Modifications

Types of Modifications

Post-translational

- Phosphorylation, acetylation

Artefacts

- Oxidation, acetylation

Derivatisation

- Alkylation of cysteine, ICAT, SILAC

Sequence variants

- Errors, SNPs, other variants.

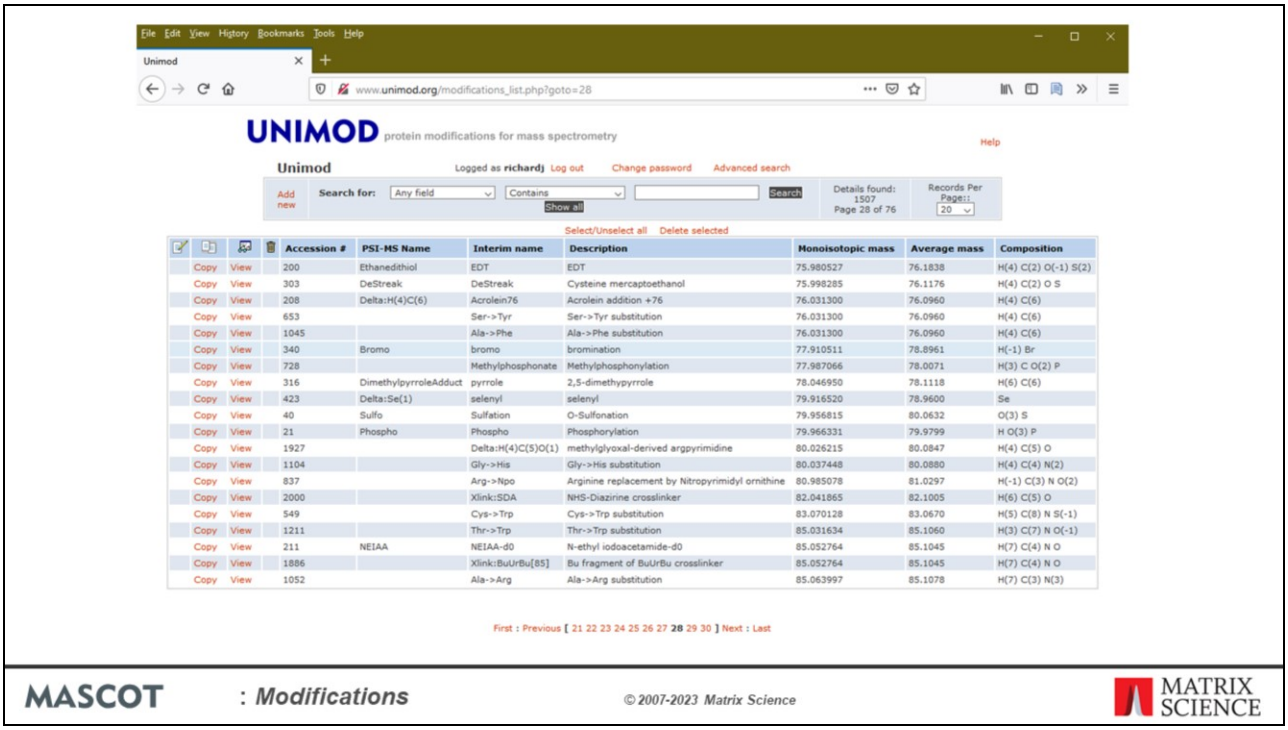
Modifications are a very important topic in database searching.

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.

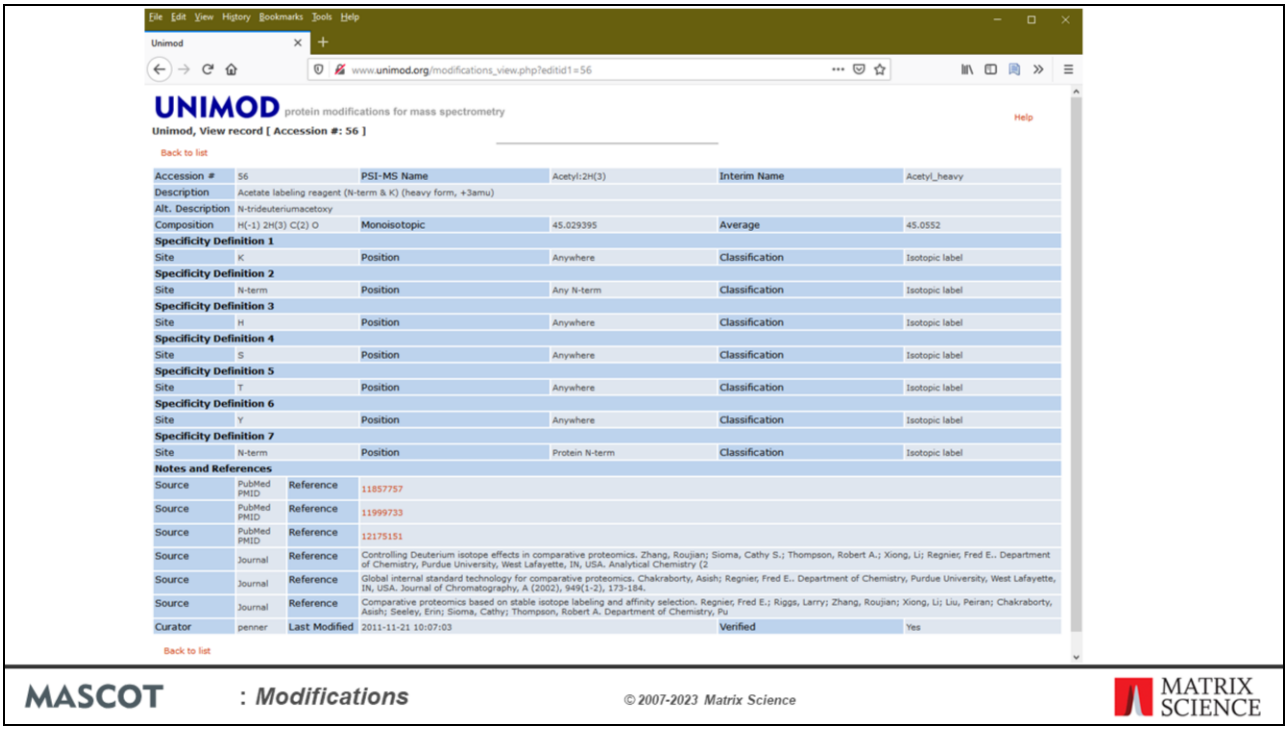
In other cases, the modification may not be of interest in itself, but you need to allow for it in order to get a match. Oxidation during sample preparation would be an example.

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.

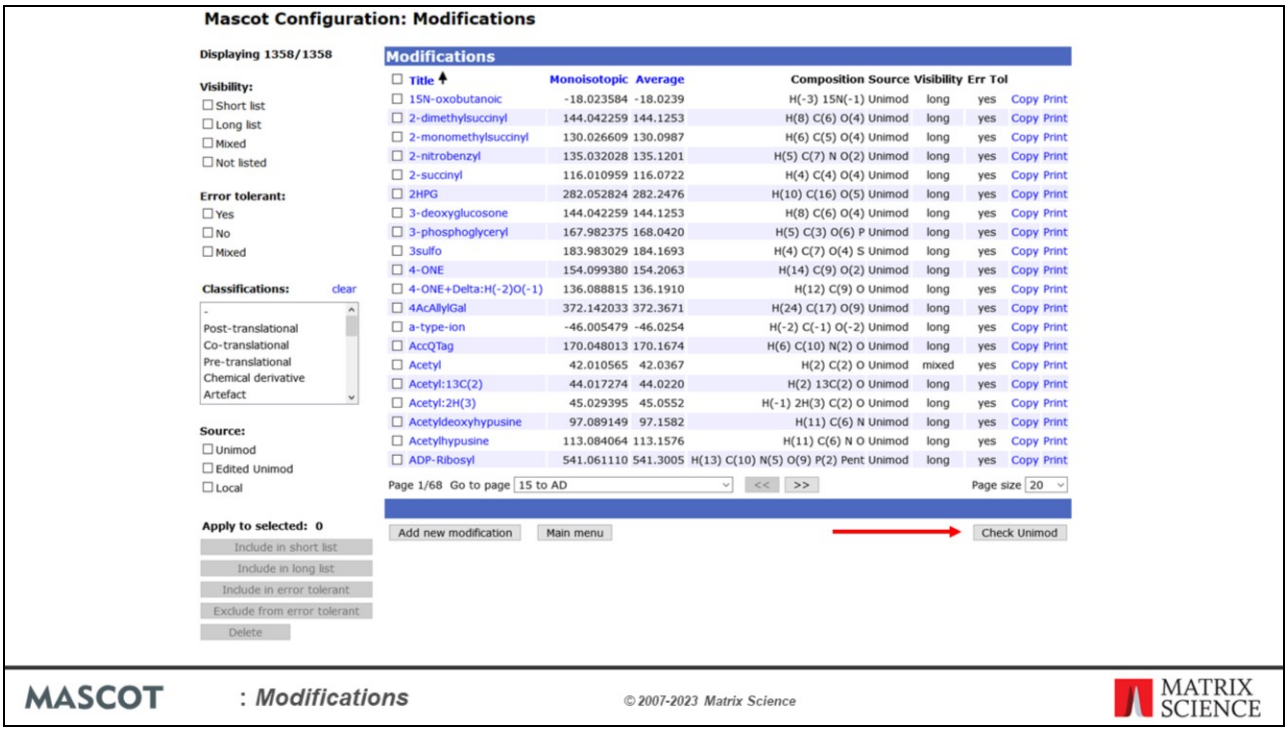


Comprehensive and accurate information about post translational and chemical modifications is an essential factor in the success of protein identification. In Mascot, we take our list of modifications from Unimod, which is an on-line modifications database.

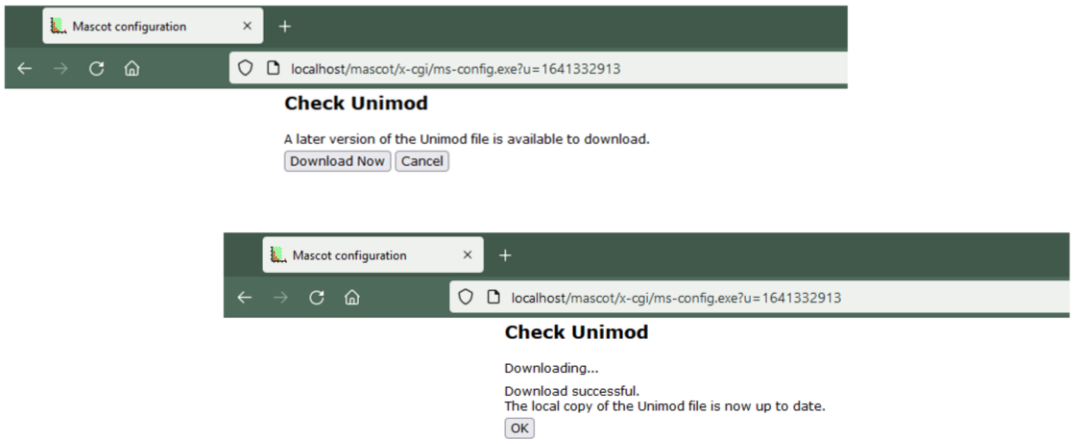


There are other lists of modifications on the web, like DeltaMass on the ABRF web site and RESID from the EBI, but none is as comprehensive as Unimod.

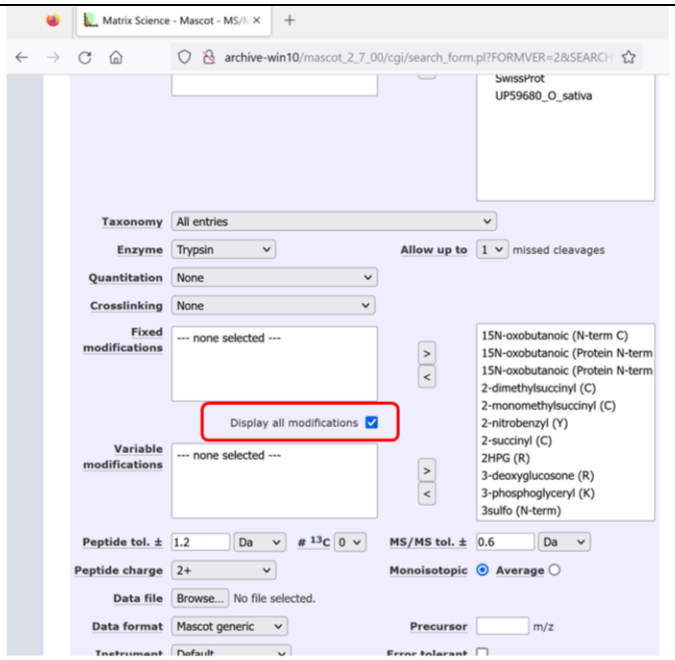
Mass values are calculated from empirical chemical formulae, eliminating the most common source of error. Specificities can be defined in ways that are useful in database searching, and there is the option to enter mass-spec specific data, such as neutral loss information. This screen shot shows one 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 Mascot Server Modification editor, there is a link to check to see if there is an updated Unimod file.

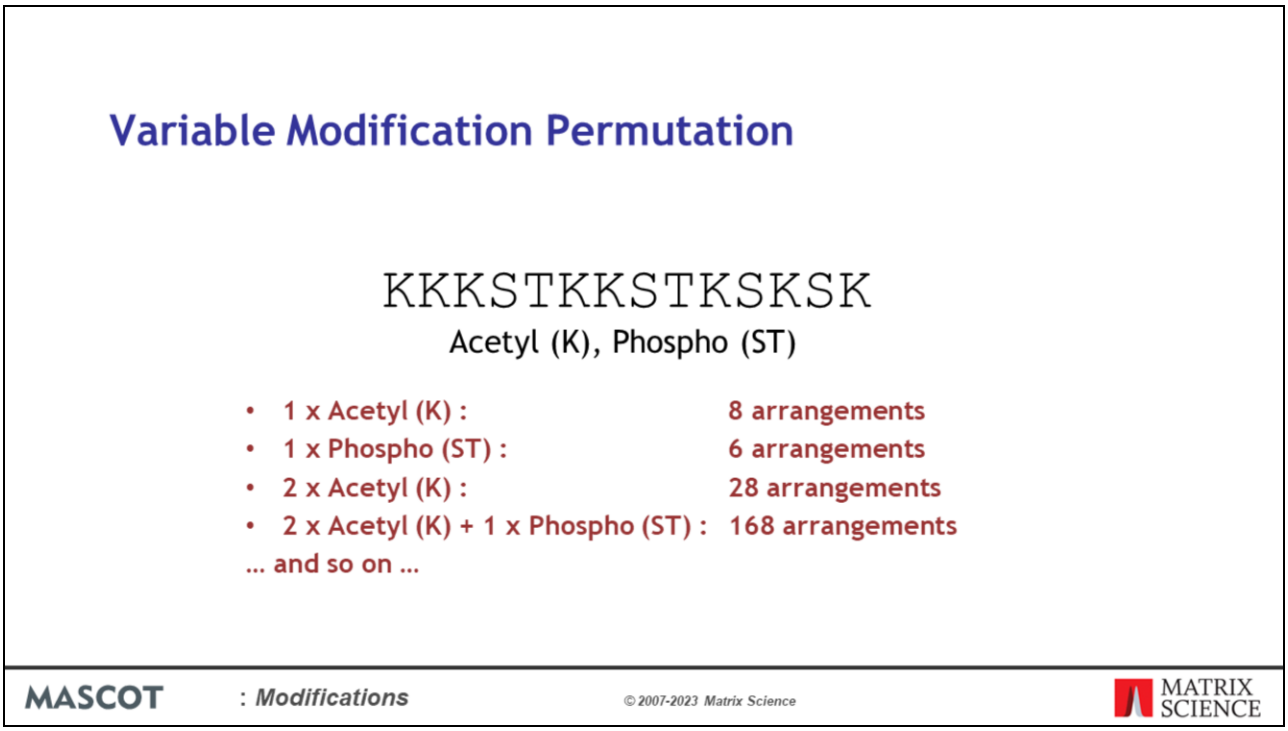


If there is a newer version available, click on the “Download Now” button and “OK” once the download is complete. This is the easiest way to keep the modifications list on an in-house Mascot server up-to-date. Note that updating the Unimod modifications does not affect or change your local modifications.



MASCOT : *Modifications* MATRIX SCIENCE

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, check the Display all modifications box.



Now let's define what it is the software is having to do when looking for modified sites.

Let's consider this slightly unusual peptide comprised of 8 Lysine, 4 serine and 2 threonine residues as an example. If we carried out a search with Lysine acetylation and Phospho Serine & Threonine:

If the precursor mass and tolerance allowed for a single Lysine acetylation, there are just 8 possible arrangements of this.

Likewise, if a single phosphorylation was possible, there are just 6 possible arrangements. However, the number of possible arrangements increases rapidly as we consider more modifications and modifiable sites.

If we need two lysine residues to be acetylated to match the precursor, we now have 28 possible arrangements.

Add in a single phosphorylation and we now have the possible 28 acetylation sites combined with 6 possible phosphorylation sites for a grand total of 168 possible arrangements.

This is the so-called combinatorial explosion, and it's one reason why some searches with a large number of frequent modifications can take a long time.

Modification iterator

Single, consistent, permutation method

- No switching between methods

Controlled by 3 user definable parameters:

- MaxPepNumVarMods
  Max no. of different variable modifications per peptide
- MaxPepNumModifiedSites
  Max no. of modified residues per peptide
- MaxPepModArrangements
  Max no. of arrangements of an individual varmod composition

We use a single, consistent permutation method – there's no switching between different methods. The permutation iterator samples arrangements using a uniformly random scheme. The operation of this is controlled by 3 user definable settings.

MaxPepNumVarMods – this specifies the maximum number of different variable modifications which can be applied to a peptide

MaxPepNumModifiedSites – this specifies the maximum number of residues which can be modified on a peptide

MaxPepModArrangements – "this specifies the maximum number of arrangements of an individual variable modification composition to test"

Modification iterator

Defaults chosen to give balanced speed and depth of search

- MaxPepNumVarMods 3
- MaxPepNumModifiedSites 5
- MaxPepModArrangements 64

Two main cases for changing the defaults

- Decrease limits to reduce search time
- Increase limits to improve site analysis, or if you're looking at a highly modified protein

Mascot 2.8 ships with the following default values for these parameters:

MaxPepNumVarMods 3, MaxPepNumModifiedSites 5 and MaxPepModArrangements 64.

There are two main cases where you might want to change these defaults. Decreasing any of these values will reduce the search space, as fewer arrangements will be tested. This will decrease the search time – so if you're looking at a sample which is not highly modified and where definitive site analysis is not the aim of the study, you may wish to decrease some of these values.

However, if site analysis is important, or if you're looking at a highly modified protein such as Histone, then you may need to increase these limits in order to gain accurate modification localisation results.

Be sparing with variable modifications

Some modifications are worse than others

- Mods that affect a terminus are less of a problem, e.g. Pyro-glu
- Mods that apply to residue(s) with a high fractional abundance and at any position are BIG problem, e.g. Phospho (ST) = 13%

Use an error tolerant search to pick up uncommon modifications

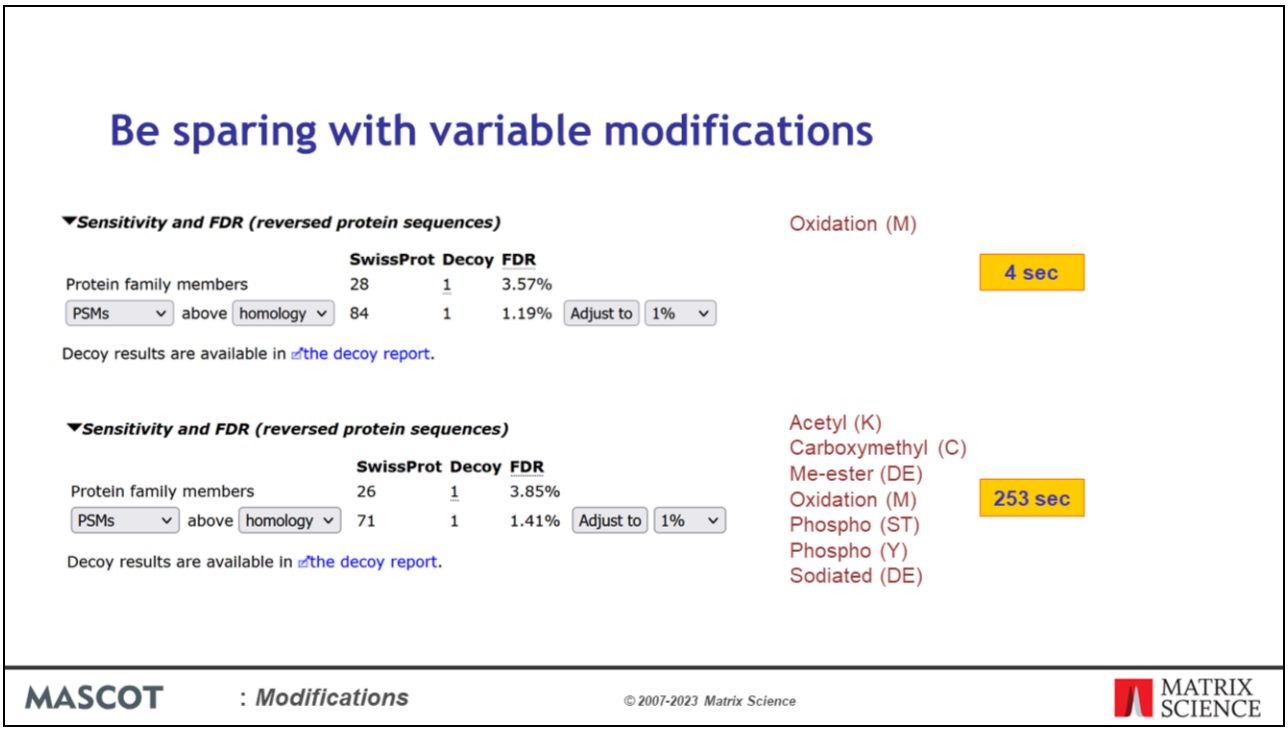
- Efficient
- Also catch non-specific peptides

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, 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.

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 the error tolerant example data from Mascot help, using one variable mod, results in 84 statically significant matches.

If the search is repeated with 7 variable mods, the individual matches have identical scores, but the significance threshold is higher and there are fewer matches overall.

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

What's worse, the search takes over 50 times as long!

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

Why is phosphorylation such a challenge?

Site heterogeneity

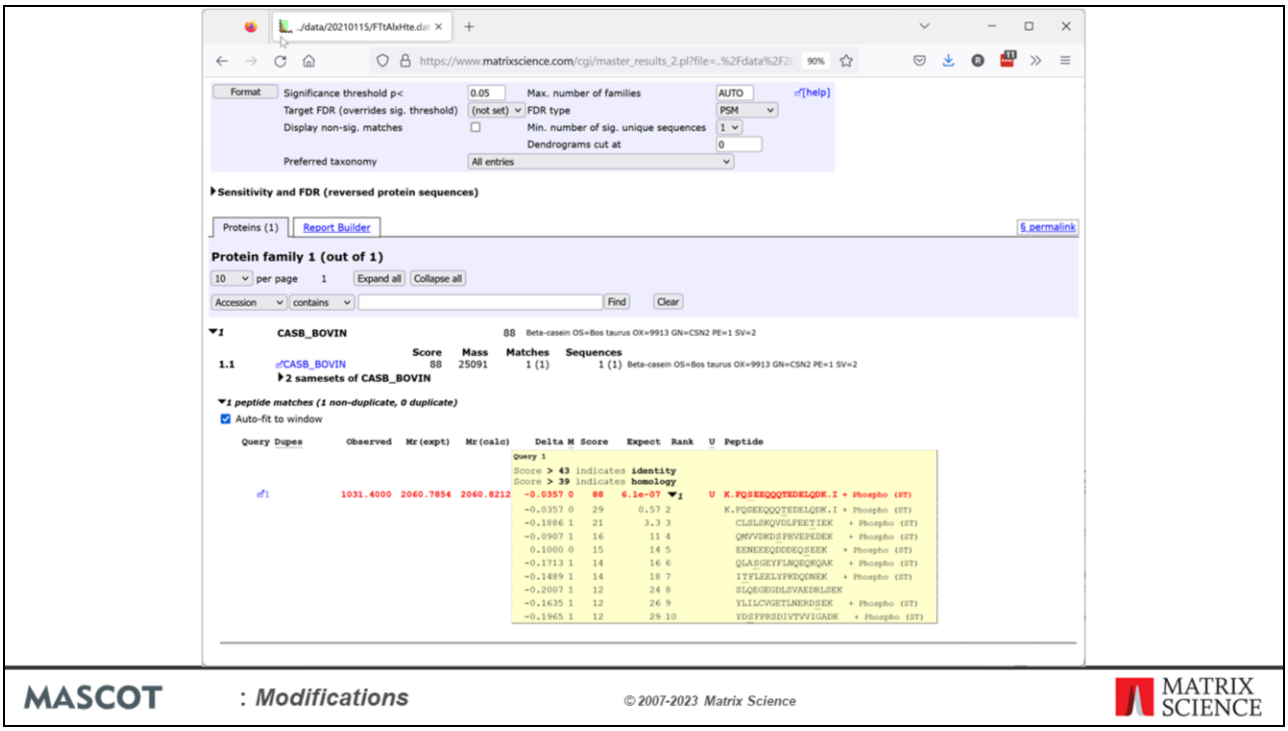
Poor ionisation efficiency

3 fragmentation channels

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

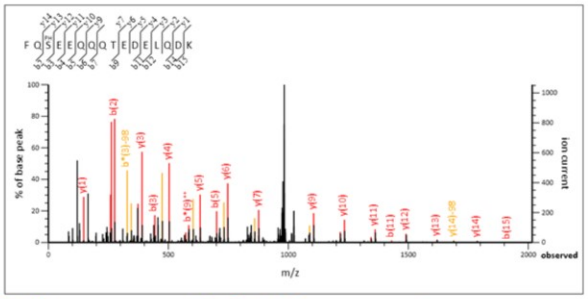
Can occur at STY - ~16% of residues.

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?



Let's 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.



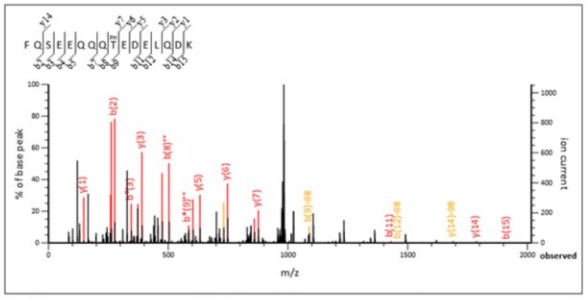
Label all possible matches ○ Label matches used for scoring ◉

Monoisotopic mass of neutral peptide Mr(calc): 2060.8212
Variable modifications:
83 : Phospho (ST), with neutral losses 0.0000(shown in table), 97.9769
Ions Score: 88 **Expect:** 1.7e-06
Matches : 39/262 fragment ions using 56 most intense peaks (help)

#	b	b^{++}	b*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	148.0757	74.5415					F							16
2	**276.1343**	138.5708	259.1077	130.0575			Q	1914.7601	957.8837	1897.7335	949.3704	1896.7495	948.8784	15
3	**443.1326**	222.0700	426.1061	213.5567	425.1221	213.0647	S	***1786.7015***	893.8544	1769.6749	885.3411	1768.6909	884.8491	14
4	**572.1752**	286.5912	555.1487	278.0780	554.1647	277.5860	E	***1619.7031***	810.3552	1602.6766	801.8419	1601.6926	801.3499	13
5	**701.2178**	351.1125	684.1913	342.5993	683.2072	342.1073	E	***1490.6605***	745.8339	1473.6340	737.3206	1472.6500	736.8286	12
6	829.2764	415.1418	812.2498	406.6286	811.2658	406.1366	Q	***1361.6179***	681.3126	1344.5914	672.7993	1343.6074	672.3073	11
7	957.3350	479.1711	940.3084	470.6578	939.3244	470.1658	Q	***1233.5594***	617.2833	1216.5328	608.7700	1215.5488	608.2780	10
8	1085.3935	543.2004	1068.3670	534.6871	1067.3830	534.1951	Q	***1105.5008***	553.2540	1088.4742	544.7408	1087.4902	544.2487	9
9	1186.4412	593.7243	1169.4147	585.2110	1168.4307	584.7190	T	977.4422	489.2247	960.4156	480.7115	959.4316	480.2195	8
10	1315.4838	658.2455	1298.4573	649.7323	1297.4733	649.2403	E	***876.3945***	438.7009	859.3680	430.1876	858.3840	429.6956	7
11	**1430.5108**	715.7590	1413.4842	707.2457	1412.5002	706.7537	D	***747.3519***	374.1796	730.3254	365.6663	729.3414	365.1743	6
12	1559.5534	780.2803	1542.5268	771.7670	1541.5428	771.2750	E	***632.3250***	316.6661	615.2984	308.1529	614.3144	307.6608	5
13	1672.6374	836.8223	1655.6109	828.3091	1654.6269	827.8171	L	***503.2824***	252.1448	486.2558	243.6316	485.2718	243.1396	4
14	**1800.6960**	900.8516	1783.6694	892.3384	1782.6854	891.8464	Q	***390.1983***	195.6028	373.1718	187.0895	372.1878	186.5975	3
15	**1915.7229**	958.3651	1898.6964	949.8518	1897.7124	949.3598	D	***262.1397***	131.5735	245.1132	123.0602	244.1292	122.5682	2
16							K	***147.1128***	74.0600	130.0863	65.5468			1

MASCOT : *Modifications* MATRIX SCIENCE

Beautiful spectrum; long run of y ions.



Label all possible matches ○ Label matches used for scoring ◉

Monoisotopic mass of neutral peptide Mr(calc): 2060.8212
Variable modifications:
T9 : Phospho (ST), with neutral losses 0.0000(shown in table), 97.9769
Ions Score: 29 Expect: 1.6
Matches : 27/262 fragment ions using 56 most intense peaks (help)

#	b	b^{++}	b^*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y^*	y^{*++}	y^0	y^{0++}	#
1	148.0757	74.5415					F							16
2	**276.1343**	138.5708	259.1077	130.0575			Q	1914.7601	957.8837	1897.7335	949.3704	1896.7495	948.8784	15
3	363.1663	182.0868	346.1397	173.5735	345.1557	173.0815	S	***1786.7015***	893.8544	1769.6749	885.3411	1768.6909	884.8491	14
4	492.2089	246.6081	475.1823	238.0948	474.1983	237.6028	E	1699.6694	850.3384	1682.6429	841.8251	1681.6589	841.3331	13
5	621.2515	311.1294	604.2249	302.6161	603.2409	302.1241	E	1570.6269	785.8171	1553.6003	777.3038	1552.6163	776.8118	12
6	749.3101	375.1587	732.2835	366.6454	731.2995	366.1534	Q	1441.5843	721.2958	1424.5577	712.7825	1423.5737	712.2905	11
7	877.3686	439.1880	860.3421	430.6747	859.3581	430.1827	Q	1313.5257	657.2665	1296.4991	648.7532	1295.5151	648.2612	10
8	1005.4272	503.2172	988.4007	494.7040	987.4167	494.2120	Q	1185.4671	593.2372	1168.4406	584.7239	1167.4565	584.2319	9
9	1186.4412	593.7243	1169.4147	585.2110	1168.4307	584.7190	T	1057.4085	529.2079	1040.3820	520.6946	1039.3980	520.2026	8
10	1315.4838	658.2455	1298.4573	649.7323	1297.4733	649.2403	E	***876.3945***	438.7009	859.3680	430.1876	858.3840	429.6956	7
11	**1430.5108**	715.7590	1413.4842	707.2457	1412.5002	706.7537	D	***747.3519***	374.1796	730.3254	365.6663	729.3414	365.1743	6
12	1559.5534	780.2803	1542.5268	771.7670	1541.5428	771.2750	E	***632.3250***	316.6661	615.2984	308.1529	614.3144	307.6608	5
13	1672.6374	836.8223	1655.6109	828.3091	1654.6269	827.8171	L	***503.2824***	252.1448	486.2558	243.6316	485.2718	243.1396	4
14	**1800.6960**	900.8516	1783.6694	892.3384	1782.6854	891.8464	Q	***390.1983***	195.6028	373.1718	187.0895	372.1878	186.5975	3
15	**1915.7229**	958.3651	1898.6964	949.8518	1897.7124	949.3598	D	***262.1397***	131.5735	245.1132	123.0602	244.1292	122.5682	2
16							K	***147.1128***	74.0600	130.0863	65.5468			1

MASCOT : *Modifications*

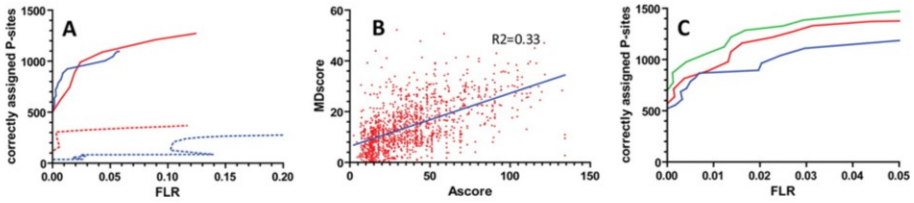
MATRIX SCIENCE

Move site to T9 and many matches would disappear.

Confident Phosphorylation Site Localization Using the Mascot Delta Score

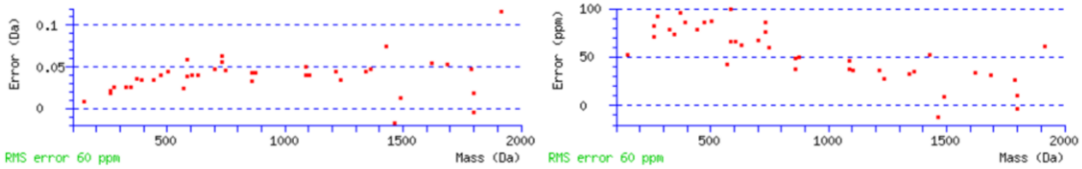
Mikhail M. Savitski‡, Simone Lemeer§, Markus Boesche‡, Manja Lang‡, Toby Mathieson‡, Marcus Bantscheff‡‖, and Bernhard Kuster§¶‖

Molecular & Cellular Proteomics 10.2 10.1074/mcp.M110.003830–1



MASCOT MATRIX SCIENCE

Mascot 2.4 and later report site localisation probabilities using the delta score method published in MCP by Bernhard 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.



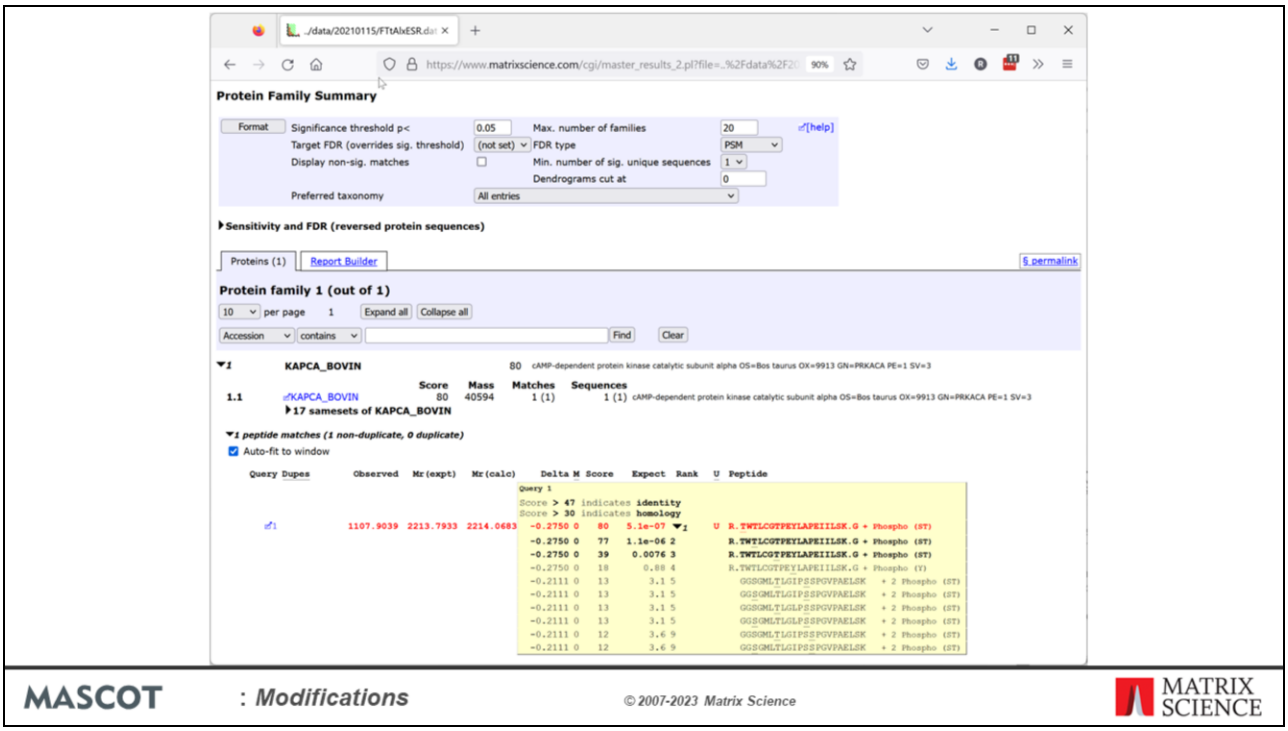
NCBI **BLAST** search of FQSEEQQQTEDELQDK
(Parameters: blastp, nr protein database, expect=20000, no filter, PAM30)
Other BLAST web gateways

All matches to this query

Score	Mr(calc)	Delta	Sequence	Site Analysis
88.3	2060.8212	-0.0357	FQSEEQQQTEDELQDK	Phospho S3 100.00%
28.6	2060.8212	-0.0357	FQSEEQQQTEDELQDK	Phospho T9 0.00%
21.0	2060.9741	-0.1886	CLSLSKQVDLFEETIEK	
15.9	2060.8762	-0.0907	QMVVDKDSPHVEPEDEK	
14.5	2060.6855	0.1000	EENEEEQDDDEQSEEK	
14.1	2060.9568	-0.1713	QLASGEYFLNQEQKQAK	
13.6	2060.9343	-0.1489	ITFLEELYPKDQDNEK	
12.4	2060.9862	-0.2007	SLQEGEGDLSVAEDRLSEK	
11.9	2060.9489	-0.1635	YLILCVGETLNERDSEK	
11.6	2060.9820	-0.1965	YDSFPRSDIVTVVIGADK	

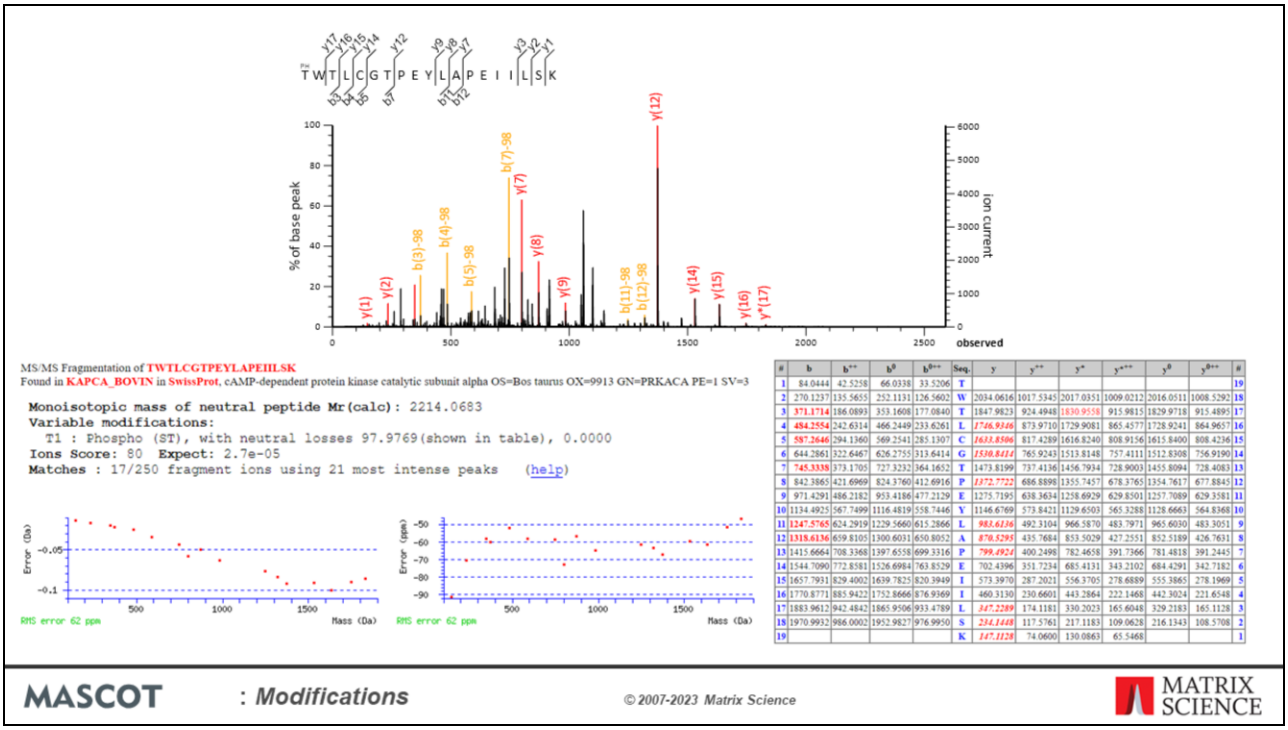
MASCOT : *Modifications* MATRIX SCIENCE

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



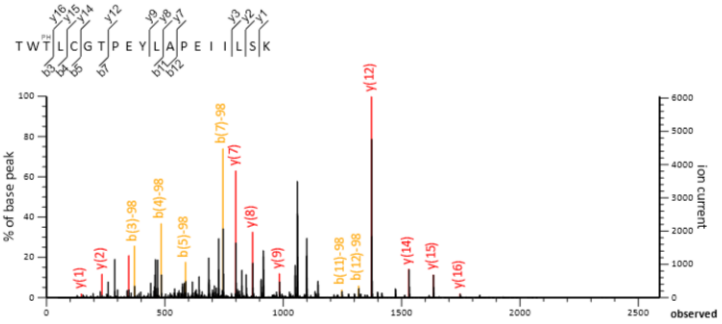
However, casein peptides are unusually easy to analyse. Here is a more typical example of what you can expect to find - a strong match to a phosphopeptide from a protein kinase.

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



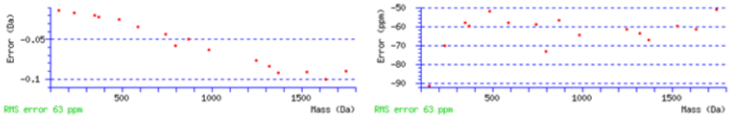
#	b	b++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	84.0444	42.5258	66.0338	33.5206	T							19
2	270.1237	135.5655	252.1131	126.5602	W	2034.0616	1017.5345	2017.0351	1009.0212	2016.0511	1008.5292	18
3	371.1714	186.0893	353.1608	177.0840	T	1847.9823	924.4948	1830.9558	915.9815	1829.9718	915.4895	17
4	484.2554	242.6314	466.2449	233.6261	L	*1746.9346*	873.9710	1729.9081	865.4577	1728.9241	864.9657	16
5	587.2646	294.1360	569.2541	285.1307	C	*1633.8506*	817.4289	1616.8240	808.9156	1615.8400	808.4236	15
6	644.2861	322.6467	626.2755	313.6414	G	*1530.8414*	765.9243	1513.8148	757.4111	1512.8308	756.9190	14
7	745.3338	373.1705	727.3232	364.1652	T	1473.8199	737.4136	1456.7934	728.9003	1455.8094	728.4083	13
8	842.3865	421.6969	824.3760	412.6916	P	*1372.7722*	686.8898	1355.7457	678.3765	1354.7617	677.8845	12
9	971.4291	486.2182	953.4186	477.2129	E	1275.7195	638.3634	1258.6929	629.8501	1257.7089	629.3581	11
10	1134.4925	567.7499	1116.4819	558.7446	Y	1146.6769	573.8421	1129.6503	565.3288	1128.6663	564.8368	10
11	1247.5765	624.2919	1229.5660	615.2866	L	*983.6136*	492.3104	966.5870	483.7971	965.6030	483.3051	9
12	1318.6136	659.8105	1300.6031	650.8052	A	*870.5295*	435.7684	853.5029	427.2551	852.5189	426.7631	8
13	1415.6664	708.3368	1397.6558	699.3316	P	*799.4924*	400.2498	782.4658	391.7366	781.4818	391.2445	7
14	1544.7090	772.8581	1526.6984	763.8529	E	702.4396	351.7234	685.4131	343.2102	684.4291	342.7182	6
15	1657.7931	829.4002	1639.7825	820.3949	I	573.3970	287.2021	556.3705	278.6889	555.3865	278.1969	5
16	1770.8771	885.9422	1752.8666	876.9369	I	460.3130	230.6601	443.2864	222.1468	442.3024	221.6548	4
17	1883.9612	942.4842	1865.9506	933.4789	L	*347.2289*	174.1181	330.2023	165.6048	329.2183	165.1128	3
18	1970.9932	986.0002	1952.9827	976.9950	S	*234.1448*	117.5761	217.1183	109.0628	216.1343	108.5708	2
19					K	*147.1128*	74.0600	130.0863	65.5468			1

We can see why there is little difference in score between placing the phosphate on T1 or T3.



MS/MS Fragmentation of **TWTLCGTPEYLAPEIILSK**
Found in **KAPCA_BOVIN** in **SwissProt**, cAMP-dependent protein kinase catalytic subunit alpha OS=Bos taurus OX=9913 GN=PRKACA PE=1 SV=3

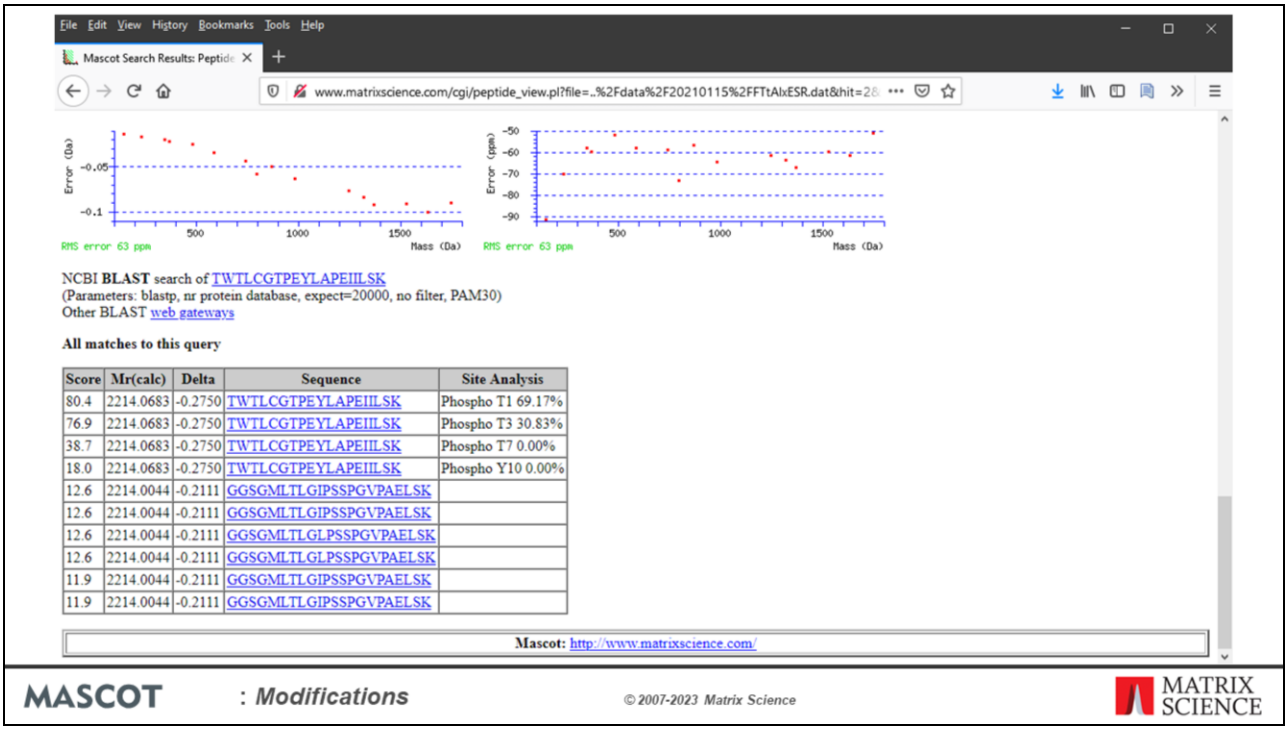
Monoisotopic mass of neutral peptide Mr(calc): 2214.0683
Variable modifications:
T3 : Phospho (ST), with neutral losses 97.9769(shown in table), 0.0000
Ions Score: 77 **Expect:** 6e-05
Matches : 16/254 fragment ions using 21 most intense peaks (help)



#	b	b**	b0	b0**	Seq.	y	y**	y*	y***	y0	y0**	#
1	102.0550	51.5311	84.0444	42.5258	T							19
2	288.1343	144.5708	270.1237	135.5655	W	2016.0511	1008.5292	1999.0245	1000.0159	1998.0405	999.5239	18
3	371.1714	186.0893	353.1608	177.0840	T	1829.9718	915.4895	1812.9452	906.9762	1811.9612	906.4842	17
4	484.2554	242.6314	466.2449	233.6261	L	*1746.9346*	873.9710	1729.9081	865.4577	1728.9241	864.9657	16
5	587.2646	294.1360	569.2541	285.1307	C	*1633.8506*	817.4289	1616.8240	808.9156	1615.8400	808.4236	15
6	644.2861	322.6467	626.2755	313.6414	G	*1530.8414*	765.9243	1513.8148	757.4111	1512.8308	756.9190	14
7	745.3338	373.1705	727.3232	364.1652	T	1473.8199	737.4136	1456.7934	728.9003	1455.8094	728.4083	13
8	842.3865	421.6969	824.3760	412.6916	P	*1372.7722*	686.8898	1355.7457	678.3765	1354.7617	677.8845	12
9	971.4291	486.2182	953.4186	477.2129	E	1275.7195	638.3634	1258.6929	629.8501	1257.7089	629.3581	11
10	1134.4925	567.7499	1116.4819	558.7446	Y	1146.6769	573.8421	1129.6503	565.3288	1128.6663	564.8368	10
11	1247.5765	624.2919	1229.5660	615.2866	L	*983.6136*	492.3104	966.5870	483.7971	965.6030	483.3051	9
12	1318.6136	659.8105	1300.6031	650.8052	A	*870.5295*	435.7684	853.5029	427.2551	852.5189	426.7631	8
13	1415.6664	708.3368	1397.6558	699.3316	P	*799.4924*	400.2498	782.4658	391.7366	781.4818	391.2445	7
14	1544.7090	772.8581	1526.6984	763.8529	E	702.4396	351.7234	685.4131	343.2102	684.4291	342.7182	6
15	1657.7931	829.4002	1639.7825	820.3949	I	573.3970	287.2021	556.3705	278.6889	555.3865	278.1969	5
16	1770.8771	885.9422	1752.8666	876.9369	I	460.3130	230.6601	443.2864	222.1468	442.3024	221.6548	4
17	1883.9612	942.4842	1865.9506	933.4789	L	*347.2289*	174.1181	330.2023	165.6048	329.2183	165.1128	3
18	1970.9932	986.0002	1952.9827	976.9950	S	*234.1448*	117.5761	217.1183	109.0628	216.1343	108.5708	2
19					K	*147.1128*	74.0600	130.0863	65.5468			1

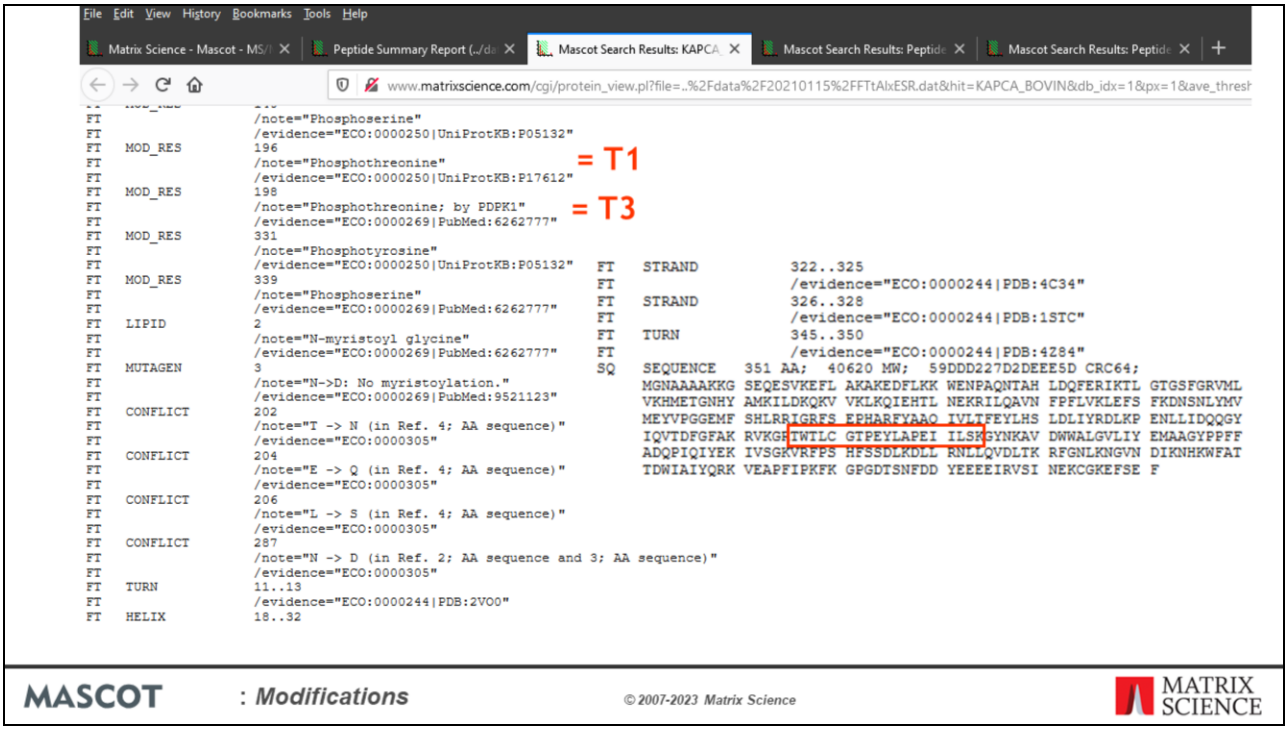
MASCOT : *Modifications* MATRIX SCIENCE

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.



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.

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 a preferred site by inspection
- Often, you just can't differentiate between closely spaced sites, even with great data.

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.

Error Tolerant Search

First pass: standard search of specified database(s)

Second pass: database entries that contain one or more significant peptide matches are selected and searched with

- Relaxed enzyme specificity
- Comprehensive list of chemical and post-translational modifications
- Single residue substitutions or single nucleotide substitutions, insertions and deletions.

Reference

➢Creasy, D. M. and Cottrell, J. S., Error tolerant searching of uninterpreted tandem mass spectrometry data, Proteomics 2 1426-1434 (2002)

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. 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.

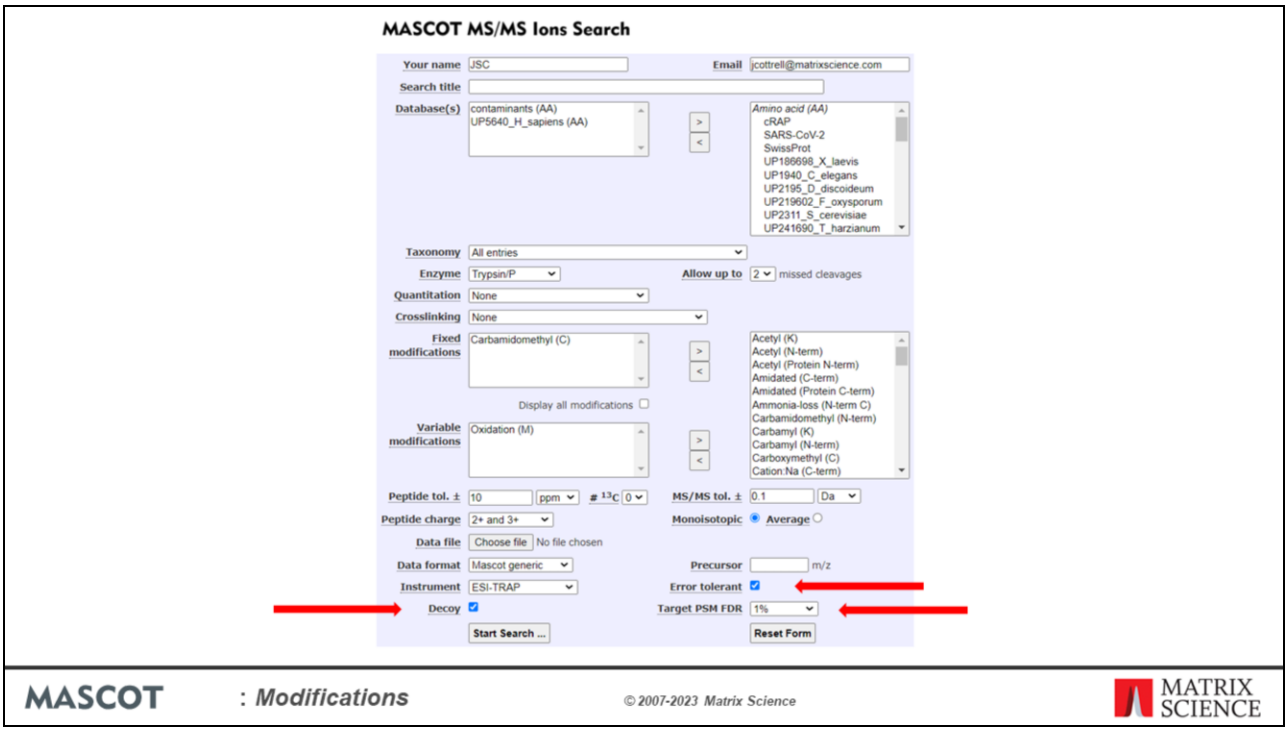
For a protein database, we can't look for the consequences of inserted or deleted bases, because these give rise to frame shifts, and the entire sequence changes from that point on.

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 quantitation
4. Search cannot include error tolerant sequence tag

There are some constraints on the standard, first pass search.

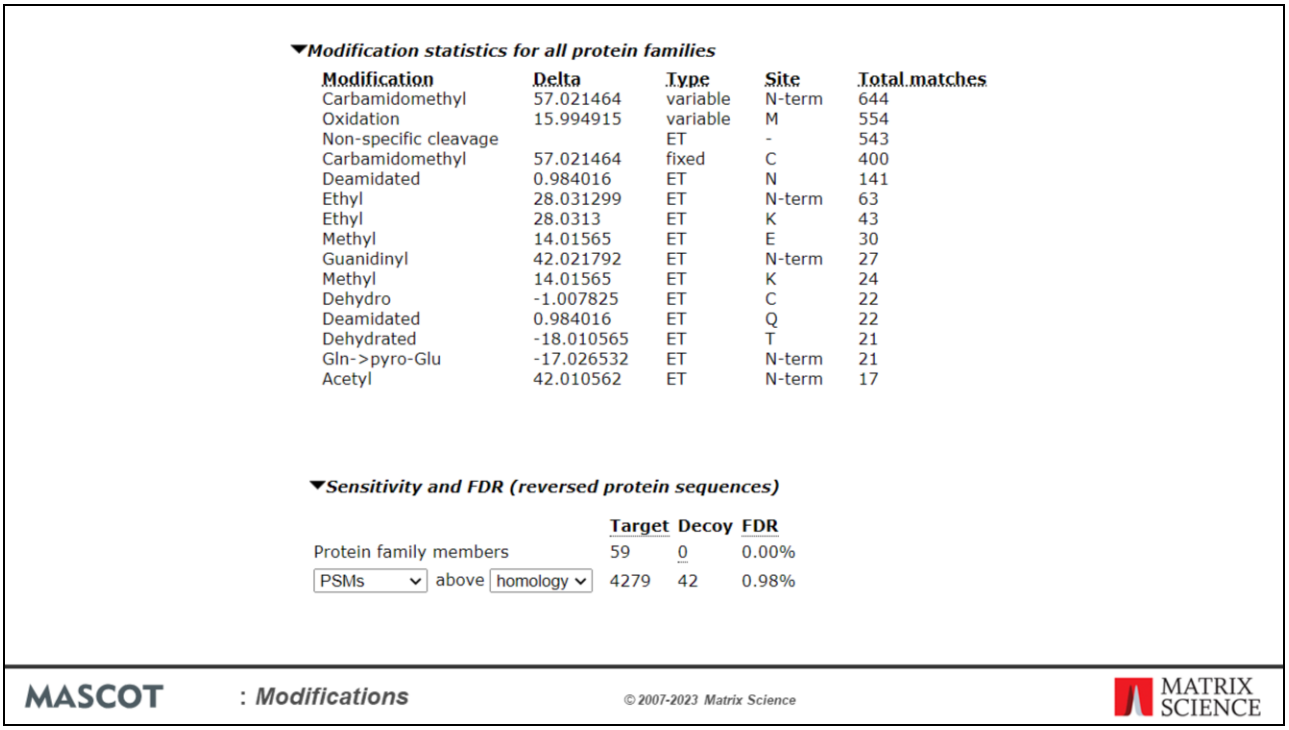


In Mascot Server 2.8 and later, we use target-decoy searching to assign significance to all matches, including those found in the second pass search.

As in previous releases, all you need to do to perform an error tolerant search is to check a box on the search form.

You can (and should) also check the box to use target-decoy. Without a decoy, expect values are derived from counting trials – that is, the number of candidate peptides that have been tested. This estimate is not always accurate; particularly when there is something wrong with the choice of database or search parameters, making a large fraction of potential matches unavailable. Ticking the checkbox to search a decoy database gives a solid, empirical basis for the statistics.

There is also a control to specify the required false discovery rate. The reason we ask for is up front is that the FDR determines the set of proteins selected for the second pass search. For example, the first pass search might identify significant peptide matches to 500 proteins at an FDR of 5%, and these are sent through to the second pass. If the FDR was reduced to 1%, the number of proteins selected for the second pass might drop to 400. Although the FDR can be tweaked at the report stage, this will not give perfectly identical results to setting the required FDR in the search form.



▼*Modification statistics for all protein families*

Modification	Delta	Type	Site	Total matches
Carbamidomethyl	57.021464	variable	N-term	644
Oxidation	15.994915	variable	M	554
Non-specific cleavage		ET	-	543
Carbamidomethyl	57.021464	fixed	C	400
Deamidated	0.984016	ET	N	141
Ethyl	28.031299	ET	N-term	63
Ethyl	28.0313	ET	K	43
Methyl	14.01565	ET	E	30
Guanidinyl	42.021792	ET	N-term	27
Methyl	14.01565	ET	K	24
Dehydro	-1.007825	ET	C	22
Deamidated	0.984016	ET	Q	22
Dehydrated	-18.010565	ET	T	21
Gln->pyro-Glu	-17.026532	ET	N-term	21
Acetyl	42.010562	ET	N-term	17

▼*Sensitivity and FDR (reversed protein sequences)*

	Target	Decoy	FDR
Protein family members	59	0	0.00%
PSMs above homology	4279	42	0.98%

MASCOT : *Modifications* *© 2007-2023 Matrix Science* MATRIX SCIENCE

When the results come back, you have a single report that combines the results from both passes.

At the top of the protein family report there is an expandable section that summarises the modifications. You can see counts of the number of queries matching fixed and variable modifications used in the first pass of the search and then for the results of the Error Tolerant search.

The sensitivity and FDR section can also be expended to see the number of decoy matches.

Error Tolerant Search: Target-Decoy

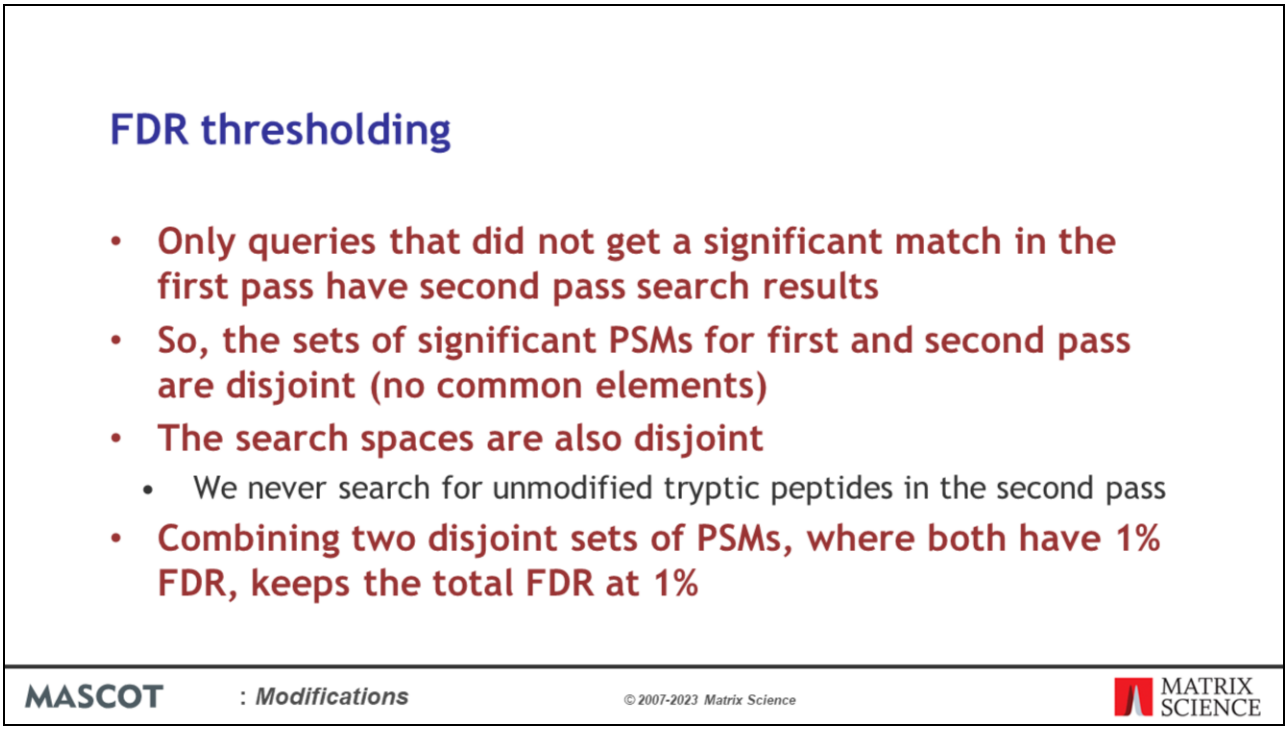
The target and decoy proteins are treated as pairs

- Target and decoy databases are of identical size
- All significant peptide matches (PSMs) from the first pass are represented

Blindly discard second pass results for queries that get a significant match in the first pass search

The way it works is that target and decoy proteins are treated as pairs. After the first pass search, when proteins are selected, each significant match, whether target or decoy, causes the relevant pair of target and decoy proteins to be selected for the second pass. This means that the target and decoy databases in the second pass search are of identical size and contain all significant peptide matches (PSMs) from the first pass.

If a query gets a significant match in the first pass search, this is what we report, and we blindly discard the second pass results for this query. Sometimes, this means a stronger match is missed, but to do otherwise would be statistically dishonest. For example, if the significance threshold for a particular query in the first pass search corresponds to a score of 40, and we get a match with a score of 52, this is what we report, even if the second pass search might give us an even better match. This is not ideal, but the alternative is to burden all matches with statistics based on both passes. To illustrate why this is a problem, imagine we were to look at the second pass results and find nothing better. Now, we have a larger search space and the score threshold has increased to 55, so we have to discard our first pass match with a score of 52 because it is no longer significant.

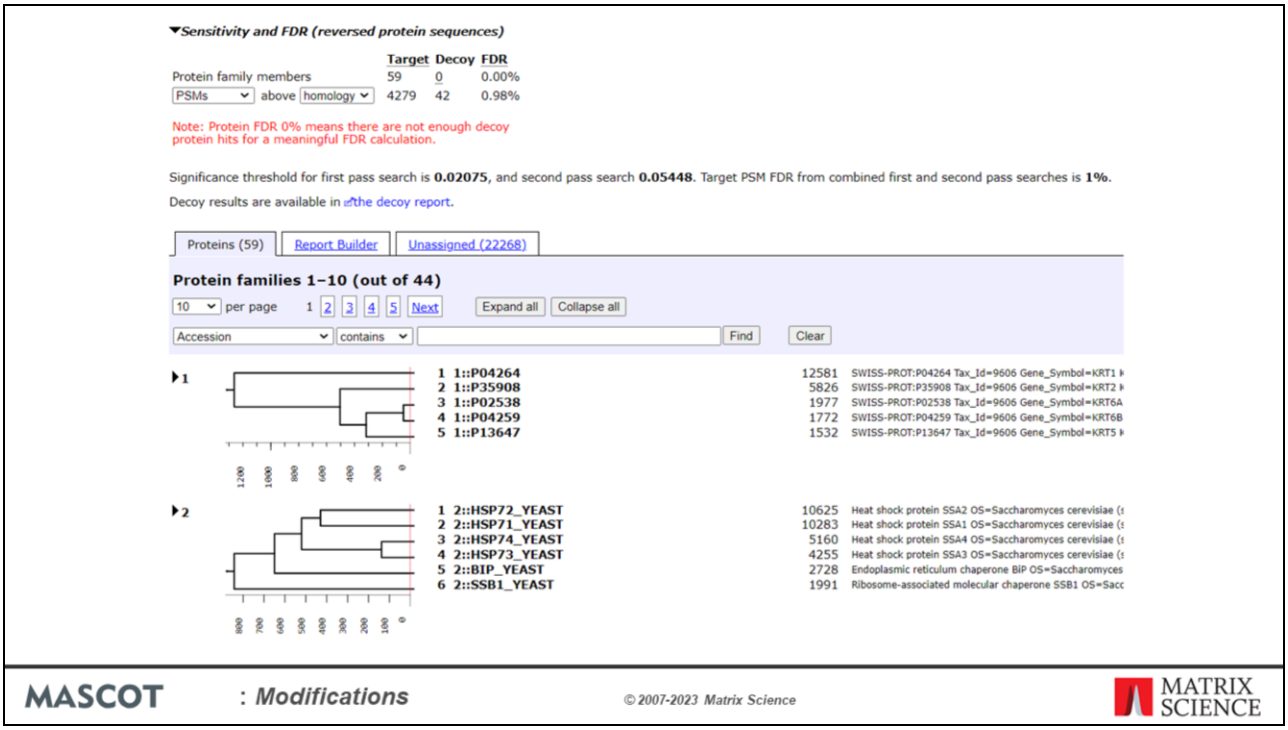


In earlier versions of Mascot, an error tolerant search could not be combined with the target-decoy option, and expect values based on counting trials were only reported for first pass matches.

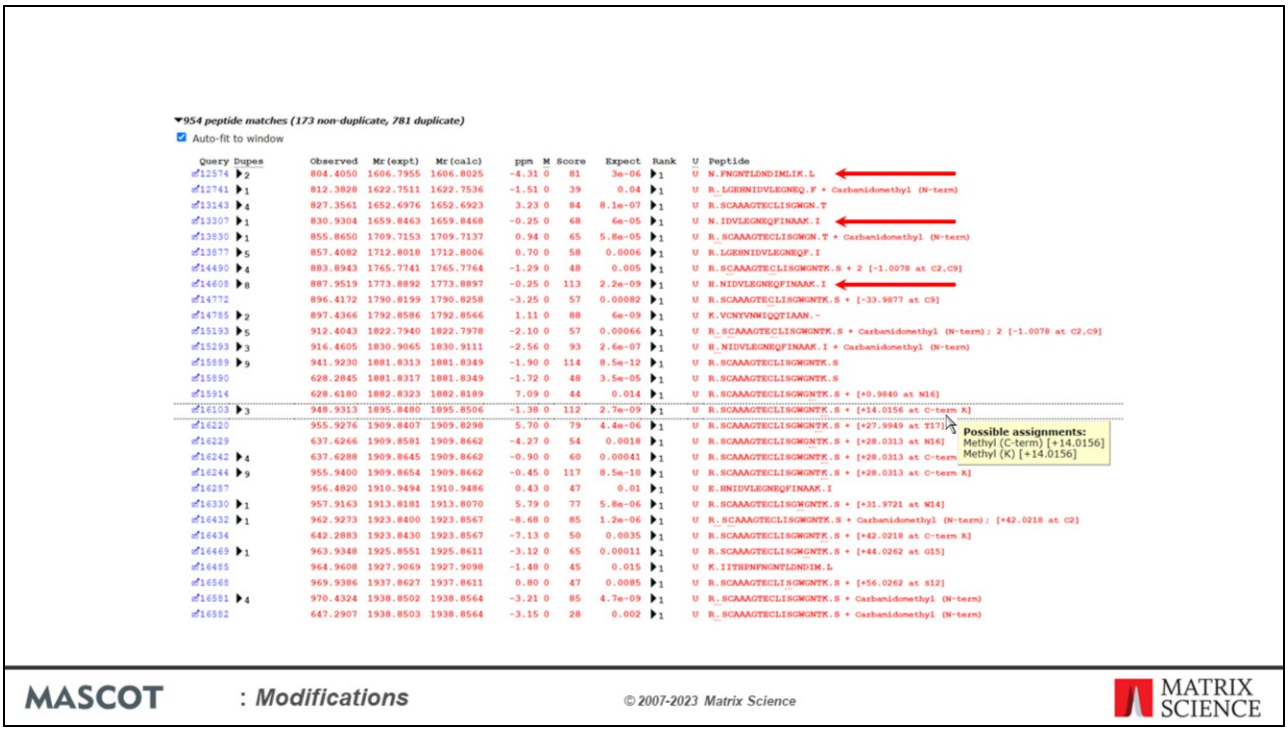
However, it is possible to combine the two passes and calculate the FDR using the following logic. Only queries that did not get a significant match in the first pass have second pass search results, so the sets of significant PSMs for first and second pass are disjoint. A disjoint set is a pair of sets which does not have any common element.

The first pass and second pass search spaces are also disjoint as we never search for unmodified tryptic peptides in the second pass.

This means there is no double counting. Combining two disjoint sets of PSMs, where both have 1% FDR, keeps the total FDR at 1%.



The required FDR is applied independently to the results from the first and second pass searches.



And the expect values are reported for both first and second pass matches.

The additional matches, found in the error tolerant search, are the ones with a mass difference in square brackets or that are a non-specific cleavage. One of these, query 12574, is a simple, non-specific peptide with a very good score. There's another example for queries 13307 and 14608. 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.
- The first pass match must be insignificant in its smaller search space while the second pass match must be a significant match in its larger search space.

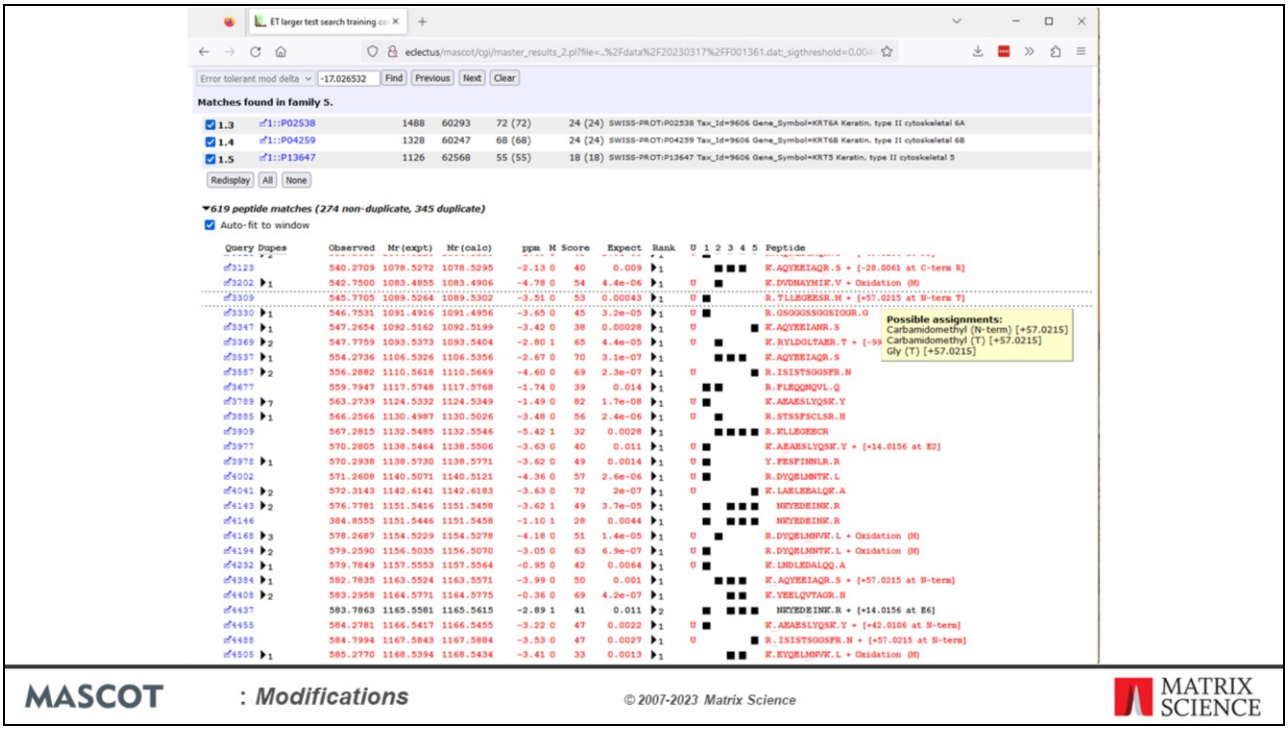
The matches from an error tolerant search are aggressively filtered to remove junk matches.

A peptide match is only allowed one error tolerant change, whether becoming semi-specific or getting one unsuspected modification.

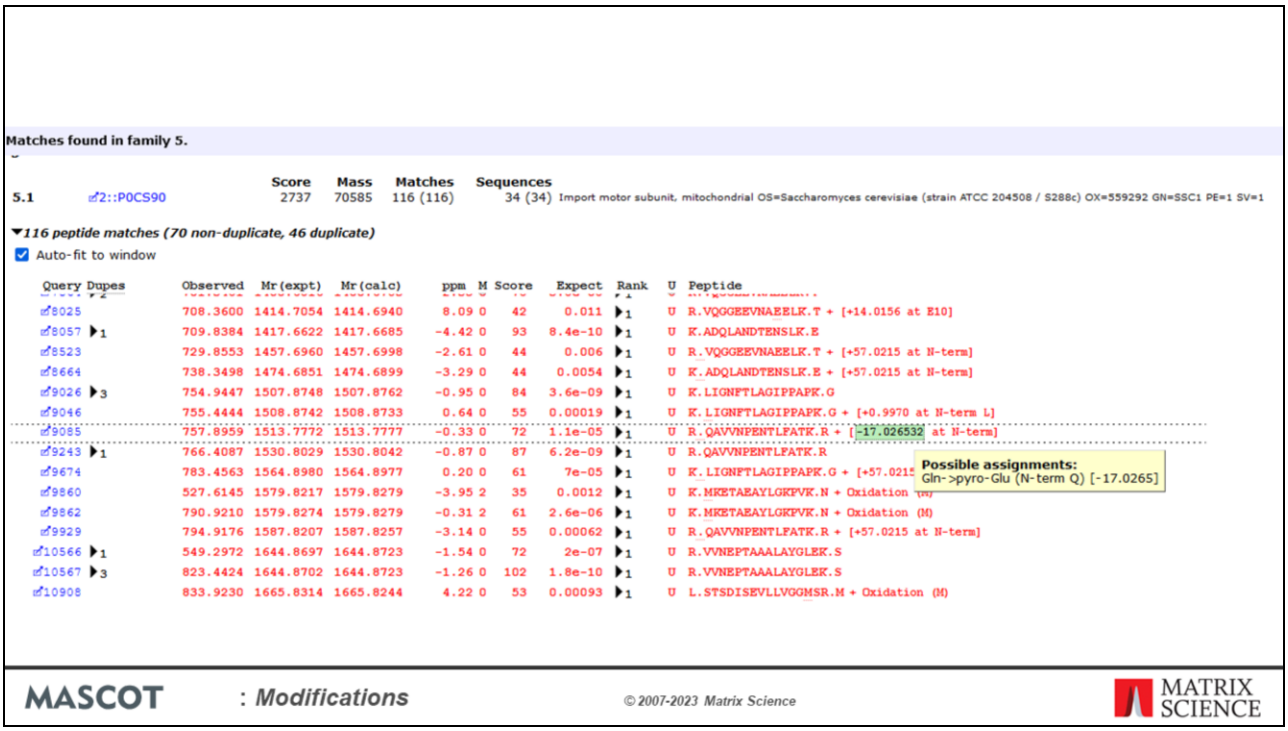
Modifications are rejected based on precursor and mass tolerance to eliminate meaningless possibilities.

And, because the second pass search space is much larger, the second pass match must have a very good score to be statistically significant, which throws away a lot of 'junk' matches.

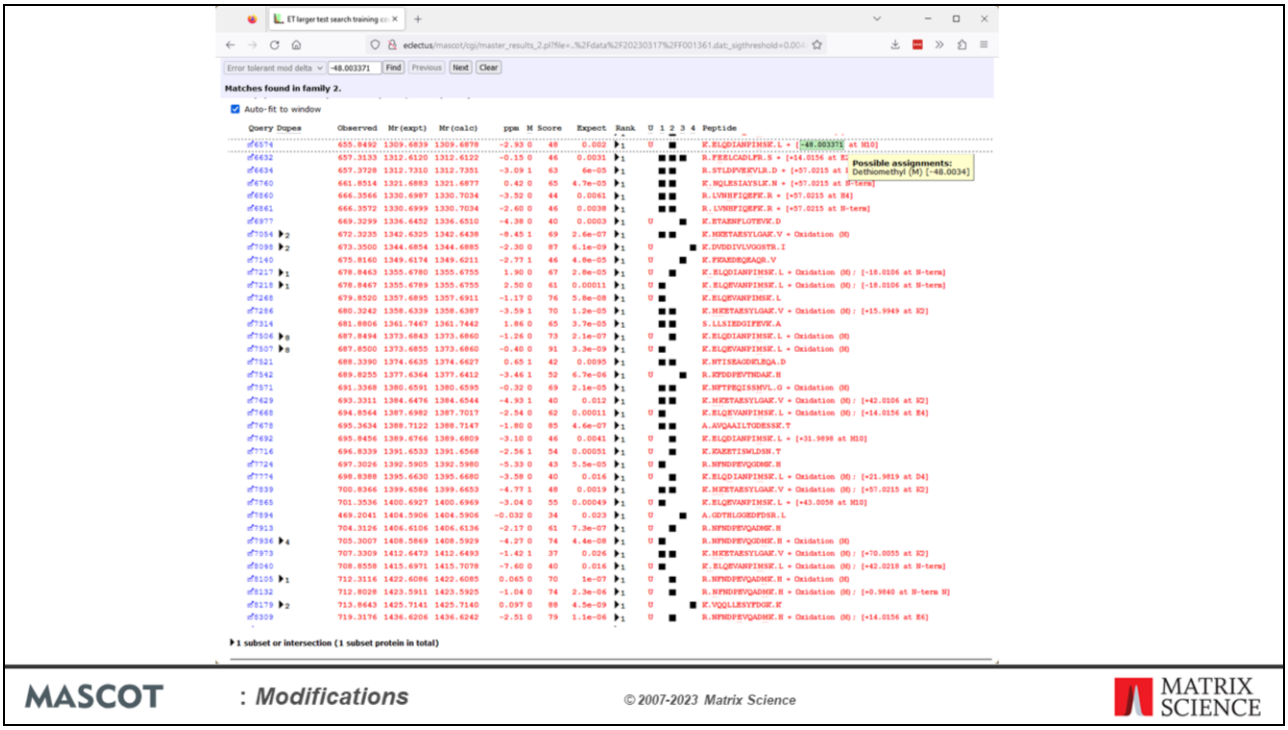
Usually, the search space for the second pass search is larger than for the first pass. This means that the significance threshold is more stringent for second pass matches. Here, for example, query 16581 gets a score of 85 in the first pass search which corresponds to an expect value of 4.7E-9. Query 16432 gets the same score in the second pass search, but the expect value is 1.2E-6, worse by a factor of 250.



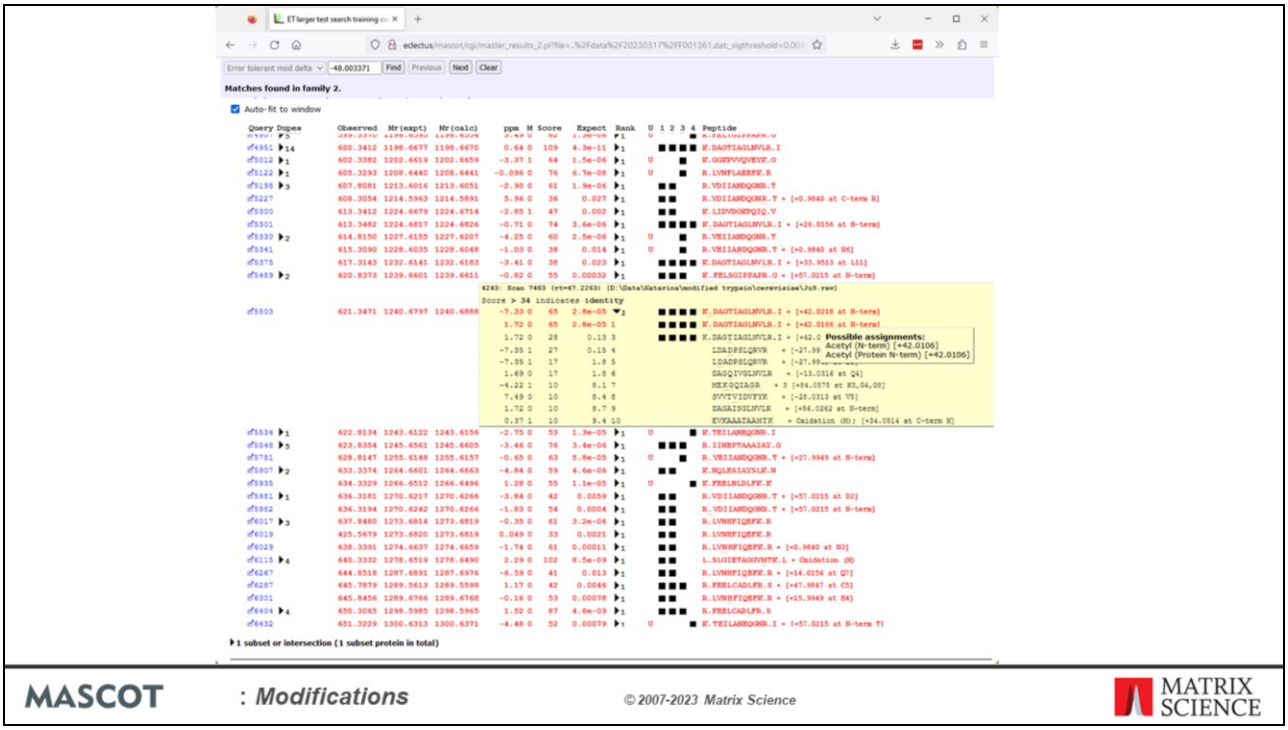
Take a look at the match to query 3309, which has three possible assignments. Since this sample was alkylated with iodoacetamide, we would choose carbamidomethylation as the more likely suspect. The assignment to carbamidomethylation at the N-terminal is also very believable, because this is a known artefact of over-alkylation. The same modification can be seen in this screen shot for three other queries.



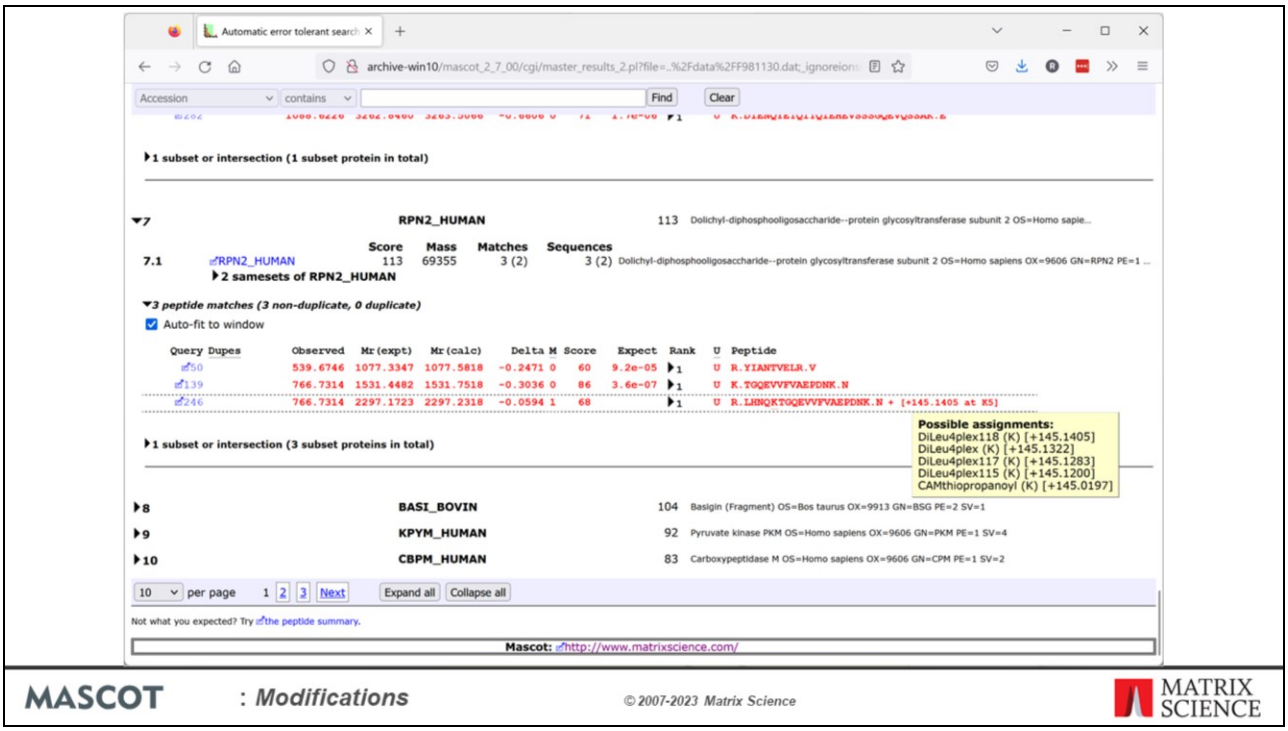
Another easily believable assignment is pyro-Glu for the match to query 11185.



Query 6574 is an interesting case. There are multiple other matches to the same peptide, oxidised at M10, oxidised at M10 with pyro-Glu at the N-term, dioxidation at M10 and others. The minus 48 modification occurs when the methionine is oxidised and then loses the side chain as methanesulfenic acid.



You should also look at the other matches to the same query when trying to decide whether to accept a match or not. For this query the top three matches are essentially to the same sequence. The error tolerant match is to a peptide that has undergone a guanidinylation or acetylation at the N-terminal or acetylation of the threonine. Although the guanidinylation PTM is know to occur the most likely interpretation for this match is Protein N-terminal acetylation. The precursor measurement error slightly favors acetylation, 1.72ppm, vs guanidinylation at -7.33ppm.



I do not have a good solution for this match. The modifications are either from DiLeu labelled quantitation experiments or a side product in crosslinking while the sample will have been exposed to neither of these two scenarios. Most likely the +145 at the C terminal is a combination of modifications in the C-terminal region but I don't know exactly what.

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.

Error Tolerant Search: Tips

In most cases, best to exclude

- Isotopic labels
- Modifications larger than 1 kDa

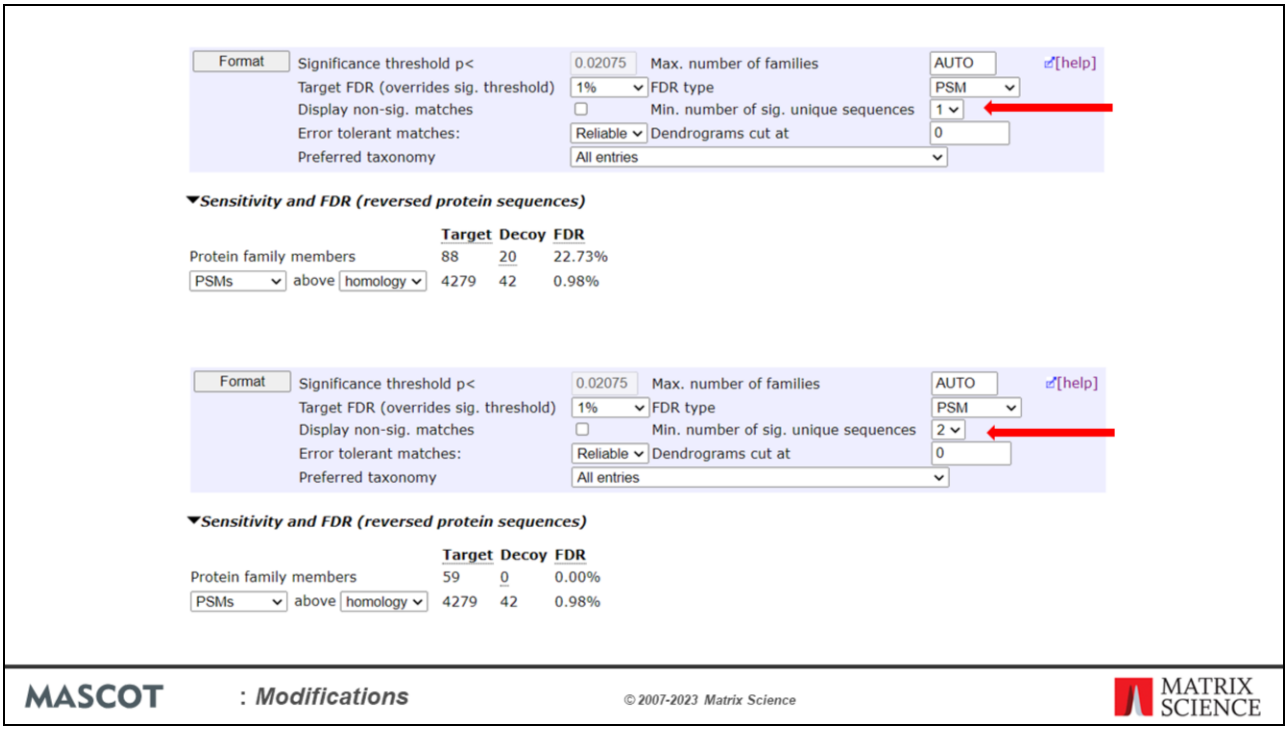
Very abundant modifications should be specified as variable

- Otherwise, miss doubly modified peptides

Adjust *Min. number of sig. unique sequences* for good protein FDR

To summarise, it's best to exclude isotopic labels and very large modifications from error tolerant searching. You should specify very abundant modifications as variable modifications, so you don't miss doubly modified peptides.

Even when the FDR for PSMs is well controlled, the FDR for proteins will often be high for an error tolerant search because only a few entries are searched in the second pass.



In this example search, at 1% FDR for PSMs, the protein FDR is 23%, which sounds awful. This is simply because the 42 significant decoy matches are scattered randomly across 20 decoy proteins. If we increase the ‘Min. number of sig. unique sequences’ from 1 to 2 and choose ‘Format’, we eliminate one hit wonders, and the protein FDR drops to a more satisfactory 0%.

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.

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