9-28-2010

A Semi-Nonparametric Mixture Model for Selecting Functionally Consistent Proteins.

Lianbo Yu
Ohio State University

Rebecca W. Doerge
Purdue University, doerge@purdue.edu

Recommended Citation
Yu, Lianbo and Doerge, Rebecca W., "A Semi-Nonparametric Mixture Model for Selecting Functionally Consistent Proteins." (2010). *Department of Statistics Faculty Publications.* Paper 6.
http://dx.doi.org/10.1186/1471-2105-11-486

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A semi-nonparametric mixture model for selecting functionally consistent proteins

Lianbo Yu[1], RW Doerge[2*]

Abstract

Background: High-throughput technologies have led to a new era of proteomics. Although protein microarray experiments are becoming more common place there are a variety of experimental and statistical issues that have yet to be addressed, and that will carry over to new high-throughput technologies unless they are investigated. One of the largest of these challenges is the selection of functionally consistent proteins.

Results: We present a novel semi-nonparametric mixture model for classifying proteins as consistent or inconsistent while controlling the false discovery rate and the false non-discovery rate. The performance of the proposed approach is compared to current methods via simulation under a variety of experimental conditions.

Conclusions: We provide a statistical method for selecting functionally consistent proteins in the context of protein microarray experiments, but the proposed semi-nonparametric mixture model method can certainly be generalized to solve other mixture data problems. The main advantage of this approach is that it provides the posterior probability of consistency for each protein.

Background

Over the last decade or longer, microarray technology has been used for measuring gene expression and has greatly impacted biomarker discovery [1], transcription factor identification [2], the assessment of gene interactions [3], and the detection of biological pathways [4]. Despite the massive application of microarrays to transcriptome applications there are limitations to the extent of the conclusions that can be made. Messenger RNA (mRNA) is the intermediate product of genes, with proteins being the final products and the key factors of metabolism. Although the levels of mRNA and protein for a gene are related they are not always highly correlated, which can be due to many reasons, e.g., translation rate, protein stability, and post-translational modification, etc. [5]. Given that the motivation and goal of many experiments is to understand not only the function of genes, but the network of genes that encode proteins, the abundance of proteins themselves are of increasing interest. Toward this end, microarray technology when adapted to proteins, are known as protein microarrays, and have been developed and widely used to assess the abundance of proteins [6-12]. The similarities between microarray technology as applied to gene expression [13], and as applied to protein abundance, are the same in that improved accuracy and precision, as well as design issues and normalization techniques for protein microarrays have been established [14,15].

Screening and identifying proteins as potential medical diagnostics and disease classification biomarkers is the main motivation of many protein microarray experiments [16-21]. The precursor to any successful screening application, and an essential issue that must be resolved to ensure that the accurate protein abundance measurements can be obtained by protein microarrays, is the consistency of a protein to report hybridization abundance. The protein itself is the probe on the array, and since proteins have a complex three dimensional structure, the structure itself, as well as the orientation of a protein, need to be retained. Toward this end, it is highly unlikely that every protein will be functional since different proteins often require different environment conditions for maintaining structures, and are typically much less stable than DNA. If the three dimensional nature of the structure is lost, or the required functional portion of the protein is not available to bind its target protein (i.e., the sample), the target protein

* Correspondence: doerge@purdue.edu
[2]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
Full list of author information is available at the end of the article

abundance measurement will be much smaller than it should be, or missed all together. Proteins whose structure or function are not maintained when attached to the array as probes are called inconsistent proteins, and if used provide inflated biomarker error rates (i.e., false positive rate and false negative rate). Alternatively, proteins that retain their structure and function are called consistent proteins and are desirable as probes on the array, and ultimately potential biomarkers. As such the selection of proteins that maintain functional consistency across experiments is a major and necessary requirement in the design and analysis of protein microarray experiments [17].

Certainly, high-throughput chemical validation of protein consistency is possible, but it is expensive and time consuming. Toward this end it is possible to statistically estimate protein consistency. In its simplest form, Pearson's correlation coefficient has been employed as a consistency measure in an antibody microarray study by Miller *et al*. [17], but it only measures the linearity of repeated measurements, and therefore is limited in its usefulness. A concordance correlation coefficient that is able to measure the consistency of repeated measurements was proposed by Lin [22], and later expanded to a total deviation index (TDI) [23], which provides a boundary within which a certain required percentage of differences between paired observations is obtained while controlling the error rate. As described by Lin [24], TDI and the concordance correlation coefficient provide the same information, but from different perspectives, and thus share their limitations. Namely, both the concordance correlation coefficient and TDI only demonstrate good asymptotic properties under the assumption of normality; a reality that is often questionable in application. Furthermore, the comparability of concordance correlation coefficients across proteins requires the ranges of the abundance measurements of proteins to be similar, which is not practical in large scale experiments [25]. To address the challenges and issues that are associated with identifying functionally consistent proteins, we propose a new statistic based on variance components from an analysis of variance (ANOVA) model. We rely on a mixture model to achieve this goal. Applications of mixture models in biology have proven to be excellent for separating data into the correct number of classes. For example, Efron *et al*. [26] proposed a two-component mixture model for testing differential expression. In this application the distributions of the t-statistics from both differentially expressed genes and non-differentially expressed genes were estimated by a nonparametric method, but the tail probabilities were not able to be estimated accurately. Toward this end the accuracy of estimating the tail probability was improved by using a two-component mixture model Pan *et al*. [27] where a finite normal mixture was assumed for each component. For microarray data it certainly is possible to simulate test statistics under the null hypothesis (i.e., a single component) using permutation theory since the treatment conditions for testing differences are known. However, for protein array data the first challenge is to identify proteins that are consistent, and then work only with these data. In other words, we are focusing on separating proteins into inconsistent and consistent classes, and then using only the informative proteins (i.e., consistent proteins) to address the biological question(s). To achieve this we propose a novel two-component semi-nonparametric mixture model. Simulations demonstrate the performance of the proposed approach and provide food for thought when designing future protein microarray experiments. We also apply the proposed approach to real data for the purpose of demonstrating its usefulness.

Results and Discussion

Simulations were conducted for the purpose of providing insight into the performance and value of the proposed semi-nonparametric approach. Data were simulated from known consistency classifications. Data were analysed with the proposed approach and the number of times proteins are correctly classified is recorded. From these simulation results, false discovery rate, as well as false non-discovery rate were calculated and are discussed.

A power study

Simulations were designed to study the statistical power of the approach under different sample sizes and different underlying two-component mixture distributions. Data were simulated directly from nine unique two-component semi-nonparametric mixture distributions with specified parameter values (Table 1). The tuning parameter K took on values 0, 1, or 2 for each semi-nonparametric density in each mixture. Sample sizes are 50, 100, 300, or 500. The proportions of the first mixture component with smaller mean, λ_0, are 0.20, 0.50, or 0.80. The distance between the two mixture components are 1 or 2, where the distance is defined as

$$D = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2}, \quad (1)$$

and μ_1 and μ_2 represent the means of two components respectively, while σ_1 and σ_2 represent the standard deviations of two components, respectively. Under each combination of model settings, 1000 data sets were generated.

For each simulated data scenario, a two-component mixture model was fit to the data. The Expectation-maximization (EM) quasi-Newton algorithm was

Table 1 Nine different simulation scenarios

		Component 1				Component 2			
model	distance	ϕ_{11}	ϕ_{12}	μ_1	σ_1	ϕ_{21}	ϕ_{22}	μ_2	σ_2
1	D = 1	$\pi/2$	$\pi/2$	12	2	$\pi/2$	$\pi/2$	17	3
	D = 2	$\pi/2$	$\pi/2$	12	2	$\pi/2$	$\pi/2$	24	4
2	D = 1	$\pi/2$	$\pi/2$	12	2	2.17	$\pi/2$	19.4	3
	D = 2	$\pi/2$	$\pi/2$	12	2	2.17	$\pi/2$	26.7	4
3	D = 1	$\pi/2$	$\pi/2$	12	2	2	2.75	18	2
	D = 2	$\pi/2$	$\pi/2$	12	2	2	2.75	30.3	4
4	D = 1	0.97	$\pi/2$	9.8	2.3	$\pi/2$	$\pi/2$	17	3
	D = 2	0.97	$\pi/2$	9.8	2.3	$\pi/2$	$\pi/2$	24	4
5	D = 1	0.97	$\pi/2$	10	2	2.17	$\pi/2$	18.15	2.5
	D = 2	0.97	$\pi/2$	10	2	2.17	$\pi/2$	26.15	4
6	D = 1	0.97	$\pi/2$	9.8	2.3	2	2.75	19.8	2.5
	D = 2	0.97	$\pi/2$	9.8	2.3	2	2.75	31.3	4
7	D = 1	4.1	0.9	9.7	1.8	$\pi/2$	$\pi/2$	17	3
	D = 2	4.1	0.9	9.7	1.8	$\pi/2$	$\pi/2$	24	4
8	D = 1	4.1	0.9	9.7	1.8	2.17	$\pi/2$	20.5	3.5
	D = 2	4.1	0.9	9.7	1.8	2.17	$\pi/2$	28.5	4.6
9	D = 1	4.1	0.9	10	2	2	2.75	19.8	2.5
	D = 2	4.1	0.9	10	2	2	2.75	31.3	4

Each simulation model is based on a two-component semi-nonparametric (SNP) mixture distribution that has eight parameters: ϕ_{11}, ϕ_{12}, μ_1, σ_1, ϕ_{21}, ϕ_{22}, μ_2, σ_2. The distance D of two components is 1 or 2.

employed to estimate the model parameters. Model selection criteria Akaike's Information Criterion (AIC) [28], Schwartz Bayesian Information Criterion (BIC) [29] and Hannan-Quinn Criterion (HQ) [30] were used to select the best model. A likelihood ratio test (19; see Methods) was employed to determine whether the mixture distribution was identifiable as two-components (18). A bootstrap method approach approximated the null distribution of the likelihood ratio test statistics, and provided a significance threshold for the likelihood ratio test statistic (see Methods). Power was calculated by estimating the proportion of correctly rejected hypotheses for each of 1000 data sets. Power comparisons for all parameter settings using BIC model selection criteria are provided in Figures 1. The general trend across all three model selection criteria is that well separated mixture distributions (D = 2) outperform mixtures that are not well identified (D = 1). When the mixtures are well defined, there is obvious increased power for situations where the mixing proportion (λ_0) of the functionally consistent component of the mixture distribution is 50% or greater. Recall, when the tuning parameters K_1 and K_2 are both zero the semi-nonparametric densities in the mixture distribution are both standard normal densities.

As expected, higher power is associated with larger sample size. Dramatically higher power is achieved when the distance between the two components is increased from 1 to 2 simply because the null hypothesis (18) is easier to reject when the mixture components are well separated. Furthermore, AIC tends to choose a larger model that has a larger likelihood ratio test statistic (19) when compared to the smaller model chosen by BIC or HQ [31], therefore the use of AIC yields higher power than BIC or HQ.

Simulated Data Scenario

The performance of the proposed mixture model with semi-nonparametric densities is evaluated for selecting functionally consistent proteins in a simulation setting based on a real experiment. Since we are interested in understanding the performance of the proposed approach it is necessary to rely on simulated data, rather than actual data since the truth for real data is unknown. Protein microarray data were simulated based on the data scenario described in Zhou *et al.* [32]. Specifically, there are three groups of patients with different stages of disease, and one group of healthy patients. Each group consists of 10 patients (40 patients total). For each patient, hybridization abundance was measured on 300 proteins. Each of the 300 proteins was represented as a probe on the array. Onboard probe (technical) replication allowed each protein to be represented 6 times on the array. Forty samples were individually mixed with a reference sample, hybridized to an array, and the entire experiment was repeated twice. Protein microarray data were simulated as follows

$$log_2(\mu_{jk}) = \mu + G_j + S_{k(j)}, \tag{2}$$

where μ_{jk} represents hybridization abundance of individual or patient k in group j, μ represents the overall mean abundance, G_j represents the fixed effect of group j, $S_{k(j)}$ represents the random effect of patient k in group j following a normal distribution with mean 0 and variance σ^2_{Sj}, and

$$\begin{aligned} \gamma_{ijkl} &= \theta_{jk} + \delta_{ijk} + \epsilon_{ijkl} \\ &= log_2(\mu_{jk}) - log_2(\bar{\mu}..) + \delta_{ijk} + \epsilon_{ijkl} \\ &= C + G_j + S_{k(j)} + \delta_{ijk} + \epsilon_{ijkl}, \end{aligned} \tag{3}$$

where $C = -log_2 \frac{\sum_{j,k} 2^{G_j + S_{k(j)}}}{40}$, i = 1,2, j = 1,2,3,4, k = 1,2, ..., 10, l = 1,2, ..., 6. y_{ijkl} represents the lth log signal ratio of patient k to the reference sample in group j for experiment i, θ_{jk} represents the mean log signal ratio of the patient k sample to the reference in group j, $\bar{\mu}..$ represents the average of μ_{jk}'s over j and k, δ_{ijk} represents the random error of experiment i for

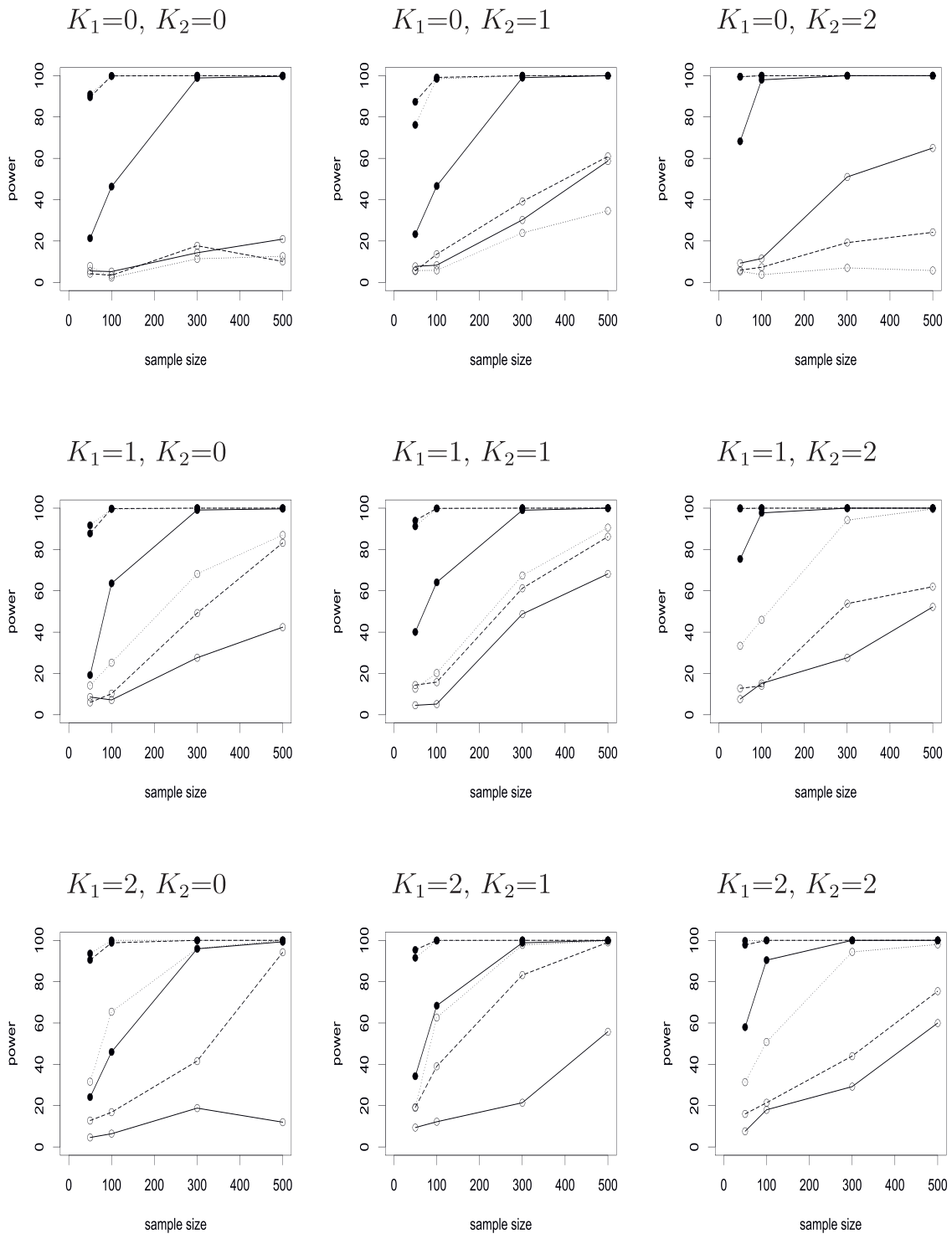

Figure 1 Power results for nine simulation settings. Schwartz Bayesian Information Criterion (BIC) provides the model selection criterion. Data were simulated under nine semi-nonparametric (SNP) mixture distributions with the tuning parameter K taking values 0, 1, or 2 for each SNP density. Sample sizes are 50, 100, 300, or 500. λ_0 is 0.2, 0.5, or 0.8. The distance between the means of the component distributions is *D* and has values of 1 or 2. Power was calculated as the proportion of correctly rejected hypothesis for 1000 simulated data sets. Solid curves represent λ_0 = 0.20 and *D* = 1(○) or *D* = 2(●). Dashed curves represent λ_0 = 0.50 and *D* = 1(○) or *D* = 2(●). Dotted curves represent λ_0 = 0.80 and *D* = 1(○) or *D* = 2(●).

patient k in group j, and ϵ_{ijkl} represents the lth random error within experiment i for patient k in group j. Assume that δ_{ijk} is from a normal distribution with mean zero and variance σ_δ^2, ϵ_{ijkl} is from a normal distribution with mean zero and variance σ_e^2.

The model parameter settings for the simulation were taken from the aforementioned Zhou et al. antibody microarray data [32], such that the $G_j's$ were sampled from uniform distribution U[-1, 1], $\sigma_{Sj}^2 = (0.4\nu)^2$, where ν were sampled from U[0.5, 2] for different j, $\sigma_\delta^2 = 0.2^2$, and $\sigma_e^2 = 0.15^2$. The hybridization abundance data for 300 functionally consistent proteins on each array were simulated from model (2) and (3) (see Methods). Fifty percent of these simulated proteins were randomly chosen to be functionally inconsistent proteins by adding a random between-array deviation with mean 0 and standard deviation drawn from U [0.05, 0.5], as well as a random within-array deviation with mean 0 and standard deviation taken from U[0.1, 0.4], to a randomly chosen number of separate arrays. Protein classification resulted from estimating the variance components in the ANOVA model (4; see Methods), and modelling the between and within-array variance component statistic with a semi-nonparametric mixture model. The main advantage of the proposed mixture model approach is that it provides the posterior probability of consistency for each protein which in turn establishes the classification rule, as well as estimates the respective error rates.

One thousand data simulations were performed under the same simulation setting, and for each the sum of the between- and within-array variation (see Methods) provides the statistic that is ultimately modelled and used for classifying each of the 300 proteins by fitting to the semi-nonparametric mixture model (5). Posterior probabilities defined in Equation (21) and Equation (22) were computed, and then used to calculate the estimated false discovery rate (FDR) in Equation (24), and the estimated false non-discovery rate (FNR) in Equation (25). Since these are simulated data for which we know the true classification, the true FDR and FNR were calculated and compared to the respective estimated values from the simulated data analyses. The estimated and true FDR and FNR were averaged over 1000 simulations, respectively. The average FDR (Figure 2) and FNR (Figure 3) were plotted against the number of inconsistent proteins. The conservative nature of this approach is illustrated in the downward bias of the FDR estimates and the corresponding upward bias of the FNR estimates.

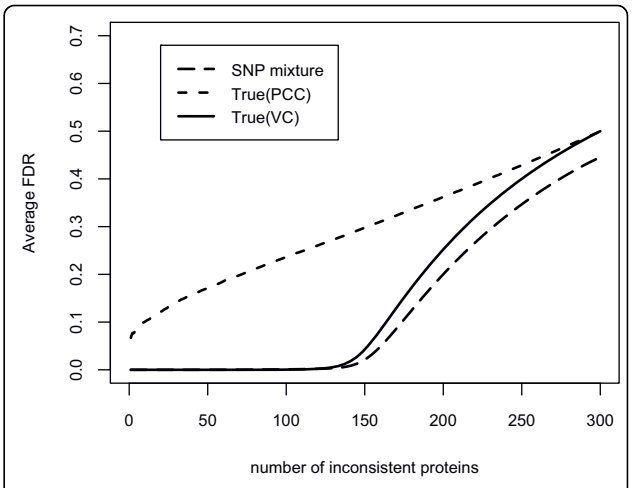

Figure 2 Comparison of false discovery rates (FDR) using simulated data based on 40 patients and 6 onboard probe replicates. False discovery rates are averaged over 1000 simulations. Solid curves are the true false discovery rates based on the between- and within-array variance component (VC) statistic, short-dashed curves are the true false discovery rates based on Pearson's correlation coefficient (PCC), and long-dashed curves are the estimated false discovery rates based on the proposed semi-nonparametric (SNP) mixture model method.

We compare the proposed semi-nonparametric approach to the work of Miller et al. [17] who selected functionally consistent proteins using an arbitrary cutoff value for Pearson's correlation coefficient. It is important to realize that their cutoff value is not statistically

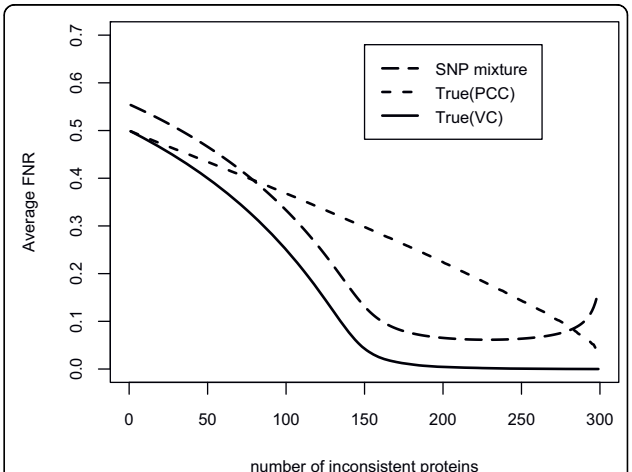

Figure 3 Comparison of false non-discovery rates (FNR) for the simulated data with 40 patients and 6 onboard probe replicates. False non-discovery rates are averaged over 1000 simulations. Solid curves are the true false non-discovery rates based on the between- and within-array variance component (VC) statistic, short-dashed curves are the true false non-discovery rates based on Pearson's correlation coefficient (PCC), and long-dashed curves are the estimated false non-discovery rates based on the proposed semi-nonparametric (SNP) mixture model method.

justified, nor does it provide error rate control. We calculated Pearson's correlation coefficients (PCC) for each of the 1000 simulated data sets, and reported in the average FDR and FNR results in Figures 2 and 3, respectively. Not surprisingly, larger true error rates are experienced for the Pearson's correlation coefficient when compared to the variance component (VC) statistic that is based on the between- and within-array variation. Essentially, the variation in the random error(s) captures the difference between consistent and inconsistent proteins allowing the variance estimate based on between- and within-array variation to provide information about protein consistency. Based on this rationale, the misclassification error rates of the proposed approach are expected to be smaller than the Pearson's correlation coefficient. As can be seen for Pearson's correlation coefficient, when the number of inconsistent proteins is 160, the false discovery rate is 0.310 (Figure 2) and the false non-discovery rate is 0.283 (Figure 3). By comparison, based on the between- and within-array variation statistics the false discovery rate is 0.083 and the false non-discovery rate is 0.024. The same phenomena occur at any other number of inconsistent proteins (Figures 2 - 3).

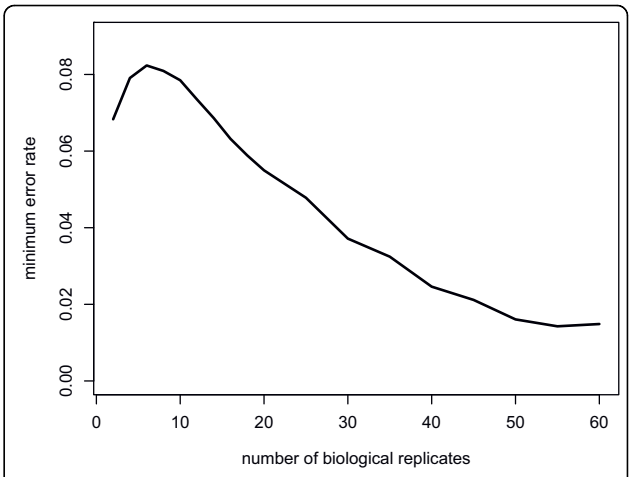

Figure 4 Minimum classification error rate for increasing numbers of biological replicates (i.e., increasing patient number). Minimum classification error rate was computed for each number of replicates ranging from 2 to 60 for a fixed number (6) of protein probes. Larger numbers of replicates/patients achieve greater classification results.

Biological and technical replication

We explored the influence of the number of biological replicates (or, total patient number) and technical replicates (or, onboard per protein probe replicate) on the proposed semi-nonparmetric mixture model approach for selecting functionally consistent proteins using two different simulation settings. Data were simulated from model (2) and (3) (see Methods). The first simulation focused on the number of biological replicates or patients (2 to 60) while fixing the number of onboard probe replicates representing each protein at 6. Six onboard replicates is a relatively large number and is in keeping with many of the current protein microarray investigations. The classification error rate was computed by minimizing (26; see Methods) for each number of replicates (Figure 4) under consideration, and it can be seen that rate drops off quickly as the number of replicates increases from 6 to 50. The second simulation evaluated the number of onboard per protein probe replicates while fixing the number of biological replicates (or patients) at 40. Figure 5 illustrates the classification error rates dropping as the number of onboard replicates increases. Clearly, the largest decrease is most dramatic in the range from 2 to 4.

A case study

We applied our method to data from an antibody microarray experiment from Zhou *et al*. [32]. Two-color rolling-circle amplification (RCA) was used to assess thirty five antibody proteins from duplicate sets of twenty four serum samples using antibody microarrays prepared on nitrocellulose. The twenty four serum samples consist of six liver cancer patients, six pre-cirrhotic patients, six cirrhotic, and six normals. Each antibody has 5 replicates on the array.

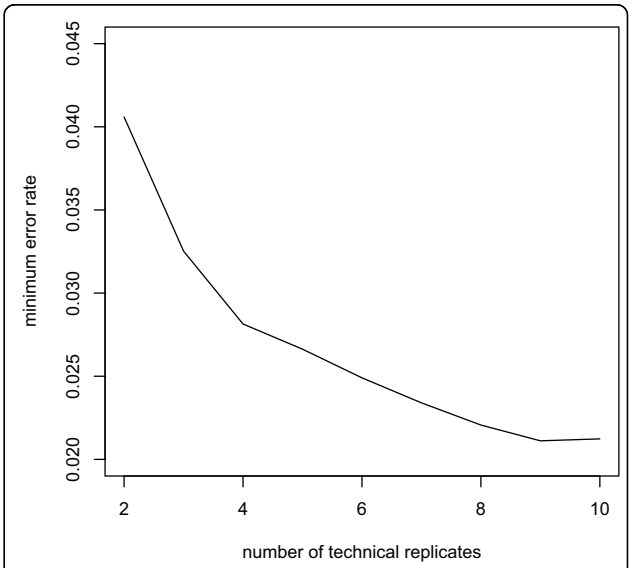

Figure 5 Minimum classification error rate for increasing numbers of technical replicates (i.e., number of onboard replicates for each protein probe). Minimum classification error rate was computed for each number of technical replicates (2 to 10) for a fixed number (40) of biological replicates (or patients). Larger numbers of technical replicates achieve better classification results.

In the analysis, one antibody was removed due to no signal. The ANOVA model (4) was employed to calculate the total variation due to random error. The range of the consistency statistics is shown in a histogram (Figure 6), from which we can see there are two clusters. To test whether the two components of the mixture model (i.e., consistent and inconsistent proteins) are separable we calculated the likelihood ratio test (19) and found it to be 13.65, for which significance was determined by comparing with a critical value under the null hypothesis. To do that, we bootstrapped the null hypothesis likelihood ratio test statistics and obtained a p-value of 0.022 for the likelihood ratio test statistics that corresponds to the value 13.65, therefore we rejected the null hypothesis and concluded that the consistent and inconsistent proteins are separable. Both components, as determined by the BIC criterion, under the alternative hypothesis have $K = 0$, and have the model parameter estimate $\hat{\mu}_0 = 0.092$, $\hat{\sigma}_0 = 0.037$, $\hat{\mu}_1 = 0.267$, $\hat{\sigma}_1 = 0.089$ and $\hat{\lambda}_0 = 0.588$. Because the sample size (number of proteins) is small, model selection tends to choose simpler models where $K = 0$. We illustrate the complex densities where $K = 2$ for both components in Figure 6. In order to determine the optimal number of functionally consistent proteins, we estimated the FDR (24) and FNR (25) and obtained the error rate (26) by using a 2:1 ratio for the cost of FDR and FNR. Figure 7 shows that the minimum error rate occurs when there are 13 inconsistent proteins and 21 consistent proteins. In Zhou *et al.* [32], Pearson's correlation coefficient was calculated for each protein to evaluate measurement reproducibility. Unfortunately, Pearson's correlation coefficient does not provide a classification of consistent and inconsistent proteins, so it is not possible to compare the results of our approach with the published results from Zhou *et al.* [32].

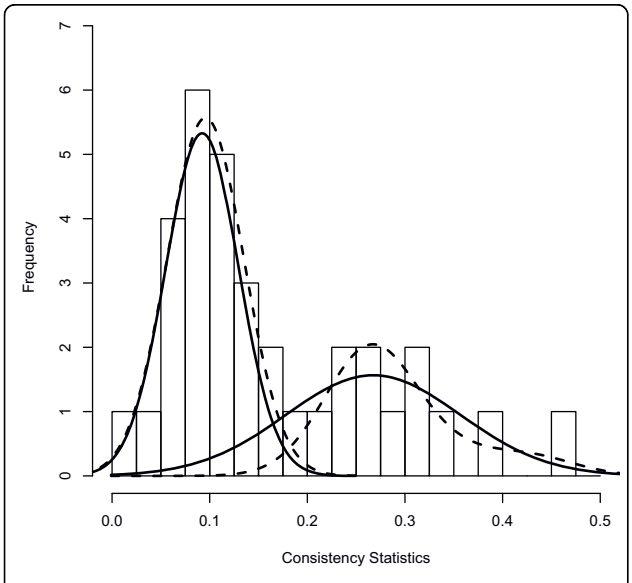

Figure 6 Histogram of consistent statistics. Solid curves are the estimated densities where $K = 0$ for both components. These results are based on the BIC model selection criterion. Dashed curves are the estimated densities where $K = 2$ for both components.

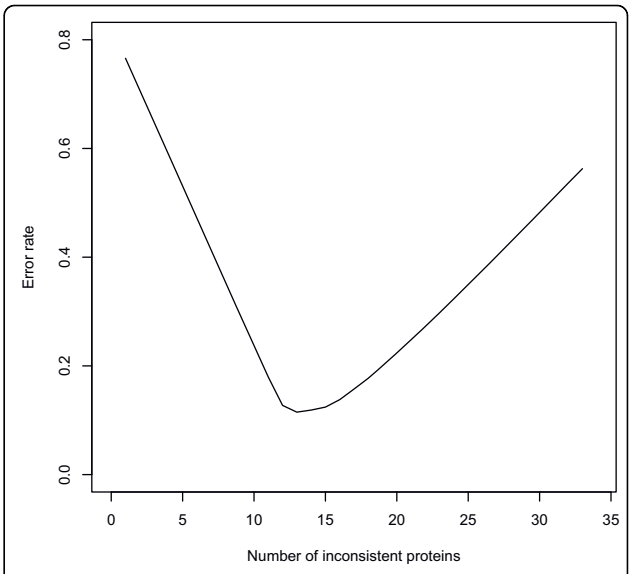

Figure 7 Error rate for claiming the number of inconsistent proteins. The estimated classification error rate plotted against each number of potential inconsistent proteins. The estimated FDR (24) and FNR (25) was calculated using a 2:1 ratio for the cost of FDR and FNR. The minimum error rate occurs when 13 proteins are found as inconsistent while 21 protein are consistent.

Discussion

The challenge of selecting and employing functionally consistent proteins for protein microarray experiments is complicated by the three-dimensional structure of the protein itself. Specifically, the proteins that are spotted on to the array as probes (during the fabrication of the array) need to maintain functional consistency for each sample hybridized to the array, as well as across experiments. Identifying and employing functionally consistent proteins continues to be a major and necessary concern in both the design and analysis of protein microarray experiments. To address this concern, a novel statistical approach based on modelling the between- and within-array variation, using a semi-nonparametric mixture model, is presented for the purpose of discriminating functionally consistent proteins. Of course, once functionally consistent proteins have been identified and the array fabricated, it is then necessary to develop additional statistical methods that can detect proteins of differing abundance.

After classifying proteins as consistent and inconsistent proteins, the abundance data from functionally consistent proteins can be used for differential protein abundance/expression analysis. The semi-nonparametric mixture model that was initially proposed to select functionally consistent proteins (5) can also be adapted for detecting differentially expressed proteins. Specifically, one component of the mixture identifies the non-differentially expressed proteins, while the other component acknowledges the differentially expressed proteins. The semi-nonparametric mixture model lies between parametric and nonparametric approaches since it does not put distributional assumption on the data themselves, but on the test statistics. The semi-nonparametric mixture model as applied to differential expression analysis was investigated and shows great performance [33].

The proposed semi-nonparametric mixture model is a novel and broadly applicable approach in the mixture model literature. For applications to either identifying functionally consistent proteins, or testing for differential protein abundance between samples, only two-component mixture models are employed. The extension of the semi-nonparametric mixture model to a multiple-component and multivariate mixture model has potential to address high-dimensional problems for the purpose of classification, and it has potential to work for a variety of data problems since it provides the flexibility necessary for model fitting.

Conclusions

A novel semi-nonparametric mixture model is proposed for the purpose of selecting functionally consistent proteins that can be used for protein microarray experiments. The proposed approach is able to attach a posterior probability of being inconsistent to each protein, from which false discovery and false non-discovery rates can be estimated. We validated the performance of our method through simulations. Additionally, the characteristics of the semi-nonparametric mixture model were studied by a power analysis. Our novel method provides an improvement in the accuracy of proteins that are selected as probes on a protein microarray, as well as an alternative approach to studying a variety of additional mixture data problems.

Methods

ANOVA model

Consider a repeated protein microarray experiment. There are m proteins (probes) spotted on n arrays. These n arrays are used to hybridize material for n test samples from J different patient groups. The same amount of a reference sample is mixed with each test sample, and each mixture is hybridized on one of n arrays. The background corrected abundance ratios of sample to reference are obtained for each probe on each array and properly normalized. There are several unique normalization methods proposed for protein microarray data, and the comparison of them are presented by Hamelinck [14].

An analysis of variance (ANOVA) model can be used to partition the sources of variation of the normalized abundance data. The ANOVA model for each protein is

$$Y_{ijkl} = \mu + T_j + S_{k(j)} + \delta_{ijk} + \epsilon_{ijkl}, \quad (4)$$

where Y_{ijkl} represents the protein abundance ratio between sample and reference of replicate l for sample k within group j in experiment i, μ represents the overall mean of the expression ratios, T_j represents the fixed effects of group j with constraint $\Sigma_j\ T_j = 0$, $S_{k(j)}$ represents the random effects of sample k within group j with mean 0 and variance σ^2_{Sj}, δ_{ijk} represents the normally distributed random between-experiment effect of experiment i for sample k in group j with mean 0 and variance σ^2_δ, ϵ_{ijkl} represents the normally distributed random error with mean 0 and variance σ^2_ϵ.

The total of the between-array variation σ^2_δ and the within-array variation σ^2_ϵ represents the variation due to random error. Inconsistent proteins inflate both the between-array and within-array variation. By least-squares estimation of the ANOVA model (4), the estimation of $\sigma^2_\delta + \sigma^2_\epsilon$ is obtained for each protein and used for classification via a novel semi-nonparametic mixture model approach.

Semi-nonparametric mixture model

The total of the between-array variation, σ^2_δ, and the within-array variation, σ^2_ϵ, represents the variation due to random error in the ANOVA model (4) and can be estimated for each protein. To select functionally consistent proteins, we assume that all spotted proteins on the arrays represent both functionally consistent and functionally inconsistent proteins with certain proportions that are not too small to be negligible. By modelling the collection of consistency statistics ($\hat{\sigma}^2_\delta + \hat{\sigma}^2_\epsilon$) from each protein, using a mixture distribution, it is possible to estimate the consistent or inconsistent status for every protein that is represented on the array. Biologically and technically, consistent proteins are very reliable and are able to generate reproducible measurements between experiments. Because the proposed consistency statistic

captures the differences in both consistent and inconsistent proteins, it should be smaller for consistent proteins simply because they have less variation (i.e., are reliable and reproducible) than the inconsistent proteins. Furthermore, statistically, when fitting a two-component mixture model and estimating two components simultaneously, the components have to be identifiable. Therefore, we assume that the mean of the statistics for consistent proteins is smaller than the mean of the statistics for inconsistent proteins, and that the statistics from the same class will be aggregated. For this application, defining a selection criterion is equivalent to finding the classification rule between two classes.

A mixture model with semi-nonparametric densities is proposed, and the Expectation-maximization (EM) quasi-Newton algorithm [34,35] is employed to estimate the parameters. Inferences are then drawn from the estimated mixture model.

Consider the generalized setting where $z_1, z_2, ..., z_m$ represent m consistency statistics ($\hat{\sigma}_\delta^2 + \hat{\sigma}_\epsilon^2$), and $f(z|\theta)$ represents their density function. Under the assumption of a two-component mixture, the density $f(z|\theta)$ is equal to a weighted sum as follows

$$f(z \mid \theta_0, \theta_1, \lambda_0, \lambda_1) = \lambda_0 f_0(z \mid \theta_0) + \lambda_1 f_1(z \mid \theta_1), \quad (5)$$

where θ_0 and θ_1 are the parameters for two densities, $f_0(z|\theta_0)$ is the density of the z_i's that are the statistics for functionally consistent proteins, $f_1(z|\theta_1)$ is the density of the z_i's that are the statistics for functionally inconsistent proteins, λ_0 is the proportion of the functionally consistent proteins, λ_1 is the proportion of the functionally inconsistent proteins, and the sum of λ_0 and λ_1 is 1. For a mixture model in (5), an order between the means of the two components is assumed. Specifically, let μ_0 and μ_1 represents the means of two components respectively, and assume $\mu_0 \le \mu_1$.

We assume that the density $f_0(z|\theta_0)$ and $f_1(z|\theta_1)$ belong to a class of semi-nonparametric (SNP) density used by Gallant and Nychka [36]. This smooth class of densities can be represented by a truncated Hermite series expansion, and contain densities that can be skewed, thin or heavy tailed, and multi-modal. The density is represented by

$$f_i(z \mid \theta_i) = \left[\sum_{j=0}^{K} a_j \left(\frac{z-u_i}{v_i} \right)^j \right]^2 \frac{1}{v_i} \phi\left(\frac{z-u_i}{v_i} \right), \quad (6)$$

where $i = 0, 1$, K represents a tuning parameter that is nonnegative, $\phi(\cdot)$ represents a standard normal density.

In order to have $f_i(z|\theta_i)$ as a density, a restriction is imposed

$$E\left[\sum_{j=0}^{K} a_j \left(\frac{z-u_i}{v_i} \right)^j \right]^2 = 1, \quad (7)$$

where z is a normally distributed random variable. When $K = 0$, a_0 has to be 1 so that $f_i(z|\theta_i)$ is exactly the standard normal density. Fortunately, there exists a transformation of coefficients to satisfy the above restriction when K is larger than 0. For $K = 1$, the transformation is represented by

$$\begin{aligned} a_0 &= sin(\phi), \\ a_1 &= cos(\phi). \end{aligned} \quad (8)$$

In this case, $\theta_i = (\phi, u_i, v_i)$, where $i = 0, 1$. For $K = 2$, the transformation is denoted by

$$\begin{aligned} a_0 &= sin(\phi_1) - \frac{cos(\phi_1)cos(\phi_2)}{\sqrt{2}}, \\ a_1 &= cos(\phi_1)sin(\phi_2), \\ a_2 &= \frac{cos(\phi_1)cos(\phi_2)}{\sqrt{2}}. \end{aligned} \quad (9)$$

Here $\theta_i = (\phi_1, \phi_2, u_i, v_i)$, where $i = 0, 1$.

The latent variable R_i takes value 0 if protein i is functionally consistent, or 1 otherwise. The likelihood of the complete data (Z, R) is

$$L = \prod_{i=1}^{m} [\lambda_0 f_0(z_i \mid \theta_0)]^{1-R_i} [\lambda_1 f_1(z_i \mid \theta_1)]^{R_i}. \quad (10)$$

The log-likelihood is then obtained as follows

$$\begin{aligned} logL = \sum_{i=1}^{m} [&(1-R_i)log\lambda_0 + (1-R_i)log\, f_0(z_i \mid \theta_0) \\ &+ R_i log\lambda_1 + R_i log f_1(z_i \mid \theta_1)]. \end{aligned} \quad (11)$$

Based on the log-likelihood (11), maximization techniques can be employed to find the estimates of model parameters, and then classification methods can be implemented based on the estimates of the mixture model.

EM-algorithm

To estimate the parameters in the log likelihood function (11)the EM-algorithm [34] is employed. There are two steps in the EM-algorithm. In the E-step, the conditional expectation given the data is calculated for missing values R_i

$$\begin{aligned} E(R_i) &= P(R_i = 1 \mid z_i) \\ &= \frac{\lambda_1 f_1(z_i|\theta_1)}{\lambda_0 f_0(z_i|\theta_0) + \lambda_1 f_1(z_i|\theta_1)}. \end{aligned} \quad (12)$$

After substituting in the expectations of missing values, the log likelihood in (11) is maximized (the M-step) by a gradient algorithm that is accelerated by a quasi-Newton method [35]. Given initial values of the parameters, the EM-algorithm iterates between the E-step and the M-step until a convergence criterion is met or until a maximum iteration number is reached.

In the M-step at the $(n + 1)^{th}$ iteration, the two parts in the log-likelihood function can be represented by

$$\begin{aligned} Q_0(\theta_0 \mid \theta_0^n) &= \sum_{i=1}^{n} [(1 - \hat{R}_i) log\lambda_0 + (1 - \hat{R}_i) log\, f_0(z \mid \theta_0)], \\ Q_1(\theta_1 \mid \theta_1^n) &= \sum_{i=1}^{n} [\hat{R}_i log\lambda_1 + \hat{R}_i log\, f_1(z \mid \theta_1)], \end{aligned} \tag{13}$$

where θ_0^n and θ_1^n are the estimated parameters in n^{th} step. The EM-gradient algorithm [35] updates the parameters as follows,

$$\begin{aligned} \theta_0^{n+1} &= \theta_0^n - \left(d^2 Q_0(\theta_0^n \mid \theta_0^n) - B_0^n \right)^{-1} d^1 Q_0(\theta_0^n \mid \theta_0^n), \\ \theta_1^{n+1} &= \theta_1^n - \left(d^2 Q_1(\theta_1^n \mid \theta_1^n) - B_1^n \right)^{-1} d^1 Q_1(\theta_1^n \mid \theta_1^n), \end{aligned} \tag{14}$$

where d^1 represents the first partial derivatives with respect to the parameters, and d^2 represents the second partial derivatives with respect to the parameters. B_0^n and B_1^n are updated in each iteration by applying Davidon's [37] update

$$B_i^n = B_i^{n-1} + a_i^n c_i^n (c_i^n)', \tag{15}$$

where i = 0, 1, constant a_i^n and vector c_i^n are defined as

$$\begin{aligned} a_i^n &= \frac{1}{(g_i^n - B_i^{n-1} s_i^n)' s_i^n}, \\ c_i^n &= g_i^n - B_i^{n-1} s_i^n, \end{aligned} \tag{16}$$

with

$$\begin{aligned} s_i^n &= \theta_i^{n-1} - \theta_i^n, \\ g_i^n &= d^1 Q(\theta_i^{n-1} \mid \theta_i^n) - d^1 Q(\theta_i^{n-1} \mid \theta_i^{n-1}). \end{aligned} \tag{17}$$

Determining the number of mixture components

Before applying the two-component mixture model to classify proteins (as consistent or inconsistent), we need to test that the number of components is compatible with the two component mixture model, and that the components can be identified.

Let g denote the number of mixture components. The hypothesis to test is

$$H_0 : g = 1 \quad \text{vs.} \quad H_\alpha : g = 2. \tag{18}$$

Ledwina [38] introduced the idea of a data-driven test for Neyman's smooth test of fit. Here the idea is generalized to the likelihood ratio test of mixture components. The likelihood ratio test statistic is defined as

$$-2log\lambda = -2log \frac{L\widehat{\theta_0}}{L\widehat{\theta_\alpha}}, \tag{19}$$

where $\hat{\theta}_0$ and $\hat{\theta}_\alpha$ represent the estimated parameters under the null and alternative hypothesis, respectively. $\hat{\theta}_0$ and $\hat{\theta}_\alpha$ are obtained by choosing the best model via model selection when the two-component mixture model is fit to the data. A bootstrap method is performed to approximate the null distribution of $-2log\lambda$ [39], and to provide a significance threshold for the likelihood ratio test statistic. Specifically, when estimating the null distribution of the likelihood ratio test statistics, we first bootstrap 500 data sets from the estimated distribution under the null hypothesis, and then perform a likelihood ratio test (19) for each simulated data set.

If the test statistic is significant, the two-component mixture model is suitable to fit the data in order to select functionally consistent proteins. Failure to reject the null hypothesis (18) indicates that consistent and inconsistent proteins are not separable, or that there is only one type of protein on the array. For this situation, specific chemical validation techniques have to be employed in order to provide additional consistency information.

Model selection

For the density in Equation (6), the tuning parameter K can be set equal to 0, 1, or 2. To balance the size of parameters and the suitability of the model fit, information criteria are applied to choose the mixture representation that fits the data best. Akaike's Information Criterion (AIC) [28], Schwartz Bayesian Information Criterion (BIC) [29] and Hannan-Quinn Criterion (HQ) [30] are applied, and they all share a penalized log-likelihood in the form of

$$-2logL + C(N)p, \tag{20}$$

where $logL$ is the log-likelihood, p is the number of free parameters in the model, and $C(N)$ is a function of

sample size N. AIC requires $C(N)$ equals constant 2, BIC takes $C(N) = logN$, and HQ has $C(N) = 2loglogN$.

Classification rule and error rate control

Given the estimation of the semi-nonparametric mixture model (5) parameters the posterior probability of protein i being functionally inconsistent is calculated by

$$P(R_i = 1 \mid z_i) = \frac{\hat{\lambda}_1 \widehat{f_1}(z_i|\theta_1)}{\hat{\lambda}_0 \widehat{f_0}(z_i|\theta_0) + \hat{\lambda}_1 \widehat{f_1}(z_i|\theta_1)}, \quad (21)$$

where, $\hat{\lambda}_0, \hat{\lambda}_1 \hat{f}_0(z_i \mid \theta_0)$, and $\hat{f}_1(z_i \mid \theta_1)$ are the estimates of λ_0, λ_1, $f_0(z_i|\theta_0)$, and $f_1(z_i|\theta_1)$, respectively. The posterior probability of protein i being functionally consistent is then obtained by

$$P(R_i = 0 \mid z_i) \quad = \quad 1 - P(R_i = 1 \mid z_i). \quad (22)$$

The classification rule that specifies protein i as functionally inconsistent protein is defined as

$$P(R_i = 1 \mid z_i) \geq c^*, \quad (23)$$

where c^* is the critical value. The selection of the critical value c^* is determined by evaluating the estimated false discovery rate (FDR) in Equation (27) and the estimated false non-discovery rate (FNR) in Equation (28). As in Newton *et al.* [40], FDR is estimated by

$$\widehat{FDR} = \frac{\sum_{i=1}^{m} P(R_i = 0 \mid z_i)\delta_i}{\sum_{i=1}^{m} \delta_i}. \quad (24)$$

Similarly, FNR is estimated by

$$\widehat{FNR} = \frac{\sum_{i=1}^{m} P(R_i = 1 \mid z_i)(1 - \delta_i)}{\sum_{i=1}^{m} (1 - \delta_i)}. \quad (25)$$

The indicator δ_i is used for declaring protein i as functionally inconsistent protein by the classification rule (23) for the specific critical value c^*. For any specific protein microarray experiment, the misclassification penalty can be specified. The critical value is obtained by minimizing the following error:

$$\gamma \widehat{FDR} \frac{d}{m} + (1 - \gamma) \widehat{FNR} \frac{m - d}{m}, \quad (26)$$

where $\gamma \in [0, 1]$ is the penalty for false positive, $(1 - \gamma)$ is the penalty for false negative, and d is the number of declared inconsistent proteins by the critical value c^*.

Table 2 Classification outcomes: consistent and inconsistent proteins

	Classified as consistent	Classified as inconsistent	Total
Consistent	U	V	m_0
Inconsistent	T	S	$m - m_0$
Total	$m - R$	R	m

The total number of proteins is denoted by m. m_0 is the number of consistent proteins. R is the total number of classified inconsistent proteins. U, V, T, S are the corresponding number of classification results.

Classification error rates

Suppose there are m proteins (of which m_0 proteins are truly consistent) that need to be simultaneously classified as consistent and inconsistent (Table 2). Let R (an observable variable) denote the number of classified inconsistent proteins, while U, V, S, T are unobservable variables. Similar to the false discovery rate (FDR) [41] and false non-discovery rate [42] proposed for multiple testing problems, the misallocation error rates: false discovery rate (FDR) and false non-discovery rate (FNR) are defined as follows. False discovery rate [41], the expected proportion of falsely classified inconsistent proteins among all classified inconsistent proteins, can be represented by

$$E(Q) = E(\frac{V}{R} \mid R > 0)P(R > 0). \quad (27)$$

False non-discovery rate [42], the proportion of falsely classified consistent proteins among all classified consistent proteins, can be represented by

$$E(N) = E(\frac{T}{m - R} \mid m - R > 0)P(m - R > 0). \quad (28)$$

Acknowledgements

We wish to thank Brian B. Haab (The Van Andel Research Institute) for stimulating discussions on exploring the number of replicates in design of experiments.

Author details

[1]Center for Biostatistics, The Ohio State University, Columbus, OH 43221, USA. [2]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.

Authors' contributions

RWD initiated the interest in developing statistical methods for protein microarray data, coordinated the research, and finalized the manuscript. LY developed the method, conducted the analysis and the simulations, and drafted the original manuscript. Both authors read and approved the final manuscript.

Received: 11 April 2010 Accepted: 28 September 2010
Published: 28 September 2010

References

1. Halvorsen O, Oyan A, Bo T, Olsen S, Rostad K, Haukaas S, Bakke A, Marzolf B, Dimitrov K, Stordrange L, Lin B, Jonassen I, Hood L, Akslen L, Kalland K: **Gene expression profiles in prostate cancer: association with patient subgroups and tumour differentiation.** *International Journal of Oncology* 2005, **26**:329-336.
2. Lee S, Huang K, Palmer R, Truong V, Herzlinger D, Kolquist K, Wong J, Paulding C, Yoon S, Gerald W, Oliner J, Haber D: **The Wilms tumor suppressor WT1 encodes a transcriptional activator of amphiregulin.** *Cell* 1999, **98**:663-673.
3. Nakahara H, Nishimura S, Inoue M, Hori G, Amari S: **Gene interaction in DNA microarray data is decomposed by information geometric measure.** *Bioinformatics* 2003, **19**:1124-1131.
4. Darvish A, Najarian K: **Prediction of regulatory pathways using rnRNA expression and protein interaction data: application to identification of galactose regulatory pathway.** *Biosystems* 2006, **83**:125-135.
5. Gygi S, Rochon Y, Franza B, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Molecular Cell Biology* 1999, **19**:1720-1730.
6. Lueking A, Horn M, Eickhoff H, Bussow K, Lehrach H, Walter G: **Protein microarrays for gene expression and antibody screening.** *Anal Biochem* 1999, **270**:103-111.
7. Ge H: **UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions.** *Nucleic Acids Res* 2000, **28**:e3.
8. MacBeath G, Schreiber S: **Printing proteins as microarrays for high-throughput function determination.** *Science* 2000, **289**:1760-1763.
9. Zhu H, Klemic J, Chang S, Bertone P, Casamayor A, Klemic K, Smith D, Gerstein M, Reed M, Snyder M: **Analysis of yeast protein kinases using protein chips.** *Nature Genetics* 2000, **26**:283-289.
10. Kusnezow W, Banzon V, Schroder C, Schaal R, Hoheisel J, Ruffer S, Luft P, Duschl A, Syagailo Y: **Antibody microarray-based profiling complex specimens: systematic evaluation of labeling strategies.** *Proteomics* 2007, **7**:1786-1799.
11. Domnanich P, Sauer U, Pultar J, Preininger C: **Protein microarray for the analysis of human melanoma biomarkers.** *Sensors and Actuators B: Chemical* 2009, **139**:2-8.
12. Rimini R, Schwenk J, Sundberg M, Sjoberg R, Klevebring D, Gry M, Uhlen M, Nilsson P: **Validation of serum protein profiles by a dual antibody array approach.** *Journal of Proteomics* 2009, **73**:252-266.
13. Yang Y, Speed T: **Design issues for cDNA microarray experiments.** *Nature Reviews - Genetics* 2002, **3**:579-588.
14. Hamelinck D, Zhou H, Li L, Verweij C, Dillon D, Feng Z, Costa J, Haab B: **Optimized normalization for antibody microarrays and applications to serum-protein profiling.** *Molecular and Cellular Proteomics* 2005, **4**:773-784.
15. Daly D, Anderson K, Seurynck-Servoss S, Gonzalez R, White A, Zangar R: **An Interal Calibration Method for Protein-Array Studies.** *Statistical Applications in Genetics and Molecular Biology* 2010, **9**:Article 14.
16. Sreekumar A, Nyati M, Varambally S, Barrette T, Ghosh D, Lawrence S, Chinnaiyan A: **Profiling of cancer cells using protein microarrays: Discovery of novel radiation-regulated proteins.** *Cancer Research* 2001, **61**:7585-7593.
17. Miller J, Zhou H, Kwekel J, Cavallo R, Burke J, Butler E, Teh B, Haab B: **Antibody microarray profiling of human prostate cancer sera: Antibody screening and identification of potential biomarkers.** *Proteomics* 2003, **3**:56-63.
18. Belov L, Mulligan S, Barber N, Woolfson A, Scott M, Stoner K, Chrisp J, Sewell W, Bradstock K, Bendall L, Pascovici D, Thomas M, Erber W, Huang P, Sartor M, Young G, Wiley J, Juneja S, Wierda W, Green A, Keating M, Christopherson R: **Analysis of human leukaemias and lymphomas using extensive immunophenotypes from an antibody microarray.** *British Journal of Haematology* 2006, **135**:184-197.
19. Ingvarsson J, Wingren C, Carlsson A, Ellmark P, Wahren B, Engstrom G, Harmenberg U, Krogh M, Peterson C, Borrebaeck C: **Detection of pancreatic cancer using antibody microarray-based serum protein profiling.** *Proteomics* 2008, **8**:2211-2219.
20. Han M, Oh Y, Kang J, Kim Y, Seo S, Kim J, Park K, Kim H: **Protein profiling in human sera for identification of potential lung cancer biomarkers using antibody microarray.** *Proteomics* 2009, **9**:5544-5552.
21. Song Q, Liu G, Hu S, Zhang Y, Tao Y, Han Y, Zeng H, Huang W, Li F, Chen P, Zhu J, Hu C, Zhang S, Li Y, Zhu H, Wu L: **Novel autoimmune hepatitis-specific antoantigens identified using protein microarray technology.** *Journal of proteome research* 2010, **9**:30-39.
22. Lin L: **A concordance correlation coefficient to evaluate reproducibility.** *Biometrics* 1989, **45**:255-268.
23. Lin L: **Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence.** *Statistics in Medicine* 2000, **19**:255-270.
24. Lin L, Hedayat A, Sinha B, Yang M: **Statistical methods in assessing agreement: models, issues, tools.** *Journal of the American Statistical Association* 2002, **97**:257-270.
25. Lin L, Chinchilli V: **Rejoinder to the letter to the editor from Atkinson and Nevill.** *Biometrics* 1997, **53**:777-778.
26. Efron B, Tibshirani R, Storey J, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *J Amer Statist Assoc* 2001, **96**:1151-1160.
27. Pan W, Lin J, Le C: **A mixture model approach to detecting differentially expressed genes with microarray data.** *Funct Integr Genomics* 2003, **3**:117-124.
28. Akaike H: **Information theory and an extension of the maximum likelihood principle.** *2nd International Symposium on Information Theory* 1973, 473-476.
29. Schwartz S: **Estimating the dimension of a model.** *Annals of Statistics* 1978, **6**:461-464.
30. Hannan E: **Rational transfer function approximation.** *Statistical Science* 1987, **2**:1029-1054.
31. Zhang D, Davidian M: **Linear mixed models with flexible distributions of random effects for longitudinal data.** *Biometrics* 2001, **57**:795-802.
32. Zhou H, Bouwman K, Schotanus M, Verweij C, Marrero J, Dillon D, Costa J, Lizardi P, Haab B: **Two-color, rolling-circle amplification on antibody microarrays for sensitive, multiplexed serum-protein measurements.** *Genome Biology* 2004, **5**:R28.
33. Yu L: **Statistical issues in protein microarray analysis.** *PhD thesis* Purdue University, West Lafayette, IN, USA 2006.
34. Dempster A, Laird N, Rubin D: **Maximum likelihood from incomplete data via the EM algorithm.** *Journal of the Royal Statistical Society. Series B* 1977, **39**:1-38.
35. Lange K: **A quasi-newton acceleration of the EM algorithm.** *Statistica Sinica* 1995, **5**:1-18.
36. Gallant A, Nychka D: **Seminonparametric maximum likelihood estimation.** *Econometrica* 1987, **55**:363-390.
37. Davidon W: **Variable metric methods for minimization.** *AEC Research and Development Report ANL-5990, Argonne National Laboratory* 1959.
38. Ledwina T: **Data-driven version of Neyman's smooth test of fit.** *Journal of the American Statistical Association* 1994, **89**:1000-1005.
39. McLachlan G: **On bootstrapping the likelihood ratio test statistics for the number of components in a normal mixture.** *Journal of the Royal Statistical Society Series C* 1987, **36**:318-324.
40. Newton M, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture method.** *Biostatistics* 2004, **5**:155-176.
41. Storey J: **A direct approach to false discovery rates.** *Journal of the Royal Statistical Society, Series B* 2002, **64**:479-498.
42. Genovese C, Wasserman L: **Operating characteristics and extensions of the false discovery rate procedure.** *Journal of Royal Statistical Society, Ser B* 2002, **64**:499-517.

doi:10.1186/1471-2105-11-486
Cite this article as: Yu and Doerge: **A semi-nonparametric mixture model for selecting functionally consistent proteins.** *BMC Bioinformatics* 2010 **11**:486.