Analysis of short- and long-range interactions within potential binding sites notably extends the fraction of verified peaks in ChIP-seq data

Anton Tsukanov, Victor Levitsky, Tatyana Merkulova

We developed pipeline for integrative application of various de novo motif search tools to massive sequencing data

Our pipeline includes four de novo models:

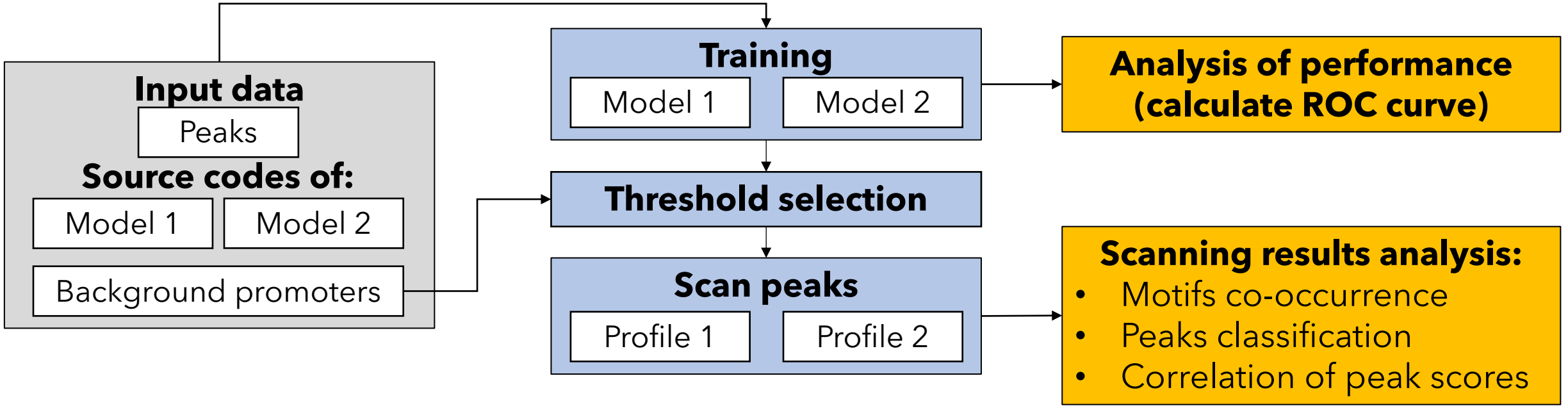
- Position Weight Matrix PWM (ChIPMunk, Kulakovskiy et al., 2010),
- InMoDe (Eggeling et al., 2017),
- BaMM (Siebert and Söding, 2016), SiteGA (adopted for de novo search from Levitsky et al., 2007)

Pipeline also integrates methods to search and combine TFBS profiles

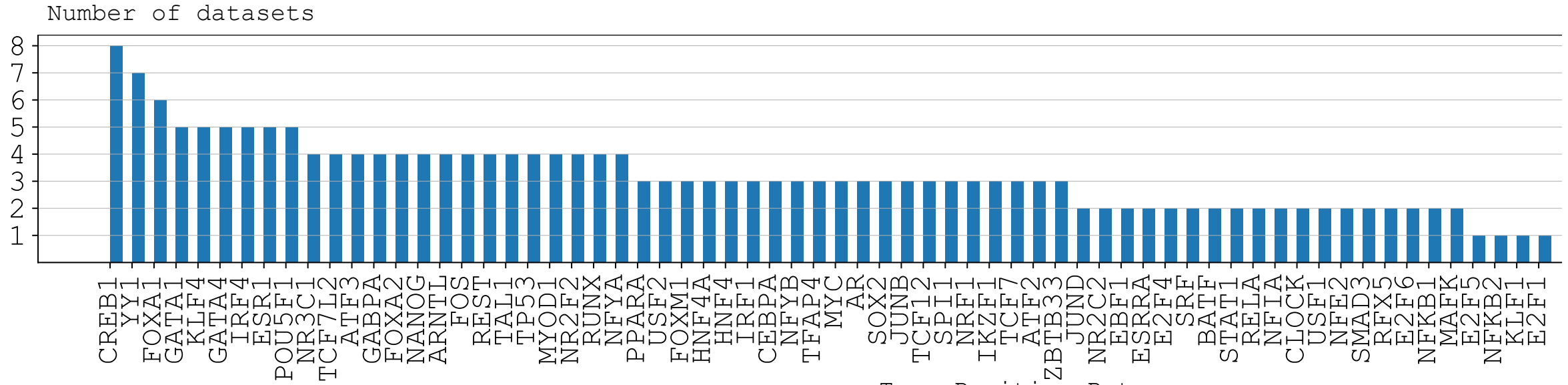
- Traditional PWMs neglect dependencies between positions of motifs,
- 'Short-range interactions' markov models BAMM/InMode permit only local dependencies within closest 5-6 bp,
- 'Long-range interaction' model SiteGA allows dependencies between arbitrary positions

The pipeline includes several main steps:

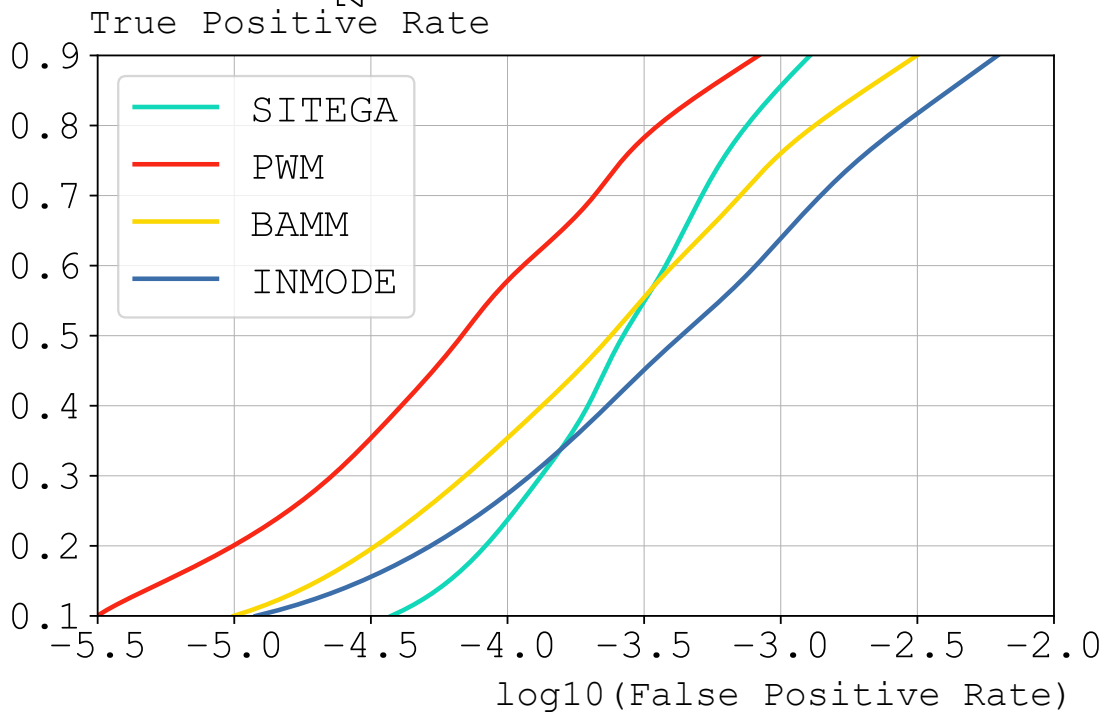
- training models;
- estimating models accuracy;
- calculating models thresholds;
- recognition of potential sites in peaks;
- classification of peaks according the presence of sites and their overlaps within a peak sequence



Totally, we took in analysis 211 ChIP-seq datasets for 66 human/mouse transcription factors from the Cistrome database (http://cistrome.org/)



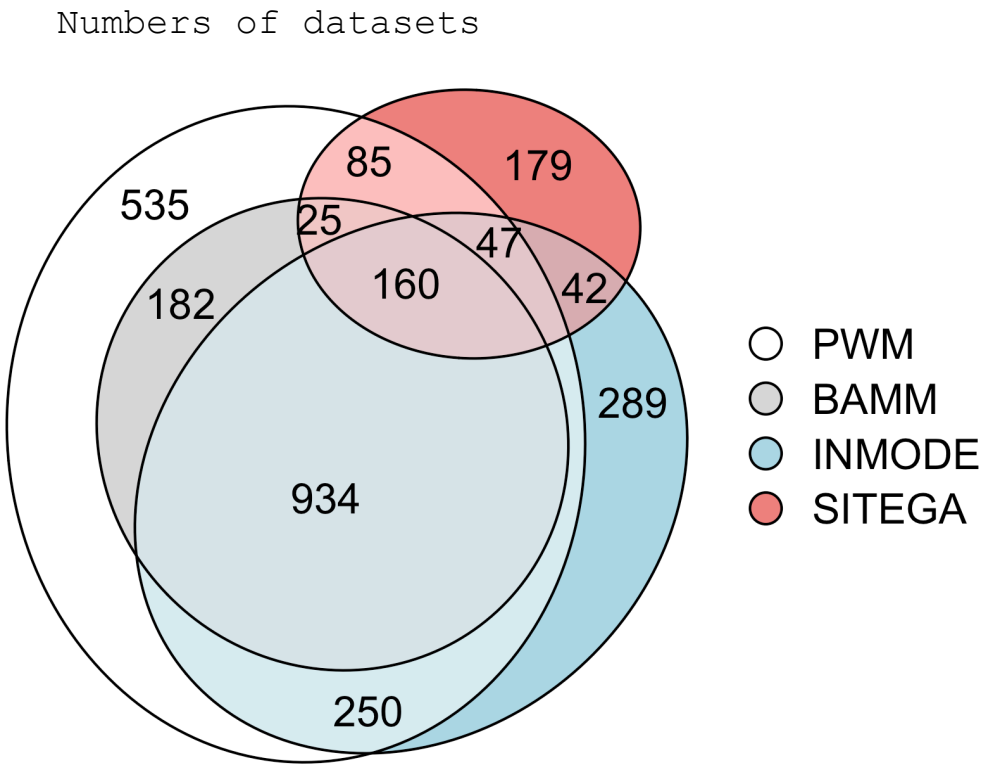
We performed the bootstrap cross-validation procedure to estimate the performance of PWM, BaMM, InMode and SiteGA models.
The median results of bootstrap cross-validation for all datasets (ROC curves) are shown on the right figure

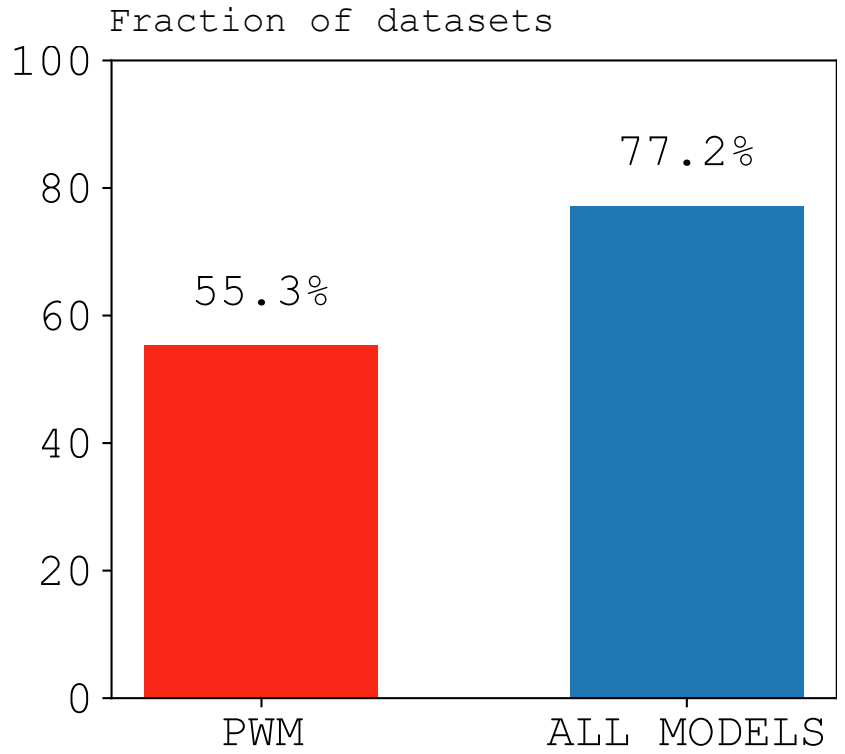


For each models were calculated list of recognition thresholds through the computation of FPR for all protein coding genes promoters. After that peaks of each datasets were scanned by de novo models. Results of scanning by different models were compared.

The average fractions of peaks containing hits of sole models PWM, BaMM, InMoDe and SiteGA are equal to 4.5%, 5.4%, 3.8% and 4.6%, respectively.

The example results of pipeline application for ESRRA ChIP-seq dataset (GEO: GSM2424191, CistromeDB: 100495, cell line: MCF-7) are shown in the right figure.





The sole PWM model verifies the average fraction of almost 55.3% of verified peaks, see left figure. The combination of all four models PWM, InMoDe, BaMM and SiteGA models provides the average fraction of 77.2% of verified peaks.

Hence, application of BAMM/InMode/SiteGA models, on average, verifies about 21.9% of additional peaks in comparison with the respective fraction of sole PWM model.