PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2013, 40(4): 309~318
www.pibb.ac.cn

非限制翻译后修饰鉴定方法的研究进展 *

张成普 李 宁 马 洁 吴松锋 朱云平 **
(军事医学科学院放射与辐射医学研究所，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)

摘要 蛋白质翻译后修饰在真核生物细胞内广泛存在，对蛋白质的结构和功能有着十分重要的影响．串联质谱技术的快速发展为翻译后修饰鉴定提供了高通量、高灵敏度和高分辨率的分析平台，但传统搜索引擎鉴定修饰的方法无法满足数据分析的需求，非限制翻译后修饰鉴定已成为目前蛋白质组修饰分析的重要手段之一．非限制翻译后修饰鉴定不需要在分析前指定修饰类型，可以直接从样品中找出大量已知或未知的修饰，对提高质谱图谱解析率以及揭示蛋白质的生物学功能具有十分重要的意义．本文首先介绍了非限制翻译后修饰鉴定的定义和发展历程，然后从序列匹配和谱图匹配两个方面详细综述了目前非限制翻译后修饰鉴定的主流算法，分析了非限制翻译后修饰鉴定的质量控制问题，最后结合非限制翻译后修饰鉴定的实际应用讨论了修饰鉴定算法的不足和发展方向．

关键词 串联质谱，蛋白质翻译后修饰，生物信息学，蛋白质组学
学科分类号 Q51，Q811.4 **DOI**: 10.3724/SP.J.1206.2012.00205

生物质谱技术的发展为蛋白质组学研究提供了高通量、高灵敏度和高分辨率的分析平台，并直接促成了大规模蛋白质组研究的开展．随着人类蛋白质组计划的蓬勃发展，蛋白质组学研究向标准化、定量化、动态化和功能化发展，蛋白质翻译后修饰谱、相互作用网络等研究也逐步深入．蛋白质翻译后修饰(post-translational modification，PTM)在真核生物细胞内广泛存在[1-2]，对其生命活动起到至关重要的作用．多种重要修饰如磷酸化(phosphorylation)、甲基化(methylation)、乙酰化(acetylation)、泛素化(ubiquitination)等，是生物体内信号转导网络不可或缺的机制．因此，采用合理的策略对蛋白质翻译后修饰进行鉴定十分重要，已成为蛋白质组研究的重大挑战之一[3]．

数据库搜索是大规模蛋白质组研究中最常用的质谱数据分析策略，但由于质谱数据的复杂性，通常只有5%～30%谱图得以解析[4]，其中一个重要的原因就是蛋白质翻译后修饰的存在[5-6]．结合抗体富集法、固相金属亲和色谱法、亲和标签标记、亲水相互作用色谱法以及 TiO_2 亲和等修饰富集方法[7-9]，传统的数据库搜索引擎，如 Mascot[10]、SEQUEST[11]、X!Tandem[12]等，能在一定程度上实现修饰肽段的鉴定[13-14]．但是这种常规修饰鉴定策略存在两点不足：首先，必须在搜库前指定修饰的类型和数量，对于未经过特定修饰富集的样品，很难预先确定其中含有哪些修饰，尤其是低丰度修饰，并且该策略无法鉴定新修饰；其次，随着设定可变修饰种类和数量的增加，会出现候选肽段组合爆炸的问题，严重影响数据库搜索的鉴定速度和效率，降低肽段鉴定的灵敏度并且产生大量假阳性鉴定．因此在使用 Mascot 或 SEQUEST 进行数据库

* 国家重点基础研究发展计划(973)(2011CB910601, 2010CB912700)，国家高技术研究发展计划(863)(2012AA020409, 2012AA020201)和国家自然科学基金(21105121)资助项目．
** 通讯联系人．
Tel: 010-80705225, E-mail: zhuyunping@gmail.com
收稿日期：2012-09-08，接受日期：2012-11-28

搜索时设定的可变修饰一般不超过 10 种．有研究表明生物体内每 10 个肽段中就会有一个修饰肽段存在[1]，其中大多数修饰处于亚计量水平．目前记录在常用修饰数据库 Unimod[15]、RESID[16]中修饰类型已将近 1000 种，传统的修饰富集策略结合数据库搜索的鉴定方式远远无法满足实际的需要，因此非限制翻译后修饰鉴定应运而生．

本文首先介绍了非限制翻译后修饰鉴定的定义和发展历程，然后从序列匹配和谱图匹配这两种最基本的修饰鉴定方法出发，详细讨论了目前非限制翻译后修饰鉴定的主流算法，并分析了非限制翻译后修饰鉴定的质量控制问题，最后结合非限制翻译后修饰鉴定的实际应用，讨论了修饰鉴定算法的不足和发展方向．

1　非限制翻译后修饰鉴定概述

非限制翻译后修饰鉴定是指在进行质谱数据分析时，不对修饰的类型和数量进行限制，直接通过序列比对或谱图匹配的策略实现一定质量范围内翻译后修饰的鉴定．该策略可以在短时间内鉴定到样品中大量的修饰类型，包括已经报道的和未知的修饰类型，对提高质谱数据谱图解析率并揭示蛋白质的生物学功能具有十分重要的意义．

非限制翻译后修饰鉴定最早出现于 Pevzner 等 2000 年提出的 spectral alignment 算法[17-18]，这种考虑了修饰质量的图谱比对策略在理论上证明了非限制翻译后修饰鉴定的可行性．基于该算法，2001 年 Pevzner 等开发工具 MS-alignment[5, 18]，并采用实际数据集进行测试，首次实现了翻译后修饰的规模化鉴定．2005 年，随着质谱技术的不断完善，翻译后修饰鉴定引起人们越来越多的关注．结合从头测序(*de novo*)及数据库搜索等肽段鉴定方法[19]，各种应用于非限制翻译后修饰鉴定的算法及软件不断涌现，实现了大规模数据集的非限制翻译后修饰分析[5, 20]．

蛋白质的翻译后修饰表现为氨基酸序列上某些位点发生翻译后的分子质量改变，因此，基于质谱的蛋白质翻译后修饰分析的核心问题是在仪器精度范围内快速准确地找寻修饰类型(质谱数据中体现为某氨基酸上的质量偏差 Δm)，并在此基础上进一步确定修饰肽段和修饰位点(图 1)．当前的非限制翻译后修饰算法主要分为两个方面：一是蛋白质翻译后修饰的鉴定；二是蛋白质翻译后修饰的质量控制．常见的非限制翻译后修饰鉴定软件或算法及其下载地址如表 1 所示．

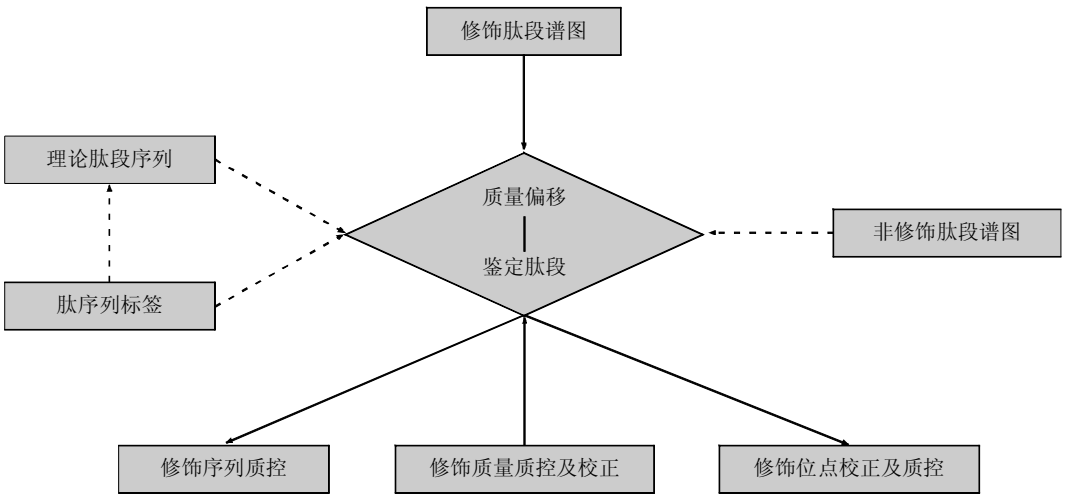

Fig. 1　The process of unrestricted post-translational modifications search

图 1　非限制翻译后修饰鉴定的总体流程

Table 1 The common tools for unrestricted post-translational modifications search

表 1 常用非限制修饰鉴定的算法及软件

类别	工具名称	网址	参考文献
序列匹配	Inspect	http://proteomics.ucsd.edu/Software/	[20]
	MODi	http://prix.uos.ac.kr/research.jsp	[21−22]
	TagRecon	http://fenchurch.mc.vanderbilt.edu/software.php	[23−24]
	SIMS	http://webprod1.ccbr.utoronto.ca	[25]
	MS-alignment	http://proteomics.ucsd.edu	[5]
	P-Mod	http://www.mc.vanderbilt.edu/lieblerlab/p-mod.php	[26]
	TwinPeaks	http://www.utoronto.ca/emililab/twinpeaks.htm	[27]
	SeMoP	http://biomed.umit.at/upload/semop.zip	[28]
	PeaksPTM	http://www-novo.cs.uwaterloo.ca:8080/PeaksPTM/	[29]
	Protein Prospector	http://prospector2.ucsf.edu/prospector/mshome.htm	[30−31]
图谱匹配	Spectral networks	http://proteomics.ucsd.edu/Software/	[32−33]
	pMatch	http://pfind.ict.ac.cn/pmatch/	[34]
	QuickMod	http://javaprotlib.sourceforge.net/	[35]
	Modifiicomb	http://www.bmms.uu.se/Software.htm	[36]
	DeltAMT	http://pfind.ict.ac.cn/pcluster/	[3, 37]
修饰质控	PTMFinder	http://proteomics.ucsd.edu/	[38]
	PTMClust	http://www.psi.toronto.edu/PTMClust/	[39]
	PIE	http://pie.giddingslab.org/	[40]
	IDPicker	http://fenchurch.mc.vanderbilt.edu/software.php	[41]

2 非限制翻译后修饰鉴定的算法

现有的非限制翻译后修饰鉴定算法主要可分为两大类：一类是通过将实验图谱与蛋白质序列理论碎裂图谱比对的方式鉴定翻译后修饰，即序列匹配算法；另一类为通过修饰肽段实验图谱与未修饰肽段实验图谱的比较发现翻译后修饰，即谱图匹配算法.

2.1 基于序列匹配的非限制翻译后修饰鉴定算法

基于序列匹配的非限制翻译后修饰鉴定算法可以进一步细分为两类：一是通过图谱和未修饰肽段理论图谱的匹配进行修饰鉴定，即基于数据库搜索的非限制修饰鉴定策略；二是采用从头测序算法构建肽序列标签(peptide sequence tag，PST)[42]以实现候选肽段筛选的序列匹配策略.

2.1.1 基于数据库搜索策略的非限制翻译后修饰鉴定算法.

作为针对修饰鉴定的数据库搜索策略，这类算法在检索实验谱图的候选肽段时，需要同时考虑与实验谱图母离子质量相同的修饰和非修饰候选肽段. 假定设置的修饰质量范围为−100～300 u，若以 1 u 为步长，则可能存在 401 种不同母离子质量的候选肽段. 过多的候选肽段不仅会带来计算资源的巨大消耗，还会产生很多假阳性结果[43]. 因此，基于数据库搜索策略的非限制修饰鉴定算法会采用多种约束原则，以减少候选修饰肽段的数量. 如 P-Mod[26]以肽段序列中质量最大氨基酸侧链的质量作为上限对修饰质量进行过滤；Interrogator[44]则通过设定单个肽段上只发生一个修饰或者多个修饰位点必须相邻的规则来减少候选修饰肽段的数量；而 MS-alignment 采用 spectral alignment 算法实现候选修饰位点的快速选取，避免了修饰位点的逐一列举. 此外，利用二次搜库策略缩小蛋白质范围[45]，也是这类非限制修饰鉴定算法常用的减少候选肽段数量的方法.

获得候选肽段后，搜索引擎需要采用合理的算法快速准确找寻修饰质量. 2007 年推出的 TwinPeaks 采用类似 SEQUEST 的交叉关联(cross-correlation)算法计算实验谱与候选非修饰肽段理论二级谱的质量偏移，取其中频率最高并与母离子质量差契合的质量偏移作为修饰质量，该思想与 Pevzner 等于 2001 年提出的 MS-convolution 相似，

但具有更强的实用性．与 TwinPeaks 相似，SeMoP 也采用实验谱图和理论肽段二级谱图中与母离子质量差相同的高丰度质量偏移来确定修饰，但 SeMoP 的修饰鉴定算法仅限于找到修饰质量，修饰位点的确认需通过将鉴定的修饰质量设为 SEQUEST 的可变修饰搜库获得．SeMoP 与 SEQUEST 相结合的三步修饰分析策略，即常规搜库获得候选蛋白质、找寻修饰质量、设置可变修饰再次搜库，已成为蛋白质翻译后修饰鉴定中广为使用的方法．

随着质谱鉴定能力的不断提高，修饰鉴定的准确性成为了评价算法性能的重要指标．2009 年由 Chen 等[46]推出的 PTMap 采用未匹配的子离子峰信息评判鉴定结果的准确性，并在打分上考虑了肽段长度的影响，以避免统计分析带来的误差．为保证修饰质量的准确性，PTMap 采用子离子信息矫正母离子质量；为保证修饰位点的准确性，PTMap 要求修饰位点两边的 b、y 离子的打分均高于非修饰肽段该位点两边 b、y 离子的打分．2011 年发布的 PeaksPTM[29]在打分时考虑了修饰肽段与非修饰肽段的关联信息，即如果一个修饰肽段对应的非修饰肽段也被鉴定到，那么该修饰肽段的可信度将被提升，以此保证修饰鉴定的准确性．

除了修饰鉴定的准确性和灵敏度，搜索引擎的跨平台应用性能也是评价算法的重要指标．比如 Protein Prospector[30-31]可以应用于 QSTAR、LTQ 等各类质谱仪器平台，并且可以与常规修饰鉴定结合实现单肽段多修饰位点的非限制翻译后修饰搜索，以提高鉴定结果的准确性．PILOT_PTM[47]则使用 ILD(integer linear optimization)模型[48-49]对一个候选肽段上可能的修饰进行筛选，避免了修饰组合和修饰位点的枚举．PILOT_PTM 不仅可以适用于 Orbitrap-LTQ、Q-TOF 等各种精度等级的质谱仪，还可以应用于电子转移解离(electron transfer dissociation，ETD)、电子捕获解离(electron capture dissociation, ECD)、碰撞诱导解离(collision-induced dissociation，CID)[50]等多种离子碎裂方式，体现了当前质谱分析的发展方向．

随着质谱鉴定能力的不断提高和算法的不断改善，基于非限制翻译后修饰的数据库搜索策略在灵敏度、准确性等方面均有了明显的提升．但由于缺少有效的候选肽段过滤机制，这类算法始终难以在大规模的数据集和数据库上取得广泛的应用[51]．

2.1.2 基于肽序列标签的非限制翻译后修饰鉴定算法．

基于肽序列标签的非限制修饰鉴定策略通过从头测序算法(*de novo*)构建肽序列标签筛选肽段，以减少候选肽段的数量，进而可以采用更为严格复杂的算法实现修饰的鉴定，避免消耗过多的计算资源，提高了鉴定结果的准确性和灵敏度．

OpenSea[52]和 SPIDER[53]为最早将肽序列标签的思想应用于非限制翻译后修饰鉴定的两种算法，它们通过计算 *de novo* 测序序列与数据库序列的质量偏差确定可能的修饰，成功地将从头测序和序列匹配打分相结合并用于修饰鉴定．但对 *de novo* 构建序列标签准确性的过多依赖使得这两种算法在图谱质量不好时很容易产生错误的鉴定结果．因此，后期的肽序列标签算法通常会限定标签的长度，尽可能保证标签构建的准确性，并且采用针对修饰的特异性打分，提高修饰鉴定的准确性．由 Tanner 等[20]于 2005 年推出的搜索引擎 InsPect 采用 *de novo* 测序构建长度为 3～4 个氨基酸的肽序列标签，以树状搜索方式快速检索含有这些标签的肽段，并使用基于动态规划的谱图联配算法(spectral alignment)定位修饰，在图谱匹配打分中采用动态打分算法，同时考虑大质量修饰对肽段长度的影响．InsPect 最早实现了蛋白质翻译后修饰的规模化鉴定，至今仍有着十分广泛的应用．与 InsPect 分析流程类似，Liu 等[54]采用点过程模型(point process model)实现类似 spectral alignment 的功能，以获得含有修饰的候选肽段．为更好地利用多个标签间的关联信息，MODi[21-22]采用肽序列标签链与候选肽段对比寻找质量差以定位修饰的方法，有效地避免了 *de novo* 测序引入的误差．而 MODi 的扩展工具 MODmap[55]则通过引入了类似 Ascore[56]的打分机制，对 MODi 鉴定的高可信结果重新评估以发现新修饰．2011 年，Na 等[51]进一步优化了肽序列标签链算法，将动态规划算法加入修饰鉴定中，不仅实现了单肽段多修饰类型的鉴定，还可以将算法应用于大规模蛋白质数据库的搜索．

与上述典型的应用肽序列标签的算法不同，SIMS[25]与 ByOnic[57]采用了相差一个氨基酸质量的特征峰筛选候选肽段的方法，并根据特征氨基酸两边的剩余质量确定修饰的存在，以避免 *de novo* 过程中产生的误差．但是由于筛选原则减弱，这类算法限定一个肽段上只能有一个修饰，以保证结果的

准确度和灵敏度.

基于肽序列标签的数据库搜索策略有效的候选肽段过滤机制使其在修饰鉴定领域有着相对广泛的应用. 这类算法的关键在于如何通过 *de novo* 准确地获得肽序列标签，因而对谱图质量要求较高. 随着质谱精度和准度的不断提升，尤其是高能碰撞诱导解离(high-energy collision induced dissociation，HCD)等高精度离子碎裂技术的流行[58]，*de novo* 的灵敏度和准确度已经可以接近数据库搜索策略[59]. 因此，应用 HCD 的肽序列标签策略，可以更准确地定位修饰的存在，有着较为广阔的发展空间.

2.2 基于谱图匹配的非限制翻译后修饰鉴定算法

基于谱图匹配的非限制翻译后修饰鉴定算法有效利用了图谱匹配策略的优势，直接将实验谱图进行对比，可以考虑子离子峰强度、离子特性等信息，相比序列数据库搜索可以更加准确地找到一个非修饰肽段对应的修饰肽段. 正因为依赖于实验图谱之间的匹配，这类算法对于谱图的质量和噪声峰过滤等谱图预处理过程有着较高的要求.

基于修饰肽段与非修饰肽段在样品中会同时存在的假定[36]，基于图谱匹配策略的非限制翻译后修饰鉴定算法同样可以进一步分为两类. 第一类算法先通过常规数据库搜索获得具有高可信鉴定结果的谱图并将其构建成图谱库，再对比剩余图谱与库中图谱的相似性，进而找到非修饰肽段所对应的修饰肽段. 这类典型的算法包括 Bonanza[60]、pMatch[34]、QuickMod[35]以及 Modificomb[36]等. 在计算图谱对时，Bonanza 先计算谱图间母离子的质量差，然后在匹配子离子峰时不仅考虑质量相近的子离子峰，还考虑用母离子质量去解释子离子峰对，最后用改进的点积打分找寻最优匹配，并采用正反库策略对结果进行评估. pMatch 的特点在于构建图谱库时考虑序列信息，将序列的理论谱与实验谱图整合用于构建谱图库，可以在一定程度上降低算法对谱图质量的要求. QuickMod 采用支持向量机(support vector machine, SVM)来计算各打分要素间的权重，并利用图谱间子离子峰的匹配信息与 Unimod 数据库信息快速定位修饰位点，避免修饰位点的穷举，其相对较快的运行速度使其在血浆蛋白质组的大规模数据中取得了应用. Modificomb 则利用修饰肽段与非修饰肽段的色谱保留时间存在规律性相似的原理，采用母离子质量差(Δm)和色谱保留时间(retention time)相结合的方法分析修饰的存在，以提高鉴定的准确性.

第二类基于图谱匹配策略的非限制翻译后修饰鉴定算法不需要事先搜库获得肽段与谱图匹配的信息，直接从实验谱图中搜索修饰肽段与非修饰肽段的图谱对以获得修饰信息. 由 Pevzner 团队推出的 Spectral networks[32-33]是这类图谱匹配策略的典型代表. 该算法采用 Spectral alignment 寻找谱图对构建成谱图网络，然后统计网络中谱图的质量差获得可能的修饰信息，再通过 *de novo* 测序构建网络中的核心序列，进而与数据库序列对比获得修饰信息. 该算法具有两个优势：一是无需蛋白质数据库信息就可以获得修饰质量；二是结合了网络中多个谱图的信息，可以更准确地鉴定出修饰肽段. 2011 年由 Fu 等推出的 DeltAMT[3, 37]简化了这一过程，通过直接提取谱图中母离子质量差 Δm 和色谱保留时间差 ΔRT 的信息，采用贝叶斯高斯混合模型区分修饰谱图对和随机谱图对，进而发现修饰的存在. 由于未使用二级谱图信息，DeltAMT 不受二级谱图质量的影响，能快速准确地鉴定出高可信的修饰质量，其不足之处在于只能鉴定高丰度修饰，且无法独自判定修饰位点.

总体而言，基于图谱匹配的非限制翻译后修饰算法得益于质谱技术精度的不断提高，可以相对准确地确定一个非修饰肽段所对应的修饰肽段，从而有效地提高图谱解析率. 这类算法的不足之处在于：一是不能通过翻译后修饰的鉴定提升肽段的鉴定数量，进而也无法提升蛋白质的鉴定数量；二是对于那些能较大程度改变图谱形态的修饰，比如富含中性丢失的磷酸化修饰等，这类算法易产生错误的鉴定. 但随着质谱技术的不断发展，尤其是 ETD、HCD 等新离子碎裂技术的广泛应用，修饰的中性丢失问题得到了一定的改善[50, 58]，也为基于图谱匹配的非限制翻译后修饰算法提供了进一步发展的空间.

3 非限制翻译后修饰鉴定的质控

同肽段鉴定一样，翻译后修饰鉴定也需要在保证结果可靠性的基础上有效地区分正确 / 错误的鉴定[61]. 常规肽段质控所使用的基于“target-decoy”策略计算结果错误发现率(false discovery rate，FDR)的方法[62]，未考虑修饰类型和修饰位点的假阳性，会低估修饰鉴定的 FDR，无法保证结果的可靠性[38]. 在肽段鉴定中存在从图谱、肽段到蛋白质鉴定可靠性不断降低的现象，以此类推，翻译后修饰鉴定需要较肽段鉴定更严格的质控[63]. 但过于

严格的质控标准会降低结果的灵敏度，不利于低丰度修饰的鉴定．因此，非限制修饰质量控制算法需要在从修饰肽段、修饰类型、修饰位点三个方面进行，并保证结果的灵敏度．

高可信的修饰鉴定首先来源于高可信的肽段鉴定，修饰肽段的质控一般沿用常规肽段鉴定质控方法．现有的非限制翻译后修饰鉴定及质控方法对修饰肽段鉴定结果过滤或通过一个分类器区别正确和错误的肽段鉴定(如支持向量机 SVM、线性判别函数等)[21, 63-64]，或通过计算修饰肽段与理论肽段的 Xcorr 值[28, 65]，或通过计算未匹配离子峰数目[46]．正伪库策略也被应用到非限制翻译后修饰以及磷酸化蛋白质组鉴定的质控上．另一种应用广泛但较为粗略的修饰肽段质量控制方法是使用多种翻译后修饰鉴定工具进行交叉验证[66-67]．

修饰类型的质控和校正方面，除了 Fu 等利用高斯混合模型计算 Δm 正确的概率外，ModifiComb、PTMap 和 SeMop 都寻找母离子与子离子一致的质量偏差作为正确的修饰类型[28, 36, 46]．MS-alignment 则使用了修饰概率矩阵(PTM frequency matrix)来判定修饰鉴定的准确性．在修饰位点的质控方面，Tanner 等[64]利用贝叶斯理论计算修饰位点 p 值．修饰位点的校正方面，通常认为相同肽段所有位点中肽段图谱匹配(peptide spectrum match，PSM)最好的位点为正确的修饰位点[46, 64]，而相邻位点具有相同的 PSM 打分时可以通过每种修饰位点下 b、y 离子峰加和来确定修饰位点[46]．另外可利用 UniProt、HPRD 等数据库先验信息辅助位点判定[21-22]．

除了非限制修饰鉴定算法本身的质量控制外，近些年还产生了一些专门针对非限制翻译后修饰鉴定结果的质控算法，如 PTMFinder[38]、PTMClust[39]、PIE[40]以及 IDPicker[41]等．这些算法可以整合不同肽段谱图匹配间的修饰信息，更好地提升结果的准确性和灵敏度．PTMFinder 作为最早应用于修饰质控的算法由 Tanner 等[38]于 2008 年推出，该算法采用鉴定为同一个修饰肽段的全部谱图构建一致性谱图并重新搜库，以整合多谱图信息、提高信噪比．PTMFinder 采用支持向量机进行特征筛选和整合，进而获得统一打分对结果进行过滤，有效提高了修饰肽段和修饰类型鉴定的准确性．PTMclust 采用机器学习的方法聚集具有相同修饰质量的肽段，形成修饰集合，并基于每个集合找到每个修饰最有可能的质量、底物肽段以及位点．PIE (protein inference engine)采用马尔科夫蒙特卡洛方法结合模拟退火算法整合 top-down 蛋白信息、bottom-up 肽段信息[68]以及修饰相关的先验信息，给出关于修饰的综合打分，并寻找最优解，即打分最高的匹配结果．IDPicker 则主要针对 TagRecon[23-24]设计，用来对肽段鉴定结果进行过滤卡值并完成蛋白质组装．

修饰质控作为非限制修饰分析的重要组成部分，已引起了越来越多的关注．修饰质控所整合的信息，也从单实验、单搜索引擎的结果逐渐向多实验、多搜索引擎甚至多平台发展．随着各种仪器平台的不断发展以及多搜索引擎整合的普及[69]，像 PIE 这类应用于修饰质控的多信息整合工具，将有着广阔的应用前景．

4 非限制翻译后修饰鉴定的应用

从 2005 年开始，非限制翻译后修饰鉴定开始在各种实际数据集上取得应用．基于序列匹配的非限制翻译后修饰鉴定算法通过修饰类型及其位点的确定，不仅可以更好地揭示蛋白质的分子功能，还可以提升质谱鉴定的谱图解析率，甚至发现新的蛋白质．

受谱图质量、计算时间等因素的影响，非限制翻译后修饰鉴定首先主要应用于小规模的数据集，以保证修饰鉴定的准确性．2005 年, Dekel Tsur 等[52]将 MS-alignment 应用于人晶状体蛋白(human lens proteins)质谱数据集．鉴定结果显示，43 518 张二级谱中有 3 271 张谱图对应修饰肽段，占谱图总数的 7.5%．在鉴定到的修饰类型中，有 10 余种是与晶体蛋白相关的化学修饰，证实了修饰鉴定的可靠性．人晶状体蛋白数据集修饰类型丰富并且研究较为透彻，至今仍是评估修饰鉴定算法常用的数据集．以 MS-alignment 为核心算法的 InsPect 采用了基于肽序列标签的候选肽段过滤机制，可以应用于大规模数据集和数据库中，但必须结合相应的质控算法才能保证结果的准确性．2007 年，Tanner 等[70]将 InsPect 结合 PTMfinder 应用于包含 1 700 万张谱图的 H293 细胞数据集，搜索含有 13 840 条序列的蛋白质数据库，共鉴定了 26 850 个高可信的修饰肽段，其中 77.61%的结果与已知化学修饰或体内修饰相对应．2009 年，Chen 等[71]将 PTMap 应用于酵母组蛋白新修饰的研究中，鉴定了 14 种全新的修饰质量，对于阐明组蛋白分子功能意义重大．2011 年 Xi Han 等[29]使用 PeaksPTM 分析了 LTQ- Orbitrap Velos 质谱仪产出人心脏组织高精度质谱数据，相对 Mascot 和 SEQUEST 结果而言，

采用了非限制翻译后修饰鉴定算法的 PeaksPTM 在修饰肽段的灵敏度和准确性上均有着显著的优势. 除了应用于专门针对修饰的数据集，非限制翻译后修饰鉴定在实际的蛋白质组数据中也取得了一定的应用. 2007 年，TwinPeaks 被应用于血浆蛋白质组的数据分析，在 3.5%的 FDR 条件下，共得到 539 张常规限制性修饰搜索中未鉴定到的高可信修饰谱图，谱图解析率在原有基础上提升了 17.6%. 2011 年, Surendra Dasari 等[23]将 TagRecon 结合质控工具 IDPicker 应用于毒理蛋白质组学(toxico-proteomics)的研究中，不仅提升了 15%的谱图解析率，还鉴定出可能与药物代谢相关的多种 修饰，如半胱氨酸上的高可信修饰质量(+134 u、+25 u、−2 u).

相对于上述序列匹配算法，谱图匹配算法的应用主要在于修饰本身的高可信鉴定. 比如 ModifiComb 通过分析基于 Mascot 搜库结果的人类唾液质谱数据，不仅找到了多种已知的修饰类型，还找到了如脯氨酸上+12 u 的新修饰类型. DeltAMT 则通过分析 ISB 标准蛋白数据集[72]，快速准确地鉴定到了 32 种高可信的修饰类型.

综合当前的文献，无论是序列匹配算法还是谱图匹配，非限制翻译后修饰鉴定在分析大规模高精度质谱数据方面应用越来越广泛，在鉴定的准确性、灵敏度不断提升的基础上，不断揭示修饰相关的生物学功能.

5 结　　语

蛋白质中广泛存在的翻译后修饰对于其结构和功能有着十分重要的影响. 大规模蛋白质翻译后修饰的准确鉴定和定位是目前蛋白质组重要的研究方向之一. 基于串联质谱的非限制翻译后修饰鉴定可以实现生物样品规模化已知修饰的鉴定和未知修饰的寻找，进而提高谱图的解析率和蛋白质的鉴定数量，并在修饰层面上揭示生物学功能. 目前非限制翻译后修饰鉴定方法的研究已经成为质谱数据分析领域的热点之一，相关的算法与工具也不断涌现，并朝着快速化、准确化、高通量化方向发展. 随着质谱仪器各方面性能的不断改进以及 ETD、HCD 等裂解技术的广泛使用，串联质谱数据不仅在二级谱中有更高的精度，还有更为丰富的修饰碎片信息，令非限制翻译后修饰鉴定有更高的灵敏度和准确性，有能力应用于大规模数据集的修饰鉴定.

随着算法的不断完善，未来非限制翻译后修饰鉴定不仅可以用于常规的修饰分析，还可以在某些特殊方面行使其功能. 比如在非限制修饰分析的基础上，对某一类具有特定质谱行为的修饰进行更精确的搜索. 目前已经有一些算法针对氨基酸突变或者糖基化修饰[73]进行设计，并取得了一定的成果. 而像 FAT10[74]这种会在质谱中发生碎裂的大质量修饰，不适合采用常规的修饰分析手段进行鉴定，改进的非限制修饰鉴定方法则可以解决这一问题. 另外将非限制修饰鉴定与现有的有标、无标定量[75]以及多重反应监测(multiple reaction monitoring，MRM)[76]技术相结合，实现修饰定量，进行基于修饰的网络分析，可以更好地揭示修饰及其对应底物蛋白质的生物学功能.

随着计算机处理能力的不断提高，运行时间已不再是制约算法发展的主要瓶颈，非限制翻译后修饰鉴定面临的问题主要有两个方面：一是缺少大规模的修饰数据集用于算法评估，导致对算法的评价缺乏统一的标准，降低了算法或工具的通用性；二是受到图谱质量、低丰度修饰、不稳定修饰的影响，以及缺少针对修饰鉴定错误发现率的通用评价指标，非限制翻译后修饰鉴定在准确性上仍存在一定的不足，尤其是修饰位点的判定，往往需要人工确认谱图匹配的正确性，导致算法的实用性降低. 目前较为常用的质谱数据汇集中心如 PeptideAtlas (http://www.peptideatlas.org/repository/)、Proteome Commons (https://proteomecommons.org/)等已提供了较为丰富修饰相关串联质谱数据集，合理地整合这些数据并建立针对修饰鉴定准确性的评价标准，将成为推动非限制修饰鉴定算法发展的重要动力来源.

参 考 文 献

[1] Nielsen M L, Savitski M M, Zubarev R A. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. Mol Cell Proteomics, 2006, **5**(12): 2384−2391

[2] Wold F. *In vivo* chemical modification of proteins (post-translational modification). Annu Rev Biochem, 1981, **50**: 783−814

[3] 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

[4] Ahrné E, Müller M, Lisacek F. Unrestricted identification of modified proteins using MS/MS. Proteomics, 2009, **9999** (999A): NA

[5] Tsur D, Tanner S, Zandi E, *et al*. Identification of post-translational modifications by blind search of mass spectra. Nat Biotechnol, 2005, **23**(12): 1562−1567

[6] 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

[7] Shi Y, Xu P, Qin J. Ubiquitinated proteome: ready for global?. Mol Cell Proteomics, 2011, **10**(5): R110 006882

[8] Tissot B, North S J, Ceroni A, *et al*. Glycoproteomics: past, present and future. FEBS Lett, 2009, **583**(11): 1728-1735

[9] Pinkse M W, Mohammed S, Gouw J W, *et al*. Highly robust, automated, and sensitive online TiO_2-based phosphoproteomics applied to study endogenous phosphorylation in *Drosophila* melanogaster. J Proteome Res, 2008, **7**(2): 687-697

[10] 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

[11] Jimmy K E, Ashley L M, John R Y. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. American Society for Mass Spectrometry, 1994, **5**(11): 976-989

[12] Craig R, Beavis R C. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 2004, **20**(9): 1466-1467

[13] 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

[14] 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

[15] Creasy D M, Cottrell J S. Unimod: Protein modifications for mass spectrometry. Proteomics, 2004, **4**(6): 1534-1536

[16] Garavelli J S. The RESID database of protein modifications as a resource and annotation tool. Proteomics, 2004, **4**(6): 1527-1533

[17] Pevzner P A, Dancik V, Tang C L. Mutation-tolerant protein identification by mass spectrometry. J Comput Biol, 2000, **7**(6): 777-787

[18] Pevzner P A, Mulyukov Z, Dancik V, *et al*. Efficiency of database search for identification of mutated and modified proteins *via* mass spectrometry. Genome Res, 2001, **11**(2): 290-299

[19] 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

[20] Tanner S, Shu H, Frank A, *et al*. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem, 2005, **77**(14): 4626-4639

[21] 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**(12): 2452-2463

[22] Kim S, Na S, Sim J W, *et al*. MODi: a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra. Nucleic Acids Res, 2006, **34**(Web Server issue): W258-263

[23] Dasari S, Chambers M C, Codreanu S G, *et al*. Sequence tagging reveals unexpected modifications in toxicoproteomics. Chem Res Toxicol, 2011, **24**(2): 204-216

[24] Dasari S, Chambers M C, Slebos R J, *et al*. TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res, 2010, **9**(4): 1716-1726

[25] Liu J, Erassov A, Halina P, *et al*. Sequential interval motif search: unrestricted database surveys of global MS/MS data sets for detection of putative post-translational modifications. Anal Chem, 2008, **80**(20): 7846-7854

[26] Hansen B T, Davey S W, Ham A J, *et al*. P-Mod: an algorithm and software to map modifications to peptide sequences using tandem MS data. J Proteome Res, 2005, **4**(2): 358-368

[27] Havilio M, Wool A. Large-scale unrestricted identification of post-translation modifications using tandem mass spectrometry. Anal Chem, 2007, **79**(4): 1362-1368

[28] Baumgartner C, Rejtar T, Kullolli M, *et al*. SeMoP: a new computational strategy for the unrestricted search for modified peptides using LC-MS/MS data. J Proteome Res, 2008, **7**(9): 4199-4208

[29] Han X, He L, Xin L, *et al*. PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. J Proteome Res, 2011, **10**(7): 2930-2936

[30] Chalkley R J, Baker P R, Medzihradszky K F, *et al*. In-depth analysis of tandem mass spectrometry data from disparate instrument types. Mol Cell Proteomics, 2008, **7**(12): 2386-2398

[31] Chalkley R J, Baker P R, Huang L, *et al*. Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: Ⅱ. New developments in protein prospector allow for reliable and comprehensive automatic analysis of large datasets. Mol Cell Proteomics, 2005, **4**(8): 1194-1204

[32] Bandeira N, Tsur D, Frank A, *et al*. Protein identification by spectral networks analysis. Proc Natl Acad Sci USA, 2007, **104**(15): 6140-6145

[33] Bandeira N. Spectral networks: a new approach to *de novo* discovery of protein sequences and posttranslational modifications. Biotechniques, 2007, **42**(6): 687, 689, 691 passim

[34] 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**(12): i399-406

[35] Ahrne E, Nikitin F, Lisacek F, *et al*. QuickMod: A tool for open modification spectrum library searches. J Proteome Res, 2011, **10**(7): 2913-2921

[36] Savitski M M, Nielsen M L, Zubarev R A. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol Cell Proteomics, 2006, **5**(5): 935-948

[37] Fu Y, Jia W, Lu Z, *et al*. Efficient discovery of abundant post-translational modifications and spectral pairs using peptide mass

and retention time differences. BMC Bioinformatics, 2009, **10** (Suppl 1): S50

[38] Tanner S, Payne S H, Dasari S, *et al*. Accurate annotation of peptide modifications through unrestrictive database search. J Proteome Res, 2008, **7**(1): 170−181

[39] Chung C, Liu J, Emili A, *et al*. Computational refinement of post-translational modifications predicted from tandem mass spectrometry. Bioinformatics, 2011, **27**(6): 797−806

[40] Jefferys S R, Giddings M C. Baking a mass-spectrometry data PIE with McMC and simulated annealing: predicting protein post-translational modifications from integrated top-down and bottom-up data. Bioinformatics, 2011, **27**(6): 844−852

[41] 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**(8): 3872−3881

[42] Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem, 1994, **66**(24): 4390−4399

[43] Ahrne E, Muller M, Lisacek F. Unrestricted identification of modified proteins using MS/MS. Proteomics, 2010, **10**(4): 671−686

[44] Tang W H, Halpern B R, Shilov I V, *et al*. Discovering known and unanticipated protein modifications using MS/MS database searching. Anal Chem, 2005, **77**(13): 3931−3946

[45] Craig R, Beavis R C. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom, 2003, **17**(20): 2310−2316

[46] Chen Y, Chen W, Cobb M H, *et al*. PTMap——a sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. Proc Natl Acad Sci USA, 2009, **106**(3): 761−766

[47] Baliban R C, DiMaggio P A, Plazas-Mayorca M D, *et al*. A novel approach for untargeted post-translational modification identification using integer linear optimization and tandem mass spectrometry. Mol Cell Proteomics, 2010, **9**(5): 764−779

[48] DiMaggio P A Jr, Floudas C A, Lu B, *et al*. A hybrid method for peptide identification using integer linear optimization, local database search, and quadrupole time-of-flight or OrbiTrap tandem mass spectrometry. J Proteome Res, 2008, **7**(4): 1584−1593

[49] DiMaggio P A Jr, Floudas C A. *De novo* peptide identification *via* tandem mass spectrometry and integer linear optimization. Anal Chem, 2007, **79**(4): 1433−1446

[50] Syka J E, Coon J J, Schroeder M J, *et al*. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci USA, 2004, **101**(26): 9528−9533

[51] Na S, Bandeira N, Paek E. Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics, 2011, **11**(4): M111010199

[52] Searle B C, Dasari S, Wilmarth P A, *et al*. Identification of protein modifications using MS/MS *de novo* sequencing and the OpenSea alignment algorithm. J Proteome Res, 2005, **4**(2): 546−554

[53] 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**(3): 697−716

[54] Liu C, Yan B, Song Y, *et al*. Peptide sequence tag-based blind identification of post-translational modifications with point process model. Bioinformatics, 2006, **22**(14): e307−313

[55] Paek E, Na S. Prediction of novel modifications by unrestrictive search of tandem mass spectra. J Proteome Res, 2009, **8**(10): 4418−4427

[56] Beausoleil S A, Villen J, Gerber S A, *et al*. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol, 2006, **24**(10): 1285−1292

[57] Bern M, Cai Y, Goldberg D. Lookup peaks: a hybrid of *de novo* sequencing and database search for protein identification by tandem mass spectrometry. Anal Chem, 2007, **79**(4): 1393−1400

[58] Olsen J V, Macek B, Lange O, *et al*. Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods, 2007, **4**(9): 709−712

[59] 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

[60] 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**(11): 4614−4622

[61] Carr S, Aebersold R, Baldwin M, *et al*. The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. Mol Cell Proteomics, 2004, **3**(6): 531−533

[62] Wang G, Wu W W, Zhang Z, *et al*. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. Anal Chem, 2009, **81**(1): 146−159

[63] 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

[64] Tanner S, Payne S H, Dasari S, *et al*. Accurate annotation of peptide modifications through unrestrictive database search. J Proteome Res, 2008, **7**(1): 170−181

[65] Havilio M, Wool A. Large-scale unrestricted identification of post-translation modifications using tandem mass spectrometry. Analytical Chemistry, 2007, **79**(4): 1362−1368

[66] Wilmarth P A, Tanner S, Dasari S, *et al*. Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility?. J Proteome Res, 2006, **5**(10): 2554−2566

[67] Thompson M R, Thompson D K, Hettich R L. Systematic assessment of the benefits and caveats in mining microbial post-translational modifications from shotgun proteomic data: the response of shewanella oneidensis to chromate exposure. J Proteome Res, 2008, **7**(2): 648−658

[68] Resing K A, Ahn N G. Proteomics strategies for protein identification. FEBS Letters, 2005, **579**(4): 885−889

[69] Tharakan R, Edwards N, Graham D R. Data maximization by

multipass analysis of protein mass spectra. Proteomics, 2010, **10**(6): 1160-1171

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

[71] Zhang K, Chen Y, Zhang Z, *et al*. Identification and verification of lysine propionylation and butyrylation in yeast core histones using PTMap software. J Proteome Res, 2009, **8**(2): 900-906

[72] Klimek J, Eddes J S, Hohmann L, *et al*. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. J Proteome Res, 2008, **7**(1): 96-103

[73] Jia W, Lu Z, Fu Y, *et al*. A strategy for precise and large scale identification of core fucosylated glycoproteins. Mol Cell Proteomics, 2009, **8**(5): 913-923

[74] Kalveram B, Schmidtke G, Groettrup M. The ubiquitin-like modifier FAT10 interacts with HDAC6 and localizes to aggresomes under proteasome inhibition. J Cell Sci, 2008, **121**(Pt 24): 4079-4088

[75] Vaudel M, Sickmann A, Martens L. Peptide and protein quantification: a map of the minefield. Proteomics, 2010, **10**(4): 650-670

[76] Anderson L, Hunter C L. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics, 2006, **5**(4): 573-588

The Research and Progress of Unrestricted Post-Translational Modifications Search Based on Tandem Mass Spectrometry*

ZHANG Cheng-Pu, LI Ning, MA Jie, WU Song-Feng, ZHU Yun-Ping**

(*State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing* 102206, China)

Abstract Protein post-translational modifications (PTM) are widespread in eukaryotic cells and have a significant influence on the structure and function of proteins. The rapid development of tandem mass spectrometry (MS/MS) has provided a sensitive and accurate platform in high throughput PTM identification. However, the traditional database search engines could not meet the needs of modification data analysis, which has made the unrestricted modification search become one of the major methods/strategies to identify modifications in proteomics. Without requirement to specify the type of modifications in advance, unrestricted modification search can detect a large number of known and unanticipated modifications from the samples, which is of great significance to improve the identification rate of tandem mass spectra and reveal the biological function of proteins. In this paper, we first described the definition and development of unrestricted modification search. Then we discussed the algorithms of unrestricted PTM identification based on two major approaches, sequence matching and spectral matching, as well as the quality control of modification identification. Finally, we summarized several applications of unrestricted PTM identification, and the challenges and strategies discussed here could benefit the future research.

Key words tandem mass spectrum, protein post-translational modifications, bioinformatics, proteomics

DOI: 10.3724/SP.J.1206.2012.00205

* This work was supported by grants from National Basic Research Program of China (2011CB910601, 2010CB912700), National High Technology Research and Development Program of China (2012AA020409, 2012AA020201) and The National Natural Science Foundation of China (21105121).

**Corresponding author.

Tel: 86-10-80705225, E-mail: zhuyunping@gmail.com

Received: September 8, 2012 Accepted: November 28, 2012