DOI 10.18699/BGRSSB-2018-067

Prediction of gene expression level by using ChIP-Seq-derived data from the GTRD database and transcription start sites identified by the Fantom5 project

I. Yevshin[1], Yu. Kondrakhin[1, 2], R. Sharipov[1, 3]*, F. Kolpakov[1, 2]
[1] *BIOSOFT.RU, Ltd, Novosibirsk, Russia*
[2] *Institute of Computational Technologies SB RAS, Novosibirsk, Russia*
[3] *Novosibirsk State University, Novosibirsk, Russia*
** e-mail: shrus79@gmail.com*

Key words: ChIP-Seq, transcription factors, binding regions, GTRD database, transcription start sites, gene expression levels

Motivation and Aim: Since the beginning of the current millennium, ChIP-Seq has become the most powerful experimental technique for the genome-wide study of interactions between transcription factors (TFs) and DNA. For our analysis, we have used 4232 datasets of human transcription factor binding regions (TFBRs) identified by ChIP-Seq and collected in a GTRD database [1]. These datasets represented TFBRs of 694 distinct TFs. We also have exploited the expression profiles measured in 1829 main human primary cell types and tissues [2]. For each cell type or tissue, these expression levels were measured for collection of 209911 transcription start sites (TSSs) identified in the frames of the FANTOM5 project [2].

Methods and Algorithms: For each considered cell line or tissue, we developed a prediction model that consisted of classification and regression submodels. The goal of the classification submodel was to discriminate between expressed and non-expressed genes while the aim of the regression submodel was to predict the real-valued expression levels. For construction of the classification submodel, we used such machine learning approaches as random forest, support vector machine, perceptron, and Fisher's discriminant analysis. To construct regression submodel we used random forest, support vector machine, least squares regression, and regression on principal components. The features of regression submodels were defined as presence/absence of TFBRs in promoter regions (–5000, –1000), (–1000, –500), (–500, –200), (–200, –100), (–100, –1), (–1, +1), (+1, +100), and (+100, +500).

Results: We constructed the set of accurate models for prediction of expression levels in distinct cell lines and tissues. In particular, in case of HepG2 cell line, the accuracy of the prediction model was estimated by high correlation (0.866) between observed and predicted expression levels as well as by high value of the explained variance (74.9 %). In this case, the most significant feature for prediction was the occurrence of TFBRs within the following promoter regions: HNF-4a in (–1, +1), TAF-1 in (+100, +500), TAF-1 in (–100, –1), TAF-1 in (+1, +100), HNF-4a in (+100, +500), ZNF274 in (–5000, –1000), and TR4 in (–1, +1).

References

1. Yevshin I., Sharipov R., Valeev T. et al. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 45:D61-D67.
2. The FANTOM Consortium and The RIKEN PMI and CLST. (2014) A promoter-level mammalian expression atlas. Nature. 507:462-471.