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: Pascal Duenk

Finding interactions between genes and other genomic elements

Connecting WUR data

Using machine learning

Objective

The objective of this study was to find interactions between genes and other genomic elements from highly dimensional genomic data using machine learning.

Method

We used a publicly available dataset from a human cell line. The dataset contained information on genomic signals such as gene expression and DNA accessibility. We used this information to derive 16 features of every gene-element pair, and we used these features in a random forest model to predict whether pairs interacted or not.

Results

The results showed that interactions could be reasonably well predicted, with an F1 score of 0.42 and a weighted correlation coefficient of 0.49 (Table 1). These results were obtained with a random validation strategy, where the genome was randomly divided into a train and test set. In addition, we used a validation strategy where data from two chromosomes were used as the test set. With this validation strategy, interactions could not be predicted at all (F1 score and correlation of coefficient close to zero). This suggests that predictions of our model mainly rely on overlapping gene and element features across the test and training set, rather than on biological signatures of genome interactions.

Table 1 F1 score and weighted correlation coefficient of predictions in the test set, using either random validation or validation based on chromosomes.

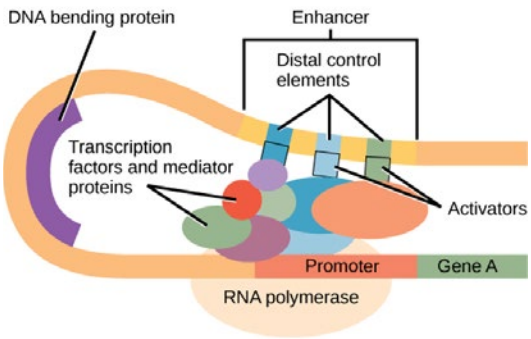
	F1 score	Correlation coefficient
Random validation	0.42	0.49
Chromosome validation	0.00	0.00

Impact

The results of this research shows that the chosen validation method can strongly influence the performance of machine learning models that aim to predict genomic interactions. This conclusion is in line with those from earlier studies. Thorough validation is an important step to ensure that the model uses biologically relevant information to make predictions, rather than overlapping information between training and test sets.

Future plans

The lessons learnt will be transferred to the researchers within GeneSwitch project, who will work with similar datasets and research questions. Furthermore, we aim to do a follow-up study where we try to improve our derived features, based on the approach of an earlier study.



Further information

This research was conducted by dr. Pascal Duenk, who is a lecturer and researcher at the Animal Breeding and Genomics group of WUR. More information about the GeneSwitch project can be found at www.gene-switch.eu.

For more information, contact pascal.duenk@wur.nl.

WAGENINGEN
UNIVERSITY & RESEARCH