A FULLY AUTOMATED 2-DE GEL IMAGE ANALYSIS PIPELINE FOR HIGH THROUGHPUT PROTEOMICS

Panagiotis Tsakanikas and Elias S. Manolakos

Department of Informatics and Telecommunications
University of Athens, Greece
{tsakanik,eliasm}@di.uoa.gr

ABSTRACT

Image analysis is still considered as the bottleneck in 2D-gel based expression proteomics analysis for biomarkers discovery. We are presenting a new end-to-end image analysis pipeline of operations that can be fully automated. The pipeline includes image denoising and enhancement based on contourlets, image segmentation into Regions of Interest (ROIs) based on active contours, followed by the analysis of the extracted ROIs for spot detection and quantification using mixture modeling, model selection and unsupervised machine learning methods. The proposed system is shown to match the sensitivity and exceed the precision of commercial spot detection software when analyzing real 2D gel images. It is suitable for high throughput proteomics analysis of image stacks since, unlike commercial software, it does not require any manual re-calibration of parameters every time a new image is to be processed.

Index Terms—Proteomics, 2D gel electrophoresis, image denoising, image segmentation, spot modeling, spot detection.

1. INTRODUCTION

Proteomics [1] is the field concerned with the study of the time varying proteome of species by using technologies of protein separation and identification. Two-dimensional gel electrophoresis (2DGE) is a widely used method for the analysis of complex protein mixtures extracted from cells, tissues, or other biological samples. It is used to separate proteins in two dimensions according to two independent properties: their molecular weight (MW) and their isoelectric point (pI) i.e. the pH at which the proteins have no net charge. In this way, complex mixtures, with thousands of proteins, can be spatially resolved on a 2D gel and the relative expression level of each protein can be determined. A key objective in a proteomic study is the identification of differentially expressed proteins, i.e. the identification of the protein spots that have been induced or inhibited i.e. have changed abundance considerably (spots increased or decreased in size and intensity). Once these protein spot features are found, the corresponding proteins can be identified accurately using mass spectrometry (MS).

In order to detect protein spot features, dedicated image analysis software is employed. There exist several mature commercial packages used in this field e.g. PDQuest, ImageMaster, Delta2D etc. Despite their capabilities, currently available commercial software packages are still limited when it comes to reliable spot data generation. As demonstrated in a recent study for all the commercial software packages considered “…less than 3% of the total processing time is automated. The remainder of the time is spent in manual, subjective editing of detected spots…” [2]. However, in order to achieve high throughput in proteomics experiments that require the processing of stacks of images the image analysis pipeline needs to be fully automated, without compromising the quality and reliability of the results. In this paper, we present such an end-to-end image analysis pipeline for fully automated 2DGE image analysis, i.e. requiring from the user to input only the gel image(s). We should note that while we evaluated its performance relatively to PDQuest [3], our main objective is to show that its accuracy and precision does not degrade despite its full automation.

Figure 1 presents a summarizing schematic view of the implemented pipeline. First we employ the Contourlet Transform in order to efficiently suppress the inherent noise [4] and to enhance the protein spots characteristics [5]. Next we segment the gel image into Regions of Interest (ROIs) i.e. regions that enclose the spots present in the gel image, using an Active Contours (ACs) based approach [5]. The extracted ROIs are then processed independently of each other so as to detect the individual protein spots they contain. In order to remove the local background and “streaks” (a common issue in gel images), a histogram based classification technique is first applied, using 1-dimensional (1D) mixture modeling. Then the resulting ROI pixels are transformed, by appropriate random sampling, into an isomorphic set of data points where 2D mixture modeling along with model selection is applied for spot detection and quantification. The result of the best mixture model selection and model fitting is the final spot list including also all spot quantities. It must be noted that the whole pipeline is fully automated, requiring from the user to input only the 2DGE image to be analyzed. Whatever parameters are needed they are estimated automatically on the fly from the image data. The performance of each part of the pipeline has been thoroughly evaluated and the results are presented elsewhere [4,5,6]. Here we present only an overview of the whole pipeline and its end-to-end performance evaluation. We should mention that the same pipeline of operations can be applied to either single gel images or group average (category) gel images without any modification.

The rest of the paper is organized as follows: In Section 2 we present briefly the elements of the pipeline. Section 3 describes the experimental evaluation method used and presents the results obtained. Finally, in Section 4 we summarize our findings and point to interesting future directions.
2. 2DGE IMAGE ANALYSIS PIPELINE

2.1. Preprocessing
Before spot detection 2DGE images need to be smoothed to suppress the extrinsic noise inherent in them (due to the sample preparation and acquisition process imperfections etc.). The most commonly used smoothing techniques are based on spatial filters such as: local Gaussian, various versions of mean filtering (power-mean, contra-mean), diffusion or polynomial convolution, and median filter. Recently the Wavelet Transform (WT) has been shown to outperform spatial filtering in 2DGE image denoising [7]. In order to overcome some disadvantages of the WT, we have introduced the use of the Contourlet Transform (CT) [8] and have demonstrated its superiority relatively to spatial filters and wavelet-based methods in [4].

We have also used the CT in order to enhance the image and make faint spot features “stick out” in areas of high noise and/or high local background variability. So, we apply an enhancement function [5] to the “middle” frequency scales transform coefficients since the objective is not only to enhance the signal but also avoid enhancing background structures and/or high frequency components (noise or high frequency edges). For more details please refer to [5].

2.2. Spot areas Segmentation
Segmentation of 2DGE images requires partitioning them into areas of “foreground” (that with high probability include protein spots) and “background” (with no protein spots). We have introduced a method based on Active Contours [9] that separates effectively and in an automatic way the pixel in those two areas of a gel image, in a way that: (i) reduces the number of missed faint spots, (ii) extracts correct and tight borders for areas containing spots, (iii) avoid over-segmentation (a major problem of many commercial packages using Watershed transform based spot segmentation [10]). As shown in [5] the extracted Regions of Interest (ROIs) include the large majority of the true spots and have “tight boundaries”, i.e. they do not enclose large background areas (please see [5] for more details).

2.3. Spot Detection & Quantification
Previously developed methods for protein spot detection suffer from various disadvantages, such us sensitivity to noise and artifacts, spot border distortions, over-segmentation and poor performance in areas with overlapping spots [10]. Furthermore, they require careful post-processing and a lot of manual effort, to finally produce reliable detection results [2]. To address these limitations, a novel approach for 2DGE automatic spot detection and quantification has been developed. Here, the informative ROI pixels are treated as sample data generators and Gaussian Mixture Modeling (GMM) is applied in an unsupervised manner [11,12], i.e. no pre-training is required and the whole process can be fully automated.

2.3.1. Local Background and Streaks removal
Although the ROI extraction step results in areas that include the vast majority of the protein spots present in a gel image, they may also include some local image background pixels and/or streak segments. In order to eliminate those pixels that may confuse spot modeling downstream from further processing, we classify the object pixels into 3 possible classes: class-1 represents the “strong” and/or saturated spot pixels, class-2 the “faint” spot pixels and tails of “strong” spots, and finally class-3 correspond to the pixels of the local background and/or streaks. The classification is performed on the histogram of pixel intensities using 1-dimensionnal Gaussian mixture modeling and a modified Expectation-Maximization (EM) algorithm [12]. The Minimum Message Length (MML) criterion (proposed in [12] for model selection) is applied to determine the optimal number of components in the mixture model that best fits the histogram data.

In summary, the use of the modified EM fits 3 pixel classes to the histogram while the MML criterion tests this fit and discards any class that is not essential for the histogram’s interpretation. So, if the histogram is well explained by 3 classes we can then threshold out the class-1 pixels that correspond to the background and/or streaks (light gray in Figure 1 (4)), else (if we end up with less than 3 classes) we keep the object intact. As one can notice, the developed method is performing local background and streak removal task only in areas and to the extent that it is really needed.

2.3.2. Initial Spot centers Estimation
At the next step we estimate the number of protein spots existing inside each ROI. To do so, a 5x5 (pixels) spatial filter is employed that finds the local minima (zero intensity corresponds to black pixels, maximum intensity to white) in the image (see green asterisks in Figure 1 (6)). Due to the pixel intensity saturation effect, it is possible that the filter identifies several closely located local minima. These are actually replicates of the same candidate spot centre and need to be grouped appropriately so as we do not end up with a very large and misleading number of candidate spot centers per object. This is accomplished by applying agglomerative Hierarchical Clustering (HC) [13] of the minima points using the Manhattan distance, the single linkage method for merging formed clusters, and cutting the resulting dendrogram tree at distance 7.
2.3.2. 2D Spot Modeling

The next and most distinguishing characteristic of our pipeline is the idea of using the pixel intensities as data generators. The total number of data points $N$ generated by random sampling for each ROI is kept proportional to the number of estimated candidate spot centres, and not to the area of the ROI. The $N$ points to be drawn are distributed among the pixels of the object according to their relative intensities (a “stronger” pixel “throws” more points in its neighbourhood). This is in accordance to the view that pixel intensity ideally represents the quantity of protein molecules concentrated at that particular gel location. Specifically, each pixel $i$ with intensity $I_i$ acts as the centre $\mu=(\xi, \eta)$ of a 2D Gaussian component $N(\mu, \Sigma)$ in a Gaussian Mixture Model [11] having as many components as the number of pixels and a predetermined fixed covariance matrix $\Sigma$.

As we can notice from Figure 1 (7) the generated set of data points may include points that are far from all candidate spot centres (light grey points). These outlier points (due to remaining background pixels) can be identified since they have low likelihood for all components of the mixture (no component “feels strongly” about them) and are removed at this stage since they may adversely affect the subsequent step of spot detection and quantification.

The last step in the pipeline is the application of Gaussian Mixtures Models (GMM) [11] in 2-dimensions on the generated dataset for spot modelling. So far we have established: (i) a good estimate of the maximum expected number of the mixture components and their means (candidate spot centres), (ii) an initial estimate of the covariance matrices of each candidate component (this can be easily obtained by applying nearest neighbour classification of data points to the candidate spot centres), and of the mixing weights of the GMM. So, using this information as an initial GMM configuration (see Figure 1 panel (8)) a similar approach as in Section 2.3.1 is applied but in 2D in order to identify the 2D mixture model that fits best the survived data points. Figure 1 (top panel) presents the final spot detection result.

3. RESULTS AND DISCUSSION

**Evaluation method:** Two real images have been used in the evaluation, available at [http://www.umbc.edu/proteome/image_analysis.html](http://www.umbc.edu/proteome/image_analysis.html). These images have also been used in [14] for the comparative evaluation of commercial software packages. They are accompanied by expert-user annotation, meaning that we are given the ground truth (number and center positions of true protein spots).

For each method evaluated we counted the number of missed spots (FNS), the number of introduced artifacts (FPs) and the True Positive Fraction: TPF=TP/(TP+FN), i.e. the percentage of the true spots in the gel that have been found (it can be done since we know the ground truth). Furthermore, we computed the Positive Predictive Value, PPV=TP/TP+FP, i.e. the probability that a spot detected by the software is a true one. Finally, we have also computed the F-measure defined as $2\times TP/(2\times PPV + TP+FP)$, that is a harmonic mean commonly used to compare methods in terms of how well they handle the sensitivity vs. precision trade-off.

![Figure 2. Spot detection performance comparison.](http://www.umbc.edu/proteome/image_analysis.html)

We compare next results obtained by the designed pipeline with those obtained using the mature commercial software package PDQuest (version 8.0.1). Before using PDQuest for spot detection we had to determine the best “sensitivity” parameter value for each image using Free-response Receiver Operating Characteristic curves (FROCs). We considered as “best” the sensitivity value that maximizes the ratio TPF/(FP+FN). This is a laborious process that has to be repeated for every new image when using PDQuest.

Figure 2 summarizes the comparative evaluation results. We notice that the proposed method exhibits high TPF (over 92%) and at the same time high PPV (over 93%), suggesting that it is both sensitive and precise. Furthermore, the achieved near perfect F-measure (over 93%) indicates that although the method is fully automated it works close to the “golden ratio” in terms of the sensitivity vs. precision trade-off.

Figure 3 presents an image area (RamanA image). Panel (a) shows the constructed spot models (the true spots are 114) overlaid on the original image. Arrows indicate the extraneous spot artifacts (there exist 6) while red ellipses the missed true spots (2 faint spots). In panel (b) (background subtracted image) the magenta borders represent the extracted ROIs using Active Contours. The red circles denote the initial estimation of spot centers and the green crosses the final positions of spot centers. The missed spots are due to AC-based segmentation (cases where we failed to include true but faint spots into the active contours) rather than in the spot modeling stage. So, a further improvement of this stage of the pipeline is needed and is under investigation. Further, Figure 4 presents an image area from RamanB gel image. Here the true spots are 99, the introduced artifacts are 5 and the missed spots are 4.

Finally in Table 1 we present the processing times for each pipeline step (using non-optimized Matlab code and RamanB image) and the total time needed for processing an 1024x1024 image on a PC with Intel Core2, 4 GB of RAM running at 2.4GHz. As we can notice the most time-consuming step is the AC based segmentation. Nevertheless, since the proposed pipeline is fully automated, the total processing time needed is not an important issue since it gives to the scientists the flexibility to load their images and continue with their work without any further involvement. Furthermore the time can be substantially reduced by recoding the pipeline in C and/or spot detecting the extracted ROIs in parallel using a multi-core CPU.
4. CONCLUSIONS

In this paper we presented an overview of a novel, fully automated image analysis pipeline for 2DGE images. The proposed method does not require any recalibration every time a new image is processed. It has been found to be both highly sensitive and precise while it seems to perform near the “golden ratio” according to the F-measure. The estimated spot quantities were found to be close to the true spot quantities (we have used synthetic images for this purpose, results not shown here but in [6]). The method still needs to be evaluated with a larger dataset of images in order to extract statistically significant measures for its performance (work in progress). Finally, further improvements in AC-based image segmentation stage, is expected to further increase the spot detection performance of the pipeline.

ACKNOWLEDGEMENTS

This work has been supported by Research Grant 03ED306, implemented within the framework of the "Program of Human Research Manpower Reinforcement" (PENED) co-financed by the Greek General Secretariat of Research and Technology and EU funds, and by EU FP7 Regpot grant TranSMed (Ref. Number 245928).

5. REFERENCES