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

Zuiddam, M.

Freedom of Additional Signals on Genes: on the combination of DNA mechanics, genetics and translation speed

Freedom of Additional Signals on Genes: on the combination of DNA mechanics, genetics and translation speed

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op woensdag 6 april 2022

klokke 15.00 uur

door

Martijn Zuiddam

geboren te Leiderdorp

Promotor:	Prof. dr. Helmut Schiessel
Promotiecommissie:	Prof. dr. Annick Lesne (Sorbonne Université, Paris, Frankrijk)
	Prof. dr. Enrico Carlon (Katholieke Universiteit, Leuven, België)
	Prof. dr. John van Noort
	Prof. dr. Luca Giomi
	Prof. dr. Jan Aarts

The work in this thesis was supported by the Delta ITP consortium, a program of the Netherlands Organisation for Scientific Research (NWO).
Casimir PhD series 2022-01
ISBN 978-90-8593-512-4

Contents

Chapter 1

Introduction

Essential for all known forms of life is the DNA molecule. This molecule is the main carrier of information on the development, reproduction and functioning of living organisms. Popularly known for its applications in forensics and family heritage research, DNA (deoxyribonucleic acid) is a central topic of fundamental theoretical research in biology, chemistry, physics, medicine, pharmaceutics, and bioinformatics. It was discovered in 1869 by the biochemist Friedrich Miescher[1]. Miescher did consider the possiblity for DNA to be involved in the transfer of hereditary information. However, he thought that it would be unlikely for one specific molecule to be the source of such a wide variety of different organisms. According to him, the amount of information contained in DNA, if any, would be small, too small even to explain the differences between individuals of the same species [1].

Of course, now we know that Miescher hugely underestimated the importance of this molecule. It is now quite commonly known how DNA is able to contain all this genetic information while still being a relatively simple molecule. DNA contains four different bases, adenine, thymine, cytosine and guanine, abbreviated as A, T, C and G. Like the twenty-six letters of our alphabet are more than enough to form this sentence, and like zeroes and ones can constitute a doctoral thesis on a computer, the four bases suffice to contain genetic information. A sequence of these bases may constitute a *gene*, a piece of DNA which codes for a protein[2].

One might think that this key role in genetics is more than enough responsibility for such a simple four-letter system as DNA. One would be wrong; this would be a mistake similar to Miescher's. In this dissertation we investigate a second — and even a third! — layer of information carried by DNA. We mainly focus on the second layer, which is expressed mechanically: the DNA molecule's three-dimensional folding, twisting and bending in space changes depending on the choice of bases because of their mechanical properties. We specifically investigate the role of mechanical information in the positioning of nucleosomes. The third layer of information is translation speed: the rate at which a protein is constructed, which has important consequences for the function of proteins. Below we will expand upon these layers of information on DNA.

[1] who named the substance "nuclein" after the nucleus of the cell [1]

[2] or, to be exact, for RNA

1.1 The genetic layer of information on DNA

Genes on the DNA can be copied to RNA (ribonucleic acid), which performs a wide variety of functions. One of these functions is to code for proteins. This is performed by messenger RNA (mRNA), which carries a series of codons (three-base sequences that code for specific amino acids). A molecular machine called a ribosome "translates" the mRNA to create a protein consisting of a chain of the encoded amino acids. Examples of other RNAs are ribosomal RNAs (rRNAs, which become part of ribosomes) and transfer RNAs (tRNAs, which are responsible for bringing the correct amino acids to the ribosomes), as well as different kinds of regulatory RNAs. DNA also contains information that does not need to be transcribed to function. Consider, for instance, the promoter: a promoter is a sequence which determines where the RNA polymerase binds to a gene (RNA polymerase copies the information on DNA to an RNA sequence). Also enhancers and silencers exist: these are sequences that attract transcription factors, which will either stimulate or inhibit transcription [2, 3]. Multiple different mRNA sequences may code for the same protein because of *synonymous codons*. The existence of synonymous codons is a consequence of the degeneracy of the genetic code: multiple codons can code for the same amino acid. In fact, 18 out of 20 amino acids are coded for by multiple synonymous codons, see Table 1.1. Therefore, for any protein, multiple sequences exist that code for its exact amino acid sequence. As a result, a protein-coding sequence may encode additional layers of information.

1.2 The mechanical layer of information on DNA

The second layer of information we discuss is a mechanical layer. After all, one must not forget that DNA is a physical object, that has long been studied by polymer physicists. The intrinsic shape, as well as the flexibility of this object, depends on the choice of its bases. A piece of DNA consisting of only A's and T's will bend differently than a sequence of G's and C's. This allows for a second layer of information on the DNA in addition to the genetic layer. This mechanical layer is particularily interesting in the case of the nucleosome, a DNA-protein complex.

Nucleosomes consist of 147 base pairs (bp) of DNA wrapped around a protein core, like a string around a spool. They are considered the fundamental building blocks of chromatin and are responsible for compactifying the DNA and serve to form its higher-order structures. The DNA is connected to the protein core via fourteen binding sites. The core consists of eight so-called histones. These histones have tails, which can be chemically modified to induce certain behaviours in nucleosomes, for example making transcriptional control sequences on the DNA accessible to proteins involved in the transcription of DNA to RNA [2]. Because of this, nucleosomes are involved in epigenetic regulation of transcription, where the term epigenetic refers to inherited changes in how cells function that do not result from changes in DNA sequence. The modifications to the nucleosomes, as well as the corresponding changes in gene expression, can be inherited by the offspring of an organism. The epigenetic function of nucleosomes is one of the reasons why the location of a nucleosome on the DNA matters, for instance near control sequences such as promoters [2].

Alanine	GCT	GCC	GCA	GCG		
Arginine	CGT	CGC	CGA	CGG	AGA	AGG
Asparagine	AAT	AAC				
Asparic acid	GAT	GAC				
Cysteine	TGT	TGC				
Glutamic acid	GAA	GAG				
Glutamine	CAA	CAG				
Glycine	GGT	GGC	GGA	GGG		
Histidine	CAT	CAC				
Isoleucine	ATT	ATC	ATA			
Leucine	CTT	CTC	CTA	CTG	TTA	TTG
Lysine	AAA	AAG				
Methionine	ATG					
Phenylalanine	TTT	TTC				
Proline	CCT	CCC	CCA	CCG		
Serine	TCT	TCC	TCA	TCG	AGT	AGC
Threonine	ACT	ACC	ACA	ACG		
Tryptophan	TGG					
Tyrosine	TAT	TAC				
Valine	TAA	TAG	TGA			

Table 1.1: This table depicts synonymous codons: all codons (combinations of three bases) that code for the same amino acids (the building blocks of proteins). These codons have different mechanical properties, which enables the existence of a mechanical layer on top of protein-coding DNA. Furthermore, different codons may be translated at varying rates. This leads to the existence of a translation speed layer of information.

Since the locations of nucleosomes are important, there is a biological incentive to position a nucleosome on the DNA. As it turns out, this positioning is partially caused by the type and order of the bases (A, T, C and G) on the DNA sequence. Various experiments have shown that nucleosomes have sequence preferences [4–6]. These preferences lead to a "nucleosome positioning code" or "nucleosome positioning signals". There are two types of nucleosome positioning signals on the DNA: translational and rotational [7]. Translational positioning signals come from DNA stretches with a relatively high affinity for nucleosomes. It turns out that this affinity is correlated with the GC content of DNA, which means that DNA with a lot of G's and C's attracts nucleosomes [8, 9]. Rotational positioning signals are more complicated and to understand them we need to explain more about the DNA and the nucleosome. DNA consists of two strands, which connect together to form a double helix with a periodicity of 10 bp. Bases on opposite strands can form four different base pairs (A-T, T-A, G-C, C-G). A pair of neighbouring bases on the same strand are called dinucleotides. The double helix has a so-called mayor groove and minor groove, on opposite sites of each other, which face the protein core every 10th base pair. It has been experimentally shown that sequences with a high affinity

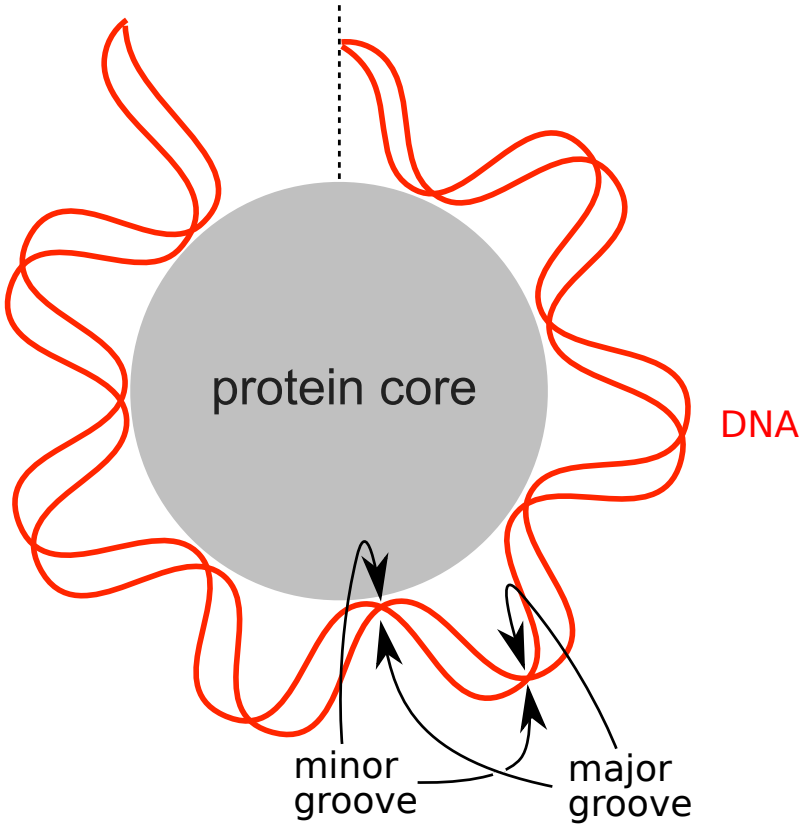


Figure 1.1: A nucleosome consists of 147 base pairs (bp) of DNA wrapped around a protein core. Here we show an exaggerated depiction of the major and minor grooves. These grooves face the protein core at 10 bp intervals, which leads to the nucleosome positioning rules. Sequences with a high affinity for nucleosomes follow nucleosome positioning rules related to this major and minor groove. For instance, the probability to find a dinucleotide TT, AA, or TA is highest where the minor groove faces the protein core, and the probability to find a GC step is highest where the major groove faces the protein core.

for nucleosomes follow nucleosome positioning rules related to this major and minor groove, see Fig. 1.1. The most important of the nucleosome positioning rules are the following: the probability to find a dinucleotide TT, AA, or TA is highest where the minor groove faces the protein core, and the probability to find a GC step is highest where the major groove faces the protein core. Because of the helical periodicity of the DNA, these probabilities have a period of 10 bp [4–6].

There exist many models that aim to replicate the nucleosome positioning code. One of these models is especially important in this dissertation: we will use it in Chapters 3 and 4. It is the model by Tompitak et al., which is an excellent approximation [10] to the Eslami-Mossalam et al. nucleosome model [11]. Underlying the Eslami-Mossalam model, and many other models, is the rigid base pair model [11–13]. In the rigid base pair model, all base pairs of the DNA are modelled as rigid plates. The energy of a conformation of DNA depends on the positions and orientations of these bp plates with respect to the neighbouring bp plates (for a visual representation, see Chapter 2, Fig. 2.2). Neighbouring plates have six (coupled) degrees of freedom: three rotational, three translational. For all degrees of freedom, the plates have equilibrium positions with respect to each other, and stiffnesses, the values of which have been obtained from experiments and molecular dynamics simulations. The model describes sequence-dependent DNA mechanics using these values, since the equilibrium values and stiffnesses are generally different for different dinucleotides.

These degrees of freedom, together with the fact that a nucleosome contains 147 bp, makes the model too complicated to solve analytically. However, it is doable by performing a Monte Carlo simulation. Eslami-Mossalam et al. even take it a step further and use Mutation Monte Carlo, which moves and mutates the DNA at the same time, such that both the configuration space as well as the sequence space are evaluated [11]. This method was made even more practical by Tompitak et al., who built an approximative scheme to the Eslami-Mossalam model that is a factor 10^5 faster and is very successful in predicting nucleosome positioning [9]. While successfull in reproducing nucleosome positioning rules, these models are unable to explain them. For instance: as stated above, the probability to encounter GC steps is highest where the major groove faces the protein core. Surprisingly, these positions cause the highest energy costs for GC. Such counterintuitive results are difficult to answer using simulations. In Chapter 2 we approach this problem by using a much simpler, analytically-tractable model. The most important simplification is that we idealize the shape of the DNA wrapped around the protein core as a perfect superhelix, and that we keep it immobile (unlike other models where DNA can move around [14–19]). We come up with an approach (in principle applicable to any model that freezes DNA in a fixed shape [15, 16, 20–22]) to precisely investigate what factors determine the nucleosome positioning code. In the case of GC, we find that the counterintuitive results can be explained by the average energy of all possible neighbours of dinucleotide GC, which favour the positions GC dislikes most.

It has been established that different sequences can have different affinities for nucleosomes, through the nucleosome positioning code, a second layer of information on DNA. In Chapter 3 we study the amount of freedom of this second layer. We do so by answering the following questions: How malleable is the mechanical information? How can we construct sequences that like or dislike nucleosomes the most? Can

such sequences be used to position nucleosomes on genomic DNA? Earlier attempts to answer these questions exist. The Mutation Monte Carlo method (MMC), for example, can be used to find sequences with high nucleosome affinity. Also, in the case of positioning nucleosomes on coding DNA, Eslami-Mossalam et al. have created a synonymous Mutation Monte Carlo method (sMMC). This method is the same as MMC but it may only make mutations that replace codons by synonymous codons (Table 1.1). Using sMMC, Eslami-Mossalam et al. have shown that it is possible to position nucleosomes on a range of positions on a gene of yeast [14].

In Chapter 3 we take their approach a step further. We map all possible DNA sequences on a weighted graph and use a shortest path algorithm to find the sequences with highest and lowest possible nucleosome wrapping energy. The two huge advantages of this method is that shortest path algorithms are in principle exact and fast. In this dissertation, we use a version of Dijkstra's shortest path algorithm [23], as well as Yen's k-shortest path algorithm [24], which we use to obtain the k-th highest or lowest energy sequences. As weights of this graph we use the probabilistic model of Tompitak et al. [10] (although any short-range probabilistic/energy model would work). We can even apply this method on coding sequences, by using graphs that contain all synonymous ways to code for the same protein. Using this method, we investigate *all* positions on *all* protein-coding genes on yeast. We manage to create nucleosome positioning signals with single-bp resolution for 99.897% of all positions.

1.3 Translation speed as a third layer of information

Our finding that there is a considerable amount of freedom for nucleosome positioning signals to be encoded on top of protein-coding DNA (Chapter 3) inspired us to investigate the possiblity of a third layer of information to exist on the same piece of DNA. This third layer is the *translation speed*, a parameter that affects co-translational folding of proteins. The term co-translational folding refers to the folding of the amino acid chain which occurs at the same time as it is being created by the ribosome. The function of proteins depends on how it is folded. With co-translational folding, this process can be regulated [25]. The folding depends on the speed at which new amino acids are attached to the growing amino acid chain by the ribosome. Different synonymous codons (Table 1.1) have different translation rates, which means that there are multiple ways to code for the same protein with different translation speed landscapes. Also, translation speed is species-specific and cell-specific [26, 27]. In general, the relationship between translation speed and proper protein construction is not straightforward. 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].

In Chapter 4, we will discuss the malleability of both the mechanical and translation speed layers on top of a gene. To achieve this, we can still rely on graph representations of genes and shortest path algorithms. We use a model for translation speed created by Rudorph et al. [29]. In this model the translation speed

depends on codons only, not on neighbouring codons (as is the case for the energy in the probabilistic nucleosome model by Tompitak et al. [10]). We can add translation speed to our graphs using two different methods. The first one is *pruning*: we again start with a graph containing all sequences that code for the same protein, but we remove all nodes that would lead to a translation speed landscape that is too different from the original translation speed landscape of a gene. Secondly, we use translation speed as part of the weight of a gene in addition to nucleosome energy. We use this to investigate whether it is possible to put the human gene Tumor Necrosis Factor (one of the most-cited genes [30]) in yeast, while modifying the codons such that the nucleosome energy landscape and the translation speed landscape are similar to those in humans, while keeping the genetic code intact. We do the same for a random selection of human genes.

1.4 Multiplexing genetics and mechanics in real genomes

In the final part of the thesis we discuss *how* genetics and mechanics are multiplexed. *Multiplexing* is a term that refers to having two or more layers of information coexist on one medium. The medium, in this case, is DNA. We have shown that nucleosome signals are free to exist on top of coding regions on genes (and can even coexist with translation speed signals), but we wonder whether this really happens in nature, or whether these signals are simply encoded by noncoding parts of the genes.

This part of the dissertation is heavily influenced by the work of Tompitak et al. [9]. Their paper discusses nucleosome positioning signals in promoters among many different organisms. They investigated the nucleosome signals around the transcription start sites and found them to be very different for different organisms. They show that the strength of the nucleosome-attracting regions appears to increase with organism complexity. We too study the level of attraction of nucleosomes at the promoter sites by studying the transcription start site, but look at it from a different angle, namely: how are these signals encoded by the DNA?

In Chapters 3 and 4 we investigate nucleosome positioning on coding parts of genes, which are called exons. Genes also contain introns and UTRs, which have different functions but do not code for proteins. In theory, these regions may be responsible for the mechanical signals on DNA. To investigate the effects of exons vs. introns vs. UTRs, we have created a classification scheme for the different types of multiplexing that may exist on nucleosomes and genetics. We show that for some organisms, the signals exist mostly because the noncoding DNA is responsible, but for others we find that the coding DNA actually does code for the signal as well.

Interestingly, we find that the relative difference in nucleosome attraction between coding and noncoding regions on the DNA is, for many organisms, responsible for a considerable part of the total nucleosome signal. With the use of our classification scheme we will even provide evidence for the importance of the nucleosome signal. For the organism *rice* we find a strong nucleosome-attracting signal on top of the coding parts on its genome. Surprisingly, a large part of this signal does not come from the choice of synonymous codon but from the amino acid encoded by the DNA. We suggest that in some cases, the choice of amino acid in a protein is

not for the benefit of the protein itself but for the mechanical signal on the gene instead. It seems that genetics is *not* the all-deciding signal, on top of which the mechanical and translation speed signals would be allowed to exist only without changing the underlying genetics. For human, on the other hand, we find a much weaker mechanical signal on exons but a significant mRNA signal likely related to translation speed. We propose that, from an evolutionary point of view, all three layers of information on DNA compete with each other.

1.5 Overview

This dissertation is organized as follows. In Chapter 2 we introduce a tractable nucleosome model which qualitatively reproduces the nucleosome positioning code. We solve the model by using the transfer matrix method. Furthermore we present a method to find exactly what the contribution is of far-away neighbours to the probability of encountering a dinucleotide at any position on the nucleosome. Finally, we introduce the average neighbour energy approximation, which we use to explain the nucleosome positioning rules. In Chapter 3 we introduce and demonstrate a novel method to find sequences with special affinity for nucleosomes given any short-ranged energy/probabilistic model. This method relies on weighted graph representations of all possible nucleosome sequences. Using a k-shortest path algorithm we find the sequences with the k-th highest or lowest probability to attract a nucleosome. We demonstrate how genetics and mechanics can be multiplexed by evaluating paths in graphs of synonymous codons. By cleverly choosing the weights of these graphs, we find that nucleosomes can be placed almost anywhere on the genome of yeast by mechanical signals. Chapter 4 takes the study of multiplexing a step further by combining the analysis of genetics, mechanics and translation speed. We achieve this by adding translation speed to our graphs, either by pruning graphs or adding translation speed to our weights. These graphs enable us for example 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. Chapter 5 studies multiplexing on genomes of real organisms. By introducing a classification scheme we find and analyze the different mechanisms used by organisms to encode mechanical signals on DNA.

Chapter 2

Physics behind the mechanical nucleosome positioning code

This chapter is based on Zuiddam, Everaers and Schiessel, 2017, Phys. Rev. E. [31]

The positions along DNA molecules of nucleosomes, the most abundant DNA-protein complexes in cells, are influenced by the sequence dependent DNA mechanics and geometry. This leads to the "nucleosome positioning code", a preference of nucleosomes for certain sequence motives. In this chapter we introduce a simplified model of the nucleosome where a coarse-grained DNA molecule is frozen into an idealized superhelical shape. We calculate the exact sequence preferences of our nucleosome model and find it to reproduce qualitatively all the main features known to influence nucleosome positions. Moreover, using well-controlled approximations to this model allows us to come to a detailed understanding of the physics behind the sequence preferences of nucleosomes.

2.1 Introduction

The DNA double helix carries, in addition to the classical genetic information (the genes encoding for the proteins), a mechanical layer of information. This is possible because the mechanical properties of DNA depend on the underlying sequence of base pairs (bp). Certain combinations of letters (especially bp steps) are softer than others and some cause intrinsic bends on the DNA molecule [32]. So unlike in a book where the stiffness of the paper does not depend on the text printed, DNA elasticity and geometry is intimately linked to the text it carriers.

Possibly the most important biological consequence of sequence dependent DNA mechanics is its impact on the positioning of DNA spools, called nucleosomes. The core of each spool is a cylinder composed of eight histone proteins and it is wrapped by a DNA stretch of 147 bp length. A short stretch of unbound DNA, the linker DNA, connects to the next protein spool. It is known from the nucleosome crystal structure [33] that the DNA is bound to the protein core at 14 locations where the minor groove of the DNA double helix faces the cylinder. This defines the binding

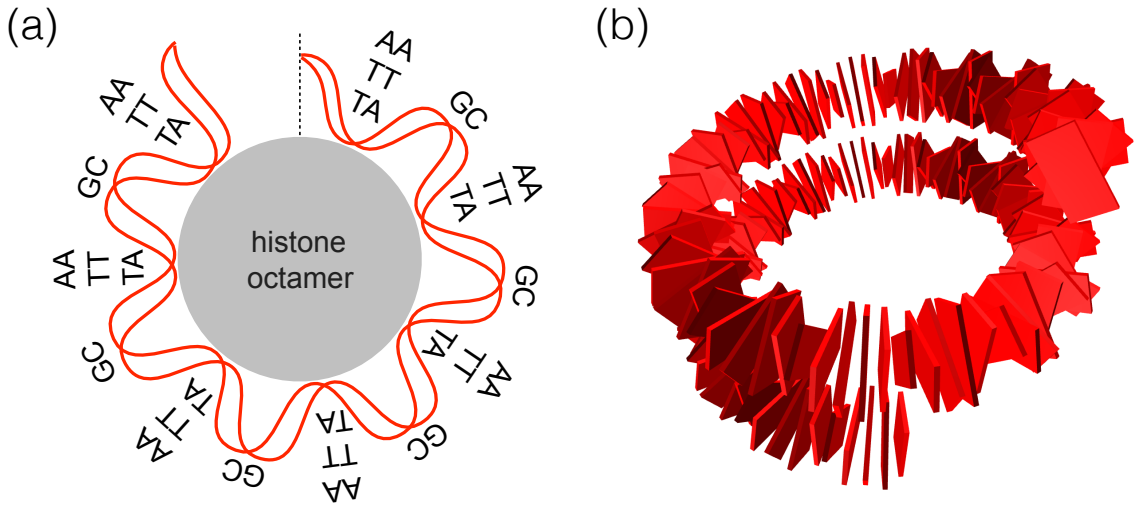


Figure 2.1: (a) The probability of finding GC steps peaks at positions where the major groove of the DNA faces the histone octamer (every 10^{th} bp) whereas TT, AA and TA are all in phase and have their peaks in between where the minor groove faces the cylinder. These are key rules from the so-called "nucleosome positioning code". (b) Visual representation of our model for nucleosomal DNA. Base pairs represented by rigid plates are frozen in an idealized superhelical shape.

path, a left-handed superhelix of one and three quarter turns.

This structure makes nucleosomes ideal "readers" of mechanical cues. Firstly, the length that is wrapped in a nucleosome is about one persistence length, 50 nm. It follows that the bending energy is much larger (about 60 times [34]) than the thermal energy. Thus even a small change in the wrapped bp sequence is expected to have a strong effect on the nucleosome affinity. Secondly, as the binding to the histone octamer occurs mostly with the two backbones of the DNA double helix, there is no direct readout of the sequence but instead the nucleosome affinity results from the elasticity and geometry of the involved DNA stretch.

It is indeed known from various experiments that nucleosomes have sequence preferences [4–6]. High affinity sequences show certain motifs along the wrapped DNA. This "nucleosome positioning code" is typically formulated in terms of bp steps or, looking along one strand, dinucleotides: most importantly, the probability of finding GC steps (nucleotide G followed by nucleotide C) peaks at positions where the major groove faces the protein cylinder (every 10^{th} bp) whereas TT, AA and TA are all in phase and have their peaks in between where the minor groove faces the cylinder (see Fig. 2.1(a)).

Over evolutionary time scales mechanical signals have evolved along genomes. Examples are nucleosome depleted regions at transcription start sites in yeast facilitating transcription initiation [6, 9], mechanically encoded retention of a small fraction of nucleosomes in human sperm cells allowing transmission of paternal epigenetic information [9, 35] or the positioning of six million nucleosomes around nucleosome inhibiting barriers in human somatic cells [8].

However, what is still missing is a deeper understanding of the physics underlying nucleosome positioning rules. An example, mentioned in [14], are the positions where GC steps typically occur in high-affinity sequences. These correspond to positions

that GC steps dislike the most. Even more remarkably, of *all* 16 bp steps it is the GC step that is energetically most costly at these positions.

A first step toward understanding the nucleosome positioning rules is using coarse grained DNA models with sequence dependent elasticity and force them into shapes that resemble the wrapped DNA portions in nucleosomes. Several such models use the so-called rigid base pair model [12, 13] in which the conformation of a DNA molecule is described by the positions and orientations of its base pairs that are modelled as rigid objects. These nucleosome models have been used to predict nucleosome stability and positioning [15–17, 20–22, 36], forces and torques on the wrapped DNA [37], nucleosome mobility along DNA [18] and the response of nucleosomes to external forces [19]. One recent study [14] specifically addresses the question whether such models can predict the above mentioned rules of the nucleosome positioning code. This was achieved by introducing the Mutation Monte Carlo method, which mixes conformational and sequence moves. This method automatically produces the sequence preferences along the wrapped DNA and it was indeed found that it reproduces the nucleosome positioning rules. However, the model is still far too complex to really come to a clear interpretation of how the rules result from the underlying elasticity and geometry of the DNA.

Here we overcome this complexity by reducing the model to its bare essentials: we consider a piece of DNA that is forcibly curved and idealize the shape by placing it on a superhelical path (Fig. 2.1(b)). Assuming such an idealized shape (as done in [20–22, 36]) instead of trying to imitate details of the crystal structure (as done in [14–19]) makes our model analytically tractable and allows us to pinpoint the dominant contributions that underlie the positioning code. Moreover we freeze the model into this configuration, unlike in some models where the base pairs are free to move with respect to others (at some energy cost) [14, 17–19, 36]. Variants of our approach are in principle applicable to any model that freezes the DNA into a fixed configuration like it is done in [15, 16, 20–22].

The goal of this chapter is not to come up with yet another tool for nucleosome positioning. Based on the more complete model [14] we were able to build a probabilistic model that is as fast as the model introduced here and is very successful in predicting nucleosome positioning [9]. The goal of the current work is instead to come to a deep understanding of the positioning rules. For instance, we will be able to explain what cause GC steps to "favour" the most costly positions on the wrapped DNA. To achieve this an analytical approach as presented here is indispensable.

In the next section we introduce our model. In Section 2.3 we explain how it can be solved using transfer matrices. This is followed by two sections that develop approximations that allow to come to a detailed understanding of the nucleosome positioning rules: in Section 2.4 we take a limited number of neighbours around the given base pair step into account to derive upper and lower bounds for the probabilities of its occurrence, and in Section 2.5 we introduce the average neighbour energy approximation, an effective approximation for interpreting nucleosome positioning rules. The exact dinucleotide probabilities, approximations to them and an interpretation of the rules is presented in Section 2.6, and a conclusion is provided in the final section.

	q^{roll} [rad]	Q^{roll} $\left[\frac{k_B T_r}{\text{rad}^2}\right]$	q^{tilt} [rad]	Q^{tilt} $\left[\frac{k_B T_r}{\text{rad}^2}\right]$
AA	0.012410451	126.98464	-0.024820902	207.73324
AT	0.019409417	148.42141	0	216.86174
AC	0.012372536	143.15931	-0.0017675051	221.16218
AG	0.079562987	123.91326	-0.030057128	200.28179
TA	0.058653564	73.527282	0	129.10674
TT	0.012410451	126.98464	0.024820902	207.73324
TC	0.03372236	113.06128	0.026622916	210.62471
TG	0.083496463	97.396194	-0.0088826025	146.17762
CA	0.083496463	97.396194	0.0088826025	146.17762
CT	0.079562987	123.91326	0.030057128	200.28179
CC	0.063703201	130.1586	0.0017695334	225.01953
CG	0.095824007	83.019248	0	150.88272
GA	0.03372236	113.06128	-0.026622916	210.62471
GT	0.012372536	143.15931	0.0017675051	221.16218
GC	0.0053117746	146.67053	0	214.38125
GG	0.063703201	130.1586	-0.0017695334	225.01953

Table 2.1: Parametrization used to calculate the dinucleotide energy, Eq. 2.5 to Eq. 2.7. The symbols q and Q denote the intrinsic value and the stiffness of roll or tilt.

2.2 Model

Our model is based on the rigid base pair model [12, 13], a coarse grained representation of the DNA double helix, that treats the base pairs as rigid plates. Neighbouring plates differ by six degrees of freedom called shift, slide, rise, roll, tilt and twist. The rotational degrees of freedom, roll, tilt, and twist, are shown in Fig. 2.2. We force this DNA model into a superhelix to mimic the bending of the DNA inside a nucleosome, neglecting the non-uniform bending of the nucleosomal DNA observed in its crystal structure [33]. As the general nucleosome positioning rules hold all along the wrapped part [5], we expect that these simplifications do not affect the rules whose origin we aim to understand here. In addition, motivated by the observation that the basic nucleosome positioning rules can be rationalized by discussing energy costs involved in the roll and tilt degrees of freedom [14], we only account for them and neglect contributions from the other degrees of freedom. This makes the model easier to analyse. The contribution of twist and any cross terms between the rotational degrees of freedom will be discussed in Appendix A.1. There we will show that neglecting these terms does not affect the main positioning rules of our model.

The rigid base pair model assumes only nearest-neighbor interactions and places a quadratic deformation energy between successive base pairs with bp step dependent stiffnesses and intrinsically preferred configurations. We use in the following the hybrid parametrization, where the intrinsic values are derived from protein-DNA crystals and the stiffnesses from atomistic molecular simulations [38], see Table 2.1 for a list of the parameters for roll and tilt.

In order to calculate the difference between the preferred and the *actual* configuration, we need to formally define the shape of our superhelix. We consider a superhelix with pitch P and radius R (similar to Morozov et al., Ref. [36]):

$$\vec{r}(s) = [R\cos(s/R_{\text{eff}}), R\sin(s/R_{\text{eff}}), -(P/2\pi R_{\text{eff}})s], \tag{2.1}$$

with $R_{\text{eff}} = \sqrt{R^2 + (P/2\pi)^2}$. The set of Frenet-Serret vectors at position s on the superhelix are given by

$$[\hat{t}(s), \hat{n}(s), \hat{b}(s)] = \left[\frac{d\vec{r}}{ds}, \frac{d\vec{t}}{ds} \Big/ \left|\frac{d\vec{t}}{ds}\right|, \vec{t} \times \vec{n}\right] \tag{2.2}$$

where $\hat{t}$ is the tangent unit vector, $\hat{n}$ the principal normal unit vector and $\hat{b}$ the binormal unit vector.

The rotational orientation of a base pair plate, compared to the origin, can be described using the three orthonormal vectors $\hat{x}, \hat{y}, \hat{z}$, see Fig. 2.2(a). We place the double helical shape of the DNA on the superhelix by defining the orthonormal vectors with respect to the Frenet-Serret vectors, such that the double helix revolves (twists) right-handedly around the superhelix:

$$\begin{aligned}[\hat{x}(p), \hat{y}(p), \hat{z}(p)] =& [\hat{n}(s)\cos(\theta p + \phi) - \hat{b}(s)\sin(\theta p + \phi), \\ & -\hat{n}(s)\sin(\theta p + \phi) - \hat{b}(s)\cos(\theta p + \phi), \hat{t}(s)],\end{aligned} \tag{2.3}$$

with $p = s(L-1)/(2\pi R_{\text{eff}}\alpha) + 1/2$ the positions of the dinucleotide (right in between two plates), where α denotes the number of superhelical turns and L the number of base pairs wrapped around the nucleosome. The constants θ and ϕ determine how much the double helix is twisted, and which positions correspond to maximum/minimum roll and tilt. To reflect the approximately 10 bp helical pitch of the DNA inside the nucleosome, we set $\theta = 2\pi/10$. The phase ϕ is set to $-147\pi/10$ such that the bp at the central position between dinucleotide steps 73 and 74 corresponds to the position of maximal roll, in accordance with the fact that at that position the major groove faces the histone octamer. This is also the place where the tilt changes sign from negative to positive values.

The convention we use to calculate the roll, tilt, and twist degrees of freedom from the orientation of the plates has been well-explained in the literature [39] and will not be discussed here. We will provide the (numerical) results of this method, as well as a short explanation of the values. Using $P = 25.9\,\text{Å}$, $R = 41.9\,\text{Å}$, $\alpha = 1.84$, and $L = 147$ [36], we find expressions for the angles q_p^i, $i \in \{\text{roll}, \text{tilt}, \text{twist}\}$ given by:

$$\begin{aligned}[q_p^{\text{roll}}, q_p^{\text{tilt}}, q_p^{\text{twist}}] =& [\Gamma\cos(2\pi p/10 - 147\pi/10), \\ & \Gamma\sin(2\pi p/10 - 147\pi/10), q^{\text{twist}}],\end{aligned} \tag{2.4}$$

with $\Gamma \approx 0.0796$ rad and $q^{\text{twist}} \approx 10.17/(2\pi)$ rad. These values can be rationalized the following way. Our superhelix has constant curvature, and as a result, a constant angle between each dinucleotide pair, to which roll and tilt make equal contributions [36]. This angle is given by $\arccos\{\vec{t}[s(p)] \cdot \vec{t}[s(p+1)]\} \approx 0.0788$, which is a great

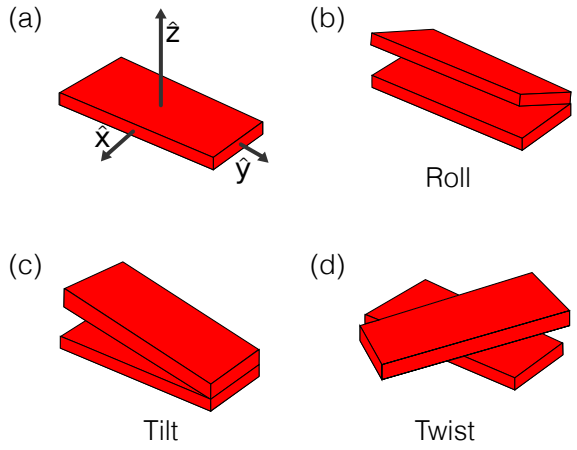


Figure 2.2: The rotational degrees of freedom between neighboring bp in the rigid base pair model. Each base pair has a coordinate system (a) which can be used to describe the relative orientation between two plates. In our model we account only for energy contributions from (b) roll and (c) tilt but neglect contributions from (d) twist. Also the translational degrees are not considered.

approximation for our value of Γ. The twist q^{twist} we report is lower than the value for θ we defined. While roll and tilt are 10 bp periodic, the twist corresponds with a periodicity of 10.17. This may seem counter-intuitive. However, if the twist were equal to $2\pi/10$, the configuration of the plates would be a ring instead of a superhelix.

As mentioned before, we only account for two degrees of freedom and also neglect cross terms between them. Hence the energy of placing a dinucleotide step $a, b \in \{A, T, C, G\}$ at position p is the sum of the roll and tilt energies:

$$E_p(a,b) = E_p^{\text{roll}}(a,b) + E_p^{\text{tilt}}(a,b) \tag{2.5}$$

with

$$E_p^{\text{roll}}(a,b) \equiv \frac{1}{2} Q^{\text{roll}}(a,b) \left[q_p^{\text{roll}} - \bar{q}^{\text{roll}}(a,b)\right]^2, \tag{2.6}$$

and

$$E_p^{\text{tilt}}(a,b) \equiv \frac{1}{2} Q^{\text{tilt}}(a,b) \left[q_p^{\text{tilt}} - \bar{q}^{\text{tilt}}(a,b)\right]^2. \tag{2.7}$$

The bp-step dependent stiffnesses in the roll and tilt degrees of freedom are given by $Q^{\text{roll}}(a,b)$ and $Q^{\text{tilt}}(a,b)$ and the corresponding intrinsic values by $\bar{q}^{\text{roll}}(a,b)$ and $\bar{q}^{\text{tilt}}(a,b)$.

2.3 Dinucleotide probabilities

Here we calculate the dinucleotide probability distribution along our nucleosome model. Base pair steps are the mechanical units in our model and also the experimentally observed nucleosome sequence preferences are typically formulated in terms of dinucleotides [5, 32]. We therefore aim to obtain the probability of having nucleotides a and b at dinucleotide position p on the DNA molecule of length $L = 147$. The nucleotides are numbered from 1 to L, such that the p^{th} dinucleotide

position contains nucleotides p and $p+1$. The probability does not merely depend on the energy stored between a and b. These bases are connected to other bases as well. In order to find the probability we need to sum over all possible DNA strands containing a and b at position p, and divide by the partition sum. Therefore the probability is given by

$$P_p(a,b) = \frac{\sum\limits_{\substack{n_1,\ldots,n_L \\ n_p=a, n_{p+1}=b}} \exp\left[-\beta \sum\limits_{i=1}^{L-1} E_i(n_i, n_{i+1})\right]}{\sum\limits_{n_1,\ldots,n_L} \exp\left[-\beta \sum\limits_{i=1}^{L-1} E_i(n_i, n_{i+1})\right]} \tag{2.8}$$

where we sum over all possible states $n_i \in \{A, T, C, G\}$, with β the inverse temperature. The probability given by Eq. 2.8 corresponds to the case where the nucleosomal DNA sequence mutates freely. This is distinct from the scenario where various DNA stretches compete for nucleosomes, as it is typically the case in experiments such as Ref. [4–6]. Then also entropic effects play a role (e.g., softer bp steps prefer to reside outside nucleosomes for entropic reasons). However, our model is also a reasonable approximation to this case since this system is energy-dominated for physiological temperatures (and lower). In this study, we therefore consider only energies but neglect entropic contributions associated with conformational degrees of freedom.

This type of probabilities can be evaluated using *transfer matrices*. Transfer matrix formalisms have been used both in the context of calculating dinucleotide probabilities for a single nucleosome and evaluating many-nucleosome systems [5, 36, 38, 40] (see Ref. [41] for an overview).

We define the position-dependent transfer matrix T_i in the basis $B = \{|A\rangle, |T\rangle, |C\rangle, |G\rangle\}$ such that

$$\langle n|T_i|m\rangle \equiv \exp\left[-\beta E_i(n,m)\right] \tag{2.9}$$

with $|n\rangle, |m\rangle \in B$. This allows us to rewrite the probability as

$$P_p(a,b) = \frac{\sum\limits_{n_1,n_L} \langle n_1| T_1 \ldots T_{p-1} |a\rangle \langle a|T_p|b\rangle \langle b| T_{p+1} \ldots T_{L-1} |n_L\rangle}{\sum\limits_{n_1,n_L} \langle n_1| T_1 \ldots T_{p-1} T_p T_{p+1} \ldots T_{L-1} |n_L\rangle}. \tag{2.10}$$

Finding this probability involves multiplying $L-1 = 146$ four-by-four transfer matrices in the nominator and denominator.

While this quantity is easy to calculate, the sheer number of terms makes it hard to determine which terms influence the probability most and which terms can be neglected. It seems reasonable that bases at positions far away from position p are not as important to the probability as its close neighbours, e.g. at positions $p+1$ and $p-1$. In the next section we will show this by quantifying the effect that far-away bases can possibly have on the probability.

2.4 Bounds of dinucleotide probabilities

Here we show how much the probability $P_p(a,b)$ can be affected by the energies of nucleotides some steps away from the position p. In the following we quantify the effect by calculating k^{th}-order bounds of the probability, which we obtain using only the energies of k bases to the left and k bases to the right of the dinucleotide at position p. We assume that all the 'unused' bases either try to make the probability $P_p(a,b)$ as high or as low as possible. This is done by substituting all terms related to the unused bases on the left by $\langle x_k|$, and the terms related to unused bases on the right by $|y_k\rangle$. The probability for $k \geq 1$ is then given by

$$P_p(a,b) = \frac{\langle x_k| \prod_{i=p-k}^{p-1} T_i \, |a\rangle \, \langle a|T_p|b\rangle \, \langle b| \prod_{j=p+1}^{p+k} T_j \, |y_k\rangle}{\langle x_k| \prod_{i=p-k}^{p+k} T_i \, |y_k\rangle} \tag{2.11}$$

with

$$\langle x_k| \equiv \frac{1}{c_k} \sum_n \langle n| \, T_1 T_2 ... T_{p-k-2} T_{p-k-1} \tag{2.12}$$

and

$$|y_k\rangle \equiv \frac{1}{d_k} \sum_n T_{p+k+1} T_{p+k+2} ... T_{L-2} T_{L-1} \, |n\rangle \, , \tag{2.13}$$

where c_k and d_k are normalization constants such that $|\, \langle x_k|x_k\rangle \,| = 1$ and $|\, \langle y_k|y_k\rangle \,| = 1$. Note that $\langle x_k|$ and $|y_k\rangle$ implicitly depend on p.

To find the k^{th}-order bounds on the probability, we assume that we know nothing about $\langle x_k|$ or $|y_k\rangle$ other than that they represent physically possible states. We formally define the k^{th}-order upper/lower bound by taking the maximum/minimum of Eq. 2.11 where we let $\langle x_k|$ and $|y_k\rangle$ run over all their possible states. Because the transfer matrix contains Boltzmann weights only, all entries in the transfer matrices T_i are positive. From this it follows that $|x_k\rangle = \sum_{n\in\{A,T,C,G\}} x_{n,k} \, |n\rangle$ and $|y_k\rangle = \sum_{n\in\{A,T,C,G\}} y_{n,k} \, |n\rangle$ with $0 < x_{n,k} \leq 1$, $0 < y_{n,k} \leq 1$. These equations are equivalent to the quantum mechanical representation of *mixed states*. The probabilities to encounter the four possible bases k positions to the left and right of dinucleotide a,b are weighted by $x_{n,k}$ and $y_{n,k}$, parameters that depend on the energy costs of bases further away.

It turns out that one finds the minimally and maximally possible value of the probability when $|x_k\rangle$ and $|y_k\rangle$ are *pure states*, states from the basis $B = \{|A\rangle \, , |T\rangle \, , |C\rangle \, , |G\rangle\}$. Pure states correspond to *exactly* knowing which bases are present k bases to the left and to the right of the dinucleotide a,b. (Strictly speaking, this happens only when the energy costs of encountering the other possible bases are infinitely high. In other words, this is a limiting case.)

Since, as we prove below, the minimally and maximally possible value of the probability is found when $|x_k\rangle$ and $|y_k\rangle$ are pure states, one can compute the k^{th}-order upper and lower bounds of the probability, $P^{(k)}_{\text{max},p}(a,b)$ and $P^{(k)}_{\text{min},p}(a,b)$, by

simply evaluating the probability for all 16 possible combinations of pure states. This leads to the expressions

$$P^{(k)}{}_{\max,p}(a,b) = \max_{|x_k^*\rangle,|y_k^*\rangle\in B} \frac{\langle x_k^*| \prod_{i=p-k}^{p-1} T_i \,|a\rangle \langle a|T_p|b\rangle \langle b| \prod_{j=p+1}^{p+k} T_j \,|y_k^*\rangle}{\langle x_k^*| \prod_{i=p-k}^{p+k} T_i \,|y_k^*\rangle} \tag{2.14}$$

and

$$P^{(k)}{}_{\min,p}(a,b) = \min_{|x_k^*\rangle,|y_k^*\rangle\in B} \frac{\langle x_k^*| \prod_{i=p-k}^{p-1} T_i \,|a\rangle \langle a|T_p|b\rangle \langle b| \prod_{j=p+1}^{p+k} T_j \,|y_k^*\rangle}{\langle x_k^*| \prod_{i=p-k}^{p+k} T_i \,|y_k^*\rangle}. \tag{2.15}$$

We prove now the expression for the k^{th}-order upper bound of the probability (the proof for the lower bound can be obtained analogously). We substitute $|x_k\rangle = \sum_{n\in\{B\}} x_{n,k}\,|n\rangle$ and $|y_k\rangle = \sum_{m\in\{B\}} y_{m,k}\,|m\rangle$ into Eq. 2.11. To prove Eq. 2.14, we need to show that one finds the largest possible value for the probability when $x_{n,k}$ and $y_{m,k}$ are zero for all n, m except for one value of n and m. For convenience, we define $\bar{\mathcal{T}}_{nm} \equiv \langle n| \prod_{i=p-k}^{p-1} T_i \,|a\rangle \langle a|T_p|b\rangle \langle b| \prod_{j=p+1}^{p+k} T_j \,|m\rangle$ and $\mathcal{T}_{nm} \equiv \langle n| \prod_{i=p-k}^{p+k} T_i \,|m\rangle$ for $n, m \in B$. The probability can then be stated as

$$P_p(a,b) = \frac{\sum_{n\in B}\sum_{m\in B} x_{n,k}\bar{\mathcal{T}}_{nm}y_{m,k}}{\sum_{n\in B}\sum_{m\in B} x_{n,k}\mathcal{T}_{nm}y_{m,k}}. \tag{2.16}$$

Without loss of generality, we assume that

$$\frac{\bar{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}} = \min\left(\frac{\bar{\mathcal{T}}_{AA}}{\mathcal{T}_{AA}}, \frac{\bar{\mathcal{T}}_{AT}}{\mathcal{T}_{AT}}, ..., \frac{\bar{\mathcal{T}}_{GG}}{\mathcal{T}_{GG}}\right) \tag{2.17}$$

holds for some $i, j \in B$, which does not have to be unique. We evaluate the sign of the derivative of $P_p(a,b)$ with respect to $x_{i,k}y_{j,k}$:

$$\begin{aligned}\frac{\partial P_p(a,b)}{\partial(x_{i,k}y_{j,k})} =& \frac{\sum_{n\in B}\sum_{m\in B} x_{m,k}\mathcal{T}_{nm}y_{n,k}\mathcal{T}_{ij}\left(\frac{\bar{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}} - \frac{\bar{\mathcal{T}}_{nm}}{\mathcal{T}_{nm}}\right)}{\left(\sum_{n\in B}\sum_{m\in B} x_{n,k}\mathcal{T}_{nm}y_{m,k}\right)^2} \\ \leq& 0.\end{aligned} \tag{2.18}$$

The less-than or equal to sign follows from the fact that $\tilde{\mathcal{T}}_{nm}$, $\mathcal{T}_{nm}$, $x_{m,k}$ and $y_{n,k}$ are non-negative for all n, m and from Eq. 2.17. Because the derivative is non-positive, the probability is non-increasing as a function of $x_{i,k}y_{j,k}$, thus a maximum can be found when $x_{i,k}y_{j,k}$ is minimal, i.e., in the limit of $x_{i,k}y_{j,k} \to 0$. Now we have 'eliminated' one combination of variables: $x_{i,k}y_{j,k}$, and the corresponding ratio $\frac{\tilde{\mathcal{T}}_{ij}}{\mathcal{T}_{ij}}$ from Eq. 2.16) (this can be checked by inserting $x_{i,k}y_{j,k} = 0$ in Eq. 2.16). This process can be performed iteratively until only one combination of variables is left. Now we assume, again without loss of generality, that this final combination is $x_{r,k}y_{s,k}$ for some $r, s \in B$. The probability is now independent of these variables:

$$P^{(k)}_{\max,p}(a,b) = \frac{x_{r,k}\bar{\mathcal{T}}_{rs}y_{s,k}}{x_{r,k}\mathcal{T}_{rs}y_{s,k}} = \frac{\bar{\mathcal{T}}_{rs}}{\mathcal{T}_{rs}}. \tag{2.19}$$

This does not mean we can freely assign a number to $x_{r,k}y_{s,k}$. Recall that $|x_k\rangle$ and $|y_k\rangle$ are unit vectors. Since $x_{m,k}y_{n,k} \to 0$ for all $m \neq r, n \neq s$, it is required that $x_{r,k} \to 1$ and $y_{s,k} \to 1$, and $x_{m,k} \to 0$ and $y_{n,k} \to 0$ for all $m \neq r, n \neq s$. Therefore, we find the k^{th} upper bound of the probability when $|x_k\rangle$ and $|y_k\rangle$ are pure states from the basis B, as we stated in Eq. 2.14.

For the zeroth-order bounds, where no neighbours are taken into account, a similar result holds. This can be obtained in the same manner as Eq. 2.14 and Eq. 2.15, therefore no proof is provided. These bounds are given by

$$P^{(0)}_{\max,p}(a,b) = \max_{|x_0^*\rangle,|y_0^*\rangle \in B} \frac{\langle x_0^*|a\rangle \langle a|T_p|b\rangle \langle b|y_0^*\rangle}{\langle x_0^*| T_p |y_0^*\rangle} = 1 \tag{2.20}$$

and

$$P^{(0)}_{\min,p}(a,b) = \min_{|x_0^*\rangle,|y_0^*\rangle \in B} \frac{\langle x_k^*|a\rangle \langle a|T_p|b\rangle \langle b|y_k^*\rangle}{\langle x_k^*| T_p |y_k^*\rangle} = 0. \tag{2.21}$$

These bounds are 1 and 0 because $\min_{|x_0^*\rangle,|y_0^*\rangle \in B} \langle x_k^*|a\rangle = 0$ and $\max_{|x_0^*\rangle,|y_0^*\rangle \in B} \langle x_k^*|a\rangle = 1$. This shows that one needs to take at least one neighbour into account to obtain non-trivial results.

Furthermore, one can show that the bounds on the probability get sharper at higher order, i.e., increasing k:

$$P^{(k)}_{\max,p}(a,b) \geq P^{(k+1)}_{\max,p}(a,b) \geq P_p(a,b) \tag{2.22}$$

$$P_p(a,b) \geq P^{(k+1)}_{\min,p}(a,b) \geq P^{(k)}_{\min,p}(a,b). \tag{2.23}$$

An intuitive explanation is that adding the information on more and more bases to our calculation should lead to sharper bounds on the probability. It is straighforward to prove. Consider Eq. 2.14 for the $(l+1)^{\text{th}}$ order upper bound (such that $k = l+1$) with its two maximizing pure states $\langle x_{l+1}^*| = \langle n|$ and $|y_{l+1}^*\rangle = |m\rangle$. This bound is smaller or equal to the upper bound for $k = l$ for the following reason: one finds exactly the same expression as above if one inserts into Eq. 2.14 the states $\langle x_l^*| = \langle n| T_{p-l}$ and $|y_l^*\rangle = T_{p+l}|m\rangle$. We find the l^{th}-order upper bound if $\langle x_l^*|$ and $|y_l^*\rangle$ are pure states. Using other values, i.e., $\langle n| T_{p-l}$ and $T_{p+l}|n\rangle$, can only result in probabilities equal to or lower than this l^{th}-order maximum. Therefore, the $(l+1)^{\text{th}}$-order upper bound cannot be higher than the l^{th}-order upper bound. The same reasoning holds for the lower bounds.

2.5 The average neighbour energy approximation

The method of finding bounds on the probability in the previous section allows us to quantify how much nucleotides a given number of steps away from a given position can affect the dinucleotide preferences at that position. By comparing the results of the bounds on the probability at different orders we will show in the next section that long-range interactions are unimportant. On the other hand, we will also find that a purely local picture where the probability of a dinucleotide is determined only by its own elastic properties is not predictive. Even the first-order bounds on the probability that take the nearest neighbours into account, are too far apart to confine sufficiently the position-dependent variations of the probabilities. It is the difference between the second-order upper and lower bounds that is much smaller than these variations. This demonstrates that only a limited number of neighbours determines the nucleosome positioning rules.

Here we further expand on this idea by showing that, for our model at room temperature, the probability of finding a dinucleotide at a given position p mostly depends on only two parameters: the energy of the dinucleotide at position p, and the sum of the averages of the energies of their possible neighbours at positions $p+1$ and $p-1$. Looking at these two parameters allows us to interpret the base pair step preferences in our nucleosome model. We will call the corresponding approximation the *average neighbour energy approximation*. This approximation will be used later, not to calculate probabilities but to give a physical interpretation of our findings from the exact treatment.

Since the first-order bounds on the probability are not good enough to confine the dinucleotide preferences, it may seem counter-intuitive to use only the nearest neighbours. This can be explained by the fact that the upper and lower bounds on the probability take extreme scenarios into account where the neglected nucleotides have the highest possible impact on the probability, whereas the actual system does not behave as extremely.

We introduce now the approximated probability that we indicate by a superscript (e) as follows: $P_p^{(e)}(a,b)$. Using the notation

$$\langle f(x)\rangle_x = \frac{1}{4}\sum_{x\in\{A,T,C,G\}} f(x), \tag{2.24}$$

$$\langle g(x,y)\rangle_{x,y} = \frac{1}{16}\sum_{x,y\in\{A,T,C,G\}} g(x,y), \tag{2.25}$$

we define the average neighbour energy approximation of the probability as

$$P_p^{(e)}(a,b) \equiv \frac{\begin{array}{l}\exp\left[-\beta\langle E_{p-1}(n_{p-1},a)\rangle_{n_{p-1}}\right] \\ \quad\times\exp\left[-\beta E_p(a,b)\right] \\ \quad\quad\times\exp\left[-\beta\langle E_{p+1}(b,n_{p+2})\rangle_{n_{p+2}}\right]\end{array}}{\sum\limits_{n_p,n_{p+1}}\begin{array}{l}\exp\left[-\beta\langle E_{p-1}(n_{p-1},n_p)\rangle_{n_{p-1}}\right] \\ \quad\times\exp\left[-\beta E_p(n_p,n_{p+1})\right] \\ \quad\quad\times\exp\left[-\beta\langle E_{p+1}(n_{p+1},n_{p+2})\rangle_{n_{p+2}}\right]\end{array}}. \tag{2.26}$$

Note that this approximation depends on $E_p(a,b)$, the energy of the dinucleotide step ab at position p, and on $\langle E_{p-1}(n_{p-1},a)\rangle_{n_{p-1}} + \langle E_{p+1}(b,n_{p+2})\rangle_{n_{p+2}}$, an average of the energies of possible nearest neighbours of ab. We have calculated the error introduced by using the average neighbour energy approximation and found it not to be larger than 3.5 percent at any position for any dinucleotide, see Appendix A.2.

Next we provide an explanation why this approximation works so well for our model. Our strategy is to bring the approximated probability, Eq. 2.26, and the full probability, Eq. 2.8, into a similar form. Comparison of the two similar expressions allows then to explain the nature of this approximation that is otherwise not straightforward to see. We start by rewriting the approximation such that it resembles more the exact probability (Eq. 2.8):

$$P_p^{(e)}(a,b) = \frac{\sum\limits_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}: \\ n_p=a,n_{p+1}=b}} \exp\Big[-\beta\sum_{i=1}^{L-1}\langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\Big]}{\sum\limits_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}}} \exp\Big[-\beta\sum_{i=1}^{L-1}\langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\Big]}. \tag{2.27}$$

The hats above n_{p-1} and n_{p+2} denote that these variables are not to be summed over. The nominator factorises in three terms: (1) a sum of terms where each term depends explicitly on at least one of the variables n_1 to n_{p-2}, (2) a sum of terms where each term depends explicitly on at least one of the variables n_{p+3} to n_L, and terms independent of those variables. The first and second factors cancel out with the exact same expressions in the denominator leading back to Eq. 2.26.

We will now make the exact probability (Eq. 2.8) look more like the approximation in the form of (Eq. 2.27). By substituting the function $C_p(i,j)$, defined as

$$C_p(m,o) \equiv \frac{\frac{1}{4}\sum\limits_n \exp\left[-\beta E_{p+1}(m,n) - \beta E_{p+2}(n,o)\right]}{\exp\left[-\beta\langle E_{p+1}(m,n) + E_{p+2}(n,o)\rangle_n\right]}, \tag{2.28}$$

into Eq. 2.8 twice, we obtain

$$P_p(a,b) = \frac{\sum\limits_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}: \\ n_p=a,n_{p+1}=b}} \exp\Big[-\beta \sum_{i=1}^{L-1} \langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\Big] \times C_{p-2}(n_{p-2},a) C_{p+1}(b,n_{p+3})}{\sum\limits_{\substack{n_1,\ldots,n_L \\ \widehat{n}_{p-1},\widehat{n}_{p+2}}} \exp\Big[-\beta \sum_{i=1}^{L-1} \langle E_i(n_i,n_{i+1})\rangle_{n_{p-1},n_{p+2}}\Big] \times C_{p-2}(n_{p-2},n_p) C_{p+1}(n_{p+1},n_{p+3})}, \tag{2.29}$$

which is indeed very similar to Eq. 2.27, apart from the functions C_p. The approximation $P_p(a,b) \approx P_p^{(e)}(a,b)$ is *exact* if $C_{p-2}(n_{p-2},a)$ does not depend on a, and if $C_{p+1}(b,n_{p+3})$ does not depend on b. The approximation works well if these functions show only a weak dependence on a and b. It turns out that (for our model) the latter is true, see Appendix A.2 for details.

The approximation gets worse with decreasing temperature. We can see this by performing a Taylor expansion in β of $C_p(m,o)$:

$$C_p(m,o) \approx 1 + \frac{1}{2}\beta^2 \big\langle \big[E_{p+1}(m,n) + E_{p+2}(n,o) - \langle E_{p+1}(m,n') + E_{p+2}(n',o)\rangle_{n'}\big]^2 \big\rangle_n. \tag{2.30}$$

Only the higher-order terms depend on m and o; these terms become increasingly important with decreasing temperature (increasing β). At room temperature the higher-order terms are not important as the various dinucleotide energies lie close to each other compared to the thermal energy. As a result the exponential of the averages is a good approximation to the average of the exponentials and $C_p(m,o)$ shows only a weak dependence on m and o.

2.6 Results

2.6.1 The dinucleotide probability

Using the transfer matrix approach we calculate here the preferences of dinucleotide steps along our nucleosome model. We focus in this section on the "nucleosome positioning code" [5] which claims that high affinity sequences are characterized by the proper positioning of four dinucleotides: the probability of finding GC steps (a G followed by a C) peaks at positions where the major groove faces the protein cylinder (every 10^{th} bp) whereas AA, TA, and TT are all in phase and have their peaks in between where the minor groove faces the cylinder.

Fig. 2.3(a) shows the combined probability to encounter AA, TA, TT along the nucleosome and, separately, that of GC calculated using transfer matrices, Eq. 2.10. Both signals are 10 bp periodic in accordance with the experimental observation. Moreover, the two probabilities show the right phases: the GC signal has a peak in the center (at the nucleosomal dyad) which corresponds to a place where the major groove faces inward and the same holds for all other peaks of GC. The combined

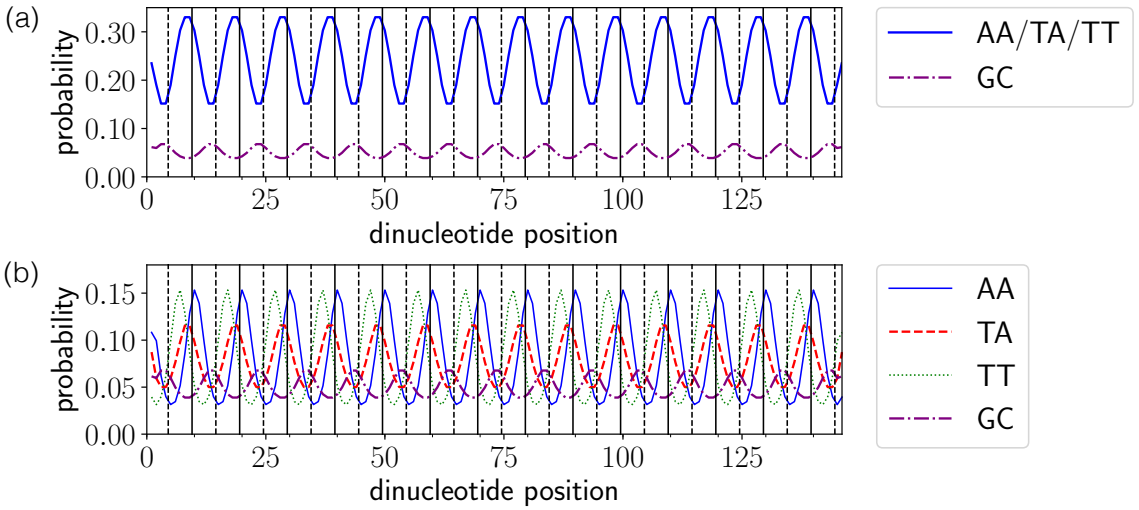


Figure 2.3: (a) The probability to find AA, TA or TT, and the probability to encounter GC at the full range of dinucleotide positions is shown. The solid and dashed vertical lines indicate minor and major bending sites (maximum negative and positive roll, respectively). The probabilities are in qualitative agreement with the well-known nucleosome positioning rules [5]. (b) Same as (a) but showing all four dinucleotide probabilities individually.

signal of AA, TA, and TT is out of phase with the GC signal and peaks at the places where the minor groove is compressed. In short, our model reproduces qualitatively the well-known nucleosome positioning rules.

More details provides Fig. 2.3(b) where all four dinucleotides are plotted separately. The figure shows that indeed AA, TA, and TT are all in phase with each other. Strictly speaking, however, TT peaks slightly before, and AA slightly after maxima in TA. This should be expected since TA bridges TT and AA steps. This leads to the question whether TA steps peak at the minor groove roll position because they just "happen" to bridge TT and AA steps or whether there is an intrinsic advantage for TA to peak at this position. As we explain further below, our model allows to give precise answers to such kind of questions.

Finally, we mention that the 10 bp periodicities of the signals displayed in Fig. 2.3 are, of course, simply a consequence of 10 bp periodicity in our model, see Eq. 2.6 and Eq. 2.7. However, very close to the termini of the nucleosomal DNA the probabilities deviate from this periodic signal. The short range of this boundary effect suggests that the probability of finding a dinucleotide is not affected much by far-away nucleotides. This can be demonstrated (and quantified) using the upper and lower bounds of the probability to which we now turn.

2.6.2 The bounds on the probability

Fig. 2.4(a)-(b) show the first- and second-order bounds on the probability to encounter AA, TA, TT or GC at dinucleotide position 58 through 88 using Eq. 2.14 and Eq. 2.15). Note that the energy as defined by Eq. 2.5 to Eq. 2.7 allows also for non-integer bp positions. Even though these non-integer positions have no physical meaning due to the discrete nature of bp sequences, we plot them here as well, as

they are a useful guide for the eye. Strictly speaking, however, only the integer positions are physically meaningful.

By using only one neighbour to the left and right (first-order bounds) the bounds indicate already the qualitative behaviour of the system for some of the dinucleotides (AA, TA and TT but not GC), see Fig. 2.4(a). Accounting for two neighbours on each side (second-order bounds) provides already an excellent estimate of the dinucleotide probabilites as the differences between the upper and lower bounds are much smaller than the observed overall variations in the probabilities at different positions, see Fig. 2.4(b).

The effect of far-away bases can be characterized by one number as follows. The difference between the upper and lower bounds decays exponentially with increasing order of the bounds (i.e., increasing the number of neighbours involved), see Fig. 2.4(c). This allows us to define an effective order κ, similar to a correlation length:

$$P^{(k)}_{\max,p}(a,b) - P^{(k)}_{\min,p}(a,b) \approx e^{-k/\kappa}. \tag{2.31}$$

The value of κ is found to be approximately equal to 1.2. This shows that increasing the order of the bounds has a huge effect around $k = 1$. It also explains why only the probabilities very close to the edges of the nucleosome are not following the 10 bp periodicity. Probabilities at positions far away from the boundaries are (exponentially) less influenced by the edge and will not 'feel' its presence.

While the results shown here are obtained at room temperature, the bounds remain an effective method at all possible temperatures, see Appendix. A.3.

2.6.3 Explaining the dinucleotide positioning rules

So far we have presented the probability distributions of a few key dinuclotides along the nucleosome model and found good agreement with the general positioning rules. We also demonstrated, by looking at upper and lower bounds of various orders, that long-range interactions are not important, but nearby neighbours matter. This is one of the reasons why the probabilities are well captured by the average neighbour approximation. Using this approximation we explain in the following how the nucleosome positioning rules in our model emerge from the elasticities and intrinsic shapes of the various dinucleotides.

Fig. 2.3 shows that the probability (calculated using Eq. 2.10) of TA dinucleotides peaks at positions of maximal negative roll (e.g., at positions 78 and 79) whereas the one of GC dinucleotides peaks at positions of maximal positive roll (e.g., at 73-74). Moreover, TT peaks at positions of maximal positive tilt (such as position 77), while AA peaks at maximal negative tilt (e.g., at 70). We first discuss the rules from a purely local perspective, i.e., just considering the elasticity and geometry of the dinucleotide under consideration. From this perspective only some of these findings make sense.

A local perspective on the dinucleotide probability fails

Table 2.1 presents all the parameters that were used in our model. Inspecting this table one finds that TT and AA have large positive and negative intrinsic tilt,

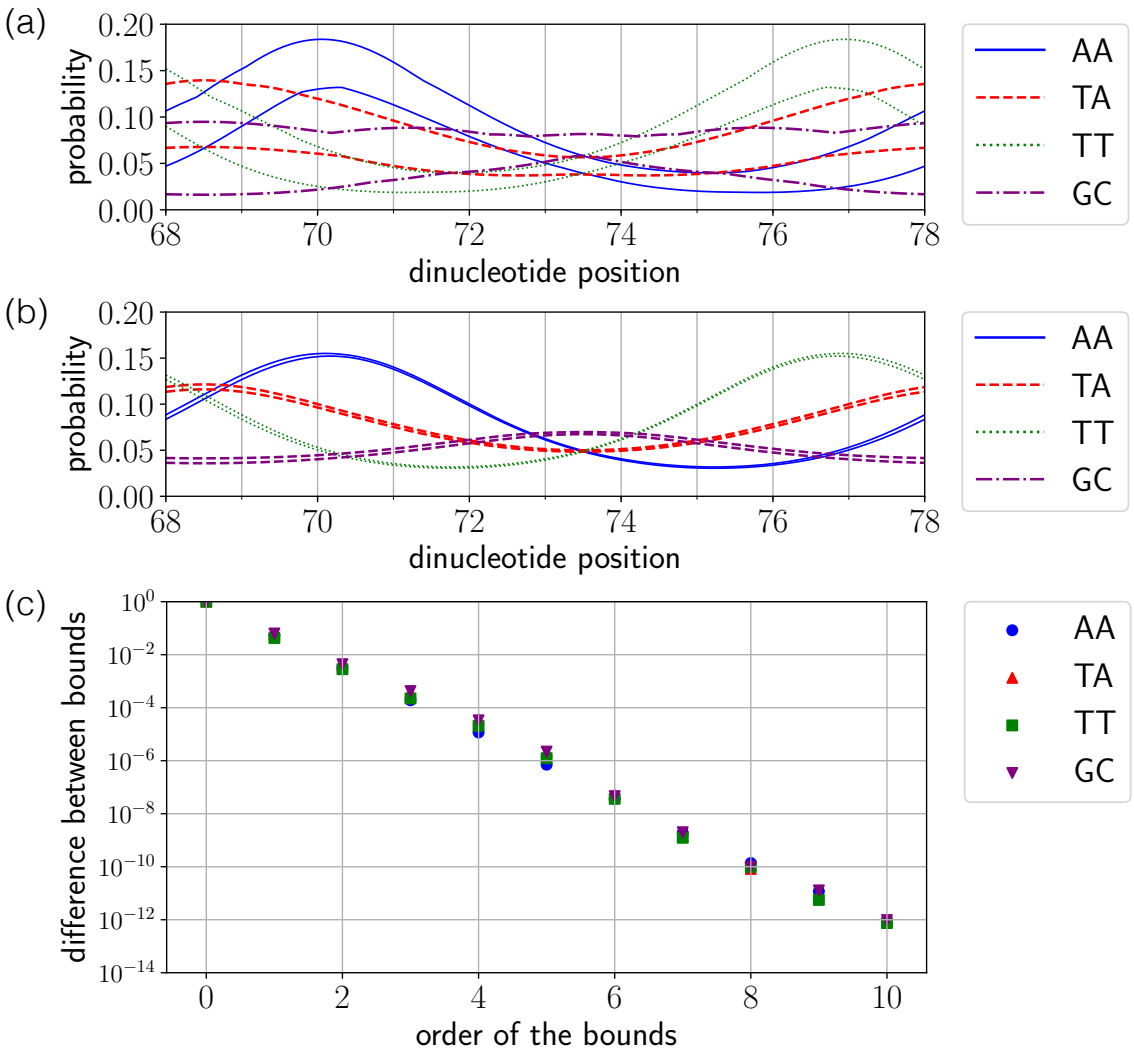


Figure 2.4: (a)-(b) Upper and lower bounds on the probabilities to have the dinucleotide AA, TA, TT or GC at several dinucleotide positions on a nucleosome. Specifically (a) depicts the first-order bounds and (b) the second-order bounds of the probability. The upper and lower bounds of the same dinucleotide have the same colour (line style). (c) Difference between the upper and lower bounds of the probabilities to encounter AA, TA, TT, and GC at position 79 at increasing order. The difference, and thereby the effect of the neighbours k steps away from the dinucleotide of interest, decreases exponentially as the order k increases.

TA										
Position	69	70	71	72	73	74	75	76	77	78
Probability	0.116	0.098	0.077	0.059	0.050	0.050	0.059	0.077	0.098	0.116
Dinucleotide Energy [k_BT_r]	0.703	0.676	0.535	0.273	0.050	0.050	0.273	0.535	0.676	0.703
-Roll	0.664	0.409	0.126	0.005	0.011	0.011	0.005	0.126	0.409	0.664
-Tilt	0.039	0.268	0.409	0.268	0.039	0.039	0.268	0.409	0.268	0.039
Average Neighbour Energy [k_BT_r]	1.573	1.566	1.472	1.275	1.102	1.102	1.275	1.472	1.566	1.573
-Roll	1.250	0.914	0.517	0.305	0.265	0.265	0.305	0.517	0.914	1.250
-Tilt	0.323	0.652	0.955	0.971	0.837	0.837	0.971	0.955	0.652	0.323

Figure 2.5: The probability of dinucleotide TA, its energy and the average of the energies of its possible neighbours are shown for 10 different positions along the nucleosome, i.e., for one full DNA helical repeat. The numbers give absolute values whereas the colours indicate how the corresponding value of the TA step compares with the values of all other possible dinucleotides at the same position. Yellow (light gray) colours represent relatively favourable values, red (dark gray) indicates unfavourable values. The probability follows mainly from a 'mixing' of the colours of the corresponding dinucleotide energies and average neighbour energies. The table also provides subdivisions of the TA energies into roll and tilt contributions.

respectively, which is consistent with their preferred positions. In contrast to that, TA has a large positive intrinsic roll, which makes positions of maximal negative roll like 78-79 highly unfavorable, even though this is where this step peaks. Even more surprising are the peaks for GC at positive roll positions as this is the dinucleotide step with the smallest intrinsic roll among all dinucleotide steps, see Table 2.1.

These findings are consistent with what we have learned from the bounds on the probabilities: zeroth-order bounds, which correspond to a purely local perspective, are not useful at all to obtain estimates of the probabilities, while first-order bounds, which include the energies of the nearest neighbours, suffice for some of the dinucleotides to have rather good estimates of the probability, see Fig. 2.4(a).

Neighbouring steps are equally important

The effect of the neighbours can be best understood using the average neighbour energy approximation, see Eq. 2.26. Since this is an excellent approximation, see Appendix A.2, the only terms important for the behaviour of the probability are the *energy of the dinucleotide itself*, and the *average of the energies of its possible neighbours*. To understand the nucleosome positioning rules we need thus to compare the energy of the dinucleotide *ab* with the energies of the 15 other dinucleotides *and* the average of the energies of all possible neighbours of *ab* with the averages of the energies of all possible neighbours of the 15 other dinucleotides.

Such information can be best presented in tabular form. Fig. 2.5 provides the relevant information for the TA dinucleotide. It presents (as numbers) the probability (obtained using Eq. 2.10) to find this dinucleotide, its energy (Eq. 2.5) and the average of the energies of its possible neighbours (see Eq. 2.26) for a 10 base pair stretch in one table (and some further information that we discuss further below). More relevant, however, are the colours assigned to each box as they indicate how these numbers compare to the values of all other possible dinucleotides. If the colour is yellow (light gray), the value is relatively favourable compared to the

GC										
Position	69	70	71	72	73	74	75	76	77	78
Probability	0.039	0.042	0.050	0.060	0.068	0.068	0.060	0.050	0.042	0.039
Dinucleotide Energy [k_BT_r]	0.546	0.644	0.681	0.571	0.428	0.428	0.571	0.681	0.644	0.546
-Roll	0.481	0.199	0.002	0.126	0.363	0.363	0.126	0.002	0.199	0.481
-Tilt	0.065	0.445	0.679	0.445	0.065	0.065	0.445	0.679	0.445	0.065
Average Neighbour Energy [k_BT_r]	2.816	2.429	1.723	0.921	0.376	0.376	0.921	1.723	2.429	2.816
-Roll	2.167	1.648	0.926	0.352	0.070	0.070	0.352	0.926	1.648	2.167
-Tilt	0.649	0.781	0.797	0.569	0.306	0.306	0.569	0.797	0.781	0.649

Figure 2.6: Same as Fig. 2.5 but for GC.

ones of other dinucleotides *at the same dinucleotide position* (i.e., the probability is relatively high, while the energy cost is relatively low). Red (dark gray) denotes unfavourable values, while orange (gray) indicates that this value is average.

First consider in Fig. 2.5 row "Probability": At positions 69 and 78, both associated with *negative* roll and zero tilt, TA is favourable, as we have seen in Fig 2.3. Next consider row "Dinucleotide energy": The dinucleotide energy of TA goes against this preference having the lowest values at positions 73 and 74, and its highest at positions 69 and 78, both in absolute values (numbers) and relative values (colours). Next turn to row "Average neighbour energy": the absolute values (numbers) have their lowest values at 73 and 74 but the relative values (colours) strongly prefer the opposite. Therefore, what causes the TA preference for negative roll positions is the average energy of the possible neighbours relative to the average energy of the forbidden neighbours.

Now we turn to the other rows in Fig. 2.5. These extra rows provide a subdivision of the dinucleotide energies and the average neighbour energies into roll and tilt components. Inspecting these four extra rows reveals that the main cause for the TA preference for positions 69 and 78 lies in the average tilt contribution of the possible neighbour steps. This overrides TA's own preference (relative and absolute) for positive roll.

The same analysis as the one on TA can be performed on GC by inspecting Fig. 2.6. This is another non-trivial dinucleotide in the sense that its behaviour is heavily affected by its possible neighbours. Positions 73 and 74, associated with large positive roll and bending towards the major groove, lead to a high dinucleotide energy of GC (compared to other steps), which has a low (positive) intrinsic roll. However, the possible neighbours cause the probability of encountering GC to be highest at these positions and lowest at positions 69 and 78.

The complete picture

In Fig. 2.7 tables are shown that present the probabilities and relative energies for all 16 dinucleotides (again in colour code). There are seven tables corresponding to the seven rows in Fig. 2.5 and Fig. 2.6. Using these tables one can analyse preferences for each dinucleotide step individually, just as explained for TA and GC above. Moreover, for cases where the average neighbour energies dominate the positional preferences of dinucleotides (like for TA and GC), these tables allow to look up which of the possible neighbours of a given dinucleotide are favourable.

As an example, we consider again the dinucleotide TA. In Fig. 2.7(a) we see

that the probability of TA peaks at positions 69 and 78, which is not TA's intrinsic preference, Fig. 2.7(b), but that of its neighbours on average, Fig. 2.7(c). We need now to inspect the intrinsic preferences of all the possible neighbours. At position 70, three of the four possible neighbours (dinucleotides starting with an A) are favourable, namely AA, AT, and AC, see Fig. 2.7(b). Due to symmetry TT, AT, and GT are favourable at position 77, see also Fig. 2.7(b). Further details are revealed by Fig. 2.7(d), and (e) that present the roll and tilt contributions to the dinucleotide energies. It shows that AA at 70, and TT at 77 are favourable due to both their roll and tilt preferences whereas the other favourable steps, AT and AC at 70, and AT and GT at 77, prefer those positions due to roll alone. Inspecting the contributions of roll and tilt to the average neighbour energies for TA at positions 69 and 78, Fig. 2.7(f) and (g), one learns that both degrees of freedom matter but tilt is the dominant factor. This reflects the very strong tilt preference for AA and TT but also the fact that the only unfavourable neighbours (AG at 70, and CT at 77) are unfavourable due to roll whereas the tilt contributions are favourable.

Note that these considerations also explain preferred occurrences of larger motives, like e.g., TTAA centered around negative roll positions. In addition, similar lines of arguments can be used to understand why TA is unfavored at high roll positions like 73 and 74, or the preferences of any other dinucleotide for that matter.

Shape is more important than stiffness

The roll and tilt terms of the energy can be subdivided even further. As can been seen from equations Eq. 2.5 to Eq. 2.7 the sequence dependences enter the roll and tilt energies both through the intrinsic geometries and through the stiffnesses related to these two degrees of freedom. We show now that the stiffnesses are not very important to the behaviour of our system. In Fig. 2.8 we compare two tables for dinucleotide energies: the original table on the left (identical to Fig. 2.7(b)) and on the right a table that is produced when we set all stiffnesses of roll and tilt to the same small value, namely to 1. Even though the specific value of the stiffness affects strongly the absolute values of the dinucleotide energies (not shown), it does hardly affect the relative values of the energies (colour code). This reveals that, at least in our simplified model, the sequence preferences are largely governed by the intrinsic roll and tilt (and not the stiffnesses) of the dinucleotides. Note that this observation is consistent with the findings reported in [42] where molecular dynamics simulations performed on rather detailed nucleosome models revealed that nucleosome affinity is dominated by the shape of the wrapped DNA.

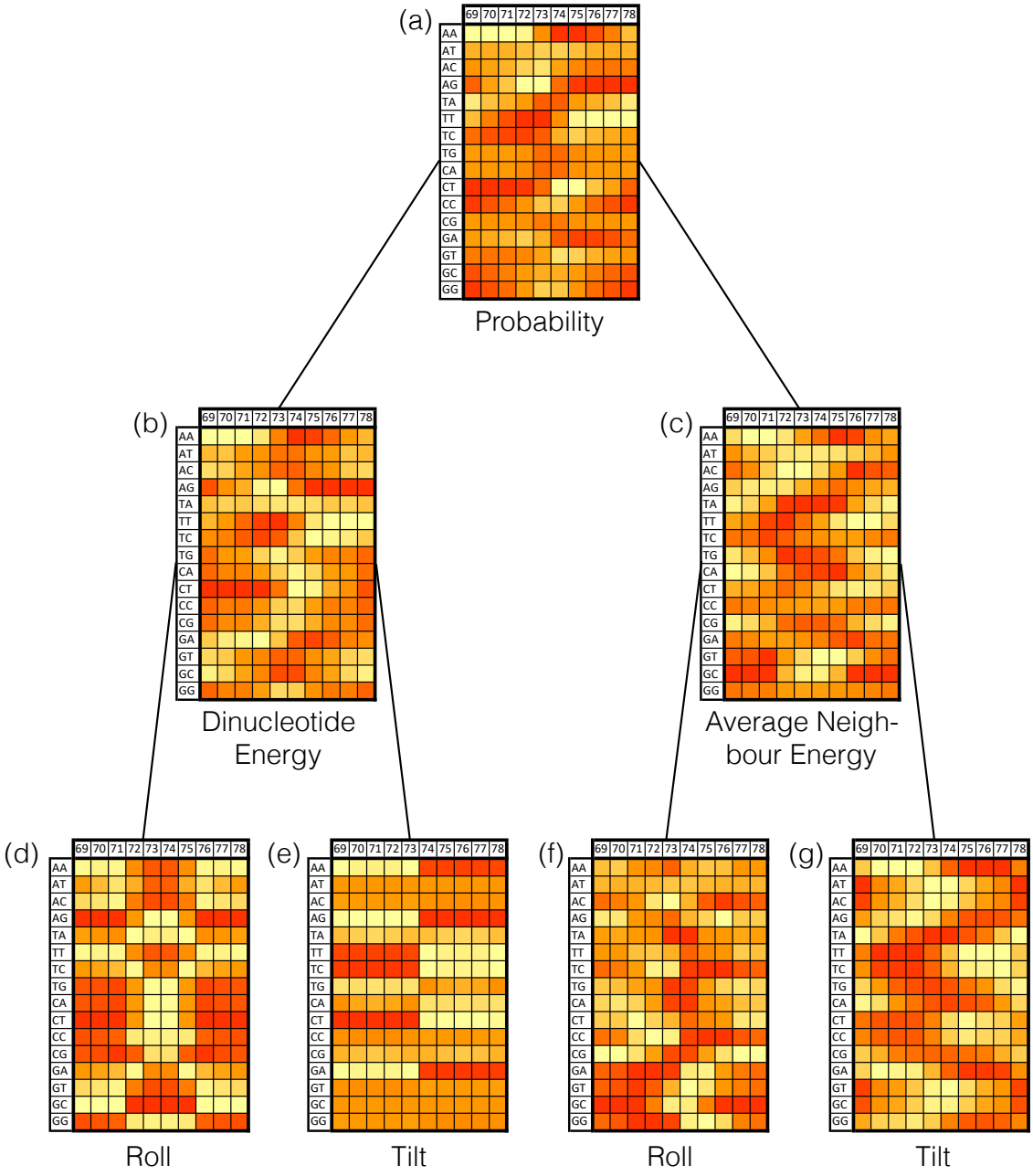


Figure 2.7: (a) Probability, (b) dinucleotide energy, and (c) average neighbour energy of all 16 dinucleotides for one DNA helical repeat. Yellow (light gray) denotes high probability/low energy, red (dark gray) low probability/high energy (relative to all other dinucleotides at the corresponding location). In addition provided are subdivisions of dinucleotide energies into (d) roll and (e) tilt, and of neighbour energies into (f) roll and (g) tilt. The colours representing the probabilities can be seen as a 'sum' of the colours of the dinucleotide energies and the average neighbour energies. The colours corresponding to the dinucleotide energies are the 'sum' of the colours for roll and tilt energies. The same holds for the neighbour energies.

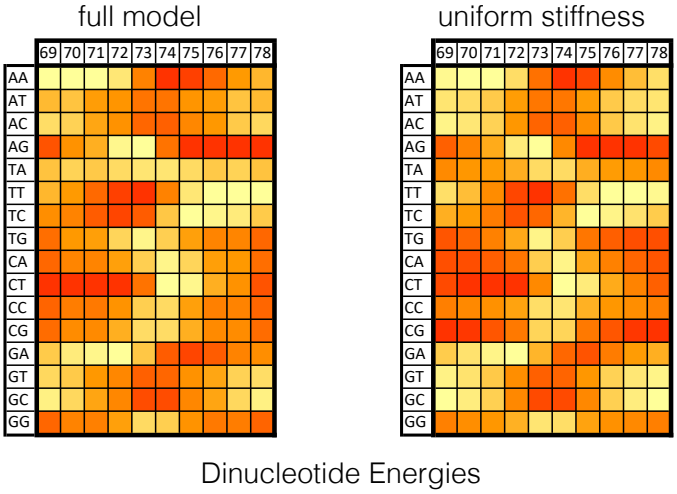


Figure 2.8: Original dinucleotide energy costs (left; same as Fig. 2.7(b)), and energy cost with all stiffnesses set to 1 (right) are shown side by side. The strong similarity between the two tables reveals that stiffnesses play only a minor role for the dinucleotide positional preferences.

Conclusion

In this chapter we obtained a detailed understanding of the physics behind nucleosome sequence preferences as they arise from the sequence dependent geometry and elasticity of the DNA double helix. Our strategy was to build a model that is simple enough so that it can be solved analytically and complex enough to reproduce the experimentally known nucleosome positioning rules. This was achieved by forcing a coarse grained DNA model (the rigid base pair model) along a circular path and accounting for the sequence dependent mechanics of only the most important degrees of freedom (roll and tilt). With the help of transfer matrices we were able to calculate the dinucleotide probabilities along our nucleosome model. These reproduce, at least qualitatively, the rules found when nucleosome position themselves freely along a long stretch of DNA (e.g., the yeast genome [6]).

However, to really understand the dinucleotide rules in detail, exactly solving the model (or simulating a more detailed version of it [14]) is not sufficient, as this system behaves rather complex. For instance, of the four "important" dinucleotides only two (AA and TT) prefer locations that correspond to their own intrinsic preferences whereas the other two (TA and GC) peak at their most unfavourable locations. To solve this puzzle, we first introduced an approximation that, by taking a limited number of neighbours around a given dinucleotide into account, provides upper and lower bounds to its probability distribution. From this we learned that the nearest neighbours influence strongly the preferences of a given dinucleotide whereas the influence of nucleotides further away is small, decreasing exponentially.

With this information at hand, we finally introduced an approximation tailored for interpreting the dinucleotide preferences. According to this average neighbour energy approximation dinucleotide preferences are dominated by two contributions: the intrinsic energy cost to place a given dinucleotide at a given position and the average energy of the possible neighbours before and after that given dinucleotide. This is an excellent approximation and allows to explain all the dinucleotide preferences found in our model. Depending on the dinucleotide at hand, a given dinucleotide is found preferentially at certain positions mainly due to its own preferences (e.g., AA and TT) or due to bringing in "good" neighbours (e.g., TA and GC).

Knowing the dinucleotide preferences of nucleosomes allows genome wide calculations of nucleosome positioning [10]. Therefore understanding how dinucleotide preferences along nucleosomes emerge from the sequence dependent DNA mechanics, means ultimately to understand the physical underpinnings of biological processes at much larger scales as the depletion of nucleosomes at gene start sites in yeast [6, 9] or the retention of nucleosomes in human sperm cells [9, 35].

Chapter 3

Shortest paths through synonymous genomes

This chapter is based on: Zuiddam and Schiessel, 2019, Phys. Rev. E. [43]

The elasticity of the DNA double helix varies with the underlying base pair sequence. This allows to put mechanical cues into sequences that in turn influence the packaging of DNA into nucleosomes, DNA-wrapped protein cylinders. Nucleosomes dictate a broad range of biological processes, ranging from gene regulation, recombination, and replication, to chromosome condensation. In this chapter, we demonstrate how genetic and mechanical information can be multiplexed by introducing a novel method. This method maps DNA sequences onto graphs and use shortest paths algorithms to determine which DNA stretches are easiest or hardest to bend inside a nucleosome. We further demonstrate how genetic and mechanical information can be multiplexed by studying paths through graphs of synonymous codons. Using this method we find that nucleosomes can be placed by mechanical cues nearly everywhere on the genome of baker's yeast. We abandon the simple physical model for the nucleosome used in the previous chapter and use a more sophisticated, trinucleotide energy model resulting from Mutation Monte Carlo simulations [10].

3.1 Introduction

The geometrical and mechanical properties of DNA double helices depend on their underlying base pair (bp) sequences. Certain bp combinations lead to intrinsically curved DNA and other combinations to DNA that is stiffer or softer than average. This allows for a second layer of information to be written along DNA molecules in addition to the classical layer, the genes that encode for the proteins.

An important biological consequence of sequence dependent DNA mechanics is its impact on the positioning of nucleosomes that sequester a large fraction of eukaryotic DNA (e.g. 3/4 for humans). Each nucleosome consists of 147 bp of DNA wrapped almost two times around a globular octamer of histone proteins leading

to a DNA spool of about 10 nm in diameter [33]. The wrapped piece of DNA is about one persistence length long; thus bending energies are substantial [34]. As a result, nucleosome stability greatly depends on sequence-dependent differences in the elasticity and shape of the wrapped DNA double helix. In addition, the DNA molecule makes mainly contact with the histone octamer via its backbones [33], which are chemically independent of its bp sequence. All this suggests that the affinity of a sequence to be part of a nucleosome is mainly reflected by the ease with which the DNA can be wrapped into a nucleosome. The total number of possible affinities is huge: there are $4^{147} \sim 10^{88}$ distinct DNA sequences that could be part of a nucleosome.

The sequence-dependent affinity leads to a non-random positioning of nucleosomes along genomic DNA. This can be clearly seen by reconstituting nucleosomes on long DNA from their pure components via salt dialysis and then producing nucleosome maps using genome wide assays that extract DNA stretches which were stably wrapped in nucleosomes (see e.g. [6]). One determines the nucleosome occupancy at each bp position which is the probability that the corresponding bp is covered by a nucleosome. There are two types of nucleosome positioning along DNA: rotational and translational positioning [7]. Rotational positioning is caused by the fact that a given DNA stretch is typically not intrinsically straight due to the intrinsic geometries of the involved bp steps. This causes a preference for the nucleosome to sit in a certain orientation on the DNA, i.e. it prefers a set of positions 10 bp apart (as the histone binding occurs via the DNA backbones and DNA is a helix with an about 10 bp periodicity). Translational positioning is caused by DNA stretches that have overall a higher affinity for nucleosomes. This correlates well with their GC content [8, 9].

Histone octamers are known to spontaneously "slide" along DNA [44] and therefore to sample different positions, allowing for the equilibration of nucleosomes, at least locally. Two mechanisms have been suggested and both are based on thermally induced defects in the nucleosome: single bp twist defects (an extra or a missing bp) [45, 46] and 10 bp bulges [47, 48]. New simulation studies [49, 50] strongly suggest that both mechanisms occur and that it depends on the underlying DNA sequence which one is preferred for a given DNA stretch. *In vivo* there are, in addition, chromatin remodellers that use ATP to move nucleosomes along DNA. New experiments [51] and simulations [52] suggest that at least some of them actively induce twist defects in the nucleosome. Chromatin remodellers might help nucleosomes to equilibrate their location along DNA [53] but might also, together with other proteins that compete for the DNA, perturb the intrinsically preferred positioning of nucleosomes [54].

In vitro nucleosome maps show clearly that bp sequences influence the positions of nucleosomes, see e.g. [6] for yeast. It has been claimed that even *in vivo* about 50% of the nucleosome positions on the yeast genome can be predicted based on the bp sequence alone [5]. However, it should be stressed that many nucleosomes are not really positioned individually by dedicated mechanical cues but rather indirectly by GC-poor regions with low nucleosome affinity, especially around transcription start and termination sites. These regions effectively act like barriers for nucleosomes. Close to such a barrier, at sufficiently high nucleosome densities, a statistically ordered pattern is formed by the nucleosomes, a scenario already suggested by Ko-

rnberg and Stryer [55]. In fact, short enough genes form crystal-like configurations between the barriers [56]. The situation is dramatically different for humans [8] and other higher vertebrates [57]. Genomes of these organisms contain well-positioned nucleosomes around nucleosome-inhibiting barriers. These barriers are spread all over the genome of those organisms and nearby nucleosomes are not just statistically ordered as in yeast, but instead they are positioned by characteristic patterns of GC- and TA-rich regions. In humans these positioned nucleosomes alone account for about 30% of all the nucleosomes mapped *in vivo*.

The purpose of the current study is to demonstrate the extreme malleability of DNA mechanics and geometry allowing for mechanical cues for nucleosomes along the bp sequence. For instance, we demonstrate that such cues can even be multiplexed with classical genetic information. In Ref. [14] and [58] we had already presented some first results for putting mechanical cues on top of genes and for creating special nucleosomes. However, we still missed a fast method to do this systematically. Nevertheless we were able to demonstrate that multiplexing was possible due to the simultaneous occurrence of three effects: the sequence properties of genomes, the degeneracy of the genetic code and the plasticity of the mechanical code (see Ref. [14] for details).

3.1.1 Overview

In this chapter we present a novel set of methods that allows to find special nucleosomes for any short-range 1D energy/probabilistic nucleosome model. In Section 3.2 we present our specific model of choice. Then in Section 3.3 we demonstrate how for given integer k one obtains the k lowest and k highest energy sequences. For DNA molecules longer than 147 bp we construct the deepest possible energy well leading to the best-positioned nucleosome (Section 3.4). Next we modify bp sequences on genes to position nucleosomes almost everywhere on the yeast genome without modifying the encoded proteins (Sections 3.5 and 3.6). All this is achieved by mapping the corresponding bp sequences on appropriately weighted graphs and using a (k-)shortest path algorithm. Earlier attempts to obtain lowest energy nucleosome positioning sequences [36], or to reposition nucleosomes on a DNA molecule [14] rely on Monte Carlo simulations, which carry serious disadvantages compared to our new methods. Such simulations do not allow to prove which sequences have the lowest or highest energy without evaluating the huge set of all possible sequences. A shortest path algorithm however, is not only deterministic and exact, but also extremely efficient (for example, Dijkstra's algorithm with Fibonacci heap implementation has a complexity of $O(|M| + |N| \log |N|)$, where M and N denote the number of edges and vertices, respectively [23]). The final section provides a conclusion.

3.2 Model

We showcase our methods by using the recent probabilistic trinucleotide model [10] that was obtained through Monte Carlo simulations of a coarse-grained nucleosome model with sequence dependent DNA elasticity [14]. In this nucleosome model the DNA is represented by the rigid bp model [32] which treats each bp as a rigid plate, the spatial position and orientation of which are described by six (three translational

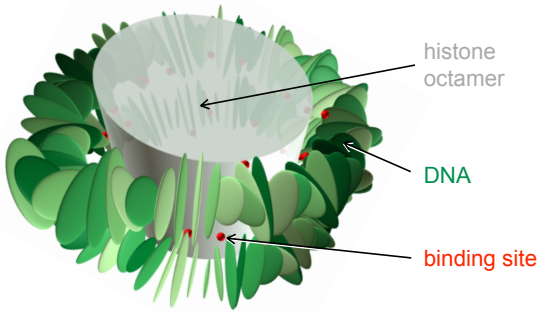


Figure 3.1: Nucleosome model [14] used to construct the probabilistic trinucleotide model [10]. Each rigid plate represents a bp, the locations of the constraints (corresponding to bound phosphates) are shown by beads, two per binding site. The cylinder is a rough representation of the protein core but was not simulated explicitly (except through the binding sites).

and three rotational) degrees of freedom. It assumes only nearest-neighbor interactions with a quadratic deformation energy between successive bps. The sequence-dependence of the model comes into play because the stiffness and intrinsic shape of a given bp step depend on its chemical identity. The DNA is forced into a superhelix through a set of 28 constraints that represent the 14 binding sites to the histone octamer (see Fig. 3.1) which were extracted from the nucleosome crystal structure without introducing free parameters [14]. These constraints correspond to bound phosphates in the DNA backbone (see Ref. [14] for details). This model has been widely tested against experiments, e.g. it successfully predicts relative nucleosome affinities of various sequences [14] (as measured in [5, 59, 60]), the rotational positioning rules of nucleosomes [14, 61] (see [4, 5]), translational positioning [9] (see [6, 62, 63]), sequence dependent nucleosome breathing [64] (see [65, 66]) and force induced unwrapping [19, 58] (see [67]).

To construct the probabilistic trinucleotide model [10] from the coarse-grained nucleosome model [14] we performed a Monte Carlo simulation that randomly mixes conformational and sequence moves (mutation Monte Carlo method [14]). With this method we created a large number of high affinity sequences allowing to accurately determine the occurrence probabilities of mono-, di- and trinucleotides along the nucleosomal DNA. The overall probability of a sequence to be part of a nucleosome can then be estimated by a two-step Markov process [10] (see **Appendix B.1**). Moreover, the energy cost of wrapping a sequence S of nucleotides $S_i \in \mathcal{B} = \{A, T, C, G\}$, $i = 1, ..., L$ with $L = 147$, into a nucleosome is given by

$$E(S) = \sum_{n=1}^{L-2} E_n(S_{n+2}, S_{n+1}, S_n). \tag{3.1}$$

The E_n's are 'conditional' trinucleotide energies (see **Appendix B.1**) which serve as weights of our graphs below. We set the energy of the ground state sequence to zero. For convenience we define $E_n = 0$ for $n < 1$ and $n > L - 2$.

3.3 Lowest and highest energy sequences

We aim to find the ground state sequence in the set of all possible sequences. These can be described as paths through graph $\mathcal{G}$ in Fig. 3.2(a). $\mathcal{G}$ consists of the nodes *source*, *sink*, and $(XY)_i$ for all $X, Y \in \mathcal{B}$, and $i \in \{1, 2, ..., L-1\}$. We draw the following directed edges (with $X, Y, Z \in \mathcal{B}$): from *source* to $(XY)_1$ with weight zero, from $(XY)_{L-1}$ to *sink* with weight zero, and for all $i \in \{1, 2, ..., L-2\}$ from $(XY)_i$ to $(YZ)_{i+1}$ with weight $E_i(X, Y, Z)$.

A path from *source* to *sink* corresponds to a sequence, and its length equals the energy cost of that sequence. Therefore, the lowest energy sequence corresponds to the shortest path from *source* to *sink*, which can be found using a shortest path algorithm. Because the graph contains no cycles, the shortest path algorithm can also be used to find the longest path, i.e., the highest energy sequence. Using a k-shortest path algorithm, we can even find $k \in \mathbb{N}^+$ of the lowest and highest energy sequences. We use Yen's algorithm with Dijkstra's as the underlying shortest path algorithm leading to a time complexity of $O(kN(M+N) + N \log N)$ [24]. The energies corresponding to the 5000 best and worst sequences are shown in Fig. 3.2(b). They resemble the tails of a Gaussian error function, suggesting that the probability density function of the energies resembles a (somewhat skewed) Gaussian.

Ten of both the lowest and highest energy sequences are depicted in Fig. 3.3, see also Fig. 3.2(b). Because L is odd, there is a bp in the center of the nucleosome leading to two ground state sequences and two highest energy state sequences. The lowest energy sequences a1 to a10 have a very high C/G content (about 80%) which is favoured by nucleosomes [8, 9]. The most common dinucleotides are CC/GG, GC and CG. We find GC steps mainly where the major groove bends towards the histone octamer which agrees with the nucleosome positioning rules [4, 5] but they appear also at many other positions. On the other hand, the highest energy sequences b1 to b10 feature a high A/T content. The most common dinucleotides are AA and TT. We find A tracts with a length of 5 to 6 bps, which are known to repel nucleosomes [68]. Moreover A/T "disobeys" the position rules [5] by avoiding locations where the minor groove faces inward.

As a cautionary remark we stress here that these extreme sequences might not outperform high affinity sequences found experimentally (like e.g. the Widom 601 sequence [69]), since errors in the underlying parametrization may be amplified when studying extreme cases, see also Ref. [36].

3.4 The best positioned nucleosomes

After finding the ground state sequences we determine next the most strongly positioned nucleosome. We consider a sequence longer than $L = 147$ and call a nucleosome positioned at a particular location when all the energies of a set of neighbouring positions are higher. Specifically, as energy landscapes show typically undulations with a 10 bp period [14, 36] we introduce sequences $\mathcal{S}$ of length $L+10$ or longer and aim to find a position that has a much lower energy than its ten closest positions.

Let S^p be a subsequence of $\mathcal{S}$ of length L, from position p to $p+L$. We call a nucleosome positioned at p if the energy $E(S^p)$ is lower than the energies at positions $p-5, p-4, ..., p+5$ (excluding p). Its energy difference to the smaller value of the two

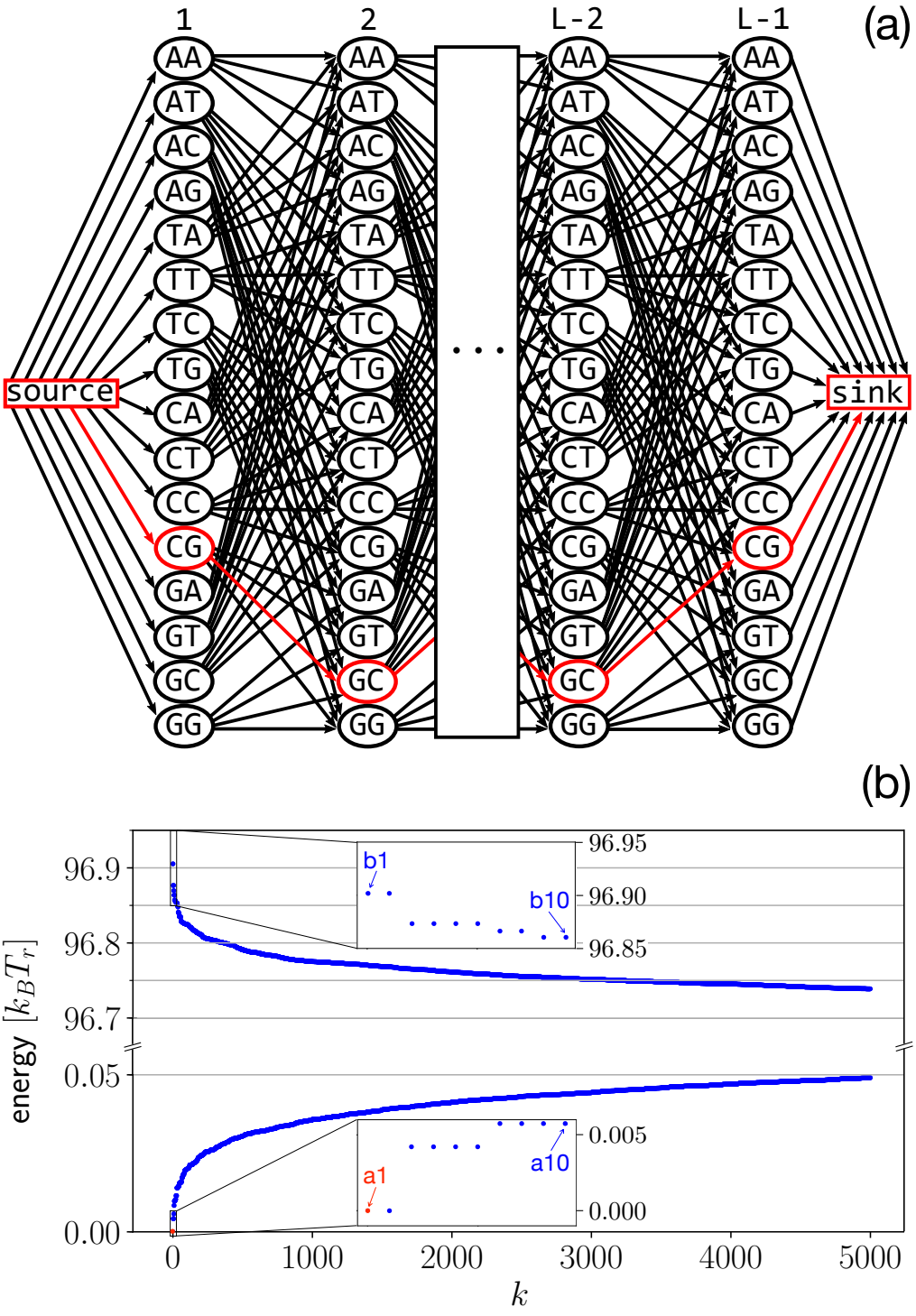


Figure 3.2: (a) Graph representation of the 4^{147} DNA sequences that can be wrapped into a nucleosome. Weights are assigned such that each path from *source* to *sink* has a length equal to the total energy of the corresponding sequence. The path in red (grey in greyscale) corresponds to the ground state sequence a1 from Fig. 3.3. (b) Energies of the 5000 cheapest (bottom) and the 5000 most expensive sequences (top). The insets show the 10 best and worst sequences, see Fig. 3.3.

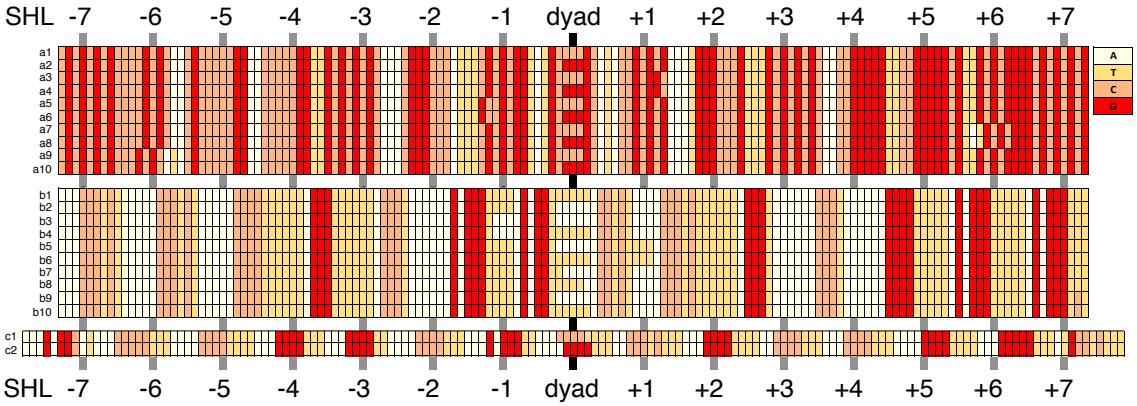


Figure 3.3: a1-a10: ten lowest energy sequences, b1-b10: ten highest energy sequences, c1/c2: best positioned nucleosomes. Because of symmetry, sequences a1 and a2 have the same energy and so on. At integer superhelical locations (SHL) the major groove of the DNA bends towards the histone octamer. SHL 0 is the nucleosome dyad.

energy maxima (to the left and to the right) we call its depth $\mathcal{D}$ (formally defined in **Appendix B.2**). As an example, consider a nucleosome on ground state sequence a1, extended by placing it in a tandem repeat. As the black curve in Fig. 3.4 shows, this leads to deep minimum with depth $\mathcal{D}$ close to $30\,k_BT$.

It turns out that one can find even deeper minima. To obtain narrow bounds on the deepest possible minimum, we introduce new graphs, with different weights, such that we minimize the quantity:

$$\min_S \left[2E(S^p) - E(S^{p+h}) - E(S^{p+j})\right] \tag{3.2}$$

with $h \in \{-5,-4,..,-1\}$, $j \in \{1,2...,5\}$. What allows us to find the deepest possible minimum is the symmetry of our system, caused by the DNA helical shape: placing a nucleosome i positions to the left or right from a local minimum will have comparable energy costs. Because of this, when we perform Eq. 3.2 for $h = -j$ we obtain $E(S^{p+h}) \approx E(S^{p+j})$, which, combined with Eq. 3.2, allows us to find a great estimate for the deepest minimum.

We now define the graphs $\mathcal{G}^+_{h,j}$ (depicted **in Fig. B.1 of Appendix B.4**), extensions and modulations of $\mathcal{G}$, for $h,j \in \{-5,-4,..,-1,1,2...,5\}$ as follows: The graph $\mathcal{G}^+$ consists of the nodes *source*, *sink*, and $(XY)_i$ for all $X,Y \in \mathcal{B}$, and for all $i \in \{-4,-3,...,L+3,L+4\}$. For all $X,Y,Z \in \mathcal{B}$ we draw the following directed edges: from *source* to $(XY)_{-4}$ with weight zero, from $(XY)_{L+4}$ to *sink* with weight zero, and for all $i \in \{-4,-3,...,L+3\}$ from $(XY)_i$ to $(YZ)_{i+1}$ with weight $E'_{i,h,j}(X,Y,Z)$, where we define the function

$$E'_{i,h,j} \equiv 2E_i - E_{i-h} - E_{i+j}.$$

A shortest path through $\mathcal{G}^+_{h,j}$ minimizes the quantity $2E(S^p) - E(S^{p+h}) - E(S^{p+j})$. By looking at all possible graphs $\mathcal{G}^+_{h,j}$, for $h,j \in \{-5,-4,..,-1,1,2...,5\}$ we can show that the maximum depth is achieved by taking the shortest path through $\mathcal{G}^+_{-5,5}$, which is given by $83.47 \pm 0.03\ k_BT_r$ (the tiny possible error is due to $E(S^{p+h})$ only being approximately equal to $E(S^{p+j})$, see **Appendix B.3**). The resulting

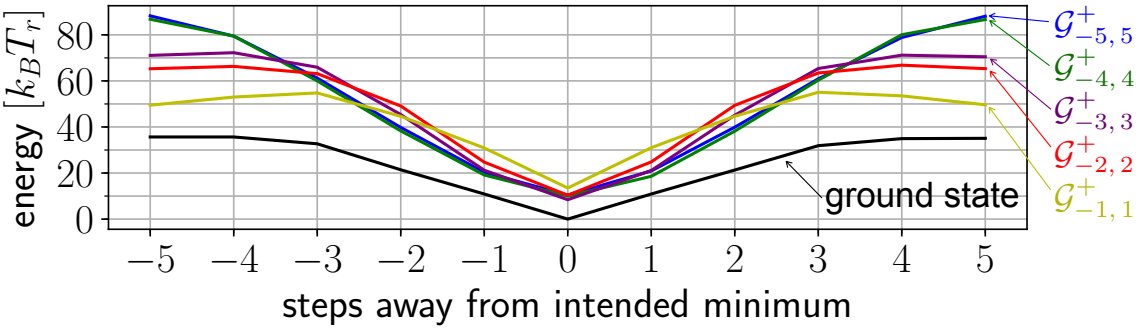


Figure 3.4: The energy landscapes as a result of a shortest path through graphs $\mathcal{G}^+_{-i,i}$, $i = 1, 2, ..., 5$, are shown, as well as the landscape of ground state sequence a1.

shape, as well as those for other graphs, is shown by Fig. 3.4. Remarkably a mere 5 bp shift leads to an energy change of 86% of the total energy range of 96.9 k_BT_r, Fig. 3.2(b).

This path (c1/c2 in Fig. 3.3) is very different from the ground state sequence a1/a2. It contains few GC or CG steps and has a much higher A/T content which is concentrated around half-integer superhelical locations (SHLs). At most such locations one finds the motive TTAA which is known to strongly position nucleosomes in a certain rotational setting by intrinsically bending the DNA double helix [31]. The dinucleotides along c1/c2 share closely the nucleosome positioning rules [5] which are in fact rotational positioning rules caused by an intrinsic DNA shape [31, 42].

3.5 Lowest and highest energy on genes

We found that the difference between the lowest and highest possible energy is very high, suggesting that DNA mechanics allows for substantial mechanical cues to position nucleosomes. Now we ask to which extent such mechanical information can still be present under an important biological constraint: conservation of genetic information. Protein coding sequences are highly degenerate with 18 of the 20 amino acids being represented by not just one codon but by a set of synonymous codons. A 147 bp-stretch consist of ℓ codons, where ℓ can be either 49 or 50 (in the latter case two codons are only partially inside that stretch). To find the lowest and highest energy sequences that code for the same protein, we use a graph $\mathcal{G}_{\text{gene}}$ which contains all synonymous codons of the given gene section. An example is depicted in Fig. 3.5(a).

$\mathcal{G}_{\text{gene}}$ is defined as follows. Let R^i denote the set of all synonymous codons at the i^{th} codon position. R^i contains at least one and at most six elements. The node-set of $\mathcal{G}_{\text{gene}}$ consists of the elements *source*, *sink* and C^i for all $C^i \in R^i$, $i \in \{1, 2, ..., \ell\}$. For all these nodes we draw the following directed edges: from *source* to C^1 with weight zero, from C^ℓ to *sink* with weight $w_\ell(C^\ell, C^x)$ (C^x can be any codon: by definition, its energy will always be zero), and for all $i \in 1, 2, ..., \ell - 1$ from C^i to C^{i+1} with weight $w_i(C^i, C^{i+1})$. The weight w_i is given by

$$\begin{aligned} w_i(C, D) = E_{3i-2}(C_1, C_2, C_3) + E_{3i-1}(C_2, C_3, D_1) \\ + E_{3i}(C_3, D_1, D_2). \end{aligned} \tag{3.3}$$

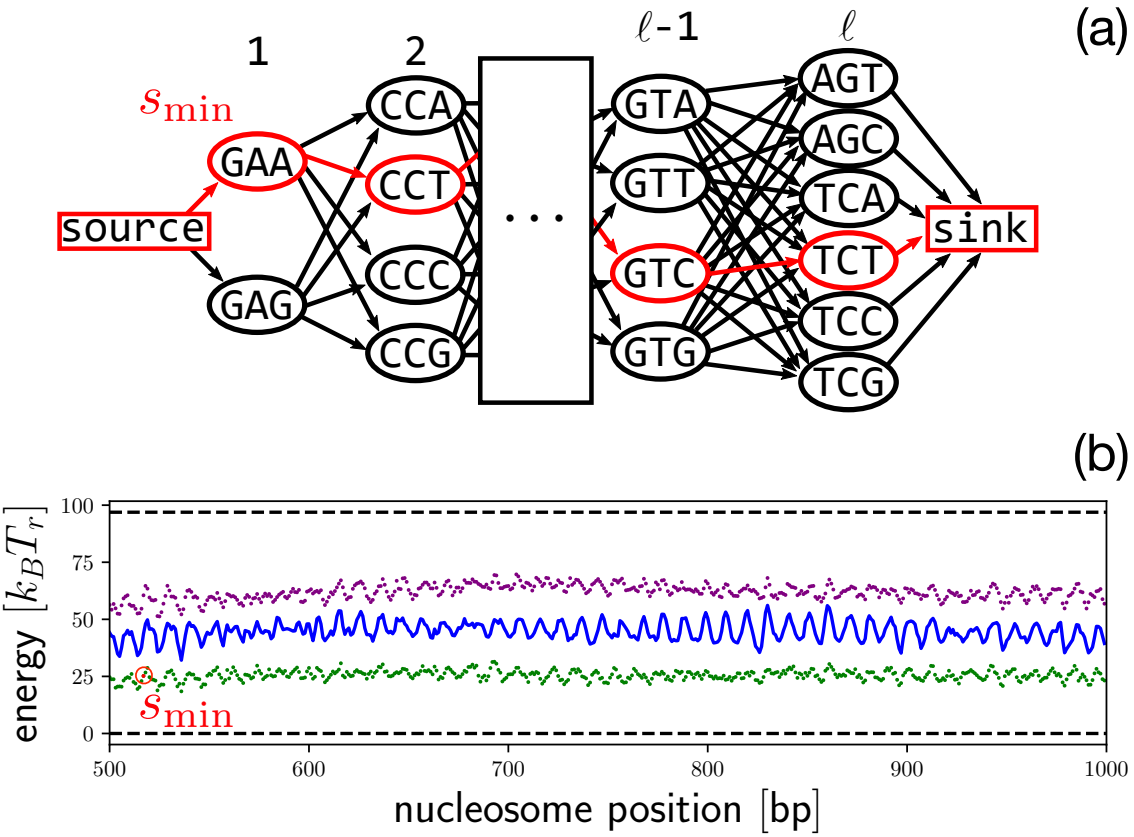


Figure 3.5: (a) Graph $\mathcal{G}_{\text{gene}}$ shows all synonymous ways to encode for a given amino acid sequence (ℓ is either 49 or 50). The shortest energy path $s_{\min}$ is highlighted. (b) Energy landscape of a 500 bp stretch of gene YAL002W from yeast (solid curve), pointwise minimal and maximal energies through synonymous mutations (dotted curves) and total minimum and maximum (dashed lines). $s_{\min}$ is the same sequence as in (a).

where C_j and D_j denote the j^{th} base of the codons C and D. The length of a path from *source* to *sink* in $\mathcal{G}_{\text{gene}}$ equals the energy of the corresponding sequence.

We now apply the shortest path algorithm to find the lowest and highest energy at each position on a 500 bp long stretch of gene YAL002W in baker's yeast, see Fig. 3.5(b). We find at each position synonymous paths that substantially lower or increase the original energy such that the available energy range is about one half of the total energy range in nucleosome affinities. Note that the 10 bp undulations of the original landscape are still visible in undulations of the lowest and highest energies.

3.6 Nucleosome positioning on genes

We have presented a method to obtain the lowest and highest energy sequences while conserving genetic information. Now we ask whether it is possible to create a minimum (of given depth $\mathcal{D}$) at any bp position on the yeast genome. To answer this question we introduce graph $\mathcal{G}^{+}_{\text{gene}}$, a modification of $\mathcal{G}_{\text{gene}}$, which includes some neighbouring positions and keeps the gene intact (see **Appendix B.4, Fig. B.2** for

an example). The node-set of $\mathcal{G}^{+}_{\text{gene}}$ consists of the elements *source*, *sink* and C^i for all $C^i \in R^i$ and for all $i \in \{-1, 0, ..., \ell + 2\}$.

For all $C^j \in R^j$ with $j \in \{-1, 0, ..., \ell + 3\}$ we draw the following directed edges: from *source* to C^{-1} with weight zero, from $C^{\ell+2}$ to *sink* with weight $w'_{\ell+2}(C^{\ell+2}, C^{\ell+3})$, and for all $i \in 1, 2, ..., \ell + 1$ from C^i to C^{i+1} with weight $w'_i(C^i, C^{i+1})$. The weight w'_i is given by

$$w'_i(C, D) = \sum_{i=-5}^{5} c_i w_i(C, D). \tag{3.4}$$

where we set $c_0 = 1$ and $c_i \leq 0$ for $i \neq 0$.

Our previous methods to create minima fail at many bp locations where minima appear at wrong positions, because genetic sequences are asymmetric (red dashed curve from Fig. 3.6(a) depicts results for ground state sequences). We resolve this by systematically changing the c_i-values at each iteration step. If e.g. a minimum appears i bp to the right of the desired position, we decrease the constant c_i by 0.1 and run the shortest path algorithm again with the modified weights. This gives the algorithm an 'incentive' to increase the energy at that position. For details see **Appendix B.5**.

The resulting depths $\mathcal{D}$ (in units of $k_B T_r$) after performing this analysis on all genes of yeast that contain no introns (7640994 nucleosome positions in total) are shown in figure 3.6(a). In 99.9943% of the cases we find a minimum, $\mathcal{D} \geq 0$; for only 438 positions we do not succeed, i.e. $\mathcal{D} < 0$. Minima are deeper than $\mathcal{D} \geq 10$ for 99.897%, $\mathcal{D} \geq 20$ for 85.67% and $\mathcal{D} \geq 30$ for 14.71%.

What about the small fraction where we fail to produce a minimum? In fact, theoretically it is possible to construct sequences with an unchangeable energy landscape, e.g. a chain made up entirely from methionine units, that can only by encoded by ATG's. As can be seen Fig. 3.6(b), gene sections where we fail to create a minimum, $\mathcal{D} < 0$, reflect the presence of amino acids with low degeneracy N_{syn}, $N_{\text{syn}} \leq 2$, whereas the presence of amino acids with six synonymous codons, $N_{\text{syn}} = 6$, allows for deep minima with $\mathcal{D} > 30$.

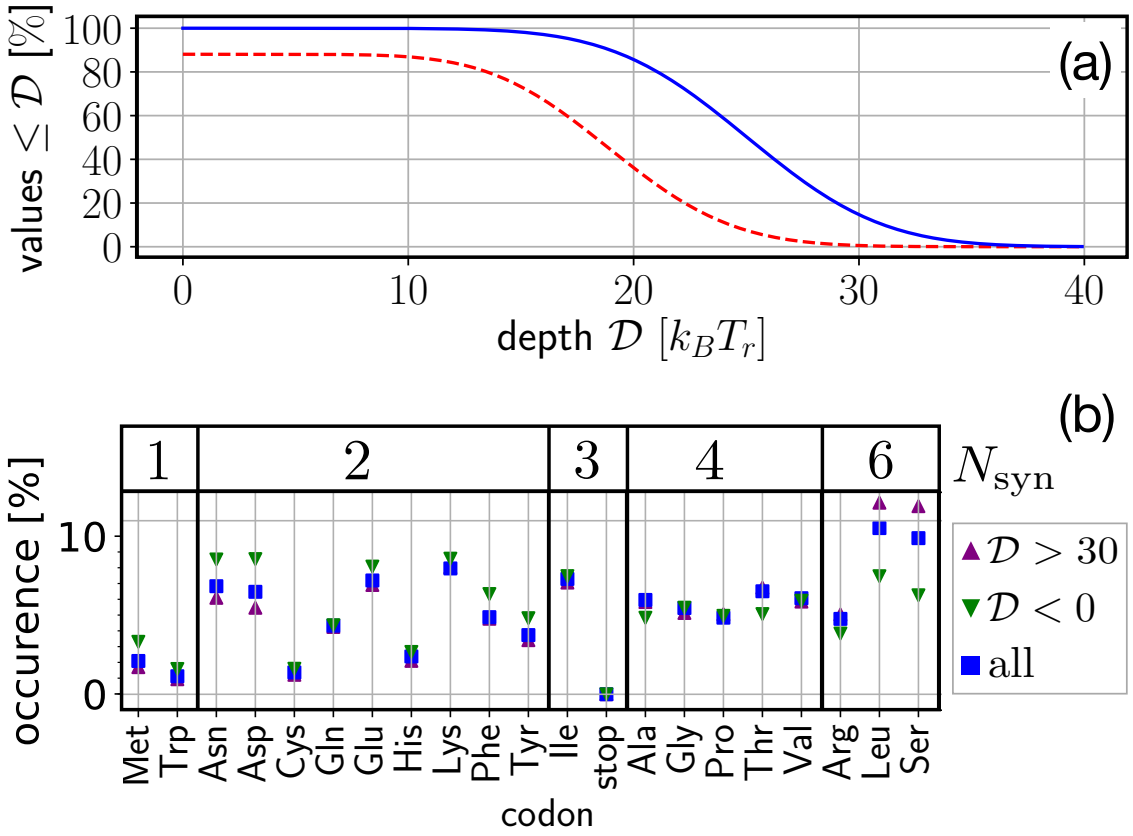


Figure 3.6: (a) Percentage of all positions on genes from yeast *S. cerevisiae* where we created at least a minimum of depth $\mathcal{D}$. Full method (blue continuous curve) and synonymous ground state sequence only (red dashed curve). (b) Full method: probability of a given amino acid to occur for positions with deep minima, $\mathcal{D} > 30\ k_B T_r$ and no minima, $\mathcal{D} < 0$, compared to all values. N_{syn} denotes the number of synonymous codons.

3.7 Conclusion

We have presented a powerful approach to study DNA mechanics, namely to describe sequences by paths through graphs. The weights along the edges of the graphs need to be derived from a mechanical model (as done here) or from experimental data. Specifically we used this approach to determine the best and the worst sequences to be wrapped into nucleosomes and to construct the best positioned nucleosome. Importantly we showed that the degeneracy of the genetic code allows to put mechanical cues even on top of genes to position stable nucleosomes almost anywhere on the genome of yeast with single bp resolution. The very small fraction of places where this is not possible corresponds to gene stretches that contain a higher than average fraction of codons that have no or only one synonymous variant.

Even though we focus here on nucleosomes, we stress that the same set of methods can be applied to any other system featuring bent DNA configurations, e.g. indirect readout of DNA binding proteins [38], protein-induced DNA loops [70], DNA with an affinity to form rings [61, 71] or bent linker DNA in chromatin fibers [72, 73].

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-

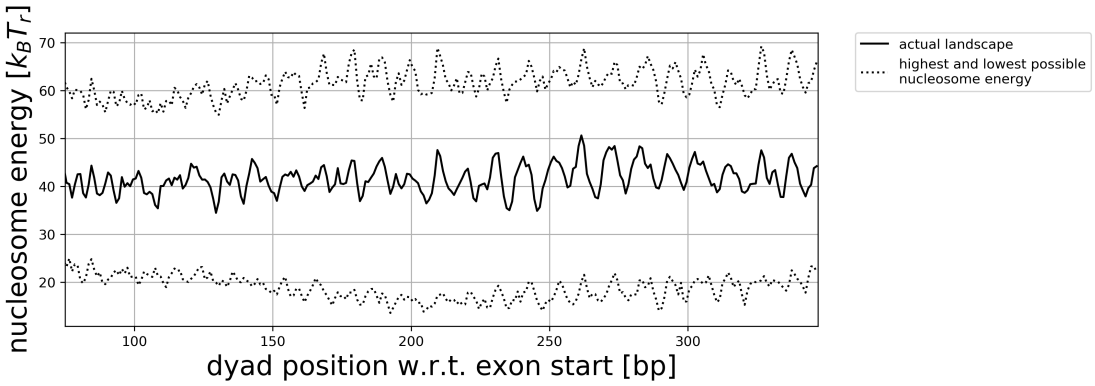


Figure 4.1: The energy landscape for the fourth exon of the human gene Tumor Necrosis Factor (TNF) is depicted by the solid line. The dotted lines depict the highest and lowest possible energy at these positions for any theoretical exon coding for the same amino acid chain, obtained by using a graph 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.

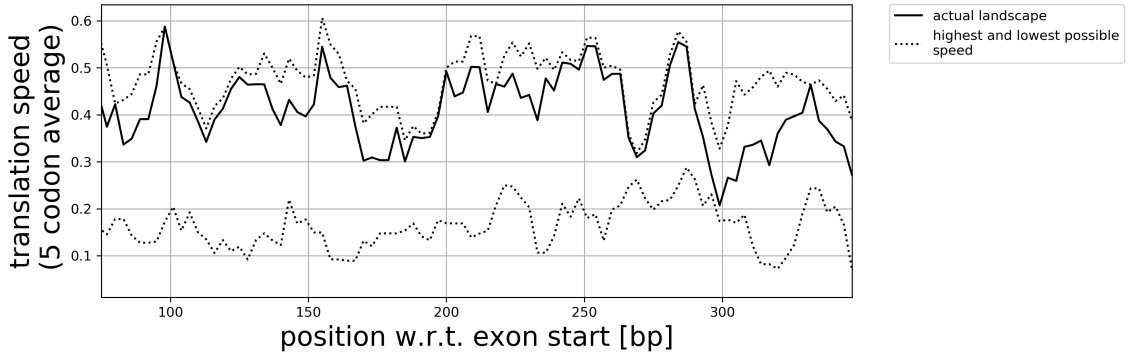


Figure 4.2: The translation speed landscape for the fourth exon of the human gene Tumor Necrosis Factor (TNF) is depicted by the solid line. The dotted lines denote the highest and lowest possible translation speed when codons may be replaced by synonymous codons. We average over five codons to obtain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

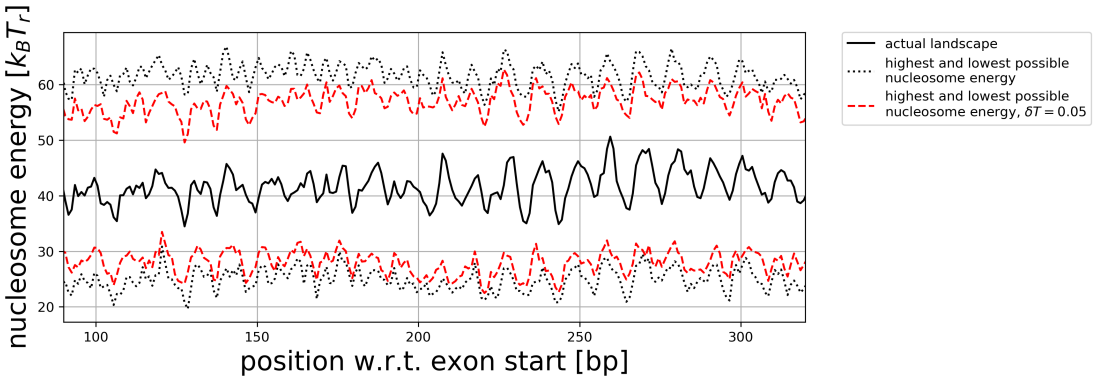


Figure 4.3: Same as figure 4.1 but with the addition of the highest and lowest possible nucleosome energy with a translation speed restriction of $\delta T = 0.05$.

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.

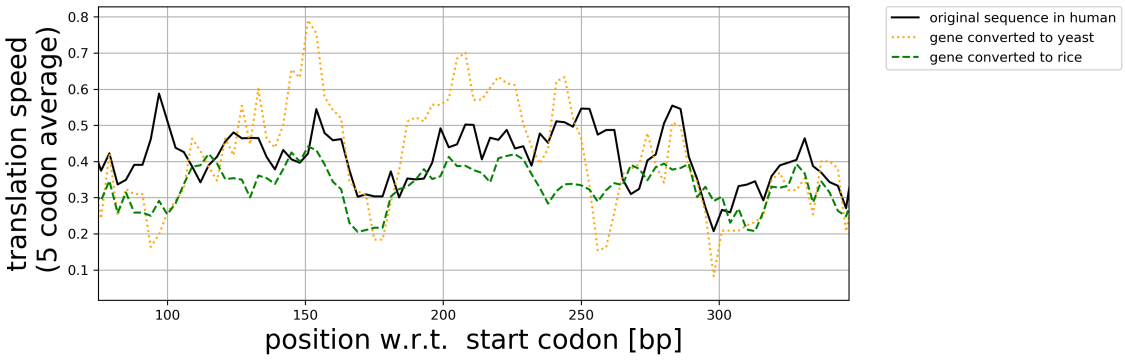
Let A_G be the set of all sequences that code for the same amino acid chain as G. We choose the closest sequence G' such that

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

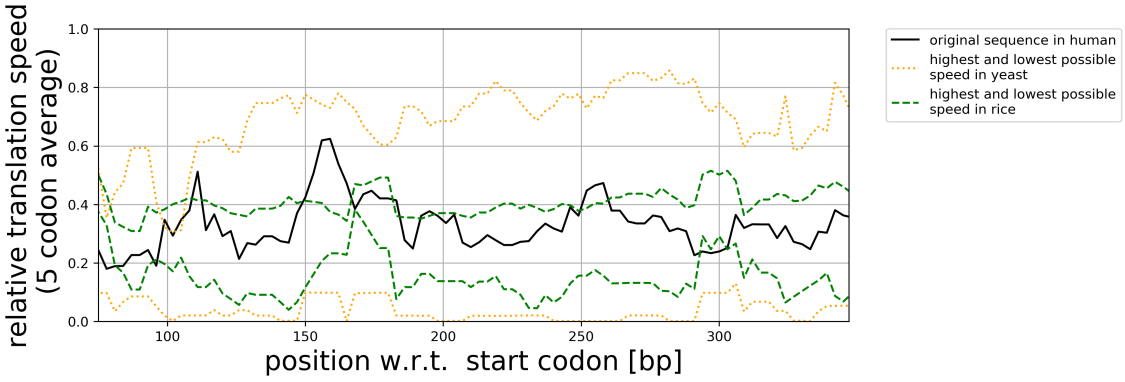
with

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

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

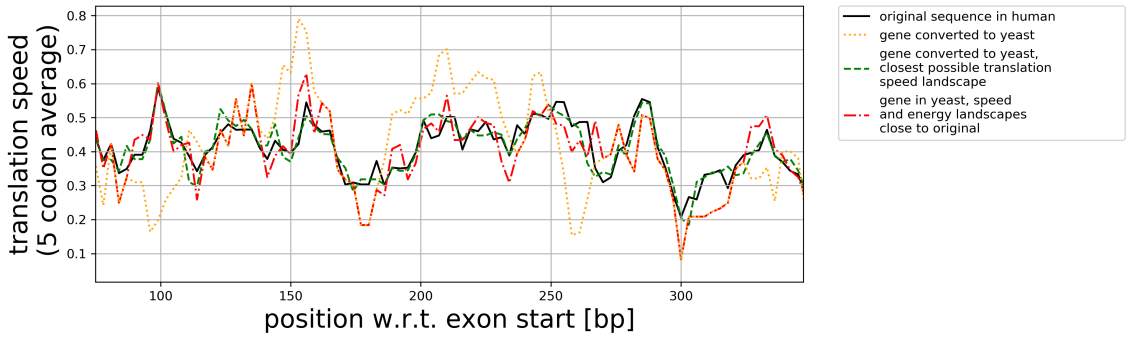


(a)

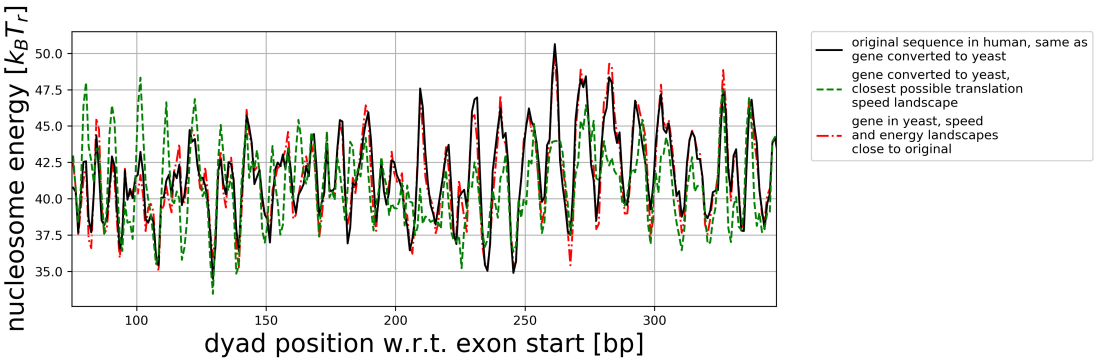


(b)

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

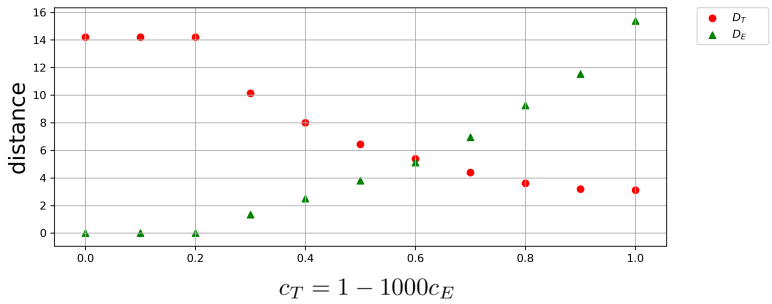


Figure 4.6: The distances D_T and D_E (defined by equations 4.5 and 4.10) are depicted as a function of c_T and c_E. Distance D_T represents the difference between the original translation speed landscape in human of a genetic sequence G (in this case, the fourth exon of TNF) and an altered sequence 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.

Chapter 5

How mechanical information is multiplexed on the transcribed regions of protein-coding genes

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

In previous chapters we demonstrated the extent to which mechanical and genetic information can be multiplexed. We studied the theoretical limits of the nucleosomal energy landscape with no restrictions, as well as restrictions based on the conservation of genetic information and translation speed. In Chapter 3 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. In Chapter 4 we have discussed that the genetic code is not truly degenerate, since different codons may incur differrent translation speed landscapes. However, we have shown that there is still room for the mechanical and translation speed layers of information to co-exist on top of a genetic sequence. Now the question remains whether this actually happens in real genomes. Following in the footsteps of Tompitak et al. [9] we investigate the nucleosome positioning signals in many different organisms around their transcription start sites. By introducing a classification scheme for different types of multiplexing we will show which organisms encode mechanical information on top of protein-coding DNA, and which organisms use different tactics.

5.1 Introduction

We have created a method to investigate multiplexing, which reveals that different organisms use different tactics to create a nucleosome signal near the transcription start sites (TSS). Before we introduce this method, we will shortly discuss the role of nucleosomes in the process of transcription, and how the TSS relates to this role.

5.1.1 Introduction to transcription

Transcription is the process where a sequence of DNA gets copied to RNA. One of the most essential machineries involved in the creation of proteins is RNA polymerase, an enzyme responsible for creating these RNA sequences. There exist many types of RNA, the type that contains the information to create proteins is called messenger RNA (mRNA). This mRNA is created by RNA polymerase, which moves from a transcription start site (TSS) to a transcription terminating site (TTS) in order to copy the bases (and therefore information) on the DNA. A TSS is located within a so-called promoter, a DNA sequence which determines where the RNA polymerase binds a gene. Promoters may contain core promoters, sequences on the DNA to which the transcription machinery can bind, which consists of RNA polymerase and general transcription factors. These general transcription factors aid the RNA polymerase by positioning it at a TSS and direct the initiation of transcription [95]. Different sequences can act as promoters, such as so-called TATA-boxes, initiator sequences and CpG islands [2]. A gene on the DNA can have long-distance transcription-control elements such as enhancers. These enhancers are sequences on the DNA that attract sequence-specific DNA-binding transcription factors, which stimulate transcription [2]. Silencers are the opposite of enhancers; these are sequences on the DNA that inhibit transcription by attracting transcription factors [3].

5.1.2 The role of nucleosomes in transcription

Nucleosomes, too, have an important role in transcription. They are even involved in epigenetic regulation of transcription, where the term epigenetic refers to inherited changes in how cells function that do not result from changes in DNA sequence. The so-called histone tails of nucleosomes can be modified, which can lead to inaccessible genes and therefore inhibited construction of specific proteins. These modifications can be inherited by the offspring of an organism [2].

We will specifically investigate the region around the TSSs. Around these sites, the nucleosome affinity of the DNA sequence seems to be an important factor in positioning nucleosomes [96]. Already in 1988, it has been demonstrated in vivo that nucleosome loss in yeast can lead to increased transcription initiation through activation of promoter elements [97]. In animals, sequence-dependent nucleosome positioning seems to be a mechanism of TSS selection by the RNA polymerase in the absense of core promoters [98]. It seems that different genomic positions may employ different 'tactics' with regard to nucleosome positioning and TSSs. On some genes, nucleosomes may have either a strong or weak affinity to occupy transcription factor binding sites at some locations, making these sites either intrinically inaccessible or accessible to transcription factors [5]. Some genes on some organisms have nucleosome-depleted regions (NDRs) in their promoters, while other promoters contain nucleosome-attracting regions (NARs). Tompitak et al. [9] have shown, using their trinucleotide model, that high nucleosome occupancy near the TSS is encoded on the DNA of multicellular organisms, and that the strength of the nucleosome positioning signals correlates with the complexity of the organism. This supported the hypothesis by Tillo et al. [99], who suggest that NARs are beneficial to organisms with differentiated cell types, since they could suppress genes by default. Cells that need specific proteins would be required to activate corresponding genes. On

the other hand, Vavouri and Lehner suggest that the main reason that these NARs exist is to position nucleosomes in sperm cells. While most of the genetic material in sperm is packaged by protamines, some nucleosomes are retained at GC-rich sequences. These nucleosomes make it possible to transfer paternal epigenetic information encoded on their histone tails to their offspring. Also, these nucleosomes prevent CpG islands from methylation, which keeps these core promoters accessible to transcription factors [35].

By introducing a classification scheme for different types of multiplexing we will show which organisms use the degeneracy of the genetic code to encode mechanical information on top of protein-coding DNA. This scheme incorporates three different DNA regions that exist on genes. These regions are exons, introns and UTRs[1] (UnTranslated Regions). Exons are the parts of a gene that code for a protein. Introns are cut out from the transcribed RNA in a process called splicing, which is required to turn pre-mRNA into mRNA. They do have biologically advantageous functions. For instance, they enable alternative splicing, where exons are ordered differently to code for different proteins [100]. Furthermore, they can modify the expression level of the gene by containing enhancers, which increase transcription of genes, or silencers, which do the opposite [3]. The introns, when cut out of the pre-mRNA, may even help regulate the expression of genes by containing regulatory non-coding RNAs. The positions of introns on genes seem to be important, since they are sometimes conserved throughout long evolutionary times [100].

After splicing, the only regions that remain on the mRNA are the exons and the UTRs. There are two types of UTRs, which exist on the two sides of the exons: in the order of transcription, the 5′UTR is followed by the stitched-together exons, which are followed by the 3′UTR. These UTRs, like the introns, do not code for the protein, but have other important functions such as the post-transcriptional regulation of gene expression, which includes the modulation of the transport of mRNAs out of the cell nucleus, the translation efficiency of the mRNA and the subcellular localization of the mRNAs [101]. Exons, introns and UTRs exist between the TSS and TTS. We will see later that all three regions bear responsibility for any mechanical signals that may exist on protein-coding DNA.

5.1.3 Overview

In the next section we introduce a scheme to dissect the mechanical signals on the DNA into different categories. Section 5.3 provides the technical details on how to perform this scheme, and uses several animal species to demonstrate its results. In sections 5.4 and 5.5 the signals of many animals and plants are compared. These sections show that animals and plants employ different 'tactics' to create mechanical signals on the DNA. For instance, some organisms use mainly coding, and other use mainly noncoding DNA to encode nucleosome positioning signals. Section 5.6 demonstrates for *Oryza sativa*, rice, that not just the degeneracy of the genetic code, but even the amino acid sequence itself creates a nontrivial mechanical signal on the DNA. Section 5.7 discusses the possibility that the nucleosome signal of coding DNA is a result of a signal on the mRNA, such as translation speed. For human we show that this hypothesis is extremely likely, while for rice it seems to be the other way

[1] It would be more consistent to call them utrons

around. Finally we end this chapter with a conclusion and outlook.

5.2 Multiplexing: Intraregional signals and interregional signals

Now we can discuss different types of multiplexing in genomes. Multiplexing simply refers to the existence of multiple signals on the same medium. There are many types of multiplexing, some more trivial than others. Signals can be separated on a medium by space (e.g. two people writing on two sides of a single piece of paper) or time (e.g. people only speaking when it's their turn during a debate). Signals do not need to be separated by time nor space, as in the case of DNA, where information on the nucleosomal energy and genetics exists on the very same base pair. It is our goal to investigate the types of multiplexing involved in the multiplexing of mechanics and genetics. We will use perhaps the simplest model for nucleosome positioning possible: GC content. GC content correlates well with DNA stretches that have a higher affinity for nucleosomes [8, 9].

To do this we will define two relevant types of multiplexing. The two types of multiplexing we consider are the following:

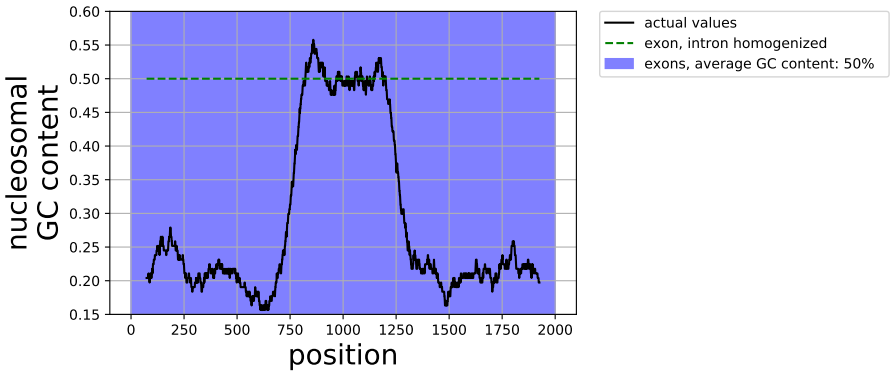
- intraregional multiplexing/intraregional positioning signals
- interregional multiplexing/interregional positioning signals

5.2.1 Intraregional signals

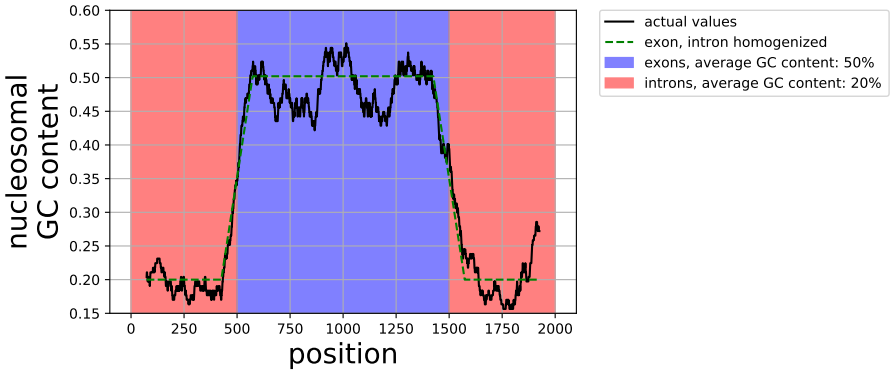
We will refer to intraregional multiplexing as intraregional positioning signals. These are signals that exist on a single region, see Fig. 5.1a for a theoretical example. In the case of exons, we get multiplexing of protein-coding information and mechanical information. Since non-exonic DNA can be functional, it would be difficult beforehand to estimate how much freedom introns and UTRs have to affect the nucleosome positioning landscape. On exons, the freedom to code for mechanical signals comes from the degeneracy of the genetic code. Introns possibly have more freedom to incorporate mechanical signals as well. We hypothesize that this relative freedom influences or even determines how much mechanical information the different regions contain on average.

5.2.2 Interregional signals

Interregional positioning signals are based on the fact that different regions on DNA have, on average, different mechanical properties. Exons generally have, on average, a higher GC content -and therefore higher flexibility- than introns. Therefore, alternating introns and exons can lead to a positioning signal, without a positioning signal existing on any separate region. See Fig. 5.1b for a theoretical example of interregional positioning signals.



(a)



(b)

Figure 5.1: In (a), we show a theoretical example of an intraregional positioning signal. In (b), we see a theoretical example of an interregional positioning signal. In both cases, the black line depicts a nucleosomal GC landscape on a range of bp positions. In green (interrupted line) the landscape is shown where all intron and exon values are homogenized by replacing them by their average values. For the intraregional signal in (a), a peak exists on a single region: be it an intron, an exon or a UTR, in this example on an exon. In (b), the interregional signal exists mostly because of differences in the average GC content of introns and exons. This example was created by having the introns contain G or C with a 20% chance, whereas the bases in the exon were chosen with equal probabilities.

5.3 Intraregional and interregional signals in a real genome

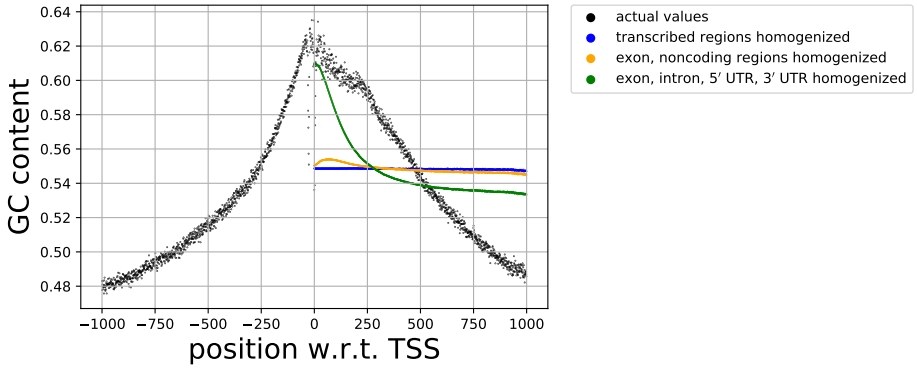
In this section we demonstrate for *Homo sapiens* how one can find these intraregional and interregional signals. All genomic data used throughout this chapter was acquired from the Ensembl Project website (www.ensembl.org), using their web-based tool Biomart. For a step-by-step guide on how to obtain this data, see appendix D.1.

5.3.1 Distinguishing the positioning signals by homogenizing

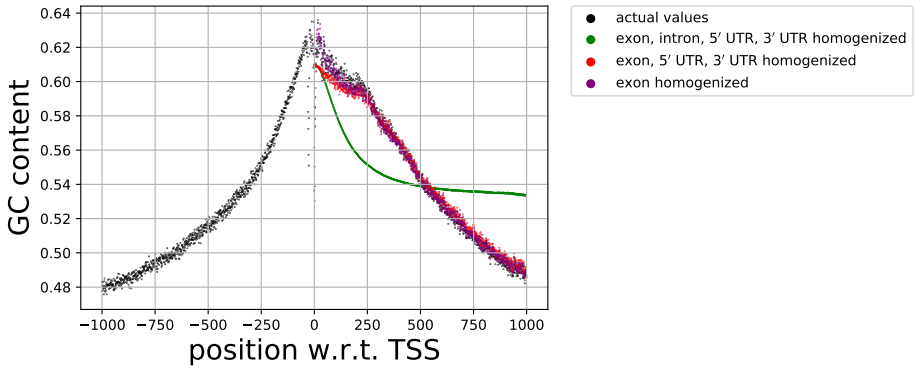
Now we introduce a method to find out whether a signal is intraregional or interregional. Interregional signals are caused by differences in averages between regions. When we replace all regions by their averages, we obtain the interregional signal size. The intraregional signal is simply the difference between the interregional signal and the actual signal. We call replacing all values of a region by their average value *homogenizing*. The results of this trick is visible in Figs. 5.2-5.5.

Fig. 5.2 depicts the GC landscapes near the TSS of human. The actual average GC content around the TSS of human genes is shown in black. In blue, we see what happens when all transcribed regions (lumped together) are homogenized, in orange we homogenized exons and noncoding regions separately, in green exons, introns and the 5′UTRs and 3′UTRs are homogenized. We can see that it is important to homogenize the introns and UTRs separately, since the green curve is much closer to the actual values than the orange curve. We can also see that the interregional positioning signal only partially explains the overall signal. In Fig. 5.2(b), more curves are depicted, where, compared to the green curve, the actual values of the introns are used to obtain the red values, instead of an average. To get from red to purple we use the real values of the UTRs as well. We go from purple to black by also including the actual values of exons. It seems that, for human, introns have the biggest effect out of all intraregional signals. This is a somewhat incomplete statement however, since humans have more intronic base pairs than exonic base pairs near the TSS. We will evaluate this effect in section 5.3.2. To obtain a more 'realistic' depiction of nucleosome signals, we will look at the *nucleosomal* GC content, i.e. the GC content per nucleosome position. Since nucleosomes contain 147 bp of DNA, we need to average over stretches of 147 bp. The result for human is depicted by Fig. 5.3. There are no qualitative differences between Fig. 5.2 and Fig. 5.3, so our analysis is valid for nucleosomal GC content as well. We will continue looking at nucleosomal GC content since it is more physically relevant.

Fig. 5.4 depicts the same as Fig. 5.3 but for *Gallus gallus*, chicken. No qualitative distinctions between chicken and human are visible. In fig. 5.5 we see *Tetraodon nigroviridis*, a pufferfish. In striking contrast to chicken and human, this fish has a signal almost entirely caused by interregional positioning signals, i.e. caused by the differences in average values of its regions (the green and black curves are practically the same). Fig. 5.6 depicts *Caenorhabditis elegans*, a nematode. For this animal, too, the signal is mostly caused by interregional positioning signals. It is possible that higher organisms have evolved to contain intraregional positioning signals in addition to the interregional signals.

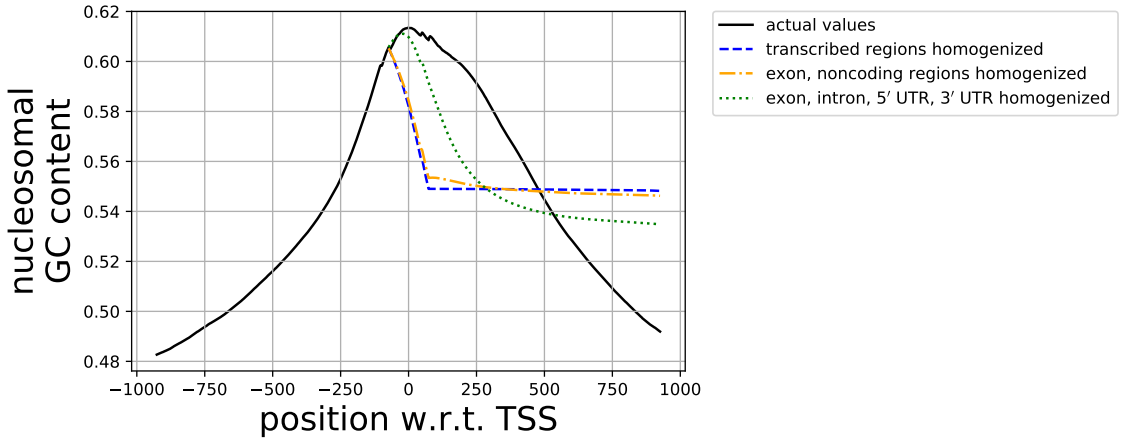


(a)

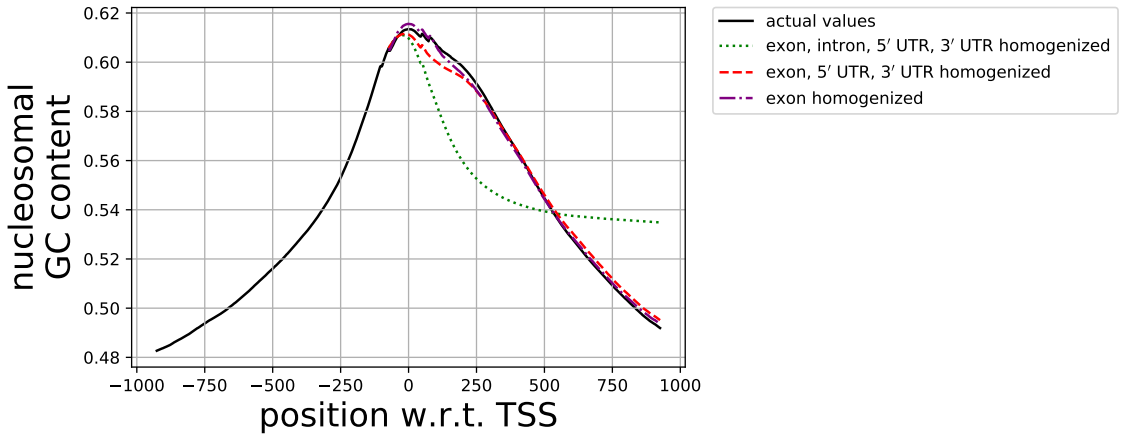


(b)

Figure 5.2: This figure describes *Homo sapiens*. In (a), we see in black a depiction of the actual average GC content around the TSS of human genes. In blue, we see what happens when all transcribed regions are homogenized, in orange and green subsets of these regions are homogenized. These curves reveal that it is important to homogenize the introns and UTRs separately, since the green curve is much closer to the actual values than the orange curve. Still we can see that the interregional positioning signal only partially explains the overall signal. In (b), more curves are depicted, where, compared to the green curve, the actual values of the introns are used to obtain the red curve. To get from red to purple we use the real values of the UTRs as well. We go from purple to black by also using the actual values of exons. For human, the introns have the biggest effect out of all intraregional signals.

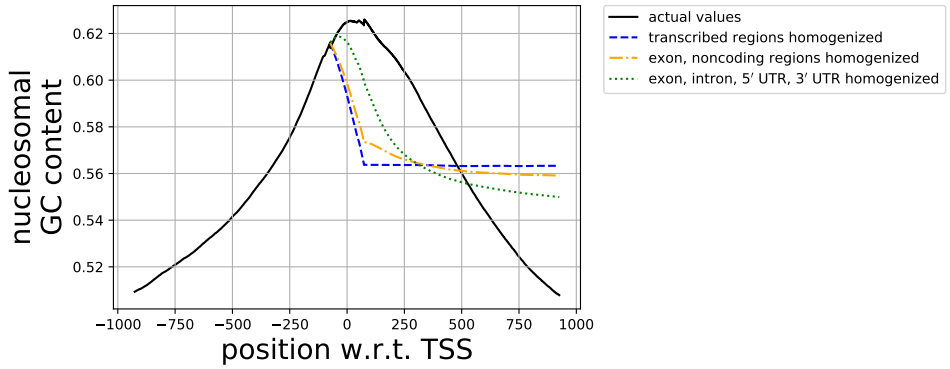


(a)

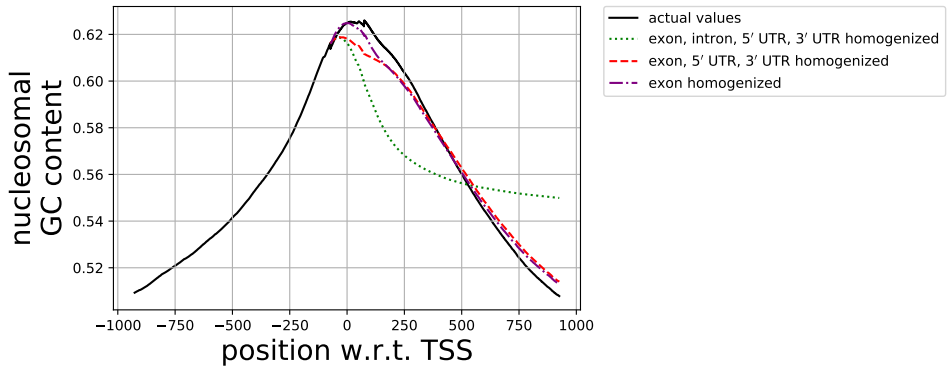


(b)

Figure 5.3: Same as Fig. (5.2) but depicting nucleosomal GC content (again for human), i.e. the GC content per nucleosome position: stretches of 147 bp.

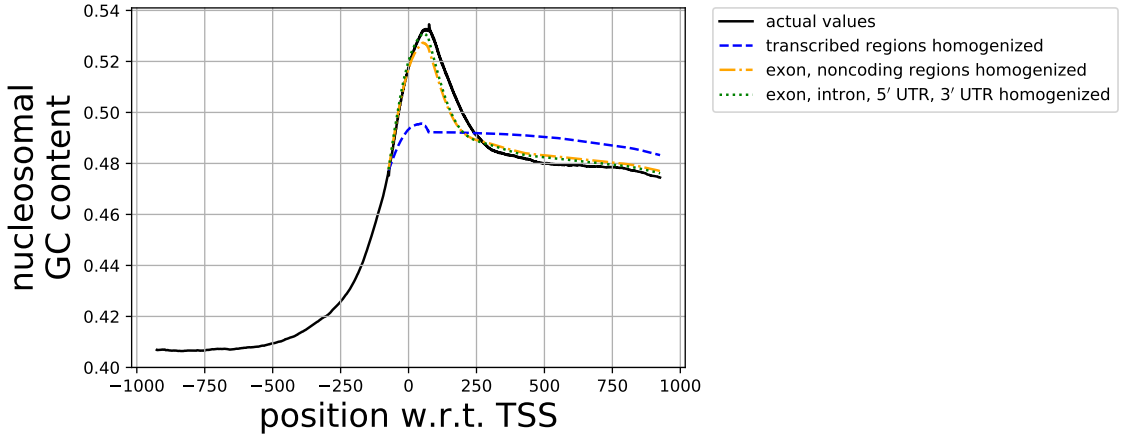


(a)

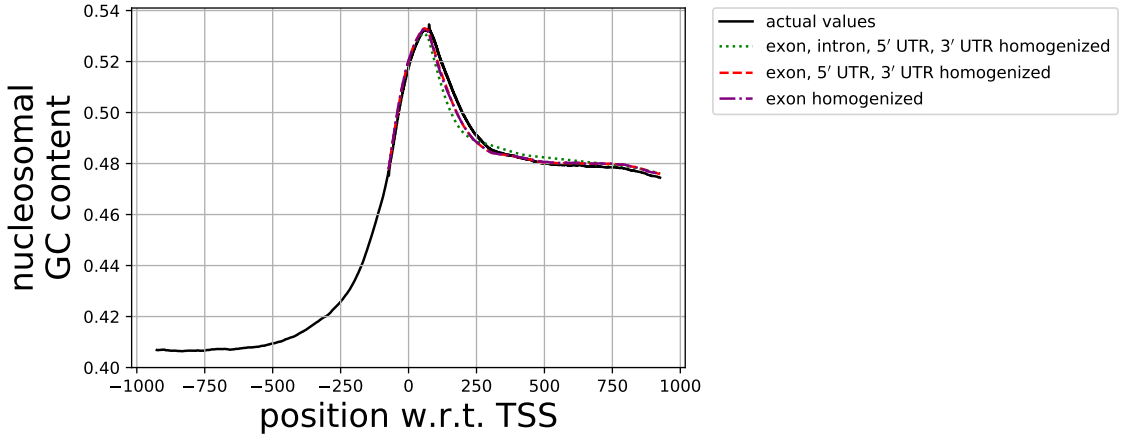


(b)

Figure 5.4: Same as Fig. (5.3) but for *Gallus gallus*, chicken. No qualitative differences are visible between human and chicken.

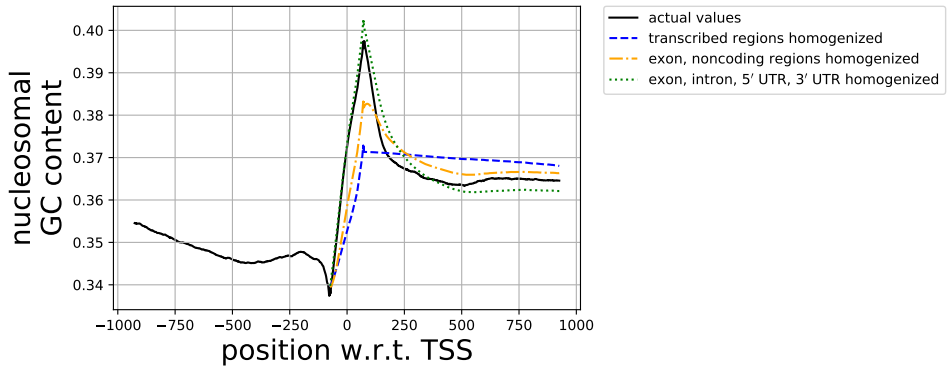


(a)

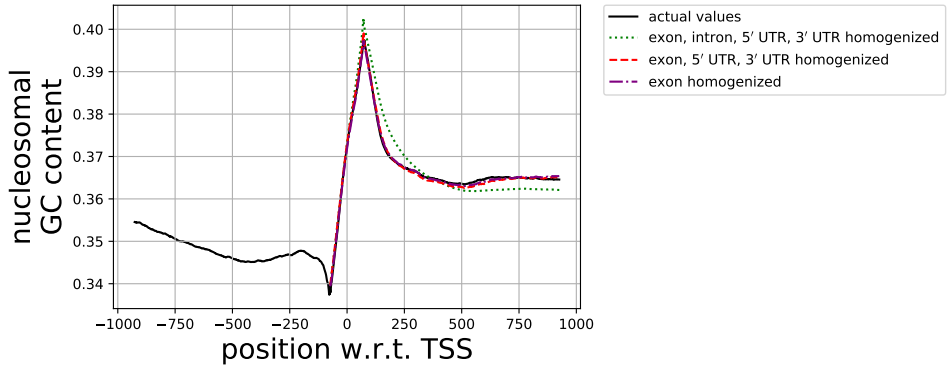


(b)

Figure 5.5: Same as Fig. (5.3) and Fig. (5.4) but for *Tetraodon nigroviridis*, a pufferfish. The signal of this animal is caused by interregional positioning signals, i.e. caused by the differences in average values of its regions.



(a)



(b)

Figure 5.6: Same as Fig. (5.3) through Fig. (5.5) but for *Caenorhabditis elegans*, a nematode. For this animal, as well as *Tetraodon nigroviridis*, the signal is mostly caused by interregional positioning signals.

5.3.2 Obtaining the strength of the positioning signals

Homogenization provided us a qualification scheme for the signals. Now we will compare the different types of signals in a single organism. Therefore we need a way to quantify signal sizes. Additionally, a proper quantification scheme enables us to compare signal sizes between organisms as well. We came up with a basic (and therefore relatively unbiased) formula for signal size. Let $L = \{L_{-1000+147/2}, L_{-999+147/2}, ..., L_{1000-147/2}\}$ be the original nucleosome GC landscape, where for every L_i, i depicts the position on the landscape relative to the TSS. The term 147/2 comes from the fact that these are nucleosome positions. We introduce $h_{\text{regions}}(L)$ to depict the homogenization of a landscape by replacing one or multiple regions by their corresponding average values. Then the signal size of the regions is given by the linear difference between the real landscape and the homogenized landscape.

$$\text{signal size of regions} = \sum_{x=-1000}^{1000} |L(x) - (h_{\text{regions}}(L))(x)| \,. \tag{5.1}$$

For example, the signal size of exons is given by $\sum_x |L(x) - (h_{\text{exons}}(L))(x)|$ and the signal size of all transcribed regions is given by $\sum_x |L(x) - (h_{\text{transcribed regions}}(L))(x)|$. The total signal size is found by homogenizing all regions as one region, i.e. the difference between the landscape and its average value:

$$\text{total signal} = \sum_x |L(x) - (h_{\text{everything}}(L))(x)| \tag{5.2}$$

and the signal size for exons and introns homogenized separately is given by

$$\text{exon, intron signal size} = \sum_x |L(x) - (h_{\text{exon, intron}}(L))(x)|. \tag{5.3}$$

Results for human are shown in Fig. 5.7. In Fig. 5.7(a) we can see that the intraregional signals (orange) are much larger than the interregional signals (blue). Exon and UTR signals (green and purple) seem nonexistent next to the intron signal size. This is because introns are much more prevalent near the TSS than exons and UTRs. Fig. 5.7(b) corrects for the occurence of exons, introns and UTRs by showing the intraregional signal sizes per bp. The signal size per bp of introns is only twice as large as that of exons, showing that exons are contributing to intraregional multiplexing as well! It was just obfuscated by the fact that humans do not have many exon bps near the TSS. From an evolutionary point of view, it seems that exons, too, have evolved to contain mechanical information. The fact that introns have a much higher signal per bp could suggest that introns are easier to evolve without them losing other functionalities.

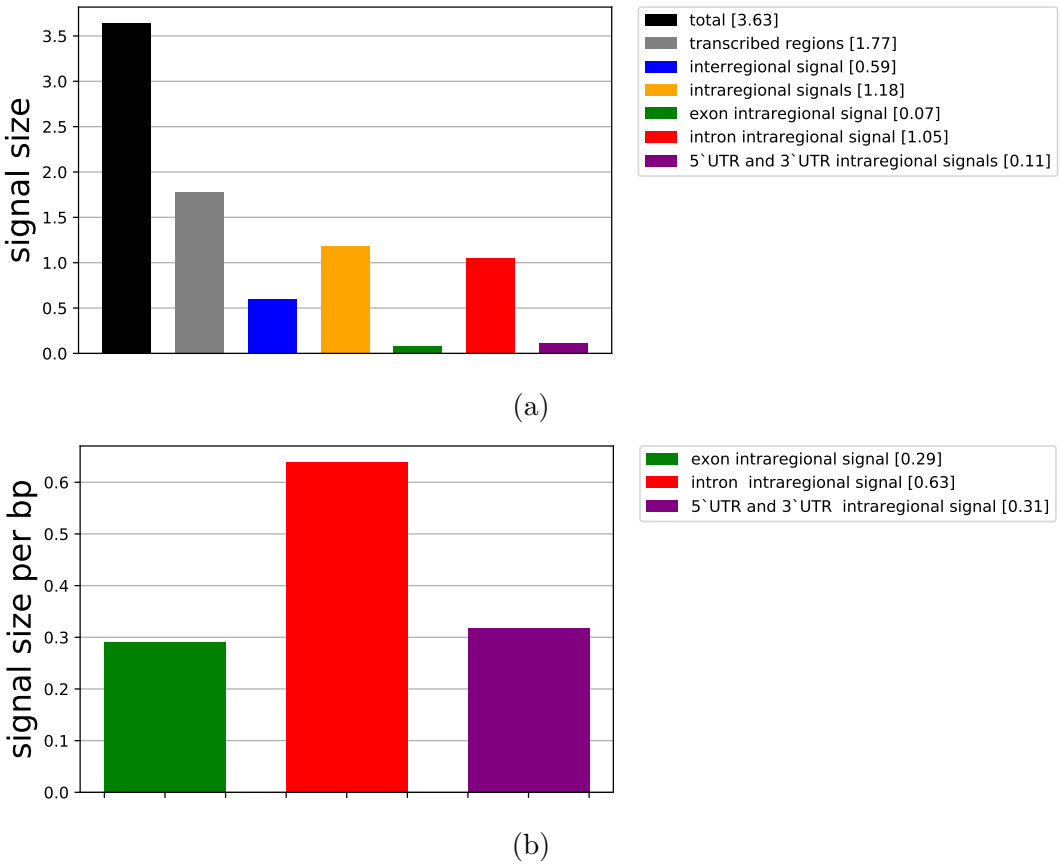


Figure 5.7: In (a) we see the signal sizes for several types of signals in human. The intraregional signals are much larger than the interregional signals. Exon and UTR signals seem nonexistent next to the intron signal size. In (b) we see the signal size per bp for the three separate intraregional signals. Because introns are much more prevalent near the TSS than exons and UTRs, its relative effect compared to exons and UTRs is only twice as big. This shows that exons and UTRs do contribute to intraregional multiplexing.

5.4 Signals on many animals

We can extend our research to many animals. Table D.1 of appendix D.2 depicts the list of animals used to obtain data. It depicts the names of the organisms as they appear in Biomart, as well as their real Latin names and a short description of the animal. The order of this list is based on the species tree as maintained by the Compara team of Ensembl [102]. It roughly reflects the distance between animals on a phylogenetic tree. Animals without UTR data were excluded from analysis and do not appear in this list.

In Fig. 5.8, the top figure depicts the total signal sizes for these animals and the middle depicts the signal sizes of their transcribed bases only. Roughly speaking, the closer an organism is (genetically) to human (or other higher-order animals) the higher the transcribed region signal size. The bottom figure of Fig. 5.8 depicts the fraction (in percent) that the intraregional signal contributes to the total signal in the transcibed region. Roughly speaking this percentage goes up for higher-order animals. Animals such as *D. melanogaster* and *C. intestinalis* do not follow this

trend. This warrants a closer look at these organisms, see Fig. 5.9. In this figure we see that, for *D. melanogaster* and *C. intestinalis* the interregional signal is stronger than the overall signal. Apparently the intraregional and interregional signals oppose each other.

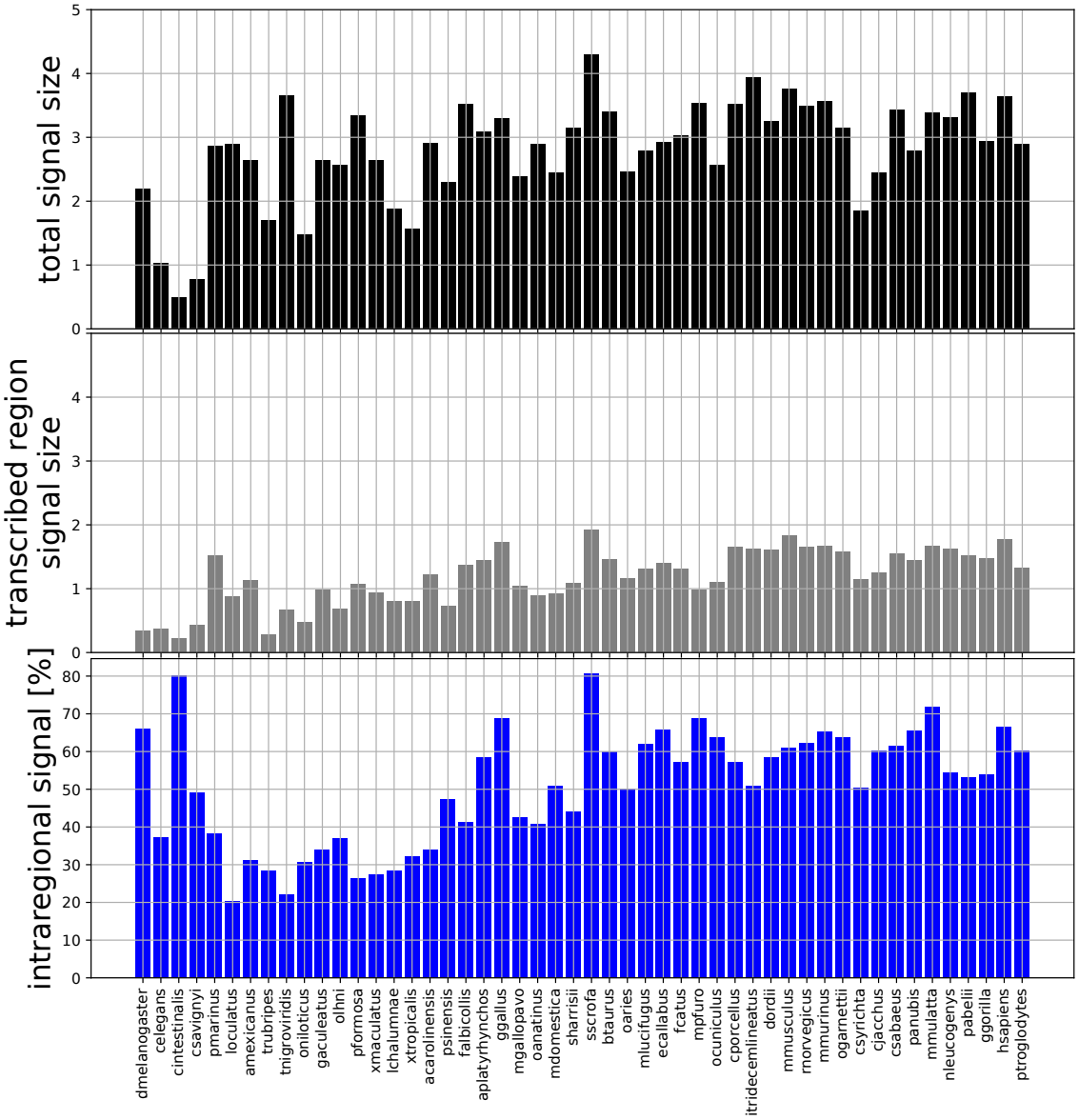
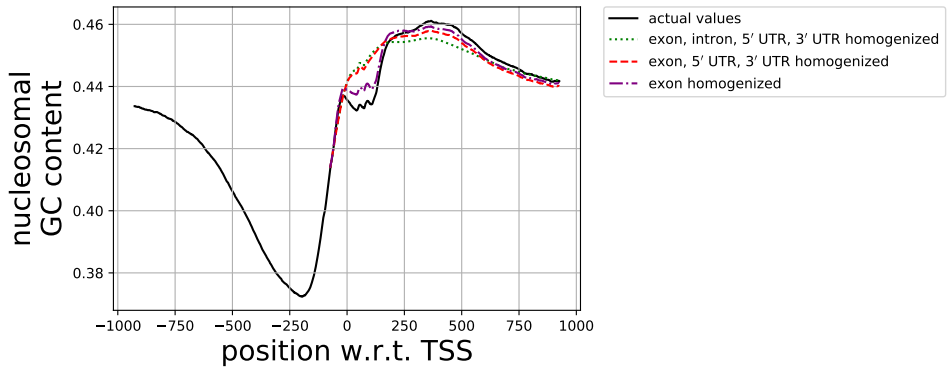
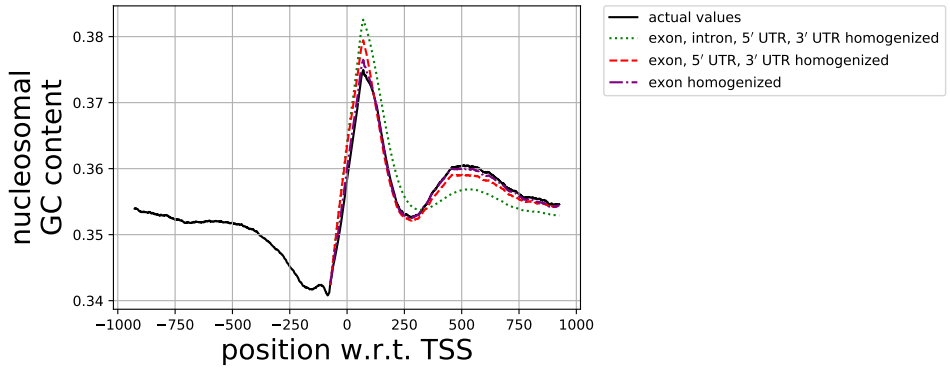


Figure 5.8: This figure depicts the signal sizes of many animals. At the top we see the total signal sizes of the animals, the middle figure depicts the signal sizes of the transcribed bases only, at the bottom we find the fraction (percent) that the intraregional signal contributes to the total signal in the transcibed region.

We can again distinguish the three separate intraregional signals for exons, introns and UTRs. This is depicted by Fig. 5.10(a). Fig. 5.10(b) depicts the signal sizes per base pair, showing that the intron values are not that exceptionally large compared to UTR and exon values. To get a better overview of the relationship between exon and intron intraregional signal we combine our results in scatter plots. In Fig. 11(a) we plot for each animal the intraregional signal sizes of exons vs. introns.



(a)



(b)

Figure 5.9: Same as sub-figures (b) from Figs. 5.2-5.5, but for (a) *d. melanogaster* and (b) *c. intestinalis*. Note the interregional signals being stronger than the overall signal for both organisms.

In general, when exon signal size increases, intron signal size increases much more. Using a least-squares approach, we find a relation of intron signal size= 4·exon signal size plus a constant, and a Pearson correlation coefficient of 0.57. In Fig. 5.11(b) we see the signal sizes per base pair. The relationship then is intron signal size per bp= 2·exon signal size per bp plus a constant and a correlation coefficient of 0.88. The correlation coefficient is much higher when the signal sizes are depicted per bp, which suggests that this is a more relevant representation of the data.

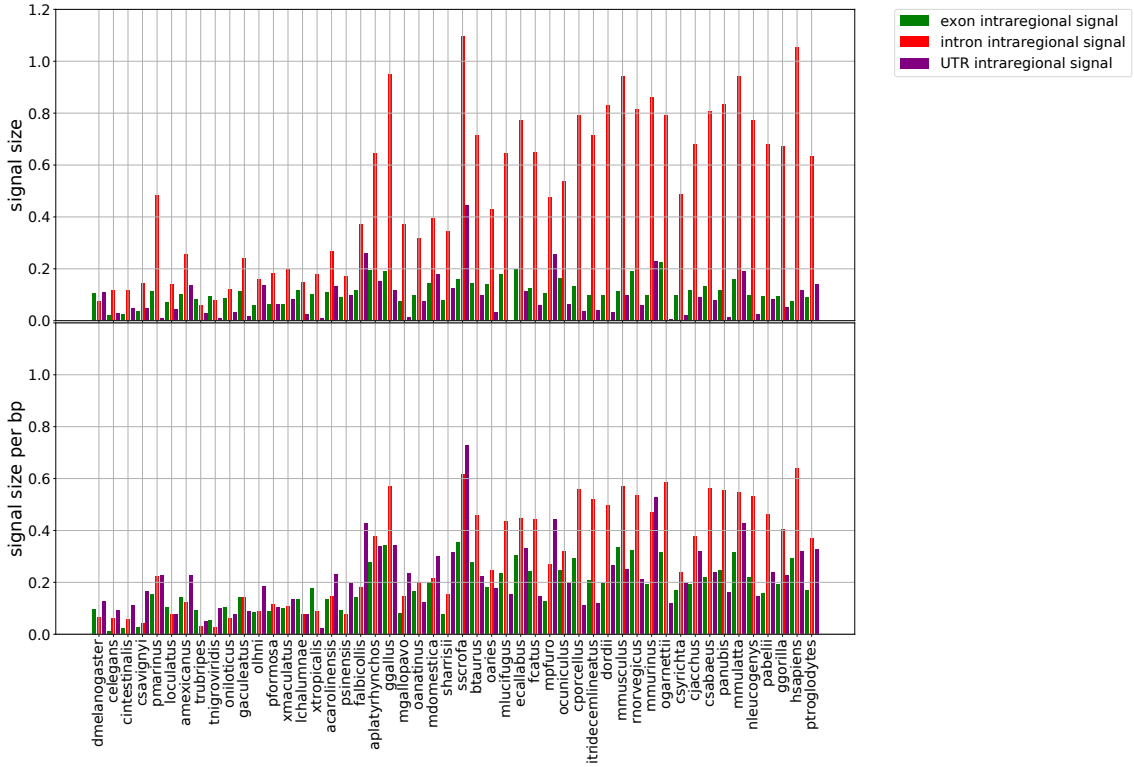
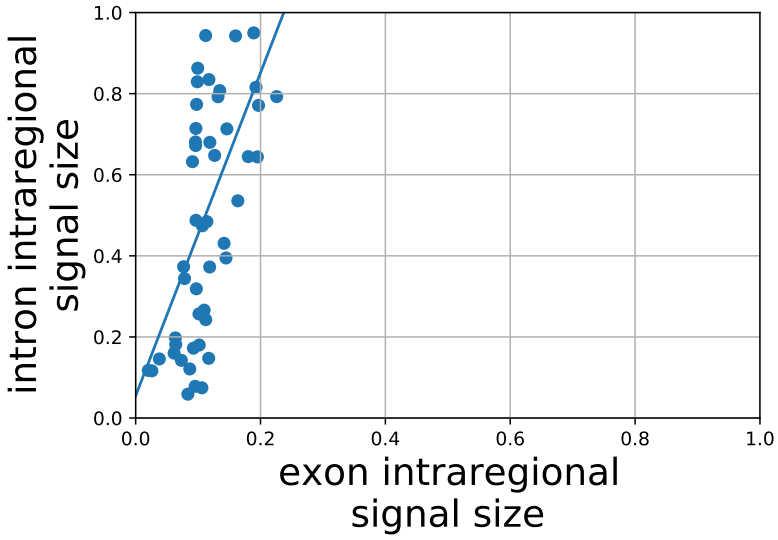
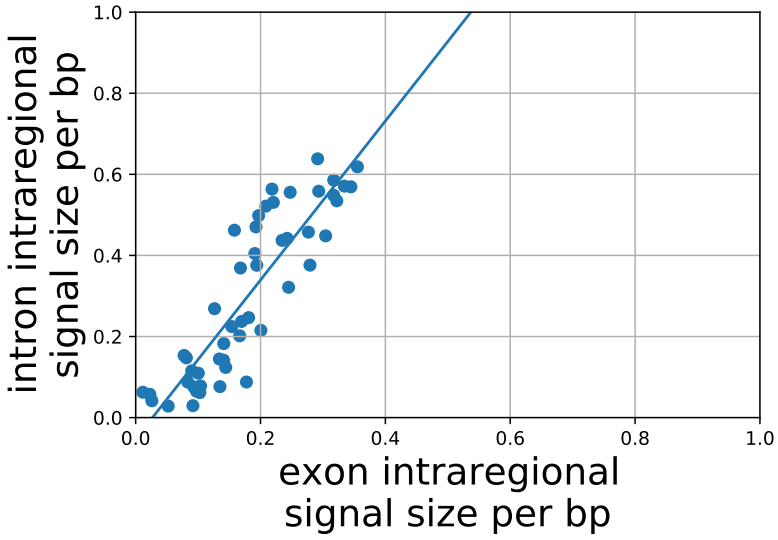


Figure 5.10: This figure depicts intraregional signal sizes of many animals. In the top figure we see that the intron signals dominate. In the bottom figure we see the signal sizes per bp, where the differences between exon, intron and UTR sizes are much closer to each other, introns still being more important in general.



(a)



(b)

Figure 5.11: This figure depicts the relationship between exon and intron intraregional signals. In (a) we see data for many animals. Generally speaking, when exon signal size increases, intron signal size increases much more. Using a least-squares approach, we find a relation of intron signal size= 4·exon signal size plus a constant, and a Pearson correlation coefficient of 0.57. In (b) we see the signal sizes per base pair. The relationship then is intron signal size per bp= 2·exon signal size per bp plus a constant and a correlation coefficient of 0.88. The correlation coefficient is much higher when the signal sizes are depicted per bp, which suggests that this is a more relevant representation of the data.

5.5 Signals on plants

Above we discussed animal genomes. Here we will discuss plants. We will discuss a few types of plants. Due to time constraints, our collection of plant genomes is considerably smaller than the set of animal data. Therefore we investigate a smaller

but diverse range of plant organisms.

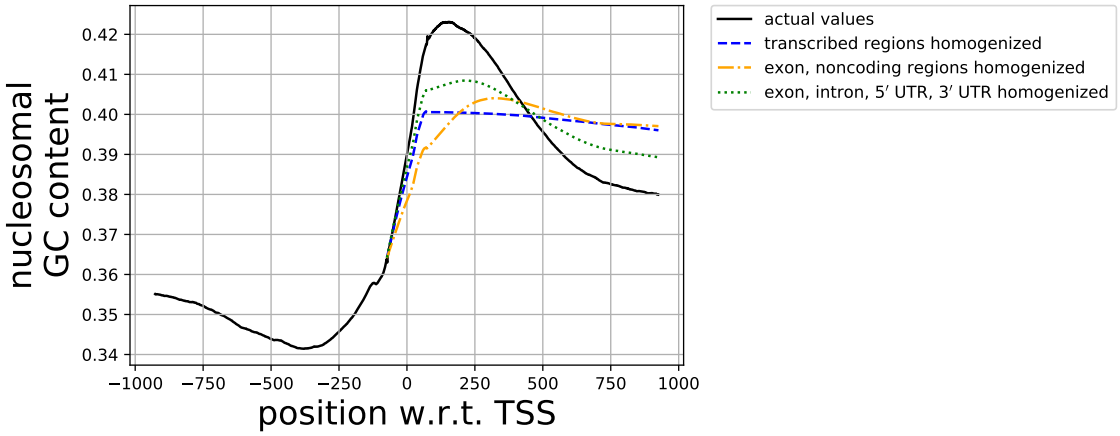
The first plant we will discuss is a tree, *P. persica*, the peach tree. Fig. 5.12 shows that a large part of the signal is caused by interregional signals, see the green dotted curve. The difference between this curve and the black curve, i.e. the intraregional signals, is almost entirely caused by exons and UTR, not by introns. We can see this because the difference between the green dotted curve and the red interrupted curve is negligible.

Fig. 5.13 depicts *A. thaliana*, thale cress, which is a small flowering plant and model organism. This plant is similar to *P. persica*, except that the effect of the UTRs is much lower, and the peak is further away from the TSS. This raises the question what the reason could be for positioning a nucleosome further upstream of the TSS. It might affect the function of nucleosomes as switches for genes.

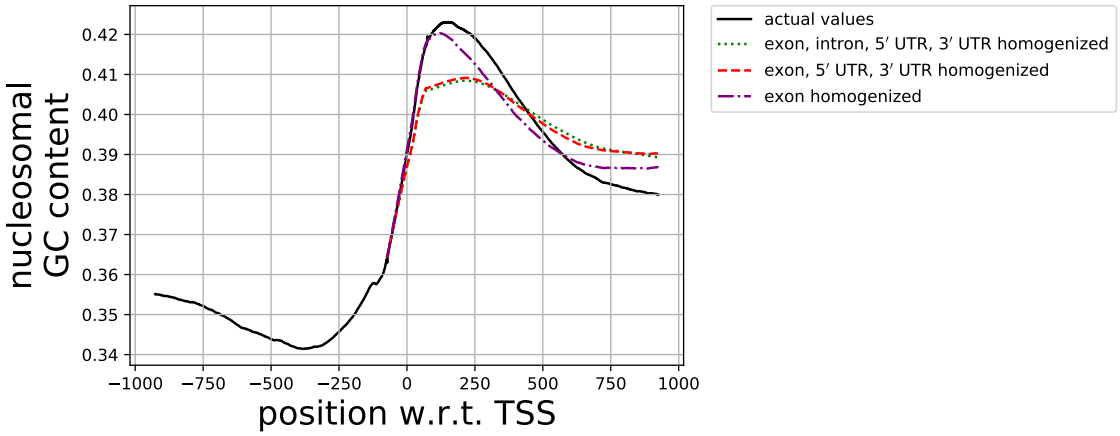
Fig. 5.14 depicts *S. tuberosum*, potato, which behaves very different from peach tree and thale cress. While the overall signal seems similar, it is almost entirely caused by interregional signals.

Fig. 5.15 depicts *O. sativa*, Japanese rice. Its signal is much stronger than the signals for the other plants. The intraregional and interregional signals are both strong and create a signal that is stronger than the signals for the animals in this thesis. Rice is a member of a group of plants called cereal grains, cultivated grasses. It turns out that other cereal grains, such as wheat and maize (Figs. 5.16 and 5.17), also have such a large GC signal near the TSS. Possibly, the strong GC signal are related to stronger nucleosome positioning signals, therefore stronger retention of epigenetic information in the offspring of these plants, which may have been a factor in breeding the wide variety of grains we consume today. It may be possible that either this signal appeared when humans started cultivating the grains, or that they already existed before. This hypothesis is tested by looking at Fig. 5.18, which depicts *L. perrieri*, a cutgrass from Madagascar not used for consumption. This plant also shows a strong signal like the cultivated grasses (i.e. cereal grains), only slightly weaker. Also, a cultivated plant such as potato does not have such a strong signal, suggesting that grasses simply have a strong GC signal, independent on whether they are cultivated or not. It may be that these pre-existing signals enhanced the inheritance of epigenetic information.

Fig. 5.19 depicts C. reinhardtii, a single-celled alga. It turns out that this alga has a strong GC signal near its TSS which is very different from the signals we have seen for the other plants or for the animals we have seen so far. The signal is periodical, see Fig. 5.19(a) but this periodicity may be unrelated to nucleosomes, since it disappears when studying nucleosomal GC content, see Fig. 5.19(b). The periodical undulations are mostly caused by the intraregional signals of UTRs but are also slightly enhanced by the exons, suggesting that the signal is not some quirk of this organism's UTRs but an evolutionary advantageous signal on the DNA.

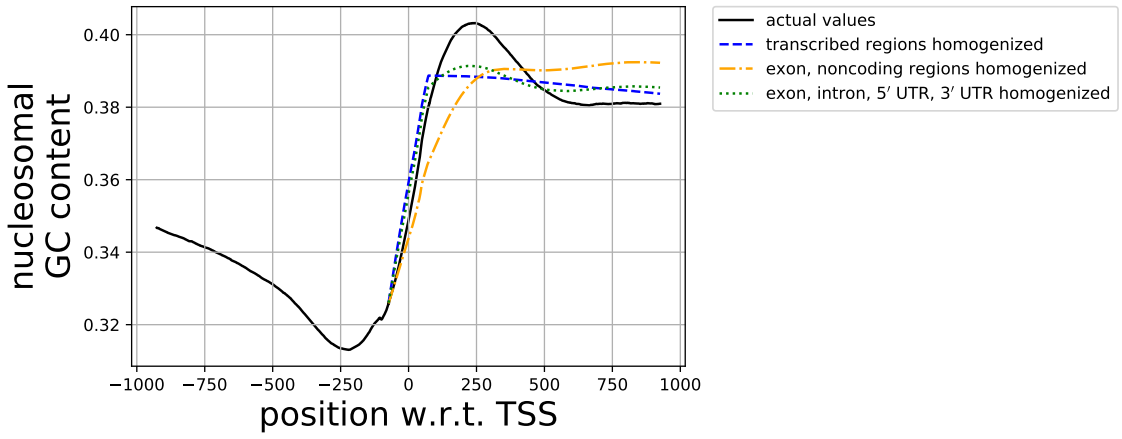


(a)

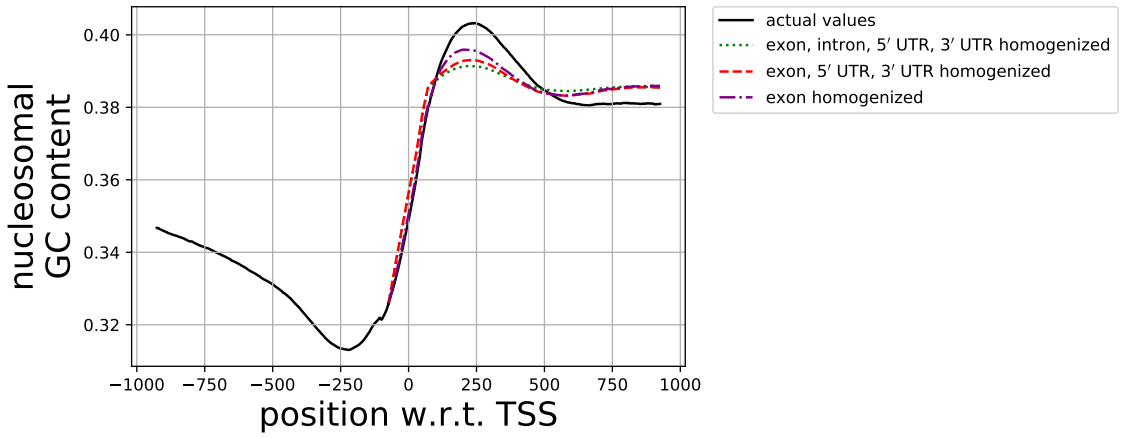


(b)

Figure 5.12: Nucleosomal GC content of *P. persica*, peach tree. As in Fig. 5.2, (a) depicts in black the actual average GC content around the TSS of human genes. In blue, we see what happens when all transcribed regions are homogenized, in orange and green subsets of these regions are homogenized. In (b), more curves are depicted, where, compared to the green curve, the actual values of the introns are used to obtain the red curve. To get from red to purple we use the real values of the UTRs as well. The intraregional signal of introns turns out to be neglegible, while the intraregional signals of UTRs and exons play a big role in creating the GC peak.

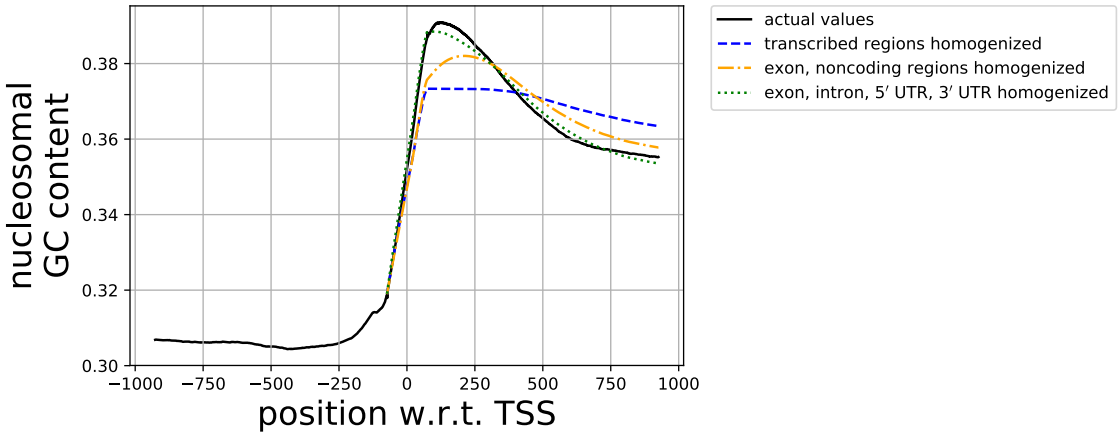


(a)

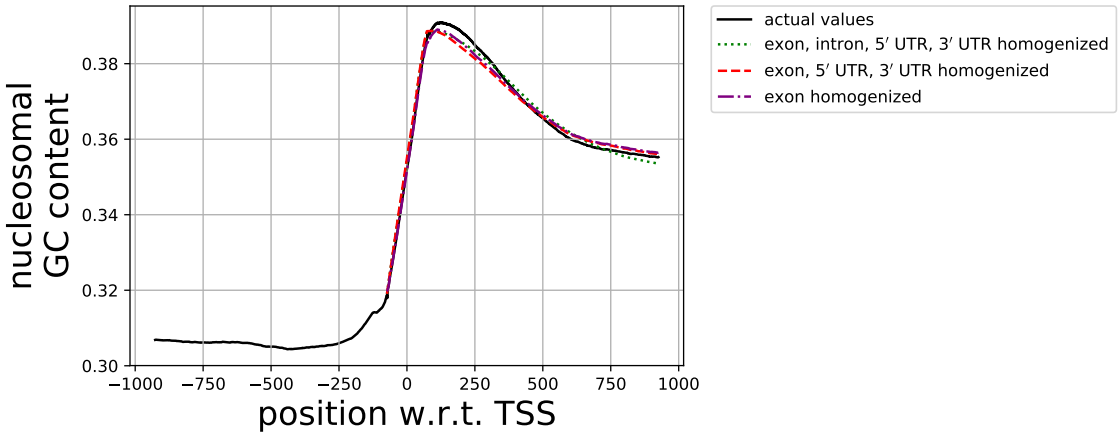


(b)

Figure 5.13: Same as Fig. 5.12 but for *A. thaliana*, thale cress. Out of the three intraregional signals, the exon signal dominates. For this plant, the introns do have some small contribution to the signal, visible around position 250. This contribution helps create the peak at bp position 250.

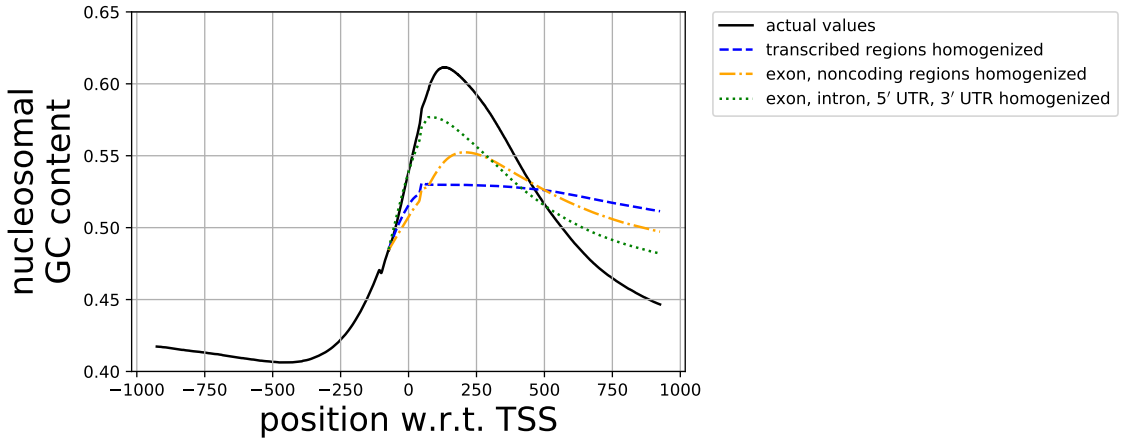


(a)

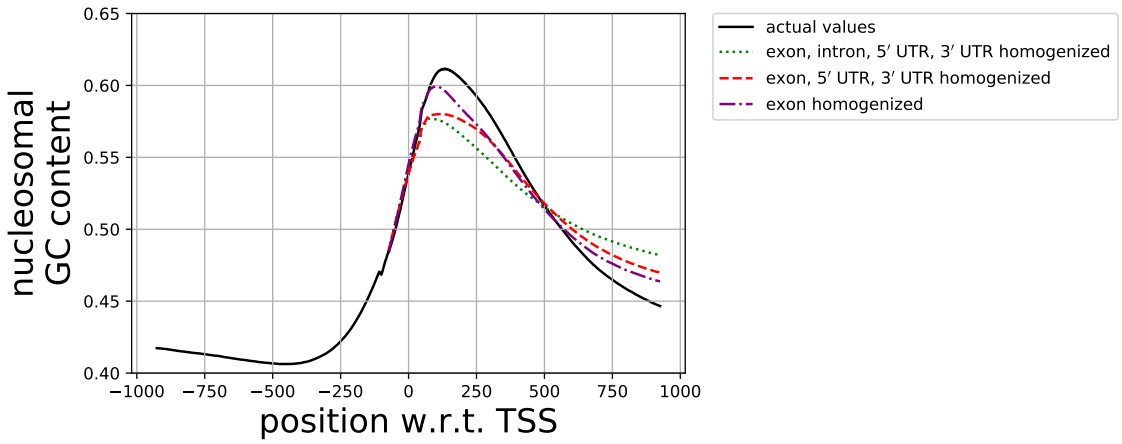


(b)

Figure 5.14: Same as Figs. 5.12 and 5.13 but for *S. tuberosum*, potato. While the overall signal seems similar to *P. persica* and *S. tuberosum*, this signal is almost entirely caused by interregional signals.

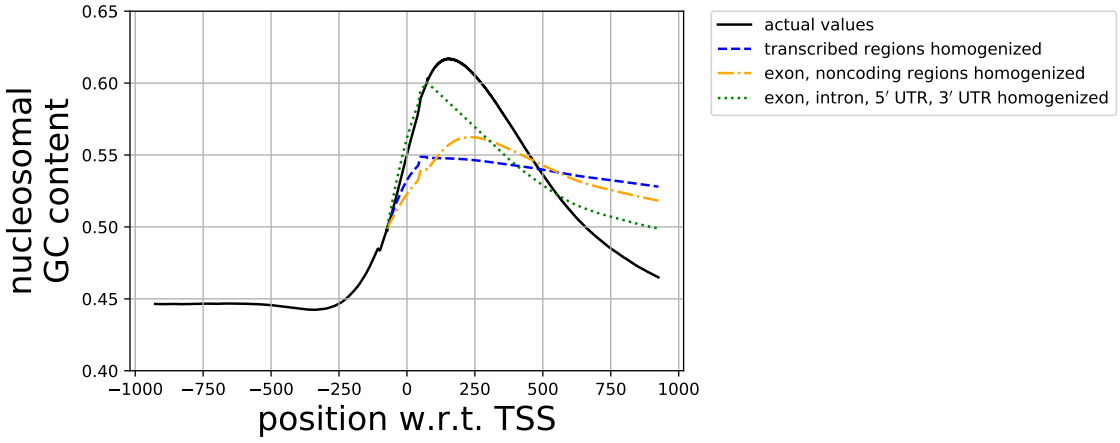


(a)

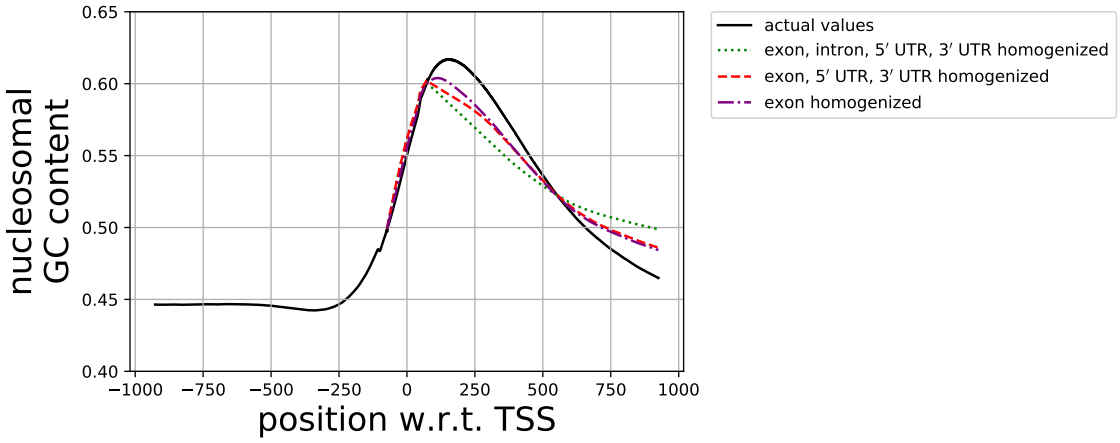


(b)

Figure 5.15: Same as Figs. 5.12 and 5.14 but for *O. sativa*, rice. Its signal is much stronger than the signals for the other plants. The interregional signal is strong and the intraregional signal of introns is not neglegible.

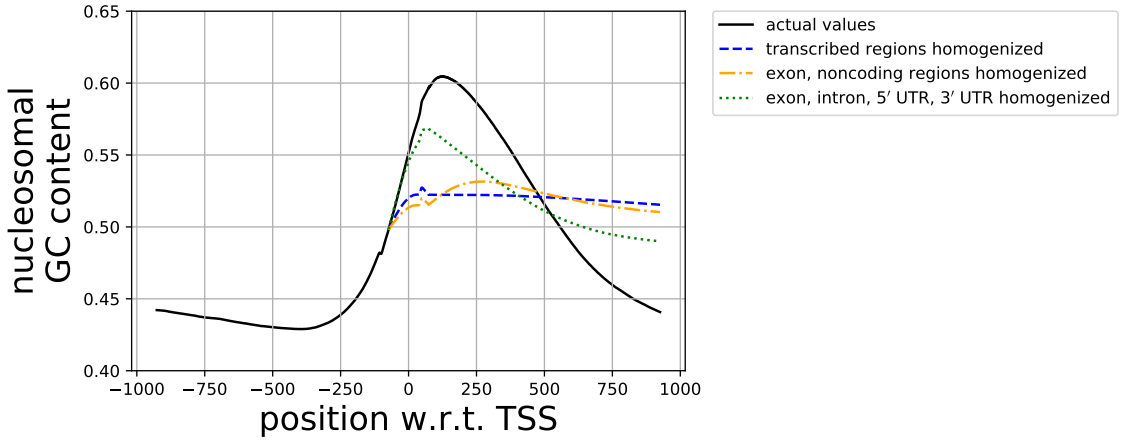


(a)

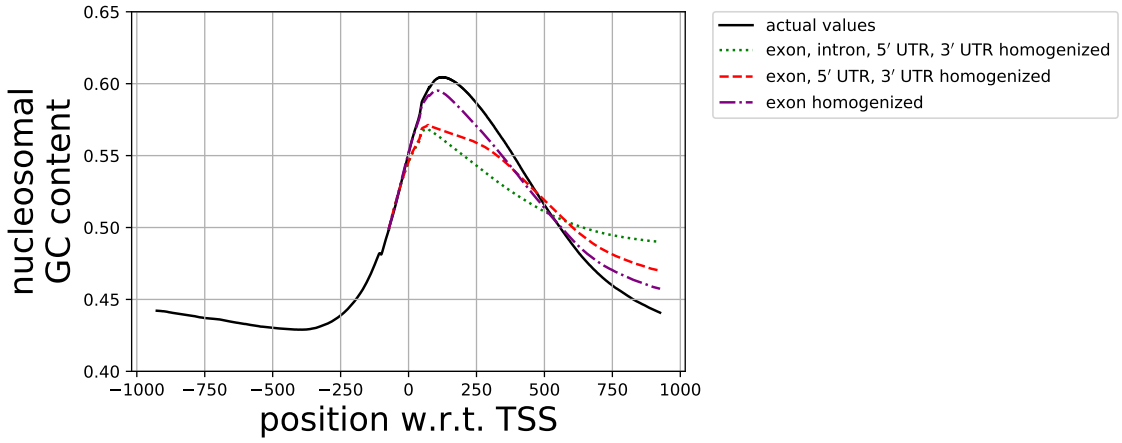


(b)

Figure 5.16: Same as Figs. 5.12 and 5.15 but for *T. aestivum*, wheat. Its signal, like the signal of rice, is much stronger than the signals for the other plants.

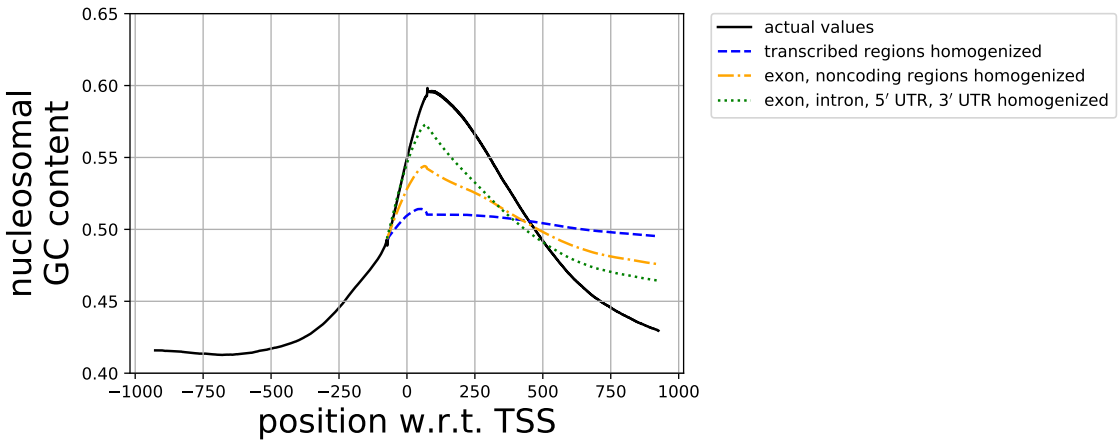


(a)

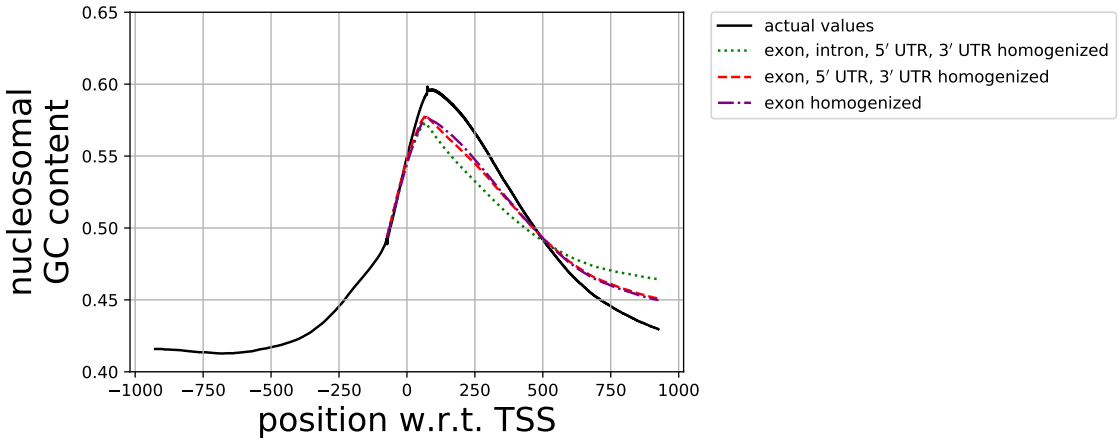


(b)

Figure 5.17: Same as Figs. 5.12 and 5.16 but for *Z. mays*, maize. It has a strong signal like rice and wheat, other grains.

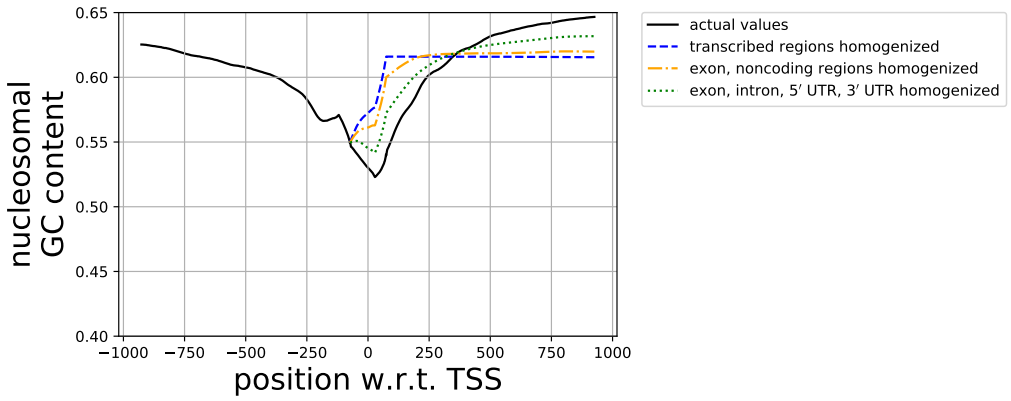


(a)

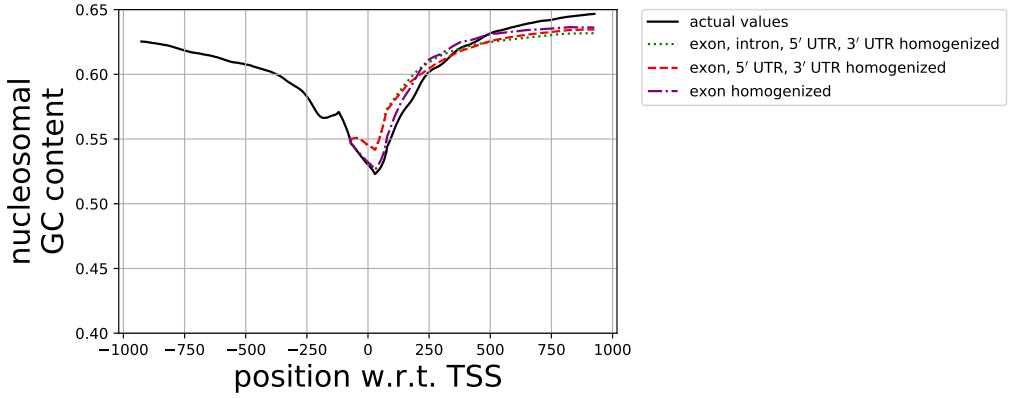


(b)

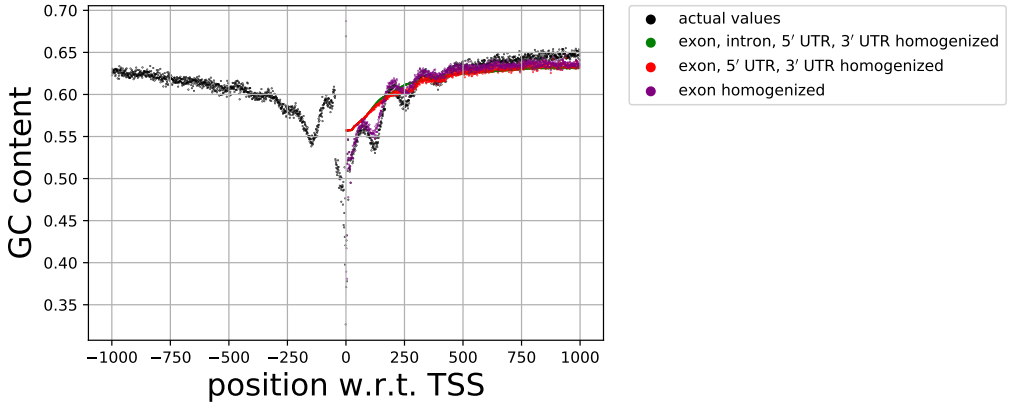
Figure 5.18: Same as Figs. 5.12 and 5.17 but for *L. perrieri*, a cutgrass from Madagascar closely related to rice but not used for consumption. This plant shows only a slightly weaker signal compared to the cultivated grains such as rice, suggesting that cultivation has not lead to significantly stronger nucleosome positioning signals.



(a)



(b)



(c)

Figure 5.19: Figs. (a) and (b) are the same as Fig. 5.12-5.18 but for *C. reinhardtii*, an alga. Fig. (c) is the same as Fig. 5.2(b), depicting the average GC content for each base pair. This algae have a strong GC signal near their TSS which is very different from the signals we have seen for the other plants or for the animals we have seen so far. The signal is periodical, see (c) The periodicical undulations are mostly caused by the intraregional signals of UTRs but are also slightly enhanced by the exons, suggesting that the signal is not some quirk of this organism's UTRs but an evolutionary advantageous signal on the DNA. However, this periodicity may be unrelated to nucleosomes, since it disappears when studying nucleosomal GC content, see (b).

5.6 Even the amino acid sequence contains nucleosome signals

In Chapters 3 and 4 we investigated the range of possible nucleosome energies on exons. We used a hard constraint: the base pair sequence could only be changed without changing the sequence of amino acids. This constraint may not be realistic: some amino acids might be altered during evolution to accommodate the second, mechanical layer. Now we are finally able to put this constraint to the test. We do so by answering the question: does the choice of amino acids affect the exon intraregional signal? And if so, is this related to nucleosomes or the result of some signal on the mRNA?

We investigate the effect of the amino acid sequence by bringing our analysis of multiplexing to a deeper level. We divide the exon intraregional signal in two kinds:

- exon intraregional positioning signal as a result of synonymous codons
- exon intraregional positioning signal as a result of the amino acid sequence

The signal caused by the amino acid sequence is the part of the exon signal that is caused by the *average* GC content of the codons that code for the same amino acid weighted by the occurrence of each of the codons. These weights ensure that the effect of the average GC content of exons is included. The rest of the signal is caused by the specific codons (chosen out of the synonymous codons).

Now we will demonstrate the two different types of intraregional exon signals for *Oryza sativa*, rice, chosen for its large intraregional exon signal. Fig. 5.20 depicts the original nucleosome GC content in the black curve and the dotted purple curve as the curve where all exons are (again) homogenized. New is the green interrupted curve where the GC content of codons is replaced by the average GC content of all synonymous codons, weighted by the occurrence of the codons. This curve is similar in both shape and size to the actual landscape. We find that for rice the choice of amino acids has a large influence on the average GC signal, about 50% of the overall signal. We find that the exon intraregional signal is significantly affected by the choice of amino acids. This means that exons may not be as restricted as thought before. In addition to synonymous codons, the exons may even use codons that code for different amino acids to create nucleosome positioning signals.

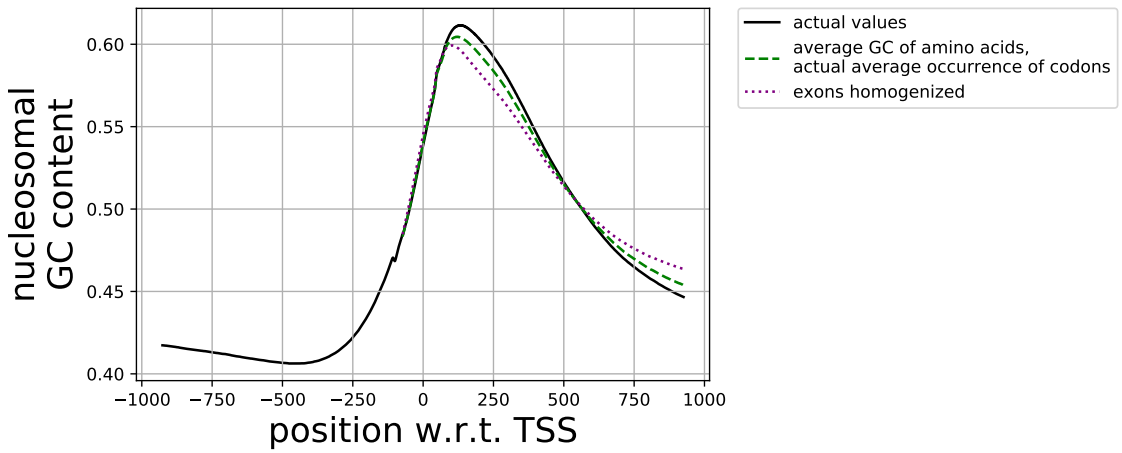


Figure 5.20: This figure describes *Oryza sativa*, rice. We see in black (solid line) a depiction of the actual average nucleosomal GC content around the TSS of protein-coding genes. In green (interrupted line), all codons on all exons have been replaced by the weighted average GC content of the amino acids they encode, weighted by the frequency of the synonymous codons. In purple (dotted line) all exons are homogenized (i.e. the exons have been replaced by the average GC content of exons). The difference between the purple and green line depicts the positioning signal as a result of the amino acids encoded on the DNA. The difference between the green and black line is the effect of the choice of synonymous codons.

5.7 Exon intraregional signals: a function on DNA or mRNA?

We have seen for some organisms, such as human, that the exon intraregional signals are much weaker than their intronic counterparts. For rice however, the exon intraregional signal dominates. This raises an important question: is this exon signal actually related to nucleosomes or is it the result of some function on the mRNA? The function on the mRNA could be related to an important bias in the choice of amino acids related to the final protein product, or to a translation speed signal. This question can be investigated quite elegantly. In Fig. 5.21 we depict in black the average GC content (not the nucleosomal GC content) for *O. sativa* around the TSS. We also depict the average GC content of exons in blue. This curve is quite different from the actual GC landscape since it excludes interregional and noncoding intraregional signals. We can compare this with the green curve, which depicts the average GC content of mRNA after the start codon (ATG). There are now two reasons to believe that the signal we see for the mRNA is a result of the signal on the DNA and not the other way around. One, the mRNA signal is less pronounced than the exon signal on the DNA, suggesting that the mRNA signal is merely a scrambled version on the DNA signal. It is scrambled since the locations of the exons on the mRNA are different from the locations on the DNA because the DNA includes noncoding bases. While it could still be possible that this GC landscape is meaningful on the mRNA, these arguments suggest otherwise. These results for rice stand in great contrast with the GC landscapes for human. Fig. 5.22 depicts the same as discussed before, but for human. Now we see that the mRNA signal is

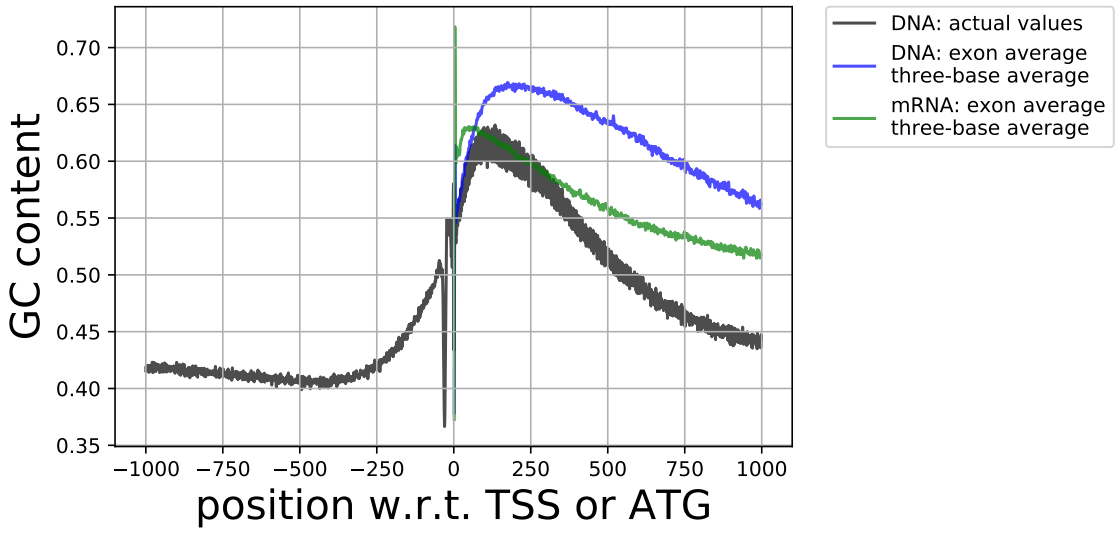


Figure 5.21: In black we show the average GC content per base pair for *O. sativa*, rice, around the TSS. We also depict the average GC content of exons in blue. This curve is quite different from the actual GC landscape since it excludes interregional and noncoding intraregional signals. The green curve depicts the average GC content of mRNA after the start codon (ATG). It is likely that the GC landscape of mRNA is a result of a functional GC landscape on DNA

much stronger than the actual signal on the DNA, and that the average exon signal for human is a (much) weaker version of the mRNA signal. Also, the exon signals do not connect to the downstream GC values. This suggests that the exon signals on human DNA are but a result of the GC landscape of mRNA, and absolutely not the other way around. The shape of the mRNA curve may very well be related to translation speed. It resembles the translation speed ramp that has been suggested to 'reduce ribosomal traffic jams', thus minimizing the cost of protein expression [103]. In the animals *D. melanogaster* and *C. elegans*, a ramp in tRNA-adaptation index (a predictor of translation speed) depicts a ramp of approximately 300 bp [103], which is similar to the 300 bp ramp in the mRNA curve for human. We can easily evaluate whether these ramps are related by calculating the translation speed landscape for human, using the model of Rudorph et al. [29] (see section 4.3). The result is depicted by Fig. 5.23. The translation speed ramp is similar to the ones depicted by Tuller et al. [103]. Possibly, since the introns in human have such a large effect on the nucleosome positioning signal, the exons have more freedom to code for genetics and the translation speed ramp. In rice, where the exons are responsible for a much larger part of the signal, they have less freedom to code for translation speed, resulting in a less-pronounced translation speed signal. Or possibly there is less 'need' for a translation speed signal, resulting in more opportunity to encode mechanical signals. Whatever the reason may be, it seems plausible that translation and mechanical signals need to compete over the course of evolution.

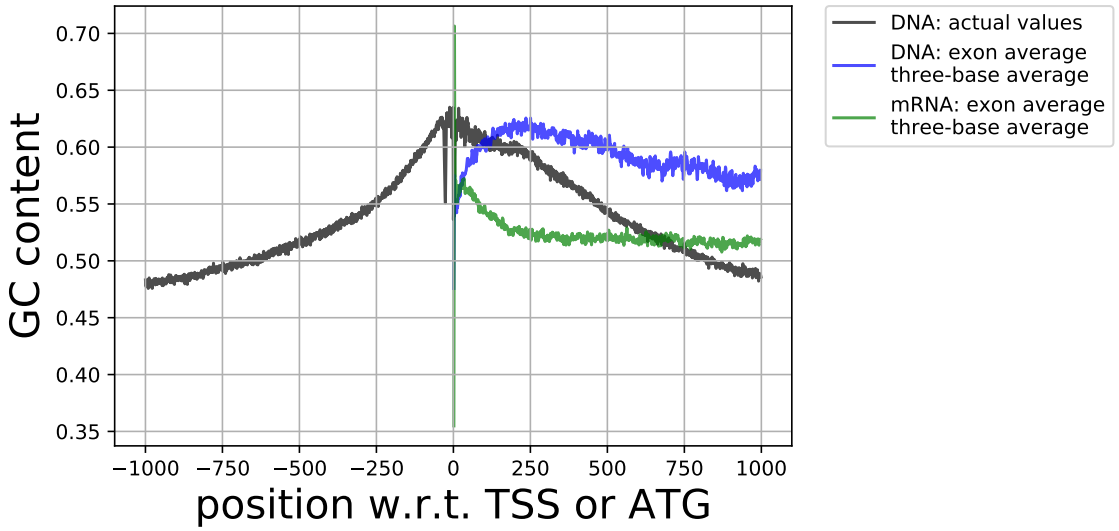


Figure 5.22: Same as Fig. 5.21 but for human. There we find an opposite result compared to rice: the mRNA signal dwarfs the DNA signal. This suggest that the exon signal on the DNA is a result of a functional GC signal on the RNA, likely related to translation speed.

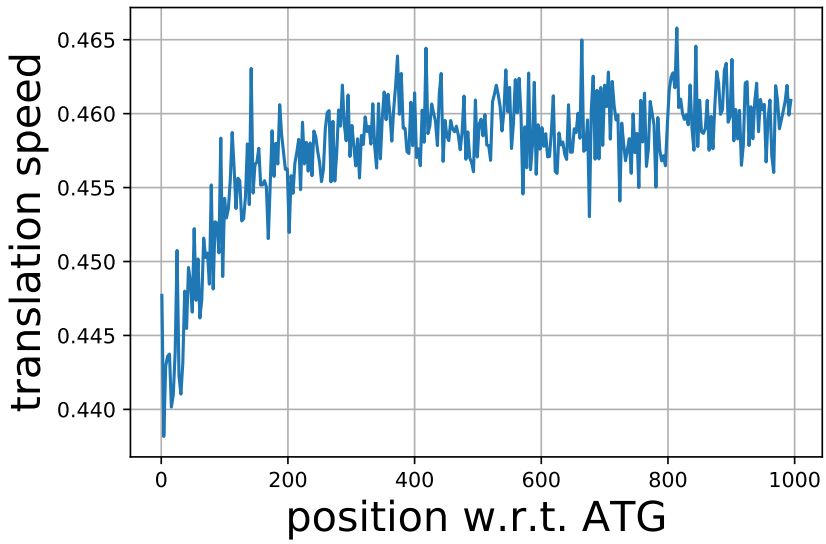


Figure 5.23: The average translation speed landscape for human is shown for a range of positions after (not including) the start codon. The mRNA GC signal in Fig. 5.22 seems related to this landscape, which contains a ramp in the first 300 bp. It resembles the translation speed ramp that has been suggested to reduce ribosomal traffic jams, thus minimizing the cost of protein expression [103].

5.8 Conclusions and Outlook

In this chapter we discussed multiplexing in real genomes. We divided multiplexing into two types, intraregional and interregional multiplexing. We have shown that, for many organisms, such as fish and many plants, interregional signals dominate. For these organisms, the fact that exons, introns and UTRs have different GC levels is enough to explain the overall signal. For other organisms, such as animals and cereal grains, we see significant intraregional signals as well as interregional signals. It is possible that higher-order organisms have evolved to contain intraregional positioning signals in addition to the interregional signals. For human and many other animals, the intron part of the intraregional signal dominates, even after taking the fraction of introns versus exons into account. This may mean that introns have, on average, more freedom to code for mechanical information alongside its other functions as compared to exons (which have to code for the amino acid chain) and UTRs. On the other hand, for rice and other grains we see that, even though the exon part of the intraregional signal dominates, it still has a signal larger than that of human. This is a perfect example of what we see as the most profound type of multiplexing: the combination of protein information and a (relatively) strong mechanical signal on the very same base pair.

Interestingly, we have shown that, for rice, a large part of the signal can be attributed not to the choice of synonymous codons but to the choice of amino acid. By subdividing exon signals into nucleosome positioning signals resulting from the encoded amino acids and positioning signals caused by synonymous codon choice, we have shown that the amino acid sequence has a significant effect on the average GC landscape of rice. It seems that, from an evolutionary point of view, enhancing the GC signal near the TSS was not restricted by a need to keep amino acid sequences intact. This suggests that mechanical and protein-coding information can compete over the course of evolution. To put this result in perspective: in Chapter 3 we have shown that there is much freedom for a mechanical layer of information to exist on top of genes, using the degeneracy of the genetic code. Chapter 4 explains that there is an additional restriction on the mechanical layer in the form of the translation speed landscape. The results from this chapter actually relax the genetic constriction by demonstrating that, in some organisms, not only the degeneracy of the genetic code is utilized. Even specific amino acids seem to be encoded on the DNA to ensure a strong nucleosome positioning signal.

On rice we have found that the strong mechanical signal caused by its exons is unlikely to originate from a functional signal on the mRNA. On the other hand, we find for human, which has a very weak nucleosome positioning signal encoded on its exons, that the mRNA signal dwarfs the mechanical signal of the exons on the DNA. The mRNA signal is possibly related to a translation speed ramp [103] such that proteins can be created efficiently by the ribosomes. Taken together, the results from rice and human suggests a competition between mechanical information and translation speed signals on exons. When we include what we have learned about amino acid nucleosome positioning signals, we suggest that nucleosome positioning signals, translation speed signals and protein-coding information all three may compete with each other. This competition should be investigated further by studying single genes. One should find out whether, for example, human genes without

introns have nucleosome positioning signals encoded on their exons.

Further research should investigate whether the choice of amino acids impacts other organisms in the same manner. How do the translation speed landscape and mechanical information compete in a wider range of organisms? We also suggest the creation of an evolutionary model, using, for example, a Mutation Monte Carlo simulation, which could incorporate replacing amino acids by amino acids with similar function. Another step could be to use the methods presented here to obtain information on the intraregional and interregional multiplexing on single genes instead of the average of genomes. Also, more types of organisms could be investigated. Furthermore, using an energy model, (instead of GC content) such as the trinucleotide model [10] used in the previous chapters, could include rotational positioning signals to our analysis, in addition to translational positioning signals.

Appendix A

The physics behind the mechanical nucleosome positioning code

A.1 Energy contributions of twist and cross terms

In section 2.2 we define the energy of a dinucleotide as the sum of the energy of roll and tilt. Keeping in mind the observation that the basic nucleosome positioning rules can be rationalized by discussing energy costs involved in the roll and tilt degrees of freedom [14], as well as our goal to reduce our model to its bare essentials, we chose to neglect the contribution of twist and of the cross terms between roll, tilt, and twist. In this appendix we will show that including the twist and cross terms does not change the qualitative agreement of our model with well-known positioning rules (see Fig. 2.3).

We start by defining the energy of twist and the cross-terms:

$$E^{\text{twist}}(a,b) \equiv \frac{1}{2} Q^{\text{twist}}(a,b) \left[q^{\text{twist}} - \bar{q}^{\text{twist}}(a,b) \right]^2 , \tag{A.1}$$

and

$$E_p^{\text{cross}}(a,b) \equiv \sum_{\substack{i,j \in \{\text{roll,tilt,twist}\} \\ i \neq j}} \frac{1}{2} Q^{i,j}(a,b) \left[q_p^i - \bar{q}^i(a,b)_p \right] \left[q_p^j - \bar{q}^j(a,b)_p \right] . \tag{A.2}$$

The bp-step dependent stiffnesses are now given by $Q^i(a,b)$, $i \in \{\text{roll}, \text{tilt}, \text{twist}\}$ and the corresponding intrinsic values by $\bar{q}^i(a,b)$, $i \in \{\text{roll}, \text{tilt}, \text{twist}\}$. The cross terms depend on the cross stiffnesses $Q^{i,j}(a,b)$, $i,j \in \{\text{roll}, \text{tilt}, \text{twist}\}$, $i \neq j$. (Note that, because of the constant twist, the energy associated with twist does not depend on position p but only on the dinucleotide step.) For the twist and cross terms, too, the hybrid parametrization [38] is used. We can redefine our energy as

$$E_p(a,b) = E_p^{\text{roll}}(a,b) + E_p^{\text{tilt}}(a,b) + E^{\text{twist}}(a,b) + E_p^{\text{cross}}(a,b). \tag{A.3}$$

Fig. A.1 was created using this redefined energy. We see that the relative behaviour of the dinucleotide probabilities at different positions is the same as without the cross terms, see Fig. 2.3.

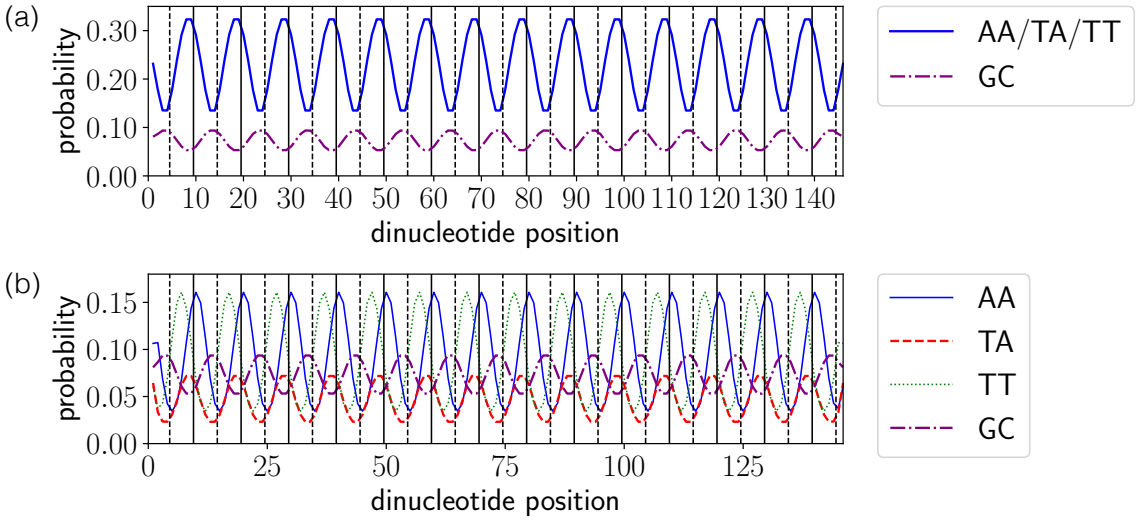


Figure A.1: Same as Fig. 2.3 but with including the cross-terms (see Eq. A.3). The positioning rules (i.e. the relative behaviour of the dinucleotide probabilities at different position) has stayed the same.

A.2 Validity of the average neighbour energy approximation

The average neighbour energy approximation of the probability works extremely well. We checked it for all dinucleotides and found that the largest error of this approximation occurs for the probability distribution of dinucleotide AA. Fig. A.2 depicts both the exact probability and its approximation for this dinucleotide. The difference between the values is always smaller than 3.5%.

To understand why this error is so small, one needs to consider the function $C_p(x, y)$, defined in Eq. 2.28. The average neighbour energy approximation is exact if this function is a constant (i.e., independent of x and y for each p). The approximation works well if the function is almost constant. That this is true is best seen by inspecting the standard deviation of $C_p(x, y)$, divided by its mean, and checking whether this quantity is much smaller than one. Here the standard deviation and mean are defined as:

$$\mathrm{std}[C_p] \equiv \sqrt{\langle \{C_p(x,y) - \mathrm{mean}[C_p]\}^2 \rangle_{x,y}} \tag{A.4}$$

with

$$\mathrm{mean}[C_p] \equiv \langle C_p(x,y) \rangle_{x,y}. \tag{A.5}$$

Fig. A.3 shows that this ratio is indeed much smaller than one for all dinucleotide positions.

A.3 Effect of temperature on the probability

The probabilities shown in the results section have all been obtained at room temperature $\beta = 1/k_B T_r$. Here we study how these probabilities change with temperature,

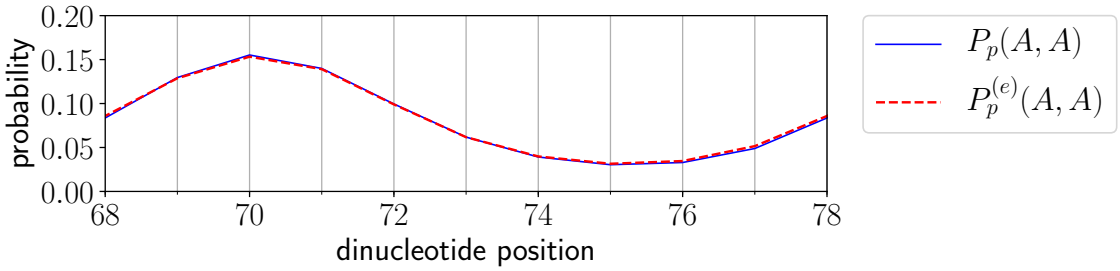


Figure A.2: The exact probability and its average neighbour energy approximation to find AA steps at all dinucleotide positions. The approximation introduces an error that is nowhere larger than 3.5%.

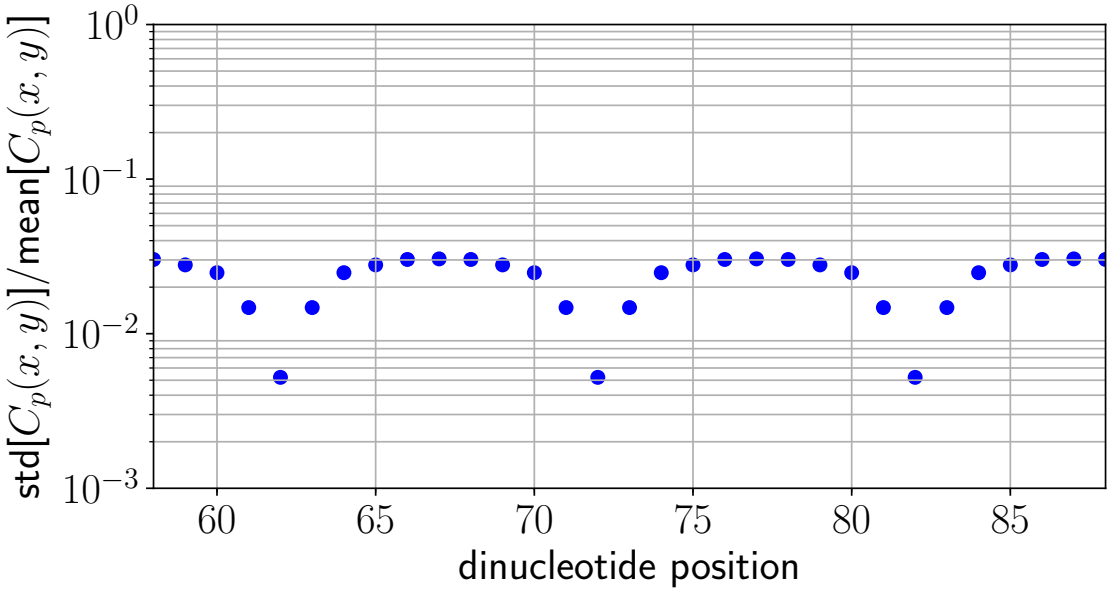


Figure A.3: The standard deviation (Eq. A.4) divided by the mean (Eq. A.5) of $C_p(x,y)$. As this ratio is very small at all positions p the function $C_p(x,y)$ is nearly constant, explaining the high accuracy of the average neighbour energy approximation.

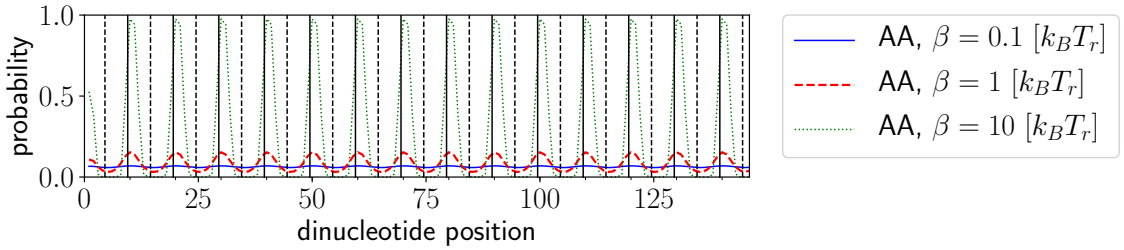


Figure A.4: The probability to obtain AA at all dinucleotide positions at several temperatures.

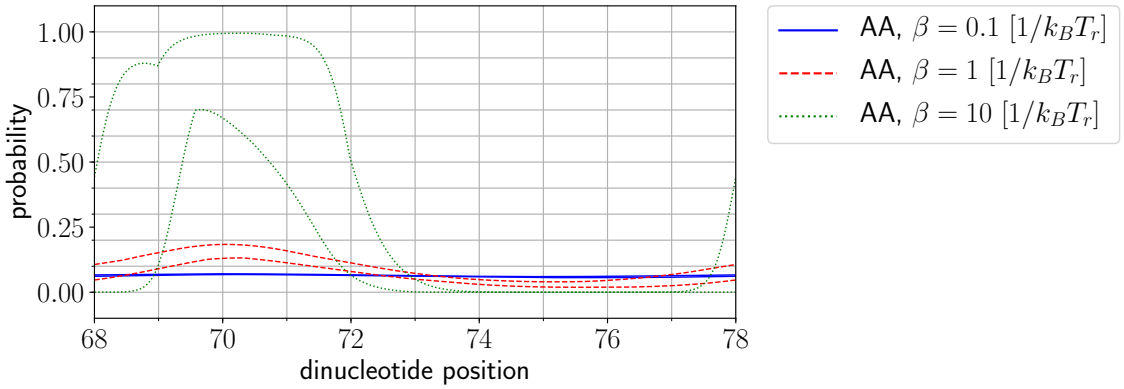


Figure A.5: The first-order bounds of the probability of encountering an AA step are shown at five different temperatures. At low temperatures, the bounds become significantly far apart from each other and only provide a qualitative description of the behaviour of the probability as a function of position.

focusing on dinucleotide AA. Its exact probability distribution for different temperatures is shown in Fig. A.4. We find at temperature $\beta = 0$ a constant value $1/16$ for the probability. This is the high temperature limit where all steps are equally probable. At low temperatures the probability varies between values close to 0 and 1, reflecting the fact that the ground state sequences becomes exceedingly important.

We also evaluated the first- and second-order bounds of the AA probability distribution at five different temperatures: $\beta = 0$, 0.1, 1, 10, and 100 (in units of $[1/k_B T_r]$), see Fig. A.5, and Fig. A.6. At high temperatures (low β) the bounds for both orders are very close to each other enclosing values close to $1/16$. With decreasing temperature the quality of the first-order bounds becomes poorer, giving only a rough qualitative estimate whereas the second-order bounds continue to work well for relatively low temperatures. Note that at $\beta = 100$ the probability takes values close to 0 and 1 at most places.

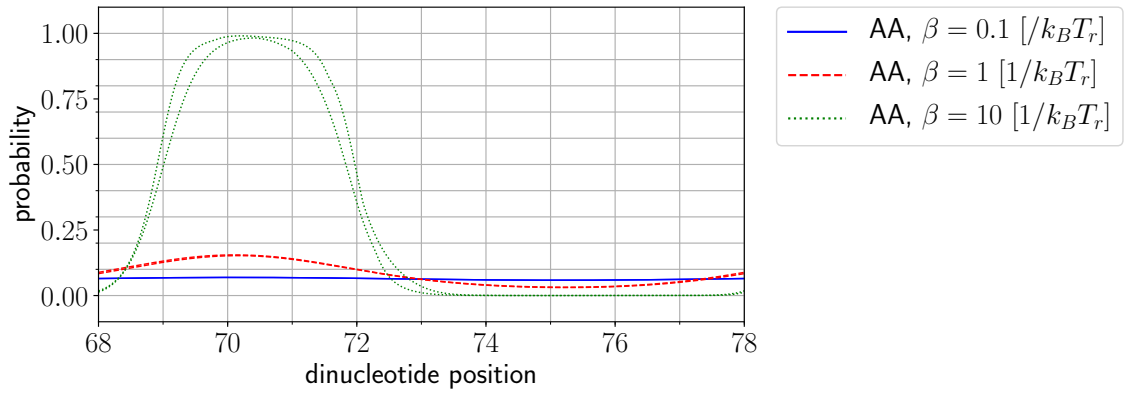


Figure A.6: The second-order bounds of the probability of encountering an AA step are shown at five different temperatures. At all temperatures the method provides a quantitative description of the probability, clearly outperforming the second-order bounds.

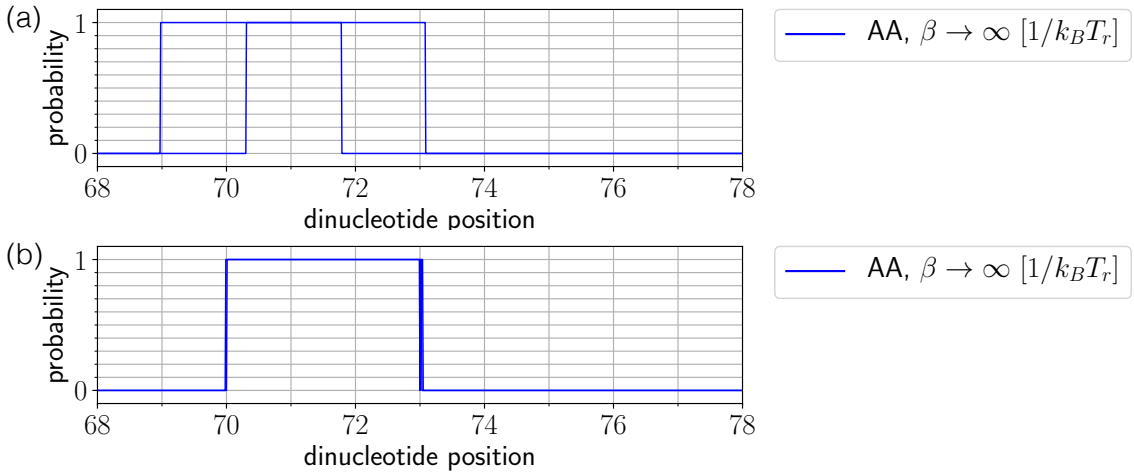


Figure A.7: (a) First-order and (b) second-order bounds on the probability to find dinucleotide AA at several dinucleotide positions on a nucleosome, in the limit of zero temperature. Higher-order bounds get increasingly sharper. At zero temperature the only possible DNA sequences are ground state sequences, hence the bounds provide us statistics on the ground states of our system. When the lower and upper bounds are 0 and 1, AA is part of an unknown number of ground states at this position. If AA is part of all possible ground states the bounds are 1 and 1, and if it is not a part of the ground state they are 0 and 0.

We finally take the limit $\beta \to \infty$, see Fig. A.7. This figure shows the second-order and third-order bounds on the probability to encounter AA at zero temperature. The only possible sequences are now ground state sequences (due to the high level of symmetry in our model we expect many different ground states). This explains why the probability of AA can take the values 0 and 1: at several positions AA is not part of any ground state sequence (probability is zero), while at other positions AA is part of all possible ground state sequences (probability is 1). At some positions the method cannot determine the percentage of ground state sequences AA is part of, resulting in bounds of 0 *and* 1.

The method of obtaining upper and lower bounds remains effective at all possible temperatures for our model, and even provides insight into the possible ground states. Going to higher-order bounds (i.e., taking neighbours that are further away into account as well) or using the exact probability should eliminate the discrepancy between the upper and lower value. However, the method employed in Chapter 3 (which uses a graph representation of all possible sequences in combination with a shortest path algorithm) is much more efficient in obtaining ground states.

Appendix B

Shortest paths through synonymous codons

B.1 Definition of the energy

In Chapter 3, we aim to find sequences with 'special' energies, e.g. the sequences with the lowest and highest possible energies. To calculate the energy of a sequence, we use the probabilistic trinucleotide model by Tompitak et al. [10] which is based on the sequence preferences of a coarse grained nucleosome model, parametrized by experimental parameters derived from protein-DNA crystals [32]. Because it is a trinucleotide model, we are able to represent the total energy of a sequence as a sum of 'conditional' trinucleotide energies, which function as the (main ingredients of the) weights in our graphs. Here we will formally define these energies.

Let $\mathcal{B}$ be the set of all nucleotides, $\mathcal{B} = \{A, T, C, G\}$. For the trinucleotide model, it is assumed that the probability of a nucleotide depends only on the previous two. Defining S as a sequence of length L, consisting of nucleotides $S_i \in \mathcal{B}$ with i from 1 to 147, this gives a probability for the full sequence:

$$P(S) = \frac{\prod_{n=1}^{L-2} P_n(S_{n+2} \cap S_{n+1} \cap S_n)}{\prod_{n=1}^{L-3} P_n(S_{n+2} \cap S_{n+1})} \tag{B.1}$$

where $P_n(S_{n+2} \cap S_{n+1} \cap S_n)$ is the joint (trinucleotide) probability to obtain S_{n+2}, S_{n+1}, and S_n at position n, and $P(S_{n+2} \cap S_{n+1})$ the joint (dinucleotide) probability to obtain S_{n+2}, S_{n+1} at position n. However, since the original trinucleotide model by Tompitak et al. does not enforce the symmetry of the coding and noncoding strand, we introduce symmetrized probabilities:

$$\begin{aligned} P'_n(S_n \cap S_{n-1} \cap S_{n-2}) = &\frac{1}{2}\left[P_n(S_n \cap S_{n-1} \cap S_{n-2})\right] \\ &+ \frac{1}{2}\left[P_n(S'_{n-2} \cap S'_{n-1} \cap S'_n)\right] \end{aligned} \tag{B.2}$$

and

$$P'_n(S_n \cap S_{n-1}) = \frac{1}{2}\left[P_n(S_n \cap S_{n-1}) + P_n(S'_{n-1} \cap S'_n)\right] \tag{B.3}$$

where

$$S'_n \equiv \begin{cases} A_{148-n} \text{ if } S_n = T \\ T_{148-n} \text{ if } S_n = A \\ C_{148-n} \text{ if } S_n = G \\ G_{148-n} \text{ if } S_n = C \end{cases} \tag{B.4}$$

such that

$$P'_n(S) = \frac{\prod_{n=1}^{L-2} P'_n(S_{n+2} \cap S_{n+1} \cap S_n)}{\prod_{n=1}^{L-3} P'_n(S_{n+2} \cap S_{n+1})}. \tag{B.5}$$

Following Tompitak et al., we use the probability to calculate a free energy, using $E(S) = -k_B T_r \ln[P(S)] + const.$ We rewrite the energy as:

$$E(S) = \sum_{n=1}^{L-2} E_n(S_{n+2}, S_{n+1}, S_n) + const. \tag{B.6}$$

where

$$E_n(S_n, S_{n+1}, S_{n+2}) = \begin{cases} -k_B T_r \ln[P'(S_{n+2} \cap S_{n+1} \cap S_n)] \text{ if } n = 1 \\ -k_B T_r \ln\left[\frac{P'(S_{n+2} \cap S_{n+1} \cap S_n)}{P'(S_{n+1} \cap S_n)}\right] \text{ if } 1 < n < 146 \\ 0 \text{ else.} \end{cases} \tag{B.7}$$

We define *const.* such that the energy E is zero if S is the ground state.

For $n = 1$, E_n is the energy cost related to the first three bases of a sequence S, for $1 < n < 146$, it is a 'conditional' energy, and it is zero elsewhere. We use these terms as weights of our graph, while keeping in mind that the sum of these weights will provide the well-defined total energy E.

B.2 Definition of the depth of a minimum

In the main text of Chapter 3, we use the depth of a minimum $\mathcal{D}$ as a measure for how well the nucleosome is positioned at this minimum. Here we will formally define $\mathcal{D}$.

Let $\mathcal{S}$ be some sequence of length greater than $L + 10$ (with $L = 147$). Let S^p be a subsequence of $\mathcal{S}$ of length L starting at position p.

We call a nucleosome positioned at p if the energy $E(S^p)$ is lower than the energies at positions $p - 5, p - 4, ..., p + 5$ (excluding p). We denote the energy corresponding to a nucleosome containing the sequence S^p by $\mathscr{E}_p \equiv E(S^p)$. For a minimum at $p_{\min}$ of sequence S we are interested in its depth, $\mathcal{D}(S^{p_{\min}})$. Now we can formally define the depth as

$$\mathcal{D}(S^{p_{\min}}) \equiv \min\left[\mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}), \mathscr{E}^{\max}_{\text{right}}(S^{p_{\min}})\right] \tag{B.8}$$

where

$$\mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}) \equiv \max\left[\mathscr{E}_{p_{\min}-i}(S) \text{ for } i \in \{1, 2, ..., 5\}\right] - \mathscr{E}_{p_{\min}}(S), \tag{B.9}$$

$$\mathscr{E}^{\max}_{\text{right}}(S^{p_{\min}}) \equiv \max\left[\mathscr{E}_{p_{\min}+i}(S) \text{ for } i \in \{1, 2, .., 5\}\right] - \mathscr{E}_{p_{\min}}(S). \tag{B.10}$$

B.3 The deepest possible minimum

Here we show how to obtain the deepest possible minimum, with only a tiny possible error, by taking the shortest paths through the graphs $\mathcal{G}^+_{h,j}$ defined in the main text.

A nucleosome is best positioned at a minimum $p_{\min}$ if $\mathcal{D}(S^{p_{\min}})$ is maximal. We assume that the deepest possible minimum $\mathcal{D}(S_{\text{deepest}})$ is found for a sequence S_{deepest}. Furthermore, we assume that

$$\mathscr{E}^{\max}_{\text{left}}(S_{\text{deepest}}) = \mathscr{E}_{p_{\min}+h}(S_{\text{deepest}}) - \mathscr{E}_{p_{\min}}(S_{\text{deepest}}) \tag{B.11}$$

and

$$\mathscr{E}^{\max}_{\text{right}}(S_{\text{deepest}}) = \mathscr{E}_{p_{\min}+j}(S_{\text{deepest}}) - \mathscr{E}_{p_{\min}}(S_{\text{deepest}}) \tag{B.12}$$

for $h \in \{-5,-4,..,-1\}$, $j \in \{1,2...,5\}$.

Let us denote the shortest path through $\mathcal{G}^+_{h,j}$ by $S_{h,j}$ with the minimum at $p_{\min}$. A shortest path through $\mathcal{G}^+_{h,j}$ will minimize the quantity $2\mathscr{E}_{p_{\min}} - \mathscr{E}_{p_{\min}+h} - \mathscr{E}_{p_{\min}+j}$. Because of this, we have

$$\mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}_{h,j}) + \mathscr{E}^{\max}_{\text{right}}(S^{p_{\min}}_{h,j}) \geq \mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}_{\text{deepest}}) + \mathscr{E}^{\max}_{\text{right}}(S^{p_{\min}}_{\text{deepest}}). \tag{B.13}$$

Since $S^{p_{\min}}_{\text{deepest}}$ is the sequence with the greatest depth, we have

$$\begin{aligned} &\min\left[\mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}_{h,j}), \mathscr{E}^{\max}_{\text{right}}(S^{p_{\min}}_{h,j})\right] \leq \\ &\quad \min\left[\mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}_{\text{deepest}}), \mathscr{E}^{\max}_{\text{right}}(S^{p_{\min}}_{\text{deepest}})\right]. \end{aligned} \tag{B.14}$$

Combining Eq. B.13 and B.14 leads to bounds on the depth of the deepest possible minimum:

$$\begin{aligned} &\min\left[\mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}_{h,j}), \mathscr{E}^{\max}_{\text{right}}(S^{p_{\min}}_{h,j})\right] \leq \\ &\quad \mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}_{\text{deepest}}) && \text{(B.15)} \\ &\quad \leq \frac{1}{2}\left[\mathscr{E}^{\max}_{\text{left}}(S^{p_{\min}}_{h,j}) + \mathscr{E}^{\max}_{\text{right}}(S^{p_{\min}}_{h,j})\right]. && \text{(B.16)} \end{aligned}$$

We took the shortest path through all graphs $\mathcal{G}^+_{h,j}$ for all $h \in \{-5,-4,..,-1\}$, $j \in \{1,2...,5\}$. Of all the graphs, $\mathcal{G}^+_{-5,5}$ provided the deepest minimum. Using the above equation, we obtained $83.47 \pm 0.03\ k_B T_r$ as the deepest possible minimum.

B.4 Graphs

We have defined the graphs $\mathcal{G}^+_{h,j}$, extensions of $\mathcal{G}$ with differently assigned weights, for $h,j \in \{-5,-4,..,-1,1,2...,5\}$. A visual depiction is shown by Fig. B.1. The graph $\mathcal{G}^+_{\text{gene}}$, an extended version of $\mathcal{G}_{\text{gene}}$, is depicted by Fig. B.2.

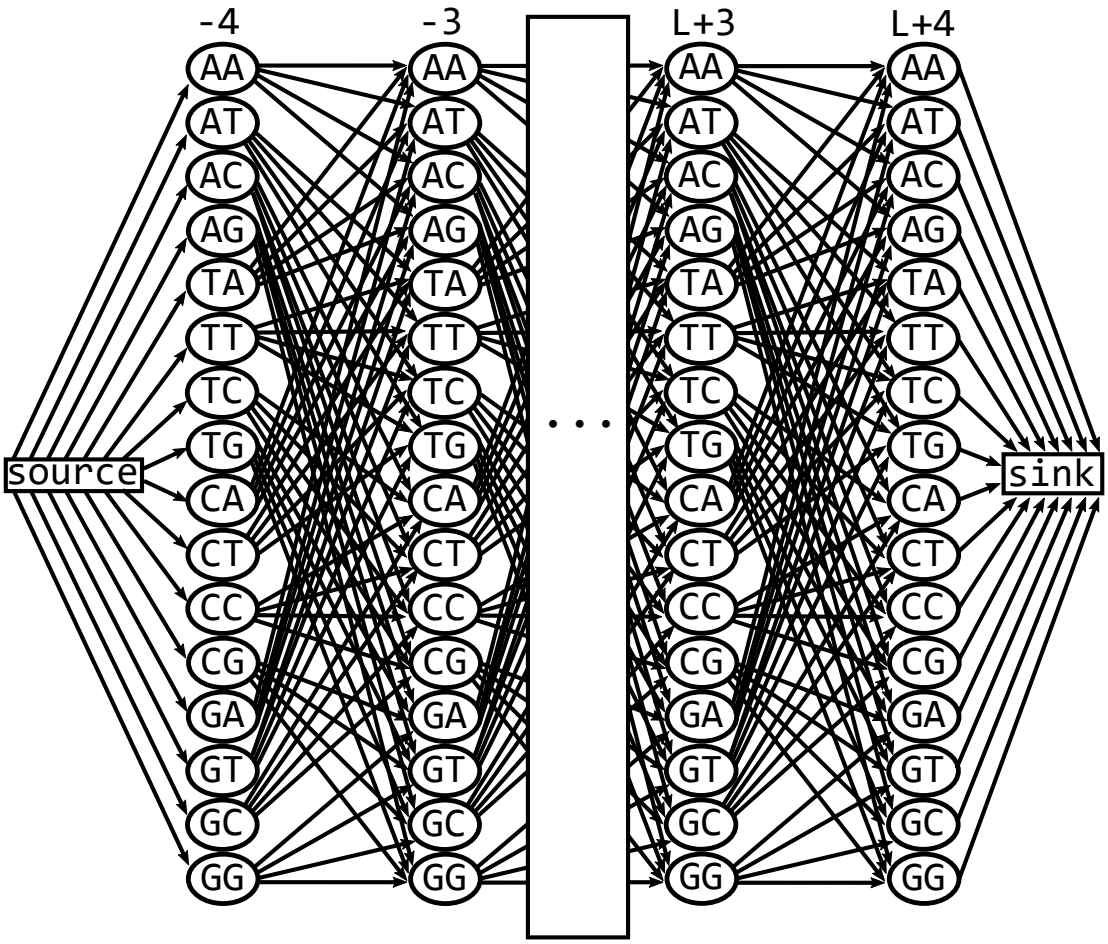


Figure B.1: Visualisation of a graph $\mathcal{G}^{+}_{h,j}$

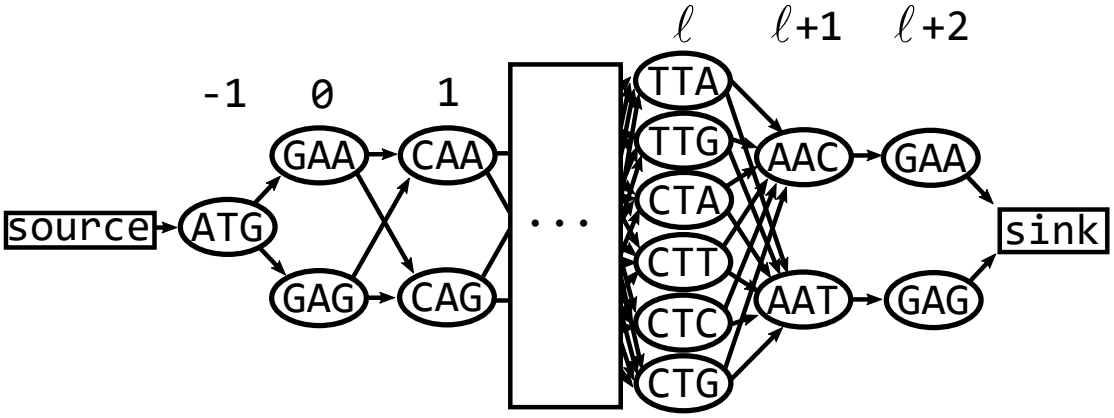


Figure B.2: Visualisation of a graph $\mathcal{G}_{\text{gene}}$. This graph corresponds to creating a minimum at the 7^{th} nucleosome position on the gene YAL002W of yeast.

B.5 Create local minima on top of genes

To create local minima at a position on a gene, we came up with a specifically tailored method where we alter the values of the constants c_i with each iteration.

iteration	**starting conditions**	**action**
all iterations	$c_0 = 1$ $c_i = 0$ for $i \notin \{-5, 0, 5\}$	minimum and depth check: $\mathcal{D} \geq 10\ k_B T_r$
1-20	$c_{-5} = c_5 = -0.3$	regular decrement
21-40	$c_{-5} = c_5 = -0.3$	neighbor decrement
41-60	$c_{-5} = c_5 = -0.2$	regular decrement
61-80	$c_{-5} = c_5 = -0.2$	neighbor decrement
81-100	$c_{-5} = c_5 = -0.1$	regular decrement
101-120	$c_{-5} = c_5 = -0.1$	neighbor decrement
121-140	$c_{-5} = c_5 = 0$	regular decrement
141-160	$c_{-5} = c_5 = 0$	regular decrement
if all fail	-	take best solution

Table B.2: Schematic form of specifically tailored method to create deep local minima at a position on a gene. The method works by altering the weights w_i' of graph $\mathcal{G}^{\text{gene}}$ by changing the constants c_i, see Eq. 4 of the main text.

This will result in a differently weighted graph each iteration and different shortest paths. The algorithm uses at most 160 iterations per position. The iterations are grouped in eight parts, with differing starting conditions and different increment rules. See Table B.2 for an overview of this method.

All iterations start with $c_0 = 1$, $c_i = 0$ for $i \notin \{-5, 0, 5\}$. Iterations 1-20 start with $c_{-5} = c_5 = -0.3$. At the start of iteration 21-40, all constants are reset and we again begin with $c_{-5} = c_5 = -0.3$. Iterations 41-60 and 61-80 have $c_{-5} = c_5 = -0.2$, 81-100 and 100-120 have $c_{-5} = c_5 = -0.1$, and 121-140 and 141-160 have $c_{-5} = c_5 = 0$. The different starting conditions are intended to first try to create deep minima through a larger incentive to have high walls, but if this fails, settle for lower minima.

At the beginning of each and every iteration a check is performed. The energy landscape corresponding to the shortest path is evaluated to find whether a local minimum has been created at the right position. If there is such a local minimum, we evaluate how deep it is. If it is deeper than $10\ k_B T_r$, we accept the corresponding sequence. If the local minimum is not deep enough, we evaluate which side of the energy well has the lowest wall. If the left or right wall is lowest, we set $c_{-5} \to c_{-5} - 0.1$ or $c_5 \to c_5 - 0.1$, respectively, and move to the next iteration. If there is no local minimum, we perform one of the two distinct schemes: 'regular decrement' and 'neighbor decrement', introduced below. We perform a 'regular decrement' at iterations 1-20, 41-60, etc., and a 'neighbor decrement' at all other iterations.

The regular decrement is defined as follows: if the position with the lowest energy is $p_{\min} + i$ instead of the intended position $p_{\min}$, we perform $c_i \to c_i - 0.1$. Differently stated, we give our algorithm an incentive to raise the energy at positions where the energy is lower than at $p_{\min}$. The main problem of the regular decrement is that the lowest energy position often alternates between $p_{\min} + 1$ and $p_{\min} - 1$. Making the decrements smaller turned out to be ineffective in solving this problem, so instead we define the 'neighbour decrement'.

The neighbour decrement is the same as the regular decrement, with one dif-

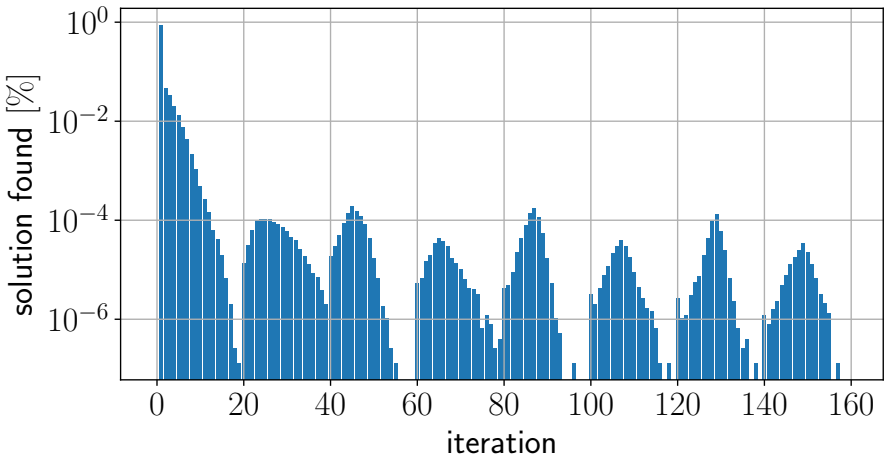


Figure B.3: For each possible iteration, the percentage of positions solved (i.e. with a deep enough minimum found) is depicted. All positions on genes from yeast *S. cerevisiae* (ignoring genes with introns) where evaluated. The bulk of the positions were completed at the first iteration.

ference: if the position with the lowest energy is $p_{\min} \pm 1$ instead of the intended position $p_{\min}$, we perform $c_{i\pm2} \to c_{i\pm2} - 0.1$.

It is possible that, after 160 iterations, no deep enough minimum is found. Then we take the deepest minimum we encountered (if any exists) as our result. The percentage of positions resolved at which iteration is depicted by Fig. B.3. It shows that the bulk of the positions were completed at the first iteration.

Appendix C

Multiplexing mechanical and translational cues on genes

C.1 Graph to obtain highest and lowest possible nucleosome energy

In Chapter 4 we use a graph representation of all possible sequences that code for the same protein. To understand the new method we use in this chapter, we first shortly summarize the method we used in Chapter 3[1], where we were able to obtain the highest and lowest possible nucleosome energies on all positions of a gene. To obtain these energies we use a graph containing all synonymous codons of the gene section corresponding to one nucleosome position.

The DNA on a nucleosome consists of 147 base pairs, which corresponds to either 49 or 50 codons. Suppose we have a sequence of 50 codons. These codons encode a sequence of amino acids $p_0, p_1, p_2, ..., p_{49}$. The number of different codons coding for the same amino acid is 6 at most. Therefore, the most general representation of all possible ways to code for the same protein at one nucleosome position is given by figure C.1 (we use the most general representation to make it easier to understand the graphs related to three layers of information).

In this figure, under each amino acid p_n, six numbers are shown representing the (at most) six possible codons, which we will refer to in the following as $p_n(1), p_n(2), ..., p_n(6)$. The actual base pairs of the codons depend on the amino acid in question. To obtain this graph we draw the following weighted edges: from start to $p_0(i)$ with weight zero for any i, from $p_{49}(i)$ to end with weight $w_{\text{end}}(p_{49}(i))$ for any i, and from $p_n(i)$ to $p_{n+1}(j)$ with weight $w_n(p_n(i), p_{n+1}(j))$ for any i, j and $n = 0, 1, ..., 48$. The weight w_i is given by

$$w_i(C, D) = E_{3i-2}(C_1, C_2, C_3) + E_{3i-1}(C_2, C_3, D_1) + E_{3i}(C_3, D_1, D_2) \tag{C.1}$$

[1]There are two main advantages to summarizing this method again, as opposed to simply referring to the previous chapter/appendix. It makes Chapter 4, in combination with its appendix, readable (and hopefully comprehensable) as a single unit. Secondly: the notation we use here is quite different, such that we can more easily incorporate translation speed in the graph (see appendix C.2).

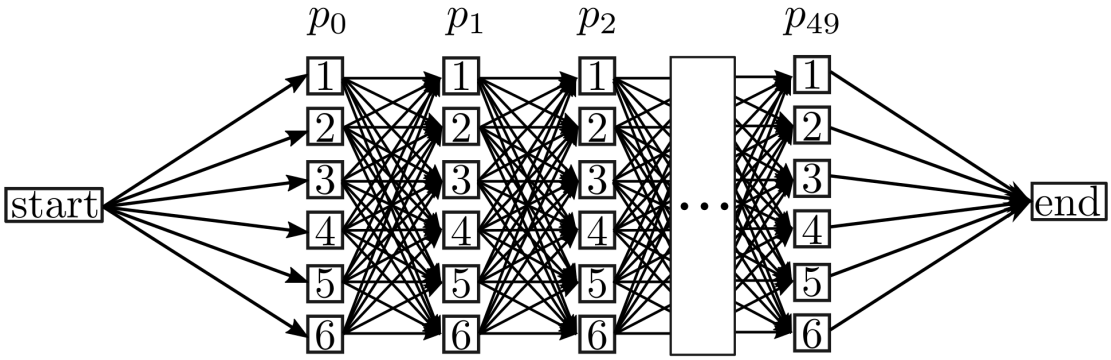


Figure C.1: Graph $\mathcal{G}_{\mathrm{E}}$ shows all synonymous ways to encode a given amino acid sequence $p_0, p_1, ..., p_{49}$ in the most general case. For each amino acid six options are shown, representing the at most six possible ways to code for the same amino acid. The actual bases depend on the amino acid in question. When there are less than six options, one can simply leave out the surplus of nodes. Weights are assigned such that each path from *start* to *end* has a length equal to the total energy of the corresponding codon sequence.

and the weight w_{end} by

$$w_{\mathrm{end}}(D) = E_{145}(D_1, D_2, D_3) \tag{C.2}$$

where C_k and D_k denote the kth base of codons C and D. Now the length of a path from *start* to *end* in the graph equals the energy of a corresponding sequence. The lowest and highest energy can be found using a shortest path algorithm.

C.2 Obtaining the highest and lowest possible nucleosome energy, incorporating translation speed

Here we describe the method used to obtain the highest and lowest possible nucleosome energy, incorporating a restriction on the translation speed. The method uses a graph $\mathcal{G}_{\mathrm{T\&E}}$, which is similar to graph $\mathcal{G}_{\mathrm{E}}$. Since we study in chapter 4 five-codon averages of the translation speed, $\mathcal{G}_{\mathrm{T\&E}}$ incorporates translation speed by using nodes consisting of five codons, see figure C.2. These nodes are connected such that any node

$$p_n(x_n)p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})$$

can only be connected to nodes

$$p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})p_{n+5}(x_{n+5})$$

for any $x_i \in \{1, 2, ..., 6\}$, with weight $w_{n+4}(p_{n+4}(x_{n+4}), p_{n+5}(x_{n+5}))$. All other edges have zero weight. Now, to ensure that one does not alter the translation speed landscape too much when changing the nucleosome energy, one can, for each node, calculate the difference between the translation speed of that node and the original speed. When the difference exceeds a certain threshold, the node needs to be pruned, such that each path through the graph corresponds to a sequence that does not change the underlying amino acid sequence and the translation speed landscape remains the same up to the threshold. Again, the lowest and highest energy can be found using a shortest path algorithm.

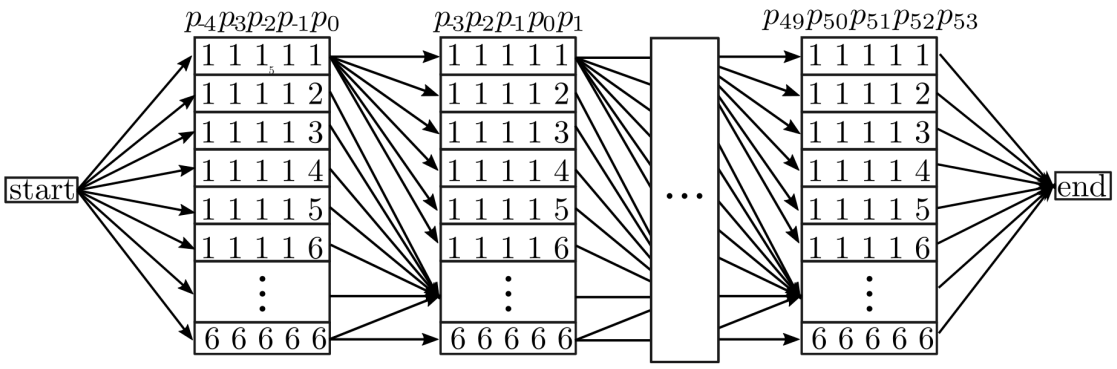


Figure C.2: Graph $\mathcal{G}_{\mathrm{T\&E}}$ is similar to graph $\mathcal{G}_{\mathrm{E}}$ from figure C.1. It incorporates translation speed by using nodes consisting of five codons. These nodes are connected such that any node $p_n(x_n)p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})$ can only be connected to nodes $p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})p_{n+5}(x_{n+5})$ for any $x_i \in \{1, 2, ..., 6\}$. When the translation speed of five codons (a node) is too different from the original speed, it is pruned. The weights of the graph are again chosen such that any path length corresponds to the nucleosome energy of the corresponding sequence.

C.3 Recovering the original nucleosome energy and translation speed landscapes in host organisms

To create the closest possible translation speed landscape in a different organism, we modify graph $\mathcal{G}_{\mathrm{T\&E}}$ to become gene-wide and obtain graph $\mathcal{G}_{\mathrm{gene}}$ see figure C.3.

We also change the weights. We denote the weights corresponding to the closest possible translation speed landscape by w^{T}. Again these nodes are connected such that any node $p_n(x_n)p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})$ can only be connected to nodes $p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})p_{n+5}(x_{n+5})$ for any $x_i \in \{1, 2, ..., 6\}$, but now with weight w^{T} given by

$$w^{\mathrm{T}} = \left| \sum_{i=1}^{i=5} T_{\mathrm{original}}(p_{n+i}(s_{n+i})) - T_{\mathrm{host}}(p_{n+i}(x_{n+i})) \right| \tag{C.3}$$

where s_i denote the original codon choices in the original organism. Now this weight denotes the linear difference between five original codon choices in the organism human, and five (possibly different) choices in host organism yeast. Note that the translation speed functions T now explicitely denote for which organism they are calculated, the original or host.

To find the sequence G'' where both the translation speed landscape and the nucleosome energy landscape in a host organism are close to their original counterparts, we only need to change the weights of $\mathcal{G}_{\mathrm{gene}}$. The weight $w^{\mathrm{T\&E}}$ of edges between

$$p_n(x_n)p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})$$

and

$$p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})p_{n+5}(x_{n+5})$$

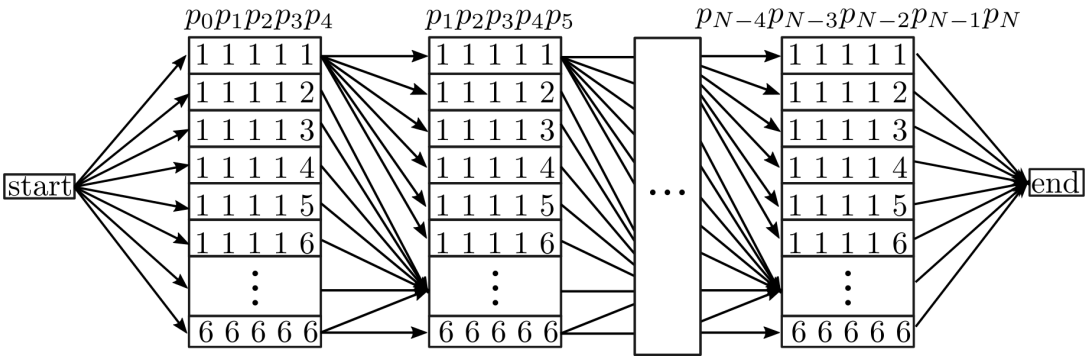


Figure C.3: Graph $\mathcal{G}_{\text{gene}}$ is similar to graph $\mathcal{G}_{\text{E}}$, see figure C.2. The graph includes the entire gene with N the number of codons on the gene. The weights of the graph depend on its purpose: the weights can be defined such that the closest possible translation speed landscape is found for a gene in a host organism, or a combination of the closest translation speed and nucleosome energy landscapes.

for any $x_i \in \{1, 2, ..., 6\}$ are now given by

$$w^{\text{T\&E}} = c_T w^{\text{T}} + c_E w^{\text{E}} \tag{C.4}$$

with

$$w^{\text{E}} = \sum_{j=-7}^{147+7-2} \left| \sum_{i=-7}^{i=7-2} E_{i+j}(S_{p+2+i}, S_{p+1+i}, S_{p+i}) - E_{i+j}(X_{p+2+i}, X_{p+1+i}, X_{p+i}) \right| \tag{C.5}$$

where X is a sequence of 15 base pairs, the sequence corresponding to

$$p_{n+1}(x_{n+1})p_{n+2}(x_{n+2})p_{n+3}(x_{n+3})p_{n+4}(x_{n+4})p_{n+5}(x_{n+5})$$

and S denotes the 15 base pairs in the original organism corresponding to

$$p_{n+1}(s_{n+1})p_{n+2}(s_{n+2})p_{n+3}(s_{n+3})p_{n+4}(s_{n+4})p_{n+5}(s_{n+5}).$$

C.4 Genetically modified organisms: many genes

In section 4.5, we introduced a method to, when one puts a gene in a different organism, this all three layers of information on the gene would be close to the original. Fig. 4.5 showed the results for one exon of the gene TNF. To remove possible bias from our results, we use the same method on a variety of human genes, randomly selected with a few non-biasing features: the exons of each transcript are fully translated and the first exon has a length ≥ 500 (the latter condition ensures a nucleosome landscape of significant size). The results are depicted in Figs. C.4-C.14.

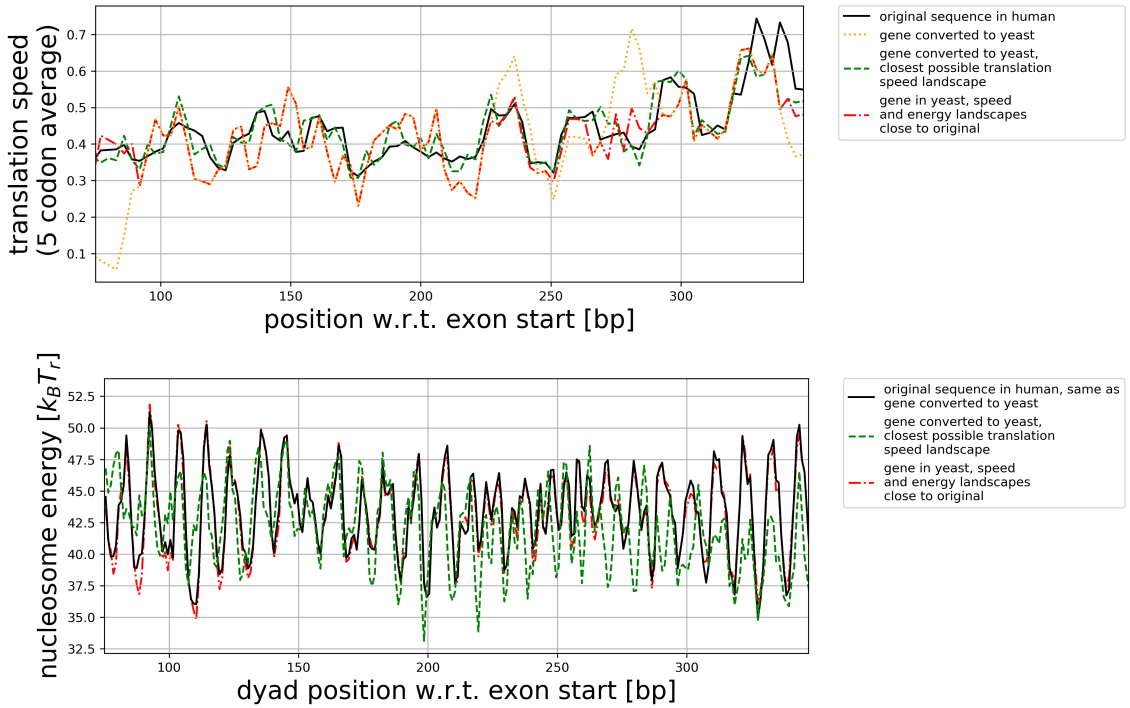


Figure C.4: Same as Fig. 4.5 but for the first exon of transcript OR6P1-001 of gene OR6P1 from human. This transcript was randomly selected with a few non-biasing features: the exons of each transcript are fully translated and the first exon has a length ≥ 500. As in Fig. 4.5, (a) depicts the translation speed landscape of this exon in three organisms: the original (human) and two possible host organisms: yeast and rice. Again, (b) shows the original landscape as well as the highest and lowest possible translation speed values in the hosts.

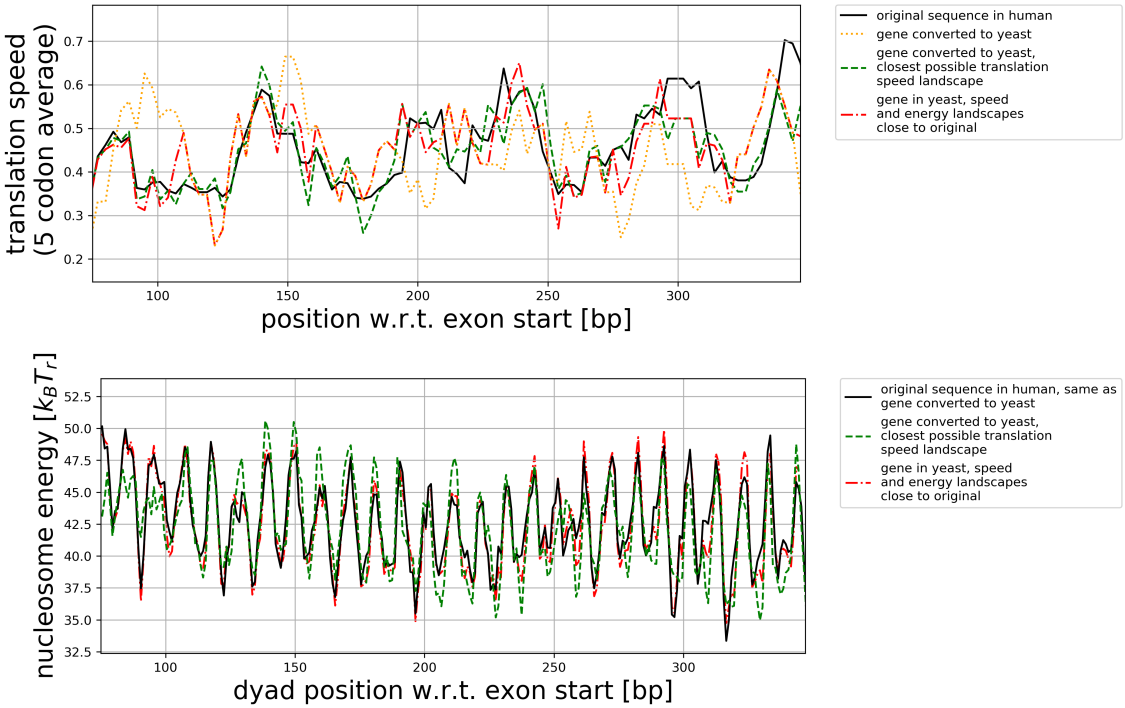


Figure C.5: Same as Fig. C.4 but for the first exon of transcript OR10J3-201 of gene OR10J3 from human.

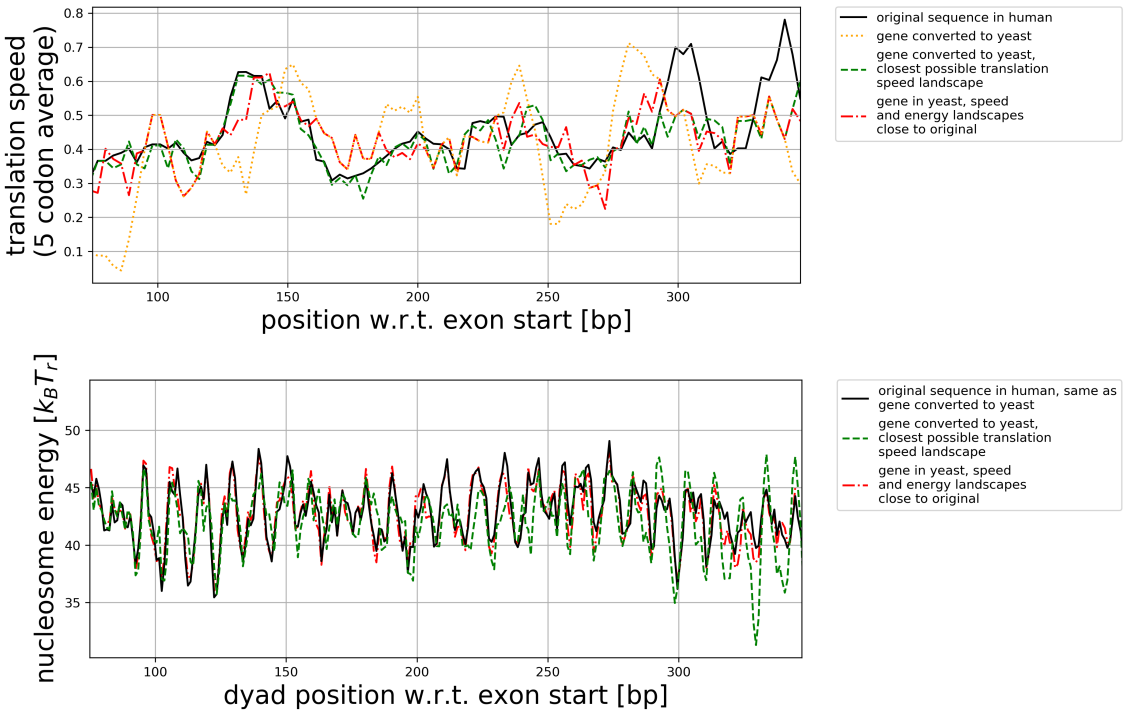


Figure C.6: Same as Fig. C.4 but for the first exon of transcript OR10T2-201 of gene OR10T2 from human.

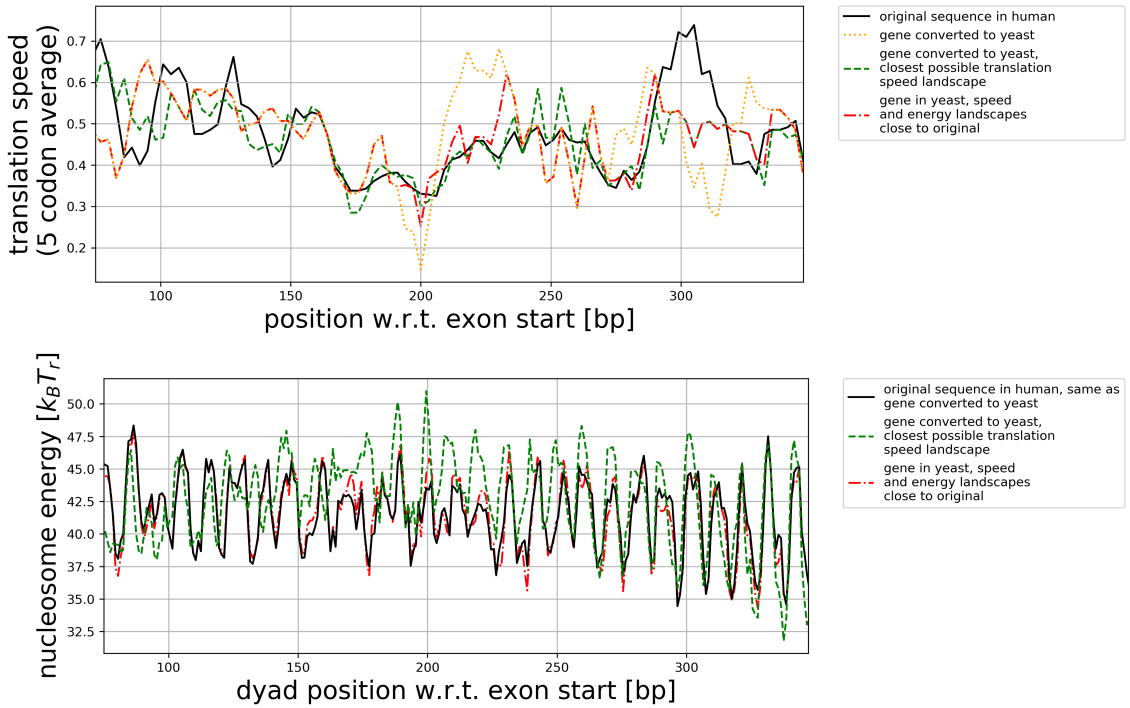


Figure C.7: Same as Fig. C.4 but for the first exon of transcript OR2T6-201 of gene OR2T6 from human.

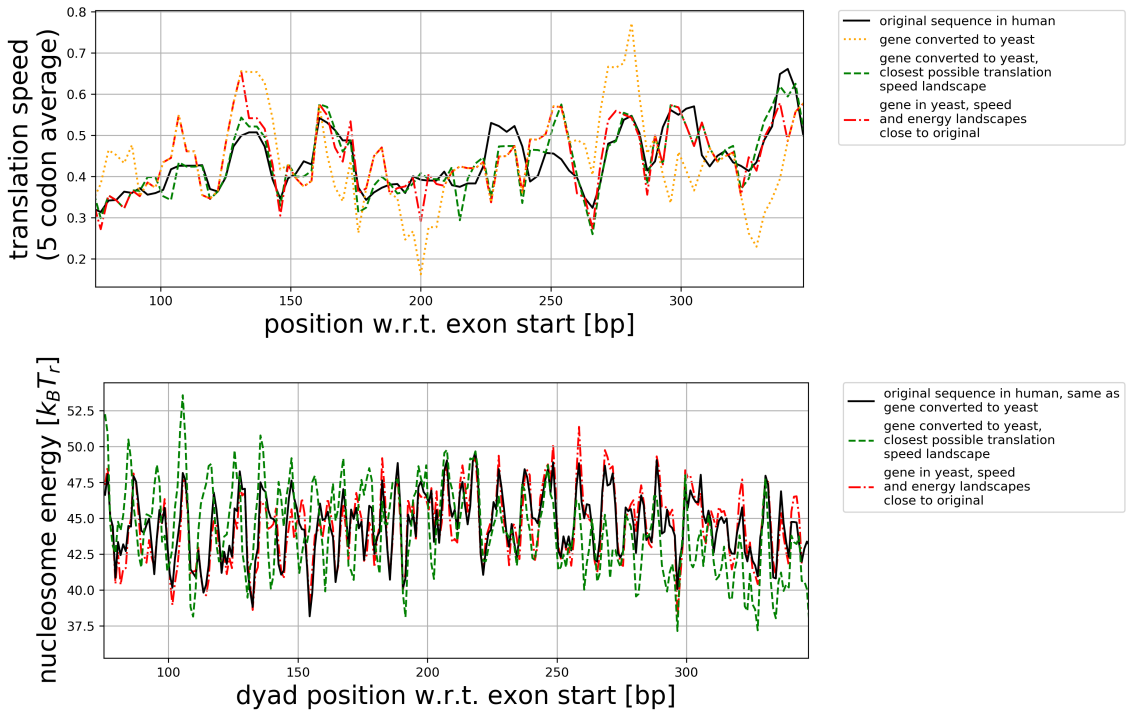


Figure C.8: Same as Fig. C.4 but for the first exon of transcript OR2M4-201 of gene OR2M4 from human.

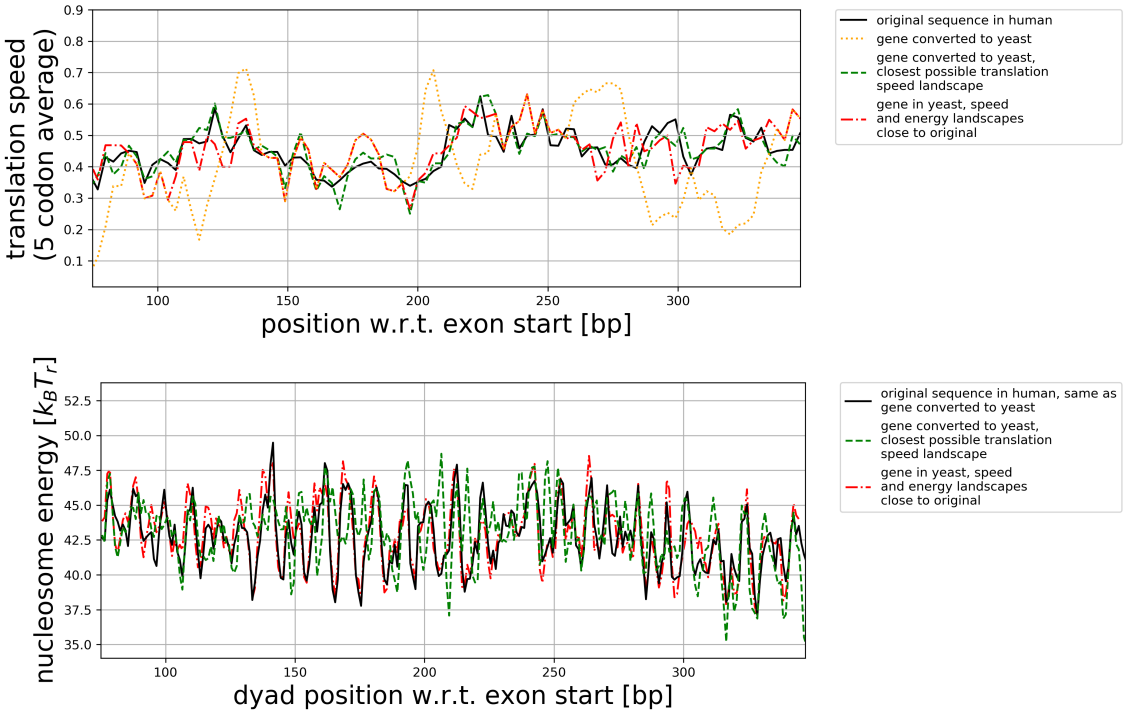


Figure C.9: Same as Fig. C.4 but for the first exon of transcript OR14K1-201 of gene OR14K1 from human.

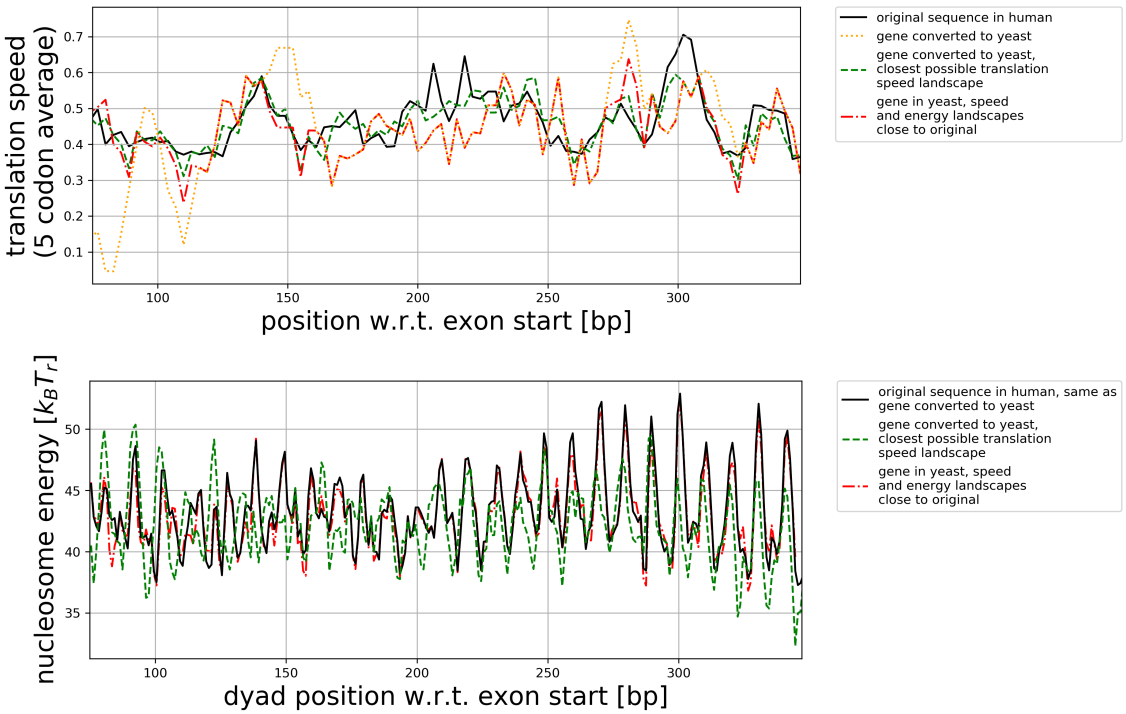


Figure C.10: Same as Fig. C.4 but for the first exon of transcript OR10K2 of gene OR10K2-201 from human.

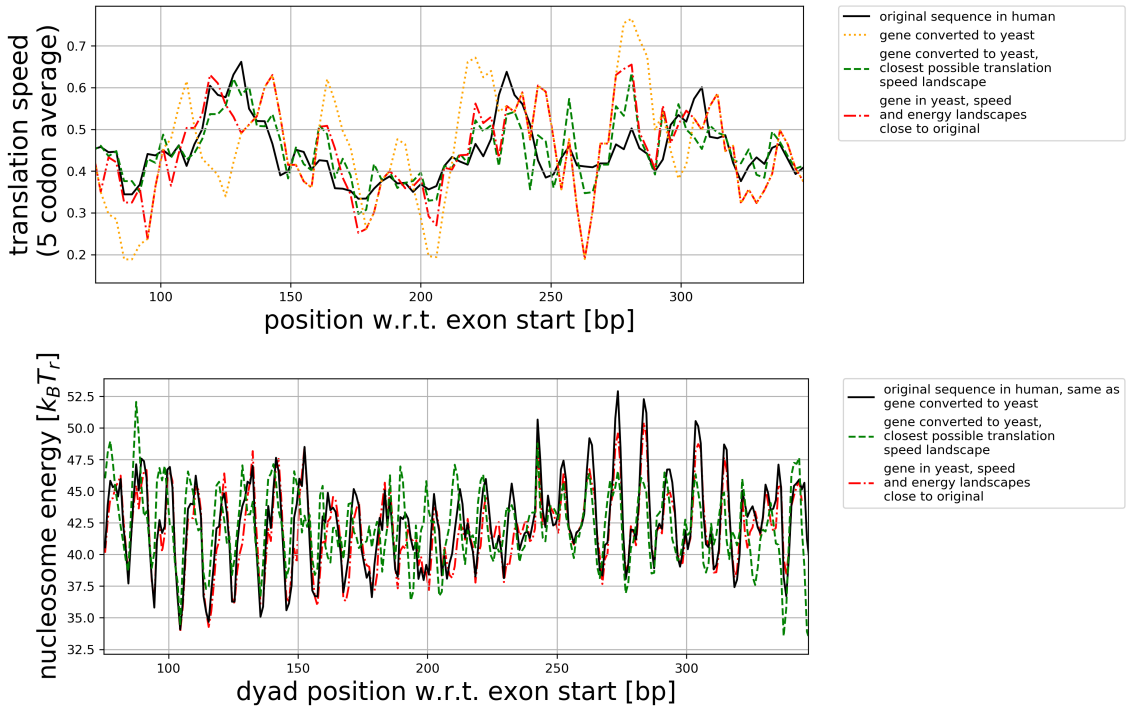
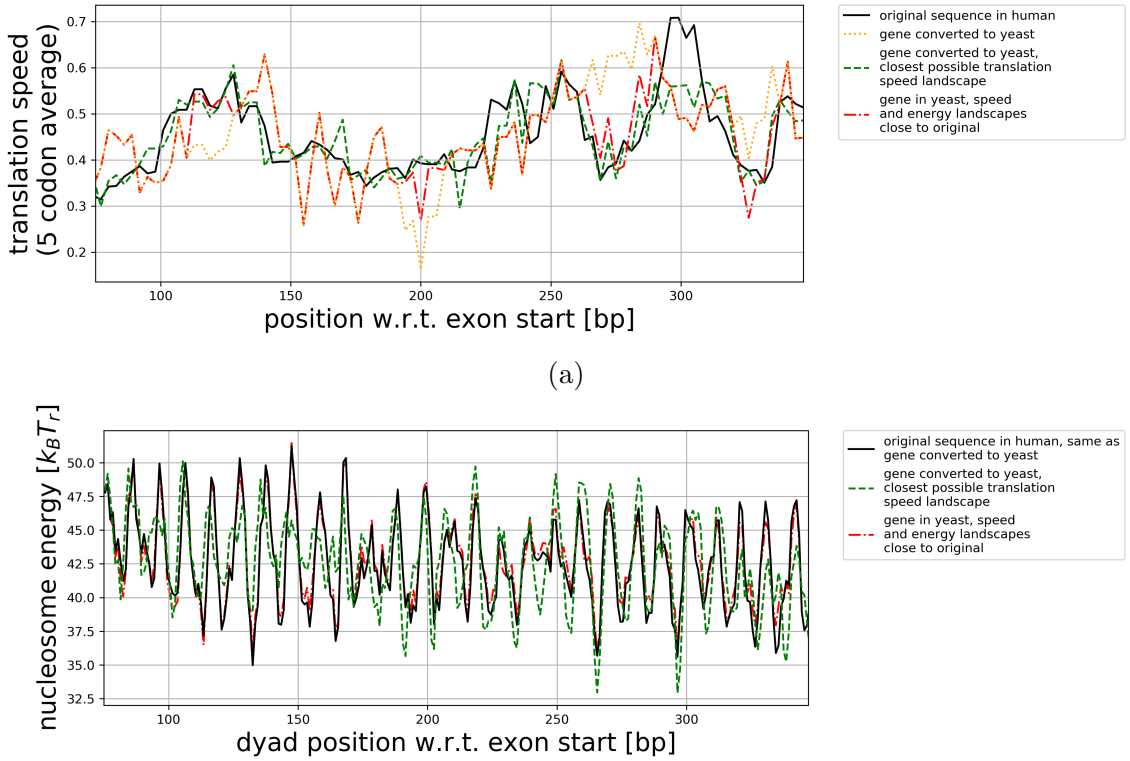


Figure C.11: Same as Fig. C.4 but for the first exon of transcript OR2T35-201 of gene OR2T35 from human.



(a)

Figure C.12: Same as Fig. C.4 but for the first exon of transcript OR2M7-201 of gene OR2M7 from human.

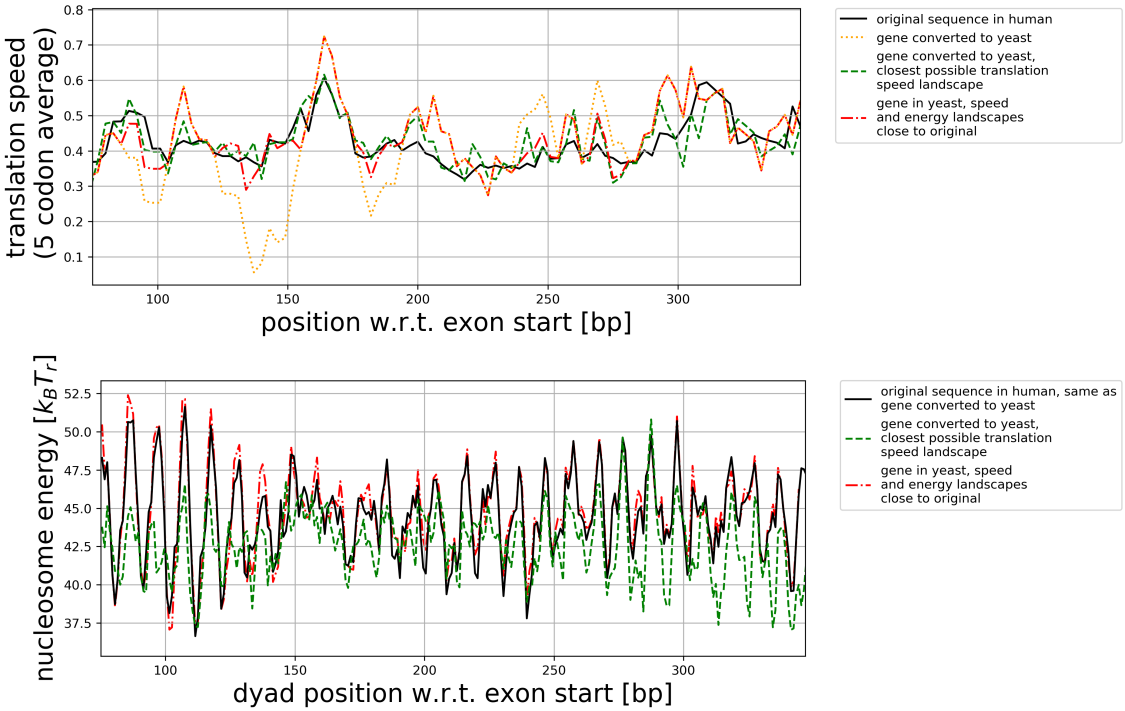


Figure C.13: Same as Fig. C.4 but for the first exon of transcript OR6Y1 of gene OR6Y1-201 from human.

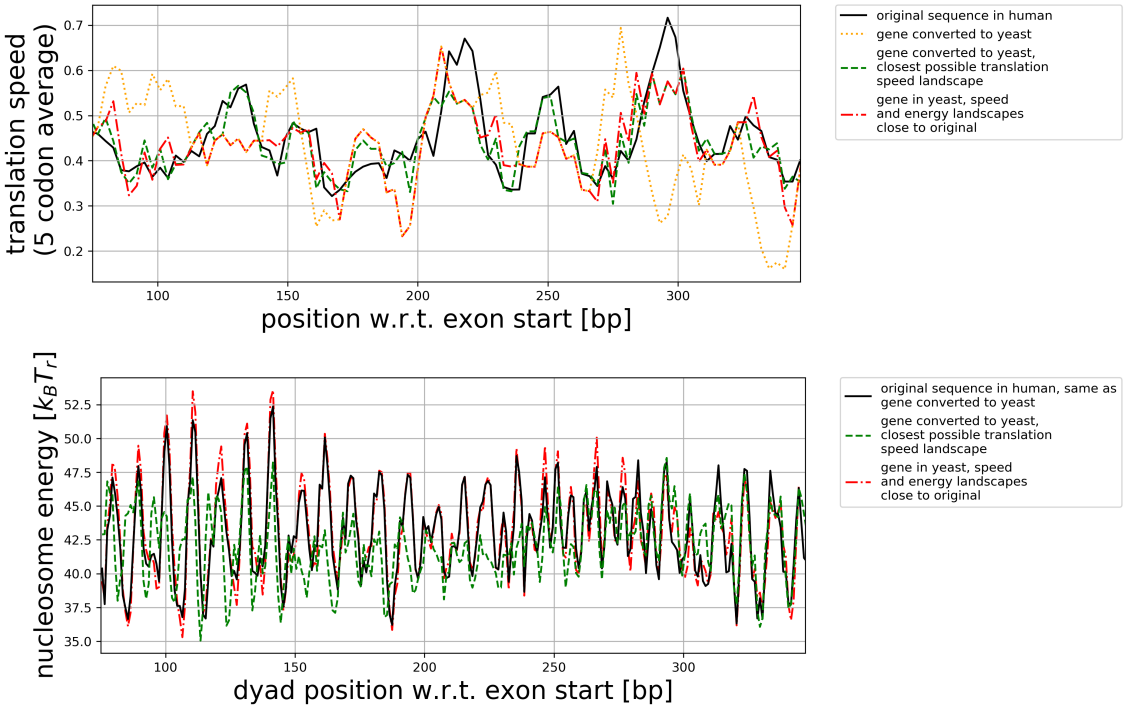


Figure C.14: Same as Fig. C.4 but for the first exon of transcript OR14C36 of gene OR14C36-201 from human.

Appendix D

How mechanical information is multiplexed on the transcribed regions of protein-coding genes

D.1 Data acquisition using Biomart

Here we provide a manual of sorts to obtain genome data the same way we did. We have acquired the genome data from the Ensembl Project website (www.ensembl.org), using their web-based tool Biomart. From Biomart, we always used the database Ensembl Genes 101. From the database we could pick an organism (such as Human genes). Under Filters, we expand the Gene menu and set the transcript type to protein-coding. Under Attributes, we chose Sequences, and under the SEQUENCES menu we chose Unspliced (Transcript), set Upstream flank to 1000 and Downstream flank to 1000. Under Attributes, in the HEADER INFORMATION menu, we chose, in this exact order: Transcript stable ID, Strand, Transcription start site (TSS), Genomic coding start, Genomic coding end, Exon region start (bp), Exon region end (bp). Using this header information we could determine the positions of the exons, introns and UTRs on any gene. By pressing the results button, and subsequently the Go button, one obtains the data in a plain text file.

Alternatively, one can download the data more efficiently. By pressing the XML button one can show the query in XML Web Service Format, which can then be used to download the data using a software package such as *wget*. We downloaded the required genomes by simply replacing the name of the organism in the XML string by the name of any available organism. An example of a command line to download the data for human is

```
wget -O human.txt 'http://www.ensembl.org/biomart/martservice?query=<?xml
    version="1.0" encoding="UTF-8"?><!DOCTYPE Query><Query
    virtualSchemaName = "default" formatter = "FASTA" header = "0"
    uniqueRows = "0" count = "" datasetConfigVersion = "0.6" ><Dataset
    name = "hsapiens_gene_ensembl" interface = "default" ><Filter name =
    "downstream_flank" value = "1000"/><Filter name = "upstream_flank"
    value = "1000"/><Filter name = "transcript_biotype" value =
    "protein_coding"/><Attribute name = "ensembl_transcript_id"
    /><Attribute name = "transcript_exon_intron" /><Attribute name =
    "strand" /><Attribute name = "transcription_start_site" /><Attribute
    name = "genomic_coding_start" /><Attribute name =
    "genomic_coding_end" /><Attribute name = "exon_chrom_start"
    /><Attribute name = "exon_chrom_end" /></Dataset></Query>'
```

For plants, e.g. for *Oryza sativa*, the command line can be given by

```
wget -O osativa.txt
    'http://plants.ensembl.org/biomart/martservice?query=<?xml
    version="1.0" encoding="UTF-8"?><!DOCTYPE Query><Query
    virtualSchemaName = "plants_mart" formatter = "FASTA" header = "0"
    uniqueRows = "0" count = "" datasetConfigVersion = "0.6" ><Dataset
    name = "osativa_eg_gene" interface = "default" ><Filter name =
    "downstream_flank" value = "1000"/><Filter name = "upstream_flank"
    value = "1000"/><Filter name = "transcript_biotype" value =
    "protein_coding"/><Attribute name = "ensembl_transcript_id"
    /><Attribute name = "transcript_exon_intron" /><Attribute name =
    "strand" /><Attribute name = "transcription_start_site" /><Attribute
    name = "genomic_coding_start" /><Attribute name =
    "genomic_coding_end" /><Attribute name = "exon_chrom_start"
    /><Attribute name = "exon_chrom_end" /></Dataset></Query>'
```

The Python code we use to turn the raw data into usable data is depicted below. An early version of this code was provided by Rhys Bird.

```
#this function requires a data file from the Biomart webtool. It provides
    three lists: cds2 is a list containing header information such as the
    name of the transcript, seq2 is a list containing 2000 bases
    corresponding to any transcipt, starting 1000 bp before the TSS. The
    list codingseq2 contains the same, but some of the base pairs have
    been replaced: all intronic bp are replaced by ";", 5'UTRs are
    replaced by "<", 3'UTR by ">".

def GetGenesShort(inputfile_string,upstream=1000, downstream=1000):
cds = []
seq = []
tempseq = ''
x=0
#here seq will be a list containing ALL bases corresponding to any
    transcipt
with open(inputfile_string) as inputfile:
```

```
for line in inputfile:
if line[0] == '>':
x+=1
if line.strip() != '>':
cds.append(line.replace('>','').split('|'))
if tempseq != '':
seq.append(tempseq)
tempseq = ''
else:
tempseq += line.replace('\n','')
for i in range(len(cds)):
cds[i][3] = cds[i][3].split(';')
cds[i][4] = cds[i][4].split(';')
cds[i][5] = cds[i][5].split(';')
cds[i][6] = cds[i][6].split(';')
codingseq = ['']*len(seq)
for i in range(len(seq)):
Xseq = [';']*len(seq[i])
#cds[i][1] tells us whether the raw data is 5' to 3', or 3' to 5'. In the
    latter scenario, the data is flipped such that everything is 5' to 3'.
#all UTRs are first replaced by ">", later we substitute it by "<" for
    5'UTRs.
if int(cds[i][1]) == 1:
for j in range(len(cds[i][3])):
start = int(cds[i][3][j]) - int(cds[i][2])+upstream
end = int(cds[i][4][j]) - (int(cds[i][2])-1)+upstream
Xseq[start:end] = [">" for _ in range(abs(end-start))]
list(seq[i][start:end])
for j in range(len(cds[i][5])):
start = int(cds[i][5][j]) - int(cds[i][2])+upstream
end = int(cds[i][6][j]) - (int(cds[i][2])-1)+upstream
Xseq[start:end] = list(seq[i][start:end])
elif int(cds[i][1]) == -1:
for j in range(len(cds[i][3])):
start = int(cds[i][2]) - int(cds[i][4][j])+upstream
end = int(cds[i][2]) - (int(cds[i][3][j])-1)+upstream
Xseq[start:end] = [">" for _ in range(abs(end-start))]
for j in range(len(cds[i][5])):
start = int(cds[i][2]) - int(cds[i][6][j])+upstream
end = int(cds[i][2]) - (int(cds[i][5][j])-1)+upstream
Xseq[start:end] = list(seq[i][start:end])
codingseq[i] += ''.join(Xseq)
for i in range(len(codingseq)):
codingseq[i]=upstream*","+codingseq[i][upstream:len(codingseq[i])-downstream]
    + downstream*","
seq2 = []
codingseq2 = []
cds2 = []
x=0
#here ">" is substituted by "<" for 5'UTRs. Also, all sequences are cut
```

```
    off after length 2000.
for i in range(len(seq)):
if len(seq[i])==len(codingseq[i]):
seq2.append(seq[i][0:2000])
cds2.append(cds[i][0:2000])
codingseq2.append([])
codingseq2[x]=codingseq[i][0:1000]
for j in range(1000,2000):
if codingseq[i][j] in ["A","T","C","G"]:
codingseq2[x]+=codingseq[i][j:2000]
break
if codingseq[i][j]==">":
codingseq2[x]+="<"
else:
codingseq2[x]+=codingseq[i][j]
x+=1
return cds2,seq2,codingseq2
```

D.2 List of animals used to obtain data

This section serves to provide a table with the list of animals of which data was obtained from biomart.

1	dmelanogaster	Drosophila melanogaster	A fruitfly
2	celegans	Caenorhabditis elegans	A nematode
3	cintestinales	Ciona intestinalis	Vase tunicate
4	csavignyi	Ciona savignyi	Solitary sea squirt
5	pmarinus	Petromyzon marinus	Sea lamprey
6	loculatus	Lepisosteus oculatus	Spotted gar
7	amexicanus	Astyanax mexicanus	Mexican tetra
8	trubripes	Takifugu rubripes	Japanese puffer
9	tnigroviridis	Tetraodon nigroviridis	Green spotted puffer
10	oniloticus	Oreochromis niloticus	Nile tilapia
11	gaculeatus	Gasterosteus aculeatus	Three-spined stickleback
12	olhni	Oryzias latipes	Japanese rice fish
13	pformosa	Poecilia formosa	Amazon molly
14	xmaculatus	Xiphophorus maculatus	Southern platyfish
15	lchalumnae	Latimeria chalumnae	West Indian Ocean coelacanth
16	xtropicalis	Xenopus tropicalis	Western clawed frog
17	acarolinensis	Anolis carolinensis	Green anole
18	psinensis	Pelodiscus sinensis	Chinese softshell turtle
19	falbicollis	Ficedula albicollis	Collared flycatcher
20	aplatyrhynchos	Anas platyrhynchos	Mallard
21	ggallus	Gallus gallus	Red junglefowl
22	mgallopavo	Meleagris gallopavo	Wild turkey
23	oanatinus	Ornithorhynchus anatinus	Platypus
24	mdomestica	Monodelphis domestica	Gray short-tailed opossum
25	sharrisii	Sarcophilus harrisii	Tasmanian devil
26	sscrofa	Sus scrofa	Wild boar
27	btaurus	Bos taurus	Cow
28	oaries	Ovis aries	Sheep
29	mlucifugus	Myotis lucifugus	Little brown bat
30	ecallabus	Equus caballus	Horse
31	fcatus	Felis catus	Cat
32	mpfuro	Mustela putorius furo	Ferret
33	ocuniculus	Oryctolagus cuniculus	European rabbit
34	cporcellus	Cavia porcellus	Guinea pig
35	itridecemlineatus	Ictidomys tridecemlineatus	Thirteen-lined ground squirrel
36	dordii	Dipodomys ordii	Ord's kangaroo rat
37	mmusculus	Mus musculus	House mouse
38	rnorvegicus	Rattus norvegicus	Brown rat
39	mmurinus	Microcebus murinus	Gray mouse lemur
41	ogarnettii	Otolemur garnettii	Northern greater galago
42	csyrichta	Carlito syrichta	Philippine tarsier
43	cjacchus	Callithrix jacchus	Common marmoset
44	csabaeus	Chlorocebus sabaeus	Green monkey
45	panubis	Papio anubis	Olive baboon

continues on the next page

46	mmulatta	Macaca mulatta	Rhesus macaque
47	nleucogenys	Nomascus leucogenys	Northern white-cheeked gibbon
48	pabelii	Pongo abelii	Sumatran orangutan
49	ggorilla	Gorilla gorilla gorilla	Western lowland gorilla
50	hsapiens	Homo sapiens	Human
51	ptroglodytes	Pan troglodytes	Chimpanzee

Table D.1: This table depicts the list of animals used to obtain data. It depicts the names of the organisms as they appear in Biomart, as well as their Latin names and a short description of the animal.

Bibliography

[1] R. Dahm, "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research," *Human Genetics*, vol. 122, no. 6, pp. 565–581, 2008.

[2] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, A. Bretscher, H. Ploegh, A. Amon, and M. P. Scott, *Molecular Cell Biology, Seventh Edition*. W. H. Freeman and Company, 2013.

[3] G. A. Maston, S. K. Evans, and M. R. Green, "Transcriptional regulatory elements in the human genome," *Annual Review of Genomics and Human Genetics*, vol. 7, no. 1, pp. 29–59, 2006.

[4] S. C. Satchwell, H. R. Drew, and A. A. Travers, "Sequence periodicities in chicken nucleosome core DNA," *Journal of Molecular Biology*, vol. 191, pp. 659–675, 1986.

[5] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thaström, I. K. M. Y. Field, J. Z. Wang, and J. Widom, "A genomic code for nucleosome positioning," *Nature*, vol. 442, pp. 772–778, 2006.

[6] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature*, vol. 458, pp. 362–366, 2009.

[7] P. T. Lowary and J. Widom, "Nucleosome packaging and nucleosome positioning of genomic DNA," *Proceedings of the National Academy of Sciences*, vol. 94, no. 4, pp. 1183–1188, 1997.

[8] G. Drillon, F. A. B. Audit, and A. Arneodo, "Evidence of selection for an accessible nucleosomal array in human," *BMC Genomics*, vol. 17, p. 526, 2016.

[9] M. Tompitak, C. Vaillant, and H. Schiessel, "Genomes of multicellular organisms have evolved to attract nucleosomes to promoter regions," *Biophysical Journal*, vol. 112, pp. 505–511, 2017.

[10] M. Tompitak, G. T. Barkema, and H. Schiessel, "Benchmarking and refining probability-based models for nucleosome-DNA interaction," *BMC Bioinformatics*, vol. 18, p. 157, 2017.

[11] B. Eslami-Mossallam, H. Schiessel, and J. van Noort, "Nucleosome dynamics: Sequence matters," *Advances in Colloid and Interface Science*, vol. 232, pp. 101–113, 2016.

[12] C. R. Calladine and H. R. Drew, "A base-centred explanation of the B-to-A transition in DNA," *Journal of Molecular Biology*, vol. 178, pp. 773–782, 1984.

[13] B. D. Coleman, W. K. Olson, and D. Swigdon, "Theory of sequence-dependent DNA elasticity," *Journal of Chemical Physics*, vol. 118, pp. 7127–7140, 2003.

[14] B. Eslami-Mossallam, R. D. Schram, M. Tompitak, J. van Noort, and H. Schiessel, "Multiplexing genetic and nucleosome positioning codes: A computational approach," *PLoS ONE*, vol. 11, p. e0156905, 2016.

[15] M. Y. Tolstorukov, A. V. Colasanti, D. M. McCandlish, W. K. Olson, and V. B. Zhurkin, "A novel role-and-slide mechanism for DNA folding in chromatin: implications for nucleosome positioning," *Journal of Molecular Biology*, vol. 371, pp. 725–738, 2007.

[16] T. Drsata, N. Spackova, P. Jurecka, M. Zgarbova, S. Sponer, and F. Lankas, "Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning," *Nucl. Acids Res.*, vol. 42, pp. 7383–7394, 2014.

[17] D. Norouzi and F. Mohammad-Rafiee, "DNA conformation and energy in nucleosome core: a theoretical approach," *J. Biomol. Struct. Dyn.*, vol. 32, pp. 104–114, 2014.

[18] A. Fathizadeh, A. B. Besya, M. R. Ejtehadi, and H. Schiessel, "Rigid-body molecular dynamics of DNA inside a nucleosome," *European Physical Journal E*, vol. 36, p. 21, 2013.

[19] L. de Bruin, M. Tompitak, B. Eslami-Mossallam, and H. Schiessel, "Why do nucleosomes unwrap asymmetrically?," *J. Phys. Chem. B.*, vol. 120, pp. 5855–5863, 2016.

[20] C. Anselmi, G. Bocchinfuso, P. D. Santis, M. Savino, and A. Scipioni, "A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability," *Biophysical Journal*, vol. 79, pp. 601–613, 2000.

[21] C. Vaillant, B. Audit, and A. Arneodo, "Experiments confirm the influence of genome long-range correlations on nucleosome positioning," *Physical Review Letters*, vol. 99, p. 218103, 2007.

[22] S. Balasubramanian, F. Xu, and W. K. Olson, "DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences," *Biophysical Journal*, vol. 96.

[23] M. L. Fredman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms," *J. ACM*, vol. 34, no. 3, p. 596–615, 1987.

[24] J. Y. Yen, "Finding the k shortest loopless paths in a network," *Management Science*, vol. 17, no. 11, pp. 712–716, 1971.

[25] A. Bitran, W. M. Jacobs, X. Zhai, and E. Shakhnovich, "Cotranslational folding allows misfolding-prone proteins to circumvent deep kinetic traps," *Proceedings of the National Academy of Sciences*, vol. 117, no. 3, pp. 1485–1495, 2020.

[26] H. Gingold, D. Tehler, N. Christoffersen, M. Nielsen, F. Asmar, S. Kooistra, N. Christophersen, L. L. Christensen, M. Borre, D. Karina, L. Dyrskjøt, C. Andersen, E. Hulleman, T. Wurdinger, E. Ralfkiær, K. Helin, K. Grønbaek, T. Orntoft, S. Waszak, and Y. Pilpel, "A dual program for translation regulation in cellular proliferation and differentiation," *Cell*, vol. 158, pp. 1281–92, 2014.

[27] K. C. Stein and J. Frydman, "The stop-and-go traffic regulating protein biogenesis: how translation kinetics controls proteostasis," *Journal of Biological Chemistry*, vol. 294, no. 6, pp. 2076–2084, 2019.

[28] K. Stein and J. Frydman, "The stop-and-go traffic regulating protein biogenesis: How translation kinetics controls proteostasis," *Journal of Biological Chemistry*, vol. 294, p. jbc.REV118.002814, 2018.

[29] S. Rudorf, M. Thommen, M. V. Rodnina, and R. Lipowsky, "Deducing the kinetics of protein synthesis in vivo from the transition rates measured in vitro," *PLoS Comput Biol*, vol. 10, no. 10, p. e1003909, 2014.

[30] E. Dolgin, "The most popular genes in the human genome," *Nature*, vol. 551, no. 7681, pp. 427–431, 2017.

[31] M. Zuiddam, R. Everaers, and H. Schiessel, "Physics behind the mechanical nucleosome positioning code," *Physical Review E*, vol. 96, p. 052412, 2017.

[32] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 11163–11168, 1998.

[33] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 åresolution," *Nature*, vol. 389, pp. 251–260, 1997.

[34] H. Schiessel, "The physics of chromatin," *J. Phys.: Condens. Matter*, vol. 15, pp. R699–R774, 2003.

[35] T. Vavouri and B. Lehner, "Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome," *PLoS Genet.*, vol. 7, p. e1002036, 2011.

[36] A. V. Morozov, K. Fortney, D. A. Gaykalova, V. M. Studitsky, J. Widom, and E. D. Siggia, "Using DNA mechanics to predict in vitro nucleosome positions and formation energies," *Nucl. Acids Res*, vol. 37, pp. 4707–4722, 2009.

[37] N. B. Becker and R. Everaers, "DNA nanomechanics in the nucleosome," *Structure*, vol. 17, pp. 579–589, 2009.

[38] N. B. Becker, L. Wolff, and R. Everaers, "Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials," *Nucl Acids Res.*, vol. 34, pp. 5638–5649, 2006.

[39] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, "Conformational analysis of nucleic acids revisited: Curves+," *Nucleic Acids Res.*, vol. 37, pp. 5917–5929, 2009.

[40] V. B. Teif, "General transfer matrix formalism to calculate DNA-protein-drug binding in gene regulation: application to o_r operator of phage λ," *Nucleic Acids Res.*, vol. 35, p. e80, 2007.

[41] G. Chevereau, A. Arneodo, and C. Vaillant, "Influence of the genomic sequence on the primary structure of chromatin," *Frontiers in Life Science*, vol. 5, pp. 29–68, 2011.

[42] G. S. Freeman, J. P. Lequieu, D. M. Hinckley, J. K. Whitmer, and J. de Pablo, "DNA shape dominates sequence affinity in nucleosome formation," *Physical Review Letters*, vol. 113, p. 168101, 2014.

[43] M. Zuiddam and H. Schiessel, "Shortest paths through synonymous genomes," *Physical Review E*, vol. 99, p. 012422, 2019.

[44] G. Meersseman, S. Pennings, and E. Bradbury, "Mobile nucleosomes–a general behavior.," *The EMBO Journal*, vol. 11, no. 8, pp. 2951–2959, 1992.

[45] I. M. Kulic and H. Schiessel, "Chromatin dynamics: Nucleosomes go mobile through twist defects," *Physical Review Letters*, vol. 91, p. 148103, 2003.

[46] G. B. Brandani, T. Niina, C. Tan, and S. Takada, "DNA sliding in nucleosomes via twist defect propagation revealed by molecular simulations," *Nucleic Acids Research*, vol. 46, no. 6, pp. 2788–2801, 2018.

[47] H. Schiessel, J. Widom, R. F. Bruinsma, and W. M. Gelbart, "Polymer reptation and nucleosome repositioning," *Physical Review Letters*, vol. 86, pp. 4414–4417, 2001.

[48] I. M. Kulic and H. Schiessel, "Nucleosome Repositioning via Loop Formation," *Biophysical Journal*, vol. 84, no. 5, pp. 3197–3211, 2003.

[49] J. Lequieu, D. C. Schwartz, and J. J. de Pablo, "In silico evidence for sequence-dependent nucleosome sliding," *Proceedings of the National Academy of Sciences*, vol. 114, no. 44, pp. E9197–E9205, 2017.

[50] T. Niina, G. Brandani, C. Tan, and S. Takada, "Sequence-dependent nucleosome sliding in rotation-coupled and uncoupled modes revealed by molecular simulations," *PLOS Computational Biology*, vol. 13, p. e1005880, 2017.

[51] J. Winger, I. Nodelman, R. Levendosky, and G. Bowman, "A twist defect mechanism for ATP-dependent translocation of nucleosomal DNA," *eLife*, vol. 7, 2018.

[52] G. B. Brandani, T. Niina, C. Tan, and S. Takada, "DNA sliding in nucleosomes via twist defect propagation revealed by molecular simulations," *Nucleic Acids Research*, vol. 46, no. 6, pp. 2788–2801, 2018.

[53] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thaström, Y. Field, I. Moore, J. Wang, and J. Widom, "A genomic code for nucleosome positioning," *Nature*, vol. 442, pp. 772–8, 2006.

[54] K. Struhl and E. Segal, "Determinants of nucleosome positioning," *Nature structural & molecular biology*, vol. 20, pp. 267–73, 2013.

[55] R. D. Kornberg and L. Stryer, "Statistical distributions of nucleosomes: non-random locations by a stochastic mechanism," *Nucleic Acids Research*, vol. 16, no. 14, pp. 6677–6690, 1988.

[56] G. Chevereau, L. Palmeira, C. Thermes, A. Arneodo, and C. Vaillant, "Thermodynamics of intragenic nucleosome ordering," *Physical Review Letters*, vol. 103, p. 188103, 2009.

[57] F. Brunet, B. Audit, G. Drillon, F. Argoul, J. Volff, and A. Arneodo, "Evidence for DNA sequence encoding of an accessible nucleosomal array across vertebrates," *Biophysical Journal*, vol. 114, 2018.

[58] M. Tompitak, L. de Bruin, B. Eslami-Mossallam, and H. Schiessel, "Designing nucleosomal force sensors," *Physical Review E*, vol. 95, p. 052402, 2017.

[59] T. E. Shrader and D. M. Crothers, "Artificial nucleosome positioning sequences," *Proc. Natl. Acad. Sci. USA*, vol. 86, pp. 7418–7422, 1989.

[60] T. E. Shrader and D. M. Crothers, "Effect of DNA sequence and histone-histone interactions on nucleosome placement," *Journal of Molecular Biology*, vol. 216, pp. 69–84, 1990.

[61] J. A. Wondergem, H. Schiessel, and M. Tompitak, "Performing selex experiments in silico.," *The Journal of chemical physics*, vol. 147 17, p. 174101, 2017.

[62] A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow, "Determinants of nucleosome organization in primary human cells," *Nature*, vol. 474, pp. 516–520, 2011.

[63] S. Ercan, S. Lubling, E. Segal, and J. D. Lieb, "High nucleosome occupancy is encoded at x-linked gene promoters in c. elegans," *Genome Research*, vol. 21, p. 237–244, 2011.

[64] J. Culkin, L. de Bruin, M. Tompitak, R. Phillips, and H. Schiessel, "The role of DNA sequence in nucleosome breathing," *European Physical Journal E*, vol. 40, p. 106, 2017.

[65] K. J. Polach and J. Widom, "Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation," *Journal of Molecular Biology*, vol. 254, pp. 130–149, 1995.

[66] J. D. Anderson and J. Widom, "Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites," *Journal of Molecular Biology*, vol. 296, pp. 979–987, 2000.

[67] T. T. M. Ngo, Q. Zhang, R. Zhou, J. G. Yodh, and T. Ha, "Asymmetric unwrapping of nucleosomes under tension directed by DNA local flexibility," *Cell*, vol. 160, pp. 1135–1144, 2015.

[68] E. Segal and J. Widom, "Poly(dA:dT) tracts: major determinants of nucleosome organization," *Current Opinion in Structural Biology*, vol. 19, pp. 65–71, 2009.

[69] P. T. Lowary and J. Widom, "New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning," *Journal of Molecular Biology*, vol. 276, pp. 19–42, 1998.

[70] S. Todolli, P. Perez, N. Clauvelin, and W. Olson, "Contributions of sequence to the higher-order structures of DNA," *Biophysical Journal*, vol. 112, 2016.

[71] G. Rosanio, J. Widom, and O. C. Uhlenbeck, "In vitro selection of DNAs with an increased propensity to form small circles," *Biopolymers*, vol. 103, no. 6, pp. 303–320, 2015.

[72] H. Meng, K. Andresen, and J. van Noort, "Quantitative analysis of single-molecule force spectroscopy on folded chromatin fibers," *Nucleic Acids Research*, vol. 43, no. 7, pp. 3578–3590, 2015.

[73] B. E. de Jong, T. B. Brouwer, A. Kaczmarczyk, B. Visscher, and J. van Noort, "Rigid basepair monte carlo simulations of one-start and two-start chromatin fiber unfolding by force," *Biophysical Journal*, vol. 115, no. 10, pp. 1848–1859, 2018.

[74] E. N. Trifonov, "The multiple codes of nucleotide sequences," *Bulletin of Mathematical Biology*, vol. 51, no. 4, pp. 417–432, 1989.

[75] E. N. Trifonov and J. L. Sussman, "The pitch of chromatin DNA is reflected in its nucleotide sequence," *Proceedings of the National Academy of Sciences*, vol. 77, no. 7, pp. 3816–3820, 1980.

[76] D. Tillo and T. Hughes, "G+C content dominates intrinsic nucleosome occupancy," *BMC Bioinformatics*, vol. 10, p. 442, 2009.

[77] J. Neipel, G. Brandani, and H. Schiessel, "Translational nucleosome positioning: A computational study," *Physical Review E*, vol. 101, p. 022405, 2020.

[78] F. Mohammad-Rafiee, I. M. Kulić, and H. Schiessel, "Theory of nucleosome corkscrew sliding in the presence of synthetic DNA ligands," *Journal of Molecular Biology*, vol. 344, no. 1, pp. 47–58, 2004.

[79] A. Z. Guo, J. Lequieu, and J. J. de Pablo, "Extracting collective motions underlying nucleosome dynamics via nonlinear manifold learning," *The Journal of Chemical Physics*, vol. 150, no. 5, p. 054902, 2019.

[80] S. Rudnizky, H. Khamis, O. Malik, P. Melamed, and A. Kaplan, "The base pair-scale diffusion of nucleosomes modulates binding of transcription factors," *Proceedings of the National Academy of Sciences*, vol. 116, no. 25, pp. 12161–12166, 2019.

[81] M. Li, X. Xia, Y. Tian, Q. Jia, X. Liu, Y. Lu, M. Li, X. Li, and Z. Chen, "Mechanism of DNA translocation underlying chromatin remodelling by snf2," *Nature*, vol. 567, p. 409–413, 2019.

[82] A. Sabantsev, R. Levendosky, X. Zhuang, G. Bowman, and S. Deindl, "Direct observation of coordinated DNA movements on the nucleosome during chromatin remodelling," *Nature Communications*, vol. 10, p. 1720, 2019.

[83] H. Schiessel and R. Blossey, "Pioneer transcription factors in chromatin remodeling: The kinetic proofreading view," *Physical Review E*, vol. 101, p. 040401, 2020.

[84] H. Dong, L. Nilsson, and C. G. Kurland, "Co-variation of tRNA abundance and codon usage in Escherichia *coli* at different growth rates," *Journal of molecular biology*, vol. 260, no. 5, pp. 649–663, 1996.

[85] M. A. Collart and B. Weiss, "Ribosome pausing, a dangerous necessity for co-translational events," *Nucleic acids research*, vol. 48, no. 3, pp. 1043–1055, 2020.

[86] D. López and F. Pazos, "Protein functional features are reflected in the patterns of mrna translation speed," *BMC genomics*, vol. 16, no. 1, pp. 1–13, 2015.

[87] S. Pechmann and J. Frydman, "Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding," *Nature structural & molecular biology*, vol. 20, no. 2, p. 237, 2013.

[88] 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*, vol. 495, no. 7439, pp. 111–115, 2013.

[89] E. P. O'Brien, P. Ciryam, M. Vendruscolo, and C. M. Dobson, "Understanding the influence of codon translation rates on cotranslational protein folding," *Accounts of chemical research*, vol. 47, no. 5, pp. 1536–1544, 2014.

[90] M. Liutkute, E. Samatova, and M. V. Rodnina, "Cotranslational folding of proteins on the ribosome," *Biomolecules*, vol. 10, no. 1, p. 97, 2020.

[91] W. Chu, "Tumor necrosis factor," *Cancer letters*, vol. 328, no. 2, pp. 222–225, 2013.

[92] J. Frank and R. L. Gonzalez Jr, "Structure and dynamics of a processive brownian motor: the translating ribosome," *Annual review of biochemistry*, vol. 79, pp. 381–412, 2010.

[93] I. Wohlgemuth, C. Pohl, J. Mittelstaet, A. L. Konevega, and M. V. Rodnina, "Evolutionary optimization of speed and accuracy of decoding on the ribosome," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, no. 1580, pp. 2979–2986, 2011.

[94] D. N. Wilson and J. H. D. C. Cate, "The structure and function of the eukaryotic ribosome," *Cold Spring Harbor Perspectives in Biology*, vol. 4, no. 5, p. a011536, 2012.

[95] V. Haberle and A. Stark, "Eukaryotic core promoters and the functional basis of transcription initiation," *Nature reviews. Molecular cell biology*, vol. 19, no. 10, pp. 621–637, 2018.

[96] R. I. Vishwanath, "Nucleosome positioning: bringing order to the eukaryotic genome," *Trends in Cell Biology*, vol. 22, no. 5, pp. 250–256, 2012.

[97] M. Han and M. Grunstein, "Nucleosome loss activates yeast downstream promoters in vivo," *Cell*, vol. 55, no. 6, p. 1137—1145, 1988.

[98] R. Dreos, G. Ambrosini, and P. Bucher, "Influence of rotational nucleosome positioning on transcription start site selection in animal promoters," *PLoS computational biology*, vol. 12, no. 10, pp. e1005144–e1005144, 2016.

[99] D. Tillo, N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, Y. Field, J. D. Lieb, J. Widom, E. Segal, and T. R. Hughes, "High nucleosome occupancy is encoded at human regulatory sequences," *PloS one*, vol. 5, no. 2, pp. e9129–e9129, 2010.

[100] M. Chorev and L. Carmel, "The function of introns," *Frontiers in genetics*, vol. 3, pp. 55–55, 2012.

[101] F. Mignone, C. Gissi, S. Liuni, and G. Pesole, "Untranslated regions of mrnas," *Genome biology*, vol. 3, no. 3, 2002.

[102] C. team of Ensembl, "Species tree." `http://Feb2021.archive.ensembl.org/info/about/speciestree.html`. Accessed: 1-3-2021.

[103] 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*, vol. 141, no. 2, p. 344—354, 2010.

Summary

DNA carries various forms of information. Out of these forms of information, the most well-known is classical genetic information, which features genes that code for RNAs and proteins. This genetic code consists of sequences of bases, shorthandedly called A, T, C and G. A sequence of these bases can code for a specific protein. Throughout this dissertation we discuss what is often referred to as the second layer of information on DNA: DNA mechanics. A sequence consisting of only A's and T's will bend differently from a sequence of G's and C's. This is because the intrinsic shape of DNA, as well as its flexibility, depend on the choice of its bases. An important consequence of this mechanical layer of information is the positioning of nucleosomes.

Nucleosomes consist of 147 base pairs (bp) of DNA wrapped around a protein core, like a string around a spool. They are considered the fundamental building blocks of chromosomes and are responsible for making the DNA compact and serve to form its higher-order structures. Nucleosomes also play an important role in the regulation of DNA. They may physically restrict transcription factors (proteins that regulate transcription) from reaching transcription factor binding sites on the DNA, simply by being situated on such a site. A cell can chemically modify a nucleosome (or more precisely its 'histone tails') such that a binding site becomes accessible again. These modifications may even be inherited by the offspring of an organism, which is an example of so-called epigenetics. By either allowing or restricting access to a binding site, a nucleosome may serve as an on/off switch, of which the location is extremely important.

The location of nucleosomes on the DNA is affected by the mechanical information on the DNA. The DNA on a nucleosome must be curved in a specific way, favoured by some, disliked by other sequences. This leads to a nucleosome positioning code, often represented by the probability to find a dinucleotide step (a base followed by another base) at different locations on the nucleosome. For example, the probability to find a TT, AA, or TA dinucleotide step is highest where the minor groove of DNA faces the protein core, while finding a GC step is most likely to happen at positions where the major groove faces the protein core.

Models have been created that can reproduce but cannot explain these rules. Especially the rules concerning the GC step turned out to be counterintuitive, since its probability peaks where its deformation energy peaks, too. The first main result of this dissertation is an explanation of these rules. In Chapter 2 we approach this problem by using an analytically-tractable model that reproduces the main rules. Using the Transfer Matrix Method and a novel approximation, we demystify the rules of GC and methodetically explain the other rules.

Another approach is not to investigate these dinucleotide rules, but to look at

the malleability of long stretches of DNA to contain mechanical information. In Chapter 3 we do this by investigating the very best and very worst nucleosome-attracting sequences. We map all possible DNA sequences on a graph, weighted using the probabilistic model of Tompitak et al. [10]. By using a shortest path algorithm on this graph, we find the sequences with highest and lowest possible nucleosome wrapping energy. Using a k-shortest path algorithm we find the k-highest and lowest possible energies. The two huge advantages of this method (over other methods such as Monte Carlo simulations) are that shortest-path algorithms are in principle exact and fast.

DNA in real, living organisms does not have the luxury to only code for mechanical signals. Real DNA also needs to code for RNAs and proteins. Theoretically, these two layers of information (genetic and mechanic) can exist on the same piece of DNA. A protein consists of one or more amino acid chains. The properties of a protein depend on the type and order of amino acids in such a chain, which is encoded by so-called codons (a triplet of bases) on the DNA. In many cases, multiple synonymous codons exist that code for the same amino acid. This is called the degeneracy of the genetic code. Because of this degeneracy, multiple sequences may code for the same protein while having a diverse range of mechanical properties. As a result, a protein-coding sequence may contain mechanical signals as well as genetic information.

Using the degeneracy of the genetic code we evaluated the freedom of DNA to contain both forms of information. We do so by using graphs that contain all synonymous ways to code for the same protein. On these graphs, a shortest path algorithm provides sequences with highest and lowest possible nucleosome wrapping energy that still codes for the original protein. Furthermore we present a heuristic method to create well-positioned nucleosomes on top of a gene, and showcase this method using the genome of *Saccharomyces cerevisiae*, baker's yeast. This method, too, relies on graph representations of genes and shortest path algorithms. We investigate *all* positions on *all* protein-coding genes on yeast, and manage to create nucleosome positioning signals with single-bp resolution for 99.897% of all positions.

So far we have mentioned two layers of information on DNA. By doing so, we ignored an *additional* layer of information: translation speed. Translation speed refers to the rate at which a protein is created. This rate has important consequences for the resulting proteins, as it does not only influence the rate at which proteins can be produced, but it also affects how the amino acid chain folds during translation. This folding affects the function and fidelity of a protein. The rate at which an amino acid is added to a chain depends on the codon on the DNA (and is cell-specific and species-dependent). Even though synonymous codons code for the same amino acid, the choice of codon does have an effect on the translation speed landscape. Therefore, changing a gene while keeping the genetic code intact may lead to dysfunctional proteins. Because of this, in Chapter 4 we include the conservation of translation speed in our analysis. We present several approaches that incorporate translation speed by using the graph representations of genes introduced in Chapter 3. These approaches either use pruning, where nodes in a graph are cut out that would lead to translation speed landscapes that are too different, or alter the weights of a graph to contain translation speed as well as nucleosome energy costs.

We even use the latter approach in the context of genetically modified organisms. When one puts a protein-coding sequence of one organism in a different organism, the translation speed landscape can be very different. Altering the DNA may restore this translation speed landscape to resemble the original landscape, but this altered sequence may then have a mechanical signal very different from the original signal. Therefore, we describe 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. This method is again powered by graphs and shortest-path algorithms.

In the final part of the dissertation we discuss how genetics and mechanics of actual organisms are multiplexed (a term that refers to having two or more layers of information on a single medium, in this case, DNA). We have shown that nucleosome signals are free to exist on top of protein-coding regions of genes, and can even coexist with translation speed signals. Genes, however, also contain noncoding parts, such as introns. Therefore, the mechanical signals may simply be encoded by noncoding parts of the genes. We introduce a classification scheme for different types of multiplexing to show which organisms encode mechanical information on top of protein-coding DNA, and which organisms use different strategies. The scheme can separate the contributions of exons (which are coding regions), introns and UTRs (which are both noncoding regions) to the nucleosome positioning signal. It also introduces the concept of intraregional signals, which occur on a single region, and interregional signals, which are signals that arise from inherent differences between regions.

We study the Transcription Start Sites (TSSs), sites known to have strong mechanical signals, of a wide range of organisms. Using our scheme we show that, for many organisms, such as fish and many plants, interregional signals dominate near the TSSs. For human and many other animals, the intron (noncoding) part of the intraregional signal dominates. For rice and other cereal grains we see that the exon (coding) part of the intraregional signal dominates. The positioning signal of rice, which is greater than that of human, shows that even coding sequences in real genomes may contain strong mechanical signals.

For rice we show that a large part of this nucleosome positioning signal can be attributed not to the choice of synonymous codons but to the choice of amino acid. We show that the amino acid sequence has a significant effect on the average mechanical landscape of rice. This makes us hypothesize that mechanical information is sometimes more important than preserving protein-coding information, and that these two types of information may even compete over the course of evolution.

We were also able to include translation speed to our analysis. Results from rice and human suggest a competition between mechanical information and translation speed signals. Summerizing, we suggest that genetics, mechanics and translation speed all three may compete with each other.

Samenvatting

DNA is de drager van verschillende vormen van informatie. De meest bekende van deze vormen is klassieke genetische informatie: genen op het DNA coderen voor eiwitten en RNAs. Deze genetische code bestaat uit reeksen van basen, afgekort A, T, C en G. Een basenreeks kan informatie bevatten die nodig is om een specifiek eitwit te bouwen. In deze dissertatie wordt een tweede informatielaag bestudeerd: de mechanica van DNA. Een reeks bestaande uit A's en T's zal anders buigen dan een reeks G's en T's. Dit komt doordat de intrinsieke vorm en flexibiliteit van DNA afhangen van de gekozen basen. Een belangrijk gevolg van deze mechanische informatielaag betreft de positionering van nucleosomen.

Nucleosomen bestaan uit 147 basenparen (bp) DNA gewikkeld om een eiwitkern, als een draad om een spoel. Nucleosomen zijn de fundamentele bouwblokken waaruit chromosomen bestaan. Ze zorgen ervoor dat het DNA compact is, en ze zijn verantwoordelijk voor de hogere-ordestructuren van DNA. Voor nucleosomen is ook een belangrijke taak weggelegd voor de regulatie van DNA. Ze kunnen fysiek voorkomen dat transcriptiefactoren (eiwitten die transcriptie reguleren) de transcriptiefactorbindingsplaatsen op het DNA bereiken, simpelweg voor zich op zo'n plaats te bevinden. Een cel kan de nucleosomen (of, om precies te zijn, de 'histonstaarten' van nucleosomen) scheikundig modificeren, waardoor een bindingsplaats weer beschikbaar wordt. Deze modificaties kunnen zelfs geërfd worden door de nakomelingen van een organisme. Dit is een voorbeeld van zogeheten epigenetica. Doordat een nucleosoom een bindingsplaats kan blokkeren of vrij kan maken, kan het dienen als een aan/uit-knop. De locatie van zulke knoppen is zeer belangrijk.

De locaties van nucleosomen op het DNA worden beïnvloed door de mechanische informatie op DNA. Het DNA op een nucleosoom moet op een specifieke manier gebogen zijn, een buiging die bij sommige DNA-reeksen makkelijk bereikt wordt, maar bij andere juist niet. Dit feit leidt tot een nucleosoompositiecode, vaak gerepresenteerd door de kans om een dinucleotidestap (een base gevolgd door een andere base) aan te treffen op verschillende locaties in het nucleosoom. Zo zijn bijvoorbeeld de kansen om dinucleotidestappen TT, AA of TA aan te treffen het hoogst daar waar de kleine groef van het DNA gericht is op de eiwitkern, terwijl de kans op een GC-stap het hoogst is op plekken waar de grote groef op de eiwitkern gericht is.

Er bestaan modellen die deze regels kunnen reproduceren maar ze niet kunnen verklaren. Vooral de regels over de GC-stap blijken tegenintuïtief te zijn, omdat de kansen om GC aan te treffen het hoogst zijn daar waar de buigingsenergie van GC maximaal is. Het eerste belangrijke resultaat in deze dissertatie is een fysische verklaring voor deze regels. In Hoofdstuk 2 wordt dit probleem aangepakt door gebruik te maken van een model dat de voornaamste nucleosoompositieregels reproduceert en ook analytisch op te lossen is. Gebruikmakende van de transfermatrixmethode

en een nieuwe benaderingsmethode slagen we er in om de mechanica achter de GC-regels te ontsluieren, alsook om methodisch de andere regels te verklaren.

Een andere aanpak is om niet deze dinucleotideregels te bestuderen, maar om te kijken naar hoeveel mogelijkheid er is om lange stukken DNA mechanische informatie te laten bevatten. In Hoofdstuk 3 pakken wij dit aan door de DNA-reeksen te bestuderen die óf het allerbest óf het allerslechtst zijn in het aantrekken van nucleosomen. We beschrijven alle mogelijke DNA-reeksen in een gewogen graaf. Voor de gewichten maken we gebruik van het probabilistische model van Tompitak et al. [10]. Door gebruik te maken van een kortstepadalgoritme vinden we de DNA-sequenties die de hoogste en laagste energiekosten hebben om onderdeel te zijn van een nucleosoom. Gebruikmakende van een k-kortstepadalgoritme vinden we k van de hoogste en laagste mogelijke energieën. De twee grote voordelen van deze methoden (ten opzichte van bijvoorbeeld Monte-Carlosimulaties) zijn dat kortstepadalgoritmes exact en snel zijn.

DNA in echte, levende organismen heeft niet de luxe om alleen mechanische signalen te bevatten. Echt DNA moet informatie bevatten voor RNA's en eiwitten. Theoretisch gezien kunnen deze twee informatielagen, genetisch en mechanisch, naast elkaar bestaan op hetzelfde stuk DNA. Een eiwit bestaat namelijk uit één of meer aminozuurketens. De eigenschappen van een eiwit hangen af van het type aminozuur en de volgorde van de aminozuren in de ketens. De aminozuren zijn geëncodeerd op het DNA in zogeheten codons (een drietal van basen). In veel gevallen zijn er synonieme codons, verschillende codons die coderen voor hetzelfde aminozuur. Dit wordt de ontaarding van de genetische code genoemd. Door deze ontaarding kunnen verschillende DNA-sequenties voor hetzelfde eiwit coderen én verschillende mechanische eigenschappen hebben. Daarom kan een eiwitcoderend stuk DNA ook mechanische informatie bevatten.

Gebruikmakende van de ontaarding van de genetische code hebben we onderzocht hoeveel vrijheid DNA heeft om beide vormen van informatie te bevatten. We hebben dit gedaan door grafen te gebruiken die alle synonieme manieren bevatten om te coderen voor hetzelfde eiwit. Een kortste pad in zo'n graaf geeft nu de DNA-sequenties met de hoogste en laagste energiekosten, terwijl het DNA nog steeds codeert voor het oorspronkelijke eiwit. Bovendien presenteren wij een heuristische methode die gunstige nucleosoomposities kan creëren bovenop een gen. Deze methode demonstreren we op het genoom van *Saccharomyces cerevisiae*, bakkersgist. Ook deze methode gebruikt graafrepresentaties van genen en korststepadalgoritmes. We onderzoeken *alle* posities op *alle* eiwitcoderende genen van gist, en slagen erin om nucleosoompositioneringssignalen te creëren met een precisie van één enkel basepaar voor $99,897\%$ van alle mogelijke nucleosoomposities.

Tot dusver hebben we twee verschillende informatielagen voorbij zien komen. Hierbij hebben we een *derde* informatielaag over het hoofd gezien: translatiesnelheid. Translatiesnelheid is de snelheid waarmee een eiwit gemaakt wordt. Deze snelheid heeft belangrijke gevolgen voor de resulterende eiwitten, aangezien het niet alleen beïnvloed hoe snel eiwitten geproduceerd kunnen worden, maar ook hoe de aminozuurketen zich vouwt tijdens translatie. De vouwing beïnvloed de functie en kwaliteit van een eiwit. De snelheid waarmee een aminozuur wordt toegevoegd aan de keten hangt af van de codon op het DNA (en ook van het celtype en het soort organisme). Ook al coderen synonieme codons voor hetzelfde aminozuur, toch maakt

de keuze van het codon uit voor het resulterende translatielandschap. Hierdoor kan het aanpassen van een gen, ook al blijft de genetische code intact, leiden tot niet-werkende eiwitten. Om deze reden voegen we in Hoofdstuk 4 de translatiesnelheid toe aan onze analyse. We presenteren verschillende aanpakken die translatiesnelheid incorporeren gebruikmakende van de graafrepresentaties die we in Hoofdstuk 3 geïntroduceerd hebben. Deze aanpakken gebruiken hetzij *pruning*, waarbij knooppunten uit een graaf worden weggesnoeid die zouden lijden tot een te verschillend translatiesnelheidlandschap, hetzij veranderen de gewichten in de graaf zodat die zowel translatiesnelheid als nucleosoomenergie bevatten.

We kunnen de laatste aanpak zelfs gebruiken in de context van genetisch gemodificeerde organismes. Wanneer een eiwitcoderend stuk DNA van het ene organisme in een ander organisme geplaatst wordt, kan het translatielandschap erg verschillend zijn. Het aanpassen van dit stuk DNA kan ervoor zorgen dat het translatiesnelheidlandschap enigszins hersteld wordt, maar hierdoor kan dit aangepaste DNA nu juist zijn mechanische signaal verloren zijn. Daarom beschrijven we in deze dissertatie een heuristische methode die laat zien hoe het DNA aangepast kan worden zodanig dat, in een ander organisme, de genetische informatie behouden blijft terwijl het mechanische landschap en het translatiesnelheidslandschap sterk lijken op de versies in het oorspronkelijke organisme. Ook deze methode is gestoeld op het gebruik van grafen en kortstepadalgoritmes.

In het laatste deel van de dissertatie bespreken we hoe genetica en mechanica in genomen van echte organismen zijn gemultiplexed (bij multiplexing is er sprake van meerdere signalen op hetzelfde medium, in dit geval: DNA). Eerder hebben we laten zien dat nucleosoomsignalen kunnen bestaan in eiwitcoderende regio's van genen, zelfs als translatiesnelheid meegenomen wordt. Het is echter belangrijk om aan te merken dat genen ook niet-coderende stukken DNA bevatten, zoals introns. Het is dus mogelijk dat de mechanische signalen op DNA simpelweg geëncodeerd worden door deze niet-coderende stukken. Om dit uit te zoeken introduceren we een klassificatiemethode voor verschillende typen multiplexing. Deze methode stelt ons in staat erachter te komen welke organismen mechanische informatie bevatten op eiwitcoderend DNA, en welke organismen andere strategieën gebruiken. De methode maakt onderscheid tussen de contributies van verschillende regio's aan het totale nucleosoompositioneringssignaal. Deze regio's zijn exonen (coderende regio's), intronen en UTRs (beide niet-coderende regio's). Ook introduceren we het concept van intraregionale signalen, die zich bevinden op één enkele regio, en interregionale signalen, de signalen die voortkomen uit de inherente verschillen tussen regio's.

We bestuderen de Transcription Start Sites (TSSs), plaatsen op het DNA waarvan bekend is dat ze sterke mechanische signalen bevatten, van een groot aantal organismen. Met onze klassificatiemethode laten we zien dat, voor veel organismen zoals vissen en veel planten, interregionale signalen dominant zijn in de buurt van een TSS. Voor mensen en veel andere dieren blijkt het introngedeelte (niet-coderend) van het intraregionale signaal te domineren. Bij rijst en andere granen overheest het exongedeelte (coderend) van het intraregionale signaal. Het mechanische signaal van rijst, wat groter is dan dat van mens, laat zien dat zelfs in echte genomen coderend DNA sterke mechanische signalen kan bevatten.

Voor rijst laten we zien dat een groot gedeelte van dit nucleosoompositioneringssignaal niet veroorzaakt wordt door synonieme codons maar door de aminozu-

urkeuze. We zijn erachter gekomen dat de aminozuursequentie een significant effect heeft op het gemiddelde mechanische landschap van rijst. Hierdoor vermoeden wij dat mechanische informatie in sommige gevallen belangrijker is dan eiwitcoderende informatie, en dat er zelfs evolutionaire competitie kan zijn tussen deze twee informatielagen.

We waren ook in de gelegenheid om translatiesnelheid aan deze analyse te verbinden. Resultaten van rijst en mens doen ons vermoeden dat er nog een competitie bestaat: tussen mechanische informatie en translatiesnelheid. Alles bij elkaar leidt dit tot de hypothese dat genetica, mechanica en translatiesnelheid een competitie met elkaar aangaan.

Curriculum Vitae

I was born on the 14th of July, 1993, in Leiderdorp, The Netherlands. In 2014, I obtained my BSc degree in physics at Leiden University. After that, I continued my education at Leiden and in 2016 I obtained an MSc degree in Theoretical Physics *cum laude*. Being fascinated by the idea of a secondary, mechanical layer of information on DNA, I worked on my Master's research project at the Theoretical Biophysics group of Helmut Schiessel. The name of the corresponding thesis was *Schemes for evaluating DNA mechanics and nucleosome positioning*. After obtaining my Master's degree, I continued working in the group of Helmut Schiessel. During my PhD project, I continued working on nucleosomes, focussing on understanding nucleosome signals and investigating their viability and occurence in nature. Three times I have been assistent at a Theoretical Physics MSc course: Theoretical Biophysics. In 2018, I have presented my work at a CECAM-Lorentz workshop in Lausanne. In 2019, I visited a summer school in Princeton: PiTP, Great Problems in Biology for Physicists.

List of publications/manuscripts

- M. Zuiddam, R. Everaers, and H. Schiessel. Physics behind the mechanical nucleosome positioning code. *Phys. Rev. E*, 96:052412, 2017.
- M. Zuiddam and H. Schiessel. Shortest paths through synonymous genomes. *Phys. Rev. E*, 99, 012422 2019.
- M. Zuiddam, B. Shakiba and H. Schiessel. Multiplexing mechanical and translational cues on genes. *Manuscript in preparation.*
- M. Zuiddam and H. Schiessel. How mechanical information is multiplexed on the transcribed regions of protein-coding genes. *Manuscript in preparation.*

Acknowledgements

First I would like to thank my supervisor, Helmut Schiessel. The freedom, support, criticism and encouragement to pursue any seemingly-promising avenue during my research are greatly appreciated.

Furthermore I would like to thank everyone who has worked at the Schiessel group in Leiden, especially Marco Tompitak, Lennart de Bruin and Bahareh Shakiba.

Next I would like to acknowledge everyone who supported me throughout the arduous task of tackling manuscripts and self-doubt, in particular Jasper van der Vaart and Mara Smeele.

Finally I want to thank my parents and my brothers for their endless support.