From ChIP-Seq data to improved transcription factor binding sites models

Ivan V. Kulakovskiy[1,2], Victor G. Levitsky[3,4], Dmitry G. Oshchepkov[3], Ilya E. Vorontsov[2,5], Vsevolod J. Makeev[1,2,6,7]

[1]*Laboratory of Bioinformatics and Systems Biology, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov str. 32, Moscow, 119991, GSP-1, Russia*

[2]*Department of Computational Systems Biology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina str. 3, Moscow, 119991, Russia*

[3]*Laboratory of Molecular Genetics Systems, Institute of Cytology and Genetics, Siberian Division of Russian Academy of Sciences, Lavrentiev Prospect 6, Novosibirsk, 630090, Russia*

[4]*Faculty of Natural Sciences, Novosibirsk State University, Pirogova str. 2, Novosibirsk, 630090, Russia*

[5]*Yandex Data Analysis School, Data Analysis Department, Moscow Institute of Physics and Technology, Leo Tolstoy Str. 16, Moscow, 119021, Russia*

[6]*State Research Institute of Genetics and Selection of Industrial Microorganisms, 1st Dorozhny proezd, 1, Moscow, 117545, Russia*

[7]*Moscow Institute of Physics and Technology, Institutskii per. 9, Dolgoprudny, 141700, Moscow Region, Russia*

ivan.kulakovskiy@gmail.com

Sequence motif analysis is one of the base components of transcriptional regulation studies in higher eukaryotes. In particular, motif finding methods are utilized to predict putative transcription factor binding sites (TFBS) in genomic regions. This requires a TFBS model, which is often represented as a positional weight matrix (PWM) based on a gapless multiple local alignment of experimentally identified TFBS sequences. Existing tools for TFBS prediction still mainly rely on PWMs based on nucleotide positional frequencies. Our PWM-based algorithm ChIPMunk [1] was able to successfully compete with other tools in several independent benchmarks including recent study by DREAM consortium [2].

Modern high-throughput methods, including chromatin immunoprecipitation followed by deep sequencing, ChIP-Seq, provide large amounts of data which can be utilized for more advanced models accounting for positional dependencies in TFBS. We present our new tool, diChIPMunk [3], http://autosome.ru/dichipmunk/, that can produce dinucleotide PWMs taking into account dependencies between neighboring nucleotides in TFBS.

Using several public ChIP-Seq datasets we show that dinucleotide PWMs produced by

diChIPMunk clearly outperform existing published PWMs and novel PWMs constructed by ChIPMunk from the same data.

This work was supported by a Dynasty Foundation Fellowship [to I.V.K.]; Russian Foundation for Basic Research [12-04-32082 to I.V.K.] and [12-04-01736-a to D.G.O.].

1. I.V. Kulakovskiy *et al.* (2010) Deep and wide digging for binding motifs in ChIP-Seq data, *Bioinformatics*, **26**(20):2622-3.

2. M.T. Weirauch et al. (2013) Evaluation of methods for modeling transcription factor sequence specificity, *Nat Biotechnol*, **31**(2):126-34.

3. I. Kulakovskiy *et al.* (2013) From binding motifs in chip-seq data to improved models of transcription factor binding sites, *J Bioinform Comput Biol*, **11**(1):1340004.