Lights, Camera, Action! - Systematic Variation in Difference Gel Electrophoresis

Kimberly F. Sellers
Carnegie Mellon University

Jeffrey Miecznikowski
Carnegie Mellon University

Surya Viswanathan
Carnegie Mellon University

Jonathan S. Minden
Carnegie Mellon University, minden@cmu.edu

William F. Eddy
Carnegie Mellon University, bill@stat.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/statistics

Part of the Statistics and Probability Commons

This Technical Report is brought to you for free and open access by the Dietrich College of Humanities and Social Sciences at Research Showcase @ CMU. It has been accepted for inclusion in Department of Statistics by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
Lights, Camera, Action! –

Systematic Variation in Difference Gel Electrophoresis

Kimberly F. Sellers*, Jeffrey Miecznikowski*, Surya Viswanathan†
Jonathan S. Minden†, William F. Eddy*

March 15, 2004

Abstract

Two-dimensional Difference Gel Electrophoresis (DIGE) circumvents many of the problems associated with gel comparison via the traditional approach, two-dimensional gel electrophoresis. DIGE’s accuracy and precision, however, is compromised by the existence of other significant sources of systematic variation, including that caused by the apparatus used for imaging proteins (location of the camera and lighting units, background material, imperfections within that material, etc.). Through a series of experiments, we estimate some of these factors, and account for their effect on the DIGE experimental data, thus providing improved estimates of the true protein intensities. The model presented here includes classical, single-dye gel electrophoresis images.

*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213
†Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213


1 Introduction

Proteome analysis has two basic components: separation of proteins into individual species and identification of the individual protein species. Identification is generally accomplished by mass spectrometry; two-dimensional gel electrophoresis (2DE; sometimes known as polyacrylamide gel electrophoresis, or PAGE) is often used for separation. A complex protein mixture is separated by 2DE into its components first by isoelectric focusing (IEF), which separates proteins by pH, and then by sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis, which separates proteins by the length of the polypeptide chain (O’Farrell, 1975).

The sample gel is first separated by its isoelectric points (pI) using commercially available IEF methods and isoelectric polyacrylamide gel (IPG) strips (Amersham Pharmacia Biotech). After IEF, the IPG strip is then washed in equilibrium buffers to prepare for SDS-polyacrylamide gel electrophoresis. To separate the proteins by molecular weight, the IPG strip is placed on top of a slab gel and electrophoresed to separate the proteins according to their size. (O’Farrell, 1975).

The conventional 2DE method is attractive because of its resolving power, sensitivity, and the low equipment cost; however, heterogeneities in different gels, the electric fields, pH gradients, thermal fluctuations, etc. are all factors that make reproducible spot matching between gels an extremely difficult task (Garrels, 1989; Conradsen and Pedersen, 1992; Gustafsson et al., 2002). Further, the 2DE-mass spectrometry approach does not optimally detect medium to low abundance proteins; it only detects proteins expressed at relatively high levels (Gygi et al., 2000). Thus, 2DE has limited potential for large-scale proteome analysis (Gygi et al., 2000; Greenbaum et al., 2003).

Ünlü, Morgan, and Minden (1997) developed a novel modified 2DE technique called
Difference Gel Electrophoresis (DIGE), which circumvents many of the problems associated with comparison of separate gels. In DIGE, two different protein samples are pre-labeled with a pair of matched cyanine dyes which fluoresce at different wavelengths of light but have identical charge and almost identical mass, thus allowing two different samples to be run within one gel mixture in both dimensions. Because the two samples have been subjected to essentially the same environment during separation, the “between gel” variation is significantly reduced; however, the “within gel” variation is essentially unaffected (Tonge et al., 2001; Alban et al., 2003).

This paper focuses on the systematic variation which occurs due to the DIGE data imaging process and the system used to image the proteins; for example, see Figure 1. Its purpose is to aid in the automatic detection of differential protein expression and modification. On the surface, this problem may appear straightforward; however, its difficulty lies in developing a model that accurately represents the complexity of the experimental conditions. Only through extensive experimentation and deliberation have we successfully identified and modelled several nuisance effects influencing the overall experimental data. Section 1.1 describes the imaging apparatus and details the sequence of steps performed in creating a protein image. The system includes a cabinet and platform fixture that houses the protein gel, and a camera and lighting structure to image the gels. This structure helps to establish a statistical model that describes the effect caused by the system components; see Section 1.2. The general approach is widely applicable, although the specific model is only appropriate for this particular type of imaging system. We are currently developing similar models for somewhat different gel imaging systems.
Figure 1: DIGE experiment image excited at two different wavelengths. A substantial number of nuisance effects (of considerable size) contribute to the images. This paper identifies and removes the most prominent factors, to isolate the true relative protein intensity corresponding to each light source.

1.1 System Description

Figure 2 is a schematic diagram of the fluorescence gel imager. The system contains a cabinet and platform, and a scientific-grade, Charge-Coupled Device (CCD) camera focused on the platform and situated approximately five inches above the platform with a lamp approximately three inches away from the camera on each side. There are currently five machines of this specific type, while several other similar types of commercial fluorescent gel imagers exist. Each of the major components is described in more detail in the sections below.

Cabinet and Platform

The gel is placed in a shallow black tray (hereafter referred to as the “platform”) and covered with a gel-storage aqueous solution of methanol and acetic acid. Ideally, the platform should not be fluorescent. Unfortunately, all materials tested to date are fluorescent at various wavelengths of interest, with the exception of quartz or fused silica glass. The gel
is placed on a quartz plate that was coated on the underside with black cyanoacrylate.

The camera image area is approximately 35mm×35mm, while a gel is approximately 140mm×175mm. Thus, to capture an image of the whole gel, a 5×4 array of tiled images is taken; we will henceforth refer to one of the 35mm square camera (sub-)images as a “tile.” The platform moves so that the tiles are collected in boustrophedonic order (back-and-forth as the ox plows the field): the first row from left to right, the second row from right to left, etc; see Figure 3. Each tile is 256 pixels×256 pixels, so the entire image of 20 tiles is 1280 pixels×1024 pixels, and each pixel is approximately 135 microns squared.

Lights

The gels are illuminated by two 250W quartz-tungsten-halogen lamps, set on either side of the camera. The lamp housings are fitted with motorized filter wheels that hold several bandpass filters for fluorescence excitation. This light is directed onto the gel via liquid light guides, and it covers the imaging area. Despite our best efforts, the illumination is not perfectly uniform.
Camera

DIGE images are recorded with a scientific-grade Peltier-cooled, 16-bit Charge-Coupled Device (CCD) camera, which the manufacturer reports has a linear response over five orders of magnitude.

A CCD is a silicon chip composed of an array of electron potential wells. When a photon hits the silicon dioxide layer above the electron well, electrons are released and trapped in the well. The contents of each well are read-out at the end of the exposure by transferring the charge, first in parallel rows, and then serially within row to an analog-to-digital converter (ADC) and stored digitally. To prevent negative values, an electronic biasing voltage is applied to each potential well (which means an image with an exposure time of zero will have positive values). Also, our CCD has a relatively long read out time so charge (known as “dark current”) will continue to accumulate during the readout.

The CCD imaging array consists of a rectangular array of pixels that are aligned vertically. These pixels are readout in the following manner: charge packets are shifted one pixel toward the serial register, causing the bottom row to transfer into the serial register. The charge packets are then transferred along the serial register toward the output amplifier.

Figure 3: Imaging order for tiles. The camera and lights remain fixed while the platform moves in the order shown.
where they are detected. The resulting data is a pixel-by-pixel, row-by-row representation of the image collected by the CCD.

**Action**

Each protein sample is labelled/dyed (i.e. a covalent link is established between the dye and the protein) with one of two cyanine dyes (Cy3 and Cy5). Each dye is excited by a unique wavelength of light selected by a filter in a moving wheel. There is a dichroic multi-bandpass, two-colored filter affixed to the front of the camera lens. The Cy3 dye is excited at 545 ± 10 nm, and imaged at 587.5 ± 17.5 nm. In contrast, the Cy5 dye is excited at 635 ± 15 nm, and imaged at 695 ± 30 nm. Since no bandpass filter is perfect, there will always be light leakage or “bleed through” from one cyanine channel to another. All of these factors should be taken into account.

The process by which DIGE images are created is as follows. Beginning with Tile 1, the camera images the DIGE sample using the Cy3 filter, then changes filters and images the same sample under the Cy5 light. After both images have been recorded, the platform moves in position to image Tile 2, first under the Cy3 light and then under the Cy5 light. A typical exposure time for DIGE images is 30 seconds for each tile under each lighting filter. Meanwhile, the transition time to move from one tile to the next is about five seconds. Thus, the total time to create the DIGE images corresponding to both cyanine dyes is approximately 25 minutes.

We have two images coming from one DIGE gel experiment, namely one image under the Cy3 light filter, and one image under the Cy5 light filter. Due to a significant amount of systematic and random noise introduced through the experimental procedure, DIGE images cannot be replicated exactly.
1.2 Overview: The model

Let \( l = \{3, 5\} \) denote the light filter setting to Cy3 or Cy5, respectively, and \( \bar{l} \) denote the complementary light. The following model is believed to describe the systematic effects caused by various components in the experimental and imaging process, including the camera, exposure time, lighting, and bleed-through:

\[
Y_l = \beta + \tau [\delta_l + \rho_l (1 + \alpha_i \cdot \gamma_l \cdot \pi_i + \kappa_l \cdot \gamma_l \cdot \pi_i)] + \epsilon_l,
\]

where \( Y_l \) is the count read from the CCD for a two-dimensional difference gel electrophoresis experiment under light \( l \), \( \beta \) represents the camera bias, \( \tau \) denotes the exposure time associated with the image, \( \delta_l \) is the effect of the lighting system on the platform under light \( l \), \( \rho_l \) measures the relative light effect that is introduced by gel excitation under light \( l \), \( \alpha_i \cdot \gamma_l \cdot \pi_i \) and \( \alpha_l \cdot \gamma_l \cdot \pi_l \) are the additional lighting effects when a uniformly labelled gel is excited with its corresponding light, \( \kappa \) denotes the effect of exciting a labelled gel with the complementary light \( \bar{l} \), and \( \gamma_l \) and \( \gamma_l \) are scaling constants, \( \pi_i \) and \( \pi_l \) represent the true relative intensity (on a 0-1 scale) of the proteins, respectively, and \( \epsilon \) is random error. We interpret \( \gamma_l \) and \( \gamma_l \) as the amount of protein loading corresponding to each dye. This paper focuses on the two dye, two light framework—\( \alpha_i \cdot \gamma_l \cdot \pi_i \) is the effect due to the protein when the light matches the dye, and \( \alpha_l \cdot \gamma_l \cdot \pi_l \) refers to the light intensity caused by the complementary light-dye relationship. A setup with more than two dyes and lights requires an even more complicated model.

The parameters, \( \beta, \delta, \rho, \gamma_l, \pi_l, \) and \( \epsilon \) may vary spatially. We let \( i \) and \( j \) denote the pixel row and column location on the CCD, respectively, within tile \( k \); \( i = 1, \ldots, 256; j = 1, \ldots, 256; k = 1, \ldots, 20 \). Thus, Equation (1) becomes

\[
Y_{l,ijk} = \beta_{ijk} + \tau [\delta_{l,ijk} + \rho_{l,ijk} (1 + \alpha_{l,ijk} \cdot \gamma_{l,ijk} \cdot \pi_{l,ijk} + \kappa_{l,ijk} \cdot \alpha_{l,ijk} \cdot \gamma_{l,ijk} \cdot \pi_{l,ijk})] + \epsilon_{l,ijk},
\]
Our ultimate goal is to estimate the within-sample relative intensities, $\pi_3$ and $\pi_5$, for the proteins when excited under Cy3 and Cy5, respectively, from a single gel. Thus, we consider two models of the form (2) where, given DIGE images $Y_{3,ijk}$ and $Y_{5,ijk}$, we want to estimate $\pi_{3,ijk}$ and $\pi_{5,ijk}$ and compare those resulting images. Any spots that appear in one image and not in the other imply that the associated protein is differentially expressed, whereas spots that significantly change in location signify differential modification. Figure 4 shows an example of DIGE images where differential expression and modification, respectively, are detected.

We did not arrive at the model given in Equation (2) by introspection and gedanken experiments. Rather, we have worked as an interactive team, collecting a substantial number of gels under a variety of experimental conditions. We are interested in $\pi_3$ and $\pi_5$; all other terms in the model are nuisance parameters of considerable size, contributing to the complexity of the model. The model has been developed iteratively; examination of residuals at each stage has suggested further experimentation and the form of additional terms in the model. In contrast, most biologists simply take a picture of the platform for a time equal to that for the DIGE experiment and subtract this result from the experimental data to determine the “preprocessed data”. Our work has accounted for large amounts of variation above and beyond that which is currently practiced, and related it to specific sources that are not attributable via the standard preprocessing technique. The model was not developed for the specific datasets. We have collected and studied hundreds of images and have arrived at this model as a complex compromise. The specific values apply only to this particular dataset but the general model is applicable to all of the images we have examined over the last couple of years.

Note that this model can also be applied to standard, single-dye, two-dimensional gel
electrophoresis (2DE) images. 2DE images are created by exciting a labelled gel with its corresponding light source, e.g. a Cy3 (Cy5) gel excited with a Cy3 (Cy5) light. Thus, under a particular light source $l$, these images can be modeled via Equation (1) with $\pi_l = 0$.

In order to estimate $\pi_{i,ijk}$ and $\pi_{l_{i,ijk}}$, we must first estimate $\beta_{i,ijk}$, $\delta_{i,ijk}$, $\rho_{i,ijk}$, $\alpha_{l_{i,ijk}}$, $\kappa_{l_{i,ijk}}$, and $\alpha_{l_{i,ijk}}$. A variety of experiments have been performed to achieve this task. Table 1 outlines the conditions of the experiments used to estimate each of the parameters. Section 2 describes the experiment performed to estimate the camera bias, and the model used to estimate $\hat{\beta}_{i,ijk}$. Section 3 explains the model for the lighting system and its effect when imaging the empty platform, and the data used to estimate this effect. Section 4 provides an estimate of the additional effect due to the existence of a proteinless gel on the platform. Section 5 discusses single-dyed uniform protein gel experiments performed to estimate $\alpha$, the relative effect of light excitation due to the gel on the platform.

Section 6 gives the final model, and considers doubly-labelled uniform protein gel experiments to check the accuracy of the model. Section 7 uses that model in an example to estimate the true relative expression of the proteins in the two samples, $\pi_3$ and $\pi_5$, and

Figure 4: Example of DIGE images under Cy3 and Cy5 exposures. The location circled in red indicates a differentially expressed protein spot, and the location circled in green indicates modified proteins.
Table 1: Experiments performed to estimate nuisance parameters; they are performed in the order listed in the table. Each row in the table corresponds to an experiment or set of experiments. The parameter in the column labelled “Condition” is set to the value supplied in the table. The parameter in the last column is estimated from the experimental results. All other parameters are set to their corresponding estimates obtained from all previously performed experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Condition</th>
<th>Estimated parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Camera bias</td>
<td>$\tau = 0$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>(B) Dark field, light $l$</td>
<td>$\rho_i = 0$</td>
<td>$\delta_i$</td>
</tr>
<tr>
<td>(C) No dye, light $l$</td>
<td>$\pi_3 = \pi_5 = 0$</td>
<td>$\rho_i$</td>
</tr>
<tr>
<td>(D) Bright, light $l$</td>
<td>$\pi_i = 1$, $\pi_i = 0$</td>
<td>$\alpha_i$, $\gamma_i$</td>
</tr>
<tr>
<td>(E) Bleed through, light $l$</td>
<td>$\pi_i = 0$, $\pi_i = 1$</td>
<td>$\kappa_i$, $\gamma_i$</td>
</tr>
<tr>
<td>(F) Double dye</td>
<td>$\pi_i = 1$, $\pi_i = 1$</td>
<td>$\gamma_i$, $\gamma_i$</td>
</tr>
</tbody>
</table>

their associated scaling constants, $\gamma_3$ and $\gamma_5$. Section 8 describes the effect on the protein estimates caused by removing the systematic variation.

We can think of this procedure as a stepwise experimental process. The estimates from previous experiments are applied to subsequent models to estimate all of the nuisance parameters describing the systematic variation in our data. We can use this approach because any DIGE image under consideration comes with associated images used to model the nuisance parameters in question.

At each stage, we first Winsorize the data to account for outliers (Tukey, 1962). Outliers in the raw image are due to any number of factors, including dust falling on the apparatus. Winsorization is a technique where extreme values are replaced with bounded values that are more representative of the data. Here, raw images were systematically Winsorized to the first and 99th percentile values. As a result, the parameter estimates are more robust.
Throughout the paper, we present many images of data, of estimates, and of residuals. While we think these are informative, we caution the reader that the appearance of any image is strongly influenced by the choice of grayscale: the mapping from numbers to gray level. We have been misled sufficiently often that our team mantra is “Don’t be deceived by the grayscale.”

2 Camera Bias

In order to model the camera bias, we assume no exposure time, i.e. $\tau = 0$ (Experiment A in Table 1). Further, our data consists of one $1280 \times 1024$ image of 20 tiles. Thus, Equation (2) reduces to

$$Y_{ijk} = \beta_{ijk} + \epsilon_{ijk}$$

$$= (m_{i} + a_{i} + b_{j} + c_{ij} + d_{k}) + \epsilon_{ijk},$$

where the subscript $k$ distinguishes the tiles ($k = 1, \ldots, 20$), and the subscripts $i$ and $j$ index the individual pixels in a tile with respect to row ($i = 1, \ldots, 256$) and column ($j = 1, \ldots, 256$). The image data is denoted by $Y$, $m$ is a typical pixel value, $a_{i}$ and $b_{j}$ represent the row and column effects respectively within the CCD, $c_{ij}$ is the additional CCD (i.e. within-tile) effect, $d_{k}$ is the tile effect, and $\epsilon$ is the unexplained random error.

We Winsorize the data to eliminate outliers, and fit the model to the Winsorized image using least squares. Our estimated value of $m_{i}$ is 507.9. We estimated the 256 values of $a_{i}$ and recognized a linear trend, comparable to that expected given the CCD readout process described in Section 1.1. On investigation, we found that the pixel readout rate from the analog-to-digital converter (ADC) is 40 KHz, implying that it takes approximately 6.5 seconds to read out the entire image because it is binned from $512 \times 512$ to $256 \times 256$ on
<table>
<thead>
<tr>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outliers</td>
<td>10</td>
<td>12386297.0</td>
</tr>
<tr>
<td>Row gradient</td>
<td>1</td>
<td>2163479.0</td>
</tr>
<tr>
<td>Column direction</td>
<td>255</td>
<td>84410.0</td>
</tr>
<tr>
<td>Error</td>
<td>1310453</td>
<td>10506868.0</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1310719</td>
<td>25141054.0</td>
</tr>
</tbody>
</table>

Table 2: Analysis of Variance table for Camera Bias

During that time, additional dark current accumulates on the CCD with the later rows having larger accumulations than the earlier rows. We modeled this by assuming that $a_i = a \cdot i$. Our estimate for the row effect is $\hat{a} \approx 0.017$, implying an estimated dark current accumulation of 0.68 electrons/row/second; see Figures 5 and 6. This parameter describes the largest amount of systematic variation, with MS=2163479; see Table 2. The camera documentation indicates a value of 1.0 electrons/row/second.

Given the readout process, we hypothesized that $b_j$ could similarly be represented as $b_j = b \cdot j$ with $b$ approximately equal to $a/256$. Upon inspection, however, we recognized a deviation from that trend on both sides of the CCD, most striking on the left-hand side of the CCD; see Figures 5 and 6. Thus, we retain 256 individual estimates, $\hat{b}_j, j = 1, \ldots, 256$ to model the transfer toward the output amplifier. After removing the row and column effects, we noted that the additional CCD effect and tiling effect were not significant and, therefore, were deleted from the model. The resulting model error image has a random white-noise pattern (Figure 6) with MSE=8.02; see Table 2.
Figure 5: Panel (a) is a plot of the 5120 within-tile row means versus row number. The estimated slope is 0.017, corresponding to the dark current of 0.68 electrons/row/second. Panel (b) is a plot of the 5120 within-tile column means versus column number. The most striking feature is the reduced intensity within the three columns on the left of the CCD.

Thus we conclude that an adequate fit for the bias is

\[
\hat{\beta}_{ij} = \hat{m}_a + \hat{a} \cdot i + \hat{b}_j
\]

\[
= 507.9 + 0.017i + \hat{b}_j,
\]

where the estimates and residuals \(e_{ijk}\) are shown in Figure 6. We will assume \(\beta_{ij} = \hat{\beta}_{ij}\) henceforth.

### 3 Dark Field

A “dark field” image is collected with the light on and the filter set to either Cy3 or Cy5, but no gel (and therefore no protein) present in the device. The platform is then imaged with an exposure time of approximately 30s for each tile. The purpose of such an experiment is to estimate \(\delta\) (Experiment B in Table 1).
Figure 6: Bias Image and Its Components – each of the images is separately scaled. The $Y_{ijk}$ are the Winsorized values for an image acquired with 0.1 second exposure time. The $\hat{a}_i$ show the linear accumulation of dark current during the readout period. The $\hat{b}_j$ show the decreased intensity values at the edges of the CCD. The $\epsilon_{ijk}$ appear to be white-noise, indicating that the model is a good fit.
Because no dyed gel is considered in this experiment, we consider Equation (2) with \( \rho = 0 \), i.e.

\[
Y_{i,j,k} = \beta_{i,j,k} + 30 \cdot \delta_{i,j,k} + \epsilon_{i,j,k},
\]

where

\[
\delta_{i,j,k} = m_{i,\delta} + f_{i,j} + g_{i,k} + h_{i,j,k},
\]

\( Y \) is the dark field image, \( \beta \) is the system bias estimated by Equation (4), \( \delta \) is the dark field, \( \epsilon \) is the random error, and \( l = \{3, 5\} \) denotes the light filter setting (Cy3 or Cy5). The dark effect contains a typical pixel value associated with the dark field under light \( l \) (\( m_{i,\delta} \)), the effect of the light \( f_{i,j} \) at pixel \((i,j)\), the effect of a particular tile \( g_{i,k} \), and the effect on intensity due to nonuniformities on the platform (e.g., scratches), \( h_{i,j,k} \). Estimating \( h_{i,j,k} \) requires multiple dark field images. For this experiment, we collected two such images under each of the Cy3 and Cy5 lights.

Modeling the dark field image corresponding to Cy3 and Cy5 occurs in a stepwise fashion using Winsorized data. The estimated value for \( m_{i,\delta} \) is the average of the resulting \( 1280 \cdot 1024 \cdot 2 = 2621440 \) pixels from the two dark field images under light \( l \). Then, after subtracting \( \hat{m}_{i,\delta} \), we estimate the effect of the light by averaging the 40 pixels associated with location \((i,j)\) across tiles for both replicates to get \( \hat{f}_{i,j} \). After we subtract this within-tile effect, we average over the \( 256 \cdot 256 \cdot 2 = 131072 \) pixels within tile \( k \) from both replicates to get \( \hat{g}_{i,k}, k = 1, \ldots, 20 \). Finally, we average over the resulting images for each replicate to estimate the platform effect, \( \hat{h}_{i,j,k} \). The tile and platform effects are not significant. Such an estimate is applicable only for our device.

We determine the estimated dark field image

\[
\hat{\delta}_{i,j,k} = \hat{m}_{i,\delta} + \hat{f}_{i,j},
\]
associated with the Cy3 and Cy5 light, respectively, and use this information to understand
the extra variation that arises when a gel is introduced to the scenario.

3.1 Cy3 Dark

Our estimate for \( m_{3,5} \) is approximately 43.419 counts/second, while the estimated lighting
effect is shown in Figure 7.

The Cy3 lighting effect can be roughly described as a convex, three-dimensional paraboloid
such that the center of each tile contains the lowest intensities, and the values increase with
the Euclidean distance from the center. We do not have a physical explanation for this
pattern.

Regarding the tile effect, we see that the top and bottom rows of tiles are brighter than
the others, with the first imaged tile within each of these rows significantly brighter than
the remaining tiles in that row. This is believed to be because of the sequence in which
the images are taken. The camera images the first tile of the platform using the Cy3 filter
for 30s, then changes to a Cy5 filter to take the same image, before the platform moves
to image the next tile location. This alternating sequence continues until the 20 tiles are
imaged both under Cy3 and Cy5 to create two overall images of the platform using each
filter. To test the significance of the ordering sequence, we also ran images where Cy5 was
imaged before Cy3 at every tile. The difference between the two resulting Cy3 images did
not appear to be significant; likewise, the Cy5 images were not significantly different. We
therefore suspect that the light somehow reflects off of the upper edge of the platform to
cause the brighter intensities in the top row of the Cy3 image.

The platform contains dust and scratches on the glass surface, as detected by the asso-
ciated platform estimate. The imaging process does not occur in a dust-free environment,
therefore dust falls onto the gel and is included in the image. The scratches on the glass are believed to be due to the cleaning process administered by the operators. They use detergent and paper towels to scrub and cleanse the platform surface between uses. Although this factor is easily explained, its effect is not significant, therefore it is not included in the model. The resulting residual images \( e_{3,ijk1} \) and \( e_{3,ijk2} \) in Figure 7 thus include the nonsignificant tiling effect and platform effect.

We conclude that an adequate fit for the Cy3 dark field is

\[
\hat{\delta}_{3,ijk} = 43.419 + \hat{f}_{3,ij},
\]

as shown in Figure 7, which will be subtracted from each subsequent image collected using a Cy3 lighting filter. The model accounts for a substantial amount of variation (SST = \( 8.697 \times 10^{12} \); SSOutliers = \( 8.356 \times 10^{8} \); SSE = \( 2.012 \times 10^{7} \)).

### 3.2 Cy5 Dark

Our estimate for \( m_{5,i} \) is approximately 22.749 counts/second, while the estimated lighting effect is provided in Figure 8.

The lighting effect appears approximately as a convex parabola in the \( y \)-direction, crossed with a constant function in the \( x \)-direction so that the intensities associated with the middle rows contain the lowest values, and they increase systematically as the distance to the top and bottom borders of the tile decreases.

There appears to be no systematic pattern corresponding to the imaging order of the individual tiles. Further, the hypothesis that light reflects from the upper edge of the platform to create a brighter top row is weakened; see Figure 8. We now see that the bottom row contains the brighter sequence of tiles; however, this area of the platform is
Figure 7: Cy3 Dark Images and Their Components – each of the images is separately scaled. The $Y_{3,ijk}$ images are the Winsorized values for images acquired with 30 second exposure time. The $\hat{f}_{3,ij}$ shows the circular lighting effect when exciting the platform. The $e_{ijk}$ images show a nonsignificant tiling effect, and a scratch or streak in the platform.
not in close proximity with the lower edge of the platform.

Again from the platform estimate image, we detect scratches and dust on the glass surface. This is believed to be caused by the environment surrounding the apparatus, and the cleaning process administered by the operators. Both the tiling and platform effects are nonsignificant and, thus, are included in the residuals. The resulting residual images, $e_{5,ijk1}$ and $e_{5,ijk2}$ shown in Figure 8.

We conclude that an adequate fit for the Cy5 dark field is

$$\delta_{5,ijk} = 22.749 + \hat{f}_{5,ij},$$

(9)
as shown in Figure 8, which is subtracted from each subsequent image collected using a Cy5 lighting filter. The model accounts for a substantial amount of variation ($\text{SST} = 3.731 \times 10^{12}; \text{SSOutliers} = 3.882 \times 10^8; \text{SSE} = 2.03 \times 10^6$).

4 Gel Images with No Protein

In this experiment, a gel is prepared containing no proteins and imaged with the lights on and filter set to either the Cy3 or Cy5 filter for 30 seconds (Experiment C in Table 1). Accordingly, we consider Equation (2) with $\pi_i = \pi_f = 0$, i.e.

$$Y_{l,ijk} = \beta_{ijk} + 30(\delta_{l,ijk} + \rho_{l,ijk}) + \epsilon_{l,ijk},$$

(10)

where $Y$ is the raw data produced from the experiment, $\beta$ is the camera bias estimate determined from Equation (4), $\delta$ is the estimate determined from Equation (7) of the lighting system on the platform, $\rho$ is the relative light introduced by gel excitation, and $\epsilon$ is the random error.

Upon inspection of the raw images associated with this experiment, we note that their intensities are not much greater than that for the respective dark field images, and each
Figure 8: Cy5 Dark Image and Its Components – each of the images is separately scaled. The $Y_{5,ijk}$ images are the Winsorized values for images acquired with 30 second exposure time. The $\hat{f}_{5,ij}$ shows the horizontal lighting effect when exciting the platform. The $e_{5,ijk}$ images also show the nonsignificant tiling effect, along with some evidence of the scratch or streak in the platform.
of the corresponding images has the same appearance. The difference between the Cy3
gel and Cy3 dark images, however, is greater than that for the corresponding Cy5 images.
After correcting for the bias and dark field effects, we find that the resulting gel image
associated with each dye appears approximately constant, therefore we consider \( \rho_{t,ijk} = \rho_t \).
In particular, \( \hat{\rho}_3 = 2.756 \) counts/second and \( \hat{\rho}_5 = 0.877 \) counts/second. See Figure 9 for
the experimental data under a Cy3 and Cy5 filter, respectively, and the corresponding
residual image. The residual images each have relatively larger intensities in the lower-left
tile. The total sum of squares associated with the Cy3 data equals \( 5.108 \times 10^{12} \) (SSOutliers
= \( 6.76 \times 10^9 \)), while the resulting residual sum of squares is \( 5.607 \times 10^7 \), so the model
accounts for a large amount of variation in the data. Similarly, the Cy5 data have a total
sum of squares of \( 1.995 \times 10^{12} \) (SSOutliers = \( 1.818 \times 10^9 \)), while SSE = \( 2.943 \times 10^6 \).

## 5 Single-dye Gel Images

Single gel images are collected with the light on, a special gel uniformly loaded with protein
labelled with a single dye, and imaged under either a Cy3 or Cy5 lighting filter. The
proteins on this gel are not separated via the electrophoresis process. “Bright” images refer
to single gel images that are excited under the associated cyanine light filter (for example,
a Cy3-labelled gel excited under a Cy3 light filter), while “bleed through” refers to images
where a single-dyed gel is excited under a complementary light source (for example, a Cy3-
labelled gel excited under a Cy5 light filter). For each light, the exposure time for these
images is dependent on the time needed to obtain near saturation in the “bright” image.

We model these images as

\[
Y_{t,ijk} = \beta_{ijk} + \tau [\delta_{l,ijk} + \rho_{l,ijk} (1 + \alpha_{l,ijk} \cdot \gamma_l \cdot \pi_{l,ijk} + \kappa_{l,ijk} \cdot \alpha_{l,ijk} \cdot \gamma_l \cdot \pi_{l,ijk})] + \epsilon_{t,ijk}, \tag{11}
\]
Figure 9: Gel Images and Respective Residual Images – each of the images is separately scaled. No protein is present in the gels. Each of the $Y_{ijk}$ images is the Winsorized values of the image acquired with a 30 second exposure time. Each of the $e_{ijk}$ shows a nonsignificant tiling effect.
where bright images are such that \( \pi_i \equiv 1 \) and \( \pi_i \equiv 0 \) (Experiment D in Table 1), and bleed through images set \( \pi_i \equiv 0 \) and \( \pi_i \equiv 1 \) (Experiment E in Table 1). Sections 5.1.1 and 5.1.2 provide estimates of \( \alpha_i \) and \( \alpha_i \), respectively. Sections 5.2.1 and 5.2.2 estimate \( \kappa_{l,ijk} \), \( l = \{3, 5\} \). We will use these estimates to quantify the variation that exists in double-dyed gels, namely that described via \( \gamma_l \) and \( \gamma_l \).

### 5.1 Bright Coefficients

“Bright” images refer to single dye images that are excited under a matching light source (eg. Cy3 dye, Cy3 light). Such datasets serve to represent the ideal notion of no bleed through from the other dye, and the gel is assumed to have a uniform composition. Thus, we consider Equation (2) where \( \pi_{l,ijk} \equiv 1 \) and \( \pi_{l,ijk} \equiv 0 \), i.e.

\[
Y_{l,ijk} = \beta_{ijk} + \tau[\delta_{l,ijk} + \rho_{l,ijk}(1 + \alpha_{l,ijk} \cdot \gamma_l)] + \epsilon_{l,ijk},
\]  

(12)

where we want to estimate \( \alpha_i \cdot \gamma_l \) for \( l = \{3, 5\} \); \( \beta, \delta \) and \( \rho \) were estimated previously, and \( \pi_i \) is assumed to equal 1. In order to estimate these parameters, we Winsorize the raw data, and adjust to account for the estimates from the previous experiments. Using this corrected data, we set \( \gamma_i \) equal to the resulting (Winsorized) maximum pixel intensity, and let \( \alpha_i \) equal the normalized lighting effect, \( 0 \leq \alpha_{l,ijk} \leq 1 \).

#### 5.1.1 Cy3 dye, Cy3 light

From the procedure described in Section 5.1, we find \( \gamma_3 = 417.6 \) and \( \alpha_{3,ijk} \) is shown in Figure 10. The effect due to tile was not significant, thus \( \alpha_{3,ijk} = \alpha_{3,ij} \). The resulting residual image, \( \epsilon_{3,ijk} \), shows that the gel is actually nonuniform; however, the gel was poured uniformly to the best of our ability (see Figure 10). The estimates, \( \hat{\alpha}_{3,ij} \) and \( \hat{\gamma}_3 \),
will be applied in Section 5.2 to estimate \( \kappa \). The model accounts for a substantial amount of variation in the data; \( \text{SST} = 6.272 \times 10^{18}, \text{SSOutliers} = 9.31 \times 10^{12}, \) and \( \text{SSE} = 8.96 \times 10^{9} \).

5.1.2 Cy5 dye, Cy5 light

Using the procedure described in Section 5.1, we get \( \gamma_5 = 1909.8 \) and \( \alpha_{5,ijk} \) provided in Figure 10. As with the analogous experiment with Cy3 dye and light, the tile effect was not significant thus \( \alpha_{5,ijk} = \alpha_{5,ij} \) and the residuals show a nonuniform distribution of the gel on the platform. Further, we can see that the right edge of the gel did not lay flat on the platform while being imaged, resulting in larger intensity readings. The model accounts for less variation with this data set than for the Cy3 data: \( \text{SST} = 7.526 \times 10^{18}, \text{SSOutliers} = 3.482 \times 10^{10}, \) and \( \text{SSE} = 3.477 \times 10^{11} \). The estimates, \( \hat{\alpha}_{5,ij} \) and \( \hat{\gamma}_5 \), will be applied in Section 5.2.

5.2 Bleed Coefficients

To estimate the bleed coefficients, we collect images made with the same single dye gel used in Section 5.1, and imaged with the complimentary light (e.g., Cy3 dye, Cy5 light). Thus, we can use the same estimates for \( \gamma \) that were obtained previously. The objective here is to model the added effect caused by having both dyes excited under a particular light. In particular, we want to estimate the additional intensity caused by exciting the dye with the complementary light source. We consider Equation (11) where \( \pi_{l,ijk} \equiv 0 \) and \( \pi_{l,ijk} \equiv 1 \), i.e.

\[
Y_{l,ijk} = \beta_{ijk} + \tau[\delta_{l,ijk} + \rho_{l,ijk}(1 + \kappa_{l,ijk} \cdot \alpha_{l,ijk} \cdot \gamma_l)] + \epsilon_{l,ijk}.
\]  

The estimates, \( \beta_{ijk}, \delta_{l,ijk}, \rho_{l,ijk}, \alpha_{l,ijk}, \) and \( \gamma_l \) are provided via the previous experiments, and we want to estimate \( \kappa_{l,ijk} \).
Figure 10: Bright Images and Their Respective Components – each of the images is separately scaled. Each of the $Y_{ijk}$ are the Winsorized values of the images acquired from an exposure time to reach saturation. The exposure time for the $Y_{3,ijk}$ image is 20 seconds, and the exposure time for the $Y_{5,ijk}$ is 15 seconds. Each of the $\alpha_{ij}$ images shows the lighting effect due to the presence of proteins in the gel. The corresponding $e_{3,ijk}$ images demonstrate the nonuniformity of the gel loading process.
We Winsorize the resulting raw data to the first and 99th percentiles to account for outliers, and adjust the data by the bias, dark, lighting and gel-induced lighting estimates. The resulting image is an estimate of $\kappa_{i,j,k}$, which is approximately constant with respect to $i, j, k$; thus, we average over the inner $3 \times 2$ matrix of tiles to obtain $\kappa_{i,j,k}$.

5.2.1 Cy3 dye, Cy5 light

From Equation (13), we estimate $\kappa_{3,i,j,k} \approx 0.019$. We see that the added intensity caused by a bleed-through of Cy5 light on a Cy3-dyed gel is minimal. In particular, the increase of intensity for this image is not much larger than that for the Cy3 dark filter, shown in Figure 7. The resulting residual image appears as white noise within the area of the gel, with the exception of values along the top edge of the gel where the gel originally reached saturation; see Figure 11. A substantial amount of variation is explained through the model: $\text{SST} = 2.353 \times 10^{16}$, and $\text{SSE} = 1.907 \times 10^4$. The bleed-through estimate will be applied in Section 7 to estimate the intensities of protein images.

5.2.2 Cy5 dye, Cy3 light

Using Equation (13), we estimate $\kappa_{5,i,j,k} = 0.045$. Thus, we see that the added intensity caused by a bleed-through of Cy3 light on a Cy5-dyed gel is significantly greater than that for Cy5 bleed-through on a Cy3-dyed gel. Meanwhile, we again see approximate white-noise in the residuals, with the exception of the right edge where the gel curled during the imaging process; see Figure 11. Again, our bleed-through model accounts for a substantial amount of variation: $\text{SST} = 2.633 \times 10^{17}$, and $\text{SSE} = 268.415$. This bleed-through estimate will be applied in Section 7 to determine the true intensities of protein images.
Figure 11: Cy3 dye, Cy5 light Bleed Image and Its Residuals – each of the images is separately scaled. The top row displays the results of the Cy3 dye, Cy5 light experiment, and the bottom row shows the images from the Cy5 dye, Cy3 light experiment. The $Y_{5,ijk}$ are the Winsorized values of the image acquired by exposing the gel for 15 seconds; the $Y_{3,ijk}$ are analogous results from a 20 second exposure. Each of the $e_{ijk}$ show a nonsignificant tiling effect. The $Y_{3,ijk}$ and $e_{3,ijk}$ both show that the gel curled upward during the imaging process.
6 Double-Dye Gel Images

In this experiment, a gel is uniformly dyed with both Cy3 and Cy5, and placed on the platform for imaging. The gel is then excited for a time to reach near saturation under each respective light source. To model this experiment, we consider Equation (2) with $\pi_{i,ijk} = 1$ and $\tau_{i,ijk} = 1$ (Experiment F in Table 1), i.e.

\[ Y_{i,ijk} = \beta_{ijk} + \tau[\delta_{i,ijk} + \rho_{i,ijk}(1 + \alpha_{i,ijk} \cdot \gamma_i + \kappa_{i,ijk} \cdot \alpha_{i,ijk} \cdot \gamma_i)] + \epsilon_{i,ijk}, \]  

where estimates for $\beta$, $\delta$, $\rho$, $\alpha$, and $\kappa$ have been obtained from previous sections. Therefore, this experiment serves to check the accuracy of the previous estimates and to estimate $\gamma_i$ and $\gamma_i$ for this experiment.

In determining $\gamma_3$ and $\gamma_5$, we find that there is no significant variation within the platform, thus we let $\gamma_3$ and $\gamma_5$ equal the average over the six inner tiles within the platform under their respective lights to ensure that the estimate is determined from gel values; see Figure 12. As a result, the estimates associated with this experimental data are $\gamma_3 = 1310.31$ and $\gamma_5 = 4622.07$. The corresponding residuals, $e_{i,ijk}$ in Figure 12, illustrate that the gel is actually not uniformly loaded. The amount of variation within the gel, however, is relatively small and will be ignored.

7 Overall Model

The basic model for an image is

\[ Y_{i,ijk} = \beta_{ij} + \tau[\delta_{i,ijk} + \rho_{ij}(1 + \alpha_{i,ij} \cdot \gamma_i \cdot \pi_{i,ijk} + \kappa_{i,j} \cdot \alpha_{i,ij} \cdot \gamma_i \cdot \pi_{i,ijk})] + \epsilon_{i,ijk}, \]  

where the leading subscript $l = \{3, 5\}$ denotes the color of the filtered light (Cy3 or Cy5). The trailing subscript, $k$, indexes the tiles which compose the entire image, while $i$ and $j$
Figure 12: Double-Dye Gel Images and Their Respective Residuals – each of the images is separately scaled. The top row contains the image excited with a Cy3 light, and the bottom row shows the image excited with a Cy5 light. The $Y_{3,ijk}$ are the Winsorized values of the image acquired with a Cy3 light for 10 seconds; $Y_{5,ijk}$ are analogous results with a 10-second, Cy5 light excitation. Each of the $e_{3,ijk}$ shows a nonsignificant, nonuniform loading of the gel.
index the row and column location of the individual pixels within a tile. $Y$ is the image data associated with a gel from any DIGE experiment, $\beta$ represents the camera bias, $\tau$ is the exposure/excitation time, $\delta$ is the dark field, $\rho$ is the effect of exciting a gel with no proteins, $\alpha$ is the lighting effect due to the presence of proteins, and $\kappa$ is the bleed coefficient. All of these parameters were estimated through the previous experiments, while the true protein loadings and relative intensities ($\gamma_l \cdot \pi_{l,ijk}$ and $\gamma_l \cdot \pi_{l,ijk}$) are yet to be estimated.

To construct these estimates, we first Winsorize the protein images to their first and 99th percentiles, and let $Y_{3,ijk}$ and $Y_{5,ijk}$ denote the DIGE experiment images under Cy3 and Cy5 light respectively. From Equation (15), we solve for $\gamma_l \cdot \pi_l$, using the estimates obtained in Experiments A-F in Table 1.

After constructing the images, $\gamma_l \cdot \pi_{l,ijk}$ and $\gamma_l \cdot \pi_{l,ijk}$, we crop each of the images to the inner $768 \times 1024$ pixels so that we avoid edge effects. We feel that this step is appropriate because interesting protein spots do not generally lie in this outer area, and there are large “artifacts” in the image outside of this area. From these reduced images, we set $\gamma_l$ and $\gamma_l$ equal to the associated maximum value over each respective image, and $\pi_l$ and $\pi_{l,ijk}$ are the corresponding relative protein images. Figure 13 shows the result of removing the associated systematic variation on a DIGE experiment image exposed under Cy3 and Cy5, respectively, with $\gamma_3 = 124.654$ and $\gamma_5 = 46.845$.

Much of the systematic variation in the DIGE experimental data is reduced or removed in the estimated true protein images. For these images, however, additional factors contributing to further variation in the data (including edge effects and other phenomena that took place during the DIGE imaging process) appear in the images for $\pi_3$ and $\pi_5$. This is because the DIGE experimental images are created without replication. The data, $Y_l$
Figure 13: Estimated true relative protein images: (a) Cy3 protein image, $\hat{\pi}_3$; (b) Cy5 protein image, $\hat{\pi}_5$. Because there were no replicates for the DIGE experiment, random phenomena are also present in these images.

and $Y_i$, are used in the two equations of the form (15) to solve for $\pi_3$ and $\pi_5$. Thus, the estimated value of the residuals is approximately zero, and any random error in the DIGE experiment is contained in the estimates for $\pi_3$ and $\pi_5$. With replication, however, we can estimate $\pi_3$ and $\pi_5$ via least squares; this will lead to non-zero residuals. By “replication”, we mean that the biologists create, electrophorese and image more than one gel from a specific protein mixture. This will allow for a more accurate estimate of the relative proteins corresponding to each light, and for other random phenomena to be accounted for in the residual images.

For the Cy3 DIGE image here, $\text{SST} = 1.654 \times 10^{13}$, while $\text{SST} = 9.628 \times 10^{12}$ for the Cy5 DIGE image. To account for the variation in the data, we provide the sums of squares associated with each of the estimates in Equation (15) in Table 3. Given the structure of the model, as well as the procedures used to determine the estimates, we note that the sums of squares do not add to the total sums of squares. The purpose here is to indicate the relative amount of variation stemming from each of the factors in the model.
<table>
<thead>
<tr>
<th>Factor</th>
<th>Parameter</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cy3 Outliers</td>
<td></td>
<td>$9.315 \times 10^{12}$</td>
</tr>
<tr>
<td>Cy5 Outliers</td>
<td>$\beta_{ij}$</td>
<td>$7.017 \times 10^{12}$</td>
</tr>
<tr>
<td>Bias</td>
<td>$\delta_{3,ijk}$</td>
<td>$1.46 \times 10^{7}$</td>
</tr>
<tr>
<td></td>
<td>$\delta_{5,ijk}$</td>
<td>$2.508 \times 10^{9}$</td>
</tr>
<tr>
<td></td>
<td>$\delta_{5,ijk}$</td>
<td>$6.846 \times 10^{6}$</td>
</tr>
<tr>
<td>Dark field</td>
<td>$\rho_3$</td>
<td>$9.956 \times 10^{6}$</td>
</tr>
<tr>
<td></td>
<td>$\rho_5$</td>
<td>$1.008 \times 10^{6}$</td>
</tr>
<tr>
<td>Gel</td>
<td>$\alpha_{3,ij}$</td>
<td>$2.508 \times 10^{9}$</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{5,ij}$</td>
<td>$6.846 \times 10^{6}$</td>
</tr>
<tr>
<td>Bright</td>
<td>$\kappa_5$</td>
<td>$9.945 \times 10^{5}$</td>
</tr>
<tr>
<td></td>
<td>$\kappa_3$</td>
<td>$1.008 \times 10^{6}$</td>
</tr>
<tr>
<td>Bleed-through</td>
<td>$\kappa_5$</td>
<td>$476.902$</td>
</tr>
<tr>
<td></td>
<td>$\kappa_3$</td>
<td>$2608.563$</td>
</tr>
<tr>
<td>Loading factors</td>
<td>$\gamma_3$</td>
<td>$2.037 \times 10^{10}$</td>
</tr>
<tr>
<td></td>
<td>$\gamma_5$</td>
<td>$2.876 \times 10^{9}$</td>
</tr>
<tr>
<td>Proteins</td>
<td>$\pi_3$</td>
<td>$1.132 \times 10^{5}$</td>
</tr>
<tr>
<td></td>
<td>$\pi_5$</td>
<td>$8.763 \times 10^{5}$</td>
</tr>
</tbody>
</table>

Table 3: Sums of squares (over pixels) associated with each of the estimates in Equation (15).
8 Discussion

We have developed an automated procedure to account for all of the aforementioned factors contributing to the variation that exists in DIGE experimental data. Our program is supplied with data from the experiments to estimate the nuisance parameters, along with the experimental data of interest. We obtain image estimates for the true protein intensities under the Cy3 and Cy5 light filters, respectively, along with estimates of the variances from the original and final images. The automated code allows for routine use. To further improve on the current procedure, we continue to investigate other less important effects including the lamp age, lamp location, movement and drying of the gel, edge effects from the gel, etc.

The grayscale associated with the images influences our impression of the model, as well as our ability to detect protein spots in the final image estimates. Winsorization of the raw data accounts for the influence of outliers in the raw data, and “cropping” the gel images accounts for the influence due to edge effects; however, there remain other internal influences from the grayscale that might lead one to overfit the data. For example, excessive Winsorization of an image may suggest the inclusion of more parameters (e.g., to account for a tiling effect).

Removal of systematic variation in protein images is the first step to addressing broader issues relating to differential proteomics. Our future work will include locating and quantifying protein spots, spot alignment to allow for more accurate comparison of images, and addressing gel-to-gel variation.

It is currently difficult to compare the two relative protein images because of the large influence of the unusual phenomenon in Tile 18 of Figure 13a. Resolving the gel-to-gel variation problem will permit us to obtain replicated images and thus obtain better
estimates of the relative protein images. As a result, random events such as that shown in
Tile 18 of Figure 13a can be included in the estimate of the residuals, and an even more
precise estimate of $\pi_3$ can be obtained. This will allow for more informative comparisons
between the two relative protein images.
References


