Analyzing Two-Dimensional Gel Images

Anindya Roy
Françoise Seillier-Moiseiwitsch
Department of Mathematics and Statistics
University of Maryland, Baltimore County

Kwan R. Lee
Data Exploration Sciences
GlaxoSmithKline Pharmaceuticals R&D

Yaming Hang
Department of Mathematics and Statistics
University of Maryland, Baltimore County

Mark Marten
Babu Raman
Department of Chemical and Biochemical Engineering
University of Maryland, Baltimore County

1. Introduction

Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) is one of the methods able to separate thousands of proteins. Mammalian cell samples exhibit up to 2,000 proteins. On such a two-dimensional gel, two coordinates characterize each protein: its isoelectric point and its molecular weight. Along one dimension, proteins are sorted electrophoretically according to their pH gradient. They stabilize at points where their net charge is zero. Along the other dimension, proteins, soaked in sodium dodecyl sulphate, separate according to their molecular weight. Thus the isoelectric point and the molecular weight uniquely identify a protein spot on a polyacrylamide gel. The separated proteins are then stained with fluorescent dyes so that they are amenable to imaging. The images are scanned and stored in a database. The process, though lengthy and subject to enormous experimental uncertainty, is still much cheaper than other technologies such as mass spectrometry that also quantifies protein expressions. Therefore, it is of interest to see whether the existing methods for analyzing two-dimensional gels can be improved to provide reliable inference.

The challenges plaguing gel analysis are mostly statistical. It is thus surprising that contributions to this field have mainly come from computer scientists. The existing methodologies for gel analysis are mostly in form of computer algorithms that in many cases have been implemented in software package developed specifically for analyzing two-dimensional gel images. Early examples of automated gel analysis techniques can be found in Skolnick (1982), Rowlands et al. (1988) and Appel et al. (1991). Figure 1 shows a typical image of a two-dimensional gel. Just by glancing at it, the reader can imagine how hard a task it is for any automated algorithm to accurately identify hundreds of protein spots among the various kinds of noise, and also to compare and match proteins over several gels when presented with multiple copies of gels made from similar groups of samples. To enhance the quality of gel analysis, current automated algorithms need significant human intervention, which is time consuming and defeats the purpose of automation. Thus, 2D-PAGE, though a thirty years old technique, still has a number issues to be resolved. Because of the nature of the problems statisticians can play a vital role in the analysis of two-dimensional gels.

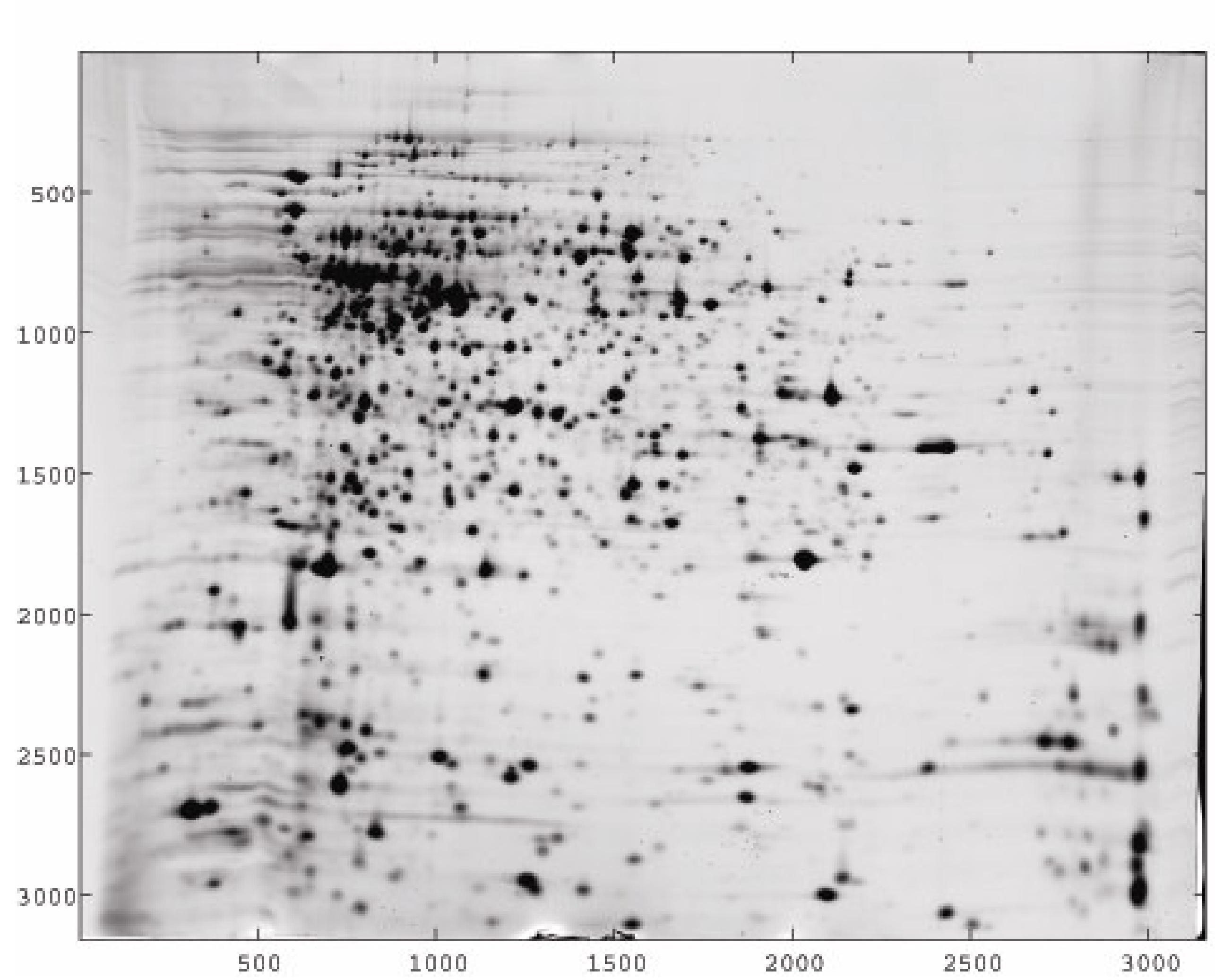

Figure 1: Original Image

The main steps in the analysis of two-dimensional gels are as follows. First, one applies procedures to filter out various kinds of noise present on the images. Image segmentation techniques are then utilized to detect the protein spots. Then one needs to align the images under study: gels may have been differentially warped due to physical transformations involved in the lengthy process of gel preparation. Once the gels have been aligned according to some set rules (usually some landmarks on a reference gel), spots on different gels need to be matched. Once this process is completed, a master gel is created for each group involved in the experiment (e.g., for each individual with replicate samples, or for each treatment group). The data also need to be extracted in terms of spot locations, volumes and shapes before the variational analysis can proceed. The final step is the analysis of qualitative features such as appearance/disappearance of spots along with quantitative features such as the changes in intensity level over individuals.

In this paper we discuss the main problems involved in gel analysis, mainly from a statistician's viewpoint. Reference to the literature is by no means complete, more so because of the fast emerging nature of the field. In Section 2 we focus on noise reduction, spot detection techniques.

Issues in gel alignment and spot matching are the topic of Section 3. The scope for statistical analysis of the aligned gels is described in Section 4. Section 5 contains a sample data analysis. Concluding remarks appear in Section 6.

2. Noise Reduction and Spot Detection

Statistical image analysis is a well-developed field. However, techniques need to be adapted to the subtleties of features of two-dimensional gel images before they can be successfully applied in proteomic research. The images are messy and contaminated with noises from various sources. Scanners used to produce the images often pick up dust particles that show up as black specs on the images. The other prominent features that are not spots are the streaks. Most streaks are horizontal appearing as a result of nucleic-acid contamination or due to the presence of salt on the strips. Sometimes there are water bubbles or hair on the images.

2.1 Single Thresholding for Removing Noise

Thresholding can remove noise and enhance the contrast between the background and the features. However, the intensity of external noise particles potentially covers the whole range of spot intensities. Thus simple thresholding of images cannot get rid of the noise alone. Depending on the degree of thresholding one can loose a significant amount of spots along with the noise or leave out a significant amount of noise along with the spots. Figure 2 is the result of a single thresholding on the original image. One can see that there is significant loss of protein spots along with the noise. Adaptive regional thresholding is more appropriate. Methods where the optimal threshold value is chosen statistically need to be developed. The irregular nature of the noise is certainly a big obstacle in developing such methods.

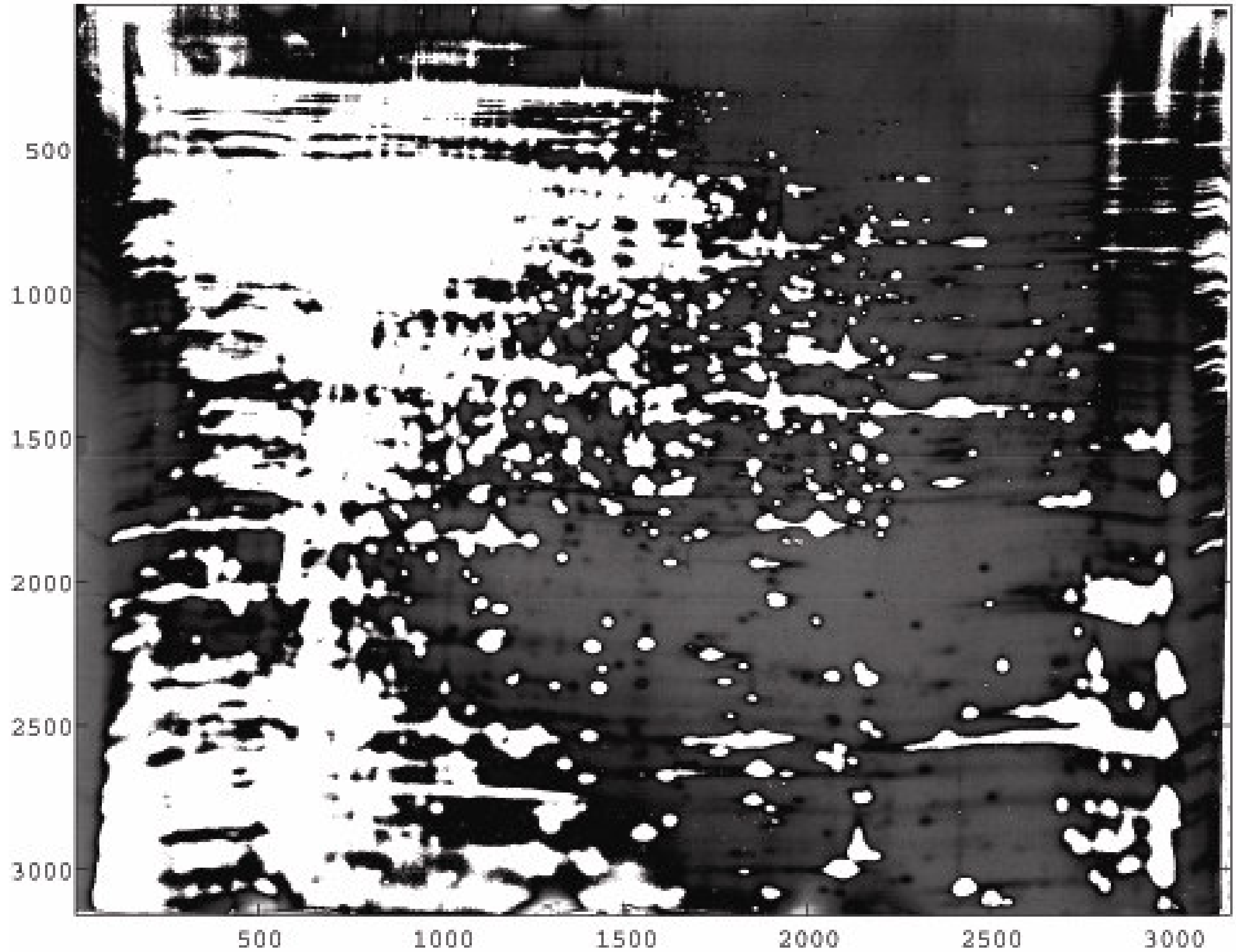

Figure 2: Single Threshold

2.2 Smoothing

Smoothing techniques are applied to remove noise from the images. Gaussian smoothing is often used to remove noise by convolving the gel image with a Gaussian kernel. Isotropic and anisotropic diffusion smoothing are also applied: these replace the image with smooth solutions of diffusion equations. Another popular method is to replace pixel intensities with the median intensity of an appropriately chosen neighborhood of the pixel (usually a 9×9 window around the pixel). Because most of the noise is transient such filters essentially replace the noise with the background. However, the method is sensitive to the distribution of the base area of the noise. Polynomial smoothing fits local smooth polynomials to the entire image. Seillier-Moiseiwitsch et al. (2002) used wavelets to smooth the gel images. Though wavelets are flexible enough to remove much of the noise, they still cannot remove all the unwanted features in their entirety. Figure 3 shows a wavelet reconstruction of the original image with multiresolution level set to 2. Although most of the noise is gone, the streaks remain, and some of the spots are either lost or are very faint. A multiresolution level of 3 brings back the lost spots but some noise reappears as well. Another troubling feature is that the streaks, being quite smooth, reside on all the coarser scales that make up the spots. Thus, simple two-dimensional wavelet cannot adequately remove specs and streaks and retain all the spots. Additional information is needed to distinguish the spots from streaks and spikes.

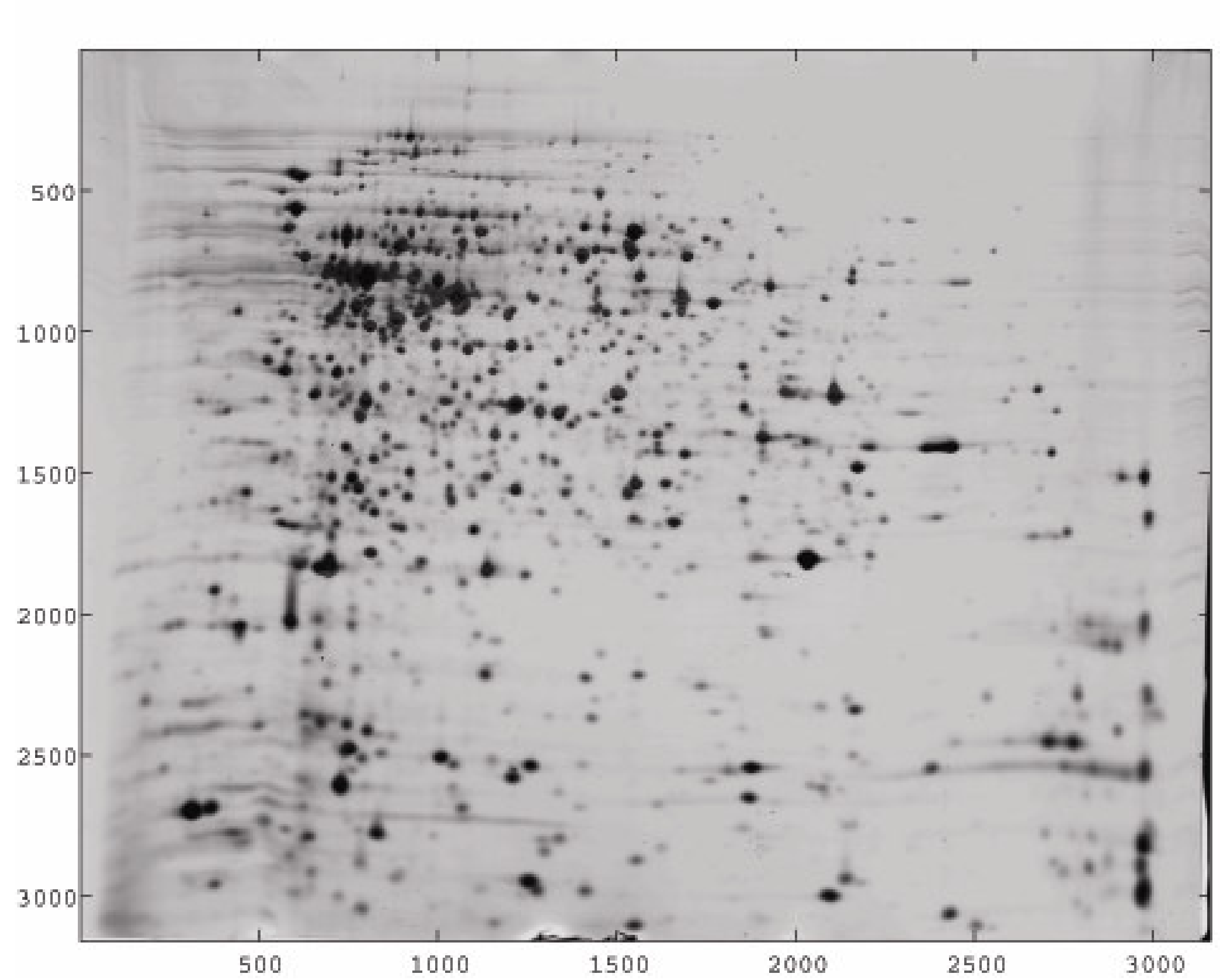

Figure 3: Wavelet Reconstruction: Multiresolution 2.

2.3 Spot Shapes

Another feature that makes it rather hard to distinguish spots from the noise is the irregularity in the shape of the spots themselves. Many current methods for spot detection assume regular shapes such as elliptical contours for the spot (Horgan & Glasbey, 1995). Also, spots are often modeled as bivariate Gaussian density (Appel et al., 1997). However, a closer look reveals that such assumptions are not valid for all spots. Figure 4 shows few spot shapes that are quite common on gel images. The irregularity may come from merging two spots into one. Single spots can have flat tops due to saturation. Figures 5 and 6 show three-dimensional plots of streaks and spikes.

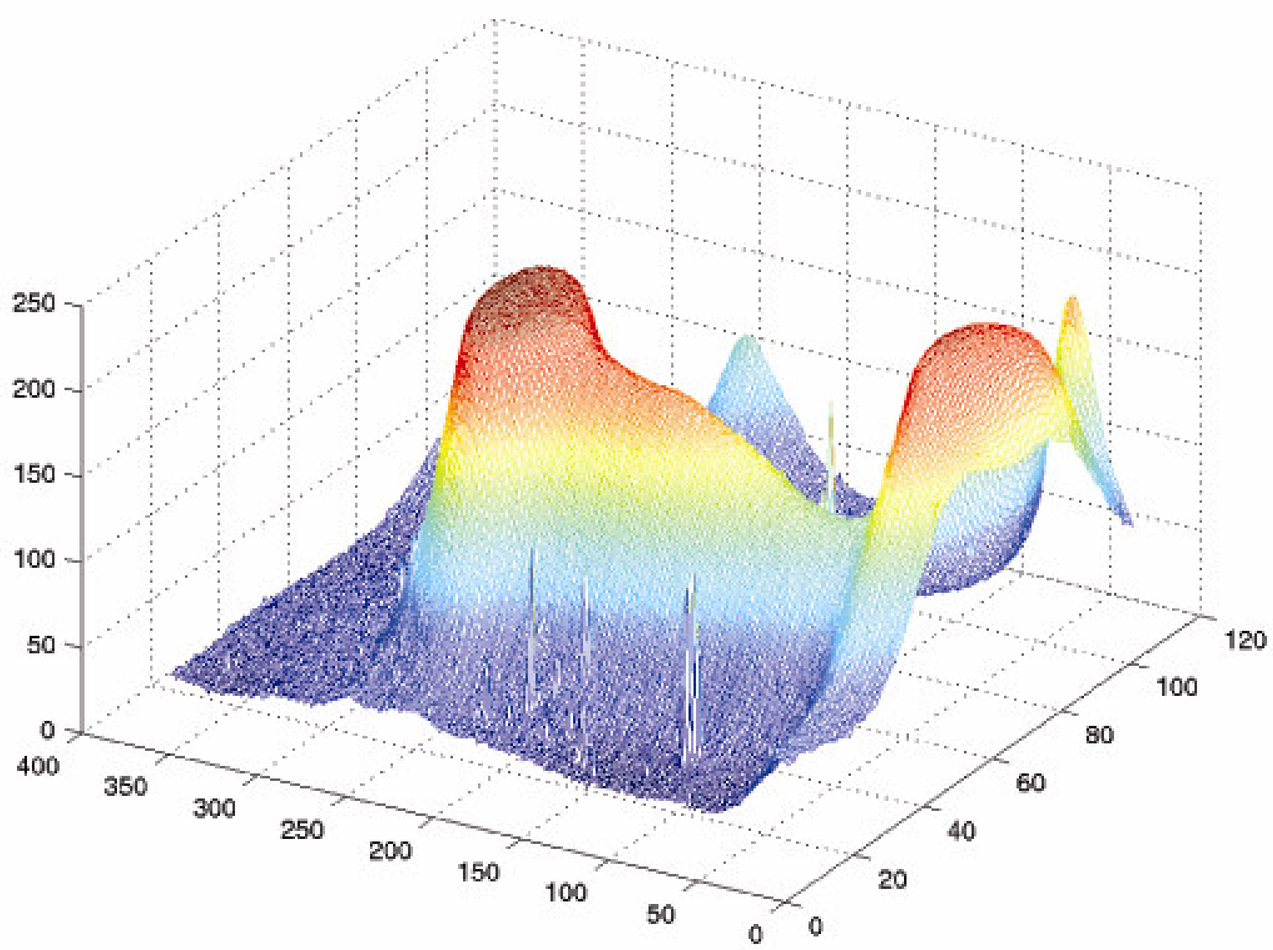

Figure 4: Spot Shapes.

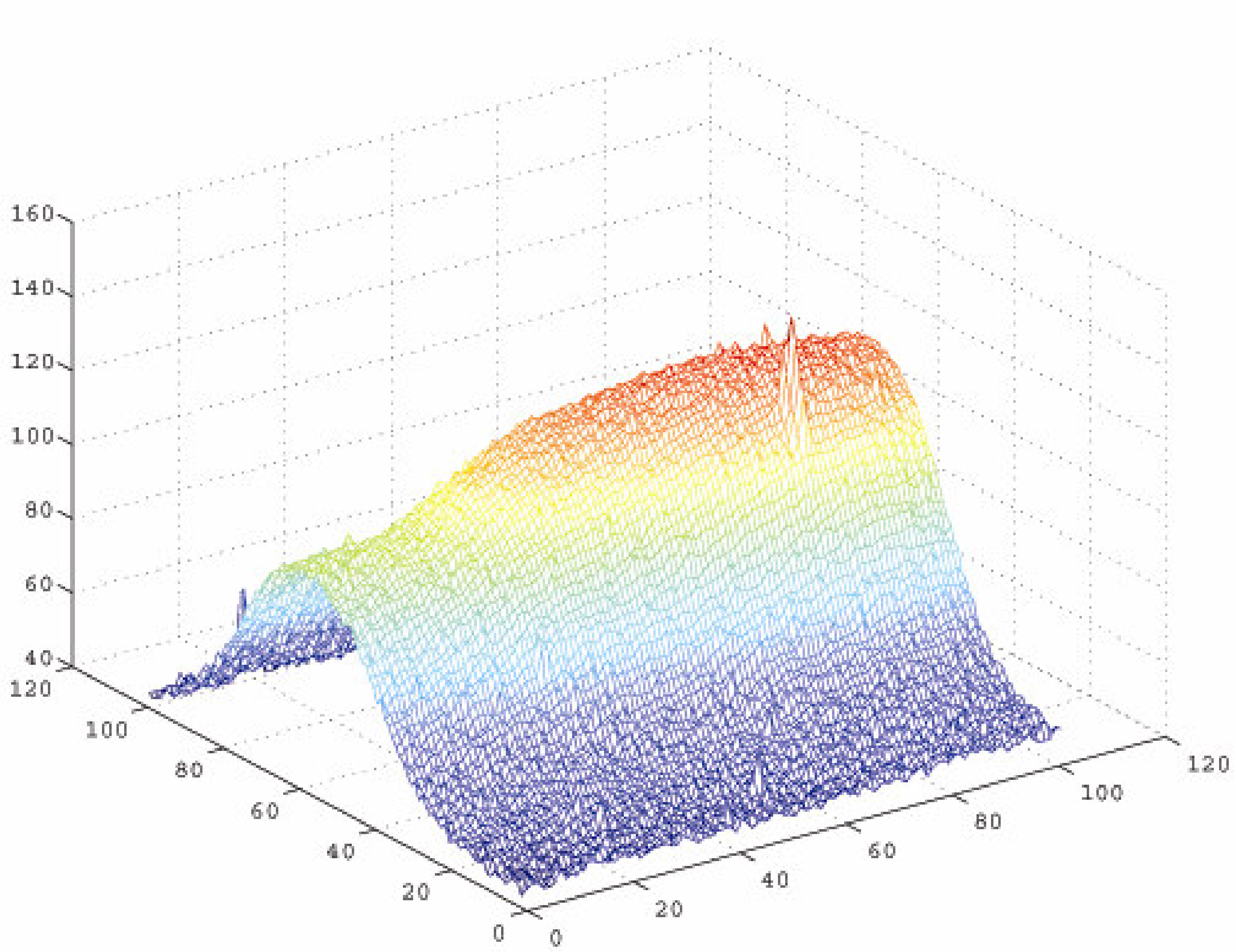

Figure 5: Streaks.

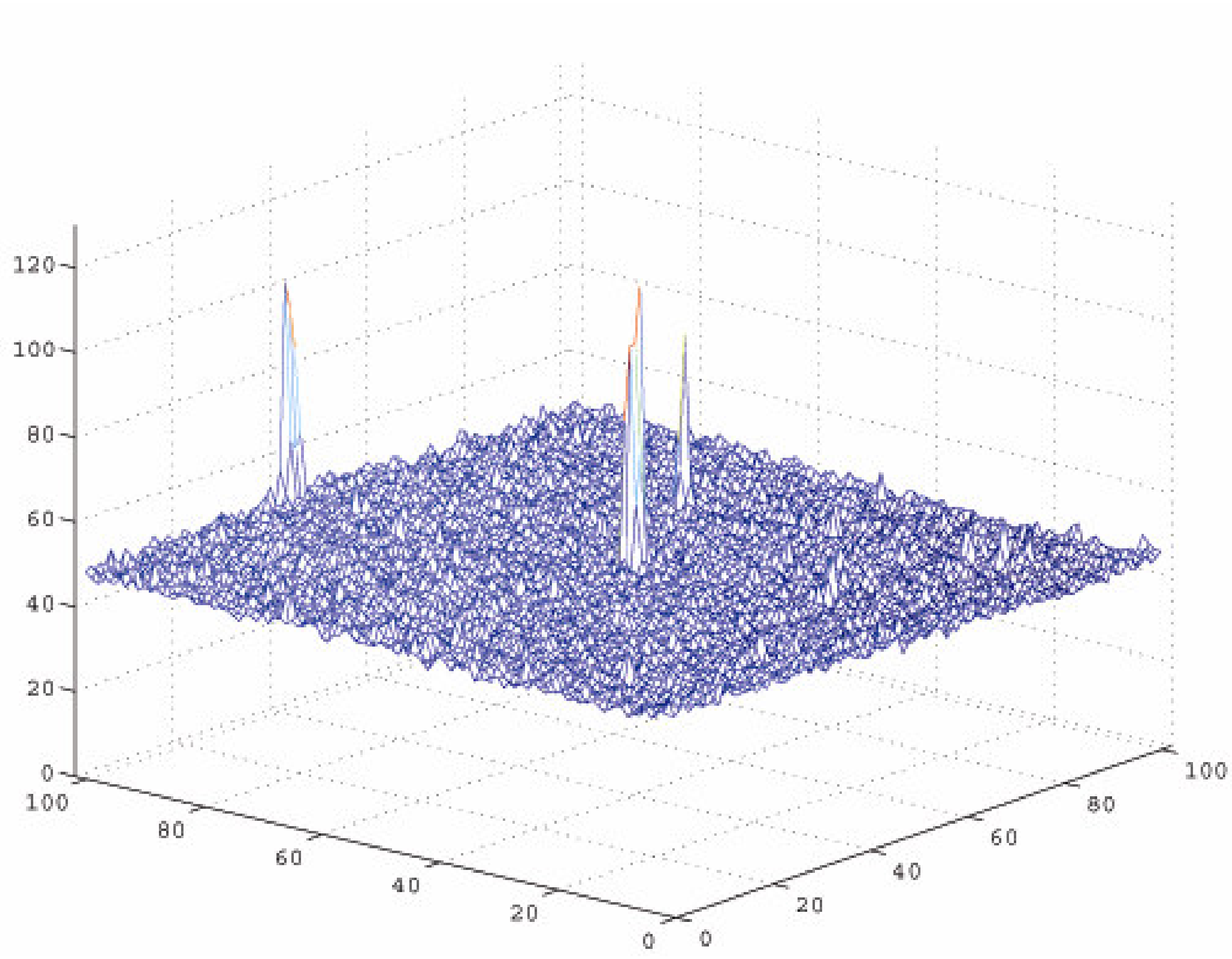

Figure 6: Spikes

2.4 Spot Detection

There are many software packages that perform spot detection in two-dimensional gel images: Melanie, PDquest, Phoretix, Kepler, Gellab to name the most popular ones. Melanie II, for example, uses nonparametric procedures based on the Laplacian of the pixel intensities to decide whether a pixel belongs to a spot or not. Conradsen & Pedersen (1998) uses successive application of second derivative filters and binary erosion for detecting spot edges. Efrat et al. (2002) uses morphological features such as elliptical shapes to fit the spots on gel images. Bettens et al. (1996) uses anisotropic diffusion to build an automatic segmentation method that separates spots from noise and background. Other popular methods are variations of the watershed segmentation algorithm that is used to segment topological reliefs in mathematical morphology. Vincent & Soille (1991) developed fast algorithms for image segmentation using watershed techniques. Pleissner et al. (1999) built hierarchical watershed algorithms along with feature extraction methods that allow for partial merging of spots and thus prevents oversegmentation of the images.

2.5 Quantitative Characteristics for Spots, Streaks and Spikes

In spite of the irregularities in the objects on gel images, the human eye can distinguish between spots, streaks and spikes with much greater certainty than any automated algorithm. It is therefore important to find characteristics that show maximal separation between spots, streaks and spikes. For a spot the following are the commonly used measures for spot quantification:

$$\text{area} = \text{AREA} = \text{number of pixels} \times \text{pixel area}$$

$$\text{optical density} = \text{OD} = \max_{x, y \in \text{spot}} I(x, y)$$

$$\text{percent optical density} = \%\text{OD} = 100 \times \frac{\text{OD}}{\sum_{s=1}^{n} \text{OD}_s}$$

$$\text{volume} = \text{VOL} = \sum_{x, y \in \text{spot}} I(x, y)$$

$$\text{percent volume} = \%\text{VOL} = 100 \times \frac{\text{VOL}}{\sum_{s=1}^{n} \text{VOL}_s}$$

where OD_s and VOL_s are respectively the optical density and the volume of spot s in a gel containing n spots. Empirical studies need to be performed to ascertain distributional properties of these measures.

Streaks run horizontally or vertically, and their intensities tend to be constant along their main axis. The spikes have peaked, symmetric, unimodal shapes, and their base areas are usually much smaller compared to those of spots. Thus, the orientation and axial gradient of the objects are promising characteristics that can be used in statistical analysis. Figure 7 shows the horizontal gradient plot of the wavelet-reconstructed image (we used a smoothed version of the image in order to have stability in the digital gradient of the image). We applied a combination of horizontal and vertical gradient procedures to remove the streaks. Also, the second difference operator was the most successful method for distinguishing spots from spikes. The disadvantage of smoothing over classification is that vital spot information is partially lost along with the spikes. Figure 8 shows the end result of gradient feature extraction and second difference operation. Another possibility for separating streaks from spikes could be to use recently developed function bases such as ridgelets (Candes & Donoho, 1999) which are tailor made for modeling linear features in images.

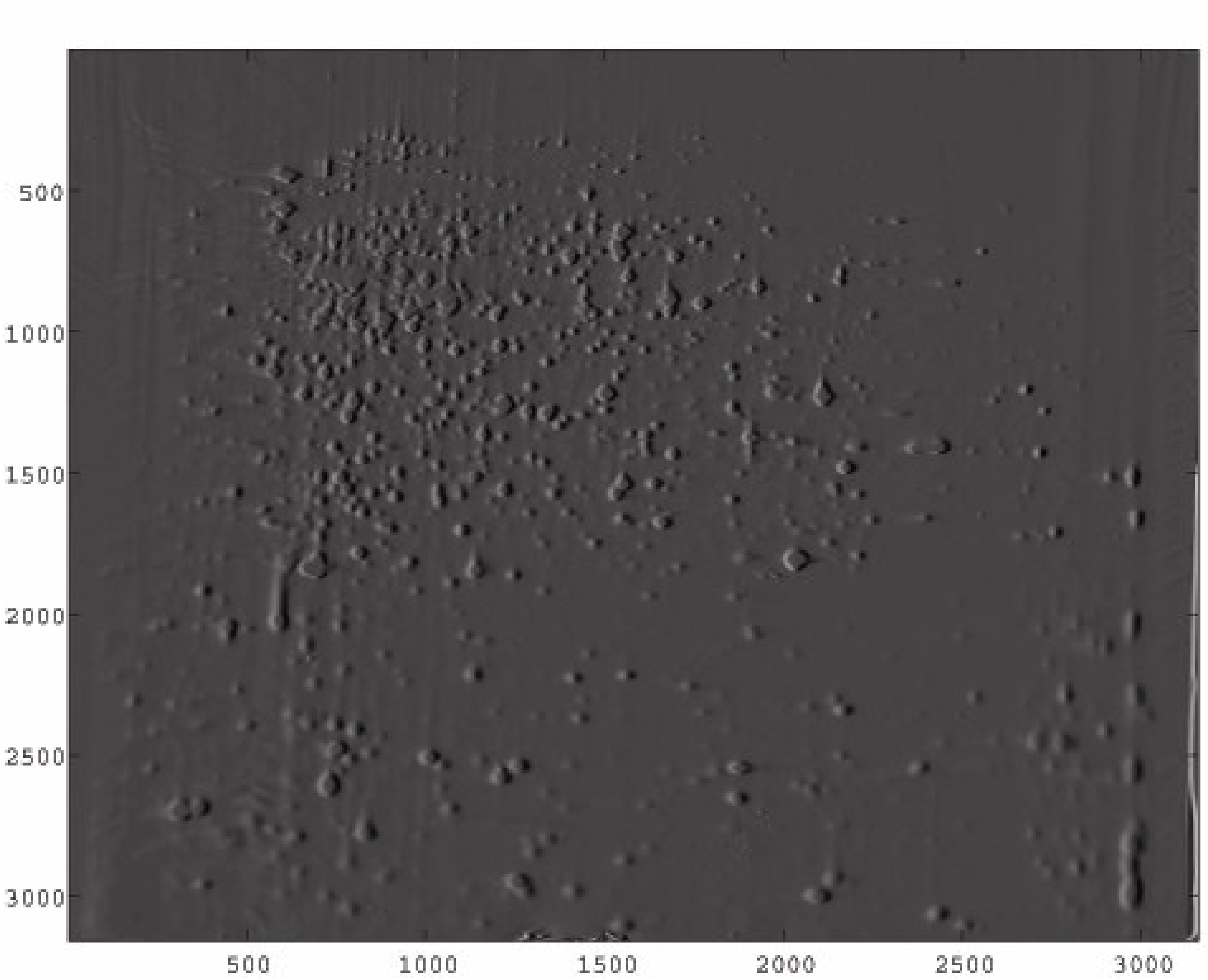

Figure 7: Horizontal Gradient.

Figure 8: Cleaned Image.

3. Gel Alignment and Spot Matching

Gels may undergo several transformations in the process of gel preparation and image scanning. Two gels prepared from the same sample can result in quite different images due to warping. The goal in gel alignment is to identify the warping function $(x,y) \rightarrow (u(x,y),v(x,y))$ from a reference gel to a source gel where (x,y) are the pixel coordinates. Unfortunately, the classes of function one can explore in order to find a good model for the physical transformation is usually quite limited. Due to the locally smooth nature of the warping a natural course for modeling is to segment the images and use simple smooth functions (e.g, lower-order polynomials) to model each segment. Typically, such algorithms assume that the distortions are direction independent, i.e., $(x,y) \rightarrow (u(x),v(y))$. Drawbacks of such algorithms include discontinuities along segment boundaries, arbitrariness of the reference gels and overfitting of local segments.

Initial aligning techniques relied on visual comparison of warped images (Horgan et al., 1992; Appel et al., 1997; Hoffman et al., 1998, 1999; Lemkin, 1997; Pleissner et al., 1997; Smilansky, 2001). Image warping and spot matching has a rich history in the image analysis literature. An excellent review of warping methods can be found in Glasbey & Mardia (1998). Conradsen &

Pedersen (1992) addresses the problem of gel alignment by minimizing disparity between two gels using cross-correlation techniques. Glasbey & Mardia (2001), in a more general image analysis set-up, developed a penalized likelihood method for gel alignment where the likelihood measures the similarity between gels and the penalty function prevents overfitting by restricting the distortion due to warping. The penalized likelihood method is applied to the context of two-dimensional gels in Gustafsson et al. (2002). Other approaches could be to develop warping-invariant methods for modeling gel images where the class of global warping is restricted to simple classes such as affine transformations. Methods such as translation-invariant wavelet fitting (Coifman & Donoho, 1995) may well be suited to such modeling purposes. One can also implement recently popularized optimization techniques such as interior-point methods (Freund & Mizuno, 1996; Potra & Wright, 2000) for optimally choosing the warping transformation.

4. Statistical Analysis

Once the gels are cleaned and aligned, the next step is to analyze the gel-to-gel variation. Unfortunately, the study of gel variation is still in infancy. To our knowledge, there has been no rigorous statistical analysis of gel variation, though partial attempts have been made. Several questions remain unanswered. What are the most appropriate statistical models for two-dimensional gels? If a mixed-effect model is assumed then what are the distribution of the errors? What are the random effects and what are fixed effects? Can parametric classes adequately model gel images? Are there spatial correlations? How does one perform model selection? What is the most appropriate break up of the sources of variation? Many of these questions have definitive answers in the statistical literature for large flexible classes that can handle very general situations. Therefore it seems that, apart from development of statistical methods that are specific to gel analysis, there is immediate scope for novel application of existing statistical methodologies in this context.

5. A Case Study

A 2-d gel experiment may be performed as part of a clinical trial to study biological factors of interest. A biological factor can have two different disease state (sick vs. normal) of patients and/or two different levels of treatment (control vs. treated with a drug). Consider an example of a lung cancer study where a total of 41 different serum samples was obtained from the clinical trial. Among those 41 samples, 24 of them (labeled A1, A2, …, A24) were from the cancer patients and 17 of them (labeled as B25, B26,…, B41) were from the normal patients.

Once the image processing of the gel has been done, the volume of each protein spot can be summarized as a matrix in a popular spreadsheet together with the spot ID, molecular weight (Mw) and isoelectric charge (pI). Typically different spot IDs appear as rows and different gel IDs appear as columns. For our example, there were more than 5,000 rows for the spots and 41 columns corresponding to different sample IDs. Each cell of the spreadsheet contains the percent volume of the protein spots. The percent volume was computed excluding the empty (“missing”?) cells in the spreadsheet.

Typical data mining of these data involves both exploratory data analysis and biomarker selection. Exploratory data analysis may use popular chemometrics tools such as principal component analysis (PCA) to assess reproducibility in the data. Biomarker selection could be the final goal of the experiment. We are interested in the biomarkers (identified proteins for the spots) that are responsible for explaining the biological factor(s). In our example, the disease status is the biological factor of interest.

Many different statistical tools can be used to select biomarkers and there are two approaches: univariate and multivariate approach. The univariate approach may include popular multiple t-test, analysis of variance (ANOVA) or nonparametric test on individual spot. The multivariate approach may involve various different classification analysis tools such as decision trees, support vector machine or classical discriminant analysis. Generally analysis of two-dimensional gel data is not much different from that of gene expression data since both cases involve the analysis of extremely high dimensional data, with the additional problem of many cells being empty for the gels.

These empty cells are not "missing" values in statistical sense since they are not missing at random. Those empty cells (blanks may be a better name than missing values) were created mainly as the result of subtracting background noise from the gel images. If the subtraction of the background noise resulted in a negative or zero value, an empty cell is created instead of an actual value. Other reasons may be the lack of precision in the alignment of several images or spots running into another another. So, any statistical algorithm that handles missing values would be misleading since they assumed the empty cells are missing at random. Those empty cells do not cause fundamental problems for univariate analysis since we can simply ignore the missing cells and perform an unbalanced test. For example, the t-test can deal with unequal sample sizes. However, these empty cells present a serious problem for exploratory multivariate analysis (to check reproducibility of the data, for example). We don't want the algorithm to treat the empty cells as missing values since they are again not missing at random but all algorithms need that assumption to work. One simple way is to replace the empty cells by zeroes. This idea is in line with other high-dimensional data analysis for instruments such as spectroscopy or chromatography. When you select a peak from those data, a zero is substituted in place of the original value when the peak area (or height) is less than the baseline noise. In our context, the proportion of the empty cells can be as high as 50 ~ 60 % of the total number of spots. On the other hand, for spectroscopy and chromatography, those numbers are minimal and rarely affect the analysis significantly.

Until a better treatment for those empty cells can be developed, we decided to replace them with zeroes and to utilize PCA. For the univariate tests, the unbalanced version of the tests can be used. However some analysts believe zero values should be used for the empty cells even for the univariate case since it is the better representation of data. Those empty cells are indeed informative. In this way, the analysis provides more power for detecting subtly expressed proteins. Sometimes important proteins are not abundantly expressed but have a very low volume. Those spots can lead to the identification of proteins that are really important for the problem at hand.

For the lung cancer example, we performed PCA after replacing the empty cells by zeroes and resulted in the patterns given in the Figure 9. The plot is a simple scatter plot of the first two principal components: together they explain 84% of the total variation. Two distinct groups can be seen, and correspond to disease group (black label) and normal group (red label) respectively. Some samples are "mixed" up. They are A1, B36, B39, B40, B41. We may have to check the quality of the data for those samples or may apply different pre-processing steps such as whole data normalization (Bolstad et al., 2003) to make further sense of those unusual samples.

Final biomarker selection for the lung cancer data has been done using multiple t-tests but the result was not given here because of proprietary reason. Usual multiplicity issue can be tackled using a couple of different approaches. Family-wise error-rate (FEW) control can be done using resampling ideas to improve the quality of the error estimate. However FEW can be too conservative in selecting markers since they may miss true positives (differentially expressed proteins). More practical approach can be the control of so called false discovery rate (FDR).

Recent reference and software for FDR can be found, for example, from Storey & Tibshirani (2003).

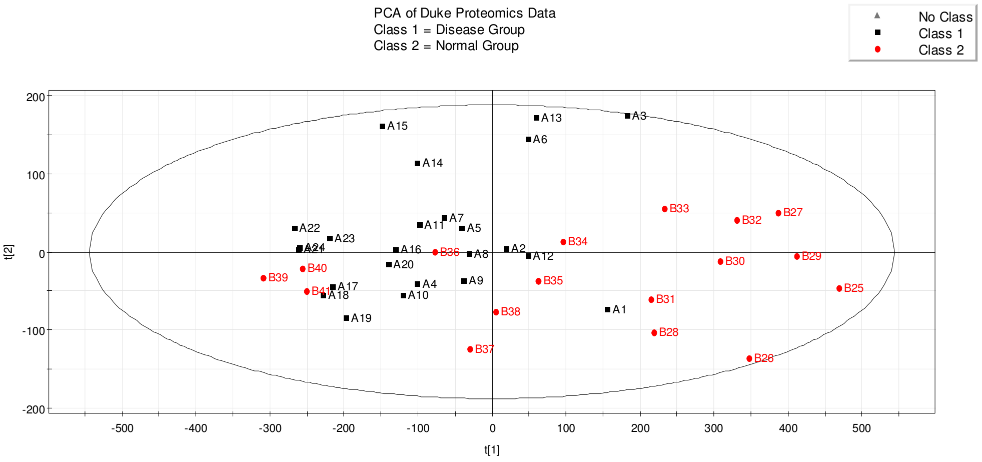

Figure 9. Scatter Plot for the First Two PCA Scores for the Lung Cancer Data (Class 1 = Disease group, Class 2 = Normal Group)

6. Conclusion

In this article we have attempted to highlight the immediate need for sophisticated statistical methods in analyzing two-dimensional gel images. There is also a compelling need for a comprehensive investigation of the quantitative features of two-dimensional gel images. There are many distributional assumptions that are being regularly made about these features. The success of any spot detection methods largely lies on the accuracy of the statistical assumption and thus they need to be verified. The variability in gel preparation is much higher than in other protein separation techniques such as mass spectrometry. However, due to the relatively cheaper cost of the 2-D PAGE technique, the reward can be great if this uncertainty can be statistically modeled and incorporated in the analysis to get optimal results.

References

Appel, R., Hochstrasser, D.F., Funk, M., Vargas, J.R., Pellegrini, C., Muller, A.F. and Scherrer, J.R. (1991) The MELANIE project: From a biopsy to automatic protein map interpretation by computer. *Electrophoresis*, **12**, 722-735.

Appel, R., Palagi, P.M., Walther, D., Vargas, J.R., Sanchez, J.C., Ravier, F., Pasquali, C. and Hochstrasser, D.F. (1997) MELANIE II -- A third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. *Electrophoresis*, **18**, 2724-2734.

Bettens, E., Scheunders, P., Sijbers, J., Van Dyck, D. and Moens, L. (1996) Automatic segmentation and modelling of two-dimensional electrophoresis-gels. In: Proceedings ICIP'96: vol. 2 IEEE International Conference on Image Processing, 665-668.

Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. In press, *Bioinformatics*.

Candes, E.J. and Donoho, D.L. (1999) Ridgelets: a key to higher-dimensional intermittency? *Phil. Trans. R. Soc. Lond. A.*, **357**, 2495-2509.

Coifman, R.R. and Donoho, D.L. (1995) Translation-invariant de-noising. In: A. Antoniadis and G. Oppenheim, ed., *Wavelet and Statistics*, Lecture Notes in Statistics, Springer-Verlag.

Conradsen, K. and Pedersen, J. (1992) Analysis of two-dimensional electrophoresis gels. *Biometrics*, **48**, 1273-1287.

Efrat, A., Hoffmann, F., Kriegel, K., Schultz, C. and Wenk, C. (2002) Geometric algorithms for the analysis of 2D-electrophoresis gels. *Journal of Computational Biology*, **9**, 299-315.

Freund, R.M. and Mizuno, S. (1996) Interior point methods: Current status and future directions. *Optima*, **51**, 1-9.

Glasbey, C.A. and Mardia, K.V. (1998) A review of image-warping methods. *Journal of Applied Statistics*, **25**, 155- 171.

Glasbey, C.A. and Mardia, K.V. (2001) A penalized likelihood approach to image warping. *Journal of the Royal Statistical Society, Series B*, **63**, 465-514.

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

Hoffmann, F., Kriegel, K. and Wenk, C. (1998) Matching 2D patterns of protein spots. Symposium on Computational Geometry, 231-239.

Hoffmann, F., Kriegel, K. and Wenk, C. (1999) An applied point pattern matching problem: Comparing 2D patterns of protein spots. *Discrete Applied Mathematics*, **93**, 75-88.

Horgan, G., Creasey, A. and Fenton, B. (1992) Superimposing two-dimensional gels to study genetic variation in malaria parasites. *Electrophoresis*, **13**, 871-875.

Horgan, G.W. & Glasbey, C.A. (1995) Uses of digital image analysis in electrophoresis. *Electrophoresis*, **16**, 298-305.

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

Pleissner, K.P., Hoffman, F., Kriegel, K., Wenk, C., Wegner, S., Sahlström, A., Oswald, H., Alt, H. and Fleck, E. (1999) New algorithmic approaches to protein spot detection and pattern matching in two dimensional electrophoresis gel databases. *Electrophoresis*, **124**, 281-302.

Pleissner, K.P., Oswald, H. and Wegner, S. (2001). *Proteomics*. BIOS Scientific Publishers, Oxford, chapter Image analysis of two-dimensional gels, pp. 131-149.

Potra, F.A. and Wright, S.J. (2000) Interior point methods. *Journal of Computational and Applied Mathematics*, **124**, 281-302.

Rowlands, D.G., Flook, A., Payne, P.I., van Hoff, A., Niblett, T. and McKee, S. (1988) GESA - a two-dimensional processing system using knowledge based techniques. *Electrophoresis*, **9**, 820-830.

Seillier-Moiseiwitsch, F., Trost, D.C. and Moiseiwitsch, J. (2002) Statistical methods for proteomics. *Methods in Molecular Biology*, **184**, Humana Press Inc. NJ.

Skolnick, M.M. (1982) An approach to completely automatic comparison of two-dimensional electrophoresis gels. *Clinical Chemistry*, **28**, 979-986.

Smilansky, Z. (2001) Automatic registration for images of two-dimensional protein gel. *Electrophoresis*, **22**, 1616-1622.

Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genome-wide experiments. Available at http://www.stat.berkeley.edu/~storey/.

Vincent, L. and Soille, P. (1991) Watersheds in digital spaces: An efficient algorithm based on immersion solutions. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, **13**, 583-598.

Voss, T. and Haberl, P. (2000). Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: Consequences for comprehensive data analysis. *Electrophoresis*, **21**, 3345-3350.