Freedom of additional signals on genes: on the combination of DNA mechanics, genetics and translation speed

Zuiddam, M.

Chapter 4

Multiplexing mechanical and translational cues on genes

This chapter is based on a manuscript by Zuiddam, Shakiba and Schiessel.

In the previous chapter we have demonstrated that the degeneracy of the genetic code can be used to create positioning signals on virtually any position on any gene of yeast. By doing so, we have demonstrated the huge extent to which additional information can be placed on top of protein-coding DNA sequences. This can be done using the degeneracy of the genetic code (18 out of 20 amino acids are encoded for by more than one codon). The purpose of this chapter is to show that it is possible to carry more than one additional layer of information on top of a gene. In particular, we show how much translation speed and nucleosome positioning can be adjusted simultaneously without changing the encoded protein. We again utilize the technique we introduced in Chapter 3, which maps genes on weighted graphs that contain all synonymous genes and then finds shortest paths through these graphs. We include translation speed in the analysis by either pruning graphs or incorporating the speed in the weights of the graphs. This enables us e.g. to readjust the translational speed profile after it has been disrupted when a gene has been introduced from one organism (e.g. human) into another (e.g. yeast) without greatly changing the nucleosome landscape intrinsically encoded by the DNA molecule.

4.1 Introduction

As early as 1989 it was suggested by Edward N. Trifonov that DNA could carry several codes in addition to the classical genetic code [74]. In particular, he mentioned a translation framing code (an excess of G in the first codon position), a chromatin code (caused by curved DNA) and a putative loop code (so as not to allow RNA secondary structure). In addition, overlapping genes were mentioned. Typically, however, the various scientific communities focus only on one additional layer of information. To give two examples: there exists a large body of work on DNA mechanics and geometry and how they influence the positioning of nucleosomes along

DNA (mentioned in Ref. [74] as chromatin code) and another large body of work on the translational speed in ribosomes and how it affects co-translational folding. The question remains, however, to which extent such different codes can really co-exist on top of one another. This chapter answers this question using the examples of nucleosome positioning and translation speed. We first introduce nucleosomes and their positioning before discussing translation speed and its role in co-translational folding.

4.1.1 Introduction to nucleosome positioning

The nucleosome is the repeated basic structure in chromatin. It is a stretch of DNA with a length of 147 base pairs (bp) wound 1 and 3/4 turns around a cylindrical aggregate made up of eight histone proteins [33]. The resulting disk-like complex is connected to the next such DNA spool by a short stretch of linker DNA. Notably, the wrapping length in the nucleosome is close to the DNA persistence length of about 150 bp or 50 nm. Bending a persistence length of DNA nearly two turns is quite expensive. Furthermore, the free energy of bending depends on the bp sequence, which reflects the fact that the geometry and elasticity of the DNA double helix depends on sequence [32]. This enormous sequence-dependent bending cost is compensated by the binding of the DNA molecule to the histone octamer at 14 binding sites [33]. The binding is mainly to the DNA backbones, the chemistry of which is not dependent on the sequence. Taken together, this suggests that the affinity of a given DNA sequence to be part of a nucleosome compared to another sequence is directly related to differences in the sequence-dependent bending costs. This makes it possible to write mechanical cues along DNA molecules to direct nucleosomes to occupy or to avoid certain positions. This has been referred to as the "nucleosome positioning code" [5] (for earlier versions of this idea see e.g. Refs. [75] and [4], and for a review see [11].)

After reconstituting nucleosomes from DNA and histone proteins using salt dialysis, position preferences of nucleosomes along genomic DNA can be clearly observed. By creating nucleosome maps using genome-wide assays that extract DNA stretches that were stably wrapped in nucleosomes (see e.g. [6]), one gets the nucleosome occupancy at each bp position, which is the probability that the corresponding bp is covered by a nucleosome. Two types of nucleosome positioning along DNA are found: rotational and translational positioning [7]. Rotational positioning mainly reflects the fact that a given DNA stretch is typically not inherently straight because of the intrinsic geometries of the bp steps involved. Nucleosomes therefore prefer positions where the DNA is pre-bent in the wrapping direction, resulting in sets of positions 10 bp (the DNA helical repeat) apart. The specific bp rules for nucleosome positioning are typically formulated in terms of dinucleotides; rotationally positioned nucleosomes have an increased probability to feature GC steps (nucleotide G followed by nucleotide C) at positions where the major groove faces the protein cylinder (every 10th bp), and TT, AA, and TA where the minor groove faces the cylinder [5]. A simulation of a nucleosome model that takes sequence-dependent DNA properties into account actually predicted these rules [14], and a simplified version of this nucleosome model made it possible to show analytically that these rules follow from the intrinsic shapes of the different bp steps together with the fact

that every bp is part of a longer bp sequence [31]. Interestingly, rotational positioning cues can even be freely placed on top of genes without altering the resulting amino acid chains, since the genetic code is degenerate [43].

On the other hand, the translational positioning of nucleosomes is caused by DNA stretches that, overall, have a higher affinity for nucleosomes. It is known that this correlates well with their GC content [8, 9, 54, 76]. The physics behind the translational positioning is less clear than that of the rotational one; a recent study suggests that it is more about entropy than energy [77]. There are various examples for translational mechanical cues, e.g. nucleosome-depleted regions before transcription start sites in unicellular organisms, which facilitate transcription initiation [6, 9], mechanically encoded retention of a small fraction of nucleosomes in human sperm cells, which allows for the transmission of paternal epigenetic information [35], and the positioning of six million nucleosomes around nucleosome-inhibiting barriers in human somatic cells [8].

Important is also the fact that histone octamers can spontaneously change their position along DNA, a phenomenon called nucleosome sliding [44]. This way nucleosomes sample different positions, allowing for a rather slow equilibration of nucleosomes, *in vitro* at least locally [77]. Two mechanisms have been suggested, both are based on thermally induced defects inside the nucleosome: single bp twist defects (a missing or an extra bp) [45, 46, 78] and 10 bp bulges [47, 48]. Recent simulation studies [49, 50, 79] found that both mechanisms can be at play and that it depends on the underlying bp sequence which one is preferred mechanism. Also a new experiment [80] indicates two types of movements of nucleosomes along DNA, small scale repositioning on short time scales and longer ranged repositioning events on the time scale of minutes.

Importantly, *in vivo* there are chromatin remodellers present that use ATP to move nucleosomes along DNA. New experiments [51, 81, 82] and simulations [52] suggest that at least some of them induce twist defect pairs inside the nucleosome. Chromatin remodelers might help nucleosomes to equilibrate their locations along DNA [68], but they might also perturb the intrinsically preferred positioning of nucleosomes, together with other proteins that compete for DNA target sites [54]. In addition, pioneer transcription factors that can bind to nucleosomal DNA might play a role in recruiting remodelers [83].

4.1.2 Introduction to translation speed and cotranslational folding

A gene on the DNA is transcribed and spliced such that it becomes mRNA, which is then translated one codon at a time by the ribosomes, which creates amino acid chains by facilitating the attachment of tRNAs containing the correct anticodon to the corresponding codons.

The rate at which amino acids are attached to the growing amino acid chain is codon-dependent and can be changed (over the course of evolution) since synonymous codons can have different attachment rates. This is because translation speed of codons depends on the concentrations of corresponding tRNAs. This concentration are correlated with the number of genes coding for the tRNAs [84]. It is species-specific, cell-specific and it depends on the circumstances of the cells [26, 27].

Translation speed has important consequences for the resulting proteins. Faster translation leads to larger amounts of protein, increased translational fidelity, less frameshifting, less amino acid misincorporation, less protein degradation and less mRNA decay, while slower translation enhances co-translational protein folding by giving more time for the protein to fold [28]. Translation speed can affect the quality and quantity of proteins in many different ways. For instance: ribosome pausing can lead to ribosome collisions and co-translational degradation of both mRNA and nascent chain. [85] Lopez and Pazos [86] showed that a number of protein functional and structural features are reflected in the patterns of ribosome occupancy, secondary structure and tRNA availability along the mRNA. They also showed specific examples where patterns of translation speed point to the protein's important structural and functional features. Pechmann and Frydmanhis' analysis of codon optimality in ten closely related yeasts reveals universal patterns of conserved optimal and nonoptimal codons, often in clusters, which associate with the secondary structure of the translated polypeptides independent of the levels of expression [87]. Mian Zhou et al. replaced the original codons of a clock protein with the most preferred synonymous codon, i.e. the one with the highest translation speed. This mutation reduced the quality of the final protein, a proof that tuning the translation speed is necessary for the protein folding [88]. More examples can be found in recent reviews of O'Brien et al.[89] and Luitkute et al. [90].

4.1.3 Overview

In the next sections we again use graph representations of DNA in combination with a shortest path algorithm, as encountered in the previous chapter. We discuss the multiplexing of genetics and mechanics by providing a short recap of Chapter 3 on how to find the lowest and highest possible nucleosome energies on top of genes. This is followed by section 4.3, which provides a short description of the translation speed model we use and how to find the highest and lowest posible translation speeds. In section 4.4 we combine all three layers of information. We find how much the highest and lowest possible nucleosome energy on a gene are influenced by restrictions on translation speed. Section 4.5 brings this subject to its logical conclusion by discussing genetically modified organisms. It describes a heuristic method on how to change the DNA sequence of a gene, such that, when one puts this gene in a different organism, the genetical information is conserved while the mechanical information and translation speed landscape are close to their counterparts in the original organism. We again end this chapter with a conclusion.

4.2 Multiplexing of genetics and mechanics

To find out how genetics and this nucleosome energy are multiplexed, we revisit a method presented in the previous chapter, where we showed how to obtain the lowest and highest possible nucleosome energy for a position on a gene without changing the resulting amino acid chain. We represented all possible sequences coding for the same amino acid chain as paths through a weighted directed graph in combination with a shortest path algorithm. The weights were given by a probabilistic trinucleotide model obtained through Monte Carlo simulations of a coarse-grained

nucleosome model with sequence-dependent DNA elasticity [10], though any short-range probability or energy model may be used. In this chapter we use the same trinucleotide nucleosome energy model where the energy cost of wrapping a sequence S of nucleotides $S_i \in \{A, T, C, G\}$, $i = 1, ..., L$ with $L = 147$ into a nucleosome is given by

$$E(S) = \sum_{n=1}^{L-2} E_n(S_{n+2}, S_{n+1}, S_n) \tag{4.1}$$

where the E_n's are energy costs associated with a trio of nucleotides.

As we stated in the previous chapter, the simulations that generate the trinucleotide model use a coarse-grained nucleosome model, where the DNA is restricted by 26 constraints corresponding to bound phosphates in the DNA backbone. These constraints represent the 14 binding sites of the DNA to the protein core which were extracted from the nucleosome crystal structure without introducing free parameters. The DNA base pairs are treated as rigid plates by using the rigid base pair model. The rigid base pair model assumes nearest-neighbour interactions with energy costs incurred by the square of the deformations from the intrinsically preferred geometry in any of the degrees of freedom and its cross-terms [32]. The relative orientations of the plates can be described using three translational and three rotational degrees of freedom. As a result one obtains a superhelix. For details on how to use equation 4.1 to obtain upper and lower limits of the nucleosome energy of a gene, see appendix C.1.

In this chapter we study a gene from human: the gene TNF, Tumor Necrosis Factor, which codes for a cytokine. A cytokine is a signaling molecule involved in the immune response of mammals [2]. TNF has an imporant role for both innate and adaptive immune responses, and is related to cancer progression and metastasis [91]. TNF was chosen because it is the second-most cited gene [30]. The most cited gene, p53 [30], was not used because it has no exon significantly longer than the nucleosomal wrapping length. The fourth exon of TNF is much longer than the nucleosomal wrapping length, allowing us to safely ignore the effect of noncoding DNA on the nucleosome energy landscape.

Figure 4.1 depicts the energy landscape for the fourth exon of TNF. The dyad position is the position of the base pair in the middle of the nucleosome. It also depicts the highest and lowest possible energy at these positions for any theoretical exon coding for the same amino acid chain.

This provides us with an indication of the malleability of the energy landscape: only a relatively small part of the attainable energy space is being used. This provides room for other layers of information on the same piece of DNA, such as translation speed.

4.3 Multiplexing of genetics and translation speed

We can do the same analysis for the multiplexing of genetics and translation speed. For this we require a model for translation speed.

To add a single amino acid to the polypeptide chain, the ribosome goes through a cycle of chemomechanical reactions. A summary of distinct states and reversible/ir-

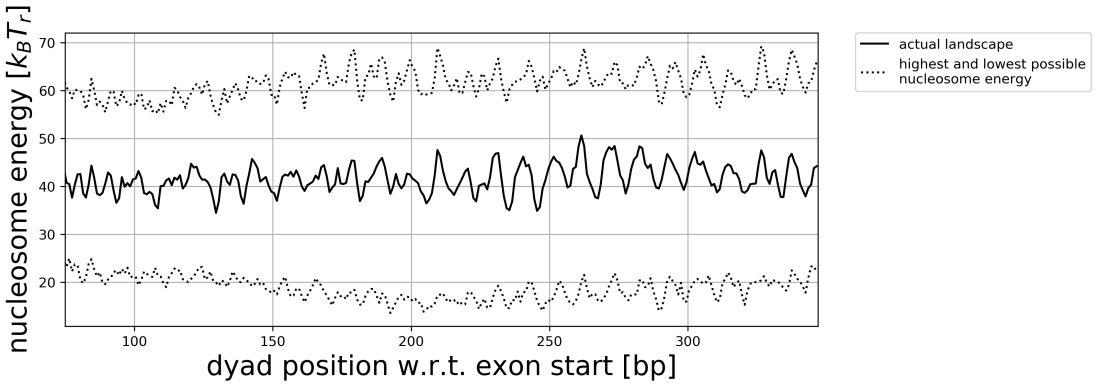


Figure 4.1: The energy landscape for the fourth exon of the human gene Tumor Necrosis Factor (TNF) is depicted by the solid line. The dotted lines depict the highest and lowest possible energy at these positions for any theoretical exon coding for the same amino acid chain, obtained by using a graph representation of all possible synonymous codons and a shortest path algorithm. The dyad position is the position of the central base pair on the nucleosome.

reversible steps of the decoding and peptidyl transfer processes can be found in the reviews by Frank and Gonzalez [92] and Wohlgemuth et al. [93]. Knowing the corresponding rates for these steps [29] and tRNAs concentrations, one can calculate the average translation rate of different codons in different organisms. A detailed calculation of the translation time can be found in the work of Rudorph et al. [29]. In their model, the translation rate of a codon C depends on concentrations of cognate, near-cognate and non-cognate tRNAs which we denote by X_C^{co}, X_C^{nr} and X_C^{no}. For each codon a tRNA is cognate if there is no mismatch in the codon-anticodon complex, the near-cognate tRNAs have one mismatch and noncognate ones have more than one mismatch. Rewriting their result, we can see that the translation rate $T(C)$ for codon C can be written as follows as a function of the concentrations of cognate, near-cognate and non-cognate tRNAs which we denote by X_C^{co}, X_C^{nr} and X_C^{no} respectively:

$$T(C) = \frac{a' X_C^{\mathrm{co}} + b' X_C^{\mathrm{nr}}}{a X_C^{\mathrm{co}} + b X_C^{\mathrm{nr}} + c X_C^{\mathrm{no}} + d}. \tag{4.2}$$

Here a, b, c and d are dimensionless factors, a' and b' have dimension of one over time and all factors are functions of translation rates. These factors are independent from the type of codon and only depend on the internal dynamics of the ribosome. They depend on ρ and τ, where ρ is a dimensionless function of transition probabilities in a specific branch, cognate or near cognate (ρ_{co} or ρ_{nr}), and τ is a timescale for a tRNA going through a cognate or near cognate branch (τ_{co} or τ_{nr}). They also depend on ω_{pro}, ω_{off} and κ_{on}, which are the processing rate, dissociation rate and association rate of a tRNA at a ribosome [29]. For the explicit dependencies and values for *E. coli*, see Table 4.2.

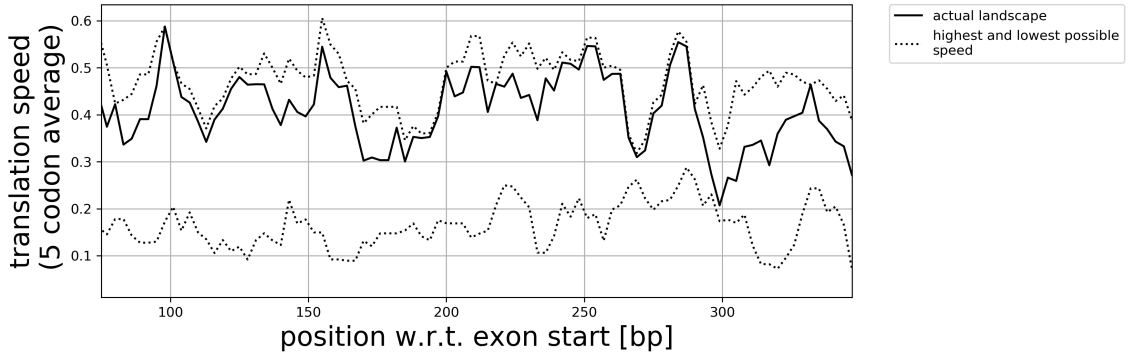


Figure 4.2: The translation speed landscape for the fourth exon of the human gene Tumor Necrosis Factor (TNF) is depicted by the solid line. The dotted lines denote the highest and lowest possible translation speed when codons may be replaced by synonymous codons. We average over five codons to obtain a clearly visible signal.

Parameters	**Description**	**E. coli** $37°C$
a'	$\rho_{\mathrm{co}}\omega_{\mathrm{pro}}$	$100 \pm 40\ s^{-1}$
b'	$\rho_{\mathrm{nr}}\omega_{\mathrm{pro}}$	$0.12 \pm 0.09\ s^{-1}$
a	$\rho_{\mathrm{co}}(\tau_{\mathrm{co}}\omega_{\mathrm{pro}} + 1)$	1.6 ± 0.4
b	$\rho_{\mathrm{nr}}(\tau_{\mathrm{nr}}\omega_{\mathrm{pro}} + 1)$	1.0 ± 0.6
c	$\omega_{\mathrm{pro}}/\omega_{\mathrm{off}}$	0.21 ± 0.11
d	$\omega_{\mathrm{pro}}/\kappa_{\mathrm{on}}$	0.86 ± 0.31

Table 4.2: Values of a, b, c, d, a' and b' for *E. coli* at 37 degrees Celsius.

The overall process of translation is conserved between the eukaryotic and prokaryotic ribosomes [94] therefore the same formula applies to both of them. However, the parameters can be different in different organisms and also in different situations such as different growth rates of cells. Here we assume that these differences do not change the overall shape of the translation speed profile along a gene. We specially prefer this scale over the tRNA adaptation index, tAI, because the later does not consider the time consumption due to the near-cognate tRNAs.

It has been shown that the tRNA concentration corresponds to the genome copy number of that tRNA [84]. These copy numbers can be found for many species, in the tRNA genome database [http://gtrnadb.ucsc.edu]. To calculate the concentration of each tRNA, we multiply the genome copy number of that type with the average total concentration of all tRNAs in a cell, which can be around 10 micro molar [84].

The translation speed in this model does not depend on the neighbours of codons. Therefore, to obtain the highest and lowest possible translation speed (keeping the protein intact) we can simply pick the codons with the highest and lowest speeds. The result for the fourth exon of TNF is depicted by figure 4.2. We average over five codons to obtain a clearly visible signal. This signal uses almost the full possible range of the translation speed. Even though TNF strongly favours high translation speed, around position 300 it is very close the lowest possible value.

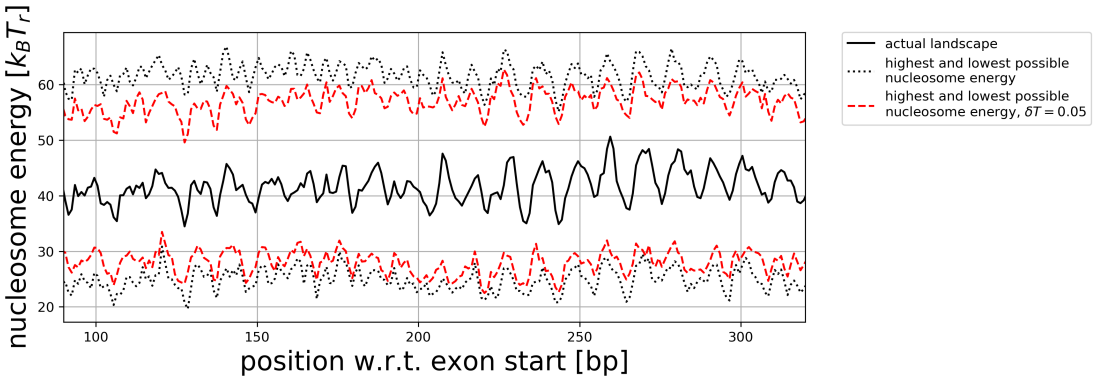


Figure 4.3: Same as figure 4.1 but with the addition of the highest and lowest possible nucleosome energy with a translation speed restriction of $\delta T = 0.05$.

4.4 Multiplexing three layers of information: genetics, mechanics and translation speed

We will now study the multiplexing of the three types of information. We have seen that the space of possible nucleosome energies for a gene is large. Now we investigate the very same while including the translation speed landscape. What are the lowest and highest possible nucleosome energies when the translation speed landscape at any position may only change by some fixed constant δT?

We calculate the energy cost of wrapping a codon sequence C around a nucleosome. A nucleosome of 147 base pairs corresponds to either 49 or 50 codons. We denote the codon sequence by $C = (C_0, C_1, ..., C_{49})$. We look at the set of sequences where the translation speed at any codon position (averaging over five codons) may only be altered by no more than some value δT:

$$\frac{1}{5}\left|\sum_{i=-2}^{i=2} T(C_{n+i}) - T(C_{n+i}^{\text{new}})\right| \leq \delta T, \quad \text{for } n = -2, -1, ..., 51, \tag{4.3}$$

where C^{new} denotes any sequence of synonymous codons. We have included four neighbouring codons on each side of the codon sequence, denoting them by C_i for $i < 0$, $i > 49$. (Including more codons did not make a difference for the results.)

Applying this restriction to a graph is not difficult. In the previous Chapter we implicitly used that genetic information can be considered as a restriction on the possible nodes of a graph: one can simply disallow nodes corresponding to nonsynonymous codons. We apply the same strategy for the translation speed: we disallow (or prune) nodes that do not conform to the speed restriction, see appendix C.2. Again one can find the lowest and highest energy by calculating its shortest and longest paths. The result for TNF is depicted by figure 4.3. It depicts that a strong restriction, $\delta T = 0.05$, results in only a small change in the highest and lowest possible energies.

4.5 Genetically modified organisms

Since we have observed some theoretical flexibility for the three layers of information -genetical information, mechanical information and translation speed, the next step is to study this flexibility for a scenario with biological relevance. We want to put a gene in a different organism - a host organism - and see what happens to the three layers of information. Since the conversion of codons to amino acids is practically universal, a gene in a host organism will almost surely encode the same amino acid chain. Secondly, since the nucleosome energy landscape depends only on the physical properties of the sequence, the nucleosome energy landscape, too, remains unchanged. However, the translation speed landscape, our third layer of information, may be very different in a host organism. This is due to differences in tRNA concentrations between organisms. In figure 4.4a we show that the shape of the translation speed landscape of TNF is qualitatively different in hosts yeast and rice. Our goal is for the host organism to have all three layers of information close to the original. More specifically, we want to make the translation speed landscape resemble the original landscape, without changing the amino acid sequence and while making only minor changes to the nucleosome energy landscape.

4.5.1 Translation speed in host organisms

Our first goal is to find out exactly how close the translation speed landscape in a host organism can get to the original landscape, ignoring for the moment the nucleosome energy landscape. It turns out that this can be a problem. See, for example, the highest and lowest values of the translation speed for the gene TNF in host organisms in figure 4.4b. We see that the original translation speed landscape fits almost everywhere inside the limits of host organism yeast. For the host rice on the other hand it is at many positions impossible to restore the translation speed of this gene without changing some of the amino acids.

Now we will show how close the translation speed landscape of yeast can get to the original while keeping amino acid information intact. Formally, we will minimize the distance D_T between the original translation speed landscape in human of a gene $G = (G_0, ..., G_{3N})$ and the translation speed landscape in *yeast* of gene G', a sequence that codes for the same amino acids. Here N is the number of codons in G and G_i denotes the i^{th} base pair.

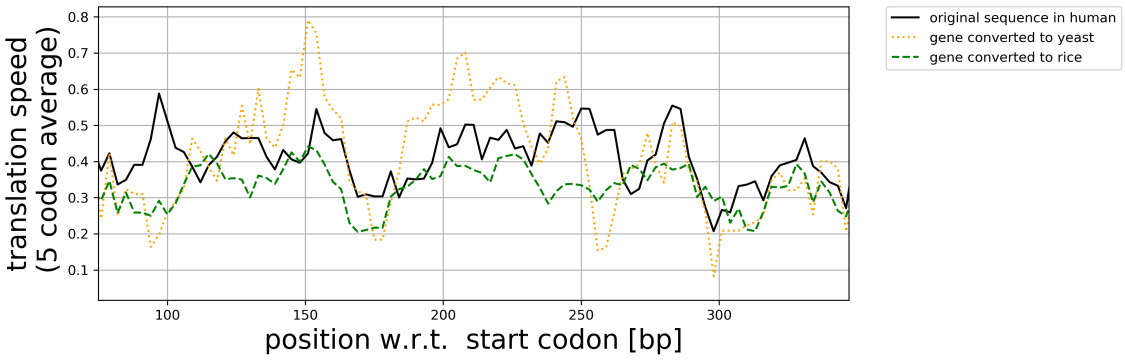
Let A_G be the set of all sequences that code for the same amino acid chain as G. We choose the closest sequence G' such that

$$D_T(G, G') \leq D_T(G, X) \text{ for all } X \in A_G \tag{4.4}$$

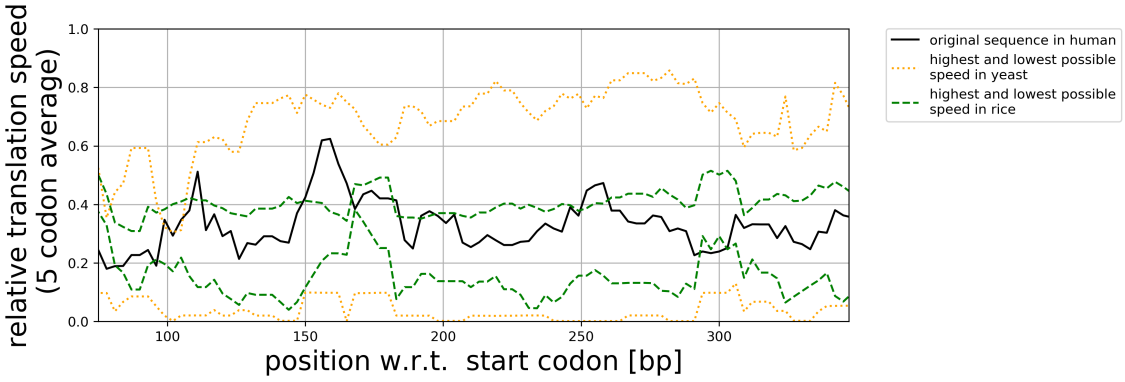
with

$$D_T(G, X) \equiv \sum_{p=2}^{N-3} \Delta T^{\text{human}}_{\text{yeast}}(G, X, p), \tag{4.5}$$

where $\Delta T^{\text{human}}_{\text{yeast}}(G, X, p)$ describes the difference between the average translation speed of an altered sequence X in yeast and the original sequence G in human, five

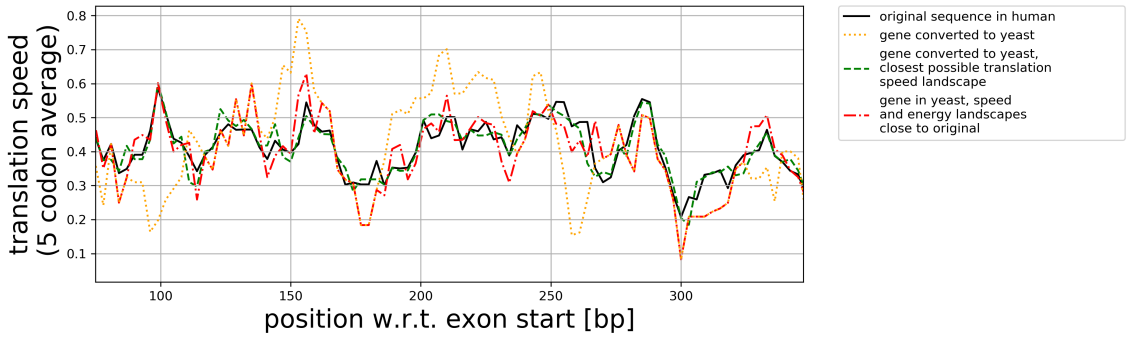


(a)

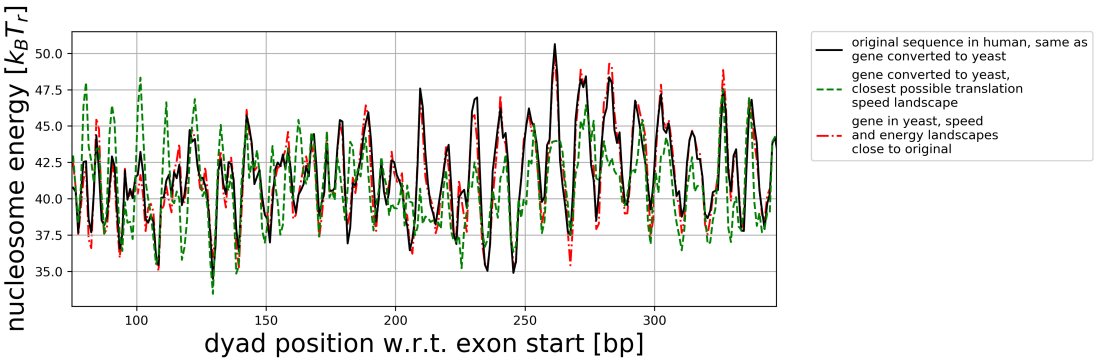


(b)

Figure 4.4: Fig. (a) depicts the translation speed landscape of the fourth exon of TNF in three organisms: the original (human) and two possible host organisms: yeast and rice. Fig. (b) shows the original landscape as well as the highest and lowest possible translation speed values in the hosts. We see that the original landscape cannot be reproduced in rice by looking at the highest and lowest values alone.



(a)



(b)

Figure 4.5: For the fourth exon of gene TNF, (a) depicts several translation speed landscapes and (b) the corresponding nucleosome energy landscapes. The original landscapes in human are depicted by a solid line. The translation speed landscape in yeast of the original sequence is depicted by the orange dotted line. The closest possible translation speed landscape is depicted by the green dashed line. The corresponding nucleosome energy landscape is now quite different from the original landscape. A compromise is made for the red slash-dotted curves, where both landscaped highly resemble the original landscapes, using equation 4.8 with $c_T = 1$ and $c_E = 1/1000$ $[1/k_BT_r]$.

codons centered around a codon position p:

$$\Delta T_{\text{yeast}}^{\text{human}}(G, X, p) \equiv \left| \sum_{i=-2}^{i=2} T_{\text{human}}(G_{3(p+i)} G_{3(p+i)+1} G_{3(p+i)+2}) - T_{\text{yeast}}(X_{3(p+i)} X_{3(p+i)+1} X_{3(p+i)+2}) \right|. \tag{4.6}$$

Here T_{organism} denotes for which organism the translation speed is calculated.

The resulting sequence G' corresponds to a translation speed landscape depicted by the green interrupted line in figure 4.5a for TNF. The altered translation speed landscape in yeast is extremely close to the original landscape in human. As a side effect, changing the base pair sequence - even by using only synonymous codons - will likely alter the nucleosome energy landscape, as shown for TNF by the green interrupted line in figure 4.5b. For examples using other genes, see C.4.

4.5.2 Restoring all layers of information

This brings us to our final method. We will attempt to restore the translation speed landscape while keeping the nucleosome energy landscape into consideration. To do so we compare ranges of 5 codons, the same length of DNA we study for the translation speed averages. (To do this perfectly, one should compare ranges of 147 base pairs, the length of a nucleosome. This would be impossible to do using our method: the graphs would consist of too many nodes. Fortunately we will see that it is not necessary to be so precise.) Formally, we will minimize the distance $D_{T\&E}$ between a combination of the translation speed and nucleosome energy landscape of G and G''. We want to find a sequence G'' such that

$$D_{T\&E}(G, G'') \leq D_{T\&E}(G, X) \text{ for all } X \in A_G \tag{4.7}$$

with

$$D_{T\&E}(G, X) \equiv \sum_{p=2}^{N-3} c_T \Delta T^{\text{human}}_{\text{yeast}}(G, X, p) + c_E \Delta E(G, X, p). \tag{4.8}$$

The constants c_T and c_E can be freely chosen, depending on which quantity, translation speed or nucleosome energy, one finds more important to be close to the original. The function $\Delta T^{\text{human}}_{\text{yeast}}(G, X, p)$ was defined by equation 4.5 and still describes the difference between the translation speed of sequence G in human and sequence X in yeast of five codons around codon position p. We have introduced a function $\Delta E(G, X, p)$ which describes the same but for energy. To properly define this function, it needs to reflect that we want to know the effect of the change of sequence on the *entire* nucleosome energy landscape. Therefore, we find $\Delta E(G, X, p)$ by summing over all possible positions of this 15 bp stretch on $147 + 14$ possible positions on a nucleosome. We sum over $147 + 14$ positions, since this is the number of positions where at least one of the possibly changed base pairs is contained within a nucleosome, i.e. the number of positions where the nucleosome energy could be affected by substitution of codons.

This leads to the definition:

$$\Delta E(G, X, p) \equiv \sum_{j=-7}^{147+7-1} \left| \sum_{i=-7}^{i=7-2} E_{j+i}(G_{p+i}, G_{p+i+1}, G_{p+i+2}) - E_{j+i}(X_{p+i}, X_{p+i+1}, X_{p+i+2}) \right|. \tag{4.9}$$

Note that, since the nucleosome energy is invariant under a change of organism, this function too does not depend on the organisms chosen. This is an amusing quirk of this system which comes from the fact that, while a sequence may have different translation speeds in different organisms (caused by differences in tRNA concentrations), the physical properties of DNA are the same between species. Note that this $\Delta E(G, X, p)$, like $\Delta T^{\text{human}}_{\text{yeast}}(G, X, p)$, is related to a total distance between the original sequence G and altered sequence X, but in this case, the total distance between the nucleosome energy landscapes. This distance $D_E(G, X)$ is defined by

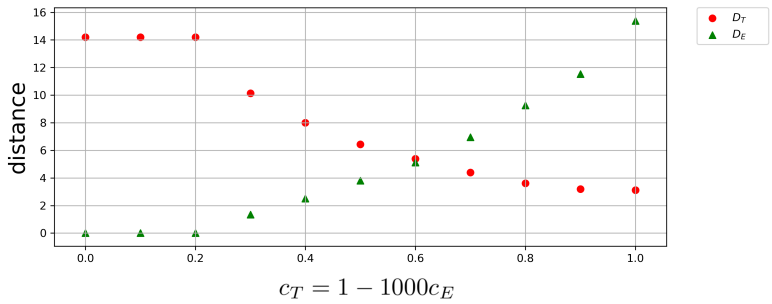


Figure 4.6: The distances D_T and D_E (defined by equations 4.5 and 4.10) are depicted as a function of c_T and c_E. Distance D_T represents the difference between the original translation speed landscape in human of a genetic sequence G (in this case, the fourth exon of TNF) and an altered sequence G'' in yeast. Distance D_E represents the same but for the nucleosome energy. The relative values of c_T and c_E represent how important it is for a quantity, translation speed or nucleosome energy,to be close to the original in a host organism. For a range of values of c_T and c_E, the combined distance $D_{T\&E}(G, X)$ is minimized. For $c_T = 1$, this is equivalent to minimizing $D_T(G, X)$, and for $c_T = 0$ it is the samne as minimizing D_E. For Fig. 4.5 and all other figures we chose $c_E = 1/1000$ $[1/(k_B T_r)]$ and $c_T = 1$, which is equivalent to $c_T = 0.5$ and $c_E = 1/500$ in this figure.

$$D_E(G, X) \equiv \sum_{p=2}^{N-3} \Delta E^{\text{human}}_{\text{yeast}}(G, X, p). \tag{4.10}$$

Returning to equation 4.8, we choose $c_E = 1/1000$ $[1/(k_B T_r)]$ and $c_T = 1$, which brings the quantities of speed and energy to the same order of magnitude while fixing the units. Appendix C.3 describes how to create a graph with the correct weights to obtain G''.

The result for TNF is depicted by red dash-dotted line in figures 4.5a and 4.5b, where we see that both the nucleosome energy and the translation speed landscape are now close to the original. Fig. 4.6 depicts how the distances D_T and D_E are affected by the choices of c_T and c_E. For $c_T \ll c_E$, $D_{T\&E}$ becomes similar to D_E, and for $c_T \gg c_E$, $D_{T\&E}$ becomes similar to D_T. For $c_E = 1/1000c_T$, we strike a balance between keeping the values of D_T and D_E low.

4.6 Conclusion

We have presented a novel approach to study the multiplexing of genetics, mechanics and translation speed. In the previous chapter we found the highest and lowest possible nucleosome energies on top of a gene, when one can only replace codons with synonymous codons such that the sequence codes for the exact same amino acid chain. In this chapter we have included the translation speed in our analysis, since this speed can be an important factor for the proper function of the final protein. We did so by adding an additional restriction to the analysis: any altered sequence must have a translation speed landscape close to the landscape corresponding with the unaltered sequence. This restriction was applied by pruning nodes from a graph.

A second approach we used was to incorporate translation speed in the weights of graphs. When one puts a gene of one organism in a host organism, the translation speed landscape in the host may be very different from the landscape in the original species. Using this second approach, we demonstrated how to to change the genetic sequence such that the host will produce a protein with a translation speed landscape, as well as a nucleosome energy landscape, very similar to the landscapes in the original organism.