



This work is on a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>. Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

Design of Machine Learning Models for the Prediction of Transcription Factor Binding Regions in Bacterial DNA [†]

Sara Alvarez-Gonzalez ^{1,2}  and Ivan Erill ^{3,*} 

¹ Departamento de Ciencias de la Computación y Tecnologías de la Información, Universidade da Coruña, 15071 A Coruña, Spain; sara.alvarezg@udc.es

² CITIC-Centro de Investigación en Tecnologías de la información y las Comunicaciones, Universidade da Coruña, 15071 A Coruña, Spain

³ Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD 21250, USA

* Correspondence: erill@umbc.edu

[†] Presented at the 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: Transcription Factors (TFs) are proteins that regulate the expression of genes by binding to their promoter regions. There is great interest in understanding in which regions TFs will bind to the DNA sequence of an organism and the possible genetic implications that this entails. Occasionally, the sequence patterns (motifs) that a TF binds are not well defined. In this work, machine learning (ML) models were applied to TF binding data from ChIP-seq experiments. The objective was to detect patterns in TF binding regions that involved structural (DNAShapeR) and compositional (kmers) characteristics of the DNA sequence. After the application of random forest and Glmnet ML techniques with both internal and external validation, it was observed that two types of generated descriptors (HelT and tetramers) were significantly better than the others in terms of prediction, achieving values of more than 90%.

Keywords: transcription factor; machine learning; protein binding



Citation: Alvarez-Gonzalez, S.; Erill, I. Design of Machine Learning Models for the Prediction of Transcription Factor Binding Regions in Bacterial DNA. *Eng. Proc.* **2021**, *7*, 59. <https://doi.org/10.3390/engproc2021007059>

Published: 29 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The prediction of specific transcription factor (TF) binding regions in the DNA of bacterial organisms is a challenging task, especially when the TF binding motifs are not well defined or there are certain structural parameters of the DNA structure at play that classical models do not take into account. Advances in machine learning (ML) techniques have made it possible to create models capable of incorporating DNA structural parameters in the prediction of TF binding sites in genomic sequences.

In this work, using ML techniques, we were able to predict with high accuracy whether a nucleotide sequence would be a region where the TF would bind or not. Our work has been previously published as a Master's Thesis at the Universitat Oberta de Catalunya (UOC) [1]. The DNA sequences used as positive data were extracted from the article by Adhikari et al. [2], obtained using the ChIP-seq technique with the GcrA TF in the bacterial organism *Brevundimonas subvibrioides*.

2. Materials and Methods

2.1. Creation of Negative Sequence Set

From the 879 peaks (nucleotide sequences where the target TF had bound), two types of nucleotide sequences were created from these in order to generate the database of negative sequences.

On the one hand, biologically plausible replicates, here termed Replicates, were created. For these, the *peaks* were located in the reference genome, and their subsequences were classified according to whether they were located relative to a gene: intergenic, intragenic, upstream, or downstream. From this, 879 sequences homologous to the positive

dataset were generated matching the composition of each one but using regions of the organism's genome where the TF had not been bound.

On the other hand, a negative dataset was generated with a pseudo-replicate process, termed Boots replicates. In this case, the generation of negative data was carried out by extracting trimers that existed in the original peak until the target length was completed.

2.2. Descriptor Extraction

The FASTA file of each of the datasets (879 peaks + 879 replicates for each of the cases) was introduced to the R DNashapeR library, and vectors of values were obtained for each of the 4 selected elements: HelT, MGW, ProT, and Roll. From the data vectors obtained for each sequence, histograms generated for each sequence were used as descriptors. In this case, 25 descriptors were chosen by DNashape and dataset algorithms.

Descriptors were also calculated by counting the appearance of k-mers in each of the sequences. Monomers, dimers, and tetramers (4, 16, and 256 combinations, respectively) were studied.

2.3. Machine Learning Models

Random forest [3] and generalized linear model (Glmnet) [4] were used in this work. A nested cross-validation was carried out, using a validation for the search of the best hyperparameters through a holdout and a second validation with a 10-fold CV to validate the model, taking the average of 5 iterations of this technique. Thus, the performance values reflected in Figure 1 represent the average of the 50 models generated for each descriptor and each ML model.

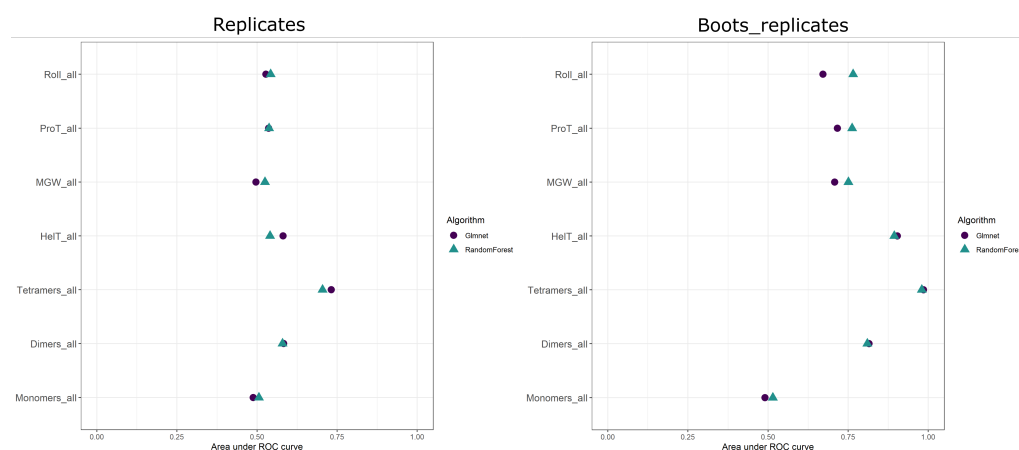


Figure 1. Dot plot showing the mean of each of the models. The mean corresponds to the 50 replicates obtained using the CV in replicates. The image on the left corresponds to the models in the dataset of Replicates, while the image on the right matches the models in the dataset of Boots_replicates.

3. Results

The total number of datasets to carry out the ML models was seven sets for each of the datasets (Replicates and Boots_replicates), resulting in a total of fourteen sets of descriptors to be trained with the two algorithms. The models obtained by the algorithms were evaluated using the *area under the receiver operating characteristics curve* (AUCROC). The main results are described in the subsection presented below.

HelT and Tetramers as Key Descriptors

Figure 1 illustrates the performances of the AUCROC for both datasets and all descriptors. On the left side, we can observe the results obtained for the Replicates, while the right side shows those corresponding to Boots_replicates. The first four sets represented belong to the structural descriptors extracted by the DNashape library, while the last three correspond to k-mers counts.

As can be observed, both the HelT and tetramers descriptors produce the ML models with the highest performance, reaching up to 0.75 in the Replicates and almost 0.9–1 in the Boots_replicates.

4. Discussion

The use of these descriptors is due to the fact that we seek to represent the sequences not only in terms of base reading but also in terms of their structural features. These have been described as two different modes for protein–DNA recognition that are key and necessary to predict the behavior of TFs in DNA [5].

Of the seven descriptors studied, it was observed that different ML models with a high level of prediction could be developed to determine the TF binding regions across the genome. In particular, both HelT and tetramer counts proved to be particularly predictive.

Similarly, it has been found that the most significant tetramers for ML algorithms largely correspond to the motif that was found in the reference paper for the data used [2] with the MEME program, as presented in Figure 2.



Figure 2. Motif found in reference article [2] for GcrA TF in *Brevundimonas subvibrioides*.

Funding: This work has received financial support from the Xunta de Galicia and the European Union (European Social Fund (ESF)).

Data Availability Statement: The database used for the development of this study it is available at the Gene Expression Omnibus (GEO) under accession number GSE138845.

Acknowledgments: This project was also supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431G/01, ED431D 2017/16).

References

1. Alvarez-Gonzalez, S. *Desarrollo de un sistema de Machine Learning para obtener modelos de unión a Factores de Transcripción en datos ChIP-seq*; Universitat Oberta de Catalunya (UOC): Barcelona, Spain, 2021. Available online: <http://hdl.handle.net/10609/133102> (accessed on 23 September 2021).
2. Adhikari, S.; Erill, I.; Curtis, P.D. Transcriptional rewiring of the GcrA/CcrM bacterial epigenetic regulatory system in closely related bacteria. *PLoS Genet.* **2021**, *17*, e1009433. [CrossRef] [PubMed]
3. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1. [CrossRef] [PubMed]
4. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
5. Abe, N.; Dror, I.; Yang, L.; Slattey, M.; Zhou, T.; Bussemaker, H.J.; Rohs, R.; Mann, R.S. Deconvolving the recognition of DNA shape from sequence. *Cell* **2015**, *161*, 307–318. [CrossRef] [PubMed]