Evolutionary Constraints on Coding Sequences at the Nucleotidic Level: A Statistical Physics Approach

Didier Chatenay, Simona Cocco, Benjamin Greenbaum, Rémi Monasson and Pierre Netter

Abstract Selection at the molecular level is generally measured by amino-acid alterations, for instance, through the ratio of non-synonymous and synonymous substitutions. While it is known that codons coding for identical amino acids are not perfectly identical in terms of fitness cost, e.g. due to differences in the kinetics of the associated t-RNAs, mechanisms exist for selection acting at the nucleotide level rather than the amino-acid level. In this work, we consider two such mechanisms. The first is the action of the innate immune system, with pattern recognition receptors capable of recognizing small nucleotidic motifs, such as CpG dinucleotides. Pathogens such as viruses are under this selective pressure while strongly constrained by the fact that their short genomes must code for essential proteins. A second tentative mechanism, referred to as the Ambush Hypothesis, suggests that codons are optimized to favor the presence of off-frame stop codons, which are useful to abort translation of non-functional proteins in case of accidental ribosomal

D. Chatenay
Laboratoire Jean Perrin (LJP), CNRS UMR8237, Sorbonne Universités, UPMC University Paris 06, 4 place Jussieu, Case Courrier 114, 75005 Paris, France

S. Cocco
Laboratoire de Physique Statistique, Ecole Normale Supérieure and CNRS-UMR8550, PSL Research University, Sorbonne Universités UPMC, 24 Rue Lhomond, 75005 Paris, France

B. Greenbaum
Icahn School of Medicine at Mount Sinai, Tisch Cancer Institute, 1190 One Gustave L. Levy Place, 1st Floor Box 1128 Icahn Building, New York, NY 10029, USA

R. Monasson (✉)
Laboratoire de Physique Théorique, Ecole Normale Supérieure and CNRS-UMR8549, PSL Research University, Sorbonne Universités UPMC, 24 Rue Lhomond, 75005 Paris, France
e-mail: monasson@lpt.ens.fr

P. Netter
Sorbonne Universités, UPMC University Paris 06, CNRS UMR7138, Evolution Paris Seine, IBPS, 7 quai Saint-Bernard, 75005 Paris, France

P. Pontarotti (ed.), *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*,
DOI 10.1007/978-3-319-61569-1_18

frame-shift. We show how the same statistical physics inspired formalism can be applied to both questions to compute selective pressure or make predictions in a null model, called random codon model, in which the coding nature of the genomic sequence and its essential statistical features are retained. Our formalism is based on the notion of transfer matrix, developed in statistical physics to deal with systems of particles with short-range interactions; here, particles are codons and interactions result from the presence of selection mechanism acting at the nucleotidic level, possibly on contiguous codons along the sequence. Our approach is computationally efficient as it requires a computation time growing only linearly with the length of the sequence under study.

1 Introduction

Selection is generally measured in terms of modifications to proteins. A popular approach to estimate the level of evolutionary pressure on a protein is the ratio K_a/K_s for amino acid residues, which estimates the ratio between the number of non-synonymous substitutions at a particular site over the number of synonymous mutations. This approach allows one to estimate how much amino acid evolution at that site is dictated by natural selection, versus how much change an be expected randomly (Li et al. 1985; Nei and Gojobori 1986). However there are other patterns of natural selection that cannot be captured by looking at amino acid changes. In particular, synonymous mutations may not actually be equivalent, but are themselves influenced by natural selection. For instance, codon usage depends on the tissue under consideration and varies across genes. One possible explanation is that the kinetics of corresponding t-RNA varies. This can create a codon usage bias, where more favorable codon usage can offer an organism a replicative fitness advantage (Plotkin and Kudla 2011; Sharp and Li 1987). In the case of, say, an amino acid which is coded for by four codons, synonymous changes at the third position that would be assumed neutral could have a fitness cost.

A clear case where synonymous changes may have a fitness cost is when the genome of a pathogen is targeted by the innate immune system. The innate immune system is a non-specific set of receptors that may target sequence features found in pathogens, but rare or absent in host genomic material found in the receptor's location (Medzhitov and Janeway 2000). Such features may be sequence specific, such as nucleic acid motifs or structural features, and as a result nucleotide changes that alter the presence of such features will have a consequence for pathogen fitness. For instance, the CpG dinucleotide is avoided in the DNA of many genomes, and hence has become a target of the innate immune system which can detect its presence in pathogen genomes (Hemmi et al. 2000). This is just one example of sequence specific patterns which can be sensed (Vabret et al. 2016). In the case of the genomes of RNA viruses, their compact genome is mostly devoted to protein coding. Hence, if one wants to detect the evolution of recognizable patterns, the protein coding aspects of a genome become a constraint (Greenbaum et al. 2014, 2008).

To capture these evolutionary processes in a theoretical framework, we developed a formalism where selective evolutionary forces on motifs and structures are pitted against randomizing forces of constrained nucleotide sequences (Greenbaum et al. 2014). Hence, a viral genome, such as influenza, will avoid a recognizable pattern due to innate immune mediated forces, even when randomizing patterns in codon usage are accounted for in a genome constrained by protein coding and codon usage. To calculate selective and entropic forces we utilized a transfer matrix formalism from statistical physics, which was originally developed to treat systems with short-range interactions in low dimension. Here, the dimension of the "system" is one as a coding sequence can be seen as a linear chain of codons, and the effective interactions between nearest codons along the coding sequence are produced by the selective pressure acting on motifs overlapping contiguous codons. The payoff for the formal development is a reward in terms of computational speed, which allows such forces to be calculated efficiently in large datasets. We showed the forces on CpG dinucleotides in influenza, a motif predicted to be stimulatory in RNA viruses, have the greatest selective forces in influenza and HIV, and created dynamical models based on these principles (Jimenez-Baranda et al. 2011).

Here, after reviewing briefly applications of this framework, we present new results detecting abnormal short nucleotidic motifs. In particular, we present new simultaneous calculations of forces acting on different motifs. This allows us to decide whether the pressures acting on those motifs are independent or not. We also show Monte Carlo (MC) simulations of simple mutational dynamical models that reproduce the equilibrium calculations. We also better characterize the nature of the space of sequences under pressure from the immune system, in particular how similar two randomly picked up sequences are. This information can be useful to understand how constrained are viral sequences by selective pressure, and how the virus can evolve in the constrained space.

The generality of our statistical-physics formalism allows us to adapt it to detect and measure any kind of pressure acting at the nucleotidic level, not necessarily related to the immune system. An example of interest is the so-called Ambush Hypothesis introduced by Seligmann and Pollock (2004). According to the Ambush Hypothesis deleterious effects (production of long and non-functional proteins) due to ribosome frame-shifts during translation can be avoided by increasing the frequency of off-frame STOP codons. This hypothesis is similar, in spirit, to the pressure exerted by the immune system evoked above, as it acts at the nucleotidic level (to produce excess STOP codons in shifted frames by virtue of the genetic code degeneracy) under the constraint of having coding sequences (in the right frame). In the present work, we introduce a new estimator of the presence of off-frame STOP codons, which is not sensitive to the genomic AT content (contrary to most estimators). Our statistical analysis of $\sim$1800 bacterial genomes shows no evidence at all in favor of the Ambush Hypothesis. In addition, extending our transfer-matrix formalism to the study of off-frame STOP codons, we compute the distribution of distances between the position at which the frameshift takes place and the first off-frame STOP codon in the same random codon model used to estimate the immune system pressure. We obtain that the average distance is small

(less than 10 codons), giving further statistical evidence for the fact that, even if the Amubush hypothesis does not hold, off-frame translation rapidly aborts.

The plan of the paper is as follows. In Sect. 2 we review previous works on the estimation of selective pressure based on our statistical physics formalism. New results for nucleotidic motifs under immune pressure and the Ambush hypothesis are reported in, respectively, Sects. 3 and 4. A short discussion with perspectives is given in Sect. 5.

2 Statistical Physics Framework for Detecting Aberrant Short Nucleotide Motifs

2.1 Viral Evolution and Pressures on Nucleotide Usage

The particular problem we are studying is what drives the evolution of a virus which changes its host, and, therefore, its environment. In addition to "local pressures" whose fitness effects derive from the consequences of changing residues to protein function, there are "global pressures", such as the codon bias of the new host, or changes in the innate immune system from one host to the next. Separating these two effects can be challenging.

For example, suppose a DNA virus were to change from a non-mammalian host to a human host. That virus, if it contained many CpG dinucleotides, could stimulate the human innate immune system via Toll-like receptor 9. Such feedback could generate a selective pressure to eliminate CpG dinucleotides. At the same time, altering the number of CpGs could effect the codon usage bias of arginine codons, since two thirds of these codons start with CpGs. If such a pressure were strong enough and arginine not particularly essential, one might even imagine cases where the amino acid itself would change, in a way that might be mistaken for positive selection at the protein level if that site were examined in isolation. As shown in Greenbaum et al. (2014), such a pressure may also exist in an RNA virus, where elimination of the CpG dinucleotide was detectable in the sequence history of influenza and where the codon bias of arginine also was altered as a consequence. This non-random evolution was associated with avoiding motifs that may be detectable (Jimenez-Baranda et al. 2011).

Hence there are at least three possible selective effects: a virus may alter replication efficiency by adopting host codon usage, detectability by altering chemical signatures that bind to host immune receptors, and adaptation via mutations that alter amino acids. We have recently developed an approach from statistical physics which is particularly useful in quantifying the first two of these effects, while offering a general program for analyzing sequences evolving under these global pressures and, therefore, broadly separating the contributions from all three types of effects. The goal is to quantify how much information one can superimpose the nucleotide sequence, at fixed amino acid sequence, thanks to the degeneracy of the

genetic code. The virus has to avoid a global pressure, such as an innate immune receptor targeting a given nucleotide word or phrase, while keeping its capability to make both viable and fit proteins, and, at the same time, operating under a host codon bias that may differ from its own.

To quantify this selective pressure acting in a coding "context" we use a random codon model (RCM) with a given codon usage and fixed amino-acid sequence. The degeneracy of the genetic code allows a number of possible genomes (sequences of codons compatible with the fixed amino-acid sequence) to code for the same protein. We associate to this number an entropic force allowing multiple synonymous mutational paths to the viral sequences in the course of evolution. We then quantify the change in entropy associated with an alteration in the number of possible genomes once a reasonable set of biological and physical constraints are imposed on a virus, such alteration is the pressure associated with moving the virus from an entropically favored configuration to a less favored one due to the external pressure exerted by the innate immune system on nucleotide phrases. In this way, we can infer when a virus is operating under a significant external pressure, since it will be in a lower probability state than the maximum entropy configuration.

In the following we review the statistical physics approach we have introduced in Greenbaum et al. (2014) to characterize the pressure associated to the number of occurrence of small nucleotidic motifs. We will start by computing for the RCM the entropy of sequences as a function of the number of occurrences of one particular dinucleotide motif. Then we draw the occurrences of the motifs sampled on the true sequence, which will correspond to a point in the distribution. The corresponding entropy will tell us how much the set of sequences is reduced or constrained by the presence of the motifs. We will define a 'pressure', equal to the derivative of the distribution in that point, to quantify the degree of such a constraint. We will study the selective pressures on all the dinucleotidic motifs in influenza and HIV viruses of different subtypes for a set of coding regions. The characterization of a given genomic viral sequence in term of the selective pressure, which is an extensive parameter and in particular does not depend on the length of the sequence, will allow us to compare all such cases. Moreover, as detailed in Greenbaum et al. (2014) the selective pressure can be followed during the evolution of a virus which adapts to a human host, and it can be shown to evolve to reach an equilibrium value. We will finally focus on *CpG* motifs and compare the selective pressures on different viruses.

In a second part of the chapter which contain new results we will extend the approach in several directions: First we will introduce a technique based on Monte-Carlo simulation to evolve in silico a sequence, starting from an initial, non-equilibrium selective pressure, to the final equilibrium value. Secondly we will also extend the approach to more motifs. In this way we will obtain a surface in a multi-dimensional space. Finally we will discuss how a selective pressure alters the space of coding sequences, in particular the loss in entropy due to a selective pressure can be associated to an increase of homology between two random sequences under the same selective pressure.

2.2 Random Codon Model: Definitions and Notations

We review here the approach introduced in Greenbaum et al. (2014). The idea is to quantify the motif frequencies in a given sequence with respect to what is expected from a random model (RCM) where the only constraints are the fixed amino acid sequence and the codon bias. We start with particular coding sequence:

$$\bar{\mathbf{C}} = \{\bar{C}_1, \bar{C}_2, \ldots, \bar{C}_L\} \ , \tag{1}$$

where $\bar{C}_i$ the ith codon coding for the ith amino-acid $\bar{a}_i$, and L is the number of amino-acids in the sequence. $\bar{\mathbf{C}}$ can be seen as a sequence of $3 \times L$ nucleotides. Let $\bar{c}_{i,\ell}$ denote the ℓ nucleotide in codon i, with $\ell = 1, 2, 3$, i.e. $\bar{C}_i = \{\bar{c}_{i,1}, \bar{c}_{i,2}, \bar{c}_{i,3}\}$. In the following we will label a nucleotide c with two indices, e.g. $c_{i,\ell}$ to indicate the codon position i and the position ℓ of the nucleotide in the codon, or, alternatively, with only one index to refer to its absolute position along the sequence, e.g. c_j, $j = 1 \ldots 3L$. We therefore have:

$$\bar{\mathbf{C}} = \{\bar{c}_{1,1}, \bar{c}_{1,2}, \bar{c}_{1,3}, \bar{c}_{2,1}, \bar{c}_{2,2}, \bar{c}_{2,3} \ldots, \bar{c}_{L,1}, \bar{c}_{L,2}, \bar{c}_{L,3}\} = \{\bar{c}_1, \bar{c}_2, \ldots \bar{c}_{3L}\} \ . \tag{2}$$

We generate random sequences $C = \{C_1, C_2, \ldots, C_L\}$ coding for the same amino acids as $\bar{\mathbf{C}}$, such that each codon in the random sequence, $C_i = \{c_{i,1}, c_{i,2}, c_{i,3}\}$ (coding for a_i), has a probability equal to the codon bias $p(C_i|a_i)$. At most six codons C_i have a non-zero probability for a given a_i. Codons are drawn independently and at random, and the probability of C is simply the product of the probabilities of the codons,

$$p(C) = \prod_{i=1}^{L} p(C_i|a_i) \ . \tag{3}$$

A motif of length K is a sequence of K characters among $\{A, C, G, T\}$, which we denote by $m = (m_1, m_2, \ldots, m_K)$. We want to compare the number of occurrences of this motif in the natural sequence,

$$\bar{N}_m = \sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{\bar{c}_{j+k}, m_k} \ , \tag{4}$$

to the average number of occurrences of the same motif in the RCM model,

$$\langle N_m \rangle = \sum_{C} p(C) \sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{c_{j+k}, m_k} \ . \tag{5}$$

Here, $\delta_{c,m}$ is the Kronecker function: $\delta_{c,m} = 1$ if the nucleotides c and m are identical, 0 otherwise. The first sum in Eq. (5) is computed over all possible codon

sequences compatible with the amino-acid content. As this number is enormous (typically, exponential–in–L), Monte Carlo simulations were used to compute such average number in Li et al. (1985); in the following we will review the faster method introduced in Greenbaum et al. (2014), based on the transfer matrix approach (Onsager 1944). We will also need to determine whether any difference between $\hat{N}_m$ and $\langle N_m \rangle$ is statistically meaningful or not. To do so, we will consider

$$\langle N_m^2 \rangle = \sum_C p(C) \left(\sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{c_{j+k},m_k} \right)^2, \tag{6}$$

and compare $\langle N_m \rangle - \bar{N}_m$ to the statistical fluctuation $\sqrt{\langle N_m^2 \rangle - \langle N_m \rangle^2}$ within the random codon model.

2.3 Statistical Physics Approach: Partition Function

A way to calculate the moments of the distribution of the number of motifs in the random model, borrowed from statistical physics, is to introduce the so-called partition function:

$$Z(x) = \sum_C p(C) \exp \left(x \sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{c_{j+k},m_k} \right). \tag{7}$$

The derivative

$$N_m(x) = \frac{\partial \log Z(x)}{\partial x}, \tag{8}$$

gives the average number of occurrences of the motif for the fixed parameter x. In particular,

$$\langle N_m \rangle = \left. \frac{\partial \log Z(x)}{\partial x} \right|_{x=0} \tag{9}$$

is the average number of times the motif is found in the unbiased RCM, as can be verified by comparing with Eq. (5). Similarly, the second derivative of the partition function gives access to the variance of the number of motifs:

$$\langle N_m^2 \rangle - \langle N_m \rangle^2 = \left. \frac{\partial^2 \log Z(x)}{\partial x^2} \right|_{x=0}, \tag{10}$$

as can be verified by comparing with Eq. (6). More generally all the moments of the distribution of the number of motifs can be calculated from the derivatives of the partition function in $x = 0$.

2.4 Constrained Model, Maximum Entropy Approach, Legendre Transform and Selective Force

In this section the analogy with statistical physics is further developed, and we show that the partition function introduced above can be considered for arguments $x \neq 0$. Parameter x will play the role of a (selective) force, constraining the distribution of the codons in the RCM to have a given average number of occurrence of the motif under consideration. Following the maximal entropy principle introduced by Jaynes (1957) the least constrained, or maximal entropy distribution $P(C|x)$ capable of reproducing the average number $N_m(C)$ of occurrence of a motifs has an exponential form of the type

$$P(C|x) = \frac{1}{Z(x)} \prod_{i=1}^{L} p_i(C_i|a_i) \times \exp\left(xN_m(C)\right), \tag{11}$$

where, for simplicity, we have assumed that the codon biases are not much affected by the constraint. For $x = 0$ one recovers the unconstrained case of Eq. (3). Our aim is to find, for any given genomic sequence $\bar{C}$, the value of x for which the average number of the number of occurrences of a motif with the distribution $P(C|x)$ corresponds to the number of motifs $\bar{N}_m$ present in the sequence. Parameter x therefore satisfies the equation:

$$\sum_{C} P(C|x) \sum_{j=1}^{3L-K+1} \prod_{k=0}^{K-1} \delta_{c_{j+k},m_k} = \bar{N}_m \tag{12}$$

which is the generalization of Eq. (5) to the biased case, $x \neq 0$.

In statistical physics a Legendre transform allows one to change the description of a system containing a fixed number of particles (Canonical Ensemble) to a system in which the number of particle can fluctuate around an average value determined by the choice of the chemical potential (Grand Canonical Ensemble). Using the same description, here, we can describe the RCM by the free energy potential, i.e. minus the logarithm of the partition function, at fixed number of occurrence of a motifs N_m, or by the entropy at fixed value of the parameter x. x is an intensive parameter, similar to the chemical potential, which we call selective pressure. In the following we show how the Legendre transform relates the two potentials and how they are equivalent in the limit of long sequences. One can

rewrite the partition function in Eq. (7) by summing together all sequences having the same number of occurrences of a motif:

$$Z(x) = \sum_{N_m \geq 0} \Omega(N_m) \exp(x N_m). \tag{13}$$

where $\Omega(N_m)$ is the weighted number of nucleotide sequences (at fixed amino acid content) having N_m motifs, as each sequence is weighted by the product of the codon biases of its codons. We consider the logarithm of $\Omega(N_m)$, denoted by $\sigma(N_m) = \log \Omega(N_m)$. In the case of very long sequences the sum over N_m in (13) is dominated by its maximal contribution, obtained for the value of N_m such that

$$\frac{\partial \sigma(N_m(x))}{\partial N_m} = -x. \tag{14}$$

We therefore obtain

$$\log Z(x) \approx x N_m(x) + \sigma(N_m(x)). \tag{15}$$

or equivalently

$$\sigma(N_m(x)) = \log Z(x) - x N_m(x). \tag{16}$$

which expresses the Legendre relation between the function $\sigma(N_m)$ and minus the free energy, $\log Z(x)$.

What is the interpretation of $\sigma(N_m)$ defined above? If the sequences were not weighted by the product of their codon biases, Ω would a number of sequences, and σ would be an entropy. Due to the presence of the multiplicative weights, σ defined above is a relative entropy with respect to the unbiased distribution. Indeed, it is easy to check from Eq. (16) that σ vanishes for $x = 0$. We therefore introduce the absolute entropy of the unconstrained RCM,

$$\sigma_0 = -\sum_{i=1}^{L} \sum_{C_i} p_i(C_i) \log p_i(C_i) = \sum_{a=1}^{20} N_a \left(-\sum_{C_\alpha} p(C_\alpha|a) \log p(C_\alpha|a) \right) \tag{17}$$

where C_α are all the codons coding for the amino acid a, $\alpha = 1 \ldots deg(a)$, where $deg(a)$ is the degeneracy of the amino acid. A simple upper bound of σ_0 is obtained by considering all amino acids as having the maximal degeneracy of 6 and all the corresponding codons as equiprobable; in this case $p(C_\alpha|a) = 1/6$ and $\sigma_0 \leq L \log 6$. A more precise upper bound is to take into account the degeneracy of each amino acid $deg(a)$ but still considering each codon coding for the same amino acid as equiprobable; we then obtain the upper bound $\sigma_0 = \sum_a N_a \log deg(a)$.

The absolute entropy of sequences, defined as the logarithm of the typical number of sequences available under pressure x, is then given by

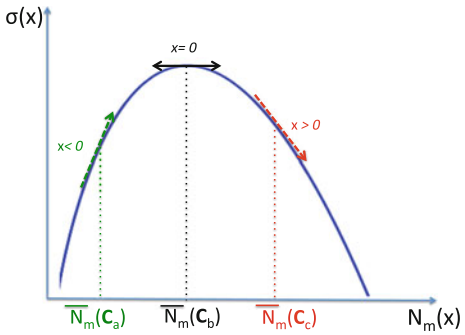

Fig. 1 Sketch of the entropy σ in the random codon model as a function of the number of occurrences of the motif, N_m. The selective pressure x associated to a given genomic sequence C with a number of motifs $\bar{N}_m$ is the derivative of the entropy σ in $N_m = \bar{N}_m$. Three cases are shown: a typical value $\bar{N}_m$ corresponding to the unconstrained case $x = 0$ (*black*, top of entropy curve); $\bar{N}_m$ atypically small, corresponding to a selective pressure $x < 0$; atypically large $\bar{N}_m$, corresponding to a selective pressure $x > 0$

$$\sigma_{tot}(x) = \sigma_0 + \sigma(N_m(x)) \ . \tag{18}$$

A sketch of the absolute entropy curve is plotted as a function of N_m in Fig. 1. The selective pressure x associated to a specific number of occurrence of motifs $\bar{N}_m$ is minus the derivative of the curve $\sigma(N_m)$ in $\bar{N}_m$, see Eq. (14). As shown in Fig. 1 the maximal value of the curve corresponds to the unconstrained case $x = 0$ and is the unconstrained entropy σ_0. Negative values of x constrain the distribution to a smaller number of occurrence of the motif with respect to the unconstrained case, while positive values of it constrain the distribution to a larger number of occurrences of the motif.

In the following section we show how to derive the curve sketched in Fig. 1 by computing, using the transfer matrix technique, the partition function and its derivative, the number of motifs, as a function of x and use Eqs. (16, 18) to obtain the entropy curve. The selective force $\bar{x}$ for a given genome is then obtained from minus the derivative of the entropy curve in $\bar{N}_m$.

2.5 *Practical Implementation with the Transfer Matrix Approach*

We calculate the normalization constant $Z(x)$, Eq. (7), using the transfer matrix formalism. We denote by $C[n : n + K - 1]$ the subsequence of K nucleotides in C, starting at position n and ending up at position $n + K - 1$. The number of occurrences of the motif $m = (m_1, m_2, \ldots, m_K)$ in a random sequence C, see Eq. (5), can be written as

$$N_m(C) = \sum_{n=1}^{3L-K+1} \delta_{C[n:n+K-1],m} \tag{19}$$

The subsequence $C[n:n+K-1]$ spreads over at most $K_c = \mathrm{Int}((K+1)/3)+1$ contiguous codons C_i in C, where Int denotes the integer part. Consider for instance the case of dinucleotide motifs m, for which $K=2$ and $K_c = 2$ according to the formula above. The two nucleotides of such a motif can indeed be found

- at the positions 1, 2 of a single codon, say, C_i; then we have $m_1 = c_{i,1}$, $m_2 = c_{i,2}$.
- at the positions 2, 3 of codon C_i; then we have $m_1 = c_{i,2}$, $m_2 = c_{i,3}$.
- at the position 3 of codon C_i, and position 1 of codon C_{i+1}; then we have $m_1 = c_{i,3}$, $m_2 = c_{i+1,1}$.

For the sake of simplicity we assume that $K=2$; the case of longer motifs can be treated similarly. According to the discussion above we can write

$$N_m(C) = \sum_{i=1}^{L-1} F(m, C_i, C_{i+1}) \ , \tag{20}$$

where

$$F(m, C_i, C_{i+1}) = \delta_{m_1,c_{i,1}}\delta_{m_2,c_{i,2}} + \delta_{m_1,c_{i,2}}\delta_{m_2,c_{i,3}} + \delta_{m_1,c_{i,3}}\delta_{m_2,c_{i+1,1}} \tag{21}$$

for all $i = 1, \ldots, L-2$ and

$$\begin{aligned} F(m, C_{L-1}, C_L) = {} & \delta_{m_1,c_{L-1,1}}\delta_{m_2,c_{L-1,2}} + \delta_{m_1,c_{L-1,2}}\delta_{m_2,c_{L-1,3}} + \delta_{m_1,c_{L-1,3}}\delta_{m_2,c_{L,1}} \\ & + \delta_{m_1,c_{L,1}}\delta_{m_2,c_{L,2}} + \delta_{m_1,c_{L,2}}\delta_{m_2,c_{L,3}} \ . \end{aligned} \tag{22}$$

The expression for F in the bulk of the sequence ($i \leq L-1$) avoids double counting of the motif occurrences.

We now rewrite $Z(x)$ as a sum over the possible codons corresponding to the same amino acids as in the viral sequence C_0:

$$Z(x) = \sum_C \left(\prod_{i=1}^{L} p_i(C_i|a_i) \right) \exp[x \sum_{i=1}^{L-1} F(m, C_i, C_{i+1})] \tag{23}$$

$$= \sum_C \prod_{i=1}^{L-1} (p_i(C_i|a_i) \exp[x F(m, C_i, C_{i+1})]) \, p_L(C_L|a_L), \tag{24}$$

where $p_i(C_i|a_i)$ is the codon bias for codon C_i (coding for the ith amino acid a_i). Let us now define L 'transfer' matrices M_i, $i = 1, \ldots, L$. The dimension of matrix M_i is $\deg(C_i) \times \deg(C_{i+1})$, where $\deg(C)$ is the degeneracy of codon C. The entries of M_i are given by, for all $i = 1, \ldots, L-2$,

$$M_i(C_i, C_{i+1}) = p_i(C_i|a_i) \exp[x F(m, C_i, C_{i+1})] \,, \tag{25}$$

and

$$M_{L-1}(C_{L-1}, C_L) = p_{L-1}(C_{L-1}|a_{L-1}) \exp[x F(m, C_{L-1}, C_L)] p_L(C_L|a_L) \,. \tag{26}$$

Then, we observe that

$$\begin{aligned} Z(x) &= \sum_{C_1, C_2, \ldots, C_{L-2}, C_{L-1}} M_1(C_1, C_2) M_2(C_2, C_3) \ldots M_{L-2}(C_{L-2}, C_{L-1}) M_{L-1}(C_{L-1}, C_L) \\ &= \sum_{C_1, C_L} (M_1 \times M_2 \times \ldots \times M_{L-2} \times M_{L-1})(C_1, C_L) \,, \end{aligned} \tag{27}$$

where $\times$ denotes the matrix product in the formula above. This formula shows that Z can be computed in a time growing linearly with L only. This is a huge gain compared to the original expression of Z, Eq. (7) in main text, which sums up an exponentially large–in–L number of codon configurations.

In practice we define the $\deg(C_L)$-dimensional vector v_L, with entries $v_L(C_L) = 1$ for all codons C_L coding for amino-acid a_L. Then we compute the vector

$$v_{L-1}(C_{L-1}) = \sum_{C_L} M_{L-1}(C_{L-1}, C_L) v_L(C_L) \,. \tag{28}$$

Then, we sum over all possible values for the $(L-1)$th codon, C_{L-1}:

$$v_{L-2}(C_{L-2}) = \sum_{C_{L-1}} M_{L-2}(C_{L-2}, C_{L-1}) \, v_{L-1}(C_{L-1}). \tag{29}$$

The process is iterated until the first codon:

$$v_1(C_1) = \sum_{C_2} M_1(C_1, C_2) \, v_2(C_2). \tag{30}$$

Finally, we obtain the value of the normalization constant through

$$Z(x) = \sum_{C_1} v_1(C_1). \tag{31}$$

When the motif is of longer length, and overlap with K_c contiguous codons, Eq. (20) has to be modified. In general one can write

$$N_m(C) = \sum_{i=1}^{L-K_c+1} F(m, C_i, C_{i+1}, \ldots, C_{i+K_c-1}) \,, \tag{32}$$

where the function F is an obvious extension of Eqs. (21) and (22). The transfer matrix method, shown above can still be used, but at a price of introducing larger transfer matrices M_i.

2.5.1 Example on Two Very Short Sequences

We will first apply the above framework on two simple examples: the derivation of the entropy associated to the number of motifs CpU (the letter p indicates that the nucleotide C and U are consecutive on the phosphate backbone) for the sequences $L = 2$ or $L = 3$ amino acid of type proline, which we will indicate as $C_1 = Pro - Pro$ and $C_2 = Pro - Pro - Pro$. The proline is a $\alpha = 1 \ldots deg(Pro) = 4$ time degenerate amino acid coded by the following codons: $C_1 = CCU$, $C_2 = CCC$, $C_3 = CCA$, $C_4 = CCG$. Considering an uniform codon bias $p(C_\alpha) = 1/4$ the average numbers of occurrence of the motif CpU in the unconstrained case is $\langle N_m \rangle = 0.5$ for C_1 and $\langle N_m \rangle = 0.75$ for C_2.

In Fig. 2 we plot the total entropy $\sigma_{tot}(N_m)$ versus the number N_m of occurrences of CpU for C_1 and C_2. The maximum of the entropy always corresponds to the unconstrained case $x = 0$, and we obtain $\sigma_0 = L \log(4)$ giving 2.77 and 4.16 for the two sequences. In Fig. 2 (left) we plot the entropy for C_1. The two extreme points of the entropy curve corresponds to $\langle N_m \rangle = 0$, $\sigma = 2.197$: there are $e^{2.197} = 9$ sequences compatible with ProPro without CpU, and for $\langle N_m \rangle = 2$, $\sigma = 0$: there is a single sequence compatible with ProPro and including 2 CpU. For $\langle N_m \rangle = 1$ we obtain $\sigma = 2.472$ and e^σ is larger than 6 (the number of sequences compatible with

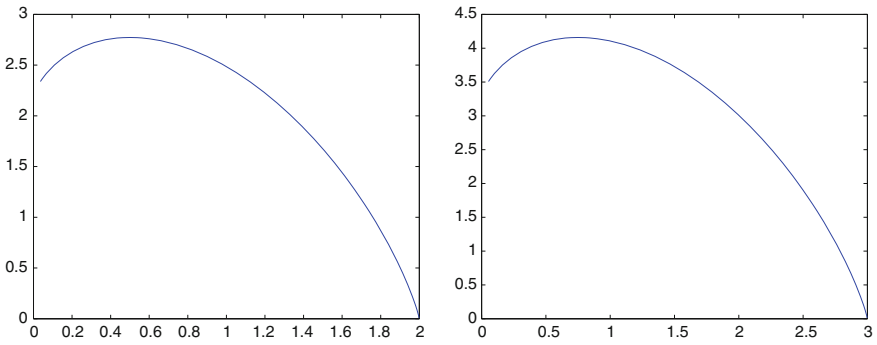

Fig. 2 Entropy σ_{tot} of sequences $C_1 = Pro - Pro$ (*left*) and $C_2 = Pro - Pro - Pro$ (*right*) as functions of the average number of occurrences of the motif CpU

ProPro with one CpU). This is because $\langle N_m \rangle$ does not coincide with N_m. As illustrated above we calculate the entropy of sequences that contain in average $\langle N_m \rangle$ repetitions of the motif, and not exactly N_m repetitions of the motif. Only for large values of N we expect that N_m will coincide with $\langle N_m \rangle$ up to negligible relative fluctuations. The entropy of sequences containing exactly 0 times the motif or two times the motif coincides with what we calculate because there is only one way to obtain zero time the motif (neither in the first nor in the second codons) or two times the motif (both in the first and in the second codons). In Fig. 2 (right) we plot the entropy curve for C_2. The total entropy of sequences with zero occurrence of the motifs is $\sigma \simeq 3.3$ and the number of sequences with zero occurrence of the motif is $e^{3.3} = 27$. The number of sequences with 3 times the motif is $\exp(\sigma)$, with $\sigma \simeq 0$.

2.5.2 Illustration on a Influenza B Sequence

In Fig. 3 we show the entropy curve obtained for an influenza B sequence with respect to the dinucleotide motifs *CpG* (left) and *ApC* (right) and with the segment codon bias. Influenza B is a virus for which humans have been a natural host for many centuries. As expected the number of CpG dinucleotides varies little over time. The green line correspond to the maximal unconstrained entropy $\sigma_0 \simeq \sum_a N_a deg(a)$ which is the same in the two cases. The red value correspond to the occurrence of number of *CpG* and *ApC* motifs in a typical sequence for Influenza B. For *ApC* the curve is quite flat (weak pressure x), hence the number of occurrences of ApC dinucleotides may largely and randomly vary. On the contrary for the *CpG* motif the selective force corresponding to the influenza B genomic sequence is large and negative, indicating that there is an important selective pressure to reduce the number of *CpG* in the sequence. The entropy of random sequences with the same number of *CpG* motifs and the same selective pressure is largely reduced with respect to the maximal, unconstrained value.

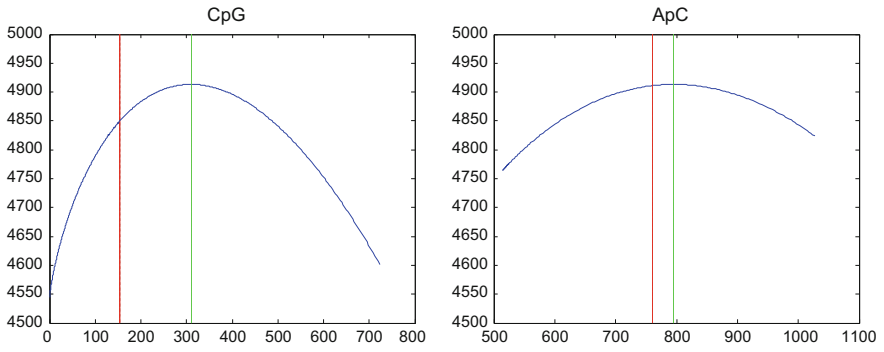

Fig. 3 *Left* Entropy σ of a influenza B isolate with its own codon bias for the dinucleotide CpG. *Right* Entropy σ of an influenza B isolate with its own codon bias for the dinucleotide ApC

2.5.3 Finding Quickly the Right Value for x

An important problem is to find the values of the entropy and of x, hereafter called $\bar{x}$, corresponding to the number $\bar{N}_m$ of occurrences of the motif in the real virus sequence. One way to do this is to compute the entropy, $\sigma(x)$, and the average number of occurrences, $N_m(x)$, for many values of x on a grid and try to be as close as possible to the data, i.e. choose $\bar{x}$ such that $N_m(x) \simeq \bar{N}_m$. A much faster procedure is the following. Consider the function

$$G(x) = \log Z(x) - x\bar{N}_m. \tag{33}$$

Two important facts about G are:

- G is a convex function of x, as its second derivative is positive:

$$\frac{d^2}{dx^2} G(x) = N_m^2(x) - N_m(x)^2 \geq 0. \tag{34}$$

- the first derivative of G vanishes when x takes the value we are looking for, since

$$\frac{d}{dx} G(\bar{x}) = N_m(\bar{x}) - \bar{N}_m = 0\ . \tag{35}$$

Hence, G has a unique minimum in $x = \bar{x}$, and we can find it very quickly with standard optimization techniques, e.g. the Newton-Raphson algorithm. Here is the procedure:

1. Start with $x = 0$
2. Compute the first and second derivatives of G in x, that is, $D_1 = N_m(x) - \bar{N}_m$ and $D_2 = N_m^2(x) - N_m(x)^2$.
3. compute the new value of x (which would be exact if G were a parabolic function)

$$x \to x - \frac{D_1}{D_2}\ . \tag{36}$$

4. Iterate step 2 until convergence is achieved.

As the parabolic approximation is generally good, we can expect that the procedure will converge very fast, in a few iterations.

2.6 *Results on Selective Pressures on Viral Sequences*

In Greenbaum et al. (2014) we have applied the above approach to influenza and HIV viral sequences. Here we recall some of the main results.

2.6.1 Influenza

We have first computed the selective force on all 16 possible dinucleotide motifs for the eight longest open reading frames from the lineage of H1N1 viruses that descend from the 1918 pandemic influenza. In Fig. 4 we show the results focusing on four dinucleotides most frequently found to be anomalous motifs and only on the PB2 gene influenza, which is the longest gene. We observe that

- The motif with the largest negative selective pressure is dinucleotide *CpG*; for this motif there is a clear evolution of the selective pressure from year 1918 when H1N1 entered the human population to much lower values, corresponding to influenza B, which has been in the human population since hundreds of years. The selective pressure has become more and more negative and the number of *CpG* dinucleotides has been lowered in the course of the viral evolution to adapt the viral sequence to the human host and avoid recognition by the immune system, which would recognize large numbers of *CpG* motifs.
- The vast majority of motifs, not represented in Fig. 4, see Fig. 2a of Greenbaum et al. (2014), have $x = 0$ when using the segment codon bias and x going from

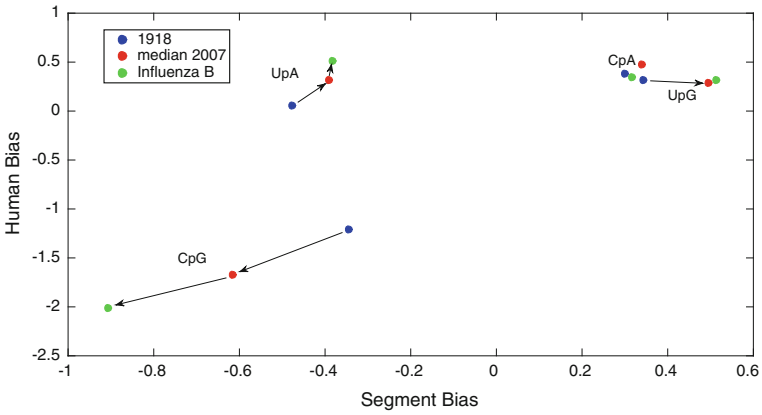

Fig. 4 A comparison of the selective pressures when calculated using the segment and human codon biases for the four dinucleotides CpA, CpG, UpA and UpA for the PB2 gene in influenza. These quantities are calculated for the 1918 H1N1, the H1N1 segments from 2007 and for Influenza B. In the later two cases the median values are shown. The arrows follow the evolution of the flu from the H1N1 1918 influenza through 2007 to influenza B (present in humans for a very long time)

$x = -1$ to $x = 1$ when using the human codon bias. This result shows that even if the virus codon bias is very similar to the one of the host it is not yet completely equivalent.

- The dependence of the selective force on the segment similarity is not very large, as shown here for PB2, it is only noticeable for CpG dinucleotides.

2.6.2 HIV

For HIV we show in Fig. 5 the selective force on six dinucleotide motifs for the Pol gene. Points of interest include:

- As for influenza sequences the motif with largest and negative pressure is CpG.
- Likewise, the vast majority of motifs have $x = 0$ when using the human codon bias and x going from $x = -1$ to $x = 1$ when using the human codon bias.
- There is some dependence on the type of protein and on the region of the sequence (not shown here, see Fig. 4d and Supplementary material in Greenbaum et al. (2014)), likely reflecting that HIVs genome codes for multiple proteins and, as a retrovirus, is targeting by many innate defense mechanisms (Vabret et al. 2016).
- There is not much dependence on the HIV subtype, showing that there is not a large evolutionary trend between different types of HIV virus which therefore seems to be already in equilibrium with respect to the small dinucleotide motif usage. This likely reflects that whereas influenza entered humans from avian and swine hosts, HIV came from primates, which are closer evolutionary species.

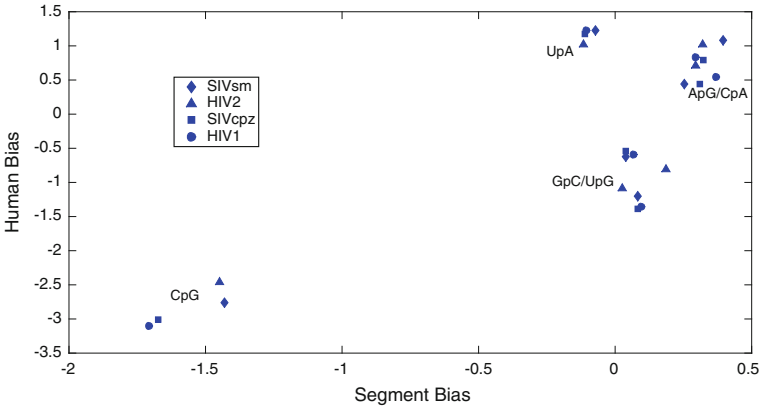

Fig. 5 A comparison of the selective pressures when calculated using the segment and human codon biases for six dinucleotides for the for the Pol genes in HIV. These quantities are calculated for the HIV1, HIV2, SIVcpz and SIVsm

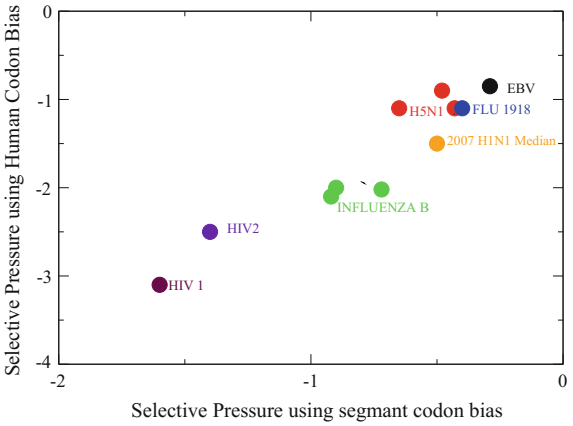

Fig. 6 Comparison of selective pressures for CpG dinucleotides using both segment codon bias and human codon bias for different viruses: influenza virus (segment PB2) with the 1918 H1N1 sequence, and the median values for all 2007 H1N1 and three influenza B segments. We also show results for Ebola virus and HIV pol (showing median values for HIV-1 and HIV-2)

2.6.3 Comparison of Different Viruses: Relationship Between the Selective Pressure and the Virulence of the Virus

The advantage of the approach presented here is that the forces associated with a given genomic sequence is an intensive variable; it is then independent of the length of the sequence and therefore different viral sequences can be compared. In Fig. 6 we compare the selective forces on *CpG* motifs for the 1918 H1N1 influenza sequence, for the median sequence from 2007 H1N1, and for the median sequence of recent Ebola virus and for the HIV1 and HIV2 median Pol sequences. Interestingly Ebola, 1918 H1N1 and 2007 H1N1 cluster together at values of the selective force which are weakly negative, while for influenza B and HIV they are much larger and negative. There is therefore a large correlation between a value of the selective pressure larger than the 'stationary' equilibrium value for influenza B and the degree to which these sequences have evolved in humans or closely related species, which may also be associated with an aberrant innate response.

3 Further Applications of the Statistical Physics Approach to Detect Anomalous Motif Usage

3.1 Monte Carlo Simulations of the Evolutionary Dynamics of Sequences

In Greenbaum et al. (2014) we have investigated a simple general dynamical model which describes the evolution of the selective pressure in the H1N1 flu virus to reach the equilibrium value:

$$\tau \frac{dN}{dt} = -x(N_m(t)) + x_{eq} \tag{37}$$

where $N(t)$ is the number of occurrences of motif m at time t. The underlying idea was directly inspired from the so-called Langevin relaxation equation of statistical physics: the dynamical variable (here, the number of motifs) relaxed to an equilibrium value where the forces acting on this variable (here, the selective and entropic pressures) balance each other. We assumed that influenza B is at equilibrium, given that the number of CpG motifs in that virus did not change much over the same time scales under which a substantial change was observed in H1N1. We therefore estimated the equilibrium pressure x_{eq} as the mean value of the pressures computed for the set of influenza B sequences. We chose for initial condition the H1N1 sequence from 1918, which had a well defined number of motifs, N_0, and the corresponding pressure, x_0.

We have solved Eq. (37) and obtained the instantaneous selective pressure $x(t) \equiv x(N(t))$, where t is the years of evolution from 1918. The time scale τ was tuned to make $x(t)$ fit best with H1N1 data over the available time range. As the pressures were (in absolute value) of the order of the unity, τ could be interpreted as the typical times it takes for the virus to decrease or increase its number of motifs by unity (see Fig. 3 in Greenbaum et al. (2014) and the values of x_B, x_0, and τ given in Table 1 of this reference).

Here we report new Monte Carlo (MC) simulations of a microscopic mutational model for the sequence of codons (with fixed amino-acid content) under constant selective pressure, denoted by x_s and supposed to be negative. The MC algorithm works in discrete time $T = \Delta t, 2\Delta t, 3\Delta t, \ldots$ as follows, from an initial sequence $C = (c_1, c_2, \ldots, c_L)$ of codons at time $T = 0$:

1. at each time step $T \to T + \Delta t$ a site i is chosen uniformly at random between 1 and L;
2. a codon C' corresponding to the ith amino acid a_i is chosen at random with probability $p_i(C'|a_i)$. If $C' = C_i$ the algorithm loops to step 1.
3. if $C' \neq C_i$ we compute the change in the number of motif occurrences ΔN_m. The move $C_i \to C'$ is always accepted if $\Delta N \leq 0$, and is accepted with probability $\exp(x_s \Delta N_m)$ if $\Delta N_m > 0$. The algorithm then loops to step 1.

This microscopic dynamics obeys detailed balance (i.e. corresponds to a general time-reversible process) and is guaranteed to converge to equilibrium at large enough times. We show in Fig. 7 typical runs of the MC algorithm for various values of the pressure (see caption). We compare the behaviour of $N_m(T)$ with the solution of (37), and observe a very good agreement of the two curves provided the elementary time-step is chosen to be $\Delta t \simeq \tau/250$.

The Monte Carlo algorithm can be used to artificially evolve sequences, starting from an initial sequence, say, the 1918 H1N1. As time goes on, the content in amino acids remains fixed, but the nucleotidic sequence changes. When the MC dynamics is stopped the resulting codon sequence may have very different

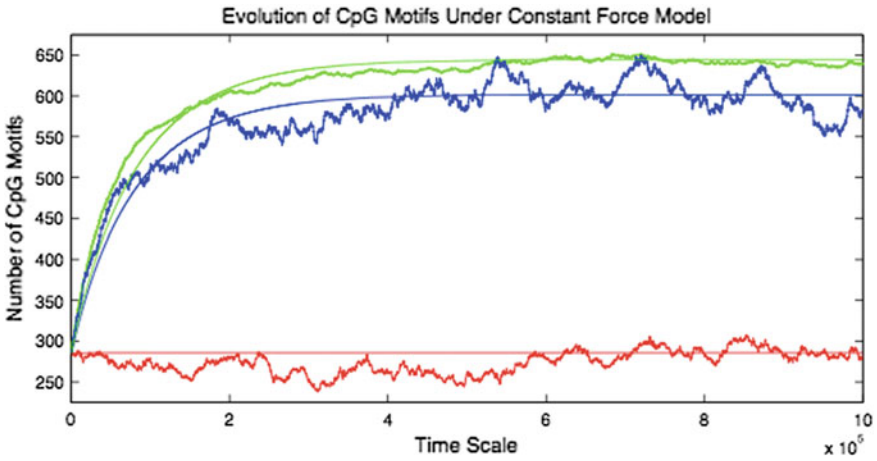

Fig. 7 Monte Carlo dynamics compared to average number of CpG motifs for three constant selective pressure values: 0, −0.119, and −1.19. These pressure values are shown in green, blue, and red respectively. In the last case the selective pressure was roughly the same as the one of the 1918 H1N1, which is the initial condition for all three trajectories

properties (compared to the initial sequence) in term of stimulation of the immune response, and can in particular be much less immuno-stimulatory, if the number of *CpG* motifs has been reduced under the action of the selective pressure.

3.2 *Entropy of Multiple Motifs*

To calculate the entropy associated with the number of occurrences of several motifs, one can extend the formalism of Sect. 2. As an example, for two dinucleotides the partition function will vary over two parameters (x_1, x_2) corresponding to dinucleotide motifs $m_1 = (m_{11}, m_{12})$ and $m_2 = (m_{21}, m_{22})$. The partition function naturally becomes

$$Z(x_1, x_2) = \sum_C p(C) \exp\left[x_1 \sum_{i=1}^{L-1} M_{1i}(C_i, C_{i+1}) + x_2 \sum_{i=1}^{L-1} M_{2i}(C_i, C_{i+1})\right], \quad (38)$$

where $M_{1i}(C_i, C_{i+1})$ is the previously defined matrix $M_i(C_i, C_{i+1})$ for the motif m_1., and M_{2i} its counterpart for motif m_2. The Legendre transformation will become

$$\sigma(x_1, x_2) = \log Z(x_1, x_2) - x_1 N_{m1}(x_1, x_2) - x_2 N_{m2}(x_1, x_2), \quad (39)$$

where

$$N_{m_1}(x_1, x_2) = \frac{\partial}{\partial x_1} \log Z(x_1, x_2) \tag{40}$$

and likewise for $N_{m2}(x_1, x_2)$. Then the average number of occurrences of motif m_1 can be computed from the partial derivative of Z with respect to x_1,

$$\langle N_{m_1} \rangle = \frac{\partial}{\partial x_1} \log Z(x_1, x_2)\Big|_{x_1=x_2=0}. \tag{41}$$

Similarly, the joint moments of the numbers of occurrences of m_1 and m_2 can be obtained from higher derivatives with respect to x_1 and x_2.

An application of the di-motif formalism is shown in Fig. 8, where we plot the entropy surface as a function of N_{UpA} and N_{CpG}. The value of the entropy constrained to the measured number of occurrence N_{UpA} and N_{CpG} in a particular sequence is smaller than the unconstrained, maximal value. The pressures $x_{ApC+CpG}$ and $x_{CpG+ApC}$ are the derivative of the entropy curve along the two axes.

An interesting question is if the selective pressures for multiple motifs are coupled, i.e. are different from the values obtained by considering one motif at a time. In Fig. 9 we compare the uncoupled (red dots) and coupled (blue) pressures for four motifs in PB1 segment. Results show that the UpA motif is essentially independent from the CpG one, as the values of the pressure for the uncoupled RCM are very similar to the one found for the coupled UpA + CpG RCM. On the

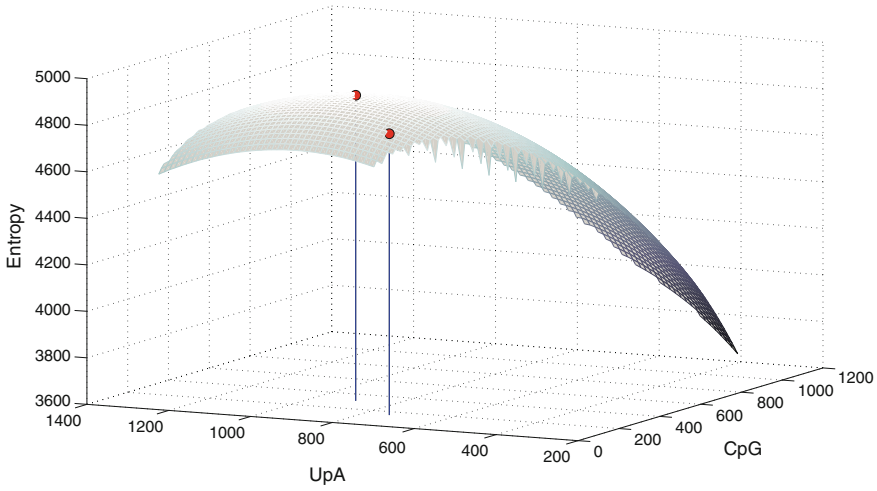

Fig. 8 Entropy σ of influenza sequences with their own codon bias for the dinucleotides CpG and UpA. Results were obtained from the eight longest coding regions of the influenza B virus (B/Cordoba/2979/1991)

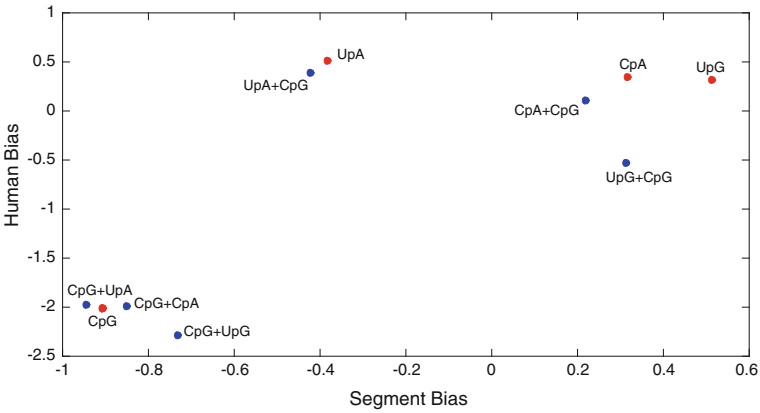

Fig. 9 Selective pressure calculated with the human codon bias and the segment codon bias for four dinucleotidic motifs in the PB1 segment of influenza B virus, calculated with RCM with two coupled motifs (*blue dots*) compared to the ones calculated with RCM with one single motif (*red dots*). For the two–motif model the selective pressure refers to the first motif in the label

contrary the selective pressures on CpA and UpG are not independent from the one of CpG. This coupling presumably originates from the fact that CpA and UpG are the mutational partners of CpG: diminishing the number of CpG motifs naturally increases the number of its mutational partners.

3.3 Geometrical Nature of the Sequence Space

So far, we have computed the entropy, that is, the log of the effective number of sequences (under some pressure). However, we do not have any information about the way those sequences are arranged in the configuration space. Are they spread over the whole configuration space or are they clustered in one tiny region? Our statistical physics formalism can however help us gain some intuition about the spatial organization of sequences as shown below.

3.3.1 Two-Sequence Formalism

Consider the following partition function, for a two-sequence system (instead of one-sequence system we have focused on so far):

$$Z_2(x,x',y) = \sum_{\{C,C'\}} \prod_{i=1}^{L} p_i(C_i|a_i)p_i(C_i'|a_i) \exp\left[x\sum_{i=1}^{L-1} M_i(C_i,C_{i+1}) + x'\sum_{i=1}^{L-1} M_i(C_i',C_{i+1}') + y\sum_{i=1}^{L}\delta_{C_i,C_i'}\right] \tag{42}$$

When $y = 0$, we simply have two independent sequences, one under pressure x and one under pressure x':

$$Z_2(x, x', y) = Z(x) \times Z(x') \ , \tag{43}$$

where $Z(.)$ is the partition function we have considered so far.

When y is not equal to zero, the two sequences are coupled according to their similarity. The weight associated to a set of two sequences is proportional to $\exp(y n_2)$; here n_2 is the number of codons equal on both sequences, it is also equal to $L - D$ where D is the Hamming distance between the two sequences (measured at the codon level, not at the base level).

We now define the average values of the number of motifs in each sequence, the average value of common codons, n_2, and a new entropy, σ_2:

$$\begin{aligned} N_m(x, x', y) &= \frac{\partial \log Z_2}{\partial x}(x, x', y), \quad N'_m(x, x', y) = \frac{\partial \log Z_2}{\partial x'}(x, x', y), \\ n_2(x, x', y) &= \frac{\partial \log Z_2}{\partial y}(x, x', y), \\ \sigma_2(x, x', y) &= \log Z_2(x, x', y) - x N_m(x, x', y) - x' N'_m(x, x', y) - y n_2(x, x', y) \ . \end{aligned} \tag{44}$$

If we choose the two pressures x and x', and we let y vary, then we can plot in a parametric way the entropy σ_2 as a function of n_2. This way, we will know how many pairs of sequences are located at a distance $d = L - n_2$. In the next paragraph we will see how this distance-dependent entropy changes as the pressures change. In general, we can choose $x = x'$ as both sequences are under the same pressure.

From a practical point of view, the calculation of Z_2 can be done along the same lines as the one of Z. The only difference is that the vectors v to be iterated are not functions of C_i only, but are now functions of both C_i, C'_i. So the maximal number of components of v is 36 instead of 6, making the computation only slightly slower.

3.3.2 Practical Implementation: Entropy as Function of Distance Between Sequences

We consider the following problem. We choose the codon bias, say, the human one, and one virus sequence, say, 1918 H1N1, and one motif, say, *CpG*. Let $\bar{N}_m$ be the number of motifs in the viral sequence, which defines the amino-acid set and the allowed codons, i.e. the probabilities $p_i(C_i|a_i)$ for all i. We want to know how many sequences (weighted by the codon bias) there are a that share n_2 codons. We consider the function

$$G(x, y; n_2) = \log Z_2(x, x, y) - 2x\bar{N}_m - yn_2. \tag{45}$$

Note that we have chosen $x = x'$ here and note also the presence of the factor 2. The variable n_2 is a positive parameter, smaller than the sequence length (measured in codons). Now, for any n_2, we can optimize G over x and y using Newton's method. The result is

$$\sigma_2(n_2) = \min_{x,y} G(x, y; n_2) \ . \tag{46}$$

The interpretation is that $\sigma_2(n_2)$ is the entropy of sequences with similarity (number of equal codons) n_2 (we neglect here the contributions coming from the fact that the average number of motifs depends on y). The maximum of the curve will be reached in n_2^*, corresponding to $y = 0$ and to the same value of x and the same entropy found in the standard one-sequence calculation. If $n_2 \neq n_2^*$, x will take a different value.

As an example of how one can interpret our results in terms of the geometry of a space of sequences, we calculate the sequence similarity for the genes of HIV and influenza. This measure shows the typical number of shared codons for two sequences drawn randomly from the distribution of possible sequences. In this case, the quantity is computed for each individual sequence when these sequences are under the derived entropic force. The average similarity (number of identical codons) between two random sequences drawn from the same codon distribution is defined as

$$n_2(x) = \sum_{C,C'} P(C|x)\ P(C'|x)\ \sum_{i=1}^{L} \delta_{C_i,C'_i} \tag{47}$$

where δ_{C_i,C'_i} equals one if the two codons at the i-th position are equal and is zero otherwise. Sequences with a large degree of similarity are close together in the space of possible sequences. In our case, for individual sequences, this would measure how close together sequences are with the same amino-acid distribution once a pressure is applied to a motif, or a set of motifs.

We plot the sequence similarity as a function of the entropy for the PB2 segment of the H1N1 virus in Fig. 10 and in Fig. 11 for the Pol gene in HIV. In much the same way as what was previously observed for the selective pressures, the similarity between sequences calculated with the RCM using the human codon bias are different to the ones obtained using the virus codon bias. The similarity is generally lower when the human codon bias is used for the background distribution rather than the bias for that segment. Overall while there is more similarity between random sequences when the segment bias is used, the difference in similarity

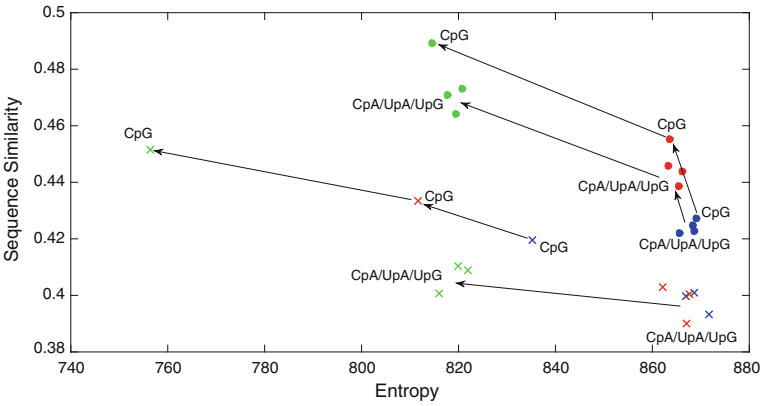

Fig. 10 Normalized sequence similarity n_2/L versus Entropy for PB2 from H1N1 flu virus (*blue* 1918 H1N1 sequence, *red* 2007 H1N1 sequence, *green* Flu B sequence for comparison). Crosses indicate the human codon bias while circles the segment codon bias

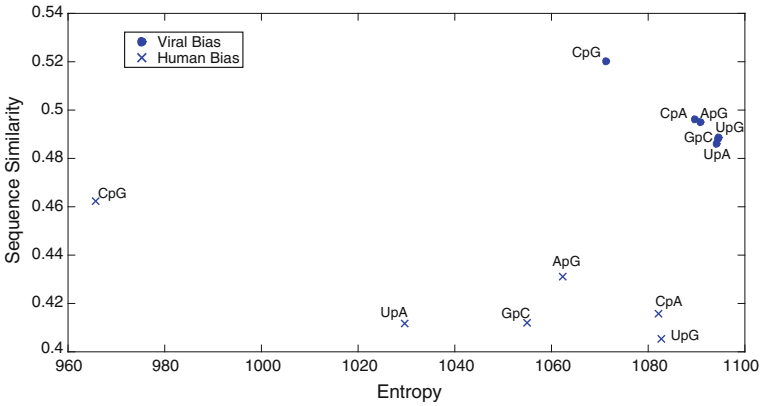

Fig. 11 Normalized sequence similarity n_2/L versus Entropy for the HIV genome from Pol HIV-1. Crosses indicate the human codon bias while circles the segment codon bias

between motifs is much larger when the human bias is used. In influenza B, with respect to the segment codon bias, the difference in similarity between CpG and other dinucleotides is much lower than the difference for the human bias.

As a general trend, for a fixed codon bias, large selective pressures lead to greater degree of similarity between sequences. The pressures, by making sequences less random, make the resulting distribution of sequences more concentrated. As expected, this effect is strong for *CpG*.

4 Out-of-Frame Stop Codons and the Ambush Hypothesis

4.1 The Ambush Hypothesis: Brief Review of Literature

Considering the deleterious effects of ribosome frame-shifts during translation Seligmann and Pollock (2004) introduced the Ambush Hypothesis according to which such deleterious effects can be avoided owing to the existence of off-frame STOP codons (OSC). This hypothesis was initially tested by Seligmann and Pollock in vertebrate mitochondrial genes (Seligmann 2010; Seligmann and Pollock 2004) and later extended to the case of prokaryotic genomes (Morgens et al. 2013; Tse et al. 2010; Wong et al. 2008). The latest study of the abundance of OSCs in prokaryotic genomes (Morgens et al. 2013) led to the conclusion that there was no statistical evidence for the existence of a correlation between a codon's usage and its propensity to form OSCs which would have been a strong evidence for the validity of the Ambush Hypothesis. Indeed, in all previous studies, the occurence of OSCs was largely dominated by the AT content of the studied genomes, and clear-cut conclusions were difficult to extract.

Here, we re-address this question along two different lines. First, we adopt a different approach in comparison with previous statistical studies. Our starting point is that apparition of an OSC involves 2 adjacent codons and thus measurement of their abundance should involve the use of the statistics of apparition of dicodons instead of mere single codons. We therefore introduce the notion of dicodon bias analogous to the well-known codon bias and refer this dicodon bias to a null model in which successive codons appear in a non-correlated way (Coleman et al. 2008; Long et al. 1998). We will adopt conventional notations for the frameshift of an OSC: within a dicodon an OSC is of type

$$\begin{cases} +1, & \text{if the OSC}'\text{S first nucleotide is the second nucleotide of the dicodon;} \\ -1, & \text{if the OSC}'\text{S first nucleotide is the third nucleotide of the dicodon.} \end{cases}$$

The study presented here is based on the use of the bacteria RefSeq database of NCBI, from which 1852 genomes of single chromosome bacterial species have been analyzed (the reduction in number of the RefSeq database was performed in order to avoid over-representation of specific bacterial species since for instance *Escherichia coli* species is represented by 173 strains in the initial database).

Secondly, since the outcome of the statistical analysis does not show any significant bias supporting the Ambush Hypothesis across all genomes, we ask whether modifying the statistics of nucleotides is actually necessary to have many OSC. To do so, we consider the random codon model of Sect. 4.2, and compute analytically within this model the distribution of distances to the first OSC after a frameshift equal to $+1$ or -1. We show that the distribution of distances decay very quickly as the distance increases, with an average distance of less than ten codons for both frameshifts. Note that this value is robust against the choice of the initial condition, i.e. also corresponds to the average distance to an OSC even if the

frameshift takes place at any location in the coding sequence (not necessarily at the beginning). Our theoretical result is corroborated by the statistical analysis of genomic sequences, and thus strongly suggests that the Ambush Hypothesis is not required to have many OSC.

4.2 Statistical Analysis of Dicodons Biases

4.2.1 Definitions and Notations

In order to quantitatively assess the occurrence of OSC within a genome we introduce the general notion of an average dicodon bias $\langle DCB_\alpha \rangle$ for dicodons belonging to a particular class α; this average dicodon bias is defined as:

$$\langle DCB_\alpha \rangle = \sum_{a,a'} p(a,a') \sum_{c,c'} (dcb(c,c') - cb(c)cb(c')) I_\alpha(c,c') \tag{48}$$

Here c (resp. c') stands for a codon and $cb(c)$ (resp. $cb(c')$) stands for the corresponding codon bias according to its usual definition, i.e. for a given amino acid a, if c codes for a, $cb(c)$ is the probability of c being chosen over all possible codons coding for a; (c,c') stands for the dicodon formed by c followed by c' and $dcb(c,c')$ stands for the dicodon bias of (c,c'). The notation a (resp. a') stands for the amino acid coded by c (resp. c'); $p(a,a')$ stands for the probability of occurence of the diamino acid (a,a'). $I_\alpha(c,c')$ is an indicator of the membership of dicodon (c,c') to a specific class α (to be specified below), and takes values 0 and 1 according to whether or not dicodon (c,c') belongs to class α. At fixed (a,a') the sum is performed over all codons c and c' coding respectively for a and a'. The definition of a dicodon bias is entirely analogous to the definition of a codon bias, i.e. for a given diamino acid (a,a') coded by (c,c') the dicodon bias for (c,c') is the probability for (c,c') to be chosen over all possible dicodons coding for (a ,a').

It should be pointed out that definition (48) of an average dicodon bias for dicodons belonging to a specific class α is a direct measure of the excess of appearance of dicodons belonging to class α with respect to the hypothesis of uncorrelated appearance of codons forming the dicodons. In addition, this estimator does not make any assumption about the statistics of di-amino acids, likely to be correlated in real coding sequences. $\langle DCB_\alpha \rangle$ can be conveniently rewritten as the dot product of 2 vectors $\vec{Y}$ and $\vec{C}_\alpha$:

$$\langle DCB_\alpha \rangle = \vec{Y} \cdot \vec{C}_\alpha \tag{49}$$

where the components of $\vec{Y}$ and $\vec{C}_\alpha$ are given by:

$$Y(c,c') = \sqrt{\omega(c,c')}X(c,c'), \quad C_\alpha(c,c') = \sqrt{\omega(c,c')}I_\alpha(c,c'), \tag{50}$$

with:

$$X(c,c') = \frac{dcb(c,c')}{cb(c)cb(c')} - 1, \quad \omega(c,c') = p(a,a')cb(c)cb(c'). \tag{51}$$

$\vec{Y}$ and $\vec{C}_\alpha$ are vectors of size $63 \times 63 = 3969$ corresponding to the formation of all possible dicodons once excluded the codon TAG which codes for non standard pyrrolysine amino acid only found in methanogenic archaea.

In order to calculate this average dicodon bias for each genome in the collection of the 1852 genomes selected from the RefSeq database, we have extracted the codon content of each CDS as well as its dicodon content; from those contents it is then easy to deduce the quantities of interest in our analysis: codon bias, dicodon bias and probability of appearance of (a,a'). In analyzing the CDS sequences the initial START codon and the sense STOP codon were excluded.

4.2.2 Statistical Significance of Calculated Values of $\langle DCB_\alpha \rangle$

Due to the limited number of codons belonging to a specific class α, it is of interest to be able to test the statistical significance of the calculated value of $\langle DCB_\alpha \rangle$. In order to perform such a test we adopt the following procedure. If n_α is the number of dicodons belonging to class α we perform **N** random permutations amongst the n_α non-zero values of the indicator $I_\alpha(c,c')$ and calculate the **N** obtained values of $\langle DCB_{\alpha,test} \rangle$; from this distribution of values of $\langle DCB_{\alpha,test} \rangle$ we then calculate a standard deviation and normalize the value of $\langle DCB_\alpha \rangle$ for the considered class with respect to this standard deviation (z-score). Following this normalization procedure a value of $\langle DCB_\alpha \rangle$ is considered as statistically significant if it is greater than 2 (in absolute value), which means away from the mean by more than twice the standard deviation of the distribution of n_α randomly chosen dicodons.

In the following we will introduce 4 classes of dicodons:

1. Class $+1$ for which (c,c') contains an OSC in the frame $+1$ associated to $\langle DCB_{+1} \rangle$;
2. Class -1 for which (c,c') contains an OSC in the frame -1 associated to $\langle DCB_{-1} \rangle$;
3. Class ± 1 for which (c,c') contains an OSC in any frame ($+1$ or -1) associated to $\langle DCB_{\pm 1} \rangle$;
4. Class *identical* for which $c = c'$ associated to $\langle DCB_{id} \rangle$.

The first (resp. second) class refers to all dicodons containing an OSC in the frame $+1$ (resp. in the frame -1); the third class refers to all dicodons containing

an OSC in whichever frame. The fourth class refers to all dicodons constituted of 2 identical codons. As a matter of example we give below the values of $I_{id}(c, c')$ for the fourth class:

$$I_{id}(c, c') = \begin{cases} 1, & \text{if } c = c'; \\ 0, & \text{otherwise.} \end{cases}$$

This fourth class is not related to the Ambush Hypothesis but will be used to validate our statistical analysis below.

In order to illustrate the statistical test explained above we present in Fig. 12 the probability density function of the $\mathbf{N}\,(=10,000)$ random permutations amongst the n_α non-zero values of the indicator $I_\alpha(c, c')$ ($n_{id} = 63$, $n_{+1} = n_{-1} = 192$ for, respectively, classes *identical*, $+1$ and -1) in the case of 2 specific genomes (*E. coli* and *Lysteria monocyogenes*).

4.2.3 Results

In a first step we report in Fig. 13 the normalized average dicodon biases for Class $+1$, Class -1, Class ± 1 and Class *identical* across all bacterial genomes. As explained above all values of $\langle DCB_\alpha \rangle$ for each genome are normalized by the standard deviation of similar distributions obtained for each studied genome and will be denoted by $\langle DCB_\alpha \rangle_{\text{norm}}$. The bottom panel of Fig. 13 refers to Class *identical*; for this class of dicodons the average bias is overall positive meaning that for the coding of 2 successive identical amino acids there is a bias towards choosing 2 identical codons. One should point out that this effect is rather weak and at the limit of being statistically significant.

The 2 upper panels refer to Class $+1$ and Class -1; quite obviously for a vast majority of genomes the Class $+1$ dicodons exhibit a bias that can be considered as showing no statistically significant deviation from 0. More interestingly the situation is quite different for Class -1 dicodons, which exhibit a statistically significant overall negative value. Grouping these 2 classes gives the third class Class ± 1, for which the overall tendency of dicodon bias values is negative (as shown on third panel from top on Fig. 13).

Before further discussing these first results we still have to test our estimator of the dicodon biases against any strong bias with respect to AT content of the considered sequences. We present in Fig. 14 the same quantities as above plotted against AT contents of the genomes used for our analysis. Quite obviously for the 4 classes of dicodons tested here there is no evidence of a strong bias of our estimator with respect to the AT content of the investigated genomes; this seems to justify our claim that our estimator for dicodon bias is a better estimator as compared to previously used estimators, see Sect. 4.1.

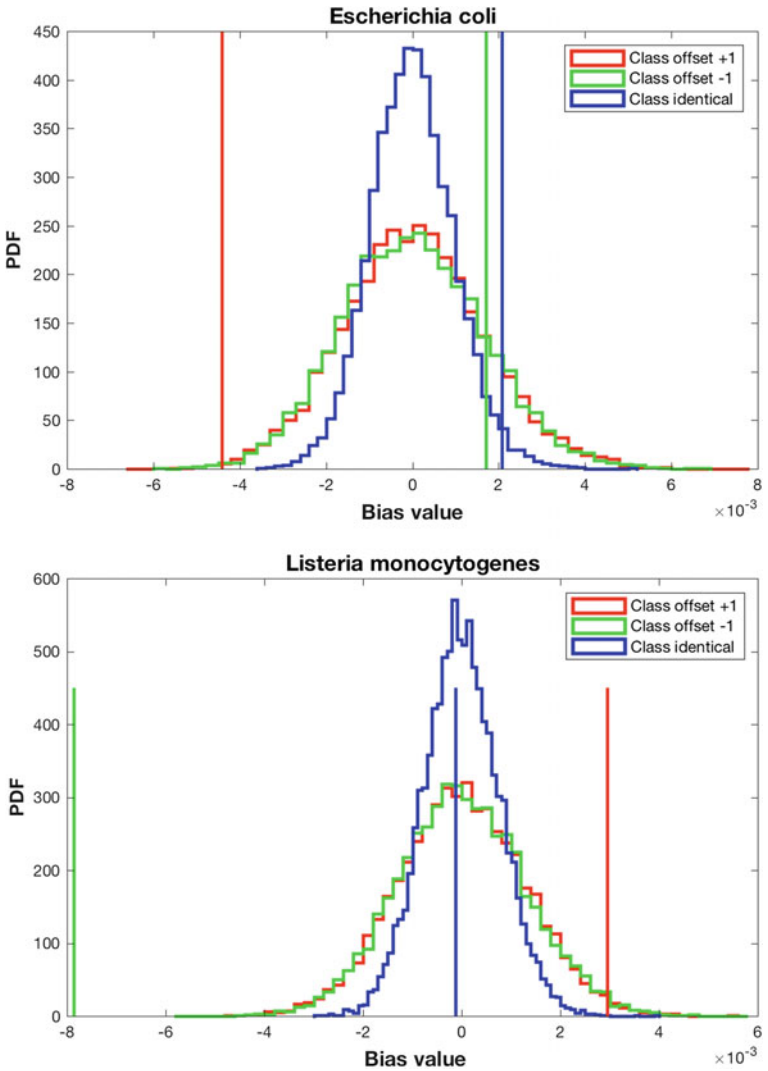

Fig. 12 Distributions of dicodon biases $\langle DCB_\alpha \rangle$ for the 3 classes *Identical*, $+1$, and -1 for two bacterial genomes, obtained by randomly reshuffling the components of vectors $\vec{C}_\alpha$, see text. *Vertical colored bars* give the values of $\langle DCB_\alpha \rangle$ for each class computed from the data. Clearly the measured value of $\langle DCB_{id} \rangle$ for *Lysteria monocyogenes* is statistically not meaningful (see position of the *vertical blue line* in the bottom panel), whereas for the same genome the value of $\langle DCB_{-1} \rangle$ is statistically meaningful (see *vertical green line* in the same bottom panel)

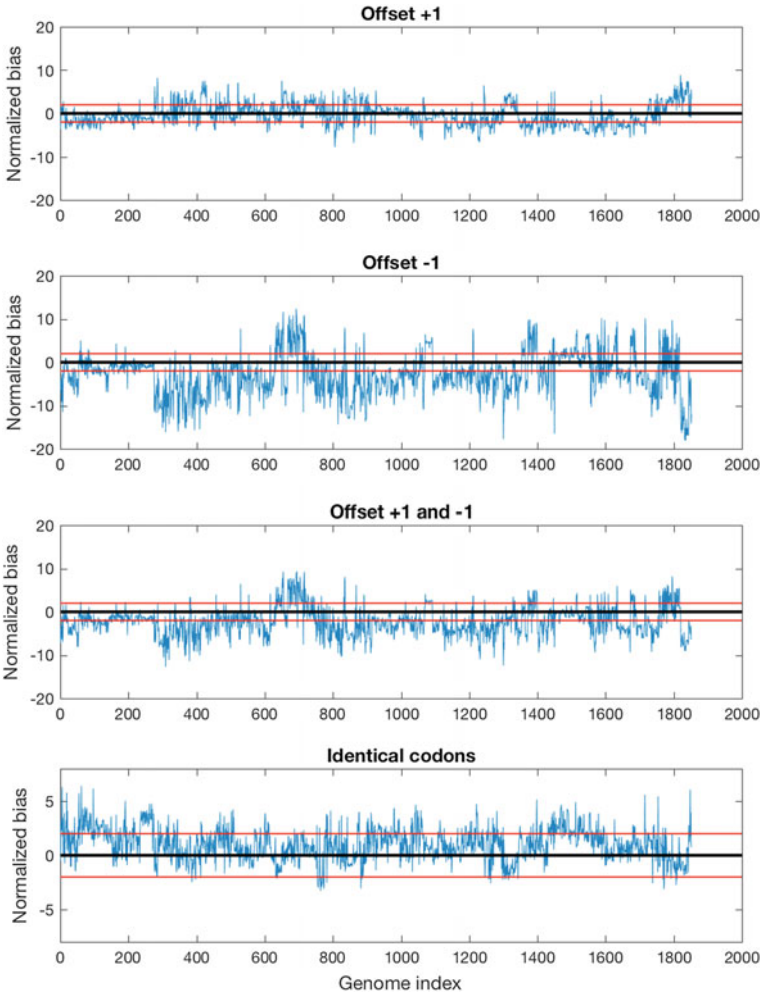

Fig. 13 Values of $\langle DCB_\alpha \rangle_{\text{norm}}$ for the 4 classes mentioned in the text. The abscissa refers to indexes of bacterial genomes in databases and *red horizontal lines* are given by $\langle DCB_\alpha \rangle_{\text{norm}} = \pm 2$; the *continuous blue lines* serve as guides to the eye. As explained in the text values of $\langle DCB_\alpha \rangle_{\text{norm}}$ above or below those *red lines* are statistically significant

We may sum up our results in the following way:

1. We have introduced an unbiased (with respect to genomic AT content) statistical indicator in which the deviations in the probability of having a stop codon out of frame are calculated with respect to the probability based on the dicodon frequencies at fixed codon bias and fixed diamino acids frequencies;
2. From this estimator we evidence a slight positive bias (at the limit of being statistically significant) for the presence of dicodons formed by identical codons

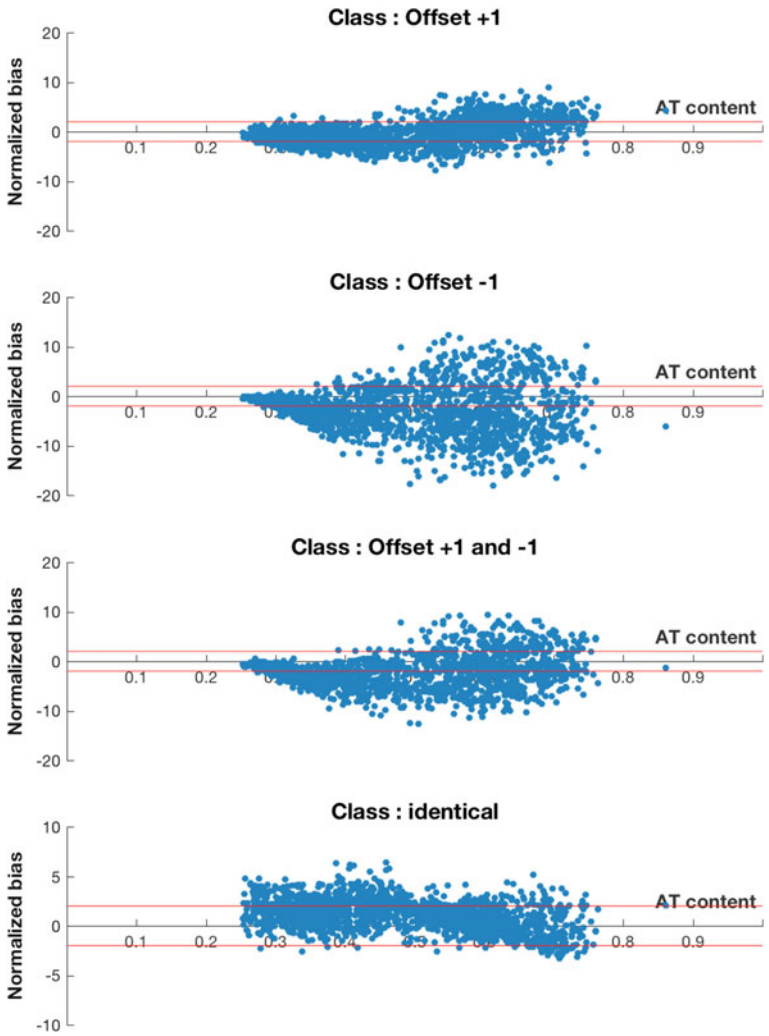

Fig. 14 Values of $\langle DCB_\alpha \rangle_{\text{norm}}$ for the 4 classes mentioned in the text versus AT content of genomes. Each point represents one bacterial genome. Again *red horizontal lines* are given by $\langle DCB_\alpha \rangle_{\text{norm}} = \pm 2$

for the coding of 2 successive identical amino acids. As the presence of correlations favoring identical successive codons was expected from literature (Shao et al. 2012), see Sect. 5, this finding shows that our approach is able to detect relevant statistical signals;

3. We also evidence an overall negative bias for the presence of dicodons containing an OSC (estimator $\langle DCB_{\pm 1} \rangle$ associated to Class ± 1). This result strongly suggest that the Ambush Hypothesis does not hold, at least for the bacterial genomes studied here;

4. This overall trend can be attributed mainly to Class -1 dicodons which present an overall negative bias, whereas Class $+1$ dicodons present an overall null bias.

4.3 Distribution of Distances to Off-Frame Stop Codons in the Random Codon Model

We analyze here whether the Ambush Hypothesis is actually necessary to prevent translation of long abnormal protein chains resulting from frameshift. In this regard, we compute the distance to the first encountered off-frame STOP after a frameshift to $+1$ or -1, starting from definiteness from the start AUG codon in the random codon model (RCM). In practice, we compute the codon usage from the genome of a given species, and draw random codons from this distribution, omitting any correlation between codons. This model therefore generates sequences of random codons. We then estimate the probabilities $Q(\ell)$ that this sequence, in frames $+1$ and -1, produces a STOP codon. To compute the distributions $Q_{+1}(\ell)$ and $Q_{-1}(\ell)$, we have to sum over sequences with ℓ off-frame codons ending up in one of the 3 possible STOPS. The summation over the exponential–in–ℓ number of compatible sequences can be easily carried out with the transfer-matrix formalism shown in Sect. 2.5. We do not report details here; note however that, as STOP codons are defined from 3 nucleotides only, the effective interaction between codons is short-range: only nearest neighbor codons interact along the sequence.

We show in Fig. 15 the outcome of this calculation for one specific bacterial species, *Thermodesulfobium-narugense*. Apart from differences at small ℓ reflecting the influence of the start codon (after the frameshift), both distribution apparently decay exponentially with ℓ. Actually the decay is not a pure exponential, as the transfer matrix is of dimension 4×4, and the number of exponentials is generically given by the size of the transfer matrix, minus one. We obtain that the average distance to the first OSC is about 8–9 in both frames. Hence, even *without any optimization* over the correlations between successive codons along the CDS, OSCs are very quickly found after a frameshift. This result raises doubts about the necessity of selecting codons to make the distance even smaller, as postulated by the Ambush Hypothesis.

5 Discussion and Perspectives

5.1 Nucleotide Motif Usage and Selective Pressures

Viruses have a rapid evolutionary rate, relatively small genomes, and, in many cases, databases of both genomic and phenotypic data that one can use to test

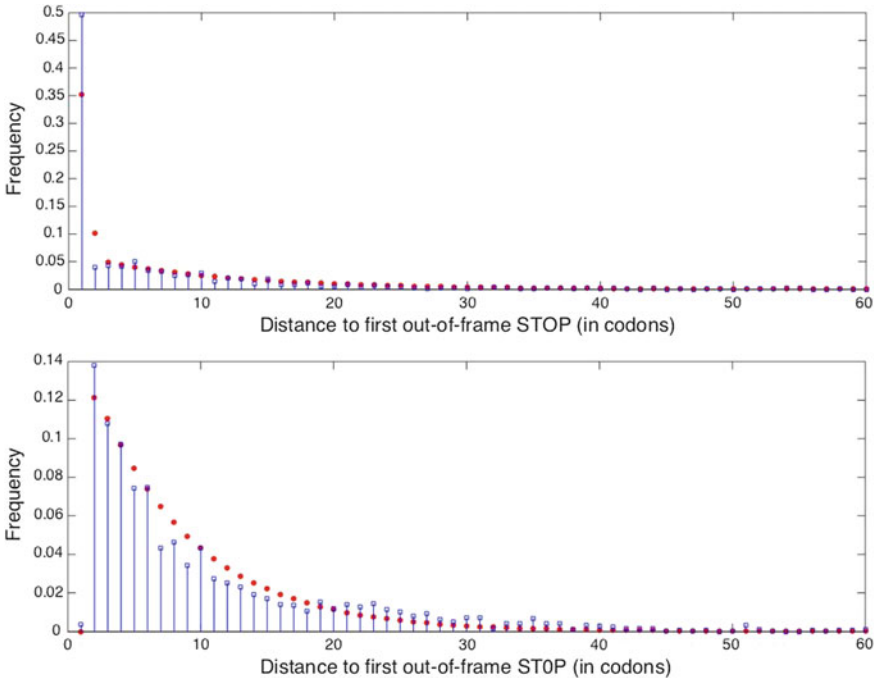

Fig. 15 Distributions of distances to first out-of-frame STOP codon after the start AUG codon and a frameshift equal to $+1$ (top panel) and -1 (bottom panel), measured in codons. *Blue* impulses and *squares* show the experimental distributions computed from all CDS of *Thermodesulfobium-narugense-DSM-14796*. *Red full circles* show the predictions from the random codon model (RCM), obtained with the transfer-matrix formalism, with codon usage estimated from the CDS of the same species (in frame). The average distances are: $\ell_{+1} \simeq 7.9$ and $\ell_{-1} \simeq 9.0$ codons

theoretical approaches. In this work we introduce a mathematical framework, inspired by an analogy with statistical physics, for a class of problems related to the evolution of viruses. The notions of entropy and pressure (or force) evoke the classical concepts of mutation-selection balance in population genetics. A major advantage of our approach is that these notions can be made quantitatively precise, with a very limited computational effort (scaling linearly with the sequence length). This approach is quite versatile, and could be extended to other evolutionary problems. Note that, while we have concentrated here on short nucleotidic motifs, our formalism can be extended to deal with longer motifs. If the motif contains from 2 to 4 nucleotides the transfer matrix M is given by Eqs. (25, 26). There are 63×63 possible matrices, which can be calculated once for all prior to the calculation of $Z(x)$ for several values of x. If the motif contains from 5 to 7 nucleotides the matrix M is $M(C_1, C_2, C_3)$ is "tridimensional", and there are 63^3 possible matrices. The vectors v_i are now functions of two codons. The calculation is slightly more complicated but can be done anyway.

While we have shown applications mainly to Influenza and HIV, many other viruses could be studied. An example is provided by Dengue virus, which goes back and forth between humans and insects. The time scales involved its evolution and the possible presence of mixed pressures acting on different motifs would be worth being studied.

A potentially interesting issue is whether the presence of pressures limits the accessibility of sequences through random mutations in the sequence space. In the absence of pressure codons are independent in our model, and may rapidly evolve under single nucleotide mutations. Hence, any possible sequence can be easily reached from another sequence. When a pressure acting on one motif is considered neighboring codons along the chain start to interact, as the motif may cover two or more contiguous codons, depending on its length. The resulting model is therefore a particular case of the short-range one-dimensional Potts model (Wu 1982), which is known in statistical physics to quickly thermalize. Therefore, as in the independent codon case, the sequence space is sampled efficiently by local moves (such as point mutations). We have checked this statement by running Monte Carlo simulations, and have verified that the relaxation times to the average values of various quantities, such as similarity between sequences and number of motifs, are independent of the value of the pressure. It is however possible that multiple pressures may lead to more complex sequence space structures, less efficiently sampled by local moves. Further studies of this point would be interesting to characterize how much pressures dynamically constrain the evolution of the virus sequence.

Another important application of our formalism is the case of non-coding sequences. In a related work (Tanne et al. 2015) we have extended our approach to non coding RNA, overexpressed in cancer cells compared to healthy tissues. Our analysis has allowed us to show that those overexpressed sequences, such as GSAT and HSATII, correspond to abnormal values of the forces acting on CpG and UpA motifs, and are likely to trigger a large auto-immune response. This prediction was confirmed experimentally, both in human and murine cells (Tanne et al. 2015).

5.2 *Ambush Hypothesis*

In the present work, we have analyzed Coding DNA Sequence (CDS) regions in all bacterial genomes to better investigate the validity of the so-called Ambush Hypothesis. We have introduced a statistical indicator in which the deviations in the probability of having a stop codon one or two nucleotides (1nt or 2nt shift) out of frame are calculated with respect to the probability based on the dicodon frequencies at fixed codon bias and fixed di-amino acids frequencies. With this unbiased indicator we found no systematic deviation across bacterial genomes favoring out of frame stop codons. On the contrary some significant statistical deviations are found for 2nt shifts, in which the probability of out frame stop is smaller than what expected in random sequences.

Our study has focused on four specific classes of di-codons. We will first discuss our result concerning Class *identical*, consisting of pairs of identical codons. Though the effect may seem weak, there is little doubt that there is a slight positive bias $\langle DCB_{\mathrm{id}} \rangle_{\mathrm{norm}}$ which means that translation of a pair of successive identical amino acids slightly favors the use of identical successive codons. This observation can be related to previously reported importance of synonymous codon ordering in yeast (Cannarozzi et al. 2010) and in bacteria (Shao et al. 2012); furthermore a recent study of archaeal aminoacyl-tRNA synthetases (aaRS) has shown that there was evidence for interactions between aaRS and the ribosome thus allowing to recycle tRNAs (Godinic-Mikulcic et al. 2014). Altogether these observations support a mechanism in which, due to colocalization of some aminoacyl-tRNA synthetases and ribosomes, in case of translation of 2 identical successive codons the ribosome, once the first codon translated, may use the same aaRS to translate the next codon.

Concerning our results for the 3 other classes (Class $+1$, Class -1, Class ± 1) one may first observe that the net result for Class ± 1 is at odds with previous results which may have seemed to support the Ambush Hypothesis, though this support was already questioned (Morgens et al. 2013). Indeed the overall negative values of $\langle DCB_{\pm 1} \rangle_{\mathrm{norm}}$ show that presence of dicodons containing an OSC is rather disadvantaged; furthermore comparison of $\langle DCB_{+1} \rangle_{\mathrm{norm}}$ and $\langle DCB_{-1} \rangle_{\mathrm{norm}}$ shows that these overall negative values can be mainly attributed to Class -1 dicodons, Class $+1$ dicodons exhibiting no specific trend in term of signed bias.

One may get further insight into our results examining Figs. 16 and 17. Figure 16 clearly shows the overall negative trend for $\langle DCB_{-1} \rangle_{\mathrm{norm}}$ and also shows that there is no obvious grouping of the genomes as characterized by their values of $\langle DCB_{+1} \rangle_{\mathrm{norm}}$ and $\langle DCB_{-1} \rangle_{\mathrm{norm}}$. Such an observation prompts to examine our results taking into account the phylogeny of our database which has been performed

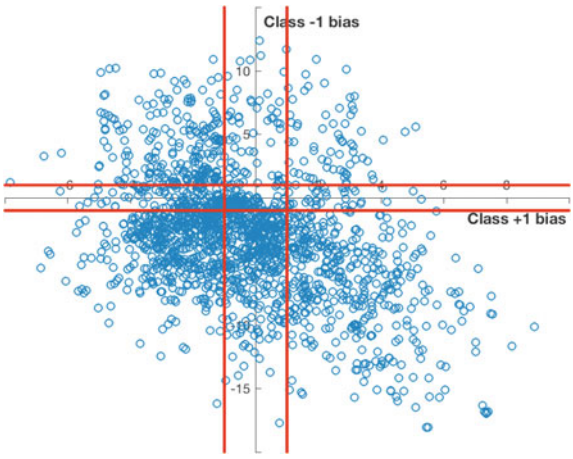

Fig. 16 Two-dimensional plot of values of $\langle DCB_{+1} \rangle_{\mathrm{norm}}$ and $\langle DCB_{-1} \rangle_{\mathrm{norm}}$ for the 1852 studied genomes. Again *red horizontal* and *vertical lines* are given by $\langle DCB_{\alpha} \rangle_{\mathrm{norm}} = \pm 2$ and define regions of statistical significance as explained in the text

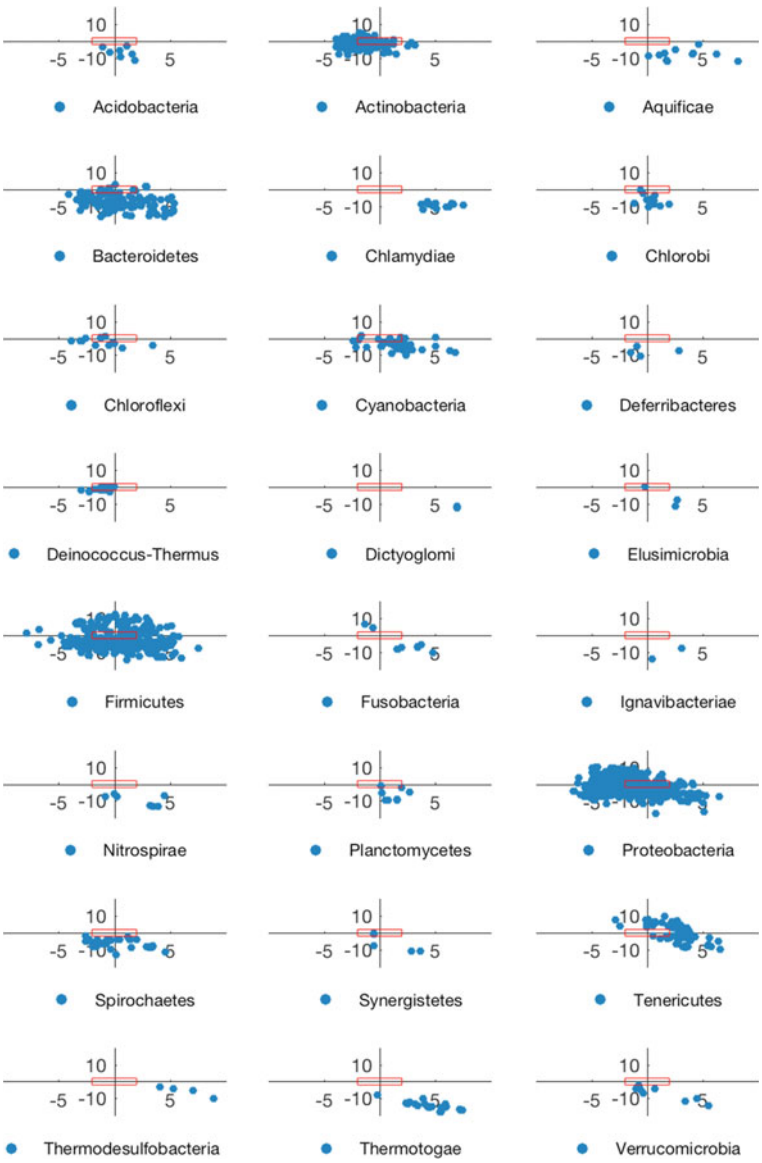

Fig. 17 Two-dimensional plot of values of $\langle DCB_{+1}\rangle_{\text{norm}}$ and $\langle DCB_{-1}\rangle_{\text{norm}}$ for the 1852 studied genomes grouped by phylum. *Red boxes* define regions of statistical significance as explained in the text

in Fig. 17. Indeed, Fig. 17 clearly shows that most phyla exhibit a negative value of $\langle DCB_{-1}\rangle_{\text{norm}}$ with the notable exception of the phyla *Actinobacteria*, *Firmicutes*, *Proteobacteria* and *Tenericutes*.

Quite obviously our results deserve further future analysis. Indeed, at this stage we can reject the Ambush Hypothesis as a general rule for prokaryotic genomes; nevertheless, refining the analysis as shown in Fig. 17, one reaches the conclusion that the situation is somehow more complex and specific phyla would deserve more detailed analysis (see the data in Fig. 17 concerning *Firmicutes* which show that within the same phylus one observes classes of opposite signs for $\langle DCB_{-1} \rangle_{\mathrm{norm}}$). Furthermore, at the present level of analysis, we did not take into account the status of each OSC (TAA, TAG and TGA) which would also deserve more detailed analysis as previously suggested (Morgens et al. 2013); indeed such analysis is probably needed if, as in the case of the observed positive values of $\langle DCB_{\mathrm{id}} \rangle_{\mathrm{norm}}$, one wishes to give a meaningful interpretation in terms of biological processes to the measured values of the various $\langle DCB_{\alpha} \rangle_{\mathrm{norm}}$.

Acknowledgements We are grateful to A. Levine for many enlightening discussions. This work was partly funded by the ANR Coevstat project (ANR-13-BS04-0012-01).

References

Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y (2010) A role for codon order in translation dynamics. Cell 141:355–367

Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S (2008) Virus attenuation by genome-scale changes in codon pair bias. Science 320(5884):1784–1787

Godinic-Mikulcic V, Jaric J, Greber BJ, Franke V, Hodnik V, Anderluh G, Ban N, Weygand-Durasevic I (2014) Archaeal aminoacyl-trna synthetases interact with the ribosome to recycle trnas. Nucleic Acids Res 42(8):5191

Greenbaum BD, Cocco S, Levine AJ, Monasson R (2014) Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. Proc Natl Acad Sci 111(13): 5054–5059

Greenbaum BD, Levine AJ, Bhanot G, Rabadan R (2008) Patterns of evolution and host gene mimicry in influenza and other rna viruses. PLoS Pathog 4(6):e1000079

Hemmi H, Takeuchi O, Kawai T, Kaisho T, Sato S, Sanjo H, Matsumoto M, Hoshino K, Wagner H, Takeda K et al (2000) A toll-like receptor recognizes bacterial dna. Nature 408(6813):740–745

Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106(4):620

Jimenez-Baranda S, Greenbaum B, Manches O, Handler J, Rabadán R, Levine A, Bhardwaj N (2011) Oligonucleotide motifs that disappear during the evolution of influenza in humans increase ifn-α secretion by plasmacytoid dendritic cells. J Virol

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2(2):150–174

Long M, De Souza SJ, Rosenberg C, Gilbert W (1998) Proc Natl Acad Sci USA 95(1):219–223

Medzhitov R, Janeway C Jr (2000) Innate immunity. N Engl J Med 343(5):338–344

Morgens DW, Chang CH, Cavalcanti ARO (2013) Ambushing the ambush hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes. BMC Genomics 14(1): 1–8

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5):418–426

Onsager L (1944) Crystal statistics I: two dimensional model with an order disorder transition. Phys Rev 65:117
Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12(1):32–42
Seligmann H (2010) The ambush hypothesis at the whole-organism level: off frame, 'hidden' stops in vertebrate mitochondrial genes increase developmental stability. Comput Biol Chem 34(2): 80–85
Seligmann H, Pollock DD (2004) The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol 23(10):701–705
Shao Z-Q, Zhang Y-M, Feng X-Y, Wang B, Chen J-Q (2012) Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency. PLoS One 7(3): e33547
Sharp PM, Li W-H (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15(3):1281–1295
Tanne A, Muniz LR, Puzio-Kuter A, Leonova KI, Gudkov AV, Ting DT, Monasson R, Cocco S, Levine AJ, Bhardwaj N et al (2015) Distinguishing the immunostimulatory properties of noncoding rnas expressed in cancer cells. Proc Natl Acad Sci 112(49):15154–15159
Tse H, Cai JJ, Tsoi HW, Lam EP, Yuen KY (2010) Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. BMC Genomics 11(1): 491
Vabret N, Bhardwaj N, Greenbaum BD (2016) Sequence-specific sensing of nucleic acids. Trends Immunol 38(1):53–65
Wong TY, Fernandes S, Sankhon N, Leong PP, Kuo J, Liu JK (2008) Role of premature stop codons in bacterial evolution. J Bacteriol 190 (20):6718–6725
Wu FY (1982) The potts model. Rev Mod Phys 54(1):235–268