Complex Proteomes Analysis Using Label-Free Mass Spectrometry-Based Quantitative Approach Coupled with Biomedical Knowledge

Chao Pan, Wenxian Peng, Huilong Duan, Ning Deng

Complex Proteomes Analysis using Label-free Mass Spectrometry-based Quantitative Approach Coupled with Biomedical Knowledge

Chao Pan[1], Wenxian Peng[2], Huilong Duan[1], Ning Deng[1]*

[1]*College of Biomedical Engineering and Instrument Science, Key Laboratory of Biomedical Engineering of Ministry of Education of china,* Zhejiang University, Hangzhou, China
[2]Department of Radiology, Zhejiang Medical College, Hangzhou, China
*Correspondence: *zju.dengning@gmail.com*

ABSTRACT: *Label-free quantitative proteomics based on mass spectrometry plays an essential role in large-scale analysis of complex proteomes. Meanwhile, quantitative proteomics is not only a way for data processing, but also an important approach for exploring protein functions and interactions in a large-scale manner. An effective method combining quantitation and qualification should be built. To systematically overcome this challenge, we proposed a new label-free quantitative method using spectral counting in the proposed method, the count of shared peptides was considered as an optimized factor to accurately appraise abundance of Isoforms for complex proteomes. Large-scale functional annotations for complex proteomes were extracted by g:Profiler and were assigned to functional clusters. To test the effect of the methods, three groups of mitochondrial proteins including mouse heart mitochondrial dataset, mouse liver mitochondrial dataset and human heart mitochondrial dataset were selected for analysis. According to the biochemical properties of mitochondrial proteins, all functional annotations were assigned to various signalling pathway or functional clusters. We came to draw a conclusion that the strategy with shared peptides overcame inaccurate and overestimated results for low-abundant isoforms to improve accuracy, and quantitative proteomics coupled with biomedical knowledge can thoroughly comprehend functions and relationships for complex proteomes, and contribute to providing a new method for large-scale comparative or diseased proteomics.*

KEYWORDS: ***Complex Proteomes, Label-free Quantitation, Mass Spectrometry, Biomedical Knowledge***

1. Introduction

Relative quantitative proteomics is aiming at quantifying and detecting differential protein expression between various biological samples of interest, such as biomarkers discovery, signalling pathway or drug discovery [1-3]. Generally, quantitative proteomics can be separated into two major approaches: the use of stable isotope labelling and label-free techniques[4]. Particularly, label-free approaches, which directly use MS feature of abundance such as spectral or peptide count or chromatographic peak area, is a reliable, versatile, and cost effective alternative to labelled quantitation [5]. However, issues arise with peptides that are shared between multiple proteins[6]. Which protein did they originate from and how should these shared peptides be used in a quantitative proteomics workflow[6]? In addition, post-translational modifications, isoforms, and splice variants are not captured by the mere analysis of transcript abundances. Protein mixtures today can routinely be characterized in terms of proteins present in the sample, but in order to allow biological interpretation, quantitative analyses are necessary[4].

In this paper, we used Normalized Spectral Abundance Factor (NSAF) based peptide count as starting point for our analysis[7] and proposed a new method for shared peptides to accurately evaluate abundance of Isoforms[6]. Label-free quantitative approaches can accurately describe abundance of complex proteins, and quantitative proteomics is not only a way for data processing, but a method for comprehending and explaining functions and relationships of proteins. Therefore, large-scale functional annotations were extracted from biomedical knowledge by g:Profiler[9] and were assigned to 12 functional clusters due to the biochemical properties of mitochondrial proteins. We found that the new strategy with shared peptides overcame inaccurate and overestimated results for low-abundant isoforms to improve accuracy, and analysis of biomedical knowledge based quantitative proteomics contributed to discovering biomarkers and targets.

2. Data and Methods

2.1. *Data acquisition*

All MS/MS spectral we used were from the preliminary work of authors[10-12]. Mitochondria were treated with 0.5% DDM to extract membrane proteins, separated by SDS-PAGE followed by CBB G250 staining. Bands were sequentially cut from the continuum of the gel lane and were labelled to obtain much more accurate results in the peptide shared quantitation. Proteins were digested with trypsin, and peptides were analysed by LTQ-Orbitrap.

2.2. *Data preparation*

All MS/MS spectra including mouse heart mitochondrial dataset, mouse liver mitochondrial dataset and human heart mitochondrial dataset were searched against the IPI mouse database (version 3.47) and IPI human database (version 3.68)[8] using the pFind software kit (version 2.6)[13]. Detailed search parameters were performed using as follows: partial tryptic digest allowing two missed cleavages; fixed modification of cysteine with carbamidomethylation (57.021 Da) and variable modification of methionine with oxidation (15.995 Da), the precursor and fragment mass tolerances were set up at 1.5 and 0.5 Da, respectively. Peptides matching the following criteria were used for protein identification: DeltaCN$\geqslant$0.1; FDR$\leqslant$1.0%; peptide mass was 600.0~6000.0; peptide length was 6~60.

2.3. *Label free quantitative algorithm*

NSAF which was described by Old *et al*[7] gained popularity because it used protein length to rectify spectral counts to improve accuracy. We used the normalized spectral abundance factor (NSAF) based peptide count for quantitative proteomics. All peptide spectral counts were summed for each identified protein, and then divided by protein length, generating the values of Spectral Abundance Factor (SAF); the SAF value of each identified subunit was then normalized against the sum of all SAFs within an individual biological sample, resulting in the Normalized SAF (NSAF) value; all NSAF values were then calculated separately for all biological samples. The average value of NSAF for each identified protein was used for further quantitative and biological analyses. The normalization process, as a routine operation to eliminate systematic errors, can only be applied in some certain circumstances, for instance, when comparing relative changes between two complex mixture samples[14]. Meanwhile, we proposed a new method with shared peptides to explore how to accurately estimate abundance of isoforms for complex proteomes. We used distinct peptides as a proportional factor and allocated shared peptides to isoforms. Corresponding with NSAF, we similarly used protein length to rectify distinct peptides and obtained the proportional factor by normalizing distinct peptides, then allocated shared peptides to isoforms to obtain the final spectral count.

$$\left(NSAF_s\right)_J = \frac{\left(Sc / L\right)_J}{\sum_{i=1}^{N}\left(Sc / L\right)_i} \tag{1}$$

$$Sc = \sum_{band=1}^{B}\left(Scu_{band} + Scs_{band} \times R_{band}\right), \quad \text{and} \quad \mathrm{R}_{\mathrm{band}} = \frac{Scu / L}{\sum_{k=1}^{n}\left(Scu_k / L_k\right)} \tag{2}$$

In first equation, where Sc was the spectral count for protein J and L was the length of protein J, N was the total proteins; in the second equation, B was the total

bands, Scu was the count of distinct peptide, Scs was the count of shared peptide, R was the proportional factor.

2.4. *Functional analysis*

g:Profiler is a web-based toolset for functional profiling of gene lists from large-scale experiments[9]. It adopts the Benjamin-Hochberg statistic method to control false discovery rate (FDR)[15] to improve accuracy. According to these properties, g:Profiler was used to obtain functional annotation of complex proteomes in large-scale experiments.

We extracted gene name of each protein from IPI fasta database, and analysed these files including gene names by g:Profiler, then outcomes were analysed by in-house software toolkit to obtain functional annotation of complex proteomes. In this paper, we used *p*Value as a key factor to filter functional annotation. For the protein with multiple functions, the functional annotation corresponding to the smallest *p*Value was filtered. According to the biochemical properties of mitochondrial proteins, all mitochondrial proteins were assigned to 12 functional clusters.

3. Results and Discussion

In order to completely test the above methods, we selected mitochondrial proteins for analysis. Mitochondria have received extensive attention due to their importance in cellular function and known causative role in diseases. Mammalian mitochondria are double-membrane organelles, serving as the metabolic power houses of eukaryotic cells[16-18]. In this paper, mouse heart mitochondrial dataset, mouse liver mitochondrial dataset and human heart mitochondrial dataset were selected to obtain all MS/MS spectra, then each group was analysed by NSAF and its optimization algorithm. We found that the new strategy with shared peptides overcame inaccurate and overestimated results for low-abundant isoforms. Meanwhile, functional annotations of mitochondrial proteins in large-scale experiments were assigned to 12 functional clusters due to the biochemical properties of mitochondrial proteins. The work flow was shown as Figure 1.

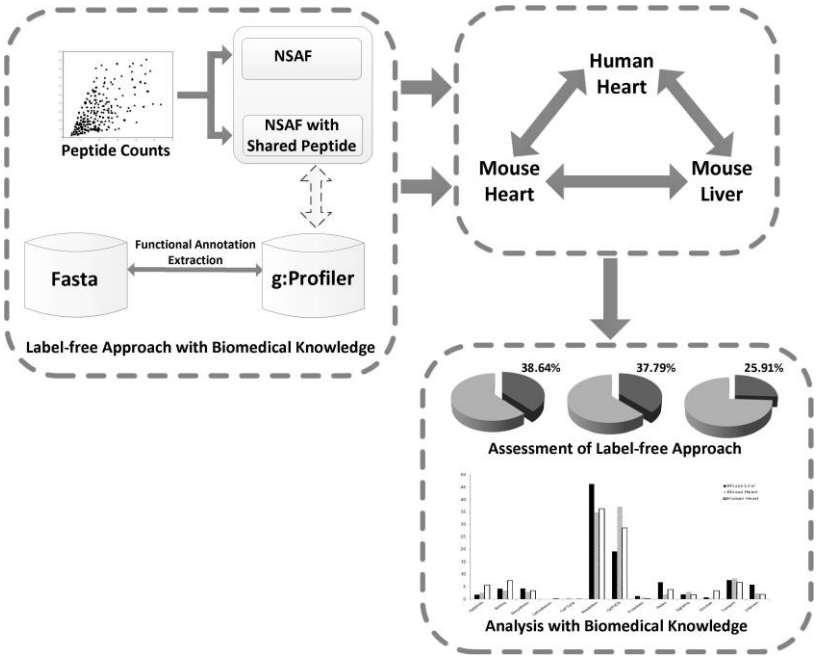

Fig 1. Workflow of label free quantitative proteomics with biomedical knowledge

3.1. *Assessment of label -free quantitative algorithm*

Spectral count, defined as the total number of spectra identified for a protein, has recently gained acceptance, therefore we evaluated the label-free approach based spectral count, especially the new strategy with shared peptides. We selected a group of mouse heart mitochondrial proteins to obtain MS/MS spectra by LTQ-Oribitrap, and this group of proteins were repeated three times, named as Group A, Group B, Group C. All proteins were searched by pFind toolkit to identify total counts of proteins and the counts of proteins with shared peptides. Then each group was quantified by NSAF and the new method, and quantitative results were sorted in descending order. We found that proteins with shared peptide were accounted twenty five to forty percent of total proteins, as table 1 shows; especially those proteins which rank have dramatic changes nearly had shared peptides. Importantly, we found that such identified proteins even came from a same family, such as proteins belonging to acyl-CoA dehydrogenase family. Proteins in the family play an important role in life event due to their biochemical properties of fatty acid metabolism and lipid metabolism. As table 2 shows, proteins with shared peptides reached 90%. Ranking of these proteins in the family generally ascended after approaching by the new strategy. Additionally, if all peptides of a protein were shared, the quantitative results were extremely different, such as IPI00331251. Therefore, we concluded that normalized processes we designed eliminated

systematic errors and should be considered when dealing with MS/MS spectra. Simultaneously, new strategy with shared peptide overcame inaccurate and overestimated results for low-abundant isoforms.

Table 1 Analysis of proteins with shared peptides in sample

Sample	Total Count of Proteins	Total Count of Proteins with Shared Peptides	Rate (%)
Group A	1589	614	38.64
Group B	1569	593	37.79
Group C	1397	362	25.91

Table 2 Analysis of proteins in acyl-CoA dehydrogenase family

Protein ID	Gene Symbol	Total Count of Peptides	Total Count of Shared Peptides	Rate (%)	Rank (NSAF Only)	Rank (with Shared Peptides)
IPI00119203	Acadvl	3033	3000	98.91	8	8
IPI00119114	Acadl	1449	1419	97.93	11	12
IPI00134961	Acadm	1150	1131	98.35	51	39
IPI00116591	Acads	705	692	98.16	63	63
IPI00274222	Acad8	200	188	94.00	140	100
IPI00331251	Acads	180	180	100	157	1444
IPI00331710	Acad9	155	144	92.90	182	111
IPI00119842	Acadsb	113	105	92.92	228	165
IPI00170013	Acad10	85	80	94.12	395	289

3.2.*Functional annotation of identified proteins*

According to the biochemical properties of mitochondrial proteins, all mitochondrial proteins were assigned to 12 functional clusters including apoptosis, DNA/RNA/protein synthesis, metabolism, oxidative phosphorylation, protein binding/folding, proteolysis, redox, signal transduction, structure, transport, cell adhesion and cell cycle. As table 3 shows, metabolic proteins have highest abundance in mouse liver mitochondrial dataset, while oxidative phosphorylation proteins show highest abundance in cardiac mitochondrial dataset. This explains that liver is important in metabolic process including nutrients synthesis, transformation and decomposition, however, heart promotes blood flowing to provide adequate blood to the organs or tissues, supplies oxygen or various nutrients and takes metabolic products away. Functional clustering for complex proteomes contributes to comprehending physiological and pathological characteristics of mitochondrial proteins.

Table 3 Analysis of functional clustering for mitochondrial proteins

Functional Clusters	% of Total Abundance (Mouse Liver)	% of Total Abundance (Mouse Heart)	% of Total Abundance (Human Heart)
OXPHOS	19.13	37.20	28.62
Metabolism	46.27	34.86	36.35
Transport	7.66	8.39	6.79
Apoptosis	1.82	2.56	5.70
Redox	6.82	1.94	3.94
Binding	4.20	3.47	7.40
Signaling	1.83	2.95	1.82
Biosynthesis	4.35	2.91	3.44
Structure	0.73	0.31	3.49
Proteolysis	1.22	0.74	0.37
Cell Adhesion	0.17	0.03	0.00
Cell Cycle	0.03	0.06	0.01
Unknown	5.76	2.40	1.89

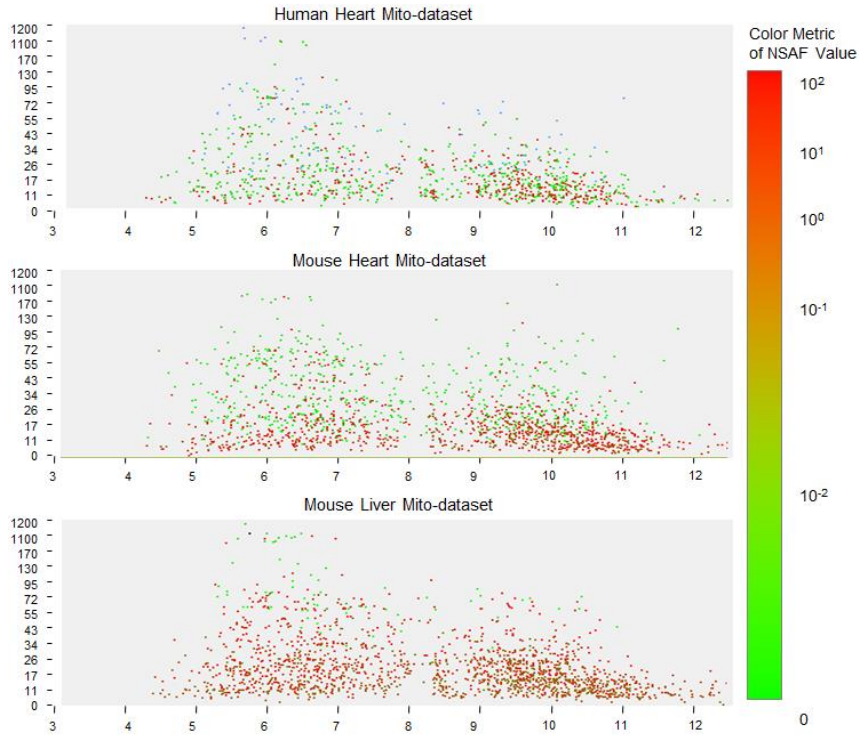

Fig 2. Individual Heatmaps Show Protein Distribution. X-axis represents *p*I and Y-axis represents molecular weight. Point color shows the abundance of proteins.

3.3. *Biochemical properties of identified proteins*

We analysed the biochemical proprieties of identified proteins including: MW (in kDa) and IEF point (*p*I) based NSAF value to discovery some new regularity using heat map. The heat map generated from software which was designed by us, as figure 2 shows. We found that *three* groups of proteins did not express much more differences, and most of abundant proteins with low molecular weight fall into the area of high *p*I value.

4. Conclusions

The method using peptide count can accurately describe abundance of proteins, in addition, the strategy dealing with shared peptides can estimate relative abundance of isoforms for complex proteomes and overcame inaccurate and overestimated results for low-abundant isoforms. According to the biochemical properties of mitochondrial proteins, large-scale functional annotations which were extracted from biomedical knowledge were assigned to 12 functional clusters. We provided a new method based on quantitative analysis to explain functions and relationships of complex proteomes and contribute to bioinformatics research including quantitative expression, difference comparison and diseased proteomics. Even though, NSAF could achieve the best precision using spectral as abundant features of isoforms, it seriously underestimate the actual fold change[14].Therefore, in order to precisely estimate quantitative results of low-abundant isoforms and further explore the deep relationship between peptides and MS/MS spectral, developing a new method becomes extremely important for complex proteomes.

5. Acknowledgements

This work was supported by the National High Technology Research and Development Programs of China (863 Programs, no. 2012AA02A601, no. 2012AA02A602, no. 2012AA020201), the National Science and Technology Major Project of China (No. 2013ZX03005012), and the National Natural Science Foundation of China, no. 31100592.

6. References

[1] Zhao Y, Lee W N P, Xiao G G. Quantitative proteomics and biomarker discovery in human cancer[J]. Expert Rev Proteomics, 2009,6(2): p. 115-118.

[2] Dong M Q, Venable J D, Au N, et al. Quantitative mass spectrometry identifies insulin signaling targets in C. elegans[J]. Science, 2007, 317(5838): 660-663.

[3] Lill J. Proteomic tools for quantitation by mass spectrometry[J]. Mass spectrometry reviews, 2003, 22(3): 182-194.

[4] Schulze W X, Usadel B. Quantitation in mass-spectrometry-based proteomics[J]. Annual review of plant biology, 2010, 61: 491-516.

[5] Zhu W, Smith J W, Huang C M. Mass spectrometry-based label-free quantitative proteomics[J]. Journal of Biomedicine and Biotechnology, 2009, 2010.

[6] Zhang Y, Wen Z, Washburn M P, et al. Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins[J]. Analytical chemistry, 2010, 82(6): 2272-2281.

[7] Zybailov B, Mosley A L, Sardiu M E, et al. Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces c erevisiae[J]. Journal of proteome research, 2006, 5(9): 2339-2347.

[8] Kersey P J, Duarte J, Williams A, et al. The International Protein Index: an integrated database for proteomics experiments[J]. Proteomics, 2004, 4(7): 1985-1988.

[9] Reimand J, Kull M, Peterson H, et al. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments[J]. Nucleic acids research, 2007, 35(suppl 2): W193-W200.

[10] Zhang J, Li X, Mueller M, et al. Systematic characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria[J]. Proteomics, 2008, 8(8): 1564-1575.

[11] Zhang J, Liem D A, Mueller M, et al. Altered proteome biology of cardiac mitochondria under stress conditions[J]. The Journal of Proteome Research, 2008, 7(6): 2204-2214.

[12] Zhang J, Lin A, Powers J, et al. Mitochondrial proteome design: From molecular identity to pathophysiological regulation[J]. The Journal of General Physiology, 2012, 139(6): 395-406.

[13] Wang L, Li D Q, Fu Y, et al. pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry[J]. Rapid Communications in Mass Spectrometry, 2007, 21(18): 2985-2991.

[14] Wu Q, Zhao Q, Liang Z, et al. NSI and NSMT: usages of MS/MS fragment ion intensity for sensitive differential proteome detection and accurate protein fold change calculation in relative label-free proteome quantification[J]. Analyst, 2012, 137(13): 3146-3153.

[15] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency[J]. Annals of statistics, 2001: 1165-1188.

[16] McDonald T G, Van Eyk J E. Mitochondrial proteomics[J]. Basic research in cardiology, 2003, 98(4): 219-227.

[17] Weiss J N, Korge P, Honda H M, et al. Role of the mitochondrial permeability transition in myocardial disease[J]. Circulation research, 2003, 93(4): 292-301.

[18] Honda H M, Korge P, Weiss J N. Mitochondria and ischemia/reperfusion injury[J]. Annals of the New York Academy of Sciences, 2005, 1047(1): 248-258.