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

Zuiddam, M.

Chapter 5

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

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

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

5.1 Introduction

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

5.1.1 Introduction to transcription

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

5.1.2 The role of nucleosomes in transcription

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

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

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

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

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

5.1.3 Overview

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

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

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

5.2 Multiplexing: Intraregional signals and interregional signals

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

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

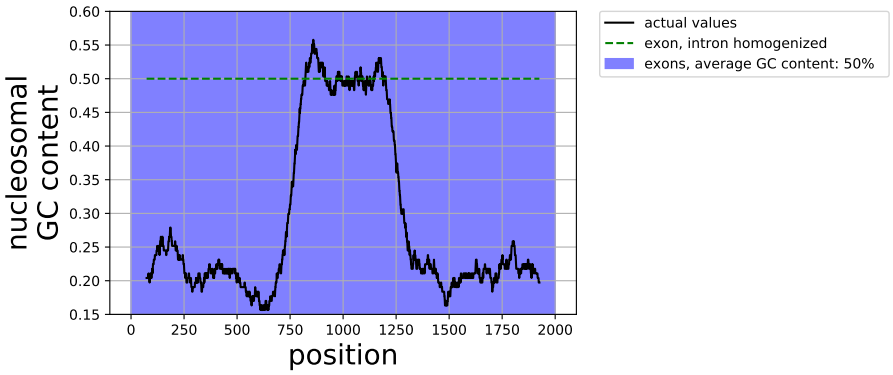
- intraregional multiplexing/intraregional positioning signals
- interregional multiplexing/interregional positioning signals

5.2.1 Intraregional signals

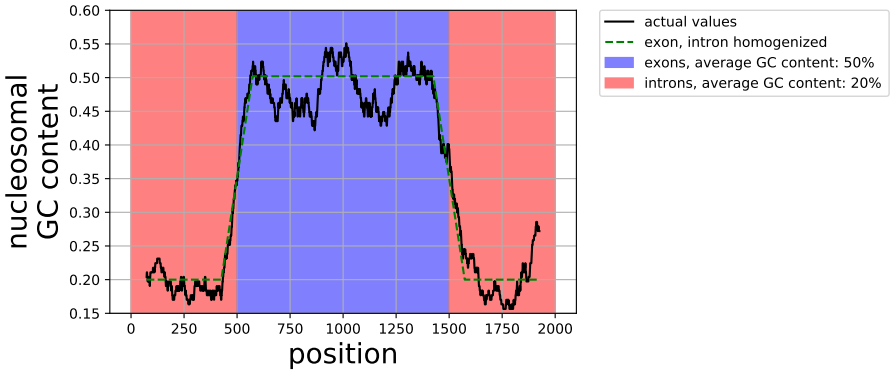
We will refer to intraregional multiplexing as intraregional positioning signals. These are signals that exist on a single region, see Fig. 5.1a for a theoretical example. In the case of exons, we get multiplexing of protein-coding information and mechanical information. Since non-exonic DNA can be functional,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

5.2.2 Interregional signals

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



(a)



(b)

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

5.3 Intraregional and interregional signals in a real genome

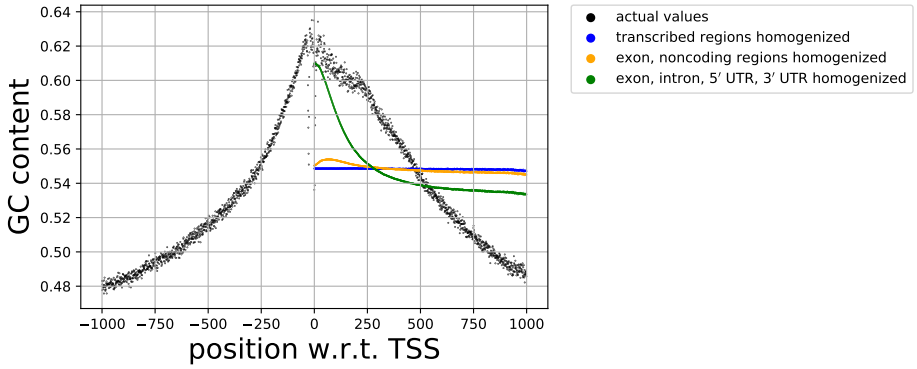
In this section we demonstrate for *Homo sapiens* how one can find these intraregional and interregional signals. All genomic data used throughout this chapter was acquired from the Ensembl Project website (www.ensembl.org), using their web-based tool Biomart. For a step-by-step guide on how to obtain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

5.3.1 Distinguishing the positioning signals by homogenizing

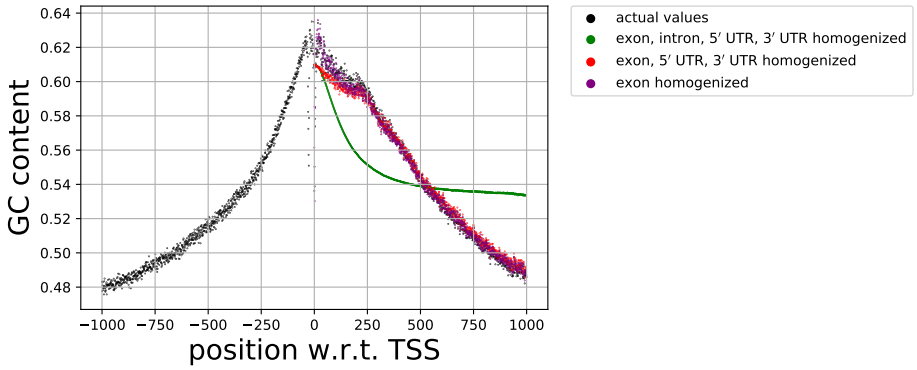
Now we introduce a method to find out whether a signal is intraregional or interregional. Interregional signals are caused by differences in averages between regions. When we replace all regions by their averages, we obtain the interregional signal size. The intraregional signal is simply the difference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

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

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

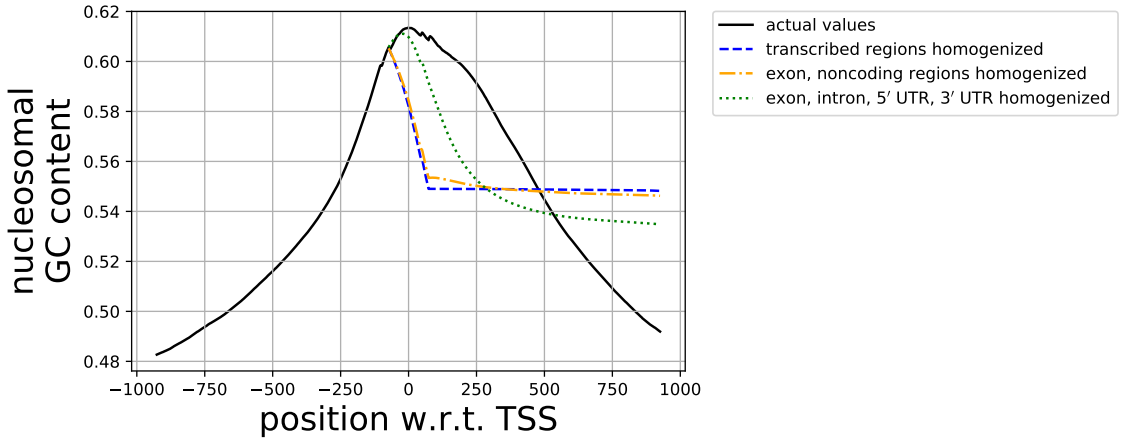


(a)

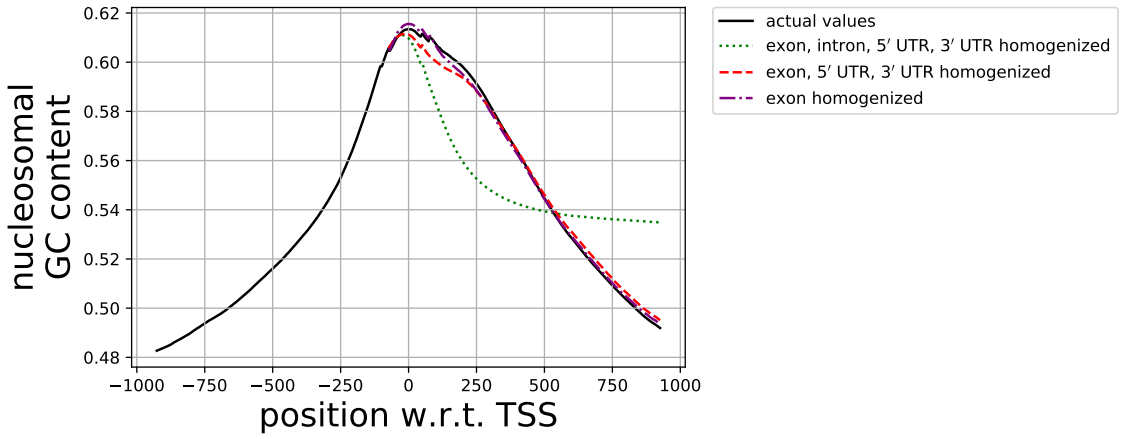


(b)

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

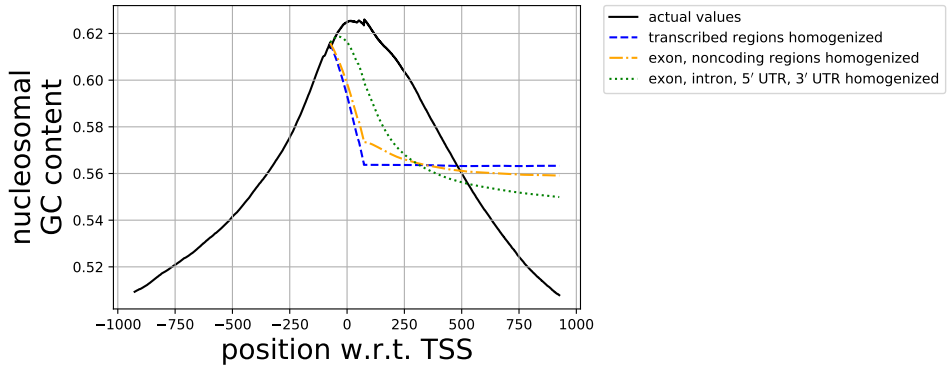


(a)

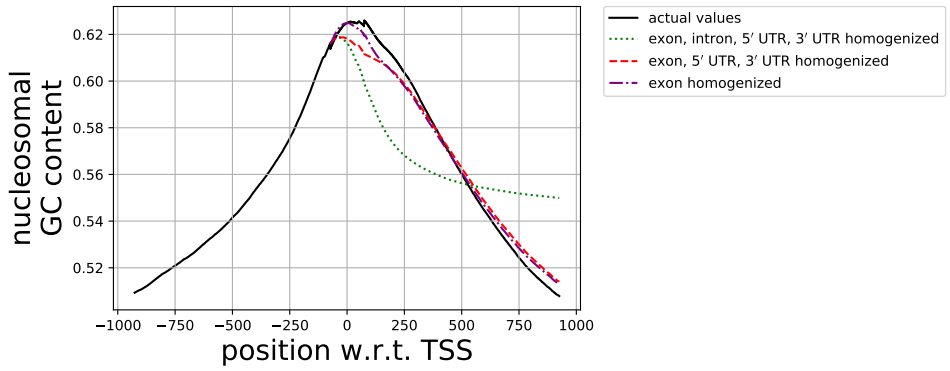


(b)

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

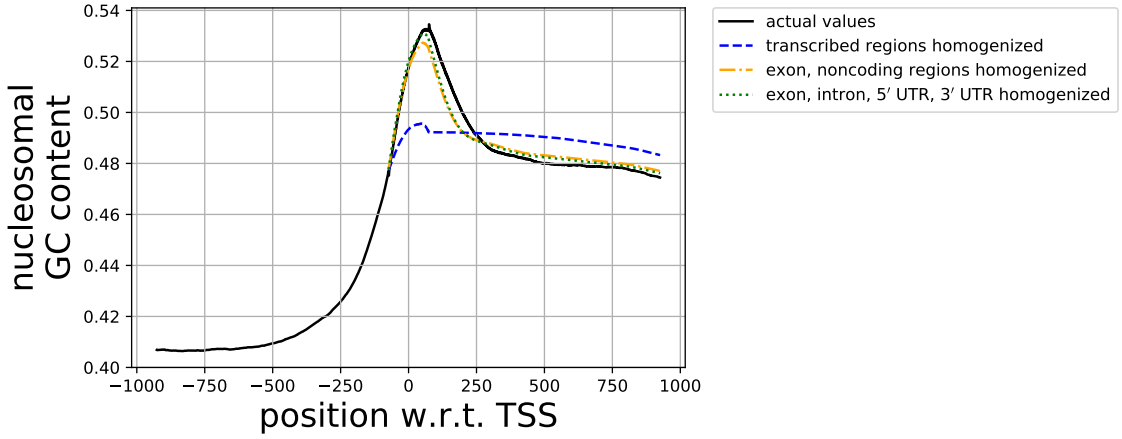


(a)

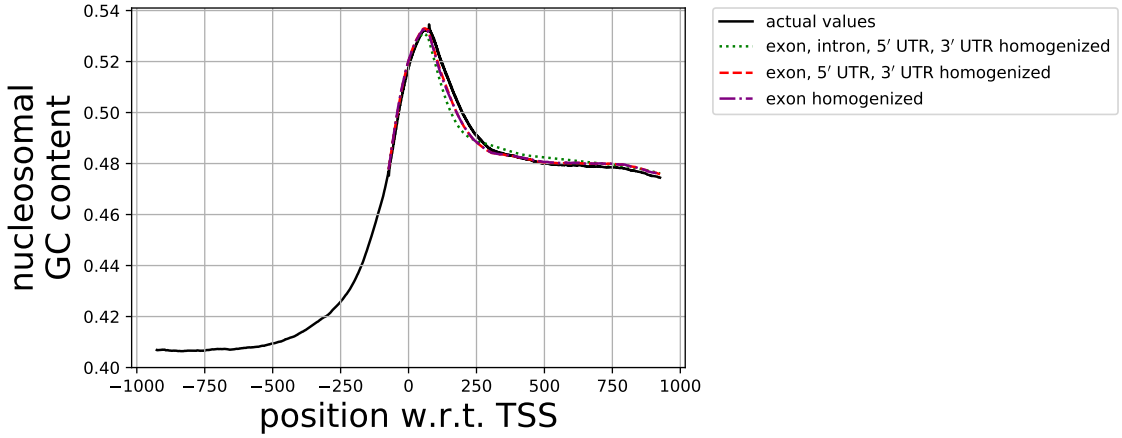


(b)

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

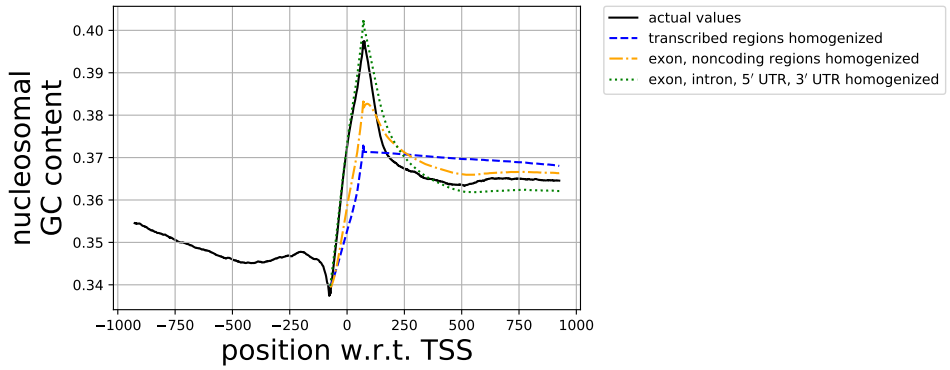


(a)

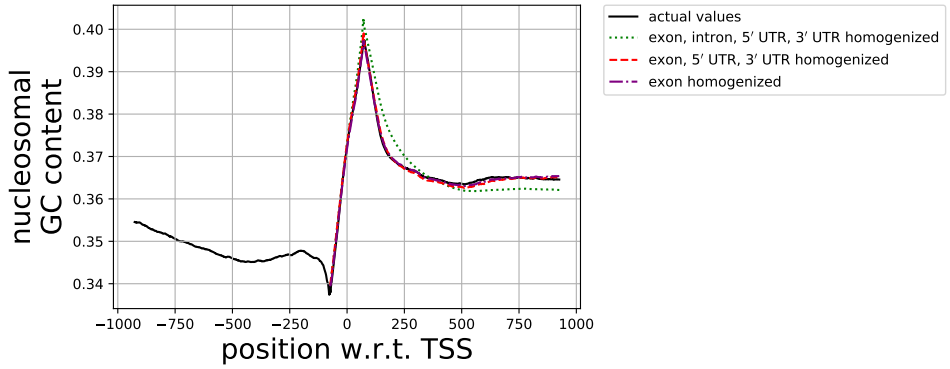


(b)

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



(a)



(b)

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

5.3.2 Obtaining the strength of the positioning signals

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

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

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

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

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

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

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

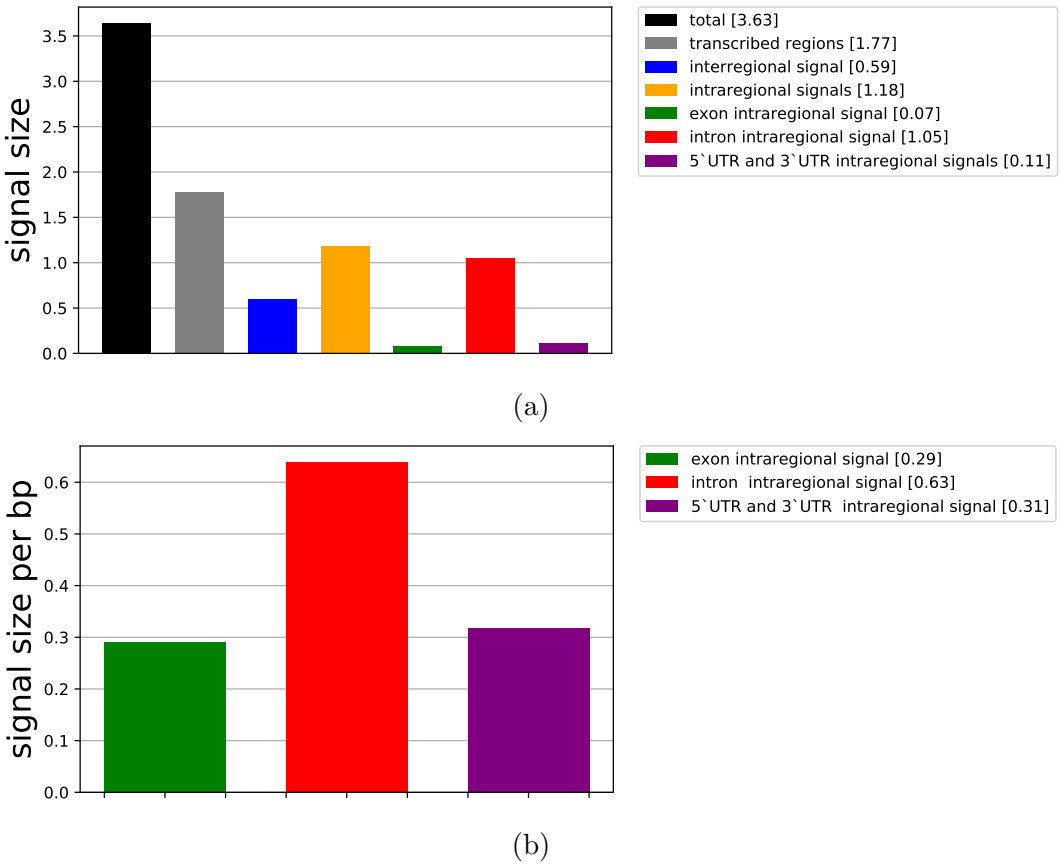


Figure 5.7: In (a) we see the signal sizes for several types of signals in human. The intraregional signals are much larger than the interregional signals. Exon and UTR signals seem nonexistent next to the intron signal size. In (b) we see the signal size per bp for the three separate intraregional signals.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

5.4 Signals on many animals

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

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

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

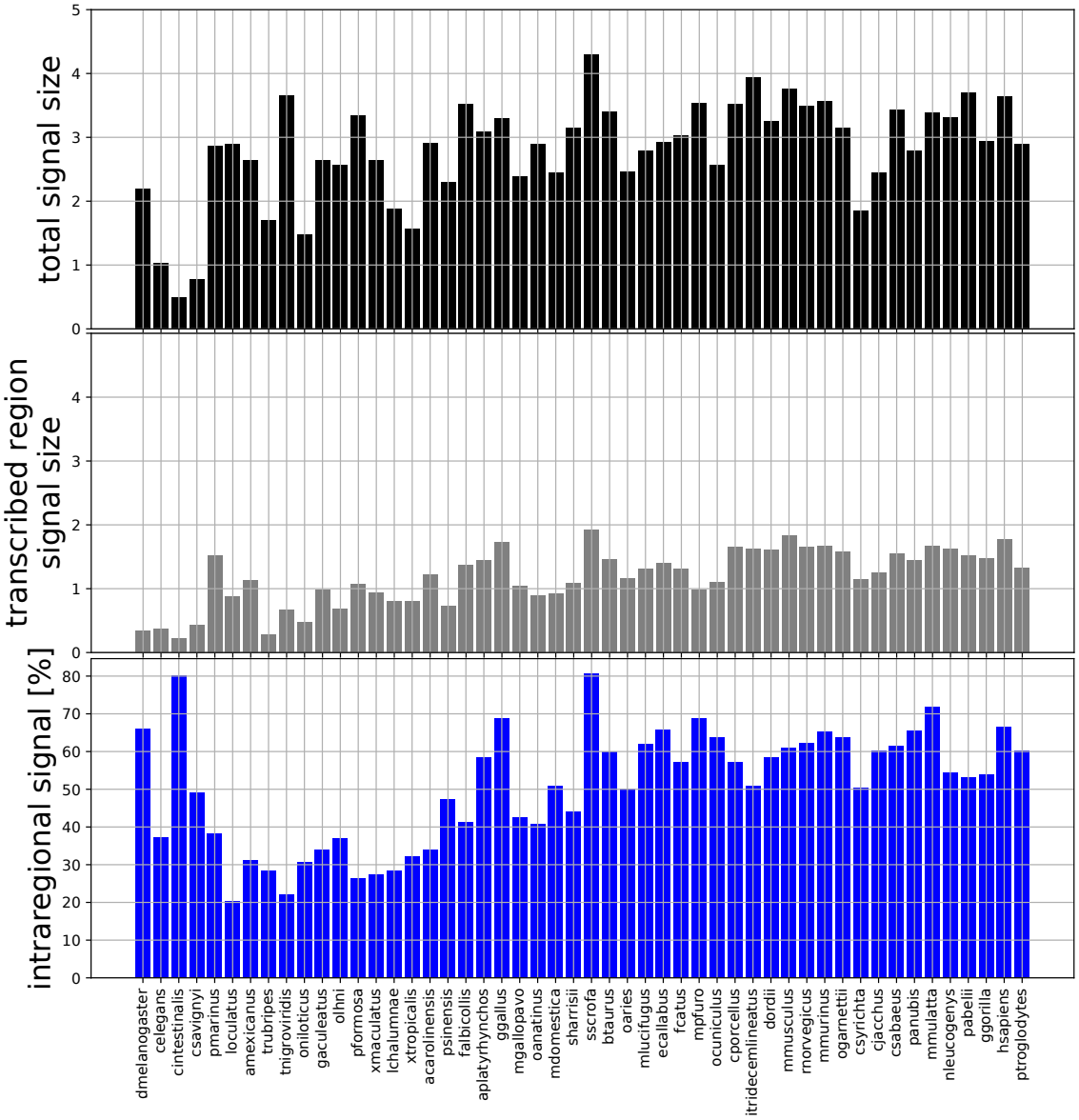
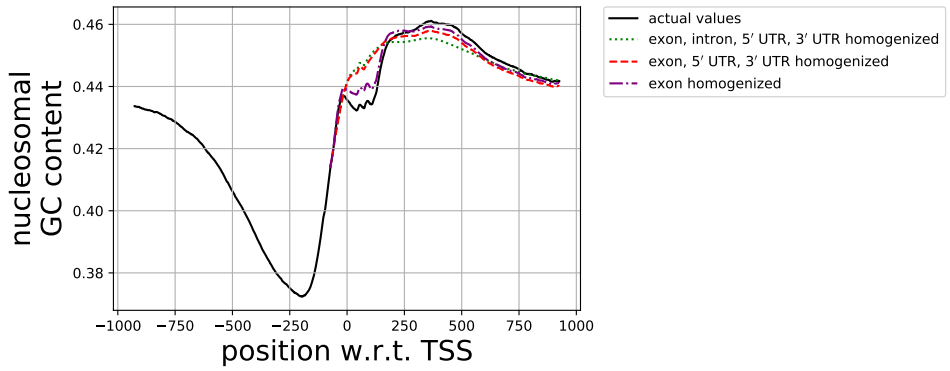
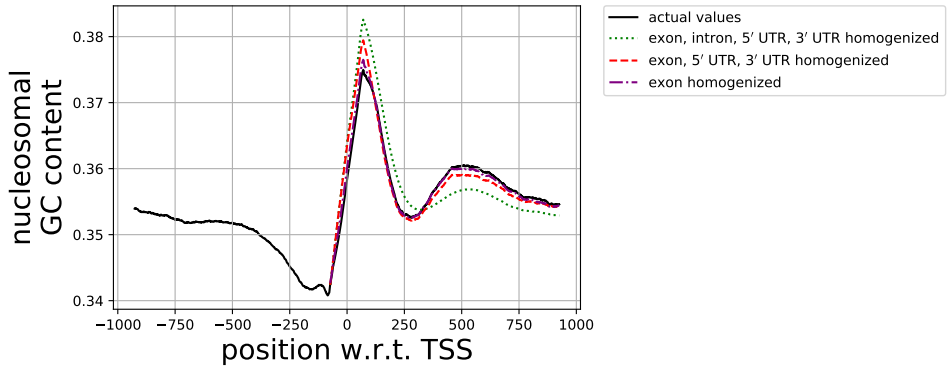


Figure 5.8: This figure depicts the signal sizes of many animals. At the top we see the total signal sizes of the animals, the middle figure depicts the signal sizes of the transcribed bases only, at the bottom we find the fraction (percent) that the intraregional signal contributes to the total signal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

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



(a)



(b)

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

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

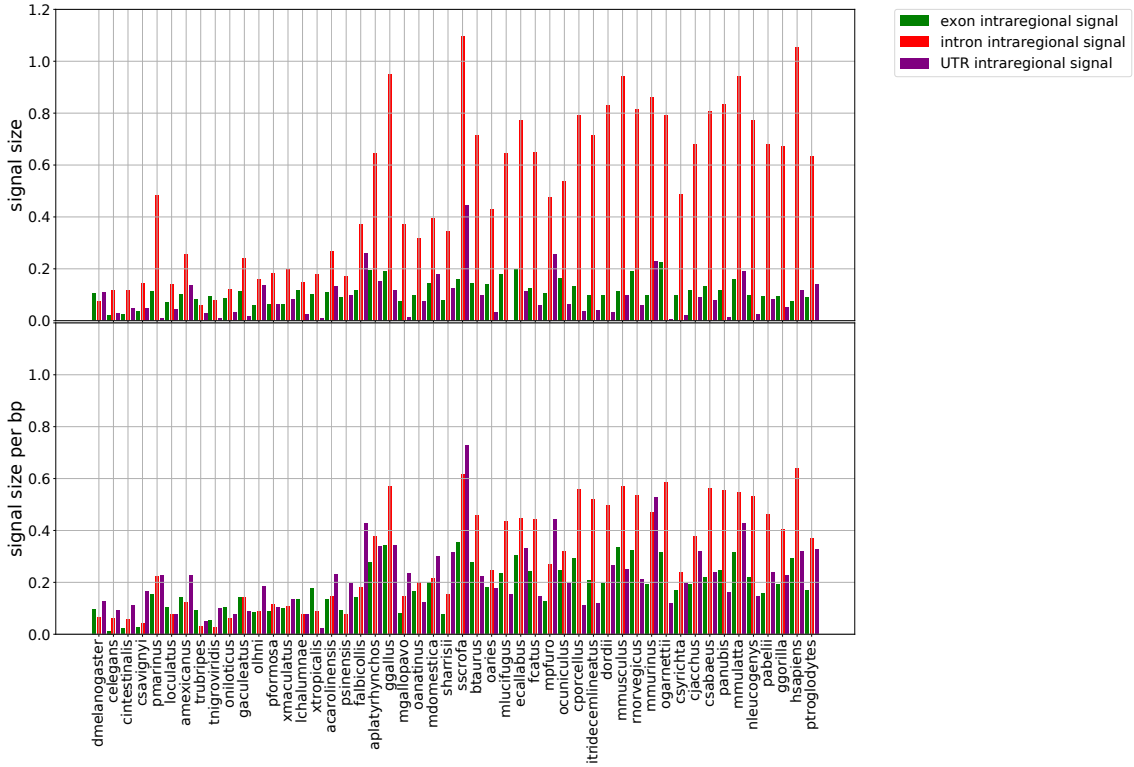
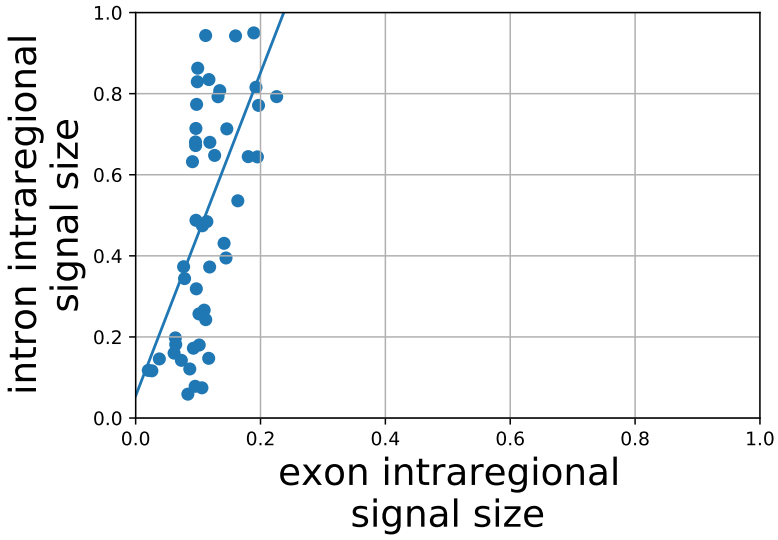
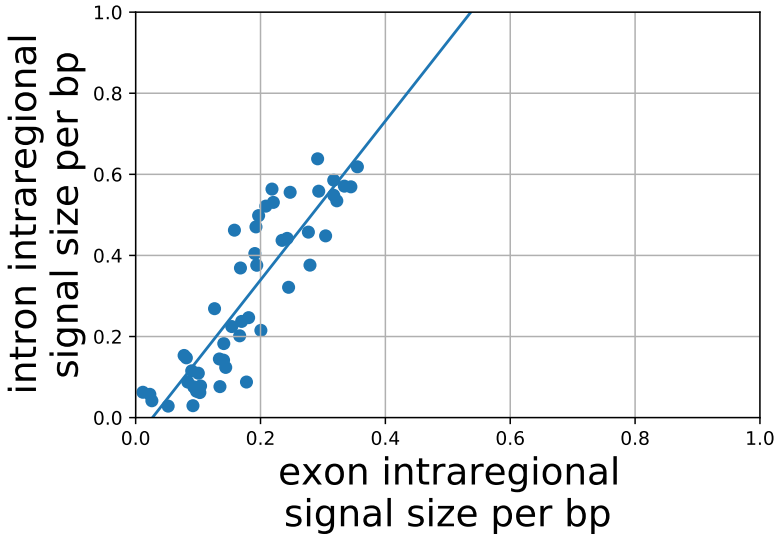


Figure 5.10: This figure depicts intraregional signal sizes of many animals. In the top figure we see that the intron signals dominate. In the bottom figure we see the signal sizes per bp, where the differences between exon, intron and UTR sizes are much closer to each other, introns still being more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



(a)



(b)

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

5.5 Signals on plants

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

but diverse range of plant organisms.

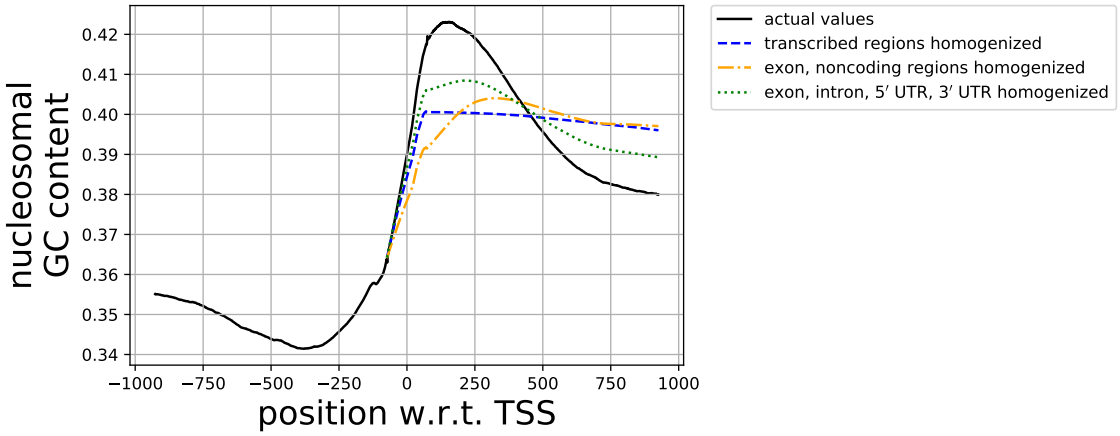
The first plant we will discuss is a tree, *P. persica*, the peach tree. Fig. 5.12 shows that a large part of the signal is caused by interregional signals, see the green dotted curve. The difference between this curve and the black curve, i.e. the intraregional signals, is almost entirely caused by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

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

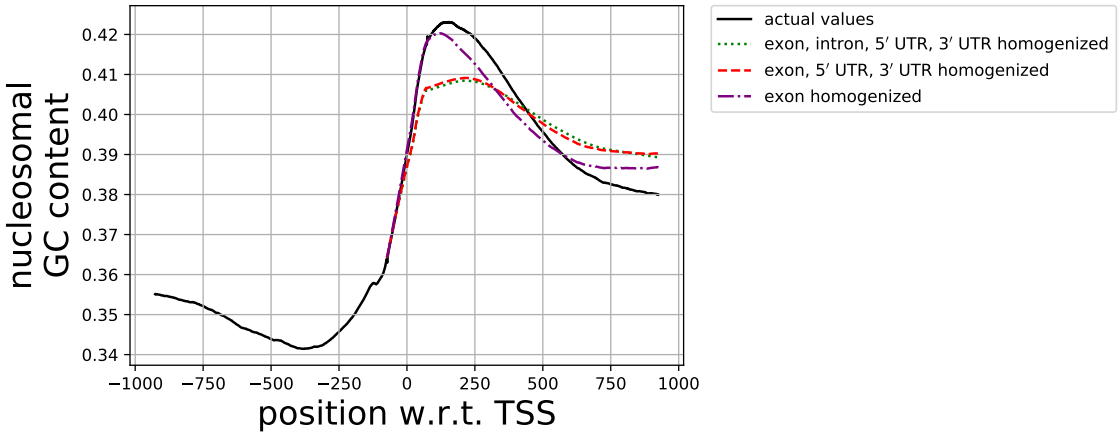
Fig. 5.14 depicts *S. tuberosum*, potato, which behaves very different from peach tree and thale cress. While the overall signal seems similar, it is almost entirely caused by interregional signals.

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

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

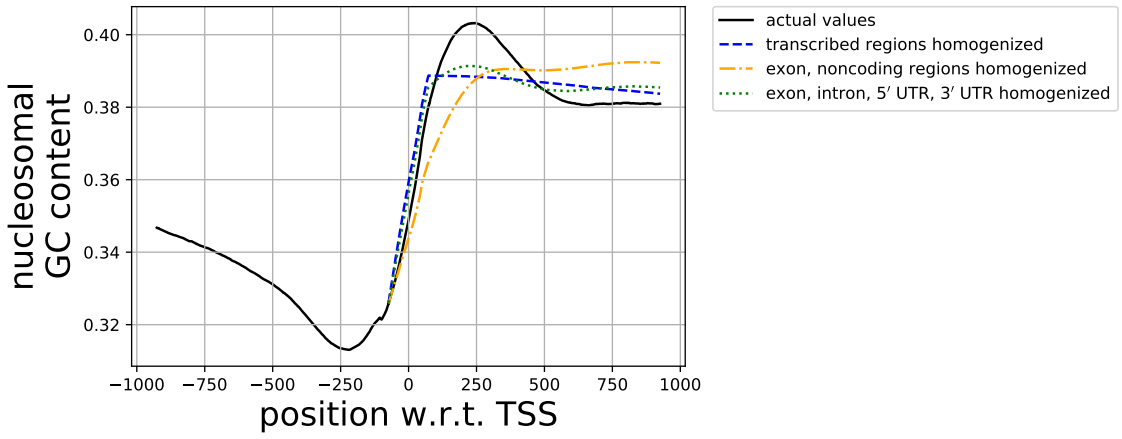


(a)

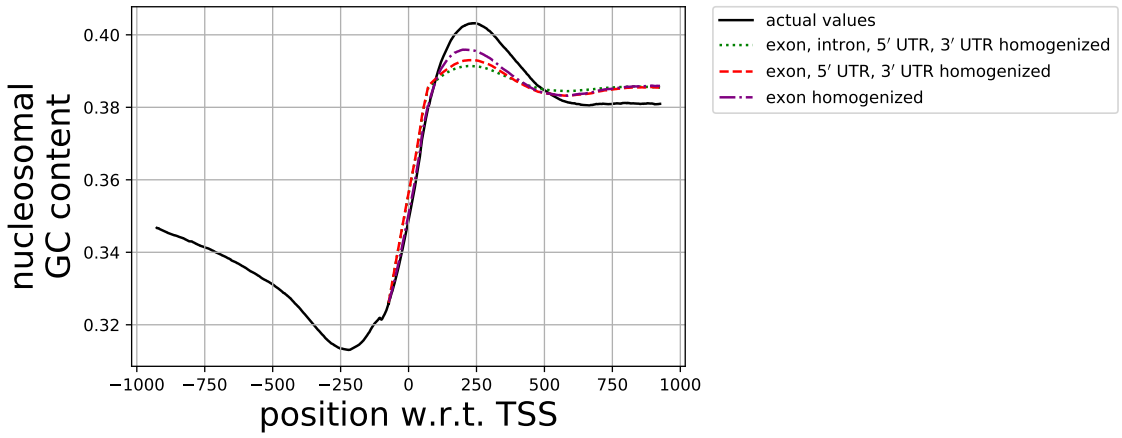


(b)

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

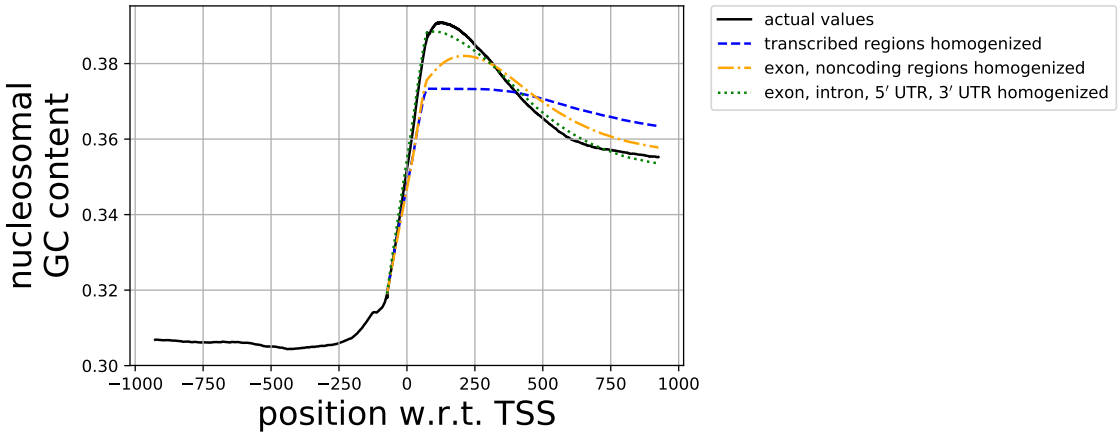


(a)

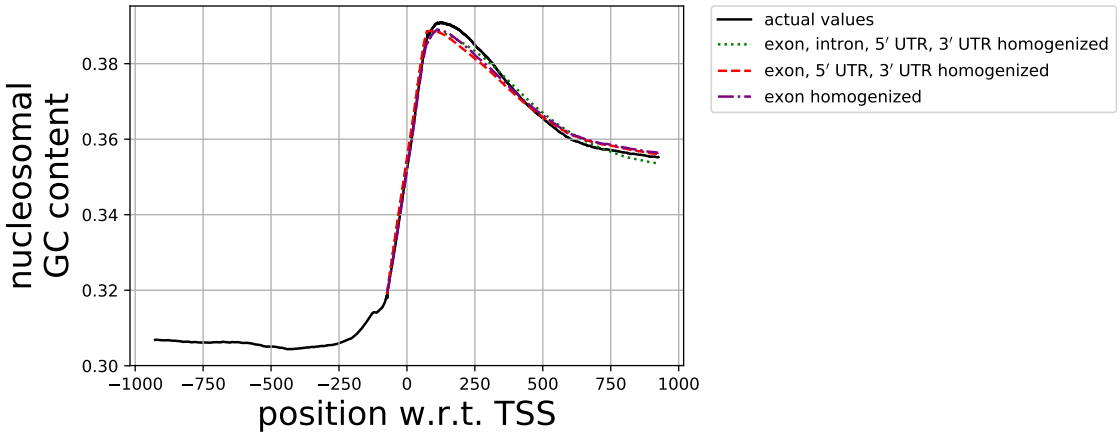


(b)

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

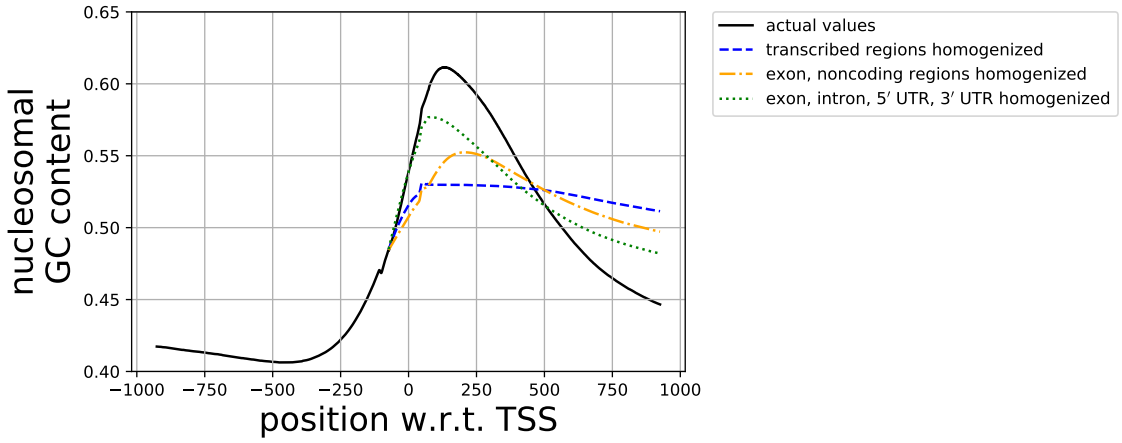


(a)

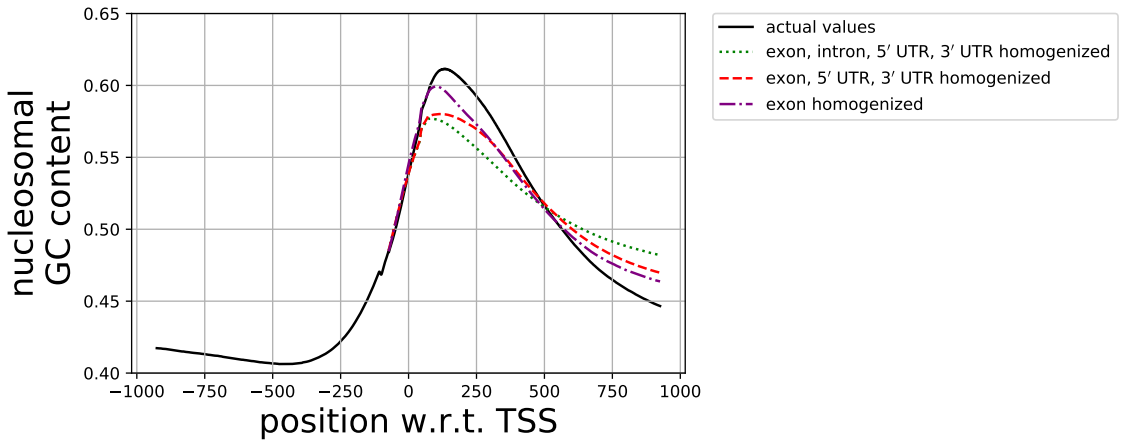


(b)

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

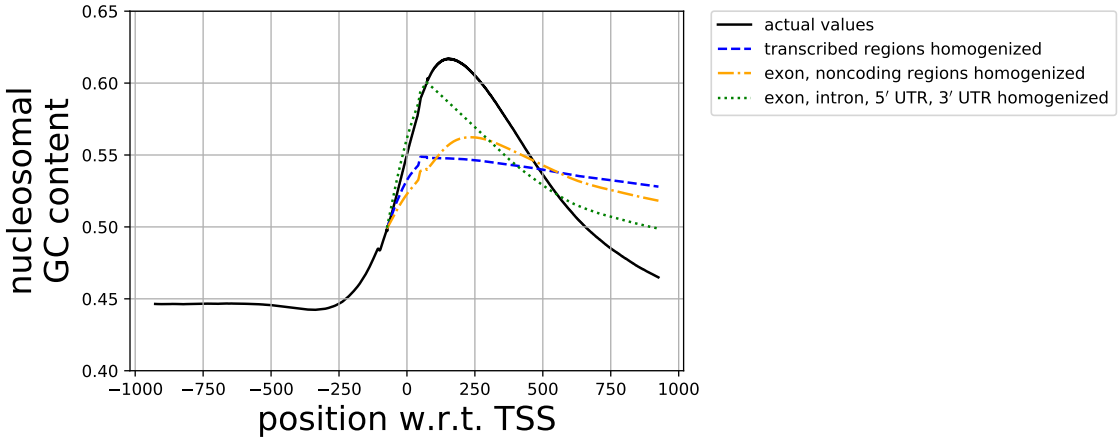


(a)

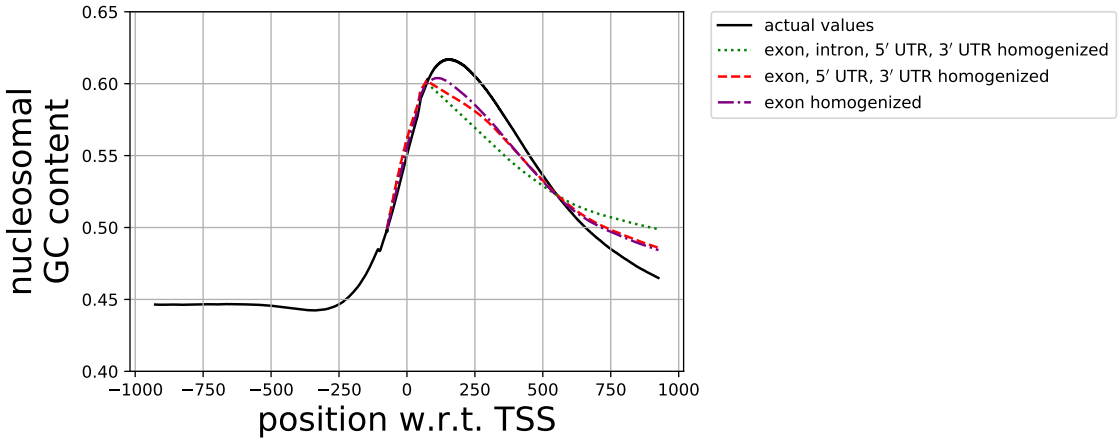


(b)

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

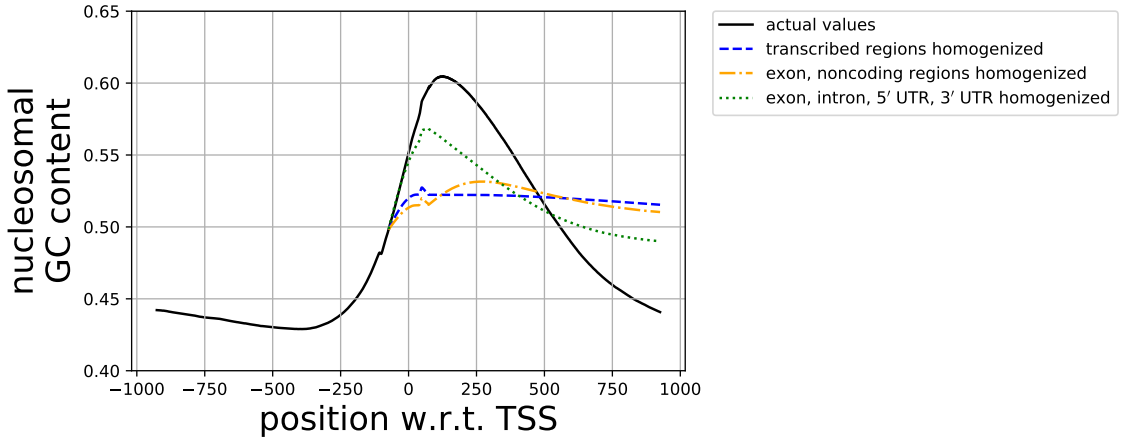


(a)

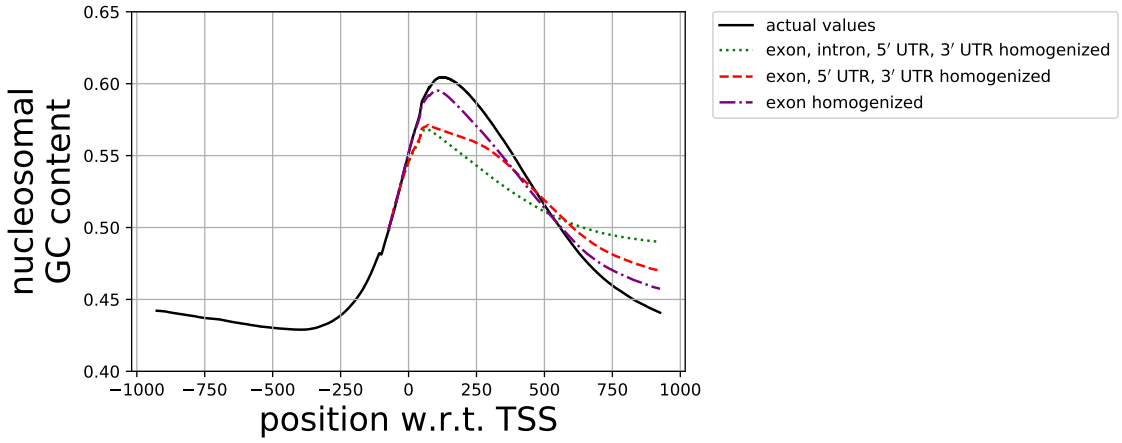


(b)

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

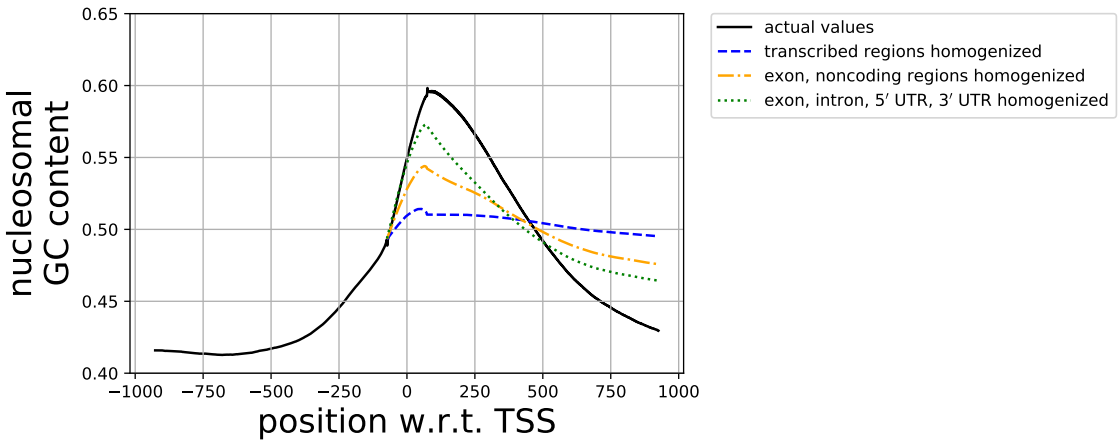


(a)

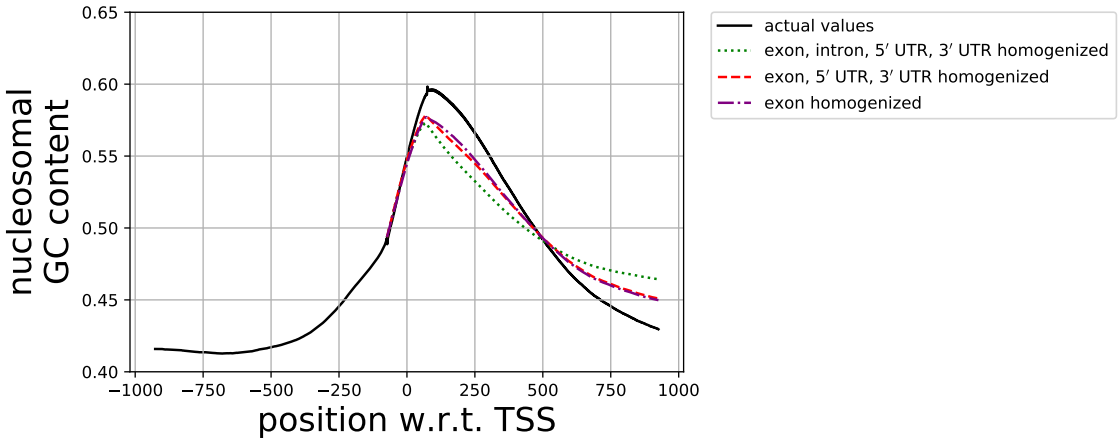


(b)

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

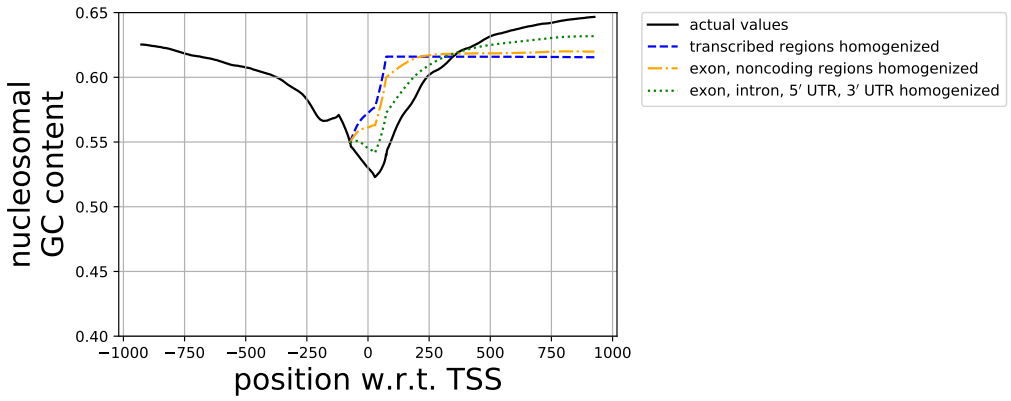


(a)

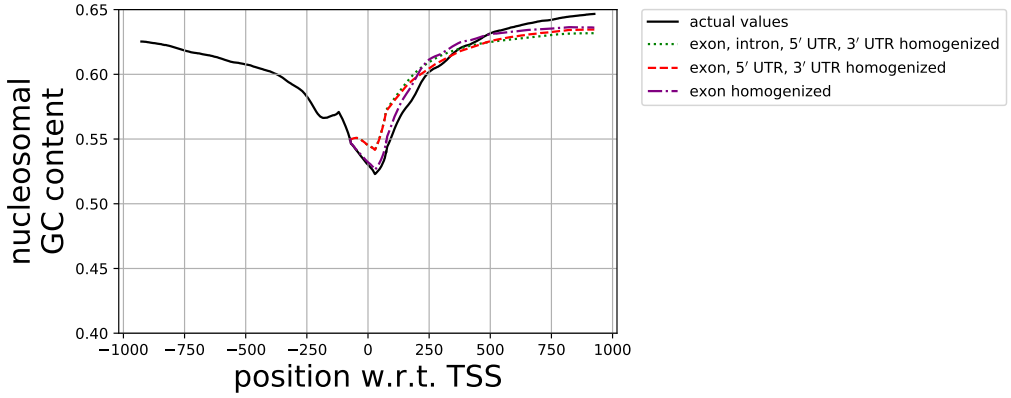


(b)

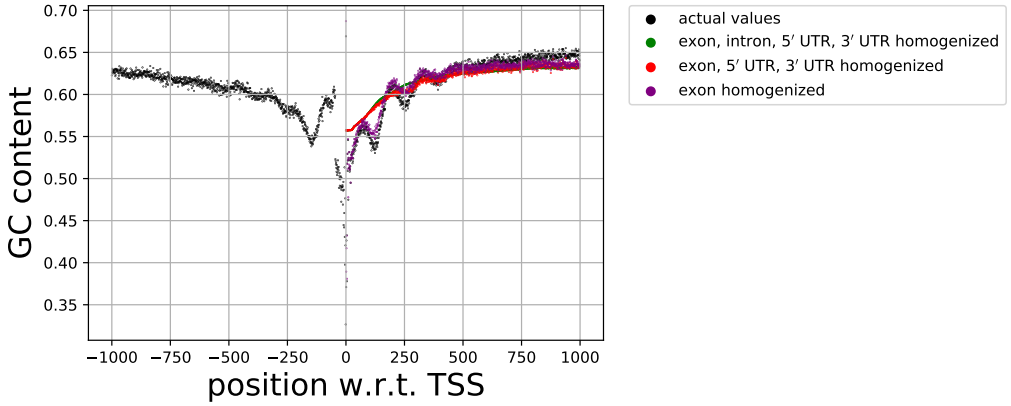
Figure 5.18: Same as Figs. 5.12 and 5.17 but for *L. perrieri*, a cutgrass from Madagascar closely related to rice but not used for consumption. This plant shows only a slightly weaker signal compared to the cultivated grains such as rice, suggesting that cultivation has not lead to significantly stronger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



(a)



(b)



(c)

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

5.6 Even the amino acid sequence contains nucleosome signals

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

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

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

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

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

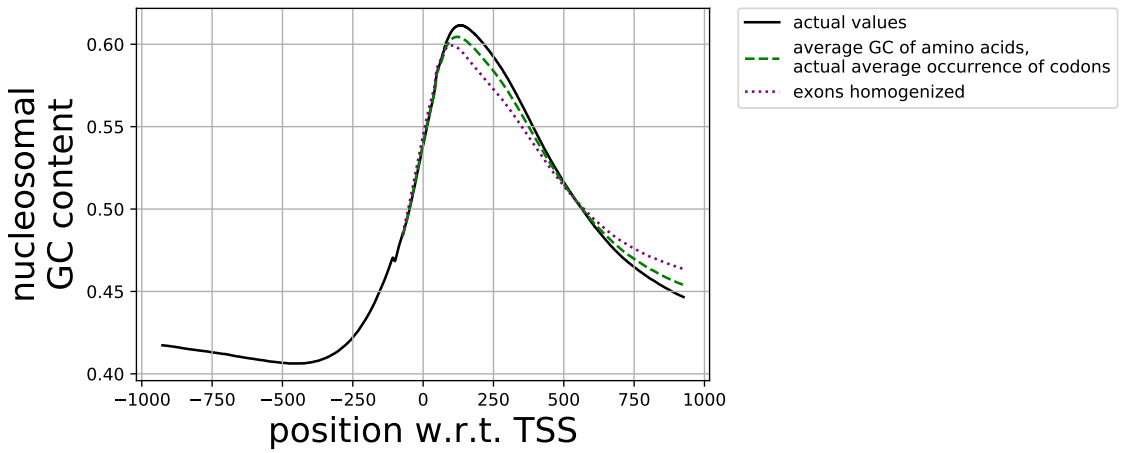


Figure 5.20: This figure describes *Oryza sativa*, rice. We see in black (solid line) a depiction of the actual average nucleosomal GC content around the TSS of protein-coding genes. In green (interrupted line), all codons on all exons have been replaced by the weighted average GC content of the amino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

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

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

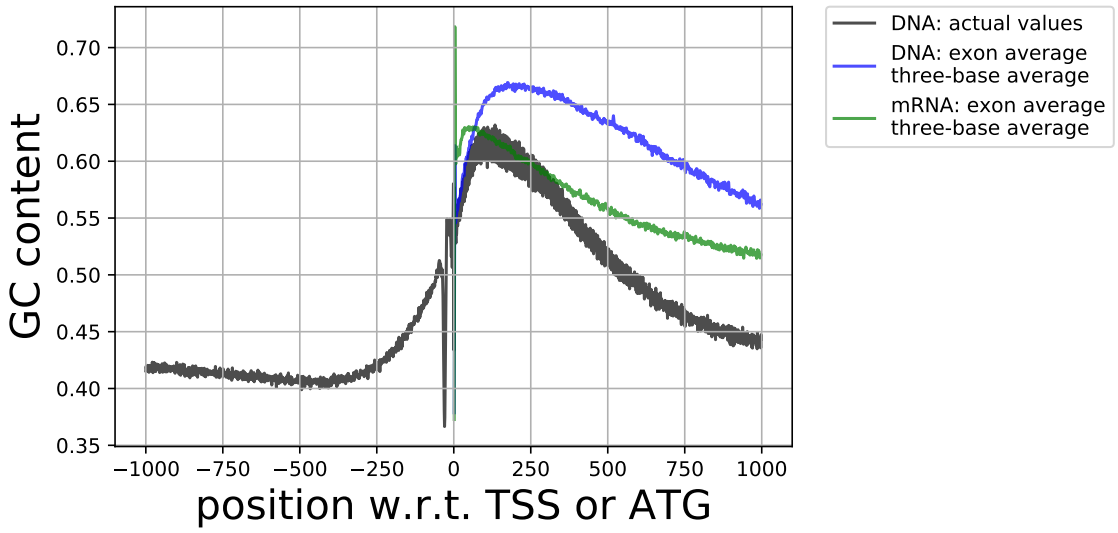


Figure 5.21: In black we show the average GC content per base pair for *O. sativa*, rice, around the TSS. We also depict the average GC content of exons in blue. This curve is quite different from the actual GC landscape since it excludes interregional and noncoding intraregional signals. The green curve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

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

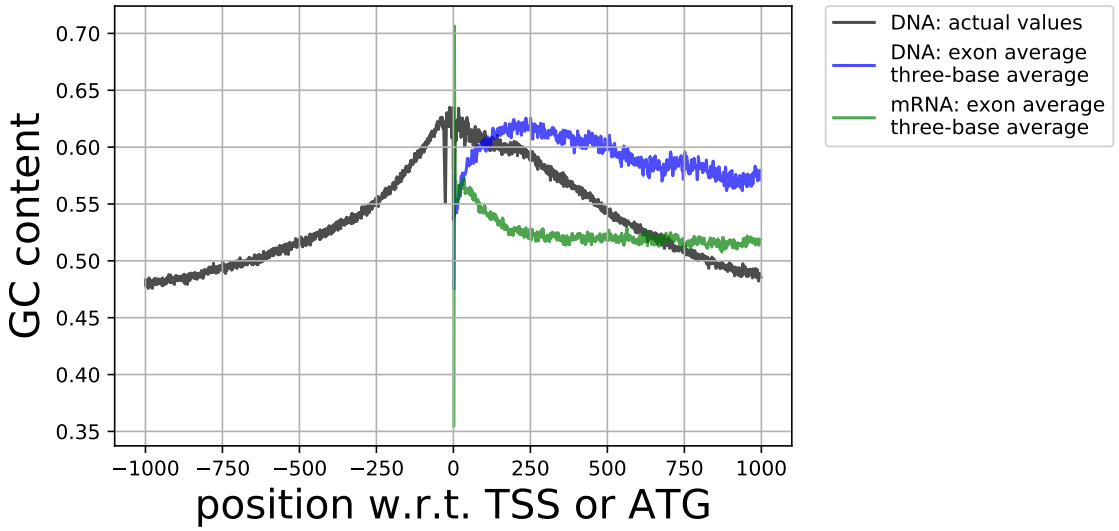


Figure 5.22: Same as Fig. 5.21 but for human. There we find an opposite result compared to rice: the mRNA signal dwarfs the DNA signal. This suggest that the exon signal on the DNA is a result of a functional GC signal on the RNA, likely related to translation speed.

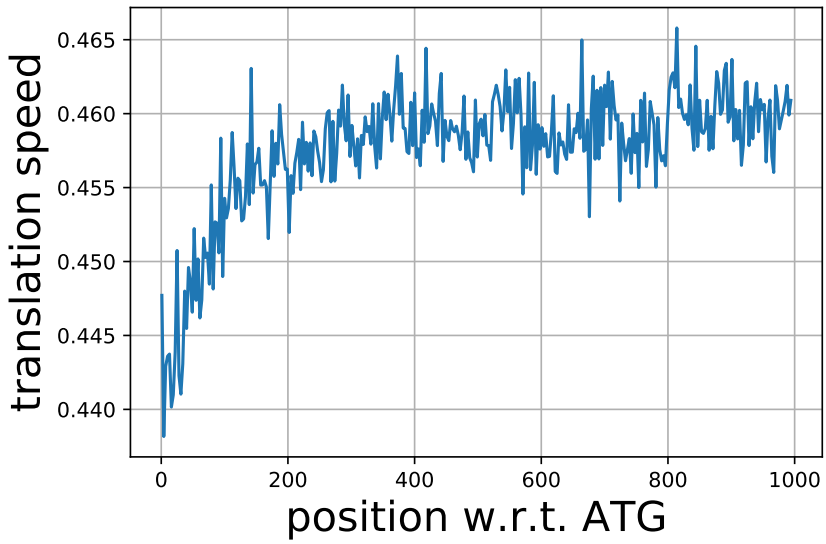


Figure 5.23: The average translation speed landscape for human is shown for a range of positions after (not including) the start codon. The mRNA GC signal in Fig. 5.22 seems related to this landscape, which contains a ramp in the first 300 bp. It resembles the translation speed ramp that has been suggested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

5.8 Conclusions and Outlook

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

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

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

introns have nucleosome positioning signals encoded on their exons.

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