PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2013, 40(4): 297~308
www.pibb.ac.cn

蛋白质基因组学：运用蛋白质组技术注释基因组 *

张 昆[1, 2] 王乐珩[1] 迟 浩[1, 2] 卜德超[1, 2] 袁作飞[1, 2] 刘 超[1, 2] 樊盛博[1, 2]
陈海丰[1, 2] 曾文锋[1, 2] 罗海涛[1] 孙瑞祥[1] 贺思敏[1] 谢 鹭[3] 赵 屹[1]**
([1]中国科学院计算技术研究所，中国科学院智能信息处理重点实验室，北京 100190；[2]中国科学院大学，北京 100049；
[3]上海生物信息技术研究中心，上海 200235)

摘要 随着高通量 DNA 测序技术的飞速发展，越来越多的物种完成了基因组测序．定位编码基因、确定编码基因结构是基因组注释的基本任务，然而以往的基因组注释方法主要依赖于 DNA 及 RNA 序列信息．为了更加精确地解读完成测序的基因组，我们需要整合多种类型的组学数据进行基因组注释．近年来，基于串联质谱技术的蛋白质组学已经发展成熟，实现了对蛋白质组的高覆盖，使得利用串联质谱数据进行基因组注释成为可能．串联质谱数据一方面可以对已注释的基因进行表达验证，另一方面还可以校正原注释基因，进而发现新基因，实现对基因组序列的重新注释．这正是当前进展较快的蛋白质基因组学的研究内容．利用该方法系统地注释已完成测序的基因组已成为解读基因组的一个重要补充．本文综述了蛋白质基因组学的主要研究内容和研究方法，并展望了该研究方向未来的发展．

关键词 蛋白质基因组学，基因组注释，蛋白质组学，质谱技术
学科分类号 Q51，TP39 **DOI**: 10.3724/SP.J.1206.2012.00263

截至 2012 年 4 月，完成基因组测序的物种共 3 173 个，其中包括原核生物 3 000 个，真核生物 173 个，另外有 10 506 个基因组测序工程正在进行当中(Genome Online Database，http://www.genomesonline.org)．确定基因的位置和结构是进一步研究其生物功能的基础，但要在庞大的基因组中，尤其是复杂的真核生物基因组中精确地寻找编码序列并确定其结构，仍然是一项宏大且复杂的工程．

相对于基因组学和转录组学，蛋白质组学直接研究编码基因翻译出的蛋白质产物，比转录组学注释基因组获得的结果更直接，而且可以发现由于知识不足导致的基因从头预测算法遗漏的基因和基因结构注释的错误；此外，蛋白质存在特有的翻译后处理现象(如翻译后修饰和信号肽)，这些现象作为对基因组的进一步注释是基因组学和转录组学方法所无法替代的．这两点使得蛋白质组学在提供基因表达产物证据、确认和校正编码基因、解析翻译后处理现象，以及发现新的编码基因及其规律上拥有先天的优势．近年来，兴起一个利用蛋白质组学数据进行基因组注释的新研究方向——蛋白质基因组学(Proteogenomics)，本文就该研究方向做一综述：首先介绍基因组注释的概念，阐述蛋白质基因组学对基因组注释的重要性，然后总结蛋白质基因组学的研究内容和研究方法，最后展望蛋白质基因组学未来的发展．

1 基因组注释与蛋白质基因组学

1.1 基因组注释

基因组注释是在基因组上确定基因及其他元件的位置和结构，并赋予这些基因和元件生物功能的过程．按照 Stein[1]的观点，基因组的注释分为三个

* 国家重点基础研究发展计划(973)(2010CB912701, 2010CB912702)和中国科学院知识创新计划(KGCX1-YW-13)资助项目．
** 通讯联系人．
Tel: 010-62601016, E-mail: biozy@ict.ac.cn
收稿日期：2012-05-31，接受日期：2013-01-11

层次：核酸层注释，蛋白质层注释，代谢层注释．核酸层注释的主要任务是对基因组进行标注，即在基因组上标明编码基因、非编码基因以及对应调控区域的位置和结构；蛋白质层注释的主要任务是对编码基因进行分类及分配功能；代谢层注释则需要对基因和蛋白质如何参与生命代谢给出解释．

要完成对基因组的完整注释，首先要在基因组上获得编码基因，即核酸层注释．以往获得编码基因序列的标准方法是结合表达序列标签(expression sequence tag，EST)测序(如今已发展为 RNA-seq 技术)、基因预测和设计引物进行目标序列 RT-PCR 测序的方法[2]．基于串联质谱技术的蛋白质组学在 2008 年仍然被看作是少数派的做法[3]，但随着质谱技术的发展，越来越多的基因组注释研究开始采用核酸数据与蛋白质组学数据相结合的方法．

1.2　蛋白质基因组学：利用高通量质谱技术进行基因组注释研究

高压液相色谱分离耦合串联质谱技术已经逐渐成为大规模研究蛋白质组学的常用方法[4-6]，发展相对成熟，该技术称为鸟枪法蛋白质组学．串联质谱数据有三种高通量解析方法：搜索蛋白质数据库、从头测序、搜索谱图库，其中最常用的是搜索蛋白质数据库的方法(图 1)．搜索蛋白质数据库的方法依赖数据库的完整性和准确性，然而常用的蛋白质数据库来源于基因组和转录组的注释结果，虽然该注释结果具有一定的指导意义，但是不能反映蛋白质组的全部信息(如本文开始所述)．基因组、转录组和蛋白质组从复杂性的角度来看呈逐级递增的趋势，所以使用蛋白质组学数据进行基因组注释变得更加重要．

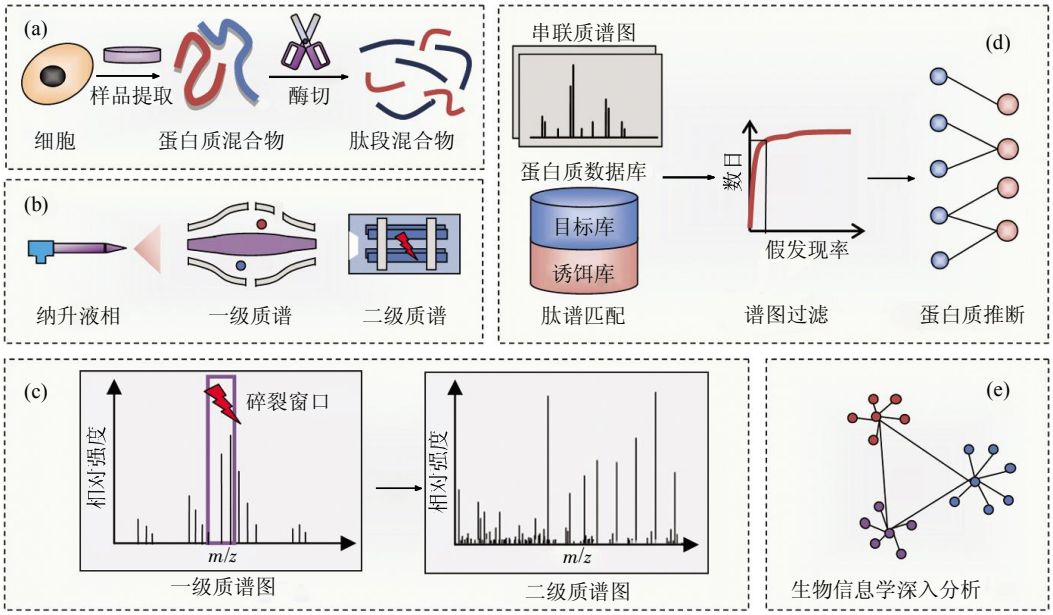

Fig. 1　Shotgun proteomics: experiments and data analyses

图 1　鸟枪法蛋白质组学的实验和数据分析流程

(a) 样品制备：从样品中分离出蛋白质，再经过酶切产生肽段混合物．(b, c) 液相分离、质谱碎裂与检测：肽段混合物通过液相分离，由质量检测器检测带电肽段质荷比，产生一级谱图 MS^1；挑选信号较强的肽段进行碎裂，产生二级谱图 MS^2．(d) 肽谱匹配(peptide spectrum matching，PSM)、肽谱匹配评价与蛋白质推断：二级谱图与数据库中肽段进行肽谱匹配打分；所有肽谱匹配通过评价算法保留高可信结果；高可信的肽谱匹配结果用于蛋白质推断，最终获得样品中蛋白质列表．(e) 进一步的生物信息学分析．

利用蛋白质组学数据，结合基因组数据(DNA)、转录组数据(RNA)来研究基因组注释问题，被称为蛋白质基因组学[7]．“蛋白质基因组学”一词由 Jaffe 等[8]于 2004 年首次提出，作者采用串联质谱数据匹配 DNA 翻译得到氨基酸序列的方法，在仅有 810 kb 大小的细菌基因组上直接鉴定开放阅读框(open reading frame，ORF)，验证并补充、修订了约 10%的 ORF．后来这种质谱数据

结合 DNA 和 RNA 数据的分析方法被应用到注释病毒基因组[9]. 原核生物基因组[10–22]以及真核生物基因组[23–39].

2 蛋白质基因组学的研究内容

目前，越来越多的基因组注释研究开始采用蛋白质基因组学的方法. 原因之一是蛋白质直接催化和调控生命活动，是对编码基因注释的最终体现(尽管部分 RNA 也参与调控活动，但对于编码基因，蛋白质组学数据显然有助于它们的注释)；其次是鸟枪法蛋白质组学技术灵敏度、精确度的不断提升，使得完整覆盖蛋白质组成为可能. 从 2010 年综述[40]的总结来看，使用高压液相色谱分离耦合串联质谱技术在不同物种的蛋白质组上均能获得 60%以上的覆盖率，以酵母为例，2008 年的研究[41]获得了 67%的全蛋白质组覆盖率，而 2011 年已经上升至 86%[42]. 由此看来，蛋白质组学数据独立于转录组、基因组数据，可以为基因组注释提供另一维丰富的信息.

从近些年的研究来看，蛋白质基因组学对基因组的注释主要集中于核酸层，也有部分蛋白质层的注释研究. 类似于 Ahrens 等[40]对蛋白质组覆盖的分层模型，蛋白质基因组学对基因组的注释可分为三个注释过程：编码基因的注释，编码基因结构的注释和翻译后处理的注释. 编码基因及其结构的注释属核酸层注释范畴，而翻译后处理的注释属蛋白质层注释范畴.

蛋白质基因组学与传统蛋白质组学的主要不同在于蛋白质基因组学结合了原始的 DNA 和 RNA 序列更为完整的信息. 蛋白质基因组学使用 DNA 和 RNA 序列建立数据库，对串联质谱图进行鉴定和评价，最终获得用于基因组注释的高可信肽段. 这些高可信肽段可分为两类：一类来源于原注释蛋白质数据库，这一类肽段用于验证已注释编码基因的表达和结构；另一类不包含在原注释蛋白质数据库中，但可与核酸序列信息匹配(genome search-specific peptide，GSSP)，该类肽段用于发现新基因和校正已注释基因的结构.

2.1 编码基因的注释

编码基因注释的目的，是要在基因组上获得所有表达(编码)基因的列表. 验证基因预测算法从核酸序列数据中预测得到的编码基因是否有正确表达的蛋白质产物，以及发现基因从头预测算法遗漏的编码基因——新基因，是蛋白质基因组学在编码基因注释方面的主要任务.

验证预测基因是蛋白质基因组学的重要内容. 基因从头预测(*ab initio* gene prediction)获得的假设蛋白质(hypothetical protein)需要通过高通量的手段进行验证. 类似地，基于相近基因组预测的结果也面临同样的问题. 再者，区分编码基因与非编码基因是一项比较困难的任务，按照长度进行区分的方法在一些长的非编码基因上会失效[43–45]，但是能否翻译成蛋白质是区分编码和非编码基因的一个主要标准. 另外一个困扰大家的基因组注释问题是假基因(pseudogene)，这类“基因”是否真正具有编码蛋白质的能力，也可以通过蛋白质基因组学的方法给出验证[16, 29, 46]. 高通量鸟枪法蛋白质组学是解决上述验证问题的利器.

发现新编码基因是蛋白质基因组学另一项重要内容. 许多研究[27, 29, 37]利用串联质谱数据结合蛋白质基因组学方法，不同程度上发现了新的编码基因，并扩充了原基因组注释数据库. Oshiro 等[23]在酵母中发现，原先按照氨基酸长度大于 100 注释得到的蛋白质数据库存在遗漏，进而使用蛋白质基因组学方法验证了 50 个氨基酸长度小于 100 的开放阅读框，补充了基因组注释.

2.2 编码基因结构的注释

获得了编码基因列表后，就需要对基因的精确结构进行研究. 对原核生物来说，编码基因的起始和终止位点注释是否准确，或者对真核生物来讲，外显子、内含子边界注释是否准确，有多少可变剪接体表达成蛋白质等，是编码基因结构注释要解决的主要问题. 图 2 详细展示了串联质谱鉴定所得的肽段如何对基因组注释进行验证和校正，具体来说有如下内容：a. 验证已注释编码基因结构. 在基因组注释工作中，如何确定编码基因的翻译起始位点，仍然是一个具有挑战性的问题. 有些基因具有多个翻译起始位点，甚至存在罕见起始子[10, 16]. 在高等真核生物中，可变剪接现象更增加了基因注释的复杂性. 因此在蛋白质组层次上验证编码基因的起始位点，可变剪接位点和不同的可变剪接体，是众多蛋白质基因组学研究的内容[25, 27, 29, 47]. b. 校正已注释编码基因边界(基因间边界). 编码基因的翻译起始位点容易存在注释错误，例如，两个独立的研究所 TIGR 和 Sanger Institute 在对结核分枝杆菌基因组预测时，预测的 ORF 有 12%存在不同，而在预测相同的 ORF 中，46%存在翻译起始位点错误[12]. 在真核生物中，蛋白质 N 端乙酰化修饰可以

用来作为编码基因翻译起始位点的一个证据[27]．另外，近年来发展起来的 N 端蛋白质组学[10, 24, 48-50]可以较好地解决该问题．Gallien 等[48]使用一种在酶解前封闭蛋白质 N 端的方法，可使蛋白质 N 端肽段在保留时间维度上更好地分离，并可以提高这些肽段的离子化效率，最终在结核分枝杆菌上发现 19%的翻译起始位点注释是错误的．在极少数蛋白质中，终止子 TGA 可以编码第 21 号氨基酸——硒代半胱氨酸，使得这些编码基因的终止子注释存在错误[51]．c．校正外显子边界(基因内边界)．尽管多数内含子是 GT-AG 形式的，但是也存在其他的形式，比如 AT-AC 形式的内含子，这就给基因预测软件带来不便．即使是 GT-AG 形式的内含子，目前基因注释结果在确定内含子外显子边界时也会存在不同程度的错误[27]．d．发现新的外显子和新的可变剪接体．落入内含子区域的肽段往往意味着新外显子的发现，同时也意味着原注释结果遗漏了新的可变剪接体．当然在原注释结果中发现新的跨越剪接位点肽段，也是对可变剪接体的补充．e．其他注释．比如一些重要的基因组突变或多态性研究：cSNP(coding single-nucleotide polymorphism)注释[52-53]，移码突变注释[15]，RNA 编辑(RNA editing)注释[54]，融合基因注释[29]等．

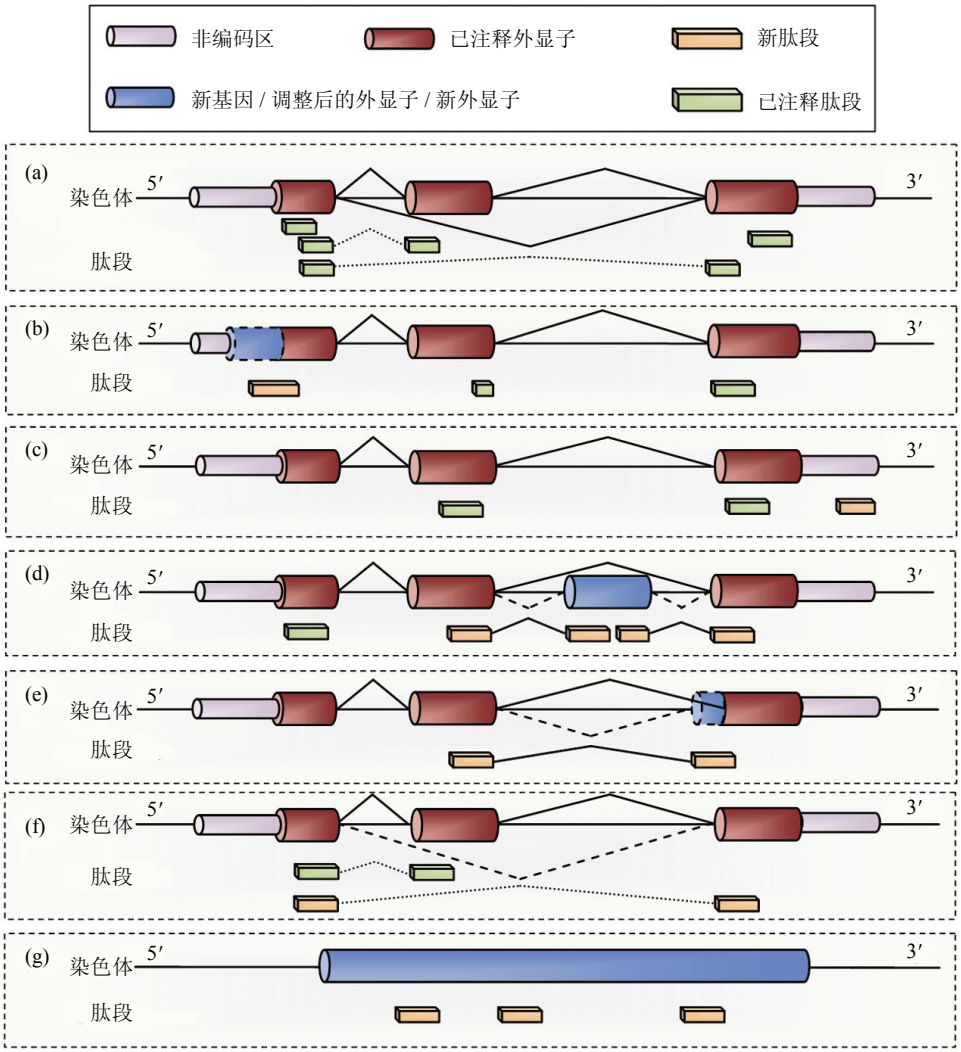

Fig. 2　Validation and refinement of gene annotation by proteogenomic methods

图 2　通过蛋白质基因组学方法验证与校正编码基因结构注释

(a) 验证编码基因注释结构．(b) 基因翻译起始位点纠正．一条 GSSP 落入已注释基因的 5′-UTR(Untranslated region)区域，支持该基因的翻译起始位点应该位于已注释起始位点的 5′端．(c) 基因 3′-UTR 表达．一条 GSSP 落入已注释基因的 3′-UTR 区域，支持该基因存在 3′-UTR 翻译的现象．(d) 发现新的外显子．GSSP 落入已注释基因内含子区域，表明原注释结果遗漏了一个外显子．(e) 校正外显子边界．GSSP 落在外显子 - 内含子边界，原注释外显子边界存在错误．(f) 发现新的可变剪接体．一条 GSSP 跨越 2 个外显子边界，而该剪接位点没有被注释．(g) 发现新基因．GSSP 落入基因间区域，支持新基因的发现，表明原注释数据库遗漏了该编码基因．

2.3 翻译后处理的注释

广义的基因组注释包括蛋白质功能和结构注释，尽管目前蛋白质基因组学研究主要集中在核酸层，即编码基因注释和基因结构注释，一些在蛋白质组学层面特有的现象，如翻译后修饰、信号肽，对蛋白质功能研究非常重要，近年来有研究[12, 16-17]逐渐拓展到翻译后处理的注释．事实上，目前鸟枪法蛋白质组学研究多数集中在蛋白质序列，或者说蛋白质一级结构的注释上．a．信号肽注释．信号肽是细胞外运蛋白质 N 端的一段长约 20 个氨基酸的序列．因为蛋白酶会切除细胞外运蛋白质 N 端的信号肽，所以成熟的细胞外运蛋白质在经过常规的样品制备过程中的胰酶酶切处理后，会在 N 端形成一段半特异酶切肽段．一些研究[12, 16]结合信号肽预测工具 SignalP 进行信号肽的注释．b．翻译后修饰(post translational modification，PTM)注释．翻译后修饰对基因组功能注释具有重要的作用，一些酶在发生修饰后才能行使功能，比如磷酸化修饰．在蛋白质组学中，非限制性鉴定和修饰发现[16, 55]可以指导蛋白质基因组学注释新的翻译后修饰．c．其他翻译后处理现象．比如 Gupta 等[16]详细讨论了在西瓦氏菌中的蛋白质 N 端甲硫氨酸切除(N-terminal methionine excision，NME)和蛋白质水解现象．

3 蛋白质基因组学的研究方法

蛋白质基因组学依赖蛋白质组学技术，但分析方法略有不同，其研究流程大致如下．

3.1 样品制备与质谱数据采集

样品制备和质谱数据采集是提取研究对象全蛋白质组的关键．为了更好地覆盖蛋白质组，蛋白质基因组学研究需要在不同条件下对研究对象的多个组织进行采样，以获得条件特异[20]和组织特异[25-26]表达的蛋白质．含量低的蛋白质(比如磷酸化蛋白质等)甚至需要一些特殊的富集手段来捕捉并鉴定．对于质谱数据采集来说，更好的色谱分离、更高的数据精度、更快的采集速度和更宽的动态范围是使更多的肽段获得鉴定的基础．新的蛋白质提取和分离技术、多酶切技术、肽段色谱分离技术、以及高精度高灵敏度质谱技术的发展确保了较高的蛋白质组覆盖[56]．

3.2 数据库构建与鉴定

蛋白质基因组学数据库通常结合已注释的蛋白质数据库以及核酸数据和数据库(基因组和转录组数据、数据库)来构建，主要有以下方法．

3.2.1 利用基因组数据构建蛋白质基因组学数据库．比较常用的方法是六阅读框翻译[57]，该方法是原核生物蛋白质基因组学研究的基本方法．真核生物利用六阅读框翻译的方法可以校正部分外显子和基因边界，但是由于可变剪接带来的复杂性，六阅读框翻译方法无法覆盖跨越剪接位点的肽段，所以发展出了以下几种方法．a．直接枚举的方法．PepSplice[58]可以在真核基因组上识别内含子保守序列，在规定长度内枚举可能的外显子剪接位点，这种方法在拟南芥上发现了 57 个新的可变剪接体，且发现 7 个预测的内含子区域有表达[25]．b．基于序列标签的方法．从谱图中获得肽段序列标签，再利用其基因座(locus)信息获得跨越外显子的肽段．比如 GPF 方法[33-34]综合使用串联质谱图从头预测和数据库鉴定策略，通过从头预测软件给出的序列标签在基因组上查找外显子片段，结合基因组序列获得可能的跨越外显子肽段，并提交数据库鉴定软件进行鉴定．类似的方法还有 PepLine[59]．c．借助已注释结果或基因从头预测软件．Mo[47]和 Xing[60]等从已注释数据库中获得每个基因的外显子并进行两两拼接，产生肽段进行鉴定．另外的一些工作直接采用基因从头预测软件预先对基因组进行注释[27, 29]，然后对预测结果进行蛋白质基因组学验证．

3.2.2 利用转录组数据构建蛋白质基因组学数据库．

转录组数据是对成熟 mRNA 的测序结果，不需要考虑任意可变剪接带来的数据库膨胀，所以转录组数据比全基因组序列复杂度低．但是由于测序结果存在冗余，也一定程度上增加了数据库的规模．

Ning 等[61]和 Oshiro 等[23]直接使用转录组测序数据来注释基因组．Edwards[62]使用 *de brujin* graph 的数据结构表示 EST 数据，可以将检索空间压缩至原来的 1/30．Tanner 等[31]将 EST 数据比对到基因组上，并使用 exon graph 的数据结构表示比对结果，将序列相互重叠的比对结果合并，拆分跨越外显子和 SNP 的比对结果，达到了压缩数据的目的．

虽然上述提到的一些构建数据库的方法已经整合了肽谱匹配打分和评价，但是通常的做法是将两者拆开，即先构建数据库，后进行肽谱匹配打分和评价．数据库构建完成之后，SEQUEST[63]、Mascot[64]、pFind[65]等数据库鉴定引擎将采集的质谱数据与数据库中的肽段进行肽谱匹配打分，PeptideProphet[66]、Percolator[67]等评价引擎会对打分

结果进一步评价，以获得可靠的肽谱匹配结果，用以进行基因组注释.

最近也有研究[68-70]讨论在没有基因组的情况下，如何利用近源物种基因组或蛋白质组数据库进行串联质谱图从头预测．与传统的串联质谱图从头预测方法不同的是，使用近源物种基因组或蛋白质组数据库可以弥补从头预测方法面临的谱峰信息不足和近源物种数据库不完整的缺陷．Zhao 等[71]使用串联质谱图从头预测软件 pNovo[72]，结合 EST 数据和 BLAST 分析，在没有参考基因组的猪蛔虫精细胞中鉴定到丝氨酸蛋白酶及其抑制剂这两个关键蛋白质，并用两者成功解释了猪蛔虫精子活化的机理．这是一个依靠蛋白质组学方法指导未测序物种基因功能研究的一个典范.

快速鉴定是蛋白质基因组学面临的一个重大挑战．以人类基因组为例，六阅读框翻译数据库大小约是传统蛋白质数据库的 230 倍左右，更不用考虑翻译后修饰、非特异酶切带来的搜索空间的膨胀．虽然 PepSplice 方法可以直接搜索基因库，并综合考虑了翻译后修饰、非特异酶切、突变以及可变剪接等变异形式，但是为了防止组合爆炸，PepSplice 算法在其迭代产生候选肽段的过程中加入了变异惩罚因子进行控制，会在一定程度上丢失候选肽段．GPF 方法虽然使用序列标签进行过滤和索引，可以提升鉴定速度，但同样存在着候选肽段丢失的问题．应对数据库和质谱数据膨胀的方法有以下三大类[73]：谱图预处理及聚类的方法[74]；数据库预处理及索引的方法[75-76]；单机加速、计算机硬件和底层编程[77]、并行计算技术[78].

3.3 基因组注释与重注释

经评价后的高可信肽段可分为两类：一类肽段来自原注释蛋白质数据库，用于验证已注释编码基因的表达和结构；另一类肽段(GSSP)来自核酸数据库，用于发现新基因和校正已注释基因结构．一般方法是将 GSSP 回贴到基因组上(BLAST 方法或字符串匹配方法[79])，结合基因组序列信息对结果进行注释．如本文 **2.1**～**2.3** 节所示，有关基因组注释内容的讨论已经清晰，但是如何合理评价基因组注释结果仍然面临挑战：a．在蛋白质质谱鉴定的目标诱饵库(target decoy approach，TDA)评价体系中，需要在数据库搜索空间中加入诱饵库，用来估计肽谱匹配的错误率．蛋白质基因组学数据库相比传统蛋白质数据库，在规模上有所增大．在这种情况下，TDA 的评价方法如何保证目标库和诱饵库不存在共享肽段，是蛋白质基因组学需要解决的问题[9, 73]．b．为了发现新基因、校正原注释基因，在蛋白质基因组学的研究当中尽可能引入更多的假设肽段，这样往往会造成灵敏度的下降[73]，即在相同假发现率的情况下获得的肽段鉴定结果较少，可利用的注释信息也随之减少．c．将肽段回贴基因组的问题与蛋白质推断问题[80-82]类似，也会遇到蛋白质推断问题中的难题，比如一条肽段对应多个基因座，由 GSSP 推断新基因假发现率扩大等问题．在真核生物中，同一基因因为可变剪接产生多个蛋白质表达产物，而且它们之间的序列相似性很高；另外同一家族的蛋白质也会存在相似序列，这给验证和校正带来困难．然而单个肽谱匹配的统计显著性在某些搜索引擎中估计不准甚至没有给出，这使得单肽谱匹配推断新注释结果往往可靠性不足[83-84]，比如 OHW(one hit wonder，只有一条肽段获得鉴定的蛋白质)的推断．最近发表的 PeptideClassifier[85]和 MIPGEM[86]可以结合基因模型推断蛋白质，而 MSpresso[87]则可利用 mRNA 表达数据辅助蛋白质推断．综合多组学的信息可以提高注释结果的可信度.

比较蛋白质基因组学(comparative proteogenomics)也是基因组注释与重注释的一种方法．它从比较基因组学发展而来，通过相近物种基因组间的比较和确证，在一定程度上提升基因组注释结果的可信度．Gupta 等[15]对西瓦氏菌的 3 个亚种进行比较蛋白质基因组学研究发现，分别有 21%，28%和 27%的蛋白质鉴定结果是 OHW，其中 3 个亚种同源的 OHW 有 50%以上同时得到鉴定．对每个亚种来说，10%以上的同源 OHW 会在其他亚种中鉴定到同一条肽段，这样随机发生的可能性非常低，从而肯定了 OHW 的鉴定结果．比较蛋白质基因组学在确认翻译起始位点[48]、区分移码突变和测序错误[88-89]等都有较多应用.

3.4 多组学数据整合与可视化工具

完整理解基因组需要整合多组学、多次实验的分析结果，并从不同的层次完善注释数据库．成熟的蛋白质基因组学注释流程应该是一个不断循环的过程(图 3)，在这个过程中，多组学数据和可视化工具可以为生物学家提供数据库支持，因此扮演了非常重要的角色．目前文献报道的可视化注释工具有 PeptideAtlas[32, 90]和 VESPA[91]，其中 PeptideAtlas 是结合 Ensembl Genome Browser 研发的一个分布式注释系统(distributed annotation system，DAS)，

VESPA 是运行在 PC 上的原核生物可视化注释工具．以 PeptideAtlas 为例，到目前为止，PeptideAtlas 已经收集了包括酵母、线虫、果蝇、小鼠在内的模式生物和人类的蛋白质组学数据，提供了编码基因的表达证据，并丰富了获得鉴定的蛋白质肽段、谱图匹配等信息．

作者认为这样的数据整合和可视化平台应该得到广泛的应用：方便数据的统一管理和信息的共享，有助于完善基因组注释；不同来源的数据整合和质量控制有利于基因组统一注释与评价，比如 PeptideAtlas 就整合了 iProphet[92]算法，达到了多来源数据质量控制的目的．

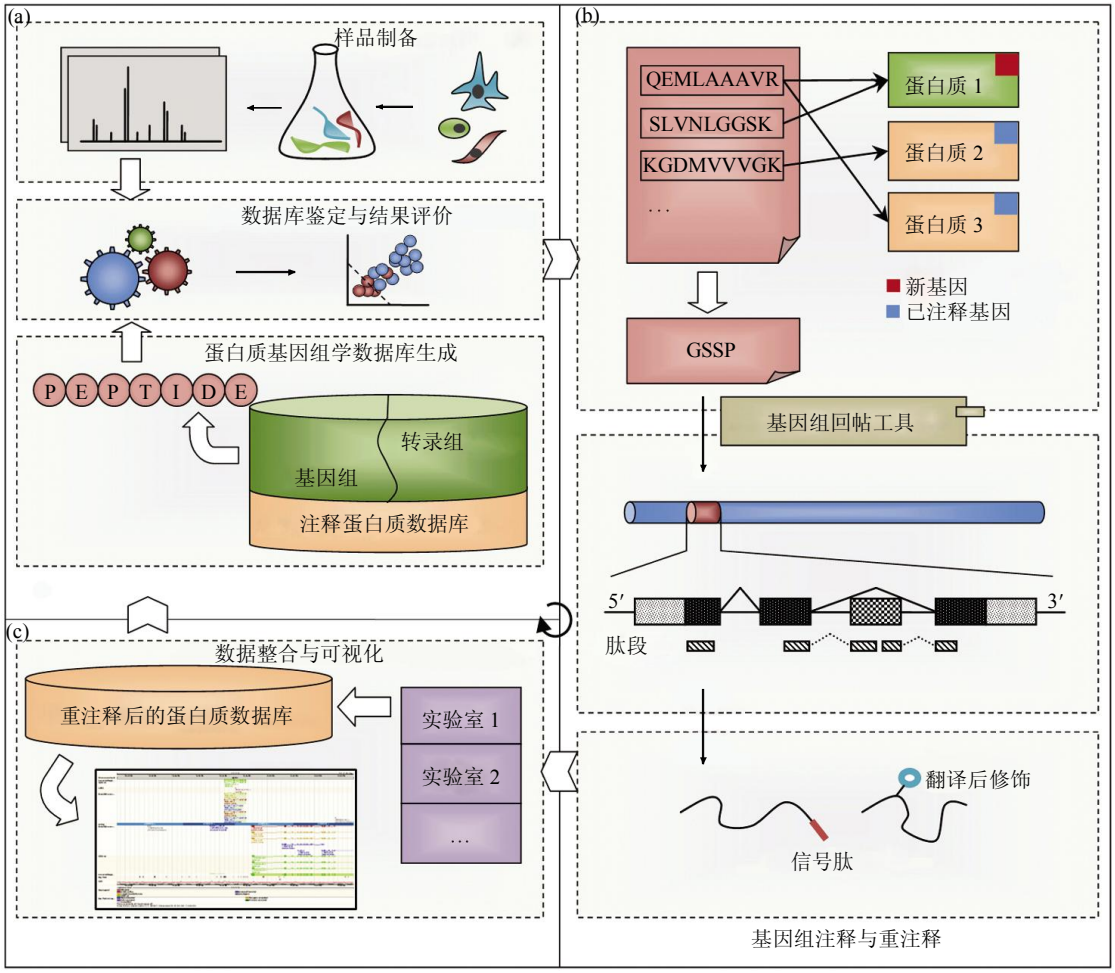

Fig. 3　A proteogenomic annotation pipeline

图 3　蛋白质基因组学注释流程

(a) 样品制备与质谱数据采集、数据库构建与鉴定．这里与传统蛋白质组学的不同点在于搜索用数据库结合了基因组、转录组数据和数据库．(b) 基因组注释与重注释．高可信肽段用于验证已注释基因组，校正原注释基因组的错误，解释翻译后处理现象．(c) 数据整合与可视化．多实验室结果经整合和处理后，可以不断完善蛋白质数据库，最终达到对基因组的完整注释．

虽然蛋白质基因组学的研究发展较快，但是也存在以下三方面的问题：a．在数据库构建方面，直接使用基因组构建真核生物蛋白质基因组学数据库比较困难，也无法应对大数据、大数据库的挑战，所以目前应用并不广泛．转录组相比基因组复杂度低，所以越来越多的研究工作开始使用转录组数据进行数据库构建，但如何使用较好的存储结构来去除数据冗余性是非常值得研究的问题．b．大部分研究文献存在数据发表的质量控制问题，比如使用 PSM 水平的 FDR 直接获得鉴定蛋白质集合，没有进行蛋白质水平的 FDR 控制．c．多数据整合和质量控制工具非常缺乏，无法实现增量式的基因

组注释，这在很大程度上阻碍了蛋白质基因组学的发展．数据量的逐渐增大，以及数据共享和传输的不便，也限制了蛋白质基因组学的推广．

4 未来展望

蛋白质基因组学直接对编码基因的表达产物——蛋白质进行研究，从而验证和校正基因组注释结果，发现新基因，该方法有别于基因组学、转录组学和功能基因组学，对解释基因组注释结果和理解生命现象来说是非常重要的．在无法获得转录组数据的研究内容中，比如动物的血液和其他组织液中，蛋白质组学或者蛋白质基因组学更是起到了不可替代的作用[93]．越来越多的研究人员提倡在基因组注释工程中加入标准蛋白质组分析作为互补，甚至直接采用蛋白质基因组学的方法对完成测序的基因组进行注释[50, 94-96]，体现了蛋白质基因组学在蛋白质组层面上注释基因组的优势．

蛋白质基因组学的发展依赖蛋白质组学相关技术的发展，主要包括质谱技术和数据处理软件两方面性能的提升．质谱技术最近发展较快，比如 Orbitrap Elite 和 Triple TOF 5600 仪器的出现，使得质谱仪采集一级和二级双高精度质谱数据的速度进一步加快，在此基础上结合一二级谱的智能数据采集系统的研发[97]，将会扩展质谱仪的动态范围，使更多的肽段得到碎裂．ETD 和 HCD 等新的碎裂模式也会丰富二级谱图信息，有利于数据库搜索和评价引擎获取更多的可靠肽谱匹配结果，同时降低随机匹配概率，提高基因组注释的可靠性．数据处理软件也有新的进展，比如串联质谱图从头预测算法的发展一定程度上解决了物种基因组缺失的问题[72, 98]；定量蛋白质组学方法使得在蛋白质组水平上测量表达量成为可能，进而结合其他组学定量信息可以在系统生物学层面上注释该基因的功能；多组学信息整合的方法，比如综合使用近源基因组序列或者转录组数据，提高了注释的可靠性[22]．但同时我们也看到，数据库构建和单次实验注释结果错误率控制的认识和研究依然需要进一步加强，也是有望优先解决的问题．而多次实验结果整合和错误率控制问题，是实现增量式注释基因组最核心的问题，但是在该方向上现存的算法工具非常少，希望在未来一段时间内能有标志性的进展．

基因组学、转录组学、蛋白质组学的通量在不断提升，数据信息也更加丰富，蛋白质基因组学作为一种组学级别的注释工具，体现了其在广度上、精度上和通量上的优越性．在不久的将来，蛋白质基因组学注释工具必将纳入到基因组注释工程当中，成为其中一项标准的注释流程，并结合基因组、转录组技术和方法，更好地为基因组注释服务．

2011 年国际人类蛋白质研究组织(HUPO)启动了人类蛋白质组计划(Human Proteome Project，HPP) [99-100] 及人类染色体蛋白质组计划(Chromosome-Centric Human Proteome Project，C-HPP)[101-102]．该项目以蛋白质组学数据为基础，整合基因组学、转录组学数据，对可变剪接、SNP、以及三类主要翻译后修饰(磷酸化、乙酰化、糖基化)进行注释，以加深人们对蛋白质(基因)功能的理解，从而指导疾病的研究．这将是蛋白质基因组学一个重要的尝试和应用，同时也对数据整合、共享及可视化工具提出了挑战．

致谢 感谢北京基因组所王全会博士和刘斯奇研究员对本工作的大力支持，感谢北京蛋白质组研究中心徐平研究员、姜颖研究员对本文的修改建议．

参 考 文 献

[1] Stein L. Genome annotation: from sequence to biology. Nat Rev Genet, 2001, **2**(7): 493-503

[2] Brent M R. Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res, 2005, **15**(12): 1777-1786

[3] Brent M R. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet, 2008, **9**(1): 62-73

[4] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature, 2003, **422**(6928): 198-207

[5] Cravatt B F, Simon G M, Yates J R, 3rd. The biological impact of mass-spectrometry-based proteomics. Nature, 2007, **450** (7172): 991-1000

[6] Nesvizhskii A I, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods, 2007, **4**(10): 787-797

[7] Renuse S, Chaerkady R, Pandey A. Proteogenomics. Proteomics, 2011, **11**(4): 620-630

[8] Jaffe J D, Berg H C, Church G M. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics, 2004, **4**(1): 59-77

[9] Kunec D, Nanduri B, Burgess S C. Experimental annotation of channel catfish virus by probabilistic proteogenomic mapping. Proteomics, 2009, **9**(10): 2634-2647

[10] Baudet M, Ortet P, Gaillard J C, *et al*. Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use

of non-canonical translation initiation codons. Mol Cell Proteomics, 2010, **9**(2): 415−426

[11] de Groot A, Dulermo R, Ortet P, *et al*. Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium Deinococcus deserti. PLoS Genet, 2009, **5**(3): e1000434

[12] de Souza G A, Malen H, Softeland T, *et al*. High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using *Mycobacterium tuberculosis* as an example. BMC Genomics, 2008, **9**: 316

[13] de Souza G A, Arntzen M O, Fortuin S, *et al*. Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. Mol Cell Proteomics, 2011, **10**(1): M110.002527

[14] Kelkar D S, Kumar D, Kumar P, *et al*. Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. Mol Cell Proteomics, 2011, **10**(12): M111.011627

[15] Gupta N, Benhamida J, Bhargava V, *et al*. Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res, 2008, **18**(7): 1133−1142

[16] Gupta N, Tanner S, Jaitly N, *et al*. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res, 2007, **17**(9): 1362−1377

[17] Ansong C, Tolic N, Purvine S O, *et al*. Experimental annotation of post-translational features and translated coding regions in the pathogen *Salmonella Typhimurium*. BMC Genomics, 2011, **12**: 433

[18] Wei C, Peng J, Xiong Z, *et al*. Subproteomic tools to increase genome annotation complexity. Proteomics, 2008, **8**(20): 4209−4213

[19] Zhao L, Liu L, Leng W, *et al*. A proteogenomic analysis of *Shigella* flexneri using 2D LC-MALDI TOF/TOF. BMC Genomics, 2011, **12**: 528

[20] Ishino Y, Okada H, Ikeuchi M, *et al*. Mass spectrometry-based prokaryote gene annotation. Proteomics, 2007, **7**(22): 4053−4065

[21] Payne S H, Huang S T, Pieper R. A proteogenomic update to Yersinia: enhancing genome annotation. BMC Genomics, 2010, **11**: 460

[22] Schrimpe-Rutledge A C, Jones M B, Chauhan S, *et al*. Comparative omics-driven genome annotation refinement: application across Yersiniae. PLoS One, 2012, **7**(3): e33903

[23] Oshiro G, Wodicka L M, Washburn M P, *et al*. Parallel identification of new genes in *Saccharomyces cerevisiae*. Genome Res, 2002, **12**(8): 1210−1220

[24] Helsens K, Van Damme P, Degroeve S, *et al*. Bioinformatics analysis of a *Saccharomyces cerevisiae* N-terminal proteome provides evidence of alternative translation initiation and post-translational N-terminal acetylation. J Proteome Res, 2011, **10**(8): 3578−3589

[25] Baerenfaller K, Grossmann J, Grobei M A, *et al*. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science, 2008, **320**(5878): 938−941

[26] Castellana N E, Payne S H, Shen Z, *et al*. Discovery and revision of *Arabidopsis* genes by proteogenomics. Proc Natl Acad Sci USA, 2008, **105**(52): 21034−21038

[27] Chaerkady R, Kelkar D S, Muthusamy B, *et al*. A proteogenomic analysis of *Anopheles gambiae* using high-resolution Fourier transform mass spectrometry. Genome Res, 2011, **21**(11): 1872−1881

[28] Kalume D E, Peri S, Reddy R, *et al*. Genome annotation of *Anopheles gambiae* using mass spectrometry-derived data. BMC Genomics, 2005, **6**: 128

[29] Brosch M, Saunders G I, Frankish A, *et al*. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. Genome Res, 2011, **21**(5): 756−767

[30] Fermin D, Allen B B, Blackwell T W, *et al*. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol, 2006, **7**(4): R35

[31] Tanner S, Shen Z, Ng J, *et al*. Improving gene annotation using peptide mass spectrometry. Genome Res, 2007, **17**(2): 231−239

[32] Desiere F, Deutsch E W, Nesvizhskii A I, *et al*. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol, 2005, **6**(1): R9

[33] Allmer J, Naumann B, Markert C, *et al*. Mass spectrometric genomic data mining: Novel insights into bioenergetic pathways in *Chlamydomonas reinhardtii*. Proteomics, 2006, **6**(23): 6207−6220

[34] Specht M, Stanke M, Terashima M, *et al*. Concerted action of the new Genomic Peptide Finder and AUGUSTUS allows for automated proteogenomic annotation of the *Chlamydomonas reinhardtii* genome. Proteomics, 2011, **11**(9): 1814−1823

[35] Chang K Y, Georgianna D R, Heber S, *et al*. Detection of alternative splice variants at the proteome level in *Aspergillus flavus*. J Proteome Res, 2010, **9**(3): 1209−1217

[36] Chang K Y, Muddiman D C. Identification of alternative splice variants in *Aspergillus flavus* through comparison of multiple tandem MS search algorithms. BMC Genomics, 2011, **12**: 358

[37] Merrihew G E, Davis C, Ewing B, *et al*. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res, 2008, **18**(10): 1660−1669

[38] Bindschedler L V, Burgis T A, Mills D J, *et al*. In planta proteomics and proteogenomics of the biotrophic barley fungal pathogen *Blumeria graminis* f. sp. hordei. Mol Cell Proteomics, 2009, **8**(10): 2368−2381

[39] Adamidi C, Wang Y, Gruen D, *et al*. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Res, 2011, **21**(7): 1193−1200

[40] Ahrens C H, Brunner E, Qeli E, *et al*. Generating and navigating proteome maps using mass spectrometry. Nat Rev Mol Cell Biol, 2010, **11**(11): 789−801

[41] de Godoy L M, Olsen J V, Cox J, *et al*. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature, 2008, **455**(7217): 1251−1254

[42] Lavigne R, Becker E, Liu Y, *et al*. Direct iterative protein profiling

(DIPP) - an innovative method for large-scale protein detection applied to budding yeast mitosis. Mol Cell Proteomics, 2012, **11**(2): M111 012682

[43] Mercer T R, Dinger M E, Mattick J S. Long non-coding RNAs: insights into functions. Nat Rev Genet, 2009, **10**(3): 155-159

[44] Bu D, Yu K, Sun S, *et al*. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res, 2012, **40**(Database issue): D210-215

[45] Liao Q, Xiao H, Bu D, *et al*. ncFANs: a web server for functional annotation of long non-coding RNAs. Nucleic Acids Res, 2011, **39**(Web Server issue): W118-124

[46] Venter E, Smith R D, Payne S H. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One, 2011, **6**(11): e27587

[47] Mo F, Hong X, Gao F, *et al*. A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. BMC Bioinformatics, 2008, **9**: 537

[48] Gallien S, Perrodou E, Carapito C, *et al*. Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res, 2009, **19**(1): 128-135

[49] Gevaert K, Goethals M, Martens L, *et al*. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat Biotechnol, 2003, **21**(5): 566-569

[50] Armengaud J. A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr Opin Microbiol, 2009, **12**(3): 292-300

[51] Ansong C, Purvine S O, Adkins J N, *et al*. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic, 2008, **7**(1): 50-62

[52] Schandorff S, Olsen J V, Bunkenborg J, *et al*. A mass spectrometry-friendly database for cSNP identification. Nat Methods, 2007, **4**(6): 465-466

[53] Wang X, Slebos R J, Wang D, *et al*. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res, 2012, **11**(2): 1009-1017

[54] Zougman A, Ziolkowski P, Mann M, *et al*. Evidence for insertional RNA editing in humans. Curr Biol, 2008, **18**(22): 1760-1765

[55] Fu Y, Xiu L Y, Jia W, *et al*. DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data. Mol Cell Proteomics, 2011, **10**(5): M110.000455

[56] Krug K, Nahnsen S, Macek B. Mass spectrometry at the interface of proteomics and genomics. Mol Biosyst, 2011, **7**(2): 284-291

[57] Yates J R, 3rd, Eng J K, McCormack A L. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem, 1995, **67**(18): 3202-3210

[58] Roos F F, Jacob R, Grossmann J, *et al*. PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. Bioinformatics, 2007, **23**(22): 3016-3023

[59] Ferro M, Tardif M, Reguer E, *et al*. PepLine: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. J Proteome Res, 2008, **7**(5): 1873-1883

[60] Xing X B, Li Q R, Sun H, *et al*. The discovery of novel protein-coding features in mouse genome based on mass spectrometry data. Genomics, 2011, **98**(5): 343-351

[61] Ning K, Nesvizhskii A I. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. BMC Bioinformatics, 2010, **11**(Suppl 11): S14

[62] Edwards N J. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol, 2007, **3**: 102

[63] Yates J R, 3rd, Eng J K, McCormack A L, *et al*. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem, 1995, **67**(8): 1426-1436

[64] Perkins D N, Pappin D J, Creasy D M, *et al*. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999, **20**(18): 3551-3567

[65] Fu Y, Yang Q, Sun R, *et al*. Exploiting the kernel trick to correlate fragment ions for peptide identification *via* tandem mass spectrometry. Bioinformatics, 2004, **20**(12): 1948-1954

[66] Keller A, Nesvizhskii A I, Kolker E, *et al*. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem, 2002, **74**(20): 5383-5392

[67] Kall L, Canterbury J D, Weston J, *et al*. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods, 2007, **4**(11): 923-925

[68] Ma B, Johnson R. De novo sequencing and homology searching. Mol Cell Proteomics, 2012, **11**(2): O111 014902

[69] Liu X, Han Y, Yuen D, *et al*. Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. Bioinformatics, 2009, **25**(17): 2174-2180

[70] Pawar H, Sahasrabuddhe N A, Renuse S, *et al*. A proteogenomic approach to map the proteome of an unsequenced pathogen - Leishmania donovani. Proteomics, 2012, **12**(6): 832-844

[71] Zhao Y, Sun W, Zhang P, *et al*. Nematode sperm maturation triggered by protease involves sperm-secreted serine protease inhibitor (Serpin). Proc Natl Acad Sci USA, 2012, **109**(5): 1542-1547

[72] Chi H, Sun R X, Yang B, *et al*. pNovo: de novo peptide sequencing and identification using HCD spectra. J Proteome Res, 2010, **9**(5): 2713-2724

[73] Castellana N, Bafna V. Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics, 2010, **73**(11): 2124-2135

[74] Nesvizhskii A I, Roos F F, Grossmann J, *et al*. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics, 2006, **5**(4): 652-670

[75] Zhou C, Chi H, Wang L H, *et al*. Speeding up tandem mass spectrometry-based database searching by longest common prefix. BMC Bioinformatics, 2010, **11**: 577

[76] Helmy M, Tomita M, Ishihama Y. Peptide identification by searching large scale tandem mass spectra against large databases bioinformatics methods in proteogenomics. Genes, Genomes and Genomics, 2012, **6**(Special Issue 1): 76−85

[77] Milloy J A, Faherty B K, Gerber S A. Tempest: GPU-CPU computing for high-throughput database spectral matching. J Proteome Res, 2012, **11**(7): 3581−3591

[78] Wang L, Wang W, Chi H, *et al*. An efficient parallelization of phosphorylated peptide and protein identification. Rapid Commun Mass Spectrom, 2010, **24**(12): 1791−1798

[79] Sanders W S, Wang N, Bridges S M, *et al*. The proteogenomic mapping tool. BMC Bioinformatics, 2011, **12**: 115

[80] Nesvizhskii A I, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics, 2005, **4**(10): 1419−1440

[81] Nesvizhskii A I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics, 2010, **73**(11): 2092−2123

[82] Huang T, Wang J, Yu W, *et al*. Protein inference: a review. Brief Bioinform, 2012, **13**(5): 586−614

[83] Kim S, Gupta N, Pevzner P A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J Proteome Res, 2008, **7**(8): 3354−3363

[84] Gupta N, Bandeira N, Keich U, *et al*. Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom, 2011, **22**(7): 1111−1120

[85] Qeli E, Ahrens C H. PeptideClassifier for protein inference and targeted quantitative proteomics. Nat Biotechnol, 2010, **28**(7): 647−650

[86] Gerster S, Qeli E, Ahrens C H, *et al*. Protein and gene model inference based on statistical modeling in k-partite graphs. Proc Natl Acad Sci USA, 2010, **107**(27): 12101−12106

[87] Ramakrishnan S R, Vogel C, Prince J T, *et al*. Integrating shotgun proteomics and mRNA expression data to improve protein identification. Bioinformatics, 2009, **25**(11): 1397−1403

[88] Deshayes C, Perrodou E, Gallien S, *et al*. Interrupted coding sequences in *Mycobacterium smegmatis*: authentic mutations or sequencing errors?. Genome Biol, 2007, **8**(2): R20

[89] Christie-Oleza J A, Miotello G, Armengaud J. High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better genomic annotation for the whole marine Roseobacter clade. BMC Genomics, 2012, **13**: 73

[90] Desiere F, Deutsch E W, King N L, *et al*. The PeptideAtlas project. Nucleic Acids Res, 2006, **34**(Database issue): D655−658

[91] Peterson E S, McCue L A, Schrimpe-Rutledge A C, *et al*. VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data. BMC Genomics, 2012, **13**: 131

[92] Shteynberg D, Deutsch E W, Lam H, *et al*. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics, 2011, **10**(12): M111 007690

[93] Cox J, Mann M. Is proteomics the new genomics?. Cell, 2007, **130**(3): 395−398

[94] Zhong Y, Chang X, Cao X J, *et al*. Comparative proteogenomic analysis of the *Leptospira* interrogans virulence-attenuated strain IPAV against the pathogenic strain 56601. Cell Res, 2011, **21**(8): 1210−1229

[95] Jaffe J D, Stange-Thomann N, Smith C, *et al*. The complete genome and proteome of *Mycoplasma* mobile. Genome Res, 2004, **14**(8): 1447−1461

[96] Lazarev V N, Levitskii S A, Basovskii Y I, *et al*. Complete genome and proteome of *Acholeplasma* laidlawii. J Bacteriol, 2011, **193**(18): 4943−4953

[97] Graumann J, Scheltema R A, Zhang Y, *et al*. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. Mol Cell Proteomics, 2012, **11**(3): M111 013185

[98] 马 洁, 吴松峰, 朱云平. 蛋白质组学中新蛋白质鉴定的研究方法和策略. 生物化学与生物物理进展, 2007, **34**(8): 791−799
Ma J, Wu S F, Zhu Y P. Prog Biochem Biophys, 2007, **34**(8): 791−799

[99] Hancock W, Omenn G, Legrain P, *et al*. Proteomics, human proteome project, and chromosomes. J Proteome Res, 2011, **10**(1): 210

[100] Legrain P, Aebersold R, Archakov A, *et al*. The human proteome project: current state and future direction. Mol Cell Proteomics, 2011, **10**(7): M111 009993

[101] Paik Y K, Omenn G S, Uhlen M, *et al*. Standard guidelines for the chromosome-centric human proteome project. J Proteome Res, 2012, **11**(4): 2005−2013

[102] Paik Y K, Jeong S K, Omenn G S, *et al*. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol, 2012, **30**(3): 221−223

Proteogenomics: Improving Genomes Annotation by Proteomics*

ZHANG Kun[1,2], WANG Le-Heng[1], CHI Hao[1,2], BU De-Chao[1,2], YUAN Zuo-Fei[1,2], LIU Chao[1,2], FAN Sheng-Bo[1,2], CHEN Hai-Feng[1,2], ZENG Wen-Feng[1,2], LUO Hai-Tao[1], SUN Rui-Xiang[1], HE Si-Min[1], XIE Lu[3], ZHAO Yi[1]**

([1] *Key Laboratory of Intelligent Information Processing-Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, China;
[2] *University of Chinese Academy of Sciences, Beijing* 100049, China;
[3] *Shanghai Center for Bioinformation Technology, Shanghai* 200235, China)

Abstract With the rapid development of high throughput DNA sequencing, genomes of more and more species have been sequenced. Identifying and determining the structures of coding genes are the basic tasks of genome annotation. To understand the sequenced genome precisely, it is necessary to integrate muilti-"omics" data to annotate genomes. However, the annotation methods developed in the past decade are mainly based on genome and transcriptome data. Recently, mass spectrometry based proteomics has come of age, which can cover proteomes nearly completely and make it possible to use mass spectrometry data to annotate genomes. Mass spectrometry data can verify annotated genes on one hand, and refine annotated genes, discover novel genes on the other, which achieve the goal of re-annotating genomes. This is exactly the research content of proteogenomics, and using proteogenomic techniques to systematically annotate the sequenced genome is becoming increasingly important. This article reviews the research content, methods and recent trends of proteogenomics.

Key words proteogenomics, genome annotation, proteomics, mass spectrometry
DOI: 10.3724/SP.J.1206.2012.00263

* This work was supported by grants from National Basic Research Program of China (2010CB912701, 2010CB912702) and CAS Knowledge Innovation Program under Grant (KGCX1-YW-13).
**Corresponding author.
Tel: 86-10-62601016, E-mail: biozy@ict.ac.cn
Received: May 31, 2012 Accepted: January 11, 2013