B 5 DNA and Chromatin

Helmut Schiessel
Institute Lorentz for Theoretical Physics,
Leiden University
Niels Bohrweg 2, 2333 CA Leiden, The Netherlands

Contents

Lecture Notes of the 49^{th} IFF Spring School "Physics of Life" (Forschungszentrum Jülich, 2018).

1 Introduction

DNA molecules are the carriers of genetic information for all life forms. DNA is typically found as a right-handed double helix with its two strands running antiparallel and its bases A, T, G, C forming pairs, A with T and G with C. Each stretch of DNA that encodes for a protein is called a gene. A protein is build from 20 different building blocks, called amino acids. With its much smaller alphabet DNA encodes for amino acids by grouping sets of three consecutive bases into information units, the codons. The precise rules how codons encode for amino acids is called the genetic code. As there are $4^3 = 64$ codons but only 20 amino acids the genetic code is degenerate. In other words, there are multiple ways to encode for one and the same amino acid (in 18 of the 20 cases). This degeneracy will be crucial in the following.
The point that these Lecture Notes want to make is that in addition to the genetic information (the genes encoding for the proteins) there is a second layer of information which is mechanical in nature. This is possible because the mechanical and geometrical properties of the DNA double helix depend on the underlying sequence of base-pairs. By choosing the "right" sequence of base-pairs a stretch of DNA can be made softer than average or stiffer than average. It is also possible to choose sequences that make the DNA molecule bent in a certain direction. The claim I want to make is that organisms have evolved their genomes to put mechanical cues along DNA molecules. Especially exciting is the fact – demonstrated below – that the classical genetic and the mechanical information can be multiplexed freely, allowing to put mechanical cues on top of genes at will, and not just on top of stretches of "junk" DNA. (We know multiplexing from daily life technologies, e.g. having two phone conversations on the same wire.)
What could be the meaning of such mechanical cues? I will argue that the cues guide the packaging of DNA molecules inside cells and by this indirectly the access to its genes. What I need to describe next is what we know about the packaging of DNA inside cells.

1.1 The hierarchical structure of chromatin

We focus here on eukaryotes (which include animals, plant and fungi). Cells of eukaryotes keep their DNA in a separate compartment, the nucleus. Eukaryotic DNA is packaged with the help of proteins into a DNA-protein complex called chromatin. Each individual DNA molecule together with the complexed proteins is called a chromosome (human somatic cells have 46 chromosomes). The structure of chromatin is hierarchical, see Fig. 1 adapted from Ref. [1], but many details of the different levels are not well understood. The first level of compaction is the wrapping of DNA molecules around protein cylinders, leading to DNA spools called nucleosomes. These are the main players of these Lecture Notes.
Just for completeness let me provide a short discussion of the higher levels. Traditionally the next level is believed to be the chromatin fiber, a rather compact structure into which the string of nucleosomes is folded. Fibers are easily seen in the test tube and lots of energy has been spent in figuring out their precise microscopic structure. But just as the debate between various competing detailed models of chromatin fibers raged at its fullest, some new experiments put serious doubts on the generally accepted believe that chromatin fibers exist in living cells [2]. Even though they are readily seen *in vitro* [3], also very recent work does not find them *in vivo* [4]. That is why I put a big question mark on top of my picture of the chromatin fiber in Fig. 1. Also the structures beyond that level are not well understood. But it is worthwhile to mention that there is currently tremendous progress in the understanding of the larger scales thanks to a new experimental method called chromosome conformation capture [5]. This leads

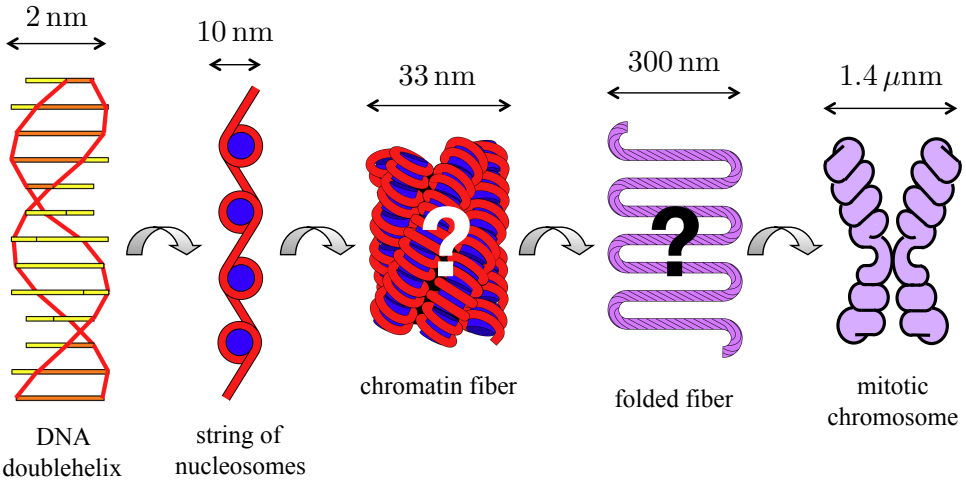

Fig. 1: *The hierarchical structure of a chromosome: the DNA double helix is wrapped around protein cylinders to form nucleosomes, the string of nucleosomes packs into a 30 nm wide chromatin fiber that folds into the chromosome. Here the well-known X-shaped mitotic chromosome is shown with its two identical copies of the DNA molecule which forms before cell division. Details (and not only details!) of the structures beyond the nucleosome are still a matter of debate.*

to new exciting developments, e.g. the idea of the loop extrusion mechanism [6]. This would have been an interesting and timely subject to speak about in a lecture. But the goal here is to discuss mechanical cues in DNA molecules. And such mechanical cues are most important at those places where DNA is bent most. This happens on the smallest compaction scale, the nucleosome, to which we now turn.

1.2 "A genomic code for nucleosome positioning"

The nucleosome is an ideal reader of mechanical information. The reason for this is twofold and can be best seen by inspecting the crystal structure of the nucleosome core particle [7]. The nucleosome core particle is the complex formed by DNA of exactly the nucleosomal wrapping length, 147 base-pairs, and the core of histone proteins, see Fig. 2(a). In the cell millions of such complexes are connected by a given DNA molecule into a string of nucleosomes connected by non-complexed stretches of so-called linker DNA, about 0 to 80 base-pairs in length. In each nucleosome 147 base-pairs, corresponding to about one DNA persistence length, are wrapped in one and three quarter turns around an octamer of histone proteins. This means that the energy of bending DNA into a nucleosome is large, about $60\,k_{\mathrm{B}}T$ [1]. Even small difference in the sequence will lead to large differences in the bending energy between those sequences. This is one reason why nucleosomes are ideal readers of mechanical information. The second reason is related to the way the DNA is bound to the histone octamer. It is bound at 14 locations where the two backbones touch the surface of the histone octamer (at the so-called minor groove of the double helix). As the sugar-phosphate backbones are independent of the underlying sequence, the pure binding energy is only weakly sequence dependent. Taken together, these two features make the nucleosome the master of the so-called indirect readout. Whereas most DNA binding proteins find their target by reading the sequence directly, the affinity of a 147 base-pair long stretch to be complexed in a nucleosome reflects the ease with which it is wrapped into it.

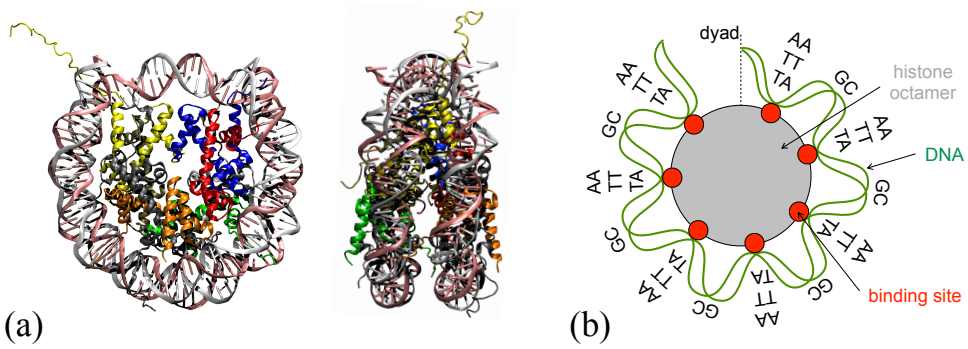

Fig. 2: *(a) Crystal structure of the nucleosome core particle (top and side view): 147 base-pairs are wrapped around an octamer of histone proteins [7]. (b) Widom's "genomic code for nucleosome positioning" [9]: key base-pair steps that increase the affinity of a sequence to the nucleosome are displayed relative to the structure of one-half of the nucleosome.*

It is indeed known since a long time that the affinity to nucleosomes varies with sequence. To learn which sequences have a strong affinity to nucleosomes, Travers and coworkers extracted in 1986 chromatin from chicken [8]. They added an enzyme, DNAase, that digested all the freely available DNA leaving just DNA intact that was very stably wrapped into nucleosomes. 177 of those intact sequences, about 150 base-pairs long, were sequenced. When looking at individual sequences it was not obvious why they had a higher affinity than average. But by looking at statistical properties of these sequences, i.e. averages over those sequences, certain pattern arose. It turned out that certain base-pairs steps are more likely to be found at certain positions along the nucleosome and less likely at other positions (one looks at base-pair steps because these are the objects that are deformed when wrapping the DNA around the octamer, as explained in detail further below). Looking along one of the strands in its so called 5' to 3' direction, the base-pair steps of importance are GC (a G followed by a C), AA, TT and TA. These four steps show characteristic oscillations in their occurrence frequency along the more stable nucleosomes, with a period of 10 base pairs, the DNA helical repeat, see Fig. 2(b). Specifically GC steps are found where the DNA's minor groove faces outward, whereas AA, TT and TA are all in phase with each other and peak at the positions in between, namely where the minor groove faces the histone octamer.
This suggests the possibility that genomic sequences have evolved over evolutionary time scales to position nucleosomes at certain positions and to encode for their stability or other physical properties. The late Jonathan Widom suggested that there is a "genomic code for nucleosome positioning" [9]. He said that "genomes care where nucleosomes are on average and so genomes code explicit information to bias [their positions]." After Jon's untimely death in 2011 I decided to follow up on his ideas using methods from statistical physics.

1.3 The space of all nucleosomal sequences

As a starting point to think about this complex problem, it is useful to introduce the concept of the sequence space of all DNA stretches that can be wrapped into a nucleosome. How many distinct sequences exist? There are four different bases and the length of the wrapped DNA is 147 base-pairs. This leads to 4^{147} different sequences, a huge number on the order 10^{88} (strictly speaking, due to a rotation symmetry of the nucleosome around the so-called dyad, see Fig. 2(b), this number needs to be divided by two). If you wanted to synthesise all this DNA,

you would need a big lab as all this DNA would fill five Milky Ways densely.
So how do scientists study this gigantic space? Well, strictly speaking they don't. This is because no matter if they are biologists, bioinformaticians or physicists, no matter whether they are experimentalists or theorists, what they tend to look at are typically genomes of certain organisms. Very popular is baker's yeast whose genome is about 12 million base-pairs long. This means that when one studies nucleosome positioning on that genome, one accesses only the 10^{-80}'s fraction of the nucleosomal sequence space. Even highly complex organisms like humans with their 3.2 billion base-pairs long genome can only scratch the surface of sequence space.
If we wanted, for instance, to answer the question: "which is the sequence with the highest affinity to reside in a nucleosome?" we would have no chance to find it by scanning the whole human genome (assuming that our genome has not evolved for highest nucleosome affinity which is a known fact). A slightly better approach is to start from a huge pool of random DNA molecules and then fish out of this pool the sequences with the highest affinity. This has been done in 1998 in the Widom lab. Starting from a huge pool of 5 trillion random DNA molecules (slightly longer than the nucleosomal wrapping length) the molecules were mixed with a much smaller number of histone proteins (namely one octamer per 10 DNA molecules) [10]. After the complexes had formed, the non-complexed DNA molecules were discarded and the winners were multiplied. This process was repeated 15 times. At the end of this so-called SELEX experiment there were only a few dozen types of DNA molecules left, all with a much higher affinity than average. The best of those sequences was called 601 (I do not know why, I wish I did) and it is nowadays the most common sequence used in the lab when working with nucleosomes.
You might wonder why one is left with just a handful sequences after starting with 5 trillion sequences. Why so few? The reason is that this experiment started from a random pool of sequences. If we order our sequence space such that the highest affinity sequences are in the center of the "Milky Way" and the lower affinity sequences toward the outskirts, then the starting sequences will fill randomly the whole space. There will then be only a small fraction of sequences close to the center and those few sequences are the only ones who have a chance to win the competition for the histone octamers. In short, one has a lot of waste DNA that needs to be discarded.
In these Lecture Notes I will introduce a different type of approach where practically all sequences to be "produced" will be automatically high affinity sequences. It is a computational approach that we call Mutation Monte Carlo (MMC) method [11]. It can be used in principle for any computer model of the nucleosome as long as it accounts for DNA sequence effects. Before I explain how MMC works I will introduce the model that we have tested and used [11].

1.4 A coarse-grained nucleosome model

Our nucleosome model is depicted in Fig. 3(a). It consists of a coarse-grained representation of the DNA molecule, 147 base-pairs long. The DNA molecules is forced into the configuration in which it is found in the nucleosome crystal structure, using 28 constraints that mimic the 14 binding sites in a real nucleosome. The protein core is not modelled explicitly, its presence is only accounted for by the constraints on the DNA.
The DNA double helix is represented by the so-called rigid base-pair model [12]. This model accounts only for the base-pairs of the DNA molecule that are modelled as rigid blocks. This leaves six degrees of freedom between neighbouring base-pairs, called shift, slide, rise, tilt,

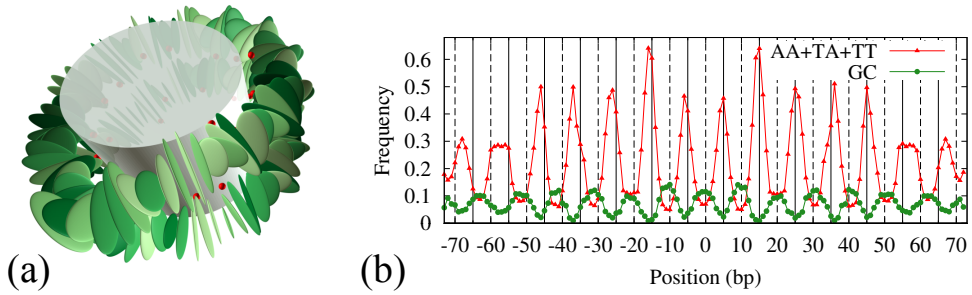

Fig. 3: *(a) The coarse-grained nucleosome model from Ref. [11]. Same colour scheme as in Fig. 2(b). (b) Base-pair step probability distributions (average over 10 million sequences) produced by MMC at one third of room temperature using the model from (a) [11]. The model reproduces the standard nucleosome sequence preferences, Fig. 2(b).*

roll, twist, see middle of Fig. 4. If you want to create an ordinary straight piece of DNA double helix, all you need are two degrees of freedom, rise and tilt. A rise of 0.34 nm combined with a twist of about 36 degrees leads to a twisted stack of base-pairs that looks similar to the real DNA double helix (in its common B-form), see Fig. 4 left. Note that because the two sugar-phosphate backbones are attached to one long site of the base-pairs, one has two grooves, a major and a minor one, going around the DNA double helix. When one looks at space-filling figures of the DNA double helix, one can clearly identify these two types of grooves. Also note that one can see clearly in such figures the parallel twisted stack of base-pairs through the gap created by the major groove. It is through this gap that DNA binding proteins typically "read" the DNA sequence by reaching inside that groove. This is how direct readout works. As mentioned above, this is *not* how the sequence preferences of nucleosomes come about.

In order to bend the DNA around the nucleosome, other degrees of freedom have to be invoked. We will mention here only the most important one, roll. Roll is the rotation around the long axis of the base-pair step. It is convention to call the roll positive if the base-pair stack is compressed towards the major groove. By periodically changing the roll from positive to negative and back with the DNA helical repeat (about 10 base-pairs) the twisted stack of base-pairs bends in one direction, see Fig. 4 right. This allows to rephrase the nucleosome positioning rules: high affinity sequences feature GC steps at positive roll positions, and AA, TT and TA steps at negative roll positions. Where do these sequence preferences come from?

In order to make any prediction one needs to go beyond a purely geometrical model by introducing also energy into the system. In fact, the rigid base-pair model has been fully parametrized in the literature. One assumes only nearest-neighbor interactions with a quadratic deformation energy between successive base-pairs [12]:

$$E = \frac{1}{2}(q - q_0) \cdot K \cdot (q - q_0). \tag{1}$$

Here q is a six-component vector that describes the relative degrees of freedom between two base-pairs. The intrinsic, preferred values of these degrees of freedom are given by q_0. The properties of the (six-dimensional) springs connecting the base-pairs are given by K, a six-by-six stiffness matrix. The sequence-dependence of the model comes into play because the stiffness (K) and intrinsic shape (q_0) of a given base-pair step depend on its chemical identity. In other words K and q_0 are different for different types of base-pair steps.

These parameters have been determined in the literature, either by looking at the conformations and fluctuations of DNA-protein cocrystal structures [12] or by performing all-atom molecular

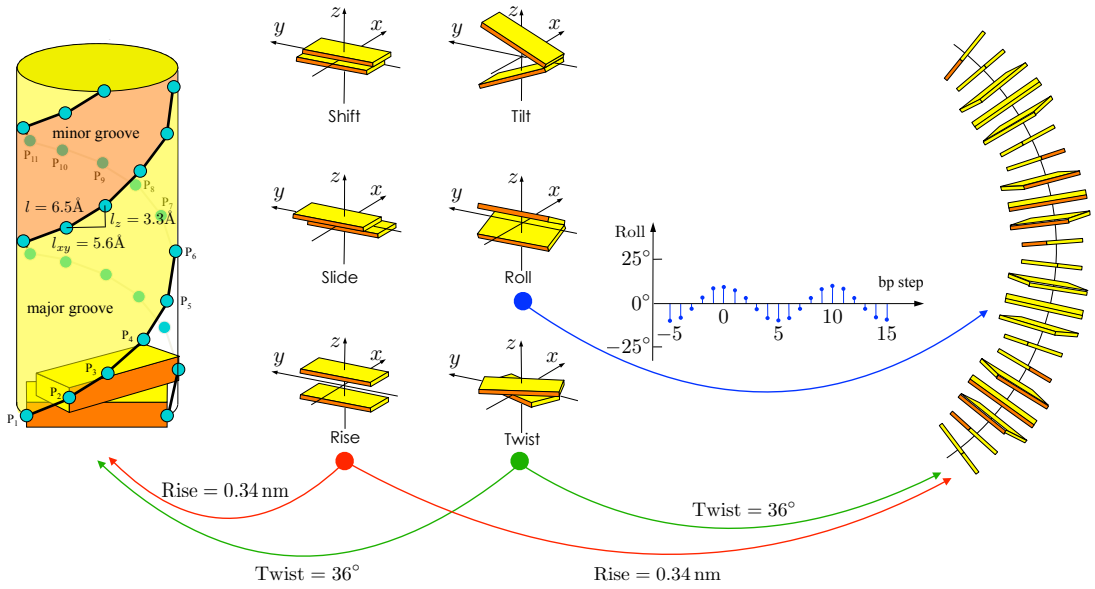

Fig. 4: *The rigid base-pair model is a coarse-grained DNA representation that leaves six degrees of freedom per base-pair step (middle). A base-pair step with 0.34 nm rise and about 36 degrees twist produces a straight standard DNA double helix (left). When in addition the roll is changed periodically with the DNA's helical repeat one obtains a bent stack (right).*

dynamics simulations of short DNA molecules with various sequences [13]. We are learning currently which of the parameter sets works best, often we use a hybrid version that uses both sources [14].

In summary, our nucleosome model consists of 147 base-pairs of DNA represented by the rigid base-pair model wrapped into a superhelix that mimics its configuration in the nucleosome crystal structure. This is achieved via 28 rigid constraints, two per binding site. One last additional detail: each rigid constraint consists of a fixed mid-plane for a consecutive base-pair step (corresponding to the bound phosphate of the involved backbone).

1.5 The Mutation Monte Carlo (MMC) method

Having now a nucleosome model with sequence dependent energetics we can introduce the MMC method. This method allows to scan regions in the nucleosomal sequence space that are special with respect to their elastic properties. But first let us discuss how a nucleosome with a given fixed sequence can be studied using a standard Monte Carlo scheme. What we would like to achieve is to sample the configurational space of the nucleosome according to the Boltzmann distribution. This is achieved as follows. Pick a random base pair. Perform a small rotation around a random axis together with a small translational shift in a random direction. According to Eq. 1 this changes the mechanical energies of the two involved base-pair steps by a (small) amount ΔE. If $\Delta E < 0$ accept the move. If $\Delta E > 0$ accept it only with a probability $e^{-\beta \Delta E / k_{\mathrm{B}} T}$. Continuing this process one obtains an equilibrium distribution of the nucleosomal DNA configurations from which one can determine e.g. the average elastic energy. (One technicality: whenever the chosen base-pair happens to be next to a rigid constraint, move the base-pair across the fixed mid-plane symmetrically to keep it fixed).

So far we are stuck at one point in sequence space. How can we explore that space? The trick is

extremely simple and very effective. The MMC method developed by Eslami-Mossallam [11] uses in addition to the conformational moves also mutation moves. A mutation move consists of randomly picking a base-pair and attempting to change its chemical identity. This affects again, as for the conformational moves mentioned above, the mechanical energy, Eq. 1, of the two involved base-pair steps. But instead of changing q, this changes K and q_0 of the corresponding steps. The move is accepted or rejected according to the energy change using the same rules as above.
By randomly mixing conformational and mutational moves, the system moves through sequence space and quickly arrives at nucleosome sequences (and corresponding conformations) that are much cheaper than average. One can then easily create 10 million independent high affinity sequences, rather than just a few as it is the case in experiments [10].
What is the role of temperature in such a simulation? MMC produces a set of conformations and sequences distributed according to the Boltzmann distribution at the chosen temperature. The lower the temperature the smaller is the section in sequence space that is explored focusing on sequences with higher and higher affinities. The temperature can thus be seen as a tool that allows to adjust the volume in sequence space that will be probed. By cooling the system close to zero temperature, it is even possible to identify the ground state sequence of our model nucleosome.

2 The mechanical genome

We can now ask the question: Is there – in addition to the classical genome (the genes that encode for the proteins) – a "mechanical genome", i.e. a set of mechanical cues written along DNA molecules that have formed over evolutionary time scales in parallel and independent of the classical genome? To answer this question we need at least to show three things: that the nucleosome positioning rules are mechanical in nature, that the mechanical cues can be multiplexed with the classical genetic information and that such mechanical cues do actually exist on real genomes. The next three sections demonstrate these three fundamental aspects of the mechanical genomic code using the tools introduced above.

2.1 DNA mechanics dictates nucleosome positioning rules

We have described earlier the nucleosome positioning rules, see also Fig.2(b). Assuming that the rules are caused by DNA mechanics and that the nucleosome model we introduced above, Fig. 3(a), is realistic enough to make reasonable predictions (as various tests suggest [11, 15, 16]), we need to show that sequences which follow these rules more than average sequences have indeed a higher affinity than average. The MMC approach allows to answer this question in a straightforward way. All what needs to be done is to run such a simulation on the model nucleosome and then to check whether the produced sequences follow on average the rules. We produced 10 million independent high-affinity sequences by performing a MMC simulation at 1/3 of room temperature. We then looked at base-pair step distributions obtained by averaging over all those sequences. Figure 3(b) displays the distribution for GC steps and the combined distribution for AA, TT and TA steps. This procedure indeed recovers the standard nucleosome positioning rules, Fig. 2(b). This suggests that these well-known rules are caused to a large extent by the sequence dependent elasticity and geometry of the DNA double helix.
It is worthwhile to mention that these rules are not straightforward to understand. For instance,

the model predicts that GC steps peak at positive roll position. When one inspects the underlying geometrical preference of GC steps and compares it to all other steps, one learns that it is the step with the lowest value of roll. This together with the fact that it is one of the stiffest steps shows that GC is the step most "unhappy" to occupy large roll positions. So why does it peak at these positions against its own preferences? The reason is that each base-pair step is part of a larger sequence. When a GC step occupies a given position, the previous step has to end on a G and the following step starts with a C. As it happens, these neighbouring steps feature on average a low elastic energy if GC sits at large positive roll position. So it is the neighbours of GC but not GC itself that cause the peak of GC at high roll positions.

2.2 Genetic and mechanical information can be multiplexed

The next question to be considered is whether mechanical cues can be written freely on top of genes. To demonstrate this we start by looking at some randomly picked gene from a standard model organism, baker's yeast. Figure 5(a) shows a 500 base-pairs long stretch of gene YAL002W. It depicts the energy landscape that the nucleosome experiences as it is moved along that stretch of DNA. This has been calculated by a simple Monte Carlo simulation (without mutations). Note the strong undulations with a period of about 10 base-pairs. These are caused by the fact that – as one moves the DNA molecule through the nucleosome – it has to perform a corkscrew motion such that the DNA minor groove is always in contact with the binding sites. Typically a given DNA molecule has locally a preferred bending direction, just by accident. So about every 10 base-pairs along the sequence there is typically a minimum in the energy landscape, five base-pairs further a maximum and so on. This leads to the so-called rotational positioning of nucleosomes. This positioning might be important as it guides the higher order arrangement of nucleosomes. Another type of positioning, translational positioning, will be discussed in the next section. Note the vertical lines in this plot; these correspond to nucleosomes that have been mapped via a chemical method in yeast *in vivo* [17]. Most of the mapped nucleosomes fall in minima of our energy landscape, all along the yeast genome. This shows again that this model predicts properly the nucleosome positioning rules.

In the following we demonstrate that it is possible to change this rotational positioning at will without affecting the protein that the gene encodes for. How is this possible? As mentioned in the introduction 64 codons encode for only 20 amino acids. This degeneracy of the genetic code can be employed to change the mechanical properties of the DNA molecule keeping the encoded protein unchanged. Figure 5(b) displays a short stretch of the YAL002W gene (top row). The sequence is already broken into codons. Below that sequence of codons you find in red the sequence of amino acids that the gene encodes for. Below each amino acid there is a list of all the synonymous codons that represent this specific amino acid. In 18 of 20 cases there is in fact more than one codon available.

We make use of this degeneracy of the genetic code in a modified version of the MMC method. We now allow only synonymous mutations, i.e. we swap between codons that form a synonymous set. This way we can change the mechanical properties of the DNA molecule without affecting the sequence of amino acids that the base-pair sequence encodes for. More specifically, we focus on the well-positioned nucleosome on base-pair position 826 in Fig. 5(a). This nucleosome has also been mapped *in vivo* at precisely that position [17]. We would like to demonstrate that one can shift this local energy minimum to any position one likes, e.g. base-pair position 827 and so on, by synonymous mutations.

To achieve this we perform synonymous MMC simulations for the nucleosome placed on the

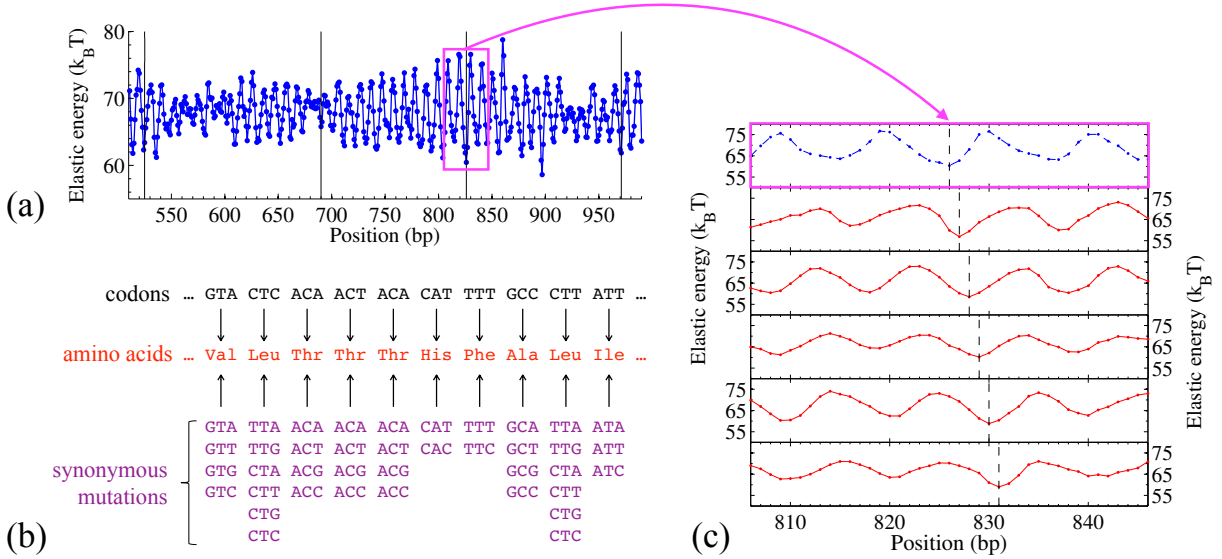

Fig. 5: *(a) Energy landscape for a nucleosome on a stretch of gene YAL002W from baker's yeast calculated using the model from Fig. 3(a). The vertical lines correspond to nucleosomes mapped in vivo [17]. (b) Top row: a stretch of the same gene as in (a) broken into a sequence of codons. Middle row (red): the sequence of encoded amino acids. Bottom (purple): list of synonymous codons for each given amino acid. (c) Synonymous energy landscapes below the original landscape (inside magenta box). This plot shows that a local minimum can be placed anywhere on that stretch of gene from base-pair 826 to 831 [11].*

positions where we want to create new local minima. Figure 5(c) demonstrates that this method works by displaying energy landscapes obtained by this method for the positions 827 to 831 [11]. Using a more sophisticated method we have in the meantime extended this calculation genome wide and were able to show that nucleosomes can be rotationally positioned anywhere on the yeast genome in at least 99.95% of the cases. This is ongoing work and we are not yet sure whether the remaining 0.05% correspond really to locations where one cannot position a nucleosome at all or whether we have to improve our method further.

2.3 Evidence for a mechanical evolution of DNA molecules

So far we have shown that in principle mechanical cues could have been written into DNA molecules and that even on top of genes. But the question remains whether this has really happened on actual genomes. And if the answer is yes: what are then the biological functions of such cues? These questions are not straightforward to answer. For instance, when you look again at Fig. 5(a) you can see a wildly oscillating energy landscape that a nucleosome would experience as one pushes it along the DNA molecule. But does this landscape constitute some kind of meaningful signal? As a matter of fact, the landscape of a completely random base-pair sequence looks pretty much the same.

To isolate meaningful signals out of genomes it turns out to be crucial to look at genome wide averages. Only then one beats the (possibly random) oscillations and starts to find in fact some very strong mechanical cues along DNA molecules. However, our nucleosome model is too slow to calculate genome-wide energy landscapes as one would need to perform a Monte Carlo simulation at each position along the genome, in order to allow the wrapped DNA stretch to

sample equilibrium configurations. Surprisingly the MMC method comes at our rescue also for this seemingly unrelated problem. The idea is to perform one long MMC simulation to learn about the sequence preferences of the nucleosome and then to use these sequence preferences as input in a simplified probabilistic model [18].

We have already determined these sequence preferences in the form of base-pair step probabilities along nucleosomes, see Fig. 3(b) for some examples. Specifically, we can use the MMC approach to learn about the probability to have base S_i at position i along the nucleosome with $i = 1, ..., 147$ and $S_i = A, T, G, C$ and the joint probability $P(S_i \cap S_{i-1})$ to have base S_i following base S_{i-1}. From these probabilities we can calculate conditional probabilities $P(S_i|S_{i-1}) = P(S_i \cap S_{i-1})/P(S_{i-1})$. We then estimate that the probability of the 147 base-pair long sequence S to be occupied by a nucleosome is given by

$$P(S) = P(S_1)P(S_2|S_1)\prod_{i=3}^{147} P(S_i|S_{i-1}). \tag{2}$$

This equation assumes that there are no longer-ranged effects along the DNA molecule, i.e. the probability of a given base to appear in a nucleosome only depends on the previous base but not on the precise nature of bases further away. That this is a reasonable approximation can be rigorously tested by performing an improved analysis starting from the probability distributions for triplets of bases which only slightly improve the predictions [18].

Which predictions do we actually speak about? One can estimate the energy of a sequence from the probability by taking the logarithm: $E(S) = -k_{\mathrm{B}}T \ln P(S)$. This way one can calculate the energy landscape of e.g. the YAL002W gene from above and compare it to the actual energy landscape as calculated from the full model. The deviations between the "real" energy landscape and the probabilistic one (based on duplets or triplets of bases) is on the order of one $k_{\mathrm{B}}T$, much smaller than the typical energy undulations in the energy landscape, see Fig. 5(a). So we make only a small error but what do we gain from it? It turns out that the speed-up using the probabilistic model is of the order of 10^5. This means that we can now perform genome wide calculations.

What needs to be done to obtain clean signals is somehow to average over the genome. A well-known way to do this is to align the same type of functional sites from all over the genome. The most promising candidate to look at is the beginning of genes as these are the places where a cell decides whether a gene is read out or not. Fig. 6(a) shows gene start sites of baker's yeast averaged over all its genes (about 6000). More specifically we show a 2000 base-pair long interval with the genes starting in the middle and going toward the right. The quantity depicted is the so-called nucleosome occupancy which is the probability that a given base-pair is covered by a nucleosome. This quantity is chosen as it is experimentally accessible. We assume that there is one nucleosome and calculate its occupancy for this 2000 base-pair wide window. From our calculation (based on Eq. 2) we produce the blue curve [16] which fits astonishingly well with the experimentally determined occupancy (green curve). In the experiment [19] nucleosomes are reconstituted on yeast DNA and their positions are determined by digesting the DNA with DNAase. The excellent agreement between data and model is certainly only a fortunate coincidence; what is important here is that both approaches give qualitatively the same overall signal.

It can be clearly seen that there is a strong depletion of nucleosomes just in front of the genes. This depletion signal in yeast has been speculated to be "partially encoded in the genome's intrinsic nucleosome organisation, and that this intrinsic organisation may facilitate transcription

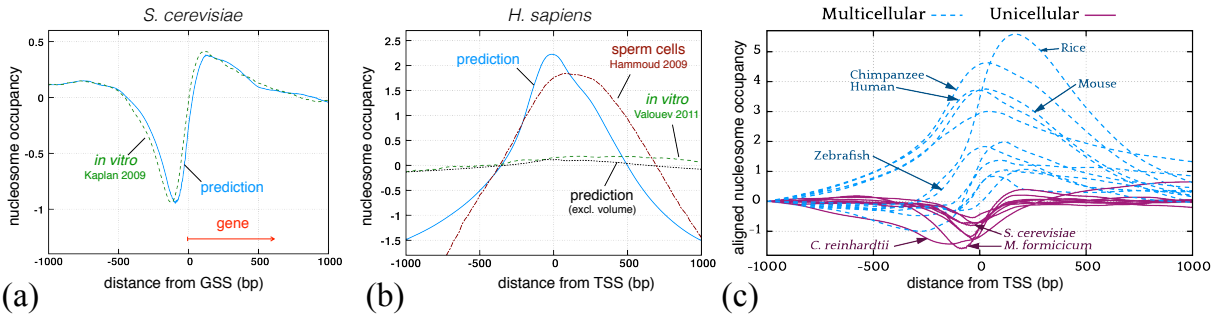

Fig. 6: *Nucleosome occupancies around the beginning of genes in various organisms. All plots are averages over all genes, either aligned at the gene start sites, (a), or the transcription start sites, (b) and (c). (a) Baker's yeast (in vitro data [19] and prediction), (b) humans (in vitro data [20], retained nucleosomes in sperm cells [23] and prediction) and (c) prediction for various unicellular and multicellular organisms. All predictions are based on Eq. 2 [16].*

initiation and assist in directing transcription factors to their appropriate sites in the genome" [19]. In short, DNA is stiffer than average before genes to keep its DNA free of nucleosomes so that the transcription machinery can always access that region if it wants to produce the corresponding protein.

This is what has been found for yeast. What about other organisms? Figure 6(b) shows the nucleosome occupancy averaged over all genes and aligned at the transcription start site (which is close to the gene start site) for the human genome. Surprisingly our model (blue curve in Figure 6(b)) shows a completely different signal, featuring a large and wide peak in the nucleosome occupancy [16]. How does this compare to experiments? At first not very favourably. The green curve in Fig. 6(b) are *in vitro* data showing a much smaller peak [20]. However, there is a profound difference between the experiment and the calculation. In the calculation we consider the probability distribution of only *one* nucleosome in the 2000 base-pair wide window. In the experiment there is a rather large nucleosome density. Since nucleosomes cannot sterically overlap, there is a saturation in the density around the peak. In fact, accounting in our calculations for a similar density as in the experiment, preventing steric overlap, we find a curve (dotted in Fig. 6(b)) similar to the experimental curve. But even if the nucleosome density cannot increase much around the transcription start sites, the signal is still contained in the affinity and thus stability of the corresponding nucleosomes.

What could be the biological function of these mechanical cues in the human genome? The following speculation [21] is based on the fact that humans – unlike yeast – are multicellular organisms: "[...] high nucleosome preference is directly encoded at regulatory sequences in the human genome to restrict access to regulatory information that will ultimately be utilised in only a subset of differentiated cells." So the idea is that many genes are only meant for specialised cells and that those genes should be closed off in all other types of cells. And this is achieved by encoding for stable nucleosomes around the start sites of those genes.

An exciting question to ask is whether this distinction between yeast and human is an example of a general rule in biology. This is hard to answer on the basis of experiments as there are not so many nucleosome maps available and, even if they were, it is not so easy to detect a signal because of the density saturation due to the excluded volume between nucleosomes. Using our model we looked at 50 different organisms and calculated the occupancy signal (for a single nucleosome) around all transcription start sites [16]. As you can see in Fig. 6(c), it is

indeed generally true that the DNA elasticity around transcription start sites is entirely different between unicellular organisms like baker's yeast or the green alga *Chlamydomonas reinhardtii* and multicellular lifeforms like zebrafish, mouse, human, chimpanzee and rice.
Is this the end of the story? Not quite. At least for humans the above given biological speculation turns out to be wrong. Dividing genes between house keeping genes and tissue specific genes and looking at the mechanical cues separately, one discovers the opposite of what one would have expected: the strong mechanical cues stem from the house keeping genes that all cell types need [22]. Even worse, when looking at actual *in vivo* nucleosome occupancies it was found that they do not reflect at all underlying DNA mechanics. Instead the transcription start sites of house keeping genes are typically depleted of nucleosomes [22]. This is, of course, expected, but it goes against the mechanical cues.
What is happening here? Apparently other processes, the binding of transcription factors to their specific target sites, the transcription by RNA polymerase and/or the action of chromatin remodellers (motor proteins that push and pull nucleosomes using ATP) overrule the mechanical cues around transcription start sites. The mechanical cues must therefore have a different function. The most logical explanation would be that they are of importance in a cell type that is transcriptionally not active.
Are there such cells in multicellular organisms? In fact, each animal no matter how big it is (think of an elephant!) needs eventually make itself very small when passing through the germ line into the next generation. Especially in sperm cells elephants shrink substantially (even smaller than the sperm cells of fruit flies or mice!). Small sperm cells are good swimmers and can be produced in larger numbers, a fact especially important for species where there is a competition between different males. That might be the reason why in sperm cells DNA is tightly packed with the help of protamines and all nucleosomes are evicted. But not quite: a recent finding shows that about 4% of the nucleosomes are retained in human sperm cells [23].
How does a sperm cell know which nucleosomes to keep? As we found out, it is the mechanical cues in the DNA molecules that determine which nucleosomes are retained: Regions where our model predicts the most stable nucleosomes correspond to regions where sperm cells retain nucleosomes, see Fig. 6(b) (brown curve) [16].
What is the evolutionary driving force for retaining a fraction of nucleosomes in sperm cells instead of getting rid of all of them? We can only speculate. But a likely reason is to allow for the transmission of epigenetic information via the father to the offspring (and not only by the mother where the nucleosomes are kept in the egg cell). Epigenetics is information in addition to and shorter-lived than genetic information. It is scribbled along the margins of the book of life. One way this can be achieved is by chemically modifying the histone proteins that form the octamer of the nucleosomes. This changes e.g. their stickiness affecting the accessibility to the associated DNA. And it is the genes that are important for the early embryonic development that are singled out for receiving this extra information; these carry the mechanical cues and thus retain the nucleosomes. A concrete (though controversial) experiment trained male mice using mild foot shocks to fear cherry blossom smell. Their offsprings had an aversion to this specific odour [24].

3 Conclusions

We have come a long way from the mechanics of base-pairs to the smell of cherry blossoms. The point that these Lecture Notes wanted to make is that if there are some degrees of free-

dom (here the DNA elasticity) that evolution can play with, it very likely makes use of it. It is, however, far from obvious what comes out of such an evolution. For example, so far the main interest in the field is to learn which nucleosomes are positioned by mechanical cues. This is done by assigning one number, the affinity of the sequence, to a given 147 base-pair long sequence. This does, however, overlook the fact that this is a much richer problem. In principle, the mechanical properties of 147 base-pairs wrapped into a nucleosome could give some nucleosomes distinct sets of physical properties, setting them far apart from standard nucleosomes. For instance, it has been shown that nucleosomes which are strongly asymmetric with respect to their two DNA halves act as polar barriers for transcribing RNA polymerases [25]. Using our model together with the MMC approach we have started to build designer nucleosomes which e.g. show an entirely different response to external forces than standard nucleosomes and might be used as "force sensors" [26]. An exciting question to ask is whether and where such special nucleosomes have evolved on real genomes and to what purpose.

References

[1] H. Schiessel, *Biophysics for Beginners: a Journey through the Cell Nucleus* (Pan Stanford, Singapore, 2014)

[2] K. Maeshima, S. Hihara, M. Eltsov, Curr. Op. Cell Biol. **22**, 291 (2010)

[3] P.J.J. Robinson, L. Fairall, V.A.T. Huynh, D. Rhodes, Proc. Natl. Acad. Sci. U. S. A. **103**, 6506 (2006)

[4] H.D. Ou *et al.*, Science **357**, 370 (2017)

[5] E. Lieberman-Aiden *et al.*, Science **326**, 289 (2009)

[6] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, Cell Reports **15**, 2038 (2016)

[7] K. Luger, A.W. Mäder, R.K. Richmond, D.F. Sargent, T.J. Richmond, Nature **389**, 251 (2016)

[8] S.C. Satchwell, H.R. Drew, A.A. Travers, J. Mol. Biol. **191**, 659 (1986)

[9] E. Segal *et al.*, Nature **442**, 772 (2006)

[10] P.T. Lowary, J. Widom, J. Mol. Biol. **276**, 19 (1998)

[11] B. Eslami-Mossallam, R.D. Schram, M. Tompitak, J. van Noort, H. Schiessel, PLoS ONE **11**, e0156905 (2016)

[12] W. K. Olson, A.A. Gorin, X.-J. Lu, L. M. Hock, V.B. Zhurkin, Proc. Natl. Acad. Sci. U. S. A. **95**, 11163 (1998)

[13] F. Lankaš, J. Šponer, J. Langowski, T.E. Cheatham III, Biophys. J. **85**, 2872 (2003)

[14] N.B. Becker, L. Wolff, R. Everaers, Nucl. Acids. Res. **34**, 5638 (2006)

[15] L. de Bruin, M. Tompitak, B. Eslami-Mossallam, H. Schiessel, J. Phys. Chem. B **120**, 5855 (2016)

[16] M. Tompitak, C. Vaillant, H. Schiessel, Biophys. J. **112**, 505 (2017)

[17] K. Brogaard, L. Xi, J.-P. Wang, J. Widom, Nature **486**, 496 (2012)

[18] M. Tompitak, G.T. Barkema, H. Schiessel, BMC Bioinformatics **18**, 157 (2017)

[19] N. Kaplan *et al.*, Nature **458**, 363 (2009)

[20] A. Valouev *et al.*, Nature **474**, 516 (2011)

[21] D. Tillo *et al.*, PLoS ONE **5**, e9129 (2010)

[22] T. Vavouri, B. Lehner, PLoS Genetics **7**, e1002036 (2011)

[23] S.S. Hammoud *et al.*, Nature **460**, 473 (2009)

[24] B.G. Dias, K.J. Ressler, Nature Neuroscience **17**, 89 (2014)

[25] V.A. Bondarenko *et al.*, Mol. Cell **24**, 469 (2006)

[26] M. Tompitak, L. de Bruin, B. Eslami-Mossallam, H. Schiessel, Phys. Rev. E **95**, 052402 (2017)