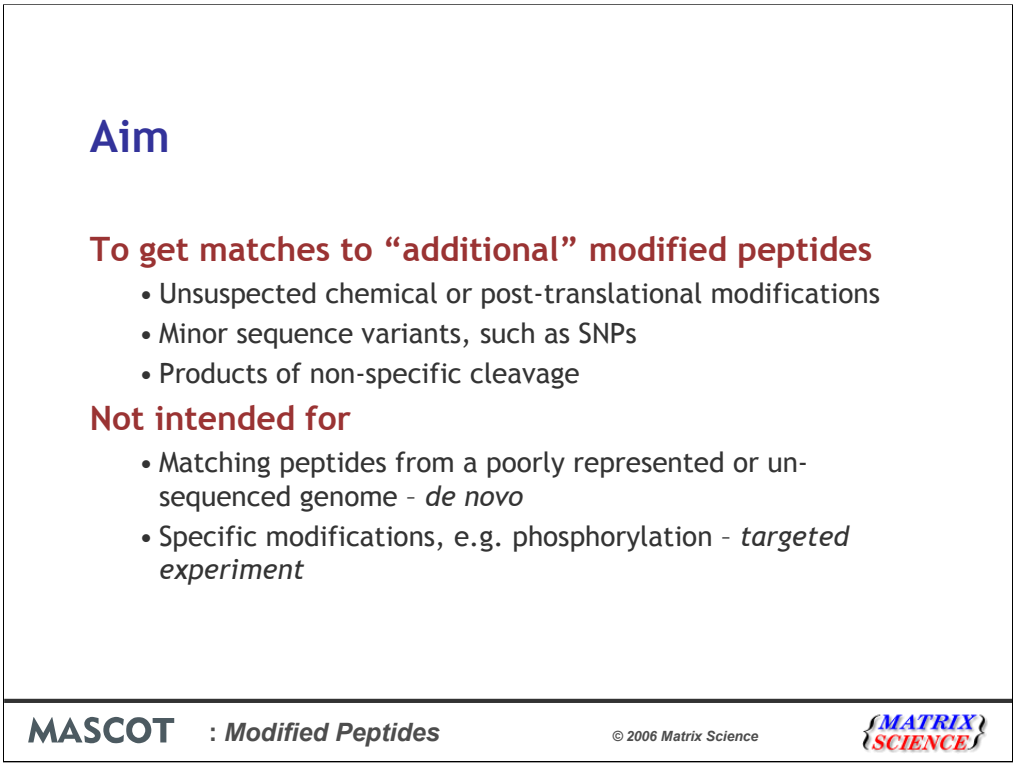
A Strategy for the Efficient Identification of Modified Peptides

MASCOT

MATRIX SCIENCE



I want to discuss the most efficient way to get as many matches as possible from an LC-MS/MS run. That is, to find matches to peptides with unsuspected chemical or post-translational modifications, with minor sequence variants, such as SNPs, and peptides which are the products of non-specific cleavage.

This strategy is not applicable to identifying peptides from protein that have little similarity to those in the database. As the databases fill up, this is becoming less common. If you are in this situation, the primary tool has to be de novo of high quality MS/MS spectra

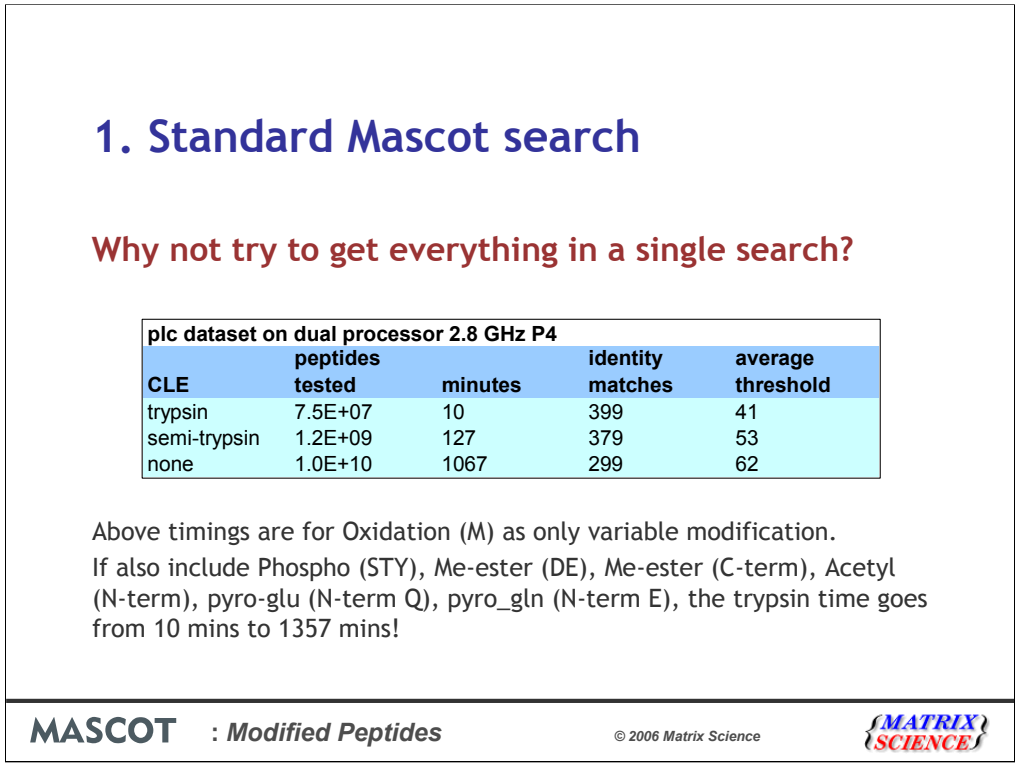
Neither is it applicable to investigations focused on a particular modification, such as phosphorylation. The key here would be a targeted experiment. Maybe using a neutral loss scan to identify and select the phosphorylated peptides, or an IMAC column to isolate the phosphopeptides

Search strategy

1. **Standard Mascot search**
   - Returns the easy matches
2. **Error tolerant search**
   - Returns additional matches, but only for proteins where we have at least one good peptide match already
   - Limited to a single additional SNP or modification per peptide
3. **De novo**
   - If data very high quality, can return novel full-length peptide sequences
     - Use Blast to find likely parent proteins
   - More often, returns partial / ambiguous peptide sequences
4. **Error tolerant tag search**
   - To find matches to
     1. Isolated peptides that have a SNP or unsuspected modification
     2. Peptides with multiple SNPs or unsuspected modifications
   - (No reason to expect additional matches from a standard tag search)

MASCOT : *Modified Peptides* MATRIX SCIENCE

If you simply want to get as many identifications as possible, so as to minimise the number of unmatched spectra and maximise protein coverage, you might come up with a strategy similar to this. I'll now go through the four steps in some detail. Step 1 is, of course, a standard Mascot MS/MS ions search.



Should we try to get as many matches as possible in the first pass search? Well, lets look at some numbers for a typical ion trap dataset when we search using loose trypsin, semi-specific trypsin, and no enzyme specificity

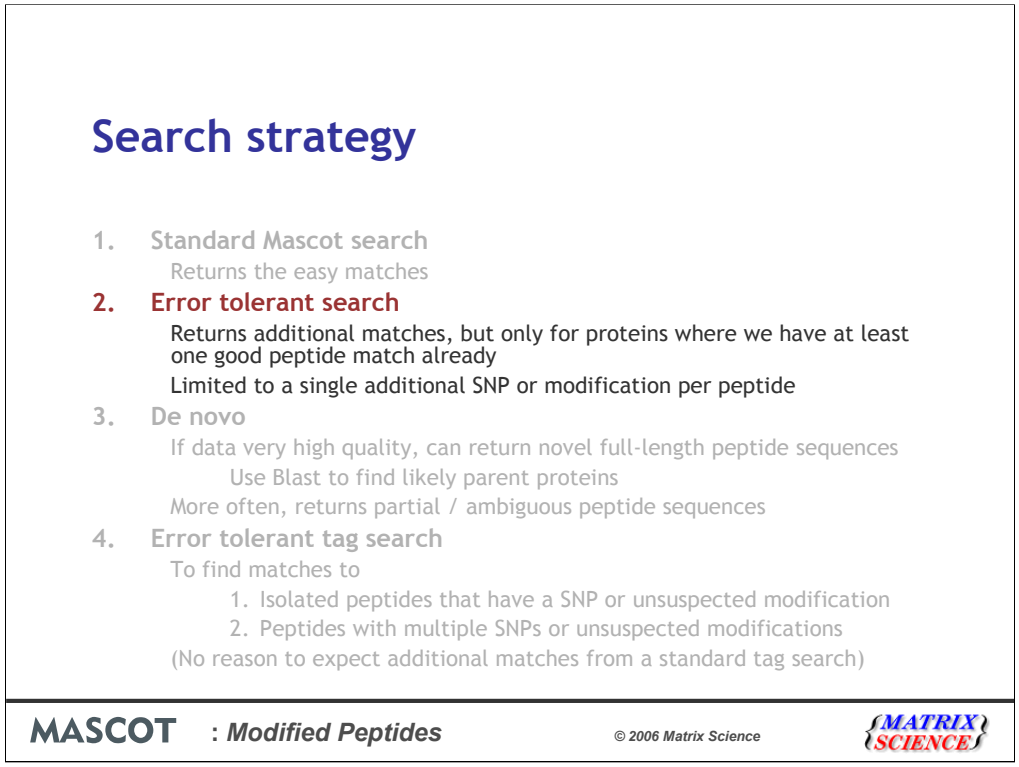
As you can see, the search time increases by an order of magnitude as we go from trypsin to semi-specific trypsin, and a further order of magnitude as we go to a completely non-specific search.

The reason is simple, the search space, that is the number of candidate peptides, is increasing by a factor of 10 each time. Having to wait 10 or 100 times as long for the results is bad enough. A more fundamental problem is that the significance threshold score is a function of the number of candidates, so this increase by 10 each time, and we lose marginal matches. Unless you have a high level of non-specific peptides in the sample, 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.

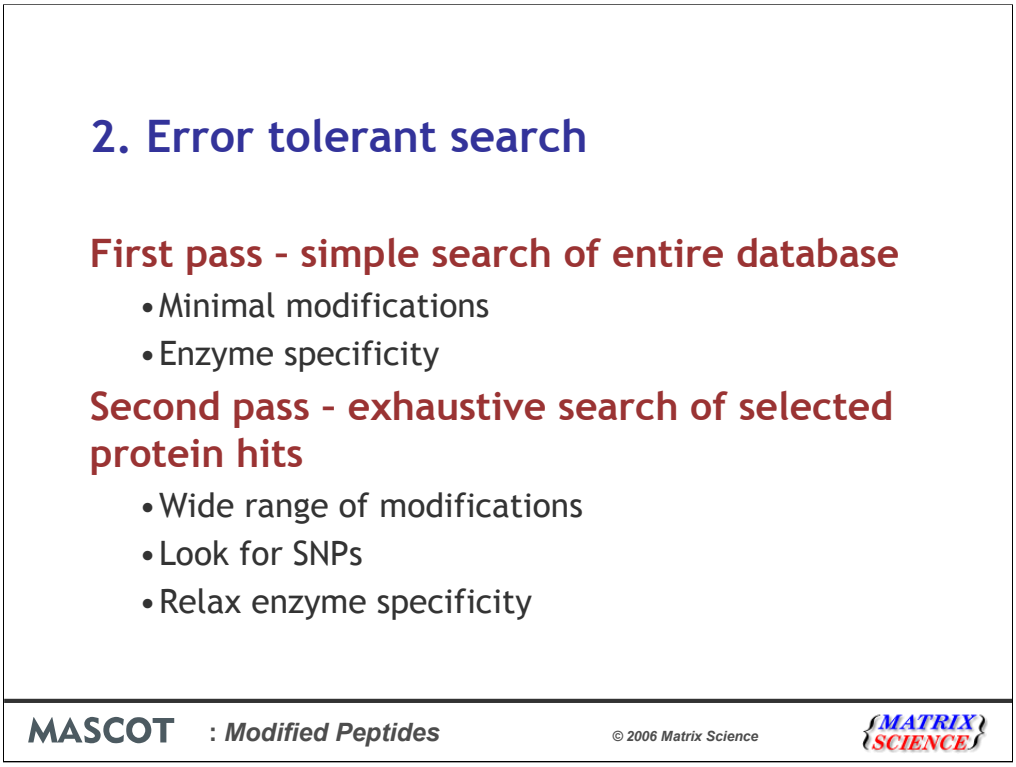
Identical considerations apply to modifications. If we go from 1 variable mod to 7, the search time is even worse than for no enzyme. This is because of the combinatorial explosion. Having to test all the combinations and permutations of these variable mods.

So, the answer is no, do not try to get as many matches as possible in the first pass search. It just makes the search very slow and very insensitive.



If you want to get as many identifications as possible, as efficiently as possible, the first pass search must be kept simple. Usually, strict or loose trypsin. Zero or one variable modifications. Certainly not more than two unless you know for sure they really are present.

Step 2 of our strategy is an error tolerant search. This is the efficient way to find unusual modifications, as well as variations in the primary sequence and peptides from non-specific cleavage



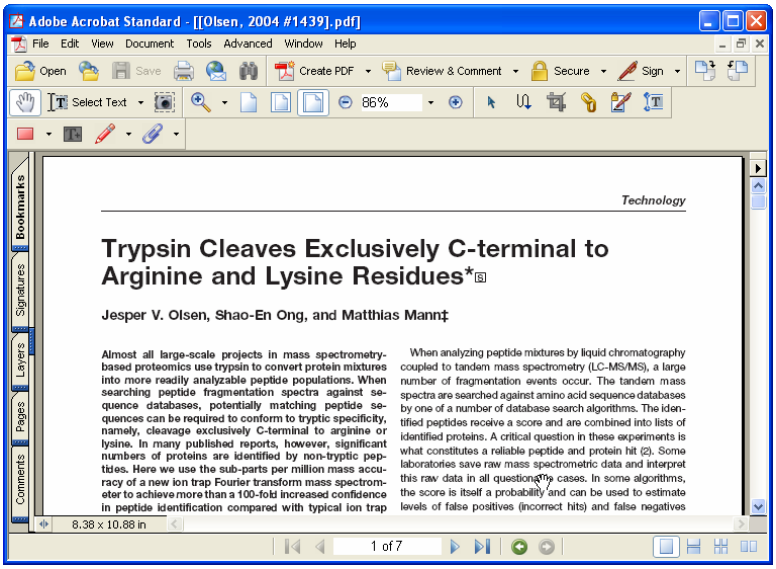
All the protein hits found in the first pass search are selected for an exhaustive second pass search.

Because only a handful of entries are being searched, search time is not an issue.

For modifications, an error tolerant search looks for one unsuspected modification per peptide in addition to those mods specified as fixed or variable. This is sufficient because it will be very, very rare to get two unsuspected mods on a single peptide.

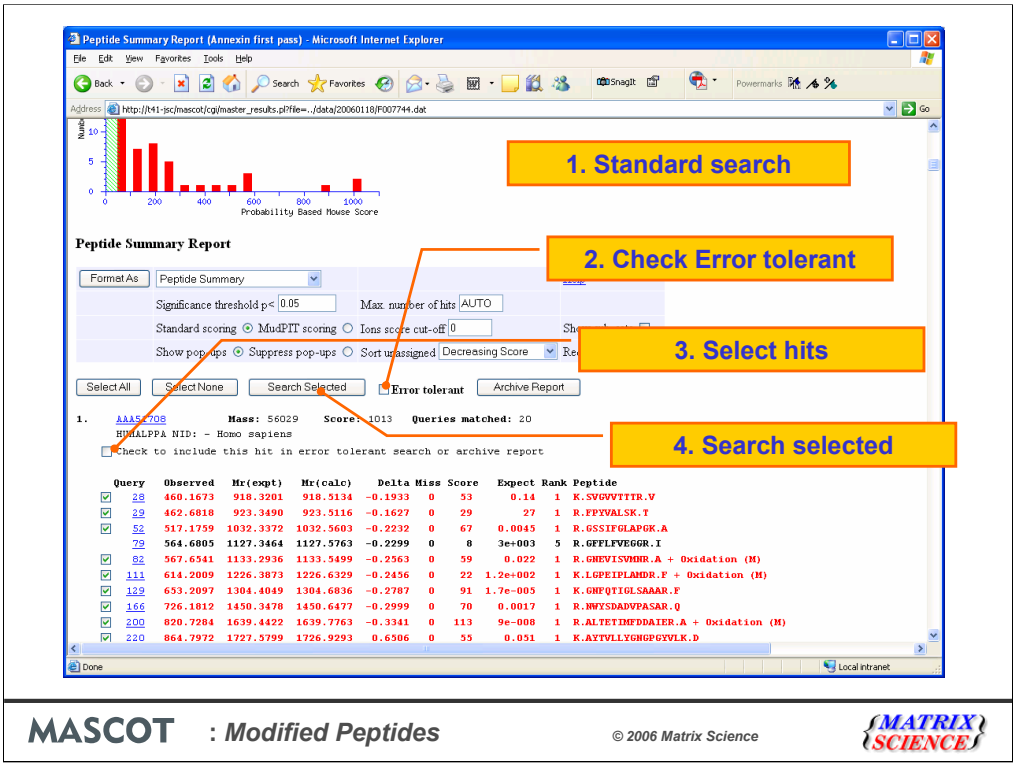
The error tolerant search also looks for sequence variants, such as single nucleotide polymorphisms (SNPs) or sequencing errors.

You can remove enzyme specificity completely, but you have to ask yourself whether you would believe a match that was doubly non-specific.

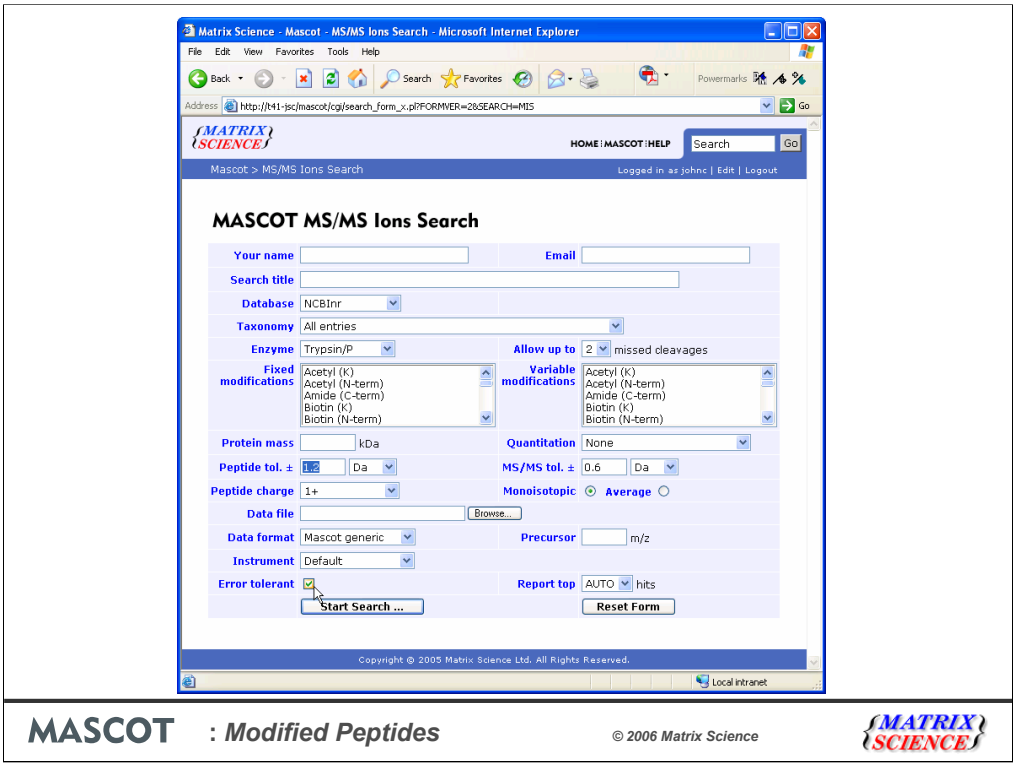


Olsen, J. V., Ong, S.-E. and Mann, M., Mol. and Cellular Proteomics, 3, 608-14 (2004)

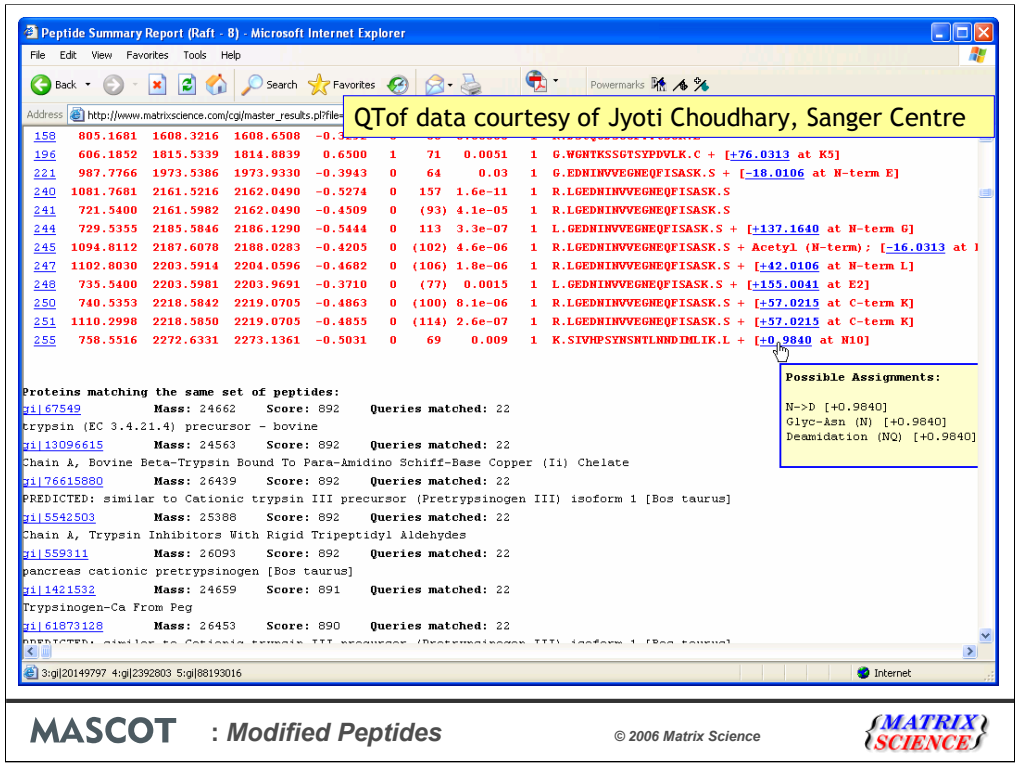
I think in most cases the answer is no. Our experience is that the levels of non-specific peptides are very low, less than 3%, unless there is something seriously wrong with the trypsin or the protocol. This is also the conclusion of a very careful study by Matthias Mann's group. So, in general, I prefer to use semi-trypsin in an error tolerant search



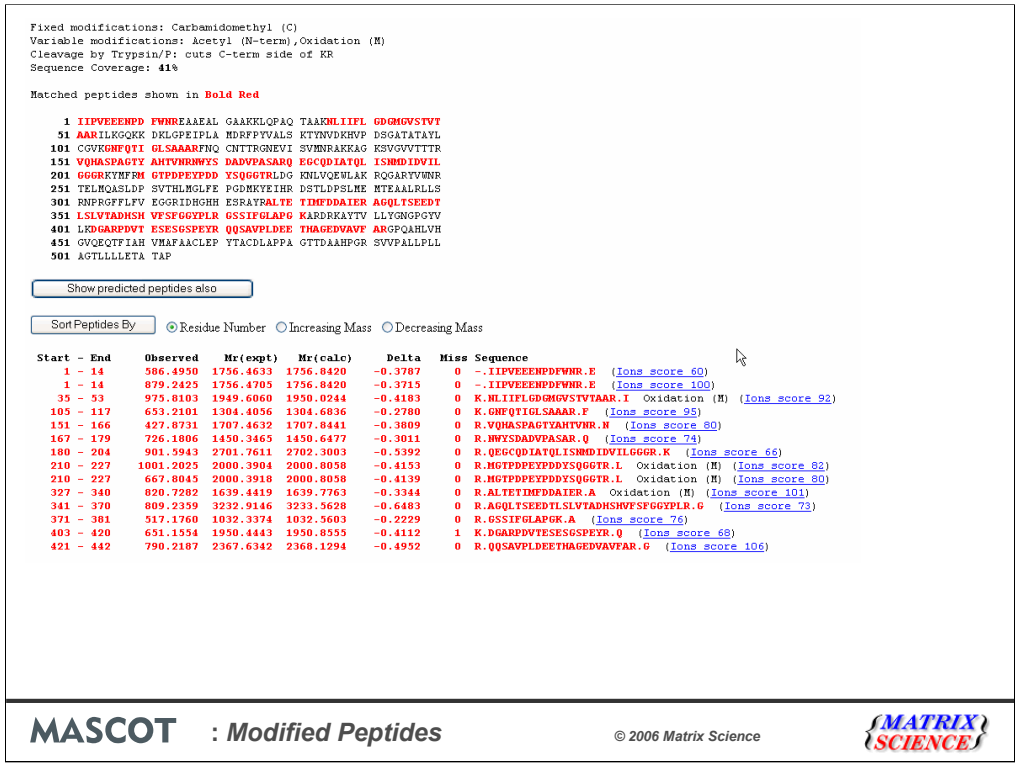
In the current version of Mascot, an error tolerant search is literally a second pass search. You have to select the protein hits you want to search by manually checking them off in the results report



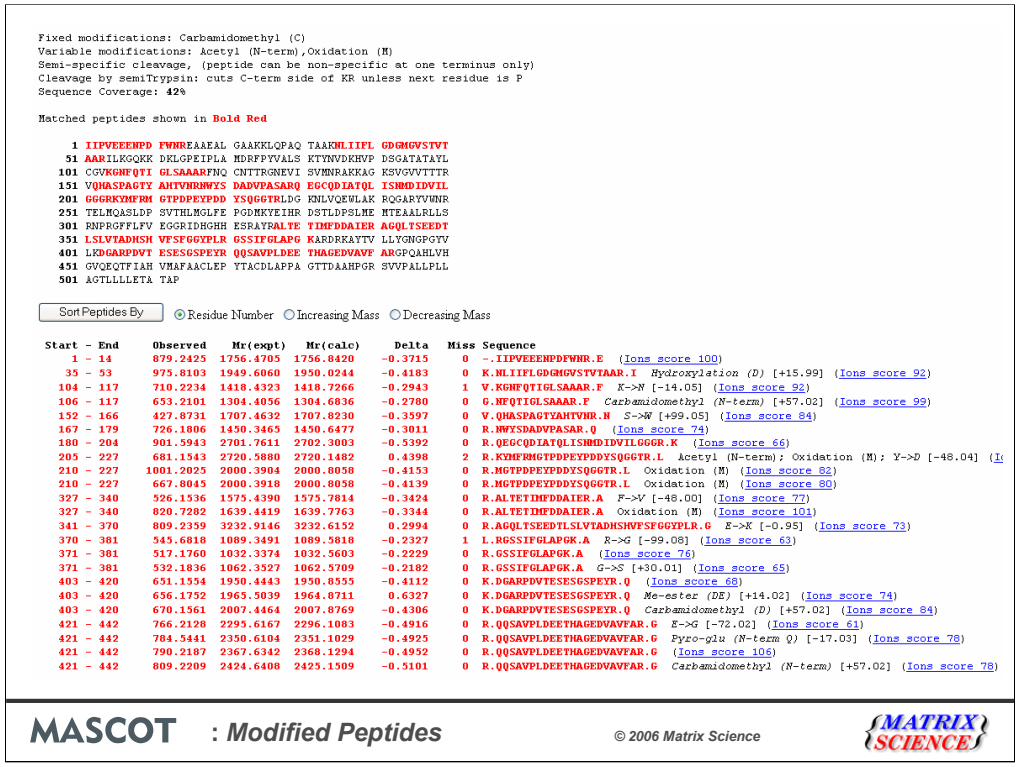
In Mascot 2.2, we will make the process integrated and automatic. You just have to check the Error tolerant box on the search form. This will perform a first pass search using the enzyme and modifications you specify in the search form. It will then automatically perform an error tolerant search on all of the proteins that contain significant peptide matches. David will say more about this in a later talk



We need to do more work on trying to filter out the unlikely matches. One rule we plan to introduce is that you can't have an unsuspected modification in a non-specific peptide. Another is to ignore a modification if it only gives a tiny increase in score over an unmodified and specific peptide. It will still be necessary to decide between alternative assignments of observed mass differences



The good news about the error tolerant search is that it substantially increases the number of matches. In this particular hit, from 14 to 22.



Fixed modifications: Carbamidomethyl (C)
Variable modifications: Acetyl (N-term),Oxidation (M)
Semi-specific cleavage, (peptide can be non-specific at one terminus only)
Cleavage by semiTrypsin: cuts C-term side of KR unless next residue is P
Sequence Coverage: **42%**

Matched peptides shown in **Bold Red**

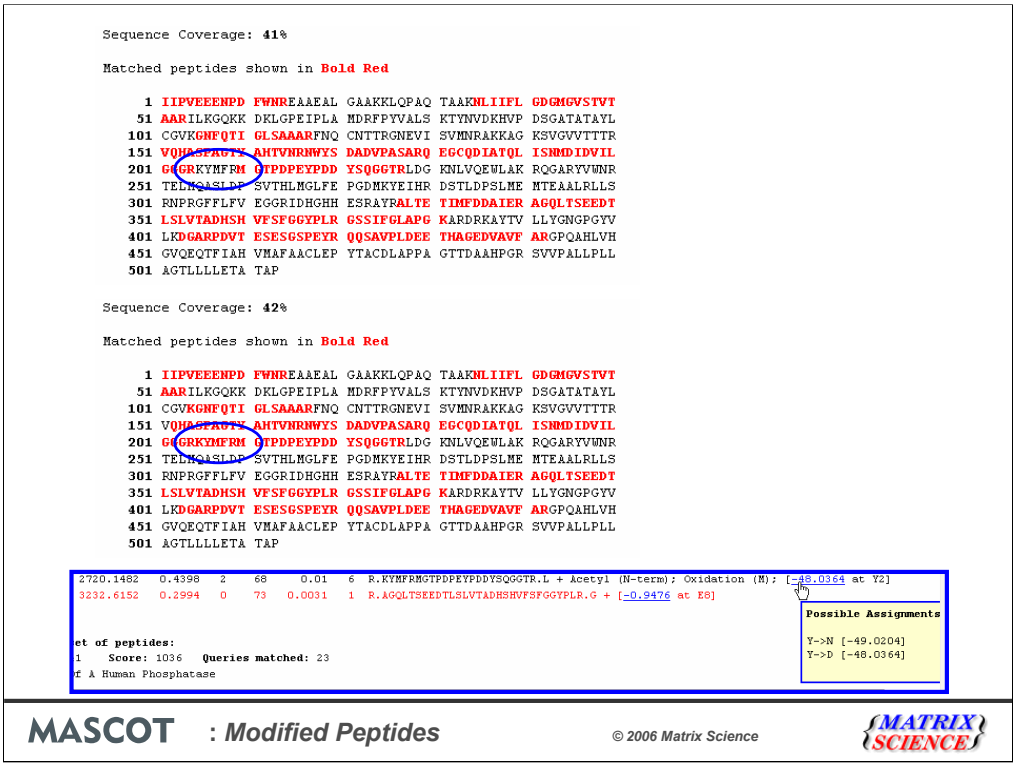
```
    1 IIPVEEENPD FWNREAAEAL GAAKKLQPAQ TAAKNLIIFL GDGMGVSTVT
   51 AARILKGQKK DKLGPEIPLA MDRFPYVALS KTYNVDKHVP DSGATATAYL
  101 CGVKGNFQTI GLSAAARFNQ CNTTRGNEVI SVMNRAKKAG KSVGVVTTTR
  151 VQHASPAGTY AHTVNRNWYS DADVPASARQ EGCQDIATQL ISNMDIDVIL
  201 GGGRKYMFRM GTPDPEYPDD YSQGGTRLDG KNLVQEWLAK RQGARYVWNR
  251 TELMQASLDP SVTHLMGLFE PGDMKYEIHR DSTLDPSLME MTEAALRLLS
  301 RNPRGFFLFV EGGRIDHGHH ESRAYRALTE TIMFDDAIER AGQLTSEEDT
  351 LSLVTADHSH VFSFGGYPLR GSSIFGLAPG KARDRKAYTV LLYGNGPGYV
  401 LKDGARPDVT ESESGSPEYR QQSAVPLDEE THAGEDVAVF ARGPQAHLVH
  451 GVQEQTFIAH VMAFAACLEP YTACDLAPPA GTTDAAHPGR SVVPALLPLL
  501 AGTLLLLETA TAP
```

Sort Peptides By ◉ Residue Number ○ Increasing Mass ○ Decreasing Mass

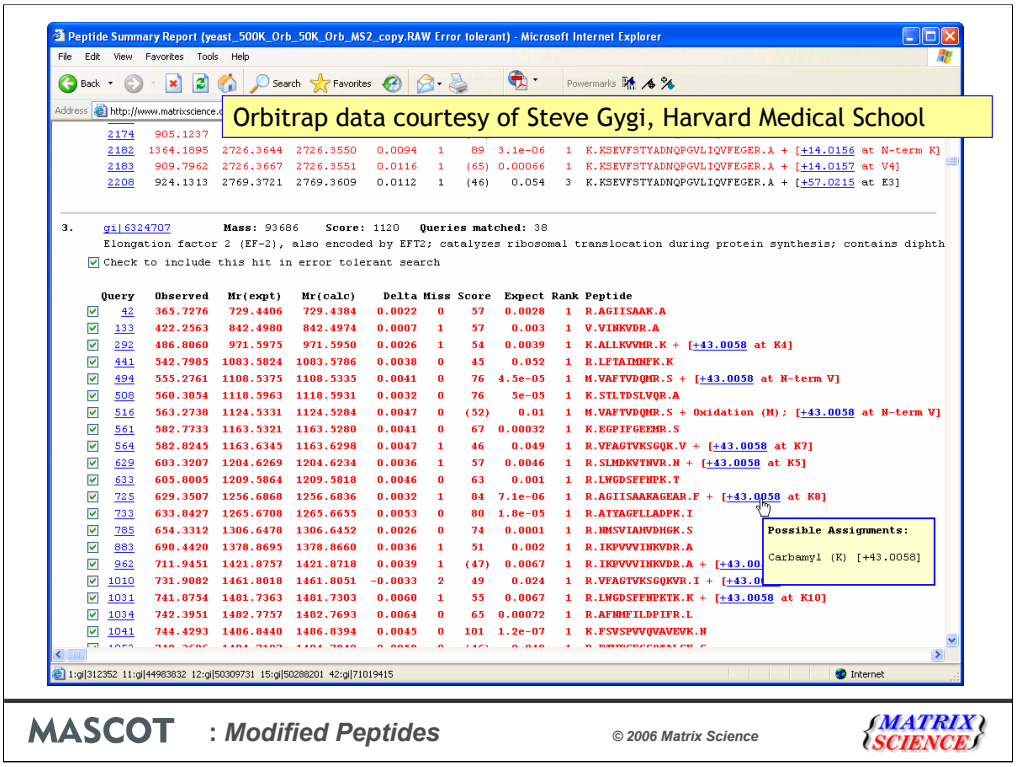
Start - End	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Sequence
1 - 14	879.2425	1756.4705	1756.8420	-0.3715	0	-.IIPVEEENPDFWNR.E (Ions score 100)
35 - 53	975.8103	1949.6060	1950.0244	-0.4183	0	K.NLIIFLGDGMGVSTVTAAR.I *Hydroxylation (D)* [+15.99] (Ions score 92)
104 - 117	710.2234	1418.4323	1418.7266	-0.2943	1	V.KGNFQTIGLSAAAR.F *K->N* [-14.05] (Ions score 92)
106 - 117	653.2101	1304.4056	1304.6836	-0.2780	0	G.NFQTIGLSAAAR.F *Carbamidomethyl (N-term)* [+57.02] (Ions score 99)
152 - 166	427.8731	1707.4632	1707.8230	-0.3597	0	V.QHASPAGTYAHTVNR.N *S->W* [+99.05] (Ions score 84)
167 - 179	726.1806	1450.3465	1450.6477	-0.3011	0	R.NWYSDADVPASAR.Q (Ions score 74)
180 - 204	901.5943	2701.7611	2702.3003	-0.5392	0	R.QEGCQDIATQLISNMDIDVILGGGR.K (Ions score 66)
205 - 227	681.1543	2720.5880	2720.1482	0.4398	2	R.KYMFRMGTPDPEYPDDYSQGGTR.L Acetyl (N-term); Oxidation (M); *Y->D* [-48.04] (Io
210 - 227	1001.2025	2000.3904	2000.8058	-0.4153	0	R.MGTPDPEYPDDYSQGGTR.L Oxidation (M) (Ions score 82)
210 - 227	667.8045	2000.3918	2000.8058	-0.4139	0	R.MGTPDPEYPDDYSQGGTR.L Oxidation (M) (Ions score 80)
327 - 340	526.1536	1575.4390	1575.7814	-0.3424	0	R.ALTETIMFDDAIER.A *F->V* [-48.00] (Ions score 77)
327 - 340	820.7282	1639.4419	1639.7763	-0.3344	0	R.ALTETIMFDDAIER.A Oxidation (M) (Ions score 101)
341 - 370	809.2359	3232.9146	3232.6152	0.2994	0	R.AGQLTSEEDTLSLVTADHSHVFSFGGYPLR.G *E->K* [-0.95] (Ions score 73)
370 - 381	545.6818	1089.3491	1089.5818	-0.2327	1	L.RGSSIFGLAPGK.A *R->G* [-99.08] (Ions score 63)
371 - 381	517.1760	1032.3374	1032.5603	-0.2229	0	R.GSSIFGLAPGK.A (Ions score 76)
371 - 381	532.1836	1062.3527	1062.5709	-0.2182	0	R.GSSIFGLAPGK.A *G->S* [+30.01] (Ions score 65)
403 - 420	651.1554	1950.4443	1950.8555	-0.4112	0	K.DGARPDVTESESGSPEYR.Q (Ions score 68)
403 - 420	656.1752	1965.5039	1964.8711	0.6327	0	K.DGARPDVTESESGSPEYR.Q *Me-ester (DE)* [+14.02] (Ions score 74)
403 - 420	670.1561	2007.4464	2007.8769	-0.4306	0	K.DGARPDVTESESGSPEYR.Q *Carbamidomethyl (D)* [+57.02] (Ions score 84)
421 - 442	766.2128	2295.6167	2296.1083	-0.4916	0	R.QQSAVPLDEETHAGEDVAVFAR.G *E->G* [-72.02] (Ions score 61)
421 - 442	784.5441	2350.6104	2351.1029	-0.4925	0	R.QQSAVPLDEETHAGEDVAVFAR.G *Pyro-glu (N-term Q)* [-17.03] (Ions score 78)
421 - 442	790.2187	2367.6342	2368.1294	-0.4952	0	R.QQSAVPLDEETHAGEDVAVFAR.G (Ions score 106)
421 - 442	809.2209	2424.6408	2425.1509	-0.5101	0	R.QQSAVPLDEETHAGEDVAVFAR.G *Carbamidomethyl (N-term)* [+57.02] (Ions score 78)

MASCOT : *Modified Peptides* © 2006 Matrix Science MATRIX SCIENCE

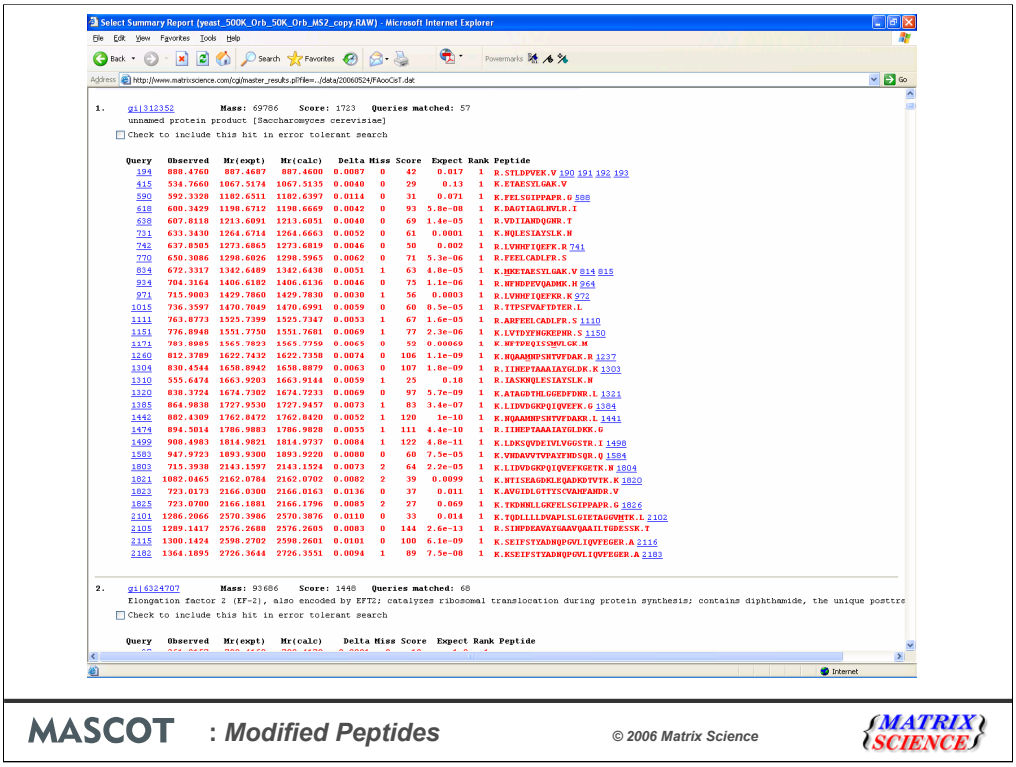
The bad news is that the sequence coverage hardly changes, 41% to 42%



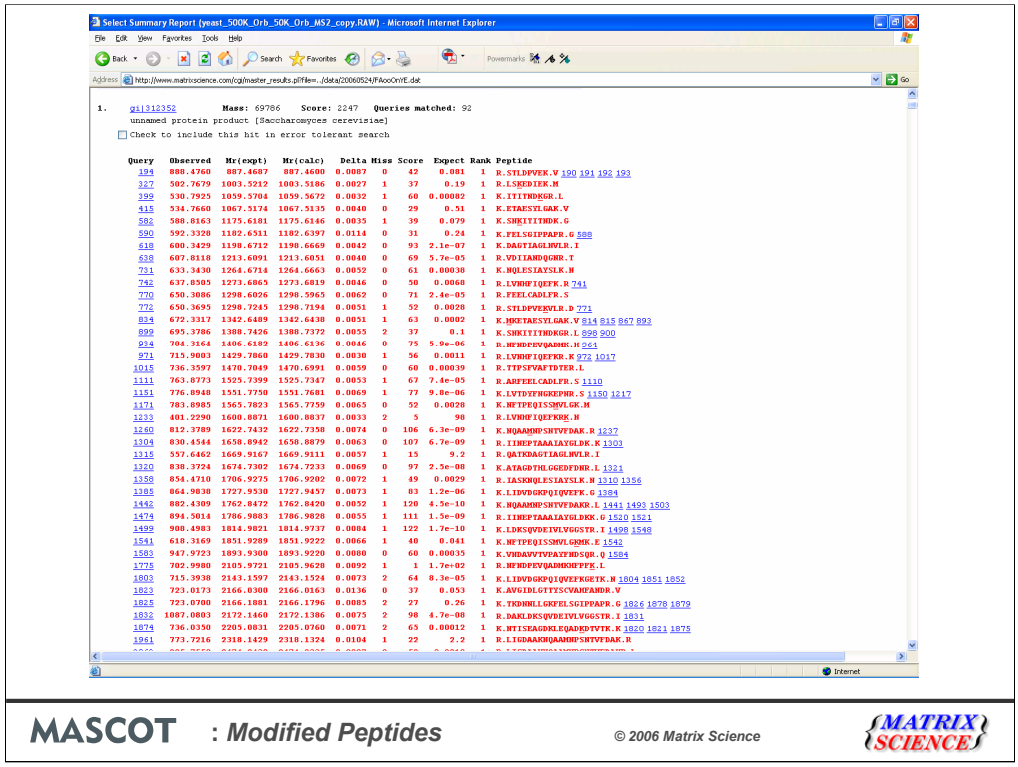
If we look more closely, we see that this is just an additional 5 residues here, KYMFR. However, when we look at the matches, there is only one match spanning this peptide and it is a weak and dubious match, requiring two mods and a SNP. In all honesty, I would not want to accept this match, so the coverage is actually unchanged. For this particular protein, the error tolerant search just gives us additional matches to the same peptides we saw in the standard search.



If you don't want to spend a lot of time studying the matches from an error tolerant search and deciding which you accept, you can use it as a quick way of spotting whether there are modifications which should be included in the first pass search as variable mods. Here's a nice example. Lots of matches for a modification of 43 Da, almost certainly carbamylation.

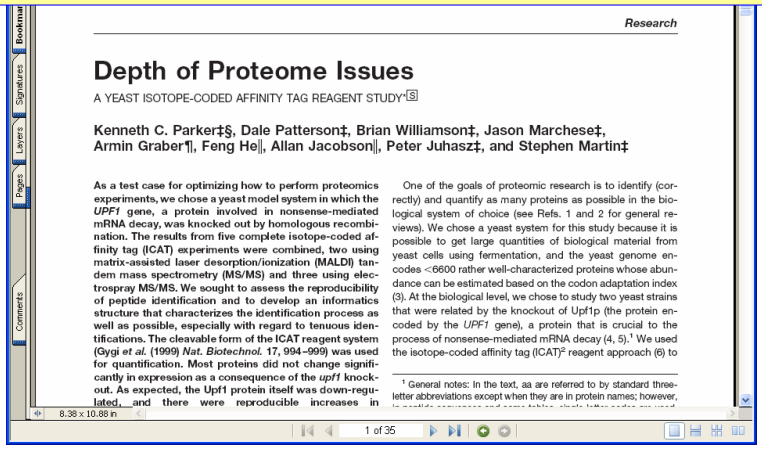


Adding this in as a variable mod increases the number of peptide matches for hit 1 from 57



To 92. However, coverage only increases from 57% to 64%. Basically, just one new peptide, that was only present in carbamylated form

"One of the main conclusions from this investigation is that as a general rule, chemically modified peptides are only likely to be correctly identified if they derive from the most abundant peptides in the sample."



Research

Depth of Proteome Issues

A YEAST ISOTOPE-CODED AFFINITY TAG REAGENT STUDY[S]

Kenneth C. Parker‡§, Dale Patterson‡, Brian Williamson‡, Jason Marchese‡, Armin Graber¶, Feng He‖, Allan Jacobson‖, Peter Juhasz‡, and Stephen Martin‡

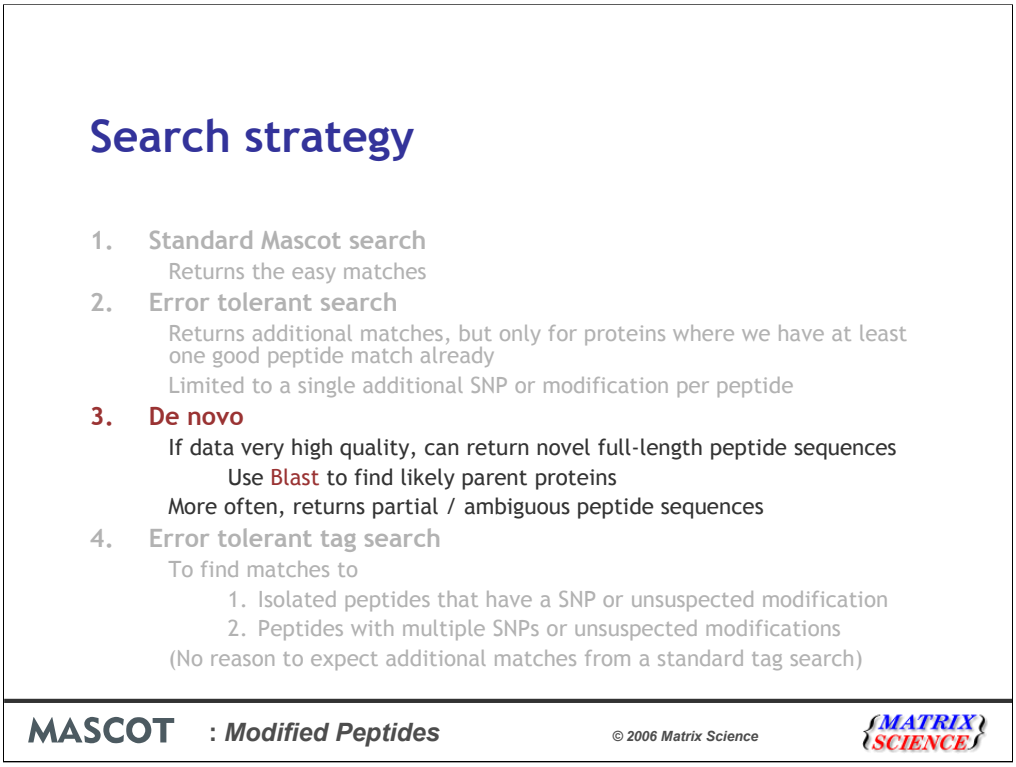
As a test case for optimizing how to perform proteomics experiments, we chose a yeast model system in which the *UPF1* gene, a protein involved in nonsense-mediated mRNA decay, was knocked out by homologous recombination. The results from five complete isotope-coded affinity tag (ICAT) experiments were combined, two using matrix-assisted laser desorption/ionization (MALDI) tandem mass spectrometry (MS/MS) and three using electrospray MS/MS. We sought to assess the reproducibility of peptide identification and to develop an informatics structure that characterizes the identification process as well as possible, especially with regard to tenuous identifications. The cleavable form of the ICAT reagent system (Gygi *et al.* (1999) *Nat. Biotechnol.* 17, 994–999) was used for quantification. Most proteins did not change significantly in expression as a consequence of the *upf1* knockout. As expected, the Upf1 protein itself was down-regulated, and there were reproducible increases in

One of the goals of proteomic research is to identify (correctly) and quantify as many proteins as possible in the biological system of choice (see Refs. 1 and 2 for general reviews). We chose a yeast system for this study because it is possible to get large quantities of biological material from yeast cells using fermentation, and the yeast genome encodes <6600 rather well-characterized proteins whose abundance can be estimated based on the codon adaptation index (3). At the biological level, we chose to study two yeast strains that were related by the knockout of Upf1p (the protein encoded by the *UPF1* gene), a protein that is crucial to the process of nonsense-mediated mRNA decay (4, 5).[1] We used the isotope-coded affinity tag (ICAT)[2] reagent approach (6) to

[1] General notes: In the text, aa are referred to by standard three-letter abbreviations except when they are in protein names; however,

Parker, K. C. *et. al.*, Mol. and Cellular Proteomics, 3, 625-59 (2004)

This observation is in agreement with a very careful study by Ken Parker and colleagues. As you go deeper, you tend to find modified versions of peptides that you already identified. They did detailed manual validation of matches to non-specific peptides and found approximately 2% were semi-tryptic and zero were fully non-specific.



Where can we go from here? Well, maybe there are some peptides which can't be picked up by the error tolerant search. Maybe a peptide that spans a splice site or a peptide with a modification that is not in our list of modifications. The next step is de novo.

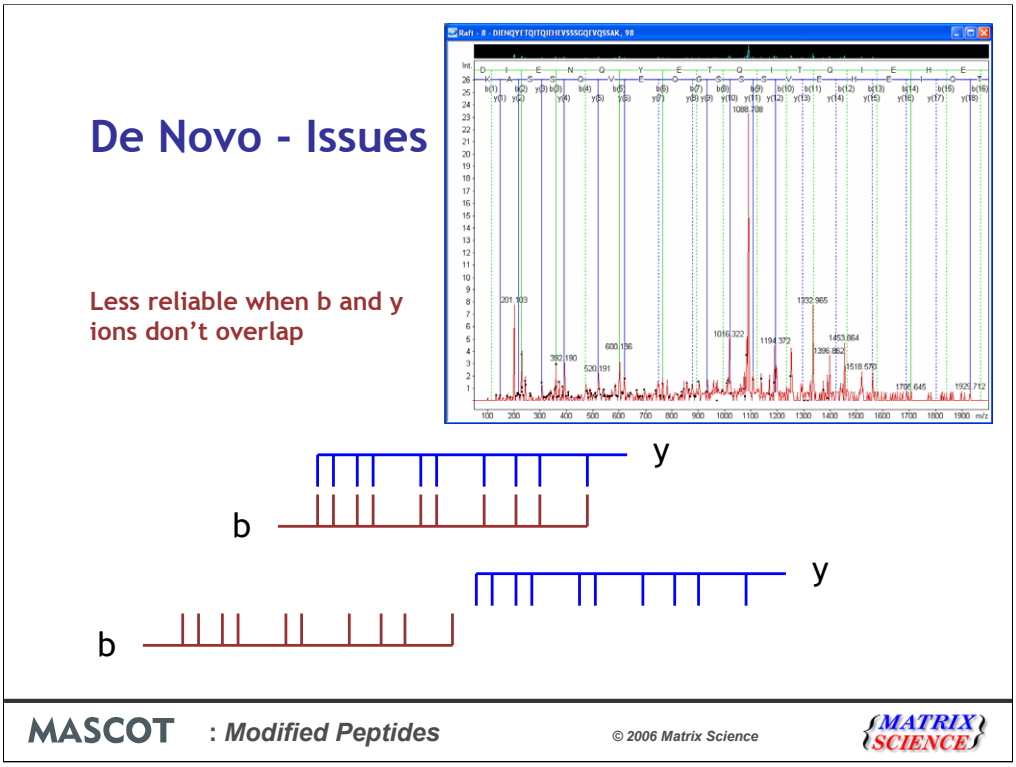
De Novo - Issues

```
Database               : NCBInr 20060518 (3647739 sequences)
Enzyme                 : Trypsin/P
Fixed modifications    : Carbamidomethyl (C)
Variable modifications : Acetyl (N-term),Oxidation (M)
Peptide Mass Tolerance : ± 0.8 Da
Fragment Mass Tolerance: ± 0.4 Da
Max Missed Cleavages   : 2
```

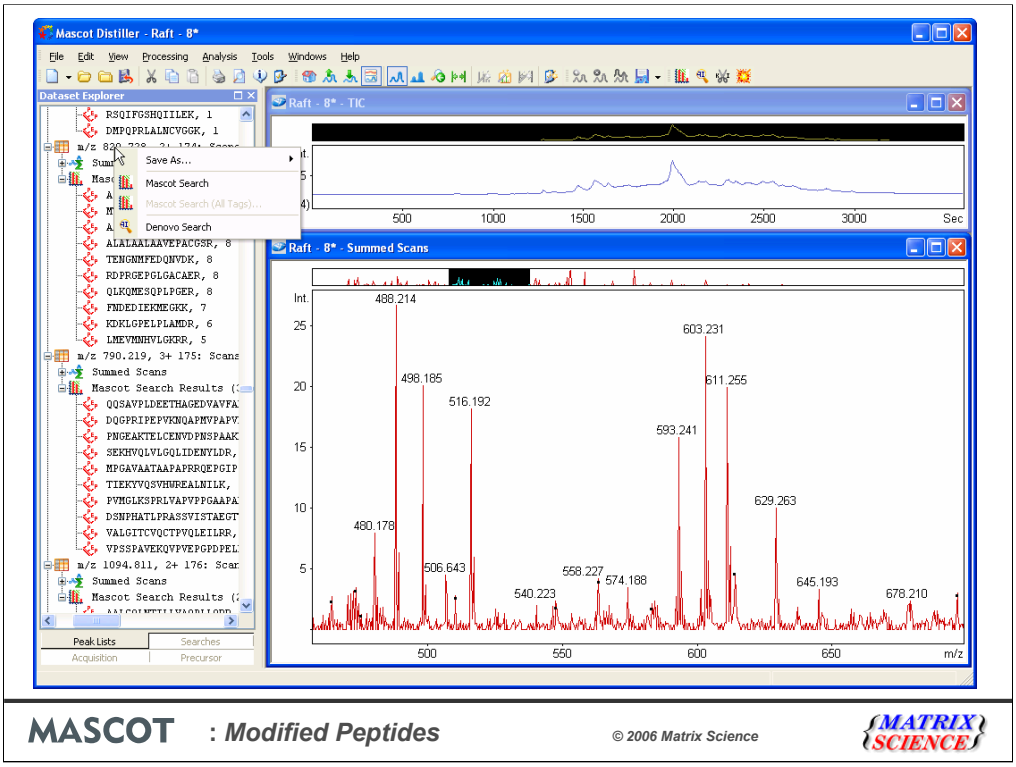
Peptide mass	Database search	De novo
~ 1000 Da	~ 6×10^5	~ 1×10^8
~ 2000 Da	~ 5×10^5	~ 4×10^{19}

MASCOT : *Modified Peptides* MATRIX SCIENCE

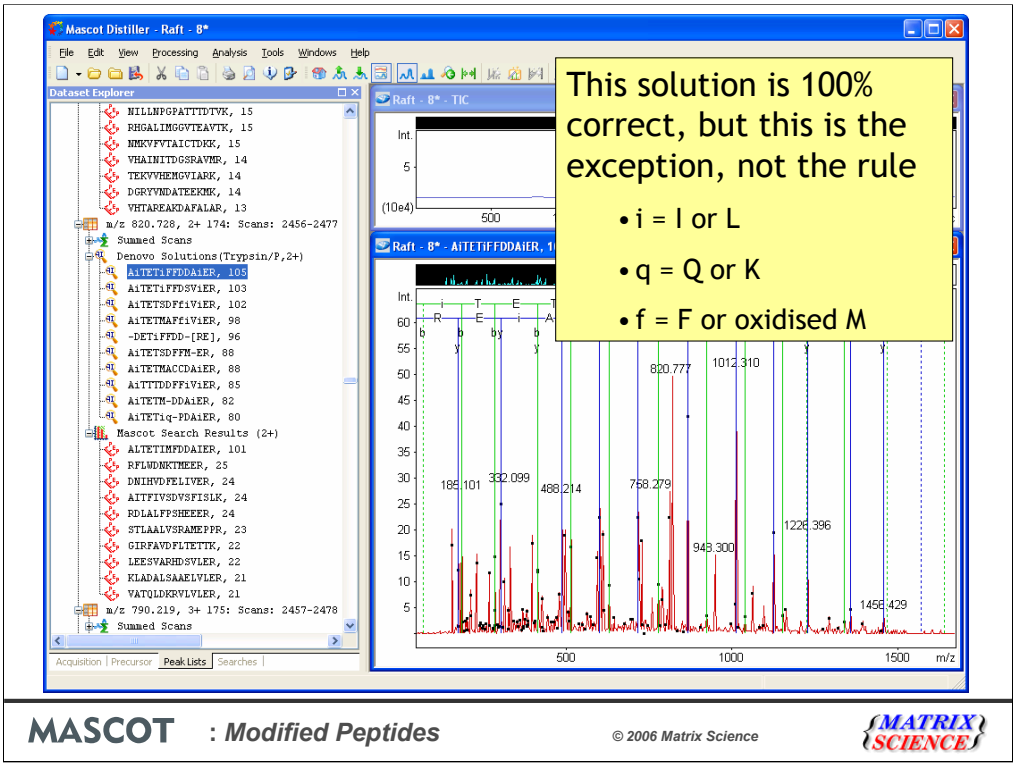
The problem with de novo is that the search space is huge. If we assume tryptic specificity, the bigger the peptide, the fewer the candidates in a database search. With de novo, the number of candidate sequences grows geometrically with peptide length. In reality, things aren't so bad. Any practical de novo algorithm explores only a small portion of this search space. Nevertheless, you cannot expect to get de novo solutions from large peptides unless the signal to noise and mass accuracy are both very good.



Another important factor is coverage. It is hard to over emphasise the importance of getting both N-term and C-term matches for a stretch of de novo sequence to be reliable. This is a particular problem with large peptides, where the spectrum is often only good at (say) the low mass end. If the C-term ladder and the N-term ladder do not overlap, this is a much, much less constrained situation.



De novo is implemented in Mascot Distiller, because it requires very reliable peak picking. The starting point can be any MS/MS scan that has been processed to create a peak list. Right click the peak list node in Dataset Explorer and choose ‘de novo Search’, or choose the de novo button from the toolbar when a Summed Scans node is selected.

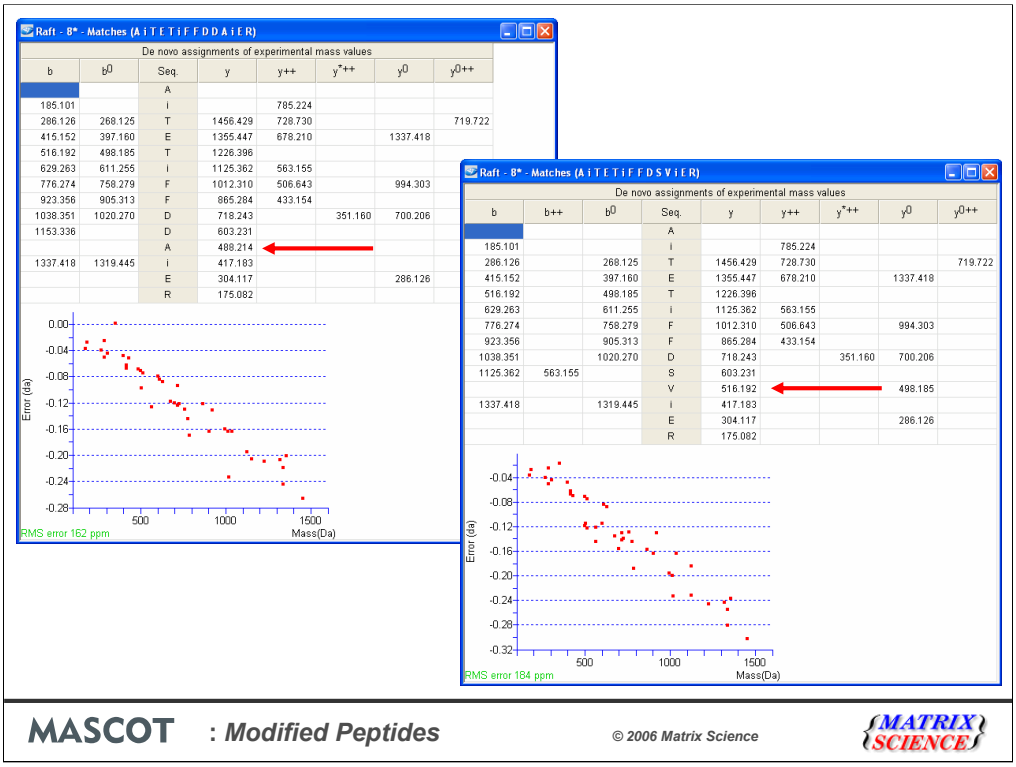


Good signal to noise and good mass accuracy are critical for successful de novo sequencing; much more so than in database searching. GIGO (garbage in - garbage out) is guaranteed.

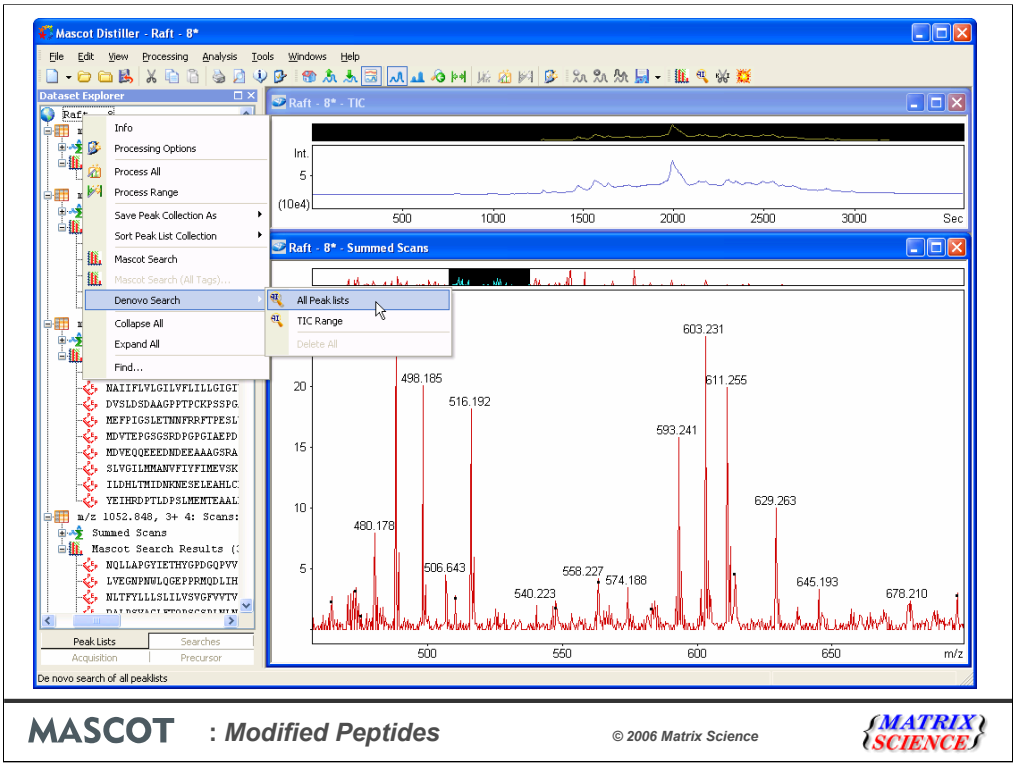
In a de novo solution, i always represents I or L. q represent Q or K, when the mass tolerance does not allow these residues to be distinguished, although K is assumed at the C terminus of a peptide when tryptic specificity applies. f represents F or Met-Ox.

Ambiguity is indicated by a dash in the sequence. The tooltip shows details of the ambiguity in square brackets, using pipe symbols to separate alternatives. Note that the order of the pairs and triplets is undefined, so that SP could also be PS.

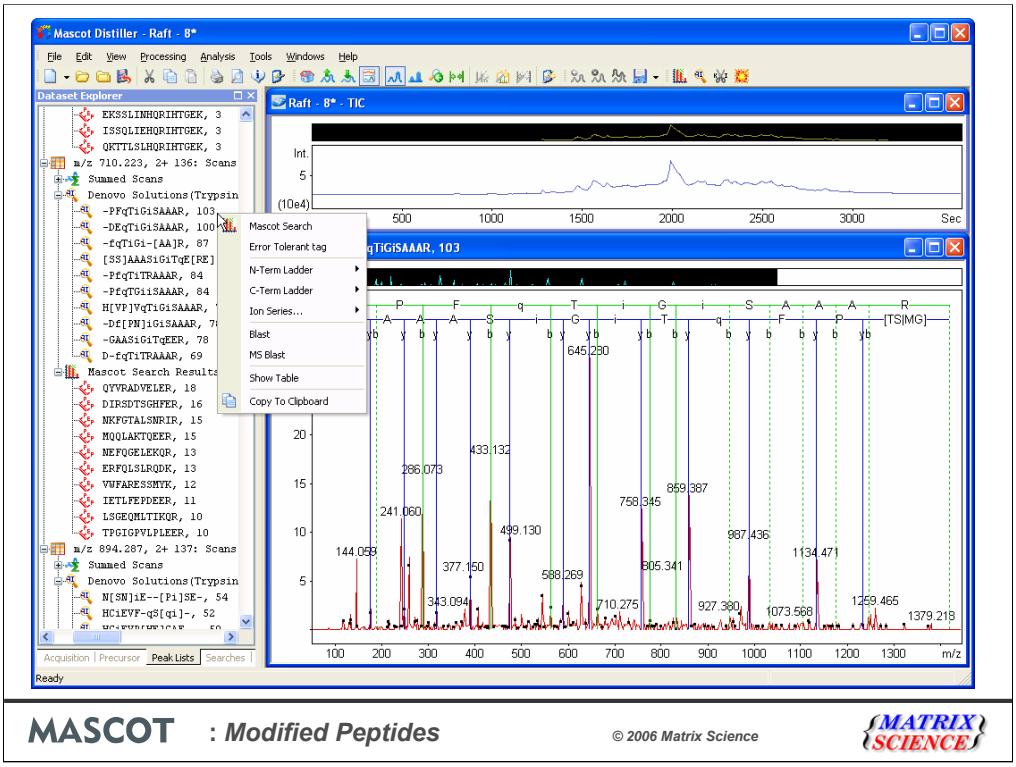
Although the example shown here looks very different to the Mascot database match, they are actually in perfectly agreement. Some uncertainty is unavoidable in de novo, because the search space is so very much larger. For example, the score hardly changes when DA is replaced by SV



If we look at details of these two matches, we can see why. There is a y ion peak at 488.214 to support the sequence being DA, but there is also a y ion peak at 516.192 to support the alternative. This is just the nature of de novo on non-ideal data



To de novo sequence a complete peak list collection, or the peak lists in the currently displayed TIC range, use the context menu obtained by right-clicking the root (world) node

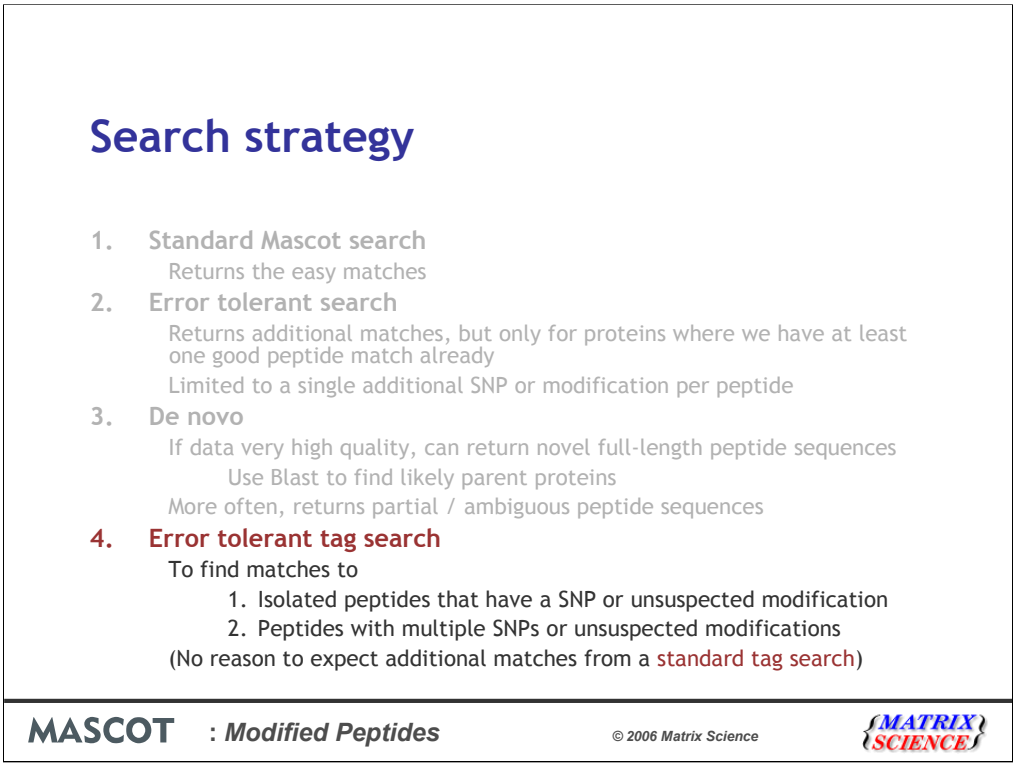


MASCOT : *Modified Peptides* © 2006 Matrix Science MATRIX SCIENCE

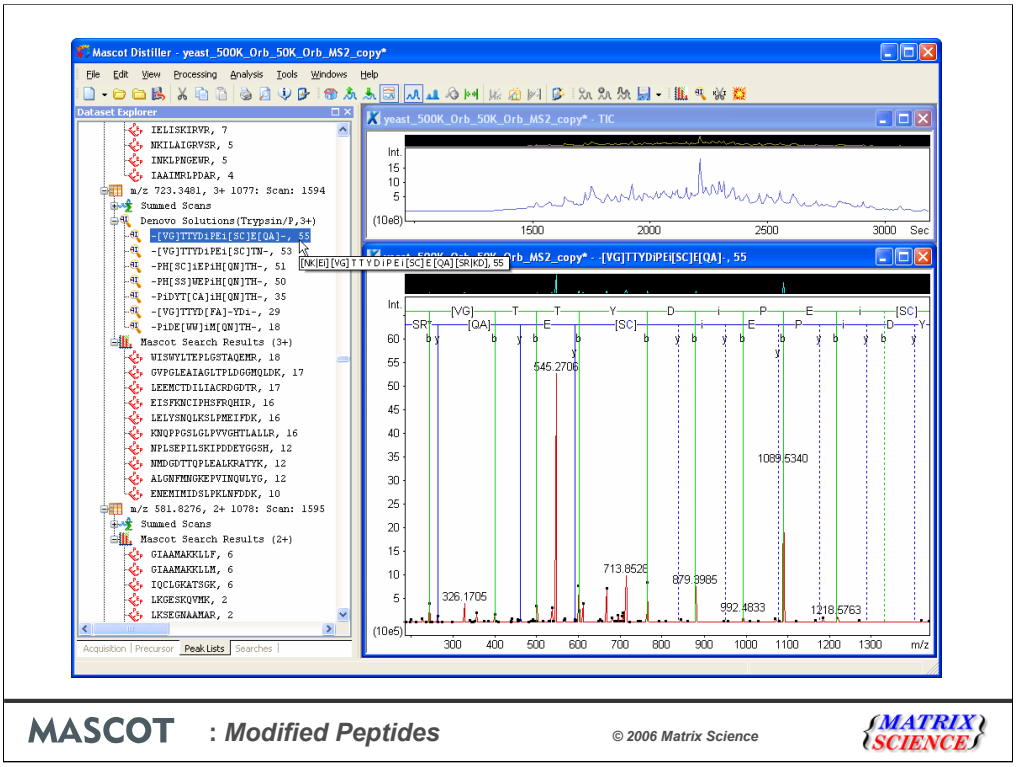
You can then browse down the tree, looking for cases where the database search failed and de novo has a high score.

This looks like a promising case. The Mascot search didn't get a significant match, but de novo has a very high score.

But, is it right? And, how do we resolve the ambiguity at the N-terminus?



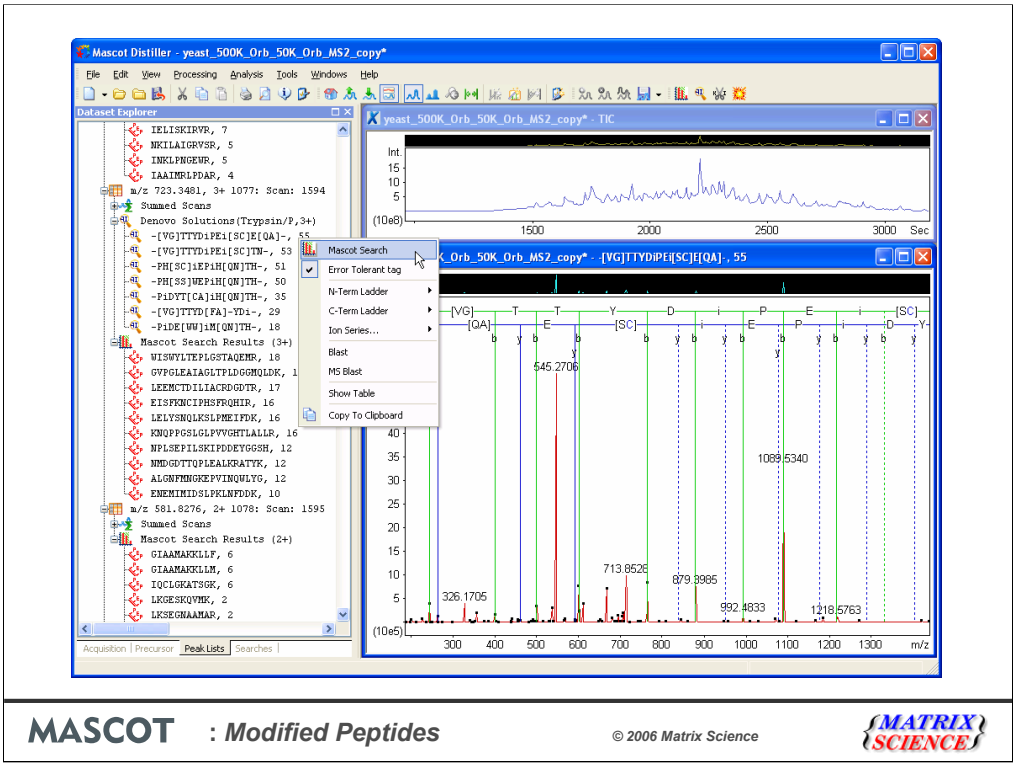
This brings us to step 4 of our strategy. As long as we are not dealing with an un-sequenced genome, the best way to test a de novo solution is an error tolerant tag search. This can often get a match even when there are multiple differences between the analyte peptide and the database sequence



MASCOT : *Modified Peptides* © 2006 Matrix Science MATRIX SCIENCE

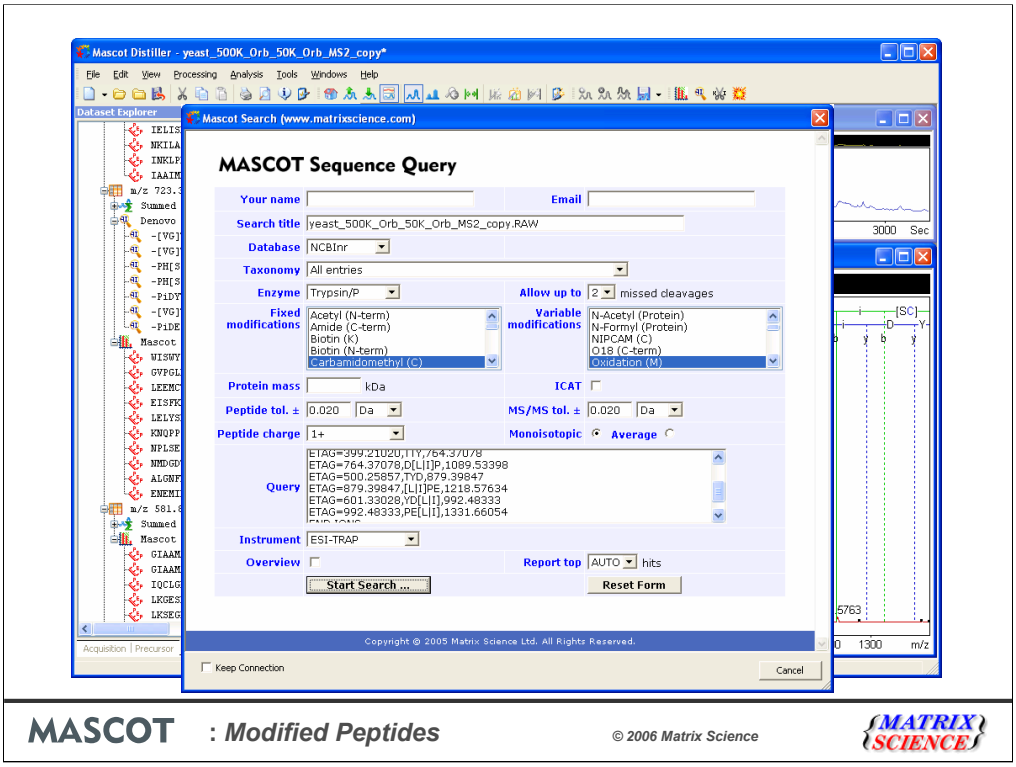
Here's another example, from the Orbitrap data, where the Mascot database search has failed to find a match

The de novo solution is not a great score, and there's ambiguity at each terminus



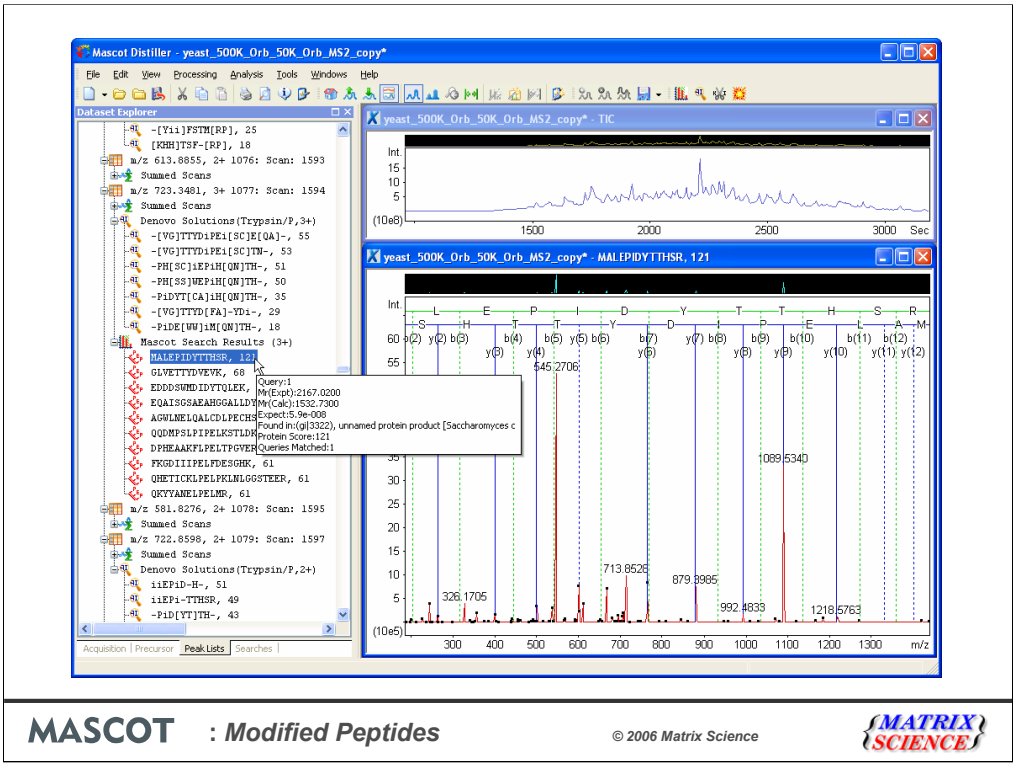
MASCOT : *Modified Peptides* *© 2006 Matrix Science* MATRIX SCIENCE

Right click the solution and choose Mascot search from the context menu. Note that we have already toggled the tag type to error tolerant



MASCOT : *Modified Peptides* *© 2006 Matrix Science* MATRIX SCIENCE

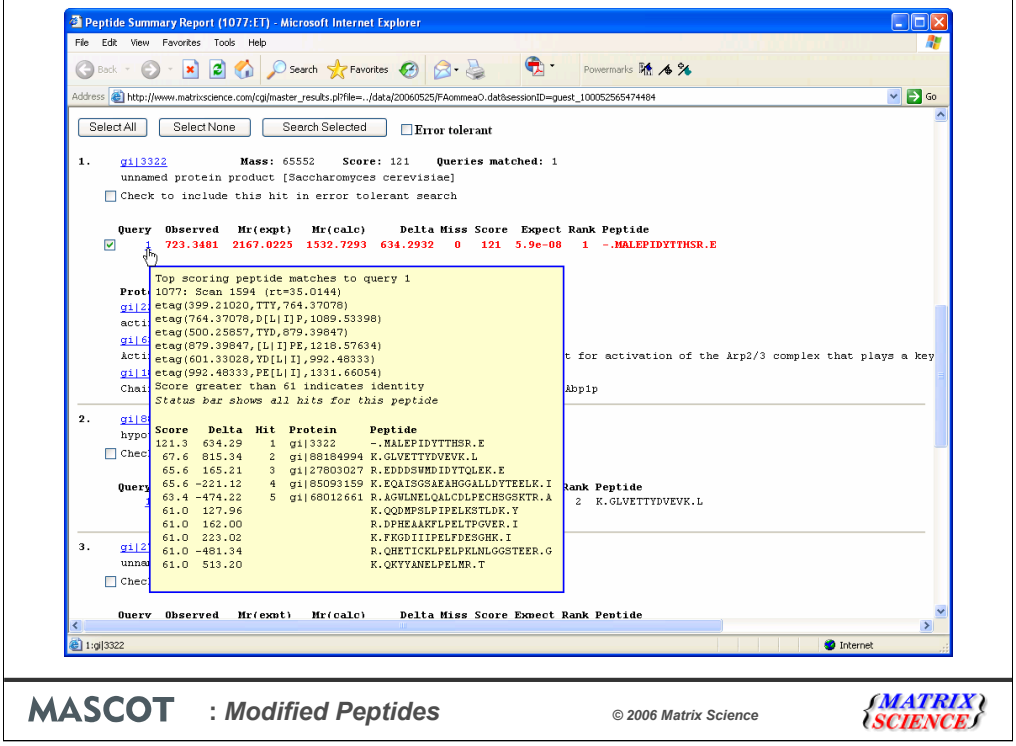
Distiller populates the query field with the tags taken from the non-ambiguous parts of the de novo solution. We submit the search …



And back comes the result. Note that the results from this most recent search have replaced the original database search. You can switch back to the previous results by selecting them on the searches tab.

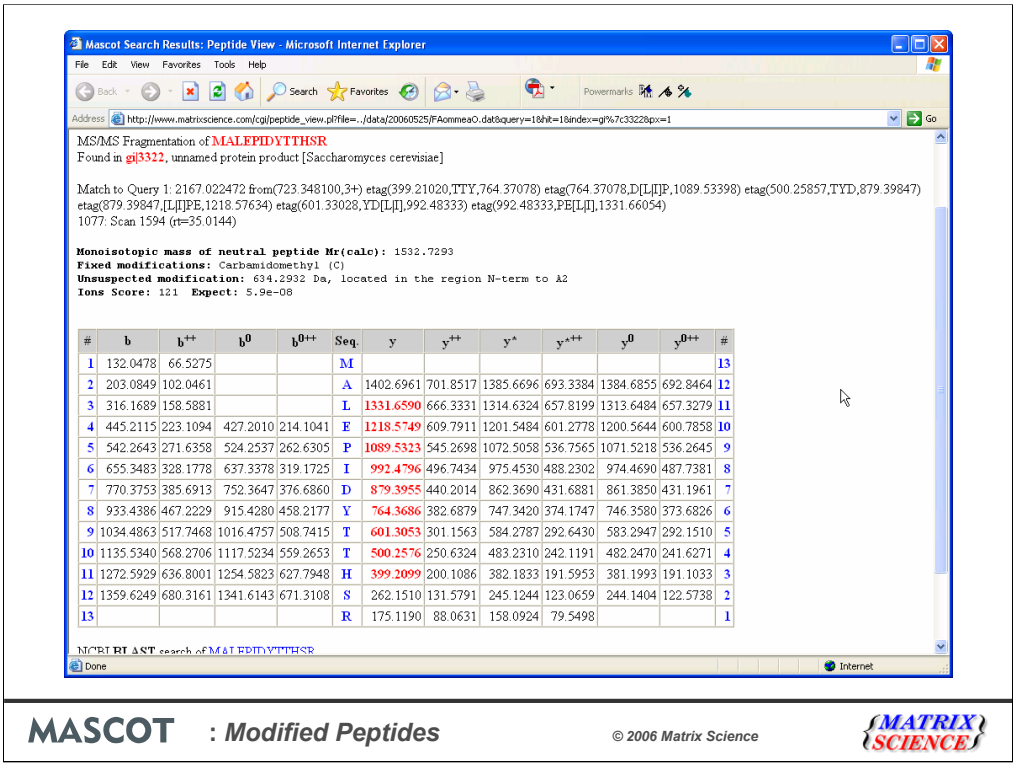
This match looks very promising. It's a high score, and it's a protein to which we already have other good matches. Notice that the de novo solution wasn't bad, but it was reversed. TTYDiPEi should have been iEPiDYTT. Unless the de novo manages to reach a terminus, there's a 50:50 chance that it will be the wrong way round

If we right click and choose to view the full Mascot report in a browser …



The reason we didn't get a match from Mascot is that there is a modification, giving a delta of 634 Da. The peptide forms the protein N-terminus

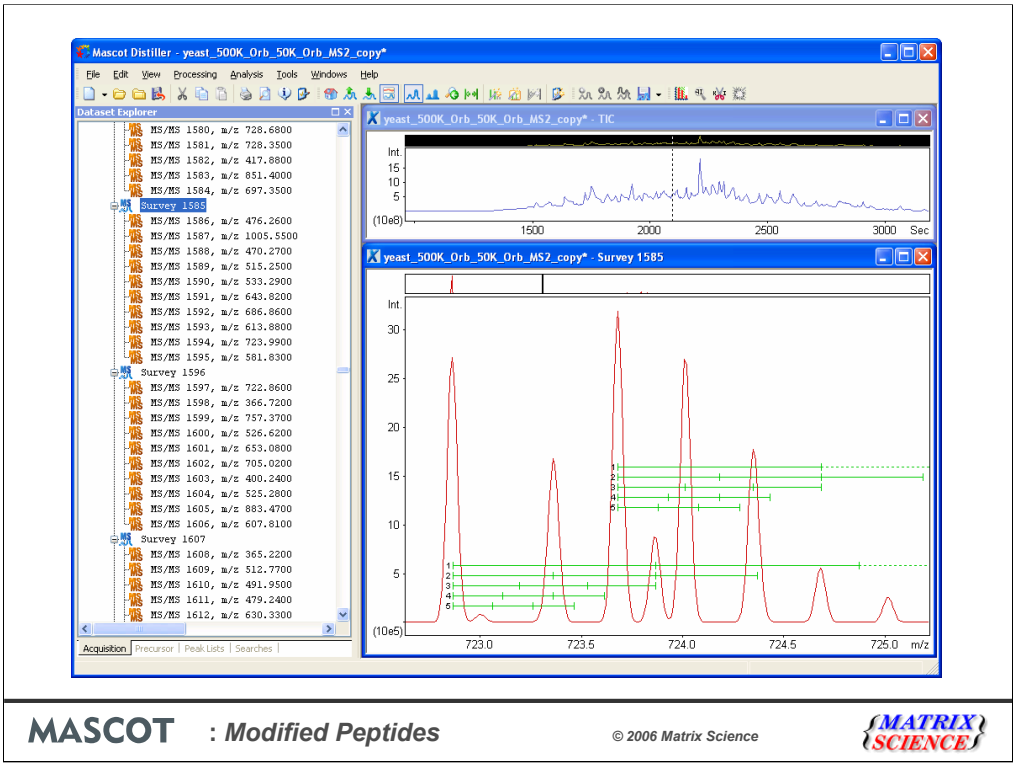
If we click on the hyperlink to see the peptide view …



#	b	b^{++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	132.0478	66.5275			M							13
2	203.0849	102.0461			A	1402.6961	701.8517	1385.6696	693.3384	1384.6855	692.8464	12
3	316.1689	158.5881			L	1331.6590	666.3331	1314.6324	657.8199	1313.6484	657.3279	11
4	445.2115	223.1094	427.2010	214.1041	E	1218.5749	609.7911	1201.5484	601.2778	1200.5644	600.7858	10
5	542.2643	271.6358	524.2537	262.6305	P	1089.5323	545.2698	1072.5058	536.7565	1071.5218	536.2645	9
6	655.3483	328.1778	637.3378	319.1725	I	992.4796	496.7434	975.4530	488.2302	974.4690	487.7381	8
7	770.3753	385.6913	752.3647	376.6860	D	879.3955	440.2014	862.3690	431.6881	861.3850	431.1961	7
8	933.4386	467.2229	915.4280	458.2177	Y	764.3686	382.6879	747.3420	374.1747	746.3580	373.6826	6
9	1034.4863	517.7468	1016.4757	508.7415	T	601.3053	301.1563	584.2787	292.6430	583.2947	292.1510	5
10	1135.5340	568.2706	1117.5234	559.2653	T	500.2576	250.6324	483.2310	242.1191	482.2470	241.6271	4
11	1272.5929	636.8001	1254.5823	627.7948	H	399.2099	200.1086	382.1833	191.5953	381.1993	191.1033	3
12	1359.6249	680.3161	1341.6143	671.3108	S	262.1510	131.5791	245.1244	123.0659	244.1404	122.5738	2
13					R	175.1190	88.0631	158.0924	79.5498			1

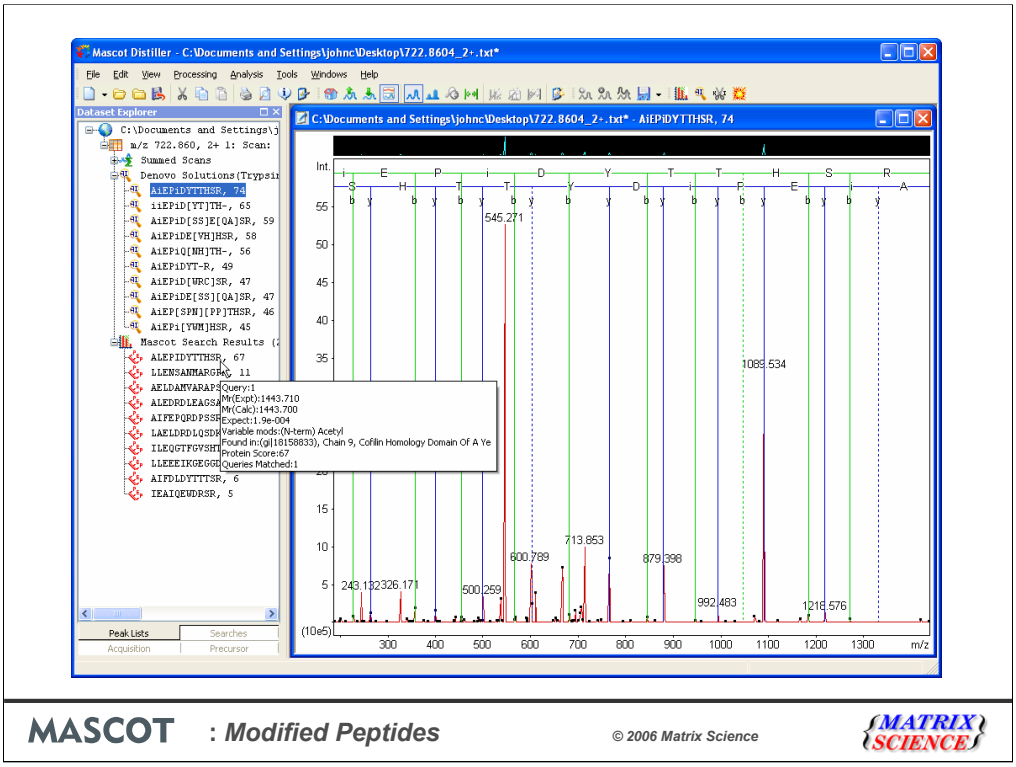
The match was obtained by placing this modification delta on or close to the N-terminus. Remember that this peptide forms the N-terminus of the protein, so in all probability, the initiator Met is lost in the mature protein, making our actual mass delta 765.33. To further complicate the picture, the annotations for this protein report that the N-term Alanine is normally acetylated. If so, our unknown mod is actually 723.32

None of these deltas correspond to anything in Unimod, which is why this match wasn't picked up by the error tolerant search. This looked like a very solid match to me, so I started trying to figure out what the mod might be. Then the penny dropped. The delta is awfully close to the m/z value, which immediately suggests a precursor charge error.



MASCOT : *Modified Peptides* *© 2006 Matrix Science* MATRIX SCIENCE

If we go back to the survey scan, this is what we find. A 3+ peptide at 723.6803 and a 2+ peptide at 722.8604. The instrument thought it was going for 723.99, so Distiller used the 3+ peptide, which is both the closest and the more intense. If we take the correct mass and charge …



MASCOT : *Modified Peptides* *© 2006 Matrix Science* MATRIX SCIENCE

We get the correct match from Mascot and even the de novo falls straight out. The initiator Met has indeed been removed and the Alanine acetylated. Not an unknown modification after all, but nice to get to the bottom of a small mystery. The take home message is that de novo plus sequence tag can often take you further than an error tolerant search of the uninterpreted data.

Search strategy

1. **Standard Mascot search**
   - Returns the easy matches
2. **Error tolerant search**
   - Returns additional matches, but only for proteins where we have at least one good peptide match already
   - Limited to a single additional SNP or modification per peptide
3. **De novo**
   - If data very high quality, can return novel full-length peptide sequences
     - Use Blast to find likely parent proteins
   - More often, returns partial / ambiguous peptide sequences
4. **Error tolerant tag search**
   - To find matches to
     1. Isolated peptides that have a SNP or unsuspected modification
     2. Peptides with multiple SNPs or unsuspected modifications
   - (No reason to expect additional matches from a standard tag search)

So, there we have it. Four powerful tools to help us find modified peptides. The challenge going forward is to make the workflow more integrated. It is still a bit too manual for large data sets. The other thing we need to address is the speed of an error tolerant tag. It would be great if we could find a way to speed this up so as to allow them to be fired off automatically for every de novo solution