Published in final edited form as:

Comput Stat Data Anal. 2011 November 1; 55(11): 3059–3072. doi:10.1016/j.csda.2011.05.013.

Region-based Statistical Analysis of 2D PAGE Images

Feng Li^{a,1,*}, Françoise Seillier-Moiseiwitsch^b, and Valeriy R. Korostyshevskiy^c

Feng Li: feng.li@fda.hhs.gov; Françoise Seillier-Moiseiwitsch: fseillier@idcrp.org; Valeriy R. Korostyshevskiy: vrk@georgetown.edu

^aDepartment of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland, USA

^bInfectious Disease Clinical Research Program, Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

^cDepartment of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, Washington, DC, USA

Abstract

A new comprehensive procedure for statistical analysis of two-dimensional polyacrylamide gel electrophoresis (2D PAGE) images is proposed, including protein region quantification, normalization and statistical analysis. Protein regions are defined by the master watershed map that is obtained from the mean gel. By working with these protein regions, the approach bypasses the current bottleneck in the analysis of 2D PAGE images: it does not require spot matching. Background correction is implemented in each protein region by local segmentation. Two-dimensional locally weighted smoothing (LOESS) is proposed to remove any systematic bias after quantification of protein regions. Proteins are separated into mutually independent sets based on detected correlations, and a multivariate analysis is used on each set to detect the group effect. A strategy for multiple hypothesis testing based on this multivariate approach combined with the usual Benjamini-Hochberg FDR procedure is formulated and applied to the differential analysis of 2D PAGE images. Each step in the analytical protocol is shown by using an actual dataset. The effectiveness of the proposed methodology is shown using simulated gels in comparison with the commercial software packages PDQuest and Dymension. We also introduce a new procedure for simulating gel images.

Keywords

proteomics; 2D PAGE; watershed region; spot matching; statistical image analysis; high dimension

1. Introduction

Proteomics is the science that deals with high-throughput separation and quantification of proteins. Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) is a powerful

* Corresponding address: Division of Biometric II, Office of Biostatistics, Office of Translational Science, Center for Drug Evaluation and Research, Food and Drug Administration, 10903 New Hampshire Ave, Room 3665, Building 21, Silver Spring, MD, 20993, USA. Tel.: 3017964974.

¹This article reflects the views of the author and should not be construed to represent FDA's view or policies.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

technique for separating complex mixtures, where thousands of proteins are separated and measured simultaneously (O'Farrell, 1975). In 2D PAGE, proteins are separated in one dimension according to their isoelectric charge and in the orthogonal dimension according to their molecular weight. Each protein is uniquely determined by its location along these two dimensions. For the first dimension, proteins are focused electrophoretically along a pH gradient. The movement of a protein stops when it reaches a position at which it has no net isoelectric charge. For the second dimension, proteins are soaked in sodium dodecylsulfate (SDS) so that all proteins acquire the same charge density, and placed on a polyacrylamide gel for electrophoretic separation. As a final step, the separated proteins are stained with fluorescent dyes so that they are amenable to imaging. The images are scanned and stored in a database. Proteomic differences across multiple samples are studied by comparing the expression profiles of proteins across sets of gels.

Figure 1 shows typical images of 2D gels. Each dark spot with a smooth contour represents a different protein. The gray intensity of a spot is proportional to the quantity of the corresponding protein present on the gel. By comparing spot intensities across images, we are able to compare amounts of the same protein under different conditions (e.g., treatments, diseases, exposures) and identify the protein spots whose volume changes across groups. It would be unwieldy to do this manually since there are thousands of spots to compare, and gels undergo distortions during the experimental process.

The main steps in the differential analysis of 2D gels are as follows (Roy et al., 2003). First, one applies procedures to filter out various kinds of noise present on the images. Image segmentation techniques are then utilized to detect protein spots. Next, one aligns the images: gels may have been differentially warped due to physical phenomena occurring during the lengthy process of the assay. Once the alignment has been completed according to some set rules (usually involving landmarks on a reference gel), spots on different gels are matched. The volume (i.e., the sum of the pixel intensities) or the mean pixel intensity of each spot is computed. The final step is the analysis of these quantitative features and their variability to identify changes across conditions. Although there are commercial software packages for 2D gel image analysis (e.g., PDQuest, Dymension), considerable human intervention is needed for spot localization and matching. Spot matching is the process by which one maps a spot on a particular gel to the corresponding spots on the other gels so that spots corresponding to the same protein are identified and labeled. Spot matching has indeed been recognized as the bottleneck in 2D PAGE image analysis (Morris et al., 2008). Moreover, in commercial software, the comparison of the quantitative features relies on basic tests, such as the t-test.

Statistical differential analysis of 2D gel images is still in its infancy. The main difficulties are the discrimination between actual protein spots and noise, the quantification of protein expression levels thereafter, and spot matching for comparison between groups of samples. Models based on image pixels (Conradsen and Pedersen, 1992) are not practical given the huge number of pixels, high variation in the background (across images and within an image) and sensitivity to misalignment. Instead of comparing spot volumes, Seillier-Moiseiwitsch et al. (2011) developed a systematic procedure for statistical analysis of 2D PAGE images based on wavelet coefficients. A gradient method was used for background correction and a fixed constant proportion of the wavelet coefficients for each gel image was extracted. An analysis of variance (ANOVA) was applied to all elements of the aggregate wavelet coefficient vector. Since there are thousands of wavelet coefficients, this approach leads to a large-scale simultaneous hypothesis testing problem. Adjustment procedures were implemented. Using only those wavelet coefficients with significant p-values, the gel images were reconstructed and differences due to a specific factor in the experimental design were therefore automatically highlighted. This procedure suffers from the following limitations.

The threshold intensity for background correction is not easily determined. Unless a powerful alignment algorithm is utilized, the wavelet coefficients at the same location across images may not correspond to the same protein spot because of misalignment of the images. Therefore, the highlighted differences may not be meaningful. Moreover, the strong spatial correlation among wavelet coefficients spanning the same protein spot render methods for multiplicity adjustment, such as the control of the false-discovery rate (FDR) via the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995), less powerful.

Instead of working with wavelet coefficients or individual spots, we consider protein regions as our analytical units. The whole gel image is divided into different regions using the watershed algorithm (Vincent and Soille, 1991). A key feature of our region-based analysis is that protein regions are defined by a master watershed map, which consists of the watershed boundaries obtained from the mean gel image. The master watershed map is overlaid onto each individual gel image for protein separation and quantification. While most of the proteins are separated into different watershed regions, a few regions may contain overlapping spots. Two overlapping spots are considered as a single entity since the objective of the analysis is to identify differences due to a factor of interest (e.g., disease status, exposure). The pixels in each watershed region are then classified as being part of either the object or the background. A summary statistic for each protein region is obtained as a means of quantification for the comparative analysis. We investigate systematic trends in the summary statistics and propose a new normalization procedure. After normalization, hypothesis testing is conducted on the regional statistics. The regions showing interesting differences across groups are highlighted for subsequent analysis. The usual BH procedure for controlling the FDR is not efficient here since there are high correlations among subsets of proteins. Evidence shows that such correlations can have a strong impact on statistical inference (Qiu et al., 2005; Efron, 2004, 2007a, 2007b). Tibshirani and Wasserman (2006) averaged the test statistics corresponding to correlated genes to get more reliable inference. We implement a permutation method to select regions that are highly correlated. Based on the detected correlations, we separate proteins into independent sets. We use MANOVA tests for correlated protein regions and ANOVA tests for independent regions. The final p-values are considered together to find protein regions with significant changes across groups.

To illustrate the procedures, we utilize twelve 2D PAGE images that were generated to investigate the effect of nicotine exposure on the proteome of spleen cells of female and male rats. The experiment follows a 2×2 balanced factorial design with gender and exposure as factors and 3 subjects for each treatment-gender combination. Figure 1 shows four images from this dataset: the top row from the control group, the bottom row from the exposure group, left column from female rats and the right column from male rats. The images were aligned using the algorithm developed by Potra et al. (2006). The objective is to find the proteins whose expression is either sensitive to nicotine or gender dependent. Subsequently, the sequence composition of differentially-expressed proteins will be determined by mass spectrometry and these sequences will be compared to a sequence database to identify them and investigate their functions.

The organization of this paper is as follows. In Section 2, the separation and quantification of protein regions are discussed. In Section 3, the normalization of summary statistics for protein regions is described. In Section 4, we proceed with identifying mutually independent protein sets (each set containing correlated proteins), and with the statistical testing of these sets. Actual gel images are often not suitable for comparing different methodologies since the ‘interesting’ proteins are not known and thus rates of false-positive and false-negative discovery cannot be estimated. Therefore, simulated 2D PAGE images are useful for the purpose of evaluating analytical procedures, but only if the simulated images are similar to

real ones. Simulating images that mirror reality is a challenging task. In Section 5, we introduce a simulation method based on random perturbation of a set of real gels. On simulated images, the proposed analytical protocol outperforms two commercial software packages (PDQuest 8.01 and Dymension 2.0.3.0).

2. Delineation and Quantification of Protein Regions

2.1. Protein Regions

Once the images are aligned, dark spots corresponding to the same protein are approximately at the same location on different images. However, some displacement may be evident, as in Figure 2. Spot matching is the procedure by which one identifies the same protein on different gel images. Due to the spots' irregular shapes and wide intensity range, this step requires considerable scrutiny and manual adjustment. Instead of striving to get a perfect alignment and accurate quantification of protein spots, we focus on finding meaningful differences between gel images due to factors of interest.

To avoid the above-mentioned limitations and to improve the reliability of the overall analysis, we separate proteins into different regions and compare protein regions across experimental groups. Regions with statistically significant differences will be highlighted for further investigation. Ideally, there should be only one protein spot in each region, but this is not necessarily the case since some spots overlap. However, since we compare regional features, this will not affect our procedure. Indeed, if one of the two spots in a region changes significantly, this will be reflected in the regional summary statistic.

The first step in our procedure is to separate protein spots into different regions. Considering spots as local minima of the images, a natural choice for image segmentation is the watershed algorithm (Vincent and Soille, 1991). This is a morphological image-processing algorithm that is effective at separating local minima. The intuitive concept underlying the watershed algorithm comes from the field of topography. The pixel intensities of a dark spot are lower than those of the local background. Consider a gel image as a 3D surface with the protein spots as local minima. Imagine the surface being flooded from holes pierced at the local minima. Water fills up basins starting at these local minima, and, at points where waters coming from different basins meet, dams are built. As a result, the surface is partitioned into regions or basins separated by dams, called watershed lines. Figure 3 shows the separation into protein regions using the watershed algorithm, from which we can see that the watershed boundary can separate spots well even when they overlap.

Note that when we construct watershed regions for each gel image, the watershed boundaries vary in location across gels given the variation in background and possibly different number of proteins. Therefore, the same proteins may be in different watershed regions on different gel images. To make sure the same proteins are compared, one would need a region-matching procedure similar to the time-consuming spot-matching procedure used in commercial software. To avoid such procedure, we propose to use a master watershed map on all images. The chosen master watershed map consists of the watershed boundaries constructed by applying the watershed algorithm to the mean image of all gels (Li, 2007; Li and Seillier-Moiseiwitsch, 2011). The Pinnacle method developed by Morris et al. (2008) shares the same idea in that it also uses the mean gel image for spot detection. The main difference is that the Pinnacle method uses the local minimum (i.e., pinnacle) within a moving pre-specified rectangular region for spot quantification. In our method, instead of extracting spot features using a moving area of fixed dimensions, the watershed algorithm is used to automatically define the region surrounding a spot.

Gel images may have very different intensity ranges due to, for instance, the image scanning procedure. To ensure that all gel images contribute equally to the mean image, it is often necessary to scale them to the same intensity range (such as [0,1], [0, 255] or [0, 4095]) before calculating the mean image. The mean image contains protein spots from all gel images. We smooth the mean image using a one-level wavelet approximation and apply the watershed algorithm to the smoothed mean image to get the master watershed boundaries. Then, these boundaries are imposed on each gel image. In this way, watershed regions with a common identification number correspond to the same location on different gel images. Since the mean image contains spots with mean intensity, the master watershed boundaries should separate spots on each gel image well. By good separation, we mean that an isolated spot should be contained fully within the corresponding master watershed region.

Figure 4 shows the same sections of the images as in Figure 1 after separation using the master watershed regions. Clearly, the master watershed boundaries provide equally good separation for images from different experimental conditions.

2.2. Protein Region Quantification

To compare protein regions across experimental conditions, a summary statistic can be used to capture the intensity and reduce data dimensionality. Since the intensity of a spot is affected by its local background, we first need to perform a local background correction on each region.

Background correction consists of two steps: classification and quantification. All pixels in the selected region are classified as either background or object pixels. Then, either the difference in medians or means between the object and the background is used as the quantification of a spot. The classification techniques are based on histograms or morphology. Because of the irregular shape of the watershed regions and reliability as well as simplicity of methods based on histograms, we opt to classify pixels by means of the pixel-intensity distribution within each watershed region, which necessitates the selection of a threshold.

Let $\bar{I}(x, y) = I_{max} - I(x, y)$ be the reversed intensity at (x, y) , where I_{max} is the maximum intensity and $I(x, y)$ is the original intensity at pixel (x, y) . Assume t is a specified threshold and let C_0 and C_1 denote the set of pixels belonging to the background and to the spot, respectively. The thresholded image, discriminating the object from the background, is

$$I_t(x, y) = \begin{cases} 0 & \text{if } \bar{I}(x, y) \leq t, \text{ pixel } (x, y) \in C_0 \\ 1 & \text{if } \bar{I}(x, y) > t, \text{ pixel } (x, y) \in C_1 \end{cases} \quad (1)$$

The threshold t is found by maximizing the ratio of the between-class sum of squares SS_B to the within-class sum of squares SS_W (Otsu, 1979), which are defined as

$$SS_B = N_1(\bar{I}_{\mu C_1} - \bar{I}_{\mu})^2 + N_0(\bar{I}_{\mu C_0} - \bar{I}_{\mu})^2$$

and

$$SS_w = \sum_{(x,y) \in C_1} (\bar{I}(x,y) - \bar{I}_{\mu C_1})^2 + \sum_{(x,y) \in C_0} (\bar{I}(x,y) - \bar{I}_{\mu C_0})^2.$$

Here N_1 and N_0 are the number of pixels in each class, $\bar{I}_{\mu C_1}$ and $\bar{I}_{\mu C_0}$ are the corresponding mean intensities, and $\bar{I}_{\mu}$ is the overall mean pixel intensity in the protein region. The optimal threshold in (1) is given by

$$t^* = \arg \max_t \frac{SS_B}{SS_W}.$$

This approach is equivalent to k -mean clustering with $k = 2$. And, this criterion is actually Fisher's criterion for linear discriminant analysis. As we know, Fisher's linear discriminant is optimal when data in both classes have been generated from normal distributions with the same variance. Here, for a watershed region containing a spot, the histogram of the reversed pixel intensities is usually left skewed as there is a large number of pixels belonging to the background. This histogram is more symmetrical after log-transformation. So, we apply Otsu's method to the log intensities for classification of the pixels, rather than to the original values as is usually done.

Figure 5 shows the classification results from applying the above method to the intensity values and the log intensity values. Clearly, a more suitable threshold is found when the log intensities are used.

For local background correction, we choose, as the summary statistic, the difference on the log scale between the means of the background pixels and object pixels,

$$\log_2(\bar{I}_{\mu C_1} - \bar{I}_{\mu C_0}).$$

Now we are able to set up a statistical model for comparing the summary statistics from each region across experimental groups. Denote the log of the summary statistic for region k , image j , experimental group i by s_{ijk} , where $i = 1, \dots, n_c$, $j = 1, \dots, n$, $k = 1, \dots, K$. Assume that

$$s_{ijk} = \mu_{ik} + \varepsilon_{ijk},$$

where ε_{ijk} 's are $N(0, \sigma_k^2)$ -distributed. For the dataset described above, we have $n_c = 4$, $n = 3$, $K = 439$, and the experimental conditions ($i = 1, \dots, 4$) correspond to the factorial combinations of treatment and gender.

Before using this linear model, we need to check whether there are systematic biases caused by the experimental procedure and whether the summary statistics for the protein regions are independent. If a systematic bias exists, normalization is needed to remove it.

3. Normalization

There are many sources of variation in 2D gel experiments that affect the measured protein levels (Nishihara et al., 2002). These sources of systematic variation include: (1) sample

preparation (differences between reagent batches, variations in sample preparation protocols), (2) sample loading (pipetting errors, sample loss during gel loading), (3) staining (inconsistent staining times, differences between stain batches), (4) image acquisition (inconsistent exposure times during scanning, images captured under different settings, differences in laser power and subsequent fluorescent dye signal emission). Systematic variation can obscure changes due to treatment (drug or exposure) and thus complicate identification of true expression changes. For accurate comparison of the abundance of a protein across gels, it is essential to account for these variations.

Normalization is the process by which one removes some experimental variation. Little work has been done to study the normalization of 2D PAGE images. In some commercial software packages (such as Progenesis), normalization by the total spot volume is recommended for single-stain experiments: the volume of each spot divided by the total spot volume on the gel is the normalized value.

One-dimensional LOESS smoothing (Cleveland, 1993), also known as local regression, is one of the methods commonly used in normalization of two-channel cDNA microarrays to correct for the systematic bias due to dyes. In the two-dimensional case, given a set of (x, y, z) values, LOESS fits a smooth function of x, y to z . A window of neighboring points about each (x, y) is selected. Within the window, a robust regression technique is applied, in which observations are down weighted the further they are from (x, y) . Observations outside the window are excluded. A smoothing parameter is used to select the window, by specifying the proportion of data to be included. The wider the window, the smoother the curve.

Since proteins in a 2D gel are separated in one coordinate by their molecular mass and in the second coordinate by their isoelectric charge, there may exist some systematic spatial bias. The possibility of such systematic bias can be investigated by looking at the overall trend in the difference in the regional summary statistics along each dimension between any gel and the overall mean gel. The top two graphs in Figure 6 show the trend in $s_{42k} - \bar{s}_{..k}$ against the Y (image row) and X (image column) coordinates of the center of region k , where $(s_{42k} - \bar{s}_{..k})$ is the difference in the summary statistics of region k between gel 2 in the exposed male group and the mean gel. If no spatial bias exists in the X - or Y -direction, there should be no obvious trend in these differences. The red solid line in each is the smooth local regression curve obtained by using one-dimensional LOESS for the X - or Y -coordinate with window size of 0.95. The curvature of the red lines depicts the overall spatial trend in each direction. The top left figure clearly shows a systematic bias in the row direction (Y) of the gel. The bias in the column direction (X) is negligible.

While the one-dimensional LOESS procedure applied in each direction removes trends along the X - and Y -axes, it leaves untouched any bias arising from interactions in the two directions. We therefore propose a technique that tackles spatial bias of any type. To continue with our example, let $d_k = s_{42k} - \bar{s}_{..k}$ and assume $d_k = f(x_k, y_k) + \varepsilon_k$, where $f(x_k, y_k)$ is an unknown function and ε_k is the error term representing random residual variation and any other non-spatial effects in the observations. Correcting for the spatial bias involves estimating the function $f(x_k, y_k)$ and subtracting it from the observed difference d_k . A two-dimensional LOESS procedure is used to estimate $f(x_k, y_k)$. Suppose the coordinates of the center of region k are (X_k, Y_k) and for a neighboring region i they are (X_i, Y_i) . A local quadratic approximation to $f(X_i, Y_i)$ is

$$\widehat{f}(X_i, Y_i) = \beta_0 + \beta_1 (X_i - X_k) + \beta_2 (Y_i - Y_k) + \beta_3 (X_i - X_k)(Y_i - Y_k) + \frac{1}{2}\beta_4 (X_i - X_k)^2 + \frac{1}{2}\beta_5 (Y_i - Y_k)^2.$$

Denote by S_k the set of all pixels contained in the square centered at (X_k, Y_k) with side length equal to the integer value of $s \min(m, n)$, where $0 < s < 1$ is a specified window size ratio, and m and n are the height and width of the image. To obtain $\hat{f}(X_k, Y_k)$, we need to find $\beta = (\beta_0, \dots, \beta_5)'$ that minimizes

$$\sum_{i \in S_k} w_{ik} (d_k - \hat{f}(X_i, Y_i))^2$$

with w_{ik} a weight function related to the distance between the centers (X_i, Y_i) and (X_k, Y_k) . We define the weights as

$$w_{ik} = (1 - |u_{ik}|^3)^3,$$

where

$$u_{ik} = \frac{\sqrt{(X_i - X_k)^2 + (Y_i - Y_k)^2}}{\max_{i \in S_k} \sqrt{(X_i - X_k)^2 + (Y_i - Y_k)^2}}.$$

Using the LOESS estimate $\hat{f}(X_k, Y_k)$, the spatial trend in $(s_{42k} - \bar{s}_{..k})$ is removed by subtracting $\hat{f}(X_k, Y_k)$ from s_{42k} , thus bringing s_{42k} closer to $\bar{s}_{..k}$.

The lower panel of Figure 6 shows the spatial trend in $(s_{42k} - \bar{s}_{..k})$ along each direction after the above two-dimensional LOESS normalization. Again, the red solid line is the one-dimensional LOESS estimate along each direction, which depicts the spatial trend along the corresponding direction after the two-dimensional LOESS normalization. Clearly, there is no obvious spatial trend after two-dimensional LOESS correction.

After normalization, provided the experimental design does not introduce confounding, we assume that differences in protein expression levels across gels are due to factors of interest, here, treatment, gender, and subject effects.

4. Differential Analysis of Protein Regions

According to the factorial experimental design for the data under consideration, we write the following linear model:

$$\begin{aligned} s_{1jk} &= \mu_k - \tau_k - \gamma_k - (\tau\gamma)_k + \varepsilon_{1jk} \\ s_{2jk} &= \mu_k - \tau_k + \gamma_k + (\tau\gamma)_k + \varepsilon_{2jk} \\ s_{3jk} &= \mu_k + \tau_k - \gamma_k + (\tau\gamma)_k + \varepsilon_{3jk} \\ s_{4jk} &= \mu_k + \tau_k + \gamma_k - (\tau\gamma)_k + \varepsilon_{4jk} \end{aligned} \quad (2)$$

where τ_k , γ_k , $(\tau\gamma)_k$ are, respectively, the treatment, gender, and interaction effect for region k ($j = 1, \dots, n$, $k = 1, \dots, K$). With the assumption that $\varepsilon_{ijk} \sim N(0, \sigma_k^2)$, the test statistic for the treatment effect on region k is

$$t_k = \frac{\bar{s}_{3,k} + \bar{s}_{4,k} - \bar{s}_{1,k} - \bar{s}_{2,k}}{2\sqrt{S_k/n}},$$

where S_k is the pooled sample variance. t_k follows the t -distribution with $4(n-1)$ degrees of freedom under the null hypothesis that $\tau_k = 0$. To validate the assumption of normally distributed error, the Kolmogorov-Smirnov test was performed on the standardized

observations $\frac{s_{ijk} - \bar{s}_{i,k}}{\sqrt{(1 - \frac{1}{n})S_k}}$, with the null hypothesis that these standardized observations follow the t -distribution with 8 degrees of freedom. The Kolmogorov-Smirnov test suggested that the model assumption is acceptable.

Since proteins may be functionally correlated, some of the s_{ijk} 's may not be independent. Next, we need to implement a method to identify the highly correlated s_{ijk} 's.

4.1. Testing for Correlations

Let $S_{ij} = (s_{ij1}, s_{ij2}, \dots, s_{ijK})'$ and $S_{ij} \sim N(\mu_i, \Sigma)$ ($i = 1, \dots, n_c; j = 1, \dots, n$). If the s_{ijk} 's are mutually independent, the covariance matrix Σ is diagonal. We want to test the hypothesis

$$H_0: \rho_{kl} = 0 \text{ for } k \neq l, \quad k, l = 1, \dots, K$$

against the alternative hypothesis

$$H_1: \rho_{kl} \neq 0 \text{ for at least one pair } (k, l) \quad k \neq l.$$

The pooled sample covariance estimator for Σ is

$$V = \frac{1}{n_c(n-1)} \sum_{i=1}^{n_c} \sum_{j=1}^n (S_{ij} - \bar{S}_i)(S_{ij} - \bar{S}_i)',$$

which is a singular matrix since the number of degrees of freedom $df = n_c(n-1)$ is much smaller than the dimension K . Let V_{ij} denote the entries of V . Then the sample correlation coefficient is

$$r_{kl} = \frac{V_{kl}}{\sqrt{V_{kk} V_{ll}}}.$$

To identify those proteins that are significantly correlated, we need to perform multiple tests controlling the proportion of false-positive inferences. The false discovery rate at a chosen threshold r_t is estimated as

$$\widehat{FDR}(r_t) = \frac{\widehat{\pi}_0 N_0(r_t)}{N_1(r_t)}$$

where $N_0(r_t)$ is the expected number of sample correlation coefficients more extreme than r_t under the null hypothesis of no correlation, and $N_1(r_t)$ is the observed number of sample correlation coefficients more extreme than r_t (Efron, 2003; Storey and Tibshirani, 2003). $\hat{\pi}_0$ is a smoothed estimator of the proportion of null hypotheses and is obtained by smoothing

the raw estimators $\hat{\pi}_0(\lambda) = \frac{n_1(\lambda)}{n_0(\lambda)}$ for a selected set of $\lambda \in (0, 1)$, where $n_1(\lambda)$ is the observed number of sample correlation coefficients that satisfy $|r_{kl}| < \lambda$ and $n_0(\lambda)$ is the expected number when there is no correlation. The final smoothed estimator for π_0 is given by

$$\hat{\pi}_0 = \lim_{\lambda \rightarrow 0} \hat{\pi}_0(\lambda).$$

To estimate $N_0(r_t)$, we use a permutation procedure. Randomly permute the observed s_{ijk} 's within each condition i for each region k many times. The average number of observed sample correlation coefficients more extreme than r_t is used as the estimate of $N_0(r_t)$. We note that, with the intention of obtaining a rough estimate, the correlations among the sample correlation coefficients are not taken into account. For the dataset under study, if we want to control the FDR so that it is smaller than 0.25 (0.35), we need to choose r_t greater than 0.96 (0.90).

4.2. Statistical Differential Analysis of Protein Regions

Based on the chosen threshold for the sample correlation coefficients, we separate protein regions into independent sets of correlated proteins. Regions $\{k_1, \dots, k_i, \dots, k_n\}$ will be called a *correlation set* of size n if each member k_i is correlated with one or more regions in the same set and not correlated with any other region not in that set. So, a protein region which is found not to be correlated with any other protein region is a set of size 1. Within each correlation set, since the protein regions are highly correlated with one another, we use a multivariate approach to analyze the overall effect of the factor of interest (e.g., treatment, exposure). For example, if regions k_1 and k_2 are found to be correlated, but they are not correlated with any other region, regions k_1 and k_2 constitute a correlation set and will be analyzed using a bivariate model. If region k_1 or k_2 is correlated with another region k_3 , and k_3 is not correlated with any other region, then regions k_1, k_2, k_3 will be in the same set.

We apply MANOVA to those sets of size greater than 1 and ANOVA to those sets with only one member. Since the correlation sets are assumed independent or weakly dependent, the resulting p-values are considered independent. MANOVA will not be feasible when the size of a set is larger than the number of degrees of freedom of the sample covariance matrix. Correlations among proteins are expected to be sparse, which means that most of the independent sets are of small size. For example, for the dataset under consideration, if we select the threshold for r_t at 0.96, we obtain 10 sets of size 2 and all others are of size 1. Simulations (see Section 5) show that using a multivariate analysis for correlated proteins improves power and has a false-discovery rate similar to the univariate analysis.

Assume the p-values obtained from the multivariate analysis are $p_1, \dots, p_c$. We apply the BH-FDR procedure to find the threshold that controls the false-discovery rate at the desired level. With the false discovery threshold chosen as 0.05, there are 13 proteins detected as differentially expressed including 5 pairs of correlated proteins. Figure 7 shows the 3D shape of a protein region with a significant treatment/exposure effect: the volume of the protein has decreased under nicotine exposure.

5. Simulations

In this section, we describe the method we used to generate simulated gel images for investigating the performance of our proposed methodology in comparison to the commercial software packages PDQuest and Dymension. We tackle the difficult task of creating realistic simulated gels by randomly perturbing an actual gel image as described below.

We chose, as templates, two gel images, G_1 and G_2 , from the same treatment/exposure group and define $D = G_1 - G_2$ as the difference image. Images D , G_1 and G_2 are matrices of pixel intensities of the same size. We used the matrices D and G_1 to generate a random gel G by letting

$$G(i, j) = G_1(i, j) - \text{sign}(D(i, j)) |R_D(i, j)|$$

where $G(i, j)$ is the pixel intensity at row i , and column j of image G , and $R_D(i, j)$ follows a normal distribution with mean 0 and standard deviation $|D(i, j)|$. We generated $2n$ such images as replicates for 2 treatment conditions with $n = 10$ for each condition. These 20 replicates were then smoothed by averaging the pixel intensities within a 5×5 window, followed by a cubic spline smoothing interpolation. We applied the watershed algorithm to the sum of the two template images, $G_1 + G_2$, and used the resulting watershed region for random perturbation.

Next, we performed perturbations by transforming pixel intensities within each watershed region. Suppose the maximum, minimum and range of the intensity I_{ij} of a protein spot in a watershed region are M , m and R_0 respectively. To create a random change, let the new range be $R_1 \geq 0$ and the new intensity at pixel (i, j)

$$I'_{ij} = \left(\frac{I_{ij} - m}{R_0} \right)^\beta R_1 + M - R_1$$

where $\beta > 0$ is randomly chosen. Specifically, we let $R_1 = \max(0, R)$ with R normally distributed with mean R_0 and standard deviation $0.1R_0$, and choose β uniformly in the range (0.9, 1.1). Figure 8 shows a selected protein spot and a perturbation using $R_1 = 0.5R_0$ and $\beta = 2$. We performed such random perturbation on each watershed region to get $2n$ simulated gels.

To simulate treatment effects, we further perturbed 11 spots at corresponding locations on each gel in the treatment group. Among the 11 spots, in comparison to the control group, 4 spots became fainter, 4 darker and 3 were removed. We set $R_1 = 0.5R_0$ and $\beta = 0.9$ to create the 4 fainter spots, and $R_1 = 1.5R_0$ and $\beta = 1.1$ to generate the 4 darker spots. We fitted a smoothing cubic spline surface to the watershed region using a 10-pixel window to erase 3 spots. We chose the above numbers so that the changes are significant in comparison with the simulated random variation.

Figures 9(a) and (b) show examples of simulated gel images for the two groups, where the altered spots are highlighted. Figure 10 displays regions where spots become darker, fainter or are missing in the treatment group.

We applied the methodology described in the previous sections to the simulated gel images, controlling the false discovery rate at 0.05: all and only the 11 significantly changed spots were detected (p-values < 0.0001).

To validate our methodology, we analyzed the same set of simulated gels with the well-known commercial software packages PDQuest and Dymension. For this paper, PDQuest Advanced version 8.0.1 Build 055 and Dymension version 2.0.3.0 were used. For both software packages, hypothesis testing was performed via a t-test with a pre-specified p-value threshold. The parameters and options in both software packages were selected to obtain the best possible results. The results are summarized in Table 1 and are depicted in Figures 11 and 12.

These figures show the detection results using the usual p-value thresholds, 0.05 and 0.01. The red circles or points identify the protein spots detected as statistically different. False-positive results are enclosed in black rectangles and false-negative results in blue rectangles.

When compared with the 11 proteins spots that truly changed in Figure 9, among the 21 detections by PDQuest (including the top-left corner) there are 10 false positives using 0.05 as the p-value threshold. There is no differentially expressed protein that was not detected. With a p-value threshold of 0.01, the number of false positives reduces to 3 (including the top-left corner), but there is now one differentially expressed protein that was not detected. Dymension with the p-value level of 0.05 gives 10 false positives among the 16 detections, and misses 5 spots. With the threshold of 0.01, there are 6 false positives among the 9 detections and 8 spots were missed. Overall, PDQuest performed better than Dymension on the set of simulated gels but was sensitive to background noise.

Next, we want to investigate the extent to which a multivariate analysis, compared to a univariate analysis, improves power and affects the resulting FDR. Without loss of generality and for simplicity, we consider the following set-up. Assume we observe y_{ij} 's ($i = 1, \dots, N$ and $j = 1, \dots, n$) in N groups of n observations, where $y_{ij} \sim N(\mu_i, \sigma_i^2)$ independently for each i . We want to test the hypothesis that $H_i : \mu_i = 0$ for each i when strong pairwise correlations exist. For convenience, the first N_1 pairs follow a bivariate normal distribution, which means that $(y_{k1j}, y_{k2j})' \sim N((\mu_{k1}, \mu_{k2})', \Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma_{k1}^2 & \rho\sigma_{k1}\sigma_{k2} \\ \rho\sigma_{k1}\sigma_{k2} & \sigma_{k2}^2 \end{pmatrix},$$

$k_2 = k_1 + 1$ and $k_1 = 1, 3, 5, \dots, 2N_1 - 1$. All other y_{ij} 's are assumed independent; there are N_2 groups for which the null hypothesis is true and N_3 groups for which the null hypothesis is false ($N = 2N_1 + N_2 + N_3$).

We consider both scenarios of known and unknown correlation structure. Suppose, first, that the correlation structure is known (Table 2). For simplicity, we simulate the data for half of the N_3 independent groups, for which the alternative hypothesis is true, from $N(-1, 1)$ and the other half from $N(2, 1)$. The data for the N_2 independent groups for which the null hypothesis holds are generated from $N(0, 1)$. For the N_1 pairs, we consider either all alternative hypotheses as true or all null hypotheses as true for 3 values of ρ , i.e. -0.9, 0, 0.9. For each setting, the simulation is repeated 500 times. We set $N_1 = 20$, $N_2 = 150$, $N_3 = 10$ and $n = 12$ in all simulations. We perform a bivariate normal analysis (MANOVA) for the N_1 pairs (i.e., Hotelling's test), and, a univariate analysis (i.e., a t-test or F-test) for all the other groups. The BH-FDR procedure is applied to the resulting p-values with the desired

FDR set at $\alpha = 0.05$. For comparison, the BH-FDR procedure is also applied directly to the individual p-values obtained from the univariate analyses (ANOVA) on all N groups (Table 2). The criteria for comparison are the average false-discovery proportion and the power.

In practice, the dependence structure is usually unknown (Table 3). We then compute the sample correlation coefficients to estimate population correlations. Based on the observed correlations, we separate the data into independent sets of correlated observations. For simplicity, we use $r_t = 0.9$ as a single threshold. We repeat the simulations, now assuming that the correlation structure is unknown and estimated from the observed sample. From Table 2, when the correlated data follow the distribution corresponding to the alternative hypothesis, using the BH-FDR procedure in conjunction with a multivariate analysis greatly increases power while keeping the false-discovery rate below the control level. Even when $\rho = 0$, using a multivariate analysis still improves power. This results from the fact that there are fewer p-values being computed for the multivariate analysis than for the univariate analyses. When all the correlated data are generated from the null distribution, the power and the false-discovery rate for both approaches are very similar, whether there is strong correlation or not. From Table 3, we see that when the observations are highly correlated and generated from the alternative hypotheses, the multivariate analyses for the constructed correlation sets still yield better power than the univariate analyses, although it is slightly decreased compared to the case where the correlation structure is known. Note that when highly correlated data are generated from the null hypotheses, the false-discovery rate for the multivariate approach is higher than the control level. This may be due to the test of correlation with the fixed threshold $r_t = 0.9$.

6. Discussion

Although there are several emerging technologies for protein quantification and separation, 2D PAGE remains an important and popular assay in proteomics. It is often used by researchers as a screening tool to locate interesting proteins and combined with mass spectra for sequence identification. Unlike for the analysis of microarrays, there is quite limited published research on statistical differential analysis of 2D PAGE images. The main difficulties are the discrimination between actual protein spots and noise, the quantification of protein expression levels, and spot matching for individual comparison. To date, some laboratory researchers are still manually selecting protein spots by visual comparison. Though commercial software packages are available, they are expensive, interaction-intensive, non-open source and include very basic statistical tools. Therefore, there has been a longstanding need to develop a procedure, along with software, that is simple, robust and easy to utilize.

The procedure introduced in this paper focuses on detecting differentially expressed proteins via numerical summaries of image regions separated by watershed boundaries rather than via accurate quantification of each protein spot, acknowledging the fact that 2D PAGE is mainly used as a screening tool. Our methodology aims at improving analytical efficiency by minimizing user input required to locate interesting proteins: current software necessitates many rounds of trials and errors and yields numerous instances of missing information though the spots are clearly visible on the images. This is due to ad-hoc spot-matching procedures. We avoid this bottleneck by comparing corresponding proteins regions defined by a master watershed map, which makes our approach fully automated. The master watershed map consists of the watershed boundaries obtained by applying the watershed algorithm to the mean image of all gels. Since a spot on the mean image is either of similar shape and size as the corresponding spot on each gel, the master watershed regions separate the proteins well on each image.

As gel images often contain faint spots, it may be necessary to further enhance the mean image to improve the contrast. For various image enhancing techniques, see Gonzalez and Woods (2002). Image processing techniques may change the shape and intensity of spots to varying extent. In our procedure, this kind of image transformation is not a concern since we only use the enhanced mean image to obtain the master watershed boundaries. Gel images usually contain spikes with small footprint (due to dust particles on the scanner, for instance). Thus, a watershed region may just contain a spike. Since it can be difficult to assess whether a faint spot is a spike or protein spot, we don't recommend spending much effort on removing those spikes. Indeed, we don't expect such regions to show statistically significant changes: spikes are artifacts that are unlikely to be confounded with a grouping of interest (e.g. treatment vs. control) as they will appear either consistently on all images or at random on some and not on others.

The final results will, of course, depend on the type of summary statistics used in the analysis. We selected the mean intensity of the object portion in each region. But, it may not capture either all the features of a spot nor the real differences between spots. Other statistics such as the intensity range and volume may be used, too. Another way to utilize these features is to perform an analysis for each feature separately and keep all interesting findings for final investigation.

The two-dimensional LOESS method to remove spatial bias should be used with caution as any normalization will modify the values for analysis and true effects may be removed. We do not have a rigorous method to determine whether spatial bias exists other than visual inspection and estimating the overall trend using LOESS methods.

Our final hypothesis testing accounts for the correlation among proteins to improve power by using a multivariate analysis. In practice, many studies involve few samples. Thus, resulting parameter estimates may not be reliable. Empirical Bayes methods, where information is combined across genes, have been shown to produce reliable results in microarray analysis of small experiments. Here, we would combine information across regions.

We validated our procedure by comparing it with two commercial software packages on a set of simulated gel images. We agree that our simulation depicts a limited scenario, though it clearly demonstrates the effectiveness of the method. Creating simulated gels images that reflect real gels is an arduous task. We tackled it by randomly perturbing real gel images. This introduces a new method for generating simulated gels that can be utilized in future research.

While some aspects of the procedure we proposed are not new in terms of statistical methodology, it is a novel combination of methods from different fields with a creative and effective implementation in the analysis of 2D PAGE images.

We developed a custom Matlab-based software package with graphical user interface (GUI). It provides both a fully automated and an interactive procedure. The user can also select a subset of protein regions to investigate in various ways. The software is introduced in detail in Li and Seillier-Moisewitsch (2011). It is available together with a user manual upon request from the corresponding author.

Acknowledgments

This research was supported in part by NIH grant 5R01GM075298. The authors thank Carol Whisnant for the use of her data.

References

- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*. 1995; 57:289–300.
- Cleveland W. A model for studying display methods of statistical graphics. *Journal of Computational and Graphical Statistics*. 1993; 2:323–364.
- Conradsen K, Pedersen J. Analysis of two-dimensional electrophoretic gels. *Biometrics*. 1992; 48:1273–1287.
- Efron B. Robbins, empirical Bayes, and microarrays. *The Annals of Statistics*. 2003; 24:366–378.
- Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*. 2004; 99:96–104.
- Efron B. Correlation and large-scale simultaneous significance testing. *Journal of American Statistical Association*. 2007a; 102:93–103.
- Efron B. Size, power, and false discovery rates. *Annal of Statistics*. 2007b; 35:1351–1377.
- Gonzalez, R.; Woods, R. *Digital Image Processing*. Prentice Hall; 2002.
- Li, F. *Empirical Bayes Methods for Proteomics* Ph D Dissertation. University of Maryland; Baltimore County: 2007.
- Li, F.; Seillier-Moiseiwitsch, F. Differential Analysis of 2D Gel Images. In: Abelson, John N.; Simon, Melvin I., editors. *Methods in Enzymology, Vol 487, Computer Methods, Part C*, Michael L Johnson, Ludwig Brand. San Diego: 2011. p. 596-609.
- Morris JS, Clark BN, Gutstein HB. Pinnacle: A fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data. *Bioinformatics*. 2008; 24:529–536. [PubMed: 18194961]
- Nishihara J, Champion K. Quantitative evaluation of proteins in one and two-dimensional polyacrylamide gels using a fluorescent stain. *Electrophoresis*. 2002; 23:2203–2215. [PubMed: 12210224]
- Otsu N. A threshold selection method from gray level histograms. *IEEE Transactions on Systems, Man and Cybernetics*. 1979; 9:62–66.
- O'Farrell, Patrick. High resolution two-dimensional electrophoresis of proteins. *Journal of Biol Chem*. 1975; 250:4007–4021. [PubMed: 236308]
- Potra F, Liu X, Seillier-Moiseiwitsch F, Roy A, Hang Y, Marten M, Raman B. Protein image alignment via piecewise affine transformations. *Journal of Computational Biology*. 2006; 13:614–630. [PubMed: 16706715]
- Qiu X, Klebanov L, Yakovlev A. Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*. 2005; 4:1–13.
- Roy A, Seillier-Moiseiwitsch F, Lee K, Hang Y, Marten M, Raman B. Analyzing two-dimensional gel images. *Chance*. 2003; 16:13–18.
- Seillier-Moiseiwitsch F, Roy A, Zhang Y, Hang Y, Li F, Korostyshevskiy V, Potra F, Liu X. Statistical analysis of differential expression of proteins on two-dimensional gel electrophoresis images. *Journal of the Royal Statistical Society*. 2011 submitted to.
- Storey J, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of National Academy of Sciences*. 2003; 100:9440–9445.
- Tibshirani, R.; Wasserman, L. Correlation-sharing for detection of differential gene expression, technical Report, Stanford University. 2006. URL <http://www-stat.stanford.edu/tibs/research.html>
- Vincent L, Soille P. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1991; 13:583–598.

Figure 1.
Example gel images for the study of exposure of spleen tissue to nicotine in rats.

Figure 2.
Same protein on different gels.

Figure 3.
Separation into protein regions using watershed boundaries.

Figure 4.
Sections of images with proteins separated into regions using the master watershed map.

(a) 2D contour of a selected spot.

(b) 3D shape of the 'reversed' spot.

(c) Threshold based on intensity.

(d) Threshold based on log intensity.

Figure 5.

Classification of pixels into object and background (the original pixel intensity range was [0, 4095]).

Figure 6.
Normalization of spatial trends.

Figure 7.
A region with a significant exposure effect (pixel intensities are reversely rescaled to [0,1] in this figure).

Figure 8.
A selected spot region and its perturbation (pixel intensities are reversed).

Figure 9.
Two simulated gel images.

(a) Control group.

(b) Darker spot in treatment group.

(c) Control group.

(d) Fainter spot in treatment group.

(e) Control group.

(f) Missing spot in treatment group.

Figure 10.
Three changed spots.

Figure 11.

PDQuest results using t-tests. False positives are enclosed in black rectangles and missed spots in blue (the top-left corner is not depicted though it was identified as a significant spot).

Figure 12.

Dymension results using t-tests. False positives are enclosed in black rectangles and missed spots in blue (the top-left corner is not depicted though it was identified as a significant spot).

Table 1

Summary of significantly changed spots identified by PDquest and Dymension.

Software	Level	Total	correct detection	False positive	False negative
PDQuest	0.05	21	11	10	0
	0.01	13	10	3	1
Dymension	0.05	16	6	10	5
	0.01	9	3	6	8

Table 2

Comparison of multivariate and univariate analyses (known correlation structure).

		$\rho = -0.9$		$\rho = 0$		$\rho = 0.9$	
(μ_1, μ_2)	Methods	FDR	Power	FDR	Power	FDR	Power
(-1, 2)	MANOVA	0.0249	0.9605	0.0251	0.9604	0.0243	0.9613
	ANOVA	0.0364	0.8243	0.0376	0.8300	0.0364	0.8298
(0, 0)	MANOVA	0.0514	0.6866	0.0504	0.6798	0.0464	0.6860
	ANOVA	0.0473	0.6748	0.0496	0.6690	0.0459	0.6744
(1, 2)	MANOVA	0.0249	0.9598	0.0274	0.9600	0.0261	0.9606
	ANOVA	0.0375	0.8291	0.0404	0.8260	0.0384	0.8251

Table 3
Comparison of multivariate and univariate analyses (unknown correlation structure).

		$\rho = -0.9$		$\rho = 0$		$\rho = 0.9$	
(μ_1, μ_2)	Methods	FDR	Power	FDR	Power	FDR	Power
(-1, 2)	MANOVA	0.0484	0.9170	0.0572	0.8342	0.0490	0.9228
	ANOVA	0.0377	0.8260	0.0384	0.8315	0.0401	0.8235
(0, 0)	MANOVA	0.1058	0.6814	0.1015	0.6632	0.1062	0.6788
	ANOVA	0.0472	0.6690	0.0493	0.6556	0.0452	0.6690
(1, 2)	MANOVA	0.0504	0.9248	0.0548	0.8303	0.0477	0.9202
	ANOVA	0.0383	0.8302	0.0353	0.8276	0.0368	0.8310