Codon usage bias

Dominique Chu

School of Computing, University of Kent, CT2 7NF, Canterbury, UK
d.f.chu@kent.ac.uk

Abstract

We model codon usage bias

Keywords: biological computing, entropy, computational performance

1 Introduction

Hypotheses:

1. Selection is at the level of the individual codons only.
2. Selection is at the level of entire sequences.

Hypotheses about the origin of the data:

1. The data could have been created by an underlying bias on the codon distribution %
2. The data could have been created by an underlying mutantional bias.
3. There is no underlying bias; all selection happens only at the level of the sequence.
4. There is selection at both levels.

2 Results

Each gene g in the genome we split up in (up to) 18 subsequences consisting of codons for each of the amino acids. We then quantified the length of each subsequence, which corresponds to the number $|L^{\mathrm{A},g}|$ of occurrences of this amino acid in gene g. The subsequences thus obtained consist of sequences of codons. We characterised each of these subsequences by its total length $|L^{\mathrm{A},g}|$, and the number of occurrences $k_i^{\mathrm{A},g}$ of each codon type in the sequence.

Having quantified codons in this way, we can now consider the evolution of codon usage as a continuous time, discrete space random walk in sequence space. Each step of the random walk corresponds to a mutation from one codon to a different synonymous codon. To keep the model simple, we make a number of assumptions about the random walk. Firstly, we assume that non-synonymous mutations are negligible, i.e. the rate of mutation from a codon to a non-synonimous codon is zero. Secondly, we assume that the mutation rates between synonymous codons are *a priori* the same. We thus posit that there is no chemical bias towards one or the other codon. Deviations from an equal codon usage pattern are thus either due to some random drift, or some explicit selection pressure. Thirdly, the way we quanmtify codons means that we ignore spatial correlations of codon usage, either across genes or correlations of codon usage within a gene. Fourthly, we assume that the random walk is in steady state.

If there are consistent selective pressures across an organisms, at least for some amino acids, then one could cosider each codon sequence of an instantiation of a random walker. The statistical distribution of these random walkers could then reveal information about the underlying processes that drive the random walk.

In the simplest imaginable case, there is no selection bias at all and every sequence is equally likely. There are two specific proedictions one can make about this case. Firstly, the relative frequency of all codons of a partiicular codon is up to statistical fluctuations equal to $\sum_g L^{\mathrm{A},g}/|C^{\mathrm{A}}|$, where $|C^{\mathrm{A}}|$ is the number of codons for amino acid A and gene g. Secondly, the number of number of codons on individual subsequences are distributed according to a multinomial distribution where each codon has a probability of $q_i = 1/|C^{\mathrm{A}}|$; in the case of $|C\mathrm{A}| = 2$ the multinomial distribution becomes a binomial distribution with with $q_i = 1/2$.

This assumptions behind this simplest scenario can be broken in two ways. Firstly, it could be that that there is a selection pressure to favour some codons such that codon usage is biased, but codons are still randomly distributed over the genome. This beanbag model would result in codon frequencies that deviating from equal distrubtion. Yet, the distribution of sequences would still follow the multinomial distribution, but now with probabilities $q_i \neq 1/|C^{\mathrm{A}}|$.

Alternatively, it could be that there is a selection pressure for sequences of a particular type and this pressure could be such that there is no underlying change in the codon frequencies. The result would be that the distribution of sequences is no longer multinomial. To illustrate this, one could image an extreme case whereby every gene consists of exactly one type of codon only, but each subsequences has a completely random codon. This scenario is compatible with a codons being equally distributed, while at the same time the distribution of sequences is non-multinomial. At the same time, one could imagine that there is a selection pressure for sequences that leads to a change of the underlying fequency of codons as well. Note that this could also be caused by a scenario where selection for sequences and codons is acting simultaneously.

If the random walk interpretation of the codon usage bias is correct, then we can now conceptualise each sequence i as having an energy E_i. A postulate from statistical physics is that in equilibrium the frequency of observing a sequence with energy E_i should be proportional to $\exp(-E/T)$, where T is a constant that in a physical system would correspond to the temperature of the system. The energy depends on the transition rates of the random walk, and as such calculating the energy depends on knowledge of the selection pressures. In the case of the a selection pressure that acts at the level of the individual codons and for amino-acids with $|C^{\mathrm{A}}| = 2$ the energy of a sequence turns out to be the logarithm of a binomial distribution, and the Boltzmann distribution becomes a reflection of the trivial fact that the the observation frequency is directly proportional to the probability of a sequence.

We found no clear realtionship between the energies based on a model with no codon selection. However, when we calculated the energies based on the a binomial model with a bias that corresponds to the codon usage bias, then we found that the logarithm of empirically observed frequencies, when plotted against the energies are indeed consistent with a straight line. However, the slope of the line is not 1, as one would be expected if the underlying distribution were 1; instead the slope is usually smaller than 1.

This observation is consistent with a model whereby there is a selection pressure acting at the level of the sequences. Furthermore, for the vast majority of species we considered, this selection pressure is the same for all amino acids, irrespective of the number of codons. It is, however, different for different species.

As a simple model that is cosistent with the observed selection pressure, we propose a model whereby the selection is bias such that the energy of a sequence is related to the energy of an unbiased model as follows (see SI for derivation):

$$\bar{E} = \underbrace{\xi E}_{\text{entropic}} + \underbrace{-\xi T \frac{\gamma - \xi}{\xi} \ln(k!)}_{\text{bias selection}}. \tag{1}$$

Here ξ and γ are parameters that determine the selection pressure and $k!$ is the number of the less frequent codon in the sequence. The first term E is a purely entropic term which arises from a model with no underlying selection pressure. The parameter ξ can be interpreted as a generalised temperature. When $\xi = \gamma$, then the there will be no global codon usage bias, but the distribution of codons over individual sequences not be according to a binomial distribution with modified temperature ξT. Such a modifed temperature cannot be achieved by selection pressures acting at the level of individual codons, but pre-supposes a selection at a higher level, for example at the level of the subsequences.

When $\gamma - \xi \neq 0$ then there is also an additional selection potential which biases selection towards sequences with uneven codon distribution, i.e. with k high or low depending on the values of ξ and γ.

3 Discussion

4 Methods

4.1 Where did we obtain the data from?

4.2

5 Supplementary Information

5.1 The beanbag model

The simplest model is where selection operates at the level of the individual codons. For each codon assignment there will be a probability p_i that the i-th codon is chosen among all possible $|C_A|$ codons $C_1^{\mathrm{A}}, \ldots, C_{|C|_A}^{\mathrm{A}}$ for amino acid A. These codons are then randomly distributed across the genome according to probabilities p_i.

This model would entail a specific distribution of sequence compositions, namely a multinomial distribution. Specifically, the probability to observe a particular gene that has $L = L^{\mathrm{A,g}}$ occurences of a particular amino acid A with $m := |C_A|$, and k_i occurences of the i-th codon is then given by

$$P = \frac{L!}{k_1! \cdots k_m} p_1^{k_1} \cdots p_m^{k_m} \tag{2}$$

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

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 $k := k_1$, $k_2 = L - k$, $p := p_1$ and $p_2 = 1 - p$.

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

5.2 Mixture models

A variant of the codon model is the multinomial mixture model, whereby selection still happens at the level of the individual sequences, but the probability to select a codon depends on the particular sequence. If we assume that there are M different groups of genes that each share the same selection pressure. If we assume that group i of 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 sequence 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 M of classes grows, the probability P_k approaches the uniform probability distrubution, since

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

5.3 A sequence selection model

An alternative to the beanbag model is a model where not individual codons are selected, but specific sequences. As an extreme example, consider the scenario where there is no underlying codon usage bias at all and all codons occurr equally often, but all codons for an amino acid A are strictly always the same. In this case, then, the probability $p_i^A = 1/|C^A|$, but the probability to observe a particular sequence would not follow the binomial distribution. In fact, there would be only C^A possibiities to code the amino acids of a particular gene. Even though each gene in this example has an extremely strong bias, across the entire genome all codons occur with equal probability.

5.3.1 Codon usage bias 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 sequence. This analysis remains thus insensitive to correlations of codon usage within genes. To model sequence evolution, we consider the codon usage problem as a random walk. Focussing at first on amino acids with 2 codons, codon evolution can be represented as a 1D random walk in discrete space and continuous time. The number of sites corresponds to $L^{A,g}a$. Each site corresponds to a particular sequence composition, i.e. a pair $(k_1^{A,g}, L^{A,g} - k_2^{A,g})$, which defines a state. In the case of 2 codons the state is entirely characterised by k_1. States are connected when a single synonymous codon change is sufficient to get from one sequence to the other. In the case of 2 codons this means that there are two states with one connected state $((0, L^{A,g})$ and $(L^{A,g}, 0))$. All other states are connected to two other states states, such that k_1 is connected to $k_1 - 1$ and $k_1 + 1$. 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. 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.

$$r(l_1, k_2, \ldots \to k_1 + 1, k_2 - 1, \ldots) \sim k_2 \tag{5}$$

The evolution of codons is a complex process that may depend on chemical characteristics of the codons, but could also depend on selection pressures that derive from organism-level constraints and may even depend on characteristics at the level of the ecosystem. It would be hopeless to include all level of complexities involved, and we will therefore not attempt to do this. Instead, we choose the simplest model that we can conceive and conceptualise evolution as a simple random walk of a particle that has $L^{A,g}$ sites available and randomly transitions between these sites. We will positulate transition rates between the sites, and we conceptualised the random walker as being in contact with a large heatbath that remains at a fixed temperature T. We stress that this temperature T is a purely conceptual temperature that should not be confused with the actual physical temperature that is experience by organisms.

Having made this conceptualisation, we can now calculate the long-term equilibrium probabilities $\pi(k_1, k_2, \ldots)$ for various configurations. In particular, the sort of system we imagine fulfulls the detailed balance condition, which means that in equilibrium there are no net-flows of probability between any two states that are connected.

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

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 physical systems of the type described here obey the Boltzmann distribution in equilibrium. This means that the probabnility to find a particle in a given configuration depends on the energy of the configuration $E(k_1, k_2, \ldots)$

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

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 for the systems under consideration one usually postulates the local detailed balance condition, which relates transition rates between states and their energy differences ΔE.

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

This relationship then implies a dependence between the two rates, namely:

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

In effect this means that, as the temperature increases the relative strength of the transitions that take the sequence back to the more likely sequence decreases.

Calculation of the energy of a specific codon sequence In order to calculate the energies of the model, we start from the state where all L codons are of type 1. This state we define as having energy $E_0 = 0$. A mutation can reach the next state (1) by chagning one of the L codons of type 1 to a codon of type 2. This happens with a rate proportional to L. From this state 1, the system can then further move 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:

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

The energy assoicated with state (0) is then E_0, and the energy difference between state (0) and state (1) is $\Delta E_1 = \ln(L/1)$. Generally, the energy difference ΔE_k between state (k) and state $(k - 1)$ is given by $\Delta E_k = \ln\left((L - k + 1)/k\right)$. Consequently, the energy of state (k) is then.

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

This entails that the (Boltzmann-)probability to observe a specific configuration k is proportional to the boinomial coeffient, which indicates that this system is distributed according to the binomial distribution and represents a model with no selection pressure at the level of sequences.

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:

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

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

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

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

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

5.4 Changing the rates of the model

We first modify the rates such that there is a constant bias C into one direction. The random walk is then modified as follows:

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

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

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

The partition function for this case can be calculated as follows:

$$Z = 1 + \sum_{k=1}^{N}\exp\left(-\frac{E'_k}{T}\right) = 1 + \sum_{k=1}^{N}\binom{N}{k}C^{-k} = (\frac{C+1}{C})^N \tag{18}$$

By the same token we can calculate

$$A := \sum_{k=0}^{N} k\binom{N}{k}C^k = N\frac{(C+1)^{N-1}}{C} \tag{19}$$

The average codon usage ϵ is then given by

$$\epsilon = \frac{A}{NZ} = \frac{1}{C+1} \tag{20}$$

Note, however, that in this scenario, the temperature remains unaffected by the change of rates.

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 versio is now no longer proportional to the number $N-k+1$ and k of codon 2, but propotional to a power of the number. This then changes the above model, as follows:

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

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

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

This change thus only affects the temperature, but does not affect the average occurence of codons.

We now further modify the rate of the model. Specificially, we assume that there is a selection pressure on the genome such that the rate with which codon 2 mutates to codon 1 is now no longer proportional to the number k of codon 2, but propotional to k^γ, i.e. a power of the number of codon 2. This then changes the above model, as follows:

$$(N,0) \underset{1^\gamma}{\overset{N}{\rightleftharpoons}} (N-1,1) \underset{2^\gamma}{\overset{N-1}{\rightleftharpoons}} \cdots \underset{k^\gamma}{\overset{N-k+1}{\rightleftharpoons}} (N-k,k) \underset{(k+1)^\gamma}{\overset{N-k}{\rightleftharpoons}} (N-k-2,k+2) \cdots \underset{N^\gamma}{\overset{1}{\rightleftharpoons}} (0,N) \tag{23}$$

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

$$\begin{aligned}\tilde{E}_k &= \sum_{i=1}^{k} \Delta \tilde{E}_i = -\ln \left(\prod_{i=1}^{k} \frac{N-i+1}{i^\gamma} \right) = \\ &= -\ln \left(\frac{N!}{(N-k)!k!(k!)^{\gamma-1}} \right) = \\ &= E_k + T(\gamma-1)\ln(k!) \end{aligned} \tag{24}$$

The corresponding partition function is given by:

$$\begin{aligned}\tilde{Z} &= 1 + \sum_{k=1}^{N} \exp\left(-E_k - (\gamma-1)\ln(k!)\right) \\ &= 1 + \sum_{k=1}^{N} \exp(-E_k)(k!)^{(\gamma-1)} \end{aligned} \tag{25}$$

Finally, we now introduce a parametrised model as follows with tuneable bias to the left and to the right.

$$(N,0) \underset{1^\gamma}{\overset{N^\xi}{\rightleftharpoons}} (N-1,1) \underset{2^\gamma}{\overset{(N-1)^\xi}{\rightleftharpoons}} \cdots \underset{k^\gamma}{\overset{(N-k+1)^\xi}{\rightleftharpoons}} (N-k,k) \underset{(k+1)^\gamma}{\overset{(N-k)^\xi}{\rightleftharpoons}} (N-k-2,k+2) \cdots \underset{N^\gamma}{\overset{1^\xi}{\rightleftharpoons}} (0,N) \tag{26}$$

Here $\xi := \frac{l\gamma}{M} + \frac{M-l}{M}$.

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

$$\begin{aligned}\bar{E}_k &= \sum_{i=1}^{k} \Delta \bar{E}_i = -T \ln \left(\prod_{i=1}^{k} \frac{(N-i+1)^\zeta}{i^\gamma} \right) \\ &= -T\xi \ln \binom{N}{k} + T(\gamma-\xi)\ln(k!) \\ &= \xi E_k + T(\gamma-\xi)\ln(k!) \end{aligned} \tag{27}$$

6 Empirical results

The entropy can be calculated for real sequences. To this end, we split up each gene into (at most) 18 subsequences of codons for each amino acid. Each subsequence can then be assigned two theoretical entropies. Firstly, for amino acids with 2 codons the entropy relative to a an unbiased base-distribution of the codons. For a gene with altogether N amino acids of the partiuclar type and k codons of type 1, the entropy is given by

$$S_N(k) := S_N^{1/2}(k) = \ln\left(\binom{N}{k} 2^{-N} \right) \tag{28}$$

Typically, the particular amino acid has a codon usage bias, such that codon 1 occurs with a normalised frequency of q. It is interesting to calculate the entropy with respect to the codon usage bias of a sequence.

$$S_N^{\text{cub}}(k) := S_N^{q}(k) = \ln\left(\binom{N}{k} q^k (1-q)^{N-k} \right) \tag{29}$$

It is straightforward to generalise this entropy to the case of more than two codons. In which case the binomial distribution is replaced by the *multinomial* distribution in the calculation of the entropies. To calculate this, we need to note that an amino acid with l codons will have a codon usage bias characterised by the normalised frequencies of occurence $\mathbf{q} = \{q_1, q_2, \dots q_l\}$ and $\sum_{i=1}^{l} q_i = 1$. These quantify the fractions of codon 1 and 2 and so on and will be specifici for a particular species. Using these species specific codon usage frequences, it is possible to calculate the entropy of a particular gene that contains k_1 codons of type 1 and k_2 codons of type 2 and so forth

$$S_N^{\text{cub}}(k) := S_N^{\mathbf{q}}(\mathbf{k}) = \ln\left(\frac{N!}{k_1! \cdots k_l!} q_1^{n_1} \cdots q_l^{n_l}\right) \tag{30}$$

Similarly, we can also calculate the base entropy $S_N(k)$ by setting $q_i = 1/l$ in equation 30.

To make this specific, it is useful to consider as a specific example the genome of a particular species. Concentrating on the amino acid E, which has only 2 codons. Let us assume that the codon usage bias of this codon is $q = 1/4$, such that a quarter of the codons is of type 1 and 3 quarters of type 2 across the whole genome. In each of the genes of the particular species the amino acid E will occur a number of times. Let us now consider only those genes that contain exactly 3 copies of this amino acid. There are altogether $N_3 = 40$ genes that fullfill the condition. There are now exactly 4 ways in which the codons can be distributed across the sequences, these are: The first possible sequence (SEQ1) consists of all three codons are codon 1. The second possible sequence (SEQ2) is wher3 2 codons are of type 1 and 1 codon is of type 2; then there is SEQ3 with 2 codons of type 2 and SEQ4 with 3 codons of type 2. In our hypothetical example, we can now assume that 10 sequences are of tyep SEQ1, 25 of type SEQ2, 4 sequences are of type SEQ3 and 1 sequence of tyep SEQ4. Note that each of the possible sequences SEQ1 - SEQ4 are associated with an entropy, which is given by equation 29. To be concrete, we calculate the entropy associated with SEQ1:

$$S_{N=3}^{\text{cub}}(\text{SEQ1}) := S_{N=3}^{q=1/4}(k=3) = \ln\left(\binom{3}{3}(\frac{1}{4})^3(1-\frac{1}{4})^0\right) \tag{31}$$

Similarly, the entropy of SEQ2 can be calclated using the same equatuion 29.

$$S_{N=3}^{\text{cub}}(\text{SEQ2}) := S_{N=3}^{q=1/4}(k=2) = \ln\left(\binom{3}{3}(\frac{1}{4})^2(1-\frac{1}{4})^1\right) \tag{32}$$

Analogously for the other two sequences.

Next, one can compare these theoretical entropies with the empirical entropies, i.e. the logarithm of the occurence frequencies. In our example, the occurence frequency of SEQ1 is $\tilde{q}_1 = 10/40$ of SEQ2 it is $\tilde{q}_2 = 25/40$ and similarly for the other sequences. The associated "empirical" entropies are then $\tilde{S}(\text{SEQ1}) = \ln(\tilde{q}_1)$ and $\tilde{S}(\text{SEQ2}) = \ln(\tilde{q}_2)$ and so on.

Finally, we can now plot these empirical entropies $\tilde{S}$ against the theoretical ones S^{cub} for each sequence. In our case, this would result in a plot consiting of 4 points, one for each of SEQ1-SEQ4. If the particular subsequence we considered — in our case all the genes that have exactly 3 amino acids of the particular type E — then we would expect that the 4 points fall on a straight line with slope 1 and intercept 1, up to some statistical error that results from the finite sample size of 40. If, on the other hand, the codons are not distributed according to a multinomial distribution, then there are no *a priori* expectations for regular relationship between obserbed and theoretical entropies.

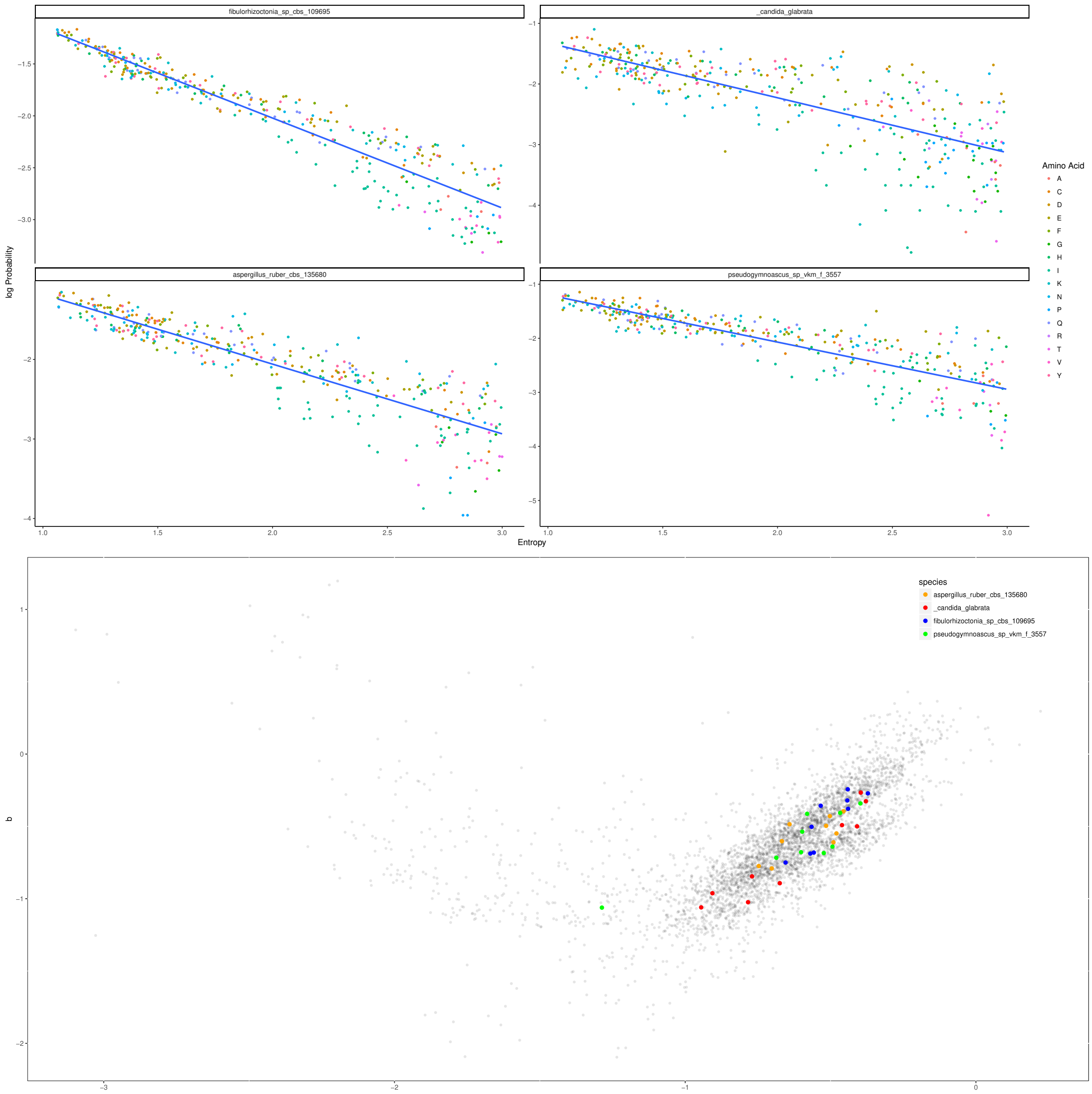

Fig. 1: (**Top:**) Empirical entropies $\tilde{S}$ as a function of the theoretical entropies S^{cub} for 4 different species. Each plot shows the subsequences of length 5-10 for all amino acids. In each of the examples the slope is significantly larger than the predicted value of -1, which implies that the sequences are not distributed according to a binomial/multinomial distribution. (**Bottom:**) We fitted the parameters ξ and η to the distributions of codons for each of the 2-codon amino acids of the 462 fungal species limiting ourselves to those amino acid sequences that have a sublength of 15. Each dot shows the ξ and η value thus obtained for a particular amino acid. The colored dots highlight the species of the curve on the top.

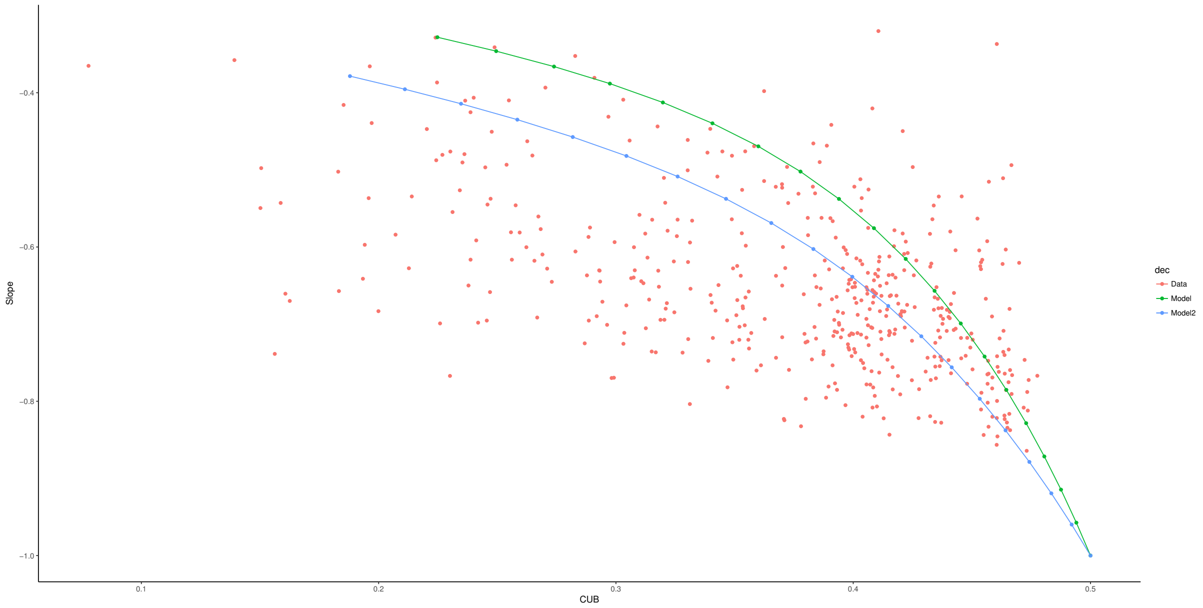

Fig. 2: The slope of the entropy as a function of the codon usage bias. The dots are 463 species of fungus. The CUB is the avergae codon usage bias q for amino acids with 2 codons. The y-axis is the slope of the entropy curves obtained by fitting theoretical entropies to empirical entropies to a straight line, as in fig 1 but only for the nine amino acids with 2 codons. The plot reveals a correlation between the codon usage bias and the temperature. The solid curves were obtained by plotting the entropies obtained from model equation 27 against a binomial distribution and fitting the resulting points. The line "Model" correspmds to $l = 3/4$ and the blue line corresponds to $l = 2/3$.