BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 5 2009, pages 650–654
doi:10.1093/bioinformatics/btn663

Data and text mining

Serial dilution curve: a new method for analysis of reverse phase protein array data

Li Zhang[1,*], Qingyi Wei[2], Li Mao[3], Wenbin Liu[1,4], Gordon B. Mills[4] and Kevin Coombes[1]

[1]Department of Bioinformatics and Computational Biology, [2]Department of Epidemiology, [3]Department of Thoracic and Head and Neck Medical Oncology and [4]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler street, Unit 1410, Houston, TX 77030, USA

Received on April 25, 2008; revised on December 03, 2008; accepted on December 27, 2008

Advance Access publication January 28, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Reverse phase protein arrays (RPPAs) are a powerful high-throughput tool for measuring protein concentrations in a large number of samples. In RPPA technology, the original samples are often diluted successively multiple times, forming dilution series to extend the dynamic range of the measurements and to increase confidence in quantitation. An RPPA experiment is equivalent to running multiple ELISA assays concurrently except that there is usually no known protein concentration from which one can construct a standard response curve. Here, we describe a new method called ‘serial dilution curve for RPPA data analysis’. Compared with the existing methods, the new method has the advantage of using fewer parameters and offering a simple way of visualizing the raw data. We showed how the method can be used to examine data quality and to obtain robust quantification of protein concentrations.

Availability: A computer program in R for using serial dilution curve for RPPA data analysis is freely available at http://odin.mdacc.tmc.edu/~zhangli/RPPA.

Contact: lzhangli@mdanderson.org

1 INTRODUCTION

The reverse phase protein array (RPPA) is an emerging high-throughput technique in proteomics (for reviews, see Borrebaeck and Wingren, 2007; Charboneau *et al*., 2002; Lv and Liu, 2007; Poetz *et al*., 2005; Sheehan *et al*., 2005). This technology has been successfully applied in a number of basic and clinical studies (Amit *et al*., 2007; Aoki *et al*., 2007; Fan *et al*., 2007; Pluder *et al*., 2006; Sahin *et al*., 2007; Tibes *et al*., 2006; Yokoyama, *et al*., 2007). A single array slide can be used to measure hundreds of samples for a protein. The protein level across the slide is detected by binding of a highly specific and sensitive primary antibody followed by detection using amplification linked to fluorescence, dye deposition, near infrared or nanoshells. Because protein concentrations can vary over many orders of magnitude in patient or cell line samples, it is desirable to have accurate measurements of protein concentrations over a wide dynamic range. To extend the dynamic range of the measurements, each sample is diluted multiple times successively and spotted on an RPPA slide so that if a protein concentration in the original sample is close to saturation, the sample can still be measured at diluted spots.

Multiple methods are available for analysis of RPPA data (Hu *et al*., 2007; Kreutz *et al*., 2007; Mircean *et al*., 2005). Typically, the methods are based on modeling the response curve, which describes the relationship between the observed signal and the protein concentration. Mircean *et al.* (2005) realized that since it is the same protein being measured for all the samples spotted on an RPPA slide, the same response curve should be suitable for all these samples. Based on this assumption, Microean *et al.* proposed a robust linear-square method to quantify the protein levels. However, an obvious drawback of the method is that it fails to recognize saturation effects for proteins at high levels. Recently, Hu *et al.* (2007) developed an alternative method using a non-linear, non-parametric approach to model the response curve.

In this study, we show an alternative approach to RPPA data analysis. Instead of modeling the response curve, we construct a new model, serial dilution curve, which characterizes the relationship between signals in successive dilution steps. The advantage of this approach is two fold: (i) the signals in successive dilutions can be related to each other in explicit formula in which the underlying unknown protein concentrations do not appear. This allows a low-dimensional non-linear optimization to estimate the key parameters of the map between protein concentration and signal intensity. The estimated map can then be applied to the observed signals to estimate the underlying abundances; (ii) it leads to an intuitive display of raw data, which is very useful for checking data quality and interpreting the model.

2 METHODS

2.1 Serial dilution curve

Our new method is based on the recognition that the relationship between signals in successive dilution steps uniquely determines the response curve. Typically, a response curve is a monotonic, s-shaped curve. It can be described by the Sips model (Sips, 1948):

$$S = a + bx^{\gamma}/[1 + x^{\gamma}/(M - a)] \tag{1}$$

where a is the background noise; b is the response rate in the linear range; M is the maximum or saturation level, x is the concentration of the protein. Sips model has been widely used to describe adsorption including binding

[*]To whom correspondence should be addressed.

of DNA (Glazer *et al*., 2006) and proteins on solid surface (Vijayendran and Leckband, 2001). Generally, $\gamma \neq 1$ applies to conditions in which the free energy of binding of the solute molecules can take a range of values instead of a unique value (Sips, 1948), i.e. there is some hereterogeneity in the solute molecules or the surface receptors. When the range of the free energy of binding shrinks to a singular point, γ approaches to 1, in which case it is equivalent to the conventional Langmiur model. With RPPA technology, one can only determine the relative protein concentration. Thus, x can be chosen on an arbitrary scale. For simplicity, we set x on a scale (i.e. a physical unit of x) so that $b=1$. Thus,

$$S=a+x^{\gamma}/[1+x^{\gamma}/(M-a)] \tag{2}$$

On this scale, protein concentration equals the background subtracted signal $(S-a)$ when $\gamma=1$ and saturation effect can be ignored.

Starting from Equation (2), we can see that if the protein concentration is diluted from x to x/d^k at the k-th dilution step, where $d>1$, the expected signal would be:

$$S_k=a+(x/d^k)^{\gamma}/[1+(x/d^k)^{\gamma}/(M-a)] \tag{3}$$

Combine the cases for S_{k+1} and S_k and eliminate x, we have:

$$S_k=a+d^{\gamma}(S_{k+1}-a)/[1+(d^{\gamma}-1)(S_{k+1}-a)/(M-a)] \tag{4}$$

Equation (3) describes S_k as a function of S_{k+1}, which we call the serial dilution curve, with three unknown parameters: a, M and γ (d is known). These parameters have graphical interpretations from the plot. As shown in Figure 1, the curve has two intersection points with identity line: one at background level, $S_k=S_{k+1}=a$, the other at the saturation level, $S_k=S_{k+1}=M$. At the left side in Figure 1, the saturation effect is of no concern and the relationship between S_k and S_{k+1} is approximately linear,

$$S_k-a\approx d^{\gamma}(S_{k+1}-a). \tag{5}$$

Thus, d^{γ} corresponds to the slope in the linear range in the serial dilution plot.

Equation (4) suggests a new model for displaying and analyzing RPPA data. It is important to note that Equation (4) does not contain protein concentration. Thus, it permits an appealing way of displaying the raw data without model specification or parameterization. Based on the plot like Figure 1, we can infer the parameters (a, M and γ) from the graph or through model fitting without knowing the protein concentrations in the samples. Model fitting with Equation (4) is relatively simpler than that with model fitting with Equation (2), which involves much more unknown parameters as in the existing methods of RPPA data analysis. Altogether, the number of unknown parameters in the model with Equation (2) is three plus the number of protein samples (each dilution series count as one sample), which can be in the hundreds. In contrast, Equation (4) only involves three unknown parameters.

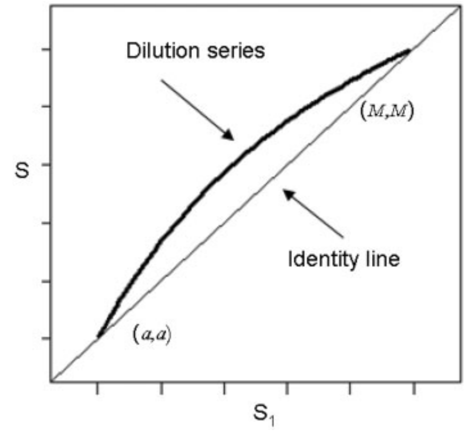

Fig. 1. Serial dilution plot. Each point in the serial dilution plot is composed of an observed signal S_k at dilution step k (on x-axis) and a corresponding signal S_k+1 of the same sample at the dilution step $k+1$ (on y-axis). The curve was produced using Equation (3). The curve has two intersection points with the identity line: (a,a) and (M,M).

2.2 Parameterization of the serial dilution curve

To find the optimal parameters, we used a weighted non-linear regression model using Equation (4) as the model and taking $a, D=d^{\gamma}, M$ as parameters. We assumed the observed signals have multiplicative errors except for the signals close to zero. The weight used in the regression model is $1/(m+|S|)$, where $m=5$, which is taken as the minimal error from signal quantification from the scanner used to obtain RPPA data. The starting values of a, D and M were taken to be max(m, min(S)), d, max(S), respectively. The *nls* function implemented in R-language (Ihaka and Gentleman, 1996) was used to optimize the parameters. The m is set to be the lower bound of a.

2.3 Estimating protein concentrations

Given the parameters in Equation (4) and signals of a dilution series of a particular sample (let these be $S_0, S_1, S_2, \ldots, S_K$), to obtain protein concentration $\hat{x}$ in the original undiluted sample, we used the following procedure. First, if all these signals are greater than M/r, the protein concentration $\hat{x}$ is marked to be saturated.

This threshold value of M/r is set according to an approximate estimate of the 95% confidence interval (CI) of the signals at the saturated spots. Under multiplicative error model, assume that the error rate of the observed signals is $\varepsilon=10\%$, and the saturation level is M, we expect the CI to be $[M/(1+2\times\varepsilon),\ M(1+2\times\varepsilon)]$ = $[M/1.2, 1.2M]$. Similarly, at background level a, we expect the 95% CI to be $[a/(1+2\times\varepsilon),\ a(1+2\times\varepsilon)]$ = $[a/1.2, 1.2a]$. In general, r should be >1 and can be reduced if precision of signals is improved.

If all the signals except one are $>M/r$ and the exception is not S_K, $\hat{x}$ is also marked to be saturated. Similarly, if all the signals are $<ar$, $\hat{x}$ is marked to be undetected. If all of them except one are $>M/r$ and the exception is not S_0, $\hat{x}$ is also marked to be undetected. The minimum and maximum of $\hat{x}$ are set to be

$$x_{\min}=[1/(ar-a)-1/(M-a)]^{-1/\gamma} \text{ and} \tag{6}$$

$$x_{\max}=[1/(M/r-a)-1/(M-a)]^{-K/\gamma}, \tag{7}$$

respectively. The above steps were taken to stabilize the protein concentration estimates for out of linear-range measurements.

If $\hat{x}$ is not marked saturated or undetected, we proceed to make an estimate of $\hat{x}$. We choose to remove signals $>M/r$ or $<ar$. Then, we convert each of the remaining signals S_j to x_j as

$$x_j=d^j[1/(S_j-a)-1/(M-a)]^{-1/\gamma} \tag{8}$$

where j denotes the j-th dilution step. To remove outliers among x_js, we identify an outlier among x_js as

$$|x_j-\text{ median}(x)|>3\times\text{ mad}(x)$$

where mad(x) is the median absolute deviation of x. Here, x is the vector of all x_js. Note that the outliers can also be identified from the serial dilution plot as points far away from the dilution curve (e.g. Fig. 3A).

Finally, we give the estimate of the dilution series as a weighted average of x_js:

$$\hat{x}=\frac{\Sigma(x_j w_j)}{\Sigma w_j} \tag{9}$$

where

$$w_j=\frac{1}{\left(\frac{\partial x_j}{\partial a}\Delta a\right)^2+\left(\frac{\partial x_j}{\partial M}\Delta M\right)^2+\left(\frac{\partial x_j}{\partial \gamma}\Delta\gamma\right)^2} \tag{10}$$

the partial derivatives are derived and computed according to Equation (8); Δa, ΔM *and* $\Delta\gamma$ are standard deviations of a,M, γ, respectively, which are obtained from the *nls* function in R. The estimated error of $\hat{x}$ is obtained from $(\Sigma w_j)^{-1/2}$.

3 RESULTS

To test the utility of the serial dilution curve for analyzing RPPA data, we first applied the method to simulated data, which was composed according to the Sips model [See Equation (2) in Section 2], with background level $a = 100$, saturation level $M = 50\,000$ and $\gamma = 1$, dilution factor $d = 2$. We added multiplicative noise (error rate = 0.15) to nominal signals and generated data as shown in Figure 2A. The multiplicative error model has been previously suggested (Kreutz *et al*., 2007). The samples were diluted to 1/2, 1/4 and 1/8 of their original concentrations serially. Figure 2B shows the serial dilution plot, which contains all data in the dilution series. Each point in the serial dilution plot is composed of an observed signal at dilution step k (on y-axis) and a corresponding signal of the same sample at the dilution step $k+1$ (on x-axis).

We found that our algorithm was able to recover the 'true' parameters from the simulated signals accurately. The values of a, M and γ were found to be 98 ± 5, $49\,800 \pm 520$, 1.05 ± 0.01, respectively. The estimated protein concentrations are also accurate (Figure 2C), except for the cases which are clearly out of the linear range. The lower and the upper bound of the range were calculated using Equations (6) and (7) and shown as dashed lines in Figure 2C. Note that setting the lower and upper bound helps to stabilize the estimates of protein concentration on logarithm scale, so that small changes in observed signals do not incur large changes in the estimates. Compare Figure 2A and C, one can also see that the linear range is much wider in the latter, showing that the dilution series can greatly expand the linear response range of the measurements.

We have also tested our algorithm with experimental data. Figure 3 shows a typical example of RPPA dataset. The experimental methods used to produce the array data were described by Fan *et al*. (2007). From the serial dilution plot (Fig. 3A), we notice many outliers (marked by red plus signs) near both x- and y-axis. Inspection of the original scanned image revealed that these outliers

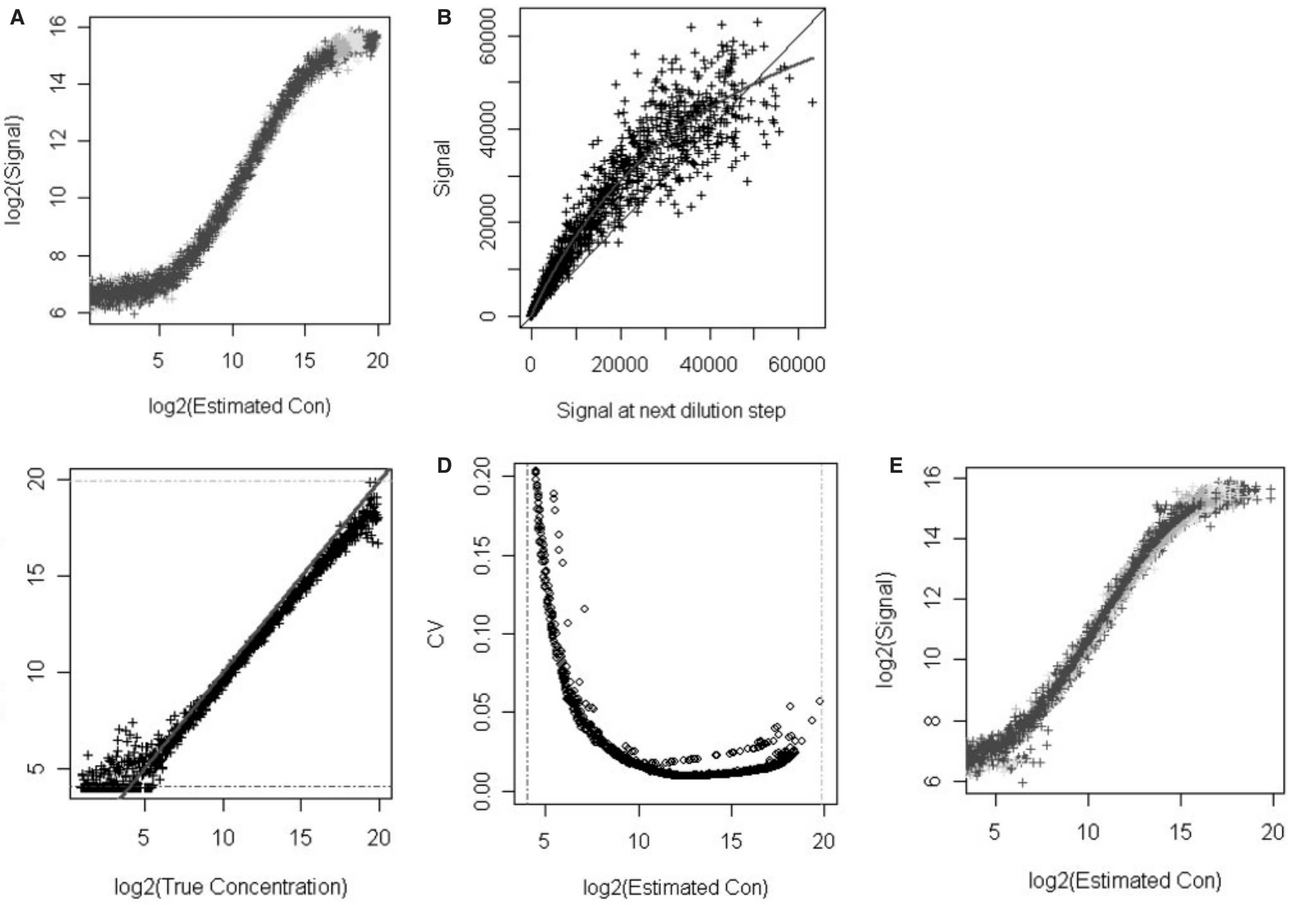

Fig. 2. Computer simulations. (**A**) Computer generated data with serial dilutions. Red, yellow, green, blue represent undiluted concentrations, 1/2, 1/4, 1/8 original concentrations, respectively. (**B**) Serial dilution plot. The blue line shows the estimated serial dilution curve. (**C**) The estimated versus the 'true' concentrations. The dashed lines show the upper (shown in green) and lower (shown in blue) bounds of the estimated concentrations according to Equations (5) and (6). The red line shows the identity lines. (**D**) Estimated error rates. CV = estimated error/estimated concentration. (**E**) Signal versus estimated concentrations. Red, yellow, green, blue represent undiluted concentrations, 1/2, 1/4, 1/8 original concentrations, respectively.

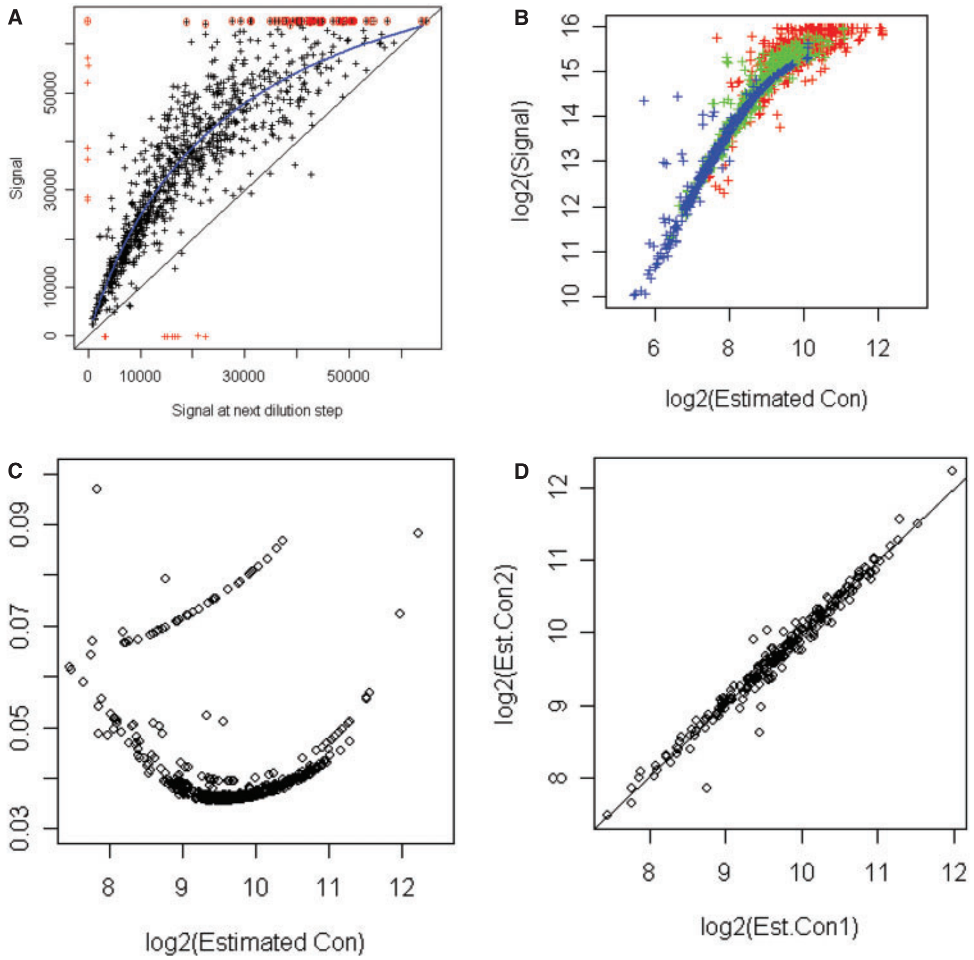

Fig. 3. Example of a practical dataset. The measured protein is beta actin, which serves as a control standard for measurements. (**A**) Serial dilution plot. Points shown in red were regarded as outliers or saturated (circled). (**B**) Signal versus estimated concentration. The signals of undiluted samples are shown in red, 1/2 diluted samples in green and 1/4 diluted samples in blue. (**C**) Estimated error rates. CV = estimated error/estimated concentration. Each point represents result from one serial dilution. (**D**) Estimated protein concentrations from replicated dilution series of the same samples.

were produced by a faulty background subtraction method that extracted signals from the scanned image. The image quantification method took median pixel intensities from local regions outside the spotted area as the background level. However, occasionally the protein samples seemed to spill over the spotted area, which caused grossly overestimated background levels, which in turn led to grossly underestimated signals.

Figure 3A also showed that all the signals are bounded below 65 000 (the points close to the upper bound are marked by the red circles). This was caused by imaging software that set the maximum pixel intensity to be 65 536. Thus, the real signals must have been truncated for these spots. We therefore removed the points shown in red in Figure 3A before fitting the serial dilution curve. The estimated parameters are $a = 5$, $M = 63\,602$, $\gamma = 0.57$. The estimated protein concentrations were shown in Figure 3B.

Sometimes RRPA experiment may fail to yield meaningful measurements of proteins. In Figure 4, we show an example that has quality problems. The experimental methods used to produce the array data was described by Tibes *et al.* (2006). Using methods as described in Section 2.2, the background was estimated to be 1000, saturation level: 4751, dilution factor: 1.11. The black line is the identity line and the blue line is the serial dilution curve. The serial dilution curve (blue) is very close to the identity line (black), indicating that after dilution, the signals tend to stay at the same levels as before. This implies that the dilution had failed to produce the expected reduction of signals. The exact cause of this effect is unclear. From our observations, such pattern often occurs in the slides that have faint signals. Furthermore, because the serial dilution curve is approximately linear, the saturation level cannot be accurately determined.

To evaluate data quality on an array, we find the following two measures to be most important according to our empirical experience.

(i) V1 = Percentage of data points in linear range (as defined by the interval $[ar, M/r]$) of all data points on the array, where a is the background level, M is the saturation level, r is the threshold value (as described earlier). High V1 value indicates good quality of data. When V1 is low, the data points are out of the linear range, in which cases extra manipulation of protein

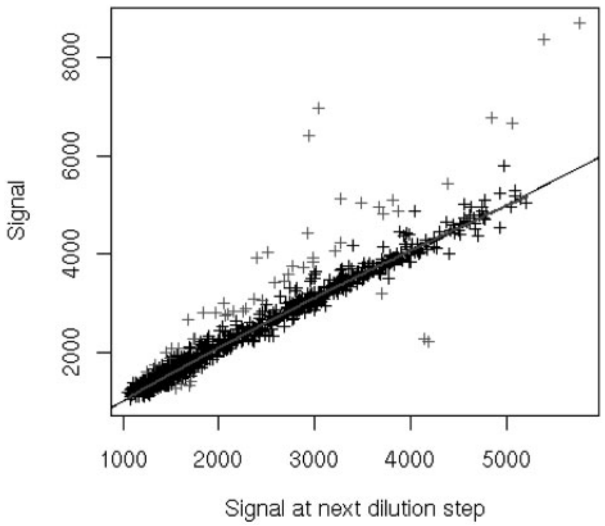

Fig. 4. Example of data with quality problems. This is a serial dilution plot. The measured protein is GAPDH. The red symbols show the outliers. The background is estimated to be 1000, saturation level: 4751, dilution factor: 1.11. The black line is the identity line and the blue line is the serial dilution curve.

concentration in the samples is needed prior to hybridization on arrays. Alternatively, the level of antibody can be adjusted so that more data points will may fall in the linear range. In addition, note that the distribution of the data points can also inform the significance of non-linear effects. When most data points are far below the saturation level, the serial dilution curve approaches a straight line, in which case the saturation level is uncertain (for example, see Fig. 4).

(ii) V2 = median CV on an array, where CV = estimated error/estimated protein concentration. V2 represents estimated error rate. High precision of protein concentration measurements is represented by low V2 values.

4 DISCUSSION

Graphical display of data plays a very important role in data analysis. For RPPA data, it is conventional to plot the observed signals against the estimated protein concentrations. However, because the estimated protein concentrations depend on the models as well as the estimated parameters, when the signals seem to fit poorly to the estimated concentrations, it is not clear whether it is due to a suboptimal model or to noisy data. Making the serial dilution plot *per se* requires no model selection or parameter fitting. The plot presents the entire set of observables on an array in their original values. From the plot one can identify the background level, saturation level, which signals are in the linear range, and which signals are outliers (as in Fig. 3A). Fitting a serial dilution curve needs only three parameters, which is much simpler than fitting the response curve, which requires estimating the protein concentrations as additional parameters.

From simulated RPPA data, we showed that our algorithm can yield robust and accurate estimates of protein concentrations. From practical RPPA data, we saw some of the data points did not follow the serial dilution curve. There may be multiple causes of the abnormal points, such as saturation or failure of binding. It should be noted that the response curve in RPPA technology is sensitive to a large number of factors, including the amount and duration of sample incubation, specific and non-specific interactions of reporter molecules and surface chemistry in the microarrays (Seurynck-Servoss *et al.*, 2007). These factors complicate the interpretation of RPPA data. Non-parametric models (Hu *et al*, 2007) take fewer assumptions about the hybridization kinetics in RPPA technology. Hence, the non-parametric models are more flexible, and in some cases they may fit better with observed RPPA data. The disadvantage of non-parametric models is that the parameters are less interpretable, while the parameters in Sips model are physically meaningful and can be used to optimize the conditions for RPPA experiments. We believe the method developed in this study will have broad utility in RRPA applications.

Funding: M. D. Anderson Cancer Center start-up fund; MDACC Institutional Research Grant (to L.Z.).

Conflict of Interest: none declared.

REFERENCES

Amit,I. *et al*. (2007) A module of negative feedback regulators defines growth factor signaling. *Nat. Genet.*, **39**, 503–512.

Aoki,H. *et al*. (2007) Telomere $3'$ overhang-specific DNA oligonucleotides induce autophagy in malignant glioma cells. *Faseb. J.*, **21**, 2918–2930.

Borrebaeck,C.A. and Wingren,C. (2007) High-throughput proteomics using antibody microarrays: an update. *Expert. Rev. Mol. Diagn.*, **7**, 673–686.

Charboneau,L. *et al*. (2002) Utility of reverse phase protein arrays: applications to signalling pathways and human body arrays. *Brief Funct. Genomic. Proteomic.*, **1**, 305–315.

Fan,Y.H. *et al*. (2007) In vitro expression levels of cell-cycle checkpoint proteins are associated with cellular DNA repair capacity in peripheral blood lymphocytes: a multivariate analysis. *J. Proteome Res.*, **6**, 1560–1567.

Glazer,M. *et al*. (2006) Kinetics of oligonucleotide hybridization to photolithographically patterned DNA arrays. *Anal. Biochem.*, **358**, 225–238.

Hu,J. *et al*. (2007) Non-parametric quantification of protein lysate arrays. *Bioinformatics*, **23**, 1986–1994.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Kreutz,C. *et al*. (2007) An error model for protein quantification. *Bioinformatics*, **23**, 2747–2753.

Lv,L.L. and Liu,B.C. (2007) High-throughput antibody microarrays for quantitative proteomic analysis. *Expert Rev. Proteomics*, **4**, 505–513.

Mircean,C. *et al*. (2005) Robust estimation of protein expression ratios with lysate microarray technology. *Bioinformatics*, **21**, 1935–1942.

Pluder,F. *et al*. (2006) Proteome analysis to study signal transduction of G protein-coupled receptors. *Pharmacol. Ther.*, **112**, 1–11.

Poetz,O. *et al*. (2005) Protein microarrays: catching the proteome. *Mech. Ageing Dev.*, **126**, 161–170.

Sahin,O. *et al*. (2007) Combinatorial RNAi for quantitative protein network analysis. *Proc. Natl Acad. Sci. USA*, **104**, 6579–6584.

Seurynck-Servoss,S.L. *et al*. (2007) Evaluation of surface chemistries for antibody microarrays. *Anal. Biochem.*, **371**, 105–115.

Sheehan,K.M. *et al*. (2005) Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. *Mol. Cell Proteomics*, **4**, 346–355.

Sips,R. (1948) On the structure of a catalyst surface. *J. Chem. Phys.*, **16**, 490–495.

Tibes,R. *et al*. (2006) Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol. Cancer Ther.*, **5**, 2512–2521.

Vijayendran,R.A. and Leckband,D.E. (2001) A quantitative assessment of heterogeneity for surface-immobilized proteins. *Anal. Chem.*, **73**, 471–480.

Yokoyama,T. *et al*. (2007) Roles of mTOR and STAT3 in autophagy induced by telomere $3'$ overhang-specific DNA oligonucleotides. *Autophagy*, **3**, 496–498.