DEPARTMENT OF BIOLOGY RESEARCH:

UNIVERSITY OF TURKU

Identifying thermophilic adaptation in Prokaryotes with machine learning and deep learning

Researchers: Akseli Reunamo (akseli.y.reunamo@utu.fi) and Pere Puigbò (pepuav@utu.fi)

Machine learning and deep learning are subfields of artificial intelligence. In these systems algorithms are rather trained than explicitly coded to find statistically meaningful structure of data. The learning can be defined as autonomous search for better results.
Learning is based on algorithms capability to measure one's performance. Performance is used as feedback to optimize algorithm's parameters to maximize the accuracy.
Optimization can be implemented several ways. In this research gradient decent is used.

Aims

This research aims to define and predict optimal growth temperature of Prokaryotes with AI based methods from gene usage (COGs) found on Clusters of Orthologous Groups (COGs) of proteins database. Adaption to temperature can be gained or lost in evolution.
Prokaryotes can be divided in four growth temperature classes that are hyperthermophiles, thermophiles, mesophiles and psychrophiles.

Results

Algorithm defined seventy-five COGs that affects to temperature adaptation (Fig 2.) Two most important functional categories were Nucleotide transport and metabolism and is Defense mechanisms (Fig 3.).

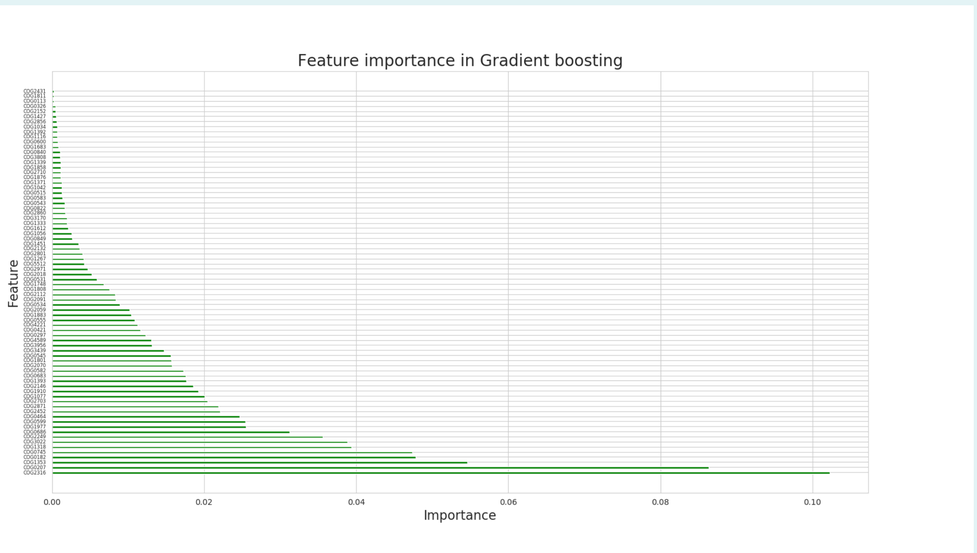

Figure 2. COGs that affected to class definition and their importance's

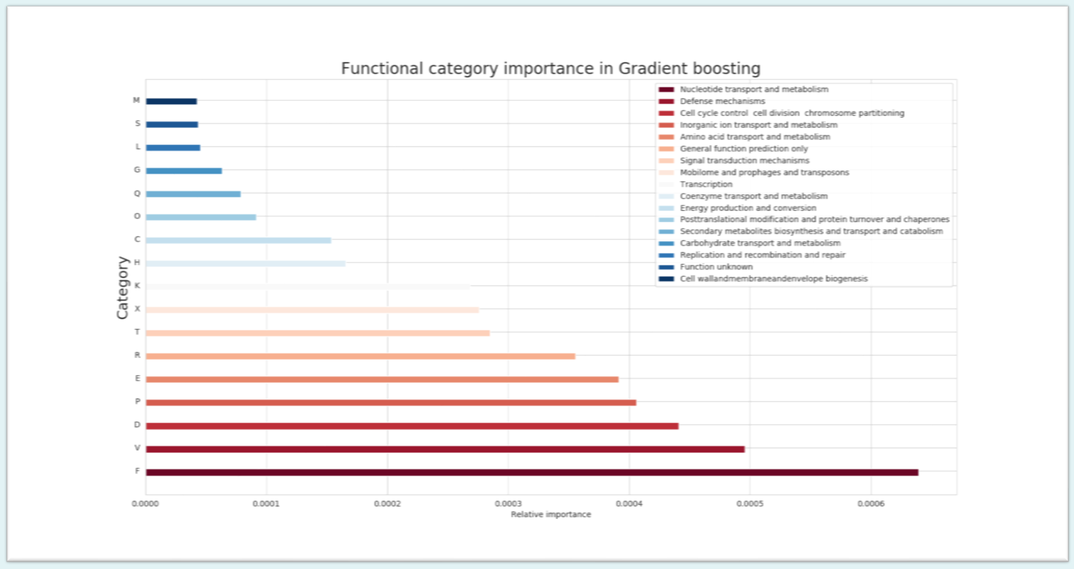

Figure 3. Importance of COGs according their functional class

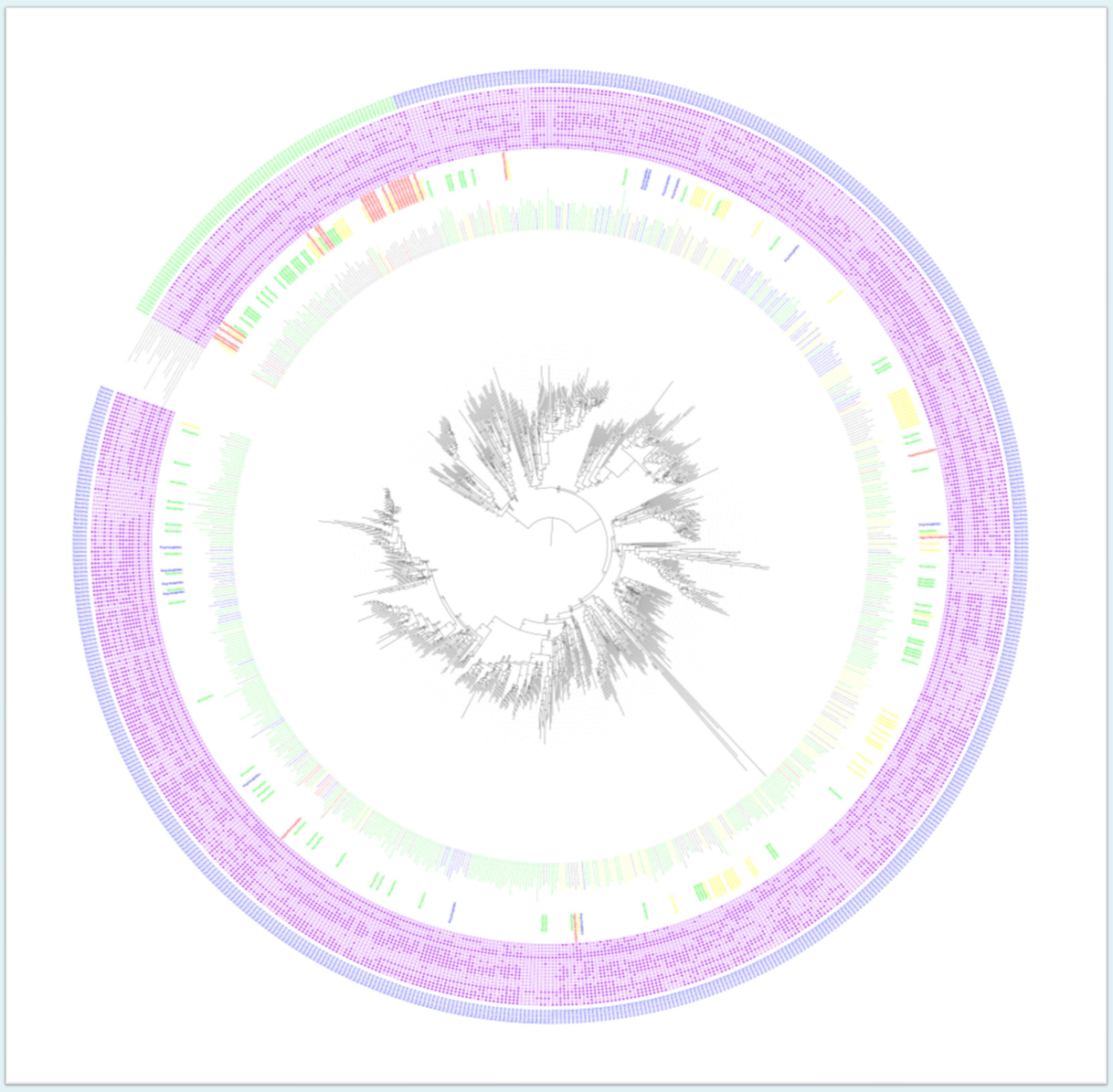

Figure 1. Phylogenic relations of species and their taxonomy, COG usage, known adaptation and predicted adaptation.

Methods

Prokaryotes are classified in four temperature classes with ensemble of machine learning methods including deep learning, random forests and gradient boosting. Main libraries used were TensorFlow, Skicit-learn and XGBoost.

Conclusions

Machine learning is powerful tool to determine temperature adaptation because it takes account of gene interactions. Algorithm is capable to classify Prokaryotes to right temperature class with more than 80% accuracy, including the separation of thermophiles and hyperthermophiles. Algorithm also allows to spot set of genes that might define thermophilic gene signature. Prediction power can be increased with combination of machine learning (gene content analysis and codon usage based methods.

References

Chen, T. and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: ACM Press, pp. 785–794.

Chollet, F. 2018. *Deep Learning With Python*. 1st ed. Shelter Island, New York: Manning Publications.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I. and Koonin, E.V. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research* 43(Database issue), pp. D261-9.

Puigbò, P., Pasamontes, A. and Garcia-Vallve, S. 2008. Gaining and losing the thermophilic adaptation in prokaryotes. *Trends in Genetics* 24(1), pp. 10–14.

Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28(1), pp. 33–36.

Turun yliopisto
University of Turku

Department of Biology, 20014 Yliopistonmäki, Turku

Poster was made by *Akseli Reunamo* for CompLifeSci Annual Meeting 2019