Universiteit
Leiden
The Netherlands

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

Zuiddam, M.

Citation
Zuiddam, M. (2022, April 6). *Freedom of additional signals on genes: on the combination of DNA mechanics, genetics and translation speed*. *Casimir PhD Series*. Retrieved from https://hdl.handle.net/1887/3281818

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3281818

Note: To cite this publication please use the final published version (if applicable).

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

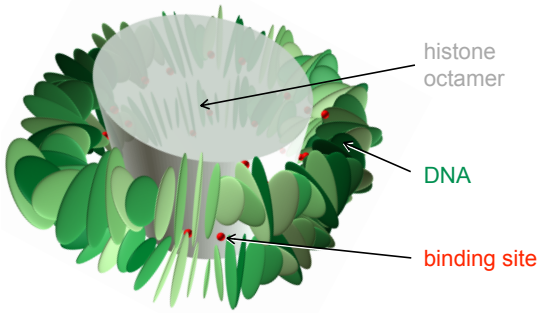

Figure 3.1: Nucleosome model [14] used to construct the probabilistic trinucleotide model [10]. Each rigid plate represents a bp, the locations of the constraints (corresponding to bound phosphates) are shown by beads, two per binding site. The cylinder is a rough representation of the protein core but 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

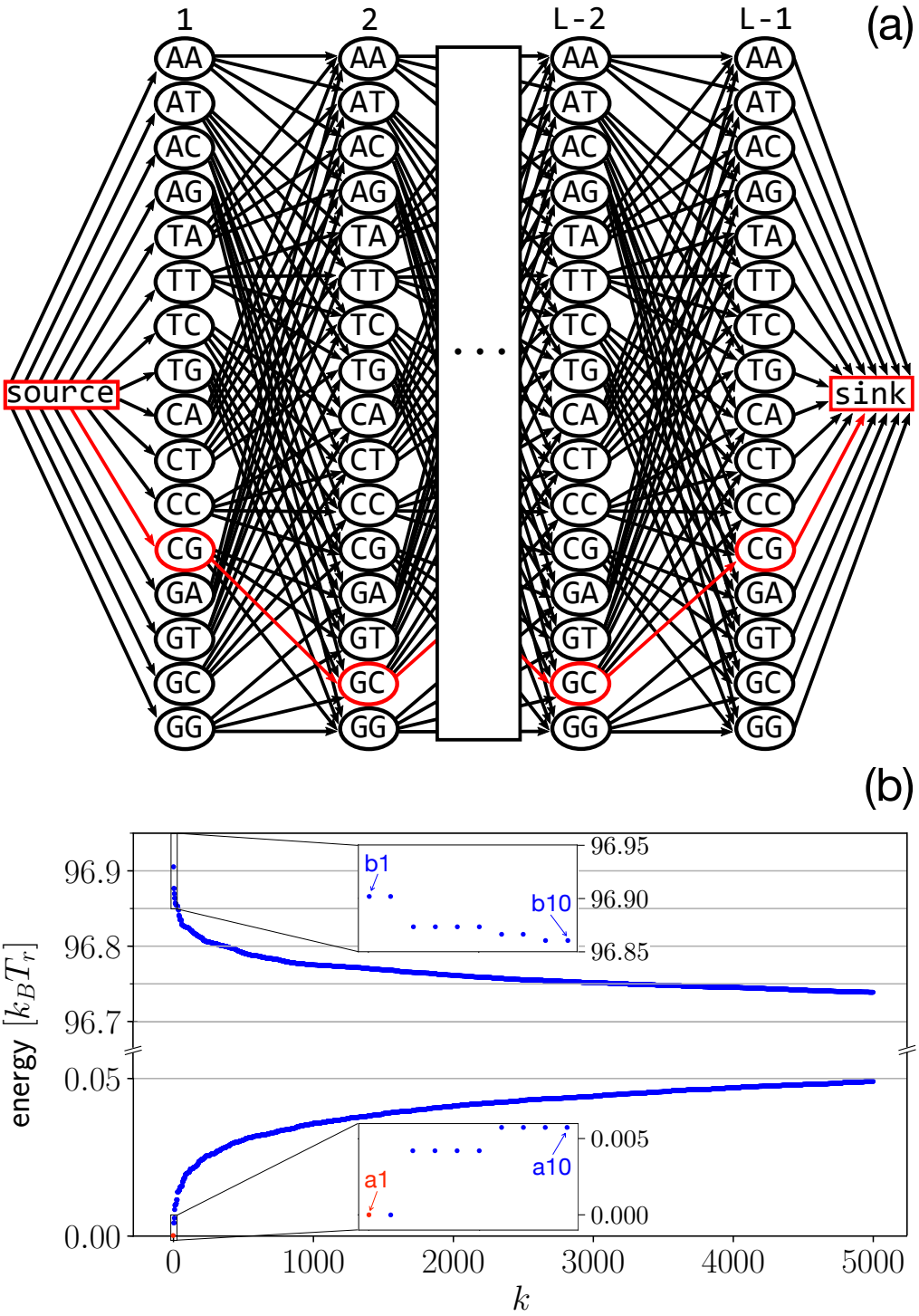

Figure 3.2: (a) Graph representation of the 4^{147} DNA sequences that can be wrapped into a nucleosome. Weights are assigned such that each path from *source* to *sink* has a length equal to the total energy of the corresponding sequence. The path in red (grey in greyscale) corresponds to the ground 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.

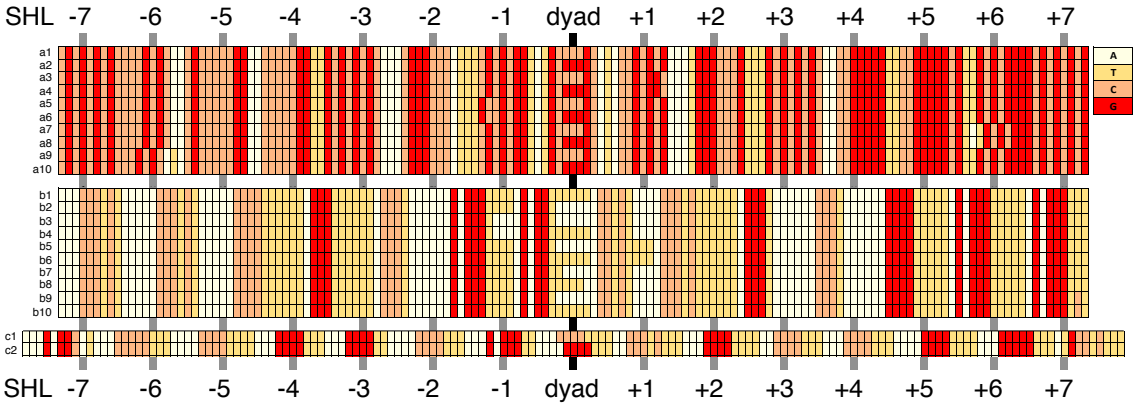

Figure 3.3: a1-a10: ten lowest energy sequences, b1-b10: ten highest energy sequences, c1/c2: best positioned nucleosomes. Because of symmetry, sequences a1 and a2 have the same energy and so on. At integer superhelical locations (SHL) the major groove of the DNA bends towards the histone octamer. SHL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

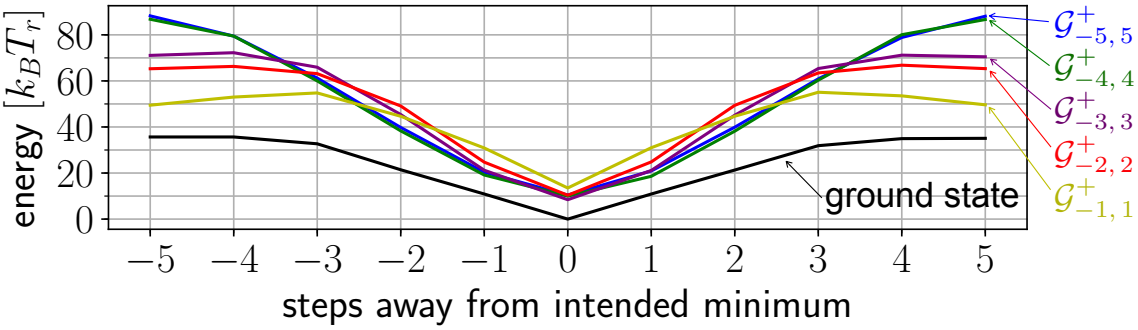

Figure 3.4: The energy landscapes as a result of a shortest path through graphs $\mathcal{G}^+_{-i,i}$, $i = 1, 2, ..., 5$, are shown, as well as the landscape of ground state sequence a1.

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

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

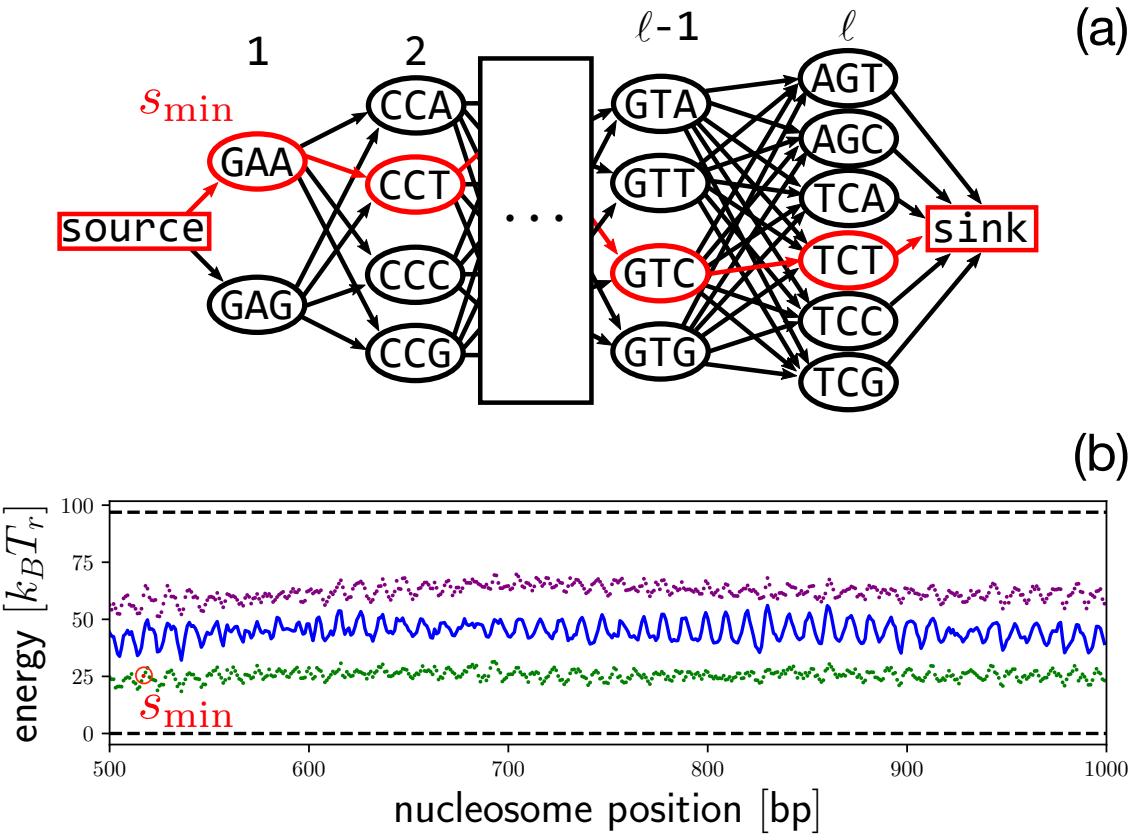

Figure 3.5: (a) Graph $\mathcal{G}_{\text{gene}}$ shows all synonymous ways to encode for a given amino acid sequence (ℓ is either 49 or 50). The shortest energy path $s_{\min}$ is highlighted. (b) Energy landscape of a 500 bp stretch of gene YAL002W from yeast (solid curve), pointwise minimal and maximal 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 $\mathcal{D} > 30$.

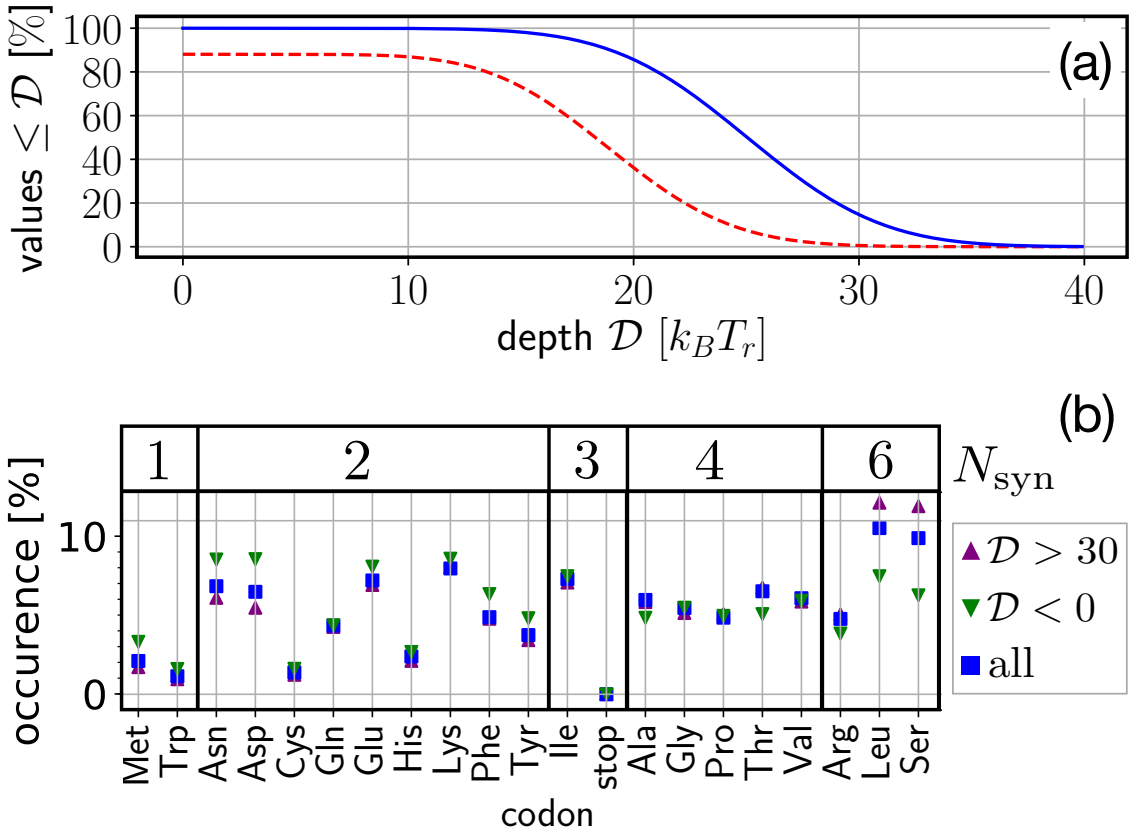

Figure 3.6: (a) Percentage of all positions on genes from yeast *S. cerevisiae* where we created at least a minimum of depth $\mathcal{D}$. Full method (blue continuous curve) and synonymous ground state sequence only (red dashed curve). (b) Full method: probability of a given amino acid to occur for 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].