doi 10.18699/bgrs2024-1.2-17

MetArea tool for predicting structural variability and cooperative binding of transcription factors in ChIP-seq data

Levitsky V.G.[1, 2]*, Tsukanov A.V.[1]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Novosibirsk State University, Novosibirsk, Russia*

** levitsky@bionet.nsc.ru*

Key words: transcription factor binding site motifs; performance of motif recognition; composite elements; recognition performance of motif models

Motivation and Aim: The rapidly growing amount of data on chromatin immunoprecipitation followed by massive sequencing (ChIP-seq) facilitates the search for the motifs of transcription factors (TFs) specific to a given cell or tissue type and condition [1]. Typically, two or more TFs work together to induce transcriptional changes [2]. However, progress in the development of methods for identifying the motifs of TFs involved in functionally important co-operative interactions still lags far behind the development of tools for finding the motifs of individual TFs.

Methods and Algorithms: For the model of TF binding site motif recognition, we calculate the recognition accuracy as the partial area under ROC (Receiver Operating Characteristic) curve (pAUC, [3]). Generalizing the ROC curves and pAUC values of the two separate models to a single joint model reveals whether the joint performance of two motifs competes with those of the participating motifs. Testing multiple possible combinations of motifs reveals the motifs most strongly reinforcing each other. These pairs of motifs can represent either structurally different types of binding sites for the same TF, or binding sites of different TFs acting as part of a single multi-protein complex. Thus, MetArea predicts motifs with synergistically related functions in gene transcription regulation. The synergy criterion requires for the pair of motifs 1 and 2 the higher pAUC1&2 value of the joint model 1&2 compared to the pAUC values of both participating motif models pAUC1 and pAUC2, Ratio = pAUC(1&2) / Max(pAUC1, pAUC2) > 1. MetArea considers ready ROC curves for TF binding site motif models, thereby it is applied to the motifs of the traditional model (Position Weight Matrix, PWM) or an alternative one, e.g. SiteGA [3].

First, MetArea algorithm computes the ROC curves for both motif models. For each motif model this curve represents the relationship between the rates of true positive (TP) and false positive (FP) predictions. For each recognition threshold, the TP rate denotes the fraction of nucleotide sequences from the foreground set (ChIP-seq peaks) containing predicted sites of a motif model. For each recognition threshold, the FP rate is defined as the frequency of the motif in the background set of nucleotide sequences [3]. The background set means genomic sequences with A/T content matching that for the sequences from the foreground set. Preliminary, for each motif model, the promoter sequences of protein coding genes from whole genomes are used to compute – Log_{10}(ERR) values (Expected Recognition Rate, ERR). These ERR values measure various motif models in the uniform scale [3, 4]. Thus, in terms of the FP rate computed through ERR values, for any pair of sites of two models, we can estimate which of two sites respects higher affinity.

Next, MetArea algorithm computes the ROC curves for the combination of two motif models. For them, the joint TP rate is the fraction of sequences containing predicted sites of at least one of two motif models. The joint FP rate requires the correction since predicted sites of two motif models may overlap each other in a nucleotide sequence. Only the three distinct cases are possible in one DNA sequence for two sites of two motif models: two sites do not overlap; they have full overlap, i.e. one of them is located entirely inside the other; they have partial overlap, i.e. neither one can be completely located within another. Correspondingly, MetArea corrects the FP rate for the joint motif model as follows. If two sites do not overlap, then the joint FP rate increases by two, $\Delta FP = 2$. If two sites have full overlap, then the joint FP rate increases by one, $\Delta FP = 1$. If two sites have partial overlap, then the joint FP rate increases by the non-integer impact between one and two ($1 < \Delta FP < 2$), it depends on the ratios between the overlap length (L) of two motifs, and the lengths of motifs W_1 and W_2,

$$\Delta FP = \frac{1}{2}\left(\frac{L}{W_1} + \frac{L}{W_1}\right) = \frac{L}{2}\left(\frac{W_1 + W_1}{W_1 * W_1}\right)$$

The cases of full or partial overlaps imply that the joint FP rate possesses the best ERR score among the two sites,

$$-Log_{10}(ERR_{1\&2}) = Max\{-Log_{10}(ERR_1), -Log_{10}(ERR_2)\}$$

Results: We took in analysis the ChIP-seq dataset of 1000 top-scored peaks for BHLHA15 TF from GTRD [5], ID PEAKS039234, GEO ID GSE86289 [6], pancreas of adult mice. We performed *de novo* motif search with the traditional PWM [7], categorized [1, 8] seven top-scored motifs to the motifs of known transcription factors and sorted these motifs by the pAUC measure (Table 1). The first two and third motifs represent BHLA15 binding sites, E-boxes of GC and AT types [6], correspondingly; the rest motifs are designated by TF names. We approved that the joint models {BHLHA15_GC_1 & BHLHA15_AT}, {BHLHA15_GC_2 & BHLHA15_AT} and {PRD13 & FOXA2} achieved an increase in accuracy of about 20 % compared to the respective single models (Table 1, Fig. 1A). Thus, the first two pairs of motifs represent the combinations of GC and AT E-boxes motifs for BHLA15 (Table 1). For the first pair pAUC values for single motifs are 7.45E-4 and 4.48E-4, while the joint model shows the pAUC value of 8.91E-4 (Fig. 1B).

Table 1. Seven top-scored motifs for PEAKS039234 dataset for murine BHLHA15 TF found by *de novo* motif search [7]. For each motif its name [5, 6], logo [8] and the performance pAUC are shown

Motif	Logo	pAUC, ×10000
BHLHA15_GC_1		7.45
BHLHA15_GC_2		6.45
BHLHA15_AT		4.48
CTCF		3.31
PRD13		3.23
FOXA2		2.66
GATA4		0.73

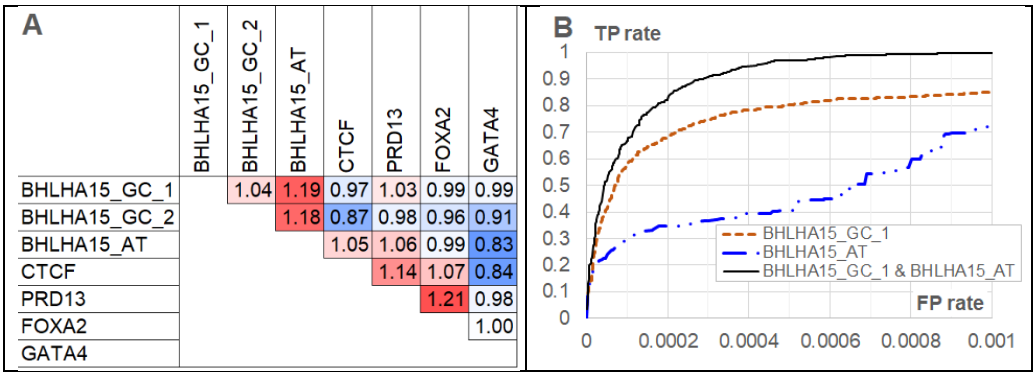

Fig. 1. Application of MetArea tool for the dataset PEAKS039234 for murine BHLHA15 TF from GTRD [5], GEO ID GSE86289 [6]. **A.** Ratio values for all 21 possible pairwise combinations of seven motifs from *de novo* motif search (Table 1, [7]). Shades of red/blue mark the Ratio values with/without the synergy of motifs. **B.** ROC curves for the motifs BHLHA15_GC_1, BHLHA15_AT (non-solid red and blue lines), and their joint model (solid black) achieving the higher performance

Conclusion: MetArea propose an approach to clarify the participants of multiprotein complexes regulating gene transcription. It can be used either to confirm distinct structural types of binding sites for the same TF, or to reveal the motif co-occurrence respecting to composite elements formed by the same or distinct motifs, homo- and heterotypic composite elements, correspondingly. Besides, MetArea tool can point to the false positive results of *de novo* motif search, since for the functional motifs the synergetic growth of the performance pAUC in the pairwise joint models is expected.
Funding: The study is supported Russian Science Foundation project No. 20-14-00140.

References

1. Vorontsov I.E., Eliseeva I.A., Zinkevich A., Nikonov M., Abramov S., Boytsov A. et al. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* 2024;52(D1):D154-D163. doi 10.1093/nar/gkad1077
2. Morgunova E., Taipale J. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol.* 2017;47:1-8. doi 10.1016/j.sbi.2017.03.006
3. Tsukanov A.V., Mironova V.V., Levitsky V.G. Motif models proposing independent and interdependent impacts of nucleotides are related to high and low affinity transcription factor binding sites in Arabidopsis. *Front Plant Sci*. 2022;13:938545. doi 10.3389/fpls.2022.938545
4. Levitsky V., Zemlyanskaya E., Oshchepkov D., Podkolodnaya O., Ignatieva E., Grosse I. et al. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Res*. 2019;47(21):e139. doi 10.1093/nar/gkz800
5. Kolmykov S., Yevshin I., Kulyashov M., Sharipov R., Kondrakhin Y., Makeev V.J. et al. GTRD: an integrated view of transcription regulation. *Nucleic Acids Res*. 2021;49(D1):D104-D111. doi 10.1093/nar/gkaa1057
6. Jiang M., Azevedo-Pouly A.C., Deering T.G., Hoang C.Q., DiRenzo D., Hess D.A. et al. MIST1 and PTF1 collaborate in feed-forward regulatory loops that maintain the pancreatic acinar phenotype in adult mice. *Mol Cell Biol*. 2016;36(23):2945-2955. doi 10.1128/MCB.00370-16
7. Bailey T.L. STREME: accurate and versatile sequence motif discovery. *Bioinformatic*s. 2021;37:2834-2840. doi 10.1093/bioinformatics/btab203
8. Gupta S., Stamatoyannopoulos J.A., Bailey T.L., Noble W.S. Quantifying similarity between motifs. *Genome Biol*. 2007;8(2):R24. doi 10.1186/gb-2007-8-2-r24