

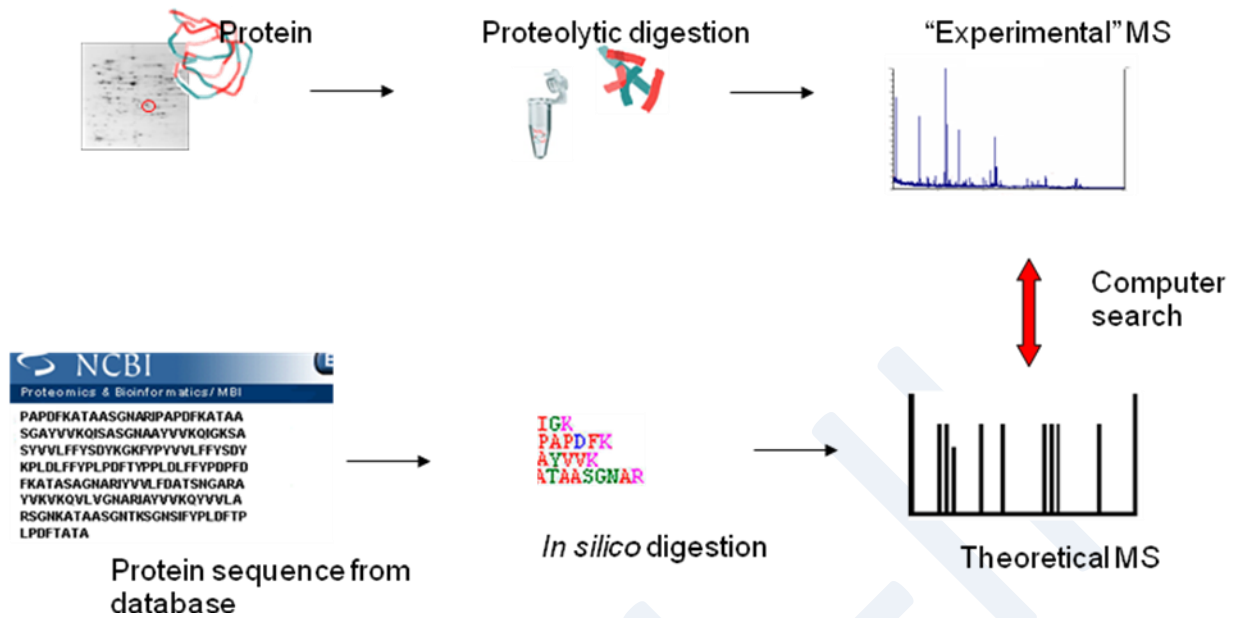
## Lecture 17: Strategies for protein identification

Protein identification may be done by various methods:

1. Peptide mass fingerprinting
2. Protein sequencing by mass spectrometry
3. Protein sequencing by chemical methods

### Peptide mass fingerprinting

We have studied two-dimensional gel electrophoresis in earlier lectures. We have also studied how over or under expression of a protein spot is identified. Once protein spot of interest is identified, the same is excised and digested by specific protease such as trypsin. As trypsin is very specific in the cleavage each protein sequence will result in specific set of peptides of varying masses that are characteristic of that protein. The mass of each peptide will be the sum of the amino acids present including any modifications that those amino acids might have undergone. The masses of these peptides are measured by mass spectrometry and then *in silico* compared to either a database containing known protein sequences or even the genome. Computer programs translate the known genome of the organism into proteins, then theoretically cut the proteins into peptides with the same protease (for example trypsin), and calculate the absolute masses of the peptides from each protein. They then compare the masses of the peptides of the unknown protein to the theoretical peptide masses of each protein encoded in the genome. The results are statistically analyzed to find the best match (Fig. 1)



**Figure 1:** Various steps of Peptide mass fingerprinting

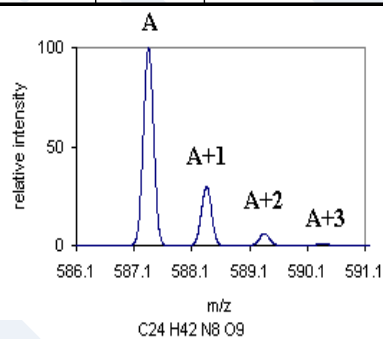
Although, the method looks very simple, there are certain limitations

- Peptide mass fingerprinting algorithms assumes the peptides come from a single protein. If there is any impurity the search result will not work.
- If protein has post translational modification, identification by Peptide mass fingerprinting does not work. As the methods assumes that the peptide mass is coming from amino acids only. Moreover, when genome data is virtually converted to protein, post-translational modification is difficult to precisely predict to match with experimental data.
- Although, specific proteases are available for protein cleave but sometime they cleave at wrong site (Mis-cleavage) which results in false prediction.

Each data peak in mass spectra may have more than one peak as atoms have several isotopes (Table 1). These peaks called isotopic peaks. For example a fragment of protein with mass  $M$ , when analyzed by mass spectrometry will give a major peak corresponding to mass  $M$  (Monoisotopic peak) but a small fraction will show  $M+1$  mass because a small fraction of protein molecule may have  $^{13}\text{C}$  isotope. While searching database for protein identification (Fig. 2), one must mention the peak corresponds to monoisotopic or average. Some additional details about the protein, if known, like molecular mass, modifications etc may further help in protein identification.

**Table 1:** Isotopes of few common atoms

Isotope (A)	mass	Abundance, %	Isotope (A+1)	mass	Abundance, %	Isotope (A+2)	mass
$^{12}\text{C}$	12	98.93	$^{13}\text{C}$	13.0033548378	1.07	$\text{C}^{14}$	14.003241988
$^1\text{H}$	1.0078250321	99.9885	$^2\text{H}$	2.0141017780	0.0115	$^3\text{H}$	3.0160492675
$^{14}\text{N}$	14.0030740052	99.632	$^{15}\text{N}$	15.0001088984	0.368		
$^{16}\text{O}$	15.9949146221	99.757	$^{17}\text{O}$	16.99913150	0.038	$^{18}\text{O}$	17.9991604



## MASCOT Peptide Mass Fingerprint

<b>Your name</b>	Vikash Kumar Dubey	<b>Email</b>	vdubey@iitg
<b>Search title</b>	Teaching		
<b>Database(s)</b>	SwissProt NCBIInr contaminants cRAP MSDB	<b>Enzyme</b>	Trypsin
		<b>Allow up to</b>	1 missed cleavages
<b>Taxonomy</b>	All entries		
<b>Fixed modifications</b>	--- none selected ---	> <	Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-term) Ammonia-loss (N-term C) Biotin (K) Biotin (N-term) Carbamidomethyl (C) Carbamyl (K) Carbamyl (N-term)
	Display all modifications <input type="checkbox"/>		
<b>Variable modifications</b>	--- none selected ---	> <	
<b>Protein mass</b>	<input type="text"/> kDa	<b>Peptide tol. ±</b>	1.2 Da
<b>Mass values</b>	<input checked="" type="radio"/> MH <sup>+</sup> <input type="radio"/> M <sub>r</sub> <input type="radio"/> M-H <sup>-</sup>	<b>Monoisotopic</b>	<input checked="" type="radio"/> Average <input type="radio"/>
<b>Data file</b>	<input type="text"/> Browse...		
<b>Query</b>	Cut and paste you peak values or load data file NB Contents of this field are ignored if a data file is specified. 1051.54 1086.52 1094.56 1111.59 1244.64		
<b>Decoy</b>	<input type="checkbox"/>	<b>Report top</b>	AUTO hits
<b>Start Search ...</b>		<b>Reset Form</b>	

**Figure 2:** A snapshot of MASCOT search window

([http://www.matrixscience.com/cgi/search\\_form.pl?FORMVER=2&SEARCH=PMF](http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF))

# *MATRIX* SCIENCE Mascot Search Results

User : Proteomics & Bioinformatics  
 Email : MBI@Helsinki.fi  
 Search title : Peptide Mass Fingerprint Example  
 Database : SwissProt 52.1 (261513 sequences; 95638062 residues)  
 Timestamp : 29 Mar 2007 at 12:10:41 GMT  
 Top Score : 112 for **PML\_HUMAN**, Probable transcription factor PML (Tripart

## Probability Based Mowse Score

Protein score is  $-10 \cdot \log(P)$ , where P is the probability that the observed match is a random event. Protein scores greater than 67 are significant ( $p < 0.05$ ).

## Mascot PMF results

Entry name	Mass	Score	Expect	Queries matched
<a href="#">PML_HUMAN</a>	97455	112	1.7e-06	15
Probable transcription factor PML (Tripartite motif-containing protein 19) (RING finger domain) (RING finger protein 19) (RING finger protein 19)				
<a href="#">ANMK_RHOP2</a>	38245	51	2.2	7
Anhydro-N-acetylmuramic acid kinase (EC 2.7.1.-) (AnhMurNAc kinase) - Rhodospirillum rubrum				
<a href="#">PURA_THEVO</a>	18601	47	4.9	6
Purification complex GINS protein 2 (Yeast) (Tetrahymena thermophila)				
<a href="#">PML_HUMAN</a>	97455	112	1.7e-06	15
Probable transcription factor PML (Tripartite motif-containing protein 19) (RING finger domain) (RING finger protein 19) (RING finger protein 19)				
<a href="#">ANMK_RHOP2</a>	38245	51	2.2	7
Anhydro-N-acetylmuramic acid kinase (EC 2.7.1.-) (AnhMurNAc kinase) - Rhodospirillum rubrum				
<a href="#">PSF2_DEBHA</a>	18601	47	4.9	6
Purification complex GINS protein 2 (Yeast) (Tetrahymena thermophila)				
<a href="#">PURA_THEVO</a>	48976	40	26	6
Adenylosuccinate synthetase (EC 6.3.4.4) (IMP--aspartate ligase) (AdSS) (AMPSase) - Thermoplasma volcanum				
<a href="#">KSGA_BRUME</a>	30303	40	26	6
Dimethyladenosine transferase (EC 2.1.1.-) (S-adenosylmethionine-6-N', N'-adenosyl-tRNA synthetase) (KsgA) (KsgA)				
<a href="#">GNAS3_BOVIN</a>	27477	38	42	6
Neuroendocrine secretory protein 55 (NESP55) [Contains: LSAL tetrapeptide; GAIPRRH peptide] (NESP55) (NESP55)				
<a href="#">Y4GB_RHISN</a>	16103	37	52	4
Hypothetical 16.1 kDa protein y4gB - Rhizobium sp. (strain NGR234)				
<a href="#">SYQ_HUMAN</a>	87743	36	74	9
Glutamyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS) - Homo sapiens				
<a href="#">YDIU_SHISS</a>	54287	35	75	6
UPF0061 protein ydiU - Shigella sonnei (strain Ss046)				

# Mascot protein view

## Protein View

Match to: [PML\\_HUMAN](#) Score: 112 Expect: 1.7e-06  
 Probable transcription factor PML (Tripartite motif-containing protein 19)

Nominal mass ( $M_r$ ): 97455; Calculated pI value: 5.88

NCBI BLAST search of [PML\\_HUMAN](#) against nr

Unformatted [sequence string](#) for pasting into other applications

Taxonomy: [Homo sapiens](#)

Cleavage by Trypsin: cuts C-term side of KR unless next residue is P

Number of mass values searched: 18

Number of mass values matched: 15

Sequence Coverage: 22%

coverage

Matched peptides shown in **Bold Red**

```

1  MEPAPARSSPR  PQQDPARPQE  PTMPPPETPS  EGRQPSPSPS  PTERAPASEE
51  EFQFLRCQQC  QAEAKCPKLL  PCLHTLCSGC  LEASGMQCPI  CQAFWPLGAD
101 TPALDNVFFE  SLQRRLSVYR  QIVDAQAVCT  RCKESADFWC  FECEQLLCAK
151 CFEAHQWFLK  HEARPLAELR  NQSVREFLDG  TRKTNNIFCS  NPNHRTPTLT
201 SIYCRGCSKP  LCCSCALLDS  SHSELKCDIS  AEIQQRQEEL  DAMTQALQEQ
251 DSAFGAVHAQ  MHAAVGQLGR  ARAETEELIR  ERVRQVVAHV  RAQERLEEA
301 VDARYQRDYE  EMASRLGRLD  AVLQIRITGS  ALVQRMKCYA  SDQEVLDMHG
351 FLRQALCLRL  QEEPQSLQAA  VRTDGFDEFK  VRLQDLSSCI  TQGKDAVSK
401 KASPEAASTP  RDPIDVDLPE  EAERVKAQVQ  ALGLAEAQPM  AVVQSVPGAH
451 PVPVYAFSIK  GPSYGEDVSN  TTTAQKRKCS  QTQCPRKVIK  MESEEGKEAR
501 LARSSPEQPR  PSTSKAVSPP  HLDGPPSPRS  PVIGSEVFLP  NSNHVASGAG
551 EAERVVVVIS  SSEDSDAENS  SSRELDDSSS  ESSDLQLEGP  STLRVLDENL
601 ADPQAEDRRPL  VFFDLKIDNE  TQKISQLAAV  NRESKFRVVI  QPEALFSIYS
651 KAVSLEVGLQ  HFLSFLSSMR  RPILACYKLW  GPGLPNFFRA  LEDINRLWEF
701 QEAISGFLAA  LPLIRERVPG  ASSFKLKNLA  QTYLARNMSE  RSAMAAVLAM
751 RDLCRLLEVS  PGPQLAQHVY  PFSSLQCFAS  LQPLVQAAVL  PRAEARLLAL
801 HNVSFMELLS  AHRRDRQGGL  KKYSRYLSLQ  TTTLPPAQPA  FNLQALGTYF
851 EGLEGPALA  RAEGVSTPLA  GRGLAERASQ  QS
    
```

% of protein length covered by the experimental peptides

## Home Assignment

**From a peptide mass fingerprinting experiment you have following monoisotopic peaks. Using peptide mass fingerprinting programs identify the protein. What is the confidence level of your identification (with explanation)?**

1051.54  
1086.52  
1094.56  
1111.59  
1244.64  
1421.7  
1476.67  
1542.84  
1613.88  
1664.97  
1763.79  
1777.82.

***Submit your assignment to course developer by e-mail;  
you shall get e-mail reply with grading and feedback in a  
week time.***

NPTEL