12

Statistical Analysis of Gel Electrophoresis Data

Kimberly F. Sellers[1] and Jeffrey C. Miecznikowski[2]
[1]*Georgetown University*
[2]*SUNY University at Buffalo*
USA

1. Introduction

Two-dimensional gel electrophoresis (2-DE) methods such as two-dimensional polyacrylamide gel electrophoresis (2D-PAGE; O'Farrell (1975)) and two-dimensional difference gel electrophoresis (2D-DIGE; Ünlü et al. (1997)) are popular techniques for protein separation because they allow researchers to characterize quantitative protein changes on a large scale. Thus, 2-DE is frequently used as an initial screening procedure whereby results obtained generate scientific ideas for study. These technologies revolutionized the field of proteomics and biomarker discovery in their ability to detect protein changes either in differential expression or modification (Huang et al., 2006; Rai & Chan, 2004; Wulfkuhle et al., 2003; Zhou et al., 2002). Further, they are attractive because of their resolving power and sensitivity. 2-DE analyses, however, require personnel with significant wet laboratory expertise and can be time-consuming, thus potentially limiting the sample size for gels.

This chapter describes the statistical implications associated with gel electrophoresis data, and statistical methods used for analysis. Section 2 describes various low-level analysis techniques used to preprocess and summarize the electrophoresis image data. Section 3 uses the preprocessed data to address the biological question(s) of interest. In this section, two statistical issues are addressed: the choice of an appropriate statistical test, and the matter of multiple testing. Section 4 discusses data missingness, and describes proposed methods for imputation. Finally, Section 5 illustrates the above analyses via a case study example, and Section 6 concludes the chapter with discussion.

2. Preprocessing

Similar to the methods used to analyze gene expression microarrays, the general steps in preprocessing 2-DE data include outlier detection, baseline or background subtraction, signal distribution normalization, protein (or peptide) alignment, feature (i.e. spot) detection and quantification, and biomarker evaluation (Sellers & Miecznikowski, 2010). Concerns regarding these procedures are significant because all subsequent analyses relating to the proteomics data are contingent on these first steps being performed appropriately and optimally. Thus, the goal in preprocessing 2-DE data is to create an unbiased, reproducible, and automated approach toward identifying differentially expressed and modified proteins via spot differences.

Although the statistical work here is analogous to that for microarray data analysis, applying these methods to 2-DE data is more complex due to the added randomness that exists in the image spots due to the process by which the data is created. Spots in microarray images are systematically placed in identical locations for simple comparisons. Meanwhile, the isoelectric focusing and sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) procedures rely on proteins being stabilized at their isoelectric point (where their net charge is zero). As a result, the protein location includes inherent variation, thus making spot comparisons a more difficult task.

This section describes the various techniques that are used to best quantify and summarize the image data, including outlier detection (Section 2.1), background correction (Section 2.2), data denoising and normalization (Section 2.3), image alignment or spot matching (Section 2.4), and spot detection (Section 2.5). Finally, Section 2.6 elaborates on issues that one must consider when performing such analyses. Many of the aforementioned methods are easily enacted using the statistical programming language, *R*, and the Bioconductor suite of software packages (Gentleman et al., 2004; R Development Core Team, 2008).

2.1 Outlier detection

Before analyzing each protein and sample in the study, it is worthwhile to detect (sample or protein) outliers that may be candidates for removal from the study. If the presence of outliers is determined, the analyst has an important decision regarding whether or not to retain the relevant observation(s). Such a decision hinges on the cause of the outlier. Outliers that occur due to measurement or other technical errors can be justifiably removed in order to improve inference and decision making. Observations that are true measurements of biological signal, however, should remain in the data; their removal will bias the downstream results.

There are numerous methods for outlier detection (e.g., Jackson & Chen (2004); Schulz-Trieglaff et al. (2009)). One of the most common methods used to detect outlying protein spots or samples (gels) is principal component analysis (PCA; Jolliffe (2002)). PCA reduces data dimensionality by performing a covariance analysis between the gels/proteins. As such, if a gel/protein is determined to have little covariance with its counterparts, then it is "flagged" as an outlier. Quickly implemented, PCA analyses are easily performed and the results can be visualized with standard statistical softwares, e.g. *R* (R Development Core Team, 2008). While there are "rules of thumb" for determining outliers using PCA, it is usually satisfactory to determine outliers from figures that display the first three principal components. See Rao (1964) and Cooley & Lohnes (1971) for further details regarding outlier detection via PCA.

2.2 Background correction

A general problem that occurs in image data is the matter of background correction. Image data are inherently inflated to avoid the potential for nonsensical, negative image values. Depending on the application, backgrounds can be inflated at a general as well as localized basis, thus making the matter of significant interest. One must remove the background noise in a proper way so that the "true data" or biological signal is summarized properly.

Background adjustments are required to correct for measured intensities resulting from non specific effects in the electrophoresis process and noise in the optical detection system (e.g.

camera or laser scanner). Other sources of noise may include unintended differences in culture growth conditions, sample preparation, and reagent quality (Keeping & Collins, 2011). Irrespective of the cause, the background should be adjusted or removed in order to more accurately quantify the proteins present. Suggested approaches for background removal include subtracting the global minimum, or performing some low percentile Winsorization; local removal based on spot boundaries and outer region(s); filtering in the wavelet domain; and asymmetric least squares splines regression. See Kaczmarek et al. (2004) for further discussion of these and other linear and non-linear techniques for background removal.

2.3 Data denoising and normalization

Gel experiments contain various sources of systematic variation, masking the true data. Data denoising and normalization serve to remove these effects in order to give accurate signal measurements from the 2-DE data.

Sellers et al. (2007) addressed this issue, focusing on factors caused by the apparatus used to image 2D-DIGE data. Through a series of experiments, they estimate the associated factors and establish a model to remove these estimates from the raw gel images to obtain respective images for the true relative protein intensities.

Relative scaling is another common practice used to normalize 2-DE data, although the approach may vary. Images may be rescaled by dividing the image by the maximum value, or total spot volume. While approaches may differ, normalization allows for better comparisons since the resulting data are on the same scale.

2.4 Alignment of group replicate profiles

2DE gel images appear distorted for any number of procedural causes (e.g. casting, polymerizing, and running the gels). Accordingly, unmatched spot pairs appear as either a result of misalignment or protein phosphorylation. Profile alignment can take place either by matching spots between the two images (preferred by some for its reduced complexity), or aligning the full images.

Various approaches have been proposed to address gel alignment. Lemkin (1997) provides two image warping options, affine and polynomial transformations, for gel alignment in his computational gel comparison tool, Flicker (available at http://www-lmmb.ncifcrf.gov/flicker). The transformations require three and six landmarks, respectively, for operation; see details in Lemkin (1997). Appel et al. (1997) also use polynomial warping to perform image alignment in their gel analysis tool, Melanie II. Here, the two polynomial functions for the respective axes are determined via least squares to minimize the error distance between respective landmarks in the two images. Note that the Melanie II system warps independently in the horizontal and vertical directions, respectively, based on their associated landmarks. Under this setup, a polynomial of degree n requires $n + 1$ landmarks, where Melanie II performs first-order, second-order, third-order, and inverse third-order polynomial warping. Meanwhile, Appel et al. (1997) have a six-step algorithm that locates neighboring spots, thus identifying clusters around the spot; matches and compares spot clusters across gels; compares secondary clusters; performs a consistency check to detect possible spot mismatching; and finally performs a transformation to match remaining spots.

Conradsen & Pedersen (1992) introduce a warping procedure where they apply a linear transformation at progressively finer scales to minimize the sum of squared differences between pixels. Smilansky (2001) instead considers a pixel-based approach to align pixels within sub-rectangles that cover the images such that the resulting shift vectors are combined to comprise the overall warping scheme using Delauney triangulation. Gustafsson et al. (2002) developed a two-step algorithm to align 2DE images. In the first step, they apply a warping model to correct for current leakage along the sides of the gel. These images are individually warped without consideration for spot matching. Secondly, the images are aligned by maximizing a penalized likelihood. There are various similarities and differences among these alignment procedures. Smilansky (2001) and Gustafsson et al. (2002) agree in using a pixel-based approach to align gel images. Meanwhile, the matching criteria of Gustafsson et al. (2002) and Conradsen & Pedersen (1992) are similar, while that of Smilansky (2001) is more involved. Finally, the Smilansky (2001) approach is performed much faster (seconds or minutes per gel, according to Gustafsson et al. (2002)) than those of Conradsen & Pedersen (1992) and Gustafsson et al. (2002) (approximately one hour per gel) because the Smilansky (2001) approach works in a parallel fashion aligning several subimages, while the methods of Conradsen & Pedersen (1992) and Gustafsson et al. (2002) operate on the full image, resulting in a longer computation period.

2.5 Feature detection and quantification

Feature detection and quantification are important issues because they reduce data dimensionality and complexity to allow for more feasible statistical analyses. This process is significant since analysts want to optimally capture protein information for subsequent study. "The difficulties and consequences involving the image processing of 2DE gels have been reported by several authors, but no general or optimal procedure for quantification of protein spots on 2DE gels is in use." (Jensen et al., 2008). Spot detection and quantification remains an open problem because proposed methods contain various tradeoffs and the scientific application of protein analysis does not allow a means for checking accuracy.

There are several classes of algorithms that can be applied for spot detection and quantification of 2-DE data. Such methods are categorized under the theme of "image segmentation". Within image segmentation, there are four main approaches: threshold techniques, boundary-based techniques, region-based methods, and hybrid techniques that combine boundary and region criteria (Adams & Bischof, 1994). Threshold techniques are based on the theory that all pixels whose values lie within a certain range belong to one class. This method neglects spatial information within the image and, in general, does not work well with noisy or blurred images. Boundary-based methods, on the other hand, are derived from the idea that pixel values change rapidly at the boundary between two regions. Such methods apply a gradient operator in order to determine rapid changes in intensity values. High values in a gradient image provide candidates for region boundaries which must then be modified to produce closed curves that delineate the spot boundaries. The conversion of edge pixel candidates to boundaries of the regions of interest is often a difficult task. The complement of the boundary-based approach is to work within the region of interest.

Region-based methods work under the theory that neighboring pixels within the region have similar values. This leads to the class of algorithms known as "region growing", of which the "split and merge" techniques are popular. In this technique, the general procedure is to

compare one pixel to its neighbor. If some criterion of homogeneity is satisfied, then that pixel is said to belong to the same class as one or more of its neighbors. As expected, the choice of the homogeneity criterion is critical for even moderate success and can be highly deceiving in the presence of noise.

The class of hybrid techniques that combine boundary and region criteria includes morphological watershed segmentation and variable-order surface fitting. The watershed method is generally applied to the gradient of the image. In this case, the gradient image can be viewed as a topography map with boundaries between the regions represented as "ridges". Segmentation is then equivalent to "flooding" the topography from local minima with region boundaries erected to keep water from different minima exclusive. Unlike the boundary-based methods above, the watershed is guaranteed to produce closed boundaries even if the transitions between regions are of variable strength or sharpness. Such hybrid techniques, like the watershed method, encounter difficulties with 2-DE images in which regions are both noisy or have blurred or indistinct boundaries. A popular alternative is seeded region growing (SRG). This method is based on the similarity of pixels within regions but has an algorithm similar to the watershed method. SRG is controlled by choosing a small number of pixels or regions called "seeds". These seeds will control the location and formation of the regions in which the image will be segmented. The number of seeds determines what is a feature and what is irrelevant or noise-embedded. Once given the seeds, SRG divides the image into regions such that each connected region component intersects with exactly one of the seeds. The choice of the number of seeds is crucial to this algorithm's success.

Early attempts for spot detection assumed that the spots were sufficiently modelled via two-dimensional Gaussian distributions; this, however, is no longer believed to be true. Thus, various algorithms instead integrate the above method(s) for spot detection and quantification in 2-DE data without the two-dimensional Gaussian assumption (Sellers & Miecznikowski, 2010). Alternative spot detection techniques include the use of diffusion equations, linear programming, and wavelet modeling. Srinark & Kambhamettu (2008) employ the watershed method, along with region filtering (via *k*-means clustering), spot extraction, and centroid estimation to locate protein spot centers and quantify spot areas. Their elaborate, seven-step algorithm works to eliminate potential oversegmentation and remove noise and extraneous features (e.g. dust) in order to re-estimate the protein spot center via the two-dimensional Gaussian distribution function. While the algorithm is robust for geometrically distorted simulated images, it has difficulties in practice with real gels.

Langella & Zivy (2008) instead use image topography via a "surface criterion," where pixels travel in the direction of maximal uphill slope. The algorithm supplies the final image illustrating spot boundaries, along with other grayscale images of potential interest. The algorithm performs well in simulated gels, but faces difficulties with regard to diffuse or saturated spots. Further, this algorithm does not account for spot matching and, thus, cannot be used for comparative analysis across gels. See Langella & Zivy (2008) for details; associated computer code is available at http://moulon.inra.fr/beads/beads.html.

Miecznikowski et al. (2009) have a feature detection procedure for finding bumps by applying a cross-shaped median smoother (of some defined size) across the gel image, and analyzing the resulting residual image that is the difference between the original and smoothed images. The residual image visually displays crosses and feature outlines that easily identify

respective spot centers and sizes. The novelty of this approach is its applicability to multidimensional datasets, i.e. it serves as a spot detection algorithm for two-dimensional datasets, and a peak detection algorithm for one-dimensional data. Meanwhile, Appel et al. (1997) developed a gel analysis tool, Melanie II, that detects spots via a nonparametric algorithm, and quantifies the spots by direct quantification or Gaussian modeling. See Appel et al. (1997) for details regarding the spot detection procedure. Direct quantifications include determination of the spot area, volume, percent volume, optical density, and percent optical density, where each result is computed using the pixels defining the shape of the detected spot.

2.6 Issues and implications

Several algorithms are presented above for preprocessing 2-DE data. While the ultimate goal is consistently enforced with each algorithm, none has been generally accepted in the scientific community. Much of the hesitation to accept one approach as a gold-standard stems from several issues associated with such analyses, particularly as they relate to 2-DE data. These concerns include (1) addressing an accepted order of operations for preprocessing, (2) automating these procedures, and (3) success in detecting low-lying protein spots, and potential false positives.

Any or all low-level analysis procedures can be performed to obtain summary information on the raw 2-DE data. The order of operations for these algorithms, however, are inconsistent and generally unrecoverable (Coombes et al., 2005). As a result, the preprocessed gel data can vary, thus potentially causing severe repercussions in the high-level analysis. To this end, one should be mindful of the low-level analyses performed (along with their order of operations) and comfortable with their use in data preprocessing.

Establishing an automated preprocessing procedure is ideal in order to remove variation due to analyst subjectivity. This is not currently possible, however, because preprocessing algorithms generally require additional analyst input in order to determine thresholding parameters or local window ranges for consideration. Inputting different parameters can likewise have an impact on the resulting preprocessed data, thus potentially affecting high-level analyses.

Initial proteomic studies have already given rise to the study of easily detected protein spots containing statistically significant differential expression. Smaller proteins (represented as smaller spots), however, may likewise provide valuable information about diseases and associated cures. One's ability to detect these spots, however, is under question. Depending on the spot size, it is possible that the spot is removed along with the image noise, thus eliminating the opportunity for further investigation. On the other hand, limiting the amount of noise removed results in a high false discovery rate – what is believed to be true signal image actually contains additional noise. Thus, detecting low-lying or small protein spots is a concern deserving some attention.

These issues are not easily solvable, illustrating the nonexistence of a uniform approach towards the low-level analysis of proteomic data. Nonetheless, data preprocessing results in the $P \times N$ summary matrix, $X = (x_{pn})$, where x_{pn} denotes the normalized measure of protein p in sample n. This data matrix will be used for subsequent high level statistical analysis.

	Reject H_0	Fail to reject H_0
H_0 true	Type I error	Correct decision
H_0 false	Correct decision	Type II error

Table 1. Possible outcomes for a null hypothesis, H_0, and associated outcome (rejecting or failing to reject H_0).

3. Answering the biological question

Irrespective of one's choice in procedure, the image data are summarized in a matrix fashion, say $X_{P\times N}$, where P denotes the number of proteins and N denotes the sample size. Various statistical issues arise at this point where we have now obtained the "proteomic expression matrix" and are interested in addressing biological questions. They most commonly include statistical inference via hypothesis testing, and multiple testing.

Hypothesis testing seeks to perform statistical inference regarding a question of interest, where the null hypothesis (H_0) defines the status quo statement and the alternative hypothesis (denoted H_1 or H_a) is that which seeks to be proven. Naturally, one wants to make a correct decision when performing a hypothesis test, however, there are four possible scenarios that can occur when performing such a test; see Table 1. Two scenarios represent correct decisions, while the other two are considered errors: (1) when one rejects the null hypothesis when it is actually true, and (2) when one does not reject the null hypothesis when it is actually false. The probability associated with the first scenario is referred to as Type I error (denoted α), and the second scenario's probability is termed Type II error (denoted β). Statistical power refers to the probability of rejecting the null hypothesis when (in fact) the null hypothesis is false; i.e. statistical power equals one minus the Type II error (i.e. $1-\beta$).

Section 3.1 discusses the statistical tests used in the analysis of differential expression in gel electrophoresis, outlining these options and comparing their impact. Meanwhile, when performing even one hypothesis test, analysts want to minimize the error probabilities. The number of proteins studied, however, can be quite large and the proteins should be simultaneously analyzed. Section 3.2 describes the multiple testing problem in general and in relation to the application of gel electrophoresis data, discussing various proposed methods for resolving this issue. In this situation, there are usually many more variables (proteins) than samples (gels). Hence, rather than developing a global model containing all of the proteins, the analysis for this data commonly consists of testing each protein for significance.

3.1 Choice of statistical test

Various statistical tests can be performed, depending on the biological question of interest. The choice of statistical test is important because one faces potential inferential consequences based on the test selected.

A parametric test refers to a test that assumes an underlying distribution. Classical hypothesis testing approaches often assume data that are normal and homoskedastic. As with any application, and particularly here with gel electrophoresis data, such assumptions would need to be justified to attain proper inference. Thus, assuming normality, one can perform a t-test to compare two groups, or an analysis of variance (ANOVA) or covariance (ANCOVA) to compare more than two groups (Sheskin, 2004). Pedreschi et al. (2008) further note that the variance among low intensity spots is smaller than that of high intensity spots, thus some

	H_0 Retained	H_0 Rejected	Total
H_0 True	C	A	m_0
H_0 False	B	D	$m - m_0$
	$m - R$	R	m

Table 2. A summary of results from analyzing multiple hypothesis tests.

form of data transformation should be applied to attain homoskedasticity. The Q-Q plot or Shapiro-Wilk test can be used to assess normality.

Alternatively, nonparametric tests do not presume a distributional assumption. To compare two groups, for example, one can perform a Mann-Whitney U test to compare two groups or, for more than two groups, a nonparametric analog to the ANOVA (e.g. the Kruskal-Wallis ANOVA) can be performed. Jensen et al. (2008) discuss proteomic data analysis based on a small number of gels, where they combine a nonparametric randomization test, and a multivariate method involving a partial least squares regression using jack-knifing for parameter estimation. They advise using a nonparametric randomized test for comparing protein spots across two groups, because the distribution of spots tends to be non-normal. Multivariate methods used were principal component analysis (PCA) and partial least squares (PLS) regression, where parameter estimation and statistically significant differences between the spots were identified via a modified jack-knife method. Jack-knifing in PLS is effective for variable selection, however using this approach does not ensure that all relevant variables are selected. Further, variable selection based on the PLS regression is impacted by the choice of scaling.

3.2 Multiple testing

Protein study is a complex task, e.g. trying to understand how proteins respond to various diseases and to each other. Accordingly, the potential underlying statistical complexities are significant and likewise complex. A naïve yet manageable approach toward proteomic data analysis is to first consider each protein separately, ignoring any possible protein interdependence. The large number of hypothesis tests, however, leads to a potentially high number of false positive results, i.e. proteins being falsely identified as differentially expressed. We want a hypothesis testing approach that maintains a high level of sensitivity and specificity while performing these numerous inferences simultaneously. Bonferroni correction is a conservative approach for controlling Type I error where, given m hypothesis tests, the measure for statistical significance is now attained if the associated p-value is less than α/m. In other words, the significance level is now scaled by the number of hypothesis tests. While this approach successfully adjusts for multiple tests, the procedure is far too conservative, that is, it fails to detect a large number of true positives. Less conservative Type I error rates include the (generalized) familywise error rate [(g)FWER], and false discovery rate (FDR). Table 2 aids in the following discussion, generalizing the hypothesis testing procedure for m tests.

The gFWER is a generalized version of the familywise error rate (FWER), where one wants to control the probability of committing one or more false discoveries. If we let A denote the number of false positives from m hypothesis tests, gFWER is expressed as

$$Pr(A \geq k) \leq \alpha, \tag{1}$$

where $k \geq 1$ and α are usually determined prior to the analysis; FWER defines the special case where $k = 1$ (see Miecznikowski et al. (2011) for an overview of gFWER methods).

Meanwhile, the false discovery rate (FDR),

$$FDR \equiv E[A/R], \tag{2}$$

is an alternative Type I error such that $FDR \leq \alpha$ (Benjamini & Hochberg, 1995). The Benjamini and Hochberg (BH) method is popular for controlling the FDR. In the BH multiple testing procedure, the FDR is controlled by the following:

1. let $p_{(1)} < \ldots < p_{(m)}$ denote the m ordered p-values (smallest to largest);
2. denote $\hat{t} = p_{(k)}$ for the largest k such that $p_{(k)} \leq \frac{k\alpha}{m}$;
3. reject all null hypotheses, H_{0i}, for which $p_i \leq \hat{t}$.

Storey (2002) show that, for p-value threshold t,

$$\text{FDR}(t) = \frac{(1-\pi)t}{(1-\pi)t + \pi F(t)}, \tag{3}$$

where π is the probability that an alternative hypothesis is true, and $F(t)$ is the distribution of p-values given the alternative. FDR analysis does not control the realized FDR, i.e. the number of false rejections A divided by the number of rejections R (Genovese & Wasserman, 2004; Gold et al., 2009).

4. Data imputation

Missingness can exist in the expression matrix if spots are not detected across a required number of gels, or (more simply) when a spot detected in one gel is not detected in other gels. Data missingness can be caused by technical issues or biological variation. Pedreschi et al. (2008) particularly attribute missingness to spots falling below a threshold, mismatches caused by distortions in the protein pattern, absent spots due to bad transfer from the first to the second dimension, or absent spots from the samples. Random causes of data missingness include differences in protein expression across experimental groups, background noise, insufficient spot resolution, or detection. However, "missing values may imply a decrease in the levels of proteins or a shift in the migration of proteins due to post-translational modifications" (Chang et al., 2004). No matter the cause, at least 30% of the data points (and as much as around 50%) may be missing in a 2-DE analysis (Grove et al., 2006; Miecznikowski et al., 2010; Pedreschi et al., 2008).

Various approaches for handling missingness include substituting missing values with zeroes, omitting protein spots that contain missing values, and running replicate samples. Each of these solutions, however, can have a detrimental impact on 2-DE analyses. While one can simply remove the protein spots containing missing data, this is not ideal as valuable information is being discarded. Omitting protein spots severely limits the amount of potential data for analysis, thus reducing the statistical power. Substituting missing values with zeroes can be justified for spots that are below the threshold value, however in cases where the missingness is due to mismatching, substitutions with zeroes leads to improper inferences. Running replicate samples is costly and possibly ineffective because newly generated samples introduce added variation compared with the original samples. Thus data imputation

procedures are an attractive solution for the missingness in gel electrophoresis problems. Data imputation seeks to replace the missing data with reasonable estimates so that investigators may obtain as much information via inference as possible.

Miecznikowski et al. (2010) compare four approaches for data imputation: row averaging, k nearest neighbors, least squares, and nonlinear iterative partial least squares (NIPALS). In row averaging, a missing protein value is replaced by the corresponding average value across nonmissing values associated with that protein. The k nearest neighbors approach instead uses a distance metric to identify closely related proteins, and imputes the average of those nearby protein elements. Both methods can be computed via the `impute` package in *R* (R Development Core Team, 2008). Originally designed for microarrays, the least squares method estimates missing values via correlations between spots and gels. This method is implemented using the JAVA package, `LSimpute` (http://www.ii.uib.no/ trondb/imputation). Finally, the NIPALS method works in a manner similar to principal component analysis, using projections to latent structures to find optimal regression equations and transforms back to impute the missing value. This computation is implemented in *R* using the package, `pcaMethods`. The least squares imputation approach with expectation maximization (EM) used to estimate missing values with an array covariance structure produces the best results in terms of root mean squared error. Further, the bootstrapped versions of the statistical tests are most liberal for determining protein spot significance. Meanwhile, Pedreschi et al. (2008) compare the NIPALS algorithm with k nearest neighbors (with $k = 20$) and Bayesian principal component analysis (BPCA). Their analysis found the BPCA imputation method to be most consistent in its selection of proteins that would be selected if there was no missingness.

5. Case study

In this section, we apply the previously discussed methods to illustrate the research strategy on a real dataset. Our dataset is designed to study the proteomic effects from placing a clamp on an artery feeding the myocardium of swine (pig) hearts. The swine were sacrificed and examined after two months and three months post insertion of the clamp designed to simulate hibernating myocardium. A third (SHAM) condition was also used where the swine did not receive a clamp. This SHAM condition can be considered as a control.

5.1 Experimental design

The dataset consists of an 11-gel experiment to examine the effects at two and three months, respectively, after the insertion of a clamp on an artery feeding the myocardium of swine. Table 3 shows the arrangement of the conditions with each dye (channel) of the gel. We see that this experiment implements dye flips, i.e., the Cy3 and Cy5 channels both contain two-month and three-month conditions, respectively. The primary question of interest is to determine the proteins that are differentially expressed in the following comparisons: two-month versus three-month, two-month versus SHAM, and three-month versus SHAM.

5.1.1 Preprocessing

The DeCyder (GE Healthcare) software was used for image processing where a logarithm normalized spot volume representing the expression level for each protein is obtained for each protein spot (approximately 2230 spots on each gel). The data were organized into an

Gel #	Cy3	Cy5	Cy2
192	three-month	two-month	SHAM
290	two-month	three-month	SHAM
306	three-month	two-month	SHAM
310	three-month	two-month	SHAM
522	three-month	two-month	SHAM
698	three-month	two-month	SHAM
712	two-month	three-month	SHAM
728	two-month	three-month	SHAM
745	two-month	three-month	SHAM
746	two-month	three-month	SHAM
763	three-month	two-month	SHAM

Table 3. Design of the gel experiment. 11 gels with *dye swaps* for the two-month and three-month phenotypes.

expression matrix where the rows correspond to proteins and the columns correspond to a gel/condition combination. The missing data were imputed using the row average method. This imputation choice is selected for illustrative purposes only. While this method is not optimal in terms of normalized root mean squared error (Miecznikowski et al., 2010), it is easy to implement and is one of the few methods with the ability to estimate a missing condition, e.g. implementing the missing SHAM condition for Gel 306. Subsequent to missing data imputation, the data were normalized using the quantile normalization method (see Bolstad et al. (2003)). Figure 1 displays the density for the gel channels before and after missing data imputation and quantile normalization.

5.2 Methods

We assume a linear model, $E[\mathbf{y}_j] = X\boldsymbol{\alpha}_j$, for protein j, where $\mathbf{y}_j$ contains the associated normalized log protein volume, X is the design matrix, and $\boldsymbol{\alpha}_j$ is a vector of coefficients. Here, $\mathbf{y}_j^T$ is the jth row of the normalized protein data matrix and contains the log-volumes for protein j across the three phenotypes (SHAM, month 2, and month 3) and the 11 gels (see Table 3).

Let $X_{j,i}^g$ denote the normalized log protein volume for protein j, phenotype i, and gel g; e.g. $X_{2,3}^4$ is the normalized log protein volume for protein spot 2, under the three-month condition, obtained from the fourth gel. For protein j, the model is given by (subscript j is suppressed for clarity):

$$E\begin{pmatrix} X_S^1 \\ X_2^1 \\ X_3^1 \\ \vdots \\ X_S^{11} \\ X_2^{11} \\ X_3^{11} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ & \ddots & \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}. \tag{4}$$

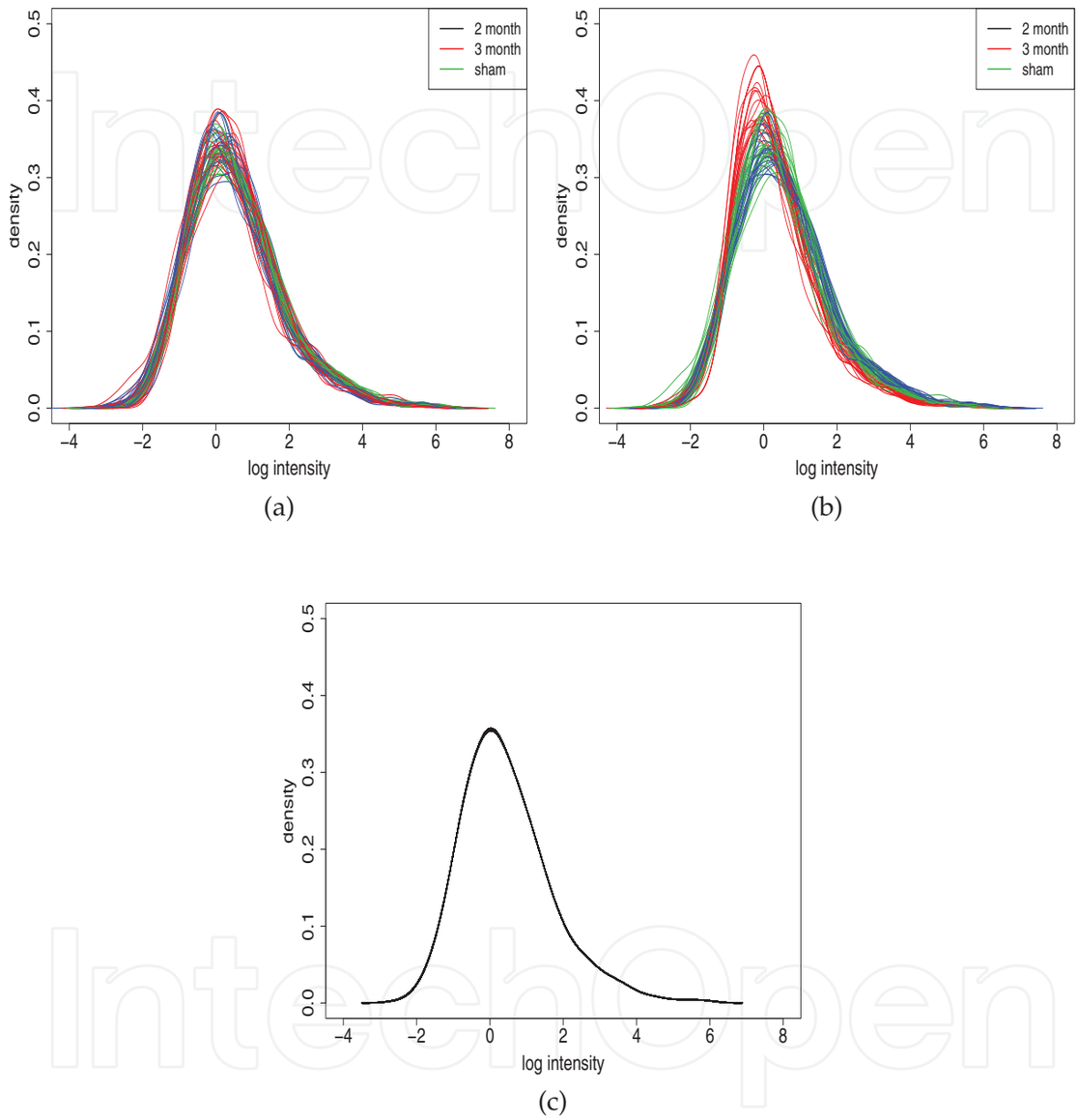


Fig. 1. **Imputation and Normalization**: (a) The density for each log transformed gel condition prior to missing data implementation and quantile normalization. (b) The density for each gel after using a row average method to implement the missing data, but prior to quantile normalization. (c) The density for each gel after missing data implementation and quantile normalization. Each gel/channel combination is normalized to have the same distribution.

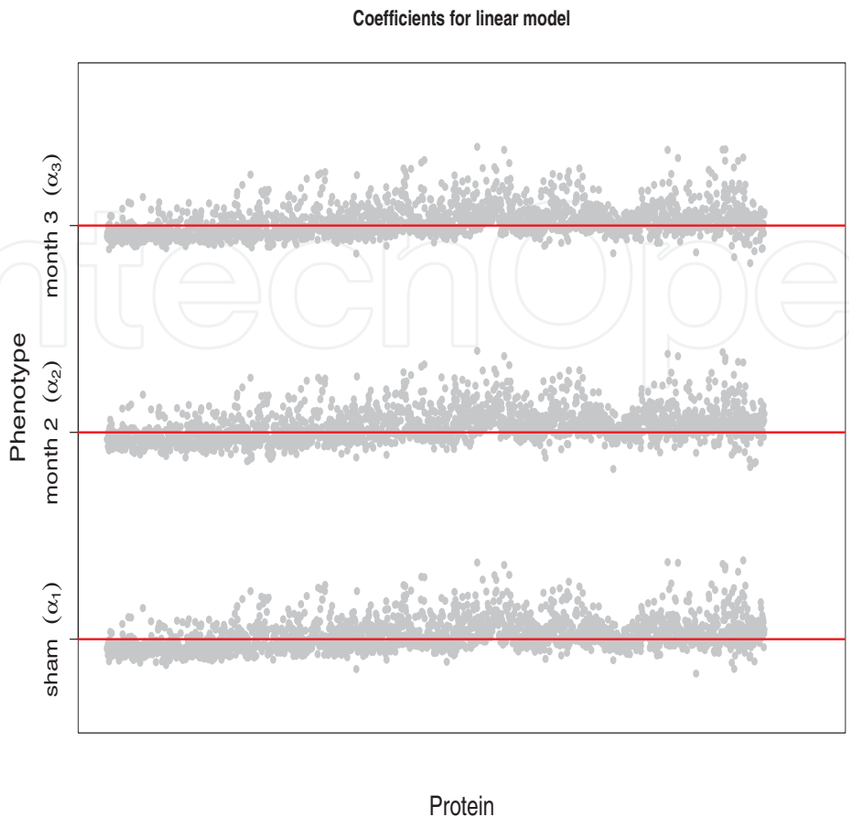


Fig. 2. **Coefficients from a linear model**: The coefficients from a linear model showing the mean estimates for each protein under each condition. The red line is at zero for each set of coefficients.

The vector, $\boldsymbol{\alpha}_j = (\alpha_1, \alpha_2, \alpha_3)$, is the vector of means for protein j under the SHAM, two-month, and three-month conditions, respectively.

The contrasts for protein j are given by $\boldsymbol{\beta}_j = C^T \boldsymbol{\alpha}_j$, where C is the contrast matrix. The contrasts of interest for protein j are specified via

$$\boldsymbol{\beta}_j = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} \alpha_1 - \alpha_2 \\ \alpha_2 - \alpha_3 \\ \alpha_1 - \alpha_3 \end{pmatrix}. \tag{5}$$

Since $\boldsymbol{\alpha}_\cdot$ measures the log volume of protein, the contrasts are equivalent to a log ratio of protein volumes under two specific conditions. The statistical hypothesis test for protein j is given by

$$H_{0,j} : \boldsymbol{\beta}_j = 0$$
$$H_{1,j} : \boldsymbol{\beta}_j \neq 0.$$

We fit the model specified in (4) via least squares and likewise use our estimates of α to obtain estimates for β_j as defined in (5). The estimated coefficients for each protein are given in Figure 2. Note that proteins near the right edge of Figure 2 have larger, more variable coefficients. This anomaly is expected since these proteins are near the bottom edge of the gel and are

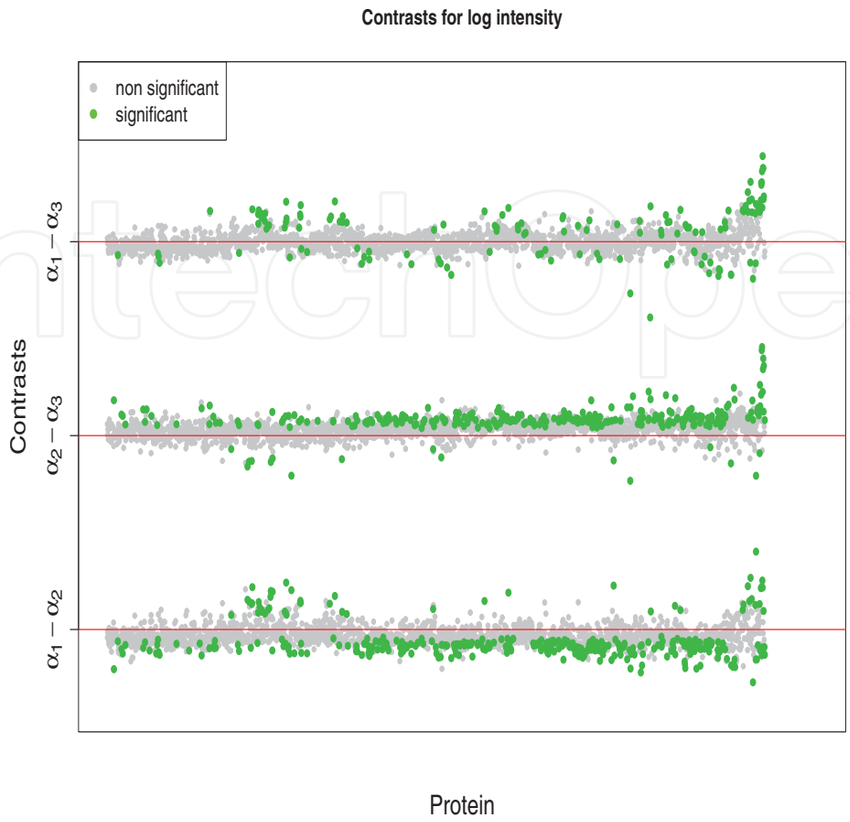


Fig. 3. **Contrasts of Interest**: The estimated contrasts of interest for each protein. The red line is at zero for the contrasts, and the spots in green indicate significant contrasts (see Methods). The potential bias in the contrasts involves α_2. Many of the significant proteins in the $\alpha_1 - \alpha_2$ and $\alpha_2 - \alpha_3$ contrasts are overexpressed in α_2 (two-month) condition. This may be due to the imputation procedure (see Results).

expected to be diffuse, thus making quantification difficult. We fit each set of contrasts as shown in Figure 3.

With the estimated contrasts in Figure 3, we can answer the scientific question of determining differential expression. Here, the issue of multiple comparisons arises where we must evaluate the significance of each single contrast in light of the whole set of contrasts. Note that there are two sets of multiple comparisons in the sense of *multiple testing* across proteins as well as *multiple comparisons* across contrasts for a given protein. We will employ a moderated t-statistic to examine each individual contrast. The statistic is "moderated" in the sense that an empirical Bayes approach is employed to estimate the variance for a given protein. The protein specific variance is augmented with a global variance estimator computed from the data on all proteins (Smyth, 2004).

To examine the question of multiple comparisons across contrasts, we use an approach based on a nested F score (see Chapter 14 in Gentleman et al. (2005)). In this method, the moderated t-statistic for a particular contrast is called significant at level α (resulting from multiple testing across proteins) if the moderated F-test for that protein is still significant at level α when

setting all the larger t-statistics for that gene to the same absolute value as the t-statistic in question. In this case, we want the proteins that are differentially expressed when fixing the significance level α of the moderated F-test so that it corresponds with a false discovery rate (FDR) of 0.05.

5.3 Results

When controlling the moderated F-test so that it corresponds to an FDR of 0.05, we declare 345, 300, and 137 proteins, respectively, as statistically significant for the two-month versus SHAM ($\alpha_1 - \alpha_2$), two-month versus three-month ($\alpha_2 - \alpha_3$), and three-month versus SHAM ($\alpha_1 - \alpha_3$) conditions. Table 4 gives the first 10 significant proteins and the specific contrasts that were significant (in bold), as well as the F-test statistic for each protein spot. This large number of proteins with contrasts involving the two-month condition may be due to the relatively large amount of imputation required in that condition. Using a row average method to impute the data for the two-month group may have artificially shrunk the variance for that condition leading to a potentially large number of false positives when estimating the contrasts with the two-month condition.

Protein	$\alpha_1 - \alpha_2$	$\alpha_2 - \alpha_3$	$\alpha_1 - \alpha_3$	F
23	**-1.025**	**0.910**	-0.116	40.862
61	**-0.399**	0.286	-0.113	7.953
120	**-0.480**	**0.673**	0.192	12.476
131	-0.396	**0.674**	0.278	8.938
166	**-0.392**	0.272	-0.120	9.692
171	**-0.586**	0.207	**-0.378**	20.951
174	**-0.656**	0.110	**-0.546**	18.296
272	**-0.546**	0.344	-0.202	7.721
339	**-0.485**	0.293	-0.192	10.284
360	**-0.476**	0.330	-0.146	8.923

Table 4. **Summary Table**: Table shows the first 10 significant contrasts when controlling the FDR at 0.05. The bolded contrasts indicate the significant differences as determined using the "nestedF" method described in the Methods section. Also included is the F-test for the hypothesis test described in the Methods section.

6. Discussion

Gel electrophoresis experiments serve to understand the relationship between sample groups and protein change, either via differential expression and/or modification. This chapter has outlined many of the statistical issues associated with analyzing such data, but there are other areas that may be of interest to readers.

Cluster analysis seeks to find patterns in data, thus identifying similar qualities that are shared among proteins or samples. There are various types of clustering algorithms that can be categorized as partitioning, hierarchical, or hybrid algorithms. Gentleman et al. (2005) introduce these various types of algorithms in greater detail, and describe how to determine the number of clusters for analysis, and a visualization tool helpful for analysis (heat maps).

Protein networks and interactions are also an area of great interest. Such interactions regulate cellular function, and thus understanding how they interact is a source of interest

to systems biologists. See Urfer et al. (2006) for an overview of statistical methods for protein interaction, including discussion and comparison of various graphical models. The interested reader can also refer to Jung (2010) for discussion regarding experimental design options, particularly one experimental factor with two or more categories, and design with two or more experimental factors.

Data visualization is a statistical subarea rich in techniques for detecting protein changes in proteomic research. While such tools do not alter the inherent data, they can prove fruitful in displaying the data, e.g. in exploratory data analysis and in providing visual representations of statistical results. See Cleveland (1993) and Cleveland (1994) for discussion of graphical tools used to visualize data, as well as other strategies for better data comprehension and study.

7. Acknowledgements

The authors are grateful to Dr. John M. Canty, Jr., the Albert and Elizabeth Rekate Professor of Medicine in the Department of Medicine at SUNY University at Buffalo, for generously allowing us to use the swine data in our case study.

8. References

Adams, R. & Bischof, L. (1994). Seeded region growing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Appel, R., Vargas, J., Palagi, P., Walther, D. & Hochstrasser, D. (1997). Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms, *Electrophoresis* 18: 2735–2748.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

Bolstad, B., Irizarry, R., Åstrand, M. & Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics* 19(2): 185.

Chang, J., Van Remmen, H., Ward, W., Regnier, F., Richardson, A. & Cornell, J. (2004). Processing of data generated by 2-dimensional gel electrophoresis for statistical analysis: missing data, normalization, and statistics., *Journal of Proteome Research* 3(6): 1210.

Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press.

Cleveland, W. S. (1994). *The Elements of Graphing Data*, Hobart Press.

Conradsen, K. & Pedersen, J. (1992). Analysis of two-dimensional electrophoretic gels, *Biometrics* 48(4): 1273–1287.

Cooley, W. & Lohnes, P. (1971). *Multivariate data analysis*, J. Wiley.

Coombes, K., Tsavachidis, S., Morris, J., Baggerly, K. & Kuerer, H. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform, *Proteomics* 5: 4107–4117.

Genovese, C. & Wasserman, L. (2004). A stochastic process approach to false discovery control, *The Annals of Statistics* 32(3): 1035–1061.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch,

F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. & Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics, *Genome Biology* 5: R80.

Gentleman, R., Carey, V., Huber, W., Dudoit, S. & Irizarry, R. (2005). *Bioinformatics and computational biology solutions using R and Bioconductor*, Springer Verlag.

Gold, D., Miecznikowski, J. & Liu, S. (2009). Error control variability in pathway-based microarray analysis, *Bioinformatics* 25(17): 2216.

Grove, H., Hollung, K., Uhien, A. K., Martens, H. & Faergestad, E. M. (2006). Challenges related to analysis of protein spot volumes from two-dimensional gel electrophoresis as revealed by replicate gels, *Journal of Proteome Research* 5: 3399–3410.

Gustafsson, J. S., Blomberg, A. & Rudemo, M. (2002). Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern, *Electrophoresis* 23(11): 1731–1744.

Huang, H., Stasyk, T., Morandell, S., Dieplinger, H., Falkensammer, G., Griesmacher, A., Mogg, M., Schreiber, M., Feuerstein, I., Huck, C. et al. (2006). Biomarker discovery in breast cancer serum using 2-d differential gel electrophoresis/maldi-tof/tof and data validation by routine clinical assays, *Electrophoresis* 27(8): 1641–1650.

Jackson, D. & Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data, *Environmetrics* 15(2): 129–139.

Jensen, K., Jessen, F. & Jorgensen, B. (2008). Multivariate data analysis of two-dimensional gel electrophoresis protein patterns from few samples, *J. Proteome Res* 7: 1288–1296.

Jolliffe, I. (2002). Principal component analysis, *Encyclopedia of Statistics in Behavioral Science* .

Jung, K. (2010). Statistics in experimental design, preprocessing, and analysis of proteomics data, *Data Mining in Proteomics*, Humana Press Inc., pp. 259–272.

Kaczmarek, K., Walczak, B., de Jong, S. & Vandeginste, B. (2004). Preprocessing of two-dimensional gel electrophoresis images, *Proteomics* 4(8): 2377–2389.

Keeping, A. & Collins, R. (2011). Data variance and statistical significance in 2d-gel electrophoresis and dige experiments: Comparison of the effects of normalization methods, *Journal of Proteome Research* 10(3): 1353–60.

Langella, O. & Zivy, M. (2008). A method based on bead flows for spot detection on 2-D gel images, *Proteomics* 8: 4914–4918.

Lemkin, P. F. (1997). Comparing two-dimensional electrophoretic gel images across the Internet, *Electrophoresis* 18: 461–470.

Miecznikowski, J., Damodaran, S., Sellers, K. & Rabin, R. (2010). A comparison of imputation procedures and statistical tests for the analysis of two-dimensional electrophoresis data, *Proteome Science* 8(1): 66.

Miecznikowski, J., Gold, D., Shepherd, L. & Liu, S. (2011). Deriving and comparing the distribution for the number of false positives in single step methods to control k-fwer, *Statistics & Probability Letters* 81(11): 1695–1705.

Miecznikowski, J., Sellers, K. & Eddy, W. (2009). Multidimensional median filters for finding bumps, *Technical Report 0907*, University of Buffalo School of Public Health and Health Professions.

O'Farrell, P. (1975). High resolution two-dimensional electrophoresis of proteins, *The Journal of Biological Chemistry* 250(10): 4007–4021.

Pedreschi, R., Hertog, M., Carpentier, S. C., Lammertyn, J., Robben, J., Noben, J.-P., Panis, B., Swennen, R. & Nicolaï, B. M. (2008). Treatment of missing values for multivariate statistical analysis of gel-based proteomics data, *Proteomics* 8(7): 1371–1383.

R Development Core Team (2008). R: A language and environment for statistical computing. ISBN 3-900051-07-0.

Rai, A. & Chan, D. (2004). Cancer proteomics: serum diagnostics for tumor marker discovery, *Annals of the New York Academy of Sciences* 1022(1): 286–294.

Rao, C. (1964). The use and interpretation of pca in applied research, *Sankhya A* 26: 329–358.

Schulz-Trieglaff, O., Machtejevas, E., Reinert, K., Schlüter, H., Thiemann, J. & Unger, K. (2009). Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments, *BioData mining* 2(1): 4.

Sellers, K. & Miecznikowski, J. (2010). Feature detection techniques for preprocessing proteomic data, *Journal of Biomedical Imaging* 2010: 16.

Sellers, K., Miecznikowski, J., Viswanathan, S., Minden, J. & Eddy, W. (2007). Lights, camera, action! systematic variation in 2-D difference gel electrophoresis images, *Electrophoresis* 28(18): 3324–3332.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical procedures*, CRC Pr I Llc.

Smilansky, Z. (2001). Automatic registration for images of two-dimensional protein gels, *Electrophoresis* 22(9): 1616–1626.

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical applications in genetics and molecular biology* 3(1): 3.

Srinark, T. & Kambhamettu, C. (2008). An image analysis suite for spot detection and spot matching in two-dimensional electrophoresis gels, *Electrophoresis* 29: 706–715.

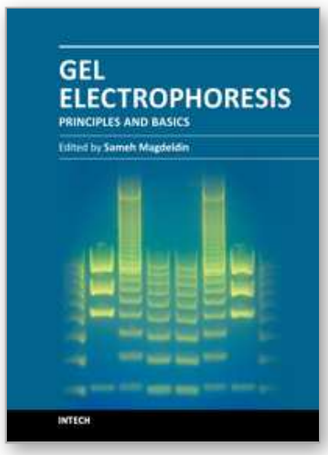
Storey, J. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society. Series B, Statistical Methodology* pp. 479–498.

Ünlü, M., Morgan, M. & Minden, J. (1997). Difference gel electrophoresis. a single gel method for detecting changes in protein extracts, *Electrophoresis* 18(11): 2071–2077.

Urfer, W., Grzegorczyk, M. & Jung, K. (2006). Statistics for proteomics: A review of tools for analyzing experimental data, *Proteomics* 6(S2): 48–55.
URL: *http://dx.doi.org/10.1002/pmic.200600554*

Wulfkuhle, J., Liotta, L., Petricoin, E. et al. (2003). Proteomic applications for the early detection of cancer, *Nature reviews cancer* 3: 267–275.

Zhou, G., Li, H., DeCamp, D., Chen, S., Shu, H., Gong, Y., Flaig, M., Gillespie, J., Hu, N., Taylor, P. et al. (2002). 2d differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers, *Molecular & Cellular Proteomics* 1(2): 117.



Gel Electrophoresis - Principles and Basics

Edited by Dr. Sameh Magdeldin

ISBN 978-953-51-0458-2

Hard cover, 346 pages

Publisher InTech

Published online 04, April, 2012

Published in print edition April, 2012

Most will agree that gel electrophoresis is one of the basic pillars of molecular biology. This coined terminology covers a myriad of gel-based separation approaches that rely mainly on fractionating biomolecules under electrophoretic current based mainly on the molecular weight. In this book, the authors try to present simplified fundamentals of gel-based separation together with exemplarily applications of this versatile technique. We try to keep the contents of the book crisp and comprehensive, and hope that it will receive overwhelming interest and deliver benefits and valuable information to the readers.

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen