FIRST-PASSAGE DYNAMICS OF RANDOM WALKS ON COMPLEX AND MUTATIONAL CODON NETWORKS

By

WILLOW B. KION-CROSBY

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Physics & Astronomy

Written under the direction of

Prof. Alexandre V. Morozov

And approved by

____

____

____

____

____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

First-Passage Dynamics of Random Walks on Complex and Mutational Codon Networks

By WILLOW B. KION-CROSBY

Dissertation Director:

Prof. Alexandre V. Morozov

The first three chapters of this dissertation describe a novel Bayesian methodology which uses random walk sampling for rapid inference of the statistical properties of undirected networks with weighted or unweighted edges. The statistics of interest include, but are not limited to, the node degree distribution, the average degree of nearest-neighbor nodes, and the node clustering coefficient. Our formalism yields high-accuracy estimates of the probability distribution of any network node-based property, and of the network size, after only a small fraction of network nodes has been explored. The Bayesian nature of our approach provides rigorous estimates of all parameter uncertainties. We demonstrate our framework on several standard examples, including random, scale-free, and small-world networks, and apply it to study epidemic spreading on a scale-free network. We also infer properties of the large-scale network formed by hyperlinks between Wikipedia pages.

During our analysis of complex networks, the connection between the frequencies of codons and the first-passage dynamics on the underlying single-point mutational network, which describes the evolution of gene sequences, was investigated. Viewing codon evolution as a random walk with deleterious sequences representing absorbing states inevitably led to the development of a detailed biophysical model for the investigation of codon usage bias. Frequencies of synonymous codons are typically non-uniform, despite the fact that such codons correspond to the same amino acid in the genetic code. This phenomenon, known as codon usage bias, is believed to be due to a combination of factors including genetic drift, mutational effects, and selection for speed and accuracy of codon translation; however, its quantitative modeling has been elusive. Here we develop a biophysical population genetics model capable of explaining genome-wide codon frequencies. Our model implements codon-level treatment of mutations with transition/transversion biases, and includes two contributions to codon fitness which describe codon translation speed and accuracy. Furthermore, it allows wobble pairing – codon-anticodon base pairing mismatches at the 3' nucleotide position of the codon. We find that the observed patterns of genome-wide codon usage are consistent with a strong selective penalty for mistranslated amino acids. In contrast, the dependence of codon fitness on translation speed is weaker on average compared to the strength of selection against mistranslation. Although no constraints on codon-anticodon pairing are imposed *a priori*, a reasonable hierarchy of pairing rates, which conforms to the wobble hypothesis and is consistent with available structural evidence, is predicted by the model. Finally, we estimate mutation rates per nucleotide directly from the coding sequences by treating the translation process explicitly in the context of a finite ribosomal pool, and predict that mutation rates are inversely proportional to the number of genes. Overall, our approach offers a unified biophysical and population genetics framework for studying codon bias.

Acknowledgements

I acknowledge departmental support in the form of an Excellence Fellowship during my fifth year as a graduate student. This eased my workload from teaching responsibilities and also reduced my cynicism. In contrast, my teaching experience as a graduate student has been invaluable, granted it provided me with many opportunities to connect with and be a part of the lives of so many people.

I am grateful to my research advisor, Alex Morozov, for allowing me to explore a variety of topics by not discouraging my distractibility. I am also grateful to Gyan Bhanot for consistently being a source of support and anecdotes, and to Anirvan Sengupta for the many times I halted him with research questions while he was trying to get coffee. My gratitude also extends to Michael Manhart for being such a dedicated research colleague and mentor, and to Humna Awan whose motivation toward her own research and education has been a significant source of personal drive and growth.

Finally, I would like to acknowledge all of the faculty, graduate students, undergraduate students, and everyone else who contributed to my graduate career and life in so many profound ways!

The majority of the material presented in this dissertation is original work from Refs. [1] and [2] unless stated otherwise.

It was the secrets of heaven and earth that I desired to learn; and whether it was the outward substance of things or the inner spirit of nature and the mysterious soul of man that occupied me, still my inquiries were directed to the metaphysical, or in its highest sense, the physical secrets of the world.

– Mary Shelley, Frankenstein

All models are wrong, but some are useful.

– George E. P. Box

A mind not to be changed by place or time.
The mind is its own place, and in itself
Can make a heav'n of hell, a hell of heav'n.

– John Milton, Paradise Lost

A teacher affects eternity, he can never tell, where his influence stops.

– Henry B. Adams

Table of Contents

Chapter 1

Introduction to Random Walks on Complex Networks

Over the past few years, our lives have become increasingly dependent on large-scale networks, often available through our computers and smartphones. In addition to the original computer-based networks such as the World Wide Web and the Internet, many online social networks have emerged, notably Twitter and Facebook. Our professional and personal activities are influenced daily by knowledge-sharing online services such as Wikipedia and YouTube. More generally, complex networks describe a broad spectrum of systems in nature, science, technology, and society [3]. Many of these networks are large and constantly changing, making an investigation of their statistical properties a challenging task. In particular, estimating the network size becomes non-trivial if the network is too large to resort to brute-force methods such as visiting every node. Consequently, predicting various network statistics, typically from random samples of limited size, has attracted considerable attention in the literature [4–11].

Here we present the development of a Bayesian theoretical framework for network sampling by random walks (RWs) [6, 9]. Unlike previous results, this framework can be used to build posterior probability distributions for any network node-based quantity of interest. This approach reproduces several previously known global network statistics estimators within a single formalism, automatically removes statistical biases caused by RW sampling [6,7], and yields standard results in the uniform sampling limit. Surprisingly, accurate estimates of various network properties, including its size, are obtained after examining only a small fraction of all network nodes. The effectiveness of this formalism is demonstrated in Chapter 2 not only on standard *in silico* networks, but additionally with applications in epidemiology, and has produced known and new statistics of the network formed by links between pages on Wikipedia in Chapter 3.

Consequentially the network property estimators which are developed by this formalism show faster convergence than their uniform sampling counterparts.

1.1 Formulation of the Mean Return Time

Consider a RW on a network of N nodes with weighted edges: $\{w_{ji}\}$, where w_{ji} is the rate of transition from node i to node j. The number of edges or links connecting i to neighboring nodes is denoted k_i and is known as the node degree [3]. Such a network is illustrated in Fig. 1.1.

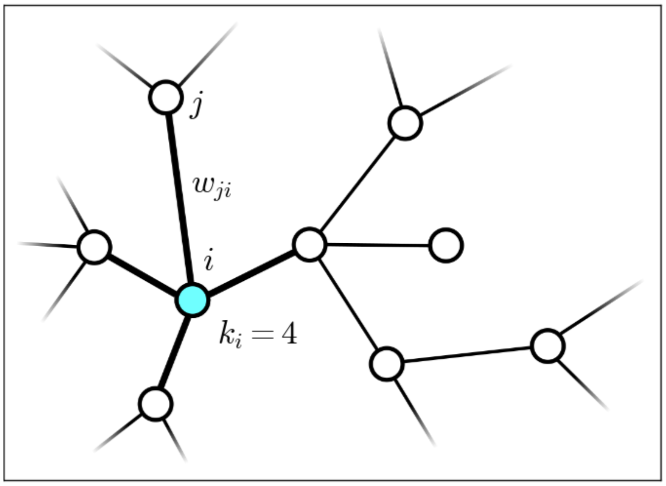

Figure 1.1: **Microscopic structure of a complex network of nodes and edges.** Node i has edge number $k_i = 4$ and is connected to node j via the edge weighted w_{ji}.

At each step, the walker will transition to a neighboring node with a probability given by the edges weights through

$$P(i \to j) = \frac{w_{ji}}{\sum_{k \in \{nn\}_i} w_{ki}}, \tag{1.1}$$

where the sum is over all nearest neighbors of node i. We subdivide all network nodes into sets S_x based on the value of some property x, such as the number of links connected to the current node, k_i; there are N_x nodes in each set. We assume that the property in question is either discrete or can be discretized by binning if continuous.

Focusing on undirected networks with symmetric rates, $w_{ji} = w_{ij}$, the stationary probability for the RW to occupy node i, π_i, can be determined using the steady-state master equation [38,39]:

$$\sum_{j \in \{nn\}_i} \left[\pi_j P(j \to i) - \pi_i P(i \to j)\right] = 0. \tag{1.2}$$

Equation (1.2) is satisfied if $\pi_i \sim w_i = \sum_{k \in \{nn\}_i} w_{ki}$, where w_i is the total outward rate from node i. It follows that for unweighted networks, the node's stationary probability is simply proportional to its degree k_i [40]. With normalization, the stationary probabilities become

$$\pi_i = \frac{w_i}{\sum_{i=1}^{N} w_i}. \tag{1.3}$$

If the walker starts from a node with property x, the average number of steps between subsequent visits to any node within the set S_x, also known as the mean return time (MRT), is given by [41]:

$$\langle \ell \rangle_x = \frac{1}{\sum_{i \in S_x} \pi_i}. \tag{1.4}$$

In the case of undirected networks this is

$$\langle \ell \rangle_x = \frac{\langle w \rangle}{p_x \langle w \rangle_x}, \tag{1.5}$$

where $p_x = N_x/N$ is the fraction of nodes with property x, $\langle w \rangle = N^{-1} \sum_{i=1}^{N} w_i$, and $\langle w \rangle_x = N_x^{-1} \sum_{i=1}^{N_x} w_i$.

1.2 Return Time Distribution

To recover the full return-time distribution, we follow the approach of [42] and define the jump matrix $\mathbf{Q}$ where the entries of this matrix are given by Eq. (1.1),

$$\mathbf{Q}_{ji} = P(i \to j). \tag{1.6}$$

Introducing the notation that $|\sigma_i\rangle$ is a column vector with a single non-zero entry at tha ith position equal to 1, the probability that the RW process is found on node j

after ℓ steps, or the occupation probability, is given by

$$P(\ell) = \langle \sigma_j | \mathbf{Q}^{\ell} | \sigma_i \rangle. \tag{1.7}$$

To describe the return process to nodes with property x, we start the RW a single step off a node from set S_x, $|\sigma_0\rangle \equiv \mathbf{Q}|\sigma_i\rangle$, and replace the single destination node row vector $\langle \sigma_j|$ with a sum nodes in S_x,

$$\langle \sigma_x | = \sum_{j \in \mathcal{S}_x} \langle \sigma_j |. \tag{1.8}$$

To only treat first returns, the jump matrix must also be modified so that the nodes in S_x act as absorbing states, $\mathbf{Q} \to \mathbf{Q}_x$. This is done by setting the probability to transition out of any of these nodes to zero,

$$\langle \sigma_j | \mathbf{Q}_x | \sigma_i \rangle = 0, \forall i \in S_x. \tag{1.9}$$

Therefore the probability that a RW process will return on exactly the $\Delta\ell$th step to any node within S_x after starting from a node in this set is given by

$$P(\Delta\ell | \mathbf{Q}_x) = \begin{cases} \langle \sigma_x | \mathbf{Q}_x^{\Delta\ell - 1} | \sigma_0 \rangle, & \text{if } \Delta\ell \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{1.10}$$

Expanding $|\sigma_0\rangle$ in the eigenbasis of $\mathbf{Q}_x$, the non-zero portion of Eq. (1.10) becomes

$$P(\Delta\ell | \mathbf{Q}_x) = \langle \sigma_x | \mathbf{Q}_x^{\Delta\ell - 1} \sum_i a_i |\psi_i\rangle = \sum_i a_i \lambda_i^{\Delta\ell - 1} \langle \sigma_x | \psi_i \rangle, \tag{1.11}$$

where each λ_i is an eigenvalue of $\mathbf{Q}_x$ with corresponding eigenvector $|\psi_i\rangle$. If the eigenvalues are ordered from largest to smallest, $\lambda_0 > \lambda_1 > ...$, then after a sufficient number of steps, $\Delta\ell^*$, defined by

$$\frac{a_1 \langle \sigma_x | \psi_1 \rangle \lambda_1^{\Delta\ell^* - 1}}{a_0 \langle \sigma_x | \psi_0 \rangle \lambda_0^{\Delta\ell^* - 1}} \ll 1, \tag{1.12}$$

the first term will dominate the sum in Eq. (1.11),

$$P(\Delta\ell|\mathbf{Q}_x) \approx a_0 \langle \sigma_x|\psi_0\rangle \lambda_0^{\Delta\ell-1}. \tag{1.13}$$

With the definition $q_x \equiv -\ln\lambda_0$, this leads to our central ansatz,

$$P(\Delta\ell|q_x) \propto \begin{cases} e^{-q_x\Delta\ell}, & \text{for } \Delta\ell \geq 1. \\ 0, & \text{otherwise} \end{cases} \tag{1.14}$$

With the normalization condition $\sum_{\Delta\ell=0}^{\infty} P(\Delta\ell|q_x) = 1$, this yields exactly

$$P(\Delta\ell|q_x) = (e^{q_x} - 1)e^{-q_x\Delta\ell} \approx q_x e^{-q_x\Delta\ell} \;\text{ for }\; \Delta\ell \geq 1, \tag{1.15}$$

where the approximation is valid for $q_x \ll 1$. Note that with this condition, the average number of steps between returns is simply

$$\langle\ell\rangle_x = \frac{1}{1-e^{-q_x}} \approx \frac{1}{q_x} \tag{1.16}$$

This exponential ansatz for the return time distribution is supported by several observations: The behavior of the solutions to the diffusion equation in d dimensions yield a first-passage distribution with exponential long-time behavior for finite systems [43]. Additionally, the return-time distribution is known to be asymptotically exponential in arbitrary finite networks [44]. We also find empirically that an exponential ansatz for $P(\Delta\ell|q_x)$ is sufficiently accurate for our purposes (Fig. 2.1(a)–3.1(a)), although the following analysis is not limited to it.

1.3 Bayesian Formalism

Equations (1.5) and (1.16) provide a connection between the RT distribution parameter q_x and several network properties: the average outward rate over all nodes, $\langle w\rangle$, over nodes with property x, $\langle w\rangle_x$, as well as the fraction of nodes with property x, p_x. It then follows that predicting q_x from the dynamics of a RW would provide some insight into these global network properties. We start this analysis by constructing the likelihood function of q_x given the characteristics of a RW on a complex network.

With the distribution given by Eq. (1.15), the survival probability, or the probability to have no return events in $\Delta\ell$ consecutive steps, is given by

$$S(\Delta\ell|q_x) = 1 - \sum_{\Delta\ell'=1}^{\Delta\ell} (e^{q_x} - 1)e^{-q_x\Delta\ell'} = e^{-q_x\Delta\ell}. \tag{1.17}$$

Therefore the likelihood that during a single RW of length ℓ steps the walker has visited the nodes in S_x at intervals $\Delta\ell_1 = \ell_1, \Delta\ell_2 = \ell_2 - \ell_1, ..., \Delta\ell_{\mathcal{K}_x} = \ell_{\mathcal{K}_x} - \ell_{\mathcal{K}_x-1}$, and has not returned to S_x for the remaining $\Delta\ell_{\mathcal{K}_x+1} = \ell - \sum_{i=1}^{\mathcal{K}_x} \Delta\ell_i$ steps, is

$$P(\{\Delta\ell_i\}|q_x) = e^{-q_x(\ell-\sum_{i=1}^{\mathcal{K}_x}\Delta\ell_i)} \prod_{i=1}^{\mathcal{K}_x} q_x e^{-q_x\Delta\ell_i} = q_x^{\mathcal{K}_x} e^{-q_x\ell}. \tag{1.18}$$

Granted that this is independent of the intervals between returns, $\{\Delta\ell_i\}$, it then follows that the likelihood of $\mathcal{K}_x$ visits to nodes in S_x for a RW with a total of ℓ steps independent of the intervals between subsequent visits is

$$P(\mathcal{K}_x|q_x) = \sum_{\Delta\ell_1+\Delta\ell_2+...\Delta\ell_{\mathcal{K}_x+1}=\ell} q_x^{\mathcal{K}_x} e^{-q_x\ell} = \binom{\ell}{\mathcal{K}_x} q_x^{\mathcal{K}_x} e^{-q_x\ell}. \tag{1.19}$$

This function has a maximum likelihood (ML) value of

$$\hat{q}_x = \frac{\mathcal{K}_x}{\ell}. \tag{1.20}$$

Assuming a uniform prior for q_x in the $[0,1]$ range, the posterior probability for q_x becomes

$$P(q_x|\mathcal{K}_x) = \frac{P(\mathcal{K}_x|q_x)P(q_x)}{P(\mathcal{K}_x)} = \frac{dq_x}{P(\mathcal{K}_x)} q_x^{\mathcal{K}_x} e^{-q_x\ell}. \tag{1.21}$$

The evidence, $P(\mathcal{K}_x)$, can be determined granted that Eq. (1.21) must be normalized over q_x,

$$P(\mathcal{K}_x) = \int_0^1 dq_x q_x^{\mathcal{K}_x} e^{-q_x\ell} = \frac{\mathcal{K}_x!}{\ell^{\mathcal{K}_x+1}}\left(1 - e^{-\ell}\right) - \frac{e^{-\ell}}{\ell}\left(\sum_{j=0}^{\mathcal{K}_x-1} \frac{\mathcal{K}_x!}{(\mathcal{K}_x - j)!\ell^j}\right). \tag{1.22}$$

The summation in the second term is bounded from above by

$$\sum_{j=0}^{\mathcal{K}_x-1} \frac{\mathcal{K}_x!}{(\mathcal{K}_x-j)!\ell^j} \leq \sum_{j=0}^{\mathcal{K}_x-1} \left(\frac{\mathcal{K}_x}{\ell}\right)^j = \frac{1-\left(\frac{\mathcal{K}_x}{\ell}\right)^{\mathcal{K}_x}}{1-\frac{\mathcal{K}_x}{\ell}}, \tag{1.23}$$

which approaches a nonzero constant value as ℓ increases given that $\mathcal{K}_x \leq \ell$. Therefore Eq. (1.22) is well approximated by

$$P(\mathcal{K}_x) \approx \frac{\mathcal{K}_x!}{\ell^{\mathcal{K}_x+1}} \tag{1.24}$$

even after a small number of steps. In this regime, Eq. (1.21) becomes a gamma distribution,

$$P(q_x|\mathcal{K}_x) \approx dq_x\Gamma(q_x;\mathcal{K}_x+1,\ell) = dq_x\ell\frac{(q_x\ell)^{\mathcal{K}_x}}{\mathcal{K}_x!}e^{-q_x\ell} \tag{1.25}$$

which rapidly approaches a Gaussian as ℓ increases. This posterior probability distribution is then completely characterized by a mean of $\bar{q}_x = \hat{q}_x$ and standard deviation of $\sigma_{q_x} = \hat{q}_x/\sqrt{\mathcal{K}_x}$.

1.4 Network Property Estimators

The resulting posterior probability distribution recovered in the previous section defines an ML value and standard error for q_x (Eq. (1.25)). This quantity is connected to the fraction of nodes with property x, p_x, through Eqs. (1.5), (1.16), and (1.20). Together this yields a maximum likelihood estimate (MLE) and a standard error for the probability p_x of the property x:

$$\hat{p}_x = \frac{\mathcal{K}_x}{\ell}\frac{\langle w\rangle}{\langle w\rangle_x} \quad \text{and} \quad \sigma_{p_x} = \frac{\hat{p}_x}{\sqrt{\mathcal{K}_x}} \tag{1.26}$$

Generally, $\langle w\rangle$ and $\langle w\rangle_x$ are not known. However, imposing normalization recovers an estimator which is independent of these two quantities,

$$\hat{p}_x = \frac{\mathcal{K}_x/\langle w\rangle_x}{\sum_{x'}\mathcal{K}_{x'}/\langle w\rangle_{x'}}. \tag{1.27}$$

As an example: if the property x is the outward rate w, Eq. (1.27) yields

$$\hat{p}_w = \frac{\mathcal{K}_w/w}{\sum_{w'} \mathcal{K}_{w'}/w'}, \tag{1.28}$$

where $\mathcal{K}_w$ is the number of visits to nodes with total outward rate w.

For an arbitrary node property x, each set S_x can be additionally subdivided by the values of w, such that

$$\hat{p}_x = \sum_w \hat{p}_{x,w} = \sum_w \frac{\mathcal{K}_{x,w}}{w} / \sum_{w'} \frac{\mathcal{K}_{w'}}{w'}, \tag{1.29}$$

where Eq. (1.28) was employed to compute $\hat{p}_{x,w}$. Here, $\mathcal{K}_{x,w}$ is the number of visits to nodes with property x and total outward rate w. Thus, the knowledge of $\mathcal{K}_w$, $\mathcal{K}_{x,w}$, and w is sufficient to reconstruct the MLE of the distribution of any property x, estimate the error in this reconstruction (Eq. (1.27)), and compute moments to arbitrary order. Note that the division by the outward rates in Eq. (1.29) naturally corrects for the bias known to be introduced by RW sampling [6–8]. For unweighted networks ($w_{ij} = 1,\ \forall ij$), $\hat{p}_w$ reduces to $\hat{p}_k$, the network degree distribution [3].

It follows from the estimate of the outward rate distribution (Eq. (1.28)) that the MLE of the average outward rate is given by

$$\langle \hat{w} \rangle = \sum_w w\hat{p}_w = \frac{\ell}{\sum_{w'} \mathcal{K}_{w'}/w'}, \tag{1.30}$$

where we used $\sum_w \mathcal{K}_w = \ell$. Note that the imposed normalization in Eq. (1.29) is identical to using this average outward rate estimator directly in Eq. (1.26) to find the fraction of nodes with property x. The uncertainty of this estimate can be evaluated *in quadrature* assuming independence in each p_w,

$$\sigma^2_{\langle w \rangle} = \left\langle \left(\sum_w w p_w - \langle \hat{w} \rangle \right)^2 \right\rangle = \sum_w w^2 \sigma^2_{p_w}, \tag{1.31}$$

where the outer $\langle ... \rangle$ represent an average of over the posterior probability distributions.

This result in combination with Eqs. (1.27) and (1.28) yields $\sigma_{\langle w \rangle} = \langle \hat{w} \rangle / \sqrt{\ell}$, in accordance with the central limit theorem. For an arbitrary property, it follows similarly that

$$\langle \hat{x} \rangle = \sum_x x \hat{p}_x \quad \text{and} \quad \sigma^2_{\langle x \rangle} = \sum_x x^2 \sigma^2_{p_x}. \tag{1.32}$$

1.5 Network Size Estimation

We will now focus on estimating a specific property: the full network size, N. Let us suppose now that the network nodes are divided into two sets: N_p randomly chosen nodes, which we will refer to as *pseudotargets*, and all other nodes. The pseudotarget nodes are drawn prior to exploring the network, so that the average pseudotarget outward rate, $\langle w \rangle_p$, is known. Equations (1.5) and (1.19) can now be used to construct the posterior probability for the network size (assuming a uniform prior in the $[N_p, N_{max}]$ range, where N_{max} denotes an expected upper limit on N):

$$P(N|\mathcal{K}_p) = \frac{N^{-\mathcal{K}_p} \exp\left\{-\frac{N_p \langle w \rangle_p}{N \langle w \rangle} \ell\right\}}{\sum_{\tilde{N}=N_p}^{N_{max}} \tilde{N}^{-\mathcal{K}_p} \exp\left\{-\frac{N_p \langle w \rangle_p}{\tilde{N} \langle w \rangle} \ell\right\}}, \tag{1.33}$$

where $\mathcal{K}_p$ is the number of visits to pseudotargets. Note that using non-uniform priors in Eqs. (1.21) and (1.33) will not significantly affect the results, as long as $\mathcal{K}_x$ and $\mathcal{K}_p$ are sufficiently large. Similar to Eq. (1.21), we find that this posterior probability quickly becomes Gaussian as $\mathcal{K}_p$ increases, with

$$\hat{N} = \frac{\ell N_p \langle w \rangle_p}{\mathcal{K}_p \langle w \rangle} \quad \text{and} \quad \sigma_N = \frac{\hat{N}}{\sqrt{\mathcal{K}_p}}. \tag{1.34}$$

Using Eq. (1.30), we obtain

$$\hat{N} = \frac{N_p \langle w \rangle_p}{\mathcal{K}_p} \sum_w \frac{\mathcal{K}_w}{w}. \tag{1.35}$$

Note that the error in $\hat{N}$ can be reduced either through increasing N_p or assigning highly-connected nodes (network hubs) to be pseudotargets. In the $N_p = 1$ limit, Eq. (1.34) recovers the network size estimator found in Ref. [9]. This process of counting returns to pseudotargets is similar to the methods for computing network size estimators discussed in both [45] and [9].

An alternative approach is to start at the level of Eq. (1.26) where the property x is either 1 or 0 if the node is or is not a pseudotarget, respectively. The fraction of pseudotargets is then given by

$$\hat{p}_p = \frac{\mathcal{K}_p \langle w \rangle}{\ell \langle w \rangle_p}. \tag{1.36}$$

Granted that N_p is known, this estimate for the fraction of nodes which are pseudotargets can be used to estimate the network size: $\hat{N} = N_p / \hat{p}_p$. Brief inspection of Eq. (1.36) demonstrates that this yields an identical estimator to Eq. (1.34).

Chapter 2

Validation and Synthetic Networks

We have implemented the above network statistics acquisition framework as follows for several examples: for each network, N_p pseudotargets are randomly drawn when node sampling is possible, and their $\langle w \rangle_p$ is computed. Commencing the RW from one of these pseudotargets, we record ℓ, $\mathcal{K}_p$, $\{\mathcal{K}_w\}$, and $\{\mathcal{K}_{x,w}\}$ for a desired set of node properties x. At any point during the RW, Eqs. (1.27)–(1.35) can then be used to find various network statistics.

A minimal form of this RW sampling process could be followed in the case of only computing $\langle \hat{w} \rangle$ and its standard error. Under this scheme, only two variables would need to be saved in physical memory: the number of steps ℓ, and the sum of the reciprocal out rates seen thus far by the walker, $\sum_{i=1}^{\ell} w_i^{-1}$. The estimator would then be simply

$$\langle \hat{w} \rangle = \frac{\ell}{\sum_{i=1}^{\ell} 1/w_i} \tag{2.1}$$

with standard error

$$\sigma_{\langle w \rangle} = \frac{\sqrt{\ell}}{\sum_{i=1}^{\ell} 1/w_i}, \tag{2.2}$$

in contrast to Eq. (1.30) in which both the visits and outward rates are stored separately. This formula is just the harmonic mean of the outward rates seen by the walker, and is notably different from what might be the more intuitive choice of taking an algebraic mean,

$$\langle \hat{w} \rangle = \frac{1}{\ell} \sum_{i=1}^{\ell} w_i \tag{2.3}$$

2.1 Synthetic Network Examples

We have used this algorithm to study three unweighted, undirected networks: an Erdős-Rényi (ER) random graph [46], a scale-free (SF) random graph [3], and a small-world (SW) network [47]. Each network has $N = 10^6$ nodes. The ER network was constructed by randomly assigning $\lceil N \log(N)/2 \rceil$ edges between nodes, the SF network by the preferential attachment method [3] with $m = 2$ edges attached to new nodes, and the SW network as described in Ref. [48], with the shortcut probability $p = 1/2$.

For each network, $N_p = 10^3$ pseudotargets were randomly drawn and the network was subsequently explored with a random walk for $\ell = 10^5$ steps, visiting at most 10% of all nodes. Besides network sizes and degree distributions, we tracked posterior probabilities of the average degree of nearest-neighbor nodes,

$$\langle k_{nn} \rangle_i \equiv k_i^{-1} \sum_{j \in \{nn\}_i} k_j, \tag{2.4}$$

the clustering coefficient [4],

$$C_i \equiv \frac{2y}{k_i(k_i - 1)}, \tag{2.5}$$

where y is the total number of links shared by the nearest neighbors of node i, and a measure of the degree inhomogeneity [7]

$$\rho_i \equiv \sum_{j \in \{nn\}_i} \left(k_i^{-1/2} - k_j^{-1/2} \right)^2. \tag{2.6}$$

This final quantity averaged over all nodes in the network is the heterogeneity index defined in [7] from which each network's *Randić index* can be determined.

A summary of the ER system statistics is provided in Fig. 2.1. Fig. 2.1(a) shows that the exponential ansatz for the RT distribution, Eq. (1.15), is accurate for this system. Fig. 2.1(b) demonstrates the convergence of the average degree and the network size to the exact values during 5 representative runs. The predicted degree distribution, p_{k_i}, known to be Poisson [46], is shown in Fig. 2.1(c). Finally, in Fig. 2.1(d), we demonstrate the evolution of the posterior distribution for the network size as more data is collected. Additional statistics for the ER, SF and SW systems are summarized in Table 2.1. The reconstruction (red points with vanishingly small error bars) overlays the true distribution (green line) almost exactly as these statistics converge long before the error in the network size diminishes due to the choice of pseudotargets. This point is further emphasized in Fig. 2.1(b) which shows that even early in the walk, the estimate for $\langle k\rangle$ is close to its true value for all 20 trials shown, while $\hat{N}$ is still variable, although the true network values are well within the confidence intervals enveloping each trajectory. Note that the assumed exponential form of the RT distribution does indeed appear to be accurate for this system, as is evidenced by Fig. 2.1(c). Additionally,

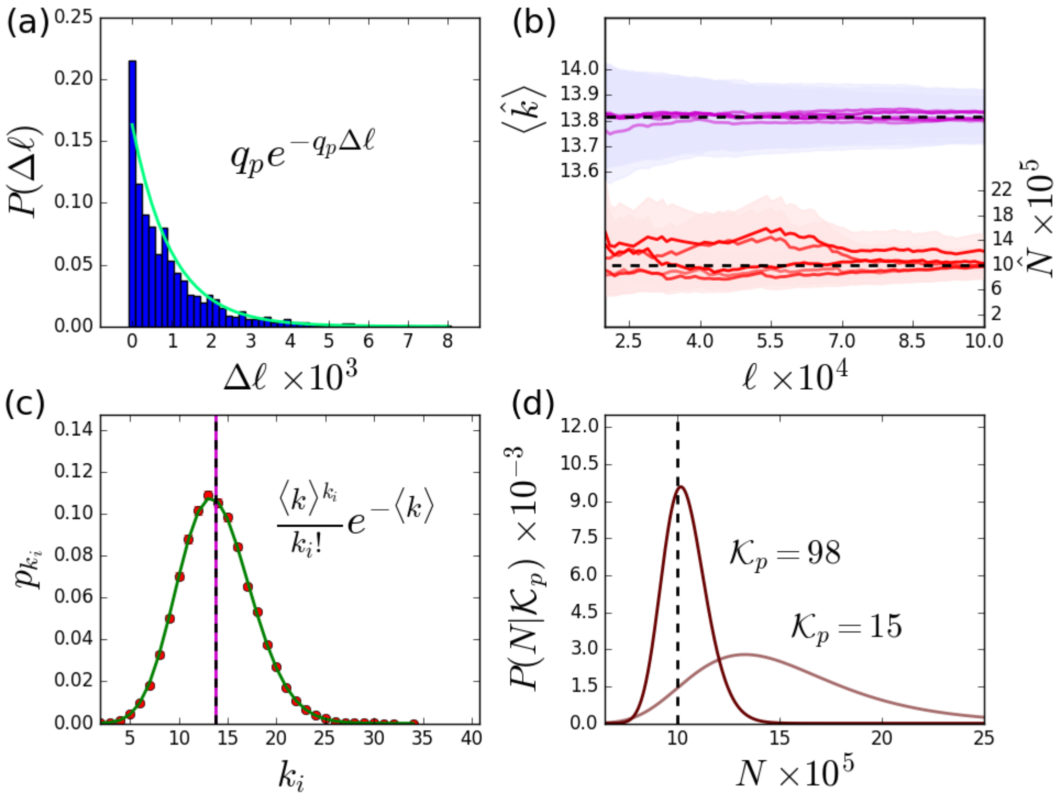

Figure 2.1: **Erdős-Rényi network statistics.** (a) Pseudotarget RT distribution (20 independent trials). Equation (1.15) parameterized by exact $q_p = \langle\ell\rangle_p^{-1}$ (Eq. (1.5)) is shown in cyan. (b) MLEs for $\langle k\rangle$ and N (purple and red lines, respectively) for 5 representative trials, with envelopes representing ±2σ intervals and exact values shown as dashed lines. (c) MLE for p_{k_i} with ±2σ intervals (red circles with error bars) (single trial); exact distribution shown in green. Exact $\langle k\rangle$ and its MLE shown is as dashed and solid lines, respectively. (d) Posteriors for N at two $\mathcal{K}_p$ values (single trial). Exact N shown as dashed line.

Table 2.1: **Network statistics summary for the Erdős-Rényi, Scale-Free, and Small-World systems.** Shown are MLE and 95% confidence interval (2σ) for each quantity, followed by exact values in parenthesis. All predictions are based on single trials with $\ell = 10^5$ steps.

	Erdős-Rényi		Scale-Free		Small-World	
N	1.25×10^6 $\pm.40 \times 10^6$	(10^6)	8.91×10^5 $\pm 1.65 \times 10^5$	(10^6)	9.48×10^5 $\pm 1.85 \times 10^5$	(10^6)
$\langle k \rangle$	13.8 $\pm.1$	(13.8)	4.00 $\pm.03$	(4.00)	4.01 $\pm.03$	(4.00)
$\langle\langle k_{nn} \rangle\rangle$	14.8 $\pm.1$	(14.8)	24.5 $\pm.6$	(24.5)	4.12 $\pm.03$	(4.12)
$\langle C \rangle$	1.46×10^{-5} $\pm.45 \times 10^{-5}$	(1.38) $\times 10^{-5}$	9.05×10^{-5} $\pm 7.18 \times 10^{-5}$	(9.89) $\times 10^{-5}$	.882 $\pm.006$	(.883)
$\langle \rho \rangle$	.0374 $\pm.0005$	(.0373)	.386 $\pm.003$	(.387)	.0142 $\pm.0001$	(.0141)

and to demonstrate a full application of Eq. (1.33), the posterior distribution for the network size has been constructed at two separate times in Fig. 2.1(d). These two times correspond to early in the walk, $\mathcal{K}_p = 15$, and after the walk has completed, $\mathcal{K}_p = 98$. Further results calculated for this system as well as the SF and SW systems are summarized in Table 2.1. Although the topologies of these three systems are quite different, we recovered the network-wide averages with high fidelity.

2.2 Generalized Erdős-Rényi Network

Next, we have constructed a generalized ER network with $N = 10^6$ nodes and weighted edges. After placing all the edges as in the unweighted ER network, a loop was added to each node with probability $p = 1/2$. All loops and edges were then assigned a symmetric weight $w_{ij} = w_{ji}$ drawn from an exponential distribution with unit mean. For this system, we have collected statistics on each node's total outward rate, w_i, loop weight, $w_i^{\text{loop}} = w_{ii}$ (note that $w_{ii} = 0$ for nodes without loops), outward rate averaged

Table 2.2: **Network statistics summary for the Generalized Erdős-Rényi system.** Shown are MLE and 95% confidence interval (2σ) for each quantity, preceded by exact values in parenthesis. All predictions are based on single trials with $\ell = 10^5$ steps.

Generalized Erdős-Rényi				
N	$\langle w \rangle$	$\langle\langle w_{nn} \rangle\rangle$	$\langle w^{\text{loop}} \rangle$	$\langle\langle w_{nn}^{\text{loop}} \rangle\rangle$
(10^6)	(14.3)	(15.3)	(.501)	(.519)
1.18×10^6	14.3	15.3	.503	.518
$\pm.26 \times 10^6$	$\pm.1$	$\pm.1$	$\pm.006$	$\pm.004$

over all nearest neighbors of node i, $\langle w_{nn} \rangle_i$, and average nearest-neighbor loop weight, $\langle w_{nn}^{\text{loop}} \rangle_i$.

We have explored the statistics of these quantities using a RW with $\ell = 10^5$ steps and $N_p = 10^3$ randomly drawn pseudotargets (Table 2.2, Fig. 2.2). Note that the RT distribution for this system deviates from purely exponential since many returns occur after a single step due to loops (Fig. 2.2(a)). Nonetheless, all the network statistics we have considered are predicted accurately (Fig. 2.2(b)–(d)), except for the tail of the Fig. 2.2(d) distribution since those rare events were not observed. Thus our methodology is equally applicable to studies of weighted networks with loops.

2.3 Traffic-Driven Epidemiological Model

After validating our approach on model systems, we have demonstrated its effectiveness in a more realistic setting, by tracking an epidemic spreading on a scale-free network in the traffic-driven epidemiological (TDE) model [49]. Following Ref. [49], we have generated the underlying network using a hidden-metric approach, which employs a tunable parameter α to control the degree of local node clustering [50, 51], and the degree distribution follows a power-law, $p_{k_i} \sim k_i^{-\gamma}$. For our network, we have chosen $N = 10^5$, $\gamma = 2.6$, and $\alpha = 2$ (which leads to significant clustering).

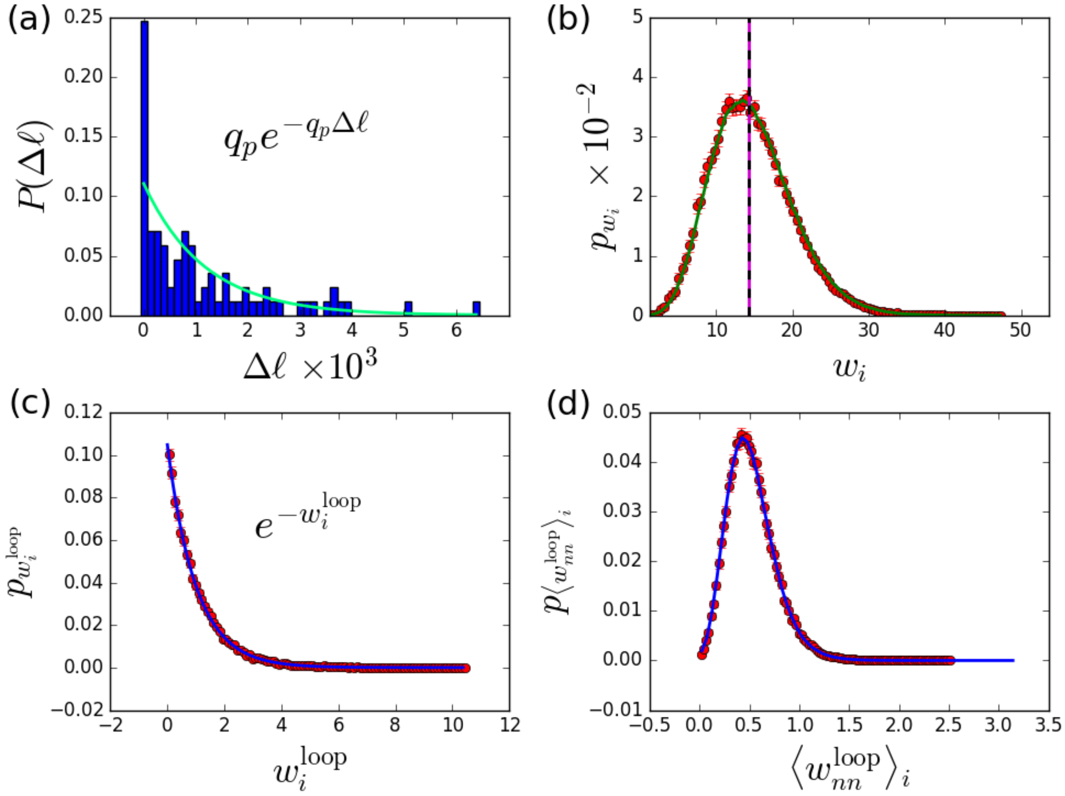

Figure 2.2: **Network statistics of a generalized Erdős-Rényi network.** (a) Pseudotarget RT distribution. Equation (1.15) parameterized by exact q_p is shown in cyan. (b) MLE $\pm 2\sigma$ (red circles with error bars) for the distribution of total outward rates; exact distribution shown in green. Predicted average: solid line, exact average: dashed line. (c) MLE $\pm 2\sigma$ (red circles with error bars) for the distribution of loop weights; exact distribution shown in blue. (d) MLE $\pm 2\sigma$ (red circles with error bars) for the distribution of loop weights averaged over all nearest neighbors; exact distribution shown in blue. In (b)–(d), all values were grouped into 100 bins.

2.3.1 Hidden-Metric Model

The hidden metric consists of a 1-dimensional circle [50]. After assigning each node a uniformly drawn location on this hidden metric, $\theta \in [0, 2\pi)$, each node is given an expected degree, κ, drawn from the power-law distribution, $p(\kappa) \sim \kappa^{-\gamma}$. Each pair of nodes is then linked with a probability based on the node locations on the metric and their expected degrees:

$$p \sim \left(1 + \frac{d(\theta, \theta')}{\eta \kappa \kappa'}\right)^{-\alpha}, \tag{2.7}$$

where $\eta \equiv (\alpha - 1)/2\langle k \rangle$, and $d(\theta, \theta')$ is the geodesic distance between the two nodes on the hidden metric. This method of generation results in a network with not only the Scale-Free and Small-World properties, but additionally develops local cluster structures through the parameter α. The resultant degree distribution has been shown to asymptotically have the same power-law behavior as the expected degree distribution with characteristic exponent γ. Figure 2.3(a) demonstrates that even with the local structures present in this network, the RT distribution is very well approximated by the exponential ansatz for our set of $N_p = 1000$ pseudotargets.

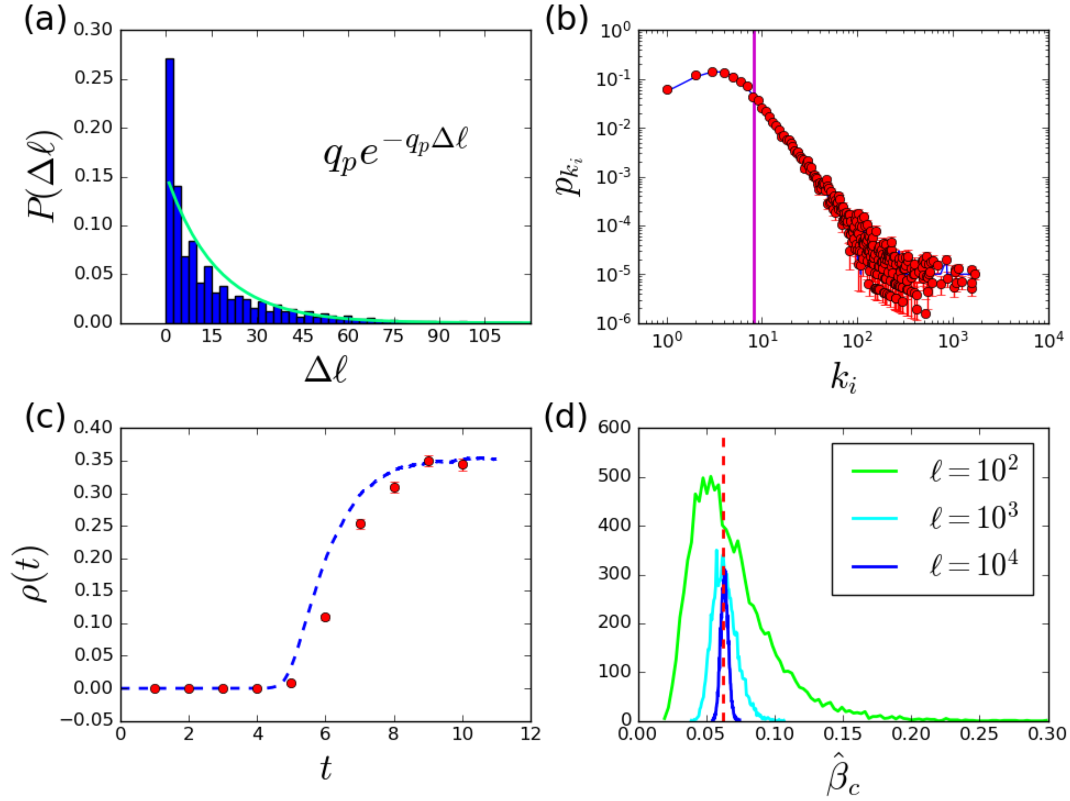

Figure 2.3: **Epidemic spreading statistics.** (a) Pseudotarget RT distribution. Equation (1.15) parameterized by exact q_p is shown in cyan. (b) MLE $\pm\, 2\sigma$ (red circles with error bars) for the node degree distribution; exact distribution is shown in blue and its average is shown as a vertical line. (c) MLE $\pm\, 2\sigma$ (red circles with error bars) for the fraction of infected nodes $\rho(t)$ computed at unit time intervals, with the exact value shown as a dashed blue curve. (d) Histograms of β_c MLEs obtained using 10^4 independent runs with $\ell = 10^2$, 10^3, 10^4 steps. Exact value is shown as a vertical dashed line.

2.3.2 Epidemic Simulation

Epidemic propagation was simulated through the exchange of W contagion packets between nodes (see Ref. [49] for details). Briefly, each node can be in either a susceptible or infected state; the simulation starts with a single infected node. When a packet moves from node i to node j on the network, node j becomes infected with the spreading probability β if node i was infected; infected nodes can also recover with rate μ, set to 1 without loss of generality. We have focused on the case in which contagion packets perform RWs between randomly assigned initial and destination nodes. Once a packet reaches its destination, it is removed and a new packet is added to keep W constant. The rate of packet movement, ν, is set such that on average each packet moves once per unit simulation time. Under this choice of packet dynamics, the mean-field equation for the relative fraction of infected nodes with degree k at time t, denoted $\rho_k(t)$, in the simulation is given by [49]

$$\partial_t \rho_k(t) = -\mu \rho_k(t) + W\nu \frac{k}{\langle k \rangle N}[1-\rho_k(t)]\Theta(t)\beta, \tag{2.8}$$

where $\Theta(t)$ is the probability that a packet is transferring from an infected node and is given by

$$\Theta(t) = \sum_k \frac{k}{\langle k \rangle} p_k \rho_k(t) \tag{2.9}$$

At steady-state, $\partial_t \rho_k(t) = 0$, this probability can be shown, through Eqs. (2.8) and (2.9), to follow the consistency equation

$$\Theta = \sum_k \frac{W\nu \frac{k^2 p_k}{\langle k \rangle^2 N}\Theta\beta}{\mu + W\nu \frac{k}{\langle k \rangle N}\Theta\beta}, \tag{2.10}$$

which has a solution for Θ if

$$\frac{d}{d\Theta} \sum_k \frac{W\nu \frac{k^2 p_k}{\langle k \rangle^2 N}\Theta\beta}{\mu + W\nu \frac{k}{\langle k \rangle N}\Theta\beta}\Bigg|_{\Theta=0} > 1. \tag{2.11}$$

Table 2.3: **TDE model statistics summary.** Shown are MLE and 95% confidence interval ($\pm 2\sigma$) for each quantity, followed by exact values for the TDE model system. All predictions are based on a single representative RW with $\ell = 10^4$ steps corresponding to the unit time interval in the TDE model.

TDE Model Network									
N		$\langle k \rangle$		$\langle\langle k_{nn} \rangle\rangle$		$\langle C \rangle$		W/N	
1.01×10^5 $\pm .08 \times 10^5$	(10^5)	8.02 ±.16	(8.14)	64.6 ±4.2	(67.1)	.251 ±.011	(.255)	2.00 ±.05	(2.00)

This defines a critical value of the infection probability given by

$$\beta_c = \frac{\langle k \rangle^2}{\langle k^2 \rangle} \frac{N}{W}, \tag{2.12}$$

above which a sustained epidemic outbreak is possible [49]. We have set $W = 2N$ and $\beta = 7 \times 10^{-1} \gg \beta_c = 6.24 \times 10^{-2}$ in the simulation.

2.3.3 Random Walk Sampling and Epidemic Tracking

We have used a single RW with $\ell = 10^4$ steps and $N_p = 10^3$ pseudotargets to verify the validity of our exponential ansatz (Fig. 2.3(a)) and predict the node degree distribution (Fig. 2.3(b)); several other statistics relevant to the study of epidemics on networks [52] are listed in Table 2.3. In addition, we have tracked time-dependent evolution of the fraction of infected nodes $\rho(t)$ (Fig. 2.3(c)). We have assumed that nodes can be queried much faster than the time scales on which the epidemic spreads, and thus matched ℓ steps of our RW sampling to the unit time interval in the TDE model (Fig. 2.3(c), Table 2.2). Finally, we have predicted β_c using the evolving system's snapshot, again under the assumption that RW sampling is fast compared to the time scales of the epidemics (Fig. 2.3(d)).

The MLE and 95% confidence intervals for the network size, average degree, average nearest-neighbor degree, clustering coefficient, and average packet occupancy from a single representative sampling of $\ell = 10^4$ steps are displayed in Table 2.2 with the full degree distribution shown in Fig. 2.3(b). Additionally the estimation of $\rho(t)$ at the end of each of the intervals alongside the true simulation value are shown in Fig. 2.3(c). Note that during the interval $t = 4...8$ when the epidemic spread is the most rapid, the estimation is below the true value as the statistic is based on a range of values of $\rho(t)$.

As a final observation, we have computed β_c by sampling nodes uniformly for a sample of size ℓ and noted the slow convergence when compared to RW sampling. For uniform node sampling, the estimates for $\langle k\rangle$, $\langle k^2\rangle$, and W/N are given by the algebraic mean:

$$\langle \hat{x}\rangle = \frac{1}{\ell}\sum_x \mathcal{K}_x x, \tag{2.13}$$

where $\mathcal{K}_x$ is the number of occurances of nodes with property x in the sample of size ℓ. The histograms of $\hat{\beta}_c$ values obtained from uniform sampling with the same three sample sizes as in Fig. 2.3(d) are shown in Fig. 2.4. For all three cases, the histograms are clearly more sharply peaked around the true value of β_c for RW sampling.

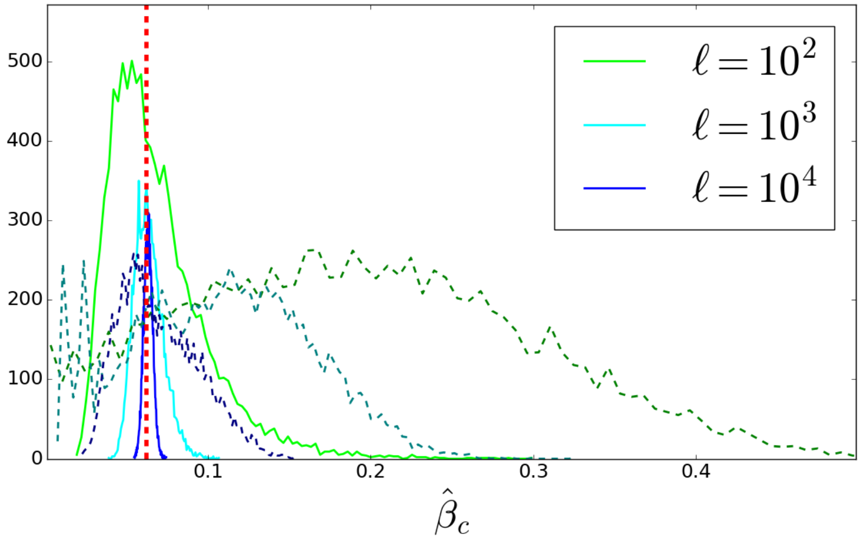

Figure 2.4: **Comparison of RW to uniform-node sampling.**Histograms of β_c MLEs obtained using 10^4 independent runs with $\ell = 10^2$, 10^3, 10^4 steps. Exact value is shown as a vertical, red dashed line. Also included (dashed curves) are the histograms of β_c MLE values for the same three sample sizes obtained through uniform node sampling.

2.4 Modular Network

As an extreme example of network size inference in a highly disjoint system, we have considered two clusters connected by a single link. Accurate prediction of the total network size is still possible in such a system if (i) pseudotargets are chosen as a random subset of all network nodes to minimize correlation effects and (ii) $\langle k \rangle$ is similar in each cluster. The latter requirement can be relaxed if pseudotargets are chosen e.g. among network hubs within a narrow range of k.

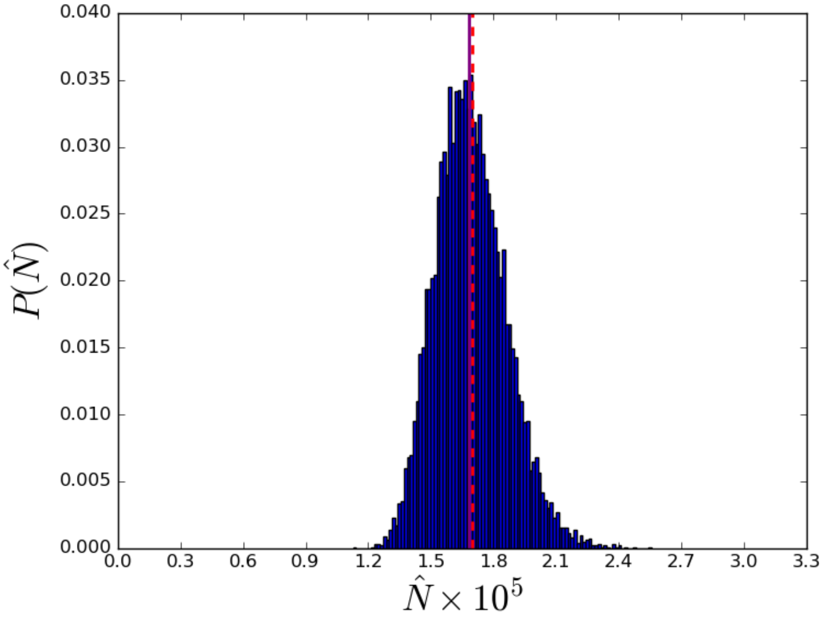

Figure 2.5: **Inference of the network size in a two-component system.** Distribution of $\hat{N}$, the MLE of the total network size, for 10^4 RWs on an unweighted network composed of two clusters that are connected by a single link. Both clusters were generated using a hidden-metric approach (Ref. [50,51]), and contain 1×10^5 ($\langle k \rangle = 8.14$) and 7×10^4 ($\langle k \rangle = 8.54$) nodes, respectively. $N_p = 1.7 \times 10^3$ pseudotargets were uniformly distributed in both clusters. Each random walk had $\ell = 1.7 \times 10^4$ steps. Exact network size is shown as a dashed red line, average predicted value is shown as a purple line.

2.4.1 Two Disconnected Cliques

Taking the case of a highly modular network, and assuming the initial pseudotargets can still be drawn uniformly from the full network, the number of returns to pseudotargets will tend towards

$$\mathcal{K}_p \to \frac{\ell N'_p \langle w \rangle'_p}{N' \langle w \rangle'}, \tag{2.14}$$

where the primes indicate that these are local rather than global quantities. Additionaly $\langle \hat{w} \rangle$ is also being computed from local statistics, and so this estimate will tend towards $\langle w \rangle'$ rather than the global $\langle w \rangle$. The estimator for the full network size will then yield

$$\hat{N} = \frac{\ell N_p \langle w \rangle_p}{\mathcal{K}_p \hat{\langle w \rangle}} = N' \frac{N_p \langle w \rangle_p}{N'_p \langle w \rangle'_p}. \tag{2.15}$$

Provided that enough pseudotargets are placed on the network uniformly, the local fraction of pseudotargets is close to the global fraction, $N'_p/N' \approx N_p/N$. This leaves Eq. (2.15) only biased by the pseudotarget outward rates,

$$\hat{N} = N \frac{\langle w \rangle_p}{\langle w \rangle'_p}. \tag{2.16}$$

If then the average outward rate in this region is close to the global average, the estimator will provide the correct answer for the full network size for $\ell \gg 1$, even if this region is completely disconnected from the network. If this is not the case, the estimator will remain biased. Fortunately this error can be detected if more than a single RW is used. Starting each RW from a random pseudotarget will result in a variety of values for the estimations of each statistic examined in this latter case.

To illustrate this point, we ran 10^4 RWs each of length $\ell = 10^5$ steps with $N_p = 10^3$ pseudotargets on a network consisting of two cliques joined by a single link. To capture the case in which $\langle w \rangle_p \neq \langle w \rangle'_p$, the two clique sizes (which determine the two average outward rates) were set to $N' = 2 \times 10^5$ and $N - N' = 8 \times 10^5$. The values obtained from Eq. (2.15) at the end of each walk formed a bimodal distribution shown in Fig. 2.6 demonstrating that the value of the estimate is highly dependent on which clique the RW started from. As a contrast, we repreated this sampling on a network with equal clique sizes, $N' = N - N' = 5 \times 10^5$, leading to $\langle w \rangle = \langle w \rangle'$ such that the network size was obtained accurately. This can be seen in Fig. 2.7 where the distribution is no longer bimodal.

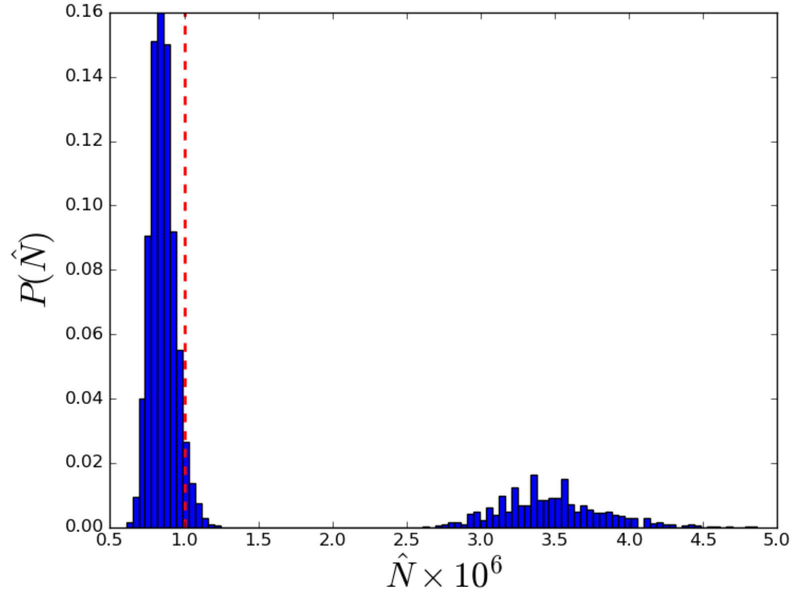

Figure 2.6: $\hat{N}$ **distribution for** 10^4 **RWs on the unequal clique network.**

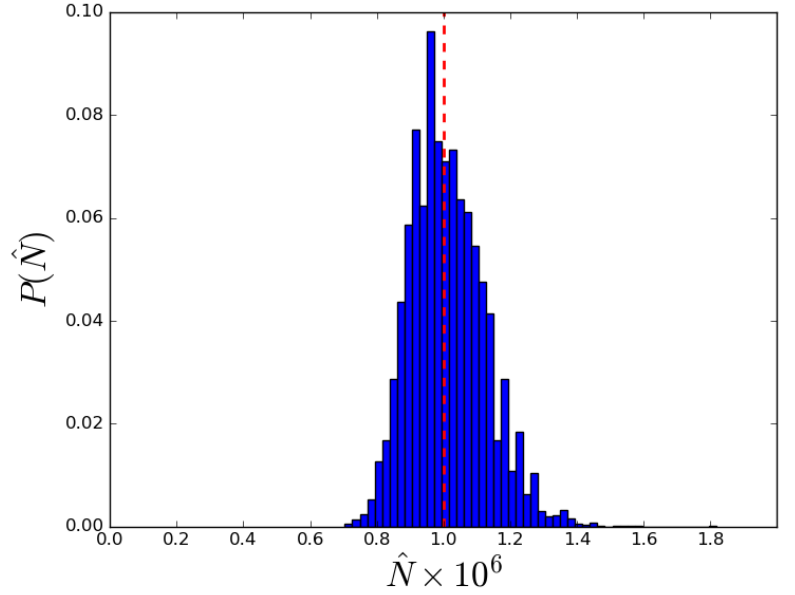

Figure 2.7: $\hat{N}$ **distribution for** 10^4 **RWs on the equal clique network.**

Chapter 3

Random Walk Sampling of Wikipedia

Finally, we have examined the network formed by hyperlinks between English articles on Wikipedia. Links connecting an article to itself were disregarded, multiple links between articles were counted as one, and automatic redirects were disallowed, resulting in an unweighted, undirected, loopless network consisting of all English articles, redirect pages, and disambiguation pages [53]. To assign pseudotargets, the first 5000 pages were drawn from Wikipedia's static HTML dumps. A single randomly chosen link was then taken from each of these pages and the node it pointed to was designated as a pseudotarget, resulting in $N_p = 4769$. This procedure increases the likelihood that the pseudotargets are hubs with a large number of links, facilitating collection of the network statistics since $\mathcal{K}_p$ grows more rapidly [5, 9, 40]. Artificially increasing $\langle k \rangle_p$ in this way significantly augments the rate of $\mathcal{K}_p$ accumulation and so causes the error in the network size estimate to diminish rapidly. In theory, further hubs could be sought to increase the rate at which $\sigma_N \to 0$. However, eventually this would lead to the exponential approximation of the RT distribution to break down. Even so, a pseudotarget set with q_p close to 1 would still yield an accurate network size estimate as is clear upon examination of Eq. (1.34).

We have focused on several statistics that facilitate comparison with known properties of Wikipedia: the size of each page in bytes, ν, and two variables $\chi_r, \chi_d \in \{0, 1\}$ representing whether a page is a redirect or a disambiguation page, respectively. The quantities $\langle \chi_r \rangle$, $\langle \chi_d \rangle$, $\langle \chi_r \chi_d \rangle$, and $\langle \nu_a \rangle \equiv \langle (1 - \chi_r)\nu \rangle$ then give the fraction of redirect pages, disambiguation pages, both redirect and disambiguation pages, and the average storage space in bytes of English articles (Wikipedia excludes redirect pages from its estimates of the number of articles [53]), respectively. The RW was run for $\ell = 5 \times 10^4$

Table 3.1: **Wikipedia statistics.** Shown are MLE and the 95% confidence interval (2σ) for each quantity. All predictions are based on a single trial with $\ell = 5 \times 10^4$ steps.

N	$\langle k \rangle$	$\langle \chi_r \rangle$	$\langle \chi_d \rangle$	$\langle \chi_r \chi_d \rangle$
13.4×10^6 $\pm 1.2 \times 10^6$	47.7 ±.4	.6009 ±.0197	.0399 ±.0047	.0165 ±.0059
$\langle \nu_a \rangle$	$\langle \nu \rangle$	$N(1 - \langle \chi_r \rangle)$	$N \langle \chi_r \rangle$	$N \langle \nu_a \rangle$
2670 ±40 bytes	2720 ±40 bytes	5.35×10^6 $\pm .56 \times 10^6$	8.05×10^6 $\pm .79 \times 10^6$	35.8 ±3.3 GB

steps, with the resulting predictions shown in Table 3.1 and Fig. 3.1.

We find that Wikipedia contains 13.4 million pages, each of which is connected to 48 other pages on average. The majority of Wikipedia pages, 60%, are redirect pages, and 4% are disambiguation pages. We estimate the total number of English articles (including disambiguation pages) to be 5.35 million, and the total number of redirect pages to be 8.05 million, within the confidence intervals of the values reported by Wikipedia: 5.5 and 8.0 million, respectively [54]. We find the total size of English articles in Wikipedia to be 35.8 gigabytes (GB), in reasonable agreement with the Wikipedia statement that text alone accounts for 27.6 GB of the storage space of English articles [55].

Fig. 3.1(a) demonstrates that the assumption of the exponential RT distribution is reasonable for Wikipedia, with some enrichment for short RTs due to the choice of network hubs as pseudotargets. Fig. 3.1(b) shows how the estimate of the total number of Wikipedia pages evolves as $\mathcal{K}_p$ increases. As in many other Internet-based networks [56], the degree distribution of Wikipedia pages is scale-free (Fig. 3.1(c)). In contrast, the distribution of page sizes is not scale-free, and the size of an average Wikipedia page is only 2.7 kB (Fig. 3.1(d), Table 3.1).

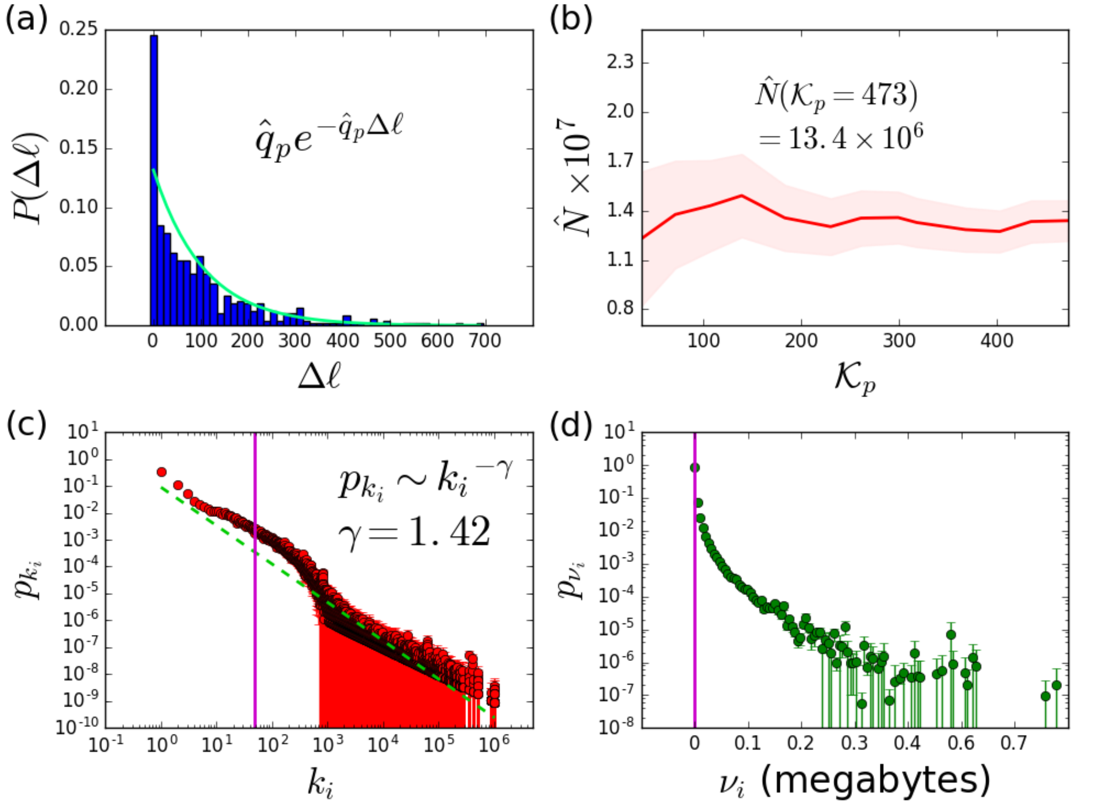

Figure 3.1: **Wikipedia network statistics.** (a) Pseudotarget RT distribution. Equation (1.15) parameterized by $\hat{q}_p$ is shown in cyan. (b) MLE $\pm 2\sigma$ for N as a function of $\mathcal{K}_p$. (c) MLE $\pm 2\sigma$ for the degree distribution of Wikipedia pages of all types. Power-law fit, $p_{k_i} \sim {k_i}^{-\gamma}$, is shown as a green dashed line. Average degree shown as vertical line. (d) MLE $\pm 2\sigma$ for the distribution of Wikipedia page sizes. Average size shown as vertical line.

Chapter 4

Concluding Remarks for Random Walk Sampling on Complex Networks

In conclusion, we have presented a general Bayesian approach to collecting various network statistics, including the size of the network, using RWs that visit only a small fraction of all network nodes. Our approach works for both weighted and unweighted undirected networks, and remains accurate in the presence of loops. Our main assumption, that of the exponentiality of the RT distribution, appears to hold in all the cases we have examined explicitly, and can be relaxed if necessary. Our future work will focus on extending this methodology to directed and time-dependent networks.

Chapter 5

Introduction to Codon Usage Bias

A further application of random walks on networks in the life sciences arises in our investigation of the codon bias. The central dogma of molecular biology states that consecutive triplets of nucleotides called codons are translated into amino acids during protein production [12,13]. As there are 64 codons and 20 amino acids, the translation code is degenerate, with as many as 6 codons translated into a single amino acid. Pronounced differences in synonymous codon usage are observed in any organism for which protein coding sequences are available and therefore codon frequencies can be reliably computed. These genome-wide differences are known as codon bias [14–18]. Since codon usage is one of the most fundamental features of genomes, a quantitative understanding of its evolution is critical to molecular biology.

Because the function of a protein is determined solely by its amino acid sequence, arguably the most basic mechanism for dictating the choice of synonymous codons is neutral evolution on the network formed by single-point mutations between codons subject to a fitness landscape shaped by selective penalties for amino acid mistranslation [19, 20]. The codons which translate into a suboptimal amino acid then act as absorbing states in this first-passage process of sequence evolution. In this approach, non-uniform codon frequencies are produced due to mutational robustness [21] and transition/transversion mutational biases [22].

Another popular explanation for the global codon bias involves selection and postulates that certain codons are translated more efficiently than others, resulting in higher protein production rates and therefore higher cellular growth rates or fitness [20,23–25]. This translation efficiency can be characterized as a balance between translation speed and accuracy [26]: a particular codon may be more rapidly translated due to a higher concentration of the corresponding tRNAs (a hypothesis supported by the correlation between tRNA gene copy numbers and codon frequencies [27]), but may also cause more translation errors. The translation errors can be viewed through the lens of the wobble hypothesis, which states that each codon can be recognized by non-cognate tRNA species, with mispairings that occur at the 3' nucleotide position in the codon [28, 29].

Codon bias has been previously examined through population genetic models which incorporate mutation, selection, and drift in a system of two codon types [23, 30–33]. Since a complete treatment of a multi-allelic mutation-selection-drift model is prohibitively complex, especially in the polymorphic limit [34], previous work has attributed the difference in codon frequencies to a balance between selection and drift, with mutations playing a subordinate role [14]. However, because selection strength has to be inversely proportional to the effective population size to reproduce the observed genomic codon frequencies, this approach leads to the "fine-tuning" problem in which selective advantages of the preferred codons have to vary through many orders of magnitude in order to reflect a broad range of effective population sizes [35]. It is challenging to provide a biophysical explanation for this behavior.

In contrast, our model focuses on the interplay between mutational and selective forces acting on individual codons: the observed codon frequencies emerge as a steady-state balance between mutational forces on one hand, and selection on translation speed and accuracy on the other. We explicitly modeled the evolutionary process on the full 64-codon mutational network in a population of organisms whose fitness is determined by genomic codon content (multi-allelic mutation-selection-drift models are in any case prohibitively complex in the polymorphic limit [34]). Our approach is based on a realistic codon-level mutation model which includes transition/transversion biases and mutational robustness, and allows for non-cognate tRNA-mRNA pairings consistent with the wobble hypothesis. The full model details, as well as our model selection process, is presented and concluded here and in Chapter 6.

Using this selection-mutation framework, we are able to accurately predict genome-wide codon frequencies in a variety of organisms spanning both prokaryotic and eukaryotic domains (Chapter 7). Our predictions of the codon-anticodon pairing rates are largely consistent with previously postulated wobble rules [28] and with the crystallographic analysis of wobble base pairs in the context of the ribosomal decoding center [36]. In Chapter 8, we incorporate Bulmer's biophysical model, which explicitly describes the details of the translation process given a finite ribosomal pool [23], into our approach, and estimate single-nucleotide mutation rates using biophysical model parameters such as ribosomal on-rates and codon translation times. Finally, in Chapter 9 we present the fitting algorithm used to determine the unknown biophysical parameters.

5.1 Biophysical Model of Codon Evolution Overview

We consider the fitness of each organism, w, given the presence of a codon c at a particular genomic location and the optimal amino acid or STOP instruction j at that location, as the product of two terms modeling translation speed and accuracy, respectively (see Section 5.2):

$$w_j(c) = \left(1 - \frac{T_0}{C_c^{\text{eff}}}\right)(1 - s_j(c)) \simeq 1 - \frac{T_0}{C_c^{\text{eff}}} - s_j(c), \tag{5.1}$$

where T_0 sets the overall scale of the selection coefficient in the first term, which penalizes for slow codon translation, and C_c^{eff} is the effective tRNA gene copy number. The approximation in Eq. (5.1) is valid when the two selection terms T_0/C_c^{eff} and $s_j(c)$ are small, as is generally expected for selection on a single codon. Since, according to the wobble hypothesis, non-cognate codon-anticodon pairing is allowed at the 3' codon position, C_c^{eff} is computed as a weighted sum over all possible codon-anticodon pairings,

$$C_c^{\text{eff}} \equiv \sum_{n' \in \{\text{A,U,C,G}\}} r_{n'/n} C_c(n'), \tag{5.2}$$

where $r_{n'/n}$ is the codon-anticodon pairing rate associated with the nucleotide pairing n'/n at the 5' anticodon and 3' codon positions, respectively, and $C_c(n')$ is the corresponding anticodon tRNA gene copy number, which we assume is proportional to the total number of tRNA molecules in the cell. For brevity, we shall refer to $r_{n'/n}$ as "pairing rates" from now on. Note that the pairing rates are defined to be dimensionless and unnormalized.

In the second term on the right-hand side of Eq. (5.1), $s_j(c)$ is the amino-acid-level selection coefficient which penalizes for incorrect amino acid translations due to wobble pairing:

$$s_j(c) = \frac{\sum_{n' \in \{\mathrm{A,U,C,G}\}} r_{n'/n} C_c(n') \bar{s}_j^c(n')}{\sum_{n' \in \{\mathrm{A,U,C,G}\}} r_{n'/n} C_c(n')}, \tag{5.3}$$

where $\bar{s}_j^c(n')$ is either zero when the tRNA bound to codon c is charged with the optimal amino acid j, or a constant penalty, s, for any other amino acid. Thus, our model assumes that all codons in the genome evolve under purifying selection at the amino acid level: as a result, all amino acid substitutions are considered to be deleterious. In other words, each codon position is assigned either an optimal amino acid given by cognate tRNA pairing with the codon currently observed at that genomic position, or a STOP instruction, such that $j = 1, \ldots, 21$. According to Eq. (5.3), even codons that predominantly produce the optimal amino acid will be penalized if there are non-zero pairing rates for translation into suboptimal amino acids. Similarly, a mutation into a codon for which the rates for translation into suboptimal amino acids are enhanced (for example, mutations of a codon which predominantly produces arginine (Arg) into a predominantly non-Arg codon at a position where the optimal amino acid is Arg) is considered deleterious. Since at each codon position evolutionary dynamics depends on the optimal amino acid, we obtain 21 distinct diagonal matrices containing fitness values for each codon, for 20 amino acids and the STOP instruction (i.e., translating stop codons into amino acids is also considered deleterious in our model).

Equation (5.1) implements the idea that additional tRNA gene copies should increase the available pool of tRNA molecules which can be paired with the codon c, reducing translation times and therefore increasing the fitness of the organism (i.e., as C_c^{eff} increases, $w_j(c)$ also increases). However, changes in the tRNA pool may also result in more translation errors, which will be reflected in the increased $s_j(c)$ (Eq. (5.3)).

5.2 Connection Between Fitness and Codon Content

We model the cell's fitness, w, as proportional to the product of its total protein production rate, $P_{\text{tot}}\,(c,q,\ell)$, which depends on the presence of codon c at location ℓ on gene q (explicit dependence on all the other codons is suppressed for brevity), and a mistranslation penalty:

$$w_j(c,q,\ell) \propto P_{\text{tot}}\,(c,q,\ell)\,(1-s_j(c))\,, \tag{5.4}$$

where $s_j(c)$ is the selection coefficient for codon mistranslation, which we assume to be dependent on the codon's genomic location only through the optimal amino acid or STOP instruction, j, at that location (Eq. (5.3)).

The change in P_{tot} upon mutating the current codon, c, at genomic coordinates (q,ℓ) into codon c' is expected to be small compared to the total protein production rate. The new protein production rate, $P_{\text{tot}}(c',q,\ell)$, can then be approximated by a first-order expansion,

$$w_j(c',q,\ell) \propto \left[P_{\text{tot}}(c,q,\ell) + \frac{dP_{\text{tot}}}{dt^{c(q,\ell)}}\left(t^{c'} - t^{c(q,\ell)}\right)\right]\left(1-s_j(c')\right)\,, \tag{5.5}$$

where the single-codon translation time $t^{c'}$ is assumed to be independent of the codon's location, and $t^{c(q,\ell)}$ is the translation time of codon c at genomic coordinates (q,ℓ).

Next, Eq. (5.5) is averaged over all codon positions for which $s_j(c')$ is the same (that is, over all positions which have the same optimal amino acid or STOP instruction j and therefore evolve under the same fitness matrix):

$$w_j(c') = \frac{1}{G}\sum_{q=1}^{G}\frac{1}{|S_q^j|}\sum_{\ell\in S_q^j} w_j(c',q,\ell) \equiv \langle w_j(c',q,\ell)\rangle, \tag{5.6}$$

where G is the total number of genes, S_q^j is the set of codon locations with the same optimal amino acid or STOP instruction j on gene q, and $|S_q^j|$ is the number of such locations. Note that all instances for which $|S_q^j| = 0$ are excluded from the average. We obtain

$$w_j(c') \propto \left[1 + t^{c'}\left\langle \frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle - \left\langle t^{c(q,\ell)}\frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle\right]\left(1 - s_j(c')\right)$$

$$= \left(1 - \left\langle t^{c(q,\ell)}\frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle\right)\left[1 - t^{c'}\frac{\left\langle \frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle}{\left\langle t^{c(q,\ell)}\frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle - 1}\right]\left(1 - s_j(c')\right)$$

$$\propto \left[1 - t^{c'}\frac{\left\langle \frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle}{\left\langle t^{c(q,\ell)}\frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle - 1}\right]\left(1 - s_j(c')\right). \tag{5.7}$$

We model the translation time, $t^{c'}$, as inversely proportional to the tRNA cellular counts:

$$t^{c'} = \frac{\tau}{\sum_{n\in\{\text{A,U,C,G}\}} r_{n/c'_3} V_{\text{cell}}\left[\text{tRNA}_{n+\bar{c}'_{23}}\right]}, \tag{5.8}$$

where V_{cell} is the cell volume, τ is the characteristic time scale for tRNA molecules to be acquired by the ribosome for translation, r_{n/c'_3} are the pairing rates at which tRNAs with n as their 5' anticodon nucleotide bind to the 3' nucleotide of codon c', denoted c'_3 (the other two anticodon nucleotides are always cognate to c'), and $\left[\text{tRNA}_{n+\bar{c}'_{23}}\right]$ are concentrations of tRNAs with anticodon $n + \bar{c}'_{23}$, where $\bar{c}'_{23}$ denotes the second and third nucleotides of the reverse complement of c'. We assume that the tRNA gene copy number, denoted as $C_{n+\bar{c}'_{23}}$, is proportional to the tRNA cellular counts:

$$V_{cell}\left[\text{tRNA}_{n+\bar{c}'_{23}}\right] = \alpha C_{n+\bar{c}'_{23}}, \tag{5.9}$$

where α is a proportionality constant, leading to

$$t^{c'} = \frac{\tau}{\alpha C^{\text{eff}}_{c'}}, \tag{5.10}$$

with the effective gene copy number $C^{\text{eff}}_{c'}$ given by Eq. (5.2). Finally, with

$$T_0 = \frac{\tau}{\alpha}\frac{\left\langle \frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle}{\left\langle t^{c(q,\ell)}\frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}}\right\rangle - 1} \tag{5.11}$$

Eq. (5.7) reduces to Eq. (5.1). The 64 fitness values for each codon, computed using Eq. (5.1) and conditioned on the optimal amino acid or STOP instruction j, provide the diagonal entries of the fitness matrix $\mathbf{W}_j$.

5.3 Codon Mutational Network

To describe mutations between codons, we have adapted the model of [57]. To determine the mutation rates between codons, we assume that detailed balance has been reached in intergenic regions, which are considered to evolve under the influence of mutational forces only [22, 57]:

$$\mu_{c'c}\pi_c = \mu_{cc'}\pi_{c'}, \tag{5.12}$$

where $\mu_{c'c}$ is the mutation rate per generation from the nucleotide trimer c to c', π_c is the steady-state frequency of the nucleotide trimer c, and β is a scale factor. The no-selection assumption is supported by the observation that trimeric nucleotide frequencies are very similar in the intergenic regions of all the species we have examined (Fig. 5.1).

Additionally, two transition/transversion rate biases are included when the trimer substitution involves a pyrimidine-to-pyrimidine ($C \leftrightarrow T$) exchange (κ_1), or a purine-to-purine ($A \leftrightarrow G$) exchange (κ_2). For example, the mutation rate from codon CGT to codon CGC is given by $\beta\kappa_1\pi_{\mathrm{CGC}}$, whereas the CGA→CGC mutation rate is given by $\beta\pi_{\mathrm{CGC}}$. Mutation rates corresponding to multiple nucleotide substitutions are set to zero. Our codon mutational model is an adaptation of the nucleotide substitution model of [57].

5.4 Population Genetics Model

Our selection-mutation approach allows us to predict genome-wide codon frequencies through a steady-state population genetics model. The major features of the approach are illustrated in Fig. 5.2 using *Escherichia coli* as an example. Figure 5.2A shows three initial *E. coli* populations which are genetically identical except for a single codon:

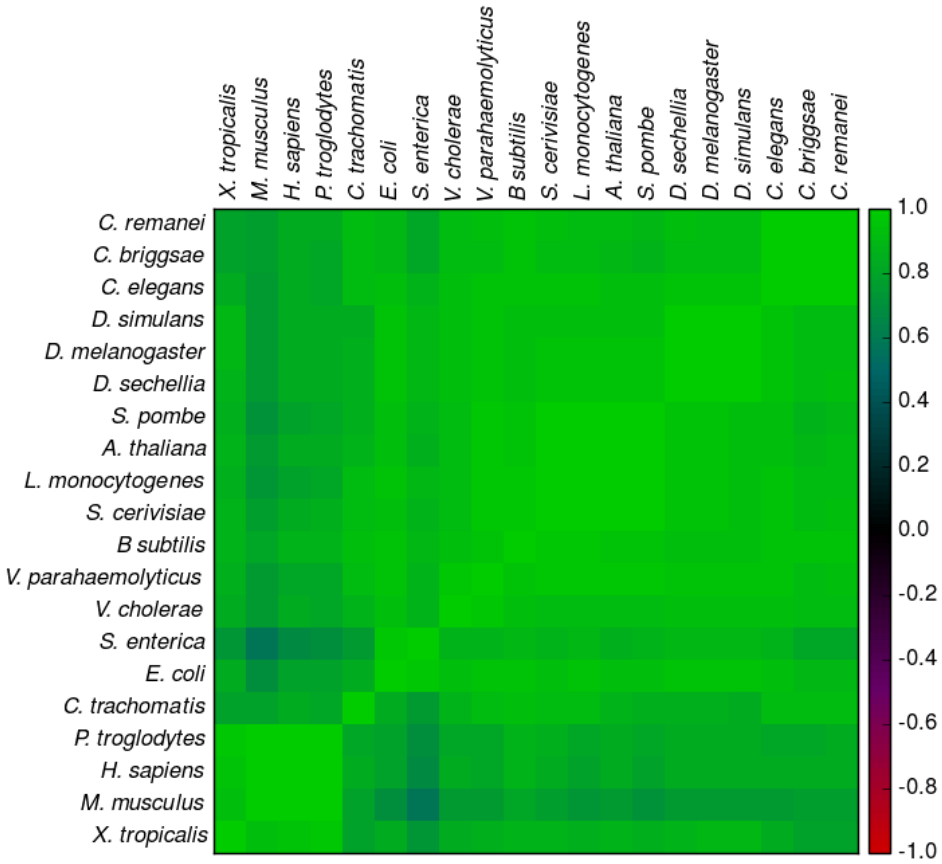

Figure 5.1: **Covariance matrix of nucleotide trimeric frequencies in intergenic regions for all pairs of organisms considered in this study.** Each entry in this matrix shows the Pearson correlation coefficient between pairs of species-specific trimer frequencies. The lowest entry in this matrix, with the Pearson correlation coefficient $\rho = 0.57$, corresponds to the *S. enterica* – *M. musculus* pair.

one population contains the wild-type codon ATA at position 101 in the thrA gene (position 1 is the start codon), whereas the other two contain codons with single-nucleotide mutations: ATG and AAA, respectively. After a fixed period of time, the three progeny populations have different sizes due to differences in their growth rates (Fig. 5.2B). The thrA codon under consideration is at a location which, according to the genetic code and the fact that the wild-type codon is ATA, codes optimally for isoleucine (Ile). Figure 5.2C shows mRNA transcripts produced in the three *E. coli* strains, with colored boxes around codons corresponding to the predominantly translated amino acid in each case: Ile (green), Met (blue), and Lys (red). The lowest-fitness strain has

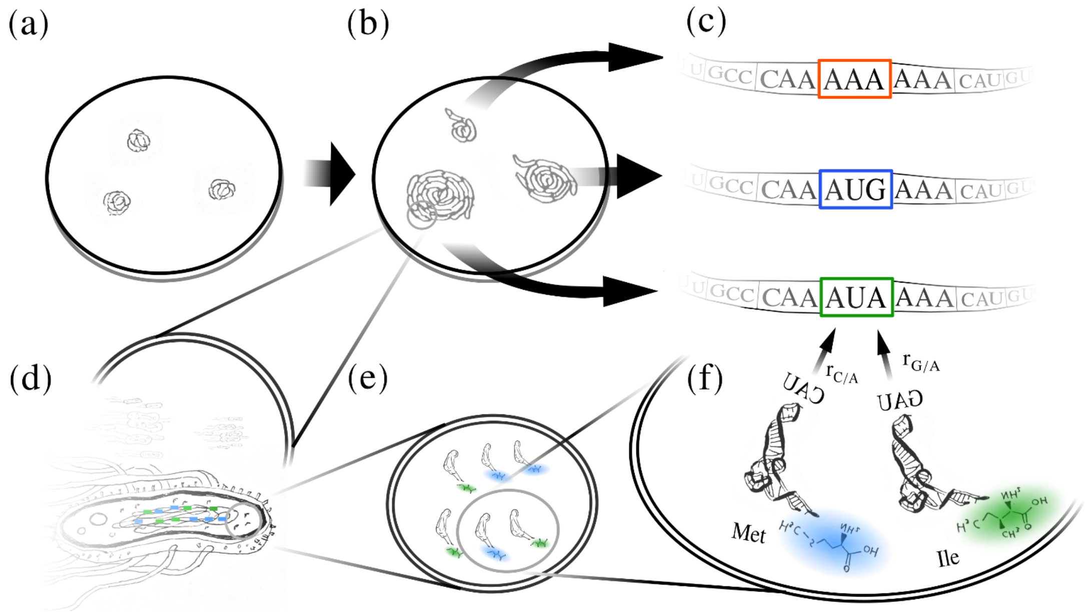

Figure 5.2: **Illustration of the biophysical fitness model on *E. coli* populations**. (a) Three initial *E. coli* populations: one wild-type and two with a single-nucleotide mutation (ATA→ATG, ATA→AAA) at codon 101 in the thrA gene. (b) The same *E. coli* populations after a fixed period of growth. (c) RNA transcripts of the thrA gene from all three strains. Each colored box around the codon in question indicates the amino acid that is primarily translated, with green, blue and red corresponding to Ile, Met, and Lys, respectively. (d) An *E. coli* cell with the tRNA gene copies for Ile (green) and Met (blue) shown as colored rectangles. (e) A magnified portion of the cell with three tRNA molecules charged with Ile (green) and three more charged with Met (blue). The proportions of each type of tRNA molecule roughly match the proportions of gene copies in (d), as assumed in our model. (f) A further magnification of (e) with two representative tRNAs shown in molecular detail. The two tRNA molecules shown, one charged with Met and the other with Ile, are present in the K-12 MG1655 *E. coli* and can bind AUA through wobble pairing, with wobble rates $r_{\mathrm{C/A}}$ and $r_{\mathrm{G/A}}$, respectively. Note that there is no cognate tRNA for this codon.

experienced an ATA→AAA mutation, resulting in a codon which cannot be translated into the optimal amino acid, Ile, even through wobble pairing. In comparison, the ATG strain has higher fitness since it can produce Ile through wobble pairing: however, the ATG codon is primarily translated into Met through cognate pairing. The wild-type ATA strain has the highest fitness as it predominantly produces Ile, even though the cognate tRNA of ATA is in fact not present in *E. coli* (Fig. 5.2D-F).

In order to predict genome-wide codon frequencies, we have employed a mutation-selection population genetics model. We represent codon counts in a population of N organisms as a vector with 64 entries, $|N(t)\rangle$, and evolve the state of the population from one generation to the next using the deterministic equation:

$$|N(t+1)\rangle_j = (\mathbf{I}+\mathbf{M})\mathbf{W}_j|N(t)\rangle_j, \tag{5.13}$$

where $\mathbf{W}_j$ is a diagonal matrix of fitness values conditioned either on the optimal amino acid or the STOP instruction (i.e., $j = 1, \ldots, 21$), $\mathbf{M}$ is the mutation matrix, and $\mathbf{I}$ is the identity matrix. The off-diagonal entries of the mutation matrix, $\mathbf{M}_{c'c}$, are the mutation rates from codon c to c', and diagonal entries are fixed through $\sum_c \mathbf{M}_{c'c} = 0$. Equation (5.13) can be rewritten in terms of the codon frequencies in a population evolving under the same fitness matrix, $|p(t)\rangle_j = |N(t)\rangle_j/\langle 1|N(t)\rangle_j$ ($|1\rangle$ is a vector with 1 in every entry),

$$\begin{aligned}|p(t+1)\rangle_j &= \frac{|N(t+1)\rangle_j}{\langle 1|N(t+1)\rangle_j} = \frac{(\mathbf{I}+\mathbf{M})\mathbf{W}_j|N(t)\rangle_j}{\langle 1|(\mathbf{I}+\mathbf{M})\mathbf{W}_j|N(t)\rangle_j} \\ &= \frac{(\mathbf{I}+\mathbf{M})\mathbf{W}_j|N(t)\rangle_j}{\langle 1|\mathbf{W}_j|N(t)\rangle_j} = \frac{(\mathbf{I}+\mathbf{M})\mathbf{W}_j|p(t)\rangle_j}{\langle 1|\mathbf{W}_j|p(t)\rangle_j}.\end{aligned} \tag{5.14}$$

Eventually these frequencies will reach a steady-state $|p^{ss}\rangle_j$ determined by

$$(\mathbf{I}+\mathbf{M})\mathbf{W}_j|p^{ss}\rangle_j = \bar{w}_j|p^{ss}\rangle_j, \tag{5.15}$$

where $\bar{w}_j = \langle 1|\mathbf{W}_j|p^{ss}\rangle_j$ is the average fitness of the corresponding population.

Finally, if each fitness matrix $\mathbf{W}_j$ operates at C_j codon locations in the genome, steady-state codon frequencies are given by the genome-wide average:

$$|p^{ss}\rangle_{\text{gen}} = \frac{\sum_j C_j |p^{ss}\rangle_j}{\sum_j C_j}, \tag{5.16}$$

where each $|p^{ss}\rangle_j$ is found using Eq. (5.15) with the corresponding $\mathbf{W}_j$. Note that the mutation rates are assumed to be independent of the fitness matrix j, yielding a universal $\mathbf{M}$ for each species.

5.5 Exact Solution in the Two-Codon Case

In order to gain insight into our biophysical model and its dependence on the various model parameters, here we provide an exact solution for a simplified system which consists of two codons, each of which corresponds to a distinct optimal amino acid. With 2 fitness matrices, Eq. (5.15) yields

$$\begin{aligned}(\mathbf{I}+\mathbf{M})\mathbf{W}_i|p^{ss}\rangle_i &= \begin{pmatrix} 1-\mu_{21} & \mu_{12} \\ \mu_{21} & 1-\mu_{12} \end{pmatrix} \begin{pmatrix} w_1^i & 0 \\ 0 & w_2^i \end{pmatrix} \begin{pmatrix} p_1^i \\ p_2^i \end{pmatrix} \\ &= \begin{pmatrix} (1-\mu_{21})w_1^i & \mu_{12}w_2^i \\ \mu_{21}w_1^i & (1-\mu_{12})w_2^i \end{pmatrix} \begin{pmatrix} p_1^i \\ p_2^i \end{pmatrix} = \bar{w}^i \begin{pmatrix} p_1^i \\ p_2^i \end{pmatrix} \end{aligned} \tag{5.17}$$

where w_c^i and p_c^i denote the fitness and the steady-state frequency of codon $c \in \{1,2\}$ evolving under fitness matrix $\mathbf{W}_i$ ($i \in \{1,2\}$), respectively, and $\bar{w}^i = w_1^i p_1^i + w_2^i p_2^i$ is the corresponding mean fitness. The steady-state frequencies are then given by

$$\begin{aligned} p_1^i = \frac{1}{2} - \frac{\mu_{21}w_1^i + \mu_{12}w_2^i}{2\Delta w^i} \\ + \frac{1}{2\Delta w^i}\sqrt{(\Delta w^i)^2 - 2(\mu_{21}w_1^i - \mu_{12}w_2^i)\Delta w^i + (\mu_{21}w_1^i + \mu_{12}w_2^i)^2} \end{aligned} \tag{5.18}$$

and

$$p_2^i = \frac{1}{2} + \frac{\mu_{21} w_1^i + \mu_{12} w_2^i}{2\Delta w^i} - \frac{1}{2\Delta w^i}\sqrt{(\Delta w^i)^2 - 2(\mu_{21} w_1^i - \mu_{12} w_2^i)\Delta w^i + (\mu_{21} w_1^i + \mu_{12} w_2^i)^2}, \quad (5.19)$$

where $\Delta w^i = w_1^i - w_2^i$.

If there are C_i genomic locations evolving under fitness matrix $\mathbf{W}_i$, the genome-wide frequencies for each codon c are given by Eq. (5.16):

$$p_{c,\mathrm{gen}} = \frac{C_1 p_c^1 + C_2 p_c^2}{C_1 + C_2}. \quad (5.20)$$

To examine the dependence of steady-state frequencies on the model parameters β, κ, s, and T_0, the following fitnesses and mutation rates are assumed:

$$\begin{aligned}
\mu_{21} &= \beta\kappa\pi_2, \quad \mu_{12} = \beta\kappa\pi_1, \\
w_1^1 &= \left(1 - \frac{T_0}{C_1^{\mathrm{eff}}}\right), \quad w_2^1 = \left(1 - \frac{T_0}{C_2^{\mathrm{eff}}}\right)(1-s), \\
w_1^2 &= \left(1 - \frac{T_0}{C_1^{\mathrm{eff}}}\right)(1-s), \quad w_2^2 = \left(1 - \frac{T_0}{C_2^{\mathrm{eff}}}\right),
\end{aligned} \quad (5.21)$$

where π_c is the steady-state frequency of codon c in the absence of selection. Note that mutations between the two codons are assumed to be transitions. With these specifications, p_1^1 in Eq. (5.18) becomes

$$p_1^1 = \frac{1}{2} - \frac{\beta\kappa\left(1 - \frac{T_0\pi_2}{C_1^{\mathrm{eff}}} - \frac{T_0\pi_1}{C_2^{\mathrm{eff}}}\right)}{2\left[T_0\frac{C_1^{\mathrm{eff}} - C_2^{\mathrm{eff}}}{C_1^{\mathrm{eff}} C_2^{\mathrm{eff}}} + s\left(1 - \frac{T_0}{C_2^{\mathrm{eff}}}\right)\right]}$$

$$+\frac{1}{2}\sqrt{1 - 2\frac{\beta\kappa\pi_2\left(1 - \frac{T_0}{C_1^{\mathrm{eff}}}\right) - \beta\kappa\pi_1\left(1 - \frac{T_0}{C_2^{\mathrm{eff}}}\right)(1-s)}{T_0\frac{C_1^{\mathrm{eff}} - C_2^{\mathrm{eff}}}{C_1^{\mathrm{eff}} C_2^{\mathrm{eff}}} + s\left(1 - \frac{T_0}{C_2^{\mathrm{eff}}}\right)} + \left[\frac{\beta\kappa\left(1 - \frac{T_0\pi_2}{C_1^{\mathrm{eff}}} - \frac{T_0\pi_1}{C_2^{\mathrm{eff}}}\right)}{T_0\frac{C_1^{\mathrm{eff}} - C_2^{\mathrm{eff}}}{C_1^{\mathrm{eff}} C_2^{\mathrm{eff}}} + s\left(1 - \frac{T_0}{C_2^{\mathrm{eff}}}\right)}\right]^2}. \quad (5.22)$$

To $\mathcal{O}(\beta)$, $\mathcal{O}(s)$, and $\mathcal{O}(T_0)$, Eq. (5.22) simplifies to

$$p_1^1 \approx \frac{1}{2} - \frac{\kappa/2}{\frac{T_0}{\beta}\frac{C_1^{\text{eff}}-C_2^{\text{eff}}}{C_1^{\text{eff}}C_2^{\text{eff}}} + \frac{s}{\beta}} + \frac{1}{2}\sqrt{1 - \frac{2\kappa(\pi_2 - \pi_1)}{\frac{T_0}{\beta}\frac{C_1^{\text{eff}}-C_2^{\text{eff}}}{C_1^{\text{eff}}C_2^{\text{eff}}} + \frac{s}{\beta}} + \frac{\kappa^2}{\left[\frac{T_0}{\beta}\frac{C_1^{\text{eff}}-C_2^{\text{eff}}}{C_1^{\text{eff}}C_2^{\text{eff}}} + \frac{s}{\beta}\right]^2}}. \qquad (5.23)$$

A similar expression can be obtained for p_1^2, from which $p_2^1 = 1-p_1^1$ and $p_2^2 = 1-p_1^2$ follow by normalization. Note that under this approximation all steady-state frequencies, including the genome-wide frequencies $p_{c,\text{gen}}$ in Eq. (5.20), only depend on the ratios s/β and T_0/β.

Chapter 6

Model Selection

To determine which biophysical factors contribute most to the codon bias and what level of detail is necessary to predict genome-wide codon frequencies, we have constructed a hierarchy of models which include from 3 to 19 free parameters (see Table 6.1 for detailed descriptions), and fit the models to *E. coli* (K-12 MG1655) genomic data. Specifically, each model was fit to minimize the L^1 distance:

$$L^1 = \frac{1}{2}\sum_{c=1}^{64} |\hat{p}_c - p_c|, \tag{6.1}$$

where $\hat{p}_c$ and p_c are predicted and observed genome-wide codon frequencies, respectively (see Chapter 9 for a detailed description of the global optimization algorithm). Each model was subjected to 5-fold cross-validation: all genomic codons were randomly divided into 5 subsets of equal size, and the model was fitted separately on each subset, with $\bar{L}^1$ denoting the average L^1 distance resulting from these 5 fits. For the purposes of cross-validation, L^1 distances were computed between codon frequencies predicted by each of the 5 fits and codon frequencies observed in each of the other 4 codon subsets which were not used to fit the model in the current round. The cross-validation score, $\bar{L}^1_{\mathrm{CV}}$, was then computed by averaging first over the other 4 subsets left out of the current fit and, finally, over the 5 independent fits.

6.1 Codon Bias from Mutational Network Structure

The first model we have examined is a minimal model which does not consider wobble pairing or the fitness penalty for slow translation and therefore only includes transition/transversion mutational parameters κ_1 and κ_2 and the amino acid selection parameter s/β. Note that the codon frequencies are affected only by the ratio of the amino acid selection coefficient s, which penalizes translation into suboptimal amino acids, and the overall mutation scale β (see Section 5.5 for an additional discussion). Under this 3-parameter model, genome-wide codon frequencies are determined by a combination of mutation rate biases and mutational proximity to deleterious sequences (i.e., mutational robustness). We illustrate this point using 6 Arg codons as a representative example (Fig. 6.1A,B). Under the 3-parameter model there is a marked enrichment of the frequencies of CGC, CGG, and CGT codons and a suppression of AGA and AGG codon frequencies, even though all 6 codons have the same fitness. These trends, with the exception of the CGG enrichment, match genome-wide codon frequency data, and are not present in the intergenic regions (Fig. 6.1A).

6.2 Addition of Selection for Translation Speed and Accuracy

Next, we examined a family of models which, in addition to κ_1, κ_2, and s/β, include a fitness penalty for slow codon translation, T_0/β, with T_0 defined in Eq. (5.1). In addition, each model in the hierarchy includes an increasingly diverse set of pairing rates (Table 6.1). Specifically, the 5-parameter model has a single parameter describing all non-cognate pairing rates. In this model, cognate pairings are assumed to occur at a rate of $r_{n'/n} = 1$, while four pairings are suppressed ($r_{n'/n} = 0$) based on the crystallographic analysis of wobble base pairs in the context of the ribosomal decoding center [36]. The remaining 8 rates are described by a single free parameter, r. The 7-parameter model replaces this single parameter with three rates: r_0, which accounts for pairings across nucleotide types (purine to pyrimidine) expected to be closest to cognate pairing; r_1, which characterizes all same-base pairings that are not already suppressed on the basis of crystallographic evidence; and r_2, which accounts for the two remaining pairings. In the 12-parameter model, rates for 8 pairings that are neither cognate nor suppressed are allowed to vary individually. In the 16-parameter model, the assumption that some of the wobble pairings are suppressed is relaxed, resulting in 4 additional pairing rates. Finally, in the 19-parameter model the assumption that all cognate pairings have a rate of $r_{n'/n} = 1$ is relaxed, and each of the possible 16 pairings is assigned an individual rate. Since there is now a degeneracy in the model related to the fact that the T_0/C_c^{eff} ratio remains invariant in Eq. (5.1) if both T_0 and all wobble rates are scaled by the same factor, we have chosen to set $T_0/\beta = 1$, resulting in 19 independent parameters. An alternative approach in which one of the cognate rates was set to 1.0 and T_0/β was allowed to vary yielded numerically inferior solutions.

6.3 Model Selection Procedure

Since 63 independent codon frequencies are fit to models containing from 3 to 19 independent parameters, it is important to ensure that there is no overfitting. Figure 6.1C demonstrates the quality-of-fit scores $\bar{L}^1$ and $\bar{L}^1_{\mathrm{CV}}$ for each of the models described above. A standard way of checking the extent of overfitting, 5-fold cross-validation, has limited applicability here since codon frequencies are very similar in all 5 subsets, as manifested by the high degree of similarity between $\bar{L}^1$ and $\bar{L}^1_{\mathrm{CV}}$ in all Fig. 6.1C fits. Thus, to investigate the issue of overfitting from a different angle, we have carried out model fits not only on genomic codon frequencies (blue lines), but also on synthetically generated data for which models previously fit on genomic data were used to generate artificial codon counts. These counts were then used in a subsequent round of model fitting (red lines). The idea is to provide a score baseline in which a given model type is employed to both generate the synthetic data and carry out subsequent parameter inference. This two-step procedure leads to consistent recovery of all model parameters used in generating the synthetic data (Table 9.1). As can be seen in Fig. 6.1C, there is no trend in the model scores of fits on synthetic data as the model complexity increases, and for each model type genomic fit scores are significantly above synthetic fit scores, indicating the absence of overfitting. Furthermore, model scores of fits on genomic data improve with model complexity, suggesting that overall increasing the model complexity is beneficial. Note however that the genomic scores become worse in going from the 3- to 5-parameter model, showing that an increase in the number of model parameters does not necessarily guarantee an improvement in fitting performance.

We have also investigated the effects of intentional overfitting on synthetic data. To this effect, an “old” model with lower complexity, previously fit on genomic data, was used to generate the codon counts, which were subsequently used to fit a “new,” higher-complexity model (red bars in Fig. 6.1D). Surprisingly, this overfitting always resulted in worse model scores, again underscoring that increasing model complexity does not necessarily lead to better scores, due to both essential differences in model parameterization and the lack of numerical convergence. However, this effect becomes very slight on the higher-complexity end of the model spectrum. To investigate this issue further, we have generated synthetic data using the 7-parameter model, and fit all model types to it (Fig. 6.2). We observe that, as expected, lower-complexity models are not able to fit the synthetic dataset as well as the “native” 7-parameter model. Furthermore, fitting more complex models does not offer any marked improvements in model scores.

In contrast to the results based on synthetic data, there is a significant improvement in model performance on genomic data with each increase in model complexity (blue bars in Fig. 6.1D), with a sole exception of the 3- and 5-parameter model pair. However, the gains in model scores diminish gradually, indicating that increasing the number of parameters beyond 19 is unlikely to lead to further significant improvements in model performance. Given that the 19-parameter model yields the best performance, we have chosen it for all further analysis carried out in this study. The model’s predictions in *E. coli* are shown in Fig. 6.3 and Fig. 7.3A, indicating that our approach is capable of reproducing all the major features observed in genome-wide codon frequencies in this organism.

Table 6.1: **Hierarchy of models fit to *E. coli* genomic data.** All models share the same three parameters κ_1, κ_2, and s/β, and all except the 3-parameter model also include T_0/β. The pairing rates are parameterized according to various categories of nucleotide base pairing: Watson-Crick (Cognate); disallowed according to [36] (Suppressed); different in type, purine or pyrimidine, and are not disallowed or cognate (Alternate); and two nucleotides of the same type that are not disallowed (Same base). ρ: Pearson correlation coefficient between predicted and observed frequencies, p: the corresponding p-value.

Parameter	Description
κ_1	Pyrimidine transition/transversion rate bias (T→C and C→T).
κ_2	Purine transition/transversion rate bias (G→A and A→G).
s/β	Ratio of amino acid selection coefficient to mutation scale.
T_0/β	Ratio of translation speed selective penalty to mutation scale.

# model parameters	Wobble rate	Nucleotide pairing	Paired nucleotides (anticodon 5'/codon 3')		Model prediction ρ (p-value)
3	1	Cognate:	A/U, C/G, G/C, and U/A.		0.79 (1.1×10^{-14})
	0	All else.			
5	1	Cognate:	A/U, C/G, G/C, and U/A.		0.79 (6.2×10^{-15})
	0	Suppressed:	C/C, C/U, G/A, and G/G.		
	r	All else:	A/A, A/C, A/G, C/A, G/U, U/C, U/G, and U/U.		
7	1	Cognate:	A/U, C/G, G/C, and U/A.		0.76 (1.9×10^{-13})
	0	Suppressed:	C/C, C/U, G/A, and G/G.		
	r_0	Alternate:	A/C, C/A, G/U, and U/G.		
	r_1	Same base:	A/A and U/U.		
	r_2	All else:	A/G and U/C.		
12	1	Cognate:	A/U, C/G, G/C, and U/A.		0.86 (1.6×10^{-19})
	0	Suppressed:	C/C, C/U, G/A, and G/G.		
	$r_{\mathrm{A/A}}$	All else:	A/A		
	$r_{\mathrm{A/C}}$		A/C		
	$\vdots$		$\vdots$	8 parameters	
	$r_{\mathrm{U/U}}$		U/U		
16	1	Cognate:	A/U, C/G, G/C, and U/A.		0.93 (2.0×10^{-29})
	$r_{\mathrm{A/A}}$	All else:	A/A		
	$r_{\mathrm{A/C}}$		A/C		
	$\vdots$		$\vdots$	12 parameters	
	$r_{\mathrm{U/U}}$		U/U		
19	$r_{\mathrm{A/A}}$		A/A		0.97 (4.4×10^{-41})
	$r_{\mathrm{A/C}}$		A/C		
	$\vdots$		$\vdots$	16 parameters	
	$r_{\mathrm{U/U}}$		U/U		

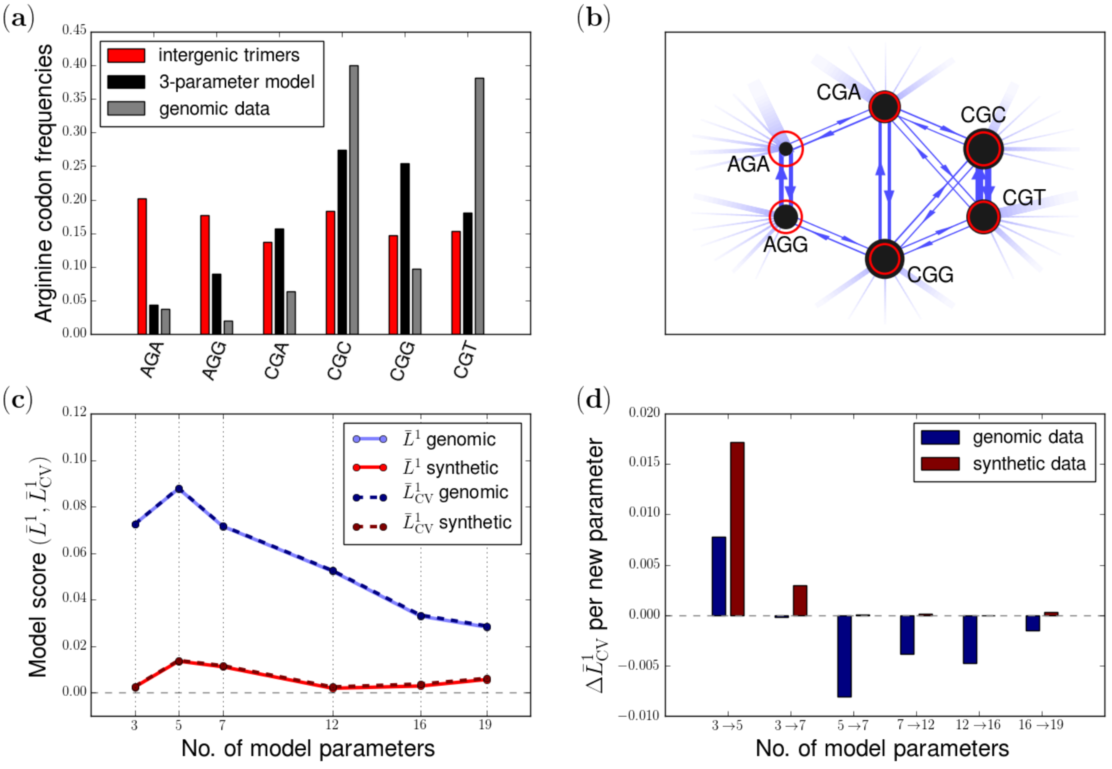

Figure 6.1: **Prediction of genome-wide codon frequencies in *E. coli.*** (a) Codon frequencies of the arginine (Arg) group predicted by the 3-parameter model (black) and found in coding regions (grey), and nucleotide trimer frequencies in the intergenic regions (red). (b) The single-point mutational network formed by the codons which translate into Arg according to the standard genetic code. The width of each line is proportional to the mutation rate, with an arrow indicating the direction of mutation. The fading lines represent all mutation rates from Arg to the corresponding non-Arg codons. The size of each circle indicates the frequency at which each codon sequence occurs in intergenic trimers (red) and when mutation and selection against non-Arg codons are taken into account (3-parameter model; black). (c) Model scores $\bar{L}^1$ (solid lines) and $\bar{L}^1_{\rm CV}$ (dashed lines) as a function of model complexity. Each model was fit to genomic data (blue lines) and synthetic data (red lines). (d) Normalized difference of $\bar{L}^1_{\rm CV}$ model scores, $\Delta\bar{L}^1_{\rm CV} = (\bar{L}^1_{\rm CV}(\text{new}) - \bar{L}^1_{\rm CV}(\text{old}))/(N_{\rm new} - N_{\rm old})$, in going from a less complex ("old") to a more complex ("new") model. $N_{\rm new}$ and $N_{\rm old}$ denote the number of model parameters in the old and new models, respectively.

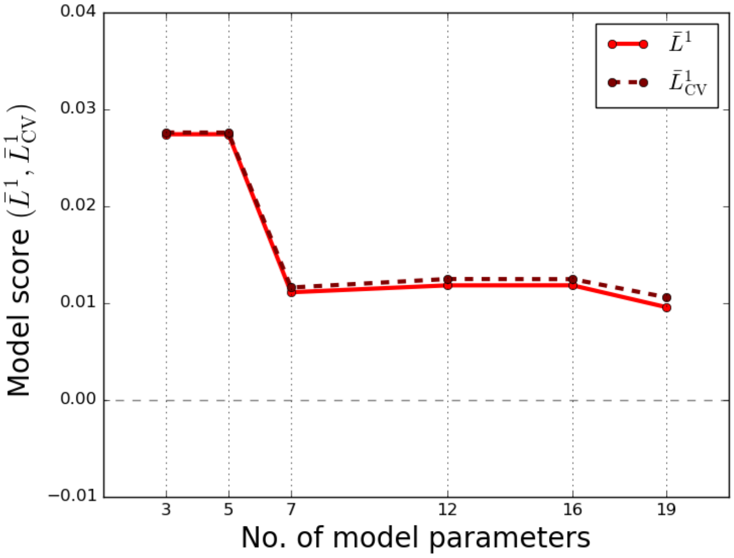

Figure 6.2: **Model scores for all model types in the hierarchy fitted to synthetic data generated by the 7-parameter model.** Fitting scores $\bar{L}_1$ (solid lines) and cross-validation scores $\bar{L}^1_{\mathrm{CV}}$ (dashed lines) are shown as a function of model complexity. All model parameters in the 7-parameter model used to generate the synthetic data were set to values previously found in fitting the model to the codon frequencies from the *E. coli* genome (Table 9.1).

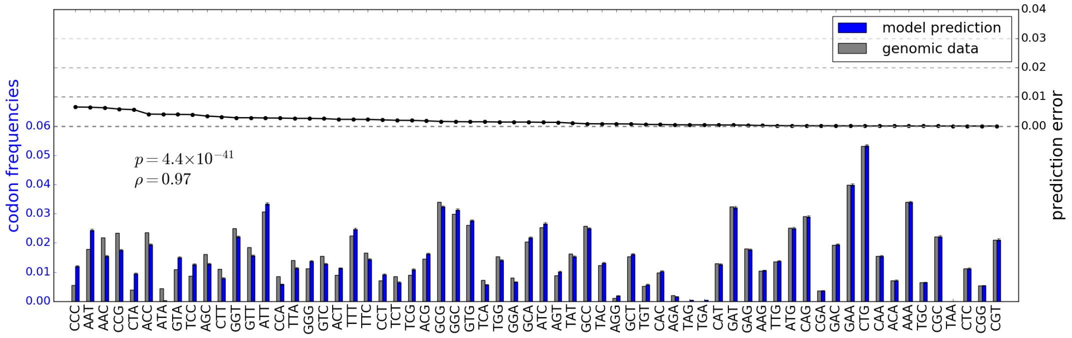

Figure 6.3: **Prediction of codon frequencies in *E. coli.*** Codon frequencies predicted using the 19-parameter model (blue), and genome-wide frequencies observed in *E. coli* (grey). All codons are sorted by the absolute magnitude of the prediction error, defined as the absolute magnitude of the difference between predicted, $\hat{p}_c$, and observed, p_c, frequencies of each codon c: $|\hat{p}_c - p_c|$. The Pearson correlation coefficient ρ between predicted and observed frequencies is also shown, along with the corresponding p-value.

Chapter 7

Multispecies Analysis

We have fit the 19-parameter model to 20 organisms spanning both unicellular and multicellular life forms (Fig. 7.1; see Section 7.5 for details of genomic data acquisition). A representative subset of these organisms is displayed in Fig. 7.2 where a metric for codon bias has been computed on 50 randomly selected genes following the style of Ref. [24]. The metric, known as the relative synonymous codon usage (RSCU), is defined for each codon, c, as [58]

$$\mathrm{RSCU}_c \equiv \frac{\mathcal{C}_c}{\frac{1}{n}\sum_{c'} \mathcal{C}_{c'}}, \tag{7.1}$$

where $\mathcal{C}_c$ is the number of occurances of codon c, n is the number of synonymous codons according to the standard genetic code, and the sum is over all synonymous codons.

For each of the 20 species, the model fits the data to a high degree of accuracy, with the Pearson correlation coefficients in the $[0.80, 0.98]$ range, with an average of 0.91 (Fig. 7.3). Distributions of these parameters are summarized in Fig. 7.4.

7.1 Consistencies in the Transition/Transversion Rate Biases

As might be expected, the values of the two transition/transversion rate biases, κ_1 and κ_2, are fairly conserved, especially in eukaryotes, with larger values generally found in bacteria (Fig. 7.5, Fig. 7.4A). This observation is consistent with the fact that trimeric nucleotide frequencies found in intergenic regions, which on average are likely to evolve only under the influence of mutational forces, are nearly organism-independent (Fig. 5.1). The values of the κ_1 and κ_2 biases are strongly correlated with each other, with $\kappa_1 > \kappa_2$ in all cases. Note that the biases are not always > 1, in agreement with

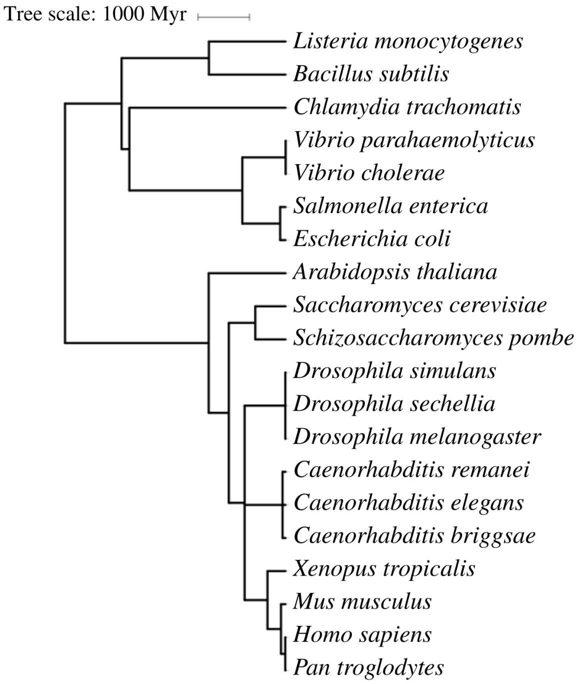

Figure 7.1: **Phylogenetic relationships between all organisms included in this study.** The divergence times between all species examined in this study were set to the estimated values reported in the TimeTree database [59, 60]. These divergence times were then used to construct the phylogenetic tree via the Interactive Tree Of Life [61].

a previously reported result [62].

7.2 Comparison of Codon Usage Bias Influences

We observe strong selection against missense mutations ($s/\beta = 5.84$ on average), indicating that amino acids translated from the genomic codons on the basis of the standard genetic code are generally optimal and their mutations are deleterious (Fig. 7.4B). The value of the selection coefficient s is closely correlated with the distribution of $s_j(c)$ values in each organism (Fig. 7.6A), indicating that it is a good measure of the strength of selection against amino acid mistranslations. Moreover, the strength of selection for the speed of codon translation is generally weaker than the strength of mutational forces, as measured by the overall mutational scale β, although there are notable exceptions (Fig. 7.6B). Correspondingly, in the majority of cases, selection for mistranslation dominates selection for translation speed (Fig. 7.6C).

7.3 Trends in Wobble Rate Numerical Results

We have found that in all organisms the rates corresponding to the A/G, C/A, C/C, G/A and G/G pairings are vanishingly small compared to all other rates (Fig. 7.4C). According to crystallographic evidence [36], C/U, C/C, G/A, and G/G pairings should be sterically disallowed, which is consistent with our findings except for C/U, for which only 3 out of the 20 organisms yield non-vanishing $r_{\mathrm{C/U}}$ rates: *S. pombe*, *V. cholerae*, and *A. thaliana*. Additionally, purine-pyrimidine pairings are consistently assigned higher rates than purine-purine and pyrimidine-pyrimidine pairings, with cognate pairings being predominant compared to non-cognate pairings: for example, averages across all species of the $\overline{r_{\mathrm{A/A}}}$, $\overline{r_{\mathrm{A/C}}}$, $\overline{r_{\mathrm{A/G}}}$, and $\overline{r_{\mathrm{A/U}}}$ pairing rates are 0.9, 3.3, 0.3, and 8.7, respectively. However, a notable exception is the $r_{\mathrm{G/U}}$ rate, which is considerably higher than $r_{\mathrm{G/C}}$ and in fact assumes unrealistically large values for $\sim$ 50% of the species considered. We do not have a satisfactory explanation for this finding at the moment.

7.4 Correlations Between Various Biophysical Parameters

Finally, we have examined a matrix of correlations among 19 model parameters and several additional key values characterizing either the genome (genome size, total number of codons, total number of genes) or the population (effective population size) (Fig. 7.7). We find that, as expected, the genome size, the total number of codons and the total number of genes are all correlated with each other and anti-correlated with the effective population size and the κ_1, κ_2 mutational biases, the latter observation being consistent with the fact that these biases are higher in prokaryotes (Fig. 7.5). In contrast, the selection coefficient s/β is not strongly correlated with any other parameter, including the effective population size. Finally, we observe that some of the pairing rates (e.g. $r_{\mathrm{A/A}}$ and $r_{\mathrm{A/G}}$) are strongly correlated with each other, reducing the effective number of model parameters.

7.5 Genome Sequences and Annotation

All genomic codon, amino acid, and intergenic nucleotide trimer frequency information was extracted for each species from sequence and annotation data in the GenBank file format, downloaded from the NCBI database on 07.13.2018. tRNA gene copy numbers were obtained from the GtRNAdb database [68, 69]. To compute intergenic trimer frequencies, we have removed all nucleotide sequences corresponding to known features, leaving only DNA segments with no currently known functions.

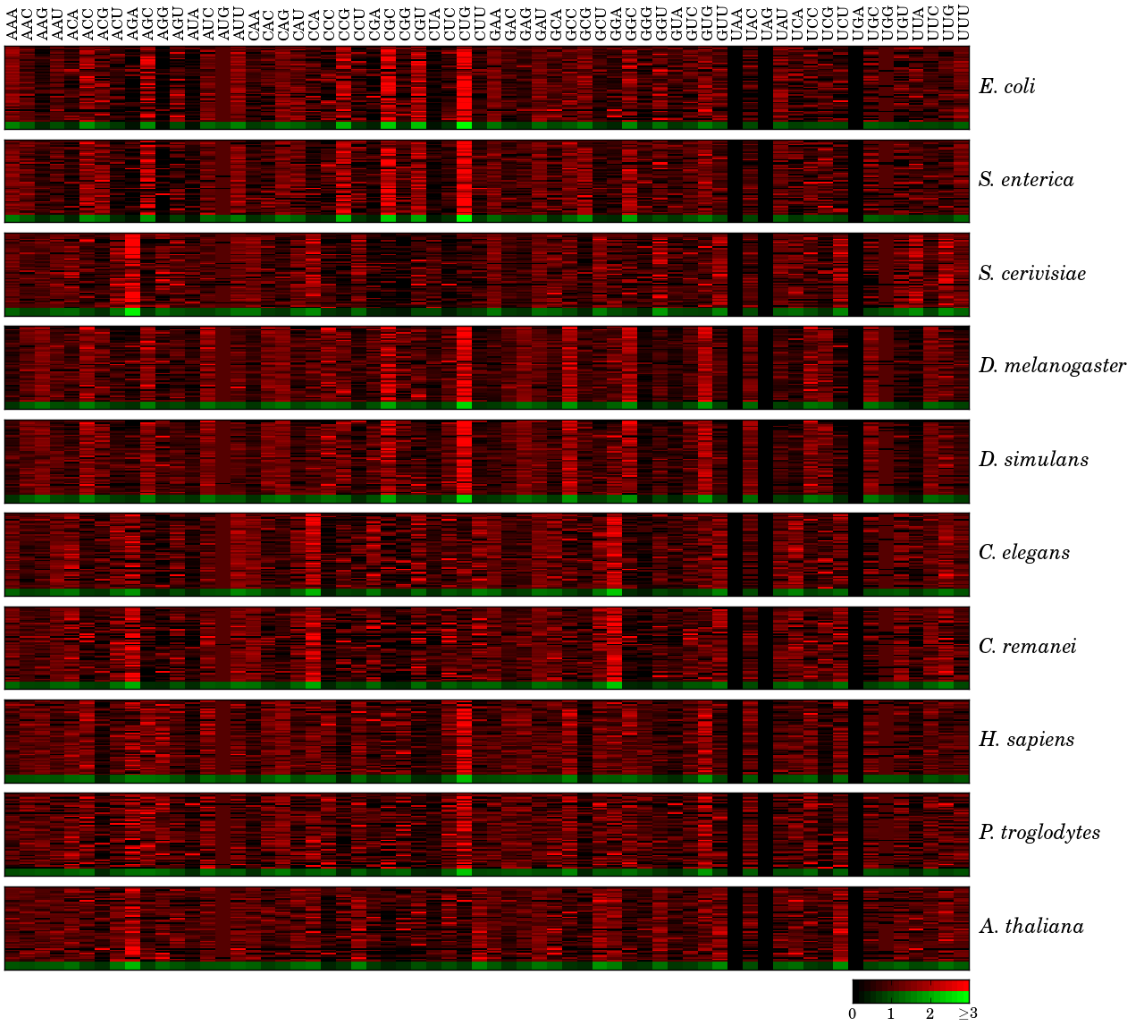

Figure 7.2: **Codon bias diversity across 10 representative species.** The value of RSCU computed for each codon in 50 randomly selected genes (red entries) and for the genome-wide frequencies of codons (green entries). Brighter entries correspond to more bias as indicated by the colorbar. A value of 1 for RSCU corresponds to no bias.

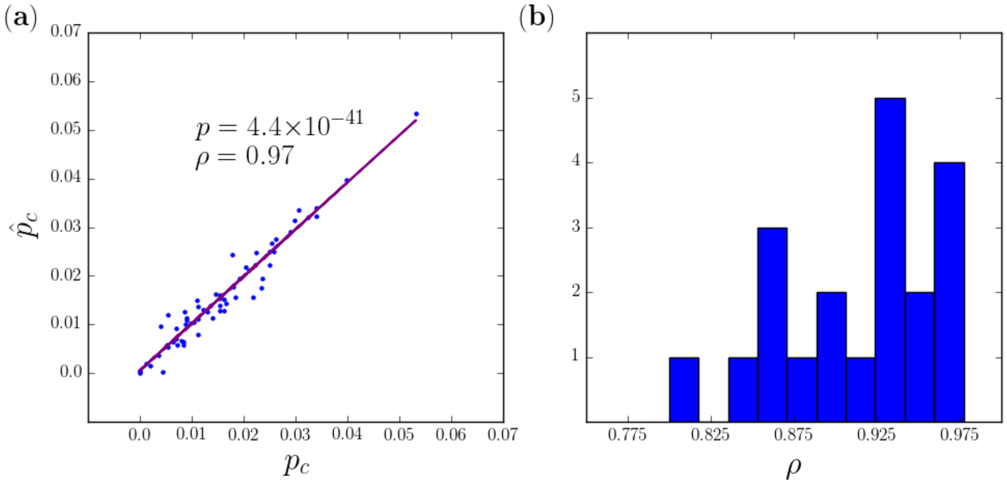

Figure 7.3: **19-parameter model performance on genomic data.** (a) Predicted versus genomic codon frequencies for *E. coli*, with the Pearson correlation coefficient and the corresponding p-value. (b) Distribution of Pearson correlation coefficients between predicted and observed genome-wide codon frequencies for all 20 species included in this study (Fig. 7.1).

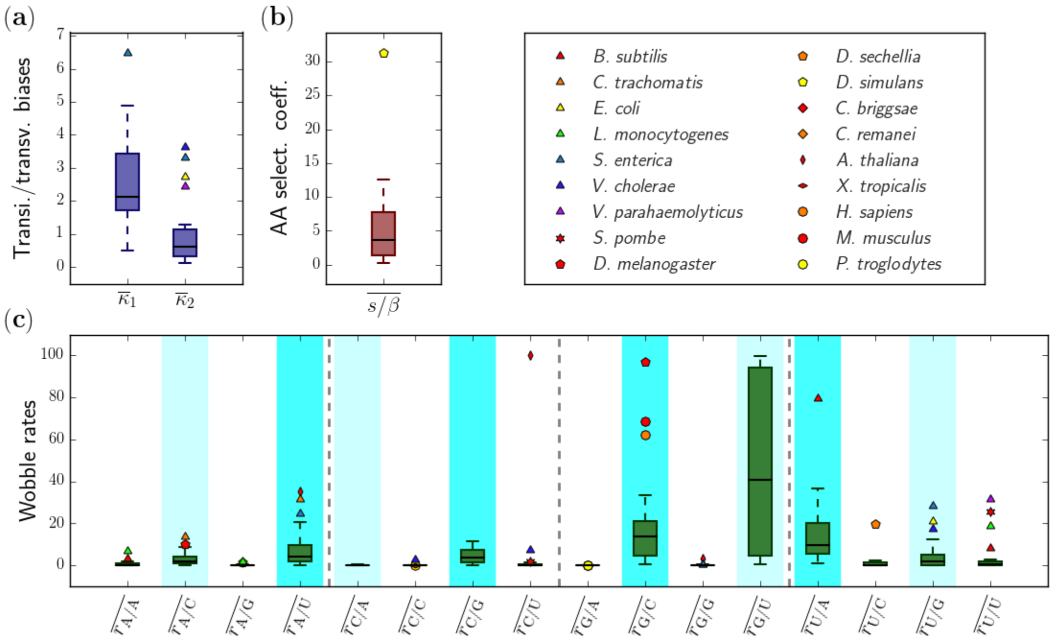

Figure 7.4: **Distributions of inferred biophysical and population genetics parameter values across 20 organisms**. All models are fitted separately on 5 codon subsets and the resulting parameters averaged, as indicated by the overbar. For each parameter averaged in this way, median values across all organisms as well as the first, Q_1, and third, Q_3, quartiles are plotted using box-and-whisker plots. The locations of upper and lower whiskers are given by the largest data point below $Q_3 + 1.5(Q_3 - Q_1)$ and the smallest data point above $Q_1 - 1.5(Q_3 - Q_1)$. Data points which extend outside of this range are considered outliers and plotted explicitly using species-specific symbols. (a) Transition/transversion rate biases $\bar{\kappa}_1$ and $\bar{\kappa}_2$, (b) amino acid selection coefficients $(\overline{s/\beta})$, and (c) wobble rates $\overline{r_{n'/n}}$. The wobble rates are separated into four sets by vertical dashed grey lines, one for each anticodon nucleotide. The cognate pairings are highlighted in solid cyan, and non-cognate pairings with alternate nucleotide types (purine to pyrimidine pairings) are highlighted in faded cyan.

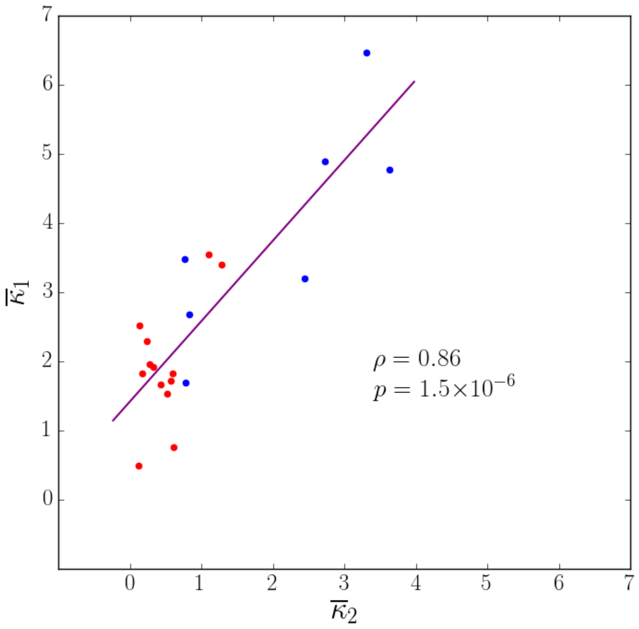

Figure 7.5: **Correlation between transition/transversion rate biases.** Transition/transversion rate bias parameters $\bar{\kappa}_1$ and $\bar{\kappa}_2$ inferred by fitting the 19-parameter model to genome-wide codon frequencies from 20 species and averaged over 5 distinct subsets of codons. Blue dots: prokaryotes, red dots: eukaryotes. Also shown is the Pearson correlation coefficient ρ and the corresponding p-value.

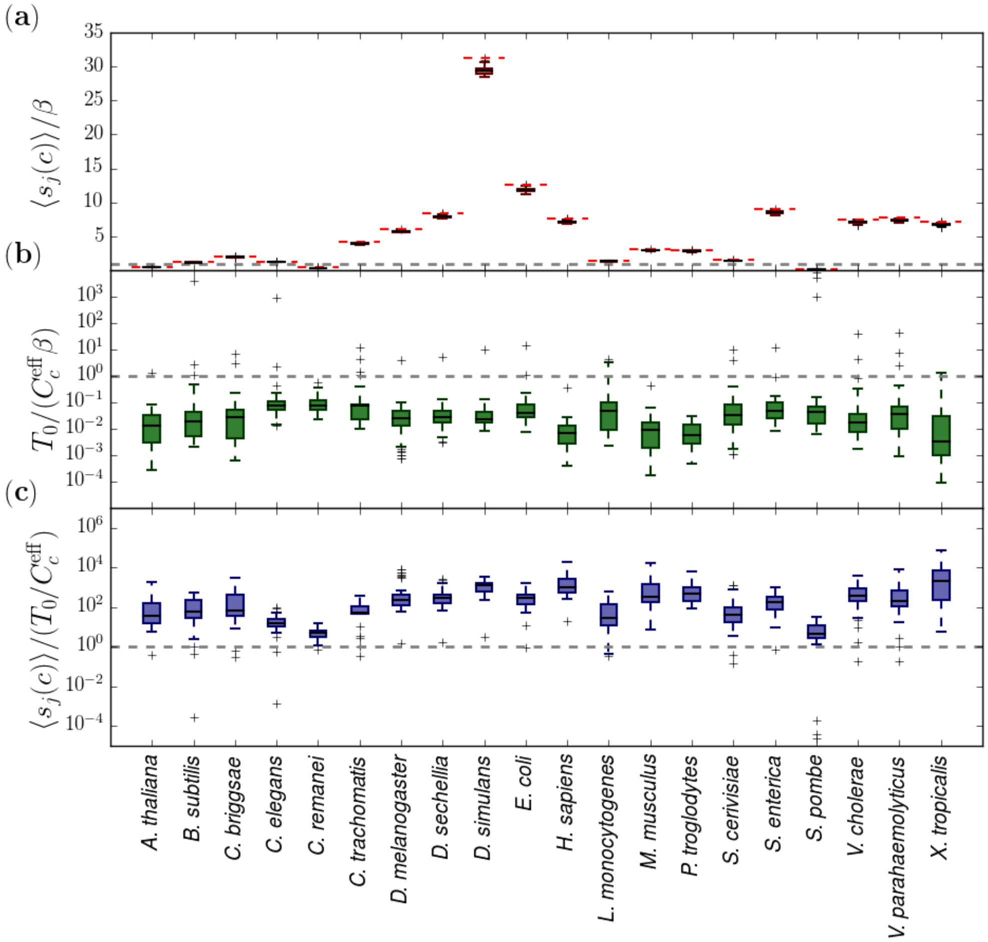

Figure 7.6: **Comparison of selection strengths for speed and accuracy of codon translation**. All selection coefficients have been computed by fitting the 19-parameter model to genomic data from 20 organisms. (a) Ratios of the selection coefficients for amino acid mistranslation, $s_j(c)$ (cf. Eqs. (5.1) and (5.3)), averaged over optimal amino acids/STOP instruction as indicated by angle brackets, to the overall mutation scale β, shown as box-and-whisker plots for each organism. Horizontal dashed red lines indicate the corresponding value of s. (b) Ratios of the selection coefficient for the speed of codon translation, T_0/C_c^{eff} (cf. Eqs. (5.1) and (5.2)), to the overall mutation scale β, shown as box-and-whisker plots for each organism. (c) Ratios of the two selection coefficients from (a) and (b), shown as box-and-whisker plots for each organism. Horizontal dashed grey lines in panels (a)-(c) indicate where each quantity equals 1.

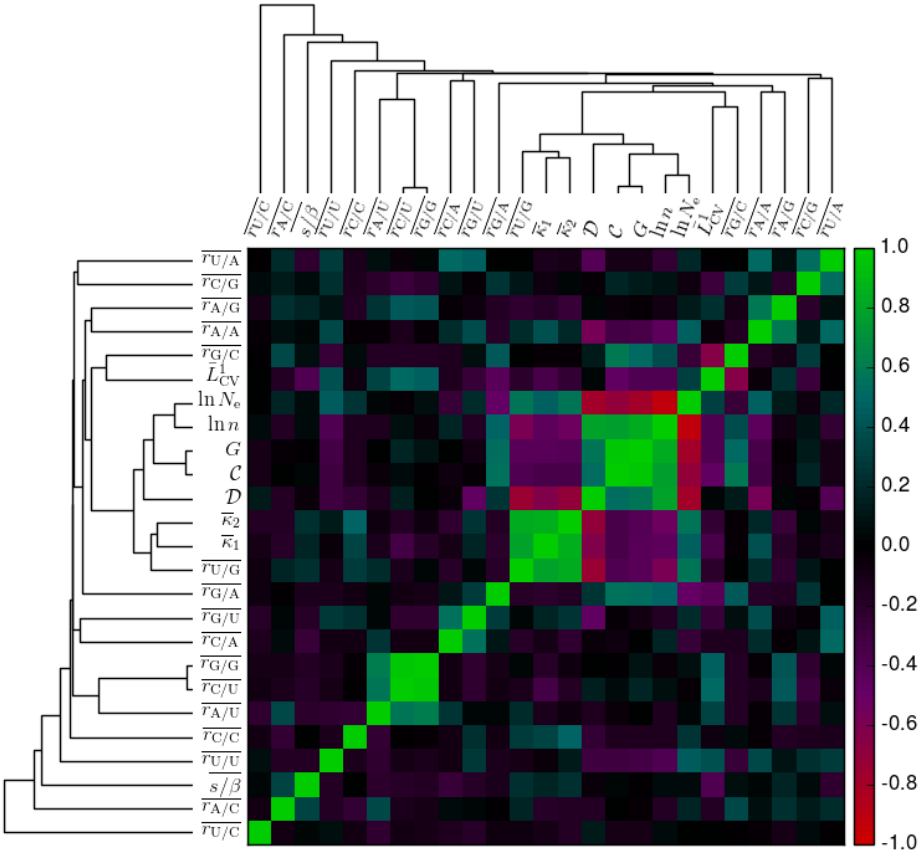

Figure 7.7: **Covariance matrix between various model and additional parameters**. Each entry in the matrix shows the Pearson correlation coefficient between a pair of parameters, with each parameter available for all 20 species included into this study (Fig. 7.1). In addition to the 19 model parameters (Table 6.1), the total number of codons, $\mathcal{C}$, genome size in nucleotides, n, effective population size, N_e, the cross-validation model score, $\bar{L}^1_{CV}$, the total number of genes, G, and the domain label, $\mathcal{D} \in \{\text{Bacteria}, \text{Eukarya}\}$, are included. N_e estimates have been obtained from Refs. [35] (*C. elegans*, *C. remanei*, *D. melanogaster*, *E. coli*, *H. sapiens*, *P. troglodytes*), [63] (*C. briggsae*), [64] (*B. subtilis*), [65] (*M. musculus*), [66] (*S. enterica*), and [67] (*A. thaliana*, *C. trachomatis*, *D. sechellia*, *D. simulans*, *L. monocytogenes*, *S. cerevisiae*, *S. pombe*, *V. cholerae*, *V. parahaemolyticus*, *X. tropicalis*). All parameters were hierarchically clustered using the linkage package from the SciPy Hierarchical clustering library with the 'single' method and default settings.

Chapter 8

Estimation of the Genome-Wide Mutation Rate

Our biophysical approach has also enabled us to estimate the genome-wide mutation rate per nucleotide per generation as an average over all codon types:

$$\langle \mu \rangle = \frac{1}{3} \sum_{c} \sum_{c' \neq c} \mu_{c'c} p_c. \tag{8.1}$$

Indeed, following the approach developed by [23], we can estimate T_0 in Eq. (5.1) directly from the explicit biophysical model of ribosome-mediated translation. Here, we have focused our attention on *E. coli* and *S. cerevisiae*, for which all the requisite values of biophysical parameters are available in the literature. For both of these organisms, we find that

$$T_0 \approx \frac{\tau}{\alpha} \frac{P_{\text{tot}} \gamma}{G t_I k_I}, \tag{8.2}$$

where P_{tot} is the total protein production rate in the cell, G is the total number of genes, t_I is the average ribosome initiation time, k_I is the average initiation on-rate per free ribosome, and τ and α are defined in Eqs. (5.8) and (5.9), respectively.

Using Eq. (8.2) and our assumption of $T_0/\beta = 1$ in the 19-parameter model, we can estimate β and, consequently, $\langle \mu \rangle$ via Eq. (8.1), using intergenic trimer frequencies and predicted values of κ_1 and κ_2. Note that although the $T_0 = \beta$ assumption is arbitrary, we estimate τ/α from the predicted pairing rates in a way that makes our procedure invariant with respect to rescaling both T_0 and all pairing rates by an arbitrary factor (cf. Eqs. (5.1) and (5.2), (8.2)). We have estimated the value of β using both the most up-to-date data available in the literature and Bulmer's original data (see Table 8.1 for input parameter values). Both sets of parameters yield very similar estimates for the average effective mutation rate: 2.4×10^{-6} and 1.1×10^{-6} mutations per nucleotide per generation, respectively. These estimates differ from independent estimates of the genomic mutation rate in *E. coli* [37, 70], which yield values on the order of 10^{-10} mutations per nucleotide per generation.

The same calculation in *S. cerevisiae*, which has a similar effective population size (Table 8.1), has resulted in $\langle \mu \rangle = 7.1 \times 10^{-7}$ mutations per nucleotide per generation, which is also higher than the independently estimated mutation rate of 3.3×10^{-10} mutations per nucleotide per generation [71].

A possible explanation for the observed discrepancy, which is reminiscent of the difficulties encountered by Bulmer in trying to reconcile a population genetics model with the biophysics of mRNA translation [23], is that the codon diversity seen in *E. coli* and *S. cerevisiae* genomic data is affected by linkage and may require an explicit treatment of genetic drift, as $\mu N_\mathrm{e} \ll 1$ for both organisms (Table 8.1). Indeed, genetic drift can contribute to allele diversity observed across multiple sites, even if each individually evolving site is in the monomorphic regime [72, 73]. Note that our model describes the frequencies of $N_c \simeq G\mathcal{L}/20 = \mathcal{O}(10^5)$ codon sites per individual for each fitness landscape, where $\mathcal{L}$ is the average gene length in codons (494 in *S. cerevisiae* and 319 in *E. coli*), and 20 accounts for the number of distinct amino acid types (positions where the STOP instruction has the highest fitness are excluded from the estimate), so statistical noise is likely not a strong contributor to diversity.

Finally, our analysis yields an inverse relationship between $\langle \mu \rangle$ and the total number of genes G, which in turn is strongly correlated with the total number of nucleotides in the genome (Fig. 7.7). This is consistent with Drake's rule, which states that organisms with larger genomes tend to have smaller mutational rates [37]. Multiple-species biophysical data of the type displayed in Table 8.1 will be required to confirm the trend and estimate its significance quantitatively.

8.1 Translation Speed Penalty Connection to Biophysical Quantities

In order to estimate the average mutational rate $\langle \mu \rangle$ (Eq. (8.1)), we first need to determine the mutational scale parameter β. Recall that in the 19-parameter model, we have set $T_0/\beta = 1$ without loss of generality because all wobble rates are free parameters not constrained by normalization. Therefore, using Eq. (5.11) to estimate T_0 automatically determines β. Specifically, to compute the right-hand side of Eq. (5.11) we have followed a biophysical approach originally developed by Bulmer [23]. In this approach, ribosome translation kinetics are modeled explicitly under the assumption of steady-state translation of each mRNA transcript. We start by evaluating

$$\left\langle \frac{d \log P_{tot}}{dt^{c(q,\ell)}} \right\rangle = \left\langle \sum_{i=1}^{G} \frac{P_i}{P_{tot}} \frac{d \log P_i}{dt^{c(q,\ell)}} \right\rangle, \tag{8.3}$$

where P_i is the protein production rate of gene i, $P_{tot} = \sum_{i=1}^{G} P_i$ is the total protein production, G is the total number of genes, and $t^{c(q,\ell)}$ is the translation time of codon c found in gene q at position ℓ. The $\langle \ldots \rangle$ average is over all codon positions which evolve under the same fitness matrix. For a single mRNA transcript, the total ribosomal on-rate will be equal to the total ribosomal off-rate, or the rate at which ribosomes complete translation, in steady-state. If the average time for translation initiation for the ith gene is t_{Ii}, the rate at which translation is completed and proteins are produced is given by $1/t_{Ii}$ per mRNA. If there are m_i mRNA transcripts per cell for gene i, the protein production rate for this gene is $P_i = m_i/t_{Ii}$, yielding

$$\frac{d \log P_i}{dt^{c(q,\ell)}} = -\frac{1}{t_{Ii}} \frac{dt_{Ii}}{dt^{c(q,\ell)}}. \tag{8.4}$$

Further noting that the initiation time, t_{Ii}, is the sum of the time for a single ribosome to bind to the transcript and the time for this ribosome to translate far enough for another ribosome to bind, we obtain

$$t_{Ii} = (k_{Ii} R_f)^{-1} + \sum_{j=1}^{L} t^{c(i,j)}, \tag{8.5}$$

where R_f is the number of free ribosomes in the cell, k_{Ii} is the on-rate per free ribosome for gene i, and L is the ribosomal footprint. Substitution of Eq. (8.5) into Eq. (8.4) yields

$$\frac{d\log P_i}{dt^{c(q,\ell)}} = \frac{1}{t_{Ii}k_{Ii}R_f^2}\frac{dR_f}{dt^{c(q,\ell)}} - \frac{1}{t_{Ii}}\delta_{iq}\delta_{\ell\leq L}. \tag{8.6}$$

The number of free ribosomes, R_f, can be expressed as

$$R_f = R_{tot} - \sum_{i=1}^{G} m_i R_{bi}, \tag{8.7}$$

where R_{tot} is the total number of ribosomes in the cell (assumed to be constant), and R_{bi} is the number of ribosomes bound to a single mRNA transcript of gene r. If t_{Ti} is the average time for a single ribosome to translate,

$$t_{Ti} = \sum_{j=1}^{\mathcal{L}_i} t^{c(i,j)}, \tag{8.8}$$

where $\mathcal{L}_i$ is the total number of codons in gene i, then R_{bi}/t_{Ti} is the total ribosomal off-rate, and therefore

$$\frac{R_{bi}}{t_{Ti}} = \frac{1}{t_{Ii}} \;\Rightarrow\; R_{bi} = \frac{t_{Ti}}{t_{Ii}}. \tag{8.9}$$

under the steady-state assumption in which the ribosomal on and off rates are equal.

The derivative of R_f in Eq. (8.6) is then given by

$$\begin{aligned}
\frac{dR_f}{dt^{c(q,\ell)}} &= -\sum_{r=1}^{G} m_r \frac{dR_{br}}{dt^{c(q,\ell)}} = -\sum_{r=1} m_r \left[\frac{1}{t_{Ir}}\frac{dt_{Tr}}{dt^{c(q,\ell)}} - \frac{t_{Tr}}{t_{Ir}^2}\frac{dt_{Ir}}{dt^{c(q,\ell)}}\right] \\
&= -\sum_{r=1}^{G} m_r \left[\frac{\delta_{rq}}{t_{Ir}} - \frac{t_{Tr}}{t_{Ir}^2}\left\{\delta_{rq}\delta_{\ell\leq L} - \frac{1}{k_{Ir}R_f^2}\frac{dR_f}{dt^{c(q,\ell)}}\right\}\right] \\
&= (R_{bq}\delta_{\ell\leq L} - 1)P_q - \frac{1}{R_f^2}\frac{dR_f}{dt^{c(q,\ell)}}\sum_{r=1}^{G}\frac{R_{br}P_r}{k_{Ir}},
\end{aligned} \tag{8.10}$$

yielding

$$\frac{dR_f}{dt^{c(q,\ell)}} = \frac{(R_{bq}\delta_{\ell\leq L} - 1)P_q}{1 + \frac{1}{R_f^2}\sum_{r=1}^{G}\frac{R_{br}P_r}{k_{Ir}}}, \tag{8.11}$$

where $\delta_{\ell\leq L} = 1$ if $1 \leq \ell \leq L$, and 0 otherwise.

Using this result, Eq. (8.6) becomes

$$\frac{d\log P_i}{dt^{c(q,\ell)}} = \frac{(R_{bq}\delta_{\ell\leq L}-1)P_q}{\left[R_f^2+\sum_{r=1}^{G}\frac{R_{br}P_r}{k_{Ir}}\right]t_{Ii}k_{Ii}} - \frac{1}{t_{Ii}}\delta_{iq}\delta_{\ell\leq L}. \tag{8.12}$$

Finally, Eq. (8.3) can be written as

$$\left\langle \frac{d\log P_{tot}}{dt^{c(q,\ell)}} \right\rangle = \left\langle \sum_{i=1}^{G} \frac{P_i}{P_{tot}} \left\{ \frac{(R_{bq}\delta_{\ell\leq L}-1)P_q}{\left[R_f^2+\sum_{r=1}^{G}\frac{R_{br}P_r}{k_{Ir}}\right]t_{Ii}k_{Ii}} - \frac{1}{t_{Ii}}\delta_{iq}\delta_{\ell\leq L} \right\} \right\rangle. \tag{8.13}$$

To evaluate Eq. (8.13) numerically, we have replaced all single-codon translation times and gene-specific initiation rates with typical values, $t^{c(q,\ell)} \to t$ and $k_{Iq} \to k_I$. Equation (8.13) now simplifies to

$$\left\langle \frac{d\log P_{\text{tot}}}{dt^{c(q,\ell)}} \right\rangle \approx -\frac{P_{\text{tot}}\gamma}{Gt_I k_I} + \sum_{q=1}^{G} \frac{P_q(R_{bq}\gamma P_{\text{tot}} - k_I)}{G|S_q^s|t_I k_I P_{\text{tot}}} \sum_{\ell\in S_q^s} \delta_{\ell\leq L}, \tag{8.14}$$

where

$$\gamma = \left(R_f^2 + \sum_{r=1}^{G} \frac{R_{br}P_r}{k_I} \right)^{-1}. \tag{8.15}$$

We have evaluated the final expression in Eq. (8.14) to estimate T_0 and subsequently the average mutational rate for two organisms where data is available.

8.1.1 Modern *E. coli* Dataset.

Protein production rates P_q were assumed proportional to the relative cellular abundance of proteins produced from gene q (see Ref. [74] for protein abundance data). Using $t = 0.12$ s, $t_I = 62$ s from the Transimulation Web Server (Ref. [75]; *E. coli* K-12 MG1655) and $L = 10$ codons (Ref. [76]), we obtain

$$k_I R_f = \frac{1}{t_I - Lt} = 1.6\times10^{-2} \text{ initiations per second} \tag{8.16}$$

for each mRNA using Eq. (8.5). The number of ribosomes in a single cell of *E. coli* K-12 is $R_{tot} \sim 55\times10^3$ (Ref. [77]) 85% of which are bound to mRNA (Ref. [23]) such that $R_f = 8300$ and $k_I = 2.0\times10^{-6}$ initiations per ribosome per second for each mRNA.

According to Ref. [75], there are on average 47 proteins produced per mRNA, which each have an average lifespan of 7.5 minutes, and 3.6 mRNAs per gene are present in the cell. With $G \simeq 4300$ genes in the *E. coli* K-12 MG1655 reference genome, we obtain

$$P_{\text{tot}} = \frac{\left(3.6\frac{\text{mRNAs}}{\text{gene}}\right) \times \left(47\frac{\text{proteins}}{\text{mRNA}}\right)}{(7.5\text{ mins }) \times \left(60\frac{\text{secs}}{\text{min}}\right)} \times 4300\text{ genes } \simeq 1600\text{ proteins/sec.} \tag{8.17}$$

With a constant translation time per codon and a gene-independent initiation time, Eq. (8.9) becomes $R_{br} = t\mathcal{L}_r/t_I$, where $\mathcal{L}_r$ is the number of codons in gene r. Then Eq. (8.15) yields

$$\gamma = \left(R_f^2 + \sum_{r=1}^{G} \frac{t\mathcal{L}_r P_r}{t_I k_I}\right)^{-1} = 2.5 \times 10^{-9} \tag{8.18}$$

and, consequently, the first term on the right-hand side of Eq. (8.14) is estimated as

$$-\frac{P_{\text{tot}}\gamma}{G t_I k_I} = -7.5 \times 10^{-6}\ s^{-1}. \tag{8.19}$$

Since the value of the second term changes depending on the fitness matrix, this term has been evaluated for all 20 amino acids, with an average value of $\mathcal{O}(10^{-8})\ s^{-1}$ and a standard deviation of $\mathcal{O}(10^{-8})\ s^{-1}$.

Consistent with the expectation that T_0 should be independent of amino acid selection, the first term dominates in *E. coli*, yielding

$$\left\langle \frac{d \log P_{\text{tot}}}{dt^{c(q,\ell)}} \right\rangle \approx -7.5 \times 10^{-6} s^{-1}. \tag{8.20}$$

Next, we focus on estimating the τ/α prefactor in Eq. (5.11). We have averaged Eq. (5.10) over all codons using the tRNA gene copy number data from Refs. [68] and [69], in combination with the pairing rates fit using the 19-parameter model (Table 9.1), and set this average equal to $t = 0.12$ s (Ref. [75]): $t = \sum_c t^c p_c$, resulting in

$$\frac{\tau}{\alpha} = t\left(\sum_c \frac{p_c}{\sum_{n\in\{\text{A,U,C,G}\}} r_{n/c_3} C_{n+\bar{c}_{23}}}\right)^{-1} \simeq 1.2\text{ s.} \tag{8.21}$$

Finally, the denominator in Eq. (5.11) is approximated by

$$\left\langle t^{c(q,\ell)} \frac{d \log P_{\text{tot}}}{dt^{c(q_\ell)}} \right\rangle - 1 = t \left\langle \frac{d \log P_{\text{tot}}}{dt^{c(q_\ell)}} \right\rangle - 1 \approx -1, \tag{8.22}$$

such that the predicted value for the mutation scale is

$$\beta = T_0 \approx \frac{\tau}{\alpha} \frac{P_{\text{tot}} \gamma}{G t_I k_I} = 8.8 \times 10^{-6}. \tag{8.23}$$

This result can be used to estimate the average mutation rate per nucleotide per generation:

$$\begin{aligned} \langle \mu \rangle &= \frac{\text{expected \# nucleotide mutations per generation}}{\text{\# nucleotides}} \\ &= \frac{\sum_c \sum_{c' \neq c} \mu_{c'c} \mathcal{C}_c}{3 \sum_c \mathcal{C}_c} = \frac{1}{3} \sum_c \sum_{c' \neq c} \mu_{c'c} p_c = 7.8 \times 10^{-7}. \end{aligned} \tag{8.24}$$

where $\mathcal{C}_c$ is the total number of codons of type c. For ease of reference, key biophysical parameters discussed above have been summarized in Table 8.1.

8.1.2 Bulmer's Original Dataset

To examine the robustness of our findings, we have repeated the mutation rate calculation using the values of biophysical quantities originally reported by Bulmer [23] (Table 8.1). Following the above analysis and using the additional assumption that $P_r = m_r / t_I$ (Ref. [23]), where m_r is the average number of mRNAs for gene r, we obtain

$$\gamma = \left(R_f^2 + \frac{R_{\text{tot}} - R_f}{k_I t_I} \right)^{-1} \simeq 3.3 \times 10^{-8}, \tag{8.25}$$

leading to $-P_{\text{tot}} \gamma / G t_I k_I \simeq 7.3 \times 10^{-6}\ s^{-1}$.

Next, we follow Bulmer in replacing gene-specific average numbers of bound ribosomes with an average over all genes, $R_{br} \to G^{-1} \sum_{r=1}^{G} R_{br}$. This simplification leads to

$$\left\langle \frac{d \log P_{\text{tot}}}{dt^{c(q,\ell)}} \right\rangle \approx -\frac{P_{\text{tot}} \gamma}{G t_I k_I} + \frac{(R_b \gamma P_{\text{tot}} - k_I) L}{G t_I k_I \mathcal{L}}, \tag{8.26}$$

where $\mathcal{L} = \frac{1}{G}\sum_r \mathcal{L}_r$ is the average gene length in codons, and the approximation

$$\sum_{q=1}^{G} \frac{P_q}{|S_q^s|} \sum_{\ell \in S_q^s} \delta_{\ell \leq L} \approx P_{\text{tot}} \frac{L}{\mathcal{L}} \tag{8.27}$$

has been made. Equation (8.26) can be evaluated directly, resulting in

$$\left\langle \frac{d \log P_{\text{tot}}}{dt^{c(q,\ell)}} \right\rangle \approx -8.2 \times 10^{-6}\ s^{-1}, \tag{8.28}$$

in close agreement with Eq. (8.20). Next, we use the same effective gene copy numbers C_c^{eff} and codon frequencies p_c as before to find $\tau/\alpha = 0.56\ s$ via Eq. (8.21). This leads to $\beta = 4.6 \times 10^{-6}$ and $\langle \mu \rangle = 1.1 \times 10^{-6}$, close to the estimate in Eq. (8.24) which employed more up-to-date parameters.

8.1.3 Modern *S. cerevisiae* Dataset.

For baker's yeast, an estimate of $R_f = 2.8 \times 10^4$ ribosomes per cell was obtained by taking 15% (Ref. [78]) of R_{tot}, reported to be 18.7×10^4 ribosomes in Ref. [79]. Next, Eq. (8.5) was used with $t = 0.10$ s (Ref. [80]), $L = 10$ codons (Ref. [81]), and $t_I = 54$ s (Ref. [75]) to find $k_I = 6.7 \times 10^{-7}$ initiations per ribosome per mRNA per second. According to Ref. [79], $P_{\text{tot}} = 1.3 \times 10^4$ proteins per second, and $G \simeq 6000$ genes in the *S. cerevisiae* reference genome downloaded from NCBI. These values were used in conjunction with the *S. cerevisiae* protein abundance data from Ref. [74] to find $\gamma = 6.2 \times 10^{-11}$ (Eq. (8.15)). The first term on the right-hand side of Eq. (8.14) was estimated to be $-P_{\text{tot}}\gamma/Gt_I k_I = -3.7 \times 10^{-6}\ s^{-1}$, and once again was found to dominate the second term. Finally, just as before, Eq. (5.10) was used in combination with the wobble rates predicted by the 19-parameter model to yield $\tau/\alpha = 2.6\ s$, resulting in $\beta = 9.7 \times 10^{-6}$ (Eq. (8.23)) and $\langle \mu \rangle = 7.1 \times 10^{-7}$ mutations per nucleotide per generation (Eq. (8.24)). For ease of reference, key biophysical parameters have been summarized in Table 8.1.

Table 8.1: **Summary of key quantities used in mutation rate estimation.**

	t	t_I	L	R_f	k_I	P_{tot}	R_{tot}	G	N_{e}
This study									
E. coli	0.12	62	10	8300	8.2×10^{-5}	1600	55000	4300	2.5×10^{7}
S. cer.	0.10	54	10	2.8×10^{4}	6.7×10^{-7}	1.3×10^{4}	18.7×10^{4}	6000	10^{7}
Bulmer [23]									
E. coli	0.056	2.0	10	2800	3.6×10^{-4}	680	18700	–	–

Chapter 9

Numerical Determination of Model Parameters

9.1 Global Optimization Algorithm.

To fit codon frequencies, we have developed a heuristic optimization algorithm whose objective is to minimize the L^1 distance between model predictions and data. The algorithm proceeds as follows:

1. Initialize a set of random model parameters: $\vec{x} = (\kappa_1, \kappa_2, s/\beta, ...)$, where all parameters are drawn from a uniform distribution in the $[0, 10^2]$ range.

2. Generate a list of all possible moves through parameter space, $\vec{x}'_i = \vec{x} + \Delta\vec{x}_i$, in which two changes to parameter values are applied simultaneously as shown below:

$$\begin{aligned}
\vec{x}'_1 &= (\kappa_1 + 2\Delta, \kappa_2, s/\beta, ...), \\
\vec{x}'_2 &= (\kappa_1 - 2\Delta, \kappa_2, s/\beta, ...), \\
\vec{x}'_3 &= (\kappa_1 + \Delta, \kappa_2 + \Delta, s/\beta, ...), \\
\vec{x}'_4 &= (\kappa_1 - \Delta, \kappa_2 + \Delta, s/\beta, ...), \\
&\cdots
\end{aligned}$$

   In each move, wobble rates are enforced to be in the $[0, 10^2]$ range and all other parameters in the $[0, 10^6]$ range, with all moves resulting in out-of-range values excluded from subsequent evaluation. Initially, $\Delta = 10$.

3. From top to bottom of the list, sequentially evaluate the L^1 score of each move until $L^1(\vec{x}'_i) < L^1(\vec{x})$ is obtained for ith move; accept this beneficial move.

4. If a beneficial move is found in step 3, attempt to move in the direction defined by $\Delta\vec{x}_i = \vec{x}'_i - \vec{x}$ repeatedly until no further gain is observed. Specifically, consider

$$\vec{x}'_{\text{new}} = \vec{x}'_i + m\Delta\vec{x}_i,$$

where m is initialized to 10^6 and subsequently reduced by a factor of 10 for each L^1 evaluation which does not result in a score improvement. If an improvement is found, the move is accepted and the evaluations continue from the new position, starting with the same value of m that led to the score improvement. The entire process is terminated when $m = 1$ and subsequent evaluations of $\vec{x}'_{\text{new}}$ yield no further gains. The algorithm then goes back to step 3 and continues the search from the next entry in the move set listed in step 2, but with the new $\vec{x}$ resulting from all the beneficial moves accepted during the extended search in the $\Delta\vec{x}_i$ direction (note that $m = 1$ at this point).

5. If no score improvement is found after a full iteration through the set of moves listed in step 2, the step size Δ is reduced by a factor of 2 and the search procedure described in steps 3 and 4 is repeated. When this reduction results in $\Delta < 10^{-4}$, the step size is reinitialized to the initial value, $\Delta = 10$, and the algorithm restarts from step 2.

6. The run is terminated after 10^3 evaluations of the L^1 score.

Throughout the run, the average rate at which L^1 decreases per function evaluation is computed using a list of 25 most recent L^1 function evaluations (computations of the average rate commence when at least 25 function evaluations have been performed). This average rate is used to estimate the final expected L^1 score by linear extrapolation, given the remaining budget of function evaluations. Note that the minimum L^1 score is 0. Thus, if the extrapolated estimate of L^1 becomes greater than 0 as a result of accepting the latest (slightly) beneficial move, the algorithm is no longer expected to find an optimal set of parameters in the remaining allotted time. Then the algorithm executes step 5, clears the L^1 history list, and proceeds to the next move in the list shown in step 2. This allows the algorithm to avoid repeating moves which result in very small score improvements.

The algorithm described above was independently run 10^3 times starting from randomized initial conditions, and the set of parameters with the lowest L^1 was selected. These parameter values were then uniformly randomized by ±5% to generate another 10^3 starting points, and the algorithm was run again for 10^3 evaluations of the L^1 score. This process was repeated until no further improvement in the L^1 score was observed, at which point the best parameter set recorded throughout the run was reported. A set of representative global optimization runs for our hierarchy of models, using *E. coli* genome-wide codon frequencies as input, is shown in Fig. 9.1.

The average optimal score, $\bar{L}^1$, and cross-validation score, $\bar{L}^1_{\mathrm{CV}}$, (both averages over the 5 data sets) as functions of the total number of function evaluations for this algorithm is shown in Fig. 9.1 for all models considered. To compute L^1_{CV} for each data set so as to determine $\bar{L}^1_{\mathrm{CV}}$ for this figure, the optimal parameter set found after each 1000 function evaluations is used to make a prediction on the other 4 data sets. These predicted frequencies are then scored and the scores averaged to compute L^1_{CV}. We compute the L^1_{CV} as a metric of over-fitting: if the recovered parameter values perform poorly on the other 4 data sets, this is taken as an indication of over-fitting. For all cases considered, the comparative difference between L^1 and L^1_{CV} is small with $L^1 < L^1_{\mathrm{CV}}$ indicating only small overfitting, although this is not considered substantial as there are minor differences between the data set frequencies given that each data set consists of a large number of data points.

The algorithm was constructed and its auxiliary parameters fine-tuned empirically using *E. coli* codon frequency data set as input; the same procedure has been used in all subsequent fitting.

Once the algorithm completed on all 5 data sets, the optimal parameters for each set are averaged and the RMS deviations from these average values are computed. Since the RMS deviation is calculated from only 5 numbers for each parameter, it should therefore only be interpreted as a way to quantify the variation between the fitting results and not as an estimate of the standard deviation or a confidence interval.

9.2 Algorithm Validation.

To validate our global optimization procedure, we have generated 5 synthetic data sets with sizes equal to those of the *E. coli* genomic data sets by multinomial sampling of the codon frequencies predicted using model parameters obtained by fitting to *E. coli* genomic data. These synthetically generated counts were then used as input data in a subsequent optimization run. The convergence of the algorithm on synthetic data is demonstrated in Fig. 9.2 for our hierarchy of models, and the parameters recovered during these subsequent optimization runs are compared with the original parameters in Table 9.1.

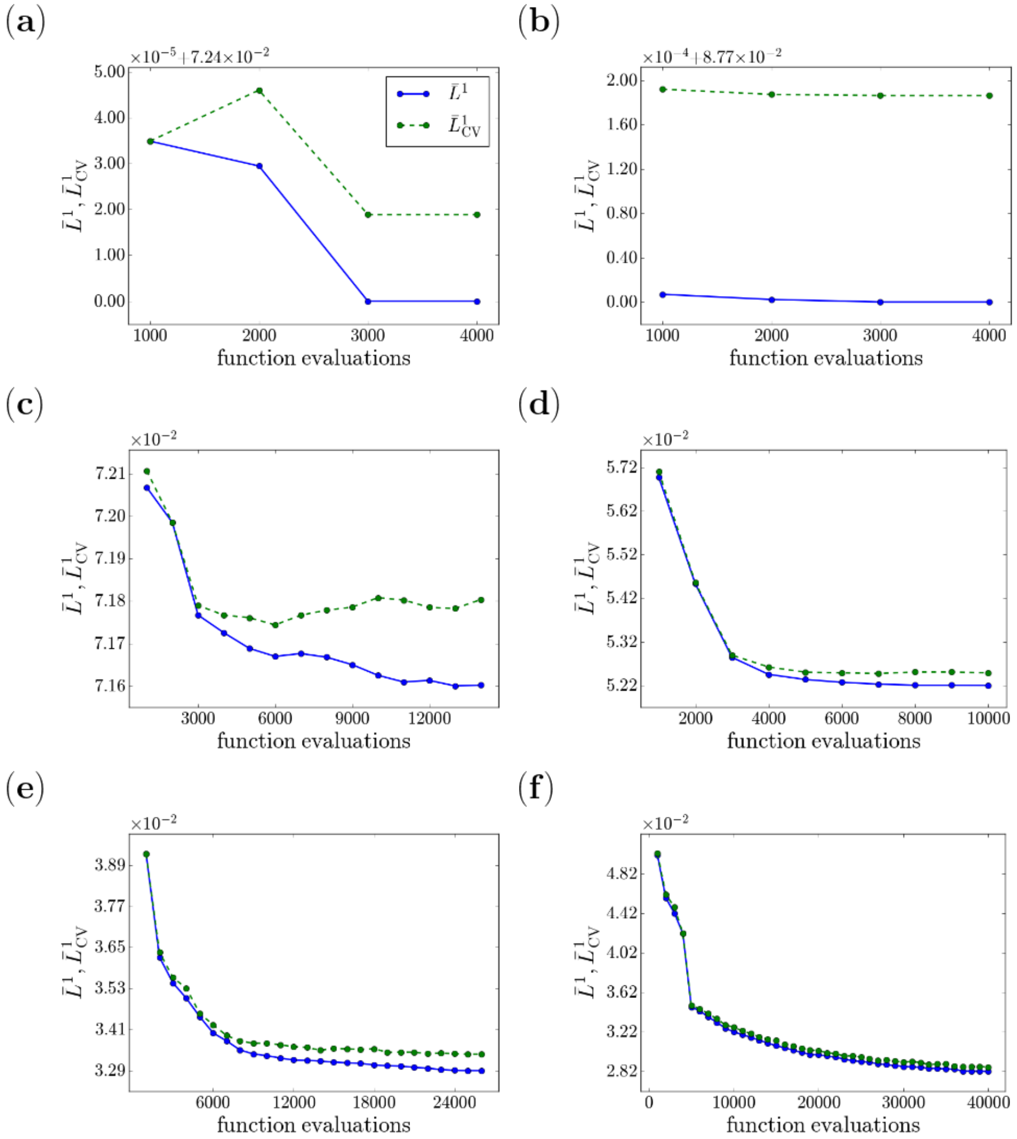

Figure 9.1: **Convergence of the algorithm fitted to the codon frequencies from the *E. coli* genome.** Model scores $\bar{L}_1$ (solid lines) and $\bar{L}^1_{\mathrm{CV}}$ (dashed lines) are shown vs. the total number of function evaluations in the 3-parameter (a), 5-parameter (b), 7-parameter (c), 12-parameter (d), 16-parameter (e), and 19-parameter (f) models.

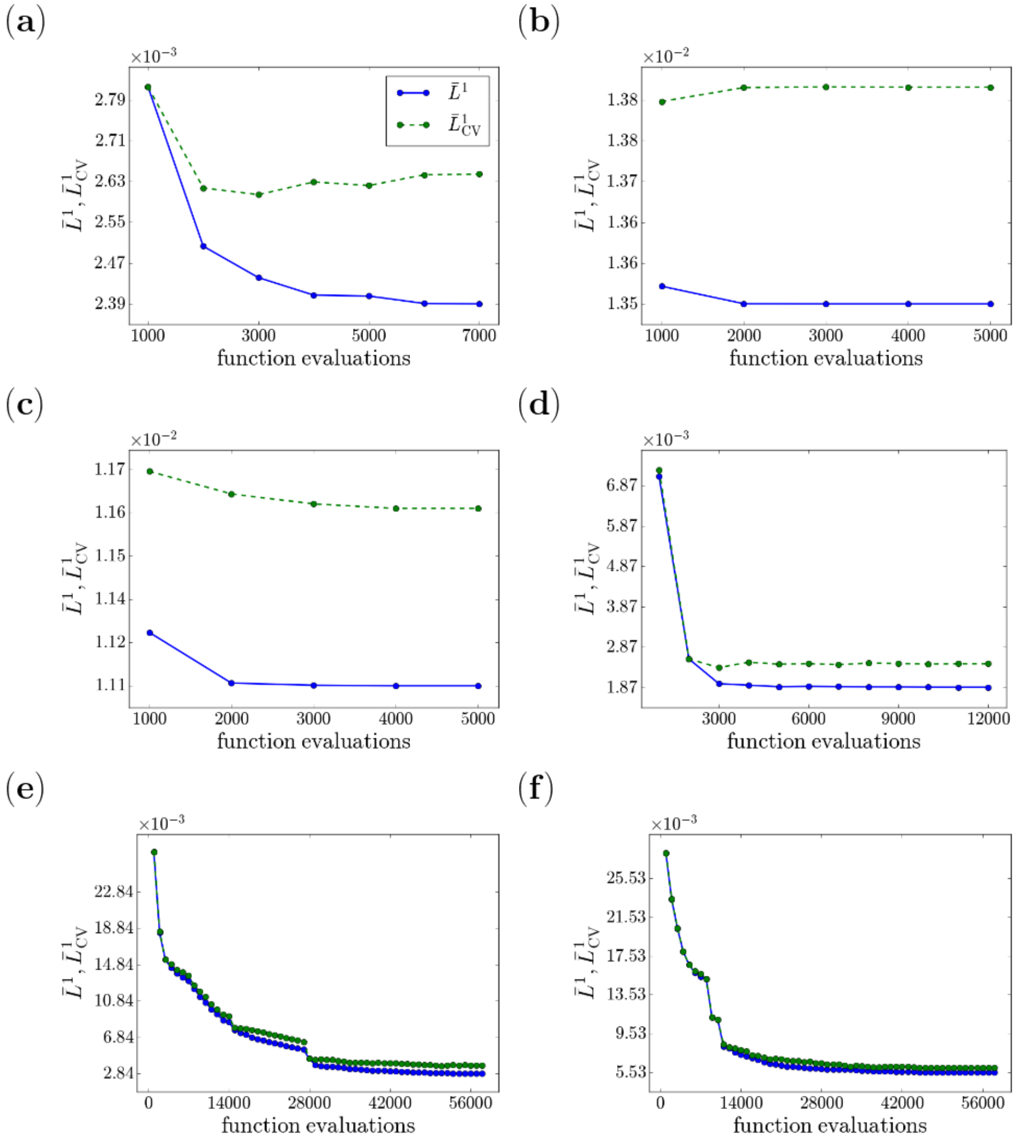

Figure 9.2: **Convergence of the algorithm fitted to the codon frequencies based on synthetic datasets.** Model scores $\bar{L}_1$ (solid lines) and $\bar{L}^1_{\rm CV}$ (dashed lines) are shown vs. the total number of function evaluations in the 3-parameter (a), 5-parameter (b), 7-parameter (c), 12-parameter (d), 16-parameter (e), and 19-parameter (f) models. Each fit was performed on synthetic data generated by the same model, with all the model parameters set to values previously found in fitting the model to codon frequencies from the *E. coli* genome (Table 9.1).

Table 9.1: **Sets of parameters used to generate synthetic datasets, and subsequent parameter predictions.** Each parameter set was obtained by fitting the corresponding model to *E. coli* genomic data (first row in each model subsection). Second and third rows show the average and the RMS deviation of model parameters obtained by re-fitting the same model on 5 synthetic datasets which were independently generated using the values in the first row.

3-parameter model

κ_1	κ_2	s/β
4.18	3.44	644
4.16	3.42	2.00×10^5
.0478	.0502	3.87×10^5

5-parameter model

κ_1	κ_2	s/β	T_0/β	r
3.38	.211	.581	.237	.659
3.41	.214	.633	.230	.650
.0901	.0192	.0105	2.62×10^{-3}	1.45×10^{-3}

7-parameter model

κ_1	κ_2	s/β	T_0/β	r_0	r_1	r_2
6.59	4.74	868	.0702	.702	.000	.000
6.65	4.76	1.39×10^4	.0712	.708	.000	.000
.222	.160	7360	1.54×10^{-3}	.0151	.000	.000

12-parameter model

κ_1	κ_2	s/β	T_0/β	$r_{\mathrm{A/A}}$	$r_{\mathrm{A/C}}$	$r_{\mathrm{A/G}}$	$r_{\mathrm{C/A}}$
5.23	2.85	644	.139	.216	.490	.000	.000
5.23	2.85	8980	.139	.209	.487	3.00×10^{-3}	.000
.0553	.0357	6480	1.25×10^{-3}	7.42×10^{-3}	.0124	5.56×10^{-3}	.000
$r_{\mathrm{G/U}}$	$r_{\mathrm{U/C}}$	$r_{\mathrm{U/G}}$	$r_{\mathrm{U/U}}$				
1.01	.000	.605	.000				
1.00	.000	.603	.000				
.0145	.000	.0105	.000				

16-parameter model

κ_1	κ_2	s/β	T_0/β	$r_{\mathrm{A/A}}$	$r_{\mathrm{A/C}}$	$r_{\mathrm{A/G}}$	$r_{\mathrm{C/A}}$
5.47	3.49	22.8	.0997	.114	.348	.000	1.22×10^{-5}
5.45	3.46	20.5	.0999	.114	.355	1.96×10^{-3}	6.10×10^{-5}
.0898	.0950	1.39	1.40×10^{-3}	4.20×10^{-3}	8.07×10^{-3}	2.85×10^{-3}	1.22×10^{-4}
$r_{\mathrm{C/C}}$	$r_{\mathrm{C/U}}$	$r_{\mathrm{G/A}}$	$r_{\mathrm{G/G}}$	$r_{\mathrm{G/U}}$	$r_{\mathrm{U/C}}$	$r_{\mathrm{U/G}}$	$r_{\mathrm{U/U}}$
.000	9.05×10^{-5}	1.10×10^{-3}	.0126	1.48	5.40×10^{-4}	1.26	.000
.000	1.40×10^{-4}	1.18×10^{-3}	.0138	1.46	6.20×10^{-4}	1.24	.000
.000	8.18×10^{-5}	1.94×10^{-4}	1.28×10^{-3}	0.0527	4.32×10^{-5}	.0385	.000

19-parameter model

κ_1	κ_2	s/β	$r_{\mathrm{A/A}}$	$r_{\mathrm{A/C}}$	$r_{\mathrm{A/G}}$	$r_{\mathrm{A/U}}$	$r_{\mathrm{C/A}}$
4.90	2.73	12.6	1.08	1.68	8.06×10^{-3}	2.38	8.80
4.92	2.78	8.41	1.13	1.72	.0127	2.42	6.81
.0650	.0272	.301	.0751	.0459	8.51×10^{-3}	.0878	1.81×10^{-3}
$r_{\mathrm{C/C}}$	$r_{\mathrm{C/G}}$	$r_{\mathrm{C/U}}$	$r_{\mathrm{G/A}}$	$r_{\mathrm{G/C}}$	$r_{\mathrm{G/G}}$	$r_{\mathrm{G/U}}$	$r_{\mathrm{U/A}}$
1.80×10^{-3}	4.26	.000	.000	18.6	.313	12.0	8.19
.000	4.37	8.17×10^{-5}	1.92×10^{-3}	18.5	.470	12.0	8.11
.000	.157	1.63×10^{-4}	1.95×10^{-3}	.834	.0381	.207	.122
$r_{\mathrm{U/C}}$	$r_{\mathrm{U/G}}$	$r_{\mathrm{U/U}}$					
.0559	21.3	.000					
.0838	20.5	.000					
.0117	1.03	.000					

Chapter 10

Concluding Remarks on Biophysical and Population Genetics Model for the Analysis of Codon Usage Bias

We have developed a population genetics treatment of the biophysical model of codon bias. We assume that genome-wide codon frequencies have reached steady state and model the codon population using a selection-mutation framework in which codons evolve independently of one another. Our model includes a detailed description of codon-level mutations which takes transition/transversion biases into account [22, 57]. Furthermore, there are two kinds of selective forces in the model. We assume that most protein coding regions in the genome evolve under purifying selection and that for each codon, translation into amino acids different from the optimal one (which corresponds to the codon in the standard genetic code) carries a selective penalty. Thus our model incorporates mutational robustness, in which steady-state allele frequencies in a polymorphic population of equal-fitness alleles can be non-uniform, with more robust alleles, separated on average by a higher number of mutational steps from the deleterious alleles, being relatively enriched [21]. Interestingly, even the minimal 3-parameter model, which takes only mutation and selection against mistranslation into account and considers only cognate codon-anticodon pairings, is capable of reproducing genome-wide codon frequencies with $\rho = 0.79$ in *E. coli* (Table 6.1).

In addition to the factors described above, we assume that cellular fitness is proportional to the total protein production rate, which leads to selective penalties for codons with longer translation times. A major factor which determines translation speed is the cellular tRNA concentration, which in our model is assumed to be proportional to the tRNA gene copy numbers in the genome [27]. Finally, codon-anticodon pairing rates are computed on the basis of the wobble hypothesis, such that a mutation in the

3' nucleotide of a given codon may bring about a complicated set of changes in which the effective tRNA gene copy number may increase or decrease simultaneously with the change in the codon's mistranslation rate. Thus the final contribution of the codon to the total cellular fitness depends on the delicate balance between speed and accuracy of the codon's translation, and the genome-wide codon frequencies depend on the steady-state balance between selection and mutation forces. While we have neglected other possible mechanisms of selection on codon usage, such as mRNA toxicity [82], mRNA transcription [83], translation initiation [84], and co-translational folding [85], the ability of our model to empirically explain observed patterns of codon usage across many organisms suggests that these mechanisms, while undoubtedly important in some cases, do not play a dominant role in shaping genome-wide codon usage.

We have fit our biophysical model to genomic codon frequencies from 20 organisms. Overall, the model reproduces observed genome-wide patterns of codon usage to a high degree of accuracy (Fig. 6.3). When codons are ranked based on the accuracy of the model prediction, the codon CTA appears in 8 of the 20 organisms as one of the top 4 least accurately predicted codon frequencies. No such pattern emerges for amino acids. In terms of the predicted model parameters, the values of mutational biases κ_1 and κ_2 are fairly conserved as expected, with larger values typically found in prokaryotes and with $\kappa_1 > \kappa_2$ in all organisms. The universality of mutational rate biases across organisms is consistent with the fact that nucleotide trimer frequencies are strongly conserved in the intergenic regions (Fig. 5.1).

Furthermore, we observe that codons are under strong selection against mistranslation, with $s/\beta = 5.84$ when averaged over all organisms (Fig. 7.4B), and $s/\beta < 1$ only in *S. pombe*, *C. remanei*, and *A. thaliana*. We have found that in each organism the fitted value of the selective penalty s, introduced in Eq. (5.3), is nearly equal to the mean of the corresponding distribution of the $s_j(c)$ selection coefficients, defined in Eq. (5.1) (Fig. 7.6A). On the other hand, in both *E. coli* and *S. cerevisaie* β is several-fold larger than $\langle \mu \rangle$, the genome-wide mutation rate per nucleotide per generation averaged over all codon types. Thus we expect $s/\langle \mu \rangle$ to be > 1 in all organisms, making selection against mistranslation a dominant evolutionary force in comparison with mutational effects.

In contrast, the ratio of the selection coefficient associated with the translation speed to the mutation scale, $T_0/(\beta C_c^{\text{eff}})$, is < 1 on average (Fig. 7.6B). Thus our model predicts that fitness costs associated with slow translation are often subordinate to the mutational effects, and are much less pronounced than selection against mistranslation (Fig. 7.6C). Nonetheless, we expect $T_0/(\langle\mu\rangle C_c^{\text{eff}})$ to be > 1 for a nonzero fraction of all codons, indicating that at least in some cases selection against slow translation is an important factor which shapes observed codon frequencies.

Finally, despite the fact that pairing rates are unrestricted in the 19-parameter model, the rates follow well-established patterns consistent with both empirical rules of the wobble hypothesis [28] and atom-level details of codon-anticodon binding on the ribosomal template [36]. For example, rates of cognate pairing are much higher than rates of wobble pairing (Fig. 7.4C), with the sole exception of the G/U pairing whose rates are predicted to be anomalously large. Note that in our framework the first two codon positions are assumed to have no effect on the pairing rates.

As an additional test of our approach, we have estimated T_0, defined in Eq. (5.1), directly using an explicit biophysical model of ribosome-mediated translation originally developed by [23]. Bulmer's model relies on biophysical parameters such as single-codon translation times and translation initiation rates, whose values are available in the literature for *E. coli* and *S. cerevisaie* (Table 8.1). Estimating T_0 has enabled us to find the average mutation rate per nucleotide, $\langle\mu\rangle$, in the coding regions, and compare it with previously published estimates of genome-wide mutation rates [35, 37, 67, 70, 71]. Our estimates of $\langle\mu\rangle$ are several orders of magnitude higher than the values of μ available in the literature. A model of codon evolution which includes genetic drift and linkage between multiple codon loci is necessary to investigate these discrepancies further. Additional refinements of the model could also replace s with several fitness penalties which would depend on the physicochemical similarity of the mistranslated amino acid to the optimal one.

Finally, we note that according to our biophysical framework, $\langle\mu\rangle$ is inversely proportional to the number of genes (Eq. (8.2)). This is reminiscent of the observation, due to Drake, that organisms with larger genomes tend to have smaller mutational rates [37]. We intend to extend our mutation-selection model to all conserved and non-conserved regions of the genome in order to study this correlation in more detail.

References

[1] W. B. Kion-Crosby and A. V. Morozov. Rapid Bayesian Inference of Global Network Statistics Using Random Walks. *Phys Rev Lett*, 121:038301, 2018.

[2] W. B. Kion-Crosby, M. Manhart, and A. V. Morozov. Inferring biophysical models of evolution from genome-wide patterns of codon usage. *bioRxiv*, 2019.

[3] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Rev Mod Phys*, 74:47–97, 2002.

[4] M. E. J. Newman. Mixing patterns in networks. *Phys Rev E*, 67:026126, 2003.

[5] S. H. Lee, P.-J. Kim, and J. Hawoong. Statistical properties of sampled networks. *Phys Rev E*, 73:016102, 2006.

[6] S. Yoon, S. Lee, S.-H. Yook, and Y. Kim. Statistical properties of sampled networks by random walks. *Phys Rev E*, 75:046114, 2007.

[7] E. Estrada. Quantifying network heterogeneity. *Phys Rev E*, 82:066102, 2010.

[8] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *Proc 29th Conf Inform Comm*, INFOCOM'10, pages 2498–2506, Piscataway, NJ, USA, 2010. IEEE Press.

[9] C. Cooper, T. Radzik, and Y. Siantos. Estimating network parameters using random walks. In *2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN)*, pages 33–40, Nov 2012.

[10] C. A. Bliss, C. M. Danforth, and P. S. Dodds. Estimation of global network statistics from incomplete data. *PLoS One*, 9:e108471, 2014.

[11] Y. Zhang, E. D. Kolaczyk, and B. D. Spencer. Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *Ann Appl Stat*, 9:166–199, 2015.

[12] F. H. C. Crick. On protein synthesis. In F. K. Sanders, editor, *Symposia of the Society for Experimental Biology, Number XII: The Biological Replication of Macromolecules*, pages 138–163. Cambridge University Press, Cambridge, 1958.

[13] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* Garland Science, New York, NY, 2015.

[14] R. Hershberg and D. A. Petrov. Selection on codon bias. *Annu Rev Genet*, 42:287–299, 2008.

[15] M. G. Napolitano, M. Landon, C. J. Gregg, M. J. Lajoie, L. Govindarajan, J. A. Mosberg, G. Kuznetsov, D. B. Goodman, O. Vargas-Rodriguez, F. J. Isaacs, D. Söll, and G. M. Church. Emergent rules for codon choice elucidated by editing rare arginine codons in Escherichia coli. *Proc Natl Acad Sci USA*, 113:E5588–E5597, 2016.

[16] R. Nielsen, V. L. Bauer DuMont, M. J. Hubisz, and C. F. Aquadro. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. *Mol Biol Evol*, 24:228–235, 2007.

[17] P. M. Sharp, E. Bailes, R. J. Grocock, J. F. Peden, and R. E. Sockett. Variation in the strength of selected codon usage bias among bacteria. *Nucl Acids Res*, 33:1141–1153, 2005.

[18] P. M. Sharp, L. R. Emery, and K. Zeng. Forces that influence the evolution of codon bias. *Phil Trans R Soc Lond B*, 365(1544):1203–1212, 2010.

[19] M. Kimura. Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. *Proc Natl Academy Sci USA*, 78:5773–5777, 1981.

[20] H. Akashi. Gene expression and molecular evolution. *Curr Opin Genet Dev*, 11:660–666, 2001.

[21] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proc Natl Academy Sci USA*, 96:9716–9720, 1999.

[22] Z. Yang. *Computational Molecular Evolution.* Oxford University Press, Oxford, UK, 2006.

[23] M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129:897–907, 1991.

[24] J. B. Plotkin and G. Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, 12:32–42, 2011.

[25] L. Duret. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*, 12:640–649, 2002.

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

[27] S. Kanaya, Y. Yamada, Y. Kudo, and T. Ikemura. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238:143–155, 1999.

[28] F. H. C. Crick. Codon-anticodon pairing: The wobble hypothesis. *J Mol Biol*, 19:548–555, 1966.

[29] N. Stoletzki and A. Eyre-Walker. Synonymous codon usage in Escherichia coli: Selection for translational accuracy. *Mol Biol Evol*, 24:374–381, 2007.

[30] W. H. Li. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J Mol Evol*, 24:337–345, 1987.

[31] G. A. T. McVean and B. Charlesworth. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genetical Res*, 74:145–158, 1999.

[32] M. Kimura. The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet*, 66:367–386, 1991.

[33] A. S. Kondrashov. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J Theor Biol*, 175:583–594, 1995.

[34] V. Mustonen and M. Lässig. Fitness flux and ubiquity of adaptive evolution. *Proc Nat Acad Sci USA*, 107:4248–4253, 2010.

[35] B. Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, 10, 2009.

[36] F. V. Murphy IV and V. Ramakrishnan. Structure of a purine-purine wobble base pair in the decoding center of the ribosome. *Nat Struct Mol Biol*, 11:1251–1252, 2004.

[37] J. W. Drake. A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA*, 88:7160–7164, 1991.

[38] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry.* Elsevier, Amsterdam, 2007.

[39] P. L. Krapivsky, S. Redner, and E. Ben-Naim. *A Kinetic View of Statistical Physics.* Cambridge University Press, 2010.

[40] J. D. Noh and H. Rieger. Random walks on complex networks. *Phys Rev Lett*, 92:118701, 2004.

[41] S. Condamin, O. Benichou, and M. Moreau. Random walks and Brownian motion: a method of computation for first-passage times and related quantities in confined geometries. *Phys Rev E*, 75:021111, 2007.

[42] M. Manhart, W. Kion-Crosby, and A. V. Morozov. Path statistics, memory, and coarse-graining of continuous-time random walks on networks. *J Chem Phys*, 143:214106, 2015.

[43] S. Redner. *A Guide to First-Passage Processes.* Cambridge University Press, 2001.

[44] E. M. Bollt and D. ben-Avraham. What is special about diffusion on scale-free nets? *New J Phys*, 7:26–47, 2005.

[45] S. J. Hardiman and L. Katzir. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 539–550, New York, NY, USA, 2013. ACM.

[46] P. Erdos and A. Renyi. On the evolution of random graphs. *Bull Inst Internat Stat*, 38:343–347, 1961.

[47] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[48] M. E. J. Newman, C. Moore, and D. Watts. Mean-field solution of the small-world network model. *Phys Rev Lett*, 84:3201–3204, 2000.

[49] S. Meloni, A. Arenas, and Y. Moreno. Traffic-driven epidemic spreading in finite-size scale-free networks. *Proc Nat Acad Sci USA*, 106:16897–16902, 2009.

[50] M. A. Serrano, D. Krioukov, and M. Boguna. Self-similarity of complex networks and hidden metric spaces. *Phys Rev Lett*, 100:078701, 2008.

[51] M. Boguna, D. Krioukov, and K. C. Claffy. Navigability of complex networks. *Nat Phys*, 5:74–80, 2009.

[52] L. Pellis, F. Ball, S. Bansal, K. Eames, T. House, V. Isham, and P. Trapman. Eight challenges for network epidemic models. *Epidemics*, 10:58–62, 2015.

[53] `https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article?`

[54] `https://stats.wikimedia.org/EN/TablesWikipediaEN.htm`.

[55] `https://en.wikipedia.org/wiki/Wikipedia:Size_in_volumes`.

[56] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput Commun Rev*, 29:251–262, 1999.

[57] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10:512–526, 1993.

[58] P. M. Sharp, T. M. Tuohy, and K. R. Mosurski. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucl Acids Res.*, 14:5125–5143, 1986.

[59] S. Kumar, G. Stecher, M. Suleski, and S. B. Hedges. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol*, 34:1812–1819, 2017.

[60] S. B. Hedges, J. Dudley, and S. Kumar. TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics*, 22:2971–2972, 2006.

[61] I. Letunic and P. Bork. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucl Acids Res*, 44:W242–W245, 2016.

[62] I. Keller, D. Bensasson, and R. A. Nichols. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLoS Genet*, 3:e22, 2007.

[63] C. G. Thomas, W. Wang, R. Jovelin, R. Ghosh, T. Lomasko, Q. Trinh, L. Kruglyak, L. D. Stein, and A. D. Cutter. Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. *Genome Res*, 25:667–678, 2015.

[64] H. Maughan, V. Callicotte, A. Hancock, C. W. Birky, W. L. Nicholson, and J. Masel. The population genetics of phenotypic deterioration in experimental populations of Bacillus subtilis. *Evolution*, 60:686–695, 2006.

[65] M. Phifer-Rixey, F. Bonhomme, P. Boursot, G. A. Churchill, J. Pialek, P. K. Tucker, and M. W. Nachman. Adaptive evolution and effective population size in wild house mice. *Mol Biol Evol*, 29:2949–2955, 2012.

[66] J. Charlesworth and A. Eyre-Walker. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol*, 23:1348–1356, 2006.

[67] M. Lynch, L.-M. Bobay, F. Catania, J.-F. Gout, and M. Rho. The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genom Hum Genet*, 12:347–366, 2011.

[68] P. P. Chan and T. M. Lowe. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucl Acids Res*, 37:D93–D97, 2009.

[69] P. P. Chan and T. M. Lowe. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucl Acids Res*, 44:D184–D189, 2016.

[70] S. Wielgoss, J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, C. Médigue, R. E. Lenski, and D. Schneider. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with Escherichia coli. *G3 (Bethesda)*, 1:183–186, 2011.

[71] M. Lynch, W. Sung, K. Morris, N. Coffey, C. R. Landry, E. B. Dopman, W. J. Dickinson, K. Okamoto, S. Kulkarni, D. L. Hartl, and W. K. Thomas. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA*, 105:9272–9277, 2008.

[72] G. Sella and A. E. Hirsh. The application of statistical physics to evolutionary biology. *Proc Natl Academy Sci USA*, 102:9541–9546, 2005.

[73] M. Manhart, A. Haldane, and A. V. Morozov. A universal scaling law determines time reversibility and steady state of substitutions under selection. *Theor Pop Biol*, 82:66–76, 2012.

[74] M. Wang, C. J. Herrmann, M. Simonovic, D. Szklarczyk, and C. Mering. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, 15:3163–3168, 2015.

[75] M. Siwiak and P. Zielenkiewicz. Transimulation - protein biosynthesis Web service. *PLoS ONE*, 8:e73943, 2013.

[76] M. Kozak. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol Rev*, 47:1–45, 1983.

[77] S. Bakshi, A. Siryaporn, M. Goulian, and J. C. Weisshaar. Superresolution imaging of ribosomes and RNA polymerase in live Escherichia coli cells. *Mol Microbiol*, 85:21–38, 2012.

[78] D. Zenklusen, D. R. Larson, and R. H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*, 15:1263–1271, 2008.

[79] T. von der Haar. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol*, 2:87, 2008.

[80] M. A. Gilchrist and A. Wagner. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *J Theor Biol*, 239:417–434, 2006.

[81] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324:218–223, 2009.

[82] P. Mittal, J. Brindle, J. Stephen, J. B. Plotkin, and G. Kudla. Codon usage influences fitness through RNA toxicity. *Proc Natl Acad Sci USA*, 115:8639–8644, 2018.

[83] Z. Zhou, Y. Dang, M. Zhou, L. Li, C. h. Yu, J. Fu, S. Chen, and Y. Liu. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci USA*, 113:E6117–E6125, 2016.

[84] S. Bhattacharyya, W. M. Jacobs, B. V. Adkar, J. Yan, W. Zhang, and E. I. Shakhnovich. Accessibility of the Shine-Dalgarno sequence dictates N-terminal codon bias in E.coli. *Mol Cell*, 70:894–905, 2018.

[85] W. M. Jacobs and E. I. Shakhnovich. Evidence of evolutionary selection for co-translational folding. *Proc Natl Acad Sci USA*, 114:11434–11439, 2017.