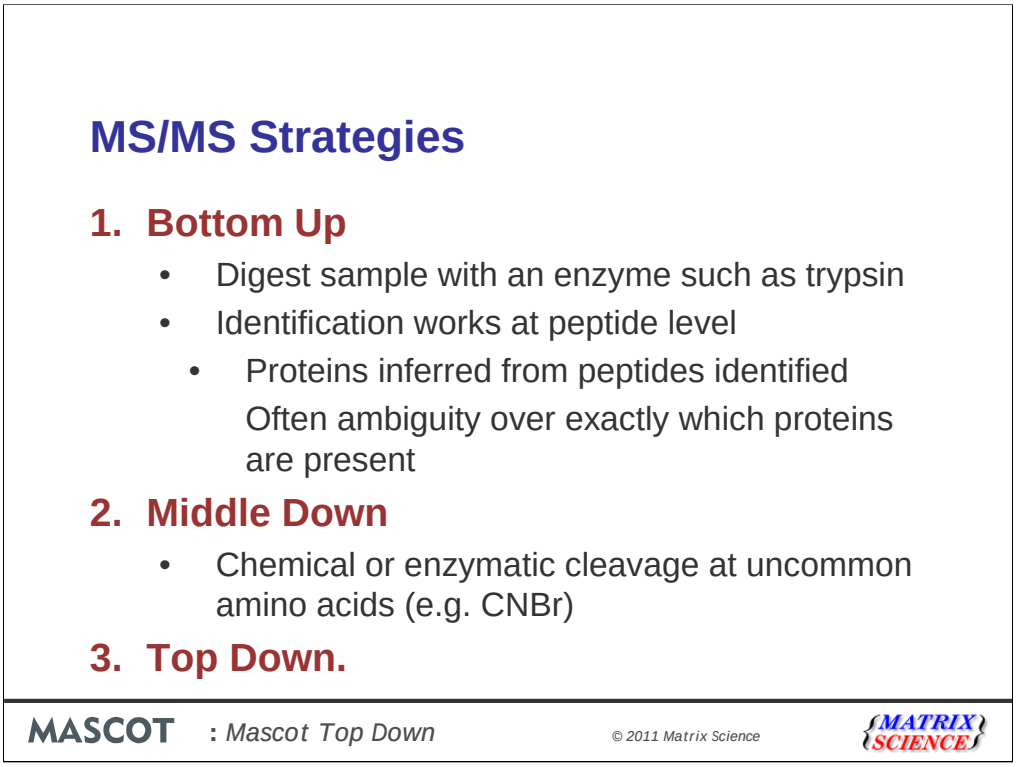
Mascot Top Down

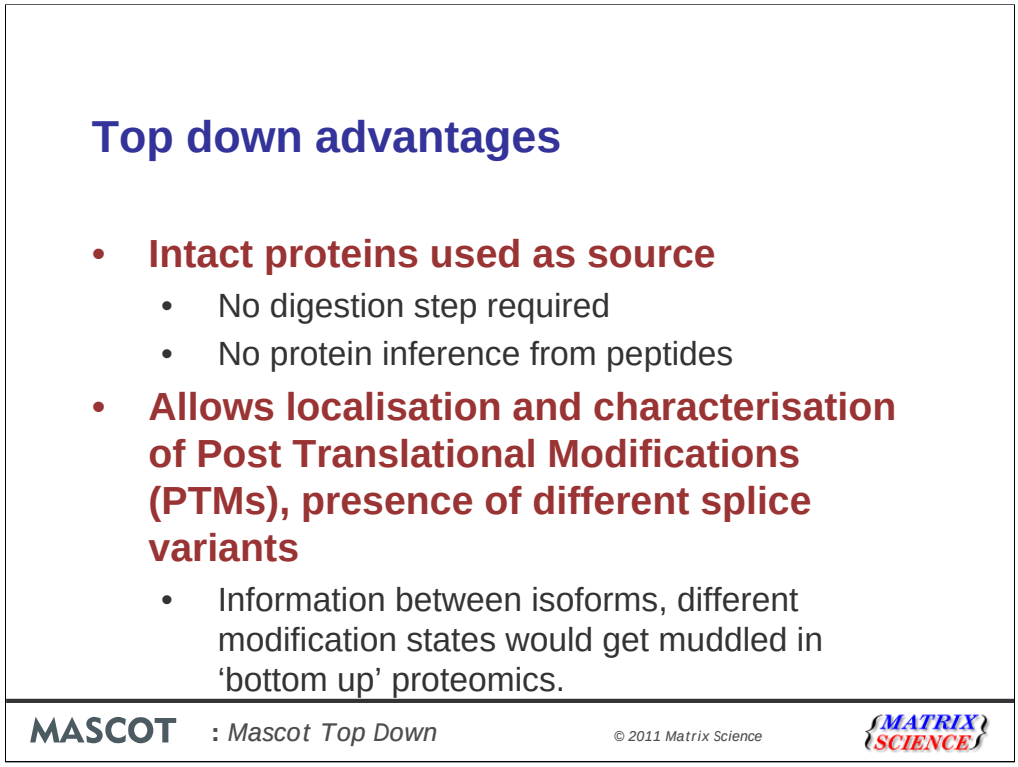
AKA “Big Mascot”



The standard approach to protein identification using MS/MS data is the bottom up approach where the target proteins are digested into relatively small fragments and MS/MS and subsequent sequence identification is of individual peptides. The bottom up approach has proven to be a very successful, robust strategy, allowing for a high throughput. The main disadvantages of this approach are that you are identifying the individual peptide sequences and the proteins present in your mixture are then inferred from the peptides identified. Because peptides are shared between homologous proteins, there is often ambiguity concerning which proteins are present.

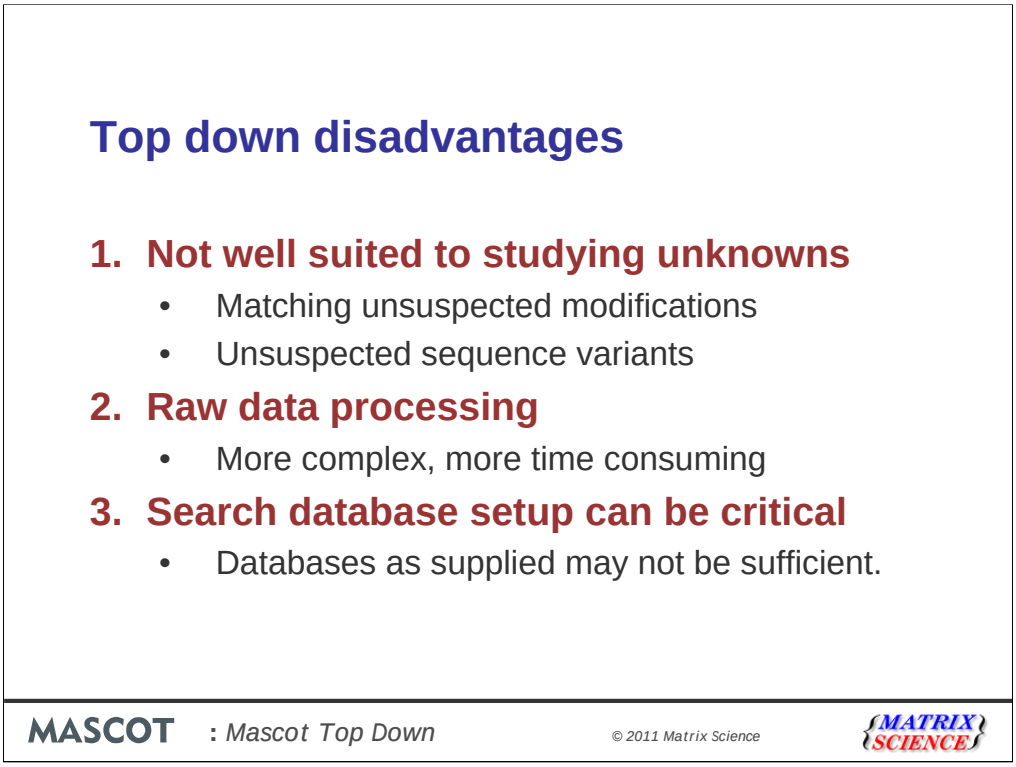
Middle down approaches are intermediate approaches where the protease used to digest the source proteins cuts at less common amino acids, resulting in longer peptides being generated.

In a top down experiment, MS/MS is carried out on the intact protein.



The main advantage of a top down approach is that you're using intact proteins as your source sample in the mass spectrometer. This means that you don't need to carry out a, potentially time consuming, digestion step (with associated purification steps). It also means that there are no issues with protein inference – in bottom up proteomics you're inferring the presence of proteins from the analysed peptides which can be a tricky step to get right, especially when considering peptides which are found in multiple proteins, or peptides shared between different isoforms of the same protein.

This allows you to characterise post translational modifications at a protein level and the presence of different splice variants in a more co-ordinated fashion – information which could easily get muddled up when carrying out bottom up analysis. Therefore, Top Down is a powerful approach if you are studying the presence of different isoforms of a protein, or of differential post translation modification of proteins – information that would be lost in a bottom up experiment.

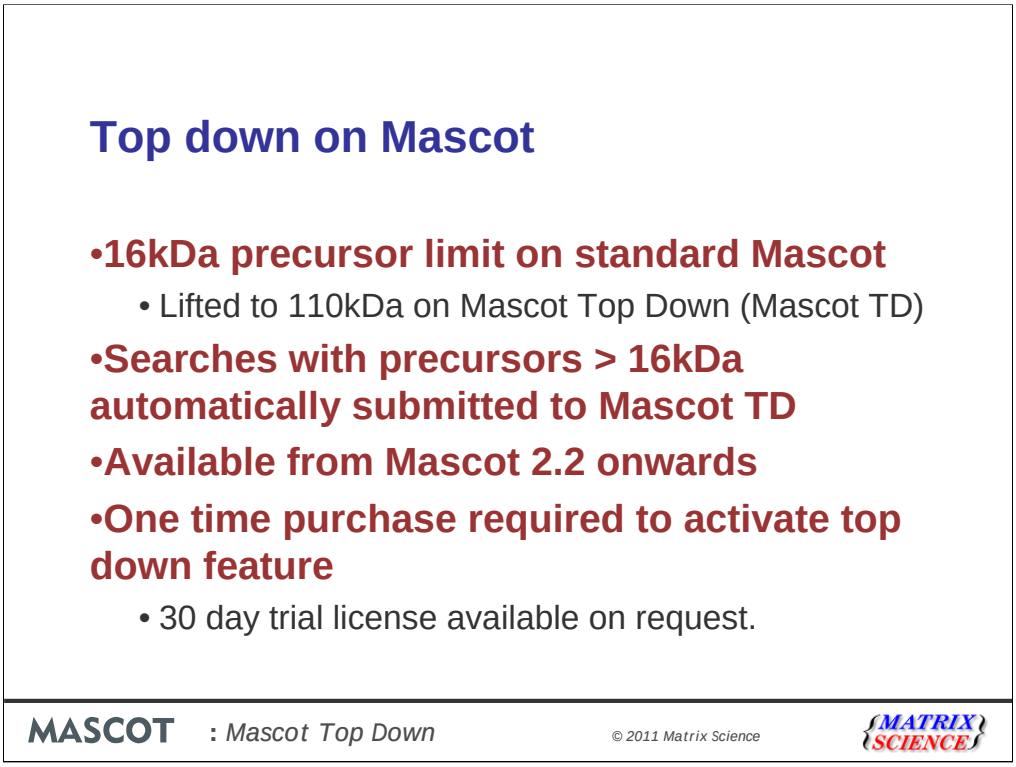


However, top down approaches have a number of disadvantages.

For a variety of reasons, Top Down approaches tend not to be well suited to studying unknown proteins since unsuspected modifications or sequence variants can stop you from getting a match to the protein and the work involved in then identifying the cause of the failure can be time consuming. If you're doing a bottom up approach you may lose a certain percentage of your peptide matches because of unsuspected chemical or post-translational modifications but you'll normally still be able to get matches to the proteins in your sample because you'll get matches from other, non-modified, peptides. However, in a top-down experiment you'll often only have a single spectrum for a protein and an unsuspected modification could well prevent you from getting a match.

Because you have to select high charge state precursors in order to work on a precursor with an m/z within the mass range of the analyser, the raw MS/MS data from a top down dataset will typically have many different charge state series within it, often with overlapping envelopes and shared peaks. Deconvoluting the peaklist is therefore normally required and this can be a time consuming step.

Searching the standard sequence databases with top-down data is probably not a good idea because your data may represent a mutant or splice variant not represented by default in the database. So some effort is needed to build up a database containing all the required sequences or you could miss a large percentage of your matches, so you'll probably need to build up a database containing splice variants and known mutants of the sequences of interest.



Standard Mascot has a precursor ion limit of 16kDa. Mascot Top Down lifts this to 110kDa. If you have Mascot Top Down licensed, then you use your Mascot server in the normal way, but searches with precursor masses of greater than 16kDa are automatically submitted to the Mascot Top Down executable. In order to enable Mascot Top Down you require a new license and a 30 day trial license is available from us on request.

Peak list preparation

•Mascot only searches for 1+ and 2+ fragments

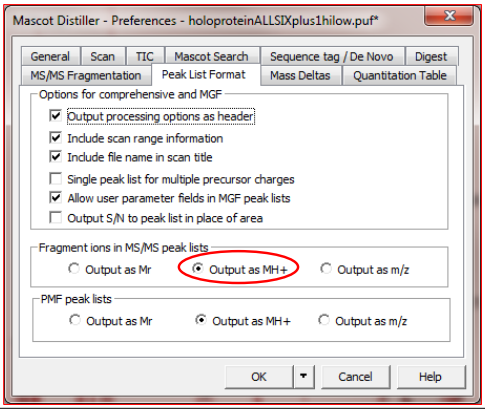
- Most ECD and FTICR instruments will produce fragments with much higher charge states
- It is therefore important to 'decharge' the fragment masses.

Mascot Server will only try and match singly or doubly charged fragment ions, but because you normally have to select a precursor with a high charge state for top down work, most ECD and FTICR instruments will produce fragments with much higher charge states. It is therefore important to 'decharge' the fragment masses.

Peaklist preparation

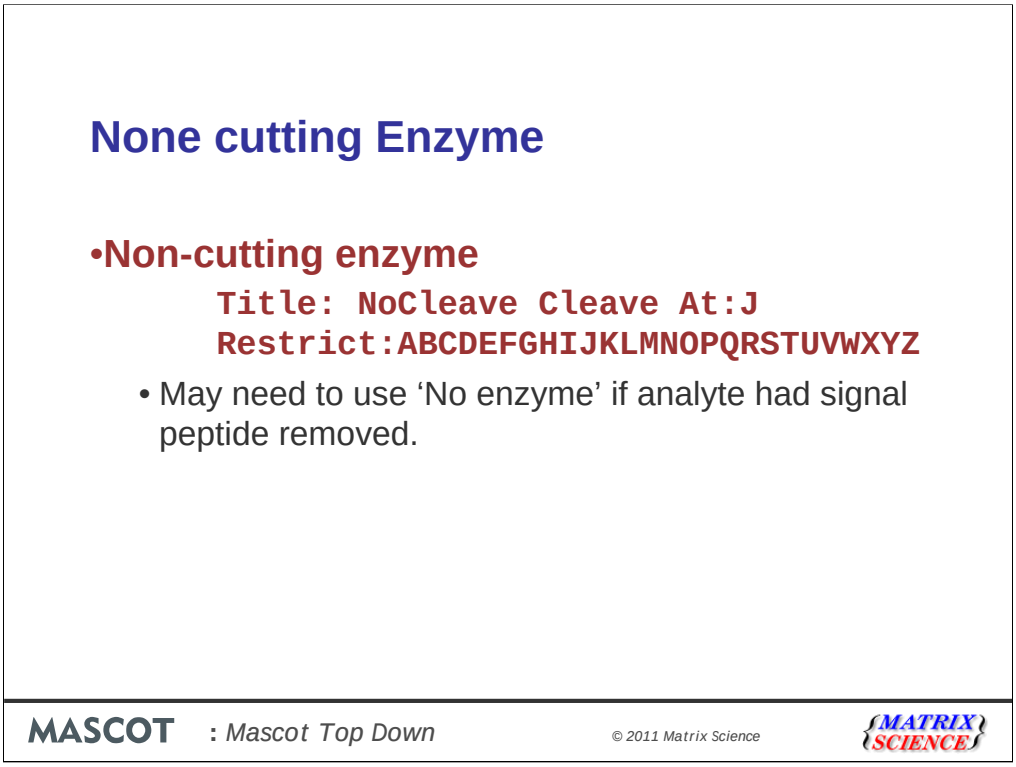
•Mascot Distiller:

- Tools->Preferences->Peak List Format tab
- Fragment ions in MS/MS peaklists->Output as MH+.



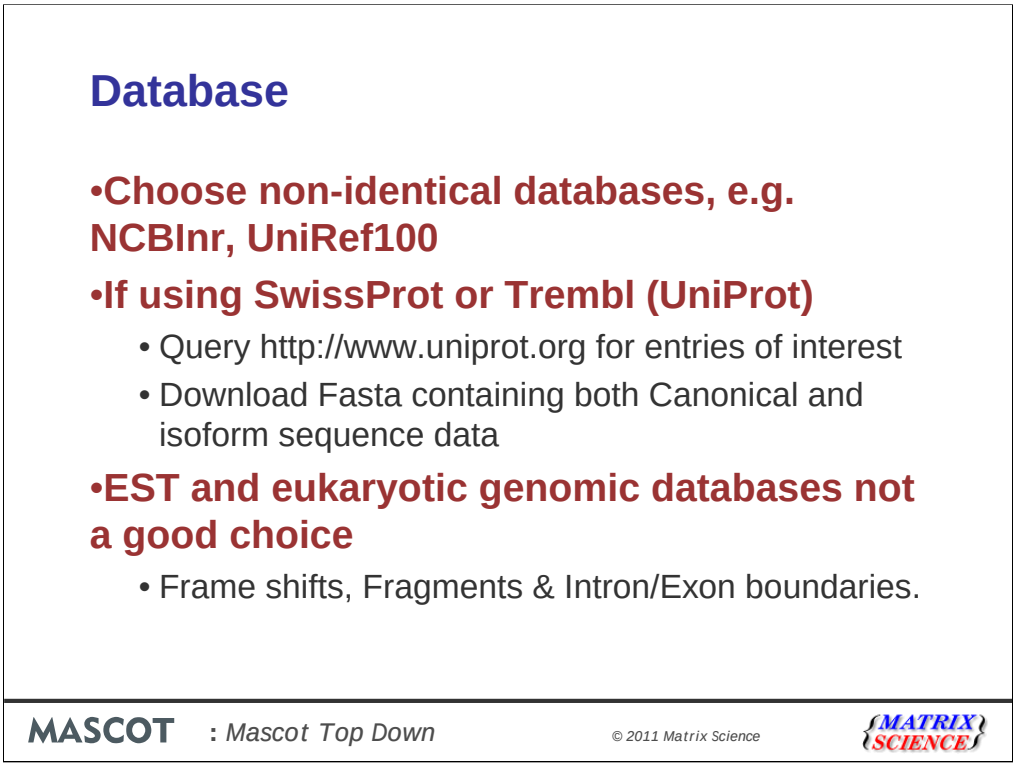
If you are using Mascot Distiller to produce the peak list and perform the search then, on the Tools menu, select Preferences and click on the *Peak List Format* tab. Select the *Output as MH+* for the *Fragment ions in MS/MS peak lists* option. To make this the default, choose *Save as defaults* from the down arrow next to the OK button.

If you are preparing your peak list in this way, you may wish to create a special instrument definition with only 1+ fragments.



In an ideal world, the analyte for a top down experiment will be identical to the sequence entry in the database. An enzyme must be specified when performing a Mascot search, so it is necessary to specify a 'non cutting' enzyme. Using the configuration utility available from the home page on your local Mascot server, select the enzymes link, add a new enzyme and, for example add in a NoCleave definition such as the one shown.

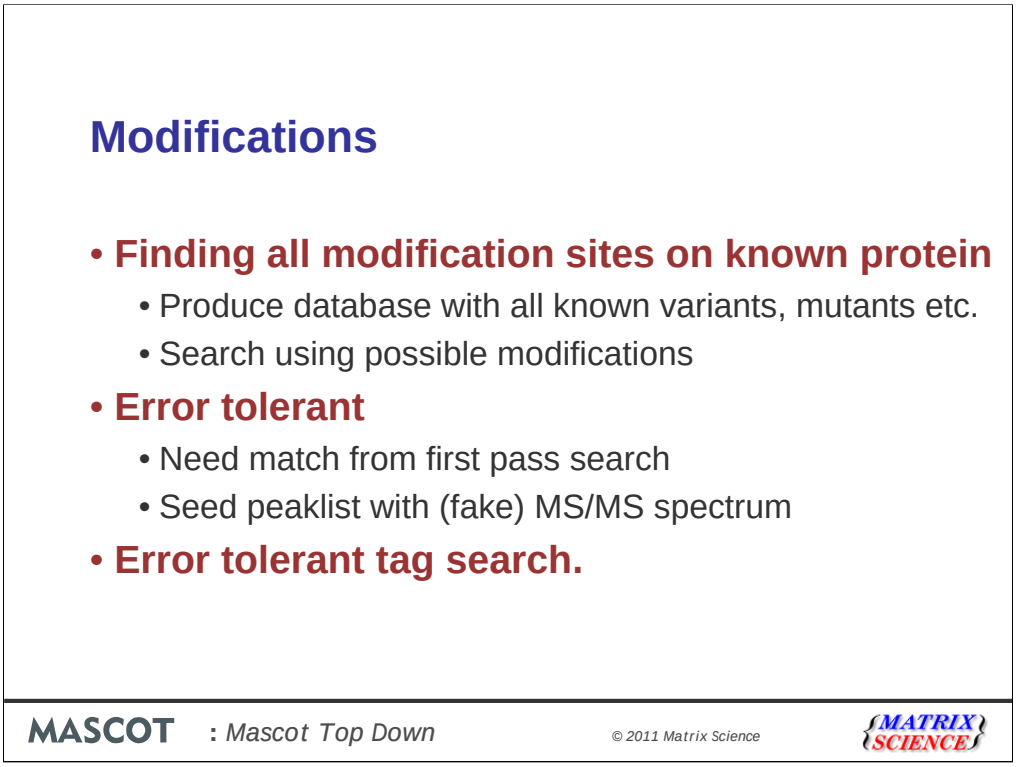
Unfortunately, the sequence in the database may not be identical to the sample. For example, the database sequences may contain signal peptides. In this case, the only alternative is to search using no enzyme, which is of course somewhat slower.



If there is a large discrepancy between the sequence database entry and your analyte – for example due to variable splicing – you may well fail to get a match from a top down search. For this reason, a non-redundant database such as SwissProt is not a good choice for carrying out top-down database searching. Non-identical databases, such as NCBInr and UNIREF100, have a wider range of sequences explicitly represented and would be a better choice.

Normally you know the proteins you are working on when carrying out top down work and are interested in seeing which isoforms, or what modification patterns, are present. SwissProt contains annotation information which includes primary sequence variants (such as splice variants, residue substitutions), and information about sequence regions that are usually removed from the mature protein (such as signal peptides and propeptides). Therefore, another option would be to download the sequences of interest from SwissProt and choose to include isoforms.

EST and unprocessed genomic DNA databases are not good choices for Top Down searching of course. Many EST entries will contain sequencing errors which lead to frame shifts, many only represent fragments of proteins. Genomic databases often contain intron/exon boundaries. These can lead you to missing many peptide matches in a bottom up search, but will stop you getting any matches at all from a top down dataset.

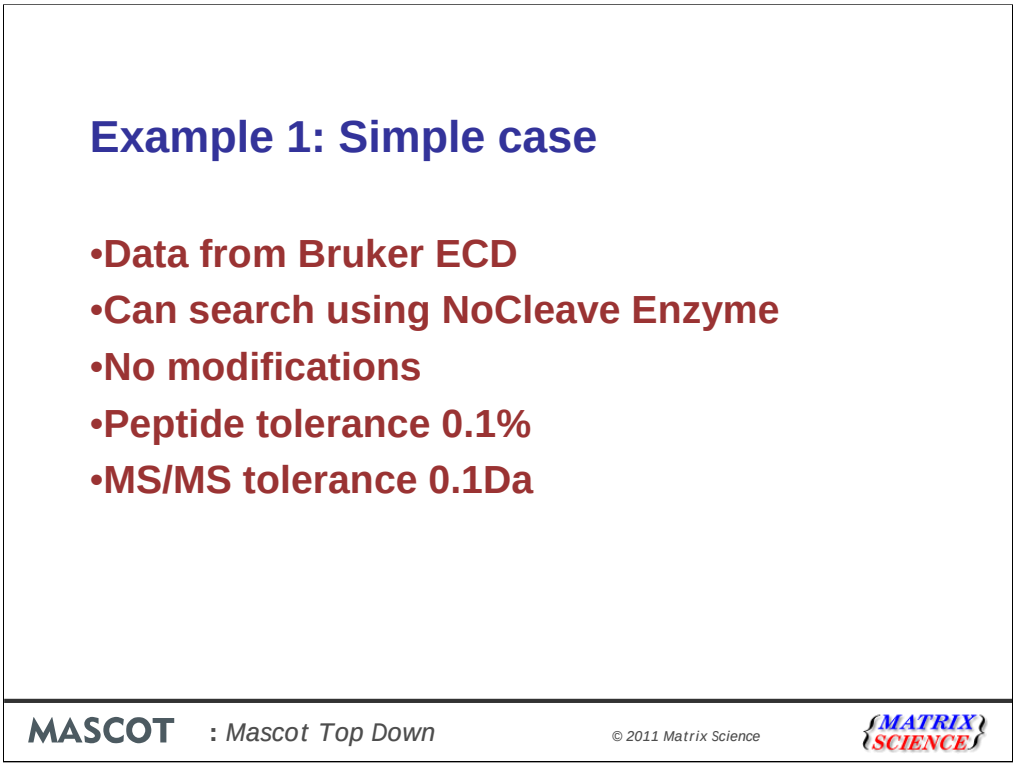


Identifying all modified sites in a known protein is a common reason for carrying out a top down experiment. One strategy for doing this is to produce a database with all known mutant, splice variant etc. forms of the sequences of interest and then to search this with a number of possible modifications – for example, you may know or suspect that the protein is phosphorylated , N-terminally acetylated etc and also allow for common processing modifications. As with a standard Mascot search, you are limited to 9 variable modifications in this standard search, but that would normally be more than sufficient. You should really only take this approach with modifications you have reason to suspect are present though, not a wide range of unlikely modifications just to get a match.

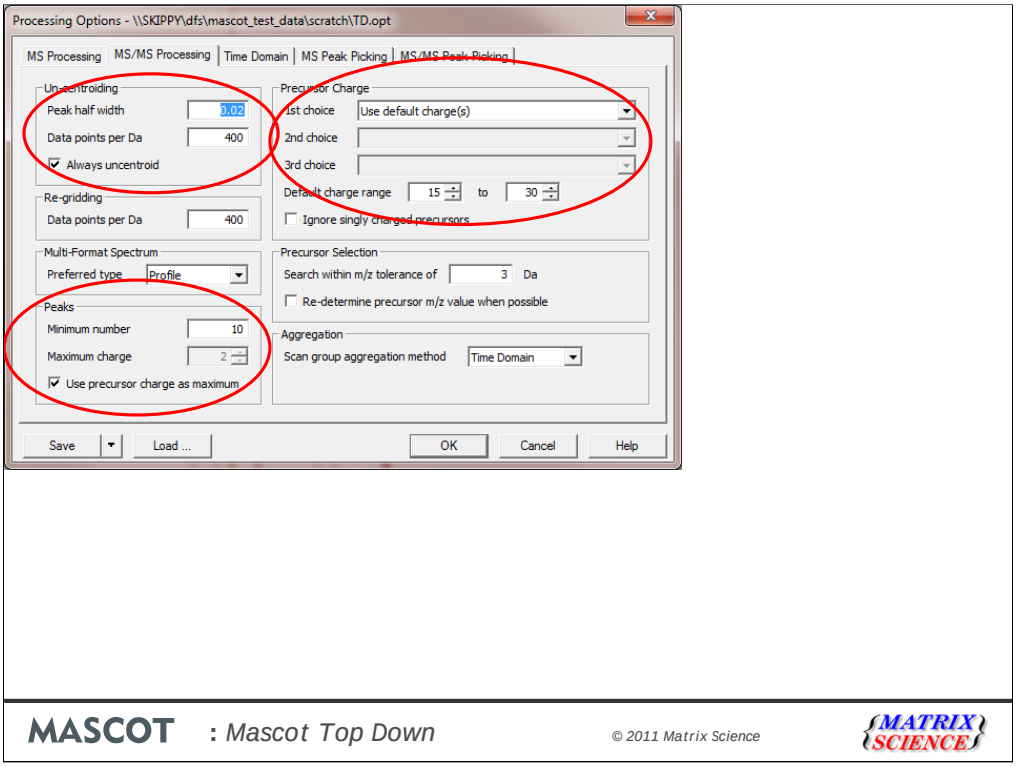
The error tolerant search iterates through a large number of possible modifications and is therefore well suited to top down searches where you are looking for modifications. However this only considers proteins with at least one significant peptide match from the standard, first pass, search. If you know which protein you are analysing you could seed your peaklist with a single ms/ms spectrum that matches the protein to trigger off the second pass search. This seed peaklist can be a fake spectrum, of course.

If these methods don't produce believable results, you can carry out error tolerant tag searches using the NoCleave enzyme. This allows the mass value to 'float' enabling you to get a match when you have an unknown modification or mutation.

Both the error tolerant and the error tolerant tag search are only going to find a match if you have one unsuspected modification or sequence change. Getting a match from Mascot TD where you have multiple changes, say multiple unsuspected modifications or SNPs is going to be difficult.

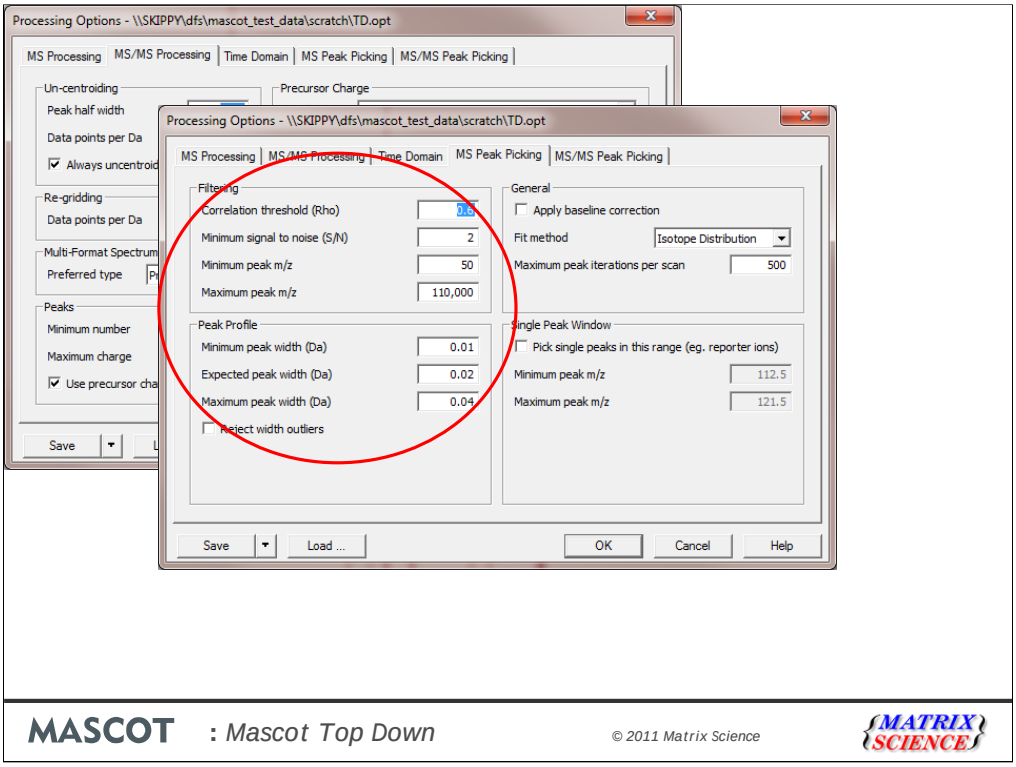


Lets take a look at some example searches and results from top down. Our first example is using a dataset supplied by Bruker. In this case, it has been possible to search using the NoCleave enzyme, using a reasonable peptide tolerance and a tight ms-ms tolerance and no modifications.

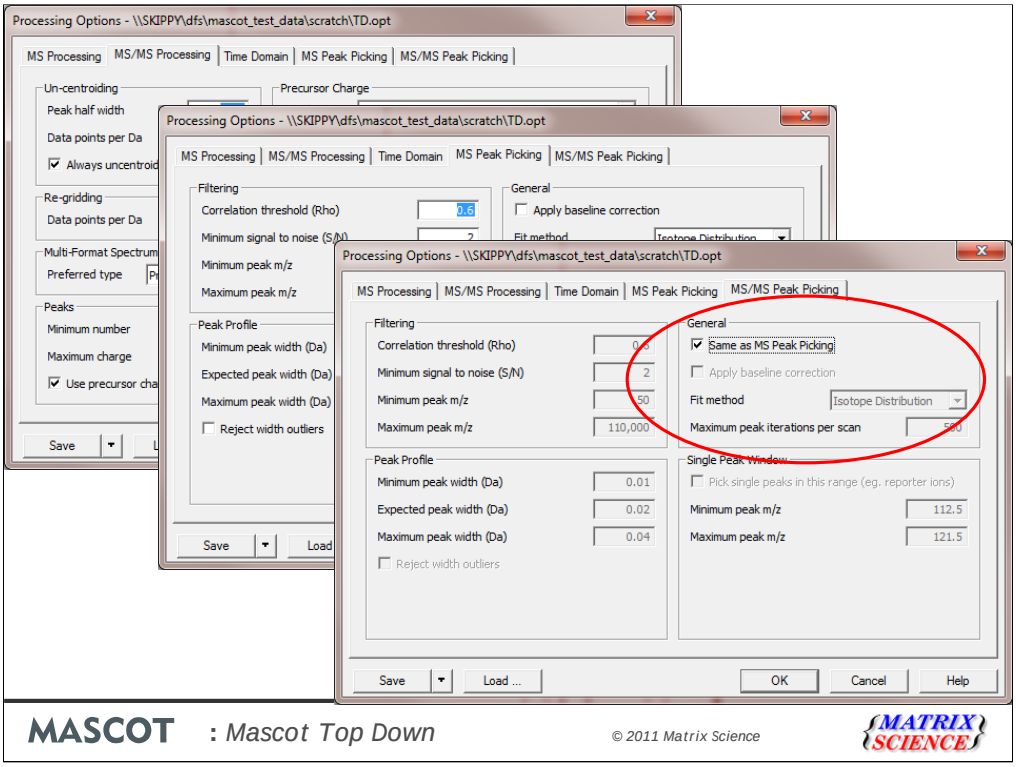


Peak detection for a top down dataset can be quite tricky to get right and you'll probably need to work on the parameters for each dataset, even if you're using Mascot Distiller. I'm just going to run through some of the critical settings to look at in Distiller – please note these are not suggested value for any particular dataset. These are just some of the parameters you may need to look at and change when processing a top down raw data file:

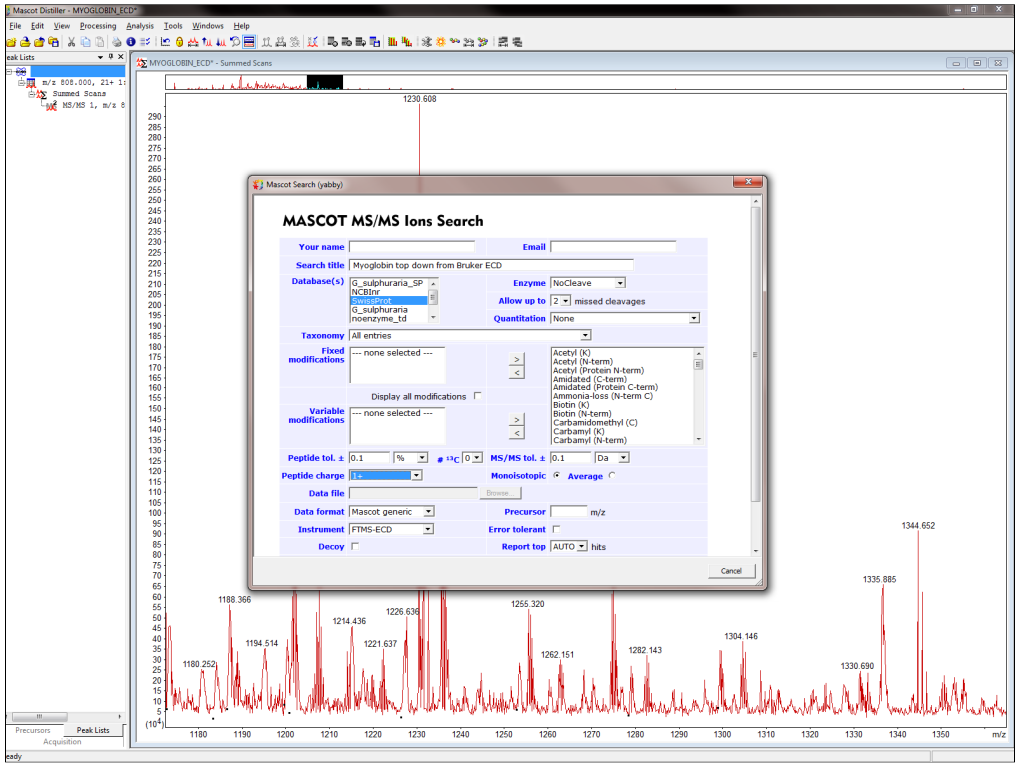
On the MS/MS Processing tab, if you have centroided data make sure you have 'Always uncentroid' checkbox checked, set the 'Peak half width' to something like 0.02 and the 'Data points per Da' to say 400. Specify a default precursor charge range, say 15+ to 30+, and on the Peaks section make sure the 'Use precursor charge' as maximum checkbox is checked.



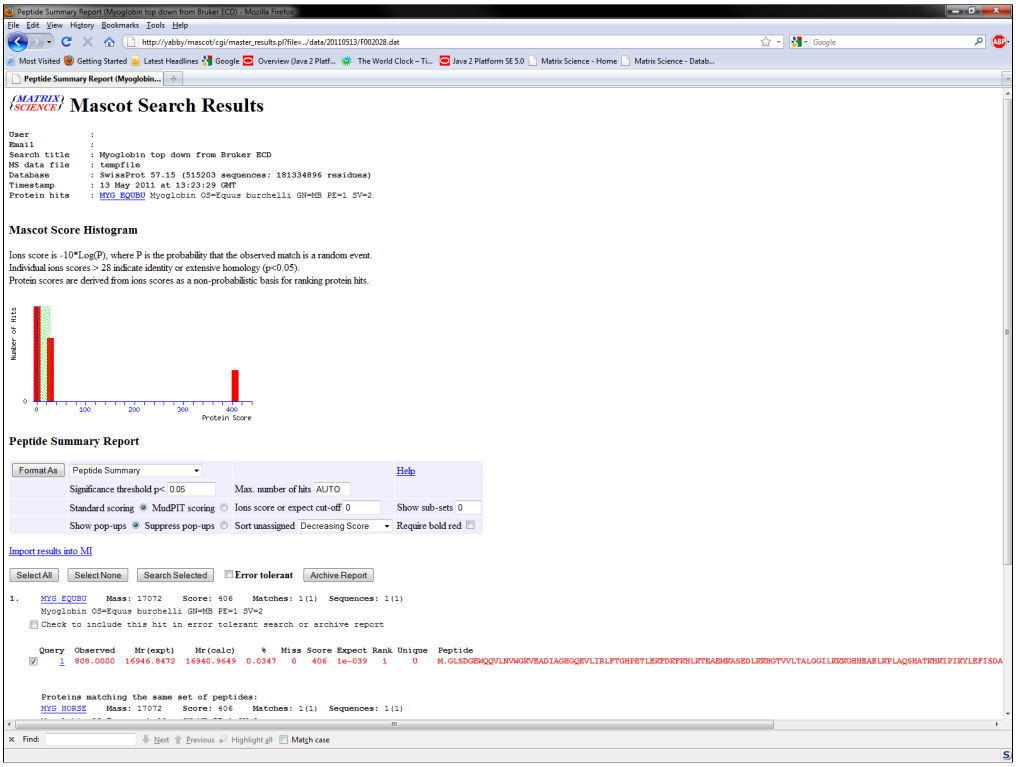
On the MS Peak Picking tab, you may want to slightly decrease the Correlation threshold from 0.7, say to 0.6 and increase the maximum peak m/z to 110,000. If you've uncentroided a centroided dataset change the expected peak width to the same value and alter the minimum and maximum peak widths accordingly.



Finally on the MS/MS Peak Picking tab, make sure you have the ‘Same as MS Peak Picking’ checkbox checked.



We submit the search to Mascot as usual

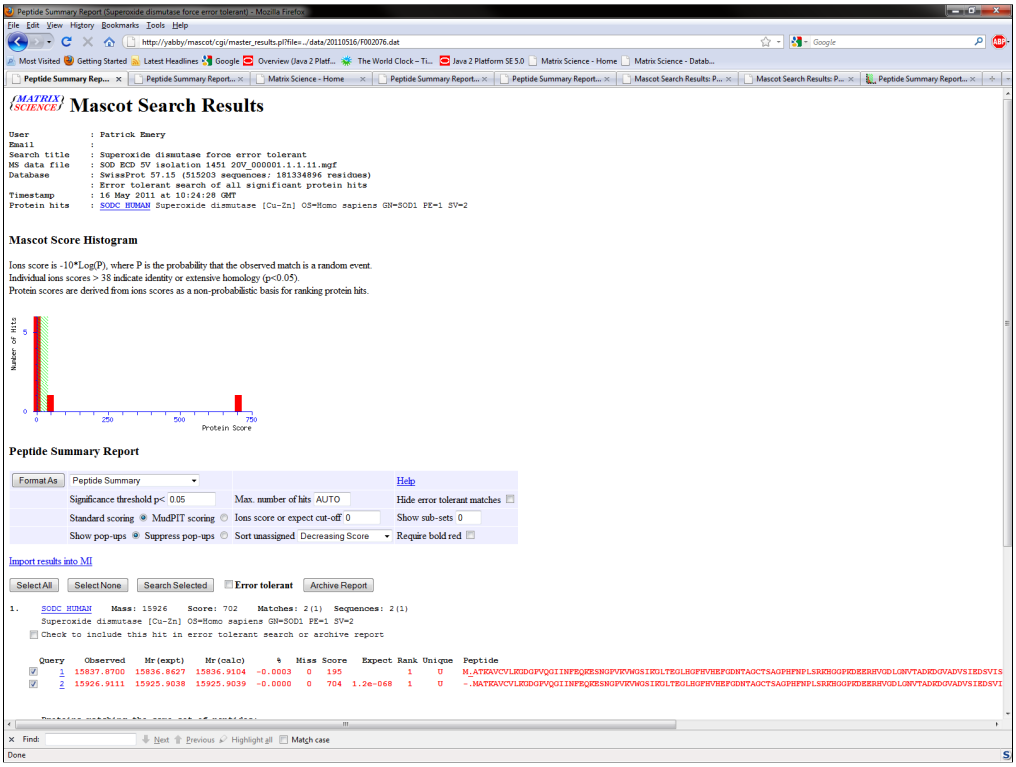


And here are the results – a match to Myoglobin. Just like a standard peptide summary report for a match to a single peptide, although the score is somewhat higher.

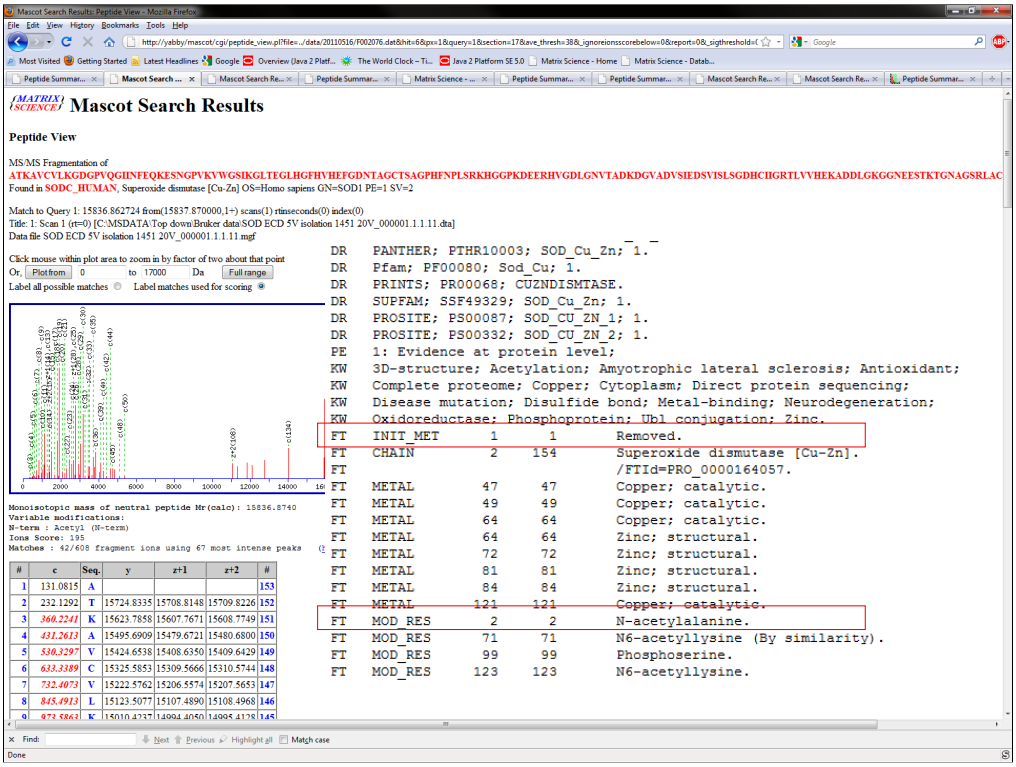
Example 2: Forcing error tolerant search

- **Source protein Superoxide dismutase**
  - No match from NoCleave or No enzyme
- **Error tolerant requires first pass match**
  - Working on a known, add a spectrum which will match in the 1st pass search
  - Can be fake spectrum
  - Mascot Parser has methods to generate predicted fragments
    - C++, PERL and Java.

In our second example, we're working on a known protein – in this case human Superoxide dismutase. However, neither a NoCleave nor No Enzyme search produces a match from SwissProt, possibly because of an unsuspected modification. One method of identifying modifications and sequence variants like this in Mascot is to carry out the two pass Error tolerant search, which attempts to identify unsuspected modifications and primary sequence variants by testing a long list of modifications against protein hits identified in a standard, first pass, search. To use this, we need to get a match to our protein of interest in the first pass search – so we need to include a spectrum guaranteed to get us a match included in the peaklist. Because we're working on a known protein we can do this – there are various tools available for generating theoretical spectra based on a given protein/peptide sequence. If you wish to write something yourself to generate the fake spectrum then our free Mascot results library, Mascot Parser, includes methods which allow you to do this. You can access Parser from C++, PERL and Java. For this example, I wrote a small piece of Java code to output a fake spectrum for Superoxide dismutase and included it in an MGF peaklist file with the experimental data and then submitted the peaklist file to a Mascot error tolerant search.

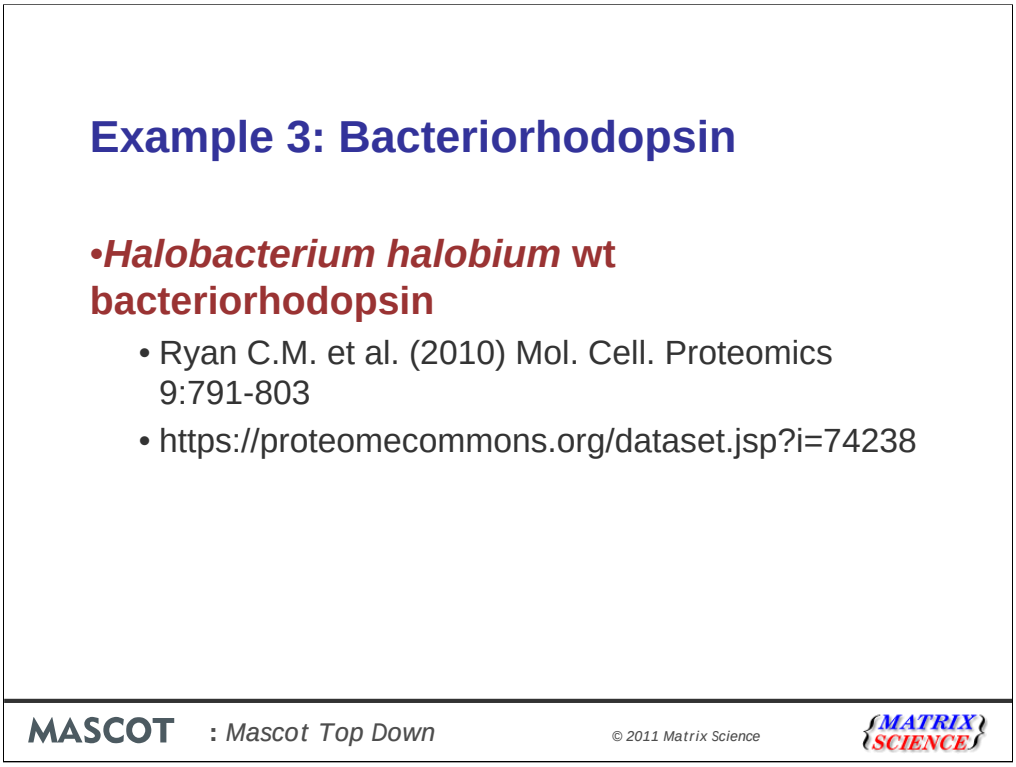


Ignore the match to query 2 – that is the match to the fake spectrum which enabled us to carry out the error tolerant pass. Query 1 is our experimental spectrum and the Mascot error tolerant search managed to get a match to it by allowing an n-terminal modification.

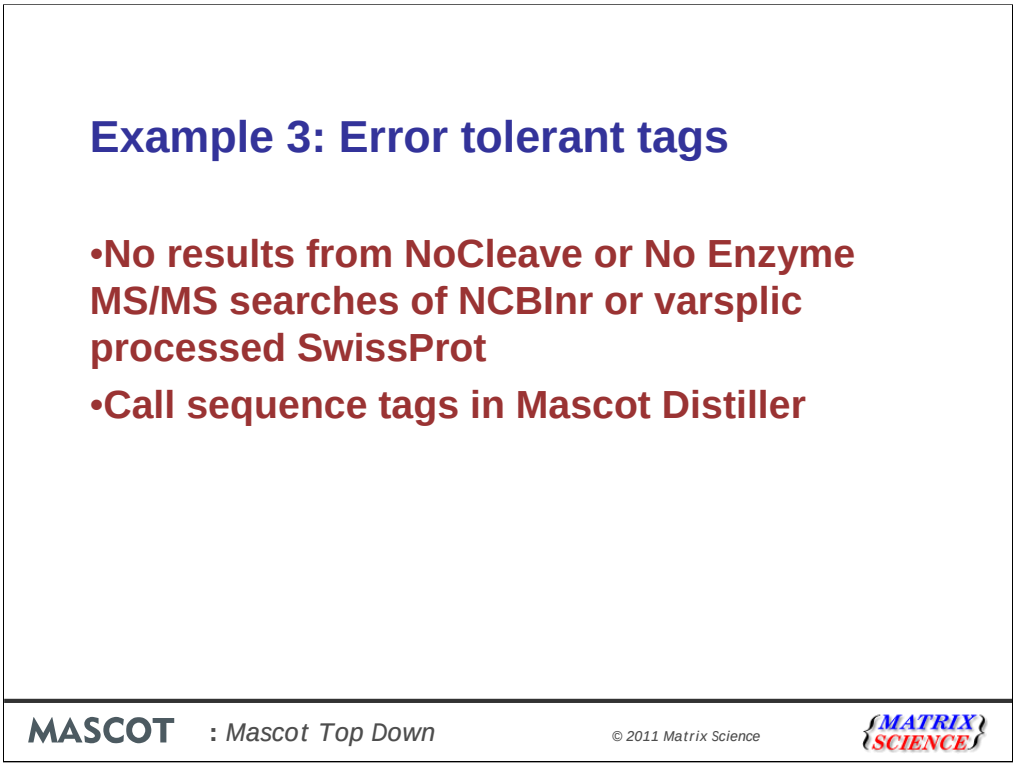


#	c	Seq.	y	z+1	z+2	#
1	131.0815	A				153
2	232.1292	T	15724.8335	15708.8148	15709.8226	152
3	360.2241	K	15623.7858	15607.7671	15608.7749	151
4	431.2613	A	15495.6909	15479.6721	15480.6800	150
5	530.3297	V	15424.6538	15408.6350	15409.6429	149
6	633.3389	C	15325.5853	15309.5666	15310.5744	148
7	732.4073	V	15222.5762	15206.5574	15207.5653	147
8	845.4913	L	15123.5077	15107.4890	15108.4968	146
9	972.5863	K	15010.4237	14994.4050	14995.4128	145

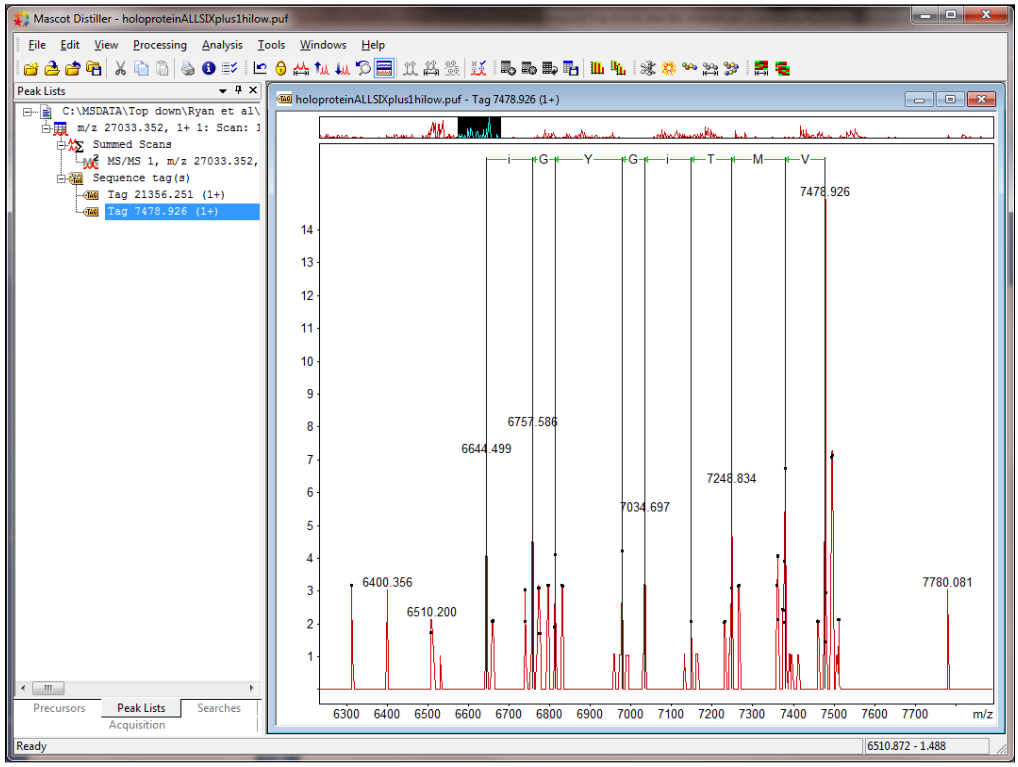
The error tolerant match identified a number of possibilities for this, but the match with the lowest precursor delta is N-terminal Acetylation. N-terminal acetylation is a common post-translational modification after loss of the initiator methionine, as we have in this match, and this assignment is supported by the SwissProt annotation for the protein.



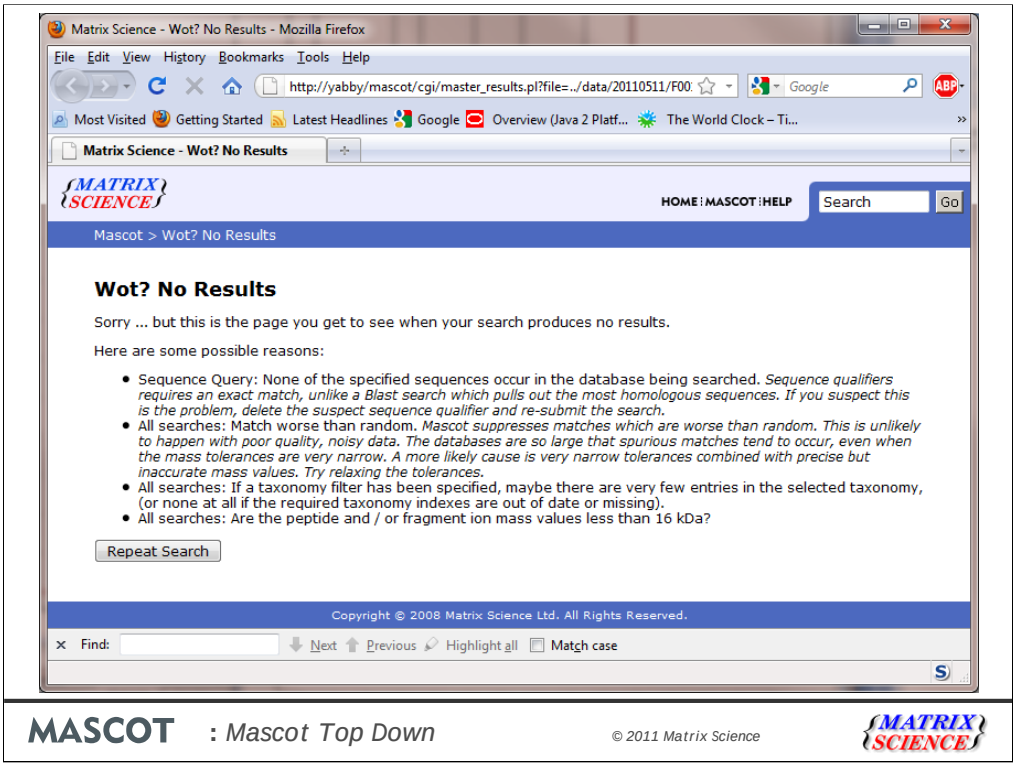
The next example is top-down analysis of bacteriorhodopsin taken from Ryan *et al.* (2010) Raw data and peaklists from the paper can be downloaded from Proteome Commons.



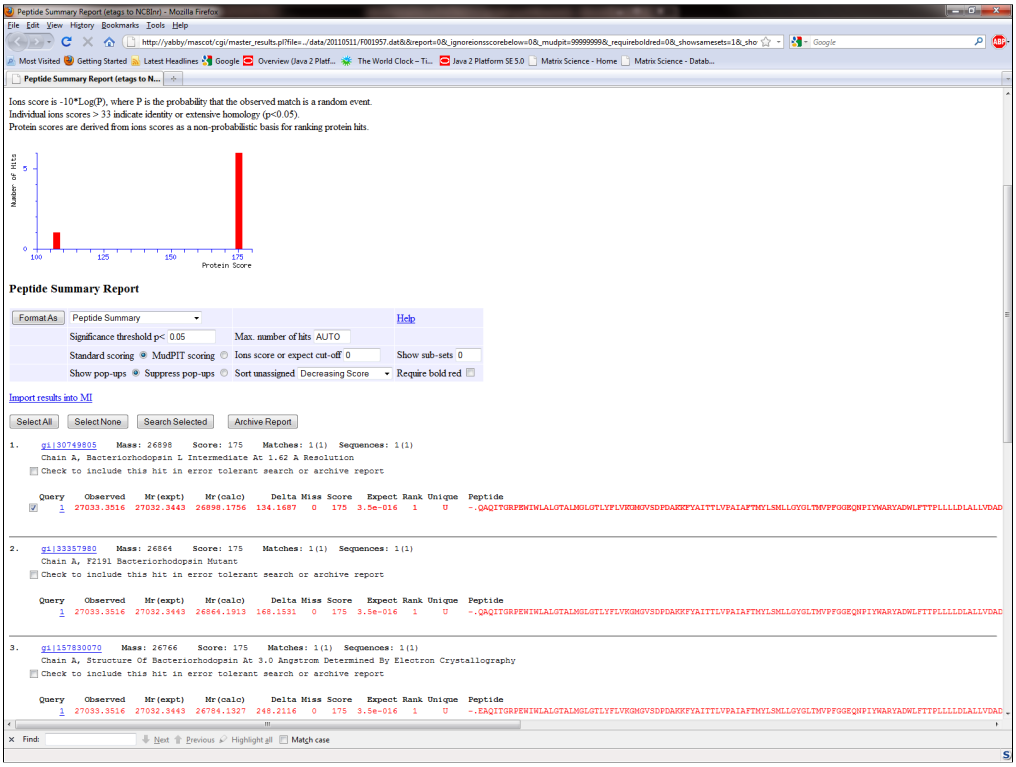
Initial MS/MS searches using either NoCleave or no enzyme specificity yielded no significant matches, so the next step is to try an error tolerant tag search. This is an even more flexible approach than the standard error tolerant search, but does require you to call enough accurate sequence tags to carry out the search.



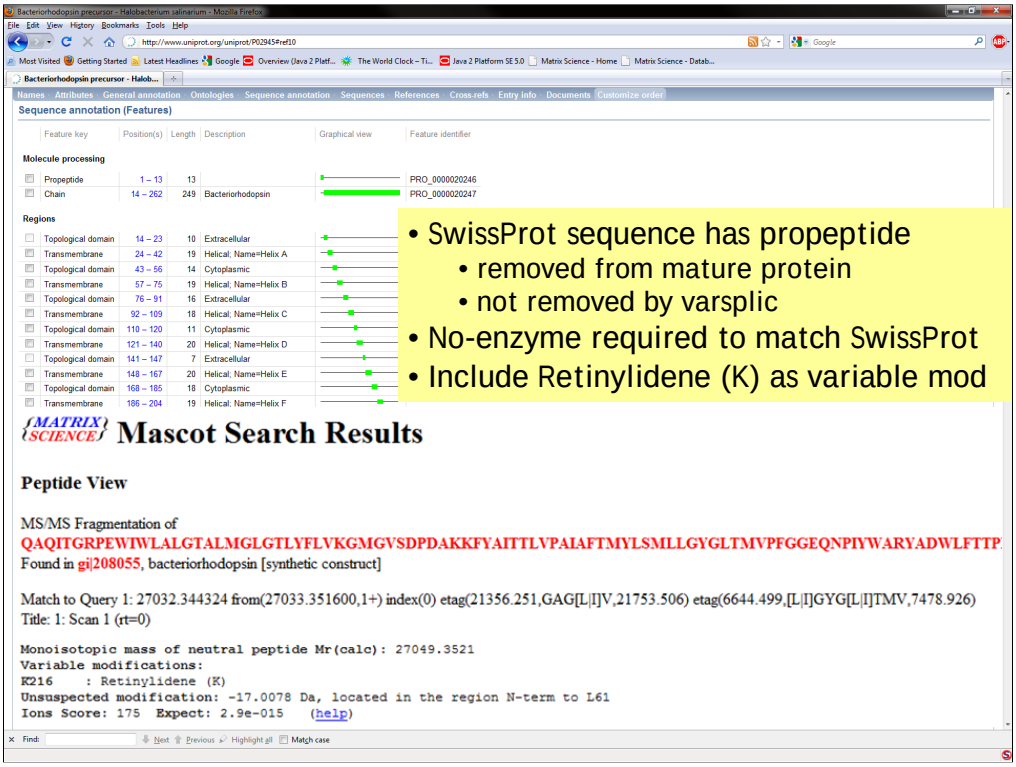
In Mascot Distiller we can call some sequence tags and submit an error tolerant tag search to Mascot TD.



A search of varsplic processed SwissProt still produced no matches.

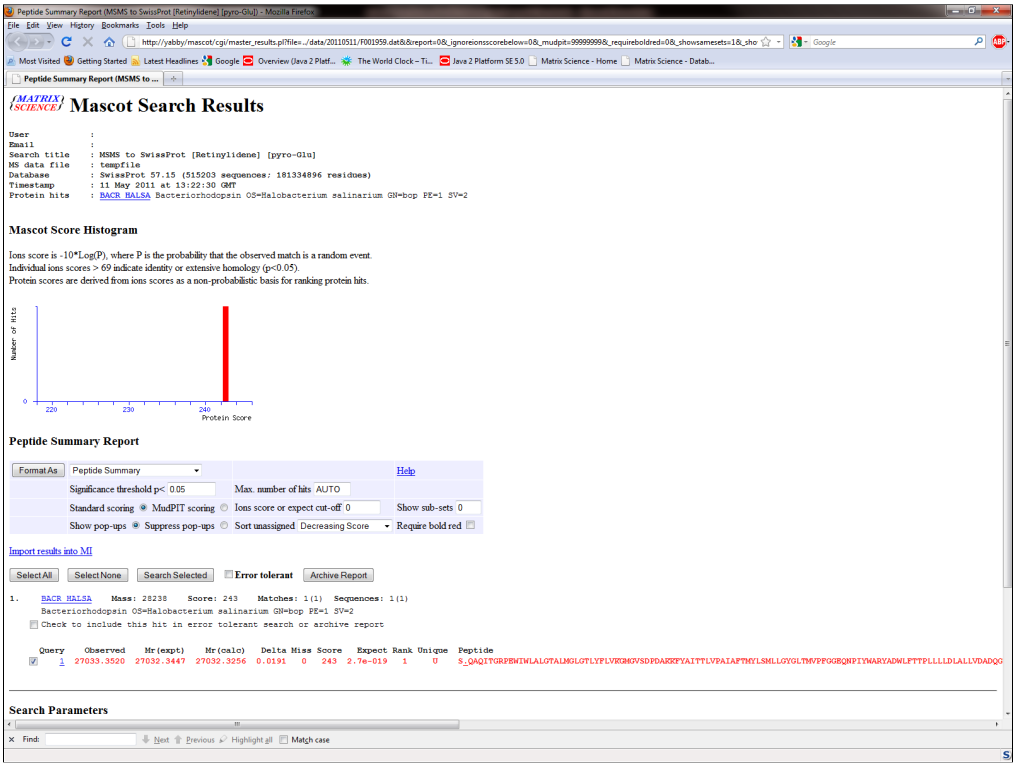


However, a search of NCBInr gives us a number of good matches to various Bacteriorhodopsin entries, with some large mass delta values as allowed for by the error tolerant tag search. Looking at these in more detail it quickly becomes clear why we're not getting a match to varsplic processed SwissProt.

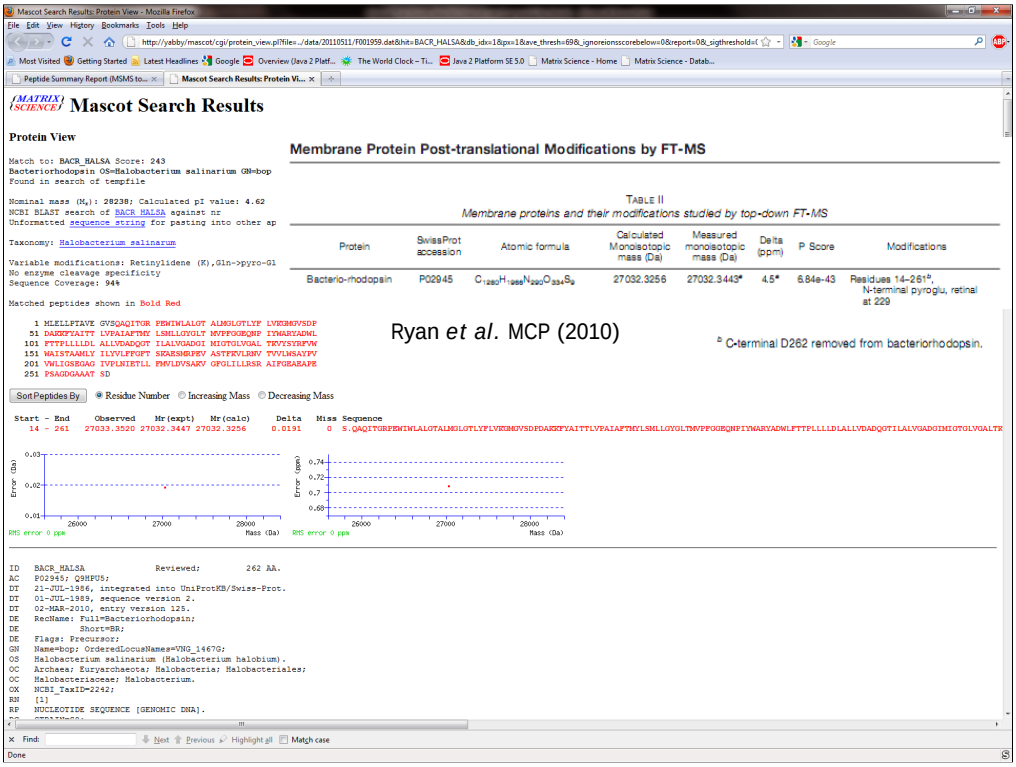


The matches from NCBInr are all from cloned, mutant or mature protein sequences which have had a 13 residue pro-peptide removed. Unfortunately varsplic does not process signal or pro–peptides so the only way to match the sequence in SwissProt would be to either run another round of processing to remove the pro-peptide or to drop enzyme specificity – and we cannot do for an error tolerant tag search. So for this example we'll carry on using nr for our tag searches while we try and explain the cause(s) of the mass deltas identified in the error tolerant tag search.

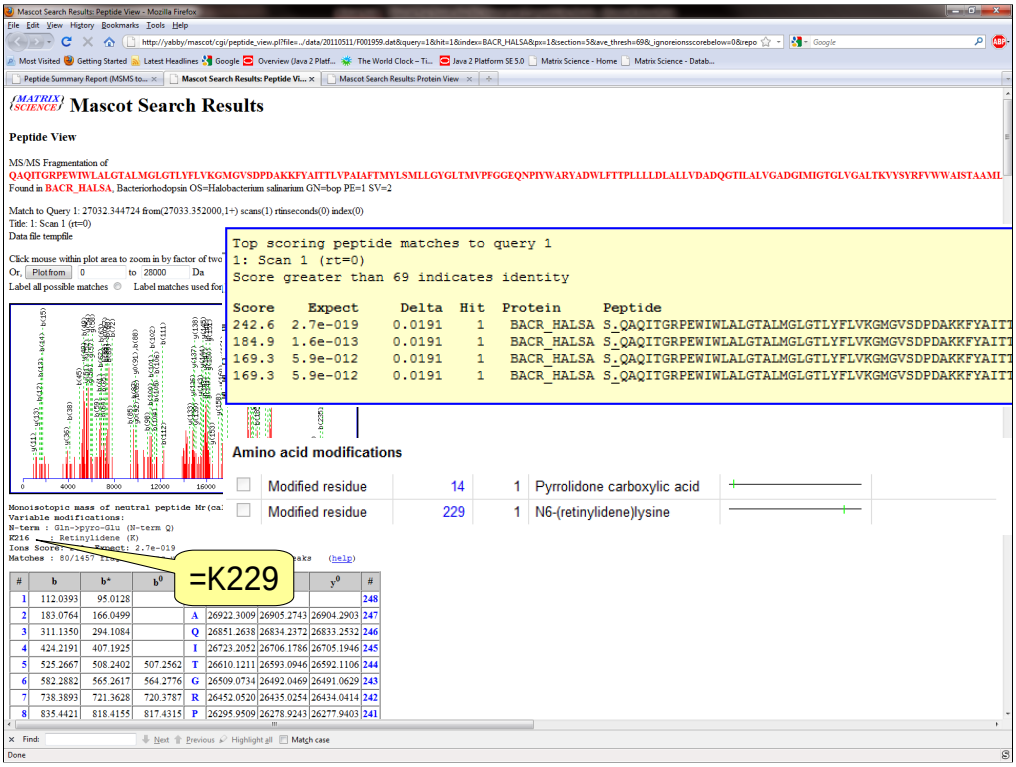
According to the paper, the precursor for both the apo- and holo- protein species of bacteriorhodopsin were observed, and the holoprotein species was deliberately selected for MS/MS i.e. the Retinylidene co-factor was still attached to the protein, and precursor selection was specifically done to select this species. So we should have included Retinylidene (K) as a variable modification in the search conditions. If you do that then the error tolerant tag search finds an unsuspected modification of approximately -17Da N terminal to L61 – which can be explained as N-terminal pyro-Glu



If we now carry out a no-enzyme MS/MS search using Retinyleidene and pyro-Glu as variable modifications we can get a good match to the SwissProt entry.



Taking a look at the protein view report, we can see that the match starts at Glutamine 14 – after the pro-peptide sequence. The match ends at Serine 261, which is in agreement with the results presented in Ryan *et al.* (2010) who also found that the C-terminal Aspartic Acid was removed.



Looking at the potential site of Retinylidene, the top match from Mascot is to K216 in the peptide with a score of 242.6. The second highest scoring match is to K129 modified and has a score of 184.9. That is still a good match, but I'd feel pretty confident in assigning K216 as the site of modification with the drop in score of approximately 58 between the matches. Confidence in that assignment grows even further when you see that K216 in the peptide is K229 in the full protein sequence including the pro-peptide – which is where SwissProt says the modification site is (and is even telling us about the pyro-glu). Again, these assignments are in agreement with the results presented in Ryan *et al.*

Summary

- **Mascot Top Down raises the precursor mass limit to 110kDa**
- **Use it in the same way as standard Mascot**
  - Same search form, same reports
- **Use etag or force error tolerant search to match unsuspected modifications**
- **One time purchase required**
  - 30 day trial license available on request

Acknowledgements

We are grateful to Bruker Daltonics Inc. for the Myoglobin and Superoxide dismutase example data files and extensive feedback during the development of Mascot Top Down