



There are three proven ways of using mass spectrometry data for protein identification. The first of these is known as a peptide mass fingerprint. This was the original method to be developed, and uses the molecular weights of the peptides resulting from digestion of a protein by a specific enzyme.



Peptide mass fingerprinting can only be used with a pure protein or a very simple mixture. The starting point will often be a spot off a 2D gel. The protein is digested with an enzyme of high specificity; usually trypsin, but any specific enzyme can be used. The resulting mixture of peptides is analysed by mass spectrometry. This yields a set of molecular mass values, which are searched against a database of protein sequences using a search engine. For each entry in the protein database, the search engine simulates the known cleavage specificity of the enzyme, calculates the masses of the predicted peptides, and compares the set of calculated mass values with the set of experimental mass values. Some type of scoring is used to identify the entry in the database that gives the best match, and a report is generated. I will discuss the subject of scoring in detail later.



If the mass spectrum of your peptide digest mixture looks as good as this, and it is a single protein, and the protein sequence or something very similar is in the database, your chances of success are very high.

Before searching, the spectrum must be reduced to a peak list: a set of mass and intensity pairs, one for each peak.

In a peptide mass fingerprint, it is the mass values of the peaks that matter most. The peak area or intensity values are a function of peptide basicity, length, and several other physical and chemical parameters. There is no particular reason to assume that a big peak is interesting and a small peak is less interesting. The main use of intensity information is to distinguish signal from noise.

Mass accuracy is important, but so is coverage. Better to have a large number of mass values with moderate accuracy than one or two mass values with very high accuracy.



There is a wide choice of PMF servers on the web. I hope this is a complete list, in alphabetical order. If I am missing a public server, please let me know, and I will add it to the list.

Many other PMF programs have been described in the literature. Most packages are either available for download from the web or are commercial products.



This is the Mascot search form for a peptide mass fingerprint. Besides the MS data, a number of search parameters are required. Some search engines require fewer parameters, others require more. We'll be discussing most of these search parameters in detail in later talks.

To perform the search, you paste your peak list into the search form, or upload it as a file, provide values for the search parameters, and press the submit button.



A short while later, you will receive the results.

A peptide mass fingerprint search will almost always produce a list of matching proteins, and something has to be at the top of that list. So, the problem in the early days of the technique was how to tell whether the top match was "real", or just the top match ... that is, a false positive.

There have been various attempts to deal with this problem, which I will describe when we come to discuss scoring.

Focus: Account and	PROTEOMICS
Protein Identification: The Origins of Peptide Mass Fingerprinting	
William J. Henzel and Colin Watanabe Protein Chemistry Department and Bioinformatics Department, Genentech, Inc., South San Francisco, California, USA	
John T. Stults Analytical Sciences Department, Biospect, Inc., South San Francisco, California, USA	
Peptide mass fingerprinting (PMF) grew from a need for a faster, more efficient method identify frequently observed proteins in electrophoresis gels. We describe the genesis of idea in 1989, and show the first demonstration with fast atom bombardment mass spect metry. Despite its promise, the method was seldom used until 1992, with the coming significantly more sensitive commercial instrumentation based on MALDI-TOF-MS. Y recount the evolution of the method and its dependence on a number of technical bree throughs, both in mass spectrometry and in other areas. We show how it laid the foundat for high-throughput, high-sensitivity methods of protein analysis, now known as proteom: We conclude with recommendations for further improvements, and speculation of the role PMF in the future. (J Am Soc Mass Spectrom 2003, 14, 931–942) © 2003 American Society Mass Spectrometry	to the ro- of ak- ion ics. of for
≻Henzel, W. J., Watanabe, C., Stults, J. T., <i>JASMS</i> 2003, <i>14</i> , 931	-942.
COT : Introduction © 2007-2010 Matrix Science	ce (MATRIX SCIENCE)

If you want to learn more about the origins of peptide mass fingerprinting, I strongly recommend this review by the Genentech group. They discuss the history and the methodology in a very readable style.



One of the strengths of PMF is that it is an easy experiment that can be performed using just about any mass spectrometer. The whole process is readily automated and MALDI instruments, in particular, can churn out high accuracy PMF data at a very high rate.

In principal, it is a sensitive technique because you are analysing all of the peptides from the digest. It doesn't matter too much if a small part of the protein fails to digest or some of the peptides are insoluble or don't fly very well.

One of the limitations is that you need a database of proteins or nucleic acid sequences that are equivalent to proteins, e.g. mRNAs. In most cases, you will not get satisfactory results from an EST database, where most of the entries correspond to protein fragments, or genomic DNA, where there is a continuum of sequence, containing regions coding for multiple proteins as well as non-coding regions.

This is because the statistics of the technique rely on the set mass values having originated from a defined protein sequence. If multiple sequences are combined into a single entry, or the sequence is divided between multiple entries, the numbers may not work.

If the protein sequence, or a near neighbour, is not in the database, the method will fail. It is not a method for protein characterisation, only for identification.

The most important limitation concerns mixtures. If the data quality is good, then one or two, possibly three, major components can be identified. But if the data are poor, it can be difficult to get any match at all out of a mixture, and it is never possible to identify a minor component with any confidence. This is the Achilles' heel of PMF.

To identify proteins from mixtures reliably, it is necessary to work at the peptide level. That is, using MS/MS data.



The experimental workflow for database matching of MS/MS data is similar to that for PMF, but with an added stage of selectivity and fragmentation.

Again, we start with protein, which can now be a single protein or a complex mixture of proteins. We use an enzyme such as trypsin to digest the proteins to peptides. If it is a complex mixture, such as a whole cell lysate, we will probably use one or more stages of chromatography to regulate the flow of peptides into the mass spectrometer. We select peptides one at a time using the first stage of mass analysis. Each isolated peptide is then induced to fragment, possibly by collision, and the second stage of mass analysis used to collect an MS/MS spectrum.

Because we are collecting data from isolated peptides, it makes no difference whether the original sample was a mixture or not. We identify peptide sequences, and then try to assign them to one or more protein sequences. One consequence is that, unless a peptide is unique to one particular protein, there may be some ambiguity as to which protein it should be assigned to.

For each MS/MS spectrum, we use software to try and determine which peptide sequence in the database gives the best match. This will involve simulating the cleavage specificity of the enzyme, followed by calculation of the mass values we expect to result from the gas phase fragmentation of the peptide.

Unlike a peptide mass fingerprint, use of a specific enzyme is not essential. By looking at all possible sub-sequences of each entry that fit the precursor mass, it is possible to match peptides when the enzyme specificity is unknown, such as endogenous peptides.



Database matching of MS/MS data is only possible because peptide molecular ions fragment at preferred locations along the backbone. In many instruments, the major peaks in an MS/MS spectrum are b ions, where the charge is retained on the N-terminus, and y ions, where the charge is retained on the C-terminus.

However, this depends on the ionisation technique, the mass analyser, and the peptide structure. Electron capture dissociation, for example, produces predominantly c and z ions.



Peptide fragmentation is rarely a clean process, and the spectrum will often show significant peaks from side chain cleavages and internal fragments, where the backbone has been cleaved twice.

This slide shows the most common structures.

There is a "ready reckoner" here for fragment ion masses. N is mass of the N-terminal group, (hydrogen for free amine). C is the mass of the C-terminal group, (hydroxyl for free acid). M is the sum of the residue masses. To determine the neutral mass of, say a 'b' ion with just two glycines, add the mass of the n terminal group, which is normally just a hydrogen, so '1', the mass of two glycines is 114 and subtract a hydrogen which leaves a mass of 114. To get the singly charged ion, we need to add a proton, which gives a mass/charge of approximately 115.

The ion series we see in ECD and ETD are actually z+H and z+2H

The best introduction to peptide dissociation is still this review by Ioannis Papayannopoulos



Which brings us to the second method of using mass spectrometry data for protein identification: a sequence query in which mass information is combined with amino acid sequence or composition data. The most widely used approach in this category is the sequence tag, developed by Matthias Mann and Matthias Wilm at EMBL.



In a sequence tag search, a few residues of amino acid sequence are interpreted from the MS/MS spectrum.

Even when the quality of the spectrum is poor, it is often possible to pick out four clean peaks, and read off three residues of sequence. In a sequence homology search, a triplet would be worth almost nothing, since any given triplet can be expected to occur by chance many times in even a small database.

What Mann and Wilm realised was that this very short stretch of amino acid sequence might provide sufficient specificity to provide an unambiguous identification if it was combined with the fragment ion mass values which enclose it, the peptide mass, and the enzyme specificity.



Picking out a good tag is not trivial, and often involves making judgements based on experience. In this spectrum, we can see a promising four residue tag.



As with PMF, I have limited my list to sequence query servers that are publicly available on the web. If I have missed any, please let me know.

C Matrix Science - Masco	S × 🔶	6	ED 200540.01 CO	Google 👝 I		
(MATRIX) SCIENCES	http://tb1-jsc/mascot/cg/search	rorm.pin+ORM	HOME : MASCOT : HELP	Search	GO	
Mascot > Sequer	ce Query Sequence Query					
Your name	Lou Scene	Email	leu@res.edu			
Search title						
Database(s)	MSIPI_mouse A NCBInr	Enzyme	Trypsin/P			
	SwissProt I	Allow up to	1 Missed deavage	s		
	UniRef100	Quantitation	None	*	J	
Taxonomy	All entries		~			
Fixed modifications	Carbamidomethyl (C)	>	Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-te	rm)		
Variable modifications	Display all modifications	>	Ammonia-loss (N-term Biotin (K) Biotin (N-term) Carbamyl (K) Carbamyl (N-term) Carboxymethyl (C)	c)		
Peptide tol. ±	0.6 Da 💌 # 13C O 💌	MS/MS tol. ±	0.6 Da 💌			
Peptide charge	Mr	Monoisotopic	⊙ Average ○			
Query	1489.430 tag(650.213,6WSV,1079	.335)		4		
Instrument	ESI-TRAP					
Decoy		Report top	AUTO 💌 hits			
	Start Search		Reset Form			
	Copyright @ 2008 Matrix S	sience Ltd. All Rig	hts Reserved.			
: Introductio	n		© 2007-2010 N	latrix Scie	nce	(MATRI SCIENCI

I entered the tag shown earlier into the Mascot Sequence Query search form. As with a PMF, several search parameters are required, such as the database to be searched and an estimate of the mass accuracy.



This is the result report from the search. There is just one peptide in the database that matches: LQGIVSWGSGCAQK from bovine trypsinogen.

The score is good, but even if it wasn't, you are on <u>very</u> safe ground accepting any match to trypsin, keratin, or BSA ;-)



A sequence tag search can be rapid, because it is simply a filter on the database.

Without doubt, the most important advantage of this approach is the so-called "error tolerant" mode. This consists of relaxing the specificity, usually by removing the peptide molecular mass constraint. When this is done, the tag is effectively allowed to float within the candidate sequence, so that a match is possible even if there is a difference in the calculated mass to one side or the other of the tag. This is one of the few ways of getting a match to a peptide when there is an unsuspected modification or a variation in the primary amino acid sequence.

Tags can be called by software. But, in most cases, they are called manually, which requires time and skill.

If the tag is not correct, then no match will be found. With some search engines, ambiguity is OK, as long as it is recognised and the query is formulated correctly. Obviously, I=L and, in most cases, Q=K and F=MetOx. Software or a table of mass values can help identify the more common ambiguities.



Which brings us to the third category: Searching uninterpreted MS/MS data from a single peptide or from a complete LC-MS/MS run. That is, using software to match lists of fragment ion mass and intensity values, without any manual sequence calling.



This approach was pioneered by John Yates and Jimmy Eng at the University of Washington, Seattle. They used a cross correlation algorithm to compare an experimental MS/MS spectrum against spectra predicted from peptide sequences from a database. Their ideas were implemented as the Sequest program.

MS/MS lons Search Servers on the Web

MA	SCOT : Introdu	ction © 2007-2010 Matrix Science	SCIENCE
	Not on-line	ByOnic, Crux, MassMatrix, Myrimatch, Paragon, PepSplice ProLuCID, ProbID, ProteinLynx GS, SIMS, Sequest, Spectro greylag, pFind, etc.	e, umMill,
	XProteo	http://xproteo.com:2698/	
	X!Tandem (The GPM)	http://thegpm.org/TANDEM/index.html	
	Sonar (Knexus)	http://hs2.proteome.ca/prowl/knexus.html	
	RAId_DbS	http://www.ncbi.nlm.nih.gov/CBBResearch/qmbp/RAId_E html	bS/index.
	Phenyx	http://phenyx.vital-it.ch/pwi/login/login.jsp	
	PepProbe	http://bart.scripps.edu/public/search/pep_probe/search.	jsp
	PepFrag (Prowl)	http://prowl.rockefeller.edu/prowl/pepfrag.html	
	Omssa	http://pubchem.ncbi.nlm.nih.gov/omssa/index.htm	
	MS-Tag (Protein Prospector)	http://prospector.ucsf.edu/prospector/cgi- bin/msform.cgi?form=mstagstandard	
	Mascot	http://www.matrixscience.com/search_form_select.html	l
	Inspect	http://proteomics.ucsd.edu/LiveSearch/	

There is a wide choice of search engines on the web for performing searches of uninterpreted MS/MS data. I've also listed some of the packages that are not on the web, which includes Sequest. Again, if I've missed one, please advise.

As with a peptide mass fingerprint, the starting point is a peak list. There are several different formats for MS/MS peak lists, and this may constrain your choice of search engine

this Some - Mascet - N	r						
C fill http://t61-jsc/mascot/cg/search_form.pIFGRMMER=28SEARCH-MIS Image: Control of the search in the search	Matrix Science - Mascot	-M × 🕀					
Vericele Display all modifications Peptide tol. ± 20 Peptid	< → C fi s	http://t61-jsc/mascot/cgi/search_	form.pl?FORMV	ER=28.SEARCH=MIS		B• 🔑•	
Allow up to image and a second i	<i>{MATRIX</i> (SCIENCE)			HOME MASCOT HELP	Search	Go	
Your name Expert User Email smartle@matrixslence.com Search title Database(s) MSIP_human Exzyme Troppin/P MSIP_human Allow up to 1 1 missed deavages Allow up to 1 1 missed deavages Transbil Quantitation None Field andomothyl (C) C Acetyl (N+arm) Allow up to 1 1 missed deavages Acetyl (N+arm) Acetyl (N+arm) modifications Carbamidometryl (C) C Acetyl (N+arm) Andated (C+arm) Amidated (C+arm) Amidated (C+arm) Amidated (C+arm) Amidated (C+arm) Amidated (C+arm) Amidated (N+arm) C Acetyl (N+arm) Amidated (C+arm) Amidated (C+arm) Amidated (N+arm) Amidated (C+arm) Amidated (C+arm) Amidated (N+arm) Amidated (N+arm) C Acetyl (N+arm) Amidated (N+arm) Amidated (N+arm) C Acetyl (N+arm) Amidated (N+arm) Amidated (N+arm) Poptide tol. ± 20 ppm W #19C Wonoisotopic © Average Earbamyli (C) V Data file Choose file No file choosen Poptidated (N+arm) Error	Mascot > MS/MS i	ions Search					
Torrane Exam Exam Exam Exam Database(s) MSEP_human Exam Exam Trypsin/P M Allow up to 1 Missed deavages Quantitation None W Taxonomy Allow up to 1 Missed deavages Quantitation None W Taxonomy Allow up to 1 Missed deavages Quantitation None W Transb Quantitation None W Allow up to 1 Missed deavages Transb Quantitation None W Allow up to 1 Missed deavages Transb Quantitation None W Allow up to 1 Missed deavages Missed deavages Quantitation None W Allow up to Missed deavages Missed deavages Quantitation Allow up to 1 Allow up to	MASCOT	M3/M3 Ions Search	for all				
Search title Database(s) MSIPL-human MSIPL-human MSIPL-human MSIPL-human MSIPL-human MSIPL-human MSIPL-human MSIPL-human MSIPL-human MSIPL-human MSIPL-human Quantitation Quantitation None Quantitation None Marchyl (Protein N-karm) Activyl (Protein N-karm) Ac	Tour name	Expert Oser	Email	smarue@mauxscience	e.com		
Database(s) MSIPL_mouse Maintown Texture Enzyme Mileven Tropsin/P M Allow up 1 Missed deavages Quantitation None M Taxonony Allow up 1 Missed deavages Quantitation None M Taxonony Allow up 1 Missed deavages Quantitation None M Taxonony Allow up 1 Activit (N M Taxonony Carbamidomethyl (C) Activit (N Activit (N Carbamidomethyl (C) Activit (N Activit (N Activit (N Activit (N Biotin (N Biotin (N Biotin (N Biotin (N Biotin (N Biotin (N Biotin (N <t< th=""><th>Search title</th><th></th><th>1</th><th></th><th></th><th></th><th></th></t<>	Search title		1				
Taxonomy All entries modifications Carbamidomethyl (C) Acetyl (K) wariable Display all modifications Acetyl (N-term) Acetyl (Porton N-term) Acetyl (Porton N-term) Acetyl (Porton N-term) Acetyl (Porton N-term) wariable Display all modifications Display all modifications posidation (M) S Carbamy (K) addition (M) S Carbamy (K) peptide tol. ± 20 Pom # #3C 0 Nonoisotipic Acetyl (Porton N-term) Ammonia-toss (N-term) (A pation (M) S Carbamy (K) carbamy (K) S S pato file Choose File No file choose File No file Coose pato file No Report top Mrcasot percersor m/2 Report top instrument Est-QuAD-TOF Report top Start Search Reset Form	Database(s)	MSIPI_human MSIPI_mouse NCBInr SwissProt Trembl	Enzyme Allow up to Quantitation	Trypsin/P V 1 V missed deavage None	es	~	
madifications Carbamidomethyl (C) Carbamidomethyl (C) madifications Display all modifications Carbamyl (Arbam) madifications Display all modifications Arbamyl (Arbam) madifications Display all modifications Arbamyl (Arbam) madifications Display all modifications Arbamyl (Arbam) Peptide tol. ± 20 ppm # #1°C 0 W MS/MS tol. ± 0.1 Da w reptide tolare 2 + and 3 + w Monoisotips (Marcan) Wonoisotips (Marcan) Wonoisotips (Marcan) Data fill Choose File No file tohosen Freeursor m/z Data format Mascot generic w Precursor m/z Instrument ESI-QUAD-TOF w Error tolerant Report top Al/TO w hts Start Search Reset Form Reset Form	Taxonomy	All entries		¥			
Variable madifications Oxidiation (M) Sectors (M	Fixed modifications	Carbamidomethyl (C)	> <	Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (C-term)	i) ierm)		
Peptide tol. ± 20 ppm w # 1°C 0 w MS/MS tol. ± 0.1 Da w eptide tol. ± 20.1 Da w Monoisotopic 0 Average 0 Data format Mascot generic w Precursor m/z Instrument Est-QuAD-TOF w Error tolerant 1 Decoy Report top AUTO w hts Reset Form	Variable modifications	Oxidation (M)	>	Biotin (K) Biotin (N-term) Carbamyl (K) Carbamyl (N-term) Carboxymethyl (C)		×	
Deta file Choose File No file choose Data file Choose File No file choose Data format Mascot generic m/z Instrument ESLQUAD-TOF Error tolerant Decoy Report top Autro Start Search Reset Form	Peptide tol. \pm	20 ppm 💙 🗰 13C 0 💌	MS/MS tol. ±	0.1 Da 💌			
Data file Choose File No file chosen Data format Mascot generic Precursor Instrument ESI-QUAD-TOF Error tolerant Decoy Report top Autor Start Search Reset Form	Peptide charge	2+ and 3+ 💌	Monoisotopic	💿 Average 🔘			
Data format Mascot generic Precursor m/z Instrument ESI-QUAD-TOF Error tolerant Decoy Report top AUTO hits Start Search Reset Form	Data file	Choose File No file chosen					
Instrument ESI-QUAD-TOF	Data format	Mascot generic 💌	Precursor	m/z			
Decoy Report top AUTO white Start Search Reset Form	Instrument	ESI-QUAD-TOF	Error tolerant				
Start Search Reset Form	Decoy		Report top	AUTO 💌 hits			
		Start Search		Reset Form			
Copyright @ 2008 Matrix Science Ltd. All Rights Reserved.	Peptide tol. ± Peptide charge Data file Data format Instrument Decoy	20 ppm * **C * 2+ and 3+ * Choose File I No file chosen Mascet generie * ESI QUAD-TOF * : Start Search	KS/MS tol. ± Monoisotopic Precursor Error tolerant Report top	Carbanyi (VA-kerm) Carbanyi (VA-kerm) Carbanyi (VA-kerm) O.1 Da Average Average Average Auto M hits Reset Form		8	
							I
	troductio	on		© 2007-2010	Matrix S	cience	(MATRIX) Sciences

This is the search form for an MS/MS search.

As before, you specify the database, mass accuracy, modifications to be considered, etc., and associate the peak list file with the search form.

Address 🔕 http://v	www.matrixscience	.com/cgi/master_i	results.pl?file=/da	ata/F981123.d	at				💙 🄁 Go
1. <u>132800</u>	onin Croff	Mass: 610	16 Score:	1195	Queries	; mat	ched: 31		1
Chaper	to include	this hit is	- numan n error tole	aront cas	rch				
_ oncer	co morade	chib hit i							
Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss S	core	Expect	Rank	Peptide
✓ <u>11</u>	417.1822	832.3498	832.3827	-0.0329	0	45	0.1	1	K.APGFGDNR.K
✓ <u>12</u>	422.7433	843.4720	843.5065	-0.0345	0	46	0.11	1	K.VGEVIVTK.D
<u>13</u>	430.7328	859.4510	859.4837	-0.0327	0	12	2.5e+02	3	K.IPAMTIAK.N + Oxidation (M)
✓ <u>15</u>	451.2499	900.4853	900.5280	-0.0427	0	52	0.025	1	K.LSDGVAVLK.V
✓ 16	456.7806	911.5467	911.5803	-0.0337	0	59	0.0041	1	K.VGLQVVAVK.A
21	480.7447	959.4748	959.5036	-0.0288	0	45	0.11	1	R. VTDALNATR. A
24	595.7855	1189.5565	1189.6012	-0.0447	0	57	0.0068	1	K.EIGNIISDAMK.K
25	603.7720	1205.5294	1205.5961	-0.0668	0	(50)	0.027	1	K.EIGNIISDAMK.K + Oxidation (M)
26	608.3099	1214.6052	1214.6506	-0.0454	U	73	0.00015	1	K. NAGVEGSLIVEK. I
27	617.2857	1232.5569	1232.5884	-0.0315	U	81	2.76-05	1	K.VGGTSDVEVNEK.K
✓ 31	672.8375	1343.6605	1343.7085	-0.0480	0	64	0.001	1	R. TVIIEUSWGSPR. V
V 34	714.0004	1427.7623	1427.0057	-0.0434	0	(83)	0.00000	-	R. GUMLAUDAUTAFLK.K
✓ 35	714.0930	1427.7730	1427.0037	-0.0327		73	0.00013	1	R. COMLANDAVIAELK. K
V 30	722.0049	1443.7332	1443.0000	-0.0434		/201	0.00014	- 1	R. GURLAUDAVIAELK.K + Oxidation (N)
V <u>57</u>	752 8643	1503 7141	1503 7490	-0.0204	0	00	2 78-06	- 1	K TINDELETTECHE E
V <u>40</u>	760.8461	1519.6777	1519.7439	-0.0662	ñ	(84)	8.9e-06	÷.	K.TLNDELETTEGMK.F + Oxidation (M)
¥ 45	640.3281	1917.9625	1918.0636	-0.1010	0	102	1.3e-07	1	K. ISSTOSTVPALETANAHR. K
V 72	960 0327	1018 0500	1018 0636	-0 0127	0	(87)	3.2e-06	1	K.ISSIQSIVPALEIANAHR.K
V 48	Top scoring	peptide ma	tches to que	ery 45			0.01	1	R.IQEIIEQLDVTTSEYEK.E
V 51	Score great	er than 32	indicates ho	omology			4.6e-09	1	R.ALMLQGVDLLADAVAVTNGPK.G
✓ 52	Score great:	er than 45	ite for this	nentide			0.00022	1	R.ALMLQGVDLLADAVAVTNGPK.G + Oxidation (M)
✓ 53	out and the set of the	SHOWD UIT H	100 101 001	, 2020100			3.9	1	R.ALMLQGVDLLADAVAVTNGPK.G + Oxidation (M)
✓ 54 5	Score Del	ta Hit Pr	otein Peptie	ie			0.00018	1	R.ALMLQGVDLLADAVAVINGPK.G + 2 Oxidation (M)
✓ 58 ⁻¹	101.5 -0.	10 1 A3	2800 K.ISSI	SIVPALEI	ANAHR. I	K I	0.0038	1	R.KPLVIIAEDVDGEALSTLVLNR.L
✓ 59	13.6 -0.1	10	R. ILLDI	NONE IHTN	ILR.T VPLTR 1		0.00038	1	R.KPLVIIAEDVDGEALSTLVLNR.L
✓ 60	11.1 0.0	01	K. TEVT	IYCIAPVGE	APEAR.	i.	4.5e-07	1	R.KPLVIIAEDVDGEALSTLVLNR.L
61	10.2 -0.0	06	K.GLVL	DADGGIHLE	IVNDR.	5	2.8	1	R. TALLDAAGVASLLTTAEVVVTEIPK.E
✓ 62	6.5 -0.1	16	R.LAPTO	GALGEVIIQ	GPTILR.	. Е	0.02	1	R. TALLDAAGVASLLTTAEVVVTEIPK.E
✓ 64	5.1 -0.1	11	R.RTIG	LLIARVPVI ICEGAPOFN	TASOK.0	2	3.4e-05	1	K.LVQDVANNTNEEAGDGTTTATVLAR.S
<									>
1:A32800									Internet

The results from this type of search tend to be more complicated to report. This is because the results usually represent a number of MS/MS spectra, rather than a single spectrum. Hence, there is an additional dimension to the data.

For each spectrum, there may be multiple possible peptide matches. Each peptide match may be assignable to multiple proteins. This makes the results more difficult to represent in two dimensions and leads to a wide variety of reporting styles. Mascot uses a pop-up window to show the alternative peptide matches to each spectrum.

MS/MS lons Search		
Easily automated for high	throughput	
Can get matches from mar	ginal data	
Can be slow No enzyme Many variable modifications Large database Large dataset		
MS/MS is peptide identifica	ation	
Proteins by inference.		
MASCOT : Introduction	© 2007-2010 Matrix Science	{MATRIX } {SCIENCE}

Searching of uninterpreted MS/MS data is readily automated for high throughput work. Most "proteomics pipelines" use this approach.

It offers the possibility of getting useful matches from spectra of marginal quality, where it would not be possible to call a reliable sequence tag.

On the down side, such searches can be slow. Particularly if performed without enzyme specificity or with several variable modifications.

Finally, always remember that it is peptides that are being identified, not proteins. From the peptides that have been identified, we can infer which proteins may have been present.

Information content20 to 200 mass values20 to 200 mass valuesBoundary conditionSingle protein sequenceSingle peptide sequenceCleavage specificityEnzymeGas-phase dissociationMajor unknownProtein lengthFragmentation channelsUniqueShotgun proteinResidue level		PMF	MS/MS
Boundary conditionSingle protein sequenceSingle peptide sequenceCleavage specificityEnzymeGas-phase dissociationMajor unknownProtein lengthFragmentation channelsUniqueShotgun proteinResidue level	Information content	20 to 200 mass values	20 to 200 mass values
Cleavage specificityEnzymeGas-phase dissociationMajor unknownProtein lengthFragmentation channelsUniqueShotgun proteinResidue level	Boundary condition	Single protein sequence	Single peptide sequence
Major unknownProtein length lengthFragmentation channelsUniqueShotgun proteinResidue level	Cleavage specificity	Enzyme	Gas-phase dissociation
Unique Shotgun protein Residue level	Major unknown	Protein length	Fragmentation channels
strength identification characterisation	Unique strength	Shotgun protein identification	Residue level characterisation

To complete this overview of the methods of protein identification, I'd like to compare the fundamental characteristics of database searching using MS data versus MS/MS data.

The mass spectrum of a tryptic digest of a protein of average size might contain 50 peptide masses, not dissimilar from the MS/MS spectrum of an average sized tryptic peptide. Thus, the "information content" of the individual spectra is similar. The reason an MS/MS search can be more powerful is mainly that the data set can contain many such spectra, so multiplying the information content. However, at the single spectrum level, there is little to choose.

In a peptide mass fingerprint, the boundary condition is that the peptides all originate from a single protein. In an MS/MS search, the boundary condition is that the fragments all originate from a single peptide. The weakness of the peptide mass fingerprint is that this boundary condition is often violated, and the spectrum actually represents the digest products of a protein mixture. The MS/MS boundary condition can also be violated, when we analyse co-eluting, isobaric peptides. If this happens, and we have a mixture, the MS/MS search is just as likely to fail as the PMF.

In the peptide mass fingerprint, the specificity comes from the predictable cleavage behaviour of the proteolytic enzyme. Thus, we want an enzyme with a low cutting frequency, such as trypsin. In the MS/MS ions search, the specificity comes from the mostly predictable gas-phase fragmentation behaviour of peptide molecular ions.

Arguably, the major strength of PMF is that it really is shotgun protein identification. The higher the coverage, the more confident one can be that the protein in the database is the one in the sample. The unique strength of searching MS/MS data is that one gets residue level information. A good match can reveal the presence and location of post translational modifications, which is not possible with a PMF.



Even a 2 day course has to skim over a lot of detail. This recent book provides greater depth in some areas that we don't have time to deal with.