14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

CHOOSING THE DESIGN PARAMETERS FOR PROTEIN LYSATE ARRAYS

Andrea Hategan, Ioan Tabus, and Jaakko Astola

Institute of Signal Processing, Tampere University of Technology
P.O. Box 553, FIN-33101 Tampere, Finland
phone: +(358)331153974, fax: +(358)331153817, email: ioan.tabus@tut.fi, andrea.hategan@tut.fi

ABSTRACT

Protein lysate array is a new technology for measuring the relative expressions of proteins, where the array image provides information about the concentrations (expressions) of a given protein for tens of patients or tissues. The array consists of replicated and serially diluted versions of the protein concentration for the biological samples at several spots. When producing the lysate array the experimenter has to set several parameters, such as: the concentration of the sample solution to be printed at a certain spot, the number of dilutions, the number of replicates for each biological sample, and the dilution factor. Having the resulting image of intensities at all spots one can assume a nonlinear model and estimate the values of the relative protein expression levels for all biological samples. In this paper we study how the obtained model can be used to improve the design of the experiment, such that if a second lysate array will be produced, better design parameters will be selected. We propose a methodology for choosing the design parameters, and illustrate it with results for several lysate array data sets.

1. INTRODUCTION

The estimation of the ratios of protein expressions in different biological samples finds important applications to the molecular profiling of various diseases, allowing to observe how a protein is expressed in an diseased tissue versus a normal tissue, or a tissue that is under treatment versus a normal tissue.

The production of a lysate array follows a complicated protocol where the solution of protein at each spot reacts with a solution containing a specific antibody, marked with a dye. Many experimental design parameters have to be set [1, 2], resulting finally in an image where each spot has the average gray level essentially dependent of two factors: the protein concentration in the solution printed at that spot and the antibody concentration. A detailed estimation procedure for the relative protein expression level for two biological samples was introduced in [3]. The accuracy of the estimates is strongly influenced by the values chosen for the design parameters, e.g., number of replicates and of dilution points.

Out of the many factors involved in the preparation of a lysate array, we will consider in this study as design parameters the following quantities: the concentration of the solution printed at each spot, the number of dilution and replicates for each biological sample, and finally the dilution factor from one spot to the next for the sequentially diluted spots.

Since in many occasions the production of a lysate array is repeated if a poor quality is obtained after the first array production, it is important to derive a procedure by which the experimenter can choose new values for the design parameters when preparing the second lysate array, with the goal of improving the accuracy observed after the second experiment.

The paper is organized as follows: in the next section we describe the typical layout of a lysate array and the estimation procedure for nonlinear calibration curves, in the third section several problems relevant to the experimental design are presented, in the fourth section the procedure is illustrated with real data, and we end with conclusions.

2. A LYSATE ARRAY EXPERIMENT

2.1 Lysate array layout

A lysate array contains $N = p \times k \times r$ spots, where p is the number of biological samples, k is the number of dilutions and r is the number of replicates. For each biological sample, a solution that contains the protein of interest is prepared and this solution is serially diluted using a dilution factor of $b = 2$ and printed at successive spots. For each biological sample, $k = 6$ such dilutions are produced and each diluted solution is printed $r = 3$ times, resulting in $k \times r = 18$ spots for each biological sample. The solution concentration printed at each spot can be controlled by the number of touches, i.e. how many times the robot will touch each spot and it was observed that 5 touches yield the best overall quality. See [1, 2] for a detailed explanation of how a lysate array is produced.

2.2 Relative protein expression level: modelling and estimation

A nonlinear modeling procedure is presented in [3], where the parameters of a nonlinear calibration curve are estimated together with the protein concentrations for different biological samples. We review here only the main steps of the modeling procedure, and refer to [3] for the detailed description of the estimation algorithms.

Denote by c_j the concentration corresponding to the biological sample j, $j = 1, \ldots, p$, $\tilde{y}_i$ the measurement of the gray level intensity at spot i, where $i \in \{1, \ldots, N\}$, and $s_i \in \{1, \ldots, p\}$ the biological sample index whose solution is printed at spot i. It was found more convenient to model the dependency intensity - protein concentration in a log-log domain. We denote $y_i = \log_2 \tilde{y}_i$ the logarithm of the intensity at one spot and x_i the logarithm of the unknown protein concentration in the solution printed at the same spot i, which is given by

$$\begin{aligned} x_i &= \log_2 \frac{c_{s_i}}{2^{d_i}} \\ &= \log_2 c_{s_i} - d_i = q_{s_i} - d_i, \end{aligned} \tag{1}$$

where $d_i \in \{0, \ldots, k-1\}$ is the dilution number.

We note that the log-concentration q_{s_i} of any given biological sample will appear in the definition of abscissas x_i's at a number of $k \times r$ spots. Finally the base of the logarithm will be assumed to be always identical to b, and in this paper $b = 2$ except Section 3.2., where b is subject to optimization.

A conditional Gaussian distribution is assumed for the measured intensities $y_i|\theta \sim \mathscr{N}(g(x_i,\beta), \sigma_0^2 g(x_i,\beta)^{2\alpha})$, where $\theta = [q^T \beta^T \sigma_0 \alpha]^T$ is the vector of unknown parameters, $g(x_i,\beta)$ is the nonlinear model for the mean, also called a calibration curve (typically a sigmoidal or polynomial model), and $\sigma_0^2 g(x_i,\beta)^{2\alpha}$ is the heteroscedastic model for the variance. If we denote by $y^N = y_1,\ldots,y_N$ the string with all measured data and assume that the measurements are independent, the overall likelihood is given by:

$$L(y^N,\theta) \quad = \quad \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_0^2 g(x_i,\beta)^{2\alpha}}} e^{-\frac{(y_i - g(x_i,\beta))^2}{2\sigma_0^2 g(x_i,\beta)^{2\alpha}}} \tag{2}$$

Denote by $\hat{\theta} = [\hat{q}^T \hat{\beta}^T \hat{\sigma}_0 \hat{\alpha}]^T$ the ML estimates for the vector of parameters that maximizes the overall likelihood in (2), see [3]. Of interest are the ratios of the concentrations

$$\frac{\hat{c}_j}{\hat{c}_{ref}} = 2^{(\hat{q}_j - \hat{q}_{ref})} \tag{3}$$

between the concentration with current index, $j \in \{1,\ldots,p\}$, and that with a reference index, $ref \in \{1,\ldots,p\}$. The reference is either set according to the biological problem studied, or if there is no biological preference, it will be chosen such that the accuracy is maximized.

The accuracy of the estimation procedure has been evaluated in [3] by extensive Monte Carlo simulations. The variances of the estimated parameters obtained in the simulations have been compared with the Cramér-Rao lower bound and they were found to agree very well, validating the estimation procedure.

A measure of interest when evaluating the accuracy of the experiment is the sum of variances of the log-ratios,

$$\begin{aligned} \sum_{j=1}^{p} Var(\hat{q}_j - \hat{q}_{ref}) \quad &= \quad \sum_{j=1}^{p} (Var(\hat{q}_j) + Var(\hat{q}_{ref}) \\ &- \quad 2cov(\hat{q}_j, \hat{q}_{ref})) \end{aligned} \tag{4}$$

The variances involved in (4) can be assumed to reach the Cramér-Rao lower bound (as experimentally found in [3]) and thus, if we denote by $R(\bar{\theta})$ the inverse of the Fisher information matrix, the criterion in (4) can be conveniently approximated by:

$$J(\bar{q},\bar{\eta}) = \sum_{j=1}^{p} R_{j,j}(\bar{\theta}) + pR_{ref,ref}(\bar{\theta}) - 2\sum_{j=1}^{p} R_{j,ref}(\bar{\theta}) \tag{5}$$

where $\bar{\theta} = [(\bar{q})^T (\bar{\eta})^T]^T$ is the vector of the model parameters with which the data was generated and $\bar{\eta} = [(\bar{\beta})^T \bar{\sigma}_0 \bar{\alpha}]^T$.

3. CHANGING THE EXPERIMENT DESIGN PARAMETERS

In the following we assume that the experimenter intends to produce a new lysate array, and he is free to change some of the parameters of the design, in light of the results obtained in the first experiment. In the current technology some of the parameters may be easier to change than others, and for this reason we address separately the choice of each design parameter. The goal is to obtain after the second experiment a new set of parameter estimates $\hat{\theta}^{[2]} = [(\hat{q}^{[2]})^T (\hat{\beta}^{[2]})^T \hat{\sigma}_0^{[2]} \hat{\alpha}^{[2]}]^T$ with an improved total accuracy as measured by (5), or an average of it.

3.1 Changing the initial dilution for each sample

We focus first on the possibility that in the second experiment the experimenter changes the concentration of the biological sample j by an amount (in log-scale) of $\hat{\Delta}q_j$, such that the initial unknown concentration $\bar{q}_j^{[1]}$ becomes $\bar{q}_j^{[2]} = \bar{q}_j^{[1]} - \hat{\Delta}q_j$. The value of $\hat{\Delta}q_j$ needs to be determined only based on the available information, i.e., on the estimates $\hat{\theta}^{[1]} = [(\hat{q}^{[1]})^T (\hat{\beta}^{[1]})^T \hat{\sigma}_0^{[1]} \hat{\alpha}^{[1]}]^T = [(\hat{q}^{[1]})^T (\hat{\eta}^{[1]})^T]^T$.

Intuitively, if a concentration estimate is located on the saturated branches of a sigmoidal calibration curve, then it is tempting to dilute it in the second experiment in such a way that the 6 diluted versions are likely to fall around the center of symmetry of the sigmoid, where the sensitivity $\frac{dy}{dx}$ is the highest. Our goal is to find such a location of high sensitivity $q_j^{ideal}(\hat{\eta}^{[1]})$, which is a function of the estimated calibration curve and of the the estimated parameters for the variance model. So, once we have an estimate $\hat{q}_j^{[1]}$, the additional dilution $\hat{\Delta}q_j$ will be decided to be

$$\hat{\Delta}q_j = \hat{q}_j^{[1]} - q_j^{ideal}(\hat{\eta}^{[1]}). \tag{6}$$

We observe that negative dilutions may be possible to be implemented, if $\bar{q}_j^{[1]}$ was originally obtained by dilution from a concentrated biological sample, but if that is not the case we will additionally constrain $\hat{\Delta}q_j \geq 0$ during the optimization process. The correction policy (6) should be carefully analyzed since $\hat{q}_j^{[1]}$ will differ in general of $\bar{q}_j^{[1]}$, and instead of moving into the point of high sensitivity $q_j^{ideal}(\hat{\eta}^{[1]})$, the real concentration will move to $\bar{q}_j^{[2]} = \bar{q}_j^{[1]} - \hat{q}_j^{[1]} + q_j^{ideal}(\hat{\eta}^{[1]})$, which depends on the estimation error resulted with the data in the first experiment. We propose to choose the vector $q^{ideal}(\hat{\eta}^{[1]})$ by minimizing the criterion $J(q^{ideal}, \hat{\eta}^{[1]})$ and then we use simulations to analyze how the correction based on the values in this vector affects the accuracy in the second experiment.

To establish the likely change of the accuracy in the second experiment due to the additional dilution $\hat{\Delta}q_j$, we want to produce a simulation scenario for statistically evaluating the accuracy changes, if one would have access to the true parameters of the model, $\bar{\theta}^{[1]} = [(\bar{q}^{[1]})^T (\bar{\eta}^{[1]})^T]^T$. We assume that the measured (simulated) data in the first experiment are given by

$$y_i = g(\bar{q}_{s_i}^{[1]} - d_i, \bar{\beta}) + \varepsilon_i, \tag{7}$$

where $\varepsilon_i \sim \mathscr{N}\left(0, \bar{\sigma}_0^2 g(\bar{q}_{s_i}^{[1]} - d_i, \bar{\beta})^{2\bar{\alpha}}\right)$, and we use the estimation procedure to produce the estimates $\hat{\theta}^{[1]} =$

$[(\hat{q}^{[1]})^T(\hat{\eta}^{[1]})^T]^T$. With the limited information available in $\hat{\theta}^{[1]}$ we propose to perform an additional dilution $\hat{\Delta}q_j$ for each biological sample, in order to correct the value $\hat{q}_j^{[1]}$ to the value $q_j^{ideal}(\hat{\eta}^{[1]}) = \hat{q}_j^{[1]} - \hat{\Delta}q_j$. In effect, the original concentration $\bar{q}^{[1]}$ will change to $\bar{q}^{[2]} = \bar{q}^{[1]} - \hat{q}^{[1]} + q^{ideal}(\hat{\eta}^{[1]})$, leading to a "true state of nature" vector $\bar{\theta}^{[2]} = [(\bar{q}^{[2]})^T(\bar{\eta}^{[1]})^T]^T$. We compute now the criterion $J(\bar{q}^{[2]}, \bar{\eta}^{[1]})$ which measures the accuracy in the second experiment. We are interested if in average the accuracy in the second experiment is better than the accuracy in the first experiment $J(\bar{q}^{[1]}, \bar{\eta}^{[1]})$, although the correction is done based on the estimated vector of parameters $\hat{\theta}^{[1]}$. This simulation scenario can be implemented in the following procedure:

Procedure 1 Evaluation of the accuracy improvement when changing the concentrations of biological samples:

Step 1 Choose a "true state of nature" vector of parameters $\bar{\theta}^{[1]} = [(\bar{q}^{[1]})^T(\bar{\eta}^{[1]})^T]^T$ and compute the criterion $J(\bar{q}^{[1]}, \bar{\eta}^{[1]})$ using (5).

Step 2 For $i = 1 : N_r$ do the following:

Step 2.1. Generate a data set according to (7) and estimate the parameter vector $\hat{\theta}^{[1]} = [(\hat{q}^{[1]})^T(\hat{\eta}^{[1]})^T]^T$.

2.2. Find the vector of log concentrations, $q^{ideal}(\hat{\eta}^{[1]})$ by minimizing the criterion $J(q^{ideal}, \hat{\eta}^{[1]})$.

2.3. Determine the true vector of log concentrations for the second experiment, $\bar{q}^{[2]} = \bar{q}^{[1]} - \hat{q}^{[1]} + q^{ideal}(\hat{\eta}^{[1]})$ and compute the criterion $J(\bar{q}^{[2]}, \bar{\eta}^{[1]})$.

Step 3 Collect the empirical distribution of $J(\bar{q}^{[2]}, \bar{\eta}^{[1]})$.

3.2 Changing the number of dilutions, replicates and the dilution factor

The protein of interest can be expressed over a broad dynamic range, for example up to a factor of 10^{10} [4]. To increase the measurement accuracy, the solution for each biological sample is serially diluted with a factor of two, yielding k diluted versions of the same solution and each of these diluted versions is printed for r times. The total number of measurements for each biological sample will be $k \times r$. The accuracy of estimation procedure will clearly change with the number of dilutions and the dilution factor used, especially for proteins with low expression levels, because many of the measured data will result in the low saturation range of the calibration curve, where the accuracy is low. The choice of the factors k and r is quite limited, since they both need be integers, and furthermore, the slide has a fixed number of spots, which imposes to keep constant the product $k \times r$. For the slides we discuss now $k \times r = 18$ and the only possibilities of choice are $(k,r) \in \{(1,18),(2,9),(3,6),(6,3),(9,2),(18,1)\}$.

Like in the previous case, we can use a similar simulation scenario to evaluate how the accuracy in the second experiment will be modified using an optimal combination $(r(\hat{\theta}^{[1]}), k(\hat{\theta}^{[1]}), b(\hat{\theta}^{[1]}))$ that is found using only the information in $\hat{\theta}^{[1]}$. Based on this optimal combination we compute the accuracy for the second experiment $J(\bar{\theta}^{[1]}, r(\hat{\theta}^{[1]}), k(\hat{\theta}^{[1]}), b(\hat{\theta}^{[1]}))$. The bulk of the distribution of these values should be located significantly at the left of the value $J(\bar{\theta}^{[1]}, 3, 6, 2)$ in order to decide to change the initial combination $(3,6,2)$.

However the change of (r,k,b) from one slide to the next is quite inconvenient technologically, so we contend here to illustrate how this change affects the accuracy of the experiment by computing the optimal accuracy for a given $\bar{\theta}^{[1]}$.

4. RESULTS

The presented procedures are illustrated here with the three lysate arrays data used in [3] and publicly available at [5]. Three different proteins are analyzed: pThr308AKT, pSer473AKT, and β-actin for $p = 96$ biological samples, where the solution for each biological sample is spotted in three replicates with six two-fold dilutions. Using the estimation procedure described in [3], we obtain for each slide the parameter vector estimate, $\hat{\theta} = [\hat{q}^T \hat{\beta}^T \hat{\sigma}_0 \hat{\alpha}]^T$. Here we are using only the sigmoidal model of the calibration curve as described in [3]. The sigmoidal model

$$g(x,\beta) = \beta_1 + \frac{\beta_2}{1+\exp(-\beta_3 x)} \tag{8}$$

has three free parameters $\beta = [\beta_1 \beta_2 \beta_3]^T$.

The true parameter vector $\bar{\theta}$, for each lysate array is given by the one estimated from the real data sets, and we will apply the simulation scenario described previously to illustrate the gains which can be obtained by changing the specified parameters. The calibration curve, the heteroscedastic variance, and the estimated log-concentration are illustrated in Figure 1.

4.1 Results - changing the initial dilution for each biological sample

We applied *Procedure1*, where the elements of the vector $q^{ideal}(\hat{\eta}^{[1]})$ are additionally constrained to be equidistantly spread in the interval $[q_{min}, q_{min} + L]$, so that the search is done in a two dimensional parameter space (q_{min}, L), conveniently selected to cover the non-saturated part of the sigmoidal model. The criterion $J(\bar{q}^{[2]}, \bar{\eta}^{[1]})$ was computed for $N_r = 200$ realizations. In Table 1 we show the results of the simulation scenario presented in *Procedure 1*. By looking at the last two columns in the table, we can conclude that the correction made based on the estimated vector $\hat{\theta}^{[1]}$ will lead on average to a better accuracy in the second experiment.

Although the main focus is in optimizing the overall criterion $J(\bar{q}^{[2]}, \bar{\eta}^{[1]})$, it is interesting to know how close the concentration values $\bar{q}^{[2]} = \bar{q}^{[1]} - \hat{q}^{[1]} + q^{ideal}(\hat{\eta}^{[1]})$ determined in Step 2.3. are to the optimal concentration values $q^{ideal}(\bar{\eta}^{[1]})$. The distribution of these values $\bar{q}^{[2]}$ is illustrated in Figure 2 by mean of the average $avg(\bar{q}_i^{[2]})$ and standard deviation $std(\bar{q}_i^{[2]})$ of $\bar{q}_i^{[2]}$ for each biological sample, superposed over the ideal concentrations $q_i^{ideal}(\bar{\eta}^{[1]})$.

4.2 Results - changing the number of dilutions, replicates and the dilution factor

For each lysate array $N_r = 200$ realizations of the $\hat{\theta}^{[1]}$ were obtained and for each realization the optimal combination $(r(\hat{\theta}^{[1]}), k(\hat{\theta}^{[1]}), b(\hat{\theta}^{[1]}))$ was found. Using this optimal combination we have computed the accuracy in the second experiment $J(\bar{\theta}^{[1]}, r(\hat{\theta}^{[1]}), k(\hat{\theta}^{[1]}), b(\hat{\theta}^{[1]}))$ which we denote for short $\hat{J}^{[2]}$ Table 2. We have also computed

proteinName	$J(\bar{q}^{[1]},\bar{\eta}^{[1]})$	$J(q^{ideal}(\bar{\eta}^{[1]}),\bar{\eta}^{[1]})$	$avg\{J(\bar{q}^{[2]},\bar{\eta}^{[1]})\}$	$std\{J(\bar{q}^{[2]},\bar{\eta}^{[1]})\}$
pThr308AKT	4.80	3.56	3.56	3.32e-003
pSer473AKT	8.05	2.57	2.57	2.15e-003
β-actin	7.60	4.00	4.03	2.57e-002

Table 1: **Accuracy measures comparison**: The criterion $J(\bar{q}^{[1]},\bar{\eta}^{[1]})$, reflecting the accuracy in the first experiment; The criterion $J(q^{ideal}(\bar{\eta}^{[1]}),\bar{\eta}^{[1]})$ reflecting the best achievable accuracy if one would know the true $\bar{\eta}^{[1]}$; The criterion $J(\bar{q}^{[2]},\bar{\eta}^{[1]})$, where $\bar{q}^{[2]} = \bar{q}^{[1]} - \hat{q}^{[1]} + q^{ideal}(\hat{\eta}^{[1]})$ depends on the noise in (7). The last two columns present the average and the empirical variance of $J(\bar{q}^{[2]},\bar{\eta}^{[1]})$ for the $N_r = 200$ realizations.

proteinName	$J^{[1]}$	$\bar{J}^{[2]}$	$avg\{\hat{J}^{[2]}\}$	$std\{\hat{J}^{[2]}\}$	$(\bar{r},\bar{k},\bar{b})$	$n_a(r)$	$n_a(k)$	$avg\{b\}$	$std\{b\}$
pThr308AKT	4.80	4.17	4.17	9.34e-003	(9, 2, 5.00)	200	200	5.28	4.98e-001
pSer473AKT	8.05	5.29	5.50	2.47e-001	(9, 2, 11.00)	200	200	14.70	4.24e+001
β-actin	7.60	5.78	5.89	1.25e-001	(1, 18, 1.40)	192	192	1.50	1.89e-002

Table 2: **Accuracy measures and parameter values comparison**: The criterion $J^{[1]}$ reflecting the accuracy in the first experiment; The criterion $\bar{J}^{[2]}$ reflecting the true accuracy in the second experiment; The average accuracy $avg\{\hat{J}^{[2]})$ and the accuracy variance $std\{\hat{J}^{[2]})\}$ for the second experiment; The true combination $(\bar{r}(\bar{\theta}^{[1]}),\bar{k}(\bar{\theta}^{[1]}),\bar{b}(\bar{\theta}^{[1]}))$; The number of times the true number of repetitions was found; The number of times the true number of dilutions was found; The average of the dilution factor and its variance.

for the true parameter vector $\bar{\theta}^{[1]}$ the optimal combination $(\bar{r}(\bar{\theta}^{[1]}),\bar{k}(\bar{\theta}^{[1]}),\bar{b}(\bar{\theta}^{[1]}))$ and the accuracy obtained when using these values, $J(\bar{\theta}^{[1]},\bar{r}(\bar{\theta}^{[1]}),\bar{k}(\bar{\theta}^{[1]}),\bar{b}(\bar{\theta}^{[1]}))$ which we denote for short $\bar{J}^{[2]}$. The accuracy in the first experiment $J(\bar{\theta}^{[1]},3,6,2)$ is denoted by $J^{[1]}$. The results are presented in Table 2 and from the column two to five we can see that in average we will attain the true accuracy in the second experiment. In columns seven and eight we show the success in determination of the optimal structure: $n_a(r)$ and $n_a(k)$ are the numbers of times of correct recovering for r and k, i.e. $r(\hat{\theta}^{[1]}) = \bar{r}(\bar{\theta}^{[1]})$ and $k(\hat{\theta}^{[1]}) = \bar{k}(\bar{\theta}^{[1]})$ and in the last two columns the average and variance of b are given, showing a good agreement with the value $\bar{b}$ from column 6.

In Figure 3, for all realizations we can see that for all proteins a better accuracy is obtained when the optimal combination is used in the second experiment.

5. CONCLUSIONS

We have presented methods for choosing new parameters for a second experiment of a lysate array that can improve significantly the accuracy of the estimated concentrations. Applying the procedures in laboratory will be the next stage of this research, in order to check how much of the predicted accuracy gains can be obtained in practice. In the past inference of the protein network was not studied as much as inference of genetic network, mostly due to the lack of accurate protein expression measurements. Lysate array may be the tool of the future for protein expression probing, similar to probing gene expressions by DNA microarrays.

REFERENCES

[1] C. Mircean, I. Smulevich, D. Cogdell, W. Choi, Y. Jia, I. Tabus, S.R. Hamilton, W. Zang, ”Robust estimation of protein expression ratios with lysate microarray technology”, *Bioinformatics*, 21(9):1935-42, May 2005.

[2] R. Jiang , C. Mircean, I. Shmulevich, D. Cogdell, Y. Jia, I. Tabus, K. Aldape, R. Sawaya, J. Bruner, G.N. Fuller, W. Zhang, ”Pathway alterations during glioma progression revealed by reverse phase protein lysate arrays”. *Proteomics* (in press, 2006).

[3] I. Tabus, A. Hategan, C. Mircean, J. Rissanen, I. Shmulevich, W. Zhang, J. Astola, ”Nonlinear modeling of protein expressions in protein arrays”, *IEEE Transactions on Signal Processing*, in press, 2006 (preprint available at [5]).

[4] V. Espina, A.I. Mehta, M.E. Winters, V. Calvert, J. Wulfkuhle, E.F. Petricoin 3rd, L.A. Liotta, ”Protein microarrays: molecular profiling technologies for clinical specimens”. *Proteomics*, 3(11):2091-100, Nov. 2003.

[5] http://www.cs.tut.fi/ tabus/lysate/Supplemental.pdf

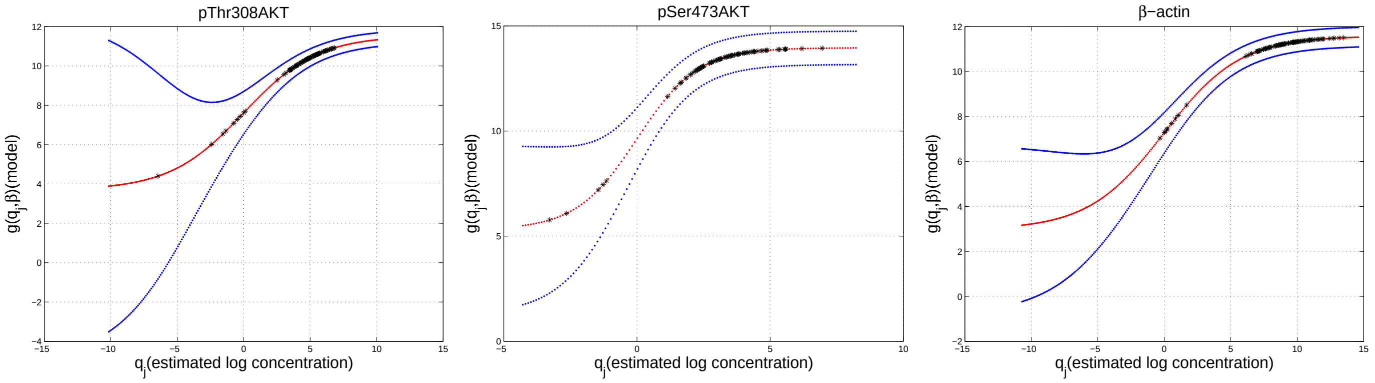

Figure 1: **Illustration of the true parameter vector $\bar{\theta}^{[1]}$ in the first experiment**: The calibration curve $g(x,\bar{\beta})$ is plotted in red; The 96 black stars are the pairs $(\bar{q}_j, g(\bar{q}_j,\bar{\beta}))$; The optimal model is heteroscedastic, the changing variance being visible from the two blue lines showing the bounds $g(x,\bar{\beta}) \pm 2\bar{\sigma}_0 g(x,\bar{\beta})^{\bar{\alpha}}$.

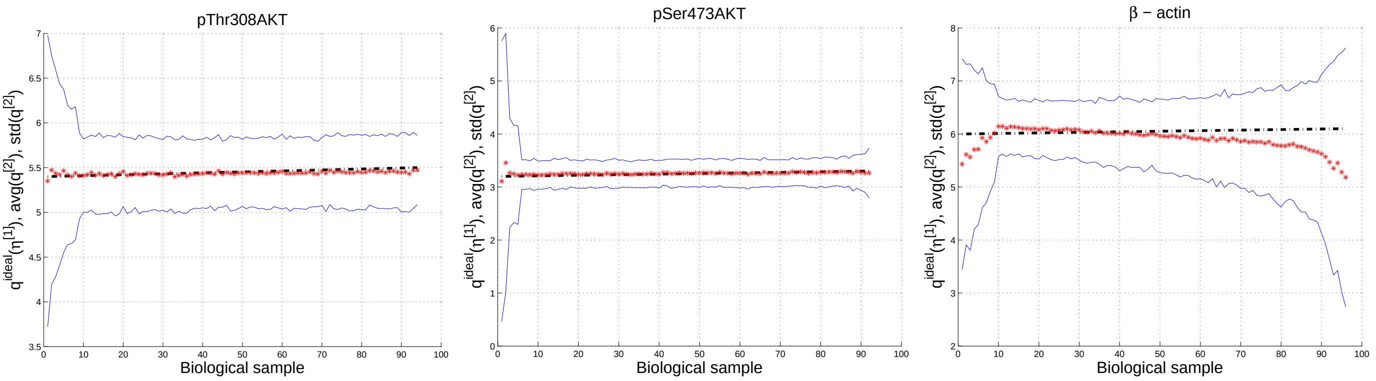

Figure 2: **Concentration values comparison**: The ideal log concentration values $q^{ideal}(\bar{\eta}^{[1]})$ plotted in black; The average values $avg(\bar{q}^{[2]})$ plotted in red, almost covering the black points; The bounds $avg(\bar{q}^{[2]}) \pm 2std(\bar{q}^{[2]})$ plotted in blue.

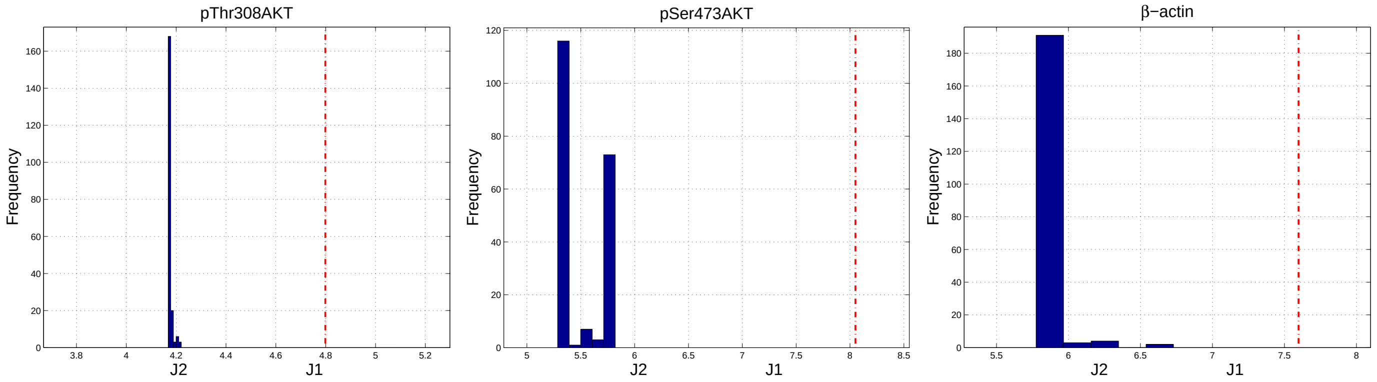

Figure 3: **Empirical distributions**: The accuracy in the first experiment $J(\bar{\theta}^{[1]},3,6,2)$ is shown by the vertical red dotted line; The distribution of the criterion in the second experiment, $J(\bar{\theta}^{[1]}, r(\hat{\theta}^{[1]}), k(\hat{\theta}^{[1]}), b(\hat{\theta}^{[1]}))$ when setting the optimal combination to $(r(\hat{\theta}^{[1]}), k(\hat{\theta}^{[1]}), b(\hat{\theta}^{[1]}))$ plotted in blue.