The codon information index: a quantitative measure of the information provided by the codon bias

J. Stat. Mech. (2013) P04031

(http://iopscience.iop.org/1742-5468/2013/04/P04031)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

The codon information index: a quantitative measure of the information provided by the codon bias

Luca Caniparoli[1], **Matteo Marsili**[2] **and Michele Vendruscolo**[3]

[1] International School for Advanced Studies (SISSA), via Bonomea 265, I-34136 Trieste, Italy
[2] CMSP Section, The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, I-34014 Trieste, Italy
[3] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
E-mail: luca.caniparoli@sissa.it, marsili@ictp.it and mv245@cam.ac.uk

Received 11 January 2013
Accepted 27 March 2013
Published 26 April 2013

Online at stacks.iop.org/JSTAT/2013/P04031
doi:10.1088/1742-5468/2013/04/P04031

Abstract. The genetic code is redundant, as there are about three times more codons than amino acids. Because of this redundancy, a given amino acid can be specified by different codons, which are therefore considered synonymous. Despite being synonymous, however, such codons are used with different frequencies, a phenomenon known as codon bias. The origin and roles of the codon bias have not yet been fully clarified, although it is clear that it can affect the efficiency, accuracy and regulation of the translation process. In order to provide a tool to address these issues, we introduce here the codon information index (C_{II}), which represents a measure of the amount of information stored in mRNA sequences through the codon bias. The calculation of the C_{II} requires solely the knowledge of the mRNA sequences, without any other additional information. We found that the C_{II} is highly correlated with the tRNA adaptation index (tAI), even if the latter requires the knowledge of the tRNA pool of an organism. We anticipate that the C_{II} will represent a useful tool to study quantitatively the relationship between the information provided by the codon bias and various aspects of the translation process, thus identifying those aspects that are most influenced by it.

Keywords: sequence analysis (theory), systems biology, bioinformatics, statistical inference

Contents

1. Introduction

Molecular biology is undergoing profound changes as major advances in experimental techniques are offering unprecedented amounts of data about the molecular components of living systems [1]–[4]. Most notably, the analysis of the thousands of genomes that have been sequenced [4] is revealing the mechanisms and principles through which the genetic information is maintained and utilized in living organisms [5]–[8].

A question of central importance concerns the causes and consequences of the redundancy of the genetic code. During translation, each amino acid is specified by a triplet of nucleotides (a codon). A given amino acid, however, may correspond to more than one codon, so that 61 codons correspond to 20 amino acids. For instance, lysine is encoded by two codons, valine by four and arginine by six. The way in which these synonymous codons are used shows a marked bias, a phenomenon known as codon bias [9, 10]. For instance, in humans, for the amino acid alanine the codon GCC (guanine–cytosine–cytosine) is used four times more frequently than the codon GCG. The codon bias is characteristic of a given organism and has been associated with three major aspects of mRNA translation, which are efficiency, accuracy and regulation [10, 11].

The first aspect is efficiency. The codon bias and the tRNA abundance in a given organism appear to have co-evolved for optimum efficiency [12]–[14]. Since synonymous codons can be recognized by different tRNAs and translated with different efficiencies, the codon bias is related to the translation rates [15]–[18]. Moreover, codon usage has been shown to correlate with expression levels [19]–[26], so that the use of particular codons can increase the expression of a gene by up to two or three orders of magnitude [27, 22].

The second aspect is accuracy. The codon bias can be used to control the accuracy in the translation, an effect that appears to have been optimized to reduce misfolding and aggregation [28].

The third aspect is regulation. Different codon choices can produce mRNA transcripts with different secondary structure and stability thus affecting mRNA regulation and translation initiation [22, 29]. The codon usage has also been associated with the folding behaviour of the nascent proteins, by timing the co-translational folding process [30].

Since all these aspects of the translation process rely in some form on the information provided by the codon bias, in this work we address the question of establishing a measure of the amount of information encoded in the codon bias itself. For this purpose, by using a combination of statistical mechanics and information theoretical techniques, we introduce the codon information index (C_{II}).

An immediate question is about the need of a new measure associated with the codon bias, when several measures for it have already been proposed, including the 'effective number of codons' ($\hat{N}_{\mathrm{c}}$) [31], the 'frequency of optimal codons' (F_{op}) [12], the 'codon bias index' (CBI) [20], the 'codon adaptation index' (CAI) [32], and the 'tRNA adaptation index' (tAI) [33]. Among these measures, the C_{II} is specifically designed to describe the amount of information encoded in mRNA sequences through the codon bias, and it is the only one with the following properties: (i) it requires only information about the mRNA sequences and does not depend on any additional data, i.e. the C_{II} is self-contained, and (ii) it produces a codon-wise profile for each sequence which is sensitive to the spatial organization of the codon. As examples, we mention in particular that, for a given organism, the tAI requires the knowledge of the tRNA pool, and that the CAI requires the knowledge of the most expressed genes.

We first analyse the general properties of the C_{II} and then we apply it to a pool of 3371 genes of yeast. We find that the C_{II} correlates with protein and mRNA abundances, as well as with the tRNA adaptation index (tAI) [33]. The latter result shows that two independent forms of information, which are stored in different parts of the genome, the tRNA copy number and the codon bias in the coding region, are remarkably dependent on one another. These results also show that the spatial organization (i.e. the order) of the codons inside the transcript is a relevant part of the information stored in the gene.

2. Construction of the C_{II}

A natural representation of the information contained in the codon bias can be given in terms of strings of bits (i.e. $(0, 1)$ variables) or of distributions over bit strings. In other words, we associate a bit to each codon. This procedure corresponds to binning the codons into two classes, each with its own codon usage distribution, given by the frequencies of the codons in that class. The first ingredient to build the C_{II} is an assignment of bits to codons that is maximally informative, in a way to be specified later. This also implies that the information encoded in this way is optimally retrievable. The second ingredient is a local codon organization along the sequence (see figure 1). We analyse these two contributions separately below.

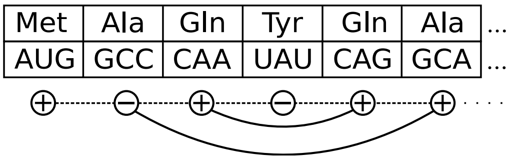

Figure 1. In the calculation of the $C_{\rm II}$ each codon has an associated binary variable. Synonymous codons interact via the information theoretic part of the Hamiltonian (solid lines), nearest neighbour sites interact via the Ising interaction (dashed line).

2.1. Maximum information

We first consider the special case of a protein of length L composed of only one amino acid type, which can be translated by K different codons [34]; the generalization to the full set of amino acids is described below. The sequence $\{c_1 \ldots, c_L\}$, $c_i = 1, \ldots, K$ is given, where c_i is the ith codon. We associate a binary variable $s_i = \pm 1$ (i.e. a *spin* variable, rather than a *bit* ($\{0, 1\}$) variable) to each site of the sequence; this spin variable identifies the class each codon is assigned to.

Let $p_{c|s}$ be the probability for the codon c to be used on a site with spin s. *A priori*, the only information available is that an amino acid can be encoded by any of its possible codons. This state of ignorance is described by the choice of a uniform prior distribution for the probabilities of the parameters $p_{c|s}$, with the normalization ensured by a δ function

$$P_0(\hat{p}) = \prod_{s=\pm1} \Gamma(K)\delta\left(\sum_{c=1}^{K} p_{c|s} - 1\right). \tag{1}$$

For any assignment $\vec{s} = (s_1, \ldots, s_L)$ of the spins, the statistical information contained in the sequences is encoded in the codon counts $n_s(c) = \sum_{i=1}^{L} \delta_K(s_i - s)\delta_K(c_i - c)$, i.e. the number of times codon c is used on a site with spin s. The probability of observing $\vec{n}_s = (n_s(1) \ldots n_s(K))$ is modelled using a product of two multinomial distributions

$$P(\vec{n}_s|\hat{p}) = \prod_{s=\pm1} \frac{\Gamma\left(N_s + 1\right)}{\prod_{\rm c} \Gamma(n_s(c) + 1)} \prod_{\rm c} p_{c|s}^{n_s(c)}, \tag{2}$$

where $N_s = \sum_1^K n_s(c)$ and obviously $N_+ + N_- = L$.

Using the Bayes formula $P(\theta|x) \propto P(x|\theta)P_0(\theta)$, we obtain the posterior distribution

$$P(\hat{p}|\vec{n}_s) = \prod_{s=\pm1} \frac{\Gamma\left(N_s + K\right)}{\prod_{\rm c} \Gamma(n_s(c) + 1)} \prod_{c=1}^{K} p_s(c)^{n_s(c)} \delta\left(\sum_{\rm c} p_s(c) - 1\right). \tag{3}$$

An important quantity that can be derived from the prior and the posterior is how much information is gained by observing the codon frequencies. This quantity is the symmetrized Kullback–Leibler divergence between the two distributions

$$\begin{aligned} I(\vec{n}_s) &= D_{\rm KL}(P\|P_0) + D_{\rm KL}(P_0\|P) \\ &= \left\langle \log \frac{P(\hat{p}|\vec{n}_s)}{P_0(\hat{p})} \right\rangle_{P(\hat{p}|\vec{n}_s)} + \left\langle \log \frac{P_0(\hat{p})}{P(\hat{p}|\vec{n}_s)} \right\rangle_{P_0(\hat{p})}, \end{aligned}$$

where the averages are performed with respect to the posterior, equation (3), and the prior, equation (1), respectively. The integration can be performed analytically and leads to

$$I(\vec{n}_s) = -\sum_{s=\pm 1}\sum_{c=1}^{K} n_s(c)\left[\psi\left(N_s + K\right) - \psi\left(n_s(c)+1\right)\right] + \mathrm{Const}, \tag{4}$$

where $\psi(x)$ is the digamma function.

The generalization to the whole set of amino acids is simply the sum of the $I(\vec{s}, \vec{n}_s)$ for each amino acid

$$I(\{\vec{n}_{s,a}\}) = -\sum_{a=1}^{20}\sum_{s=\pm 1}\sum_{c_a=1}^{K_a} n_{s,a}(c_a)[\psi(N_{s,a} + K_a) - \psi(n_{s,a}(c_a)+1)] + \mathrm{Const}, \tag{5}$$

where K_a is the number of codons encoding for the amino acid a, $n_{s,a}(c_a)$ is the number of times a spin s is associated to the codon c_a of the amino acid a and $N_{s,a} = \sum_{c_a=1}^{K_a} n_{s,a}(c_a)$.

To extract as much information as possible from the codon counts we have to maximize equation (5). However, the contributions of different amino acids are independent and each amino acid sector is invariant under a spin flip. Therefore, the minima of equation (5) are highly degenerate. Moreover, the amino acids with only one codon, methionine and tryptophan, do not contribute to equation (5).

These issues can be cured observing that the previous derivation does not use any information about how the codons are arranged along the sequence. Therefore, information carried by the codon order can be used to weight the minima of equation (5).

2.2. Codon spatial organization

In order to remove the degeneracy, we add to equation (5) an interaction between nearest neighbour spins that favours their alignment. This coupling is also consistent with the observation of the existence of a ‘codon pair bias’ [35, 36]. We thus define the cost function

$$H\{\vec{s}\} = -J\sum_{i=1}^{L-1} s_i s_{i+1} - I(\{\vec{n}_{s,a}\}), \tag{6}$$

where J is a parameter to be tuned which accounts for the degree of spatial homogeneity of the sequence. In statistical physics terms, equation (6) can be regarded as the Hamiltonian of a spin system that, besides the 1D Ising interaction J, also has a long range interaction I as shown in figure 1.[4] This analogy makes it possible to apply techniques used to study spin systems in statistical physics to the present model.

We are interested in the spin arrangements minimizing the Hamiltonian (6), given the codon sequence. Numerically, energy minimization was performed by simulated annealing Monte Carlo [37]. When the states of minimal H were found to be degenerate, an average over all of them was considered.

The optimization of the cost function H is carried out simultaneously on a pool of genes of the same organism, but clearly the nearest neighbour interaction is defined only for neighbouring codons within the same gene.

[4] The Hamiltonian (6) is invariant under a global spin flip, thus the average magnetization would be zero. To break this symmetry it is sufficient to add a term $H_h = h\sum_i s_i$, which favours the configuration aligned with the external field h. The field h will be taken vanishingly small ideally, very small in practice.

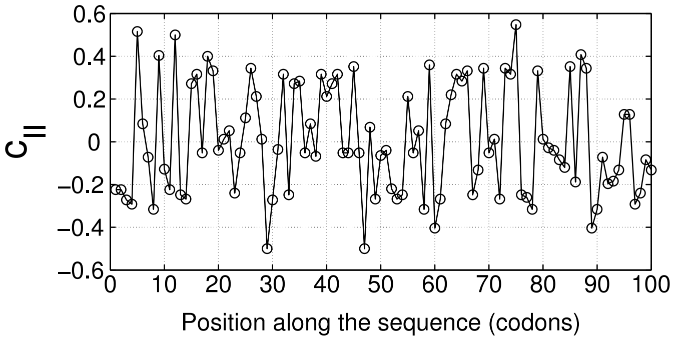

Figure 2. Local c_{II}, as in equation (7), for the first 100 codons of the TFC3/YAL001C gene.

We define the local codon information index as the magnetization at site i on the transcript g, i.e. the thermodynamic average of the spin at site i

$$c_{\mathrm{II}}^{(g)}(i) = \langle s_i^{(g)} \rangle \tag{7}$$

(an example is given in figure 2), and the codon information index of the gene g as the average of the local c_{II} on the gene codons

$$C_{\mathrm{II}}^{(g)} = \frac{1}{L_g} \sum_{i=1}^{L_g} c_{\mathrm{II}}^{(g)}(i). \tag{8}$$

3. Phase diagram of the Hamiltonian (6)

In this section we characterize the properties of H in equation (5) and its minima. We will also prove the existence of a phase transition in temperature at $J = 0$. Finally, we will describe the effect of the nearest neighbour interaction.

3.1. Maximum of equation (5)

As a first step in the characterization of (5), we want to show that each term of the form (4) has a minimum and is convex. Let us rewrite (4) setting $n_\pm(c) = n(c)/2 \pm \delta_{\mathrm{c}}$, $\delta_{\mathrm{c}} \in [-n(c)/2; n(c)/2]$ and $\Delta = \sum_{\mathrm{c}} \delta_{\mathrm{c}}$

$$\begin{aligned} I(\vec{\delta}) &= -\sum_{s=\pm1}\left[\left(\frac{N}{2}+s\Delta\right)\psi\left(\frac{N}{2}+s\Delta+K\right) - \sum_{c=1}^{K}\left(\frac{n_{\mathrm{c}}}{2}+s\delta_{\mathrm{c}}\right)\psi\left(\frac{n_{\mathrm{c}}}{2}+s\delta_{\mathrm{c}}+1\right)\right] \\ &= \sum_{s=\pm1}\left[-g_K\left(\frac{N}{2}+s\Delta\right) + \sum_{c=1}^{K} g_1\left(\frac{n_{\mathrm{c}}}{2}+s\delta_{\mathrm{c}}\right)\right] \\ &= -\ G_K\left(\frac{N}{2},\Delta\right) + \sum_{c=1}^{K} G_1\left(\frac{n_{\mathrm{c}}}{2},\delta_{\mathrm{c}}\right), \end{aligned} \tag{9}$$

where $g_i(x) = x\ \psi(x+i)$ and $G_i(n,x) = g_i(n+x) + g_i(n-x)$.

We observe that I is symmetric with respect to a transformation $\vec{\delta} \to -\vec{\delta}$. Since $g_i(x)$ is continuous and differentiable in the domain, the derivative must be zero in $\vec{\delta} = \vec{0}$. This point corresponds to a uniformly distributed posterior and, since I is computed as the KL divergence of the posterior from a uniform prior (which is a non-negative quantity and is zero iff the two distributions are equal), it is an absolute minimum. It is the only critical point since the system of equations

$$\frac{\partial I}{\partial \delta_i} = -G'_K\left(\frac{N}{2}, \Delta\right) + G'_1\left(\frac{n_i}{2}, \delta_i\right) = 0 \qquad i = 1 \dots K \tag{10}$$

has the $\delta_i = 0$ solution only, observing that $\partial_{\delta_i} G'_1\left(n_i/2, \Delta\right) > \partial_{\delta_i} G'_K\left(N/2, \Delta\right)$.

This implies that the maxima must reside on the boundary. Repeating the argument on the boundary faces, we end up concluding that the maxima must lie on the boundary vertices, i.e. the points such that $\delta_i^* = \pm n_i/2$. On these points I becomes

$$I(\vec{\delta}^*) = \sum_{c=1}^{K} n_c \psi\left(n_c + 1\right) - \left(\frac{N}{2} + \Delta\right) \psi\left(\frac{N}{2} + \Delta + K\right) - \left(\frac{N}{2} - \Delta\right) \psi\left(\frac{N}{2} - \Delta + K\right).$$

The function is now only dependent on Δ and observing again that it is symmetric and concave we see that there is a maximum in $\Delta = 0$.

The maximum is thus obtained on the vertices which minimize the difference $|N_+ - N_-|$ (e.g., for four codons with $\{n(c)\} = (5, 3, 4, 2)$ the maximum is obtained for $(+, -, -, +)$ or $(-, +, +, -)$, since I is symmetric under a global spin flip). This is an instance of the number partition problem which belongs to the NP-complete class. However, we are dealing with sets which contain at most six elements.

Considering the full set of amino acids, we can finally ask how many states have the same $I(\{\vec{\delta}_a\})$. Using the fact that the contribution for each amino acid is invariant under a spin flip and that the amino acids with one codon only are not considered, there are at least 2^{18} states in addition to the trivial degeneracy coming from the amino acids methionine and tryptophan which do not contribute to (5).

3.2. Phase transition in temperature at $J = 0$

It is possible to analytically work out the thermodynamics of equation (6) at $J = 0$. At high temperatures we expect a disordered paramagnetic phase, while at low temperatures the system falls into one of its many minima, which correspond to the maxima of equation (5) described in the previous section. Here we prove that a phase transition exists by showing that the concavity of the free energy changes sign at a critical temperature T_c in the large n_c limit.

At $J = 0$ we can easily compute the entropy of a state specified by $\vec{n}_+$. The number of different configurations is simply the number of permutations of n_c elements, given that the $n_+(c)$ and $n_-(c)$ are equivalent. The entropy is thus the logarithm of the product of binomials

$$S(\vec{n}_+|\vec{n}) = \log \prod_c \binom{n_c}{n_+(c)} \tag{11}$$

and we can easily write the free energy $F = H - TS$,

$$F = \left[G_K(N/2, \Delta) - \sum_{c=1}^{K} G_1(n_\mathrm{c}/2, \delta_\mathrm{c}) \right] - T \sum_{c=1}^{K} \log \left(\frac{n_\mathrm{c}}{\frac{n_\mathrm{c}}{2} + \delta_\mathrm{c}} \right). \tag{12}$$

At high temperature the thermodynamics of the system is governed by the entropic term which has a minimum at $\vec{\delta}_\mathrm{c} = 0$ (paramagnetic phase).

To prove that this minimum becomes repulsive at a critical temperature we study the concavity of the free energy. Taking the second derivatives gives

$$\frac{\partial^2 F}{\partial \delta_i^2} = G''_K(N/2, \Delta) - G''_1(n_i/2, \delta_i) + T(\psi_1(n_i/2 + \delta_i + 1) + \psi_1(n_i/2 + \delta_i + 1)),$$
$$\frac{\partial^2 F}{\partial \delta_i \partial \delta_j} = G''_K(N/2, \Delta).$$

At high temperature we expect that $|\delta_i| \ll n_i$. Expanding in large $\vec{n}_\mathrm{c}$, we find

$$\frac{\partial^2 F}{\partial \delta_i \partial \delta_j} \sim \frac{4}{N} + \delta_{ij}^{KR} \left[-\frac{4}{n_i} + T \left(\frac{4}{n_i} + \frac{4}{n_i^2} \right) \right] + O(n^{-3})$$

which can be written as $\partial^2 F = b + a_i \delta_{ij}^{KR} + O(n^{-3})$, where both $a_i = -4(1 + T + T/n_i)/n_i$ and $b = 4/N$ are independent of δ_i up to terms of order n_c^{-3}.

The free energy F is convex (concave) if the Hessian matrix is positive (negative) definite, i.e. if each eigenvalue is positive (negative). Its characteristic polynomial can be easily computed using the Sylvester's determinant theorem and reads as

$$P_K(\lambda) = \prod_{i=1}^{K} (a_i - \lambda) + b \sum_{i=1}^{K} \prod_{j \neq i}^{K} (a_j - \lambda).$$

Using some combinatorics, we obtain

$$P_K(\lambda) = (-\lambda)^K + \sum_{i=1}^{K} (-\lambda)^{K-i} \left[\sum_{j_1 < \cdots < j_i} a_{j_1} \ldots a_{j_i} + b(K - i + 1) \sum_{j_1 < \cdots < j_{i-1}} a_{j_1} \ldots a_{j_{i-1}} \right] \tag{13}$$

with the convention $\alpha_0 = 1$. If $T > T_\mathrm{c}^{+} = \max_\mathrm{c}(1 - n_\mathrm{c}^{-1})$ we immediately see that each coefficient is positive and thus the Hessian is positive definite, while if $T < T_\mathrm{c}^{(-)} = \min_\mathrm{c}(1 - n_\mathrm{c}^{-1})$ the Hessian is negatively defined because of the Descartes rule of signs[5]. At $T < T_\mathrm{c}^{(-)}$ the free energy becomes convex and the ground state moves discontinuously far from the paramagnetic ($\vec{\delta} = \vec{0}$) state.

The order parameter which captures this phase transition is the codon coherence

$$\phi_{J=0}(T) \equiv \sum_{c \neq M,W} \left\langle \left(\frac{2\delta_\mathrm{c}}{n_\mathrm{c}} \right)^2 \right\rangle_T, \tag{14}$$

where the sum is intended on every codon except methionine and tryptophan and the thermodynamic average is performed at temperature T. This parameter is small in the

[5] The Descartes rule of signs states that the number of positive roots of a polynomial (known to have all real roots, like in this case, since we are computing the eigenvalues of a symmetric matrix) is equal to the number of sign differences between consecutive non-zero coefficients.

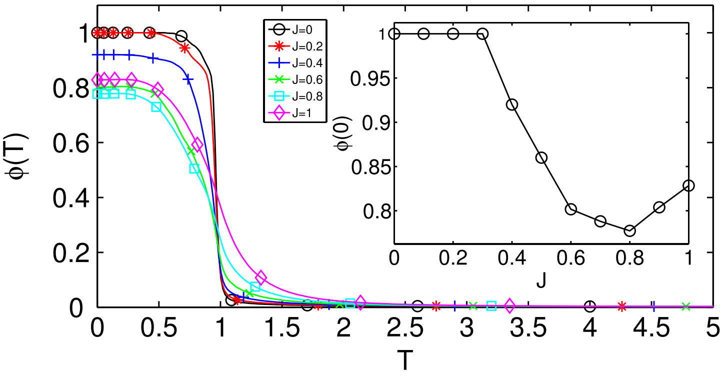

Figure 3. Codon coherence $\phi_J(T)$ as a function of the temperature for increasing J and codon coherence at $T = 0$ (inset). As the nearest neighbour interaction is turned on, the phase transition becomes smoother. At $T = 0$, the codon coherence is preserved up to a critical value of $J \sim 0.3$.

paramagnetic phase (which has $2\delta_c = n_+(c) - n_-(c) = o(n_c)$) while it is one in the partitioning phase. Its plot is given in figure 3: the phase transition at $T = 1$ is evident, albeit smoothed out due to finite size effects.

3.3. Effects of the nearest neighbour interaction: $J > 0$

The introduction of the nearest neighbour interaction makes the analytical treatment much more difficult. Nevertheless, we expect that the phase transition will become smoother and smoother as J is raised, since the Ising model in 1D does not exhibit any phase transition. This observation is numerically tested in figure 3, where the profiles of $\phi_J(T)$ are plotted for increasing J.

The $T > 1$ behaviour is easily understandable by observing that in the paramagnetic phase ($\delta_c \ll n_c$) the information theoretical part of the Hamiltonian is flat around $\vec{\delta} = 0$ in the large n_c limit. We expect the high temperature ($T > 1$) behaviour to be dominated by the magnetic field and the nearest neighbour interaction terms: excluding the information theoretical part, the Hamiltonian reduces to the Ising model one, $H_{\text{Ising}} = -J\sum_{i=1}^{L-1}\sigma_i\sigma_{i+1} - h\sum_{i=1}^{L}\sigma_i$. Thus, the thermodynamics at $T > 1$ should be described by the phenomenology of the Ising model.

To check this hypothesis, we numerically computed the magnetization for the Hamiltonian (6)

$$m = \frac{1}{\sum_c n_c}\left\langle \sum_c n_+(c) - n_-(c) \right\rangle \tag{15}$$

as well as, analytically, the magnetization for the Ising model,

$$m_{\text{Ising}} = \frac{1}{L}\left\langle \sum_{i=1}^{L} \sigma_i \right\rangle = \frac{\sinh(h/T)}{\sqrt{\mathrm{e}^{-4J/T} + (\sinh h/T)^2}}. \tag{16}$$

These quantities are plotted in figure 4, where is clearly shown that the Ising model correctly describes the $T > 1$ behaviour of (6).

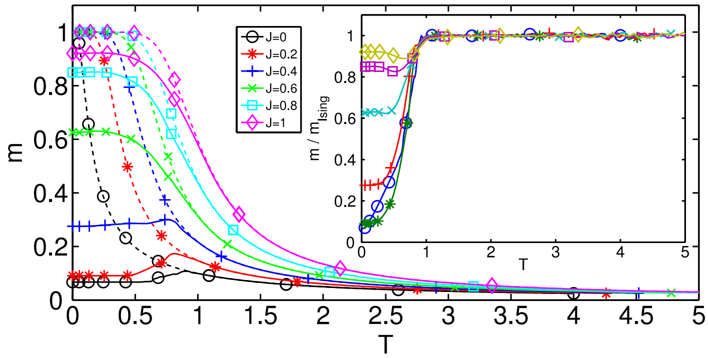

Figure 4. Solid lines: numerically computed magnetization as a function of temperature (equation (15)). Dashed lines: analytically computed magnetization of the Ising model in 1D (equation (16)). The inset shows the ratio m/m_{Ising}. The magnetization for $T > 1$ is correctly described by the Ising model, and as J is raised the behaviour at $T < 1$ becomes more and more similar to that of the Ising model.

We introduced the nearest neighbour interaction to weight the many minima of the information theoretical part of the Hamiltonian and to extract information from the spatial arrangement of the codons. Observing the inset of figure 3, we see that at $J > 0.3$ the codon coherence at $T = 0$ is lost. This means that the same codon is assigned a different c_{II} on different positions. Since there is no *a priori* biological reason for this differentiation, we restrict the admissible J to those such that $\phi_J(0) = 1$. Moreover, since we want to maximize the information extracted from the spatial arrangement of the codons, we fix J as the maximum value such that $\phi_J(0) = 1$. Interestingly, we find that for this value of J the correlation of $C_{\mathrm{II}}^{(g)}$ with the tAI exhibits a maximum.

4. Analysis of the C_{II}

4.1. C_{II} correlates with protein and mRNA abundance

We computed the C_{II} for a set of 3371 transcripts of *S. cerevisiae* and compared it with the logarithms of the measured protein [38] and mRNA [39] abundances, finding a significant correlation ($C \simeq 0.60$ and $C \simeq 0.69$ for proteins and mRNAs, respectively, figure 5).

Indeed, computing the same quantities for the half of the set comprising the most abundant proteins and mRNAs we observe a sharp increase in the correlation coefficients $C \simeq 0.70$ and $C \simeq 0.79$ for proteins and mRNAs, respectively.

4.2. C_{II} correlates with tAI

A common method to address the translational efficiency of a gene is given by the tRNA adaptation index (tAI) [33], defined as

$$\mathrm{tAI} = \left(\prod_{i=1}^{L_g} w_{c_i} \right)^{1/L_g}, \qquad (17)$$

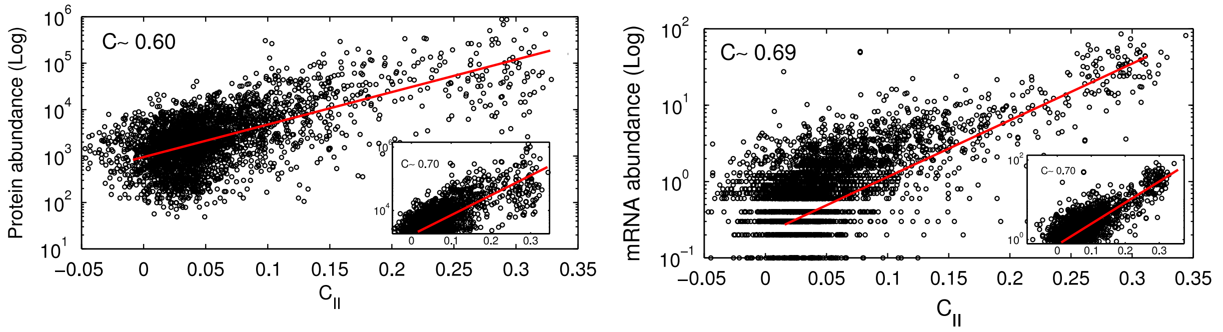

Figure 5. The C_{II} correlates with the logarithms of protein abundance (left) and mRNA abundance (right). The correlation is most evident for the most abundant half of the set (insets).

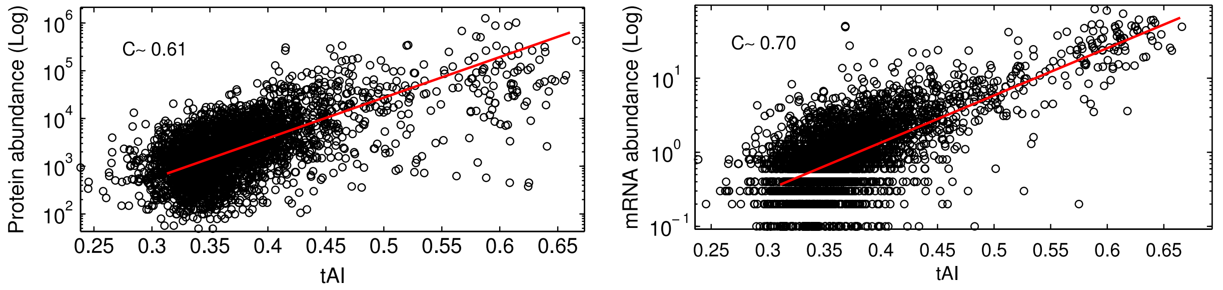

Figure 6. The tAI is highly correlated with protein (left) and mRNA (right) abundance.

where w_{c_i} is the weight associated to the c_ith codon in the gene g. These w_c are defined as

$$w_i = \begin{cases} \dfrac{W_i}{W_{\max}} & \text{if } W_i \neq 0 \\ w_{\text{avg}} & \text{otherwise} \end{cases}, \qquad W_i = \sum_{j=1}^{n_i} (1 - s_{ij})\mathrm{tGCN}_{ij},$$

where n_i is the number of tRNA isoacceptors recognizing codon i, tGCN_{ij} is the number of the jth tRNA recognizing codon i and s_{ij} gauges the efficiency of codon–anticodon coupling. The s_{ij} are then optimized to maximize the correlation with protein abundance. The computation of this index requires information about two of the most influential factors affecting translation efficiency, namely tRNA abundance (tRNA copy number is highly correlated to tRNA abundance [40]) and codon coupling efficiency.

Computing the tAI for the same 3371 transcripts and comparing it with the C_{II} we observe an extremely high correlation ($\rho \sim 0.93$, see figure 7). We thus are able to reproduce all the results obtained from the tAI without needing any additional information beyond the codon sequences and without any parameter optimization: the C_{II} depends only upon the parameter J which can be fixed from thermodynamics considerations, as explained in section 3.3.

The tAI is known to correlate well with protein abundance ($\rho \simeq 0.61$) and mRNA abundance ($\rho \simeq 0.70$), see figure 6. Moreover, also in this case the correlation improves for the most abundant proteins ($\rho \simeq 0.70$) and mRNAs ($\rho \simeq 0.77$) but to a significantly smaller extent with respect to the C_{II} case.

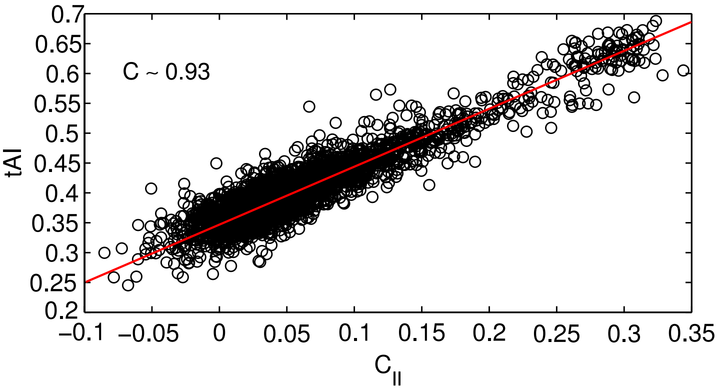

Figure 7. The $C_{\rm II}$ is highly correlated with the tAI.

Table 1. Pearson correlation coefficients between numerical and experimental quantities. The values between parentheses, when present, refer to correlations computed for the most abundant half of the set.

	$C_{\rm II}$	tAI	Protein levels (log)	Half-life (log)	mRNA levels (log)
$C_{\rm II}$	1	0.93	0.60 (0.70)	0.18	0.69 (0.79)
tAI		1	0.61 (0.68)	0.20	0.70 (0.77)
Abundance (log)			1	0.30	0.60
Half-life (log)				1	0.22

Another quantity usually taken into consideration in this kind of study is protein half-life. We computed correlations among all these quantities; the results are summarized in table 1. Between parentheses are the correlations computed for the 1600 most abundant proteins and mRNAs. All these values are statistically significant (P-value $\sim 10^{-9}$ at most). Those involving $C_{\rm II}$, tAI, protein and mRNA abundance are highly significant (P-value $< 10^{-20}$).

It has been suggested that the correlation between the $C_{\rm II}$ and the mRNA levels can be caused by evolutionary forces acting more effectively on highly expressed genes [22], as beneficial codon substitutions are more likely to be fixed on these genes because the gain in fitness is likely to be higher, although a fully causal relationship can be more complicated and involve other determinants [11, 10].

4.3. Average $C_{\rm II}$ profile along the proteins

In the previous sections we analysed the properties of the global $C_{\rm II}$ value for whole genes, but the $c_{\rm II}(i)$ gives another local layer of information. We thus ask whether the local $c_{\rm II}(i)$ can be interpreted as a local measure of translational optimality. Unfortunately too few data exist to confirm or falsify this hypothesis, but we can explore whether a common behaviour at the beginning of the transcripts exists. Similarly to the case of the tAI [29],

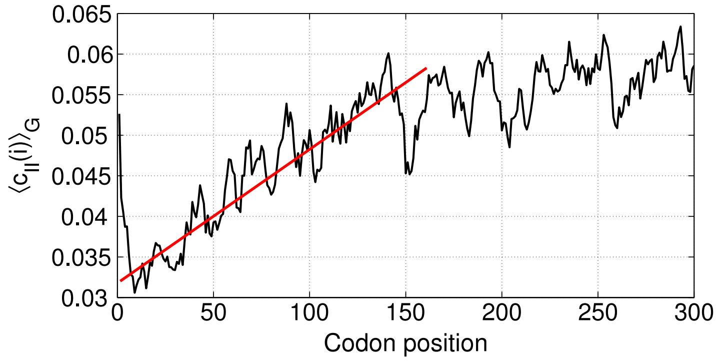

Figure 8. Local c_{II} averaged along the proteins, as in equation (18). The straight line is a guide for the eye.

it is possible to compute the average of the $c_{\mathrm{II}}^{(g)}(i)$ across the transcripts,

$$\langle c_{\mathrm{II}}(i)\rangle_G = \frac{1}{N_G}\sum_{g=1}^{N_G} c_{\mathrm{II}}^{(g)}(i). \tag{18}$$

The plot in figure 8 reveals the presence of a 'ramp' roughly 120 codons long followed by a plateau.

This result is consistent with the findings in [29] for the tAI, where this procedure reveals a signal at the beginning of the transcript: the average local tAI has a minimum at the beginning of the sequence and rises up to the average value in ~100 codons. Since the authors claim that the tAI carries information about codon translational efficiency, they hypothesize that this feature helps translation stabilization by avoiding ribosome jamming.

5. Conclusions

In this work we have introduced the codon information index (C_{II}) as a measure of the amount of information stored in mRNA sequences through the codon bias. We have shown that C_{II} can capture at least as much complexity as previously introduced codon bias indices, but its computation does not require additional data beyond transcript sequences. In order to calculate the C_{II} we do not make any assumption on the origins and roles of the codon bias, but quantify the amount of information associated with it in an unbiased manner, with a procedure that enables us to fully quantify the amount of such information.

We calculated the C_{II} for a set of over 3000 yeast transcripts and found values highly correlated with the tAI scores, as well as with experimentally derived protein and mRNA abundances. Furthermore, we were able to reproduce the result that the first 70–100 codons are, on average, translated with low efficiency, a feature which is thought to help translational stabilization [29].

We anticipate that by using the C_{II} it will be possible to investigate open questions about the role of the codon bias in optimizing translational efficiency, improving codon reading accuracy, and minimizing the risk of misfolding and aggregation.

Data sets

The set of RNA transcripts for *S. Cerevisiae* was downloaded from [41]. Protein half-lives were extracted from [42]. Protein and mRNA abundances were found in [39, 43]. Protein abundances were found in [44, 38].

Source code and details of the algorithm

The thermodynamic averages (and thus the C_{II}) were computed using a Monte Carlo algorithm implemented with simulated annealing [37] and frequent reannealings: the temperature is a function of the simulation time and is slowly lowered. Provided that the cooling schedule is sufficiently slow, this method is guaranteed to sample the whole space, a vital feature if the free energy landscape is rough (meaning that the Hamiltonian has many metastable states).

The algorithm run time for the set of 3371 proteins was of the order of 1 day on a dual core workstation. The algorithm is massively parallelizable. A major speedup was obtained by observing that the energy differences of the information theoretical part of the Hamiltonian (6) (which are required in the Monte Carlo step) can be efficiently and locally computed using the properties of the digamma function. The code is available at the web page www-vendruscolo.ch.cam.ac.uk/CII/index.php.

Acknowledgments

Work partially supported by the GDRE 224 GREFI MEFI, CNRS-INdAM.

References

[1] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H and Kanehisa M, *Kegg: Kyoto encyclopedia of genes and genomes*, 1999 *Nucl. Acids Res.* **27** 9

[2] Kozomara A and Griffiths-Jones S, *miRBase: integrating microRNA annotation and deep-sequencing data*, 2011 *Nucl. Acids Res.* **39** (Suppl. 1) D152

[3] Benson D A, Karsch-Mizrachi I, Lipman D J, Ostell J and Sayers E W, *Genbank*, 2011 *Nucl. Acids Res.* **39** (Suppl. 1) D32

[4] Pagani I, Liolios K, Jansson J, Chen I-M A, Smirnova T, Nosrat B, Markowitz V M and Kyrpides N C, *The genomes online database (gold) v.4: status of genomic and metagenomic projects and their associated metadata*, 2012 *Nucl. Acids Res.* **40** D571

[5] The Arabidopsis Genome Initiative, *Analysis of the genome sequence of the flowering plant arabidopsis thaliana*, 2000 *Nature* **408** 796

[6] McCarthy M I, Abecasis G R, Cardon L R, Goldstein D B, Little J, Ioannidis J P A and Hirschhorn J N, *Genome-wide association studies for complex traits: consensus, uncertainty and challenges*, 2008 *Nature Rev. Genet.* **9** 356

[7] Gibson G, *Rare and common variants: twenty arguments*, 2012 *Nature Rev. Genet.* **13** 135

[8] Brown T A, 2006 *Genomes* (New York: Garland Publishing)

[9] Ikemura T, *Codon usage and tRNA content in unicellular and multicellular organisms*, 1985 *Mol. Biol. Evol.* **2** 13

[10] Plotkin J B and Kudla G, *Synonymous but not the same: the causes and consequences of codon bias*, 2011 *Nature Rev. Gen.* **12** 32

[11] Gingold H and Pilpel Y, *Determinants of translation efficiency and accuracy*, 2011 *Mol. Syst. Biol.* **7** 481

[12] Ikemura T, *Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system*, 1981 *J. Mol. Biol.* **151** 389

[13] Bulmer M, *Coevolution of codon usage and transfer RNA abundance*, 1987 *Nature* **325** 728

[14] Dong H, Nilsson L and Kurland C G, *Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates*, 1996 *J. Mol. Biol.* **260** 649

[15] Varenne S, Buc J, Lloubes R and Lazdunski C, *Translation is a non-uniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains*, 1984 *J. Mol. Biol.* **180** 549

[16] Sørensen M A, Kurland C G and Pedersen S, *Codon usage determines translation rate in Escherichia coli*, 1989 *J. Mol. Biol.* **207** 365

[17] Cannarozzi G, Cannarrozzi G, Schraudolph N N, Faty M, von Rohr P, Friberg M T, Roth A C, Gonnet P, Gonnet G and Barral Y, *A role for codon order in translation dynamics*, 2010 *Cell* **141** 355

[18] Tuller T, Waldman Y Y, Kupiec M and Ruppin E, *Translation efficiency is determined by both codon bias and folding energy*, 2010 *Proc. Nat. Acad. Sci. USA* **107** 3645

[19] Grantham R, Gautier C, Gouy M, Jacobzone M and Mercier R, *Codon catalog usage is a genome strategy modulated for gene expressivity*, 1981 *Nucl. Acids Res.* **9** 213

[20] Bennetzen J L and Hall B D, *Codon selection in yeast*, 1982 *J. Biol. Chem.* **257** 3026

[21] Gouy M and Gautier C, *Codon usage in bacteria: correlation with gene expressivity*, 1982 *Nucl. Acids Res.* **10** 7055

[22] Kudla G, Murray A W, Tollervey D and Plotkin J B, *Coding-sequence determinants of gene expression in Escherichia coli*, 2009 *Science* **324** 255

[23] Zhang G, Hubalewska M and Ignatova Z, *Transient ribosomal attenuation coordinates protein synthesis and co-translational folding*, 2009 *Nature Struct. Mol. Biol.* **16** 274

[24] Komar A a, *A pause for thought along the co-translational folding pathway*, 2009 *Trends Biochem. Sci.* **34** 16

[25] Deane C M and Saunders R, *The imprint of codons on protein structure*, 2011 *Biotechnol. J.* **6** 641

[26] Tuller T, Kupiec M and Ruppin E, *Determinants of protein abundance and translation efficiency in S. cerevisiae*, 2007 *PLoS Comput. Biol.* **3** e248

[27] Gustafsson C, Govindarajan S and Minshull J, *Codon bias and heterologous protein expression*, 2004 *Trends Biotechnol.* **22** 346

[28] Drummond D A and Wilke C O, *Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution*, 2008 *Cell* **134** 341

[29] Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I and Pilpel Y, *An evolutionarily conserved mechanism for controlling the efficiency of protein translation*, 2010 *Cell* **141** 344

[30] Clarke T and Clark P, *Increased incidence of rare codon clusters at* 5′ *and* 3′ *gene termini: implications for function*, 2010 *BMC Genomics* **11** 118

[31] Wright F, *The 'effective number of codons' used in a gene*, 1990 *Gene* **87** 23

[32] Sharp P M and Li W-H, *The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications*, 1987 *Nucl. Acids Res.* **15** 1281

[33] dos Reis M, Savva R and Wernisch L, *Solving the riddle of codon usage preferences: a test for translational selection*, 2004 *Nucl. Acids Res.* **32** 5036

[34] Bailly-Bechet M, Danchin A, Iqbal M, Marsili M and Vergassola M, *Codon usage domains over bacterial chromosomes*, 2006 *PLoS Comput. Biol.* **2** e37

[35] Gutman G A and Hatfield G W, *Nonrandom utilization of codon pairs in Escherichia coli*, 1989 *Proc. Nat. Acad. Sci.* **86** 3699

[36] Coleman J R, Papamichail D, Skiena S, Futcher B, Wimmer E and Mueller S, *Virus attenuation by genome-scale changes in codon pair bias*, 2008 *Science* **320** 1784

[37] Kirkpatrick S, Gelatt C D and Vecchi M P, *Optimization by simulated annealing*, 1983 *Science* **220** 671

[38] Newman J R S, Ghaemmaghami S, Ihmels J, Breslow D K, Noble M, DeRisi J L and Weissman J S, *Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise*, 2006 *Nature* **441** 840

[39] Holstege F C P, Jennings E G, Wyrick J J, Lee T I, Hengartner C J, Green M R, Golub T R, Lander E S and Young R A, *Dissecting the regulatory circuitry of a eukaryotic genome*, 1998 *Cell* **95** 717

[40] Percudani R, Pavesi a and Ottonello S, *Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae*, 1997 *J. Mol. Biol.* **268** 322

[41] Cherry J M *et al*, *Saccharomyces genome database: the genomics resource of budding yeast*, 2012 *Nucl. Acids Res.* **40** D700

[42] Belle A, Tanay A, Bitincka L, Shamir R and O'Shea E K, *Quantification of protein half-lives in the budding yeast proteome*, 2006 *Proc. Nat. Acad. Sci. USA* **103** 13004

[43] Dudley A M, Aach J, Steffen M A and Church G M, *Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range*, 2002 *Proc. Nat. Acad. Sci.* **99** 7554

[44] Ghaemmaghami S, Huh W-K, Bower K, Howson R W, Belle A, Dephoure N, O'Shea E K and Weissman J S, *Global analysis of protein expression in yeast*, 2003 *Nature* **425** 737