A Classification-based Quantitative Approach for SILAC Data

Seongho Kim*
Biostatistics Core
Karmanos Cancer Institute
Department of Oncology
Wayne State University
Detroit, MI 48201, USA

Joohyoung Lee
Department of Family Medicine and Public Health Sciences
Wayne State University
Detroit, MI 48201, USA

ABSTRACT

A practical and powerful approach for stable isotope labeling is stable isotope labeling by amino acids in cell culture (SILAC). A key advantage of SILAC is the ability to detecting simultaneously the isotopically labeled peptides in a single instrument run and so guarantees relative quantitation for a large number of peptides without introducing any variation caused by separate experiments. In this work, we introduce a new quantitative approach to dealing with SILAC 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 particular, our approach uses Gaussian mixture model and a stochastic, metaheuristic global optimization algorithm, particle swarm optimization (PSO), to avoid either a premature of convergence or being stuck in a local optimum. Our simulation studies show that the proposed method performs the best in terms of F1 score.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical computing; [**Special Track II**]: Bioinformatics

General Terms

Algorithm, Classification

Keywords

Classification, Gaussian mixture model, Proteomics, PSO, SILAC

1. INTRODUCTION

Stable isotope labeling, which either introduces a mass difference between two proteomes or provides an internal standard for relative quantification, has been widely used as quantitative proteomics methods [12]. A practical and powerful approach for stable isotope labeling is stable isotope labeling by amino acids in cell culture (SILAC), introduced first by the laboratory of Matthias Mann [11], which is a metabolic labeling strategy using stable isotope-labeled amino acids in growth medium for 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 [9].

A key advantage of SILAC is the ability to detecting simultaneously the isotopically labeled peptides in a single instrument run and so guarantees relative quantitation for a large number of peptides without introducing any variation caused by separate experiments.

Several software packages are available for stable isotope labeling proteomics, such as MaxQuant [2], XPRESS [6], Isoquant [8], Proteome Discoverer [4], etc. These mainly focus on automated data extraction and subsequent data analysis to refine the data, however. A universal fold-change threshold is one of the common methods to identify differentially abundant proteins, ignoring the differences in variability among experiments [10]. A simple method to account for both systematic variation and multiple testing is the use of standard student's t tests along with multiple comparison correction, but it requires large numbers of replicates or peptides for each protein.

In this work, we introduce a classification-based approach to dealing with SILAC protein-level summary. Unlike existing methods, our method relies mainly on the protein ratio summary and is not restricted only to the proteins with two or more peptide hits. Moreover, our method uses Gaussian mixture model and a stochastic, metaheuristic global optimization algorithm, particle swarm optimization (PSO), to avoid either a premature of convergence or being stuck in a local optimum. To evaluate the proposed method on the quantitative analysis of SILAC data, we perform simulation studies.

2. METHODS

2.1 Classification methods

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 Gaus-

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

.

BICT 2015, December 03-05, New York City, United States

DOI 10.4108/eai.3-12-2015.2262391

sian distribution. Nonetheless, a general technique to parameter estimations for both approaches is the expectation-maximization (EM) algorithm [3, 1].

Suppose there are N protein-level abundances for case and control, $\{x_1,x_2,\cdots,x_N\}$ and $\{y_1,y_2,\cdots,y_N\}$, respectively, summarized by their peptide-level abundances. Then the protein-level ratios are constructed by $r = x/y$, resulting in N protein-level ratios, $\{r_1, r_2, \cdots, r_N\}$, where $r_i, i = 1, \cdots, N$, are log-transformed ratios. Note that the protein-level ratio represents the ratio between heavy-labeled and light-labeled proteins.

The primary objective 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.,$r_i = 0$), $i = 1, \cdots, N$, 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 gives us an idea to use a classification-based approach to differentially abundant proteins.

We employ the K-means clustering to classify the protein-level ratios into three groups. Consider N protein-level ratios, $\{r_1, r_2, \cdots, r_N\}$, where r_i,$i = 1, \cdots, N$, are log-transformed ratios, as described above. Then the goal of the K-means clustering is to estimate values for the two vectors, $\{z_{ik}\}$ and $\{_k\}$, where $i = 1, 2, \cdots, N$ and $k = 1, 2, 3$, by minimizing the following objective function, given by

$$J = \sum_{i=1}^{N} \sum_{k=1}^{3} z_{ik}(r_i - \mu_k)^2, \tag{1}$$

where μ_k, $k = 1, 2, 3$, represent the centers of each cluster and $z_{ik} \in \{0, 1\}$, $i = 1, 2, \cdots, N$ and $k = 1, 2, 3$, are binary indicator variables describing which of three clusters the ratio $r_i, i = 1, 2, \cdots, N$, is assigned to. As mentioned earlier, the estimation of $\{z_{ik}\}$ and $\{\mu_k\}$ can be carried out by the EM algorithm (e.g., [1]).

The K-means clustering can be considered as a nonparametric approach due to assuming no underlying distribution of the clusters. On the contrary, Gaussian mixture models (GMMs), a parametric approach, assume that the protein-level ratios are composed of a mixture of three Gaussian distributions. The GMM aims to maximize the following likelihood function to estimate the parameters, $\theta = \{\lambda_k, \mu_k, \sigma_k; k = 1, 2, 3\}$, given by

$$L_r(\theta | r_i, i = 1, \cdots, N) = \prod_{i=1}^{N} \sum_{k=1}^{3} \lambda_k \phi(r_i; \mu_k, \sigma_k), \tag{2}$$

where $\lambda_k > 0$, $k = 1, 2, 3$, is a mixing coefficient and $\sigma_{k=1}^3 \lambda_k = 1$; $\phi(x; \mu, \sigma)$ stands for a probability density function of Gaussian distribution with mean μ and standard deviation σ. To employ the EM algorithm, we introduce a set of trinary latent variable $z = \{z_i, i = 1, \cdots, N\}$ into Equation (2), such that:

$$r_i | z_i = k \sim Normal(\mu_k, \sigma_k); \tag{3}$$

$$Prob(z_i = k) = \lambda_k. \tag{4}$$

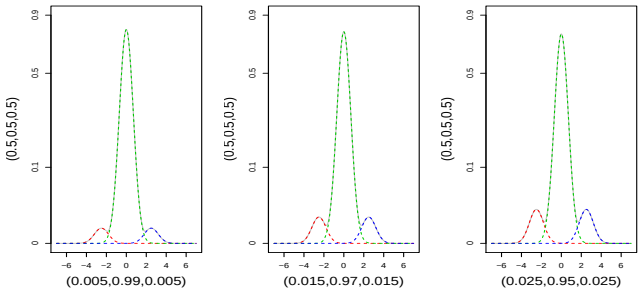

Figure 1: The density plots of simulated data.

where $Normal$ stands for a Gaussian distribution.

Each ratio then is assigned to a cluster based on z_{ik} for the K-means clustering and $Prob(z_i = k)$ for GMM. In other words, the ith ratio $r_i, i = 1, 2, \cdots, N$, is assigned to the cluster such that:

$$\arg \max_{k=1,2,3} z_{ik}; \quad \arg \max_{k=1,2,3} Prob(z_i = k); \tag{5}$$

for the K-means clustering and GMM, respectively.

2.2 Particle swarm optimization (PSO) classification

The EM algorithm [3] is popular in various areas because of its simplicity and stability, but it is a local optimization so that convergence to the global optimum cannot be guaranteed or expected. Therefore, none of the aforementioned approaches will guarantee convergence to the global optimum. To circumvent this difficulty, a stochastic global optimization, particle swarm optimization (PSO), is utilized in addition to the EM algorithm.

PSO, introduced by [7], 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 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 particle moves 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 proved to be an efficient method for many global optimization problems, and not suffering from the difficulties encountered by other evolutionary computation techniques. For an overview of PSO and its variants, see [5].

In this work, PSO is employed to enhance the performance of the M-step of EM algorithms. Namely, 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. To implement PSO within the M-step, we use the function 'psoptim' in the R package *pso*.

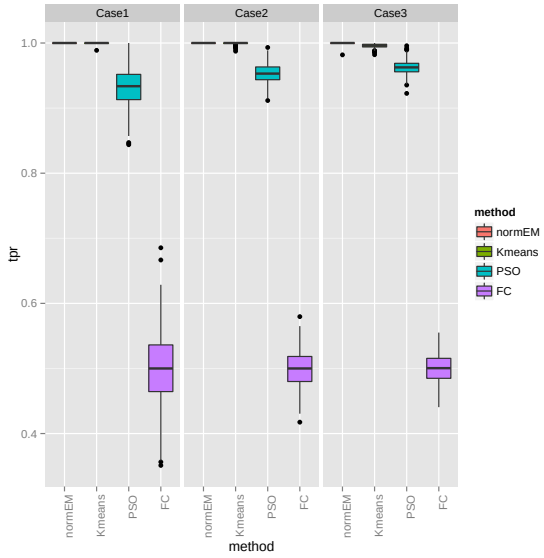

Figure 2: The boxplots of TPRs of each method according to three cases

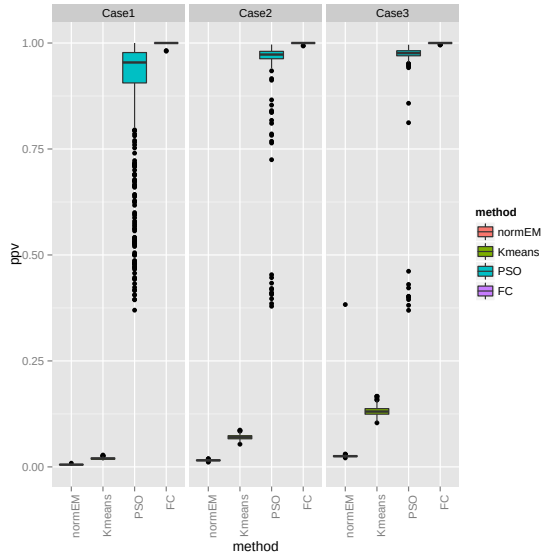

Figure 3: The boxplots of PPVs of each method according to three cases

2.3 Implementation

We use the functions available in the statistical software R packages (Version 3.0.1) in order to implement the aforementioned methods. The K-means classification is implemented using the function 'kmeans' available in the R package *stats*. For Gaussian mixture models, the function 'normalmixEM' in the R package *mixtools* is used. To implement the PSOM-step, the function 'psoptim' available in the R package *pso* is used for the PSO-based classification.

2.4 Performance criteria

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

The goal of the classification is to classify 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, $\{r_1, r_2, \cdots, r_N\}$. Suppose there are P down-regulated ratios, Q no-difference ratios, and R up-regulated ratios, where $N = P + Q + R$. Then, the performance measures are obtained by

$$\begin{aligned} TPR &= (E+K+G+M)/(P+R); \\ PPV &= (E+K+G+M)/(S+U); \\ F1\ score &= (2 \cdot TPR \cdot PPV)/(TPR+PPV) \\ &= (2 \cdot (E+K+G+M))/(P+R+S+U). \end{aligned}$$

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

3. RESULTS

We compared the performance of the proposed methods with the aforementioned methods using simulated data. As stated before, the used methods are normEM (Gaussian mixture models using the R package *mixtools*), Kmeans (K-means classification), and PSO. In addition to these three methods, we also included the fold-change-based method as a control: FC (fold change-based cut-off of 2.5), resulting in a total of four approaches.

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, 0, and 2, respectively. For the standard deviation (SD), the following set of values was used, (0.5, 0.5, 0.5), where (σ_1,σ_2,σ_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 order 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 were generated randomly using a mixture of three Gaussian components for each of three cases. The true density plots of each of the three cases are depicted in Figure 1. In the figure, from the left to the right, the index of each plot is 'Case1', 'Case2', and 'Case3'. 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 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 2 displays the boxplots of TPRs of each method according to three cases. Across all cases, normEM performs the best, followed by Kmeans. FC achieves the worst

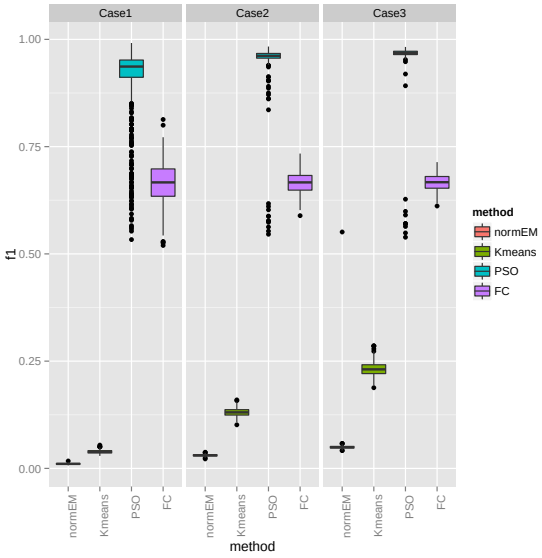

Figure 4: The boxplots of F1 scores of each method according to three cases

performance, while the performance of PSO is comparable to either normEM or Kmeans although it shows relatively larger variation.

The PPVs of each method are depicted by each of three cases in Figure 3. Interestingly, normEM and Kmeans achieve the worst PPV even though their TPRs are the best among four methods, and PSO and FC show the relatively larger PPVs compared to normEM and Kmeans, but PSO has the largest variation in PPV.

Figure 4 displays the boxplots of F1 scores by each case. As expected, PSO achieve the best performance in terms of F1 score. In particular, although normEM and Kmeans perform the best in terms of TPR, their F1 scores become the worst due to their poor performance of PPV.

4. CONCLUSIONS

We developed a new quantitative approach to dealing with SILAC protein-level summary using classification-based methodologies. Differently from the other methods, our approach relies mainly on the protein ratio summary and does not require the proteins with two or more peptide hits. As a result, no matter how many peptide hits the protein has, the developed method can be employed, rescuing many proteins doomed to removal. Another advantage 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 of convergence or being stuck in a local optimum. Our simulation studies clearly demonstrate that the PSO-based approach achieves the best performance among the four methods in terms of F1 score.

5. ACKNOWLEDGMENTS

The research reported in this paper was partially supported by the National Science Foundation (NSF) under Award Number DMS-1312603. 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 contents in this paper are solely the responsibility of the authors and do not necessarily represent the official views of NIH and NSF.

6. REFERENCES

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, NY, 2006.

[2] J. Cox and M. Mann. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, 26:1367–1372, 2008.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[4] E. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, and et al. A guided tour of the trans-proteomic pipeline. *Proteomics*, 10:1150–1159, 2010.

[5] A. Engelbrecht. *Particle Swarm Optimization, in Computational Intelligence: An Introduction*. John Wiley and Sons, Chichester, UK, 2007.

[6] D. Han, J. Eng, and R. Aebersold. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology*, 19:946–951, 2001.

[7] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proc. of IEEE International Conference on Neural Networks (ICNN)*, pages 1942–1948. IEEE, 1995.

[8] Z. Liao, Y. Wan, S. Thomas, and A. Yang. Iosquant: a software tool for stable isotope labeling by amino acids in cell culture-based mass spectrometry quantitation. *Anal Chem*, 15:4535–4543, 2012.

[9] M. Mann. Functional and quantitative proteomics using silac. *Nature Reviews Molecular Cell Biology*, 7:952–958, 2006.

[10] A. Margolin, S. Ong, M. Schenone, R. Gould, S. Schreiber, S. Carr, and et al. Empirical bayes analysis of quantitative proteomics experiments. *PLoS ONE*, 4(10):e7454, 2009.

[11] S. Ong and et al. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics*, 1:376–386, 2002.

[12] S. Ong and M. Mann. A practical recipe for stable isotope labeling by amino acids in cell culture (silac). *Nature Protocols*, 1:2650–2660, 2007.