This factsheet is a result of the first Wageningen University & Research (WUR) Data Science and AI Fellowship program. With this program we aim to increase and integrate our expertise in DS/AI throughout the entire organisation. The variety of projects highlights the potential for DS/AI across the WUR domains.

Author: Barbara Terlouw

Finding determinants of protein translation efficiency

Via codon randomisation and machine learning

Objective

To identify which parts of the encoding of a gene are most relevant for protein production levels.

Method

We created ~1500 genes which all encode the same fluorescent protein, but are encoded in different ways, and for each gene measured fluorescence to see how much protein was produced. We used explainable machine learning to predict protein production from encoding, and determined which parts of the encoding were important for protein production levels.

Results

We found that primarily the start of the coding region of the gene and the base pairs right before the coding region were most important in determining how high protein production was going to be. Using the encoding of just this region, we were able to create a predictor that showed a correlation between actual and predicted protein production levels of around 80% (Figure 1).

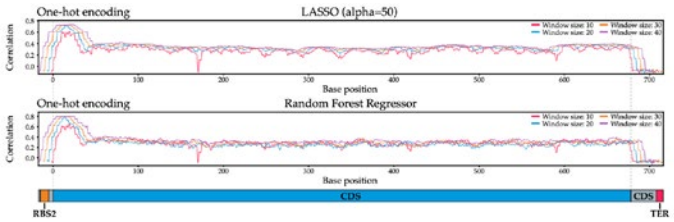

Figure 1 The correlation between actual and predicted protein production levels for predictors trained for parts of encoding across the entire length of the gene. Predictors trained on the start of the gene clearly perform much better than predictors trained on the rest of the gene.

Impact

It is highly interesting that only the encoding of the start of the gene and not the rest is critical for protein production. Currently, many protein production optimisation strategies involve changing the sequence of the entire gene. Our research shows that it may be much more effective and affordable to optimise only the start of the gene, rather than the entirety of it. We expect that, in the near future, methods for protein production optimisation will change drastically as a result.

Future plans

While we have demonstrated that the encoding of the start of the gene is most important for protein production, we have not yet been able to determine why certain encodings perform better than others. One reason for this is that we only looked at a single protein, and from this limited dataset it is impossible to create a generalised algorithm that would work for all proteins. Ultimately, we would want to be able to engineer the start of any gene in such a way that protein production would be optimal. To achieve this, more proteins will need to be analysed in similar fashion.

Further information

Contact persons:
Prof. John van der Oost (Laboratory of Microbiology) and
Prof. Dick de Ridder (Bioinformatics Group)

Project duration: April 15th 2021 – June 30th 2021

WAGENINGEN
UNIVERSITY & RESEARCH