Bioinformatics Advance Access published September 25, 2008

Genomics

Improving 2D-DIGE protein expression analysis by two-stage linear mixed models: Assessing experimental effects in a melanoma cell study.

Fernández Elmer A.[1,*], Girotti María R.[2*], López del Olmo Juan A[3], Llera Andrea S.[2*], Podhajcer Osvaldo L.[2*], Cantet Rodolfo J. C.[4*] and Balzarini Mónica[5,*]

[1]School of Engineering, Intelligent Data Analysis Group, Catholic University of Córdoba, Argentina, [2]Laboratory of Molecular and Cellular Therapy, Fundación Instituto Leloir, Argentina, [3]Unidad de Proteómica, Centro Nacional de Investigaciones Cardiovasculares, Spain, [4] Facultad de Agronomía, UBA (University of Buenos Aires), [5] Biometric Department, National University of Córdoba and [*]CONICET, Argentina

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: DIGE-based protein expression analysis allows assessing the relative expression of proteins in two biological samples differently labeled (Cy5, Cy3 CyDyes). In the same gel, a reference sample is also used (Cy2 CyDye) for spot matching during image analysis and volume normalization. The standard statistical techniques to identify differentially expressed (DE) proteins are the calculation of fold-changes and the comparison of treatment means by the *t*-test. The analyses rarely accounts for other experimental effects such as CyDye and gel effects, which could be important sources of noise while detecting treatment effects.

Results: We propose to identify DIGE DE proteins using a two-stage linear mixed model. The proposal consists of splitting the overall model for the measured intensity into two interconnected models. First, we fit a normalization model that accounts for the general experimental effects such as gel and CyDye effects as well as for the features of the associated random term distributions. Second, we fit a model that uses the residuals from the first step to account for differences between treatments in protein-by-protein basis. The modeling strategy was evaluated using data from a melanoma cell study. We found that a heteroskedastic model in the first stage, which also account for CyDye and gel effects, best normalized the data while allowing for an efficient estimation of the treatment effects. The Cy2 reference channel was used as a covariate in the normalization model to avoid skewness of the residual distribution. Its inclusion improved the detection of DE proteins in the second stage.

Supplementary information: R and SAS codes to analyze DIGE data with the proposed approach are available at http://www.uccor.edu.ar/modelo.php?param=3.8.5.15.2

Contact: elmer.fernandez@ucc.edu.ar

1 INTRODUCTION

Nowadays it is possible to afford a global view of the state of a proteome by means of two-dimensional (2D) gel electrophoresis (2DE). The 2DE technique is a high-throughput option for measuring changes in expression levels of hundreds of individual proteins simultaneously. The comparison of 2DE gel images from different biological samples (treatments) is a common method used to study protein expression. Traditional experiments rely on comparing images from at least two different gels.

In 1997, a new method for protein expression analysis known as Difference Gel Electrophoresis (DIGE) was introduced (Ünlü *et al*., 1997). In this technique, up to three different biological samples are examined in parallel on the same gel. They are labeled with spectrally resolvable fluorescent cyanine CyDyes Cy2, Cy3 and Cy5. Samples are then mixed prior to isoelectrofocusing (IEF) and resolved on the same 2D gel. Each CyDye will give an independent channel of measurement (Marouga *et al.*, 2005). Usually Cy2 is used to label a reference sample (a mix of all the experimental samples) (Alban *et al.* 2003). This reference pool is commonly used for spot matching during image analysis. It has also been applied as a normalization channel for spot comparison within and between gels. As claimed by creators of this methodology, using an internal standard labeled with a resolvable CyDye allows avoidance of major gel running effects, providing a more accurate comparison of spot volume.

The primary goal of this kind of experiments is the detection of proteins showing a statistically significant difference on expression under different experimental conditions. This should be accomplished in such a way that it would be possible to have an optimal control of both false positives and false negatives differentially expressed proteins. In DIGE the usual statistical analyses are based on Student's *t* test and simple ANOVA in a spot by spot (protein-by-protein) basis. Some properties about the distribution of the data are assumed to apply the referred tests. However, in many experimental situations such distributional requirements are difficult to meet. For this reason, analytical software for DIGE, provides tools for data normalization. Particularly, DeCyder software (GE Healthcare) (Amersham, 2003) uses an ad-hoc normalization procedure based on the use of the Cy2 CyDye as a reference channel. However, some controversy exists about the impact of this internal standard on experimental

[*]To whom correspondence should be addressed.

variance and noise (Karp *et al.*, 2005), which suggest the need for further research.

In a typical DIGE experiment several sources of variations can be identified *a priori*. Some of them act at a biological level –i.e. treatments-, and others depend on the technology itself, such as gel, CyDye and spot (protein) variation. A CyDye effect could arise when one CyDye reagent may be more efficiently coupled to proteins than the others, or one of the CyDyes may render consistently different quantum yields (Mujumdar *et al.,* 1993). Significant CyDye effects were described in the protein expression profile from DeCyder normalization data (Krogh *et al*., 2007). Variations due to experimental effects are not biologically interesting. They should be estimated and removed in some way before differential expression analysis is conducted. Otherwise, the differential expression test can lose power and miss biologically relevant information.

The goal of this research is to introduce an alternative framework for the statistical analyses of DIGE experiments. The framework is general enough to handle an arbitrary number of treatments and experimental effects that affect the development of the proteomic experiment. In doing so, we approach the problem by means of a two-stage statistical mixed model. A linear mixed model (Demidenko, 2005) in the first stage accounts for the underlying variance and covariance data structure. In this first step we attempt to remove noisy experimental effects (normalization model) over the log raw data. The internal standard Cy2 is introduced as a covariate in the normalization model, and its impact on the detection of differentially expressed proteins is analyzed. In this way the use of Cy2 as an internal standard and can be statistically evaluated. In the second stage we deal with testing protein-by-treatment interactions - i.e. the different expression of a given protein related to the treatment effects- in a protein-by-protein (P-by-P) basis. The P-by-P model (the second analytical stage) proposed here is similar to that used by Krogh et al. (2007). However, in our approach the residuals from the selected normalization model are used. On the contrary, the approach in Krogh et al. 2007, uses the residuals from the DeCyder ad-hoc normalization. According to the DeCyder documentation the resulting residual distributions are homogeneous in variances and do not contain dye nor gel effects (Amersham, 2003). In addition, our approach involves the use of statistically based techniques, such as the Akaike and Bayesian Information Criteria and the Likelihood Ratio Test, to evaluate the normalization run in the first stage. Therefore, the selection of the normalization model, as proposed here, is more objective and easy to control by the researcher.

To prove the usefulness of this model-based framework, we performed a step by step analysis of the various effects in the model to quantify their impact on the residual distribution. This modeling strategy was tested with data from the analysis of secretome (i.e the proteome of conditioned medium) from melanoma cells that express differential levels of the tumorigenic protein SPARC (Secreted Protein Acid and Rich in Cysteine). SPARC is a secreted glycoprotein overexpressed in melanoma and other tumors (Bos *et al*. 2004). For example, SPARC expression by melanoma and glioma cells has been linked to an aggressive phenotype in vivo (Ledda *et al.* 1997, Rempel *et al.* 1999). However, little is known about the molecular mechanisms that are affected by SPARC during tumor growth. In pursue of molecular mediators of SPARC protumoral activity, we have design DIGE experiments to compare the expression levels of secreted proteins in two cell lines (treatments) with differential expression of SPARC. By using small interfering RNA (siRNA) a stable cell clone (L2F6) of human melanoma MEL-LES have been developed in which SPARC expression was downregulated. SPARC downregulation at L2F6 abolished tumor growth in a murine in vivo model (Sosa *et al*. 2007). In the current research, we have analyzed quantitative data from A DIGE experiment comparing protein levels of four different conditioned media from L2F6 with matched media from control cell line, LBLAST. Previous experiments using Western blotting (an independent quantitative technique) had proved quantitative differences between treatments for four proteins present in our study (Sosa *et al*., 2007). These four proteins, including SPARC itself, were known to be differentially expressed in the L2F6 extracellular medium with respect to that of LBLAST. We have used such four proteins in this work as gold standards to evaluate the power of the proposed model strategy for detecting differentially expressed proteins.

2 METHODS

2.1 Sample and Data Preparation

Melanoma cell lines and clones were grown following protocols described in Sosa *et al*. 2007. For preparation of the conditioned media of human melanoma, cells were seeded according to similar percentages of confluence and grown for 24hrs in serum-containing medium, washed three times with PBS and kept in serum-free medium for additional 24 hrs. Conditioned media were collected and processed as in Sosa *et al*. 2007. After quantification, proteins were labeled with CyDyes as suggested by manufacturers (GE Healthcare). Table 1 summarizes the labeling and gel experimental design. Cy2, Cy3 and Cy5 labeled samples were mixed according to the experimental design. For this purpose, the reference pool was prepared by mixing equal amounts of proteins from each biological sample in the experiment and labelling them with Cy2 Dye. IEF was performed in an Ettan IPGphor isoelectrofocusing system (GE Healthcare), using 18-cm strips covering pH 4-7. The resulting strips were then loaded and run on 12.5% acrylamide gels using the Ettan Dalt Six system (GE Healthcare).

For image acquisition, labelled proteins were visualized using the Typhoon 9400 Imager (GE Healthcare). Cy2 images were scanned using 488 nm laser and an emission filter of 520nm. Cy3 images were scanned using a 532nm laser and an emission filter of 580nm. Cy5 images were scanned using a 633nm laser and a 670nm Band Pass 30Hz emission filter. A narrow BP emission filter ensures that there was negligible cross-talk between fluorescence channels. Photomultiplier voltage was selected for each channel to ensure no spot was signal-saturated. All gels were scanned at 100 µm resolution, and images were cropped using ImageQuant V5.0 (GE Healthcare) prior to analyses.

Fisrt, gel image analysis was performed using DeCyder 6.5V (GE Healthcare). Spot detection was carried out on image pairs consisting of the pooled standard and each sample from the same gel. These two images overlay and allow direct measurement of volume ratios of spots between the standard and the sample. Standard spot maps were matched between gels in order to identify the same spot/protein across gels. DeCyder Batch Processor (Amersham, 2003) provides two different data sources: the raw data, which are the volume measured over each intensity channel, and the corresponding normalized volume based on its internal standard. Raw data were log transformed.

2.2 Experimental Design

Table 1 displays the experimental design that yielded the data used in this paper. Four gels were used with three biological samples in each gel. In all cases, the Cy2 channel corresponds to the same reference pool composed by equal amounts of all samples analyzed in the experiment. Biological variation was addressed by conditioning four samples of each cell line at different times. Gels 3 and 4 are CyDye swaps of gels 1 and 2. From the experimental design one can identify the factor effects that could be included in the statistical model. In this design Cy3 and Cy5 are neither confounded with treatment, nor with gel effects. However, the Cy2 effect is not separable from the gel effects.

Table 1: Experimental design of a DIGE proteomic study.

Gel	Cy2	Cy3	Cy5
1	Reference Pool	LBLAST	L2F6
2	Reference Pool	LBLAST	L2F6
3	Reference Pool	L2F6	LBLAST
4	Reference Pool	L2F6	LBLAST

If rows 2 and 3 of Table 1 are interchanged, the design can be easily interpreted as a repeated Latin Square design (Cochran *et al*. 1957).

2.3 Statistical modeling

Classical statistical linear models assume that the observed variable can be described as $Y \sim N(\mu, \sigma^2)$ where the mean μ can be decomposed as $\boldsymbol{X\beta}$, a linear combination of experimental effects $\boldsymbol{\beta}$ related to Y by means of covariables $\boldsymbol{x}$, which are the columns of matrix $\boldsymbol{X}$. In DIGE the observed variable Y can be expressed as $Y_d = log(I_d)$ where I_d is the measured intensity on channel d and $d=Cy3$ or $Cy5$. The vector $\boldsymbol{\beta}$ is composed of the fixed effects that could provide an additive contribution to the mean.

From the experimental design in Table 1, the model is written as

$$Y_{tdgp} = \mu + T_t + D_d + G_g + P_p + TP_{tp} + \varepsilon_{tdgp} \quad \text{(Eq.1)}$$

where Y_{tdgp} is the observation for Treatment "t", within CyDye "d" in Gel "g" for Protein (spot) "p" . The constant μ represents an overall mean value; T, D, G and P represent the main effects of Treatment, CyDye, Gel, and Protein, respectively. The term TP represents the interactions between main effects. The last term is an stochastic error for which we assume a $N(0,\sigma^2)$ distribution.

The overall model on Eq. 1 can be decomposed into two interconnected equations, as proposed by Wolfinger *et al*. (2001) for the microarray technology. The first equation will be referred as the *normalization model*. This first model is expected to remove all the experimental effects not related to differentially expressed proteins, and includes the following terms:

$$Y_{tdgp} = \mu + T_t + D_d + G_g + \varepsilon_{tdgp} \quad \text{(Eq.2)}$$

The difference between the observed value Y_{tdgp} and the model predicted value is the residual term,

$$r_{tdgp} = Y_{tdgp} - \hat{\mu} + \hat{T}_t + \hat{D}_d + \hat{G}_g$$

where the hat over the symbol for the effect means "the estimate of". In a second stage, and assuming that residuals from the first model have a normal distribution, we define a *protein model* as:

$$r_{tdgp} = P_p + TP_{tp} + \gamma_{tdgp} \quad \text{(Eq.3)}$$

where $\gamma_{tdgp} \sim N(0, \sigma_\gamma^2)$.

From this model, the interaction term TP allows testing differential expression between treatments for each protein in the experiment.

In the classical approach provided by DeCyder, the observed data used for analysis is $Y_g^* = \log\left(I / Cy2_g^{\delta_g}\right)$, where g refers to gels and "δ" is a "centering" constant (Amersham, 2003). Here we decompose the previous expression as $Y_g^* = Y_g - \delta_g \cdot \log\left(Cy2_g\right)$. In this way the normalization in Eq. 2 can be extended by the inclusion of the covariate *log(Cy2),*

$$Y_{tdgp} = \mu + T_t + D_d + G_g + \delta_g \cdot log(Cy2_g) + \varepsilon_{tdgp} \quad \text{(Eq.4)}.$$

In all equations above, it was assumed that $\varepsilon_{tdgp} \sim N(0,\sigma^2)$, meaning that all error terms have the same mean and variance.

We observed in our experiment that the standard assumptions about the distribution of the error terms (homoskedasticity) were not fulfilled by either the DeCyder Normalized volumes, or by the log raw data. Therefore, we suggest that the error distribution of both, the log raw data and the normalized data, is best described by several $N(0,\sigma^2_{tg})$ densities, where "tg" refers to a specific treatment × gel (TG) combination, i.e heteroskedastic or non-homogeneous variance models are more realistic for the data. The observed/suggested heteroskedasticity across TG combinations can be modeled by a linear mixed model accounting for the residual covariance structure. A proper estimation method of the dispersion parameters in these models is restricted or residual maximum likelihood (REML, Patterson and Thomposn, 1971). In this paper, several mixed models based on different combinations of fixed and random effects and different co-variance structures were tested as normalization models (Table 2). The fitted models were compared by means of likelihood based statistics such as the Akaike Information Criteria (AIC), the Bayesian Information Criteria (BIC), and the likelihood ratio test (LRT). The first two criteria attempt to select the most informative and parsimonious model (i.e. the one that explains the variability in the data with the fewest possible number of parameters) by means of a penalized likelihood. In both cases, and either in the current versions of R and SAS software languages, smaller values of the criteria imply better models. But they do not provide a value for the significance of the difference between alternative models. The LRT can be used to statistically test the difference between two nested models, i.e. the parameters in one model are a subset of the parameters in the other model. The LRT can be applied to test hypotheses on the fixed-effects, or on the parameters associated to the covariance matrix. ML estimates are recommended if model comparison is based on fixed-effects. Instead REML estimates are preferred in comparing covariance structures as the method takes into account the loss in degrees of freedom due to the need of estimating the variance components along with the fixed-effects (Harville, 1977)

Boxplots of the residuals were used to check model assumptions for the protein-by-protein model in the second stage of our proposal. The model that best fitted the data was selected to test the TP interaction. The best model will produce better standard errors for the parameter estimates and consequently an improvement of the statistical efficiency of treatment comparisons for each protein.

3 RESULTS

From the right panel of Fig. 1 it is possible to see that the normalized volumes, calculated by DeCyder, show a symmetric distribution but the variances across treatments and gels are not homogeneous. Therefore, the standard assumption about the distribution of the error terms could be not fulfilled. Fig.1 suggests that the log raw data as well as the DeCyder-normalized data could be better represented by several normal distributions for the errors. Each distribution is characterized through its mean and variance, i.e $N(0,\sigma^2_{tg})$ where "tg" refers a specific TG combination. In Fig. 1 it is possible to observe, mainly for the log raw data, that the mean values for CyDye Cy5 are lower than for CyDye Cy3 in each gel.

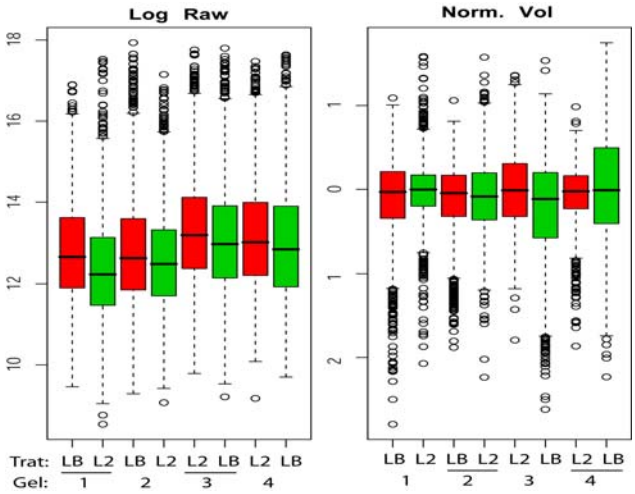

Figure 1: Left, boxplots of log raw data distributions for TG combinations (Trat:Treatment, LB:LBLAST, L2:L2F6 and Gel 1 to 4) . Right, boxplots for the same distributions but using DeCyder normalized log volumes.

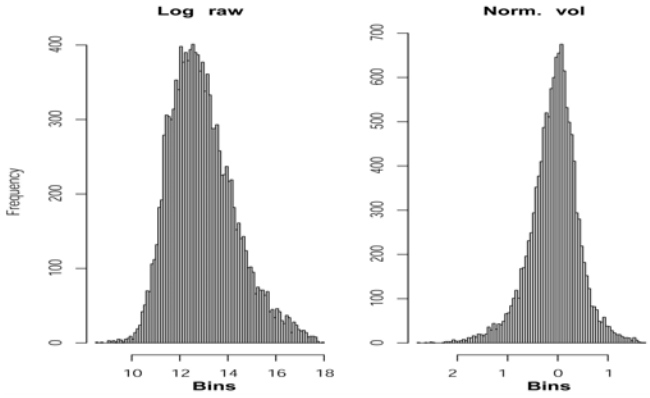

Figure 2: Histograms showing the distribution of the logged raw data (left) and the normalized volumes from DeCyder (right)

Fig. 2 shows the logged raw data distribution and residual distribution after DeCyder normalization for the whole data set. It is possible to observe that the distributions are

not symmetric for the logged raw data, showing a right tail. To overcome this observation, a set of normalization models were evaluated here (Table 2). They allowed us to analyze the incidence of the different experimental factors in the protein analysis.

After fitting each normalization model, we analyzed the model residuals by each TG combination. Boxplots of the residual distribution from all tested normalization models, M1 to M6, are shown in Fig. 3.

Table 2: Linear fixed and mixed models for normalization of DIGE data

Model	Equation	Stochastic assum.
M1	$Y_{tdgp}=\mu+T_t+\varepsilon_{tdgp}$	$\varepsilon \sim N(0,\sigma^2)$
M2	$Y_{tdgp}=\mu+T_t+D_d+\varepsilon_{tdgp}$	$\varepsilon \sim N(0,\sigma^2)$
M3	$Y_{tdgp}=\mu+T_t+D_d+G_d+\varepsilon_{tdgp}$	$\varepsilon \sim N(0,\sigma^2)$
M4	$Y_{tdgp}=\mu+T_t+D_d+G_d+\delta_g\cdot\log(Cy2_{gp})+\varepsilon_{tdgp}$	$\varepsilon \sim N(0,\sigma^2)$
M5	$Y_{tdgp}=\mu+T_t+D_d+\delta_g\cdot log(Cy2_{gp})+G_g+\varepsilon_{tdgp}$	$\varepsilon \sim N(0,\sigma^2), G\sim N(0,\sigma_g^2)$
M6	$Y_{tdgp}=\mu+T_t+D_d+\delta_g\cdot log(Cy2_{gp})+\varepsilon_{tdgp}$	$\varepsilon \sim N(0,\sigma_{tg}^2)$

Y: log(I) from DIGE technology; T: treatment effect; D=CyDye effect; G=gel effects, TD treatment-by-CyDye interaction effect; Cy2 continuous covariate containing reference channel values and ε random error terms.

Models M1, M2 and M3 do not include the covariate log(Cy2). Boxplots of the three first models (top panels in Fig. 3) show a progressive improvement towards the presence of homoskedastic errors, when adding the effects of treatments, CyDye and Gel as fixed effects in the model. For instance, model M1 only takes into account an overall mean and the treatments effects. Note that the treatment mean value was successfully removed, but CyDye and Gel effects are still present (notice that boxes are biased between CyDyes and between gels). The effect of treatments was not significant under M1 (p-value = 0.21), but significant under M2 (p=0.037) and M3 (p=0.034). The fitting criteria (Table 3) show that M3 is preferred over the two other previous models. Therefore, treatments, as well as CyDye and gel effects, should be included in the model. This conclusion is not evident from the histograms obtained from the distribution of the residuals in Fig. 4. However, boxplots (Fig.3) and histograms (Fig. 4) of these three models suggest that the residuals follow a non symmetrical distribution. When the covariate *log(Cy2)* was entered into the model (M4, M5 and M6), the distribution of the residuals became more symmetric (bottom panels in Fig. 3). The inclusion of the covariate produces a significant decrease in the Akaike, Bayesian and log-likelihood information criteria (Table 3). However, M6 that accounts for different residual variances in each TG combination had a better fit (lowest AIC, BIC and loglikelihood values) than the other models (Table 3). The model with heterogeneous error variance (M6) produced the best normalization of the melanoma data.

The estimates of the residual variances for each TG combination are shown in Table 4. The estimated coefficients for *log(Cy2)* in M6 were statistically different from 1 ($p < 0.001$). The pair-wise comparisons of these coefficients between gels displayed significant differences. Once the "normalization model" was applied over the whole data set, spot-by-spot (protein-by-protein) analyses were run on the residuals (Eq. 3). In this second stage, we model the protein main effect (Pp) and the Protein-by-Treatment interaction (TP - our biological target). Fig. 5 shows the boxplots and the histograms of the residuals $(\hat{\gamma})$ for the protein-by-protein analysis after M6 normalization. It is verifiable that in both cases the new residuals obtained from Eq.4 follow normal distributions, assumption which is needed to test the hypothesis on treatment differences. Table 5 displays the estimated Treatment by Protein (TP) effects for the four known differentially expressed proteins (SPARC, N-Cadherin, 1169, HSP27 - see Sosa et al., 2007 and unpublished results). The estimates were obtained using the protein-by-protein analyses from M3, M6 and DeCyder normalization strategies.

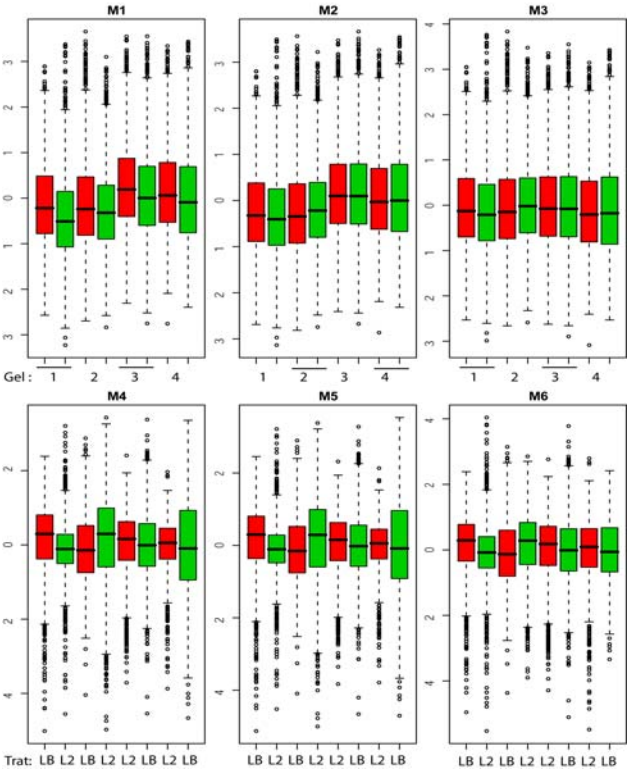

Figure 3: Boxplot of residual distributions by TG combinations from the six normalization models in Table 2.

Observe that the use of residuals from models not including the covariate *log(Cy2)* induces the possibility of missing some differentially expressed proteins. All the target proteins were detected using the residuals from the het-

eroskedastic model (M6). The normalized log volume from DeCyder also allowed the detection of the target proteins.

Table 3: Information criteria for each of the fitted models.

Model	Df	AIC	BIC	logLik
M1	3	43255	43277	-21624
M2	4	43143	43173	-21567
M3	5	42712	42749	-21351
M4	9	17297	17364	-8639
M5	10	17248	17322	-8614
M6	16	16198	16317	-8083

Df: degree of freedom. AIC: Akaike information criteria, BIC: Bayesian Information Criteria. LogLik: Log likelihood. (all the values were provided by R code)

Using the residuals from model M6, the estimated fold change between treatments can be calculated by means of $\exp(b_{TP})$, where b_{TP} is the estimated coefficient for the Treatment by gel interaction term in Eq. 3 of a particular protein.

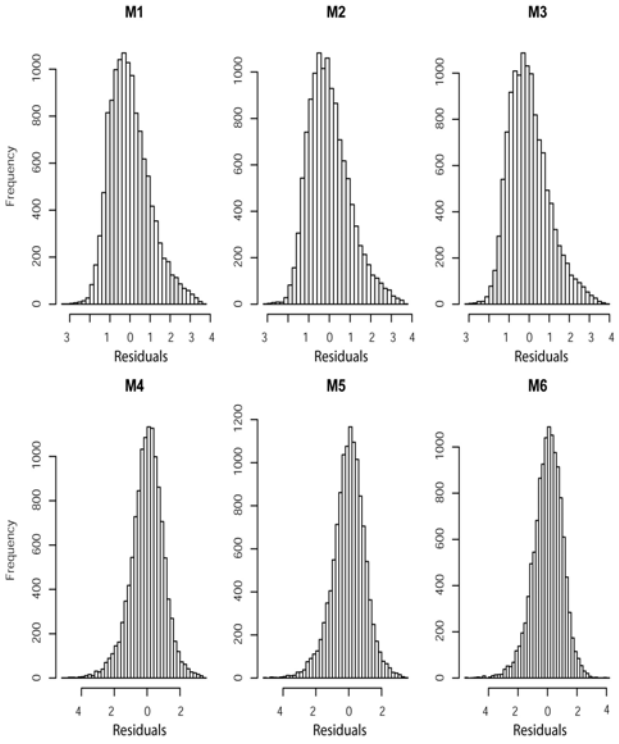

Figure 4: Histograms of the overall residuals for the different normalization models.

Table 4: Estimated variances at each TG combination under M6 normalization model.

	Gel 1		Gel 2		Gel 3		Gel 4	
Treat	LB	L2	LB	L2	LB	L2	LB	L2
σ^2_{tg}	0.220	0.169	0.191	0.263	0.187	0.182	0.292	0.147

LB:LBLAST, L2:L2F6

4 DISCUSSION

The use of DIGE technology for massive protein expression analysis has been growing since their introduction in the late 90's. Based on the similarity with microarray technology, several approaches have been applied to DIGE in order to remove some specific non-biological effects, such as the CyDye-intensity dependence (Fodor *et al.*, 2005), variance stabilization (Kreil *et al.*, 2004), and some location specific effects (Kultima *et al.*, 2006). However, lack of established analyses of protocols causes that different pre-processing strategies (normalization), result in different proteins to be significant (Meleth *et al.*, 2005). We propose a mixed model based framework for statistical analysis of DIGE data. This framework is flexible enough to be applied over a wide range of experimental designs, with respect to the number of treatments and other sources of variation. It was shown that the application of the two stage linear mixed model allows an overall data normalization, and the estimation of treatments effects in a protein by protein basis. The identification of those non-biological effects that contribute the most to the normalization stage is a particularly relevant finding for quality control of the lab (Draghichi *et al.*, 2003). The removal of noisy effects was also addressed as an important design issue in cDNA microarray experiments (Yang *et al.*, 2002).

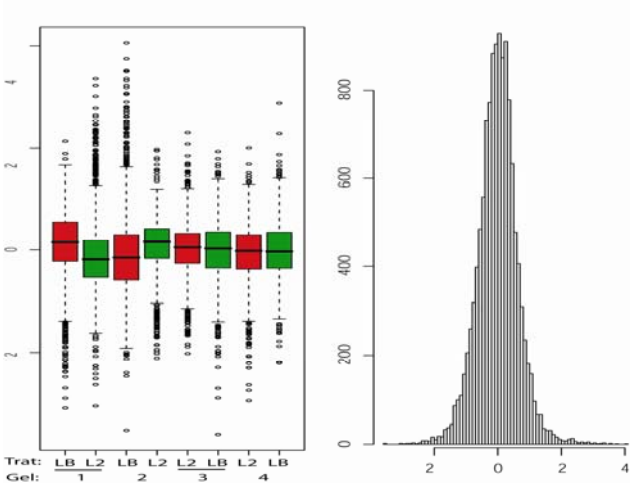

Figure 5: Boxplots and histogram of the protein-by-protein residuals (Eq. 3) under M6 normalization model.

The statistical modeling approach to 2D-DIGE data presented here, allows verifying model assumptions prior to the protein analysis in a unified way. Model selection can be done by means of statistically sound fitting criteria,

such as the Akaike and Bayesian Information Criteria, as well as the Likelihood Ratio Test (Littell *et al*., 2006).
Accounting for an appropriate covariance structure of the data provides suitable backdrops by which to assess the significance of interesting effects. Thus, better estimations of "p" value are expected to be obtained. This will impact positively on any function of the p-values such as the false discovery rate (FDR), which is a common practice in experimental settings involving multiple tests of hypotheses (Storey *et al.*, 2003; Qian *et al.*, 2005; Karp *et al.*, 2007).

Table 5: Estimated coefficients for TP interaction for the known differentially expressed proteins using the protein-by-protein model under M3, M6 and DeCyder® normalization.

	Normalization strategy					
	Linear model M3		Linear mixed model M6		Software DeCyder	
Protein	b_{TP}	P	b_{TP}	P	AR	p
SPARC	0.986	0.064	3.324	0.015	4.265	0.003
N-Cadherin	0.686	0.699	2.131	0.021	2.893	0.025
1169[c]	0.537	0.006	1.584	0.003	2.166	0.0003
HSP27	-0.48	0.298	-1.27	0.008	-1.81	0.001

b_{TP} = the estimated coefficient for the TP interaction term in the protein model. p: p value. [a] SPARC, [b] *N*-Cadherin (Sosa et al2007), [c]unpublished [d] HSP27 (Sosa et al2007). AR: Log Average Ratio.

The use of the modeling strategy in two-stages (Eq. 2 and 3) instead of fitting a protein model (Eq.1), has the advantage of using more information in the normalization of the data. In addition, it provides more efficient estimation of the CyDye and gel effects. It was found in the melanoma cell experiment analyzed here that the Cy3 channel consistently displayed higher values than the Cy5 channel. The proposed approach can account for this effect and handle it appropriately.
The inclusion of the covariate *log(Cy2)* in the normalization model was helpful in order to obtain symmetric residual distributions. After using the covariate, the distributions were similar to those achieved by DeCyder software. This fact can be understood if we identify $_g$ as the estimated regression coefficient for the covariate *log(Cy2$_g$)* in the gel "*g*" and rewrite the normalization model as follows

$$Log\left(I_{..g}\right) = ... + \delta_g \cdot Log\left(Cy2_g\right) + \varepsilon_{..g} \Rightarrow$$
$$\Rightarrow Log\left(I_{..g}\right) - \delta_g \cdot Log\left(Cy2_g\right) = ... + \varepsilon_{..g} =$$
$$= Log\left(I_{..g} \Big/ Cy2_g^{\delta_g} \right) = ... + \varepsilon_{..g}$$

where $g = 1..4$, with the dots (…) indicating Treatment, CyDye and Gel effects. The above expression resembles the normalization equation published in the DeCyder® documentation. The main difference between both is that the coefficients $_g$ is here estimated by means of likelihood based procedures using all the data. The procedure employed by DeCyder to estimate the *log(Cy2)* coefficient is based on Least Square Means, and implemented for each gel-CyDye combination separately.
When the *log(Cy2)* covariate was not included (models M1 to M3) the residual distributions showed homogeneity of variances but they were strongly skewed. The use of the *log(Cy2)* produce symmetric distribution of the residuals, but introduce data correlations within gels. Such data correlation could be explained by the fact that for a particular protein the *Cy2* effect varies across gels. The estimated variance components for models including random gels effects, with and without the covariate *log(Cy2)*, were 0.455 to 0.079, respectively (data not shown). Therefore, higher correlation between data in the same ge should be expected when using the covariate *log(Cy2)*. Thus, the heteroskedastic model for the residual covariance could be a good choice during the normalization stage. With the data set used here, the best normalization model (lowest BIC and AIC values - Table 4) was the one accounting for a heteroskedastic residual covariance structure in the error terms (model M6), i.e a model accounting for non-constant variances across TG combinations (Table 4).
The correlation induced by the inclusion of the *Cy2* channel was also suggested by the analysis of the distribution of the "p" values in simulated and real self-self experiments (Karp *et al.*, 2007). In spite of this observation, the inclusion of the *Cy2* channel seems to be relevant in DIGE experiments as it makes it possible to fulfill the required distributional assumptions in order to infer differential expression. Although including the *Cy*2 channel induces a correlation structure, these covariances can be appropriately modeled using the mixed model normalization approach. In the study of melanoma cells, most of the known differentially expressed proteins were not identified in those models lacking the covariate *log(Cy2).*
The model accounting for the internal reference and for heteroskedasticity of the distribution during normalization (M6), allowed the proper identification of known differentially expressed proteins in the melanoma cell study. The advantage of the use of statistical modeling for normalization is the opportunity to control source of well known variation prior to protein identification. The ad-hoc normalization from DeCyder, even removing experimental effects, does not allow the identification of the relative contribution of them. Even more, the statistical significance of these sources of variation can be formally tested. Our approach allows fitting specific models that best describe the data at hand and verify the fulfillment of the required distributional assumptions for the protein analysis. All the normalization coefficients are estimated by means of likelihood based procedures being in this way less sensitive to missing data. Moreover, the knowledge gained by fitting an experimental design-based model is a welcome input in the design of new 2D-DIGE experiments. A more complex experimental design, where extra terms could be added, would undoubtedly yield more in-

teresting data to show the advantages of this mixed model approach. A guide for the application of mixed model approach in the normalization stage is available in the Supplementary Material.

ACKNOWLEDGEMENTS

This project has been partially supported by Catholic University of Córdoba, and CONICET under grant PIP 5338. We wish to thank Marisol Fernández Rodríguez for her help on DIGE experiments. This work has been supported in part by grants from the Agencia de Promoción Científica y Tecnológica (to OLP and ASL), Fundación Antorchas (to ASL) and the Third World Academy of Sciences (to ASL). We acknowledge the continuous support to OLP of Fundación René Baron and the Amigos de la Fundación Leloir para la Investigación en Cáncer (AFULIC). We would like also thank to the reviewers for their valuable suggestions that help us to improve the manuscript.

REFERENCES

Alban BA. David SO. Bjorkesten L. Andersson C. Sloge E. Lewis S. Currie I. (2003) A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*. 3(1):36-44.

Alvarez MJ, Prada F, Salvatierra E, Bravo AI, Lutzky VP, Carbone C, Pitossi FJ, Chuluyan E, Podhajcer OL. (2005) Secreted Protein Acid and Rich in Cysteine produced by human melanoma cells modulates polymorphonuclear leukocyte recruitment and umor cytotoxic capacity *Cancer Res*. 65:12: 5123-5131

Amersham (2003): DeCyder Differential Analysis Sotware User manual, version 5.0. Amersham Biosciences.

Bos, TJ., Cohn SL., et al. (2004). International Hermelin brain tumor symposium on matricellular proteins in normal and cancer cell-matrix interactions. *Matrix Biol* 23(1): 63-9

Chu T, Weir BS, Wolfinger RD (2004) Comparison of Li-Wong and loglinear mixed models for the statistical analysis of aligonucleotide arrays. *Bioinformatics* 20:4:500-506.

Cochran WG. Cox GM. (1957) Experimental Designs. 2nd Ed. Wiley & Sons Inc. Canada.

Demidenko E. (2005) Mixed Models: Theory and Applications. Wiley-IEEE Hoboken, New Jersey. USA

Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, Tainsky MA. (2003) Noise Sampling Method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics* 19:11, 1348-1359

Fernández EA, Alvarez M. Podhajcer OL. Stolovitzky G. (2007) Genómica Funcional: en busca de la función de genes. *National Academy of Sciences Argentina* in press (in Spanish)

Fievet J, Dillman C, Lagniel G, Davanture M, Negroni L, Labarre J, de Viene D. (2004) Assessing factors for reliable quantitative proteomics based on two dimensional gel electrophoresis. *Proteomics* 4: 1939-1949

Fodor IK. Nelson DO. Alegria-Hartman M. Robbins K. Langois RG. Turteltaub KW. Corzett TH, McCutchen-Maloney S. (2005) Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyder. *Bioinformatics* 21 3733-3740

Harville, DA. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc*. 72, 320-340.

Karp NA, Lilley KS (2005) Maximizing sensitivity for detecting changes in protein expression: Experimental design using minimal CyDyes. *Proteomics* 5: 3105-3115

Karp NA, McCormick P, Russel M, Lilley KS. (2007). Experimental and Statistical Considerations to avoid False Conclusions in Proteomic Studies Using Differential in-Gel Electrophoresis. *Mol. & Cell. Proteomics*. DOI 10.1074/mcp.M600274-MCP200

Kreil PD. Karp NA. Lilley KS.: (2004) DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. *Bioinformatics* 2026-2034

Kultima K, Scholz B. Alm H. Sköld K. Svensson M. Crossman AR. Bezard E. Andrén PE. Lönnstedt I. : (2006) Normalization and expression changes in predefined sets of proteins using 2D gel electrophoresis: A proteomic study of L-DOPA induced dyskinesia in an animal model of Parkinson's disease using DIGE. *BMC Bioinformatics* 7:475

Krogh M. Liu Y. Waldemarson S. Valastro B. James P. (2007) Analysis of DIGE data using a linear mixed model allowing for protein-specific CyDye effects. *Proteomics* 7,4235-4244

Ledda MF. Adris S. Bravo AI. Kairiyama C. Bover L. Chernajovsky Y. Mordoh J. Podhajcer, OL.: (1997) Suppression of SPARC expression by antisense RNA abrogates the tumorigenicity of human melanoma cells. *Nat Med* 3, 171-175.

Littell RC. Milliken GA. Stroup WW. Wolfinger RD, Schaubenberger O. (2006) *SAS for Mixed Models*. SAS Inst. NC USA.

Marouga R, David S, Hawkins E (2005) The development of the DIGE system: 2D fluorescence difference gel analysis technology. *Anal. Bioanal. Chem*. 382,669-678.

Meleth S, Deshane J, Kim H. (2005) The assessment for well-conducted experiments to validate statistical protocols for 2D gels: different pre-processing = different list of significant proteins. *BMC Biotechnology*. Doi:10.1186/1472-6750-5-7

Mujumdar RB, Ernst LA, Mujumdar SR, Lewis CJ, Waggoner AS. (1993). Cyanine dye labeling reagents: sulfoindocyanine succinimidyl esters. *Bioconjug Chem*. 1993. 2:105-11

Qian HR, Huang S (2005) Comparison of false discovery rate methods in identifying genes with differential expression. *Genomics* 86, 495-503

Patterson HD, Thompson, R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545–554.

Rempel SA. GE S. Gutiérrez JA. (1999) SPARC: a potential diagnostic marker of invasive meningiomas. *Clin Cancer Res* 5, 237-241

Sosa MS, Girotti MR, Salvatierra E, Prada F, Lopez de Olmo JA, Gallango SJ, Albar JP, Podhajcer OL, Llera AS. (2007) Proteomics analysis identified N-cadherin, clusterin and HSP27 as mediators of SPARC activity in melanoma cells. *Proteomics* doi: 10.1002/pmic.200700255

Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. *Proc. Nat. Acad. Sci, USA* 100,9449-9445

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

Wolfinger RD, Gibson G, Wolfinger ED, Bennet L, Hamadeh H, Bushel P, Afshari C, Paules RS. (2001) Assessing Gene Significance from cDNA Microarray Expresión Data via Mixed Models. *Jour. Comp. Biol*. 8:6, 625-637

West BT, Welch KB, Gatecki AT. Linear Mixed Models: A Practical Guide Using Statistical Software. (2007) Chapman & Hall/CRC. Boca Ratón, USA

Yang YH, Speed T (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet*. 3:579-588