

Database Searching for Protein Identification and Characterization

John Cottrell
Matrix Science

<http://tinyurl.com/e4brg>

jcottrell@matrixscience.com

Topics

Methods of database searching

Practical tips

Scoring

Validation & reporting tools

Why searches can fail

Modifications

Sequence databases

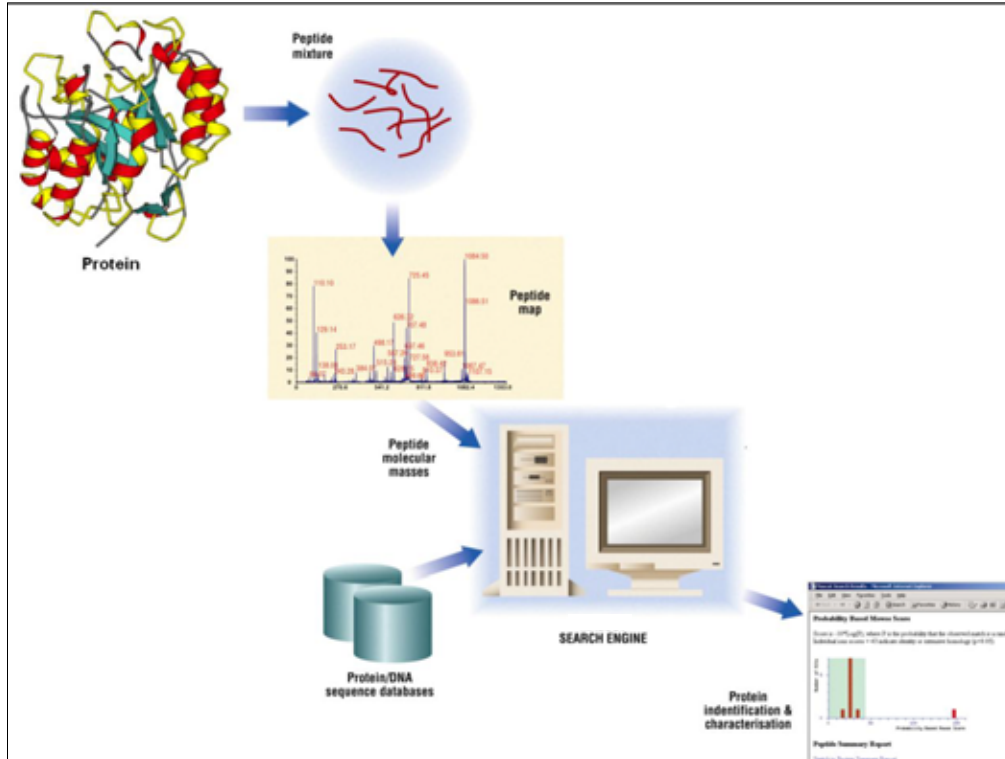
Future directions

Three ways to use mass spectrometry data for protein identification

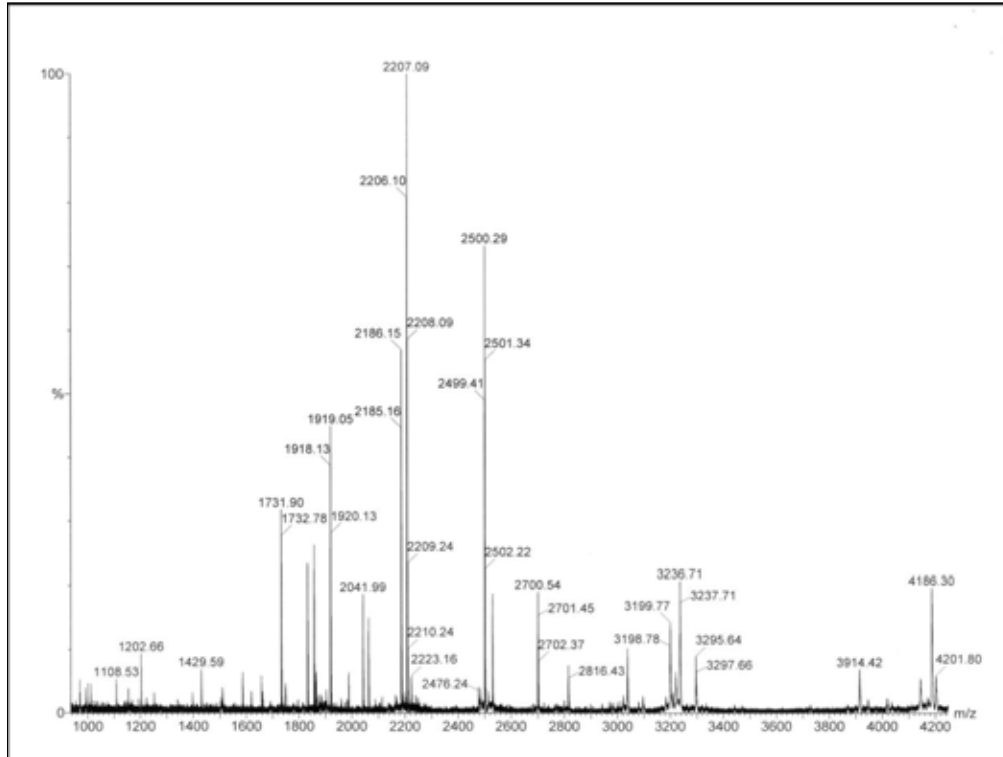
1. Peptide Mass Fingerprint

A set of peptide molecular masses from an enzyme digest of a protein

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.

PMF Servers on the Web

Aldente (Phenyx)

- <http://www.expasy.org/tools/aldente/>

Mascot

- http://www.matrixscience.com/search_form_select.html

MassSearch

- <http://cbrg.inf.ethz.ch/Server/MassSearch.html>

Mowse

- <http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse>

MS-Fit (Protein Prospector)

- <http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm>

PepMAPPER

- <http://wolf.bms.umist.ac.uk/mapper/>

PeptideSearch

- <http://www.mann.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html>

Profound (Prowl)

- <http://bioinformatics.genomicsolutions.com/service/prowl/profound.html>

XProteo

- <http://xproteo.com:2698/>

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.

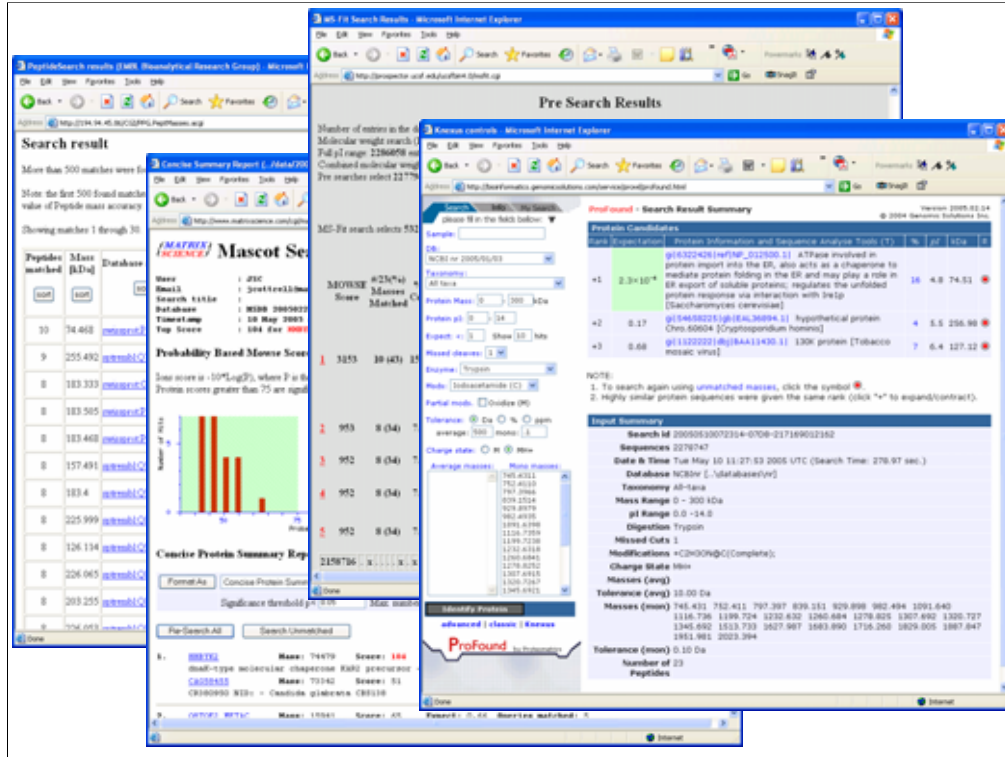
Many other PMF programs have been described in the literature. Most packages are either available for download from the web or are commercial products.

Search Parameters

- database
- taxonomy
- enzyme
- missed cleavages
- fixed modifications
- variable modifications
- protein MW
- protein pI
- estimated mass measurement error

This is the search form for MS-Fit, part of Karl Clauser's Protein Prospector package. Besides the MS data, a number of search parameters are required. Some search engines require fewer parameters, others require more. I'll be discussing common search parameters in detail in the practical tips section of this talk.

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. The reports shown here come from PeptideSearch, Mascot, MS-Fit, and Profound.

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.



1993: Vintage Year for PMF

- Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C. and Watanabe, C. (1993). Proc Natl Acad Sci USA 90, 5011-5.
- James, P., Quadroni, M., Carafoli, E. and Gonnet, G. (1993). Biochem Biophys Res Commun 195, 58-64.
- Mann, M., Hojrup, P. and Roepstorff, P. (1993). Biol Mass Spectrom 22, 338-45.
- Pappin, D. J. C., Hojrup, P. and Bleasby, A. J. (1993). Curr. Biol. 3, 327-32.
- Yates, J. R., 3rd, Speicher, S., Griffin, P. R. and Hunkapiller, T. (1993). Anal Biochem 214, 397-408.

On a historical note, the discovery of peptide mass fingerprinting was unusual in that it wasn't just one or two groups that first demonstrated the feasibility of the technique. In 1993, no less than five groups independently published descriptions of the method. Bill Henzel and colleagues at Genentech, Peter James's group at ETH Zurich, Matthias Mann's lab at EMBL Heidelberg, Darryl Pappin's group at ICRF London, and John Yates et al. at U Washington.

Clearly, an idea whose time had come.

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 to identify frequently observed proteins in electrophoresis gels. We describe the genesis of the idea in 1989, and show the first demonstration with fast atom bombardment mass spectrometry. Despite its promise, the method was seldom used until 1992, with the coming of significantly more sensitive commercial instrumentation based on MALDI-TOF-MS. We recount the evolution of the method and its dependence on a number of technical breakthroughs, both in mass spectrometry and in other areas. We show how it laid the foundation for high-throughput, high-sensitivity methods of protein analysis, now known as proteomics. We conclude with recommendations for further improvements, and speculation of the role of PMF in the future. (J Am Soc Mass Spectrom 2003, 14, 931-942) © 2003 American Society for Mass Spectrometry

➤Henzel, W. J., Watanabe, C., Stults, J. T., *J. Am. Soc. Mass Spectrom.* 2003, 14, 931-942.

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

Peptide Mass Fingerprint



Fast, simple analysis

High sensitivity

Need database of protein sequences

- not ESTs or genomic DNA

Sequence must be present in database

- or close homolog

Not good for mixtures

- especially a minor component.

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.

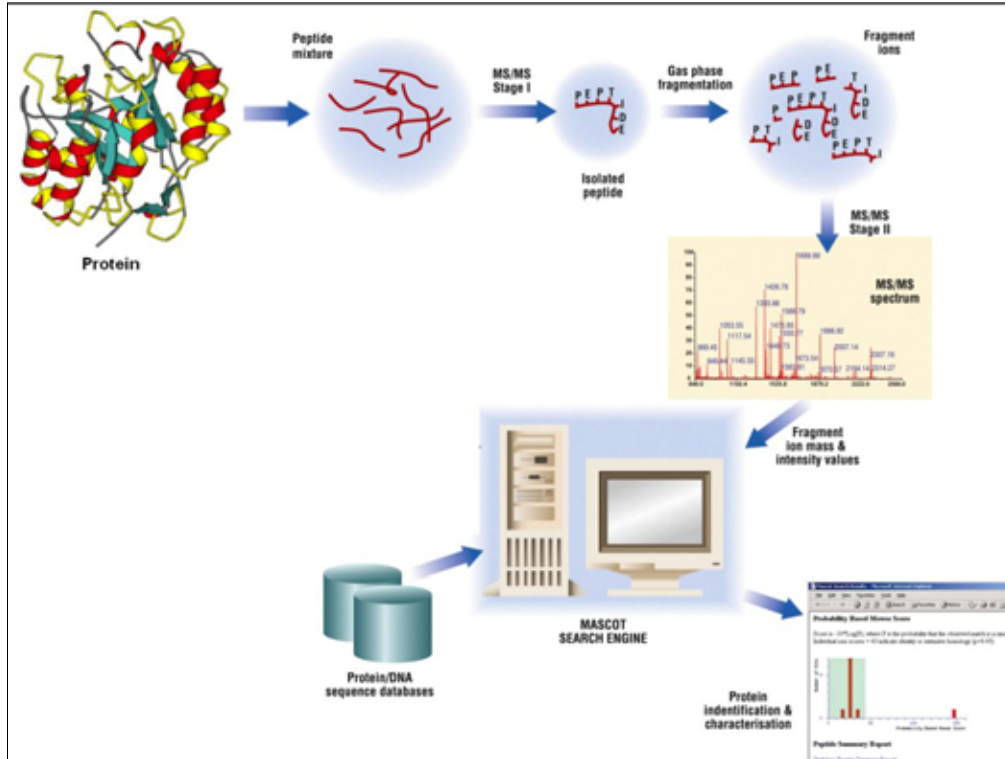
One of the limitations is that you need a database of proteins or nucleic acid sequences that are equivalent to proteins, e.g. mRNAs. In most cases, you will not get satisfactory results from an EST database, where most of the entries correspond to protein fragments, or genomic DNA, where there is a 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.



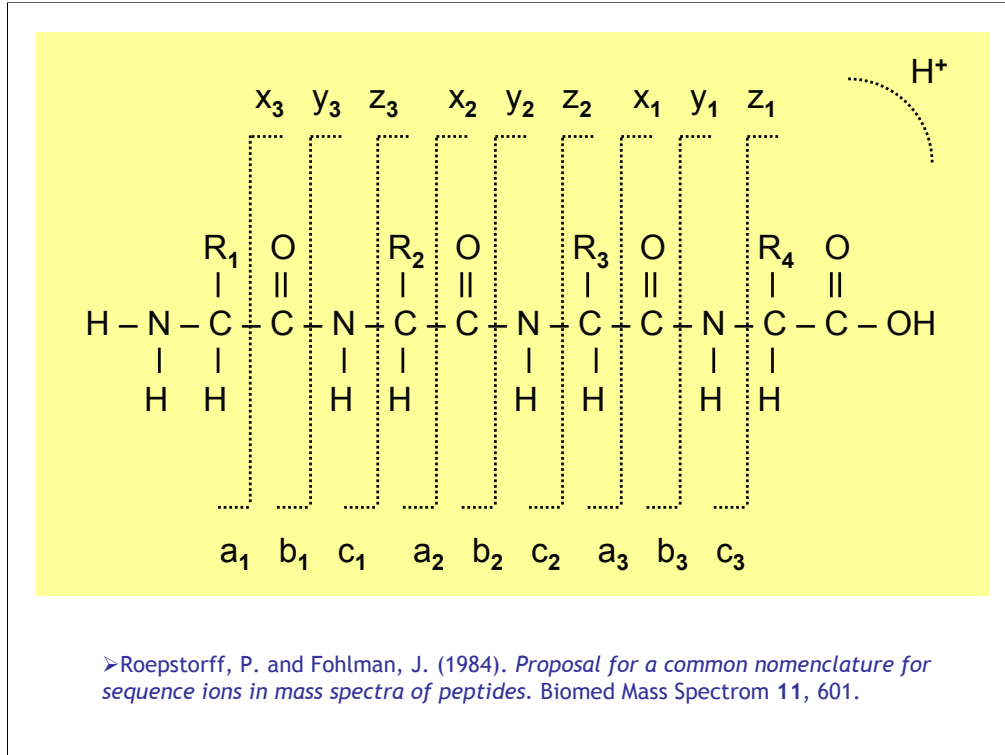
The experimental workflow for database matching of MS/MS data is similar to that for PMF, but with an added stage of selectivity and fragmentation.

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. We select the 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.

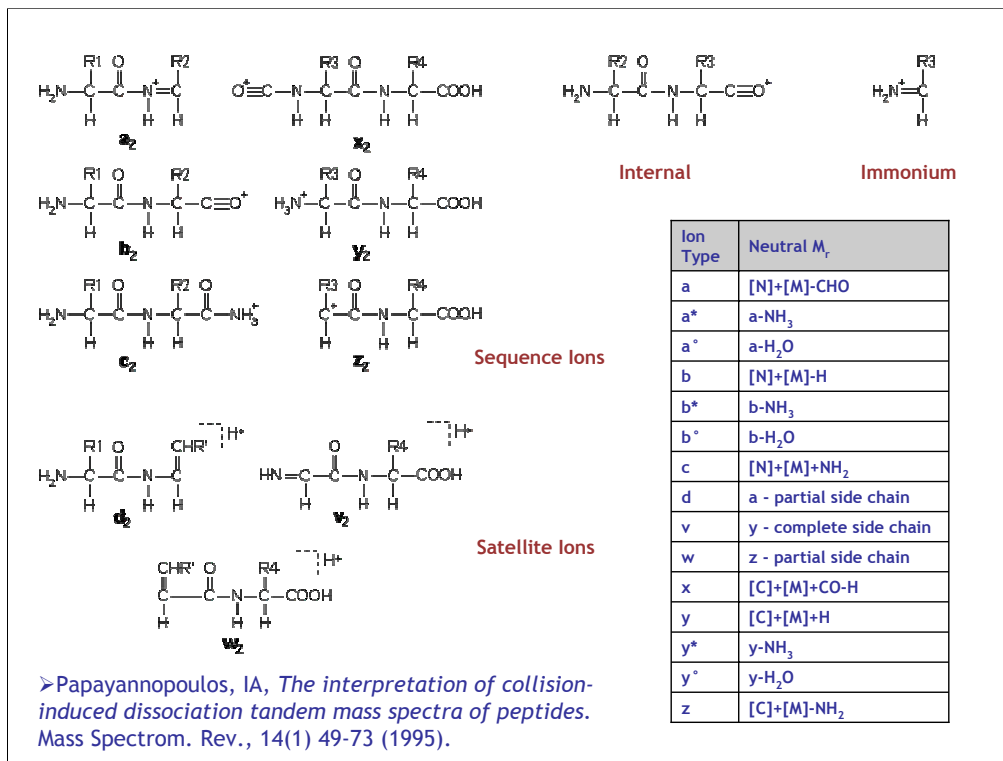
For each MS/MS spectrum, we use software to try and determine which peptide sequence in the database gives the best match. This will involve simulating the cleavage specificity of the enzyme, followed by calculation of the mass values we expect to result from the gas phase fragmentation of the peptide.

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 and includes a “ready reckoner” for fragment ion masses. N is mass of the N-terminal group, H for free amine. C is the mass of the C-terminal group, OH for free acid. M is the sum of the residue masses

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

Three ways to use mass spectrometry data for protein identification

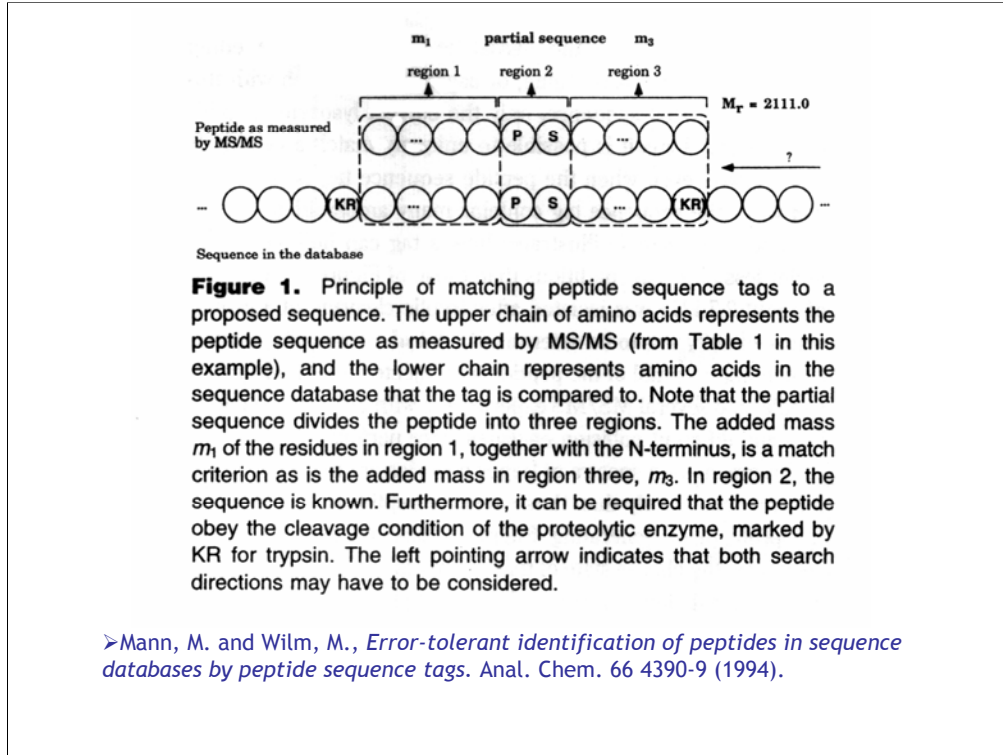
1. Peptide Mass Fingerprint

A set of peptide molecular masses from an enzyme digest of a protein

2. Sequence Query

Mass values combined with amino acid sequence or composition data

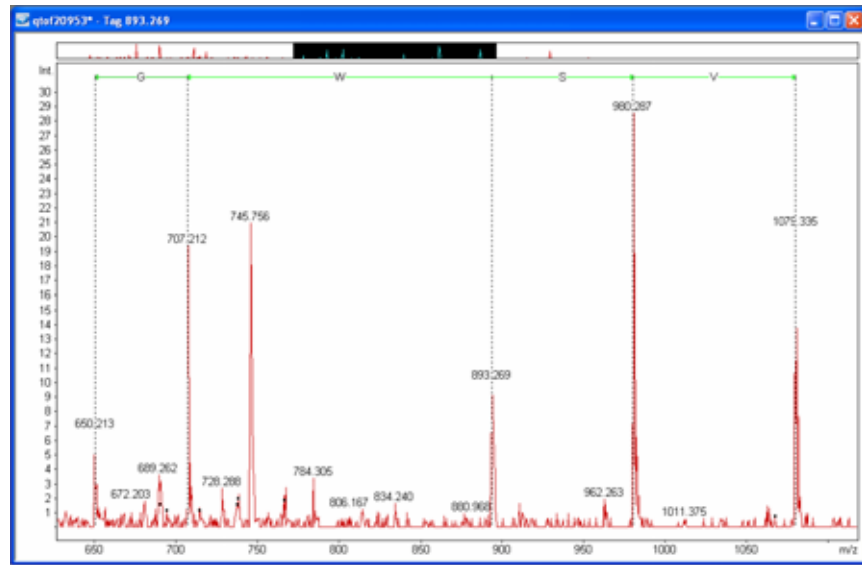
Which brings us to the second method of using mass spectrometry data for protein identification: a sequence query in which mass information is combined with amino acid sequence or composition data. The most widely used approach in this category is the sequence tag, developed by Matthias Mann and Matthias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



1489.430 tag(650.213,GWSV,1079.335)

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.

Sequence Query Servers on the Web

Mascot

- http://www.matrixscience.com/search_form_select.html

MS-Seq (Protein Prospector)

- <http://prospector.ucsf.edu/ucsfhtml4.0/msseq.htm>

Multident (TagIdent, etc.)

- <http://www.expasy.org/tools/multiident/>

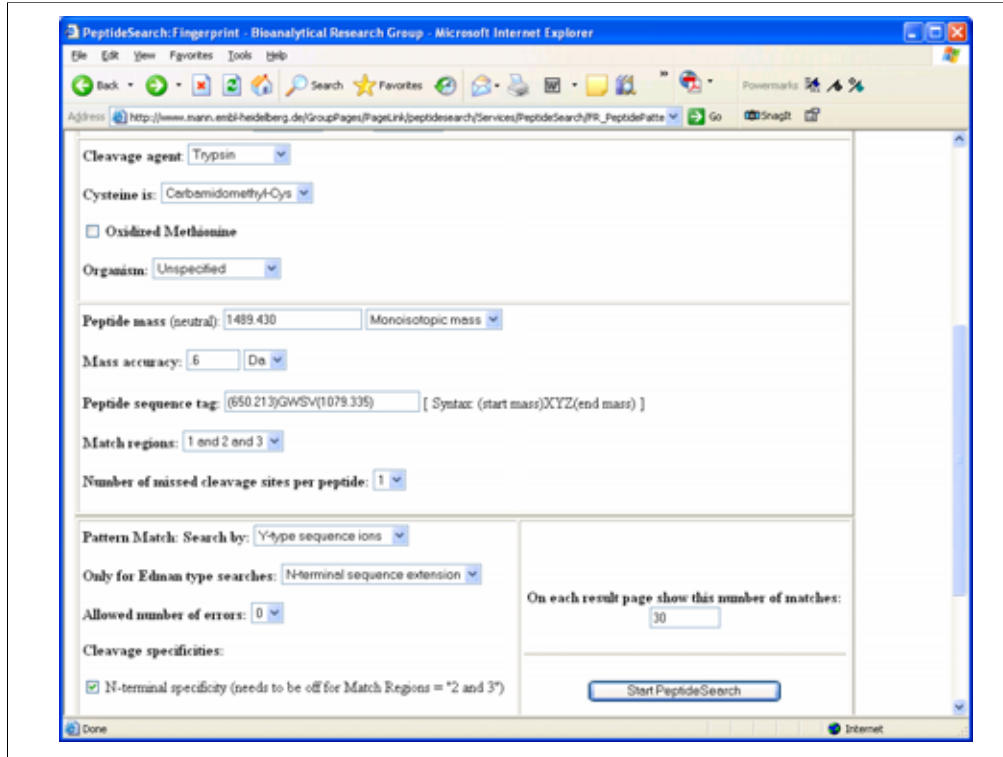
PeptideSearch

- <http://www.mann.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html>

Spider

- <http://proteome.sharcnet.ca:8080/spider.htm>

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. My email address is on the first slide.



For my example, I entered the tag shown earlier into the original PeptideSearch program at EMBL. As with a PMF, several search parameters are required, such as the database to be searched and an estimate of the mass accuracy.

Peptide Sequence matched/Peptide found	Mass [kDa]	Database accession	Internal Sequence	Organism	Protein Name	Digest
LQGIVSWGSGCAQK	25.425	rwisaprot.P00760	●	Bos taurus	Trypsinogen, cationic precursor	☐
LQGIVSWGSGCAQK	23.305	pdb:1AQ7 (not known by SRS)	●		1AQ7 TRYP SIN WITH INHIBITOR A	☐
LQGIVSWGSGCAQK	23.993	pdb:1BTP (not known by SRS)	●		1BTP MOL_ID: 1; MOLECULE BET	☐
LQGIVSWGSGCAQK	23.894	pdb:1G3B (not known by SRS)	●		1G3B-A BOVINE BETA-TRYP SIN BOU	☐
LQGIVSWGSGCAQK	23.308	pdb:1NTP (not known by SRS)	●		1NTP MODIFIED BETA TRYP SIN (M	☐
LQGIVSWGSGCAQK	25.409	pdb:1OPH (not known by SRS)	●		1OPH-B NON-COVALENT COMPLEX BE	☐
LQGIVSWGSGCAQK	23.291	pdb:1TAW (not known by SRS)	●		1TAW-A BOVINE TRYP SIN COMPLEXE	☐
LQGIVSWGSGCAQK	24.72	pdb:1ZZZ (not known by SRS)	●		1ZZZ-A TRYP SIN INHIBITORS WITH	☐
LQGIVSWGSGCAQK	23.329	pdb:5PTP (not known by SRS)	●		5PTP STRUCTURE OF HYDROLASE (☐

This is the result report from the search. There are 9 hits, but the peptide is the same in all cases: LQGIVSWGSGCAQK from bovine trypsinogen.

Although you don't get a score from this particular search engine, in my experience, you are on very safe ground accepting any match to trypsin , keratin, or BSA ;-)

Sequence Tag

Rapid search times

- Essentially a filter

Error tolerant

- Match peptide with unknown modification or SNP

Requires interpretation of spectrum

- Usually manual, hence not high throughput

Tag has to be called correctly

- Although ambiguity is OK

2060.78 tag(977.4,[Q|K][Q|K][Q|K]EE,1619.7).

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.

Sequence Tag / Sequence Homology

MultiTag

- Sunyaev, S., et al., *MultiTag: Multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry*, Anal. Chem. 75 1307-1315 (2003).

GutenTag

- Tabb, D. L., et al., *GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model*, Anal. Chem. 75 6415-6421 (2003).

MS-Blast

- Shevchenko, A., et al., *Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time of flight mass spectrometry and BLAST homology searching*, Analytical Chemistry 73 1917-1926 (2001)

FASTS, FASTF

- Mackey, A. J., et al., *Getting More from Less - Algorithms for rapid protein identification with multiple short peptide sequences*, Molecular & Cellular Proteomics 1 139-47 (2002)

OpenSea

- Searle, B. C., et al., *High-Throughput Identification of Proteins and Unanticipated Sequence Modifications Using a Mass-Based Alignment Algorithm for MS/MS de Novo Sequencing Results*, Anal. Chem. 76 2220-30 (2004)

CIIdentify

- Taylor, J. A. and Johnson, R. S., *Sequence database searches via de novo peptide sequencing by tandem mass spectrometry*, Rapid Commun. Mass Spectrom. 11 1067-75 (1997)

Sequence queries can be extremely powerful. These references are a good starting point if you are interested in learning more about the potential of combining mass and sequence information.

Three ways to use mass spectrometry data for protein identification

1. Peptide Mass Fingerprint

A set of peptide molecular masses from an enzyme digest of a protein

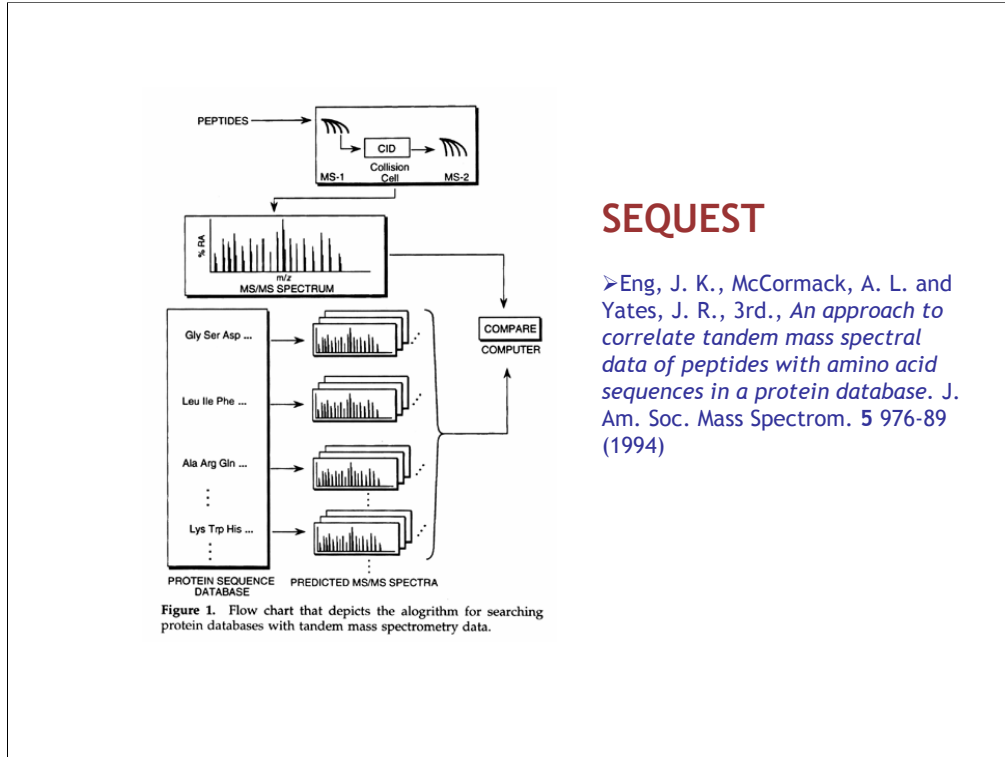
2. Sequence Query

Mass values combined with amino acid sequence or composition data

3. MS/MS Ions Search

Uninterpreted MS/MS data from a single peptide or from a complete LC-MS/MS run

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 Ions Search Servers on the Web

Mascot

- http://www.matrixscience.com/search_form_select.html

MS-Tag (Protein Prospector)

- <http://prospector.ucsf.edu/ucsfhtml4.0/mstagfd.htm>

Omssa

- <http://pubchem.ncbi.nlm.nih.gov/omssa/index.htm>

PepFrag (Prowl)

- <http://prowl.rockefeller.edu/PROWL/pepfragch.html>

Phenyx

- <http://www.phenyx-ms.com/index.html>

Sequest

- N/A

Sonar (Knexus)

- <http://bioinformatics.genomicsolutions.com/service/prowl/sonar.html>

X!Tandem (The GPM)

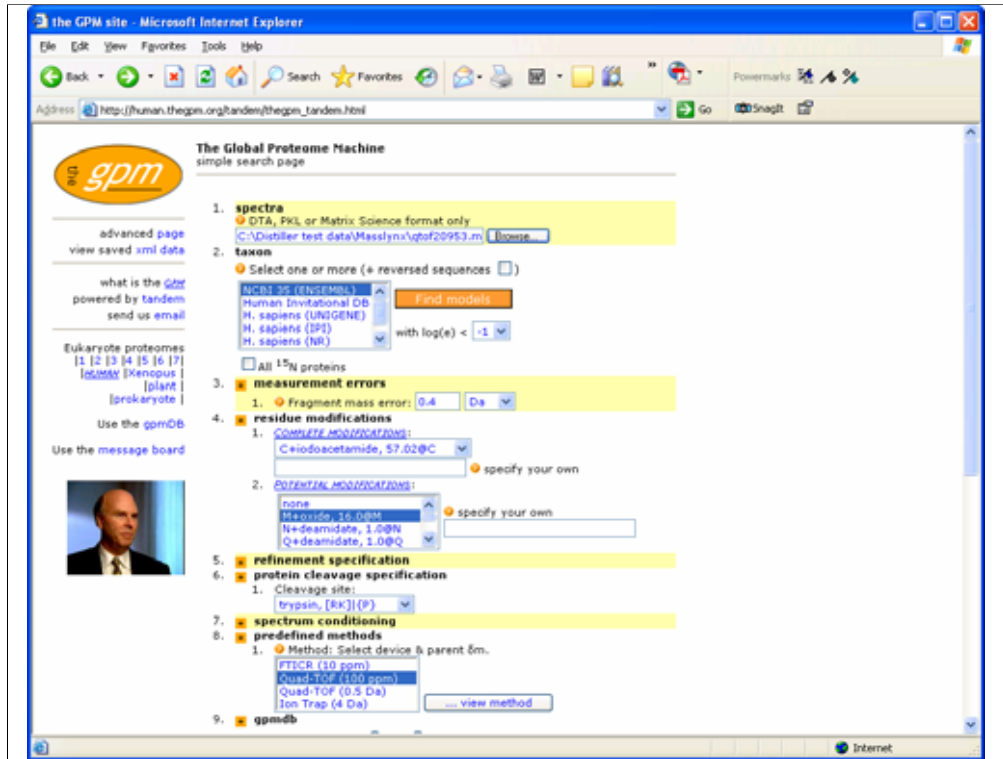
- <http://thegpm.org/TANDEM/index.html>

XProteo

- <http://xproteo.com:2698/>

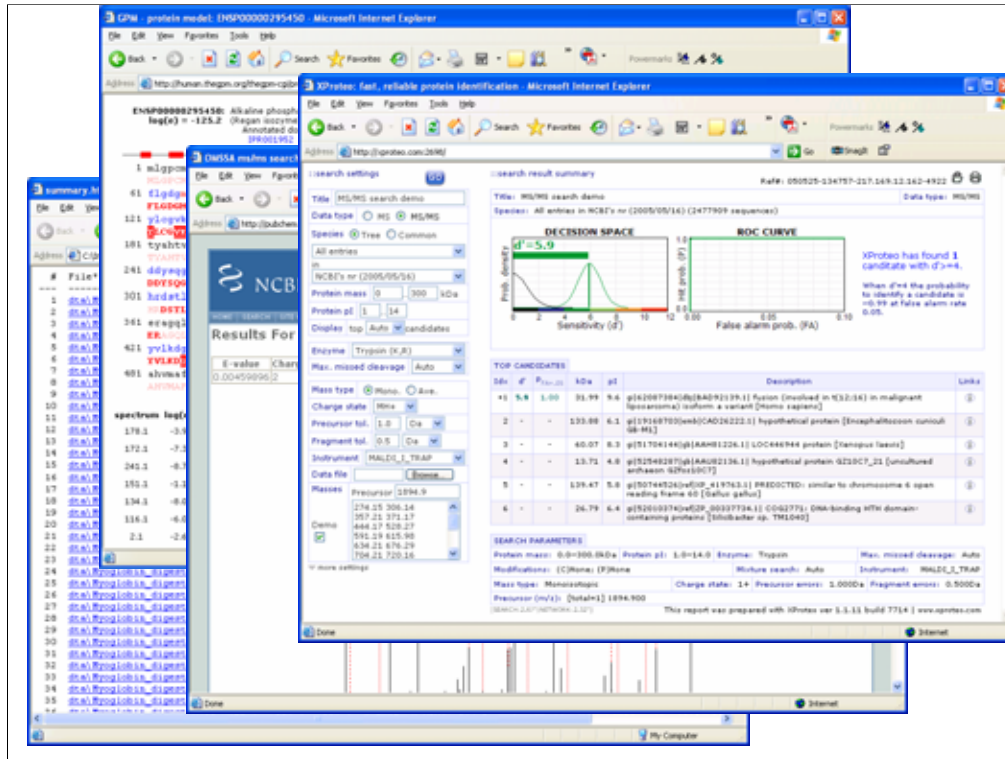
There is a wide choice of search engines on the web for performing searches of uninterpreted MS/MS data. Funnily enough, Sequest is not one of them, but I've listed it anyhow because it was the original. Again, if I've missed one, please advise.

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 is the search form for X!Tandem, from Ron Beavis and colleagues. This particular form is for searching human sequences, and Ron has adopted this useful pictorial method of choosing the correct organism.

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



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. The examples shown here are from Sequest, X!Tandem, Omssa, and XProteo.

MS/MS Ions Search

Easily automated for high throughput

Can get matches from marginal data

Can be slow

No enzyme

Many variable modifications

Large database

Large dataset

MS/MS is peptide identification

- Proteins by inference.

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

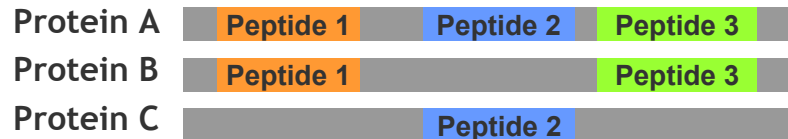
It offers the possibility of getting useful matches from spectra of marginal quality, where it would not be possible to call a reliable sequence tag.

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.

MS/MS matching identifies peptides, not proteins

Assigning peptide matches to protein hits can be arbitrary



Principal of parsimony / Occam's razor prefers Protein A

Imagine that we have three peptide matches, which can be assigned to three proteins, as illustrated here. Do we have evidence for all three proteins, or just one?

Many reports apply the so-called principal of parsimony, also called Occam's razor. This chooses the minimum number of proteins that can account for the observed peptides. Hence, most search engines will report that the sample contained protein A. They may add that proteins B and C contain sub-sets of the same peptides.

This is certainly a reasonable decision, but there is no guarantee that it is correct. It is possible that the sample actually did contain just proteins B and C. Another thing to watch for is the possibility that peptide 2 is a very weak match, maybe spurious. If so, then there is nothing to choose between Proteins A and B.

This ambiguity is exacerbated by shotgun proteomics or MudPIT, where the proteins from a cell lysate are digested to peptides without any prior fractionation or separation.

	PMF	MS/MS
Information content	20 to 200 mass values	20 to 200 mass values
Boundary condition	Single protein sequence	Single peptide sequence
Cleavage specificity	Enzyme	Gas-phase dissociation
Major unknown	Protein length	Fragmentation channels
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.

Practical Tips



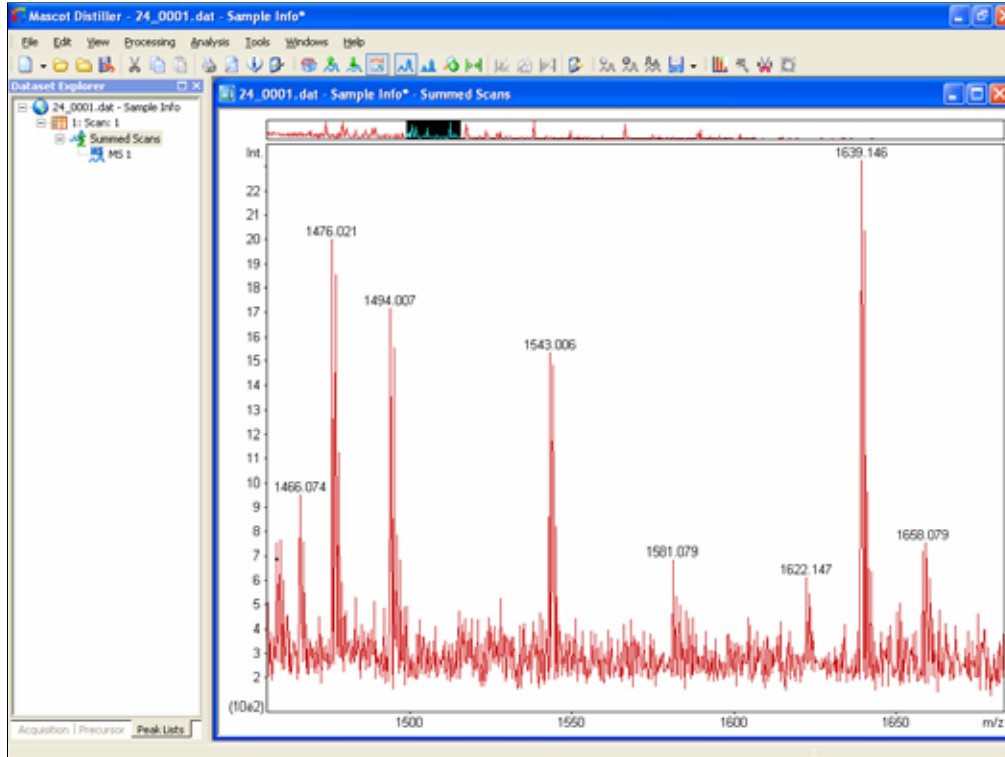
Peak detection, peak detection, peak detection

Especially critical for Peptide Mass Fingerprints

- A tryptic digest of an “average” protein (30 kDa) should produce of the order of 50 de-isotoped peptide peaks

Everyone is familiar with the phrase “garbage in garbage out”, usually applied to software. Which reminds us that the results of a database search are only as good as the peak list.

This is especially critical for Peptide Mass Fingerprint, because the higher mass peaks, which are the most discriminating, are often weak compared with the low mass peaks. When looking at a PMF peak list, bear in mind that a tryptic digest of an “average” protein (30 kDa) should produce something of the order of 50 de-isotoped peptide peaks.



If your peak list has only 2 or 3 peaks then you either have a very small protein or a sensitivity problem. At the other extreme, if you have 1000 peaks, most of them have to be noise, which will destroy the identification statistics.

Peak detection, peak detection, peak detection

Especially critical for Peptide Mass Fingerprints

- A tryptic digest of an “average” protein (30 kDa) should produce of the order of 50 peptide peaks

Time domain summing of LC-MS/MS data is very important

If you are working with LC-MS/MS data, don't neglect processing in the time domain

Mascot Search Results - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/master_results.pl?file=.../data/20030602/FTrcCicin.dat&PEPTYPE=peptide

Peptide Summary Report

[Switch to Protein Summary Report](#)

To create a bookmark for this report, right click this link: [Peptide Summary Report \(M. Moss/D. Becherer Sample\)](#)

Select All Select None Search Selected Error tolerant

1. [q11443370](#) Mass: 25583 Total score: 320 Peptides matched: 15
Chain A, Concanavalin A (Native)

Check to include this hit in error tolerant search

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
<input checked="" type="checkbox"/> 27	959.39	958.38	958.51	-0.13	0	(40)	1	LLGLFPDAN
<input checked="" type="checkbox"/> 28	480.22	958.43	958.51	-0.08	0	43	1	LLGLFPDAN
<input checked="" type="checkbox"/> 44	659.76	1317.50	1317.63	-0.13	0	80	1	VSSHGSPQSSVGR
<input checked="" type="checkbox"/> 52	524.90	1571.67	1571.84	-0.17	1	53	1	VGTAMLYRSVDR
<input checked="" type="checkbox"/> 61	1051.86	2101.70	2102.05	-0.35	0	(84)	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 62	1051.86	2101.70	2102.05	-0.35	0	(51)	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 63	1051.86	2101.71	2102.05	-0.34	0	(78)	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 64	1051.86	2101.71	2102.05	-0.34	0	(77)	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 65	1051.87	2101.73	2102.05	-0.32	0	90	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 66	781.60	2101.78	2102.05	-0.27	0	(56)	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 67	781.60	2101.79	2102.05	-0.26	0	(52)	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 68	781.62	2101.82	2102.05	-0.23	0	(46)	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 69	1051.93	2101.85	2102.05	-0.20	0	(67)	1	DLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 76	825.29	2472.84	2473.23	-0.39	1	(45)	1	DQKDLILQGDATTGTDGHELETR
<input checked="" type="checkbox"/> 77	825.30	2472.86	2473.23	-0.37	1	53	1	DQKDLILQGDATTGTDGHELETR

If your results look like this, with the same peptide identified over and over again, then it is likely that something is wrong with the data reduction settings. Ideally, there should only be two matches here, one for the 2+ precursor and one for the 3+. By summing together identical spectra, you gain in three ways: (i) the signal to noise of the summed spectrum is improved, making the identification more reliable, (ii) the report is more concise, (iii) the search is faster

Peak detection, peak detection, peak detection

Especially critical for Peptide Mass Fingerprints

- A tryptic digest of an “average” protein (30 kDa) should produce of the order of 50 peptide peaks

Time domain summing of LC-MS/MS data is very important

If in doubt, throw it out

- MS/MS spectra from low mass precursors (< 700 Da)
- And any spectrum with less than ~ 10 peaks

There is little point in searching MS/MS spectra from low mass precursors. Short peptides can occur by chance in large databases, so carry limited value for identification purposes. I recommend setting a cut-off at 700 Da, (not m/z 700).

To make searches as efficient as possible, it is also worth filtering out any MS/MS spectrum with less than around 10 peaks. You’re unlikely to get a meaningful match from a very sparse spectrum, so why waste time searching it?

Practical Tips

Modifications

- Fixed / static modifications cost nothing
- Variable / differential modifications are very expensive
- Use minimum variable modifications, especially for PMF
 - Maybe oxidation of M
 - Maybe alkylation of C

Modifications in database searching are always handled in two ways. First, there are the fixed or static or quantitative modifications. An example would be the efficient alkylation of cysteine. Since all cysteines are modified, this is effectively just a change in the mass of cysteine. It carries no penalty in terms of search speed or specificity.

In contrast, most post-translational modifications do not apply to all instances of a residue. For example, phosphorylation might affect just one serine in a peptide containing many serines. These variable or differential or non-quantitative modifications are expensive, in the sense that they increase the time taken for a search and reduce its specificity. This is because the software has to permute out all the possible arrangements of modified and unmodified residues that fit to the peptide molecular mass. As more and more modifications are considered, the number of combinations and permutations increases geometrically. The so-called combinatorial explosion.

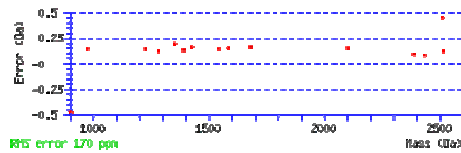
Hence, it is very important to be as sparing as possible with variable modifications. Especially in a peptide mass fingerprint, where the increase in the number of calculated peptides quickly makes it impossible to find a statistically significant match.

Obviously, if the point of the search is to find a particular modification, you have no choice. But, if the aim of the search is to identify as many proteins as possible, the best advice is to use a minimum of variable modifications.

Practical Tips

Make a reasonable estimate of mass error

- Don't just guess, run a standard



Search a comprehensive, non-identical database

- NCBI nr, UniRef100, MSDB
- More about databases later

Making an estimate of the mass accuracy doesn't have to be a guessing game. Most search engines include graphs of mass errors in the search report. Just search a standard and look at the error graphs for the strong matches. You'll normally see some kind of trend. Add on a safety margin and this is your error estimate.

I'll be returning to the subject of databases later. For now, the executive summary is: search a comprehensive, non-identical database, not a non-redundant one. NCBI nr, UniRef100, and MSDB are ideal for general purpose searching.

Practical Tips

Enzyme

- Loose trypsin (cleaves KP, RP)
- Semi-specific trypsin
- Only use “no enzyme” if strictly necessary
- Set missed cleavages by inspection of standards

Protein MW

- Processed protein may be much shorter than database sequence

Protein pI

- Adds little specificity.

The vast majority of searches are of tryptic digests. Although some people like to perform searches without enzyme specificity, and then gain confidence that a match is correct if the match is tryptic, I don't think this is a good idea. I normally choose loose trypsin, which cuts after K or R even when the next residue is P. If there is evidence for non-specific cleavage, then a semi-specific enzyme would be my next choice. This allows one end of the peptide to be non-specific, but not both. Only abandon enzyme specificity if you must, such as when searching endogenous peptides.

A missed cleavage parameter should be set by looking at the successful search results to see how complete your digests are. Setting it far too high or far too low is nearly as bad as setting the wrong mass tolerance.

It is possible to constrain PMF searches by the mass or pI of the protein. I'm not a big fan of doing this. It adds little specificity to the search, and there is the risk of excluding the correct match because the processed protein was very different in mass or pI from the database entry.

Practical Tips

Don't cheat!

- Iteratively adjusting search parameters to get a better score can give misleading results
- Beware of
 - Narrowing the taxonomy
 - Reducing mass tolerances
 - Removing modifications
 - Selecting spectra or mass values

Set search parameters using standard samples

It is easy to distort the search results without realising.

Basically, it is risky to adjust the search parameters interactively to get a better score for an unknown.

For example, you search the complete database and don't get a significant match. However, a very interesting looking protein is near the top of the list, surrounded by some others that are clearly wrong. You change the taxonomy filter so as to exclude the "wrong" proteins. Sorry, but this is cheating.

Search parameters should be set using standards. Broadening the search if you get a negative result is usually OK, but not narrowing the search.

Scoring



➤Elias, J. E., et al., *Intensity-based protein identification by machine learning from a library of tandem mass spectra*, Nature Biotechnology 22 214-219 (2004)

I borrowed this figure from a publication by the Steve Gygi group because it has to be the most colourful of all the scoring algorithms

Finding a match

Filter

- Sequence tag

Relative / arbitrary score

- Count number of matches
- Cross correlation

Absolute score

- *A priori* classical or Bayesian probability
- Post-search score normalisation.

The way in which we find or judge or score a match is the most critical part of the whole method. There are three fundamentally different ways of finding a match:

First, a search can be a simple filter on the database. The original sequence tag was a filter. There may be no matches, there may be one, there may be hundreds. All matches are equal. If there are multiple matches, it is up to the user to make a decision which, if any, is the preferred match.

A second way to find a match is to use an arbitrary score. Anything which is a good discriminator between matches classified as correct and matches classified as incorrect. An example of an arbitrary score in PMF is counting the number of matched mass values. An example of an arbitrary score in MS/MS is the Sequest cross-correlation coefficient. Arbitrary scores may be very sensitive, and very good discriminators. The difficulty is defining a criterion to judge whether a match is real or not.

The third mechanism is an absolute score. These scoring schemes are invariably based on Bayesian or classical probability. The scoring algorithm can be intrinsically probabilistic, or it can be a normalisation procedure applied to an arbitrary score.

PMF: Perfectly Meaningless Fingerprint

- Dates of birth of first 10 US Presidents

- Convert to “peptide mass” using DDMM.YY

President	Date of birth	“Peptide mass”
<i>George Washington</i>	<i>22 February 1732</i>	<i>2202.32</i>
<i>John Adams</i>	<i>30 October 1735</i>	<i>3010.35</i>
<i>Thomas Jefferson</i>	<i>13 April 1735</i>	<i>1304.35</i>
<i>James Madison</i>	<i>16 March 1751</i>	<i>1603.51</i>
<i>James Monroe</i>	<i>28 April 1758</i>	<i>2804.58</i>
<i>John Quincy Adams</i>	<i>11 July 1767</i>	<i>1107.67</i>
<i>Andrew Jackson</i>	<i>15 March 1767</i>	<i>1503.67</i>
<i>Martin van Buren</i>	<i>5 December 1782</i>	<i>512.82</i>
<i>William Harrison</i>	<i>9 February 1773</i>	<i>902.73</i>
<i>John Tyler</i>	<i>29 March 1790</i>	<i>2903.90</i>

To illustrate the difference between the three ways of finding a match, I want to use a data set that is clearly rubbish. This PMF isn't a peptide mass fingerprint, it's a perfectly meaningless fingerprint.

Rather than choose random numbers, which you might think I selected to suit my argument, I've taken the dates of birth of the first ten US presidents and converted them to mass values as shown. Hopefully, no-one expects to get a positive protein identification from this data set.

Filter

Require all masses to match

- Search conditions
 - Trypsin, 2 missed cleavages
 - Mass tolerance ± 0.5 Da
 - Average masses
 - No modifications
- Swiss-Prot
 - Failure (maximum number matched is 8)
- NCBI nr
 - Success (2 entries, nesprin 1 and titin)

For our filter, let's be stringent, and require all 10 mass values to match. The PMF search conditions are fairly standard.

If we search Swiss-Prot, our filter fails. The maximum number of matches of any entry is 8

If we search NCBI nr, our filter succeeds. All 10 mass values can be matched by nesprin and titin!

gi|55627590|ref|XP_518815.1| PREDICTED: nesprin 1 [Pan
troglyodytes] MW: 1038380 Da pI: 5.5

President	"Peptide mass"	Matched mass	Sequence
<i>George Washington</i>	2202.32	2202.47	<i>MQNLNRHWSLISSQTTER</i>
<i>John Adams</i>	3010.35	3010.45	<i>LLDPEDVDVDKPKDEKSIMTYVAQFLK</i>
<i>Thomas Jefferson</i>	1304.35	1304.49	<i>QLKSVKEEQSK</i>
<i>James Madison</i>	1603.51	1603.65	<i>GGSDSSLSEPGGRSGR</i>
<i>James Monroe</i>	2804.58	2804.19	<i>SCQVALQEHEALEEALQSMWSWVK</i>
<i>John Quincy Adams</i>	1107.67	1107.26	<i>VWIEQFER</i>
<i>Andrew Jackson</i>	1503.67	1503.77	<i>LQNLQDAADMKK</i>
<i>Martin van Buren</i>	512.82	512.54	<i>DYSK</i>
<i>William Harrison</i>	902.73	903.07	<i>LEEQKKK</i>
<i>John Tyler</i>	2903.90	2904.22	<i>EEITIQVHEETANTIQRKLEQHK</i>

Which is the problem with a filter. It leaves us to do all the work, deciding whether the match is meaningful or not. If you didn't know the source of the data, and needed to make a judgement on whether this match was real, you have to start pounding your calculator.

The sharp eyed will notice that nesprin is a pretty big protein, ~ 1 MDa. Titin is even bigger, at ~ 3 MDa. Which is why simply counting the number of matches is not such a great way of scoring a peptide mass fingerprint. Throw in a few modifications and you can match almost anything to these mega proteins

Relative / arbitrary score

MOWSE

- <http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse>

```
Database: /data/mowse/owl
Data file: /tmp/44701116780804.data
Reagent: Trypsin
Tolerance: 0.10
Sequence MW: 0.0
MW filter: 0.00%
Pfactor: 0.20
```

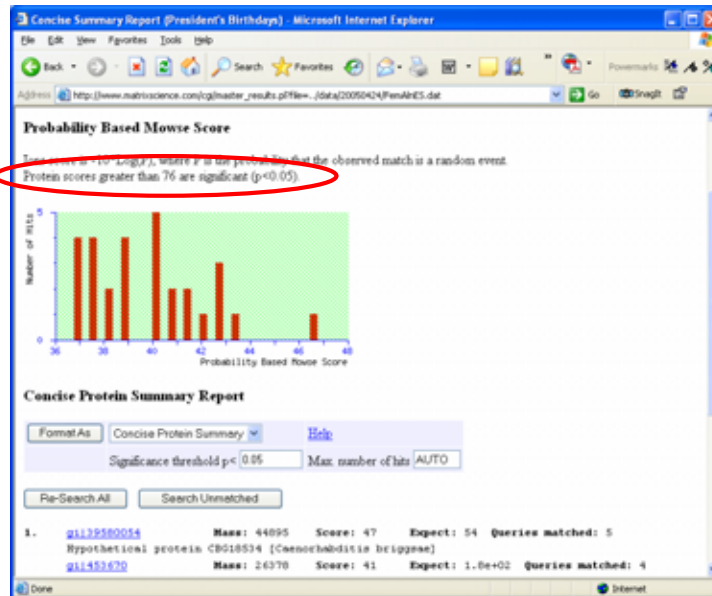
```
1 : PUTA_ECOLI      5.799922e+04 143814.8  0.600
    PROLINE DEHYDROGENASE (EC 1.5.99.8) (PROLINE OXIDASE)
    MW      START  END      SEQ
    2905.2  1148   1173   EWAANRPELQALCTQYGELAQAGTOR
    2804.3  645    670    WQALPMLQPFVAAGEMSPVINPAEFK
    1604.9  410    423    CPLVIDYLIDLATR
    1503.7  591    604    LAQQEGQTGLPHPK
    1107.2  735    744    TFSNAIAEVR
    903.1   1267   1275   ALCEA VAAR
    NO MATCH      3010.3  2202.3  1304.3  512.8
```

This problem drove Darryl Pappin and colleagues to develop the Mowse scoring scheme, in which each mass match in a PMF contributes to the score according to the size of the peptide and the size of the protein. A small peptide from a large protein carries the lowest score, a large peptide from a small protein carries the highest. This counteracts the “titin effect”, where the largest proteins will always pick up the most mass matches.

If we submit our meaningless data set to Mowse, the top match is proline oxidase, with 6 matches out of 10. This is an improvement in the sense that nesprin and titin now have lower scores, despite getting 10 mass matches each. However, we still have the problem of deciding whether the match to proline oxidase is correct.

The difficulty is that the scores for good matches vary widely from data set to data set. Scores depend on factors such as the number of mass values and the mass tolerance, which precludes the use of a fixed threshold. One has to judge whether a match might be interesting according to whether it is an outlier. That is, whether there is a large enough score gap between the best match and the second best match.

Probability based scoring



If we submit the same search to an algorithm that uses probability based scoring, such as Mascot, we still get a list of matches, but the report tells us that these matches are not statistically significant. The score threshold for this search is 76, and the top scoring match from our meaningless fingerprint is 47. The graph is a histogram of the scores of the top ten matches and, as you see, all of them are in the area shaded green to indicate random, meaningless matches.

What is probability based scoring?

We compute the probability that the observed match between the experimental data and mass values calculated from a candidate protein or peptide sequence is a random event.

The 'correct' match, which is not a random event, has a very low probability.

What exactly do I mean by probability based scoring?

We calculate, as accurately as possible, the probability that the observed match between the experimental data, and mass values calculated from a candidate peptide or protein sequence, is a random event.

The real match, which is not a random event, then has a very low probability.

Probability based scoring enables standard statistical tests to be applied to results

In a database of 1,000,000 entries, a 1 in a 100 chance of getting a false positive match is a probability of
 $P = 1 / (100 \times 1,000,000)$

The calculated probability may be converted into a score or may be reported as an expectation value: how often you would expect to get a match this good or better purely by chance. This is possible because, if we are working with true probabilities, we just have to multiply the acceptable false positive rate by the number of trials. In the case of a PMF, the number of trials is the number of entries in the database. In the case of an MS/MS search, the number of trials is the number of candidate peptides.

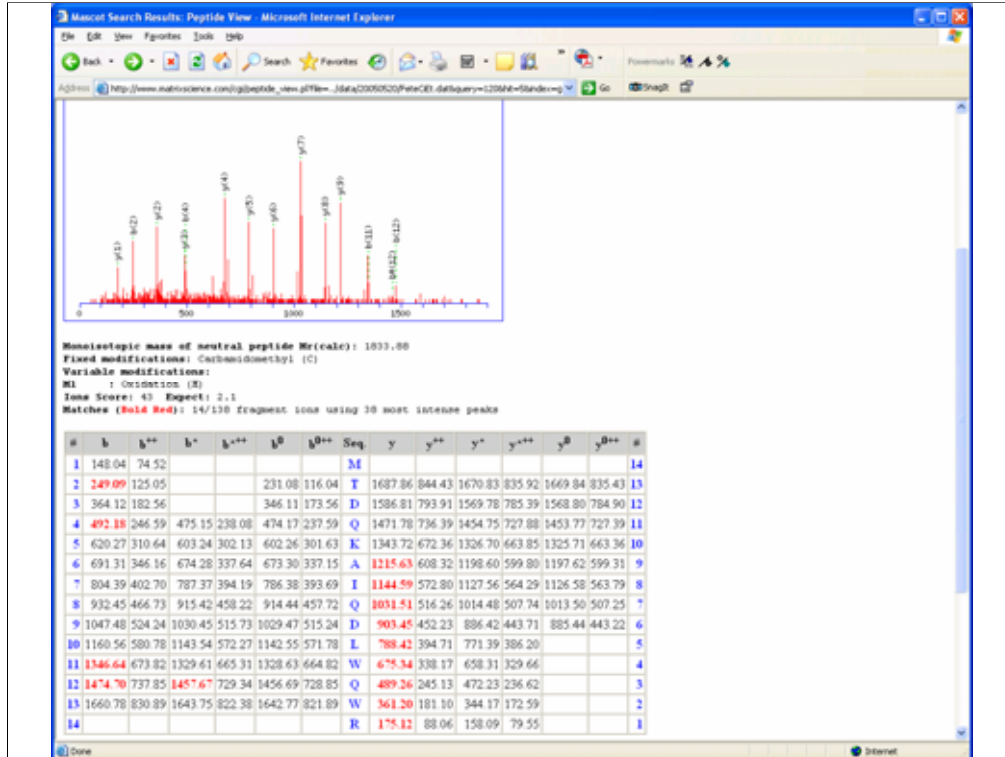
So, for a PMF, if our acceptable false positive rate is 1%, and we have a million entries in the database, we are looking for probabilities of less than $10E-8$.

Why is probability based scoring important?

- Human (even expert) judgment is subjective and can be unreliable

Why is probability based scoring important?

First and foremost, because it is very difficult to judge whether a match is significant or not by looking at the spectrum. Let me illustrate this with an example



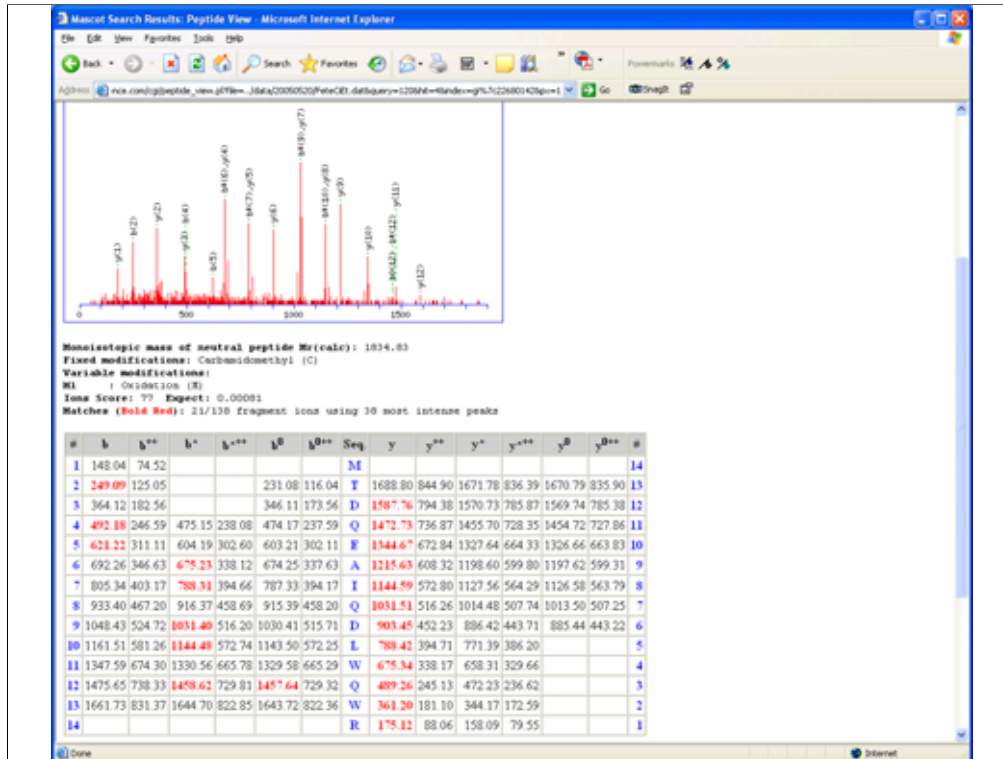
This match has a good, unbroken run of y ions plus a few b ions. All the major peaks seem to be labelled. Could such a good match have occurred by chance?

You cannot tell, because you can match anything to anything if you try hard enough.

If I say that I tossed a coin ten times and got ten heads in a row, does that mean there was something strange about the coin, like it had two heads? You cannot tell, because you need to know how many times I tossed the coin in total. If I picked it up off the table, tossed it ten times, then put it down, yes, that would suggest this was not a fair coin. However, if I tossed it ten thousand times, I would expect to get ten heads in a row more than once.

So, it isn't just a matter of how good the match is, i.e. how many y or b ions you found, it's a case of how hard you tried to find the match. In the case of a database search, this means how large is the database, what is the mass tolerance, how many variable modifications, etc., etc. These are very difficult calculations to do in your head, but they are easy calculations for the search engine.

If we look at the expectation value for this match, it is 2.1. That is, we could expect to get this match purely by chance. It looks good, but it's a random match.



If I show you a better match, then it is easy to dismiss the previous one as inferior. We can all make that judgement very easily. This match has an expectation value of less than 1 in 1000. It is definitely not random.

The challenge is, what if you don't have the better match to compare against? Maybe this sequence wasn't in the database. If you only had the inferior match, how would you decide by looking at it whether it was significant or not?

The other interesting question is whether this is the "correct" match. Who can say that a better match isn't possible, where we get the last y ion or some more of the b ions fall into line?

Why is probability based scoring important?

- Human (even expert) judgment is subjective and can be unreliable
- Standard, statistical tests of significance can be applied to the results
- Arbitrary scoring schemes are susceptible to false positives.

If we use probability based scoring, we can apply standard, statistical tests of significance to the results.

If we don't do this, then how do we know what the level of false positives is? It could be 1 in 1000 or 5% or 50%. In any statistical process, such as database matching, there will always be some level of false positives. The important thing is to know what it is.

Can we calculate a probability that a match is correct?

Yes, if it is a test sample and you know what the answer should be

- Matches to the expected protein sequences are defined to be correct
- Matches to other sequences are defined to be wrong

If the sample is an unknown, then you have to define “correct” very carefully:

- The best match in the database?
- The best match out of all possible peptides?
- The peptide sequence that is uniquely and completely defined by the MS data?

Probability based scoring tells you the probability that the match is random. This is, the probability that the match is meaningless. Many people would prefer a probability that the match is correct. Is this possible?

It is certainly possible if you are analysing a known protein or standard mixture of proteins. If you know what the sequences are, or think you know, then the matches to the known sequences are defined to be correct and those to any other sequence are defined to be wrong.

If the sample is an unknown, then it is difficult even to define what is meant by a correct match.

Is the correct match the best match in the database? Certainly not ... this would be a false positive if the correct sequence was not in the database.

What about the best match out of all possible peptides. Yes, a reasonable definition, but not a very practical one. This is what we try to find in de novo sequencing. The reason for searching a database is that the data quality are not good enough for reliable de novo, so we reduce the size of the search space to the content of the chosen sequence database.

How about the peptide sequence that is uniquely and completely defined by the MS data? This is equally impractical. One rarely, if ever, sees a mass spectrum perfect enough to meet this criterion

Peptide Summary Report (Annexin) - Microsoft Internet Explorer

Address: http://41-jsc/nascol/jgi/master_results.pl?file=.../data/2001016/P209940.dat

Mass: 35874 Score: 700 Queries matched: 14

601512345F1 NIH_H0C_71 Homo sapiens cDNA clone IMAGE:3913811 5'

Check to include this hit in archive report

Query	Observed	Hr(expt)	Hr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
12	415.19	828.36	828.51	-0.14	0	33	58	1	NALLSLAK
45	607.16	1212.31	1212.53	-0.21	0	70	0.015	1	DITSDTSGDFFR
52	631.70	1261.30	1261.59	-0.29	0	69	0.019	1	TPAQFPADELK
62	694.25	1386.49	1386.76	-0.27	0	73	0.0065	1	GVDEATIIDILTK
21	515.20	1542.58	1542.86	-0.28	1	46	3.8	1	GVDEATIIDILTKR
28	547.49	1639.45	1639.77	-0.32	1	(41)	11	1	DLAKDITSDTSGDFFR
29	820.75	1639.40	1639.77	-0.29	1	52	0.09	1	DLAKDITSDTSGDFFR
102	851.77	1701.52	1701.88	-0.36	0	82	0.00083	1	GLGTDIEDLIEILASR
105	870.21	1738.41	1738.73	-0.32	0	82	0.00092	2	SEDFGVNEDLGDSDAR + Methyl ester (DE)
15	476.92	1903.67	1904.03	-0.36	1	22	9.9e+002	1	AAFLQETGRPLDELTK
13	Top scoring peptide matches to query 105								
13	Score greater than 64 indicates identity								
13	Status bar shows all hits for this peptide								
Score	Delta	Hit	Protein	Peptide					
99.0	-0.32	2+	gi 10347940	SEDFGVNEDLADSDAR	Expect 1.8E-5				
82.0	-0.32	1	gi 10348033	SEDFGVNEDLGDSDAR	Expect 9.2E-4				
66.0	-0.32	5	gi 10330826	SEDFGVNEDLGDSDGR	Expect 0.037				
45.6	-0.35	0	gi 10345301	SEDFGVNEDLADSDAK	Expect 4.0				
24.1	-0.45			SFNKASINMLRQCR					
23.0	0.48			LIPVKAIDSEKQQR					
21.9	0.54			ECPYGLIMLRPDK					
20.5	0.63			CPNCLLICKPTSR					
20.4	-0.31			ECWRECEWFCAR					
5				CFVSEGLWASVSR					
5									
21	515.20	1542.58	1542.86	-0.28	1	46	3.8	1	GVDEATIIDILTKR
28	547.49	1639.45	1639.77	-0.32	1	(41)	11	1	DLAKDITSDTSGDFFR

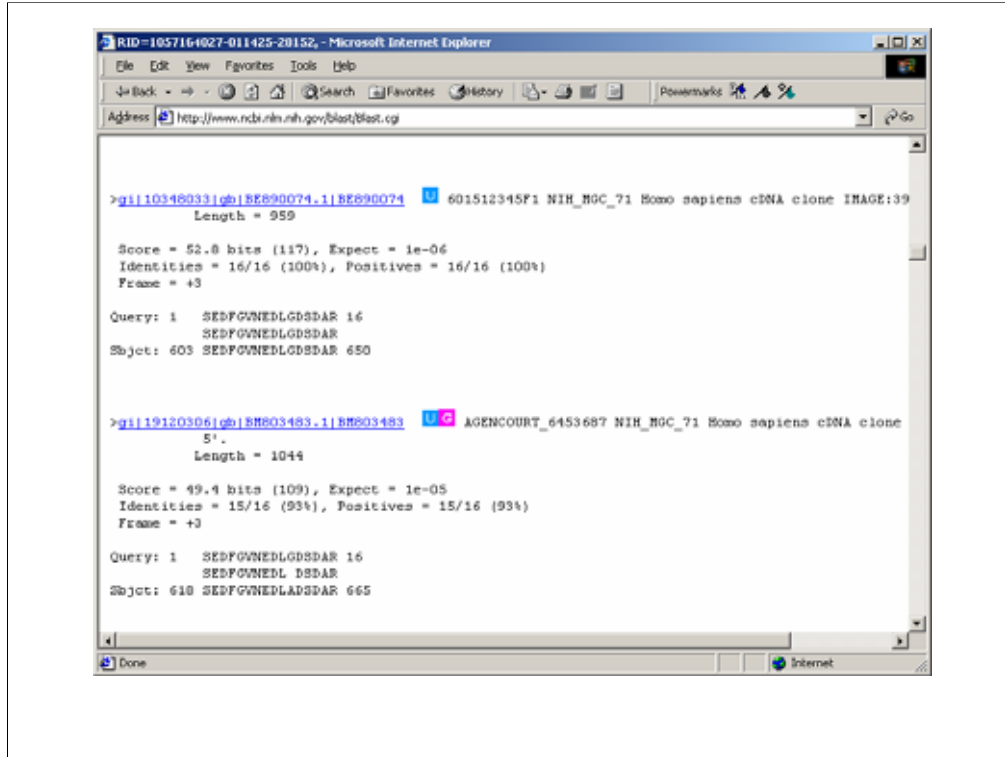
2. 601512345F1 NIH_H0C_71 Homo sapiens cDNA clone IMAGE:3913811 5'

Check to include this hit in archive report

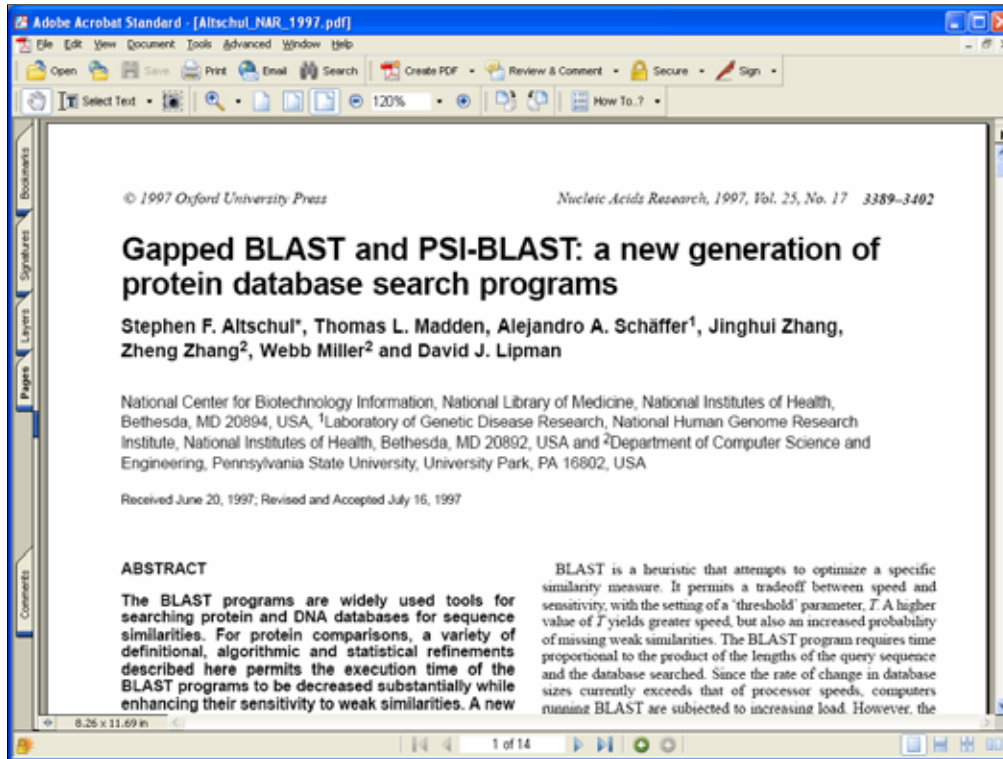
Query	Observed	Hr(expt)	Hr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
41	607.16	1212.31	1212.53	-0.21	0	70	0.015	1	DITSDTSGDFFR
5	631.70	1261.30	1261.59	-0.29	0	69	0.019	1	TPAQFPADELK
6	694.25	1386.49	1386.76	-0.27	0	73	0.0065	1	GVDEATIIDILTK
21	515.20	1542.58	1542.86	-0.28	1	46	3.8	1	GVDEATIIDILTKR
28	547.49	1639.45	1639.77	-0.32	1	(41)	11	1	DLAKDITSDTSGDFFR

This is a typical MS/MS search result, where we see a series of high scoring homologous peptides. The sequences of the top four matches are very similar, and their expectation values vary from random through to very unlikely to be random. The best match has an expectation value of $2E-5$. However, we cannot be sure that this is an identity match to the analyte peptide. It is simply the best match we could find in the database. There is always the possibility that a better match exists, that is not in the database, so to call it the correct match can be misleading.

The important thing is that we have a mechanism to discard matches that are nothing more than random matches.



It is a similar situation in Blast, except that you have the luxury of seeing when you have a perfect identity match. The identity match has an expectation value of $1E-6$, which reminds us that it would be a random match if the database was a million times larger. The match with one different residue is not worthless, it has an expectation value of $1E-5$ and is a very good match. It just isn't as good a match as the one above.



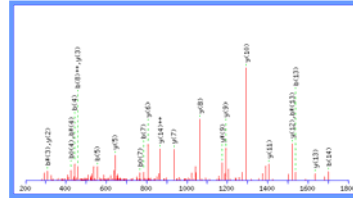
We can learn a lot from sequence homology searching, which deals with many similar issues on a sound, statistical basis.

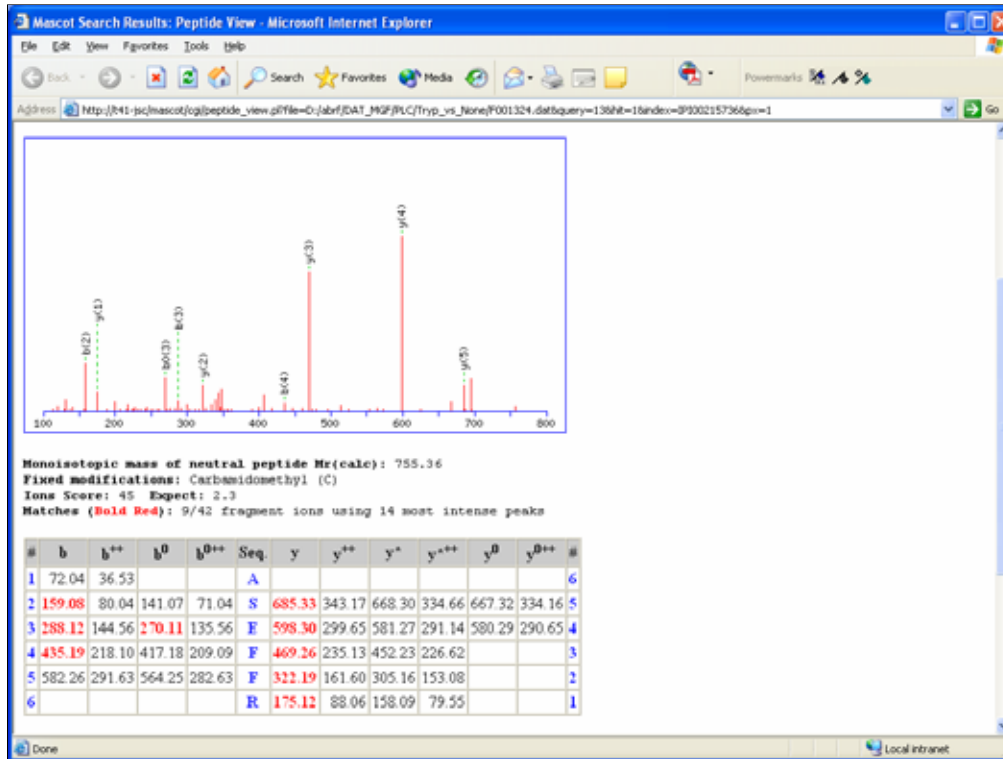
BLAST / FASTA

```
>gi145221061|gb|AA527066.1| 90 kDa heat shock protein (Megalocle rotundata)
Length = 235
Score = 54.5 bits (121), Expect = 2e-07
Identities = 15/15 (100%), Positives = 15/15 (100%)
Query: 1  NFDDITQEETGEFFK 15
          NFDDITQEETGEFFK
Sbjct: 161 NFDDITQEETGEFFK 175
```

- Sequence against sequence
- Can be used to find weak / distant similarity
- Can make gapped alignments

MS/MS-based ID





If we are doing probability based matching, we are not scoring the quality of the spectrum, we are scoring whether the match is random or not.

Even when the mass spectrum is of very high quality, if the peptide is so short that it could occur in the database by chance, then you will not get a very good score.

```
>gi11847921|sp|P22712|MPB1\_HUMAN C-myc promoter-binding protein (MPB-1) (MDP-1)
Length = 335

Score = 22.7 bits (46), Expect = 48
Identities = 6/6 (100%), Positives = 6/6 (100%)

Query: 1 ASEFFR 6
      ASEFFR
Sbjct: 150 ASEFFR 155

>gi11847921|sp|P06733|ENOA\_HUMAN Alpha enolase (2-phospho-D-glycerate hydro-lyase) (Non-neural
enolase) (NNE) (Enolase 1) (Phosphopyruvate hydratase)
Length = 434

Score = 22.7 bits (46), Expect = 48
Identities = 6/6 (100%), Positives = 6/6 (100%)

Query: 1 ASEFFR 6
      ASEFFR
Sbjct: 248 ASEFFR 253

>gi11847921|sp|P34147|RAC1\_DICDI RAS-related protein racA
Length = 598

Score = 22.7 bits (46), Expect = 48
Identities = 6/6 (100%), Positives = 6/6 (100%)

Query: 1 ASEFFR 6
      ASEFFR
Sbjct: 261 ASEFFR 266
```

The situation in a Blast search is identical. Even though this is a perfect identity match, the expectation value is 48. This is just a random match. Hence, the earlier tip to discard spectra from low mass precursors.

Post-search score normalisation

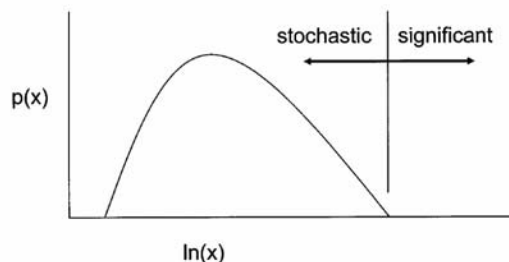


Figure 1. Schematic representation of a stochastic score distribution, as described in the text. Any peptide corresponding to a score within the body of the stochastic distribution cannot be confidently assigned as being a valid identification. A score higher than the right-hand boundary of the stochastic distribution may be assigned as potentially valid, with an associated expectation value.

➤Fenyo, D. and Beavis, R. C., *A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes*, Analytical Chemistry 75 768-774 (2003)

One method of transforming an arbitrary score into a probability is to determine the distribution of scores experimentally, and look for an outlier. As long as the number of matches is large, the score distribution will approximate to a normal distribution. An interesting match, with a high score, will be an outlier. The distance of the outlier from the tail of the distribution can be used to calculate an expectation value, which is the number of times we expect to get that score or better by chance.

Post-search score normalisation

Sequest-Norm

- MacCoss, M. J., et al., *Probability-based validation of protein identification using a modified Sequest algorithm*, Anal. Chem., 74(21) 5593-5599 (2002).

PeptideProphet

- Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search*, Anal. Chem., 74(20) 5383-5392 (2002).

SVM

- Anderson, D. C., et al., *A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores*, Journal of Proteome Research 2 137-146 (2003)

ANN

- Baczek, T., et al., *Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics*, Analytical Chemistry 76 1726-1732 (2004)

Prot_Probe

- Sadygov, R. G., et al., *Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases*, Analytical Chemistry 76 1664-1671 (2004)

There are very many publications concerning the transformation of Sequest scores into probabilities. Some are listed here. The first and last are from the Yates group.

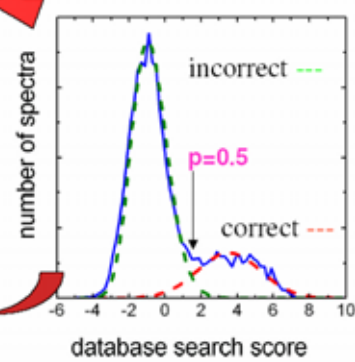
Statistical Model for Estimating the Accuracy of Peptide Identifications

entire dataset:

Spectrum	Peptide	Score	Probability
Spectrum 1	LGEYGH	4.5	1.0
Spectrum 2	FQSEEQ	3.4	0.97
Spectrum 3	FLYQE	1.3	0.01
...
Spectrum N	EIQKKF	2.2	0.3

probability

unsupervised learning



EM mixture model algorithm learns the most likely distributions among correct and incorrect peptide assignments given the observed data

Image courtesy of Institute for Systems Biology

One of the more widely discussed transforms is PeptideProphet from the Institute for Systems Biology. This derives an empirical discriminant function from a training set in which matches can be classified as correct or incorrect. This discriminant function is then fitted to other data sets using a technique known as expectation maximisation.

Validation & reporting tools



Validation

Search a “decoy” database

- Reversed entries sufficient for MS/MS with enzyme
- Randomized required for MS/MS without enzyme or PMF

Direct estimate of false positive rate

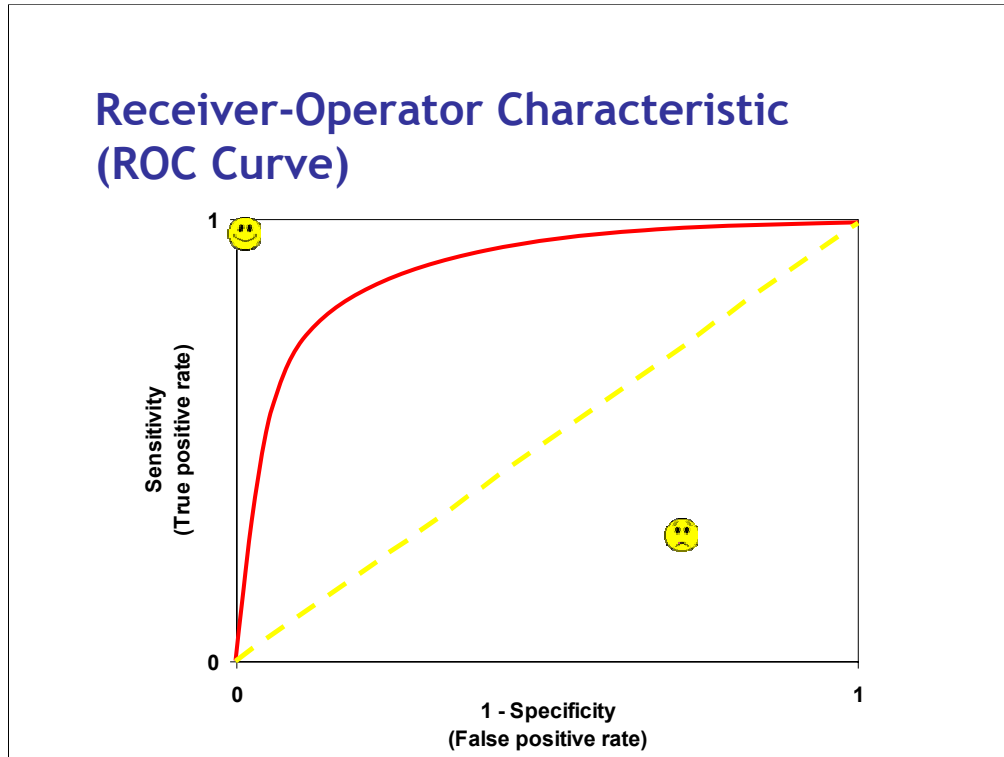
Whether you use an arbitrary scoring scheme or probability based scoring, it is good scientific practice to verify the actual level of false positives, so validating the scoring scheme.

One approach is to repeat the search, using identical search parameters, against a so-called decoy database. This is a database in which the sequences have been reversed or shuffled.

You do not expect to get any significant matches from the decoy database. So, the number of matches that are found is an excellent estimate of the false positive rate in the original search.

This is an excellent validation method for MS/MS searches of large data sets. It is not as useful for a search of a small number of spectra, because the numbers are too small to give an accurate estimate. Also, this method will not work for a two pass searches.

Receiver-Operator Characteristic (ROC Curve)

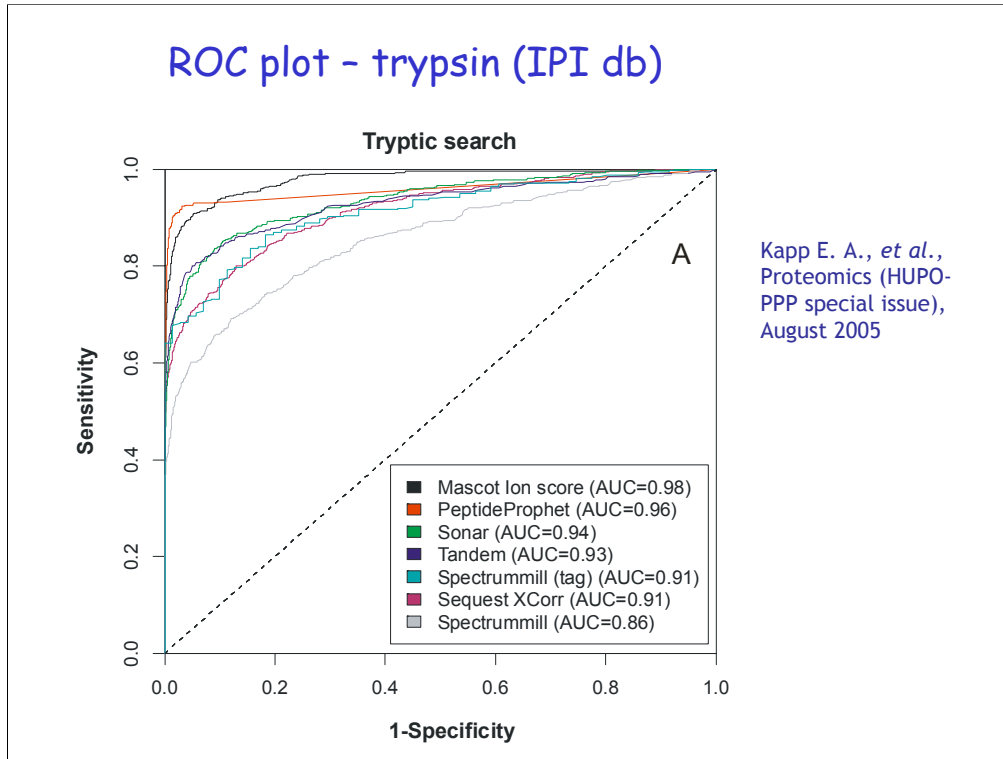


The accepted way to report true and false positive rates is a Receiver-Operator Characteristic curve.

If your scoring scheme was useless, and provided no discrimination between good and bad matches, then your ROC curve would follow the yellow dashed diagonal line. So, if your ROC curve wanders into the area below this diagonal, you have a serious problem. Tossing a coin would give better results

The curve shows what true positive rate we can expect, given an acceptable false positive rate. A good scoring scheme will try to follow the axes, as illustrated by the red curve, pushing its way up into the top left corner.

A real life scoring scheme can never actually reach the smiley face at the top left hand corner. The only point on the curve that has a zero false positive rate is the one at the origin, which has a zero true positive rate.



Real-life ROC curves come in a variety of shapes. This plot shows curves from the reprocessing of the data collected for the HUPO plasma proteome project, being performed at the Ludwig Institute in Melbourne.

Reporting Tools - tabulation

DTASelect

- <http://fields.scripps.edu/DTASelect/>

Interact

- <http://www.systemsbiology.org/Default.aspx?pagename=proteomicssoftware>

Silver

- <http://llama.med.harvard.edu/~fgibbons/cgi/SILVER/silver.cgi>

ProteinProphet

- <http://www.systemsbiology.org/Default.aspx?pagename=proteomicssoftware>

ROC curves are just one aspect of reporting search results. There are many tools available on the web that allow search results to be tabulated in different ways. This is just a selection.

Reporting Tools - relational database

DBParser

- Yang, X., et al., *DBParser: web-based software for shotgun proteomic data analyses*, J. Proteome Res. 3 1002-8 (2004)

EPIR

- Kristensen, D. B., et al., *Experimental Peptide Identification Repository (EPIR): An Integrated Peptide-Centric Platform for Validation and Mining of Tandem Mass Spectrometry Data*, Mol Cell Proteomics 3 1023-38 (2004)

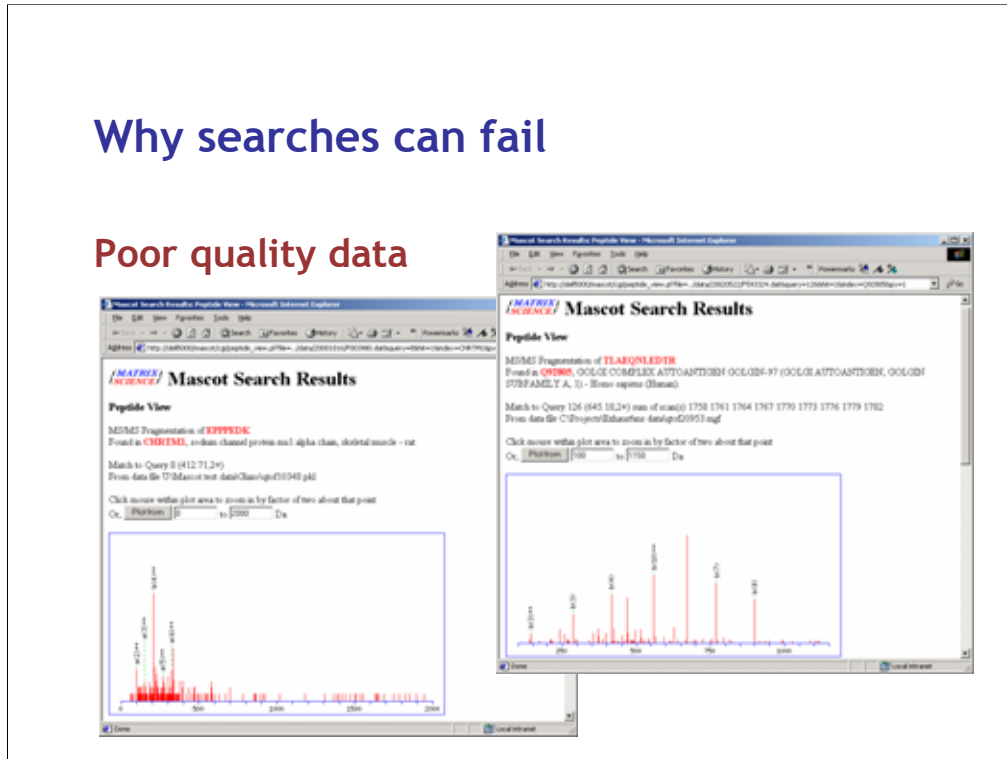
For the ultimate flexibility in reporting, the search results need to be imported into a relational database. DBParser is free, EPIR is commercial. There are several commercial relational database products from the instrument vendors and the search engine developers

Why searches can fail



Why searches can fail

Poor quality data

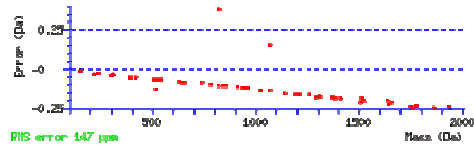


The number one cause of failure is poor quality data. This is always the first place to look when there is a problem.

However, a search can fail even though the data look great

Why searches can fail

- **Incorrect determination of precursor charge**
 - Particularly for higher charge states
 - Sodiated ions?
- **Under-estimated mass measurement error**
 - Peak detection may sometimes choose ^{13}C peak
 - Accuracy, not precision



Calling the wrong precursor charge is a common problem for low resolution instruments. It can also be a problem for higher charge states on instruments with good resolution. The instrument data system has very little time to make a decision on this, so may use a relatively crude algorithm to guess the charge state. Re-processing the data to determine the precursor m/z and charge may give better results.

And, maybe the charge isn't a proton ... high levels of sodium may produce sodiated ions. Peak detection routines sometime choose the ^{13}C peak rather than the ^{12}C peak, leading to unexpectedly large errors.

Another common problem is being over optimistic about mass accuracy. Remember, it is accuracy that matters, not precision. If your error graph looks like this, wouldn't it be worth calibrating? Then, it may be possible to search with a tolerance of $\pm 0.1\text{Da}$ rather than $\pm 0.5\text{Da}$.

Why searches can fail

Protein or peptide sequence not in the database

- Less of a problem in PMF, unless working with unusual organism
- MS/MS: SNPs, sequencing errors, splice variants

Unsuspected chemical or post-translational modifications

➤ Karty, J. A., *et al.*, *Artifacts and unassigned masses encountered in peptide mass mapping*, *J. Chrom. B* 782 363-83 (2002)

An obvious cause for a search to fail is if the sequence is not in the database. As the databases get larger, this is becoming less of a problem for well represented organisms. A PMF will often give a reasonable match to a homologous protein from a related organism, because a sufficient number of the peptides will be the same. For an individual MS/MS spectrum, the exact peptide sequence is required, and a single base change, due to a SNP (single nucleotide polymorphism) or a sequencing error, can prevent matching.

Unsuspected modifications have identical consequences. This paper from the Reilly group provides an excellent analysis of the origins of many unexplained mass values in peptide digests.

Why searches can fail

•Enzyme non-specificity

➤Olsen, J. V., *et al.*, *Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues*, *Mol. and Cellular Proteomics* 3 608-14 (2004)

- In-source fragmentation - ragged ends?
- Contamination; endogenous proteases?
- Worth using “loose” trypsin, cleaves RP, KP

Mixture

- Protein mixture in PMF
- Isobaric, co-eluting peptide mixture in MS/MS

Enzyme non-specificity became the topic of a very lively debate recently on the ABRF email discussion list. The trigger was this paper from Mann’s group.

They used the high accuracy of an FT instrument to study some 1000 tryptic peptides. There was no evidence for non-specific cleavage apart from some peptides with an N-terminal proline, which had originally been tryptic, but had lost N-terminal residues either in acid conditions in solution or by “nozzle-skimmer” fragmentation in the ion source.

Not everyone agrees with this point of view, but I think there is general agreement that trypsin is very accurate when the digest is properly conducted. Obviously, if the trypsin is contaminated, the picture may be different. Yates has pointed out that a cell lysate may contain endogenous proteases.

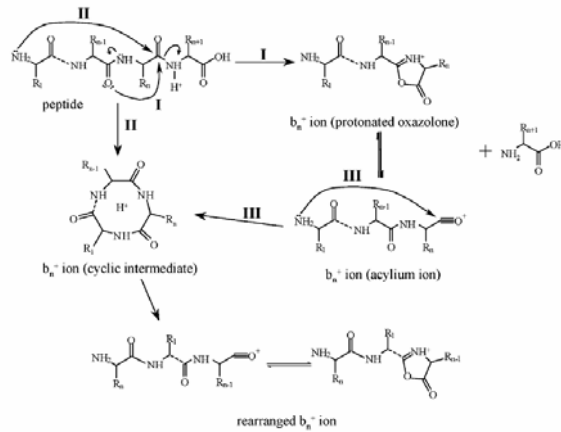
“Loose” trypsin, where cleavage is allowed before proline, often picks up some additional matches.

PMF searches can fail if the sample is a mixture. Similarly, a mixed MS/MS spectrum is unlikely to get a good match

Why searches can fail

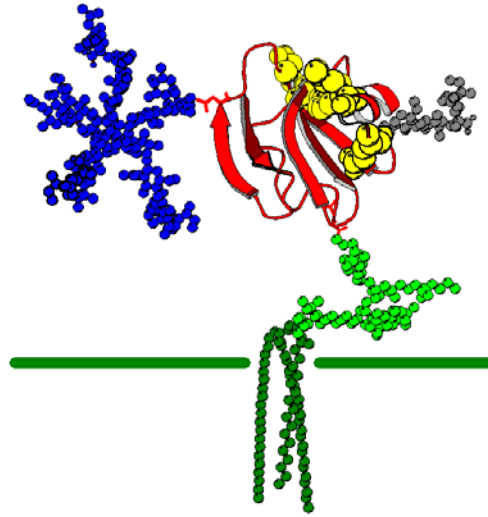
In-source re-arrangement

➤ Yague, J., *et al.*,
Peptide Rearrangement during Quadrupole Ion Trap Fragmentation: Added Complexity to MS/MS Spectra, *Anal Chem.* 75 1524-35 (2003)



This is the one that frightens me. b ions cyclising in the instrument and then ring opening at a different place to scramble the sequence. Hopefully, not a very common phenomenon.

Modifications



Oxford Glycobiology Institute

More about modifications

Unimod: Basic Data View - Microsoft Internet Explorer

Address: http://www.unimod.org/unimod.cgi?records_per_page=25&columns_to_view=full_name&columns_to_view=code

UNIMOD protein modifications for mass spectrometry

View All Records | Add Record

beta | Home | Options | Advanced Search | Logout

Search: Go

Result Set

Modification	Short name	Monoisotopic	Average	Composition	Details
Propionaldehyde +42	Propionald+42	42.046950	42.0797	H(6) C(3)	⌵
Carbamylation	Carbamyl	43.005814	43.0247	H C N O	⌵
Gamma-carboxylation	Gamma-carboxyl	43.989829	44.0095	C O(2)	⌵
Carboxylation	carboxyl	43.989829	44.0095	C O(2)	⌵
Ethanolation of Cys	EtOH	44.026215	44.0526	H(4) C(2) O	⌵
Oxidation to nitro	Nitro	44.985078	44.9976	H(-1) N O(2)	⌵
Acetate labeling reagent (N-term & K) (heavy form, +3amu)	AcetyL_heavy	45.029395	45.0582	H(-1) H2(3) C(2) O	⌵
Methyl methanesulfonate	MMTS	45.987721	46.0916	H(2) C S	⌵
Beta-methylthiolation	b-methylthiol	45.987721	46.0916	H(2) C S	⌵
Selenium replaces sulphur in Methionine	SeMet	47.944449	46.8950	S(-1) Se	⌵
cysteine oxidation to cysteic acid	Cysteic_acid	47.984744	47.9982	O(3)	⌵
MDA adduct +54	MDA54	54.010565	54.0474	H(2) C(3) O	⌵
Acrolein addition +56	Acrolein56	56.026215	56.0633	H(4) C(3) O	⌵
Propionate labeling reagent light form (N-term & K)	Propionyl_light	56.026215	56.0633	H(4) C(3) O	⌵
Iodoacetamide derivative	Carbamidomethyl	57.021464	57.0513	H(3) C(2) N O	⌵
Iodoacetic acid derivative	Carboxymethyl	58.005479	58.0361	H(2) C(2) O(2)	⌵
Hydroxyethanone	hydroxyethanone	58.005479	58.0361	H(2) C(2) O(2)	⌵
Propionate labeling reagent heavy form (+3amu), N-term&K	Propionyl_heavy	59.036279	59.0412	H(4) C13(3) O	⌵

It will be clear that comprehensive, accurate information about post translational and chemical modifications is a very important factor in the success of protein identification. I want to give a plug for Unimod, which is an on-line modifications database.

The screenshot shows the UNIMOD website in a Microsoft Internet Explorer browser window. The page title is "UNIMOD protein modifications for mass spectrometry". The address bar shows the URL: http://www.unimod.org/cgi/unimod.cgi?sort_field=mono_mass&sort_field2=full_name&first_record_to_display=0&v. The page content includes a search bar, navigation links, and a detailed record for "Acetyl_Light".

Record Details

Accession #	57	Short name	Acetyl_Light
Modification	Acetate labeling reagent light form (N-term & K)		
Composition	H(2) C(2) O	Monoisotopic	42.010565
		Average	42.0367
Specificity Definition 1			
Site	K	Position	Anywhere
Neutral Loss		Monoisotopic	
		Average	
Classification	Isotopic label		
Comment			
Specificity Definition 2			
Site	N-term	Position	Any N-term
Neutral Loss		Monoisotopic	
		Average	
Classification	Isotopic label		
Comment			
Notes and References			
Source	Journal Reference	Controlling Deuterium isotope effects in comparative proteomics. Zhang, Roujian; Sioma, Cathy S.; Thompson, Robert A.; Xiong, Li; Regnier, Fred E.. Department of Chemistry, Purdue University, West Lafayette, IN, USA. Analytical Chemistry (2002), 74(12), 2649-2654.	
Source	Journal Reference	Global internal standard technology for comparative proteomics. Chakraborty, Asish; Regnier, Fred E.. Department of Chemistry, Purdue University, West Lafayette, IN, USA. Journal of Chromatography, A (2002), 949(1-2), 173-184.	
Source	Journal Reference	Comparative proteomics based on stable isotope labeling and affinity selection. Regnier, Fred E.; Riggs, Larry; Zhang, Roujian; Xiong, Li; Liu, Peiran; Chakraborty, Asish; Seeley, Erin; Sioma, Cathy; Thompson, Robert A. Department of Chemistry, Pu	
Notes			
Curator	perner	Last Modified	2002-10-20 10:50:36

At the bottom right of the record details, there is a link for [Email Change Request](#).

Mass values are calculated from empirical chemical formulae, eliminating the most common source of error. Specificities can be defined in ways that are useful in database searching, and there is the option to enter mass-spec specific data, such as neutral loss information. This screen shot shows one of the better annotated entries, I can't pretend that all of them are this detailed. Nevertheless, it is a very useful, public domain resource that beats having to create your own list in an Excel spreadsheet or on the back of an envelope.

Two pass searching

First pass - simple search of entire database

- Minimal modifications
- Enzyme specificity

Second pass - exhaustive search of selected protein hits

- Wide range of modifications
- Look for SNPs
- Relax enzyme specificity

Examples

- Mascot - Error tolerant search
 - Creasy, D. M. and Cottrell, J. S., Error tolerant searching of uninterpreted tandem mass spectrometry data, *Proteomics* 2 1426-1434 (2002)
- X!Tandem - Model refinement
 - Craig, R. and Beavis, R. C., A method for reducing the time required to match protein sequences with tandem mass spectra, *Rapid Communications in Mass Spectrometry* 17 2310-2316 (2003)


There are many hundreds of modifications in Unimod, yet I've emphasised the importance of using the minimum number of variable modifications in a search. So, how are we supposed to find these unusual modifications?

If you are searching uninterpreted MS/MS data, the efficient way to find unusual modifications, as well as variations in the primary sequence, is a two pass search. The first pass search is a simple search of the entire database with minimal modifications. The protein hits found in the first pass search are then selected for an exhaustive second pass search.


Because only a handful of entries are being searched, search time is not an issue. The downside is that it is difficult to apply any kind of threshold to the results, or calculate expectation values, because the entries being searched have been pre-selected.

This type of search can be performed by manually selecting entries to create a new fasta database. More recently, it has been directly implemented into search engines such as Mascot and X!Tandem

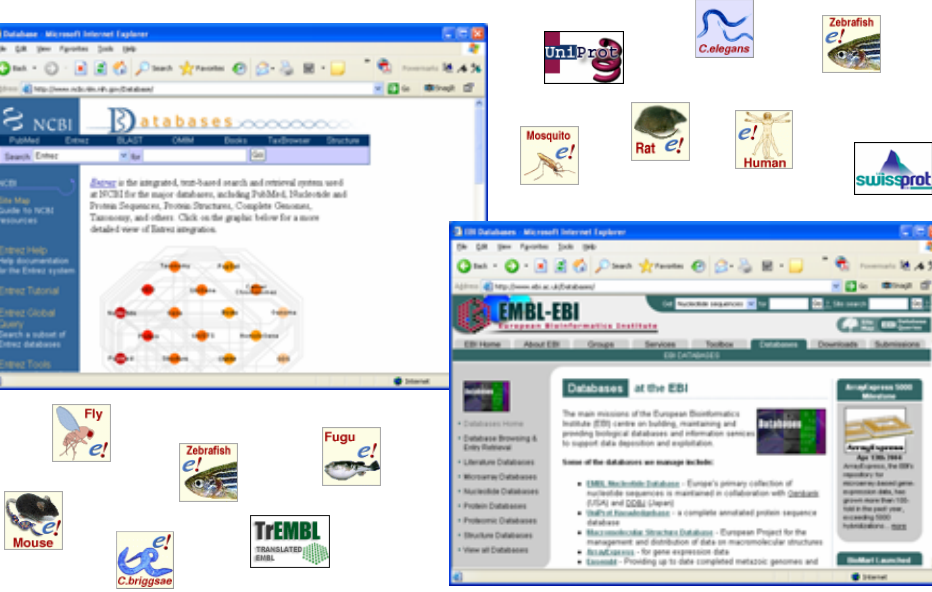
Sequence Databases



The screenshot shows the NCBI Databases homepage in a Microsoft Internet Explorer browser. The page features a search bar, navigation tabs (PubMed, Entrez, BLAST, OMIM, Books, Taxonomy, Structure), and a central graphic of a network diagram. The URL is http://www.ncbi.nlm.nih.gov/Databases/.



The screenshot shows the EMBL-EBI website in a Microsoft Internet Explorer browser. The page features a search bar, navigation tabs (Home, About EBI, Groups, Services, Tools, Conferences, Downloads, Submissions), and a central section titled 'Databases at the EBI'. The URL is http://www.ebi.ac.uk/Databases/.



A collection of logos for various species and databases, including UniProt, C. elegans, Zebrafish, Mosquito, Rat, Human, swissprot, Fly, Zebrafish, Fugu, Mouse, and TrEMBL (EMBL TRANSLATED). Each logo typically includes the species name and a stylized 'e!' symbol.

More about sequence databases

Sequence Databases

NCBI nr, UniRef100, MSDB (>2,000,000 entries)

- Comprehensive, non-identical

UniRef90, UniRef50, etc.

- Avoid non-redundant databases; need explicit sequences

Swiss-Prot (~180,000 entries)

- High quality, non-redundant; good for PMF

EST databases (>100,000,000 entries)

- Very large and very redundant
- Not suitable for PMF

Sequences from a single genome

- Very small databases may give misleading results
- Play safe by appending sequences to larger database

There are a huge number of database, and often it is not clear which is the appropriate one to choose for a search.

As mentioned earlier, the large, comprehensive, non-identical databases are the best choice for general purpose searching. Examples are NCBI nr, UniRef100, and MSDB.

Non-redundant databases are not ideal for database searching because you need the exact protein or peptide sequence to be explicitly represented in the database.

Swiss-Prot is acknowledged to be the best annotated database, but it is non-redundant. The Phenyx search engine actually reads Swiss-Prot annotations, so side-steps this problem. Swiss-Prot can also be a good choice for fast PMF searches, where the loss of one or two peptides may not be a concern.

The EST databases are huge. Worth trying with high quality MS/MS data if a good match could not be found in a protein database. Not advisable for PMF, because many sequences correspond to protein fragments.

Single genome databases are good for protein characterisation using MS/MS data. Small genomes are not such a good choice for protein identification. The statistics will be better if the sequences are appended to a larger database

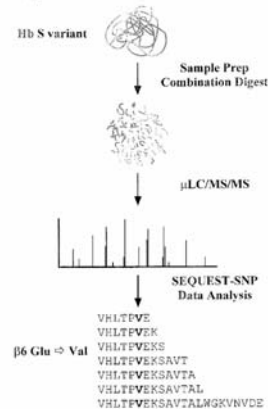
Searching Genomic DNA / SNP's

SNP's

- Digest with three different enzymes for multiple sequence coverage
- Search engine (Sequest) dynamically generates all possible SNP's

✦ Gatlin, C. L., *et al.*, *Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry*, *Anal. Chem.* 72 757-763 (2000)

Scheme 1. Approach Used to Identify Amino Acid Sequence Variations in Proteins^a



With most search engines, you are also able to search nucleic acid databases. Usually, these will be translated in all six reading frames.

Searching a nucleic acid database of ORFs (open reading frames) or coding sequences is not very different from searching a protein database.

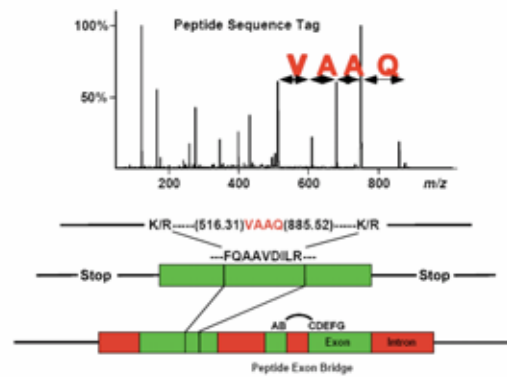
Searching a nucleic acid database is the best way to search for SNP's. One of the earliest examples of this was from the Yates group, where Sequest was used to identify haemoglobin mutants

Searching Genomic DNA / SNP's

Confirm and refine predicted genes

- Overpredict genes
- Use sequence tag to identify exon
- Use accurate MW data to refine exon - intron boundaries

➤ Kuster, B., *et al.*, *Mass spectrometry allows direct identification of proteins in large genomes*, *Proteomics* 1 641-650 (2001)

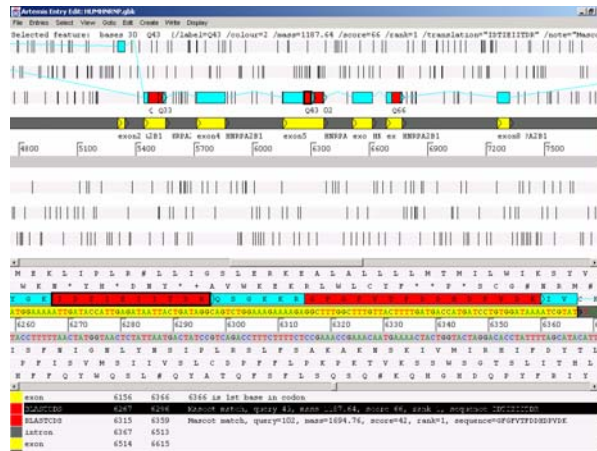


With genomic DNA sequence, there is the possibility to confirm the coding sequences of predicted genes, possibly finding errors. In this example from the Mann group, sequence tags were used to locate exons and then large numbers of accurate molecular mass values used to locate the exon / intron boundaries; like splattering paint through a stencil

Searching Genomic DNA / SNP's

Search genomic DNA with uninterpreted MS/MS data

- Look for splice variants, very small genes, etc.
- Lose ~ 20% of potential matches at exon - intron boundaries
- Choudhary, J. S., *et al.*, *Interrogating the human genome using uninterpreted mass spectrometry data*, *Proteomics* 1 651-667 (2001)



This is an early example of searching the human genome assembly using uninterpreted MS/MS data. The graphic shows the peptide matches displayed using a genome browser. The red bands are the peptide matches and the blue bands are the predicted coding sequences.

The disadvantage of searching raw genomic DNA with an exon / intron structure is that a fraction of the potential matches are lost because they straddle boundaries. In the case of the human genome, the fraction is about 20%.

Future Directions



Future directions

Fully integrated searching

- First & second pass searching of uninterpreted spectra
- Error tolerant Sequence Tag
- De novo
- Sequence homology

Data management

- Automation
- Data mining

We still have some way to go to integrate the many tools and techniques for extracting useful information from protein mass spectrometry data. Tools for automation, data management and data mining are limited in scope, and mostly support just one vendor's products.

Future directions

Structure-based spectrum prediction

- Tabb, D. L., et al., *Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides*, Anal. Chem. 75 1155-1163 (2003)
- Kapp, E. A., et al., *Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation*, Anal. Chem. 75 6251-6264 (2003)
- Elias, J. E., et al., *Intensity-based protein identification by machine learning from a library of tandem mass spectra*, Nature Biotechnology 22 214-219 (2004)

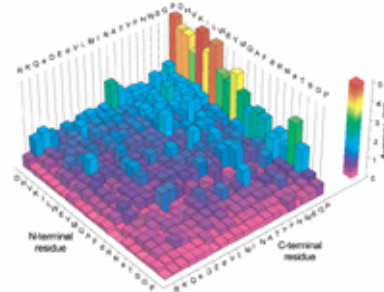


Figure 5. Three-dimensional plot summarizing relative intensity information for all residue combinations based on calculated average ARI values for doubly protonated peptides identified as peptides from the NCBI database. The color of the N- and C-terminal residue along each axis is determined by the specific residue's individual contribution to charge state. Residues grouping to enhanced cleavage are positioned toward the back of the plot such that synergistic cleavage interactions are highlighted.

There is a great deal of interest in trying to make better use of the intensity information in an MS/MS spectrum. Several groups have tried to extract useful rules from compilations of spectra. These three papers are a good introduction, and contain references to other work in this area. The figure is from the Kapp paper.

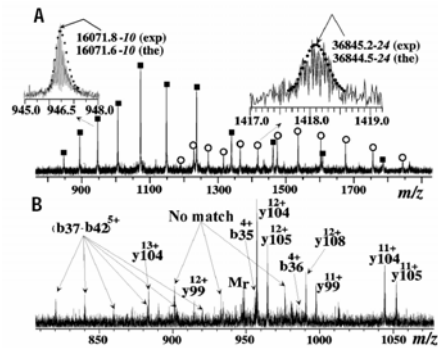
Future directions

Top-down proteomics

- Meng, F. Y., *et al.*, *Informatics and multiplexing of intact protein identification in bacteria and the archaea*, *Nature Biotechnology* 19 952-957 (2001)

Prosight PTM

- <https://prosigthptm.scs.uiuc.edu/>



I've focused on the matching of MS/MS data from peptides, but there is also interest in matching data from intact proteins, so-called top-down proteomics. The Kelleher group is one of the most active, and has developed a software package called Prosight

Future directions

Better use of accurate mass

- Schlosser, A. and Lehmann, W. D., *Patchwork peptide sequencing: Extraction of sequence information from accurate mass data of peptide tandem mass spectra recorded at high resolution*, *Proteomics*, 2(5) 524-533 (2002)
- Spengler, B., *De Novo Sequencing, Peptide Composition Analysis, and Composition-Based Sequencing: A New Strategy Employing Accurate Mass Determination by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*, *J. Am. Soc. Mass Spectrom.* 15 703-14 (2004)

One of the strengths of mass spectrometry for small molecules is the possibility to get an empirical chemical formula direct from an accurate mass measurement. Wolf Lehmann and colleagues pointed out that we could do a similar thing with the low mass fragment ions in an MS/MS spectrum. If the mass accuracy is high enough, the amino acid composition of a fragment can be reduced to just a few possibilities, maybe only one. I don't think any of the mainstream search engines take advantage of this yet.

Bernhard Spengler has applied the same concept to improve de novo sequence interpretation

Future directions

Data Repositories

- Pride
<http://www.ebi.ac.uk/pride/>
- gpmDB
<http://www.thegpm.org/GPMDB/index.html>
- Open Proteomics Database
<http://bioinformatics.icmb.utexas.edu/OPD/>
- PeptideAtlas
<http://www.peptideatlas.org/>

A few prototype repositories for proteomics data are starting to get off the ground. gpmDB is interesting because it makes it easy to compare your own search results with the accumulated results of other people's searches. Peptide Atlas addresses the issue of presenting identifications in an integrated, hierarchical fashion.



Figure 6
 Example of peptides confirming a case of alternative splicing of the lamin A/C gene (LMNA). PAp00038023 was identified as part of protein ENSP00000310687 from the SiHa human cell line experiment. PAp00042742 was identified as part of protein ENSP00000292304 from a human B-cell experiment.

This enables the peptide matches to be displayed against the genomic DNA sequence, annotated with genes and coding sequences. Proteins from a variety of databases, plus translations from mRNAs and EST's are aligned so as to provide a comprehensive picture of the context for each peptide match.

Future directions

Standards

➤ Pergola, P. G., *et al.*, *A common open representation of mass spectrometry data and its application to proteomics research*, *Nat Biotechnol.* 22 1459-66 (2004)

- Proteomics Standards Initiative

mzData, mzIdent

<http://psidev.sourceforge.net/ms/>

- Institute for Systems Biology

pepXML, mzXML

http://sashimi.sourceforge.net/software_tpp.html

Guidelines

➤ Carr, S., *et. al.*, *The need for guidelines in publication of peptide and protein identification data*, *Molecular & Cellular Proteomics* 2004, 3, 531-3.

Finally, there are strong moves towards standards and guidelines for the exchange and publication of proteomics data, the majority of which relates to database searching. The Proteomics Standards Initiative and the Institute for Systems Biology are both very active in this area. All those involved are well aware that these efforts can only succeed by being inclusive, and achieving a consensus among all interested parties.

Selected Reviews

- Aebersold, R. and Mann, M., *Mass spectrometry-based proteomics*, Nature 422 198-207 (2003)
- Ashcroft, A. E., *Protein and peptide identification: the role of mass spectrometry in proteomics*, Natural Product Reports 20 202-215 (2003)
- Baldwin, M. A., *Protein identification by mass spectrometry - Issues to be considered*, Mol. & Cellular Proteomics 3 1-9 (2004)
- Henzel, W. J., *et al.*, *Protein identification: The origins of peptide mass fingerprinting*, Journal of the American Society For Mass Spectrometry 14 931-942 (2003)
- Johnson, R. S., *et al.*, *Informatics for protein identification by mass spectrometry*, Methods 35 223-36 (2005)
- Mann, M., *et al.*, *Analysis of proteins and proteomes by mass spectrometry*, Ann. Rev. of Biochem. 70 437-473 (2001)
- Yates, J. R., 3rd, *Database searching using mass spectrometry data*, Electrophoresis 19 893-900 (1998)

<http://tinyurl.com/e4brg>

A presentation such as this has to be brief and superficial. I'll finish by listing a few relevant reviews, which make a good starting point for anyone wishing to explore the field in greater depth.