Appeared in Proc. of 4th International Conference on Bioinformatics and Computational Biology (BICoB), (Las Vegas, USA), March 2012.

Detecting STR peaks in degraded DNA samples

Emanuela Marasco*, Arun Ross*, Jeremy Dawson*, Tina Moroose†, Tanya Ambrose†
*Lane Department of Computer Science and Electrical Engineering,
West Virginia University, PO Box 6109 Morgantown, WV 26506
email:{emanuela.marasco, arun.ross, jeremy.dawson}@mail.wvu.edu
†Forensic and Investigative Science Program
West Virginia University, PO Box 6121 Morgantown, WV 26506
email:{tina.moroose, tanya.ambrose}@mail.wvu.edu

***Abstract*—Human identification from DNA is typically based on 13 short-tandem repeat (STR) alleles. Commercial kits used in forensic casework rely on the detection of these alleles in DNA samples acquired from an individual. However, the process itself is slow (it can take up to 2 days when conducting a laboratory analysis or 1 hour when using Rapid DNA systems) and has been designed to operate on pristine DNA samples. The need for achieving fast and accurate DNA processing has spurred efforts in developing portable systems that can reduce the processing time to less than 1 hour. But such systems are expected to operate on degraded DNA samples due to the architecture and process used by the instrument. Consequently, detecting the alleles in such degraded DNA samples can be a challenging problem. In this paper, we present an algorithm to detected allelic peaks from degraded DNA signals based on an adaptive signal-processing scheme. The performance of the algorithm is evaluated on two datasets: 1) data collected at the WVU Department of Forensic and Investigative Sciences, obtained by performing a controlled DNA degradation using ultraviolet radiation, 2) data provided by NIST obtained by varying cycle counts for the PCR processing step. Experiments indicate the efficacy of the algorithm in allelic peak detection and reiterate the need for approaching the problem in a systematic manner.**

I. Introduction

In both battlefield (U.S. Title 10) and criminal justice (U.S. Title 18) forensic scenarios, the importance of fast, accurate DNA analysis has driven the development of portable sequencing systems that can accept whole cells as input and produce a high-base-pair-resolution allelic separation as output [1],[2]. Major advancements have been leveraged by miniaturization and/or procedural modification of well-known bench-top processes, with interpretation of data performed using standard fluorescence spectroscopy and human expert data analysis. Current rapid DNA system development efforts focus on identification based on the 13 human short-tandem repeat (STR) alleles used in forensic casework for which commercial kits are readily available. These technologies could be viewed as first-generation molecular biometric systems that have the potential to enhance current automated systems based on physical and behavioral biometrics. Hardware development has greatly outpaced efforts aimed at automated DNA signal analysis. Novel approaches to DNA signal processing are necessary to fully utilize any level of available information in DNA signatures, either from pristine or degraded data both at the allelic andgenomic level. Due to degradation, STR peaks are often shifted in relation to the size standard, and either present lower relative fluorescence unit (RFU) measures or may be missing all together (i.e., drop-out). For example, the DNA typically found in outdoor environments is often severely degraded due to heat and ultraviolet radiation from the sun. Exposing DNA to UV light (10-400nm) induces the formation of cis-syn cyclobutane-pyrimidine dimers while heat denatures it [3]. This degradation will typically result in a sample that cannot be easily used for human identification purposes. However, allele information can be inferred from such purposefully degraded samples, amplified, and be made suitable for use in human identification. Degraded samples sometimes contain less than 100 picograms (pg) of template DNA and the presence of such low copy number (LCN) samples could be due to several factors including damaged or degraded DNA. Recovery of DNA profiles from LCN samples is difficult using standard STR methods, and such attempts often result in total failure or recovery of partial profiles. This is an expected outcome since commercial STR kits have been optimized to produce good quality, balanced profiles with 1 nanogram (ng) of DNA subjected to 28-30 PCR cycles. Hence, special LCN methods, based upon increasing the PCR cycle number in order to enhance allelic signal intensity, have been developed to permit profile recovery from limited quantity samples [4],[5]. These lab methodologies could be further strengthened if signal enhancement schemes are used to improve the quality of the input signal.

This paper focuses on addressing the failure in detecting drop-out STR peaks in degraded DNA samples. The proposed approach is able to increase the number of correctly detected peaks thereby facilitating human identification from LCN DNA. This study compares the performance of the proposed approach to that obtained using the commercial DNA analysis software *GeneMapper ID* (currently adopted in most labs) on challenging biological samples that contain less than 100pg of template DNA or degraded with UV light. The paper is organized as follows. In Section 2, we describe the current methods for DNA typing. Section 3 presents the proposed approach based on signal processing technology. Section 4 reports the experimental procedure and compares results against those obtained by *GeneMapper ID*. Section 5 presents our conclusions.

II. Traditional DNA analysis via software

Following the separation of amplified DNA products, the information from the DNA separation must be converted into a common language that is standard across laboratories and instruments. Software programs provide the means to perform the necessary data analysis and standardization of the output. Data produced in the separation and characterization of amplified DNA is displayed as fluorescence peaks (capillary electrophoresis) or bands (slab gel electrophoresis) as seen in Fig. 1. The steps for converting fluorescent data/peaks into

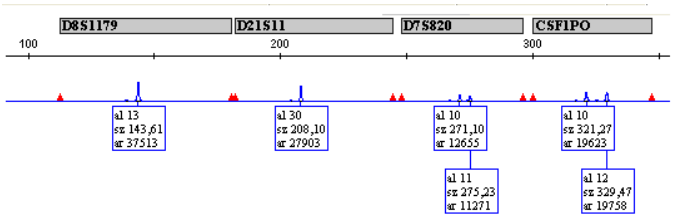

Fig. 1. The DNA fragments are sized, which includes an indirect assessment of quantity present (loci peak area/height or band density), and genotypes are assigned. The standard for conversion of sized DNA fragments to genotypes is based on the CODIS system.

allele calls are shown in Fig. 2 along with the corresponding software.

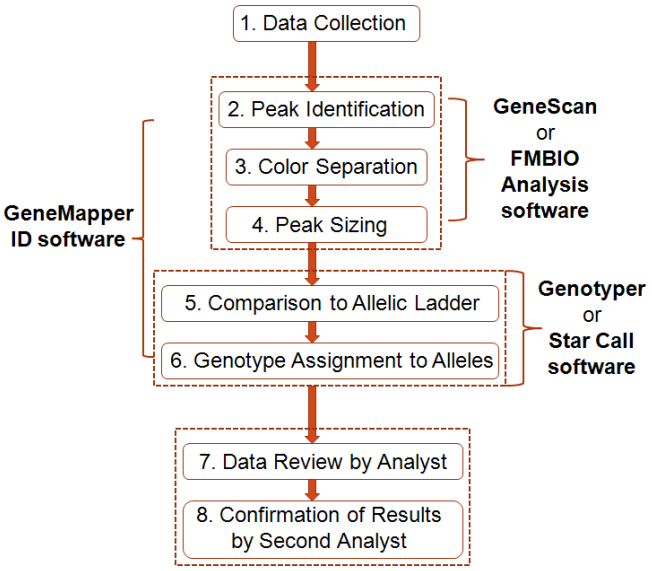

Fig. 2. The steps for converting fluorescent data/peaks into allele calls.

GeneScan is a sophisticated software program that converts raw data to useful data through the application of a size standard, a matrix file, and specific parameter settings. Based on threshold values, GeneScan software determines peaks, separates colors using the matrix file, and sizes peaks using the internal size standard added to the sample prior to separation, as shown in Fig. 3 and Fig. 4. Genotyper software converts GeneScan sized peaks into genotype calls using macros, by

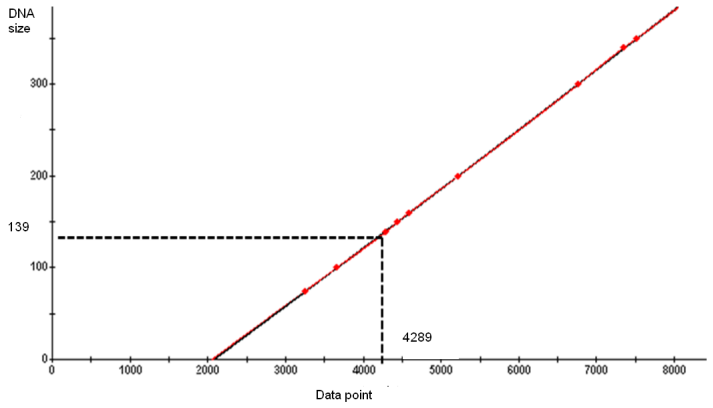

Fig. 3. DNA fragment peaks are sized based on the sizing curve produced from the points on the internal size standard.

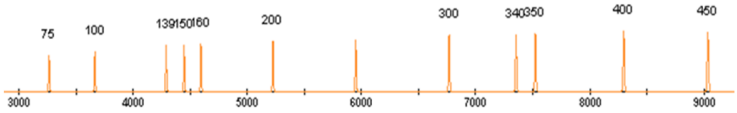

Fig. 4. Internal size standard.

comparing allele sizes in allelic ladder to sample alleles. GeneMapper ID (GMID) combines both GeneScan and Genotyper with the advantage of being 20%-40% faster than separately using GeneScan and Genotyper. It is an automated genotyping program that designates peaks in electropherograms by sizing, and makes allele calls through size comparisons with an allelic ladder. There are two analysis modes (classic Macintosh and advanced Windows NT modes). The differences between these modes are found in the sizing method and the flexibility of peak sensitivity settings. In the classic mode, size calling is performed by matching the actual size standard fragments of the sample with a defined size standard that must be accurately labeled; it utilizes scan number to assign sizes. In the advanced mode, size calling is performed using a function known as ratio matching. Ratio matching uses an algorithm to determine the distance between the size fragments based on a set of size fragment values, where sizing is based on the relative distance between the neighboring loci. It is important to note that the approach employed by these systems uses a fixed threshold. Any peak falling below this threshold is considered unusable in criminal justice applications[1]. Biometric applications of rapid DNA systems in tiered screening scenarios may have a relaxed standard for acceptance of sub-threshold peaks. The framework presented in this paper is aimed at providing identification based on a variable matching threshold accompanied by confidence measures (such as false positives

[1]http://www.statepolice.wv.gov/about/Documents/CrimeLab/9thmanual.pdf

and false negatives) rather than a tight, fixed threshold.

III. The Proposed Approach

The derivative signal processing technique has been widely used to detect a pristine signal that is mixed with interfering noisy signals [6]. In our problem, the signal represents a DNA sample, referred to as DNA signal in which the x-axis indicates the data point and the y-axis the amplitude; Fig. 5 shows an example of a DNA signal corresponding to raw data for blue and green dye, respectively. In the current analysis, we denote a DNA signal as $x(t)$ where t indicates the vector of data points. A degraded DNA sample is represented by a

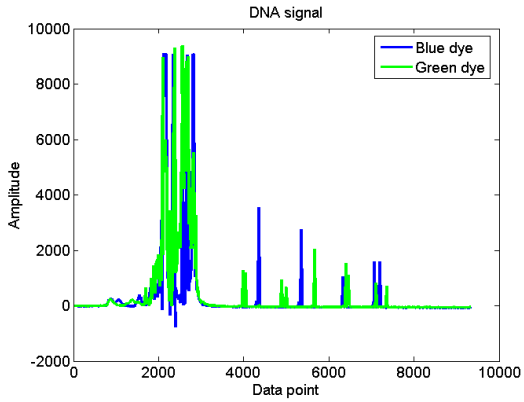

Fig. 5. A DNA signal representing raw data for blue and green dye.

weak signal confounded by several noise sources due to the instrument and biological processes. These noise sources make it difficult to measure the intensity of STR loci peaks. In this paper, finding the location of peaks is formulated as a problem of predicting the shape of signals using differentiation. Below, we first draw some fundamental properties of derivatives and then describe our procedure where such properties are exploited.

A. Differentiation of signals

The first derivative of the signal represents the slope of a given signal. It is positive corresponding to the points where the signal slopes up, it is negative corresponding to the points where the signal slopes down, and it is zero where the signal has no slope. This basic property of differentiation helps in predicting the shape of a signal. In our study, we focus on the fact that the first derivative of a peak has a zero-crossing point at the maximum. Let us consider the sigmoidal signal function characterized by some very useful mathematical properties. These include the presence of one *inflection* point which has the maximum slope at the center of the x-axis range (see Fig. 6); this point corresponds to a peak in the first derivative of the signal (see Fig. 7), and to a point where the signal crosses the x-axis, referred to as *zero-crossing* point in the second derivative (see Fig. 8) [7][8].

Further, an important property of the differentiation regards the effect of the peak width on the amplitude of the derivative signal. When considering peaks with the same height but

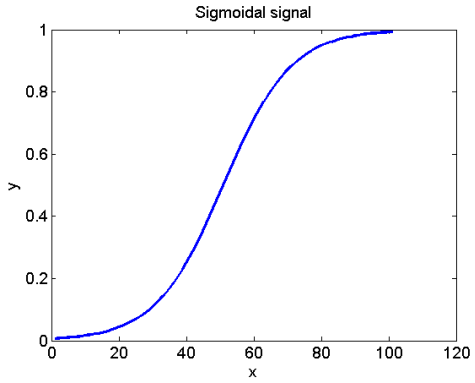

Fig. 6. The sigmoidal signal. It presents the maximum slope at the center of the x-axis range; such a point is referred to as inflection point.

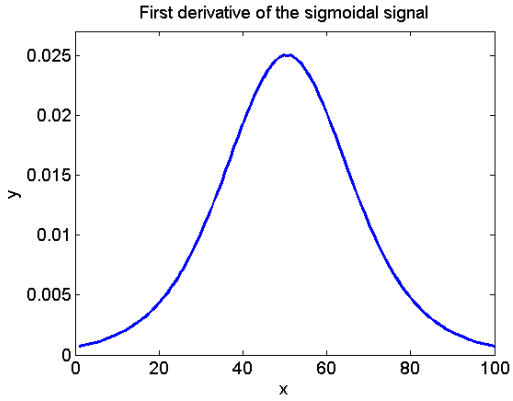

Fig. 7. First derivative of the sigmoidal signal. A peak in the first derivative corresponds to a point where the original signal has the maximum slope; it is the inflection point when considering the sigmoidal signal.

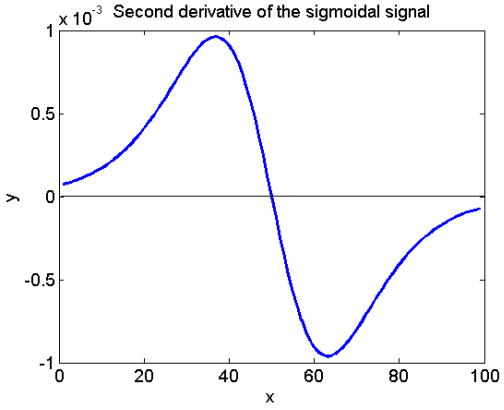

Fig. 8. Second derivative of the sigmoidal signal. The second derivative crosses the x-axis when going from its maximum to its minimum at a point referred to as zero-crossing point; the location of the zero-crossing point in the second derivative corresponds to the location of the maximum in the first derivative.

different width, wider peaks such as those corresponding to a noisy signal, present a lower derivative amplitude.

B. The proposed peak finding algorithm

Given a *peak-type* signal such as those used in STR analysis, the location of the maximum can be computed as location of the zero-crossing points in its first derivative. When dealing with real scenarios, the presence of noise may cause the detection of false zero-crossing points resulting in a number of false peaks [9]. This problem is addressed by smoothing the derivative of the signal, causing an attenuation of the amplitude of the resulting signal. Because differentiation and smoothing are both linear techniques, the amplitude of the smoothed first derivative is exactly proportional to the amplitude of the original signal [10]. The adopted technique detects peaks by selecting zero-crossing points in the smoothed first derivative; however, smoothing may distort the *peak-type* signal, reducing peak height and increasing peak width. To address this, the position, height and width of each peak are determined by re-fitting a segment of the original unsmoothed signal in the proximity of the selected zero-crossing point. So peak details (i.e., position in the x-axis and height in the y-axis) are determined using the proximity of the peak in the unsmoothed signal and a *curve fitting* function which does not distort it [2]. The positive peaks are detected by looking for zero-crossing points in the smoothed first derivative. The discrimination is based on an adaptive amplitude threshold based on the loci peak amplitude, which disregards any peaks with amplitude less than a threshold $\eta(x)$ computed as follows:

$$\eta(x) = \mu(x) + 2\sigma(x) \tag{1}$$

where μ and σ are, respectively, the mean and standard deviation of the signal x. The defined threshold varies with the input signal based on its mean and standard deviation. When considering signals that do not have a very high standard deviation on amplitude, the height value of the peaks is close to the mean; thus, by setting the threshold to a value proportional to the sum of the mean and standard deviation, true peaks are detected and most of the noise peaks are discarded. The first part of the DNA signal, corresponding to smaller x values, may contain some high-amplitude noise peaks that arise due to the byproducts of the amplification and labeling steps (see Fig. 14); such peaks increase the standard deviation of the signal. Therefore, in order to detect true STR loci peaks, it is not convenient to use a threshold based on the standard deviation value. Since we do not have to detect the capillary noise present in the first part of the signal, we assume that the mean of the true peak height a fraction of the maximum of the signal and adopt a value proportional to that maximum as the threshold. Given the maximum amplitude of the signal which corresponds to the maximum peak height, we use a threshold equal to a fraction of that maximum as follows:

$$\eta(x) = \alpha * Max(x) \tag{2}$$

where α was experimentally determined as the value that results in the maximum number of true peaks based on the ground truth provided in the positive control Gene Mapper scans. The discrimination of the peaks also takes into account the peak width, by selecting only those peaks whose slope in the smoothed first derivative exceeds a fixed threshold. Such a threshold is high enough to discard broad characteristics of the signal. Further, since wider peaks have a smaller derivative amplitude, the differentiation in general is able to discriminate against wider peaks since noisy peaks are wider than true peaks [3].

C. De-noising

In order to assist signal de-noising, the notion of a Negative Control (NC) is used. The NC contains only reagents (Taq polymerase, primers and buffer), and no DNA and Internal Lane Standard. Passing this sample through the DNA analysis instrument allows for the detection of contamination events that may have affected the reagents used after the amplification step. Since NC does not contain DNA, there should be no amplification. If an amplified signal is present, it is most likely due to some form of contamination in the reagents. Given a degraded signal x, the NC signal is subtracted from it, as follows:

$$y_j = x_j - NC_j, \quad j = 1 \ldots N. \tag{3}$$

This step improves the quality of the signal before peak finding.

Algorithm A illustrates the process flow of the proposed algorithm. Step 1 describes the de-noising of the signal; step 2 computes the first derivative of the latter portion of the signal; step 3 smoothes the derivative of the signal to avoid the detection of a large number of false peaks; step 4 gives details about the discrimination of peaks based on a slope threshold and an amplitude threshold; step 5 looks for zero-crossing points in the derivative of the signal; step 6 re-fits a segment of the original unsmoothed signal in the vicinity of the detected zero-crossing points.

IV. Experimental results

A. Dataset

The performance of the proposed approach was evaluated on two datasets. The first dataset was collected at the WVU Department of Forensic and Investigative Sciences by performing a controlled DNA degradation using ultraviolet radiation. Two types of DNA were used in this study. Initially the DNA used was the AmpFlSTR Control DNA 9947A (Lot number 1004074) from the Applied Biosystems AmpFlSTR Identifiler Kit. After some initial results using the 9947A DNA, three buccal swabs were collected from a Caucasian female. The buccal swabs were extracted using phenol-chloroform following the procedures of the West Virginia State Police Forensic Laboratory (WVSPFL) DNA Analysis Manual (West Virginia State Police Forensic Laboratory, 2009). The extracted DNA was then quantified using Real-time PCR and

[2] http://terpconnect.umd.edu/~toh/spectrum/CurveFitting.html

[3] http://terpconnect.umd.edu/~toh/spectrum/Differentiation.html

Algorithm A. The proposed algorithm for peak detection

Let $x = \{x_j\}_{j=1}^{N}$ be the input DNA signal.
Let $NC = \{NC_j\}_{j=1}^{N}$ be the signal due to the Negative Control.
Let $t = \{t_j\}_{j=1}^{N}$ be the time instances in which the signals are sampled.
Output: A list containing the number of peaks and the estimated position, height and width of each detected peak.

1) De-noise the DNA signal x:

$$y_j = x_j - NC_j, \quad j = 1 \ldots N. \tag{4}$$

2) Compute the first derivative of the enhanced signal:

$$\mathrm{y}_j = \frac{y_{j+1} - y_j}{t_{j+1} - t_j} \tag{5}$$

for $2800 \leq j \leq N$

3) Smooth the derivative of the signal by replacing each point in the signal with the average of m:

$$s_j = \frac{\sum_{j=-\frac{m-1}{2}}^{\frac{m-1}{2}} \mathrm{y}_j}{m} \tag{6}$$

for $2800 + \frac{m-1}{2} \leq j \leq N - \frac{m-1}{2}$, where m is a positive odd integer called the *smooth factor*.

4) Detect peaks:

$$M_j = \begin{cases} True, & \text{if } \operatorname{sgn}(s_j) \geq \operatorname{sgn}(s_{j+1}) \\ & \quad \& \ (s_j - s_{j+1} \geq \gamma \ y_j) \\ & \quad \& \ (y_j \geq \eta(x)) \\ False, & otherwise \end{cases} \tag{7}$$

for $2800 + m \leq j \leq N - m$, where $\gamma = 0.5 * w^{-2}$ is the slope threshold, $w = 50$ is the average number of points in half-width of peaks, $\eta(x) = \mu(x) + 2\sigma(x)$ is the amplitude threshold, and μ and σ are the mean and the standard deviation, respectively, of the signal x.

5) Compute the initial estimate of peak location and peak height:
k=0;
For j=1 to N
If M_j = True then
$k = k + 1$; $p_k = j$; $h_k = y_j$
(p_k is the peak location and h_k is the peak height)
endif
endfor
K=k;

6) Recompute the location p'_k, the height h'_k and the width w'_k of these peaks as follows:
For k=1 to K
Fit a polynomial of degree 2, $h = c_0 + c_1 j + c_2 j^2$, through a set of points in the vicinity of peak p_k and estimate the coefficients c_0, c_1 and c_2.
Let μ_P and σ_P be the mean and standard deviation, respectively, of these set of points.

$$p'_k = -\sigma_P \left(\frac{c_1}{2c_2}\right) - \mu_P \tag{8}$$

$$h'_k = \exp\left(c_0 - c_2 \left(\frac{c_1}{2c_2}\right)^2\right) \tag{9}$$

$$w'_k = norm\left(\sigma_P \frac{2.35703}{\sqrt{2} * \sqrt{-c_2}}\right) \tag{10}$$

endfor

7) Output the list $\{(p'_k, h'_k, w'_k)\}_{k=1}^{K}$.

Applied Biosystems Quantifiler Human DNA Quantification Kit following WVSPFL procedures and diluted, using sterile water, to the target range of 0.05-0.125 ng/uL needed for amplification. Amplification was performed using the Applied Biosystems AmpFlSTR Identifiler PCR Amplification Kit following manufacturer's protocols with the thermal cycler set at 28 cycles. The DNA was then analyzed on a 3130 Genetic Analyzer, following manufacturer's protocols, to determine the full profile of the DNA used for the experiments. The 3130 analysis used POP-4 Polymer, AmpFlSTR Identifiler Kit Allelic Ladder, GeneScan-500 LIZ Size Standard, a 3.0 kV injection voltage, 5 second injection time, 15.0 kV run voltage, and a 1500 second run time[11]. All samples used in this study were prepared in the following manner: 15μL of the extracted DNA was placed into 0.5 mL yellow tube and then briefly centrifuged to collect all liquid at the bottom of the tube. The crosslinker is factory set at 254 nm which correlates to an Intensity of 3500-4500 μW/cm^2. The irradiance display resolution is +/- 5 μW/cm^2 over the entire range. Inner chamber dimensions are 13.5 W x 7 H x 7.5 D in. There are 5 8-watt UV tubes set at 254 nm. The energy per unit area for each of the samples can be calculated as follows:

$$Energy(\mu J/cm^2) = Intensity(\mu W/cm^2) * Time(sec) \tag{11}$$

The intensity varies slightly but generally stays around 3700 μW/cm^2. The intensity was checked each day before any samples were irradiated to verify that it was within range. The

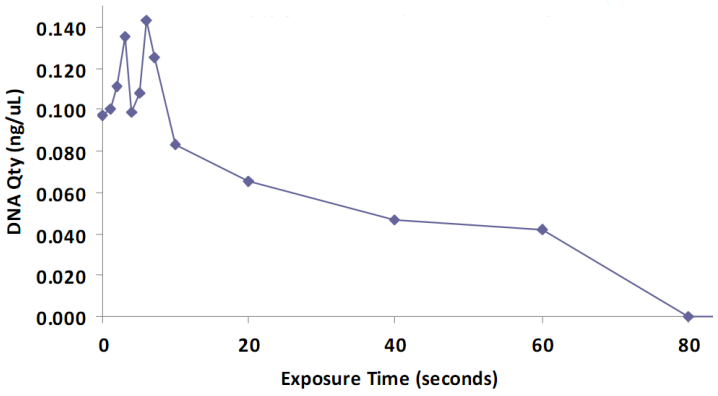

Fig. 9. DNA quantity (ng/μL) vs UV Exposure.

tubes were placed horizontally on the floor of the crosslinker about 1 inch apart from each other and in the middle of the chamber about 6.25 inches away from the UV lights [3]. Fig. 9 illustrates DNA degradation as a function of UV exposure time, indicating a sharp decrease after as little as 10s exposure. After 80s, the quantity of DNA present does not produce enough fluorescence to reach the RFU cut-off of the instrument. Signal irregularities before 10s can be attributed to the PCR amplification.

The second dataset was provided by NIST. Data were obtained using standard Identifiler reagent kits and variable

cycle counts for the PCR processing step. Positive controls at 1 ng/μL consist of 2 samples (MT and PT) with 10 replicates at 100pg, 30pg, and 10pg of DNA using Identifiler at 28 cycles (normal conditions) and 2 samples (MT and PT) with 10 replicates at 100pg, 30pg, and 10pg of DNA using Identifiler at 31 cycles (LCN conditions).

B. Results

For the purpose of our analysis, we considered each signal from t=2800 onward, since the first part of the signal is highly affected by the noise introduced by the amplification phase, as discussed previously. The signal shown in Fig. 10 represents the Positive Control sample which is the output of a reaction involving a pristine DNA sample of 1ng, reagents (Taq polymerase, primers and buffer) and Internal Lane Standard. It is taken from the first dataset. Fig. 11 shows an example

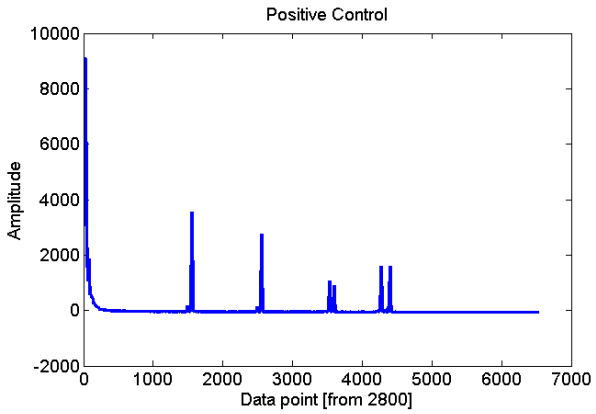

Fig. 10. Positive Control (Blue dye).

of degraded signal after exposing it to ultraviolet radiation for 75 seconds. It shows that the degradation causes lower, shifted and missing peaks. Fig. 12 reports the number of peaks

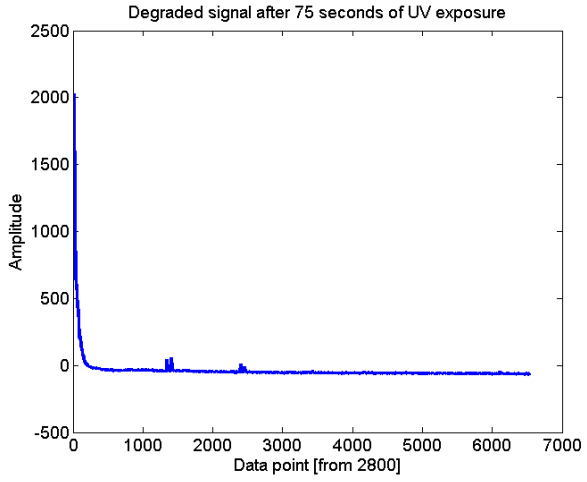

Fig. 11. Degraded signal after 75 seconds of ultraviolet radiation exposure (Blue dye).

detected by GeneMapper typing system for different fixed threshold values. In the presence of non-degraded samples, the number of actual peaks correctly detected is high and it remains constant even when decreasing the RFU value from 100 (value adopted by most of the laboratories) to 45, by reporting a limited number of false positive ranging from 2 (th=100) to 4 (th=45). However, in the presence of degraded samples, for standard threshold values (close to 100), the number of detected peaks is 1 for a degradation level corresponding to an UV exposure of 75 seconds and it becomes 0 for a degradation level corresponding to an exposure of 150 seconds onward. In the case of highly degraded samples (i.e., 240 seconds of UV exposure), for low threshold values the system detects only one or two actual peaks with a high percentage of false positives (about 65%). Table I reports the

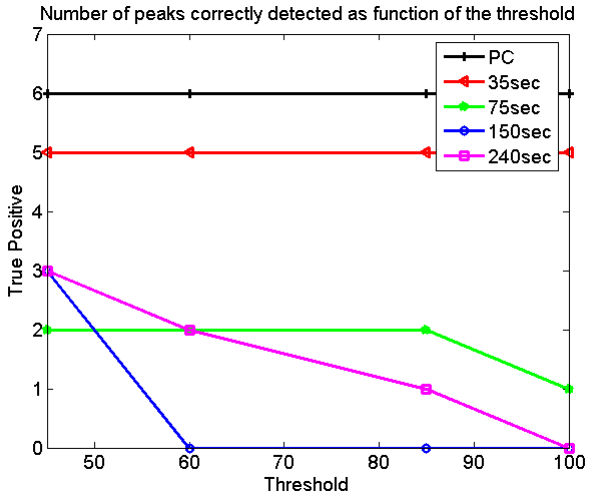

Fig. 12. Number of actual peaks correctly detected by GeneMapper at different threshold values.

set of peaks detected by the proposed algorithm by employing a threshold equal to $0.37 * Max(x(t))$ where $x(t)$ is the degraded data input. Results are reported for each DNA sample starting from the non-degraded one by increasing its level of degradation; peaks are defined by their label, position, height and width. In the presence of highly degraded data, such as those obtained after 150 seconds or 240 seconds of UV exposure, the proposed peak finding algorithm is still able to detect an acceptable number of actual peaks (7 peaks for the sample obtained after an UV exposure of 150 seconds and 3 peaks for the sample obtained after an UV exposure of 240 seconds). For the DNA sample degraded with an UV exposure of 150 secs, the corresponding signal presents peaks lower than 60; subsequently, by employing a fixed threshold greater than 60, those peaks are not detected by GeneMapper, while they are detected by the proposed approach which uses an adaptive threshold that assumes a lower value in the presence of a signal having peaks with lower heights. The signal shown in Fig. 13 represents the Positive Control taken from the second dataset (NIST). The represented sample contains the amount of DNA recommended by Applied Biosystems. Fig. 14 shows

Appeared in Proc. of 4th International Conference on Bioinformatics and Computational Biology (BICoB), (Las Vegas, USA), March 2012.

TABLE I
DETECTED PEAKS FOR DYE 1 BY RUNNING THE PROPOSED ALGORITHM

Sample	Peaks	Position	Height	Width
PC	1	4355.6	3157.0	10.9
	2	5353.3	2505.8	10.283
	3	6331.2	1004.8	12.031
	4	6393.4	850.8	12.6
	5	7073.7	1546.3	12.2
	6	7195.7	1556.7	12.8
35sec	1	4044.1	242.85	12.0
	2	4105.4	206.9	12.5
	3	5061.4	180.1	12.78
	4	5120.7	125.0	14.0
	5	6770.2	130.5	14.7
75sec	1	4146.7	87.8	15.3
	2	4209.8	94.3	14.7
	3	5196.3	65.2	18.8
	4	5258.4	56.4	19.1
	5	6228	40.7	33.5
150sec	1	4127.4	47.4	16.4
	2	4189.5	61.2	15.5
	3	5158.3	57.0	17.3
	4	5217.9	40,9	20.3
	5	6771.5	32.3	22,5
	6	6881.8	39.0	26.8
	7	9089.8	31.1	20.5
240sec	1	4186.4	305.4	9.5
	2	4250.3	211.1	10.1
	3	5254.2	88.6	11.3

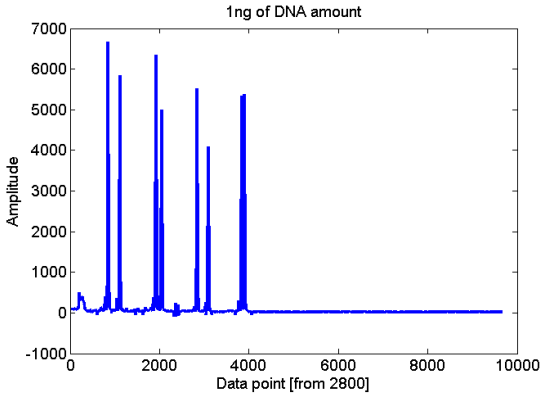

Fig. 13. Positive Control which contains 1 ng of DNA (Blue dye).

an example of low copy number samples with only 10 pg of DNA.

Table II reports results obtained by using the GeneMapper typing system with a threshold value equal to 100. It shows that, the success rate of the typing system decreases when decreasing the DNA amount present in the analyzed samples, leading to 0 detected peaks in the presence of samples containing only 10pg of DNA. This result indicates that the effects of DNA degradation on STR genotyping cannot be overcome by simply adding more DNA template, but they are better addressed by using the recommended 1 ng of DNA template.

Table III reports results of our peak finding algorithm. The amount of DNA factoring the sample presents a non-significant impact on the performance detection. We observed a false positive increase in the presence of the samples with 10 pg

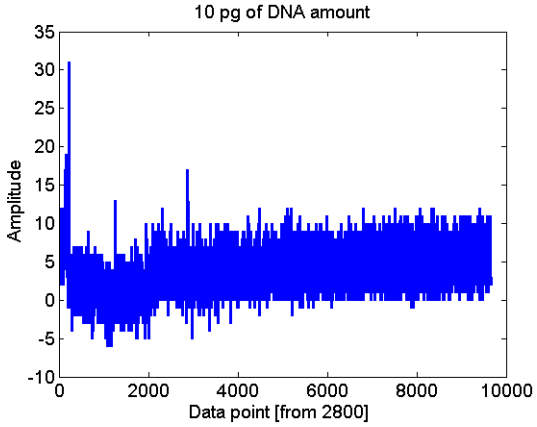

Fig. 14. Signal related to a Sample which contains 10 pg of DNA.

TABLE II
NUMBER OF PEAKS DETECTED BY THE GENEMAPPER SOFTWARE FOR DIFFERENT DYES

Sample	Number of cycles	DNA amount	O dye	B dye	G dye	Y dye	R dye
MT	PC	1 ng	16	18	28	34	11
PT	PC	1 ng	13	14	23	17	10
MT	28	100 pg	12	7	9	6	4
PT	28	100 pg	12	8	10	8	3
MT	28	30 pg	12	0	0	1	1
PT	28	30 pg	12	3	6	5	4
MT	28	10 pg	12	0	0	0	0
PT	28	10 pg	12	0	0	0	0
MT	31	100 pg	13	13	27	24	9
PT	31	100 pg	12	8	10	8	3
MT	31	30 pg	13	8	10	8	6
PT	31	30 pg	12	3	6	5	4
MT	31	10 pg	12	0	0	0	0
PT	31	10 pg	12	0	0	0	0

TABLE III
NUMBER OF PEAKS DETECTED BY THE PROPOSED APPROACH FOR DIFFERENT DYES

Sample	Number of cycles	DNA amount	O dye	B dye	G dye	Y dye	R dye
MT	PC	1 ng	13	8	10	8	6
PT	PC	1 ng	13	8	10	8	6
MT	28	100 pg	12	8	9	6	4
PT	28	100 pg	12	8	10	8	4
MT	28	30 pg	12	9	10	8	8
PT	28	30 pg	12	8	8	7	4
MT	28	10 pg	12	4	5	3	1
PT	28	10 pg	12	4	6	3	2
MT	31	100 pg	13	8	10	8	6
PT	31	100 pg	13	8	10	8	6
MT	31	30 pg	13	9	10	8	6
PT	31	30 pg	13	8	10	7	6
MT	31	10 pg	13	7	10	7	6
PT	31	10 pg	13	8	8	7	5

of DNA; in particular, the number of false positive peaks detected respectively for Blue, Green, Yellow and Red was 12, 3, 10, 13, in case of MT samples, while it was 1, 4, 8, 5 in case of a PT sample. For both datasets, in the presence of high quality samples, the typing system works as well as the GeneMapper software, while in the presence of degraded samples, the proposed algorithm significantly improves the true peak detection rate achieved by the GeneMapper software. The peak finding algorithm has been designed to deal with noisy signals, and the obtained results match our expectations.

V. Conclusions

We presented a method for peak detection when estimating STR alleles in degraded DNA samples where variations in peak height preclude a constant level of loci peak intensities across the entire range of bp-values in STR profiles. The adaptive threshold utilized in our approach and the discrimination power against wider peaks of the proposed algorithm allow for a robust peak detection performance in the presence of low DNA templates. Our experiments show that the success rate achieved by our technique is similar in both scenarios when dealing with high-quality samples which contain the recommended DNA amount (1 ng) and when dealing with critical amount of DNA (less than 100pg). Our experiments also show the robustness of the proposed peak finding algorithm to high level ultraviolet degradation. A limitation of the proposed peak detection algorithm lies in the usage of a global threshold; a future direction to extend this work will focus on designing a local threshold. Further, since the adopted derivative was first-order, we will carry out experiments with higher order derivatives in future work. Finally, we will extend this research by incorporating an additional procedure to automatically adjust parameters (i.e., the amplitude threshold) to process signals representing different DNA samples from different instruments.

VI. Acknowledgement

This work was funded by the Center for Identification Technology Research (CITeR). We would like to thank Raghunandan Pasula, West Virginia University, for his assistance during the development of the project; Prof. Thomas C. O'Haver, University of Maryland for his assistance with our queries regarding peak detection; and National Forensic Science Technology Center (NFSTC) for providing scientific training services. The Peak Finding code developed by Prof. O'Haver was used in this work.

References

[1] C. Easley, J. Karlinsey, J. Bienvenue, L. Legendre, M. Roper, S. Feldman, M. Hughes, E. Hewlett, T. Merkel, J. Ferrance, and J. Landers. A fully integrated microfluidic genetic analysis system with sample-in-answer-out capability. *PNAS*, 103(51), 2006.

[2] K. Horseman, J. Bienvenue, K. Blaiser, and J. Landers. Forensic DNA analysis on microfluidic devices: a review. *Forensic Science International*, 52:784–799, 2007.

[3] B. Pang. One-step generation of degraded DNA by UV irradiation. *Analytical Biochemistry*, pages 163–165, 2007.

[4] J. Butler and C. Hill. Scientific issues with analysis of low amounts of DNA. *http://www.promega.com/resources/articles/profiles-in-dna/2010/scientific-issues-with-analysis-of-low-amounts-of-dna/*, 2010.

[5] J. Butler. *Fundamentals of Forensic DNA Typing*. Elsevier, 2010.

[6] P. Wentzell and C. Brown. Signal processing in analytical chemistry, in encyclopedia of analytical chemistry., 2000. http://myweb.dal.ca/pdwentze/papers/c2.pdf.

[7] D. Skoog and D. West. *Principles of Instrumental Analysis*. Third Edition, Saunders, 1984.

[8] H. Malmstadt, C. Enke, S. Crouch, and G. Horlick. *Electronic Measurements for Scientists*. W. A. Benjamin, Menlo Park, 1974.

[9] S. Smith. The scientist and engineer's guide to digital signal processing. http://terpconnect.umd.edu/ toh/spectrum/Differentiation.html.

[10] A. Felinger. *Principles of Instrumental Analysis. Data Analysis and Signal Processing in Chromatography*. Elsevier Science, 1998.

[11] Applied Biosystems. Quantifiler Kits User's Manual.