doi 10.18699/bgrs2024-12.3-24

Analysis of the relationships of significant contextual signals in ChIP-Seq peaks

Vishnevsky O.[1, 2, 3]

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

[2] *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

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

* *oleg@bionet.nsc.ru*

Key words: motif discovery; composite elements; transcription regulation; ChIP-Seq

Motivation and Aim: The development of ChIP-seq approaches has led to the accumulation of a huge body of information about the *in vivo* genome-wide localization of thousands of transcription factors in various cellular conditions on DNA. ChIP-seq peak sequences are characterized by a high-level complexity of contextual organization and may either contain several potential binding sites for target TFs or not contain them at all. Additionally, they may contain quite a few potential binding sites for partner TFs. A wealth of experimental data provide evidence that transcription factors regulate transcription in close cooperation with each other. Thus, taking into account pairwise and groupwise interactions between transcription factors is necessary for a deeper understanding of gene regulation and for scoring the observed ChIP-seq peaks.

Methods and Algorithms: In the present work, we analyzed the contextual organization of ChIP-seq peak sequences in experiments with 10 TFs from CistromeDB [1] belonging to six main types of DNA-binding domains [2]. Using the original *de novo* GPU-based motif discovery method Argo_CUDA [3], we revealed, in the peak sequences from each ChIP-seq experiment, both sets of significant IUPAC motifs corresponding to the target TF binding sites studied in each experiment and specific sets of motifs corresponding to the binding sites for the partner TFs. Unlike heuristic methods, Argo_CUDA evaluates the significance of all possible IUPAC motifs of a given length, which guarantees finding a global optimum. The annotation of the identified motifs was carried out using the Tomtom [4] software package based on the Hocomoco [5] and Jaspar [6] positional weight matrix databases.

Results: A significant correlation was shown between the presence of motifs corresponding to the binding sites for the target transcription factors and the partner motifs corresponding to the binding sites for the partner transcription factors, which was experimentally confirmed by the scientific literature. For all the ChIP-seq experiments considered, multiple regression models were constructed, demonstrating a significant dependence of the ChIP-seq peak sequence scores on the presence of sets of specific IUPAC motifs in these sequences. It has been shown that the most significant target motifs make a substantial contribution to the observed dependence. At the same time, the prediction quality can be improved through the use of less significant motifs as well as partner motifs.

Conclusion: The contextual features of ChIP-seq peaks that we have identified can be used to set up experiments aimed at testing potential partner interactions of TFs, the motifs of which are reliably co-represented in the sequences of ChIP-Seq peaks and also help in building potential regulatory gene networks involved in subtle developmental

processes and tissue-specific gene expression. In addition, the significant IUPAC motifs we identified can be used to develop new methods for predicting the localization of potential TFBSs in genomic sequences.

Funding: The study is supported by Russian government project (No. FWNR-2022-0020).

References

1. Zheng R., Wan C., Mei S., Qin Q., Wu Q., Sun H., Chen C.H., Brown M., Zhang X., Meyer C.A., Liu X.S. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* 2019;47:D729-D735. doi 10.1093/nar/gky1094
2. Wingender E., Schoeps T., Haubrock M., Krull M., Dönitz J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic acids res.* 2018;46:D343-D347. doi 10.1093/nar/gkx987
3. Vishnevsky O.V., Bocharnikov A.V., Kolchanov N.A. Argo_CUDA: Exhaustive GPU based approach for motif discovery in large DNA datasets. *J Bioinform Comput Biol.* 2018;16:1740012. doi 10.1142/S0219720017400121
4. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., Noble, W. S. Quantifying similarity between motifs. *Genome Boil.* 2007;8:R24. doi 10.1186/gb-2007-8-2-r24
5. Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Sharipov R.N., Fedorova A.D., Rumynskiy E.I., Medvedeva Y.A., Magana-Mora A., Bajic V.B., Papatsenko D.A., Kolpakov F.A., Makeev V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 2018;46:D252-D259. doi 10.1093/nar/gkx1106
6. Castro-Mondragon J.A., Riudavets-Puig R., Rauluseviciute I., Lemma R.B., Turchi L., Blanc-Mathieu R., Lucas J., Boddie P., Khan A., Manosalva Pérez N., Fornes O., Leung T.Y., Aguirre A., Hammal F., Schmelter D., Baranasic D., Ballester B., Sandelin A., Lenhard B., Vandepoele K., … Mathelier A. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2022;50:D165-D173. doi 10.1093/nar/gkab1113