**Promoter library analysis and sequence prediction with Jupyter-notebook assisted random forest regressor**

Ulf W. Liebal1†, Sebastian Köbbing1, Lars M. Blank1

1 Institute of Applied Microbiology-iAMB, Aachen Biology and Biotechnology-ABBt, RWTH Aachen University, Worringerweg 1, 52074 Aachen, Germany

KEYWORDS: Biotechnology, Synthetic Biology, Promoters, Pseudomonas, Machine Learning, Random Forest, Python

**ABSTRACT**

In metabolic engineering heterologous enzymes are expressed to introduce new metabolic pathways to micro-organisms. Tight control of gene expression is beneficial to fine-tune enzyme activity, possibly across multiple hosts to either identify the optimal host or to separate cloning and production host. Among the main determinants of heterologous gene expression is the promoter region. However, detailed knowledge of the translation of a promoter sequence to an expression strength is missing. Here we present a Jupyter notebook workflow that uses a promoter library from *P. putida* and trains a random forest regressor. The random forest regressor allows for the activity prediction of thousands of novel promoter sequences that cover a broad range of expression profiles while minimizing genetic adjustments to a reference sequence. The associated statistical analyses identify the sequence exploration space of the promoter library within which reasonable predictions are possible. The random forest regression extracts the position-specific nucleotide-base importance towards the activity prediction and reproduces the outstanding effects of -35 and -10 promoter domains. The workflow identifies novel promoter sequences with defined activities thereby enlarging the promoter toolbox for controlled gene expression. Moreover, the notebook is flexible to also analyze multiple promoter libraries across multiple species to predict defined cross-host activity. It can be adapted to investigate any sequence library to identify sequence factors that determine desired quantitative outcomes.

**INTRODUCTION**

**Expression activity prediction from sequence**

ML prediction of quantitative phenotypes from nucleic acid sequences 1

Scoring of sequence to activity with position weight matrix (PWM) 2

sigmaE in E. coli with different ML-approaches 3,4

**Library construction developments**

Standardized expression elements 5

244,000 sequences for translation optimization 6

Effect of 5’ untranslated region on expression with 500,000 sequences 7

**Jupyter notebook development**

Guidelines for Jupyter notebooks 8

Existing approaches with MFA in JBEI 9,10

Good coding practices for data analysis scripts 11

**REFERENCES**

1 Sauer, D. & Wang, D.-N. Using machine learning to predict quantitative phenotypes from protein and nucleic acid sequences. *bioRxiv* (2019).

2 Rhodius, V. A., Mutalik, V. K. & Gross, C. A. Predicting the strength of UP-elements and full-length E. coli σE promoters. *Nucleic acids research* **40**, 2907-2924 (2011).

3 Meng, H. *et al.* Quantitative design of regulatory elements based on high-precision strength prediction using artificial neural network. *PLoS One* **8**, e60288-e60288 (2013).

4 Meng, H., Ma, Y., Mai, G., Wang, Y. & Liu, C. Construction of precise support vector machine based models for predicting promoter strength. *Quantitative Biology* **5**, 90-98 (2017).

5 Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature methods* **10**, 354-354 (2013).

6 Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nature biotechnology* **36**, 1005-1005 (2018).

7 Cuperus, J. T. *et al.* Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res* **27**, 2015-2024 (2017).

8 Rule, A. *et al.* Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Comp Biol* (2019).

9 Birkel, G. W. *et al.* The JBEI quantitative metabolic modeling library (jQMM): a python library for modeling microbial metabolism. *BMC bioinformatics* **18**, 205-205 (2017).

10 Suárez, A., Alvarez-Feijoo, M. A., Fernández González, R. & Arce, E. Teaching optimization of manufacturing problems via code components of a Jupyter Notebook. *Computer Applications in Engineering Education* **26**, 1102-1110 (2018).

11 van Vliet, M. Guidelines for data analysis scripts. *ArXiv* **abs/1904.06163** (2019).