中国科学: 生命科学 2014 年 第 44 卷 第 11 期: 1113 ~ 1124
SCIENTIA SINICA Vitae www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述

利用串联质谱鉴定氨基酸突变的生物信息学算法

余庆[①②], 吴松锋[②③], 马洁[②③], 朱云平[②③*], 舒坤贤[①*]

① 重庆邮电大学生物信息学研究所, 重庆 400065;
② 北京蛋白质组研究中心, 蛋白质组学国家重点实验室, 北京 102206;
③ 中国人民解放军军事医学科学院放射与辐射医学研究所, 北京 100850
* 联系人, E-mail: zhuyunping@gmail.com; shukx@cqupt.edu.cn

收稿日期: 2014-03-05; 接受日期: 2014-04-22; 网络版发表日期: 2014-09-17
国家自然科学基金(批准号: 21105121, 21275160)、国家高技术研究发展计划(批准号: 2012AA020409, 2012AA020201)和北京市自然科学基金(批准号: 5122013)资助项目
doi: 10.1360/N052014-00099

摘要 氨基酸突变能够改变蛋白的结构和功能, 影响生物体的生命过程. 基于串联质谱的鸟枪法蛋白质组学是目前大规模研究蛋白质组学的主要方法, 但是现有的质谱数据鉴定流程为了提高鉴定结果的灵敏度往往会有意压缩数据库中的氨基酸突变信息. 因此, 如何挖掘数据中的氨基酸突变信息成为当前质谱数据鉴定的一个重要部分. 当前应用于氨基酸突变鉴定的串联质谱鉴定方法大致可以分为 3 大类: 基于序列数据库搜索的方法、基于序列标签搜索的算法以及基于图谱库搜索的算法. 本文首先详细介绍了这 3 种氨基酸突变鉴定算法, 并分析了各种方法的特点和不足, 然后介绍了氨基酸突变鉴定的研究现状和发展方向. 随着基于串联质谱的蛋白质组学的不断发展, 蛋白序列中的氨基酸突变信息将被更好地解析出来, 从而得以深入探讨由氨基酸突变引起的蛋白结构和功能改变, 为揭示氨基酸突变的生物学意义奠定基础.

关键词
氨基酸突变
串联质谱
鸟枪法蛋白质组学
氨基酸突变鉴定

单核苷酸变异(single nucleotide variations, SNVs)是由 DNA 序列上单个碱基变异产生的, 包括碱基的缺失、插入、转换及颠换等. SNVs 是基因组序列变异的主要形式[1], 同时也是生物体生理和病理变异的遗传基础[2]. 从遗传学的角度看, SNVs 既可以存在于具有遗传性的生殖细胞中, 也可以存在于不具有遗传性的体细胞中. 其中, 只有位于基因编码区的 SNVs 能够影响蛋白的编码. 位于编码区的 SNVs 可以分为 3 类: (i) 同义 SNVs, 不改变相应的氨基酸种类; (ii) 无义 SNVs, 突变成为终止密码子, 提早结束编码; (iii) 非同义 SNVs(nonsynonymous SNVs, nsSNVs), 改变氨基酸的种类. nsSNVs 能够改变蛋白的结构、功能、表达以及亚细胞定位等[3], 进而对多种遗传性的特征、疾病以及癌症等产生影响[4~9], 如人类耳垢的类型[6]、腋窝的气味[7]、癌症与肿瘤的发生[8]、阿尔茨海默病[9]以及镰刀形红细胞贫血症[10]等. 因此, 对 SNVs 展开研究可以揭示出基因与表型多样性和基因与疾病间的关系, 并且有可能研发出治疗疾病的新方法. 目前, 全基因组关联研究(genome-wide association studies, GWAS)[11]虽然在基因变异与

引用格式: 余庆, 吴松锋, 马洁, 等. 利用串联质谱鉴定氨基酸突变的生物信息学算法. 中国科学: 生命科学, 2014, 44: 1113–1124
Yu Q, Wu S F, Ma J, et al. Bioinformatics algorithms for identification of amino acid mutations in tandem mass spectrometry. SCIENTIA SINICA Vitae, 2014, 44: 1113–1124, doi: 10.1360/N052014-00099

表型多样性的研究中产出了许多能够用来解释特异性疾病分子途径的结果, 但是仍然难以对绝大部分具有复杂特征的分子机制以及SNVs与复杂疾病表型间的关系进行解释[12]. 在这种情况下, 对突变蛋白的研究提供了另一种了解基因型与表型间关联的方法[13].

由 SNVs 引起的单个氨基酸的变异称为单氨基酸变异(single amino acid variations, SAVs), 因此 SAVs 是 SNVs 在蛋白水平上的表现. 对 SAVs 的研究, 有助于了解基因型与表型间的关系, 进而从本质上了解基因是怎样在蛋白水平上影响生物体的生命过程的[14]. 目前, 基于串联质谱的鸟枪法蛋白质组学(shotgun proteomics)技术由于其自动化、高通量、高灵敏度和高分辨率等特点, 已成为大规模蛋白质研究的主要方法. 序列数据库搜索算法由于具有较高的可靠性以及灵敏度而成为当今鸟枪法蛋白质组学中蛋白鉴定的主要生物信息学方法. 然而, 通常蛋白质数据库在构建时为了减小数据库的冗余程度, 往往有意压缩对 SAVs 信息的收录(如 Swiss-Prot 数据库[15,16], IPI 数据库[17]等), 从而使得常用的数据库搜索策略不能有效地鉴定出样本中的氨基酸突变信息. 为此, 研究人员提出了一系列鉴定突变蛋白的方法, 如构建包含有突变信息的蛋白质数据库、构建相似性图谱库等.

在基于串联质谱进行 SAVs 鉴定时, 可以采用与蛋白质翻译后修饰(post-translational modifications, PTMs)鉴定[18]相同的方法, 这是因为肽段的突变和修饰在质谱图中的表现都是质量迁移, 如甲硫氨酸(Met)氧化与丙氨酸(Ala)突变为丝氨酸(Ser)在质量上都是增加 16 Da[19], 所以鉴定 PTMs 的算法和流程通常也能够鉴定 SAVs(如 Bonanza 算法[20]). 虽然 PTMs 和 SAVs 的质谱鉴定方法非常相似, 但由于其来源上的差别, 在实际的鉴定策略中有所不同. (ⅰ) PTMs 的种类远比 SAVs 要多, 鉴定 PTMs 所需的搜索空间一般会比鉴定 SAVs 所需的大, 在质量控制方面具有更大的挑战; (ⅱ) 蛋白水平的 SAVs 大部分是从基因组或转录组延续过来的, 充分利用 SNVs 的数据能大大降低搜索空间, 从而得到更可靠的结果. 因此在计算方法与策略方面, SAVs 和 PTMs 的鉴定具有一定的相似性, 也有其独有的特点.

本文从序列数据库搜索算法、序列标签搜索算法以及图谱库搜索算法 3 个大方面, 详细地介绍了目前基于生物质谱数据鉴定SAVs的各种生物信息学方法, 并分析了各种突变鉴定方法的不足之处, 最后介绍了基于生物质谱的 SAVs 鉴定研究现状及其发展方向.

1 氨基酸突变鉴定的算法

当前基于生物质谱的 SAVs 鉴定算法都是由常规鉴定算法改进而来的, 因此根据常规串联质谱鉴定算法中对数据库的依赖程度以及使用的数据库种类, 可以将基于生物质谱的 SAVs 鉴定算法分为 3 大类(表 1): (ⅰ) 完全依赖序列数据库的搜索算法, 即基于序列数据库搜索的氨基酸突变鉴定算法. 此算法利用前体离子质量从序列数据库中筛选出候选肽段, 然后将候选肽段的理论图谱与目标图谱进行比对, 从而鉴定出样品中的突变肽段; (ⅱ) 将从头测序算法(*de novo*)与序列比对结合的算法, 即基于序列标签的氨基酸突变鉴定算法. 此算法首先通过 *de novo* 测序算法推导出目标图谱中的肽序列标签(peptide sequence tags, PSTs), 然后利用PSTs过滤数据库筛选出候选肽段, 最后结合 PSTs 对理论谱图与目标图谱进行比较打分, 从而鉴定出样品中的突变肽段; (iii) 依赖于图谱库的搜索算法, 即基于图谱库的氨基酸突变鉴定算法. 此算法将实验图谱与图谱库中的一致性图谱进行比对, 从而鉴定出样品中的突变肽段. 这 3 类方法和策略在实施过程中各有其优劣(表 1), 相互之间暂无法替代, 因此在不同的目的下各有其适用性.

1.1 基于序列数据库搜索的氨基酸突变鉴定算法

基于序列数据库搜索的氨基酸突变鉴定算法, 根据不同的数据库构建方法可以细分为 3 类: (ⅰ) 基于穷举法的氨基酸突变鉴定算法, 即通过枚举数据库中氨基酸残基的所有可能突变种类进行突变肽段的鉴定; (ⅱ) 结合已知氨基酸突变信息对突变肽段进行鉴定, 即结合当前变异数据库(如 dbSNP 数据库[21]、COSMIC 数据库[22]等, 表 2 列举了常用的氨基酸与基因突变数据库)中的变异信息构建数据库进行突变肽段的鉴定; (iii) 基于样本特异性的数据库鉴定突变肽段, 即结合样本数据中可能存在的突变肽段信息构建数据库进行突变肽段的鉴定. 以下将对这 3 种方式进行逐一详细地说明.

表 1 氨基酸突变鉴定的算法

类别	优点	不足
基于序列数据库搜索的氨基酸突变鉴定算法	现有的蛋白突变序列信息相对比较丰富, 覆盖度比较广; 理论上能够鉴定到所有可能的突变肽段	依赖于数据库所包含的信息; 搜索空间较大、灵敏度较低
基于序列标签的氨基酸突变鉴定算法	能够有效地过滤数据库, 降低搜索空间; 具有较强的处理能力和用于突变肽段鉴定的潜能	依赖 *de novo* 算法以及过滤算法
基于图谱库的氨基酸突变鉴定算法	利用图谱的真实信息; 有较高的灵敏度	图谱库的覆盖范围小; 图谱搜索引擎对图谱的解析度较低

表 2 常用氨基酸及基因突变数据库

数据库	数据库简介	网址	参考文献
COSMIC	收录和展示体细胞突变的信息和相关细节, 以及与人类癌症有关的信息	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/	[22]
OMIM	收录遗传疾病、表型与人类基因, 主要关注遗传变异与表型间的关系	http://www.ncbi.nlm.nih.gov/omim/	[23]
MS-CanProVar	收录各个数据库中有关癌症的蛋白变异数据和 dbSNP 数据库中的编码 SNVs	http://bioinfo.vanderbilt.edu/canprovar/	[24]
HPMD	收录各个大型数据库中的蛋白突变信息	http://www.exocarta.org/HPMD	[25]
HGMD	收集已知的与人类遗传疾病有关的基因变异信息	http://www.hgmd.cf.ac.uk/ac/index.php	[26]
dbSNP	收录各种变异基因的信息	http://www.ncbi.nlm.nih.gov/snp/	[21]

(1) 基于穷举法的氨基酸突变鉴定算法. 在序列数据库搜索中, 最早对突变肽段进行鉴定的自动化方法是穷举法, 此方法不仅原理简单而且理论上能够鉴定出样品中所有可能的突变肽段. 这类算法的大体步骤是: 通过穷举法罗列出所有可能的突变肽段序列, 然后用常规鉴定方法进行比对打分筛选出最有可能的突变肽段序列. 此类算法的代表有 SEQUEST-SNP 算法[27]和 Sipros v2.0 算法[18]等. Gatlin 等人[27]在 2000 年, 利用改进的 SEQUEST 算法(SEQUEST-SNP)率先实现了利用自动化的数据库搜索对突变肽段进行鉴定. 此方法特点在于动态生成所有可能的核苷酸突变序列, 将其翻译成肽段并构建成一个数据库用于对突变肽段的鉴定. 此后, 通过穷举蛋白序列中所有可能的氨基酸突变进行肽段突变鉴定的方法在 Mascot[28]和 X!Tandem[29]相继采用. 2012 年, Hyatt 和 Pan[18]提出了不受数据库约束的穷举法突变肽段鉴定算法 Sipros v2.0, 此算法通过肽段产生模块和肽段打分模块实现对 CPU 和内存效率的优化以应对穷举法产生的大数据库. 理论上, 穷举法能够鉴定出样品中所有的突变肽段, 但肽段中的每一个氨基酸残基都有 18 种可能的突变, 因此利用此方法会大大增加搜索空间[18,24], 延长搜索时间, 并且会增加假阳性风险从而降低结果的灵敏度.

(2) 结合已知氨基酸突变信息对突变氨基酸进行鉴定. 为了避免穷举法引起搜索空间过大的问题, 一些团队提出结合已知的编码 SNVs 信息或是与疾病等有关的突变信息构建蛋白质数据库, 以减小突变肽段的搜索范围. 此类数据库的代表有 MSIPI[17]和 MS-CanProVar[24]等. 2007 年, Schandorff 等人[17]将一些来自 dbSNP 数据库[21]的编码 SNP(single nucleotide polymorphism) 以及与 IPI(the international protein index)数据库中数据有冲突的序列等整合到 IPI 数据库[30]中构建了质谱友好型的变异数据库 MSIPI. 其质谱友好型体现在, 在保留原始 IPI 条目完整性的基础上, 将后加的肽段序列附加到原有序列中, 用不代表任何氨基酸的字母“J”将原始条目与附加肽段区分开来, 并且将在原始条目的表头信息中加入附加肽段信息. 同年, Bunger 等人[31]也利用 dbSNP 数据库中人类基因变异信息构建变异蛋白质数据库 K-SNPdb, 并构建相应的常规数据库. 然后对分开搜库结果进行比对打分, 筛选出高可信的变异肽段. Li 等人[24]在 2011 年基于人类癌症蛋白质变异数据库 CanProVar[32]构建了一个 MS-CanProVar 数据库, 此数据库中不仅包含了 dbSNP 数据库中的编码的 SNP 信息, 还包括了 COSMIC[22]和 OMIM[23]等数据库中与癌症相关的体细胞变异信息.

除了自定义构建突变数据库以外, 氨基酸突变信息也被一些在线平台收录、整合, 如 Swiss-Var[33],

SysPIMP[34]和 RAId_DbS[35]等. Swiss-Var 网站搜集的是 Swiss-Prot 数据库[36]中突变肽段的信息, 主要为用户提供 Swiss-Prot 数据库中的突变肽段信息及其与疾病间的关系. SysPIMP 主要用于鉴定与人类疾病有关的突变肽段序列, 它的数据主要来源于 OMIM 数据库中等位基因突变信息、蛋白质突变数据库(protein mutation database, PMD)[37]以及 Swiss-Prot 数据库中与人类疾病和多态性有关的序列信息. 而在 RAId_DbS 数据库中不仅整合了 SAVs 与疾病的信息, 同时也收录了 PTMs 与疾病有关的信息.

2012 年, Mathivanan 等人[25]提出的 iMASp 策略即是利用现有的突变信息对突变肽段进行鉴定. 这种策略利用了分步搜索的方法, 即是第一次通过常规搜索鉴定出样本中的常规蛋白, 第二次利用突变数据库对第一次没有鉴定出的质谱图进行搜索鉴定样品中的突变肽段. 相比穷举法, 结合已知氨基酸突变信息对突变氨基酸进行鉴定的方法虽然在一定程度上缩小了搜索空间, 但在数据库中添加的上万条突变肽段序列绝大部分不会在样品数据集中出现. 因此, 这种方法并没有十分有效地规避假阳性升高以及鉴定结果灵敏性降低的缺点[14].

(3) 基于样本特异性的数据库鉴定突变肽段. 除了直接利用公共数据库中的突变数据外, 利用 DNA/RNA 等信息提供的样本特异性突变构建的数据库能更好地贴合实际样本数据, 提高鉴定效率. 目前利用样本特异性鉴定突变肽段的方法有 2 种: 两次搜索数据库的方法以及利用转录组数据构建数据库的方法. 两次搜索数据库的方法与 iMASp 策略中所使用的分步搜索以及 Mascot 和 X!Tandem 中的容错搜索相似, 不同的地方在于两次搜索数据库中所使用的突变数据库依赖于样本特异性的 DAN/RAN 信息, 而 iMASp 策略中的突变数据库是整合所有已知的蛋白突变信息, 不具有样本特异性; Mascot 和 X!Tandem 则是对第一次搜索所得的蛋白序列进行穷举从而鉴定出突变或修饰肽段. Chernobrovkin 等人[38]提出的二次迭代法以及 Su 等人[39]构建样本特异性突变数据库的策略都是样本特异性的两次搜索方法的代表.

另一种方法是利用转录组数据构建样本特异性数据库用于突变肽段的鉴定. 相对于利用公共的突变数据库, 利用转录组数据构建蛋白质数据库可以由样品转录组数据直接推导样本中可能存在的蛋白及其突变序列并由其构建数据库[40]. 用此方法构建的数据库所包含的蛋白质信息更加接近样品中真实信息, 因此这种无偏性的数据库能高效地鉴定出样品中存在的突变序列[16,41]. 由于转录组数据十分庞大, 在现有的计算能力下要想利用转录组数据构建数据库就必须要对转录组数据进行压缩. 2007 年, Edwards[16]提出了一个压缩表达序列标签(expressed sequence tags, ESTs)数据的策略, 实现了利用 EST 数据库进行常规化的肽段序列和变异位点的鉴定. 此压缩策略的特点在于选用某种方法来表示肽段, 确保绝大多数的重复肽段序列被消除, 并且不影响肽段序列的鉴定. 随着下一代测序(next generation sequencing, NGS)技术的出现, RNA 测序(RNA-sequecing, RNA-Seq)的成本越来越低[14], 并且克服了 EST 测序存在的克隆偏性和高花费等缺点[42], 因此利用 RNA-Seq 数据构建样本特异性数据库逐渐受到人们的重视. Wang 等人[41]在 2012 年提出了一个利用 RNA-Seq 数据构建样本特异性数据库的策略, 此策略通过两步来实现: (ⅰ) 利用一个经验性的 RPKM (reads per kilo bases per million reads)值排除不表达或低表达基因以减小数据库中的条目; (ⅱ) 将由 RNA-Seq 数据鉴定得来的高可靠性 SNVs 的相应肽段添加到数据库中, 以寻找变异肽段. 此后, Wang 和 Zhang[43]为生成自定义 RNA-Seq 数据库编写了 R 程序包 customProDB, 能够生成含有突变、插入、缺失等变异肽段的 RNA-Seq 数据库. 2013 年, Sheynkman 等人[14]实践了 Wang 和 Zhang[43]的方法, 利用 Jurkat 细胞系的 RNA-Seq 数据构建一个自定义的变异蛋白质数据库, 并成功地应用在 Jurkat 细胞系的质谱数据突变鉴定中. 同年, Woo 等人[44]在尽量不影响鉴定结果灵敏性的基础上, 将秀丽隐杆线虫(*Caenorhabditis elegans*)的 RNA-Seq 数据压缩了近 1000 倍, 并利用此数据库成功地鉴定到了新型蛋白. 由于并不是所有的样本都同时拥有蛋白质数据和 RNA-Seq 数据, 因此, Wang 和 Zhang[43]利用 64 个大肠癌的 RNA-Seq 数据构建了一致性蛋白质数据库, 并成功地将此数据库应用在蛋白鉴定中. 样本特异性的数据库, 特别是利用 RNA-Seq 数据构建的样本数据库不仅能够有效地缩减搜索空间, 而且能够鉴定出样品中所有已知类型的蛋白种类以及新型的变异肽段序列. 随着计算方法的不断改进, 通过 RNA-Seq 数据对样本进行突变肽段的鉴定方法有望成为常规的突变鉴定方法.

(4) 基于序列数据库搜索的氨基酸突变鉴定算法的缺点. 在鉴定突变肽段的方法中, 虽然通过构建含有突变信息的序列数据库鉴定突变肽段的方法是目前被最广泛采用的方法, 但它的缺点也是不容忽视的. (ⅰ) 除了利用穷举法构建的突变数据库以外, 利用其他方法构建的突变数据库对突变信息包含得都不够全面, 如公共数据库通常会有意忽略对变异数据的收录, 而样本特异性数据库为了减小搜索空间通常也会去除低表达的蛋白质; (ⅱ) 序列数据搜索中, 当图谱中的碎裂信息不够完整、信噪比较低时, 搜索引擎就不能将候选肽段正确地区分开[45], 因而会增加假阳性的概率. 为了避免序列数据库的上述缺点, 提出了其他鉴定突变肽段的方法, 如序列标签算法、图谱库搜索算法等.

1.2 基于序列标签的氨基酸突变鉴定算法

相比序列数据库搜索算法利用肽段母离子质量从数据库中筛选候选肽段, 序列标签算法利用 *de novo* 测序算法推导的 PSTs 能够更有效地过滤数据库, 减少候选肽段的数目以缩小搜索空间, 使得更复杂和计算更密集的方法能够应用到对候选肽段的突变打分算法中[45], 从而提高了突变鉴定结果的灵敏性并且减少了结果中的假阳性率. 下面从序列标签搜索算法与 *de novo* 测序算法之间的关系以及当前结合 PSTs 进行氨基酸突变鉴定的主流工具两个方面对序列标签算法鉴定突变氨基酸进行介绍.

(1) 序列标签搜索算法与 *de novo* 测序算法. 相比序列数据库搜索算法, *de novo* 算法在对质谱图进行氨基酸序列推导时不依赖蛋白质数据库, 因此它在鉴定氨基酸突变方面有独特的优势[45~47]. 当前使用 *de novo* 测序算法的代表性工具有 SHERENGA[48], PEAKS[49~51]以及 PepNovo[52]等. 这些工具所使用的算法都是通过生成前缀残基质量图谱(prefix residue mass spectra)重构整个图谱进行肽段序列推导的, 因此这些算法对质谱图的质量具有较高的要求[45]. 但通过诱导碰撞解离(collision-induced dissociation, CID)产生的串联图谱中不可避免地含有不完整的碎裂离子系列、噪音离子和精度较差的碎裂离子质量, 这使得 *de novo* 算法常常产生一些不确定的序列区域, 导致 *de novo* 算法通常只能准确地推导出肽段序列中的部分序列[46]. 因此, 结合 *de novo* 算法鉴定的部分肽段序列进行数据库搜索的序列标签算法应运而生, 这种算法不仅可以利用 *de novo* 推导出的 PSTs 作为筛选候选肽段时的过滤指标, 有效地减少搜索空间, 而且可以通过改变 PSTs 与候选肽段匹配的打分算法, 提高对突变肽段的鉴定效率.

(2) 结合肽序列标签的氨基酸突变鉴定算法. 最早结合 PSTs 进行数据库搜索的方法是由 Mann 和 Wilm[53]在 1994 年提出的, 此方法不仅能有效地对常规图谱进行鉴定, 而且能够鉴定出带有突变或修饰图谱的肽段序列. 当前结合肽序列标签对氨基酸突变进行鉴定的算法或程序有 GutenTag 程序[54], Opensea 工具[55], SPIDER 程序[56,57], InsPecT 搜索引擎[45], DirecTag 算法[58]以及 MoDa 算法[59]等. 鉴定突变氨基酸常用的序列标签软件及其网址见表 3.

GutenTag 是由 Yates 实验室开发出来的能够自动推导+2 电荷母离子串联图谱 PSTs 用于数据库搜索的算法, 其特点是利用碎片离子峰强度经验模型并结合相邻氨基酸和碎片离子的相对质量对肽段碎裂的影响推导 PSTs, 之后用多个 PSTs 进行搜库, 同时放宽对 PST 两端质量匹配的限制, 从而能够有效地进行突变肽段的鉴定. 但由于 GutenTag 算法没有考虑同源突变或修饰, 所以此算法只能对数据库中已存在的突变序列进行鉴定, 并且由于在打分方面存在漏洞[55], 所以鉴定出来的结果中存在较高的假阳性.

在 GutenTag 算法发表后的第 2 年, Searle 等人[55]首次将序列标签算法的思想应用于非限制翻译后修饰, 并提出了基于质量的序列比对算法工具 Opensea. Opensea 的特点是利用基于质量的宽度优先的算法(“breadth-first search” algorithm)鉴定出突变位点或修

表 3 鉴定氨基酸突变常用的序列标签软件及其下载地址

软件名称	网址	参考文献
GutenTag	http://fields.scripps.edu/downloads.php	[54]
SPIDER	http://www.bioinfor.com/peaks/features/spider.html	[56,57]
DirecTag+TagRecon+IDPicker	http://fenchurch.mc.vanderbilt.edu./software.php	[47,58,60]
InsPecT	http://proteomics.ucsd.edu/Software.html	[45]

饰位点. 但宽度优先的算法是一种贪婪的匹配算法, 并且在 Opensea 中没有考虑在一个位点上同时存在 *de novo* 的测序错误和同源突变的情况, 所以它不能保证最终结果的可靠性. SPIDER 方法与 Opensea 工具有相似的序列标签算法思想, 但与 Opensea 不同的是, 它能够在一个位置上同时考虑 *de novo* 的测序错误和同源突变的情况, 并且利用动态规划算法进行比对打分. SPIDER 算法已被整合进 PEAKS 软件中, 专门用来对突变肽段和跨物种的同源性肽段进行鉴定.

在 GutenTag 算法推出后, Pevzner 实验室迅速推出了 InsPecT 序列标签算法搜索引擎[45], 它是最早实现规模化鉴定翻译后修饰肽段的搜索工具, 现在仍然被广泛使用. InsPect 搜索引擎推导 PSTs 的算法的特点在于利用改进的 *de novo* 算法推导出 PSTs 作为过滤器缩小候选肽段的范围, 并利用树状快速搜索方法(fast tree-based search)找出与 PSTs 匹配的候选肽段, 用基于动态规划算法(dynamic programming)的图谱比对方法鉴定修饰肽段, 并在打分算法中考虑肽段的碎裂模式. 在推导 PSTs 时, InsPecT 需要构建前缀残基质量图, 而 DirecTag 算法则是直接利用串联图谱的质核比值和峰强度信息对可能的标签进行打分. 由于 DirecTag 只能用来推导 PSTs, 因此其团队后续开发了 TagRecon 算法[47]并将 DirecTag, TagRecon 和 IDPicker 工具[60]整合成鉴定突变和修饰肽段的流程, 其大致过程为: (ⅰ) 利用 DirecTag 生成 PSTs; (ⅱ) TagRecon 利用 PSTs 对常规数据库进行候选肽段过滤, 并且定位数据集中的突变或修饰肽段; (iii) 利用 IDPicker 工具对鉴定结果进行质量控制并且装配成蛋白. 此流程算法在 2013 年由 Abraham 等人[19]在鉴定胡杨树(*Populus*)单氨基酸多态性的实验中被成功地使用.

目前序列标签算法都依赖于 *de novo* 测序构建 PSTs, 但是由 *de novo* 算法测出的肽片段往往存在部分构建错误的序列[56]. MoDa 算法[59]在搜索候选肽段时, 由于采用序列标签链算法(tag chain algorithm)[61], 能有效地避免由 *de novo* 测序引起的错误匹配. 在 MoDa 算法中, 将序列标签算法和动态规划算法结合, 同时利用多条序列标签与候选肽段进行比对, 找出存在质量差的位点, 然后利用基于动态规划算法的图谱比对算法找出最佳的肽段序列. 此方法能够大规模地鉴定出存在多个修饰位点或突变位点的肽段.

(3) 基于序列标签的氨基酸突变鉴定算法面临的问题. 基于肽段序列标签的氨基酸突变序列鉴定算法虽然能够有效地利用 PSTs 过滤数据库, 弥补 *de novo* 测序算法的测序错误并且提高对突变或修饰肽段鉴定的效率和准确性, 但目前已有的 PSTs 算法仍然存在着许多不足, 如在 GutenTag 算法中没有考虑同源突变或修饰, 所以不能鉴定出数据库中不存在的突变序列, 而在 Opensea 软件中没有考虑到突变位点的出现可能是由 *de novo* 的测序错误引起的等. 但是图谱质量是限制序列标签算法的主要因素, 因为低能 CID 碎裂模式通常很难将质量相同或相近的碎裂离子区分开来, 如亮氨酸(Leu)和异亮氨酸(Ile)、赖氨酸(Lys)和谷氨酰胺(Gln)以及苯丙氨酸(Phe)和氧化的甲硫氨酸(Met)等[46]. 近年来, 随着电子转移解离(electron transfer dissociation, ETD)和高能碰撞解离(high-energy collision induced dissociation, HCD)的出现, 越来越多的比 CID 质谱图质量高的、含有丰富的碎裂离子信息的高精度质谱图被产出, 这些高精度的质谱图能更好地适用于序列标签算法, 提高其准确性.

1.3 基于图谱库搜索的氨基酸突变鉴定算法

在肽段鉴定领域, 图谱库搜索是一种有望取代序列数据库搜索的鉴定策略[62]. 相比序列数据库搜索策略, 图谱库搜索策略有以下优点: (ⅰ) 直接利用图谱库中每一张真实图谱的各种不同的特征信息进行比对, 如碎片离子峰的峰强度信息、碎裂模式等, 使图谱比对算法具有更高的灵敏性; (ⅱ) 能够在一个更小、更精确的搜索空间内进行搜索, 可以比序列搜索速度快好几个数量级; (iii) 能够轻松地鉴定出图谱库中已存在的变异肽段[63]. 对于依赖于图谱库搜索的蛋白突变鉴定来讲, 目前最大的限制来源于图谱库的覆盖范围, 尤其是对突变和修饰肽段图谱的包含[63,64]. 由于在相似的条件下, 肽段的图谱具有可再生性[65]并且相似序列的肽段通常能够产生相似的质谱图[20,66], 因此一批利用图谱库中已收录的肽段图谱来扩大图谱库对肽段的覆盖范围, 以实现对氨基酸突变进行鉴定的算法或工具应运而生. 目前常用的图谱搜索软件及其网址见表 4.

在蛋白质组学中, 图谱库搜索概念早在 1998 年就由 Yates 等人[70]率先提出, 但由于质谱仪通量不高、生物质谱数据缺乏以及质谱数据的自动化分析

表 4 常用的图谱搜索软件

工具名称	网址	参考文献
SpectraST	http://www.systemsbiology.org/spectrast	[64,67]
X!Hunter	http://xhunter.thegpm.org/tandem/thegpm_hunter.html	[68]
pMatch	http://pfind.ict.ac.cn/pmatch/	[63]
DeltAMT	http://pfind.ict.ac.cn/pcluster/	[69]
Pepitome	http://fenchurch.mc.vanderbilt.edu/software.php	[62]

方法不完善等[71]原因使得图谱库搜索策略发展缓慢. 直到最近 10 年, 随着质谱和计算机技术的快速发展, 鉴定出的肽段图谱匹配对(peptide spectrum match, PSM)的数目与日俱增, 图谱库搜索策略才逐渐被应用到大规模数据集和数据库中. 最近, 图谱库搜索策略更是被用于发掘样品中的突变肽段. 要用图谱搜索策略来鉴定样品中的突变肽段, 就必须要扩大图谱库对突变肽段的覆盖范围. 目前用于扩大图谱库覆盖范围的算法有 pMatch[63]、半经验算法[72,73]以及 Ji 等人[66]提出的相似性算法等. pMatch 在构建图谱时, 利用肽段已知的实验图谱和理论图谱混合构建图谱, 用来缓冲由修饰或突变氨基酸残基引起的肽段碎裂模式的变化[64]. 由 Hu 等人[72]在 2011 年提出的半经验方法通过利用图谱库中已收录的 PSMs 构建突变肽段的质谱图以扩大对突变肽段的覆盖范围. 这种算法把图谱库中图谱对应的肽段序列替换为相应的突变肽段序列, 并将突变肽段的碎裂离子的质核比值替换到图谱中. 2013 年, Ji 等人[66]提出的相似性算法通过利用相似序列肽段的图谱来推断目标肽段的图谱, 以达到扩充图谱库的覆盖范围的目的. 这种算法的特点是, 通过加权 K 邻近相似算法[66](weighted K-nearest neighbor method)和支持向量机(support vector machine, SVM)[74], 利用与目标肽段序列相似且长度相等的肽段的图谱来精确地预测目标肽段序列的优势碎裂离子(如 b, y 离子类型以及其中性丢失离子类型等)的峰强度, 并且利用 SpectraST[64,67]创建的模型构建诱饵数据库进行数据过滤. 同时, Ji 等人[66]指出, 将此算法应用于扩建美国国家标准与技术研究院(National Institute of Standards and Technology, NIST)图谱数据库, 能有效地将 NIST 图谱库的覆盖率提高 20%~60%, 并且用此数据库能够鉴定到样品中更多的突变肽段.

除了通过扩大图谱库覆盖范围以提高图谱库搜索对样品突变肽段的鉴定率以外, 通过改善图谱-图谱匹配(spectrum-spectrum match, SSM)的打分算法也是一条有效提高突变肽段鉴定效率的途径. 目前点积法是 SSMs 打分的主流算法, 如 SpectraST 和 X!Hunter[68]等主流的图谱搜索工具都是利用点积算法进行匹配打分的. 近年来, 一些基于点积法、用于搜索变异肽段的图谱库搜索工具或算法也逐渐被开发出来, 如 pMatch 工具[63], Bonanza 算法[20]等. pMatch 工具的特点在于, 利用电荷依赖型的质量位移进行离子峰匹配, 并且将常规的点积法与基于概率的模型相结合对图谱间的匹配进行打分. Bonanza 算法特点在于, 在筛选候选图谱时, 不限制母离子质量, 可以将不同母离子质量的图谱聚在一起作为候选图谱; 在对离子峰进行匹配时, 不仅将质量相近的子离子峰考虑进去, 还将母离子间的质量差考虑进去; 最后, 利用改进的点积法进行打分. 最近, 考虑到点积法不能提供一个清晰的统计学上的解释并且在打分中忽略了碎裂离子质核比值的差异等缺点, Dasari 等人[62]构建了一个利用概率评分标准对 SSMs 的质量进行评估的搜索引擎 Pepitome, 并且在错误发现率(false discovery rate, FDR)为 2%的情况下, 成功地鉴定到比 SpectraST 多 10%~12%的肽段数目. 前面所提到的图谱鉴定方法都需要依赖图谱库, Fu 等人[69]在 2011 年提出了一个不需要搜索图谱库就能直接对突变肽段进行鉴定的统计学算法 DeltAMT, 此算法通过二维高斯混合模型利用高精确度的母离子质量差和保留时间信息对变异肽段进行鉴定.

总体而言, 基于图谱库的蛋白质突变鉴定算法能够有效地缩小搜索空间, 降低搜索时间, 提高搜索的灵敏度. 目前, 由于存在谱图搜索软件对质谱图的整体解析度无法达到传统数据库搜索策略的程度以及谱库的覆盖范围小等原因, 谱图搜索更多的是作为传统数据库搜索策略的互补策略被使用. 但是随着算法的改进以及 PeptideAtlas[75]计划的进行[62], 相信在不久的将来, 利用图谱库对串联质谱进行鉴定的方法会越来越广泛地被使用.

2 氨基酸突变鉴定的应用

当调节细胞增殖、分化、死亡的蛋白序列突变累积到一定程度就会引起癌变[76]. DNA 测序显示, 在复杂的癌症基因组中通常包含 40~100 个可能的氨基酸突变位点[25], 然而这些突变中只有小部分会对癌症与肿瘤的发生产生作用. 因此如果能够鉴定出与癌症或肿瘤发生有关的突变肽段, 进而对能够真正引发癌症的基因进行重注释就有机会从更深的层次上了解癌症或肿瘤病发的机理, 找寻到治疗癌症或肿瘤的新方法. 所以, 提高图谱的解析率, 鉴定出更多的高质量的突变图谱是找寻突变肽段的关键.

受限于质谱数据的质量、计算能力以及当前已知 SAVs 的覆盖范围等因素, SAVs 鉴定首先应用在小规模的样品数据集中. 2000 年, Gatlin 等人[27]通过动态构建人类血红蛋白变异数据库首次成功地对人类血红蛋白样品进行了突变鉴定. 随后, 2003 年 Tabb 等人[54]利用序列标签算法对 32950 张人类晶状体蛋白质样品(human lens sample)质谱数据中的突变肽段鉴定作出了尝试, 成功地鉴定出 742 条肽段, 其中 134 条与突变有关. 随着科技的进步和算法的优化, SAVs 鉴定逐渐被应用到大规模数据集的鉴定中. 2007 年, Bunger 等人[31]通过搜索结合 dbSNP 数据库的自建蛋白质数据库从 DU4475 乳腺肿瘤细胞样品质谱数据中鉴定出 629 个 nsSNVs. 同时他们指出, 在大规模数据集中, 要想鉴定出高可信的 SAVs, 不仅要依赖鉴定算法还要对假阳性鉴定结果进行过滤, 如通过诱饵数据库去除假阳性鉴定等. Tanner 等人[77]利用 InsPecT 对 1850 万张人类蛋白质样品 HEK293 质谱图进行鉴定, 并结合 PTMfinder[78]算法对鉴定结果进行了假阳性过滤, 从中发现了与 308 个 nsSNVs 有关的肽段. 之后, SAVs 的鉴定被广泛地应用于组织、器官等复杂样品数据集中. 2012 年, Hyatt 和 Pan[18]将 Sipros v2.0 算法应用于鉴定酸性矿坑水(acid mine drainage)环境中的微生物群落蛋白质的突变氨基酸, 在含有 57001 个蛋白的数据库中进行搜索, 鉴定出 1683 张图谱对应的 755 个突变肽段. 同时, Hyatt 和 Pan[18]指出, 氨基酸突变中有些可能来自于氨基酸的修饰作用, 如在鉴定出的频率最高的突变氨基酸中, 谷氨酰胺(Gln)与谷氨酸(Glu)以及天冬酰胺(Asn)与天冬氨酸(Asp)之间都能经过脱氨基作用进行转换. Su 等人[39]利用自定义的人类血浆蛋白质突变数据库, 从 290 个亚洲人血浆样品中鉴定出 2029 个 SAVs, 并挑选出 3 对与糖尿病和肥胖有关的 SAVs 进行了绝对定量分析, 指出表型不仅和 SAVs 的浓度有关, 也和 SAVs 变种的相对率有关系. Mathivanan 等人[25]通过构建人类蛋白质突变数据库(HPMD), 从直肠癌细胞系中鉴定出 2728 个蛋白, 其中有 57 个突变蛋白是首次在直肠癌中被鉴定出来的. 这些新鉴定出来的突变蛋白在发展新的直肠癌生物标志物和研究治疗直肠癌的靶蛋白方面将会发挥巨大的作用.

近年来, 利用 RNA-seq 数据对蛋白质组数据进行鉴定逐渐受到人们的青睐. 2012 年, Wang 等人[41]基于 RNA-Seq 数据构建了蛋白质数据库并对 2 个直肠癌细胞系 SW480 和 RKO 进行了鉴定, 分别鉴定出 18760 和 22623 张质谱图. 这些图谱中共包含 23 条不存在于 dbSNP54 中的变异肽段, 其中鉴定到的 $TP53^{P309S}$ 突变能够增加 SW480 细胞的增殖能力, 并且能够增强对细胞抗癌药物的耐受性; $HSP90AA1^{D393N}$ 突变对致癌蛋白的构象和稳定性有着巨大的影响.

虽然利用质谱数据结合序列数据库搜索是目前主流的蛋白鉴定的策略, 但在传统的数据库搜索中, 即使利用最好的质谱平台和最优的分析软件, 也有相当一部分质谱图不能被解析出来[79,80]. 随着越来越多的 PSMs 被鉴定出来, 人们开始利用质谱图数据库来鉴定突变肽段, 并且成功地鉴定到了比序列数据库搜索更多的 SAVs. 在 FDR=0.0001%的条件下, Hu 等人[72,73]利用 SpectraST 搜索半经验图谱库并结合 PeptideProphet[81]对结果进行检验, 成功地从人类血浆样品中鉴定出了与 SAVs 有关的 2045 条肽段, 而相同条件下, X!Tandem 则只从序列数据库中鉴定出来 623 条与 SAVs 有关的肽段.

3 结语

随着 DNA 测序成本大幅降低, 越来越多个体的基因组序列被鉴定出来[82]. 但即便在知道人类全基因序列信息的情况下, 科学家们对基因型与分子表型间关系的了解也只是冰山一角[83]. 而对分子表型的了解有助于科学家们对人类疾病发生机理的理解, 比如由 RNA、蛋白质以及翻译后修饰数据能够容易地推断出信号通路是否被激活. 虽然目前出现了许多能够预测基因突变对蛋白分子结构及功能影响的软件和在线工具, 如 IntOGen[84], SIFT[85]和 Poly-

Phen-2[86]等, 但这些预测工具只能辅助性地对突变氨基酸进行筛选和排序, 以便减少实验验证的候选者[87]. 而结合了变异蛋白信息的基因信息能够有效地帮助科学家对特定生物学过程的分子途径以及疾病发生的机制等进行理解, 进而增加预防、诊断、治疗疾病的手段[88].

本文从数据库搜索、数据库搜索与 *de novo* 结合的序列标签搜索以及新兴的图谱比对搜索方法 3 个方面对大规模鉴定突变蛋白的方法作出了比较全面的介绍. 目前, 无论哪一种搜索方法都受到离子碎裂模式理解程度的深入、计算能力高低以及数据库覆盖范围大小等因素的限制, 而结合不同搜索方法能够实现不同方法间的互补, 能有效地提高鉴定结果的灵敏度. Dasari 等人[62]发现, 将序列数据库搜索和图谱库搜素结合起来对样品进行搜索能有效地提高搜索结果的覆盖范围, 并且成功地将此方法应用在了对 MMR 细胞系的鉴定中. 相似地, 在 PEAKS 软件中, 将 *de novo* 测序、序列数据库搜索以及同源性搜索等方法整合到一起形成一个工作流程, 结合多个搜索引擎产出高可信的结果, 并且使得鉴定结果对样本数据库的覆盖范围最大化[49~51]. 随着质谱技术的不断发展和新型计算方法的出现, 序列数据库搜索算法和图谱库搜索算法以及 *de novo* 测序算法的不断地改善、提高, 将来会有越多的突变蛋白被鉴定出来, 这些鉴定结果在寻找生物标记物、个性化医疗以及生理病理机制研究等方面将发挥重要的作用.

参考文献

1 Collins F S, Brooks L D, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res, 1998, 8: 1229–1231
2 Frazer K A, Ballinger D G, Cox D R, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature, 2007, 449: 851–861
3 Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res, 2011, 39: e118
4 Nakamura Y. DNA variations in human and medical genetics: 25 years of my experience. J Hum Genet, 2009, 54: 1–8
5 Yin H, Liang Y, Yan Z, et al. Mutation spectrum in human colorectal cancers and potential functional relevance. BMC Med Genet, 2013, 14: 32
6 Martin A, Saathoff M, Kuhn F, et al. A functional ABCC11 allele is essential in the biochemical formation of human axillary odor. J Invest Dermatol, 2010, 130: 529–540
7 Yoshiura K, Kinoshita A, Ishida T, et al. A SNP in the *ABCC11* gene is the determinant of human earwax type. Nat Genet, 2006, 38: 324–330
8 Vogelstein B, Kinzler K W. Cancer genes and the pathways they control. Nat Med, 2004, 10: 789–799
9 Di Fede G, Catania M, Morbin M, et al. A recessive mutation in the *APP* gene with dominant-negative effect on amyloidogenesis. Science, 2009, 323: 1473–1477
10 Driscoll M C. Sickle cell disease. Pediatr Rev, 2007, 28: 259–268
11 McCarthy M I, Abecasis G R, Cardon L R, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Rev Genet, 2008, 9: 356–369
12 Do R, Kathiresan S, Abecasis G R. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet, 2012, 21: R1–R9
13 Hecht M, Bromberg Y, Rost B. News from the protein mutability landscape. J Mol Biol, 2013, 425: 3937–3948
14 Sheynkman G M, Shortreed M R, Frey B L, et al. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. J Proteome Res, 2014, 13: 228–240
15 Apweiler R, Bairoch A, Wu C H, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res, 2004, 32: D115–D119
16 Edwards N J. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol, 2007, 3: 102
17 Schandorff S, Olsen J V, Bunkenborg J, et al. A mass spectrometry-friendly database for cSNP identification. Nat Methods, 2007, 4: 465–466
18 Hyatt D, Pan C. Exhaustive database searching for amino acid mutations in proteomes. Bioinformatics, 2012, 28: 1895–1901
19 Abraham P, Adams R M, Tuskan G A, et al. Moving away from the reference genome: evaluating a peptide sequencing tagging approach for single amino acid polymorphism identifications in the genus *Populus*. J Proteome Res, 2013, 12: 3642–3651

20 Falkner J A, Falkner J W, Yocum A K, et al. A spectral clustering approach to MS/MS identification of post-translational modifications. J Proteome Res, 2008, 7: 4614–4622
21 Sherry S T, Ward M H, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res, 2001, 29: 308–311
22 Forbes S A, Tang G, Bindal N, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res, 2010, 38: D652–D657
23 Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, 2005, 33: D514–D517
24 Li J, Su Z, Ma Z Q, et al. A bioinformatics workflow for variant peptide detection in shotgun proteomics. Mol Cell Proteomics, 2011, 10: M110.006536
25 Mathivanan S, Ji H, Tauro B J, et al. Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. J Proteomics, 2012, 76: 141–149
26 Stenson P D, Mort M, Ball E V, et al. The human gene mutation database: 2008 update. Genome Med, 2009, 1: 13
27 Gatlin C L, Eng J K, Cross S T, et al. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. Anal Chem, 2000, 72: 757–763
28 Creasy D M, Cottrell J S. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics, 2002, 2: 1426–1434
29 Craig R, Beavis R C. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 2004, 20: 1466–1467
30 Kersey P J, Duarte J, Williams A, et al. The International Protein Index: an integrated database for proteomics experiments. Proteomics, 2004, 4: 1985–1988
31 Bunger M K, Cargile B J, Sevinsky J R, et al. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. J Proteome Res, 2007, 6: 2331–2340
32 Li J, Duncan D T, Zhang B. CanProVar: a human cancer proteome variation database. Hum Mutat, 2010, 31: 219–228
33 Mottaz A, David F P, Veuthey A L, et al. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics, 2010, 26: 851–852
34 Xi H, Park J, Ding G, et al. SysPIMP: the web-based systematical platform for identifying human disease-related mutated sequences from mass spectrometry. Nucleic Acids Res, 2009, 37: D913–D920
35 Alves G, Ogurtsov A Y, Yu Y K. RAId_DbS: mass-spectrometry based peptide identification web server with knowledge integration. BMC Genomics, 2008, 9: 505
36 Yip Y L, Famiglietti M, Gos A, et al. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat, 2008, 29: 361–366
37 Kawabata T, Ota M, Nishikawa K. The protein mutant database. Nucleic Acids Res, 1999, 27: 355–357
38 Chernobrovkin A L, Mitkevich V A, Popov I A, et al. Identification of single amino acid polymorphisms in MS/MS spectra of peptides. Dokl Biochem Biophys, 2011, 437: 90–93
39 Su Z D, Sun L, Yu D X, et al. Quantitative detection of single amino acid polymorphisms by targeted proteomics. J Mol Cell Biol, 2011, 3: 309–315
40 Evans V C, Barker G, Heesom K J, et al. *De novo* derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods, 2012, 9: 1207–1211
41 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: 1009–1017
42 Sultan M, Schulz M H, Richard H, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008, 321: 956–960
43 Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics, 2013, 29: 3235–3237
44 Woo S, Cha S W, Merrihew G, et al. Proteogenomic database construction driven from large scale RNA-seq data. J Proteome Res, 2014, 13: 21–28
45 Tanner S, Shu H, Frank A, et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem, 2005, 77: 4626–4639
46 Ma B, Johnson R. *De novo* sequencing and homology searching. Mol Cell Proteomics, 2012, 11: O111.014902
47 Dasari S, Chambers M C, Slebos R J, et al. TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res, 2010, 9: 1716–1726
48 Dancik V, Addona T A, Clauser K R, et al. *De novo* peptide sequencing via tandem mass spectrometry. J Comput Biol, 1999, 6: 327–342
49 Ma B, Zhang K, Hendrie C, et al. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. Rapid

Commun Mass Spectrom, 2003, 17: 2337–2342
50 Zhang J, Xin L, Shan B, et al. PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics, 2012, 11: M111.010587
51 Han X, He L, Xin L, et al. PeaksPTM: mass spectrometry-based identification of peptides with unspecified modifications. J Proteome Res, 2011, 10: 2930–2936
52 Frank A, Pevzner P. PepNovo: *de novo* peptide sequencing via probabilistic network modeling. Anal Chem, 2005, 77: 964–973
53 Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem, 1994, 66: 4390–4399
54 Tabb D L, Saraf A, Yates J R 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem, 2003, 75: 6415–6421
55 Searle B C, Dasari S, Turner M, et al. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS *de novo* sequencing results. Anal Chem, 2004, 76: 2220–2230
56 Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with *de novo* sequencing error. J Bioinform Comput Biol, 2005, 3: 697–716
57 Yuen D. SPIDER: reconstructive protein homology search with *de novo* sequencing tags. Master Thesis. Waterloo: University of Waterloo, 2011
58 Tabb D L, Ma Z Q, Martin D B, et al. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. J Proteome Res, 2008, 7: 3838–3846
59 Na S, Bandeira N, Paek E. Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics, 2012, 11: M111.010199
60 Ma Z Q, Dasari S, Chambers M C, et al. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. J Proteome Res, 2009, 8: 3872–3881
61 Na S, Jeong J, Park H, et al. Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. Mol Cell Proteomics, 2008, 7: 2452–2463
62 Dasari S, Chambers M C, Martinez M A, et al. Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. J Proteome Res, 2012, 11: 1686–1695
63 Ye D, Fu Y, Sun R X, et al. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. Bioinformatics, 2010, 26: i399–i406
64 Lam H, Deutsch E W, Eddes J S, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics, 2007, 7: 655–667
65 Hoopmann M R, Moritz R L. Current algorithmic solutions for peptide-based proteomics data generation and identification. Curr Opin Biotechnol, 2013, 24: 31–38
66 Ji C, Arnold R J, Sokoloski K J, et al. Extending the coverage of spectral libraries: a neighbor-based approach to predicting intensities of peptide fragmentation spectra. Proteomics, 2013, 13: 756–765
67 Lam H, Deutsch E W, Eddes J S, et al. Building consensus spectral libraries for peptide identification in proteomics. Nat Methods, 2008, 5: 873–875
68 Craig R, Cortens J C, Fenyo D, et al. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res, 2006, 5: 1843–1849
69 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: M110.000455
70 Yates J R 3rd, Morgan S F, Gatlin C L, et al. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal Chem, 1998, 70: 3557–3565
71 Lam H. Building and searching tandem mass spectral libraries for peptide identification. Mol Cell Proteomics, 2011, 10: R111.008565
72 Hu Y, Li Y, Lam H. A semi-empirical approach for predicting unobserved peptide MS/MS spectra from spectral libraries. Proteomics, 2011, 11: 4702–4711
73 Hu Y, Lam H. Expanding tandem mass spectral libraries of phosphorylated peptides: advances and applications. J Proteome Res, 2013, 12: 5971–5977
74 Joachims T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Dordrecht: Kluwer Academic Publishers, 2002
75 Desiere F, Deutsch E W, King N L, et al. The PeptideAtlas project. Nucleic Acids Res, 2006, 34: D655–D658
76 Davies H, Bignell G R, Cox C, et al. Mutations of the *BRAF* gene in human cancer. Nature, 2002, 417: 949–954

77 Tanner S, Shen Z, Ng J, et al. Improving gene annotation using peptide mass spectrometry. Genome Res, 2007, 17: 231–239
78 Tanner S, Payne S H, Dasari S, et al. Accurate annotation of peptide modifications through unrestrictive database search. J Proteome Res, 2008, 7: 170–181
79 Resing K A, Meyer-Arendt K, Mendoza A M, et al. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal Chem, 2004, 76: 3556–3568
80 Yen C Y, Russell S, Mendoza A M, et al. Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. Anal Chem, 2006, 78: 1071–1084
81 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: 5383–5392
82 Snyder M, Du J, Gerstein M. Personal genome sequencing: current approaches and challenges. Genes Dev, 2010, 24: 423–431
83 Ng P C, Levy S, Huang J, et al. Genetic variation in an individual human exome. PLoS Genet, 2008, 4: e1000160
84 Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods, 2013, 10: 1081–1082
85 Ng P C, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res, 2001, 11: 863–874
86 Adzhubei I A, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods, 2010, 7: 248–249
87 Liu X, Jian X, Boerwinkle E. dbNSFP v2. 0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat, 2013, 34: E2393–E2402
88 Snyder M, Weissman S, Gerstein M. Personal phenotypes to go with personal genomes. Mol Syst Biol, 2009, 5: 273

Bioinformatics Algorithms for Identification of Amino Acid Mutations in Tandem Mass Spectrometry

YU Qing[1,2], WU SongFeng[2,3], MA Jie[2,3], ZHU YunPing[2,3] & SHU KunXian[1]

1 Institute of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;
2 State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing 102206, China;
3 Beijing Institute of Radiation Medicine, Beijing 100850, China

Amino acid mutations can change the structures and functions of proteins, which affect life processes of organisms. Currently, shotgun proteomics based on tandem mass spectrometry is the main strategy for large-scale proteomics research, but the existing interpretation methods of mass spectrometry data focus more on improving the sensitivity of the identification results, and tend to compress the information of amino acid mutations in protein databases. Therefore, how to explore the mutations of identified proteins becomes an important task of proteomics research. The current available methods to identify protein mutations using mass spectrometry can be roughly divided into three categories: sequence database search based methods, sequence tags search based methods and spectral library search based methods. Here, firstly, we describe the three kinds of algorithms to identify protein mutations, and then the characteristics and limitations of each method were evaluated in details. Finally, we introduce the research status and development direction of the identification of amino acid mutations by mass spectrometry strategy. With the continuous development of tandem mass spectrometry-based proteomics, the mutations in the protein sequence information would be better interpreted, and it would be helpful for in-depth studies about the diseases caused by protein mutations.

amino acid variation, tandem mass spectrometry, shotgun proteomics, identification of amino acid mutations

doi: 10.1360/N052014-00099