Forensic Science International: Genetics 30 (2017) 114–126

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

An artificial neural network system to identify alleles in reference electropherograms

Duncan Taylor[a,b,*], Ash Harrison[b], David Powers[b]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia

ARTICLE INFO

Article history:
Received 3 April 2017
Received in revised form 20 May 2017
Accepted 6 July 2017
Available online 8 July 2017

Keywords:
Electropherogram
Gel reading
Artificial neural network
Allele detection

ABSTRACT

Electropherograms are produced in great numbers in forensic DNA laboratories as part of everyday criminal casework. Before the results of these electropherograms can be used they must be scrutinised by analysts to determine what the identified data tells them about the underlying DNA sequences and what is purely an artefact of the DNA profiling process. This process of interpreting the electropherograms can be time consuming and is prone to subjective differences between analysts. Recently it was demonstrated that artificial neural networks could be used to classify information within an electropherogram as allelic (i.e. representative of a DNA fragment present in the DNA extract) or as one of several different categories of artefactual fluorescence that arise as a result of generating an electropherogram. We extend that work here to demonstrate a series of algorithms and artificial neural networks that can be used to identify peaks on an electropherogram and classify them. We demonstrate the functioning of the system on several profiles and compare the results to a leading commercial DNA profile reading system.

1. Introduction

A common task for any forensic DNA laboratory is the generation of short tandem repeat (STR) DNA profiles. Before these profiles can be used in interpretations they must be scrutinised by analysts to determine whether the information in the profile is representative of some component of DNA in the extract used to generate it, or if it is an artefactual product of the DNA profiling process. This task of 'reading' the electropherogram (EPG) can be time consuming and often leads to subjective differences between analysts. A recent work by Taylor et al. [1] demonstrated an artificial neural network (ANN) that was trained on two good quality reference EPGs to classify data in the 6-FAM dye lane and then applied to a third (also good quality) EPG with reasonable success. Taylor et al. [1] provided a proof of concept that ANN could be used to interpret EPGs, which we extend here by:

1) Increasing the amount of training data
2) Increasing the range of training EPG quality from completely blank to highly overloaded
3) Improving on the architecture of the ANN used
4) Training a series of ANN that are used on different areas of the EPG
5) Coupling the predictions of the ANNs with a peak detection algorithm originally designed for LCMS data [2,3] and recently extended to DNA EPG data [4] to produce a peak detection and classification system

Having created the peak detection and classification system we trial it on several profiles and demonstrate the results, which we compare to the peaks flagged by Genemapper® ID-X (Life Technologies).

2. Method

2.1. ANN input data

An EPG consists of a measure of fluorescence (called relative florescence units, RFU) for a number of dye lanes at various timepoints (called scans). To classify each 'scan' it was deemed that the input data for ANN training would be the scan in question and

[*] Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2017.07.002
1872-4973/© 2017 Elsevier B.V. All rights reserved.

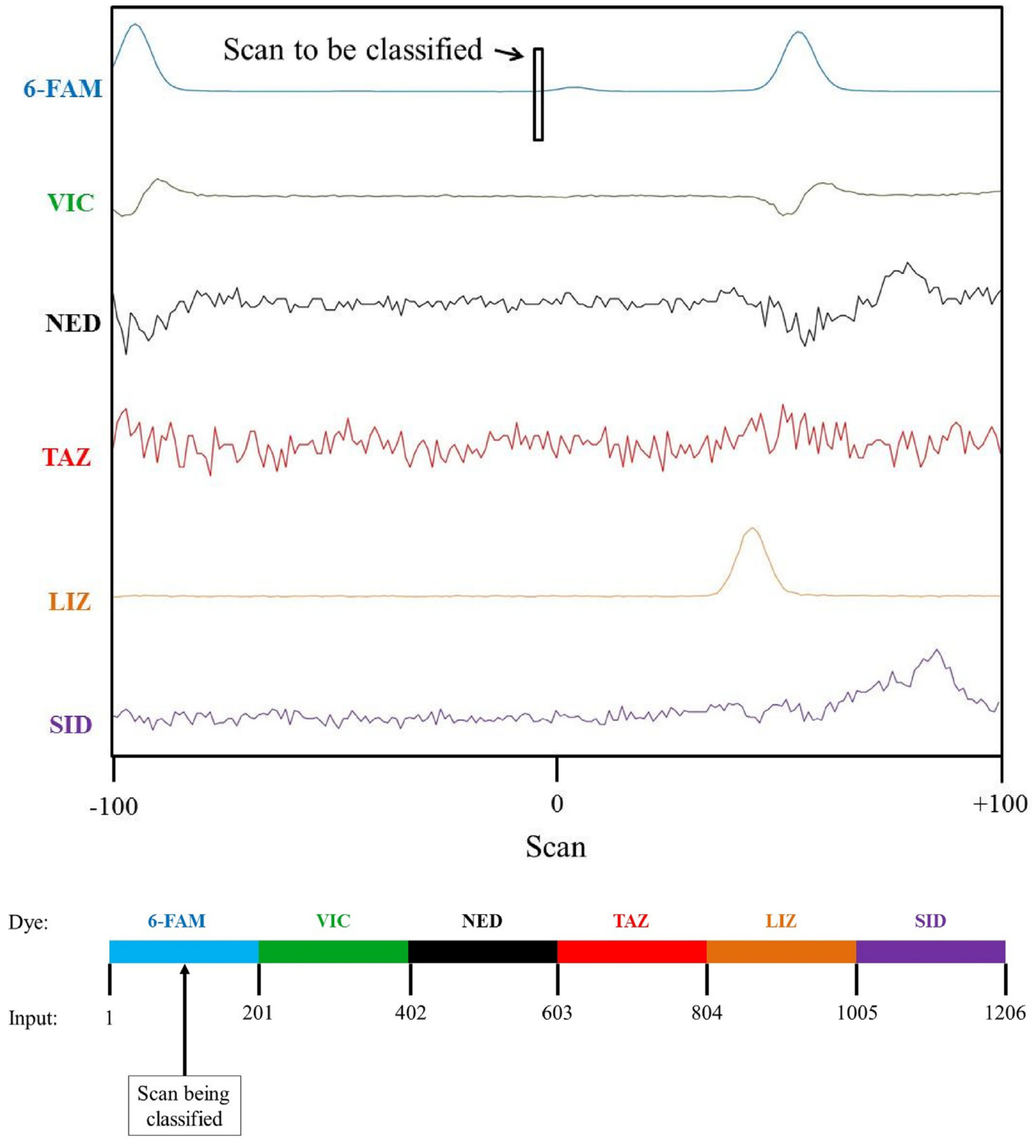

Fig. 1. Data used as input to classify central scan point in the 6-FAM dye lane in EPG context (upper) and in the context of an input for an ANN (lower).

100 scans in either direction, in all dye lanes, which corresponds to approximately 8 base pairs (bp) in both directions. This information is presented diagrammatically in Fig. 1 for a GlobalFiler^{TM} DNA profile (which possesses six dyes). The result is 201 scans in each of six dyes, leading to 1206 inputs for each training set.

All profiles used in this work have been produced using the GlobalFiler^{TM} DNA profiling system and run on a 3130xl genetic analyser (Life Technologies).

The number of outputs in the ANN depends on the number of categories of feature that we wish to classify. There is a trade-off between distinguishing large numbers of categories so as to provide ANN with many distinct (potentially diagnostic) data patterns, and the increase in the required training data to generalise well using a large number of classifications. In this work we consider the following features:

- Baseline
- Allele
- Back Stutter (one repeat unit shorter than the allele)
- Pull-up
- Forward stutter (one repeat unit longer than the allele)
- Half Stutter (half a repeat unit, typically 2 base-pairs, shorter than the allele)

We break up the category of ‘stutter’ into three categories for two reasons. Firstly, they are each distinct in their relative position to allelic peaks and secondly, different loci in the EPG have different combinations of these stutter features. In our training we also classify double back stutter in the ‘Back Stutter’ category and stutter that is one and a half repeat units shorter than the allele as Half Stutter. There were 10 ANN trained for reading all GlobalFiler^{TM} EPG data (Table 1).

There are a couple of points to note from the information provided in Table 1. Scans in each dye require their own training data as the position of the scan being classified within the 1206 input into the ANN varies (i.e. for the LIZ ANN the scan being classified would sit at position 905 in Fig. 1, lower panel) and the pattern of pull-up from the surrounding dye lanes is different. In the cases of VIC, TAZ and SID, multiple ANN are required. In the case of VIC and SID this is due to the different stutters exhibited by different loci within the dyes, specifically while all STR loci exhibit back stutter and forward stutter, the amelogenin and Y-indel loci within VIC are not STR and so exhibit no stutters. Within SID, D1 is