Modified SuperCurve Method for Analysis of Reverse-Phase Protein Array Data

MIAO SUN¹, DEJIAN LAI¹, LI ZHANG² and XUELIN HUANG²

ABSTRACT

Reverse-phase protein arrays (RPPAs) are widely used in biological and biomedical fields of study. One of the most popular analytic methods in RPPA data analysis is the SuperCurve method, which requires estimation of the background fluorescence level. This estimation is usually not accurate and has sample bias and spatial bias. Here, we propose a *taking-the-difference* method to overcome this problem. Briefly, for each two consecutive RPPA cycles, we subtract the later cycle from the earlier cycle, transforming the m -cycle data into $m - 1$ cycle of data. This removes most of the background fluorescence noise. We then use the $m - 1$ cycle of data to fit a new model accordingly derived from the SuperCurve model. To evaluate our proposed method, we compare the accuracy and precision between our proposed model and the original SuperCurve model by testing them on both real and simulated datasets. For both situations, our modified model shows improved results. The modified SuperCurve method is easy to perform and the *taking-the-difference* idea is recommended for application to all current methods of RPPA data analysis.

Key words: background subtraction, regression model, reverse-phase protein array, SuperCurve model.

1. INTRODUCTION

THE REVERSE-PHASE PROTEIN ARRAY (RPPA) is commonly used as an analytical method in proteomics and has been successfully applied in capturing disease progression (Paweletz et al., 2001), signal pathway detection (Charboneau et al., 2002), analysis of proteomic profiling (Nishizuka et al., 2003), biomarker discovery in drug development (Wilson and Nock, 2003), and in other types of studies. RPPA provides several advantages. First, it does not require direct labeling on the sample or two-site antibody. Therefore, there will be no experimental variability due to labeling yield (Liotta et al., 2003). Second, with RPPA, all samples are spotted at the same time, which enables this method to represent the underlying disease signal transactions for the whole cell (Tabus et al., 2006). Third, as each sample is exposed to the same concentrations of primary and secondary antibodies for the same period of time, subtle differences in analytes can be captured (Liotta et al., 2003). In addition, each protein sample in the reverse-phase array has a calibration curve. This provides a better means for matching the antibody concentration and the analyte concentration (Sheehan et al., 2005).

¹Division of Biostatistics, The University of Texas School of Public Health, Houston, Texas.

²Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas.

Multiple methods have been applied for the analysis of RPPA data. The sigmoidal shape has been chosen as it is consistent with the expected relationship of observed signal and true protein concentration (Tabus et al., 2006). This method has been developed as the “SuperCurve” method and is publicly available online through the Department of Bioinformatics and Computational Biology at the University of Texas MD Anderson Cancer Center. Hu et al. (2007) proposed a more flexible nonparametric joint sample model for RPPA analysis. Troncale et al. (2012) proposed a new method, called NormaCurve, which allows for the normalization of RPPA data. All the above analytical methods require the step of background correction. However, there are problems caused by background corrections (Rao et al., 2013). First, incorrect background subtraction leads to miscalculation of protein efficiency (Shain and Clemens, 2008). Second, background noise varies when different models are applied (Guescini et al., 2008). Third, all the methods assume a constant background for all samples within one RPPA experiment; however, this assumption usually does not hold in reality. Different areas of an RPPA array may have different background fluorescence levels.

In this study, we modify the SuperCurve method to minimize the influence introduced by background noise. The idea is that the data are in a time series structure, and differentiating among the data is an approach that is commonly used for this kind of data (Diggle, 1990). For each sample, we subtract each later dilution cycle from its previous dilution cycle, transforming the m -cycle data into $m-1$ cycles. By subtracting, most of the background noise is removed. The SuperCurve model for m -cycle data implies a model for the $m-1$ cycle of data. We simplify the model by using Taylor’s expansion so that linear regression techniques can be applied to fit the model. We demonstrate through real and simulated datasets that the proposed method improves data analysis results.

2. METHODS

SuperCurve is a logistic model with three global parameters (Hu et al., 2007).

$$Y_{ij} = \alpha + \beta \frac{e^{[\gamma(x_{ij} + u_j)]}}{1 + e^{[\gamma(x_{ij} + u_j)]}} + \varepsilon_{ij} \quad (1)$$

For notation, Y_{ij} is the observed signal intensity for dilution i of sample j , where $i = 1, \dots, m$ and $j = 1, \dots, n$. Typically, $x_{ij} = (1 + m)/2 - i$ for all j . The median effective protein concentration is u_j , which is the 50th percentile of the series protein dataset. It is the parameter of interest. The known dilution offset step is x_{ij} . The global parameters for the curve are α , β , and γ , where α represents the background fluorescence in the model, and the error term is $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Based on our idea, we build an $m-1$ cycle of data Z_{ij} from Y_{ij} so that we can remove α .

$$Z_{ij} = Y_{ij} - Y_{i+1,j} \quad (2)$$

RPPA data are serial diluted. The data in the later step should be theoretically smaller than the data in the previous step. Thus, the Z ’s should always be nonnegative. Negative Z ’s are not used in the data analysis below.

We then use the new $m-1$ cycle of data to fit the SuperCurve model. Denoting $x_{ij} + u_j = t$, we obtain the log form of Z_{ij} as

$$\log(Z_{ij}) = \log(\beta) + \log(1 - e^{-\gamma}) + \gamma t - \log(1 + e^{-\gamma t}) - \log(1 + e^{-\gamma(t+1)}) + \zeta_{ij} \quad (3)$$

By Taylor’s series, we obtain $\log(1 + e^{-\gamma t}) \approx \log(2) + \frac{\gamma t}{2} + \frac{(\gamma t)^2}{8} - \frac{(\gamma t)^4}{192}$. Then, Equation (3) becomes

$$\log(Z_{ij}) = k_1 + k_4 x_{ij}^4 + k_4(4\mu_j + 2)x_{ij}^3 + k_2 x_{ij}^2 + k x_{ij} + \zeta_{ij} \quad (4)$$

It is a linear regression function of $\log(Z_{ij})$ on x_{ij}^4 , x_{ij}^3 , x_{ij}^2 , and x_{ij} . We assume that θ_1 is the coefficient for x_{ij}^4 and θ_2 is the coefficient for x_{ij}^3 . Then, after calculations, we obtain u_j , as shown below:

$$u_j = \frac{\theta_2 - 2\theta_1}{4\theta_1} \quad (5)$$

We may conduct a separate linear regression for each sample. However, to borrow strength between samples, we use a linear mixed model (Stroup et al., 2012) to estimate the above parameters. By

implementing the mixed model, we obtain the median protein effect concentration for each sample. Also, we assume that there is a constant noise level on a Z scale. Then, by the delta method, the variance of the noise level at the $\log(Z)$ level is proportional to the inverse of Z^2 . To further improve our method, we incorporate a weight with $W_i = Z_i^2$ into the mixed model. Then, we can obtain u_j for each sample, which is the median effective concentration.

We apply the above modified SuperCurve method and the original SuperCurve method to simulated and real data sets. We used three criteria to compare them: relative error (RE), mean squared error (MSE), and coefficient of variation (CV), as defined below.

$$RE = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_0^*}{x_0} - 1 \right) \quad (6)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_0^* - x_0)^2 \quad (7)$$

where x_0^* is the estimated protein concentration effect and x_0 is the true protein concentration effect of each of the n samples. A model with smaller RE and MSE is more accurate. We also used the CV to test the precision of the methods, as shown below.

$$CV = \frac{s}{\bar{x}} \times 100\% \quad (8)$$

In Equation (8), s and $\bar{x}$ are the respective standard deviation and the mean of the protein concentrations. A smaller CV represents a more precise method. We used the CV to compare the precision of the models in the practical dataset and used RE, CV, and MSE to compare both precision and accuracy in the simulated datasets.

3. RESULTS

The simulated dataset we used was developed by Hu et al. (2007). The signals were generated based on the SuperCurve model in Equation (1) with $\beta = 10,000$ and $\gamma = 3$. A common set of true u_j was selected from a normal distribution, with mean 1.0 and standard deviation 1.5. There were 1000 samples and 5 dilution cycles for each sample. The background level α was generated from a normal distribution, with mean 4000 and standard deviation 1000. We compared the value of the RE, CV, and MSE between the result obtained from the SuperCurve model and that from the modified SuperCurve model as described in the Methods section.

Table 1 shows the comparison of the mean values over the 1000 samples of RE, CV, and MSE for the SuperCurve method and the modified SuperCurve model as applied to the simulated dataset. The RE obtained by the modified SuperCurve method was 1.16, which was smaller than that obtained by the SuperCurve method (6.33). The MSE obtained by our modified model was 4.07, which was smaller than the corresponding value obtained by the SuperCurve model (7.63). The absolute value of CV for our modified method was 0.56, which was much smaller than that obtained in the original SuperCurve model (12.55). The original SuperCurve method yielded 14 outlier results (estimated $\mu_j > 50$) when analyzing the simulated dataset. These outliers were not included in the calculation of the above RE, MSE, and CV. Our proposed method did not give any such outliers.

We used a real experimental dataset provided by Troncale et al. (2012). We selected 390 samples from the dataset, all of which have the same initial protein inputs (1 mg/ml) and fifteen 2-fold serial dilutions

TABLE 1. COMPARISON OF RELATIVE ERROR, COEFFICIENT OF VARIATION, AND MEAN SQUARED ERROR BETWEEN THE SUPERCURVE METHOD AND THE MODIFIED SUPERCURVE METHOD WHEN APPLIED TO THE SIMULATED DATASET

<i>Criteria</i>	<i>SuperCurve</i>	<i>Modified SuperCurve</i>
Relative error	6.33	1.16
Mean squared error	7.63	4.07
Coefficient of variation	12.55	−0.56

(1, $\frac{1}{2}$, $\frac{1}{4}$, ..., $1/16384$). For convenience, we used only the first five data dilution cycles in our data analysis. Most RPPA experiments use only five dilutions.

The samples in this dataset had the same initial protein inputs and dilution steps, and so it was convenient for us to compare the CVs using different algorithms. The absolute value of CV for the SuperCurve model was 17.47, which was much larger than that in our modified model (8.18), indicating that our modified model can improve precision compared to the original model.

4. DISCUSSION

This study describes the incorporation of the *taking-the-difference* strategy in the SuperCurve method for RPPA data analysis. The modified method is theoretically more advantageous as it does not require the step of background correction. We compared our modified method with the original method in terms of accuracy and precision. Applied to the real dataset, our modified method showed more precision than the original method. Applied to the simulated dataset, our modified method showed better results in terms of accuracy and precision, even after removing the outliers generated by the SuperCurve model.

The advantages of our model are the ability to avoid the background noise effect and to generate results with less variance after fitting the mixed model. This advantage is particularly important when analyzing datasets with relatively large variance and background noise levels.

The original SuperCurve method includes a process of removing outliers, whereas our modified method does not. The NormaCurve method introduced a helpful normalization process (Troncale et al., 2012). Further work may incorporate a process of diagnostics and data normalization in our modified method to further improve its accuracy and precision.

Overall, our modified SuperCurve method is valuable for RPPA data analysis. It is easy to perform and can avoid errors introduced by background subtraction. We recommend the application of this *taking-the-difference* approach in all current methods for RPPA data analysis.

ACKNOWLEDGMENTS

This research was supported by the U.S. National Institutes of Health Grants U54 CA096300, U01 CA152958, and 5P50 CA100632.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Charboneau, L., Scott, H., Chen, T., et al. 2002. Utility of reverse phase protein arrays: applications to signaling pathways and human body arrays. *Brief. Funct. Genomics Proteomics* 1, 305–315.
- Diggle, P.J. 1990. *Time Series: A Biostatistical Introduction*. Oxford University Press, Inc., New York; pp. 31–39.
- Guescini, M., Sisti, D., Rocchi, M.B., et al. 2008. A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. *BMC Bioinform.* 9, 326.
- Hu, J., He, X., Baggerly, K.A., et al. 2007. Non-parametric quantification of protein lysate arrays. *Bioinformatics* 23, 1986–1994.
- Liotta, L.A., Espina, V., Mehta, A.I., et al. 2003. Protein microarrays: meeting analytical challenges for clinical applications. *Cancer Cell* 3, 317–325.
- Nishizuka, S., Charboneau, L., Young, L., et al. 2003. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc. Natl. Acad. Sci. USA* 100, 14229–14234.
- Pawletz, C.P., Charboneau, L., Bichsel, V.E., et al. 2001. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* 20, 1981–1989.
- Rao, X., Lai, D., and Huang, X. 2013. A new method for quantitative real-time polymerase chain reaction data analysis. *J. Comput. Biol.*, 20, 703–711.

- Shain, E.B., and Clemens, J.M. 2008. A new method for robust quantitative and qualitative analysis of real-time PCR. *Nucleic Acids Res.* 36, e91.
- Sheehan, K.M., Calvert, V.S., Kay, E.W., 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.
- Stroup, W.W. 2012. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, Boca Raton, FL.
- Tabus, I., Hategan, A., Mircean, C., et al. 2006. Nonlinear modeling of protein expressions in protein arrays. *IEEE Trans. Signal Process.* 54, 2394–2407.
- Troncale, S., Barbet, A., Coulibaly, L., et al. 2012. NormaCurve: a SuperCurve-based method that simultaneously quantifies and normalizes reverse phase protein array data. *PloS One* 7, e38686.
- Wilson, D.S., and Nock, S. 2003. Recent developments in protein microarray technology. *Angew. Chem. Int. Ed.* 42, 494–500.

Address correspondence to:

Dr. Xuelin Huang

Department of Biostatistics, Unit 1411

The University of Texas MD Anderson Cancer Center

PO Box 301402

Houston, TX 77230-1402

E-mail: xluhuang@mdanderson.org