计算分子生物学(网络版), 2015 年, 第 4 卷, No.1, 1−12
Jisuan Fenzi Shengwuxue (Online), 2015, Vol.4, No.1, 1−12

http://cmb.biopublisher.cn Doi: 10.5376/cmb.cn.2015.04.0001

特邀评述
Invited Review

# 蛋白质基因组学研究中的质谱仪与生物信息学方法

巩鹏涛 [1,3], 徐润生 [2], 方宣钧 [1,2,3]

1. 东北林业大学盐碱地生物资源环境研究中心, 哈尔滨, 150040
2. 浙江农林大学暨阳学院, 暨阳国际先进技术研究中心, 诸暨, 311800
3. 海南省热带农业资源开发利用研究所, 三亚, 572025

通讯作者：jim.xj.fang@hitar.org； 作者

计算分子生物学, 2015 年, 第 4 卷, 第 1 篇 doi: 10.5376/cmb.cn.2015.04.0001



引用格式(中文):
巩鹏涛等, 2015, 蛋白质基因组学研究中的质谱仪与生物信息学方法, 计算分子生物学(online), 4(1): 1-12 (doi: 10.5376/cmb.cn.2015.04.0001)
引用格式(英文):
Gong et al., 2015, The Spectrometry and Bioinformatics in the Studies of Proteogenomics, Jisuan Fenzi Shengwuxue (online), 4(1): 1-12 (doi: 10.5376/cmb.cn.2015.04.0001)

**摘 要** 随着串联质谱的蛋白质组学技术和高通量基因组测序技术的迅速发展，大大促进了组学之间的交叉和融合，作为基因组学和蛋白组学的一个集合，蛋白质基因组学(Proteogenomics)应运而生。蛋白质基因组学的核心任务是获得足够深度和广度的蛋白质组，转录组和基因组等数据, 并将所获得的组学数据进行整合分析，以便从整体的层面进行系统生物学研究。显然，蛋白质基因组学的研究涉及到蛋白质组和基因组数据的获取及后续的交叉大数据分析。针对蛋白质基因组学的分析研究中，本文总结了质谱仪和蛋白质基因组学流程分析软件的选择，重点评述了常用的生物信息学计算工具，如 PepLine、Proteogenomic Mapping Tool InsPecT、iPiG、PGP、Peppy、Bacterial Proteogenomic Pipeline、SearchGUI、ProteoAnnotator、Genosuite 等。

**关键词** 蛋白质基因组学; 质谱仪; 蛋白质组; 基因组注释; 生物信息学

# The Spectrometry and Bioinformatics in the Studies of Proteogenomics

Gong Pengtao [1,3], Xu Runsheng [2], Fang Xuanjun [1,2,3]

1. Alkali Soil Natural Environmental Science Center (ASNESC), Northeast Forestry University, Harbin, 150040, P.R. China
2. Jiyang International Center for Advanced Technology (JICAT), Jiyang College of Zhejiang A&F University, Zhuji, 311800, P.R. China
3. Hainan Institute of Tropical Agricultural resources (HITAR), Sanya, 572025, P.R. China

Corresponding author, jim.xj.fang@hitar.org; Authors

**Abstract** With the rapid development of proteomics technology based tandem mass spectrometry and high-throughput genome sequencing technology, it has greatly facilitated the cross and the integration among the omics, therefore proteogenomics came into being as an area at the interface of genomics and proteomics. The core task of proteogenomics is to obtain the data of proteome, genome and transcription in sufficient depth and breadth and to integrate the obtained omic data for the studies of systems biology in the overall level. Clearly, the study of proteogenomics involves in acquisition of proteomic and genomic data as well as subsequent cross analysis of big data. This review summarized the choosing of spectrometers and proteogenomic pipeline softwares, and mainly commented on some frequently-used bioinformatic softwares such as PepLine, Proteogenomic Mapping Tool InsPecT, iPiG, PGP, Peppy, Bacterial Proteogenomic Pipeline, Genosuite, SearchGUI, and ProteoAnnotator, etc. .

**Keywords** Proteogenomics; Mass spectrometry; Genome annotation; Proteomics; Bioinformatics

## 研究背景

应用串联质谱的数据搜索蛋白质数据库，已经发展成为成熟的研究生物样本中蛋白质组成的方法。然而，如果被搜索的参考蛋白质数据库只是包含了已知的蛋白质，那么利用该方法发掘新蛋白或修饰蛋白的能力就有一定的局限。

在人类疾病的分析中，如分析癌症，异常的表达模式可能产生与疾病相关的蛋白质，而此类蛋白质在参考蛋白质数据库中是往往是不存在的(Xu and Lee, 2003)。因此，搜索常用的标准蛋白质数据，如 ProtKB 或 Swiss-Prot 等，就极可能错失那些与癌症相关的新蛋白质或修饰模式。蛋白质基因组学

收稿日期：2015 年 01 月 11 日
接受日期：2015 年 02 月 12 日
发表日期：2015 年 02 月20 日
基金项目：本研究由海南省热带农业资源研究所微生物基因组测序及生物信息学研究项目资助

(Proteogenomics)分析策略的使用极大的改善了这种情况(Jaffe, 2004)。

蛋白质基因组学的分析不仅可以发现基因组上的新的编码区和修正基因正确的起始和终止位置，还可以用来鉴定真核生物普遍存在的剪切变异体。在构建人类全蛋白质组草图的研究中蛋白质基因组学发挥了重要的作用，Kim 等(2014)人使用该策略发掘出了 808 个新的人类基因组注释，其中包括 140 个假基因的翻译，44 个新的 ORFs，106 个在原有注释基因结构内部的新的编码区或外显子，110 个基因/蛋白质/外显子扩展事件，198 个新的蛋白质 N 末端和 201 个新的信号肽切割位点。

Zhang 等人对结肠和直肠癌蛋白质基因组学研究更是表明，虽然 mRNA 和蛋白质的水平有一定的相关性，但在这些组织体中蛋白质的丰度不能简单的依据 DNA 或 RNA 水平推导出来，因为有超过三分之二的相关性是不显著的(Zhang et al., 2014)。拟南芥，玉米和原核生物中，蛋白质基因组的分析也展现了良好的发展潜力(Kucharova, 2014; Castellana et al., 2014; Castellana et al., 2008)。

蛋白质基因组学的核心任务是获得足够深度和广度的蛋白质组，转录组和基因组等数据，并将所获得的组学数据进行整合分析，以便从整体的层面进行系统生物学研究。蛋白质基因组学的研究过程中涉及到蛋白质组和基因组数据的获取的工具适合程度和复杂的分析流程及后续的交叉大数据分析(Gong et al., 2014)。本文综述了在蛋白质基因组学的研究中的质谱仪的选择和蛋白质基因组学流程分析软件的研究进展。

## 1 蛋白质基因组学中的质谱仪的使用

目前，质谱仪种类繁多，适合蛋白质和肽段分析的质谱仪的原理可以参考已有的相关综述文献(Ahmed,. 2008; Thelen and Miernyk, 2012)，表 1 列出了各种质谱仪的指标。从质谱的分辨率方面大致可以分为两类，低分辨率如离子阱质谱仪(ion traps)、三级四极质谱仪(triple quadrupoles)，高分辨率如飞行时间质谱仪(Q-TOF)，傅立叶变换离子回旋共振质谱仪(FTICR)，轨道阱质谱仪(Orbitrap)。

### 1.1 离子阱质谱仪

离子阱质谱仪，由于其具有高灵敏性和快速的扫描速度可以确保蛋白质组的高覆盖率，在蛋白质基因组学研究中使用最为广泛。然而，应用离子阱质谱仪得到的数据解析度和准确性范围在 0.2~0.5 Da (200~500 ppm 在 1000 m/z)，存在一定的局限性。较低的灵敏度使离子阱质谱仪需要高的质量允差度(mass tolerance)，因此增加了搜索空间并降低了搜索的准确度。

另外，大多数离子阱质谱仪不具备同时存储所有的碎片离子的能力，对低质量的碎片离子覆盖率比较差，并导致最终的模糊不清的肽段分配。这一局限性已经在新的离子阱质谱仪中得到改进，例如，通过 pulsed-Q 解离方法。然而在与前体离子的低质量准确性结合时依然可能出现偏差，从而可能增大搜索空间并增加蛋白质数据库搜索时的错误发现率(false discovery rate) (Krug et al., 2011)。

### 1.2 傅立叶变换离子回旋共振质谱仪与轨道阱质谱仪

高灵敏度的质谱仪器，例如 FT ICR 和 Q-TOF 都具有很好的解析度，可以达到亚-ppm 级的质量灵敏度，这将可以减低搜索空间并增加蛋白质数据库搜索时的准确性。然而，高灵敏度的质谱仪存在一个缺点就是相对较低的数据获取速度，导致在分析复杂肽段混合物时的采样不足，并进而影响蛋白质组的覆盖率。

新一代的混合质谱仪则通过低解析和高解析质谱仪并用的肽段检测方法，解决这一问题。典型的实例就是线性离子阱静电场轨道阱组合质谱仪(LTQ-Orbitrip)，前体肽离子质量在轨道阱质谱仪中以高的解析度和灵敏度得到测定，而裂解后的肽在线性离子阱质谱仪中以较高的速度和灵敏性得到测定(Wenger et al., 2010)。最后得到的质谱数据不仅在肽段测序速度和质量灵敏度方面，而且在蛋白质组覆盖度和数据库搜索空间方面，取得了很好的平衡。

### 1.3 LTQ-Oibitrap Velos

最新型的 LTQ-Oibitrap Velos 质谱仪通过混合一个更快速、更灵敏的双压线性离子阱质谱系统(dual-pressure linear ion trap)和一个高灵敏度的 Orbitrap 质谱分析仪，在蛋白质组分析的深度和准确性方面显示出巨大的潜力。改良后的更高能的碰撞诱导解离碰撞室(higher energy collision dissociation, HCD)能确保获得高速和高质谱灵敏度的 MS 和 MS/MS 质谱，并且在质量范围内有良好的碎片离子覆盖度(Olsen et al., 2009)。该设备也适合于利用化学离子源来确保电子转移裂解(electron transfer dissociation, ETD)，这种方法可以获得更全面的多电荷离子的裂解(Wenger et al., 2010; Syka et al., 2004)，提供适合蛋白质基因组学研究使用的数据。

## 2 蛋白质基因组学中的计算工具

由于蛋白质基因组学具有相对固定的分析流程，发展一套完整的类似基因组测序和注释的蛋白质基因组学的软件是很有必要的，这样不仅可以将基因组学和蛋白质组学完整的结合，而且方便基因组的注释(Renuse et al., 2011)。适用于串联质谱为基础的蛋白质组工具众多(Nesvizhskii, 2010)，但并不是所有都适合于蛋白质基因组学。因此，开展蛋白质基因组学的研究最早开发起来的实用工具的相

表 1 可用于蛋白质分析的质谱仪(Thelen and Miernyk, 2012)
Table 1 Overview of mass spectrometers available for protein analysis (Thelen and Miernyk, 2012)

| 型号 | 厂商 | 分析仪 | 精度(PPM) | 灵敏度 | 范围(M/Z) | 分辨率(FWHM) | 扫描速度 | 碎片化 | 电离 |
|---|---|---|---|---|---|---|---|---|---|
| Orbitrap (with LTQ XL or Velos) | Thermo Scientific | Hybrid | <1 | Attomole1 | 50–4 000 | >100 000 | 1 spectra/s | CID, HCD, ETD, PQD | MALDI, HESI, ESI, nano, API, APCI, APCI/APPI |
| TSQ Vantage | Thermo Scientific | Triple quadrupole | 50* | Attomole1 | 10–1500 | 7500 | 5 000 units/s | CID | HESI, APCI/APPI, nano |
| Velos | Thermo Scientific | Linear ion trap | 50* | Attomole1 | 50–4 000 | >25 000 | 33 333 units/s | CID, ETD, PQD | ESI, APCI, APCI/APPI, nano |
| LTQ XL | Thermo Scientific | Linear ion trap | 50* | Attomole1 | 50–4 000 | >25 000 | 16 000 units/s | CID, ETD, PQD | ESI, APCI, APCI/APPI, nano |
| Xevo G2 Qtof | Waters | Quadrupole TOF | <1 | Femtomole2 | <100 000 | >22 500 | 30 spectra/s | CID | ESI/APCI/ESCi, APCI, APPI/APCI, nano, ASAP, APGC, TRIZAICTM |
| Xevo TQ-S | Waters | Tandem quadrupole | 50* | Attomole1 | 2–2 048 | Not provided | 10 000 units/s | CID | ESI/APCI/ESCi, APCI, APPI/APCI, nano, ASAP, APGC, TRIZAICTM |
| Synapt G2 HDMS | Waters | Quadrupole TOF | <1 | Femtomole2 | <100 000 | 40 000 | 20 spectra/s | CID | ESI/APCI/ESCi, APCI, APPI/APCI, nano, MALDI, ASAP, APGC, TRIZAICTM |
| SolariX FTMS | Bruker Daltonics | FTMS | <1 | Attomole3 | <100 000 | >1 000 000 | Not provided | CID, ETD, ECD, (SORI)-CID | ESI, nano, APCI, APPI |
| Amazon ETD | Bruker Daltonics | Linear ion trap | 50* | Attomole1 | 50–3 000 | 20 000 | 52 000 units/s | CID, ETD/PTR | ESI, APCI, APPI, nano, HPLC-Chip, ESI/MALDI |
| micrOTOF-Q II | Bruker Daltonics | Quadrupole TOF | <2 | Attomole4 | 50–20 000 | 20 000 | Not provided | CID | ESI, APCI, ESI/APCI, APPI, nano, CE/MS |
| 6 490 Triple Quadrupole | Agilent | Triple quadrupole | 50* | Zeptomole5 | 10–2 000 | Not provided | 150 MRM/s | CID | HPLC-chip, ESI, APCI, APPI, MMI |
| 6 500 Q-TOF | Agilent | Quadrupole TOF | <2 | Femtomol1 | 20–20 000 | 40 000 | 20 spectra/s | CID | ESI, APCI, ESI/APCI, APPI, MALDI, HPLC-Chip |
| TripleTOF 5 600 | ABSciex | Triple TOF | <1 | Femtomol1 | <40 000 | 40 000 | 100 spectra/s | CID | ESI/APCI, Turbo V, nano |
| Qtrap 5 500 | ABSciex | Triple quadruple | 100* | Femtomol1 | 5–1 250 | Not provided | 12 000 units/s | CID/ETD | APCI, Turbo V, nano, ESI/APCI, photo |
| | | Linear ion trap | 100* | Femtomol1 | 50–1 000 | 9 200 | 20 000 units/s | | |

注: *使用道尔顿为质量精度的仪器都通过 1000 的 m/z 使用转换为 p.p.m 单位; 灵敏度使用(1)利血平; (2) [Glu1]-血纤维蛋白肽 B; (3)泛素; (4) BSA 消化物或; (5)维拉帕米来判定; 所有的数据获取自制造商的官方网站(http://www.thermoscientific.com, http://www.waters.com, http://www.bdal.com, http://www.agilent.com 和 http://www.absciex.com)的仪器说明书或其销售代表提供的参数指标。APGC, 常压气相色谱(atmospheric pressure gas chromatography); APCI, 常压化学电离化(atmospheric pressure chemical ionization); API, 常压电离(atmospheric pressure ionization); APPI, 常压光电离(atmospheric pressure photoionization); ASAP, 常压固体分析探针(atmospheric solids analysis probe); CE/MS, 毛细管电泳质谱(capillary electrophoresis MS); CID, 碰撞诱导解离(collision-induced dissociation); ESCi, 电喷雾常压化学电离(electrospray atmospheric pressure chemical ionization); ESI, 电喷雾离子化(electrospray ionization); FTMS，傅里叶变换质谱(Fourier transform MS); HCD，高能量的 C−陷阱解离(higher-energy C-trap dissociation); HESI, 加热电喷雾电离(heated electrospray ionization); MMI, 多模式电离(multi-mode ionization); MRM, 多反应监测(multiple reaction monitoring); nano, 纳升喷雾(nanospray); photo, 光电离(photoionization); PQD, 脉冲 Q 碰撞诱导解离(pulsed Q collision-induced dissociation); PTR, 质子转移反应(proton transfer reaction); SORI, 持续非共振辐照碰撞诱导解离(sustained off-resonance irradiation)

关的搜索及蛋白质数据库构建相关的算法，代表性的有 InsPecT (Tanner et al., 2005)，GENQUEST (Sevinsky et al., 2008)和 PepSplice (Roos et al., 2007)，ABLCP (Zhou et al., 2010)，ByOnic (Bern et al., 2007)和 GenomicPeptideFinder (Allmer et al., 2004)等，这些可以说是蛋白质基因组学分析的基础。但在蛋白质基因组学发展的初期，整体流程化的软件仍然很匮乏，这种情况在近几年得到一定程度的改善。

通过质谱搜索鉴定得到的肽段，如何用来重新修正和注释原基因组是这类软件工具的核心任务，表 2 列出一些适用于蛋白质基因组学的开放源代码工具。早期的软件主要集中在将新发现的 PSMs 在基因组上可视化方面，随着蛋白质基因组学分析策略在基因组注释方面展现出较高的应用价值，将蛋白质基因组学分析策略集成到基因组注释中，使其成为测序物种基因组注释的固定组成部分就成为工作的目标。因此，整个蛋白质基因组学分析流程的系统化整合工具的开发就显得方兴未艾。

### 2.1 PepLine

PepLine 是一个将 MS/MS 质谱通过 *de novo* 方法鉴定的蛋白酶酶切的肽定位到基因组序列的全自动化的软件(Ferro et al., 2008)。PepLine 包含三个模块：Taggor，PMMatch 和 PMClust。这一方法是基于在第一个模块通过四极杆飞行时间(quadrupole time-of-flight, QTOF)串联质谱获得肽序列标记(peptide sequence tags, PSTs)，在第二个模块中这些 PSTs 根据其比值被定位回六码框翻译的基因组序列，在第三个模块中这些比值被聚类分析以鉴定潜在的编码区。该方法在处理大数据和大度真核生物基因组方面有足够的速度，并可以用来鉴定基因的内含子和外显子结构。需要注意到的是 Taggor 模块是特别为 QTOF 串联质谱数据设定的，因此，在分析其他类型的串联质谱数据时需要使用其他程序代替 Taggor 模块。

### 2.2 Proteogenomic Mapping Tool

Proteogenomic Mapping Tool 是基于 Aho-Corasick 字符串搜索算法的 Java 编程的单机版跨平台应用(Sanders et al., 2011)。与 PepLine 不同，Proteogenomic Mapping Tool 是利用质谱数据搜索基因组的六码框翻译数据库鉴定的独有肽段(Unique Peptides)，并将这些肽段定位回其翻译的基因组中。三个输入文件：FASTA 格式的要定位回去的肽段，FASTA 格式的肽段要定位回去的基因组序列和遗传密码表文件。三个输出文档：包含产生的 ePSTs 的 FASTA 格式文档，详细的制表符分隔的文本文档，主要包含 ePST 的在基因组上的匹配位置信息等，ePSTs 的 GFF3 格式文档，便于研究者快速将其导入基因组阅览器实现数据的可视化。

VESPA (Visual Exploration and Statistics to Promote Annotation)是基于 Java 的可互动的整合蛋白质组(肽段)和转录组数据(RNA-Seq)来注释修改原核生物基因组的单机版软件(Peterson et al., 2012)。该工具可视化基因组所有的潜在读码框，通过可查询多层次基因组信息的可视化整合来发现在某些区域的高度可能的错误注释，通过 SVM 技术评估酶切肽段(SVM technique evaluate proteotypic peptides, STEPP)的统计方法来对可视化的肽段进行过滤。序列可以直接通过 BLAST 比对公共数据库进一步分析和验证。

### 2.3 iPiG

iPiG (integrating peptide spectrum matches into genome browser visualizations)是基于 Java 的单机运行的有良好用户界面的工具，方便将鉴定的肽段在基因组阅览器中很好的可视化(Kuhring et al., 2012)。其输入三个必需文件为：mzIdentML 格式或制表符分隔文本格式的 PSMs 文档；UCSC 表格格式的参考基因组注释的文本文档和 UCSC 表格格式的对应氨基酸翻译的文本文档。两个可选文档是：UniProt 数据库阐明蛋白质和基因匹配情况的 id 映射文档(id-mapping)和 FASAT 格式的包含用来肽段鉴定蛋白质的蛋白质组文档。其输出文档包含三个文件类型：BED (browser extensible data)，GFF3 (generic feature format version 3)和文本。iPiG 的特点就是搜集了在蛋白质鉴定过程中的信息，特别是考虑了肽段和蛋白质的匹配，保证了更加特异更加快速的肽段到基因的定位。

### 2.4 PGP

PGP 是基于 Python 和 C++设计服务于消息传递接口高通量的批处理集群多核工作站的并行原核生物蛋白质基因组学流程工具(Tovchigrechko et al., 2014)。串联质谱数据通过 InsPecT 搜索基因组六码框的翻译，随后使用 PepNovo 和 MSGF 重新计算得分。那些 Pvalue 值为 e−10 或者更好的 PSMs (肽段水平发现错误率大约 0.3%, 质谱发现错误率 0.01%)的肽被定位回其基因组位置。使用五个 ORF 过滤条件对在一个 ORF 中的肽段进行聚合分析，过滤掉低复杂度的甘氨酸和丙氨酸组成大约 70%的肽段；去除超过 750 bp 来自下一个编码框肽段的；过滤掉 ORFs 缺少一个独有定位肽段的或缺少一个完全胰蛋白酶酶切的肽段；每个蛋白质至少有两个肽段。该流程可以在多水平输出分析结果，但一般便于解析和使用的有两个：GFF 格式的定位肽段文件和 PSMs 结果文档。

### 2.5 Peppy

Peppy 是基于 Java 的可单机运行的跨平台全程

表 2 蛋白质基因组学研究分析工具列表
Table 2 List of tools for proteogenomic research

| 软件名称 | 描述 | 程序语言 | 图形用户界面 | 得分系统 | 平台 | 适用对象 | 参考文献 |
|---|---|---|---|---|---|---|---|
| GenoSuite | 可使用多种搜索算法进行质谱数据分析，并将结果整合用于原核生物基因组注释改进的原核生物蛋白质基因组学分析流程 | Perl | Yes | OMSSA X!Tandem InsPecT MassWiz | Windows Linux | 原核生物 | Kumar et al., 2013 |
| iPiG | 将肽谱匹配整合入基因组阅览器实现可视化 | Java | Yes | NO | Windows Linux | 真核生物 原核生物 | Kuhring and Renard, 2012 |
| PepLine | 高通量的直接将串联质谱数据定位回其基因组序列的流程软件 | Java C | Yes | Taggor PSTs | Mac | 真核生物 原核生物 | Ferro et al., 2008 |
| Peppy | 可以方便进行基因组六码框搜索的肽段鉴定软件 | Java | NO | P-value [29] Morpheus [30] | Windows Linux Mac | 真核生物 原核生物 | Risk et al., 2013b |
| PGP | 适用于消息传递接口集群，高通量的批处理集群和多核工作站的并行原核生物蛋白质基因组学流程工具 | Python | NO | InsPecT PepNovo MSGF | Linux | 原核生物 | Tovchigrechko et al., 2014 |
| PMT | 将肽段定位或其来源基因组 | Java | Yes | Aho-Corasick 字符串搜索算法 | Windows Linux Mac | 真核生物 原核生物 | Sanders et al., 2011 |
| VESPA | 整合蛋白质组和转录组数据来改进原核生物基因组注释的工具 | Java | Yes | NO | Windows Linux | 原核生物 | Peterson et al., 2012 |
| PG Nexus | 将下一代测序产生的基因组和转录组数据与蛋白质质谱产生的蛋白质组数据进行整合的软件包 | Java | NO | Mascot | Linux | 真核生物 原核生物 | Pang et al., 2014 |
| ProteoAnnotator | 开源的并且支持蛋白质组学标准计划的标准化数据格式用于肽和蛋白质的鉴定的蛋白质基因组学注释分析流程软件 | Java | Yes | OMSSA X!Tandem MSGF+ MS-Amanda | Linux Windows | 真核生物 原核生物 | Ghali et al., 2014 |
| GAPP | 全自动的用于从串联质谱进行人类肽段可信度鉴定的软件 | NO | NO | X!Tandem Mascot | Linux | 真核生物 | Shadforth et al., 2006 |
| GenoMS | 整合了数据库搜索 *de novo* 肽段鉴定的蛋白质基因组学工具 | NO | NO | InsPecT | Linux | 真核生物 原核生物 | Castellana et al., 2010 |
| Galaxy-P | 基于 Galaxy 框架的灵活易操作的蛋白质基因组需分析流程工具 | Base on Galaxy | Yes | ProteinPilot X!Tandem | Windows Linux | 真核生物 原核生物 | Jagtap et al., 2014 |
| SearchGUI | 整合多个搜索算法的并拥有良好图像化用户参数设置界面的开源的蛋白质组鉴定搜索引擎 | Java | Yes | X!Tandem MS-GF+ MS-Amanda MyriMatch Comet OMSSA | Windows Linux Mac | 真核生物 原核生物 | Vaudel et al., 2011 |
| Neosi | 适于大及复杂基因组真核生物蛋白质基因组学分析套件 | Java Python | NO | MS-GF+ or Others | Windows Linux Mac | 真核生物 原核生物 | Castellana et al., 2014 |
| Bacterial Proteogenomic Pipeline | 可将不同实验条件下鉴定得到的肽段可视化比较的细菌的蛋白质基因组分析流程 | Java | Yes | NO | Windows Linux Mac | 真核生物 | Uszkoreit et al., 2014 |

自动化的蛋白质基因组学分析工具(Risk et al., 2013)。为了解决蛋白质基因组学中大尺寸基因组的计算问题，Peppy 在构建基因组六码框比对数据库的时间，采用的是用户自定义的基因组分段的方法，分段的序列之间直接有 120 bp 的重叠。程序内建有类似 Ascore 的质谱清除方法(Beausoleil et al., 2006)，对质谱图谱中的峰进行过滤，同样可以减轻计算压力。Peppy 的质谱匹配和得分系统有两个选项：自建的基于离子匹配，高于平均值匹配离子丰度和合适的相对 b 和 y 离子丰度可以性 P-values 的(Risk et al., 2013)；另外一个 PSM 得分系统来自 Morpheus (Wenger and Coon, 2013)。Peppy 程序的输入文档两个：DTA 或者 PKL 格式的质谱数据；FASTA 格式的基因组 DNA 或蛋白质序列数据。另外一个参数文档来设定整个流程的所有参数。

### 2.6 Bacterial Proteogenomic Pipeline

Bacterial Proteogenomic Pipeline 是基于 Java 单机运行具有图形化界面的跨平台细菌蛋白质基因组学分析工具。该细菌蛋白质基因组学分析工具供包含六个可以使用命令行或 Java Swing 图形化界面运行的模块：Parse Protein Information 模块将读取一个包含基因组读码框位置信息的 FASTA 格式的蛋白质库和一个包含已注释基因或蛋白质的所有信息的 TSV/CSV 文件，创建一个已知蛋白质的 GFF3 文件；Compare And Combin 可选模块使用另外一个 FASTA 数据库作为参考选项，进一步对 Parse Protein Information 模块创建的 GFF 文档和对应的 FASTA 文档添加信息；Genome Parser 模块依据细菌基因组序列创建六码框蛋白质数据库；Create Decoy DB 可选模块用来创建诱饵数据库；Combine Identifications 模块将外部搜索引擎以 mzTab 文档格式输入，对鉴定的 PSMs 进行验证和 FDR 过滤；Analysis 模块可以对鉴定的肽段进行分析，并可视化每个肽段对应的不同的鉴定的 PSMs 的数目。Bacterial Proteogenomic Pipeline 支持将任何鉴定搜索算法和后处理算法得到的肽段鉴定转换为 mzTab 格式输入，可以对不同实验条件下鉴定得到的肽段可视化和比较分析。并且所有的蛋白质和肽段信息都可以输出到 GFF3 格式文档中，可以利用自身模块实现可视化检验，也便于在常用的基因组阅览器上近一步分析验证。

### 2.7 Genosuite

Genosuite 是基于 Perl 跨平台单机版的全自动的基于四种开源质谱肽段鉴定算法基于质谱蛋白质组数据进行原核生物蛋白质基因组学分析的流水线工具(Kumar et al., 2013)。Genosuite 共包括三个组件：PPT (prokaryotic proteogenomic tool), ORFmapper 和 PSMplotter，具体流程如图 1 所示。在 PPT 中使用 OMSSA，X!Tandem，InsPecT 和 MassWiz 四种肽段搜索鉴定算法或任意一种组合来对基因组六码框翻译的数据库进行搜索，不同算法的组合使用提高了蛋白质组搜索时间的覆盖度，基于组合的 FDR 得分(Combined FDRScore) (Jones et al., 2009)来对不同算法的结果进行整合过滤。程序自动将过滤后的肽段定位回基因组和已知蛋白，那些仅仅定位到基因组翻译数据库的肽段并归类为新肽，并可以 GFF 的格式方便分布式注释服务器(distributed annotation system, DAS)使用。ORFmapper 使用 genbank 文档，GFF 格式或者 GeneMark 格式的 ORF 预测文档和新肽的 GFF 文档作为输入，用来将新肽和已存在的注释和 ab initio 注释进行对比，进一步将新肽分类为新蛋白质编码区(novel proteins coding region, NPCR)或者是基因模式改变。最终 ORFmapper 就可以分别输出产生新蛋白质的肽段文件，产生基因模式变化的肽段文件和 ORFs 定位到新肽的文档。ORFmapper 还创建了每个肽段在基因组的基因组图谱文档，这就提供了基因组范围的肽段的可视化，HTML 的文件格式也便于分析。PSMplotter 程序是一个肽段质谱匹配的可视化应用，其将 PPT 的 XML 文档作为输入，生产 HTML 文档。在 HTML 中所有的来自 XML 文档的质谱匹配都和其 PSM 图片超链接，这样就可以便于对 PSMs 的人工验证。

### 2.8 ProteoAnnotator

ProteoAnnotator 是基于 Java 的全自动的将质谱的蛋白质组学证据整合入基因组数据库的软件流程(Kucharova and Wiker, 2014)，其具体分析流程图如图 2 所示。ProteoAnnotator 既为终端用户如实验室科学家提供了图形化的界面，也为希望在并行环境下运行该程序的信息分析人员提供了命令行模式的分析环境。ProteoAnnotator 在每个分析模块中都使用了蛋白质组学标准计划(proteomics standards initiative, PSI)规定的 mzIdentML 标准化数据格式用于肽和蛋白质的鉴定。mzIdentML 的使用使得 ProteoAnnotator 单个模块可以和其他分析工具整合，其输出结果可以直接提交到 ProteomeXchange 中心数据库(Vizcaino et al., 2014)和 PRIDE (Vizcaino et al., 2013)。ProteoAnnotator 使用 GFF3 和 FASTA 格式的文档作为数据库输入文档。

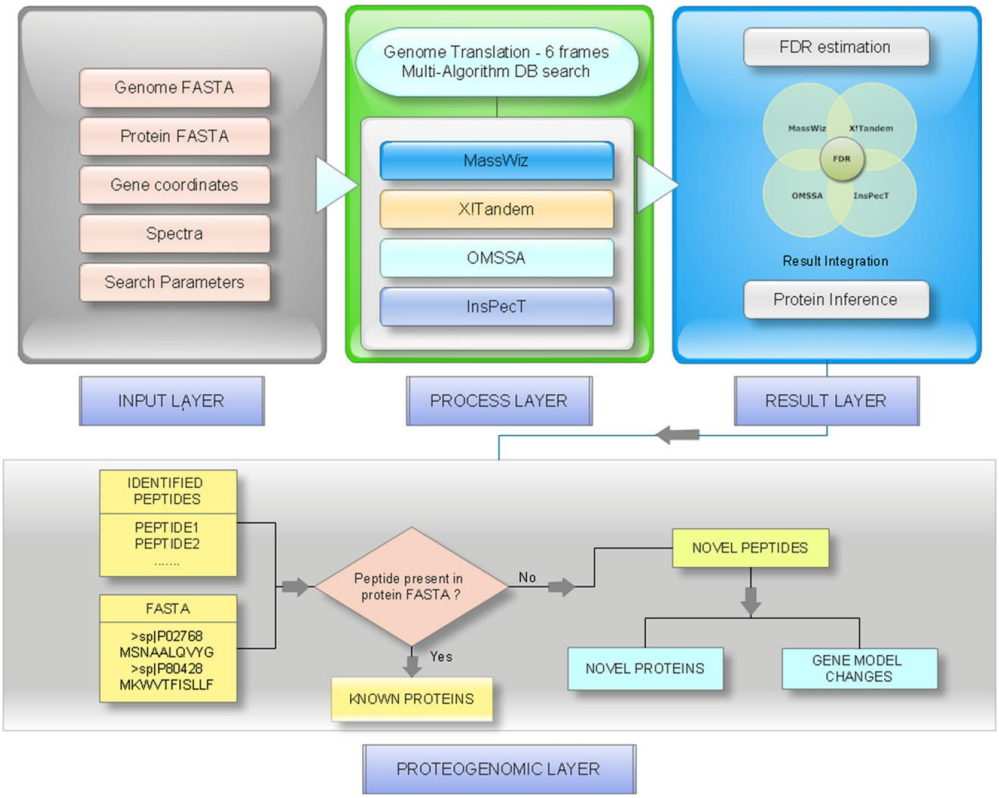


图 1 GenoSuite 流程图
Figure 1 Schematic representation of GenoSuite workflow

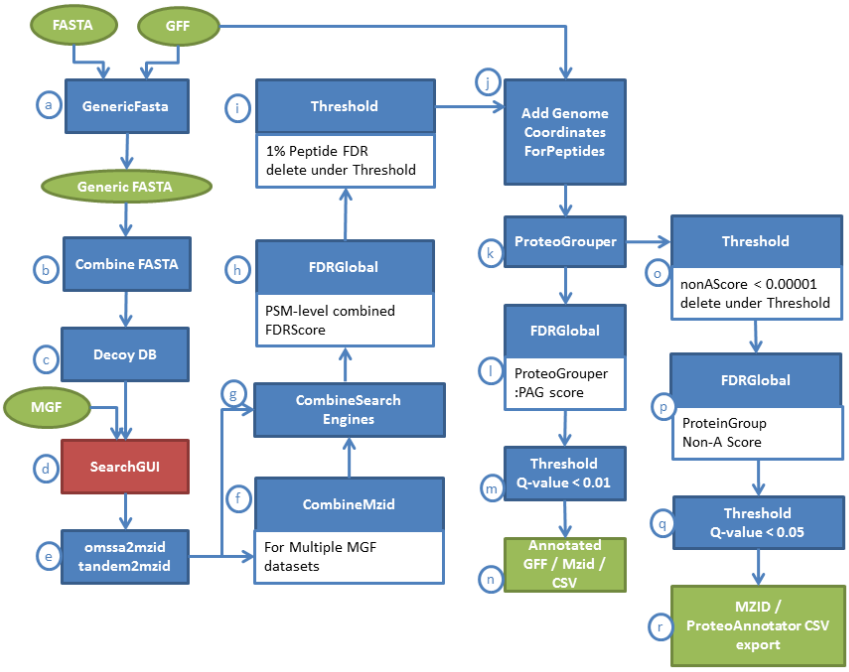


图 2 ProteoAnnotator 流程图
注: MzidLib 标识: 蓝长方形; 输入文档标识: 绿椭圆形; 输出表示: 绿色长方形; SearchGUI 标识: 红色长方形
Figure 2 The ProteoAnnotator workflow
Note: MzidLib: Blue rectangles; With file inputs: Green ovals; Outputs: Green rectangles; SearchGUI integration: Red rectangle

如果 GFF 文档已经包含 FASTA 格式数据的情况下，FASTA 格式的蛋白质序列文档就是可选择项。ProteoAnnotator 需要用户上传一个套基因组坐标(genomic coordinates)和蛋白质序列作为物种基因组的正式注释模式，并被标记为“A”套基因/蛋白质以用作进一步的分析。在某些没有正式基因组模式的测序基因组，用户应当上传一套最好质量的比如由基因预测软件预测过的基因模式，接着用户可自行决定是否按照预测基因模式质量顺序为参考上传其他的并依次标注为“B”，“C”，“D”等等。ProteoAnnotator 在创建诱饵数据用于 SearchGUI 和预处理与后处理过程中，及在使用 Omssa 和 X!Tandem 进行 MS/MS 搜索中都使用了 MzidLib (Ghali et al., 2013)。SearchGUI 作为一个开放源工具可以在一个质谱搜索中使用不同的开放源代码的搜索引擎(Omssa, X!Tandem, MSGF+和 MS-Amanda) (Vaudel et al., 2011)。ProteoAnnotaotr 下游分为两个途径，因此可以提供给用户两套不同的输出文档类型：一类是提供源自正式的基因模式指定的蛋白质和肽和/或者被鉴定的不同基因模式的证据；另外一类是提供正式基因组注释在高度确信鉴定的位点上有改善提高空间的证据。目前 ProteoAnnotator 目前整合入 Proteosuite (http://www.proteosuite.org)可视化运行。

### 2.9 Galaxy-P

Galaxy-P 是基于 Galaxy 项目(http://galaxyproject.org)的多“组学”数据特别关注于质谱为基础的蛋白质组学的分析平台，在 Galaxy-P 的基础上，研究人员开发出了一套高度灵活和宜使用的蛋白质基因组学分析流程(Jagtap et al., 2014)。基于 Galaxy-P 的蛋白质基因组学分析流程共包含大约 140 个处理步骤，可以归类为四个模块中，其具体分析流程图如图 3 所示：1、质谱数据 Peaklist 文件的生产和来自组装的 DNA 或 RNA 序列来源的蛋白质序列数据库的生成；2、序列数据库的搜索；3、数据过滤和可信度的分配；4、新蛋白质产物在基因组中的可视化及阐释。

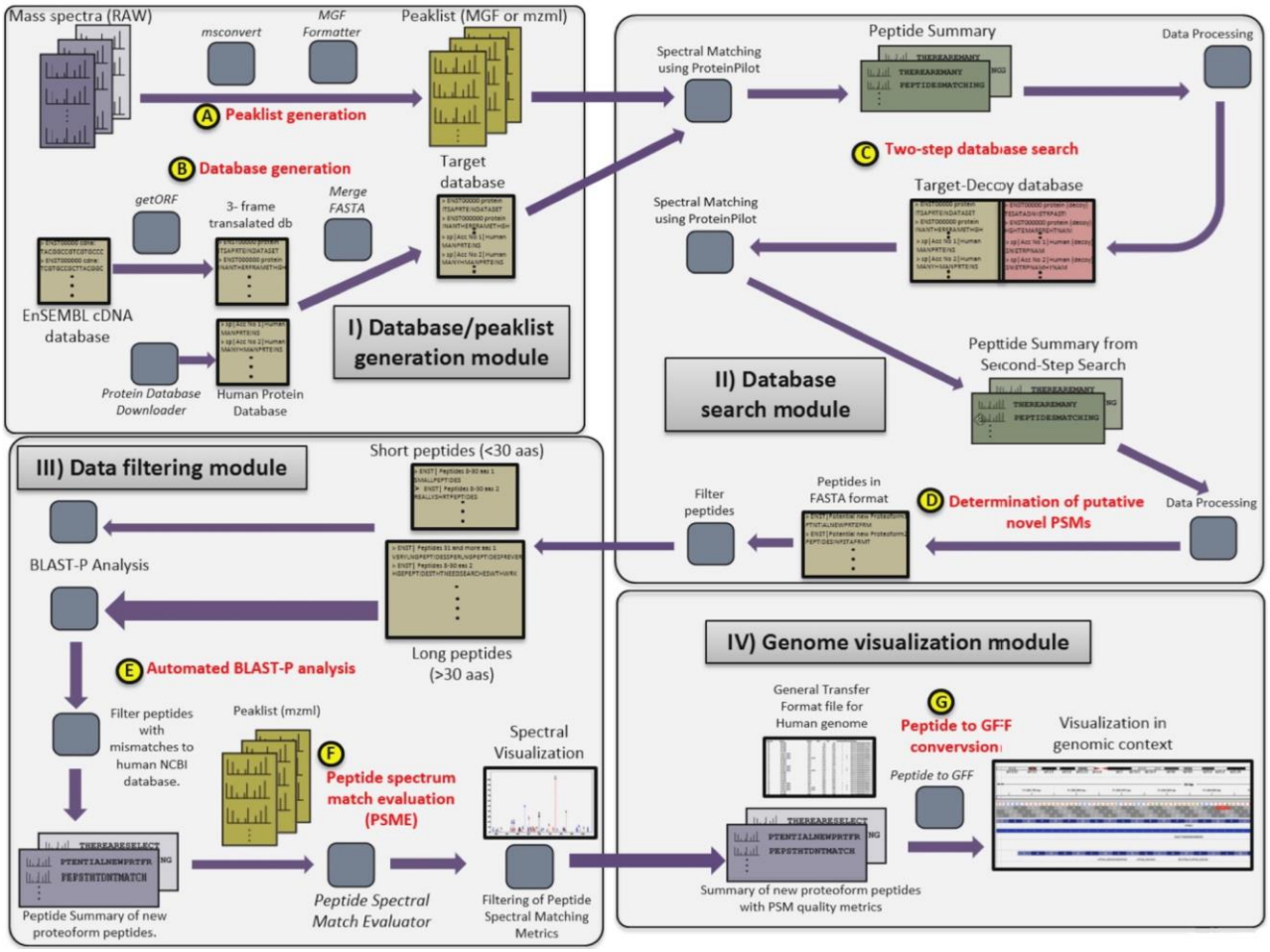


图 3 Galaxy-P 分析流程图
Figure 3 Overview of modules and subworkflows comprising the Galaxy-based proteogenomic analysis workflow

Galaxy-P 分析的灵活性的体现之一就是数据库构建的灵活性，使得 Galaxy-P 可以针对不同的样品进行不同的数据库改变。这种灵活性带来便利和针对性在研究者使用该分析流程对人类唾液的蛋白质基因组学的演示分析中得到体现。研究人员发现在 PSMs 搜索分析中使用人类蛋白质数据库加入人类口腔微生物的蛋白质数据库组成综合数据库，比单独使用人类蛋白质数据库要多发现两倍的新序列变异体。导致这一问题出现的原因就是：微生物序列的缺失会导致串联质谱得到的来自非宿主的肽被迫和宿主的蛋白质相匹配，增加了假阳性并迫使 PSMs 为得到可接受的 FDR 必须有一个更高的得分值，因此就降低了新肽序列可信匹配的数量。这一发现说明在进行蛋白质基因组学的研究中，在样品包含非宿主的蛋白质情况下进行的蛋白质基因组学分析，如果仅仅使用宿主的蛋白质序列进行数据库搜索，其他结果会受到很大影响。

Galaxy-P 灵活性的另外一个表现就是搜索时可以使用不同的质谱数据库搜索引擎，这样可以相互验证并弥补不足。在演示分析研究中 ProteinPilot 和 X!Tandem 的使用就充分证明了这一点，另外 SearchGUI (Vaudel et al., 2011)未来整合入 Galaxy-P 的蛋白质基因组学分析流程就更值得期待。序列数据库的搜索中，“明尼苏达两步法”的使用也可以解决蛋白质基因组分析中大蛋白质数据库所固有的挑战(Jagtap et al., 2013)。在两步法中，第一步的数据库搜索使用比较宽松的严谨度来鉴定那些最可能在样品中存在的蛋白质，组成一个比较小的蛋白质库；在第二步中对第一步产生的修正的更小蛋白质数据库和与其相匹配的质谱被施加了高度严密的条件进行二次分析，在可接受的 FDR 水平产生 PSMs。

蛋白质基因组学分析往往依靠单个的 PSMs 来确定潜在的新蛋白序列，考虑到单个 PSMs 带来的潜在的假阳性，Galaxy-P 的蛋白质基因组学分析流程中提供了多中水平的质量控制和过滤。除了在数据库搜索模块使用多个数据搜索引擎来改善结果的可信度外，在关键的第三模块中 BLASTP 方法的使用和自行开发的 PSME (peptide spectrum match evaluation)工具是 PSMs 质量控制的重要一环。BLASTP 中和 NCBI 数据库的比对可以进一步过滤掉和已知序列匹配的 PSMs。PSME 不仅提供了一种可视化串联质谱图谱及其对应假定序列匹配的工具，还可以用户自行设定多种 PSM 质量相关的参数来对质谱进行过滤。在 Galaxy-P 演示数据的分析中 BLASTP 的分析将 9333 个 PSMs 减少点 1630 个，PSME 高严谨度 PSM 质量标准的限定更是将新肽序列匹配的数量减少到 55 个。

Galaxy-P 的蛋白质基因组学分析流程的最后一步是通过自开发的“Peptides to GFF”工具来将肽段的氨基酸序列转换为 IGV (integrated genome viewer)兼容的格式，从而实现在基因组中的可视化和阐释。新肽片段在基因组上的可视化有利于进一步的对这些肽段所对应假定新蛋白质可能性的评估分析。

虽然，基于 Galaxy-P 的蛋白质基因组学分析流程尽管包含大约 140 多个步骤，但整个流程在参数优化设定的情况下，只需要一次单击就可以完成整个流程，最要的是每个模块中的分流程都可以按需求单独自我运行。这些整个流程可以通过一个网络连接或一个保存的 Galaxy 流程文档而与其他人分享，和流程文档一样 Galaxy-P 的历史文档也可以被分享，历史文档中包含了重现整个分析流程所需要的所有的软件和各种输入与输出数据。

### 2.10 Proteomic-Genomic Nexus

Proteomic-Genomic Nexus 是基于 Java 语言软件包，其设计目的是将下一代测序产生的基因组和转录组数据和蛋白质质谱产生的蛋白质组数据进行整合，该软件包的具体分析流程可见图 4。

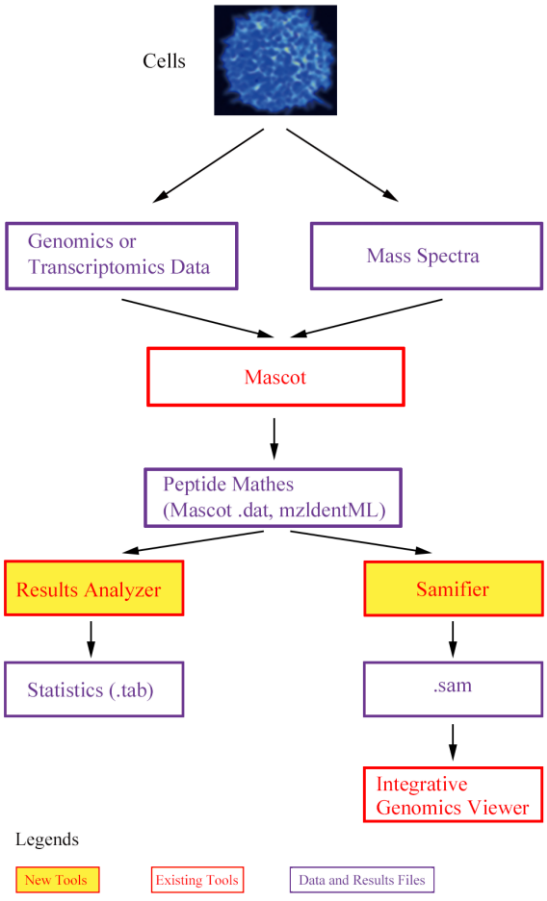


图 4 The Proteomic-Genomic Nexus 分析流程图示
Figure 4 The Proteomic-Genomic Nexus analysis workflow

PG Nexus 软件包含有两个主要的自主设计的工具：Samifier 和 Results analyser。Samifier 可以将蛋白质质谱转换为 SAM 格式，这就可以在整合基因组学查看器(integrative genomics viewer, IGV)中同时查看基因组学，转录组学和蛋白质组学的数据。Results analyser 报告肽段和蛋白质的数目和类型，并可以报告他们基于自设地订的过滤条件所对应的 Mascot 得分，跨越外显子直接连接的也被高亮显示，这可被用于验证蛋白质的不同剪切变异体。在分析原核生物的基因组时，PG Nexus 多增加了 Virtual protein generator 和 Virtual protein merger 两个工具：Virtual protein generator 用来产生基于 Glimmer 基因预测的 Mascot 序列数据；Virtual protein merger 则是通过搜索起始密码子和终止密码子的两侧，重新计算那些匹配到虚拟蛋白质的肽的 PG 开放阅读框的位置。Nexus 软件可以整合入 Galaxy 项目，这就极大的增强了其使用的方便性。

**2.11 Neosi**

Neosi 是基于 Java 和 Python 语言编程的尤其适用于分析大及复杂基因组真核生物的自动化蛋白质基因组学分析套件(Castellana et al., 2014)。Neosi最早开发的版本是使用EST 和其他的转录数据(非 RNA-seq)来创建自定义数据库，使用 InsPect 来直接搜索可变剪切图模式库(splice graphs)，用来注释模式生物拟南芥和玉米(Castellana et al., 2010)。后期改进的 Neosi 包含两个套件，第一个套件 SpliceDB 工具主要用来创建特异的数据库用来发掘变化的基因事件(gene events)，自定制构建的数据库可以使用任意的质谱搜索引擎来搜索，但程序自身整合并推荐使用 MS-GF+，MS-GF+使用一个组合的方法来为谱肽匹配(peptide spectra matches, PSMs）统计得分和赋予显著性(Kim and Pevzner, 2014)。最后 Enosi 的第二个套件用来分析鉴定的肽段。鉴定的肽段序列自身对新肽并不具有太多信息，有基于此 Enosi 将这些肽归类为已知和新的两种，并在基因组中为新肽寻找定位。这新肽的定位将和已知基因的定位相比较，并进行事件的归类，这样新肽更加可以直观的易识别。Enosi 也包含自己的方法用来过滤一些不可信的事件。这样 Enosi 就可以自动的使用所有的质谱数据搜索自定制数据库，积累所有结果并使用错误发现率(FDR)的计算方法来鉴别 PSMs，归类新肽并自动对新注释事件做出提示。这些注释事件包含了：可变剪切，新剪切，融合基因，插入，缺失，突变，翻译的非翻译区(untranslated regions, UTR)，基因边界，外显子边界，新外显子，读码框偏移，反向链，新基因。Neosi 并不仅适用于真核生物，在经过特殊的流程定制后，Neosi 可以适用于原核生物，并且在敏感性和特异性方面比 GenoSuite 更加有优势(Chapman and Bellgard, 2014)。从最早开发的相关的搜索及与数据库构建相关的算法，到将鉴定的 PSMs 如何在基因组上可视化，再到后期整个蛋白质基因组学分析流程的系统化整合工具的开发，蛋白质基因组学的实用工具的开发随着其自身概念及应用的发展而不断发展，详细情况见表 1。虽然目前已经存在多套完整的自动化的蛋白质基因组学分析流程，然而这些流程套件还是缺少统一的标准，这些流程直接注释的准确性还缺少直接全面的分析准确性方面的比较研究，因此，在进行蛋白质基因组学分析研究中如何选取合适的分析流程还是一个难题。在今后的蛋白质基因组学的发展中，确立统一的标准化规范是一个长期的目标，也是其持续发展不可或缺的一环。把相对固定的蛋白质基因组学分析流程分解模块化，提供可以相互可以衔接的标准化接口，是今后蛋白质基因组学分析工具发展的一个方向。

## 作者贡献

巩鹏涛负责论文的构思，文献调研，初稿撰写及修改；徐润生负责文献阅读、整理和确认，并对论文提出修改意见；方宣钧博士负责论文写作框架的确定、全文系统修改以及最后的定稿。

## 致谢

本研究由海南省热带农业资源研究所微生物基因组测序及生物信息学研究项目资助。

## 参考文献

Ahmed F. E. 2008, Utility of mass spectrometry for proteome analysis: part I. Conceptual and experimental approaches, Expert Rev Proteomics, 5 (6): 841-864
http://dx.doi.org/10.1586/14789450.5.6.841

Ahmed F. E. 2009, Utility of mass spectrometry for proteome analysis: part II. Ion-activation methods, statistics, bioinformatics and annotation, Expert Rev Proteomics, 6 (2): 171-197
http://dx.doi.org/10.1586/epr.09.4

Allmer J., Markert C., Stauber E. J., and Hippler M. 2004, A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases, FEBS Lett, 562 (1-3): 202-206
http://dx.doi.org/10.1016/S0014-5793(04)00212-1

Beausoleil S. A., Villen J., Gerber S. A., Rush J., and Gygi S. P. 2006, A probability-based approach for high-throughput protein phosphorylation analysis and site localization, Nat Biotechnol, 24 (10): 1285-1292
http://dx.doi.org/10.1038/nbt1240

Bern M., Cai Y., and Goldberg D. 2007, Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry, Anal Chem, 79 (4): 1393-1400
http://dx.doi.org/10.1021/ac0617013

Castellana N. E., Pham V., Arnott D., Lill J. R., and Bafna V. 2010, Template proteogenomics: sequencing whole proteins using an imperfect database, Mol Cell Proteomics, 9 (6): 1260-1270
http://dx.doi.org/10.1074/mcp.M900504-MCP200

Castellana N. E., Payne S. H., Shen Z., Stanke M., Bafna V., and Briggs S. P. 2008, Discovery and revision of Arabidopsis genes by proteogenomics, Proc Natl Acad Sci U S A, 105 (52): 21034-21038
http://dx.doi.org/10.1073/pnas.0811066106

Castellana N. E., Shen Z., He Y., Walley J. W., Cassidy C. J., Briggs S.P., and Bafna V., 2014, An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays, Mol Cell Proteomics, 13 (1): 157-167
http://dx.doi.org/10.1074/mcp.M113.031260

Chapman Brett, and Bellgard Matthew. 2014, High-throughput parallel proteogenomics: A bacterial case study, PROTEOMICS, 14 (23-24): 2780-2789
http://dx.doi.org/10.1002/pmic.201400185

Ferro M., Tardif M., Reguer E., Cahuzac R., Bruley C., Vermat T., Nugues E., Vigouroux M., Vandenbrouck Y., Garin J., and Viari A. 2008, PepLine: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences, J Proteome Res, 7 (5): 1873-1883
http://dx.doi.org/10.1021/pr070415k

Ghali F., Krishna R., Perkins S., Collins A., Xia D., Wastling J., and Jones A. R. 2014, ProteoAnnotator - Open Source Proteogenomics Annotation Software Supporting PSI Standards, Proteomics, 14 (23-24): 2731-2741
http://dx.doi.org/10.1002/pmic.201400265

Ghali F., Krishna R., Lukasse P., Martinez-Bartolome S., Reisinger F., Hermjakob H., Vizcaino J. A., and Jones A. R. 2013, Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML, Mol Cell Proteomics, 12 (11): 3026-3035
http://dx.doi.org/10.1074/mcp.O113.029777

Gong P.T., Xu R.S. Xu and X.J. Fang, Proteogenomics: Progress, Strategy and Problem Genomics and Applied Biology, 2014, Vol.33, No.6, 1169-1180

Jaffe J. D., Berg H. C., and Church G. M. 2004, Proteogenomic mapping as a complementary method to perform genome annotation, Proteomics, 4 (1): 59-77
http://dx.doi.org/10.1002/pmic.200300511

Jagtap P., Goslinga J., Kooren J. A., McGowan T., Wroblewski M. S., Seymour S. L., and Griffin T. J. 2013, A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies, Proteomics, 13 (8): 1352-1357
http://dx.doi.org/10.1002/pmic.201200352

Jagtap Pratik Dilip, Johnson James E., Onsongo Getiria, Sadler Fredrik W., Murray Kevin, Wang Yuanbo, Sheynkman Gloria M., Bandhakavi Sricharan, Smith Lloyd M., and Griffin Timothy J., 2014, Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework, Journal of Proteome Research, 13 (12): 5898-5908
http://dx.doi.org/10.1021/pr500812t

Jones A. R., Siepen J. A., Hubbard S. J., and Paton N. W. 2009, Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines, Proteomics, 9 (5): 1220-1229
http://dx.doi.org/10.1002/pmic.200800473

Kim Min-Sik, Pinto Sneha M., Getnet Derese, Nirujogi Raja Sekhar, Manda Srikanth S., Chaerkady Raghothama, Madugundu Anil K., Kelkar Dhanashree S., Isserlin Ruth, Jain Shobhit, Thomas Joji K., Muthusamy Babylakshmi, Leal-Rojas Pamela, Kumar Praveen, Sahasrabuddhe Nandini A., Balakrishnan Lavanya, Advani Jayshree, George Bijesh, Renuse Santosh, Selvan Lakshmi Dhevi N., Patil Arun H., Nanjappa Vishalakshi, Radhakrishnan Aneesha, Prasad Samarjeet, Subbannayya Tejaswini, Raju Rajesh, Kumar Manish, Sreenivasamurthy Sreelakshmi K., Marimuthu Arivusudar, Sathe Gajanan J., Chavan Sandip, Datta Keshava K., Subbannayya Yashwanth, Sahu Apeksha, Yelamanchi Soujanya D., Jayaram Savita, Rajagopalan Pavithra, Sharma Jyoti, Murthy Krishna R., Syed Nazia, Goel Renu, Khan Aafaque A., Ahmad Sartaj, Dey Gourav, Mudgal Keshav, Chatterjee Aditi, Huang Tai-Chung, Zhong Jun, Wu Xinyan, Shaw Patrick G., Freed Donald, Zahari Muhammad S., Mukherjee Kanchan K., Shankar Subramanian, Mahadevan Anita, Lam Henry, Mitchell Christopher J., Shankar Susarla Krishna, Satishchandra Parthasarathy, Schroeder John T., Sirdeshmukh Ravi, Maitra Anirban, Leach Steven D., Drake Charles G., Halushka Marc K., Prasad T. S. Keshava, Hruban Ralph H., Kerr Candace L., Bader Gary D., Iacobuzio-Donahue Christine A., Gowda Harsha, and Pandey Akhilesh. 2014, A draft map of the human proteome, Nature, 509 (7502): 575-581
http://dx.doi.org/10.1038/nature13302

Kim S., and Pevzner P. A. 2014, MS-GF+ makes progress towards a universal database search tool for proteomics, Nat Commun, 5: 5277
http://dx.doi.org/10.1038/ncomms6277

Krug Karsten, Nahnsen Sven, and Macek Boris. 2011, Mass spectrometry at the interface of proteomics and genomics, Molecular BioSystems, 7 (2): 284-291
http://dx.doi.org/10.1039/c0mb00168f

Kucharova V., and Wiker H. G. 2014, Proteogenomics in microbiology: Taking the right turn at the junction of genomics and proteomics, Proteomics: Kuhring M., and Renard B. Y. 2012, iPiG: integrating peptide spectrum matches into genome browser visualizations, PLoS One, 7 (12): e50246

Kumar D., Yadav A. K., Kadimi P. K., Nagaraj S. H., Grimmond S. M., and Dash D. 2013, Proteogenomic analysis of Bradyrhizobium japonicum USDA110 using GenoSuite, an automated multi-algorithmic pipeline, Mol Cell Proteomics, 12 (11): 3388-3397
http://dx.doi.org/10.1074/mcp.M112.027169

Nesvizhskii A. I. 2010, A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics, J Proteomics, 73 (11): 2092-2123
http://dx.doi.org/10.1016/j.jprot.2010.08.009

Olsen J.V., Schwartz J.C., Griep-Raming J., Nielsen M. L., Damoc E., Denisov E., Lange O., Remes P., Taylor D., Splendore M., Wouters E. R., Senko M., Makarov A., Mann M., and Horning S. 2009, A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed, Mol Cell Proteomics, 8 (12): 2759-2769
http://dx.doi.org/10.1074/mcp.M900375-MCP200

Pang C. N., Tay A. P., Aya C., Twine N. A., Harkness L., Hart-Smith G., Chia S. Z., Chen Z., Deshpande N. P., Kaakoush N. O., Mitchell H. M., Kassem M., and Wilkins M. R. 2014, Tools to covisualize and coanalyze proteomic

 

data with genomes and transcriptomes: validation of genes and alternative mRNA splicing, J Proteome Res, 13 (1): 84-98
http://dx.doi.org/10.1021/pr400820p

Peterson E. S., McCue L. A., Schrimpe-Rutledge A. C., Jensen J. L., Walker H., Kobold M. A., Webb S. R., Payne S. H., Ansong C., Adkins J. N., Cannon W. R., and Webb-Robertson B. J. 2012, VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data, BMC Genomics, 13: 131
http://dx.doi.org/10.1186/1471-2164-13-131

Renuse S., Chaerkady R., and Pandey A. 2011, Proteogenomics, Proteomics, 11 (4): 620-630
http://dx.doi.org/10.1002/pmic.201000615

Risk B. A., Edwards N. J., and Giddings M. C. 2013a, A peptide-spectrum scoring system based on ion alignment, intensity, and pair probabilities, J Proteome Res, 12 (9): 4240-4247
http://dx.doi.org/10.1021/pr400286p

Risk B. A., Spitzer W. J., and Giddings M. C. 2013b, Peppy: proteogenomic search software, J Proteome Res, 12 (6): 3019-3025
http://dx.doi.org/10.1021/pr400208w

Roos F. F., Jacob R., Grossmann J., Fischer B., Buhmann J. M., Gruissem W., Baginsky S., and Widmayer P., 2007, PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra, Bioinformatics, 23 (22): 3016-3023
http://dx.doi.org/10.1093/bioinformatics/btm417

Sanders W. S., Wang N., Bridges S. M., Malone B. M., Dandass Y. S., McCarthy F. M., Nanduri B., Lawrence M. L., and Burgess S. C. 2011, The proteogenomic mapping tool, BMC Bioinformatics, 12: 115
http://dx.doi.org/10.1186/1471-2105-12-115

Sevinsky J. R., Cargile B. J., Bunger M. K., Meng F., Yates N. A., Hendrickson R. C., and Stephenson J. L., Jr. 2008, Whole genome searching with shotgun proteomic data: applications for genome annotation, J Proteome Res, 7 (1): 80-88
http://dx.doi.org/10.1021/pr070198n

Shadforth I., Xu W., Crowther D., and Bessant C. 2006, GAPP: a fully automated software for the confident identification of human peptides from tandem mass spectra, J Proteome Res, 5 (10): 2849-2852
http://dx.doi.org/10.1021/pr060205s

Syka J. E., Coon J. J., Schroeder M. J., Shabanowitz J., and Hunt D. F. 2004, Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry, Proc Natl Acad Sci U S A, 101 (26): 9528-9533
http://dx.doi.org/10.1073/pnas.0402700101

Tanner S., Shu H., Frank A., Wang L. C., Zandi E., Mumby M., Pevzner P. A., and Bafna V. 2005, InsPecT: identification of posttranslationally modified peptides from tandem mass spectra, Anal Chem, 77 (14): 4626-4639
http://dx.doi.org/10.1021/ac050102d

Thelen J. J., and Miernyk J. A. 2012, The proteomic future: where mass spectrometry should be taking us, Biochem J, 444 (2): 169-181
http://dx.doi.org/10.1042/BJ20110363

Tovchigrechko A., Venepally P., and Payne S. H. 2014, PGP: parallel prokaryotic proteogenomics pipeline for MPI clusters, high-throughput batch clusters and multicore workstations, Bioinformatics, 30 (10): 1469-1470
http://dx.doi.org/10.1093/bioinformatics/btu051

Uszkoreit Julian, Plohnke Nicole, Rexroth Sascha, Marcus Katrin, and Eisenacher Martin. 2014, The bacterial proteogenomic pipeline, BMC Genomics, 15 (Suppl 9): S19
http://dx.doi.org/10.1186/1471-2164-15-S9-S19

Vaudel M., Barsnes H., Berven F. S., Sickmann A., and Martens L. 2011, SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches, Proteomics, 11 (5): 996-999
http://dx.doi.org/10.1002/pmic.201000595

Vizcaino J. A., Cote R. G., Csordas A., Dianes J. A., Fabregat A., Foster J. M., Griss J., Alpi E., Birim M., Contell J., O'Kelly G., Schoenegger A., Ovelleiro D., Perez-Riverol Y., Reisinger F., Rios D., Wang R., and Hermjakob H. 2013, The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013, Nucleic Acids Res, 41 (Database issue): D1063-1069
http://dx.doi.org/10.1093/nar/gks1262

Vizcaino J. A., Deutsch E. W., Wang R., Csordas A., Reisinger F., Rios D., Dianes J. A., Sun Z., Farrah T., Bandeira N., Binz P. A., Xenarios I., Eisenacher M., Mayer G., Gatto L., Campos A., Chalkley R. J., Kraus H. J., Albar J. P., Martinez-Bartolome S., Apweiler R., Omenn G. S., Martens L., Jones A. R., and Hermjakob H. 2014, ProteomeXchange provides globally coordinated proteomics data submission and dissemination, Nat Biotechnol, 32 (3): 223-226
http://dx.doi.org/10.1038/nbt.2839

Wenger C. D., and Coon J. J. 2013, A proteomics search algorithm specifically designed for high-resolution tandem mass spectra, J Proteome Res, 12 (3): 1377-1386
http://dx.doi.org/10.1021/pr301024c

Wenger C. D., McAlister G. C., Xia Q., and Coon J. J. 2010, Sub-part-per-million precursor and product mass accuracy for high-throughput proteomics on an electron transfer dissociation-enabled orbitrap mass spectrometer, Mol Cell Proteomics, 9 (5): 754-763
http://dx.doi.org/10.1074/mcp.M900541-MCP200

Xu Q., and Lee C. 2003, Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences, Nucleic Acids Res, 31 (19): 5635-5643
http://dx.doi.org/10.1093/nar/gkg786

Zhang B., Wang J., Wang X., Zhu J., Liu Q., Shi Z., Chambers M. C., Zimmerman L. J., Shaddox K. F., Kim S., Davies S. R., Wang S., Wang P., Kinsinger C. R., Rivers R. C., Rodriguez H., Townsend R. R., Ellis M. J., Carr S. A., Tabb D. L., Coffey R. J., Slebos R. J., Liebler D. C., the Nci Cptac, and National Cancer Institute Clinical Proteomics Tumor Analysis Consortium. 2014, Proteogenomic characterization of human colon and rectal cancer, Nature, 513 (7518): 382-387
http://dx.doi.org/10.1038/nature13438

Zhou C., Chi H., Wang L. H., Li Y., Wu Y. J., Fu Y., Sun R. X., and He S. M. 2010, Speeding up tandem mass spectrometry-based database searching by longest common prefix, BMC Bioinformatics, 11: 577
http://dx.doi.org/10.1186/1471-2105-11-577