DOI 10.18699/WIBSB-2018-18

A Kolmogorov – Smirnov based approach for predicting targets of transcription

M.J. Djordjevic[1]*, M.R. Djordjevic[2], E.Zdobnov[3]
[1] *Faculty of Biology, Institute of Physiology and Biochemistry, University of Belgrade, Belgrade, Serbia*
[2] *Institute of Physics Belgrade, University of Belgrade, Belgrade, Serbia*
[3] *Swiss Institute of Bioinformatics and Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland*
** e-mail: dmarko@bio.bg.ac.rs*

Key words: regulatory element detection, Kolmogorov–Smirnov, transcription regulation, ChIP-seq

Motivation and Aim: Accurately predicting direct targets of transcriptional regulators is necessary to understand gene expression regulation. Predicting these targets typically leads to a large number of false positives, and usually corresponds to searching for high-scoring binding sites in the upstream genomic regions. In contrast to the common approach, we here propose a novel concept, where overrepresentation of the scoring distribution that corresponds to the entire searched region is assessed, as opposed to predicting individual binding sites.
Methods: As opposed to predicting individual binding sites, we here propose a novel concept, where the entire searched region is scored, and overrepresentation of this scoring distribution is assessed [1]. We implement this concept through both Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests, where both allow straightforwardly predicting *P*-values for each target.
Results: We first apply this approach to pleiotropic bacterial regulators, including σ^{70} (housekeeping bacterial σ factor) whose target prediction is a classical bioinformatics problem characterized by high number of false positives. We show that KS based approach is both more accurate and faster compared to AD, departing from the current paradigm of AD being more accurate (though slower). Moreover, KS has a significantly higher accuracy compared to the standard approach, while straightforwardly assigning well established *P*-values to potential targets. Secondly, we apply the method to ChIP-seq data analysis, to test how it can predict bacterial transcription targets *in vivo*. While we find a good correspondence between computational predictions and *in vitro* binding data, both of them correlate significantly worse with *in vivo* data [2].
Conclusion: New KS based method proposed here, which assigns P-values to fixed length upstream regions, provides a fast and accurate approach for predicting bacterial transcription targets. Binding of transcription factors in vivo may be significantly influenced by factors other than binding energy, even in bacteria where this binding does not happen in the chromatin context.
Acknowledgement: This work is supported by the Swiss National Science Foundation under SCOPES project number IZ73Z0_152297 and by the Ministry of Education and Science of the Republic of Serbia under Project number ON173052.

References

1. Djordjevic M.J., Djordjevic M.R., Zdobnov E. (2017) Scoring targets of transcription in bacteria rather than focusing on individual binding sites. Front. Microbiol. 8:2314.
2. Djordjevic M.J., Djordjevic M.R. (2018) Regulation by bacterial transcription factors: relationship between binding energy and binding site functionality. (in preparation).