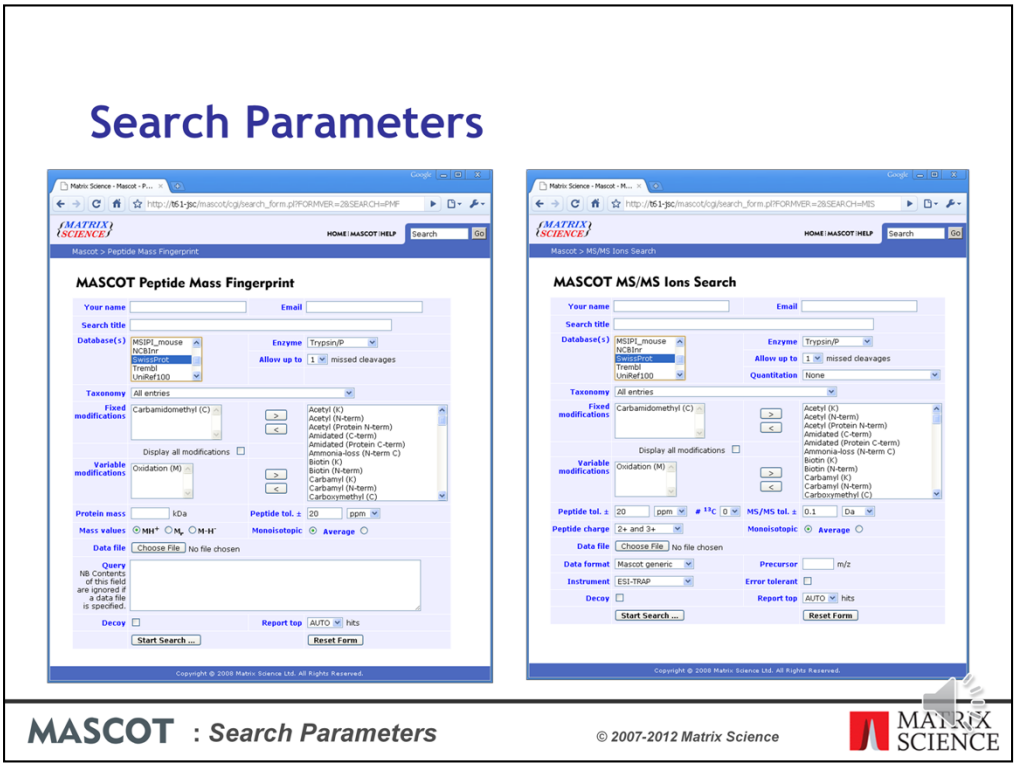
Search Parameters

MASCOT

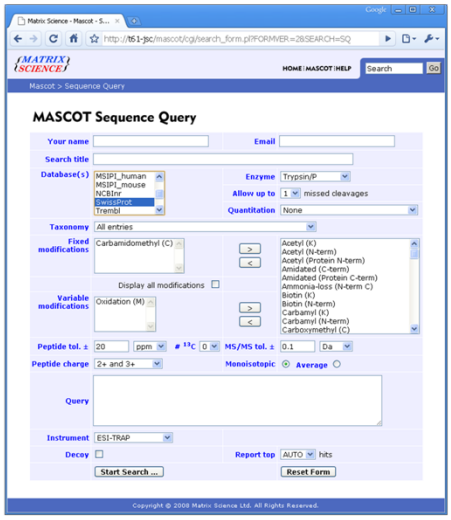
MATRIX SCIENCE



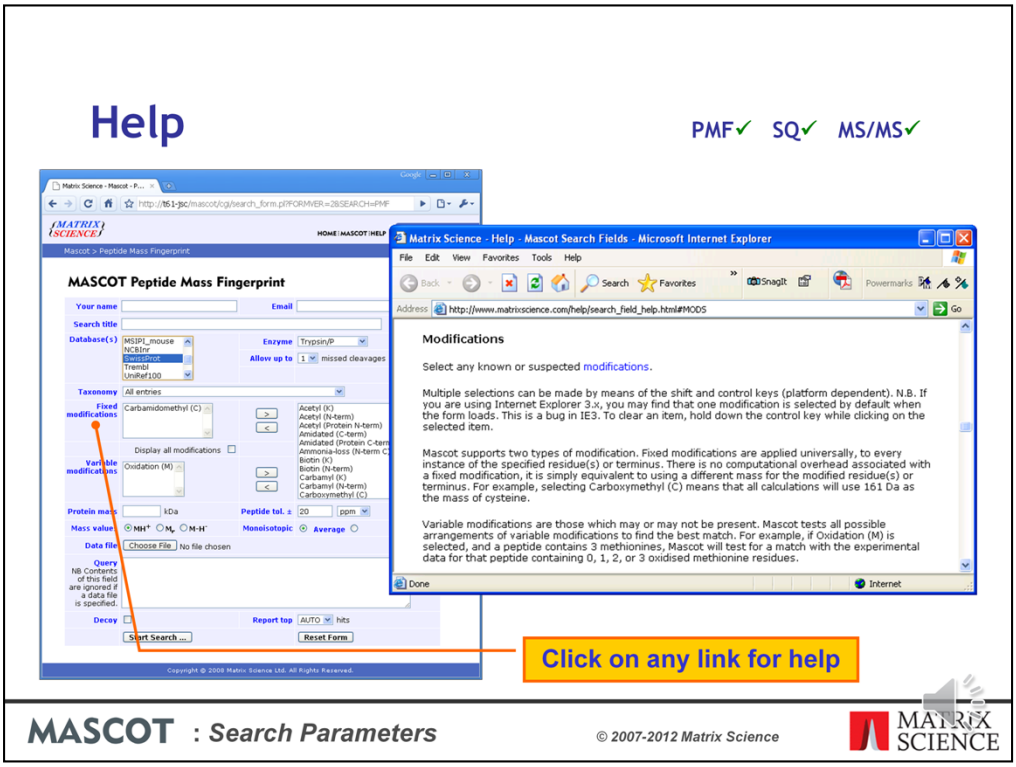
In this presentation, we will describe each of the Mascot search parameters.

If you submit a search from a web browser, you have a choice of three different search forms. All three forms submit to the same search engine, but they have been optimised for three different types of search. The form for a peptide mass fingerprint is shown on the left, and the form for a search of uninterpreted MS/MS data on the right. Most of the controls are common to both.

Search Parameters

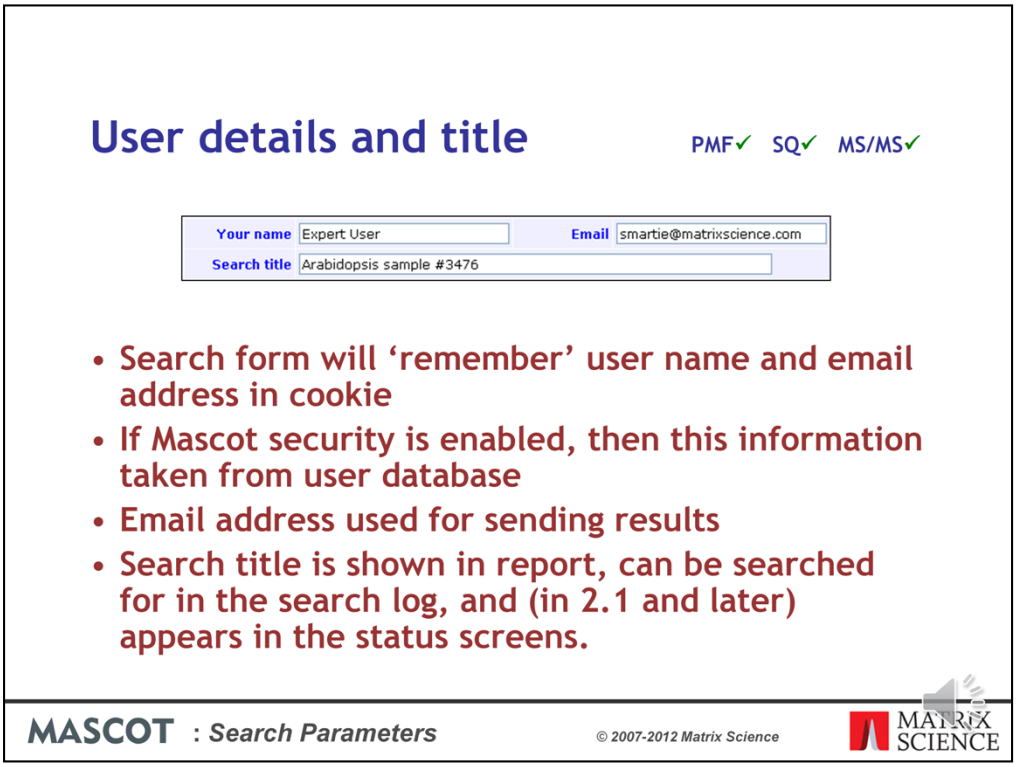


The third form is for a sequence query, such as a sequence tag search. The controls on this form are very similar to those on the MS/MS form. The main difference is that we have a text area to type in the queries, rather than a data file upload control.



At the top of each slide, there is a key to show which search parameter applies to which type of search.

The labels on the search form are hyperlinks. Just click on them to get detailed help

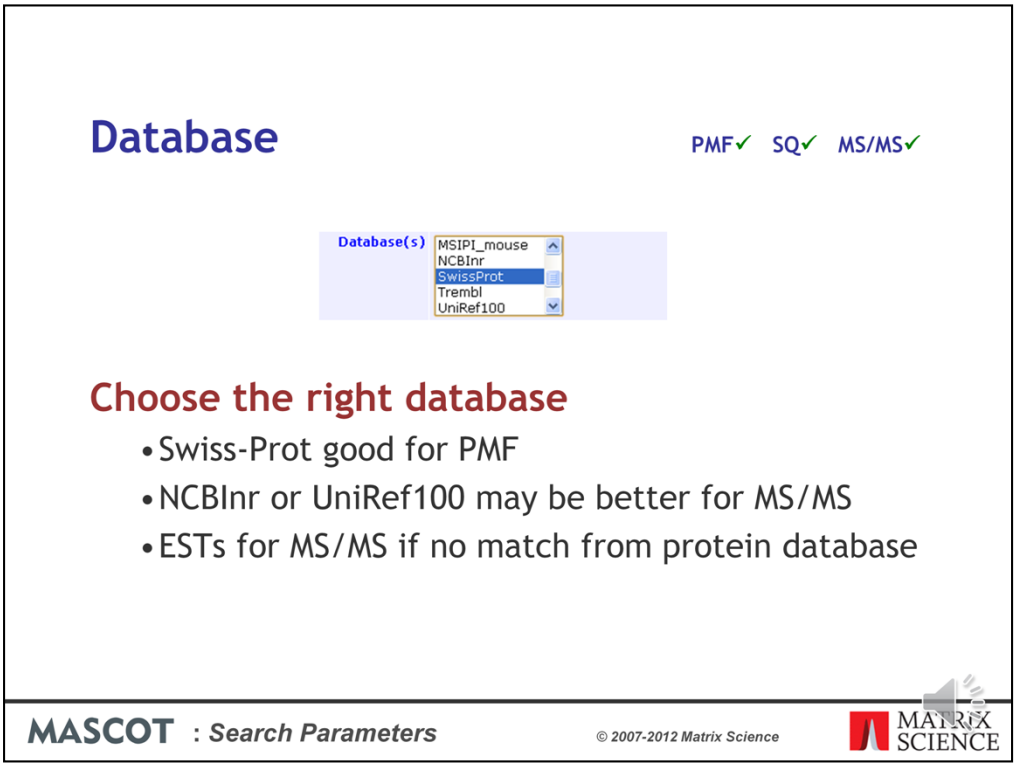


At the top of the form are a couple of fields for user information. The name and email are saved as a browser cookie when a search is submitted, so you don't need to complete them every time.

If you have an in-house server, and Mascot security is enabled, these fields will be populated automatically with the details of the user who is logged in

When you use the Matrix Science public web site, you have to supply a name and email address. This is to allow the results of a search to be returned by email. Usually, search results are returned promptly to your browser window. However, if your connection to the web site is broken before the search is complete, they will be emailed to the supplied address. If you have an in-house server, you can enable this if you wish. It is turned off by default

The search title is free text. You don't have to enter anything. However, it is a good idea to fill in all of these fields, because it makes it much easier to find your old search results in the search log.



Choosing the right database is so important that there will be a complete presentation on this topic.

Very briefly, for a peptide mass fingerprint, search a comprehensive, non-redundant database, like SwissProt. If the data are any good, it won't matter if one or two mass values fail to find matches. The advantage of searching a small database is that the search is fast and the reports are concise.

In MS/MS, the most interesting protein in the mixture might be at a very low level and only represented by a single spectrum. So, you don't want to miss a single peptide. You need a non-identical database, where every single peptide is explicitly represented, such as NCBInr or UniRef100.

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.

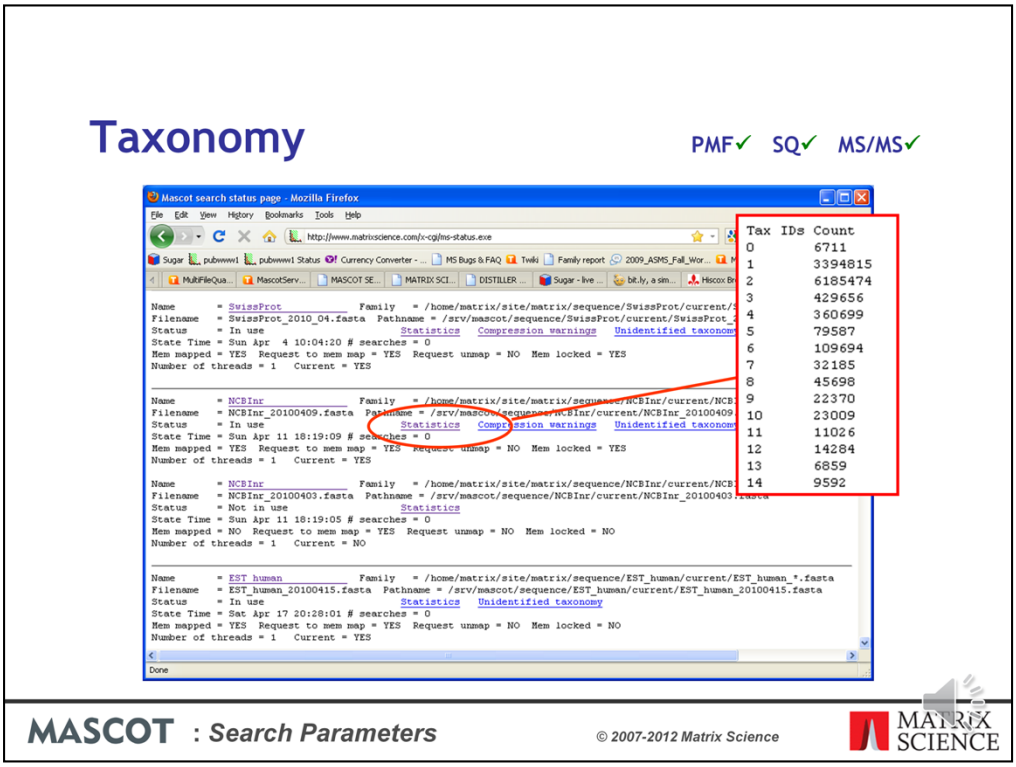
In Mascot 2.3 and later, you can select multiple databases for a search. This is particularly useful when you want to search a single organism database and include the sequences of common contaminants, such as BSA and trypsin. One restriction is that you cannot mix AA and DNA databases.

If a database contains taxonomy information, we can use this to restrict the search to entries for a particular organism or family. This speeds up the search because, in effect, it makes the database smaller.

Limiting the taxonomy simplifies the result report, because you don't see all the homologous proteins from other species.

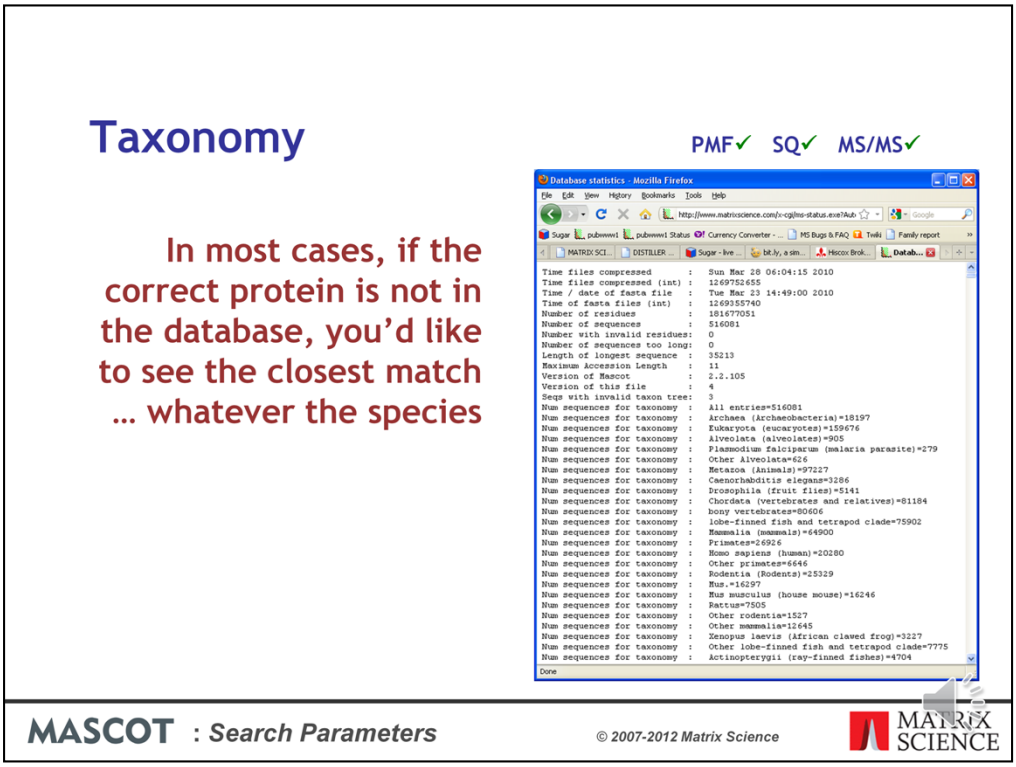
The drop down list in the search form is configurable. If you are working on a particular organism, you can easily add this to the list

It is important that the taxonomy is as accurate as possible, which means keeping the indexes up to date



From time to time, its a good idea to go to the database status page and check the stats file for each database. The stats file contains lots of useful information, like whether entries contain illegal characters or whether an entry is too long.

It also tells you how good your taxonomy is. Here are the numbers for the nr database on our web site in April 2010. There are 10.8 million entries, and 6711 have no taxonomy. In other words, better than 99.9% of the entries have a taxonomy assigned. If you look at your stats file and see that (say) 10% of the entries have no taxonomy, that's 10% of the entries that are going to be missed whenever you do a search with taxonomy specified.

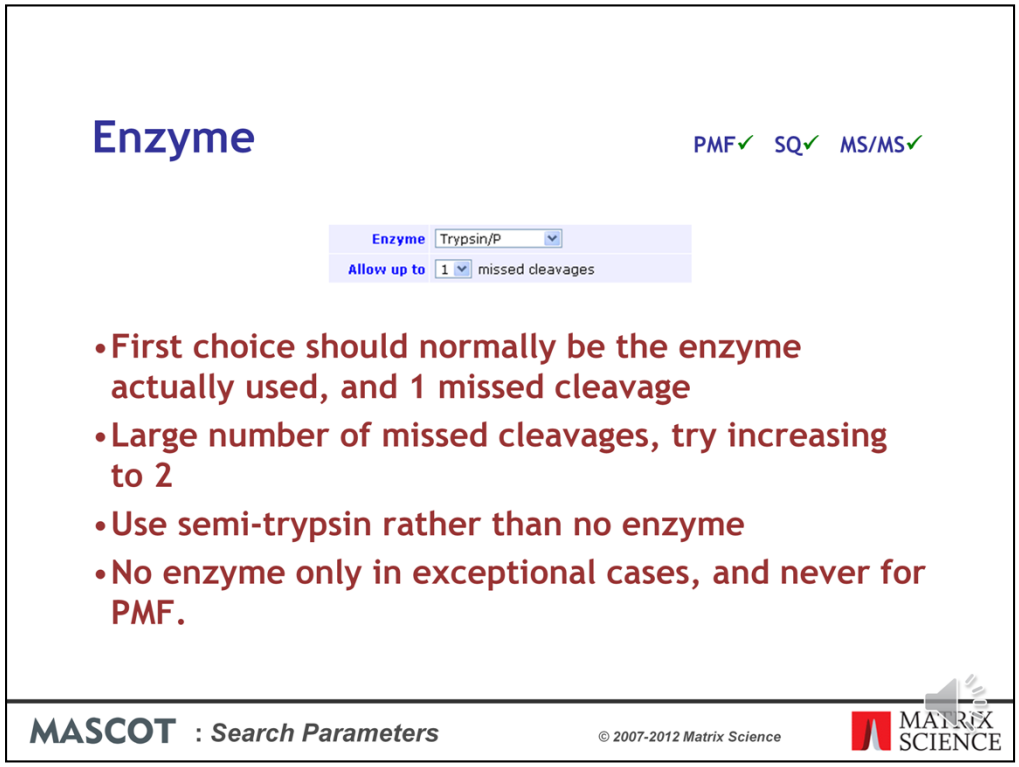


A word of warning. Don't specify a very narrow taxonomy in a search.

Think carefully about what you are trying to achieve when you do this.

If the correct protein from the correct species is not in the database, wouldn't you want to see a good match to a protein from a similar species?

This is especially important for poorly represented species. For example, look at these numbers for the Swiss-Prot 2010_04: half a million entries; 25 thousand entries for rodents, but only 1500 are not either mouse or rat. So, even if you are studying hamster or porcupine, you probably don't want to choose 'Other rodentia'.



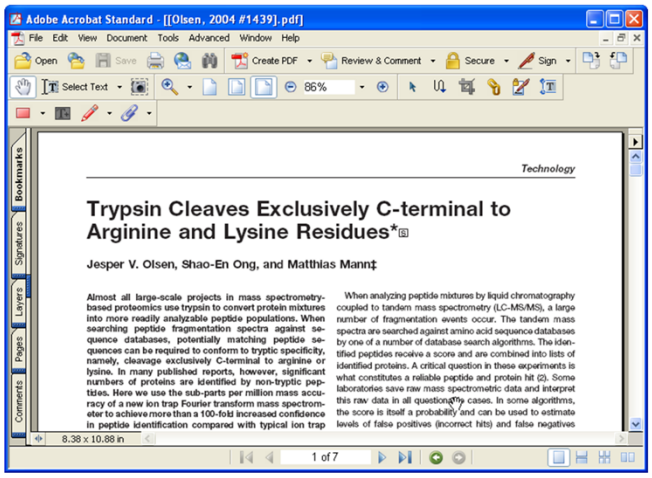
All the search forms have a drop down list for choosing an enzyme. If your peptides come from an enzyme digest, you need to know what the enzyme was and then choose it from the list.

Setting the number of allowed missed cleavage sites to zero simulates a limit digest. If you are confident that your digest is perfect, with no partial fragments present, this will give maximum discrimination and the highest score for a peptide mass fingerprint.

If experience shows that your digest mixtures usually include some partials, that is, peptides with missed cleavage sites, you should choose a setting of 1, or maybe 2 missed cleavage sites. Don't specify a higher number without good reason, because each additional level of missed cleavages increases the number of calculated peptide masses to be matched against the experimental data. In other words, the missed cleavage parameter should set by looking at some successful search results to see how complete your digests really are.

Enzyme

PMF✓ SQ✓ MS/MS✓



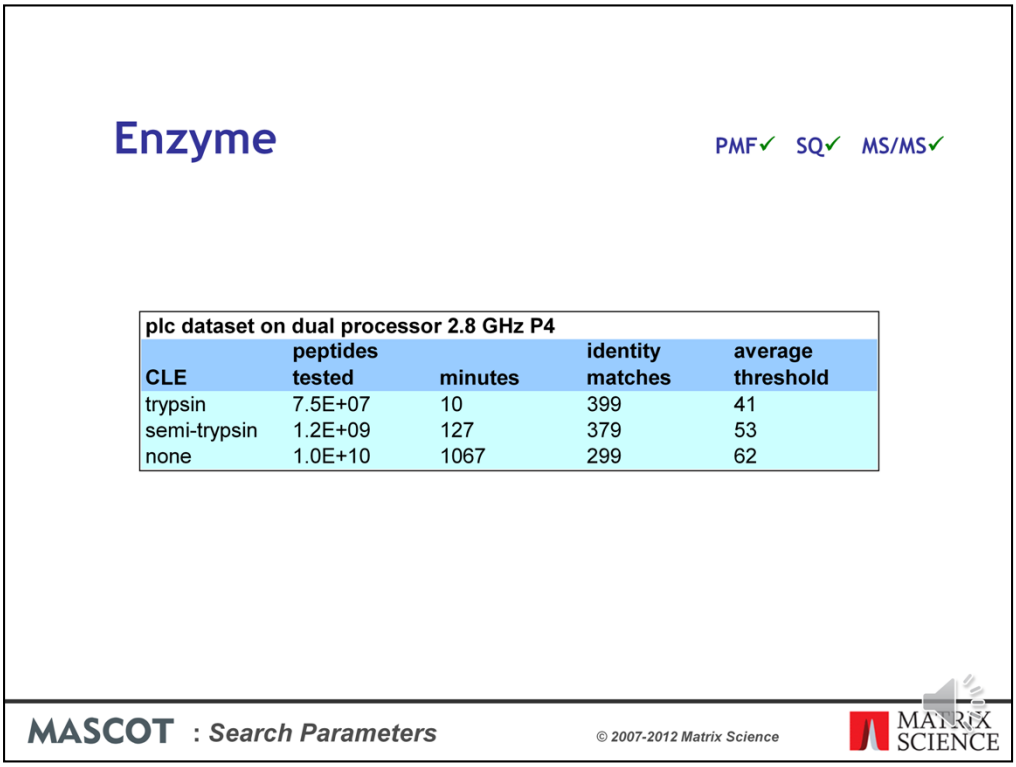
Olsen, J. V., Ong, S.-E. and Mann, M., Mol. and Cellular Proteomics, 3, 608-14 (2004)

Although some people like to perform searches without enzyme specificity, and then gain confidence that a match is correct if the match is tryptic, this isn't a good idea. If there is evidence for a lot of non-specific cleavage, then a semi-specific enzyme allows one end of the peptide to be non-specific, but not both. Only abandon enzyme specificity if you have no other choice, such as when searching endogenous peptides.

You cannot perform a no-enzyme peptide mass fingerprint. It simply won't work, even if you have good mass accuracy

There is some controversy over the level of non-specific peptides that can be expected in a tryptic digest. Our experience is that the levels of non-specific peptides are very low, less than 3%, unless there is something seriously wrong with the trypsin or the protocol.

Why do we advise so strongly against no-enzyme searches?

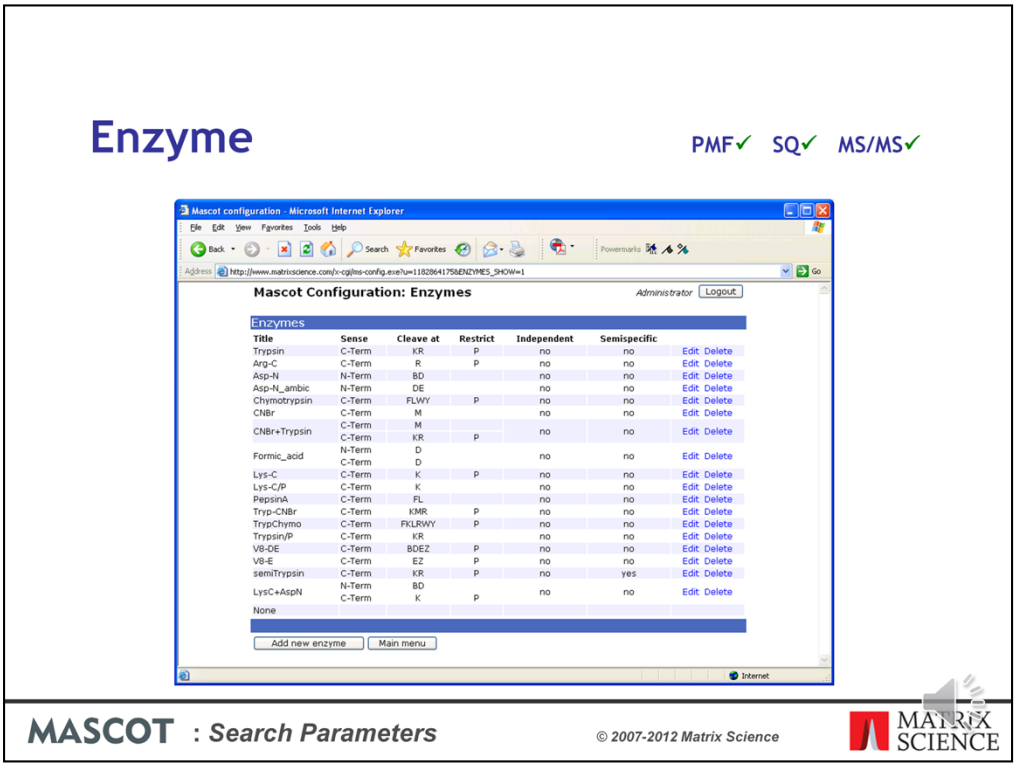


Here are some numbers for a typical ion trap dataset when we search using loose trypsin, semi-specific trypsin, and no enzyme specificity

As you can see, the no enzyme search takes a lot longer and we get fewer reliable matches.

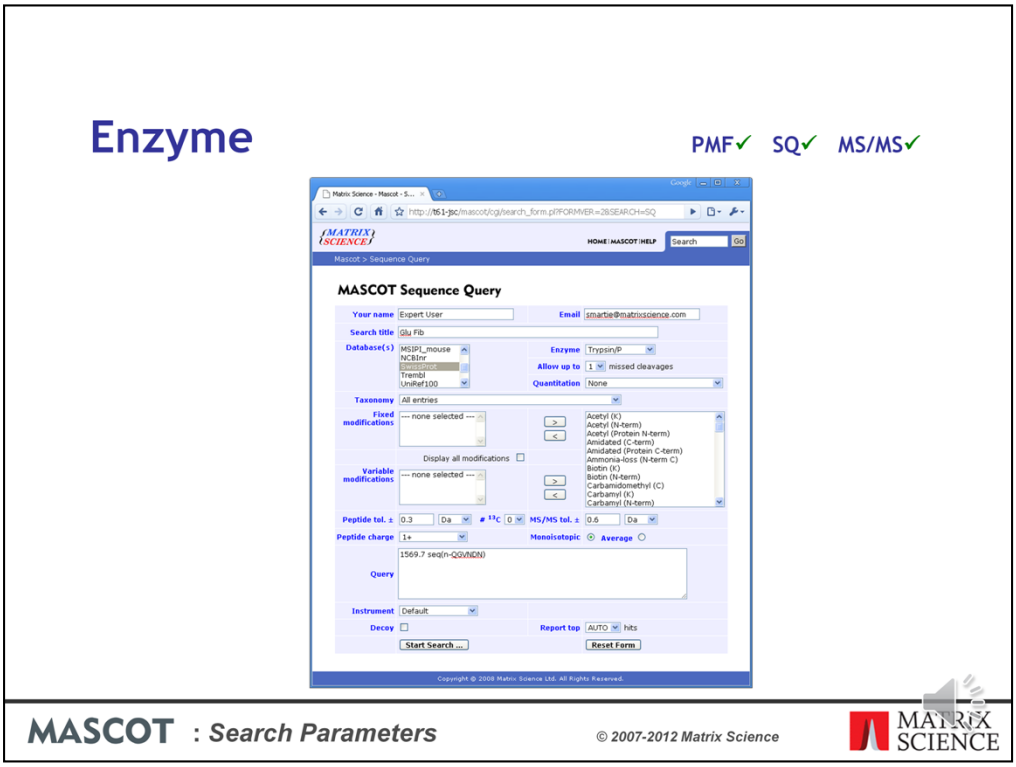
The reason is simple, the search space for a no-enzyme search is much, much larger than for a tryptic search. This means that the thresholds are higher and we lose marginal matches. Unless you have a high level of non-specific peptides, you lose more than you gain.

So, doing a no-enzyme search in Mascot is not a good idea unless there is a very high level of non-specific peptides. Semi-trypsin is almost always a better choice if the peptides came from a tryptic digest. Only use no enzyme if the peptides are not the products of a deliberate enzyme digest, e.g. MHC peptides or endogenous peptides.



The list of enzymes is user configurable. Standard entries are described in the help. If you wish, you can modify the definitions or create new ones using the configuration editor. The configuration editor was new in Mascot 2.2, so if you are using an earlier version of Mascot, you'll need to edit the configuration file called enzymes in a text editor. The format is described in the Mascot Installation and Setup Manual.

Mascot supports two categories of mixed enzyme definitions. An independent mixed enzyme is used where multiple sample aliquots have been digested separately, and the digests combined for analysis. This means that the sample could contain (say) tryptic peptides and Asp-N peptides, but no peptides that are tryptic at one end and Asp-N at the other. The second category simulates a single sample aliquot being digested simultaneously or serially by more than one cleavage agent. For example CNBr followed by trypsin.



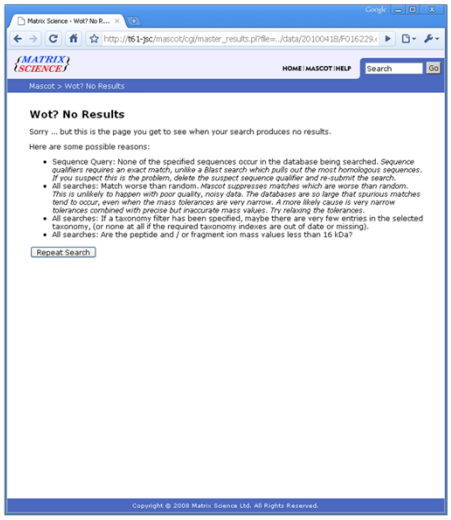
Remember that enzyme specificity also applies to Sequence Queries

One of the most common emails we receive is "Mascot is broken. I did a search for this peptide and I know its in the database but Mascot failed to find it“

For example, here's a search for glu-fib, a very common sequencing standard. The mass is correct and the sequence is correct. But, when we do a search of Swiss-Prot -

Enzyme

PMF✓ SQ✓ MS/MS✓

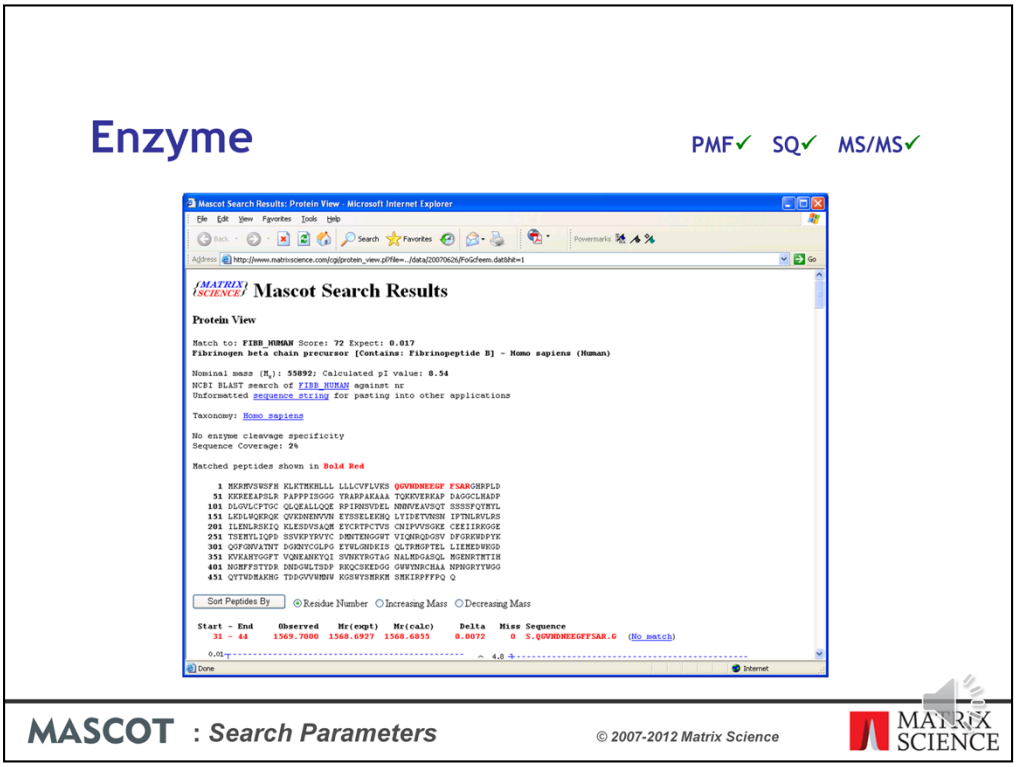


MASCOT : Search Parameters

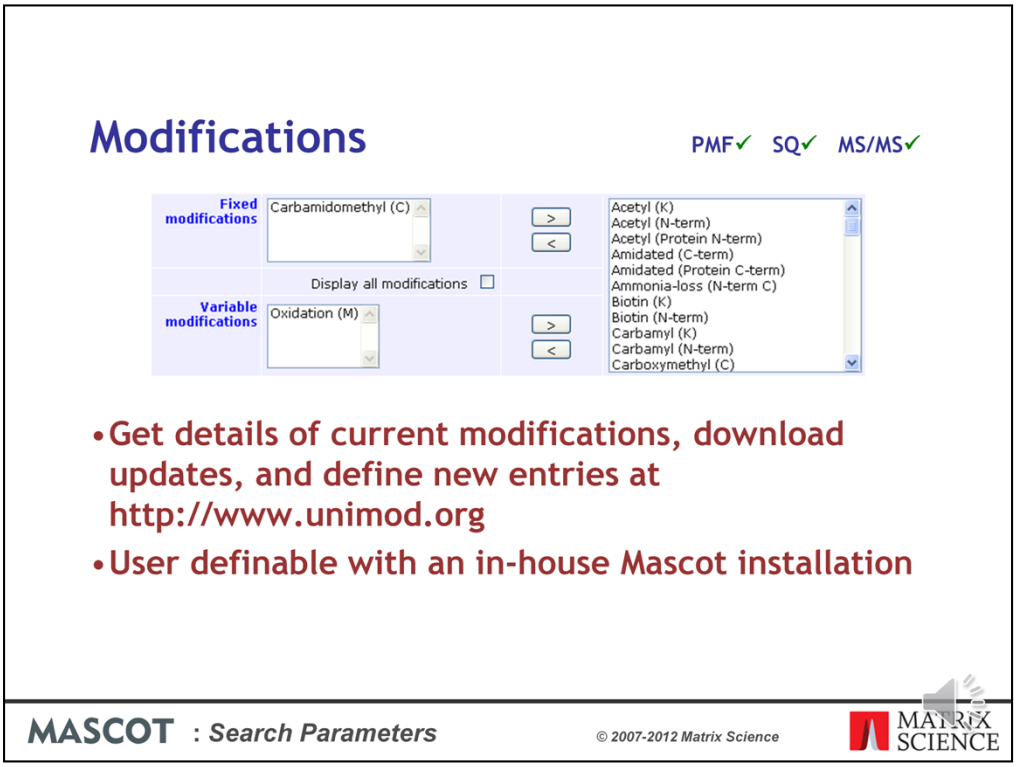
MATRIX SCIENCE

No results!

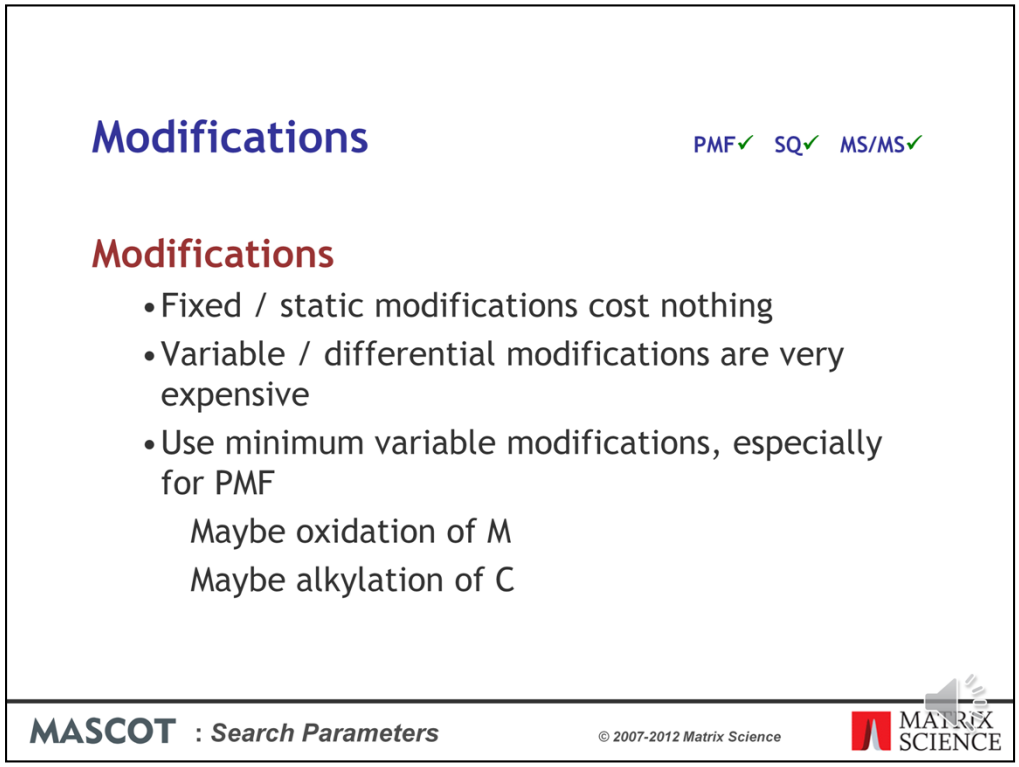
Why?



Because glu-fib in Swiss-Prot is not a tryptic peptide. The N-terminus is created by a post-translational cleavage after serine. If you now go back to the search form and select enzyme type none, bingo ... you'll get a match



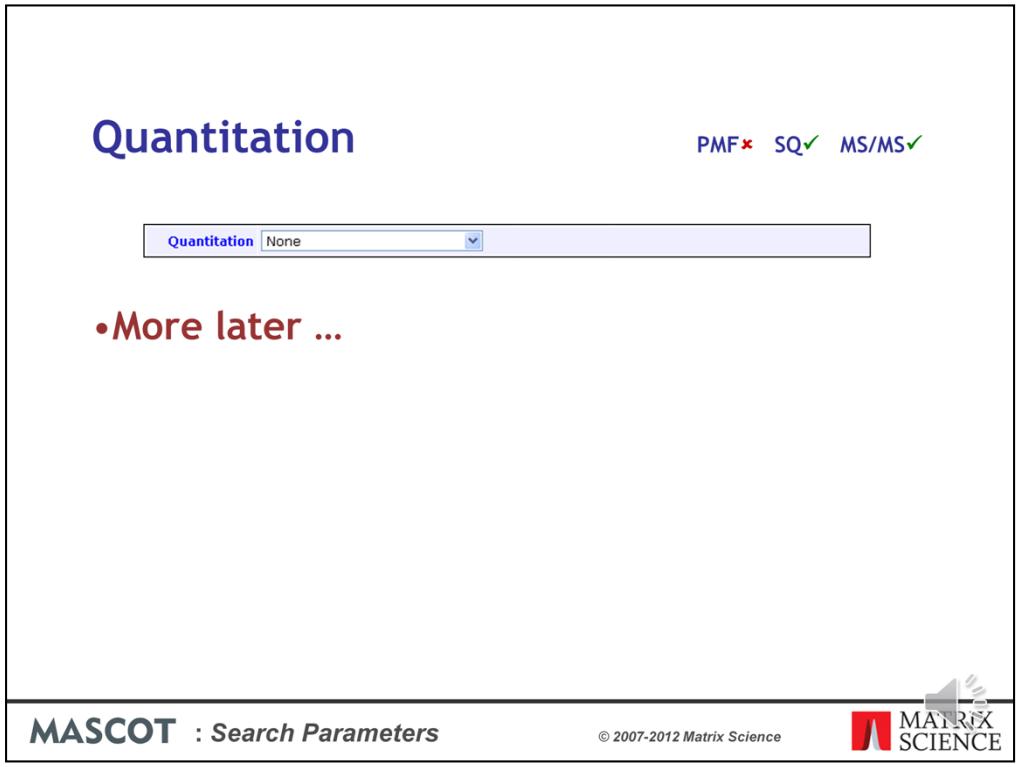
This screen shot shows how modifications are displayed in the search form in Mascot 2.3 and later. If you are using an earlier version, there are just two list boxes, one for fixed modifications and one for variable. In the newer arrangement, you move modifications from the single list on the right to and from the lists on the left. This makes it easier to see at a glance what has been selected for the search. If the checkbox labelled 'Display all modifications' is clear, as shown here, you get a relatively short list of the most common modifications. If you check the box, a much longer list is available. You can keep your list of modifications up-to-date by downloading the latest information from Unimod. If you have a modification which you don't want to share with others, you can add it to the local configuration file. We'll describe how to go about doing this in detail in the Mascot Server Administration talk.



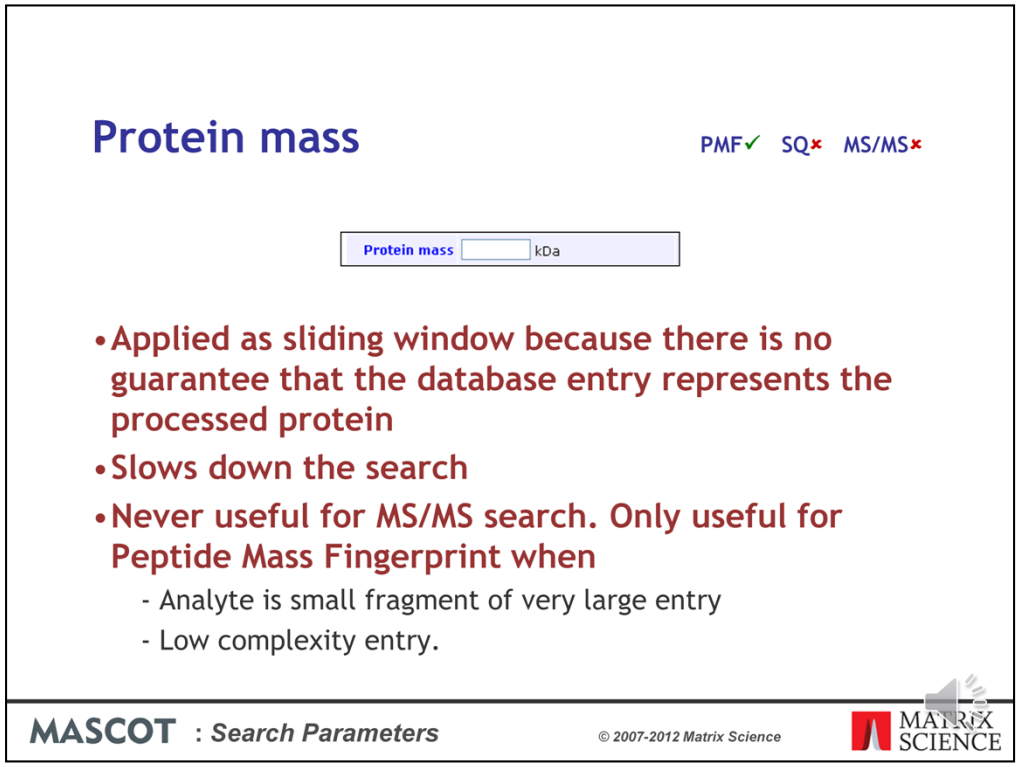
Modifications in database searching are handled in two ways. First, there are the fixed or static or quantitative modifications. An example would be a the efficient alkylation of cysteine. Since all cysteines are modified, this is effectively just a change in the mass of cysteine. It carries no penalty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



Quantitation is the subject of a separate presentation.

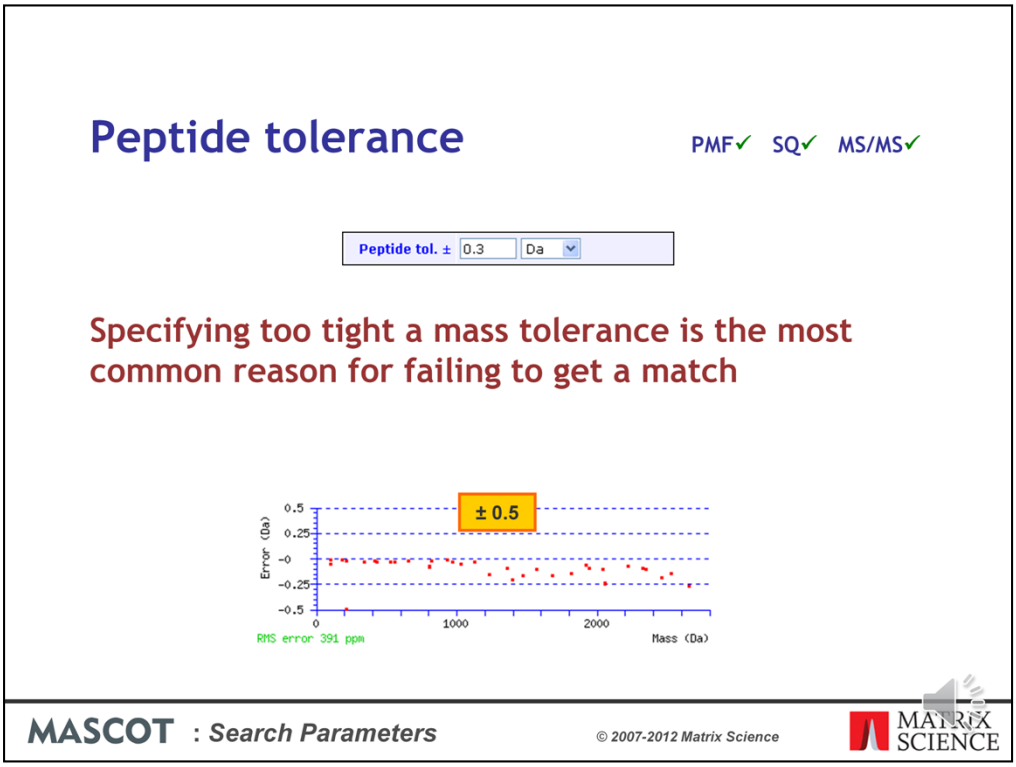


The protein mass is the mass of the intact protein in kDa applied as a sliding window. That is, the mass of the contiguous stretch of sequence which contains all of the matched peptide mass values. This will generally be less than the mass of the entire sequence entry. Consequently, if you specify a value for the protein mass, this acts only as a ceiling. Not only will you see smaller proteins on the hit list, you will also see larger ones, but all of the reported matches will be within a stretch of sequence less than or equal to the specified mass.

If this field is left blank, there is no restriction on protein mass

Specifying a protein mass will slow down the search a little.

Its hard to find examples where this parameter is useful. We include it mainly because many people requested it. It could give a better score if the analyte was small fragment of very large entry, or a low complexity protein. But, you can’t know this in advance, so our general recommendation is to leave the protein mass open



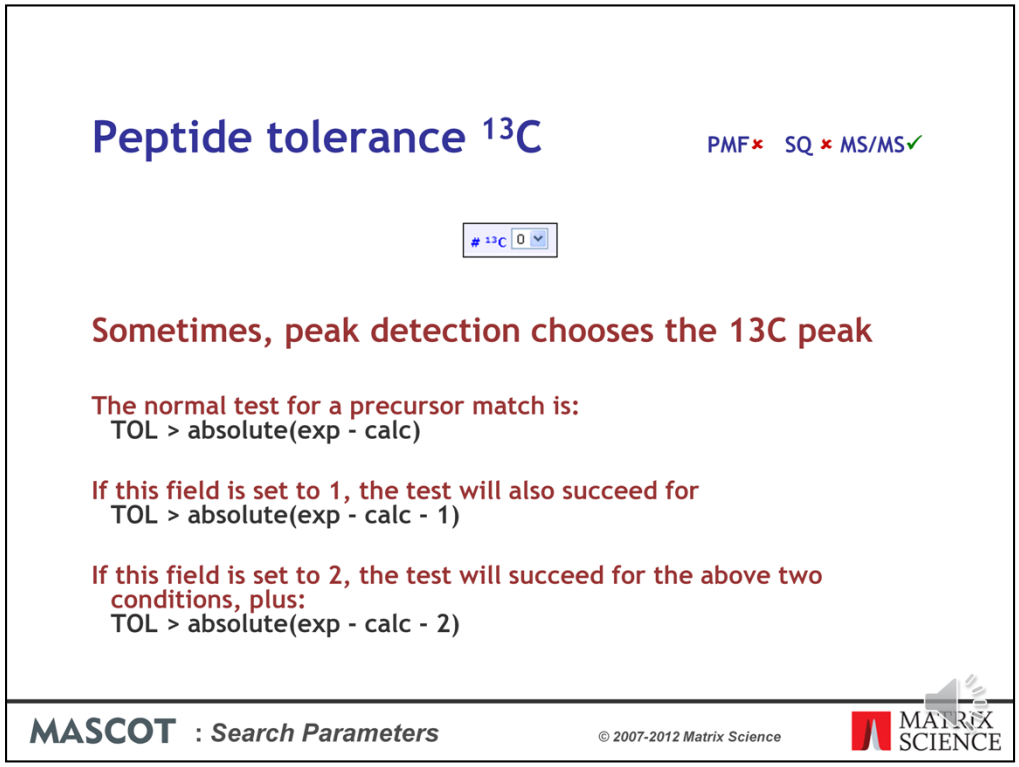
This is the error window on experimental peptide mass values, not the error window for MS/MS fragment ion mass values, which is set using the MS/MS tol. ± parameter.

Units can be selected from: percentage, milli-mass units, parts per million, or Daltons.

Specifying too tight a tolerance is a very common reason for failing to get a match.

Making an estimate of the mass accuracy doesn't have to be a guessing game. Protein View includes a graph of the mass errors for intact peptides. Just search a strong standard and look at the error graph. You'll normally see some kind of trend. Add on a safety margin and this is your error estimate. If you see something that looks like this, a mass tolerance of +/- 0.5 Da is about right. It gives some safety margin. Remember that there will always be the odd outlier, like the data point at the lower left. It is the general trend and distribution of the majority of the data points that is important.

For a peptide mass fingerprint, the score depends on the peptide tolerance. In an MS/MS search, this parameter has no effect on the ions score. However, it does affect the search time. The larger the tolerance, the longer the search will take.



Sometimes, peak detection chooses the 13C peak rather than the 12C. In extreme cases, it may pick the 13C2 peak. The normal test for a precursor match is:

TOL > absolute(exp - calc)

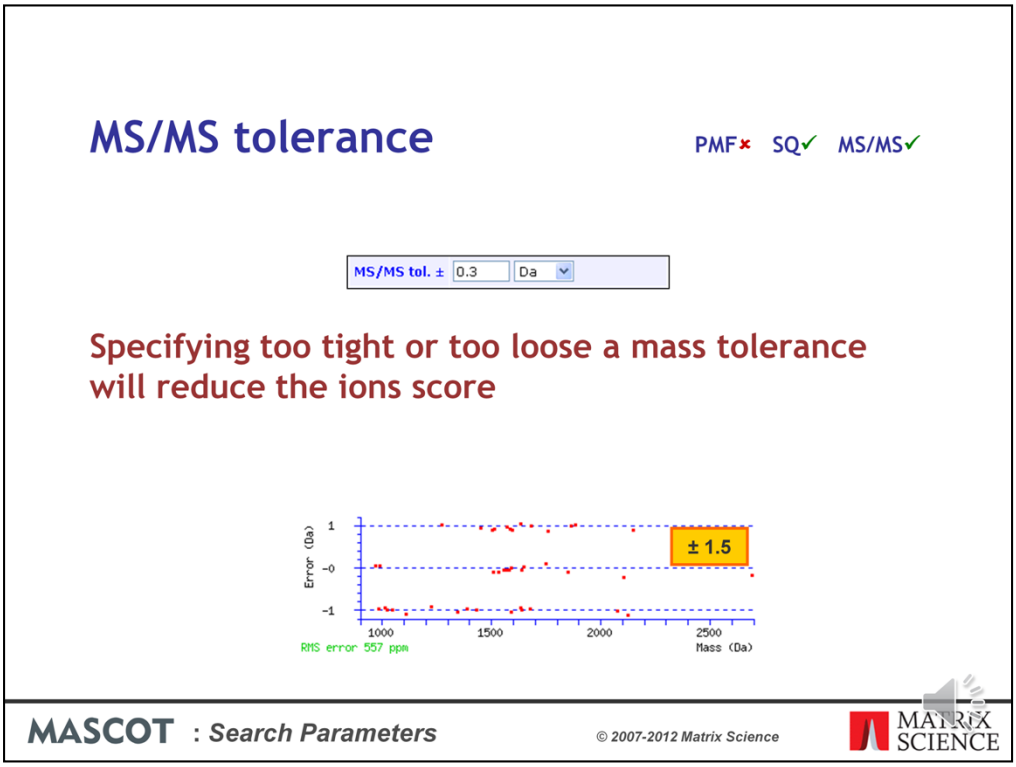
Assuming the mass values and tolerance are in Da, if this field is set to 1, the test will also succeed for

TOL > absolute(exp - calc - 1)

If this field is set to 2, the test will succeed for the above two conditions, plus:

TOL > absolute(exp - calc - 2)

This means that you can use a tight mass tolerance and still get a match to a 13C peak. If you are using a very high accuracy instrument, note that the precise shifts are the carbon isotope spacings of 1.00335 and 2.00670, rather than 1 and 2.

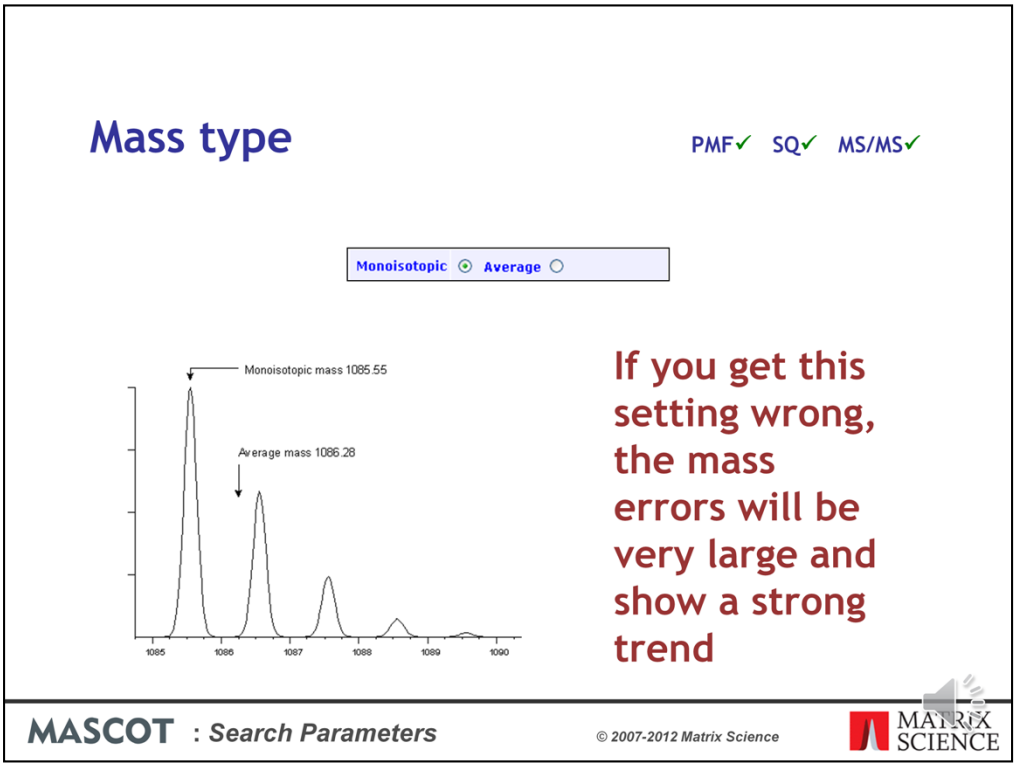


This is the error window on MS/MS fragment mass values.

Units can be milli-mass units or Daltons.

Specifying too tight or too loose a mass tolerance will reduce the ions score. Peptide View includes a graph of the mass errors for fragment ions.

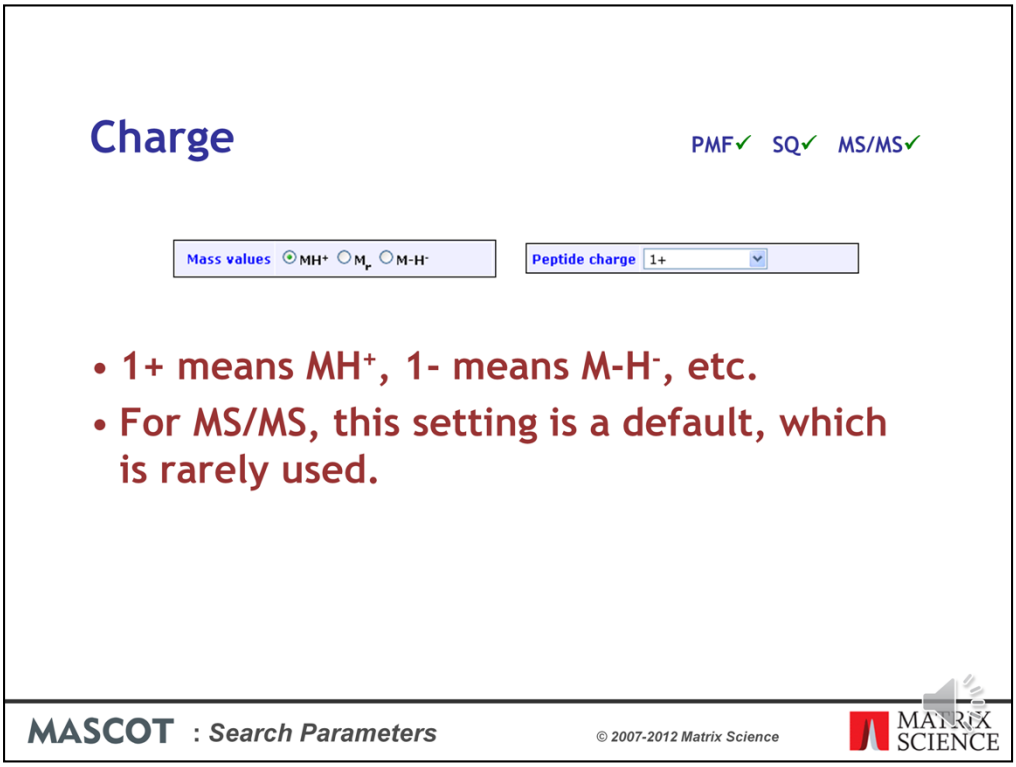
Here, the mass tolerance is much too high. A more appropriate tolerance might be +/- 0.3. Having a tolerance which is much too high can sometimes lead to artefacts and false positives



Mass type specifies whether the experimental mass values are average or monoisotopic. Monoisotopic mass is the mass of the peptide where all atoms are the most abundant natural isotopes of their elements, e.g. Carbon 12, Nitrogen 14, Hydrogen 1, etc. In most cases, this is the first peak of the natural isotope distribution. Average mass is the chemical mass, which is the centre of gravity of the isotope distribution.

In Mascot, you cannot mix the two, and have (say) average precursors and monoisotopic fragments.

Most modern instruments produce monoisotopic mass values. You will only have an average mass if the entire isotope distribution has been centroided into a single peak, which usually implies very low resolution. If you get this setting wrong, the mass errors will be very large and show a strong trend, because the difference between an average and a monoisotopic mass for peptides and proteins is approximately 0.06%.

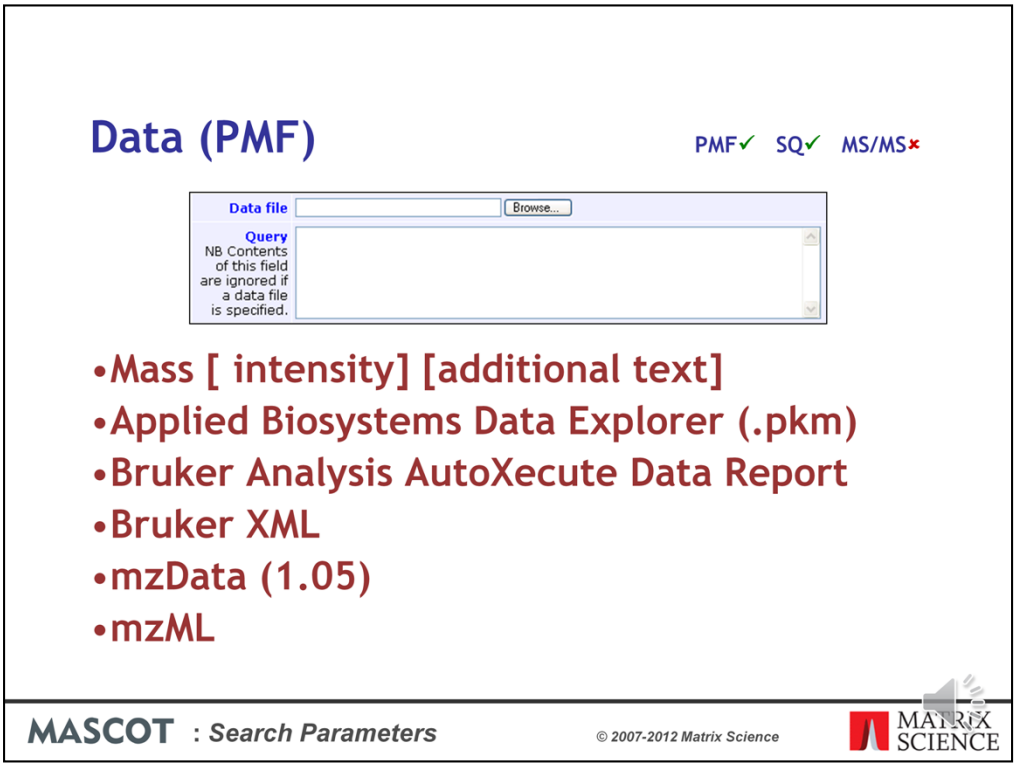


These fields are used to specify the peptide charge state. The radio buttons are from the peptide mass fingerprint form. The drop down list is used on the sequence query and MS/MS forms.

The notation "1+", "2+", etc. is used to save space and because some HTML form fields do not support the use of superscripts and subscripts. "1+" always means MH+, "1-" always means M-H-, etc.

For MALDI-PSD, the precursor peptides will generally be MH+, so the charge state should be set to "1+“

For an MS/MS search, the value specified here is a default. Most peak lists always specify a charge state, so default is never used.

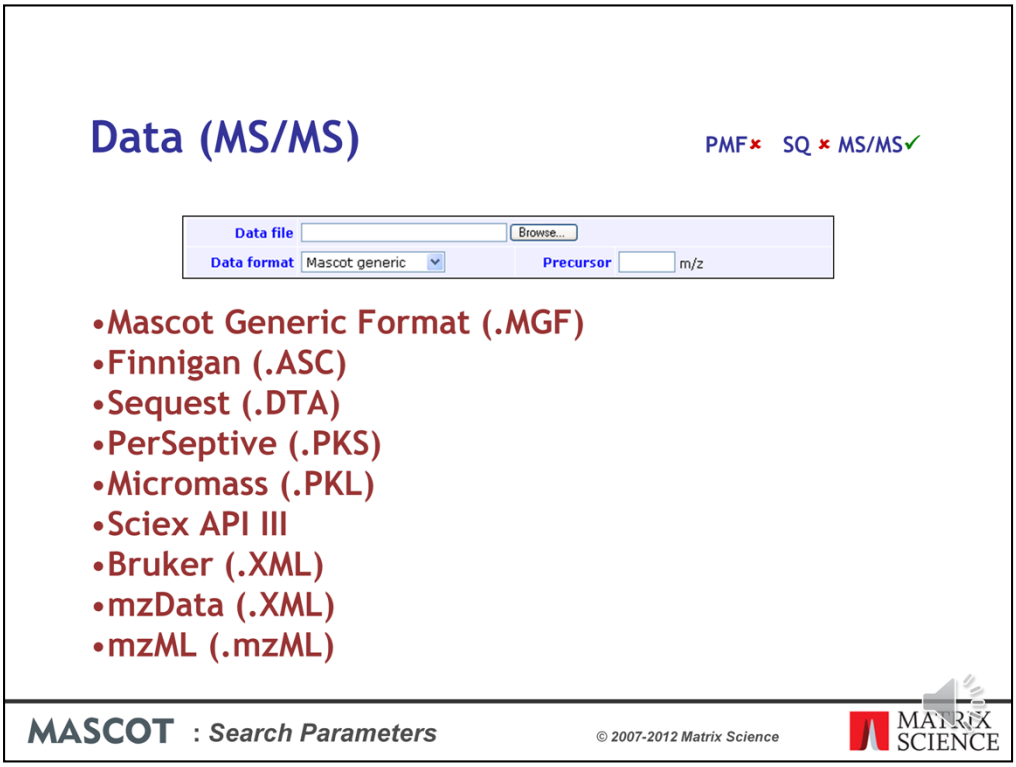


The contents of the query window on the peptide mass fingerprint form are only used when no data file has been specified.

The data format for a peptide mass fingerprint is auto detected. It can be a simple list of mass values, one per line. If a second values is present, it is assumed to be intensity. Any further values on the same line are ignored

Mascot also supports other peak list formats, as listed.

mzData is the standard interchange format sponsored by the HUPO Proteomics Standards Initiative working group

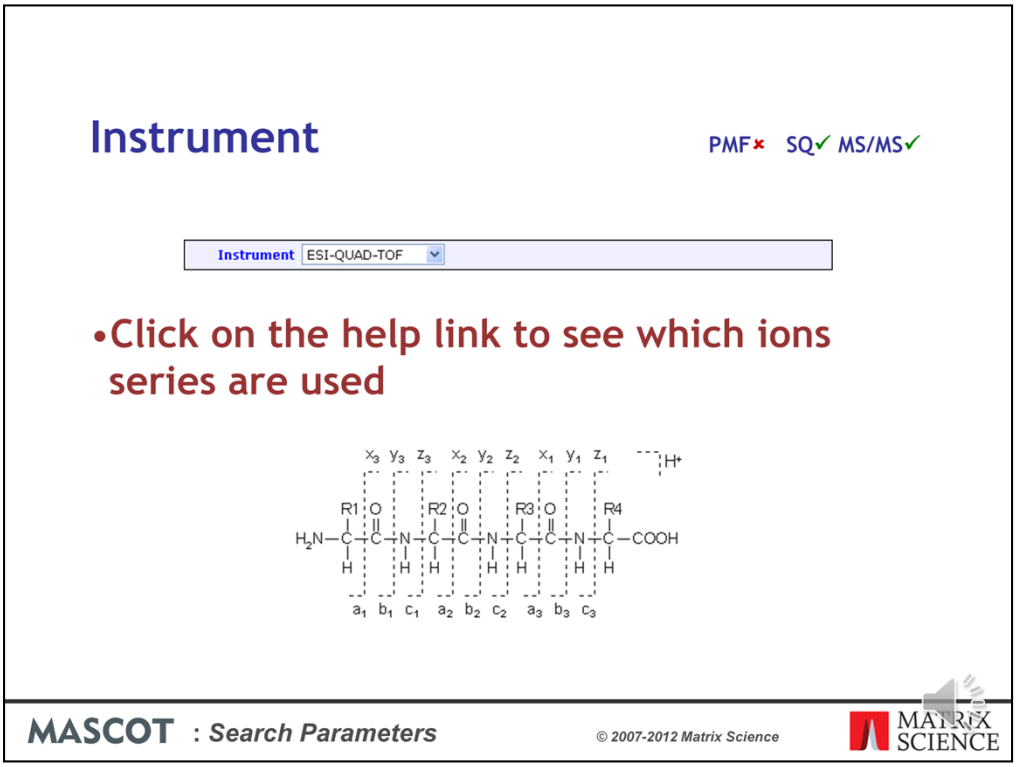


Data for MS/MS ion searches must be supplied as an ASCII file in one of these supported formats. The format cannot be auto-detected, and must be specified using the drop down list.

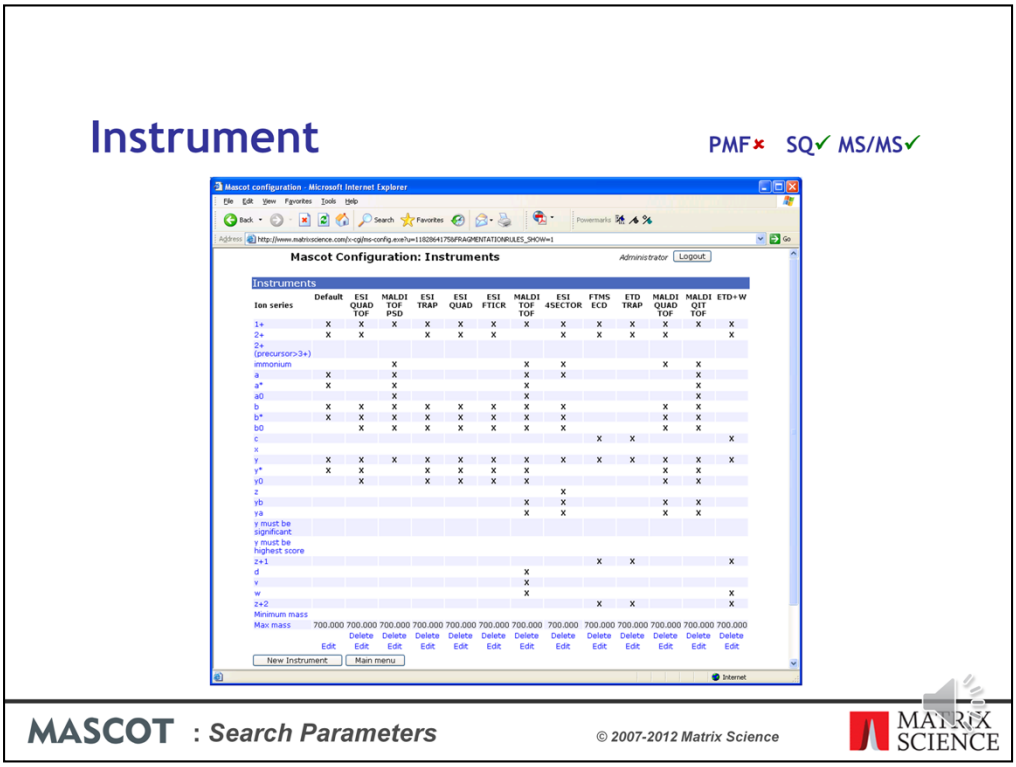
Certain data file formats, SCIEX API III, PerSeptive (.PKS), and Bruker (.XML), do not include m/z information for the precursor peptide. For these formats only, the Precursor field is used to specify the m/z value of the parent peptide.

A data file may include embedded search parameters. Most embedded parameters can only appear once, at the head of the data file. In a Mascot generic format file, a few parameters can appear within an MS/MS dataset. See the Data File Format help page for further details

If there is a conflict between the values of the embedded parameters and values entered into search form fields, the embedded parameters always take precedence. The search form fields are essentially defaults for values missing from the data file.



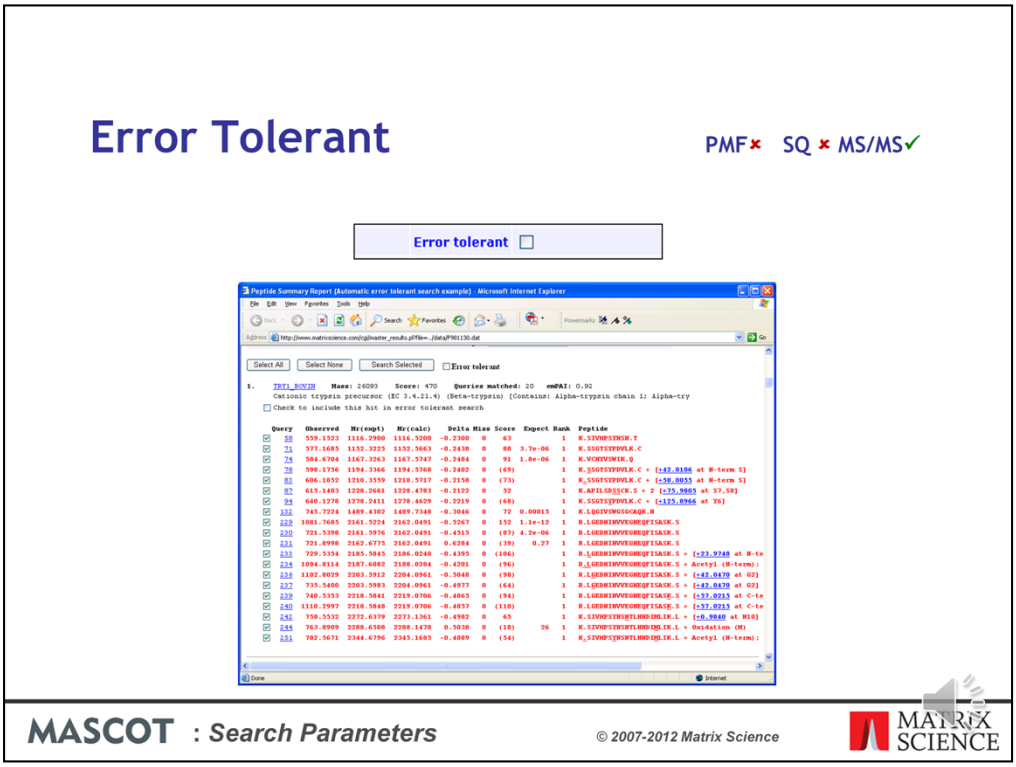
For an MS/MS Ions Search, choose the description which best matches the type of instrument used to acquire the data. This setting determines which fragment ion series will be used for scoring, according to the following table.



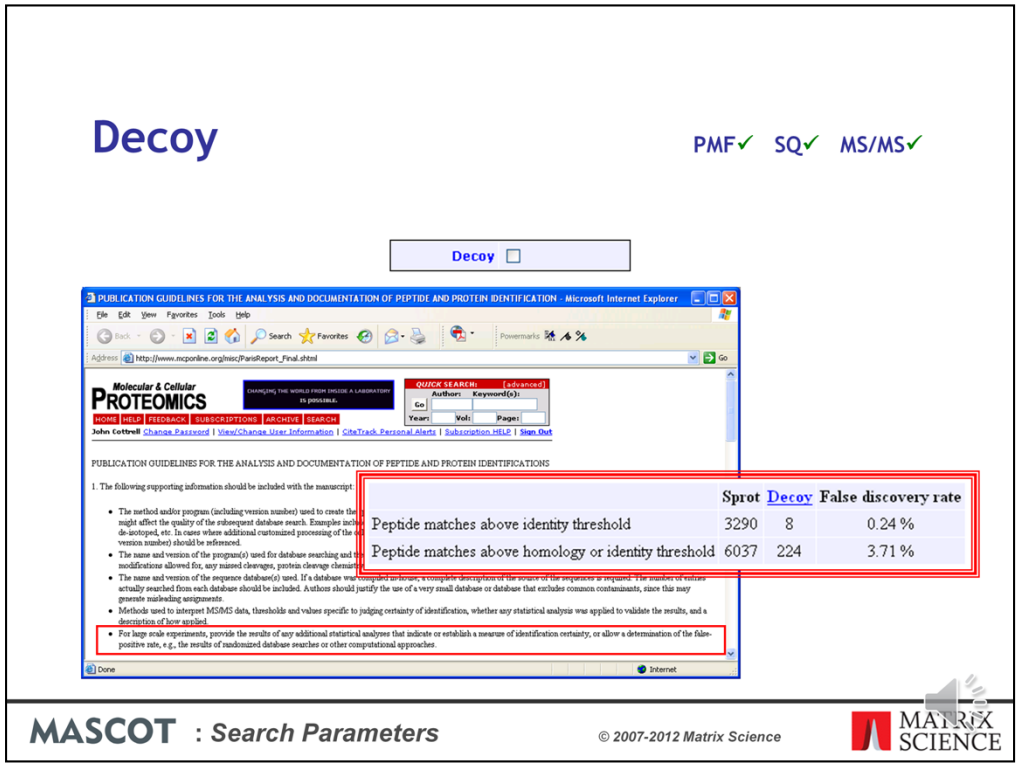
"Default" corresponds to the configuration used in Mascot version 1.7 and earlier.

Many of the instruments are very similar.

You can modify instrument settings or create new ones using the configuration editor. In this screenshot, the right hand column is an experiment to see how the addition of w ions affects ETD matching

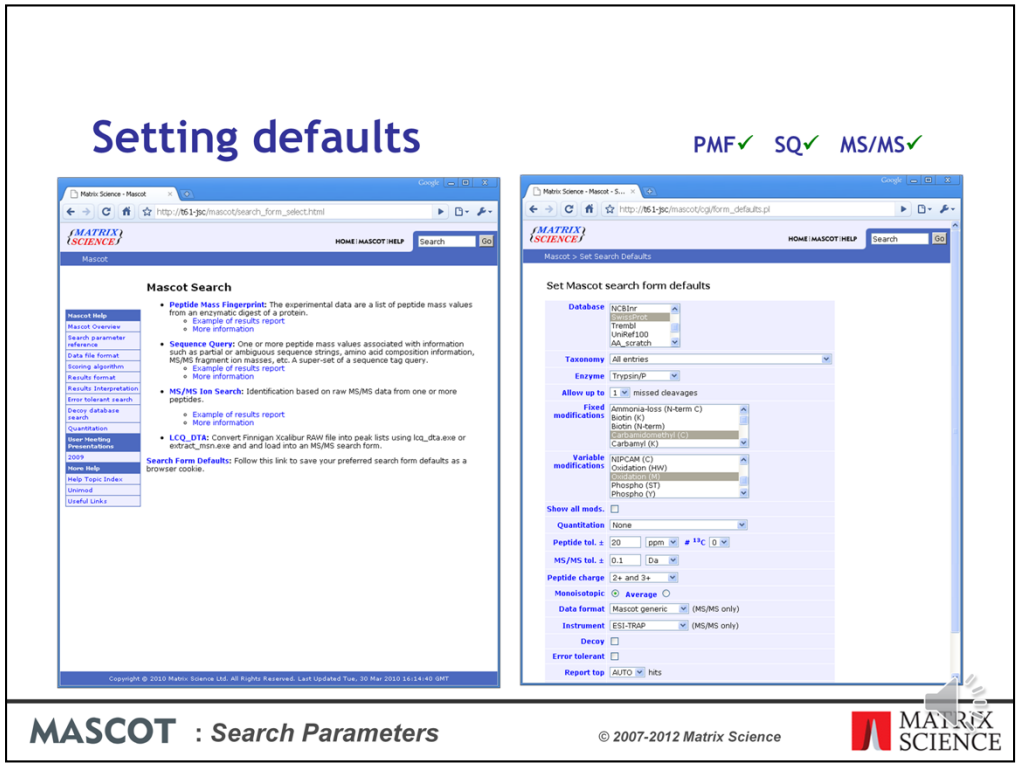


If you have MS/MS data, and are interested in finding post-translational modifications, you can perform an error tolerant search by checking this box on the search form. This is a much more efficient way to discover unusual modifications, as well as non-specific peptides and sequence variants. More about this in a later presentation.



The decoy checkbox enables you to validate the false discovery rate according to the approach recommended in the Molecular & Cellular Proteomics Guidelines for Publication: “For large scale experiments, provide the results of any additional statistical analyses that indicate or establish a measure of identification certainty, or allow a determination of the false-positive rate, e.g., the results of randomized database searches or other computational approaches”

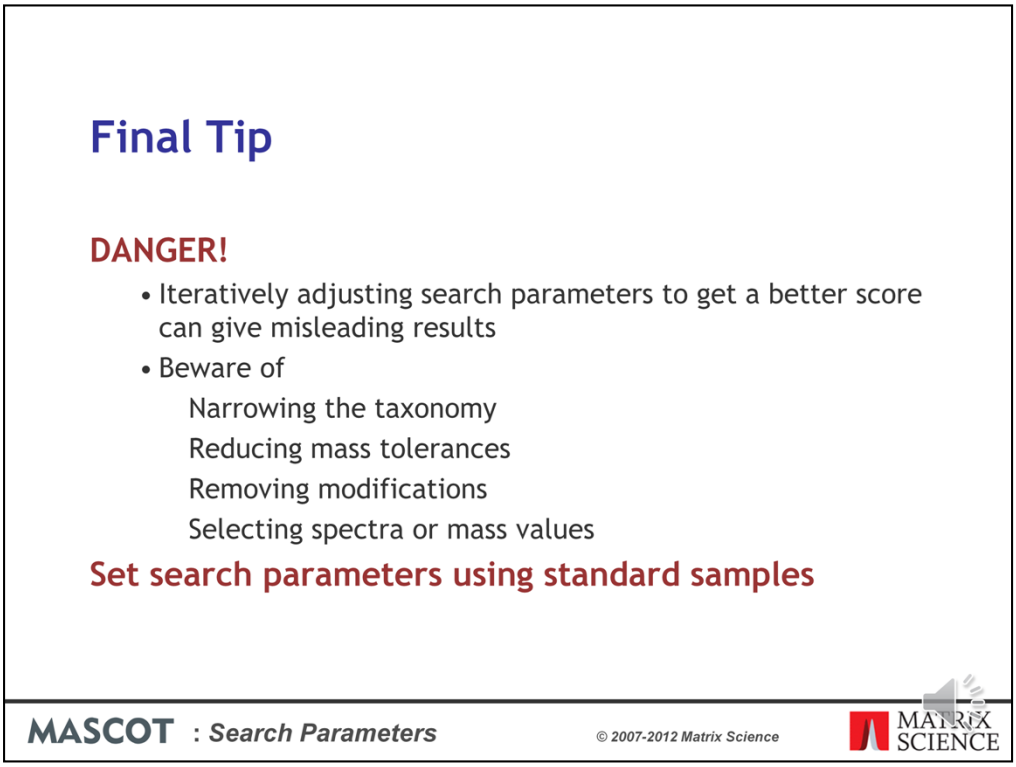
REPORT determines the *maximum* number of hits displayed in a search results report. Choose AUTO to display only protein hits with significant scores. In a protein summary report, one additional hit is reported after the cutoff at the significant score. This is to ensure that the report shows the highest scoring hit, even though it is not significant.



You can choose your own defaults for the search forms. Look for the link at the bottom of the search form selection page

In particular, if you are using Mascot 2.2 or earlier, this is where you choose whether to display the full modifications list or just a short list of the most common mods

When you save the defaults, they are saved as a browser cookie. If you go to a different PC, or switch to a different browser, you'll need to repeat this step



A final word of advice: 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.