J. Math. Biol. (2017) 74:1611–1625
DOI 10.1007/s00285-016-1067-4

Mathematical Biology

An entropy-based technique for classifying bacterial chromosomes according to synonymous codon usage

Andrew Hart[1] · Servet Martínez[2]

Received: 15 October 2015 / Revised: 30 March 2016 / Published online: 12 October 2016

Abstract We present a framework based on information theoretic concepts and the Dirichlet distribution for classifying chromosomes based on the degree to which they use synonymous codons uniformly or preferentially, that is, whether or not codons that code for an amino acid appear with the same relative frequency. At its core is a measure of codon usage bias we call the Kullback–Leibler codon information bias (KL-CIB or CIB for short). Being defined in terms of conditional entropy makes KL-CIB an ideal and natural quantity for expressing a chromosome's degree of departure from uniform synonymous codon usage. Applying the approach to a large collection of annotated bacterial chromosomes reveals three distinct groups of bacteria.

Keywords Entropy · Conditional entropy · Dirichlet distribution · Annotated bacteria

Mathematics Subject Classification 60F10 · 60G42 · 92D20 · 62P10

1 Introduction

Living cells use the genetic code to translate triples of nucleic acids called codons into amino acids, the building blocks of proteins. There are a total of 64 codons of which61

✉ Andrew Hart
ahart@dim.uchile.cl

Servet Martínez
smartine@dim.uchile.cl

[1] UMI 2071 CNRS-UCHILE, Facultad de Ciencias Físicas y Matemáticas, Centro de Modelamiento Matemático, Universidad de Chile, Casilla 170, Correo 3, Santiago, Chile

[2] Departamento de Ingeniería Matemática, UMI 2071 CNRS-UCHILE, Facultad de Ciencias Físicas y Matemáticas, Centro de Modelamiento Matemático, Universidad de Chile, Casilla 170, Correo 3, Santiago, Chile

code for 20 amino acids while the remaining 3 constitute translation stop signals. This many-to-one mapping between codons and amino acids means the genetic code is degenerate. As almost all amino acids are represented by two to six different codons, the code possesses intrinsic redundancy. This provides the cellular machinery with the ability to correctly manufacture protein products in the presence of certain kinds of transcription/translation/replication errors. However, the way in which codons are used to represent amino acids varies from gene to gene and from organism to organism. The mechanism by which variations in these usage patterns arise is not clearly understood, though a number of factors that can influence codon usage are known. These include mutational biases, translational selection pressures, gc content at the third codon site (gc3s) and gene size. The pattern in the way amino acids are represented by codons is called synonymous codon usage (SCU) and inequity in the distribution of codons which code for the same amino acid is referred to as codon usage bias (CUB).

As noted above, patterns of SCU are specific to each organism and their study is of significant biological interest. Consequently, there is a substantial body of literature devoted to studying SCU in the genes of organisms. Comeron and Aguade (1998) give a brief review of methods for measuring SCU, both general and species-specific. Two useful measures are the codon adaptation index (Sharp and Li 1987) and the ‘effective number of codons’ (Wright 1990). A number of other measures have been subsequently posed such as relative codon adaptation (Fox and Erill 2010), various modifications to the ‘effective number of codons’ (Fuglsang 2006; Banerjeee et al. 2005) and the measure of gene expression $E(G)$ which was used by Karlin et al. (2001) to characterize predicted highly expressed genes in four bacteria.

Here, we develop a method for assessing codon usage bias based on concepts from information theory and apply it to a large set of bacterial chromosomes. Our objective is to explore and identify trends in global SCU in a large collection of organisms. In order to do this, our approach differs from that which is typically taken when studying SCU in a chromosome. Rather than examining SCU in a chromosome gene by gene, we aggregate all the genes annotated for the chromosome together and compute total relative frequencies for codons and amino acids which are then plugged into our measure.

To set the scene more concretely, let {a, c, g, t} be the alphabet of nucleotides in DNA. Being triplets of nucleotides, codons are elements of $\{\text{a, c, g, t}\}^3$ and 61 of the 64 possible codons code for amino acids. The remaining 3 codons (taa, tag and tga) indicate a STOP condition that is transcribed into messenger RNA and which instructs the ribosome to halt the translation of a sequence of codons into a polypeptide chain.

Each amino acid can be represented by any of its so-called synonymous codons, which are those codons that code for it. For instance, the four codons gca, gcg, gct and gcc all code for the amino acid alanine (Ala). Thus the list of amino acids defines a partition $\mathscr{A} = \{A_d : d = 1, \ldots, 20\}$ on the set of 61 codons. Each amino acid is associated with a class A_d which is the set of its synonymous codons. For example, if A_1 corresponds to Ala, then $A_1 = \{\text{gca, gcg, gct, gcc}\}$. Further, the collection of 20 amino acids can be organized according to the number of synonymous codons each has: there are 5 amino acids with 4 synonymous codons, 3 with 6 codons, 9 with 2 codons, 1 with 3 codons and 2 with just 1 codon.

We seek a measure or statistic Δ which captures the degree to which the set of amino acids are represented equally often by their synonymous codons. If we suppose that there are k_d codons that code for amino acid A_d, then A_d will exhibit no CUB if each of its k_d synonymous codons appears with relative frequency $1/k_d$. Extending this, it should be clear that a complete chromosome will exhibit no overall CUB provided that every codon C appears in the chromosome with a relative frequency equal to the reciprocal of the number of codons that code for the same amino acid as C. In order for Δ to be useful, it should possess a number of desirable properties. Firstly, a complete absence of CUB, the ideal case just described, should be indicated by the reference value of 0. Secondly, the most extreme form of CUB where each amino acid is always represented by the same codon should correspond to the maximum value of Δ. Thirdly, larger values of Δ should correspond to greater concentrations of the codon distribution on a smaller number of codons, with the extreme case being one codon per amino acid. Finally, it should be independent of the amino acid composition so that comparisons can be made between chromosomes and/or organisms.

This paper is organized as follows. Section 2 begins by presenting information theoretic concepts such as partitions, entropy, conditional entropy and maximal entropy of probability measures when fixing a probability measure over some partition. We propose a statistic Δ for CUB which we call the Kullback–Leibler codon information bias (KL-CIB or simply CIB). This section also includes a discussion of the relationship of Δ to probabilistic measures such as the Kullback–Leibler divergence, from which it takes its name and it is also compared to homozygosity, a well-known measure of genetic similarity. Next, Sect. 3 is concerned with computing the expected value of Δ under a uniform distribution assumption on the set of all possible probability distributions of synonymous codons for each amino acid. This enables a reference value for the mean of the statistic to be computed directly from knowledge of the amino acid composition alone.

The final section applies the proposed entropy-based approach to a set of 2535 annotated bacterial chromosomes and we discover that the bacteria can be divided into three broad groups based on their SCU behavior. In the largest group, which contains 1587 bacteria, amino acids are fairly uniformly represented by their synonymous codons. The next largest group consists of 592 bacteria that exhibit a very high degree of CUB, indicating an extreme preference for a small number of codons. The third and smallest group has 356 bacteria whose SCU bias is moderate.

2 Partitions, entropy and maximal entropy

Let I be a finite set and denote its cardinality by $|I|$. A partition $\mathscr{A} = (A : A \in \mathscr{A})$ of I is such that its elements, which we shall call atoms, are non-empty, disjoint and cover I. Throughout the remainder of this work, I is the set of 61 codons which code for amino acids while $\mathscr{A}$ is the set of 20 amino acids and each amino acid $A \in \mathscr{A}$ may be viewed as the set of codons which code for it.

It will also be useful to introduce the partition $\mathscr{I} = \{\{i\} : i \in I\}$ where the atoms are the singletons of I. For every $i \in I$ there is a unique atom $A^i \in \mathscr{A}$ containing i. For $A \in \mathscr{A}$, $|A|$ is the number of elements in A.

Once and for all, we fix a probability distribution q on the set of amino acids, which can be represented by a vector $q = (q_A : A \in \mathscr{A})$ satisfying $q_A \geq 0$ for $A \in \mathscr{A}$ and $\sum_{A\in\mathscr{A}} q_A = 1$. Its Shannon entropy $h(\mathscr{A}, q)$ is given by

$$h(\mathscr{A}, q) = -\sum_{A\in\mathscr{A}} q_A \log q_A$$

with the convention that the summand is taken to be 0 whenever $q_A = 0$. Next, a probability measure p on I is given by a vector $p = (p_i : i \in I)$ and $p(B) = \sum_{i\in B} p_i$ for all $B \subseteq I$. We say that p extends q if $p(A) = q_A$ for all $A \in \mathscr{A}$. We write this extension relation as $p \succ q$. From now on, we shall assume that $p \succ q$ always holds.

For any measure p extending q from $\mathscr{A}$ to $\mathscr{I}$, we have the increasing property $h(\mathscr{A}, q) \leq h(\mathscr{I}, p)$. More generally, let $\mathscr{B}$ be a partition of I which is coarser than $\mathscr{A}$, that is, the atoms of $\mathscr{A}$ are unions of atoms of $\mathscr{B}$. Then, $h(\mathscr{A}, q) \leq h(\mathscr{B}, p) = -\sum_{B\in\mathscr{B}} p(B) \log p(B)$. The conditional entropy is

$$h_p(\mathscr{B} \mid \mathscr{A}) = h(\mathscr{B}, p) - h(\mathscr{A}, q) = -\sum_{A\in\mathscr{A}} q_A \left(\sum_{B\in\mathscr{B}:B\subseteq A} \frac{p(B)}{q_A} \log \frac{p(B)}{q_A} \right). \quad (1)$$

So, when $\mathscr{B} = \mathscr{I}$ we obtain

$$h_p(\mathscr{I} \mid \mathscr{A}) = -\sum_{A\in\mathscr{A}} q_A \left(\sum_{i\in A} \frac{p_i}{q_A} \log \frac{p_i}{q_A} \right) = -\sum_{A\in\mathscr{A}} q_A \left(\sum_{i\in A} p_{i|A} \log p_{i|A} \right),$$

where $p_{i|A} = p_i/q_A$ for $i \in A$ is the probability that an arbitrary instance of amino acid A is coded by the synonymous codon specified by i. Note that the vector $(p_{i|A^i} : i \in I)$ provides a complete picture of SCU.

Next, define the probability measure $p^{(q)}$ on I by

$$\forall i \in I : \quad p_i^{(q)} = q_{A^i} \big/ |A^i|,$$

that is, $p^{(q)}$ gives the same weight to the points of I contained in the same atom of $\mathscr{A}$. A straightforward computation shows that $p^{(q)} \succ q$ and its entropy is

$$h\big(\mathscr{I}, p^{(q)}\big) = \sum_{A\in\mathscr{A}} q_A \log |A| + h(\mathscr{A}, q).$$

Since the function $-x \log x$ is concave on the interval [0, 1], Jensen's inequality implies that $p^{(q)}$ maximizes the entropy over the set of probability measures that extend q, that is,

$$h\big(\mathscr{I}, p^{(q)}\big) = \max\big\{h(\mathscr{I}, p) : p \succ q\big\}.$$

We can now define the Kullback–Leibler codon information bias (KL-CIB or simply CIB):

$$\Delta = \Delta(p \,|\, q) = h\big(\mathscr{I}, p^{(q)}\big) - h\big(\mathscr{I}, p\big) \text{ for } p \succ q.$$

In other words, Δ is the difference between two entropies where the first is the maximum entropy (computed over the discrete partition $\mathscr{I}$) of codon distributions extending the amino acid composition q and the second is the entropy of the observed codon distribution p, which always extends q. Then,

$$\begin{aligned}\Delta(p \,|\, q) &= \sum_{A\in\mathscr{A}} q_A \log|A| + h(\mathscr{A}, q) - h_p(\mathscr{I} \,|\, \mathscr{A}) - h(\mathscr{A}, q)\\ &= \sum_{A\in\mathscr{A}} q_A \log|A| - h_p(\mathscr{I} \,|\, \mathscr{A}),\end{aligned} \tag{2}$$

which provides an alternative formulation of CIB in terms of conditional entropy. CIB possesses a number of properties which make it useful for assessing CUB. Being based on entropy, it is a natural quantity for measuring departure from equal usage. It takes the value zero if and only if p is constant on each atom of $\mathscr{A}$, a configuration concordant with unbiased SCU. The maximum value that Δ can take is $\sum_{A\in\mathscr{A}} q_A \log|A|$. By the definition of Δ, this maximum value corresponds to a conditional entropy of zero and this means there is precisely one codon representing each amino acid. Finally, fixing the distribution of amino acids in the definition of Δ has the effect of removing the influence of amino acid composition so that it can be used to compare the degree of CUB in coding sequences from different chromosomes/species.

In Sect. 4, we shall use The empirical CIB $\hat{\Delta}$ for assessing a chromosome's CUB. The empirical CIB is

$$\hat{\Delta} = \Delta(\hat{p} \,|\, \hat{q}), \tag{3}$$

where $\hat{q}$ and $\hat{p}$ are respectively the vectors of amino-acid and codon relative frequencies obtained from annotation and sequence data. So $\hat{\Delta}$ is simply Δ evaluated for empirically derived estimates of p and q rather than theoretical or simulated probability measures. Note that $\hat{\Delta}$ retains all the desirable properties of Δ outlined above. In addition, we should point out that, in any real chromosome or DNA coding sequence, the vector of codon relative frequencies *always* extends the vector of amino-acid relative frequencies (subject to computation error), that is, the assumption $\hat{p} \succ \hat{q}$ is always valid in this context and hence $\hat{\Delta}$ always has a reasonable interpretation as a summary of SCU.

2.1 Relationship to Kullback–Leibler divergence

Next, we shall establish an equivalence between $\Delta(p \,|\, q)$ and the Kullback–Leibler divergence for which it has been named. Then, we shall describe how $\Delta(p \,|\, q)$ is related to homozygosity, a standard quantity in the field of population genetics which has also been used in measuring CUB.

First, we recall the definition of the Kullback–Leibler divergence. For two probability measures μ and ν on I such that $\nu(i) = 0$ implies $\mu(i) = 0$, the Kullback–Leibler divergence of ν from μ is

$$K(\mu \mid \nu) = \sum_{i \in I} \mu(i) \log\bigl(\mu(i)/\nu(i)\bigr),$$

where by convention the summand is taken to be 0 if $\nu(i)$ [and hence $\mu(i)$] is equal to 0.

Proposition 1 *For $p \succ q$, $\Delta(p \mid q) = K\bigl(p \mid p^{(q)}\bigr)$.*

Proof For $i \in I$ we have

$$\frac{p_i}{p_i^{(q)}} = |A^i| \frac{p_i}{q_{A^i}} = |A^i| p_{i|A^i},$$

so

$$K\bigl(p \mid p^{(q)}\bigr) = \sum_{i \in I} p_i \log\Bigl(|A^i| p_{i|A^i}\Bigr) = \sum_{A \in \mathscr{A}} \sum_{i \in A} p_i \log\bigl(|A| p_{i|A^i}\bigr).$$

Then

$$\begin{aligned} K\bigl(p \mid p^{(q)}\bigr) &= h(\mathscr{A}, q) - h(\mathscr{I}, p) + \sum_{A \in \mathscr{A}} \sum_{i \in A} p_i \log |A| \\ &= h(\mathscr{A}, q) - h(\mathscr{I}, p) + \sum_{A \in \mathscr{A}} q_A \log |A| = \Delta(p \mid q). \end{aligned}$$

□

Next, statistical estimators of homozygosity have been used to construct a measure of CUB called the "effective number of codons" (Wright 1990). The homozygosity of each atom $A \in \mathscr{A}$ is

$$\kappa_A = \sum_{i \in A} p_{i|A}^2$$

and it is related to $\Delta(p \mid q)$ as follows.

Proposition 2 *The statistic $\Delta(p \mid q)$ and the homozygosity are related by the inequality*

$$\Delta^q(p) \leq \sum_{A \in \mathscr{A}} q_A |A| \kappa_A - 1.$$

Proof Firstly, the homozygosities of the atoms of $\mathscr{A}$ can be related to the χ^2 distance. The χ^2 distance between two probability measures μ and ν is

$$d^2_{\chi^2}(\mu,\nu) = \sum_{i\in I} (\mu(i)/\nu(i) - 1)^2\, \nu(i).$$

In our case,

$$\begin{aligned} d^2_{\chi^2}\big(p, p^{(q)}\big) &= \sum_{A\in\mathscr{A}} \sum_{i\in A} \big(|A| p_{i|A} - 1\big)^2 \frac{q_A}{|A^i|} = \sum_{A\in\mathscr{A}} \frac{q_A}{|A|} \sum_{i\in A} \big(|A| p_{i|A} - 1\big)^2 \\ &= \sum_{A\in\mathscr{A}} \frac{q_A}{|A|} \left(\sum_{i\in A} \big(|A| p_{i|A}\big)^2 - 2|A| + |A| \right) = \sum_{A\in\mathscr{A}} \frac{q_A}{|A|} \sum_{i\in A} \big(|A| p_{i|A}\big)^2 - 1 \\ &= \sum_{A\in\mathscr{A}} q_A |A| \sum_{i\in A} p^2_{i|A} - 1. \end{aligned}$$

Then,

$$d^2_{\chi^2}\big(p, p^{(q)}\big) = \sum_{A\in\mathscr{A}} q_A |A| \kappa_A - 1.$$

Finally, combining Proposition 1 with the well-known result (see, Gibbs and Su 2002) that

$$K(\mu \mid \nu) \le d^2_{\chi^2}(\mu,\nu)$$

gives the result.

3 The expected CIB

We would like to have a better idea of how the quantity $\Delta(p \mid q)$ behaves probabilistically. Recalling that the amino acid composition q is fixed, our initial step towards this end is to place a uniform probability law on the set of probability measures that extend q which treats all possible SCU scenarios as equally likely. We shall then compute the expected value of $\Delta(P \mid q)$ when P is chosen according to this law. Thus, $\mathbb{E}\big(\Delta(P \mid q)\big)$ will represent the mean CIB among all possible SCU scenarios where equal weight or importance is given to all scenarios. Note that all SCU scenarios imply some kind of CUB except for $p^{(q)}$ which equates to an utter lack of bias. As a result, $\mathbb{E}\big(\Delta(p \mid q)\big)$ is a measure of the average CUB one would expect to observe supposing all possible SCU scenarios are equally likely. It carries an implicit assumption of CUB.

To begin, since any extension p of q satisfies $\sum_{i\in A} p_i = q_A$ for $A \in \mathscr{A}$, p is characterized by the set of probability vectors of the form

$$\big(\big(p_i/q_A : i \in A\big) : A \in \mathscr{A}\big).$$

Any such measure takes values in $S_{|A|-1}$, the simplex of dimension $|A| - 1$.

Pursuant to the ‘equal likeliness’ supposition described above, we impose a distribution $\mathbb{P}$ on the set of probability measures on I that extend q which is a product of independent Dirichlet distributions with its support constrained to the set of extensions of q. More precisely, the random vector $P = (P_i : i \in I)$ is such that

$$X(A) = \big(P_i/q_A : i \in A\big) \sim \text{Dirichlet}\,(\underbrace{1,\ldots,1}_{|A|\text{ times}}) \tag{4}$$

and $(X(A) : A \in \mathscr{A})$ are independent random vectors. We recall that the density for $(Y_1,\ldots,Y_a) \sim \text{Dirichlet}\,(\underbrace{1,\ldots,1}_{a\text{ times}})$ is $f_{Y_1,\ldots,Y_a}(y_1,\ldots,y_a) = (a-1)!\mathbf{1}\big((y_1,\ldots,y_a) \in S_{a-1}\big)$.

We can now compute the expectation of $\Delta(P \mid q)$ with respect to this distribution.

Theorem 1 *Let P be a random probability vector extending q that is distributed according to (4). Then*

$$\mathbb{E}\big(\Delta(P \mid q)\big) = \sum_{A\in\mathscr{A}} q_A \log|A| - \sum_{A\in\mathscr{A}} q_A \xi_{|A|} + 1,$$

where $\xi_n = \sum_{j=1}^n j^{-1}$ denotes the nth *harmonic number.*

Proof From (2), it suffices to show

$$\mathbb{E}\,(h_P(\mathscr{I} \mid \mathscr{A})) = \sum_{A\in\mathscr{A}} q_A \xi_{|A|} - 1. \tag{5}$$

However, we shall prove a slightly more general statement. Let $\mathscr{B}$ be a coarser partition of I than $\mathscr{A}$ and P a random probability measure on $\mathscr{B}$ extending q. We shall prove that

$$\mathbb{E}\,(h_P(\mathscr{B} \mid \mathscr{A})) = \sum_{A\in\mathscr{A}} q_A \xi_{|A|} - \sum_{A\in\mathscr{A}} \frac{q_A}{|A|}\left(\sum_{B\in\mathscr{B}:B\subseteq A} |B|\xi_{|B|}\right). \tag{6}$$

Note that (5) immediately follows from (6) when $\mathscr{B} = \mathscr{I}$ because the atoms of $\mathscr{I}$ are all singletons. Restricting our attention to (6), we let B be an arbitrary atom of the partition $\mathscr{B}$.

Now, if $(Y_j : j = 1,\ldots,a) \sim \text{Dirichlet}\,(c_1,\ldots,c_a)$ with $c_1,\ldots,c_a > 0$, $(J_i : i = 1,\ldots,k)$ is a partition of $\{1,\ldots,a\}$ and $Z_i = \sum_{j\in J_i} Y_j$ for $i = 1,\ldots,k$, then it follows that $(Z_i : i = 1,\ldots,k) \sim \text{Dirichlet}\,(\sum_{j\in J_i} c_j : i = 1,\ldots,k)$ (see for instance Section 2.2.1 in, Bertoin 2006). As a special case of this, we have

$$Z_i \sim \text{Beta}\left(\sum_{j\in J_i} c_j, \sum_{j\in\{1,\ldots,a\}\setminus J_i} c_j\right). \tag{7}$$

Recalling that $P(B) = \sum_{i \in B} P_i$, (1) yields

$$h_P(\mathscr{B} \,|\, \mathscr{A}) = -\sum_{A \in \mathscr{A}} q_A \left(\sum_{B \in \mathscr{B} : B \subseteq A} \frac{P(B)}{q_A} \log \frac{P(B)}{q_A} \right)$$

while combining distributions (4) and (7) enables us to conclude that

$$\frac{P(B)}{q_A} \sim \text{Beta}(|B|, |A| - |B|).$$

Then,

$$\mathbb{E}(h_P(\mathscr{B} \,|\, \mathscr{A})) = -\sum_{A \in \mathscr{A}} q_A \left(\sum_{B \in \mathscr{B} : B \subseteq A} \mathbb{E}(W_{B,A} \log W_{B,A}) \right), \tag{8}$$

where $W_{B,A} \sim$ Beta $(|B|, |A| - |B|)$.

We need to compute $\mathbb{E}(W \log W)$ for $W \sim$ Beta (l, m). Firstly,

$$\mathbb{E}\,(W \;\log W) = \frac{(l+m-1)!}{(l-1)!\,(m-1)!} \int_0^1 x\,(\log x)\, x^{l-1}(1-x)^{m-1} dx.$$

For reals $s, t \geq 0$, define

$$\theta(s,t) = \int_0^1 x^s\,(1-x)^t\; \log x\, dx.$$

By using the fact that $\theta(s,t)$ is the derivative with respect to s of the Beta function $\beta(s+1, t+1) = \int_0^1 x^s(1-x)^t\, dx = \Gamma(s+1)\Gamma(t+1) \big/ \Gamma(s+t+2)$, we can write

$$\theta(s,t) = \frac{d}{ds}\beta(s+1, t+1) = \frac{d}{ds}\frac{\Gamma(s+1)\Gamma(t+1)}{\Gamma(s+t+2)}$$
$$= \beta(s+1, t+1)\left(\frac{\Gamma'(s+1)}{\Gamma(s+1)} - \frac{\Gamma'(s+t+2)}{\Gamma(s+t+2)} \right).$$

Since the Digamma function $\psi(z) = \Gamma'(z)/\Gamma(z)$ satisfies $\psi(n+1) = \psi(1) + \sum_{j=1}^{n-1} j^{-1}$ for integer values $n \geq 2$ (see formula 6.3.2 on Page 258 of, Abramowitz and Stegun 1964), we can deduce that for integers $s, t \geq 0$

$$\theta(s,t) = \frac{s!\,t!}{(s+t+1)!}\,(\xi_s - \xi_{s+t+1})\,.$$

Hence, for $(W, 1-W) \sim$ Dirichlet (l, m) we have

$$\mathbb{E}(W \;\log W) = -\frac{(l+m-1)!}{(l-1)!\,(m-1)!} \cdot \frac{l!\,(m-1)!}{(l+m)!}\,(\xi_{l+m} - \xi_l) = -\frac{l}{l+m}\,(\xi_{l+m} - \xi_l)\,.$$

Setting $W = W_{B,A}$, $l = |B|$ and $m = |A| - |B|$, we obtain

$$\mathbb{E}(W_{B,A} \log W_{B,A}) = -\frac{|B|}{|A|}\left(\xi_{|A|} - \xi_{|B|}\right)$$

and substituting this into (8) yields the result. □

Note. The calculations in the above proof also enable $\mathbb{E}\big(h_P(\mathscr{B} \,|\, \mathscr{A})\big)$ to be computed exactly when $X(A) \sim \text{Dirichlet } (\underbrace{c, \ldots, c}_{|A|\text{times}})$ where c is a positive fixed integer which is the same for all $A \in \mathscr{A}$. Here, we have fixed $c = 1$ in the statement of the theorem because this causes all distributions p that extend q to occur with the same probability.

Remark 1 To better understand the tail distribution of $\Delta(P \,|\, q)$, one can attempt to bound the expression $\mathbb{P}\left((\Delta(P \,|\, q) - \mathbb{E}(\Delta(P \,|\, q))\right) \le -\lambda)$. One way of doing this is by using the Azuma-Hoeffding large deviation inequality (see for instance Lemma 11.2 in Section 11.1.4 of, Waterman 1995) and only considering dyadic refinements. Let $(\mathscr{B}_k : k = 0, \ldots, K)$ be a dyadic sequence of partitions from $\mathscr{B}_0 = \mathscr{A}$ to $\mathscr{B}_K = \mathscr{I}$, B^k be the unique atom of $\mathscr{B}_k$ which is split into $\bar{B}^k$ and $B^k \backslash \bar{B}^k$, and A^k be the atom in $\mathscr{A}$ that contains B^k. Then define $(\zeta_k : k = 0, \ldots, K - 1)$ by

$$\zeta_k = q_{A^k} \log \max\{|\bar{B}^k|, |B^k \backslash \bar{B}^k|\} + \log 2 - q_{A^k} \log q_{A^k}. \tag{9}$$

Assuming $q_A \le 1/e$ for all $A \in \mathscr{A}$, standard arguments combined with (6) yield

$$\forall \lambda > 0 : \quad \mathbb{P}\left(\Delta(P \,|\, q) \le \mathbb{E}\big(\Delta(P \,|\, q)\big) - \lambda\right) \le \exp\left(-\lambda^2 \Big/ 2 \sum_{k=0}^{K-1} \zeta_k^2\right). \tag{10}$$

As an aside, Sect. 4 examines a collection of 2535 bacterial chromosomes downloaded from the NCBI ftp server. Among these, we found that $\max\{q_A : A \in \mathscr{A}\} \approx 0.179855 < 1/e$ always holds and so the above conclusions are valid for all the real-world chromosomes considered. Next, $\mathscr{B}_{0,K}$ can be chosen optimally to minimize the bound on the right-hand side of (10) and this bound can be shown to be greater than 0.99805. Clearly, this is not tight enough for any kind of practical application.

4 Application and comments

We downloaded a large set of 2585 bacterial DNA sequences from the NCBI ftp server. All of the sequences were marked as ‘complete genome’ or ‘nearly complete genome’, so constitute chromosomes and not plasmids. Chromosomes which were either lacking annotation data or which had fewer than 200 coding sequences, of which there were 8 and 45 respectively, were filtered out. This left a set of 2535 chromosomes. Next, the codon distribution ($\hat{p}$) was estimated from the relative frequencies of the codons for each chromosome by adding up the counts of codons in all the genes annotated in the GenBank (`.gbk`) file, excluding the terminating STOP codons, and rescaling the results to sum to one. For example, the numbers of aaa codons appearing in each gene were added together and then divided by the total number of codons contained in all the genes in the chromosome. The corresponding amino acid distribution ($\hat{q}$) was computed by summing the relative frequencies of the synonymous codons for each amino acid. Finally, the relative frequencies of the codons and amino acids were used to compute $\hat{\Delta}$ for each chromosome in accordance with (3) and (2).

In order to compare the $\hat{\Delta}$ statistic observed in a real-world chromosome with what is theoretically observable for such a chromosome, we examined the behavior of Δ in codon distributions that are compatible with the chromosome's amino acid composition. We did this for each of the 2535 chromosomes by carrying out a series of Monte Carlo simulations in which 100 codon distributions were sampled uniformly at random subject to having the same amino acid distribution as the chromosome in question. This produced a sample of size 253,500 covering all 2535 chromosomes with the characteristic that the two sets of statistics, the set of $\hat{\Delta}$'s observed from the bacterial chromosomes and the larger set of theoretically simulated Δ's, are comparable since they are based on the same mixture of amino acid compositions.

The algorithm we used for uniformly sampling codon distributions that extend a specified amino acid distribution can be described as follows. First fix the amino acid distribution $q = (q_A : A \in \mathscr{A})$. Then, a distribution $p_{\cdot|A} = \left(p_{i|A} : i \in A\right)$ for the usage of the synonymous codons that code for each amino acid A is sampled from the appropriate Dirichlet distribution. For example, a distribution for the two codons (gac and gat) that code for aspartic acid (Asp) would be obtained by sampling from the 2-dimensional Dirichlet(1, 1) distribution while a distribution for proline (Pro) which has 4 synonymous codons (cct, ccc, cca and ccg) would be obtained by sampling from a 4-dimensional Dirichlet(1, 1, 1, 1) distribution. The final codon distribution $p = (p_i : i \in I)$ is then obtained by combining these two distributions as follows:

$$\forall i \in I, \quad p_i = p_{i|A^i} q_{A^i}.$$

In other words, take an amino acid distribution and uniformly select a possible SCU distribution at random for each amino acid in turn. After multiplying each SCU distribution by the probability of the corresponding amino acid, the result is a codon distribution which is a discrete extension of the original amino acid composition. In future, we shall refer to this procedure as ‘uniformly simulating SCU scenarios’, as it generates a codon distribution compatible with a given amino acid composition by selecting a possible usage of synonymous codons uniformly at random.

Figure 1 shows kernel density estimates of $\hat{\Delta}$ for the collection of 2535 chromosomes and the codon distributions simulated as just described above. While codon distributions are known not to manifest uniformly in nature as this would be inconsistent with the amino acid compositions that have been empirically observed in organisms, the figure illustrates the unsurprising fact that the patterns of SCU in bacteria, which are essentially characterized by the codon distribution conditioned on amino acid composition, are also non-uniform. The figure shows the overall shape of CUB based on CIB for the collection of bacteria included in the study and contrasts this with the wide range of feasible SCU patterns. Observe that the range of values (roughly [0.0, 0.6]) taken by $\hat{\Delta}$ on the x-axes of the plot is similar for both the collec-

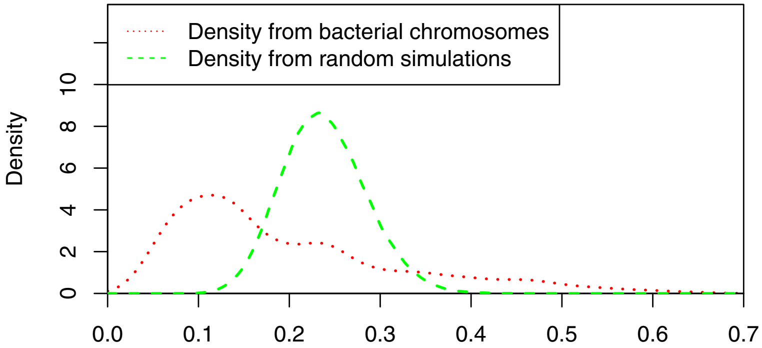

Fig. 1 Distributions of CIB for real and randomly simulated data. The *dotted red line* is a density estimated from 2535 bacterial chromosomes while the *dashed green line* is an estimate obtained from a sample produced by aggregating 100 random simulations for each of the 2535 chromosomes' amino acid compositions

tion of chromosomes and the simulated data. However, the $\hat{\Delta}$'s for the chromosomes are skewed towards the no CUB end of the spectrum with fatter tails compared to the Δ's based on uniformly simulated SCU scenarios, which appear to have a Gaussian shape.

Next, by examining where the $\hat{\Delta}$ computed for each bacterial chromosome lies among the set of all codon distributions that extend the amino acid distribution for that chromosome, an interesting phenomenon can be revealed. For each chromosome, we estimated Δ^*, which is defined as follows: Δ^* is the proportion of all codon distributions that are compatible with the chromosome's amino acid distribution and that would give rise to a Δ smaller than the chromosome's empirically computed $\hat{\Delta}$. In other words, Δ^* gives the probability of observing a chromosome with less CUB than the chromosome under consideration. The probability for each chromosome was estimated by carrying out a series of Monte Carlo simulations in which 10^4 codon distributions were sampled uniformly at random subject to having the same amino acid distribution as the chromosome. The simulations were conducted using the algorithm for simulating uniform SCU scenarios described above. The probability Δ^* was then obtained as the fraction of the 10^4 simulated codon distributions whose Δ is smaller than or equal to the chromosome's $\hat{\Delta}$.

Theoretically, if Δ^* is computed for a codon distribution obtained by uniformly simulating a SCU scenario given a particular amino acid composition, then Δ^* should itself be uniformly distributed on the interval [0, 1]. To check this empirically, we sampled a codon distribution for each chromosome by uniformly simulating a SCU scenario based on its amino acid composition and then computed Δ together with the corresponding value of Δ^* as described in the preceding paragraph.

Figure 2 displays kernel density estimates of the distribution of Δ^* for the collection of bacterial chromosomes. Three concentrations of chromosomes are apparent in the real data (the dotted red line), the central one being fairly amorphous while the extreme concentrations stand out prominently. In contrast, Δ^* is uniformly distributed in the randomly simulated data (the dashed green line) as expected.

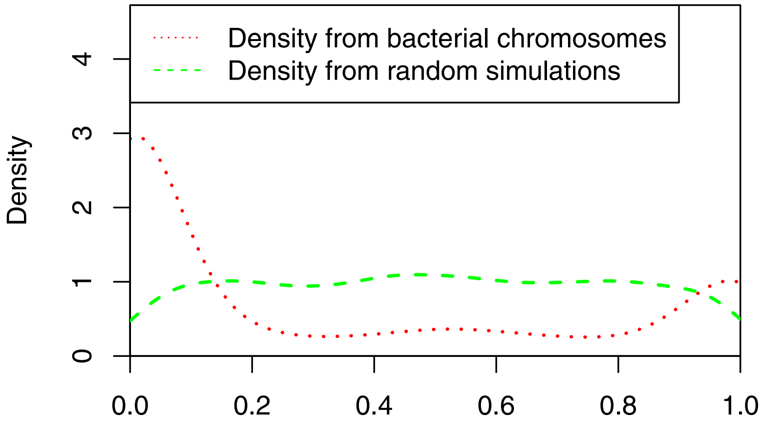

Fig. 2 Distributions of probabilities of observing a value smaller than $\hat{\Delta}$ in real and randomly simulated data. The *dotted red line* is a density estimated from 2535 bacterial chromosomes while the *dashed green line* is an estimate for Δ^* based on uniformly simulating a SCU scenario for every amino acid composition exhibited by the chromosomes

Table 1 Clustering of 2535 bacterial chromosomes into three groups according to Δ^*, the proportion of codon distributions with less CUB than the chromosome

Group	Bias	Size	Δ^*			$\hat{\Delta}$		
			Center	Min.	Max.	Center	Min.	Max.
1	Low	1587	0.0269	0.0000	0.2661	0.1182	0.0262	0.2140
2	Moderate	356	0.5086	0.2699	0.7240	0.2372	0.2042	0.2709
3	High	592	0.9556	0.7362	1.0000	0.3921	0.2645	0.6629

The grouping was obtained via k-means clustering and the between-cluster sum of squares accounts for 96.8 % of the total sum of squares

We used k-means clustering to assign each chromosome to one of three groups according to the value of Δ^*. The main characteristics of the resulting groups, which we nominally named ‘low CUB’, ‘moderate CUB’and ‘high CUB’, are summarized in Table 1. The table gives the size of each group, the range of Δ^* and $\hat{\Delta}$ spanned by each group, as well as the center (mean) value of Δ^* and $\hat{\Delta}$. Observe that the ranges of Δ^* covered by the three groups are disjoint, which is a side effect of the k-means clustering procedure. On the other hand, the ranges of $\hat{\Delta}$ that the groups encompass overlap slightly. This phenomenon is attributable to the theoretical equal weighting applied to all codon distributions during the calculation of Δ^*.

Next, a principal component analysis (PCA) of the chromosomes’ codon distributions highlights the concentration of the codon distributions for bacterial chromosomes into a lower dimensional region inside the 61-dimensional set of all theoretically possible codon distributions. The first principal component is by far the most important, accounting for approximately 75 % of variability in the codon distributions, while the first 10 principal components explain about 95 % of the total variability. Performing PCA on the amino acid distributions yields a similar picture: 80 % of the variability

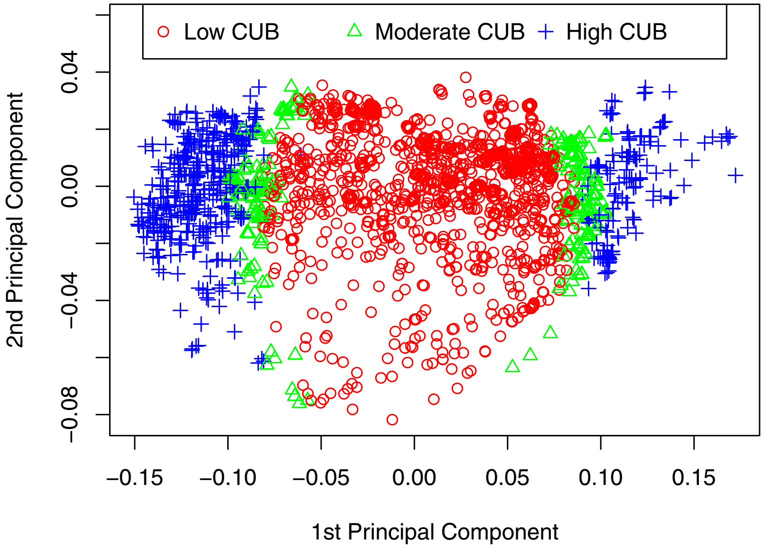

Fig. 3 Plot of the first two principal components of the codon distributions of 2535 bacterial chromosomes classified into three classes according to degree of CUB

is explained by the first principal component while the first 10 components explain 98 % of the total variability in the amino acid distributions.

This is not surprising in light of what is known about CUB in organisms, though there would seem to be something particular about the way codon distributions for bacterial chromosomes are arranged. We plotted the first two principal components of the codon distributions, indicating to which group each chromosome belongs. The most significant feature of our final plot (see Fig. 3) is that the plot is roughly broken into vertical bands according to membership in the 'low CUB', moderate CUB'or 'high CUB' group. Thus, the first principal component captures substantial information concerning CUB in the chromosome. We believe that this may be the first time that this characteristic of the first principal component has been demonstrated at the chromosome level. At present, we have no satisfactory biological explanation for the division of bacterial chromosomes into three groups based on CUB.

Acknowledgements This work was supported by the Center for Mathematical Modeling's CONICYT (Comisión Nacional de Investigación Científica y Tecnológica) Basal program PFB 03. The authors would like to thank members of the CMM Mathomics Laboratory for interesting and helpful discussions. Thanks also go to the referee whose comments helped to improve the paper, in particular, for recognizing that $\theta(s, t)$ is the derivative of the Beta function, which in conjunction with the Digamma function, considerably shortened our original proof of Theorem 1.

References

Abramowitz M, Stegun I (1964) Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Applied Mathematics Series, vol 55, 10th edn. Dover Publications, New York

Banerjeee T, Gupta S, Ghosh T (2005) Towards a resolution on the inherent methodological weakness of the "effective number of codons used by a gene". Biochem Biophys Res Commun 330:1015–1018

Bertoin J (2006) Random fragmentation and coagulation processes, studies in advanced mathematics, vol 102. Cambridge University Press, Cambridge

Comeron J, Aguade M (1998) An evaluation of measures of synonymous codon usage bias. J Mol Evol 47:268–274

Fox J, Erill I (2010) Relative codon adaptation: a generic codon bias index for prediction of gene expression. DNA Res 17(3):185–196. doi:10.1093/dnares/dsq012

Fuglsang A (2006) Estimating the "effective number of codons": the wright way of determining codon homozygosity leads to superior estimates. Genetics 172:1301–1307. doi:10.1534/genetics.105.049643

Gibbs A, Su F (2002) On choosing and bounding probability metrics. Int Stat Rev 70(3):419–435

Karlin S, Mrázek J, Campbell A, Kaiser D (2001) Characterizations of highly expressed genes of four fast-growing bacteria. J Bacteriol 183(17):5025–5040. doi:10.1128/JB.183.17.5025-5040

Sharp P, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias and its potential applications. Nucleic Acids Res 15:1281–1295

Waterman M (1995) Introduction to computational biology. Chapman and Hall, London

Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29