An Application for Semi-automatic HPV Typing of PCR-RFLP Images

Christos Maramis, Evangelia Minga, and Anastasios Delopoulos

Dept. of Electrical and Computer Engineering,
Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
{chmaramis,evang}@mug.ee.auth.gr, adelo@eng.auth.gr

Abstract. The human papillomavirus, coming in over 100 flavors/types, is the causal factor of cervical cancer. The identification of the types that have infected the cervix of a patient is a very laborious yet critical task for molecular biologists that is still performed manually. HPVTyper is a novel research software application that assists biologists by analyzing digitized images of electrophorized gel matrices that contain cervical samples processed by the PCR-RFLP technique in order to semi-automatically identify the existing types of the virus. HPVTyper has been designed to be functional under minimum user input conditions and yet to allow the user to intervene in any step of the typing procedure.

Keywords: HPV typing, gel electrophoresis, PCR-RFLP, biomedical image processing, software application.

1 Introduction

The human papillomavirus (HPV) is a double stranded DNA virus that is responsible for many forms of genital dysplasia and neoplasia [1,2] and is considered to be the causal factor for cervical cancer [3,4]. There have been identified more than 100 types of HPV having similar but slightly altered genotypes; more than 40 of these infect the anogenital tract [5]. However, not all of them are associated with the development of malignancies of the cervix [6]; there are HPV types associated with a high risk of malignant progression (high-risk types), types with a low risk of malignant progression (low-risk types) and types whose associated risk has not been determined yet (undetermined-risk types).

Given the above facts, it becomes evident that the discovery of the identity of the HPV type(s) that have infected a patient is crucial for determining the patient's risk of developing cervical lesions and cancer. This identification process is called HPV typing and remains even nowadays an inherently manual procedure. In this paper we introduce a software application that is intended to help molecular biologists in the task of HPV typing.

HPVTyper is a novel research application that has been developed within the Information Processing Laboratory of the Electrical and Computer Engineering

A. Campilho and M. Kamel (Eds.): ICIAR 2010, Part II, LNCS 6112, pp. 173–184, 2010.

Department of the Aristotle University of Thessaloniki. The application has been designed with the collaboration of and is currently under evaluation by the Molecular Biology Laboratory of the Papageorgiou Hospital of Thessaloniki. HPVTyper attempts to semi-automatically identify the types of HPV that have infected a patient by analyzing the image resulting from the gel electrophoresis of material that has been processed by the PCR-RFLP method (see Sect. 2.1 for an explanation of the molecular biology terms). In this effort, many of the steps are performed automatically while others require input from the biologist. However, the user can intervene at every step in order to adjust the miscomputed parameters of the problem.

The paper is structured as follows: In Sect. 2 we present the molecular biology techniques that comprise the current in vitro HPV typing protocol and also cite the related software applications. In Sect. 3 we describe HPVTyper and its components. Finally, in Sect. 4 we discuss the results of the preliminary use and also possible future improvements of HPVTyper.

2 Background

2.1 HPV Typing

In this section we describe step by step the in vitro protocol that is followed by molecular biologists in order to perform HPV typing on human samples.

First of all, a cervical tissue sample is being collected and is amplified with the use of the polymerase chain reaction (PCR) technique [8] by employing an appropriate set of primers. The reaction increases the concentration of any existing viral DNA molecules up to six orders of magnitude. Afterwards, the amplified material is being digested by a carefully selected restriction enzyme, which cuts the genetic material of HPV at positions of specific DNA base sequence; this is the restriction fragment length polymorphism (RFLP) technique [9], which results, due to genotype differences among HPV types, in a – known in advance – set of fragments of different lengths in base pairs (bp) for each virus type.

The next step in the protocol is gel electrophoresis. Solutions containing the genetic material from different samples are marked with a fluorescent dye and loaded into separate wells at the front end of a gel matrix. Then, in the presence of an electric field, the DNA fragments of various sizes are forced to move with different mobilities in a direction parallel to the field: the fragments of large size remain close to the well, while the more agile smaller fragments cover a much larger distance. This way, a number of *lanes*, starting from each well, are formed that contain blobs of DNA fragments of the same size shaped as *bands* perpendicular to the electric field. 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 *ladders* that help the biologist estimate the unknown lengths associated with the bands of the other lanes.

After the 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. Figure 1 depicts such an image and also emphasizes the concepts of *lane*, *band* and *ladder*. The electrophoresis image is analyzed by the biologist in order to answer to the following questions for each lane/sample: Is the sample infected by HPV? If so, by which types of the virus?

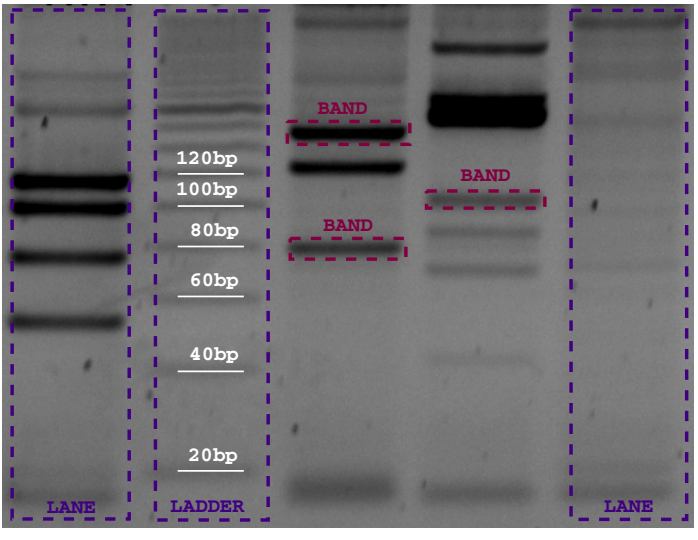

Fig. 1. Typical image of a gel matrix after electrophoresis. Samples of *lanes*, *bands* and *ladder* are enclosed in rectangles.

The first step towards answering the above questions is locating the bands of viral DNA that exist in each lane. Then, the fragment length which corresponds to each band is calculated by comparing its location with the locations of the bands of known length from the ladder(s). This is accomplished through an appropriate interpolation procedure (see Sect. 3.4). The result of this step for each lane is a set of estimated fragment lengths for the viral DNA existing in the sample. At this point, the biologist determines the combination of HPV types which is the most probable to have produced the estimated fragment lengths on each lane, having in mind for each type the set of fragment lengths that result from its digestion by the employed restriction enzyme. This is a tedious and often error-prone procedure.

2.2 Related Work

Gel electrophoresis has been at the forefront of molecular biology for many decades and it remains the most popular technique for separation of macromolecules. Thus, it comes as no surprise that there are plenty of software applications that deal with the processing and analysis of electrophoretic images:

TotalLab Quant [10], GelCompar II [11], Gel-Pro Analyzer [12] – just to name a few. However, all these applications are generic and cannot be employed directly for the typing of HPV. The most a biologist can get out of them is the estimation of the fragment lengths corresponding to the bands of a lane (see the previous section). Still, the actual typing procedure, i.e., the discovery of the combination of types that best explains the estimated lengths has to be performed manually.

On the other hand, there are application-specific programs that analyze electrophoretic images. For example, SAFA [13] which deals with DNA footprinting, GASepo [14] with Epo doping control, GelBandFitter [15] which defines the boundaries of closely spaced bands, etc. However, to the best of our knowledge, there is no software application dealing with HPV typing and this makes HPV-Typer both innovative and useful.

3 System Description

3.1 System Overview

HPVTyper is a standalone software application implemented in C++; for the development of its graphical interface we have employed the cross-platform wxWidgets library [16,17].

HPVTyper handles digital images of electrophorized gel matrices and is parametrized according to the restriction enzyme(s) used by the RFLP technique. The parametrization is accomplished through a configuration file which contains for each HPV type the list of the lengths (in bp) of the DNA fragments that result from the application of RFPL with the employed restriction enzyme(s). Genetic material that have been digested with different restriction enzymes can be analyzed as long as the information of the resulting fragment lengths per virus type for each utilized restriction enzyme is contained in the above configuration file.

Our application consists of three modules. The *Image Processing and Segmentation* module performs all the required image processing operations on the input image and also locates the boundaries of the existing lanes. The *Fragment Mobility Calibration* module deals with the ladder(s) included in the image and employs optimization techniques to map band positions (in pixels) to fragment lengths (in bp). This is achieved by optimally estimating the parameters that determine the mobility of the DNA molecules on the gel from the observed positions of the bands of the ladder(s). The *Band Selection and Type Identification* module performs the actual HPV typing procedure. For each lane, it helps the user select the existing bands of viral DNA and, based on the selected bands, it calculates for each type the probability[1] of the fact that this type is present in the sample loaded on the lane. The system architecture of HPVTyper is given in Fig. 2.

[1] It will be explained later that this is not exactly the probability but a compatibility degree.

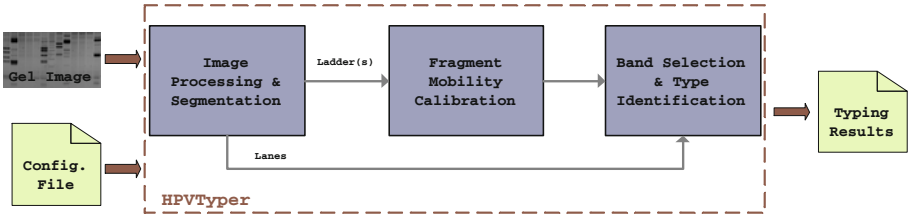

Fig. 2. HPVTyper's system architecture

The application is organized in four tabs that have a serial relation. This means that, in order to perform any action on some tab, the user has to visit the previous tabs and interact with the application so that HPVTyper can set all the prerequisite parameters. Each one of the above modules corresponds to a different tab of the application.

3.2 Image Processing and Segmentation

After loading the digitized image of an electrophoresis experiment into the application, the user can isolate the useful part of it, cropping the blank margins. This part, that contains only the lanes, looks like the one depicted in Fig. 1. Next, the application attempts to correct three types of defects that are apparent on the remaining part of the image. First, the lanes might not be exactly vertical. HPVTyper allows the user to rotate the image by small angles until the lanes are aligned to the vertical axis. Second, there exist dark stains of undetermined shape all over the area of the image due to unavoidable gel impurity. The application tries to eliminate them by applying a 3×3 median filter on the image. Finally, it is often the case that the bands appear dark on lighter background. However, it is visually better for the bands to appear light on darker background. Thus, the image can be subjected to color inversion by a simple click in order to stick to the above color convention.

After the above preprocessing actions, the application is ready to segment the image into lanes, i.e., to attempt to automatically locate the boundaries of the lanes. The only input the user has to provide is the number of the lanes. As the image is now properly oriented, the boundaries are simply vertical lines. The main idea is that, since the lane areas are covered with viral genetic material, each lane area generally appears lighter than the empty gel areas between the lanes. Therefore, we expect high intensity transitions between lanes and background when moving horizontally. This effect is magnified if we consider the entire length of a lane. Thus, the application calculates the discrete intensity derivative in the horizontal direction and sums its value across the vertical direction. The resulting one-dimensional curve has local extrema at the boundaries of the lanes with positive sign at transitions from lane to background area,[2] and

[2] When moving from the left to the right of the image.

with negative sign at the inverse transitions. The extraction of the local extrema is performed by the watershed algorithm [18].

However, in the case of noisy images, the discovered extrema do not always correspond to lane boundaries. To overcome this problem, we employ a second idea: The lanes must by design have similar – if not equal – widths. This means that the distance between the left or right boundaries of two neighboring lanes should be almost constant for all lanes. This "equadistance" property is combined with the located extrema positions so as to conclude to the actual boundaries of the lanes.

When the lane boundaries have been determined, the application displays the results by drawing blue dotted lines on the image at the positions of the left boundaries, and red dotted lines at the positions of the right boundaries. The user is allowed to modify the boundary positions by dragging the dotted lines accordingly. A snapshot of the application's tab which is used to perform the actions described in this section is given in Fig. 3.

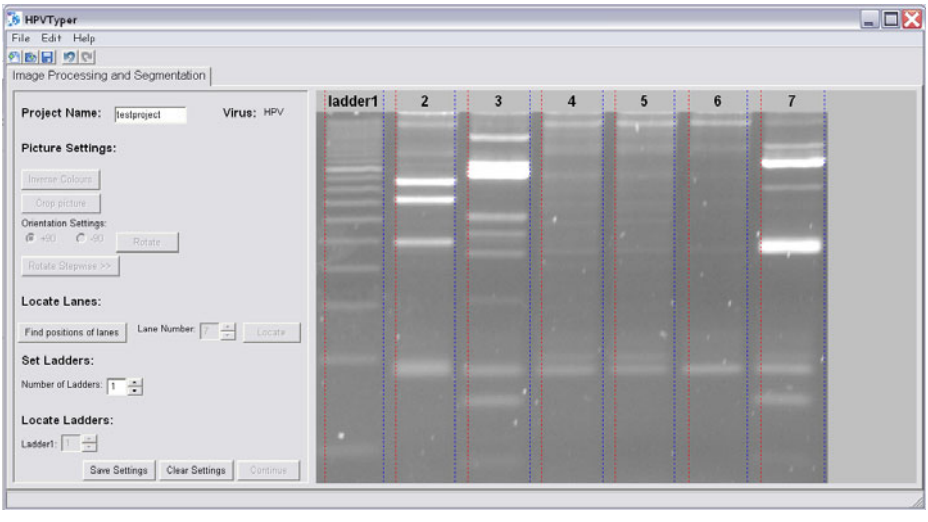

Fig. 3. A snapshot of the Image Processing and Segmentation tab

3.3 Lane Identification

In this tab, the user designates the ladder(s) that exist in the image, thus discriminating them from the other lanes and provides IDs for the lanes that correspond to patient samples. For the subsequent analysis, a ladder is assigned automatically to each lane in order to serve as a ruler of fragment lengths. By default, if there exist more than one ladders, the ladder assigned to a lane is the one closest to it. However, the user can manually alter this assignment.

3.4 Fragment Mobility Calibration

The goal of this module is to perform the mapping of pixel positions on the image to lengths of DNA fragments. This is accomplished by processing the ladder(s) that exist in the image. First, the positions of a number of ladder's bands corresponding to known fragment lengths are located. Then, the extracted pairs of positions on the image and fragment lengths are fitted into a predefined model of DNA mobility on the gel. This analysis is performed individually for each ladder and this also applies to the description that follows.

Before the application takes action, the user has to specify the step of the ladder, i.e., the constant length in bp by which the material loaded in the ladder increases. After that, the average intensity profile along the width of the lane is extracted, and the background intensity is subtracted from it. Next, the application attempts to locate a predefined number of bands on the ladder starting from the band corresponding to the smallest fragment length. These bands are basically local maxima of the extracted one-dimensional profile satisfying the following condition: Between two local maxima corresponding to successive bands the curve must fall below a near-zero intensity threshold. The number of the maxima sought depends on ladder's step. For instance, for a 20 bp ladder the lowermost 10 bands are sought.

The ladder part is detached from the gel image and displayed in horizontal orientation with the estimated positions of the bands indicated as superimposed red lines. The average intensity profile of the ladder with the located maxima is drawn just below the ladder image and serves as a visual aid for the user in case he would like to move some of the band position indicators.

According to [19,20], the theory which best describes the mobility of DNA fragments on gel under electrophoresis is the one claiming that the distance covered by a fragment on the gel is inversely proportional to the logarithm of its length. Hence, if l_i is the length of the DNA fragments forming the i-th of the N bands of a ladder and d_i is the distance they have covered from the start (i.e. the well) of the lane in pixels, then the above statement can be expressed as:

$$d_i = \theta_1 - \theta_2 \log(l_i) \quad \text{for } i = 1, 2, \ldots, N \ . \tag{1}$$

This can be treated as a linear least-squares optimization problem with respect to the unknown parameters θ_1 and θ_2. The extracted set of band positions and their corresponding fragment lengths are used in (1) to estimate θ_1 and θ_2 and when this is accomplished a ruler of fragment lengths is drawn just below the ladder's profile curve. The ruler also depicts the fragment length values that correspond to the bands of the ladder as they are calculated from the estimated mobility parameters.

Since the above estimation procedure is determined by the automatically located band positions on the ladder, it is evidently error-prone. To overcome this problem, the user may alter the band positions, thus invoking a new parameter optimization round as many times as needed. A snapshot of the application's tab in which the actions just described are performed is given in Fig. 4.

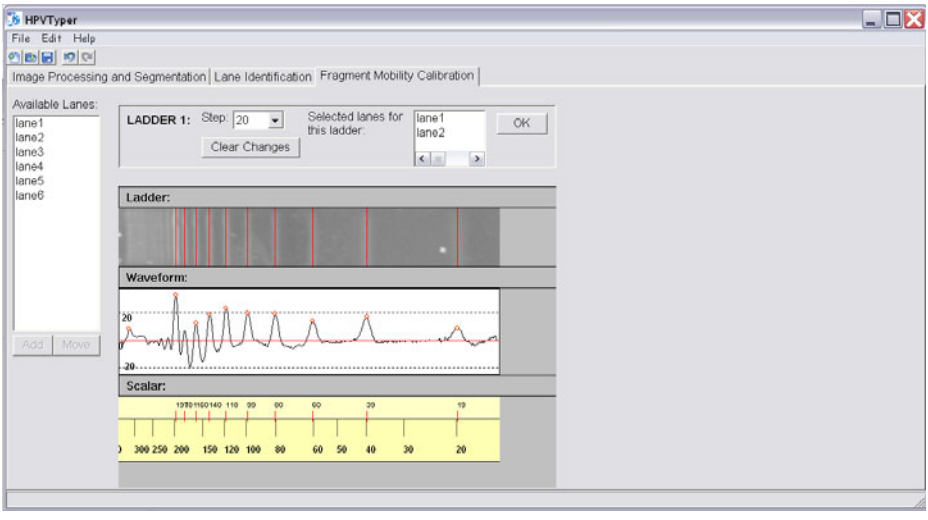

Fig. 4. A snapshot of the Fragment Mobility Calibration tab. This is the case where only one ladder exists in the image.

3.5 Band Selection and Type Identification

The last tab of the application is linked with the *Band Selection and Type Identification* module. Here, the user locates, with the guidance of the application, the bands that exist in each lane and, based on this information, the HPV types that may be present in the sample are identified. This analysis is performed for each lane separately and this also applies to the description that follows. The tab displays from top to bottom:

1. The image of the ladder that is assigned to the lane in horizontal orientation.
2. The image of the selected lane, also in horizontal orientation.
3. The background-free average intensity profile of the lane, which is extracted as explained in the previous subsection.
4. The ruler that has resulted from the estimated mobility parameters of the ladder.

At this point the user has to manually select all the bands that exist in the lane under investigation by clicking on the lane's image. The user selection is marked with a thick red line on the image and with a dotted red line on the profile curve. Moreover, the corresponding fragment length is displayed on the ruler. Although band selection is a manual procedure, HPVTyper assists the user in this task in many ways. First of all, the displayed profile of the lane can be proved very helpful during band selection, especially when the bands are thick or vague. The

same holds for the ruler. Moreover, the application can indicate with the click of a button the expected positions of the bands that correspond to all the possible fragment lengths for all the types of the virus. These virtual bands are displayed as dotted blue lines and can guide the user while selecting the band positions.

After the band selection, the HPV typing algorithm takes over. The algorithm aims to answer to the following question for each type of the virus: Is the existence of this type in the sample compatible with the image of the lane? In other words, could the genetic material of this type have caused *some* of the observed bands on the lane? Thus, a compatibility degree is calculated for each type with the algorithm that is described in the following paragraphs. The degree ranges from 0 to 1, with 0 meaning that the type is completely incompatible with the gel image and 1 meaning that the type is fully compatible with it.

Obviously the fragment lengths that are interpolated for the selected bands are not completely accurate. There are plenty of reasons for that: impurities of the gel, imperfections of the capturing device, misplaced selection of the bands by the user, etc. In order to overcome this problem, the algorithm assigns to each selected band, instead of the corresponding estimated fragment length, a range of lengths centered around it. The width of the range is determined by its center. For instance, for fragment lengths lower than 80 bp, the range spans 2 bp on each side of the center, while for fragment lengths greater than 80 bp, it spans 7 bp on each side. This length-dependent assignment of the range's width makes perfect sense if we consider the motion mechanism of the macromolecules on the gel.

Next, for each HPV type, the application counts how many of its expected fragments lengths after digestion belong in the ranges of the observed bands. Only these lengths that can be interpolated by the discovered ladder bands are considered.[3] For example, since for the case of a 20 bp ladder, the mobility parameters estimation algorithm considers the 10 smaller fragment lengths, i.e., from 20 to 200 bp, all the fragment lengths (both observed and expected) that do not belong in this range are ignored. The compatibility degree of a type is the percentage of the type's expected fragments within the considered length range that belong in the ranges of the observed lengths.

In case no type is found to be compatible with degree higher than 0.7, the application suggests that the sample loaded on the lane has no HPV infection. Otherwise, the type(s) that overpass the above threshold appear on the right pane of the tab in order of decreasing compatibility. HPVTyper displays up to 5 HPV type employing a color-coding scheme for their names. Low-risk types are typed in blue font, high-risk types in red font and undetermined-risk types in green font. A report containing both the intermediate and the final (typing) results can be stored in a human-readable format and reloaded by the application at a future time. A snapshot of the application's last tab is given in Fig. 5.

[3] Forbidding length extrapolation is the common practice among molecular biologists.

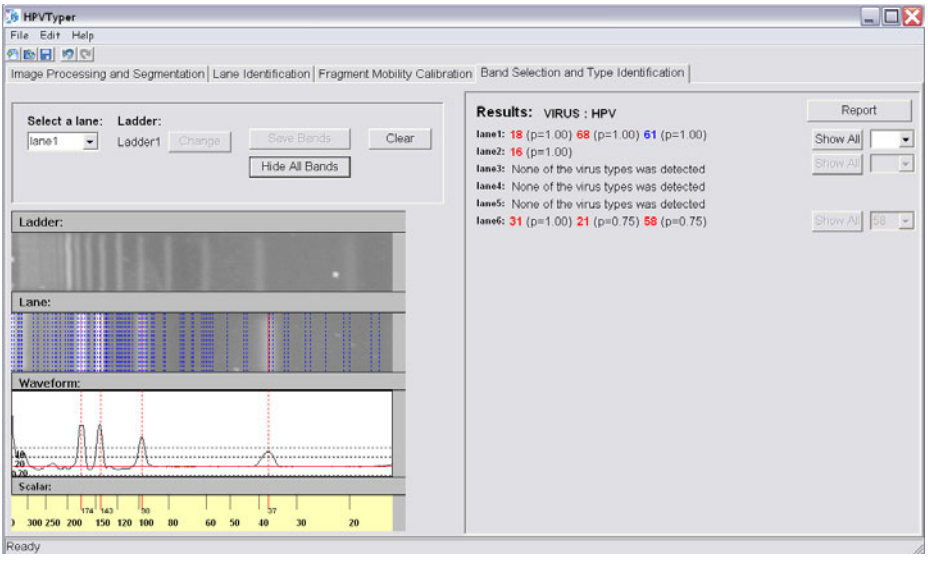

Fig. 5. A snapshot of the Band Selection and Type Identification tab

4 Discussion

HPVtyper was subjected to some early tests in the typing of images of gels that had been produced according to the materials and methods described in [7]. More specifically, cervical tissue samples from 20 individuals were collected, including 4 healthy subjects, 14 single type infections and 2 double type infections. The L1 region of the viral DNA existing in the samples was amplified using MY09/11 (pair of primers) and then was digested by HpyCH4V (restriction enzyme). The material was loaded to non-denaturating polyacrylamide gel for the electrophoresis and each gel matrix included one 20 bp ladder. Only the 41 types and subtypes of HPV given in [7] were considered. Each lane that resulted from the electrophoresis was manually typed by an expert molecular biologist and these were our ground-truth results for the comparison with HPVTyper's outcomes.

The results were very satisfactory. All the types that had been discovered by the expert were also identified by HPVTyper with very high compatibility degrees (ranging from 0.85 to 1). Moreover, all the lanes for which the expert had found no type, were also characterized healthy by the application (i.e. no type had compatibility degree higher than 0.7). At this point, we should mention that there were cases where HPVTyper pointed out as partially compatible types that had not been mentioned by the expert.

We consider HPVTyper as an application that can help molecular biologists in HPV typing as it is now, but also as a basis for a much more powerful application in the future. It is our intention to add new features that will automate some steps of the typing procedure and make more accurate some other steps. First of

all, more efficient strategies for removing the noise from the image and subtracting the background intensity have to be employed during the profile extraction procedure. Moreover, if more than one ladders exist in the same image, the information extracted from both of them should be combined for the estimation of the mobility parameters. This can improve the accuracy of length assignment to bands especially for lanes that lie far from the ladders. Furthermore, the process of locating the bands that belong to the lanes should be automated by fitting the extracted profile to a superposition of properly shaped parametric functions (e.g. superposition of Gaussian or Lorentzian functions). Finally, we have to employ typing algorithms that are sophisticated enough to actually combine the types of the virus in order to explain the observed bands on a lane and not just deal with each type separately. Such algorithms could possibly be based on the use of more than one restriction enzymes [21].

References

1. Baseman, J.G., Koutsky, L.A.: The epidemiology of human papillomavirus infections. J. Clin. Virol. 32, 16–24 (2005)
2. Wang, S.S., Hildesheim, A.: Viral and host factors in human papillomavirus persistence and progression. J. Natl. Cancer Inst. Monogr. 31, 35–40 (2003)
3. Bosch, F.X., Lorincz, A., Muñoz, N., Meijer, C.J., Shah, K.V.: The causal relation between human papillomavirus and cervical cancer. J. Clin. Pathol. 55(4), 244–265 (2002)
4. Walboomers, J.M., Jacobs, M.V., Manos, M.M., Bosch, F.X., Kummer, J.A., et al.: Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J. Pathol. 189(1), 12–19 (1999)
5. de Villiers, E.M., Fauquet, C., Broker, T.R., Bernard, H.U., zur Hausen, H.: Classification of papillomaviruses. Virology 324(1), 17–27 (2004)
6. Muñoz, N., Bosch, F.X., de Sanjosé, S., Herrero, R., et al.: Epidemiologic classification of human papillomavirus types associated with cervical cancer. N. Engl. J. Med. 348, 518–527 (2003)
7. Santiago, E., Camacho, L., Junquera, M.L., Vázquez, F.: Full HPV typing by a single restriction enzyme. J. Clin. Virol. 37(1), 38–46 (2006)
8. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., Erlich, H.: Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb. Symp. Quant. Biol. 51(1), 263–273 (1986)
9. Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., Arnheim, N.: Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230(4732), 1350–1354 (1985)
10. TotalLab Quant, `http://www.totallab.com/products/totallabquant`
11. GelCompar II, `http://www.applied-maths.com/gelcompar/gelcompar.htm`
12. Gel-Pro Analyzer, `http://www.mediacy.com/index.aspx?page=GelPro`
13. Das, R., Laederach, A., Pearlman, S.M., Herschlag, D., Altman, R.B.: SAFA: Semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. RNA 11(3), 344–354 (2005)
14. Bajla, I., Holländer, I., Minichmayr, M., Gmeiner, G., Reichel, C.: GASepo–a software solution for quantitative analysis of digital images in Epo doping control. Comput. Methods Programs Biomed. 80(3), 246–270 (2005)

15. Mitov, M.I., Greaser, M.L., Campbell, K.S.: GelBandFitter–A computer program for analysis of closely spaced electrophoretic and immunoblotted bands. Electrophoresis 30(5), 848–851 (2009)
16. The wxWidgets Cross-Platform GUI Library, http://www.wxwidgets.org/
17. Smart, J., Hock, K., Csomor, S.: Cross-Platform GUI Programming with wxWidgets. Prentice Hall PTR, Upper Saddle River (2005)
18. Vincent, L., Soille, P.: Watersheds in digital spaces: An efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13(6), 583–598 (1991)
19. Southern, E.M.: Measurement of DNA length by gel electrophoresis. Anal. Biochem. 100(2), 319–323 (1979)
20. Schaffer, H.E., Sederoff, R.R.: Improved estimation of DNA fragment lengths from agarose gels. Anal. Biochem. 115(1), 113–122 (1981)
21. Nobre, R.J., Almeida, L.P., Martins, T.C.: Complete genotyping of mucosal human papillomavirus using a restriction fragment length polymorphism analysis and an original typing algorithm. J. Clin. Virol. 42(1), 13–21 (2008)