



HAL
open science

Using positional information for predicting transcription factor binding sites

Raphaël Romero, Christophe Menichelli, Jean-Michel Marin, Sophie Lèbre,
Charles-Henri Lecellier, Laurent Brehelin

► To cite this version:

Raphaël Romero, Christophe Menichelli, Jean-Michel Marin, Sophie Lèbre, Charles-Henri Lecellier, et al.. Using positional information for predicting transcription factor binding sites. SMPGD: Statistical Methods for Post Genomic Data, Jan 2019, Barcelone, Spain. . hal-02068254

HAL Id: hal-02068254

<https://hal.science/hal-02068254>

Submitted on 14 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Using positional information for predicting transcription factor binding sites

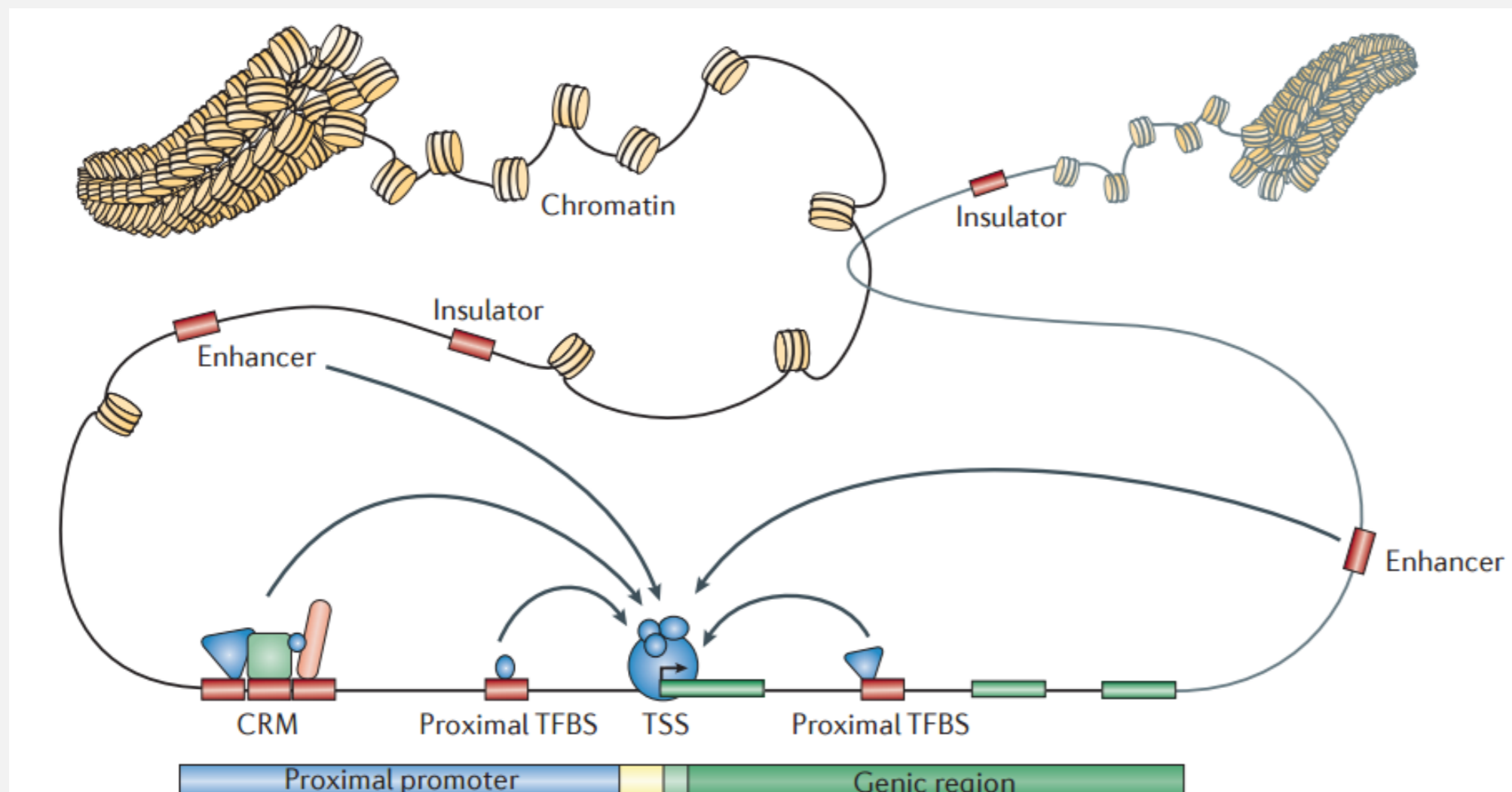
Raphaël Romero, Christophe Menichelli, Jean-Michel Marin, Sophie Lèbre*, Charles-Henri Lecellier*, Laurent Bréhélin*

IMAG, LIRMM, IBC, IGMM, Univ. Montpellier, CNRS

* These authors contributed equally to this work

Introduction

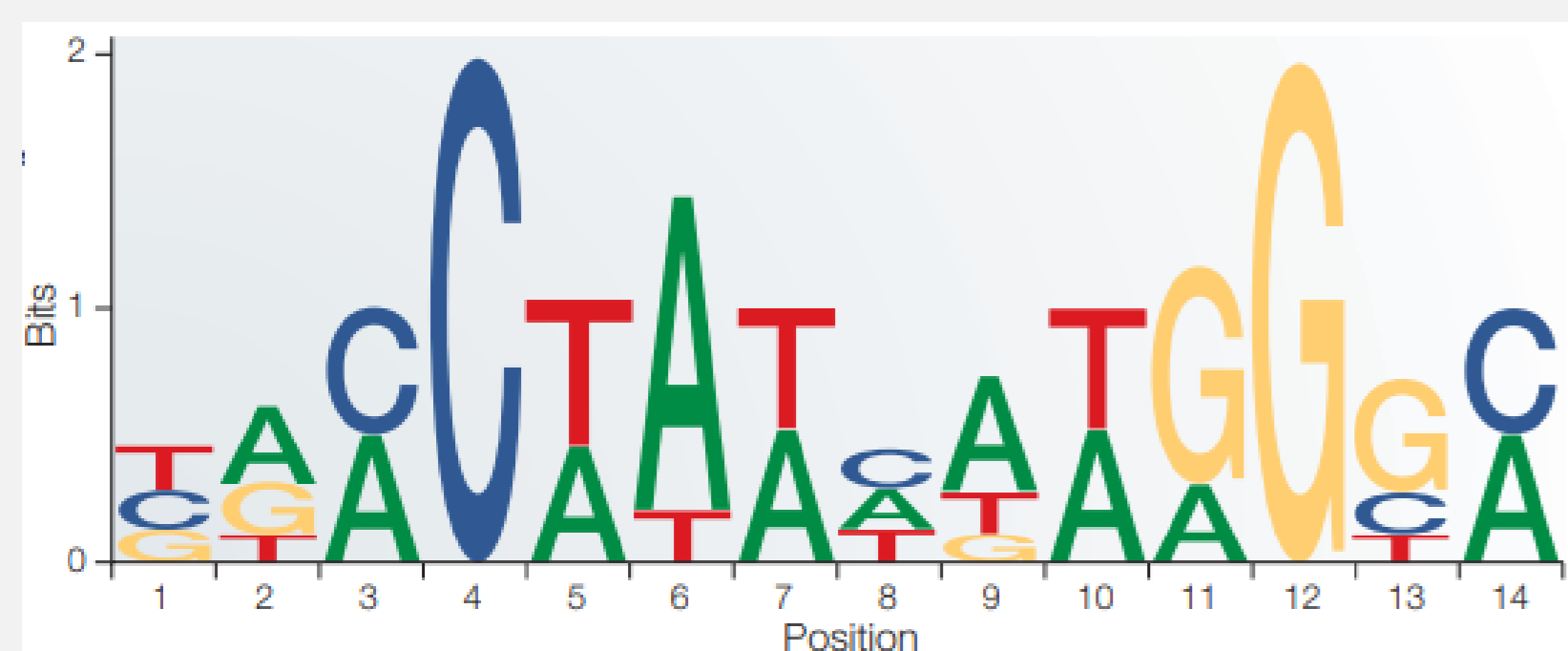
Transcription factors (TF) are proteins that play a central role in the mechanism of transcription. **Those proteins bind DNA in promoters** (the region around the Transcription Start Sites -TSS- of each gene) **or in enhancers** (a region distant from TSS but also associated with gene regulation). Each TF usually recognises and binds a specific set of k-mers. TFs often bind together with other TFs and **cooperate to regulate the transcription**.



Adapted from Lenhard *et al.*[1].

Scoring TF binding affinity

The set of k-mers recognised by a TF are usually slight variations of a common motif that can be resumed in a probabilistic model known as a binding motif or **Position Weight Matrix (PWM)**. PWMs of many TFs are available in dedicated database (JASPAR [2]).



TF: MEF2, image adapted from Wasserman and Sandelin April 2004[3]

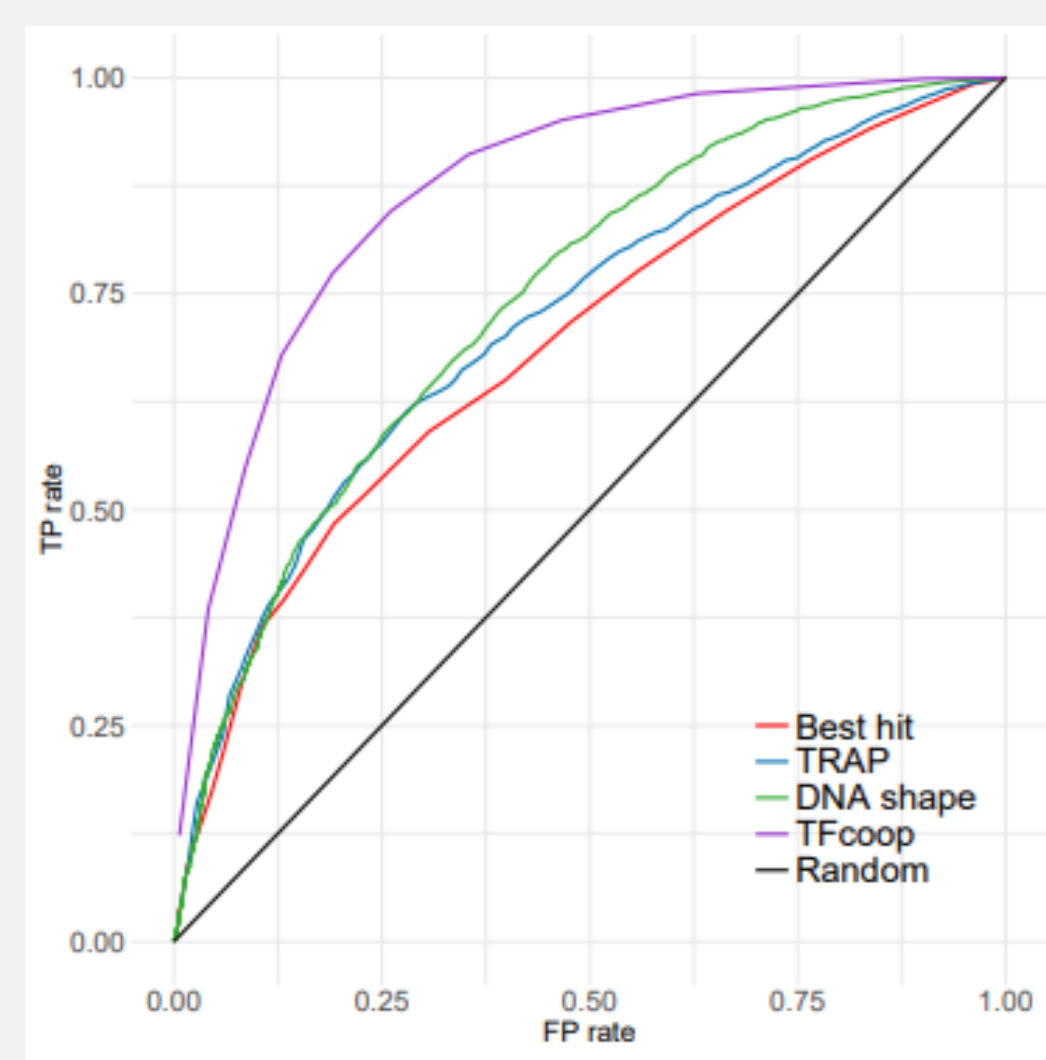
Such motif can be used to scan a DNA sequence and to identify potential binding sites of the associated TF. At each position, a score is computed on the basis of the probability to generate the k-mer that starts at this position with PWM. The affinity of the TF for a given DNA sequence is estimated by the maximal score computed at any position of the sequences (Best-hit approach). This approach is known to suffer from low accuracy, with lot of false positives.

TFcoop

- TFcoop [4] is a recent statistical approach which considers PWM scores of all TFs possibly cooperating with the target TF.
- 22.000 sequences (promoters) centred around the TSS (size = 1kbp).
- Predictive variables: binding affinities of JASPAR PWMs ($S_i, i \in \{1, \dots, p\}$ $p = 662$) and dinucleotide rates of the sequences ($R_j, j \in \{1, \dots, q\}$ $q = 12$).
- Predicted variable Y: Promoters bound or not bound by the TF.
- Logistic regression with L1 penalisation - **LASSO**

$$\ln \frac{\mathbb{P}(Y = 1 | X)}{1 - \mathbb{P}(Y = 1 | X)} = \beta_0 + \sum_{i=1}^p \beta_i S_i + \sum_{j=1}^q \gamma_j R_j + \epsilon$$

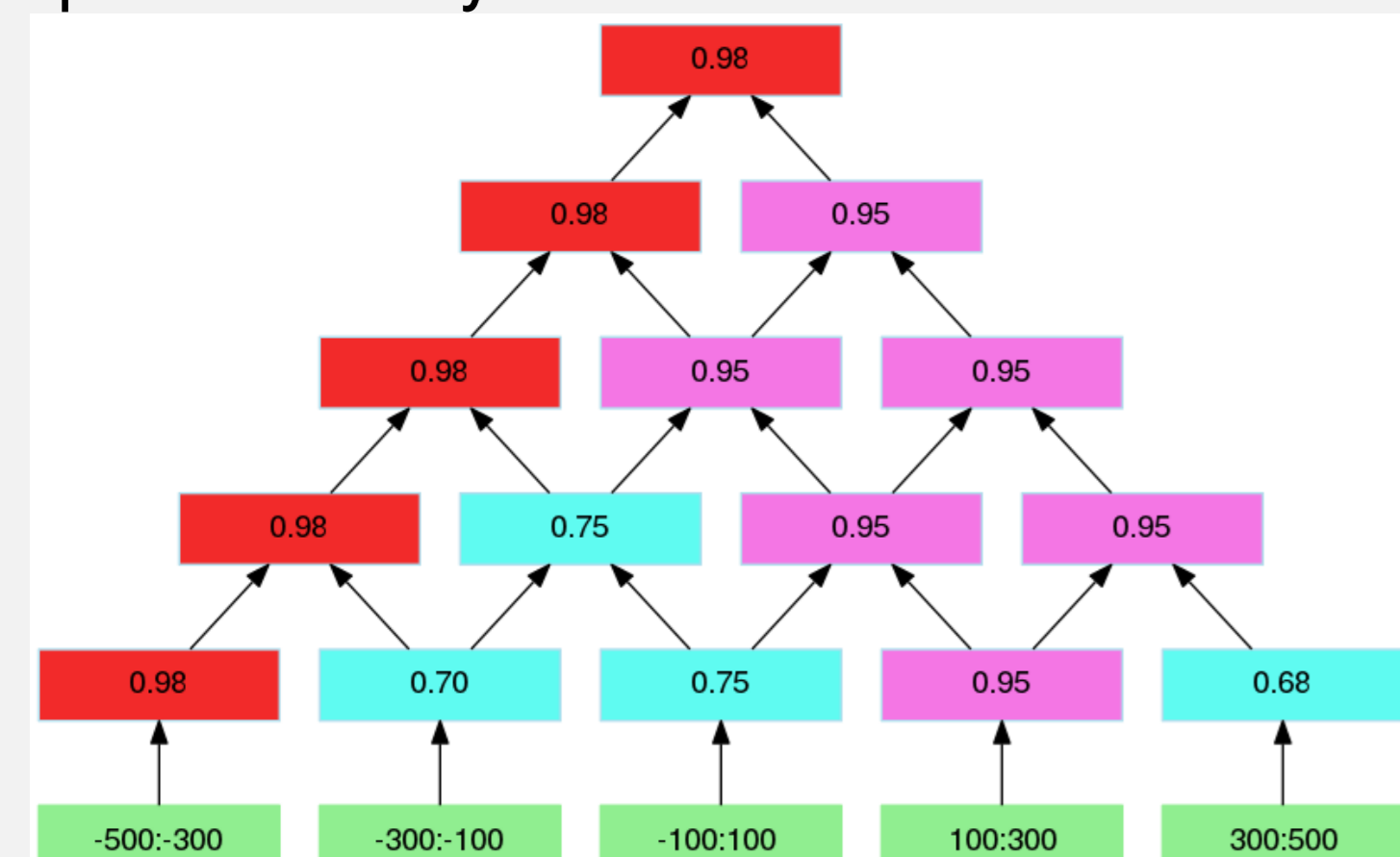
Where $X = (S_1, \dots, S_p, R_1, \dots, R_q)$ and $\epsilon \sim N(0, \sigma^2)$



- TFcoop outperforms the single PWM method (Best-Hit) in term of prediction accuracy
- Additional information may help to improve the accuracy of this approach. **We propose to use positional information of TFs occurrences**

Principle of our approach

Computing binding affinities in various sub-sequences rather than in the entire sequence solely. For each PWM and each promoter sequence a lattice is computed:



- Each node of the lattice contains the binding affinity of PWM for the associated sub-sequence.
- Allows to quickly explore many different sub-sequences.



Methods

- For each PWM, compute the lattices associated with each sequence
- We have two sets of lattices: bound or not bound sequences



- Each lattice position is assessed to identify the sub-sequence allowing the best discrimination between bound and unbound sequences.

New model

The most discriminative position of each PWM is used to create a new variable (positional variable $P_k, k \in \{1, \dots, p\}$) that is added to the TFcoop model:

$$\ln \frac{\mathbb{P}(Y = 1 | X)}{1 - \mathbb{P}(Y = 1 | X)} = \beta_0 + \sum_{i=1}^p \beta_i S_i + \sum_{j=1}^q \gamma_j R_j + \sum_{k=1}^p \alpha_k P_k + \epsilon$$

Simulations

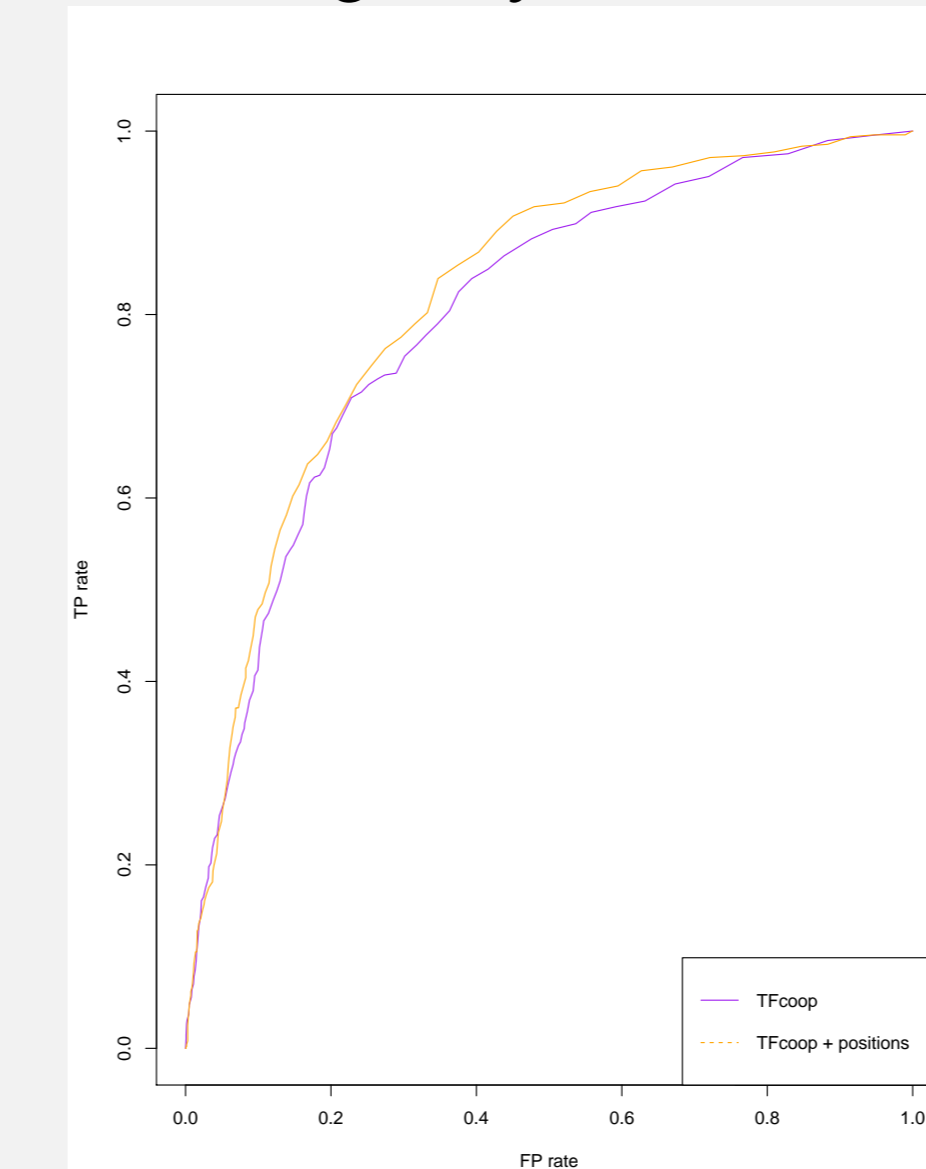
- Take one ChIP-seq experiment.
- Add some false positives occurrences in %FP of sequences, position follows a $U[0, 1000]$
- Add some true positives occurrences in %TP of sequences, position follows a $N(\mu, \sigma)$
- Run the approach and compare the identified sub-sequence R_1 to the region $R_2 = [\mu - 1, 96\sigma; \mu + 1, 96\sigma]$
- Compare both regions with Jaccard index : $J = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|}$

	TP=1%	TP=7%	TP=30%
FP=1%	0.24	0.64	0.97
FP=7%	0.24	0.75	0.97
FP=30%	0.25	0.68	0.89

Means of Jaccard indexes

Experiments

We apply our approach for discriminating the sequences bound by two different TFs sharing very similar PWMs



- A set of sequences bound by Fra1 (FOSL1) and a set of sequences bound by Fra2 (FOSL2).
- Slightly better accuracy than the TFcoop approach (AUC = 0.80 vs 0.82)

Perspectives

1. Use this method to discriminate other TF pairs with very similar motifs.
2. Improve the approach by including relative position information between TFs.
 - binary variable: the TF is upstream/downstream of another TF.
 - distance (bp) between two TF binding sites.

Huge increase of the number of variables: adequate strategies are needed.

References

- [1] B. Lenhard, A. Sandelin, and P. Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *13(4):233–245*.
- [2] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *32:D91–94*.
- [3] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics, 5(4):276–287*, April 2004.
- [4] J. Vandel, O. Cassan, S. Lebre, C. Lecellier, and L. Bréhélin. Probing transcription factor combinatorics in different promoter classes and in enhancers. *BMC Genomics - In press 2019*, March 2018.

Acknowledgements

Thanks to Fabienne Bejjani and Isabelle Jariel-Encontre from IGMM for providing us Fra1/Fra2 data.