# IA Applications in Transcription Factors Binding Site

André Borges Farias
fariasab@lncc.br

45 anos LNCC Laboratório Nacional de Computação Científica

Laboratório de Bioinformática

25 Years Advancing Bioinformatics

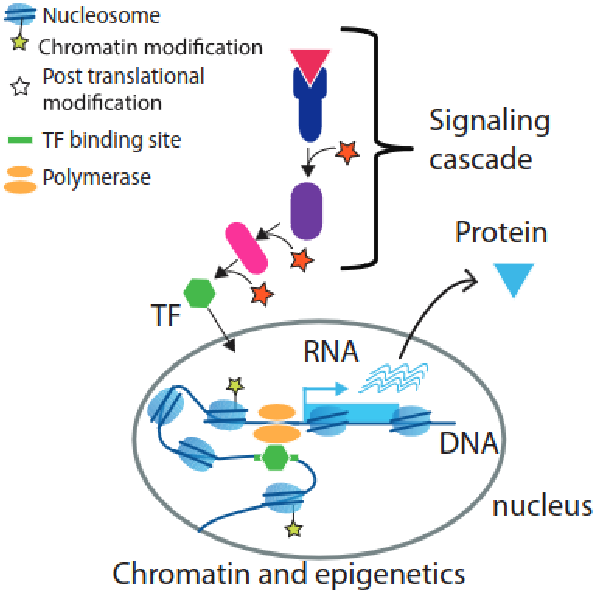
GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO POR MI RAZA HABLARA EL ESPIRITU

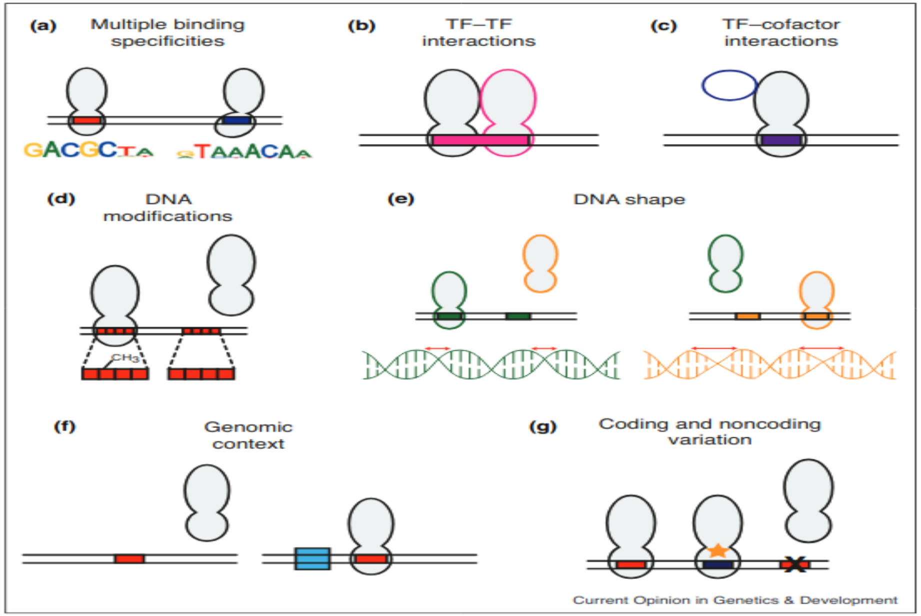
# Transcription Factor (TF)

- Proteins that bind to specific DNA sequences (TFBS) to regulate gene transcription.
- Play a fundamental role in gene expression regulation.
- Act as transcription activators or repressors.



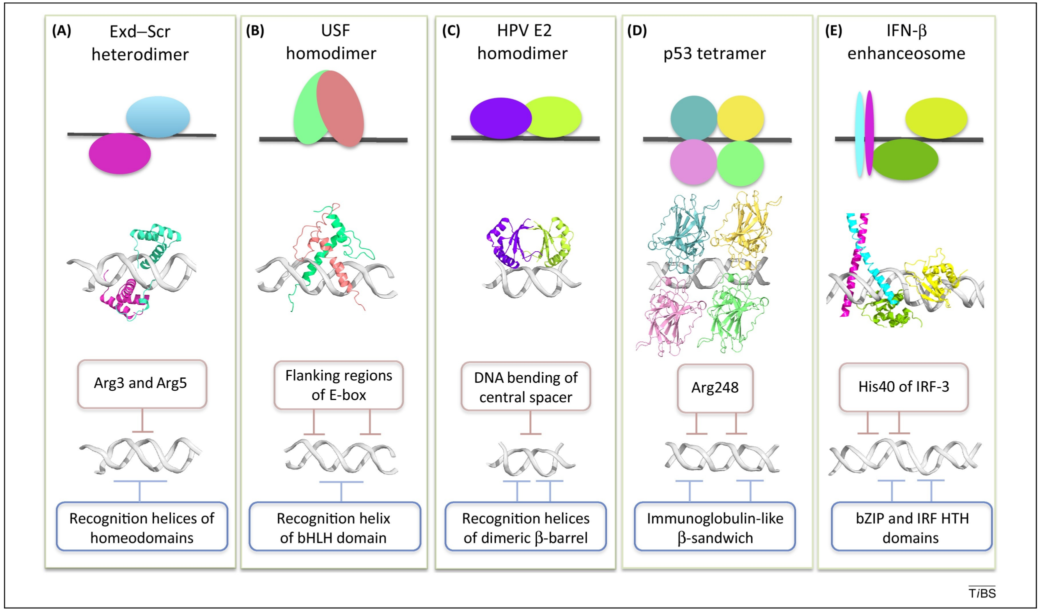
# Characteristics that influence TF binding:

The modulation of TF-DNA recognition depends on the characteristics of transcription factors (TFs) or DNA binding sites.



(a) Multiple binding specificities
GACGCTA
TAAACAA
(b) TF–TF interactions
(c) TF–cofactor interactions
(d) DNA modifications
CH3
(e) DNA shape
(f) Genomic context
(g) Coding and noncoding variation

Current Opinion in Genetics & Development



(A) Exd–Scr heterodimer
(B) USF homodimer
(C) HPV E2 homodimer
(D) p53 tetramer
(E) IFN-β enhanceosome
Arg3 and Arg5
Flanking regions of E-box
DNA bending of central spacer
Arg248
His40 of IRF-3
Recognition helices of homeodomains
Recognition helix of bHLH domain
Recognition helices of dimeric β-barrel
Immunoglobulin-like β-sandwich
bZIP and IRF HTH domains

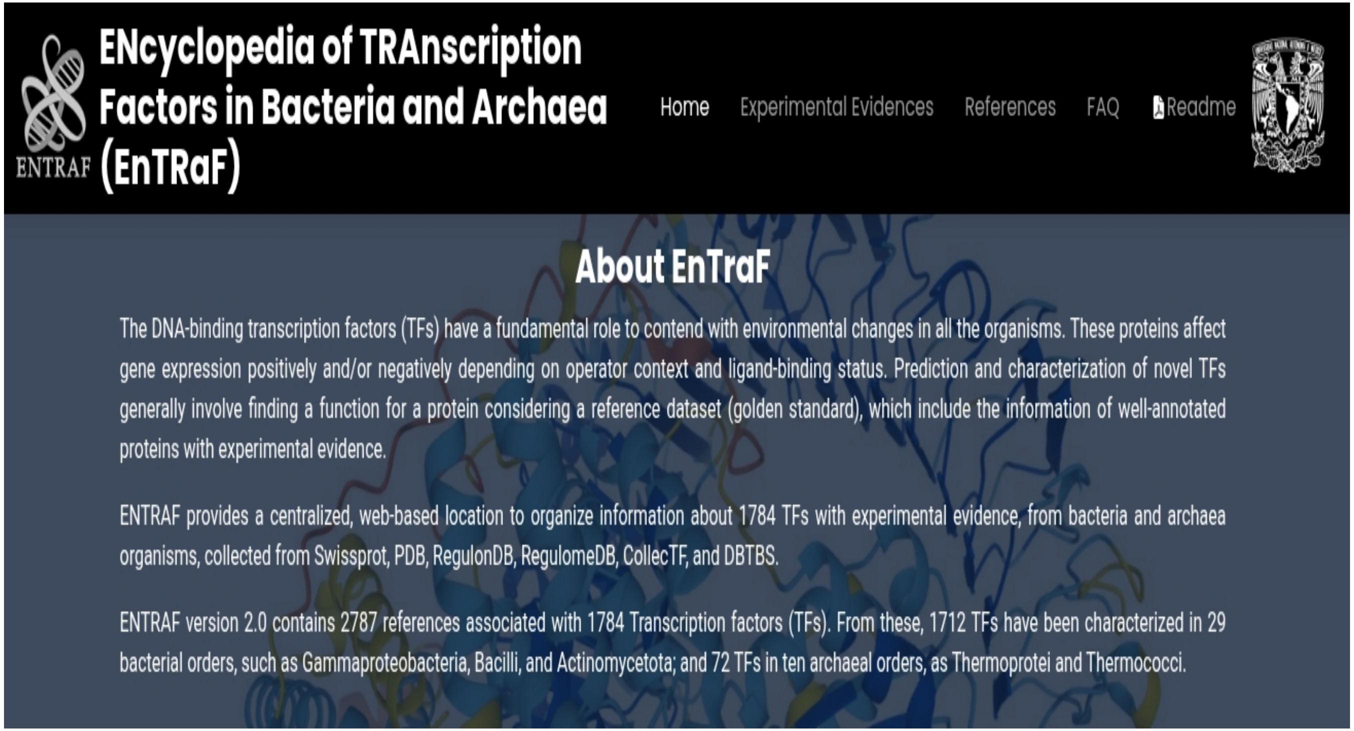
TiBS

Inukai S, Kock KH, Bulyk ML. Transcription factor–DNA binding: beyond binding site motifs. Current Opinion in Genetics & Development. 2017 Apr;43:110–9.

Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. Absence of a simple code: how transcription factors read the genome. Trends in Biochemical Sciences. 2014 Sep;39(9):381–99.

# Entraf

https://entraf.iimas.unam.mx



ENcyclopedia of TRAnscription Factors in Bacteria and Archaea (EnTRaF)

## About EnTraF

The DNA-binding transcription factors (TFs) have a fundamental role to contend with environmental changes in all the organisms. These proteins affect gene expression positively and/or negatively depending on operator context and ligand-binding status. Prediction and characterization of novel TFs generally involve finding a function for a protein considering a reference dataset (golden standard), which include the information of well-annotated proteins with experimental evidence.

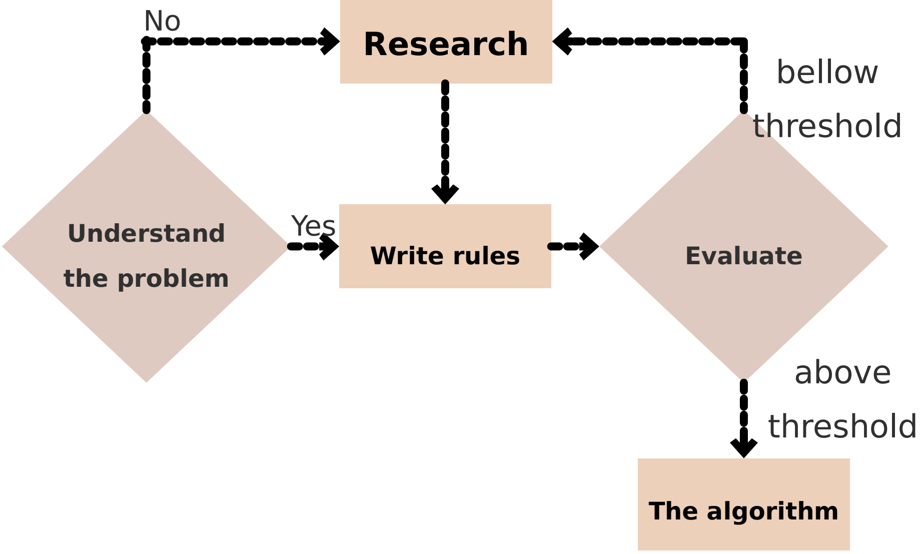
ENTRAF provides a centralized, web-based location to organize information about 1784 TFs with experimental evidence, from bacteria and archaea organisms, collected from Swissprot, PDB, RegulonDB, RegulomeDB, CollecTF, and DBTBS.

ENTRAF version 2.0 contains 2787 references associated with 1784 Transcription factors (TFs). From these, 1712 TFs have been characterized in 29 bacterial orders, such as Gammaproteobacteria, Bacilli, and Actinomycetota; and 72 TFs in ten archaeal orders, as Thermoprotei and Thermococci.

Tenorio-Salgado, S.; Ledesma-Dominguez, L.; Galan-Vasquez, E.; Farias, A. B., Alvarez, D.I. , Villalpando, J. L.; Perez-Rueda, E. ENcyclopedia of TRAnscription Factors in Bacteria and Archaea genomes (ENTRAF) version 2.0. Database. Submitted.

# WOULD BE POSSIBLE TO PREDICT TFBS BASED ON STRUCTURAL CHARACTERISTICS?

Traditional programming



No

Research

bellow threshold

Understand the problem

Yes

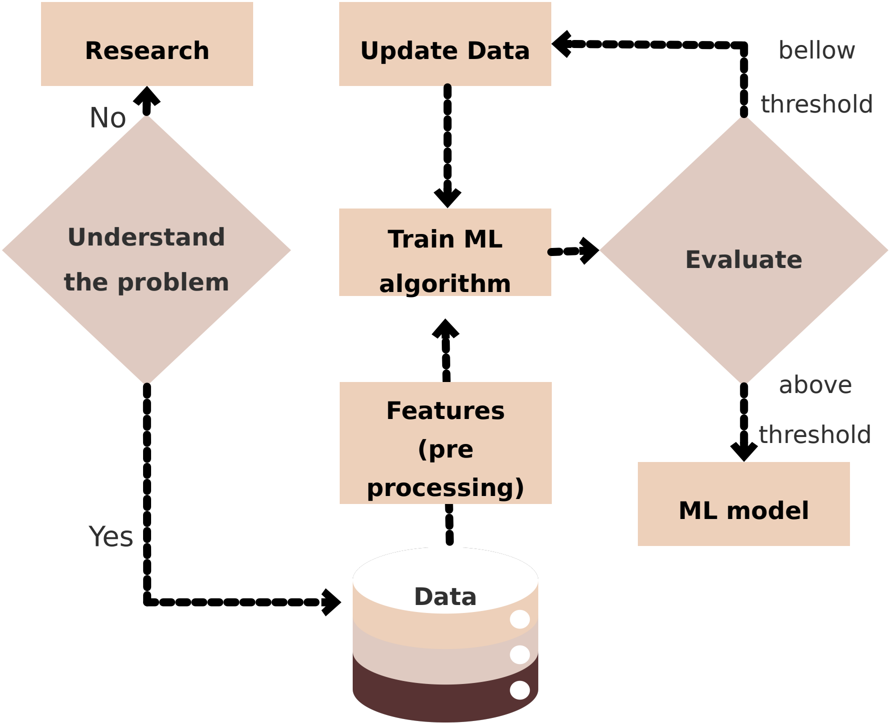
Write rules

Evaluate

above threshold

The algorithm

ML models



Research

No

Update Data

bellow threshold

Understand the problem

Train ML algorithm

Evaluate

Features (pre processing)

above threshold

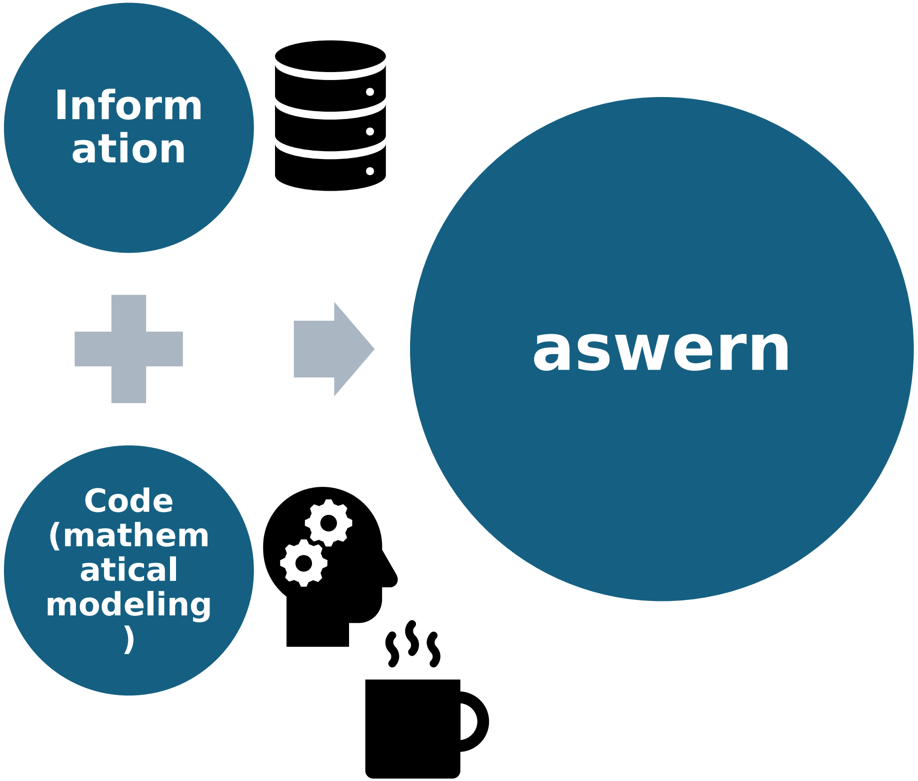
ML model

Yes

Data

How many rules should I define to get the information of TFBS?

# WOULD BE POSSIBLE TO PREDICT TFBS BASED ON STRUCTURAL CHARACTERISTICS?

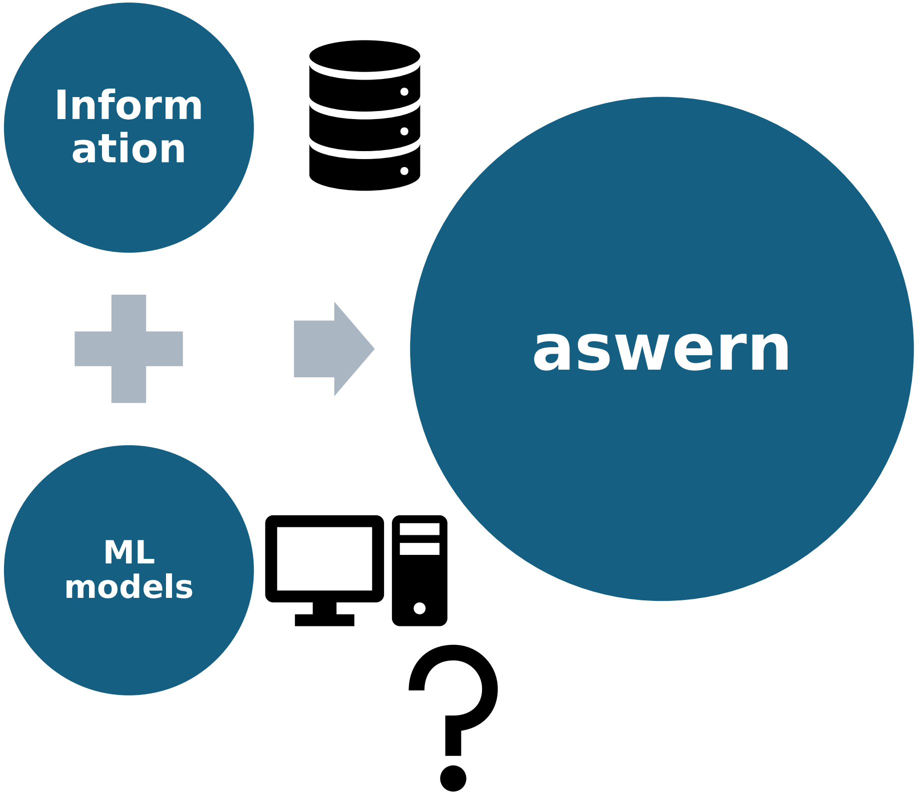


Information

Code (mathematical modeling)

aswern

Can the algorithm include all biological complexity?

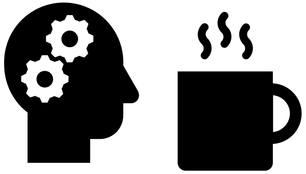


Information

ML models

aswern

What's in the black box? How does the model learn the data?



**mathematical modeling**

# Position Weight Matrices (PWMs)

PWM is a 4×L matrix, where:

- L is the length of the motif (number of positions in the binding sequence).
- Each line represents a nucleotide (A, C, G, T).

```
>seq1
ATTC
>seq2
CACT
>seq3
TTAA
>seq4
GGGG
>seq5
ACGC
```

Count Matrix

| Base | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 2 | 1 | 1 | 1 |
| C | 1 | 1 | 1 | 2 |
| G | 1 | 1 | 2 | 1 |
| T | 1 | 2 | 1 | 1 |

Probability matrix

| Base | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 2/5 | 1/5 | 1/5 | 1/5 |
| C | 1/5 | 1/5 | 1/5 | 2/5 |
| G | 1/5 | 1/5 | 2/5 | 1/5 |
| T | 1/5 | 2/5 | 1/5 | 1/5 |

# Position Weight Matrices (PWMs)

$$PWM(b,i)=\log_2\left(\frac{f(b,i)}{p(b)}\right)$$

f(b,i) is the observed frequency of nucleotide b at position i.

p(b)=0.25 assuming that each base (A, C, G or T) occurs with 25% probability

| Base | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | Log2((2/5)/0.25) | Log2((1/5)/0.25) | Log2((1/5)/0.25) | Log2((1/5)/0.25) |
| C | Log2((1/5)/0.25) | Log2((1/5)/0.25) | Log2((1/5)/0.25) | Log2((2/5)/0.25) |
| G | Log2((1/5)/0.25) | Log2((1/5)/0.25) | Log2((2/5)/0.25) | Log2((1/5)/0.25) |
| T | Log2((1/5)/0.25) | Log2((2/5)/0.25) | Log2((1/5)/0.25) | Log2((1/5)/0.25) |

# Position Weight Matrices (PWMs)

$$PWM(b,i)=\log_2\left(\frac{f(b,i)}{p(b)}\right)$$

f(b,i) is the observed frequency of nucleotide b at position i.

p(b)=0.25 assuming that each base (A, C, G or T) occurs with 25% probability

| Base | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 0,68 | -0,32 | -0,32 | -0,32 |
| C | -0,32 | -0,32 | -0,32 | 0,68 |
| G | -0,32 | -0,32 | 0,68 | -0,32 |
| T | -0,32 | 0,68 | -0,32 | -0,32 |

We can now use this PWM to predict new binding sites by scanning sequences.

# Prediction using PWMs

Suppose you want to predict an unknown sequence:

**ATGCCATGACGTAGCTAGTGCTAGC**

We choose a reference TF, represented by the matrix (PWM) below:

| Base | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1.2 | -0.8 | 0.5 | -0.3 | 1.1 | -1.0 |
| C | -1.1 | 0.2 | -0.7 | 1.5 | -0.5 | 0.9 |
| G | 0.4 | 1.7 | -0.9 | -1.3 | 0.8 | -0.2 |
| T | -0.5 | -1.0 | 1.2 | 0.3 | -1.4 | 0.5 |

# Prediction using PWMs

We define an score (S):

$$S=\sum_{i=1}^{L} PWM(b_i, i)$$

Where,
L is the length of the sequence;
PWM($b_i$, i) is the PWM value for nucleotide bi at position i

**ATGCCATGACGTAGCTAGTGCTAGC**

**A** at position 1 → 1.2
**T** at position 2 → -1.0
**G** at position 3 → -0.9
**C** at position 4 → 1.5
**C** at position 5 → -0.5
**A** at position 6 → -1.0

| Base | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1.2 | -0.8 | 0.5 | -0.3 | 1.1 | -1.0 |
| C | -1.1 | 0.2 | -0.7 | 1.5 | -0.5 | 0.9 |
| G | 0.4 | 1.7 | -0.9 | -1.3 | 0.8 | -0.2 |
| T | -0.5 | -1.0 | 1.2 | 0.3 | -1.4 | 0.5 |

S = 1.2+(–1.0)+(–0.9)+1.5+(–0.5)+(-1,0)= **-0.7**

# Predictions using PWMs

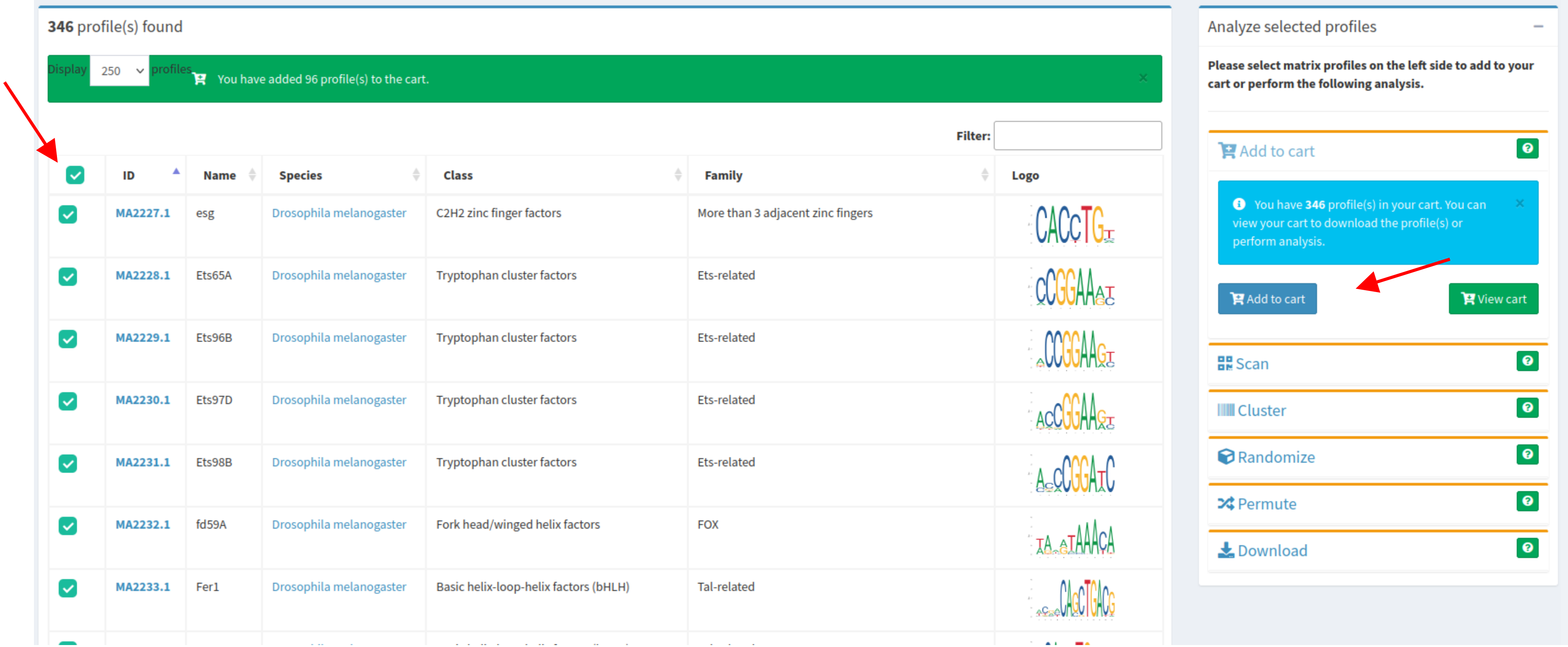
Repeat the same process for the rest of the sequence:

Position 1-6: ATGCCA → $S_1$
Position 2-7: TGCCAT → $S_2$
Position 3-8: GCCATG → $S_3$
Position 4-9: CCATGA → $S_4$

| Posição | Sequência | Score | Predição |
|---|---|---|---|
| 1-6 | ATGCCT | -0.7 | Não significativo |
| 2-7 | TGCCAT | 0.3 | Não significativo |
| 3-8 | GCCATG | 1.1 | Sítio de Ligação |
| 4-9 | CCATGA | -0.2 | Não significativo |

Valores aleatórios

# JASPAR Scan – Analysis of TFBS



**346** profile(s) found

Display 250 profiles

You have added 96 profile(s) to the cart.

Filter:

| ☑ | ID | Name | Species | Class | Family | Logo |
|---|---|---|---|---|---|---|
| ☑ | MA2227.1 | esg | Drosophila melanogaster | C2H2 zinc finger factors | More than 3 adjacent zinc fingers | |
| ☑ | MA2228.1 | Ets65A | Drosophila melanogaster | Tryptophan cluster factors | Ets-related | |
| ☑ | MA2229.1 | Ets96B | Drosophila melanogaster | Tryptophan cluster factors | Ets-related | |
| ☑ | MA2230.1 | Ets97D | Drosophila melanogaster | Tryptophan cluster factors | Ets-related | |
| ☑ | MA2231.1 | Ets98B | Drosophila melanogaster | Tryptophan cluster factors | Ets-related | |
| ☑ | MA2232.1 | fd59A | Drosophila melanogaster | Fork head/winged helix factors | FOX | |
| ☑ | MA2233.1 | Fer1 | Drosophila melanogaster | Basic helix-loop-helix factors (bHLH) | Tal-related | |

Analyze selected profiles

**Please select matrix profiles on the left side to add to your cart or perform the following analysis.**

Add to cart

You have **346** profile(s) in your cart. You can view your cart to download the profile(s) or perform analysis.

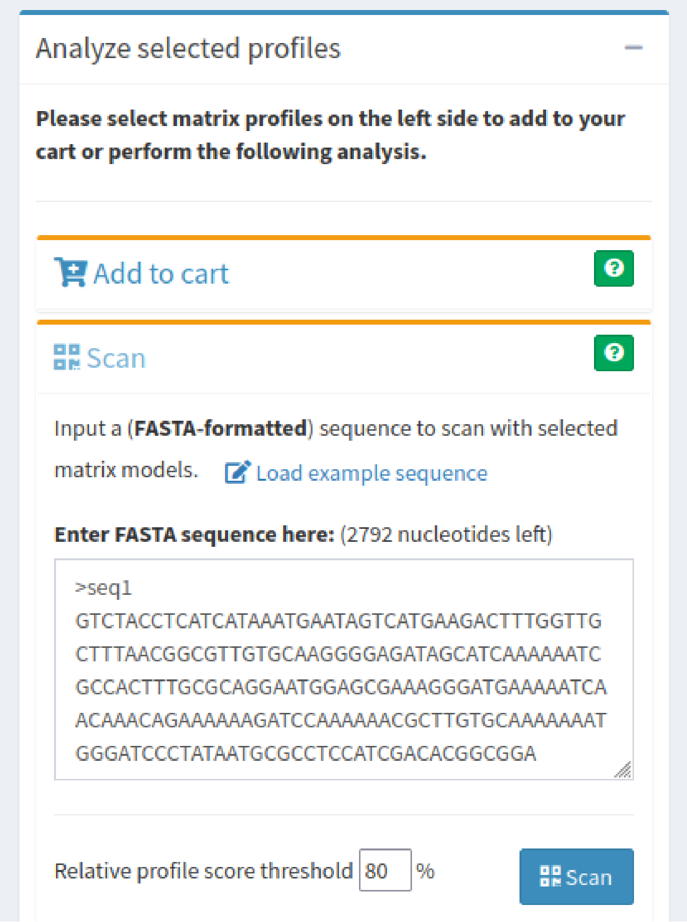
# JASPAR Scan – Analisys of TFBS

```
>Seq1

GTCTACCTCATCATAAATGAATAGTCATGAAGACTTTGGTTGCTTTAACGGCGTTGTGCAAGG
GGAGATAGCATCAAAAAATCGCCACTTTGCGCAGGAATGGAGCGAAAGGGATGAAAAATCA
ACAAACAGAAAAAAGATCCAAAAAACGCTTGTGCAAAAAAATGGGATCCCTATAATGCGCCT
CCATCGACACGGCGGA

 >Seq 2

 AAAGAAGTAAGCACACCTGCAAGGCCAGTTAACTGGCCATCGTAAATGGCCCGATAGTGTA
AGCTATTCGGGCCTGCGGTGTTTATGCCTGGGGAACGCCACCGGGTAAAACACGTTCTTCA
TTATCAATACTCTGAAACCGTCTGTTAATAACAGACGGTTTTTTGTCTATGGAAAA
```



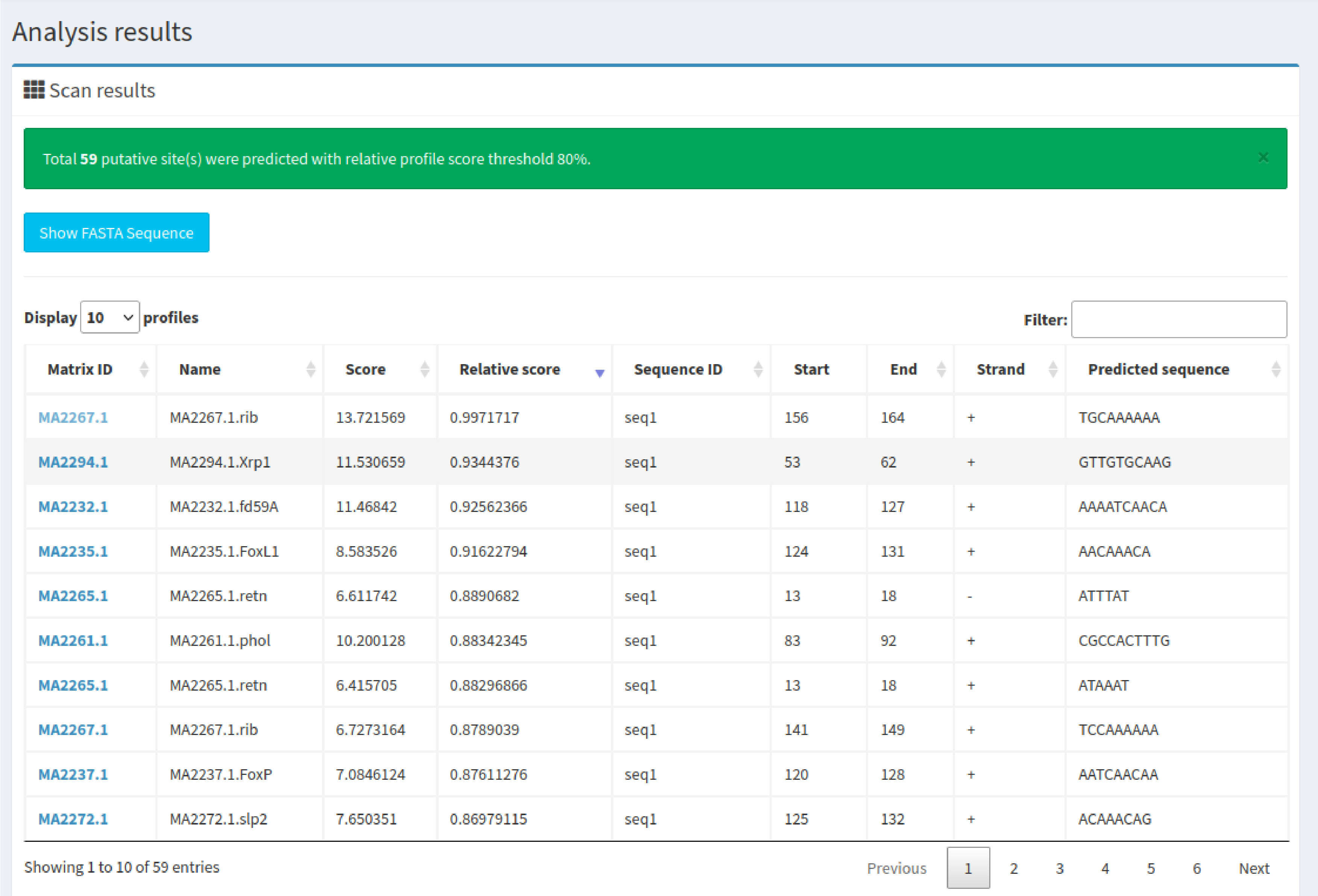
**Please select matrix profiles on the left side to add to your cart or perform the following analysis.**

Input a (**FASTA-formatted**) sequence to scan with selected matrix models. Load example sequence

**Enter FASTA sequence here:** (2792 nucleotides left)

```
>seq1
GTCTACCTCATCATAAATGAATAGTCATGAAGACTTTGGTTG
CTTTAACGGCGTTGTGCAAGGGGAGATAGCATCAAAAAATC
GCCACTTTGCGCAGGAATGGAGCGAAAGGGATGAAAAATCA
ACAAACAGAAAAAAGATCCAAAAAACGCTTGTGCAAAAAAAT
GGGATCCCTATAATGCGCCTCCATCGACACGGCGGA
```

Relative profile score threshold 80 %



Analysis results

Scan results

Total **59** putative site(s) were predicted with relative profile score threshold 80%.

Display 10 profiles

Filter:

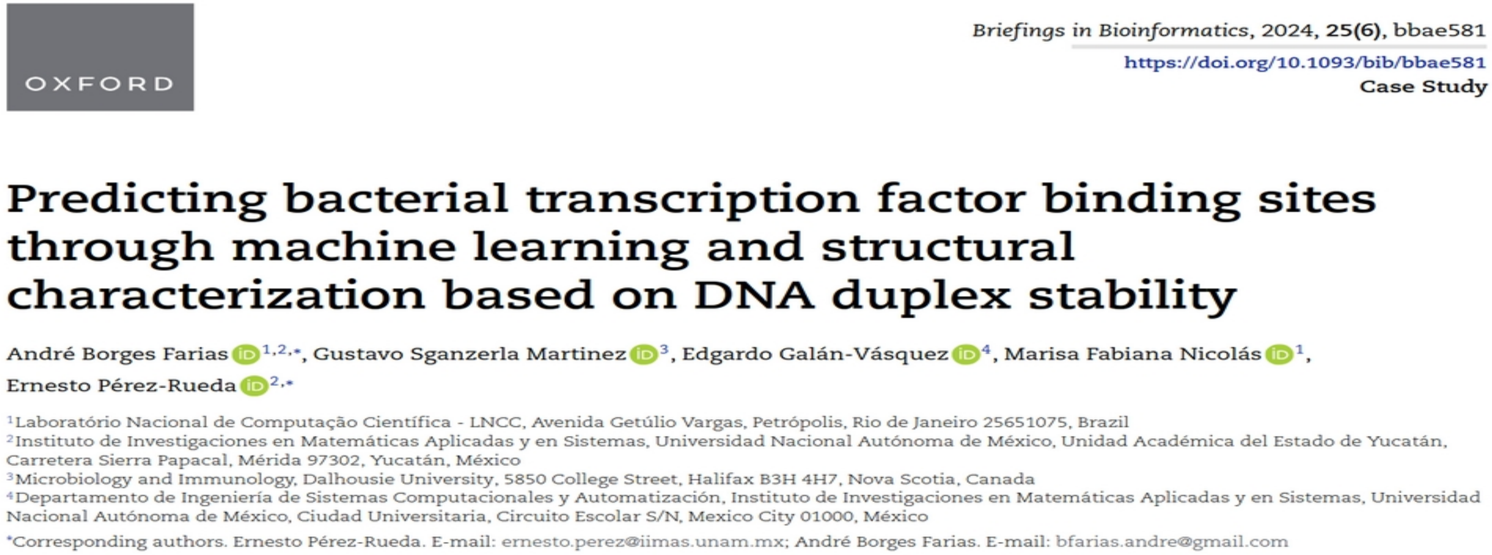
| Matrix ID | Name | Score | Relative score | Sequence ID | Start | End | Strand | Predicted sequence |
|---|---|---|---|---|---|---|---|---|
| MA2267.1 | MA2267.1.rib | 13.721569 | 0.9971717 | seq1 | 156 | 164 | + | TGCAAAAAA |
| MA2294.1 | MA2294.1.Xrp1 | 11.530659 | 0.9344376 | seq1 | 53 | 62 | + | GTTGTGCAAG |
| MA2232.1 | MA2232.1.fd59A | 11.46842 | 0.92562366 | seq1 | 118 | 127 | + | AAAATCAACA |
| MA2235.1 | MA2235.1.FoxL1 | 8.583526 | 0.91622794 | seq1 | 124 | 131 | + | AACAAACA |
| MA2265.1 | MA2265.1.retn | 6.611742 | 0.8890682 | seq1 | 13 | 18 | - | ATTTAT |
| MA2261.1 | MA2261.1.phol | 10.200128 | 0.88342345 | seq1 | 83 | 92 | + | CGCCACTTTG |
| MA2265.1 | MA2265.1.retn | 6.415705 | 0.88296866 | seq1 | 13 | 18 | + | ATAAAT |
| MA2267.1 | MA2267.1.rib | 6.7273164 | 0.8789039 | seq1 | 141 | 149 | + | TCCAAAAAA |
| MA2237.1 | MA2237.1.FoxP | 7.0846124 | 0.87611276 | seq1 | 120 | 128 | + | AATCAACAA |
| MA2272.1 | MA2272.1.slp2 | 7.650351 | 0.86979115 | seq1 | 125 | 132 | + | ACAAACAG |

Showing 1 to 10 of 59 entries

# WOULD BE POSSIBLE TO PREDICT TFBS BASED ON STRUCTURAL CHARACTERISTICS?

Machine learning is an approach to **learning** complex patterns from existing data and using those patterns to make **predictions** on unknown data.

# WOULD BE POSSIBLE TO PREDICT TFBS BASED ON STRUCTURAL CHARACTERISTICS?



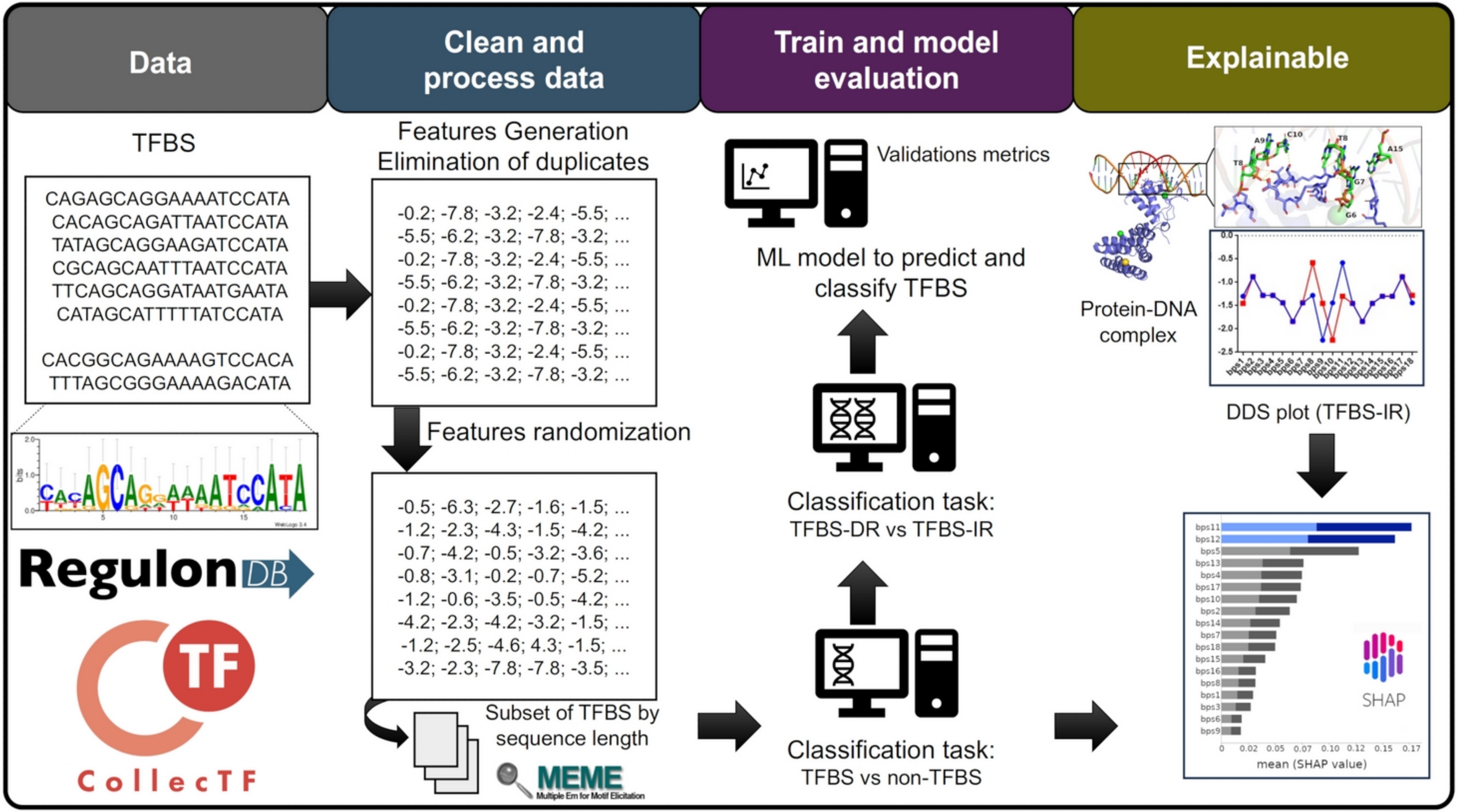
OXFORD

*Briefings in Bioinformatics*, 2024, **25**(6), bbae581
https://doi.org/10.1093/bib/bbae581
**Case Study**

## Predicting bacterial transcription factor binding sites through machine learning and structural characterization based on DNA duplex stability

André Borges Farias[1,2,*], Gustavo Sganzerla Martinez[3], Edgardo Galán-Vásquez[4], Marisa Fabiana Nicolás[1], Ernesto Pérez-Rueda[2,*]

[1]Laboratório Nacional de Computação Científica - LNCC, Avenida Getúlio Vargas, Petrópolis, Rio de Janeiro 25651075, Brazil
[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Unidad Académica del Estado de Yucatán, Carretera Sierra Papacal, Mérida 97302, Yucatán, México
[3]Microbiology and Immunology, Dalhousie University, 5850 College Street, Halifax B3H 4H7, Nova Scotia, Canada
[4]Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad Universitaria, Circuito Escolar S/N, Mexico City 01000, México
*Corresponding authors. Ernesto Pérez-Rueda. E-mail: ernesto.perez@iimas.unam.mx; André Borges Farias. E-mail: bfarias.andre@gmail.com



Data

Clean and process data

Train and model evaluation

Explainable

TFBS

CAGAGCAGGAAAATCCATA
CACAGCAGATTAATCCATA
TATAGCAGGAAGATCCATA
CGCAGCAATTTAATCCATA
TTCAGCAGGATAATGAATA
CATAGCATTTTTATCCATA

CACGGCAGAAAAGTCCACA
TTTAGCGGGAAAAGACATA

Features Generation
Elimination of duplicates

-0.2; -7.8; -3.2; -2.4; -5.5; ...
-5.5; -6.2; -3.2; -7.8; -3.2; ...
-0.2; -7.8; -3.2; -2.4; -5.5; ...
-5.5; -6.2; -3.2; -7.8; -3.2; ...
-0.2; -7.8; -3.2; -2.4; -5.5; ...
-5.5; -6.2; -3.2; -7.8; -3.2; ...
-0.2; -7.8; -3.2; -2.4; -5.5; ...
-5.5; -6.2; -3.2; -7.8; -3.2; ...

Features randomization

-0.5; -6.3; -2.7; -1.6; -1.5; ...
-1.2; -2.3; -4.3; -1.5; -4.2; ...
-0.7; -4.2; -0.5; -3.2; -3.6; ...
-0.8; -3.1; -0.2; -0.7; -5.2; ...
-1.2; -0.6; -3.5; -0.5; -4.2; ...
-4.2; -2.3; -4.2; -3.2; -1.5; ...
-1.2; -2.5; -4.6; 4.3; -1.5; ...
-3.2; -2.3; -7.8; -7.8; -3.5; ...

Subset of TFBS by sequence length

Validations metrics

ML model to predict and classify TFBS

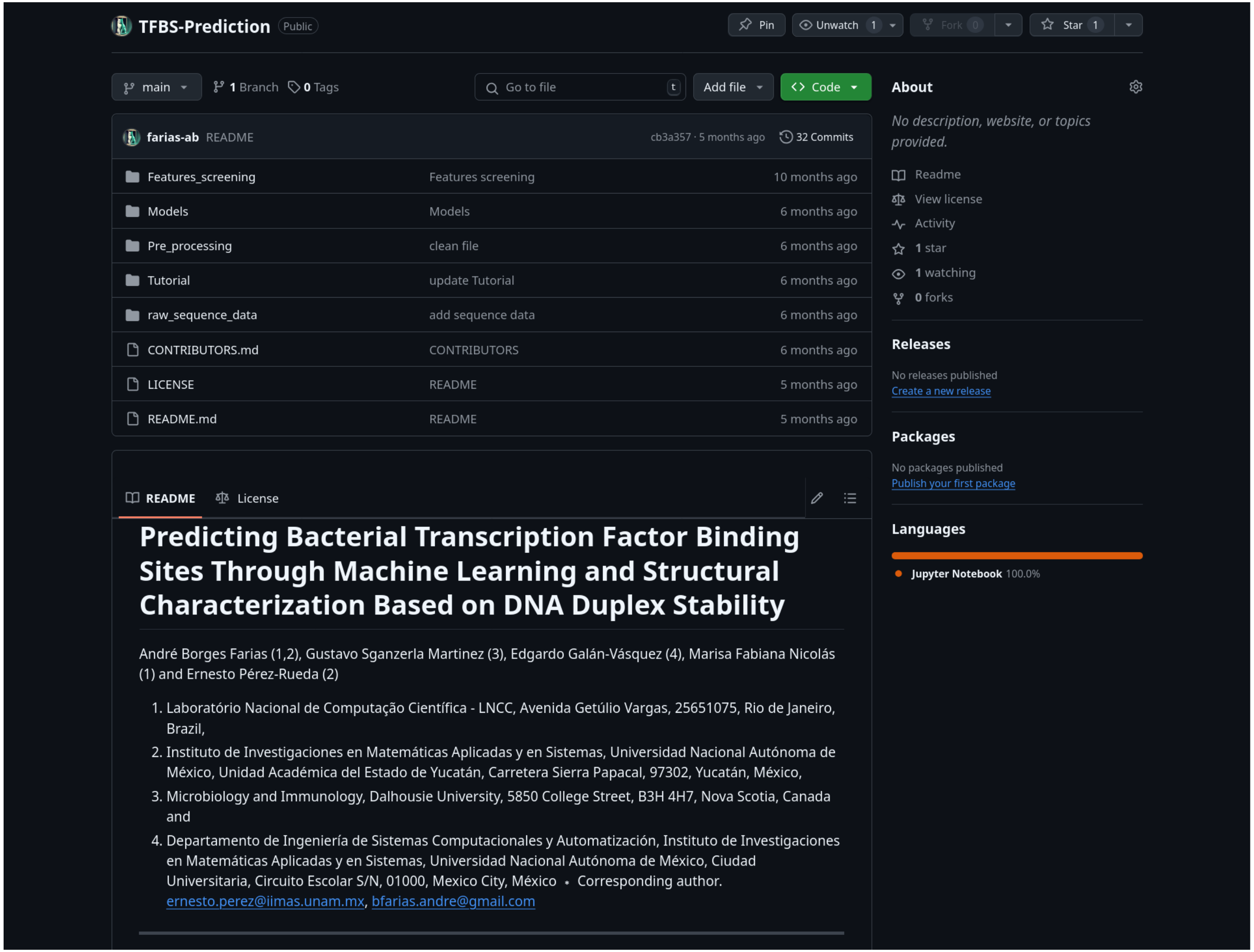
Classification task:
TFBS-DR vs TFBS-IR

Classification task:
TFBS vs non-TFBS

Protein-DNA complex

DDS plot (TFBS-IR)

# Git clone https://github.com/farias-ab/TFBS-Prediction.git



farias-ab README cb3a357 · 5 months ago 32 Commits

| | | |
|---|---|---|
| Features_screening | Features screening | 10 months ago |
| Models | Models | 6 months ago |
| Pre_processing | clean file | 6 months ago |
| Tutorial | update Tutorial | 6 months ago |
| raw_sequence_data | add sequence data | 6 months ago |
| CONTRIBUTORS.md | CONTRIBUTORS | 6 months ago |
| LICENSE | README | 5 months ago |
| README.md | README | 5 months ago |

**About**

*No description, website, or topics provided.*

- Readme
- View license
- Activity
- **1** star
- **1** watching
- **0** forks

**Releases**

No releases published
Create a new release

**Packages**

No packages published
Publish your first package

**Languages**

● **Jupyter Notebook** 100.0%

# Predicting Bacterial Transcription Factor Binding Sites Through Machine Learning and Structural Characterization Based on DNA Duplex Stability

André Borges Farias (1,2), Gustavo Sganzerla Martinez (3), Edgardo Galán-Vásquez (4), Marisa Fabiana Nicolás (1) and Ernesto Pérez-Rueda (2)

1. Laboratório Nacional de Computação Científica - LNCC, Avenida Getúlio Vargas, 25651075, Rio de Janeiro, Brazil,
2. Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Unidad Académica del Estado de Yucatán, Carretera Sierra Papacal, 97302, Yucatán, México,
3. Microbiology and Immunology, Dalhousie University, 5850 College Street, B3H 4H7, Nova Scotia, Canada and
4. Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad Universitaria, Circuito Escolar S/N, 01000, Mexico City, México • Corresponding author. ernesto.perez@iimas.unam.mx, bfarias.andre@gmail.com