PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2015, 42(1): 82~90
www.pibb.ac.cn

一种结合生物医学知识的蛋白质组非标记定量分析方法及其应用*

潘 超 苏运聪 杨 睿 段会龙 邓 宁**
(浙江大学生物医学工程与仪器科学学院，生物医学工程教育部重点实验室，杭州 310027)

摘要 基于质谱的非标记定量方法能够对复杂蛋白质组进行规模化分析，同时，在定量分析的基础上理解和解释蛋白质组的功能和相互作用关系更有意义．这需要建立一种有效的兼容定量和定性分析结果的方法．针对这一需求，本文首先借鉴了NSAF(normalized spectral abundance factor)算法采用肽段计数对蛋白质组数据进行定量，进一步结合共享肽对该方法进行优化．以此为基础，通过 g:Profiler 获取海量蛋白质组的功能注释信息，在定量分析的过程中，同步实现了对蛋白质组数据的功能性分析．本文选择来自人心脏、小鼠心脏、小鼠肝脏的三组线粒体蛋白质组数据对该方法进行验证，按照功能性分析将三组数据划分为若干功能组或信号通路，并进行相关性、功能聚类以及电子传递链分析．结果表明，结合共享肽的优化算法克服了对低丰度蛋白质的错误估计，提高了非标记定量的准确性．同时，结合生物医学知识的分析方法解释了蛋白质组的功能和相互作用关系，为差异比较蛋白质组学、疾病蛋白质组学以及功能蛋白质组学等组学研究提供了新的方法．

关键词 蛋白质组学，非标记定量分析，生物质谱，生物医学知识
学科分类号 Q51，Q811.4 **DOI**: 10.3724/SP.J.1206.2014.00191

定量蛋白质组学主要通过某种方法或技术，对生物样品(细胞、组织或体液等)在某些过程中蛋白质的含量进行比较分析，在疾病标志物发现、信号通路研究、药物发现等领域有广泛应用[1~3]．目前，以生物质谱技术为基础的定量蛋白质组学研究出现了稳定同位素标记定量和非标记定量两类主流方法[4]．其中，非标记定量方法直接利用蛋白质组质谱实验所产生的数据，比较肽段数、峰强度等谱图特征参数实现定量分析[5]，使实验流程简单化，克服了稳定同位数标记昂贵的实验费用、动态范围受到标记方法的限制等技术局限[6]，从而得到广泛应用．直接利用肽段数作为定量分析的丰度特征值，提高了定量的精确性[7]．结合共享肽的优化算法修正了当一个肽段匹配到多个蛋白质时，该肽段究竟起源于哪个蛋白质，在定量分析过程中如何分配使用该肽段的关键问题[8]．同时，随着生物医学研究的不断深入，以定性分析为主的蛋白质组鉴定、翻译后修饰分析等已无法满足需要[4]，如何充分发挥质谱技术及生物计算技术对复杂蛋白质生物样本的分析处理能力，深层次地发现和解释生命现象，依然是目前蛋白质组学研究需要思考和解决的问题．

针对上述问题，本文首先借鉴了NSAF (normalized spectral abundance factor)算法[9]对线粒体蛋白质数据进行定量，利用结合共享肽的优化算法[8]准确估计复杂蛋白质组中同源异构体之间的表达水平．蛋白质组的非标记定量分析方法能准确表示蛋白质丰度的相关信息，同时，定量蛋白质组学不仅仅是数据处理过程，在定量分析的基础上理解和解释蛋白质组的功能和相互作用关系则更有意义．因此规模化的定量分析模式需要与之相适应的蛋白质组自动功能注释．以信号上下游通路为基础的疾病特异性靶标分析和预测需要借助生物信息学

* 国家自然科学基金(31100592)，国家高技术研究发展计划(863)(2012AA02A601, 2012AA02A602, 2012AA020201)，国家科技重大专项(2013ZX03005012).
** 通讯联系人.
Tel: 0571-87951792, E-mail: zju.dengning@gmail.com
收稿日期：2014-07-04，接受日期：2014-10-28

分析手段自动获得蛋白质所参与的代谢、功能或信号通路．因此，本文首先从 IPI 蛋白质数据库中抽取出蛋白质的基因名[10]，利用 g:Profiler 获取蛋白质组的功能信息[11]，然后通过我们设计的接口软件自动解析 g:Profiler 的输出结果，从而获得海量蛋白质组的功能注释信息．通过功能性分析将不同物种不同器官的线粒体蛋白质组划分为若干功能组或信号通路，进行相关性分析、功能性聚类以及电子传递链分析．以 NSAF 算法为基础，结合共享肽的优化算法能准确地表示蛋白质的丰度信息，提高了定量结果的准确性．在定量分析的基础上对蛋白质组数据进行生物医学知识分析，为大规模蛋白质组的差异比较及疾病特异性标记物的预测与评估提供了新的方法．

1 数据与方法

1.1 数据的获取与预处理

1.1.1 数据的获取

应用于本文的所有蛋白质组数据均来自于论文作者前期工作中从线粒体蛋白质组的生物质谱实验中所产生的数据[12-14]．首先在线粒体中提取出膜蛋白，通过 CBB G250 染色后用 SDS-PAGE 进行分离．着染好的胶带顺序地进入凝胶泳道被连续切断，为了能够在共享肽的实验中对蛋白质组数据进行准确定量，对胶带进行标记，使实验数据与其相对应．接着再通过胰蛋白酶酶解获得肽段，最后将酶解好的肽段通过 LTQ-Orbitrap 进行分析得到质谱数据．

1.1.2 数据的预处理

所有质谱图的数据检索通过蛋白质鉴定引擎 pFind2.6 软件包(中国科学院计算技术研究所，北京)分析[15]，所用数据为小鼠心脏、小鼠肝脏以及人心脏的质谱数据和 IPI3.47 小鼠蛋白质 Fasta 数据库以及 IPI3.68 人蛋白质 Fasta 数据库[10]．其参数设置为：母离子质量偏差(precursor tolerance)为 ±1.5Da；碎片离子质量偏差(fragment tolerance)为 ±0.5Da；蛋白质酶解类型：胰蛋白酶；最多允许 2 个漏切位点；半胱氨酸的烷基化修饰(carbamidomethyl：C+57.021 Da)设置为肽序列氨基酸的固定修饰；甲硫氨酸氧化修饰(oxidation：M+15.995 Da)设置为可变修饰；同时，满足如下条件的肽段用于蛋白质组的鉴定：Deltcn(Sequest) ≥ 0.1；FDR ≤ 1.0%；肽质量范围为 600.0～6000.0；肽长度范围为 6～60.

1.2 方法

1.2.1 非标记定量算法设计

在本文中，首先借鉴 Florens 等[9]提出的 NSAF 算法，该方法利用蛋白质的序列长度矫正了蛋白质的图谱总数，在定量蛋白质组研究中得到了广泛的应用和验证．NSAF 算法对所获取的重复肽段采样数进行了两步归一化处理，获得谱丰度因子(spectral abundance factor，SAF)，然后累加统一实验样本中每个鉴定蛋白质 SAF 总和，并用单个蛋白质的 SAF 值与 SAF 总和值进行归一化，求出归一化的 SAF 值，即 NSAF，作为蛋白质丰度变化的表现．归一化的过程作为减少系统误差的一个基本操作，经常被用于两个复杂样本中相对变化的比较[7]．接着，在 NSAF 的基础上，本文提出了一种结合共享肽的优化算法，探索如何准确估算复杂蛋白质组中同源异构体之间的表达水平，即：以非共享肽计算值作为比例因子，将共享肽按比例分配至各同源异构体．为了能够在共享肽的实验中对蛋白质组数据进行准确定量，我们在数据获取过程中对着染好的胶带进行标记，使其与通过质谱仪分析得到的实验数据相对应．在结合共享肽的算法中，为了与 NSAF 算法相对应，同样利用了蛋白质的序列长度矫正了蛋白质的独立肽段数，对所获取的独立肽段采样数进行了归一化处理，从而得到以非共享肽为计算值的比例因子，最后再将共享肽按比例分配至各同源异构体，得到最终的肽段采样数．

$$(\mathrm{NSAF}_s)_J=\frac{(Sc/L)_J}{\sum_{i=1}^{N}(Sc/L)_i} \tag{1}$$

$$Sc=\sum_{band=1}^{B}(Scu_{band}+Scs_{band}\times R_{band}),\ \text{其中}\ R_{band}=\frac{Scu/L}{\sum_{k=1}^{n}Scu_k/L_k} \tag{2}$$

在等式(1)中，s 表示结合共享肽的 NSAF 优化算法，Sc 是检出蛋白质的有效肽段采样数，L 是蛋白质的氨基酸序列长度，N 是检出的蛋白质总数，J 是样本中被鉴定的第 J 个蛋白质．在等式(2)中，Sc 是单个蛋白质最终的谱图数，$band$ 是检出肽所在的切胶条带号，B 是切胶条带总数，Scu 是非共享肽的计算值，Scs 是共享肽的计算值，R 是共享肽在各同源异构体中的分配比例．

1.2.2 基于生物医学知识挖掘的定量蛋白质组功能分析

g:Profiler 是一个用于大规模实验中获取基因功能的工具[11]．该工具用于挖掘海量基因数据的功能

信息，采用了 Benjamini-Hochberg 统计学模型控制多次重复实验中假发现率的问题[16]，从而提高分析结果的准确性．因此，根据上述特性本文采用 g:Profiler 获取海量蛋白质组的功能注释信息．

本文首先从 IPI 蛋白质数据库中抽取出每个蛋白质的基因名，将含有基因名的文件输入 g:Profiler 中获取蛋白质组的功能信息，然后通过我们设计的接口软件自动解析 g:Profiler 的输出结果，从而获得海量蛋白质组的功能注释信息．当出现一个蛋白质有多个功能时，将选取最小 P 值所对应的功能；若同时几个功能所对应的 P 值相同且都最小，则同时筛选出这几个功能信息．依据论文作者前期工作中对线粒体蛋白质组的功能性分析结果[12, 17-18]，我们将挖掘出的不同物种不同器官的线粒体蛋白质组的功能注释信息按线粒体本身的功能特性划分为 13 个功能组或信号通路，包括：氧化磷酸化、细胞代谢、转运、细胞凋亡、氧化还原反应、蛋白质结合、生物合成、蛋白质水解、信号蛋白等，采用序列比对的方法对不同物种不同器官的 13 组功能信息进行聚类，实现功能性分析，最后结合定量的方法，对蛋白质组中的电子传递链 5 个不同复合体进行了定量比较分析．

2　结果与分析

为了验证以上方法，本文选取线粒体蛋白质组作为分析对象．线粒体作为细胞中最重要的细胞器，通过合成 ATP 为生命提供能量，是“细胞动力工厂”．同时，线粒体还参与细胞分化、信息传递和凋亡等过程，其蛋白质结构与功能的改变与人类许多疾病也密切相关，与线粒体相关的研究受到广泛关注[19-21]．本文首先将选取的人心脏、小鼠心脏、小鼠肝脏 3 组线粒体蛋白质组数据通过上述方法获得肽段数据，并对 3 组蛋白质组数据分别进行 NSAF 和结合共享肽的优化算法定量．接着利用非标记定量的结果，对同一物种的不同器官，不同物种的同一器官以及不同物种的不同器官进行相关性分析．同时，将从生物医学知识库中挖掘出的海量蛋白质组功能注释信息划分为若干功能组或信号通路，从信号通路、线粒体电子传递链复合体的蛋白质丰度等多个层次进行定量分析、评估和验证．其方法流程图如图 1 所示：

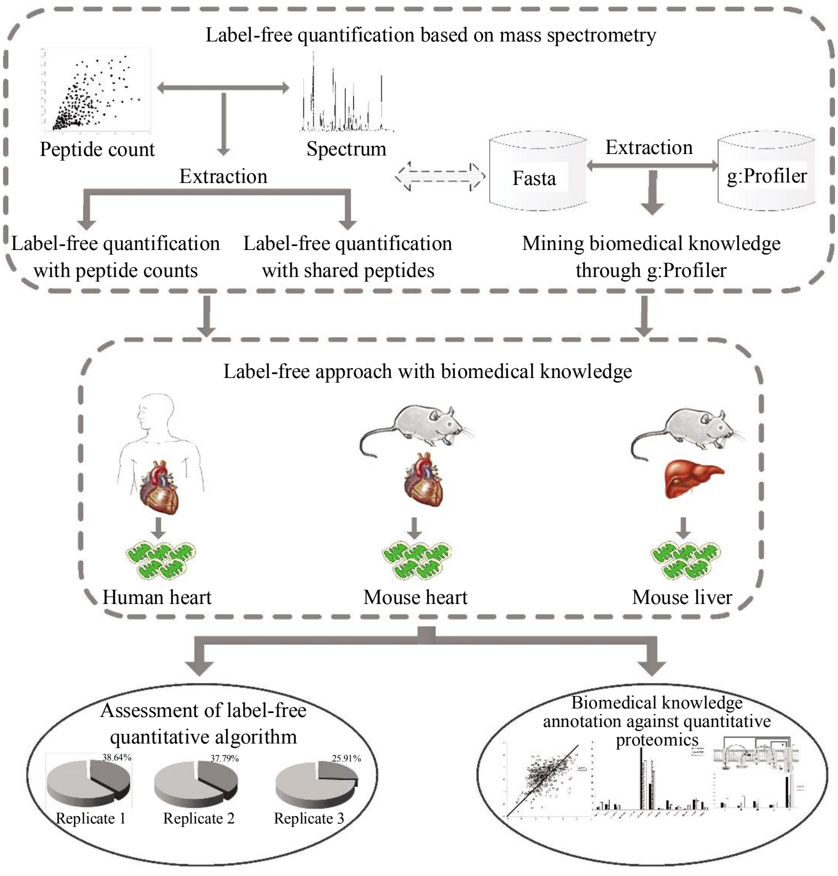

Fig. 1　Workflow of label free quantitative proteomics combined biomedical knowledge

2.1 非标记定量方法的评估

从目前蛋白质组学非标记定量方法的发展来看，采用肽段数的方法获得了广泛应用，是普遍认可的定量分析模型，本文以此为基础重点对结合共享肽的优化算法进行了评估．我们任意选取一组小鼠心脏线粒体蛋白质组数据，经高分辨率质谱仪3次重复实验后得到3组质谱数据分别为Group A，Group B和Group C．通过pFind软件分别检索出这3组数据总的蛋白质数以及其中含有共享肽的蛋白质数．接着将3组数据分别通过NSAF算法和结合共享肽的优化算法进行定量，对定量结果按降序排列．我们发现在3组重复试验中，含有共享肽的蛋白质占总蛋白质的25%～40%(表1)，其中排名变化超过100的蛋白质都含有大量的共享肽，这些含有大量共享肽的蛋白质有的来自于同一个蛋白质家族，如被鉴定出来的属于acyl-CoA dehydrogenase家族的所有蛋白质几乎都含有共享肽，如表2所示，比率达到了90%以上．该家族的蛋白质在生物体内承担着脂肪酸代谢、脂质代谢等重要功能，对维持生命活动具有重要意义．采用结合共享肽的优化算法后，该家族的蛋白质排名普遍上升．另外，当一个蛋白质的肽段全部是共享肽时，定量结果就出现了极大的偏差，如IPI00331251蛋白质所示．结果表明，本文设计的重复质谱实验的归一化统计分析方法，降低了系统误差，结合共享肽的优化算法因其以非共享肽计数值作为比例因子，将共享肽按比例分配至各同源异构体的方法避免了传统方法直接剔除共享肽计数所造成的信息缺失，克服了NSAF对于低丰度蛋白质的错误或过高估计，提高了定量结果的准确性．

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	100.00	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

2.2 结合生物医学知识的分析

2.2.1 采用相关性统计学模型对线粒体蛋白质组数据进行定量比较

我们利用非标记定量获得的蛋白质，根据BioEdit[22]的Local Blast[23]本地序列比对，通过我们设计的接口软件自动解析比对结果，以Score最高、*E*-Value值最低为原则，将人和小鼠的蛋白质进行映射，根据映射结果对3组数据通过散点图进行相关性分析，相关性公式如等式(3)所示．同时，为了使所有蛋白质均匀分布在散点图中，对定量结果进行了对数变换，得到最终的Score(4)．如图2所示：在第一组实验中，小鼠心脏的线粒体蛋白质和小鼠肝脏的线粒体蛋白质的相关性为0.6852；在第二组实验中，人心脏的线粒体蛋白质和小鼠心脏的线粒体蛋白质的相关性为0.6245；在第三组实验中，人心脏的线粒体蛋白质和小鼠肝脏的线粒体蛋白质的相关性为0.5607．分析结果表明，同一物种不同器官的相关性最高，不同物种不同器官的相关性最低．接着，利用IBM SPSS (version 20)统计分析软件对3组实验数据进行显著性分析[24-25]．结果显示，3组实验数据在0.01水平(双侧)上都显著相关，其P值基本为0．这表明根据3组实验的相关系数评价相关性与根据P值评价相关性是相互支撑的．

$$R=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}} \tag{3}$$

$$\mathrm{Score}=\lg(\mathrm{NSAF}_s) \tag{4}$$

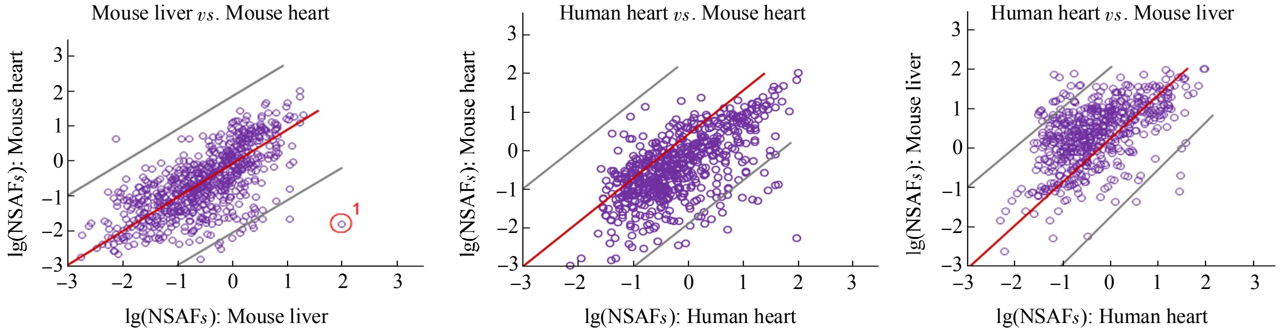

Fig. 2 Comparison of mitochondrial proteome

2.2.2 通过功能聚类理解线粒体蛋白质的生理病理特性

根据从生物医学知识库中挖掘出的海量蛋白质组的功能注释信息，我们将其划分为包括氧化磷酸化、细胞代谢、转运、细胞凋亡、氧化还原反应、蛋白质结合、生物合成、蛋白质水解、信号蛋白等在内的 13 个功能组或信号通路[12, 17-18]．如图 3 所示，肝脏的代谢蛋白质含量最多，心脏氧化磷酸化的蛋白质则最多．这就解释了肝脏是身体内以代谢功能为主的一个器官，参与营养物质的合成、转化与分解，而心脏起着推动血液流动，向器官、组织提供充足的血流量，以供应氧和各种营养物质，并带走代谢终产物(如二氧化碳、尿酸等)，使细胞维持正常的代谢和功能等作用．按照不同的功能聚类，有助于我们能更好地理解组织器官在生物体内的作用，发现不同样本中差异表达的蛋白质，从而有利于疾病等分析．

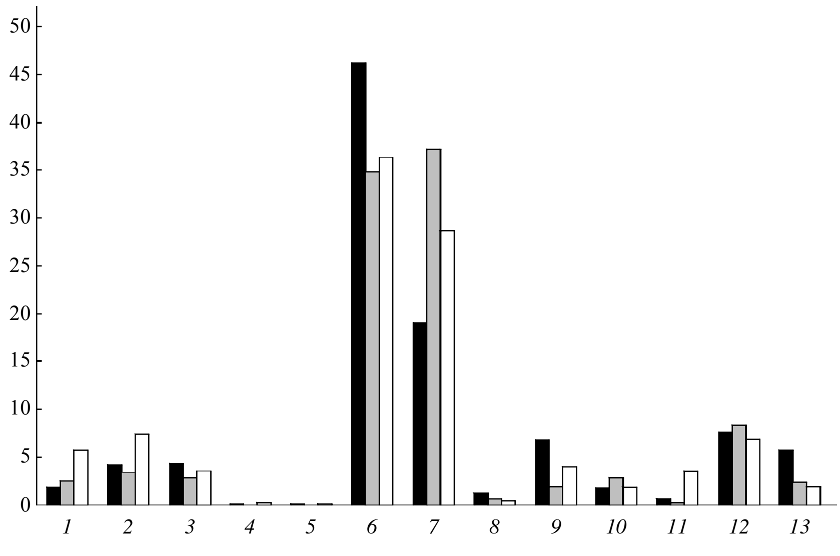

Fig. 3 Distribution of protein abundance (human heart, mouse heart & mouse liver mito-datasets)

x-Axis represents protein function and *y*-axis represents percentage. ■: Mouse liver; ▨: Mouse heart; □: Human heart. *1*: Apoptosis; *2*: Binding; *3*: Biosynthesis; *4*: Cell adhesion; *5*: Cell cycle; *6*: Metabolism; *7*: OXPHOS; *8*: Proteolysis; *9*: Redox; *10*: Signaling; *11*: Structure; *12*: Transport; *13*: Unknown.

接着分别取出 3 个样本中相对丰度位于前 10 的蛋白质做比较分析，从功能上看，排在前列的蛋白质都参与线粒体的重要功能，如电子传递链复合体、三羧酸循环、脂肪代谢等．首先，列出了人心脏线粒体蛋白质组中丰度位于前 10 位的蛋白质，这些蛋白质都参与了线粒体的重要功能，如表 3 所

示．我们发现这些蛋白质在小鼠心脏线粒体数据集中的丰度也很高，其中有 6 个蛋白质位于前 10，而在小鼠肝脏线粒体数据集中，仅有来自电子传递链的 2 个蛋白质位于前 10，其他蛋白质的含量则相对较低．接着分析位于小鼠心脏线粒体数据集中前 10 位的蛋白质，如表 4 所示，这些蛋白质的其中 4 个在人心脏线粒体数据中也位于前 10．对于肝脏的线粒体蛋白质组数据，如表 5 所示，位于前 10 位的蛋白质分布情况与心脏的差别就很大，个别蛋白质在心脏中甚至未被检测出，如第 5 行的蛋白质．另外位于小鼠肝脏中第一的蛋白质在小鼠心脏中位于 999，在人心脏中位于 1004，该蛋白质对应到图 2 的小鼠肝脏和小鼠心脏的散点图中则是 1 号蛋白质，此蛋白质在小鼠肝脏中的表达量几乎是心脏表达量的 100 倍以上．原因在于该蛋白质的主要功能是尿素循环，在哺乳动物体内，该蛋白质几乎仅存在于肝脏中．

Table 3　Top 10 most abundant proteins in human heart mitochondrial dataset

Rank	Protein ID	Gene	Function	Rank mouse heart	Rank mouse liver
1	IPI00303476	ATP5B	ETC complex	1	3
2	IPI00440493	ATP5A1	ETC complex	2	4
3	IPI00022314	SOD2	Redox	32	27
4	IPI00017855	ACO2	Metabolism	3	55
5	IPI00015141	CKMT2	Transport	14	14
6	IPI00022891	SLC25A4	Transport	5	59
7	IPI00007188	SLC25A5	Transport, Apoptosis	121	36
8	IPI00015911	DLD	TCA cycle	222	631
9	IPI00337541	NNT	Metabolism	7	202
10	IPI00291006	MDH2	TCA cycle, Metabolism	4	17

Table 4　Top 10 most abundant proteins in mouse heart mitochondrial dataset

Rank	Protein ID	Gene	Function	Rank mouse heart	Rank mouse liver
1	IPI00468481	Atp5b	ETC complex	1	3
2	IPI00130280	Atp5a1	ETC complex	2	4
3	IPI00116074	Aco2	TCA cycle	5	45
4	IPI00323592	Mdh2	Metabolism	14	14
5	IPI00115564	Slc25a4	Transport	7	48
6	IPI00223092	Hadha	Fatty acid metabolism	15	16
7	IPI00309964	Nnt	OXPHOS	13	163
8	IPI00119203	Acadvl	Metabolism	25	23
9	IPI00230351	Sdha	OXPHOS	180	41
10	IPI00226430	Acaa2	Metabolism	73	2

Table 5　Top 10 most abundant proteins in mouse liver mitochondrial dataset

Rank	Protein ID	Gene	Function	Rank mouse heart	Rank human heart
1	IPI00111908	Cps1	Urea cycle	999	1004
2	IPI00226430	Acaa2	Fatty acid metabolism	10	65
3	IPI00468481	Atp5b	ETC complex	1	1
4	IPI00130280	Atp5a1	ETC complex	2	2
5	IPI00420718	Hmgcs2	TCA cycle	N/D	N/D
6	IPI00116753	Etfa	ETC complex	27	36
7	IPI00308885	Hspd1	Binding	15	23
8	IPI00111218	Aldh2	Redox	200	91
9	IPI00114710	Pcx	Binding	114	824
10	IPI00230507	Atp5h	ETC complex	12	13

2.2.3 电子传递链的定量比较分析

以人心脏、小鼠心脏、小鼠肝脏 3 组线粒体数据集为例，通过分解电子传递链(electron transfer chain，ETC)中每一个蛋白质的定量分析结果来进一步揭示线粒体蛋白质组的功能特性．线粒体中参与电子传递链的蛋白质，无论在不同物种之间还是不同组织器官之间都表现出较高丰度．首先按照复合体进行分组，对电子传递链蛋白质的平均 $NSAF_s$ 值进行了比较(图 4)，作为 ATP 合成工厂的第 5 复合体，蛋白质的表达量远高于其他复合体．从功能上：5 个复合体中，第 1～4 复合体主要参与了线粒体膜内外的电子传递过程，第 5 复合体则是 ATP 合成的主要场所；从表达量上：第 1～4 复合体的丰度水平接近，而第 5 复合体的丰度水平则是其他的 2～3 倍．另外，存在于线粒体中的蛋白质，一部分是由细胞核基因组编码得来，而另一部分则是由线粒体本身的基因组编码得来[26]．其中由线粒体本身的基因组所编码的 13 个蛋白质全部位于电子传递链复合体之上，分别是：MTATP6、MTATP8、MTCO1、MTCO2、MTCO3、MTCYB、MTND1、MTND2、MTND3、MTND4、MTND4L、MTND5、MTND6[27]．我们通过质谱方法鉴定出了其中的 12 个蛋白质，未被鉴定出来的蛋白质 NADH-ubiquinone oxidoreductase subunit 6，在 ETC Complex Ⅰ中承担着电子转运等重要作用．我们发现，相对于由细胞核基因组编码的其他蛋白质而言，由线粒体本身的基因编码得到的蛋白质的丰度水平远远偏低．可以假设，由线粒体自身基因组编码的蛋白质是线粒体电子传递链复合体合成过程中的限制性因素，它们在线粒体中的表达量决定了电子传递链复合体的合成过程．同时也有相关研究报道指出许多由于线粒体功能缺失引起的疾病都与这些线粒体基因组编码蛋白高度相关[28]．

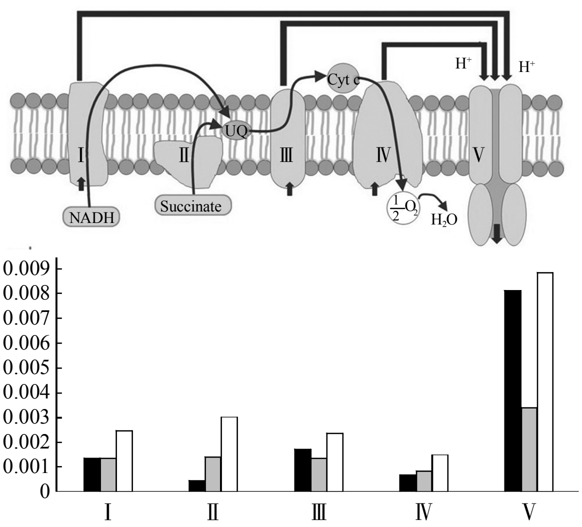

Fig. 4　Quantitative analysis of ETC complex proteins

■: Human heart; ▨: Mouse liver; □: Mouse heart.

3 总　　结

采用肽段计数的定量分析方法，能准确地表示蛋白质的丰度信息；结合共享肽的优化算法能准确估计复杂蛋白质组中同源异构体之间的表达水平，克服了对低丰度蛋白质的错误或者过高估计，提高了定量结果的准确性．以此为基础，从生物医学知识库中挖掘出海量蛋白质组的功能注释信息，以线粒体蛋白质组为分析对象，进行相关性、功能性聚类以及电子传递链分析，在定量分析的基础上进一步解释了蛋白质组的功能以及相互作用关系，为开展基于生物质谱数据的蛋白质组的定量表达、差异比较、功能和疾病蛋白质组等相关生物信息学研究提供新的方法和工具．另外，在生物质谱实验中，因胰蛋白酶作为特异性最强的蛋白酶，在决定蛋白质的氨基酸排列中有着重要的作用[29]，因此被广泛应用于大规模蛋白质组的非标记定量鉴定分析中．而对于在本次实验电子传递链的定量比较分析中，未被鉴定出来的蛋白质因不含能被胰蛋白酶酶切的氨基酸位点而无法定量分析．因而在将来的小规模特定研究中将采用其他酶进行鉴定分析．同时，基于肽段计数的非标记定量分析方法能准确估算出定量的精确性，但却低估了定量的 fold-change[7]．因此，为了对低丰度蛋白质更准确的定量，进一步挖掘肽段丰度和谱图信息之间更深层次的关系，发展新的定量指标则成为接下来研究的关键问题．

参 考 文 献

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

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

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

[4] Schulze W X, Usadel B. Quantitation in mass-spectrometry-based proteomics. Annual Rev Plant Biol, 2010, **61**: 491-516

[5] 赵慧辉, 杨　帆, 王　伟, 等. 无标记定量法研究冠心病不稳定性心绞痛血瘀证的差异蛋白质组. 高等学校化学学报, 2010(2): 285-292

Zhao H H, Yang F, Wang W, *et al*. Chem J Chin U, 2010 (2): 285-292

[6] Zhu W, Smith J W, Huang C M. Mass spectrometry-based label-free quantitative proteomics. J Biomed Biotechnol, 2009, **2010**(840518): 1-6

[7] 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. Analyst, 2012, **137**(13): 3146-3153

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

[9] Zybailov B, Mosley A L, Sardiu M E, *et al*. Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces cerevisiae*. J Proteome Research, 2006, **5**(9): 2339-2347

[10] Kersey P J, Duarte J, Williams A, *et al*. The international protein index: an integrated database for proteomics experiments. Proteomics, 2004, **4**(7): 1985-1988

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

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

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

[14] Zhang J, Lin A, Powers J, *et al*. Mitochondrial proteome design: from molecular identity to pathophysiological regulation. J General Physiology, 2012, **139**(6): 395-406

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

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

[17] Deng N, Zhang J, Zong C, *et al*. Phosphoproteome analysis reveals regulatory sites in major pathways of cardiac mitochondria. Mol Cell Proteomics, 2011, **10**(2): M110. 000117

[18] Mueller M, Marters L, Reidegeld K A, *et al*. Functional annotation of proteins identified in human brain during the HUPO brain proteome project pilot study. Proteomics, 2006, **6**(18): 5059-5075

[19] McDonald T G, Van Eyk J E. Mitochondrial proteomics. Basic Research in Cardiology, 2003, **98**(4): 219-227

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

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

[22] Hall T A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT [C]//Nucleic acids symposium series. 1999, **41**: 95-98

[23] Tatusova T A, Madden T L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiology Letters, 1999, **174**(2): 247-250

[24] Nie N H, Jenkins J G, Steinbrenner K, Bent D H. SPSS: Statistical Package for the Social Sciences [M]. New York: McGraw-Hill, 1975

[25] Green S B, Salkind N J. Using SPSS for Windows and Macintosh: Analyzing and understanding data [M]. USA: Prentice Hall Press, 2010

[26] Anderson S, Bankier A T, Barrell B G, *et al*. Sequence and organization of the human mitochondrial genome. 1981

[27] Moilanen J S, Finnilä S, Majamaa K. Lineage-specific selection in human mtDNA: lack of polymorphisms in a segment of MTND5 gene in haplogroup J. Molecular Biology and Evolution, 2003, **20**(12): 2132-2142

[28] 刘松梅, 周　新, 郑　芳, 等. 基因芯片筛查糖尿病线粒体基因突变. 中华医学杂志, 2006, **86**(40): 2853-2857
Liu S M, Zhou X, Zheng F, *et al*. Chin Med J, 2006, **86** (40): 2853-2857

[29] Huber R, Bode W. Structural basis of the activation and action of trypsin. Acc Chem Res, 1978, **11**(3): 114-122

A Mass Spectrometry-based Label-free Quantitative Approach Coupled With Complex Proteome Functional Analysis*

PAN Chao, SU Yun-Cong, YANG Rui, DUAN Hui-Long, DENG Ning**

(*College of Biomedical Engineering and Instrument Science, Key Laboratory of Biomedical Engineering of Ministry of Education of China, Zhejiang University, Hangzhou* 310027, China)

Abstract Label-free quantitative approach based mass spectrometry was used for analysis of complex proteomes, meanwhile, a method based on quantitative analysis which is used for explaining functions and interactions in a large-scale manner is of great importance. To systematically overcome this challenge, we should build a method combing with quantitation and qualification. We used Normalized Spectral Abundance Factor (NSAF) based peptide count as starting point for our analysis and proposed a new method with shared peptides to accurately evaluate abundance of Isoforms for complex proteomes. In addition, large-scale functional annotations of complex proteomes were extracted by g:Profiler and analyzed in the process of quantitation. In this paper, three groups of mitochondrial proteins including mouse heart mitochondrial proteins, mouse liver mitochondrial proteins and human heart mitochondrial proteins were selected for analysis. All MS/MS spectra t were searched against the IPI mouse database and IPI human database using the pFind software kit. 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. According to the biochemical properties of mitochondrial proteins, all functional annotations were assigned to various signaling pathway or functional clusters, such as apoptosis, DNA/RNA/protein synthesis, metabolism, oxidative phosphorylation, protein binding/folding, proteolysis, redox, signal transduction, structure, transport, cell adhesion and cell cycle, and analyzed by correlation analysis, functional clustering and electron transport chain analysis. We found that proteins which rank have enormous variation between NSAF and the new method 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 searched using the online database analysis tool available through UniProt (www.uniprot.org). From the global perspective of the three groups of mitochondrial proteins, the correlation of mouse heart mitochondrial proteins and mouse liver mitochondrial proteins shows highest, while the correlation of human heart mitochondrial proteins and mouse liver mitochondrial proteins shows lowest, it denotes that the correlation of simple species and different organs shows highest. On the aspect of functional clustering, 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 plays an important role in metabolic process including nutrients synthesis, transformation and decomposition, however, heart promotes blood flowing to provide adequate blood to the organs or tissues, supply oxygen or various nutrients and take metabolic products away. We concluded that the strategy with shared peptides overcame inaccurate and overestimated results to improve accuracy, and label-free quantitative approach coupled with complex proteome functional analysis can thoroughly explore protein functions or relationship and provide a new method for large-scale comparative or diseased proteomics.

Key words proteomics, label-free quantitative approach, mass spectrometry, biomedical knowledge

DOI: 10.3724/SP.J.1206.2014.00191

---

*This work was supported by grants from The National Natural Science Foundation of China (31100592), National High Technology Research and Development Programs of China (863 Programs)(2012AA02A601, 2012AA02A602, 2012AA020201) and National Science and Technology Major Project of China (2013ZX03005012).

**Corresponding author.

Tel: 86-571-87951792, E-mail: zju.dengning@gmail.com

Received: July 4, 2014 Accepted: October 28, 2014