SpotDSQ
Kostopoulou, Eirini; Katsigiannis, Stamos; Maroulis, Dimitris

Published in:
19th IEEE International Conference on Bioinformatics and Bioengineering

DOI:
10.1109/BIBE.2019.00015

Published: 26/12/2019

Document Version
Peer reviewed version

Link to publication on the UWS Academic Portal

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the UWS Academic Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact pure@uws.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
SpotDSQ: A 2D-gel image analysis tool for protein spot detection, segmentation and quantification
Kostopoulou, Eirini; Katsigiannis, Stamos; Maroulis, Dimitris

Published in:
19th IEEE International Conference on Bioinformatics and Bioengineering

Accepted/In press: 31/08/2019

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the UWS Academic Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact pure@uws.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
SpotDSQ: A 2D-gel image analysis tool for protein spot detection, segmentation and quantification

Eirini Kostopoulou
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece
ikostop@di.uoa.gr

Stamos Katsigiannis
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece
stamos@di.uoa.gr

Dimitris Maroulis
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece
dmaroulis@di.uoa.gr

Abstract—The field of proteomics offers powerful methods for studying and analyzing protein expression levels in cells. The Two-Dimensional Gel Electrophoresis technique is a well-established proteomics technique focusing on protein separation and identification that provides digital images containing thousands of protein spots. 2D-gel images are then segmented into spots and background in order to quantify the expression levels of proteins located on a single gel or of differentially expressed proteins between samples from a series of 2D-gels. Spot detection and segmentation are complex and arduous tasks due to the inherent characteristics of 2D-gel images. Several software packages and experimental methods are available for 2D-gel image analysis, achieving different levels of success, each one having its respective advantages and drawbacks. A common characteristic is their dependency on user intervention in order to achieve optimal results, a process that leads to subjective and usually non-reproducible results. In this work, the authors present SpotDSQ, a software tool for 2D-gel electrophoresis image analysis that incorporates novel algorithms for accurate spot detection, segmentation and quantification.

SpotDSQ provides high quality automatic protein spot detection, segmentation and quantification by employing novel algorithms that outperform state-of-the-art alternatives. Local automated multi-thresholding along with a modified version of the grow-cut segmentation algorithm is utilized in order to detect areas containing spots as well as the spot centers. The segmentation process is then guided by the information gathered during the detection step, utilizing the detected spot centers as seeds for a region growing approach that separates spot areas, while morphological operators are then utilized in order to accurately detect the protein spots boundaries. SpotDSQ offers an easy-to-use graphical user interface that requires no special training to operate. Results are exported as images and text data in order to facilitate further analysis.

The performance of SpotDSQ was evaluated on real as well as synthetic 2D-gel images using well-established statistical measures. Spot detection performance was evaluated by means of precision, sensitivity, and the F-measure. Volumetric overlap, volumetric error and volumetric overlap error were utilized for evaluating the segmentation performance. The high F-measure (94.8%) value, the low volumetric overlap error (8.3%), and the accurate spot boundaries achieved by SpotDSQ indicate its effectiveness compared to alternative methods.

Experimental results show that SpotDSQ outperforms state-of-the-art software packages as well as methods proposed in the literature, achieving high accuracy and reduced errors. The advantages of SpotDSQ indicate that it has the potential to be a powerful and reliable tool for 2D-gel image analysis in biomedical laboratories.

Keywords—Proteomics, 2D-gel electrophoresis, Segmentation, Spot detection, Protein quantification

I. INTRODUCTION

The field of proteomics offers powerful methods for studying and analyzing proteomic changes in cells [1]-[3]. One of the most widely utilized techniques in proteomics is the Two-Dimensional Polyacrylamide Gel Electrophoresis (2D-PAGE) technique that aims at protein separation and identification. 2D-PAGE employs isoelectric focusing and sodium dodecyl sulfate polyacrylamide gel electrophoresis in order to separate thousands of proteins on polyacrylamide gels by taking into consideration the differences in their net charge and their molecular mass [4]-[5]. The results of the 2D-PAGE technique are visualized into a digital image containing thousands of protein spots, which is then analyzed in order to detect and quantify the expression levels of proteins located on a single gel or of differentially expressed proteins between samples from a series of 2D-gels. In order to measure the expression level of protein spots, their position on the 2D-gel image must be first detected and then the pixels belonging to each spot must be distinguished from those belonging to the background, a process referred as “segmentation”. Spot detection and segmentation are challenging tasks due to the inherent characteristics of 2D-gel images that suffer from artifacts, inhomogeneous background and high levels of noise. Moreover, protein spots exhibit great variability in intensity, size and shape, while adjacent spots are often overlapped [6].

Available commercial software solutions for 2D-gel image analysis like for example PDQuest [7], DeCyder 2-D [8], Melanie 7 [9], ImageMaster 2D [10], Delta2D [11] and Progenesis SameSpots [12] achieve varying levels of success [13]-[14]. All these tools depend on manual parameter tuning that can inevitably lead to subjective results. Moreover, time-consuming human intervention may be needed for manual editing of the results since all these tools may: 1) fail to separate overlapping spots, 2) separate single spots into more, 3) not accurately approximate the spot boundaries, 4) miss some spots and 5) characterize artifacts as spots. The extracted protein expression levels are significantly affected by all these errors, interfering with the extraction of reliable biological conclusions. In addition to the non-reproducible results due to
human intervention, the throughput of the analysis process is also significantly reduced. Many methods for 2D-gel spot detection and segmentation have also been proposed in the literature [15]-[21], each one having its respective advantages and drawbacks. Recently, Dos Anjos et al. [22], in their proposed watershed-based method named “Scimo”, achieved more accurate estimation of close proximity spots and partially overlapping spots, but underperformed in cases of major spot overlapping. Furthermore, a method based on 2D histograms and 3D spot morphology, proposed by Kostopoulou et al. [23], achieved significantly improved results but at the cost of high computational complexity.

In this work, the authors present SpotDSQ, a 2D-gel image analysis tool which provides more accurate segmentation results than state-of-the-art alternatives and offers an easy-to-use graphical interface, advanced parameter tuning options, result visualization and creation of report files with segmentation and quantification results.

II. METHODS

SpotDSQ was developed using MatLab (MathWorks, v. 2016a) and runs on the latest Windows 64-bit operating systems. It has been compiled into a standalone application that requires the MatLab runtime in order to run, without the need of a MATLAB installation. SpotDSQ employs a novel scheme for spot detection and segmentation, recently proposed by the authors in [24]. The rest of this Section provides a brief overview of the detection and segmentation method used in SpotDSQ. The following algorithms were incorporated into a user-friendly graphical user interface (GUI) in order to facilitate their use and enable users with no technical background to apply these algorithms to 2D-gel images.

A. Spot detection

For the detection process, the image is first partitioned into overg regions and then a multi-thresholding scheme is applied in order to select pixels as spot center candidates and detect regions that have a high probability to belong to foreground (spot) or background. Initially, the image is partitioned into tiled windows \( W_i \), \( i = 1, \ldots, \text{N} \) of size \( d' \times d' \) and overlapping windows \( W_{ij} \), \( i = 1, \ldots, \text{N} \) of size \( d \times d \), where \( d' \) is selected by the user and \( d \) is calculated as \( d' = \lceil \sqrt{d'/8} \rceil \). For each overlapping window \( W_{ij} \), \( M \) thresholds \( T_{ij} \), \( j = 1, \ldots, l, \ldots, h, \ldots, M \) are automatically computed using the multiple Otsu Thresholding method [25]. Then, at each tiled window \( W_i \), pixels with intensity value higher than \( T_{ik} \) are added to the set \( F \) of foreground pixels (spot pixels), while pixels with an intensity value less than \( T_{ij} \) are added to the set \( B \) of background pixels. Fig. 1(a) illustrates an example of a tiled window of \( I \), as well as its corresponding overlapping window, whereas the results of the multi-thresholding scheme with the \( M + 1 \) thresholds and the two thresholds \( (T_{ij}, T_{ik}) \) are depicted on Fig. 1(b) and Fig. 1(c), respectively. Each pixel belonging to the set \( F \) that corresponds to a local intensity maximum of the image is subsequently considered as a spot center candidate pixel. Nevertheless, as shown in Fig. 1(d), this procedure leads to some spots containing multiple candidate spot centers. In order to address this problem, candidate spot centers are subsequently discarded by applying the local thresholding technique [26] in a region of size \( \alpha \alpha \) around each candidate spot center in order to distinguish between background and foreground pixels and then discarding the candidates whose Euclidean distance from the nearest pixel classified as “Background” inside the \( \alpha \alpha \) region is less than a user defined parameter \( k \), and b) by merging the candidates whose Euclidean distance from each other is less than \( k \) (merging is performed by replacing the candidates with their midpoint each other is less than \( k \) (merging is performed by replacing the candidates with their midpoint pixel). Fig. 1(e) depicts the final spot centers.

B. Segmentation into spots and background

The first step of the segmentation process consists of a modified grow-cut algorithm [27] utilizing a custom update rule that takes into consideration the inherent characteristics of 2D-gel images. The modified grow-cut update function used is defined as:

\[
g(q, p) = \begin{cases} 
0, & p, q: (I(q) = f) \land [I(p) < T_{ij}] \\
1, & p, q: (I(q) = f) \land [I(p) > T_{ij}]
\end{cases}
\]

\[
v(q, b) = \begin{cases} 
1 - \frac{T_{ih} - I(p)}{T_{ih} - T_{ij}}, & p, q: (I(q) = f) \\
\frac{T_{ih} - I(p)}{T_{ih} - T_{ij}}, & p, q: (I(q) = b)
\end{cases}
\]

in order to take into consideration both the difference in intensity values between each pixel \( p \) and its neighbor pixels \( q \) as well as the difference in labels (“spot” or “background”). \( I(p) \) denotes the label of pixel \( p \), \( I(p) \) the intensity of \( p \) and \( f \) and \( b \) are the labels for spot and background respectively. The pixels belonging to the sets \( F \) and \( B \), created during the spot detection process, are used as seed pixels for the spot and background regions respectively. The result of the modified grow-cut for the image region \( W_i \) depicted on Fig. 1(a), is shown on Fig. 2(a). After this initial segmentation process, each spot region formed by pixels labeled as “spot” is further segmented to the spot areas it contains, utilizing a region growing approach. Let \( R_s \) be a distinct spot region and \( K_s \), the subset of spot centers of \( I \) that are located inside \( R_s \). The pixels of \( K_s \) are utilized as seed pixels and are each labeled with a different ID, \( R_{s,b} \). Then, the rest of the pixels of \( R_s \) considered to belong to the set \( L_s \) are grouped based on their intensity values. The unique intensities \( V_z \) of pixels belonging to \( L_s \) are determined in descending intensity order and then groups of similar intensity values –each containing \( x \) distinct intensities- are formed, with \( x \) being a user-defined parameter. Pixels of higher intensity values are examined and labeled and subsequently utilized as seeds to impose their label on their neighborhood pixels of lower intensity values. Let \( u_{n} \) be the number of different intensities of \( L_s \) and \( z_n = \lfloor u_{n}/x \rfloor \) the number of groups. Each group of pixels \( G_{p,n} \), \( e = 1, \ldots, z \), contains the pixels of \( L_s \) whose intensities are within the range of \( V_{z}(e-1+1), V_{z}(e) \). The pixels of each \( G_{p,n} \) are then labeled based on their neighborhood pixels. For each pixel \( p \), its neighborhood pixels are examined and \( p \) is labeled by the application of a majority voting criterion on the labels of its already labeled neighbors.
of $[V_{(e-1)+1}^e, V_{e+1}]$. The pixels of each $G_p$ are then labeled based on their neighborhood pixels. For each pixel $p$, its neighborhood pixels are examined and $p$ is labeled by the application of a majority voting criterion on the labels of its already labeled neighbors. If none of its neighborhood pixels is already labeled, it will be reexamined and labeled at the next iteration of the procedure. An example of the application of the majority voting process is shown in Fig. 2(b). As it can be observed, a small number of low intensity pixels of spot areas are missed and some high intensity background pixels are included in the spot areas. In order to address this problem, the morphological dilation operator is applied on each area $R_{a,b}$ followed by the application of the optimal thresholding technique [26] in each dilated region $R_{a,b'}$. The optimal thresholding technique is applied on both the intensity and the gradient intensity values in order to detect the pixels near the spot centers, as well as those near the spot boundaries, whose gradient intensity value is high. The final segmentation result for the region $W_1'$ depicted in Fig. 1(b) is shown on Fig. 2(d).

III. EXPERIMENTAL EVALUATION

SpotDSQ outperforms software packages and state-of-the-art techniques in both spot detection and segmentation, as it detects and segments the majority of spots, avoids the detection and segmentation of spurious spots and achieves the best balance amongst correctly detected and segmented spots and background areas falsely segmented as spots. The detection and segmentation algorithms incorporated into SpotDSQ have been evaluated using real and synthetic 2D-gel image datasets, as reported in [24]. Each dataset consisted of 2D-gel images of 16-bit color depth, containing a total of ~10200 spots. The real dataset has been provided through the courtesy of the Biomedical Research Foundation of the Academy of Athens (BRFAA) [28] whereas the synthetic dataset consists of images generated by our research group [23] by combining spots produced using the 2-D Gaussian flat top function with background extracted from real images in order to imitate the real ones more accurately and retain their characteristics, i.e. overlapping spots, poorly contrasted spots, inhomogeneous background, artifacts and streaks. Both datasets contain inverted images (spots appear bright on a dark background instead of dark on a bright background), a technique commonly applied to 2D-gel images as shown in [29] and [30]. Performance on real 2D-gel images has been quantitatively evaluated only for the detection step, due to the unavailability of segmentation ground-truth information, whereas performance on the synthetic 2D-gel images has been quantitatively evaluated for both detection and segmentation. For the evaluation, the approach employed by SpotDSQ was compared with Delta2D [11], Melanie 7 [9], PDQuest [7] and the recently published method Scimo [22].

Spot detection performance was evaluated in terms of precision, sensitivity, and their weighted harmonic mean (F-measure) in [23] and the results are presented in Fig. 3(a). SpotDSQ achieves a mean precision value equal to 96.5% (± 2.9) compared to the second best 91.9% (± 3.7) achieved by Scimo, and comparable sensitivity to Melanie 7 (93.2% ± 2.6 versus 94.1% ± 2.4). Nevertheless, SpotDSQ outperforms all methods in terms of the F-measure (94.8% ± 2.8 compared to 83.4% ± 3.3 and 89.4% ± 2.4 for Melanie 7 and Scimo, respectively), resulting to less false-positive (spurious) spots and a high number of correctly detected spots, indicating its better overall performance. The F-measure is the weighted harmonic mean of precision and sensitivity and can be considered as more reliable measure than just sensitivity and
Segmentation performance was evaluated in (Kostopoulou, Katsigiannis, Maroulis, 2015) in terms of the Volumetric Overlap (VO), Volumetric Error (VE), and Volumetric Overlap Error (VOE) measures, defined as:

\[
VO = \frac{ASV}{(ASV + FBV)} \quad (2) \\
VE = \frac{FSV}{(ASV + FBV)} \quad (3) \\
VOE = 1 - \frac{ASV}{(ASV + FBV + FSV)} \quad (4)
\]

\(ASV, FBV\) and \(FSV\) denote the volume of “Actual Spot” (AS), “False Background” (FB) and “False Spot” (FS) pixels respectively, with \(V = \sum_{x,y} I(x,y)\) and \(I(x,y)\) the pixel intensity. The \(VOE\) is considered as the most important metric and its value is zero when there is complete overlap among the segmentation result and the ground truth, whereas it is 100 when there is no overlap. SpotDSQ achieves the lowest mean \(VE\) and \(VOE\) among the examined approaches: 5.7% \(\pm\) 1.6 and 8.3% \(\pm\) 0.5 respectively, compared to the second lowest 12.9% \(\pm\) 2.6 and 12.6% \(\pm\) 1.5 achieved by Scimo. The large difference in these segmentation errors (Fig. 3b) demonstrates the efficiency of SpotDSQ in segmenting 2D-gel images. Details on the algorithms employed by SpotDSQ and a thorough experimental evaluation are available in [24]

IV. THE SPOTDSQ SOFTWARE

The SpotDSQ GUI allows the user to select and analyze 2D-GE images of TIFF format in either their original or inverted form. Optimal parameters [24] are pre-set and the user can start the analysis process after loading up to six images by pressing the “Select Image(s)” button (Fig. 4a) and then pressing the “Run algorithm” button (Fig. 4b). Alternatively, the user may choose to modify the default parameters before initiating the analysis procedure. In addition to parameters related to the detection and segmentation methods described in the Methods section of this work, SpotDSQ offers the option to also use the original grow-cut algorithm [27], the Fast Marching Method [31], segmentation via adaptive weighted [32] and watershed [33] as an alternative to the proposed custom grow-cut algorithm at the first step of the segmentation process. After pressing the “Run algorithm” button, a progress bar indicates the progress of the analysis.

After completion, the output images are presented next to the input images and the spot boundaries are overlaid using the selected color, as shown in Fig. 5. An Identification Number (ID) is assigned to each spot and can also be overlaid to the images if the button “Show/Hide IDs” is pressed (Fig. 5d). The user can zoom in or out and pan through the input and output images using the respective tools (Fig. 5h). The display of input and output images is synchronized in order to always keep the same region of the images in view. After selecting an image tool, the mouse cursor will change accordingly. In order to zoom in or out of a region of an image, the user has to click on the input or output image as many times as needed to reach the desired zoom level. For scrolling through the regions of a zoomed image, the user can move through the image by moving the mouse while holding down the "Run algorithm" button, a progress bar indicates the progress of the analysis.
the left mouse button. The user can reset the display to the original view by pressing the “Reset to Original View” button (Fig. 5c). Using the “Spot selector” tool (Fig. 5e), the user can click on a spot or background area on each gel image. Quantification information for the spot selected from each image is shown inside a window (shown after pressing the button “Show Comparison” (Fig. 5f) that will appear after clicking on a spot) in tabular form, in order to facilitate the comparison. Quantification information for each spot includes its size in pixels, the sum of the intensity values of its pixels as explained in [34], and two normalized forms: 1) the intensity of each spot is normalized by the total intensity of all spots in the gel [35] and 2) the intensity value of each spot is replaced by its respective position in the cumulative distribution of spot intensities in the gel [35]-[36]. These normalization techniques have been shown to provide fairly reproducible results (Zhou et al., 2002; Almeida et al., 2005). Minimum, maximum, mean and median values are also shown regarding the quantification results of the spots selected. In case the user clicks on the background, the absence of the spot at that position is reported.

An output folder named as “%TIMESTAMP%” is created for each experiment and the results for each image are saved into separate folders inside the experiment folder, named as “%ImageName%”, containing: 1) the input image, 2) the grayscale and color versions of the output image in TIFF format annotated with the spot boundaries and IDs, 3) a text file with coordinates and the protein quantification results for every detected spot (Fig. 6) and 4) a text file with the parameter values. In the text file with the results, the ID of each spot is presented along its coordinates and its quantification metrics (size, sum of pixel intensities, normalized values). The user can easily access the output folder by pressing the “Open output folder” button (Fig. 6g). Average analysis times are 0.5 min, 1 min and 11 min for 16bit real 2D-GE images of resolution 400x400, 1000x1000 and 1822x2500 respectively, on a computer equipped with an Intel i7-6700K CPU (released in 2015). The improved segmentation quality offered by the algorithms incorporated in SpotDSQ [24] compared to available software, constitutes an acceptable trade-off between quality and analysis time.

Fig. 4: Main screen of SpotDSQ. Four loaded inverted 2D-GE images are shown on the left, while the parameter tuning options are shown on the right. Annotated regions correspond to the zoomed regions shown in Fig. 5.
Fig. 5: Screenshot of SpotDSQ after finishing the analysis of the four inverted 2D-GE images depicted in Fig. 5. Images have been zoomed to the region annotated in Fig. 5 and Spot IDs are overlaid on the results.

Fig. 6: Example of the text file containing the IDs, the coordinates, the size in pixels and the protein quantification result for every detected spot.

V. CONCLUSIONS

SpotDSQ is a powerful tool for 2D-gel image analysis. In contrast to other available tools, it is focused in minimizing the number of background pixels detected as spot pixels. As a result, the protein expression levels are more accurately quantified, thus enhancing the ability to extract more reliable biological conclusions. Experiments on real and synthetic 2D-gel images demonstrated that SpotDSQ outperforms state-of-the-art 2D-gel image analysis tools and techniques including Melanie 7, Delta2D, PDQuest, and Scimo. SpotDSQ offers a user-friendly standalone application that requires no specialized hardware to run. Quality results can be obtained easily by selecting an image and clicking a button. Nevertheless, it offers to the advanced user the option to modify parameter values in order to cope with images exhibiting unusual characteristics. Results are organized in separate folders for each experiment and are extracted as annotated images, as well as a text file with information for every detected spot, thus providing the required flexibility for further analysis. All these advantages indicate that SpotDSQ can be a reliable and useful tool that can significantly assist biologists in 2D-gel image analysis. SpotDSQ is available for download at http://rtsimage.di.uoa.gr/index.php?option=com_content&view=article&id=75&Itemid=69.

REFERENCES


