Analysis of PCR-RFLP Gel Electrophoresis Images for Accurate and Automated HPV Typing

Christos F. Maramis, Anastasios N. Delopoulos, and Alexandros F. Lambropoulos

Abstract— **The identification of the types of the human papilomavirus (HPV) that have infected a woman provides valuable information as regards to her risk for developing cervical cancer. HPV typing is often performed by means of manually analyzing PCR-RFLP gel electrophoresis images. However, the typing procedure that is currently employed suffers from unsatisfactory accuracy and high time consumption. In order to treat these problems we propose a novel approach to HPV typing that automates the analysis of the electrophoretic images and concurrently improves the accuracy of the typing decision. The proposed methodology contributes both to the extraction of information from the images through a novel modeling approach and also to the process of making a typing decision based on the above information by the introduction of an original HPV typing algorithm. The efficiency of our approach is demonstrated with the help of a complex worked example that involves multiple HPV infections.**

I. INTRODUCTION

According to recent epidemiological studies, the human papillomavirus (HPV) is the *causal factor* for cervical cancer [1]. Currently, there have been discovered more than 100 HPV types, i.e., variants of HPV characterized by different genotypes. However, not all of them are involved in the development of cervical malignancies; instead the HPV types are classified by virologists into four categories in terms of their associated risk for malignant progression: high-risk types, low-risk types, potentially-high-risk types, and undetermined-risk types.

When the high frequency of cervical cancer is taken into account – it is one of the leading types of cancer affecting women worldwide [2] – the importance of the above classification of the HPV types is revealed: If the specific HPV types that have infected a woman could be identified by examining a tissue sample, this would provide very valuable prognostic information to her treating doctor. This identification task is indeed achievable on the basis of the genotypic differences among the virus types and is called *HPV genotyping* or *HPV typing*.

Today, there are plenty of molecular biology methods for performing HPV typing: reverse hybridization analysis, microarray platforms, genotype sequencing – just to name a few. Among them, when accuracy and cost-effectiveness are considered, the prevailing method is *PCR-RFLP gel electrophoresis* and this is also reflected in its extensive application in molecular biology laboratories worldwide.

However, even though the above biological method outperforms its rivalry, the currently employed typing protocol suffers from several shortcomings that are mostly introduced by its partially "manual" nature. The typing procedure is laborious, time-consuming and error-prone, especially for complex cases of multiple HPV infections. In order to treat these problems, we introduce in this paper the algorithms and computational methods that are necessary to automate and make more accurate the protocol of HPV typing via PCR-RFLP gel electrophoresis.

The rest of the paper is structured as follows: In the next section the present HPV typing protocol is described along with its associated problems that we are going to treat. Then, the related work is cited and the proposed approach to the problem of HPV typing is outlined. In Sect. III we describe the extraction of fragment information from PCR-RFLP gel electrophoresis images based on a novel model of the fragment motion. In Sect. IV a novel algorithm for HPV typing from the previously extracted information is introduced. Next, we present a complete worked example of the application of the proposed approach on a complex case that involves multiple HPV infections in Sect. V. Finally, Sect. VI includes the conclusions of this paper and the work to be done in the future.

II. HPV TYPING VIA PCR-RFLP GEL ELECTROPHORESIS

A. The Present In Vitro HPV Typing Protocol

The protocol of HPV typing with the discussed method involves the conduction of a biological examination and the subsequent analysis of the resulting data by an expert molecular biologist – usually with the help of appropriate software. The aforementioned examination includes the sequential application of three well-established molecular biology techniques on biological samples collected from the tissue where the hypothesized[1] virus is supposed to reside, i.e., cervical samples. These techniques are the Polymerase Chain Reaction (PCR) amplification, the Restriction Fragment Length Polymorphism (RFLP) digestion, and one-dimensional gel electrophoresis. The result of this procedure is captured in a digital image which provides the required input to the expert biologist.

Manuscript received July 12, 2010.

C. Maramis and A. Delopoulos are with the Information Processing Laboratory, Department of Electrical and Computer Engineering, School of Engineering, Aristotle University of Thessaloniki, Thessaloniki, 54124 GREECE chmaramis@mug.ee.auth.gr, adelo@eng.auth.gr

A. Lambropoulos is with the Laboratory of Molecular Biology, First Department of Obstetrics and Gynecology, General Regional Hospital Papageorgiou, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, 54124 GREECE lambrop@auth.gr

[1]We use the word *hypothesized* because the sample can be HPV-free. The future use of the word aims to remind the reader of this possibility.

978-1-4244-6561-3/10/$26.00 ©2010 IEEE

When the cervical sample has been collected, PCR is employed to amplify – up to six orders of magnitude – a highly reserved region of the hypothesized HPV DNA. This region is a segment of the L1 gene of HPV. After that, the amplified DNA is digested by a predefined *restriction enzyme* resulting to a set of DNA fragments of various lengths.

Let us elaborate more on RFLP which is the cornerstone of the discussed HPV typing method. A restriction enzyme is an oligonucleotide which digests a genetic macromolecule at positions that are supplementary to the enzyme's nucleotide sequence. Since the nucleotide sequence for the amplified region of each HPV type is known *in advance*, the outcome of the digestion of the genetic material of each type by a specific restriction enzyme is a set of fragments whose lengths in base pairs (bp) are also known in advance. Thus, given a restriction enzyme, each HPV type is characterized by a predefined *fragment lengths pattern* (FLP).

Following RFLP digestion, solutions containing the digested hypothesized DNA from different samples are marked with a fluorescent dye and injected into separate wells at the front end of a *gel matrix*. Then, in the presence of an electric field, the molecules of various sizes are forced to move with different mobilities in a direction parallel to the electric field (*electrophoresis direction* from now on).

During this one-dimensional gel electrophoresis, the longer (i.e., larger) molecules remain close to the well while the more agile shorter molecules cover a much larger distance. This way, a number of *lanes*, starting from each well, are formed that contain blobs of DNA of the same length shaped as *bands* in the direction perpendicular to the electric field (*perpendicular direction* from now on). One or more wells are reserved to include material of known length (usually fragments constantly increasing by 20, 50, or 100 bp). These wells serve as reference *ladders* that help the expert biologists estimate the unknown length of the DNA forming the bands of the other lanes.

After the completion of gel electrophoresis, a digitized image of the electrophorized gel is acquired by an appropriate digital camera in order to obtain a permanent record of the resulting gel matrix. Fig. 1 depicts such an image emphasizing the concepts of *lane*, *band* and *ladder*.

The next step is the analysis of the acquired image by an expert biologist. The cornerstone of the analysis is the fact that the intensity of the image at some position can be related to the viral DNA concentration (viral load) at the corresponding position on the gel matrix. First, the biologist estimates – mostly with the help of specialized software packages – the fragment lengths that correspond to the observed bands on a lane of interest. This is achieved with the help of a virtual marker associating positions along the electrophoresis direction to fragment lengths, which is constructed from the corresponding information from the bands of the image's ladder(s). Finally, the biologist "manually" compares the set of estimated fragment lengths from the lane of interest with the FLPs of all the HPV types in order to judge which type or combination of types could have produced the observed pattern of fragment lengths (HPV typing decision).

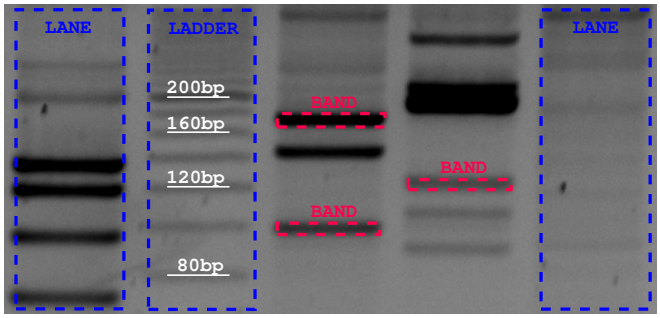

Fig. 1. Typical image of a gel matrix after one-dimensional electrophoresis. Samples of *lanes*, *bands*, and *ladder* are enclosed in rectangles.

B. Problem Statement and Intended Contribution

The ultimate objective of an HPV typing protocol is to *correctly* identify the virus type or combination of types that have infected a patient, with the *least possible effort* through an *objective* procedure. However, the current HPV typing protocol fails to fully satisfy this objective.

Regarding the correctness of the biologist's decision (the most critical requirement), there are plenty of cases where this goal is not fulfilled. The first reason for this is the *unsatisfactory accuracy* in the estimation of the fragment lengths that are found on a lane, which is triggered by the approximate definition of (i) the band positions, and (ii) the association between pixel positions on the image and corresponding fragment lengths. For instance, how could the biologist make the distinction between the high-risk type T_a and the low-risk type T_b with FLPs $\boldsymbol{l_a} = [x_1, x_2, x_3]$ and $\boldsymbol{l_b} = [x_1, x_2 + 2, x_3 - 2]$ respectively, if the uncertainty in the estimation of the observed fragment lengths is higher than 2 bp? As we will see in Sect. III, this problem can be treated with the appropriate modeling of the fragment motion on the gel matrix.

The second reason that leads to erroneous typing decisions is the *overlapping between combinations of FLPs* when multiple infections are involved. Due to the large number of HPV types and the great genotypic similarity in their L1 gene, the restriction enzymes that are currently in use result in individual FLPs that are sometimes partially overlapping. This allows for cases of multiple infections where the observed set of fragment lengths may have resulted from *more than one* combinations of types. As an example, let us consider three HPV types T_c (low-risk), T_d (high-risk) and T_e (high-risk) with FLPs $\boldsymbol{l_c}$, $\boldsymbol{l_d}$, and $\boldsymbol{l_e}$ respectively and let us assume that $\boldsymbol{l_c} = \boldsymbol{l_d} \cup \boldsymbol{l_e}$. In the case where the pattern $\boldsymbol{l_c}$ is observed, what would the typing decision of the previous protocol be? Fortunately, a large portion of real-life multiple infection cases can be resolved if extra information extracted from the lane's image is exploited. The proposed approach takes advantage of this additional information to treat such cases (see Sect. IV).

With respect to the problems of the *objectivity* and *laboriousness* of the typing protocol, the proposed approach constitutes a considerable improvement. Owing to the introduced algorithms and computational methods, the analysis

of the image can be performed automatically in a very short period of time.

C. Related Work

Regarding the research on HPV typing via PCR-RFLP gel electrophoresis, we can distinguish two types of contributions. On one hand, there are scientific and technical efforts in the form of general purpose[2] software applications that deal with the analysis of electrophoretic images: TotalLab, GelComparII, Gel-Pro Analyzer – just to name a few. However, the final goal of all these applications is the estimation of the fragment lengths that correspond to the observed bands of a lane, a task that is sometimes performed without the desired accuracy. Still, the biologist is not assisted in taking the final typing decision.

On the other hand, there are contributions that propose methodologies or/and algorithms to deal with the decision making part when the PCR-RFLP gel electrophoresis method is employed. However, these approaches either require the use of considerably more expensive in vitro protocols than the one we described in Sect. II-A or do not target the problems described in Sect. II-B. Regarding the former case, in [3] an efficient HPV typing algorithm which requires the independent digestion of the PCR product by four discrete restriction enzymes is presented. Another example is [4] where a strategy for HPV typing which involves type-specific PCR analysis is described. With respect to the latter case, [5] targets the case of single HPV infections by searching for a single restriction enzyme to discriminate each HPV type from every other type, not dealing with the problem of multiple infections.

D. Outline of the Proposed Methodology

The goal of our work is to describe in a systemic and algorithmic manner a methodology that can be employed to evaluate the combination(s) of HPV types that have infected a patient by analyzing the digitized image resulting from the PCR-RFLP gel electrophoresis examination of her cervical sample. When viewed as a system, our proposition is outlined by the block diagram of Fig. 2.

The proposed HPV typing approach includes two phases. In the first phase, two properties of the viral DNA fragments that are present on a lane are extracted (*fragment properties*); these are the length and also the concentration of the fragments (Sect. III). In order to achieve the first phase's aim, the gel matrix images are subjected to a series of image processing procedures to detect the lanes' position, subtract the background intensity, and extract, for each lane, an one-dimensional intensity curve that aggregates the intensity information of the lane's image in the perpendicular direction, namely the lane's *intensity profile*.

Then, the extracted intensity profiles are modeled by an appropriate mathematical function. The resulting intensity profile models that belong to ladders are employed to calibrate the motion of the fragments on the gel establishing an association between positions on image and fragment lengths. This association is required in order to extract from the profile model of each lane: (i) the set of fragment lengths, and (ii) the set of corresponding concentrations that characterize the observed bands of viral material.

In the second phase, the fragment properties of each lane are provided to an original HPV typing algorithm that determines the combination(s) of types that best match the input information, i.e., best explain the observed bands (Sect. IV). The algorithm first performs individually for each HPV type a compatibility check based only on the fragment length information in order to eliminate the types whose presence is not justified. The next step jointly investigates the compatible types in order to evaluate each combination's ability to explain both the estimated fragment properties.

III. FRAGMENT INFORMATION EXTRACTION

The evaluation of the set of fragment properties that characterize a lane is performed in four phases, which are described below.

A. Image Preprocessing

The intensity of a typical PCR-RFLP gel electrophoresis image at some position consists of two components: (i) the intensity component that results from the tagging of fluorescent dye to the viral DNA at this position, and (ii) the background intensity component. Since the latter component hinders the efficient extraction of information about the viral concentration, in [6] we have proposed a methodology for dealing with the problem of the background component removal. After applying the image processing algorithms described in that work (listed in Fig. 2), we exploit the one-dimensional nature of the lane's information to extract from each lane its intensity profile by aggregating with the median operator the intensities of the lane's image in the perpendicular direction.

B. Intensity Profile Model Fitting

The motion mechanism of molecules under gel electrophoresis is an inherently stochastic process, despite the deterministic treatment of the phenomenon by both past and modern related approaches. For this reason, we have proposed in [7] the use of the integrated Weibull distribution to express the position of the DNA fragments after electrophoresis.

Moreover, it is known that the number of fluorescent molecules that are tagged to a fragment is proportional to its length. Consequently, it is reasonable to assume that the contribution of a single fragment to the observed image intensity at its position is also proportional to its length.

Based on these assumptions, we introduce the following parametric model for the intensity profile of a lane with K bands:

$$m(x; \boldsymbol{A}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{x}) = \sum_{i=1}^{K} A_i \cdot \exp(-\frac{1}{\gamma_i} |\frac{x - x_i}{\beta_i}|^{\gamma_i}), \quad (1)$$

[2] I.e., not necessarily associated with PCR-RFLP.

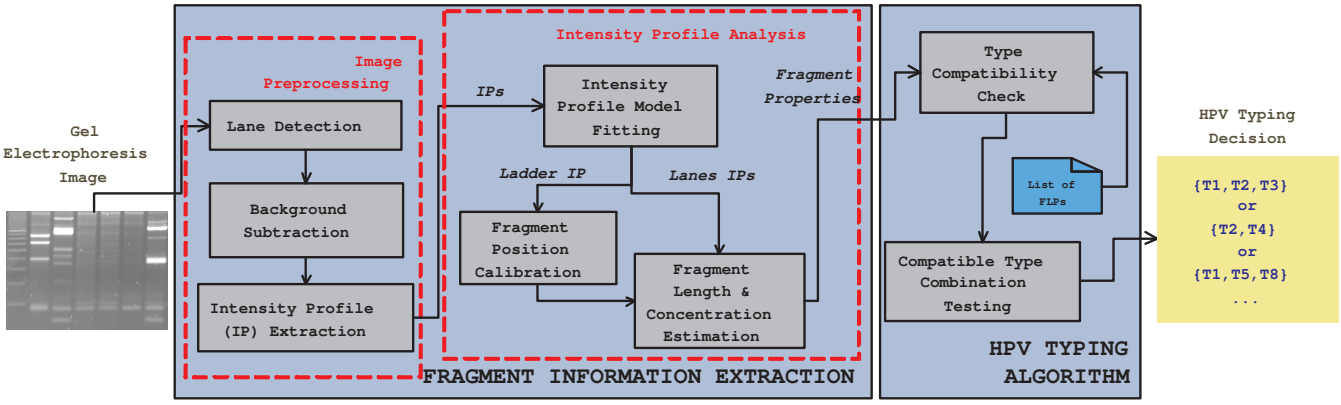

Fig. 2. The block diagram of the proposed HPV typing system.

where x denotes the position in the electrophoresis direction, $\boldsymbol{A} = [A_1, A_2, \ldots, A_K]$ and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{x}$ are defined accordingly. From the previous assumptions, the quantities A_i in the above equation are defined by the formula:

$$A_i = \frac{\gamma_i}{2\gamma_i^{1/\gamma_i}\beta_i\Gamma(1/\gamma_i)} c_i l_i \quad i = 1, \ldots, K. \tag{2}$$

where the fraction corresponds to the normalization factor of the integrated Weibull distribution and l_i, c_i denote the length and the concentration of the fragments forming the i-th band of the profile respectively. $\Gamma(\cdot)$ is the complete gamma function.

The model of (1) is fitted under the least squares optimization criterion to the intensity profile (let this be denoted by $p(x)$ for $x = 1, \ldots, D$) that was extracted from the lane in the previous phase. More formally, if we gather all the model parameters into one vector $\boldsymbol{\rho} = [\boldsymbol{A}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{x}]$, we seek the vector $\boldsymbol{\rho}_{opt}$ such that:

$$\boldsymbol{\rho}_{opt} = \arg\min_{\boldsymbol{\rho}} \sum_{x=1}^{D} \{p(x) - m(x; \boldsymbol{\rho})\}^2. \tag{3}$$

Regarding the determination of K, we obtain an initial estimation of its value by locating the local maxima of the lane's profile with the help of the watershed algorithm [8]. Various values around the initial estimation are tested and the one minimizing the fitting error is selected.

C. Fragment Position Calibration

The objective of this phase is to formally specify the mean position of the DNA fragments as a function of their length, i.e., to associate the x_i parameter of (1) with the fragment length l_i. In order to perform this task, we have experimentally tested all the functions that have been proposed in the related scientific literature. This experimentation resulted in adopting the following formula, which slightly modifies the popular mean position function in [9]:

$$x_i = d(l_i; \boldsymbol{\mu}) = \mu_1 + \mu_2 \log_2(\mu_3 + \mu_4 l + l^2), \tag{4}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2\ , \mu_3, \mu_4]$ and $\mu_2 < 0$.

In order to optimize the parameters of the adopted function, one of the image's ladders is employed. From the fitting process of the previous phase, we have estimated the positions of the bands in the ladder's profile (let these be denoted by x_i^* for $i = 1, \ldots, K$). Moreover, from the specifications of the ladder, the corresponding lengths of its bands are also known; let this be expressed by l_i^* for $i = 1, \ldots, K$. The resulting pairs (l_i^*, x_i^*) are utilized to calibrate the mean position function by optimizing the parameter vector $\boldsymbol{\mu}$ in the least squares sense:

$$\boldsymbol{\mu}_{opt} = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{K} \{x_i^* - d(l_i^*; \boldsymbol{\mu})\}^2. \tag{5}$$

D. Fragment Length and Concentration Estimation

The length of the fragments that form the bands of a lane are estimated from the optimized parameters of (1) with the help of (4). If $d^{(-1)}(\cdot)$ is the inverse of function $d(\cdot)$, then the length of the fragments of the i-th band is given by the formula:

$$l_i = d^{(-1)}(x_i; \boldsymbol{\mu}), \quad i = 1, \ldots, K. \tag{6}$$

When the fragment lengths have been estimated, we are in position to estimate also the fragment concentrations from (2):

$$c_i = \frac{2\gamma_i^{1/\gamma_i}\beta_i\Gamma(1/\gamma_i)}{l_i\gamma_i} A_i, \quad i = 1, \ldots, K. \tag{7}$$

IV. THE VIRUS TYPING ALGORITHM

In order to make an HPV typing decision, each possible combination of types has to be examined for its ability to produce the previously estimated set of fragment properties (*type combination testing*). However, due to the large number of HPV types, the above task is computationally intensive. For this reason, prior to the combination testing, each HPV type is checked individually for its compatibility with the estimated fragment lengths (*type compatibility check*) so as to eliminate the non-compatible types.

A. Type Compatibility Check

For each HPV type a probabilistic *compatibility degree* with respect to the estimated fragment lengths $\boldsymbol{l} = [l_1, l_2, \ldots, l_K]$ is calculated. Let us focus on the HPV type T with FLP $\boldsymbol{p} = [p_1, p_2, \ldots, p_N]$. First, an appropriate assignment function $a : [1, \ldots, N] \rightarrow [1, \ldots, M]$ needs to be specified where $a(i) = j$ means that the j-th estimated length is assigned to – i.e., "it is thought to actually be" – the i-th length of T's FLP. Since we wish to take into account the frequent case of bands that encompass very similar but diverse fragment lengths, $a(\cdot)$ is specified in such a way so as to minimize the quantities $|p_i - l_{a(i)}| \; \forall \, i = 1, \ldots, N$ under the condition that :

$$\nexists \, k, l : \big(a(k) = a(l)\big) \,\&\, \big(|p_k - p_l| > th_1\big), \tag{8}$$

where th_1 is an expert-defined length coincidence threshold.

Let us now define $A(x, y)$ to be the event "the fragment length that is estimated as y is actually x" and let us assume an appropriate function $p_A(x, y)$ that evaluates the probability of the above event. Then, given an assignment $a(\cdot)$, we can define the probability that the type T *could be present* on the lane:

$$Pr\{T; a(\cdot)\} = \prod_{i=1}^{N} p_A(p_i, l_{a(i)}). \tag{9}$$

Finally, the compatibility degree C_T of type T is defined by the equation:

$$C_T = log_2\Big(Pr\{T; a(\cdot)\}\Big) = \sum_{i=1}^{N} log_2\big(p_A(p_i, l_{a(i)})\big) \tag{10}$$

Type T is considered to be compatible with the investigated lane if its compatibility degree is higher than an expert-defined compatibility threshold th_2.

B. Type Combination Testing

Each combination that is formed by the compatible types is tested in terms of its goodness of fit to the estimated fragment properties. For this purpose, under the assumption of a type combination, we search for the concentrations of its component types that best explain the observed bands.

Let us assume a combination of M types T^1 to T^M. With $a^m(\cdot)$ being the previously specified assignment of type T^m, we construct the vector of the type's relative contribution to the observed bands $\boldsymbol{v^m} = [v_1^m, v_2^m, \ldots, v_K^m]$ where for each k : $v_k^m = \|\{i | a^m(i) = k\}\|$. If $\boldsymbol{c} = [c_1, c_2, \ldots, c_K]$ is the vector of estimated fragment concentrations and $\boldsymbol{t} = [t^1, t^2, \ldots, t^M]$ denotes the concentrations of the discussed types prior to RFLP digestion, we employ the quantity ε to evaluate the combination's goodness of fit:

$$\varepsilon = \min_{\boldsymbol{t}} \|\boldsymbol{c} - \sum_{m=1}^{M} t^m \cdot \boldsymbol{v^m}\|, \tag{11}$$

under the constrain that $t^m > 0$ for all $m = 1, \ldots, M$.

Apart from the goodness of fit, one additional objective is considered for the typing decision. This is the prior probability of the combinations. The two objectives are combined in a pareto optimization problem and the resulting pareto front is the final HPV typing decision.

V. A WORKED EXAMPLE

In this section we present a worked example of applying the proposed algorithms on a complex case where multiple HPV infections are involved. The investigated cervical sample was collected, processed (as described in [5]) and photographed by the Laboratory of Molecular Biology of the General Hospital Papageorgiou in Thessaloniki. The resulting lane is shown in Fig. 3(A) and the band-depicting points of its intensity profile – only these are employed for the model fitting – are given in Fig. 3(B).

The initial estimation of K by the watershed algorithm is 8. However, the split of the forth band into two partially overlapping bands improves the modeling accuracy and thus $K = 9$ is selected. The estimated fragment lengths and concentrations that result from the fitting of the intensity profile model are given in Table I.

Assuming a normal distribution $\mathcal{N}(0, 1.57)$ for the probabilities $p_A(\cdot, \cdot)$ and a compatibility threshold of 0.1, the compatible HPV types are 68, 66, 74 and 74a (an alternative genotype of 74), where we have adopted the notation in [5]; their FLPs[3] in bp are $[37, 97, 147, 174]$, $[58, 66, 72, 91, 114]$, $[61, 73, 147, 174]$, and $[118, 147, 174]$ respectively.

The circumstantial phenomenon of partial digestion produces altered FLPs for the compatible types (see [5]) which are also considered when examining the type combinations. In addition, the *primers dimers* – a visible byproduct of PCR – produce on every lane a single length FLP which is taken into account. For the investigated lane, the goodness of fit ε of the 15 possible combinations belong in the range $[0.051, 208.210]$. The solution in terms of pareto optimality is the type combination $\{66, 68, 74a\}$ with $\varepsilon = 0.051$. This is the typing algorithm's decision, made in no more than a few seconds.

The solution includes 4 FLPs: the FLPs of the three participating types plus $[58, 72, 91, 94, 114]$ which is a partial digestion instance of type 66. The concentrations and the assignments $a(\cdot)$ that have been calculated for these FLPs by setting $th_1 = 4$ are given in Table II. The model of the lane's intensity profile that is reconstructed from the evaluated FLP's concentrations is given in Fig. 3(D) and the contribution of each FLP to it is visualized in Fig. 3(C).

The proposed solution for the investigated lane subsumes the typing decision that has been provided by the expert biologists, which is the combination $\{66, 68\}$.

VI. CONCLUSIONS AND FUTURE WORK

The proposed analysis of PCR-RFLP gel electrophoresis images for the purpose of HPV typing has been currently tested on a limited number of lanes with very satisfactory

[3] Actually, the refer to the FLPs that result after discarding the lengths that are too small to be detected.

TABLE II

THE ESTIMATED FRAGMENT LENGTHS AND CONCENTRATIONS.

lengths	173.63	147.72	113.24	93.85	92.86	73.68	66.58	59.50	35.12
concentrations	6.297	6.027	7.736	4.477	7.571	4.928	1.217	4.804	19.959

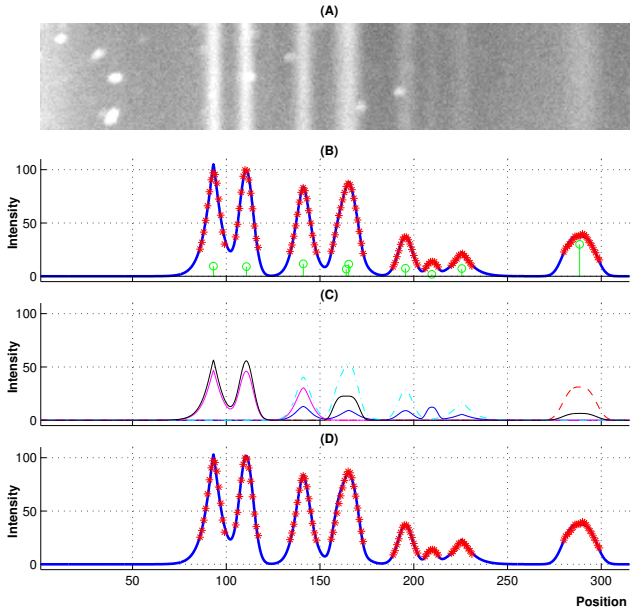

Fig. 3. (A) The image of the investigated lane. (B) The band-depicting points of the extracted intensity profile (red/stars) and the fitted profile model (blue/solid line). The estimated fragment concentrations are depicted in green/circles. (C) The contribution of the FLPs to the lane's intensity profile for the optimal combination of types. The contribution of primers dimers is depicted in red/dashed line. (D) The reconstructed intensity profile model from the optimized contributions of the FLPs of the aforementioned combination. The difference from the model in (B) can hardly be observed.

TABLE I

THE ASSIGNMENTS $a(\cdot)$ AND RELATIVE CONCENTRATIONS THAT ARE CALCULATED FOR THE FLPS OF THE OPTIMAL TYPE COMBINATION.

source	FLP (bp)	assignment	concentration
type 68	$[37, 97, 147, 174]$	$[9, 4, 2, 1]$	3.383
type 66	$[58, 66, 72, 91, 114]$	$[8, 7, 6, 5, 3]$	1.187
type 74a	$[118, 147, 174]$	$[3, 2, 1]$	2.795
part. digestion	$[58, 72, 91, 94, 114]$	$[8, 6, 5, 4, 3]$	3.724
primers dimers	$[37]$	$[9]$	16.577

results. With respect to accuracy, in all the tested cases, the typing decision according to our analysis was identical to or subsumed the typing decision that was made by the expert biologists. In terms of speed, the extraction of a typing decision with the proposed approach was a matter of few seconds leading to a considerable speedup in comparison to the currently used manual procedure. These preliminary experiments provide an encouraging indication in favor of the usefulness of our approach. However, we aim at proving the value of our proposition by conducting HPV typing experiments on a statistically significant number of lanes.

Setting aside the need for more experiments, there might be room for further improvement in the proposed approach. More specifically, the relation of integrated Weibull parameters β and γ with the fragment length needs to be investigated in order to impose additional constrains on the intensity profile model. The use of such constrains will lead to a more realistic and thus accurate modeling of intensity profiles. Moreover, an appropriate distribution for the probability $p_A(\cdot, \cdot)$ needs to be formally defined. This will make more sound the employed type compatibility check. Finally, in the same direction, the optimization of the FLPs' concentrations during the type combination testing could be performed with a fitting procedure on the primary data, i.e., the intensity profile itself and not the estimated fragment concentrations. This modification could contribute both to the accuracy and soundness of the proposed analysis.

In addition to the aforementioned improvements, we intend to incorporate the presented algorithms and methods into our research software application, namely HPV-Typer [10], that aims to assist biologists in the task of HPV typing by PCR-RFLP gel electrophoresis.

REFERENCES

[1] F. Bosch, A. Lorincz, N. Munoz, C. Meijer, and K. Shah, "The causal relation between human papillomavirus and cervical cancer," *Journal of Clinical Pathology*, vol. 55, no. 4, p. 244, 2002.

[2] S. Landis, T. Murray, S. Bolden, and P. Wingo, "Cancer statistics, 1999." *CA: A Cancer Journal for Clinicians*, vol. 49, no. 1, p. 8.

[3] R. Nobre, L. Almeida, and T. Martins, "Complete genotyping of mucosal human papillomavirus using a restriction fragment length polymorphism analysis and an original typing algorithm," *Journal of clinical virology*, vol. 42, no. 1, pp. 13–21, 2008.

[4] V. Fontaine, C. Mascaux, C. Weyn, A. Bernis, N. Celio, P. Lefevre, L. Kaufman, and C. Garbar, "Evaluation of combined general primer-mediated PCR sequencing and type-specific PCR strategies for determination of human papillomavirus genotypes in cervical cell specimens," *Journal of clinical microbiology*, vol. 45, no. 3, p. 928, 2007.

[5] E. Santiago, L. Camacho, M. Junquera, and F. Vázquez, "Full HPV typing by a single restriction enzyme," *Journal of clinical virology*, vol. 37, no. 1, pp. 38–46, 2006.

[6] C. Maramis and A. Delopoulos, "Efficient Quantitative Information Extraction from PCR-RFLP Gel Electrophoresis Images," in *20th International Conference on Pattern Recognition, ICPR 2010*, Istanbul, Turkey, 2010, pp. 2564–2567.

[7] ——, "Improved Modeling of Lane Intensity Profiles on Gel Electrophoresis Images," in *XII Mediterranean Conference on Medical and Biological Engineering and Computing, MEDICON 2010*, Porto Karas, Chalkidiki, Greece, 2010, pp. 671–674.

[8] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based onimmersion simulations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

[9] H. Schaffer and R. Sederoff, "Improved estimation of DNA fragment lengths from agarose gels," *Analytical Biochemistry*, vol. 115, no. 1, pp. 113–122, 1981.

[10] C. Maramis, E. Minga, and A. Delopoulos, "An Application for Semi-Automatic HPV Typing of PCR-RFLP Images," in *7th International Conference on Image Analysis and Recognition, ICIAR 2010, Part II*, Povoa de Varzim, Portugal, 2010, pp. 173–184.