Non-parametric quantification of protein lysate arrays

Jianhua Hu1,*, Xuming He2, Keith A. Baggerly1, Kevin R. Coombes1, Bryan T.J. Hennessy3 and Gordon B. Mills4

1Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, 2Department of Statistics, University of Illinois at Urbana-Champaign, 3Department of Gynecologic Medical Oncology and 4Department of Systems Biology, University of Texas M.D. Anderson Cancer Center, USA

Received on February 7, 2007; revised on April 10, 2007; accepted on May 19, 2007
Advance Access publication June 28, 2007

ABSTRACT

Motivation: Proteins play a crucial role in biological activity, so much can be learned from measuring protein expression and post-translational modification quantitatively. The reverse-phase protein lysate arrays allow us to quantify the relative expression levels of a protein in many different cellular samples simultaneously. Existing approaches to quantitate protein arrays use parametric response curves fit to dilution series data. The results can be biased when the parametric function does not fit the data.

Results: We propose a non-parametric approach which adapts to any monotone response curve. The non-parametric approach is shown to be promising via both simulation and real data studies; it reduces the bias due to model misspecification and protects against outliers in the data. The non-parametric approach enables more reliable quantification of protein lysate arrays.

Availability: Code to implement the proposed method in the statistical package R is available at: http://odin.mdacc.tmc.edu/jhu/lysatearray-analysis/
Contact: jhu@mdanderson.org
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Although DNA microarray technologies are able to simultaneously monitor the mRNA levels of thousands of genes, the post-translational modifications that result in the complexity of the proteome cannot be detected at the genomic level. It is often necessary to measure proteins directly in biological and medical research, because most of the known biological effector molecules and diagnostic markers are proteins. Current technologies used for detecting protein biomarkers include 2D polyacrylamide gel electrophoresis (Bernard et al., 2004; Hojlund et al., 2003), mass spectrometry (Ong et al., 2003) and enzyme-linked immunosorbent assays (Kakizawa et al., 2004; Toney et al., 2003). These technologies require large amounts of input material, and need good ways to identify markers. To overcome those shortcomings, various protein microarray technologies have been developed (Cahill et al., 2003; Ivanov et al., 2004; MacBeath and Schreiber, 2000).

Protein microarrays use forward-phase or reverse-phase approaches. The forward-phase approach allows for the simultaneous measurement of many different proteins across single samples. In contrast, the reverse-phase approach aims to measure a common protein across multiple samples. A commonly used reverse-phase assay is the tissue microarray (Camp et al., 2000; Kononen et al., 1998; Liu et al., 2002; Packeisen et al., 2003; Torhorst et al., 2001). However, its major drawback is that protein levels are typically assessed manually by a pathologist, which can potentially produce subjective bias and is only semi-quantitative.

In this article, we focus on quantification of another type of reverse-phase array which uses homogenized samples. These arrays are called tissue lysate or protein lysate arrays, as described in detail by Paweletz et al. (2001) (also see MacBeath, 2002). Mechanically, lysed tissue samples are first robotically spotted onto nitrocellulose-coated slides in a dilution series. The samples are then hybridized with a primary validated antibody tailored to recognize the protein of interest. In order to attach measurable labels, the arrays are developed by any number of techniques most frequently incubated with a biotinylated secondary antibody which recognizes the primary antibody. Finally, streptavadin-linked labels (which can be dyes, or precipitates, or others) are introduced and the streptavadin binds to the biotin. The array can then be scanned to produce an image file, and spots in the image are quantified using any of a variety of software packages.

Lysate arrays can be used to address a wide variety of questions by printing multiple samples on the same slide. For example, by printing samples from treated and untreated cell lines, the response to therapeutic agents can be assessed. Lysate arrays have been used to study signal pathway profiling related to prostate cancer (Herrmann et al., 2003; Paweletz et al., 2001). Nishizuka et al. (2003a) applied lysate arrays to profile 60 human cancer cell lines (NCI-60) and found that there was a high correlation between mRNA and protein levels for structure-related proteins. Jiang et al. (2006) used glioma protein lysate arrays to identify proteins that are involved in signaling pathways of cell proliferation, cell survival, apoptosis, angiogenesis and cell invasion. Ramaswamy et al. (2005) discussed verification of presence and quantification of human tissue samples for protein markers and compared the
results with those obtained with ELISAs. They concluded that protein lysate arrays could show high sensitivity, reproducibility and high-throughput fashion. Some other cancer and clinical-related studies using this technology and its variations include Amit (2006); Chan et al. (2004); Cheng et al. (2005); Espina et al. (2003); Espina et al. (2004); Lin et al. (2003); Murph et al. (2006); Nishi-zuka et al. (2003b); Sheehan et al. (2005); Tibes et al. (2006); Wulfkuhle et al. (2003); Yan (2003); Zha et al. (2004).

Regardless of the biological questions to be addressed, the relative amounts of protein in the arrayed samples need to be quantified as accurately as possible. The current commercial analysis package MicroVigene (http://www.vigenetech.com/product.htm) extracts the foreground and background intensity levels from the raw array image, and then estimates the protein expression level by fitting a 4-parameter logistic model to each dilution series, with the replicates being averaged or combined in the analysis. Another sample-by-sample approach (Mirecan et al., 2005) models the background-corrected spot values as linear when the log intensity is plotted against the dilution step number.

The quantification methods mentioned earlier work on the data one sample at a time. With reverse-phase tissue lysate arrays, however, the same antibody binding is taking place at each spot, so it is reasonable to assume that protein expression levels of different samples corresponding to the same antibody have similar chemistry, and can thus be fit using a common response curve increasing the confidence of the curve fit. The joint estimation approach assuming a parametric form of the mean response curve has been discussed extensively in Tabus et al. (2006). In particular, the logistic model is chosen due to its sigmoidal shape, which is believed to be consistent with the expected relationship between the observed intensity of a spot and the true protein concentration, due to quenching at high levels and background noise at low levels. This also forms the core of the newly developed R package called ‘SuperCurve’, publicly available at http://bioinformatics.mdanderson.org

The logistic model used in SuperCurve is of the form:

\[ y_{ij} = \alpha + \beta \frac{2^{g(x_i + \text{EC50})}}{1 + 2^{g(x_i + \text{EC50})}} + \varepsilon_{ij} \]  

(1)

where \( y_{ij} \) is the observed expression level at the \( i \)th dilution level of the \( j \)th sample, \( i = 1, \ldots, I \) and \( x_i \) defines the corresponding dilution level index at \( i \)th step, typically, \( x_i = i-(1 + I)/2 \) is used. The median effective concentration (denoted by EC50), which is the 50th percentile of the series, is the single quantity per sample or per dilution series to be estimated. The random errors \( \varepsilon_{ij} \) are assumed to be independently normally distributed with mean 0 and variance \( \sigma^2 \). The global model parameters \( \alpha, \beta \) and \( \gamma \) and the EC50s can be obtained via a non-linear least squares procedure.

The joint logistic modeling procedure was shown to improve the accuracy and dynamic range of the estimated protein expression levels over sample-by-sample estimation based on simulation studies and a real experiment conducted at the M.D. Anderson Cancer Center (Tabus et al., 2006). However, the rigid assumption of the logistic curve may lead to estimation bias for some samples.

We propose a more flexible non-parametric approach by assuming that the median of \( y_{ij} \) is equal to \( g(x_i + \text{EC50}) \) for some monotonically increasing function \( g \), without specifying a parametric form for \( g \). The non-parametric model can be highly adaptive to the data and avoid bias due to parameterization. Both simulation studies and real data examples show that the non-parametric approach is highly promising, as it produces consistent and robust results regardless of whether a logistic curve fits the data.

## 2 METHODS

### 2.1 Monotone smoothing

To avoid a rigid parametric form on the protein response curve \( g \), we choose a monotone smoothing technique to estimate the curve given a set of trial values on \( \text{EC50} \). The literature on monotone smoothing is rich, but we choose to use the R package \( \text{cobs} \) based on constrained b-spline approximations of He and Shi (1998); He and Ng (1999). We shall briefly describe this methodology in this section, but refer to He and Ng (1999) for the algorithmic details.

Given \( n \) pairs of realizations \((x_i, y_{ij})\) \( i = 1, \ldots, n \) with \( a = x_0 < x_1 < \cdots < x_n < x_{n+1} = b \), we wish to model the relationship as \( y_i = g(x_i) + \text{error} \) for some smooth and monotone, but otherwise unknown, function \( g \). The idea of \( \text{cobs} \) is to approximate \( g \) by a quadratic spline that minimizes the following objective function

\[ \min_{g} \text{deviation} + \lambda P(g) \]  

(2)

where

\[ \text{deviation} = \sum_{i=1}^{n} |y_i - g(x_i)| \]

measures infidelity to data, \( P(g) \) is a roughness penalty on the function \( g \) and \( \lambda \) is a smoothing parameter that controls the trade-off between infidelity to the data and roughness of the fit. When the roughness penalty \( P(g) = \max_{j} \int g(x)^2 \text{dx} \) is chosen, the solution to (2) is a quadratic natural smoothing spline. To add a monotonicity constraint on \( g \), all we need is to solve (2) subject to \( g(x) \leq g(x + 1) \). As shown in He and Ng (1999), this constrained optimization problem can be easily re-expressed as a linear program. As a result, the monotone function \( g \) can be estimated with very efficient algorithms.

The \( \text{cobs} \) package used a BIC-type criterion for selecting \( \lambda \). To further save computing time, \( \text{cobs} \) limits the number of potential knots to a prespecified value. In all our empirical work with lysate arrays, we choose 40 potential knots in the construction of splines. The effective dimensions of the splines, as determined by the \( \lambda \) selection, are typically much smaller than 10.

### 2.2 Non-parametric modeling

We propose the non-parametric model of the form:

\[ y_{ij} = g(x_i + \text{EC50}) + \varepsilon_{ij} \]  

(3)

where \( y_{ij} \) is the observed expression level at the \( i \)th dilution step of the \( j \)th sample, and \( x_i \) defines the corresponding dilution level index at \( i \)th step; we use \( x_i = i-(1 + I)/2 \) in our procedure. The median effective concentration (denoted by EC50) is the single quantity to be estimated for each dilution series or each sample. The median of the random errors, \( \varepsilon \), is assumed to be 0, while \( g \) is an unknown but monotonically increasing function.

We propose the following quantification algorithm to obtain the EC50’s, where \( g \) is estimated iteratively together with the EC50’s.

1. Obtain initial estimates of EC50 for each dilution series assuming that the protein expression intensities are linearly related to the true concentration level, i.e. \( y_{ij} = \alpha + \beta (x_i + \text{EC50}) + \text{noise} \).
The estimate \( \hat{a} \) is the minimum of the observed protein expression intensities of all samples on an array. The estimate \( \beta \) is the median slope over all the dilution series. Then an estimate of \( EC50 \) is obtained as the median of \( \{y_{ij} - \hat{a} - \beta x_i \}; i = 1, \ldots, I \).

(2) Based on the initial estimates of \( EC50 \), the \( R \) package \texttt{cobs} is used to obtain the monotonically increasing function \( g \) by regressing \( y_{ij} \) on \( x_i + EC50 \), with all \( (i, j) \) included.

(3) Conditional on the estimated curve \( g \), update the estimate of \( EC50 \) by minimizing \( \sum |y_{ij} - g(x_i + EC50)| \), the total distance for the \( j \)th dilution series. This step can be carried out using the function \texttt{nlrq} (Koenker and Park, 1994) implemented in the \( R \) package \texttt{quantreg}.

(4) To refine the estimate of \( EC50 \), we iterate Steps 2 and 3 until a convergence criterion is met. For example, an intuitive criterion is to stop the procedure if the difference of the deviance of the estimated curve to the observed data in two consecutive iterations is smaller than 0.1. However, we found in empirical studies that a small number of iterations (say, five) is sufficient to obtain reasonably good estimates of \( EC50 \).

We shall discuss further some of the issues that are worth exploring in the proposed algorithm.

(1) There are a number of ways to obtain reasonable initial values in Step 1. For example, Tabus et al. (2006) used a linear relationship between certain order statistics in computing the slopes. The SuperCurve package used a logistic curve instead of a linear relationship. Both approaches worked well in our experience.

(2) The \texttt{cobs} algorithm in Step 2 uses \( |y_{ij} - g(x_i + EC50)| \) in its objective function (for a given spline space for \( g \)). In Step 3, the same objective function is used to find \( EC50 \) given \( g \). The use of the \( L1 \)-norm in the objective functions has some important properties.

(a) We assume that the median protein expression as a function of the true protein concentration is non-decreasing and second-order differentiable, but we need to make no assumption of symmetry or additivity of the error terms. In contrast, the least squares approach minimizing \( \sum (y_{ij} - g(x_i + EC50))^2 \) could lead to biased estimates if the noise is not additive. To make this point clearer, consider the simple model of \( y_i = \epsilon g(x_i + EC50) \) for some multiplicative error \( \epsilon \), with a log-normal distribution. The true value of \( EC50 \) does minimize the expected value of \( \sum |y_{ij} - g(x_i + \epsilon)| \), over \( \epsilon \), but it does not minimize the expected value of \( \sum y_{ij} - g(x_i + \epsilon) \).

(b) When \( g \) and \( EC50 \)'s are estimated iteratively, the objective function \( \sum |y_{ij} - g(x_i + EC50)| \) decreases with every step of the iteration to ensure convergence. Although reasonable initial values are still needed to ensure convergence to the global minimum of the objective function, the monotonicity of the objectives along the iteration is a very desirable property.

(c) The use of the \( L1 \)-norm allows \texttt{cobs} to use efficient linear programming algorithms to solve for \( g \) in Step 2 with the monotonicity constraint.

(3) A parametric model interacts with the scale of \( y \). Sometimes the logistic curve fits well for protein intensity, but sometimes it fits better on the log-scale. With the non-parametric approach, the results are much less sensitive to monotone transformations on \( y \).

3 SIMULATION STUDIES

As a function of the protein concentration level, the protein response curve has been modeled by a logistic curve in the existing literature. The shape of the logistic curve makes good sense, but careful thoughts about the logistic curve reveal some limitations of this approach. In this section, the logistic fitting approach and the proposed non-parametric approach are compared through a series of simulation studies.

We carried out several simulation studies with data generated from various parametric forms of the response curve. We intended to mimic the patterns of real lysate array data. Throughout the study, we used a common set of true \( EC50 \)s of 40 samples, generated from the normal distribution with mean 1 and variance 3.5. We generated protein expression data of each sample in duplicate in most cases, unless otherwise specified.

The response curve in the first case is similar to a logistic curve but allows different scale parameters at the two tails. This is a realistic situation, because the lower end of background intensities and the upper end of saturation intensities are not necessarily symmetric. In this case, data was generated based on Model (1) with \( \alpha = 3000, \beta = 10000 \). But we took \( \gamma = 0.7 \) for the positive concentration measures (on the log2 scale, as always) and \( \gamma = 2.1 \) otherwise. We also used different noise distributions for different concentration regions. For positive concentration values, the data were generated subject to random noise from the \( t \)-distribution with three degrees of freedom and a scale parameter \( \sigma = 6000 \) (true concentration + 5) \(^{-1} \); otherwise, the random noise was generated from the normal distribution with mean 0 and variance 600\(^2\). The random noise is heterogeneous and decreases with the increasing positive true concentrations.

The scatter plot of one data set (against the true concentration levels) is exhibited in the upper left panel of Figure 1. The logistic curve (dashed curve) failed to fit the data on either tail due to its built-in symmetry on the two tails. However, the non-parametric curve we obtained (dotted curve) matches the true curve well.

The performance of the two approaches can be directly assessed by comparing the estimated \( EC50 \)s against the true \( EC50 \)s. The straight lines in the other three panels of Figure 1 correspond to the 45 degree line \( y = x \). The estimates obtained by the logistic fit (upper right panel) deviate substantially from the true \( EC50 \)s at both the lower and the upper tails; several samples showing fairly large differences in true concentrations were given similar estimates around five. This implies that some interesting differential expression patterns among the biological samples are missed by the use of logistic curves in the quantification method. In contrast, the non-parametric approach (lower left panel) shows much better agreement with the true \( EC50 \)s, except for one sample at the lower end of the intensity scale. Variations at the lower tail where the data are dominated by background noise are of less biological importance in the studies.

An important application of protein lysate arrays is to compare the protein content in different biological samples/agents on one slide. So it is important to examine if a quantification method can preserve the relative differences of
protein content among different samples. We conducted 50 simulation trials, in each of which the sample with the median EC50 was used as the baseline. Then for any other sample, we computed the difference $\hat{D}$ between its estimated EC50 and the estimated EC50 value for the reference sample to see how close it is to the true relative difference $D$. The sum of $|\hat{D} - D|$ over all samples, denoted by $D_{\text{tot}}$, is used as an overall measurement for the quantification method. The smaller $D_{\text{tot}}$ is, the better.
The box-and-whisker plot of the distribution of the measurements \( D_{\text{tot}} \) over 50 trials in this case were shown in the lower right panel of Figure 1, for both the logistic approach and the non-parametric approach. It turned out that the non-parametric approach did better in all 50 trials, with an average reduction of nearly 50% in \( D_{\text{tot}} \).

In the second case, we generated the expression data from the model:

\[
y_{ij} = \left( \alpha + \beta \frac{2^{x_{ij} + \text{EC50}}} {1 + 2^{x_{ij} + \text{EC50}}} \right) \times \epsilon_{ij}.
\]

A multiplicative random noise \( \epsilon \) with a log-normal \( \epsilon \sim N(0, \sigma^2) \) distribution was incorporated in this model instead of an additive normal random noise in Model 1. We used \( \alpha = 1000, \beta = 4000, \gamma = 0.8 \) and \( \sigma = 0.2 \) in the simulation. In the upper left panel of Figure 2, the dots in gray show the intensity data versus the true concentration in one data set. The estimated response curve by the non-parametric approach (dotted line) matched the true curve (solid line) well, while the logistic curve (dashed line) showed poor fit, especially at the upper tail. It fails to reflect the curvature shape of the true model by placing all the data in the linear range of the logistic curve. It is partly because the rigid parametric structure of the logistic model led to a sacrifice in the upper tail in order to obtain relatively good fit of the lower tail. This can be seen more clearly from the comparison between the estimated EC50s and the true EC50s (the middle left panel). The bias gets larger as the true concentration increases in the upper tail. In fact, the logistic approach yielded similar estimates for several samples with large relative differences in true concentration.

When the number of replicates increases, the variability of the estimates drops, but the bias due to mid-specified models would not diminish, so the advantage in bias for our proposed method becomes more evident. The second column of Figure 2 corresponding to the case with 6 replicates for each sample illustrated this point.

In the third case, we considered a piece-wise linear model, consisting of a constant value \( a \) for the response curve at negative concentration levels (on the log2 scale) and a straight line with the intercept \( a \) and slope \( b \) for positive concentrations. This type of curve is often observed in applications where saturation is not attained. An additive random noise from \( N(0, \sigma^2) \) was introduced. We used \( a = 3000, b = 500 \) and \( \sigma = 1000 \) in the simulation. It was shown in the top panel of Figure 3 that logistic approach produced artificial saturation at the upper tail, while the non-parametric approach provides a good fit to the data.

In the fourth case, we are interested in assessing the performance of non-parametric approach when the true response curve is logistic. In the simulation, we used Model (1) with \( \alpha = 1000, \beta = 4000, \gamma = 0.8 \) and \( \sigma = 500 \). The result as shown in the bottom panel of Figure 3 indicates a competitive performance of the non-parametric approach relative to the logistic approach when the data are in favor of the latter.

The simulation studies demonstrated that the non-parametric quantification approach produces consistent and robust relative protein concentration estimates, while the success of the logistic curves depends on the underlying relationship. When a rigid parametric model is mis-specified, bias becomes a serious concern with the parametric model. We now consider whether this concern is practically important.

## 4 ANALYSIS OF TWO LYSATE ARRAYS

We used two lysate arrays produced at M.D. Anderson for demonstration in this article. In this experiment, cell lysates were diluted at eight levels and spotted on an array in duplicate. The layout of the array consisted of a \( 4 \times 10 \) grid with each grid composed of a \( 4 \times 4 \) subgrid. Each subgrid contained the duplicate dilution series for a single sample. Each of the identical top two rows of a subgrid contained the four most concentrated dilution steps (in decreasing concentrations from left to right), and the identical bottom two rows contained the four least concentrated dilution steps. The upper panel in Figure 4 shows the spot intensity array image for protein AKT in 40 cell lines. The zoned-in subgrids for two arbitrarily selected samples are shown in the lower panel.

AKT has emerged as a critical enzyme in signal transduction pathways involved in cell proliferation, apoptosis, angiogenesis and diabetes (Dudek et al., 1997). For example, 40% of the patients with ovarian cancer show high levels of active AKT in tumor. KRT17 is involved in nail bed, hair follicle, sebaceous glands and other epidermal appendages. Abnormal activities of KRT17 lead to Jackson–Lawler type pachyonychia congenita and steatoctystoma multiplex (Schweizer et al., 2006).

We applied both the logistic (using the package ‘SuperCurve’) and the non-parametric approaches to these two lysate arrays. For the AKT array (shown in the left column of Figure 5, both curves seemed to fit the data reasonably well, and there is high concordance between these two approaches (left bottom panel).

The lysate array of protein KRT17 tells a different story, and the estimates obtained from the two approaches are quite different; see the right column of Figure 5. The estimates of protein concentrations in the lower tail is not critical because they fall into the region dominated by background intensities. However, in the upper tail, the logistic curve did not fit the intensity data well (right top panel), and as a result, the estimated EC50 for those samples were higher from the logistic approach than from the non-parametric approach.

To look into the differences, we now take a dilution series with high EC50 estimate. Table 1 lists the intensity measures for sample ref and sample ext with 2 replicates each. The quantity of interest is the relative difference between these two dilution series. The estimate based on SuperCurve is higher than that from the non-parametric approach by almost 1 unit on the log2 scale. From Table 1, we noticed that there is a single outlier in one dilution series (i.e. value 9723 in sample_{ext,rep2}) in the expression data. This single outlier lifted the EC50 estimate of the series when Super-Curve was used. If this outlier was dropped, the EC50 estimate from SuperCurve was dropped by about 0.5, whereas the non-parametric estimate was changed only by about 0.15. It demonstrates that the non-parametric approach based on the median is much more robust to outliers in the intensity measures.

When we chose the log-transformation on the expression data before using either the logistic or non-parametric models,
Fig. 2. Results for simulation study 2. The left columns correspond to 2 replicates per sample and the right columns 6 replicates. In each column, the top panel shows the fitted logistic curve (dashed curve) and non-parametric approach (dotted curve in red). The solid curve is the true response curve. The middle panel shows the EC50 estimates from the logistic approach versus the true EC50 values. The bottom panel shows the EC50 estimates from the non-parametric approach. The solid line (black) is $y = x$. Colour version of this figure is available as Supplementary material online.
we obtained a new set of estimates for EC50 for all samples. The correlation coefficient between the two sets of EC50 estimates (with y as intensity or log-intensity) is 0.86 for the logistic approach but 0.94 for the non-parametric approach, confirming better stability of the non-parametric approach against data transformations. Our experience shows that the use of neither the intensity nor the log intensity as y is better for all samples in lysate array studies.

From these two lysate array examples, we have noticed that the rigid logistic curves do not always fit the data well, and the non-parametric approach is valuable for robust quantification.

5 DISCUSSION

In this article, we proposed a non-parametric approach to quantify the protein expression intensities. This procedure aims at removing estimation bias in a parametric approach. Biased protein quantification results put scientists at the risk of drawing false biological conclusions. For example, in comparing the responses of treated and untreated cell lines to a certain therapeutic agent, the logistic approach may give similar protein concentration estimates due to its rigid and global parametric form, even when the difference of responses between the treated and untreated cell lines can be clearly captured by the non-parametric approach. Therefore, the proposed approach can improve the quality of protein expression quantification that is an essential step in making lysate arrays a useful tool in biological and medical research.

Note that the proposed non-parametric approach also differs from the SuperCurve in the objective functions, as the former uses the L1 distance (absolute difference between the estimated and observed protein intensities) while the logistic modeling uses the L2 distance (squared difference). As described in Section 2.2, the L1 distance works better for nonadditive errors, whereas the L2 distance leads to the most efficient estimates for Gaussian errors.

We recommend the non-parametric approach given in the article as a general quantification method for protein lysate arrays to reduce model bias and to protect against outliers. If the number of replicates and the length of the dilution series are small, e.g. the total number of measurements that can be used to estimate EC50 for each sample is no more than 10, we suggest using the least squares method to estimate EC50 after the proposed L1 distance has been used to obtain the estimate of g.

A possible improvement to the proposed estimation method is to use weights in the objective function to reflect heteroscedasticity in the data so that the L1 norm is weighted...
Fig. 5. The left column corresponds to AKT and the right column corresponds to KRT17. In each column, the top panel shows the logistic fits (curve in black); The middle panel shows the non-parametric fits (curve in black); The data points are represented by the circled points, based on the estimated concentrations; The bottom panel shows the EC50 estimates obtained from the non-parametric approach versus those from the logistic fit. The solid line (black) is $y = x$. 
according to variability of the intensities around the median curve with the purpose of obtaining EC50 estimates with smaller variances. Specific implementation of this approach requires further work.

The accuracy of the estimated EC50 varies with multiple factors, including the quality of the dilution series, the number of replicates and the estimation method used. In order to make inference on the relative difference in protein content across samples, the SEs need to be estimated. With either the parametric approach (e.g. logistic curve) or the non-parametric approach, the analytic derivations of the SEs are difficult, especially given the possible dependence of measurements within dilution series and across replicates. In the SuperCurve package, a bootstrap method is used to obtain confidence intervals. The same approach can be used for the non-parametric approach proposed in this article. The need to accommodate heteroscedasticity of the data in inference will be investigated in future work.

ACKNOWLEDGEMENTS

This research work is partially supported by the Kleberg Center for Molecular Markers, P50CA83639 and U24CA126477 to G.B.M. and by the National Science Foundation Grant DMS-0604229 and DMS-0706818.

Conflict of Interest: none declared.

REFERENCES

Zha, H. et al. (2004) Similarities of prosurvival signals in Bel-2-positive and Bel-2-negative follicular lymphomas identified by reverse phase protein microarray. Lab. Invest., 84, 235–244.

Table 1. Observed intensities of two samples under discussion

<table>
<thead>
<tr>
<th>Sample_ext,rep1</th>
<th>Sample_ext,rep2</th>
<th>Sample_ref,rep1</th>
</tr>
</thead>
<tbody>
<tr>
<td>12345678</td>
<td>2152 2979 4336 6460</td>
<td>1529 1446 2083 2678 3801 6546 11036 17127</td>
</tr>
<tr>
<td>9723</td>
<td>4865 12780 18119</td>
<td>2449 2771 2895 3295 2152 2979</td>
</tr>
<tr>
<td>2745 2916 2954 3165 2901 3434 4975 7252</td>
<td>1812 2151 2567 3417</td>
<td>2745 2916 2954 3165 2901 3434 4975 7252</td>
</tr>
</tbody>
</table>

Downloaded from https://academic.oup.com/bioinformatics/article/23/15/1986/205819 by guest on 14 March 2019