Identification of bacteria transcription factor targets by machine learning

I.A. Garanina, D.V. Evsutina

Scientific Research Institute of Physical-Chemical Medicine SRI PCM, Moscow, Russia

G.U. Fisunov

irinagaranina24@gmail.com

Reconstruction of transcription regulation networks for bacteria is the important problem that still have not resolved. Current methods rely on information about known transcription factors of coregulation networks. In this work we aimed to develop a new method for prediction of regulated genes by using machine learning methods.

We used random forest algorithm of machine learning to predict strength of bacterial promoters by sequence. As experimental data for learning we used published and obtained by our laboratory data on coverage and position of transcription start sites. We used these features of promoter sequence to build model: sequences of -10 box, extension of -10 box, -35 box, nucleotides on transcription start site, length of spacers between -10 and -35 boxes, GC content of spacers between -10 and -35 boxes and between -10 box and transcription start site, up element upstream of -35 box, sequence upstream and downstream -10 box.

We propose new algorithm that calculates probability of gene to be regulated. Our model predicts the bacterial promoters` power on the basis of its sequence. We applied this model for three bacteria with different genome size and number of regulators in genome. Comparing promoter power (theoretical prediction) with the data on promoters' activity (experimental data) we predicted promoters which activity is deviant from their power. We confirmed our approach on well-studied bacteria with big number of known regulators and then applied it to genome reduced bacteria Mycoplasma gallisepticum with a few number of known transcription factors. About 60 promoters of Mycoplasma gallisepticum are affected by repressors, which is 10 times higher than number of transcription factors identified by conservation. These repressed promoters don't have any sequences near promoters similar to known for Mycoplasma transcription factor binding sites. So, probably, these repressed genes represent a hidden layer of regulation that work only in specific conditions or we faced with new type of regulation in bacteria.

New method for prediction of repressed genes in bacterial genomes was developed and tested on three bacteria species. We identified many new potential targets for regulation of Mycoplasma gallisepticum. Our study thus provides insight into transcriptional regulation of bacteria.

This work was funded by the Russian Science Foundation grant 14-24-00159 “Systems research of minimal cell on a Mycoplasma gallisepticum model”