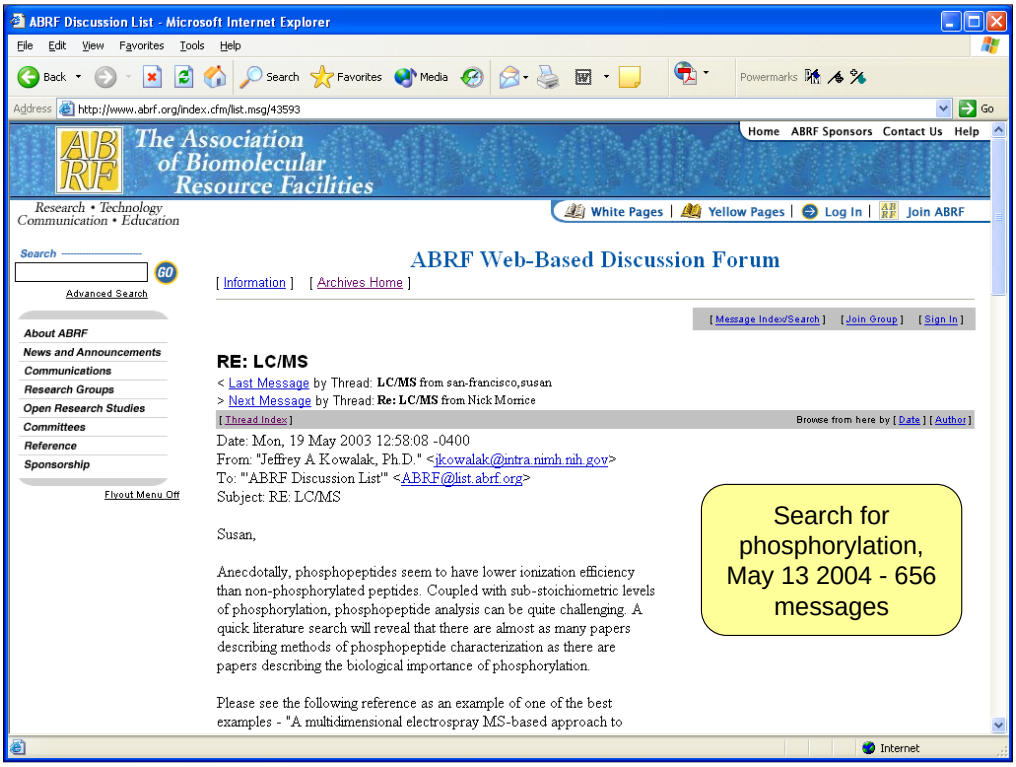
Phosphorylation Site Analysis

ASMS 2004

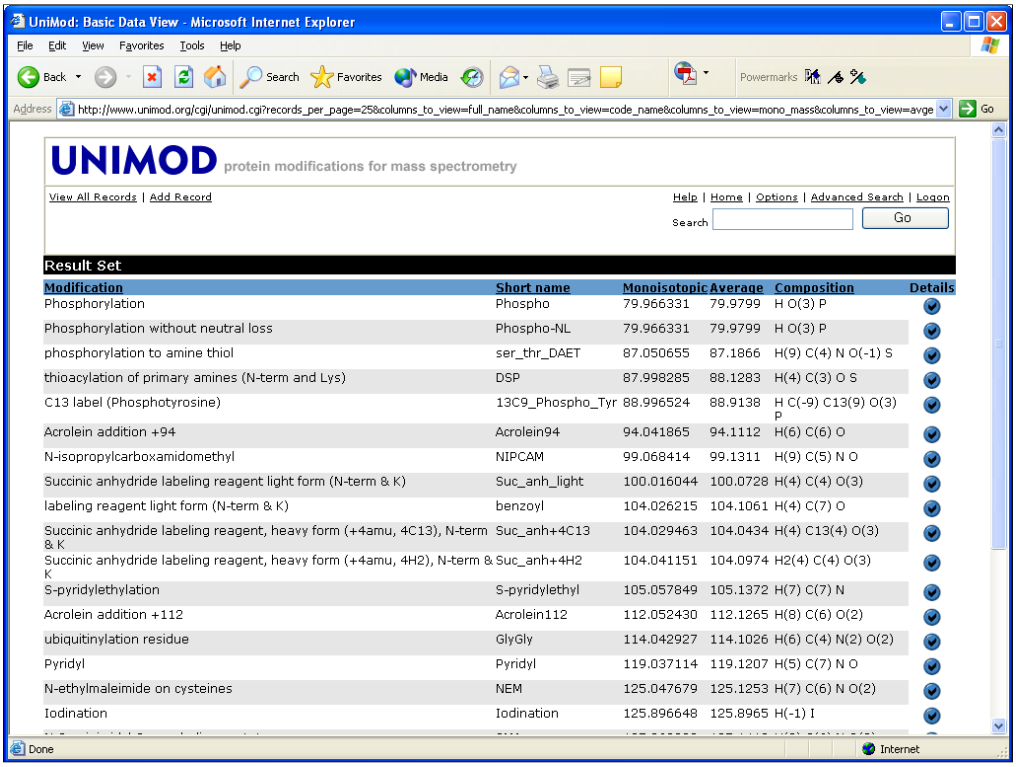
MATRIX SCIENCE

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.

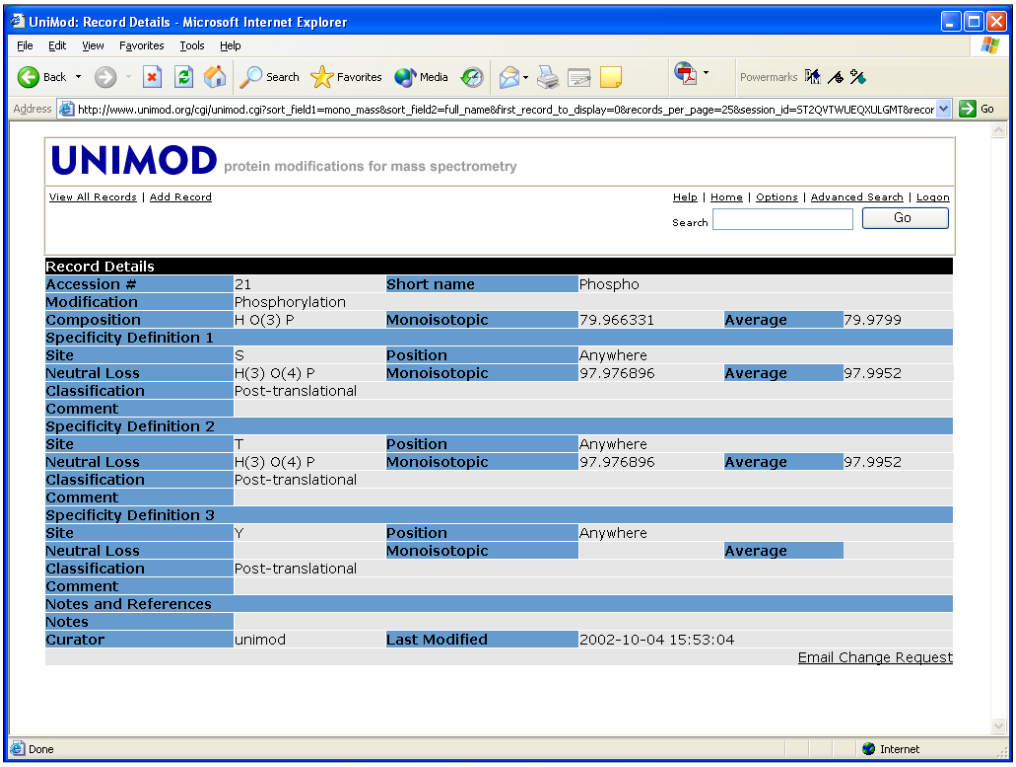


One of the best sources of information on phosphorylation is the ABRF email discussion group. When I checked a few days ago, there were 656 messages on this topic



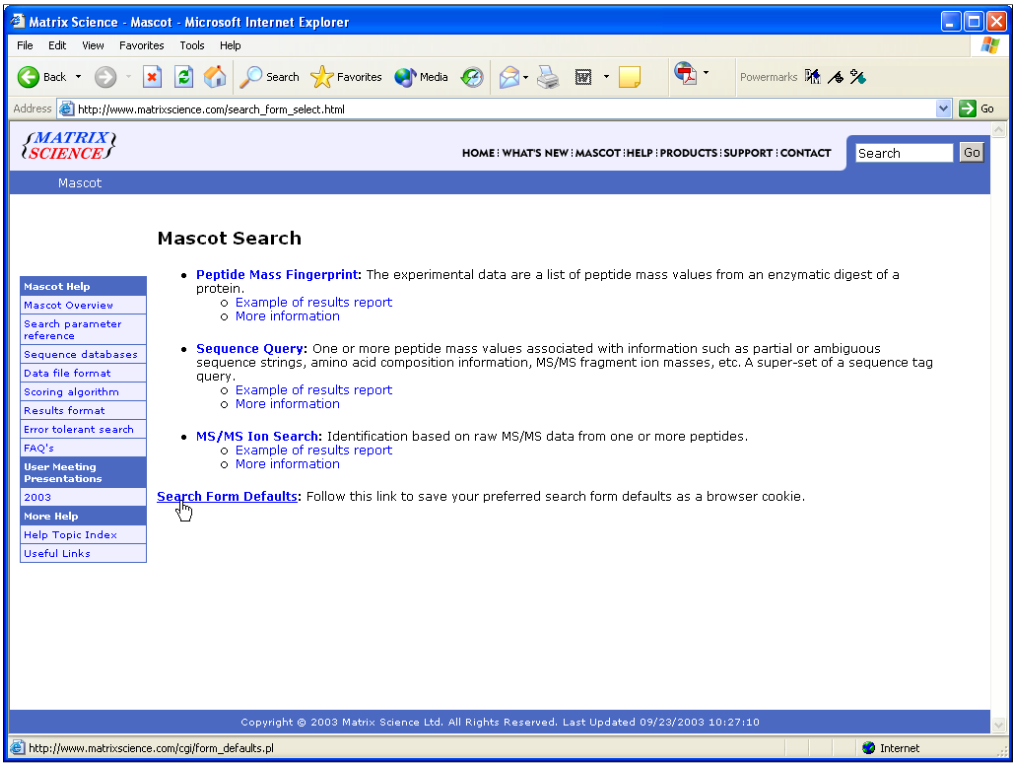
How is phosphorylation handled in Mascot?

Whether you use the public web site or whether you have Mascot in-house, we recommend using the Unimod web site to browse and update the list of modifications

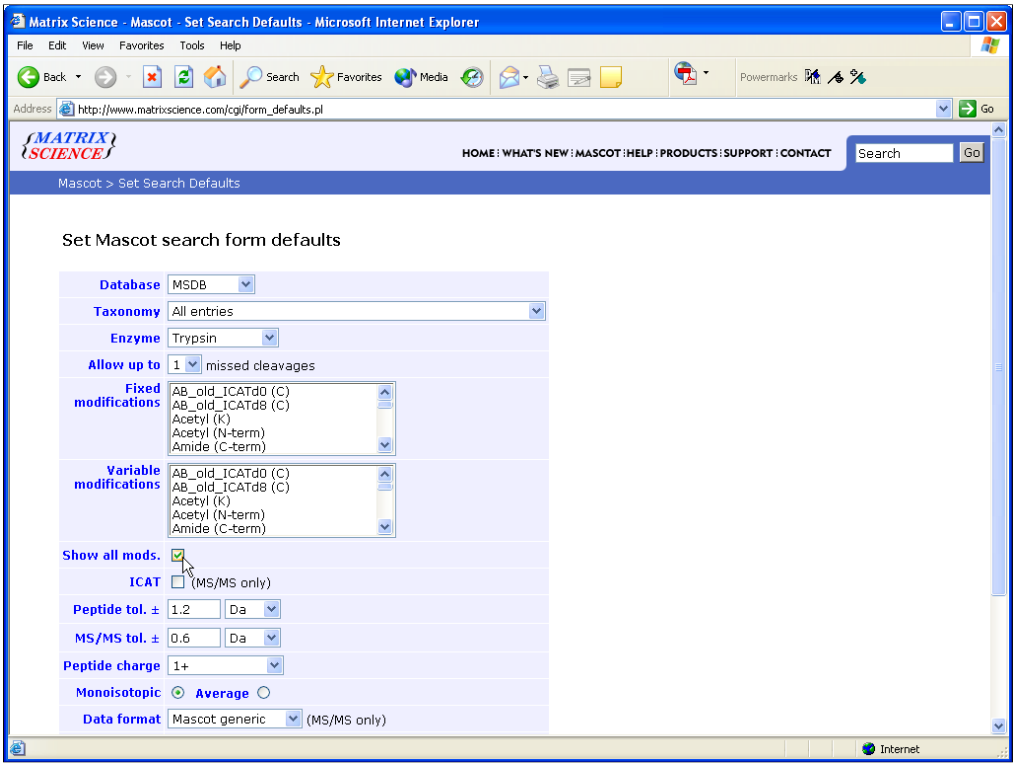


The standard modifications, that appear in the short, default list, assume that serine and threonine predominantly lose phosphate as a neutral loss of 98 Da, while tyrosine fragment ions stay intact

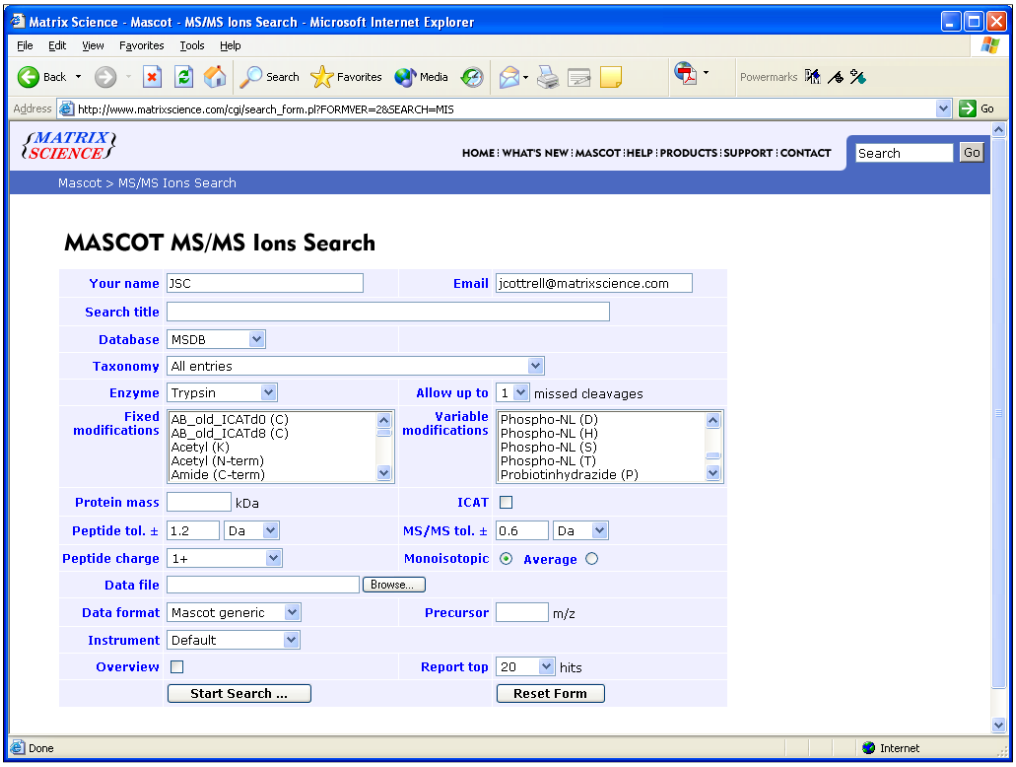
However, several other possibilities are also defined. For example, serine and threonine without any neutral loss. You don't see these listed on the short, default list of mods



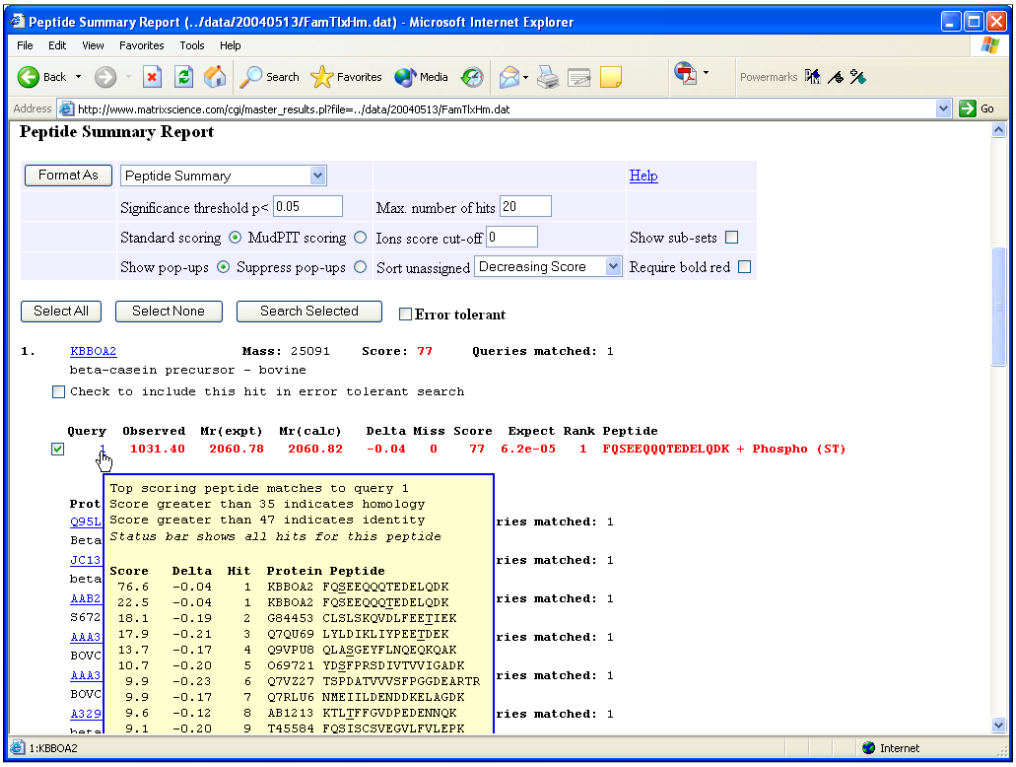
If you want to see the complete list of mods, you just need to choose this option in the search form defaults. Follow the link at the bottom of the search form selection page



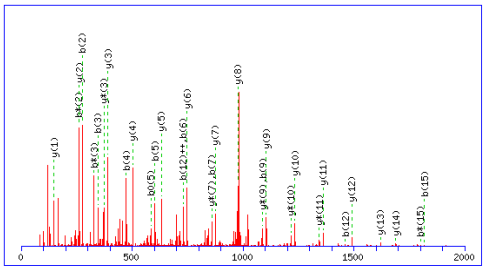
And check the box for Show all mods



Then, you'll see all of the mods listed in the search form.



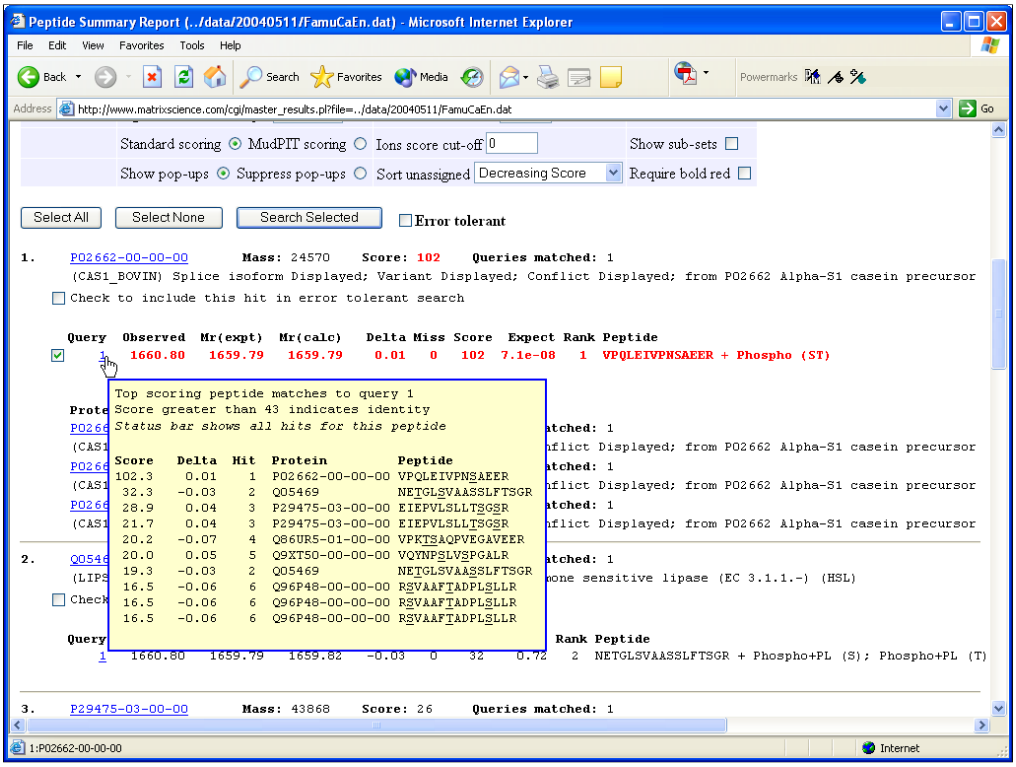
One of the most common phosphopeptides comes from the milk protein, beta casein. There are two potential phosphorylation sites, S and T, but only one is modified. Because the two sites are widely separated, there is no ambiguity, even if the spectrum is not the greatest.



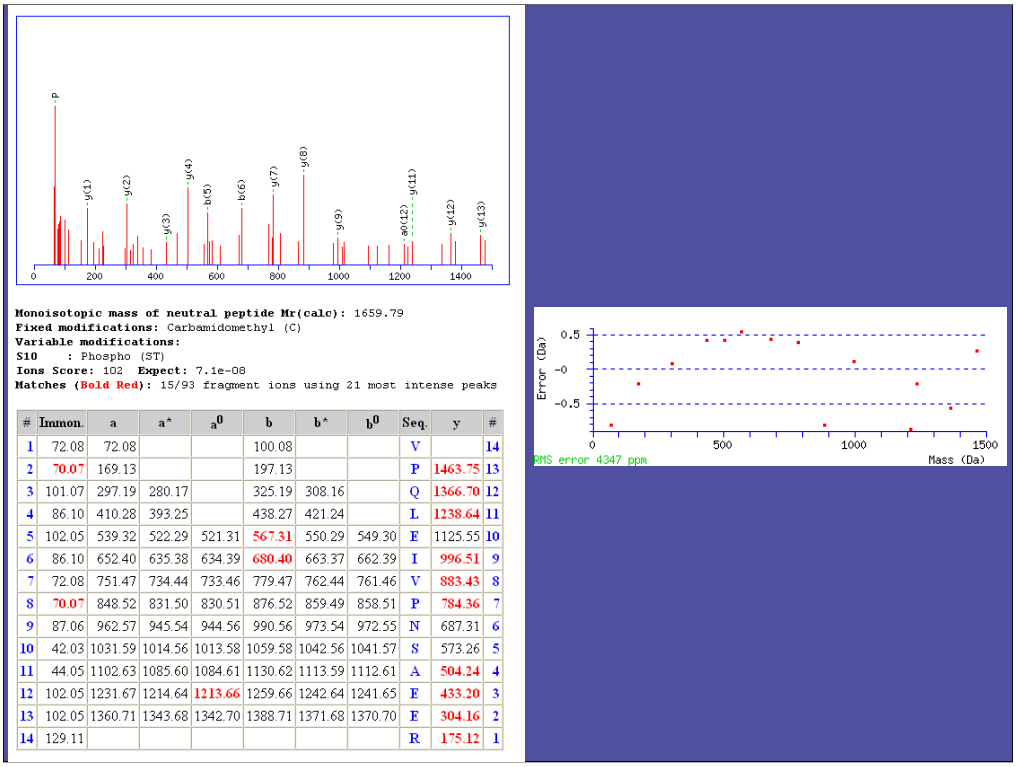
Monoisotopic mass of neutral peptide Mr(calc): 2060.82
Variable modifications:
S3 : Phospho (ST)
Ions Score: 77 **Expect:** 4e-05
Matches (Bold Red): 33/172 fragment ions using 65 most intense peaks

#	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

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



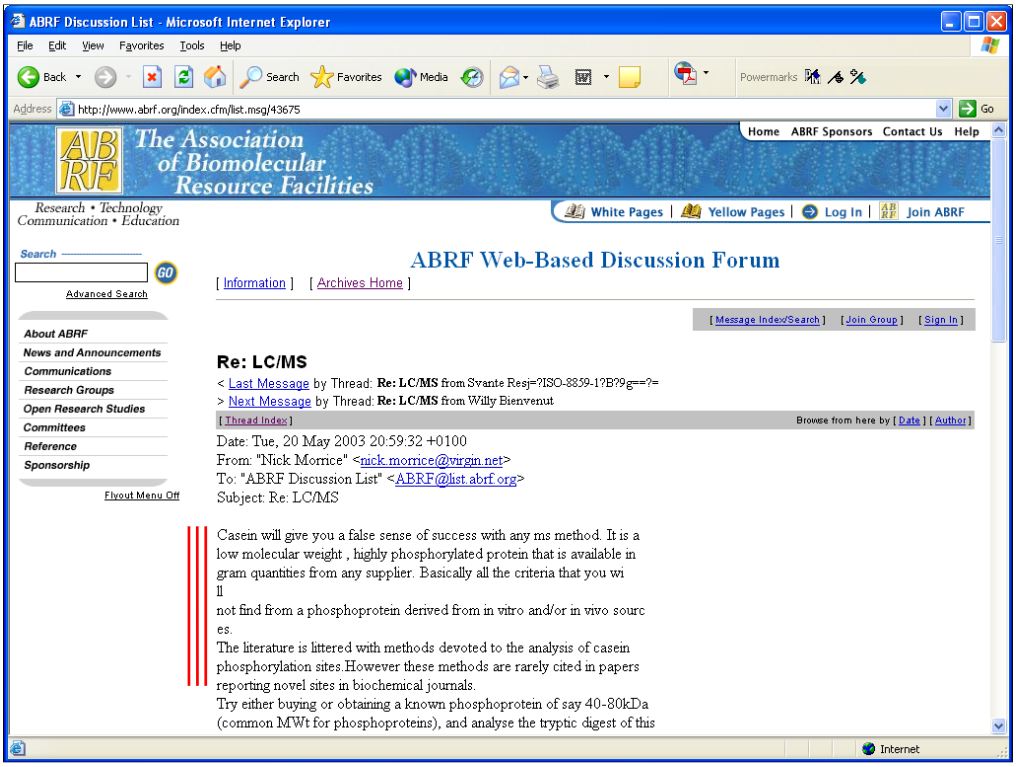
Also, when there is only one site, can get very clear cut matches. Here for example, is a peptide from alpha casein. There is only one potential site and look at the difference in score.



#	Immon.	a	a*	a^0	b	b*	b^0	Seq.	y	#
1	72.08	72.08			100.08			V		14
2	70.07	169.13			197.13			P	1463.75	13
3	101.07	297.19	280.17		325.19	308.16		Q	1366.70	12
4	86.10	410.28	393.25		438.27	421.24		L	1238.64	11
5	102.05	539.32	522.29	521.31	567.31	550.29	549.30	E	1125.55	10
6	86.10	652.40	635.38	634.39	680.40	663.37	662.39	I	996.51	9
7	72.08	751.47	734.44	733.46	779.47	762.44	761.46	V	883.43	8
8	70.07	848.52	831.50	830.51	876.52	859.49	858.51	P	784.36	7
9	87.06	962.57	945.54	944.56	990.56	973.54	972.55	N	687.31	6
10	42.03	1031.59	1014.56	1013.58	1059.58	1042.56	1041.57	S	573.26	5
11	44.05	1102.63	1085.60	1084.61	1130.62	1113.59	1112.61	A	504.24	4
12	102.05	1231.67	1214.64	1213.66	1259.66	1242.64	1241.65	E	433.20	3
13	102.05	1360.71	1343.68	1342.70	1388.71	1371.68	1370.70	E	304.16	2
14	129.11							R	175.12	1

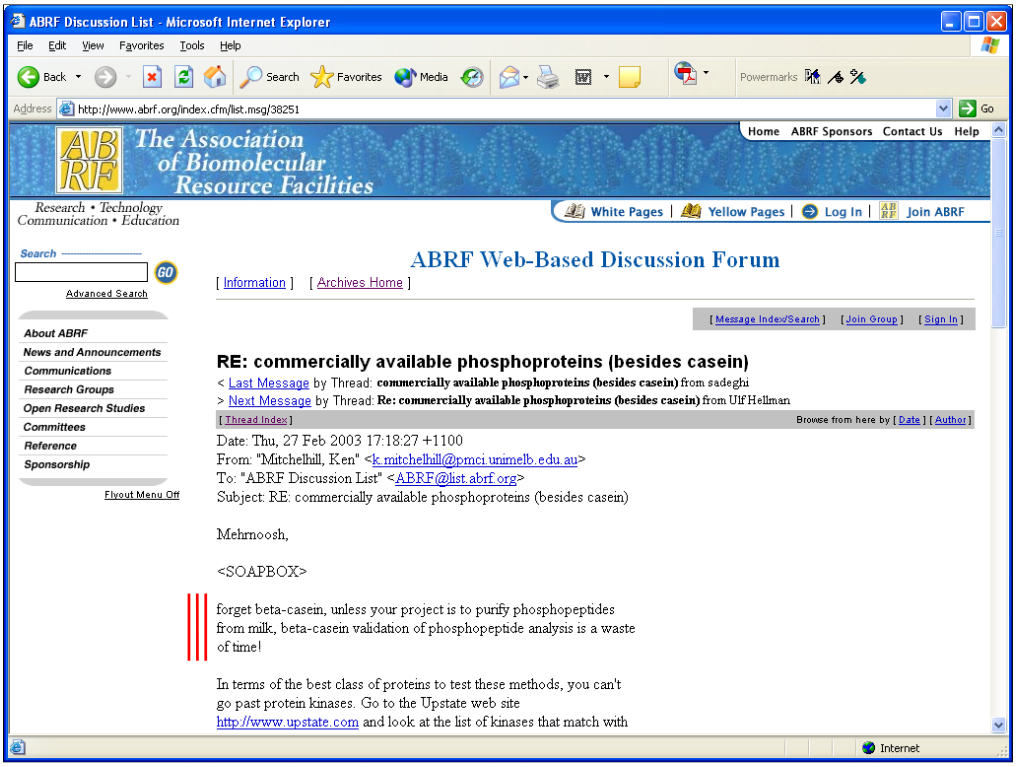
What may surprise you is that the spectrum doesn't look as pretty as the last one. This is MALDI PSD data. In this particular case, the mass accuracy is not so good, +/- 0.5 Da, but Mascot still gets a great match because the sequence coverage is good.

However, phosphopeptides from caseins are notoriously easy ...



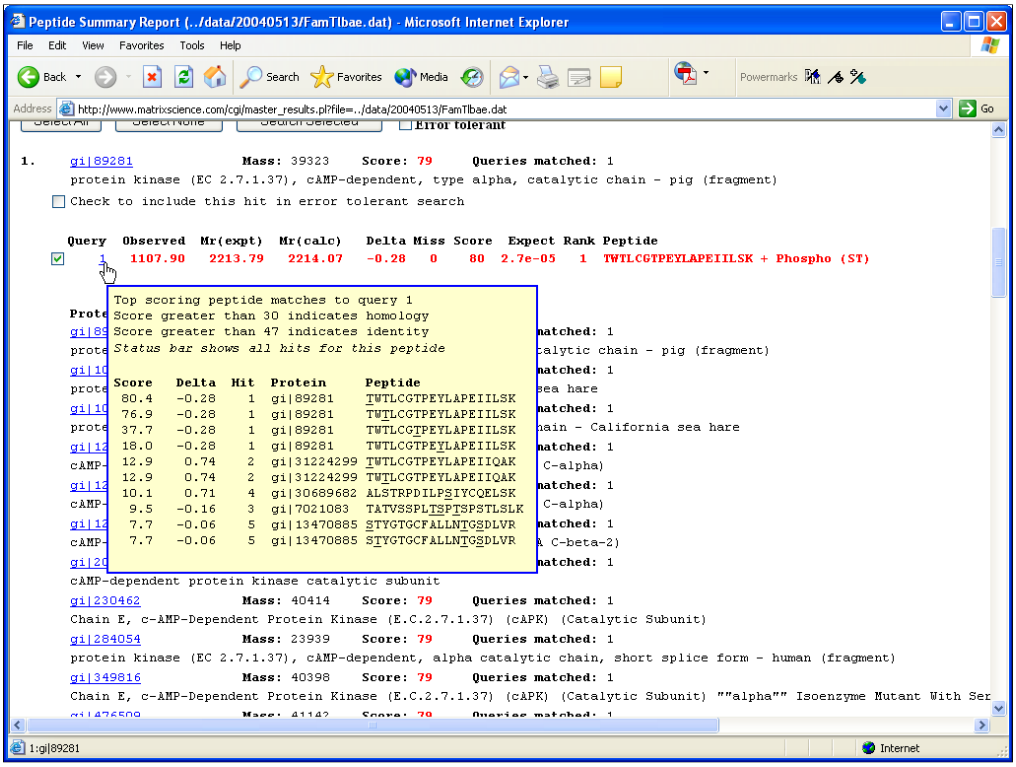
If we go the the ABRF archives, we'll find comments like these from experts in the field:

"Casein will give you a false sense of success with any ms method. It is a low molecular weight , highly phosphorylated protein that is available in gram quantities from any supplier. Basically all the criteria that you will not find from a phosphoprotein derived from in vitro and/or in vivo sources." - Nick Morrice



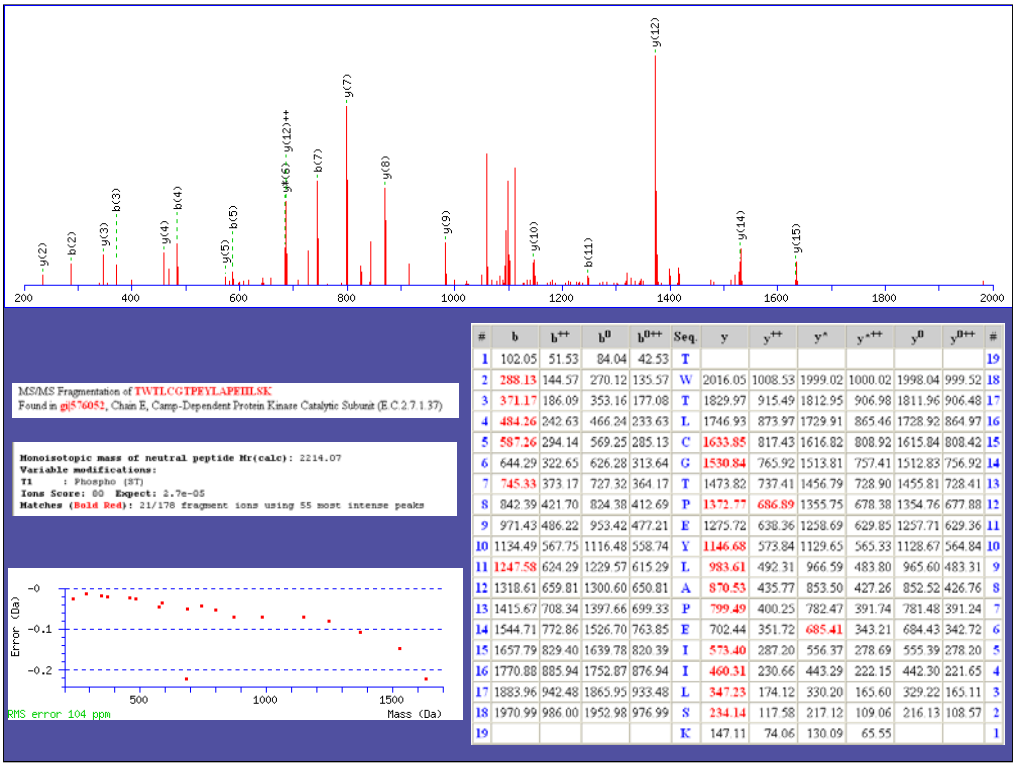
Or: “forget beta-casein, unless your project is to purify phosphopeptides from milk, beta-casein validation of phosphopeptide analysis is a waste of time!” - Ken Mitchelhill

So, no more phosphopeptides from casein in this presentation!



Here is a more realistic example - a strong match to a phosphopeptide from cAMP-dependent `protein kinase`.

This is very typical and reproducible … there is little to choose in terms of score between having the phosphate on T1 or T3. But, we can be very confident it is not on T7 or Y10 because the score drops dramatically



#	b	b^{++}	b^0	b^{0++}	Seq.	y	y^{++}	y^*	y^{*++}	y^0	y^{0++}	#
1	102.05	51.53	84.04	42.53	T							19
2	288.13	144.57	270.12	135.57	W	2016.05	1008.53	1999.02	1000.02	1998.04	999.52	18
3	371.17	186.09	353.16	177.08	T	1829.97	915.49	1812.95	906.98	1811.96	906.48	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

If we compare the two top hits carefully, we can see that there is just one peak difference.

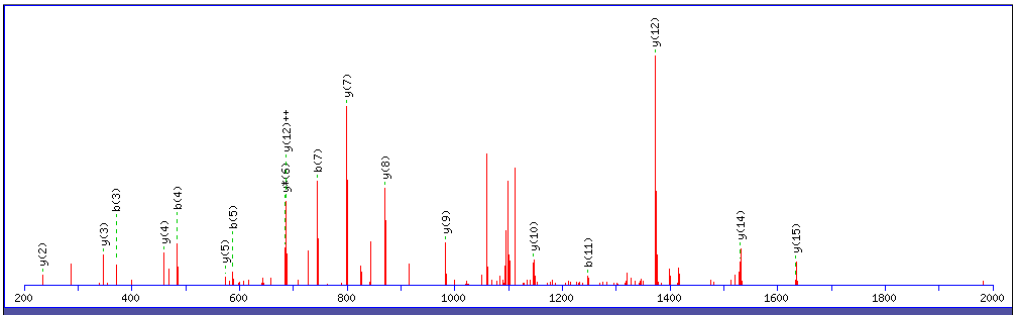
Unfortunately, in this case, the scoring comes mainly from the y ions, so this additional peak has little effect on the score

This highlights the difference between finding a match and verifying it. Mascot is doing a good job of pulling the correct peptide out of all the molecular weight matches in the database.

And, if we accept that this is the correct peptide sequence, and look for evidence to prefer one phosphorylation site over the other, then we immediately focus on this one peak.

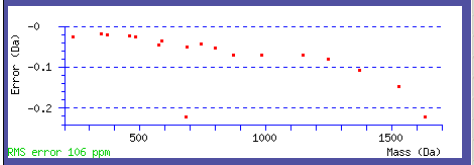
However, there are many unmatched peaks in the spectrum. Also, there is always the possibility of a random match. For example, looking at the mass errors, we might guess that y*(6) is wrong.

If we want to follow through this particular example, we go to the bottom of peptide view

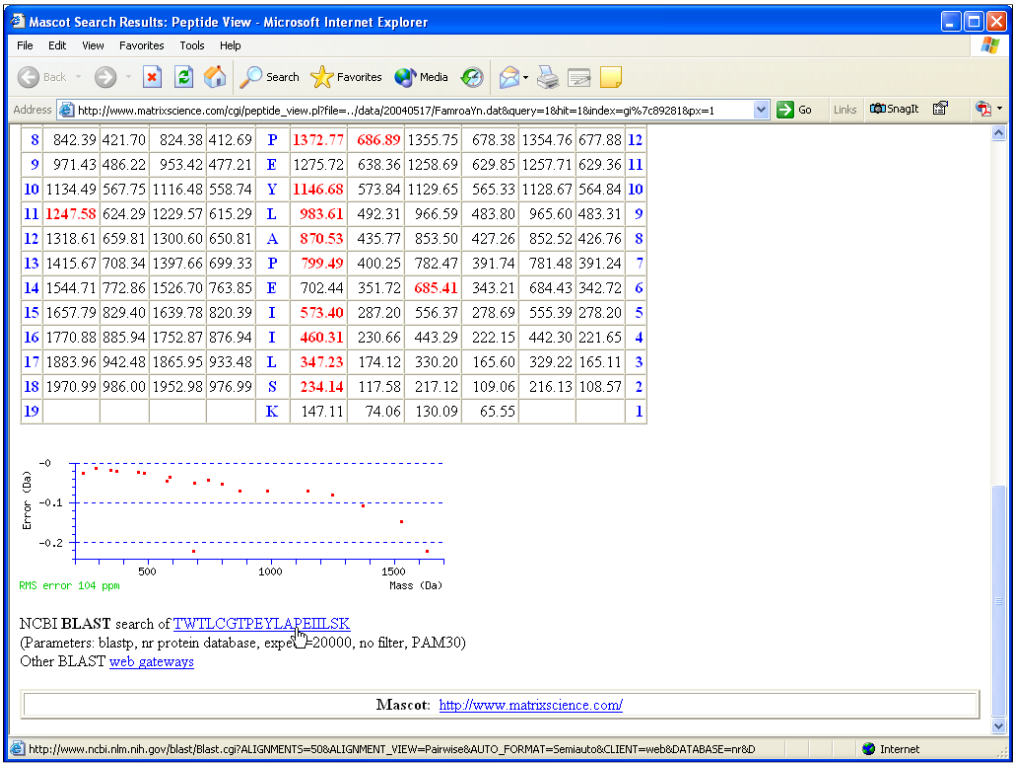


MS/MS Fragmentation of TWTLCGTPEYLAPEIILSK
Found in gi|576052, Chain E, Camp-Dependent Protein Kinase Catalytic Subunit (E.C.2.7.1.37)

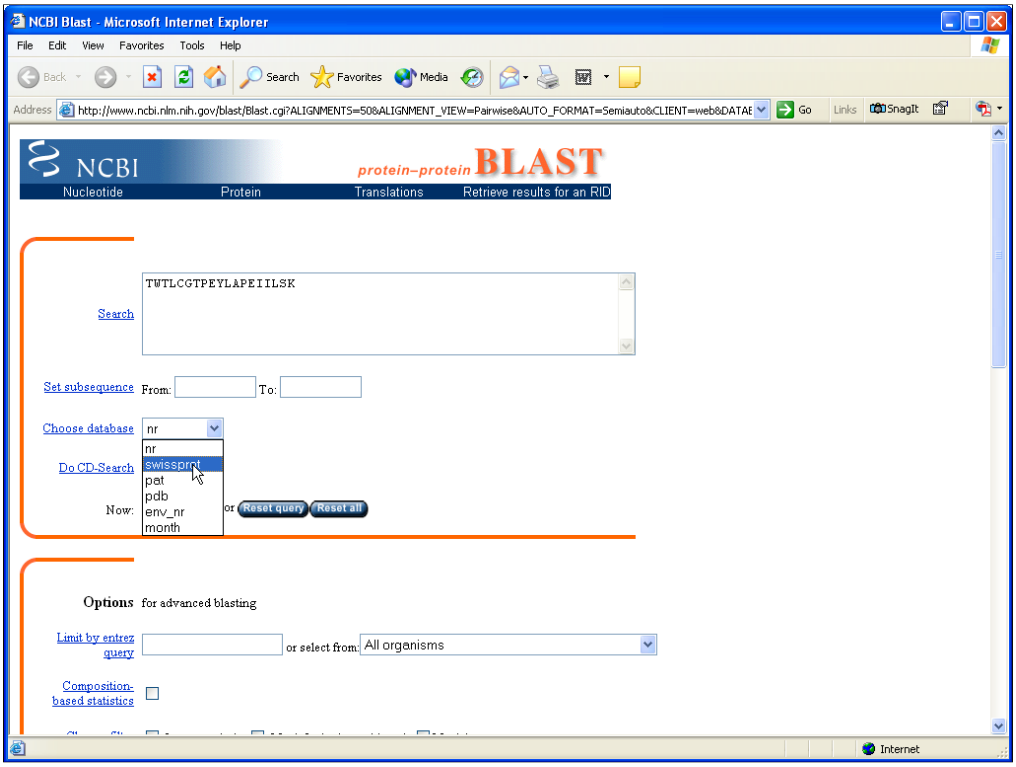
Monoisotopic mass of neutral peptide Mr(calc): 2214.07
Variable modifications:
T3 : Phospho (ST)
Ions Score: 77 **Expect:** 6.1e-05
Matches (Bold Red): 20/178 fragment ions using 55 most intense peaks



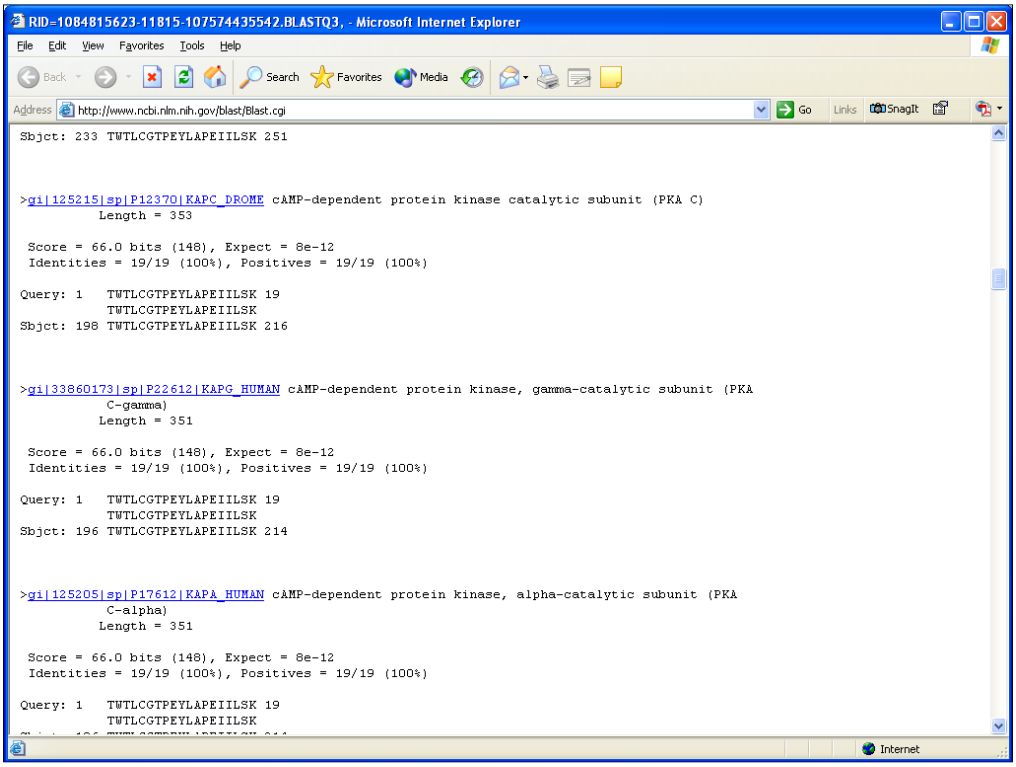
#	b	b^{++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	84.04	42.53	66.03	33.52	T							19
2	270.12	135.57	252.11	126.56	W	2034.06	1017.53	2017.04	1009.02	2016.05	1008.53	18
3	**371.17**	186.09	353.16	177.08	T	1847.98	924.49	1830.96	915.98	1829.97	915.49	17
4	**484.26**	242.63	466.24	233.63	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



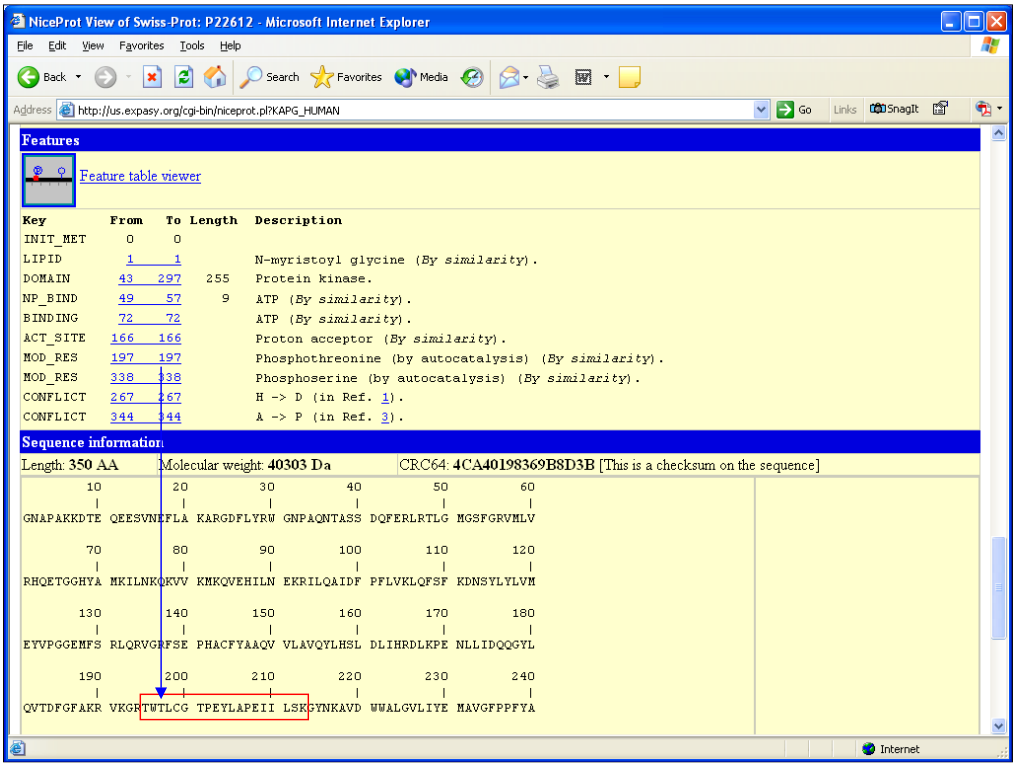
Follow the link to run a Blast search of this peptide



Choose Swiss-Prot as the database, because we want a database with good annotations



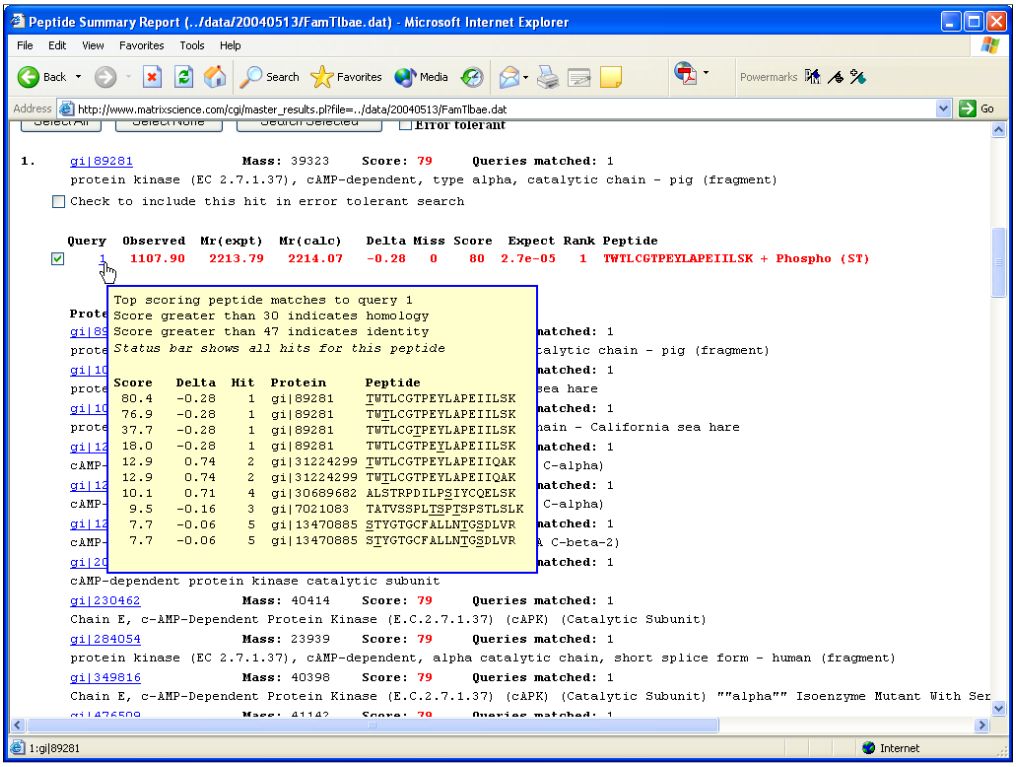
Lots of identity matches to this very common protein. Choose the relevant species, in this case human. Make a mental note of the offset of the peptide in the protein, 196



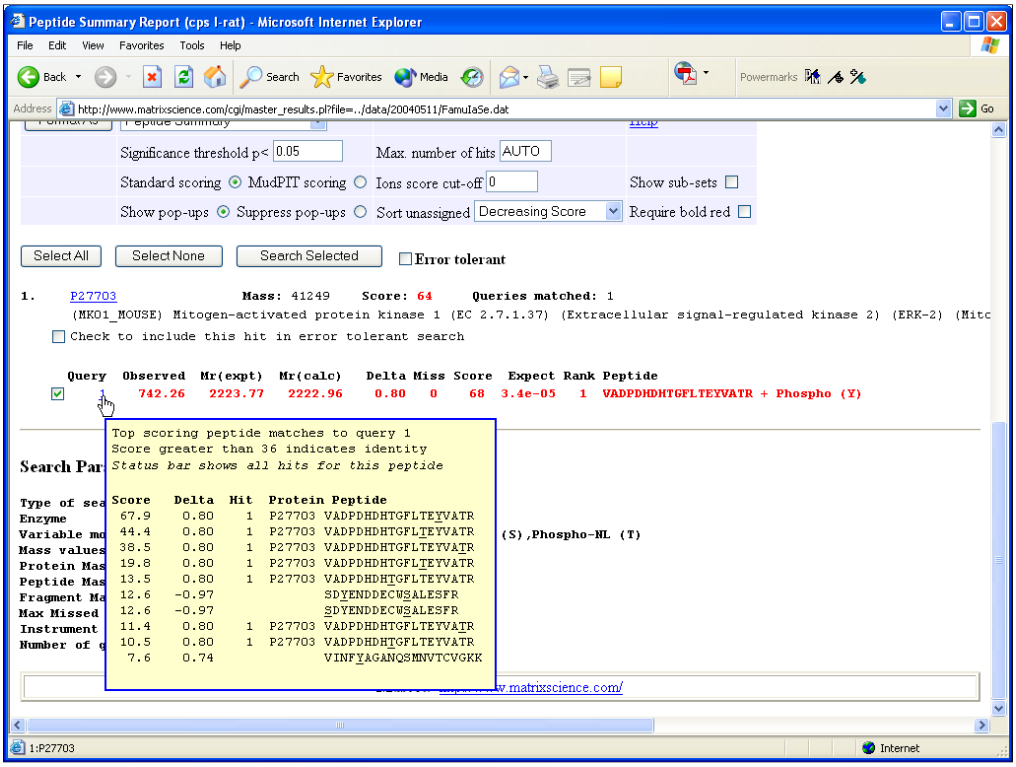
And hop over to Expasy to see the full text for this entry. Here is the peptide. The Expasy numbering has it starting at 195 because the initiator methionine is not present.

But, what is this? According to Swissprot, the phosphate is on T3, not T1!

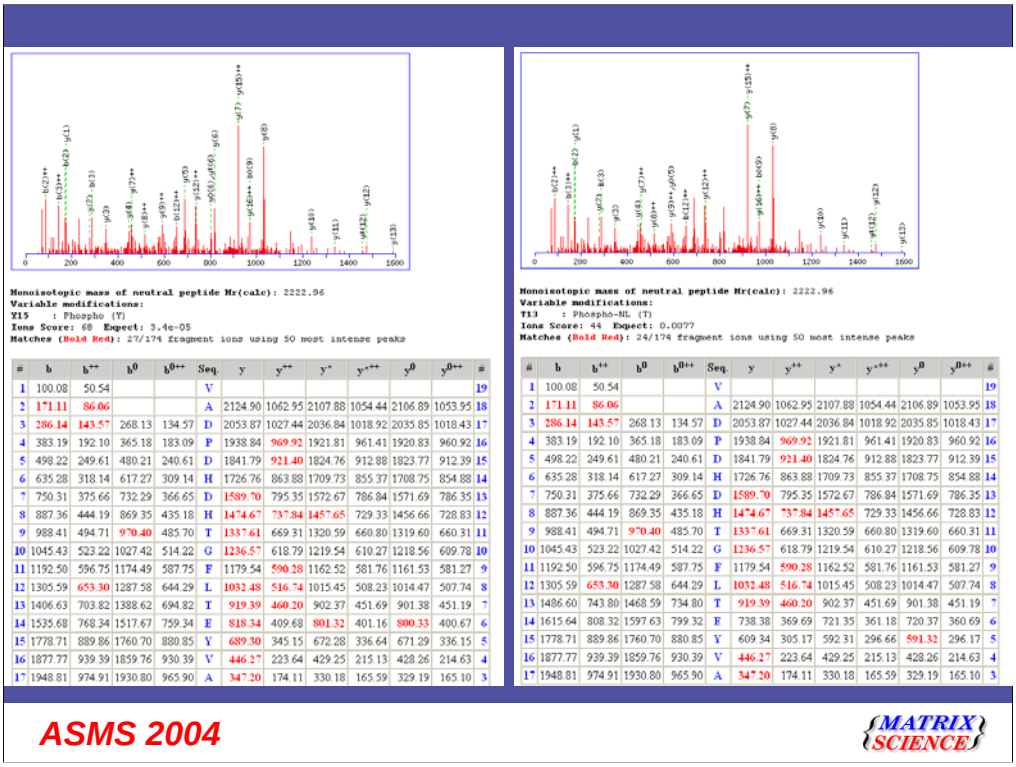
So, either Swiss-Prot is wrong or the extra match in the b series, which looked so convincing is spurious. I've no idea which. But, this does illustrate how easy it is to over-interpret noisy MS data.



The Mascot score reminds us that there is little to choose between T1 and T3. All we can say with confidence is that the phosphate is on one or the other …or maybe there is a mixture of both forms?

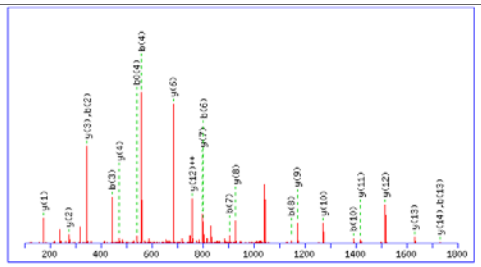


Here is an example where a similar difference, moving the site by 2 residues, causes a much bigger change in the score



This is because there are two additional peaks, rather than one, and they show up in the y series, which is where the score is coming from.

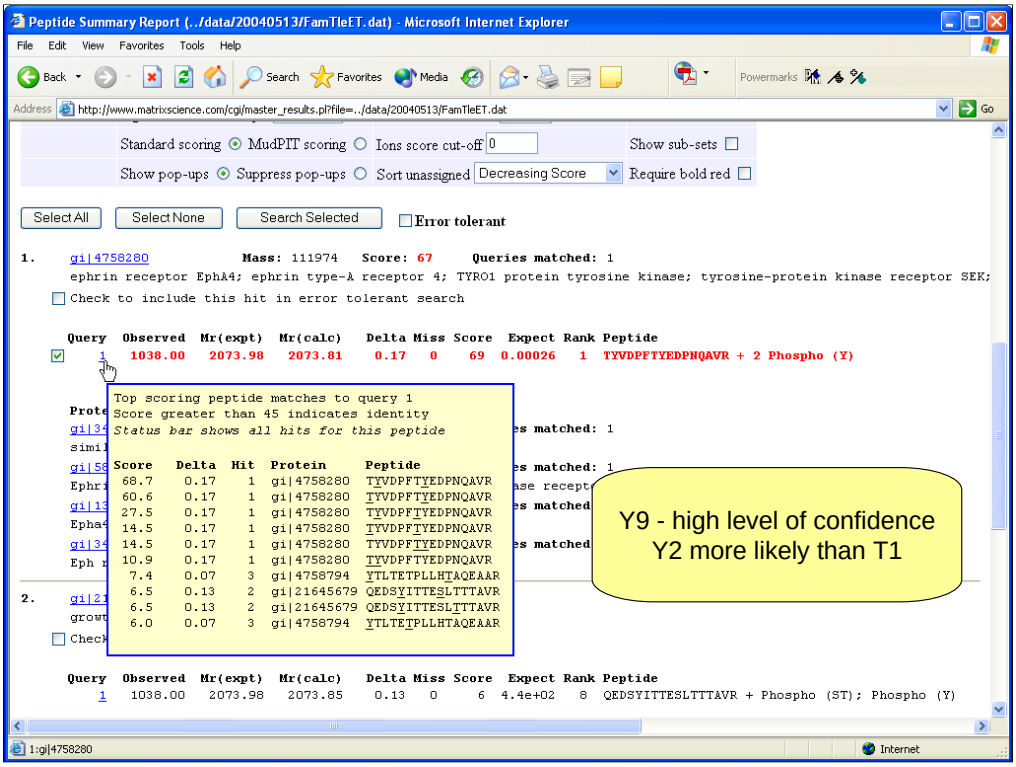
Additional evidence that the phosphate is on Y15 comes from the lack of a strong neutral loss peak for the molecular ion. However, this does not contribute to the Mascot score.



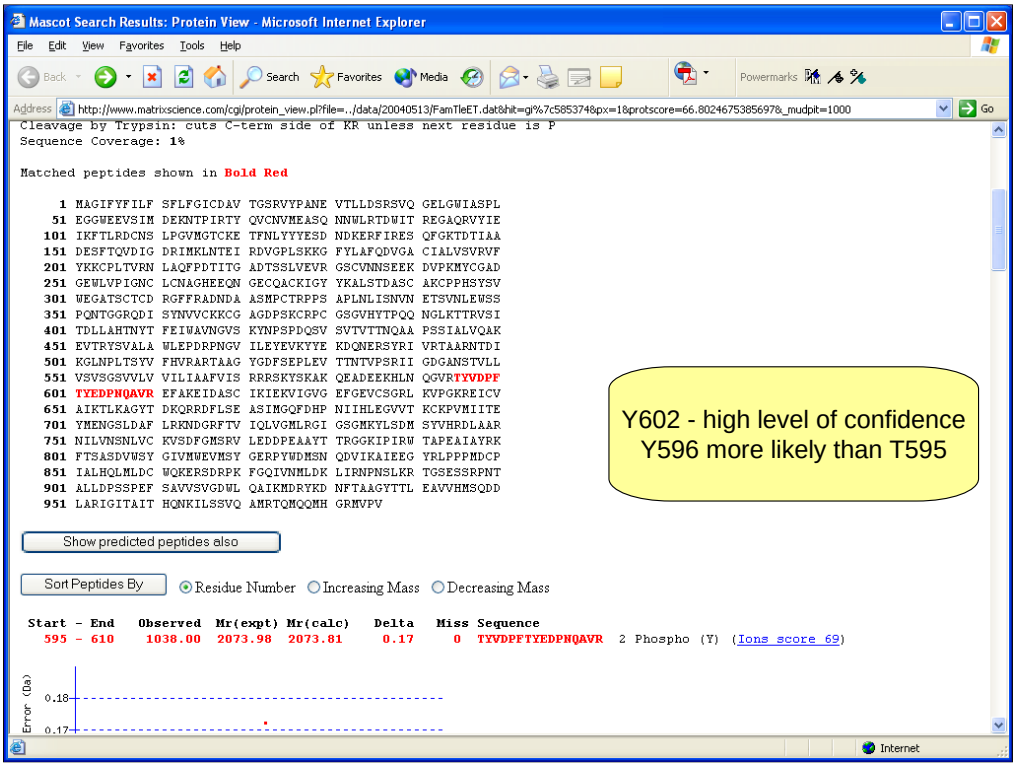
Monoisotopic mass of neutral peptide Mr(calc): 2079.81
Fixed modifications: Acetyl (K)
Variable modifications:
Y2 : Phospho (Y)
Y8 : Phospho (Y)
Ions Score: 69 **Expect:** 0.00026
Matches (Bold Red): 23/146 fragment ions using 54 most intense peaks

#	b	b^{++}	b^*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y^*	y^{*++}	y^0	y^{0++}	#
1	102.05	51.53			84.04	42.53	T							16
2	**345.08**	173.05			327.07	164.04	Y	1973.77	987.39	1956.75	978.88	1955.76	978.38	15
3	**444.15**	222.58			426.14	213.57	V	**1730.74**	865.87	1713.72	857.36	1712.73	856.87	14
4	**559.18**	280.09			**541.17**	271.09	D	**1631.67**	816.34	1614.65	807.83	1613.66	807.34	13
5	656.23	328.62			638.22	319.61	P	**1516.65**	**758.83**	1499.62	750.31	1498.64	749.82	12
6	**803.30**	402.15			785.29	393.15	F	**1419.59**	710.30	1402.57	701.79	1401.58	701.30	11
7	**904.35**	452.68			886.34	443.67	T	**1272.53**	636.77	1255.50	628.25	1254.52	627.76	10
8	**1147.38**	574.19			1129.37	565.19	Y	**1171.48**	586.24	1154.45	577.73	1153.47	577.24	9
9	1276.42	638.71			1258.41	629.71	E	**928.45**	464.73	911.42	456.21	910.44	455.72	8
10	**1391.45**	696.23			1373.44	687.22	D	**799.41**	400.21	782.38	391.69	781.40	391.20	7
11	1488.50	744.75			1470.49	735.75	P	**684.38**	342.69	667.35	334.18			6
12	1602.54	801.78	1585.52	793.26	1584.53	792.77	N	587.33	294.17	570.30	285.65			5
13	**1730.60**	865.80	1713.58	857.29	1712.59	856.80	Q	**473.28**	237.15	456.26	228.63			4
14	1801.64	901.32	1784.61	892.81	1783.63	892.32	A	**345.22**	173.12	328.20	164.60			3
15	1900.71	950.86	1883.68	942.34	1882.70	941.85	V	**274.19**	137.60	257.16	129.08			2
16							R	**175.12**	88.06	158.09	79.55			1

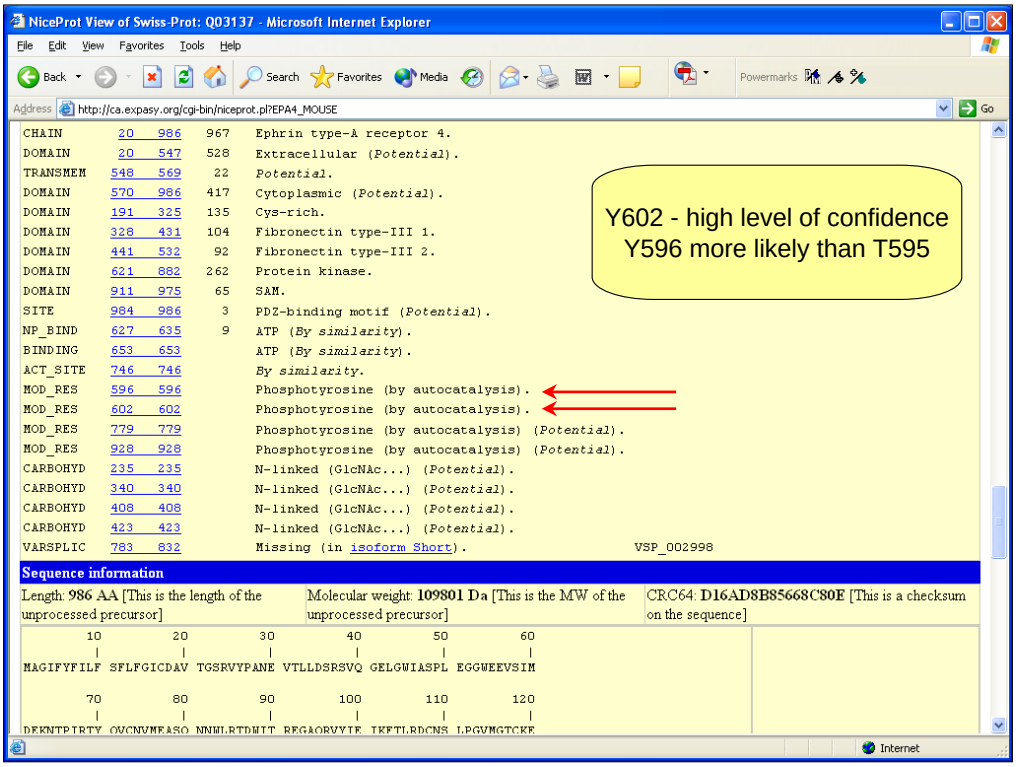
A slightly more complicated example. Two phosphates and three potential sites.



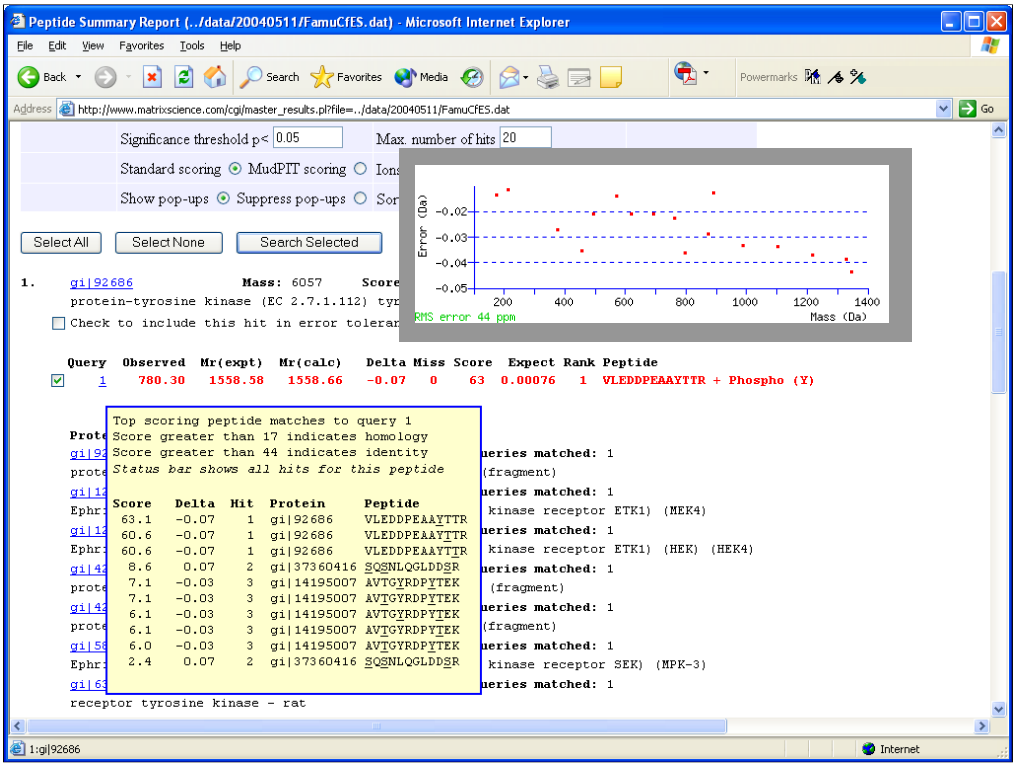
Looking at the scores, we see a big score drop when Y9 is not carrying a phosphate, so we definitely want to put one there. However, not so clear cut whether the other phosphate is on T1 or Y2



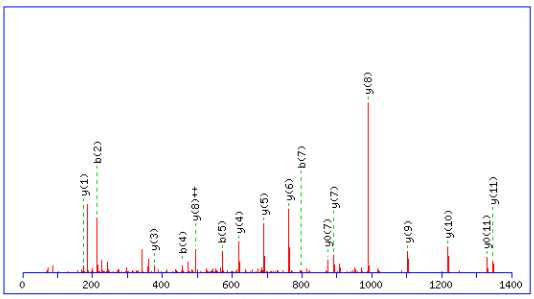
The numbering in the protein is T595 and Y596. Again, if we hop over to Expasy



We find that, this time, we agree. Again, the lack of a neutral loss peak from the molecular ion suggests Y rather than T. However, the Mascot score difference is less than 10, which can come down to just a single extra peak being matched. In the absence of additional evidence, this would be a shaky assignment. My subjective feeling, not very scientific, but based on looking at a range of examples, is that you can be reasonably confident in an assignment when the score difference is 20 or more.



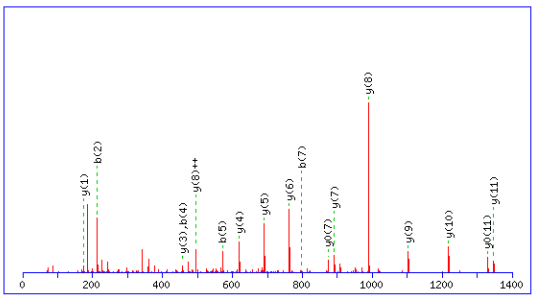
Here's a really difficult example. Great mass accuracy, good sequence coverage, but almost nothing to choose in score between the top 3 matches



Monoisotopic mass of neutral peptide Mr(calc): 1558.66
Variable modifications:
Y10 : Phospho (Y)
Ions Score: 63 **Expect:** 0.00076
Matches (Bold Red): 17/114 fragment ions using 32 most intense peaks

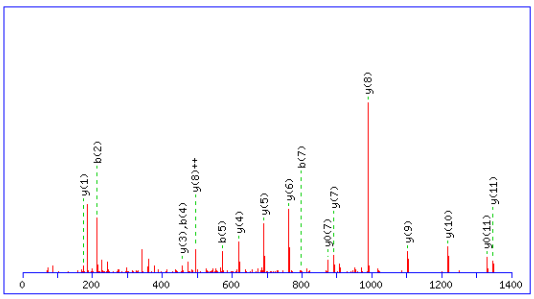
#	b	b^{++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	100.08	50.54			V							13
2	**213.16**	107.08			L	1460.59	730.80	1443.57	722.29	1442.58	721.80	12
3	342.20	171.60	324.19	162.60	E	**1347.51**	674.26	1330.48	665.75	**1329.50**	665.25	11
4	**457.23**	229.12	439.22	220.11	D	**1218.47**	609.74	1201.44	601.22	1200.46	600.73	10
5	**572.26**	286.63	554.25	277.63	D	**1103.44**	552.22	1086.41	543.71	1085.43	543.22	9
6	669.31	335.16	651.30	326.15	P	**988.41**	**494.71**	971.39	486.20	970.40	485.71	8
7	**798.35**	399.68	780.34	390.67	E	**891.36**	446.18	874.33	437.67	**873.35**	437.18	7
8	869.39	435.20	851.38	426.19	A	**762.32**	381.66	745.29	373.15	744.31	372.66	6
9	940.43	470.72	922.42	461.71	A	**691.28**	346.14	674.25	337.63	673.27	337.14	5
10	1183.46	592.23	1165.44	583.23	Y	**620.24**	310.63	603.22	302.11	602.23	301.62	4
11	1284.50	642.76	1266.49	633.75	T	**377.21**	189.11	360.19	180.60	359.20	180.11	3
12	1385.55	693.28	1367.54	684.27	T	276.17	138.59	259.14	130.07	258.16	129.58	2
13					R	**175.12**	88.06	158.09	79.55			1

If we flick between these matches, you can see that there is no evidence in the score to support preferring one of these arrangements over the other two. The absence of a neutral loss peak from the molecular ion would indicate a preference for Y10



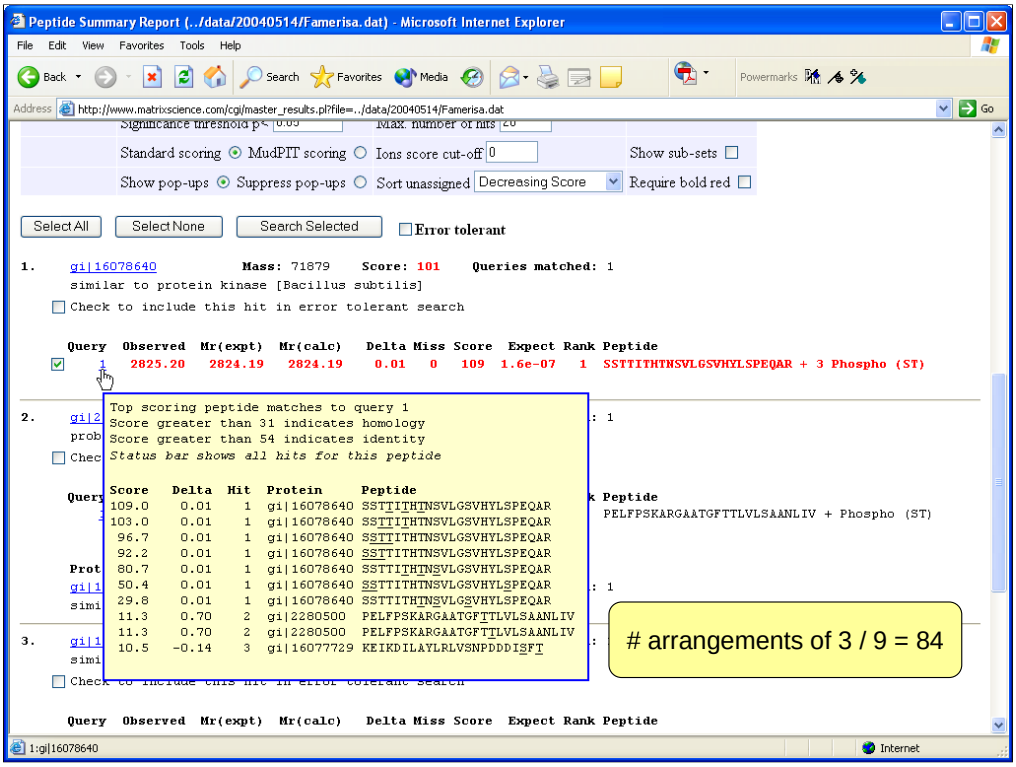
Monoisotopic mass of neutral peptide Mr(calc): 1558.66
Variable modifications:
T11 : Phospho-NL (T)
Ions Score: 61 **Expect:** 0.0013
Matches (Bold Red): 17/114 fragment ions using 32 most intense peaks

#	b	b^{++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	100.08	50.54			V							13
2	**213.16**	107.08			L	1460.59	730.80	1443.57	722.29	1442.58	721.80	12
3	342.20	171.60	324.19	162.60	E	**1347.51**	674.26	1330.48	665.75	**1329.50**	665.25	11
4	**457.23**	229.12	439.22	220.11	D	**1218.47**	609.74	1201.44	601.22	1200.46	600.73	10
5	**572.26**	286.63	554.25	277.63	D	**1103.44**	552.22	1086.41	543.71	1085.43	543.22	9
6	669.31	335.16	651.30	326.15	P	**988.41**	**494.71**	971.39	486.20	970.40	485.71	8
7	**798.35**	399.68	780.34	390.67	E	**891.36**	446.18	874.33	437.67	**873.35**	437.18	7
8	869.39	435.20	851.38	426.19	A	**762.32**	381.66	745.29	373.15	744.31	372.66	6
9	940.43	470.72	922.42	461.71	A	**691.28**	346.14	674.25	337.63	673.27	337.14	5
10	1103.49	552.25	1085.48	543.24	Y	**620.24**	310.63	603.22	302.11	602.23	301.62	4
11	1284.50	642.76	1266.49	633.75	T	**457.18**	229.09	440.15	220.58	439.17	220.09	3
12	1385.55	693.28	1367.54	684.27	T	276.17	138.59	259.14	130.07	258.16	129.58	2
13					R	**175.12**	88.06	158.09	79.55			1



Monoisotopic mass of neutral peptide Mr(calc): 1558.66
Variable modifications:
T12 : Phospho-NL (T)
Ions Score: 61 **Expect:** 0.0013
Matches (Bold Red): 17/114 fragment ions using 32 most intense peaks

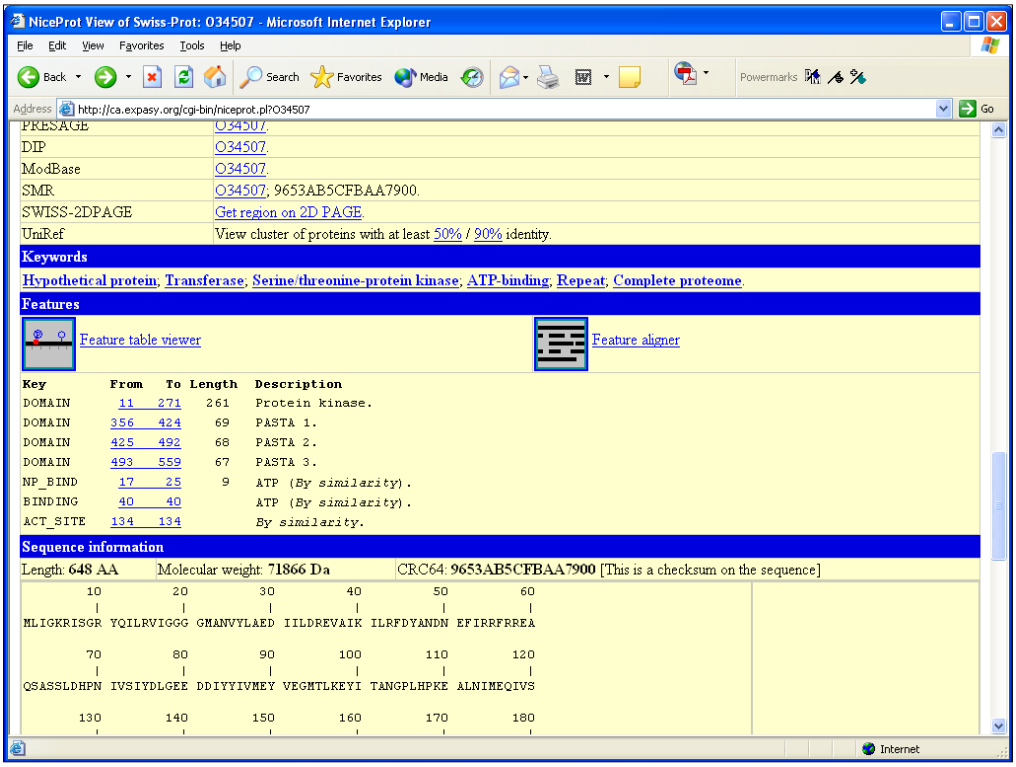
#	b	b^{++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	100.08	50.54			V							13
2	**213.16**	107.08			L	1460.59	730.80	1443.57	722.29	1442.58	721.80	12
3	342.20	171.60	324.19	162.60	E	**1347.51**	674.26	1330.48	665.75	**1329.50**	665.25	11
4	**457.23**	229.12	439.22	220.11	D	**1218.47**	609.74	1201.44	601.22	1200.46	600.73	10
5	**572.26**	286.63	554.25	277.63	D	**1103.44**	552.22	1086.41	543.71	1085.43	543.22	9
6	669.31	335.16	651.30	326.15	P	**988.41**	**494.71**	971.39	486.20	970.40	485.71	8
7	**798.35**	399.68	780.34	390.67	E	**891.36**	446.18	874.33	437.67	**873.35**	437.18	7
8	869.39	435.20	851.38	426.19	A	**762.32**	381.66	745.29	373.15	744.31	372.66	6
9	940.43	470.72	922.42	461.71	A	**691.28**	346.14	674.25	337.63	673.27	337.14	5
10	1103.49	552.25	1085.48	543.24	Y	**620.24**	310.63	603.22	302.11	602.23	301.62	4
11	1204.54	602.77	1186.53	593.77	T	**457.18**	229.09	440.15	220.58	439.17	220.09	3
12	1385.55	693.28	1367.54	684.27	T	356.13	178.57	339.11	170.06	338.12	169.56	2
13					R	**175.12**	88.06	158.09	79.55			1



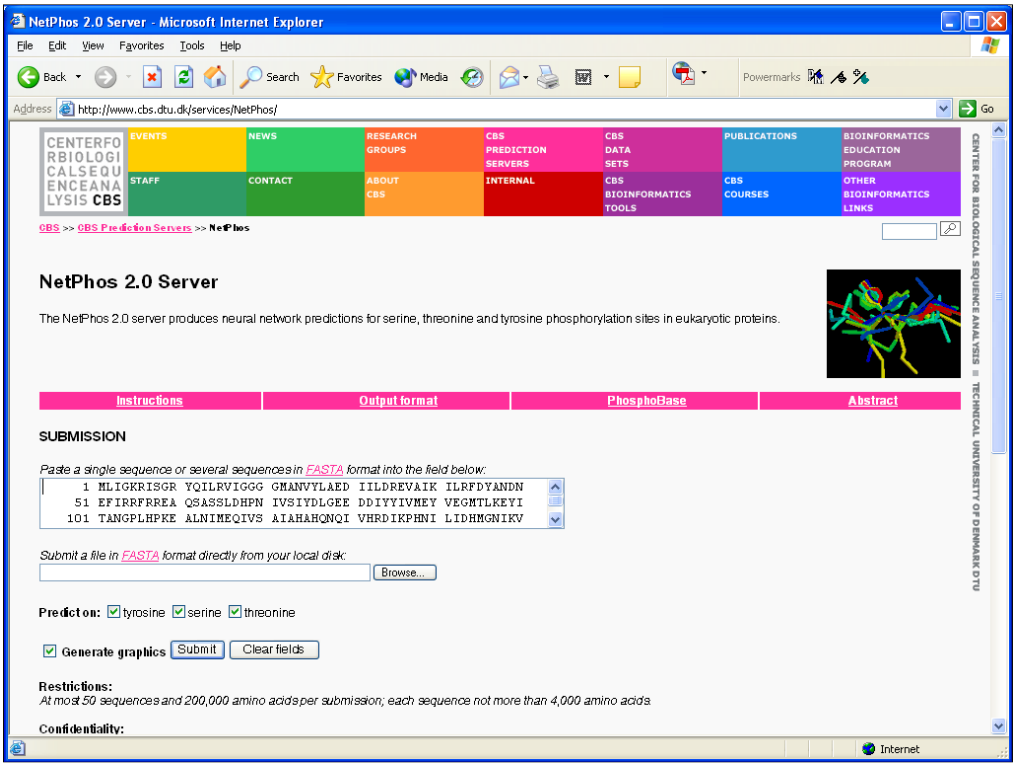
How about this? Three phosphates and 9 possible sites.

This is a nice illustration of why searches with lots of modifications take a long time. For this peptide, there are 84 possible arrangements of the phosphates. And, of course, many of the other peptides that have to be tested will be even worse. Even though they don't show up in the top 10, Mascot still had to work its way through all the different arrangements.

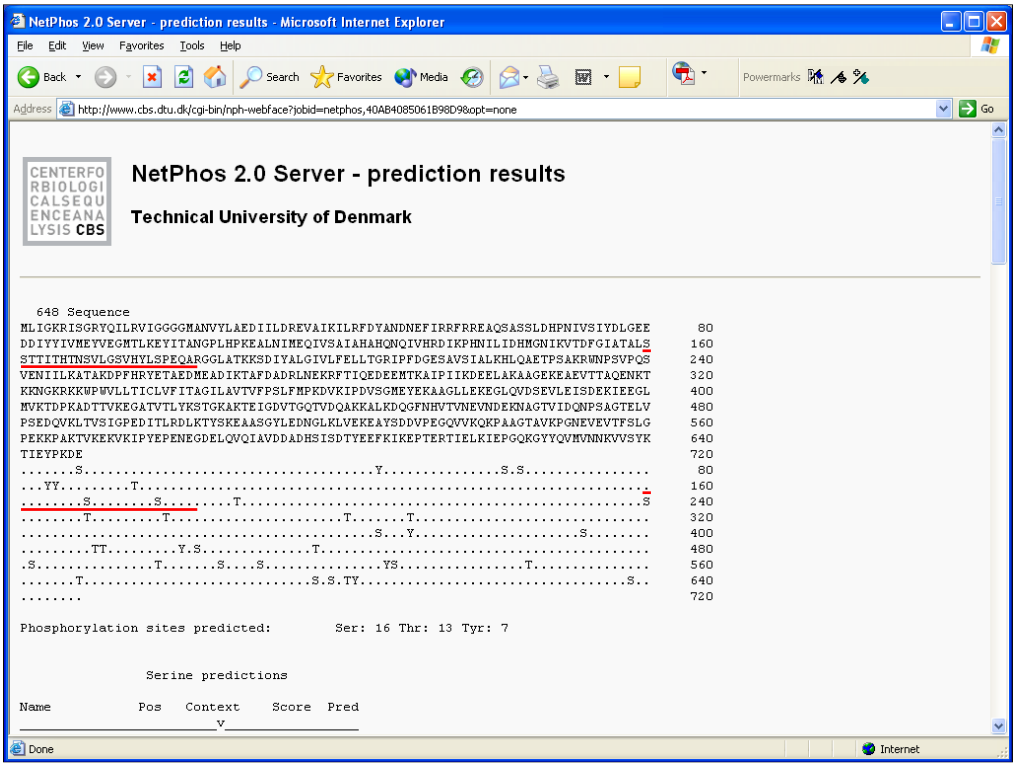
Looking at the scores, I would feel very confident that the phosphorylation is all down towards the N terminus. You can see how the score drops when we try to push phosphate up to S10 or beyond. However, it would be a brave (or foolish) person who claimed they could pin it down more precisely.



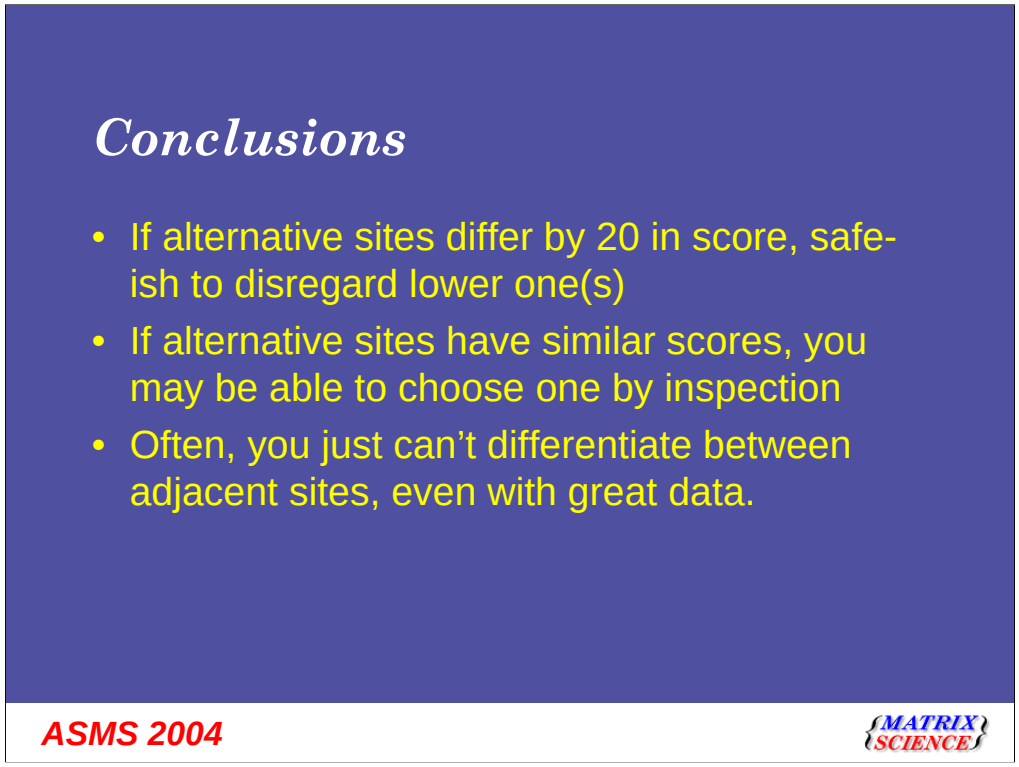
If we follow through the Blast search route and hop over to Expasy, this illustrates another reality of phosphopeptide analysis. The Swissprot entry has no information on phosphorylation for this particular protein.



We can try entering the protein sequence into NetPhos, a popular tool for predicting phosphorylation sites



Unfortunately, we get a prediction that doesn't fit at all well with the MS data The MS evidence puts the 3 phosphates close to the N terminus. It is possible that one could be at S169, but more likely not. And definitely not at S178. So, we're none the wiser in this particular case.



I had hoped to present some clean statistics on how often Mascot called phosphorylation sites correctly. However, I quickly found that, in most cases, it was not possible to find independent evidence on which to base such judgements. It would be a very interesting study … but far more work than we are able to undertake.

So, my subjective conclusions are as follows.

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.

Thank you for your attention