DOI 10.18699/SBB-2022-046

Heterogeneity of transcription factor binding sites within ChIP-Seq datasets

Sharipov R.[1, 2, 3]*, Kondrakhin Y.[2, 4], Kolmykov S.[1, 2], Yevshin I.[1, 2], Ryabova A.[1, 2], Kolpakov F.[1, 2, 4]

[1] *Sirius University of Science and Technology, Sochi, Russia*

[2] *BIOSOFT.RU, LLC, Novosibirsk, Russia*

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

[4] *Federal Research Center for Information and Computational Technologies, Novosibirsk, Russia*

** shrus79@gmail.com*

Key words: transcription factor binding sites, ChIP-Seq, heterogeneity of binding sites

Motivation and Aim: Despite wide and useful application of the powerful ChIP-Seq technology for experimental identification of transcription factor (TF) binding sites, the computational prediction of site motifs is also relevant. There are many distinct methods and approaches to predict site motifs, however position weight matrix (PWM) approach remains to be the most common and widely used. Several well-known repositories contain PWMs for many distinct TFs and species, in particular HOCOMOCO, JASPAR and UniPROBE databases are among them.

In general, the main disadvantages of PWM methods are the ignorance of possible dependencies between nucleotides at different site positions as well as ignorance of information about nucleotide content at both flanks of site cores. To avoid these disadvantages, we proposed novel method for site prediction.

Methods and Algorithms: The novel method works as follows:

- at first, it assigns individual probability scores (IPSs) to each motif in order to take into account nucleotide content at both site flanks;
- at second, several PWMs are simultaneously used for prediction of binding sites for a single TF. In other words, the existence of such matrices indicates heterogeneity of total set of binding sites. Thus, the matrices represent distinct homogeneous subsets of total heterogeneous set of binding sites. These matrices were derived beforehand by application of the special IPS-alignment method to the training sample of most reliable binding regions.

The key point of alignment method is successive application of two-component normal mixture method to training sequence sample. The training sample, in turn, was composed by meta-analysis of ChIP-Seq datasets available in a GTRD database. It consists of sequences that are the most reliable binding regions. The reliability of training sequences was measured using special scores obtained by three-steps meta-method based on Rank Aggregation approach.

Results: Currently, using the developed method, we have composed three libraries of TF matrices for *Drosophila melanogaster*, *Danio rerio* and *Arabidopsis_thaliana*. The corresponding site motifs are available in the GTRD database (http://gtrd.biouml.org/).

Acknowledgements: The study was supported by the Ministry of Science and Higher Education of the Russian Federation, grant 075-15-2021-1344.