doi 10.18699/bgrs2024-1.2-20

Transcription factors with different DNA-binding domains respect different hierarchic levels such as class, family and subfamily as sufficient for significant similarity of their motifs

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

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

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

* *radicav06@gmail.com*

Key words: transcription factor binding sites; analysis of ChIP-seq data; de novo motif search; similarity of motifs; classification of transcription factors by their DNA-binding domains

Motivation and Aim: Transcription factors (TFs) are proteins that regulate transcription by binding to genomic DNA. TF binding sites (TFBS) in genomic DNA usually varied in length from 6 to 20 bp, their common pattern for particular TF is called motif. The ChIP-seq experimental technology, based on chromatin immunoprecipitation (ChIP), allows genome-wide mapping of genomic loci (peaks) potentially containing these sites for a target TF. The important task of ChIP-Seq data analysis is to detect enriched motifs in peaks. A *de novo* motif search applied to sets of foreground (peaks) and background (non-specific DNA) sequences can learn enriched motifs that represent an exact model of a target TF motif recognition [1]. The significance of similarity of the motifs of known TFs [2] from public databases [3, 4] to enriched motifs supports their relation to particular TFs. For each TF, the structure of its DNA-binding domain (DBD) defines available variability of its motif. The hierarchical classification of TFs by DBD [5] defines top two levels as superclasses and classes based on the general type of DBD and its structure. Next two levels, families and subfamilies are defined by sequence similarity of DBDs. The main idea here is to trace gradually the connection between similar DBDs and related DNA sequences. However, this classification does not have strict criteria for the significance of similarity of TFBS motifs at certain levels of TF classification. Identification of TFBS motifs is complicated by the noticeable increase in the number of TFs with known motifs (there are already more than a thousand of them [3, 4]). However, the number significantly different TFBS motifs is substantially less, since the motifs of structurally related TFs may be practically indistinguishable [6]. For some classes, there may be two or more structurally different DNA motifs per TF. E.g., TFs can bind only as dimers, or as monomers or dimers, etc. [7].

To date, for TFs of a given DBD structure (class, family, and subfamily) it has not been determined which of these levels is sufficient to identify a set of TFs with significantly similar motifs. The aim of this study is to include the systematic analysis of similarity of motifs of known TFs from different levels of TF classification to the standard protocol of *de novo* motif search applied to the results of ChIP-seq experiment. We performed the massive analysis of ChIP-seq data, estimated the similarity of motifs of different target TFs within certain classes/families and compared these distributions with those computed for distinct families within each class.

Methods and Algorithms: We extracted ChIP-seq datasets for *M. musculus* from the GTRD [8]. We used the hierarchical classification of TFs by DBDs into classes,

families [5]. Six most numerous TF classes (Table 1) were included in the analyses. We used *de novo* motif search tool STREME [1], it applies the traditional motif model of Position Weight Matrix. We applied TOMTOM tool [2] (a) to confirm the significant similarity of enriched motifs to motifs of known TFs from JASPAR/Hocomoco [3, 4] databases, and (b) to estimate the similarity in pairs of enriched motifs. To define the distributions of significance motifs similarity within a class or family, we used only pairs of distinct TFs.

Table 1. ChIP-seq datasets classified by DBD structure of target TFs

TF classes	Total count of		
	Datasets	Families	TFs
Basic leucine zipper factors (bZIP) {1.1}	96	6	16
Basic helix-loop-helix factors (bHLH) {1.2}	113	6	18
Nuclear receptors with C4 zinc fingers {2.1}	187	3	13
C2H2 zinc finger factors {2.3}	244	3	12
Fork head / winged helix factors {3.3}	92	2	6
Tryptophan cluster factors {3.5}	72	3	13

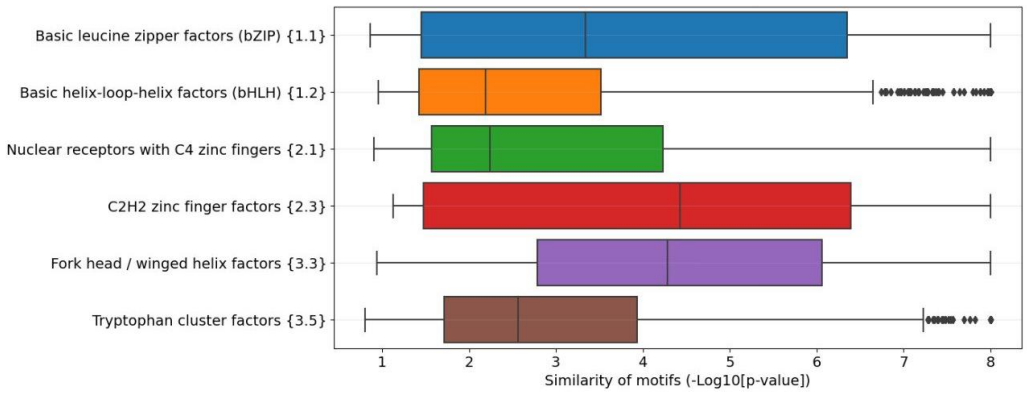

Fig. 1. Intra-class variability of motifs similarity. X-axis denotes classes. Y-axis indicates the significance of motifs similarity, –Log10(p-value). The boxplots depict the distributions of the Q1, Q2 and Q3 quartiles of the fractions of pairs of datasets with certain values of the significance. Whiskers on either side of the Q1/Q3 respect the minimum/maximum values if they were located within 1.5 interquartile ranges (IQR=Q3–Q1) from Q1/Q3, otherwise they are equal to {Q1−1.5*IQR}/{Q3+1.5*IQR}, respectively. In the latter case, we marked all other points as outliers

Results: We applied STREME and TomTom tools to define the enriched motifs of target TFs for all ChIP-seq datasets as described above. Figure 1 shows the distributions of motifs similarity for pairs of enriched motifs within each of six classes. The classes bHLH {1.2} and Nuclear receptors {2.1} known for TF dimerization [7] show the lowest significances (p > 2E-3), while C2H2 zinc finger factors {2.3} and Fork head / winged helix factors {3.3} classes lacking this functionality of TFs show the highest significances (p < 1E-4). Figure 2 compares for six classes the distributions of the median significance of motifs similarity within its families with the cross-family ones. Some pairs of families, e.g. {1.1.1} vs. {1.1.2}, {1.2.1} vs. {1.2.2}, show very high similarity, while the majority of the rest families are not quite similar to other families of the same class, e.g. {1.1.8}, {1.2.6}, and {3.5.1}. We confirmed the structural heterogeneity of TFs from C2H2 zinc finger factors {2.3} class. Only some families due to structural variation of motifs are heterogeneous, e.g. the diagonal cell for the family {1.1.2}.

Conclusion: We propose to indicate for enriched motifs in the results of *de novo* motif search the appropriate level of the hierarchical classification of TFs by DBD structure. These levels should be preliminary defined so that within each of them the motifs are significantly similar.

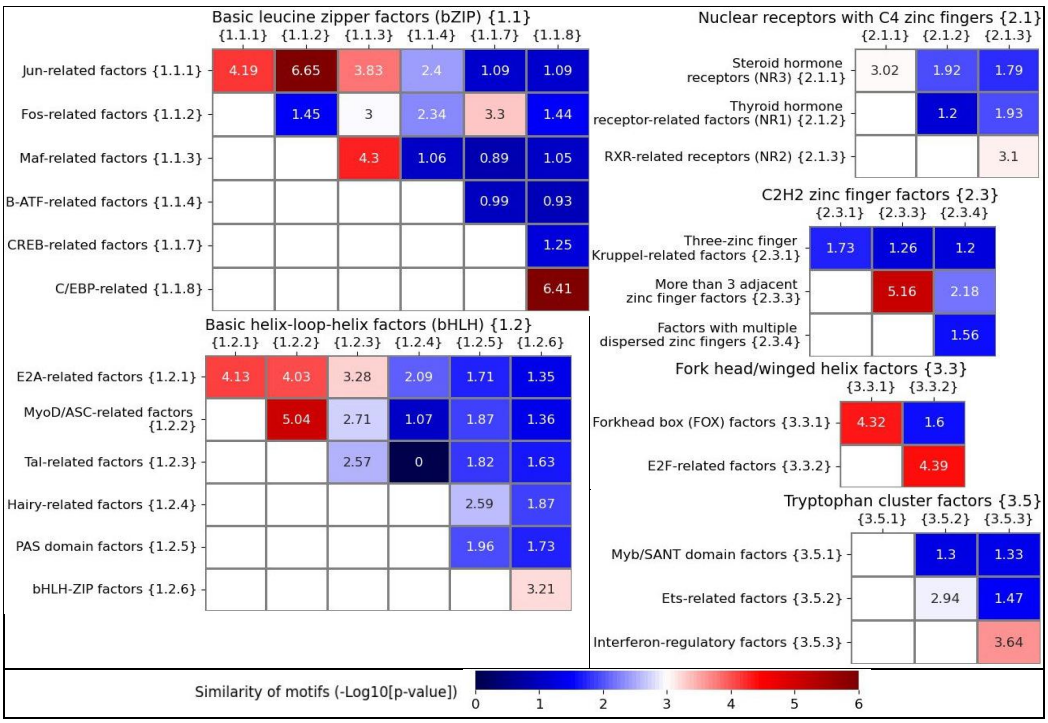

Fig. 2. Variation of the motif similarity within families and between different families for the six classes. Each heatmap entitled with the name of class. Colors/numbers in cells on the diagonals and above them show the median significance for the intra- and inter-family distributions of motifs similarity, –Log10(p-value). Rows and columns show families within the same classes. Empty diagonal cells ({1.1.4}, {1.1.7}, {1.2.4} and {3.5.1}) imply that these families contain sole TFs

Funding: The study is supported Russian Science Foundation project No. 20-14-00140.

References

1. Bailey T.L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*. 2021;37:2834-2840. doi 10.1093/bioinformatics/btab203
2. Gupta S. et al. Quantifying similarity between motifs. *Genome Biol*. 2007;8(2):R24. doi 10.1186/gb-2007-8-2-r24
3. Rauluseviciute I. et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2024;52(D1):D174-D182. doi 10.1093/nar/gkad1059
4. Vorontsov I.E. 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
5. Wingender E. et al. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.* 2018;46:D343-D347. doi 10.1093/nar/gkx987
6. Ambrosini G. et al. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol.* 2020;21:114. doi 10.1186/s13059-020-01996-3
7. Amoutzias G.D. et al. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci.* 2008;33(5):220-229. doi 10.1016/j.tibs.2008.02.002
8. Kolmykov S. et al. GTRD: An integrated view of transcription regulation. *Nucleic Acids Res.* 2021;49(D1):D104-D111. doi 10.1093/nar/gkaa1057