AMSI VACATION RESEARCH SCHOLARSHIPS 2019–20

EXPLORE THE MATHEMATICAL SCIENCES THIS SUMMER

Getting the Most Out of Electropherograms

Louise Branch-Smith

Supervised by Associate Professor Nicola Armstrong

Murdoch University

Vacation Research Scholarships are funded jointly by the Department of Education

and the Australian Mathematical Sciences Institute.

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government Department of Education

Abstract

Artificial neural networks were investigated to correctly classify the baseline and peaks from an electropherogram using both untreated DNA and degraded DNA which received UV (Ultraviolet) treatment. The results of these networks could be a valuable aid for forensic investigations with almost 99% accuracy in classification. Extracting more information from an electropherogram other than just peak height, which is the current accepted method in both forensic and judicial fields to produce a DNA profile, could be paramount in both industries. Initial tests were difficult to conduct without the correct software being unattainable at this time due to cost, however including data for peak area, kurtosis and baseline width, could provide a great avenue for further research to be conducted in the future if the researcher has access to the relevant software.

Statement of Authorship

Under the direction of my academic supervisor I produced code to execute in R Studio to analyse the electropherogram samples and to create the artificial neural networks. I sourced the samples from a publically available database of human STR profiles from lftdi.com, and I interpreted the results, made the figures and tables and drafted the report.

Contents

1. Introduction

Electropherograms allow for genetic analysis in forensic laboratories and for the extension of biomedical scientific research. They are plots developed after scans of fluorescent intensity are detected, following the correct preparation and amplification process of the original sample in the laboratory. The fluorescence produces a peak which is measured as an arbitrary unit referred to as the RFU (Relative fluorescence unit), and directly reflects the concentration of labelled and separated DNA (Deoxyribonucleic acid) molecules, measured over time. When they are plotted,

the RFU goes on the y-axis, and the x-axis is for time measured in seconds, which also represents the fragment size in base pairs of the DNA being measured (Jamieson, 2009, pp. 259). An example of an electropherogram can be seen in Figure 1 below including both green and blue dyes (Taylor & Powers, 2016).

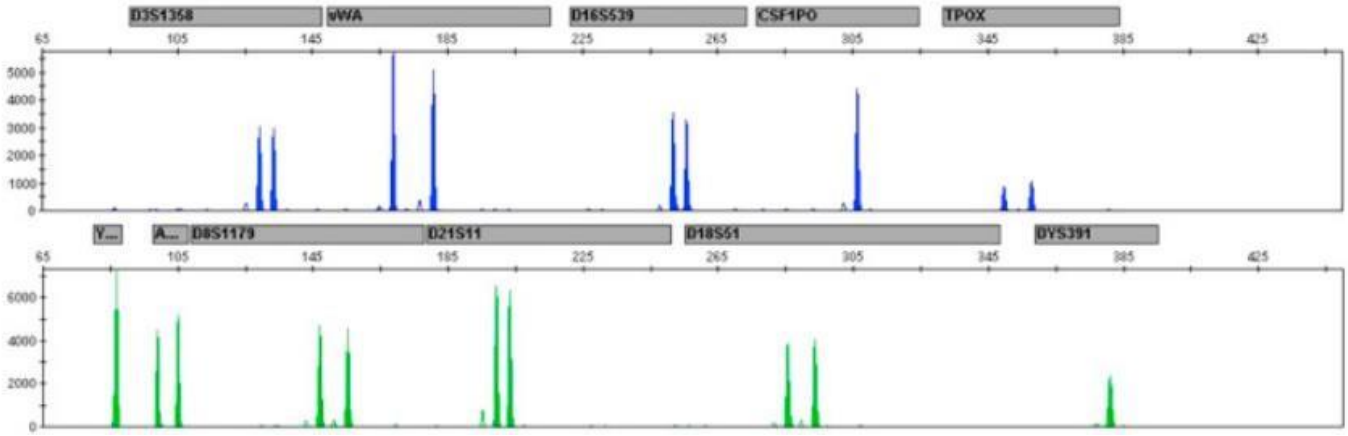

Figure 1. Image of an electropherogram including blue and green dye.

There can be many issues which can interfere with the production of a clear electropherogram, including stutter and pullups. These show up in the plots as peaks but are not actually peaks. They are common, yet sometimes inconsistent, even when analysing the same sample multiple times, which can make reading them a difficult task. PCR (polymerase chain reaction) artefacts are a common cause of problems created during the preparation of the samples and can also affect the quality of electropherograms produced (Jamieson, 2009, pp. 9495).

The peaks in an electropherogram are produced when polymorphic STR (short tandem repeat) regions are found at a specific location on the chromosome known as a locus (Butler, 2014, pp. 265). These are areas on the chromosome where the DNA base pairs repeat themselves in sets of four and are known as tetranucleotides. The number of times they repeat is called an allele and are different on each pair of chromosomes. This is due to production of the embryo resulting from one chromosome from the paternal mother and one from the paternal father. Sometimes these repeated regions, or alleles, are the same for a specific locus as each parent contributed the same number of repeated units for that STR on each chromosome. The combination of alleles is unique for each person and assists to produce a DNA profile, which is the current accepted method in the forensic and judicial communities (Butler, 2014, pp. 58).

Peaks can be therefore be homozygote where they produce one peak usually twice as high in RFU's as other peaks at different markers and is due to the individual having the same allele from each paternal parent at that loci. Alternately heterozygote peaks are found where they have produced two peaks and are usually the same height as each other on the same marker (Butler, 2014, pp. 241).

The aims of this project were to develop a method that used the raw electropherogram to gather information on peak area, base width, kurtosis, as well as peak height to classify the alleles present and investigate whether using all of the information extracted, improved allele calling in degraded DNA samples. The over-arching aim was to improve the readability and statistical evidence in court and therefore increase the number of accurate outcomes in forensic cases. This could allow more convictions of criminals who would otherwise have been released to the community, or alternately provide an innocent member of society their freedom.

2. Methods

2.1 Amplification Kit

Electropherogram peaks are read at different markers on several chromosomes using different amplification kits. The kit used in the sourced profiles for this project was the Identifiler Plus kit consisting of 15 markers and the amelogenin or sex marker. This kit uses four coloured dyes for different markers including FAM which is a blue dye and reads four loci, VIC which is a green dye and reads five loci, NED which is yellow and reads four loci, PET which is red and reads three loci, as well as LIZ an orange dye which is used for the internal size standard. In this project, the focus was on the blue FAM dye. Table 1 below details the loci for this dye, the chromosomes they are located on, and the tetranucleotide repeat sequences involved.

Table 1. Loci, chromosomal location and sequence of the blue FAM dye from the identifiler amplification kit.

Locus	Chromosomal Location	Sequence

D8S1179	8	TCTA
D21S11	21	TCTA, TCTG
D7S820	7	GATA
CSF1PO	5	AGAT

2.2 STR Profiles

STR profile samples were downloaded from the PROVED lt (Project Research Openness for Validation with Empirical Data) project available at the Laboratory for Forensic Technology Development and Integration database found at lftdi.com. The five second folder in the file named PROVEDlt_RD14-00003_1-Person Profiles_3130 5sec_IDPlus 28cycles was used. Different parts of the file name represent different variables relating to the experiment used to generate the data. RD14-0003 is the project number; 1-Person Profiles are for single contributor samples; 3130 indicates the capillary electrophoresis instrument used for these samples was the 3130 genetic analyser and 5sec refers to the injection time of five seconds. IDPlus stands for the Identifiler Plus amplification kit and 28cycles is the number of PCR cycles that the samples were processed through.

2.3 Software

R v3.5.2 was used for analysis and read the fsa files directly into the program through a package called “binner”.

2.4 Artificial Neural Network (ANN)

In order to classify each section of an electropherogram as either baseline or belonging to a true peak, an ANN was constructed with three layers. The input layer consisted of 201 neurons, the hidden layer had 20 neurons and the output layer had two neurons (representing baseline and peak classifications). Input data to both train and test sets for the ANN consisted of scans from

electropherograms starting from scan 3,000, as generally anything lower than this is just PCR artefacts. The electropherograms were continued to be read until scan 8,999 resulting in 6,000 scans per sample.

2.5 R Code for ANN

Code for constructing and testing the ANN was obtained from the article by Taylor and Powers (2016). This was altered to accept a different number of neurons per layer to reduce the time to complete each analysis, as well as to accept the STR profiles from the PROVED lt project as they had different column names.

2.6 ANN data sets

STR's from two subjects' untreated DNA profiles were used to create the training data for the artificial neural network. The same two subject's DNA profiles were used for the test data, with UV treated STR profiles. One was exposed for 15 minutes and the other for 60 minutes. For both the training and test data, the scans were annotated manually to denote if they were baseline or peak regions.

3. Discussion and Conclusion

Many programs were trialled for this project and it was found that the program Genemapper available directly from Thermofisher who supplies the Applied Biosystems genetic analyser used to process the STR samples, would have been ideal to extract the desired information from the electropherograms. This was extremely expensive and not a viable option. An application for a trial version but was not received before the project completion deadline. The files produced from the Applied Biosystems analyser are fsa files and cannot be opened in a text browser.

As an alternative to Genemapper, OSIRIS was investigated as a free option to convert the fsa files to an xml format for use in the R statistical program. OSIRIS would not recognise the fsa files.

Investigation into the fsa files downloaded from lftdi.com revealed that they each had a lock file attached to them and this may have been preventing them from being recognised by OSIRIS. Attempts were made to strip the locks from the files, but they were still unable to be read by OSIRIS.

Genemarker, software available at Murdoch University for forensics students, would not perform the operations required to extract the desired information from the raw data. Likewise, Geneious and Galaxy were also unsuccessfully investigated as alternatives to extract the raw information from the fsa files.

Peakscanner, a small downloadable program available from the Thermofisher website upon request, analysed the fsa files but would only reveal limited information of peak area and peak height. It would not convert the scans into base pair values which could then be used to interpret and predict the allele calls.

R studio was also used to analyse the data which allowed for base pair readings and peak height, as well as with some extra manipulation, peak area and kurtosis could be calculated. However, there was no clear way to identify the link between this information and which allele they corresponded to as it was a combination of information from base pairs and time and there was no way to unify the information.

This is when the approach to looking at artificial neural networks was taken as programs would do different parts of the required analysis or none at all, but not one program extracted all information required.

The electropherograms from the STR profiles used in the ANN of untreated and therefore nondegraded DNA samples are shown in Figure 2. Figure 2A on the left has a homozygote peak at the first marker as circled below, and then heterozygote peaks for the remaining markers. All peaks in Figure 2B below are heterozygote and one is shown circled also.

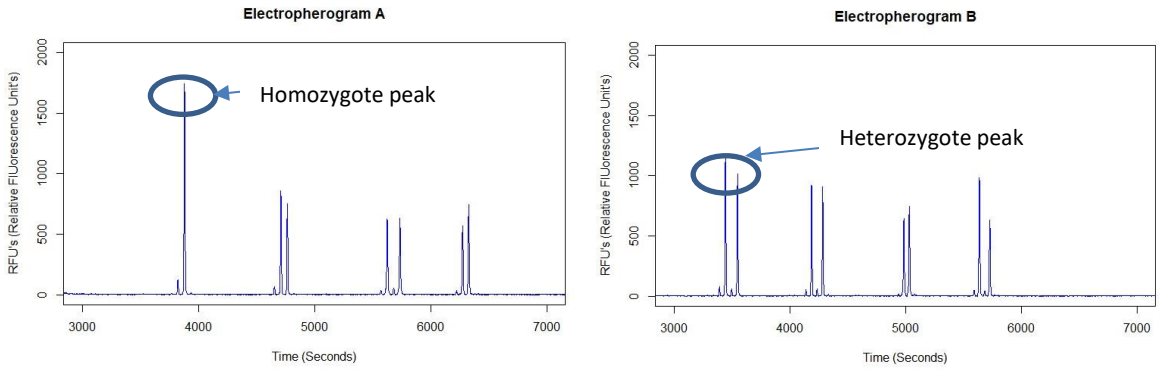

Figure 2. Electropherogram A (left) and B (right) both from untreated DNA STR profiles.

Degraded samples from the same subject as those in Figure 2 above were UV treated and produced the electropherograms in Figure 3 below. Figure 3A on the left received 15 minutes of UV exposure, while Figure 3B on the right was exposed to UV for a duration of 60 minutes.

You can easily visualise the difference in peak heights with the plot parameters being the same in each of the four plots from Figures 2 and 3. The third marker in Figure 3A at approximately 5800 seconds has completely disappeared in the UV treated DNA compared to the untreated DNA in Figure 2A. Likewise, with the same marker in Figure 3B at approximately 5000 seconds compared to Figure 2B. Both areas are pointed to with a red arrow in Figure 3 below.

The peak heights have clearly decreased in both electropherograms from Figure 3 compared to the those from each of the same subject in Figure 2. There does not appear to be a lot of difference between the two electropherograms in Figure 3 when comparing the differences in UV exposure time. Figure 3B received four times the length of time of exposure to UV and appears no more degraded than the sample in Figure 3A.

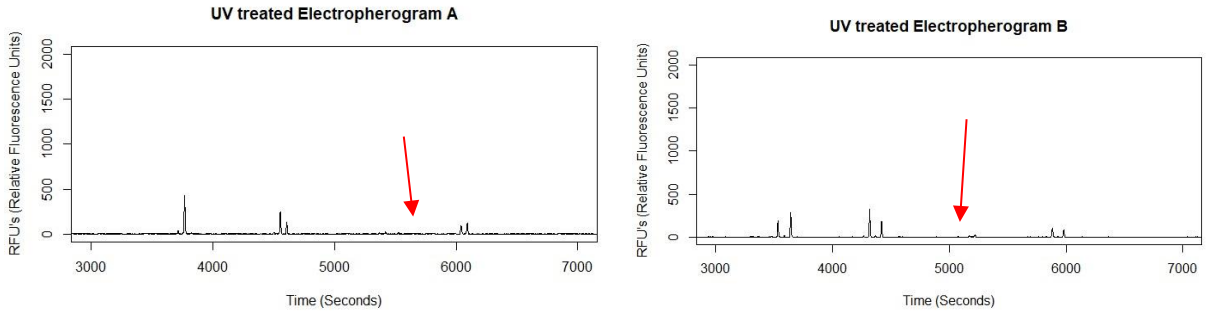

Figure 3. Electropherogram A (left) received 15 minutes and B (right) 60 minutes of UV exposure to DNA.

Artificial neural network results are summarised in table two below for the training data. This shows that sample A had the baseline called correctly 100% of the time and the peaks 98.5% of the time. This concludes that sample A was correctly classified more than 99% of the time. Sample B showed similar results with the baseline also being correctly classified 100% of the time and the peaks 98.5% of the time with an overall accuracy of more than 99% also.

Table 2. Artificial neural network results for sample A and B showing the percentage of correctly classifying either the baseline or the peaks of the electropherogram.

Training Set	Sample A		Sample B	
	Baseline	**Peaks**	**Baseline**	**Peaks**
Correct Classification	100%	98.5%	100%	98.75%
Total	99.27%		99.37%	

The test data results are in table three below and show that sample A which was UV treated for 15 minutes correctly classified the baseline 76.83% of the time and the peaks 99.66% of the time, with an overall correct classification of more than 99%. Sample B, which was UV exposed for 60 minutes correctly classified the baseline 79.22% of the time and the peaks 99.33% of the time with an overall correct classification of 98.82%.

Table 3. Artificial neural network results for sample A and B showing the percentage of correctly classifying either the baseline or the peaks of the electropherogram.

Test Set	Sample A with UV 15 minutes		Sample B with UV 60m minutes	
	Baseline	**Peaks**	**Baseline**	**Peaks**
Correct Classification	76.83%	99.66%	79.22%	99.33%
Total	99.08%		98.82%	

These are great initial findings as the importance of the baseline being called correctly is of little use to develop a profile. The peaks are the valuable information which contributes to the development of a DNA profile, and therefore would be the focus of the success of the ANN.

Further research could be conducted altering the number of inputs, hidden and output layers number of neurons, as well as choosing different samples, multi-contributor samples and different levels of degradation samples, and the possibility of including more parameters for prediction like stutter, pull ups and other PCR artefacts. This could provide a valuable tool in the future of forensics and statistics to minimise manual scanning of electropherograms, and therefore reduce human error.

Acknowledgements

Thank you to the Australian Mathematical Sciences Institute for this opportunity. I have learnt a lot from this experience and will take this valuable knowledge with me into my honours degree.

Thank you to Associate Professor Nicola Armstrong for your continued support as a supervisor.

Thank you to Dr Shane Tobe for the idea contribution and for initial access to the STR profiles.

4. References

Alfonse, L, Garrett, A, Lun, D, Duffy, K & Grgicak, C 2016, 'A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVED It', Forensic Science International: Genetics, vol. 32, no. 2016, pp. 62-70.

Butler, J. M. (2014). *Advanced topics in forensic DNA typing: interpretation*: Academic Press.

Jamieson, A. (2009). Introduction to Forensic DNA Profiling-The Electropherogram (epg). *Wiley Encyclopedia of Forensic Science*, 1-13.

Taylor, D & Powers, D 2016, 'Teaching artificial intelligence to read electropherograms', Forensic Science International: Genetics, vol. 25, no. 2016, pp. 10-18.