# Hidden patterns of codon usage bias across kingdoms

*Yun Deng*[1], *Jeremie Kalfon*[1], *Dominique Chu*[1] *and 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**
{D.F.Chu,T.von-der-Haar@}@kent.ac.uk

**Abstract**

The genetic code is necessarily degenerate with 64 possible nucleotide triplets being translated into 20 amino acids. 18 out of the 20 amino acids are encoded by multiple synonymous codons. While synonymous codons are clearly equivalent in terms of the information they carry, it is now well established that they are used in a biased fashion. There is currently no consensus as to the origin of this bias. Drawing on ideas from stochastic thermodynamics we derive from first principles a mathematical model describing the statistics of codon usage bias. We show that the model accurately describes the distribution of codon usage bias of genomes in the fungal and bacterial kingdoms. Based on it, we derive a new computational measure of codon usage bias — the distance $\mathcal{D}$ capturing two aspects of codon usage bias: (*i*) Differences in the genome-wide frequency of codons and (*ii*) apparent non-random distributions of codons across mRNAs. By means of large scale computational analysis of over 900 species across 2 kingdoms of life, we demonstrate that our measure provides novel biological insights. Specifically, we show that while codon usage bias is clearly based on heritable traits and closely related species show similar degrees of bias, there is considerable variation in the magnitude of $\mathcal{D}$ within taxonomic classes suggesting that the contribution of sequence-level selection to codon bias varies substantially within relatively confined taxonomic groups. Interestingly, commonly used model organisms are near the median for values of $\mathcal{D}$ for their taxonomic class, suggesting that they may not be good representative models for species with more extreme $\mathcal{D}$, which comprise organisms of medical an agricultural interest. We also demonstrate that amino acid specific patterns of codon usage are themselves quite variable between branches of the tree of life, and that some of this variability correlates with organismal tRNA content.

**Keywords:** stochastic thermodynamics, codon usage bias, fungi, protists, bacteria

## Introduction

Codon Usage Bias (CUB), the preferential use of some types of codon over others encoding the same amino acid, is an established phenomenon. How CUB has evolved, and in particular which evolutionary forces drive its evolution, is still being actively researched. A particular recent focus of research has been on the question whether evolution is linked to translational selection [1–4] and if so, whether selection is more likely linked to protein quantity or quality [5].

Studying CUB evolution usually involves the quantification of codon usage in some way. Early attempts to do so relied on heuristic measures such as the frequency of optimal codons $F_{opt}$ [6], the codon bias index $CBI$ [7], and the codon adaptation index $CAI$ [8]. While such measures can efficiently catalogue the codon usage preferences of organisms, they make a number of assumptions that may influence their performance. For example, all three indices rely on defining optimal codons from the outset, usually by detecting preferred codons in a set of highly expressed genes. This is challenging for organisms that are not well studied experimentally, and even in well-studied organisms $CAI$, $CBI$ and $F_{opt}$ values are quantitatively dependent on the exact choice of the highly expressed gene set.

The more recently introduced tRNA adaptation index ($tAI$) [2] evaluates the tRNA pool available to decode a codon rather than the codon frequency itself, and does not require defining a set of optimal codons. However, since tRNAs have only recently become amenable to genome-wide quantification methods, $tAI$-based approaches usually approximate tRNA abundance with tRNA gene copy numbers. This is a reasonable

approximation for organisms like *S. cerevisiae* and *E.coli* where a close correlation between the gene copy number and the tRNA abundance has been demonstrated [9], although gene-copy number based tAI analyses have also been applied to humans where this correlation is less strong or absent [10]. For most organisms the relationship between gene copy numbers and tRNA abundance is uncertain. It is also noteworthy that the $tAI$ and $CAI$ evaluate gene-specific adaptation by calculating the geometric mean of codon-specific values, which strongly emphasizes the importance of poorly adapted codons in these indices. Thus, all heuristic indices rely on a number of assumptions and parameter choices. It often remains unclear to what extent these choices bias the results one obtains.

In addition to heuristic measures, model-based analyses have been employed to test particular aspects of CUB. An early approach [11] termed the "Effective Number of Codons" ($EN_c$) essentially performs a statistical test against the null hypothesis that codon usage is solely governed by genomic GC content. Based on combined analyses of $EN_c$ and $tAI$, dos Reis *et al.* [2] concluded that translational selection is prevalent in many microbial genomes, but largely absent in metazoans. A more recent approach is based on the assumption that non-preferred codons carry a cost in terms of energy usage that must be minimised for highly expressed genes [12, 13]. Using a Bayesian framework, these models were used to make impressively detailed predictions of gene expression levels in baker's yeast from codon usage data alone [1]. Other models based around the dynamics of mutation, selection and genetic drift in populations have similarly explored aspects of codon usage evolution [3, 4]. From these studies, an overall picture emerges where CUB is driven by a balance between mutation bias and translational selection, where mutation bias is the dominant force for low expressed genes and translational selection the dominant force for highly expressed genes. While this is the most frequently stated hypothesis, it is noteworthy that this is not universally accepted, and alternative or complementary hypotheses continue to be put forward [14–19].

With the explosion in genome sequencing data, there is an emerging scope for studying CUB and sequence evolution at a scale much broader than ever before. However, for organisms where data about tRNA populations and other parameters is not available, which is the case for the vast majority of genome sequences, applying the established approaches can be problematic. Moreover, for organisms with extreme lifestyles such as parasites [20–22] or thermophiles, codon evolution may follow fundamentally different rules. We therefore sought to develop an approach for modelling codon usage bias that relies as little as possible on initial assumptions or parameter sets, and which can easily be applied to the wealth of emerging genomic information from poorly characterised organisms.

In this contribution, we address this problem by taking a fresh view on the quantification of CUB. To do so, we draw on ideas from statistical physics and stochastic processes. Specifically, we conceptualise observed codon usage patterns as the result of multiple concurrent random walks in the space of synonymous codons. The codon usage in each gene is then a snapshot of the current state of one random walker. The genome as a whole provides statistics over a larger number of random walkers, which can then reveal underlying forces acting on codon usage, or equivalently, forces that bias the random walk of the individual walkers. We use this idea to develop a novel quantification method — the distance measure $\mathcal{D}$ — that measures the amount of bias in a particular genome, expressed as a distance from a hypothetical genome that does not have any systematic codon usage bias. The calculation of this distance does not rely on any *a priori* choice of a reference set.

Having established the distance measure, we then apply $\mathcal{D}$ to a large number of genomes from across 2 microbial kingdoms of life. We show that $\mathcal{D}$ encapsulates both the genome-wide aggregate codon usage bias and sequence level (local) selection resulting in re-arrangements of codons with no aggregate effect at the level of the genome. As far as we are aware no other measure is able to capture this. The quantification of sequence level selection reveals that this is a strong driver of codon usage bias in most organisms, but that there are interesting exceptions in particular taxa within the fungal kingdom. Furthermore, our analysis shows that commonly used model organisms are near the median for values of $\mathcal{D}$, but at all taxonomic classes contain strong outliers where codon usage appears to be shaped by very different types of forces. Finally, amino acid specific patterns of codon usage are themselves quite variable between branches of the tree of life, and some of this variability correlates with the tRNA content.

| 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 $g$ (length of subsequence). |

Tab. 1: Explanation of the symbols used.

## Materials and Methods

## The dataset

All datasets were obtained from ENSEMBL `https://www.ensembl.org`. We downloaded coding sequences for 462 species from the fungal kingdom (release 36 in August 2017), and 442 randomly chosen species from the bacterial kingdom (release 40 in July 2018). Bacterial species were chosen to roughly size-match the fungal data set, while proportionally representing the taxonomic spread of sequenced bacterial genomes. All species names and corresponding download weblinks are in supplementary file "species.xlsx". We then produced clean sequence files by converting each gene sequence into a valid codon sequence and removing those genes where the number of nucleotides was not a multiple of 3 (indicating errors in the ORF annotation). This led to the exclusion of 35748 genes from 4554328 total genes in the fungal dataset, and of 6384 genes from 1286467 in the bacterial dataset.

Each gene in the clean sequence files was then split 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 subsequence thus produced, a control coding subsequence was generated 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.

For the results presented in the main text we limited the analysis to the 9 amino acids with only 2 codons because the calculation of $\mathcal{D}$ becomes statistically unreliable when subsequence sets are too small, and this can generate problem in the analysis of amino acids encoded by more than two codons. The number of *possible* subsequences grows quickly with the number of codon possibilities, whereas the *actual* number of subsequences represented in a genome does not. As a consequence, there are fewer examples per configuration which increases the statistical error.

The parameters of the model were obtained as follows:

1. For each species we split up each gene into 18 subsequences (or 9 respectively when we focussed on the 2 codon amino acids).

2. We then discarded all subsequences with more than 15 or fewer than 5 codons.

3. For each species, we then obtained a parameter for each length, resulting in 11 different fits per amino acid.

4. Species-wide distances were calculated by taking occurrence weighted averages over the individual fits.

Fitting was done using the Maple 2018 "NonlinearFit" function. The initial estimates for $\gamma$ and $\xi$ 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 $< 0.00099999$.

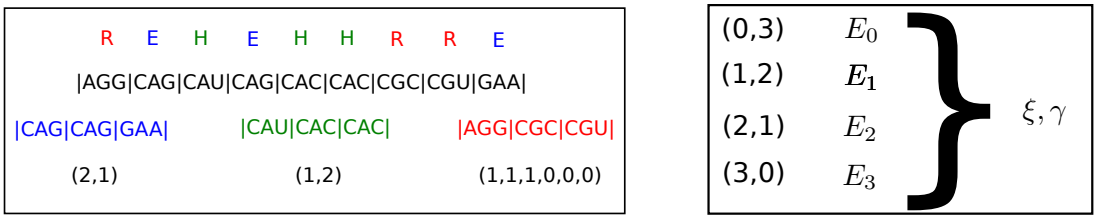


Fig. 1: (a) For each species, each mRNA is split into up to 18 subsequences, i.e. one subsequence for each amino acid with a codon choice greater than one. Here, we show the example of a hypothetical mRNA coding for `REHEHHRRRE`. This mRNA is split in three subsequences, shown in the third line. In the model the codon frequencies stated in the last line are then used. The "energy" of a subsequence is then calculated based on the number of codons of each type in each subsequence, as given in the last line. (b) For each amino acid of an organism, the subsequence configurations for each length $L$ are collected; here an example is shown for $L = 3$. Each configuration the frequency of configurations is calculated and fitted to a Boltzmann distribution based on eq. 2 to obtain $\xi$ and $\gamma$. This in turn is used to calculate the distance $\mathcal{D}$ according to the formula.

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.

**Definition of ambiguous region** Fig. 2c shows the residuals obtained from the full model versus the residuals obtained from the binomial model. If the former is low and the latter is high then the corresponding subsequence is better explained by SLS than the beanbag model. However, there is an ambiguous region where the beanbag model fits equally well as the full model (or better). There are many ways to define such a region, and we chose the following conservative approach: We first fitted a straight line to the corresponding plot of the control data in fig 2b. Then we decided that all points in the plot for the full model in fig. 2c that lie below this line are ambiguous.

## Results

We start this section by describing the theoretical model and how it was obtained. Readers who are less interested in the mathematical details of the model can, for their first reading, skip most of the next two subsections. They are encouraged though to read the first paragraph of the next subsection in connection with fig. 1, which introduces the basic notation used throughout. They should also take note of the eq. 2 which introduces the full model.

## Codon selection as a random walk

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 of the amino acids that is encoded by more than 1 codon. For the purpose of this article, we will exclusively focus on synonymous mutations, whereby a codon $a$ encoding an amino acid gets exchanged for a codon $a'$ that encodes the same amino-acid. We can then conceptualise each gene sequence as consisting of (up to) 18 independent *subsequences*, where each subsequence is a string of codons encoding the same amino acid $\mathcal{A}$ appearing in a gene $g$, where $\mathcal{A} \in \{\mathtt{E,H,Q,F,Y,C,N,K,D,I,P,T,A,V,G,L,S,R}\}$. For our random walk model, we will consider each of the subsequences as an independent random walker. Throughout this manuscript, we denote the number of codons for amino acid $\mathcal{A}$ in gene g by $L^{\mathcal{A},g}$. Each subsequence consists of $k_1$ codons of type 1, $k_2$ codons of type 2, ..., $k_{|C^{\mathcal{A}}|}$ 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 arbitrarily

assigned codons to codon 1, codon 2 and so on. This assignment remained fixed for all species we analysed. See table 1 for a summary of the symbols used.

We can now consider each possible configuration of a subsequence $\{k_1, \ldots, k_{|C^{\mathcal{A}}|}\}$ as a state of a random walker; see fig. 1 for a graphical explanation of what we mean by "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; see SI sec. 1 for more detail.

Throughout this contribution, we make a number of simplifying assumptions about the nature of the random walk. Firstly, we do not consider non-synonymous mutations, 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 arise from biasing effects that operate on the genome. These could be, for example, mutation biases, GC-bias or evolutionary selection pressures (including effects of random drift). A subtlety arises here in the context of 6 synonymous codons where the mutation rates between synonymous codons are not equal. This does not affect the discussion here, however, because we will focus mostly on amino acids with 2 synonymous codons. Thirdly, we assume throughout that 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 [23]. We conceptualise each subsequence configuration $i$ as having an energy $E_i$, where $E_i$ depends on the codon composition of the subsequence, which is the result of some a priori unknown "forces," likely including mutation bias, selection and GC contents. To calculate this energy, we arbitrarily assign an energy of 0 to subsequences that are in the most likely configurations, i.e. those where all codons appear equally often. Then we postulate that the local detail balance condition is fulfilled [24]:

$$T \ln\left(\frac{r_+}{r_-}\right) = \Delta E \tag{1}$$

Here $r_\pm$ are the (unknown) rates of the random walker to transition between a state with energy $E_i$ and $E_{i'}$, with $\Delta E := E_i - E_{i'}$. Note that $r_\pm$ will be different for different pairs of $i$ and $i'$; for notational simplicity we suppress this dependence in the equation above, but this will become important below. The key point is that this equation allows us to calculate energies of individual states, if the rates $r_\pm$ are known.

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 gas 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.

## Deriving the full model

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 the Boltzmann distributions of random walkers/sequences, which can be compared to data. We will base our reasoning on two specific selection scenarios. We refer to the first of these as *beanbag selection*, i.e. the number of codons of a type is determined by a fixed probability (analogous to selecting different coloured beans from a bag containing the different colours in a given proportion). We refer to the second as *sequence level selection* (SLS), i.e. the codon composition is determined by rules that differ between different subsets of sequences. These scenarios have biological equivalents: for example, beanbag selection includes mechanisms such as genome-wide mutation biases, whereas sequence level selection includes mechanisms such as translational selection, in the models proposed by Gilchrist and colleagues [1, 12, 13].

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 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 models a scenario where the codon evolution is driven by unbiased random mutations.

This simplest model can be readily expanded to include a global codon usage *bias* $q$, yielding an energy $\hat{E}_i = E_i + \ln\left((1-q)^i/q^i\right)$ (for the calculation see SI sec. 1). The resulting Boltzmann distribution coincides again with the binomial distribution, but this time with a bias $p = q$.

The two preceding scenarios represent beanbag models, because they assume that 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 $k_1^\xi$ 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 in that the resulting statistics can no longer be reproduced by throwing dice or tossing coins, not even unfair ones. Instead, this model entails sequence level selection. In the SI sec. 1 we show that in this scenario the energy for a subsequence with $i$ codons of type 1 is given by the *full model*:

$$\bar{E}_i = \xi E_i + T(\gamma - \xi)\ln(i!). \tag{2}$$

This energy model has two parts that lend themselves to direct interpretation. The first term on the left hand side is an "entropic" part that characterises codon usage in a no-selection scenario. This means that the first term does not lead to a global codon usage bias. The second term is an effective "selection potential." It encapsulates forces on the genome that alter the overall frequency of codons and thus 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.

Given the full model eq. 2, a Boltzmann distribution with parameters $\gamma$ and $\xi$ 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, for amino acids where the codon choice $C^{\mathcal{A}} = 2$.

Before proceeding, we discuss some special choices for the ad-hoc parameters $\xi, \gamma$ of the full model, 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 unbiased beanbag model with a modified inverse temperature $\xi$. In this case there will be no selection pressure on the global usage of codons, but there may be sequence level selection, affecting how codons are distributed across subsequences. For $\xi = \gamma = 1$ the full model 2 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.

Finally, we note that the full model eq. 2 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 truly binomial.

## Genomic data bear the signature of sequence level selection

We first address whether there is evidence for sequence level selection 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 [25] and calculated the energies for each of the subsequences contained in this database for both the beanbag model and the full model. If the codon distribution was explained by beanbag selection, then we would expect that the energies are distributed multinomially, which would reduce to a binomial distribution in the case of amino acids with 2 codons only. On the other hand, if the genomes are subject to SLS then the full model would be a better description of the data.

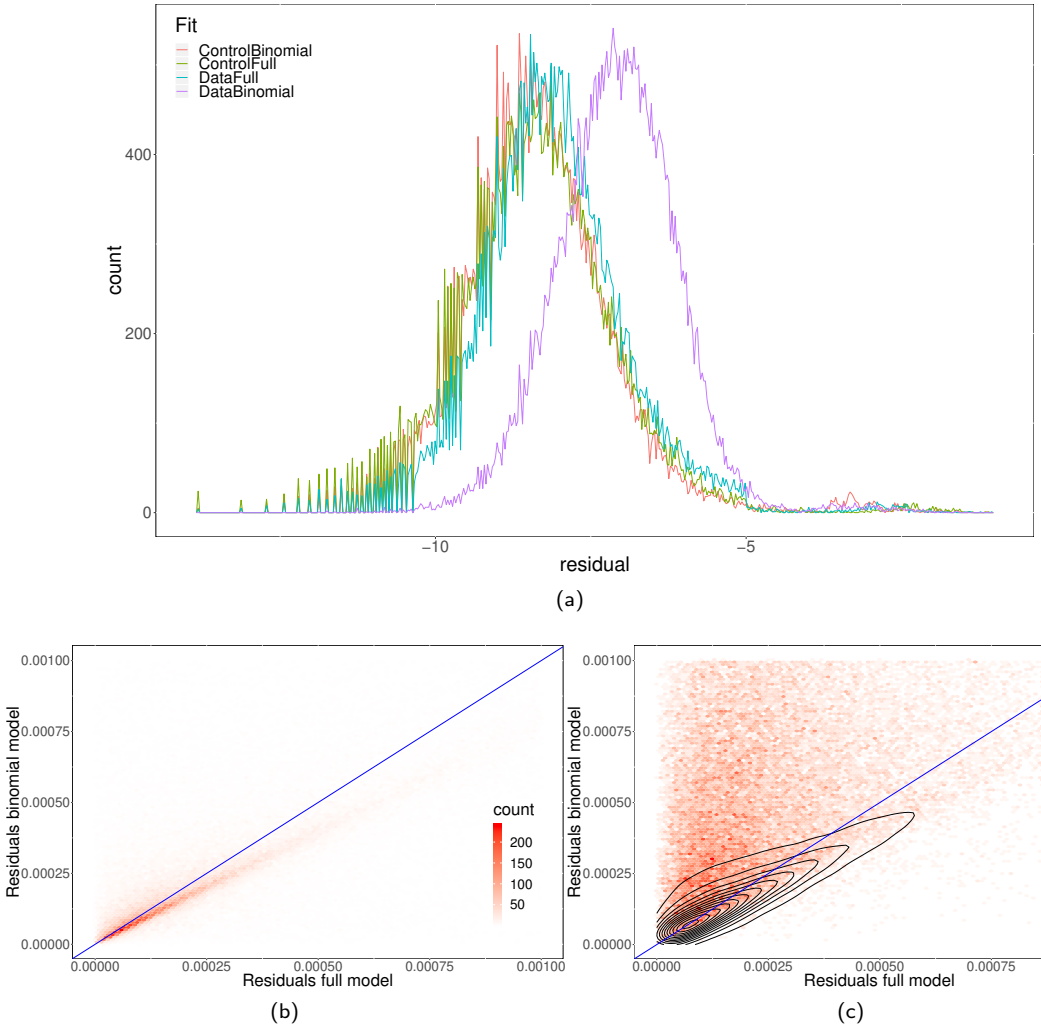


Fig. 2: (a) Histogram for the mean-residuals obtained from fitting the binomial distribution and the full model for both the real data and control data generated under the beanbag model assumption, i.e. where codons have been replaced by random synonymous codons with a bias corresponding to the global 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.

To check this, we fitted the frequencies of energies in our dataset to a binomial distribution and the full model. We limited ourselves to the nine amino acids with 2 codons, i.e. $|C^{\mathcal{A}}| = 2$, because only for those amino acids it is possible to obtain sufficiently powerful statistics. Furthermore, we only analysed subsequences with lengths $5 \leq L^{\mathcal{A},g} \leq 15$. Sequences with $L < 5$ provide too few datapoints to fit, and for sequences with $L > 15$ the statistical errors become overly large because the number of subsequences reduces quickly with increasing subsequence length. Altogether, we obtained 45702 fits of the fungal data set to the binomial model and the same number of fits to the full model. Similar results with fits to bacterial data are shown in the SI.

### Codon distribution can be fitted to the binomial distribution

In order to understand how well the genomic data fits the respective models we analysed the mean-residuals of the fits, which are indicators for how well subsequences fit the respective models. In the limit of an infinite number of samples that have been generated according to the fitting distribution, the mean-residual would vanish. When the sample size is finite, or when the data is not distributed according to the hypothesis distribution, then the mean-residuals take a positive value. The fits of the fungal data to the binomial distribution yielded mean-residuals between $\exp(-4)$ and $\exp(-9)$ peaking at about $\exp(-7)$; see fig. 2. Visual inspection of a number of examples suggests that these mean-residuals indicate a reasonably good fit to the data.

### The full model fits the fungal data better

Next, we fitted the full model eq. 2 to the same data. On the whole, this resulted in smaller mean-residuals, indicating that the full model is a better fit to the data than the binomial model. Fig. 2 summarises this quantitatively. 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. This suggests that the full model is a better description of the data than the binomial model.

The better fit of the full model could be merely a reflection of the fact that it has more parameters than the binomial model. To check this, we prepared a control set of distributions. This control set consists of the same subsequences 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, and hence the fitting error of this synthetic dataset is only due to the statistical error, i.e. a consequence of the finite (and indeed small amount of) data. When we fitted both the full model and the binomial model to these control data, we obtain two distributions of 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. Interestingly, 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 worse (i.e. shifted to the right) compared to the fit to the synthetic data. *Prima facia* this means that almost all the error of the full model fit is due to statistical error, which in turn leads to the conclusion that the full model captures almost all of the variation in the underlying real data. From the above analyses it is not clear whether the full model is a better fit for all subsequences, or whether it merely fits the majority better while there still are many subsequences that are equally well described by a beanbag model. In order to investigate this, we compared residuals of fits to the binomial and full models for individual subsequences directly; see fig. 2. It is instructive to consider control data first, which by construction follows the beanbag model. As expected, we found that most subsequences are approximately equally well fitted by the binomial and control data (see fig. 2b), 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. In contrast, for the real data the same analysis leads to a high density of points in the upper left corner of the figure, where the mean-residuals of the full model are lower than those of the binomial model; see fig. 2c. The majority of subsequences are thus better explained by sequence-level selection. However, there remains a significant minority of subsequences (less than 20%) that can be equally well explained by the beanbag model and sequence level selection.

We sought to explore the origin of the residuals showing equally good fits for beanbag and sequence level selection further. In order to produce the fits for figure 2c we split gene sequences into the subsequence for

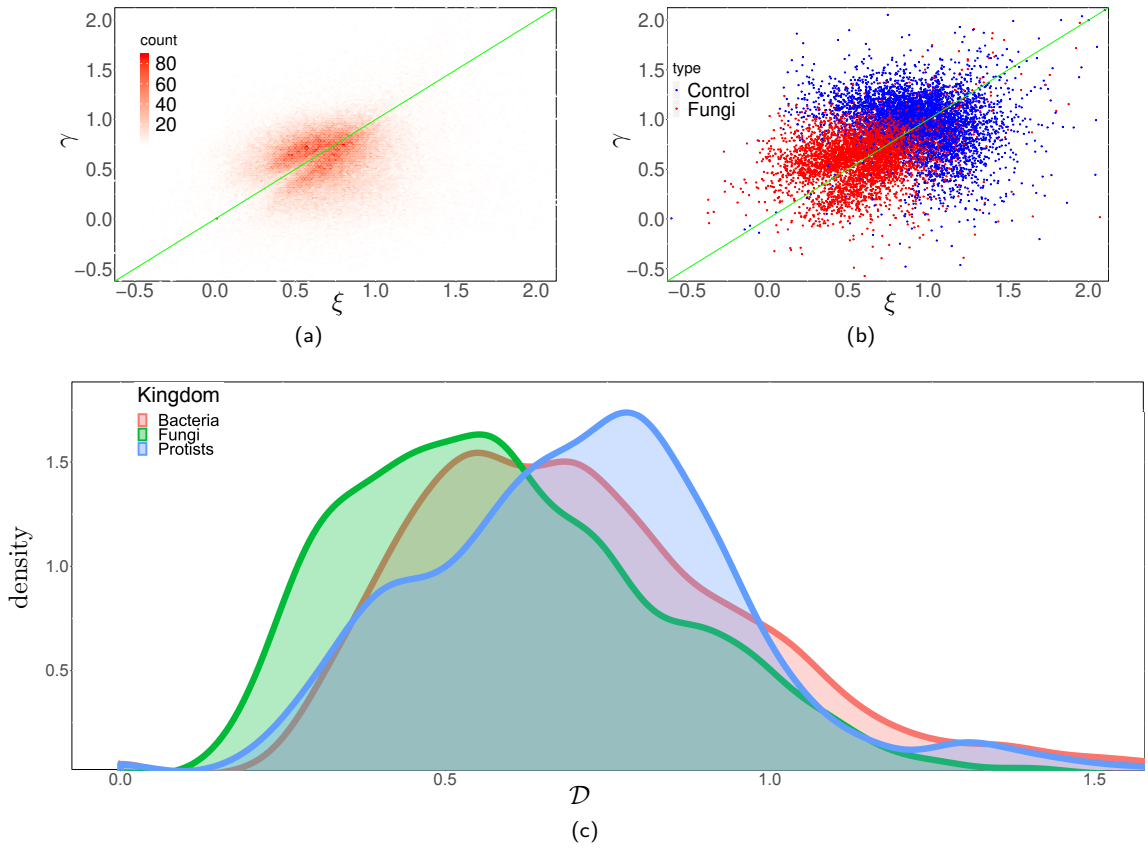


Fig. 3: (a) The density of fitted parameters $\xi$ and $\gamma$ for each of the 2-codon amino acids for all 462 fungal species in our dataset. We are limiting ourselves to fits with mean-residuals $< 0.0009999$. The fitted values largely concentrate into the interval of $[0, 1.5]$. (b) Comparing the 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 the values of $\mathcal{D}$ in bacteria, protists and fungi. The plot shows a smoothed distribution of the species averaged distances in the three kingdoms. Clearly, the distances in fungi tends to be the lowest.

each of the nine amino acids with a codon choice of 2, and then grouped all sequences of equal length in the range 5-15 as explained above. Thus, any individual organism is represented with up to $9{\cdot}11 = 99$ subsequence groups in the dataset, each corresponding to one residual in the figure. If we make the assumption that some subsequences fit a sequence level selection model and a beanbag model equally well purely by chance, we would expect that all organisms contribute data points to this ambiguous region of the plot (see Materials and Methods how we defined this region) with equal probability. We can then calculate how likely the contribution of a particular number of subsequences by an organism is.

Interestingly, the contribution of data points to the ambiguous region is very clearly not governed by chance (fig. 4). Instead, we observe that most organisms completely avoid this region, with the majority of organisms represented with seven data points or fewer (by chance, this should only occur for 0.3% of organisms). On the other hand, a minority of organisms contributes the majority of ambiguous subsequence groups, with the highest individual contribution being 51 subsequences (which should occur by chance with a frequency of less than $10^{-14}$). Thus, while in most organisms sequence level selection is a strong driver of codon usage bias that shapes the majority of sequences, in some organisms sequence level selection is weak or absent.

In the taxonomic tree, organisms where sequence level selection makes apparently weaker contribution to codon usage bias are clearly grouped (fig. 4b). Two larger groups of such organisms are indicated as "Islands" in the figure, and are located within the taxonomic classes *Agaricomycetes* (A) and *Eurotiomycetes* (E). On the other hand, species with higher than expected ambiguous subsequences are completely absent from the large *Sordariomycetes* class (S). Thus, absence of clearly detectable sequence level selection appears to be a trait that has evolved in certain taxonomic groups and which is likely linked to the organisms' lifestyle or other biological traits, although we do not currently understand the exact mechanisms by which the effect of sequence level selection on codon usage could become reduced.

Overall, these observations confirm that codon usage bias is not solely shaped by genome-wide forces, but that sequence-level selection makes substantial contributions e.g. via mechanisms linked to translational regulation [1], gene length [26, 27] or other gene specific parameters. Going beyond the current state of the art, we have established here that the distribution of codons can be described compactly by a surprisingly simple distribution, i.e. eq. 2, which can be derived from rather general principle within a statistical physics framework.

## Defining distance

For the fungal and bacterial genomes we examined, typical values of the fitted parameters $\gamma$ and $\xi$ are small and positive with 96.39% of fits resulting in $0 < \gamma, \xi < 2$. These fitting parameters do not have an immediate biological interpretation, but interesting insights can be gained by considering the distribution of sequences in $\xi, \gamma$ space. Fig. 3b reveals that for fits to fungal genomic data, $\xi$ and $\gamma$ are concentrated in a smaller part of the parameter space relative to fits to simulated genomic data based on beanbag selection only. This confirms that *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.

We define now the (Euclidean) *distance* of a genome in $\xi, \gamma$ space from a (hypothetical) unbiased, entirely random sequence located at $\xi = \gamma = 1$.

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

A vanishing distance $\mathcal{D} = 0$ would indicate that no codon usage bias whatsoever is present in the genome as a whole. However, note that individual genes may still have measurable biases in codon usage but the biases would be distributed according to the binomial distribution. When $\mathcal{D} > 0$ then this indicates that a CUB is present. The distance measure captures two types of biases: $(i)$ a statistical global over-representation of the codon frequencies, and $(ii)$ a deviation from the beanbag assumption. Either of those would be sufficient to lead to a non-vanishing $\mathcal{D}$. As such $\mathcal{D}$ is useful as a high level descriptor of codon usage that provides a quantification of the codon usage bias complementary to comparable existing measures like $F_{opt}$ or $CAI$. In particular, in addition to reflecting the absolute magnitude of codon usage bias, $\mathcal{D}$ increases with the diversity of forces that shape codon usage in a set of sequences, such as mutation bias, GC content, or selection. No assumptions are made here about the specific nature of these forces.

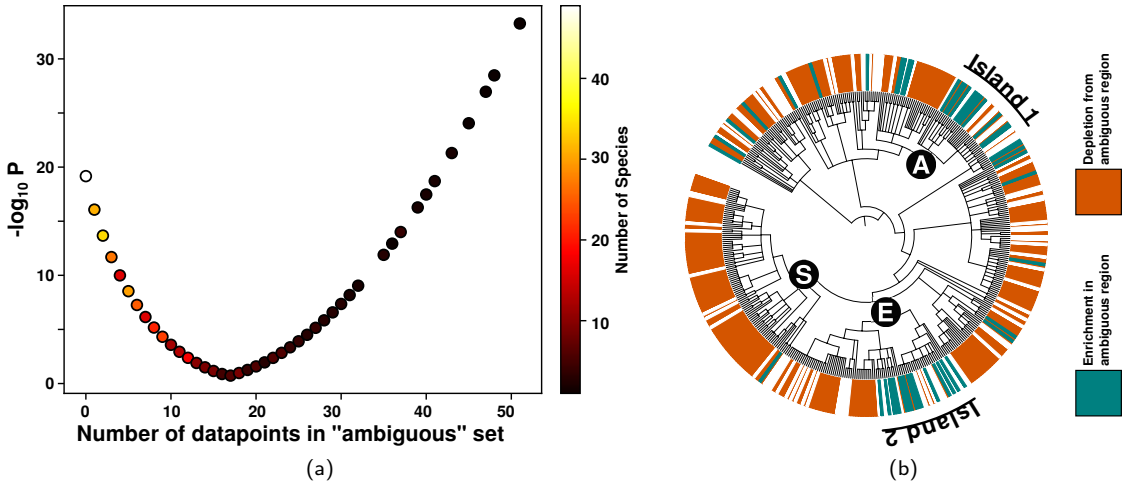


Fig. 4: (a) A "Volcano"-plot showing the distribution of points into the ambiguous versus general population classes. A single species could contribute up to 99 data points to the "ambiguous" region highlighted in figure 2c, where it is not possible to distinguish whether sequence level or beanbag selection is more likely for a set of subsequences. "$-log_{10}P$" quantifies the probability of obtaining an observed number of subsequences in the ambiguous region by chance. (b) A taxonomic tree of the species in our fungal dataset, colour coded according to whether the species is statistically under-represented or over-represented in the ambiguous group relative to chance (based on a cut-off value of $P = 10^{-3}$). Most species are under-represented indicating that sequence level selection dominates significantly. However, in the two identified "Islands", for many species sequence level selection does not contribute to shaping codon usage with statistical significance according to our test.

We illustrate this in figure 5, which displays $\mathcal{D}$ and $F_{opt}$ (the proportion of optimal codons) as a function of genome complexity. The four analysed genomes are all based on the *S. cerevisiae* genome and contain the same number of sequences encoding identical proteomes. The first genome was generated by replacing all codons with randomly chosen synonymous ones, i.e. this genome has no codon bias at all. For the second genome, codons were again replaced randomly but now with a probability proportional to the observed average codon usage in the yeast genome: this genome shows the same average codon bias as the real genome, but here the bias is generated by a single selective force which applies genome-wide. For the third genome, codons were replaced by applying a mixture of two different replacement schemes, with the actual mixture chosen for each gene as a function of its protein expression levels as reported in [28]. This procedure simulates a genome obeying the proposed balance between mutation bias and translation selection underpinning the models described by Gilchrist *et al.*, and the parameters chosen for the replacement were selected to recapitulate the data in [1] as closely as possible. The fourth genome is the actual *S. cerevisiae* genome. For these four genomes, we show the distribution of $F_{opt}$ values for each sequence, as well as amino acid-specific $F_{opt}$ values averaged over all sequences, and amino acid-specific $\mathcal{D}$-values (note that, because $\mathcal{D}$ is estimated based on comparing distributions, it is not possible to estimate this for individual sequences). Only values for amino acids encoded by two codons are shown in the latter two cases.

A salient property of $\mathcal{D}$ revealed by this analysis is its gradual increase with increasing complexity of the forces that shape codon usage (c.f. the shift of the mean $\mathcal{D}$ from top to bottom). In contrast, because $F_{opt}$ simply summarises average sequence properties, its behaviour does not reflect genome complexity and the mean $F_{opt}$ value shows a non-linear relationship with codon usage bias when complex selective forces shape this bias. Although $CAI$ and $CBI$ are calculated differently, they correlate with $F_{opt}$ and would show a similar, non-linear relationship. Interestingly, there is a notable increase in $\mathcal{D}$ between the complex simulated and the actual yeast genome, indicating that the actual yeast genome is shaped by a more complex array of forces than the mutation bias/ translational selection model alone accounts for.

One motivation for proposing $\mathcal{D}$ as a useful measure for characterising codon usage in fungal genomes was the fact that this measure does not require any context information other than the nature of the coding

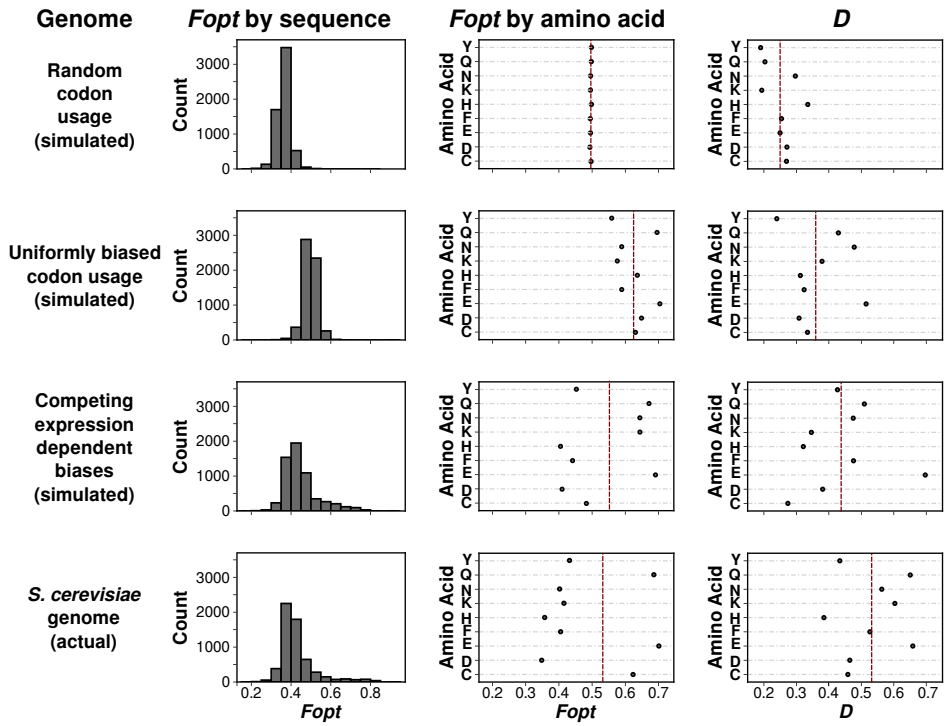


Fig. 5: Comparing $\mathcal{D}$ to $F_{opt}$. The top three rows analyse simulated genomes encoding identical proteomes as the actual *S. cerevisiae* genome, which is analysed in the bottom row. Columns display $F_{opt}$ averaged over all amino acids for each sequence (left), $F_{opt}$ averaged over all sequences for each amino acid with a codon choice $C^{\mathcal{A}} = 2$ (centre), and $\mathcal{D}$ averaged over all sequences for each amino acid with $C^{\mathcal{A}=2}$ (right). Red lines indicate average values for all analysed amino acids.

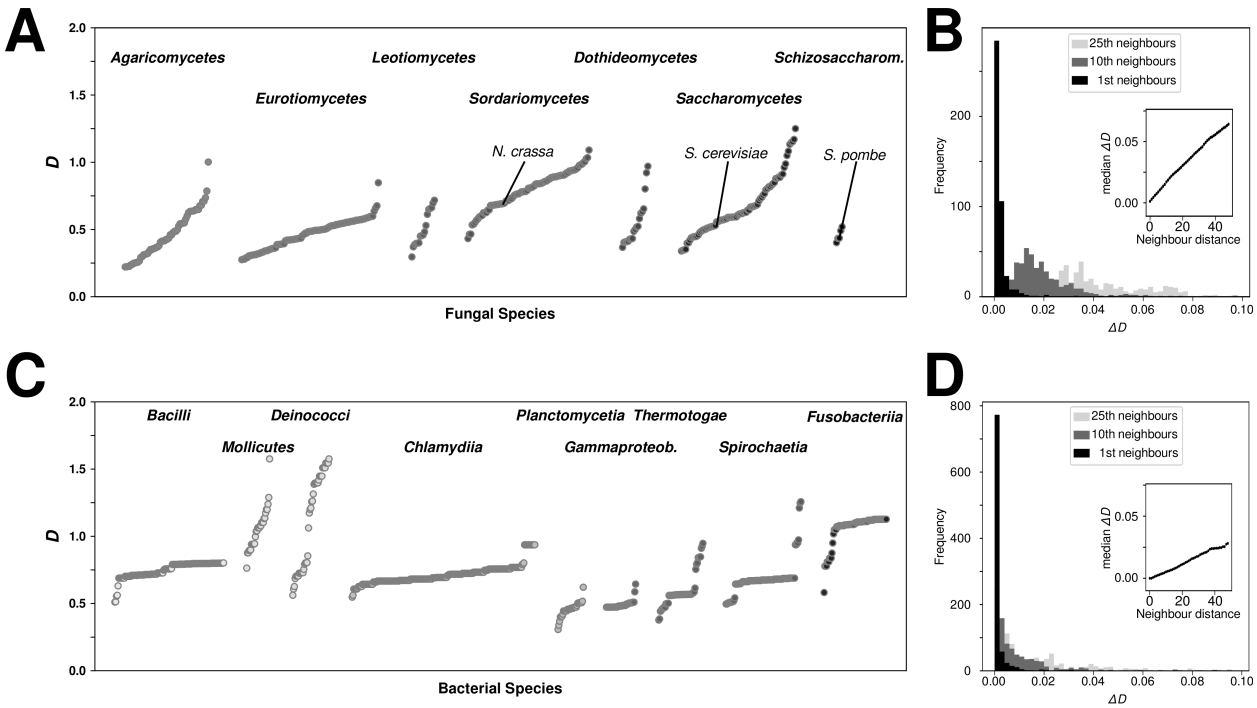


Fig. 6: Genome-wide $\mathcal{D}$ for fungal (top) and selected bacterial (bottom) species. A, Genome-wide $\mathcal{D}$ for selected fungal taxonomic classes with multiple sequenced genomes. B, the difference in $\mathcal{D}$ between species increases with taxonomic distance. Detailed histograms show values for $\Delta\mathcal{D}$ for species pairs that are immediate taxonomic neighbours (black),or that are separated by 10 (dark grey) or 25 (light grey) species. Inset, median $\Delta\mathcal{D}$ as a function of neighbour distance. C and D, as in A and B but for bacterial species.

sequences themselves, and is thus applicable to any genome irrespective of the degree of knowledge available on the corresponding organism. We calculated the average $\mathcal{D}$ for each of the many hundreds of species with genome information in the ENSEMBL Fungi database [25], as well as for a size-matched selection of bacterial genomes. Fig. 6 shows the resulting $\mathcal{D}$ values, grouped by taxonomic class of the species. This analysis highlights a number of features that are not apparent from existing analyses. Within each taxonomic class there is significant diversity in $\mathcal{D}$ values, indicating that the diversity and nature of forces that shape codon usage varies widely even within taxonomic classes. This trend is particularly noticeable within the fungi whereas most bacterial classes display more uniform $\mathcal{D}$ values (although there are exceptions, e.g. within the *Mollicutes* and *Deinococci*). Despite the high degree of variability within classes, $\mathcal{D}$ clearly arises from heritable traits, since the average difference in $\mathcal{D}$ between two species increases proportionally with the taxonomic distance between these species (fig. 6 B,D). Interestingly, the relationship between $\Delta\mathcal{D}$ and taxonomic distance is quantitatively different between fungi and bacteria (compare the slopes in the inset graphs in Fig. 6.)

### $\mathcal{D}$ 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. 7 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 define 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. 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, whereas in other groups F either behaves like other amino acids or shows atypically weak selection (eg in the *Agaricomycetes*). The fact that codon bias differs for different amino acids in any one organism has been observed before and is both apparent from inspecting average codon usage tables, and predicted from current models of codon usage bias evolution such as ref. [1]. A novel insight our analyses add is how strongly the direction of these differences can vary between taxonomic groups. Because codons are decoded by tRNAs, an immediate hypothesis could be that atypical $\mathcal{D}$ values reflect unusual features of the tRNA population. To explore this assumption, we identified tRNA genes in all analysed fungal genomes using tRNAscan-SE software [29], and then used a regression approach to explore in how far tRNA gene copy number could explain atypical $\mathcal{D}$ values for particular amino acids. The lasso regression approach chosen for this analysis [30] allows estimating both the predictive power of a dataset for particular target variables, and the importance of individual dataset features for the prediction.

The results of this analysis show that overall, the tRNA pool has variable predictive power over amino acid specific $\mathcal{D}$ (Fig. 8). For some amino acids tRNA gene copy numbers make small but appreciable contributions to explaining $\mathcal{D}$, up to $R^2$ values of 0.26 for lysine (K) and 0.25 for glutamine (Q). On the other hand, tRNA gene copy numbers cannot predict $\mathcal{D}$ for histidine (H) at all. Further investigation of the amino acid where tRNA gene copy numbers can most strongly explain $\mathcal{D}$ (lysine) showed that in this case 80% of the explanatory power lies with the gene copy number of two glutamic acid inserting tRNAs, tE(CUC) and tE(UUC) (Fig. 8). Both tRNAs can form two Watson-Crick base pairs with one or other of the lysine codons, which makes them strong candidates to be near-cognate tRNAs for these codons (ie tRNAs which inhibit efficient decoding, [31, 32]).

In sum, these data indicate that in some of the cases where atypical selection is observed for individual amino acids, this is in part in response to the particular tRNA pools in these organisms. However, currently uncharacterised influences are also at play, and for some amino acids atypical codon selection is entirely caused by tRNA-independent forces.

## Discussion

The model we propose heavily draws upon ideas from statistical mechanics and especially stochastic thermodynamics [23, 33]. Originally, statistical physics was developed to describe the properties of gasses as a

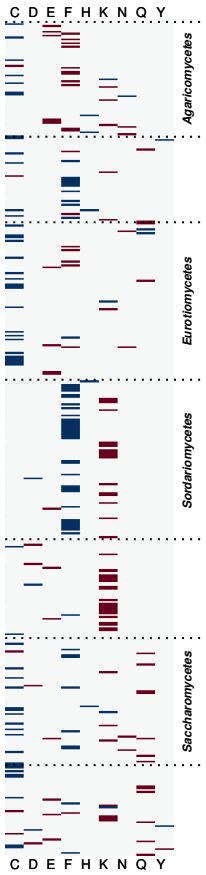


Fig. 7: Amino acid-specific patterns of codon usage bias in fungal genomes. Average $\mathcal{D}$ values were calculated for all subsequences for each amino acid in each genome. Amino acids are highlighted if their $\mathcal{D}$ value was more than $2\sigma$ above (blue) or below (red) the median $\mathcal{D}$ for that species, i.e. if they 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.

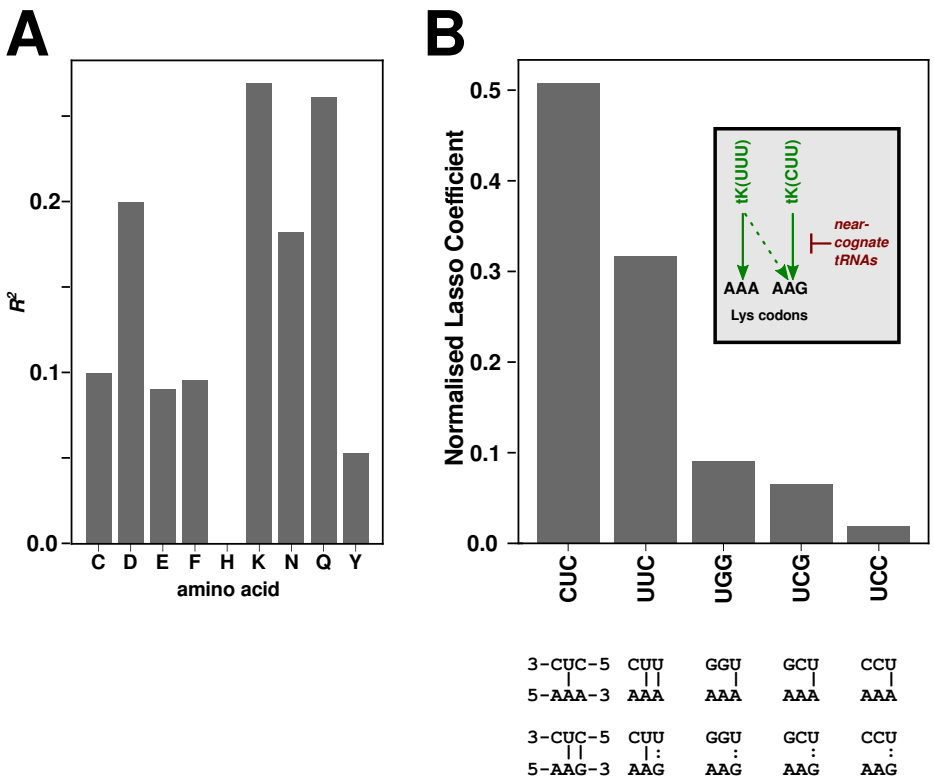


Fig. 8: tRNA gene copy number can partially explain atypical $\mathcal{D}$ values. A, $R^2$ values for lasso regression using tRNA gene copy numbers to explain $\Delta\mathcal{D}$ values (the difference between $\mathcal{D}$ for an individual amino acid and the average $\mathcal{D}$ for all amino acids). $R^2$ is proportional to the predictive power of tRNA gene copy number to explain atypical $\mathcal{D}$. B, importance of individual tRNAs for predicting atypical $\mathcal{D}$ for lysine (K) codons. The base-pairing potential of these tRNAs with the two lysine codons is indicated below this panel (| , Watson Crick base-pair; : ,wobble base-pair)

function of their constitutive particles, but has since been extended well beyond its original scope. One of the basic features of statistical mechanics is that it can derive useful and universal laws from macroscopic descriptions of systems containing many intractable microscopic interactions. Previous examples of application in biology include scaling laws [34], spatial genome structures [35] and evolutionary processes [36,37] to name but a few. Common to all applications of statistical mechanics is that a large number of difficult, or indeed intractable behaviours can, when considered at the right level, lead to simple macroscopic behaviours. Precisely this we see in our model as well. It relies on only two free parameters and does not postulate any specific forces. Instead, it describes the aggregate effect of a putatively large number of simultaneously acting evolutionary forces on the genome.

From its origins in statistical mechanics, our model inherits a number of desirable features. Firstly, its mathematical formulation eq. 2 is both compact (in that it relies on two parameters only) and remarkably accurate. Indeed, we found that the error from fitting the model to the data model was mainly statistical error, i.e. a consequence of the limited amount of data that is fitted.

Secondly, its only assumption is that genomes are the result of a random walk and that they are in steady state. This is in contrast to previous models of codon-specific evolutionary selection that typically assume a small number of specific drivers of selection, compared to the many potential drivers that have been proposed in the literature. This agnosticism with respect to the origins of CUB makes our model also more easily applicable. For example, it does not require an *a priori* reference set.

Thirdly, it directly led to a novel measure of codon usage bias, i.e. the distance $\mathcal{D}$ that encapsulates both the global codon usage bias due to global frequency imbalances ($\approx$ bias due to beanbag model) and the codon usage bias resulting from non-random distributions of codons across genes ($\approx$ bias due to SLS). Therefore a measure of $\mathcal{D} > 0$ could occur even if all synonymous codons are used equally often. We demonstrated that a higher $\mathcal{D}$ is related to the diversity of forces acting on codon usage; see fig. 5.

While $\mathcal{D}$ captures two aspects of CUB, a more detailed analysis makes it possible to disentangle these. From fig. 3a it is clear that the distribution of genomes in $\xi, \gamma$ space is very different from the distribution of hypothetical genomes that are only subject to selection by the beanbag model. Indeed, there is a sharp demarcating line in this space across which beanbag model genomes cannot be found. Complementary to this, in the SI we also compare directly the $\mathcal{D}$ values of actual genomes, with their randomised counterparts, whereby the former nearly always are larger, as expected. The difference between the two $\mathcal{D}$ values then quantifies how much of the selection pressure is due to SLS.

While $\mathcal{D}$ has the ability to explain codon selection in greater depth and does not require any contextual datasets or assumptions it is not suitable to analyse individual genes, or even small sets of genes. This is because it is based on fitting a model to a distribution of subsequences. This can only be done if the underlying data yields a sufficiently accurate statistics.

Our application of $\mathcal{D}$ to fungal genomes represented in the fungal section of the ENSEMBL database [25], as well as a size-matched selection of genomes from the bacterial section of the same database yielded interesting results. These show that the nature and diversity of selective forces acting on codon usage is surprisingly varied even in taxonomic classes within these groups, as indicated by the spread in $\mathcal{D}$ values (Fig. 6), even though $\mathcal{D}$ is clearly based on heritable traits. The fact that common model organisms tend to be located near the median of the range underlines that it may not be appropriate to simply port approaches for analysis of codon usage from such models to other organisms. For example, while models relying on the analysis of opposing forces of mutation bias and translational selection [1] appear to come relatively close to describing codon usage bias in baker's yeast ($\mathcal{D} = 0.50$), the diversity of forces acting on codon usage appears very different in other fungi of interest, including human pathogens like *Candida albicans* ($\mathcal{D} = 1.08$), *Candida glabrata* ($\mathcal{D} = 0.81$) or *Pneumocystis carinii* ($\mathcal{D} = 0.92$); and plant pests like *Verticillum* ($\mathcal{D} = 0.98$) and *Magnaporthe* ($\mathcal{D} = 0.86$).

## References

[1] Gilchrist MA, Chen WC, Shah P, Landerer CL, Zaretzki R. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. Genome Biology and Evolution. 2015;7(6):1559–1579. doi:10.1093/gbe/evv087.

[2] dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon us-

age bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Research. 2003;31(23):6976–6985.

[3] Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Research. 2005;33(4):1141–1153. doi:10.1093/nar/gki242.

[4] Ran W, Kristensen DM, Koonin EV. Coupling Between Protein Level Selection and Codon Usage Optimization in the Evolution of Bacteria and Archaea. mBio. 2014;5(2):e00956. doi:10.1128/mBio.00956-14.

[5] Ran W, Higgs PG. Contributions of Speed and Accuracy to Translational Selection in Bacteria. PLoS ONE. 2012;7(12):e51652. doi:10.1371/journal.pone.0051652.

[6] 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. Journal of Molecular Biology. 1981;151(3):389–409.

[7] Bennetzens JL, Hall BD. Codon Selection in Yeast. The Journal of Biological Chemistry. 1982;257(6):3026–3031.

[8] Sharp PM, Li WH. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. 1987;15(3):1281–1295.

[9] Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Molecular Biology and Evolution. 1985;2(1):13–34.

[10] Dittmar KA, Goodenbour JM, Pan T. Tissue-specific differences in human transfer RNA expression. PLoS Genetics. 2006;2(12):e221. doi:10.1371/journal.pgen.0020221.

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

[12] Shah P, Gilchrist MA. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(25):10231–10236. doi:10.1073/pnas.1016719108.

[13] Gilchrist MA. Combining Models of Protein Translation and Population Genetics to Predict Protein Production Rates from Codon Usage Patterns. Molecular Biology and Evolution. 2007;24(11):2362–2372. doi:10.1093/molbev/msm169.

[14] Biro JC. Does codon bias have an evolutionary origin? Theoretical Biology and Medical Modelling. 2008;5(1):16. doi:10.1186/1742-4682-5-16.

[15] Trotta E. Selection on codon bias in yeast: a transcriptional hypothesis. Nucleic Acids Research. 2013;41(20):9382–9395. doi:10.1093/nar/gkt740.

[16] Yannai A, Katz S, Hershberg R. The Codon Usage of Lowly Expressed Genes Is Subject to Natural Selection. Genome Biology and Evolution. 2018;10(5):1237–1246. doi:10.1093/gbe/evy084.

[17] Southworth J, Armitage P, Fallon B, Dawson H, Bryk J, Carr M. Patterns of Ancestral Animal Codon Usage Bias Revealed through Holozoan Protists. Molecular Biology and Evolution. 2018;35(10):2499–2511. doi:10.1093/molbev/msy157.

[18] Zhou Z, Dang Y, Zhou M, Yuan H, Liu Y. Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. eLife. 2018;7. doi:10.7554/eLife.33569.

[19] Jossé L, Singh T, von der Haar T. Experimental determination of codon usage-dependent selective pressure on high copy-number genes in Saccharomyces cerevisiae. bioRxiv. 2018;doi:10.1101/358259.

[20] Xiang H, Zhang R, Butler RR, Liu T, Zhang L, Pombert JF, et al. Comparative Analysis of Codon Usage Bias Patterns in Microsporidian Genomes. PLOS ONE. 2015;10(6):e0129223. doi:10.1371/journal.pone.0129223.

[21] Badet T, Peyraud R, Mbengue M, Navaud O, Derbyshire M, Oliver RP, et al. Codon optimization underpins generalist parasitism in fungi. eLife. 2017;6. doi:10.7554/eLife.22472.

[22] Sinha I, Woodrow CJ. Forces acting on codon bias in malaria parasites. Scientific Reports. 2018;8(1):15984. doi:10.1038/s41598-018-34404-9.

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

[24] Crooks G. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. Journal of Statistical Physics. 1998;90(5/6):1481–1487. doi:10.1023/a:1023208217925.

[25] Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, et al. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res. 2018;46(D1):D802–D808.

[26] Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci USA. 1999;96(8):4482–4487.

[27] Moriyama EN, Powell JR. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. 1998;26(13):3188–3193.

[28] Ho B, Baryshnikova A, Brown GW. Unification of Protein Abundance Datasets Yields a Quantitative *Saccharomyces cerevisiae* Proteome. Cell Systems. 2018;6(2):192–205.e3. doi:10.1016/j.cels.2017.12.004.

[29] Lowe TM, Eddy SR. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Research. 1997;25(5):955–964. doi:10.1093/nar/25.5.955.

[30] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

[31] Plant EP, Nguyen P, Russ JR, Pittman YR, Nguyen T, Quesinberry JT, et al. Differentiating between near- and non-cognate codons in Saccharomyces cerevisiae. PLoS ONE. 2007;2(6):e517.

[32] Tarrant D, von der Haar T. Synonymous codons, ribosome speed, and eukaryotic gene expression regulation. Cellular and Molecular Life Sciences. 2014;71(21):4195–4206.

[33] Parrondo J, Horowitz J, Sagawa T. Thermodynamics of information. Nature Physics. 2015;11(2):131–139. doi:10.1038/nphys3230.

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

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

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

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