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. Annual Workshop on Statistical Methods for Post Genomic Data. hal-02068254

HAL Id: hal-02068254
https://hal.archives-ouvertes.fr/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.
Introduction

Transcription factors (TF) are proteins that play a central role in the mechanism of transcription. These proteins bind DNA in promoters (the region around the Transcription Start Sites -TSSs- of each gene) or in enhancers (a region distant from TSS but also associated with gene regulation). Each TF usually recognizes a sequence solely. For each PWM and each promoter sequence a lattice is computed:

\[ P(Y = 1 | X) = \beta_0 + \sum_{i=1}^{p} \beta_i S_i + \sum_{j=1}^{q} \gamma_j R_j + \epsilon \]

Where \( X = (S_1, \ldots, S_p, R_1, \ldots, R_q) \) 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.

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) centered around the TSS (size = 1kb).
- Predictive variables: binding affinities of JASPAR PWMs \( (S_i, i \in \{1, \ldots, p\}) \)
- Predicted variable \( Y \): Promoters bound or not bound by the TF.
- Logistic regression with L1 penalisation: LASSO

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

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 \( P_i, i \in \{1, \ldots, p\} \)) that is added to the TFcoop model:

\[ \ln \left( \frac{P(Y = 1 | X)}{1 - P(Y = 1 | X)} \right) = \beta_0 + \sum_{i=1}^{p} \beta_i S_i + \sum_{j=1}^{q} \gamma_j R_j + \sum_{i=1}^{p} \alpha_i P_i + \epsilon \]

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 \( N(1.96 \sigma, \sigma) \).
- Run the approach and compare the identified sub-sequence \( R_i \) to the region \( R_2 = [\mu - 1.96 \sigma, \mu + 1.96 \sigma] \).

<table>
<thead>
<tr>
<th>%TP</th>
<th>%FP</th>
<th>%TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>30%</td>
<td>0.24</td>
</tr>
<tr>
<td>7%</td>
<td>30%</td>
<td>0.24</td>
</tr>
<tr>
<td>30%</td>
<td>30%</td>
<td>0.25</td>
</tr>
</tbody>
</table>

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


Acknowledgements

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