The mechanical genome : inquiries into the mechanical function of genetic information

Tompitak, M.; Tompitak M.

6 NUCLEOSOME POSITIONING SIGNALS IN GENE PROMOTERS

Er muss sozusagen die Leiter wegwerfen,
nachdem er auf ihr hinaufgestiegen ist.

—Wittgenstein

This chapter is based on:
Tompitak, Vaillant and Schiessel 2017 *Biophys. J.* 112.3 505–511 [79]

We now leave the full Rigid Base Pair model and the Eslami-Mossallam nucleosome model [1] behind and, in this chapter, rely on the Markov-chain model of Chapter 4. At the expense of accuracy, we gain astronomically in computational cost. With this fast method in hand, we are now able to turn towards entire genomes and analyze the nucleosome affinity of billions of different sequences, and look for real nucleosome positioning signals in nature.

6.1 INTRODUCTION

Nucleosomes are the fundamental packaging units of DNA that eukaryotic organisms employ to render their genomes compact enough to fit inside a cell, consisting of about 147 base pairs worth of DNA wrapped around a histone core. This packaging also restricts access to the genome: DNA bound to histones is unavailable for coupling to many other DNA-binding complexes, such as the transcriptional machinery. Therefore, the positioning of nucleosomes along the genome interacts with gene expression, as was already realized some three decades ago [33, 34].

This interplay suggests that nucleosomes may play a role in gene regulation, and nucleosomes are in fact actively displaced in order to regulate gene expression [158, 159]. Genomic sequences may also have evolved to position nucleosomes in specific, beneficial locations. This possibility is suggested both by the fact that the degeneracy of the genetic code in principle allows for multiplexing of such positioning signals with genetic information [1, 116–119], and by the observation that the mutation patterns of DNA bound to histones differ from those of linker DNA [48].

Research into such nucleosome positioning signals, hardcoded into eukaryotic genomes, has veritably exploded over the last decade, primarily due to the development of experimental methods that allow for efficient genome-wide nucleosome mapping [160]. This research has provided insight into the importance of nucleosomal sequence preferences for chromatin organization [161], and has allowed for the creation, refinement and testing of many models for predicting nucleosome positioning along genomes [73, 75, 162]. The intrinsic nucleosome-DNA affinity of genomic sequences appears to play a significant role *in vivo* in positioning nucleosomes in certain regions of the genome, such as transcription start sites (TSSs) and origins of replication [161], alongside other effects like the presence of proteins that compete for the same DNA stretch or the action of chromatin remodellers [31, 163].

Around the TSS of *S. cerevisiae* (baker's yeast), nucleosomes have been found to be depleted on average, both *in vitro* and *in vivo* [62–64, 130, 164–167]. The persistence of this depletion *in vitro*, in the absence of active remodeling, identifies the sequence preferences of nucleosomes as the dominant cause. Those preferences have been measured and utilized in various models to explain the observed nucleosome depletion [63, 64, 71, 72, 166]. These nucleosome-depleted regions (NDR) in gene promoters are thought to be encoded into the genomic sequence to allow RNA polymerases more ready access to the TSS, thereby facilitating transcription [62]. This is not only of interest for an understanding of the workings of natural genomes, but has also recently been put forward as an interesting engineering method to modulate transcription in synthetic genomes [168].

Since the earliest studies on baker's yeast, inquiries into nucleosome positioning have been extended to the genomes of many other organisms, such as *S. pombe* [149, 169] and various other species of yeast [170], *C. elegans* [171, 172], *Plasmodium falciparum* [173], flies [174], zebrafish [175], *Arabidopsis thaliana* [176], mice [177–179] and humans [133, 178, 180–183]. Most of these studies were conducted *in vivo*, and therefore do not allow for isolation of effects encoded into the genomic sequences. This body of research shows, however, that sequence effects alone are not generally sufficient to explain *in vivo* observations [163]. An important role is also played by the active regulation of transcription. In yeast, the promoters of actively transcribed genes show much more pronounced nucleosome depletion than those of inactive genes [169].

In human cells, as in yeast, NDRs were found *in vivo* only for actively expressed genes [180]. However, *in vitro* nucleosome mapping reveals that the human genome does not share yeast's strategy of depletion-by-default.

Instead, it was found that promoter regions in the human genome showed enhanced nucleosome occupancy. One interpretation is that this is a reflection of the differentiated nature of human cells: it may be more beneficial to keep genes relatively inaccessible by default, and to actively open up the promoter region only when needed [133, 182]. This idea seems to be countered by newer results, however, which find stronger intrinsic nucleosome-attracting regions (NARs) for housekeeping genes than for tissue-specific genes, directly opposite of what one would expect [29]. Those results indicate that the function of the NARs in the human genome may be to retain nucleosomes in sperm cells (in which most nucleosomes are removed from the chromatin) and so pass on epigenetic information to the next generation.

Whichever is the case, these ideas raise the question whether the presence of an NDR in yeast versus that of an NAR in humans might be a general distinguishing feature between unicellular and multicellular life. In order to answer this question, we utilize a purely mechanics-based model for the sequence-dependent DNA-nucleosome affinity to predict *in vitro* nucleosome positioning signals, and compare the signals encoded into the promoter regions of a wide range of genomes.

6.2 METHODS

6.2.1 Data acquisition

Let us briefly summarize the origins of all the experimental data used in this chapter. All genomic sequences and gene (cDNA) data were downloaded from ensemblgenomes.org, release 31 [184]. The *in vitro* nucleosome map produced by Kaplan *et al.* [64] was retrieved from GEO accession number GSE13622. The map from Valouev *et al.* [133] was downloaded from [185]. The map from Locke *et al.* [186] was downloaded from [187]. The data from Ercan *et al.* [172] was taken directly from Fig. 1C in that reference. TSS locations in *S. cerevisiae* were derived from [188] in the manner described in [189].

6.2.2 Model

The model used for the work in this chapter is the trinucleotide approximation to the Eslami-Mossallam nucleosome model [1] described in Chapter 4, with one major alteration. For the current work, the parameterization

of the Eslami-Mossallam nucleosome model was changed from the hybrid parameterization described in [1], to a parameterization informed solely by crystallography data [11]. We found that this improves its applicability to long-range effects. See Appendix A for more information.

6.2.3 Sequence analysis

For every genome analyzed, we calculated the averaged signal as follows. For every annotated gene, we looked up the location of the TSS, and extracted the 1146 bp before and after. For each of the resulting sequences, we calculated a probability landscape for nucleosome positioning using the trinucleotide model mentioned above. We would like to calculate occupancies from these landscapes and average over all genes. Unfortunately, because the probabilities vary over several orders of magnitude, the number of genes is generally not large enough to provide a meaningful average; it tends to be dominated by the highest probabilities. Therefore, we instead consider the average energy landscape for a given organism.

From the predicted probabilities, an energy landscape can be calculated up to a constant shift, since such a probability is the normalized Boltzmann weight of a state. We took the average of the energy landscapes of all the sequences as a representative energy landscape for a given organism. For each bp (-1000 to +1000) we then calculated the nucleosome occupancy by summing the Boltzmann probabilities of all 147 nucleosome positions that lead to that bp being covered by the nucleosome. This gives us a prediction of the intrinsic nucleosome affinity encoded in the genomic sequences.

6.3 OPPOSING NUCLEOSOME OCCUPANCY SIGNALS IN YEAST AND HUMAN GENOMES

The high-coverage *S. cerevisiae* nucleosome maps provide the standard testing ground for any model designed to predict nucleosome occupancy [51]. Applying our nucleosome affinity model from Chapter 4, we find a peak in the free energy of the nucleosome in the promoter regions of *S. cerevisiae* (Fig. 6.1), which correctly predicts experimentally observed NDRs in these regions. The comparisons, for regions centered on the TSSs and on the start codons, are shown in Fig. 6.2A and B, respectively.

For the human genome, a map of *in vitro* nucleosome occupancy has been published by Valouev *et al.* [133], and, as predicted by Tillo *et al.* [182],

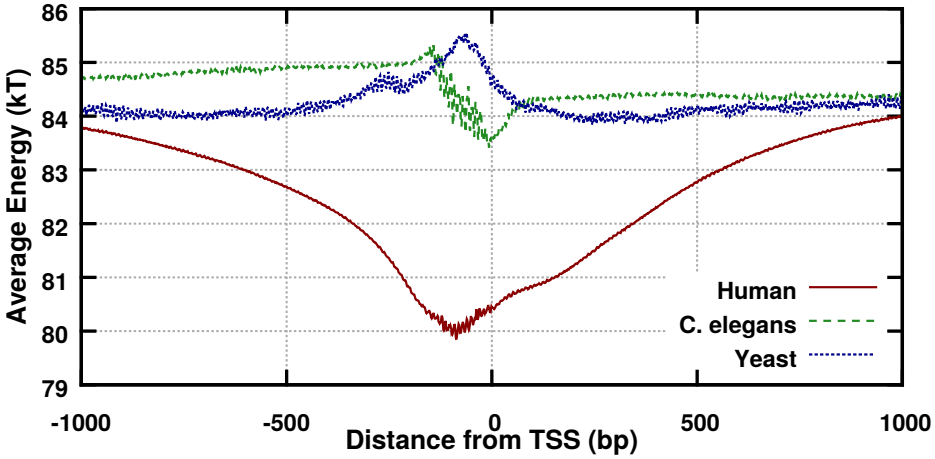

Figure 6.1: Average free energy landsapes in the promoter regions of the human, yeast and *C. elegans* genomes.

it reveals occupancy signals opposite to that of yeast: human promoters seem to encode for high, rather than low, nucleosome occupancy. Vavouri and Lehner [29] similarly find an increased retention of nucleosomes when nucleosomes are depleted in human sperm cells. Correspondingly, when applying our model to the promoter regions of the human genome, we find a very strong NAR around the TSS, due to a vast dip in the free energy (Fig. 6.1), as can be seen in Fig. 6.2C.

Initially surprisingly, the signal found by Valouev *et al.* is an order of magnitude smaller than that predicted by our model and that found by Vavouri and Lehner. This discrepancy can be explained when we consider that the nucleosome density cannot exceed 1 per 147 bp due to excluded volume. The experiment attempts to measure enrichment of nucleosomes in the promoter regions relative to the average density of nucleosomes. Unlike in experiments that look at nucleosome depletion or retention, the excluded volume between nucleosomes puts a limit on how strong the enrichment can be in practice.

This is the reason for the discrepancy between the *in vitro* results of Valouev *et al.* and ours and those of Vavouri and Lehner. In order to approximate the effects of steric interactions, we applied Percus' equation [191] to our average energy landscapes, and solved it as described in [192]. The solution depends on the chemical potential of the nucleosomes binding to the DNA (see also [190]), which we adjust to achieve a good fit with the *in vitro* data. We see that steric interactions can indeed explain the very

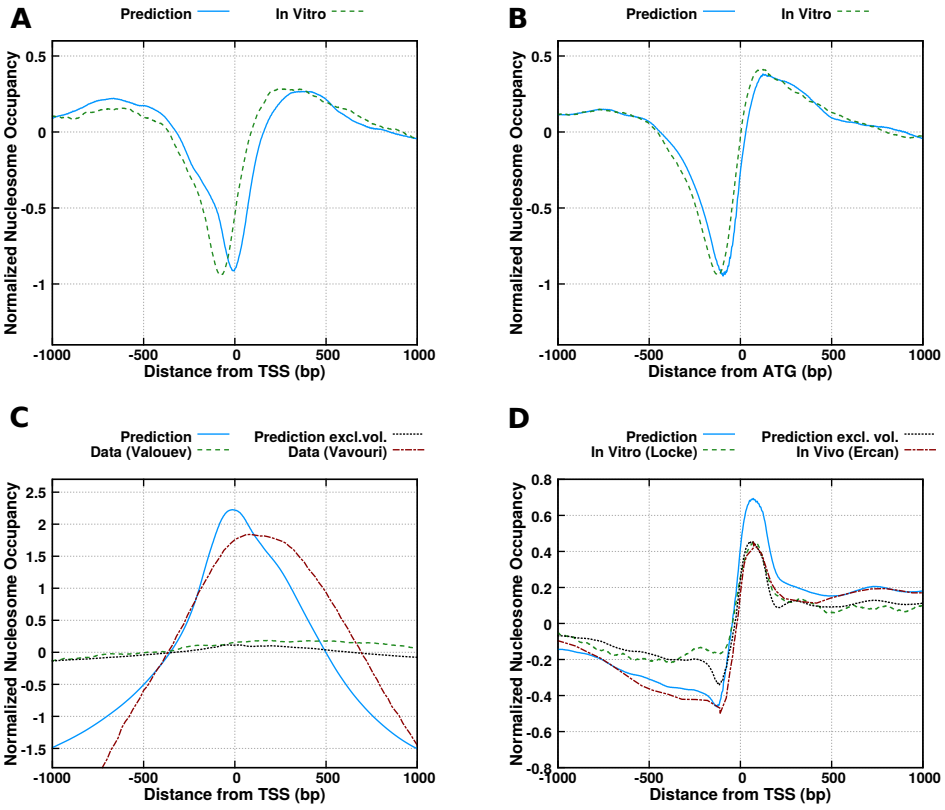

Figure 6.2: Comparison of predicted and measured intrinsic nucleosome positioning signals in promoter regions. The quantities plotted are the natural logarithms of the occupancies and the signals have been normalized such that they average to zero. In all plots, the solid blue curves are our predictions in the limit of low nucleosome density, which give an account of the strength of the signals intrinsically encoded. The dashed green curves represent *in vitro* measurements. The dotted black curves are predictions taking into account steric interactions. Using the same treatment as in [190], these curves have a free parameter $\tilde{\mu} = \mu - \langle E \rangle$, i.e. the difference between the chemical potential and the average energy of the landscape, which we determined to be -8.5 kT for yeast (curves not shown due to similarity with the low-density limit), -5.7 kT for *C. elegans* and -1.38 kT for humans. **A, B:** *S. cerevisiae*, average nucleosome occupancy centered on the TSS and start codons, respectively. Data from [64]. **C:** Like A, for *H. sapiens*. The *in vitro* data is from [133]. Additionally shown is the nucleosome retention signal from [29]. **D:** Like A, for *C. elegans*. The *in vivo* data is from [172], the *in vitro* data is from [186].

weak signal for humans (dotted black curve in Fig. 6.2C) as well as the apparent overshoot of our prediction for C. elegans (dotted black curve in Fig. 6.2D).

This means that also at physiological conditions, the nucleosome density will be saturated at much smaller values due to steric interactions. However, we stress that independent of this saturation effect, a nucleosome at the peak of the nucleosome occupancy signal will be strongly energetically bound, and so hinder transcription if it is not actively removed, as well as be more stable under a nucleosome-depleting force.

The results of Vavouri and Lehner [29] when examining where nucleosomes are retained when they are depleted from chromatin in human sperm are more in line with our predictions, as can also be seen in Fig. 6.2C. When depleting nucleosomes, excluded-volume interactions are not a constraint and our predictions can be probed. Although these authors studied a special *in vivo* situation, the nucleosome retention signals were found to correlate strongly with DNA sequence. Because the depletion of nucleosomes in sperm is an out-of-equilibrium process, and our model therefore does not make direct numerical predictions for this situation, we note the similarity between our predictions and the *in vivo* nucleosome retention signal.

We thus have interesting observations and predictions on two ends of a spectrum. A very simple, unicellular eukaryote shows nucleosome depletion as its most prominent, intrinsically encoded nucleosome positioning feature. A complex multicellular one shows high nucleosome occupancy instead. What happens in between these two extremes?

In Fig. 6.2D we present a comparison between our predicted nucleosome occupancy signal (for the underlying free energy landscape, see Fig. 6.1) for the nematode *C. elegans* and the signals found *in vitro* by Locke *et al.* [186] and *in vivo* by Ercan *et al.* [172]. We find remarkable agreement in the shape of the signal, indicating that the data is indeed indicative of intrinsically encoded nucleosome positioning. Somewhat surprisingly, the *in vitro* and *in vivo* signals are similar to each other, which is not as strongly the case for yeast, and even less so for humans (see e.g. Fig. 3 in [29]). It has been noted that an *in vivo* nucleosome occupancy map of *C. elegans* lacks many of the features that distinguish *in vivo* maps from *in vitro* maps of yeast, such as strongly phased nucleosomes. Valouev *et al.* [171] find much flexibility in nucleosome positions in *C. elegans*. Such variability may average out some of the effects of active remodeling, rendering the two maps similar.

C. elegans seems to show a nucleosome positioning signal that is a hybrid of the signals found in the yeast and human genomes. It has an NDR upstream of the TSS, like yeast, but it also shows a significant NAR just after the TSS.

6.4 INTRINSIC NUCLEOSOME POSITIONING SIGNALS ARE INDICATIVE OF MULTICELLULARITY

The hybrid behavior in *C. elegans* may be hypothetically explained. As suggested by Tillo *et al.* [182], organisms may wish to tune their genomic sequences to intrinsically deactivate genes that are active only in some cell types, while intrinsically activating those that are common to all of its cells. In unicellular life, most genes will not be permanently silenced, leading to an overall average depletion signal. In complex multicellular life, the signal may be dominated by the many genes that are intrinsically deactivated, leading to an overall attractive signal. *C. elegans* may then represent a range of organisms where the two contributions are more equal, leading to both a depleted region just before the TSS (where it is also observed in yeast) and an attractive region just after (the peak in occupancy in the human genome is also skewed towards the right).

The results of Vavouri and Lehner [29], however, suggest that, at least in the human genome, the hypothesis of Tillo *et al.* does not hold, and the function of the NARs is to retain nucleosomes in sperm cells. The hybrid signal we find in *C. elegans* may in this case similarly play a dual role of facilitating initiation of transcription but at the same time assisting in nucleosome retention.

We can extend our observation of these signals to other genomes using our model. We mapped the nucleosome positioning signals for promoters in genomes across the tree of life and discovered organisms that have intrinsically encoded NDRs and NARs, as well as many that fall into the hybrid category. The full set of signals found and described below are presented in Fig. 6.3.

Most archaea (14 genomes analyzed) show a signal similar to that of yeast, in that a nucleosome-depleted region is the most prominent feature. Archaea are unicellular organisms that do not have histone octamers, but employ only tetramers of (archaeal) histones to compactify their DNA. We expect these tetramers to obey positioning rules similar enough to nucleosomes that our model is predictive of their occupancy. We therefore analyzed the octamer affinity landscapes, for the sake of comparison to

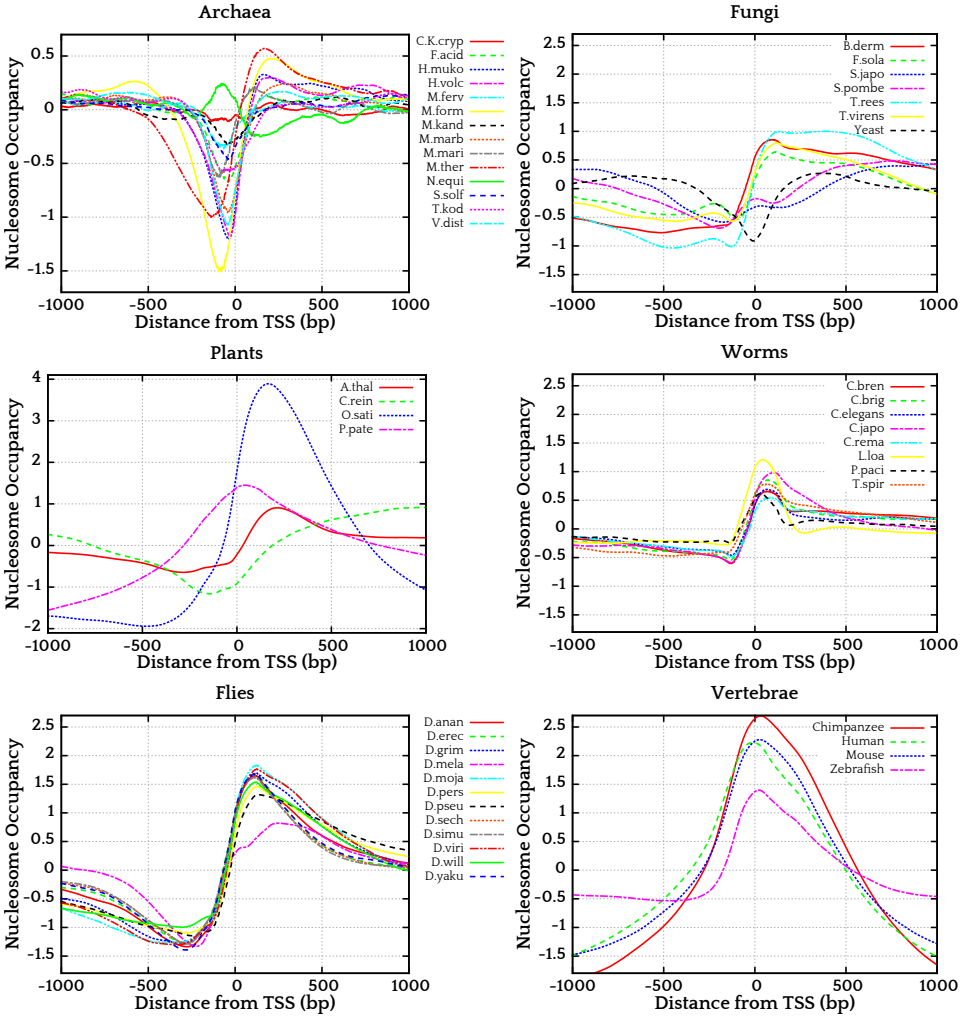

Figure 6.3: The full set of nucleosome positioning signals in the promoter regions of the organisms analyzed for this study.

eukaryotes, even though archaea do not possess them. The signals show that these simple unicellular organisms almost all fall into the depletion-by-default category.

Fungi (7 genomes analyzed) show somewhat more diverse signals than the archaea. While *S. cerevisiae* has a prominent NDR, many of the other fungi analyzed lack both a localized depleted region and a localized attractive region, but retain a step-function signal centered on the TSS. Fungal cells are not highly differentiated, but some fungi are dimorphic (they switch between unicellular and filamentous states), possibly causing these more hybrid-like signals.

Plants (4 genomes analyzed) come in many forms, from unicellular algae to complex multicellular life. As expected, we see various signals. The genome of *C. reinhardtii*, a unicellular alga, shows an NDR. Among the multicellular plants, we see two signals with a strong NAR, and one with hybrid behavior.

Among animals (24 genomes analyzed) we also find various signals. In worms, like *C. elegans*, we find both hybrid signals and more NAR-like signals. *D. melanogaster* and other members of its genus show strong hybrid signals, with a swift rise in nucleosome occupancy at the TSS. Finally, the zebrafish genome and all mammalian genomes analyzed (human, chimpanzee and mouse) have strong NARs.

We see a clear separation between unicellular and multicellular organisms. Though some signals from unicellular lifeforms show some hybrid characteristics, the dominant feature is generally an NDR. All multicellular genomes, on the other hand, either encode for high nucleosome occupancy in the promoter region, or show hybrid signals. This distinction persists across the eukaryotic phylogenetic tree and is clearly visible in Fig. 6.4, where we have plotted a representative set of signals, divided into unicellular and multicellular classes.

We finally note that, as was expected (see Section 1.3), these signals qualitatively correlate well with GC content – see Fig. 6.5 – which suggests that GC content is a prominent factor in shaping mechanical signals in promoter regions. Note however that, while GC content may be a good predictor of the nucleosome occupancy signals (the visual similarity between Figs. 6.4 and 6.5 is striking), it does not provide a numerical value for the occupancy without some sort of model.

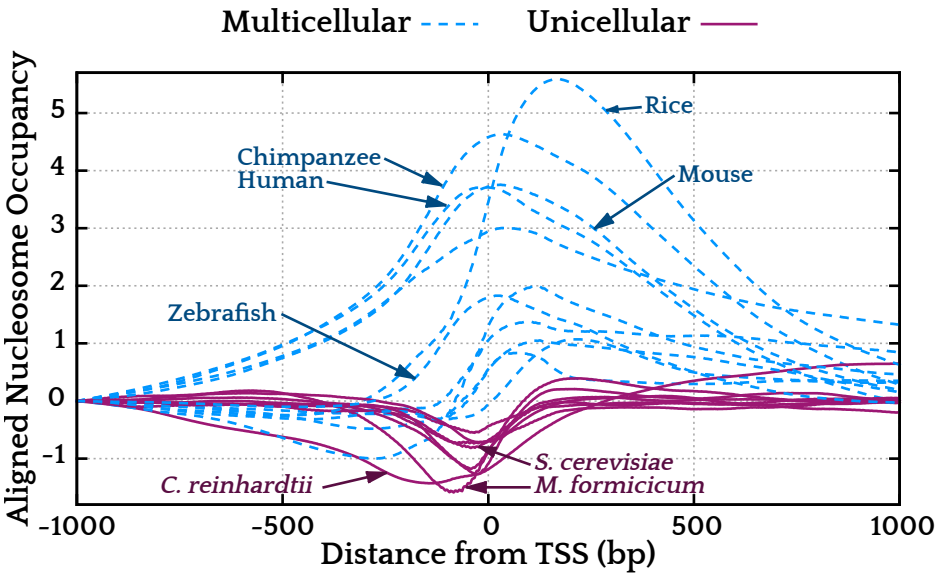

Figure 6.4: A representative selection of nucleosome positioning signals from various genomes. As a visual aid, the signals have been shifted vertically such that the logarithmic nucleosome occupancy at position -1000 is 0. The signals clearly fall into two distinct classes, based on whether the organism is unicellular or multicellular.

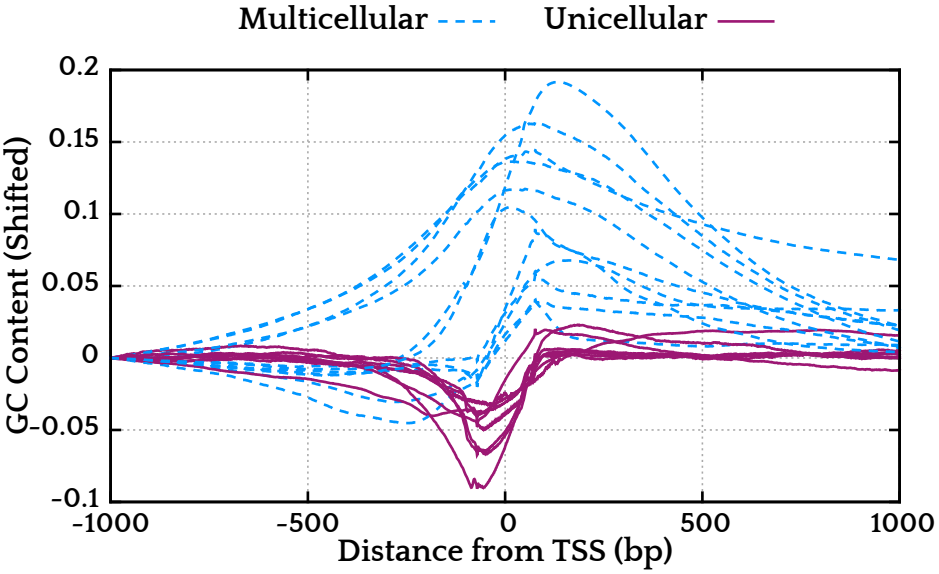

Figure 6.5: Average GC content in the promoter regions of the organisms for which the nucleosome occupancy signals were presented in Fig. 6.4.

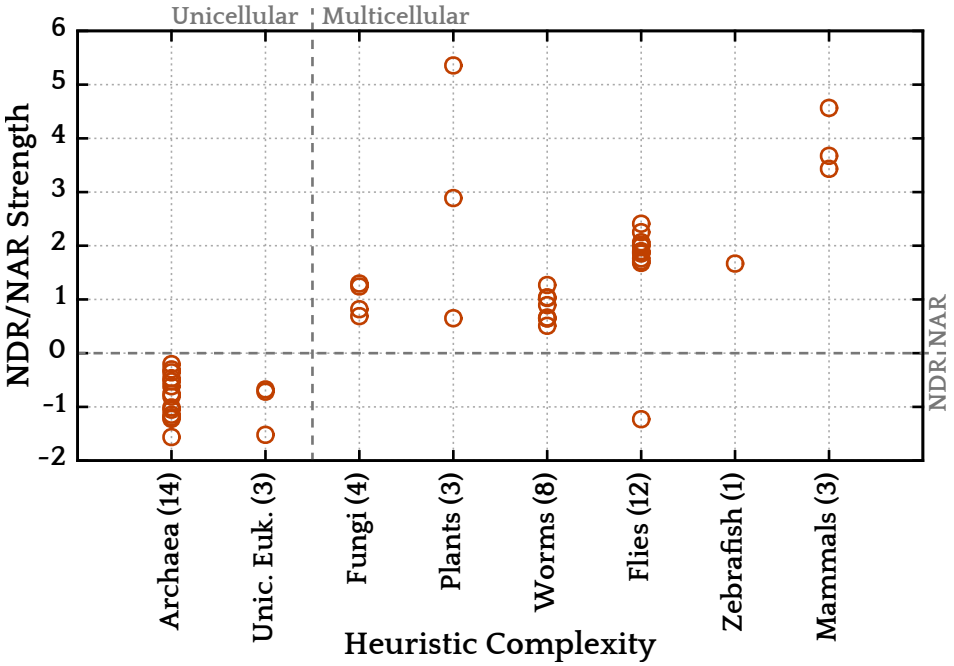

Figure 6.6: Promoter nucleosome positioning signal strength grouped by a heuristic measure of complexity of the organisms. The numbers in parentheses indicate how many genomes fall in each category.

6.5 INTRINSIC NUCLEOSOME POSITIONING SIGNALS CORRELATE WITH COMPLEXITY

One proposed measure for organism complexity is the number of different cell types an organism possesses [193], and the ideas presented here clearly have a link to this measure. Unfortunately, numerical data describing the numbers of cell types does not appear to be readily available in the literature, so we were unable to define a numerical measure of complexity. Therefore, we have restricted ourselves to ordering the organisms, by making assumptions about the cell type numbers. From simple to complex, we list: archaea, unicellular eukaryotes, filamentous and dimorphic fungi, multicellular plants, nematodes, Drosophila flies, zebrafish, and mammals.

We then considered the strength and direction of the NDR/NAR signals. To quantify this, we calculated the maximum and minimum of the signal and took the difference with the signal value at position -1000 relative to the TSS. We then took the largest of these two values (in the

absolute sense) and designated this value as the signal's strength (not in the absolute sense; a dominant NDR gives a negative signal strength).

The signal strength as thus defined clearly distinguishes unicellular and multicellular lifeforms (Welch's $t(39.051) = 10.5512$, p-value 5.4×10^{-13}) and the signals for multicellular organisms show correlation with our complexity ordering (Spearman $r_s = 0.52$, p-value 82.3×10^{-3}), as shown in Fig. 6.6. The ordering of the organisms is almost certainly imperfect, for example because all multicellular plants have been lumped together; without more accurate knowledge of the cell type numbers, there is no way to place them more realistically. However, the NDR/NAR strengths show a tentative trend. All unicellular eukaryotes have a negative signal strength, indicating an NDR, as noted in the previous section. All multicellular eukaryotes (with one exception, *D. melanogaster*) have a stronger NAR than NDR, and the strength of this NAR roughly increases with complexity. This observation concurs with the hypothesis of Tillo *et al* [182]. Our expectation based on that hypothesis would be that a more differentiated organism will have more genes that are nucleosome-occupied by default, leading to a higher NAR signal. It is not clear what purpose this correlation might serve in the context of nucleosome retention in the germline.

6.6 CONCLUSIONS

We found that the recently discovered fact that the human genome, unlike the yeast genome, encodes (on average) for an NAR rather than an NDR in the promoter region, is in fact a universal feature of multicellular life. The hypothesis put forth by Tillo *et al.* [182] is that this NAR suppresses gene transcription and that this suppression helps an organism with differentiated cell types manage its gene expression. Genes that are not needed in every cell type are suppressed by default, and only activated in those cells where they are necessary. In unicellular lifeforms, however, most genes will be in constant use, and keeping those genes easily accessible is more favorable.

On the other hand, Vavouri and Lehner [29] have found that the NARs found in humans in fact serve a different purpose, namely the retention of certain nucleosomes in sperm cells, and their study of the signals found for housekeeping genes versus tissue-specific genes directly contradicts the hypothesis of Tillo *et al.* The NARs we find in multicellular life may therefore instead be indicative of the need to retain nucleosomes in the germ cells of multicellular organisms.

NARs are common to complex multicellular lifeforms, while almost all unicellular lifeforms we analyzed have NDRs. In-between there is a range of organisms with hybrid positioning signals. In almost all of these signals, however, the NAR is a more prominent feature than the NDR. This leads to a clear distinction between uni- and multicellular life based on the type of nucleosome positioning signals found in the promoter regions.

Furthermore, the strength of the NAR appears to increase with organism complexity. This fits the hypothesis of Tillo *et al.* [182], since organisms with more cell differentiation will have more genes suppressed by an NAR (and possibly by stronger ones). If the purpose of the NARs is solely to retain nucleosomes in the germline, it seems that more complex life cares more strongly about retaining its nucleosomes and passing on epigenetic information. More research will be needed to explore this idea.

Given the presence of hybrid signals, we speculate that the encoding of NARs versus NDRs in promoter regions is not an all-or-nothing choice for organisms. Whether the NARs serve to close off genes by default, or to retain nucleosomes in the germline, they compete with an apparent need to create an NDR to facilitate the initiation of transcription. The organisms showing hybrid signals seem to strike a balance between the two.

We hope that our results will motivate the experimental community to expand the available catalog of *in vitro* nucleosome maps to a greater number and variation of organisms. This will help not only verify our findings but also be of great service to any follow-up inquiries into the deeper nature and meaning of the signals we have found. We also suggest that nucleosome maps be generated at lower nucleosome densities, because steric hindrance will hide strong enrichment signals.

We also hope to encourage further examination of housekeeping versus tissue-specific genes in other organisms to further test the hypothesis of Tillo *et al.* [182], and an expansion of the results of Vavouri and Lehner [29] to other organisms, in order to test whether or not nucleosome retention in the germline is a goal served by the mechanical signals we find in the genomes of other complex organisms. If so, our results raise an intriguing question: why do more complex organisms tend to favor stronger nucleosome retention?