Novel algorithms implemented in the gel image analysis system GAS2

Ivan Bajla, Igor Holländer, Matej Kollár

ARC Seibersdorf research GmbH, A-2444 Seibersdorf, Austria

E-mail: ivan.bajla@arcs.ac.at

Abstract

A novel implementation of the second generation Gel Analysis System (GAS) with several improvements is presented. The main novelties are: a new module for correction of the smoothing image artifacts using the notch filter in Fourier domain, a new module of the image rectification which speeds up the calculations considerably, and a novel band detection operator based on the template approach which solves some crucial cases of band detection.

1. Introduction

In the recent decade a number of software systems of image analysis for electrophoretic gels (e.g., of DNA fragments) have been developed. Most of defects occurring in gel images can be corrected by tools incorporated into these systems. However, some of image characteristics, in particular low contrast and very closely located bands (deposits of the DNA fragments), the mobility distances of which are to be calculated, cause essential problems with proper band detection.

In [1, 2] we presented a novel philosophy of the gel image analysis (GIA) which changes the common processing paradigm and instead of processing 1D cumulated intensity profile of a gel image, it introduces a two-stage two-dimensional image analysis. We illustrated the major advantages of our approach on a pilot software system GAS1, which we developed using the *MATLAB*^{TM} programming environment. The extensive practical testing of the GAS1 by biologists of the Molecular Biology Group, the Biotechnology Unit of the ARC Seibersdorf research, justified the two-stage philosophy proposed, and confirmed the anticipated advantages of the methods proposed. The interactive tools incorporated into the GAS1 have been improved and some add-ons to the system have been developed. The system testing also revealed particular problems associated with the specific type of gel images. Based on the analysis of these problems the following research and development goals were formulated:

- to enable the analysis of silver-stained gel images which exhibit heavy vertical strip-like artifacts
- to considerably speed up the operation of the image rectification aimed at correction of geometrical distortions
- to further improve the performance of band boundary (BB) detection algorithms, thereby to increase the number of automatical proper detections in two crucial cases, low contrast and closely located bands
- to improve the graphical user interface (GUI)
- to ensure practical applicability of the system by re-implementing it in Visual Basic programming language.

In the paper the results of the research and development in these directions are summarized.

2. Preprocessing of gel images by notch filter in the Fourier domain

The silver-stained gel images suffer from a serious degradation characterized by vertical stripes caused by smearing effect in the gels (Fig. 1). As a consequence considerable distortions of individual bands, largely at the location of the lane boundaries, are observed. These distortions make proper detection of band boundaries even more difficult. It is known that periodic nature of image distortions produces bursts of concentrated energy in the horizontal or vertical axis (according to the orientation of the periodicity) of the Fourier spectrum of the image. A basic approach for reducing the effect of such kind of distortions is to use a notch filter $H(u, v)$ which attenuates the values of the Fourier transform in the coordinate axes and multiplies all other values of the transform (Fourier spectral coefficients) by 1. The corrected image $p(x, y)$ of the input corrupted image $g(x, y)$ is then obtained by the inverse Fourier transform

$$p(x, y) = \mathcal{F}^{-1}\{H(u, v) \cdot G(u, v)\} \ ,$$

where $G(u, v)$ is the Fourier transform of the input image $g(x, y)$.

The notch filter is characterized by two parameters: (i)the *width* w_1 of the horizontal (or vertical) rectangular notches, and (ii)the size w_2 of the gap between two notches (Fig. 1, bottom right window). Since optimum values of these parameters for the images of the given class can only be found by computer experiments, we have developed a special program module *fftstudy* in *MATLAB*^{TM}. The layout of this experimental tool is displayed in Fig. 1.

The original image of a silver-stained gel is first converted from the color domain into gray scale. Then a ROI is selected and the clipped image is the input image for the application of the notch filter. It is displayed in the upper left window in Fig. 1. The output filtered image is displayed in the upper right window. The bottom left window in the same Figure serves for visualization of the Fourier (magnitude) spectrum of the input image. In the bottom right window a particular size of the notch filter is displayed. We have accomplished a set of computer experiments with several images selected from the class of

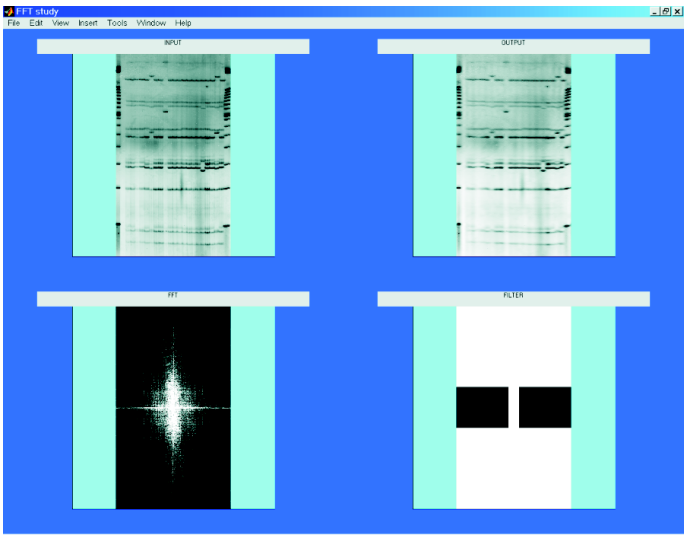

Figure 1: The layout of the experimental program modul *fftstudy* for notch filtering.

silver-stained gel images using various values of two parameters of the filter. The first parameter is responsible for changing the average intensity value and overall level of blurring. The filtering effect (attenuation of the vertical stripes) is significantly affected by the second parameter, the gap between two rectangular notches. We have found the optimum values of the parameters of the notch filter

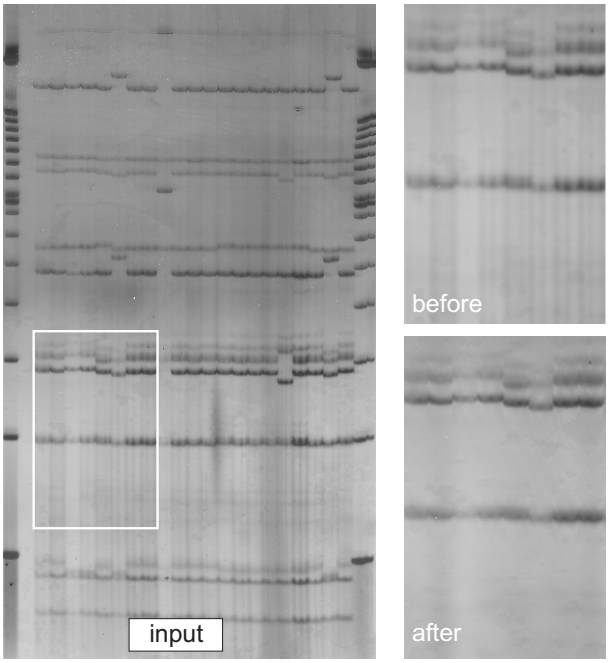

Figure 2: Original input image with two fragments before and after filtering by the notch filter with the optimal parameters $w_1 = 0.1, w_2 = 0.1$.

$w_1 = 0.1, w_2 = 0.1$. In Fig. 2 the corrupted original input image and the result of filtering by the notch filter with the optimum parameter values is displayed.

3. Improvement of gel image rectification algorithms

In most of contemporary GIA software systems geometrical distortions (nonlinear in general) are corrected on the basis of the following paradigm: estimated correction curves are superimposed on the underlying gel image to enable the calculations of the proper quantities of mobility distances; the image itself is not transformed and its distorted appearance is preserved. We chose another approach, namely, the rectification (warping) of the entire image. It means that all subsequent operations, starting with lane separations, are applied to an already corrected image. The rectification algorithm we implemented is based on superpositon of a regular rectangular grid on the input image. The grid consists of interconnected control nodes. As default (Fig. 3), control nodes are arranged in 6 rows by 9 columns (additional rows and columns can be added interactively). The user moves the control nodes to target positions to follow the shape of distortions. The resulting distorted grid has to be transformed to the regular output positions (Fig. 4). The parameters of the geometrical transformation calculated for the grid nods are then applied to the entire image raster and the final corrected image is calculated.

From the mathematical point of view two problems need to be solved: first, the computation of target point coordinates, second, the interpolation of the new intensity in the output (rectified) image. As the rectification module represented the slowest function implemented under *MATLAB*^{TM} in the GAS1, research into faster algorithms and their implementation in the GAS2 was needed. The reduction of the computational cost of the rectification transformation we proposed has been reached by the following three ways: 1)by replacing the *MATLAB*^{TM} functions by a compact program written in Visual Basic, 2)by limiting the computations only to those image areas where the points of the input rectification grid have been adjusted by the user (local corrections), and 3)by simplification of the formula for the

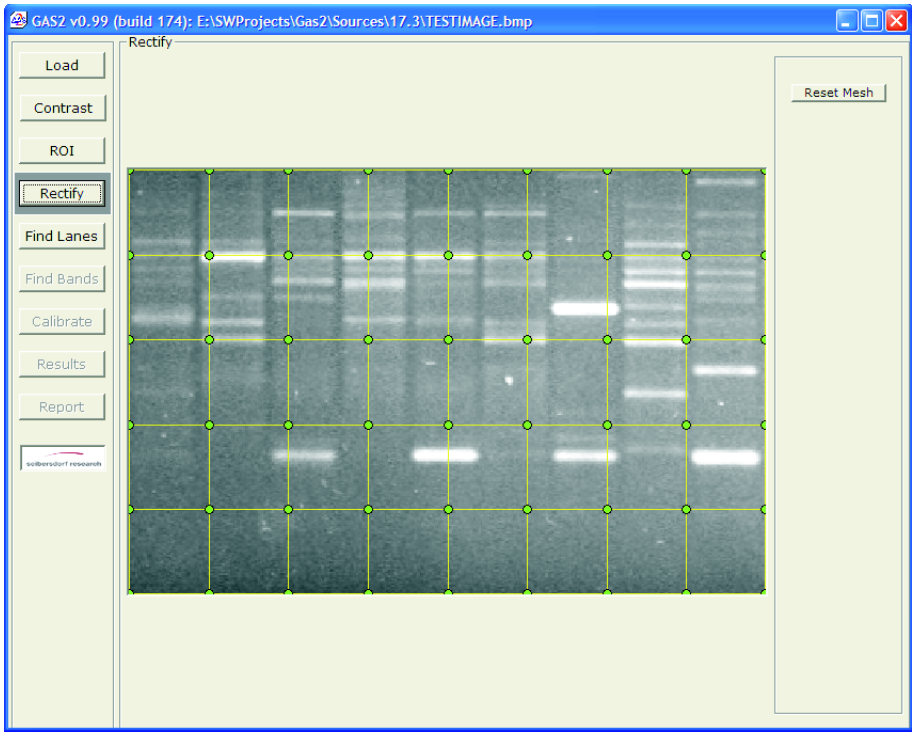

Figure 3: The image rectification window of GAS2.

2-D linear interpolation. In particular, we have tested three interpolation algorithms, namely, nearest neighborhood interpolation, bilinear interpolation and linear interpolation. A special testing application *RectifyDevelope* has been developed for this research ([3]). The new implementation in Visual Basic has shortened the computational time from tens of seconds to seconds, the local application of the rectification transformation (for linear interpolation) reduced the processing time from 62 s to 7 s. The simplification of the interpolation formula yielded 1.5 % saving of the processing time.

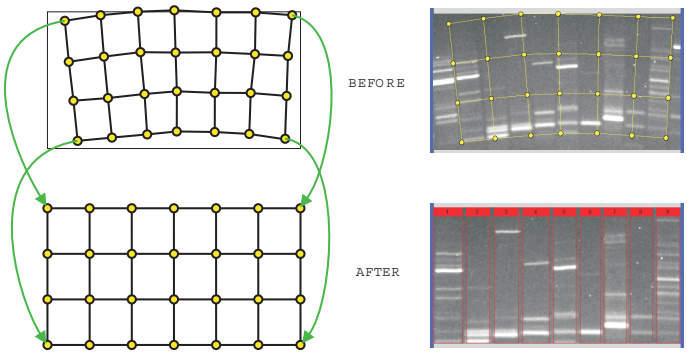

Figure 4: Image rectification principle: the source and target positions of the grid control points (left), and the appearance of the image before and after rectification (right).

4. A novel template approach to band detection

The band detection operation represents the crucial point of the GIA. The implementation of this operation in the GAS2 follows the two-stage philosophy we proposed in [1, 2]. We explored several approaches to detection of band edges (boundaries) using standard 2D edge detectors and other operators which measure the level of intensity nonhomogeneity in rectangular neighborhoods. However, the results achieved by these operators in the preliminary computer experiments were not satisfactory. Since the band boundaries manifest prevailing horizontal orientation, an operator accumulating vertical intensity differences has been tested. It proved to work quite satisfactorily in a number of situations except the gel images containing low-contrast bands and the bands located closely to each other. The further research lead us to the development of a *TEmplate DEtector of Band BoundarY* (*TEDEBBY*).

The new *TEDEBBY* operator is characterized by the following basic steps:

- generation of BB templates (graphical illustration in Fig. 5A)
- calculation of the cumulative difference along each of the given templates and searching for their local maxima (a fragment of an input image is displayed in Fig. 5B and the result of the first step is displayed in Fig. 5C)
- deleting BBs which are not compatible with their neighbors (Fig. 5D)
- deleting BBs for low-contrast neighboring regions (the resulting BBs are displayed in Fig. 5E).

For the description of the main elements of this operator we denote by $[g(i,j)]$ an $m \times n$ pixel matrix representing the image of the lane. Further, we will call (i,j) *the index pair*. The index pair represents *the pixel position*. If we say "pixel (i,j)" we mean the pixel with position (i,j). *The value of the pixel* in position (i,j) is $g(i,j)$. Let $\{(i,j)\}$ be the set of all possible index pairs in the matrix $[g(i,j)]$. We will call *image region* any subset of $\{(i,j)\}$. *The (digital) curve* in the image is also regarded to be an image region.

The possible shapes of BBs are given by *templates*. *The template of type* s for lane width n is a vector $\mathbf{t}^s = (t_1^s, t_2^s, \cdots, t_n^s)$, where t_j^s are small integers. The number t_j^s represents the relative row position of the pixel belonging to the band boundary curve in the column j, with respect to the reference position. It is reasonable to require $t_1^s = 0$ for any s, so that the position of the band boundary curve in the first column is identical to the reference position of the band boundary. In particular, we have two basic types of templates: (i) *slanted line*, and (ii) *arc*. These template types can be further specified by an integer parameter defining their *skewness*. If the skewness is limited by number S, we have in total $4S + 1$ different templates (for both basic types, *skewness* can be $\pm 1, \pm 2, \cdots, \pm S$, and we have one trivial template type for *skewness* $= 0$). The templates are generated by regular repeating of components whose absolute values does not exceed skewness. In Fig. 5 the set of all templates for $S = 3$ is graphically represented by stepwise lines (note the trivial template is displayed twice).

For each template we define *the cumulative vertical difference* $d^s(i)$:

$$d^s(i) = \sum_{j=1}^{n} |g(i + t_j^s, j) - g(i + t_j^s - 1, j)| \quad . \tag{1}$$

This quantity represents the strength of a potential band boundary of the given shape with the reference position in the row i.

For each selected (designed) template $\mathbf{t}^s$ we can characterize the lane $[g(i,j)]$ by means of the column vector $\mathbf{d}^s = (d^s(1), d^s(2), \cdots, d^s(m))'$. We search for all local maxima of the components of the vector

$\mathbf{d}^s$, i.e., all such $d^s(i)$, for some $i \in \{1, 2, \cdots, m\}$ for which $d^s(i-1) < d^s(i) > d^s(i+1)$, since these maxima may indicate band boundaries. Then we arrange these maxima in a matrix $[M_{i,s}]$

$$M_{i,s} = \begin{cases} d^s(i) & \text{if } d^s(i-1) < d^s(i) > d^s(i+1), \\ 0 & \text{otherwise.} \end{cases}$$

In each row i of the matrix $[M_{i,s}]$, we are interested in the template yielding maximum response, as well as in the value of the response itself. Thus, we define the $m-$dimensional column vectors $\mathbf{m}_{\text{init}} = (m_{\text{init}}(1), m_{\text{init}}(2), \cdots, m_{\text{init}}(m))$ and $\mathbf{a}_{\text{init}} = a_{\text{init}}(1), a_{\text{init}}(2), \cdots, a_{\text{init}}(m))$ such that

$$m_{\text{init}}(i) = \max_s(M_{i,s}),$$

$$a_{\text{init}}(i) = \arg\max_s(M_{i,s}).$$

If there are more equal maxima in the row i, we simply take a randomly selected s of one of them for $a_{\text{init}}(i)$. If $m_{\text{init}}(i) = 0$, we define $a_{\text{init}}(i) = 0$.

A nonzero value $a_{\text{init}}(i)$ represents the most probable shape of the band boundary (simply: *the BB candidate*) positioned in the row i. If $a_{\text{init}}(i) = 0$, we do not expect a band boundary in the row i. The corresponding value $m_{\text{init}}(i)$ represents the "strength" (likelihood) of the potential band boundary.

The digital curve representing the band boundary for the given vectors $\mathbf{m}_{\text{init}}$ and $\mathbf{a}_{\text{init}}$ is denoted by $B_{\text{init},k}$. If $1 \le k_1 < k_2 < \cdots < k_p \le m$ are indices of all nonzero components in the vector $\mathbf{m}_{\text{init}}$, then the BB candidates are

$$\begin{aligned} B_{\text{init},1} &= \{(b_1^1, 1), (b_2^1, 2), \ldots, (b_n^1, n)\}, \\ B_{\text{init},2} &= \{(b_1^2, 1), (b_2^2, 2), \ldots, (b_n^2, n)\}, \\ &\vdots \\ B_{\text{init},p} &= \{(b_1^p, 1), (b_2^p, 2), \ldots, (b_n^p, n)\}, \end{aligned}$$

where

$$\begin{aligned} b_1^1 &= k_1 + t_1^{a_{\text{init}}(k_1)}, \\ b_2^1 &= k_1 + t_2^{a_{\text{init}}(k_1)}, \\ &\vdots \\ b_n^1 &= k_1 + t_n^{a_{\text{init}}(k_1)}, \\ &\cdots \\ b_j^k &= k + t_j^{a_{\text{init}}(k)}, \\ &\cdots \\ b_n^p &= k_p + t_n^{a_{\text{init}}(k_p)}. \end{aligned}$$

The adjacent BB candidates $B_{\text{init},k}$ and $B_{\text{init},k+1}$ must not intersect or touch each other. Two adjacent BB candidates satisfying this condition are called *mutually compatible*. Therefore it is necessary to check mutual compatibility of all BB candidates. The goal is to build new vectors $\mathbf{m}_{\text{comp}}$, $\mathbf{a}_{\text{comp}}$, (and thereby new BB candidates) which are guaranteed to be mutually compatible. The new vectors are built from the initial vectors $\mathbf{m}_{\text{init}}$ and $\mathbf{a}_{\text{init}}$ by setting some of the nonzero components to zero (this effectively means deleting some BB candidates). Due to the limited extent of the templates, it is sufficient to check only the compatibility of the BB candidates whose reference positions do not differ by more that $\pm 2S$. If two BB candidates turn out to be mutually incompatible, one of them (preferrably the one which is less likely) has to be deleted. The resulting BBs are illustrated in Fig. 5D.

In Fig. 6 the basic layout of the GAS2 with the Stage I window is displayed. The processing is organized lane by lane. Moving to the arbitrary lane is accomplished by the slider located in the bottom part of the band finding panel (BFP). For semi-interactive work of the user in both band detection stages

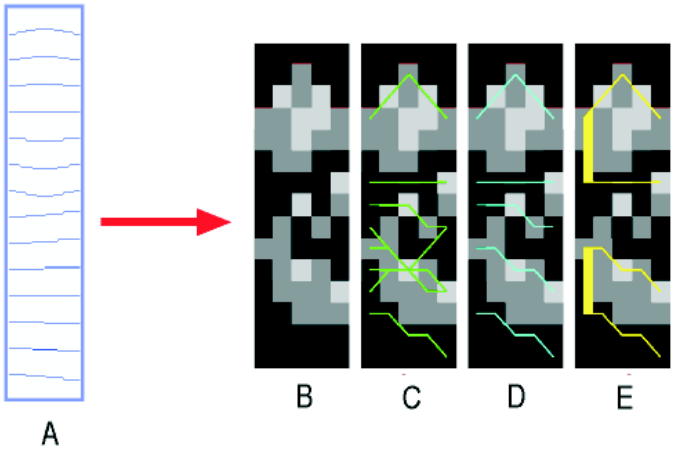

Figure 5: The set of digital geometrical templates of band boundaries (A), and the illustration of the *TEDEBBY* and *TOODIS* operator results (B, C, D, E).

we have designed a triple of the adjacent image (graphical) panes which occupy the main part of the BFP panel. The left pane (LP) contains the original (unfiltered) image of the selected lane. In the **Set** mode, the image is overlaid with markers of the positions of the potential band boudaries. The boundaries are shown only if they are paired (which is marked by a green rectangle connecting the two boundary lines. The middle pane (MP) contains the lane image smoothed by GDD filtering and its primary purpose is to display the lane image without any overlays so the user can examine it in detail with no obstruction. The green line passing through this lane is a marker of the actual local maximum of the boundary detector responses for the given lane. Finally, the third, rightmost pane (RP) of the BFP, contains a plot of the band boundary detector values *TEDEBBY* for the given lane. Only local maxima are displayed (their peaks point to the left). These maxima represent the potential candidates of band boundaries. The pairing of the detected BBs was accomplished in GAS1 by the operator *TOODIS* (*TOOth DIScrimination*). It is based on the idea of measuring the homogeneity of the regions between individual BB candidates by some statistical characteristic and searching for positive teeth in a step-like function of these characteristic. Introduction of the novel *TEDEBBY* operator into GAS2 system was accompanied by modification of the operator *TOODIS*. We describe the main elements of the modification.

If the flexible threshold thr is set to some local maximum of the values of the *TEDEBBY* operator, the components of the vector $\mathbf{m}_{\text{comp}}$ which are smaller or equal to the threshold thr are deleted. The resulting vector $\mathbf{m}_{\text{comp}}^{thr}$ is obtained. The ultimate goal of the *TOODIS* operator is to couple the appropriate BB candidates. It operates on the vectors $\mathbf{m}_{\text{comp}}^{thr}$ generated by the particular values of the threshold thr. For each set $\{B_{\text{comp},k}\}_{thr}$, of the BB candidates, represented by the vector $\mathbf{m}_{\text{comp}}^{thr}$, that corresponds to the threshold thr, a set of image regions R_k, each bounded by two adjacent BB candidates $B_{\text{comp},k}$, $B_{\text{comp},k+1}$, $k \in \{1, 2, \cdots, m-1\}$ can be considered. These regions can again be characterized by appropriate statistical characteristic $\mathbf{f}$ of region intensities $f(R_k)$ and the entire lane image by the vector $\mathbf{f}_{\text{lane}} = (f(R_1), f(R_2), \cdots, f(R_p))$, for $p \leq m-1$. The coupling of neighboring BB candidates is based on the following assumption. Let us denote the three adjacent components (triple) of the vector $\mathbf{f}_{\text{lane}}$: $f(R_{k-1}), f(R_k), f(R_{k+1})$. Provided band image structures are brighter than background (this is the standard situation in gel image analysis) and the relation $f(R_{k-1}) < f(R_k) > f(R_{k+1})$ is true, the central region is assumed to be brighter than two neighboring regions. So, we interpret the region R_k as a band surrounded by background and call the component $f(R_k)$ a *positive tooth*. Again, to avoid influence of neighboring image structures we do not apply this scheme to the regions R_k directly. Instead, we decompose them, similarly as in the previous step, into two approximately symmetrical

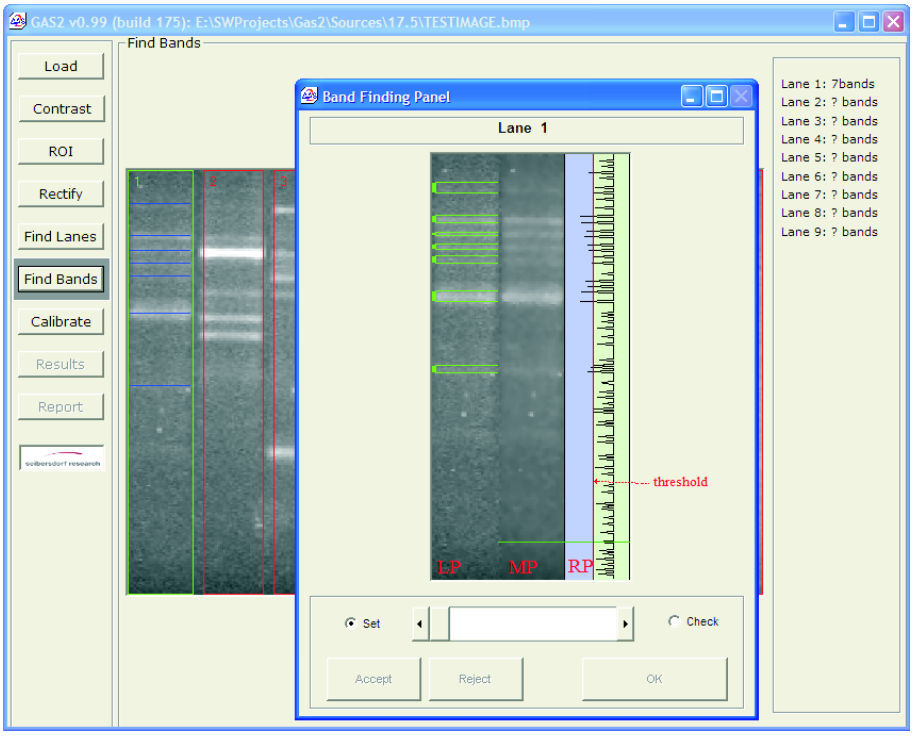

Figure 6: The GAS2 window for band detection.

subregions $S_{k,u}, S_{k,l}$ and select the characteristic of one of these subregions as representative for the whole region R_k. The medians *med*$(S_{k,u})$ *med*$(S_{k,l})$ for these subregions are calculated. We can use minimum or maximum in each median pair as a representative value for the whole region R_k. Since we maximize the detection of true band boundaries (at the expense of additional false detections, which will be ultimately deleted in the Stage II), the following scheme for representative median selection has been proposed. In the following we abbreviate $m_{ku} = med(S_k, u)$ and $m_{kl} = med(S_k, l)$ and we denote $f_{k-1} = \min\{m_{k-1,u}, m_{k-1,l}\}$, $f_k = \max\{m_{ku}, m_{kl}\}$, $f_{k+1} = \min\{m_{k+1,u}, m_{k+1,l}\}$. Then, for all relevant k we

- generate a triple of values (f_{k-1}, f_k, f_{k+1}) for the central position k ;
- check, whether $f_{k-1} < f_k > f_{k+1}$;
- if so, the BB candidate pair $B_{\mathrm{comp},k}$, $B_{\mathrm{comp},k+1}$, corresponds to a positive tooth and the next triple is generated for the central position $k + 2$;
- if not, the BB candidate pair $B_{\mathrm{comp},k}$, $B_{\mathrm{comp},k+1}$ does not correspond to a positive tooth and the next triple is generated for the central position k+1.

The main part of the false detection rejection is done implicitly by the *TOODIS* operator before starting the Stage II. However, due to fluctuations of background intensity homogeneity in lanes (often occuring in imperfect gel images) we have to admit some false detections also in pairs of detections which have been accepted for the Stage II. So, the basic purpose of the BB indicators in this stage is to generate estimates of curvilinear band boundaries that represent improved estimates of the band boundaries in the Stage I.

In [4] we proposed and explored three BB indicators for Stage II. Finally we have implemented into the GAS2 the indicator *MAMBO* (*Median Above Median BelOw*). Briefly, this indicator is based

on the improved curvilinear estimate of BB We compute the intensity median for all the neighborhood pixels above the boundary and intensity median below it. The indicator is then defined simply as the absolute difference of these medians. Based on the values of the *MAMBO* indicator (its values are ordered in the decreasing ordwer of the probability of false detections) the Stage II can be organized as a semi-automatical checking procedure. The detections with lower probability are checked first. The user can decide, if it is necessary to continue the procedure or to accept all remaining detections at once. Furthermore, the improved BB indicators can serve at the same time for more precise calculation of the mobility distances.

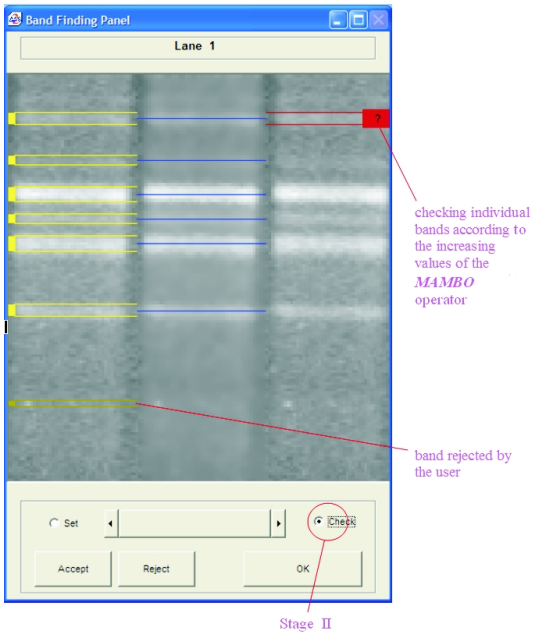

Figure 7: The GAS2 window in Stage II.

The Stage II is implemented in the BFP panel which is switched over to the **Check** mode (see Fig. 7). In this mode the rightmost graphical pane is replaced automatically with the replica of the original lane image. A marker with a question mark overlaid in this pane points to the actually checked band position. The user can reject the detection under question by corresponding button or he/she can accept such a BB detection. The goal of this interactive phase is to finally reject all false BB detections. Usually in case of some false BB detections present, there is a very short way how to reject them, because all detections are ordered according to the specifically designed operators yielding a probability of false detections.

The information acquired during the process of gel mage analysis can be either printed out or recorded in files, which can in turn be used for further work, e.g., as an input for another software package or for documentation. An example of the graphical report is given in Fig. 8.

5. Conclusions

Based on the pilot software system GAS1 further research into improved algorithms of gel image analysis has been carried out. The second generation of this system, GAS2, has been written in Visual Basic and it incorporates a possibility to analyse the silver-stained gel images by making use of specific Fourier filter for correction of smoothing artifacts. The GAS2 comprises a completely new module of image rectification which is considerably faster than the previous one. The original template approach has been implemented into GAS2 which improves the process of band boundary detection. The systematic testing

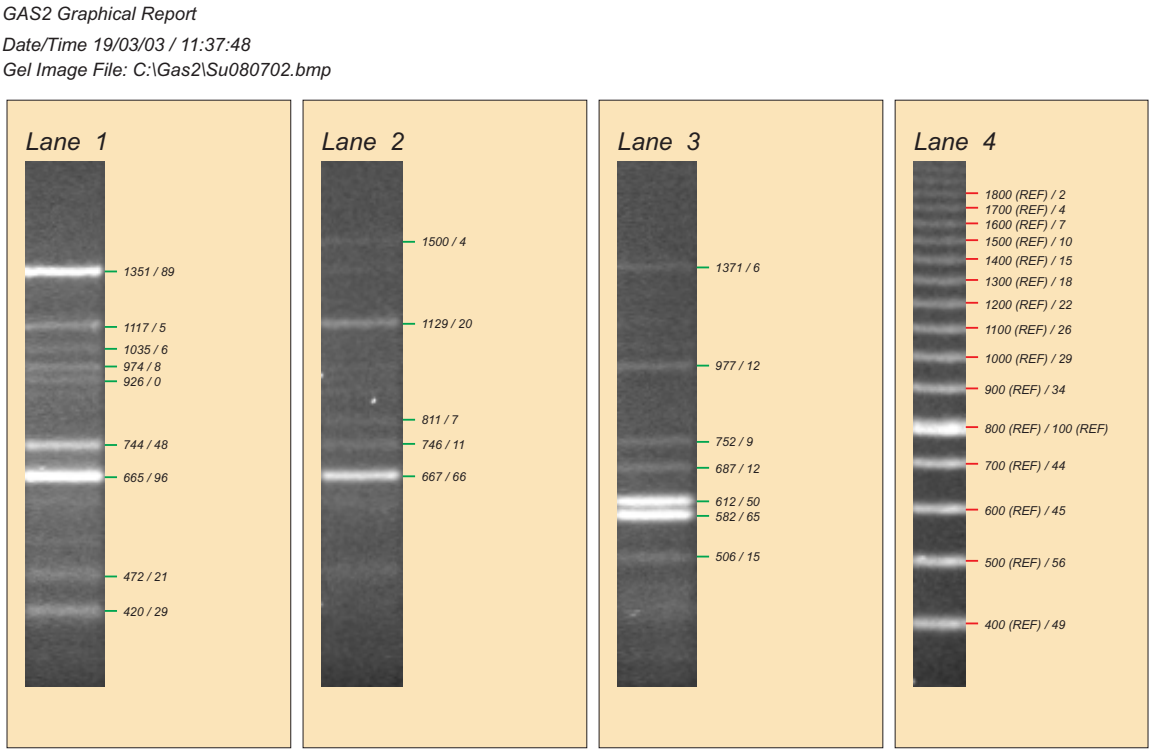

Figure 8: The report example.

of the GAS2 in the Molecular Biology Group is in progress.

References

[1] I. Bajla, I. Holländer, and K. Burg: A study on DNA-Gel image analysis improvement. In: I. Frollo, M. Tyšler, and A. Plačková (eds) Proc. of the 3rd Int. Conf. on Measurement, Smolenice, Slovak Republic, May 14-17, 2001. VEDA Bratislava, 2001, 223-226.

[2] I. Bajla, I. Holländer, and K. Burg: Improvement of electrophoretic gel image analysis. Online edition of the Measurement Science Review, Journal of the Institute of Measurement Science, Slovak Academy of Sciences, No.5-10, Paper Section 2: Measurement in Biomedicine. “http://www.measurement.sk”.

[3] M. Kollár: Software tool for processing electrophoretic gel images. MSc diploma thesis, FEI STU Bratislava, 2003.

[4] I. Bajla, I. Holländer, K. Burg, and S.Fluch: Novel approach to quantitative analysis of electrophoretic gel images of DNA fragments. In: Proc. of the IEEE Int.Symposium on Biomedical Imaging, 7-11 July 2002, Washington, pp.889-902, CD ROM.