J Stat Phys (2012) 146:408–422
DOI 10.1007/s10955-011-0377-6

A Gibbs Approach to Chargaff's Second Parity Rule

Andrew Hart · Servet Martínez · Felipe Olmos

Received: 15 July 2011 / Accepted: 8 October 2011 / Published online: 27 October 2011

Abstract Chargaff's second parity rule (CSPR) asserts that the frequencies of short polynucleotide chains are the same as those of the complementary reversed chains. Up to now, this hypothesis has only been observed empirically and there is currently no explanation for its presence in DNA strands. Here we argue that CSPR is a probabilistic consequence of the reverse complementarity between paired strands, because the Gibbs distribution associated with the chemical energy between the bonds satisfies CSPR. We develop a statistical test to study the validity of CSPR under the Gibbsian assumption and we apply it to a large set of bacterial genomes taken from the GenBank repository.

Keywords Reverse complementary relation · Chargaff's parity rules · Gibbs measure · Central Limit Theorem

1 Introduction

Double helical DNA is made up of two complementary polynucleotide chains, the primary and the secondary strands, each having opposing polarities. Chargaff's first parity rule is that "the numbers of A's and T's and the numbers of C's and G's match exactly in every DNA duplex" [4]. Chargaff's second parity rule (CSPR) states that this is valid when looking at a single strand, see [13], and that this happens not only for mononucleotides but also for short polynucleotide chains. Chargaff's first parity rule is a simple consequence of the double-stranded organization of genomic sequences and the chemistry of nucleic acids which only

A. Hart (✉) · S. Martínez · F. Olmos
Departamento Ingeniería Matemática and Centro Modelamiento Matemático, UMI 2071
CNRS-UCHILE, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 170,
Correo 3, Santiago, Chile
e-mail: ahart@dim.uchile.cl

S. Martínez
e-mail: smartine@dim.uchile.cl

F. Olmos
e-mail: felipe@olmos.cl

permits A to bond with T and C to bond with G. In this work we argue that the reverse complementary relationship between nucleic acids on opposing strands can also explain Chargaff's second parity rule provided that the distribution of nucleotides throughout the genome is governed by a Gibbs distribution.

CSPR was first observed experimentally in *Bacillus subtilis* [13] and was subsequently confirmed in sufficiently long sequences available in GenBank for small polymer chains of 3 to 6 bases [12]. More recent empirical studies assessing its validity can be found in [1, 10] and [7]. In addition, the empirical study of various symmetries carried out in [8] shows that CSPR is prevalent in complex patterns found in genomic sequences.

A number of possible mechanisms explaining strand symmetry have been proposed, for example, no strand biases for mutation and selection [9, 14] and selection of step-loop structures [6]. Further discussion of various mechanisms that could support the origins of this intrastrand symmetry are discussed in [15] and references therein. In [2] and [11], a number of mechanisms causing violation of CSPR in short polymers are described.

Here we propose that CSPR arises directly from the effect that reverse complementarity has on the Gibbs distribution. In Sect. 2 we give the framework. There, the empirical polymer frequencies are replaced by polymer occurrence probabilities on a translation invariant probability distribution. We then express CSPR using this notion and prove that CSPR written in this way follows from the fact that energy symmetry is preserved for the Gibbsian distribution. This is done in Theorem 1. In Sect. 3 we give a characterization of CSPR for dinucleotides, we prove an extension of the Central Limit Theorem for Gibbs measures to vector-valued random variables and we derive an explicit expression for the asymptotic covariance matrix. In Sect. 4 we supply a statistical test for the validity of CSPR for dinucleotides under the hypothesis that the nucleotides of the strand are distributed as a stationary Gibbsian process. We have applied the test extensively to bacterial genomes available from GenBank. The hypothesis of CSPR in the Gibbsian setting is confirmed for a large number of genomes. Further analysis would be necessary in order to determine whether genomes rejected by the test were because they fail to comply with CSPR, because they are not Gibbsian, or both.

2 Chargaff's Second Parity Rule

2.1 Preliminaries

Let $\mathcal{A}$ be a finite set (alphabet) endowed with an involution $\Gamma : \mathcal{A} \to \mathcal{A}$: Γ is one-to-one and $\Gamma^{-1} = \Gamma$. In the genomic setting $\mathcal{A} = \{A, C, G, T\}$ and Γ is an involution given by the complementary function $\Gamma(A) = T$ and $\Gamma(C) = G$.

Let $x = (x_j : j = 0, \ldots, n-1) \in \mathcal{A}^n$ be the sequence of nucleotides on a strand of the genome (for bacterial DNA $n \approx 10^6$). The sequence x complies with CSPR whenever the frequencies of all short polymers agree with the frequencies of their reverse complements. In other words, for all k small (order of 10) and all polymers $(a_0, \ldots, a_{k-1}) \in \mathcal{A}^k$:

$$\begin{aligned} &\#\left\{j \le n-k : x_j = a_0, \ldots, x_{j+k-1} = a_{k-1}\right\} \\ &\quad = \#\left\{j \le n-k : x_j = \Gamma(a_{k-1}), \ldots, x_{j+k-1} = \Gamma(a_0)\right\}. \end{aligned} \tag{1}$$

(Here $\#B$ denotes the cardinality of the set B.) Observe that the frequency is computed by moving a window of length k along the strand.

2.2 CSPR as a Symmetric Probability Relation

We shall use a theoretical framework in which the strands are modeled by bi-infinite sequences and the frequencies of a word are the probabilities that they appear at an arbitrary place. We restrict ourselves to translation invariant probability distributions that are Gibbs measures with respect to the chemical energy.

The strands are modeled by sequences in $\mathcal{A}^{\mathbb{Z}}$. Thus, $x = (x_j : j \in \mathbb{Z})$ represents the primary strand in the sense $5'$ to $3'$ while $y = (y_j : j \in \mathbb{Z})$ represents the complementary strand in the sense $3'$ to $5'$. They are related by reverse complementarity: $y_j = \Gamma(x_{-j})$ for $j \in \mathbb{Z}$. Let us write this rule in another way. Let $\mathcal{I} : \mathcal{A}^{\mathbb{Z}} \to \mathcal{A}^{\mathbb{Z}}$ be the space reversal involution given by $(\mathcal{I}(x))_j = x_{-j}$ and let $\overline{\Gamma} : \mathcal{A}^{\mathbb{Z}} \to \mathcal{A}^{\mathbb{Z}}$ be such that $(\overline{\Gamma}(x))_j = \Gamma(x_j)$, for $x \in \mathcal{A}^{\mathbb{Z}}$, $j \in \mathbb{Z}$. Then, the rule of reverse complementarity may be written as $y = \overline{\Gamma} \circ \mathcal{I}(x)$.

A genome duplex is the pair (x, y) and we denote by $\widetilde{\Psi}(x, y)$ its chemical energy, which results from the interactions between the nucleotides on both strands. Since the interactions between the nucleotides are symmetric we assert that

$$\widetilde{\Psi}(x, y) = \widetilde{\Psi}(y, x). \tag{2}$$

Insight into this equality may be obtained from the discussion of energy on finite pieces which appears in [3]. Let $\Psi_l(x[-l, l]; y[-l, l])$ be the energy in the portion $[-l, l] = \{-l, \ldots, l\}$ of the duplex. In analogy with [3], Page 5, this energy can be assumed to be given by

$$\begin{aligned}\Psi_l(x[-l, l]; y[-l, l]) = &\sum_{-l \le j \le k \le l} \psi^s(k - j; x_k, x_j) + \sum_{-l \le j \le k \le l} \psi^s(k - j; y_{-j}, y_{-k}) \\ &+ \frac{1}{2} \sum_{-l \le j, k \le l} \psi^o(|k - j|; x_k, y_{-j}).\end{aligned} \tag{3}$$

The first two summations are due to the interactions between sites on the same strand while the last one expresses the interactions between sites on opposite strands. The quantity $\psi^s(r; a, b)$ is the interaction between the nucleotides a, b at distance r on the same strand and $\psi^o(r; a, b)$ is the interaction between the nucleotides a, b in opposite strands such that the distance from the site containing one to the site in front of the other is r (recall that y_{-j} is in front of x_j, so the distance from site k containing x_k to the site j, which is in front of the site containing y_{-j}, is $|k - j|$). The expression (3) is clearly symmetric in x and y.

Let us express the symmetry relation (2) in another way. Since $y = \overline{\Gamma} \circ \mathcal{I}(x)$, the energy can be simply expressed as $\Psi(x) = \widetilde{\Psi}(x, \overline{\Gamma} \circ \mathcal{I}(x))$ and the symmetric dependence $\widetilde{\Psi}(x, y) = \widetilde{\Psi}(y, x)$ between the strands implies that Ψ satisfies the invariance property

$$\forall x \in \mathcal{A}^{\mathbb{Z}}: \quad \Psi(x) = \Psi(\overline{\Gamma} \circ \mathcal{I}(x)); \quad \text{or equivalently} \quad \Psi = \Psi \circ \overline{\Gamma} \circ \mathcal{I}.$$

Next, the set $\mathcal{A}^{\mathbb{Z}}$ is endowed with the product σ-algebra and let $T : \mathcal{A}^{\mathbb{Z}} \to \mathcal{A}^{\mathbb{Z}}$ be the translation operator given by $(T(x))_j = x_{j+1}$ for all $j \in \mathbb{Z}$. Let $\mathbb{P}$ be a translation invariant distribution on $\mathcal{A}^{\mathbb{Z}}$, that is

$$\mathbb{P}(T^{-1}B) = \mathbb{P}(B), \quad \forall \text{ measurable } B \subseteq \mathcal{A}^{\mathbb{Z}}.$$

In the spirit of (1), we say that $\mathbb{P}$ satisfies CSPR if

$$\begin{aligned}&\forall x \in \mathcal{A}^{\mathbb{Z}},\ \forall k \ge 1,\ \forall (a_0, \ldots, a_{k-1}) \in \mathcal{A}^k: \\ &\quad \mathbb{P}(x_0 = a_0, \ldots, x_{k-1} = a_{k-1}) = \mathbb{P}(x_0 = \Gamma(a_{k-1}), \ldots, x_{k-1} = \Gamma(a_0)).\end{aligned} \tag{4}$$

We claim that if $\mathbb{P}$ is a translation invariant distribution on $\mathcal{A}^{\mathbb{Z}}$, then property (4) is equivalent to $\mathbb{P}$ being $\overline{\Gamma} \circ \mathcal{I}$-invariant, that is, it satisfies $\mathbb{P}((\overline{\Gamma} \circ \mathcal{I})^{-1} B) = \mathbb{P}(B)$ for all measurable subsets B of $\mathcal{A}^{\mathbb{Z}}$. Indeed, from the equality

$$\overline{\Gamma} \circ \mathcal{I}^{-1}\{x : x_j = a_j, \ldots, x_k = a_k\} = \{x : x_{-k} = \Gamma(a_k), \ldots, x_{-j} = \Gamma(a_j)\} \tag{5}$$

taken together with the translation invariance property, one can show that if $\mathbb{P}$ is $\overline{\Gamma} \circ \mathcal{I}$-invariant then (4) holds. Conversely, the same translation invariance property combined with equality (5) may be used to prove that (4) implies $\mathbb{P}_\Psi((\overline{\Gamma} \circ \mathcal{I})^{-1} B) = \mathbb{P}(B)$ for all cylinders B. Carathéodory's extension theorem then shows that this holds for all measurable sets B and the claim follows.

In the next section, we use the symmetry of energy to imply that all genomes comply with CSPR (4) under a Gibbsian hypothesis.

2.3 Gibbs Measures and CSPR

We will derive CSPR from the complementary relation in the thermodynamical formalism. We begin by introducing some basic notions in this formalism.

Let $\mathcal{A}$ be a finite alphabet and $\Xi = (\Xi(a, b) : a, b \in \mathcal{A})$ be an aperiodic $0-1$-valued matrix. The shift of finite type defined by Ξ is the set $\mathcal{X}_\Xi = \{x \in \mathcal{A}^{\mathbb{Z}} : \Xi(x_j, x_{j+1}) = 1\ \forall j \in \mathbb{Z}\}$ endowed with the metric $\Delta_\theta(x, z) = \theta^{K(x,z)}$, where $K(x, z) = \sup\{k \geq 0 : x_i = z_i\ \forall |i| \leq k\}$ and $\theta \in (0, 1)$ is an arbitrary but fixed value. This metric induces the product topology.

Let $\theta \in (0, 1)$ be fixed. Consider the set of Hölder (continuous) functions in $(\mathcal{X}_\Xi, \Delta_\theta)$,

$$F_\theta = \{g \in C(\mathcal{X}_\Xi) : |g|_\theta < \infty\} \quad \text{where } |g|_\theta = \sup\left\{\frac{|g(x) - g(z)|}{\Delta_\theta(x, z)} : x, y \in \mathcal{X}_\Xi, x \neq z\right\}. \tag{6}$$

The linear set F_θ is a Banach space when it is endowed with the norm $\|g\|_\theta = \|g\|_\infty + |g|_\theta$.

Each Gibbs measure on $\mathcal{X}_\Xi$ is defined by an energy function $\Psi \in F_\theta$. The value $\Psi(x)$ represents the energy of the system in state $x \in \mathcal{X}_\Xi$. In the thermodynamic formalism of shifts of finite type, it has been shown (see Theorem 1.2 in [3], Pages 5–6) that there exists a unique translation invariant probability measure $\mathbb{P}_\Psi$ that satisfies

$$\exists 0 < c_1 < c_2 < \infty, p \in \mathbb{R},\ \forall z \in \mathcal{X}_\Xi, \forall k \geq 0: \quad c_1 \leq \frac{\mathbb{P}_\Psi(x : x_0 = z_0, \ldots, x_k = z_k)}{e^{-pk + \sum_{i=0}^{k-1} \Psi(T^i z)}} \leq c_2. \tag{7}$$

A detailed proof of this result as well as a complete exposition of this topic is given in [3], Pages 3–16. In this reference, it is also proven that the constant $p = p(\Psi)$ is the pressure of Ψ and that the probability measure $\mathbb{P}_\Psi$ is the unique translation invariant probability measure satisfying the variational principle $p(\Psi) = h_{\mathbb{P}_\Psi}(T) + \int \Psi d\mathbb{P}_\Psi$, where $h_{\mathbb{P}_\Psi}(T)$ is the entropy of T for the translation invariant distribution $\mathbb{P}_\Psi$.

Theorem 1 *Assume that the aperiodic matrix Ξ satisfies*

$$\forall a, b \in \mathcal{A}: \quad \Xi(a, b) = \Xi(\Gamma(b), \Gamma(a)). \tag{8}$$

Let $\Psi \in F_\theta$. Assume that Ψ is $\overline{\Gamma} \circ \mathcal{I}$-invariant: $\Psi(x) = \Psi(\overline{\Gamma} \circ \mathcal{I}(x))$ for all $x \in \mathcal{X}_\Xi$. Then the unique translation invariant Gibbs probability measure $\mathbb{P}_\Psi$ is $\overline{\Gamma} \circ \mathcal{I}$-invariant and hence complies with CSPR:

$$\forall k \geq 0, (z_0, \ldots, z_k) \in \mathcal{A}^{k+1}:$$
$$\mathbb{P}_\Psi(x : x_0 = z_0, \ldots, x_k = z_k) = \mathbb{P}_\Psi(x : x_0 = \Gamma(z_k), \ldots, x_k = \Gamma(z_0)).$$

Proof To begin, let $\mathbb{P} = \mathbb{P}_\Psi$ denote the unique T-invariant probability measure on $\mathcal{X}_\Xi$ that satisfies (7). Define the probability measure $\widetilde{\mathbb{P}}$ as $\widetilde{\mathbb{P}}(B) = \mathbb{P}((\overline{\Gamma} \circ \mathcal{I})^{-1} B)$ for all measurable sets B in $\mathcal{X}_\Xi$.

Claim 1 $\widetilde{\mathbb{P}}$ is translation invariant. This can be proved as follows. Note that $\mathcal{I}^{-1} = \mathcal{I}$ and $\overline{\Gamma}^{-1} = \overline{\Gamma}$, while $\overline{\Gamma}$ commutes with $\mathcal{I}$, T and T^{-1}. So $(\overline{\Gamma} \circ \mathcal{I})^{-1} = (\overline{\Gamma} \circ \mathcal{I})$. We also have $\mathcal{I} \circ T^{-1} = T \circ \mathcal{I}$ and hence

$$(\overline{\Gamma} \circ \mathcal{I})^{-1} \circ T^{-1} = T \circ \overline{\Gamma} \circ \mathcal{I}.$$

Since $\mathbb{P}$ is T-invariant, it is also T^{-1}-invariant, so

$$\begin{aligned}\widetilde{\mathbb{P}}(T^{-1}(B)) &= \mathbb{P}((\overline{\Gamma} \circ \mathcal{I})^{-1} \circ T^{-1}(B)) = \mathbb{P}(T \circ \overline{\Gamma} \circ \mathcal{I}(B)) = \mathbb{P}(\overline{\Gamma} \circ \mathcal{I}(B)) \\ &= \mathbb{P}((\overline{\Gamma} \circ \mathcal{I})^{-1}(B)) = \widetilde{\mathbb{P}}(B),\end{aligned}$$

which yields the claim.

Claim 2 $\widetilde{\mathbb{P}}$ satisfies

$$\exists 0 < \tilde{c}_1 < \tilde{c}_2 < \infty,\ \forall z \in \mathcal{X}_\Xi,\ \forall k \geq 0\colon \quad \tilde{c}_1 \leq \frac{\widetilde{\mathbb{P}}(x : x_0 = z_0, \ldots, x_k = z_k)}{e^{-pk + \sum_{i=0}^{k-1} \Psi(T^i z)}} \leq \tilde{c}_2.$$

Note that once this claim has been shown, the result will immediately follow because uniqueness of $\widetilde{\mathbb{P}}$ implies $\mathbb{P} = \widetilde{\mathbb{P}}$, and so $\mathbb{P}$ is $\overline{\Gamma} \circ \mathcal{I}$-invariant. To prove the claim, first observe that since $\Gamma^{-1} = \Gamma$ and $\mathbb{P}$ is T-invariant,

$$\begin{aligned}&\widetilde{\mathbb{P}}(x : x_0 = z_0, \ldots, x_k = z_k) \\ &\quad = \mathbb{P}\left(x : (\overline{\Gamma}(\mathcal{I}(x)))_0 = z_0, \ldots, (\overline{\Gamma}(\mathcal{I}(x)))_k = z_k\right) \\ &\quad = \mathbb{P}(x : \Gamma(x_0) = z_0, \ldots, \Gamma(x_{-k}) = z_k) = \mathbb{P}(x : x_0 = \Gamma(z_0), \ldots, x_{-k} = \Gamma(z_k)) \\ &\quad = \mathbb{P}(x : x_0 = \Gamma(z_k), \ldots, x_k = \Gamma(z_0)) = \mathbb{P}\left(x : x_0 = (\overline{\Gamma} \circ \mathcal{I}(z))_{-k}, \ldots, x_k = (\overline{\Gamma} \circ \mathcal{I}(z))_0\right) \\ &\quad = \mathbb{P}\left(x : x_0 = (T^{-k}(\overline{\Gamma} \circ \mathcal{I}(z)))_0, \ldots, x_k = (T^{-k}(\overline{\Gamma} \circ \mathcal{I}(z)))_k\right).\end{aligned}$$

On the other hand, from the equality $T^{i-k}(\overline{\Gamma} \circ \mathcal{I}(z)) = \overline{\Gamma} \circ \mathcal{I}(T^{k-i}(z))$ and using the fact that Ψ is $\overline{\Gamma} \circ \mathcal{I}$-invariant, we obtain

$$\begin{aligned}\sum_{i=0}^{k-1} \Psi(T^i T^{-k}(\overline{\Gamma} \circ \mathcal{I}(T^{-1} z))) &= \sum_{i=0}^{k-1} \Psi(\overline{\Gamma} \circ \mathcal{I}(T^{k-i-1} z)) \\ &= \sum_{i=0}^{k-1} \Psi(T^{k-i-1} z) = \sum_{i=0}^{k-1} \Psi(T^i(z)).\end{aligned}$$

Hence

$$\begin{aligned}&\frac{\widetilde{\mathbb{P}}(x : x_0 = z_0, \ldots, x_k = z_k)}{e^{-pk + \sum_{i=0}^{k-1} \Psi(T^i z)}} \\ &\quad = \frac{\mathbb{P}(x : x_0 = (T^{-k}(\overline{\Gamma} \circ \mathcal{I}(z)))_0, \ldots, x_k = (T^{-k}(\overline{\Gamma} \circ \mathcal{I}(z)))_k)}{e^{-pk + \sum_{i=0}^{k-1} \Psi(T^i T^{-k}(\overline{\Gamma} \circ \mathcal{I}(T^{-1} z)))}}. \qquad (9)\end{aligned}$$

We note that

$$\forall z \in \mathcal{X}_\Xi,\ \forall k \geq 0: \quad \tilde{c}_1 \leq \frac{e^{-pk+\sum_{i=0}^{k-1}\Psi(T^i T^{-1}(z))}}{e^{-pk+\sum_{i=0}^{k-1}\Psi(T^i(z))}} \leq \tilde{c}_2,$$

with $\tilde{c}_1 = e^{\min\Psi - \max\Psi}$ and $\tilde{c}_2 = e^{\max\Psi - \min\Psi}$. Then,

$$\tilde{c}_1 \leq \frac{e^{-pk+\sum_{i=0}^{k-1}\Psi(T^i T^{-k}(\overline{\Gamma}\circ\mathcal{I}(T^{-1}z)))}}{e^{-pk+\sum_{i=0}^{k-1}\Psi(T^i T^{-k}(\overline{\Gamma}\circ\mathcal{I}(z)))}} \leq \tilde{c}_2. \tag{10}$$

Hence from (9), (7) and (10), we deduce that Claim 2 holds,

$$\forall z \in \mathcal{X}_\Xi,\ \forall k \geq 0: \quad c_1\tilde{c}_1 \leq \frac{\widetilde{\mathbb{P}}(x : x_0 = z_0, \ldots, x_k = z_k)}{e^{-pk+\sum_{i=0}^{k-1}\Psi(T^i z)}} \leq c_2\tilde{c}_2\,.$$

Hence, $\widetilde{\mathbb{P}} = \mathbb{P}$ and the proof is complete. □

Remark 2 We note that in the genomic framework, $\mathcal{A} = \{A, C, G, T\}$ and $\Xi(a, b) = 1$ for all $a, b \in \mathcal{A}$. Hence, Ξ always satisfies condition (8) and the Theorem 1 may be directly applied to genome sequences.

3 CSPR for Dinucleotides

3.1 A 5-Dimensional Characterization of CSPR

Henceforth, we shall focus on the dinucleotide distributions under CSPR. Let $\mathbb{P}$ be a translation invariant distribution on $\mathcal{A}^{\mathbb{Z}}$. As stated, CSPR means that for all $R \geq 1$, we have

$$\begin{aligned}&\forall\,(a_0, \ldots, a_{R-1}) \in \mathcal{A}^R:\\ &\quad \mathbb{P}(x : x_0 = a_0, \ldots, x_{R-1} = a_{R-1}) = \mathbb{P}(x : x_0 = \Gamma(a_{R-1}), \ldots, x_{R-1} = \Gamma(a_0)).\end{aligned} \tag{11}$$

If the set of equalities (11) holds for some $R = R_0$, we say that CSPR holds for R_0. In this case, by taking appropriate marginals, the equalities also hold for all positive integers $R \leq R_0$.

Now, for discussing CSPR for $R = 2$, it is convenient to introduce the following notation. Let $[ab]_k$ be the event $\{x : x_k = a,\ x_{k+1} = b\}$. Since $\mathbb{P}$ is translation invariant, we have $\mathbb{P}([ab]_k) = \mathbb{P}([ab]_0)$ for all $k \in \mathbb{Z}$ and $a, b \in \mathcal{A}$. Therefore, CSPR for $R = 2$ reduces to

$$\forall a, b \in \mathcal{A}: \quad \mathbb{P}([ab]_0) = \mathbb{P}([\Gamma(b)\Gamma(a)]_0). \tag{12}$$

This equality implies CSPR for $R = 1$: $\mathbb{P}([a]_0) = \mathbb{P}([\Gamma(a)]_0)$ for $a \in \mathcal{A}$, where $[a]_0 = \{x : x_0 = a\}$.

We want to test the hypothesis H_0: CSPR holds for $R = 2$. In order to construct such a test, it is useful to introduce the following quantities:

$$f = (f(a, b) : (a, b) \in \mathcal{A}^2) \quad \text{where } f(a, b) := \mathbb{P}([ab]_0) - \mathbb{P}([\Gamma(b)\Gamma(a)]_0). \tag{13}$$

From (12), CSPR for $R = 2$ is satisfied if and only if $f = 0$.

We remark that 4 of the above 16 quantities $f(a, b)$ vanish. More precisely, whenever $(a, b) = (c, \Gamma(c))$ for some $c \in \mathcal{A}$, we see that $f(a, b) = 0$. Moreover, among the remaining 12 terms, only 5 are meaningful since $f(a, b) = -f(\Gamma(b), \Gamma(a))$ for any $a, b \in \mathcal{A}$, and $\sum_{c\in\mathcal{A}} f(a, c) = \sum_{c\in\mathcal{A}} f(c, a)$ for all $a \in \mathcal{A}$. In the following, we fix an index set $\mathcal{K} = \{(A, A), (A, C), (A, G), (C, A), (C, C)\}$ for 5 of these values and gather them together into a vector $f^{\mathcal{K}} := (f(a, b) :\ (a, b) \in \mathcal{K})$. Using this alternative representation, the null hypothesis H_0 is satisfied if and only if $f^{\mathcal{K}} = 0$.

3.2 Covariances and the Central Limit Theorem

Here, we present some results that are essential for developing an asymptotic test for the hypothesis H_0: CSPR holds for $R = 2$, in the setting of Gibbs distributions.

Let $\mathbb{P} = \mathbb{P}_\Psi$ be Gibbsian for some $\Psi \in F_\theta$, with $\theta \in (0, 1)$ fixed. We begin by giving a simple computation. Let $\mathbb{E} = \mathbb{E}_\Psi$ denote the expectation operator associated with $\mathbb{P}_\Psi$. A function $g \in F_\theta$ is said to be of zero mean if $\mathbb{E}(g) = 0$.

In this section we assume $\varphi^1, \ldots, \varphi^l$ are zero mean functions in F_θ and set $\varphi = (\varphi^1, \ldots, \varphi^l)$. We shall consider $X_i^k := \varphi^k \circ T^i$ for $i \geq 0$ and $k = 1, \ldots, l$, and define for $n \geq 1$,

$$S_n^k := \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i^k \,. \tag{14}$$

Proposition 3 *The limits*

$$\begin{aligned} &\Sigma^\varphi(k, j) = \lim_{n\to\infty} \mathbb{E}_\Psi(S_n^k\, S_n^j) \quad \textit{exist for all } k, j \in \{1, \ldots, l\} \textit{ and} \\ &\Sigma^\varphi(k, j) = \mathbb{E}_\Psi\left(X_0^k\, X_0^j\right) + \sum_{i=1}^{\infty} \mathbb{E}_\Psi\left(X_0^k\, X_i^j\right) + \sum_{i=1}^{\infty} \mathbb{E}_\Psi\left(X_0^j\, X_i^k\right). \end{aligned} \tag{15}$$

The matrix $\Sigma^\varphi = (\Sigma^\varphi(k, j) : k, j \in \{1, \ldots, l\})$ is symmetric and semi-positive definite.

Moreover the convergence of the two summations on the right-hand side of (15) *occurs at a geometric rate, more precisely,*

$$\exists\, \bar\delta < \infty,\ \xi \in (0, 1),\ \forall k, j = 1, \ldots, l,\ \forall i \geq 1: \quad \left|\mathbb{E}_\Psi(X_0^k X_i^j)\right| \leq \bar\delta \xi^i. \tag{16}$$

Proof By expanding the terms in the sum and using the translation invariance property $\mathbb{E}(X_i^k\, X_r^j) = \mathbb{E}(X_0^k\, X_{r-i}^j)$ for all $k, j \in \{1, \ldots, l\}$ and $i < r$, we get

$$\begin{aligned} \mathbb{E}(S_n^k S_n^j) &= \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}\left(X_i^k\, X_i^j\right) + \frac{1}{n} \sum_{i=1}^{n-1} \sum_{r=0}^{i-1} \mathbb{E}\left(X_i^k\, X_r^j\right) + \frac{1}{n} \sum_{i=1}^{n-1} \sum_{r=0}^{i-1} \mathbb{E}\left(X_i^j\, X_r^k\right) \\ &= \mathbb{E}\left(X_0^k\, X_0^j\right) + \frac{1}{n} \sum_{i=1}^{n-1} (n-i)\mathbb{E}\left(X_0^k\, X_i^j\right) + \frac{1}{n} \sum_{i=1}^{n-1} (n-i)\mathbb{E}\left(X_0^j\, X_i^k\right). \end{aligned}$$

Since $\varphi^k \in F_\theta$ for each k, the exponential cluster property of Gibbs measures (see Property 1.26 on Page 23 in [3]) guarantees the existence of $\delta < \infty$, and $\xi \in (0, 1)$ only depending on θ and Ψ, such that for all $k, j \in \{1, \ldots, l\}$,

$$\left|\mathbb{E}(X_0^k\, X_i^j)\right| = \left|\mathbb{E}\big(\varphi^k \cdot (\varphi^j \circ T^i)\big)\right| \leq \delta \|\varphi^k\|_\theta \|\varphi^j\|_\theta \xi^i.$$

As a consequence, (16) is satisfied. Hence all the series are absolutely convergent. Moreover, since $\sum_{i=0}^{\infty} i\mathbb{E}(X_0^k X_i^j)$ is finite, the Cesàro mean of $i\mathbb{E}(X_0^k X_i^j)$ converges to zero and we obtain the formula

$$\lim_{n\to\infty} \mathbb{E}(S_n^k\, S_n^j) = \mathbb{E}\left(X_0^k\, X_0^j\right) + \sum_{i=1}^{\infty} \mathbb{E}\left(X_0^k\, X_i^j\right) + \sum_{i=1}^{\infty} \mathbb{E}\left(X_0^j\, X_i^k\right) = \Sigma_{kj}^\varphi.$$

Finally, we see from this explicit expression that the matrix Σ^φ is symmetric and semi-positive definite because each matrix $(\mathbb{E}(S_n^k S_n^j) : k, j \in \{1, \ldots, l\})$ is a covariance matrix. Hence, its limit Σ^φ is also semi-positive definite. □

Next, we show a Central Limit Theorem for random vectors in the Gibbs framework, which is a corollary of the Central Limit Theorem given in [5].

Proposition 4 *If $\Sigma^\varphi = (\Sigma^\varphi(k,j) : k,j = 1,\ldots,l)$ given by (15) is positive definite then the vector process $Z_n := (S_n^1,\ldots,S_n^l)$ (where S_n^k is given in (14)) converges in distribution to the multivariate normal vector $\mathcal{N}(0,\Sigma^\varphi)$.*

Proof We recall that the Central Limit Theorem shown in [5] says that if a function $g \in F_\theta$ is of zero mean and

$$\sigma(g)^2 \neq 0 \quad \text{where } \sigma(g)^2 := \lim_{n\to\infty} \frac{1}{n}\,\mathbb{E}\left(\left(\sum_{i=0}^{n-1} g\circ T^i\right)^2\right),$$

then

$$\frac{1}{\sqrt{n}}\left(\sum_{i=0}^{n-1} g\circ T^i\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma(g)^2). \tag{17}$$

Since F_θ is Banach, for all $\alpha = (\alpha_1,\ldots,\alpha_l)$ the function $\varphi_\alpha = \sum_{k=1}^l \alpha_k\varphi^k$ is in F_θ. Since φ_α has zero mean and $\sum_{i=0}^{n-1}(\varphi_\alpha\circ T^i)/\sqrt{n} = \alpha' Z_n$, where a' denotes the transpose of a vector a, the Central Limit Theorem (17) gives

$$\alpha' Z_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma_\alpha^2), \tag{18}$$

where

$$\sigma_\alpha^2 = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left(\left(\sum_{i=0}^{n-1}\varphi_\alpha\circ T^i\right)^2\right),$$

provided that $\sigma_\alpha^2 \neq 0$. Now, $\sigma_\alpha^2 \neq 0$ will be obtained as a consequence of the fact that

$$\sigma_\alpha^2 = \alpha'\Sigma^\varphi\alpha = \sum_{k=1}^l\sum_{j=1}^l \alpha_k\alpha_j\Sigma^\varphi(k,j). \tag{19}$$

This together with the assumption that Σ^φ is positive definite allows us to determine that $\sigma_\alpha^2 = \alpha'\Sigma^\varphi\alpha > 0$ for all $\alpha \neq 0$. Furthermore, (18) and (19) yield

$$\forall s\in\mathbb{R}:\quad \lim_{n\to\infty}\mathbb{E}(e^{is\alpha' Z_n}) = e^{-\frac{1}{2}s^2\sigma_\alpha^2} = e^{-\frac{1}{2}s^2\alpha'\Sigma^\varphi\alpha},$$

which is the characteristic function of an $\mathcal{N}(0,\alpha'\Sigma^\varphi\alpha)$ random vector. Convergence of Z_n in distribution to an $\mathcal{N}(0,\Sigma^\varphi)$ random vector then follows from Lévy's continuity theorem.

It only remains to prove (19). Notice that

$$\sum_{i=0}^{n-1}\varphi_\alpha\circ T^i = \sum_{k=1}^l\alpha_k\sum_{i=0}^{n-1}\varphi^k\circ T^i = \sum_{k=1}^l\alpha_k\sum_{i=0}^{n-1}X_n^k = \sqrt{n}\sum_{k=1}^l\alpha_k S_n^k$$

which implies that

$$\sigma_\alpha^2 = \lim_{n\to\infty}\mathbb{E}\left(\left(\sum_{k=1}^l\alpha_k S_n^k\right)^2\right) = \sum_{k=1}^l\sum_{j=1}^l\alpha_k\alpha_j\lim_{n\to\infty}\mathbb{E}(S_n^k S_n^j).$$

Finally, Proposition 3 asserts that $\lim_{n\to\infty}\mathbb{E}(S_n^k S_n^j) = \Sigma^\varphi(k,j)$ and hence the result follows. □

4 Testing Under the Gibbsian Assumption

4.1 A Statistical Test

Recall that the hypothesis H_0: CSPR for $R = 2$, is equivalent to $f = 0$, where f was defined in (13). Let us introduce estimators of the various quantities involved in testing this. For any finite observed sequence $X = (X_0, \ldots, X_{n-1})$, let

$$\widehat{f}_n(a,b) := \frac{N_n(a,b)}{n} - \frac{N_n(\Gamma(b),\Gamma(a))}{n},$$

where

$$N_n(a,b) := \#\{k \in \{0,\ldots,n-1\} : (X_k, X_{k+1}) = (a,b)\}$$

counts the number of occurrences of the pattern ab in the sequence. Note that we treat the sequence X as though it were circular with X_{n-1} connected to X_0, so that $X_n \equiv X_0$.

We shall show that one appropriate statistic for assessing this test is

$$\widehat{\eta}_n = n\, \widehat{f}_n^{\mathcal{K}\prime}\, \widehat{V}_n^{-1}\, \widehat{f}_n^{\mathcal{K}},$$

where $\widehat{f}_n^{K} = (\widehat{f}_n(a,b) : (a,b) \in \mathcal{K})$ is a consistent unbiased estimator of $f^{\mathcal{K}}$ and $\widehat{V}_n$ is a consistent biased estimator of the asymptotic covariance matrix V of $\sqrt{n}\,\widehat{f}_n^{\mathcal{K}}$ which we shall define shortly. Furthermore, we shall prove that $\widehat{\eta}_n$ converges asymptotically in distribution to a χ^2_5 random variable. Then, sufficiently large values of $\hat{\eta}_n$ will identify sequences that fail to comply with CSPR for $R = 2$.

More precisely, the test is set up as follows:

$$\text{Reject } H_0 \quad \text{if } \widehat{\eta}_n \geq s,$$

where s is some threshold to be chosen. If α is the type I error desired for the test (for instance $\alpha = 0.05$ or 0.01), then we require that

$$\mathbb{P}_{H_0}(\text{reject } H_0) = \mathbb{P}_{H_0}(\widehat{\eta}_n \geq s) \leq \alpha,$$

either exactly or asymptotically. Doing this exactly is not feasible in the current setting, but $\hat{\eta}_n \xrightarrow[n\to\infty]{d} \chi^2_5$, where $\xrightarrow[n\to\infty]{d}$ denotes convergence in distribution. Thus the threshold s can be fixed asymptotically by appealing to the χ^2 distribution on 5 degrees of freedom. We merely have to set s to the $1-\alpha$ quantile $\chi^2_{5,1-\alpha}$ of the χ^2_5 distribution.

4.2 Asymptotics of the Test Statistic

In order to construct this asymptotic test, we make the further assumption that the distribution $\mathbb{P} = \mathbb{P}_\Psi$ is Gibbsian for some energy $\Psi \in F_\theta$, where $\theta \in (0,1)$. Recall that $\mathbb{P}$ is ergodic. Let $\mathbb{E}$ denote the mean expected value operator associated with $\mathbb{P}$.

Firstly,

$$\begin{aligned}\mathbb{E}(\widehat{f}_n^{\mathcal{K}}) &= \mathbb{E}\left(\frac{N_n(a,b)}{n} - \frac{N_n(\Gamma(b),\Gamma(a))}{n}\right)\\ &= \frac{1}{n}\left(n\mathbb{P}([ab]_0) - n\mathbb{P}([\Gamma(b)\Gamma(a)]_0)\right) = f(a,b).\end{aligned}$$

From ergodicity the law of large numbers holds and so

$$\lim_{n\to\infty} \frac{N_n(a,b)}{n} = \mathbb{P}([ab]_0) \quad \mathbb{P}\text{-a.e. and hence } \lim_{n\to\infty} \widehat{f}_n^{\mathcal{K}} = f^{\mathcal{K}} \quad \mathbb{P}\text{-a.e.}$$

Therefore $\widehat{f}_n^{\mathcal{K}}$ is a consistent, unbiased estimator of $f^{\mathcal{K}}$.

Next define

$$\varphi = (\varphi^{a,b} : (a,b) \in \mathcal{A}^2) \quad \text{where } \varphi^{a,b} := \mathbb{1}_{[ab]_0} - \mathbb{P}([ab]_0),$$

where, as usual, $\mathbb{1}_B$ is the characteristic function of the set B. Observe that for all $(a,b) \in \mathcal{A}^2$ we have $\varphi^{a,b} \in F_\theta$. For $i \geq 0$ and $n \geq 1$, define

$$\forall\, (a,b) \in \mathcal{A}^2\colon \quad X_i^{a,b} = \varphi^{a,b} \circ T^i \quad \text{and} \quad S_n^{a,b} = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i^{a,b}.$$

A simple calculation that takes advantage of the T-invariance of $\mathbb{P}$ gives

$$X_i^{a,b} = (\mathbb{1}_{[ab]_i} - \mathbb{P}([ab]_0)) \quad \text{and} \quad S_n^{a,b} = \frac{1}{\sqrt{n}} \left(N_n(a,b) - n\mathbb{P}([ab]_0)\right).$$

A straight forward application of Proposition 3 can be used to show existence of the matrix $\Sigma^\varphi = (\Sigma^\varphi(a,b;c,d) : (a,b), (c,d) \in \mathcal{A}^2)$, whose elements are defined by

$$\Sigma^\varphi(a,b;c,d) := \lim_{n\to\infty} \mathbb{E}(S_n^{a,b}\, S_n^{c,d}).$$

(Note that for simplicity we write $\Sigma^\varphi(a,b;c,d)$ rather than $\Sigma^\varphi((a,b),(c,d))$.) Furthermore, using (15) from the same lemma, we can see that

$$\Sigma^\varphi(a,b;c,d) = \mathbb{E}\left(X_0^{a,b}\, X_0^{c,d}\right) + \sum_{i=1}^{\infty} \mathbb{E}\left(X_0^{a,b}\, X_i^{c,d}\right) + \sum_{i=1}^{\infty} \mathbb{E}\left(X_0^{c,d}\, X_i^{a,b}\right),$$

and that Σ^φ is symmetric and semi-positive definite.

Further simple computations enable us to write the elements of Σ^φ explicitly as

$$\begin{aligned}\Sigma^\varphi(a,b;c,d) = {} & \mathbb{P}([ab]_0 \cap [cd]_0) - \mathbb{P}([ab]_0)\mathbb{P}([cd]_0) \\ & + \sum_{k=1}^{\infty} [\mathbb{P}([ab]_0 \cap [cd]_k) - \mathbb{P}([ab]_0)\mathbb{P}([cd]_0)] \\ & + \sum_{k=1}^{\infty} [\mathbb{P}([cd]_0 \cap [ab]_k) - \mathbb{P}([ab]_0)\mathbb{P}([cd]_0)].\end{aligned}$$

As a corollary to Proposition 4, we obtain:

Proposition 5 *Assume Σ^φ is positive definite. Then, the joint distribution of the counts $N_n(a,b)$ asymptotically satisfy*

$$\left(\frac{N_n(a,b) - n\mathbb{P}([ab]_0)}{\sqrt{n}} : (a,b) \in \mathcal{A}^2\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma^\varphi).$$

The following is then obtained by taking appropriate marginals in the preceding result.

Corollary 6 *Assume Σ^φ is positive definite. We have*

$$\sqrt{n}(\widehat{f}_n^{\mathcal{K}} - f^{\mathcal{K}}) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V),$$

where the covariance matrix $V = (V(a,b;c,d) : (a,b),(c,d) \in \mathcal{K})$ *is given by*

$$V(a,b;c,d) = \Sigma^{\varphi}(a,b;c,d) + \Sigma^{\varphi}(\Gamma(b),\Gamma(a);\Gamma(d),\Gamma(c)) - \Sigma^{\varphi}(\Gamma(b),\Gamma(a);c,d) - \Sigma^{\varphi}(a,b;\Gamma(d),\Gamma(c)).$$

From this result we conclude under the hypothesis $H_0 : f^{\mathcal{K}} = 0$ that

$$\sqrt{n}\,\widehat{f}_n^{\mathcal{K}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,V).$$

As a consequence, $n\,\widehat{f}_n^{\mathcal{K}\prime} V^{-1}\,\widehat{f}_n^{\mathcal{K}}$ converges in distribution to a χ^2 distribution on 5 degrees of freedom, provided that V is positive definite.

Observe that $\widehat{f}_n^{\mathcal{K}} = \frac{1}{n}\lambda N_n$, where $N_n = (N_n(a,b) : (a,b) \in \mathcal{A}^2)$ and $\Lambda = (\Lambda(a,b;c,d) : (a,b) \in \mathcal{K}, (c,d) \in \mathcal{A}^2)$ is the 5×16 matrix given by

$$\Lambda(a,b;c,d) := \begin{cases} 1, & \text{if } (a,b) = (c,d), \\ -1, & \text{if } (a,b) = (\Gamma(d),\Gamma(c)), \\ 0, & \text{otherwise.} \end{cases}$$

The covariance matrix V may then be written as $V = \Lambda\,\Sigma^{\varphi}\,\Lambda'$. Since Λ is of full rank, V is positive definite whenever Σ^{φ} is positive definite.

Proposition 7 *Assume that* Σ^{φ} *is positive definite. Then, there exists a consistent estimator* $\widehat{V}_n$ *of* V *such that* $\widehat{\eta}_n := n\,\widehat{f}_n^{\mathcal{K}\prime}\widehat{V}_n^{-1}\widehat{f}_n^{\mathcal{K}}$ *converges in distribution to a* χ^2_5 *random variable.*

The proof of this proposition will be a consequence of the following constructions and intermediate results.

In order to define the estimator $\widehat{V}_n$ of the covariance matrix V, we first require an estimator of Σ^{φ}. Let $\widehat{\Sigma}_{n,m} = (\widehat{\Sigma}_{n,m}(a,b;c,d) : (a,b),(c,d) \in \mathcal{A}^2)$, where

$$\begin{aligned}\widehat{\Sigma}_{n,m}(a,b;c,d) := {} & \frac{N_n^{(0)}(a,b;c,d)}{n} - \frac{N_n(a,b)}{n}\cdot\frac{N_n(c,d)}{n} \\ & + \sum_{i=1}^{m}\left(\frac{N_n^{(i)}(a,b;c,d)}{n} - \frac{N_n(a,b)}{n}\cdot\frac{N_n(c,d)}{n}\right) \\ & + \sum_{i=1}^{m}\left(\frac{N_n^{(i)}(c,d;a,b)}{n} - \frac{N_n(a,b)}{n}\cdot\frac{N_n(c,d)}{n}\right) \end{aligned} \tag{20}$$

and

$$N_n^{(i)}(a,b;c,d) := \#\{j \in \{0,\dots,n-1\} : (X_j, X_{j+1}, X_{j+i}, X_{j+i+1}) = (a,b,c,d)\}.$$

Recall that we treat genome sequences as circular, so that $X_{n+i} = X_i$ for $i = 0,\dots,n-1$.

Now, from the law of large numbers for Gibbs measures,

$$\lim_{n\to\infty}\frac{N_n^{(i)}(a,b;c,d)}{n} = \mathbb{P}([ab]_0 \cap [cd]_i) \quad \mathbb{P}\text{-a.e.}$$

and so

$$\lim_{n\to\infty}\widehat{\Sigma}_{n,m}(a,b;c,d) = \Sigma^{\varphi}_{(m)}(a,b;c,d) \quad \mathbb{P}\text{-a.e.}$$

where

$$\Sigma^{\varphi}_{(m)}(a,b;c,d) = \mathbb{P}([ab]_0 \cap [cd]_0) - \mathbb{P}([ab]_0)\mathbb{P}([cd]_0) + \sum_{i=1}^{m} [\mathbb{P}([ab]_0 \cap [cd]_i) - \mathbb{P}([ab]_0)\mathbb{P}([cd]_0)] + \sum_{i=1}^{m} [\mathbb{P}([cd]_0 \cap [ab]_i) - \mathbb{P}([ab]_0)\mathbb{P}([cd]_0)]. \tag{21}$$

However,

$$\Sigma^{\varphi}(a,b;c,d) = \lim_{m\to\infty} \Sigma^{\varphi}_{(m)}(a,b;c,d). \tag{22}$$

Now, we claim that there exists a sequence $(m(n) : n \geq 1)$ which monotonically increases to ∞ such that

$$\widehat{\Sigma}_{n,m(n)}(a,b;c,d) \xrightarrow[n\to\infty]{\mathbb{P}} \Sigma^{\varphi}(a,b;c,d),$$

where $\xrightarrow[n\to\infty]{\mathbb{P}}$ is used to denote convergence in probability. To show this, first recall that convergence in probability is metrizable by some metric D, for instance, $D(g,h) = \mathbb{E}((|g-h|)/(1+|g-h|))$. Since $\lim_{n\to\infty} D(\widehat{\Sigma}_{n,m}(a,b;c,d), \Sigma^{\varphi}_{(m)}(a,b;c,d)) = 0$, we deduce that for all $m \geq 1$, there exists a positive integer $N(m)$ satisfying

$$\forall n \geq N(m), \forall k \in \{1,\ldots,m\}: \quad D(\widehat{\Sigma}_{n,k}(a,b;c,d), \Sigma^{\varphi}_{(k)}(a,b;c,d)) \leq 1/m.$$

The sequence $(N(m) : m \geq 1)$ is increasing. Now for all $n < N(1)$, we set $m(n) = 1$ and, for $n \geq N(1)$, we define $m(n) = \sup\{m : N(m) \leq n\}$. By construction $m(n)$ increases with n. On the other hand, since $m(n) \geq m$ for all $n \geq N(m)$, we have $\lim_{n\to\infty} m(n) = \infty$. By construction we have

$$\forall n \geq N(1): \quad D(\widehat{\Sigma}_{n,m(n)}(a,b;c,d), \Sigma^{\varphi}_{(m(n))}(a,b;c,d)) \leq 1/m(n),$$

and the claim follows by letting $m \to \infty$ and taking account of (22).

Let $\widehat{V}_{n,m} = (\widehat{V}_{n,m}(a,b;c,d) : (a,b),(c,d) \in \mathcal{K})$, where

$$\widehat{V}_{n,m}(a,b;c,d) := \widehat{\Sigma}_{n,m}(a,b;c,d) + \widehat{\Sigma}_{n,m}(\Gamma(b),\Gamma(a);\Gamma(d),\Gamma(c)) - \widehat{\Sigma}_{n,m}(\Gamma(b),\Gamma(a);c,d) - \widehat{\Sigma}_{n,m}(a,b;\Gamma(d),\Gamma(c)).$$

Since $\widehat{\Sigma}_{n,m(n)}$ converges in probability to Σ^{φ}, it follows that $\widehat{V}_n := \widehat{V}_{n,m(n)}$ must converge in probability to V. Furthermore, $\widehat{V}_n^{-1}$ will also converge to V^{-1} in probability, provided that V is positive definite. To summarize, we have shown that $\widehat{f}_n^{\mathcal{K}}$ is a consistent (unbiased) estimator of $f^{\mathcal{K}}$ while $\widehat{V}_n$ constitutes a consistent (but biased) estimator of V.

Next, we prove the asymptotic behavior of $\widehat{\eta}_n$ claimed in Proposition 7. Recall that V is positive definite since Σ^{φ} is positive definite. We have shown that $\sqrt{n}\widehat{f}_n^{\mathcal{K}}$ converges in distribution to $\mathcal{N}(0,V)$ and $\widehat{V}_n^{-1}$ converges in probability to V^{-1}. This implies that

$$n\widehat{f}_n^{\mathcal{K}\prime}\widehat{V}_n^{-1}\widehat{f}_n^{\mathcal{K}} - n\widehat{f}_n^{\mathcal{K}\prime}V^{-1}\widehat{f}_n^{\mathcal{K}} = n\widehat{f}_n^{\mathcal{K}\prime}\left(\widehat{V}_n^{-1} - V^{-1}\right)\widehat{f}_n^{\mathcal{K}} \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Combining this with the aforementioned fact that $n\widehat{f}_n^{\mathcal{K}\prime}V^{-1}\widehat{f}_n^{\mathcal{K}}$ converges in distribution to a χ^2_5 random variable, we see that $\widehat{\eta}_n$ converges in distribution to a χ^2_5 distribution as $n \to \infty$ and hence Proposition 7 has been proved.

4.3 Practical Considerations

When computing the statistic $\widehat{\eta}_n$ for a real genomic sequence, the parameter n is dictated by the length (in bases) of the genome under study. However, it is necessary to choose an appropriate value for the parameter m and this is not so straight forward. The regime $(m(n))$ derived in the previous subsection is not unique. In fact, the convergence results in the preceding subsection remain valid for any sequence that converges to ∞ more slowly than $(m(n))$. Consequently, any value $m(n) \ll n$ should satisfy the consistency criterion. On the other hand, the exponential cluster property of Gibbs measures implies that terms of the series in (21) should tend geometrically toward zero and the same should also be true for terms of the series in (20) when n is large. Eventually, there should come a point after which the terms of (20) will constitute noise of the estimators and these should be ignored. Consistency of the estimator $\widehat{\Sigma}_{n,m}$, together with the exponentially fast convergence of $\Sigma^{\varphi}_{(m)}$ to Σ^{φ}, means that satisfactory results should be obtainable by setting $m(n)$ small relative to n when computing $\widehat{\eta}_n$.

In our implementation of this test of CSPR for dinucleotides, we chose $m(n)$ to be the smallest value of i such that

$$\left|\frac{N_n^{(i)}(a,b,a,b)}{n} - \left(\frac{N_n(a,b)}{n}\right)^2\right| \leq 0.01\left(\frac{N_n(a,b)}{n} - \left(\frac{N_n(a,b)}{n}\right)^2\right) = 0.01\,\widehat{\mathrm{Var}}([ab]_0) \quad \forall (a,b) \in \mathcal{A}^2.$$

Here, $\widehat{\mathrm{Var}}([ab]_0)$ denotes a consistent estimator of the variance of the frequency of the dinucleotide ab in a genome sequence. Since $|\frac{N_n(a,b)}{n} - (\frac{N_n(a,b)}{n})^2|$ is typically on the order of 0.06, we conjecture that truncating the covariance estimators at the point where the sums composing the estimators change by less than 1% of $\widehat{\mathrm{Var}}([ab]_0)$ is reasonable.

Finally, some numerical experimentation leads us to conjecture that the test is highly powerful. Markov chains constitute a subset of the Gibbsian processes. Simulating genomic sequences from Markov chains which fail to comply with CSPR yields a rejection rate of 100% at the 5% significance level. Performing the same experiments on Markov chains that do satisfy CSPR results in a rejection rate close to the α chosen for the test, as one would expect. We obtained similar results for genomic sequences generated as realizations of Markov random fields. A Markov random field with maximal clique size k is equivalent to a Gibbs measure whose energy Ψ takes the form

$$\Psi(x) = \sum_{j=1}^{k}\sum_{i=0}^{n-1} \psi^{(j)}(x_i, \ldots, x_{i+j-1}). \tag{23}$$

In other words, the energy is a linear sum of functions $\Psi^{(j)}$, each of which depends on cliques of size j, that is, sets of j mutually neighboring sites. Simulations of such sequences can be produced using the Gibbs sampler and Ψ will be $\overline{\Gamma} \circ \mathcal{I}$-invariant if $\Psi^{(j)}$ is $\overline{\Gamma} \circ \mathcal{I}$-invariant for all $j = 1, \ldots, k$. In our numerical experiments, we simulated sequences from Markov random fields having maximal clique sizes of 3 and 4, using an energy which is not $\overline{\Gamma} \circ \mathcal{I}$-invariant.

4.4 Application of the Test

Although successful tests of CSPR in genomic sequences have already been conducted using both empirical and rigorous methods (for instance, see [1, 7, 10, 12]), we would like to test for CSPR for $R = 2$ under a Gibbsian hypothesis.

Table 1 Summary statistics for the lengths and GC-contents of a collection of 1049 complete bacterial genome sequences obtained from the GenBank repository

Property	First Quartile	Median	Third Quartile	Mean	Std Deviation
Length	1906322	2976212	4603746	3317355	1759175
GC-content	0.3769	0.4753	0.6035	0.4839	0.1326

We considered a set of 1049 complete bacterial genome sequences obtained from the GenBank 'genomes' repository. Length and GC-content statistics for the set of genomes are shown in Table 1.

To correct for multiple testing of a large number of genomes, we used the Holm-Bonferroni method, whose application posed no difficulties since p-values for the test statistic are readily obtainable from the χ^2_5 distribution. Of the 1049 genomes tested at the $\alpha = 0.01$ level of significance, the null was accepted in 410 cases and was rejected in the remaining 639 genomes. We found no relationship between GC-content, genome length and rejection of the null.

Note that the Gibbsian assumption determines the form of the covariance matrix Σ^{φ}, which exists as a consequence of the exponential cluster property. Any genomic sequence that departs significantly from exponential clustering will give rise to an $\widehat{\eta}_n$ far out in the tail of the χ^2_5 distribution, since Σ^{φ} is likely to be near singular in such cases. A caveat with the test proposed here is that when a sequence is rejected, the reason for its rejection is unclear. Rejection could be due to either violation of CSPR or lack of compliance with the Gibbsian or translation invariance assumptions. In any case, further examination is warranted in order to discover why a particular sequence is rejected. On the other hand, given the test's apparently high sensitivity to departures from a Gibbsian structure, sequences for which the null hypothesis is accepted must comply much more closely to CSPR and exhibit a much stronger Gibbsian-like structure than those that are rejected.

Acknowledgements This work was supported by the Center for Mathematical Modeling (CMM) Basal CONICYT Program PFB 03. We would like to thank researchers and engineers in the Laboratory of Bioinformatics and Mathematics of the Genome in the CMM for providing assistance and advice. We are indebted to Dr Catherine Matias from CNRS, Laboratoire Statistique et Génome, Université Évry, for invaluable discussions. Finally, our thanks go to the referees for their useful comments.

References

1. Albrecht-Buehler, G.: Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc. Natl. Acad. Sci. USA **103**(47), 17828–17833 (2006)
2. Bell, S.J., Forsdyke, R.: Deviations from Chargaff's second parity rule correlate with direction of transcription. J. Theor. Biol. **197**, 63–76 (1999)
3. Bowen, R.: Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms. Lecture Notes in Mathematics, vol. 470, revised edn. Springer, Berlin (2008). With a preface by David Ruelle, Edited by Jean-René Chazottes
4. Chargaff, E.: Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia **6**(6), 201–209 (1950)
5. Coelho, Z., Parry, W.: Central limit asymptotics for shifts of finite type. Isr. J. Math. **69**(2), 235–249 (1990). doi:10.1007/BF02937307. http://www.springerlink.com/content/g415tk6310717655/
6. Forsdyke, R., Bell, S.J.: Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. Appl. Bioinformatics **3**(3), 3–8 (2004)

7. Hart, A., Martínez, S.: Statistical testing of Chargaff's second parity rule in bacterial genome sequences. Stoch. Models **27**(2), 1–46 (2011)
8. Kong, S.-G., Fan, W.-L., Chen, H.-D., Hsu, Z.-T., Zhou, N., Zheng, B., Lee, H.-C.: Inverse symmetry in complete genomes and whole-genome inverse duplication. PLoS ONE **4**(11), e7553 (2009)
9. Lobry, J.: Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. **40**(3), 326–330 (1995)
10. Mitchell, D., Bridge, R.: A test of Chargaff's second rule. Biochem. Biophys. Res. Commun. **340**(1), 90–94 (2006)
11. Powdel, B., Satapathy, S., Kumar, A., Jha, P., Buragohain, A., Borah, M., Ray, S.: A study in entire chromosomes of violation of the intra-strand parity of complementary nucleotides. DNA Res. **16**, 325–343 (2009)
12. Prabhu, V.: Symmetry observations in long nucleotide sequences. Nucleic Acids Res. **21**(12), 2797–2800 (1993)
13. Rudner, R., Karkas, J., Chargaff, E.: Separation of B. subtilis DNA into complementary strands. III. Direct analysis. Proc. Natl. Acad. Sci. USA **60**, 921–922 (1968)
14. Sueoka, N.: Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. **40**(3), 18–325 (1995)
15. Zhang, S., Huang, Y.: Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA. Bioinformatics **26**(4), 478–485 (2010)