Machine Learning of DNA Profile Mixtures

Julia Gratsova

Supervised by Mr. A. Cobley, Dr. C. Cole

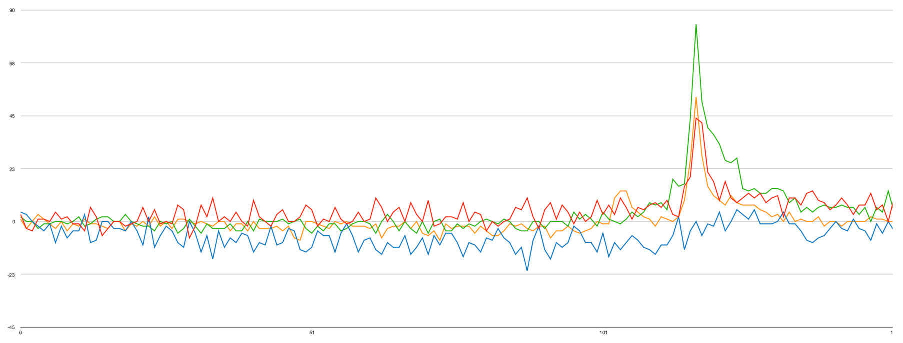
The aim of the project is to implement a Machine Learning methodology for identifying artefacts and peaks in DNA mixture profiles and aid complex mixture deconvolution while reducing the variability and the evaluation bias

Tools

- Python
- Pandas
- NumPy
- Scikit-learn
- Tensorflow
- Keras
- Scipy
- Matplotlib
- ANN
- Random Forest
- XGBoost

Research Observations

- Manual evaluation is prone to bias during interpretation
- No reliable benchmark of the accuracy of manual data evaluation by human analysts
- Results vary in different types of software due to the difference in the applied signal - processing methods
- Usage of analytical threshold in genotyping software can cause data loss



Results

- Model with 99.87% accuracy prediction on unseen dataset
- Peak data missed by genotyping software identified
- Exhaustive prediction across all of the data avoiding any hard analytical thresholds
- Benchmark results for ML usage in Forensics

Best Prediction Model Performance

Parameter	ANN	RF
Kappa Score	67.8	92.5
F1-Score	68	93
Precision	85	89
Accuracy	99.57	99.87
False Negative	15	17
False Positive	65	6
Peaks Identified	85 out of 150	144 out of 150

TensorFlow jupyter python™ Keras dmlc XGBoost THE DATALAB value from data

Leverhulme Research Centre for Forensic Science
University of Dundee
LEVERHULME TRUST

University of Dundee

Computing
MSc Projects 2018