Unifying open chromatin assays using supervised learning for transcription factor binding prediction

Florian Schmidt[1,3,4], Jonas Fischer[1,3], Karl J Nordstroem[2], Nina Gasparoni[2], Gilles Gasparoni[2], Kathrin Kattler[2], Nico Pfeifer[1], Joern Walter[2], Marcel H Schulz[1,3]

[1]Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, Germany, [2]Department of Genetics, Saarland University, [3] Cluster of Excellence on Multimodal Computing and Interaction, Saarland University, and [4]Graduate School of Computer Science, Saarland University

Transcription factors (TFs) frequently bind to DNA in open chromatin regions [4]. Therefore, using integrative models incorporating information on open chromatin might improve TF binding prediction. Different assays have been proposed to measure open chromatin genome-wide. We compare three of these techniques, NOMe, DNase1, and ATAC-seq to understand differences and commonalities between them. Despite a significant overlap, each method has its own advantages in finding open chromatin regions resulting in method specific assessment of open chromatin. Thus, predicting transcription factor binding in open chromatin regions defined by different assays can lead to contradictory results. This discrepancy could be reduced by improving open chromatin assessment, e.g. by computational post processing of the experimental results. Here, we propose an efficient computational pipeline that combines an SVM-based open chromatin classifier with motif based TF affinity predictions [1]. To test and optimise our pipeline, we use a comprehensive gold standard data set of open chromatin regions using DNase1, NOMe, and ATAC data generated in the DEEP project. Validation of our pipeline on ChIP-seq TF data from ENCODE shows that applying the classifier improves the precision recall AUC of TF binding predictions compared to common approaches.

Motivation

We compare three established methods for open chromatin assessment: DNase1, NOMe and ATAC. Due to the disagreement shown in Figure 1, the performance of TF prediction strongly depends on the chosen open chromatin assessment method. Unifying the assays using supervised learning might decrease that dependency.

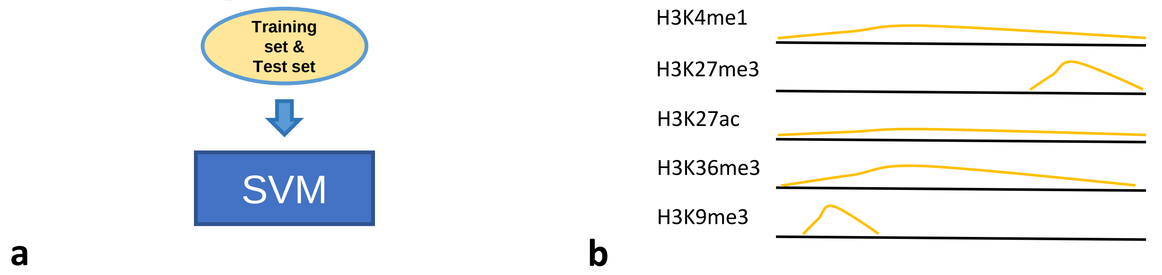

Figure 3: (a) For training and testing the SVM, we use a gold standard set, consisting of the intersection of NOMe and ATAC data. On the basis of that, a background data set representing closed regions is computed. As features, we use DNase1 and/or HM ChIP-Seq data. The data is splitted into 19690 samples for training and 59047 samples for testing. (b) We apply binning to the DNase1 signal as well as to the HM signals to construct a feature set. We use six bins, each of length 100.

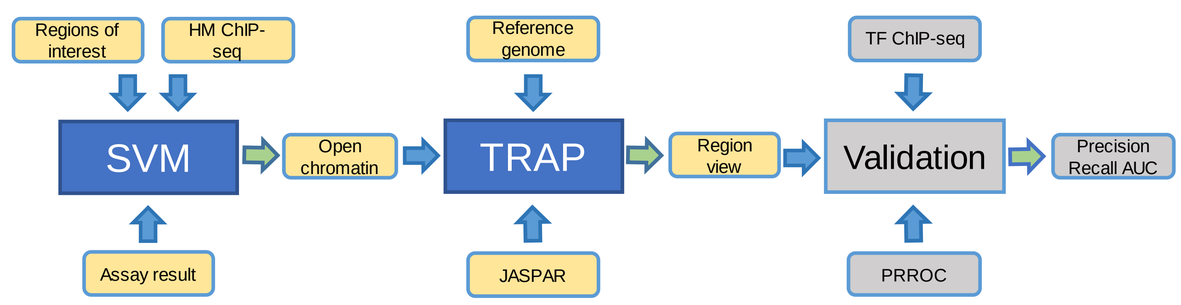

Figure 4: Workflow of our TF binding prediction pipeline. There are three main steps: classification through the SVM, TF binding prediction using TRAP, and validation. As validation is not directly related to pure usage of the pipeline, all its components are shown in grey.

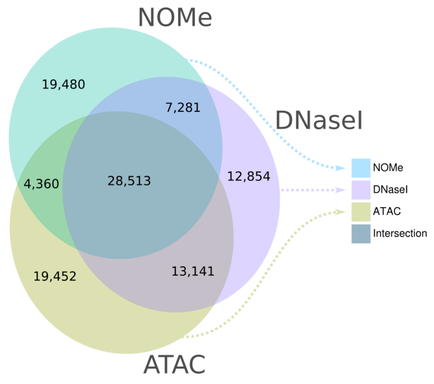

Figure 1: Open chromation regions (peaks) were predicted on HepG2 cells using three different technologies, ATAC-seq, NOMe, and DNase1. The venn diagramm shows the overlaps between peaks from the three assays.

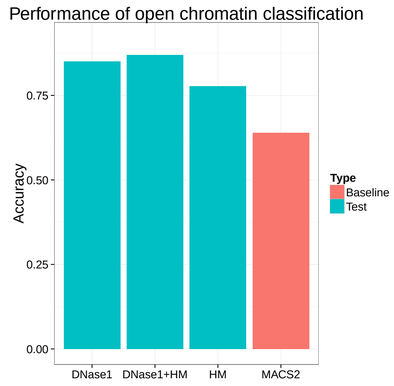

Figure 2: Test accuracy of the classifier using different feature sets. The MACS2 baseline shows the accuracy of predicting the gold standard regions using DNase1 peak calls.

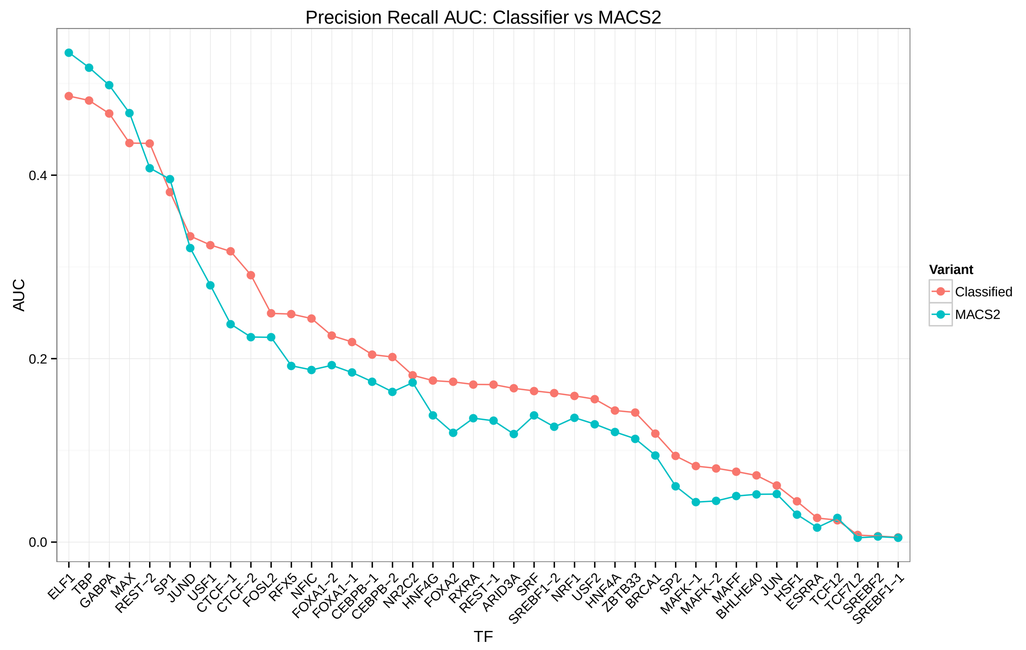

Figure 5: Preciscion recall AUC curve for TF binding predictions in 1000bp regions centered around the TSS of all protein coding genes. Regions are considered for the prediction if they are classified as open (orange) or overlap with a MACS2 [2] DNase1 peaks (cyan). In this setting, we use both DNase1 and HMs as features. In most cases, predictions on classified regions achieve better precision recall AUCs then predictions on MACS2 overlapping regions.

Methods

To classify open chromatin, we train a Support Vector Machine (SVM), using a nonlinear RBF kernel. We exploit the intersection of ATAC and NOMe data as a gold standard of open chromatin regions. On the basis of the gold standard data set a background data set representing closed chromatin is computed. To generate features, we perform binning on the signal of DNase1 and several Histone Marks (HMs). An overview of the aforementioned steps is shown in Figure 3. We implemented the classifier using LIBSVM [6]. To perform the actual TF binding prediction, we use the R-package TRAP [1]. Within this tool, we employ the current version of PWMs from JASPAR [5] to compute TF binding predictions in regions classified as open. We validate the predictions by calculating precision recall AUCs using PRROC [3]. To this end, we download 42 TF ChIP-seq data sets from ENCODE. For our analysis, we use data on HepG2. An overview on the workflow for TF binding prediction is shown in Figure 4.

Results

As a first usecase, we apply our pipeline on the gold standard data set consisting of ATAC and NOMe data using three different feature sets: DNase1, HMs, and both. We splitted the data into 19690 training samples and 59047 test samples. The best accuracy was achieved by the classifier using both DNase1 and HMs as features. Thus, we use this classifier in the next step to perform TF binding predictions. In addition, we computed the accuracy of predicting open regions in the gold standard data set using only peak calls on the DNase1 signal produced by MACS2. This accuracy is generally lower than the accuracy achieved by our classifier (Figure 2). To test the performance of our pipeline, we generate 1000bp windows centered around the TSS of all human protein coding genes. We perform TF binding predictions in a window if it is classified as open, or if it overlaps with a DNase1 peak. In the end, we compute precision recall AUCs for all 42 TF ChIP seq data sets we obtained from ENCODE. In most cases, we find that usage of our classifier leads to a better precision recall AUC (Figure 5).

Conclusion & Outlook

- The open chromatin classifier improves PR-AUC for TF binding predictions in this simple promoter setup. We want to improve the setup to make it more realistic.
- We are currently applying the same classifier setup for NOMe and ATAC-seq.
- It is necessary to further investigate the importance of HM in open chromatin classification.
- We want to add additional matrices from Transfac and UniPROBE to cover more TFs.

References

1) *Predicting transcription factor affinities to DNA from a biophysical model. Roider. H-G, Kanhere. A, Manke. T, Vingron. M. Bioinformatics, 2007.*

2) *Model-based analysis of ChIP-Seq (MACS). Zhang. Y, Liu. T, Meyer. CA, Eeckhoute. J, Johnson. DS, Bernstein. BE, Nusbaum. C, Myers. RM, Brown. M, Li. W, Liu. XS. Genome Biology, 2008.*

3) *PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Grau. J, Grosse. I, Keilwagen. J. Bioinformatics, 2015.*

4) *Predicting cell-type-specific gene expression from regions of open chromatin. Natarajan. A, Yardımcı. G, Sheffield. N, Crawford. G, Ohler. U. Genome Research, 2012.*

5) *JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles Mathelier. A, Zhao. X, Zhang. A W, Parcy. F, Worsley-Hunt. R, Arenillas. D J, Buchman. S, Chen. C-y, Chou. A, Ienasescu. H, Lim. J, Shyr. C, Tan. G, Zhou. M, Lenhard. B, Sandelin. A, and Wasserman. W W. Nucleic Acids Research, 2014.*

6) *LIBSVM: A library for support vector machines. Chang. C-C, Lin. C-J. ACM Transactions on Intelligent Systems and Technology, 2011.*

CLUSTER OF EXCELLENCE · UNIVERSITÄT DES SAARLANDES · max planck institut informatik · CBI CENTER FOR BIOINFORMATICS · SAARLAND UNIVERSITY · SAARBRÜCKEN GRADUATE SCHOOL OF COMPUTER SCIENCE