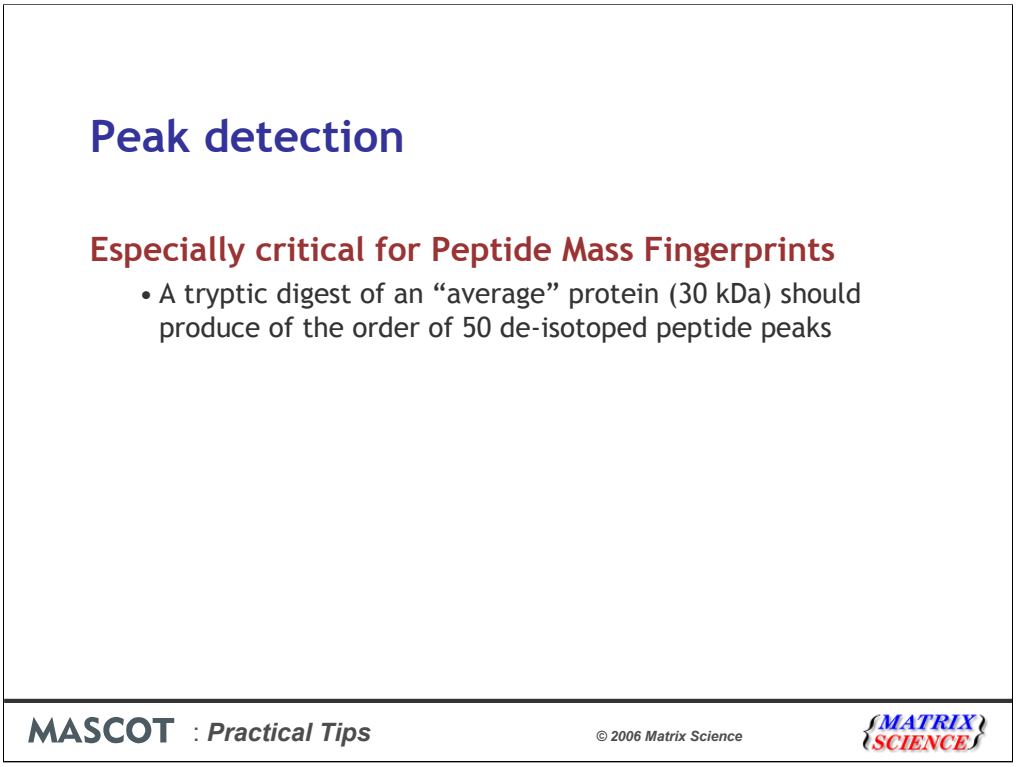
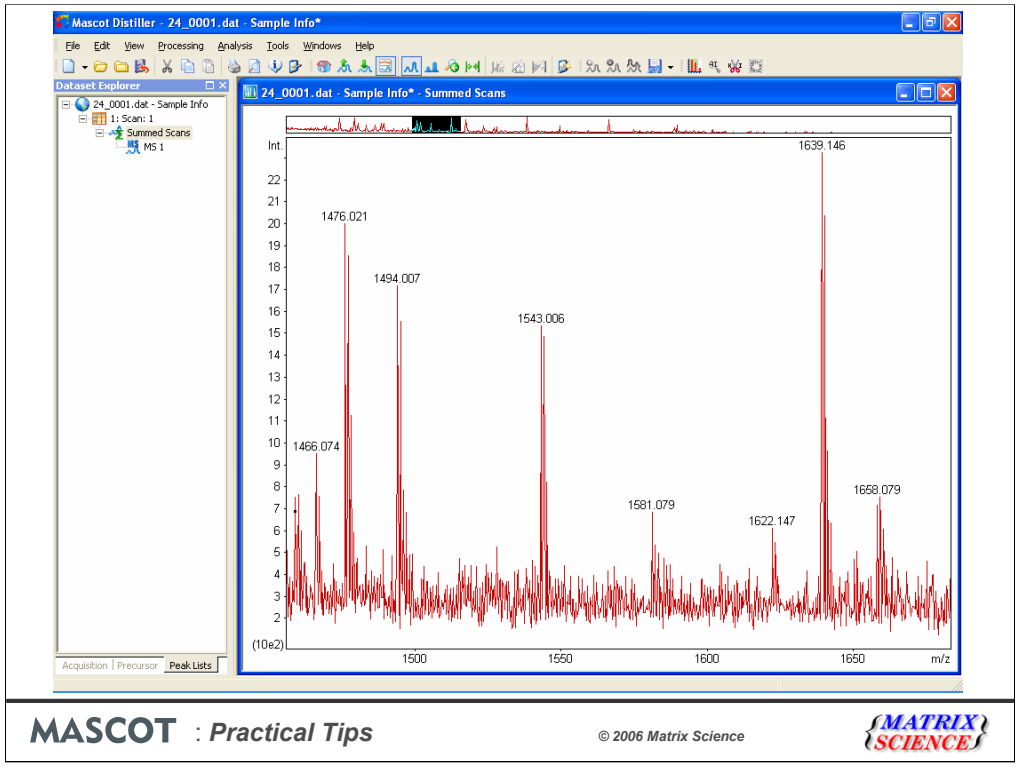
Practical Tips

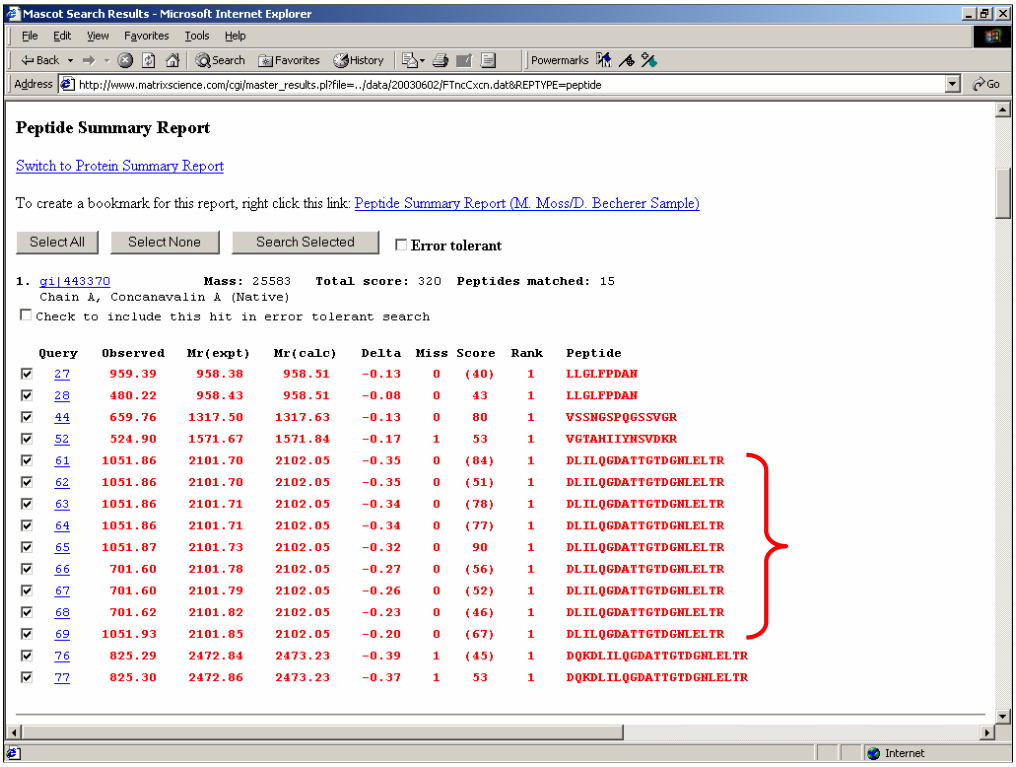


Everyone is familiar with the phrase "garbage in garbage out". In the case of database searching, this reminds us that the results are only as good as the peak list.

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, more likely, 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.



If you are searching MS/MS data, and your results look like this, with the same peptide identified over and over again, this is also a peak detection problem. 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 ratio of the summed spectrum is improved, making the identification more reliable, (ii) the report is more concise, (iii) the search is faster

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

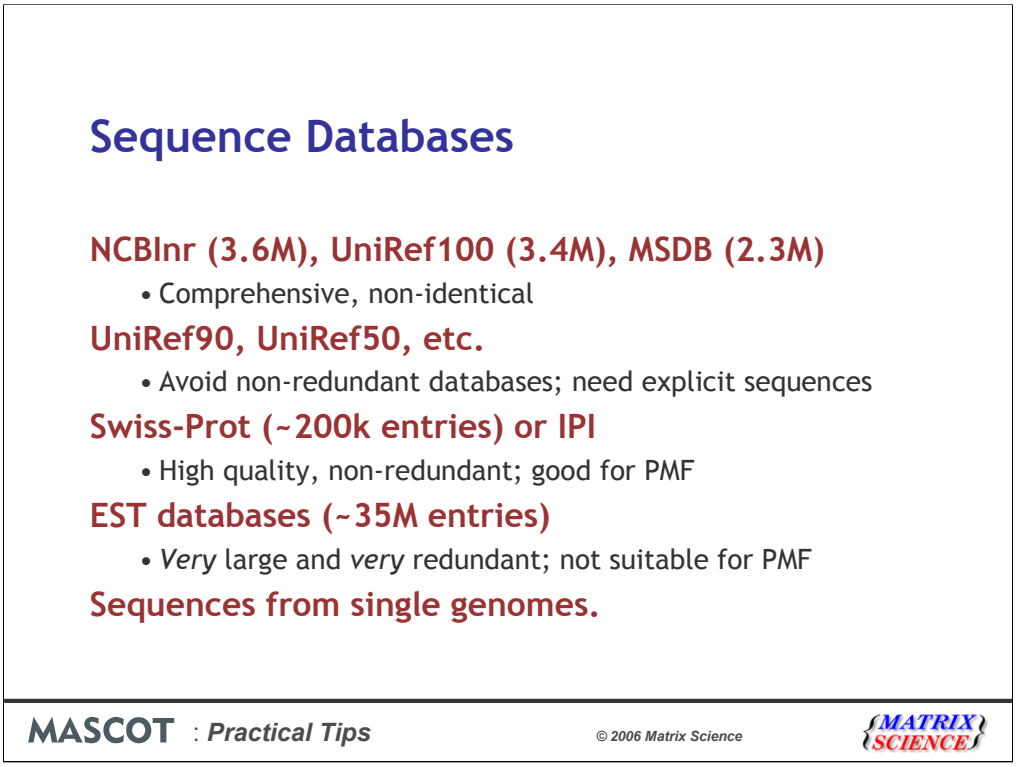
If in doubt, throw it out

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

If in doubt, throw it out.

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?



Which sequence database to choose?

The large, comprehensive, non-identical databases are usually 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 and IPI are curated, well annotated databases, but they are also non-redundant. These are good choices for PMF searches, where the loss of one or two peptides may not be a concern. Not such good choices for searching MS/MS data, because you need the exact peptide sequence to be explicitly represented in the database.

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.

If you have an in-house server, and are only interested in proteins from a particular organism, you can search a database of sequences from the genome. For MS/MS data, this can be the ORFs or even the raw genomic DNA sequence

Taxonomy

In most cases, if the correct protein is not in the database, you'd like to see the closest match … whatever the species

If the database allows you to set a taxonomy filter, 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 46.2: 172 thousand entries; 14 thousand entries for primates, but most of these are for human. So, even if you are studying chimps or orang-utans or yeti, you probably don't want to choose 'Other primates'.

Modifications

- Get details of current modifications, and define new entries at
  http://www.unimod.org
- User definable with an in-house Mascot installation

MASCOT : Practical Tips © 2006 Matrix Science MATRIX SCIENCE

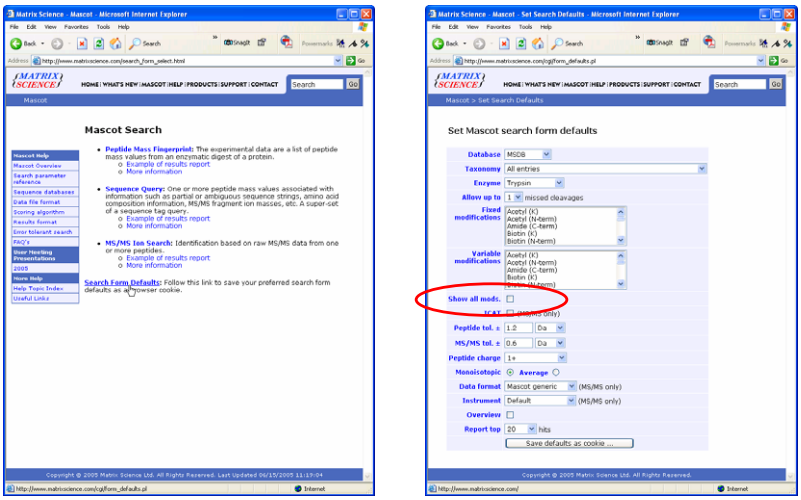
There are two identical lists of modifications on the search form. One for fixed modifications, such as alkylation of cysteine. Since all cysteines are modified, this is effectively just a change in the mass of cysteine.

The second list is for variable modifications, such as phosphorylation, which might affect just one serine in a peptide containing many serines.

You can keep your list of modifications up-to-date by downloading the latest mod_file from Unimod. This is also where you can find details of individual modifications, such as the elemental formula

If you have a modification which you don't want to share with others, then you can add it to the local mod_file, which s just a text file. The format is described in the Mascot Installation and Setup Manual.

Setting defaults

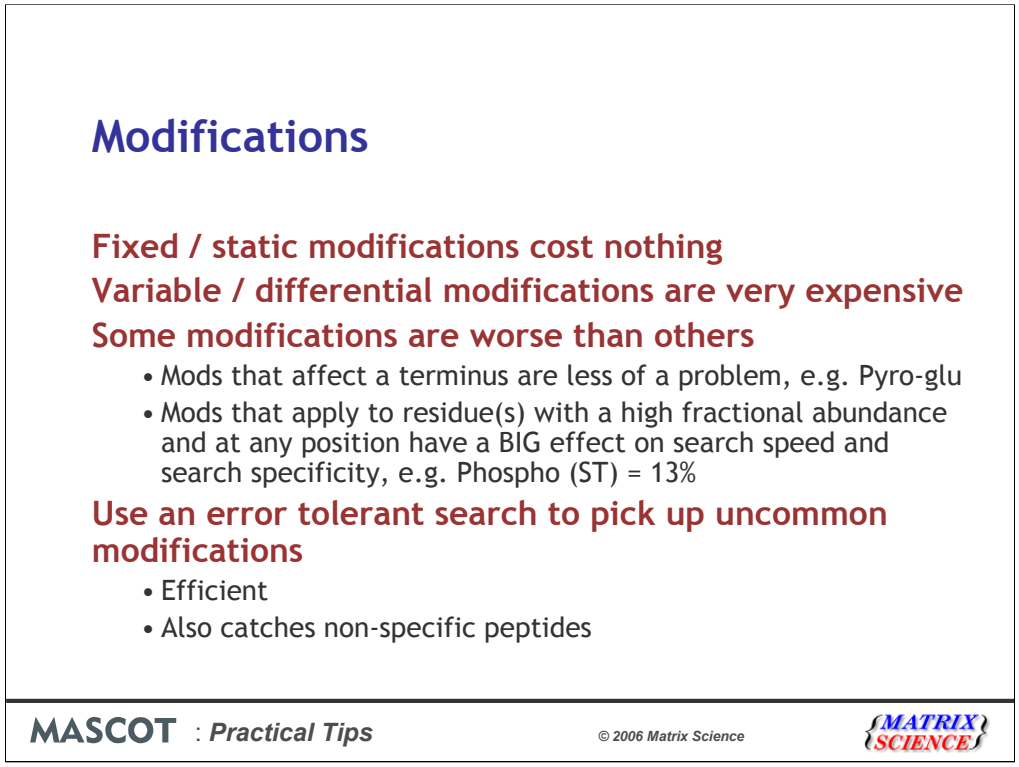


MASCOT : Practical Tips © 2006 Matrix Science MATRIX SCIENCE

Many people don't realise that the default list of modifications is a short list of just the most common ones. To see a full list of modifications, you need to go to the search form defaults page. Look for the link at the bottom of the search form selection page.

Check the checkbox for Show All Mods., and set the other defaults to typical values for your searches.

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

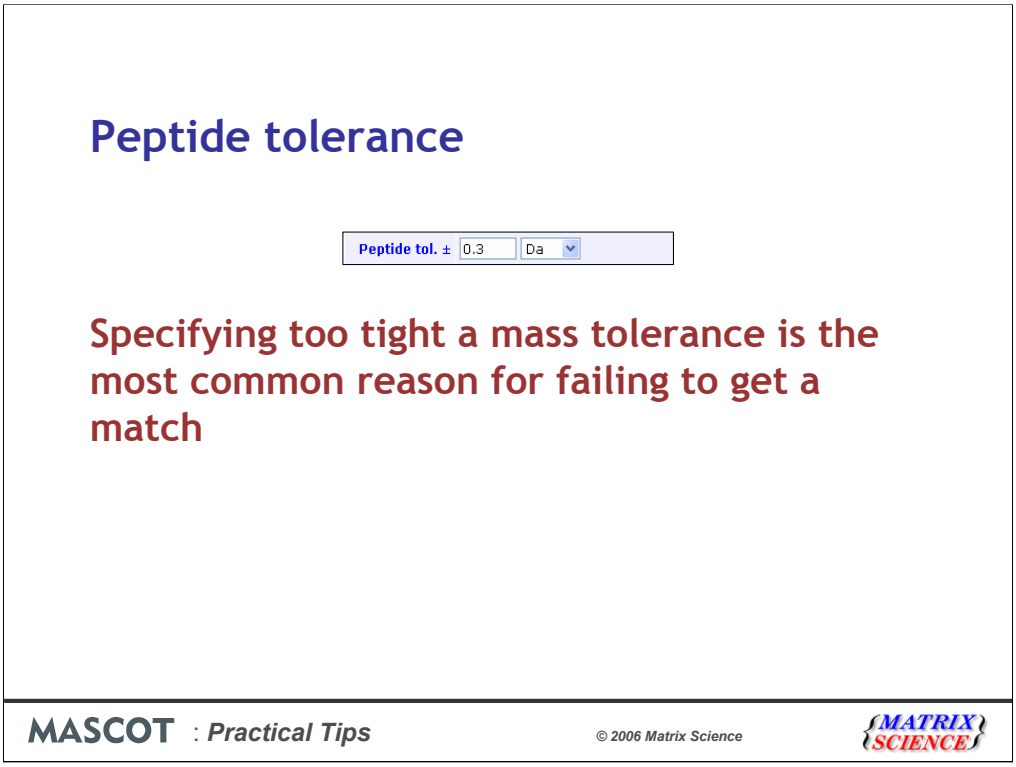


Fixed modifications cost nothing in terms of the time taken for a search or the search specificity. As described earlier these simply cause a change in the mass of all instances of the affected residue.

Variable 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. Some modifications are worse than others for this, but if a modification applies to a common residue and can occur at any position, then this can have a very large effect on the search speed and specificity.

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.

If you are hunting for post-translational modifications, use an error tolerant search. It is both more efficient and more comprehensive

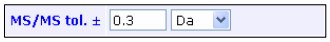


This is the error window on experimental peptide mass values, not the error window for MS/MS fragment ion mass values, which is set using the MS/MS tol. ± parameter.

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

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.

MS/MS tolerance



Specifying too tight or too loose a mass tolerance will reduce the ions score

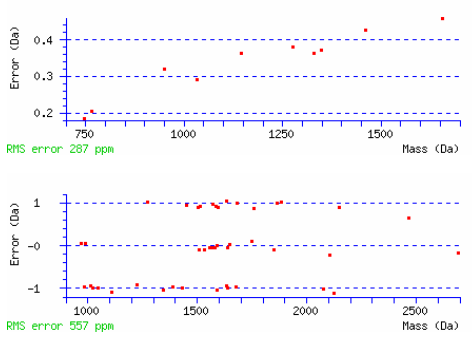
This is the error window on MS/MS fragment mass values.

Specifying too tight or too loose a mass tolerance will reduce the ions score

Practical Tips

Make a reasonable estimate of mass error

- Don't just guess, run a standard



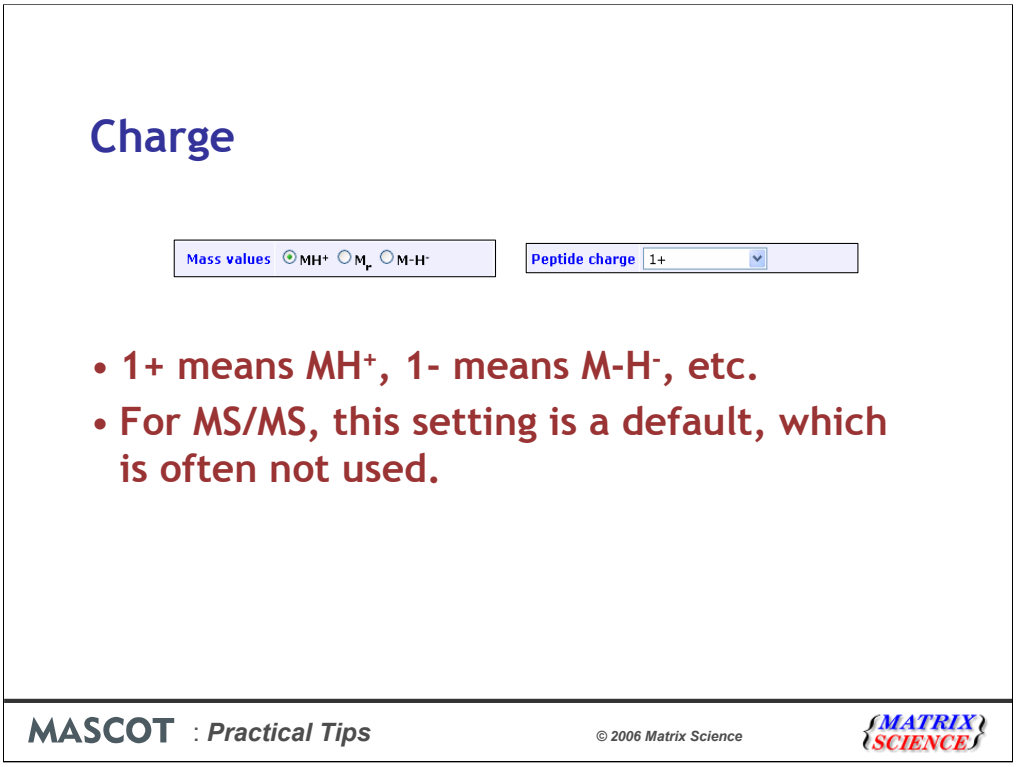
MASCOT : *Practical Tips* MATRIX SCIENCE

Making an estimate of the mass accuracy doesn't have to be a guessing game. Just search a standard digest 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.

The graph for the peptide mass errors is on the Protein View report. The graph for fragment ion errors is on the Peptide View.

This first example is from a search where the tolerance was set to 0.5 Da. It looks like we might be clipping at the high mass end, so try a search with the tolerance opened up to (say) 0.7 Da and see whether this brings in additional matches.

On the other hand, the second example shows a tolerance that is much too wide. The errors are bunched around 0 Da and +/- 1 Da. In some cases, this will be because the wrong isotope peak is being matched. Setting the tolerance much too wide can lead to scoring artefacts. If these were fragment ion errors, or peptide mass errors in a PMF, then the score will be lower than it should be. If these were peptide mass errors in a search of MS/MS data, then the specificity of the search and the search time will suffer.



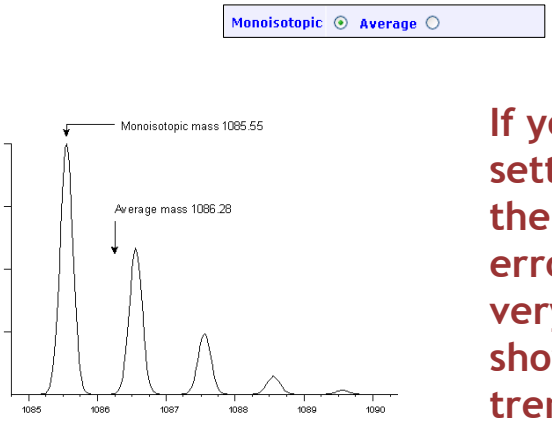
These fields are used to specify the peptide charge state. The radio buttons are from the peptide mass fingerprint form. The drop down list is used on the sequence query and MS/MS forms.

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.

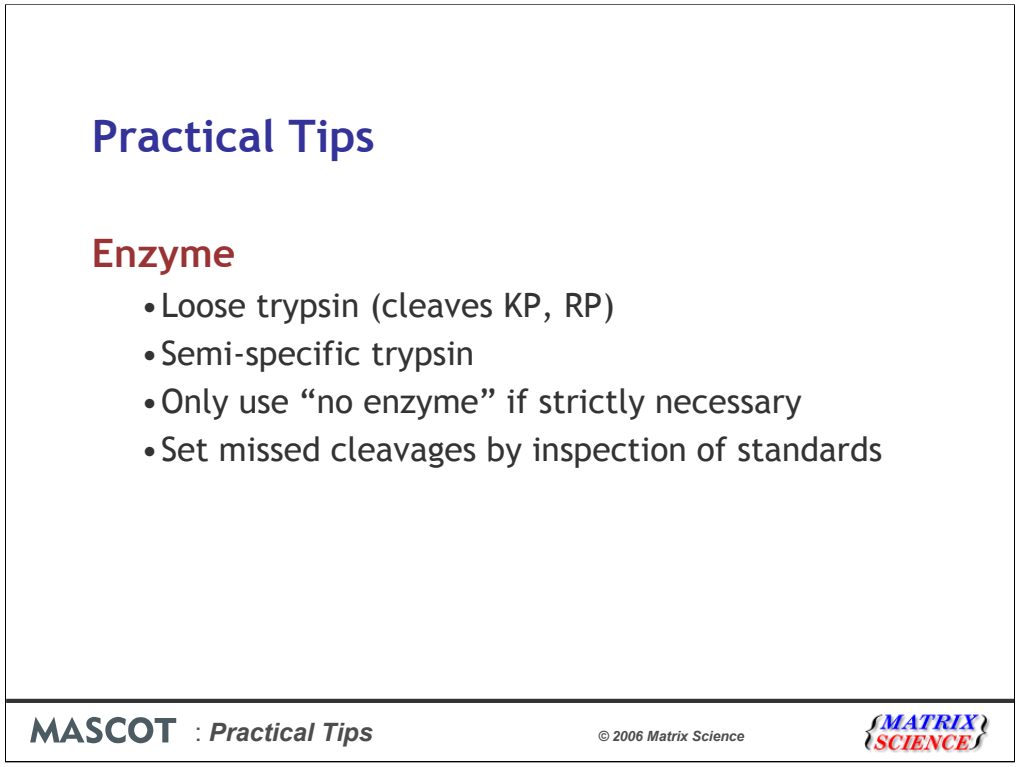
Mass type



If you get this setting wrong, the mass errors will be very large and show a strong trend

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.

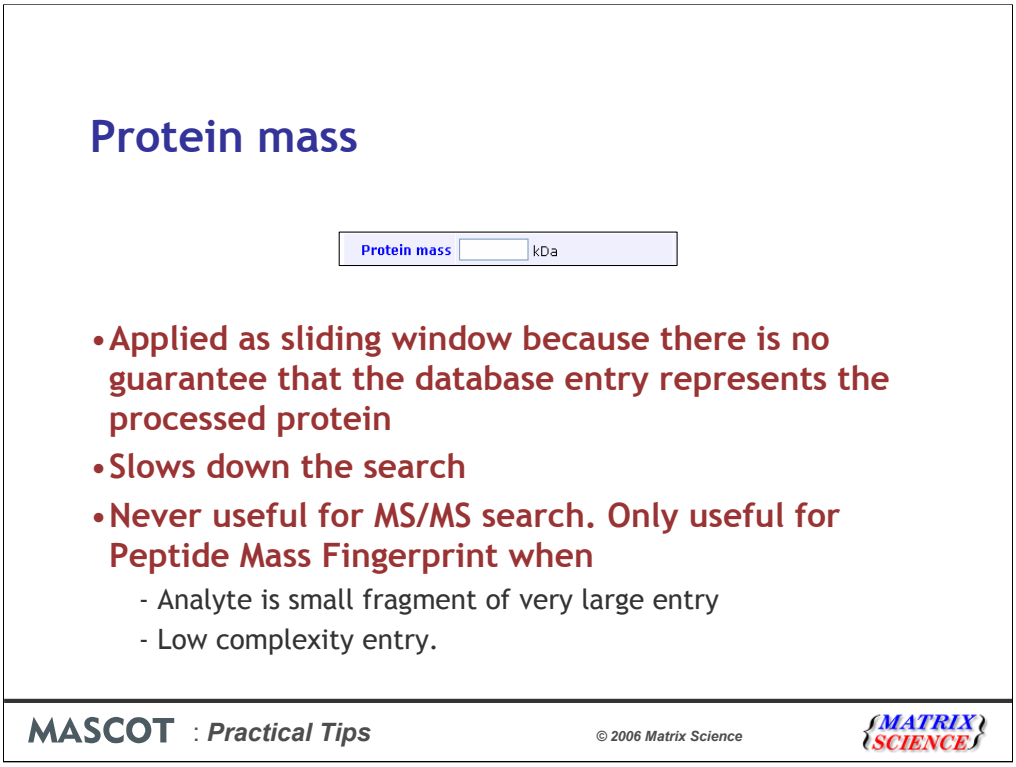
Most modern instruments produce monoisotopic mass values. You will only have an average mass if the entire isotope distribution has been centroided into a single peak, which usually implies very low resolution. If you get this setting wrong, the mass errors will be very large and show a strong trend,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



The vast majority of searches are of tryptic digests. In such cases, I normally choose loose trypsin, which cuts after K or R even when the next residue is P. If there is evidence for a high level of 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.

The missed cleavage parameter should be set by looking at 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. Setting the number of allowed missed cleavage sites to zero simulates a limit digest.

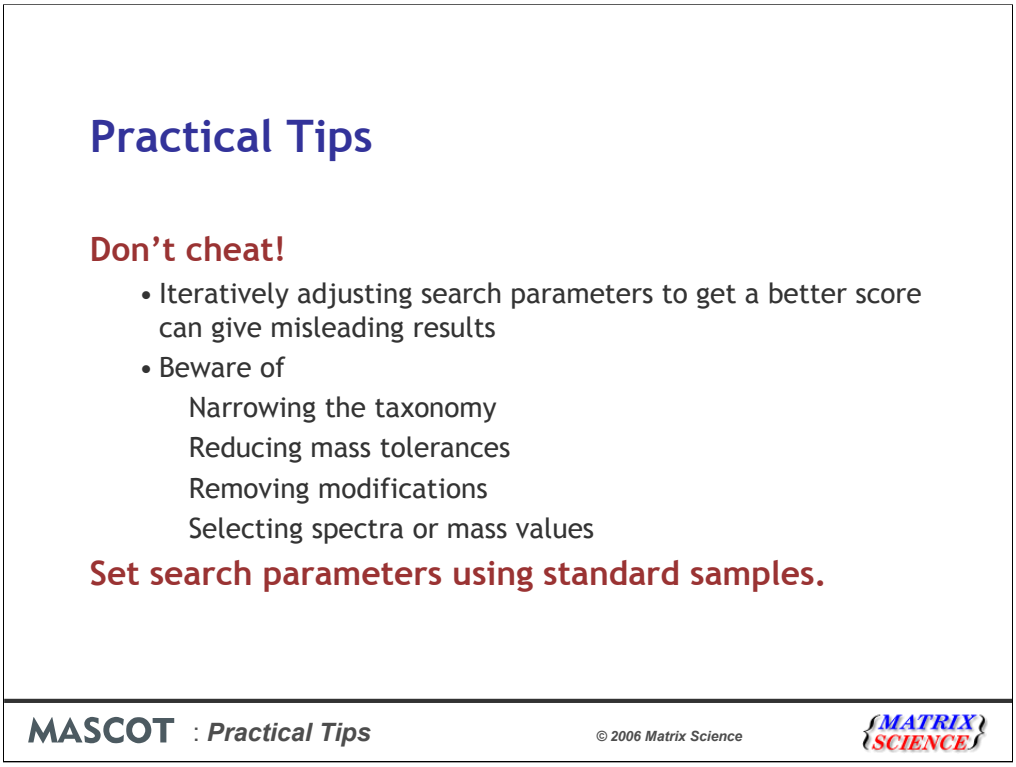
If you are confident that your digest is perfect, with no partial fragments present, this will give maximum discrimination and the highest score for a peptide mass fingerprint. If experience shows that your digest mixtures usually include some partials, that is, peptides with missed cleavage sites, you 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.



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



Don't cheat!

It is easy to distort the search results without realising.

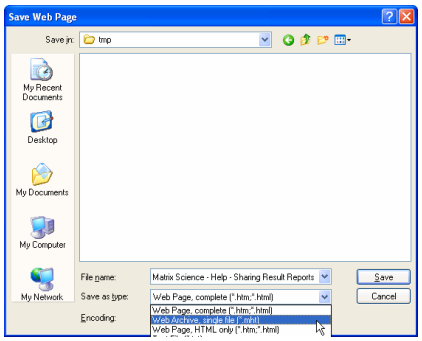
Basically, it is risky to adjust the search parameters interactively to get a better score for an unknown.

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.

How can I send a result report to a colleague?

Save a single report as web page complete or web archive



MASCOT : Practical Tips MATRIX SCIENCE

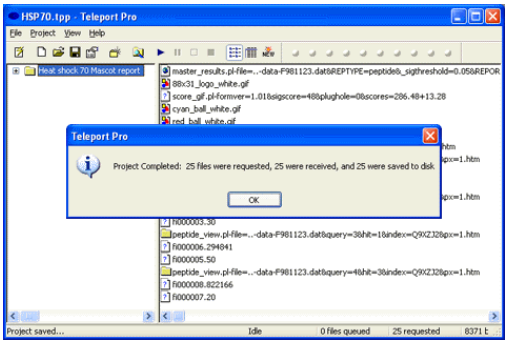
There are a number of options for sharing search results.

One is to save the single report page as a 'Web page complete' or 'web archive'. Saving as "Web page, HTML only" is no good because graphics like the score histogram will be missing. This will only contain the page that you save – the linked pages for the protein and peptide view reports will not be saved.

How can I send a result report to a colleague?

Print a single report to an Acrobat PDF file

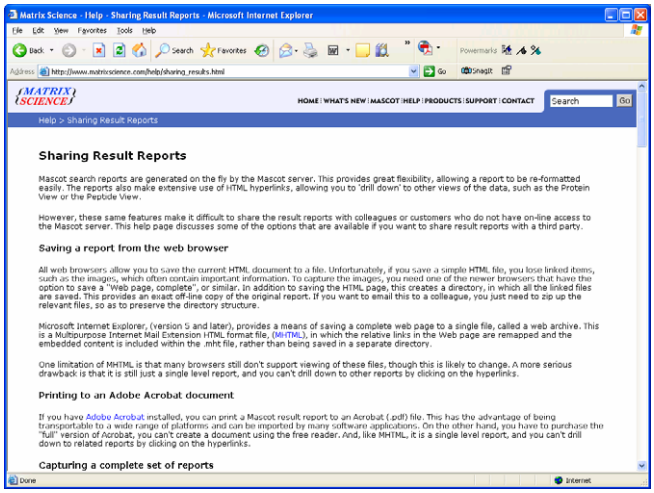
Capture a complete set of reports using an off-line browser utility



Another option is to print the single report to an Acrobat PDF file. As with saving the report as a web archive, only the selected page will be included in the file.

The most comprehensive solution is to capture a complete set of result reports for off-line viewing using any web browser. There are many tools, including freeware and shareware options, available for this purpose. Tucows (www.tucows.com) has listings of freeware and shareware for Windows, Linux, and Mac.

How can I send a result report to a colleague?



MASCOT : *Practical Tips*

A detailed answer to this question can be found on the help page at http://www.matrixscience.com/help/sharing_results.html