diChIPMunk: utilizing ChIP-Seq data to construct advanced dinucleotide models of transcription factor binding sites

I. KULAKOVSKIY [1,2,*], V. LEVITSKY [3,4], D. OSCHEPKOV [3], I. VORONTSOV [2,5], V. MAKEEV [1,2,6]

(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, GSP-1, Russia

(3) Laboratory of Molecular Genetics Systems, Institute of Cytology and Genetics of the 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) Faculty of Molecular Biology, Moscow Institute of Physics and Technology, Institutskii per. 9, Dolgoprudny, 141700, Moscow Region, Russia

* ivan.kulakovskiy@gmail.com

Abstract

Computational analysis and prediction of transcription factor binding sites (TFBS) is one of the fundamental tasks in regulatory genomics. A TFBS model can be derived from a set of experimentally determined DNA sequences, specifically recognized by a transcription factor (TF). A typical approach is to apply computational *de novo* motif discovery tools. With ChIP-Seq as the new gold standard for genome-wide detection of TFBS *in vivo* it becomes possible to construct advanced TFBS models. Here we present a special motif discovery tool, diChIPMunk, which can produce dinucleotide positional weight matrices (diPWMs) from ChIP-Seq data. We show that diPWMs produced by diChIPMunk significantly outperform existing classic mononucleotide matrices in terms of TFBS recognition quality.

The software is freely available: http://autosome.ru/dichipmunk/

Introduction

Transcription regulation in higher eukaryotes involves transcription factors (TFs) specifically recognizing binding sites (TFBS) in DNA. Experimental techniques based on chromatin immunoprecipitation produce thousands of DNA segments putatively recognized by a TF. One of typical aims is to detect a common text pattern representing preferred TFBS. Careful representation of this pattern, the TFBS model, allows computational prediction of TFBS in genomic sequences of interest.

The most widely used TFBS model is a positional weight matrix (PWM) directly computed from a gapless multiple local alignment of TFBS-containing sequences. PWM assumes independent nucleotide frequencies in different alignment columns, as there were no correlations between them.

At the same time, some more complex models based on ChIP-Seq data provided only incremental improvement over properly trained traditional PWMs [Bi2011].

A matrix of positional weights based on dinucleotide frequencies takes into account correlations of nucleotides in neighboring alignment positions and provides simple extension of the PWM model. Earlier it was already demonstrated that dinucleotide PWMs could outperform classic mononucleotide PWMs if learned from on a reasonably large set of sequences [Levitsky2007]. The remaining step is to properly utilize ChIP-Seq data for model training.

Here we present diChIPMunk, a tool able to produce dinucleotide PWMs based on ChIP-Seq data.

Results

Earlier we presented ChIPMunk [Kulakovskiy2010], an effective algorithm for construction of traditional PWM models based on ChIP-Seq data. ChIPMunk performed efficiently and accurately in several independent benchmarking studies including a recent one of the DREAM consortium [Weirauch2013]. diChIPMunk is based on the same computational engine as ChIPMunk, and thus shares several advantages including usage of ChIP-Seq peak shape (the reads pileup profile) and a support for multi-threaded computations. To utilize ChIPMunk engine diChIPMunk uses a "superalphabet" approach converting initial DNA sequences written in a mononucleotide A-C-G-T alphabet into dinucleotide sequences with a AA-AC-AT…TT alphabet (with each nucleotide included in two neighboring dinucleotides).

To test TFBS recognition quality we have used different ChIP-Seq datasets to compare diChIPMunk models with those of ChIPMunk and PWMs available from public sources.

Here, as a case study, we used top 1000 ChIP-Seq peaks of NANOG and SOX2 TFs published in [Chen2008]. Even ranked peaks were used for model training; odd ranked peaks were used as control true positive sequences. Using a strategy from [Kulakovskiy2013] we have plotted ROC-curves and calculated area-under-curve (AUC) values. Figure 1 presents results of the comparison.

Several other examples of diChIPMunk models evaluation were presented in the corresponding paper [Kulakovskiy2013].

Conclusions

diChIPMunk is able to produce dinucleotide PWMs that perform significantly better than mononucleotide PWMs. We provide diChIPMunk as a production-ready tool. diChIPMunk is going to be included in BioUML platform [Kolpakov2006] as a motif discovery algorithm along with several accompanying tools. As the dinucleotide PWM is a fairly simple model it becomes possible to adapt many existing supporting tools, such as TFBS prediction in a given sequence (i.e. motif finding), computing P-values for given score threshold levels etc for dinucleotide PWMs. We believe this will facilitate a wider usage of dinucleotide PWMs with more and moreChIP-Seq data becoming available.

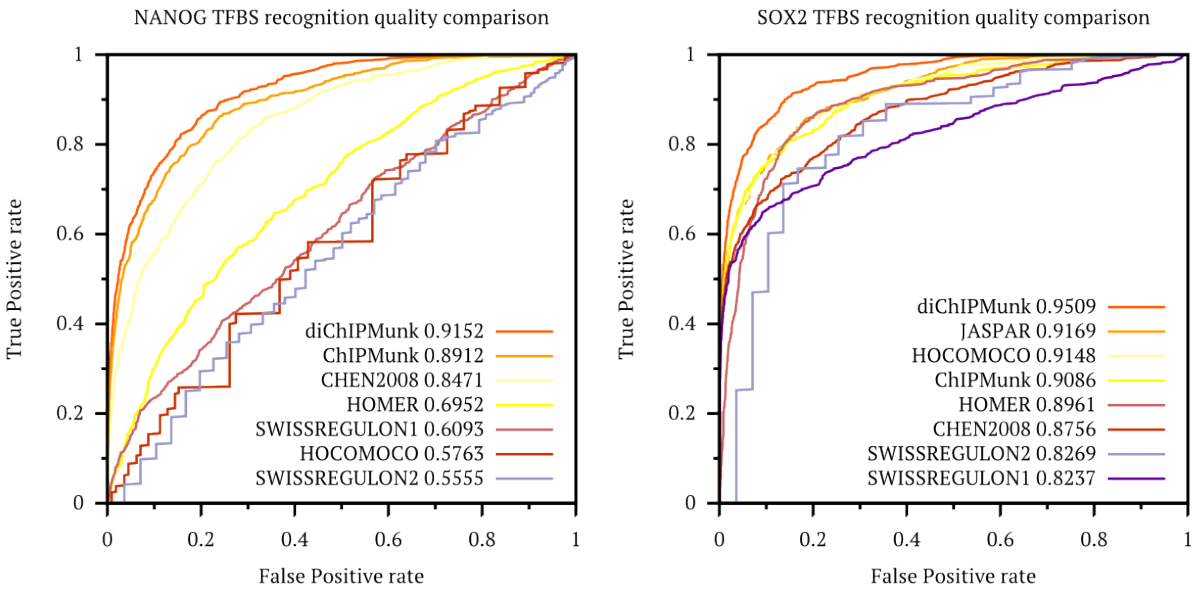

Figure 1. ROC curves of TFBS models for NANOG (left panel) and SOX2 (right panel) TFs. True positive rate was estimated using independent control subset of ChIP-Seq peaks. False positive rate was estimated based on PWM/dinucleotide PWM P-values as described in [Kulakovskiy2013]. AUC values are given in figure legends. HOMER, SwissRegulon and JASPAR PWMs were taken from corresponding collections. CHEN2008 PWM was presented in the same paper as the TF ChIP-Seq data. The SOX2 matrix from JASPAR collection was based on the same ChIP-Seq dataset.

Acknowledgements

This work was supported by a Dynasty Foundation Fellowship [to I.K.]; Russian Foundation for Basic Research [12-04-32082-mol_a to I.K.] and [12-04-01736-a to D.O.]; Presidium of the Russian Academy of Sciences program in Cellular and Molecular Biology.

References

[Bi2011] PLoS One. 2011;6(9):e24210. doi: 10.1371/journal.pone.0024210. Epub 2011 Sep 2. Tree-based position weight matrix approach to model transcription factor binding site profiles. Bi Y, Kim H, Gupta R, Davuluri RV.

[Chen2008] Cell. 2008 Jun 13;133(6):1106-17. doi: 10.1016/j.cell.2008.04.043. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH.

[Kolpakov2006] Proceedings of The Fifth International Conference on Bioinformatics of Genome Regulation and Structure; July 16–22, 2006; Novosibirsk, Russia. 2006.3 p. 281-285. BioUML: visual modeling, automated code generation and simulation of biological systems. Kolpakov F, Puzanov M, Koshukov A.

[Kulakovskiy2010] Bioinformatics. 2010 Oct 15;26(20):2622-3. doi: 10.1093/bioinformatics/btq488. Epub 2010 Aug 24. Deep and wide digging for binding motifs in ChIP-Seq data. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ.

[Kulakovskiy2013] J Bioinform Comput Biol. 2013 Feb;11(1):1340004. doi: 10.1142/S0219720013400040. Epub 2013 Jan 16. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. Kulakovskiy I, Levitsky V, Oshchepkov D, Bryzgalov L, Vorontsov I, Makeev V.

[Levitsky2007] BMC Bioinformatics. 2007 Dec 19;8:481. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. Levitsky VG, Ignatieva EV, Ananko EA, Turnaev II, Merkulova TI, Kolchanov NA, Hodgman TC.

[Weirauch2013] Nat Biotechnol. 2013 Feb;31(2):126-34. doi: 10.1038/nbt.2486. Epub 2013 Jan 27. Evaluation of methods for modeling transcription factor sequence specificity. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S; DREAM5 Consortium, Bussemaker HJ, Morris QD, Bulyk ML, Stolovitzky G, Hughes TR.