**Characterizing the structural landscape of core promoters across eukaryotes**

# Abstract

Promoter regions are crucial for gene regulation in eukaryotes. Despite their functional importance, computational methods to identify eukaryotic promoters exhibit poor accuracy when applied to diverse species. This lack of success has been attributed to weak selective pressure on sequence identity, limiting the potential for identification through conserved sequence features. Recent studies suggest that the structural features of promoter DNA, rather than the precise nucleotide sequence, underpin promoter functionality. Specifically, the three-dimensional shape of the promoter DNA itself determines its binding affinity to transcription factors and polymerases. In this project, I will investigate whether conserved DNA structural features are found in eukaryotic core promoters and if these features align with known promoter motifs and are predictive of promoter activity and gene expression level. I will further examine the conservation of 3D structural features across eukaryotes. If conserved shapes are identified, I will attempt to develop a computational method to predict promoter locations in raw DNA sequences. Thus, this project has the potential to reveal previously hidden features of eukaryotic core promoters, provide novel methods for promoter discovery, and shed new light on the evolutionary dynamics of promoter regions.

# Background

***Transcription of DNA to RNA is a key step in the manifestation of genotype as phenotype.***

Transcription of DNA into RNA is the first step in the conversion of genotype to phenotype. It is a dynamically regulated process that adapts in response to internal and external stimuli in order to control an organism’s growth and development. In eukaryotes, the transcription of genes is performed by different classes of DNA-directed RNA polymerases (RNA pol), each dedicated to synthesizing specific types of RNA. While RNA pol I and II transcribe rRNA, tRNA (and other small or long non-coding RNAs), RNA pol II transcribes protein-coding genes, and thus is primarily responsible for the conversion of genotype to phenotype. Initiation of transcription with RNA polymerases is tightly regulated at specific sites in DNA sequences known as core promoters - which recruit RNA polymerase to the transcription start site (TSS) [1].

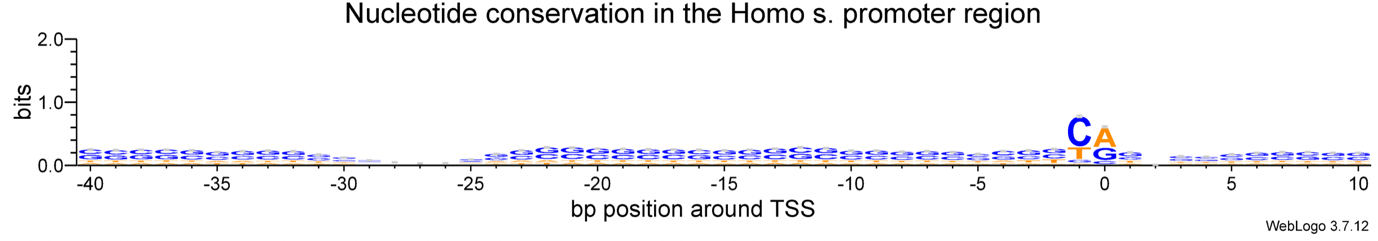
Core promoters achieve this precise control by serving as binding platforms for the assembly of preinitiation complex (PIC), which includes various general transcription factors that position transcription machinery at the TSS. By determining the precise location of the TSS, promoters control the expression of genotype to phenotype, affecting both the precise sequence and amount of RNA produced from each gene. Thus, knowledge of the location of core promoters is important for our understanding of the anatomy of the genes (including their untranslated regions) and the composition of the genome.

***Promoters are functionally diverse playing a role in differential gene expression.*** In addition to knowing where promoters are, to truly be able to predict phenotype from genotype, we need to be able to understand how promoters are regulated. The regulation of core promoters is influenced by combinations of specific motifs, which act as binding sites for transcription factors. Motifs, such as the TATA box, CCAAT box, CpG islands, TFIIB recognition elements (BRE), Initiators (Inr), and downstream promoter elements (DPE) [2, 11], act in a combinatorial manner, creating a complex regulatory landscape that influences the expression of a gene [3]. Consequently, despite sharing the same transcription machinery, certain genes exhibit distinct transcription patterns, with some genes showing higher expression rates or more ubiquitous expression across different issues [4, 10].

Despite well-documented motifs, we are still unable to predict the expression level of a gene from the presence or absence of motifs within promoters. However, there are some associations between certain motifs and promoter regulation. For instance, housekeeping genes, which are essential for basic cellular functions, often contain promoters with CpG islands, that are rich in unmethylated cytosine and guanine nucleotides, maintaining an open chromatin structure conducive to continuous transcription and on average lower expression rates of genes [5]. On the other hand, tissue-specific and inducible genes may contain promoters with TATA boxes, which bind TATA-binding proteins (TBPs) to initiate transcription in a more regulated and conditional manner [2, 6, 7]. The presence of these motifs and their combinations play a critical role in the differential regulation of gene expression [8].

Thus, the core promoter encompasses features that extend beyond merely initiating transcription; combinations of promoter elements correlate with tissue- and process-specificity and gene expression abundance, reflecting their operational diversity [9]. Thus, in addition to knowing the location of promoters, determining their composition is essential for enabling the prediction of gene transcription patterns and the conversion of genotype to phenotype. Such prediction remains the biggest goal of modern biology.

***Promoters have poor sequence conservation leading to inability to predict their location or activity.*** Despite their functional importance, core promoters exhibit a surprising degree of sequence variability and show little conservation, even among closely related species. Specifically, many known motifs are extremely degenerate, often not indicative of promoter activity [10, 11]. For example, when analysing the sequence similarity across *H. sapiens* RNA-pol II promoter regions (*n* = 16453; -40 +10 bp around TSS), we observe a low degree of conservation, with only a dominant CA(CG) motif present at the TSS (*Figure 1*). This lack of conserved sequence features substantially limits our ability to computationally locate promoters and predict their activity. **Therefore, predicting the localization of promoters and the precise point of transcription initiation on raw sequences remains one a major challenge in computational biology.**



**TSS**

Figure 1 The nucleotide conservation in H. sapiens. The height of each stack in the logo represents the degree of conservation at each position, measured in bits. Promoter sequences retrieved from EPD, analysed in WebLogo3.

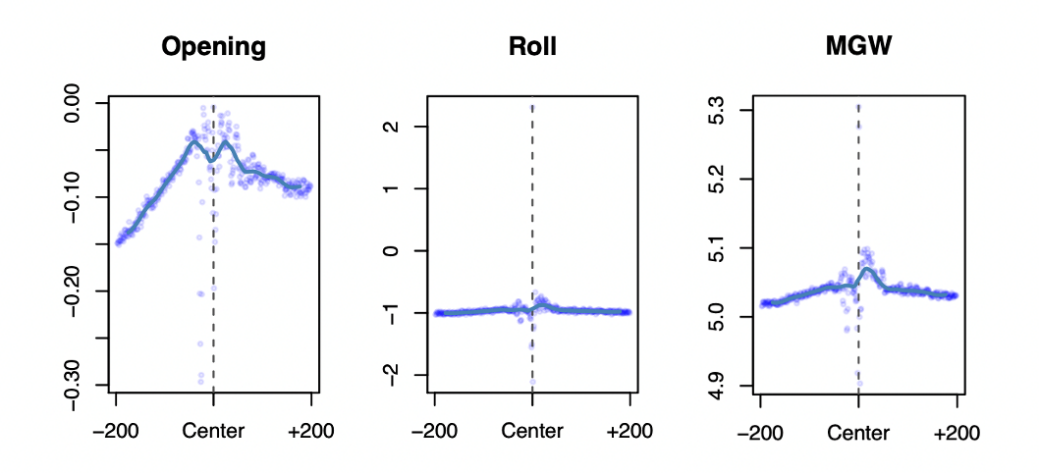
The diversity of promoter sequences, yet the precise initiation of transcription at a very specific position, emphasizes the need to consider features beyond mere sequence motifs when studying promoter function and regulation [12]. Traditional methods for identifying promoters, such as CAGE-seq and RNA-seq, are reliable but expensive and time-consuming. In the post-genomic era, with over 15,000 whole genome sequences available in databases like *Ensembl*, it is crucial to develop computational methods for efficient genome annotation. Accurate computational prediction of promoters will enable us to identify the precise location of the starts of genes and will provide a platform for predicting their regulatory mechanisms without relying on costly experimental validation.

***DNA shape prediction is a new technology that is providing insights into transcription.*** Recent advancements in DNA shape prediction have opened new avenues for understanding transcription regulation. Unlike linear sequence motifs, DNA shape features, such as minor groove width (“*MGW*”), angle (“*Roll*”) and degree of separation between nucleotides (“*Opening*”), and other structural parameters offer a three-dimensional perspective that can significantly influence transcription factor binding and promoter activity [13, 14]. Importantly, highly similar DNA shapes can be achieved with dramatically dissimilar sequences. Given that promoters are precisely defined by cells in the absence of conserved sequence features, we hypothesise that cells may be relying on DNA shape, instead of sequence identify, to specify the location of promoters and the point of transcription initiation. Accordingly, this project proposes to determine whether shape features in eukaryotic promoters are conserved and whether these shape features could be used to predict the location and activity of gene promoters [15, 16].

***Preliminary data suggests common DNA shape associated with motifs in human promoters.***

Preliminary data in the work done leading up to this proposal suggests the presence of common shape features associated with promoter regions in the human genome, indicating their potential as reliable markers for promoter prediction. More specifically, it has been revealed that certain shape features – *“Opening”* and *“Minor Groove Width (MGW)”* - change as a function of proximity to the TSS, while others – like “*Roll”* - remain constant, highlighting that some shape features, but not others, appear to show an association with transcription start sites. For example, on the plot the *Opening* parameter shows a distinct dip around the TSS, indicating tighter base pairing at this critical region. This feature suggests a conserved structural characteristic that may facilitate the binding of transcription factors and the formation of the transcription preinitiation complex.

Moreover, some local maxima and minima on the plot may be indicative of known motifs, like TATA, CCAAT, Inr, and DPE. Based on this data, many questions arise, including **whether these and other promoter shape features are conserved across genes. Whether they could be indicative of promoter function, including extent of gene expression or tissue specificity. And most importantly, whether they can be utilized to accurately predict the location of promoter regions and transcription start sites in raw DNA sequence data.** In this project, I will attempt to answer these fundamental questions.



*Figure 2 The consensus 3D structural features of RNA pol II promoter regions in Homo sapiens, predicted using deepDNAshape (n = 16,453). The -200 to +200 bp window is centred on the TSS (dashed vertical line) with. MGW – minor groove width. Opening parameter measures the degree of base pair separation in the DNA double helix, the Roll parameter reflects the inclination angle between adjacent base pairs, MGW measures the distance between the minor groove edges of the DNA helix.*

**AIM:** To assess the conservation of 3D shape in eukaryotic promoters and evaluate the role of DNA shape in the regulation of gene expression. This aim will be achieved through the following objectives:

**Objective A:** Establish the extent of 3D shape conservation in the validated set of *Homo sapiens* promoter regions.

**Objective B:** Determine whether the shape of a promoter region of a gene is predictive of its expression level and tissue-specificity in *H. sapiens.*

**Objective C:** Assess whether promoter shape features are conserved across species.

**Objective D:** Evaluate whether promoter regions and TSS can be predicted based on the DNA shape.

**Objective A: Establish the extent of 3D shape conservation in the validated set of *Homo sapiens* promoter regions. (2 weeks).** In this part of the project, I will collect and analyse a set of validated pol-II promoter regions. Promoter sequences will be downloaded from Eukaryotic Promoter Database. Next, I will predict the shape of these promoter sequences using the deepDNAshape package [14], which can accurately predict 14 key structural parameters, including minor groove width, roll, propeller twist, helical twist, and several other structural features of DNA directly from the DNA sequence itself. Given that promoter elements can be found both upstream and downstream of the TSS I will analyse promoter regions encompassing 2000 base pairs upstream and 2000 base pairs downstream of the TSS. Predicted structural features will then be analysed to determine the extent of conservation between genes and to evaluate whether conserved structural features are found proximal to core promoter sequence motifs (including the TATA-box, BRE, Inr etc). **Expected outcome:** Establish the landscape of 3D shape in human RNA pol II promoters, identifying common structural elements and their association with known sequence motifs.

**Objective B: Determine whether the shape of a promoter region of a gene is predictive of its expression level and tissue-specificity in *H. sapiens.*****(3 weeks).** This part of the project focuses on the promoters of genes that are preferentially expressed in certain tissues or exhibit ubiquitous expression. I will partition genes into expression deciles based on tissue-specific and overall expression levels, using databases like TiGER and Human Cell Atlas, which provide comprehensive human gene expression profiles. I will compare the promoters of genes within each decile to identify common structural features distinguishing high expression from low expression genes. Additionally, I will assess whether the distance from the consensus promoter shape, defined in Objective A, correlates with gene expression levels and tissue specificity. Statistical methods will be employed to analyze these correlations and determine how deviations from the canonical shape affect expression rates and tissue-specificity. **Expected Outcome:** Assess the relationship between promoter shape and gene expression levels, as well as their localization.

**Objective C: Assess the conservation of promoter shape across different species (2 weeks).** Having defined the landscape of promoter shape in objective A and identified what structural features are associated with variation in gene expression in objective B, I will next determine whether any of these newly identified structural features are conserved across species. Specifically, I will assess whether the structural features are found in 6 additional species *M. musculus, D. rerio, D. melanogaster, C. elegans, A. thaliana, and Z. mays.* To do this I will download the complete sets of experimentally validated promoter sequences from the Eukaryotic Promoter Database and determine their shape as described in objective A. I will then use distance metrics on the resultant shape profiles to determine the extent of shape conservation between species and analyse how shape varies along the phylogenetic tree that relates the species. **Expected Outcome:** Determine the extent of promoter shape conservation across species and its correlation with gene regulatory functions, highlighting evolutionary patterns and species-specific adaptations.

**Objective D: Develop predictive models for promoter regions and TSS based on DNA shape (8 weeks).** This part will combine structural parameters with sequence data in a machine learning model to predict the locations of promoters and TSS on raw DNA sequences. I will train a convolutional neural network models on datasets comprising both DNA sequence and structural data using PyTorch. The models will be trained using DNA structural parameters estimated from experimentally validated promoter sequences. Control non-promoter sequences will be generated by shuffling the nucleotide sequences of promoter sequences so that the nucleotide composition will remain constant. Randomly sampled regions of the genome of the same size will also be selected as control “non-promoter” sequences. Ten rounds of cross validation using randomly sampled data partitions will be performed to prevent over fitting. The final model will be benchmarked on its ability to detect promoter sequences in the complete genomes of the target species used in this study. **Expected Outcome:** Create and validate predictive models that can accurately identify promoter regions and TSS in genomic sequences based on DNA structural features, enhancing the accuracy of gene regulation studies.

**Project feasibility**: Preliminary analysis suggests that eukaryotic promoters exhibit several conserved DNA structural features, providing a robust basis for Objectives A, B, and C. i.e. structural features are present (*Figure 2*) and this project will describe the variation in those structural features, whether they correlate with expression level, and whether they are conserved across species. Thus, objectives A B and C are low risk and the outputs from these objectives will advance our understanding of the role of promoter DNA shape in the regulation of gene expression. In contrast, Objective D carries a higher risk of failure. This is because it may not be possible to train a predictive model that is able to identify promoter locations in raw DNA sequences based on structural features alone. However, even if this objective is unsuccessful this will not detract from the advances made in the delivery of objective A, B and C.

**Future development**: The results of this project will yield new insight into the composition and features of RNA pol II promoters in eukaryotes. However, all of the results in this project will be computational predictions that will require experimental validation. For example, if the predictive model is able to identify new promoter sites in genomes with no characterised promoters, then CAGE-seq can be used to validate these predictions. Similarly, if structural features are found that are predicted of gene expression level, then test promoters can be designed by mutating existing promoters to remove or add these structural features and then test their effect on the expression level of a reporter gene. Lastly, as the results could be helpful for genome annotation, developed methods could be combined to help in identification of non-coding regulatory elements on raw sequences and building of promoter maps for sequenced genomes that RNA-sequencing-based promoter characterisation.

**References:** 1. Haberle, V., & Stark, A. (2018). Molecular Cell Biology, 19(10), 621–637. 2. Kadonaga, J. T. (2011). WIREs Developmental Biology, 1(1), 40–51. 3. Bhattacharjee, S., Renganaath, K., Mehrotra, R., & Mehrotra, S. (2013). BioMed Research International, 2013, 1–11. 4. Zou, Y., Huang, W., Gu, Z., & Gu, X. (2011). Molecular Biology and Evolution, 28(10), 2893–2904. 5. Deaton, A. M., & Bird, A. (2011). Genes &amp; Development, 25(10), 1010–1022. 6. Basehoar, A. D., Zanton, S. J., & Pugh, B. F. (2004). Cell, 116(5), 699–709. 7. Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J. M., Bucan, M., & Stoeckert, C. J. (2005). Genome Biology, 6(4). 8. Moshonov, S., Elfakess, R., Golan-Mashiach, M., Sinvani, H., & Dikstein, R. (2008). BMC Genomics, 9(1). 9. Rach, E. A., Winter, D. R., Benjamin, A. M., Corcoran, D. L., Ni, T., Zhu, J., & Ohler, U. (2011). PLoS Genetics, 7(1). 10. Smale, S. T., & Kadonaga, J. T. (2003). Annual Review of Biochemistry, 72(1), 449–479. 11. Carninci, P., et al. (2006). Nature Genetics, 38(6), 626-635. 11. Bucher, P. (1990). Journal of Molecular Biology, 212(4), 563–578. 4. Nature Genetics, 38(6). 12. Gagniuc, P., & Ionescu-Tirgoviste, C. (2012). BMC Genomics, 13(1), 512. 8. 13. Yella, V. R., & Bansal, M. (2017). FEBS Open Bio, 7(3), 324–334. 14. Li, J., Chiu, T.-P., & Rohs, R. (2024). Nature Communications, 15(1). 15. . Misteli, T. (2007). Cell, 128(4), 787–800. 11. 16. Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., & Hume, D. A. (2007). Nature Reviews Genetics, 8(6), 424–436.