基于质谱技术的计算蛋白质组学研究*

孙瑞祥[1**] 付 岩[1,2] 李德泉[1,2] 张京芬[1,2] 王晓彪[1,2] 盛泉虎[3]
曾 嵘[3] 陈益强[1] 贺思敏[1] 高 文[1,2]

(1. 中国科学院计算技术研究所, 北京 100080; 2. 中国科学院研究生院, 北京 100039; 3.中国科学院上海生命科学研究院生物化学与细胞生物学研究所, 上海 200031)

摘要 蛋白质组是继人类基因组计划完成之后又一新兴的生命科学研究对象, 蛋白质组学研究细胞或组织内所有表达的蛋白质. 生物质谱技术已为蛋白质组学研究产生了大规模的质谱数据; 而如何从这些数据中提取和发现有关蛋白质组的重要生物学知识为计算蛋白质组学的研究提出了重大需求, 如蛋白质鉴定、翻译后修饰、定量分析, 以及疾病模式的发现等. 本文研究了如何应用计算技术来解决蛋白质组学研究中质谱信息处理的这几个关键问题.

关键词 计算蛋白质组学 质谱技术 蛋白质鉴定 算法 生物信息学

自 20 世纪 90 年代人类基因组计划(human genome project, HGP)的正式实施以来, 人们对基因组序列信息的计算分析方法研究成为生物信息学最集中的研究内容之一; 而随着 HGP 于 2003 年被宣布完成之后, 对蛋白质组的全面研究将逐渐成为 21 世纪前期的另一项重要任务.

蛋白质组学(proteomics)是研究细胞或组织内所有表达的蛋白质的一门新兴学科; 计算蛋白质组学(computational proteomics)则是研究如何应用计算技术来解决蛋白质组学中关键的生物学问题, 如蛋白质鉴定、结构预测、功能分类、亚细胞定位、翻译后修饰分析、相互作用网络、定量分析、疾病诊断与药物设计等, 它已成为计算生物学, 或生物信息学的一个主要分支[1]. 特别是最近几年, 蛋白质组学的逐渐兴起, 以及计算机硬件、信息处理技术、网络技术的快速发展为计算蛋白质组学研究的广泛开展提供了成熟的条件, 计算蛋白质组学在当代生命科学的研究中正发挥着越来越重要的作用. 计算蛋白质组学是采用计算的手段

收稿日期: 2005-09-28; 接受日期: 2005-11-17
* 国家“973”计划(2002CB713807)和国家科技攻关计划(2004BA711A21)资助项目
** E-mail: rxsun@ict.ac.cn

来解决蛋白质组学中的生物学问题, 因而需要对要解决的生物学问题的背景和相关的计算技术皆有充分深入的认识, 这体现了至少计算与生物两个发展最活跃学科的交叉与融合 [1)]. 本文的研究对象锁定在计算蛋白质组学中的质谱数据.

利用蛋白质的质谱数据可以实现蛋白质的身份鉴定、翻译后修饰分析、寻找生物标记物与疾病的早期诊断等应用, 在蛋白质组学中, 以质谱技术(mass spectrometry, MS)为核心的研究工作最近几年逐渐受到重视, 特别是获得 2002 年诺贝尔化学奖的MALDI和ESI软电离技术的发明使得生物质谱的发展与推广使用尤为迅速. 基于质谱技术的计算蛋白质组学主要研究如何分析和利用各种类型的质谱数据来发现与蛋白质相关的生物学知识, 包括蛋白质的身份鉴定、氨基酸序列信息分析、翻译后修饰分析、定量化信息提取、生物标记物发现与疾病诊断建模等过程中所涉及到的统计分析、数据挖掘、数据前后处理算法等与计算相关的问题[2~5].

本文首先在第 1 部分简介质谱数据的相关背景; 然后在第 2 和 3 部分分别介绍两个研究成果——利用串联质谱鉴定蛋白质的新算法 KSDP 和软件系统 pFind, 以及利用串联质谱中的同位素模式预测离子分子式的新方法; 第 4 部分介绍基于质谱技术的计算蛋白质组学的两个最新发展方向——“定量蛋白质组学”和“疾病蛋白质组学”的研究内容与动向; 第 5 部分是对全文的总结和对学科发展趋势的展望.

1　质谱数据的相关背景

分子生物学知识告诉我们 DNA 是生命遗传信息的载体; 蛋白质则是生命功能的执行者, 其基本组成单位是 20 种氨基酸, 氨基酸经化学键首尾相接成线性长链构成蛋白质. 蛋白质的一级结构(primary structure)指的就是构成蛋白质的氨基酸长链的序列信息, 如图 1 所示, 每个字母代表一种氨基酸.

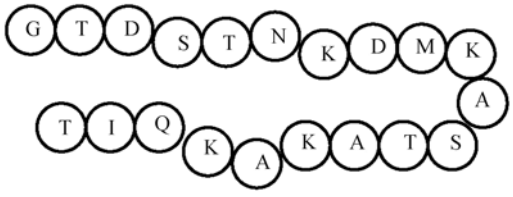

图 1　蛋白质的一级结构示意图

蛋白质的一级结构信息对于结构预测和认识蛋白质的功能, 同源性检测和分析进化关系等都具有重要的作用, 因此, 通过生化实验和计算分析获取蛋白质的序列信息是关键性的基础问题. 传统的 Edman 降解技术获得氨基酸序列信息

1) 孙瑞祥. 信息处理技术在蛋白质鉴定中的应用研究. 中国科学院研究生院博士后出站报告, 2004 年 7 月

只能逐个获取，效率太低，因而不能适应当今高通量实验的需求；质谱技术为快速准确地获取蛋白质的一级结构信息提供了新的检测技术手段：首先通过质谱仪(mass spectrometer)测量蛋白质样品的质量信息(更严格地说，应该是质量与电荷的比值)，获得样本的质谱信息；然后据此进行计算分析和推理，来获得蛋白质的一级结构信息，该过程即通常所说的蛋白质鉴定(protein identification).

简单地说，质谱仪是使大量分子带上电荷(离子)，并根据不同离子的质量与电荷比的差异而导致它们在电场或磁场中运动轨迹的不同而对这些离子进行分离，进而测量这些离子的质荷比与强度，并记录下来获得质谱数据的一种仪器. 它由离子源(ion source)、质量分析器(mass analyzer)和检测器(detector)三大部分串联而成. 对于蛋白质，由于其分子量很大，并且存在多种蛋白质的分子量非常接近的情况，所以通过质谱仪直接测量全段蛋白质的质量并不能实现可靠的蛋白质鉴定，一般是先通过酶切的方式将分离出的蛋白质切成小的片段，称为肽或肽段，然后再测量这些肽段的质量，形成的质谱称之为肽质量指纹谱(peptide mass fingerprinting, PMF)；根据需要，还常将某个肽离子经过诱导碰撞碎裂(collision-induced dissociation, CID)，碎裂成为更小的碎片离子，测量这些碎片离子的质荷比和强度，获得二级或串联质谱(MS/MS,或 Tandem MS). 本文主要研究从 MS/MS 数据出发，如何设计新的蛋白质鉴定算法(第 2 部分将重点介绍作者的工作).

MS/MS 质谱数据中包含着丰富的信息，要想利用好这些信息必须首先掌握质谱数据的基本特点. 下面以一个典型的 MS/MS 谱图为例进行说明，如图 2 所

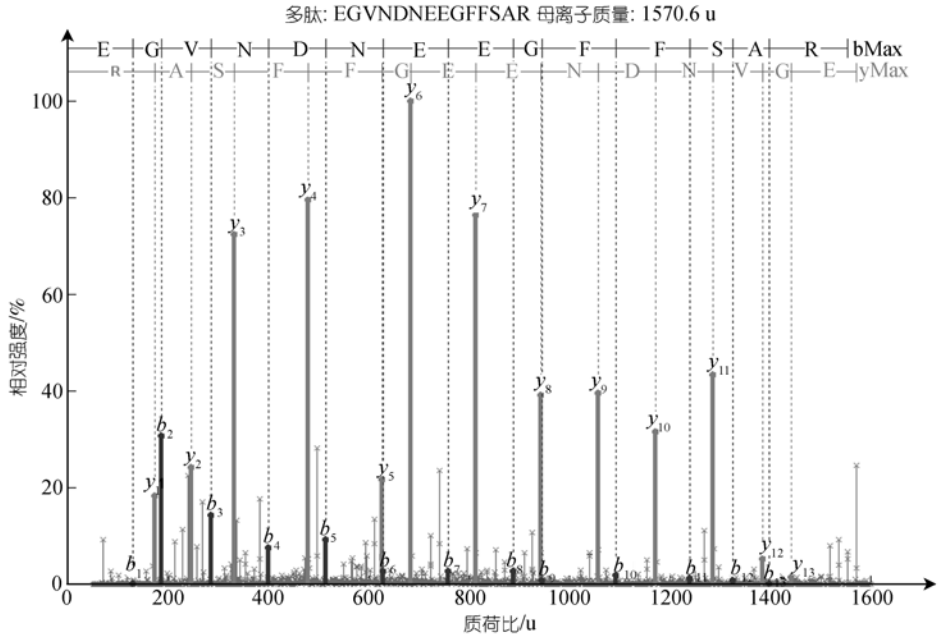

图 2 一个 MS/MS 谱图示例

示，横轴是离子的质量与电荷的比值，即质荷比 *m/z*；纵轴是对应离子的相对强度，表示检测到的离子信号的相对强弱. 该谱图是由 14 个氨基酸组成的序列 EGVNDNEEGFFSAR 的肽段离子经 CID 碎裂产生的，母离子(肽离子)的质量为 1570.6 u.

在肽链的 CID 断裂过程中，通常三种不同类型的肽键断裂方式可以产生六种类型的碎片离子，即 N 端的 *a*, *b*, *c* 型离子与 C 端的 *x*, *y*, *z* 型离子. 每种断裂类型分别生成互补的两种离子，如 *a*-*x*,*b*-*y*,*c*-*z*，如图 3 和图 4 所示. 图中离子类型符号右下标的数字表示该离子所含有的氨基酸残基数目. 另外，断裂后的离子还可能丢失一个中性的水分子或氨分子. 同时，其他更复杂的断裂产生的离子，如侧链断裂等，以及由于污染物而产生的离子，电子信号噪声等都会被检测出来，出现在图 2 所示的 MS/MS 谱图中. 图 2 中已标注出了 *b* 和 *y* 类型离子峰,而对于实验质谱，谱峰类型是未知的.

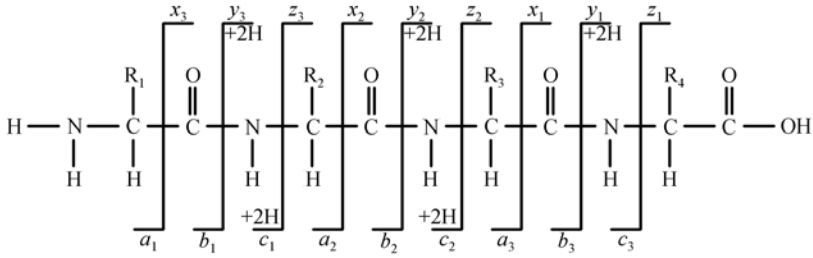

图 3　肽链断裂点与离子类型示意图

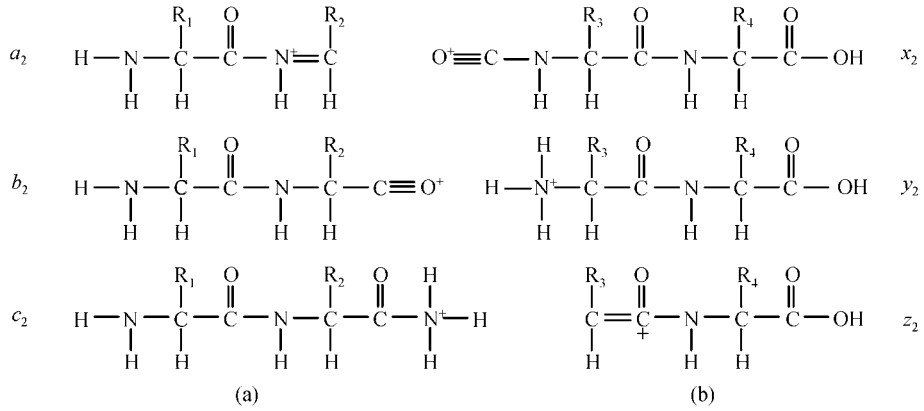

图 4　六种碎片离子类型的化学结构

2　蛋白质鉴定的新算法 KSDP 与软件系统 pFind

利用串联质谱鉴定蛋白质通常有三种方法:

1) 查询蛋白质序列数据库. 这是目前鉴定蛋白质最常用的方法.

2) 从头测序(de novo peptide sequencing). 针对数据库里没有的新蛋白或发生翻译后修饰等情况的蛋白质, 不依赖数据库而直接利用质量较好的串联质谱推理获得序列信息.

3) 首先用从头测序方法获得高可信度的序列片段(tag), 然后利用这些片段辅助查询蛋白质数据库, 是前两种方法的结合.

由于利用实验串联质谱直接查询蛋白质序列数据库的蛋白质鉴定方法使用得最普遍, 所以本文如下工作主要是针对第一种方法而提出来的, 对于后两种方法的研究也是质谱信息处理领域中非常重要的内容.

2.1 蛋白质鉴定的新算法KSDP[6]

用串联质谱查询数据库鉴定蛋白质的核心部分是实验串联质谱与理论串联质谱的匹配打分算法. 理论串联质谱是通过对数据库中的蛋白质序列做模拟酶解, 然后从酶解产生的肽序列预测生成的,当前使用的理论串联质谱预测模型还是非常简单的. 常用鉴定软件中使用的匹配打分算法有SEQUEST软件的互相关(cross correlation)分析[7]和Mascot软件的基于概率的打分算法[8]等. 这些算法在蛋白质鉴定问题上还没有达到满意的程度. 例如, 鉴定的假阳性仍然较高, 可有效利用的质谱数量也仅占实验质谱的 10%左右, 因此需要研究新的更有效的匹配打分算法.

本文作者在对该问题的深入研究和对大量质谱的观测、统计分析, 以及与质谱实验人员的广泛交流的基础上, 发现在匹配打分算法设计过程中, 如何考虑相关离子的匹配程度是一个重要因素, 特别是后来的实验结果发现连续离子匹配的打分策略对鉴定结果具有重要的影响. 所谓相关离子是指质谱中肽碎裂产生的部分离子之间并不彼此独立, 而是具有一定的关联性, 如连续离子是指断裂位点相邻的同类型离子, 它们倾向于在质谱中同时出现. 从这点出发, 作者提出了一种全新的利用离子相关性的匹配打分策略, 并设计实现了相应的算法, 称为核谱向量点积(kernel spectrum dot product, KSDP). 它是对普通打分算法谱向量点积(SDP)的扩展, 借助机器学习领域中的核函数技术, 巧妙利用连续离子匹配信息进行匹配打分算法. 其直观思想是如果连续匹配的离子峰越多, 则鉴定的可靠性会相应的越高. 通过使用核函数技术, 避免了穷举所有可能相关离子组合. 图 5 是 KSDP 算法思想的简单示意图, 图 5(b)与(c)中虽然都是 6 个 y 离子实现了与实验质谱的匹配, 但它们在鉴定结果的可信度上有显著差异, 连续匹配的(b)情况其可靠度更高, KSDP 算法正是利用了这一点而提高了鉴定的可靠度.

基于径向基核函数的KSDP算法与SEQUEST, Sonar MS/MS[9]算法的比较实验结果如表 1 所示, 表中分别对 230 个和 1323 个串联质谱进行搜索数据库鉴定, 从假阳性鉴定结果的数量上可看出, KSDP算法具有较高的优越性, 分析其主要

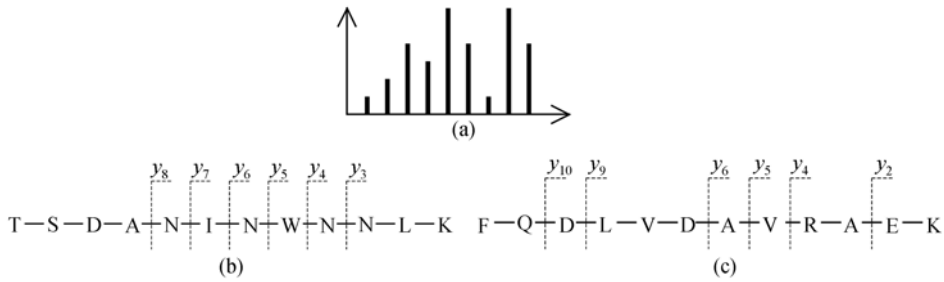

图 5　KSDP 算法的核心思想

(a) 实验质谱示意图; (b) 6 个连续 *y* 离子匹配; (c) 6 个非连续 *y* 离子匹配

表 1　KSDP 算法与其他算法的比较实验结果

数据集	质谱数量	假阳性数		
		KSDP	SEQUEST	Sonar MS/MS
A-1,2	230	10	16	46
A	1323	40	66	–[a)]

a) 由于 Sonar MS/MS 软件没有提供批处理功能, 所以无法对所有 1323 个质谱逐一进行实验

原因是连续性匹配在打分函数中被加以强调, 符合实际的匹配情况, 详细内容请参考文献[6].

2.2　蛋白质鉴定的新软件**pFind**系统[10]

以本文作者提出的上述 KSDP 新算法为基础, 作者设计了利用串联质谱通过查询数据库鉴定蛋白质的新软件系统 pFind, 该系统的主要结构如图 6 所示, 并且提供了网络计算平台(http://pfind.jdl.ac.cn), 如图 7 所示, pFind 是我国第一个具有自主知识产权的蛋白质鉴定网络计算综合系统, 目前已经提供对国内外的服务, 并正在研发新的 2.0 版本, 以进一步提升 pFind 的综合性能. 而要提高

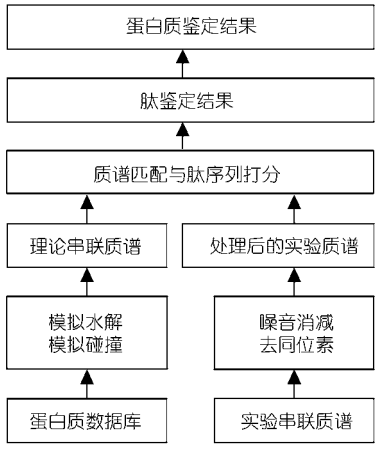

图 6　pFind 系统的主要结构图

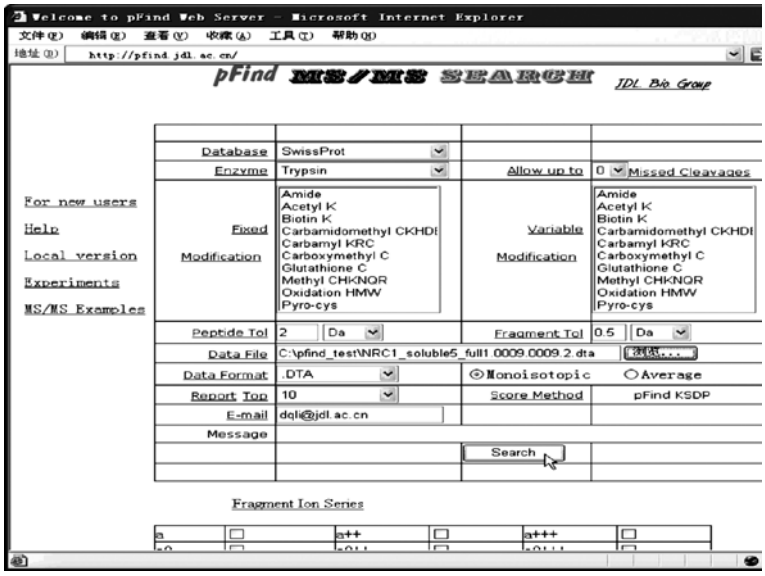

图 7 pFind 鉴定系统的网络计算综合平台

pFind 的性能，必须对质谱的特性进行进一步的挖掘，下一部分将介绍我们在质谱同位素模式信息的发现与应用问题上的研究.

3 质谱中同位素模式的计算与应用[11]

由于质谱实验过程中多种因素的影响，质谱中存在着大量的物理噪声、化学噪声，不规则碎裂的离子谱峰，部分离子谱峰缺失，谱峰重叠等复杂现象，这导致直接在原始质谱数据上识别肽序列存在着很多的干扰. 我们通过对大量串联质谱图的观察和研究分析，发现其中的同位素峰具有稳定的模式可以加以利用. 众所周知，多肽主要是由 C,H,N,O,S,P 元素组成，它们在天然情况下存在着稳定的同位素，不同元素组成的肽，存在着不同的同位素分布，我们称之为同位素模式，如图 8. 这些同位素模式可以帮助判断是否存在谱峰重叠现象. 有效的多肽碎片离子会存在一系列有规律的同位素，噪声则不然. 因此首先可以应用同位素模式来识别质谱中的有效峰和噪音.

根据质谱中的同位素模式信息，作者提出了一种利用同位素模式来预测离子分子式的新算法FFP(fragment formula prediction)[11]. 根据C,H,N,O,S,P元素的天然稳定分布，可以计算出一个分子式的理论同位素模式，与实验谱中的模式距离最小的分子式认为是正确的，加上对一个有效分子式的线性约束，可将分子式预测问题建模为一个二次规划问题. 通过求解该问题，同时结合数据库中的统计规则，可以高精度确定Q-TOF型质谱 500 u以下谱峰所对应的离子分子式.

在 50 组Q-TOF 数据上的验证结果如图 9，对(0~300 u)范围内的离子分子式，

预测首选正确率为 83%, 5 选正确率为 97%; 在(300~500 u)内 5 选正确率为 95%, 详细结果参考文献[11]. 目前正在研究如何将预测的分子式同蛋白质的鉴定, 从头测序等结合起来.

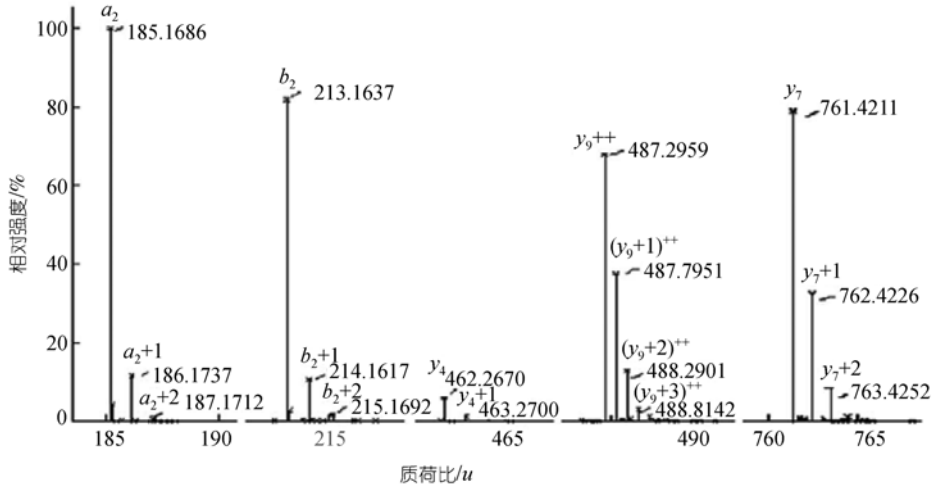

图 8　四个同位素模式

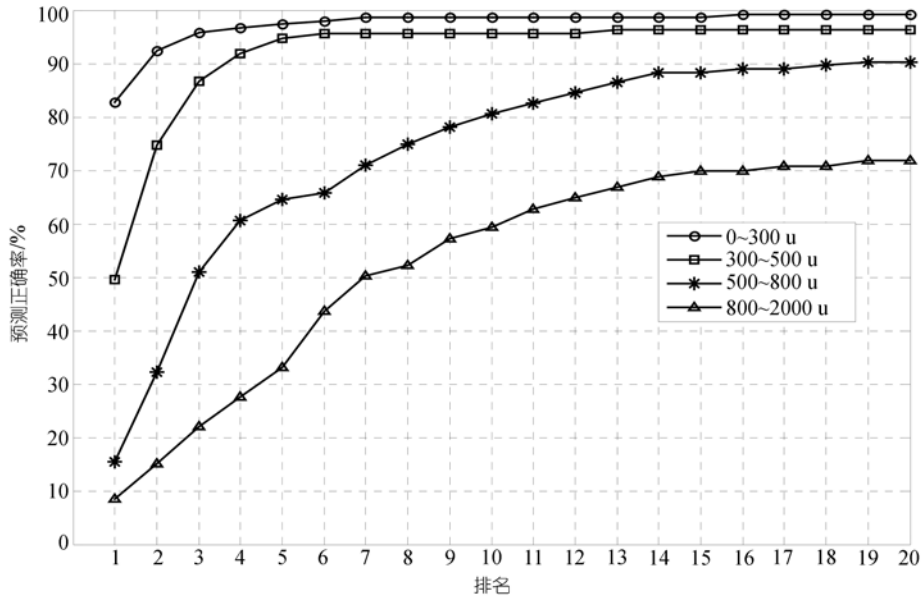

图 9　FFP 分子式预测的正确率

4　基于质谱技术的定量蛋白质组学与疾病蛋白质组学

前面介绍的蛋白质鉴定问题只是从定性上研究样品中存在哪些蛋白质, 而仅知道是哪些蛋白质还远远不够, 特别是对于疾病的研究, 需要知道对于某种疾

病,哪些特异蛋白质的表达量与正常人相比发生了显著的变化, 如减少或增加,这对于疾病的诊断与治疗具有特别重要的意义. 因此, 不仅需要定性上研究蛋白质,而且需要从定量上研究蛋白质的表达量变化信息. 定量蛋白质组学(quantitative proteomics)已经成为当前蛋白质鉴定基础上的另一个重要问题. 同时, 利用质谱技术如何发现某些重大疾病, 如癌症的生物标记物和实现疾病的早期诊断是疾病蛋白质组学(disease proteomics)的研究内容, 定量蛋白质组学与疾病蛋白质组学具有一定的联系, 前者可以为后者提供更多的信息. 下面分别介绍一下它们的研究内容.

4.1 定量蛋白质组学

定量蛋白质组学是从定量的角度研究蛋白质组, 通过定量蛋白质组学的技术和方法, 可以实现大规模鉴定蛋白质和评估蛋白质的表达水平高低, 进一步认识细胞和分子机制, 提供疾病和药物治疗的新的生物标记物[13]. 目前主要集中在相对量化信息的获取研究上, 包括实验手段和计算手段两个方面.

定量蛋白质组学主要有两种方法:

一种是通过无标记的蛋白质二维凝胶电泳(two dimensional gel electrophoresis, 2DE). 首先对蛋白质混合物进行分离, 然后通过比较不同凝胶上的表示同一个蛋白质的点的着色强度来得到其定量信息. 最后对选出的蛋白质点经过胶切、蛋白酶水解和质谱检测进行鉴定分析.

另一种是近几年发展起来的通过标记物实现量化信息的获取, 基于蛋白质的同位素标签[12]和自动化的串联质谱检测分析方法. 在这种方法中, 通过代谢标记、酶反应或化学反应, 用不同的同位素标签标记存在于不同样品中的蛋白质, 然后将不同标记的样品混合在一起, 同时进行处理分析. 在本方法具体实施时, 该混合的差异标记样品经过酶解, 得到的肽混合物再经过液相色谱(liquid chromatography, LC)和串联质谱分析. 这样, 蛋白质的定性鉴定与定量分析是通过其对应的肽的鉴定与定量完成的.

定量蛋白质组学目前的困难主要在于: ① 源于同一个蛋白质的多个肽可能被鉴定出来, 而它们的量化信息不一致. ② 同一个肽可能以不同的同位素身份或带电荷状态被多次鉴定出来. ③ 某个肽可能是错误鉴定的或翻译后修饰的. ④ 肽信号的数据质量, 即信噪比(signal-to-noise ratio, S/N)有时较低. 所以, 人们要想可靠且自动地评价蛋白质的量化结果, 必须得有高级的数据处理方法来综合考虑所有这些复杂的影响因素.

总之, 在蛋白质组学研究中, 定量蛋白质组学仍处在刚起步阶段, 许多实验材料、分析仪器、实验方法以及相应的数据分析系统都还处在不断完善、发展之中, 因此, 对计算方法, 工具的开发也提出了新的要求, 针对具体的实验方法,

具有高准确度和高速度的自动化蛋白质量化分析软件, 是当前高通量定量蛋白质组学的主要瓶颈.

4.2　疾病蛋白质组学

疾病蛋白质组学主要研究寻找各种疾病的特异性标志蛋白质, 进而应用于临床诊断和药物开发等, 因此也常称为临床蛋白质组学(clinical proteomics). 2005 年 7 月 25 日在长春召开的中国蛋白质组学第三届大会(http://www.bioon.com/biology/advance/cancer/200508/149930.html)和 2005 年 8 月 28 日在德国慕尼黑召开的 HUPO(The human proteome organization)第四届世界大会上(http://www.hupo2005.com/), 疾病蛋白质组学都成为了一个特别引人注目的专题, 特别是肿瘤蛋白质标记物成为报告的热点之一.

很多学者认为, 多数疾病在症状明显出现之前就已经在蛋白质的种类和数量上发生了一些变化, 这些发生了相对稳定表达量变化的特异蛋白质常被称为疾病的“生物标志物(biomarker)”. 如果能够及时检测到这些变化, 对正常人群和疾病人群的蛋白质组进行表达模式差异分析, 发现疾病的生物标志物, 就可以在最早期阶段发现疾病, 这成为当前疾病蛋白质组学的一个研究焦点. 因为, 找到的那些蛋白质标志物, 都可能成为疾病早期诊断的灵敏工具, 这蕴涵着难以估量的产业价值, 将会极大地提高人民的健康和国民经济的发展[13,14]. 据文献报道, 国际上在卵巢癌、前列腺癌等肿瘤的疾病蛋白质组研究中已经取得初步成果. 如发现了由 5 个肽段峰组成的特征性血清蛋白质组模式, 在小规模测试集上诊断卵巢癌的敏感性和特异性分别达到了 100%和 95%, 远高于目前临床上常用的卵巢癌标记物CA125[15].

从模式识别的角度分析, 应用质谱技术寻找重大疾病, 如各种癌症的生物标记物, 进而实现疾病的早期诊断问题, 是一个特征选择和模式分类的问题. 目前主要使用的美国 Ciphergen 公司开发的以决策树算法为核心的 SELDI-TOF(surface-enhanced laser desorption/ionization time-of-flight, 表面增强激光解吸电离飞行时间)质谱分析系统. 我国的生物或医学各大科研机构, 包括高校和医院, 很多都在应用这套质谱分析技术来进行蛋白质和肽的理论研究与应用技术开发, 也已经取得了部分成果. 如北京师范大学等, 已经积累了不少的质谱数据.

利用蛋白质的SELDI-TOF质谱技术进行疾病早期诊断的基本流程如图 10 所示[16]. 首先提取病人和正常健康人(为了对照)的血样或尿样等体液样本, 进行相关的生化处理后滴入蛋白质芯片, 由于蛋白质芯片的亲和性, 将样本中的部分蛋白质吸附在芯片上, 经过漂洗将非特异性蛋白去除再进行干燥处理; 然后送入SELDI-TOF质谱仪中进行蛋白质分子的离子化, 测量不同蛋白质或其片段对应

的质量与电荷比m/z，获得图 10 中所示意的质谱图(横轴为m/z,纵轴为对应离子的强度值)；利用模式识别技术通过对大量病人和正常人质谱的分析处理，找出可以稳定区分病人与正常人的标志蛋白集，进而应用其进行疾病的早期诊断.

从图 10 中可看出，利用 SELDI-TOF 质谱进行疾病模式识别最关键的两个环节是质谱的产生与模式识别算法，后者主要涉及质谱的预处理，特征提取与分类

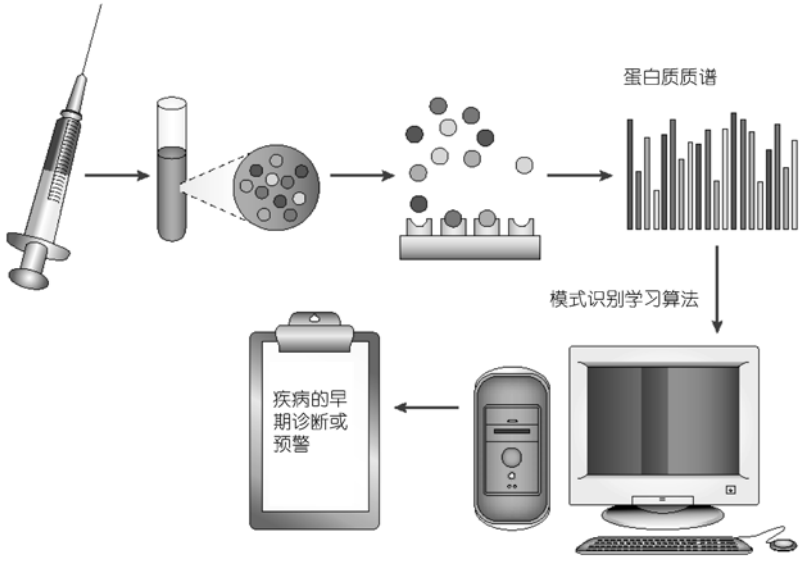

图 10 利用蛋白质SELDI-TOF质谱模式进行疾病早期诊断的基本流程[16]

器设计三步. 预处理包括基线校正，谱峰检测，强度归一化，m/z 值对中，去噪等处理；特征提取是对预处理后的谱峰进行筛选，剔除与分类不相关或冗余的特征，提取出具有高区分性能的特征集用于后面的分类器设计；分类器设计则是利用训练样本对分类器进行优化设计，通过验证和完善，最后应用进行实际新样本的分类决策，如判断测试样本是否患上某种疾病，疾病的类型与阶段等.

通常情况下，SELDI-TOF 质谱数据具有“多”与“少”两个显著特点：“多”是指每一个样本对应的谱数据多. 为了保证足够高的测量精度和宽的测量范围，获得的原始质谱每张谱具有上万，甚至几十万个谱数据，这些数据中混有基质(matrix)等化学噪声和信号 A/D 转换带来的电子噪声. 如何从这些复杂的高维质谱数据中找出具有生物本质差异的稳定标志蛋白集不是一件容易的事情，这需要深刻理解质谱的产生机理和有关蛋白质组成的生物化学知识，以及数据预处理和信号分析技术. “少”则是指获得的样本数目相对较少，一般在几十个至上百个的数量级，这就对可利用的模式识别技术和统计方法提出了较高的要求. 小样本，多特征容易造成“过拟合(overfitting)”问题. 研究如何有效地解决这些问题，将多种技

术有机地结合起来, 设计并实现完整的系统在实际应用中不断完善是目前需要重点研究和实践的内容.

5　结论与展望

质谱技术是当前蛋白质组学研究, 特别是蛋白质定性与定量分析中最重要的技术手段之一, 从采集到质谱数据的那一刻开始, 就产生了从这些数据中发现重要的蛋白质知识的需求, 这对于计算蛋白质组学的发展提供了良好的机遇, 研究利用质谱数据来鉴定蛋白质, 需要深入了解质谱数据的产生背景和质谱数据的特点, 一方面可以从物理产生机理上获得; 另一方面也可以从积累的大量质谱数据中去挖掘和发现, 获得新的规律、知识. 本文介绍了作者研究的利用串联质谱进行蛋白质鉴定的新算法 KSDP 与开发的新系统 pFind, 以及如何挖掘利用质谱中的同位素模式信息, 并对定量蛋白质组学和疾病蛋白质组学进行了介绍. 其中后两个方向是在鉴定基础上未来深入发展的方向, 在计算蛋白质组学领域中, 定量蛋白质组学和疾病蛋白质组学更接近实际应用, 因而需求更多, 这将会促进对这两个新兴方向的研究投入.

基于质谱技术的计算蛋白质组学研究虽然已经取得了良好的开端, 但当前对于质谱数据的利用, 还远远没有达到满意的程度. 从计算的角度, 质谱数据可以看作是一个两维数组, 第一维代表谱峰的质荷比, 第二维为每个质荷比所对应的谱峰强度, 目前的蛋白质鉴定系统中, 质荷比信息应用比较广泛, 而谱峰强度信息由于具有不确定性和影响因素复杂等原因, 还远远没有得到有效的利用, 这仍是一个值得深入研究的重要课题. 除此之外, 围绕以质谱鉴定蛋白质为中心的质谱信息处理方面的研究问题还有很多, 如翻译后修饰的处理, 鉴定结果的可靠评估, 定量信息的获取, 以及满足质谱分析人员要求的软件系统的研发等. 这些问题也是计算蛋白质组学今后需要研究的重点内容.

展望未来, 蛋白质组学在疾病诊断和药物设计等与人类健康息息相关的领域中必将发挥出更加出色的作用, 而计算技术也会因为这些具体的应用问题的解决自身也会得到进一步的扩展与壮大.

致谢　感谢中国科学院计算技术研究所的王海鹏、蔡津津、邹翠等在算法方面的研究与讨论, 王乐珩、魏勇刚等在 pFind 软件开发方面的突出贡献以及中国科学院上海生物化学研究所、军事医学科学院放射医学研究所等单位的大力支持.

参　考　文　献

1　Patterson S D, Aebersold R H. Proteomics: the first decade and beyond. Nature Genetics Supplement, 2003, 33:311~323[DOI]

2 Johnson R S, Davis M T, Taylor J A, et al. Informatics for protein identification by mass spectrometry. Methods, 2005, 35: 223 ~236[DOI]
3 Aebersold R H, Mann M. Mass spectrometry-based proteomics. Nature, 2003, 422:198~207[DOI]
4 Steen H, Mann M. The ABC's(and XYZ's) of peptide sequencing. Nature Reviews Molecular Cell Biology, 2004, 5:699~711[DOI]
5 Russell S A, Old W, Resing K A, et al. Proteomic informatics. International Review of Neurobiology, 2004, 61:129~157
6 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, 20:1948~1954[DOI]
7 Eng J K, McCormack A L, Yates J R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of American Society of Mass Spectrometry, 1994, 5: 976~989[DOI]
8 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: 3551~3567[DOI]
9 Fenyö D, Beavis R C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Analytical Chemisty, 2003, 75:768~774[DOI]
10 Li D, Fu Y, Sun R, et al. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. Bioinformatics, 2005, 21(13): 3049~3050[DOI]
11 Zhang J, Gao W, Cai J, et al. Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005, 2(3):217~230[DOI]
12 Gygi S P, Rist B, Gerber S A, et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature Biotechnololy, 1999, 17: 994~999[DOI]
13 何大澄, 肖雪媛. SELDI 蛋白质芯片技术在蛋白质组学中的应用. 现代仪器, 2002, 1:1~4
14 Liotta L A, Ferrari M, Petricoin E F. Clinical proteomics: written in blood. Nature, 2003, 425:905[DOI]
15 Petricoin E F, Ardekani A M, Hitt B A, et al. Use of proteomic patterns in serum to identify ovarian cancer. The Lancet, 2002, 359: 572~577[DOI]
16 Wulfkuhle J D, Liotta L A, Petricoin E F. Proteomic applications for the early detection of cancer. Nature Reviews Cancer, 2003, 3: 267-275[DOI]