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

Zuiddam, M.

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

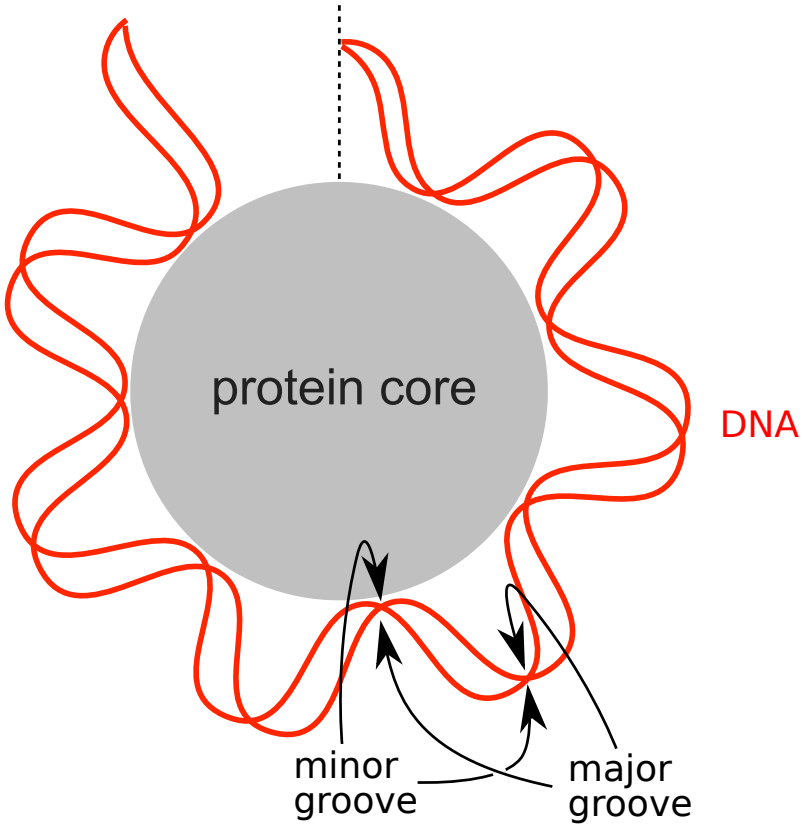

Figure 1.1: A nucleosome consists of 147 base pairs (bp) of DNA wrapped around a protein core. Here we show an exaggerated depiction of the major and minor grooves. These grooves face the protein core at 10 bp intervals, which leads to the nucleosome positioning rules. Sequences with a high affinity 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.