한국식물생명공학회지 제29권 제3호
Korean J. Plant Biotechnology
Vol. 29, No. 3, 161~170 (2002)

인삼 모상근 프로테옴 데이터 분석 : 인삼 EST database와의 통합 분석에 의한 단백질 동정

권경훈 · 김승일 · 김경욱 · 김은아 · 조 건 · 김진영 · 김영환 · 양덕춘[1] · 허철구[2] · 유종신 · 박영목*
한국기초과학지원연구원 프로테옴분석팀, [1]바이오피아, [2]한국생명공학연구원

Proteome Data Analysis of Hairy Root of *Panax ginseng* : Use of Expressed Sequence Tag Data of Ginseng for the Protein Identification

KWON, Kyung-Hoon · KIM, Seung Il · KIM, Kyung-Wook · KIM, Eun A · CHO, Kun · KIM, Jinyoung · KIM, Young Hwan · YANG, Deok-Chun[1] · HUR, Cheol-Goo[2] · YOO, Jong-Shin · PARK, Young-Mok*
Proteome Analysis Team, Korea Basic Science Institute, Daejeon 305-806, Korea
[1]Biopia, Daejeon 305-345, Korea
[2]Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea

ABSTRACT For the hairy root of *Panax ginseng*, we have got mass spectrums from MALDI/TOF/MS analysis and Tandem mass spectrums from ESI/Q-TOF/MS analysis. While mass spectrum provides the molecular weights of peptide fragments digested by protease such as trypsin, tandem mass spectrum produces amino acid sequence of digested peptides. Each amino acid sequences can be a query sequence in BLAST search to identify proteins. For the specimens of animals or plants of which genome sequences were known, we can easily identify expressed proteins from mass spectrums with high accuracy. However, for the other specimens such as ginseng, it is difficult to identify proteins with accuracy since all the protein sequences are not available yet. Here we compared the mass spectrums and the peptide amino acid sequences with ginseng expressed sequence tag (EST) DB. The matched EST sequence was used as a query in BLAST search for protein identification. They could offer the correct protein information by the sequence alignment with EST sequences. 90% of peptide sequences of ESI/Q-TOF/MS are matched with EST sequences. Comparing 68% matches of the same sequences with the nr database of NCBI, we got more matches by 22% from ginseng EST sequence search. In case of peptide mass fingerprinting from MALDI/TOF/MS, only about 19% (9 proteins of 47 spots) among peptide matches from nr DB were correlated with ginseng EST DB. From these results, we suggest that amino acid sequencing using tandem mass spectrum analysis may be necessary for protein identification in ginseng proteome analysis.

Key words : Database, expressed sequence tag, mass spectrometer, *Panax ginseng*, proteome

*Corresponding author Tel 042-865-3420
E-mail ympark@kbsi.re.kr

서 론

포스트게놈 시대에 생명체들의 유전자에서 발현되는 단백질을 조사하고 이들의 기능을 밝혀내는 프로테오믹스 분야의 연구는 신약 개발과 질병 예방을 위해 기초 연구로서의 중요성을 가진다. 프로테오믹스는 세포의 단백질을 연구하는 분야로 전기영동에 의한 2D gel을 사용하여 단백질을 분리하고, 질량분석기와 데이터베이스 검색으로 단백질을 동정(identification)하며 이들의 기능 및 구조를 분석한다.

과학기술의 발달과 더불어 실험 장비들의 성능은 크게 향상되어 대량의 데이터를 단시간에 얻게 되었다. 한편 컴퓨터 및 정보 기술의 발달은 얻어진 실험 데이터를 대용량의 데이터베이스를 이용하여 분석하고 유용한 정보를 추출해낼 수 있도록 하였다. 첨단의 분석 장비와 분석 기법, 고성능 컴퓨터 시스템 및 데이터 분석 알고리듬을 종합적으로 활용하는 프로테오믹스 분야는 이러한 과학과 기술의 발전과 더불어 세계적으로 활발하게 연구가 진행되고 있으며, 데이터의 획득 및 분석 기법이 지속적으로 개발되고 있다.

전기 영동법으로 단백질의 분자량 및 pI 값에 따라 2D gel로 분리된 단백질은 다양한 효소를 사용하여 가수분해에 의해 작은 펩타이드들로 분해한다. 분해된 펩타이드들은 질량분석기를 이용하여 얻은 질량 스펙트럼으로부터 분자량을 측정하고 이를 데이터베이스로 검색하여 단백질을 추정한다 (Nyman 2001). 질량 스펙트럼을 얻는 실험에서 한 단계 더 나아가 단백질에 대한 보다 정확한 정보를 얻으려면, 펩타이드를 다시 아미노산 단위로 분해하고 분자량을 측정하는 탄뎀 질량분석기를 사용한다. 탄뎀 질량분석기로부터 얻는 탄뎀 질량 스펙트럼을 데이터베이스로 검색하면, 각 펩타이드의 아미노산 서열을 구할 수 있다 (Ashton et al. 2001). 한편 계산 시간은 오래 걸리지만, *de novo* sequencing 방법을 사용하면 데이터베이스 없이도 탄뎀 질량 스펙트럼으로부터 펩타이드의 아미노산 서열을 계산해낸다 (Taylor and Johnson 1997). 질량분석 스펙트럼이나 탄뎀 질량분석 스펙트럼으로부터 아미노산 서열을 구하는 데에는 지금까지 축적된 데이터베이스를 활용하는데, 일반적으로 NCBI의 non-redundant protein 데이터베이스 (nr DB)를 가장 많이 사용한다.

탄뎀 질량 스펙트럼으로부터 아미노산 서열을 얻었을 때, 이로부터 단백질을 검색하려면 주로 BLAST 프로그램을 사용한다 (Altshul et al. 1990). BLAST 검색과 같이 염기서열이나 아미노산 서열을 검색어로 사용하는 검색 프로그램의 경우에는 그 서열과 유사한 서열이 단백질 안에 존재하면 검색이 가능해진다. 즉, 몇 군데에서 아미노산에 차이가 나는 다른 단백질이라도 그 차이가 중요하지 않은 경우에는 비슷한 서열로서 검색 결과에 포함되어 나타나며, 단지 서열의 유사도에 따라 정렬에 대한 점수가 달라진다.

반면에 질량분석기에서 얻은 질량스펙트럼에서 나온 분자량들의 프로파일로부터 단백질을 동정하기 위해서는 검색하려는 데이터베이스에 해당 펩타이드가 정확하게 일치하도록 존재해야 한다. 하나의 아미노산이라도 차이가 나는 펩타이드는 질의어로 입력한 분자량과 완전히 다른 분자량을 보이므로 검색 대상 데이터베이스에서 제외된다. 여기에 질량 스펙트럼에 의한 단백질 동정의 어려움이 있으며, 이를 보완하기 위해 탄뎀 질량분석기를 사용하고 아미노산 서열을 구하게 된다.

한편 이러한 질량 스펙트럼 데이터 검색에서의 한계를 보완하기 위하여 탄뎀 질량 분석기를 사용하기 전에 EST (expressed sequence tag) 데이터베이스를 검색에 활용하는 방법도 고려할 수 있다(Mathesius et al. 2001). 질량 스펙트럼의 효율적인 검색을 위해 EST 데이터베이스를 활용한 사례들을 알아보면, Mann (1996)은 인간 유전자 연구 중에 질량 스펙트럼의 분석에서 펩타이드의 위치 정보와 아주 적은 부분의 아미노산 서열정보들을 조사하고, 이 데이터와 EST 데이터베이스를 단백질 동정에 이용하였다. Porubleva는 옥수수의 프로테옴 분석에서 pdbEST 데이터베이스를 단백질 동정에 활용하여 300개의 단백질 spot 중 67개를 추가로 동정하였음을 보고하였다 (Porubleva et al. 2001). 한편 dbEST 데이터베이스로부터 EST 서열의 isoform들을 다중서열정렬 방법 (multiple sequence alignment)에 의해 비교 분석하여 단백질 동정에 활용한 사례도 발견된다 (Lisacek et al. 2001).

본 연구에서는 한국기초과학지원연구원 (KBSI)에서 얻은 인삼 모상근의 프로테옴 데이터 (Kim et al. 2001)와 바이오피아의 EST 데이터를 활용하여, 인삼에 대한 EST 서열 정보들을 질량 스펙트럼 및 탄뎀 질량 스펙트럼 데이터와 함께 비교 분석함으로써, 단백질 동정에서 인삼의 EST 서열 데이터베이스의 역할을 조사하였다.

재료 및 방법

인삼 Proteome 데이터 분석

이차원 전기영동 겔 (2D gel)로부터 분리된 인삼의 단백질은 trypsin을 이용한 In-gel digestion으로 분해한 뒤 질량분석기 MALDI/TOF/MS에 의해 절편단백질의 질량 스펙트럼을 구하였다. 질량 스펙트럼에서 얻은 peak 값들은 MS-FIT, MASCOT과 같은 검색 프로그램들을 사용하여 단백질을 밝혀내는 데에 사용하였다. 인삼 모상근의 실험에서는 MALDI/TOF/MS의 질량 스펙트럼 중 keratin, trypsin 등에 의한 peak 값들을 제거한 스펙트럼을 nr 데이터베이스에서 검색하였다. 질량 스펙트럼만으로 단백질의 동정이 되지 않은 spot에 대해서는 ESI/Q-TOF/MS에 의해 탄뎀 질량 스펙트럼을 얻어서 펩타이드의 아미노산 서열을 de novo sequencing 방법으로 계산하였다. 그러나, ESI/Q-TOF/MS를 거치지

않고 MALDI/TOF/MS 만으로 단백질을 찾은 경우에는 질량 스펙트럼과 가장 많이 분자량이 일치하는 단백질을 nr DB 중에서 선택하였으므로 이를 해당 단백질로 완전히 확신하기는 어렵다. 데이터에 대한 신뢰도를 높이려면 ESI/Q-TOF/MS 실험에 의해 아미노산 서열을 확인하던지 아니면 다른 방법으로의 확인 작업을 거쳐야 한다. 질량 스펙트럼으로부터 확보한 단백질 정보의 신뢰도를 확인하기 위한 방법으로, 본 연구에서는 인삼에서 얻은 프로테옴 데이터, 즉 *MALDI/TOF/MS*의 스펙트럼과 *ESI/Q-TOF/MS*의 아미노산 서열들을 인삼의 EST 데이터베이스에서 검색하여 nr DB에서 검색한 결과와 비교하였다. 시료는 인삼 모상근의 *CBB G-250-stained 2-D PAGE*를 이용하였으며, pH4-7 범위 내에서 spot을 분리하였다.

ESI/Q-TOF/MS 데이터와 EST sequence 와의 비교

FASTA format으로 구성된 인삼의 EST 염기서열은 각 서열별로 3개의 가능한 아미노산 서열로 변환하여 BLAST DB에 추가하였다. EST 서열의 변환 중에 나타나는 Start, Stop codon은 X라는 코드를 사용하여 비정상적인 codon으로 처리하였다. EST 데이터는 .ab1이라는 확장자로 데이터의 ID 가 표시되었는데, 각 EST 데이터별로 세 개의 아미노산 서열이 대응되므로, 각각에 대해 .ab10, .ab11, .ab12라는 확장자를 부여하여 서열을 구분하였다. BLAST 프로그램은 몇몇 프리웨어가 제공되고 있는데, 여기서는 미국의 NCBI에서 제공하는 standalone BLAST를 dual processor의 Linux server에 설치하여 사용하였다. 단백질 서열 검색은 blastp 프로그램에 PAM30 score matrix로 실행하였다. 한편 ESI/Q-TOF/MS의 아미노산 서열은 FASTA format 으로 파일에 저장하여 BLAST 검색 과정을 여러 아미노산 서열에 대해 Linux 환경에서 자동으로 실행할 수 있도록 하였다.

MALDI/TOF/MS 데이터 분석

MALDI/TOF/MS 데이터를 EST 서열 데이터베이스로 검색하기 위하여 European Molecular Biology Laboratory (EMBL)에서 공개하는 생물정보학 프로그램 패키지인 EMBOSS package에 포함되어 있는 eMOWSE 프로그램을 사용하였다. eMOWSE 프로그램은 C 언어로 개발된 프로그램으로 MOWSE라는 알고리듬을 사용한다 (Pappin et al. 1993). eMOWSE 프로그램에서는 질량 스펙트럼을 ASCII 파일로 저장한 뒤, FASTA format으로 구성한 데이터베이스에서 데이터를 검색할 수 있다. 이를 이용하여 eMOWSE 프로그램에 데이터베이스 파일로 인삼의 EST 데이터베이스를 설정하고 mass fingerprint 데이터는 MALDI/TOF에서 구한 질량스펙트럼 값들을 입력하여 검색을 실행하였다.

결 과

아미노산 서열을 nr DB 와 EST DB에서 검색한 결과의 비교

인삼 모상근 단백질을 전기영동 겔로 분리하여 MALDI/TOF/MS에서 얻은 질량 스펙트럼을 오차 범위는 50 ppm로, missed cleavage는 1개 이하, 단백질 중에 10% 이상의 펩타이드가 질량 스펙트럼에서 얻어져야 한다는 조건으로 nr DB에서 단백질 동정을 시도하였으며, 47개의 발현단백질에서 질량 스펙트럼을 분석한 결과 34%인 16개의 단백질에서 펩타이드 분자량들과 MALDI peak 값들이 일치함을 발견하였다. 탄뎀 질량 스펙트럼으로부터 얻은 펩타이드의 아미노산 서열은 BLAST 검색 결과 21개의 스펙트럼에 대해 유사 단백질을 검색해내었다. 이로서 37개의 spot에 대한 단백질을 동정하였다. 한편 인삼 EST 서열을 BLAST DB로 설치하여 탄뎀 질량 스펙트럼을 검색한 결과 28개의 단백질이 발견되었으며, 이 중 9개는 nr DB에서는 찾을 수 없었던 단백질이고, 반대로 nr DB로는 검색되었으나 EST DB로는 밝혀내지 못한 단백질은 2개이다. Table 1은 ESI/Q-TOF/MS 데이터를 인삼 EST 서열을 이용하여 BLAST 검색한 결과와 nr DB를 이용하여 검색한 결과를 보여준다.

인삼 EST 데이터베이스에서 아미노산 서열을 검색한 뒤에 여기서 검색된 EST 서열을 다시 nr DB에서 검색하여 단백질을 동정하는 분석방법을 좀더 자세히 살펴보기 위해 Table 1에서 4번 발현 단백질에 대한 EST 서열 검색 결과를 분석하였다. BLAST 프로그램의 score matrix는 질의어의 종류 및 검색 목적에 따라 다른 행렬을 사용한다. NCBI의 BLAST 검색에서 일반적인 경우에는 BLOSUM62를 사용하나, 아미노산의 개수가 35개 이하이면서 유사성이 높은 아미노산 서열을 찾는 검색에 대하여는 PAM30을 score matrix로 사용한다. *ESI/Q-TOF/MS*에서 얻은 인삼 단백질의 펩타이드 아미노산 서열에서도 PAM30을 score matrix로 사용하였을 때에 BLOSUM62에서보다 더 많은 검색 결과를 얻어 짧은 아미노산 서열에서는 PAM30이 유리함을 확인할 수 있었다. ESI/Q-TOF/MS으로 4번 단백질에서 얻은 펩타이드의 아미노산 서열 YYLTTNNN은 nr DB로 BLAST 검색을 했을 때에 E-value가 10 이하인 범위 내에서는 유사한 서열을 찾을 수 없었다. 그러나, 이 서열을 EST 데이터베이스에서 검색하였을 때에는, DC01027G01.ab12에서 유사한 아미노산 서열을 찾을 수 있었다 (Figure 1). 다음 단계로 이렇게 찾은 인삼 EST에서 유래한 DC01027G01.ab12 서열을 BLAST 프로그램에서 nr DB로 단백질을 검색한 결과 애기장대 (*Arabidopsis thaliana*)의 known ORF (F14D16.29)와 높은 상동성을 나타내었다 (Figure 2). 이 경우 특이하게 ESI 서열과 일치를 보였던 부분인 FYITTNNN 서열은 known ORF (F14D16.29)에서는 동일한 부분에 존재하지 않았다. 그렇지

Table 1. Summary of ginseng protein identification by using ESI/Q-TOF/MS sequence tag, ginseng EST DB and nr DB.

No.	MW (D)	ESI/Q-TOF/MS sequence tag	Ginseng EST search	E-value	Protein id from EST	nr DB search from ESI sequence
1	939.422	TGGPFGTMoxR	DC02030B06.ab12	0.003	Ascorbate peroxidase	L-ascorbate peroxidase
2	1399.73 1423.76	ITSPLLEPSSVEK LFQVEYAIEAIK	DC03005F03.ab12	0.026	20S proteasome subunit PAF1	Proteasome subunit alpha type 5-1
3	1954.91 1970.89 1125.6 1850.8 1710.8 1726.8	ETAAVMQEFTQSGGVRPF ETAAVMoxQEFTQSGGVRPF EAIPVTGLVR DWQYSFSLTTFSPSGK ETAAVMQEFTQSGGVR ETAAVMoxQEFTQSGGVR	DC01017G12.ab10	5.1	Chain A, structure of restriction endonuclease	Proteasome subunit alpha type 2
4	1399.9 1976.2	VYYLTTNNNMR LNLQPHPEGGFYAETFR	DC01027G01.ab12	1.5	F14D16.29	Unknown
5	1167.57 1703.67 1149.6 1176.6 1387.8 1560.8	GTDGSDYIALR ~ DSSPTDDATDDYR YTSIK ~ R TSGGLLLTETK TTNLLLP ~ EDDILGILDT ~ K	DC02008C10.ab10	1e-4	Chaperonin 21 precursor	Chaperonin 10
6	1703.7 1334.6	~ PTDDATDDYR ~ HGLYPYNEK	DC02024E04.ab12	2e-4	Ribonuclease	Unknown
7	1120.56 1432.75 1448.72	LGDTLLEQGL YMVIQGEPGAVIR YMoxVIQGEPGAVIR	DC03005F02.ab10	0.66	19S proteosome subunit 9	Profilin
8	801.464 1739.79 755.78	AacETVVLK MEGVESFDIDI ~ MoxEGVESFDIDIDQNK	DC02007H10.ab12	1e-5	Copper homeostasis factor	Copper homeostasis factor
9	625.5 1352.68 1491.76	NDA ~ K FADEVN ~ K ~ EL ~				Unknown
10	1140.54 1310.8 2930.52	EHGAPEDETR QIPLIGSGSIIGR LSGSGGVS~	DC02011H01.ab12	5e-4	no hits	Superoxide dismutase
11	1588.0	NLNQQEVLMLLEK	DC01002E10.ab11	1.6	RNA polymerase	Polygalaturonase
12	1368.66 2083.03	[AS]TGTVVQ~R SANL~	DC02001A04.ab12	6.9	Protein disulfide isomerase precursor	Aminopeptidase
13	1771.86 2953.45	~LGGLSYDTD~ ~AIQALDQGDLHGR	DC02025H02.ab12	2.1	Expressed protein	RNA binding protein gend-id
14	1713.85	VFIGGLSYGTDDNSLR	DC01006A05.ab10	4e-7	Glycine-rich RNA-binding protein	RNA-binding protein
15	1527.98 1535.98 1936.18 1980.18 2243.18 1444.75	[PS]NSDT[Y]VLF[W]AK TLEFLAWLPVTN ~DDSSQLQTQAAEQFK ~PQEELLAAHL~ ~LAQDDEDVDET~ [LD]ELVFTQAGV~	DC01005D08.ab11	0.019	Nascent poly peptide associated complex alpha	Unknown
16	1070.53 998.6 1468.8 1608.8 1841.8	LTQYL[YN]R TFESVLFT [SL]VAVPLD~ [SQ]TFPAQLSGT~ [SVL]TDYDTLGSDD[SLL]K	DC02005G10.ab12	6.6	Phosphate-repressible acid phosphatase	Unknown

No.	MW (D)	ESI/Q-TOF/MS sequence tag	Ginseng EST search	E-value	Protein id from EST	nr DB search from ESI sequence
17	982.504 1070.58 1112.38 1467.78 1607.64 1840.78	TFESVLFR LTQYL~R [DV][DV]YLMoxDK [SL]VAVPLD[MT]V[AP]R [SG]ATFPAQLSG~ [GE]LTDYDTLGSD~AK	DC03005F08.ab10	6.8	Probable cation-transporting ATPase	Unknown
18	774.445 1554.78 1671.93	AFPQAIK LVTLGEASQFNTMK LYAGLLLDIDDILPK	DC01002E03.ab11	8e-8	Ribonuclease	Ribonuclease 1
19	1447.7 1554.77 1671.95 2105.04 648.49 1283.78 1360.78 1458.98 1398.88 1524.88 1570.98 1658.18	SSEIIEGDGGVGTVK LVTLGEASQFNTMoxK LYAGLLLDIDDILPK ~TIYNTIGDAVIPEENIK AVALFK TIVPTPDG~ TEVEATSTVPAQK VSEVIEGDGGVGTLK TIGDAVIPEENIK pGVG[TIGDATIPEENIK LVTLGEASQFNTYK LYAGLLLDIDTILPK	DC01002E03.ab11	8e-8	Ribonuclease	Ribonuclease
20	735.432 761.435 1324.66 1370.69 1457.81 3160.64	SIQLFK AFPEGIK GSFLDMoxDTVVPK TETQAISPVPAEK SVQVLEGDGGVGTIK ~ITQTTIYNTIGDAVIPEENIK	DC02008A04.ab12	3e-14	Ribonuclease	Ribonuclease 2
21	1033.58 1905.98 2052.01 2081.98	FALLVDDLK ~LLLSVA~ ~LETGGEFTVS~K ~LFKE~ELVDL~R	DC01010B12.ab12	6e-3	Thioredoxin peroxidase	Unknown
22	901.58 1446.78	VLQLS[GE]R ~EDGVLTVTVPK	DC02017B09.ab10	8e-4	Hsp20.1 protein	Unknown
23	10224.48 901.58 1462.78 1446.78	ETSAFVNTR VLQLSGER ASFEDGVLTVTVPK AAFEDGVLTVTVPK	DC02017B09.ab10	7e-5	Hsp20.1 protein	Unknown
24	1033.58 1060.38 1307.58 1434.78 2067.98 2509.98	FALLVDDLK TLLDLDNTR LQFE~YR [R]LLENELQTYR ~VETNEFTVSGDV[ET]LK ~MHNIMHELCDFESSGQK	DC01010B12.ab12	6e-3	Thioredoxin peroxidase	Unknown
25	1204.7 1429.64 894.38 945.38 1011.58 1154.58 1835.78 2005.98	~LIIGGTG~ ~PSEFGNDVDR SALLESFK EATLSLPSK VVLL~PK EDDIGT~K ~EVGFLAELL~ ~VLSTVGHAQLADQDK	DC02014C03.ab11	4e-3	Phenylcoumaran benzylic ether	Isoflavine reductase homolog BET V5
26	1363.72 1649.98 1444.86 1713.78	~QGLSIDEF~ ~VVANPANTNALILK ~GNGLVAYS~ ~ENGELVDL~	DC01008D08.ab11	0.035	Aytosolic malate dehydrogenase	Malate dehydrogenase, cytoplasmic

No.	MW (D)	ESI/Q-TOF/MS sequence tag	Ginseng EST search	E-value	Protein id from EST	nr DB search from ESI sequence
27	1302.67 1548.77 1367.58 1231.6 2005.9 1470.78	LLENLPGLMFR TENFNQNLQLSLK YIFLSEASASTIR VPIEVMoxGELK LTVELAEGPASADA [LE]GLGLVP~	DC01002B10.ab12	6e-7	Auxin-induced protein	Auxin-induced protein, aldo/keto reductase
28	1774.94	~LPDGQVITI~R	DC03001G04.ab10	4e-3	Actin [imported]	Actin
29	901.48 1446.78 1115.58	VLQISGER AAMoxEDGVLTVTVPK ELLPH~R	DC01022G09.ab10	9e-7	Small heat-shock protein class I, 18.6K	18.2kD class I small heat-shock protein
30	1091.58 1462.78	[SN]QQFQALR SSSGVGSTTSTA~PK	DC02003C03.ab11	0.2	RAD23 protein, isoform I	RAD23 protein isoform 1
31	1062.58	[LD]V[EN]H~	No hit			Reversibly glycosylated polypeptide-1

A

```
Query: VYYLTTNNNMR
       an ESI/Q-TOF/MS sequence tag at Spot #4 of Table 1
Subject: ginseng EST sequence, DC01027G01.ab12 (Length = 189)

Score = 24.0 bits (49), Expect = 1.5
Identities = 6/8 (75%), Positives = 7/8 (87%)

Query: 2   YYLTTNNN 9
           +Y TTNNN
Sbjct: 106 FYITTNNN 113
```

B

```
DC01027G01.ab12

TRETFRDKSVLLAKSQLPPQYKVDRAVSTNIYFLVPSGSVSHLHRI
PCSETWNFYLGDPLTVLEMNEEDGSVKLTTLGSDITGENQLLQYT
VPPNVWFGAFPAKDFYITTNNNAVKNPPRDAENQFSLVGCTCAP
AFEFADFELAKLSELVSRFPDHKSLVTLLTLPSXMXGFNYFSXSXI
LACCFKFX
```

Figure 1. The analysis results of amino acid sequence of ginseng protein spot #4 with ginseng EST DB (A) and the amino acid sequence of EST DC01027G01 (B). At the amino acid sequence of DC01027G01.ab12, the shadowed sequences are matched with the sequence, YYLTNNN, which is the data from Tandem mass spectrum. YTTNNN, a part of the sequence YYLTTNNN is completely aligned with the selected EST sequence. And the '+' sign means the amino acid Y has the similar chemical characteristic with the amino acid F of EST sequence. The E-Value of this alignment is 1.5 .

```
(AC068602) F14D16.29 [Arabidopsis thaliana]
        Length = 197

Score =  217 bits (553), Expect = 4e-56
Identities = 107/170 (62%), Positives = 132/170 (76%), Gaps = 6/170 (3%)

Query: 3   ETFRDKSVLLAKSQLPPQ-----YKVDRAVSTNIYFLVPSGSVSHLHRIPCSETWNFYLG 57
           ETFRD SV L+ SQLPP       KVDRAVST+IYFL+PSGSVS LHRIP +ETW+FYLG
Sbjct: 28  ETFRDSSVFLSTSQLPPTCSSLPLKVDRAVSTSIYFLLPSGSVSRLHRIPMAETWHFYLG 87

Query: 58  DPLTVLEMNEEDGSVKLTTLGSDITGENQLLQYTVPPNVWFGAFPAKDFYITTNNNAVKN 117
           +PLTV+E+ + DG +K T LG D+   +Q  QYTVPPNVWFG+FP KD + + +   +K
Sbjct: 88  EPLTVVELYD-DGKLKFTCLGPDLFEGDQKPQYTVPPNVWFGSFPTKDVHFSQDGALLKA 146

Query: 118 PPRDAENQFSLVGCTCAPAFEFADFELAKLSELVSRFPDHKSLVTLLTLP 167
             RD+EN FSLVGCTCAPAF+F DFELAK S+L+SRFP H+SL+T+L+ P
Sbjct: 147 EARDSENHFSLVGCTCAPAFQFEDFELAKRSDLLSRFPQHESLITMLSYP 196
```

Figure 2. The result of homology search of DC01027G01.ab12 (amino acid sequence) using the BLAST program.

만 그 외의 다른 부분들에서 매우 유사한 서열을 가지며 펩타이드를 구성하는 170개의 아미노산 중 62%인 107개의 아미노산이 완전히 일치하였다. 여기서 측정된 E-value는 4e-56으로 DC01027G01.ab12는 F14D16.29과 매우 유사함을 알 수 있다. 지금까지의 결과를 요약하면 ESI/Q-TOF/MS의 아미노산 서열 YYLTTNNN은 EST기원인 DC01027G01.ab12와 E-value가 1.5로 일치하며, EST 서열은 애기장대 기원인 F14D16.29와 E-value가 4.0e-56로 일치한다. 이로부터 아미노산 서열 YYLTTNNN이 포함된 4번 발현은 단백질 F14D16.29와 유사한 단백질이지만 펩타이드의 아미노산 서열이 너무 짧아서 BLAST 프로그램에서 직접 검색 결과를 얻지 못한 것으로 추측할 수 있다.

이러한 예에서 볼 수 있듯이 직접 ESI 서열을 nr DB에서 검색하여 유사 단백질을 찾지 못하는 경우에 EST 서열을 이용한 검색으로 단백질의 동정이 가능한 경우가 존재한다. ESI/Q-TOF/MS에서 얻은 아미노산 서열은 대부분이 35개보다 적은 아미노산으로 구성된 서열이다. nr DB는 non-redundant protein DB로서 단백질 발현 중에 나타나는 서열의 변화들을 포함시키지 않은 데이터베이스이다. 그러나, EST DB는 cDNA로부터 얻은 EST 서열로 구성하므로, 단백질 발현 과정에서 나타나는 서열의 다양성을 포함하고 있다. 따라서, nr DB에서는 유사한 서열이 검색되지 않았으나 EST DB에서 검색되는 경우들이 있었다. 이러한 관계에 의해 31개의

발현단백질 spot 중 ESI/Q-TOF/MS 아미노산 서열을 nr DB에서 검색하여 단백질 동정을 한 경우는 68%인 데 반하여, 인삼 EST DB 검색을 거쳐서 간접적인 방법으로 BLAST 검색을 실행한 경우는 90%에서 유사한 EST 서열 및 단백질을 찾을 수 있었다.

단백질 동정에 사용한 nr DB의 데이터 개수는 860,252개였으며, EST DB는 인삼 모상근에 대한 데이터로 5,373개의 EST 서열을 가지고 있다. 데이터베이스의 개수는 nr DB가 EST DB의 160배에 달하였다. 그러나, 인삼 EST DB는 nr DB와 달리 인삼에 제한된 데이터베이스이며, 인삼에서의 단백질 발현 시 서열 변화가 적용된 데이터베이스이므로 ESI/Q-TOF/MS에서 얻은 프로테옴 데이터와 일치되는 서열들을 nr DB에서보다 다량으로 발견할 수 있었다.

MALDI/TOF/MS 스펙트럼과 EST 데이터의 비교

MALDI/TOF/MS로부터 peptide mass fingerprinting 의 검색으로 얻은 펩타이드의 아미노산 서열들은 ESI/Q-TOF/MS 에서 얻어진 아미노산 서열과 비교할 때에 많은 차이를 보였다. 그러므로 탄뎀 질량 스펙트럼으로 아미노산 서열을 얻지 않고 MALDI/TOF/MS에서의 질량 스펙트럼만으로 단백질을 동정하는 것은 믿을 만한 결과를 주지 못한다. 본 연구에서는 MALDI/TOF/MS 결과의 신뢰도를 높이고 실험을 효과적으로 수행하기 위해 가능한 한 ESI/Q-TOF/MS 실험을 거치지 않고 MALDI/TOF/MS 실험만으로 단백질을 밝혀낼 수 있는 방법을 알아보고자 하였다. EST 서열 데이터베이스를 단백질 동정에 활용할 수 있는지의 여부를 판단하고자, MALDI/TOF/MS 질량 스펙트럼의 EST 서열 데이터베이스에 의한 분석을 시도하였다.

인삼 EST 데이터베이스에서 MALDI/TOF/MS의 스펙트럼을 검색한 결과는 nr DB에서의 경우와 비슷하였다. 질량 스펙트럼은 아미노산의 구성이 비슷한 서열을 찾지 않고 분자량이 같은 서열을 찾으므로, 실제 아미노산 서열과 일치하는 EST 서열은 여러 개의 다른 서열들과 함께 검색결과 속에 포함되어 있게 된다. 따라서, 아미노산 서열에 대한 추가적인 정보 없이 MALDI/TOF/MS 데이터의 EST 검색만으로 원하는 단백질 서열 하나를 찾기는 쉽지 않았다. 단, 앞서 ESI/Q-TOF/MS의 아미노산 서열에서 언급한 바와 같이 nr DB와 EST DB의 구조적인 차이로, nr DB에서 mass fingerprint를 검색하는 것보다는 EST 데이터베이스에서의 검색이 단백질 동정 결과의 신뢰도를 높이는 효과가 있었다.

MALDI/TOF/MS의 mass fingerprint 검색에 대해서는 해당 스펙트럼을 만족시키는 가능한 아미노산 서열들 중에서 신뢰성 있는 서열을 가려내기 위한 대안으로, MALDI/TOF/MS 스펙트럼을 nr DB에서 검색하여 예상된 아미노산 서열과 EST DB에서 검색하여 예상된 아미노산 서열을 서로 비교하여 일치하는 경우들을 찾아내었다.

Table 2에서는 인삼 모상근의 Silver-stained 2-D PAGE로부터 분리한 단백질들을 trypsin으로 가수분해한 뒤 MALDI/TOF/MS로 구한 펩타이드 분자량 프로파일을 nr DB로 검색한 결과와 인삼 EST DB로 검색한 결과를 비교하였다. nr DB에 의한 검색은 University of California, San Francisco의 MS-FIT 프로그램을 사용하였으며, EST DB 에 의한 검색은 EMBOSS package의 eMOWSE 프로그램을 활용하였고, 자체 제작 프로그램에 의해 검색 결과를 확인하였다. Trypsin에 의한 단백질 분해에서 missed cleavage 는 한 개 이하인 경우만을 검색하였고, 결과 중 nr DB에서 예상된 아미노산 서열과 EST DB에서 예상된 아미노산 서열이 일치되는 9개의 spot에 대한 예상 아미노산 서열들을 Table 2에 표시하였다.

단백질에 대한 서열을 제공하는 nr DB에서와는 달리, EST 서열은 전체 단백질의 일부분에 대한 서열이다. 그러므로, 인삼 EST DB를 활용한 분자량 프로파일의 검색에서는 질량 스펙트럼에서 얻은 분자량에 해당되는 펩타이드 중의 일부분만 EST 서열에 포함되고 나머지 부분은 포함되지 않는 경우들이 존재한다. 따라서 질량 스펙트럼과 일치되는 펩타이드의 개수도 nr DB 검색결과에 비해 상대적으로 적었다. 이러한 이유로 질량 스펙트럼의 peak들을 동시에 만족시키는 서열은 EST 데이터베이스 중에서는 찾을 수 없었으며, peak들 중 일부분만이 일치하는 서열들은 EST DB에서 매우 많이 발견할 수 있었다. 단백질 정보를 얻기 위한 결과 분석에서는 여러 개의 중복된 EST 결과로부터 특정한 검색결과를 선택하여 사용하였다. 선택 기준은 우선 EST 결과 중에서 Start, Stop codon이 포함되지 않은 정상적인 아미노산 서열만을 골라내었으며, 이 중에 여러 개의 아미노산 서열이 일치되며 missed cleavage가 적은 서열로서 EST DB 중에서 많이 존재하는 것을 선택하였다.

분자량 프로파일의 DB 검색에서는 EST DB에 의한 단백질 동정 결과와 nr DB의 결과가 서로 일치하지 않는 경우가 81% 였는데, 이 때에는 탄뎀질량 스펙트럼과 같이 다른 방법으로의 추가분석에 의해 검색 결과의 신뢰도 측정이 필요하다.

고 찰

프로테옴 데이터는 유전자가 mRNA를 거쳐서 단백질로 발현된 후에 얻는 데이터로 유전자의 염기서열과는 차이가 있다. 우선 단백질 발현 단계까지 오는 동안의 변화가 생길 수 있으며, 주위 환경에 따라서 발현되는 양이 달라지므로, 유전자 정보에 대한 연구결과들과 프로테옴 데이터와는 여러 면에서 다른 양상을 보인다. 이러한 점을 감안하여 mRNA에서 cDNA를 생성하여 얻는 EST 데이터를 바탕으로 프로테옴 데이터를 분석함은 유전자 데이터베이스로 직접 프로테옴 데이터를 검색하는 방법에 비해 훨씬 효과적인 분석방법이 될

Table 2. Analysis of MADLI/TOF/MS spectrums of ginseng proteins with nr DB and ginseng EST DB.

No.	Molecular ion observed (D)	Sequence from nr DB	Sequence from ginseng EST DB	EST sequence	Protein ID
1	923.429	TGGPFGTMR	TGGPFGTMR	DC02016C 04.ab11	L-ascorbate peroxidase
	939.422	TGGPFGTMoxR			
	1611.92	ALLDDPVFRPRVEK	ALLDDPVFRPLVEK		
	2090.92	GCDHLRDVFAKQMGLSDK	YAADEDAFFVDYAESHLK		
2	1325.63	IEFPHPTEEAR	IEFPHPTEEAR	DC02010A 04.ab10	protease regulatory subunit 6A
	1453.77	KIEFPHPTEEAR	KIEFPHPTEEAR		
	1886.92	DSYLILDTLPSEYDSR			
	1969.94	TMLELLNQLDGFSSDDR			
	2257.08	DATEVNHEDFNEGIIQVQAK	DATEVNHEDFNEGIIQVQAK		
3	774.445	AFPQAIK	AFPQAIK	DC01002E 03.ab11	ribonuclease 1
	1554.78	LVTLGEASQFNTMK			
	1671.93	LYAGLLLDIDDILPK	LYAGLLLDIDDILPK		
	963.54				
	1283.78				
	1360.78				
	1447.78		SSEIIEGDGGVGTVK		
	2105.18		NTTIYNTIGDAVIPEENIK		
4	774.449	AFPQAIK	AFPQAIK	DC01002E 03.ab11	ribonuclease 1
	1447.7	SSEIIEGDGGVGTVK	SSEIIEGDGGVGTVK		
	1554.77	LVTLGEASQFNTMKLYA			
	1671.95	GLLLDIDDILPK	LYAGLLLDIDDILPK		
	2105.04	NTTIYNTIGDAVIPEENIK	NTTIYNTIGDAVIPEENIK		
			AGLIFK		
	648.49				
	1283.78				
	1360.78				
	1458.98				
	1398.88				
	1524.88				
	1570.98				
	1658.18				
5	735.432	SIQLFK	SIQLFK	DC01008A 03.ab11	ribonuclease 2
	761.435	AFPEGIK	AFPEGIK		
	1324.66	GSFLDMDTVVPK			
	1370.69	TETQAISPVPAEK	TETQAISPVPAEK		
	1457.81	SVQVLEGNGGVGTIK	SVQVLEGNGGVGTIK		
	3160.64	IVPTDGGSTITQTTIYNTI	IVPTDGGSTITQTTIYNTI		
		GDAVIPEENIK	GDAVIPEENIK		
6	1084.59	SGHPTFALVR	SGHPTFALVR	DC02014E 06.ab11	Isoflavone reductaase homolog BET V5
	1204.7	ILIIGGTGYIGK	ILIIGGTGYIGK		
	1304.73	EKVVIFGDGNAR			
	1429.64	FFPSEFGNDVDR			
	894.38		SAIIESFK		
	945.38		EATLSIPSK		
	1011.58		VVILGDGNPK		
	1154.58		EDDIGTYTIK		
	1835.78				
	2005.98				
7	873.484	ALGQISER	ALGQISER	DC03010G 10.ab10	Malate dehydrogenase, cytoplasmic
	1346.71	MELVDAAFPLLK	MELVDAAFPLLK		
	1363.72	IVQGLSIDEFSR			
	1643.76	LSSALSAACDHIR			
	1649.98	VLVVANPANTNALILK	VLVVANPANTNALILK		
	2016.15	VLVTGAAGQIGYALVPMIAR			
	1444.86				
	1713.78				

No.	Molecular ion observed (D)	Sequence from nr DB	Sequence from ginseng EST DB	EST sequence	Protein ID
8	1140.54	EHGAPEDETR	EHGAPEDETR	DC03007G 03.ab12	superoxide dismutase
	1310.8	QIPLIGSGSIIGR			
	2345.08	HAGDLGNVTVGEDGTAE FTIVDK	HAGDLGNVTVGEDGTAE FTIVDK		
	2930.52	AVTVLSGSGGVSGVIHFTQ EEDGPTTVTGK			
9	976.43	AGFAGDDAPR		DC03005B 05.ab10	actin
	1192.53	GYMFTTTAER			
	1445.68	GEYEDSGPSIVHR	GEYDESGPSIVHR		
	1884.94	NYELPDGQVITIGAER	NYELPDGQVITIGAER		
	1954.06	VAPEEHPVLLTEAPLNPK			
	2199.03	DLYGNIVLSGGSTMFPGIADR			
	3151.67	TTGIVLDSGDGVSHTVPIY EGYALPHAILR			

수 있다.

본 연구에서는 MALDI/TOF/MS에서 측정된 질량 스펙트럼을 인삼 EST 데이터베이스에서 검색하여 일치하는 EST 서열을 찾아내었으며, ESI/Q-TOF/MS의 탄뎀 질량 스펙트럼에서 얻은 펩타이드의 아미노산 서열을 인삼 EST 데이터베이스에서 검색하고 여기에서 얻은 EST 서열을 GenBank에서 검색하여 단백질을 동정하는 방법에 대한 분석을 하였다. 프로테옴 분석 방법은 이차원 전기영동 겔에 의한 spot의 비교분석, 질량 스펙트럼에 의한 peptide mass fingerprint 분석, 탄뎀 질량 스펙트럼에 의한 아미노산 서열분석으로 크게 세 가지로 구분된다.

1단계인 이차원 전기영동 겔의 분석에서는 환경 변화에 따른 단백질의 발현 정도를 측정할 수 있으나, 단백질 동정을 위해서는 질량 스펙트럼이 필요하다. 정량적인 분석도 대부분의 경우 spot의 분리 상태가 명확치 않으므로 정확한 분석이 어렵다. 2단계인 질량 스펙트럼에 의한 peptide mass *fingerprint* 분석은 아미노산 서열 중에 데이터베이스의 서열과 한 아미노산이라도 차이가 나는 경우에는 옳은 결과를 얻을 수 없으므로, 데이터의 신뢰도가 문제가 된다. 이에 대한 보완으로는 중복성이 있을지라도 EST 데이터베이스에서 질량 스펙트럼의 검색을 하여야만 해당 단백질 정보를 얻을 수 있다. 3단계인 탄뎀 질량 스펙트럼에 의한 아미노산 서열의 BLAST 검색은 1단계, 2단계의 데이터들에 비해 비교적 정확한 결과를 제공한다. 그러나, 서열의 길이가 매우 짧으므로 단백질의 극히 일부분에서 서열의 유사성을 검사하게 된다. 따라서 전체적으로는 유사한 단백질인데 탄뎀 질량 스펙트럼에서 얻은 결과만으로는 유사성을 밝힐 수 없는 경우들이 발생한다. EST 데이터베이스에 의한 아미노산 서열 검색은 이러한 문제점에 대한 해결 방안을 제시하였다. 다양한 발현 양상을 포함하는 EST 데이터베이스로부터 아미노산 서열과 유사한 데이터를 찾아낸 뒤에 이를 *nr DB*와 비교하는 방법을 사용함으로써 탄뎀 질량 스펙트럼 결과에서 펩타이드의 길이가 짧은 단점을 극복할 수 있었다.

인삼에서 단백질의 동정을 위한 프로테옴 연구는 인삼 EST 데이터를 활용하여 보다 향상된 결과를 얻을 수 있었다. 그러나, 환경 변화에 따른 단백질의 발현 정도를 비교하는 정확한 정량적 분석은 아직 기술적으로 미흡한 단계이다. 최근 몇 년간 활발하게 연구되고 있는 reagent와 tagging 기법 (Gygi et al. 1999)에 대한 적극적인 수용 및 연구개발에 의해 단백질의 정량 분석에 대한 기술개발을 추진한다면, 질량분석기를 활용한 프로테옴 실험 데이터들이 단백질 기능 연구의 기초 자료로의 역할을 충분히 수행할 수 있을 것이다.

적 요

인삼 모상근의 프로테옴 분석에 의해 얻은 질량분석 스펙트럼 데이터는 MALDI/TOF/MS에서 얻는 질량 스펙트럼과 ESI/Q-TOF/MS에서 얻는 탄뎀 질량 스펙트럼으로 구분된다. 질량 스펙트럼은 단백질이 효소에 의해 분해된 펩타이드들의 분자량 정보를 제공하며, 탄뎀 질량 스펙트럼에서는 아미노산 단위로 분해된 절편 단백질의 분자량으로부터 아미노산 서열을 결과로 얻는다. 펩타이드의 아미노산 서열을 BLAST로 검색하면 유사한 단백질을 GenBank에서 검색할 수 있다. 이러한 단백질 동정 방법은 완전한 유전체 서열이 알려진 생물체의 경우 높은 정확도로 단백질을 동정할 수 있으나, 그렇지 않은 경우는 유사한 단백질이 데이터베이스에 존재하지 않아 분석이 용이하지 않다.

본 연구에서는 질량 스펙트럼 및 절편 단백질의 아미노산 서열을 EST (expressed sequence tag) 서열과 비교하여 프로테옴 데이터와 일치하는 EST 서열을 찾아내고 이를 BLAST 검색에 의해 단백질 동정에 활용하였다. ESI/Q-TOF/MS 에서 얻은 아미노산 서열은 길이는 짧지만 데이터의 신뢰도가 높으므로 EST 서열과의 연관 관계를 밝힘으로써 단백질에

대한 정보를 보완할 수 있었다. ESI/Q-TOF/MS에서 얻은 펩타이드의 아미노산 서열을 EST 서열과 비교한 결과 90%의 아미노산 서열이 EST DB에서 발견되었다. NCBI의 nr 데이터베이스에서 아미노산 서열을 검색하여 찾은 단백질이 68%임에 비하여, 인삼 EST 서열에 의한 검색이 22% 더 많은 결과를 얻었다.

MALDI/TOF/MS의 질량 스펙트럼에서 nr 데이터베이스로 검색한 결과와 인삼 EST 데이터베이스를 검색한 결과가 일치하는 경우는 47개 중 9개인 19%에 불과하여, 탄뎀 질량분석으로 아미노산 서열을 얻지 않고, 단지 질량 스펙트럼으로부터 단백질을 동정하는 방법으로는 단백질 동정의 정확한 결과를 기대하기 어려움을 확인하였다.

사사 - 본 연구는 자생식물이용기술 개발사업단 (21C 프론티어 연구개발사업)의 연구비에 의해 수행되었음.

인용문헌

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. **215**:403-410

Ashton PD, Curwen RS, Wilson RA (2001) Linking proteome and genome: how to identify parasite proteins. Trends Parasitol. **17**:198-202

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. **17**: 994-999

Kim SI, Kim SJ, Nam MH, Seo JB, Kim S, Kwon KH, Kim YH, Choi JS, Yoo JS, Yang DC, Choi KT, Park YM (2001) Purification of Crude Protein Mixture from Panax ginseng and Hairy Root for Proteome Analysis. Korean J. Plant Tissue Culture **28**:347-351

Lisacek FC, Traini MD, Sexton D, Harry JL, Wilkins MR (2001) Strategy for protein isoform identification from expressed sequence tags and its application to peptid mass fingerprinting. Proteomics **1**:186-193

Mann M (1996) A shortcut to interesting human genes : peptide sequence tags, expressed-sequence tags and computers. Trends Biochem. Sci. **21**:494-495

Mathesius U, Keijzers G, Natera SHA, Weinman JJ, Djordjevic MA, Rolfe BG (2001) Establishment of a root proteome reference map for the model legume Medicago truncatula using the expressed sequence tag database for peptide mass fingerprinting. Proteomics **1**:1424-1440

Nyman TA (2001) The role of mass spectrometry in proteome studies. Biomol. Eng. **18**:221-227

Pappin DJC, Hojrup P, Bleasby AJ (1993) Rapid Identification of Proteins by Peptide-Mass Fingerprinting. Current Biol. **3**:327-332

Porubleva L, Velden KV, Kothari S, Oliver DJ, Chitnis PR (2001) The proteome of maize leaves: Use of gene sequences and expressed sequence tag data for identification of proteins with peptide mass fingerprints. Electrophoresis **22**:1724-1738

Taylor JA, Johnson RS (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Comm. Mass Spec. **11**:1067-1075

(접수일자 2002년 7월 22일)