Study of DNA curvature in genetic regulation

Joan Estop Cepero

Resum— Identificar amb precisió els llocs d'unió dels factors de transcripció (TFBS) és un repte fonamental en biologia computacional. Mentre que els mètodes tradicionals es basen en gran mesura en la seqüència de nucleòtids, investigacions recents suggereixen que les característiques estructurals de l'ADN, com la curvatura, també poden tenir un paper clau en el reconeixement dels factors de transcripció (TF). Aquest estudi explora la contribució de la curvatura de l'ADN en la regulació de l'expressió gènica bacteriana analitzant la seva conservació a través de les regions TFBS i avaluant la seva utilitat en models de classificació binària. Utilitzant xarxes neuronals convolucionals (CNN), avaluem els perfils de curvatura per a TFBS validats experimentalment de dos factors de transcripció, NAC i Crp, en Escherichia coli. Els resultats mostren que la curvatura és una característica conservada i amb informació per als llocs d'unió de Crp, mentre que la seva contribució és mínima per a NAC. Tot i que els models entrenats només amb curvatura van aconseguir un rendiment modest, la incorporació de la curvatura amb informació de la seqüència de nucleòtids va conduir a petites però consistents millores, especialment en condicions desbalancejades. Aquestes troballes destaquen la rellevància de les característiques estructurals de l'ADN i donen suport a futures investigacions sobre la seva integració per a la predicció de TFBS.

Paraules clau—Bioinformàtica, Xarxes Neuronals Convolucionals, Anàlisi de seqüències de DNA, Regulació Gènica, Aprenentatge Computacional, Identificació de Motius, Factors de transcripció

Abstract— Accurately identifying transcription factor binding sites (TFBS) is a fundamental challenge in computational biology. While traditional methods rely heavily on sequence motifs, recent research suggests that DNA structural features, such as curvature, may also play a key role in transcription factor (TF) recognition. This study explores the contribution of DNA curvature in the regulation of bacterial gene expression by analyzing its conservation across TFBS regions and evaluating its utility in binary classification models. Using convolutional neural networks (CNNs), we assess curvature profiles for experimentally validated TFBSs of two transcription factors, NAC and Crp, in Escherichia coli. Results show that curvature is a conserved and informative feature for Crp binding sites, while its contribution is minimal for NAC. Although models trained with curvature alone achieved modest performance, incorporating curvature with nucleotide sequence information led to small yet consistent improvements, especially under imbalanced conditions. These findings highlight the relevance of structural DNA features and support further investigation into their integration for TFBS prediction.

Index Terms— Bioinformatics, Convolutional neural networks, DNA sequence analysis, Gene regulation, Machine learning, Motif identification, Transcription factors

1 Introduction

THE biology of living organisms can be understood in terms of information flow: genetic information is stored in DNA, transcribed into RNA, and ultimately translated into proteins (Mercatelli et al., 2020). This transfer of information from DNA to protein via RNA transcripti-on is known as the "central dogma" of molecular bio-logy (Lambert et al., 2018).

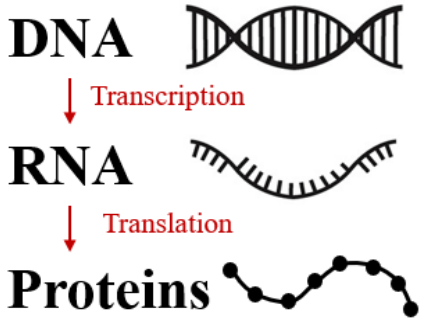

Fig. 1 The central dogma of biology describes the flow of genetic information within a cell: DNA is transcribed into RNA, and RNA is translated into protein. It explains how genes dictate cellular structure and function. This concept underpins molecular biology.

The complexity of living organisms arises from a multi-level, dynamically regulated network of gene expression (Brazhnik et al., 2002). This network not only shapes and maintains individual phenotypes but also modulates adaptive responses to environmental changes (Delgado & Gómez-Vela, 2019). A key component of this regulation is the action of transcription factors (TFs), which control gene expression by binding to regulatory DNA regions such as promoters and enhancers (Inukai et al., 2017). Each TF typically recognizes a set of similar DNA sequences, called motifs. Motifs are generalized sequence pattern representing a functional element. Transcription factor binding sites (TFBSs) are specific instances of a motif found in a particular DNA. Given the intricate nature of transcriptional regulation, accurately identifying TFBS is crucial for comprehending the regulatory mechanisms governing gene expression and cellular processes (Vahed, Ishihara, & Takahashi, 2019).

Due to the critical importance of identifying and discovering TFBSs, their discovery has been a long-standing challenge in computational biology. This has driven the development of nearly a hundred algorithms over the past 30 years (Vahed, Vahed, & Garmire, 2022), designed for both motif discovery and—most notably—motif identification. Motif discovery refers to the process of finding new

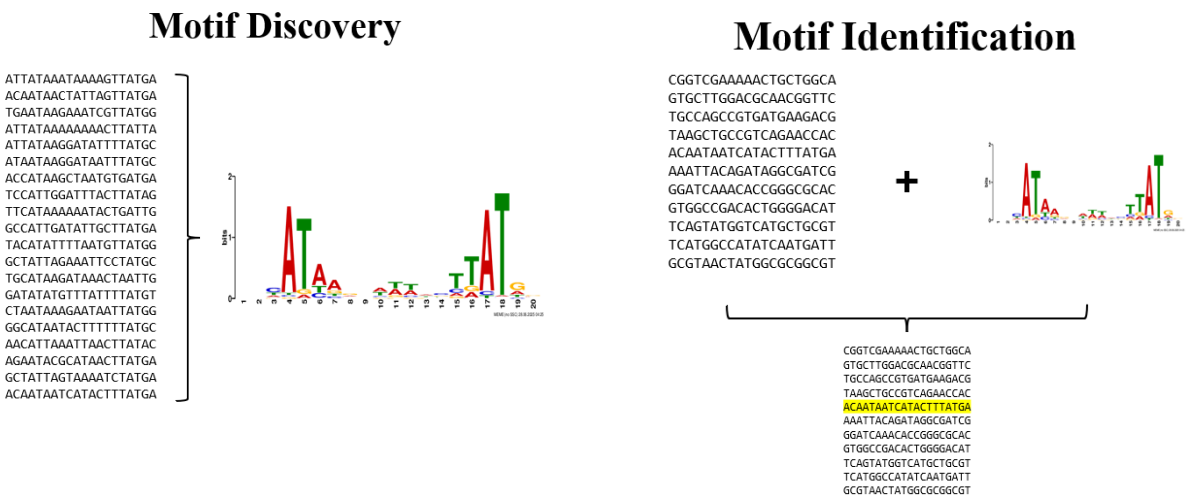

Fig. 2 Motif discovery and motif identification represent two distinct approaches in the analysis of regulatory sequence patterns. Motif discovery refers to the de novo identification of statistically overrepresented sequence motifs in a given set of sequences, without prior knowledge of the motif structure. It is typically used to uncover novel regulatory elements potentially involved in gene expression control. In contrast, motif identification (or motif scanning) involves searching for occurrences of known motifs—usually represented as position weight matrices—in target sequences, aiming to locate instances of established binding sites. While discovery is exploratory and unsupervised, identification is confirmatory and relies on existing motif databases.

sequence patterns in a set of DNA regions without prior knowledge, while motif identification involves scanning DNA sequences to locate occurrences of already known motifs (Georgiev et al.,2010). Characterizing motifs is a critical step in understanding the regulatory functions of TFs and their role in shaping gene regulatory networks (Inukai et al., 2017).

Despite advances in motif characterization, the current catalogue of TFBS motifs remains incomplete. The binding preferences of over 40% of the approximately 1,400 sequence-specific TFs in the human genome have yet to be determined, either experimentally or computationally (Hume et al., 2015). While traditional studies of TF–DNA binding have focused on TFs intrinsic preferences for specific nucleotide sequence motifs, recent research has uncovered additional layers of complexity (Siggers & Gordân, 2014), such as DNA curvature.

DNA is a double helix made by four nucleotides: Adenine, Cytosine, Guanine and Thymine. These nucleotides are paired in the double helix, as A-T and G-C pairs. The stacking (precise order in which they are in the sequence one after the other) of these nucleotide pairs on the genetic sequence creates local tensions that lead to local distortions in the shape and bendability of the DNA sequence (Bolshoy et al., 1991). Depending on how these distortions accumulate over relatively long stretches of the DNA sequence, the DNA sequence can adopt a predetermined curvature.

Experimental evidence suggests that DNA curvature plays a crucial role in regulating the transcription of multiple genes (Atlung & Ingmer, 1997). Moreover, genome-wide analyses have shown that promoter sequences tend to be significantly more curved than coding regions or randomly permuted sequences (Jáuregui et al., 2003).

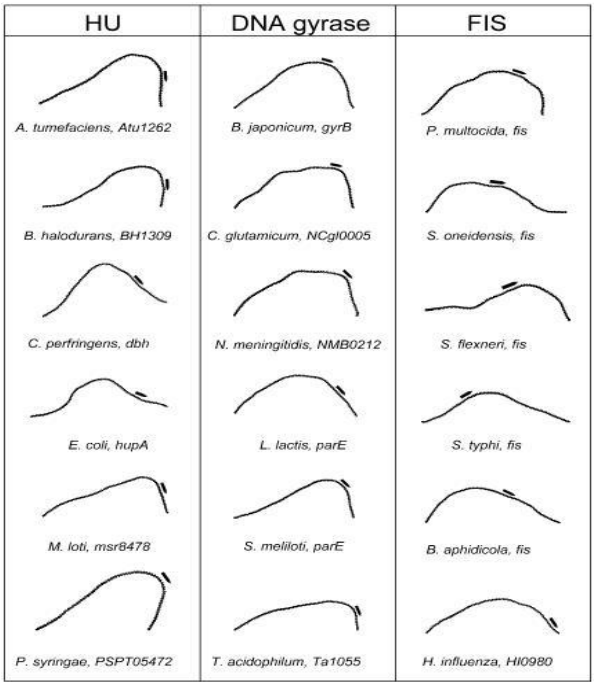

Fig. 3 DNA curvature plots of upstream regions for genes coding for DNA gyrase, HU, and FIS orthologs. A black arrow indicates putative promoter regions. As shown by Jáuregui et al. (2003), DNA curvature appears to be conserved across these orthologous genes.

In bacteria, several transcription factors have been shown to require intrinsic DNA curvature, in addition to specific nucleotide sequences, to accurately recognize and bind their target sites. For example, H-NS, a nucleoid-associated protein in Escherichia coli, preferentially binds to curved DNA elements in promoters, as demonstrated through atomic force microscopy and binding assays (Yamada, Muramatsu, & Mizuno, 1990; Dame, Wyman, & Goosen, 2000). Similarly, LRP and H-NS jointly recognize

curvature upstream of rrnB promoters, where only the phasing of the curved DNA module determines effective binding and functional repression (Pul et al., 2008). Another classic case involves CAP (CRP) in E. coli, which binds within intrinsically curved regions upstream of the lac and gal operons, facilitating transcriptional activation (Travers & Muskhelishvili, 1998). These studies collectively confirm that DNA shape—specifically curvature—is a functional determinant for bacterial transcription factor binding, complementing sequence-based recognition.

Also, in the 2010 study "Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli" (Meysman et al., 2010) intrinsic DNA curvature is recognized as a key regulatory feature. In that work, the authors developed CRoSSeD, a Conditional Random Fields model that incorporated both nucleotide sequence information and local DNA structural properties—including curvature and bendability—to predict TFBS in E. coli. They found that adding structural features significantly improved classifier performance across multiple regulons, enabling the identification of novel binding sites that were undetectable by sequence-based models alone. This confirms that, in prokaryotes, structural signals like curvature play a functional role in TFBS recognition, and that integrating them into computational models enhances motif identification.

The main objective of this project is to systematically analyse DNA curvature and its conservation in the promoter regions of Escherichia coli, and to explore its relationship with the presence of binding sites for known transcription factors. By focusing on DNA curvature, we aim to uncover patterns that may contribute to transcriptional control mechanisms in bacteria. To achieve this, the project is divided into the following sub-objectives: 1. Collect and preprocess genomic data from selected bacterial species. 2. Compute DNA curvature profiles across promoter regions using DNA curvature library. 3. Evaluate conservation of DNA curvature using machine learning models. 4. Visualize and interpret the results to better understand the potential regulatory roles of DNA structural features.

2 METHODS

2.1 Work Methodology

An adapted version of the Scrum methodology will be used to carry out the project. The process will be based on weekly or biweekly meetings to review the work completed during the previous week and plan the objectives for the upcoming one. After each meeting, it will be defined a list of tasks for the following week, specify how to approach each task, and estimate the time required for eachone. Throughout the week, daily check-ins (daily scrums) will be conducted to monitor the progress and adjust if necessary. At the end of each meeting, a short retrospective to analyse the previous sprint would be conducted, and apply those insights to improve the next cycle.

2.2 Data Collection

We obtained transcription TF binding motif data from RegulonDB, a database containing experimentally validated TFBSs for Escherichia coli. First, we identified the two TFs with the highest number of annotated binding sites: NAC and Crp. For both TFs, we retrieved the TFBSs from the TF PSSM database within RegulonDB.

For each TF, three distinct datasets were generated. The first dataset followed the procedure outlined below: using the extended sequences and the Biopython library, we obtained the consensus motif and constructed a PSSM. We then scored the original TFBS sequences using this PSSM and computed a threshold based on the 10th percentile of the real scores. Next, we scanned the entire E. coli reference genome using the PSSM, storing all sequences that scored above the defined threshold.

The second dataset consisted of random genomic sequences. Specifically, we sampled random DNA segments from the E. coli genome matching the number and length of the TFBS sequences obtained from RegulonDB. The third dataset was a combination of the first two datasets, including both predicted TFBSs and randomly selected genomic sequences.

For all three datasets, once the sequences were identified, we located their positions within the E. coli K-12 MG1655 reference genome, available in the NCBI GenBank database. We then extracted extended sequence regions by adding 100 nucleotides both upstream and downstream to enable further analysis of the surrounding sequence context. Using the DNAcurve library, we computed structural features such as DNA curvature and curvature angle for each extended sequence. We obtained 413 extended sequences corresponding to experimentally validated NAC TFBSs, and 177 extended sequences corresponding to experimentally validated Crp TFBSs.

2.3 DNA Curvature

We used the DNAcurvature library to calculate DNA curvature values for each extended TFBS sequence. This tool provides six different computational models to analyse the DNA curvature. Among them, we selected the AA Wedge model, as it has been shown to produce the most accurate results according to Tan and Harvey (1987). When applying this model, the output includes three main metrics: curvature, curvature angle, and bend angle. We focused on curvature and curvature angle, as the bend angle appeared visually noisy in the plots, while the other two metrics showed more stable and consistent patterns.

2.4 CNN

We implemented two convolutional neural networks (CNNs) to assess whether DNA curvature is a conserved feature at TFBSs, and to evaluate the contribution of curvature information in the classification of experimentally validated versus non-validated sites.

The first model is a simple CNN built using the Keras library. Its architecture consists of a one-dimensional convolutional layer with 32 filters and a kernel size of 5, using the ReLU activation function. This is followed by a max pooling layer with a pool size of 2 to reduce dimensionality. The output is then passed through a flatten layer, followed by a fully connected layer with 32 neurons and ReLU activation. Finally, the model includes a single output neuron with a sigmoid activation function for binary classification. The input to this model is a one-dimensional vector of 220 units, corresponding to the curvature angle profile obtained using the DNAcurve library.

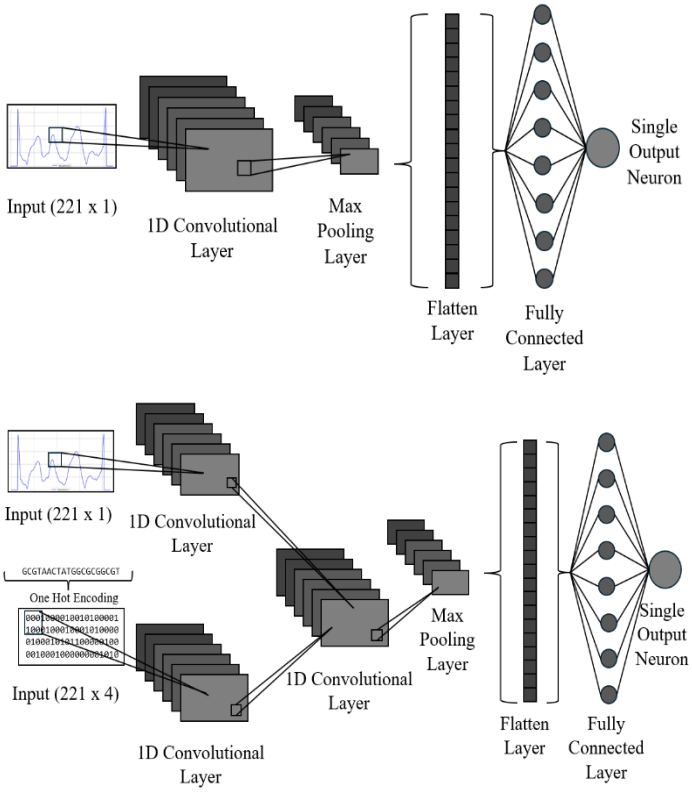

Fig. 4 Architecture of the CNN models implemented using Keras. (Top) The first model is a sequential CNN that receives a one-dimensional curvature input vector of shape (221, 1). It includes a Conv1D layer with 32 filters and kernel size of 5 (ReLU activation), followed by a max pooling layer (pool size 2), flattening, a dense layer with 32 units (ReLU), and a final sigmoid-activated output neuron for binary classification. (Bottom) The second model is a dual-input CNN that processes one-hot encoded DNA sequences (221×4) and curvature vectors (221×1) in parallel. Each input passes through a separate Conv1D layer with 32 filters and kernel size 3 (padding='same', ReLU). The resulting feature maps are concatenated along the channel axis to preserve positional alignment, followed by an additional Conv1D layer with 64 filters, max pooling, flattening, a dense layer with 32 neurons (ReLU), and a final output layer with a sigmoid activation.

This first CNN serves two purposes: (1) to evaluate whether DNA curvature is a conserved feature in the TFBS regions of the two selected TFs (NAC and Crp), and (2) to assess how well the model performs when using only nucleotide sequence information for classification. In this case, the nucleotide sequences are transformed from a one-dimensional vector into a four-dimensional format using one-hot encoding, mapping nucleotides (A, T, C, G) into numeric vectors.

The second CNN was designed to test whether the inclusion of DNA curvature information improves the model's predictive performance. This architecture takes two parallel inputs: one for the encoded nucleotide sequence and another for the curvature profile. The sequence input is a 221×4 tensor (after one-hot encoding), and the curvature input is a 221×1 tensor. Each input is processed by a separate one-dimensional convolutional layer with 32 filters and kernel size 3 (ReLU activation, padding='same'). The resulting feature maps are concatenated to preserve positional alignment. The combined output then passes through another convolutional layer (64 filters, kernel size 3, ReLU), followed by a max pooling layer, a flattening step, and two dense layers (32 neurons with ReLU and 1 neuron with sigmoid activation).

Training both models posed challenges due to the significant imbalance in the dataset, in the case of NAC with 413 positive samples and 19,342 negatives. To address this, we applied three complementary strategies. First, we used dynamic undersampling, where in each training epoch the model is trained using all positive samples and an equal number of randomly selected negative samples. Second, we applied SMOTE (Synthetic Minority Over-sampling Technique) to synthetically increase the number of positive samples. Third, we implemented weighted training, assigning a higher penalty to errors made on the minority class. This combination of approaches was key to mitigating the effects of class imbalance and improving the robustness of both models.

3 Results

During the initial testing of the model, we observed a high degree of overfitting. Anticipating this issue, we opted for the previously described CNN architectures, which include only a single convolutional layer and one dense layer, in order to keep the model simple and avoid overfitting due to the severe class imbalance. To further address this, we added a dropout layer with a rate of 0.5 after the dense layer in an effort to reduce overfitting.

The results obtained from attempting to classify the sequences using only DNA curvature are summarised in table 1. As it shows, curvature alone does not provide enough discriminatory power for effective classification when working with a highly imbalanced dataset. However, when the test set is balanced, the results for the transcription factor CRP suggest that curvature is indeed a conserved feature. This pattern is not observed for NAC, where curvature does not appear to be conserved. Among the class imbalance mitigation strategies tested, dynamic undersampling yielded the best performance. Additionally, models using curvature alone outperformed those using either the curvature angle or a combination of both curvature metrics.

Moreover, the SMOTE oversampling technique led to significant challenges related to overfitting. Throughout all experiments, we applied early stopping during training based on validation loss. In many cases where SMOTE was used, early stopping was triggered immediately after the first epoch.

Training Method	Transcription Factor	Curvature	Test	Accuracy	Precision	Recall
SMOTE	NAC	Curvature	Balanced	0.50	0.50	0.25
SMOTE	NAC	Curvature angle	Balanced	0.47	0.46	0.43
SMOTE	**CRP**	**Curvature**	**Balanced**	**0.67**	**0.92**	**0.37**
SMOTE	**CRP**	**Curvature angle**	**Balanced**	**0.62**	**0.65**	**0.54**
50-50*	NAC	Curvature	Balanced	0.48	0.48	0.38
50-50*	NAC	Curvature Angle	Balanced	0.50	0.50	0.36
50-50*	**CRP**	**Curvature**	**Balanced**	**0.72**	**0.73**	**0.71**
50-50*	CRP	Curvature Angle	Balanced	0.54	0.53	0.60
Weighted	NAC	Curvature	Balanced	0.51	0.50	0.96
Weighted	NAC	Curvature Angle	Balanced	0.48	0.49	0.68
Weighted	**CRP**	**Curvature**	**Balanced**	**0.70**	**0.69**	**0.71**
Weighted	**CRP**	**Curvature Angle**	**Balanced**	**0.68**	**0.66**	**0.74**
Weighted	**CRP**	**2**	**Balanced**	**0.65**	**0.72**	**0.51**
SMOTE	NAC	Curvature	Unbalanced	0.88	0.02	0.14
SMOTE	NAC	Curvature angle	Unbalanced	0.82	0.02	0.21
SMOTE	CRP	Curvature	Unbalanced	0.98	0.05	0.02
SMOTE	CRP	Curvature angle	Unbalanced	0.87	0.02	0.20
50-50*	NAC	Curvature	Unbalanced	0.74	0.02	0.26
50-50*	NAC	Curvature Angle	Unbalanced	0.88	0.02	0.12
50-50*	CRP	Curvature	Unbalanced	0.85	0.02	0.28
50-50*	CRP	Curvature Angle	Unbalanced	0.75	0.02	0.51
Weighted	NAC	Curvature	Unbalanced	0.85	0.02	0.15
Weighted	NAC	Curvature Angle	Unbalanced	0.60	0.02	0.38
Weighted	CRP	Curvature	Unbalanced	0.85	0.02	0.28
Weighted	CRP	Curvature Angle	Unbalanced	0.91	0.04	0.25

Table. 1 Results of the CNN using different training methods, parameter values, and test set configurations. The "Curvature Table" column indicates whether curvature or curvature angle was used for training and classification. A value of 2 indicates that both were combined during training and testing. The "Balanced" or "Unbalanced" label refers to whether the test set maintains the original class imbalance or is balanced with an equal number of instances per class. The model was compiled using the ADAM optimizer, with 50 epochs, a batch size of 16, and a validation split and validation test of 20% each.

Early stopping based on validation loss was applied in all cases. For models trained with SMOTE on balanced validation and test sets, significant overfitting was observed, with the lowest validation loss occurring at the first epoch. The Dataset used is the Not Random data set for each TF.

After identifying the most effective strategy for addressing dataset imbalance, confirming that DNA curvature is more conserved in Crp than in Nac, and observing that models perform better using curvature than curvature angle, we evaluated whether combining Crp curvature data with nucleotide sequence information would improve model performance compared to using the nucleotide sequence alone. The results obtained are gathered in table 2.

As observed, incorporating curvature information into the model significantly improves both precision and recall in almost all scenarios involving imbalanced test sets. However, despite this improvement, the overall results are still not satisfactory. In balanced test conditions, the improvement is minimal or negligible.

Dataset	Test	Input Vector	Accuracy	Precision	Recall
Not Random	Balanced	Seq	0.62	0.66	0.51
Not Random	Balanced	Curvature	0.65	0.66	0.62
Not Random	Balanced	Seq + Curvature	0.61	0.65	0.48
Not Random	Unbalanced	Seq	0.95	0.18	0.05
Not Random	Unbalanced	Curvature	0.84	0.10	0.42
Not Random	Unbalanced	Seq + Curvature	0.92	0.17	0.22
Random	Balanced	Seq	0.90	0.96	0.82
Random	Balanced	Curvature	0.70	0.68	0.74
Random	Balanced	Seq + Curvature	0.88	0.96	0.80
Random	Unbalanced	Seq	0.92	0.57	0.62
Random	Unbalanced	Curvature	0.81	0.12	0.60
Random	Unbalanced	Seq + Curvature	0.97	0.64	0.68
Mix	Balanced	Seq	0.80	0.76	0.85
Mix	Balanced	Curvature	0.78	0.75	0.85
Mix	Balanced	Seq + Curvature	0.81	0.82	0.80
Mix	Unbalanced	Seq	0.94	0.19	0.14
Mix	Unbalanced	Curvature	0.85	0.13	0.48
Mix	Unbalanced	Seq + Curvature	0.94	0.27	0.28

Table. 2 Results of the CNN using different test sets and input data types. In the "Seq" condition, the input is a one-hot encoded nucleotide sequence represented as a 221-position vector with 4 dimensions. In the "Curvature" condition, the input is the curvature 221x1 vector. In the "Seq + Curvature" condition, two input vectors are used: a 221×4 sequence vector and a 221×1 curvature vector. The optimizer (ADAM), number of epochs (50), batch size (16), early stopping strategy, and the size of the test and validation sets (20% each) are the same as those used in Table 1.

When introducing the nucleotide sequence into the model—either on its own or combined with curvature—we use the extended sequence, not just the TFBS region. This approach allows the model to learn positional relationships when combining sequence and curvature information, which would not be feasible if we only included the TFBS.

As shown in Figure 5, when the sequence is added on its own, the positions corresponding to the TFBS stand out with significantly higher importance compared to the rest of the sequence. In contrast, the curvature vector shows a much more evenly distributed positional contribution.

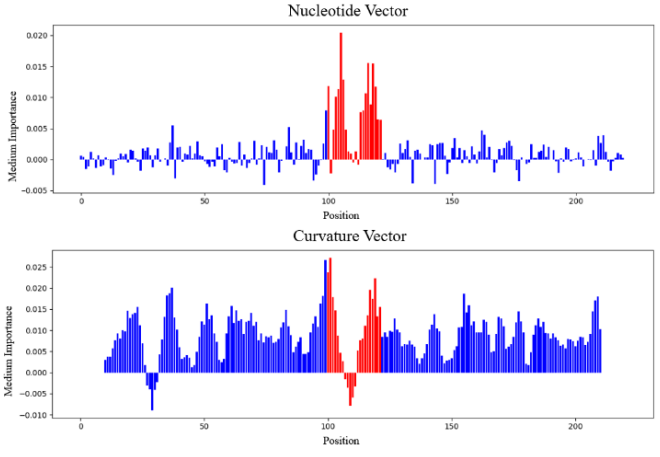

Fig. 5 Histogram showing the average positional importance for both the curvature and sequence input vectors. Red bars indicate positions corresponding to the TFBS. Importance was computed by masking each position and measuring the change in the model's binary prediction, averaged over the balanced test set from the Random dataset.

4 Conclusions

As the results show, the curvature of TFBS regions and their surrounding sequences can help distinguish these nucleotide sequences, though not in all cases—as seen with Nac—nor with particularly strong performance. This suggests that curvature around TFBS is conserved or partially conserved for some transcription factors, potentially serving as an additional layer of information that certain TFs use to recognize their binding sites.

The fact that we do not observe curvature conservation for Nac, while we do for Crp, aligns with findings in the literature: in some cases, sequence alone provides sufficient information for TFBS recognition, while in others, additional factors are required. It could be that for Nac, sequence information alone is enough, or that it relies on other features unrelated to curvature.

The results presented in Table 2 highlight the potential of DNA curvature as a key feature for the accurate classification of TFBSs. Notably, the model that achieves the best performance on datasets composed of highly similar sequences is the one that uses the curvature vector as input. This suggests that DNA curvature holds strong potential for improving TFBS classification and should be seriously considered in future motif identification models.

The poor results observed with imbalanced test sets suggest that curvature is not conserved across all binding sites of a given transcription factor. This indicates that curvature may only be necessary for recognizing a subset of binding sites, while others do not require it. When evaluating the model on a large set of sequences and curvature profiles, there is a higher likelihood of including sequences not experimentally confirmed as TFBS but that share curvature patterns with known TFBS, which can lead to misleading results. Figure 6 further supports this idea. It shows that curvature is more strongly conserved among orthologous genes across different Enterobacteria species than among different genes regulated by the same TF within a single species.

Furthermore, the severe class imbalance in our dataset posed significant challenges in developing a more complex model for accurate TFBS classification. The need to keep the architecture simple to avoid overfitting may have limited the model's ability to fully capture and leverage curvature-related information, potentially hindering performance and preventing better results.

Finally, the modest improvement observed when combining curvature with sequence data may be due to how the DNA curvature is computed by the DNAcurve library. Since the curvature is calculated as a sum of contributions from nucleotide dimers or trimers, it is inherently tied to the sequence itself, making the curvature and sequence information highly similar.

Although curvature is not conserved across all binding sites for a given TF, it is clearly conserved in some cases and has been shown to play a functional role in the accurate identification of specific TFBS. The modest improvements observed when combining curvature with sequence data suggest that, while the contribution of curvature is limited under the current modeling approach, it still adds valuable information. These findings highlight the need for further investigation and optimization, as refining how curvature is integrated could lead to more robust models with enhanced ability to detect TFBS.

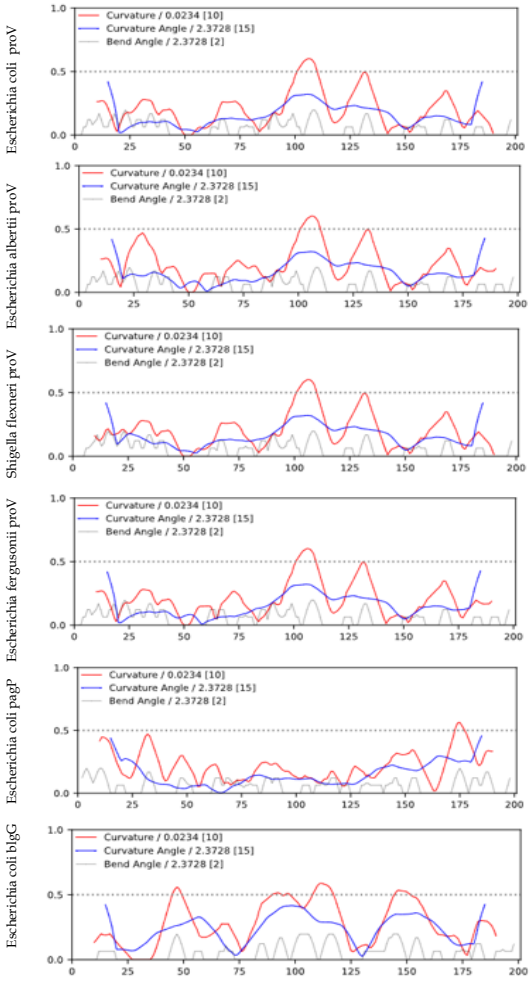

Fig. 6 Curvature, curvature angle, and bend angle profiles for six H-NS binding sites, a transcription factor known to use DNA curvature for site recognition. Four of the six plots correspond to the proV gene in different Enterobacteria species (Escherichia coli, Escherichia albertii, Shigella flexneri, and Escherichia fergusonii), where curvature-driven recognition by H-NS has been experimentally confirmed. The remaining two plots show the profiles for the pagP and blgG genes from Escherichia coli, also known to be recognized by H-NS due to their DNA curvature.

Agraïments

Als meus pares i als meus germans. A l'Ariadna Costas per ajudar-me en tot.

Bibliografia

[1] **Aptekmann, A. A., Bulavka, D., Nadra, A. D., & Sánchez, I. E.** (2022). Transcription factor specificity limits the number of DNA-binding motifs. *PloS one, 17*(1), e0263307.

https://doi.org/10.1371/journal.pone.0263307

[2] **Atlung, T., & Ingmer, H.** (1997). H-NS: a modulator of environmentally regulated gene expression. *Molecular microbiology, 24*(1), 7–17. https://doi.org/10.1046/j.1365-2958.1997.3151679.x

[3] **Bolshoy, A., McNamara, P., Harrington, R. E., & Trifonov, E. N.** (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(6), 2312–2316. https://doi.org/10.1073/pnas.88.6.2312 Bolshoy, A., McNamara, P., Harrington, R. E., & Trifonov, E. N. (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(6), 2312–2316. https://doi.org/10.1073/pnas.88.6.2312

[4] **Brazhnik, P., de la Fuente, A., & Mendes, P**. (2002). Gene networks: how to put the function in genomics. *Trends in biotechnology, 20*(11), 467–472. https://doi.org/10.1016/s0167-7799(02)02053-x

[5] **Dame, R. T., Wyman, C., & Goosen, N.** (2000). Structural basis for preferential binding of H-NS to curved DNA. *Nucleic Acids Research*, 28(18), 3504–3510. https://doi.org/10.1093/nar/28.18.3504

[6] **Delgado, F. M., & Gómez-Vela, F.** (2019). Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artificial intelligence in medicine, 95*, 133–145. https://doi.org/10.1016/j.artmed.2018.10.006

[7] **Georgiev, S., Boyle, A. P., Jayasurya, K., Ding, X., Mukherjee, S., & Ohler, U.** (2010). Evidence-ranked motif identification. *Genome biology*, *11*(2), R19. https://doi.org/10.1186/gb-2010-11-2-r19

[8] **Harrington, R. E.** (1992). DNA curving and bending in protein-DNA recognition. *Molecular microbiology, 6*(18), 2549–2555. https://doi.org/10.1111/j.1365-2958.1992.tb01431.x

[9] **Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., & Bulyk, M. L.** (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic acids research, 43*(Database issue), D117–D122. https://doi.org/10.1093/nar/gku1045

[10] **Inukai, S., Kock, K. H., & Bulyk, M. L.** (2017). Transcription factor-DNA binding: beyond binding site motifs. *Current opinion in genetics & development, 43*, 110–119. https://doi.org/10.1016/j.gde.2017.02.007

[11] **Jáuregui, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., Collado-Vides, J., & Merino, E.** (2003). Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. *Nucleic acids research, 31*(23), 6770–6777. https://doi.org/10.1093/nar/gkg882

[12] **Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T.** (2018). The Human Transcription Factors. *Cell, 172*(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

[13] **Martinez, G. S., Perez-Rueda, E., Kumar, A., Dutt, M., Maya, C. R., Ledesma-Dominguez, L., Casa, P. L., Kumar, A., de Avila E Silva, S., & Kelvin, D. J.** (2024). CDBProm: the Comprehensive Directory of Bacterial Promoters. *NAR genomics and bioinformatics*, *6*(1), lqae018. https://doi.org/10.1093/nargab/lqae018

[14] **Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., & Giorgi, F. M.** (2020). Gene regulatory network inference resources: A practical overview. *Biochimica et biophysica acta. Gene regulatory mechanisms, 1863*(6), 194430. https://doi.org/10.1016/j.bbagrm.2019.194430

[15] **Meysman, P., Marchal, K., & Engelen, K.** (2010). Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli. Nucleic Acids Research, 38(10), 3355–3367. https://doi.org/10.1093/nar/gkq063

[16] **Pul, Ü., Lux, B., Wurm, R., & Wagner, R.** (2008). Effect of upstream curvature and transcription factors H-NS and LRP on the efficiency of *Escherichia coli* rRNA promoters P1 and P2 – a phasing analysis. *Microbiology*, 154(9), 1887–1895. https://doi.org/10.1099/mic.0.2008/018408-0

[17] **Shimada, T., Tanaka, K., & Ishihama, A.** (2017). The whole set of the constitutive promoters recognized by four minor sigma subunits of Escherichia coli RNA polymerase. *PloS one*, *12*(6), e0179181. https://doi.org/10.1371/journal.pone.0179181

[18] **Siggers, T., & Gordân, R.** (2014). Protein-DNA binding: complexities and multi-protein codes. *Nucleic acids research, 42*(4), 2099–2111. https://doi.org/10.1093/nar/gkt1112

[19] **Tan, R. K., & Harvey, S. C.** (1987). A comparison of six DNA bending models. *Journal of biomolecular structure & dynamics*, *5*(3), 497–512. https://doi.org/10.1080/07391102.1987.10506410Tan, R. K., & Harvey, S. C. (1987). A comparison of six DNA bending models. *Journal of biomolecular structure & dynamics*, *5*(3), 497–512. https://doi.org/10.1080/07391102.1987.10506410

[20] **Travers, A., & Muskhelishvili, G.** (1998). DNA microloops and microdomains: a general mechanism for transcription activation by torsional transmission. Journal of molecular biology, 279(5), 1027–1043. https://doi.org/10.1006/jmbi.1998.1834

[21] **Urtecho Guillaume, Insigne Kimberly D., Tripp Arielle D., Brinck Marcia S., Lubock Nathan B., Acree**

Christopher, Kim Hwangbeom, Chan Tracey, Kosuri Sriram (2023) Genome-wide Functional Characterization of Escherichia coli Promoters and Sequence Elements Encoding Their Regulation eLife 12:RP92558 https://doi.org/10.7554/eLife.92558.1

[22] **Vahed, M., Ishihara, J. I., & Takahashi, H.** (2019). DIpartite: A tool for detecting bipartite motifs by considering base interdependencies. *PloS one, 14*(8), e0220207. https://doi.org/10.1371/journal.pone.0220207

[23] **Vahed, M., Vahed, M., & Garmire, L. X.** (2022). BML: a versatile web server for bipartite motif discovery. *Briefings in bioinformatics, 23*(1), bbab536. https://doi.org/10.1093/bib/bbab536

[24] **Wang, M., Wang, D., Zhang, K., Ngo, V., Fan, S., & Wang, W.** (2020). Motto: Representing Motifs in Consensus Sequenc-es with Minimum Information Loss. Genetics, 216(2), 353–358. https://doi.org/10.1534/genetics.120.303597

- *E-mail de contacte: joan.estop@autonoma.cat*
- *Menció realitzada: Computació*
- *Treball tutoritzat per: Ivan Erill (Àrea de Ciències de la Computació i Intel·ligència Artificial)*
- *Curs 2024/25*