Introduction to Database Searching using MASCOT

MASCOT

MATRIX SCIENCE

This presentation introduces the topics we will discuss in depth in subsequent talks. It assumes a working knowledge of protein chemistry and mass spectrometry, but no experience of database searching.

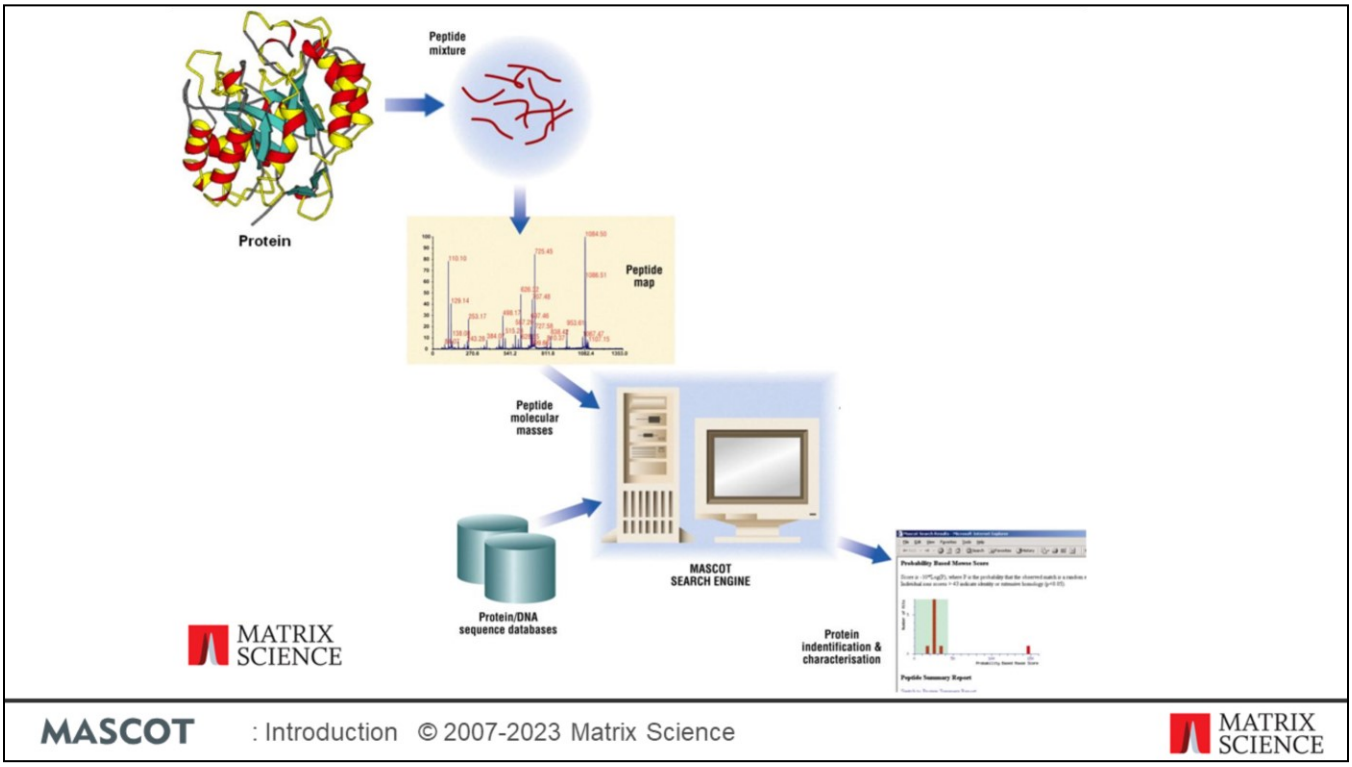
Three ways to use mass spectrometry data for protein identification

1. Peptide Mass Fingerprint

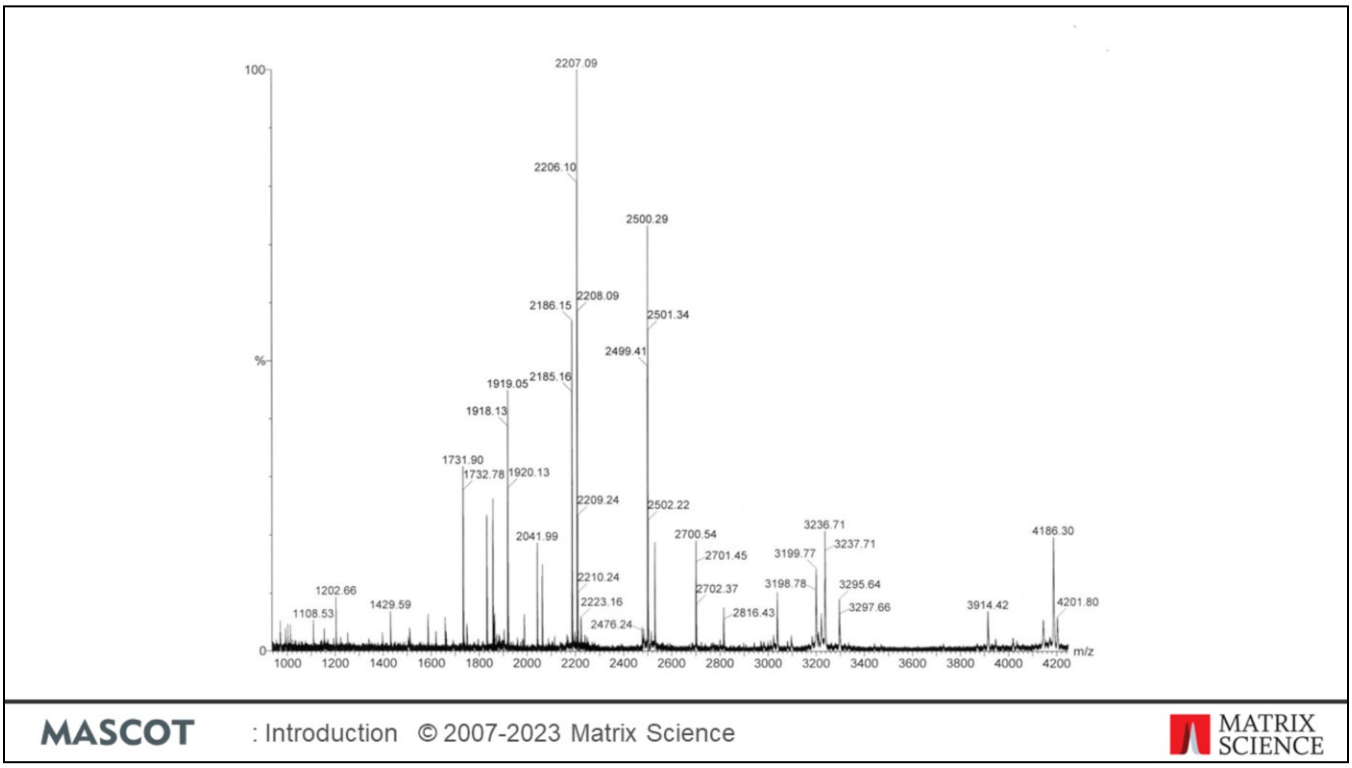
A set of peptide molecular masses from an enzyme digest of a protein

 © 2007-2023 Matrix Science

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 masses 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, so the starting point will often be a spot off a 2D gel. The protein is digested with an enzyme of high specificity; usually trypsin, because this is reliable and inexpensive and produces peptides of a suitable size, 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.

We don't submit the raw data to the search engine. First of all, the spectrum must be reduced to a peak list: a set of mass and intensity pairs, one for each peak. We call this procedure peak detection or peak picking.

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.

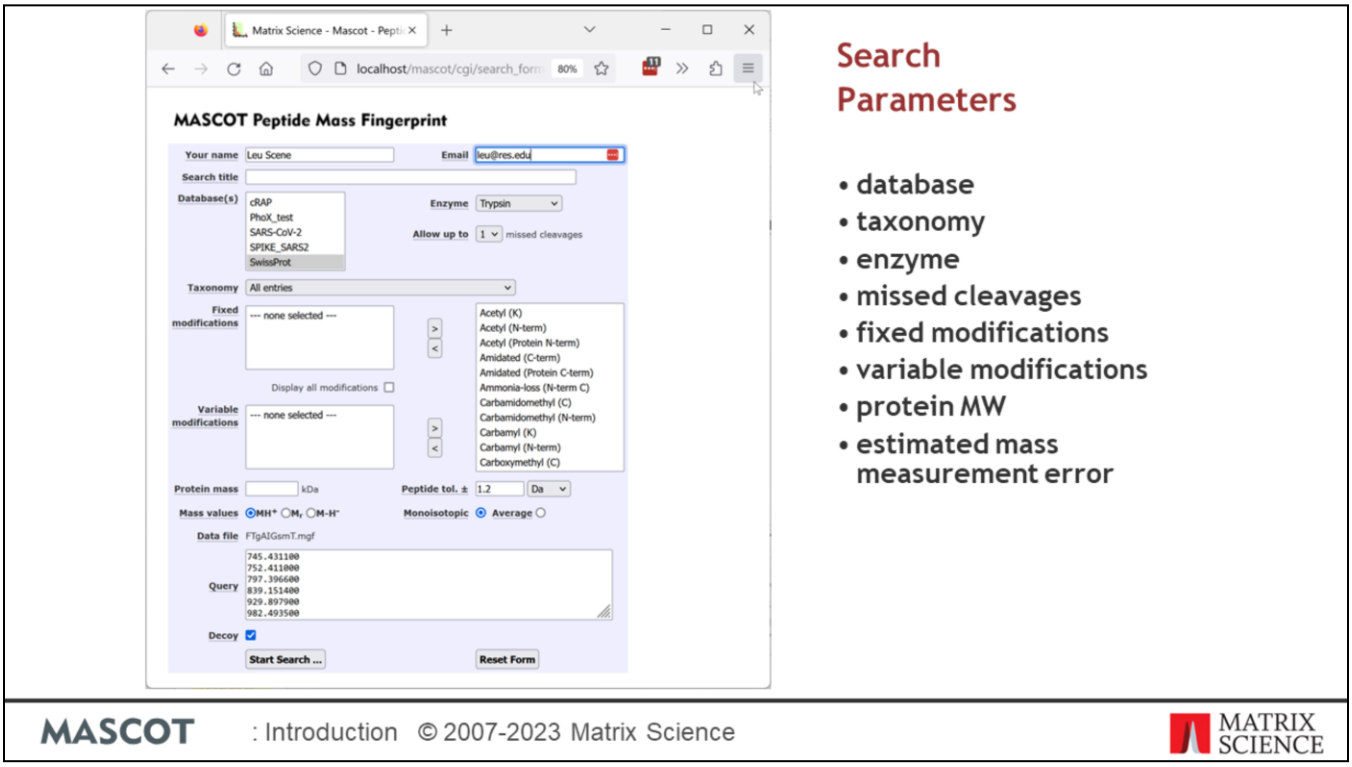
Mass accuracy is important, but so is coverage. Better to have a large number of mass values with moderate accuracy than one or two mass values with very high accuracy.

PMF Servers on the Web

Mascot: https://www.matrixscience.com/search_form_select.html
MS-Fit (Protein Prospector): https://prospector.ucsf.edu/prospector/mshome.htm
PeptideMass (Expasy): https://web.expasy.org/peptide_mass/
SpectrumMill: https://proteomics.broadinstitute.org/millhtml/msfit.htm

Mowse, PeptideSearch, Profound (Prowl), Protocall, Aldente, Xproteo, Bupid, MassSearch

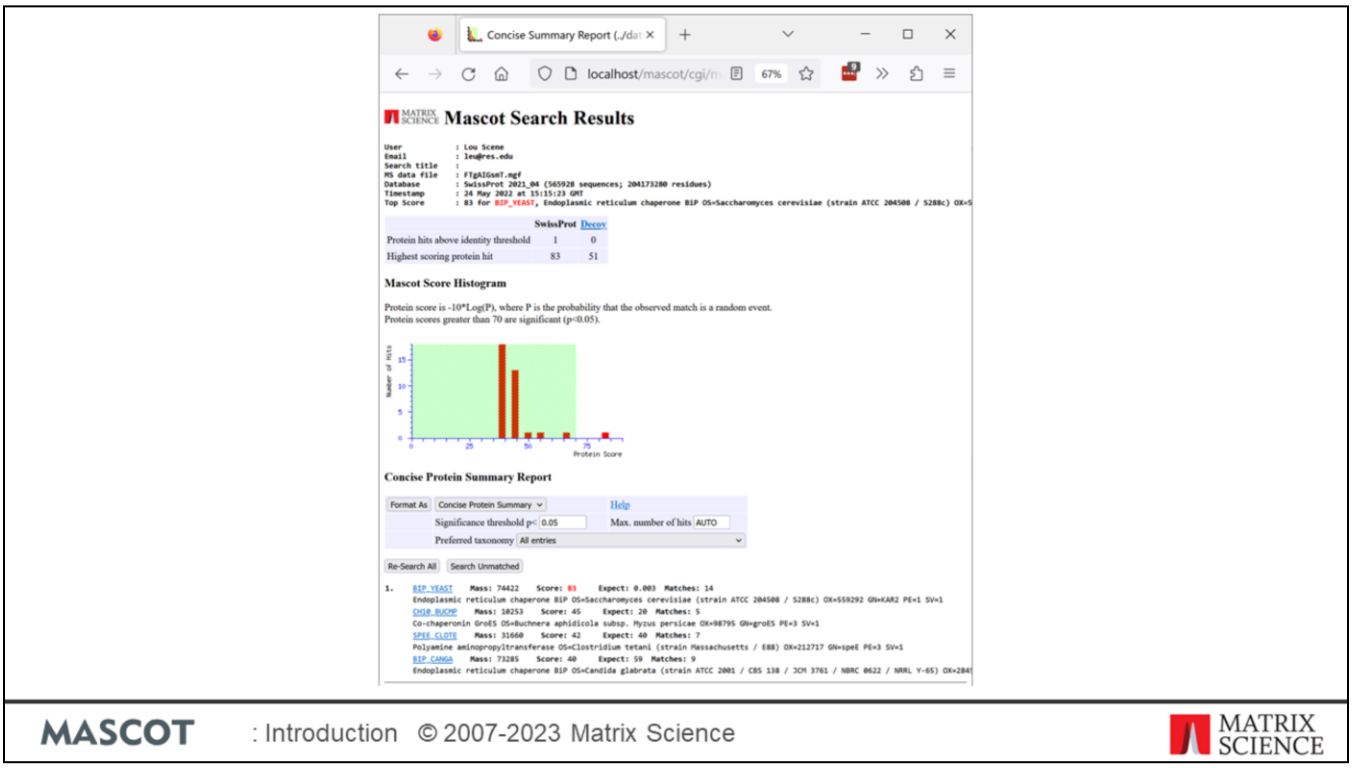
These presentations will focus on Mascot, but you should be aware that there are a few other PMF search engines on the web. There are also software packages available for download to run locally or sold as commercial products. Some of the early search engines, such as Mowse and PeptideSearch, are no longer available.



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 a later presentation.

In theory, you could design a search engine that didn't require search parameters, and tried to work everything out from the mass values, but this would be very inefficient. If you know the enzyme was trypsin, much easier to supply this information as part of the search.

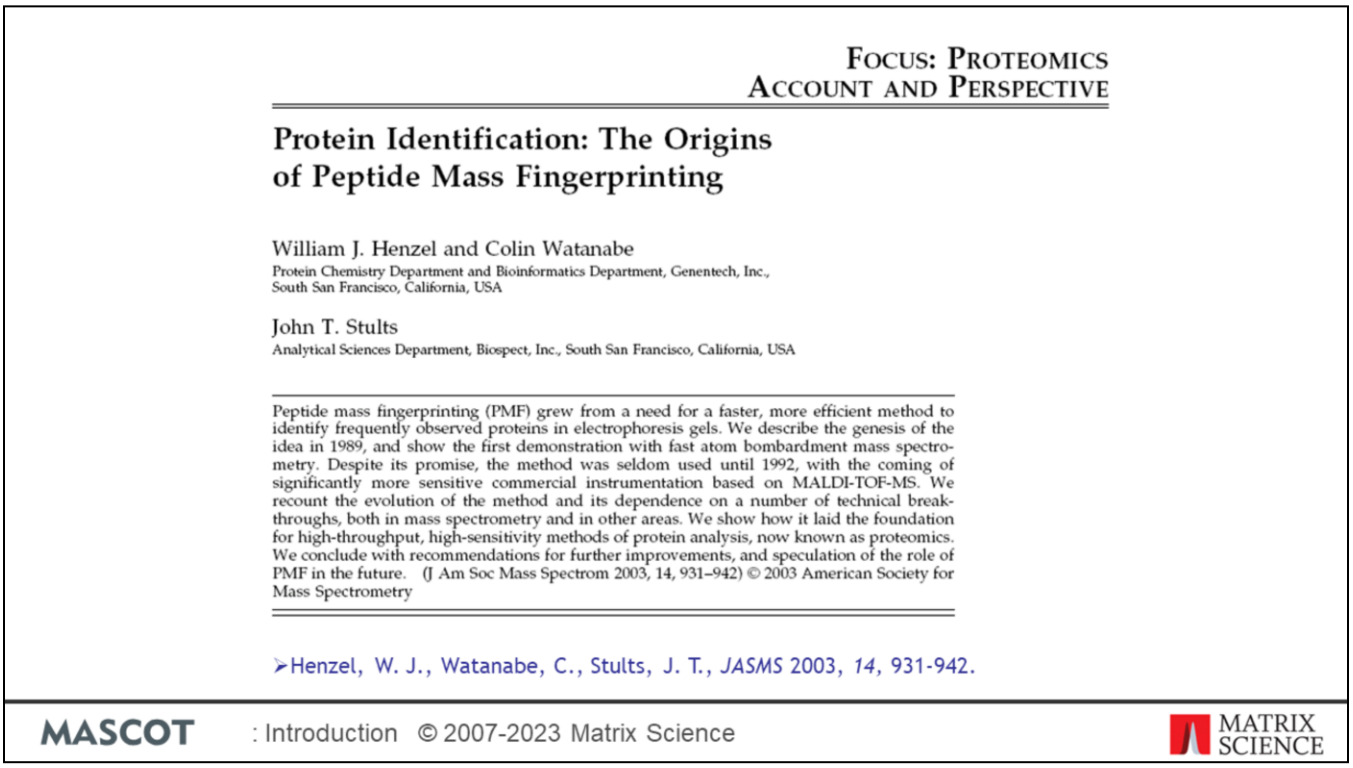
To perform a search, you paste your peak list into the search form, or upload it as a file, enter values for the search parameters, and press the submit button.



A short while later, you 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. One of the main problems in the early days of the technique was how to tell whether the top match was “real”, or just the match at the top of the list … 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: PROTEOMICS
ACCOUNT AND PERSPECTIVE

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., *JASMS* 2003, *14*, 931-942.

If you want to learn more about the origins and development of peptide mass fingerprinting, I can 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 principle, it is a sensitive technique because you don't need 100% coverage. 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.

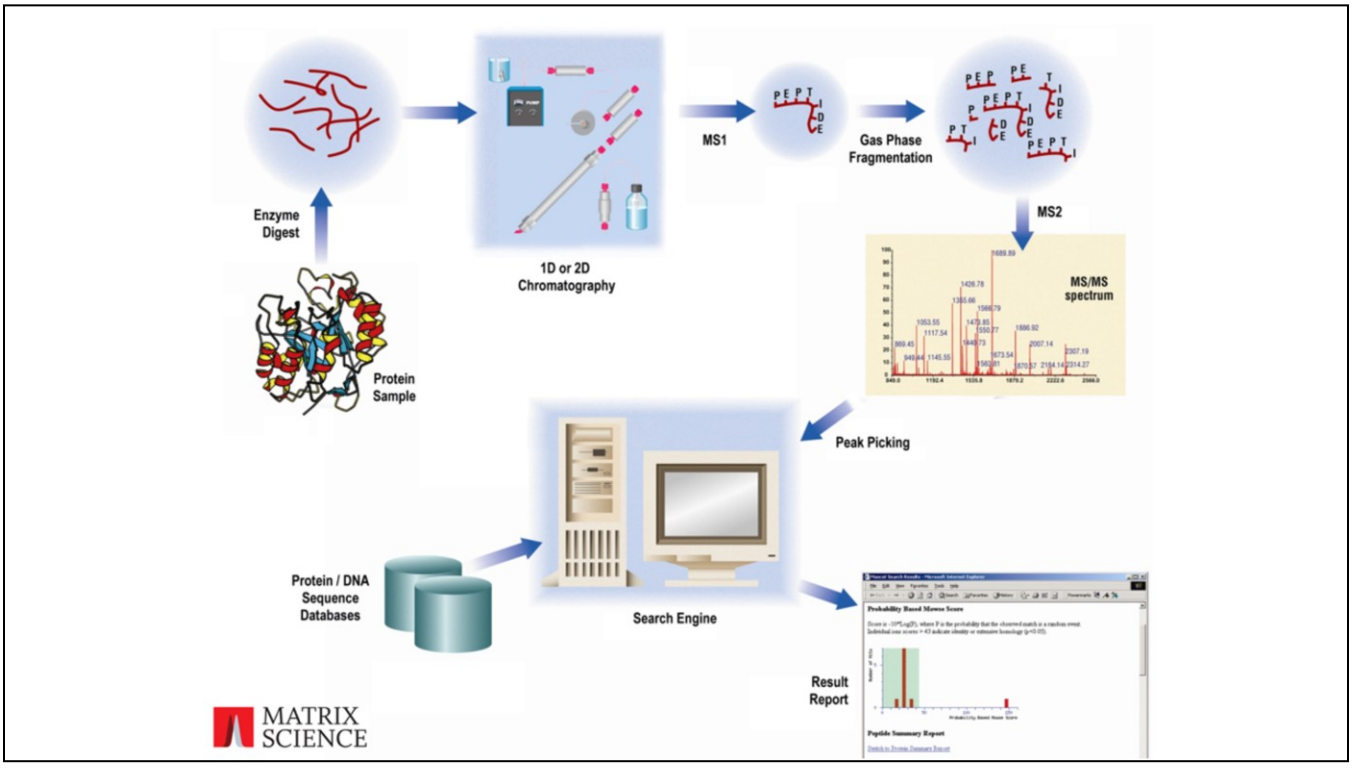
One of the limitations is that you need a database of proteins or nucleic acid sequences that are equivalent to proteins, e.g. mRNAs. In most cases, you will not get satisfactory results from an EST database, where most of the entries correspond to protein fragments, or genomic DNA, where there is a 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 something very similar, is not in the database, the method will fail. If you are studying a well characterised organism, such as human or mouse or yeast, this is unlikely to be a problem. If you are studying a virus or plant with an unsequenced genome, it can be a major problem, and you depend on getting matches to homologous proteins from related organisms.

The most important limitation concerns mixtures. If the data quality is good, then it may be possible to identify a two component mixture, where both components are at a similar level,

and on very rare occasions three. 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.

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.

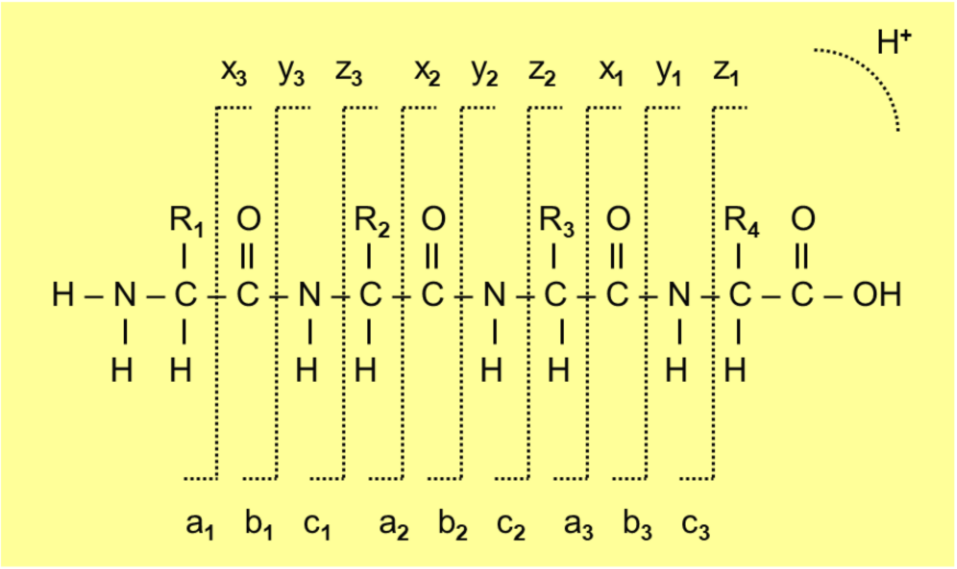
Again, we start with protein, which can now be a single protein or a complex mixture of proteins. We use an enzyme such as trypsin to digest the proteins to peptides. If it is a complex mixture, such as a whole cell lysate, we will probably use one or more stages of chromatography to regulate the flow 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. As in the case of a peptide mass fingerprint, each entry in the database is digested, in silico, and the masses of the expected peptides calculated. If a calculated peptide mass matches the experimental one, the mass values expected to result from the gas phase fragmentation of the peptide are calculated and the degree of matching to the peaks in the MS/MS spectrum scored.

Unlike a peptide mass fingerprint, use of a specific enzyme is not essential. By looking at

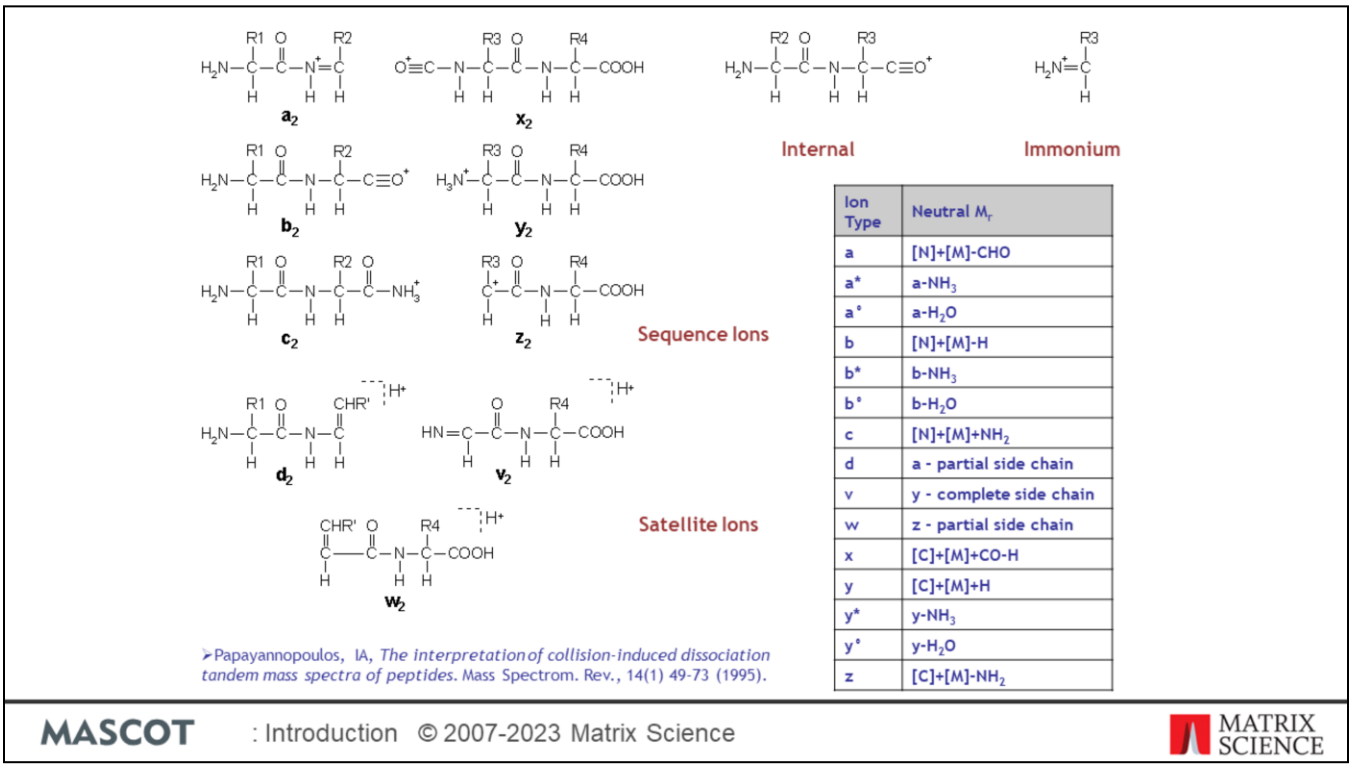
all possible sub-sequences of each entry that fit the precursor mass, it is possible to match peptides when the enzyme specificity is unknown, such as endogenous peptides.



➢Roepstorff, P. and Fohlman, J. (1984). *Proposal for a common nomenclature for sequence ions in mass spectra of peptides.* Biomed Mass Spectrom **11**, 601.

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.



Ion Type	Neutral M_r
a	[N]+[M]-CHO
a*	a-NH_3
a°	a-H_2O
b	[N]+[M]-H
b*	b-NH_3
b°	b-H_2O
c	[N]+[M]+NH_2
d	a - partial side chain
v	y - complete side chain
w	z - partial side chain
x	[C]+[M]+CO-H
y	[C]+[M]+H
y*	y-NH_3
y°	y-H_2O
z	[C]+[M]-NH_2

If peptides fragmented cleanly and uniformly along the backbone, we wouldn't need database search. We would see a ladder of peaks for each ion series, where the distance from one peak to the next was the mass of an amino acid residue, allowing the sequence to be read off the spectrum. In real life, fragmentation is rarely perfect, and the spectrum will usually show significant peaks from side chain cleavages and internal fragments, where the backbone has been cleaved twice. More importantly, the backbone may fail to cleave at certain locations, so that the MS/MS spectrum has no evidence for some of the residues.

This slide shows the most common fragment ion structures and the table is a "ready reckoner" that can be used to calculate the 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

This review by Ioannis Papayannopoulos is a good introduction to the fragmentation chemistry of peptide ions in the gas phase

---

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.

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.

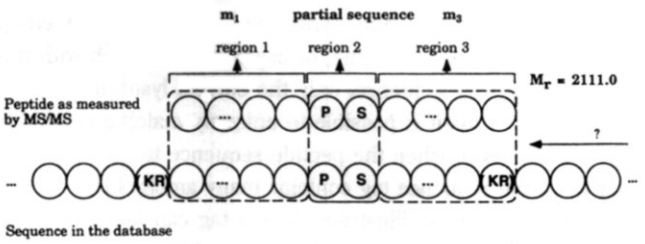


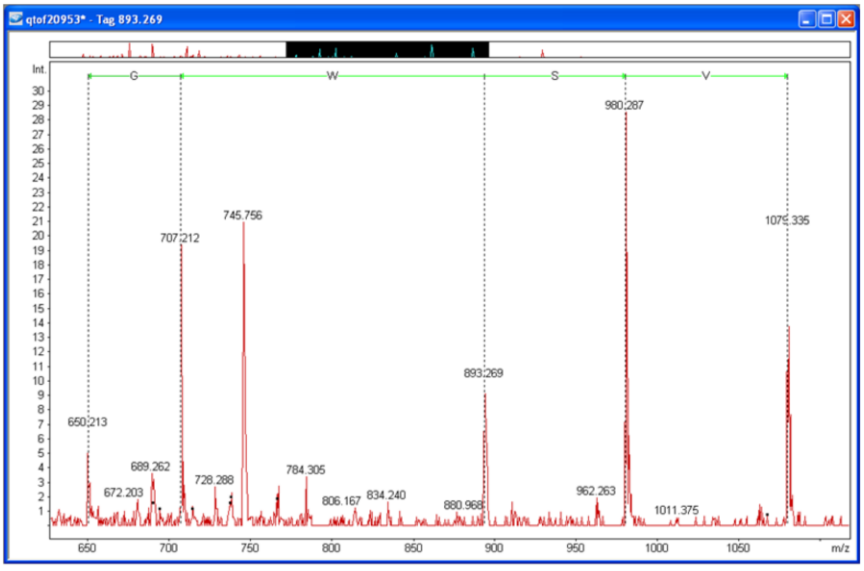
Figure 1. Principle of matching peptide sequence tags to a proposed sequence. The upper chain of amino acids represents the peptide sequence as measured by MS/MS (from Table 1 in this example), and the lower chain represents amino acids in the sequence database that the tag is compared to. Note that the partial sequence divides the peptide into three regions. The added mass m_1 of the residues in region 1, together with the N-terminus, is a match criterion as is the added mass in region three, m_3. In region 2, the sequence is known. Furthermore, it can be required that the peptide obey the cleavage condition of the proteolytic enzyme, marked by KR for trypsin. The left pointing arrow indicates that both search directions may have to be considered.

➢Mann, M. and Wilm, M., *Error-tolerant identification of peptides in sequence databases by peptide sequence tags*. Anal. Chem. 66 4390-9 (1994).

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.



1490.4 tag(650.2,GWSV,1079.3)

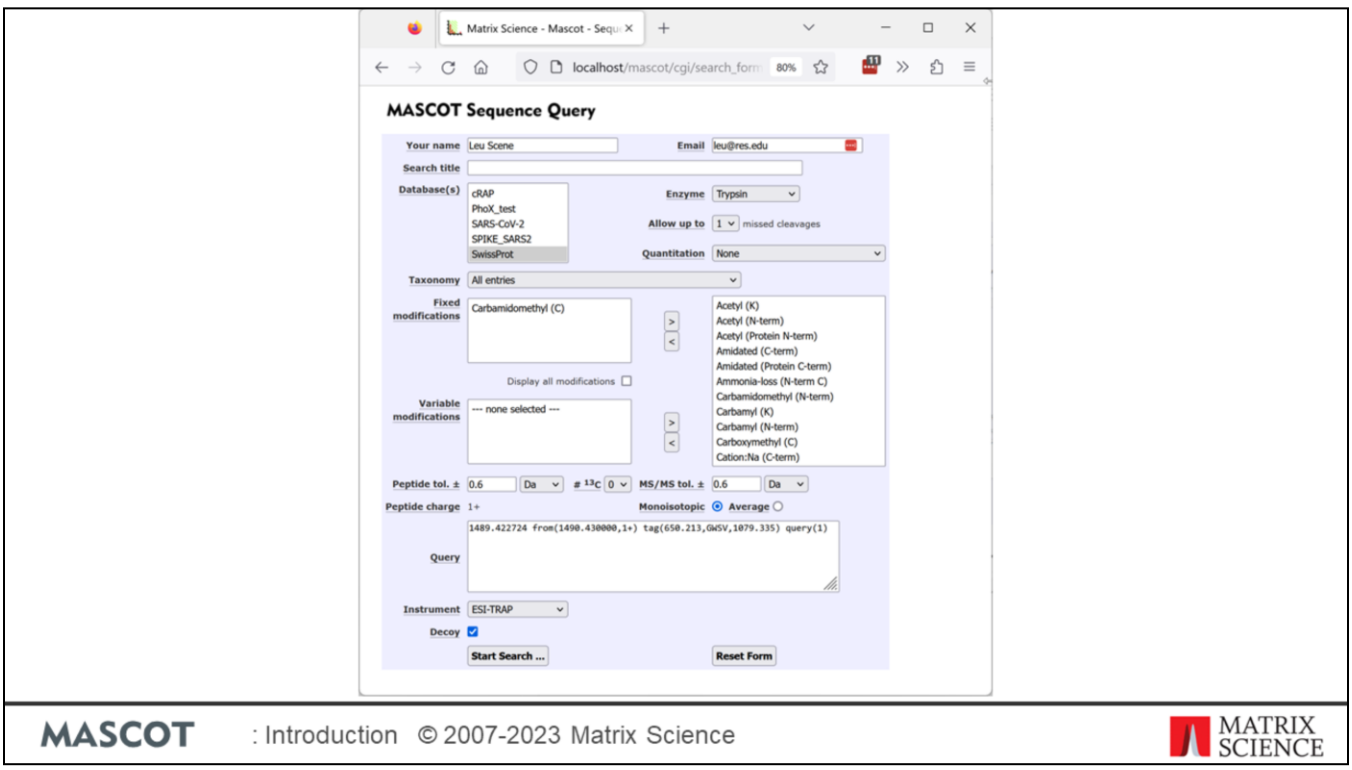
Picking out a good tag is not trivial, and requires both luck and experience. In this spectrum, we can see a promising four residue tag. The syntax used by Mascot for a sequence tag is shown below the spectrum. We'll discuss this format in greater detail in the Sequence Query presentation

Sequence Query Servers on the Web

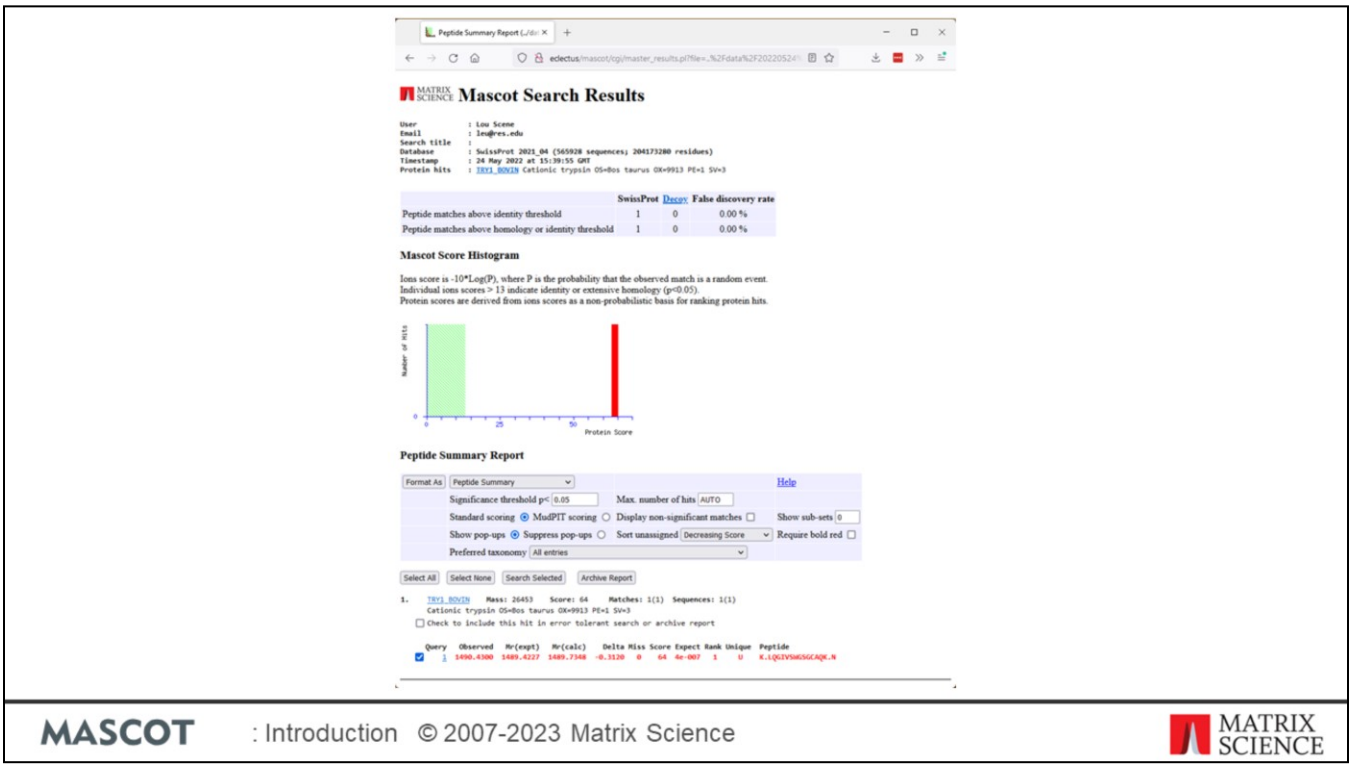
Mascot: https://www.matrixscience.com/search_form_select.html
MS-Seq (Protein Prospector): https://prospector.ucsf.edu/prospector/mshome.htm

PeptideSearch, MultiIdent

There are a few software packages for sequence query searches. As with PMF, I have limited my list to servers that are publicly available on the web. Not such a wide choice as for PMF.



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 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.

However, the standard sequence tag is essentially obsolete. It is easier and more reliable to skip the interpretation step and pass the peak list to the search engine. The reason the sequence tag is still important is because it can be used in an “error tolerant” mode. This consists of relaxing the specificity, by removing the peptide molecular mass constraint. 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. In Mascot, 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. Even so, it is very difficult to identify all possible ambiguities, especially when we allow for missing peaks.

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

 © 2007-2023 Matrix Science

Which brings us to the third category: Searching the uninterpreted MS/MS data from a single peptide or from a complete LC-MS/MS run. That is, using software to match the peak list, without any manual sequence calling.

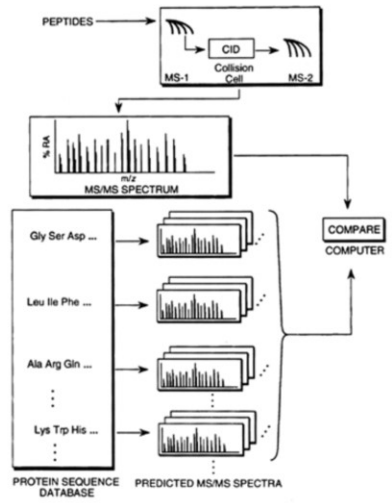


Figure 1. Flow chart that depicts the alogrithm for searching protein databases with tandem mass spectrometry data.

SEQUEST

- Eng, J. K., McCormack, A. L. and Yates, J. R., 3rd., *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.* J. Am. Soc. Mass Spectrom. 5 976-89 (1994)

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.

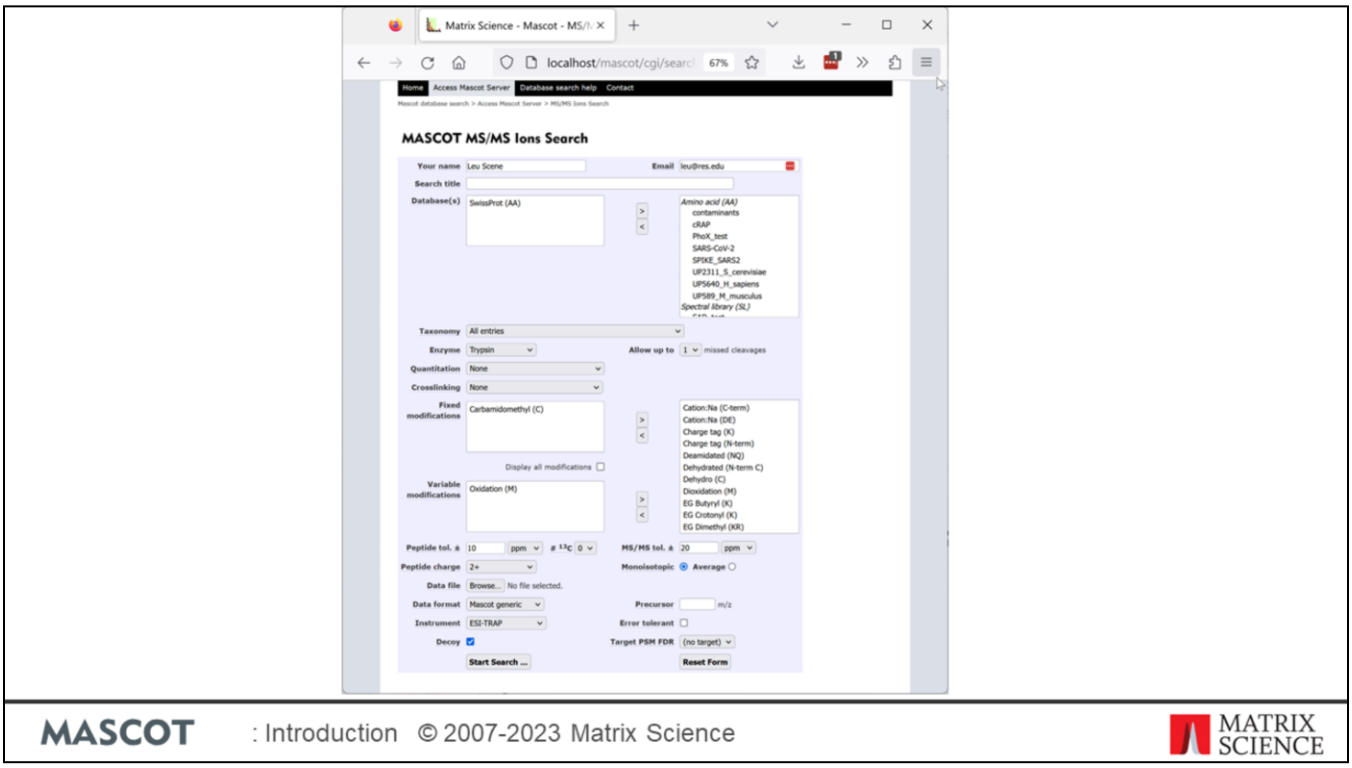
MS/MS Ions Search Servers on the Web

Inspect / MS-GFDB	https://proteomics.ucsd.edu/ProteoSAFe/
Mascot	https://www.matrixscience.com/search_form_select.html
MS-Tag (Protein Prospector)	https://prospector.ucsf.edu/prospector/mshome.htm
PepFrag (Prowl)	https://prowl.rockefeller.edu/prowl/pepfrag.html
RAId_DbS	https://www.ncbi.nlm.nih.gov/CBBResearch/qmbp/RAId_DbS/index.html
Not on-line	Andromeda, Byonic, Comet, greylag, Morpheus, Myrimatch, MSFragger, OMSSA, Paragon, Peaks DB, PepSplice, pFind, Phenyx, ProbID, ProLuCID, ProteinLynx GS, Sequest, SIMS, SpectrumMill, Tide, X!Tandem (The GPM), etc. etc.

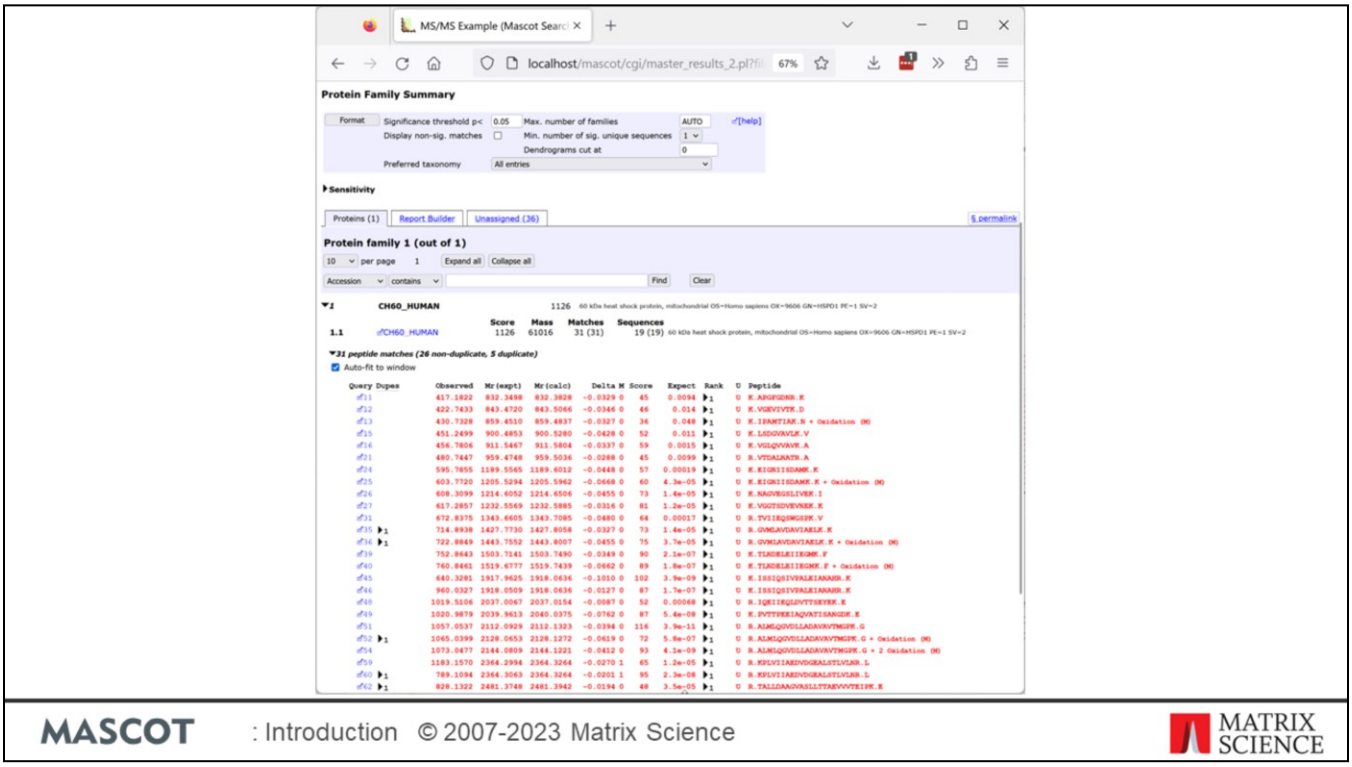
MASCOT : Introduction © 2007-2023 Matrix Science MATRIX SCIENCE

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.

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

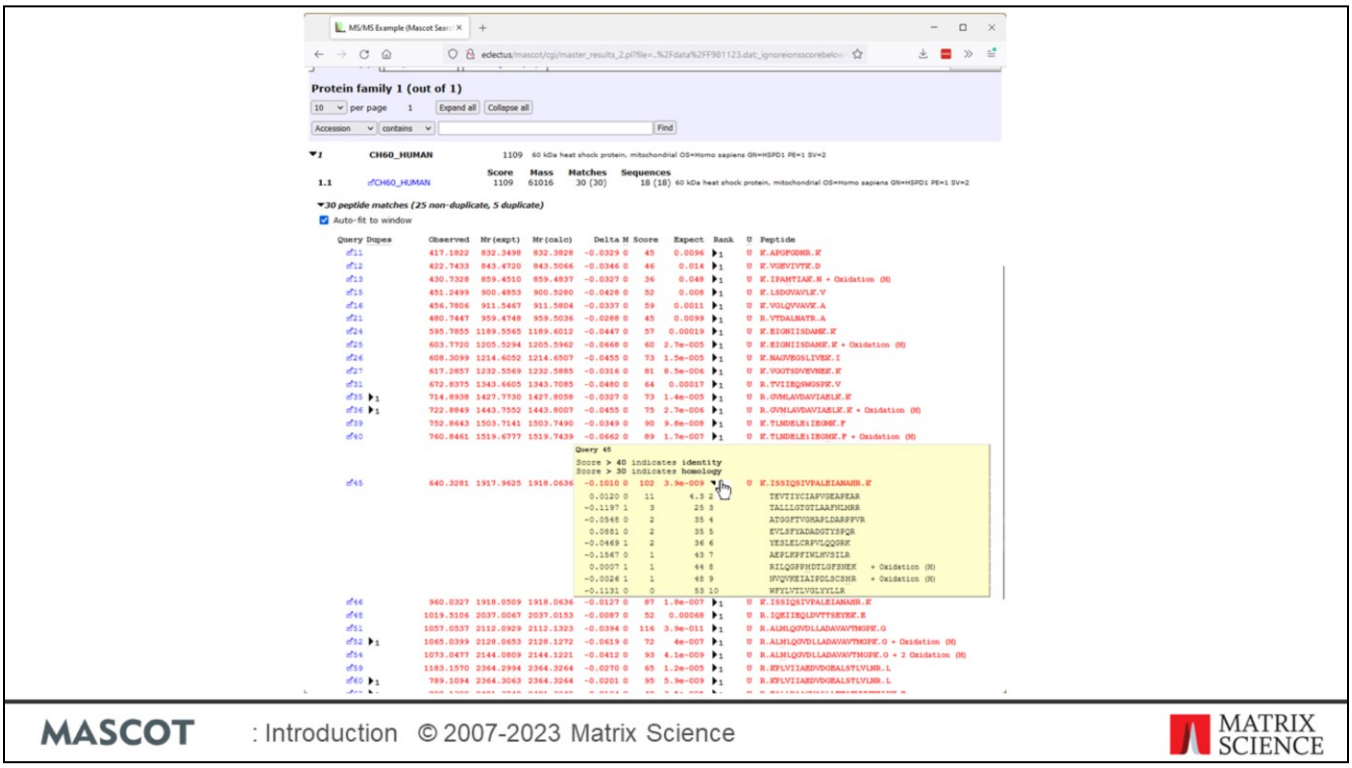


This is the Mascot search form for an MS/MS search. Fairly similar the previous two and, as before, you must also specify the database, mass accuracy, modifications to be considered, etc.



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.

We match peptide sequences to individual MS/MS spectra, then try to assign these peptide sequences to proteins. Usually, there is ambiguity, and we aim to report a minimal list of proteins. That is, the shortest list of proteins, grouped into families by shared peptides, that can account for all the observed peptide matches. So, the report lists a series of protein families and, for each protein family, the peptide matches that have been assigned. But, there is an additional dimension to the data.



For each spectrum, there may be multiple possible peptide matches. This particular Mascot report uses a pop-up window to show the alternative peptide matches to each spectrum. In this case, the top match has a high score and the other matches are low scoring, random matches, so no ambiguity. In other cases, the top two or three matches may all be interesting, such as a phosphopeptide where there are several potential phosphorylation sites and moving the phosphate from one site to another only changes the score slightly.

MS/MS Ions Search

Easily automated for high throughput

Can get matches from marginal data

Can be slow if:

- No enzyme
- Many variable modifications
- Large database
- Large dataset

MS/MS is peptide identification

- Proteins by inference.

To summarise, 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. Imagine a weak or noisy spectrum that gets a match with a poor score. In isolation, this might be insufficient evidence for the presence of a protein. But, if there are other, similar quality spectra with matches to the same peptide or to other peptides from the same protein, taken together and with the right safeguards, they can provide a degree of confidence that the protein has been identified.

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

Always remember that it is peptides that are being identified, not proteins. From the peptides that have been identified, we try to infer which proteins were present in the sample.

Another approach we can take with MS/MS peak lists is to search spectral libraries.

Spectral Library Search

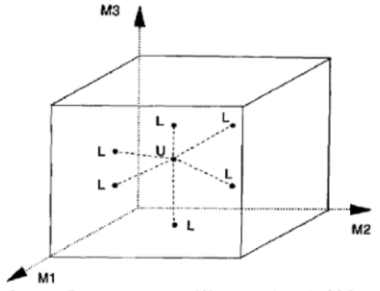


Figure 2. Point representation of library search results (L) for a hypothetical three-peak unknown (U) spectrum (masses M1, M2, and M3)

Hertz, H. S, Hites, R. A., and Biemann, K (1971)
Identification of mass spectra by computer-searching a file of known spectra
Anal. Chem. **43**, 6, 681–691

Stein, S. and Scott, D. R. (1994).
Optimization and testing of mass spectral library search algorithms for compound identification, J. Am. Soc. Mass Spectrom., **5**, 859-66

 © 2007-2023 Matrix Science

Spectral library searching of small molecules has been around for a long time as this Biemann paper from 1971 attests, but only more recently used for proteomics and peptide identification.

The earliest one was NIST MS Search, introduced in 1994, which also provided mass spectral libraries to search.

MS/MS Spectral Library Search Servers on the Web

SpectraST Spectrum Library Search	https://www.peptideatlas.org/spectrast/
Mascot	https://www.matrixscience.com/search_form_select.html
Protein Prospector Batch-Tag	https://prospector.ucsf.edu/prospector/mshome.htm

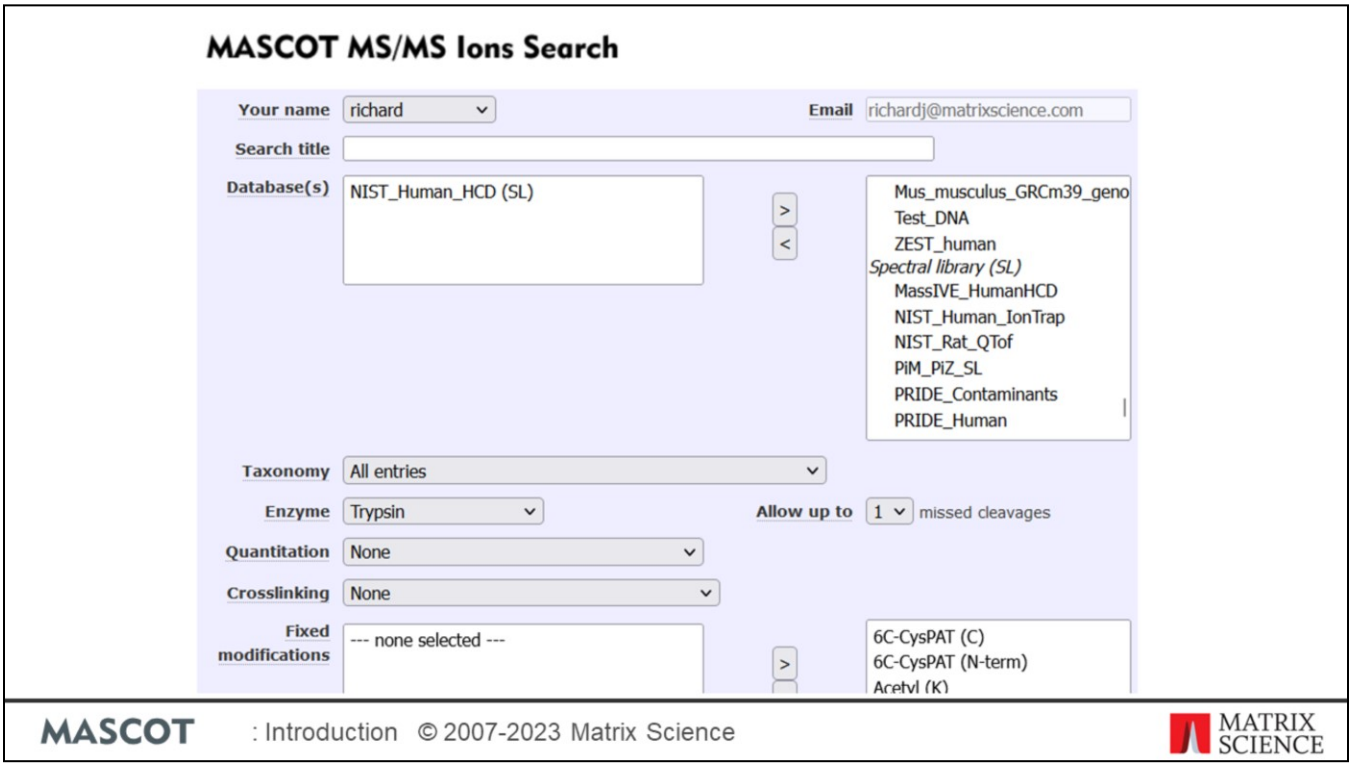
Not on-line: **BiblioSpec, COSS, Epsilon-Q, Mistle, MS PepSearch, NIST MS Search, PEAKS Studio, Pepitome, PeptideAtlas, Pmatch, Progenesis, Scaffold, Skyline, SpectraST, etc**

Initially, spectral library searching was not as popular as other methods because the peptide libraries were incomplete. Around 2006 things started to change with mass spec data becoming available on a large scale. Multiple spectral library search engine papers were published that year.

Most spectral library search engines are only available to install and run locally. Of the two originally on the web, SpectraST and X!Hunter, only SpectraST is still available.

Mascot integrates NIST MSPepSearch, which is available for use on our public website.

There are number of other SL search engines that are available for download and local searching.

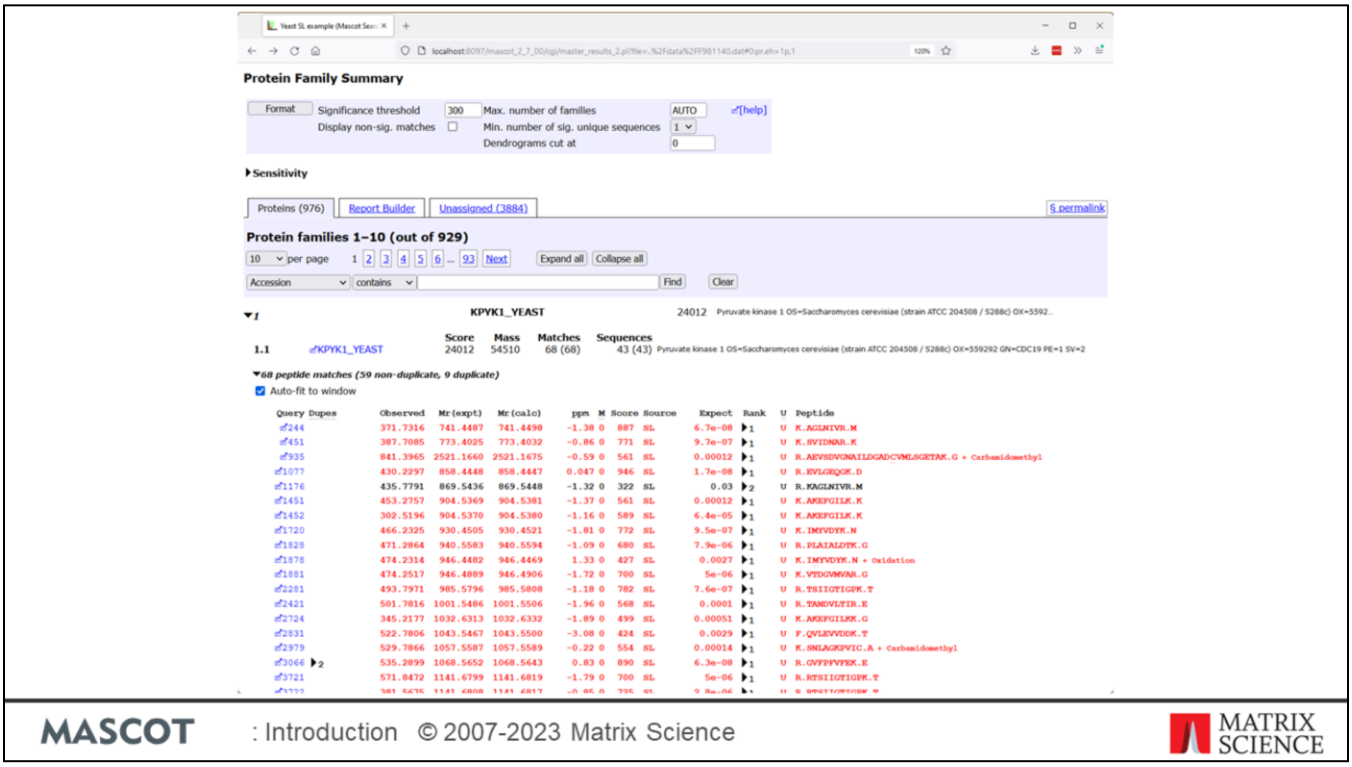


The input and output of a spectral library search uses the same input form and report format as MS/MS Ions Search. The Mascot search engine and NIST MSPepSearch are integrated. Combined searches can be run by selecting a spectral library database and a FASTA database in the search form.

Spectral libraries are in section separate from sequence databases.

Most search parameters – modifications, enzyme, missed cleavages, taxonomy, and instrument – simply don't apply to a library search, although they are used for the database search portion.

The precursor and fragment ion tolerances apply to both.

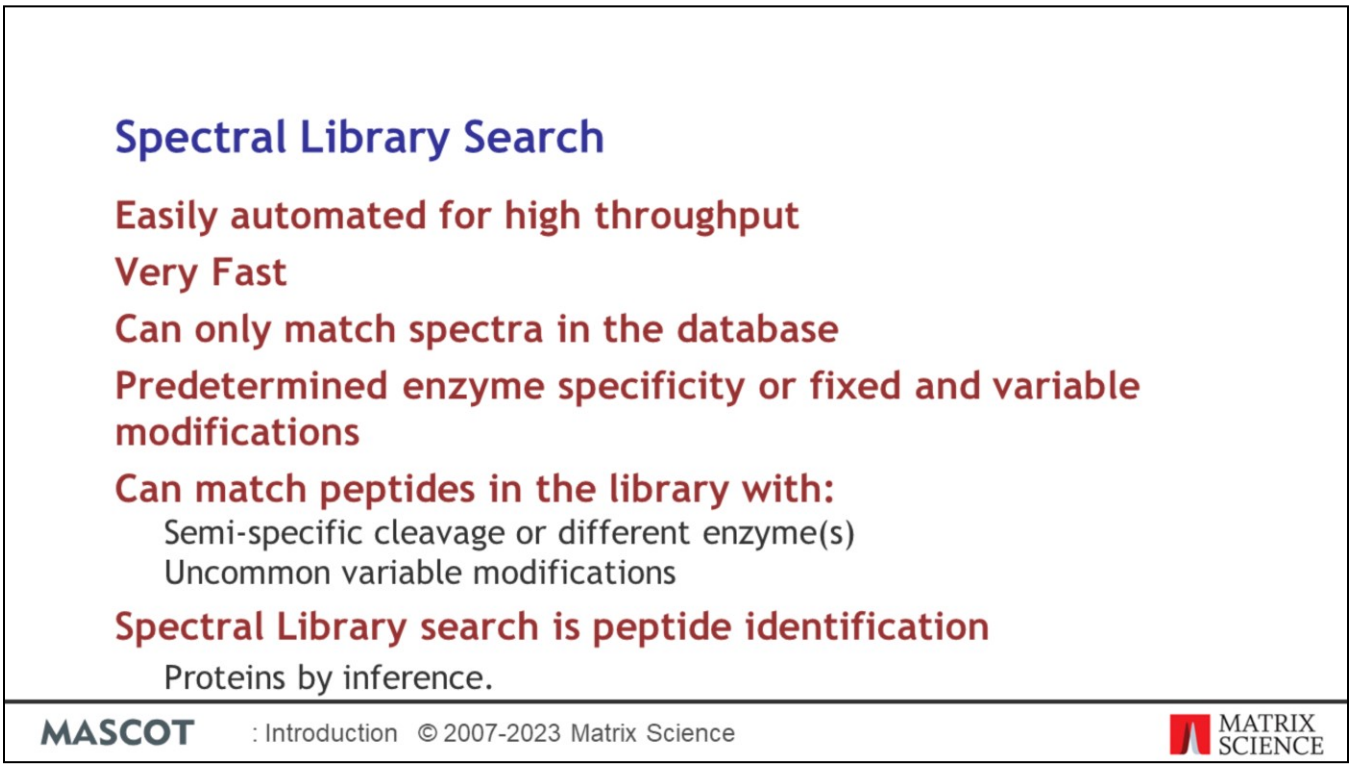


Results are presented in the same format as an MS/MS search with a list of inferred proteins and the peptide matches.

Spectral libraries typically contain only peptide-level information. Mascot maps spectral library matches to a reference sequence database, which enables robust protein inference. Each library entry is assigned to all the accessions from the reference database that contain the peptide sequence, ignoring enzyme specificity. This means that, in favourable cases, protein inference will be just as good as if the matches had been found in a search of the reference sequence database.

When an entry has no reference accession, we use the protein metadata from the spectral library, if present. When the reference file is well chosen, as in the example, few or no spectral library accessions will be visible in the summary report.

As in MS/MS reports there is an additional dimension to the data. For each spectrum, there may be multiple possible peptide matches which are displayed in the pop-up window.



Just like MS/MS searches, the process is easily automated for high throughput. This is one of the fastest ways to search uninterpreted MS/MS data. One of the reasons why it is so fast is that the search only considers spectra in the library, which is a smaller search space than traditional database searching.

Most search parameters – modifications, enzyme, missed cleavages, taxonomy, and instrument – simply don't apply to a library search. These are predetermined when the library is created. However, if the library contains peptides that are outside the database search space, such as different enzyme or uncommon variable modifications, then they will automatically be included in the search.

As in MS/MS searching, the spectral library search is identifying peptides not proteins. Unlike an MS/MS search against proteins from a FASTA file, the spectral library contains limited protein information beyond an accession number and the peptide sequence. Instead, a reference library is selected when the library is set up and protein data is taken from there.

	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, I'd like to compare the fundamental characteristics of database searching using MS data versus MS/MS data. The MS/MS analysis covers both database searching and spectral library searching.

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 MS/MS searches are perceived to be more powerful is mainly that the data set often contains many spectra, 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. We tend not to notice this, because there are many reasons why spectra fail to get matches, such as unsuspected modifications or incorrect precursor mass or charge. We don't normally investigate these failures, so don't see that some of these are due to acquiring a mixed MS/MS spectrum.

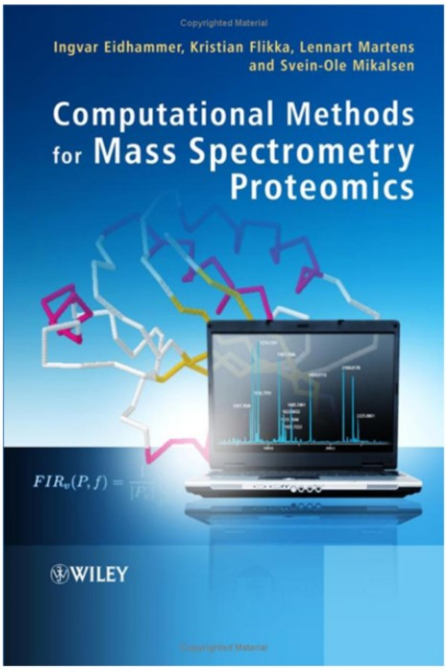
In the peptide mass fingerprint, the specificity comes from the predictable cleavage behaviour of the proteolytic enzyme. Thus, we want an enzyme with a low cutting frequency, such as trypsin. In the MS/MS ions search, the specificity comes from the mostly

predictable gas-phase fragmentation behaviour of peptide molecular ions.

In a PMF, we tend to be unsure of the protein mass. Even if the sample is a spot off a 2D gel, there is no guarantee that the database sequence corresponds to the fully processed protein. For MS/MS, we tend to be unsure which types of fragment are present. There is often a dependency on the peptide sequence or a modification. We might get b ions or y ions or a combination of b and y ions. There may be neutral losses and multiple charge states.

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 simply not possible with a PMF.

Further Reading



Finally, if you are looking for a recommendation for a text book, this one is fairly recent and covers the whole field clearly and systematically. It isn't just a loose collection of research papers.