DOI 10.18699/BGRSSB-2018-035

The overlapped motifs co-occurrence in ChIP-seq data

V.G. Levitsky[1, 2]*, D.Y. Oshchepkov[1], E.V. Zemlyanskaya[1, 2], V.V. Mironova[1, 2], E.V. Ignatieva[1, 2], O.A. Podkolodnaya[1], T.I. Merkulova[1, 2]
[1] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
** e-mail: levitsky@bionet.nsc.ru*

Key words: composite element, chromatin immunoprecipitation, transcription factor binding site prediction

Motivation and Aim: The cooperative binding of transcription factors (TFs) is the common mechanism of their functioning [1]. Recently developed collections of whole-genome datasets for ChIP-seq peaks [2] and derived motifs for TF binding models [3] require development of adequate tools for prediction of potential composite elements (CEs) consisting of anchor/partner motifs separated by relatively short spacer (not more than some tens base pairs). Existing bioinformatics approaches for prediction of motif co-occurrence in ChIP-seq datasets can't treat the motif overlapping, i.e. only motifs separated by a spacer of zero/positive length were considered as a potential CE (e.g. [4]).
Methods and Algorithms: We propose a new algorithm that can infer motif co-occurrence in ChIP-seq data without limitation for overlap. First, we compute the frequencies of anchor/partner motifs in a peak. Second, for each motif hit we count the number of overlapped motifs of the same type. These two measures help to generate the permuted sequences for a peak. We apply the Fisher's exact test to estimate the enrichment of the CE content in peaks in comparison with that in permuted data. Additionally, we use the Tomtom tool [5] to filter out possible overpredictions related to a significant match between the anchor and partner motifs.
Results: We analyzed more than hundred ChIP-seq datasets for about fifty TF types of mammals and plants and found that the majority (~95 %) of the overrepresented co-occurred motif pairs are overlapped. Our results are in a good accordance with earlier analysis of motif co-occurrence in specific cell lines [6] and the applicaion of *in vitro* SELEX modelling for cooperative TF binding [7].
Conclusion: We found that motifs overlap is widespread in ChIP-seq data. The application of our novel tool will substantially contribute to their careful annotation.
Acknowledgements: The work was supported by RFBR 18-04-01130 and ICG SB RAS budget project 0324-2018-0017.

References

1. Morgunova E., Taipale J. (2017) Structural perspective of cooperative transcription factor binding. Curr. Opin. Struct. Biol. 47:1-8.
2. Kulakovskiy I.V. et al. (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 46(D1):D252-D259.
3. Yevshin I. et al. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 45(D1):D61-D67.
4. Whitington T. et al. (2011) Inferring transcription factor complexes from ChIP-seq data. Nucleic Acids Res. 39:e98.
5. Gupta S. et al. (2007) Quantifying similarity between motifs. Genome Biology. 8(2):R24.
6. Jankowski A. et al. (2013) Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. Genome Res. 23(8):1307-1318.
7. Jolma A. et al. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 527(7578):384-388.