蛋白质组学研究中的质谱鉴定与生物信息学分析

刘 康，高起飞，万振昆，毛婵娟，张天真*

（南京农业大学棉花研究所，作物遗传与种质创新国家重点实验室，南京 210095）

摘要：以陆地棉 TM-1 开花后 6 d 的胚珠为对照，比较开花后 6、10、14、18、22 d 纤维的蛋白质组双向电泳图谱，从中选取了编号为 37 的差异表达的蛋白质点。胶内胰蛋白酶酶切的多肽，用 AB 4700 蛋白质组学分析仪进行 MALDI-TOF/TOF 分析，获得了该点内蛋白质的高品质肽质量指纹图谱(PMF)，从中挑出质量为 1517.9 的肽离子进行串联质谱分析，获得碰撞后断裂片段离子的串联质谱(FFP)。将获得的 PMF 分别用 MASCOT、ProFound、Aldente 和 MS-Fit 等常用软件对 NCBInr 和 UniProt 数据库中的绿色植物蛋白质数据库进行搜索，并用 MASCOT 搜索本地棉属 EST 数据库。将获得的 FFP 用 MASCOT 软件的离子搜索模式和序列查询模式搜索数据库结合手工质谱解析推断序列，最后确认该蛋白质是羟甲基转移酶。本文着重讨论了蛋白质组学研究中质谱分析和生物信息学软件的应用，评价常用蛋白质组学 PMF 检索工具及其检索结果。

关键词：棉花；胚珠；纤维；蛋白质；质谱

中图分类号：S562.035.3 **文献标识码**：A

文章编号：1002-7807(2008)-04-0281-08

Mass Spectrum Identification and Bioinformatics Analysis in Proteomic Research

LIU Kang, GAO Qi-fei, WAN Zhen-kun, MAO Chan-juan, ZHANG Tian-zhen*

(*Cotton Institute, Nanjing Agricultural University, State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing* 210095, *China*)

Abstract: Using ovules of 6 days postanthesis (DPA) of the upland cotton cultivar TM-1 as the control, we compared 2-dimensional electrophoresis gel profiles of the proteins from the fibers at 6, 10, 14, 18, and 22 DPA, and dug one of the differentially expressed spots of NO37 to digest-in-gel with trypsin, and subjected the recovered peptides to AB4700 Proteomics Analyzer for MALDI-TOF/TOF mass spectrometry (MS) analysis. We obtained a peptide mass fingerprint (PMF) of high quality, according to the peak abundance and quality of this PMF, we selected a peptide peak with mass of 1517.9 for tandem MS (MS/MS) analysis, and acquired a fragment fingerprint (FFP) in high quality. The PMF was used to search against NCBInr viridiplantae protein database with MASCOT, ProFound, and MS-Fit engines, and against UniProt viridiplantae protein database with Aldente program, respectively, this PMF was also searched against a local *Gossypium* EST database downloaded from NCBInr with MASCOT in-house version. Furthermore, the MS/MS data were used to search NCBInr viridiplantae protein database with MASCOT in ion search and sequence query mode respectively, followed by manual MS interpretation for *de novo* sequencing. The protein in No 37 spot was identified as hydroxmethyltransferase. In this paper we focus our discussion on MS analysis and the application of bioinformatic tools, we also provide our comments on the typical PMF alignment tools

收稿日期：2007-06-12 **作者简介**：刘 康(1965-)，男，副教授，博士，kangliu@njau.edu.cn；*通讯作者，cotton@njau.edu.cn

基金项目：国家自然科学基金(30671120)；江苏省基础研究计划(BK2006138)和高等学校创新引智计划(supported by the 111 project)(B08025)

in proteomics and their searching results.

Key words: cotton; ovule; fiber; protein; mass spectrometry

随着生物质谱和 IPG(immobilized pH Gradient)为基础的双向电泳(2-DE)技术的发展,蛋白质组学成为后基因组时代的一个重要研究领域,蛋白质组学研究迅速普及到一般规模的重点实验室,从人类以及模式生物扩展到非模式生物和农作物。质谱仪的高度自动化以及生物信息学的发展,似乎不需要很多专业知识和技术就能够"鉴定"很多蛋白质。实际上,如何解读质谱、从质谱中正确提取数据、选用合适的生物信息学软件进行分析,以及如何评价鉴定结果不仅直接关系到蛋白质鉴定结果的正确性,同时也是很多没有足够经验的实验室所面临的一个问题。本文以一个在陆地棉 TM-1 开花后 6、10、14、18、22 d 胚珠、纤维 2-DE 图谱中差异表达的第 37 号点内蛋白质的鉴定为例,讨论如何正确鉴定一个蛋白质,为蛋白质组学大规模研究打下基础。

1 材料和方法

1.1 试验材料

陆地棉遗传标准系 TM-1 自交、挂牌标好开花日期,分别取 6、10、14、18、22 DPA 的胚珠及其附着的纤维,手工剥离纤维和胚珠。

1.2 试验方法

1.2.1 *蛋白质的提取、溶解、定量、双向电泳*。按照前文[1]报道的酚法分别提取 6 DPA 的裸露胚珠以及 6、10、14、18、22 DPA 纤维总蛋白质:将组织在液氮研磨至粉末,加入 3 倍体积的提取缓冲液(500 mmol·L^{-1} Tris-HCl pH 8.65,50 mmol·L^{-1} EDTA,100 mmol·L^{-1} KCl,2 mmol·L^{-1} DTT),等体积饱和酚匀浆,12000 g 离心,收集酚相以及界面,加入 10×体积含 0.1 mmol·L^{-1} NH_4Ac 的冷甲醇,-20C 过夜后离心,沉淀悬浮于含 NH_4Ac 的冷甲醇洗涤三次,冷丙酮洗涤一次,真空干燥后密封储存于-70℃备用。电泳前按加入样品裂解液。定量用考马斯亮蓝法,双向电泳步骤同前文[1]。

1.2.2 *双向电泳聚丙烯酰胺胶的染色与图像扫描*。2-DE 胶用胶体考马斯亮蓝法进行染色。用 Bio-Rad 凝胶扫描系统采集图像,用 PDQuest 2-D 分析软件对图像进行分析。

1.2.3 *挖点与胶内酶切*。参照廖翔等[2]的方法进行。

1.2.4 *质谱分析*。参照 Bienvenutd 等[3]的参数和方法在 AB 4700 蛋白质组分析仪进行 MALDI-TOF 和 MALDI-TOF/TOF 分析。

1.3 数据分析与生物信息学鉴定

1.3.1 *质谱峰值的提取*。用 Data Explorer(Applied Biosystems)软件自动去除同位素峰,获取单同位素肽质谱峰值,手工验证和修正并参照空白对照样品去除常见的基质和角蛋白质峰、胰蛋白酶自切以及其它试剂或污染物引起的杂峰。

1.3.2 *数据库的构建和搜索*。(1)本地数据库的构建:从 NCBI 下载棉属 EST 序列(2005 年 12 月),经过格式转换,本地 MASCOT 软件可以直接对其进行质谱匹配。应用 Emboss 软件包[4],按照正反各 3 种阅读框架电子翻译每条 EST 序列,获得"推导"蛋白质序列作为本地棉属蛋白质数据库。(2)PMF 搜索。使用 MASCOT、ProFound、Aldente 和 MS-Fit 四种搜索程序进行。搜索参数的标准设置是:NCBInr(MASCOT、ProFound)、UniProtKB/Swiss-Prot(Aldente)绿色植物(Viridiplantae)或者 NCBInr 中的 Arabidopsis 和 rice (MS-Fit)蛋白质数据库;胰蛋白酶(Trypsin);最多 1 个漏切位点;固定修饰:未选;可变修饰:Carbamidomethyl(C)和 Oxidation(M);肽段误差:50 ppm;质量类型:MH^+;单同位素。肽质量:0-4000;pI 3-10。(3)MALDI-TOF/TOF 串联质谱的搜索:用 MASCOT 序列查询(sequence query)模式和离子搜索模式联网搜索。搜索参数设置:肽离子误差选 0.15 Da;MS/MS 误差:0.4;数据格式:Micromass (PKL);仪器:MALDI-TOF/TOF,其他参数与 PMF 搜索相同。

2 结果与分析

2.1 不同发育阶段的棉花纤维蛋白质 2-DE 胶图谱比较

以 6 DPA 的去毛胚珠为对照,比较了 6、10、14、18 和 22 DPA 分别代表纤维快速极性伸长期、膨胀(expansion)峰速期、初生壁合成向次生壁合成转变前期、中期、后期等 5 个时期的纤维蛋白质的 2-DE 胶图谱。图 1a 是 6 DPA 裸露胚珠和 6、10、14、18 和 22 DPA 纤维混合样品的蛋白质 2-D 电泳图谱,总计检测到 1200 多个蛋白质

点。其中有 77 个点差异性表达，第 37 号蛋白质点所在的位置用箭头标示。

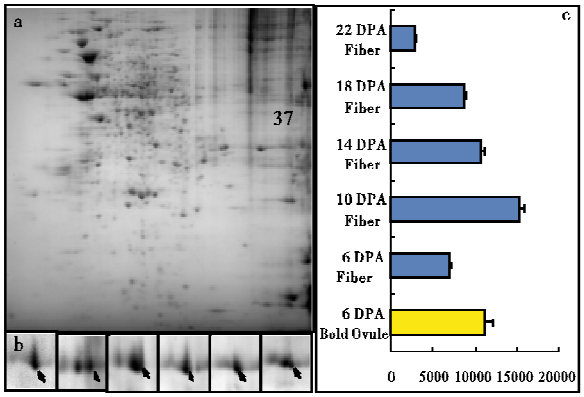


图 1　陆地棉 TM-1 开花后 6 d 裸露胚珠和 6,10,14,18,22 DPA 纤维混合样品总蛋白质的 2-DE 图谱

Fig. 1　2-DE map of the equally mixed proteins extracted from the depilated ovules of 6 DPA, and the fibers of 6、10、14、18、22 DPA, respectively, in upland cotton variety of TM-1

图 1b 显示 37 蛋白质点在各个时期胚珠或纤维中的表达情况，图 1c 该点的表达量分布图，可见该蛋白质在不同时期差异表达，而且在 6、10、22 DPA 纤维中表现为两个点，这可能是该蛋白质发生了某些修饰作用。该点中的蛋白质在纤维发育过程中表现活跃，值得进一步分析。

2.2　蛋白质胶内酶切后的质谱分析

从 2-DE 胶挖下 37 号蛋白质点，胰蛋白酶胶内酶切产物用 MALDI-TOF/TOF 质谱仪进行质谱分析：首先获得酶解多肽的 PMF，然后，选择其中强度大、峰形好的肽离子，送入高能撞池中碰碎(CID)，获得的碎片离子再进行 TOF 分析。图 2a 是 37 号蛋白质点的 PMF，图 2b 是分子量为 1517.9 的肽离子 CID 碰碎后碎片的 TOF/TOF 质谱图。这两张质谱图具有信噪比大、峰形明显、解析度高等特点，可以用于数据的分析和蛋白质数据库的搜索。

2.3　用 PMF 搜索蛋白质数据库

用 Data Explorer 软件从 MS 和 MS/MS 质谱中提取峰值：在设置好信噪比(S/N)、基线校准之后，去除同位素，获得质谱峰值用于数据库搜索。图 3、4、5 分别是将 37 号蛋白质点的 PMF 用 MASCOT、ProFound 和 MS-Fit 搜索数据库的结果。

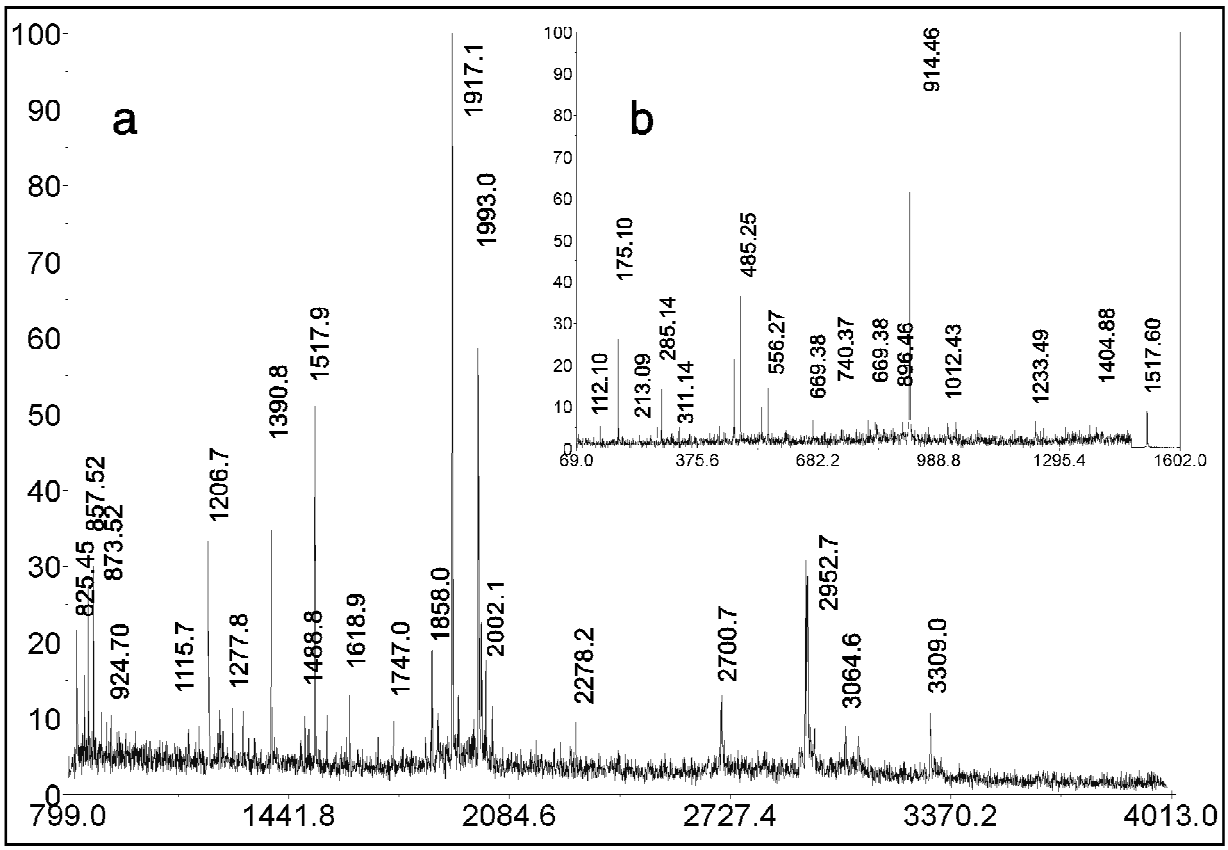


图 2　(a)第 37 号蛋白质点酶切后的 MALDI-TOF 质谱(PMF)
(b)分子量为 1517.9 Da 的肽离子 CID 碰碎后碎片的 MS/MS 质谱

Fig. 2　(a) MALDI-TOF spectrum of the tryptic digest of spot 37
(b) MS/MS spectrum of 1517.9 Da peptide ion acquired with the MALDI-TOF/TOF

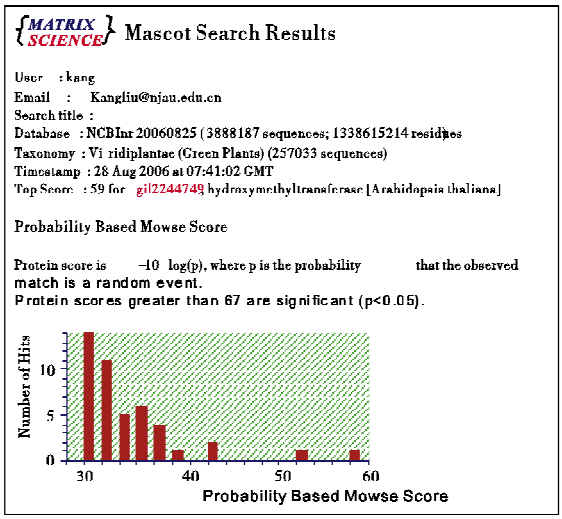


图 3 第 37 号蛋白质点的 PMF 用 MASCOT 搜索 NCBInr 绿色植物蛋白质数据库的最高匹配项

Fig. 3 Top hit of MASCOT search output conducted by using PMF of spot 37 against viridiplantae protein database in NCBInr

MASCOT 搜索的结果中的首项虽然有 9 个肽匹配，氨基酸序列覆盖率达到 23%，但最高分值为 59(图 3)，小于 5%显著水平的临界值，说明该蛋白质是随机匹配的可能性大于 5%(所有匹配肽段均在阴影区间内)。前三位匹配的蛋白质不同，但是分值差异不大，所以，该蛋白质点没有得到确定鉴定。

ProFound 搜索结果中的最高匹配的 Z 值达到 1.43，是随机匹配的概率约为 7.6%，略高于 5%，但是，排行第一和第二的概率值和 Z 值相差悬殊(图 4)。有人通过测序对 ProFound 检索结果进行验证，发现 Z 值大于 1.15 以上，就没有发现有错(即假阳性)。9 个肽匹配，序列覆盖率达到 23%，所以 37 号蛋白质点很有可能是羟甲基转移酶(EC 2.1.2.1)。

图 5 是 MS-Fit 检索拟南芥蛋白质序列数据库的结果，第一名也是羟甲基转移酶，与 MASCOT 和 ProFound 检索结果一致。MOWSE 分值比第二匹配项高出约 2000。由于没有给出显著阈值，很难判断该结果的有效性。

ProFound–Search Result Summary

Version 4.10.5
The Rochefeller University Edition

Protein Candidates for search B2295995–08E41–1CBD6DB5[172119 sequences searched

Rank	Probability	Est' d z	Protein Information and Sequence Analyse Tools(T)	%	*PI*	kDa	®
+1	1.0e+000	1.43	Tgi|7268097|emb|CAB78435.1|hydroxylmethyltransferase[Arabidopsis thaliana]	23	6.8	52.16	®
–	–		Tgi|11762130|gb|AAG40343.1|AT4g13930[Arabidopsis thaliana]	19	7.2	42.24	®
–	–		Tgi|21592544|gb|AAM64494.1|hydroxylmethytransferase[Arabidopsis thaliana]	19	6.6	52.14	®
+2	4.7e–005	0.00	Tgi|28416483|AAO42772.1|At3g49060/T2J13_100[Arabidopsis thaliana]	13	6.0	92.54	®
–	–		Tgi|6522560|emb|CAB62004.1|putative protein [Arabidopsis thaliana]	10	5.6	135.05	®
+3	2.0e–006	–	Tgi|30385668|gb|AAP23874.1|pyruvate phosphate dikinase [Sorghum bicolor]	10	5.7	103.07	®

图 4 第 37 号蛋白质点的 PMF 用 ProFound 搜索 NCBInr 绿色植物蛋白质数据库的结果

Fig. 4 ProFound search output conducted by using PMF of spot 37 against viridiplantae protein database in NCBInr

MS–Fit Search Results

Abort Search

Data Set1 Results`

MS–Fitsearchselects425 entries(resultsdisplayedfor top 5 matches).

Result Summary

	MOWSE Score	#/39(%) Masses Matched	% Cov	% TIC	Mean Err ppm	Data Tol ppm	MS–Digest Index#	Protein MW(Da)/pI	Accession#	Species	Protein name
1	7355	8(20)	22	20.5	79.6	28.4		51718/68	M	ARABIDOPSISTHALIANA	SHM4(SERINEHYDROXYMETHYLTRANSFERASE4); glycinehydroxymethyltransferase
2	5360	9(23)	13.0	23.1	52.2	56.0		91593/6.0	M	ARABIDOPSISTHALIANA	ATP binding/kinase/protein kinase/protein serine/threonine kinase/protein–tyrosine kinase/ubiquitin–protein ligase
3	5358	9(23)	13.0	23.1	52.2	56.0		91621/6.0		ARABIDOPSISTHALIANA	unknown protein
4	3509	8(20)	18.0	20.5	50.6	58.4		57874/5.8		ARABIDOPSISTHALIANA	unknown protein
5	1238	4(10)	20.0	10.3	56.1	53.2		36968/8.9		ARABIDOPSISTHALIANA	unknown

图 5 第 37 号蛋白质点的 PMF 用 MS-Fit 搜索 NCBInr 拟南芥蛋白质数据库的结果

Fig. 5 MS-Fit search output conducted by using PMF of spot 37 against Arabidopsis protein database in NCBInr

在 Aldente 的搜索结果中,15 个匹配项的 Z 值,最高 3.56(第 1 条),最低 2.12,随机匹配的概率都小于 5%。但是内部分值(internal score)最高 1.07,最低 0.49,都比较低。由此没有给出确切的鉴定结果。考虑到最佳匹配的蛋白质 Expansin 在纤维发育中起重要作用,存在该蛋白质点是混合蛋白的可能性。点击被认为是正确鉴定的匹配项后的 Validate,重新设置和放宽检索参数(主要是修饰类型),对未匹配的肽峰值再次进行搜索。我们尝试了比如磷酸化等其它修饰,但是结果都没有找到羟甲基转移酶。

2.4　用 MS/MS 数据搜索蛋白质数据库

为准确鉴定该点内的蛋白质,选取质量为 1517.9 的肽离子进行了高能碰撞获得串联质谱。用 MASCOT "序列查询"模式搜索 NCBInr 绿色植物蛋白质序列数据库,结果如图 6 所示。虽然这是个单离子 FFP 的搜索结果,但是结合手工解析质谱,发现大多数高强度峰分别与 y(y_1、y_2、y_5～y_{13})、a_3、b_3 离子吻合,还有一个 R(精氨酸)的 immonium 离子。据此可以推导出该肽段的部分序列为 FGDSSA(L/I)AP(GG/N)VR(图 7),BLAST 比对结果证实该蛋白质与羟甲基转移酶高度同源。

综合以上分析,PMF 数据用 MASCOT、Pro-Found、Ms-Fit 搜索的结果与 MS/MS 数据的 MASCOT 搜索和手工解析质谱推导氨基酸序列相一致。可以确定该蛋白质是羟甲基转移酶。

Mascot Search Results

Peptide View

MS/MS Fragmentation of　NAVFGDSSALAPGGVR

Found in **gi|11762130** , AT4g13930 [Arabidopsis thaliana]

MONOISOTOPIC mass of neutral peptide (Mr): 1516.76

Ions Score: 76　**Matches (Bold Red):** 32/117 fragment ions using 43 most intense peaks

#	Immon.	A	a*	a^0	b	b*	b^0	seq.	y	y*	y^0	#
1	87.06	87.06	70.03		115.05	98.02		N				16
2	44.05	158.09	141.07		186.09	169.06		A	1403.73	1386.70	1385.72	15
3	72.08	257.16	240.13		285.16	268.13		V	1332.69	1315.66	1314.68	14
4	120.08	404.23	387.20		432.22	415.20		F	1233.62	1216.60	1215.61	13
5	30.03	461.25	444.22		489.25	472.22		G	1086.55	1069.53	1068.54	12
6	88.04	576.28	559.25	558.27	604.27	587.25	256.26	D	1029.53	1012.51	1011.52	11
7	60.04	663.31	646.28	645.30	691.31	674.28	673.29	S	914.51	897.48	896.50	10
8	60.04	750.34	733.32	732.33	778.34	761.31	760.33	S	827.47	810.45	809.46	9
9	44.05	821.38	804.35	803.37	849.37	832.35	831.36	A	740.44	723.42		8
10	86.10	934.46	917.44	916.45	962.46	945.43	944.45	L	669.40	652.38		7
11	44.05	1005.50	988.47	987.49	1033.55	1016.47	1015.48	A	556.32	539.29		6
12	70.07	1102.55	1085.53	1084.54	1130.55	1113.52	1112.54	P	485.28	468.26		5
13	30.03	1159.57	1142.55	1141.56	1187.57	1170.54	1169.56	G	388.23	371.20		4
14	30.03	1216.60	1199.57	1198.59	1244.59	1227.56	1226.58	G	331.21	314.18		3
15	72.08	1315.66	1297.65	1297.65	1343.66	1326.63	1325.65	V	274.19	257.16		2
16	129.11							R	175.12	158.09		1

图 6　选取质量为 1517.9 的母离子进行 CID 高能碰撞后获得的 MS/MS 用 MASCOT 序列查询模式搜索 NCBInr 绿色植物蛋白质数据库的结果

Fig. 6　MASCOT search output in sequence query mode using MS/MS of the fragments produced by CID of a precursor peptide of 1517.9 against NCBInr viridiplantae protein database

2.5 EST数据库的应用

PMF数据搜索EST数据库成功鉴定蛋白质[6-9]，大大提高了基因组序列未知生物蛋白质鉴定的成功率。棉花EST可以作为研究棉花功能基因组学和蛋白质组学的重要资源。37号蛋白质点的PMF用MASCOT搜索本地棉属EST库，匹配的EST序列(gi|48817448)分值达到显著水平(图8)，有12个肽匹配，序列的覆盖率达72%。经过blastx比对，该EST正向第二框架翻译的蛋白质序列与拟南芥中的羟甲基转移酶(SHM4，gi|15236375)高度相似，虽然SHM4和MASCOT搜索NCBInr的第一匹配项(gi|2244749，图3)以及ProFound的匹配项(gi|7268097，图4)名称不同，但是蛋白质序列完全相同。将翻译的EST序列与MASCOT、ProFound和MS-Fit搜索到拟南芥羟甲基转移酶序列进行比对，并对照质谱匹配的肽段序列(图9)，发现在250个氨基酸残基的肽段上，两个序列有9.6%的氨基酸是不相同的，有8个质谱峰与EST翻译肽段上的部分序列匹配，而只有5个质谱峰与SHM4的这部分序列中的某些肽段相匹配(阴影部分)，有4个肽段与翻译的EST匹配而不能与SHM4匹配，有1个肽段与SHM4匹配而不与翻译的EST匹配(方框)。可见，同种功能的蛋白质由于不同植物之间存在着或多或少的序列差异，而PMF检索仅仅根据酶切片段的分子量信息来搜索蛋白质数据库，当匹配的是不同物种的蛋白质时，容易产生偏差。检索相同或近缘物种的EST库，由于蛋白质同源性高，得到的结果更为可靠。

3 讨论与结论

3.1 PMF搜索程序的评价

MALDI-TOF质谱产生的PMF尤其适用于蛋白质的大规模初步鉴定。MASCOT、ProFound和Aldente都给出分值。ProFound和Aldente的分值是固定的Z值，95%的置信水平的Z值是固定的1.65，不受数据库大小的影响。而MASCOT的分值则随数据库而变化。MS-Fit不提供显著水平，只有MOWSE值，有研究指出，MOWSE值达到14000仍然有随机匹配，而MOWSE值等于5时仍有正确匹配。所以很难从MS-Fit搜索的结果中判断出鉴定的准确性。虽然MASCOT、ProFound也都是以MOWSE算法为基础，但是处理方法完全不同，MASCOT计算概率，可以直接了解随机匹配的概率的大小。ProFound则根据数据库中的蛋白质序列出现的可能性进行Bayesian统计过滤，大大提高了蛋白质鉴定的灵敏性和选择性。Aldente的算法比较特殊，而且使用了Hough转换，可以预先设定分值阈值，如果放宽条件，可以搜索到更多的结果。MASCOT可供检索的数据库种类最多，ProFound只有NCBInr数据库，而且不是即时更新。MS-Fit的可供选用的数据库虽然多达18种，但也不是即时更新，而且没有整个绿色植物这样的分类选项。如果研究对象不属于这些物种则很麻烦。Aldente可供选用的数据库为Swiss-Prot和TrEMBL，该数据库即时更新，功能注释最为详细，而且与许多数据库超连接便于进一步了解搜索到蛋白质的各种性质。不少研究[10-12]通过结果表明ProFound搜索的准确率最高，检出的比例也最高。但是作者的研究(未发表资料)表明有些PMF用ProFound检索没有结果，而用MASCOT检索时却有很好的匹配。有人建议用MS-Fit检索数据库，凡是MOWSE值大于2000的匹配蛋白质都作为候选蛋白质，然后用其它软件或者试验手段对其进行验证。但是对于植物来说，由于其数据库存在上述缺陷，采纳这个建议并不方便。Aldente虽然可以给出很多结果，但是本研究以及其它研究结果都表明其准确度并不高。

3.2 基于PMF数据的蛋白质鉴定

PMF鉴定蛋白质往往遇到以下问题：(1)不同搜索引擎搜索结果不同；(2)改变搜索条件搜索结果不同；(3)搜索不到任何显著的结果；(4)搜索结果中的匹配项难以取舍。解决这些问题应该从两方面着手：一是从质谱中正确提取数据。试验中尽量避免干扰物质，设置空白胶对照便于正确去除杂峰，正确进行质谱的标定、去同位素等。智能程度高的软件(如GPS和RADAR)获得的数据一般不需要人工核对可以直接用于数据库搜索，其它软件建议手工验证之后再用于数据库搜索。二是数据库搜索：总的策略是以MASCOT、ProFound搜索为主，两者的结果相互印证相互补充，以MS-Fit搜索结果作为参考。

PMF可以用MASCOT等软件搜索本地EST库，对于非测序生物蛋白质的鉴定有时候更为可信。至于EST存在的序列不全、冗余多等缺陷可以通过去冗余和整合拼接等手段加以解决。然而，目前越来越多的人主张用串联质谱检索EST更可靠。

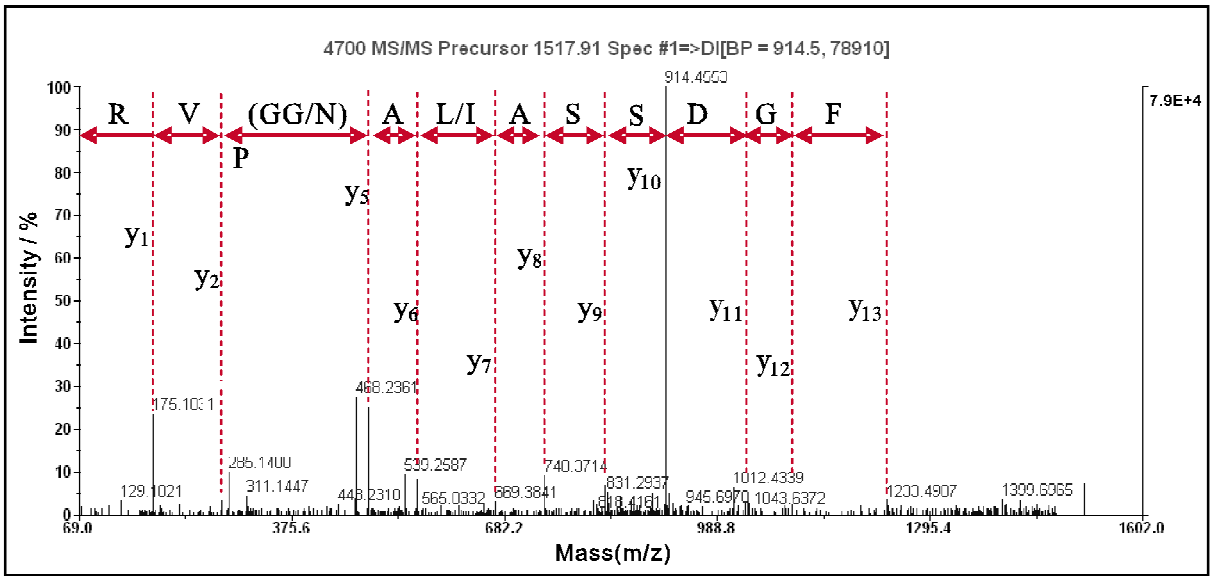
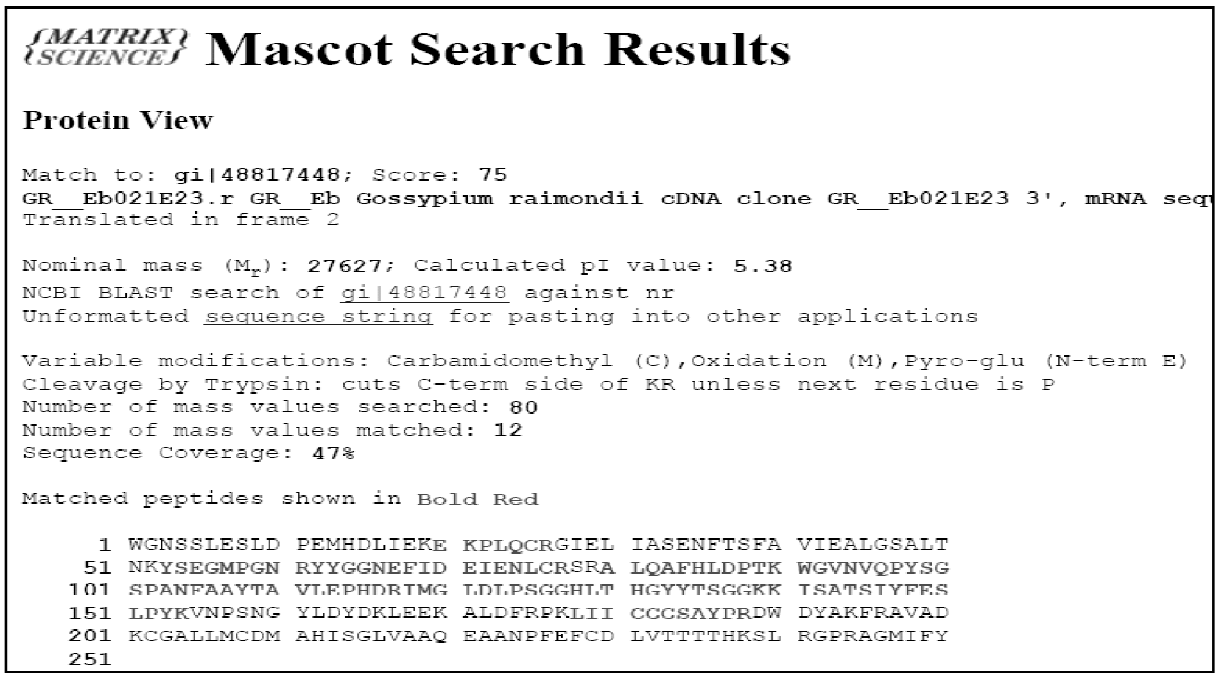


图 7　对质量为 1517.9 的母离子 MS/MS 数据的手工从头序列分析图

Fig. 7　Manual peptide de novo sequencing map based on MS/MS spectra of a precursor ion of 1517.9



MATRIX SCIENCE **Mascot Search Results**

Protein View

Match to: gi|48817448; Score: 75
GR__Eb021E23.r GR__Eb Gossypium raimondii cDNA clone GR__Eb021E23 3', mRNA seq
Translated in frame 2

Nominal mass (M_r): 27627; Calculated pI value: 5.38
NCBI BLAST search of gi|48817448 against nr
Unformatted sequence string for pasting into other applications

Variable modifications: Carbamidomethyl (C),Oxidation (M),Pyro-glu (N-term E)
Cleavage by Trypsin: cuts C-term side of KR unless next residue is P
Number of mass values searched: 80
Number of mass values matched: 12
Sequence Coverage: 47%

Matched peptides shown in Bold Red

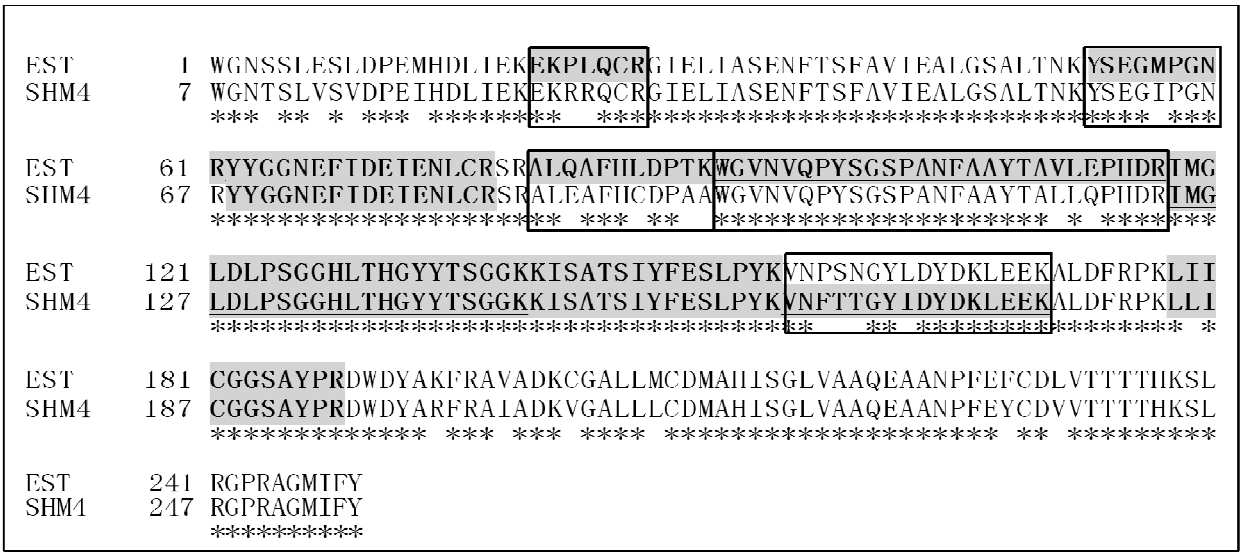
```
     1 WGNSSLESLD PEMHDLIEKE KPLQCRGIEL IASENFTSFA VIEALGSALT
    51 NKYSEGMPGN RYYGGNEFID EIENLCRSRA LQAFHLDPTK WGVNVQPYSG
   101 SPANFAAYTA VLEPHDRIMG LDLPSGGHLT HGYYTSGGKK ISATSIYFES
   151 LPYKVNPSNG YLDYDKLEEK ALDFRPKLII CGGSAYPRDW DYAKFRAVAD
   201 KCGALLMCDM AHISGLVAAQ EAANPFEFCD LVTTTTHKSL RGPRAGMIFY
   251
```

图 8　37 号样品的 PMF 用 MASCOT 搜索本地棉属 EST 数据库所得到的结果

Fig. 8　MASCOT search result of PMF of sample NO37 against local *Gossypium* EST database

3.3　串联质谱的应用

PMF 鉴定结果受到蛋白质酶切是否充分和准确性、质谱仪的参数设定、基质的选用、质谱峰值的提取和解释[12]、数据库以及搜索工具的选用和搜索条件的设置等诸多因素的影响。因此，PMF 鉴定蛋白质的成功率和可信度都不够高。串联质谱是选取酶解多肽的某一个离子进一步碰撞得到 FFP，由于提供了更多的信息，大大提高蛋白质鉴定的准确性。MASCOT 提供的 MS+MS/MS 联合分析模式来检索数据库，可以很好地匹配到蛋白质序列上去，并给出 a、b、y 以及 immonium 离子的匹配信息(图 6)。根据这些信息，返回到原始的质谱图，可以手工分析侧链修饰[3]。串联质谱搜索 EST 数据库也是目前公认的鉴定基因组序列未知生物的蛋白质的一条可靠途径。



星号代表氨基酸残基相同的座位，阴影粗体字母代表 PMF 搜索时匹配的肽段氨基酸，如果这些肽段彼此连续，则以下划线区分开；方框提示 PMF 在 EST 和 SHM4 两个序列上有不同的匹配

Asterisks represent sites with same amino acid residues, bold letters in shade show matched peptides, underlines are used to distinguish when the matched peptides are consecutive; squares indicate different matches found between EST and SHM4

图 9 EST 翻译的蛋白质序列与拟南芥 SHM4 序列的比对以及质谱峰匹配图

Fig. 9 Diagram of the alignment between the translated EST sequence and *Arabidopsis* SHM4 protein sequence and matched peptides by PMF search

参考文献：

[1] 刘 康，胡凤萍，张天真. 棉花胚珠与纤维蛋白质的两种提取方法比较研究[J]. 棉花学报，2005，17(6)：323-327.

[2] 廖 翔，应天翼，王恒樑，等. 考马斯亮蓝染色双向电泳凝胶内酶切方法的改进[J]. 生物技术通讯，2003，14(6)：509-511.

[3] BIENVENUT W V, Déon C, Pasquarello C, et al. Matrix-assisted laser desorrption/ionization-tandem mass spectrometry with high resolution and sensitivity for identification and characterization [J]. Proteomics, 2002, 2: 868-876.

[4] RICE P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite [J]. Trends in Genetics, 2000, 16 (6): 276-277.

[5] MANN M, Hendrickson R C, Pandey A. Analysis of proteins and proteomes by mass spectrometry [J]. Annu Rev Biochem, 2001, 70: 437-473.

[6] PROBLEVA L, Vander Velden K, Kothari S, et al. The proteome of maize leaves use of gene sequences and expressed sequence tag data for identification of proteins wit peptide mass fingerprints [J]. Electrophoresis, 2001, 22: 1724-1738.

[7] WATSON B S, Asirvatham V S, Wang L, et al. Mapping the proteome of barrel medic (Medicago truncatula) [J]. Plant Physiol, 2003, 131: 1104-1123.

[8] MOONEY B P, Thelen J J. High-troughput peptide mass fingerprinting of soybean seed proteins: automated workflow and utility of Unigene expressed sequence tag databases for protein identification [J]. Phytochemistry, 2004, 65: 1733-1744.

[9] CHAMRAD D C, Koerting G, Gobom J, et al. Interpretation of mass spectrometry data for high-throughput proteomics [J]. Anal Bioanal Chem, 2003, 376: 1014-1022.

[10] LUBEC G, Afjehi-Sadat L, Yang J W, et al. Searching for hypothetical proteins: Theory and practice based upon original data and literature [J]. Progress in Neurobiology, 2005, 77: 90-127

[11] TANG C, Zhang Wen-zhu, Feny D, et al. Assessing the performance of different protein identification algorithms [C]// ASMS. 48TH ASMS Conference. Long Beach, California [s. i.], 2000.

[12] CHAMRAD D. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data [J]. Proteomics, 2004, 4: 619-628.