Biophysical Journal, Volume *112*

Supplemental Information

Genomes of Multicellular Organisms Have Evolved to Attract Nucleosomes to Promoter Regions

Marco Tompitak, Cédric Vaillant, and Helmut Schiessel

Supplementary Methods

Model: The model used in this work to predict nucleosome affinity is based on that of Segal *et al.* (1), which is a model for the thermodynamic probabilities for 147-base-pair sequences to reside in a nucleosome. That is, it provides a method to calculate the probability *P(S)* of a sequence *S* related to the energy cost *E* of using a DNA molecule with this sequence to form a nucleosome:

(1) $$P(S) \propto e^{-E/kT}$$

This probability depends on every one of the nucleotides that make up the sequence *S*. If we define *S* as a set of S_i with *i* an index running from 1 to 147, we can write

(2) $$P(S) = P(\bigcap_i S_i)$$

Using the chain rule of probabilities, this can be rewritten as

(3) $$P(S) = \prod_{n=1}^{147} P(S_n | \bigcap_{i=1}^{n-1} S_i)$$

This equation expresses the probability of the whole sequence as simply the product of all the separate base pairs in the sequence. The catch is that the probabilities of the base pairs are all interdependent; the probability for S_n depends on the values of S_1 through S_{n-1}.

The way the model of Segal *et al.* is obtained is by assuming that long-range correlations between base pairs can be neglected in the expression above. Specifically, they assume that the probability distribution of S_n depends only on the value of S_{n-1} and not on any base pairs further away, so that

(4) $$P(S_n | \bigcap_{i=1}^{n-1} S_i) \approx P(S_n | S_{n-1})$$

If we apply this assumption, we obtain the model of Segal *et al.*

For the model to make predictions, it needs to be parameterized. Segal *et al.* and follow-up work (1–3) produced experimental thermodynamic ensembles of sequences with high affinity for nucleosomes. The probability of a given sequence in such an ensemble should be described by the model above, so one counts the prevalences of the dinucleotides and mononucleotides at every nucleosomal position in this average to produce the probability distributions needed to inform the model.

We here repurpose this model for a somewhat different endeavor. Another common approach to investigating nucleosome affinity is to model the energetics of the nucleosome directly. This can be done with a DNA model such as the Rigid Base Pair model (4) and a suitable model for the nucleosome. We have made use here of the nucleosome model presented in (5). This model can also be used to predict nucleosome affinity, based on the local elastic properties of base pair steps. Unfortunately, this model is computationally very expensive and cannot be used to analyze large numbers of sequences, such as entire genomes.

Such a model can, however, in a reasonable amount of time, be used to generate sequence ensembles of the same kind as employed to parameterize the Segal *et al.* model and follow-ups. Using a recently published computational method (Mutation Monte Carlo, (5)) we were able to generate ensembles large enough that probability distributions of mono-, di- and even trinucleotides could be calculated. When we plug those distributions into the Segal *et al.* model, we find that we have a good approximation of the predictions made by the full underlying nucleosome model, which is computationally far less expensive and allows us to analyze whole genomes.

We finally note that we used not the dinucleotide-based model of Segal *et al.*, but we have extended it to trinucleotides:

(5)
$$P(S) = P(S_1)P(S_2|S_1)\prod_{n=3}^{147} P(S_n|S_{n-1} \cap S_{n-2})$$

In this case, we make the assumption that the probability of S_n depends on the values of S_{n-1} and S_{n-2}. This assumption on the correlations between base pairs is less stringent than that of the dinucleotide model and should therefore provide a better approximation. The downside is that many more probability values need to be calculated, and a correspondingly larger sequence ensemble is required. However, we found that we were able to create a large enough ensemble (10^7 sequences) that the trinucleotide model provided a significant improvement over the dinucleotide model. When predicting the affinities of all 147-base-pair subsequences of the first chromosome of *S. cerevisiae*, the trinucleotide model came to a root-mean-square deviation of 0.85 kT when comparing its predictions with those of the underlying energetic model. The dinucleotide model yielded a deviation of 1.08 kT, so the trinucleotide model reduces the deviation by about 20%.

For the underlying nucleosome model, we chose the same model presented in (5). However, we have made an important alteration to the model in order to perform the analyses presented here. Previously, a hybrid parameterization was chosen for the Rigid Base Pair Model (6) that underlies the nucleosome model presented there. In this hybrid parameterization (7), the intrinsic deformations of the base pair steps are derived from crystal-structure data, and the stiffnesses of the steps from all-atom molecular dynamics similations.

This hybrid model had previously been found to approximate reality best by Becker *et al.* (7). Those authors, however, used only short sequences to test the different parameterizations. Hence they primarily tested the local accuracies of the parameterizations, for which the correct oscillatory behavior of the predicted energy with the helical repeat of DNA is most important.

However, we are interested not in the local changes in affinity, but in long-range effects on the order of tens of helical repeats. For this purpose, we found that the hybrid parameterization yields unsatisfactory results. Although it gives correctly phased dinucleotide probability distributions, the average abundances of AT-rich dinucleotide steps in high-affinity sequences are overestimated with respect to those of GC-rich steps. It is known that high GC content correlates with high affinity, but the hybrid model ascribes higher affinity to AT-rich sequences. See Fig. S1. The result is that the model is unable to detect the nucleosome-depleted regions in *S. cerevisiae* promoters.

We find that when using a parameterization where both the intrinsic deformations and the stiffnesses are derived from crystal-structure data (4), the model does correctly ascribe high affinity to high GC content. See Fig 1b. When using this pure parameterization for our model, we find we do detect the NDR in yeast.

We speculate that the two parameterizations can fulfill complementary roles. The hybrid model may be most accurate when considering local changes in affinity, but its performance in detecting long-range effects is lacking. Conversely, the pure crystallography parameterization may not be as realistic locally (7), but it is able to capture long-range effects much more accurately. For this work we therefore applied the pure parameterization.

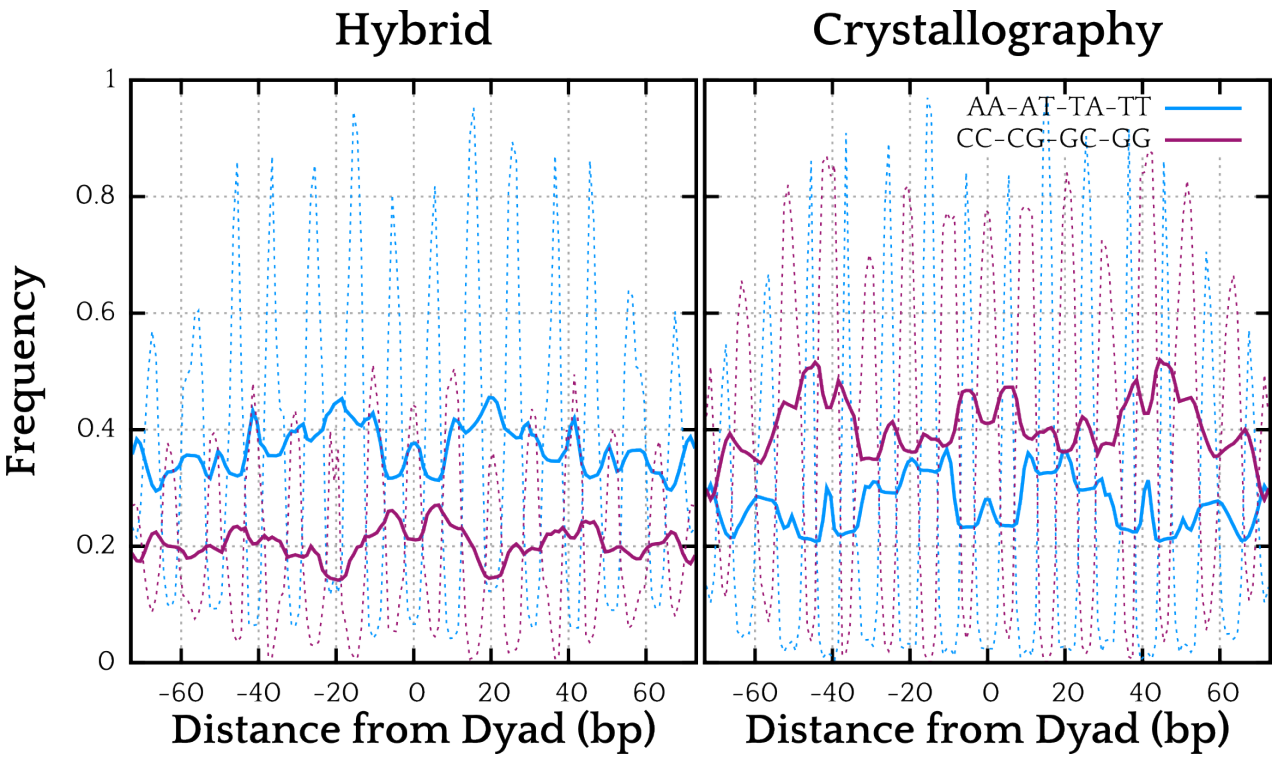

Fig. S1: Dinucleotide step frequencies and their 11-bp averages in high-affinity nucleosome ensembles. Left: Using the hybrid parameterization, AT-rich dinucleotide steps are enriched, while GC-rich steps are depleted. Right: In the pure parameterization, GC-rich steps are enriched, in line with experimental evidence.

Supplementary Results

Full set of mechanical signals: In this section we supply the full set of nucleosome positioning signals centered on transcription start sites. The signals are plotted in Figs. S2-S7, with organisms grouped together under a number of headings.

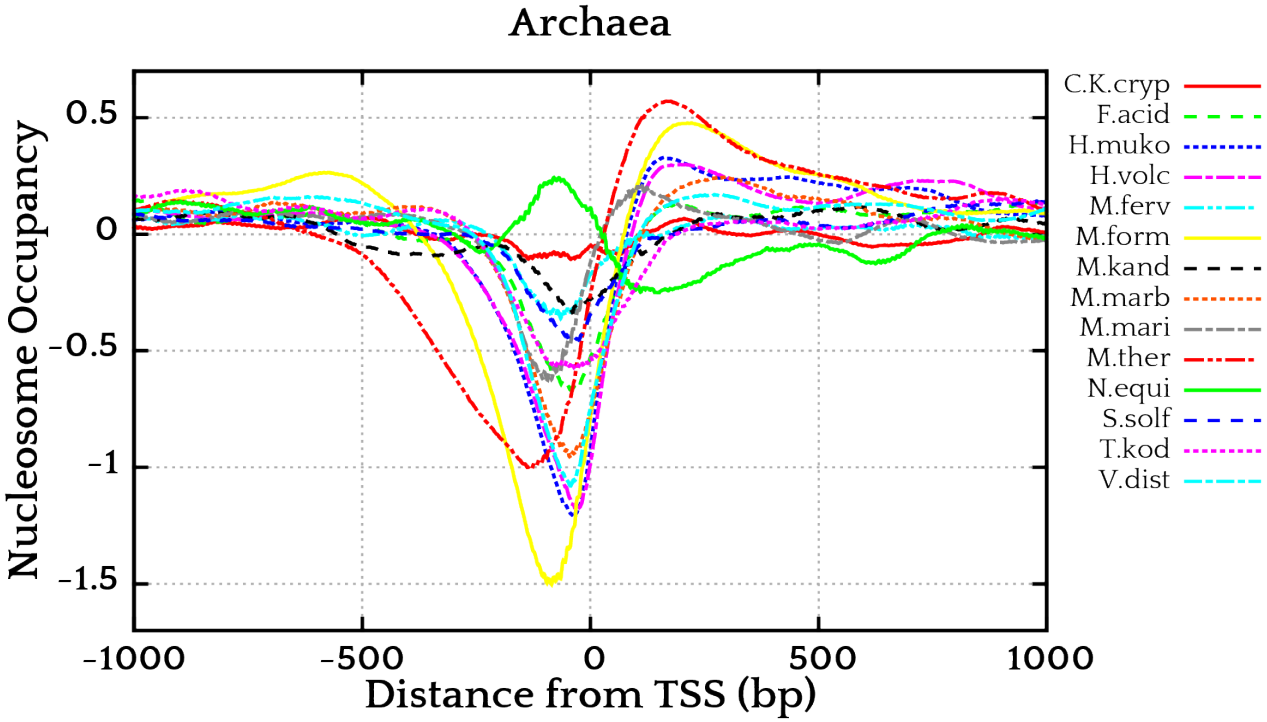

Fig. S2: Nucleosome positioning signals in the promoter regions of a number of Archaea.

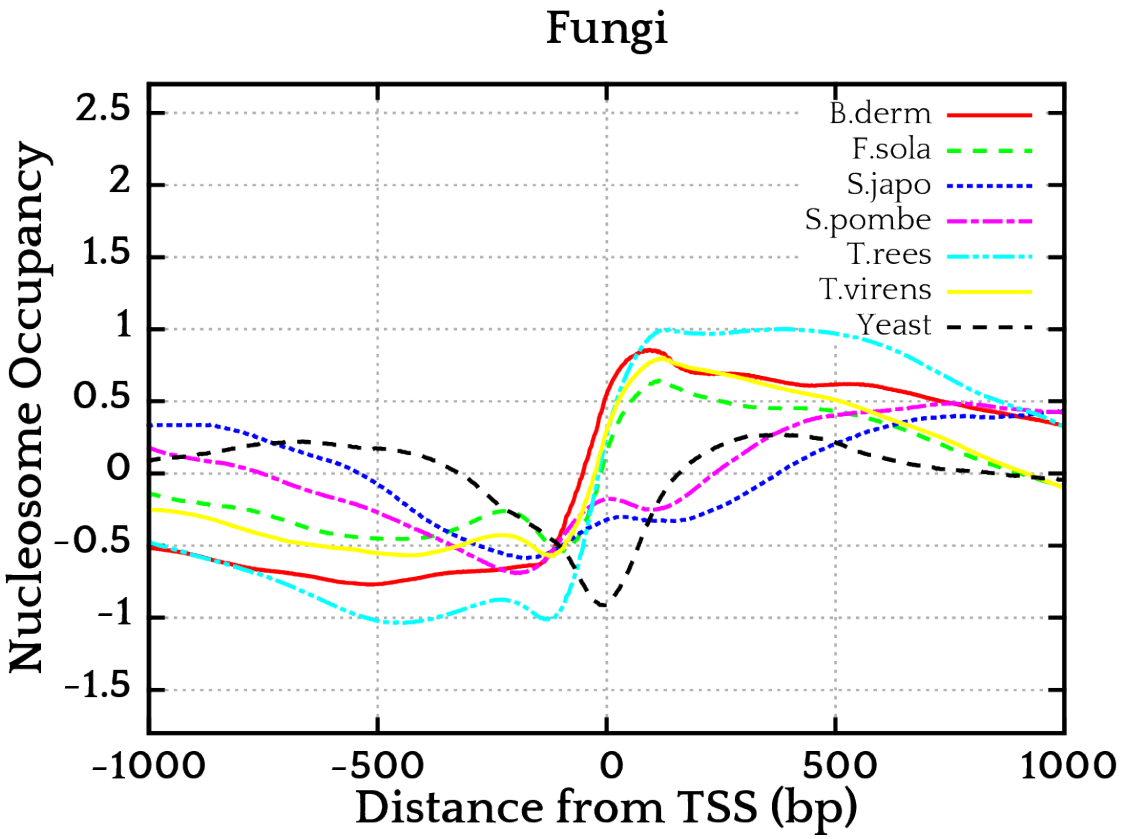

Fig. S3: Nucleosome positioning signals in the promoter regions of a number of fungi.

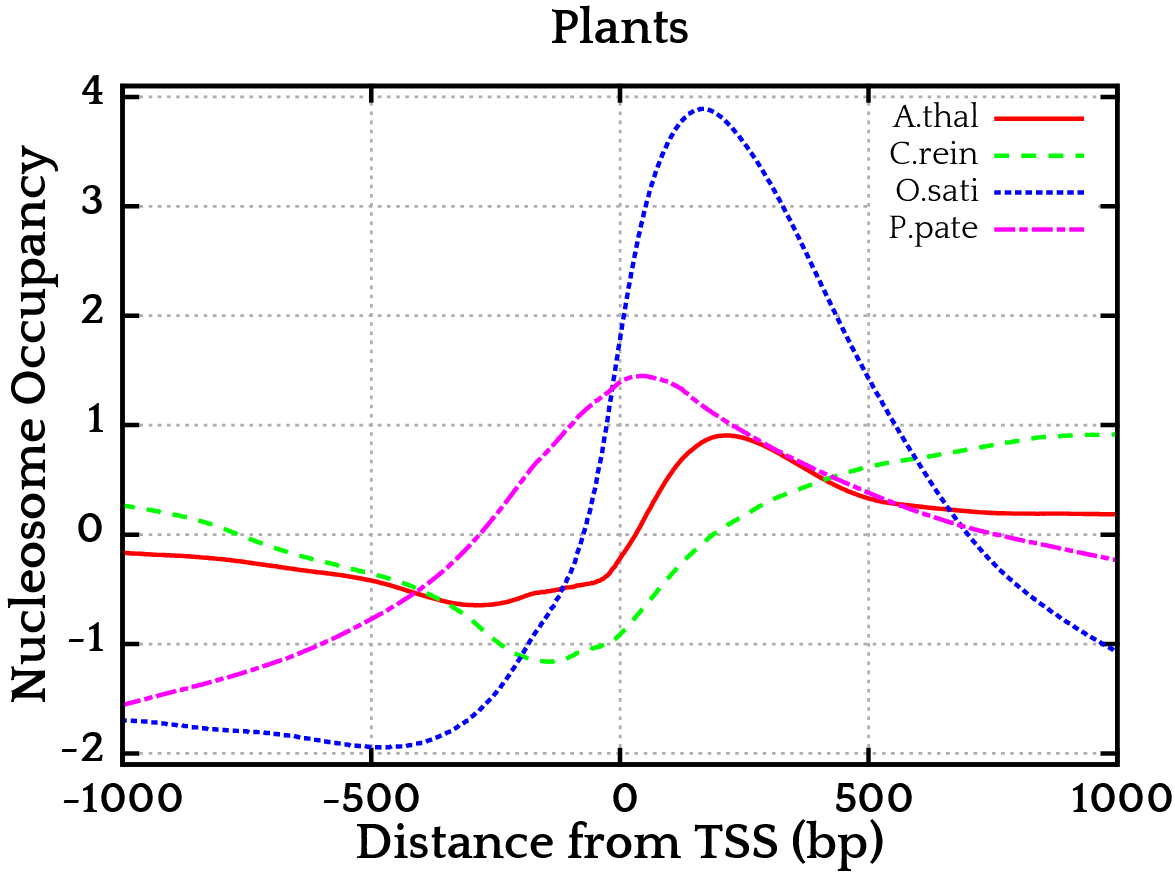

Fig. S4: Nucleosome positioning signals in the promoter regions of a number of plants.

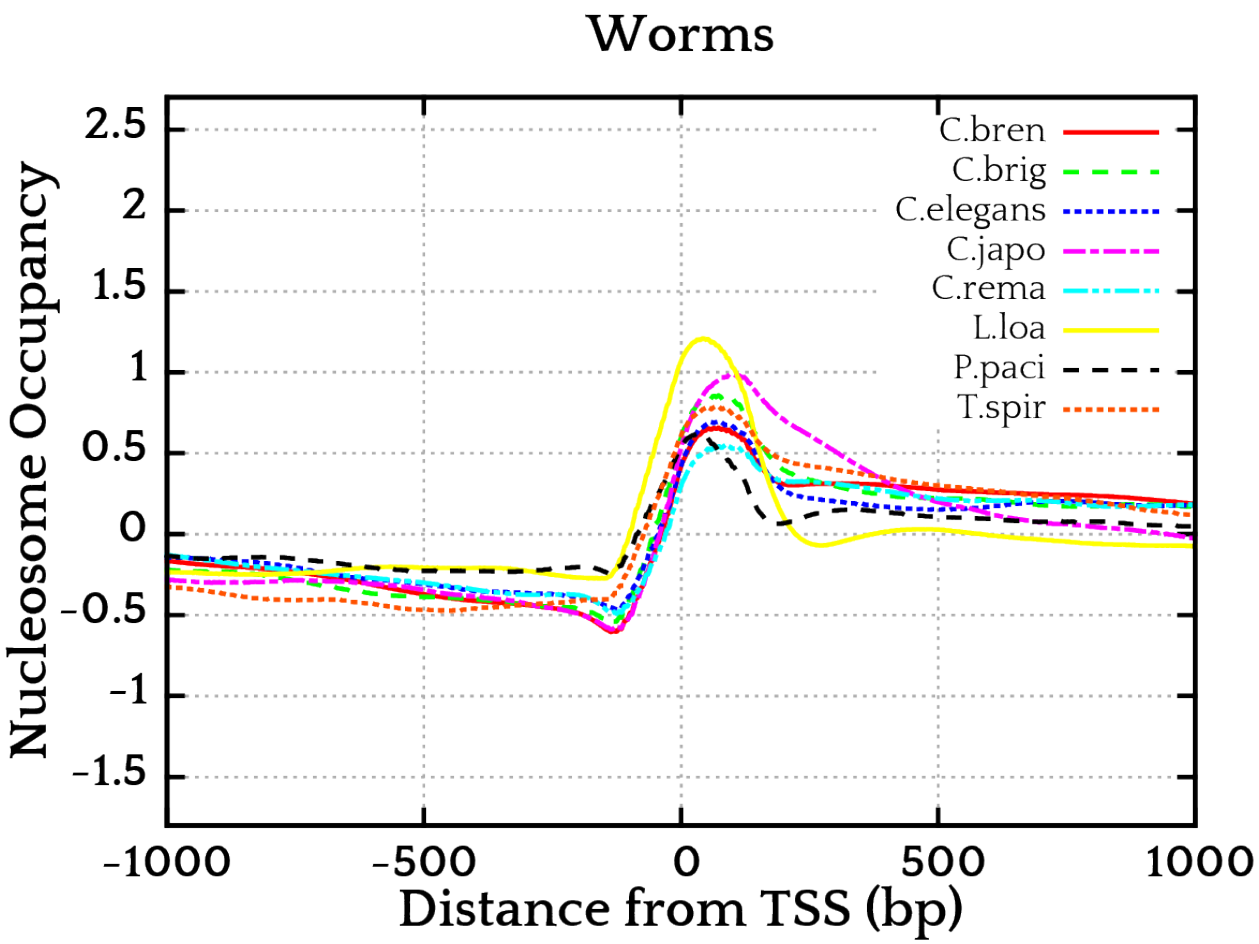

Fig. S5: Nucleosome positioning signals in the promoter regions of C. elegans and a number of other nematodes.

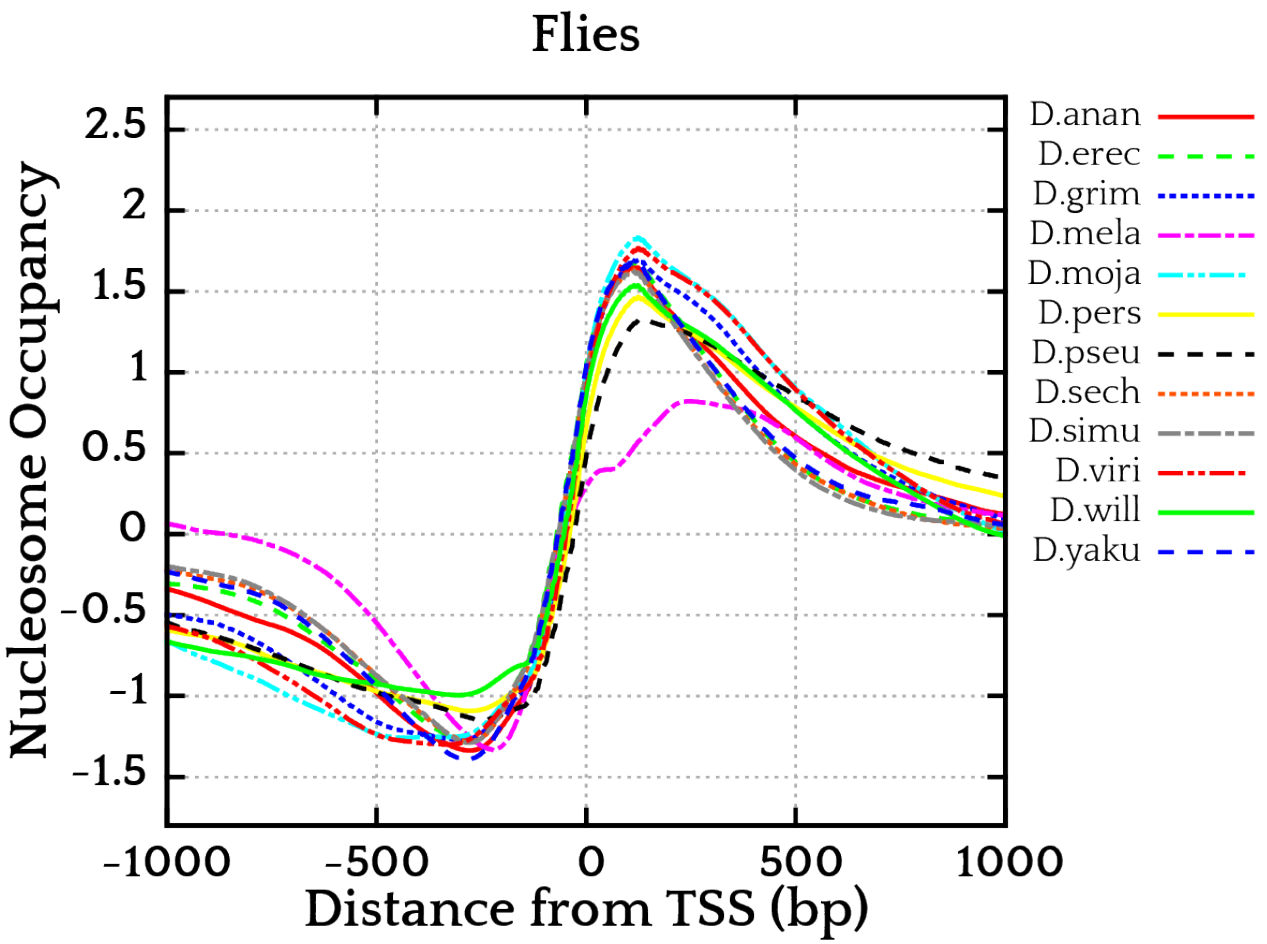

Fig. S6: Nucleosome positioning signals in the promoter regions of D. melanogaster and a number of other flies.

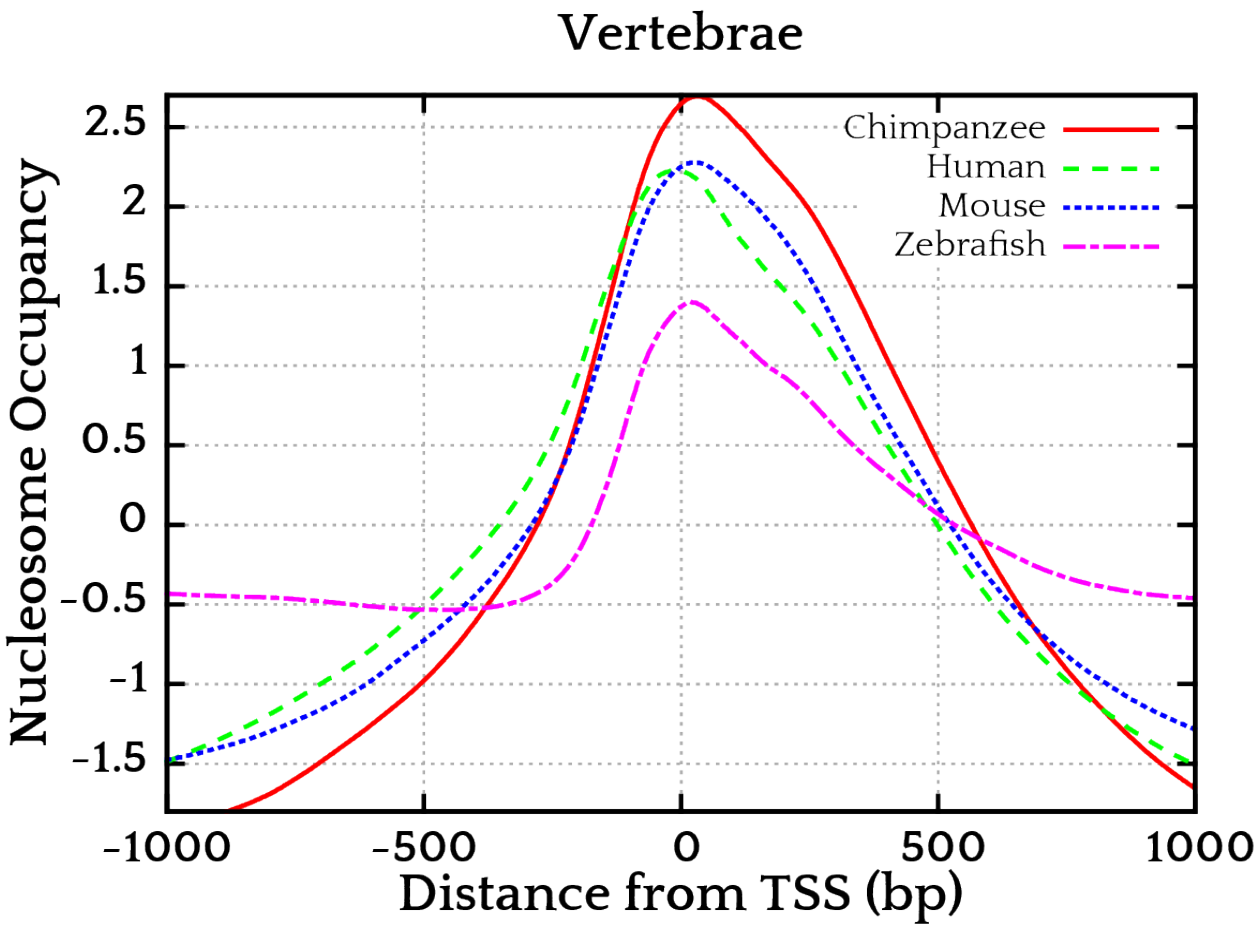

Fig. S7: Nucleosome positioning signals in the promoter regions of human, chimpanzee, mouse and zebrafish genomes.

GC content as signal predictor: Finally we wish to note that, in terms of classifying these signals as we have done in Fig. 2 in the main manuscript, one might also look at the signals in the GC content, which are depicted in Fig. S8. The visual similarity with Fig. 2 is of course striking.

We would warn against relying on GC content alone for the purpose for which we have applied our model here. The first reason is that, obviously, GC content in itself does not tell us anything about the numerical values of the nucleosome occupancy without some sort of calibration. Our model, on the other hand, has no free parameters, and is built on physical principles.

Secondly, we have also found that, using the Mutation Monte Carlo method with the Eslami-Mossallam nucleosome model [5], we can create sequences with very different mechanical properties by only changing the order of the sequence, while keeping GC content fixed, which shows that GC content is only part of the story.

That said, statistically, signals in GC content in promoter regions may also be a fruitful way to classify organisms. This will require further study.

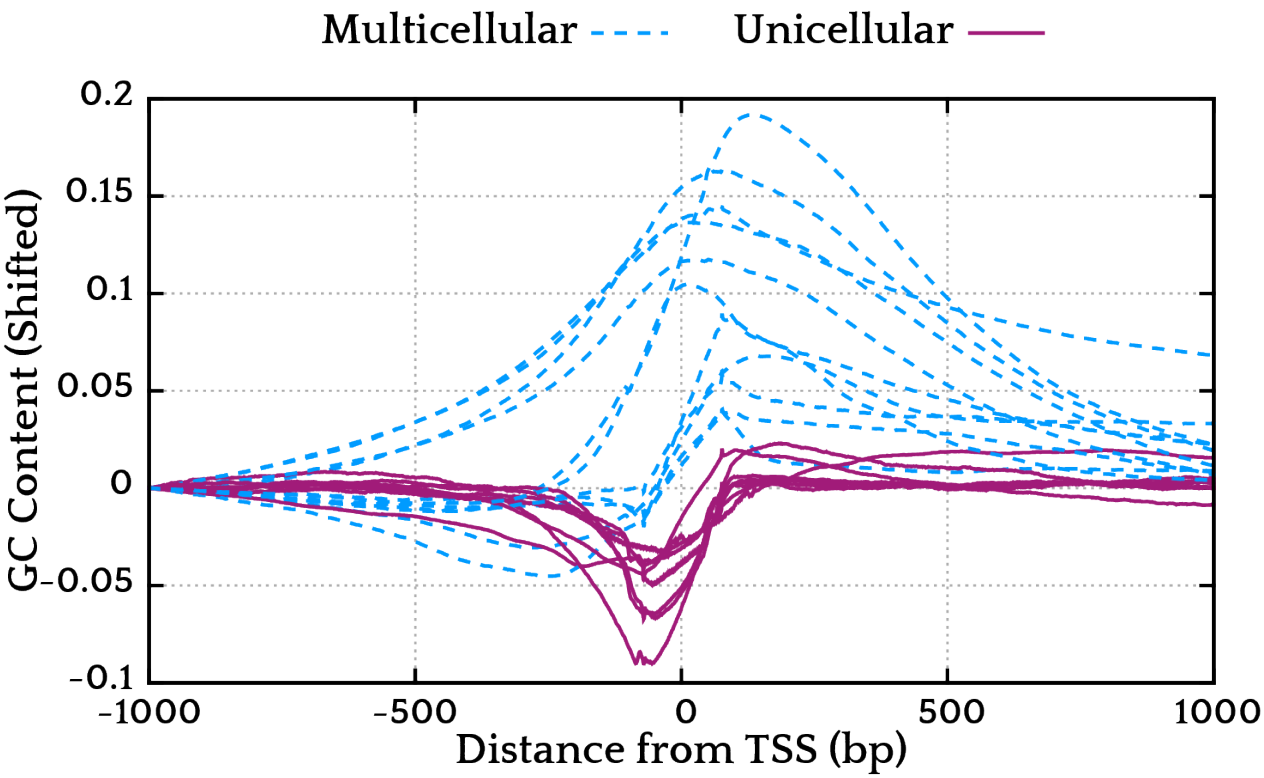

Fig. S8: Average GC content around the transcription start sites for the same organisms as presented in Fig. 2. Curves have been shifted such that the value at -1000 is zero, and have been smoothed using a 147-bp running average.

1. Segal, E., Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I.K. Moore, J.-P.Z. Wang, and J. Widom. 2006. A genomic code for nucleosome positioning. Nature. 442: 772–8.

2. Field, Y., N. Kaplan, Y. Fondufe-Mittendorf, I.K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput. Biol. 4.

3. Kaplan, N., I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, E.M. Leproust, T.R. Hughes, J.D. Lieb, J. Widom, and E. Segal. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 458: 362–366.

4. Olson, W.K., A.A. Gorin, X.J. Lu, L.M. Hock, and V.B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc. Natl. Acad. Sci. U. S. A. 95: 11163–11168.

5. Eslami-Mossallam, B., R.D. Schram, M. Tompitak, J. van Noort, and H. Schiessel. 2016. Multiplexing Genetic and Nucleosome Positioning Codes: A Computational Approach. PLoS One. 11: e0156905.

6. Calladine, C.R., and H.R. Drew. 1984. A base-centred explanation of the B-to-A transition in DNA. J. Mol. Biol. 178: 773–782.

7. Becker, N.B., L. Wolff, and R. Everaers. 2006. Indirect readout: Detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. Nucleic Acids Res. 34: 5638–5649.