Scuola Internazionale Superiore
di Studi Avanzati

PhD in Statistical Physics

Statistical physics approaches to protein translation

Luca Caniparoli

Thesis submitted for the degree of
Doctor Philosophiae

Advisors: Prof. Matteo Marsili
Prof. Michele Vendruscolo

October 2013

Abstract

In this work we present an integrated approach to the study of protein translation, based on Statistical Physics. We adopted three different but complementary perspectives: building hypothesis up from the data, modeling down from reasonable assumptions, and using computer simulations when everything else fails.

In particular, we first analyze the mRNA sequences by means of information theory. We focus on the way the redundancy of the genetic code (the 61 sense triplets of nucleotides -the codons- encode for 20 amino acids) is utilized in the actual sequences, a phenomenon known as the codon bias. We observe that it is not completely random, and encodes information in the frequencies and in the order of the codons.

With the scope of explaining these anomalies, we develop and analyze a family of stochastic models. Translation emerges as a systemic process, where the limited amount of resources in the cell couples the expression of the genes at a global level. We also suggest a game-theoretical interpretation of the codon bias.

We finally attack the problem of cotranslational folding (i.e., the folding of the protein while it is still being translated). Specifically, we ask how this process depends on the order of the codons. This question is computationally very cumbersome. We thus propose a framework based on Markov chains, which allows the efficient simulation of arbitrarily complicate cotranslational folding mechanisms.

List of published papers

1. L. Caniparoli and P. Lombardo
   A non-equilibrium, stochastic model for codon translation time statistics. *ArXiv preprint arXiv:1308.1875* (2013).

2. L. Caniparoli, M. Marsili and M. Vendruscolo
   The codon information index: a quantitative measure of the information provided by the codon bias. *J. Stat. Mech: Theory Exp.*, P04031 (2013).

3. A. Vezzani, R. Burioni, L. Caniparoli and S. Lepri
   Local and average behavior in inhomogeneous superdiffusive media. *Philos. Mag.*, 91(13-15), 1987-1997 (2011).

4. R. Burioni, L. Caniparoli and A. Vezzani
   Lévy walks and scaling in quenched disordered media. *Phys. Rev. E*, 81:060101 (2010).

In Chap. 1 we present and expand the results of the second paper. Sec. 2.3 in Chap. 2 is based on paper 1. The results in Chap. 3 are still unpublished and are the subject of the manuscript "The effect of codon translation rates on cotranslational folding mechanisms of arbitrary complexity", by L. Caniparoli and E. P. O'Brien, in preparation. The topics covered by the remaining papers are not included in this Thesis.

Acknowledgments

I am deeply grateful to M. Marsili and M. Vendruscolo. Their advice and encouragement guided me in the last three years, and made this Thesis possible.
I acknowledge the profitable and stimulating collaboration with R. Burioni, S. Lepri, P. Lombardo, E.P. O'Brien, Vu Giang Thi, and A. Vezzani. I especially thank Pierangelo and Edward. Working with them taught me lessons that cannot be learned from books.
I also would like to thank the present and former members of the group of M. Marsili at ICTP (M. Bardoscia, G. De Luca, G. Gori, G. Livan, F. Mancini, I. Mastromatteo, P. Vivo and E. Zarinelli), and A. Samal, for the many discussion we had. I especially thank Giancarlo for his obstinacy in supporting his arguments.
I thank my colleagues M. Beria, M. Marcuzzi and J. Marino, together with the Statistical Physics group staff in SISSA.

On the personal side, I would like to thank my parents for their constant support. The final thanks are dedicated to Giulia who, according to the common belief, is a very unruffled woman.

Contents

Introduction

The biological sciences are undergoing a profound revolution since the last decade, driven by major breakthrough in both the *quantity* and the *quality* of the experimental data.

The diffusion of entire genome sequencing techniques [104, 86, 73, 11] and high-throughput facilities (robotically assisted wet lab) [38, 99, 80], united with the easy accessibility of their results via web-based repositories [4, 7, 20, 64, 10, 21], is offering unprecedented amounts of data about the components of living systems. Most notably, the analysis of the thousands of genomes that have been sequenced [87] is revealing the mechanisms and principles through which the genetic information is maintained and utilized in living organisms [111, 81, 46, 16].

On the other hand, the introduction and the refinement of single molecule techniques such as the atomic force microscopy (AFM)[12, 52, 56], fluorescence resonance energy transfer (FRET) [55, 43], and optical tweezers [5, 61] have pushed the sensibility up to single molecule [98, 106, 77], allowing to probe the intrinsic stochastic nature of life at the microscopic level.

The analysis of large quantities of data on one side, and the comprehension of the molecular building blocks on the other, is revealing universal, emergent and quantitative regularities [72, 120, 94, 51, 58, 117]. This poses a grand theoretical challenge: how much of these empirical "laws" can be reproduced by simple modeling?

Statistical Physics can very relevantly contribute to this program, as it is the natural language for the stochastic modeling of the microscopic reactions and, potentially, for extrapolating the collective and macroscopic behaviors emerging out of them. Furthermore, the large size of the datasets and their intrinsic noisiness inevitably requires a statistical interpretation.

In this Thesis we use the means of Statistical Physics to approach protein translation. During this reaction the information stored in the cell (the DNA) is processed and utilized in order to correctly and reproducibly assemble the proteins. In particular, the 64 possible triplets of the 4 nu-

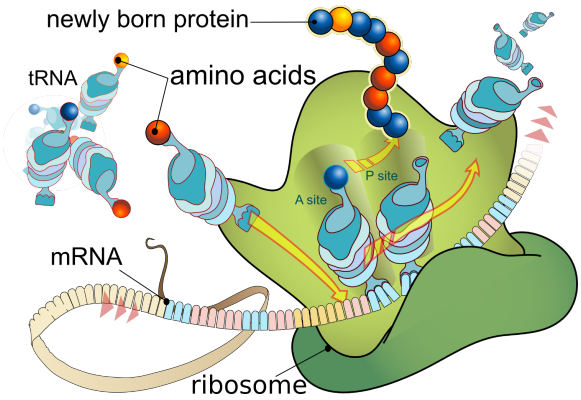

Figure 1: Cartoon of the translation process. The ribosome scans through the mRNA. The codons (triplets of nucleotides) are recognized by the tRNAs, which carry the amino acids. The newly assembled proteins exits the ribosome as the translation proceeds. (Adapted from Wikipedia)

cleotides (the codons) composing the messenger RNAs (mRNA) are sequentially scanned by the ribosomes (the molecular machines where the synthesis occurs) and translated into amino acids by the transfer RNAs (tRNA), see Fig. 1. The tRNA molecules are present in the cell in several variants (42 in Yeast), and each of them decodes one or more among the 61 sense codons (3 codons are used as the stop signals terminating the translation) into one of the 20 amino acids. Since the number of codons and tRNAs is greater than the number of amino acids, the genetic code is degenerate.

The first question we ask is whether this degeneracy is used to encode a further layer of information beyond the amino acid sequence. In Chap. 1 we analyze the mRNA sequences for detecting biases in the way the codons are used. We base our approach on Information Theory with the scope of analyzing the codon usage in the most abstract and model-free way. We are able to detect an anisotropy in the way the different codons are used across the genome, and we show that, mostly, this is a bias in the usage frequencies. The spacial organization of the codons (beyond that given by the amino acid sequence), although informative, seems to act as a second-order correction.

The following chapter (Chap.2) is devoted to models, whose scope is to explain the previously observed frequency anomaly. We analyze the effects of the finite amount of ribosomes and tRNAs by means of stochastic models. The picture which emerges from these models is that the optimization of the translation process is an intrinsically systemic problem, due to the finite amount of resources available in the cell. We also study how the interplay between the timescales of tRNA recharging, diffusion and translation can

shape the distribution of time interval between translation events, and lead to a non-exponential distribution.

A plausible hypothesis which can help in explaining the information in the codon spatial arrangement involves the folding of the protein while it is still tethered to the ribosome. The process of *cotranslational folding*, in fact, depends on the arrangement of the codons along the mRNA sequences: it was hypothesized that clusters of slow-translating codons can significantly help the folding of a protein domain by selectively slowing down the translation [71, 70, 123, 124]. The analysis of the information from the spacial organization of the codons, however, requires a level of detail and dependence on the actual sequence which cannot be easily obtained by analytical modeling. Besides, the computational burden inherent to this problem is overwhelming: the most interesting point resides in studying how the folding is affected by varying the translation rates, and it would require an immense power (a set of molecular dynamics simulations is required for each value of the parameters). In Chap. 3 we propose a method which overcomes this limitation by using a Markov chain formalism. This new methods makes the problem of cotranslational folding practically tractable.

Before plunging deep in the core of this work, let us briefly discuss Evolution and its implications because, as it has been observed [34], "Nothing makes sense in Biology except in the light of Evolution".

Selection and non-typicality

The central dogma of molecular biology [29] states that the information in the living system is stored in the genetic material (tipically, apart from RNA viruses, the DNA), that this information can be utilized to transcribe RNAs (like, e.g., mRNAs, tRNAs, miRNAs or structural components of cellular organelles), and subsequently, in the case of mRNAs, to encode proteins. It also states that the information flows unidirectionally from DNA to RNA, and from RNA to proteins.

However, this dogma does not capture the long term feedback from expression to genetic information which is at the core of evolution. In fact, on timescales longer than the lifetime of the single organism, the plasticity of the genomes[1] induces modifications on the DNA which are selectively

[1]The genetic material can change due to many different factors, whose relative importance depends on the organism. In the Prokaryotes, for instance, the main mechanisms are mutation and horizontal gene transfer (i.e., exchange of genes between different organisms). The latter is of extreme importance for antibiotic resistance, as different bacterial strains can exchange pieces of DNA, possibly containing the genes which confer the resis-

retained ("fixed") or discarded according to their benefit.

As the generations follow one another, the organisms having the highest capability to cope with the environment and reproduce rapidly -i.e., higher *fitness*- are favored. Effectively, the information flows from gene expression to DNA, and the living systems are *optimized* by selection in order to increase their fitness.[2]

The signs of the optimization process are evident, and we expect that the organisms which underwent selection for a specific feature will stand out when compared to the others. As a pictorial example, consider a specific trait, the neck length-to-height ratio, for many different animals. Most of the measures would disperse around a typical value, but at least the giraffe will stand out. In fact, unless the trait is under selective pressure, natural variation and mutations act randomly. The non-typical samples are produced when a specific trait undergoes selection for an underlying function: the giraffe has a long neck to reach out for the highest branches.

This idea has been applied successfully, for instance, to protein sequence analysis [119]. The sequences of the same protein from several different organisms were aligned and compared, keeping track of the mutations of the single amino acids. Among the many possible mutations, some of them repeatedly occur simultaneously and, since that pattern is very unlikely to be caused by chance, it is most probable that an underlying function drives the selection. In particular, it was shown that those informative couples are physically interacting during the fold of the protein, and the interaction induces the correlation of the mutations. The information extracted with this method is successfully used in assisting protein folding software.

In the following chapter we will apply this approach to one of the largest set of data, the sequenced genomes. In particular, we will study how the different codons (i.e., the different triplets of nucleotides encoding for an amino acid) are used. The synonymous mutations[3] do not alter the amino acid sequence and are often regarded as neutral and therefore random, as

tance.

[2]This qualitative representation is currently the subject of numerous quantitative studies, whose scope is in understanding how the modifications of the genotype (i.e., the DNA) affect the fitness of the organism. Intriguingly, experimental sampling of small fitness landscapes (of the order of up to 9 mutations, producing 2^9 different genotypes) is beginning to unveil the complexity of the interaction between different mutations, a mechanism known as epistasis (an in-depth review of these experiments and their theoretical interpretation can be found in Ref. [108]).

[3]The genetic code is degenerate, as 61 sense codons are translated into 20 amino acids. Most of the amino acids are encoded by more than one synonymous codons. A synonymous mutation which substitutes a codon with one of its synonyms produces the same amino acid sequence.

no selective pressure acts on them. This assumption is not completely true, as we show in the forthcoming sections, and there is information beyond the amino acid sequence. By understanding the causes and the implications driving the codon usage evolution, we believe that the comprehension of the whole translation machinery will be benefited.

Chapter 1

Learning from the data: the codon usage

1.1 The codon usage: information beyond the amino acid sequence

During translation, each amino acid is specified by a triplet of nucleotides (a codon). A given amino acid, however, may correspond to more than one codon, so that 61 codons correspond to 20 amino acids. For instance, lysine is encoded by two codons, valine by four and arginine by six. The way in which these synonymous codons are used shows a marked bias, a phenomenon know as codon bias [63, 91, 103, 47]. For instance, in humans the amino acid alanine the codon GCC (guanine-cytosine-cytosine) is used four times more frequently than the codon GCG. A question of central importance concerns the causes and consequences of the redundancy of the genetic code.

The codon bias is characteristic of a given organism and has been associated with three major aspects of mRNA translation, which are efficiency, accuracy and regulation [91, 47]. The first aspect is efficiency. The codon bias and the tRNA abundance in a given organism appear to have co-evolved for optimum efficiency [62, 17, 35]. Since synonymous codons (i.e., encoding for the same amino acid) can be recognized by different tRNAs and translated with different efficiencies, the codon bias is related to the translation rates [118, 109, 19, 114]. Moreover, codon usage has been shown to correlate with expression levels [49, 9, 48, 75, 123, 70, 31, 113], so that the use of particular codons can increase the expression of a gene by up to two or three orders of magnitude [53, 75].

The second aspect is accuracy. The codon bias affects the accuracy in

the translation, an effect that appears to have been optimized to reduce misfolding and aggregation [36].

The third aspect is regulation. Different codon choices can produce mRNA transcripts with different secondary structure and stability thus affecting mRNA regulation and translation initiation [75, 112]. The codon usage has also been associated with the folding behavior of the nascent proteins, by timing the co-translational folding process [26, 88, 70].

Since all these aspects of the translation process rely in some form on the information provided by the codon bias, it is most natural to use information theoretical techniques. In particular, in the following sections we first analyze the information content of the codon bias by analyzing the entropy and mutual information of the mRNA sequences of S.Cerevisiae, and then we introduce a measure for the information stored in the codon usage, the CII.

1.1.1 Codon usage entropy

The mRNAs are sequences of 61 different symbols, the codons, and the most natural way to analyze the properties of strings of symbols is by means of information theory.

Information theory, brought to worldwide attention by Claude Shannon in 1948 [102], deals with the question of how to quantify the information content of a message. The key quantity is the entropy, which captures the amount of uncertainty of a random variable.

In the case of mRNA sequences, we are interested in how the synonymous codons are utilized to encode each amino acid. We therefore consider each amino acid separately. Equivalently, we are supposing that each mRNA is composed by 20 independent sub-messages, whose entropy can be summed up. We can compute the entropy for the mRNA g as

$$S^{(g)} = \sum_a S^{(g)}(a) = -\sum_a \sum_{c \in a} f_c^{(g)}(a) \log_2 f_c^{(g)}(a), \tag{1.1}$$

where the first sum over a is intended over the 20 amino acids, the second sum over $c \in a$ runs over the codons c encoding for amino acid a, and $f_c^{(g)} = n_c^{(g)} / \sum_{c \in a} n_c^{(g)}$ are the frequencies of the codons encoding for the amino acid a.

The entropy Eq. (1.1) measures how much the codon usage is random for each of the amino acids. For instance, for a 2-codon amino acid like Lysine, $S^{(g)}(\text{Lys})$ reaches a maximum when $f_1^{(g)}(\text{Lys}) \simeq f_2^{(g)}(\text{Lys}) \simeq 0.5$ and is zero when $f_1^{(g)}(\text{Lys})$ is close to 0 or 1.

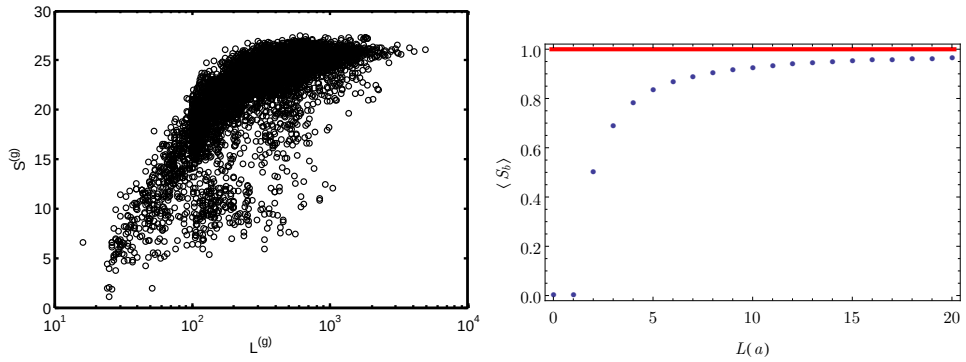

Figure 1.1: Left: entropy $S^{(g)}$ versus the length of the mRNA $L^{(g)}$. As the length of the transcript becomes shorter than ~ 200 codons, finite sampling imposes a clear cutoff for the entropy. Right: average value of the entropy of a Bernoulli distribution whose parameter p^* is estimated from L trials, Eq. (1.2) (blue dots). In this case the samples were generated with $p = 1/2$. The entropy approaches the asymptotic value 1 (red line) in the large $L(a)$ limit. On the other hand, as $L(a)$ becomes smaller, insufficient sampling effects reduce the average value of the entropy.

Let us compute the entropy $S^{(g)}$ for the entire set of mRNA of Saccharomicies Cerevisiae (from SGD database [21]), and as a first check let us plot it against the length of the mRNA (in codons), as in left panel of Fig. 1.1. This plot reveals that the entropy of the sequence is affected by a bias which grows as the length of the transcript becomes shorter than ≈ 200 codons.

The origin of this bias is inherent to the error in the estimation of the set of probabilities $\{f_c^{(g)}\}$ based on too few observations, which occurs when the length is reduced. In order to explain the mechanism, let us work out an example: suppose that L Bernoulli trials (biased coin tosses, head with probability p and tail with probability $1-p$) are produced, observing n_h heads and $n_t = L - n_h$ tails. The estimator p^* for the probability p, assuming binomially distributed n_h, is $p^* = n_h/L$.[1] The average value of the entropy of the empirical distribution as a function of the length is given by

$$\langle S_b(L) \rangle = \sum_{n_h=0}^{L} \left[P_{\rm B}(n_h|p,L) \sum_{i=h,t} \left(-\frac{n_i}{L} \log_2 \frac{n_i}{L} \right) \right], \qquad (1.2)$$

where $P_{\rm B}(n_h|p,L)$ is the binomial distribution. In the right panel of Fig. 1.1 we plot $\langle S_b(L) \rangle$ as a function of the length L. The estimated entropy exhibit a strong bias as the size of the sample is reduced.

[1] The estimator $p^* = n_h/L$ is the Maximum Likelihood estimator.

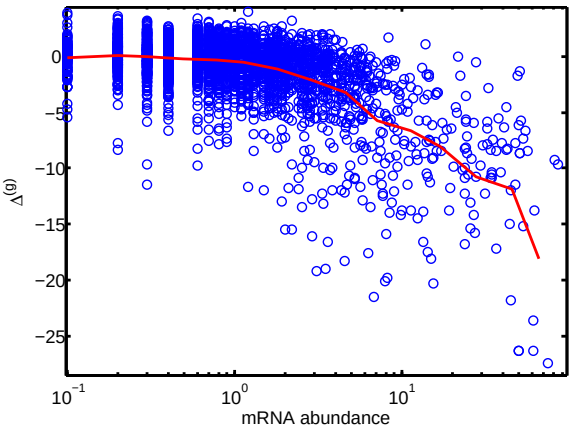

Figure 1.2: Plot of the ratio $\Delta^{(g)}$, Eq. (1.3), versus mRNA abundance. Interestingly, the entropy of the resampled sequence typically increases with respect to the true sequence as the abundance of the transcript increases. The red line is a moving average.

The bare entropy cannot be used as a measure of how much the codon usage is biased due to the previous length dependence. However, we are interested in analyzing how untypical the codon usage of a mRNA is. To do so we therefore compute the average codon usage $\{\bar{f}_c\}$ across the whole transcriptome and, for every gene g, we produce M randomized samples such that I) we keep the number of each amino acid $N^{(g)}(a)$ fixed, and II) the number of codons $\{n_c^{(g)}\}_{c\in a}$ of each amino acid a is generated from the multinomial distribution $P_{\rm M}(\{n_c^{(g)}\}|\{\bar{f}_c\}, N^{(g)}(a))$. Using this method we can estimate how far the entropy of the true sequence is from a typical sample having an average codon usage. By using the average $\bar{S}_R^{(g)}$ and the standard deviation $\sigma_R^{(g)}$ of the entropy of the resampled sequences, we can compute the ratio

$$\Delta^{(g)} = \frac{S^{(g)} - \bar{S}_R^{(g)}}{\sigma_R^{(g)}}. \tag{1.3}$$

We plot this ratio versus the experimentally measured mRNA abundance in Yeast (the data are from [37, 60]) in Fig. 1.2. Interestingly, the entropy $S^{(g)}$ of the true sequences decreases with respect to the average $\bar{S}_R^{(g)}$ as the abundance increases, clearly showing that the codon usage is not constant across the genome. Furthermore, the codon usage of abundant mRNAs tends to be significantly less random (i.e., less even) than that of the less abundant mRNAs.[2]

[2]The early observations that the bias increases with the abundance of the mRNA date

Note that entropy does not discriminate 1) the codon species (in a 2 codon case, low entropy is obtained with both $n_1 = 0, n_2 = 10$ and $n_1 = 10, n_2 = 0$), implying that mRNAs with equally low entropies might have radically different codon usages, and 2) the codon order, as it is computed from the frequencies only. In particular, codon order could carry some information, and in the next section we will analyze it.

1.1.2 Codon order and mutual information

The quantity $\Delta^{(g)}$ introduced in the previous section is a good measure of how much the codon usage is biased. However, its computation completely neglects codon order, as it relies on the frequencies only.

Let us consider the codon c_i at position i along a mRNA sequence. What can we say about the codon c_{i+d} at position $i + d$? Does the information about c_i tell us something about the following codons? Equivalently, we can ask how much the entropy of the distribution of c_{i+d} is reduced by knowing c_i. This quantity can be written as $S(c_{i+d}) - S(c_{i+d})_{|c_i}$, where $S(c_{i+d})_{|c_i}$ is computed using the probability distribution for c_{i+d} conditioned to c_i. Summing over the distribution of c_i, we obtain the mutual information, which in the case of two random variables X and Y can be written as

$$M(X,Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \tag{1.4}$$

where $p(x, y)$ is the joint distribution and $p(x)$ and $p(y)$ are the marginals. The mutual information measures how far the joint probability $p(x, y)$ is from being factorizable as $p(x)p(y)$, i.e., how much X and Y deviate from being independent.

In order to compute the mutual information between codons separated by a distance d, we need the knowledge of the joint probability for codons C_1 and C_2, which can be easily extracted from S sequences as

$$p_d(C_1, C_2) = \frac{1}{\mathcal{N}} \sum_{s=1}^{S} \sum_{i=1}^{L_s-d} \delta_{C_1,c_i(s)} \delta_{C_2,c_{i+d}(s)}, \tag{1.5}$$

where L_s is the length of sequence s, $\mathcal{N}$ the normalization, and $C_1, C_2 = 1, \ldots, 61$.

The idea of using mutual information is not new, as already in the '90s it was observed [40] that it's not vanishing. However, the "bare" mutual information is not a well-behaved measure of the information carried by the

back to the '80 [49, 63, 17, 103, 2].

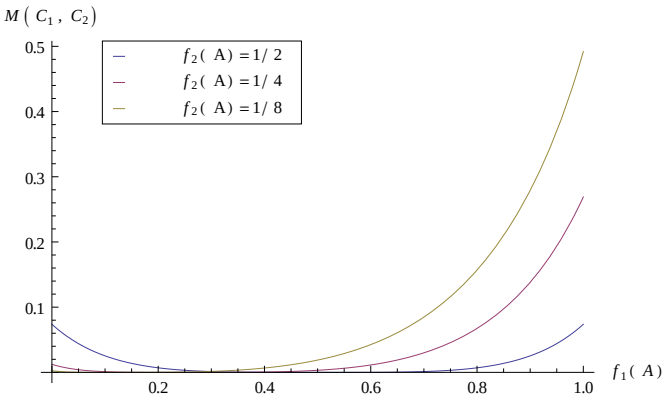

Figure 1.3: Mutual information between sites at distance d on two sequences of the same length composed by two possible symbols, computed using Eq. (1.6). $M(C_1, C_2)$ is plotted as a function of the frequency $f_1(A)$ of the symbol A on the first sequence, for several values of the frequency $f_2(A)$. When $f_1(A) \neq f_2(A)$ the mutual information is positive.

codon order for two main reasons. First, we are probing the properties of a "meta-sequence" of symbols, as the sequence of the codons is superimposed over that of the amino acids. A non zero mutual information could be trivially due to an informative sequence of amino acids with completely random codons.

Another very subtle bias is due to the inhomogeneity of the codon usage across the sequences: let us consider a simplified case, with two sequences of length L_1 and L_2, composed by only two possible symbols, A and B. Let us also suppose that the symbol A has a frequency $f_1(A)$ and $f_2(A)$, respectively on the two sequences. Assuming that the symbols are completely random and that the sequences are long enough, the joint probability for two symbols (at any distance) reads

$$p(C_1, C_2) = \frac{\sum_{s=1,2} L_s f_s(C_1) f_s(C_2)}{\sum_{s=1,2} L_s}. \tag{1.6}$$

Even though the two sequences are not spatially organized, the mutual information computed from this joint probability is non-vanishing (apart from the points $L_1 = 0$, $L_2 = 0$, and $f_1(A) = f_2(A)$), as show in Fig. 1.3.

These two biases can be accounted for by subtracting the trivial part of the information, computed on the sequences randomized such as to disrupt the spacial organization of the codons while leaving unaltered 1) the amino acid sequences, and 2) the codon frequencies inside each sequence. We there-

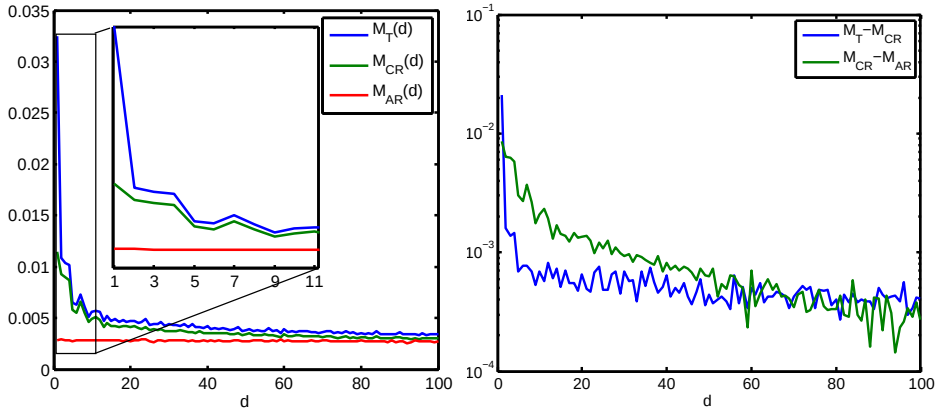

Figure 1.4: Mutual information between codons at distance d along the sequences. On the left, the mutual information for the true sequences is compared with that for the reshuffled sequences, showing that the largest deviation is for $d = 1$. The differences $M_{\rm T}(d) - M_{\rm CR}(d)$ and $M_{\rm CR}(d) - M_{\rm AR}(d)$ are plotted in the right panel. Interestingly, $M_{\rm T}(d) - M_{\rm CR}(d)$ seems to relax to a constant. On the other hand, $M_{\rm CR}(d) - M_{\rm AR}(d)$

fore define the "codon reshuffle" protocol (CR), such that the codons of each amino acid are randomly permuted inside each transcript. In order to also measure the information stored in the amino acid ordering, we also introduce the "amino acid reshuffle" protocol (AR): the codons of each mRNA are randomly permuted, therefore altering the amino acid sequence. Let us name $M_{\rm T}(d)$, $M_{\rm CR}(d)$ and $M_{\rm AR}(d)$ the "true" and the reshuffled mutual informations. The difference $M_{\rm T}(d) - M_{\rm CR}(d)$ measures how much information is encoded in the order of the codons, while $M_{\rm CR}(d) - M_{\rm AR}(d)$ measures the contribution due to the sequence of amino acids.

These quantities are plotted in Fig. 1.4 for S.Cerevisiae.[3] Interestingly, $M_{\rm T}(d)$ is much larger than $M_{\rm CR}(d)$ for $d = 1$, and it relaxes rapidly. The difference is therefore due to the spacial ordering of the codons. The residual dependence of $M_{\rm CR}(d)$ on the distance is due to the organization of the amino acids, proved by the fact that reshuffling the amino acids produces a constant $M_{\rm AR}(d)$.

This finding implies that the spacial organization of the codons is not

[3]The mutual informations for the reshuffled sequences should be regarded as random variables as their value could fluctuate from sample to sample. However, the data set is so large that their behavior can be regarded as self-averaging, as the variance is at least 2 orders of magnitude smaller than the mean.

completely random. So we can ask what this information is used for and how codon order affects translation. Plausibly, the most important effects are I) recycling of the tRNAs which have been used recently [19], II) folding of the mRNA [75], III) translation pauses induced by slow codon clusters in order to let the protein fold cotransaltionally [70, 88], and IV) flow of the ribosomes along the mRNA [82, 97, 23].

Since all these aspects of the translation process rely in some form on the information provided by the codon bias, it is very important to first address the question of establishing a measure of the amount of information encoded in the codon bias itself. For this purpose in the next section we introduce the Codon Information Index (C_{II}).

1.2 The Codon Information Index (*CII*)

In the previous sections we showed that the way the codons are used is not homogeneous across the genome. Furthermore, there is some degree of spacial organization, in the sense that the order of the codons along the sequence is less random than expected, mostly for codons at distance one. However, we only *detected* the presence of information. In order to answer the fundamental question of what the information is used for, we need to carefully measure the amount of information stored in the codon usage at each site along the sequences. This level of detail permits the analysis of the anisotropies of the information inside each mRNA.

In the following sections we introduce the Codon information Index (CII), by using a combination of statistical mechanics and information theoretical techniques. We first analyse the general properties of the CII and then we apply it to a pool of 3371 genes of yeast. We find that the CII correlates with protein and mRNA abundances, as well as with the tRNA Adaptation Index (tAI) [96]. The latter result shows that two independent forms of information, which are stored in different parts of the genome, the tRNA copy number and the codon bias in the coding region, are remarkably dependent on one another.

1.2.1 Measures of the codon bias

Several measures of the codon usage exist. Let us briefly review them:

- The "effective number of codons" ($\hat{N}$) [121]) is based on population genetics results. Each amino acid a is considered as a different locus with a number K_a of alleles equal to the number of codons. The frequencies p_i of use of each of those alleles are used to compute the

"homozygosity" $F_a = (n_a \sum_{i=1}^{K} p_i^2 - 1)/(n_a - 1)$, where n_a is the number of the amino acids a. The effective number of codons N_a for the amino acid a is simply $N_a = F_a^{-1}$, and the overall $\hat{N}$ is computed by summing the contributions of all the amino acids. This index does not depend on any data beyond the mRNA sequence.

- The "frequency of optimal codons" (F_{opt}) [62] of a sequence is defined as $F_{\text{opt}} = N_{\text{opt}}/N_{\text{tot}}$, where N_{opt} is the number of codons identified as optimal and N_{tot} is the total number of codons. The identification of the "optimal" codons is based on their interaction with the tRNA and tRNA abundances.

- The definition of the "Codon Bias Index" (CBI) [9] is similar to that of F_{opt} and, fundamentally, differs by a constant.

- The "Codon Adaptation Index" (CAI) [103]) for a sequence S is defined as the geometric mean of the weights w_c associated to each codon. The weights are computed from the frequencies as $w_c = f_c / max_{c' \in a} \hat{f}_{c'}$, where the codons c and c' are synonymously translating the amino acid a. Note, however, that $\hat{f}_c$ is computed for a set of highly expressed genes.

- The tRNA adaptation index (tAI) [96] is the most recent one. It is defined for a mRNA sequence g as

$$tAI = \left(\prod_{i=1}^{L_g} w_{c_i} \right)^{1/L_g}, \tag{1.7}$$

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

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

where t_i is the number of tRNA isoacceptors recognizing codon i. The parameter tGCN_{ij}, referring to the jth tRNA recognizing codon i, and is the number of the copies of the tRNA gene ij present in the genome of the organism under consideration. The set of parameters s_{ij} gauges the efficiency of codon-anticodon coupling and are optimized to maximize the correlation of the tAI with protein abundance. The computation of this index requires information about two of the most

influential factors affecting translation efficiency, namely tRNA abundance (tRNA copy number is highly correlated to tRNA abundance [89]) and codon coupling efficiency.

Among these measures, the CII is specifically designed to describe the amount of information encoded in mRNA sequences through the codon bias, and it is the only one with the following properties: *i)* it requires only information about the mRNA sequences and does not depend on any additional data: the CII is self-contained, and *ii)* it produces a codon-wise profile for each sequence. As examples, we mention in particular that, for a given organism, the tAI requires the knowledge of the tRNA pool, and that the CAI requires the knowledge of the most expressed genes.

1.2.2 Construction of the CII

A natural representation of the information contained in the codon bias can be given in terms of strings of bits (i.e. $(0, 1)$ variables) or of distributions over bit strings. In other words, we associate a bit to each codon. This procedure corresponds to bin the codons into two classes, each with its own codon usage distribution, given by the frequencies of the codons in that class. The first ingredient to build the CII is an assignment of bits to codons that is maximally informative, in a way to be specified later. This also implies that the information encoded in this way is optimally retrievable. Pictorially, we are supposing that each codon encodes a hidden binary property (e.g., fast or slow). The maximally informative assignment is such that, given the string of bits, the information about the string of codons is maximized: we want to compress the codon code into a 2 symbols code.

The second ingredient is a local codon organization along the sequence (see Fig. 1.5). We analyze these two contributions separately below.

Maximum information

Let us first consider the special case of a protein of length L composed by only one amino acid type, which can be translated by K different codons [6]; the generalization to the full set of amino acids is described below. The sequence $\{c_1 \dots, c_L\}$, $c_i = 1, \dots, K$ is given, where c_i is the i^{th} codon. We associate a binary variable $s_i = \pm 1$ (i.e. a *spin* variable, rather than a *bit* ($\{0, 1\}$) variable) to each site of the sequence; this spin variable identifies the class each codon is assigned to.

Let $p_{c|s}$ be the probability for the codon c to be used on a site with spin s. A priori, the only information available is that an amino acid can be

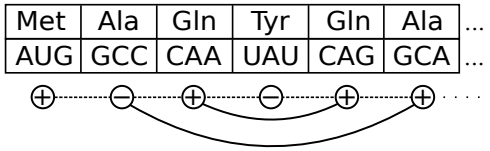

Figure 1.5: In the calculation of the CII each codon has an associated binary variable. Synonymous codons interact via the information theoretic part of the Hamiltonian (solid lines), nearest neighbor sites interact via the Ising interaction (dashed line)

encoded by any of its possible codons. This state of ignorance is described by the choice of a uniform prior distribution for the probabilities of the parameters $p_{c|s}$, with the normalization ensured by a δ function

$$P_0(\hat{p}) = \prod_{s=\pm 1} \Gamma(K)\ \delta\left(\sum_{c=1}^{K} p_{c|s} - 1\right). \tag{1.8}$$

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

$$P(\vec{n}_s|\hat{p}) = \prod_{s=\pm 1} \frac{\Gamma\left(N_s + 1\right)}{\prod_c \Gamma(n_s(c) + 1)} \prod_c p_{c|s}^{n_s(c)}, \tag{1.9}$$

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

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

$$P(\hat{p}|\vec{n}_s) = \prod_{s=\pm 1} \frac{\Gamma\left(N_s + K\right)}{\prod_c \Gamma(n_s(c) + 1)} \prod_{c=1}^{K} p_s(c)^{n_s(c)} \delta\left(\sum_c p_s(c) - 1\right). \tag{1.10}$$

An important quantity that can be derived from the prior and the posterior is how much information is gained by observing the codon frequencies. This requires to compute how many extra bits must be used to code the samples from the posterior compared to the prior. This quantity is the Kullback-Leibler divergence, which is defined, for two distributions $P(x)$ and $Q(x)$, as

$$D_{\mathrm{KL}}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

To our purposes it is useful to use the symmetrized version of the KL divergence:

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

where the averages are performed with respect to the posterior, Eq. (1.10), and the prior, Eq. (1.8), respectively. The integration can be performed analytically and leads to

$$I(\vec{n}_s) = -\sum_{s=\pm 1} \sum_{c=1}^{K} n_s(c) \Big[\psi\left(N_s + K\right) - \psi\left(n_s(c) + 1\right) \Big] + \text{Const}, \qquad (1.11)$$

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

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

$$I(\{\vec{n}_{s,a}\}) = -\sum_{a=1}^{20} \sum_{s=\pm 1} \sum_{c_a=1}^{K_a} n_{s,a}(c_a) \big[\psi(N_{s,a} + K_a) - \psi(n_{s,a}(c_a) + 1) \big] + \text{Const}, \qquad (1.12)$$

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

To extract as much information as possible from the codon counts we have to maximize Eq. (1.12). However, the contributions of different amino acids are independent and each amino acid sector is invariant under a spin flip. Therefore the minima of Eq. (1.12) are highly degenerate. Moreover, the amino acids with only one codon, methionine and tryptophan, do not contribute to Eq. (1.12).

These issues can be cured observing that the previous derivation does not use any information about how the codons are arranged along the sequence. Therefore, information carried by the codon order can be used to weight the minima of Eq. (1.12).

Codon spatial organization

In order to remove the degeneracy, we add to Eq. (1.12) an interaction between nearest-neighbour spins that favors their alignment. This coupling is also consistent with the observation of the existence of a "codon pair bias" [54, 27]. We thus define the cost function

$$H\{\vec{s}\} = -J \sum_{i=1}^{L-1} s_i s_{i+1} - I(\{\vec{n}_{s,a}\}), \qquad (1.13)$$

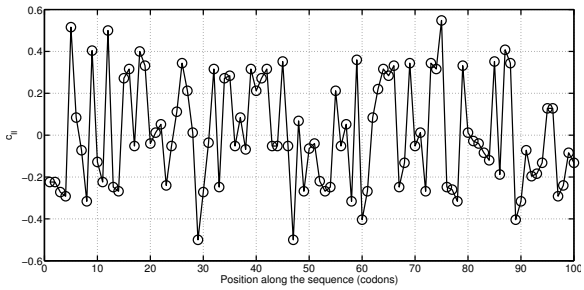

Figure 1.6: Local CII, as in equation (1.14), for the first 100 codons of the TFC3/YAL001C gene.

where J is a parameter to be tuned which accounts for the degree of spatial homogeneity of the sequence. In statistical physics terms, Eq. (1.13) can be regarded as the Hamiltonian of a spin system, that besides the 1D Ising interaction J, also has a long range interaction I as shown in Fig. 1.5 [4]. This analogy makes it possible to apply techniques used to study spin systems in statistical physics to the present model.

We are interested in the spin arrangements minimizing the Hamiltonian (1.13), given the codon sequence. Numerically, energy minimization was performed by Simulated Annealing Monte Carlo [67]. When the states of minimal H were found to be degenerate, an average over all of them was considered.

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

We define the Local Codon Information Index as the magnetization at site i on the transcript g, i.e. the thermodynamic average of the spin at site i

$$c_{II}^{(g)}(i) = \langle s_i^{(g)} \rangle \tag{1.14}$$

and the Codon Information Index of the gene g as the average of the local CII on the gene codons

$$C_{II}^{(g)} = \frac{1}{L_g} \sum_{i=1}^{L_g} c_{II}^{(g)}(i). \tag{1.15}$$

[4] The Hamiltonian (1.13) is invariant under a global spin flip, thus the average magnetization would be zero. To break this symmetry it is sufficient to add a term $H_h = h \sum_i s_i$, which favours the configuration aligned with the external field h. The field h will be taken vanishingly small ideally, very small in practice.

1.2.3 Phase diagram of the Hamiltonian (1.13)

In this section we characterize the properties of H in Eq. (1.12) and its minima. We also prove the existence of a phase transition in temperature at $J = 0$. Finally, we describe the effect of the nearest neighbor interaction.

Maximum of Eq. (1.12)

As a first step in the characterization of (1.12), we want to show that each term of the form (1.11) has a minimum and is convex. Let's rewrite (1.11) setting $n_{\pm}(c) = n(c)/2 \pm \delta_c$, $\delta_c \in [-n(c)/2; n(c)/2]$ and $\Delta = \sum_c \delta_c$

$$\begin{aligned} I(\vec{\delta}) &= -\sum_{s=\pm 1} \Bigg[\left(\frac{N}{2} + s\Delta\right) \psi\left(\frac{N}{2} + s\Delta + K\right) \\ &\qquad - \sum_{c=1}^{K} \left(\frac{n_c}{2} + s\delta_c\right) \psi\left(\frac{n_c}{2} + s\delta_c + 1\right) \Bigg] \\ &= \sum_{s=\pm 1} \Bigg[- g_K\left(\frac{N}{2} + s\Delta\right) + \sum_{c=1}^{K} g_1\left(\frac{n_c}{2} + s\delta_c\right) \Bigg] \\ &= -G_K\left(\frac{N}{2}, \Delta\right) + \sum_{c=1}^{K} G_1\left(\frac{n_c}{2}, \delta_c\right) \end{aligned} \tag{1.16}$$

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

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

$$\frac{\partial I}{\partial \delta_i} = -G'_K\left(\frac{N}{2}, \Delta\right) + G'_1\left(\frac{n_i}{2}, \delta_i\right) = 0,\ i = 1 \dots K \tag{1.17}$$

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

Therefore, the maxima must reside on the boundary. Repeating the argument on the boundary faces, we end up concluding that the maxima must lie on the boundary vertexes, i.e. the points such that $\delta_i^* = \pm n_i/2$.

On these points I becomes

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

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

The maximum is thus obtained on the vertexes which minimize the difference $|N_+ - N_-|$ (e.g., for four codons with $\{n(c)\} = (5, 3, 4, 2)$ the maximum is obtained for $(+, -, -, +)$ or $(-, +, +, -)$, since I is symmetric under a global spin flip). This is an instance of the number partition problem which belongs to the NP-complete class. However, we are dealing with sets which contain at most 6 elements.

Considering the full set of amino acids, we can finally ask how many states have the same $I(\{\vec{\delta}_a\})$. Using the fact that the contribution for each amino acid is invariant under a spin flip and that the amino acids with one codon only are not considered, there are at least 2^{18} states in addition to the trivial degeneracy coming from the amino acids methionine and tryptophan which do not contribute to (1.12).

Phase transition in temperature at $J = 0$

It is possible to analytically work out the thermodynamics of Eq. (1.13) at $J = 0$. At high temperatures we expect a disordered paramagnetic phase, while at low temperatures the system falls into one of its many minima, which correspond to the maxima of Eq. (1.12) described in the previous section. Here we prove that a phase transition exists by showing that the concavity of the free energy changes sign at a critical temperature T_c in the large n_c limit.

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

$$S(\vec{n}_+|\vec{n}) = \log \prod_c \binom{n_c}{n_+(c)} \tag{1.18}$$

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

$$F = \left[G_K(N/2, \Delta) - \sum_{c=1}^{K} G_1(n_c/2, \delta_c)\right] - T \sum_{c=1}^{K} \log \binom{n_c}{\frac{n_c}{2} + \delta_c}. \tag{1.19}$$

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

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

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

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

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

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

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

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

Using some combinatorics, we obtain

$$\begin{aligned}P_K(\lambda) = (-\lambda)^K + \sum_{i=1}^{K}(-\lambda)^{K-i}\Bigg[&\sum_{j_1<\ldots<j_i} a_{j_1}\ldots a_{j_i} \\ &+ b(K-i+1)\sum_{j_1<\ldots<j_{i-1}} a_{j_1}\ldots a_{j_{i-1}}\Bigg] \qquad (1.20)\end{aligned}$$

with the convention $\alpha_0 = 1$. If $T > T_c^+ = \max_c(1-n_c^{-1})$ we immediately see that each coefficient is positive and thus the Hessian is positive definite, while if $T < T_c^{(-)} = \min_c(1-n_c^{-1})$ the Hessian is negatively defined because of the Descartes' rule of signs.[5] At $T < T_c^{(-)}$ the free energy becomes convex and

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

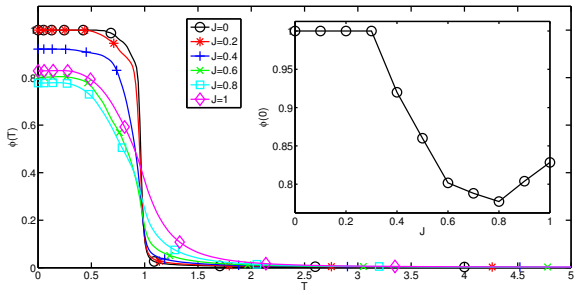

Figure 1.7: (Color online) Codon coherence $\phi_J(T)$ as a function of the temperature for increasing J and codon coherence at $T = 0$ (inset). As the nearest neighbor interaction is turned on, the phase transition becomes smoother. At $T = 0$, the codon coherence is preserved up to a critical value of $J \sim 0.3$.

the ground states moves discontinuously far from the paramagnetic ($\vec{\delta} = \vec{0}$) state.

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

$$\phi_{J=0}(T) \equiv \sum_{c \neq M,W} \left\langle \left(\frac{2\delta_c}{n_c} \right)^2 \right\rangle_T \tag{1.21}$$

where the sum is intended on every codon except Methionine and Tryptophan and the thermodynamic average is performed at temperature T. This parameter is small for paramagnetic configurations (which have $2\delta_c = n_+(c) - n_-(c) = o(n_c)$), while is one in the partitioning phase. Its plot is in figure 1.7: the transition is evident, although the presence of finite size effects smooths the step out a bit.

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

The introduction of the nearest neighbor interaction makes the analytical treatment much more difficult. Nevertheless, we expect that the phase transition becomes smoother and smoother as J is raised, since the Ising model in 1D does not exhibit any phase transition. This observation is numerically tested in figure 1.7, where the profiles of $\phi_J(T)$ are plotted for increasing J.

The $T > 1$ behavior is easily interpreted by observing that in the paramagnetic phase ($\delta_c \ll n_c$) the information theoretical part of the Hamiltonian is flat around $\vec{\delta} = 0$ in the large n_c limit. We expect the high temperature ($T > 1$) behavior to be dominated by the magnetic field and the nearest neighbor interaction terms: excluding the information theoretical part, the Hamiltonian reduces to the Ising model's one: $H_{\text{Ising}} =$

$-J\sum_{i=1}^{L-1}\sigma_i\sigma_{i+1} - h\sum_{i=1}^{L}\sigma_i$. Thus, the thermodynamics at $T > 1$ should be described by the phenomenology of the Ising model.

To check this hypothesis, we numerically computed the magnetization for the Hamiltonian (1.13)

$$m = \frac{1}{\sum_c n_c}\left\langle \sum_c n_+(c) - n_-(c) \right\rangle \tag{1.22}$$

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

$$m_{\text{Ising}} = \frac{1}{L}\left\langle \sum_{i=1}^{L}\sigma_i \right\rangle = \frac{\sinh(h/T)}{\sqrt{e^{-4J/T} + (\sinh h/T)^2}} \tag{1.23}$$

These quantities are plotted in figure 1.8, where is clearly shown that the Ising model correctly describes the $T > 1$ behaviour of (1.13).

We introduced the nearest neighbour interaction to weight the many minima of the information theoretical part of the Hamiltonian and to extract information from the spatial arrangement of the codons. Observing the inset of figure 1.7, we see that at $J > 0.3$ the codon coherence at $T = 0$ is lost. This means that the same codon is assigned a different CII on different positions. Since there is no a priori biological reason for this differentiation, we restrict the admissible J to those such that $\phi_J(0) = 1$. Moreover, since we want to maximize the information extracted from the spacial arrangement of the codons, we fix J as the maximum value such that $\phi_J(0) = 1$. Interestingly, we find that for this value of J the correlation of $C_{II}^{(g)}$ with the tAI exhibits a maximum.

1.2.4 Analysis of the CII

CII correlates with protein and mRNA abundance, as well as with the tAI

We computed the CII for a set of 3371 transcript of *S. cerevisiae* and we compared it with the logarithms of the measured protein [83] and mRNA [60] abundances[6], observing a significant correlation ($C \simeq 0.60$ and $C \simeq 0.69$ for proteins and mRNAs, respectively, Fig. 1.9). Furthermore, computing the same quantities for the half set comprising the most abundant proteins and mRNAs we observe a sharp increase in the correlation coefficients $C \simeq 0.70$ and $C \simeq 0.79$ for proteins and mRNAs, respectively.

[6]The protein and mRNA abundances are correlated. However, the non-homogeneity of the two samples (the data come from different laboratories and use different techniques) hinders the possibility of asking how much of the protein abundance is explained by the CII beyond the mRNA abundance.

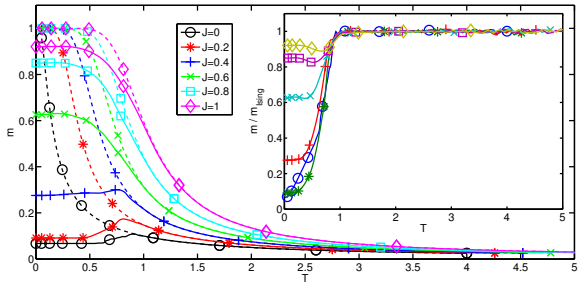

Figure 1.8: (Color online) Solid lines: numerically computed magnetization as a function of temperature (equation (1.22)). Dashed lines: analytically computed magnetization of the Ising model in 1D (equation (1.23)). The inset shows the ratio m/m_{Ising}. The magnetization for $T > 1$ is correctly described by the Ising model, and as J is is raised the behaviour at $T < 1$ becomes more and more similar to Ising model's.

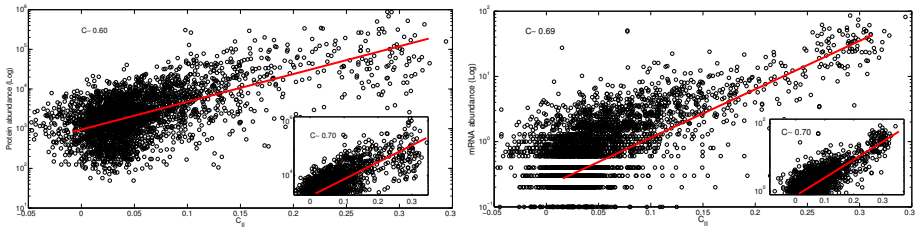

Figure 1.9: The CII correlates with the logarithm of protein abundance (left) and mRNA abundance (right). The correlation is most evident for the most abundant half of the set (insets).

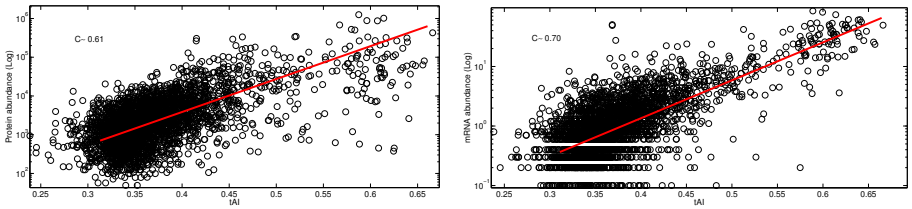

Figure 1.10: The tAI is highly correlated with protein (left) and mRNA (right) abundance.

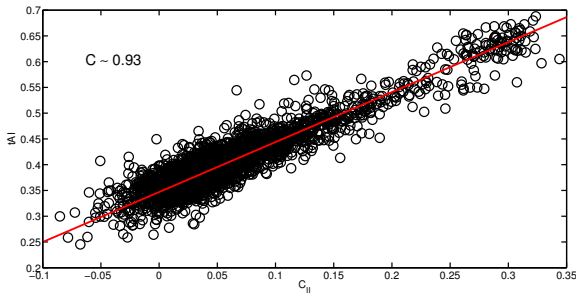

Figure 1.11: The CII is highly correlated with the tAI

Note that mRNA abundance can radically vary during the life cycle of the cell. In Ref. [13] it was shown that, during mitosis (the M-phase)[7], the expression of many genes can be up- or down-regulated by one order of magnitude or more with respect to the vegetative phase. However, the M-phase usually accounts for a small part of the whole cell cycle (usually, less than 1/10th). The abundances obtained in an experiment which samples a cell colony are therefore largely dominated by the vegetating -i.e., growing- cells.

This fits very well with another observation on the high-CII genes. We analyzed the Gene Ontology[8] of the set of the 200 highest CII genes. We found that the most over-represented terms in this set where translation- and metabolism-related (with an extremely high significance, as 124 out of 200 carried the tag Transcription, with a p-value smaller than 10^{-20}). These genes are typically expressed at a high level during cell growth.

[7] During mitosis the eucaryotic cell divides, as opposed to the interphase, when the cell grows.

[8] The genes are annotated according to their molecular function, the biological process they are involved in, and the cellular component where they are typically located. Given a set of genes of interest, one can ask whether a set of those tags are over-represented with respect to the whole set of genes. We used the Gene Ontology web-interface available at the Saccharomyces Genome Database [21].

	CII	tAI	Protein levels (log)	Half life (log)	mRNA levels (log)
CII	1	0.93	0.60 (0.70)	0.18	0.69 (0.79)
tAI		1	0.61 (0.68)	0.20	0.70 (0.77)
Abundance (log)			1	0.30	0.60
Half life (log)				1	0.22

Table 1.1: Pearson correlation coefficients between numerical and experimental quantities. The values between parentheses, when present, refer to correlations computed for the most abundant half of the set

We also computed the tAI for the same 3371 transcripts and compared it with the CII. We observe an extremely high correlation ($\rho \sim 0.93$), as shown in Fig. 1.11. We thus are able to reproduce all the results obtained from the tAI without needing any additional information beyond the codon sequences and without any parameter optimization: the CII depends only upon the parameter J which can be fixed from thermodynamics considerations, as explained in section 1.2.3.

The tAI is known to correlate well with protein abundance ($\rho \simeq 0.61$) and mRNA abundance ($\rho \simeq 0.70$), see Fig. 1.10. Moreover, as in the previous case the correlation improves for the most abundant proteins ($\rho \simeq 0.70$) and mRNAs ($\rho \simeq 0.77$), but to a significantly smaller extent.

Another quantity usually taken in consideration in this kind of studies is proteins half life. We computed correlations among all these quantities, the results are summarized in Table 1.1. Between parenthesis are the correlations computed for the 1600 most abundant proteins and mRNAs. All these values are significative (P-value $\sim 10^{-9}$ at most). Those involving CII, tAI, protein and mRNA abundance are highly significant (P-value $< 10^{-20}$).

It has been suggested that the correlation between the CII and the mRNA levels can be caused by evolutionary forces acting more effectively on highly expressed genes [75, 69], as beneficial codon substitutions are more likely to be fixed on these genes because the gain in fitness is likely to be higher, although a fully causal relationship can be more complicated and involve other determinants [47, 91],

Average CII profile along the proteins

In the previous sections we analyzed the properties of the global CII value for whole genes, but the $c_{II}(i)$ gives another local layer of information. We thus ask whether the local $c_{II}(i)$ can be interpreted as a local measure of translational optimality. Unfortunately too little data exist to confirm or falsify this hypothesis, but we can explore if a common behavior at the

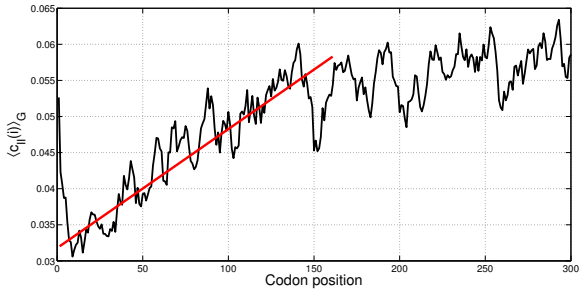

Figure 1.12: Local CII averaged along the proteins, as in equation (1.24). The straight line is a guide for the eye.

beginning of the transcript exist. Similarly to what have been done in the case of the tAI [112], it is possible to compute the average of the $c_{II}^{(g)}(i)$ across the transcripts

$$\langle c_{II}(i) \rangle_G = \frac{1}{N_G} \sum_{g=1}^{N_G} c_{II}^{(g)}(i). \tag{1.24}$$

Its plot in figure 1.12 reveals the presence of a "ramp" roughly 120 codons long followed by a plateau.

This result is consistent with the findings in [112] for the tAI, where this procedure reveals a signal at the beginning of the transcript: the average local tAI has a minimum at the beginning of the sequence and rises up to the average value in ~ 100 codons. Since the authors claim that the tAI carries information about codon translational efficiency, they hypothesize that this feature helps translation stabilization by avoiding ribosome jamming.

Final considerations about the CII

Let us conclude this long section by reviewing some of the most important points. We introduced the Codon Information Index (CII) as a measure of the amount of information stored in mRNA sequences through the codon bias. We showed that CII can capture at least as much complexity as previously introduced codon bias indexes, but its computation does not require additional data beyond transcript sequences. In order to calculate the CII we do not make any assumption on the origins and roles of the codon bias, but quantify the amount of information associated with it in an unbiased manner.

We calculated the CII for a set of over 3000 yeast transcripts and found values highly correlated with the tAI scores, as well as with experimentally-derived proteins and mRNAs abundance. A Gene Ontology analysis on the

set of the high-CII genes revealed a over-representation of metabolic and translational genes. This genes are typically highly expressed when the cell is rapidly growing.

Furthermore, we were able to reproduce the result that the first $70 \div 100$ codons have a lower average when compared to the remaining part of the ORF. This part is thought to be translated with low efficiency, a feature which should help translational stabilization [112].

1.3 CII and reshuffles. A case study: Yeast

The choice of the different synonymous codons along the mRNA sequences is thought to play an important role in gene expression, by modulating it (via controlling the flow of the ribosomes, see, e.g., [97, 82, 23]), and by mediating pauses which help the proteins to fold (this will be the subject of Chap. 3). These two regulatory layer are being very actively investigated, as a conclusive and definitive understanding is still lacking.

What hindsight can we extract from the CII about these questions? How does the CII depends on the codon order? In the following sections we show that, unexpectedly, the computation of the CII is not influenced by codon order. This observation implies that most of the information encoded in the codon bias is on the codon frequencies, and that their spacial organization is a second order correction. Furthermore, we utilize this finding to obtain a much faster and easily implementable algorithm, as shown in Sec. 1.3.2.

1.3.1 Reshuffling the codons do not alter the CII output

The spacial organization of the codons enters in the computation of the CII by the Ising-like term, which enforces a homogenizing interaction. In order to test how much this information is used, we have to compare the "true" results with those obtained on randomized sequences. A significant difference would imply that the spatial organization is a major feature of codon usage.

We therefore ran the algorithm on several realizations of 100 Codon Reshuffled sequences (as defined in section 1.1.2). Contrary to our expectations, as the plot in the left panel of Fig. 1.13 shows, there is no evident anomaly between the true results and those obtained from the reshuffled sequences. As a further check, we also computed the CII for the Amino acid Reshuffled sequences, obtaining a similar outcome (Fig. 1.13, right panel). These facts implies that the spacial organization of the codons does not affect the output of the CII algorithm, and is therefore not used in the

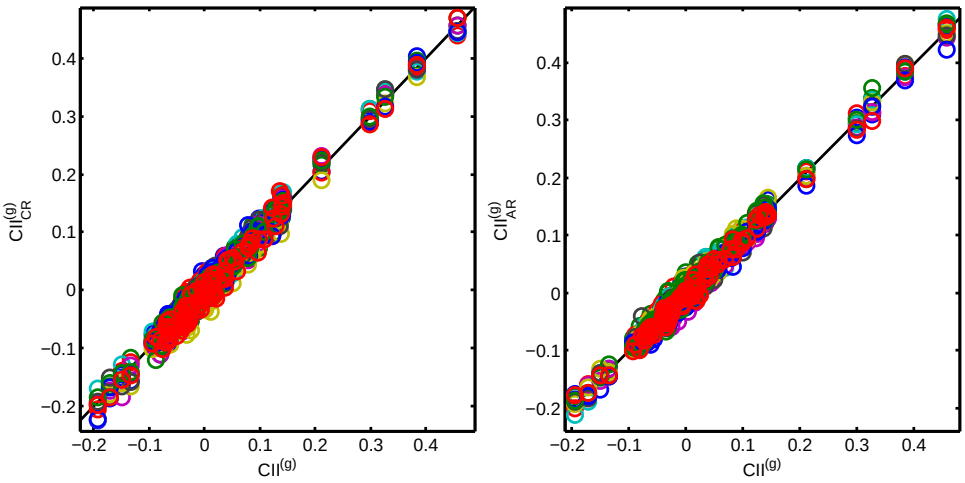

Figure 1.13: Codon information index computed for Codon Reshuffled (CR) sequences (left), and for Amino acids Reshuffled (AR) sequences (right), plotted versus the true sequences for 10 reshuffled replicas (plotted as different colors) of 100 different mRNAs. The reshuffled results are highly consistent with those of the true sequences, suggesting that the organization of the codons along the sequence, as well as that of the amino acids, do not play a relevant role in the CII computation. The black line is a guide for the eye and represent perfect proportionality.

computation of the CII. We finally computed the CII for a set of sequences whose codons were randomly reshuffled across the sequences. This Generalized Codon Reshuffle (GCR) does not preserve the codon frequencies of each mRNA. The results are plotted in Fig. 1.14, where no correlation is reported.

Furthermore, as Fig. 1.15 shows, the ramp at the beginning of the mRNAs is due to a bias in the codon composition and not to finite size or border effect: lower CII codons are chosen with a higher probability near the start of the mRNA. This is the only sequence-specific effect we were able to find. Its relatively small size, however does not affect the output of the algorithm.

In this section we showed that most of the information encoded in the codon usage is order-agnostic, as the codon frequencies are the only factor affecting the CII computation. This observation might seem in contrast with numerous evidences which support the existence of codon spatial organization [40, 123, 19, 75, 82, 23]. Codon usage is a complex phenomenon, as several factors (whose relative importance can be organism-dependent) can contribute to its determination. It is most probable, then, that the informa-

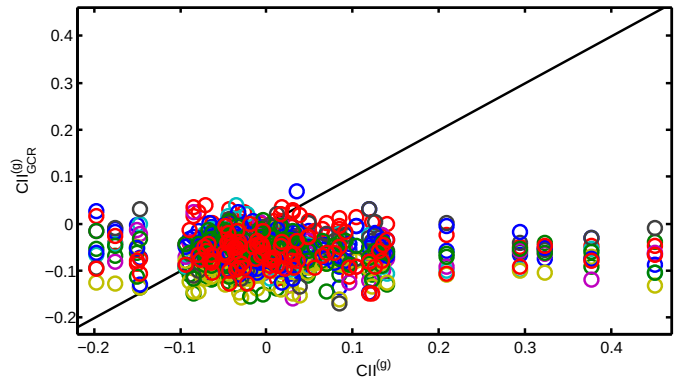

Figure 1.14: Codon information index computed for Generalized Codon Reshuffled (GCR) sequences plotted versus the true sequences for 10 reshuffled replicas (plotted as different colors) of 100 different mRNAs. There is no correlation between the two sets, showing that the codon frequencies inside each sequence are the determinant information in the CII calculation. The black line is a guide for the eye and represent perfect proportionality.

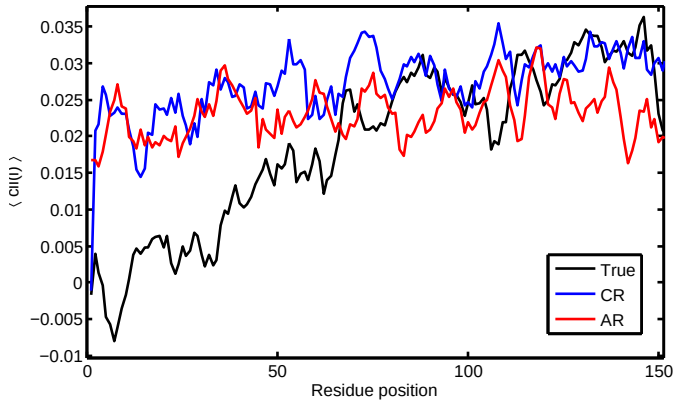

Figure 1.15: Average of the CII profiles across the true (black), Codon Reshuffled (CR, blue), and Amino acid Reseshuffled (AR, red) sequences. A ramp is present only for the true sequences, showing that the there is a codon compositional bias at the beginning of the ORFs. The profiles were smoothed with a moving average filter of size 7 for better readability.

tion contained in the codon bias is layered, with the frequencies representing the "first-order" approximation. This also implies that any exploration of the successive orders should first remove the biases coming from the frequencies effects, and the CII can be extremely helpful in this respect.

1.3.2 Mean-field theory for codon bias

In the previous section we showed that the CII is invariant under the rearranging of the codons along the sequences. Let us go back to reanalyze the symmetries of the two parts of the Hamiltonian (1.13): the information theoretical part 1.12 is invariant under CR and AR (and also GCR), while the 1D Ising term is not. Since the results of the CII computation are CR- and AR-invariant, it is reasonable to construct an Hamiltonian which explicitly implements these symmetries. As we show in the following sections, this new method greatly reduces the computational complexity of the CII.

This simplification arises due to the fact that I) the system can be treated in mean field, and II) the entropy can be written exactly. We are therefore able to compute the free energy of the system, whose large-N expansion has a particularly simple analytical expression. Furthermore, we observe that the new interaction term has the structure of a covariance matrix of codon usage across the genes.

Fully-connected model

The interaction should be invariant under CR and AR, i.e., under the exchange of two randomly chosen codons in a sequence. This invariance is verified if each codon inside the sequence g interacts with every other codon in that sequence. In the spin system lexicon, this is a fully connected Ising term, which produces the Hamiltonian

$$H = -I - \sum_g \frac{J}{L_g} \sum_{i,j\in g,\ j>i} \sigma_i\sigma_j - h\sum_g \sum_{i\in g} \sigma_i, \tag{1.25}$$

where I is the information theoretical part of the Hamiltonian, Eq. 1.12, the sums are intended on the spins belonging to the mRNA g whose length is L_g, and the J/L_g interaction parameter was introduced in order to have a well-behaved thermodynamic limit (as the number of terms in that sum scales as L_g^2). Let us recall that there are $n_c^{(g)}$ codons of type c in the sequence g (of length L_g), and that $n_s^{(g)}(c)$ among them are associated to a spin of sign s. Moreover, $n_c = \sum_g n_c^{(g)}$, $N_s(a) = \sum_{c\in a}\sum_g n_s^{(g)}(c)$ and $N_a = \sum_{c\in a} n_c$.

With some manipulation on H we have

$$\begin{aligned} H &= -I - \sum_g \frac{J}{2L_g}\left(\sum_{i,j\in g}\sigma_i\sigma_j - L_g\right) - h\sum_g\sum_{i\in g}\sigma_i \\ &= -I - \sum_g L_g m^{(g)}\left(\frac{J}{2}m^{(g)} + h\right) + \text{const.} \end{aligned} \tag{1.26}$$

where $m^{(g)} = \sum_{i\in g}\sigma_i/L_g$ is the magnetization of the sequence g.

Let us introduce codon magnetizations

$$m_c = \frac{n_+(c) - n_-(c)}{n_c} = \frac{2n_+(c) - n_c}{n_c}, \tag{1.27}$$

where we remind that $n_s(c) = \sum_g n_s^{(g)}(c)$ and use them in rewriting I:

$$\begin{aligned} I &= -\sum_a\sum_{s=\pm1}\left[N_s(a)\psi\big(N_s(a)+K\big) - \sum_{c=1}^{K_a} n_s(c)\psi\left(n_s(c)+1\right)\right] \\ &= -\sum_a\sum_{s=\pm1}\left[N_a\frac{1+sM_a}{2}\psi\left(N_a\frac{1+sM_a}{2}+K\right)\right. \\ &\qquad\left. - \sum_{c=1}^{K_a} n_c\frac{1+sm_c}{2}\psi\left(n_c\frac{1+sm_c}{2}+1\right)\right] \\ &= -\sum_a\left[G_K\left(\frac{N_a}{2},\frac{M_aN_a}{2}\right) - \sum_{c=1}^{K} G_1\left(\frac{n_c}{2},\frac{n_cm_c}{2}\right)\right], \end{aligned} \tag{1.28}$$

where $M_a = (\sum_{c\in a} n_c m_c)/N_a$ and $G_i(n,x) = g_i(n+x) + g_i(n-x)$, with $g_i(x) = x\,\psi(x+i)$.

Typically, the numbers n_c of codons c which appear in I are large (as each sequence is ≈ 450 codons long, $n_c \gtrsim 10^4$ for 1000 sequences). We therefore expect that the large-N expansion of I is a very good approximation. In the large N expansion, G_K reads

$$\begin{aligned} G_K\left(\frac{N}{2},\frac{mN}{2}\right) \sim \frac{N}{2}&\left[(1-m)\log_2(1-m) + (1+m)\log_2(1+m)\right] \\ &+ N\log_2\frac{N}{2} + 2K - 1 + O(n^{-1}), \end{aligned}$$

and by using this expression in I we have, for the amino acid a:

$$I(a) \sim -\frac{N_a}{2}\left[Q(M_a) - \sum_{c\in a}\phi_c\,Q(m_c)\right] + \text{const} + O(N_a^{-1}), \tag{1.29}$$

where we introduced the function $Q(x) = (1-x)\log_2(1-x)+(1+x)\log_2(1+x)$, and the frequencies $\phi_c = n_c/N_a$. Note that these frequencies are normalized such that $\sum_{c\in a}\phi_c = 1$. It is also extremely interesting to observe that the function $Q(x)$ can be written as

$$\frac{Q(x)}{2} = \frac{1-x}{2}\log_2\frac{1-x}{2} + \frac{1+x}{2}\log_2\frac{1+x}{2} + 1 = -H_{\rm b}(p) + 1,$$

where $H_{\rm b}(p)$ is the entropy of a binary distribution with probability $p = (1+x)/2$. The interpretation of $I(a)$ is therefore as follows: maximizing $I(a)$ requires to maximize the entropy for the whole amino acid, i.e., to reach $p_a = \frac{1+M_a}{2} \approx 1/2$, while minimizing the entropy for each codon, i.e., $p_c = \frac{1+m_c}{2} \approx \{1, 0\}$.

Fully connected term, homogeneous spin clusters approximation

Let us now consider the fully connected term

$$H_{\rm FC} = -\frac{J}{2}\sum_g L_g \left(m^{(g)}\right)^2. \tag{1.30}$$

The magnetization $m^{(g)}$ of gene g can be written as $m^{(g)} = \sum_c m_c^{(g)} n_c^{(g)}/L_g$. However, we expect that the magnetization $m_c^{(g)}$ would not fluctuate too much from gene to gene due to the homogenizing interaction I. Furthermore, from a physical point of view, the spin encodes for a hidden property of the sequence, and allowing it to fluctuate from gene to gene implies that the property is context-dependent.

We therefore impose that the same codon is equally magnetized independently from the gene where it appears, i.e., $m_c^{(g)} = m_c^{(g')} \equiv m_c$, $\forall g, g'$, and the magnetizations $m^{(g)}$ can be written as

$$m^{(g)} = \sum_c m_c f_c^{(g)}, \tag{1.31}$$

where $f_c^{(g)} = n_c^{(g)}/L_g$ is the empirical frequency of the codon c in the sequence g.

Using this approximation, we have

$$\begin{aligned} H_{\rm FC} &= -\frac{J}{2}\sum_g L_g \sum_{c_1,c_2} m_{c_1} m_{c_2} f_{c_1}^{(g)} f_{c_2}^{(g)} \\ &= -\frac{J}{2}\sum_{c_1 c_2} m_{c_1} m_{c_2} \sum_g L_g f_{c_1}^{(g)} f_{c_2}^{(g)} \\ &= -L_{\rm tot}\frac{J}{2}\left[\sum_{c_1 c_2} m_{c_1} m_{c_2}\Sigma_{c_1 c_2} + \sum_{c_1 c_2} m_{c_1} m_{c_2} f_{c_1} f_{c_2}\right], \end{aligned} \tag{1.32}$$

where $L_{\rm tot} = \sum_g L_g$ is the total length of all the sequences, $f_c = \langle f_c^{(g)} \rangle_{L_g} = n_c/L_{\rm tot}$ is the overall frequency of codon c across the sequences, $\Sigma_{c_1 c_2} = \langle f_{c_1}^{(g)} f_{c_2}^{(g)} \rangle_{L_g} - f_{c_1} f_{c_2}$ is the covariance matrix of the codon frequencies, and all the averages are intended with respect to the weight $L_g/L_{\rm tot}$, namely $\langle x^{(g)} \rangle_{L_g} = \sum_g x^{(g)} L_g / L_{\rm tot}$.

The interpretation of $H_{\rm FC}$ is simple: the two terms favor, respectively, the alignment of those codons whose usage is correlated, and those whose abundance is higher.

Entropy and free energy

In the previous sections we obtained mean-field expressions for the energy of the spin system. In order to fully characterize the thermodynamic behavior we also need to include the entropic contribution to the free energy, as a function of the magnetization.

Given a state of the system, its energy is a function of the counts $\{n_s^{(g)}(c),\ s = \pm,\ c = 1, \ldots, 61,\ g = 1, \ldots, G\}$ only. The number of inequivalent permutations of the spins leaving these numbers unaltered is $\Omega = \prod_g \prod_c \binom{n_c^{(g)}}{n_+^{(g)}(c)}$. The entropy $S = \log_2 \Omega$ is therefore given by

$$S = \log_2 \prod_g \prod_c \binom{n_c^{(g)}}{n_+^{(g)}(c)} = \log_2 \prod_g \prod_c \binom{n_c^{(g)}}{(n_c^{(g)} \frac{1+m_c^{(g)}}{2})}. \tag{1.33}$$

Let us now approximate the entropy S in the large $n_c^{(g)}$ limit:

$$\begin{aligned} S &\sim -\sum_c n_c \left(\frac{1+m_c}{2} \log_2 \frac{1+m_c}{2} + \frac{1-m_c}{2} \log_2 \frac{1-m_c}{2} \right) \\ &= \sum_c \frac{n_c}{2} Q(m_c), \end{aligned} \tag{1.34}$$

where the homogeneous spin cluster approximation was used. Note that this approximation is valid in the case of infinitely long sequences. Since the mRNAs have a finite length, some error is introduced. However, as we show in the following section, the results of the free energy minimization using Eq. (1.34) are compatible with those of the previously defined CII. Besides, this simplification is particularly convenient from the computational point of view (as it allows a dramatic reduction of the complexity of the computation of the free energy differences, from $O(G)$ to $O(1)$).

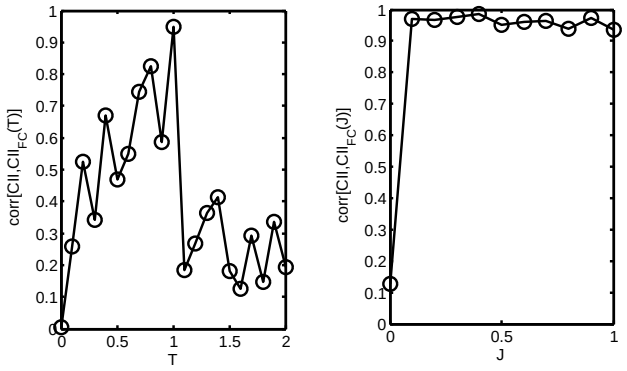

Figure 1.16: Correlation between the results from the CII as computed in Sec. 1.2 and the fully connected version. In the left panel we set $J = 0.3$ and performed a sweep in temperature, showing that the "sweet spot" lies close to $T = 1$. As the right panel shows, at $T = 1$ the results do not depend strongly on the choice of J.

The free energy $F = H - TS$ can therefore be written as

$$F = \sum_a \frac{N_a}{2} \left[Q\left(\sum_{c \in a} m_c \phi_c \right) - (1-T) \sum_{c \in a} \phi_c Q(m_c) \right]$$
$$- L_{\text{tot}} \frac{J}{2} \left[\sum_{c_1 c_2} m_{c_1} m_{c_2} \Sigma_{c_1 c_2} + \sum_{c_1 c_2} m_{c_1} m_{c_2} f_{c_1} f_{c_2} \right] - L_{\text{tot}} h \sum_c m_c f_c. \quad (1.35)$$

Let us observe that both I and H_{FC} have a critical point, respectively at $T = 1$ and $T = J$. We are not interested in fully characterizing the phase diagram of this model here, we only remind that the fully connected model can be analytically solved, as shown in App. A.2).

Comparison with the CII

In order to compare the results of this new method with the previously introduced CII, we need to fix the choice of the parameters J and T. As the thermodynamics of this new system is different from the previous one, a priori the optimal choice of the parameter might be different. For this reason we can also use the temperature as a further tuning parameter.

We therefore set up a Monte Carlo algorithm in order to compute the average magnetization of each codon $\langle m_c \rangle$ by minimizing the free energy (1.35). The fully connected codon information index $CII^{(g)}_{\text{FC}}$ is computed as

$$CII^{(g)}_{\text{FC}} = \frac{\sum_c \langle m_c \rangle n_c^{(g)}}{L^{(g)}}. \quad (1.36)$$

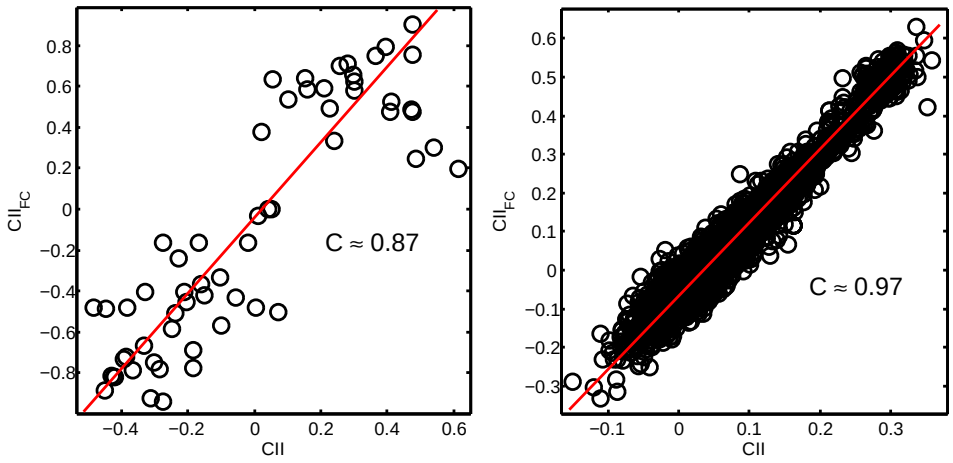

Figure 1.17: Comparison between the CII as computed in Sec. 1.2, and the fully connected version as in the present section. In the left panel we plot the codon-wise values of the magnetization, on the right panel we plot the average $CII^{(g)}$ value of the mRNAs belonging to the transcriptome of Yeast. The two versions produce highly correlated results.

For consistency reasons, we set the J parameter at the same value as in the previous case ($J = 0.3$). Let us note that, however, the choice of the parameter J is not critical, as shown in the right panel of Fig. 1.16. We ran an initial set of simulations in order to tune the temperature T and we found that optimal correlation is obtained for $T = 1$.

The comparison with the previous definition of the CII is plotted in Fig. 1.17, where it is shown that the two methods produce highly compatible results.

On the strictly computational side, we emphasize that the fully connected version runs in few minutes, with a speedup of at least 3 orders of magnitude compared to the previous version.

As a final point let us give a qualitative argument for why the two methods give the same results. The CII method in Sec. 1.15 utilizes a nearest neighbor interaction term, which captures *on average* the correlations between the codon usage frequencies. In fact, supposing that codons $c_1 \in a_1$ and $c_2 \in a_2$ are over-represented in a set of genes, the number of interactions $\sigma_{c_1}\sigma_{c_2}$ will be, on average, higher than all the other possible interactions. Due to this fact the spins of the two codons tend to be aligned, enforcing the correlations between those codons which are used in a similar way.

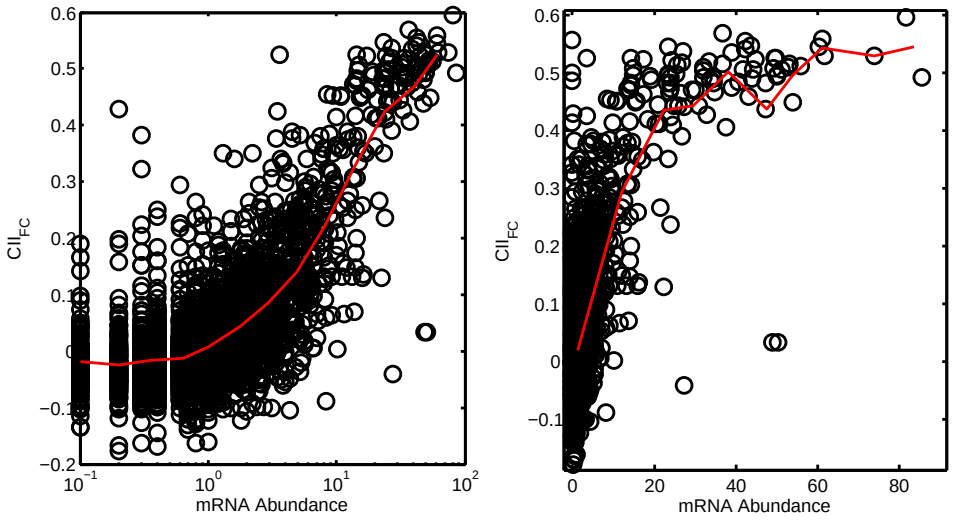

Figure 1.18: Fully connected CII, as computed in Sec. 1.3.2, versus mRNA abundance, logarithmic (left) and linear scale (right). The red line is a moving average. The CII has a plateau for very small and very large mRNA abundances.

1.4 Conclusions

In the previous sections we analyzed the sequences of mRNA of Yeast in order to characterize in the most abstract way how the different codons are used. Specifically, we showed that the codon bias is not homogeneous across the transcriptome (in Sec. 1.1.1), and that some information is stored in the codon order (in Sec. 1.1.2). We also introduced the CII as a mean to measure the local information content of a sequence. Strikingly, the CII turned out to be invariant under the reshuffling of the codons, which implies that the relevant information used in its computation is given by the codon usage frequencies. We finally used this fact to define an approximated, fully connected algorithm which explicitly depends on the correlations between the codon usages across the genes. This new method produces compatible results at a 3 orders of magnitude lower computational cost.

From these observations we can conclude that, at the leading order, the codon bias is mostly a *codon frequency* bias, and that the spacial organization enters at sub-leading orders.

This approach, however, does not answer why the codon bias exists, nor it explains the behavior of the plot of the fully connected CII versus mRNA abundance in Fig. 1.18: at small mRNA abundances the CII has a plateau, while it grows at larger abundances and seems to top out. Qualitatively,

this behavior is expected if an optimal codon bias exists and the related evolutionary pressure is small and proportional to the abundance. In fact, let us suppose that an optimal codon bias C^* exists. If the abundance is too small, selection is negligible and random synonymous mutations determine the codon usage (evidenced by the plateau in the left panel of Fig. 1.18). On the other hand, if the abundance is very large the codon usage is very close to C^*, as shown in the right panel of Fig. 1.18. The evolutionary origin of the anisotropy of the codon bias seems very likely and received wide support in the literature [18, 59, 69]. However, the underlying mechanism driving the evolution has not been clearly and univocally identified yet.

Two grand question emerge clearly from this analysis: I) which evolutionary forces shape and produce the bias in the usage frequencies of the codons? And II) does the spatial organization of the codons play a role in optimizing and/or regulating translation? Answering these questions requires a deeper analysis of the whole translation machinery. In the following chapter we will therefore analyze how the interplay between the resources used in translation (namely, mRNAs, ribosomes, and tRNAs) can affect the process.

Chapter 2

Explaining the data: models for the systemic translation of the proteins

The two most important players in the synthesis of the proteins, beyond the mRNAs, are the tRNAs and the ribosomes. Their "economics" (i.e., the relative abundance in the cell) affects the speed, the accuracy, and the regulation of the reaction, as we will show later.

The ribosomes are the molecular factory where the synthesis of the protein occurs: they bind to the mRNA and sequentially scan through its codons. At each codon, the translation event occurs when the correct tRNA diffuses in the active site of the ribosome and the new amino acid is attached to the nascent polypeptidic chain.

The tRNA molecules are at least equally fundamental. They carry the amino acids to the ribosome and effectively decode the genetic code, by recognizing the codons. Moreover, the rate at which the new amino acids are incorporated in the nascent protein depends on the abundances of the tRNAs, as tRNA waiting time is typically the rate-limiting step in translation [79].

After translation has occurred, the tRNA leaves the ribosome and must be recharged with the corresponding amino acid by the enzyme aminoacyl-syntethase[1] before it can be used again. This finite recharge rate and its interplay with the consumption by the ribosomes affects the quantity of tRNA available for translation. Due to this phenomenon, the fraction of charged tRNA can range, depending on the growth condition of the organism, from

[1]The recharge process is more complicated, as it also involves a species-dependent elongation factor, and ATP [1].

≈ 1 to $\approx 10^{-3}$, as measured experimentally in Ref. [33]. Furthermore, several among the 20 amino acids are translated by more than one codon and one tRNA, and the codon bias determines the rate at which the tRNAs encoding for the same amino acids are used.

In order to understand the implications of these facts, we develop here a series of simple, mean field models. First, we analyze the effect of a finite number of ribosomes. Then, using an extension of the same model, we observe that the optimal codon bias is a function of the environment (here modelled as the tRNA recharge rate): if the environment is rich enough, the optimum is obtained by using the most abundant tRNA only, while in starvation conditions the different species of tRNA should be used accordingly to their abundances. We finally abandon the mean field approximation in order to analyze the effects of the small concentration of tRNA on the distribution of the time intervals between subsequent translation events.

2.1 Ribosome load and translation optimality

Ribosomes are a limited resource in translation, as their number strongly affects the growth rate of the unicellular organisms [101]. It is therefore of great interest to study how this limitation overall affects the translation.

In the model introduced in the following sections we assume that the ribosomes do not interact. We neglect the possible occlusions of the binding site, or traffic jams along the sequences.[2] This effects become less important the lower the ribosome density along the sequences is, i.e., when the translation initiation rate is smaller than the elongation timescale.

2.1.1 Dynamics of the ribosome occupation

Let us consider a cell having R_T ribosomes. At a certain time, $R^{(g)}$ among them are translating the mRNA g. The rate at which a ribosome initiates translation for that mRNA is $\eta^{(g)} = n^{(g)}\chi^{(g)}$, where $n^{(g)}$ is the number of mRNA copies,[3] and $\chi^{(g)}$ is the rate of translation initiation per mRNA for gene g. Most of the variation in $\eta^{(g)}$ is due to $n^{(g)}$, whose dynamic range spans 3 orders of magnitude. The translation initiation rates $\chi^{(g)}$ were

[2] The ribosomes have a finite size and the "excluded volume" effects can be relevant. For instance, a new ribosome can bind to a mRNA only when the previous one has moved on along the sequence.

[3] Gene expression is itself a stochastic process, with bursts of transcription followed by a slow decay of the mRNAs [39]. In principle, $n^{(g)}$ fluctuates around an average value, which is the one we use here.

estimated in Ref. [23], showing a less than 2 orders of magnitude dynamic range.

Let us also suppose that with rate $1/T^{(g)}$ a ribosome terminates the translation. $T^{(g)}$ is the average time from translation initiation to ribosome release for gene g.

At each time, the pool of free ribosomes is $R_T - \sum_g R^{(g)}$, and each of them can initiate the translation on gene g. The mean field equations for this system are therefore

$$\dot{R}^{(g)} = \eta^{(g)} \left(R_T - \sum_{g'} R^{(g')} \right) - \frac{R^{(g)}}{T^{(g)}}, \tag{2.1}$$

whose stationary state is given by

$$R^{(g)} = \frac{\eta^{(g)} T^{(g)}}{1 + \sum_{g'} \eta^{(g')} T^{(g')}} R_T. \tag{2.2}$$

As expected, the number of ribosomes translating gene g is proportional to the initiation rate $\eta^{(g)}$ (which includes a dependence on the number $n^{(g)}$ of mRNAs g) and to the total translation time $T^{(g)}$.

The consequences of this formula, however, are far from being trivial. Let us first consider the rate $K^{(g)}$ of protein production for gene g, i.e., the number of proteins produced in the unit of time:

$$K^{(g)} = \frac{R^{(g)}}{T^{(g)}} = \frac{\eta^{(g)}}{1 + \sum_{g'} \eta^{(g')} T^{(g')}} R_T. \tag{2.3}$$

Interestingly, the protein production rate is proportional to $\eta^{(g)}$ and weakly depends on $T^{(g)}$, since it is mediated at the denominator. This model predicts that I) the production rate of a protein is very little influenced by its translation rate, II) the ratios between production rates of different proteins is exclusively determined by the ratios in translation initiation rates $\eta^{(g)}$, whose most important determinant is the mRNA level.

Eq. (2.3) has another important consequence on the systemic translation process. Let us consider the growth rate of the organism. Its inverse, the doubling time, is at least as great as the time it takes to double all the proteins in the cell. Supposing that the physiological (target) protein number is $N^{(g)}$ and that in fast growth conditions the degradation of the protein is negligible, the doubling time $D^{(g)}$ of gene g is given by $D^{(g)} = N^{(g)}/K^{(g)}$. Since all the proteins are translated simultaneously, the overall doubling

time D is limited by the slowest gene:

$$D \geq \max_g D^{(g)} = \max_g \frac{N^{(g)}}{K^{(g)}} = \frac{1}{R_T} \left(\max_g \frac{N^{(g)}}{\eta^{(g)}} \right) \left(1 + \sum_{g'} \eta^{(g')} T^{(g')} \right). \tag{2.4}$$

We observe that, since D is limited by the maximum of $N^{(g)}/\eta^{(g)}$, the optimal equilibrium is $\eta^{(g)} = cN^{(g)}$, $\forall g$. The optimum is reached by finely tuning the initiation rate $\eta^{(g)}$ by varying the mRNA abundances $n^{(g)}$ and the initiation rate per mRNA $\chi^{(g)}$ (the latter is influenced by a large extent by the folding of the mRNA, as shown in Ref. [75]).

Furthermore, since the doubling time is a common measure of the fitness of the organism (the lower D is, the higher the fitness), we expect that evolution pushes D towards a minimum. This condition is achieved by systemically increasing the number R_T of ribosomes, the number $n^{(g)}$ of mRNAs, and the per-mRNA initiation rates $\chi^{(g)}$ (the increase of R_T and $n^{(g)}$, however, is limited by the finite amount of resources in the cell). The term $\sum_{g'} \eta^{(g')} T^{(g')}$, moreover, suggests the global (or systemic) nature of the translation optimization problem: by increasing the speed of translation, more free ribosomes are available and the growth rate is overall increased.

The last observation has a further implication: the relative effect of a change in the translation time $T^{(g)}$ of mRNA g is proportional to $\eta^{(g)}$, and in turns to the abundance of mRNA. Importantly, therefore, the evolutionary pressure acting on a gene is proportional to the abundance of its mRNA, qualitatively explaining the set of plots Figs. 1.18, 1.9. Indeed, a population genetics calculation [69] showed that these kind of behaviors are well explained in a selection-mutation-drift framework.

Following the independent ribosomes approximation, it is reasonable to expect that the translation time of the codons in a sequence does not depend on their order and is determined by the type of the codon only. Furthermore, if the mRNA is long enough, most of the time will be spent in elongation. Let us therefore express the time of translation as $T^{(g)} = \sum_c n_c^{(g)} t_c$ where $n_c^{(g)}$ is the number of times the codon c is present in mRNA g and t_c is the average translation time for that codon. We can thus rewrite the second term of doubling time as

$$1 + \sum_g \eta^{(g)} T^{(g)} = 1 + \sum_c t_c \sum_g \eta^{(g)} n_c^{(g)}. \tag{2.5}$$

The optimization of the codon translation times is realized when the codons which are *effectively* most utilized (i.e., after weighting the codon usage of

gene g with the effective number of ribosomes translating it) are translated faster.

The picture which emerges from these considerations is that translation is a systemic process: optimizing the translation of a few, very expressed genes might affect the overall growth rate of the organism, and the reverse is also true. If some poorly optimized genes are present in the cell (as for instance an exogenous plasmid) the translation machinery can be globally hampered.

2.2 Quantitative model for tRNA dynamics at finite recharging

An accurate description of the translation process cannot neglect the tRNA dynamics, and in particular the fact that the tRNA molecules must be loaded with the corresponding amino acid in order to be used in translation.

The codon usage has a direct effect on the rates at which the different tRNAs are used, as synonymous codons are often translated by different tRNAs. As an example, let us consider two synonymous codons translated by two different species of tRNA. The consumption rate of the tRNAs is determined by the frequencies of the corresponding codons along the sequence (i.e., by the codon bias). If one of the two tRNA is recharged more slowly than the characteristic time of its usage, we expect that its charged fraction will be depleted and the translation globally hampered.

The recharge rate of the tRNAs is not a constant, as shown in Ref. [33]: the environmental conditions where the organism is growing has a large influence on the charged fraction of tRNAs, which can range from less than 1% to almost 100% depending on the richness of the media. The codon usage can help in coping with these variability.

In the following we first ask what is the optimal codon bias as a function of the recharging rate. However, as this rate depends on the random environmental conditions, we observe that the codon bias can be interpreted as an optimal strategy.

2.2.1 The model

Let us consider a cell containing R ribosomes translating simultaneously the same mRNA. The mRNA is composed using K different codons, which are used with frequencies $f_1, \ldots, f_K$ in the sequence. Let us also suppose that

each of those codons is decoded by one and only one kind of tRNA[4]. The tRNAs are expressed at $N_1, \ldots, N_K$ copies in the cell, and $n_1, \ldots, n_K$ among them are charged with the corresponding amino acid.

Furthermore, each uncharged tRNA molecule of kind i is recharged at rate $\lambda_R(i)$ (and $n_i \to n_i + 1$), while it is used in translation at rate $\lambda_T(i)$ (and $n_i \to n_i - 1$).[5] In order to simplify the calculation, in the following we will suppose that these rates are uniform across all the tRNAs and equal to λ_R and λ_T, respectively. Moreover, without loss of generality, we set $\lambda_T = 1$ as the time scale of the process and define $\lambda = \lambda_R/\lambda_T$.

At each time, R_i ribosomes are waiting for a charged tRNA of type i. The dynamics of the R_is is determined by the frequencies f_i of the codons along the mRNA, supposing that the codons do not have any spacial organization. For all the possible i and j, the rates can be written as

$$\begin{aligned} &w[(R_i, R_j; n_i) \to (R_i - 1, R_j + 1; n_i - 1)] = f_j \frac{n_i}{R} R_i, \text{ for } i \neq j, \\ &w[(R_i; n_i) \to (R_i; n_i - 1)] = f_i \frac{n_i}{R} R_i, \\ &w[n_i \to n_i + 1] = \lambda(N_i - n_i), \end{aligned} \tag{2.6}$$

where we measure the volume in units of R.

In a typical Yeast cell there are at least $\approx 10^3$ tRNA molecules of each kind and $\approx 10^4$ ribosomes. Since these numbers are large, we approximate the master equation associated to this process with the mean-field equations for the rescaled quantities $r_i \equiv R_i/R$ (fraction of ribosomes translating codon i) and $X_i \equiv n_i/N_i$ (fraction of charged tRNAs of kind i). Neglecting the spacial inhomogeneity of the tRNAs, we have

$$\begin{aligned} \dot{r}_i &= f_i \sum_{j, j \neq i} r_j \nu_j X_j - (1 - f_i) r_i \nu_i X_i \\ \dot{X}_i &= \lambda(1 - X_i) - X_i r_i, \end{aligned} \tag{2.7}$$

where $\nu_i = N_i/R$. The stationary solution of this equations can be easily

[4]This approximation is generally very accurate and has been shown to lead to good results [89].

[5]Supposing that each tRNA molecule can be used in translation with a certain rate implies that the translation rate for codon i is proportional to the abundance of the tRNA. Even thou translation is composed by many different events, it was shown [79] that the tRNA selection time (i.e., the time spent by ribosome waiting for the right to tRNA diffuse in its active site) is the rate limiting step of the process.

obtained by setting $Q = \sum_j r_j \nu_j X_j$:

$$\begin{aligned} r_i &= \frac{f_i Q \lambda}{\lambda \nu_i - f_i Q} \\ X_i &= \frac{\nu_i \lambda - f_i Q}{\nu_i \lambda}, \end{aligned} \tag{2.8}$$

where Q is fixed by solving the equation $\sum \frac{f_i}{\lambda \nu_i - f_i Q} = \frac{1}{\lambda Q}$ obtained by imposing $\sum r_i = 1$.

Optimal codon usage, 2 codons and 2 tRNAs case

The optimal codon bias for each amino acid as a function of λ can be obtained from the translation time of each codon, which is $t_i \propto (\nu_i X_i)^{-1}$.

Let us consider the simplified case of a mRNA encoding for a protein composed by several repeats of the same amino acid, which is translated by two codons each associated to one tRNA. The average translation time for that amino acid is therefore

$$T_a = f_1 t_1 + f_2 t_2 \propto \frac{f_1}{\nu_1 X_1} + \frac{f_2}{\nu_2 X_2}, \tag{2.9}$$

with $f_1 + f_2 = 1$.

The parameters X_i can be obtained by solving the set of equations (2.8) for X_i. We therefore have

$$\begin{aligned} X_1 &= \frac{\nu_1 \lambda - f_1 Q}{\nu_1 \lambda}, \\ X_2 &= \frac{\nu_2 \lambda - (1 - f_1) Q}{\nu_2 \lambda}, \end{aligned} \tag{2.10}$$

with Q given by

$$\frac{f_1}{\lambda \nu_1 - f_1 Q} + \frac{1 - f_1}{\lambda \nu_2 - (1 - f_1) Q} = \frac{1}{\lambda Q}, \tag{2.11}$$

which has the solution

$$Q = \nu_2 \lambda \frac{(1+\lambda)\left(\mu + \frac{f}{1-f}\right) - \sqrt{(1+\lambda)^2 \left(\mu + \frac{f}{1-f}\right)^2 - 4\frac{f}{1-f}(1+2\lambda)\mu}}{2f(1+2\lambda)}, \tag{2.12}$$

where $\mu = \nu_1/\nu_2$ and we dropped the index from the frequency.

In order to calculate the optimal codon bias, we need to take the derivative of the time T_a, Eq. (2.9), with respect to the frequency f. Let us observe that, using Eq. (2.11), the expression for T_a becomes

$$T_a = \frac{1}{Q}, \tag{2.13}$$

which implies that we need to find the maxima of Q. Setting the derivative to zero, we obtain the optimal frequency:

$$f^* = \begin{cases} 0 & \text{if } \mu < 1 \wedge \lambda > \frac{\sqrt{\mu}}{1-\sqrt{\mu}} \\ 1 & \text{if } \mu > 1 \wedge \lambda > \frac{1}{\sqrt{\mu}-1} \\ \frac{(1+\lambda)^2\mu^2+(1+(2-\lambda)\lambda)\mu+\lambda(1+\lambda)(\mu-1)\sqrt{\mu}}{(1+\lambda)^2(\mu^2+1)+2(1+(2-\lambda)\lambda)\mu} & \text{otherwise.} \end{cases} \quad (2.14)$$

Let us consider the two important limits $\lambda \to 0$ and $\lambda \to \infty$, respectively modeling starvation and rich growth conditions. In the $\lambda \to 0$ case

$$f_s^* = \frac{\mu}{1+\mu} = \frac{\nu_1}{\nu_1+\nu_2}, \quad (2.15)$$

expressing a proportionality rule between codon usage and relative tRNA abundance.

On the other hand, in the $\lambda \to \infty$ case,

$$f_r^* \sim \begin{cases} 0 & \text{if } \mu < 1 \\ 1 & \text{if } \mu > 1. \end{cases} \quad (2.16)$$

In this conditions only the codon coupling to the most abundant tRNA is used. We will refer to this behavior as the single-tRNA rule.

In the case of more than two codons we expect that these general rules will be still verified. However, for finite λ the expression for the optimal codon bias might be complicate.

If the codon usage of the living organisms is optimized for translation speed at the tRNA usage level, we should expect to find the true frequency f between $f_s^* < f_w < f_r^*$. A priori, however, λ is a random variable which depends on the environment (λ is large if the environment is rich and vice versa), while the codon usage is fixed in the genome. Therefore, the codon usage can be thought of as an optimal *strategy* responding to the random environment, here modeled with the variation of λ.

Codon bias as an optimal strategy

Let us proceed using the previous 2 codons, 2 tRNA case. Suppose now that the variable λ switches randomly between two values $\{\lambda_s, \lambda_r\}$, such that $\lambda_s \ll 1$ and $\lambda_r \gg 1$ respectively represent starvation and rich growth conditions, and that λ, on average, spends a fraction τ of the time in starvation conditions.

The translation speed averaged on the variation of λ is therefore

$$v = \frac{\tau}{T_a(s)} + \frac{1-\tau}{T_a(r)}. \quad (2.17)$$

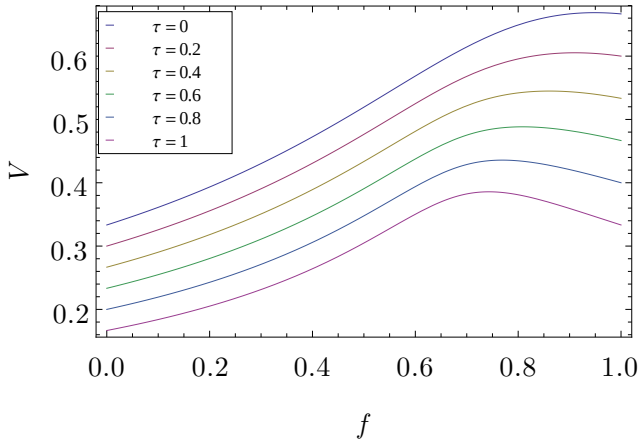

Figure 2.1: Speed of translation Eq. (2.17) as a function of the frequency of the codon coupling to the most abundant tRNA, for several values of τ. In this example $\lambda_1 = 0.5$, $\lambda_2 = 2$, $\nu_1 = 1$, $\nu_2 = 0.5$. The maximum of the curves -i.e., the optimal codon bias- interpolates between the values at $\tau = 0$ and $\tau = 1$.

Using the approximations for small and large λ, Eqs. (2.15) and (2.16), and supposing that $\nu_1 > \nu_2$, we have

$$v \sim \frac{\nu_1 \lambda}{f}\tau + \frac{\nu_1 \nu_2}{\nu_1(1-f) + f\nu_2}(1-\tau). \tag{2.18}$$

In this simple case, the codon bias which maximizes v is $f^* = \mu/(1+\mu)$ if $\frac{\mu-1}{\mu-1+2\lambda} < \tau$ and $f^* = 1$ otherwise. In less extreme conditions (i.e., for finite λ_s and λ_r) f^* interpolates between the optimal codon biases at the limiting conditions, as shown in Fig. 2.1.

Interestingly, the codon usage can carry the hallmarks of the environment where the organism spent its evolutionary history. If the harmonization with the tRNA pool plays a predominant role in shaping the codon usage, we expect to find some characteristic behavior in the codon usage.

We therefore analyzed the mRNA sequences in order to extract the codon frequencies of the amino acid with more than one tRNA in the set of the 130 most expressed genes in Yeast[6]. First, we grouped the codons according to which tRNA is used to translate them by using the "parsimony rule" as in [89], i.e., in the minimal way such that each codon is translated by one and only one tRNA. For each amino acid a we computed the frequencies of tRNA usage as $f_t(a) = \frac{\sum_{c\in t} n_c(a)}{\sum_{c\in a} n_c(a)}$, where the sum at the numerator is intended

[6]Since the selective pressure is proportional to the expression of the gene, we expect that if some effect is present it will be stronger on the highly expressed genes.

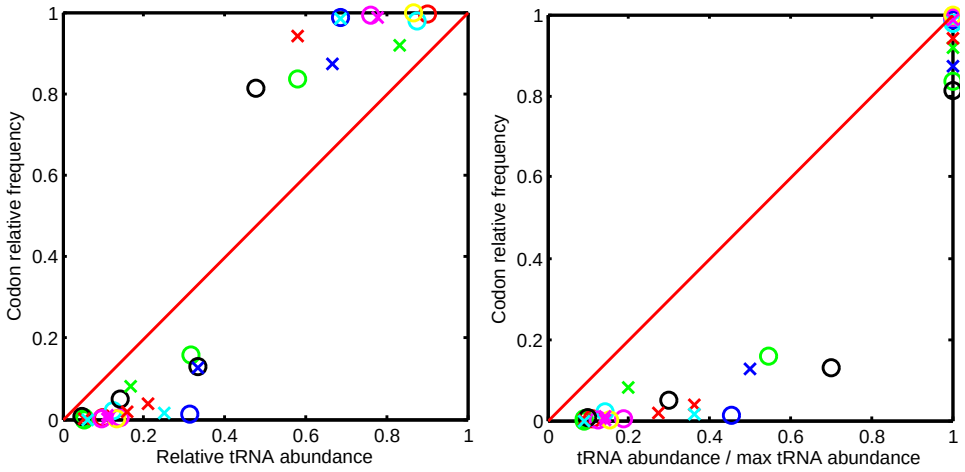

Figure 2.2: Plot of $f_t(a)$ versus $\mu_t(a)$ (left panel) and $f_t(a)$ versus $\kappa_t(a)$ (right panel) for the 130 most expressed genes in Yeast. Different symbols and colors correspond to different amino acids. The red line indicate direct proportionality. As the right panel shows, the most abundant tRNA is used at least 80% of the times, indicating that the actual codon usage is close to the one-tRNA only case for this set of genes.

on the codons translated by the tRNA t. Furthermore, we obtained the numbers ν_t of the genes of each tRNA as a proxy for the tRNA abundances.[7] From these numbers we computed, for each amino acid, the ratios $\mu_t(a) = \frac{\nu_t}{\sum_{t\in a}\nu_t}$, and $\kappa_t(a) = \frac{\nu_t}{\max_{t\in a}\nu_t}$. These two quantities are best suited to identify the proportionality and the one-tRNA rules, respectively.

In Fig. 2.2 we plot $f_t(a)$ versus $\mu_t(a)$ (left panel) and $f_t(a)$ versus $\kappa_t(a)$ (right panel) for all the amino acids translated by more than one tRNA. The most abundant tRNA is used at least 80% of the times, as the right panel shows. This observation seems to indicate that the evolutionary pressure pushes towards the "one-tRNA" case, i.e., towards the use of the most abundant tRNA.

In the light of the previous game-theoretical example, it seems that this strain of Yeast optimized its translation machinery in order to cope with rich environments. This codon choice could be supported also by another evolutionary game theoretical argument: when different strains (i.e., agents) are competing for the same limited resource, the evolution is played as an

[7] This is a well established procedure based on the experimental observation that the correlation between tRNA genes copy number and tRNA abundance in the cell is very high [89]. The data on tRNA genes copy numbers are easily available on the database [20].

adversarial game, and the ability to grow quickly in rich conditions is an effective strategy to overgrow the competitors by more effectively consuming the resources.

Let us conclude by observing that, most generally, the study of the living systems cannot ignore the environmental variability experienced by an organism during its evolutionary history. This approach can provide a very profound hindsight on the present state of the organisms.

2.3 Statistics of the codon translation time

Protein synthesis is one of the most common biochemical reactions happening in the cell and, despite this process is biologically and chemically well understood, the implications of its intrinsic stochastic nature (caused by the small number of particles involved) have not been fully elucidated yet.

An intriguing question concerns the distribution of the time intervals between two translation events (the codon translation time distribution, CTTD). This distribution, in fact, heavily influences the dynamics of the ribosomes along the sequence [82, 97, 50, 23] and can affect the efficiency, accuracy and regulation of the translation process [91, 47], as well as the process of cotranslational folding of the nascent protein [85]. This distribution was measured *in vitro* [115], however with a sub-optimal resolution. In addition, numerical simulations [122] have shown that the CTTD can significantly deviate from an exponential. The minimal mechanism which produces these deviations, however, is unclear and an analytical understanding of this phenomenon is lacking.

A key ingredient in translation is the tRNA and, as we showed in Sec. 2.2, these molecules have a non trivial recharge dynamics. Importantly, since they are present at low concentrations in the cell [35], local fluctuations in their number can have an important effect on the translation dynamics.

A further interesting observation was made in Ref. [19]: the spatial organization of the codons is toward the reuse of the same tRNA. Specifically, when the same amino acid appears on the mRNA sequence at a short distance, it tends to be encoded by codons translated by the same kind of tRNA. Describing the tRNA dynamics in the neighbor of the ribosome can therefore help in understanding this feature of the sequences.

In order to understand quantitatively and analytically how the previous facts affect the translation and the CTTD, in the next sections we introduce a stochastic model which explicitly incorporates (*i*) tRNA charging and discharging dynamics, and (*ii*) spatial inhomogeneity and stochastic fluctuations in the number of charged tRNAs around the ribosome. This minimal

model captures these two fundamental aspects of the translation process [8], and is analytically tractable. Its solution, validated using Monte Carlo numerical simulations, shows that the interplay between diffusion, recharging and translation dynamics induces a coupling between the fluctuations in the number of charged and uncharged tRNAs. Due to this phenomenon the CTTD, which we obtain analytically from the model, deviates from a pure exponential. Besides, this model reaches asymptotically a non equilibrium steady state (NESS). NESSs have attracted a lot of attention since a variety of systems in physics, chemistry, biology and engineering exhibit them, and their characterization is typically far more difficult than the equilibrium states [126, 127, 90, 22].

2.3.1 The model

We model here a ribosome translating a mRNA into a protein. Each translation event occurs when a tRNA, charged with the proper amino acid, interacts with the ribosome. We suppose that the ribosome recruits the tRNAs within an effective radius r (see Sec. 2.3.4), and that the tRNAs farther than r are considered as a part of an infinite reservoir. Moreover, the reservoir and the system can exchange tRNAs, due to diffusion.

We treat the special case of a single tRNA species translating a single type of codons. This assumption allows a very detailed analysis of the system. We anticipate that the understanding obtained in this simplified case can be helpful in describing qualitatively the general case of many types of codons and tRNAs, due to the fact that the underlying mechanisms are the same.

Let us therefore consider a system which comprises a ribosome (translating an mRNA composed by several repeats of the same codon), n charged and m uncharged tRNAs (see Fig. 2.3). Each uncharged tRNA can be recharged with rate λ_R, while each charged tRNA can be chosen for translation with rate λ_T, becoming uncharged [9]. We also suppose that there is a stochastic flux between the system and an infinite reservoir, i.e., that each tRNA (either charged of uncharged) can exit the system with rate ρ, while with rate μ ($\tilde{\mu}$) a charged (uncharged) tRNA diffuses from the reservoir into the system.

[8] We did not consider for instance the enzymatic nature of the recharging of the tRNAs, the continuous spatial dependency of the tRNA density, nor tRNA proofreading.

[9] The translation time is mostly determined by the codon selection time (i.e., the time a ribosome has to wait until the right aminoacylated tRNA diffuses in its active site [79]), which is proportional to the fraction of charged tRNA.

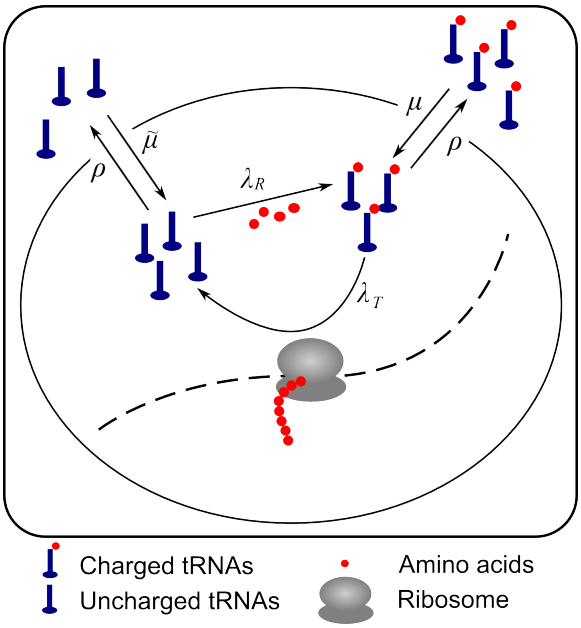

Figure 2.3: Cartoon of the model illustrating the possible reaction pathways. Uncharged tRNAs (blue, on the left) can either be exchanged with the reservoir or be recharged (with rate λ_R), illustrated by the addition of an amino acid (red dots). Similarly, the charged tRNAs can enter in or leave the system, or be used in translation (with rate λ_T) by the ribosome (gray), which is translating a mRNA (dashed line). In the state represented here, $n = 3$ and $m = 4$.

Considering an infinitesimal time step δt, the possible single-step transitions (with the corresponding probabilities) are:

- $(n,m) \overset{m\lambda_R\,\delta t}{\longrightarrow} (n+1, m-1)$: recharge, one uncharged tRNA gets charged.
- $(n,m) \overset{n\lambda_T\,\delta t}{\longrightarrow} (n-1, m+1)$: translation, one charged tRNA gets discharged and one codon is translated.
- $(n,m) \overset{\mu\,\delta t}{\longrightarrow} (n+1, m)$ and $(n,m) \overset{\tilde{\mu}\,\delta t}{\longrightarrow} (n, m+1)$: a tRNA (respectively charged, uncharged) enters the system from the reservoir.
- $(n,m) \overset{n\rho\,\delta t}{\longrightarrow} (n-1, m)$ and $(n,m) \overset{m\rho\,\delta t}{\longrightarrow} (n, m-1)$: a tRNA (respectively charged, uncharged) leaves the system.

These rates define, in general, a non-equilibrium system: the stationary state is a function of all the rates, as we show in the next section.

2.3.2 Stationary distribution of the number of charged tRNAs

The set of rates given in the previous section produces the following master equation for the probability $p_{n,m}(t)$ of being in the state (n,m):

$$\begin{aligned}\dot{p}_{n,m} = \lambda_R\big[-m\,p_{n,m} + (m+1)p_{n-1,m+1}\big] + \lambda_T\big[-n\,p_{n,m} + (n+1)p_{n+1,m-1}\big] \\ + \rho\big[(n+1)p_{n+1,m} + (m+1)p_{n,m+1} - (n+m)p_{n,m}\big] \\ - (\mu+\tilde{\mu})p_{n,m} + \mu\,p_{n-1,m} + \tilde{\mu}\,p_{n,m-1}. \quad (2.19)\end{aligned}$$

We focus on the stationary state of the system by setting $\dot{p}_{n,m} = 0$. Since the system is ergodic, the stationary state is unique and it is reached after a relaxation time that will be discussed forward in this section.

In order to determine the stationary solution of Eq. (2.19), we introduce the generating function $G(z,w) = \sum_{n,m=0}^{\infty} p_{n,m} z^n w^m$ and we obtain

$$\begin{aligned}\lambda_R(z-w)\partial_w G + \lambda_T(w-z)\partial_z G + \rho[(1-z)\partial_z G \\ + (1-w)\partial_w G] + \mu(z-1)G + \tilde{\mu}(w-1)G = 0, \quad (2.20)\end{aligned}$$

whose solution can be calculated by using the method of characteristics. After imposing the condition $G(1,1) = 1$ (normalization), we have

$$G(z,w) = e^{\frac{(z-1)[\lambda_R(\mu+\tilde{\mu})+\mu\rho]+(w-1)[\lambda_T(\mu+\tilde{\mu})+\tilde{\mu}\rho]}{\rho(\lambda_R+\lambda_T+\rho)}}, \quad (2.21)$$

and by recursive differentiation, we obtain the stationary probability

$$p_{n,m} = \left[\frac{(\partial_z)^n (\partial_w)^m}{n!\, m!} G(z,w) \right]_{\substack{z=0,\\ w=0}} = \frac{e^{-\bar{N}}\, \bar{n}^n\, \bar{m}^m}{n!\, m!}, \tag{2.22}$$

where $\bar{n}$, $\bar{m}$ and $\bar{N}$ are the average values of the quantities n, m and $N = n + m$, respectively:

$$\begin{aligned} \bar{n} &= \langle n \rangle = \frac{\lambda_R \bar{N} + \mu}{\lambda_T + \lambda_R + \rho}, \\ \bar{m} &= \langle m \rangle = \frac{\lambda_T \bar{N} + \tilde{\mu}}{\lambda_T + \lambda_R + \rho}, \\ \bar{N} &= \langle n + m \rangle = \frac{\mu + \bar{\mu}}{\rho}. \end{aligned} \tag{2.23}$$

The stationary distribution Eq. (2.22) is a factorized Poissonian in n and m [10]: the two variables are uncorrelated *at the same time*. However, as we show in App. B.1, n and m are non-trivially correlated *at different times*, and we anticipate that the structure of these correlations leads to the deviations of the CTTD from the exponential form.

The parameters μ and $\tilde{\mu}$ can be conveniently expressed in terms of the diffusion parameter ρ, the average tRNA number $\bar{N}$ and the fraction X of charged tRNA in the reservoir, assumed to be constant:

$$\begin{aligned} \mu &= X \bar{N} \rho, \\ \tilde{\mu} &= (1 - X) \bar{N} \rho. \end{aligned}$$

These parameters have a simple physical interpretation or can be measured in living systems, as we show in Sec. 2.3.4. For instance, the parameter ρ is related to the diffusivity of the tRNA, while the fraction X of charged tRNA in the cell can be measured experimentally [33].

Moreover, in order to simplify the notation, let us rescale the time such that $\lambda_T = 1$, and set $\lambda_R = \lambda$. Let us also introduce the average fraction x of charged tRNAs into the system:

$$x \equiv \frac{\lambda + X\rho}{1 + \lambda + \rho}. \tag{2.24}$$

The average values for n and m can be expressed as $\bar{n} = \bar{N}x$ and $\bar{m} = \bar{N}(1 - x)$.

These quantities and the stationary distribution Eq. (2.22) behave as expected in the limit $\rho \to \infty$: the system is at equilibrium with the cell and

[10] Note that, since detailed balance does not hold in general, the system is out of equilibrium and it is not a priory obvious to find a stationary distribution [127, 90]

the average fraction x of charged tRNA therein coincides with the fraction X in the cell: $x = X$. On the other hand, if $\rho \to 0$, the diffusion is much slower than translation, and the average number of charged tRNAs is completely determined by the internal dynamics: $x = \lambda(1+\lambda)^{-1}$. In this case the effect of diffusion amounts to a slow but not negligible fluctuation of the tRNA number $N = n + m$.

The relaxation times can be deduced from the time-dependent solution of Eq. (2.19), obtained in App. B.1. The relaxation to the stationary distribution of this solution is ruled by the two time scales as in Eq. (B.7) and, as a result, we expect that the stationary state is reached when the observation time is larger than the largest time scale: in this case, $T_{\rm obs} \gg \rho^{-1}$. This means that the convergence to the stationary distribution is guaranteed once the fluctuations in the number N of tRNAs in the system have been sampled.

Finally, as we show in App. B.2, we observe that the detailed balance condition is satisfied only for $\rho = 0$, i.e., in the absence of diffusion, or for $\lambda = X/(1 - X)$. In the latter case the stationary average values for the charged fraction of tRNA of both the internal and the diffusive dynamics coincide, and $x = X$. In all other cases the stationary state is a non-equilibrium state.

2.3.3 Statistics of translation times

The average translation time per codon predicted by this model is trivially $1/\bar{n}$. In general, however, when the distribution is not exponential, the average does not fully characterize the behavior of the random variable. We stress that the shape of this probability influences the dynamics of the whole translation machinery by affecting the translocation of the ribosomes along the mRNA. In this section we therefore compute analytically the probability density function for the intervals between two subsequent translation events.

The derivation is carried out by writing a master equation which accounts for an auxiliary variable r counting the number of time steps elapsed since the last translation event (see below). This procedure allows the calculation of the cumulative distribution of the translation intervals and finally of the CTTD.

Let us therefore consider a discrete-time dynamics where δt is the unit time step. The state of the system is now described by $(n, m; r)$, where the counter r, at each time step, is either set to zero if a translation event occurs, or increased by one otherwise. Without loss of generality, we set $\lambda_T = 1$ from the beginning. The possible transitions are:

- $(n,m;r) \xrightarrow{m\lambda\,\delta t} (n+1,m-1;r+1)$: one uncharged tRNA gets recharged
- $(n,m;r) \xrightarrow{n\,\delta t} (n-1,m+1;0)$: one codon is translated and one charged tRNA gets discharged
- $(n,m;r) \xrightarrow{\mu\,\delta t} (n+1,m;r+1)$ and $(n,m;r) \xrightarrow{\tilde{\mu}\,\delta t} (n,m+1;r+1)$: a tRNA (respectively charged, uncharged) enters the system from the reservoir
- $(n,m;r) \xrightarrow{n\rho\,\delta t} (n-1,m;r+1)$ and $(n,m;r) \xrightarrow{m\rho\,\delta t} (n,m-1;r+1)$: a tRNA (respectively charged, uncharged) leaves the system to the reservoir
- $(n,m;r) \xrightarrow{1-\delta t[n+\lambda m+\mu+\tilde{\mu}+\rho n+\rho m]} (n,m;r+1)$: nothing happens and the counter is increased.

This set of rates leads to the discrete time master equation for the probability $q_{n,m;r}(t)$ of being in the state $(n,m;r)$ at time t

$$\begin{aligned}\frac{q_{n,m;r}(t+\delta t)-q_{n,m;r-1}(t)}{\delta t} &= \lambda(m+1)\,q_{n-1,m+1;r-1}(t)+\mu\,q_{n-1,m;r-1}(t)\\ &+\tilde{\mu}\,q_{n,m-1;r-1}+\rho(n+1)q_{n+1,m;r-1}(t)+\rho(m+1)q_{n,m+1;r-1}\\ -\big[n+\lambda m+\mu+\tilde{\mu}+\rho n+\rho m\big]q_{n,m;r-1}(t)&+\delta_{r,0}\sum_{r'=0}^{\infty}(n+1)q_{n+1,m-1;r'}(t).\end{aligned} \tag{2.25}$$

The limit $\delta t \to 0$ is well defined by setting $\tau = r\delta t$ and it results in the following partial differential equation:

$$\begin{aligned}\partial_t q_{n,m}(\tau,t) &= -\partial_\tau q_{n,m}(\tau,t)+\lambda(m+1)q_{n-1,m+1}(\tau,t)-(n+\lambda m)q_{n,m}(\tau,t)\\ &+\rho\big[(n+1)q_{n+1,m}(\tau,t)+(m+1)q_{n,m+1}(\tau,t)-(n+m)q_{n,m}(\tau,t)\big]\\ +\mu\,q_{n-1,m}(\tau,t)&+\tilde{\mu}\,q_{n,m-1}(\tau,t)-(\mu+\tilde{\mu})q_{n,m}(\tau,t)+(n+1)\delta(\tau)p_{n+1,m-1}(t),\end{aligned} \tag{2.26}$$

where $p_{n,m}(t)=\int_0^\infty d\tau\, q_{n,m}(\tau,t)$ is the solution of Eq. (2.19).

The differential equation for the stationary probability is obtained by setting $\partial_t q_{n,m}(\tau,t)=0$, and reads

$$\begin{aligned}\partial_\tau q_{n,m}(\tau) &= \lambda(m+1)q_{n-1,m+1}(\tau)-(n+\lambda m)\,q_{n,m}(\tau)\\ +\rho\Big[(n+1)q_{n+1,m}(\tau)&+(m+1)q_{n,m+1}(\tau)+X\bar{N}\,q_{n-1,m}(\tau)+(1-X)\bar{N}\,q_{n,m-1}(\tau)\\ &-(n+m+\bar{N})q_{n,m}(\tau)\Big]+\delta(\tau)\,\alpha_{n+1,m-1},\end{aligned} \tag{2.27}$$

where $\alpha_{n,m} = n\, p_{n,m}$ and $p_{n,m}$ is provided by Eq. (2.22).

As in the previous case, we introduce the generating function

$$G(z,w;\tau) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} q_{n,m}(\tau) z^n w^m, \tag{2.28}$$

and Eq. (2.27) becomes

$$\partial_\tau G = \lambda(z-y)\partial_y G - z\partial_z G + \rho\Big[(1-z)\partial_z + (1-w)\partial_w + \bar{N}X(z-1) + \bar{N}(1-X)(w-1)\Big]G + \delta(\tau) f(z,w), \tag{2.29}$$

where

$$\begin{aligned} f(z,w) =& \sum_{n=0}^{\infty}\sum_{m=1}^{\infty} \alpha_{n+1,m-1} z^n w^m \\ =& \bar{n}\, w\, \exp[\bar{n}(z-1) + \bar{m}(w-1)]. \end{aligned} \tag{2.30}$$

Even though Eq. (2.29) could be solved in full generality, here we are interested in the particular value $G(1,1,\tau)$ as it coincides with the marginal distribution

$$Q(\tau) = \sum_{n,m=0}^{\infty} p_{n,m}(\tau) = G(1,1;\tau) \tag{2.31}$$

for τ. The probability $P(\tau)$ for the time interval t between two subsequent translation events to be $t > \tau$, is proportional to $Q(\tau)$. In fact, let us suppose that, at some time T during the evolution of the system, the auxiliary variable has a value $\tau = \tau^*$. In this case, the time interval t between the two subsequent translation events enclosing T is, by construction, $t > \tau^*$. It follows that $P(\tau) = Q(\tau)/Q(0)$.

By solving Eq. (2.29) with $w = z$, we obtain the generating function $G(z,z;\tau)$ and the probability $P(\tau)$, describing the probability for the time t between two consecutive translation events to be larger than τ:

$$\begin{aligned} P(\tau) = & \left[R + \frac{\lambda}{\lambda-1}\left(\frac{e^{-\tau(\rho+1)}}{\rho+1} - \frac{e^{-\tau(\rho+\lambda)}}{\rho+\lambda}\right)\right] e^{-\tau R\bar{N}x} \\ & \times \exp\Big[\frac{\lambda\bar{N}x}{\lambda-1}\left(\frac{e^{-\tau(\rho+1)}-1}{(\rho+1)^2} - \frac{e^{-\tau(\rho+\lambda)}-1}{(\rho+\lambda)^2}\right)\Big], \end{aligned} \tag{2.32}$$

where

$$R = \frac{\rho(\rho+\lambda+1)}{(\rho+1)(\rho+\lambda)} \tag{2.33}$$

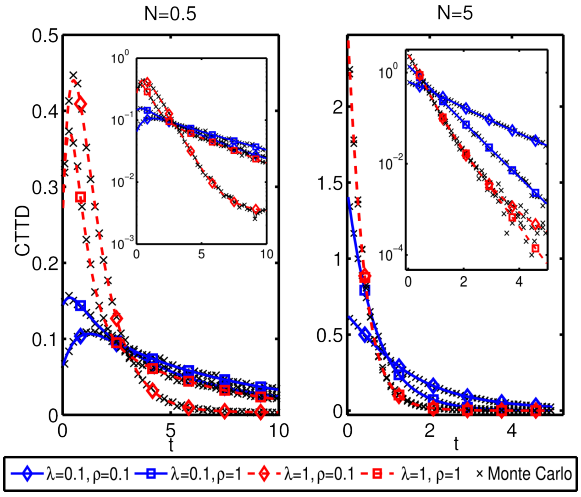

Figure 2.4: Probability density function $p(t)$ for the translation time (CTTD) for various choices of the parameters λ (blue and continuous lines for $\lambda = 0.1$, red and dashed lines for $\lambda = 1$), ρ (diamonds for $\rho = 0.1$, squares for $\rho = 1$) and $\bar{N}$ ($\bar{N} = 0.5$ in the left panel, $\bar{N} = 5$ in the right one). The parameter X is kept fixed to 0.5. The black crosses are the results of Monte Carlo simulations, and do not show any significant deviation from the theoretical predictions. As the two log-plot insets show, deviations from a pure exponential are most evident for small $\bar{N}$, where the fluctuations play a relevant role.

is always < 1, and x is the fraction of charged tRNAs in the system, Eq. (2.24).

For further reference, note that the function $P(\tau)$ can be written as $P(\tau) = \partial_\tau A(\tau)$, with

$$A(\tau) = -\frac{1}{\bar{N}x} \exp\Big[-\tau R\bar{N}x + \frac{\lambda\bar{N}x}{\lambda - 1}\left(\frac{e^{-\tau(\rho+1)} - 1}{(\rho+1)^2} - \frac{e^{-\tau(\rho+\lambda)} - 1}{(\rho+\lambda)^2}\right)\Big]. \quad (2.34)$$

Let us now observe that $P(\tau)$ is the complement of the cumulative distribution for the CTTD, defined as $p(t)$. Therefore, since $Q(\tau) = \int_\tau^\infty dt\, p(t)$, the CTTD is given by

$$p(t) = -\partial_\tau P(\tau)_{|\tau=t}. \quad (2.35)$$

Some typical realizations of $p(t)$ are shown in Fig. 2.4, where we also compare the theoretical prediction with the numerical Monte Carlo simulations. We did not observe any significant deviation between the theoretical results and

the simulations. Interestingly, for small times and small values of $\bar{N}$ the CTTD relevantly deviates from an exponential (see the log-plot insets of Fig. 2.4). On the other hand, these deviations are milder for small values of λ and large values of ρ. The main features the $p(t)$ are analyzed in the next section.

Characterization of the distribution of the translation times

In order to characterize the distribution $p(t)$, we calculate its first two moments and we compare them to an exponential distribution having the same mean, observing that the CTTD is overdispersed with respect to that distribution.

Let us first check the consistency of the average of the CTTD:

$$\langle t \rangle = \int_0^\infty dt\, t\, p(t) = \int_0^\infty d\tau\, P(\tau) = \frac{1}{\bar{N}x}. \tag{2.36}$$

As expected, it coincides with the inverse of the average number $\bar{n}$ of charged tRNAs in the system.

The second moment is given by

$$\langle t^2 \rangle = \int_0^\infty dt\, t^2\, p(t) = 2\int_0^\infty dt\, t\, P(t) = -2\int_0^\infty dt\, A(t), \tag{2.37}$$

where $A(t)$ is given by Eq. (2.34), and can be written as

$$\langle t^2 \rangle = \frac{2}{\bar{N}x} \exp\left[\frac{\lambda\bar{N}x}{\lambda-1}\left(\frac{1}{(\rho+\lambda)^2} - \frac{1}{(\rho+1)^2}\right)\right] \times \int_0^1 dy\, y^{R\bar{N}x-1} \exp\left[\frac{\lambda\bar{N}x}{\lambda-1}\left(\frac{y^{\rho+1}}{(\rho+1)^2} - \frac{y^{\rho+\lambda}}{(\rho+\lambda)^2}\right)\right]. \tag{2.38}$$

Equation (2.38) can be numerically evaluated in order to determine the variance $\sigma_t^2 = \langle t^2 \rangle - \langle t \rangle^2$.

In Fig. 2.5 we plot the ratio $\sigma_t^2/\sigma_{\rm exp}^2$ for various values of the parameters, where $\sigma_{\rm exp}$ is the variance of the exponential distribution

$$p_{\rm exp}(t) = x\bar{N}e^{-x\bar{N}t}, \tag{2.39}$$

fixed to having the same average of the CTTD. By inspection, we did not find any point in the parameter space such that $\sigma_t < \sigma_{\rm exp}$: the CTTD is overdispersed with respect to the exponential distribution, Eq. (2.39).

This observation can be further characterized by comparing the small and large t expansions of the two distributions: first, by analyzing the Taylor expansion around $t = 0$ of the two probability distributions, we observe that

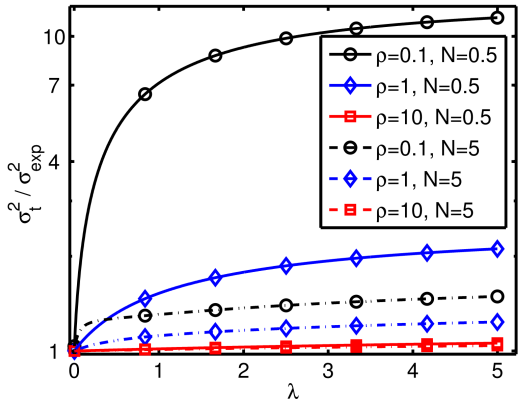

Figure 2.5: Ratio between the variances σ_t^2 of the CTTD, and $\sigma_{\rm exp}^2$ of the exponential with the same average Eq. (2.39), as a function of λ, for various values of $\bar{N}$ and ρ (the parameter X do not change qualitatively the results and is set to $X = 1/2$). For $\lambda \to 1$ the CTTD converges to an exponential distribution, as shown by Eq. (2.41), and the ratio $\sigma_t^2/\sigma_{\rm exp}^2 \to 1$. In general, however, the CTTD is over dispersed with respect to the exponential distribution.

$p(t) - p_{\rm exp}(t) \sim \lambda t + O(t^2)$. Short translation times are under represented in the exponential distribution. Also note that for $\lambda = 0$ the two distributions are the same and the ratio of the variances, as shown in Fig. 2.5, is $= 1$.

Similarly, the tails of the two distributions differ in the large t limit. In fact, Eq. (2.35) behaves for $t \to \infty$ as:

$$p(t) \propto e^{-Rx\bar{N}t} \tag{2.40}$$

with $R < 1$. Accordingly, long translation times are also under represented in the exponential distribution.

Finally, we observe that the CTTD in Eq. (2.35) reduces to an exponential both in the slow recharge limit $\lambda \to 0$, where

$$p(t) \to x\bar{N}e^{-x\bar{N}t}, \tag{2.41}$$

and in the fast diffusion limit $\rho \to \infty$:

$$p(t) \to X\bar{N}e^{-X\bar{N}t}. \tag{2.42}$$

In the latter case the charged fraction x of tRNA in the system coincides with the fraction X in the reservoir, consistently with the expectation that in the fast diffusion limit the fluctuations of charged tRNAs are determined by the exchange with the bath and are uncorrelated in time. As we show in the next section, this two limits have an interesting physical interpretation.

Time correlations in the number of charged and uncharged tRNAs induce the deviation from the exponential

The model that we introduced in Sec. 2.3.1 is Markovian and memoryless, as its time evolution depends only on the present state. Since this fact typically implies exponentially distributed intervals between transitions [41], the appearance of a non-exponential CTTD could be surprising at first sight. Here we show that the deviation from an exponential of the CTTD arises due to a nontrivial coupling between the fluctuations of n and m.

First, as we show in App. B.3, the CTTD can be written as a function of the time evolution of the average value of n after a translation event, $\langle n(t)\rangle_{\rm nt}$, conditioned to the fact that no other translation events were recorded up to time t. As Eq. (B.17) shows, the deviations from an exponential of the CTTD appear as soon as $\langle n(t)\rangle_{\rm nt}$ departs from a constant and acquires a time dependency.

In the stationary regime the probability for a translation event to occur is proportional to np_n, where $p_n = \sum_m p_{mn}$ is the marginal stationary probability for n, obtained from Eq. (2.22). Precisely, the distribution for n at the instant before a translation event is:

$$p_n^{\rm tr^-} = \frac{\bar{n}^{n-1}}{(n-1)!}e^{-\bar{n}}, \tag{2.43}$$

whose average is $\bar{n}_{\rm t^-} = \bar{n} + 1$: a translation event typically occurs when a fluctuation increases the number of charged tRNAs in the system (note that the number m of uncharged tRNAs is not influenced). During a translation event $n \to n-1$ and $m \to m+1$, therefore, immediately after the translation, $\bar{n}_{\rm t^+} = \bar{n}$ and $\bar{m}_{\rm t^+} = \bar{m} + 1$: the fluctuation on n has propagated to m. Now, if $\lambda > 0$ and $\rho < \infty$, this fluctuation can again propagate to n with a characteristic time scale, producing a loop which induces a time dependency in $\langle n(t)\rangle_{\rm nt}$. This mechanism is suppressed if $\lambda = 0$ and it is negligible if $\rho \to \infty$ since, as it can easily be seen from the rates described at the beginning of Sec. 2.3.1, the dynamics of n is not affected by the dynamics of m. Fig 2.6 shows that for $\lambda \to 0$ the average $\langle n(t)\rangle_{\rm nt}$ reduces to a constant. For $\rho \to \infty$ the fluctuations on m are immediately dissipated in the thermal bath before they can propagate back to n.

As a further check, we can quantify the influence of m at a given time on the future n dynamics by studying the two point correlator:

$$C_{mn}(t) \equiv \langle m(0)n(t)\rangle - \bar{m}\bar{n}. \tag{2.44}$$

The analytic expression for such correlator is obtained in App. B.1 and, as it is shown in Eq. (B.8), it vanishes identically when $\lambda = 0$ or $\rho \to \infty$. In

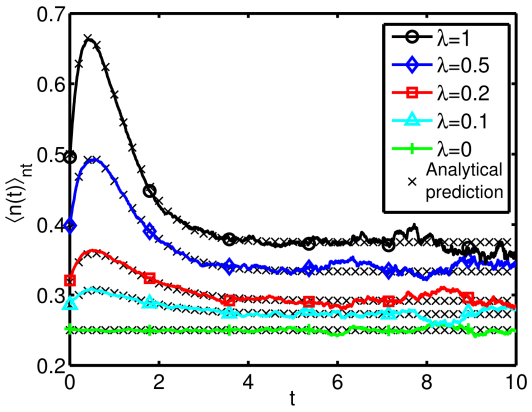

Figure 2.6: Monte Carlo evaluations of the average number of charged tRNAs in the system $\langle n(t)\rangle_{\rm nt}$ conditioned to the fact that no translation event occurred up to time t after that one at $t = 0$, plotted for various values of λ (solid lines). The other parameters were set to $\bar{N} = 1$, $\rho = 1$ and $X = 1/2$. The black crosses are the analytical predictions, based on Eq. (B.18). Interestingly, as λ approaches 0, the function $\langle n(t)\rangle_{\rm nt}$ becomes a constant.

these conditions the dynamics of n is completely decoupled from m since the fluctuations of m cannot propagate to n (note, however, that the reverse is not true, since when $\lambda = 0$ the two variables are not independent as shown by the fact that the correlator $C_{nm}(t)$ in Eq. (B.8) does not vanish).

On the other hand, for $\lambda > 0$ and $\rho < \infty$, the correlator in Eq. (2.44) is a linear combination of two exponentials with different decay times, and interestingly it is not monotonic as a function of time: at $t = 0$ it vanishes (as expected from the factorization of the stationary probability) and it has a maximum for $t = t_{\rm max}$ reported in Eq. (B.9).

Let us finally observe that the CTTD can be regarded as an observable on the timeseries of n. Now, if we consider the process projected on the n variable only, its behavior is not Markovian since the future evolution is not completely determined by the present state: $n(t)$ depends on its history because the information on m is missing. The two limits $\lambda = 0$ and $\rho \to \infty$, where the CTTD is exponential, are coherent with the absence of memory in the timeseries of n, as the exponential is the only memoryless continuous distribution [41]. We emphasize, however, that the complete model is always Markovian, as its time evolution depends only on the current state (n, m) and not on the past history.

Cell volume[1]	$V_c \approx 6 \times 10^{-19}\,\mathrm{m}^3$
Typical radius of a ribosome[2]	$L_R \approx 1 \times 10^{-8}\,\mathrm{m}$
Number of tRNA in a cell[3]	$N_t \approx (0.1 \div 5) \times 10^3$
Number of ribosomes in a cell[3]	$N_R \approx (7 \div 70) \times 10^3$
Concentration of the tRNA[3]	$C_t \approx (0.3 \div 20) \times 10^{-6}\mathrm{M}$
Diffusion constant of the tRNA[4]	$D \approx (0.2 \div 2) \times 10^{-12}\,\mathrm{m}^2/\mathrm{s}$
Average translation rate[5]	$\Theta \approx 10 \div 20$ codons/s

Table 2.1: Data for E.Coli.

2.3.4 Discussion and interpretation of the parameters

The parameters ρ, $\bar{N}$, X, λ_R and λ_T which define the model can be interpreted in terms of the physical and biological quantities given in Tab. 2.1.

Let us first suppose that the model is enclosed in a sphere of radius r centered around the ribosome. This radius is interpreted as the effective distance such that a tRNA has a non-negligible probability of diffusing towards (and being captured by) the ribosome. As shown in Ref. [95], the probability of being absorbed by a target of radius L_R centered at the origin, starting from radius r, is $P_{\mathrm{abs}}(r) = L_R/r$. Since we want the probability $P_{\mathrm{abs}}(r)$ to be finite (i.e., we want to consider those tRNAs which have a non-zero probability of diffusing into the ribosome in a short time), an order-of-magnitude estimate for the radius r is the ribosome radius itself: $r \gtrsim L_R$.

Defining $V_r = 4\pi r^3/3$ as the effective volume around the ribosome, we can estimate the average number of tRNAs by fixing the concentration in the volume V_r to be the same as in the cell (C_t). We obtain $\bar{N} = V_r C_t N_a \times 10^3 \approx 10^{-2} \div 10^0$, where N_a is the Avogadro number and the numerical factor results from the conversion of the unities of volume. The wide variation is due to the fact that different species of tRNA have very different concentrations in the cell.

The parameter X measures the fraction of charged tRNAs in the cell. Its range, measured *in vivo* for E. Coli in Ref. [33], spans the interval $X \approx 10^{-3} \div 10^0$, depending on the richness of the growth media.

The diffusion rate ρ can be interpreted in terms of the diffusion constant D of the tRNA molecules and the volume V_r. Supposing that each tRNA

[1] From Ref. [74]
[2] From Ref. [125]
[3] From Ref. [35]
[4] From Ref. [116]
[5] From Ref. [14]

performs a Brownian motion, its mean square displacement in the time T is $r^2 = \langle \sum_i \Delta x_i^2 \rangle = 6DT$, and the typical exit time from the sphere of radius r is

$$T_{\text{exit}} = \frac{r^2}{6D}. \tag{2.45}$$

The average exit time in the stochastic model is given by $1/\rho$, and equating the two times we have

$$\rho = \frac{6D}{r^2} \approx (1 \div 70) \times 10^3\,\text{s}^{-1}. \tag{2.46}$$

The average translation rate in our model, Eq. (2.36), is given by $\langle t \rangle^{-1} = \bar{N}x$. Restoring the λ_T dependence and equating with the average experimental translation rate Θ, we have

$$\langle t \rangle^{-1} = \bar{N}\lambda_T \frac{\lambda_R + X\rho}{\lambda_T + \lambda_R + \rho} \approx (10 \div 20)\,\text{s}^{-1}. \tag{2.47}$$

Since $\bar{N} \approx 10^{-2} \div 10^0$, we roughly have $\lambda_T \approx 10^1 \div 10^3$. In the previous sections we set $\lambda_T = 1$, which corresponds to using the ratio ρ/λ_T. This ratio can range between $\rho/\lambda_T \approx 10^0 \div 10^3$, depending on the tRNA species and the growth conditions of the organism.

The process of tRNA recharge involves many components (the enzyme aminoacyl-synthetase, ATP and elongation factor molecules) and we expect λ_R to be an increasing function of the concentrations of the reagents. Nevertheless, our one-parameter approximation serves the purpose of effectively modeling the range of variation of the recharge dynamics.

Interestingly, a recent analysis of mRNA sequences [19] revealed that the subsequent occurrences of the same amino acid tend to be encoded by codons translated by the same tRNA. More specifically, the probability of tRNA recycle decays as a function of the distance along the mRNA, i.e., of the average time between the two translation events. This measure seems to imply that (*i*) the timescale of translation is not larger than that of diffusion: if diffusion were very fast, no correlation along the sequences would be observed, that (*ii*) the number of tRNAs in the neighborhood of a ribosome is small since the addition of one tRNA induces a relevant fluctuation, and that (*iii*) the recharge dynamics is at least as fast as translation, otherwise that tRNA would not be charged and it would not be recycled. These observations, along with the previous order-of-magnitude estimations for ρ and λ_T, suggest that $\lambda_T \approx \rho$, $\bar{N} \lesssim 1$, and $\lambda_R \gtrsim \lambda_T$. Moreover, by equating the total recharge rate to the total translation rate in the cell, we find $\lambda_T X \approx \lambda_R(1 - X)$. Since $X < 1$ [33], we expect λ_T to be roughly of the same order of magnitude of λ_R. In this range of parameters ($\bar{N} \lesssim 1$,

$\lambda_R \approx \lambda_T \approx \rho$) our model predicts significant deviations from an exponential distribution, which could be potentially measured with the techniques employed in Ref. [115].

2.4 Conclusions

In the previous sections we developed a general model for systemic translation in the cell, progressing through three steps of increasing detail.

We began by analyzing the effects of ribosome load on translation optimality, observing that the optimization of the translation is a global problem.

We then progressed to include tRNA recharging dynamics and codon usage. Depending on the recharging rate of the tRNAs, the optimal usage of synonymous codons interpolates between the proportional rule (the tRNAs are used accordingly to their abundance, slow recharging) and the single-tRNA rule (only the most abundant tRNA is used, fast recharging). However, since the codon usage cannot vary as fast as the environment, we suggest that the codon usage might be interpreted as a strategy to deal with the randomness of the environment.

Finally, we focused on the single translation events: using a stochastic, spacial model we obtained the distribution of the time between translation events (ribosome dwell time). This distribution can deviate from an exponential due to the coupling of the fluctuations of the number n and m of charged and uncharged tRNAs, respectively. The qualitative mechanism is as follows: (i) a translation event typically occurs when the number of charged tRNAs around a ribosome is increased due to a fluctuation, on average $\bar{n}_{\mathrm{tr}^-} = \bar{n} + 1$, (ii) during the translation event, a charged tRNA gets discharged and the fluctuation on n propagates to m, $\bar{m}_{\mathrm{tr}^+} = \bar{m} + 1$, (iii) if $\lambda > 0$, this fluctuation on m can propagate again on n with a characteristic timescale, producing a "bump" in the timeseries of n as in Fig. 2.6. The size of this effect is larger the smaller the average number of tRNAs $\bar{N}$ is - i.e., the bigger the relative size of the fluctuations is. As we remark in section 2.3.4, the typical values for $\bar{N}$ are $\lesssim 1$. The other parameters can be in a range where this mechanism might be relevant and lead to a significant deviation from the exponential distribution.

Furthermore, this mechanism can contribute in explaining the spacial organization of the codons which was observed in Sec. 1.1.2. As shown in Ref. [19], this anomaly is due to the fact that close occurrences of the same amino acid tend to be encoded utilizing codons translated by the same tRNA. Since the tRNA concentration is small, a tRNA which reaches the ribosome constitute a significant fluctuation. If the tRNA, after being

used, is rapidly recharged before it diffuses away, the local concentration of charged tRNAs is effectively increased. Re-using by choosing the same codon among its synonyms can therefore significantly reduce the waiting time.

Chapter 3

Simulating the data: the effect of codon translation rates on cotranslational folding

The process which leads from a mRNA to a protein requires that the nascent polypeptide assumes its functional 3-dimensional conformation, as the failure to fold into the native structure produces inactive and potentially toxic molecules (several neurodegenerative diseases, including Parkinson's and Alzheimer's, are caused by the aggregation of misfolded proteins into insoluble amyloids). However, the fact that a protein consistently folds from a random coil to an ordered and functional structure in a very short time -seconds or less- might seem paradoxical: if the dynamics were purely exploratory (i.e., random), finding the minimum free energy configuration could take an enormous amount of time.[1] Plausibly, the folding of the protein is therefore biased toward the correct native state, and a bias of the order of a few k_BT can in fact reduce the folding time to the biological scale [128]. Qualitatively, the folding energy landscape is funnel-shaped: the state of the protein tumbles down a valley which leads to the globally folded state (for a recent review, see Ref. [32]).

[1] In a famous paper [76], Levinthal gave a rule-of-thumb estimate for that time. Suppose that the bond connecting amino acids can have several (e.g., three) possible states. A protein of, say, 101 amino acids has $3^{100} \approx 5 \times 10^{47}$ configurations. If the protein sampled new configurations at a rate of $10^{13}\, s^{-1}$, it would take 10^{27} years to try them all. Since, indeed, proteins fold on time scales of less than a second, random sampling does not explain the folding process.

Besides pure free energy minimization, the folding is facilitated by I) the presence of a set of molecules, the chaperones, (see Ref. [57] for a recent review) and by II) fine tuning of the translation process, via co-translational folding. In the following we will focus on the latter process, as it is strictly related to the choice of the codons along the mRNA sequences.

The study of this problem on even a single proteins is computationally extremely cumbersome, as the sampling at each different translation rate requires a full set of Molecular Dynamics simulations to be run. This makes practically unfeasible to extensively study how the variation of the single-codon translation rates affects the process of cotranslational folding. In the following section we will develop a Markov chain model which overcomes this severe limitation, by allowing the efficient simulation of arbitrarily chosen translation rates once the data form a highly informative subset of simulations is known.

3.1 Cotranslational folding and codon usage

In the past 10 years several evidence accumulated showing that the folding of the proteins begins while the polypetide chain is still tethered to the ribosome [71, 70, 85]. It has been shown, for instance, that about one-third of E.Coli proteome folds cotranslationally [24].

The cotranslational folding process can be strongly influenced by the rate at which the single amino acids are covalently attached to the nascent polypeptide chain [71]. In fact, there is a large variance in the average time it takes to translate the 61 different codons encoding for the amino acid that comprise an mRNA molecule. For instance, in logarithmically growing E. coli cells the fastest codons are estimated to translate with average times as short as 10 ms, while the slowest codons may take 100s of ms or longer [42, 30], with an average speed of between 45 and 100 ms per codon [78]. An order-of-magnitude or more can thus separate the fastest and the slowest translating codons.

Furthermore, the synonymous codons can be translated at very different rates, although encoding for the same amino acid [42]. An increasing number of experiments shows that, in general, synonymously altering the codon sequence affects the folding of the protein [123, 107, 105].

Analyses of synonymous codon usage across entire transcriptomes reveal systematic biases between different species. In particular, stretches enriched in rare codons (i.e., parts of the mRNA where rare codons are more abundant than the average) are also enriched in α-helical and β-strand structural motifs [88, 100, 110]. It is therefore likely that evolutionary pressures select

for patterns of translation rates along an mRNAs open reading frame. When these patterns of translation rates are altered the process of cotranslational folding can go awry, resulting in misfolding and malfunction of the nascent protein in vivo [66, 124, 28].

For these reasons a grand challenge in the *in vivo* protein folding field is to be able to analytically model the coupling between individual codon translation rates and the states (conformations) that a nascent protein populates during its cotranslational folding process. Attempts at addressing this challenge via a probabilistic approach [84] were successful in deriving equations that modeled single cotranslational folding pathways involving up to three thermodynamic states [85]. More complex situations, however, can occur in cells involving parallel pathways [93] and additional states populated during cotranslational folding [25], such as interdomain misfolding [105]. To model these biologically relevant behaviors requires a general analytic description of the coupling between translation rates and cotranslational folding.

In the following sections we will therefore introduce a general formalism which models the reaction scheme in cotranslational folding. By utilizing standard Markov chain methods we are able to analytically calculate the probability that a nascent protein is in any one of an arbitrarily large number of thermodynamic states during translation, as a function of the nascent chain length, the codon translation rates, and the rates of interconversion between states.

3.2 Absorbing random walks and cotranslational folding

Let us consider a mRNA molecule whose open reading frame (ORF) consists of M codons. We number the codons starting from 1 at the start codon (Fig. 3.1a). A ribosome molecule translating this ORF converts the genomic information encoded in the sequence by unidirectionally translocating along the ORF one codon at a time, decoding the information at the new codon, and covalently attaching the corresponding amino acid to the nascent polypeptide chain before the next translocation step (Fig. 3.1b).

For a nascent chain that is L residues long at a given point during its translation, the rate of translation of the $L+1$-th codon, which elongates the nascent chain by one residue, is denoted by $k_A^{(L+1)}$.

At length L the nascent chain can interconvert between N_L distinct thermodynamic states, e.g., folded, intermediate and misfolded states, which we denote as $S_i(L)$, with $i = 1, \ldots, N_L$ (see Fig. 3.2). These states can directly

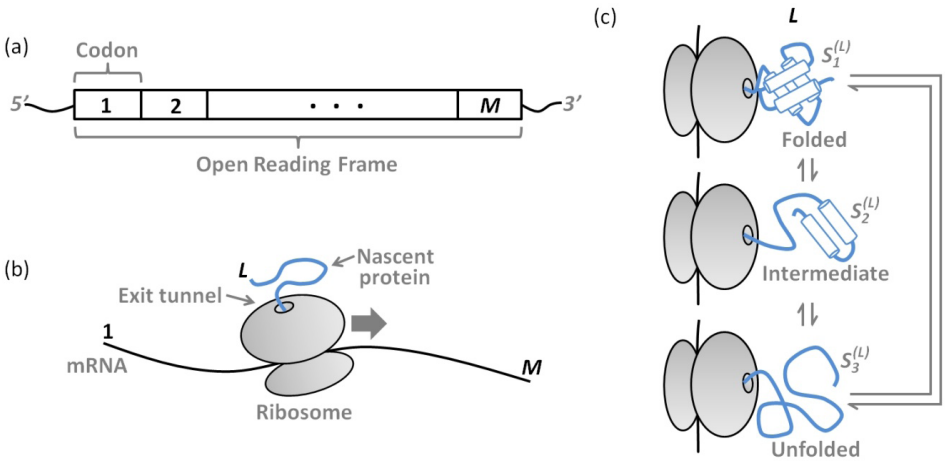

Figure 3.1: Cotranslational protein folding. (a) The ribosome translates the codons contained in a mRNA into a nascent protein. (b) Starting from the 5 end the ribosome unidirectionally translocates (large gray arrow) along the mRNA molecule and converts the genomic information in the ORF into a nascent protein (blue), which emerges through a channel known as the ribosome exit tunnel. (c) At a given nascent chain L during synthesis the nascent chain has the potential to form tertiary structure; such states may include folded, intermediate and unfolded conformations. The arrows indicate that these states may be able to interconvert at the given chain length.

and reversibly interconvert with one another at this nascent chain length. The rate of interconversion between state $S_i^{(L)}$ and $S_j^{(L)}$ is denoted by $k_{i,j}^{(L)}$ (in App. C.1 we give a simple method to estimate these rates). Note well that all the rates have an explicit length dependence, as the chemical environment experienced by a segment of the nascent changes as it elongates. Moreover, we assume that when a translation event occurs the state $S_i^{(L)}$ directly transitions to state $S_i^{(L+1)}$, provided that the timescale of the chemical step of peptide bond formation is much smaller than that of transition between different states of the protein. States $S_i^{(L)}$ and $S_i^{(L+1)}$ are therefore effectively equivalent, having the same conformation of the domains. Finally, observe that the process of amino acid addition is irreversible under physiological conditions: the states at length $L+1$ act as absorbing states for those at length L.

A nascent chain described by this model will thus randomly change its conformation due to thermal fluctuations, performing a random walk on the states $\{S^{(L)}\}$ until a new amino acid is added, at which point state $S_i^{(L)}$ transitions to and is absorbed by state $S_i^{(L+1)}$.

The evolution of the system is captured by the stochastic vector $\mathbf{p}_L(t)$ as

$$\mathbf{p}_L(t) = \left(p_1^{(L)}(t), p_2^{(L)}(t), \ldots, p_N^{(L)}(t)\right), \tag{3.1}$$

where $p_i^{(L)}(t)$ is the probability of being on state i at chain length L and time step t. To model the influence of codon translation rates on cotranslational folding the central quantity we are interested in calculating is the probability $\mathbf{p}_L(t_A)$ at the time when the translation occurs. Note that the final state at length L is the initial condition for the dynamics at length $L+1$, i.e.,

$$\mathbf{p}_{L+1}(0) = \mathbf{p}_L(t_A). \tag{3.2}$$

From this perspective, the cotranslational folding process is a biased random walk on a reaction network consisting of subsets of reactions that can reversibly interconvert, connected by irreversible transitions between those subsets (Fig. 3.2).

3.2.1 Random walks with absorbing states

The problem of calculating the probability vector $\mathbf{p}_L(t_A)$ for the protein to be in each of the N_L states when the translation occurs, admits a very general solution by utilizing the framework of Markov chain with absorbing states [65, 41]. Therefore, given the system in state i, the rates of the

possible transitions

$$\begin{aligned} w(S_i^{(L)} \to S_j^{(L)}) &= k_{ij}^{(L)}, \quad j = 1, \ldots, N_L, \\ w(S_i^{(L)} \to S_i^{(L+1)}) &= k_A^{(L+1)}, \end{aligned} \tag{3.3}$$

produce the following transition probabilities:

$$\begin{aligned} t_{i,j}^{(L)} &= k_{ij}^{(L)}/\mathcal{N}, \quad j = 1, \ldots, N_L \\ a_i^{(L)} &= k_A^{(L+1)}/\mathcal{N}, \end{aligned} \tag{3.4}$$

where $\mathcal{N} \equiv \sum_j k_{ij}^{(L)} + k_A^{(L+1)}$. We also define the two $(N_L \times N_L)$ matrices

$$(\mathbf{T}_L)_{i,j} = t_{i,j}^{(L)}, \tag{3.5}$$

$$(\mathbf{A}_L)_{i,j} = \delta_{i,j}\, a_i^{(L)}. \tag{3.6}$$

This matrices can be arranged blockwise into the stochastic transition matrix $\mathbf{P}^{(L)}$:

$$\mathbf{P}_L = \begin{pmatrix} \mathbf{T}_L & \mathbf{A}_L \\ \mathbf{0} & \mathbf{I}_{N_L}, \end{pmatrix}$$

where now the states $1, \ldots N_L$ are the $\{S_i^L\}$, the states $N_L + 1, \ldots 2N_L$ are the absorbing ones $\{S_i^{L+1}\}$, and $\mathbf{I}_{N_L}$ is the $N_L \times N_L$ identity matrix.

The initial condition can be written as the stochastic vector $\mathbf{p}_L(0) = \left(p_1^{(L)}(0), p_2^{(L)}(0), \ldots, p_N^{(L)}(0)\right)$, where each $p_i^{(L)}(0)$ is the probability of being on the state S_i^L. By taking the product $\mathbf{p}_L(1) = \mathbf{p}_L(0)\mathbf{T}_L$ we compute the probability of being on the sites S_i^L after one step of the random walk without being absorbed. Iterating, after t steps we have

$$\mathbf{p}_L(t) = \mathbf{p}_L(0)\mathbf{T}_L^t. \tag{3.7}$$

On the other hand, the probability of being absorbed exactly at step $t+1$ by any of the state S_i^{L+1} can be computed from eq. (3.7) by applying the matrix $\mathbf{A}_L$, and is equal to $\mathbf{p}_L(0)\mathbf{T}_L^t\mathbf{A}_L$.

The total probability of being absorbed (and thus the initial condition for the random walk at length $L+1$) can be computed by summing $\mathbf{p}_L(0)\mathbf{T}_L^t\mathbf{A}_L$ over t:

$$\mathbf{p}^{(L+1)}(0) = \sum_{t=0}^{\infty} \mathbf{p}^{(L)}(0)\mathbf{T}_L^t\mathbf{A}_L = \mathbf{p}^{(L)}(0)(\mathbf{1} - \mathbf{T}_L)^{-1}\mathbf{A}_L. \tag{3.8}$$

The matrix inversion operation can be performed extremely efficiently up to very large sizes, making possible the treatment of proteins with complicate folding landscapes.

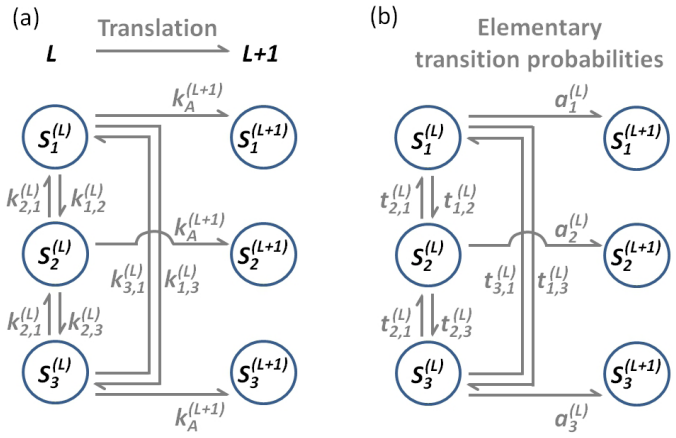

Figure 3.2: A triangular cotranslational folding reaction scheme with (a) rates and (b) elementary transition probabilities indicated. Assuming state S_1 corresponds to the folded state, then a domain that folds via this mechanism can take parallel pathways to the folded state, either directly from S_2 or S_2. At length L these three states can reversibly and directly interconvert with one another with rates $k_{i,j}^{(L)}$ and elementary transition probabilities $t_{i,j}^{(L)}$. Addition of a residue to the nascent chain shifts the system irreversibly from length L to length $L+1$ with rate $k_A^{(L+1)}$ and elementary reaction probability $a_i^{(L)}$ that state $S_i^{(L)}$ transitions to state $S_i^{(L+1)}$ after one step on this reaction network.

Eq. (3.8) is the main theoretical result of this chapter as it provides an exact expression for the probability of being in a given state (e.g., the folded state) during translation for arbitrarily complex folding mechanisms. Specifically, the i-th element of the vector $\mathbf{p}^{(L+1)}(0)$ expresses the probability of finding the nascent chain in state $S_i^{(L)}$ at length $L+1$ immediately the residue has been added in terms of the codon translation rates and interconversion rates between states.

3-state case: an example

To illustrate this method in practice consider the reaction scheme shown in Fig. 3.2 . It represents a domain that can cotranslationally fold via two-parallel pathways. In this situation the matrices $\mathbf{T}_L$ ans $\mathbf{A}_L$ explicitly read

$$\mathbf{T}_L = \begin{pmatrix} 0 & t_{1,2}^{(L)} & t_{1,2}^{(L)} \\ t_{2,1}^{(L)} & 0 & t_{2,3}^{(L)} \\ t_{3,1}^{(L)} & t_{3,2}^{(L)} & 0 \end{pmatrix}, \quad \mathbf{A}_L = \begin{pmatrix} a_1^{(L)} & 0 & 0 \\ 0 & a_2^{(L)} & 0 \\ 0 & 0 & a_3^{(L)} \end{pmatrix}, \tag{3.9}$$

with $t_{i,j}^{(L)}$ and $a_i^{(L)}$ given by Eq. (3.4).

To test the accuracy of the predictions from this approach we ran Langevin Dynamics simulations of the synthesis of the MIT protein domain, which folds into a three-helix bundle structure and can do so via a three-state parallel pathway mechanism. The three states that can be populated by the MIT domain are the unfolded state, an intermediate comprised of natively-structured helices 1 and 2 (Fig. 3.3a), and the fully folded state (Fig. 3.3b). The coarse-grained model and simulation protocol is described in App. C.2.

Two sets of simulations were carried out: I) the first was used as part of the process of making the predictions, and II) the second was used to test those predictions. In the first set of simulations a series of arrested ribosomes at nascent chain lengths ranging from 65 to 120 residues were simulated. Arrested ribosomes do not undergo translation, i.e., $k_A^{(L)} = 0$. The MIT domain is 77 residues in length, therefore this domain emerges fully from the narrow ribosome exit tunnel at a nascent chain length of around 110 residues in its fusion construct with polyglycine (Fig. 3a). At each length L the rates $k_{i,j}^{(L)}$ of interconversion between the 3 states were measured, using the method reported in App. C.1.

In this example we assumed constant $k_A^{(L)}$ along the whole sequence, and we chose 3 significant values, namely $0.83k_F^{(\mathrm{bulk})}$, $8.3k_F^{(\mathrm{bulk})}$, and $83k_F^{(\mathrm{bulk})}$, where $k_F^{(\mathrm{bulk})}$ is the folding rate of the MIT domain at 310 K in bulk solution -i.e., when no ribosome is present- and is equal to 12.4 ns^{-1}.

Using the set of rates $k_{i,j}^{(L)}$ extracted from the simulations we computed the matrices $\mathbf{T}_L$ and $\mathbf{A}_L$ for the three different values of $k_A^{(L)}$. We plugged them into Eq. (3.8) and predicted how the MIT domain behaves during continuous translation. We note that coarse-grained models and low-friction Langevin Dynamics significantly speeds up the folding rate relative to experimentally observed values [68] but preserves realistic thermodynamic properties. While these predictions are for ORFs with uniform translation rate profiles, we emphasize that our model can easily treat arbitrarily non-uniform profiles as well.

We tested these predicted state curves against explicit simulations of continuous translation. In this second set of simulations residues were stochastically attached to the C-terminus of the ribosome-bound nascent chain with the rates $k_F^{(\mathrm{bulk})}$ used to make the predictions.

We find that Eq. 3.8 yields accurate predictions of the effect of codon translation rates on the probability of the MIT domain being in the folded, intermediate and unfolded states (Fig. 3.3d). This indicates that the kinetic model, Eq. 3.8, captures the essential features present in cotranslational folding mechanism and can produce accurate predictions about the influence of codon translation rates on arbitrarily complex cotranslational folding mechanisms.

3.2.2 Conclusions and perspectives

We introduced a novel method for studying the cotranslational folding mechanism of proteins. By utilizing the rates obtained from arrested ribosome Molecular Dynamics simulations, accurate predictions for the probability of being in any of the folded state as a function of the chain length are obtained, as showed by Fig. 3.3d.

The biological importance and influence of codon translation rates on the proper folding and functioning of nascent proteins is coming to the forefront in a number of fields including molecular and cellular biology [107], cancer biology [44], personalized medicine [66], and biotechnology [3]. What is currently lacking in these fields, however, is a theoretical framework to understand, model and predict the influence of codon translation rates on these processes. Eq. (3.8) provides an integral part of that framework as it provides a general and computationally efficient methodology to make predictions about the consequences of changing individual codon translation rates for cotranslational folding and misfolding.

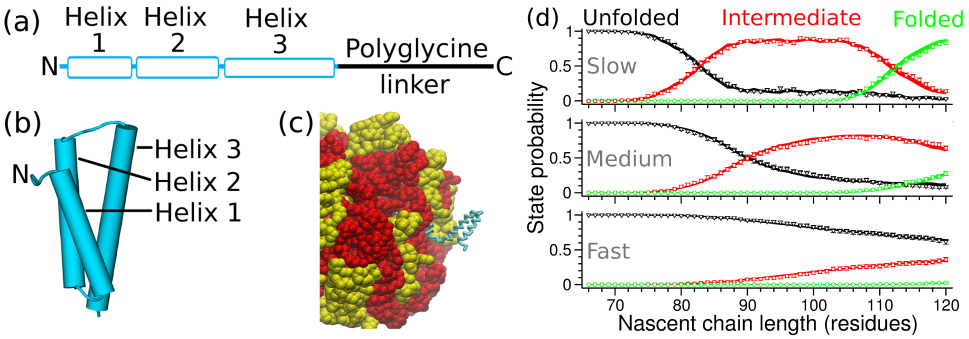

Figure 3.3: Eq. (3.8) accurately predicts the effect of codon translation rates on the probability of populating different states during cotranslational folding in the coarse-grained Langevin Dynamics simulations. The 77-residue MIT domain consists of three helices and was fused to the N-terminus of an unstructured 43-residue polyglycine linker (a). The domain forms a helix bundle in the folded state (b). The synthesis of this nascent chain was simulated using a coarse-grained model, a simulation structure of which is shown in (c) in which the intermediate is present (the large red and yellow structure is the ribosome). (d) The populations of unfolded, intermediate and folded states are shown, respectively, in black, red and green at different translation rates. The predictions from Eq. (3.8) are shown as solid lines (their width corresponds to the 68% Confidence Interval). The predictions were made at constant $k_A^{(L)}$ rates equal to $k_A^{(L)} = 0.83 k_F^{(bulk)}$ (d, top panel), $k_A^{(L)} = 8.3 k_F^{(bulk)}$ (d, middle panel), and $k_A^{(L)} = 83 k_F^{(bulk)}$ (d, bottom panel). The continuous translation results from the coarse-grained model are shown as symbols at the various $k_A^{(L)}$ value; error bars correspond to the standard error about the mean.

Conclusions

In this work we presented an integrated approach to the study of protein translation, based on Statistical Physics. We adopted three different but complementary perspectives: building hypothesis up from the data, modeling down from reasonable assumptions, and using computer simulations when everything else fails.

In Chap. 1 we utilized the data from the mRNA sequences and analyzed how the different codons are used. We showed that the codon bias is not homogeneous across the transcriptome, and that some information is stored in the codon order. We compressed some of this information in an index, the CII, which turned out to be invariant under the reshuffling of the codons. By utilizing this observation we built an approximated, fully connected algorithm, whose results are compatible with the previous one. The latter formulation also makes explicit the dependence on the correlations between the codon frequencies across the genes. We emphasize that this new method implements a speedup of at least 3 orders of magnitude when compared to the previous formulation. At the leading order, the frequencies are the relevant information in the codon bias: spatial organization exists but is relevant at sub-leading orders and needs specifically tailored methods to be detected and decoded.

We showed in Chap. 2 that the problem of translation optimization is a global one due to the finite amount of resources available in the cell (ribosomes and tRNAs). Moreover, the fact that the contribution to the overall translation optimality of the cell is proportional to the abundance of the mRNA can qualitatively explain the origin of the correlation of CII with the mRNA abundance in Fig. 1.18, in the case that an optimal codon bias exists and the related evolutionary pressure is small.

We then progressed to include tRNA recharging dynamics and codon usage. Depending on the recharging rate of the tRNAs, the optimal usage of synonymous codons interpolates between the proportional rule (the tRNAs are used accordingly to their abundance, slow recharging) and the

single-tRNA rule (only the most abundant tRNA is used, fast recharging). However, since the codon usage cannot vary as fast as the environment, we suggest that the codon usage might be interpreted as a strategy to deal with the randomness of the environment.

The study of this model reveals that the translation mechanism might be very context dependent, as the translation speed of the codons is a function of the rate at which the different tRNAs are used. A great question to ask is therefore how does the translational machinery react when a highly expressed heterologous mRNA is induced, e.g., a plasmid in a *in vivo* experiment. The answer to this question is part of the solution to the long-posed challenge of the optimization of heterologous proteins expression.

Proceeding towards deeper levels of detail, we then focused on the single translation events. By using a stochastic and spacial model we studied how the low concentration of tRNA can lead to non-exponential deviations in the time between translation events (ribosome dwell time) when the timescales of diffusion, translation and recharge are comparable (and we give an argument supporting that physiological conditions can sample those ranges of parameters). We argue that the deviation is due to the coupling of the fluctuations between the charged and uncharged tRNAs. This mechanism can contribute to explain the spatial organization of the codons observed in Sec. 1.1.2, characterized in Ref. [19] (where it was observed that close occurrences of the same amino acid tend to have a higher-than-random probability to be encoded by codons translated by the same tRNA). When a charged tRNA reaches the ribosome, it produces a significant fluctuation in the concentration (which is on average very low). If the tRNA is quickly recharged after being used, re-using it can significantly reduce the waiting time.

The spacial organization of the codons can also affect the cotranslational folding of the proteins. The analysis of the implications of variable translation rates on this process, however, constitutes a great computational challenge. As a first step towards its solution we develop in Chap. 3 a Markov chain model which allows to simulate arbitrarily complex cotranslational folding mechanisms.

Let us conclude with a few notes about the codon usage, which after more than 30 years of study continues to pose a relevant puzzle. In this Thesis we identified two different sources of bias: a bias on the frequencies, and a bias in the spacial organization.

From the considerations in Chap. 2 the frequency bias seems most likely to be caused by tRNA and ribosome optimal utilization. A recent paper [92] suggests that the weighted codon usage (where the weights are the mRNA abundances) evolves toward a *global* proportionality rule (the

weighted codon usage is proportional to the abundance of the corresponding tRNA). The higher probability of fixing a beneficial mutation in highly expressed genes explains the codon usage/mRNA abundance correlation. The same model, however, fails to explain why similar correlations are found in the 2 codons, 1 tRNA amino acids.

This example is illustrative of the problem associated to the complexity of codon usage: several effects, whose relative importance is often difficult to estimate, are jointly at work, making it very difficult to disentangle the different causes.

The same can be replicated for the spacial organization bias, where, beyond protein cotranslational folding, a role is played by mRNA folding [75] and ribosome traffic [82, 97, 23].

A grand challenge in codon usage, and more generally in Biology, is to accurately determine what is the size of the effect that we are studying. This approach is essential for developing realistic model and for understanding these complex systems. The (over)simplified model that we presented in this Thesis dissect a small set of phenomena, allowing a complete enumeration of their consequences. Due to this sharp analytical power, this approach can be very helpful in disentangling different causes and consequences, and in discriminating the main effects from the corrections.

Appendix A

Learning from the data: the codon usage example

A.1 Details about the CII

A.1.1 Data sets

The set of RNA transcript for *S. Cerevisiae* were downloaded from [21]. Protein half-life were extracted form [8], mRNAs abundances were found in [60, 37]. Protein abundances were found in [45, 83].

A.1.2 Details of the algorithm

The thermodynamic averages (and thus the C_{II}) were computed using a Monte Carlo algorithm implemented with simulated annealing [67] and frequent reannilings: the temperature is a function of the simulation time and is slowly lowered. Provided that the cooling schedule is sufficiently slow, this method is guaranteed to sample the whole space, a vital feature if the free energy landscape is rough (meaning that the Hamiltonian has many metastable states).

The algorithm run time for the set of 3371 proteins was of the order of 1 day on a dual core workstation. The algorithm is massively parallelizable. A major speedup was obtained by observing that the energy differences of the information theoretical part of the Hamiltonian (1.13) (which are required in the Monte Carlo step) can be efficiently and locally computed using the properties of the Digamma function, see App. A.1.2. The results are available at the web page `http://www-vendruscolo.ch.cam.ac.uk/CII/index.php`.

Single spin flip Monte Carlo implementation

The energy difference associated to the flipping of a single spin can be obtained in a very simple and efficient way. Let us focus on a particular site where the spin is s^*, the amino acid a^*, and the codon $c^*_{a^*}$ are used. The single-spin flip move changes $s^* \to -s^*$, and the terms affected in the Hamiltonian are those involving $n_s(c^*_{a^*}, a^*)$ and N_{s,a^*}. These quantities change as follows:

$$\begin{aligned} n_s(c^*_{a^*}, a^*) &\to n_s(c^*_{a^*}, a^*) - ss^* \\ N_{s,a^*} &\to N_{s,a^*} - ss^*, \end{aligned}$$

and I transforms from

$$I_{|c^*_{a^*},a^*} = \sum_s \Big[n_s(c^*_{a^*}, a^*)\psi\big(n_s(c^*_{a^*}, a^*) + \alpha + 1\big) - N_{s,a^*}\psi\big(N_{s,a^*} + K_{a^*}\alpha + K_{a^*}\big)\Big]$$

to

$$I'_{|c^*_{a^*},a^*} = \sum_s \Big[\big(n_s(c^*_{a^*}, a^*) - ss^*\big)\psi\big(n_s(c^*_{a^*}, a^*) - ss^* + \alpha + 1\big) - \big(N_{s,a^*} - ss^*\big)\psi\big(N_{s,a^*} - ss^* + K_{a^*}\alpha + K_{a^*}\big)\Big].$$

The digamma function enjoys the recurrence property $\psi(z+1) = \psi(z) + \frac{1}{z}$, which can be adapted to our needs as

$$\psi(z - s) = \psi(z) - \frac{s}{z - \delta_{s,1}}, \tag{A.1}$$

with $s = \pm 1$. Using this property, we can write ΔI as

$$\begin{aligned} \Delta I = \sum_s \Big[& - ss^* \frac{n_s(c^*_{a^*}, a^*)}{n_s(c^*_{a^*}, a^*) + \alpha + 1 - \delta_{ss^*,1}} - ss^*\psi\big(n_s(c^*_{a^*}, a^*) + \alpha + 1\big) \\ & + \frac{1}{n_s(c^*_{a^*}, a^*) + \alpha + 1 - \delta_{ss^*,1}} + ss^* \frac{N_{s,a^*}}{N_{s,a^*} + K_{a^*}\alpha + K_{a^*} - \delta_{ss^*,1}} \\ & + ss^*\psi\big(N_{s,a^*} + K_{a^*}\alpha + K_{a^*} - \frac{1}{N_{s,a^*}} N_{s,a^*} + K_{a^*}\alpha + K_{a^*} - \delta_{ss^*,1}\big)\Big] \\ = \sum_s \Big[& ss^*\Big(\psi\big(N_{s,a^*} + K_{a^*}\alpha + K_{a^*} - \psi\big(n_s(c^*_{a^*}, a^*) + \alpha + 1\big)\Big) \\ & + \frac{ss^* N_{s,a^*} - 1}{N_{s,a^*} + K_{a^*}\alpha + K_{a^*} - \delta_{ss^*,1}} - \frac{ss^* n_s(c^*_{a^*}, a^*) - 1}{n_s(c^*_{a^*}, a^*) + \alpha + 1 - \delta_{ss^*,1}}\Big]. \end{aligned} \tag{A.2}$$

Finally, the values on the integers of ψ can be saved in a list, and the ΔI can be evaluated extremely efficiently.

A.1.3 Distance between $\{n_+(c)\}$ and $\{n_-(c)\}$ distributions

It is interesting to observe that the maxima of (1.11) also maximize the distance between the two distributions of codon usage defined by the $\{n_+(c)\}$ and $\{n_-(c)\}$. To show this, let's compute the symmetrized K-L divergence

$$D_\pm(\vec{s},\vec{c}) = \left\langle \log \frac{P\{\hat{p}|\vec{c},+\}}{P\{\hat{p}|\vec{c},-\}} \right\rangle_{P\{\hat{p}|\vec{c},+\}} + \left\langle \log \frac{P\{\hat{p}|\vec{c},-\}}{P\{\hat{p}|\vec{c},+\}} \right\rangle_{P\{\hat{p}|\vec{c},-\}}. \quad \text{(A.3)}$$

Performing the integration we have

$$D_\pm(\vec{s},\vec{c}) = \sum_{s=\pm 1}\sum_{c=1}^{K} \big(n_s(c) - n_{-s}(c)\big)\big[\psi\left(n_s(c)+1\right) - \psi\left(N_s+K\right)\big] + \text{Const.} \quad \text{(A.4)}$$

Since we are considering a distance, there must be a minimum where the distributions are equal, i.e. in $(n(1)/2,\ldots,n(K)/2)$, and we expect to find the maxima on the boundary (on those point which make the two distributions most different). Like in the case of (1.11), $D_\pm$ computed on the vertexes evaluates to

$$\begin{aligned} D_\pm(\vec{s}_{\min},\vec{c}) &= \sum_{c=1}^{K} n(c)\big[\psi\left(n(c)+1\right) - \psi\left(1\right)\big] \\ &\qquad - (N_+ - N_-)\big[\psi\left(N_+ + K\right) - \psi\left(N_- + K\right)\big] \\ &= \text{const} - (2N_+ - L)\big[\psi\left(N_+ + K\right) - \psi\left(L - N_+ + K\right)\big]. \end{aligned} \quad \text{(A.5)}$$

This function has a maximum for $N_+ = L/2$, meaning that the maximally distant configurations are those minimizing $|N_+ - N_-|$, as in the previous case.

A.2 Fully connected Ising model

In this section we briefly review the solution of the fully connected Ising model. The Hamiltonian (1.30) introduced in Sec. 1.3.2 has in fact this structure, and its thermodynamic description is important for the characterization of the CII itself.

The fully connected Ising model is an archetype among the mean field models. In fact, in this case the mean field description is exact in the thermodynamic limit. In this section we will sketch out the phase diagram of the model by showing that a phase transition between a paramagnetic and a ferromagnetic phase is present at $T = J$.

Let us define the Hamiltonian of the fully connected Ising model as

$$H = -\frac{J}{N}\sum_{j>i}\sigma_i\sigma_j - h\sum_i \sigma_i \tag{A.6}$$

where N is the number of spins. By using the per-spin magnetization $m = \frac{1}{N}\sum_{i=1}^{N}\sigma_i$, we can recast the Hamiltonian H in

$$H = -Nm\left(\frac{J}{2}m - h\right) + \frac{J}{2}, \tag{A.7}$$

The number of different states at fixed number $N_\uparrow$ of up spins (i.e., at a given magnetization) is $\Omega(N_\uparrow|N) = \binom{N}{N_\uparrow}$, and in terms of the magnetization m it is

$$\Omega(m) = \binom{N}{\frac{m+1}{2}N}. \tag{A.8}$$

We can now write the partition function of the system:

$$\begin{aligned} Z &= \sum_{\{\sigma_i\}} e^{-\beta H(\{\sigma_i\})} \\ &= \sum_m \Omega(m) e^{-\beta H(m)} = \sum_m e^{S-\beta H(m)} = \sum_m e^{-\beta F(m)}, \end{aligned} \tag{A.9}$$

where $S(m) = \log\Omega(m)$ is the entropy at magnetization m and $F(m) = H(m) - TS(m)$ is the free energy. The probability of a state having magnetization m is therefore

$$P(m) = \frac{e^{-\beta F(m)}}{Z}. \tag{A.10}$$

Let us use the Stirling approximation to write down the large-N expansion of the entropy:

$$S(m) \sim -N\left(\frac{1+m}{2}\log\frac{1+m}{2} + \frac{1-m}{2}\log\frac{1-m}{2}\right). \tag{A.11}$$

As both H ans S grow linearly with N, the saddle point approximation can be used: the probability (A.10) is dominated by the minima of $f(m) = F(m)/N$.

Let us consider $h = 0$: at large temperatures, the only minimum of $f(m)$ is at $m = 0$, as the entropy has a maximum there. However, two more minima appear when the temperature is lowered, as $f'(0)$ becomes negative when $T < J$ (and $f(\pm 1) = \pm\infty$). The system has a phase transition at $T_c = J$. Furthermore, if we expand $S(m)$ around $m = 0$, we obtain the following equation for the stationary points of $f(m)$:

$$m\left(-J + T(1 + m^2/3)\right) = 0, \tag{A.12}$$

which has the solutions $m = 0$ and $m = \pm m^*$ with $m^* = \sqrt{3\frac{J-T}{T}}$. By studying the second derivative, we observe that for $T > T_c$ the minimum is $m = 0$, while for $T < T_c$ there are two minima at $m = \pm m^*$.

Appendix B

Explaining the data: models for the systemic translation of the proteins

B.1 Time dependent solution of Eq. (2.19) and two-points correlators

In this appendix we solve the time-dependent master equation, Eq. (2.19), in order to characterize the relaxation dynamics of the model toward the stationary state. Moreover, by using the properties of the characteristic function, we are able to compute the different-time correlators between n and m.

In order to simplify the notation, let us first introduce the quantities

$$\begin{aligned} E &= e^{-t(\lambda+\rho+1)}, \\ F &= e^{-\rho t}. \end{aligned} \tag{B.1}$$

The solution of the differential equation for the generating function $G(z,w;t)$ associated to Eq. (2.19) can be easily obtained with the method of characteristics. Using the initial condition $p_{n,m}(0) = \delta_{n,n_0}\delta_{m,m_0}$, we have

$$\begin{aligned} G(z,w;t) = \exp\left[\bar{N}\left(\frac{w+\lambda z+\rho[zX+w(1-X)]}{\lambda+\rho+1} - 1 + \left(1-\frac{w+\lambda z}{\lambda+1}\right)F\right.\right. \\ \left.\left. + \frac{\rho(w-z)[X(\lambda+1)-\lambda]}{(\lambda+1)(\lambda+\rho+1)}E\right)\right]\left(1+\frac{z-w}{\lambda+1}E+\frac{w-1+\lambda(z-1)}{\lambda+1}F\right)^{n_0} \\ \times\left(1+\frac{\lambda(w-z)}{\lambda+1}E+\frac{w-1+\lambda(z-1)}{\lambda+1}F\right)^{m_0}. \end{aligned} \tag{B.2}$$

By differentiation, $p_{n,m}(t) = \left[\frac{(\partial_z)^n}{n!}\frac{(\partial_w)^m}{m!}G(z,w;t)\right]_{z=w=0}$, we obtain the joint time-dependent probability distribution for m and n:

$$\begin{aligned} p_{n,m}(t) = &\frac{e^{\bar{N}(F-1)}}{n!m!(\lambda+1)^{n+m}}\sum_{r=0}^{n}\sum_{s=0}^{m}A^{n-r}B^{m-s}\sum_{j=0}^{r}\sum_{k=0}^{s}\binom{r}{j}\binom{s}{k} \\ &\times\frac{n_0!m_0!\,\lambda^j(E+\lambda F)^{r-j}(\lambda E+F)^k(F-E)^{s+j-k}}{(n_0-r+j-s+k)!(m_0-j-k)!}(1-F)^{n_0+m_0-r-s} \\ &\times\theta(n_0-r+j-s+k)\,\theta(m_0-j+k), \end{aligned} \tag{B.3}$$

where $\theta(x)$ is the Heaviside step function and

$$\begin{aligned} A &= (\lambda+1)\bar{n} - \frac{\bar{N}\rho[X(\lambda+1)-\lambda]}{\lambda+\rho+1}E - \bar{N}\lambda F, \\ B &= (\lambda+1)\bar{m} + \frac{\bar{N}\rho[X(\lambda+1)-\lambda]}{\lambda+\rho+1}E - \bar{N}F. \end{aligned} \tag{B.4}$$

For generic initial conditions we can write a large-t (i.e., a small E, F) expansion, in order to see how $p_{n,m}(t)$ relaxes to the stationary value $p^{\rm st}_{n,m}$, Eq. (2.22):

$$p_{n,m}(t) = p^{\rm st}_{n,m}\left[1+\alpha E+\beta F+O(E^2,F^2,EF)\right], \tag{B.5}$$

where

$$\begin{aligned} \alpha =&\frac{\bar{N}\rho[X(\lambda+1)-\lambda]}{(\lambda+1)(\lambda+\rho+1)}\left(\frac{m}{\bar{m}}-\frac{n}{\bar{n}}\right) \\ &+\frac{n_0\theta(n_0-1)-\lambda m_0\theta(m_0-1)}{\lambda+1}\left(\frac{1}{\bar{n}}-\frac{1}{\bar{m}}\right), \\ \beta =&\bar{N}-(n_0+m_0)-\frac{\bar{N}}{\lambda+1}\left(\frac{\lambda n}{\bar{n}}+\frac{m}{\bar{m}}\right) \\ &+\frac{n_0\theta(n_0-1)+m_0\theta(m_0-1)}{\lambda+1}\left(\frac{\lambda}{\bar{n}}+\frac{1}{\bar{m}}\right). \end{aligned} \tag{B.6}$$

The leading term for large times is therefore associated with E and F, i.e., with the relaxation-times:

$$\begin{aligned} t_1 &= \frac{1}{\rho}, \\ t_2 &= \frac{1}{\lambda+\rho+1}. \end{aligned} \tag{B.7}$$

The first time scale t_1 is associated with the diffusion process, while the second one is the inverse of the sum of all rates (if we restore the λ_T dependence we have $1/t_2 = \lambda_R+\rho+\lambda_T$).

Given the analytic expression of the generating function $G(z,w;t)$ in Eq. (B.2), it is straightforward to evaluate the correlators, for instance:

$$\begin{aligned}\langle n(0)m(t)\rangle &= \sum_{n_0,m_0}\sum_{n,m} n_0 m\, p_{n_0,m_0}\left[p_{n,m}(t)\right]_{\substack{n(0)=n_0\\ m(0)=m_0}} \\ &= \sum_{n_0,m_0} n_0\, p_{n_0,m_0}\left[\partial_w G(z,w;t|n_0,m_0)\right]_{\substack{z=1\\ w=1}}.\end{aligned}$$

In particular, we find:

$$\begin{aligned}C_{nn}(t) &\equiv \langle n(0)n(t)\rangle - \bar{n}^2 = \bar{n}\frac{\lambda e^{-\rho t} + e^{-t(\lambda+\rho+1)}}{\lambda+1},\\ C_{mm}(t) &\equiv \langle m(0)m(t)\rangle - \bar{m}^2 = \bar{m}\frac{e^{-\rho t} + \lambda e^{-t(\lambda+\rho+1)}}{\lambda+1},\\ C_{nm}(t) &\equiv \langle n(0)m(t)\rangle - \bar{n}\bar{m} = \bar{n}\frac{e^{-\rho t} - e^{-t(\lambda+\rho+1)}}{\lambda+1},\\ C_{mn}(t) &\equiv \langle m(0)n(t)\rangle - \bar{n}\bar{m} = \bar{m}\lambda\frac{e^{-\rho t} - e^{-t(\lambda+\rho+1)}}{\lambda+1}.\end{aligned} \tag{B.8}$$

The first two correlators in (B.8) are monotonically decreasing, as they are linear combinations (with positive coefficients) of the exponentials characterized by the decay times of Eq. (B.7).

The other two correlators are again linear combinations of the same exponentials, but the coefficients of such linear combination have different signs, which makes them non-monotonic. The maximum is at time

$$t_{\max} = \frac{1}{\lambda+1}\log\left[\frac{\lambda+\rho+1}{\rho}\right]. \tag{B.9}$$

B.2 Violation of detailed balance

The model can be interpreted as a random walk on the two dimensional (n,m) lattice, with site- and direction-dependent transition rates (Fig. B.1).

We use this analogy to check for eventual violation of detailed balance (DB) in the stationary state described by Eq. (2.22). As it can be seen from Fig. B.1, there are three directions for the single step jumps:

- Along the vertical direction $(n,m) \leftrightarrow (n,m+1)$ the DB condition is

$$(1-X)\rho\bar{N}\, p_{n,m} = (m+1)\rho\, p_{n,m+1}, \tag{B.10}$$

  which is satisfied if

$$\rho = 0 \quad \text{or} \quad \rho \to \infty \quad \text{or} \quad \lambda = \frac{X}{1-X}. \tag{B.11}$$

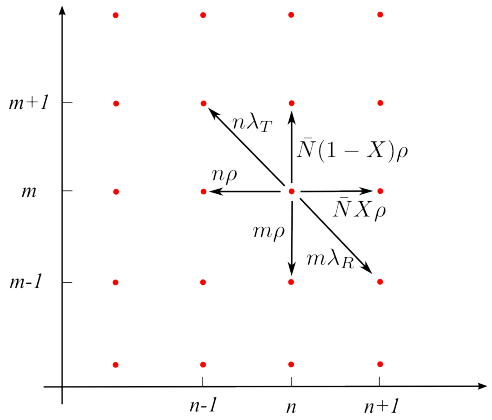

Figure B.1: The model as a random walk on the (n, m) lattice. The transition rates from one site to another depend on the direction and on the position on the lattice.

- Along the horizontal direction $(n, m) \leftrightarrow (n+1, m)$ the DB condition is

$$X\bar{N}\rho\, p_{n,m} = (n+1)\rho\, p_{n+1,m}. \tag{B.12}$$

  Again, its solution is Eq. (B.11).

- Along the diagonal direction $(n, m) \leftrightarrow (n-1, m+1)$ the DB condition is

$$n\, p_{n,m} = \lambda(m+1)\, p_{n-1,m+1}, \tag{B.13}$$

  whose solution is

$$\rho = 0 \qquad \text{or} \qquad \lambda = \frac{X}{1-X}. \tag{B.14}$$

We conclude that the only values of the parameters satisfying DB are those in Eq. (B.14), while for $\rho \to \infty$ DB is "almost satisfied", being the violation vanishingly small.

The first value of Eq. (B.14) coincides with the trivial case where diffusion is suppressed, while the second one is the value where the stationary points of the internal (recharge and translation) and diffusive dynamics coincide.

In all other cases there are current probability loops and the stationary state is out of equilibrium [127].

B.3 Relation between the CTTD and the average number of charged tRNAs

Let us consider the time dependent average of the number of charged tRNAs $\langle n(t)\rangle_{\rm nt} = \sum_{n=0}^{\infty} n\, p_n(t|\text{no translation})$, where at $t = 0^-$ a translation event occurred and no translation events were recorded between 0 and t.

In discrete time with temporal step δt, we can write the probability that a translation event happens at time $t > \tau_k = k\delta t$ as

$$P(\tau_k) = \prod_{i=0}^{k}(1 - \langle n(\tau_i)\rangle_{\rm nt}\delta t) \sim \exp\left(-\sum_{i=0}^{k}\langle n(\tau_i)\rangle_{\rm nt}\delta t\right). \tag{B.15}$$

In the continuous time limit we have

$$P(\tau) = \exp\left(-\int_0^{\tau} dt\, \langle n(t)\rangle_{\rm nt}\right). \tag{B.16}$$

Interestingly, $P(\tau)$ is exponential as long as $\langle n(t)\rangle_{\rm nt}$ is independent from t. Using Eq. (B.16), we can express the CTTD as:

$$p(t) = \langle n(t)\rangle_{\rm nt} \exp\left(-\int_0^{t} dt'\, \langle n(t')\rangle_{\rm nt}\right), \tag{B.17}$$

and inverting Eq. (B.16), we can write

$$\langle n(t)\rangle_{\rm nt} = -\partial_\tau \log P(\tau)_{|\tau=t}. \tag{B.18}$$

This relation is utilized in Fig. 2.6 to plot the theoretical predictions.

Appendix C

Simulating the data: the effect of codon translation rates on cotranslational folding

The predictions we make in Chap. 3 utilize the data obtained by a specifically tailored set of simulations, whose scope is to measure the transition rates between the possible folded states of the nascent chain at length L. This was accomplished by simulating the thermodynamic evolution of the protein tethered to the ribosome at a fixed length ("arrested ribosome"). In the following section we describe a quick and dirty method to estimate the rates given a timeseries of the protein evolution.

C.1 Extracting the interconversion rates from the arrested ribosome simulations timeseries

The output of the coarse grained, arrested translation simulations is a time series capturing the the state of the system at time intervals δt. Let us suppose that I) the protein we are studying can transition between K states, and that II) we simulated its thermodynamic evolution, observing the state occupied by the protein every δt for T times. The data is therefore the time series $\vec{X} = (x_1, x_2, \ldots, x_T)$ with $x_i \in \{1, \ldots, K\}$.

From this timeseries we can obtain the transition matrix N whose entries $n_{i\to j}$ contain the number of times a transition from state i to state j is observed. The elements on the diagonal $n_{i\to i}$ are the number of times no

transition was observed.

From the counts we can obtain the empirical probability of transitioning from i to j in a time interval δt:

$$\hat{p}_{i\to j}(\delta t) = \frac{n_{i\to j}}{\sum_j n_{i\to j}}.$$

Supposing the underlying process to be markovian, the time between the transition events will be exponentially distributed, and the probability $p_{i\to i}(\delta t)$ that no transition occurs before δt is

$$p_{i\to i}(\delta t) = \mathrm{Prob}\big(t_{i\to j} > dt,\, \forall_{j\neq i}\big) = \prod_{j(\neq i)} e^{-k_{ij}\delta t}, \tag{C.1}$$

where k_{ij} is the transition rate from state i to j. In the derivation of this equation we assumed that the timescale of the transitions, k_{ij}^{-1}, is larger than δt, i.e., that the probability of the 2-step walks $i \to j \to i$ is negligibly small.

The probability $p_{i\to j}(\delta t)$ of transitioning from i to j can be written as the complement of the previous probability times the probability $q_{i,j} = \frac{k_{ij}}{\sum_l k_{il}}$ of transitioning towards the j-th state:

$$p_{i\to j}(\delta t) = (1 - p_{i\to i})q_{ij} = \frac{k_{ij}}{\sum_l k_{il}}(1 - \prod_{l(\neq i)} e^{-k_{il}\delta t}). \tag{C.2}$$

By using Eqs. (C.1) and (C.2), we can write the relations for the rates

$$\begin{aligned} S &\equiv \sum_{j\neq i} k_{ij} = -\frac{\log p_{i\to i}}{\delta t}, \\ k_{ij} &= \frac{S}{1 - p_{i\to i}} p_{i\to j}, \end{aligned} \tag{C.3}$$

and replacing the probabilities with their empirical estimates, we finally have the estimated rates

$$\hat{k}_{ij} = -\frac{\log \frac{n_{i\to i}}{\sum_k n_{i\to k}}}{\delta t \sum_{l\neq i} n_{i\to l}} n_{i\to j}. \tag{C.4}$$

C.2 Simulation protocols

Simulations were carried out in CHARMM [15] (version c35b5) using Langevin dynamics with a 15 fs integration time step. The states were identified by the fraction of native contacts between the helices: as the free energy plot Fig. C.1, these are clearly identifiable.

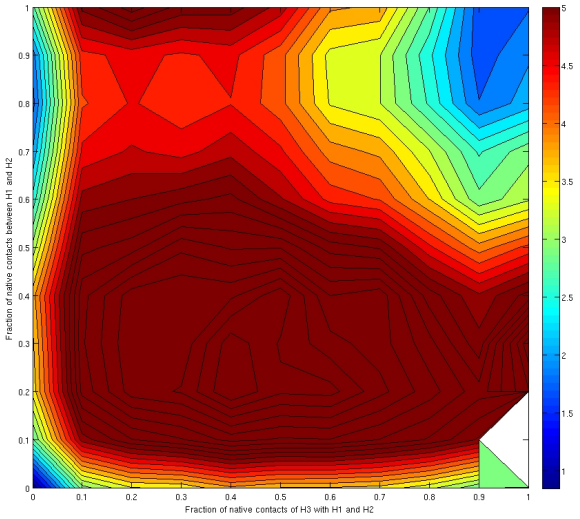

Figure C.1: Free energy surface as a function of the fraction of native contacts between the three helices of protein MIT. The three state structure is clearly identifiable, with the minima on the bottom left (unfolded), top left (partially folded), and top right (folded).

The arrested ribosome simulations where performed at a fixed length of the nascent protein. The timeseries of the states at chainlength L where analyzed as in App. C.1 in order to obtain the rates $k_{ij}^{(L)}$ of interconversion between the states.

The second set of simulations was performed in continuous translation: at exponentially distributed time intervals (determined by the rate $k_A^{(L)}$) a new amino acid was added to the protein and the state recorded. A full description of the protocol can be found in Ref. [85].

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, 2002.

[2] S. Andersson and C. Kurland. Codon preferences in free-living microorganisms. *Microbiological Reviews*, 54(2):198–210, 1990.

[3] E. Angov, C. J. Hillier, R. L. Kincaid, and J. A. Lyon. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS One*, 3(5):e2189, 2008.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[5] A. Ashkin. Optical trapping and manipulation of neutral particles using lasers. *Proc. Natl. Acad. Sci. U.S.A.*, 94(10):4853–4860, 1997.

[6] M. Bailly-Bechet, A. Danchin, M. Iqbal, M. Marsili, and M. Vergassola. Codon usage domains over bacterial chromosomes. *PLoS Comput. Biol.*, 2(4):e37, Apr. 2006.

[7] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, et al. The pfam protein families database. *Nucleic Acids Res.*, 32(suppl 1):D138–D141, 2004.

[8] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea. Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. U.S.A.*, 103(35):13004–9, Aug. 2006.

[9] J. L. Bennetzen and B. D. Hall. Codon selection in yeast. *J. Biol. Chem.*, 257(6):3026–3031, 1982.

[10] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res.*, 38(suppl 1):D46–D51, 2010.

[11] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res.*, 39(suppl 1):D32–D37, 2011.

[12] A. Borgia, P. M. Williams, and J. Clarke. Single-molecule studies of protein folding. *Annu. Rev. Biochem.*, 77:101–125, 2008.

[13] G. A. Brar, M. Yassour, N. Friedman, A. Regev, N. T. Ingolia, and J. S. Weissman. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science*, 335(6068):552–557, 2012.

[14] H. Bremer, P. P. Dennis, et al. Modulation of chemical composition and other parameters of the cell by growth rate. In F. C. Neidhardt, editor, *Escherichia coli and Salmonella: cellular and molecular biology*, volume 2, pages 1553–1569. ASM press, Washington DC, 1996.

[15] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[16] T. A. Brown. *Genomes*. Garland Publishing Inc, 2006.

[17] M. Bulmer. Coevolution of codon usage and transfer rna abundance. *Nature*, 325(6106):728–730, Feb. 1987.

[18] M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3):897–907, 1991.

[19] G. Cannarozzi, N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral. A role for codon order in translation dynamics. *Cell*, 141(2):355–67, 2010.

[20] P. P. Chan and T. M. Lowe. Gtrnadb: a database of transfer rna genes detected in genomic sequence. *Nucleic Acids Res.*, 37(suppl 1):D93–D97, 2009.

[21] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong. Saccharomyces genome database: the genomics

resource of budding yeast. *Nucleic Acids Res.*, 40(D1):D700–D705, 2012.

[22] T. Chou, K. Mallick, and R. K. P. Zia. Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport. *Rep. on Prog. Phys.*, 74(11):116601, 2011.

[23] L. Ciandrini, I. Stansfield, and M. C. Romano. Ribosome traffic on mrnas maps to gene ontology: Genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS Comput. Biol.*, 9(1):e1002866, 2013.

[24] P. Ciryam, R. I. Morimoto, M. Vendruscolo, C. M. Dobson, and E. P. OBrien. In vivo translation rates can substantially delay the cotranslational folding of the escherichia coli cytosolic proteome. *Proc. Natl. Acad. Sci. U. S. A.*, 110(2):E132–E140, 2013.

[25] P. L. Clark and J. King. A newly synthesized, ribosome-bound polypeptide chain adopts conformations dissimilar from early in vitrorefolding intermediates. *J. Biol. Chem.*, 276(27):25411–25420, 2001.

[26] T. Clarke and P. Clark. Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. *BMC Genomics*, 11(1):118, 2010.

[27] J. R. Coleman, D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, and S. Mueller. Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884):1784–1787, 2008.

[28] P. Cortazzo, C. Cerveñansky, M. Marın, C. Reiss, R. Ehrlich, and A. Deana. Silent mutations affect in vivo protein folding in¡ i¿ escherichia coli¡/i¿. *Biochem. Biophys. Res. Commun.*, 293(1):537–541, 2002.

[29] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[30] J. F. Curran and M. Yarus. Rates of aminoacyl-trna selection at 29 sense codons in vivo. *J. Mol. Biol.*, 209(1):65 – 77, 1989.

[31] C. M. Deane and R. Saunders. The imprint of codons on protein structure. *Biotechnol. J.*, 6(6):641–649, 2011.

[32] K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[33] K. A. Dittmar, M. A. Sörensen, J. Elf, M. n. Ehrenberg, and T. Pan. Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep.*, 6(2):151–7, 2005.

[34] T. Dobzhansky. Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, 35(3):125–129, 1973.

[35] H. Dong, L. Nilsson, and C. G. Kurland. Co-variation of trna abundance and codon usage inescherichia coliat different growth rates. *J. Mol. Biol.*, 260(5):649 – 663, 1996.

[36] D. A. Drummond and C. O. Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52, July 2008.

[37] A. M. Dudley, J. Aach, M. A. Steffen, and G. M. Church. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl. Acad. Sci. U. S. A.*, 99(11):7554–7559, 2002.

[38] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, 2002.

[39] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[40] R. Farber, A. Lapedes, and K. Sirotkin. Determination of eukaryotic protein coding regions using neural networks and information theory. *J. Mol. Biol.*, 226(2):471 – 479, 1992.

[41] W. Feller. *An introduction to probability theory and its applications.* Wiley, 1968.

[42] A. Fluitt, E. Pienaar, and H. Viljoen. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput. Biol. Chem.*, 31:335–346, 2007.

[43] T. Funatsu, Y. Harada, H. Higuchi, M. Tokunaga, K. Saito, Y. Ishii, R. D. Vale, and T. Yanagida. Imaging and nano-manipulation of single biomolecules. *Biophys. Chem.*, 68(1):63–72, 1997.

[44] J. J. Gartner, S. C. Parker, T. D. Prickett, K. Dutton-Regester, M. L. Stitzel, J. C. Lin, S. Davis, V. L. Simhadri, S. Jha, N. Katagiri, et al. Whole-genome sequencing identifies a recurrent functional

synonymous mutation in melanoma. *Proc. Natl. Acad. Sci. U.S.A.*, 110(33):13481–13486, 2013.

[45] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–41, Oct. 2003.

[46] G. Gibson. Rare and common variants: twenty arguments. *Nat. Rev. Genet.*, 13(2):135–145, Feb. 2012.

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

[48] M. Gouy and C. Gautier. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, 10(22):7055–7074, 1982.

[49] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.*, 9(1):213, 1981.

[50] P. Greulich, L. Ciandrini, R. J. Allen, and M. C. Romano. Mixed population of competing totally asymmetric simple exclusion processes with a shared reservoir of particles. *Phys. Rev. E*, 85:011142, 2012.

[51] J. Grilli, B. Bassetti, S. Maslov, and M. C. Lagomarsino. Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. *Nucleic Acids Res.*, 40(2):530–540, 2012.

[52] A. N. Gupta, A. Vincent, K. Neupane, H. Yu, F. Wang, and M. T. Woodside. Experimental validation of free-energy-landscape reconstruction from non-equilibrium single-molecule force spectroscopy measurements. *Nat. Phys.*, 7(8):631–634, 2011.

[53] C. Gustafsson, S. Govindarajan, and J. Minshull. Codon bias and heterologous protein expression. *Trends Biotechnol.*, 22(7):346 – 353, 2004.

[54] G. A. Gutman and G. W. Hatfield. Nonrandom utilization of codon pairs in escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.*, 86(10):3699–3703, 1989.

[55] T. Ha, T. Enderle, D. Ogletree, D. S. Chemla, P. R. Selvin, and S. Weiss. Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. U.S.A.*, 93(13):6264–6268, 1996.

[56] N. C. Harris, Y. Song, and C.-H. Kiang. Experimental free energy surface reconstruction from single-molecule force spectroscopy using jarzynski's equality. *Phys. Rev. Lett.*, 99:068101, 2007.

[57] F. U. Hartl and M. Hayer-Hartl. Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.*, 16(6):574–581, 2009.

[58] D. R. Hekstra and S. Leibler. Contingency and statistical laws in replicate microbial closed ecosystems. *Cell*, 149(5):1164–1173, 2012.

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

[60] F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717 – 728, 1998.

[61] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proc. Natl. Acad. Sci. U.S.A.*, 98(7):3658–3661, 2001.

[62] T. Ikemura. Correlation between the abundance of escherichia coli transfer rnas and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the e. coli translational system. *J. Mol. Biol.*, 151(3):389 – 409, 1981.

[63] T. Ikemura. Codon usage and trna content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, 2(1):13–34, 1985.

[64] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, 2000.

[65] J. Kemeny and J. Snell. *Finite Markov chains*. Springer-Verlag, New York, 1983.

[66] C. Kimchi-Sarfaty, J. M. Oh, I.-W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman. A "silent" polymorphism in the mdr1 gene changes substrate specificity. *Science*, 315(5811):525–528, 2007.

[67] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[68] D. Klimov and D. Thirumalai. Viscosity dependence of the folding rates of proteins. *Phys. Rev. Lett.*, 79(2):317, 1997.

[69] S. Klumpp, J. Dong, and T. Hwa. On ribosome load, codon bias and protein abundance. *PLoS One*, 7(11):e48542, 2012.

[70] A. a. Komar. A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.*, 34(1):16–24, Jan. 2009.

[71] A. A. Komar, T. Lesnik, and C. Reiss. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.*, 462(3):387–391, 1999.

[72] E. V. Koonin. Are there laws of genome evolution? *PLoS Comput. Biol.*, 7(8):e1002173, 2011.

[73] A. Kozomara and S. Griffiths-Jones. mirbase: integrating microrna annotation and deep-sequencing data. *Nucleic Acids Res.*, 39(suppl 1):D152–D157, 2011.

[74] H. E. Kubitschek. Cell volume increase in escherichia coli after shifts to richer media. *J. Bacteriol.*, 172(1):94–101, 1990.

[75] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin. Coding-sequence determinants of gene expression in escherichia coli. *Science*, 324(5924):255–258, 2009.

[76] C. Levinthal. How to fold graciously. In P. Debrunner, J. C. M. Tsibris, and E. Münck, editors, *Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Allerton House, Monticello, IL*, 1969.

[77] G.-W. Li and X. S. Xie. Central dogma at the single-molecule level in living cells. *Nature*, 475(7356):308–315, 2011.

[78] S.-T. Liang, Y.-C. Xu, P. Dennis, and H. Bremer. mrna composition and control of bacterial gene expression. *J. Bacteriol.*, 182(11):3037–3044, 2000.

[79] H. Liljenström, G. Heijne, C. Blomberg, and J. Johansson. The trna cycle and its relation to the rate of protein synthesis. *Eur. Biophys. J.*, 12(2):115–119, 1985.

[80] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W.

Paslay, et al. Impact of high-throughput screening in biomedical research. *Nature reviews Drug discovery*, 10(3):188–195, 2011.

[81] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, 9(5):356–369, May 2008.

[82] N. Mitarai, K. Sneppen, and S. Pedersen. Ribosome collisions and translation efficiency: Optimization by codon usage and mrna destabilization. *J. Mol. Bio.*, 382(1):236 – 245, 2008.

[83] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–6, June 2006.

[84] J. Ninio. Alternative to the steady-state method: derivation of reaction rates from first-passage times and pathway probabilities. *Proc. Natl. Acad. Sci. U.S.A.*, 84(3):663–667, 1987.

[85] E. P. O'Brien, M. Vendruscolo, and C. M. Dobson. Prediction of variable translation rate effects on cotranslational protein folding. *Nat. Commun.*, 3:868–, 2012.

[86] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 27(1):29–34, 1999.

[87] I. Pagani, K. Liolios, J. Jansson, I.-M. A. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. The genomes online database (gold) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, 40(D1):D571–D579, 2012.

[88] S. Pechmann and J. Frydman. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.*, 20(2):237–243, 2013.

[89] R. Percudani, a. Pavesi, and S. Ottonello. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J. Mol. Biol.*, 268(2):322–30, May 1997.

[90] T. Platini. Measure of the violation of the detailed balance criterion: A possible definition of a "distance" from equilibrium. *Phys. Rev. E*, 83:011119, 2011.

[91] J. B. Plotkin and G. Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, 12(1):32–42, Jan. 2011.

[92] W. Qian, J.-R. Yang, N. M. Pearson, C. Maclean, and J. Zhang. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.*, 8(3):e1002603, 2012.

[93] S. E. Radford, C. M. Dobson, and P. A. Evans. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature*, 358(6384):302–307, 1992.

[94] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[95] S. Redner. *A guide to first-passage processes*. Cambridge University Press, 2001.

[96] M. d. Reis, R. Savva, and L. Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, 32(17):5036–5044, 2004.

[97] S. Reuveni, I. Meilijson, M. Kupiec, E. Ruppin, and T. Tuller. Genome-scale analysis of translation elongation with a ribosome flow model. *PLoS Comput. Biol.*, 7(9):e1002127, 2011.

[98] F. Ritort. Single-molecule experiments in biological physics: methods and applications. *J. Phys.: Condens. Mat.*, 18(32):R531, 2006.

[99] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[100] R. Saunders and C. M. Deane. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.*, 38(19):6719–6728, 2010.

[101] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa. Interdependence of cell growth and gene expression: origins and consequences. *Science*, 330(6007):1099–1102, 2010.

[102] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

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

[104] J. Shendure and H. Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[105] E. Siller, D. C. DeZwaan, J. F. Anderson, B. C. Freeman, and J. M. Barral. Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J. Mol. Biol.*, 396(5):1310–1318, 2010.

[106] M. Sotomayor and K. Schulten. Single-molecule experiments in vitro and in silico. *Science*, 316(5828):1144–1148, 2007.

[107] P. S. Spencer, E. Siller, J. F. Anderson, and J. M. Barral. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *J. Mol. Biol.*, 422(3):328–335, 2012.

[108] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. G. M. de Visser. Quantitative analyses of empirical fitness landscapes. *J. Stat. Mech: Theory Exp.*, 2013(01):P01005, 2013.

[109] M. A. Srensen, C. Kurland, and S. Pedersen. Codon usage determines translation rate in escherichia coli. *J. Mol. Biol.*, 207(2):365 – 377, 1989.

[110] T. Thanaraj and P. Argos. Ribosome-mediated translational pause and protein domain organization. *Protein Sci.*, 5(8):1594–1612, 1996.

[111] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815–, Dec. 2000.

[112] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–54, Apr. 2010.

[113] T. Tuller, M. Kupiec, and E. Ruppin. Determinants of protein abundance and translation efficiency in S. cerevisiae. *PLoS Comput. Biol.*, 3(12):e248, Dec. 2007.

[114] T. Tuller, Y. Y. Waldman, M. Kupiec, and E. Ruppin. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.*, 107(8):3645–50, Feb. 2010.

[115] S. Uemura, C. E. Aitken, J. Korlach, B. A. Flusberg, S. W. Turner, and J. D. Puglisi. Real-time trna transit on single translating ribosomes at codon resolution. *Nature*, 464(7291):1012–1017, 2010.

[116] G. Van Den Bogaart, N. Hermans, V. Krasnikov, and B. Poolman. Protein mobility and diffusive barriers in escherichiacoli: consequences of osmotic stress. *Mol. Microbiol.*, 64(3):858–871, 2007.

[117] E. van Nimwegen. Scaling laws in the functional content of genomes. *Trends in Genetics*, 19(9):479 – 484, 2003.

[118] S. Varenne, J. Buc, R. Lloubes, and C. Lazdunski. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J. Mol. Biol.*, 180(3):549–76, 1984.

[119] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.*, 106(1):67–72, 2009.

[120] G. B. West and J. H. Brown. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *Journal of Experimental Biology*, 208(9):1575–1592, 2005.

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

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

[123] G. Zhang, M. Hubalewska, and Z. Ignatova. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.*, 16(3):274–80, Mar. 2009.

[124] M. Zhou, J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. S. Sachs, and Y. Liu. Non-optimal codon usage affects expression, structure and function of clock protein frq. *Nature*, Advance online publication:–, 2013.

[125] J. Zhu, P. A. Penczek, R. Schrder, and J. Frank. Three-dimensional reconstruction with contrast transfer function correction from energy-filtered cryoelectron micrographs: Procedure and application to the 70s e. coli ribosome. *J. Struct. Biol.*, 118(3):197 – 219, 1997.

[126] R. K. P. Zia and B. Schmittmann. A possible classification of nonequilibrium steady states. *J. Phys. A: Math. Gen.*, 39(24):L407, 2006.

[127] R. K. P. Zia and B. Schmittmann. Probability currents as principal characteristics in the statistical mechanics of non-equilibrium steady states. *J. Stat. Mech. Theor. Exp.*, 2007(07):P07012, 2007.

[128] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. *Proc. Natl. Acad. Sci. U.S.A.*, 89(1):20–22, 1992.