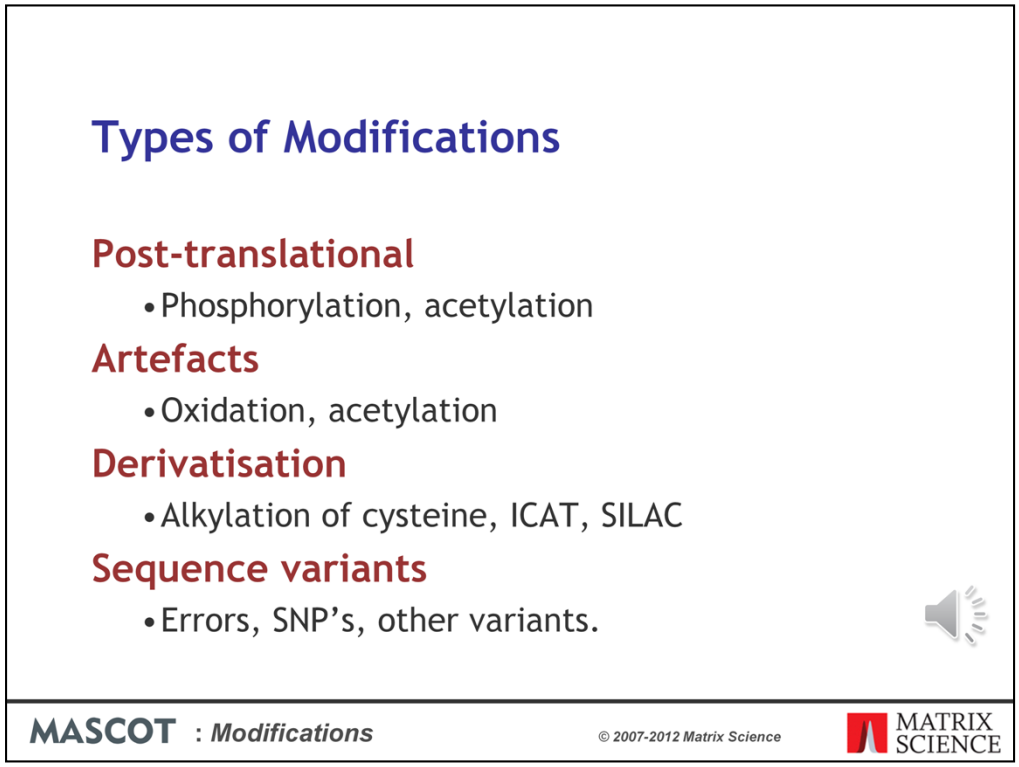
Modifications

MASCOT

MATRIX SCIENCE



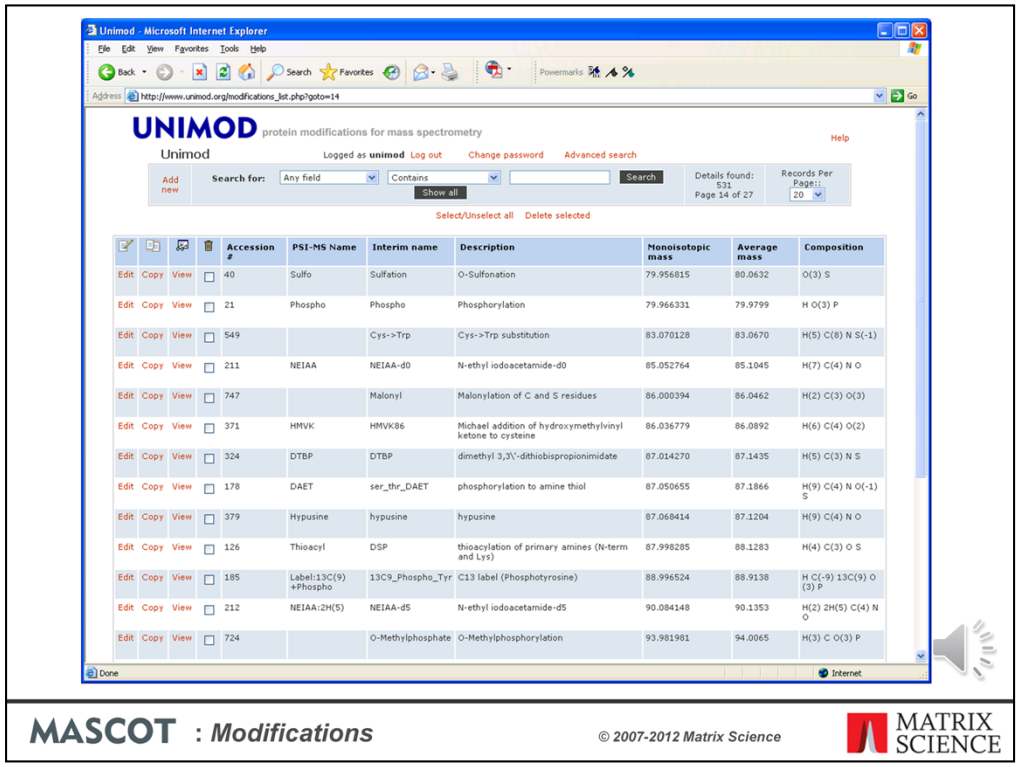
Modifications are a very important topic in database searching.

In some cases, the main focus of a study is to characterise post translational modifications, which may have biological significance. Phosphorylation would be a good example.

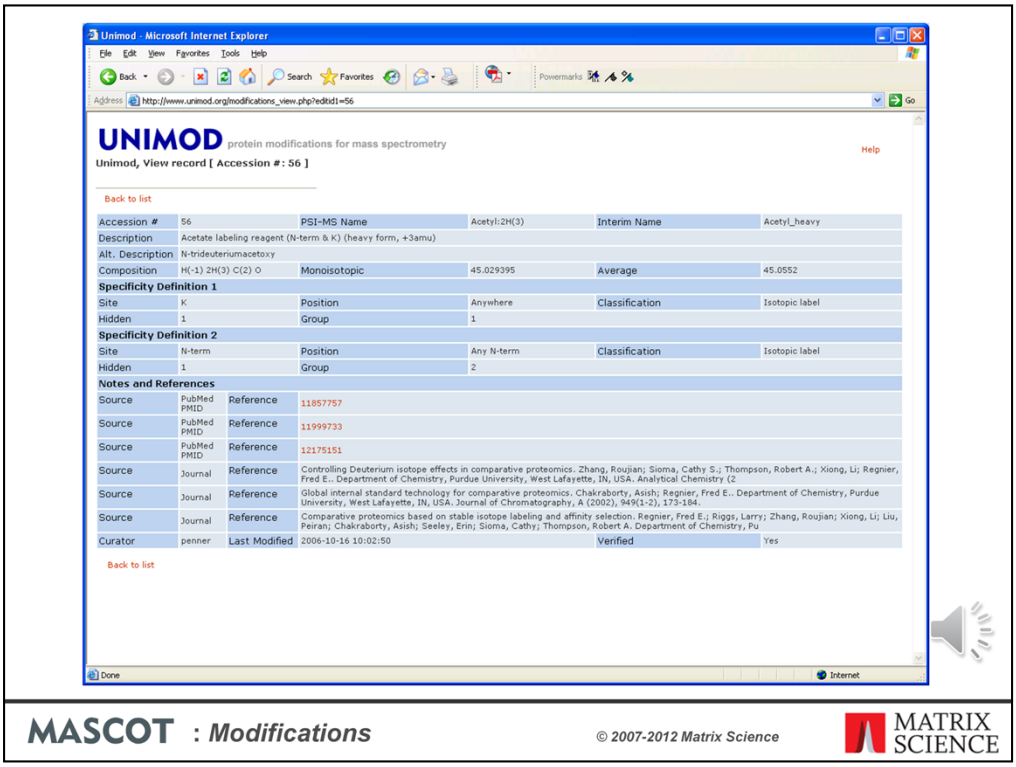
In other cases, the modification may not be of interest in itself, but you need to allow for it in order to get a match. Oxidation during sample preparation would be an example.

And, of course, many methods of quantitation involve modifications containing isotopic labels

Some sequence variants, such as the substitution of one residue by another, are equivalent to modifications, and can be handled in a similar way

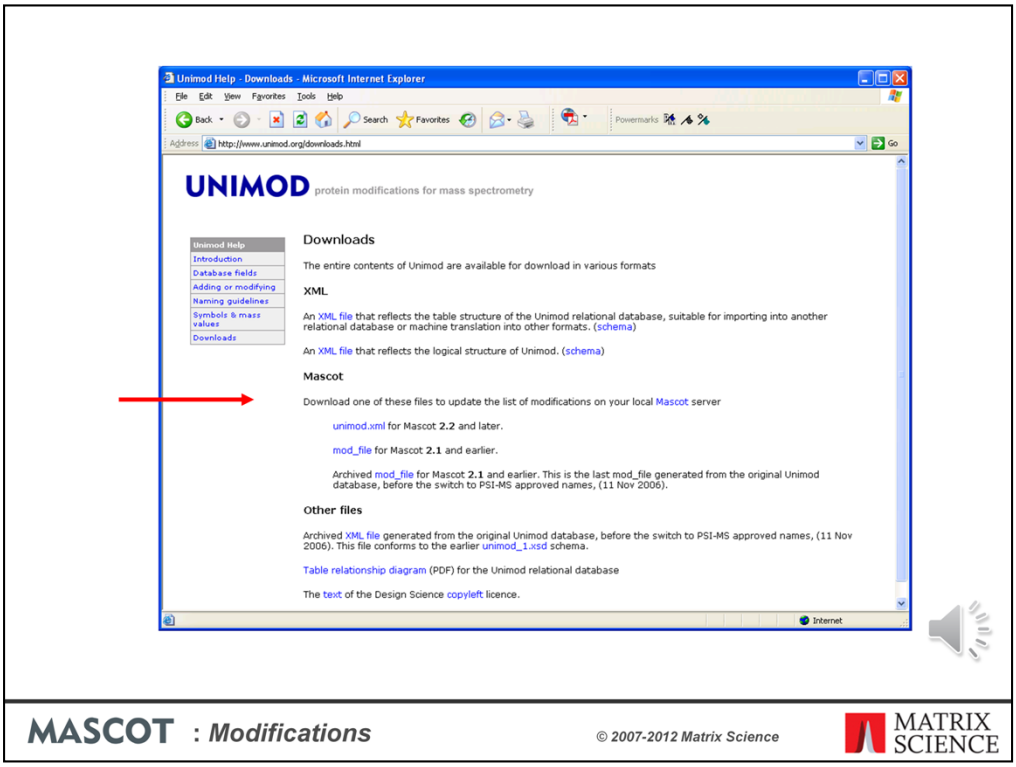


Comprehensive and accurate information about post translational and chemical modifications is an essential factor in the success of protein identification. In Mascot, we take our list of modifications from Unimod, which is an on-line modifications database.

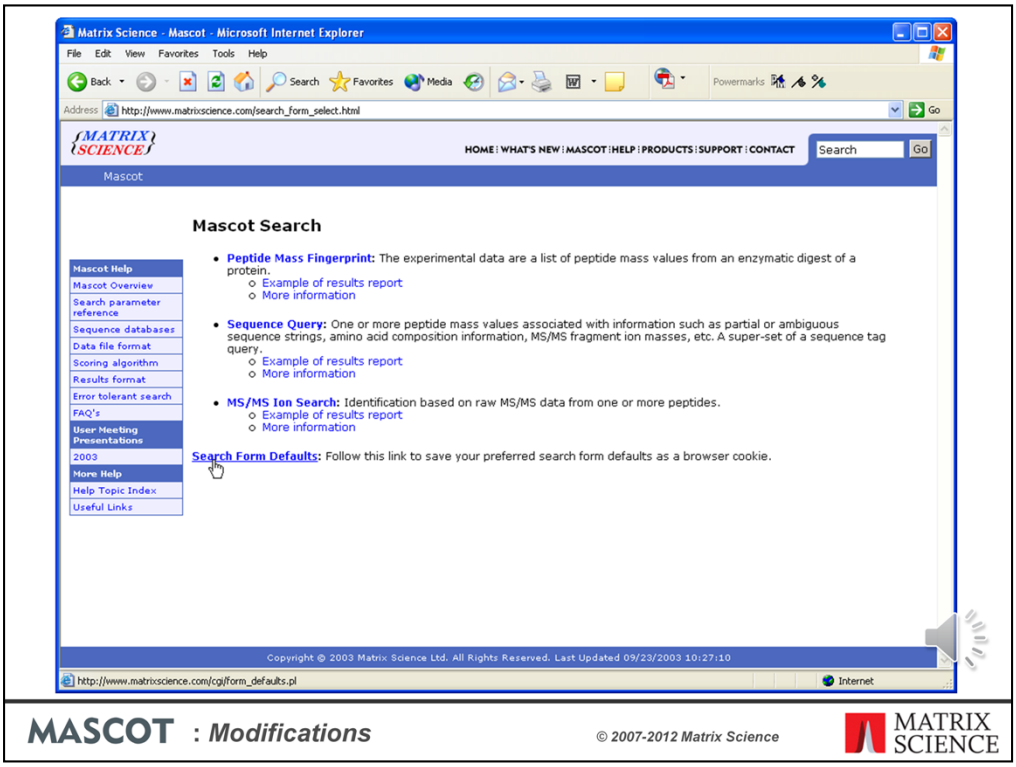


There are other lists of modifications on the web, like DeltaMass on the ABRF web site and RESID from the EBI, but none is as comprehensive as Unimod

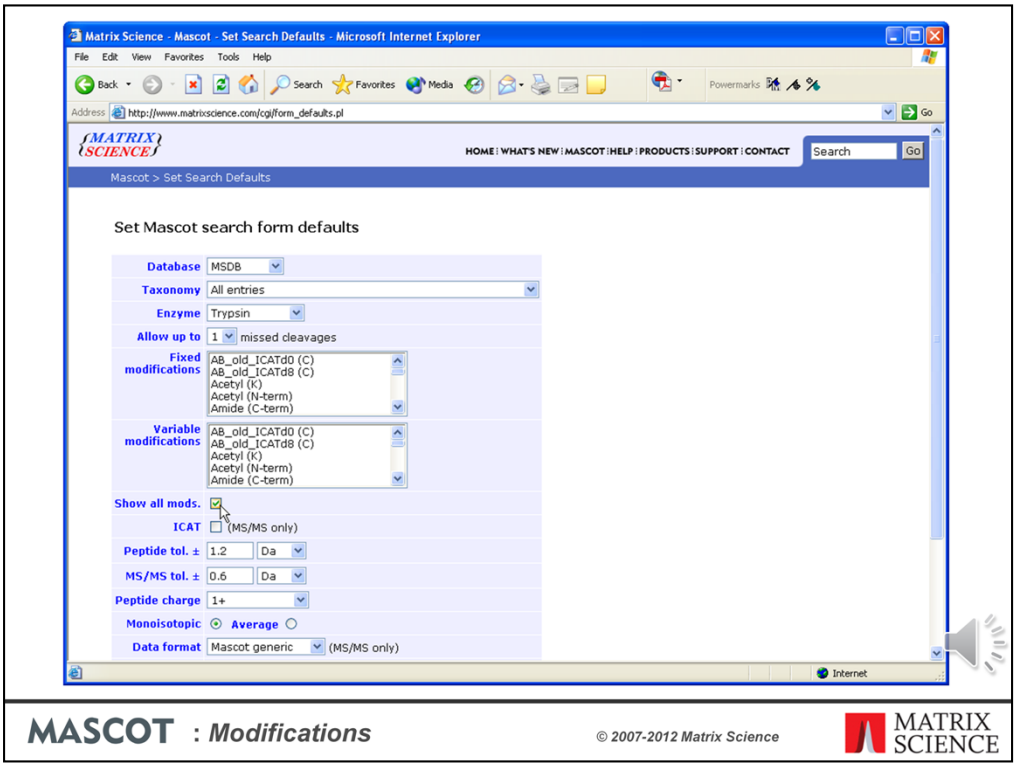
Mass values are calculated from empirical chemical formulae, eliminating the most common source of error. Specificities can be defined in ways that are useful in database searching, and there is the option to enter mass-spec specific data, such as neutral loss information. This screen shot shows one of the better annotated entries, I can't pretend that all of them are this detailed. Nevertheless, it is a very useful, public domain resource that beats having to create your own list in an Excel spreadsheet or on the back of an envelope.



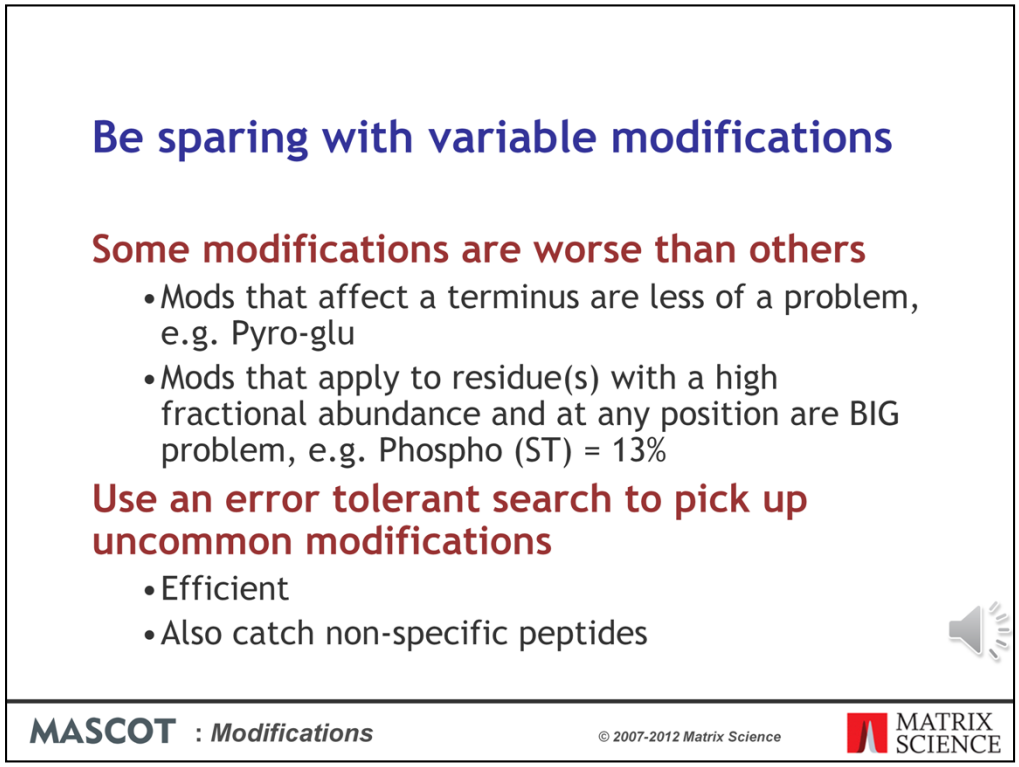
If you go to the help page, there is a link to download the contents of Unimod as a Mascot modifications file. This is the easiest way to keep the modifications list on an in-house Mascot server up-to-date



Here is a tip. The default list of modifications displayed in the Mascot search form is a short list, containing only the most common mods. If you want to see the complete list of mods, and you are using Mascot 2.2 or earlier, you need to follow the link at the bottom of the search form selection page



Check the box for Show all mods, then choose Save. This still sets the default state of the checkbox in Mascot 2.3, but we decided to place the checkbox on the search form, so as to make it easier to swap between the short and long lists.

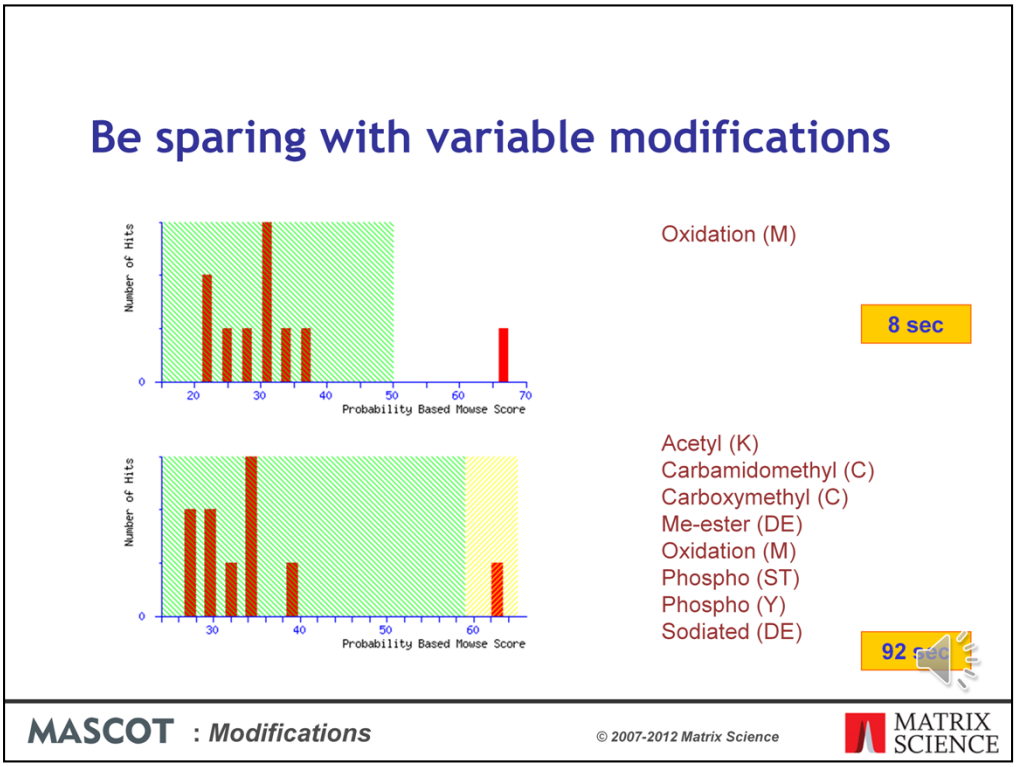


It is extremely important that you do not choose more than the absolute minimum number of variable modification in a search. We talked about this in an earlier presentation, but it is worth repeating.

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.

Some variable modifications are worse than others. Modifications that only apply to a terminus, especially if they only apply when particular residue is at the terminus, like pyro-glu, make little difference to the number of peptides to be tested. The problem modifications are the ones that apply to residues in any position, especially if they apply to multiple residues, like phosphorylation.

Unless you have enriched the sample in a particular PT-mod, e.g IMAC for phosphopeptides, it is usually not a good idea to try and catch PT-mods in a first pass search. Better to use a second pass search, which we call an error tolerant search, to catch the low abundance mods. We will come back to this later.



To illustrate this point. This search of a single MS/MS spectrum, using one variable mod, gives a nice, statistically significant match.

If the search is repeated with 8 mods, the match is the same, with an identical score, but now it is barely significant.

All of these mods have effectively increased the size of the database by a factor of 30

What's worse, the search takes over 10 times as long!

So, use variable mods sparingly. You'll get better results and faster.

By the way, the yellow region in the histogram indicates scores above the homology but below the identity thresholds. You will only see these regions highlighted in an MS/MS search report if it is a search of a single spectrum.

Why is phosphorylation such a challenge?

Site heterogeneity

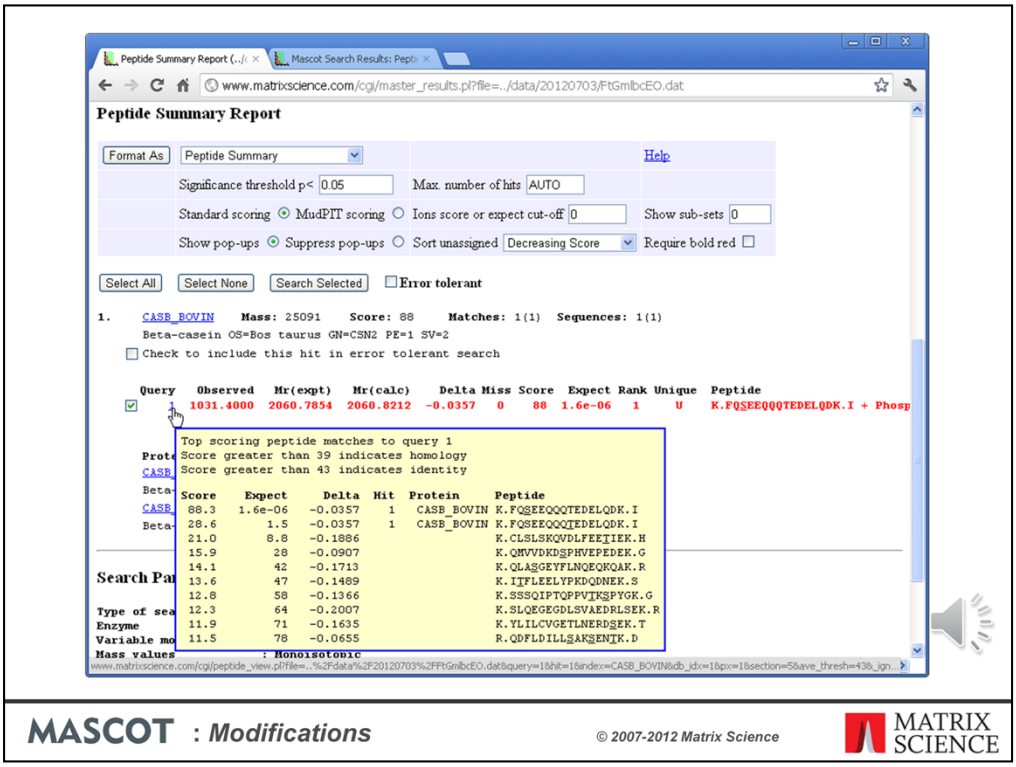
Poor ionisation efficiency

3 fragmentation channels

- intact fragments
- neutral loss of HPO_3 (80 Da)
- neutral loss of H_3PO_4 (98 Da)

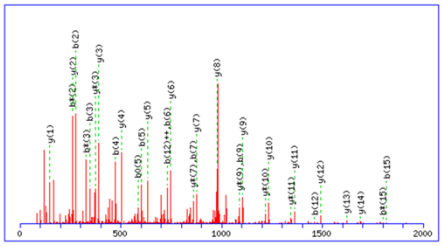
Can occur at STY - ~16% of residues.

Of all post-translational modifications, phosphorylation is one of the most interesting and also one of the most difficult. Why is it such a challenge?



Lets look at an example or two.

One of the most common phosphopeptides comes from the milk protein, beta casein. There are two potential phosphorylation sites, S and T, but only one is modified. Because the two sites are widely separated, the two arrangements get very different scores.



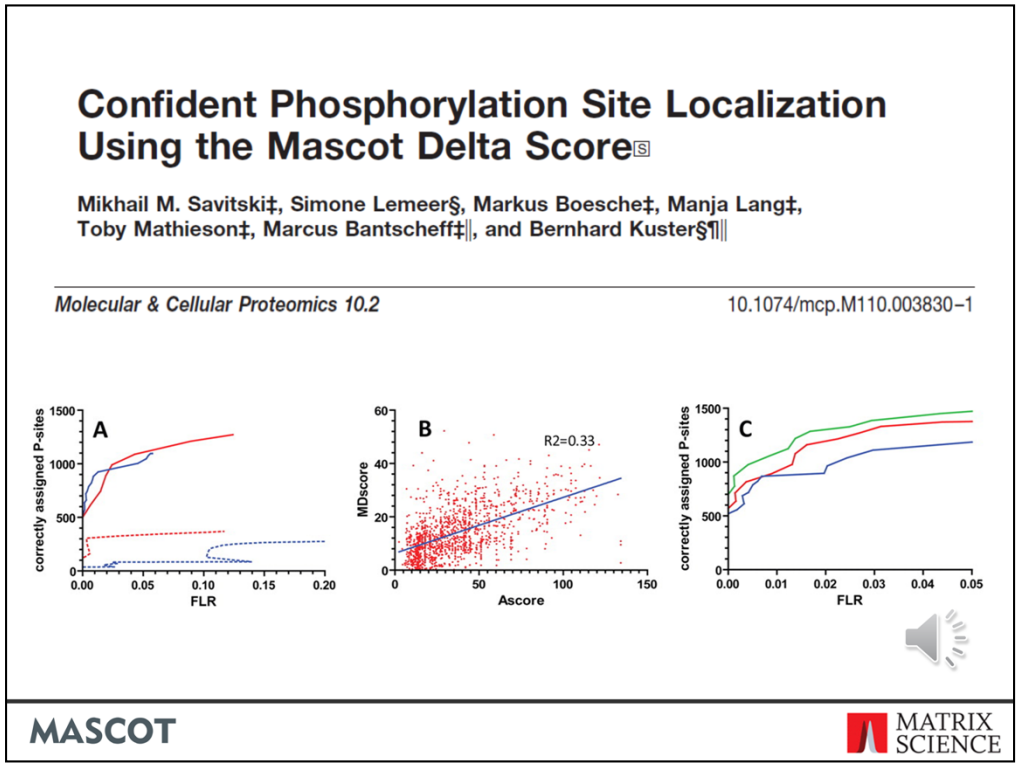
Monoisotopic mass of neutral peptide Mr(calc): 2060.82
Variable modifications:
S3 : Phospho (ST)
Ions Score: 77 **Expect:** 4e-05
Matches (Bold Red): 33/172 fragment ions using 65 most intense peaks

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	148.08	74.54					F							16
2	276.13	138.57	259.11	130.06			Q	1816.78	908.90	1799.76	900.38	1798.77	899.89	15
3	345.16	173.08	328.13	164.57	327.15	164.08	S	1688.72	844.87	1671.70	836.35	1670.71	835.86	14
4	474.20	237.60	457.17	229.09	456.19	228.60	E	1619.70	810.36	1602.68	801.84	1601.69	801.35	13
5	603.24	302.12	586.21	293.61	585.23	293.12	E	1490.66	745.83	1473.63	737.32	1472.65	736.83	12
6	731.30	366.15	714.27	357.64	713.29	357.15	Q	1361.62	681.31	1344.59	672.80	1343.61	672.31	11
7	859.36	430.18	842.33	421.67	841.35	421.18	Q	1233.56	617.28	1216.53	608.77	1215.55	608.28	10
8	987.42	494.21	970.39	485.70	969.41	485.21	Q	1105.50	553.25	1088.47	544.74	1087.49	544.25	9
9	1088.46	544.74	1071.44	536.22	1070.45	535.73	T	977.44	489.22	960.42	480.71	959.43	480.22	8
10	1217.51	609.26	1200.48	600.74	1199.50	600.25	E	876.39	438.70	859.37	430.19	858.38	429.70	7
11	1332.53	666.77	1315.51	658.26	1314.52	657.77	D	747.35	374.18	730.33	365.67	729.34	365.17	6
12	1461.58	731.29	1444.55	722.78	1443.57	722.29	E	632.32	316.67	615.30	308.15	614.31	307.66	5
13	1574.66	787.83	1557.63	779.32	1556.65	778.83	L	503.28	252.14	486.26	243.63	485.27	243.14	4
14	1702.72	851.86	1685.69	843.35	1684.71	842.86	Q	390.20	195.60	373.17	187.09	372.19	186.60	3
15	1817.75	909.38	1800.72	900.86	1799.74	900.37	D	262.14	131.57	245.11	123.06	244.13	122.57	2
16							K	147.11	74.06	130.09	65.55			1

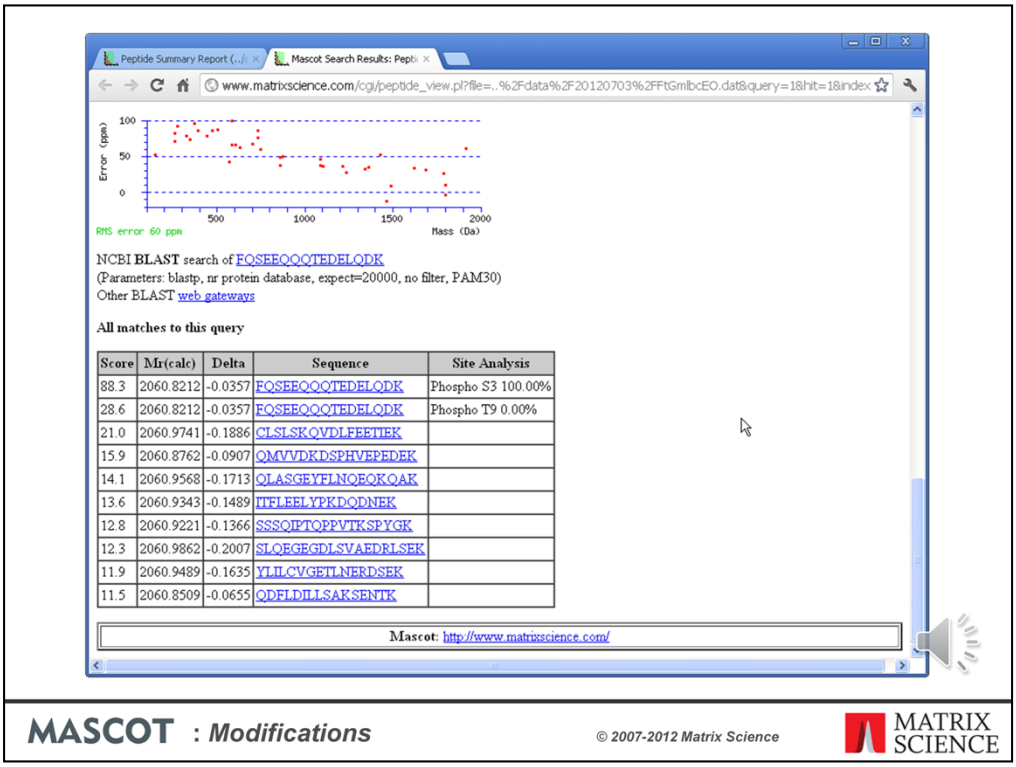
MASCOT : *Modifications*

MATRIX SCIENCE

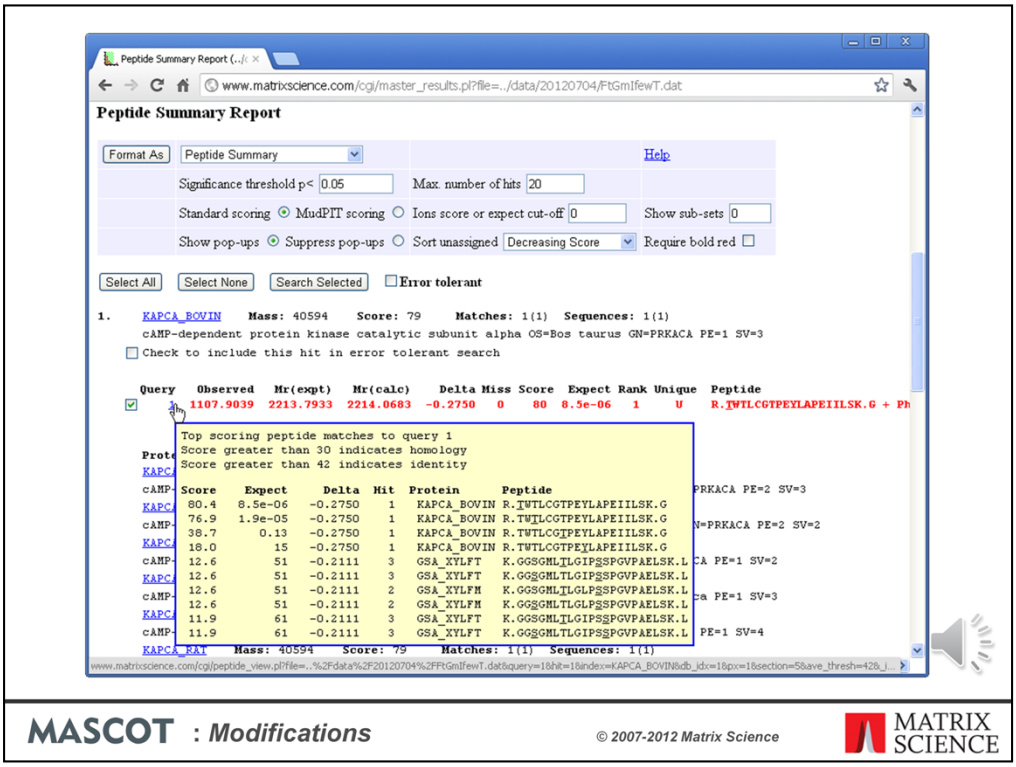
Beautiful spectrum; long run of y ions; move site to T9 and many matches would disappear



Mascot 2.4 reports site localisation probabilities using the delta score method published in MCP by Bernard Kuster's group. They analysed a collection of synthetic analogs of real phosphopeptides and determined what score difference was required to determine the correct site with an error rate of (say) 5%. Because we don't expect everyone to calibrate their data in this way, we have made the calculation slightly more conservative. A score difference of 10 would give approximately 90% probability that the higher scoring arrangement was correct.

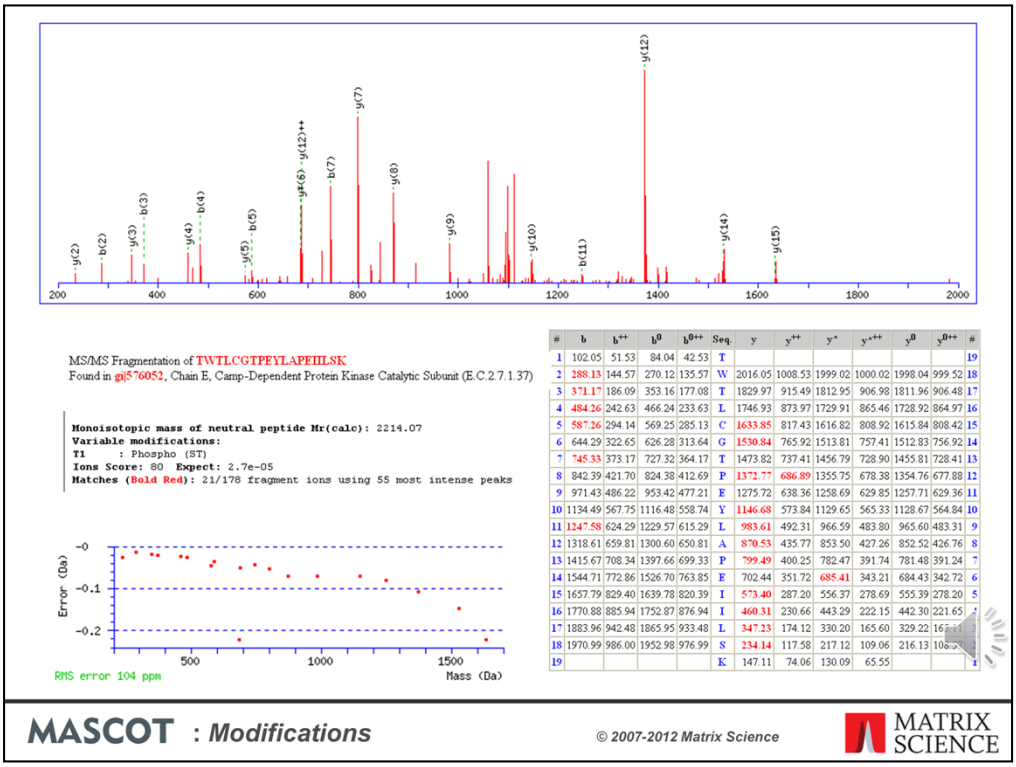


A very large score difference such as the one we were just looking at gives 100% likelihood that the phosphate is on S3.

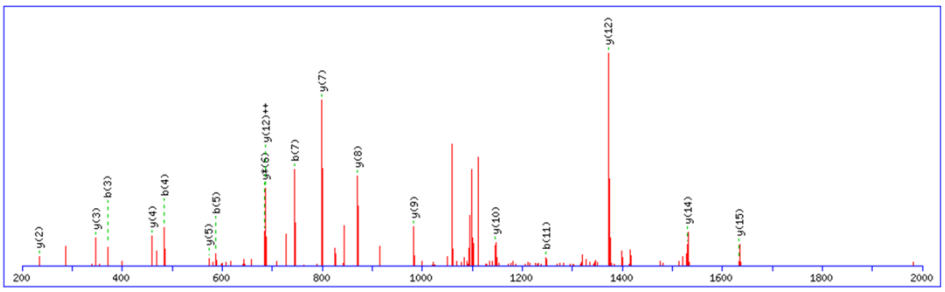


However, casein peptides are unusually easy to analyse. Here is a more typical example of what you can expect to find - a strong match to a phosphopeptide from a protein kinase.

There is little to choose in terms of score between having the phosphate on T1 or T3.

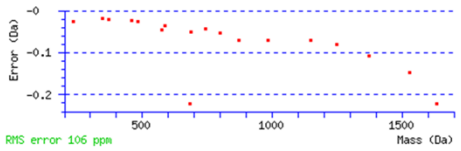


We can see why there is little difference in score between placing the phosphate on T1 or T3. There is just one extra matched peak, and in probability terms, there isn't a huge difference between 20 matches using 55 experimental peaks and 21. However, if you had to choose one or the other, you'd probably go for T1



MS/MS Fragmentation of TWTLCGTPEYLAPEIILSK
Found in gi|576052, Chain E, Camp-Dependent Protein Kinase Catalytic Subunit (E.C.2.7.1.37)

Monoisotopic mass of neutral peptide Mr(calc): 2214.07
Variable modifications:
T3 : Phospho (ST)
Ions Score: 77 **Expect:** 6.1e-05
Matches (Bold Red): 20/178 fragment ions using 55 most intense peaks

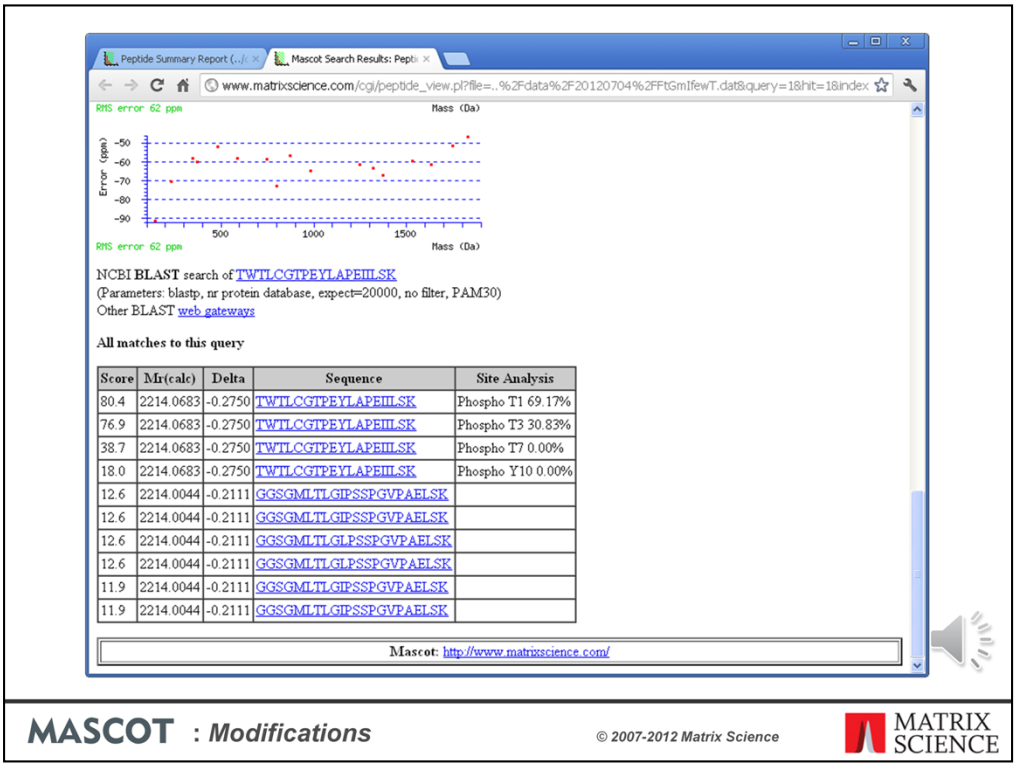


RMS error 106 ppm

#	b	b^{++}	b^0	b^{0++}	Seq.	y	y^{++}	y^*	y^{*++}	y^0	y^{0++}	#
1	84.04	42.53	66.03	33.52	T							19
2	270.12	135.57	252.11	126.56	W	2034.06	1017.53	2017.04	1009.02	2016.05	1008.53	18
3	371.17	186.09	353.16	177.08	T	1847.98	924.49	1830.96	915.98	1829.97	915.49	17
4	484.26	242.63	466.24	233.63	L	1746.93	873.97	1729.91	865.46	1728.92	864.97	16
5	587.26	294.14	569.25	285.13	C	1633.85	817.43	1616.82	808.92	1615.84	808.42	15
6	644.29	322.65	626.28	313.64	G	1530.84	765.92	1513.81	757.41	1512.83	756.92	14
7	745.33	373.17	727.32	364.17	T	1473.82	737.41	1456.79	728.90	1455.81	728.41	13
8	842.39	421.70	824.38	412.69	P	1372.77	686.89	1355.75	678.38	1354.76	677.88	12
9	971.43	486.22	953.42	477.21	E	1275.72	638.36	1258.69	629.85	1257.71	629.36	11
10	1134.49	567.75	1116.48	558.74	Y	1146.68	573.84	1129.65	565.33	1128.67	564.84	10
11	1247.58	624.29	1229.57	615.29	L	983.61	492.31	966.59	483.80	965.60	483.31	9
12	1318.61	659.81	1300.60	650.81	A	870.53	435.77	853.50	427.26	852.52	426.76	8
13	1415.67	708.34	1397.66	699.33	P	799.49	400.25	782.47	391.74	781.48	391.24	7
14	1544.71	772.86	1526.70	763.85	E	702.44	351.72	685.41	343.21	684.43	342.72	6
15	1657.79	829.40	1639.78	820.39	I	573.40	287.20	556.37	278.69	555.39	278.20	5
16	1770.88	885.94	1752.87	876.94	I	460.31	230.66	443.29	222.15	442.30	221.65	4
17	1883.96	942.48	1865.95	933.48	L	347.23	174.12	330.20	165.60	329.22	165.11	3
18	1970.99	986.00	1952.98	976.99	S	234.14	117.58	217.12	109.06	216.13	108.57	2
19					K	147.11	74.06	130.09	65.55			1

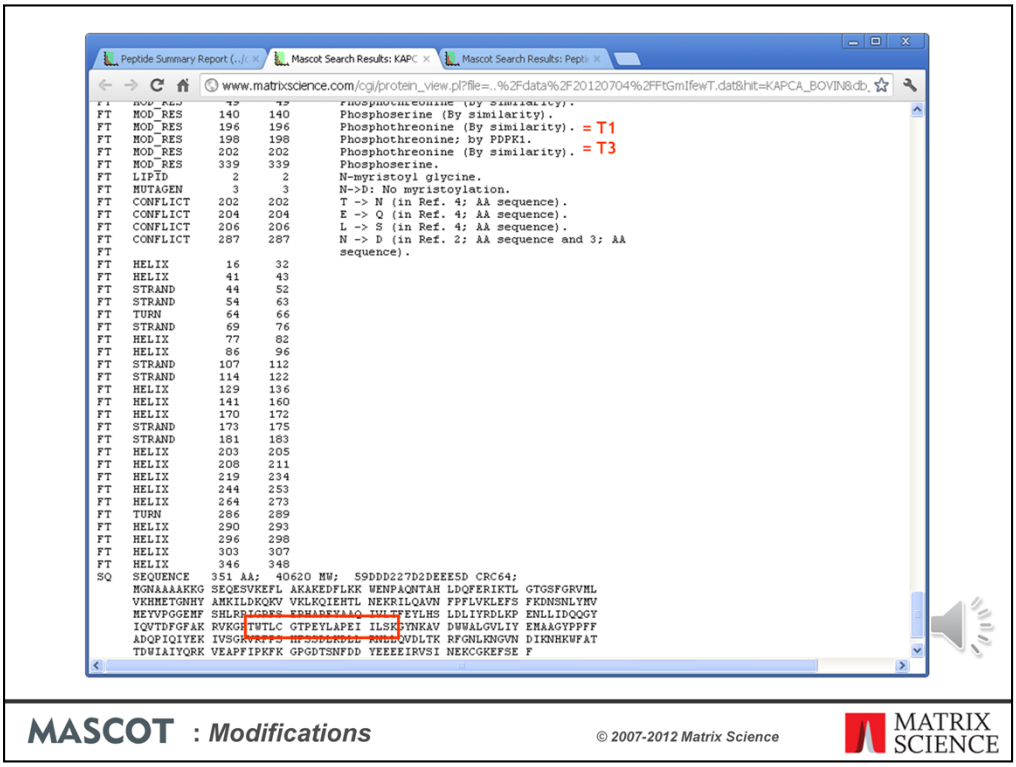
MASCOT : *Modifications*

MATRIX SCIENCE

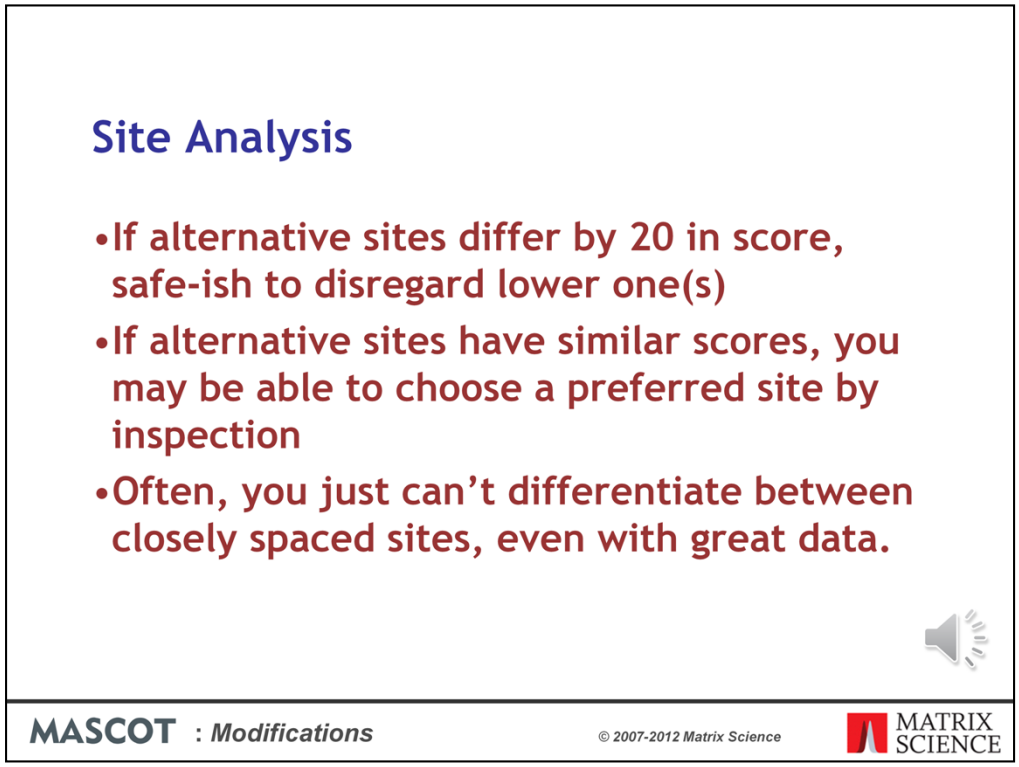


The delta score site analysis suggests 70% probability on T1 and 30% on T3 … much less clear cut. We can't be confident which site is modified, or whether there is a mixture of both isoforms. But, we can be confident it is not on T7 or Y10 because the score drops dramatically, and these are assigned 0% probability.

Sometimes, it is worth looking at the sequence annotations to see whether these are known phosphorylation sites. If the database sequence doesn't have detailed annotations, you can follow the BLAST link to try and match the peptide to an entry from a better annotated database. In this case, we're searching SwissProt, so we can go straight to the protein view report



According to Swissprot, both T1 and T3 are possible phosphorylation sites. If you really needed to know which was the case here, or whether it was a mixture, you'd have to acquire more data. Maybe try a different enzyme or target the incomplete cleavage peptide that includes the preceding KG so as to move the sites towards the centre of the peptide, where you might get stronger b and y fragments

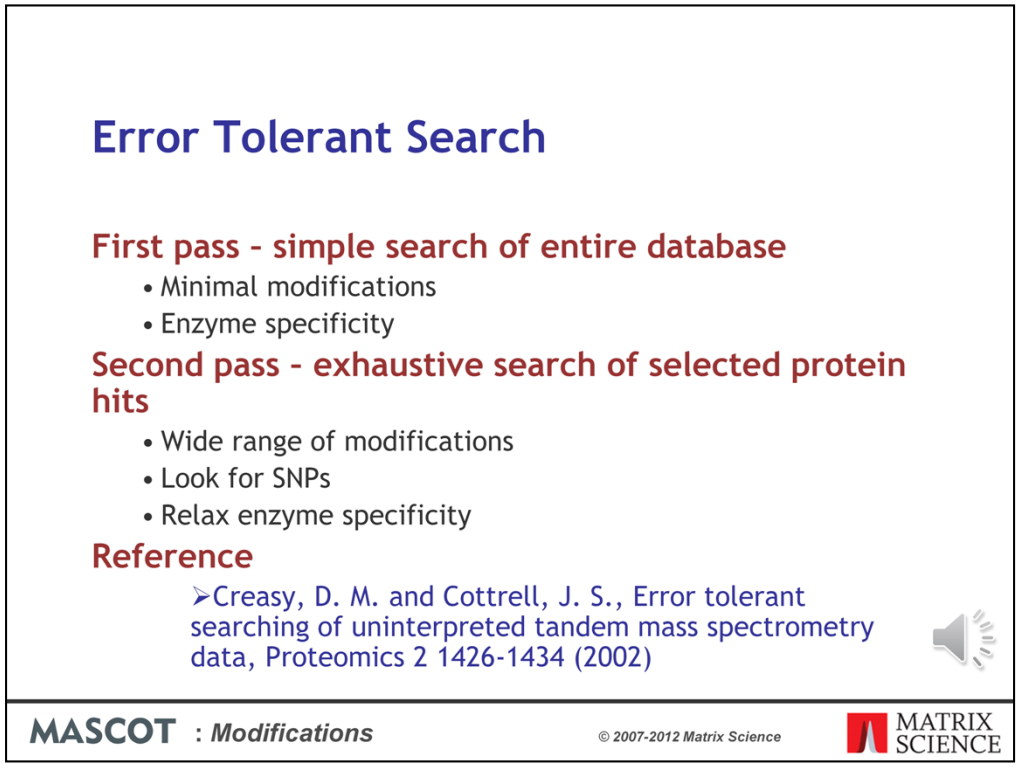


If you are using Mascot 2.3 or earlier, the delta score calculation is not performed in Peptide View. These are our suggested guidelines when using Mascot for site analysis:

If alternative sites differ by 20 in score, safe-ish to disregard lower one(s)

If alternative sites have similar scores, you may be able to choose one by inspection. But, be careful … one peak is just one peak

Often, you just can't differentiate between adjacent sites, even with great data.



Now, back to the challenge of finding PT modifications. There are many hundreds of modifications in Unimod, yet I've emphasised the importance of using the minimum number of variable modifications in a search. So, how are we supposed to find unusual modifications?

If you are searching uninterpreted MS/MS data, the efficient way to find unusual modifications, as well as variations in the primary sequence, is a two pass search. The first pass search is a simple search of the entire database with minimal modifications. The protein hits found in the first pass search are then selected for an exhaustive second pass search. During this second pass search, we can look for all possible modifications, sequence variants, and non-specific cleavage products.

Because only a handful of entries are being searched, search time is not an issue. It would be extremely difficult to calculate meaningful statistics for the additional matches in an error tolerant search, and we don't report expect values. The evidence for the presence of any particular protein are the matches from the first pass search. The additional matches from the second pass search serve to increase coverage and may discover interesting modifications or SNPs.

Error Tolerant Search

Unsuspected chemical & P-T modifications

- Iterate serially through comprehensive list
- All fixed and variable mods retained
- Allow for one additional "unsuspected" modification per peptide

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 rare to get two unsuspected mods on a single peptide

Error Tolerant Search

Primary sequence variants

- Protein database
  - Look for all residue substitutions
  - No attempt to identify single base insertions & deletions because of frame shifts
- Nucleic acid database
  - Look for all single base substitutions, insertions & deletions

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

For a protein database, we can't look for the consequences of inserted or deleted bases, because these give rise to frame shifts, and the entire sequence changes from that point on.

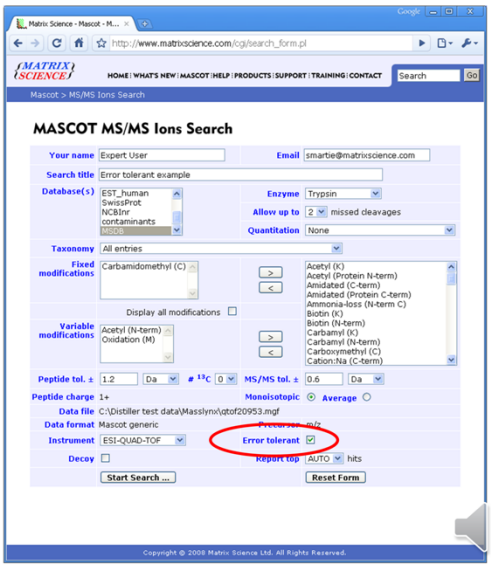
Error Tolerant Search

The following constraints apply to the standard, first pass search:

1. Enzyme must be fully specific
2. A reduced ceiling on the number of variable modifications, (default is 2, but this can be changed globally in mascot.dat or for a user group in Mascot security)
3. Cannot be combined with an automatic decoy database search
4. Cannot be combined with quantitation
5. Search cannot include error tolerant sequence tag

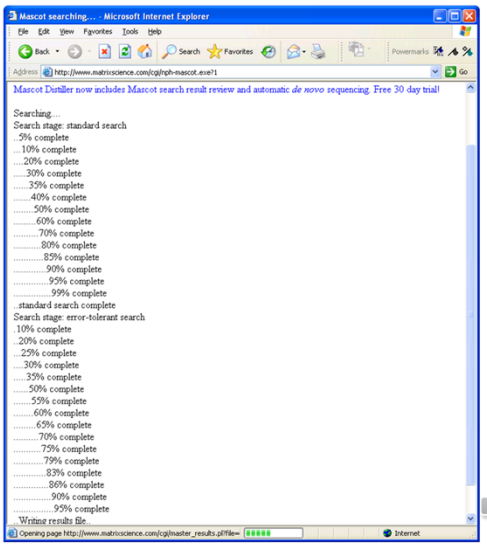
There are some constraints on the standard, first pass search

Error Tolerant Search



Otherwise, submitting the search is just like submitting a standard search except that you check the Error Tolerant Checkbox

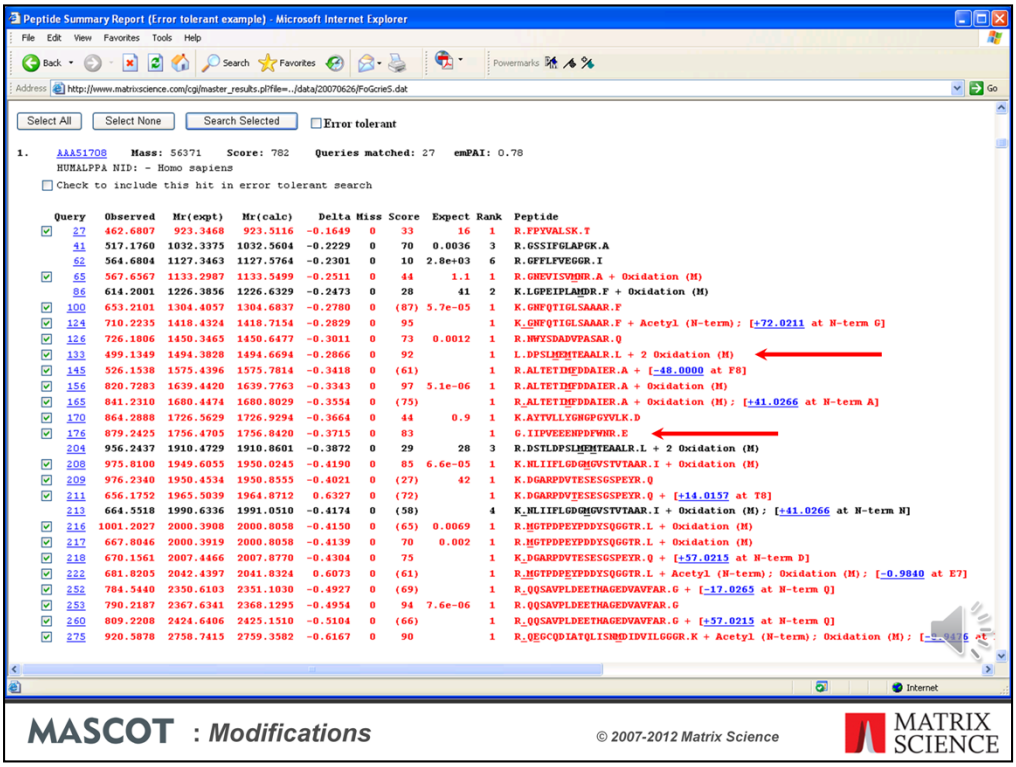
Error Tolerant Search



MASCOT : *Modifications*

MATRIX SCIENCE

You see two sets of progress reports



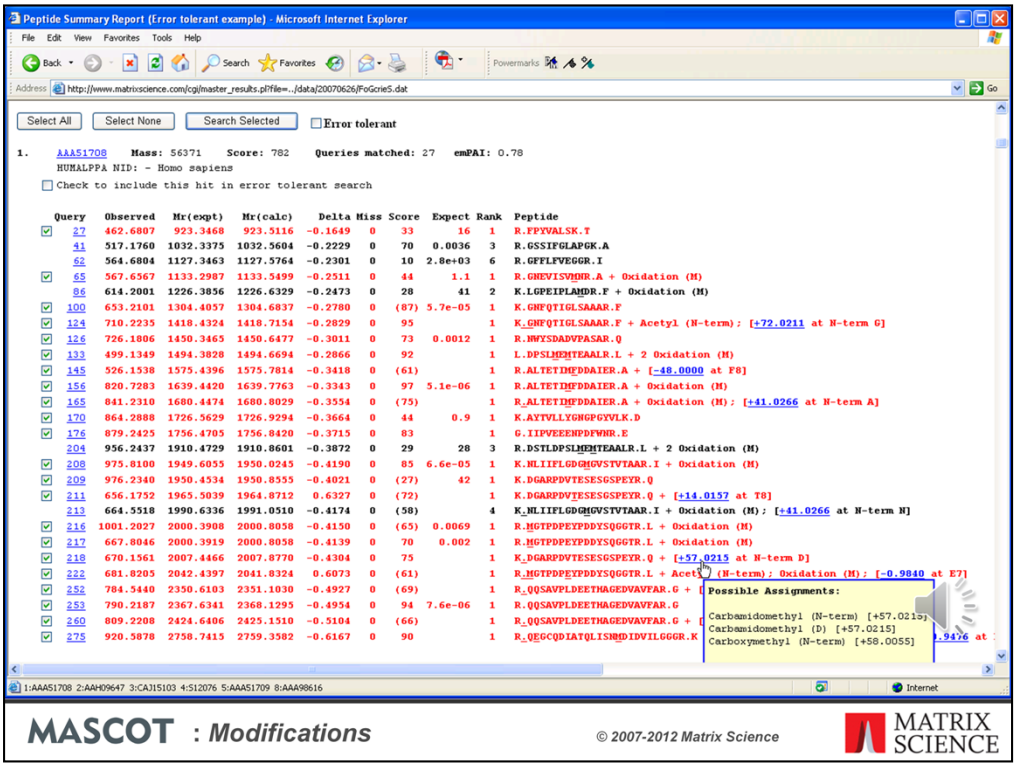
And here is the first hit of the results report. The additional matches, found in the error tolerant search, are the ones without expect values. One of these, query 133, is a simple, non-specific peptide with a very good score. There's another example for query 176. The error tolerant search is a much better way of picking up non-specific peptides than searching the entire database with semi-trypsin or no enzyme. We only fail to get such matches in an error tolerant search if there are no matches to the protein in the first pass search. However, you have to ask yourself whether you would believe a protein hit in which the only peptide match was non-specific. I think the answer is no.

Error Tolerant Search

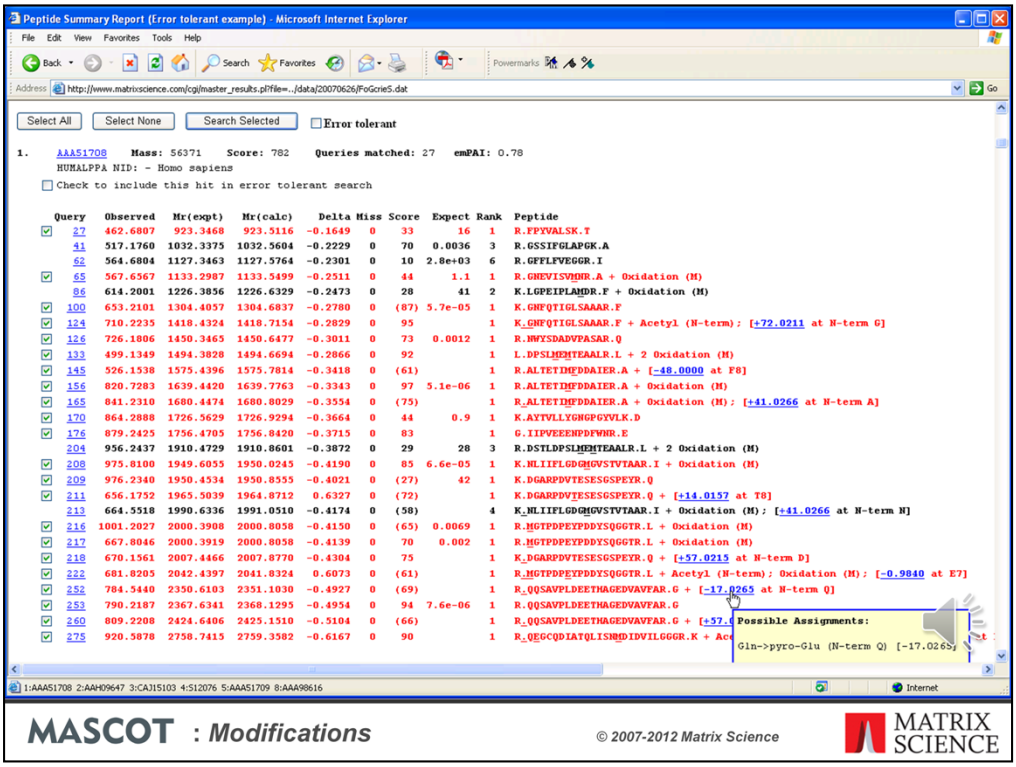
To reduce 'junk' matches

- An individual peptide can be semi-specific OR have one unsuspected modification OR have one primary sequence mutation.
- If the mass delta of the modification is less than the smaller of the precursor mass tolerance and the fragment mass tolerance, the modification is rejected. This eliminates modifications that are meaningless given the estimated mass error, like Q->K, in most cases.
- Match must have a score of at least the identity threshold for the same query in the first pass search
- Match must have a score in excess of the highest scoring match to the same query in the first pass search

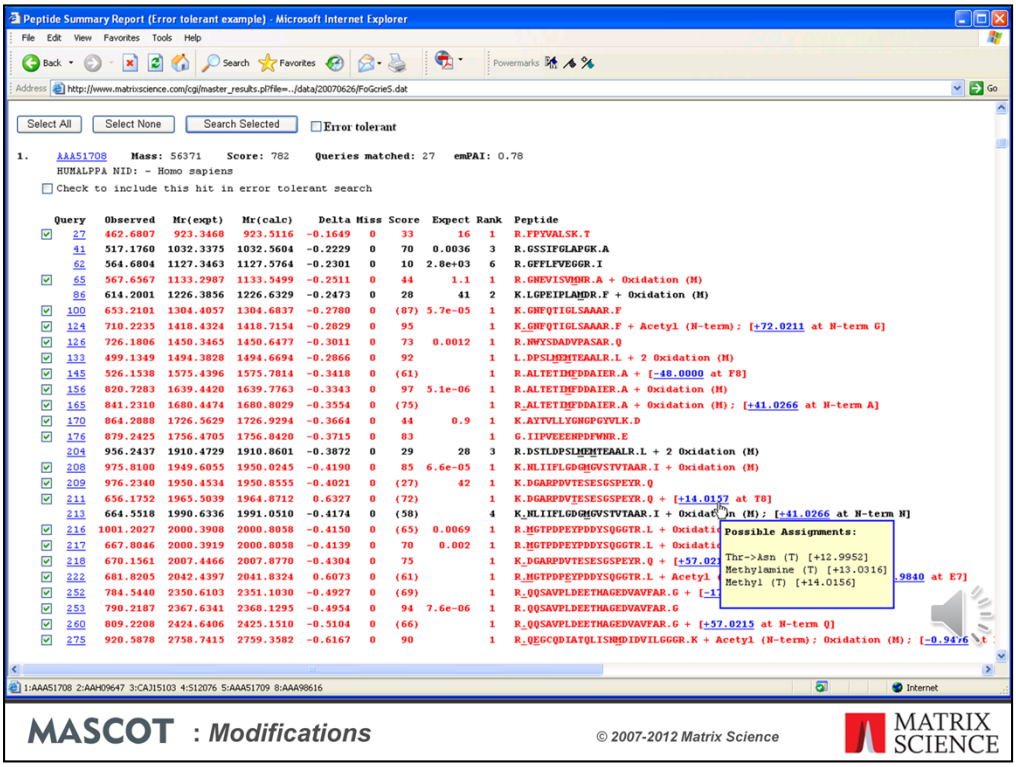
The matches from an error tolerant search are aggressively filtered to remove junk matches



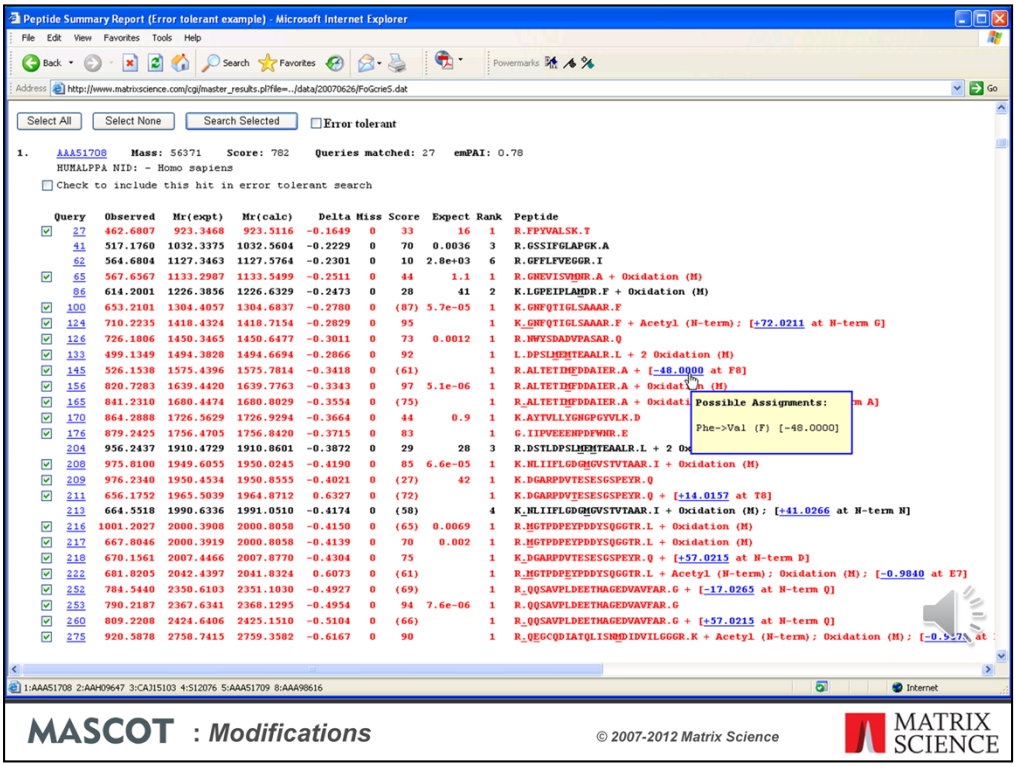
Take a look at the match to query 218. The mass tolerance for this search was fairly wide, so the observed mass difference could correspond to either carbamidomethylation or carboxymethylation at the N-terminus. Since this sample was alkylated with iodoacetamide, we would choose carbamidomethylation as the more likely suspect, especially as this brings the error on the precursor mass into line with the general trend, whereas carboxymethylation would give an error of +0.6 Da. The assignment to carbamidomethylation is also very believable, because this is a known artefact of over-alkylation. The same modification is found for query 260.



Another easily believable assignment is pyro-Glu for the match to query 252.

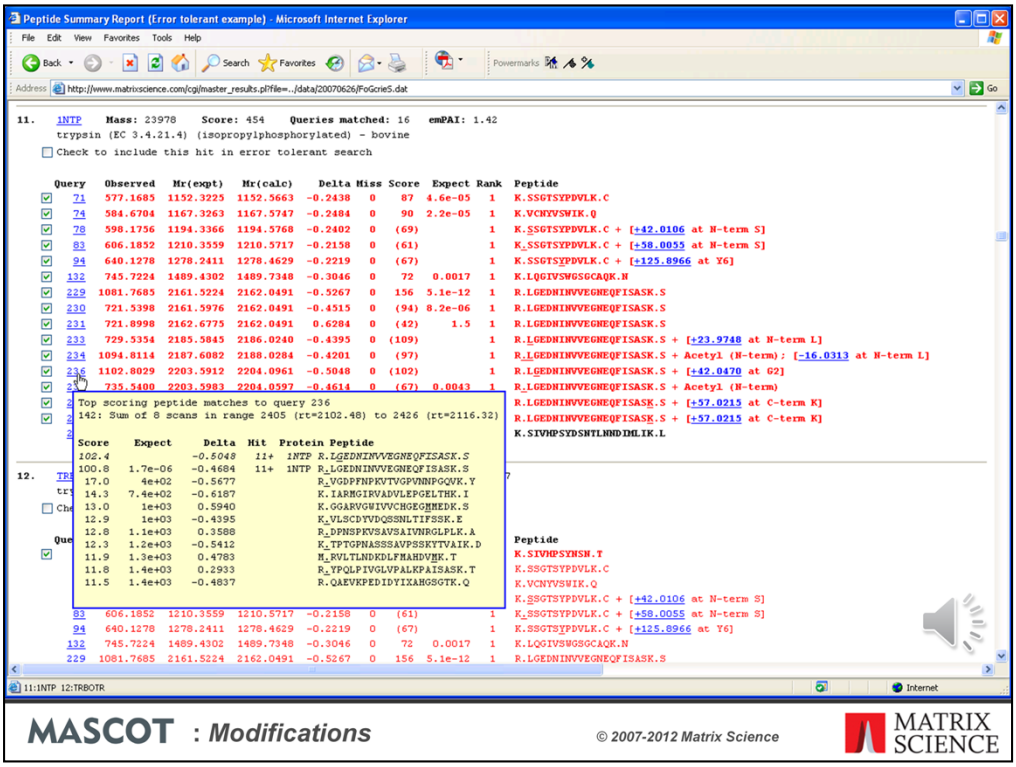


As is methylation ay T8 for query 211

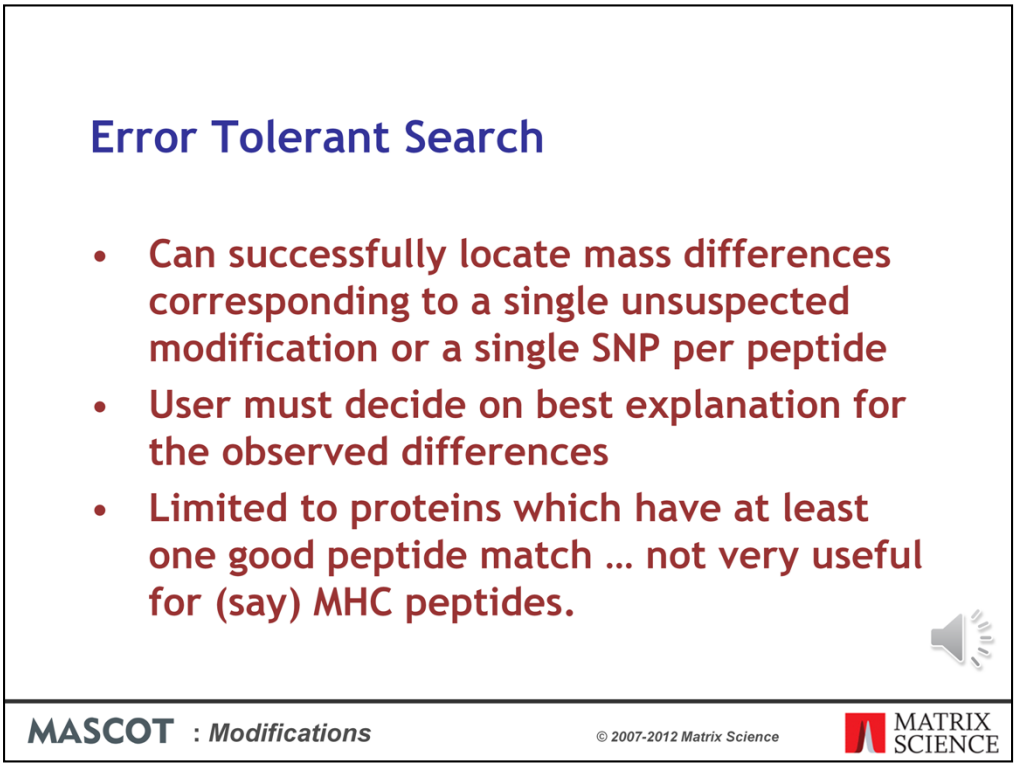


In other cases, the match may be good, but the assignment is not believable. Query 145 is listed with a substitution at F8 causing a loss of 48 Da. This seems unlikely because we have 2 other matches to the same peptide without any substitution. What else could it be? Well, notice that the other two matches are both oxidised at M7. If we suppose this peptide is also oxidised, then the mass shift becomes -64, which is a well-known loss for oxidised methionine, (loss of methanesulfenic acid). This would seem a much more likely explanation for this match.

It is important to understand that the error tolerant search finds new matches by introducing mass shifts at different positions in the database sequences. The match may be very strong, but figuring out a credible assignment can require a bit of detective work.



You should also look at the other yellow pop-up when trying to decide whether to accept a match or not. In this example, the error tolerant search was able to get a slightly higher score by shifting a modification of +42 Da from the amino terminus to the adjacent glycine. However, as score increase of 2 in 100 is negligible. Much more believeable to take the original match from the first pass search, which can be explained as N-terminal acetylation.



In summary, an error tolerant search

•Can successfully locate mass differences corresponding to a single unsuspected modification or a single SNP per peptide

•User must decide on best explanation for the observed differences

•Limited to proteins which have at least one good peptide match … not very useful for (say) MHC peptides