Accurate prediction of *in-vivo* transcription factor binding across cell types

Jan Grau[1], Stefan Posch[1], and Jens Keilwagen[2]

[1]*Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany*
[2]*Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut, Quedlinburg, Germany*
grau@informatik.uni-halle.de

Transcriptional regulation mediated by transcription factors binding to genomic DNA is one of the fundamental regulatory steps of gene expression. Most transcription factors bind to DNA with some sequence specificity, which is typically described by sequence motifs. Such sequence motifs may be inferred from experimental data, e.g., from ChIP-seq experiments, and current high-throughput techniques and large-scale projects like ENCODE allowed for determining motifs of a large collection of human transcription factors.

Although the number of *in vivo* datasets increases, it is still not possible to perform ChIP-seq experiments for every transcription factor against all cell types or tissues under all possible physiological conditions. Hence, accurate and high-resolution computational approaches are necessary to close this gap and complement experimental results. The goal of the "ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge" (`https://www.synapse.org/#!Synapse:syn6131484/wiki/402026`) was to identify the best-performing approach for predicting *in vivo* transcription factor binding across different cell types and tissues. To this end, the organizers provided for several training cell types i) ChIP-seq data for several transcription factors, ii) DNase-seq data representing chromatin accessibility and iii) RNA-seq data. The task for participating approaches was then to predict transcription factor binding in held-out test cell types on held-out chromosomes based on the corresponding DNase-seq and RNA-seq data.

Here, we present an approach for solving this task based on i) diverse features determined from DNase-seq data, ii) motif models of different complexity and origin including motifs from de-novo motif discovery on the DREAM training regions, and iii) further features derived from raw sequence, gene annotations, as well as RNA-seq data. We use sparse local inhomogeneous mixture (Slim) models [1] for representing motifs, which are capable of modeling intra-motif dependencies. In the ENCODE-DREAM challenge, this approach gained a shared first rank among 40 international teams.

In this approach, features are determined genome-wide in non-overlapping 50 bp windows represented by aggregate values, which are modelled by independent Gaussian (continuous features) and multinomial (discrete features) models. Model parameters are optimized with respect to the discriminative maximum conditional likelihood (MCL) principle. While the foreground training regions can be derived directly from ChIP-seq positive regions, it is less obvious how to determine a representative background data set. In extension of boosting-like methods, we use an iterative training approach. Starting from an initial training set, the negative training data are iteratively complemented by additional training examples of the background class that scored especially bad in the previous iteration. Models are learned for each individual combination of transcription factor and cell type of the training data, and for three individual iterations of the iterative training process. Predictions on the test data are determined as the average of the predicted a-posteriori foreground probabilities of all classifiers for a transcription factor using DNase and RNA-seq data for the target cell type.

In post-challenge studies, we found that this "wisdom of crowds" approach averaging over the prediction based on multiple training cell types typically performed better than the median performance of individual cell types. In addition, we discovered that DNase-seq-derived features and motif-based features are the most important determinants of prediction performance, which facilitates a reduction of model complexity while preserving high-quality predictions.

References

[1] Jens Keilwagen and Jan Grau. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Research*, 43(18):e119, 2015.