35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

SF-RPQ: A novel statistical framework for reliable protein quantification in label-free quantitative proteomics

Mingon Kang[1], Dongchul Kim[1], Baoju Zhang[2], Xiaoyong Wu[2], and Jean Gao[1]

Abstract— **Label-free proteomics is a promising technology that provides qualitative and quantitative high-throughput analysis for determining the differential expression levels of proteins in proteomics. Protein quantification using mass spectrometry data plays a key role in analyzing proteins quantitatively and lays a foundation for further research such as biomarker discovery and signaling pathway construction in proteomics. Current quantification approaches use spectral counting, chromatographic peak area, peptide count, or sequence coverage to compare and quantify global protein expression differences. Although existing protein quantification methods have been contributing to quantify protein expression, however, no single method has been widely acknowledged due to their pros and cons depending on data and experiment setting. To make things worse, different quantification methods often tend to produce conflicting results with each other, which make it difficult to derive a reliable conclusion. In order to obtain significant protein biomarker candidates among thousands of proteins in the samples, a well-designed method to validate quantitative protein measurements is required as well as a high quality of protein identification and protein quantification.**

In this paper, we propose a statistical framework for reliable protein quantification (SF-RPQ) adapting Dezert-Smarandache theory from the artificial intelligent. SF-RPQ is designed to validate quantitative measurements of proteins with statistical models, including probabilistic approaches to quantify the reliability of the measurements. The proposed framework SF-RPQ was assessed by the experiments with the publicly available NCI-funded data, where SF-RPQ showed a good performance with high accuracy.

I. INTRODUCTION

Label-free quantitative proteomics has been widely used to identify and quantify the large number of proteins in complex biological samples, especially in order to study the differential protein expressions [1], [2], [3]. Label-free quantitative proteomics provides efficient mechanisms to analyze biological samples rapidly. Therefore, label-free technologies such as LC-MS (Liquid chromatography-mass spectrometry) and LC-MS/MS (liquid chromatography-tandem mass spectrometry) have been developed in concert with current advances in rapid data acquisition, ultra-high sensitivity, and dynamic range for the global shotgun proteomics study [4].

Protein quantification is a significantly important step for further proteomics research in label-free quantitative proteomics. None of the further research such as biomarker discovery and signaling pathway construction can be achieved without accurate protein quantitation. However, obtaining accurate quantitative measurements of proteins from tandem mass spectrometry data is still challenging. Moreover, it is difficult to design a universal quantification method due to various equipments of the mass spectrometer, sample preparation methods, experimental environments, and different goals of biological problems. No single method has been widely acknowledged due to their pros and cons depending on data and experiment settings.

Current quantification approaches use spectral counting, chromatographic peak area, peptide count, or sequence coverage to compare and quantify global protein expression differences. Although existing protein quantification methods have been reported suitable and contribute in quantifying protein expression, those quantification methods often produce conflict in providing protein abundance in the sample, which make it difficult to analyze the relation of the protein expression between control and treated samples. Also, proteins are often not identified simultaneously across all samples. In practice, only a few proteins simultaneously identified over all the samples are typically selected among thousands of proteins in the samples in order to obtain strongly reliable experimental results. However, it may possibly remove huge amounts of putative biomarker candidates with high probability. More importantly, while a large number of proteins can be identified simultaneously in label-free proteomics, only a few numbers of proteins tend to be a biologically significant biomarker. Thus, the development of methods which deal with both such confliction and the uncertainty of samples has been strongly demanded.

II. METHODS

We developed a statistical framework for reliable protein quantification (SF-RPQ) adapting a fusion method. Dezert-Smarandache theory (DSmT) was proposed by J. Dezert and F. Smarandache to deal with conflicting evidences and to enhance reasoning systems fusing heterogeneous data and information [5], [6], [7]. DSmT controls uncertainty coming from conflicting evidences and consequently makes a rational decision one step ahead of Bayesian probability [5]. While Bayesian theory is based on the classical ideas of probability, DSmT focuses more on interpretation of uncertainty and how to manage the information arising from evidences. Due to voluminous mathematical and philosophical concepts, the description of DSmT is not included in this paper.

We apply this fusion theory to deal with mutually contradictory information provided by multiple protein quantification methods, which is considered as evidence. Then,

[1] M. Kang is with the Department of Computer Science and Engineering, University of Texas at Arlington, TX 76019, USA {mingon.kang, dongchul.kim, gao} at uta.edu

[3] Baoju Zhang and Xiaoyong Wu are with the School of Physics and Electronic Information, Tianjin Normal University, Tianjin, China

the statistical framework assigns probabilistic degrees to determine whether the evidences are reliable or not.

Suppose that there are implicit peptide assignments (number of peptides, peptide ion intensity, and so on) analyzed by peptide/protein identification tools (e.g. Mascot and SEQUEST) and associated with P proteins in N samples of the binary class, $\tilde{\mathbf{X}}$ (treated) and $\tilde{\mathbf{Y}}$ (control), comprising of tandem mass spectrometry data, i.e., $\tilde{x}_{pi} \in \tilde{\mathbf{X}}, \tilde{y}_{pi} \in \tilde{\mathbf{Y}}$ for the pth protein and the ith sample, where $\tilde{x}_{pi}$ and $\tilde{y}_{pi}$ represent a set of implicit peptide assignments corresponding to a protein. Protein quantification methods define their functions with those sets of peptide assignments to quantitate proteins. Let $F_{(q)}(\tilde{\mathbf{x}}_{pi})$ be the function of the qth quantification method with implicit inputs $\tilde{\mathbf{x}}_{pi}$, and it returns the quantity of the pth protein in the ith sample. Now, we have manipulable $\mathbf{X}$ and $\mathbf{Y}$ matrices for treated and control data sets, respectively,

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_Q], \quad \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_Q], \tag{1}$$

$$\mathbf{x}_q = \log_2 \left(\begin{bmatrix} F_i(\tilde{x}_{11}) & \cdots & F_i(\tilde{x}_{1N}) \\ \vdots & \ddots & \vdots \\ F_i(\tilde{x}_{P1}) & \cdots & F_i(\tilde{x}_{PN}) \end{bmatrix} \right) \tag{2}$$

where Q is the number of protein quantification methods utilized to combine ($1 \leq q \leq Q$), and the binary logarithm is taken for efficient manipulation.

All measurements produced by protein quantification methods are considered as evidences.We assume that the measurements can be validated by the two hypotheses: the measurement is reliable (1) or not (2). The frame of discernment is then defined $\Theta = \{\theta_p, \theta_{\neg p}\}$ in DSmT. The power set of the Θ is $2^\theta = \{\phi, \{\theta_p\}, \{\theta_{\neg p}\}, \{\theta_p \cup \theta_{\neg p}\}, \{\theta_p \cap \theta_{\neg p}\}\}$ that can be interpreted as reliable ($\{\theta_p\}$), non-reliable ($\{\theta_{\neg p}\}$), possibly both reliable or non-reliable but uncertain ($\{\theta_p \cup \theta_{\neg p}\}$), and a conflicting evidence ($\{\theta_p \cap \theta_{\neg p}\}$). A *bpa* function assigns a degree of the belief to each element of the power set; $m(\cdot) : 2^\theta \mapsto [0, 1]$. That is, $m(\theta_p)$ and $m(\theta_{\neg p})$ are *bpa* functions that assign the degree of how likely reliable the measurement of the protein is or not, respectively. We define the *bpa* functions mathematically to represent the reliability of the measurements.

A. A bpa function for repeatability

The frequency of occurrence of proteins over the samples may represent repeatability in identification. Proteins can be often identified with a low score, which represents little similarity between the given mass spectrum signals and the hypothetically obtained signals from protein databases. Even though a protein is assigned a high score by protein identification tools, the protein of a low frequency would be doubtable. The *bpa* function for repeatability, $m_p^r(\theta_p)$, for the pth protein is defined as,

$$m_p^r(\theta_p) = \frac{\sum_{i=1}^{N} \sum_{q=1}^{Q} \left(sgn(\mathbf{x}_{pi}^{(q)}) + sgn(\mathbf{y}_{pi}^{(q)}) \right)}{2QN} \tag{3}$$

$$m_p^r(\theta_p) = m_p^r(\theta_p)(1 - 2T) \tag{4}$$

$$m_p^r(\theta_{\neg p}) = (1 - m_p^r(\theta_p))(1 - 2T) \tag{5}$$

$$m_p^r(\theta_p \cup \theta_{\neg p}) = 2T, \quad m_p^r(\theta_p \cap \theta_{\neg p}) = 0, \tag{6}$$

where $\mathbf{x}_{pi}^{(q)}$ describes the abundance measurement of the pth protein produced by the jth protein quantification method in the ith sample. The $sgn(x)$ function returns one if the pth protein is identified in the sample, otherwise zero. The tolerance T is introduced to set the upper and low bounds when assigning the belief degree to avoid the perfect belief assignment. $m_p^r(\theta_{\neg p})$ function in (5) is determined by $(1 - m_p^r(\theta_p)) \times (1 - 2T)$ in this study assuming that there is no confliction between θ_p and $\theta_{\neg p}$ in (6). Note that the confliction between θ_p and $\theta_{\neg p}$ does not mean the disparity between the evidences that multiple protein quantification methods generate. The evidences of disparity will be dealt with when combining this information using PCR5. The settings of (5) and (6) are basically applied to all of the following *bpa* functions if not specified.

B. A bpa function for reproducibility

In an ideal system, a protein quantification method would produce the consistent measurements of the protein abundances over the samples. Thus, quantification methods have been ultimately developed to provide high reproducibility and been evaluated by proving their high reproducibility [8], [9]. In order to measure the degree of reproducibility, the distribution comprising of standard deviations of $\mathbf{x}_p^{(q)}$ over the samples is considered. The standard deviation distributions are computed over the samples in (9). Then, we apply a logistic function to convert them to a probabilistic system in (10).

$$s_{x_p^{(q)}} = \sqrt{\frac{\sum(\mathbf{x}_p^{(q)} - \sum \mathbf{x}_p^{(q)}/N)^2}{N}} \tag{7}$$

$$s_{y_p^{(q)}} = \sqrt{\frac{\sum(\mathbf{y}_p^{(q)} - \sum \mathbf{y}_p^{(q)}/N)^2}{N}} \tag{8}$$

$$S_p = \{s_{x_p^{(1)}}, ..., s_{x_p^{(Q)}}, s_{y_p^{(1)}}, ..., s_{y_p^{(Q)}}\} \tag{9}$$

$$m_p^s(\theta_p) = \frac{(1 - 2T)}{2Q} \sum_{i=1}^{2Q} \Big(1 - \frac{1}{1 + e^{-(S_p - \mu_S)/\sigma_S}}\Big) \tag{10}$$

where $s_{x_p^{(q)}}$ and $s_{y_p^{(q)}}$ are the standard deviation vectors that the qth quantification method produces. μ_S and σ_S are the mean and the standard deviation of the standard deviation distribution (S_p) derived from all data sets, respectively, i.e., $\mu_S = \sum_{q=1}^{2Q} (S_p)_q / 2Q$, $\sigma_S = \sqrt{\sum_{q=1}^{2Q} ((S_p)_q - \mu_S)^2 / 2Q}$.

C. A bpa function for consistency

A protein ratio ($\tilde{x}_p/\tilde{y}_p$) between a control and a treated class provides protein significance showing protein expression changing. While $m_p^s(\cdot)$ in (10) is designed for reproducibility within each single protein quantification method, the *bpa* function $m_p^c(\cdot)$ defines the consistency between multiple protein quantification methods. All possible pairwise combinations of sample data produce protein ratio distributions (e.g., triplicated samples of binary classes produce a total of nine protein ratios) in (11). The density distribution $D_p^{(q)}(x)$ of protein ratios for the qth protein quantification method is generated using a kernel density estimator in (12).

$$z_p^{(q)} = \{\mathbf{x}_{pi}^{(q)} - \mathbf{y}_{pj}^{(q)}; 1 \leq i, j \leq N\} \tag{11}$$

$$D_p^{(q)}(x) \sim \frac{1}{N^2 h} \sum_{i=1}^{N^2} K\Big(\frac{x - (z_p^{(q)})_i}{h}\Big), \tag{12}$$

where $K(\cdot)$ is a kernel function (Gaussian Kernel function is used in this study), and h is a parameter of the scaled kernel. Note that $\mathbf{x}_{pi}^{(q)}$ and $\mathbf{y}_{pj}^{(q)}$ are logarithmic so thus protein ratios are calculated by subtraction not division in (11). On this wise, protein ratio distributions are generated for each protein quantification method. The dissimilarity between the distributions that each quantification method produces are computed by symmetric Kullback-Leibler divergence (KL divergence) [10] with pairwise comparison,

$$KL_p = \{\frac{1}{2}\Big(\sum D_p^{(i)} \ln \frac{D_p^{(i)}}{D_p^{(j)}} + \sum D_p^{(j)} \ln \frac{D_p^{(j)}}{D_p^{(i)}}\Big); \quad 1 \leq i, j \leq Q, i \neq j\}. \tag{13}$$

The *bpa* function for consistency $m_p^c(\cdot)$ finally defines it with a logistic function,

$$m_p^c(\theta_{\neg p}) = \frac{(1 - 2T)}{1 + e^{(KL_p - \mu_{KL})/\sigma_{KL}}} \tag{14}$$

$$m_p^c(\theta_p) = (1 - m_p^c(\theta_p))(1 - 2T), \tag{15}$$

where μ_{KL} and σ_{KL} are the mean and the standard deviation of the KL_p distribution, respectively.

D. Fusion of uncertain and conflicting evidences

Basic probability assignment functions are defined to quantify the degrees of repeatability, reproducibility, and consistency between multiple quantification methods in the probabilistic system. The total degree of beliefs for the pth protein is obtained by the combination of $m_p^r(\cdot)$, $m_p^s(\cdot)$, and $m_p^c(\cdot)$ using a PCR5 combination rule [11],

$$m_p^t(\cdot) = m_p^r(\cdot) \oplus m_p^s(\cdot) \oplus m_p^c(\cdot). \tag{16}$$

Belief (Bel) and plausibility (Pl) functions are computed with the total belief for each power set of Θ. The belief and plausibility function provide the possible interval for true belief ([Bel(θ_p), Pl(θ_p)]). The range of the probabilities, rather than a single probabilistic number as Bayesian theory has been done, provides more capable powers to represent the uncertainty. For example, more evidences typically make narrower range of the probabilities and expose stronger beliefs while Bayesian theory lacks the power. Note that Bayesian theory deals with the problems only on average no matter what the size of data is.

In order to make a rational decision, DSmT furthermore proposed a generalized *pignistic* transformation [12], The generalized *pignistic* transformation provides the Bayesian probabilistic approach for the classification problems. The discriminant to determine whether the protein quantitation is reliable or not is defined with a threshold t,

$$\frac{P(\theta_p)}{P(\theta_{\neg p})} > t. \tag{17}$$

III. EXPERIMENT RESULTS

We assessed the proposed framework by experiments with the data sets which are well characterized and publicly available for download from the Tranche repository (https://proteomecommons.org/tranche). The data was designed by NCI-funded Clinical Proteomic Technology Assessment for Cancer (CPTAC) group to test repeatability and reproducibility in proteomics for inter-laboratory comparability. In CPTAC data sets, human proteins (Sigma UPS-1) were spiked into digested yeast (60ng/uL) with five different levels (0.25, 0.74, 2.22, 6.67, and 20.00 $\text{fmol}/\mu\text{L}$) as equimolar protein mixtures, where each sample was triplicated. CPTAC_6A∼6E for convenience sake denote the data sets corresponding the different spike levels.The samples were distributed to laboratories, and mass spectra were generated by different kinds of mass spectrometers. For this experiment, the CPTAC samples spiked into yeast with only 6.67 (CPTAC_6D) and 20.00 (CPTAC_6E) $\text{fmol}/\mu\text{L}$ were used, since the human proteins were rarely detected in CPTAC_6A, CPTAC_6B, and CPTAC_6C data sets. CPTAC_6D and CPTAC_6E were used as treated and control data set, respectively, where the protein ratio (ground truth, $\log_2(6.67/20.00) = -1.5842$) is known.

Mascot daemon (Version 2.2.2) identified proteins with the UniProt KB/Swiss Prot database (released in Jan. 2012). Peptide mass tolerance was set to ±10 ppm, fragment mass tolerance was ±0.5 Da. Carbamidomethyl and oxidation were set for variable modifications. None was set for fixed modifications. Among double and triple-charged peptides with one allowed missed cleavage identified from the data set, only Sigma UPS-1 proteins were considered.

A total of 17.75 and 25.5 proteins on average are identified in CPTAC_6D and CPTAC_6E of four data sets, respectively. Three spectral counting-based protein quantification methods such as NSC (normalized spectral count), NSAF (normalized spectral abundance factor), and SIn (normalized spectral index) were used for $F_{(q)}(x)$ functions [13], [14], [8]. However, the proposed framework is rather flexible, allowing the combination of more quantification methods and software such as Progenesis-LC/MS, MSight, MZmine, OpenMS and MSQuant.

The total degree of the belief, probability interval, and pignistic probability, as a result, are listed in order of the pignistic probability in Table I.

The reliability for the proteins is determined by the discriminant with the threshold t in (17). The discriminant filters out the measurements of the proteins that are classified as unreliable. The assessment of the proposed framework SF-RPQ was measured by root-mean-square errors (RMSE), which is the normalized squared sum between the protein ratios of the proteins identified as reliable and true protein ratios. Fig. 1 illustrates the performance of the framework with various parameter settings of the tolerance T and the discriminant threshold t. The result shows that stronger power to measure proteins abundance as higher discriminant threshold t is applied. The measurements identified as reliable appear to

TABLE I: Quantifying the reliability of proteins in 'LTQ Orbitrap@86' data set ($T = 0.1$). The 23 protein, which are observed in at least two samples, are listed.

UniProt Accession	$m^t(\theta_p)^1$	[Bel Pl]^2	$P(\theta_p)^3$
P02787	.7394	[.7394 .7451]	.7422
P55957	.7382	[.7382 .7438]	.7410
P01127	.7162	[.7162 .7223]	.7190
P04040	.7084	[.7084 .7144]	.7113
P00918	.6827	[.6827 .6887]	.6856
P12081	.6750	[.6750 .6806]	.6779
P01031	.6646	[.6646 .6704]	.6675
P02768	.6563	[.6563 .6627]	.6593
P08263	.6560	[.6560 .6621]	.6589
P06732	.6341	[.6341 .6398]	.6371
P51965	.6148	[.6148 .6208]	.6178
P10636	.5889	[.5889 .5947]	.5919
P07339	.5692	[.5692 .5752]	.5723
P02788	.5450	[.5450 .5509]	.5481
P01344	.5443	[.5443 .5503]	.5474
P10145	.5277	[.5277 .5347]	.5307
P08311	.4588	[.4588 .4648]	.4620
P00709	.3728	[.3728 .3803]	.3761
P00441	.3472	[.3472 .3535]	.3504
O00762	.3257	[.3257 .3315]	.3290
P02144	.2375	[.2375 .2432]	.2410
P02741	.2366	[.2366 .2438]	.2401
P01008	.2346	[.2346 .2424]	.2380

$m^t(\theta_p)^1$: the total degree of the beliefs using PCR5, [Bel Pl]^2: the interval for true belief, $P(\theta_p)^3$: the pignistic probability

have good approximation to the true. Note that RMSE is normalized by the number of samples to counterbalance the different numbers.

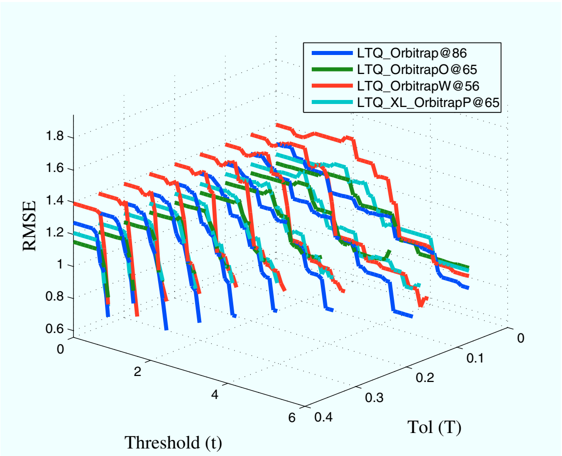

Fig. 1: The assessment of the performance of SF-RPQ with various parameter settings of the tolerance T ($0 < T < 0.4$) and the discriminant threshold t ($0 < t < 6$).

IV. Conclusions

Accurate and reliable protein quantification is indispensable in proteomics. Although large numbers of quantification methods and software have been introduced for this purpose, validating the protein quantitation is strongly necessary for reliable experiments and analysis. To this end, we set out to develop a novel statistical framework for reliable protein quantification based on Dezert-Smarandache theory. We built statistical models to validate measurements arising from multiple protein quantification methods comprehensively. We have fully evaluated the performance of the method with publicly available data sets. Plots with RMSE were provided to assess the performance of the proposed framework. The experimental result shows that this framework gives a distinct advantage of providing a probabilistic reliablity indicator of the metrics. The proposed framework SF-RPQ can be easily extended, utilizing either more quantification methods or software as well as various database protein identification tools as evidences. The proposed framework would promise successful further research such as biomarker discovery and signaling pathway construction.

References

[1] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: a critical review," *Anal Bioanal Chem*, vol. 389, pp. 1017–1031, 2007.

[2] W. Zhu, J. Smith, and C.-M. Huang, "Mass spectrometry-based label-free quantitative proteomics," *Journal of Biomedicine and Biotechnology*, 2010.

[3] M. Bantscheff, S. Lemeer, M. Savitski, and B. Kuster, "Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present," *Analytical and Bioanalytical Chemistry*, vol. 404, pp. 939–965, 2012.

[4] M. Wang, J. You, K. G. Bemis, T. J. Tegeler, and D. P. G. Brown, "Label-free mass spectrometry-based protein quantification technologies in proteomic analysis," *Briefings in Functional Genomics & Proteomics*, vol. 7, no. 5, pp. 329–339, 2008.

[5] J. Dezert and F. Smarandache, *Advances and Applications of DSmT for Information Fusion*. Rehoboth, NM: Amer. Res. Press, 2004.

[6] J. Dezert and F. Smarandache, *Advances and Applications of DSmT for Information Fusion*. Rehoboth, NM: Amer. Res. Press, 2006.

[7] J. Dezert and F. Smarandache, *Advances and Applications of DSmT for Information Fusion*. Rehoboth, NM: Amer. Res. Press, 2009.

[8] N. M. Griffin, J. Yu, F. Long, P. Oh, S. Shore, Y. Li, J. A. Koziol, and J. E. Schnitzer, "Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis," *Nature Biotechnology*, vol. 28, pp. 83–89, 2010.

[9] N. Colaert, J. Vandekerckhove, K. Gevaert, and L. Martens, "A comparison of MS2-based label-free quantitative proteomic techniques with regards to accuracy and precision," *Proteomics*, vol. 11, no. 6, 2011.

[10] B. Fuglede and F. Topsoe, "Jensen-Shannon divergence and Hilbert space embedding," in *IEEE International Symposium on Information Theory*, pp. 31–31, 2004.

[11] F. Smarandache and J. Dezert, "Information fusion based on new proportional conflict redistribution rules," in *Information Fusion, 2005 8th International Conference on*, vol. 2, p. 8 pp., july 2005.

[12] F. S. J. Dezert and M. Daniel, "The generalized pignistic transformation," in *Proc 7th Intl Conf. Information Fusion*, pp. 384–391, 2004.

[13] H. Liu, R. G. Sadygov, and J. R. Yates, "A model for random sampling and estimation of relative protein abundance in shotgun proteomics," *Analytical Chemistry*, vol. 11, no. 4, pp. 4193–4201, 2004.

[14] B. Zybailov, A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens, and M. P. Washburn, "Statistical analysis of membrane proteome expression changes in saccharomyces cerevisiae," *J. Proteome Res*, vol. 5, pp. 2339–2347., 2006.