

Program Instruction Manual

Feature Calculation Programs

Features were calculated using five Perl programs. Each program is named after the corresponding feature. For example, F_3.pl calculates the feature 3 and so on. Every program has two options for input data, UCE or WG, that will be calculated separately per input data.

Command line:

```
perl F_3.pl UCE or perl F_3.pl WG
```

Therefore, to calculate the feature for both UCE and WG sequences, the program must be ran twice with the corresponding argument as shown above. Each program saves the specified feature in a separate file. Each output file is names according to its feature and input data. For example, score3_UCE or score3_WG.

Feature table Program

Once all the feature calculation programs are ran for both UCE and WG, the final input table can be created using the F_input_tab.pl program. This program will result in a csv file named "input_table.csv" that can be used for machine learning algorithms.

Command line:

```
perl F_input_tab.pl
```

Machine Learning Models in R

Three machine learning models were trained and tested in R version 4.2.3; SVM, Random Forest, and Neural Network using the UCE3_ML_code.txt R code. This code contains the preprocessing, model training/testing, results analysis and statistics, and ROC-AUC curves.

For this code 8 packages must be installed and loaded into the environment: e1071, caTools, caret, pROC, randomForest, nnet, predtools, and ggplot2 (read materials and methods for more information).

The same three models were trained and tested in Python version 3.12.2 using the SciKit Learn package as well (see materials and methods for more information). This code is available in UCE3_ML_code2.txt and follows the workflow available on the SciKit Learn website (<https://scikit-learn.org/stable/>).

PCA

Principle component analysis was done using the raw input table, without preprocessing. This analysis can be done using the code PCA.txt For this code 2 packages must be installed and loaded into the environment; plotly and stats (see materials and methods for more information). The resulting PCA scatter plot is an animated three dimensional html file.