Statistical Testing of Chargaff's Second Parity Rule in Bacterial Genomes

Andrew Hart   Servet martínez

El Centro de Modelamiento Matemático,
Universidad de Chile

7to Encuentro Regional de Probabilidad
y Estadística Matemática
Santa Fe, Argentina,
1–3 December 2010

Hart, A.G. and Martínez, S. (2011) Statistical testing of Chargaff's second parity rule in bacterial genome sequences. *Stoch. Models* **27(2)**, to appear.

Structure of DNA

- DNA strand: A sequence of nucleotides.
- Nucleotide: Building blocks of the genome. There are four types: *a*, *c*, *g*, *t*.
- DNA comprises 2 strands: The primary (or principal) and the complementary. The two strands together are called a duplex.
- Corresponding nucleotides on each strand forma base pair.
- Within each base pair, *a* bonds with *t* while *c* bonds with *g*.
- The complementary strand is read in the opposite direction to the principal strand.

$$\begin{array}{ccccc} & & \text{Principal strand} & & \\ 5' & \leftarrow & \text{GGGATCAAGTCCATCA} & \rightarrow & 3' \\ 3' & \leftarrow & \text{CCCTAGTTCAGGTAGT} & \leftarrow & 5' \\ & & \text{Complementary strand} & & \end{array}$$

Notation

- Set of nucleotides: $\mathcal{A} = \{A, C, G, T\}$.
- Involution: $\gamma : \mathcal{A} \to \mathcal{A}$, where $\gamma(A) = T$, $\gamma(C) = G$, $\gamma(G) = C$ and $\gamma(T) = A$.
- DNA sequence: $X = (X_m : m = 1, \ldots, L)$, where $x_m \in \mathcal{A}$.
- We treat sequences as circular so that $X_{L+m} = X_m$ for all $m = 1, \ldots, L$.
- Oligonucleotide: $X_m X_{m+1} \ldots X_{l-1} X_l$.
- Frequency of r-oligonucleotide:

  $$\nu^X(a_1, \ldots, a_r) := \frac{1}{L} \sum_{m=1}^{L} \mathbf{1}_{\{(X_m, \ldots X_{m+r-1}) = (a_1, \ldots, a_r)\}} ,$$

  for all $(a_1, \ldots, a_r) \in \mathcal{A}^r$, $1 \leq r \leq M$. $\mathbf{1}_B$ takes the value one if the condition B is satisfied and zero otherwise.
- $$\pi_a := \nu^X(a) \text{ and } P_{a,b} := \frac{\nu^X(a, b)}{\nu^X(a)}.$$

More Notation

- Complementary strand: $Y = (Y_m : m = 1, \ldots, L)$, where $Y_m \in \mathcal{A}$.
- For chemical reasons, X and Y are related by $Y_m = \gamma(X_{L-m+1})$ for $m = 1, \ldots, L$.
- Frequencies for Y are given by

$$\nu^Y(a_1, \ldots, a_r) := \frac{1}{L} \sum_{m=1}^{L} \mathbf{1}_{\{(Y_m,\ldots,Y_{m+r-1})=(a_1,\ldots,a_r)\}} ,$$

  for all $(a_1, \ldots, a_r) \in \mathcal{A}^r$, $1 \leq r \leq M$.
- Hence, for all $(a_1, \ldots, a_r) \in \mathcal{A}^r$, $1 \leq r \leq M$, we have

$$\nu^Y(a_1, \ldots, a_r) = \nu^X(\gamma(a_r), \ldots, \gamma(a_1)) .$$

- Mononucleotide and conditional dinucleotide distributions of Y:

$$\rho_a := \nu^Y(a) \text{ and } Q_{a,b} := \frac{\nu^Y(a,b)}{\nu^Y(a)} .$$

Chargaff's First Parity rule

- For all $a, b \in \mathcal{A}$,

$$\rho_a = \pi_{\gamma(a)} \text{ and } \rho_a Q_{a,b} = \pi_{\gamma(b)} P_{\gamma(b),\gamma(a)} .$$

Chargaff's First Parity Rule.

In any DNA duplex, the number of A nucleotides is the same as the number of T nucleotides, while the number of C nucleotides is the same as the number of G nucleotides.

chargaff's Second Parity Rule

Chargaff's Second Parity Rule (CSPR).

On a DNA strand, the frequency of a short oligonucleotide is the same as the frequency of its reverse complement.

CSPR means that, for all $r \ll L$, $(a_1, \ldots, a_r) \in \mathcal{A}^r$,

$$\nu^X(a_1, \ldots, a_r) = \nu^X(\gamma(a_r), \ldots, \gamma(a_1)). \tag{1}$$

CSPR for $r = r_0$.

We say that CSPR holds for $r = r_0$ if (1) holds for $r = r_0$.

- if CSPR holds for $r = r_0$, then it also holds for all $r < r_0$.
- For $r = 1$, CSPR means that $\pi = \rho$, or $\pi_A = \pi_T$ and $\pi_C = \pi_G$.
- For $r = 2$, CSPR means that $\rho = \pi$ and $Q = P$, or equivalently,

$$\pi_a P_{a,b} = \pi_{\gamma(b)} P_{\gamma(b),\gamma(a)}, \ \ a, b \in \mathcal{A}.$$

A Matrix characterisation of CSPR for Dinucleotides

- Assume the order $A < C < G < T$.
- Let θ be the set of 4×4 positive stotchastic matrices,

$$P = \begin{bmatrix} P_{A,A} & P_{A,C} & P_{A,G} & P_{A,T} \\ P_{C,A} & P_{C,C} & P_{C,G} & P_{C,T} \\ P_{G,A} & P_{G,C} & P_{G,G} & P_{G,T} \\ P_{T,A} & P_{T,C} & P_{T,G} & P_{T,T} \end{bmatrix}.$$

Proposition

Chargaff's second parity rule holds for $r = 2$ if and only if the matrix P takes the form

$$\begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & 1-(\beta_1+\beta_2+\beta_3) \\ \zeta\beta_6 & \beta_4 & 1-(\zeta\beta_6+\beta_4+\zeta\beta_3) & \zeta\beta_3 \\ \zeta\beta_5 & 1-(\zeta\beta_5+\beta_4+\zeta\beta_2) & \beta_4 & \zeta\beta_2 \\ 1-(\beta_5+\beta_6+\beta_1) & \beta_5 & \beta_6 & \beta_1 \end{pmatrix}$$

where $\zeta \in (0, \infty)$ and $\beta_1, \ldots, \beta_6$ represent values in $(0, 1)$ such that P is a strictly positive stochastic matrix.

Uniformly distributed Stochastic Matrices

- set $\mathcal{A}_3 = \{A, C, G\}$ and $\mathcal{A}_2 = \{A, C\}$.
- The $n-$simplex is $\mathcal{S}_n = \{(s_1, \dots, s_{n+1}) \in \mathbb{R}_+^{n+1} : \sum_{i=1}^{n+1} s_i = 1\}$.
- The interior of the n dimensional ℓ^1 unit ball intersected with the positive orthant is $\mathcal{C}_n = \{(s_1, \dots, s_n) \in \mathbb{R}_+^{n} : \sum_{i=1}^{n} s_i < 1\}$.
- $\overline{P} := (P_{a,b} : (a, b) \in \mathcal{A} \times \mathcal{A}_3) \in \mathcal{C}_3^{\mathcal{A}}$.
- $\vec{X} = (X_1, X_2, X_3, X_4)$ taking values in $\mathcal{S}_3$ is Dirichlet$(1, 1, 1, 1)$ distributed if $\overline{X} = (X_1, X_2, X_3)$, which takes values in $\mathcal{C}_3$, has probability density function f given by $f_{\overline{X}}(x_1, x_2, x_3) = 6 \quad \text{for } (x_1, x_2, x_3) \in \mathcal{C}_3$.
- The volume of $\mathcal{C}_3$ relative to Lebesgue measure is $\mathrm{Vol}(\mathcal{C}_3) = 1/6$.
- Taking the distribution of $P \in \Theta$ to be uniform is equivalent to taking $P \sim (\text{Dirichlet}(1, 1, 1, 1))^{\otimes 4}$.
- Let $\mathbb{P}_\theta$ denote this probability measure.

CSPR for Dinucleotides

Let Θ_2 be the set of $P \in \Theta$ having the form prescribed by the Proposition.
Let $J_7 = \mathcal{A}_2 \times \mathcal{A}_3 \cup \{(G, A)\}$ and define $\widetilde{P} = (P_{a,b} : (a, b) \in J_7)$.
Then, Θ_2 is the set of $P \in \Theta$ satisfying the set of constraints $P_{G,G} = f_1(\widetilde{P}),\; P_{G,C} = f_2(\widetilde{P}),\; P_{T,G} = f_3(\widetilde{P}),\; P_{T,C} = f_4(\widetilde{P}),\; P_{T,A} = f_5(\widetilde{P})$, where

$$
\begin{aligned}
f_1(\widetilde{P}) &= P_{C,C} \\
f_2(\widetilde{P}) &= 1 - P_{G,A} - f_1(\widetilde{P}) - \frac{P_{A,C}P_{C,T}}{P_{A,G}} \\
f_3(\widetilde{P}) &= \frac{P_{C,A}P_{A,G}}{1 - P_{C,A} - P_{C,C} - P_{C,G}} \\
f_4(\widetilde{P}) &= \frac{P_{G,A}P_{A,G}}{1 - P_{C,A} - P_{C,C} - P_{C,G}} \\
f_5(\widetilde{P}) &= 1 - P_{A,A} - f_3(P) - f_4(P)
\end{aligned}
$$

Identification of Θ_2

$$P_{a,b} \geq 0 \text{ for } (a,b) \in J_7, \quad f_i(\widetilde{P}) \geq 0,\ \text{For } i = 1,2,3,4,5, \tag{2}$$

$$\sum_{b \in \mathcal{A}_3} P_{a,b} < 1 \text{ for } a \in \mathcal{A}_2\,,\ P_{G,A} + f_1(\widetilde{P}) + f_2(\widetilde{P}) < 1\,, \tag{3}$$

$$\sum_{j=3}^{5} f_j(\widetilde{P}) < 1\,. \tag{4}$$

Θ_2 can be identified with
$V_7 := \{\widetilde{P} \in \mathcal{C}_3^{\mathcal{A}_2} \times (0,1) : \widetilde{P}$ satisfies (2) and (3)$\}$.

The Test of CSPR for dinucleotides

- Since P is positive and stochastic, it can be seen that $V_7 = \{\widetilde{P} \in \mathcal{C}_3^{\mathcal{A}_2} \times (0,1) : f_2(\widetilde{P}) > 0, f_5(\widetilde{P}) > 0\}$.
- For $\epsilon > 0$, define $\Delta(h, \epsilon) := (h - \frac{\epsilon}{2}, h + -\frac{\epsilon}{2})$ for h real.
- Define

$$\begin{aligned}
C_7(\epsilon) \quad = \quad & \{\overline{P} \in \mathcal{C}_3^{\mathcal{A}} : \widetilde{P} \in V_7, P_{G,G} \in (f_1(\widetilde{P}) - \epsilon/2, f_1(\widetilde{P}) + \epsilon/2), \\
& P_{G,C} \in (f_2(\widetilde{P}) - \epsilon/2, f_2(\widetilde{P}) + \epsilon/2), \\
& P_{T,G} \in (f_3(\widetilde{P}) - \epsilon/2, f_3(\widetilde{P}) + \epsilon/2), \\
& P_{T,C} \in (f_4(\widetilde{P}) - \epsilon/2, f_4(\widetilde{P}) + \epsilon/2), \\
& P_{T,A} \in (f_5(\widetilde{P}) - \epsilon/2, f_5(\widetilde{P}) + \epsilon/2)\}.
\end{aligned}$$

Define the statistic $\eta_2 = \eta_2(P)$ as

$$\begin{aligned}
\eta_2 \quad = \quad & \max\Big\{ \left|P_{G,G} - f_1(\widetilde{P})\right|, \left|P_{G,C} - f_2(\widetilde{P})\right|, \\
& \left|P_{T,G} - f_3(\widetilde{P})\right|, \left|P_{T,C} - f_4(\widetilde{P})\right|, \left|P_{T,A} - f_5(\widetilde{P})\right| \Big\},
\end{aligned}$$

if $P \in V_7$. Otherwise, $\eta_2 = 1$.

Formulation of the Test

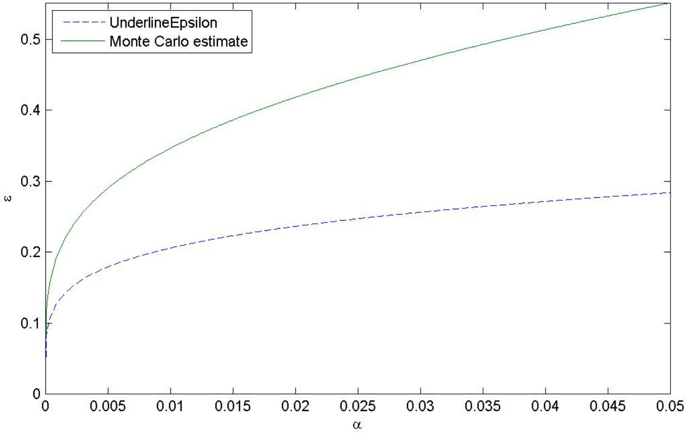
$$\begin{array}{ll} H_0\colon & P \in \Theta \setminus \Theta_2 \iff \overline{P} \notin C_7(\epsilon_\alpha) \iff \eta_2 > \epsilon_\alpha/2, \\ H_1\colon & \quad P \in \Theta_2 \quad \iff \overline{P} \in C_7(\epsilon_\alpha) \iff \eta_2 \leq \epsilon_\alpha/2. \end{array}$$

the probability of a type I error is

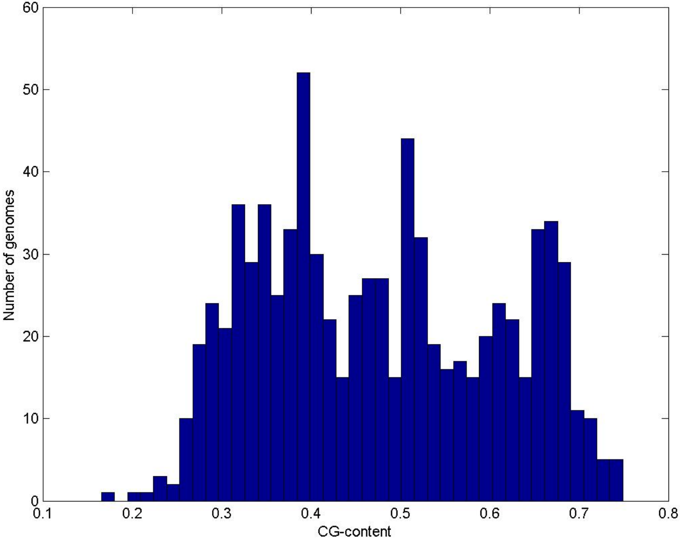
$$\begin{aligned} \mathbb{P}(H_0 \text{ is rejected} \mid H_0 \text{ is true}) &= \mathbb{P}_{\Theta\setminus\Theta_2}(C_7(\epsilon_\alpha)) \\ &= \frac{\mathbb{P}_\Theta\left(C_7(\epsilon_\alpha) \cap (\Theta \setminus \Theta_2)\right)}{\mathbb{P}_\Theta(\Theta \setminus \Theta_2)} \\ &= \mathbb{P}_\Theta(C_7(\epsilon_\alpha)) \end{aligned}$$

The significance level α of the test is fixed by choosing ϵ_α so as to guarantee $\mathbb{P}_\Theta(\eta_2 \leq \epsilon/2) = \mathbb{P}_\Theta(\overline{P} \in C_7(\epsilon_\alpha)) \leq \alpha$.
Let ϵ^* be such that $\mathbb{P}_\Theta(\overline{P} \in C_7(\epsilon^*_\alpha)) = \alpha$.
$\underline{\epsilon}_\alpha := \sqrt[5]{\alpha/27} \leq \epsilon^*_\alpha$.

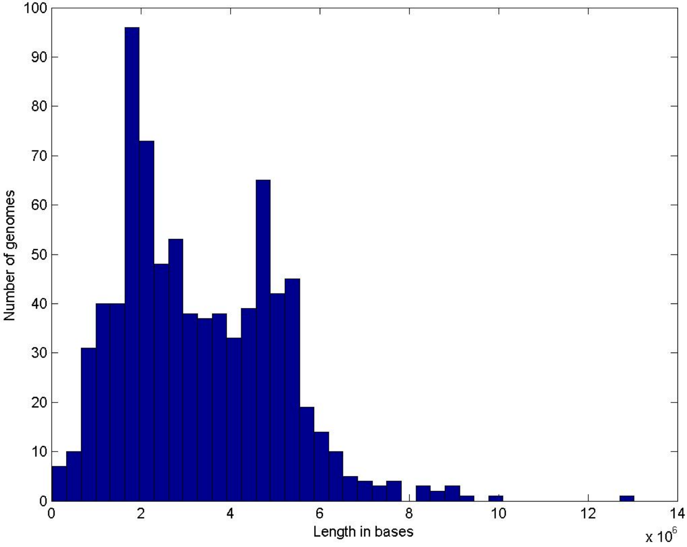
Choices of ϵ_α



Histogram of *GC*-content for 805 Bacteria



Histogram of Lengths of 805 Bacterial Genomes


Number of genomes
Length in bases
x 10^6