BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 3 2011, pages 368–375
doi:10.1093/bioinformatics/btq666

Gene expression

Advance Access publication December 2, 2010

New approach for segmentation and quantification of

António dos Anjos[1,*], Anders L. B. Møller[2], Bjarne K. Ersbøll[3], Christine Finnie[2] and Hamid R. Shahbazkia[1]

[1]Department of Electronic and Informatics Engineering, University of Algarve, Faro, Portugal, [2]Department of Systems Biology, Enzyme and Protein Chemistry and [3]Department of Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Detection of protein spots in two-dimensional gel electrophoresis images (2-DE) is a very complex task and current approaches addressing this problem still suffer from significant shortcomings. When quantifying a spot, most of the current software applications include a lot of background due to poor segmentation. Other software applications use a fixed window for this task, resulting in omission of part of the protein spot, or including background in the quantification. The approach presented here for the segmentation and quantification of 2-DE aims to minimize these problems.

Results: Five sections from different gels are used to test the performance of the presented method concerning the detection of protein spots, and three gel sections are used to test the quantification of sixty protein spots. Comparisons with a state-of-the-art commercial software and an academic state-of-the-art approach are presented. It is shown that the proposed approach for segmentation and quantification of 2-DE images can compete with the available commercial and academic software packages.

Availability: A command-line prototype may be downloaded, for non-commercial use, from http://w3.ualg.pt/~aanjos/prototypes.html.

Contact: antoniodosanjos@gmail.com

Supplementary information: Supplementary data are available at *Bioinformtaics* online.

Received on August 26, 2010; revised on November 11, 2010; accepted on November 30, 2010

1 INTRODUCTION

Developing new bioinformatics tools, both computational and experimental, such as those for membrane protein type prediction (Cai *et al.*, 2003), protein subcellular location prediction (Cai *et al.*, 2010; Chou and Cai, 2002; Chou and Shen, 2010), drug–target interaction prediction (He *et al.*, 2010), substrate–enzyme–product triad network prediction (Chen *et al.*, 2010), and so on, can timely provide very useful information and insights for both basic research and drug development and hence are widely welcome by the science community.

Using image feature or graphic approaches to study biological systems is currently a hot topic in the biological and medical science because it can provide useful intuitive insights, as indicated by many previous studies on a series of important biological topics such as enzyme-catalyzed reactions (Chou, 1989), protein folding kinetics and folding rates (Chou, 1990), inhibition kinetics of processive nucleic acid polymerases and nucleases (Chou *et al.*, 1994) and others (Chou, 2010; Xiao *et al.*, 2005, 2006). These kind of approaches are also of great importance in the field of proteomics (Palagi *et al.*, 2006). The present study is an attempt to propose a new approach for the segmentation and quantification of two-dimensional gel electrophoresis images (2-DE).

The proteome was originally defined as the protein complement of the genome (Wasinger *et al.*, 1995). The proteome is complex and highly dynamic, since the proteins present in a sample will depend on e.g. the tissue, cell, organelle and/or development stage analyzed. Analysis of proteomes requires techniques that can simultaneously separate thousands of proteins in complex mixtures. Classical proteome analysis is based on separation of proteins by 2-DE, a technique originally described decades ago (O'Farrell, 1975), which has since been a subject to continuous developments enabling high resolution and reproducibility. In the first dimension of 2-DE, proteins are separated by isoelectric focusing during which they migrate along a pH gradient to their isoelectric point (pI). The second dimension separation, perpendicular to the first, separates proteins by sodium dodecyl sulpfate—polyacrylamide gel electrophoresis (SDS–PAGE) according to their molecular size. 2-DE is unique in its ability both to resolve differently modified forms of the same gene product in separate spots, facilitating analysis of protein post-translational modifications, and to enable simultaneous analysis of synthesis and amounts of hundreds of proteins in the same gel. Thus, 2-DE is a widely used and important tool in proteome analysis (Hecker *et al.*, 2008). In a typical 2-DE-based proteomics experiment, aimed at detecting changes in the proteome related to a specific treatment, several biological replicates each for treated and control samples are separated on 2-DE. Proteins on the gels may be visualized by one of many staining techniques (e.g. Coomassie blue dye, silver staining, fluorescent dyes, radiolabeling). In 2-DE, image analysis software is required for objective and quantitative comparison of 2-DE spot patterns, and several software packages are commercially available. Usually, 2-DE image pipeline analysis consists of:

- spot detection;
- alignment of multiple 2-DE images; and
- quantification of protein spots.

*To whom correspondence should be addressed.

Ideally, the analysis should involve minimum manual edition/supervision, which is both labor intensive and introduces subjectivity to the analysis. However, each stage is subject to complications due to the nature of the 2-DE technology. Protein stains have different detection limits, posing the problem of determining the threshold for spot detection. Due to the highly complex nature of biological samples, in most cases, thousands of protein forms will be present in a protein extract, and only the most abundant of these will be detected as spots on 2-DE, the remainder of which will be part of the background. Despite the high resolution of 2-DE, protein spots will often overlap partly or completely. Therefore, 2-DE spots are rarely of uniform, well-defined shape. Technical artifacts, streaking, either horizontal or vertical, can also disturb spot shape. Staining procedures and dust particles may give rise to background speckles. Such features of 2-DE images represent challenges for spot detection and quantification software.

2 AUTOMATIC ANALYSIS

Segmentation of the image is one of the most important steps in the computational analysis of 2-DE gels. Background and foreground are separated in this step, and areas in the gel will be accounted as belonging to the protein spot or discarded as background. Because all the remaining steps rely on the quality of the segmentation step, it is extremely important that this separation is performed as accurately as possible.

Early approaches (Appel *et al.*, 1997; Olson and Miller, 1988) for spot detection and gel segmentation involved the use of Laplacian of Gaussian techniques (LoG). LoG techniques are extremely sensitive to noise requiring an aggressive noise pre-filtering, which may result in the elimination of the weakest protein spots.

Nowadays, the most popular technique for spot segmentation is the Watershed Transform (*WST*) (Beucher and Lantuejoul, 1979). The *WST* is a technique used for more than 30 years with its origin in the field of mathematical morphology. It is a very powerful tool for image segmentation. Although computational power is not an issue in the present, there are other issues that prevent the *WST* from being an optimal approach. These will be addressed in the next sections.

2.1 Watershed transform

From the different algorithms (Roerdink and Meijster, 2000) for the *WST*, watershed by immersion simulation is one of the best for dealing with image plateaus. A plateau is an area where the neighbors of a pixel have the same gray level. It is important that the chosen algorithm is able to properly handle plateaus because, in this way, there is no need to transform the images being analyzed in *lower complete* images (Soille, 2003). Vincent and Soille define a recursive algorithm for the computation of the *WST* by immersion simulation. For more details, please see Vincent and Soille (1991).

Traditional approaches using the *WST* involve the use of the direct application of the *WST* on the, previously smoothed, image. Other approaches apply the *WST* on the gradient magnitude of the image. Both techniques are discussed in the next subsections.

2.1.1 Direct WST The *WST* 'sees' the image as a topographical surface so, the idea behind the *WST* by immersion is to simulate a flood of the topological surface, with water flowing from the regional minima of it. A dam is built whenever water coming from different catchment basins meet (see Fig. 1). The catchment basins, or simply basins, are the partitions in which the *WST* breaks the topological surface. At the end of the flooding process, the set of dams are the watershed lines or, simply, the watersheds. The *WST* presented by Vincent and Soille does not produce complete watershed lines. Thus, the result of this transform is post-processed in order to get the complete watershed lines.

The result of applying the *WST* directly to a 2-DE image is shown in Figure 3b. It is noticeable that the resulting watershed lines are dividing the image in catchment basins that do not represent exclusively the protein spots. There is a lot of background included in each catchment basin. Moreover, there are basins that have no protein spots at all.

2.1.2 WST of the gradient magnitude Let f be the gel image being processed. Applying the *WST* to the gradient magnitude of f is another typical approach. The gradient magnitude of f may be defined as:*vs.* 6

$$|\nabla_f| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2} \quad (1)$$

Figure 3c shows that there is a quite good definition of the protein spot contours. One of the problems from this approach is that the resulting image is over-segmented due to the high sensitivity of the $|\nabla_f|$, resulting in excessive minima from which the flooding process starts. It is common to apply a threshold defining the minimum gradient value to solve this problem. However, this is inefficient because over-segmentation will still remain in some areas, and in some other, correct segmentation will be compromised (dos Anjos and Shahbazkia, 2009). If holes were drilled on the topological representation of the $|\nabla_f|$, and only in the desired minima (the center of each spot, in this case) and on the background, the flooding would only occur from those locations, and the watersheds would only be built at the border proximity of the desired objects. This is known as Marker Controlled Watershed Transform. The markers can be set

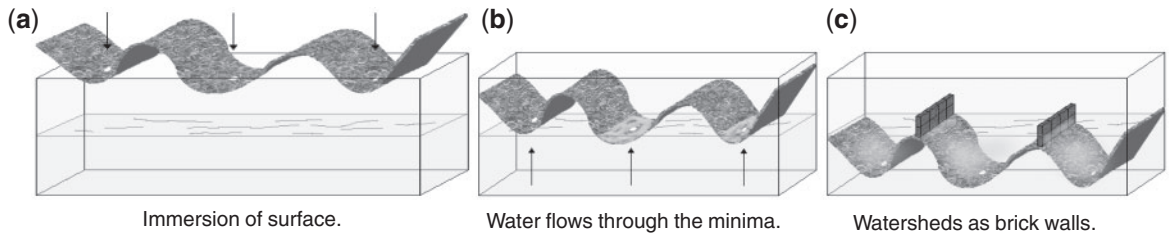


Fig. 1. Immersion simulation.

by, for example, using the mouse pointer. Nevertheless, on 2-DE images, it is impractical to mark all the spots manually due to the great number of spots in each gel and the great number of gels usually involved in an experiment. Therefore, automatic detection of quality markers is extremely important.

2.2 Other approaches

Obviously, commercial packages, such as Progenesis Samespots (Non-linear Dynamics), PD-Quest (Bio-Rad) and Melanie (GeneBio), are, or became closed source, preventing a proper analysis and comparison of their approaches with academic approaches. In these cases, the only comparison that can be made is in terms of visible results.

Gao *et al.* (2006) proposed a method for automatic detection of markers that consists of extracting minima from the low frequency components in the morphological gradient of the image. The h-minima transformation is used for that purpose. This transformation requires a parameter for minimum depth to be accounted and, although the gradient of the image is previously smoothed by a Butterworth low-pass filter of order 2, some valid regions, that have less depth than spurious ones, are inevitably destroyed.

Recently, Morris *et al.* (2008) presented an original approach named 'Pinnacle'. After the alignment of the images is performed by a third-party software, if the images are to be analyzed together, this approach consists succinctly of the following steps:

- Compute an average gel;
- Denoise the average gel;
- Detect pinnacles in the denoised gel;
- Combine pinnacles within a windowed defined proximity;
- Quantify each spot by extending a window, of defined size, centered in the pinnacle;
- Background correction and normalization of spot quantification.

The window sizes used in defining the pinnacle proximities, and in the spot quantification are manually defined parameters.

3 PROPOSED TECHNIQUE

When the gel images suffer from high-frequency noise, as is usually the case, a convolution of the image with a low-pass filter will increase the quality of the results. The images used in this study were pre-filtered by a Gaussian kernel ($0.0 < \sigma \leq 4.5$), and this is the only pre-processing needed in order to obtain good results using the proposed method.

3.1 Initial watershed

As mentioned in Section 2.1.1, the basins delimited by the watersheds include a lot of background surrounding the protein spots. Also, there are many areas that do not contain a spot (see Section 2.1.1). Even if the result of the segmentation performed by the direct *WST* of f is not the intended result, a lot of useful information can be extracted from it that may be used for further segmentation. For example, a very important piece of information provided by the direct *WST* is that there is only one protein spot in each of the basins. This may not be entirely true due to saturation, limitations of the scanner or even due to the limitations of the staining methods of the gel. When two saturated spots are partially overlapped, a separation by the *WST* will not occur. That is a problem that can be solved afterward (e.g. by watershedding the distance transform) and it is not in the scope of this article. Completely overlapped spots are impossible to separate using current image processing techniques. In these cases, only running the sample through a narrower gradient gel, or cutting the spot from the gel and submitting it to mass spectrometry analysis, will allow to resolve the overlapping proteins. Thus, from now on, it will be assumed that each basin contains only one protein spot.

3.2 Automatic basin validation

A great number of the basins without any protein spot may be rejected by setting a threshold for minimum spot area. All the basins that are smaller than the minimum spot area are ignored in the subsequent steps.

The remaining invalid basins cannot be discarded as easily. Theoretically, the lines of the initial watershed are crossing only the background. Given this, a synthetic background b is generated by interpolation of these watershed lines, providing an extra layer of validation. Linear interpolation is used horizontally and vertically, then the average between these two interpolated backgrounds is used to produce the final synthetic background (see Supplementary Appendix A). For a better approximation of the possible real background, one can interpolate the watershed lines in other directions and, finally, calculate the average. Bi-linear, bi-cubic or other orders of interpolation can be used for a more accurate background synthesis. Although it would produce better results for the possible background, the drawback is that it requires more computational power with minimal advantages. Lieber and Mahadevan-Jansen (2003) use another possible approach for background approximation, where a sequence of polynomial fittings and subtractions are performed iteratively.

Using the generated background, the validation of the basins is performed in the following way:

Let $g = \bigcup_{i=1}^{n} g_i$, be the set of all basins, where n is the number of basins present in image f. Also, let $b = \bigcup_{i=1}^{n} b_i$, be the synthetic background, with b_i being the connected component corresponding to the area of basin g_i. After the generation of the synthetic background b, the decision of whether or not a basin contains a protein spot can be done by comparing the difference between the SD σ of each basin g_i and the SD of each of the respective areas b_i of the generated background b, in the following way:

$$s(i) = \begin{cases} \text{True}, & \text{if } \sigma_{g_i} - \sigma_{bi} > \delta \\ \text{False}, & \text{otherwise.} \end{cases} \quad (2)$$

where $s(i)$ is a boolean function that indicates if basin g_i contains a valid protein spot or not. Figure 2 shows how sensitivity to protein spot detection may be controlled using different values of δ. It can be seen that the sensitivity to spots decreases as the parameter δ grows.

3.2.1 Placing the markers After the valid basins are selected, holes are 'drilled' in the topological surface of $|\nabla_f|$. It is clear that, the image of the $|\nabla_f|$ should be pierced at the same position as the center of each protein spot existing in image f. The center of the spot is usually the minima of the component but, due to the existence

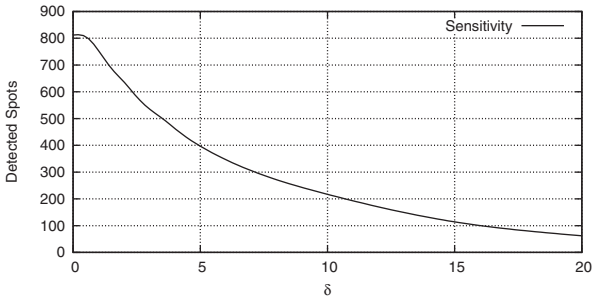


Fig. 2. Smaller δs result in higher sensitivity to spot detection.

of plateaus and, in some cases, the minima being placed side by side with the watershed lines, another approach is to apply a simple regional threshold (global with respect to the basin) to each of the valid basins and, then, find the centroid of the resulting component. Then, the centroid may be used as a marker of the spot center. It is not extremely important that the marker is set precisely at the center of the protein spot, as long as the marker is not placed on, or outside the borders (the watershed lines to be found) of the object.

3.3 Final watershed

The *WST* is used again but, this time, to flood the surface of the $|\nabla_f|$ image from the center of the protein spots and from the watershed lines of the initial *WST*. Now the water coming from different sides meets at the top of the 'craters' that represent the edges of the protein spots in image f. As there is a great number of other minima produced as result of the sensitivity to noise of the $|\nabla_f|$ image, and there is no such thing as *WST* from minima, the image has to be transformed in such a way that the only existing minima are the ones pointed by the markers, that is, the watershed lines of the initial *WST* and the markers of the valid protein spots.

The solution is to morphologically reconstruct the $|\nabla_f|$ surface by minima imposition. This technique consists of reconstructing a mask image from a marker image (See Supplementary Appendix B).

Let $R^{\epsilon}{}_{v}(u)$ be the reconstruction by erosion of a mask image v from a marker image u with domains $D_u = D_v$ and $u \geq v$. Let M be the set of markers that identify valid spots, and $h_{\max}$ be greatest possible intensity level allowed by the image in use. Let us use $|\nabla_f|$ as the mask image, and the marker image f_m be defined for each pixel p as:

$$f_m(p) = \begin{cases} 0, & p \in M \text{ or } p \in \mathrm{WST}(f) \\ h_{\max}, & \text{otherwise.} \end{cases} \quad (3)$$

The reconstruction R is now performed by erosion ϵ of the marker image f_m with respect to the mask image $|\nabla_f|$. In some situations, two or more distinct minima may fall within a plateau of $|\nabla_f|$ at the gray level 0 (Soille, 2003), so reconstruction should be made with respect to the image $(|\nabla_f|+1)$. As this may violate the requirement that $(|\nabla_f|+1) \geq f_m$, thus demanding the use of self-dual reconstruction, the final mask is defined as $(|\nabla_f|+1) \wedge f_m$, where $\wedge$ is the point-wise minimum operator. The watershed lines *WS*, that represent the edges of the protein spots, are the lines resulting from applying the *WST* in the following way:

$$\mathrm{WS} = \mathrm{WST}(R^{\epsilon}{}_{(|\nabla_f|+1)\wedge f_m}(f_m)) \quad (4)$$

The final result is presented in Figure 3e. As can be seen, if compared with the initial watershed in Figure 3b, or the watershed of the gradient magnitude of f in Figure 3c, there is a great improvement in the segmentation of the image. Moreover, the segmentation was performed without the intervention of the user to select and place markers on the protein spots existing in the image.

3.4 Spot delineation

The method described in this article for spot detection and delineation finds a closed curve around the spot. The position of this curve is where the spot has a second-order derivative of zero. Some algorithms use the integrated density inside the delineation curve as a measure of spot volume. Although consistent this clearly underestimates the true spot volume. Others try to normalize the spots to give a less biased estimate of spot volume. On the other hand, this may be sacrificing variance. Here, we propose to use a property of the 2D Gaussian as a possible normalizing factor. For a 1D Gaussian, the points with second-order derivative equal to zero are called the deflection points. These occur exactly one SD away from the mean. This is illustrated in Figure 4a where the area under the curve between the deflection is seen to be 68.27%. For a 2D Gaussian, deflection points become a curve, namely a circle or an ellipse. In Figure 4b, a 2D Gaussian has been truncated at the deflection curve. The resulting volume corresponds to 39.35% of the full 2D Gaussian. For a spot in a 2-D electrophoretic gel, a simple way to estimate the volume of the full spot is by multiplying the volume found by the method described in this article by the constant $\frac{1}{0.3935}$ or 2.54. Nevertheless, as protein spots only in very rare cases represent closely a 2D Gaussian (Rogers *et al.*, 2003), this is an operation that may be considered controversial. Therefore, we leave the choice of normalization or not up to the user.

4 RESULTS

Tests were performed in order to assess the quality of the proposed approach. A state-of-the-art commercial software for 2-DE analysis and a recent academic approach were used as base of comparison. The chosen commercial software was Progenesis Samespots (Nonlinear Dynamics, version 3.3.3420.25059), a very well-known tool in the field of proteomics. As for the academic approach, we used Pinnacle (Morris *et al.*, 2008). When analyzing images derived from biological samples, it is always necessary to validate detected spots manually to avoid inclusion of the background noise. In this study, the spots detected by any software in the whole set of images were manually checked by an experienced proteomics researcher and defined, based on experience, as protein spots ('valid positives'), or as 'false positives', resulting from staining artifacts giving rise to background noise or to erroneous definition of spot boundaries because of the overlapping spots.

4.1 Detection

For evaluation of spot detection, five gel sections of varying quality were used (see Fig. 3 in the Supplementary Appendix c), each of which contained 100–250 spots. It is important to underline that Gel #5 was of extremely low quality, having a lot of noise (see Figure 3(e) in the Supplementary Appendix c). In Progenesis and in the proposed approach, the spots were thresholded, by size, in two ways. First, the area (number of pixels) of the smallest protein

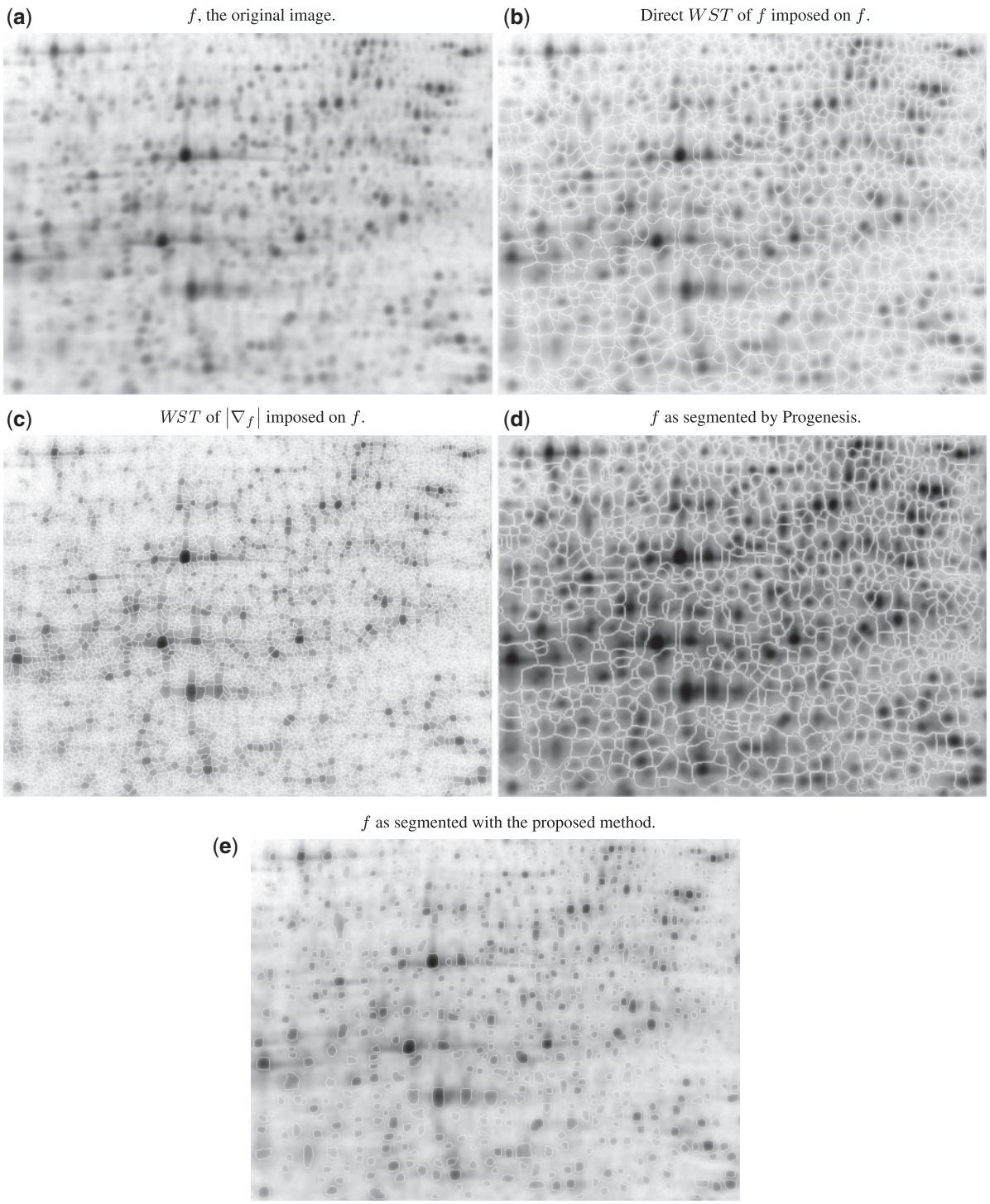


Fig. 3. Image f versus watersheds approaches imposed on f.

spot, in each gel, was chosen as a cutoff filter in both Progenesis Samespots and the proposed approach. The idea behind this option was to minimize the false positive spots.

The selected cutoff filters used for gels 1–5 were of 2, 7, 7, 7 and 4 pixels, respectively. In Pinnacle, as there is no parameter to define a threshold for the minimum spot size, the used parameters were the default. The comparison of the results can be observed in Table 1.

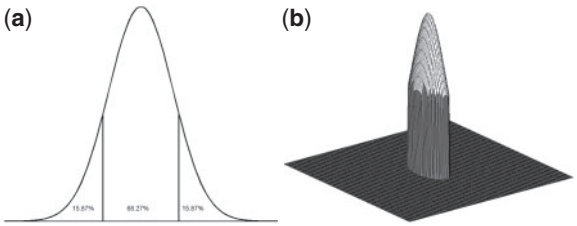
In the second test for spot detection, the 'optimal' cutoff filter was manually identified for Gel #1, separately for Progenesis Samespots and for the proposed method. The same filter was then applied to

all gels. This was of 199 pixels for Progenesis Samespots and 6 pixels for the proposed approach. As for Pinnacle, the parameters 'Minimum Pinnacle Size (MPS)' and 'Neighborhood Size (NS)' were tuned in the following way: Gel #1: MPS = 0.1; NS = 10; Gel #2: MPS = 0.4; NS = 5; Gel #3: MPS = 0.5; NS = 8; Gel #4: MPS = 0.2; NS = 15; Gel #5: MPS = 0.6; NS = 8.

The result of segmenting the gels with this parameters is presented in Table 2.

When using the real size of the smallest valid spot in the gel as a cutoff filter, Progenesis Samespots overestimates the number of possible detected spots as presented in Table 1. This is due to the fact that Progenesis Samespots exaggerates the area of the spots, in part, as result of using common spot areas for all gels. When overestimating, one has to manually remove a lot of false positives, which is laborious. If the cutoff filter is adjusted to fit the individual program, the two packages will perform roughly the same (See Table 2).

As Table 2 shows, in Gel #5, the proposed approach will detect a number of false positives that is near to Progenesis result. As stated before, this was a very low-quality gel with a lot of noise and the cutoff filter was the same that was defined for Gel #1 which was of 6 pixels of area.*vs.* -3



1-D Gaussian and deflection points.

2-D Gaussian truncated at the deflection curve.

Fig. 4. Deflection points and truncated volume.

4.2 Quantification

To evaluate the quantification of spot volumes, three gel sections, originated from two biological replicates of barley leaf extracts, were used. The plants were grown in the presence or absence of nitrate in order to induce changes in their protein profiles. The volumes of 20 randomly selected spots were quantified using Progenesis, Pinnacle and the proposed approach, and the fold-change between samples, grown with and without nitrate, was determined. The results are presented in the table provided as Supplementary Material. When using Progenesis Samespot, the user selects a reference gel against the rest of the gels that are aligned. In this case, each replicate gel was used in turn as the reference, to test whether similar results would be obtained. The analysis conducted by Progenesis Samespot depends on which gel is chosen as reference (ProgA and ProgB), whereas Pinnacle and the proposed approach stayed consistent. It is noticeable from looking at the results of Tables 1 and 2, that in any of the situations Pinnacle will miss a lot of valid spots. All the presented methods obtain different results, which is not a surprise, as reported by Stessl *et al.* (2009), where it is shown that it is common that different software usually present different results, and, as the same source mentions, different versions of the same software brand also disagree in the presented results.*vs.* -3

5 DISCUSSION

When comparing protein profiles with Progenesis Samespot, using a reference gel that defines the spot area for all gels, the end-result will

Table 1. Spot detection comparison

Gel number	G.T.	Detected			True positives			False positives			Missed		
		Prog	Pinn	Prop	Prog	Pinn	Prop	Prog (%)	Pinn (%)	Prop (%)	Prog (%)	Pinn (%)	Prop (%)
1	204	279	102	187	194	101	180	42	0	3	5	50	12
2	199	316	112	190	197	107	181	60	3	5	1	46	9
3	235	401	102	210	228	101	206	74	0	2	3	57	12
4	112	300	58	98	111	47	96	169	10	2	1	58	14
5	163	753	99	219	162	84	153	363	9	40	1	48	6

Used cutoff: progenesis and proposed = minimum spot size; pinnacle = default values. 'G.T.' is the ground truth defined by the biologist, 'Prog', 'Pinn' and 'Prop' refer to Progenesis, Pinnacle and Proposed method, respectively.

Table 2. Spot detection comparison

Gel number	G.T.	Detected			True positives			False positives			Missed		
		Prog	Pinn	Prop	Prog	Pinn	Prop	Prog (%)	Pinn (%)	Prop (%)	Prog (%)	Pinn (%)	Prop (%)
1	204	229	222	203	176	165	185	26	28	9	14	19	9
2	199	241	145	220	191	126	188	25	10	16	4	37	6
3	235	259	127	252	203	106	220	24	9	14	14	55	6
4	112	172	95	115	96	62	104	68	29	10	14	45	7
5	163	259	163	282	146	115	154	69	34	79	10	29	6

Used cut-off: progenesis and proposed = the adjusted minimum spot size that provides best results for gel number 1; pinnacle = 'optimal' parameters.

Table 3. Fold change comparison

Spot Number	Gel Number 1				Gel Number 2				Gel Number 3			
	ProgA	ProgB	*Pinn*	Our	ProgA	ProgB	Pinn	Our	ProgA	ProgB	Pinn	Our
01	1.0	0.8	0.1	0.7	1.0	N/A	0.9	0.8	1.0	1.1	1.9	1.0
02	1.0	0.8	∞	0.8	0.9	0.9	1.0	1.3	1.1	1.3	0.5	0.8
03	0.9	1.0	∞	1.0	0.9	0.9	1.3	0.9	0.8	0.9	1.5	1.4
04	1.2	1.4	1.3	0.8	0.9	N/A	0.1	1.1	0.8	1.2	−∞	2.5
05	1.2	1.0	1.1	0.9	0.9	0.7	∞	0.9	0.8	0.8	0.6	1.5
06	1.3	1.1	1.0	0.9	0.9	0.9	0.8	0.9	1.2	1.9	∞	2.2
07	0.8	0.8	6.3	0.7	1.2	1.0	0.1	1.2	1.3	0.6	0.5	1.6
08	1.3	1.3	4.5	1.0	0.8	1.0	−∞	0.8	0.7	0.7	1.0	1.5
09	1.3	1.0	−∞	1.2	0.8	N/A	∞	1.0	1.4	1.5	1.1	2.2
10	1.4	1.3	∞	1.6	1.2	1.3	1.0	1.5	0.7	0.7	0.6	1.1
11	1.4	1.1	3.5	1.0	1.2	1.4	∞	1.5	0.7	0.9	1.1	0.8
12	1.5	N/A	∞	1.4	0.8	0.8	∞	1.3	1.5	1.6	16.2	2.6
13	1.5	1.2	6.3	1.0	0.8	0.9	−∞	0.7	0.7	0.8	0.6	0.6
14	1.5	1.2	1.0	1.1	0.8	N/A	0.8	0.6	0.7	0.9	0.4	1.2
15	1.7	1.7	N/A	2.0	0.8	0.7	∞	0.8	1.6	1.7	1.1	2.3
16	1.8	1.4	∞	2.1	0.8	0.7	0.9	0.8	0.6	0.7	N/A	1.1
17	1.8	N/A	1.5	1.4	1.3	1.8	2.1	1.8	0.5	0.6	−∞	N/A
18	1.9	1.3	N/A	N/A	0.7	0.7	−∞	0.7	2.4	N/A	0.7	1.5
19	2.0	1.9	N/A	N/A	1.4	N/A	∞	2.0	2.4	1.8	∞	5.2
20	2.9	2.3	8.0	N/A	0.7	0.7	0.7	0.6	2.4	2.1	0.4	1.3

'ProgA' and 'ProgB' are the results from progenesis with different gels chosen as reference. N/A means that the spot was not detected on any of the gels. ∞ and −∞, mean that the spot was only detected in the first or the second gel, respectively.

depend on which gel is chosen as reference. In Progenesis, it is most likely that the user chooses the gel with most round uniform spots, because it will be easier to align this to the others. The problem arises when the other gels/spots vary in quality with respect to shape and size. Having defined the spots' areas in the 'best' gel, and using these definitions in the other gels, if the analysis is repeated with a new reference gel, different results will be provided to the user, as demonstrated by the results presented in Table 3. Since the proposed approach independently defines the spot area in each gel, the program is more likely to give less-biased end-results for being reference-gel independent.

Another issue is how the spot area is detected. Progenesis uses spot and background to define the spot area. Pinnacle centers a fixed size window on the pinnacle of the spot. The proposed approach defines the spot area from near the border of the spots and, therefore, avoids potential problems with less uniform spots in between gels. Furthermore, by defining the spot area from the border of the spot, it is easier to get at a more realistic estimate of spots placed close to each other as well as partially overlapping non-saturated spots, which will be present in most experiments.

In comparison with the other software packages tested, the proposed method is shown to be able to detect the majority of validated protein spots in a series of gel images, while at the same time minimizing the number of false positives. This represents a considerable improvement in terms of the time needed for manual validation and correction of spot boundaries, a common bottleneck in proteomics studies.

When analyzing samples based on biological material with unknown absolute quantities of individual protein forms, it is not possible to say which is the 'true' answer. Other software packages also give slightly different results, but again, it is not possible to say from this type of analysis which is the 'correct' answer. A complete gel image is included in the support data, segmented by Progenesis Samespot, PDQuest and proposed method for comparison. It is noticeable that PDQuest is the result that requires more user intervention and manual editing, as also witnessed by other studies (Arora *et al.*, 2005).

In summary, this article presents an efficient way of automatic detection of quality markers to use in conjunction with the Marker Controlled Watershed Transform or even with the Image Foresting Transform (Falcao *et al.*, 2004) for the segmentation of 2-DE images and quantification of protein spots. Also, the presented technique can be used in the segmentation step for approaches that use common spot boundaries such as Rye *et al.* (2008) or even Progenesis Samespots. The results can also be used by gel alignment for gel alignment methods such as the one presented in Pèrés *et al.* (2008). Post-processing the results produced by the method proposed in this article, using the domain expertise, will allow the correction of most of the discussed shortcomings but, as stated before, the topic of post-processing is not in the scope of this article. This article contributes to the field of 2-D gel-based proteomics with an alternative approach that may compete with Progenesis Samespot and other important academic approaches in the detection of spots and the estimation of the spot volume for gel comparison.

ACKNOWLEDGEMENT

We thank Drugmode (Odense, Denmark) for providing part of the gel material used for testing in this project.

Funding: Portuguese Foundation for Science and Technology (grant SFRH/BD/19233/2004); Danish Research Council for Natural Sciences (FTP).

Conflict of Interest: none declared.

REFERENCES

Appel,R.D. *et al.* (1997) Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: II. algorithms. *Electrophoresis*, **18**, 2735–2748.

Arora,P.S. *et al.* (2005) Comparative evaluation of two two-dimensional gel electrophoresis image analysis software applications using synovial fluids from patients with joint disease. *J. Orthop. Sci.*, **10**, 160–166.

Beucher,S. and Lantuejoul,C. (1979) Use of watersheds in contour detection. In *International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation*. Rennes, France.

Cai,Y.-D. *et al.* (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J*., **84**, 3257–3263.

Cai,Y. *et al.* (2010) Predicting protein subcellular locations with feature selection and analysis. *Protein Pept. Lett.*, **17**, 464–472.

Chen,L. *et al.* (2010) Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. *BMC Bioinformatics*, **11**, 293.

Chou,K.-C. and Cai,Y.-D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **277**, 45765–45769.

Chou,K.-C. and Shen,H.-B. (2010) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS ONE*, **5**, e9931.

Chou,K.C. (1989) Graphic rules in steady and non-steady state enzyme kinetics. *J. Biol. Chem.*, **264**, 12074–12079.

Chou,K.C. (1990) Applications of graph theory to enzyme kinetics and protein folding kinetics. steady and non-steady-state systems. *Biophys. Chem.*, **35**, 1–24.

Chou,K.-C. (2010) Graphic rule for drug metabolism systems. *Curr. Drug Metab.*, **11**, 369–378.

Chou,K. *et al.* (1994) Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal. Biochem.*, **221**, 217–230.

dos Anjos,A. and Shahbazkia,H. (2009) Automatic marker detection for blob images. In *WACV'09: Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 1–6.

Falcao,A.X. *et al.* (2004) The image foresting transform: theory, algorithms, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 19–29.

Gao,L. *et al.* (2006) A New Marker-Based Watershed Algorithm. *TENCON 2006. IEEE Region 10 Conference*. pp. 1–4.

Hecker,M. *et al.* (2008) Gel-based proteomics of gram-positive bacteria: a powerful tool to address physiological questions. *Proteomics*, **8**, 4958–4975.

He,Z. (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One*, **5**, e9603.

Lieber,C.A. and Mahadevan-Jansen,A. (2003) Automated method for subtraction of fluorescence from biological raman spectra. *Appl. Spectrosc.*, **57**, 1363–1367.

Morris,J.S. *et al.* (2008) Pinnacle: a fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data. *Bioinformatics*, **24**, 529–536.

O'Farrell,P.H. (1975) High resolution 2-D electrophoresis of proteins. *J. Biol. Chem.*, **250**, 4007–4021.

Olson,A.D. and Miller,M.J. (1988) Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Anal. Biochem*., **169**, 49–70.

Palagi,P.M. *et al.* (2006) Proteome informatics I: Bioinformatics tools for processing experimental data. *Proteomics*, **6**, 5435–5444.

Pèrés,S. *et al.* (2008) A new method for 2D gel spot alignment: application to the analysis of large sample sets in clinical proteomics. *BMC Bioinformatics*, **9**, 460.

Roerdink,J.B.T.M. and Meijster,A. (2000) The watershed transform: definitions, algorithms and parallelization strategies. *Fundam. Inf.*, **41**, 187–228.

Rogers,M. *et al.* (2003) Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images. *Proteomics*, **3**, 887–896.

Rye,M.B. *et al.* (2008) A new method for assigning common spot boundaries for multiple gels in two-dimensional gel electrophoresis. *Electrophoresis*, **29**, 1359–1368.

Soille,P. (2003). *Morphological Image Analysis: Principles and Applications*. Springer, New York, Inc., Secaucus, NJ, USA.

Stessl,M. *et al.* (2009) Influence of image-analysis software on quantitation of two-dimensional gel electrophoresis data. *Electrophoresis*, **30**, 325–328.

Vincent,L. and Soille,P. (1991) Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, **13**, 583–598.

Wasinger,V.C. *et al.* (1995) Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis*, **16**, 1090–1094.

Xiao,X. *et al.* (2005) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. *J. Theor. Biol.*, **235**, 555–565.

Xiao,X. *et al.* (2006) A probability cellular automaton model for hepatitis B viral infections. *Biochem. Biophys. Res. Commun.*, **342**, 605–610.