Hidden patterns of codon usage bias across kingdoms

Yun Deng[1,2], *Jeremie Kalfon*[1], *Dominique Chu*[1], *Tobias von der Haar*[2]

[1]**School of Computing, University of Kent, CT2 7NF, Canterbury, UK**
[2]**School of Biosciences, University of Kent, CT2 7NJ, Canterbury, UK**
{dfc,tv}@kent.ac.uk

Abstract

The genetic code encodes 20 amino acids using 64 nucleotide triplets or codons. 18 of the 20 amino acids are encoded by multiple synonymous codons which are used in organismal genomes in a biased fashion. Codon bias arises because evolutionary selection favours particular nucleotide sequences over others encoding the same amino acid sequence. Despite many existing hypotheses, there is no current consensus on what the evolutionary drivers are. Using ideas from stochastic thermodynamics we derive from first principles a mathematical model describing the statistics of codon usage bias and apply it to extensive genomic data. Our main conclusions include the following findings: (1) Codon usage cannot be explained solely by selection pressures that act on the genome-wide frequency of codons, but also includes pressures that act at the level of individual genes. (2) Codon usage is not only biased in the usage frequency of nucleotide triplets but also in how they are distributed across mRNAs. (3) A new model-based measure of codon usage bias that extends existing measures by taking into account both codon frequency and codon distribution reveals distinct, amino acid specific patterns of selection in distinct branches of the tree of life.

Keywords: stochastic thermodynamics, codon usage bias, fungi, protists, bacteria, Boltzmann distribution

1 Introduction

Codon usage bias (CUB), the preferred use of particular codons over others encoding the same amino acid, is an established phenomenon. The principal forces that shape CUB are thought to be mutation, selection, and random drift (reviewed in [1]). While this view appears generally accepted, there is little consensus about the precise evolutionary drivers of CUB. A number of individual drivers have been proposed either by way of correlation or through experimental evidence, but it is currently unclear whether selection has multiple causes [1, 2] or a single dominant cause, and if so, what the dominant cause may be.

The proposed selective forces can be categorised into two groups. The first group comprises forces that act at the level of codons only, i.e. they are independent of the subsequence in which the codon is located and lead to uniform bias across the entire genome. We will henceforth refer to this as *beanbag selection*. Examples include tRNA-based selection models where codon usage is matched to the supply of tRNAs [3], and GC content based models where codon usage is matched to constraints imposed by some preferred proportion of G and C bases in the DNA sequence [4, 5, 6].

The second group of mechanisms acts at the level of coding sequences and results in *sequence level selection (SLS)*. In the context of CUB, sequence-level effects have gained increasing appreciation in recent years through the demonstration that codon usage can exert control over both protein levels [7, 8] and protein quality [9, 10, 11], and that this impacts biological function. For example, specific codon usage patterns enable the functioning of biological mechanisms as diverse as the mammalian cell cycle [12], the mechanism by which sub-physiological temperatures engender neuroprotection [13] and fungal circadian clocks [14]. It has also been observed that codon usage may change along a coding sequence [15, 16, 17], may be related to spatial organisation of the genome [18], and some experimental effects of altered codon usage have been observed for which no underlying cause is known [19].

While there is thus evidence for a large number of possible evolutionary drivers for CUB, it remains unclear how the various mechanisms interact and how much they contribute to overall CUB relative to one another. There have been some recent attempts to use statistical methods to disentangle the various influences on CUB [20, 21, 22, 23, 24]. However, this was mostly done in order to quantify how much one specific mechanism contributes relative to others. Quantifying the relative importance of various drivers remains difficult because of the sheer number of mechanisms many of which are perhaps yet to be discovered.

Even if it were feasible to map out all the evolutionary forces in detail, this would not necessarily be particularly insightful. Instead, a macroscopic description of the system may provide more insight. There are many

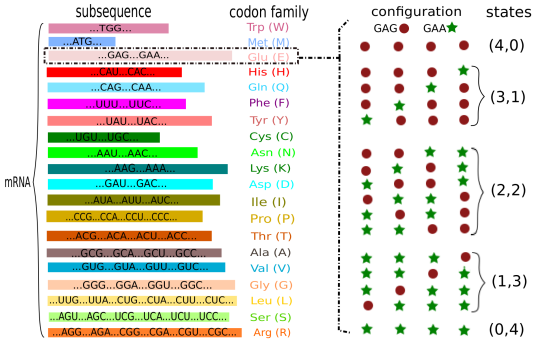


Fig. 1: For each species, each mRNA is split into up to 18 subsequences, each consisting of codons for a particular amino acid only. Each subsequence could exist in many different configurations, i.e. sequences of synonymous codons. For this paper, we only take into account the number of codons of each type. Here, all configurations are indicated by assuming that the length $L^{\mathcal{A},g} = 4$. To calculate energies, we only consider the summaries states of the subsequence counting how many of each codon type are present in an mRNA.

precedents in science, notably in statistical physics, where simple, useful and universal laws emerge from intractable microscopic interactions. Examples, include the ideal gas law that relates 4 macroscopic quantities to one another while ignoring individual positions and momenta of molecules, scaling laws in biology [25], word frequencies in texts [26], spatial structures of genomes [27] and evolution [28, 29] all of which abstract away from microscopic detail in order to arrive at robust macroscopic laws.

In this paper, we will take inspiration from this approach. As the main result, we derive from first principles a novel, parsimonious and general model of codon evolution as a random walk based on ideas of stochastic thermodynamics[30] and information thermodynamics [31]. This model has only two free parameters and does not postulate any specific evolutionary forces, but instead describes the aggregate effect of a putatively large number of simultaneously acting evolutionary forces on the genome. As a second main contribution of this study we fit this model to a comprehensive genomic dataset consisting of 462 fungal genomes from the ENSEMBL database [32]. Such datasets are only now becoming available in sufficient numbers to probe thermodynamic features of CUB in any depth. As a third contribution, we find unambiguous evidence that SLS left a pervasive signature on the distribution of codons over genes in fungal, bacterial and protist genomes. This provides evidence from an evolutionary perspective for recent observations that connect codon usage to translational control in a number of different setting including human development and diseases including cancer [33]. Finally, as a fourth contribution, we propose a new quantitative description of codon usage bias, that, while summarised as a single number, does not only take into account the relative proportion of codons but crucially also how they are distributed across the genome. We argue that this captures more accurately the selection pressure on codon usage than measures that merely quantify the relative abundance of codons, such as CAI [34] or tAI [35]. We demonstrate the usefulness of this measure by investigating amino acid-specific patterns of codon usage bias across the fungal kingdom, revealing for the first time the strong amino acid specificity of codon selection in the different clades of the fungal tre of life.

2 Results

2.1 Deriving the full model

In order to derive our model we conceptualise codon evolution as a discrete space, continuous time random walk in the space of synonymous codons. In this picture, each gene represents up to 18 independent random walkers, one for each amino acid encoded by more than 1 codon. Each random walker is thus a *subsequence* of synonymous codons for some amino acid $\mathcal{A} \in \{\texttt{E,H,Q,F,Y,C,N,K,D,I,P,T,A,V,G,L,S,R}\}$ appearing in a gene g. Each such subsequence has a length $L^{\mathcal{A},g}$ which is the number of amino acids of type $\mathcal{A}$ appearing in the gene. Each subsequence consists of k_1 codons of type 1, k_2 codons of type 2, $\ldots$, $k_{|C^{\mathcal{A}}|}$ codons of type $|C^{\mathcal{A}}|$, where $|C^{\mathcal{A}}| \in \{2, 3, 4, 6\}$ is the total number of codons for amino acid $\mathcal{A}$. For each amino acid we decided arbitrarily which one of its codons is codon 1, codon 2 and so on. This assignment remained fixed for all species we analysed.

We can now consider each possible configuration of a subsequence $\{k_1, \ldots, k_{|C^{\mathcal{A}}|}\}$ as a state. From any such state, the random walker can access all states that are 1 synonymous mutation away. For example, one of the codons in the subsequence may be mutated from codon 1 to codon 2, which would correspond to the transition from $\{k_1, k_2, \ldots, k_{|C^{\mathcal{A}}|}\}$ to $\{k_1 - 1, k_2 + 1, \ldots, k_{|C^{\mathcal{A}}|}\}$. In the case of only two codons, where $|C^{\mathcal{A}}| = 2$ this random walk reduces to a 1-dimensional discrete state random walk in continuous time with $L^{\mathcal{A},g} + 1$ states, corresponding to $L^{\mathcal{A},g}$ codons being of type 1, $L^{\mathcal{A},g} - 1$ codons being of type 1,..., 0 codons being of type 1; see

Symbol	Meaning
$\lvert\mathcal{A}^g\rvert$	Number of occurrences of amino acid $\mathcal{A}$ in gene g
$C^{\mathcal{A}}$	A type of codon of amino acid $\mathcal{A}$
$\lvert C^{\mathcal{A}}\rvert$	number of different codons for amino acid $\mathcal{A}$
$C_i^{\mathcal{A}}$	i-th codon type of amino acid $\mathcal{A}$
$\lvert C^{\mathcal{A}}\rvert \in \{1,2,3,4,6\}$	The number of codons codon for amino acid $\mathcal{A}$
$k_i^{\mathcal{A},g}, k_i^{\mathcal{A}}, k_i$	The number of codons of type i of amino acid $\mathcal{A}$ occurring in gene g.
$L^{\mathcal{A},g} := \sum_i k_i^{\mathcal{A},g}$	The number of occurrences of $\mathcal{A}$ in gene g.

Tab. 1: Explanation of the symbols used.

supplementary information for more detail on the model.

Throughout this contribution, we make a number of simplifying assumptions about the nature of the random walk. Firstly, we assume that non-synonymous mutations are negligible, i.e. the rate of mutation from a codon to a non-synonymous codon is zero. Secondly, we assume that the mutation rates between synonymous codons are *a priori* the same, i.e. the random walk is unbiased. Any deviations from this assumption are due to evolutionary selection pressures (including effects of random drift). Thirdly, the random walker is in a steady state. Continuing evolutionary pressure could therefore change individual subsequences, but will not, on the whole, change the statistics of the codon distribution. Fourthly, throughout this article we are not concerned with the spatial arrangements of codons across a gene or genome, but we only record how many codons of a particular kind are to be found in a particular subsequence.

To derive predictions for the distribution of codons across subsequences in response to specific selective pressures, we devised a theoretical model of the dynamics of codon evolution based on stochastic thermodynamics [36]. We conceptualise each subsequence configuration i as having an energy E_i, where E_i depends on the codon composition of the subsequence and the selection pressure. In steady state the probability of observing a subsequence with energy E_i, i.e. the probability to find the random walker in state i, is then given by the Boltzmann distribution $P(E_i) = \exp(-E_i/T)/\sum_i \exp(-E_i/T)$, where we have assumed that the Boltzmann constant $k_\mathrm{B} = 1$. In this model T is a constant that in a physical system would correspond to the temperature, but we will interpret this here as an abstract temperature that is not in a clear relationship with the ambient temperature experienced by the organism. Having established this conceptual framework, we are now able to determine the energy that is implied by various selection scenarios, which in turn leads to a prediction for steady state Boltzmann distributions of random walkers/sequences, which can be compared to data.

The simplest energy function can be derived for the beanbag model and in the absence of selection forces acting on codon usage. In this case, the assignment of codons would be like throwing $L^{\mathcal{A},g}$ times a die, with $|\mathcal{A}^g|$ sides, and recording how often a particular result was obtained. In this contribution, we will mainly consider the case of amino acids with 2 codons, in which case the unbiased case would be like tossing a fair coin $L^{\mathcal{A},g}$ times and counting how often tail and heads were obtained. In this case we find that that the energy becomes $E_i = -\ln\binom{L}{k_1}$, where k_1 is the frequency of the first codon; see SI for the calculation. The corresponding Boltzmann distribution coincides with the unbiased binomial distribution with $p = 1/2$, as expected. This simplest model can be readily expanded to include a beanbag model with a global codon usage bias q, yielding an energy $\hat{E}_i = E_i + \ln\left((1-q)^i/q^i\right)$. Again, the resulting Boltzmann distribution coincides with the binomial distribution with bias $p = q$.

In the beanbag models the rate of mutation from codon 1 to codon 2 is proportional to k_1, the number of codons of type 1. We now posit instead that this rate is proportional to k_1^ξ where $\xi \in \mathbb{R}$, and the rate from codon 2 to codon 1 becomes proportional to $(L-k_1)^\gamma$ and $\gamma \in \mathbb{R}$. This breaks the assumptions of beanbag selection. The resulting statistics can no longer be reproduced by throwing dice or tossing coins, not even unfair ones. Instead, this model entails SLS. In the supplementary information we show that in this scenario the energy for a subsequence with i codons of type 1 is given by the *full model* (see SI for derivation):

$$\bar{E}_i = \xi E_i + T(\gamma - \xi)\ln(i!). \tag{1}$$

Given this model, a Boltzmann distribution with parameters γ and ξ can be obtained, as above. Biologically, this Boltzmann distribution would then formulate the probability to observe a gene that has exactly $k_1 = i$ codons of type 1 and $L^{\mathcal{A},g} - k_1$ codons of type 2.

Before proceeding, we discuss some special choices for the ad-hoc parameters ξ, γ so as to clarify their biological meaning. When $\xi = \gamma \neq 1$, then the second term on the right hand side disappears and the energy is the same as in the biased beanbag model with a modified inverse temperature ξ. In this case there will be no selection pressure on the global usage of codons, but there will be a SLS affecting how codons are distributed across subsequences. For $\xi = \gamma = 1$ the full model 1 reduces to the binomial distribution with $q = 0.5$ exactly. In the most general case of $\xi \neq \gamma$ selection is affected by the second term, which can be interpreted as a selection "potential." In this case, a global codon usage bias q will emerge as a result of sequence level selection. We now define an inverse temperature $T^{-1} := (\xi + \gamma)/2$ for the model as the simplest function that is symmetric in the

two parameter and reduces to the inverse temperature in the case of $\xi = \gamma$. This temperature is unrelated to the physical temperature of the organism, but has an interpretation in terms of the width of the steady state distribution. The "colder" the distribution, the more the probability mass is concentrated around the maximum of the standard case of $T = 1$. The extreme case is $T = 0$, when the only observed subsequences would correspond to the most probable subsequence configuration. In contrast, as $T \to \infty$ all subsequence configurations become equally likely. We will find here that actual genomes tend to be moderately hot with $1 < T < 2$ for most subsequences, resulting in a flatter distribution than one would expect from a temperature of $T = 1$.

Finally, we note that the full model 1 does not reduce exactly to the binomial distribution for $q \neq 1/2$ for any choice of parameters, but we found that it can approximate it to high degrees of accuracy. Given the relatively high statistical error of determining codon distributions it will therefore not be possible to reject SLS empirically even if the underlying data was binomial (ie if the beanbag model, rather than SLS, actually applies).

2.2 Genomic data bear the signature of SLS

In order to understand whether or not there is evidence for SLS or the beanbag-model in actual genomic data, we obtained genome sequence data from 462 fungal species represented in the Fungi division of the ENSEMBL database [32]. We split all annotated protein coding sequences into subsequences for each of the nine amino acids where the codon choice $|C^{\mathcal{A}}| = 2$. For these amino acids beanbag model selection results in binomially distributed data, whereas for larger codon choices beanbag model selection yields multinomially distributed data for which the statistical errors of our analyses would be much larger. We further only analysed subsequences with lengths $5 \leq L^{\mathcal{A},g} \leq 15$, thereby excluding sequences < 5 where there are very few datapoints to fit, and excluding sequences > 15 where statistical errors become overly large because the number of subsequences reduces quickly with increasing subsequence length. We fitted each of the subsequences in this dataset to a binomial distribution as well as to the full model described above. The discussion to follow focuses on the resulting 45702 fits to the fungal data, although we note that we obtained similar results with fits to bacterial and protist data as shown in the supplementary information.

2.2.1 Distribution of codons can be fitted to binomial distributions

The distribution of codons across genes in our dataset can be fitted to a binomial distribution. Doing this for all subsequences results in a distribution of mean-residuals between $\exp(-4)$ and $\exp(-9)$ peaking at about $\exp(-7)$. Visual inspection of a number of examples suggests that these mean-residuals indicate a reasonably good fit to the data. The only fitting parameter in the model is the bias p, which is the *global codon usage bias* q. Since we fitted each subsequence length separately we obtained, for each species and each amino acid, 10 different estimates for q. Pairwise comparison of the estimates of q between length 15 and the other length yielded high correlations, with Pearson coefficient > 0.95. This means that the global codon usage bias varies little as the length of the subsequence is varied; see supplementary plots. Taken on their own, these results seem to point to codons being distributed binomially consistent with the beanbag model.

2.2.2 The full model fits the fungal data better

As a comparison we also fitted the full model eq. 1 to the data thus obtaining estimated values for the parameters ξ and γ of the full model. We also obtained for each fit a mean-residual indicating how well the full model can be fitted to the data.

For all our datasets, the typical values of the parameters γ and ξ are small and positive with 96.39% of the fits resulting in $0 < \gamma, \xi < 2$. The quality of the fits can be quantified by comparing the mean-residuals obtained from fitting the full model with those obtained from fitting the binomial model. This indicates that the former is a better description of the data *on the whole* in the sense that the distribution of mean-residuals is shifted to the left towards smaller values; see fig. 2 for a comparison of the distributions. The median for the residuals of the full model is 0.0002850, and as such about 3 times smaller than the corresponding value for the binomial fits, which is 0.000845.

The better fit of the full model could be merely a reflection of the fact that it has more parameters than the binomial model. We therefore prepared a control set of distributions. This control set consists of the same subsequences that the real dataset contains, but with each codon replaced by a random synonymous codon according to the global codon usage bias q; see supplementary information for a description and for the control dataset. By construction this control set implements the beanbag model exactly, meaning that the sequence composition of subsequences is distributed according to the binomial distribution with a global codon usage bias q corresponding to empirically measured values. Fitting both the full model and the binomial model to these control data results in mean-residuals that are visually indistinguishable from one another reflecting the above cited fact that the full model can approximate binomial data; see fig. 2a.

The quality of the fit of the binomial model to the binomial data of the control-set can be viewed as a benchmark for the best mean-residuals that can be obtained given the statistical error inherent in the dataset. An inspection of the histogram in fig. 2a reveals that the distribution of mean-residuals obtained from fitting the full model to the real data is only minimally shifted to the right of this optimal benchmark. This leads to the

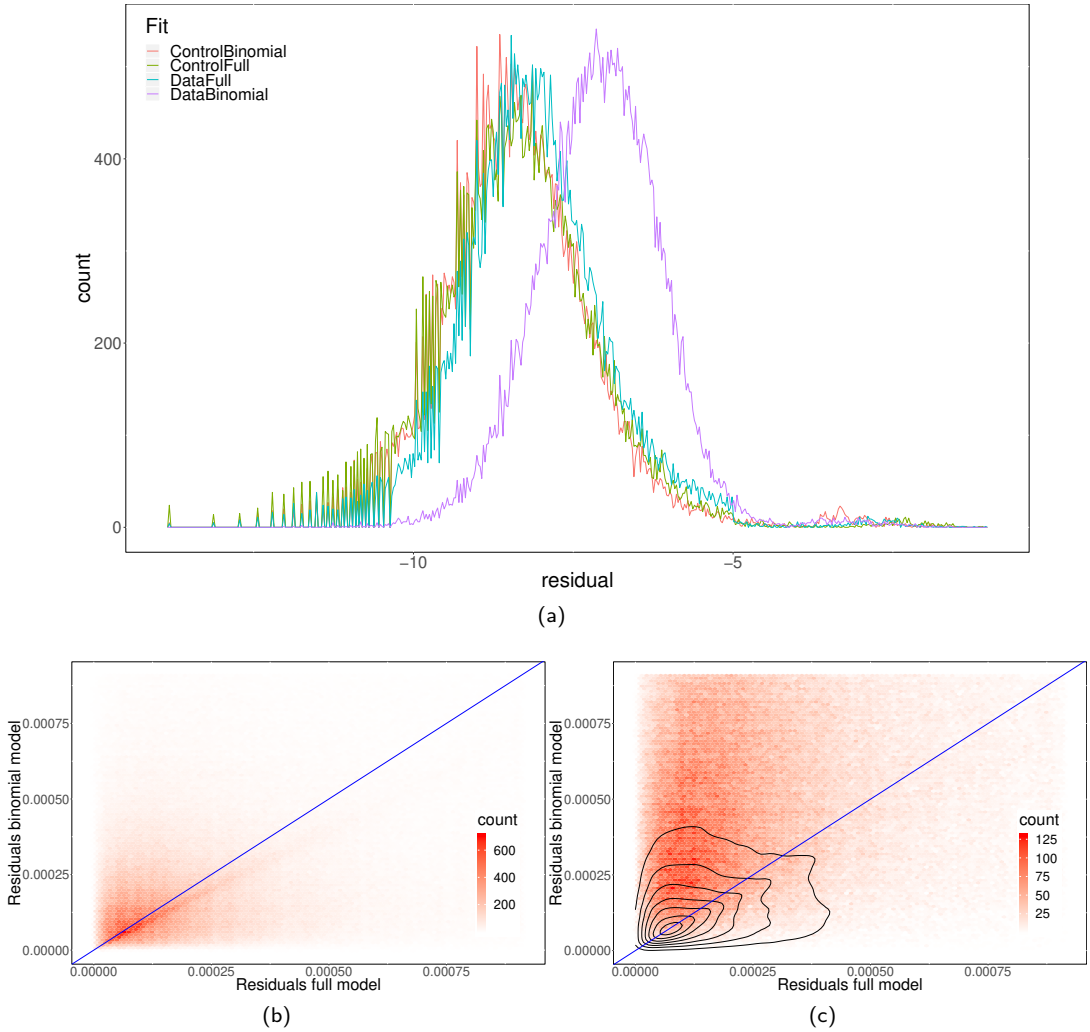


Fig. 2: (a) Histogram for the mean-residuals obtained from fitting the binomial distribution and the full model for both the real data and control data generated under the beanbag model assumption, ie where codons have been replaced by random synonymous codons with a bias corresponding to the global codon usage bias. The x-axis is shown on a logarithmic scale. The distribution of the mean-residuals of the real binomial fitted to real data is clearly shifted to the right of the fit to the full model, suggesting that the latter is a better fit on the whole. On the other hand, the mean-residuals of the full model overlap with the distribution of the mean-residuals resulting from the fit of both the binomial and the full model to the control data. (b) Comparing the mean-residuals from the full model to those of the binomial model. The plot shows the the density of points for the control data. The area above the diagonal indicates subsequences where the full model is a better fit than the binomial model. Points on the diagonal indicate that both models fit the subsequence equally well. (c) Same comparison, but for real data. The contour lines indicate the density of the control data in (b) for comparison.

conclusion that the full model captures almost all of the variation of the underlying real data. The beanbag model (which implies a binomial distribution) is not sufficient to explain how codons are distributed across the genome in fungi. Instead, it is necessary to postulate sequence-level selection in order to account for the distribution of codons over subsequences. In contrast, the full model, as formulated in eq. 1 accounts for the distribution.

2.2.3 Comparing individual subsequences

So far, we have concluded that the full model is a better fit to the distribution of codons on the whole, but we do not know whether this applies to all individual subsequences, or whether there is only a subset of subsequences that is better described by the full model, whereas the rest is equally well described by the binomial model. To decide this, we plotted the mean-residuals arising from fitting the binomial distribution against the mean-residual obtained from fitting the full model to the same subsequence. We did this for both the control dataset described above and for the real data. It is instructive to first consider the former; see fig 2b. This analysis confirms that most subsequences of the control data are approximately equally well fitted by the binomial and the control data, although the density of points appears to be higher below the diagonal indicating that the binomial model fits the control data somewhat better. This is because, as mentioned above, the full model can only approximate the binomial distribution.

Turning now to the real data the same analysis leads to a high density of points in the upper left corner of the figure. This is the area where the mean-residuals of the full model are low, but those of the binomial model are high; see fig. 2c. The corresponding subsequences are consistent with sequence-level selection, but not with the beanbag model. In contrast, the subsequences along the diagonal, can be equally well explained by the beanbag model and SLS. Note that from fig. 2c it is only possible to make statistical inferences, comparing densities of a large number of points. For any individual subsequence, it is not possible to exclude conclusively the beanbag model.

2.2.4 Defining distance

A different perspective on the difference between the models can be obtained from the distribution of subsequences in ξ, γ space; see fig. 3b. It reveals that the fits remain concentrated into a smaller part of parameter space than the fits to the real data. This is further evidence that SLS has shaped codon usage in fungi. Statistically, it means that the global codon usage bias q is not the only manifestation of selection pressures on codons, but SLS has also altered how codons are distributed across subsequences.

Established measures, such as the CAI [34] or tAI [35] that are based on the global codon usage biases miss this aspect and are therefore incomplete quantifications of the real codon usage bias. Based on the full model we now propose a new measure that quantifies in a more complete way the selection pressure on codon usage, irrespective of whether it arises from beanbag selection or SLS. As such a measure we propose the Euclidean distance of the subsequence from the no-selection case in ξ, γ space. This no-selection case corresponds to $\xi = \gamma = 1$ exactly and any deviation from that indicates a selection pressure.

$$\mathcal{D} := \sqrt{(1-\xi)^2 + (1-\gamma)^2} \qquad \text{(Selection pressure)}$$

In the case of no global codon usage bias, the selection pressure is in a simple relationship to the inverse temperature $\mathcal{D} = |1 - T^{-1}| = |1 - \xi|$.

Fig. 3c applied this measure to the entire fungal dataset and the corresponding control data. The deviation from the no-selection case in the latter is only due to the global CUB q. The actual data-set contains the same global CUB, but additionally it is also subject to SLS and therefore, on the whole, has a greater distance $\mathcal{D}$, indicating a stronger selection acting on it than would be expected from an analysis of q only. Similarly, in fig. 4 we show the distribution of $\mathcal{D}$ for sequences that have a negligible global CUB only. If the beanbag model was true, then these would have distances close to 0, but in reality the values are much higher than that, indicating that even the subsequences with $q \approx 1/2$ are under selective pressure.

We focussed the discussion above on the fungal dataset. A repetition of the same analyses for 560 species of bacteria and for 126 species of protists yielded qualitatively and quantitatively similar results; see supplementary information. Notably, the parameters of the model ξ and γ distributed into the same range and differed only in minor ways in their temperature, such that those subsequences also bear the signatures of SLS.

2.2.5 $\mathcal{D}$ reveals amino acid-specific patterns of codon selection pressure

An advantage of using $\mathcal{D}$ over other measures of codon usage bias is that it lends itself to detecting differences in selective pressure in different subsequence sets. By way of example, we compared how $\mathcal{D}$ differs for different amino acids in the fungal kingdom. Initial visual inspection of the dataset revealed that, as a general pattern, most amino acids in the same organism behave similarly in terms of $\mathcal{D}$, suggesting that they experience similar selective forces. There are, however, also exceptions to this pattern. Fig. 5 reveals that atypically stronger or weaker selection for particular amino acids is an evolutionary feature that is linked to taxonomic groups. In this analysis, we defined atypical selection as a $\mathcal{D}$ value that is more than 2 standard deviations above or below the average $\mathcal{D}$ value for that organism.

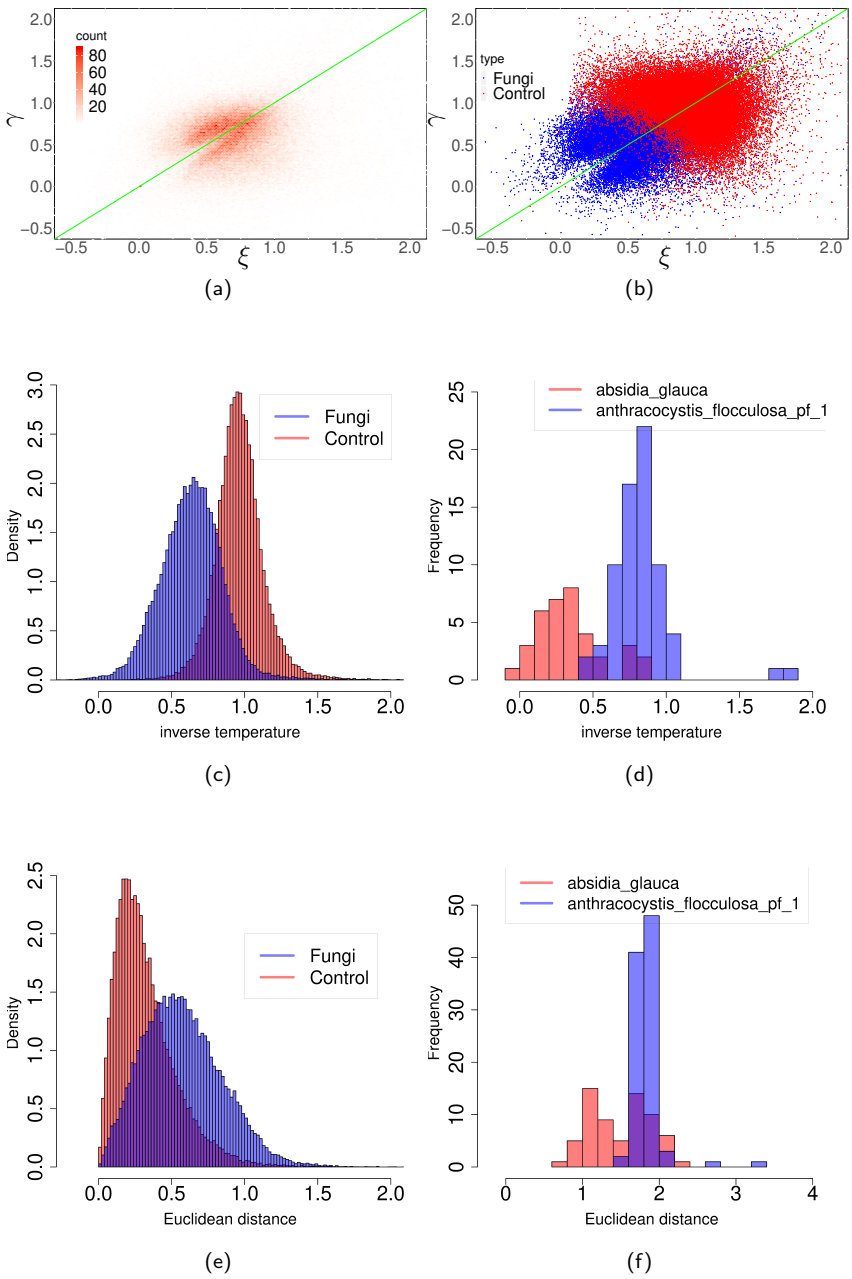


Fig. 3: (a) The density of fitted parameters ξ and γ for each of the 2-codon amino acids for all 462 fungal species in our dataset. We are limiting ourselves to fits with mean-residuals < 0.0009999. The fitted values largely concentrate into the interval of $[0, 1.5]$. (b) Comparing the fitted parameters obtained from the full model (red) to the fitted parameters obtained from the control (blue). The plot shows actual points rather than density. (c) Distribution of inverse temperature in the fungal dataset showing all sub-length and all species. The control data peaks around an inverse temperature of 1, whereas the real data is distributed around a lower inverse temperature. (d) Distribution of inverse temperature for two different species. This shows the temperature for two species including all amino acids and is a subset of (c).3e The distribution of distances $\mathcal{D}$. The control data clearly has a smaller distance on the whole than the non-selection model, indicating that considering only the global codon usage bias underestimates the selection pressure. (f) Same data, but for two species only.

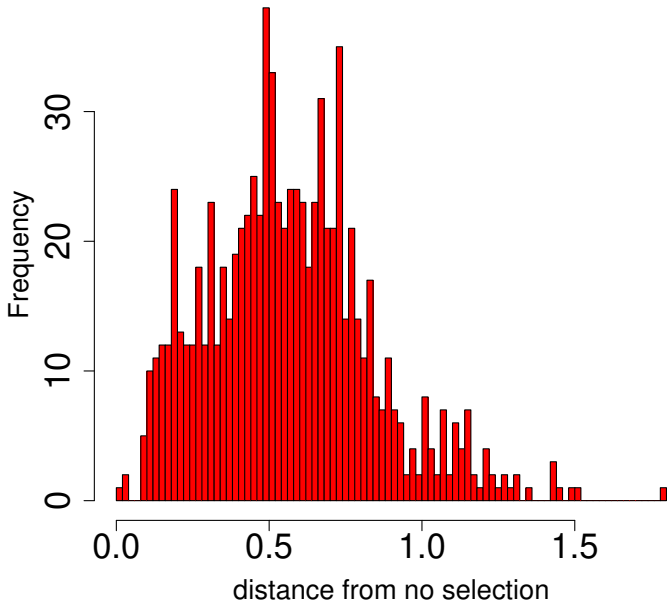


Fig. 4: Distribution of distances $\mathcal{D}$ in genomes that have no global CUB. We selected all subsequences where the global codon usage bias towards codon 1 is between 0.495 and 0.5. The beanbag model of selection would predict that these subsequences have a distance of 0. It is apparent that there are many examples of subsequences that have no global bias, but at the same time subject to a SLS pressure, as evidenced by a distance that is different from 0.

Particularly notable patterns include the *Sordariomycetes* group where the amino acid phenylalanine (F) shows atypically strong codon selection (higher than average $\mathcal{D}$ values) in most species. Interestingly, the pattern is reversed in the *Agaricomycotina* group where selective pressure on phenylalanine codon usage is weaker than for other codons, and in the *Leotiomycetes* group the selection force on phenylalanine codon usage is similar to that of other codons. Some of the observed patterns are highly interesting. For example glutamic acid (E) and aspartic acid (D) are physically very similar, negatively charged amino acids that can frequently be substituted in evolution. Nevertheless in these analyses they show quite distinct behaviour in terms of codon usage selection. The fact that these patterns have remained hidden throughout decades of analysis illustrates the usefulness of $\mathcal{D}$ as a measure of selection acting on codon usage bias.

3 Discussion

At present there is no consensus on the evolutionary drivers of codon usage bias. Many different drivers of codon selection have been described; perhaps many remain to be discovered. Here, we refrained from committing to a particular selection mechanism, but we found that the aggregate effect of all selection forces can be summarised by a parsimonious mathematical model (eq. 1) with only 2 parameters. This model is derived from first principles and it is the simplest model that is consistent with SLS. Its two parameters can be directly interpreted in terms of selection forces namely as the exponents modifying the rate of synonymous mutations from one codon to another one.

The full model eq. 1 has two parts that lend themselves to direct interpretation. The first term on the left hand side is the "entropic" part equivalent to a no-selection system at an altered temperature. The second term is an effective "selection potential" that modifies the probability distributions of the random walkers relative to the purely entropic case of no selection. We do not claim that this potential has a concrete single counterpart in biology. Instead, we interpret it as the emergent result of many evolutionary forces acting simultaneously on the genome. Listing these forces and disentangling how they act and interact is probably an intractable task, but collectively these forces seem to behave in a simple way, leading to a macroscopic description of codon usage bias.

Using this model, we could establish that fungal genomes bear the signature of sequence level codon selection. The same applies to protists and bacteria. While we cannot exclude that beanbag selection acts as well on the genomes, either simultaneously or for a subset of the genomes, mathematically this assumption is not necessary.

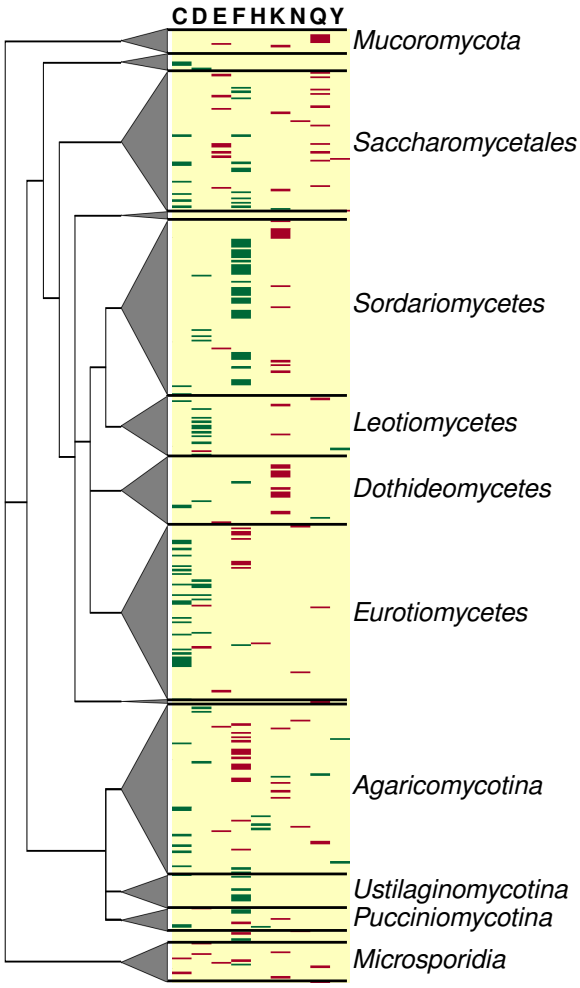


Fig. 5: Amino acid-specific patterns of codon usage bias in fungal genomes. Average $\mathcal{D}$ values were calculated for all subsequences for each amino acid and each genome. Amino acids are highlighted if their $\mathcal{D}$ value was more than 2σ above (green) or below (red) the median $\mathcal{D}$ for that species. In other words, red and green highlights indicate amino acids that are under atypical selection compared to other amino acids in the same species. Species were ordered according to the taxonomic hierarchy in NCBI taxonomy, and taxonomic groups represented with larger numbers of genomes are indicated.

SLS is sufficient to explain all the empirical data and necessary for a subset.

A consequence of SLS is that selection forces on codon usage bias do not only manifest themselves in global codon usage bias, but also, less visibly, in the way codons are distributed. This means that traditional codon usage indices, such a cAI or tAI tend to underestimate the real codon usage bias. From the full model we derived a measure of the distance from the no-selection case, which takes into account both the global codon usage bias but also deviations from the binomial distribution, i.e. SLS effects. In the special case of no global codon usage bias, traditional metrics would conclude that there is no global codon usage bias. However, we showed that in fact even those subsequences do show signatures of selection (see fig. 4).

We limited our analysis above to the 9 amino acids that are encoded by 2 codons only. In principle, there is no theoretical difficulty to extend the model to the remaining 9 amino acids that are encoded by more than 2 codons. The binomial distribution needs to be replaced according to a multinomial distribution and the full model needs to be adapted to include an extra parameter for each additional codon. In practice, the analysis becomes problematic for two reasons. Firstly, with more codons the number of possible subsequence compositions grows quickly, but the number of subsequences does not. As a consequence, there are fewer examples per configuration which increases the statistical error. Secondly, and connected to this is that fitting 3 or more parameter-models to noisy data becomes unreliable. It is still possible to gain some limited insight by comparing entire species rather than only particular subsequences of a given length. This analysis indicates that amino acids with more than 2 codons are much less impacted by SLS and more consistent with beanbag selection; see supplementary information for details. Based on general considerations this is not entirely surprising. The same statistical error that makes the analysis of amino acids with more than 2 codons difficult also affects the cells itself, in the sense that the effects of even a moderate selection pressures at the level of the sequence will remain inefficient against the high levels of mutational noise.

4 Methods

4.1 The dataset

All datasets were obtained from ENSEMBL `https://www.ensembl.org`. To obtain the results we first downloaded the coding sequences of interest (CDS files); for 462 species from the Fungi kingdom (release 36 in AUG 2017), 442 species in Bacteria kingdom (release 40 in JUL 2018), 143 species in Protist kingdom (release 40 in JUL 2018). All the species names and corresponding download weblinks are in the supplementary file "species.xlsx". From these files we produced all subsequences. To do this we converted each gene sequence into a valid codon sequence. We then removed all genes of which the number of nucleotides was not a multiple of 3, which indicates an error in the ORF annotation. There are 35748 error genes excluded from 4554328 total genes of Fungi kingdom, 6384 excluded from 1286467 of Bacteria kingdom, and 25142 excluded from 1439975 of Protist kingdom.

We then prepared the data by splitting each gene into (up to) 18 subsequences as follows: For each gene g and amino acid $\mathcal{A}$ in the dataset we found all codons that code for $\mathcal{A}$ and discarded all others. Thus, we reduced the gene g to a subsequence of codons of length $L^{\mathcal{A},g}$.

For each species, a control coding subsequence was produced by replacing each codon with a random synonymous codon (which could be the same as the one in the original subsequence). The probability of choosing a random synonymous codon was biased according to the observed global codon usage bias of the respective species and amino acid, such that in the control data the codons were distributed according to the multinomial distribution by construction.

Fitting was done using Maple 2018 "NonlinearFit" function. The initial estimates for γ and ξ were set to 1. If an initial fit resulted in a mean-residual > 0.0009999 then the fit was repeated with randomly chosen initial estimates. This was repeated up to 1000 times until a fit was found with a mean-residual < 0.00099999.

All datasets necessary to reproduce the results of this article are available via `https://www.cs.kent.ac.uk/projects/statthermcub`. This includes the links to the ENSEMBL genomes, codon usage tables and processed data on subsequences.

Author contribution

Conceived the research: DC,TVH; did the research: DC,YD,TVH, wrote the paper: DC, YD,TVH,analysed the data: DC,YD,TVH,JK

References

[1] S. Behura and D. Severson. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biol Rev Camb Philos Soc*, 88(1):49–61, Feb 2013.

[2] G. Hanson and J. Coller. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.*, 19(1):20–30, Jan 2018.

[3] T. Ikemura. Correlation between the abundance of yeast transfer rnas and the occurrence of the respective codons in protein genes. differences in synonymous codon choice patterns of yeast and escherichia coli with reference to the abundance of isoaccepting transfer rnas. *J Mol Biol*, 158(4):573–597, Jul 1982.

[4] G. Bernardi and G. Bernardi. Codon usage and genome composition. *J. Mol. Evol.*, 22(4):363–365, 1985.

[5] G. Marais, D. Mouchiroud, and L. Duret. Does recombination improve selection on codon usage? lessons from nematode and fly complete genomes. *Proceedings of the National Academy of Sciences*, 98(10):5688–5692, apr 2001.

[6] M. Zhou, J. Guo, J. Cha, M. Chae, S. Chen, J. Barral, M. Sachs, and Y. Liu. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, 495(7439):111–115, feb 2013.

[7] D. Chu, E. Kazana, N. Bellanger, T. Singh, M. Tuite, and T. von der Haar. Translation elongation can control translation initiation on eukaryotic mrnas. *EMBO Journal*, 33(1):21–34, Jan 2014.

[8] A. Diament, R. Y. Pinter, and T. Tuller. Three-dimensional eukaryotic genomic organization is strongly correlated with codon usage expression and function. *Nat Commun*, 5:5876, Dec 2014.

[9] A. K. Hess, P. Saffert, K. Liebeton, and Z. Ignatova. Optimization of translation profiles enhances protein expression and solubility. *PLoS ONE*, 10(5):e0127039, 2015.

[10] G. Zhang, I. Fedyunin, O. Miekley, A. Valleriani, Alessandro Moura, and Zoya Ignatova. Global and local depletion of ternary complex limits translational elongation. *Nucleic Acids Research*, 38(14):4778–4787, Aug 2010.

[11] M. Orešič, M. Dehn, D. Korenblum, and D. Shalloway. Tracing specific synonymous codon-secondary structure correlations through evolution. *Journal of Molecular Evolution*, 56(4):473–484, apr 2003.

[12] M. Frenkel-Morgenstern, T. Danon, T. Christian, T. Igarashi, L. Cohen, Y. Hou, and L. Jensen. Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. *Mol. Syst. Biol.*, 8:572, Feb 2012.

[13] A. Bastide, D. Peretti, J. R. Knight, S. Grosso, R. V. Spriggs, X. Pichon, T. Sbarrato, A. Roobol, J. Roobol, D. Vito, M. Bushell, T. von der Haar, C. M. Smales, G. R. Mallucci, and A. E. Willis. RTN3 Is a Novel Cold-Induced Protein and Mediates Neuroprotective Effects of RBM3. *Curr. Biol.*, 27(5):638–650, Mar 2017.

[14] M. Zhou, J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. S. Sachs, and Y. Liu. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, 495(7439):111–115, Mar 2013.

[15] H. Qin, W. Wu, J. Comeron, M. Kreitman, and W. Li. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, 168(4):2245–2260, dec 2004.

[16] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–354, Apr 2010.

[17] A. Coghlan and K. H. Wolfe. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast*, 16(12):1131–1145, Sep 2000.

[18] A. Diament, A. Feldman, E. Schochet, M. Kupiec, Y. Arava, and T. Tuller. The extent of ribosome queuing in budding yeast. *PLoS Comput. Biol.*, 14(1):e1005951, 01 2018.

[19] L. Jossé, T. Singh, and T. von der Haar. Experimental determination of codon usage-dependent selective pressure on high copy-number genes in saccharomyces cerevisiae. *bioRxiv*, 2018.

[20] F. Wright. The 'effective number of codons' used in a gene. *Gene*, 87(1):23–29, Mar 1990.

[21] H. Akashi. Inferring weak selection from patterns of polymorphism and divergence at silent sites in Drosophila DNA. *Genetics*, 139(2):1067–1076, Feb 1995.

[22] S. Karlin and J. Mrazek. What drives codon choices in human genes? *J. Mol. Biol.*, 262(4):459–472, Oct 1996.

[23] A. Urrutia and L. Hurst. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, 159(3):1191–1199, Nov 2001.

[24] J. A. Novembre. Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.*, 19(8):1390–1394, Aug 2002.

[25] G. West, J. Brown, and B. Enquist. A general model for the origin of allometric scaling laws in biology. *Science*, 276(5309):122–126, 1997.

[26] P. Zörnig and G. Altmann. Unified representation of Zipf distributions. *Computational Statistics & Data Analysis*, 19(4):461–473, apr 1995.

[27] G. Cristadoro, M. Degli Esposti, and E. G. Altmann. The common origin of symmetry and structure in genetic sequences. *Sci Rep*, 8(1):15817, Oct 2018.

[28] P. Bak and K. Sneppen. Punctuated Equilibrium and Criticality in a Simple Model of Evolution. *Phys. Rev. Lett.*, 71:4083–4086, 1993.

[29] H. Flyvbjerg, K. Sneppen, and P. Bak. Mean Field Theory for a Simple Model of Evolution. *Phys. Rev. Lett.*, 71:4087–4090, 1993.

[30] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012.

[31] J. Parrondo, J. Horowitz, and T. Sagawa. Thermodynamics of information. *Nature Physics*, 11(2):131–139, feb 2015.

[32] P. J. Kersey, J. E. Allen, A. Allot, M. Barba, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, C. Grabmueller, N. Kumar, Z. Liu, T. Maurel, B. Moore, M. D. McDowall, U. Maheswari, G. Naamati, V. Newman, C. K. Ong, M. Paulini, H. Pedro, E. Perry, M. Russell, H. Sparrow, E. Tapanari, K. Taylor, A. Vullo, G. Williams, A. Zadissia, A. Olson, J. Stein, S. Wei, M. Tello-Ruiz, D. Ware, A. Luciani, S. Potter, R. D. Finn, M. Urban, K. E. Hammond-Kosack, D. M. Bolser, N. De Silva, K. L. Howe, N. Langridge, G. Maslen, D. M. Staines, and A. Yates. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.*, 46(D1):D802–D808, Jan 2018.

[33] W. Faller, T. Jackson, J. Knight, R. Ridgway, T. Jamieson, S. Karim, C. Jones, S. Radulescu, D. Huels, K. Myant, K. Dudek, H. Casey, A. Scopelliti, J. Cordero, M. Vidal, M. Pende, A. Ryazanov, N. Sonenberg, O. Meyuhas, M. Hall, M. Bushell, A. Willis, and O. Sansom. mTORC1-mediated translational elongation limits intestinal tumour initiation and growth. *Nature*, 517(7535):497–500, Jan 2015.

[34] P. M. Sharp and W. H. Li. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3):1281–1295, Feb 1987.

[35] M. dos Reis, L. Wernisch, and R. Savva. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Research*, 31(23):6976–6985, Dec 2003.

[36] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012.

[37] G. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *Journal of Statistical Physics*, 90(5/6):1481–1487, 1998.

A Theory

A.1 Beanbag model

Beanbag models assume that selection acts at the level of codons. Assigning codons to a particular site can be thought of as selecting beans from a bag. Each time a codon is picked, the choice between the codons / bags is made with a fixed probability that does not depend on the previous choices made. For each codon assignment there will be a probability p_i that the i-th codon is chosen among all possible $|C^{\mathcal{A}}|$ codons $C_1^{\mathcal{A}}, \ldots, C_{|C^{\mathcal{A}}|}^{\mathcal{A}}$ for amino acid A. This type of selection procedure is generally known to lead to a multinomial distribution of codon subsequences. Specifically, the probability to observe a particular gene that has $L = L^{\mathcal{A},g}$ occurrences of a particular amino acid $\mathcal{A}$ with $m := |C^{\mathcal{A}}|$, and k_i occurrences of the i-th codon is given by

$$P = \frac{L!}{k_1! \cdots k_m!} q_1^{k_1} \cdots q_m^{k_m} \tag{2}$$

Here q_i is the global codon usage bias of the i-th codon. In the special case when there are only two codons in the amino acid, i.e. $m = 2$ then this probability reduces to the binomial distribution. We set $k := k_1$, $k_2 = L - k$, $p := q$ and $p_2 = 1 - q$.

$$P = \binom{L}{k} q^k (1-q)^{L-k} \tag{3}$$

A.1.1 Mixture models

A variant of the codon model is the multinomial mixture model, whereby selection still happens at the level of the individual subsequences, but the probability to select a codon depends on the particular subsequence. If we assume that there are M different groups of genes that each share the same selection pressure and the fraction of genes belonging to group i is F_i then we can calculate the expected frequencies of genes as follows (for the case of only 2 codons):

$$\begin{aligned}\mathcal{L}(k) &:= F_1 \binom{L}{k} \left(q_1^k (1-q_1)^{L-k} \right) + F_2 \binom{L}{k} \left(q_2^k (1-q_2)^{L-k} \right) + \ldots F_M \binom{L}{k} \left(q_M^k (1-q_M)^{L-k} \right) \\ &= \binom{L}{k} \langle q_i^k (1-q_i)^{L-k} \rangle_F .\end{aligned}$$

Here the brackets indicate an ensemble average over the M classes. The empirical probability to observe a subsequence with k codons of type one is then given by

$$P_k = \frac{\mathcal{L}(k)}{\sum_{i=1}^{L} \mathcal{L}(i)}. \tag{4}$$

As the number of classes M grows, the probability P_k approaches the uniform probability distribution, since

$$\int_0^1 q_1^k (1-q_1)^{L-k} dq = \frac{1}{N+1}\binom{N}{k}.$$

A.2 Codon usage bias as a random walk

An alternative to the beanbag model is a model where not individual codons are selected, but specific subsequences. As an extreme example, consider the scenario where there is no underlying global codon usage bias at all and all codons occur equally often across the genome, but within a particular gene all codons for an amino acid $\mathcal{A}$ are strictly always the same. In this case the probability $p_i^{\mathcal{A}} = 1/|C^{\mathcal{A}}|$, but the probability to observe a particular subsequence would not follow the binomial distribution. In fact, there would be only $C^{\mathcal{A}}$ different subsequences of codons for each amino acid of a particular gene.

For the derivation of the full model, it is convenient to consider codon evolution as a random walk. We consider here a model whereby the position of the codons does not matter, and we only care about the number of codons in a particular subsequence. This analysis remains thus insensitive to correlations of codon usage within genes.

For simplicity, we focus here exclusively on amino acids with 2 codons, yet extensions to more codons are straightforward. In the case of 2 codons, codon evolution can be represented as a 1D random walk in discrete space and continuous time. The number of sites is $L^{\mathcal{A},g}$. Each site corresponds to a particular subsequence composition, i.e. a pair $(k_1^{\mathcal{A},g}, L^{\mathcal{A},g} - k_2^{\mathcal{A},g})$, which defines a state. In the case of 2 codons the state is entirely characterised by k_1. States are connected (i.e. accessible to the random walker) when a single synonymous codon change is sufficient to get from one state to the other. In the case of 2 codons this means that there are two states with one connected state $(0, L^{\mathcal{A},g})$ and $(L^{\mathcal{A},g}, 0)$). All other states are connected to two other states states, such that k_1 is connected to $(k_1 - 1, L^{\mathcal{A},g} - k_1 + 1)$ and $(k_1 + 1, , L^{\mathcal{A},g} - k_1 + 1)$. When there are more than 2 codons, then each state is connected to more than two states.

A transition from one state to another always involves that a codon of a particular type is changed to a codon of a different type. The rate of such an event is proportional to the number of codons of the type that is lost. For example, the rate of transitions where codon 2 is converted to a codon of type 1 is proportional to k_2.

$$r(k_1, k_2, \ldots \to k_1 + 1, k_2 - 1, \ldots) \sim k_2 \tag{5}$$

In order to derive a model, we now take a conceptual leap and (i) model each subsequence as a site of the random walk that has a given energy E_i, where the index i counts the numbers of codons of type 1 (which we choose arbitrarily). This energy is, at least for the time being, an entropic energy in that it is the (conceptual) consequence of the number of subsequence configurations that have k_1 codons of type 1. This number is lowest when all codons are of the same type and highest when all codons are used exactly equally often. (ii) We posit that the random walker that moves between the sites is in contact with a large heat-bath that remains at a fixed temperature T. This temperature bath exchanges energy with the walker, thus enables it to transition to sites with higher energy, as well as lower energy while extracting/adding energy to the heat bath. We stress here that this idea of energy and temperature are merely conceptual devices and should not be confused with the actual physical temperature that is experience by organisms.

Having made this conceptualisation, we can now apply well established concepts from stochastic thermodynamics [36] to this problem in order calculate the long-term equilibrium probabilities $\pi(k_1, k_2, \ldots)$ for various configurations. To do this we impose that the system obeys the detailed balance condition, which means that in equilibrium there are no net-flows of probability between any two states that are connected.

$$\pi(k_1, k_2, \ldots) r(k_1, k_2, \ldots \to k_1 + 1, k_2 - 1, \ldots) = \pi(k_1 + 1, k_2 - 1, \ldots) r(k_1 + 1, k_2 - 1, \ldots \to k_1, k_2, \ldots) \tag{6}$$

Here $r(k_1, k_2, \ldots \to k_1 + 1, k_2 - 1, \ldots)$ is the rate of a mutation from any of the k_1 codons of type 1 to codon 2. This implies that

$$\frac{\pi(k_1, k_2, \ldots)}{\pi(k_1 + 1, k_2 - 1, \ldots)} = \frac{k_1 + 1}{k_2} \tag{7}$$

We also know that the occupation probabilities of a random walker of the type described here obeys the Boltzmann distribution in equilibrium. This means that the probability to find the walker in a given configuration depends on the energy of the configuration $E(k_1, k_2, \ldots)$ in the following way:

$$\pi(k_1, k_2, \ldots) = \frac{1}{Z} \exp\left(-\frac{E(k_1, k_2, \ldots)}{k_B T}\right). \tag{8}$$

Since we are not interested here in specific units, we will henceforth set the Boltzmann constant $k_B = 1$.

The energy is *a priori* unknown, but we postulate that the local detail balance condition is fulfilled [37]:

$$T \ln\left(\frac{r_+}{r_-}\right) = \Delta E, \tag{9}$$

where r_+ and r_- are the forwards and backwards transition rates respectively. This relationship then implies a dependence between the two rates, namely:

$$r_+ = r_- \exp\left(\frac{\Delta E}{T}\right) \tag{10}$$

A.2.1 Calculation of the energy of a specific codon in the completely unbiased beanbag model.

We first calculate the energies associated with a model where there is no selection pressure at all and codons perform a random walk only. This is the beanbag model where all codons have the same probability of being chosen. To calculate the energies of each subsequence, we start from a subsequence where all L codons are of type 1. We define this state as having energy $E_0 = 0$. A mutation can reach the next state (1) by changing one of the L codons of type 1 to a codon of type 2. When there are no selection forces then this happens with a rate proportional to L. From this state (1), the system can then move further to state (2). This happens now with a rate proportional to $L-1$. Alternatively, with a rate proportional to 1 it can move back to state (0). Altogether, the following transitions are thus possible:

$$(L,0) \underset{1}{\overset{L}{\rightleftharpoons}} (L-1,1) \underset{2}{\overset{L-1}{\rightleftharpoons}} \cdots \underset{k}{\overset{L-k+1}{\rightleftharpoons}} (L-k,k) \underset{k+1}{\overset{L-k}{\rightleftharpoons}} (L-k-2,k+2) \cdots \underset{L}{\overset{1}{\rightleftharpoons}} (0,L) \tag{11}$$

Using the local detailed balance condition (eq. 9) and because the energy associated with state (0) is by construction $E_0 := 0$, the energy difference between state (0) and state (1) is $\Delta E_1 = \ln(L/1)$. Generally, the energy difference ΔE_i between state (i) and state $(i-1)$ is given by $\Delta E_i = \ln\left((L-i+1)/i\right)$. Consequently, the energy of state (k) is then:

$$\begin{aligned} E_k &= \sum_{i=1}^{k} \Delta E_i = -T \ln\left(\prod_{i=1}^{k} \frac{L-i+1}{i}\right) = \\ &= -T \ln\left(\frac{L!}{(L-k)!k!}\right) = -T \ln\binom{L}{k}. \end{aligned} \tag{12}$$

This entails that the (Boltzmann-)probability

$$P(E_k) = \frac{\exp\left(-\frac{E_k}{T}\right)}{\sum_i \exp\left(-\frac{E_i}{T}\right)}$$

to observe a specific configuration k is proportional to the binomial coefficient, which indicates that this system is distributed according to the binomial distribution eq. 3 with $p = 1/2$. As such it represents a model with no selection pressure at the level of sequences.

A.2.2 The binomial distribution with $q \neq 1/2$.

We now establish an energy model of the binomial distribution when there is an underlying bias to the mutations. In this case then the model is (c.f. eq. 11):

$$(L,0) \underset{1\cdot(1-q)}{\overset{Lq}{\rightleftharpoons}} (L-1,1) \underset{2(1-q)}{\overset{(L-1)q}{\rightleftharpoons}} \cdots \underset{k(1-q)}{\overset{(L-k+1)q}{\rightleftharpoons}} (L-k,k) \underset{(k+1)(1-q)}{\overset{(L-k)q}{\rightleftharpoons}} (L-k-2,k+2) \cdots \underset{L(1-q)}{\overset{1\cdot q}{\rightleftharpoons}} (0,L) \tag{13}$$

Following the same reasoning as above, we can establish the energy differences:

$$\begin{aligned} \hat{E}_0 &= 0 \\ \Delta\hat{E}_1 &= -T \ln\left(\frac{Lq}{1\cdot(1-q)}\right) \\ \Delta\hat{E}_2 &= -T \ln\left(\frac{(L-1)q}{2(1-q)}\right) \\ \Delta\hat{E}_k &= -T \ln\left(\frac{(L-k+1)q}{k(1-q)}\right) \end{aligned} \tag{14}$$

Hence, it follows that the energy $\hat{E}_k$ is given by:

$$\begin{aligned}\hat{E}_k &= -\ln\left(\frac{(L-k+1)!}{(L-k)!k!}\cdot\frac{q^k}{(1-q)^k}\right)\\ &= -\ln\binom{L}{k}+\ln\left(\frac{(1-q)^k}{q^k}\right)\\ &= E_k+\ln\left(\frac{(1-q)^k}{q^k}\right)\end{aligned} \tag{15}$$

The corresponding Boltzmann distribution can be shown to be the binomial distribution eq. 3 with $p = q$.

A.2.3 Asymmetric bias

For completeness we also consider the case where the transition rates are biased in one direction only. To this end we modify the rates such that there is a constant bias C into one direction. The random walk is then modified as follows:

$$(L,0)\underset{C}{\overset{L}{\rightleftharpoons}}(L-1,1)\underset{2C}{\overset{(L-1)}{\rightleftharpoons}}\cdots\underset{kC}{\overset{(L-k+1)}{\rightleftharpoons}}(L-k,k)\underset{(k+1)C}{\overset{(L-k)}{\rightleftharpoons}}(L-k-2,k+2)\cdots\underset{LC}{\overset{1}{\rightleftharpoons}}(0,L) \tag{16}$$

Following the same steps as above, we can calculate energies for this model.

$$\begin{aligned}E'_k &= \sum_{i=1}^{k}\Delta E'_i = -T\ln\left(\prod_{i=1}^{k}\frac{(L-i+1)^\gamma}{i^\gamma C}\right)=\\ &= -T\ln\left(\frac{L!}{(L-k)!k!}\right)+k\ln(C)=E_k+ln(C^k)\end{aligned} \tag{17}$$

The average codon usage ϵ is then given by

$$q=\frac{A}{LZ}=\frac{1}{C+1} \tag{18}$$

A.3 Sequence selection models

We now further modify the rates of the model and assume that there is a selection pressure on the genome such that the rate with which codon 2 mutates to codon 1 and *vice versa* is now no longer proportional to the number $N-k+1$ and k of codon 2, but proportional to a power of this number. This then changes the above model, as follows:

$$(L,0)\underset{1^\gamma}{\overset{L^\gamma}{\rightleftharpoons}}(L-1,1)\underset{2^\gamma}{\overset{(L-1)^\gamma}{\rightleftharpoons}}\cdots\underset{k^\gamma}{\overset{(L-k+1)^\gamma}{\rightleftharpoons}}(L-k,k)\underset{(k+1)^\gamma}{\overset{(L-k)^\gamma}{\rightleftharpoons}}(L-k-2,k+2)\cdots\underset{L^\gamma}{\overset{1^\gamma}{\rightleftharpoons}}(0,L) \tag{19}$$

The energies then become:

$$\begin{aligned}\tilde{E}_k &= \sum_{i=1}^{k}\Delta\tilde{E}_i = -T\ln\left(\prod_{i=1}^{k}\frac{(L-i+1)^\gamma}{i^\gamma}\right)=\\ &= -T\gamma\ln\left(\frac{L!}{(L-k)!k!}\right)=\gamma E_k\end{aligned} \tag{20}$$

We conclude that, if the transition rates are as in the unbiased case, but modified by the exponent, then the energies of the model remain unaffected, but the temperature of the system changes according to γ. Note that in this model, on average, the number of codons of type 1 will always be the same as the number of codons of type 2. Thus no global codon usage bias would occur. Selection manifests itself in the way that codons are distributed across subsequences, which is quantified by the inverse temperature γ.

Finally, we now consider the most general model with different exponents to the left and to the right to arrive at the full model.

$$(L,0)\underset{1^\gamma}{\overset{L^\xi}{\rightleftharpoons}}(L-1,1)\underset{2^\gamma}{\overset{(L-1)^\xi}{\rightleftharpoons}}\cdots\underset{k^\gamma}{\overset{(L-k+1)^\xi}{\rightleftharpoons}}(L-k,k)\underset{(k+1)^\gamma}{\overset{(L-k)^\xi}{\rightleftharpoons}}(L-k-2,k+2)\cdots\underset{L^\gamma}{\overset{1^\xi}{\rightleftharpoons}}(0,L) \tag{21}$$

This changes the energies of the model in the following way:

$$\begin{aligned}\bar{E}_k &= \sum_{i=1}^{k}\Delta\bar{E}_i = -T\ln\left(\prod_{i=1}^{k}\frac{(L-i+1)^\zeta}{i^\gamma}\right)\\ &= -T\xi\ln\binom{L}{k}+T(\gamma-\xi)\ln(k!)\\ &= \xi E_k+T(\gamma-\xi)\ln(k!)\end{aligned} \tag{22}$$

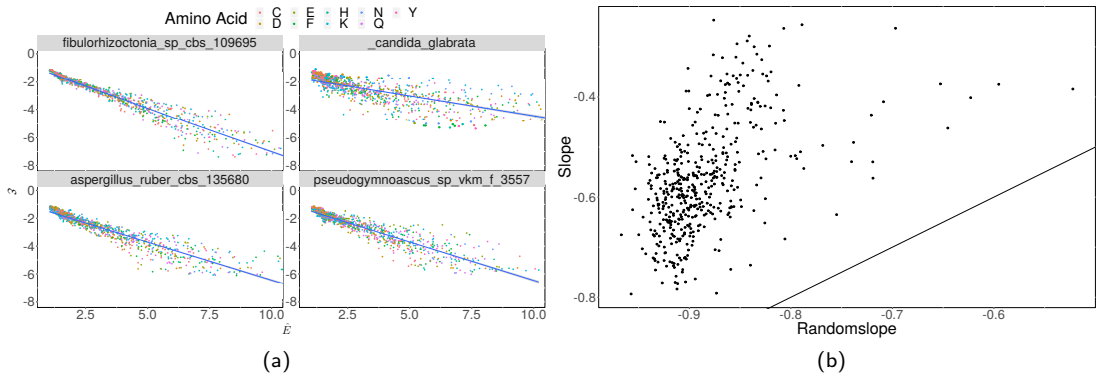


Fig. 6: (a) Empirical energies $\mathcal{E}$ as a function of the energies $\hat{E}$ derived from the multinomial model assuming the global codon usage bias for each amino acid. Each plot combines data for 9 amino acids and 10 different lengths. Each point shows the theoretical $\hat{E}_k$ along the horizontal axis, denoting the logarithm of the binomial probability to observe a exactly k codons of type one in a subsequence of length $L^{\mathcal{A}}$, given the global codon usage bias for amino acid $\mathcal{A}$ in this species. The corresponding $\mathcal{E}$ is the logarithm of the observed fraction of occurrences of subsequences with k codons of type 1 amongst all subsequences of length $L^{\mathcal{A}}$. (b) The slopes obtained for all 462 fungal species plotted against slopes obtained from randomised genes where all codons were replaced by a random synonymous codons draw according to the global codon usage bias. If codons were distributed randomly across genes, then the points would cluster along the diagonal.

It is apparent that this case changes both the inverse temperature of the system and the expected number of codons of type 1 relative to the case of no selection. The corresponding Boltzmann distribution,

$$P(\bar{E}_k) = \frac{\xi E_k + T(\gamma - \xi)\ln(k!)}{\sum_i \left(\xi E_i + T(\gamma - \xi)\ln(i!)\right)} \tag{23}$$

is a mono-modal distribution. The mean of the distribution depends on the parameters ξ and γ. For certain choices of these parameters, $P(\bar{E}_k)$ can approximate a binomial distribution with some bias $p \neq 1/2$.

B Global temperatures

For each set of subsequences corresponding to a particular length and amino acid, we next define an empirical energy $\mathcal{E}$ as

$$\mathcal{E}_{k_1,\ldots,k_{|C^{\mathcal{A}}|}} := -\ln\left(\frac{n_{k_1,\ldots,k_{|C^{\mathcal{A}}|}}}{N}\right), \tag{24}$$

where $n_{k_1,\ldots,k_{|C^{\mathcal{A}}|}}$ is the number of occurrences of subsequences with configuration $k_1,\ldots,k_{|C^{\mathcal{A}}|}$ in the species and N is the number of subsequences of amino acid $\mathcal{A}$ with length $L = \sum_{i=1}^{|C^{\mathcal{A}}|} k_i$ in the species. In order to assign a global temperature to a species as a whole we plot the empirical energy $\mathcal{E}_{k_1,\ldots,k_{|C^{\mathcal{A}}|}}$ against the energy $\hat{E}_{k_1,\ldots,k_{|C^{\mathcal{A}}|}}$.

If the codons were distributed according to the multinomial distribution, then up to statistical errors

$$\mathcal{E}^L_{k_1,\ldots,k_{k_1,\ldots,k_{|C^{\mathcal{A}}|}}} = \hat{E}^L_{k_1,\ldots,k_{k_1,\ldots,k_{|C^{\mathcal{A}}|}}}$$

, i.e. plotting the empirical energies against $\hat{E}$ would result in a straight line with slope 1. In contrast, if the distribution of subsequences was distributed by a multinomial distribution with a higher temperature, then this plot would still result in a straight line, but with a slope corresponding to $1/T$.

We plotted the empirical energies $\mathcal{E}$ against $\hat{E}$ and found that the two energies are in good approximation, related via a straight line. Fig. 6a shows 4 examples of fungal species for the 9 amino acid with exactly 2 codons. In order to check the temperature of a species, we fitted $\mathcal{E}$ versus $\hat{E}$ to a straight line. The slope of this straight line can then be taken as an estimate for the inverse temperature for each of the 462 fungal species in our dataset, see fig. 6a.

We found that the slopes are significantly and systematically different from 1 and significantly and systematically higher than slopes obtained from random controls; see fig. 6. Seen globally, the model is thus consistent with the full model and thus consistent with the hypothesis of sequence-level selection. At the same time, the significantly altered slopes are not consistent with the assumption of a purely codon-based selection, which would lead to a binomial distribution of data, and hence slopes close to 1.

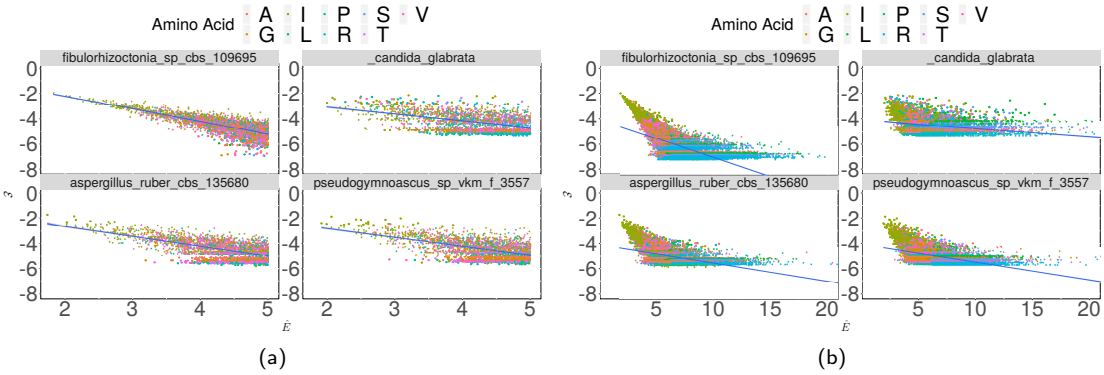


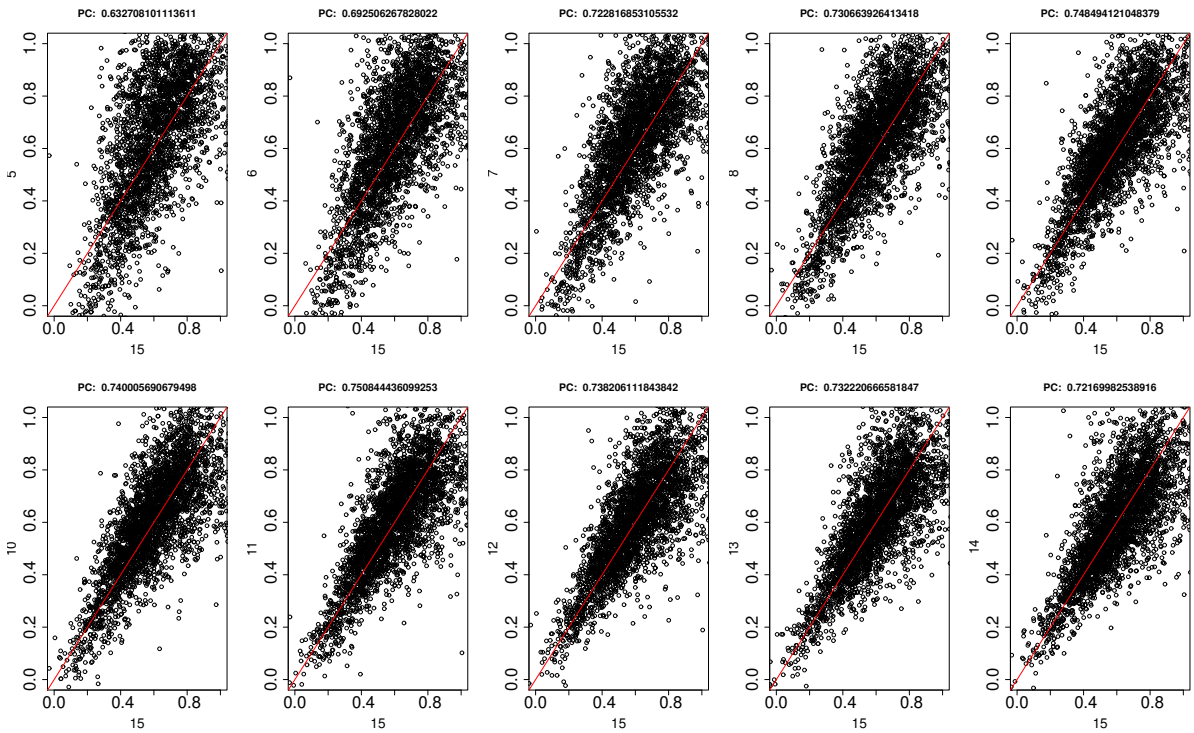
Fig. 7: (a) Empirical energies $\mathcal{E}$ as a function of the energies $\hat{E}$ derived from the multinomial model assuming the global codon usage bias for each amino acid. Each plot combines data for the amino acids with more than 2 codons for length 5-15. Each point shows the theoretical $\hat{E}_k$ along the horizontal axis, denoting the logarithm of the multinomial probability to observe a exactly k_1 codons of type 1, k_2 codons of type 2, etc.., in a subsequence of length $L^{\mathcal{A}}$, given the global codon usage bias for amino acid $\mathcal{A}$ in this species. The corresponding $\mathcal{E}$ is the logarithm of the observed fraction of occurrences of subsequences with k codons of type 1 amongst all subsequences of length $L^{\mathcal{A}}$. The data (and the fit) is restricted to the subsequences where the theoretical distribution is greater than exp(-5). The blue line represent a fit to the data. (b) Same as 7a but the probabilities the data is not restricted to the high probability subsequences. The plot is dominated by noise.

While this global data on the slope can be used qualitatively to demonstrate the signature of sequence-level selection acting on codon usage, it cannot be reliably used to quantify this selection pressure. While the result that slopes systematically deviate from the random model is robust with respect to changes of the fitting protocol, the precise numerical values are not robust to such changes. In fig. 6b we reported slopes obtained by fitting the lines to the entire dataset. If instead fits are restricted to energies $\hat{E} < 5$, which would exclude the low probability points, then somewhat different numerical results would appear. Moreover, the global slopes obtained thus are in no clear relationship with the summary statistics of the detailed fits of the full model to subsequences, even though they consist of exactly the same data.

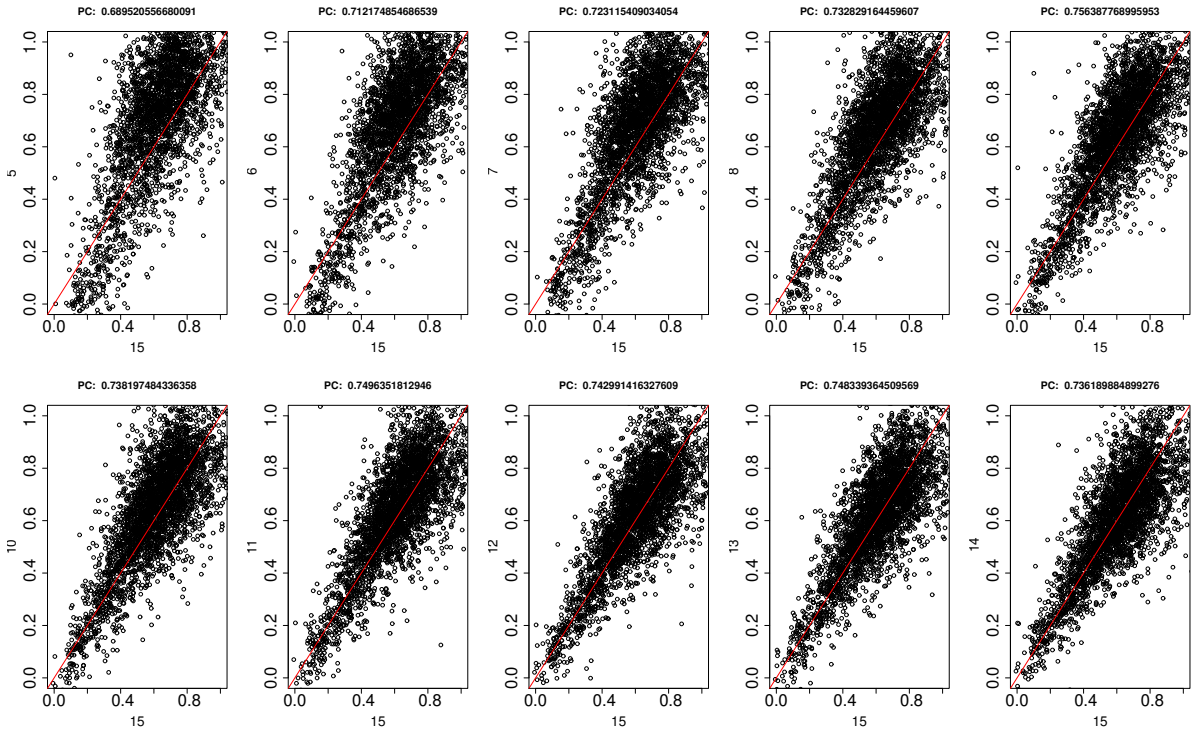
We then repeated the same plot for the same species, but this time for the amino acids that have more than 2 codons. Fig. 7 demonstrates that these also lead to approximately straight lines, but only for the subset of high probability subsequences. For those we obtained slopes that were close to 1. When all subsequences are included then noise dominates and no meaningful fitting is possible; see fig. 7b.

C Supplementary plots

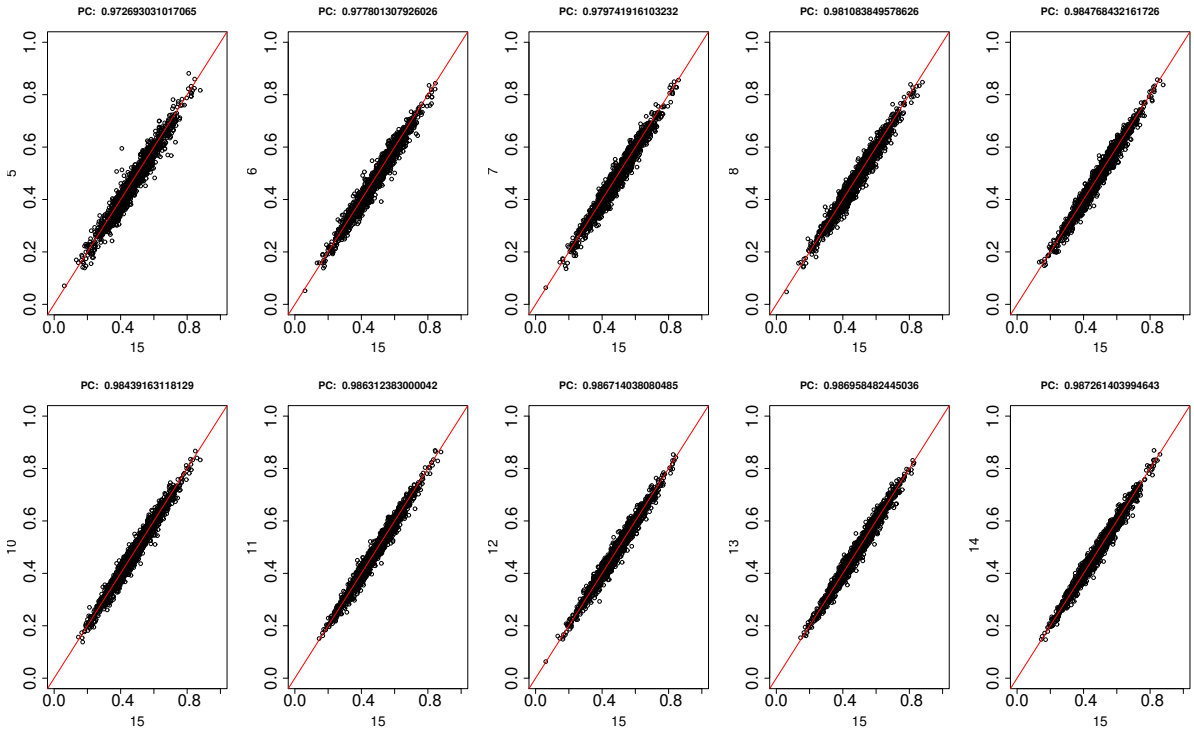
Correlation for the parameter ξ for the fungal dataset. We include only those fits where the mean-residual was smaller than 0.0009999. The horizontal axis indicates the fitted value for the subsequence length of 15 and the vertical axis for the subsequences values $5-14$. The title indicates the Pearson correlation.

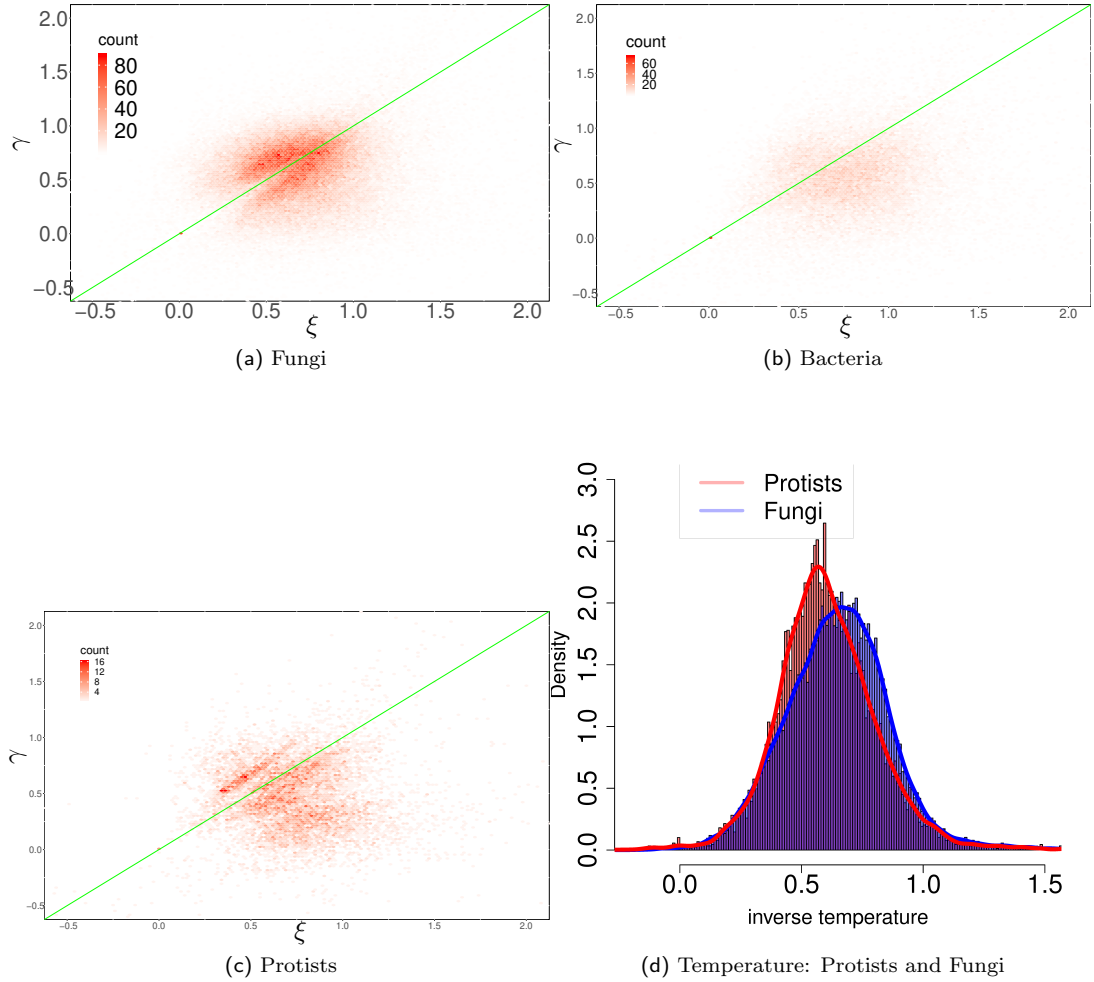


Same as above but for parameter γ.



Same as above, but for the fits to the binomial distribution.





(a) Fungi (b) Bacteria

(c) Protists (d) Temperature: Protists and Fungi

Fig. 8: (a)-(c) The fitted values of parameters ξ and γ for each of the 2-codon amino acids for bacteria, fungi and protists. The graphs show heatplots that summarise the density of points in the area. Red indicates a high density of points. We are limiting ourselves to those amino acid subsequences that have a sub-length of 15. (d) The distribution of inverse temperatures of protists and fungi. There is considerable overlap between the two groups. Density estimates for each group are overlayed on the graph to aid the eye. It appears that fungi are somewhat cooler than protists. Bacteria would lie in-between protists and fungi, but are omitted to aid graph readability.