

Motif discovery and benchmarking from diverse assays highlight the best tools for predicting DNA-binding specificity of human transcription factors

Vorontsov I.E.¹, Kozin I.^{2,3}, Abramov S.⁴, Boytsov A.⁴, Jolma A.⁵, Albu M.⁵, Ambrosini G.^{6,7}, Faltejskova K.⁸, Gralak A.^{6,7}, Gryzunov N.⁹, Inukai S.¹⁰, Kolmykov S.¹¹, Kravchenko P.¹², Kribelbauer J.^{6,7}, Lavery K.⁵, Nozdrin V.⁹, Patel Z.⁵, Penzar D.¹, Plescher M.-L.¹³, Pour S.⁵, Razavi R.⁵, Yang A.⁵, Yevshin I.¹¹, Zinkevich A.⁹, Bucher P.⁷, Deplancke B.^{6,7}, Fornes O.¹⁴, Grau J.¹³, Grosse I.¹³, Kolpakov F.¹¹, GRECO-BIT/Codebook Consortium, Makeev V.J.¹, Hughes T.⁵, Kulakovskiy I.V.^{1,3*}

¹ Vavilov Institute of General Genetics, RAS, Moscow, Russia

² Faculty of Biotechnology, Lomonosov Moscow State University, Moscow, Russia

³ Institute of Protein Research, RAS, Pushchino, Russia

⁴ Altius Institute for Biomedical Sciences, Seattle, WA, USA

⁵ Donnelly Centre and Department of Molecular Genetics, Toronto, Canada

⁶ School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

⁷ Swiss Institute of Bioinformatics, Lausanne, Switzerland

⁸ Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Praha, Czech Republic; Computer Science Institute, Faculty of Mathematics and Physics, Charles University, Praha, Czech Republic

⁹ Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

¹⁰ Chugai Pharmaceutical Co., Ltd, Tokyo, Japan

¹¹ Department of Computational Biology, Sirius University of Science and Technology, Sirius, Krasnodar region, Russia

¹² Max Planck Institute of Biochemistry, Planegg, Germany

¹³ Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany

¹⁴ Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC, Canada

* ivan.kulakovskiy@gmail.com

Key words: TF; TFBS; DNA motif discovery; benchmarking; transcription factors

Motivation and Aim: Computational methods that discover DNA motifs specifically binding proteins have been developed for over three decades, stimulated by the continuing progress in experimental methods for DNA-protein interaction profiling. Yet, the assessment of how well the computational models quantitatively reflect the binding specificities is still far from routine. Particularly, despite the availability of many advanced motif models, the position weight matrix, the traditional model of DNA binding specificity, remains the most commonly used approach, and it is still unclear if the more complex model outperforms positional weight matrices, and if so, to which degree.

Methods: As GRECO-BIT/Codebook initiative, we used the newly generated experimental data from several thousand experiments for several hundred less-studied

human transcription factors to derive binding motifs with diverse DNA motif discovery tools, from multiple sequence aligners to deep learning methods. First, we performed an extensive benchmarking study highlighting the best-performing tools and forming a rich catalog of motifs, many of which belong to previously unexplored transcription factors. Next, we used a carefully selected subset of forty TFs to call for the wisdom of crowds: we have organized an open competition in constructing the best sequence-level models of transcription factor binding sites from these yet unpublished data. The challenge in Inferring Binding Specificities, IBIS, is built on the experimental data on the TF-DNA interactions obtained from five different experimental techniques, and aimed at a fair assessment of existing and novel methods for de novo TFBS motif discovery, using both classic Position Weight Matrices and Arbitrary Advanced Approaches, “triple-A” models.

Results: We plan to report the best-performing software tools and key findings of the internal within-Consortium motif benchmarking study and tease the first insights from the open IBIS Challenge, which will end mid-summer 2024.

Funding: The study is supported by RSF 20-74-10075 to IVK.