Classification-Based Quantitative Analysis of Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) Data

Seongho Kim^{1,2,*}, Nicholas Carruthers^{3,4}, Joohyoung Lee^{1,5}, Sreenivasa Chinni^{6,&}, and Paul Stemmer^{3,4,&}

¹Biostatistics Core, Karmanos Cancer Institute, ²Department of Oncology, ³Proteomics Core, Karmanos Cancer Institute, ⁴Institute of Environmental Health Sciences, ⁵Department of Family Medicine and Public Health Sciences, ⁶Department of Urology, School of Medicine, Wayne State University, Detroit, MI 48201.

[&]Equally contributed.

*Corresponding author's email: kimse@karmanos.org

Abstract:

Background and Objective: Stable isotope labeling by amino acids in cell culture (SILAC) is a practical and powerful approach for quantitative proteomic analysis. A key advantage of SILAC is the ability to simultaneously detect the isotopically labeled peptides in a single instrument run and so guarantee relative quantitation for a large number of peptides without introducing any variation caused by separate experiment. However, there are a few approaches available to assessing protein ratios and none of the existing algorithms pays considerable attention to the proteins having only one peptide hit.

Methods: We introduce new quantitative approaches to dealing with SILAC protein-level summary using classification-based methodologies, such as Gaussian mixture models with EM algorithms and its Bayesian approach as well as K-means clustering. In addition, a new approach is developed using Gaussian mixture model and a stochastic, metaheuristic global optimization algorithm, particle swarm optimization (PSO), to avoid either a premature convergence or being stuck in a local optimum.

Results: Our simulation studies show that the newly developed PSO-based method performs the best among others in terms of F1 score and the proposed methods are further demonstrated the ability of detecting potential markers through real SILAC experimental data.

Conclusions: No matter how many peptide hits the protein has, the developed approach can be applicable, rescuing many proteins doomed to removal. Furthermore, no additional correction for multiple comparisons is necessary for the developed methods, enabling direct interpretation of the analysis outcomes.

Keywords: Classification; mass spectrometry; particle swarm optimization; proteomics; SILAC.

1. Introduction

One widely used quantitative proteomics method is stable isotope labeling that either introduces a mass difference between the proteins from two proteomes or provides an internal standard for relative quantification [1]. A practical and powerful approach for stable isotope labeling is stable isotope labeling of amino acids in cell culture (SILAC), introduced first by the laboratory of Matthias Mann [2]. This metabolic labeling strategy uses stable isotope-labeled amino acids in growth medium to uniformly incorporate stable isotopes into all the proteins in cell culture samples that will be subjected to quantitative proteomic analysis. Owing to the nature of metabolic labelling, it allows mixing of labelled and unlabeled cells so that subsequent fractionation and purification steps will not introduce any errors in quantitation [3].

SILAC allows the chemically identical peptides from heavy isotope and light isotope labeled samples to be differentiated by mass of the peptides. The abundance changes in proteins between different physiological conditions are measured by the ratios of the integrated single intensities of corresponding peptides, enabling straightforward quantitative comparison in a single experiment [4]. A key advantage of SILAC is the ability to detect simultaneously the isotopically labeled peptides in a single instrument run and so guarantee relative quantitation for a large number of peptides without introducing any variation caused by separate experiments.

There are various software packages, in either open source or commercially available, for stable isotope labeling proteomics, such as MaxQuant [5], XPRESS [6], Isoquant [7], Proteome Discoverer [8], etc. However, these informatics pipelines mainly focus on automated data extraction and subsequent data analysis to refine the data, such as, peak detection, peptide/protein identification, data normalization, summarization of peptide/protein ratios, lacking a statistical approach to assessing protein ratios. The most common method to identify

differentially abundant proteins is the use of a universal fold-change threshold, ignoring the differences in variability among experiments [9]. A simple method to account for both systematic variation and multiple testing is the use of standard student's t tests along with either Bonferroni correction or FDR control, which, for example, is implemented in Perseus software [10], but it requires large numbers of replicates or peptides for each protein.

To date, three algorithms have been developed to resolve the aforementioned issues primarily on quantitative analysis of SILAC data. Margolin et al. [9] introduced empirical Bayes modeling to the protein level ratio data generated from SILAC. Quantitative Proteomics p-value Calculator (QPPC) [11], a web-based tool (<http://qppc.di.uq.edu.au/>), was developed based on the permutation test introduced by Nguyen et al. [12]. These two methods rely essentially on the protein-level ratios summarized, discarding the reproducibility information available from the peptide-level ratios. For this reason, Koh et al. [4] recently established a mixture model-based approach, EBprot, to incorporate the reproducibility of peptide-level intensity ratios into FDR. EBprot uses the peptide-protein hierarchy structure by treating peptide ratios as repeated measurements of corresponding protein ratios.

In a SILAC experiment, it is highly unlikely that all proteins have two or more corresponding peptide hits and, in practice, many proteins have only one peptide hit, often times more than 20%, varied from experiment to experiment. Furthermore, this number becomes worse if we consider the proteins that have more than one peptide ratio available (i.e., often more than 50% of proteins that have only one peptide ratio available; e.g., see Supplementary Information Figure S1). However, none of the above algorithms pays prudent attention to the proteins having only one peptide hit, and these proteins often are discarded before analysis as having insufficient confidence to be considered. Note that Nguyen et al. [12] achieved the quantitative analysis of

the proteins with single peptide using a quantile-based approach, which does not duly isolate the null distribution.

In this work, we introduce a quantitative approach to dealing with protein-level summary using classification-based methodologies. Unlike existing methods, our approach depends mainly on the protein ratio summary and is not restricted only to the proteins with two or more peptide hits. In other words, no matter how many peptide hits the protein has, the developed approach can be applicable, rescuing many proteins doomed to removal. Another advantage of this approach is that there is no need for multiple testing corrections, enabling direct interpretation of the analysis outcomes. In particular, our approach uses Gaussian mixture model and a stochastic, metaheuristic global optimization algorithm, particle swarm optimization (PSO), to avoid either a premature convergence or being stuck in a local optimum. To evaluate the proposed methods on the quantitative analysis of SILAC data, we perform simulation studies and a real-world application using lipid rafts isolated from SILAC labeled PC3 prostate cancer cells with the CXCR4 chemokine receptor either overexpressed or underexpressed and further validate the proposed methods using additional simulation studies and a real experimental SILAC data obtained from Cox and Mann [5].

2. Methods

Experimental data

PC3 cells were grown in RPMI medium containing either heavy or natural isotopic lysine-and-arginine. Heavy isotope-labeled cells were induced to overexpress CXCR4 and natural isotopic cells' CXCR4 expression was reduced by shRNA. Cells were harvested and lipid rafts were isolated using a sucrose density gradient. Lipid rafts from the heavy and natural isotopic cells

were then mixed in equal quantities. The experiment was also carried out in the reverse orientation where natural isotopic cells overexpressed CXCR4 and heavy isotope labeled cells underexpressed CXCR4.

Lipid raft proteins were precipitated using TCA, resolubilized in 6M Urea with 2% SDS and then electrophoresed into an SDS-PAGE gel. A single protein band was excised, reduced alkylated and digested in-gel using trypsin. The resulting peptides were analyzed by LC-MS/MS using a Thermo Q-Exactive mass spectrometer. Peptides were separated on a 25 cm long Acclaim PepMap C18 column via a 1 hour 3 to 30% acetonitrile gradient and the top 15 ions in each MS scan were selected for MS/MS. Spectra were matched to peptides by searching the Uniprot reviewed, canonical human database that was downloaded on 06.24.2014 (20207 entries) using MaxQuant version 1.5.2.8 [5](Cox and Mann, 2008). Default parameters were used except that proteins were accepted on the basis of a single peptide, rather than at least 2. Match between runs and re-quantification of peptides found only in a single isotope state were enabled. Peptide matches from reversed or contaminant proteins were removed before further analysis.

Similar parameters were used for searches and optimal settings were used so as to allow comparison of spectra identified at 1% global FDR Protein FDR = 0.05 and Peptide FDR = 0.01.

Proteins and Peptides, can be filtered at a preset FDR score (0.01 or 0.05) during “Identify” analysis

Peptide and protein false discovery rate (FDR) was set to one percent.

Particle swarm optimization (PSO) classification

We consider two types of unsupervised clustering algorithms in this work. One is called K-means clustering algorithm [13] and the other is based on Gaussian mixture models. The K-means clustering algorithm can be recognized as a nonparametric approach because it assumes no underlying distribution, while the second type of approach can be considered as a parametric approach due to the use of Gaussian distribution. Nonetheless, a general technique to parameter estimations for both approaches is the expectation-maximization (EM) algorithm [14, 15, 16]. The K-means clustering method, Gaussian mixture models (GMMs) method, and a Bayesian version of GMM (BGMM) are described in Supplementary Information. In this section, we mainly focus on the developed PSO-based clustering.

Suppose there are N protein-level abundances for case and control, $\{x_1, x_2, \dots, x_N\}$ and $\{y_1, y_2, \dots, y_N\}$, respectively, summarized by their peptide-level abundances. Then the protein-level ratios are constructed by $r = x/y$, resulting in N protein-level ratios, $\{r_1, r_2, \dots, r_N\}$, where $r_i, i = 1, \dots, N$, are log-transformed ratios. Note that the protein-level ratio represents the ratio between heavy-labeled and light-labeled proteins. The primary goal of the quantitative analysis of SILAC is to identify the differentially abundant proteins that statistically reject the null hypothesis, $H_0: x_i = y_i$ i.e., $r_i = 0, i = 1, \dots, N$, after multiple comparison correction. The differentially abundant proteins are further categorized into two groups, either down-regulated (i.e., $x < y$ or $r < 0$) or up-regulated (i.e., $x > y$ or $r > 0$). In essence, it assumes that the SILAC ratios are generated from distributions of three groups, $r < 0$, $r = 0$, and $r > 0$. In this regard, the above statistical testing can be interpreted as classifying the ratios into three clusters, down-regulated ($r < 0$), no-difference ($r = 0$), and up-regulated ($r > 0$). This relationship spurs us to consider a classification-based approach to identifying differentially abundant proteins.

Although the EM algorithm is popular in various areas because of its simplicity and stability, it is a local optimization so that convergence to the global optimum cannot be guaranteed or expected. As a consequence, none of the aforementioned approaches will guarantee convergence to the global optimum. To overcome this, a stochastic global optimization, particle swarm optimization (PSO), is employed in addition to the EM algorithm.

PSO, introduced by Kennedy and Eberhart [17], is a population-based global optimization and evolves a group of solutions (particles) stochastically. It was motivated by the behavior of a flock of birds or school of fish in nature. Its main notion is to take advantage of the communication involved in such flocks or schools. PSO initially has a population randomly generated consisting of a set of solutions. Each potential solution or element of the set, which is called a particle, travels with a random velocity to find a solution through the problem space. Each particle's trace in the problem (i.e., search) space is then determined by its own memory of best fittings. Individual particles move towards a stochastically weighted average of these positions, until they converge to the global best solution. It is used to solve a wide array of different optimization problems because of its attractive advantages, such as the ease of implementation and its gradient free stochastic algorithm. It has been proven to be an efficient method for many global optimization problems, and does not suffer from the difficulties encountered by other evolutionary computation techniques. For an overview of PSO and its variants, see [18].

In particular, PSO is employed to enhance the performance of the M-step of EM algorithms. To do this, we maximize the likelihood function in Supplementary Information Equation (S5) given the estimate of Z , Z :

$$L_{\theta|z, r, i=1, \dots, N} = L_{\theta, z| r, i=1, \dots, N} = \prod_{k=1}^K \lambda_k n_k! z_i = k \phi(r_i; \mu_k, \sigma_k), \quad (1)$$

where $z_i \in \{1, 2, 3\}$ is a trinary latent variable; $n_k = \sum_i \mathbb{I}(z_i = k)$, $i = 1, 2, \dots, N$; $N = \sum_k n_k$. In other words, the latent variable z is still estimated by the E-step, while the parameters $\theta = \{\lambda_k, \mu_k, \sigma_k; k = 1, 2, 3\}$ are estimated by the M-step with PSO, in which we call it PSOM-step. To implement PSOM-step, we use the function ‘psoptim’ in the R package pso with the swarm size of 40 and with the following settings:

$$0.001 < \theta_1, \theta_3 < 0.3, \theta_2 = 1 - \theta_1 - \theta_3;$$

$$-9 < \mu_1 < -0.1, 0.1 < \mu_3 < 9, \mu_2 = 0;$$

$$0.0001 < \sigma_i < 20, i = 1, 2, 3.$$

The default settings are used for others. On the other hand, we often observed a Gaussian distribution with a long-tail so that it becomes flat. Therefore, to prevent this, we combine one of the following two constraints with the likelihood function in Equation (1):

$$\phi(0|\mu_1, \sigma_1) - \phi(0|\mu_3, \sigma_3)^2 < \epsilon; \quad (2)$$

$$\phi(0|\mu_1, \sigma_1)^2 + \phi(0|\mu_3, \sigma_3)^2 < \epsilon, \quad (3)$$

where ϵ is a positive small number and $\phi(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2}\right)$. Both constraints aim to prevent a Gaussian distribution from dispersing with a long but thick tail. The constraint in Equation (2) minimizes the square of the difference between two pdfs at zero so that it can allow each pdf to be a large value, while that in Equation (3) minimizes the sum of square of each pdf with forcing each pdf to be zero.

Implementation

To implement the aforementioned methods and the developed PSO-based method, we use the functions available in the statistical software R packages (Version 3.2.3). The K-means

classification is implemented using the function ‘kmeans’ available in the R package *stats*. For Gaussian mixture models and its Bayesian approach, the function ‘mclust’ in the R package *Mclust* is utilized [19]. In addition to the function ‘mclust’, the function ‘normalmixEM’ in the R package *mixtools* is also used for Gaussian mixture models to compare with the function ‘mclust’. To implement the PSOM-step, the function ‘psoptim’ available in the R package *pso* is used for the PSO-based classification. The proposed approaches and the real data analysis are implemented in the R package *cSILAC* available as the Supplementary Information.

Performance criteria

The performances of all the methods are compared by calculating the true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), accuracy, and F1 score of the classification.

As previously stated, the goal of the classification is to classify these N protein-level ratios into three clusters, down-regulated ($r < 0$), no-difference ($r = 0$), and up-regulated ($r > 0$), where there are N protein-level ratios, $\{r_1, r_2, \dots, r_N\}$. Suppose there are P down-regulated ratios, Q no-difference ratios, and R up-regulated ratios, where $N = P + Q + R$.

After classification using one of the aforementioned methods, a 3×3 confusion table will be obtained as depicted in Table 1. Then, using the confusion table in Table 1, the accuracy is computed by

$$\text{Accuracy} = \frac{E + I + MN}{N} \quad (4)$$

However, the performance measures TPR, FPR, PPV, and F1 score are calculated based on a 2×2 confusion table so that these measures cannot be computed directly from Table 1. For

this reason, we convert a 3×3 confusion table into its corresponding 2×2 confusion table in Table

2. As a result, the performance measures are obtained by

$$\text{TPR} = \frac{E+K+G}{E+K+G+MP+R};$$

$$\text{FPR} = \frac{F+L}{Q};$$

$$\text{PPV} = \frac{E+K+G}{E+K+G+MS+U};$$

$$\text{F1 score} = \frac{2 \cdot \text{TPR} \cdot \text{PPV}}{\text{TPR} + \text{PPV}} = \frac{2 \cdot (E+K+G) \cdot (E+K+G)}{(E+K+G+MP+R) + (E+K+G+MS+U)}. \quad (5)$$

Note that TPR is also called recall and PPV precision, and F1 score is the harmonic mean of TPR and PPV [20].

3. Results

We compared the performance of the proposed methods with the aforementioned methods using simulated data and real biological data. As stated before, the methods employed are normEM and Mclust (Gaussian mixture models using the R packages mixtools and Mclust, respectively), Bayes (Bayesian approach), Kmeans (K-means classification), PSO1 (without any constraint), PSO2 (using Equation (2)), and PSO3 (using Equation (3)). In addition to these seven methods, we included the following three methods as controls: FC (fold change-based cut-off of 2.5), QV1 (quantile-based cut-off of 1%), and QV2 (quantile-based cutoff of 5%), resulting in a total of ten approaches. That is, FC and QV, respectively, detects potential markers based on the protein-level ratios and on quantiles [21] of the empirical distribution of the protein-level ratios.

3.1 Simulated data

For the simulated data, we used a mixture of Gaussian distribution that consists of three components, up-regulated, no-difference, down-regulated. Without loss of generality, the means

of each component were fixed as -2.5, 0, and 2.5, respectively. For the standard deviation (SD), the three sets of values were used, (0.5, 0.5, 0.5), (0.5, 0.7, 0.5), and (0.5, 1, 0.5), where $(\sigma_1, \sigma_2, \sigma_3)$ represents the true SDs of each component of up-regulated, no-difference, and down-regulated, respectively. To reflect a real-world data set, the weights of each component were set to 1%, 3%, and 5% for the proportion of both up- and down-regulated proteins, i.e., (0.005, 0.99, 0.005), (0.015, 0.97, 0.015), and (0.025, 0.95, 0.025), respectively, out of 10,000 proteins, where (w_1, w_2, w_3) represents the proportion of proteins belonging to each component of up-regulated, no-difference, and down-regulated, respectively. In other words, of these 10,000 proteins, the number of proteins that are up-regulated, no-difference, and down-regulated is assumed to be (50, 9900, 50), (150, 9700, 150), and (250, 9500, 250), respectively. A total of 500 simulated data sets were generated randomly using a mixture of three Gaussian components for each of nine cases (3 SDs $\times$ 3 weights). The true density plots of each of the nine cases are depicted in Figure 2.

Figure 3 displays the boxplots of F1 scores of each method according to nine cases. In general, PSO2 and PSO3 achieve the best performance among ten methods except for Cases 7-9. In Cases 7-9, either FC or QV1 perform better than the other approaches. However, as can be seen in Figure 2, Cases 7 and 8 are less likely to occur in a real-world problem because two up- and down-regulated distributions are included under the no-difference distribution. For Case 9, PSO2 and PSO3 show a comparable performance to either FC or QV1 although they seem to include more potential outliers. Interestingly, normEM and Mclust use the same method from a theoretical perspective, but Mclust achieves a better performance. Empirical means and SDs of F1 scores are depicted in Supplementary Information Table S1.

The boxplots of accuracies of each method are displayed in Figure 4, and empirical means and SDs of accuracies are depicted in Supplementary Information Table S2. The accuracy was computed using Equation (4). Likewise to F1 scores, PSO2 and PSO3 perform better than the other approaches for Cases 1 to 7. However, QV2 shows the best performance in Case 8, while FC performs the best in Case 9.

The boxplots, empirical means, and SDs of TPRs, FPRs, and PPVs can be found in Supplementary Information Tables S3-S5 and Figures S2-S4. Interestingly, either normEM or Mclust achieve the best performance in TPR across all cases, while either FC or QV2 perform the best in terms of FPR and PPV. Nevertheless, PSO-based methods show the comparable performances for TPR, FPR, and PPV.

Overall, PSO-based methods achieve the highest F1 scores and accuracies and provide consistent and robust performance across all cases in this simulation studies.

3.2 Experimental data

The developed methods were further evaluated using a real-life data set as described in the Methods section. The samples were prepared from prostate cancer cells, PC3 shCXCR4 (CXCR4 knocked down with lentivirus shRNA) and PC3 CXCR4 (CXCR4 overexpressed using lentivirus expressing CXCR4). Then lipid raft micro domains were isolated in these two cells. The objective of this experiment was to identify proteins in lipid raft membrane micro domains which were enriched in CXCR4 expressing cells using the SILAC-based global proteomics approach. In particular, it was planned to identify CXCR4 signaling molecules in lipid raft membrane micro domains that were involved in the invasion and metastasis.

To find potential protein markers that are differentially expressed between PC3 shCXCR4 and PC3 CXCR4, we applied the following nine methods: normEM, Bayes, Kmeans, Mclust, PSO1, FC, QV, PSO2, and PSO3. In particular, FC and QV were used with the cut-off values of 1.5 fold-changes and 1%, respectively. The shCXCR4 cell was labeled with heavy isotopes, while CXCR4 cell was labeled with light isotopes. In this experiment, the SILAC ratio is the ratio of light to heavy. Thus, the up-regulated protein (i.e., SILAC ratio > 1) represents that it overexpressed in CXCR4 cell and the down-regulated protein (i.e., SILAC ratio < 1) indicates that it overexpressed in shCXCR4.

Table 3 displays the results of the SILAC data analysis using the nine methods. PSO1 (Method 5), QV (Method 7), PSO2 (Method 8) and PSO3 (Method 9) declare the least amount of proteins as potential markers, while normEM (Method 1) detected the most proteins as markers. Likewise to the simulation studies, normEM and Mclust (Method 4) perform very differently from each other although their theoretical methodologies are the same. Moreover, normEM appears the least discriminating method because it declares all 308 identified proteins as potential markers.

As stated before, this experimental data set has ~21% (66/308) of proteins that have only one peptide hit and has further ~54% (165/308) of proteins that have only one peptide ratio available. Of the detected markers, the proteins that have one peptide hit ranges from 13% to 25% and the proteins that have one peptide ratio ranges from 54% to 88%, as can be seen in Table 3 and Supplementary Information Table S6.

Figure 5 depicts the Venn-Euler diagram to see the relationship among the sets of the protein markers detected by each method. From this diagram, we can have two types of information. One is the inclusion relationship among the nine methods and the other is the

similarity among the sets. That is, the numbers in the figure indicate the center of each circle and their centers are near to each other if their sets are very similar. The markers detected by Kmeans (Method 3) are similar to those detected by normEM (Method 1), while those of Mclust (Method 4) are similar to those of FC (method 6). All PSO-based methods (Methods 5, 8, 9) have exactly the same center, meaning that they identify the same markers. Further, the markers detected by PSO-based methods are a subset of the markers detected by other methods except QV (Method 7). The more detailed inclusion relationship can be found in Supplementary Information Table S6.

The fitted curves of Gaussian mixture model and the estimated clusters are further investigated as depicted in Figure 6. Not surprisingly, normEM shows the worst fitting performance among the nine methods. In particular, the fitted curve of no-difference component is under those of the other two components, causing some proteins with $r \approx 1$ to be assigned to the up-regulated clusters. In terms of the shape of the fitted curves, those of Bayes are very similar to those of Mclust, although Bayes declares more proteins as potential markers. Except for normEM, Kmeans has the smallest variance of no-difference component, resulting in the largest number of markers. On the other hand, PSO-based methods fit the curve of no-difference component with the largest variance so that they have the smallest number of markers.

The developed PSO-based approaches declare eight proteins as potential markers from the SILAC data. Using Ingenuity Pathway Analysis (IPA; Ingenuity Systems, <http://www.ingenuity.com/>), we could further confirm that these proteins are involved in prostate cancer development. Indeed, we obtained the following five pathways in which those proteins are involved through IPA: (i) caveolar mediated endocytosis signaling, (ii) virus entry via endocytic pathways, (iii) ILK signaling, (iv) LPS/IL-1 mediated inhibition of RXR function and

(v) protein kinase A signaling. The first three are well known cancer related pathways. In the fourth pathway, a decrease of RXR-alpha in the basal cells is used as a biomarker for prostate adenocarcinoma associated field change [22]. It represents an early stage in the prostate cancer development. The fifth protein kinase A pathway is known to induce high level of androgen receptor and prostate specific antigen in the prostate cancer cells in the absence of androgen [23].

3.3 Validation data

Two data sets were applied to additionally validate the developed methods. The first data set was generated based on the estimates of the real data in Section 3.2 and is composed of two cases.

Case I was simulated based on the estimates by Bayes (Method 2) and Case II by PSO3 (Method 9). It was assumed that there were 308 proteins and the means of each component (down-regulated, no-difference, up-regulated) were (-0.53, 0, 1.88) for Case I and (-3.34, 0, 2.82) for Case II, respectively. The second data set was a real SILAC data obtained from Cox and Mann [5].

The density plots for the first data set are depicted in Supplementary Information Figure S5. Unlike the simulated data in Section 3.1 (Figure 2), both cases have the unequal proportions of down- and up-regulated protein ratios (i.e., Case I: 25%: 68%: 7%; Case II: 0.3%: 97%: 2.7%) and Case II has the distribution of the down-regulated protein ratios with very small standard deviation (i.e., 0.0004). Further, the distributions of Case I are not distinguished clearly from each other. The results after 500 simulations are summarized in Supplementary Information Table S7. The boxplots of F1 score and accuracy are depicted in Supplementary Information Figures S6 and S7, respectively. Similar to the earlier simulation in Section 3.1, PSO-based approaches in Case I perform not better than other approaches because the large portion of

distribution is overlapped with each other (Table S7), but they achieve higher PPV and accuracy. In the second case, PSO-based approaches achieve the highest F1 scores and accuracies after FC (Figures S6 and S7).

A total of 4285 proteins were available in the real SILAC data of Cox and Mann [5]. Out of 4285 proteins, 147 proteins were removed because neither names nor p-values (i.e., ‘significance B’ [5]) were available, resulting in a total of 4138 proteins. Following the procedures described in Cox and Mann [5], the ‘significance B’-based p-values were further adjusted using Benjamini-Hochberge FDR (BH-FDR) method [24]. We considered the results with BH-FDR < 5% as the true values. That is, the proteins with the ratio > 1 and BH-FDR < 5% were considered as the true up-regulated proteins, those with the ratio < 1 and BH-FDR < 5% as the true down-regulated proteins, and other proteins as the true no-difference proteins. Accordingly, there were 86 down-regulated, 3973 no-difference, and 79 up-regulated proteins.

Based on those empirical true values, we compared the performance of each of nine methods that were used in Section 3.2. The results are shown in Table 4 and Figure 7. Interestingly, normEM, Bayes, Mclust, and PSO1 achieve more than 99% of TPR, but these are in fact due to that the distributions of down- and up-regulated protein ratios are overlapping with that of no-difference ratios, as can be seen in Figure 7. In particular, the predicted density plot of down-regulated protein ratio by normEM is much different from others in the sense that it has very small standard deviation. Consequently, their accuracies are lower than those of other methods (Table 4). In case of FPR and PPV, PSO3 performs the best among the methods followed by QV. The highest F1 score is achieved by FC, while PSO3 shows the highest accuracy among the methods. Overall, FC, QV, PSO2, and PSO3 give much better performance in terms of F1 score and accuracy.

In summary, for both validation data sets, PSO-based approaches (in particular, PSO2 and PSO3) show a robust performance among the methods in terms of F1 score and accuracy. In addition, PSO-based approaches (in particular, PSO3) achieve a lower FDR (i.e., a higher PPV since $PPV = 1 - FDR$) than other approaches.

4. Discussion

We proposed a new quantitative approach to dealing with SILAC protein-level summary using four classification-based methodologies which are normEM/Mclust, Byes, Kmeans, and PSO. Among them, Kmeans is a nonparametric approach, while others are parametric approaches using Gaussian mixture models. In particular, we also developed three PSO-based approaches which use a stochastic global optimization.

The normEM and Mclust methods use the same methodology based on Gaussian mixture model and EM algorithms. However, as observed in both simulation and real data analyses, their performances are very different from each other. This difference is due to their different numerical approaches to estimating parameters and Mclust shows a better performance in this study, implying that it is a more stable and robust algorithm than normEM.

As shown in the fitting of normEM on the real SILAC data (Figures 6 and 7), normEM is often vulnerable to local optima because it is based on a local optimization. To overcome this difficulty (i.e., to avoid either a premature convergence or being stuck in a local optimum), we developed a stochastic global optimization approach using a metaheuristic global optimization algorithm PSO and further introduced two more versions using Equations (2)-(3). In the simulation studies, comparing to PSO1, the better performance of PSO2 and PSO3 over PSO1 is case-dependent. In general, there is no significant difference in performance when the true

components of up- and down-regulated are well separated. Indeed, all the PSO-based approaches detect the exactly same proteins in the real SILAC data analysis in Section 3.2.

In contrast to the other methods, our approach relies mainly on the protein ratio summary and does not require proteins with two or more peptide hits. As a result, no matter how many peptide hits the protein has, the developed methods can be employed, rescuing many proteins doomed to removal.

On the other hand, the inclusion of proteins with only one peptide hit might cause an increased false positive rate due to their low quality of quantitative information. However, the global (i.e., untargeted) proteomics approach is used as an initial screening or a pilot or feasibility study to find potential markers as a pre-step for the following targeted proteomics. In this sense, the more proteins, the better to achieve this fundamental goal in the global proteomics. Besides, we observe that proteins that have only one peptide hit or ratio cannot be easily ignored because they represent a significant proportion of the real SILAC data (e.g., see Figure 1 and Supplementary Information Figure S1). Furthermore, many of these proteins have strong signal, i.e., many of their ratios are located in either left-most or right-most (e.g., see Supplementary Information Table S6), and strong signals provide greater confidence in the information and so more important than weak signals. Therefore, deletion of these proteins ends up removing strong and important signals, causing an increase in the false discovery rate. For these reasons, the proposed approach will be of benefit to global proteomics especially to decrease the false discovery rate.

Another advantage of our approach is that there is no need for multiple testing corrections which is one of the most compelling and convenient characteristics of the proposed approach. Namely, this advantage enables users to carry out direct interpretation of the analysis outcomes.

In fact, the developed methods achieved the higher F1 score and accuracy without multiple comparison correction in the simulated data in Section 3.1 (Supplementary Information Tables S1 and S2). In particular, FDR is equal to one minus PPV, so the highest PPV reflects the lowest FDR. Supplementary Information Table S5 shows that the highest PPV for each case is more than 97% except for Cases 4, 7-9 that the true components of up- and down-regulated are not clearly distinguished. In other words, the best methods in Cases 1-3, 5-6 achieve FDR less than 3%.

Each of nine methods took 0.55 (normEM), 0.23 (Bayes), 0.03 (Kmeans), 0.22 (Mclust), 56.69 (PSO1), 0.02 (FC), 0.02 (QV), 8.57 (PSO2), and 26.20 (PSO3) seconds to analyze the SILAC real data with 4138 proteins obtained from Cox and Mann [5]. This shows that PSO-based approaches are much more computationally expensive than others, but at the same time it demonstrates that their amount of computation time is acceptable and applicable in practice as all the computation times are less than one minute.

Our simulation studies clearly demonstrate that the PSO-based approach achieves the best performance among the four methods in terms of F1 scores and accuracies. In the real SILAC experimental data analyses, the PSO-based approach declares the smallest number of proteins as potential markers that are further confirmed by pathway analysis. In addition, Mclust might also be considered as an alternative because of its comparable performance in the simulation studies and its reasonable fitting curves in the real data analysis.

Acknowledgements

This work has been partially supported by NSF grant DMS-1312603 and NIH grant 5R01CA151557. The Biostatistics Core is supported in part by NIH Cancer Center Support

Grant P30 CA022453 to the Karmanos Cancer Institute at Wayne State University. The Proteomics Core is supported in part by NIH grants P30 CA022453, P30 ES020957 and S10 OD010700.

References

- [1] S.E. Ong, M. Mann, A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC), *Nature Protocols* 1 (2007) 2650-2660.
- [2] S.E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics, *Mol. Cell Proteomics* 1 (2002) 376-386.
- [3] M. Mann, Functional and quantitative proteomics using SILAC, *Nature Reviews Molecular Cell Biology* 7 (2006) 952-958.
- [4] H.W.L. Koh, H.L. Swa, D. Fermin, S.G. Ler, J. Gunaratne, H. Choi, EBprot: Statistical analysis of labeling-based quantitative proteomics data, *Proteomics* 15 (2015) 2580-91.
- [5] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, *Nat. Biotechnol.* 26 (2008) 1367–1372.
- [6] D.K. Han, M. Eng, H. Zhou, R. Aebersold, Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry, *Nature Biotechnology* 19 (2001) 946-951.
- [7] Z. Liao, Y. Wan, S.N. Thomas, A.J. Yang, IsoQuant: a software tool for stable isotope labeling by amino acids in cell culture-based mass spectrometry quantitation, *Anal. Chem.* 15 (2012) 4535-4543.
- [8] E.W. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, *et al.*, A guided tour of the Trans-Proteomic Pipeline, *Proteomics* 10 (2010) 1150–1159.

- [9] A.A. Margolin, A.A. S-E. Ong, M. Schenone, R.Gould, S.L. Schreiber, S.A. Carr, *et al.*, Empirical Bayes Analysis of Quantitative Proteomics Experiments, *PLoS ONE* 4 (2009) e7454.
- [10] J. Cox, M. Mann, 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data, *BMC Bioinformatics* 13 (2012) S12
- [11] D. Chen, A. Shah, H. Nguyen, D. Loo, K.L. Inder, M.M. Hill, Online quantitative proteomics p-value calculator for permutation-based statistical testing of peptide ratios, *J. Proteome. Res.* 13 (2014) 4184-91.
- [12] H. Nguyen, I. Wood, M. Hill, A robust permutation test for quantitative SILAC proteomics experiments, *J. Integr. OMICS*. 2 (2012) 80-93.
- [13] S.P. Lloyd, Least squares quantization in PCM, *IEEE Transactions on Information Theory* 28 (1982) 129-137.
- [14] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B* 39 (1977) 1-38.
- [15] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
- [16] T. Benaglia, D. Chauveau, D.R. Hunter, D. Young, mixtools: An R package for analyzing finite mixture models, *Journal of Statistical Software* 32 (2009) 1-29.
- [17] J. Kennedy, R.C. Eberhart, Particle swarm optimization, *Proc. of IEEE International Conference on Neural Networks (ICNN)* Vol.IV, pp.1942-1948, Perth, Australia, 1995.
- [18] A. P. Engelbrecht, *Particle Swarm Optimization*, in *Computational Intelligence: An Introduction*, Second Edition, John Wiley & Sons, Ltd, Chichester, UK, 2007.
doi: 10.1002/9780470512517.ch16

- [19] C. Fraley, A.E. Raftery, MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering, *Technical Report* no. 504, Department of Statistics, University of Washington, 2006.
- [20] Goutte C , Gaussier E . A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *Adv Inform Retrieval* 2005, 3408:345–359.
- [21] R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980
- [22] G.E. Mao, V.E. Reuter, C. Cordon-Cardo, G. Dalbagni, H.I. Scher, J.B. deKernion, Z.F. Zhang, J. Rao, Decreased retinoid X receptor- α protein expression in basal cells occurs in the early stage of human prostate cancer development, *Cancer Epidemiol Biomarkers Prev.* 13 (2014) 383.
- [23] M. Sarwar, S. Sandberg, P.A. Abrahamsson, J.L. Persson, Protein kinase A (PKA) pathway is functionally linked to androgen receptor (AR) in the progression of prostate cancer, *Urologic Oncology: Seminars and Original Investigations* 32 (2014) 25.e1-25.e2.
- [24] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a pratical and powerful approach to multiple testing, *J. R. Statist. Soc. B* 57 (1995) 289-300.

Figure Legends.

Figure 1. The histogram of SILAC ratios generated from an experimental data set. The x-axis is the $\log_2$ ratio between heavy-labeled and light-labeled proteins. The gray histogram represents the entire 308 ratios, while the green and the blue histograms indicate the proteins that have two or more peptide hits (242) and that have only one peptide hit (66), respectively. The red dotted line represents the two fold difference.

Figure 2. The density plots of simulated data. From the top-left to the bottom-right, the index of each plot is 'Case1', 'Case2', 'Case3', 'Case4', 'Case5', 'Case6', 'Case7', 'Case8', and 'Case9'. The null distribution is depicted by green color, while those of the up and down regulated proteins are by red and blue colors, respectively. The gray color lines represent the mixture distribution. The numbers in parentheses in the left-most y-axis display the standard deviations of each distribution, Up-regulated, No-difference, and Down-regulated, respectively, and the numbers in parentheses in the x-axis show the percentage of proteins in each distribution, Up-regulated, No-difference, and Down-regulated, respectively.

Figure 3. The boxplots of F1 scores of each method according to nine cases.

Figure 4. The boxplots of accuracies of each method according to nine cases.

Figure 5. The Venn-Euler diagram. The numbers in the diagram indicate the center of each circle. Note that PSO-based methods (i.e., Methods 5, 8 and 9) share the exactly same center.

Figure 6. The fitted curves of a Gaussian mixture distribution and the estimated clusters.

The gray line indicates the fitted curve of Gaussian mixture distribution. The red, green, and blue dotted lines represent the fitted curves of each component corresponding to down-regulated ($r < 1$), no-difference ($r = 1$), and up-regulated ($r > 1$), respectively. The circles indicate the membership of each protein corresponding to each fitted component. The nonparametric approaches (Kmeans, FC, and QV) do not have the fitted curves.

Figure 7. The fitted curves of a Gaussian mixture distribution and the estimated clusters using the real SILAC data of Cox and Mann [5]. The gray line indicates the fitted curve of Gaussian mixture distribution. The red, green, and blue dotted lines represent the fitted curves of each component corresponding to down-regulated ($r < 1$), no-difference ($r = 1$), and up-regulated ($r > 1$), respectively. The circles and '+' indicate the membership of each protein corresponding to each fitted component and corresponding to Cox and Mann [5]'s results, respectively. The nonparametric approaches (Kmeans, FC, and QV) do not have the fitted curves.

Tables.

Table 1. A 3×3 confusion table.

		Decision			Total
		Down ¹	No ²	Up ³	
True	Down	E	H	K	P
	No	F	I	L	Q
	Up	G	J	M	R
Total		S	T	U	N

¹, down-regulated ($r < 0$); ², no-difference ($r = 0$); ³, up-regulated ($r > 0$).

Table 2. A 2×2 confusion table corresponding to the 3×3 confusion table in Table 1.

		Decision		Total
		No ¹	Down ² +Up ³	
True	No	I	F+L	Q
	Down+Up	H+J	E+K+G+M	P+R
Total		T	S+U	N

¹, no-difference ($r = 0$); ², down-regulated ($r < 0$); ³, up-regulated ($r > 0$).

Table 3. The number of potential protein markers detected by the proposed methods on a real experimental SILAC data.

	Data¹	Methods²								
		1	2	3	4	5	6	7	8	9
Up	144	68	15	61	11	7	25	4	7	7
Down	164	240	37	85	16	1	16	4	1	1
Total	308	308	52	146	27	8	41	8	8	8
SinglePep³	66	66	12	33	4	1	10	2	1	1
SingleRatio⁴	165	165	34	79	17	7	28	7	7	7

¹, based on the observed ratio, up if the ratio >1 and down if the ratio < 1; ², 1:normEM, 2:Bayes, 3:Kmeans, 4:Mclust, 5:PSO1, 6:FC, 7:QV, 8:PSO2, 9:PSO3; ³, the number of proteins that have only one peptide hit among the detected markers; ⁴, the number of proteins that have only one peptide ratio among the detected markers.

Table 4. Performance results by each method using the real SILAC data of Cox and Mann [5].

Method	TPR	FPR	PPV	F1 score	Accuracy
normEM	1.0000	0.2976	0.0783	0.1452	0.7098
Bayes	1.0000	0.3491	0.0675	0.1265	0.6595
Kmeans	0.6364	0.1717	0.1334	0.2206	0.8207
Mclust	1.0000	0.3444	0.0684	0.1280	0.6641
PSO1	0.9901	0.1930	0.1138	0.2041	0.8115
FC	0.5212	0.0083	0.7227	0.6056	0.9729
QV	0.4303	0.0033	0.8452	0.5703	0.9741
PSO2	0.4788	0.0053	0.7900	0.5962	0.9741
PSO3	0.4061	0.0015	0.9178	0.5630	0.9749

Figure 1. The histogram of SILAC ratios generated from an experimental data set. The x-axis is the $\log_2$ ratio between heavy-labeled and light-labeled proteins. The gray histogram represents the entire 308 ratios, while the green and the blue histograms indicate the proteins that have two or more peptide hits (242) and that have only one peptide hit (66), respectively. The red dotted line represents the two fold difference.

Figure 2. The density plots of simulated data. From the top-left to the bottom-right, the index of each plot is 'Case1', 'Case2', 'Case3', 'Case4', 'Case5', 'Case6', 'Case7', 'Case8', and 'Case9'. The null distribution is depicted by green color, while those of the up and down regulated proteins are by red and blue colors, respectively. The gray color lines represent the mixture distribution. The numbers in parentheses in the left-most y-axis display the standard deviations of each distribution, Up-regulated, No-difference, and Down-regulated, respectively, and the numbers in parentheses in the x-axis show the percentage of proteins in each distribution, Up-regulated, No-difference, and Down-regulated, respectively.

Figure 3. The boxplots of F1 scores of each method according to nine cases.

Figure 4. The boxplots of accuracies of each method according to nine cases.

Figure 5. The Venn-Euler diagram. The numbers in the diagram indicate the center of each circle. Note that PSO-based methods (i.e., Methods 5, 8 and 9) share the exactly same center.

Figure 6. The fitted curves of a Gaussian mixture distribution and the estimated clusters.

The gray line indicates the fitted curve of Gaussian mixture distribution. The red, green, and blue dotted lines represent the fitted curves of each component corresponding to down-regulated ($r < 1$), no-difference ($r = 1$), and up-regulated ($r > 1$), respectively. The circles indicate the membership of each protein corresponding to each fitted component. The nonparametric approaches (Kmeans, FC, and QV) do not have the fitted curves.

Figure 7. The fitted curves of a Gaussian mixture distribution and the estimated clusters using the real SILAC data of Cox and Mann [5]. The gray line indicates the fitted curve of Gaussian mixture distribution. The red, green, and blue dotted lines represent the fitted curves of each component corresponding to down-regulated ($r < 1$), no-difference ($r = 1$), and up-regulated ($r > 1$), respectively. The circles and '+' indicate the membership of each protein corresponding

to each fitted component and corresponding to Cox and Mann [5]’s results, respectively. The nonparametric approaches (Kmeans, FC, and QV) do not have the fitted curves.