleftarrow{\mathtt{\dots GCCTAAATCGAT\dots}}, \tag{2}$$

$$x_* = (x_n^*, \dots, x_1^*): \quad \overrightarrow{\mathtt{\dots TAGCTAAATCCG\dots}}. \tag{3}$$

Chargaff and his colleagues (Rudner *et al.*, 1968) have noticed that

$$\bigl|\bigl\{t \in [n] : x_t = \mathtt{A}\bigr\}\bigr| \approx \bigl|\bigl\{t \in [n] : x_t = \mathtt{T}\bigr\}\bigr|,$$
$$\bigl|\bigl\{t \in [n] : x_t = \mathtt{C}\bigr\}\bigr| \approx \bigl|\bigl\{t \in [n] : x_t = \mathtt{G}\bigr\}\bigr|,$$

($|A|$ is the number of elements in a set A) which actually means that

$$\bigl|\bigl\{t \in [n] : x_t = \nu\bigr\}\bigr| \approx \bigl|\bigl\{t \in [n] : x_t^* = \nu\bigr\}\bigr|, \quad \forall \nu \in \mathcal{A}.$$

Thus, if x is treated as a *random* sequence, the last expression can be interpreted and generalized as follows: a probabilistic law generating x is *invariant* with respect to the complementary transformation $x \to x_*$.

2.2. *Basics of Markov Random Fields*

Let us start with basic notation and notions. Set $\mathcal{N} = [n]$, fix some positive integer $m < n/2$ and define the m-interior $\mathcal{N}_m^\circ$, the m-boundary $\partial\mathcal{N}_m$ of $\mathcal{N}$, and a collection of neighbourhoods:

$$\begin{aligned} &\mathcal{N}^\circ = \mathcal{N}_m^\circ := [m+1, \dots, n-m], \qquad \partial\mathcal{N} = \partial\mathcal{N}_m := \mathcal{N} \setminus \mathcal{N}_m^\circ, \\ &\mathcal{N}(\ell) = \mathcal{N}_m(\ell) := [\ell - m, \ell + m] \setminus \{\ell\}, \quad \ell \in \mathcal{N}^\circ. \end{aligned} \tag{4}$$

Given $x \in \mathcal{A}^n$ and a set of indices $J \subset \mathcal{N}$, let $x_J := (x_i, i \in J)$ denote the corresponding subsequence of x treated as an element of $\mathcal{A}^{|J|}$.

Definition 1. A random sequence $x \in \mathcal{A}^n$ is called an m-order *Markov random field* (MRF) with the state space $\mathcal{A}$ and the collection of neighbourhoods (4) iff $\forall a \in \mathcal{A}^n$ and $\forall \ell \in \mathcal{N}_m^\circ$

$$\mathbf{P}\{x_\ell = a_\ell \mid x_i = a_i,\ i \in [n],\ j \neq \ell\} = \mathbf{P}\{x_\ell = a_\ell \mid x_{\mathcal{N}_m(\ell)} = a_{\mathcal{N}_m(\ell)}\}. \tag{5}$$

A MRF x is called an m-order *homogeneous* MRF (m-MRF) if its m-order marginal conditional probabilities given in the right-hand side of (5) are independent of the site $\ell \in N^\circ$.

Definition 2. For a fixed *reference value* $r \in \mathcal{A}$ and given m-order marginal conditional probabilities

$$p(v|u) := \mathbf{P}\{x_{m+1} = v \mid x_{\mathcal{N}_m(m+1)} = u\}, \tag{6}$$

the respective *generalized logit* $\Lambda_v(u) = \Lambda_{v|r}(u)$ of a state $v \in \mathcal{A}$ versus r, given the neighbouring values $u \in \mathcal{A}^{2m}$, is defined as

$$\Lambda_{v|r}(u) := \log\left(\frac{p(v|u)}{p(r|u)}\right), \tag{7}$$

where we set $\log(0/0) = 0$ and $\log(p/0) = \infty$ for $p > 0$.

Suppose that values of m-MRF x are fixed on the boundary $\partial\mathcal{N}$: $x_{\partial\mathcal{N}} = b$ a.s. for some $b \in \mathcal{A}^{2m}$. Denote

$$\mathcal{X}_b := \{w \in \mathcal{A}^n \colon w_{\partial\mathcal{N}} = b\}.$$

From *Hammersley–Clifford theorem* (Besag, 1974), we obtain the following statement.

Proposition 1. *Suppose the distribution of m-MRF x is positive on $\mathcal{X}_b$, i.e. $\mathbf{P}\{x = w\} > 0$ for all $w \in \mathcal{X}_b$. Then the distribution of x is uniquely determined by the family of gene-*

Table 1
Nucleotide recoding rule.

	Purine	(Bonds)	Pyrimidine	s
Weak (2 bonds)	A	(=)	T	$s=-1$
Strong (3 bonds)	G	(≡)	C	$s=+1$
y	$y=-1$		$y=+1$	

ralized logits $\Lambda_{v|r}(u)$, $r, v \in \mathcal{A}$, $u \in \mathcal{A}^{2m}$, which for $w \in \mathcal{A}^{2m+1}$, take the following form

$$\Lambda_{w_{m+1}|r}(w_{\mathcal{N}_m(m+1)}) = \sum_{j=1}^{m+1} \big[\lambda_m(w_{[j,m+j]}) - \lambda_m\big(w^{(r)}_{[j,m+j]}\big)\big], \tag{8}$$

and in general depend on $M = (|\mathcal{A}| - 1)|\mathcal{A}|^m$ free scalar parameters. Here λ_m: $\mathcal{A}^{m+1} \to \mathbf{R}$ is an arbitrary function and $w^{(r)}$ is obtained from w by substituting r for w_{m+1}.

The statement is well-known, it is just rewritten in the notation introduced above.

2.3. *Local Symmetry: Definition and Characterization*

Let us recall that DNA strand symmetry means that probability distribution of oligonucleotides (sequences of adjacent nucleotides) of the both complementary strands of DNA, read in the respective direction, are similar in some sense. Having in mind the definition of m-MRF, the following formal definition of DNA strand symmetry can be given in terms of complementary transformation $w \to w_*$, $w \in \mathcal{A}^{2m+1}$, defined in Subsection 2.1.

Definition 3 (See Židanavičiūtė, 2010.). A random sequence $x_{[n]}$ is *m-order locally symmetric* ($m < n/2$) iff

$$\mathbf{P}\{x_\ell = v \mid x_{\mathcal{N}_m(\ell)} = u\} = \mathbf{P}\big\{x_\ell = v^* \,\big|\, x_{\mathcal{N}_m(\ell)} = u^*\big\} \tag{9}$$

for all $\ell \in \mathcal{N}^\circ$, $v \in \mathcal{A}, u \in \mathcal{A}^{2m}$.

Thus, for locally symmetric sequence, the marginal conditional distributions given m nearest neighbours (from the each side) are *invariant* under the complementary transformation. Under the assumption that DNA sequence x is m-MRF, the local strand symmetry can be expressed in terms of the conditional distributions $p(v|u)$ and/or the generalized logits $\Lambda_v(u)$.

For characterization of local symmetry in terms of the generalized logits, it is convenient to change the initial alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ of nucleotides v to $\mathcal{A}_1^2 = \mathcal{A}_1 \times \mathcal{A}_1$, $\mathcal{A}_1 := \{-1, +1\}$, via mapping $v \to z := (s, y)$ by the rule indicated in Table 1. The components $s = s(v) \in \mathcal{A}_1$ and $y = y(v) \in \mathcal{A}_1$ of a nucleotide $v \in \mathcal{A}$ represent its bonding property *strong* versus *weak* and its hydrophobic property *pyrimidine* (large molecule, less hydrophobic) versus *purine* (small molecule, more hydrophobic), respectively.

Now, let $x = (x_1, \dots, x_k) \in \mathcal{A}^k$ be a nucleotide sequence in the leading strand of DNA and let x^* be its complement read from the left to the right but taken in the common direction. Set

$$z = z(x) := (\overrightarrow{s}, \overrightarrow{y}) \in \mathcal{A}_1^{2k}, \qquad (\overrightarrow{s})_i := s(x_i), \qquad (\overrightarrow{y})_i := y(x_i), \quad i = 1, \dots, k, \tag{10}$$

$$z_* = z_*(x) := (\overrightarrow{s}_*, \overrightarrow{y}_*) = (\overleftarrow{s}, -\overleftarrow{y}), \tag{11}$$

$$(\overleftarrow{s})_i = (\overrightarrow{s})_{k-i+1}, \ (\overleftarrow{y})_i = (\overrightarrow{y})_{k-i+1}, \quad i = 1, \dots, k. \tag{12}$$

Then $z(x_*) = z_*(x)$. To illustrate the notation we apply them to the nucleotide sequence from (1) (to save space here and below we will omit the numeral 1):

$x = (x_1, \dots, x_n)$: ...CGGATTTAGCTA...,

$s = (s_1, \dots, s_n)$: ...+++−−−++−...,

$y = (y_1, \dots, y_n)$: ...+−−+++−++−...,

$x_* = (x_n^*, \dots, x_1^*)$: ...TAGCTAAATCCG...,

$\overrightarrow{s}_* = (s_n, \dots, s_1)$: ...−++−−−+++...,

$\overrightarrow{y}_* = -(y_n, \dots, y_1)$: ...+−++−−+++−....

In what follows we identify $p(v|u)$ with $p(z(v)|z(u))$ and $\Lambda_{v|r}(u)$, $r = \mathtt{A}$, with $\Lambda_{z(v)}(z(u))$.

Let us introduce functions that are *symmetric* (*antisymmetric*) with respect to the complementary transformation $z \to z_*$, $z \in \mathcal{A}_1^{2k}$, defined in (10)–(12).

Definition 4. A function $\psi : \mathcal{A}_1^{2k} \to \mathbf{R}$ is called *symmetric* (*antisymmetric*) with respect to the complementary transformation $w \to w_*$ iff $\psi(w) = \psi(w_*)$ (respectively, $\psi(w) = -\psi(w_*)$) for all $w \in \mathcal{A}_1^{2k}$.

Proposition 2. *Let $p(\eta|w), \eta \in \mathcal{A}_1^2$, $w \in \mathcal{A}_1^{4m}$, denote the m-order conditional probabilities of a bivariate random sequence $z(x)$ obtained from the nucleotide sequence $x \in \mathcal{A}^{2m+1}$ via z-transform* (10). *The following statements are equivalent*:

(a) *the sequence x and the marginal conditional probabilities $p(\cdot|\cdot)$ are m-order locally symmetric;*

(b) *there exist a symmetric function $\psi : \mathcal{A}_1^{4m} \to \mathbf{R}$ and two antisymmetric functions $\psi_- : \mathcal{A}_1^{4m} \to \mathbf{R}$ and $\psi_+ : \mathcal{A}_1^{4m} \to \mathbf{R}$ such that*

$$\log\left(\frac{p(-,+\mid w)}{p(-,-\mid w)}\right) = \psi_-(w), \tag{13}$$

$$\log\left(\frac{p(+,+\mid w)}{p(+,-\mid w)}\right) = \psi_+(w), \tag{14}$$

$$\log\left(\frac{p(+,+\mid w)\cdot p(+,-\mid w)}{p(-,+\mid w)\cdot p(-,-\mid w)}\right)=\psi(w),\quad \forall w\in\mathcal{A}_1^{4m}. \tag{15}$$

Another form of (13)–(15) *expressed in terms of the generalized logits* $\Lambda_{s,y}(w)$*:*

$$\Lambda_{-,+}(w) = \psi_-(w), \tag{16}$$

$$\Lambda_{+,-}(w) = \frac{1}{2}\big(\psi(w)-\psi_+(w)+\psi_-(w)\big), \tag{17}$$

$$\Lambda_{+,+}(w) = \frac{1}{2}\big(\psi(w)+\psi_+(w)+\psi_-(w)\big). \tag{18}$$

Proof. From the definition of generalized logits (7) and the recoding rule defined in Table 1 and (10), (11), we obtain, for all $w\in\mathcal{A}_1^{4m}$,

$$\begin{aligned}\Lambda_{-,+}(w) &= \log\left(\frac{p(-,+\mid w)}{p(-,-\mid w)}\right)=\psi_-(w),\\ \Lambda_{+,+}(w)-\Lambda_{+,-}(w) &= \log\left(\frac{p(+,+\mid w)}{p(+,-\mid w)}\right)=\psi_+(w),\\ \Lambda_{+,+}(w)+\Lambda_{+,-}(w)-\Lambda_{-,+}(w) &= \log\left(\frac{p(+,+\mid w)\cdot p(+,-\mid w)}{p(-,-\mid w)\cdot p(-,+\mid w)}\right)=\psi(w).\end{aligned}$$

Let us check that the functions $\psi_-(w)$, $\psi_+(w)$ and $\psi(w)$ possess the respective properties. By the definition of the local symmetry

$$p(s,y\mid w)=p\big(s,-y\,\big|\,w^*\big),\quad \forall w\in\mathcal{A}_1^{4m}. \tag{19}$$

Consequently, for all $w\in\mathcal{A}_1^{4m}$,

$$\psi_-(w) = \log\left(\frac{p(-,+\mid w)}{p(-,-\mid w)}\right)=\log\left(\frac{p(-,-\mid w^*)}{p(-,+\mid w^*)}\right) \tag{20}$$

$$= -\log\left(\frac{p(-,+\mid w^*)}{p(-,-\mid w^*)}\right)=-\psi_-(w^*). \tag{21}$$

Thus, $\psi_-(u)$ is antisymmetric. Analogously, for all $w\in\mathcal{A}_1^{4m}$,

$$\psi_+(w) = \log\left(\frac{p(+,+\mid w)}{p(+,-\mid w)}\right)=\log\left(\frac{p(+,-\mid w^*)}{p(+,+\mid w^*)}\right) \tag{22}$$

$$= -\log\left(\frac{p(+,+\mid w^*)}{p(+,-\mid w^*)}\right)=-\psi_+(w^*) \tag{23}$$

and

$$\psi(w) := \log\left(\frac{p(+,+\mid w)p(+,-\mid w)}{p(-,+\mid w)p(-,-\mid w)}\right) \tag{24}$$

$$= \log\left(\frac{p(+,-\mid w^*)p(+,+\mid w^*)}{p(-,-\mid w^*)p(-,+\mid w^*)}\right) = \psi(w^*). \tag{25}$$

The proof is completed. □

When estimating the generalized logits $\Lambda_\tau(w)$ one needs some parametrization. Below convenient parametric representations for symmetric and antisymmetric functions are presented.

According to the recoding rule defined in Table 1 and (10)–(12), $z = (s, y)$, $s, y \in \mathcal{A}_1^{2m}$, and hence in the sequel we deal with functions $\psi(s, y)$, $\psi : \mathcal{A}_1^k \times \mathcal{A}_1^k \to \mathbf{R}$, $k := 2m$.

Let $J \subset K := \{1, \dots, k\}$. Define the conjugate set J_* of the set J by

$$J_* := k + 1 - J = \{k + 1 - j : j \in J\}. \tag{26}$$

For a given sequence $s = (s_1, \dots, s_k) \in \mathcal{A}_1^k$, denote

$$s^J := \prod_{i \in J} s_i, \quad s^\emptyset := 1. \tag{27}$$

Any function $\psi : \mathcal{A}_1^k \times \mathcal{A}_1^k \to \mathbf{R}$ has the unique representation

$$\psi(s, y) = \sum_{J', J \subset K} a_{J'J}\, s^{J'} y^J, \quad s, y \in \mathcal{A}_1^k, \tag{28}$$

where summation is over all subsets of K (including the empty set $\emptyset$), $a_{J'J} = a_{J'J}(\psi)$, $J', J \subset K$, are free parameters determining the function ψ. In general, there are 4^k free parameters.

For a symmetric (antisymmetric) function ψ, we have

$$\psi(\overleftarrow{s}, -\overleftarrow{y}) = \psi(s, y) \quad \big(\text{respectively, } \psi(\overleftarrow{s}, -\overleftarrow{y}) = -\psi(s, y)\big). \tag{29}$$

Consequently, in the case of the symmetric ψ, for all $s, y \in \mathcal{A}_1^k$,

$$\sum_{J', J \subset K} a_{J'J}\, s^{J'} y^J = \sum_{J', J \subset K} (-1)^{|J|} a_{J'J}\, s^{J'_*} y^{J_*} = \sum_{J', J \subset K} (-1)^{|J|} a_{J'_* J_*}\, s^{J'} y^J, \tag{30}$$

and hence

$$a_{J'J} = (-1)^{|J|} a_{J'_* J_*}, \quad J', J \subset K. \tag{31}$$

If $J_* = J$ and $J'_* = J'$ (i.e. the both subsets J' and J are self-conjugate), the set J has an even number of elements and the equations (31) become the identities. Thus, there are no restrictions on the parameter $a_{J'J}$ values in this case. Let $k_* = k_*(k)$ denote the total number of the self-conjugate subsets of K.

Let τ be some total order (enumeration of elements) in the class of pairs (J', J) of the set K. Equations (31) imply that, for not self-conjugate pairs (J', J), $(J', J) \neq (J'_*, J_*)$, values of the coefficients $a_{J',J}$, $\tau(J', J) < \tau(J'_*, J_*)$, uniquely determine values of the remaining coefficients $a_{J',J}$, $\tau(J', J) > \tau(J'_*, J_*)$. Define

$$\mathcal{K}_2 := \big\{(J', J) : \tau(J', J) < \tau(J'_*, J_*)\big\}, \tag{32}$$

$$\mathcal{K}_{20} := \big\{(J', J) : \tau(J', J) = \tau(J'_*, J_*)\big\}. \tag{33}$$

From (28), (31), (32) and (33) we derive a general parametric form of a symmetric function ψ:

$$\psi_S(s, y) = \sum_{(J',J)\in\mathcal{K}_{20}} a_{J'J} s^{J'} y^J + \sum_{(J',J)\in\mathcal{K}_2} a_{J'J}\big(s^{J'} y^J + (-1)^{|J|} s^{J'_*} y^{J_*}\big). \tag{34}$$

It has

$$k_S = k_S(k) := k_*^2 + (4^k - k_*^2)/2 \tag{35}$$

free parameters.

The case of antisymmetric function differs from that of symmetric function only in additional minus sign in equations (31). For self-conjugate pairs (J', J), these equations hold if and only if $a_{J',J} = 0$. Thus, the first summand in (34) and in (35) disappears giving the function

$$\psi_A(s, y) = \sum_{(J,J')\in\mathcal{K}_2} a_{J,J'}\big(s^J y^{J'} - (-1)^{|J|} s^{J'_*} y^{J_*}\big) \tag{36}$$

with

$$k_A = k_A(k) := \big(4^k - k_*^2\big)/2 \tag{37}$$

free parameters.

For $m = 1$, $k_* = 2$, thus $k_A = (4^2 - 2^2)/2 = 6$ and $k_S = k_*^2 + k_A = 10$. Then symmetric (34) and antisymmetric (36) functions in a general form are given by

$$\begin{aligned}
\psi_S(s, y) = {} & a_{\emptyset\emptyset} + a_{\emptyset\{12\}} y_1 y_2 + a_{\{12\}\emptyset} s_1 s_2 + a_{\{12\}\{12\}} s_1 s_2 y_1 y_2 \\
& + a_{\emptyset\{1\}}(y_1 + y_2) + a_{\{1\}\emptyset}(s_1 - s_2) \\
& + a_{\{1\}\{1\}}(s_1 y_1 - s_2 y_2) + a_{\{1\}\{2\}}(s_1 y_2 - s_2 y_1) \\
& + a_{\{1\}\{12\}}(s_1 y_1 y_2 - s_2 y_1 y_2) + a_{\{12\}\{1\}}(s_1 s_2 y_1 + s_1 s_2 y_2),
\end{aligned}$$

$$\begin{aligned}\psi_A(s,y) = {} & a_{\emptyset\{1\}}(y_1 - y_2) + a_{\{1\}\emptyset}(s_1 + s_2) \\ & + a_{\{1\}\{1\}}(s_1y_1 + s_2y_2) + a_{\{1\}\{2\}}(s_1y_2 + s_2y_1) \\ & + a_{\{1\}\{12\}}(s_1y_1y_2 + s_2y_1y_2) + a_{\{12\}\{1\}}(s_1s_2y_1 - s_1s_2y_2),\end{aligned}$$

respectively.

Remark 1. An ordered sequence of symbols $x = \overrightarrow{x}$ is said to be *palindromic* iff $\overrightarrow{x} = \overleftarrow{x}$. We refer to the mapping $\overrightarrow{x} \to \overleftarrow{x}$ as *palindromic* transformation. In particular, for a DNA sequence x, the sequence (x, x_*^*) (here $x_*^* = (x_*)^* = (x^*)_*$) is palindromic and for a palindromic DNA sequence x, we have $s(x) = s(x_*)$ and $y(x) = -y(x_*)$. Note that the mapping $\mathcal{A}_1 \to \{+1, -1\}$ is a palindromic transform of the binary alphabet $\mathcal{A}_1$. Thus, the transform $y(x) \to y(x_*)$ is a superposition of two palindromic transforms: the transform of ordering of the sequence $y(x)$ elements and the transform of their alphabet.

Palindromic distributions are defined as those invariant under some palindromic operation. For instance, palindromic Bernoulli distributions (Marchetti and Wermuth, 2016) and palindromic Ising models (Marchetti and Wermuth, 2017) are invariant with respect to palindromic transforms of the alphabet. Formulas (34) and (36) are analogues of the characterization of palindromic Bernoulli distribution in terms of log-linear parameters of multivariate Bernoulli distribution given in Marchetti and Wermuth (2016).

3. Statistical Analysis

In this section, the first-order local symmetry of the longest non-coding sequences of bacterial genomes is tested by making use of its characterization in terms of generalized logits. A special regression-type probabilistic structure is imposed on the data.

3.1. *Regression-Type Probabilistic Structure of the Data*

Let us introduce the following data structure of the observed sequence $x \in \mathcal{A}^n$ with $n = (n_m + 1) \cdot (m + 1) - 1$, the quantity n_m being an integer:

$$\mathcal{D} := \big\{(v_\ell, z_\ell), \ell \in S\big\}, \quad S = S_{n,m} = \{1, 2, \ldots, n_m\}, \tag{38}$$

where $v_\ell := x_{(m+1)\ell}$ is a response variable and $z_\ell = x_{U_m((m+1)\ell)} \in \mathcal{A}^{2m}$ is a vector of explanatory variables, $\ell \in S$.

Assumption (Am):

1. $\{v_\ell,\ \ell \in S\}$ are conditionally independent given $\{z_\ell,\ \ell \in S\}$,
2. the conditional distribution of v_ℓ when value of z_ℓ is given does not depend on the site $\ell \in S$.

Assumption (Am) ensures that usual conditions of the generalized logit model with the response variable $v \in \mathcal{A}$ and the vector $z \in \mathcal{A}^{2m}$ of explanatory variables are satisfied, see (Agresti, 1990; Stokes *et al.*, 2001).

Remark 2 (Compatible evolutionary models). Suppose that a DNA sequence x is an outcome of a "long" homogeneous Markov evolution and hence has a stationary distribution. Assumption (Am) imposed on x is compatible with some common DNA evolutionary models. In particular, assumption (Am) with $m = 2$ hold for the independent codon evolution (Goldman and Yang, 1994). Assumption (Am) is also fulfilled if x is generated by m-MRF. Thus, it is valid in case of time-reversible, site-homogeneous and context-dependent Markov evolution model with m-order nearest neighbour interactions (see, e.g. Jensen, 2005). However, it is satisfied for some non-homogeneous, say $(m + 1)$-periodic, MRF of order m as well.

In general, the introduced regression-type data structure supplemented with a saturated generalized logit model for m-order conditional probabilities does not determine the distribution of x. However, if assumption (Am) holds for $S = \mathcal{N}^\circ$ (to be precise, for all shifts $((m + 1)S + \ell) \cap \mathcal{N}^\circ$ of the set of central nucleotides $(m + 1)S$ by ℓ, $\ell = 1, \dots, m - 1$, simultaneously), then, due to Hammersley–Clifford theorem (Proposition 1), x is m-MRF, and m-order generalized logits take the form of (8) and determine the distribution of x.

3.2. *Testing of Local Symmetry*

We analyse data of bacterial genomes (1221 genomes) taken from the database *GenBank* (https://www.ncbi.nlm.nih.gov/genbank/). In order to bypass the data sparsity problem *the longest non-coding* (for the both strands) DNA sequences are extracted from each genome. Assuming that the extracted sequences satisfy assumption (Am) with $m = 2$ we test their first order local symmetry.

In Fig. 1, the length distribution density of the extracted sequences is plotted in a logarithmic scale. The sequence lengths range from 1891 to 42901 with median 6605 and mean 7721. About a half of the sequences have length between 6000 and 8000. Since we assume (Am) with $m = 2$, the logit analysis is based on three-dimensional contingency tables (64 cells) of nonintersecting triplets in the DNA sequences. The average and median of cell counts in the tables are 40 and 34, respectively. The percentage of cells with less than 6 counts does not exceed 1%. Thus we can ignore p-value approximation problems incident to statistical analysis of sparse contingency tables (Agresti, 1990).

Generalized logit model is fitted to the data and the Wald criterion is applied to test if the coefficients of the generalized logit model satisfy conditions implied by antisymmetric (36) and symmetric (34) components of generalized logits specified in Proposition 2.

In Figs. 2–4, values of the logarithmized Student statistic (the Student statistic S transformed by $S \to \operatorname{sgn}(S)\log_2(1 + |S|)$) for testing the significance of the coefficients of response functions ψ_-, ψ_+ and ψ defined in (13)–(15), respectively, are presented. For better visibility of the logarithmized Student statistic distributions, we use the violin plot (Hintze and Nelson, 1998; Wickham, 2016), which combines a box plot and a kernel density plot that is rotated and placed on each side, to show the distribution shape of the

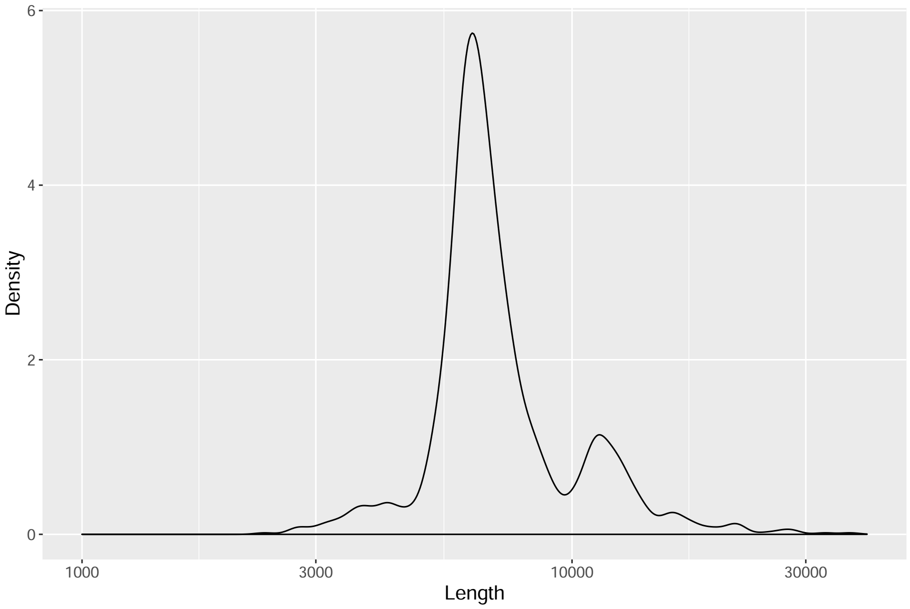

Fig. 1. The length distribution density of the longest non-coding sequences of bacteria genomes plotted in a logarithmic scale.

data. The first 6 coefficients represent the antisymmetric part of the response functions and the last 10 represent the symmetric part. According to Proposition 2, in case of the local symmetry, the first 2 response functions should be antisymmetric while the last one should be symmetric. Hence the last 10 and, respectively, the first 6 coefficients should be insignificant. In the figures, the approximate critical value obtained by 3σ rule (i.e. for the significance level ≈ 0.0054) corresponds to y-coordinates ± 2.

First response function (expected to be antisymmetric). The distributions of its coefficient estimates are represented in Fig. 2. The coefficient estimates of the antisymmetric part (white violins) have skewed distributions, especially the second, which is left-skewed and has large positive bias, and the third, which is right-skewed and has large negative bias. The distributions in the symmetric part (grey violins) are quite symmetric about zero. A large proportion of the non-coding DNA sequences ($> 40\%$) has significant (at the approximate significance level of 0.005) 7th coefficient (7th parameter) expected to be zero in the case of local symmetry.

In what follows only violations of local symmetry (grey violins) are discussed.

Second response function (expected to be antisymmetric). A major part ($> 70\%$) of the non-coding sequences has significant 7th coefficient (23rd parameter) expected to be zero in the case of local symmetry. A large proportion of the sequences also exhibits significant deviations from 0 of the 8th coefficient (24th parameter).

Third response function (expected to be symmetric). The second coefficient (34th parameter) expected to be zero in case of local symmetry shows a clear tendency to deviate significantly from 0.

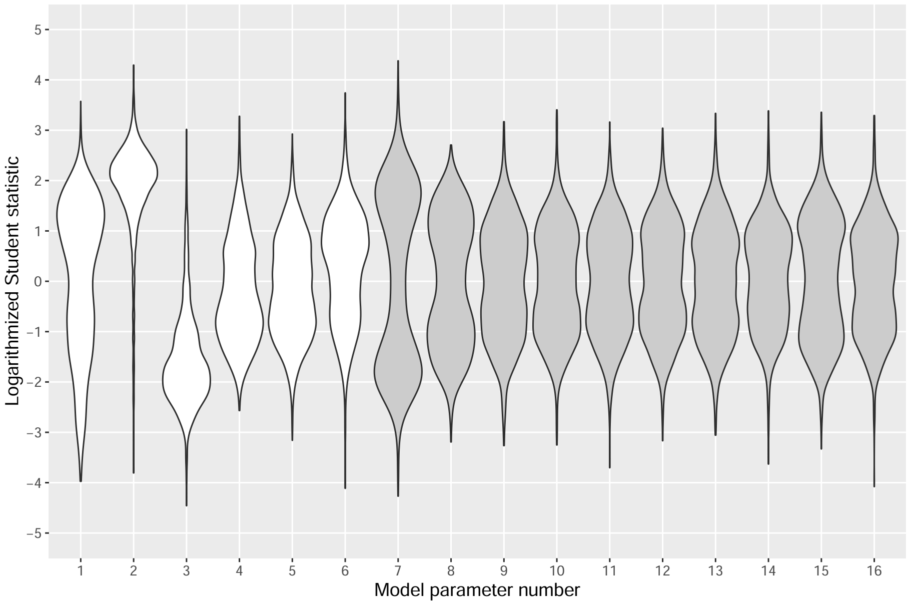

Fig. 2. Distribution of the logarithmized Student statistic of the 1st response function coefficients: the first 6 coefficients represent the antisymmetric part, the last 10 – the symmetric part (expected to be null).

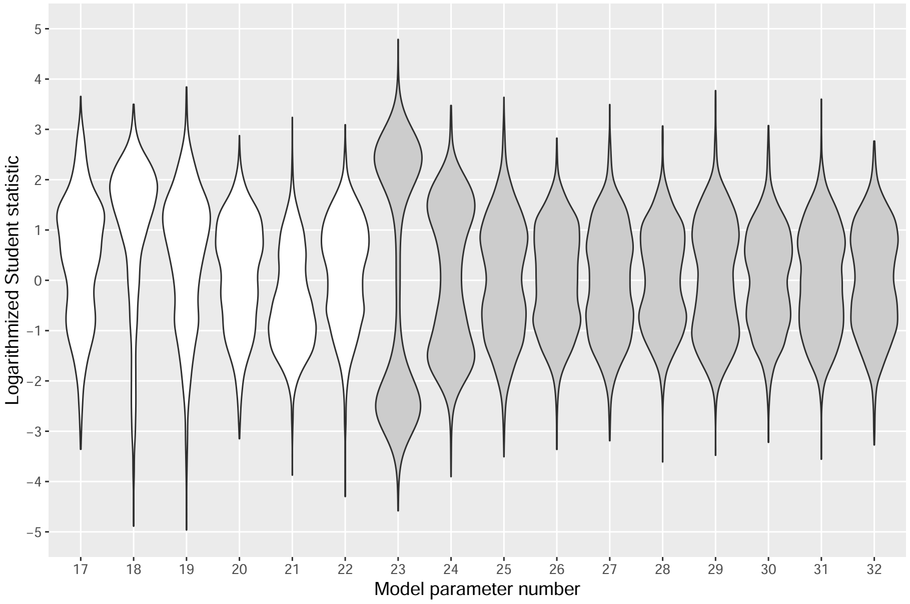

Fig. 3. Distribution of the logarithmized Student statistic of the 2nd response function coefficients: the first 6 coefficients represent antisymmetric part, the last 10 – the symmetric part (expected to be null).

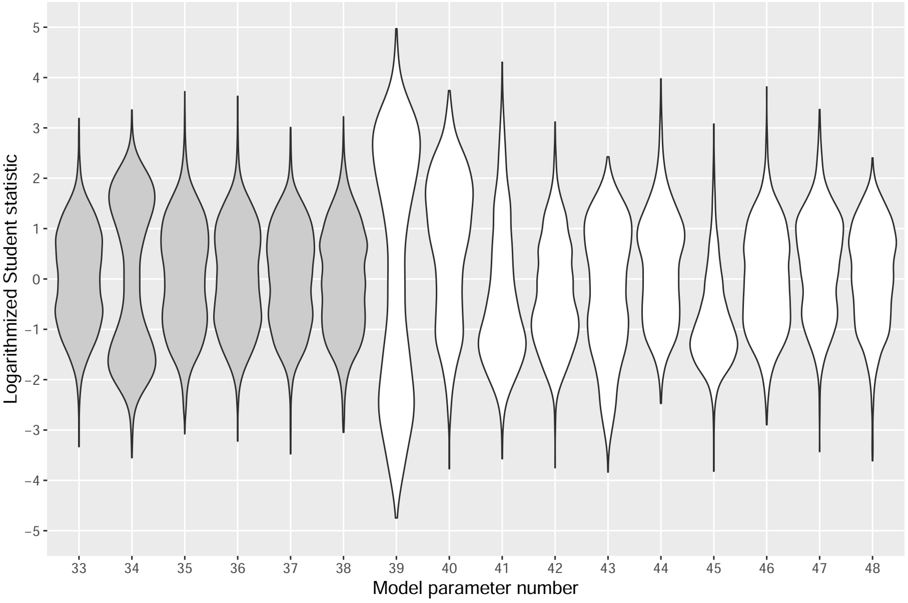

Fig. 4. Distribution of the logarithmized Student statistic of the 3rd response function coefficients: the first 6 coefficients represent antisymmetric part (expected to be null), the last 10 – the symmetric part.

In Fig. 5, centres of 8 clusters obtained using the standard R function for k-means clustering (R Core Team, 2018) of 48-dimensional vectors of the estimated model parameters (i.e. estimated coefficients of the all three response functions) are drawn. The coordinates of each centre are joint thus representing 8 different patterns of their interrelationships. The centre of the 8th cluster represents DNA sequences which approximately satisfy the local symmetry hypothesis. The sequences of the third cluster are also rather close to symmetry. Clusters 8 and 3, however, apparently differ in the regions [17, 19] and [19, 41]. All the clusters are similar in [1, 6]. In the grey zones (regions [7, 16] and [23, 38]), we have two triplets of similar clusters: $(1, 2, 6)$ and $(4, 5, 7)$. The 39th parameter for cluster 1 clearly differs from that of clusters 2 and 6 having the opposite sign. The same applies to clusters 4, 5 and 7, respectively. Clusters 2 and 6, as well as 5 and 7, exhibit some discrepancy in values of parameter 41. Cluster 5 also has specific values in the region [18, 19].

Note that the deviations of the parameter estimates in the grey region, i.e. their deviations from the DNA local symmetry hypothesis, are quite symmetric, see also Figs. 2–4. This observation is consistent with the ISFDP property noticed in (Powdel *et al.*, 2009).

3.3. *Concluding Remarks*

Elements of DNA sequences x are treated as random variables taking values from the alphabet $\mathcal{A} := \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. A definition of the local symmetry of x of order m is given and is characterized in terms of generalized logits (Židanavičiūtė, 2010). To test the first

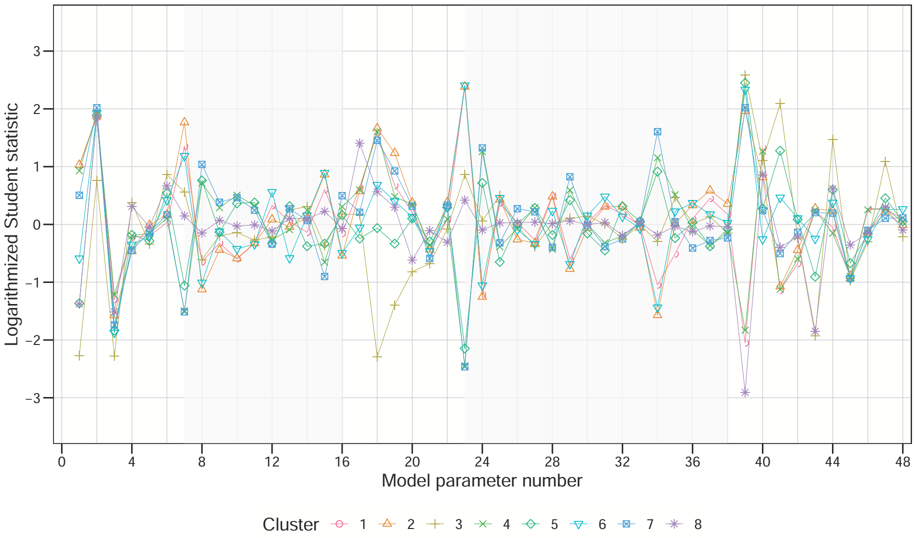

Fig. 5. Lines represent the patterns of 8 clusters obtained via k-means clustering from 48-dimensional data of the logarithmized Student statistics. The grey region indicates the model parameters vanishing under the null hypothesis of the local symmetry.

order local symmetry of non-coding sequences of bacteria genoms a special regression-type structure is imposed on probability distribution of x (assumption (Am) with $m = 2$). It defines a generalized logit model with 48 scalar parameters. In the case of the first order local symmetry, 22 of them should vanish.

The generalized logit model was fitted to the longest non-coding sequences of 1221 bacteria genomes taken from *GenBank* and Wald test was applied to check the null hypothesis of the first order local symmetry.

Conclusions:

1. Most of the non-coding sequences of bacteria genomes do not possess the first order local symmetry.
2. The deviations from the local symmetry of the non-coding sequences are pretty symmetric: the sample distributions of estimates of the model parameters that should vanish in case of the local symmetry are very close to symmetric one. Apparently this symmetry is related to intra-strand frequency distribution parity noticed in Powdel *et al.* (2009).
3. As a by-product of the statistical analysis of the local symmetry, we show that distributions of adjacent nucleotides are not independent even for the non-coding sequences of bacteria genoms. Hence independent evolution models (see, e.g. Faith and Pollock, 2003; Marin and Xia, 2008) are not consistent with the data of bacteria genomes.

Further work. A natural next step is to study higher order asymmetry patterns. Under assumptions (Am) with $m = 2$, for the statistical analysis of the second order local asymmetries the saturated generalized logit model can be applied. Then the analysis is based on 5-dimensional contingency tables (1024 cells). Hence for the data of the longest non-coding bacterial sequences, the average cell frequency in the contingency tables is less than 3, thus indicating their sparsity. A straightforward solution of the sparsity problem by joining all non-coding sequences of each genome seems to be inappropriate because of heterogeneity of DNA sequences (see, e.g. Cristadoro *et al.*, 2018). Special statistical methods are needed to deal with both the sparsity and heterogeneity.

Acknowledgement. The authors are grateful to Nanny Wermuth for relevant references and stimulating discussions on palindromic graphical models and to the anonymous referee for constructive comments.

References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons, New York.

Arndt, P.F., Burge, Ch.B., Hwa, T. (2003). DNA sequence evolution with neighbor-dependent mutation. *Journal of Computational Biology*, 10(3–4), 313–322.

Afreixo, V., Rodriges, J.M.O.S., Bastos, C.A.C., Tavares, A.H.M.P., Silva, R.M. (2017). Exceptional symmetry by genomic word: a statistical analysis. *Interdisciplinary Sciences Computational Life Sciences*, 9, 14–23.

Baisnée, P.-F., Hampson, S., Baldi, P. (2002). Why are complementary DNA strands symmetric? *Bioinformatics*, 18(8), 1021–1033.

Bérard, J., Guéguen, L. (2012). Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Systematic Biology*, 61(3), 510–521.

Besag, J. (1974). Spatial interactions and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36, 192–236.

Cristadoro, G., Esposti, M.D., Altmann, E.G. (2018). The common origin of symmetry and structure in genetic sequences. *Scientific Reports*, 8 (158171644).

Faith, J.J., Pollock, D.D. (2003). Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genome. *Genetics*, 165(2), 735–745.

Goldman, N., Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11, 725–736.

Hart, A., Martínez, S. (2011). Statistical testing of chargaff's second parity rule in bacterial genome sequences. *Stochastic Models*, 27, 272–317.

Hart, A., Martínez, S., Olmos, F. (2012). A gibbs approach to chargaff's second parity rule. *Journal of Statistical Physics*, 146, 408–422.

Hintze, J.L., Nelson, R.D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2), 181–184.

Jensen, J.L. (2005). Context dependent DNA evoliutionary models. *Research Reports*, 458. Department of Mathematical Sciences, University of Aarhus.

Kong, S.-G., Fan, W.-L., Chen, H.-D., Hsu, Z.-T., Zhou, N., Zheng, B., Lee, H.-C. (2009). Inverse symmetry in complete genomes and whole-genome inverse duplication. *PLoS ONE*, Nov. 09. doi:10.1371/journal.pone.0007553.

Lobry, J.R. (1995). Properties of a general model of DNA evolution under no-strand-bias conditions. *Journal of Molecular Evolution*, 40, 326–330; Erratum in: *Journal of Molecular Evolution*, 41, 680.

Lunter, G., Hein, J. (2004). A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20(18), 216–223.

Marchetti, G.M., Wermuth, N. (2016). Palindromic Bernoulli distributions. *Electronic Journal of Statistics*, 10(2), 2435–2460, also on arXiv:1510.09072.

Marchetti, G.M., Wermuth, N. (2017). Explicit, identical maximum likelihood estimates: for some cyclic Gaussian and cyclic Ising models. *Stat*, 6(1).

Marin, A., Xia, X. (2008). GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. *Journal of Theoretical Biology*, 253, 508–513.

Parks, S.L. (2015). *Mathematical Models and Statistics for Evolutionary Inference*. PhD thesis, University of Cambridge, Cambridge.

Petoukhov, S., Petukhova, E., Svirin, V. (2018). New symmetries and fractal-like structures in the genetic coding system. In: Hu Z., Petoukhov S., Dychka I., He M. (Eds.), *Advances in Computer Science for Engineering and Education*, *ICCSEEA 2018. Advances in Intelligent Systems and Computing*, Vol. 754. Springer, Cham, pp. 588–600.

Powdel, B.R., Satapathy, S.S., Kumar, A., Jha, P.K., Buragohain, A.K., Borah, M., Ray, S.K. (2009). A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides Chargaff's second parity rule. *DNA Research*, 16(6), 325–343.

R Core Team (2018). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org/.

Rosandić, M., Vlahović, I., Glunčić, M., Paar, V. (2016). Trinucleotide's quadruplet symmetries and natural symmetry law of DNA creation ensuing Chargaff's second parity rule. *Journal of Biomolecular Structure and Dynamics*, 34(7), 1383–1394.

Rudner, R., Karkas, J.D., Chargaff, E. (1968). Separation of B. subtilis DNA into complementary strands. III. Direct analysis. *Proceedings of the National Academy of Sciences of the USA*, 60, 921–922.

Shporer, S., Chor, B., Rosset, S., Horn, D. (2016). Inversion symmetry of DNA k-mer counts: validity and deviations. *BMC Genomics*, 17(696), 1–13.

Siepel, A., Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, 21(3), 468–488.

Simons, G., Yao, Y.-C., Morton, G. (2005). Global Markov models for eukaryote nucleotide data. *Journal of Statistical Planning and Inference*, 130, 251–275.

Sobottka, M., Hart, A.G. (2011). A model capturing novel strand symmetries in bacterial DNA. *Biochemical and Biophysical Research Communications*, 410, 823–828.

Stokes, M.E., Davis, C.S., Koch, G.S. (2001). *Categorical Data Analysis Using the SAS System*. SAS Institute, Cary, NC.

Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *Journal of Molecular Evolution*, 40, 318–325.

Tavares, A.H., Raymaekers, J., Rousseeuw, P.J., Silva, R.M., Bastos, C.A.C., Pinho, A., Brito, P., Afreixo, V. (2018) Comparing reverse complementary genomic words based on their distance distributions and frequencies. *Interdisciplinary Sciences Computational Life Sciences*, 10(1), 1–11.

Wickham, H. (2016). *R: ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2016.

Zhang, S.-H., Huang, Y.-Z. (2008). Characteristics of oligonucleotide frequencies across genomes: Conservation versus variation, strand symmetry, and evolutionary implications. *Nature Precedings*. hdl:10101/npre.2008.2146.1.

Židanavičiūtė, J. (2010). *Dependence Structure Analysis of Categorical Variables with Applications in Genetics*. Doctoral thesis., Vilnius Gediminas Technical University, Vilnius, Lithuania. Retrieved from https://vb.vgtu.lt/object/elaba:2115290/2115290.pdf.

M. Radavičius, Assoc. Prof. Dr., is a senior researcher at Institute of Data Science and Digital Technologies and a professor at Institute of Applied Mathematics, Vilnius University. He received a PhD degree (probability and statistics) in 1982 from the Steklov Institute of Mathematics of Russian Academy of Sciences (St. Petersburg Department). His major research interests include asymptotic statistics, nonparametric and adaptive estimation, dimension reduction and data sparsity, cluster analysis, applications of statistics in life sciences, medicine, linguistics and education.

T. Rekašius, Assoc. Prof. Dr., is working at Department of Mathematical Statistics, Vilnius Gediminas Technical University. He received a PhD degree (mathematics) in 2007 from Vilnius Gediminas Technical University and Institute of Mathematics and Informatics, Vilnius. His major research interests include bioinformatics, applications of statistics in life sciences and medicine.

J. Židanavičiūtė, Dr., received a master's degree in statistics from 2003 and a PhD degree in mathematics from 2010 from Vilnius Gediminas Technical University. She has been working at Vilnius Gediminas Technical University for 15 years. Her major research interests is applications of statistics in engineering, medicine and other fields.