

# Modifications

MASCOT

 MATRIX  
SCIENCE

## 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.

Accession #	PSI-MS Name	Interim name	Description	Monoisotopic mass	Average mass	Composition
200	Ethanedithiol	EDT	EDT	75.960527	76.1838	H(4) C(2) O(-1) S(2)
303	DeStreak	DeStreak	Cysteine mercaptoethanol	75.998285	76.1176	H(4) C(2) O S
208	Delta:H(4)C(6)	Acrolein76	Acrolein addition +76	76.031300	76.0960	H(4) C(6)
653		Ser->Tyr	Ser->Tyr substitution	76.031300	76.0960	H(4) C(6)
1045		Ala->Phe	Ala->Phe substitution	76.031300	76.0960	H(4) C(6)
340	Bromo	bromo	bromination	77.910511	78.8961	H(-1) Br
728		Methylphosphonate	Methylphosphorylation	77.987066	78.0071	H(3) C O(2) P
316	DimethylpyrroleAdduct	pyrrole	2,5-dimethylpyrrole	78.046950	78.1118	H(6) C(6)
423	Delta:Se(1)	selenyl	selenyl	79.916520	78.9600	Se
40	Sulfo	Sulfation	O-Sulfonation	79.956815	80.0632	O(3) S
21	Phospho	Phospho	Phosphorylation	79.966331	79.9799	H O(3) P
1927		Delta:H(4)C(S)O(1)	methyglyoxal-derived argpyrimidine	80.026215	80.0847	H(4) C(5) O
1104		Gly->His	Gly->His substitution	80.037448	80.0880	H(4) C(4) N(2)
837		Arg->Npo	Arginine replacement by Nitropyrimidyl ornithine	80.985078	81.0297	H(-1) C(3) N O(2)
2000		Xlink:SDA	NHS-Diazirine crosslinker	82.041865	82.1005	H(6) C(5) O
549		Cys->Trp	Cys->Trp substitution	83.070128	83.0670	H(5) C(8) N S(-1)
1211		Thr->Trp	Thr->Trp substitution	85.031634	85.1060	H(3) C(7) N O(-1)
211	NEIAA	NEIAA-d0	N-ethyl iodoacetamide-d0	85.052764	85.1045	H(7) C(4) N O
1886		Xlink:BuIrBu[85]	Bu fragment of BuIrBu crosslinker	85.052764	85.1045	H(7) C(4) N O
1052		Ala->Arg	Ala->Arg substitution	85.063997	85.1078	H(7) C(3) N(3)

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.

**UNIMOD** protein modifications for mass spectrometry

Unimod, View record [ Accession #: 56 ]

[Back to list](#)

Accession #	56	PSI-MS Name	Acetyl:2H(3)	Interim Name	Acetyl_heavy
Description	Acetate labeling reagent (N-term & K) (heavy form, +3amu)				
Alt. Description	N-trideuteriumacetoxyl				
Composition	H(-1) 2H(3) C(2) O	Monoisotopic	45.029395	Average	45.0552
<b>Specificity Definition 1</b>					
Site	K	Position	Anywhere	Classification	Isotopic label
<b>Specificity Definition 2</b>					
Site	N-term	Position	Any N-term	Classification	Isotopic label
<b>Specificity Definition 3</b>					
Site	H	Position	Anywhere	Classification	Isotopic label
<b>Specificity Definition 4</b>					
Site	S	Position	Anywhere	Classification	Isotopic label
<b>Specificity Definition 5</b>					
Site	T	Position	Anywhere	Classification	Isotopic label
<b>Specificity Definition 6</b>					
Site	Y	Position	Anywhere	Classification	Isotopic label
<b>Specificity Definition 7</b>					
Site	N-term	Position	Protein N-term	Classification	Isotopic label
<b>Notes and References</b>					
Source	PubMed PMID	Reference	11857757		
Source	PubMed PMID	Reference	11999733		
Source	PubMed PMID	Reference	12175151		
Source	Journal	Reference	Controlling Deuterium isotope effects in comparative proteomics. Zhang, Roujian; Sioma, Cathy S.; Thompson, Robert A.; Xiong, Li; Regnier, Fred E.. Department of Chemistry, Purdue University, West Lafayette, IN, USA. Analytical Chemistry (2002), 74(12), 2649-2654.		
Source	Journal	Reference	Global internal standard technology for comparative proteomics. Chakraborty, Ashish; Regnier, Fred E.. Department of Chemistry, Purdue University, West Lafayette, IN, USA. Journal of Chromatography, A (2002), 949(1-2), 173-184.		
Source	Journal	Reference	Comparative proteomics based on stable isotope labeling and affinity selection. Regnier, Fred E.; Riggs, Larry; Zhang, Roujian; Xiong, Li; Liu, Peiran; Chakraborty, Ashish; Siskley, Erin; Sioma, Cathy; Thompson, Robert A. Department of Chemistry, Pu		
Curator	penner	Last Modified	2011-11-21 10:07:03	Verified	Yes

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.



## Mascot Configuration: Modifications

Displaying 1358/1358

### Visibility:

- Short list
- Long list
- Mixed
- Not listed

### Error tolerant:

- Yes
- No
- Mixed

### Classifications:

- clear
- - Post-translational
  - Co-translational
  - Pre-translational
  - Chemical derivative
  - Artefact

### Source:

- Unimod
- Edited Unimod
- Local

### Apply to selected: 0

- Include in short list
- Include in long list
- Include in error tolerant
- Exclude from error tolerant
- Delete

## Modifications

<input type="checkbox"/> Title ↑	Monoisotopic	Average	Composition	Source	Visibility	Err Tol
<input type="checkbox"/> 15N-oxobutanoic	-18.023584	-18.0239	H(-3) 15N(-1)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 2-dimethylsuccinyl	144.042259	144.1253	H(8) C(6) O(4)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 2-monomethylsuccinyl	130.026609	130.0987	H(6) C(5) O(4)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 2-nitrobenzyl	135.032028	135.1201	H(5) C(7) N O(2)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 2-succinyl	116.010959	116.0722	H(4) C(4) O(4)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 2HPG	282.052824	282.2476	H(10) C(16) O(5)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 3-deoxyglucosone	144.042259	144.1253	H(8) C(6) O(4)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 3-phosphoglyceryl	167.982375	168.0420	H(5) C(3) O(6) P	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 3sulfo	183.983029	184.1693	H(4) C(7) O(4) S	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 4-ONE	154.099380	154.2063	H(14) C(9) O(2)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 4-ONE+Delta:H(-2)O(-1)	136.088815	136.1910	H(12) C(9) O	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> 4AcAllyGal	372.142033	372.3671	H(24) C(17) O(9)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> a-type-ion	-46.005479	-46.0254	H(-2) C(-1) O(-2)	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> AccQTag	170.048013	170.1674	H(6) C(10) N(2) O	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> Acetyl	42.010565	42.0367	H(2) C(2) O	Unimod	mixed	yes <a href="#">Copy Print</a>
<input type="checkbox"/> Acetyl:13C(2)	44.017274	44.0220	H(2) 13C(2) O	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> Acetyl:2H(3)	45.029395	45.0552	H(-1) 2H(3) C(2) O	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> Acetyldeoxyhypusine	97.089149	97.1582	H(11) C(6) N	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> Acetylhypusine	113.084064	113.1576	H(11) C(6) N O	Unimod	long	yes <a href="#">Copy Print</a>
<input type="checkbox"/> ADP-Ribosyl	541.061110	541.3005	H(13) C(10) N(5) O(9) P(2)	Pent Unimod	long	yes <a href="#">Copy Print</a>

Page 1/68 Go to page 15 to AD

<< >>

Page size 20

Add new modification

Main menu

Check Unimod

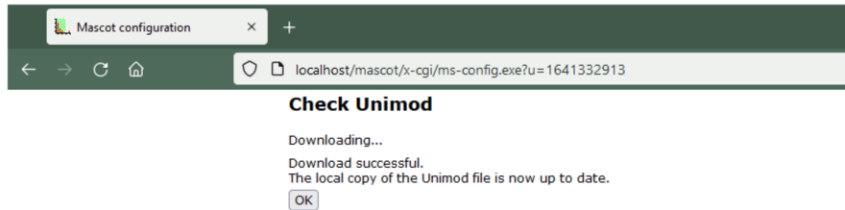
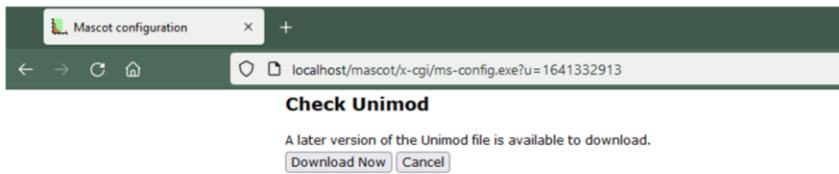
MASCOT

: Modifications

© 2007-2023 Matrix Science

MATRIX SCIENCE

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.

The image shows a screenshot of the Mascot search interface. The browser address bar indicates the URL: `archive-win10/mascot_2_7_00/cgi/search_form.pl?FORMVER=2&SEARCH=`. The search results show a protein entry: `SwissProt UP59680_O_sativa`. The search parameters are: Taxonomy: All entries; Enzyme: Trypsin; Allow up to: 1 missed cleavages; Quantitation: None; Crosslinking: None; Fixed modifications: none selected; Variable modifications: none selected. A red box highlights the `Display all modifications` checkbox, which is checked. The list of modifications displayed includes: 15N-oxobutanoic (N-term C), 15N-oxobutanoic (Protein N-term), 15N-oxobutanoic (Protein N-term), 2-dimethylsuccinyl (C), 2-monomethylsuccinyl (C), 2-nitrobenzyl (Y), 2-succinyl (C), 2HPG (R), 3-deoxyglucosone (R), 3-phosphoglyceryl (K), and 3sulfo (N-term). The bottom of the page features the Mascot logo, the text `: Modifications`, the copyright notice `© 2007-2023 Matrix Science`, and the Matrix Science logo.

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.

## Variable Modification Permutation

KKKSTKKSTKSKSK

Acetyl (K), Phospho (ST)

- 1 x Acetyl (K) : 8 arrangements
- 1 x Phospho (ST) : 6 arrangements
- 2 x Acetyl (K) : 28 arrangements
- 2 x Acetyl (K) + 1 x Phospho (ST) : 168 arrangements
- ... and so on ...

**MASCOT**

: *Modifications*

© 2007-2023 Matrix Science

 **MATRIX  
SCIENCE**

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.

# Be sparing with variable modifications

## ▼ Sensitivity and FDR (reversed protein sequences)

	SwissProt	Decoy	FDR		
Protein family members	28	1	3.57%		
PSMs	above	homology	84	1	1.19%
				Adjust to	1%

Decoy results are available in [the decoy report](#).

Oxidation (M)

4 sec

## ▼ Sensitivity and FDR (reversed protein sequences)

	SwissProt	Decoy	FDR		
Protein family members	26	1	3.85%		
PSMs	above	homology	71	1	1.41%
				Adjust to	1%

Decoy results are available in [the decoy report](#).

Acetyl (K)  
Carboxymethyl (C)  
Me-ester (DE)  
Oxidation (M)  
Phospho (ST)  
Phospho (Y)  
Sodiated (DE)

253 sec

MASCOT : Modifications

© 2007-2023 Matrix Science

MATRIX SCIENCE

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  $\text{HPO}_3$  (80 Da)
- neutral loss of  $\text{H}_3\text{PO}_4$  (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?

The screenshot shows the Matrix Science Mascot web interface. The browser address bar displays the URL: [https://www.matrixscience.com/cgi/master\\_results\\_2.pl?file=%2Fdata%2F2](https://www.matrixscience.com/cgi/master_results_2.pl?file=%2Fdata%2F2). The interface includes a search configuration panel with the following settings:

- Significance threshold p <: 0.05
- Max. number of families: AUTO
- Target FDR (overrides sig. threshold): (not set)
- FDR type: PSM
- Display non-sig. matches:
- Min. number of sig. unique sequences: 1
- Dendrograms cut at: 0
- Preferred taxonomy: All entries

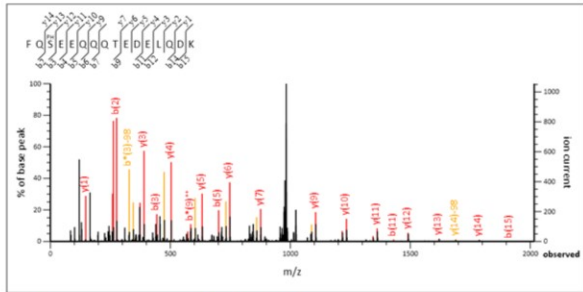
Below the configuration panel, the search results are displayed for "Protein family 1 (out of 1)". The protein is identified as CASB\_BOVIN (Accession: 1.1). The search parameters are: 88 Beta-casein OS=Bos taurus OX=9913 GN=CSN2 PE=1 SV=2. The results show 1 peptide match (1 non-duplicate, 0 duplicate) with an auto-fit to window. The peptide match table is as follows:

Query Dupes	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Expect	Rank	U	Peptide
af1	1031.4000	2060.7854	2060.8212	-0.0357	0	88	6.1e-07	1	U K.FQSERQQQTTEDELQK.I + Phospho (ST)
				-0.0357	0	29	0.57	2	K.FQSERQQQTTEDELQK.I + Phospho (ST)
				-0.1886	1	21	3.3	3	CLSLSKQVILFEETIEK + Phospho (ST)
				-0.0907	1	16	11	4	QMVYKDSPIVDFEIEK + Phospho (ST)
				0.1900	0	15	14	5	EENEEKQGRQSEK + Phospho (ST)
				+0.1713	1	14	16	6	QLASGEYFANQGRQAK + Phospho (ST)
				-0.1489	1	14	18	7	ITFLEELYKQDQNEK + Phospho (ST)
				-0.2007	1	12	24	8	SLQEGEGLFVAERLSEK
				-0.1635	1	12	26	9	YLILCVGRTLNRRSEK + Phospho (ST)
				-0.1965	1	12	29	10	YDSFFRFDIVVVIGAK + Phospho (ST)

The Mascot logo and "Modifications" text are visible in the bottom left corner. The copyright notice "© 2007-2023 Matrix Science" is in the bottom center. The Matrix Science logo is in the bottom right corner.

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.



Monoisotopic mass of neutral peptide Mr(calc): 2060.8212  
 Variable modifications:  
 83 ± Phospho (PT), 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 <sup>++</sup>	b <sup>*</sup>	b <sup>+++</sup>	b <sup>0</sup>	b <sup>0++</sup>	Seq.	y	y <sup>++</sup>	y <sup>*</sup>	y <sup>+++</sup>	y <sup>0</sup>	y <sup>0++</sup>	#
1	148.0757	74.5415					F	1914.7601	957.8837	1897.7335	949.3704	1896.7495	948.8784	16
2	<b>276.1343</b>	138.5708	<b>259.1077</b>	130.0575			Q	<b>1766.7015</b>	893.8544	1769.6749	885.3411	1768.6909	884.8491	14
3	<b>443.1326</b>	222.0700	426.1061	213.5567	425.1221	213.0647	S	<b>1619.7031</b>	810.3552	1602.6766	801.8419	1601.6926	801.3499	13
4	<b>572.1757</b>	286.5912	555.1487	278.0780	554.1647	277.5860	E	<b>1490.6605</b>	745.8339	1473.6340	737.3206	1472.6500	736.8286	12
5	<b>701.2178</b>	351.1125	684.1913	342.5993	683.2072	342.1073	E	<b>1361.6179</b>	681.3126	<b>1344.5914</b>	672.7993	1343.6074	672.3073	11
6	829.2764	415.1418	812.2498	406.6286	811.2658	406.1366	Q	<b>1223.5594</b>	617.2833	<b>1216.5328</b>	608.7700	1215.5488	608.2780	10
7	957.3350	479.1711	940.3084	470.6578	939.3244	470.1658	Q	<b>1165.5008</b>	583.2540	<b>1088.4742</b>	544.7408	1087.4902	544.2487	9
8	1085.3935	543.2004	1068.3670	534.6871	1067.3830	534.1951	Q	977.4422	489.2247	960.4156	480.7115	959.4316	480.2195	8
9	1186.4412	593.7243	1169.4147	<b>585.2110</b>	1168.4307	584.7190	T	<b>876.3945</b>	438.7009	<b>859.3680</b>	430.1876	858.3840	429.6956	7
10	1315.4838	658.2455	1298.4573	649.7325	1297.4733	649.2403	E	<b>747.3519</b>	374.1796	730.3254	365.6663	729.3414	365.1743	6
11	<b>1430.5108</b>	715.7590	1413.4842	707.2457	1412.5002	706.7537	D	<b>632.3259</b>	316.6661	615.2984	308.1529	614.3144	307.6608	5
12	1559.5534	780.2803	1542.5268	771.7670	1541.5428	771.2750	L	<b>563.2824</b>	282.1448	546.2558	274.6316	545.2718	274.1396	4
13	1672.6374	836.8223	1655.6109	828.3091	1654.6269	827.8171	L	<b>390.1982</b>	195.6028	<b>373.1718</b>	187.0895	372.1878	186.5975	3
14	<b>1800.6960</b>	900.8516	1783.6694	892.3384	1782.6854	891.8464	Q	<b>262.1397</b>	131.5735	245.1132	123.0602	244.1292	122.5682	2
15	<b>1915.7229</b>	958.3651	1898.6964	949.8518	1897.7124	949.3598	D	<b>147.1128</b>	74.0600	130.0863	65.5468			1

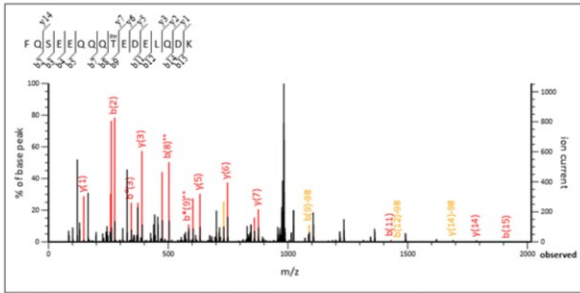
Label all possible matches  Label matches used for scoring

MASCOT : Modifications

© 2007-2023 Matrix Science

MATRIX SCIENCE

Beautiful spectrum; long run of y ions.



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 <sup>+</sup>	b <sup>+</sup>	b <sup>++</sup>	b <sup>0</sup>	b <sup>0++</sup>	Seq.	y	y <sup>+</sup>	y <sup>+</sup>	y <sup>++</sup>	y <sup>0</sup>	y <sup>0++</sup>	#
1	148.0757	74.5415					F							16
2	<b>276.1343</b>	138.5708	<b>259.1077</b>	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	<b>345.1557</b>	173.0815	S	<b>2786.7025</b>	893.8544	1769.6749	885.3411	1768.6909	884.8491	14
4	492.2089	246.6081	475.1823	238.0948	<b>474.1983</b>	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	<b>603.2409</b>	302.1241	Q	1570.6269	785.8171	1553.6003	777.3038	1552.6163	776.8118	12
6	749.3101	375.1587	732.2835	366.6454	<b>731.2993</b>	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	<b>859.3581</b>	430.1827	Q	1313.5257	657.2665	1296.4991	648.7532	1295.5151	648.2612	10
8	1005.4272	<b>503.2172</b>	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	<b>583.2110</b>	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	<b>876.3945</b>	438.7009	<b>859.3680</b>	430.1876	858.3840	429.6956	7
11	<b>1430.5108</b>	715.7590	1413.4842	707.2457	1412.5002	706.7537	D	<b>747.3519</b>	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	<b>632.3250</b>	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	<b>503.2824</b>	252.1448	486.2558	243.6316	485.2718	243.1396	4
14	<b>1800.6960</b>	900.8516	1783.6694	892.3384	1782.6854	891.8464	Q	<b>390.1983</b>	195.6028	<b>373.1718</b>	187.0895	372.1878	186.5975	3
15	<b>1915.7229</b>	958.3651	1898.6964	949.8518	1897.7124	949.3598	D	<b>262.1397</b>	131.5735	245.1132	123.0602	244.1292	122.5682	2
16							K	<b>147.1128</b>	74.0600	130.0863	65.5468			1

Label all possible matches  Label matches used for scoring

MASCOT

: Modifications

© 2007-2023 Matrix Science

MATRIX SCIENCE

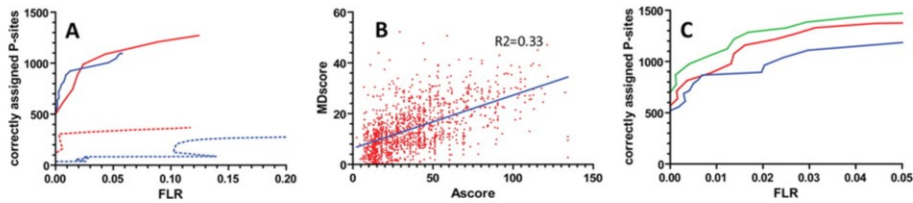
Move site to T9 and many matches would disappear.

# Confident Phosphorylation Site Localization Using the Mascot Delta Score<sup>®</sup>

Mikhail M. Savitski<sup>‡</sup>, Simone Lemeer<sup>§</sup>, Markus Boesche<sup>‡</sup>, Manja Lang<sup>‡</sup>,  
Toby Mathieson<sup>‡</sup>, Marcus Bantscheff<sup>‡||</sup>, and Bernhard Kuster<sup>§||</sup>

*Molecular & Cellular Proteomics* 10.2

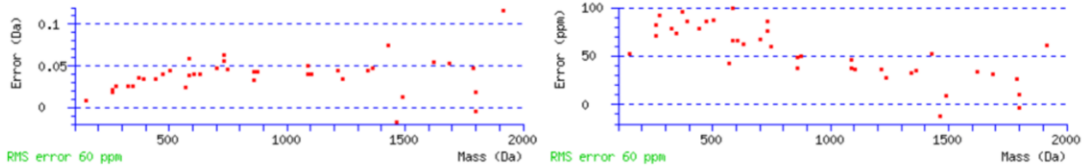
10.1074/mcp.M110.003830-1



**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 [FQSEEQQTTEDELQDK](#)  
 (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	<a href="#">FQSEEQQTTEDELQDK</a>	Phospho S3 100.00%
28.6	2060.8212	-0.0357	<a href="#">FQSEEQQTTEDELQDK</a>	Phospho T9 0.00%
21.0	2060.9741	-0.1886	<a href="#">CLSLSKQVDLFEETIEK</a>	
15.9	2060.8762	-0.0907	<a href="#">QMVVVDKDSPHVEPEDEK</a>	
14.5	2060.6855	0.1000	<a href="#">EENEEEQDDDEQSEEK</a>	
14.1	2060.9568	-0.1713	<a href="#">QLASGEYFLNQEQKQAK</a>	
13.6	2060.9343	-0.1489	<a href="#">ITFLEELYPKDQDNEK</a>	
12.4	2060.9862	-0.2007	<a href="#">SLQEGEGDLSVAEDRLSEK</a>	
11.9	2060.9489	-0.1635	<a href="#">YLILCVGETLNERDSEK</a>	
11.6	2060.9820	-0.1965	<a href="#">YDSFPRSDIVTVVIGADK</a>	

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

**Protein Family Summary**

Format: [dropdown] Significance threshold p<: 0.05 Max. number of families: 20 [help]  
 Target FDR (overrides sig. threshold): (not set) FDR type: PSM  
 Display non-sig. matches: [checkbox] Min. number of sig. unique sequences: 1  
 Preferred taxonomy: All entries Dendrograms cut at: 0

**Sensitivity and FDR (reversed protein sequences)**

Proteins (1) [Report Builder](#) [Permalink](#)

**Protein family 1 (out of 1)**

10 per page 1 Expand all Collapse all

Accession contains Find Clear

**KAPCA\_BOVIN** 80 cAMP-dependent protein kinase catalytic subunit alpha OS=Bos taurus OX=9913 GN=PRKACA PE=1 SV=3

1.1 [KAPCA\\_BOVIN](#) Score 80 Mass 40594 Matches 1 (1) Sequences 1 (1) cAMP-dependent protein kinase catalytic subunit alpha OS=Bos taurus OX=9913 GN=PRKACA PE=1 SV=3

**17 samesets of KAPCA\_BOVIN**

**1 peptide matches (1 non-duplicate, 0 duplicate)**

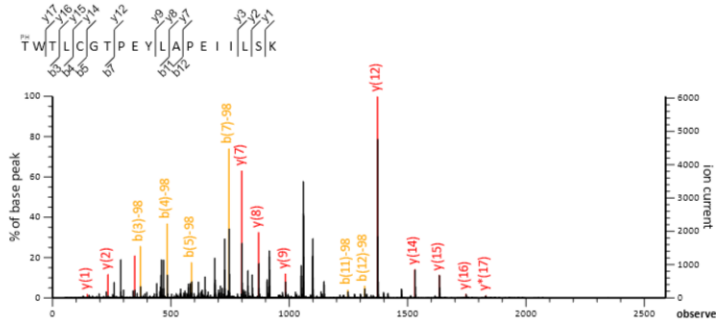
Auto-fit to window

Query Dupes	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Expect	Rank	U	Peptide
									Query 1
									Score > 47 indicates identity
									Score > 30 indicates homology
	1107.9039	2213.7933	2214.0683	-0.2750	80	5.1e-07	1	U	R.TWTLGTFEYLAEIILK.G + Phospho (ST)
				-0.2750	77	1.1e-06	2		R.TWTLGTFEYLAEIILK.G + Phospho (ST)
				-0.2750	39	0.0076	3		R.TWTLGTFEYLAEIILK.G + Phospho (ST)
				-0.2750	18	0.88	4		R.TWTLGTFEYLAEIILK.G + Phospho (T)
				-0.2111	13	3.1	5		GGSGMLTGLPESPOVFALEK + 2 Phospho (ST)
				-0.2111	13	3.1	5		GGSGMLTGLPESPOVFALEK + 2 Phospho (ST)
				-0.2111	13	3.1	5		GGSGMLTGLPESPOVFALEK + 2 Phospho (ST)
				-0.2111	12	3.6	9		GGSGMLTGLPESPOVFALEK + 2 Phospho (ST)
				-0.2111	12	3.6	9		GGSGMLTGLPESPOVFALEK + 2 Phospho (ST)

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

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.



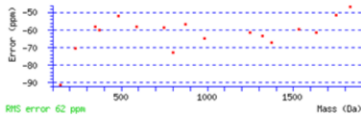
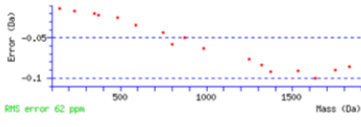
MS/MS Fragmentation of **TWTLCTPEYLAPLEILSK**  
 Found in **KAPCA\_BOVIN** in **SwissProt**, cAMP-dependent protein kinase catalytic subunit alpha OS=Bos taurus OX=9913 GN=PRKACA PE=1 SV=3

**Monoisotopic mass of neutral peptide Mr(calc)**: 2214.0683

**Variable modifications**:  
 T1: Phospho (ST), with neutral losses 97.9769 (shown in table), 0.0000

**Ions Score**: 80 **Expect**: 2.7e-05

**Matches**: 17/250 fragment ions using 21 most intense peaks ([help](#))



#	b	b**	b <sup>0</sup>	b**	Seq	y	y**	y*	y**	y <sup>0</sup>	y**	#
1	84.0444	42.5258	66.0338	33.2206	T							19
2	270.1237	135.5655	252.1131	126.5602	W	2034.0616	1017.5345	2017.0351	1009.0212	2016.0311	1008.5292	18
3	371.1714	186.0893	353.1608	177.0840	T	1847.9823	924.4948	1830.9538	915.9815	1829.9718	915.4895	17
4	484.2554	242.6314	466.2449	233.6261	L	1746.9348	873.9710	1729.9081	865.4577	1728.9241	864.9657	16
5	587.3646	294.1360	569.2541	283.1307	C	1633.8368	817.4209	1616.8240	808.9156	1615.8400	808.4236	15
6	644.2861	322.6467	638.2751	313.6414	G	1538.8412	769.9243	1513.8148	757.4111	1512.8208	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	1372.7222	686.8698	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	988.6336	492.3104	966.5870	483.7971	965.6030	483.3051	9
12	1318.6136	659.8105	1300.6031	650.8052	A	876.5293	435.7684	853.5029	427.2551	852.5189	426.7631	8
13	1415.6664	708.3368	1397.6538	699.3316	P	799.4924	400.2498	782.4658	391.7366	781.4818	391.2445	7
14	1544.7090	772.8381	1528.6984	763.8529	E	702.4396	351.2244	685.4131	343.2102	684.4291	342.7382	6
15	1657.7931	829.4002	1639.7825	820.3849	I	573.3970	287.2021	556.3705	278.6889	555.3865	278.1969	5
16	1770.8771	885.9422	1752.8666	876.9369	L	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

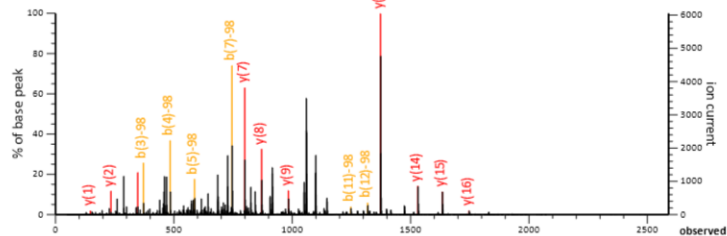
**MASCOT** : Modifications

© 2007-2023 Matrix Science



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





**MS/MS Fragmentation of TWTL**CGTPEY**LAPEIIL**LSIK****

Found in **KAPCA\_BOVIN** in **SwissProt**, cAMP-dependent protein kinase catalytic subunit alpha OS=Bos taurus OX=9913 GN=PRKACA PE=1 SV=3

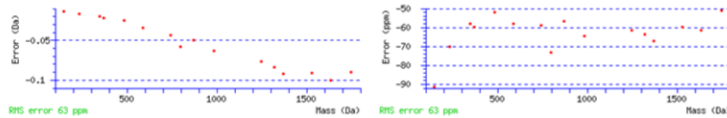
**Monoisotopic mass of neutral peptide Mr(calc): 2214.0683**

**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 <sup>++</sup>	b <sup>0</sup>	b <sup>++</sup>	Seq.	y	y <sup>++</sup>	y <sup>0</sup>	y <sup>++</sup>	y <sup>0</sup>	y <sup>++</sup>	#
1	102.0550	51.5311	84.0444	42.5258	<b>T</b>							19
2	288.1343	144.5708	270.1237	135.5653	<b>W</b>	2016.0511	1008.5292	1999.0245	1000.0159	1998.0405	999.5239	18
3	<b>371.1714</b>	186.0893	353.1608	177.0840	<b>T</b>	1829.9718	915.4895	1812.9452	906.9762	1811.9612	906.4842	17
4	<b>484.2554</b>	242.6314	466.2449	233.6261	<b>L</b>	<b>1746.9346</b>	873.9710	1729.9081	865.4577	1728.9241	864.9627	16
5	<b>587.2046</b>	294.1360	569.2541	285.1307	<b>C</b>	<b>1633.8700</b>	817.4359	1616.8240	808.9156	1615.8400	808.4236	15
6	644.2561	322.6467	626.2755	313.6414	<b>C</b>	<b>1536.8414</b>	768.9243	1513.8148	757.4111	1512.8308	756.9190	14
7	<b>745.3338</b>	373.1705	727.3232	364.1652	<b>T</b>	1473.8199	737.4136	1456.7934	728.9003	1455.8094	728.4083	13
8	842.3865	421.6969	824.3760	412.6916	<b>P</b>	<b>1372.7222</b>	686.8898	1355.7457	678.3765	1354.7617	677.8845	12
9	971.4291	486.2182	953.4186	477.2129	<b>E</b>	1275.7195	638.3634	1258.6929	629.8501	1257.7089	629.3581	11
10	1134.4925	567.7499	1116.4819	558.7446	<b>V</b>	1146.6769	573.8421	1129.6503	565.3288	1128.6663	564.8368	10
11	<b>1247.5765</b>	624.2919	1229.5660	615.2866	<b>L</b>	<b>983.6136</b>	492.3104	966.5870	483.7971	965.6030	483.3051	9
12	<b>1318.6136</b>	659.8105	1300.6031	650.8052	<b>A</b>	<b>876.5295</b>	438.7684	853.5029	427.2551	852.5189	426.7631	8
13	1415.6664	708.3368	1397.6538	699.3316	<b>P</b>	<b>799.4927</b>	400.2498	782.4638	391.7366	781.4818	391.2445	7
14	1548.7090	772.8581	1526.6984	763.8529	<b>E</b>	702.4396	351.2248	685.4131	343.2102	684.4291	343.7382	6
15	1657.7931	829.4002	1639.7823	820.3949	<b>I</b>	573.3970	287.2021	556.3705	278.4889	555.3865	278.1969	5
16	1770.8771	885.9422	1752.8666	876.9369	<b>I</b>	460.3130	230.6601	443.2864	222.1468	442.3024	221.6548	4
17	1883.9612	942.4842	1865.9506	933.4789	<b>S</b>	<b>347.2289</b>	174.1181	330.2023	165.6048	329.2183	165.1128	3
18	1970.9932	986.0002	1952.9827	976.9950	<b>S</b>	<b>234.1448</b>	117.5761	217.1183	109.0628	216.1343	108.5708	2
19					<b>K</b>	<b>147.1128</b>	74.0600	130.0863	65.5468			1

**MASCOT**

: **Modifications**

© 2007-2023 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.

NCBI BLAST search of [TWTLCGTPEYLAPEIILSK](#)  
 (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
80.4	2214.0683	-0.2750	<a href="#">TWTLCGTPEYLAPEIILSK</a>	Phospho T1 69.17%
76.9	2214.0683	-0.2750	<a href="#">TWTLCGTPEYLAPEIILSK</a>	Phospho T3 30.83%
38.7	2214.0683	-0.2750	<a href="#">TWTLCGTPEYLAPEIILSK</a>	Phospho T7 0.00%
18.0	2214.0683	-0.2750	<a href="#">TWTLCGTPEYLAPEIILSK</a>	Phospho Y10 0.00%
12.6	2214.0044	-0.2111	<a href="#">GGSGMLTLGIPSSPGVPAELSK</a>	
12.6	2214.0044	-0.2111	<a href="#">GGSGMLTLGIPSSPGVPAELSK</a>	
12.6	2214.0044	-0.2111	<a href="#">GGSGMLTLGLPSSPGVPAELSK</a>	
12.6	2214.0044	-0.2111	<a href="#">GGSGMLTLGLPSSPGVPAELSK</a>	
11.9	2214.0044	-0.2111	<a href="#">GGSGMLTLGIPSSPGVPAELSK</a>	
11.9	2214.0044	-0.2111	<a href="#">GGSGMLTLGIPSSPGVPAELSK</a>	

Mascot: <http://www.matrixscience.com/>

**MASCOT**

: *Modifications*

© 2007-2023 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.

Matrix Science - Mascot - MS/... Peptide Summary Report (/d... Mascot Search Results: KAPCA... Mascot Search Results: Peptid... Mascot Search Results: Peptid...

www.matrixscience.com/cgi/protein\_view.pl?file=...%2Fdata%2F20210115%2FFT1AlxESR.dat&hit=KAPCA\_BOVIN&db\_idx=1&px=1&ave\_thres

```

...
FT /note=="Phosphoserine"
FT /evidence="ECO:0000250|UniProtKB:P05132"
FT MOD_RES 196
FT /note=="Phosphothreonine" = T1
FT /evidence="ECO:0000250|UniProtKB:P17612"
FT MOD_RES 198
FT /note=="Phosphothreonine; by PDPK1" = T3
FT /evidence="ECO:0000269|PubMed:6262777"
FT MOD_RES 331
FT /note=="Phosphotyrosine"
FT /evidence="ECO:0000250|UniProtKB:P05132" FT STRAND 322..325
FT MOD_RES 339 /evidence="ECO:0000244|PDB:4C34"
FT /note=="Phosphoserine" FT STRAND 326..328
FT /evidence="ECO:0000269|PubMed:6262777" FT /evidence="ECO:0000244|PDB:1STC"
FT LIPID 2 FT TURN 345..350
FT /note=="N-myristoyl glycine" FT /evidence="ECO:0000244|PDB:4Z84"
FT /evidence="ECO:0000269|PubMed:6262777" FT SQ SEQUENCE 351 AA; 40620 MW; 59DDD227D2DEE5D CRC64;
FT MUTAGEN 3 MGNAAAAKKG SEQESVKEFL AKAKEDFLKK WENPAQNTAH LDQFERIKTL GTGSFGRVML
FT /note=="N->D: No myristoylation." VKHMETGNHY AMKILDKQKV VKLQIEHTL NEKRILQAVN FPFLVKLEFS FKDNSNLYMV
FT /evidence="ECO:0000269|PubMed:9521123" FT MEYVPGGEMF SHLRIGRES EPHAREYAAQ IVLTFEYLS LDLIYRDLKP ENLLIDQQGY
FT CONFLICT 202 /note=="T -> N (in Ref. 4; AA sequence)" FT IQVIDFGFAK RVKGR INTLC GTPHYLAPEI ILSH SYNKAV DNMALGVLIY EMAAGYPPFF
FT /evidence="ECO:0000305" FT ADQPIQIVEK IVSGRVRFPS HFSSDLKDLL RNLLQVDLTK RFGNLKGNVN DIKNHKWFAT
FT CONFLICT 204 TDWIAIYQRK VEAPFIPKFK GPGDTSNFDD YEEEEIRVSI NEKCGKEFSE F
FT /note=="E -> Q (in Ref. 4; AA sequence)"
FT /evidence="ECO:0000305"
FT CONFLICT 206 /note=="L -> S (in Ref. 4; AA sequence)"
FT /evidence="ECO:0000305"
FT CONFLICT 287 /note=="N -> D (in Ref. 2; AA sequence and 3; AA sequence)"
FT /evidence="ECO:0000305"
FT TURN 11..13
FT /evidence="ECO:0000244|PDB:2V00"
FT HELIX 18..32

```

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

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.



**MASCOT MS/MS Ions Search**

Your name:  Email:

Search title:

Database(s):

Taxonomy:

Enzyme:  Allow up to:  missed cleavages

Quantitation:

Crosslinking:

Fixed modifications:

Display all modifications

Variable modifications:

Peptide tol.  $\pm$   ppm  $\neq$

MS/MS tol.  $\pm$   Da

Peptide charge:  Monoisotopic  Average

Data file:  No file chosen

Data format:  Precursor:  m/z

Instrument:  Error tolerant   Target PSM FDR

Decoy

**MASCOT** : Modifications

© 2007-2023 Matrix Science



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 <input type="text" value="above"/> <input type="text" value="homology"/>	4279	42	0.98%

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 expanded 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.

## FDR thresholding

- 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 (no common elements)
- The search spaces are also disjoint
  - We never search for unmodified tryptic peptides in the second pass
- Combining two disjoint sets of PSMs, where both have 1% FDR, keeps the total FDR at 1%

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%.

▼ **Sensitivity and FDR (reversed protein sequences)**

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

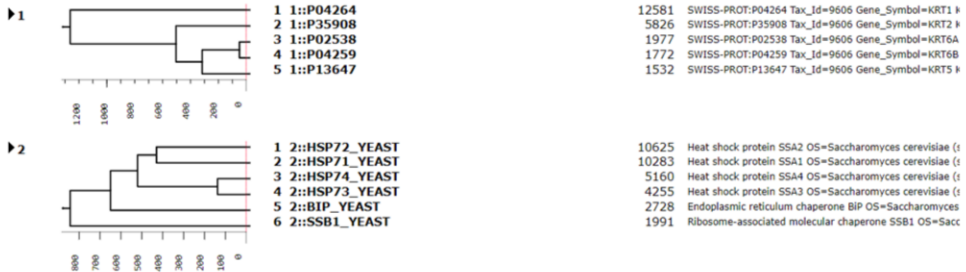
Note: Protein FDR 0% means there are not enough decoy protein hits for a meaningful FDR calculation.

Significance threshold for first pass search is **0.02075**, and second pass search **0.05448**. Target PSM FDR from combined first and second pass searches is **1%**.  
Decoy results are available in [the decoy report](#).

Proteins (59) [Report Builder](#) [Unassigned \(22268\)](#)

**Protein families 1-10 (out of 44)**

10 per page 1 2 3 4 5 [Next](#) [Expand all](#) [Collapse all](#)  
 Accession contains Find Clear



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

▼954 peptide matches (173 non-duplicate, 781 duplicate)  
 Auto-fit to window

Query Dupes	Observed	Mr(expt)	Mr(calcd)	ppm	M Score	Expect	Rank	U Peptide
#12574 ▶2	804.4050	1606.7955	1606.8025	-4.31	81	3e-06	▶1	U N.FNGNTLNDIIMLIK.L ←
#12741 ▶1	812.3828	1622.7511	1622.7536	-1.51	39	0.04	▶1	U R.LGEHNIIDVLEGNQ.F + Carbanidomethyl (N-term)
#13143 ▶4	827.3561	1652.6976	1652.6923	3.23	84	8.1e-07	▶1	U R.SCAAAGTECLISGWNT.T
#13307 ▶1	830.9304	1659.8463	1659.8488	-0.25	68	6e-05	▶1	U N.IDVLEGNQFVNAAK.I ←
#13830 ▶1	855.8650	1709.7153	1709.7137	0.94	65	5.8e-05	▶1	U R.SCAAAGTECLISGWNT.T + Carbanidomethyl (N-term)
#13877 ▶5	857.4082	1712.8018	1712.8006	0.70	58	0.0006	▶1	U R.LGEHNIIDVLEGNQF.I
#14490 ▶4	883.8943	1765.7741	1765.7764	-1.29	48	0.005	▶1	U R.SCAAAGTECLISGWNTK.S + 2 [-1.0078 at C2,C9]
#14608 ▶8	887.9519	1773.8892	1773.8897	-0.25	113	2.2e-09	▶1	U H.NIDVLEGNQFVNAAK.I ←
#14772	896.4172	1790.8199	1790.8258	-3.25	57	0.00082	▶1	U R.SCAAAGTECLISGWNTK.S + [-33.9877 at C9]
#14785 ▶2	897.4366	1792.8586	1792.8566	1.11	88	6e-09	▶1	U K.VCNVYNIQQTIAAN.-
#15193 ▶5	912.4043	1822.7940	1822.7978	-2.10	57	0.00066	▶1	U R.SCAAAGTECLISGWNTK.S + Carbanidomethyl (N-term); 2 [-1.0078 at C2,C9]
#15293 ▶3	916.4605	1830.9065	1830.9111	-2.56	93	2.6e-07	▶1	U H.NIDVLEGNQFVNAAK.I + Carbanidomethyl (N-term)
#15889 ▶9	941.9230	1881.8313	1881.8349	-1.90	114	8.5e-12	▶1	U R.SCAAAGTECLISGWNTK.S
#15890	628.2845	1881.8317	1881.8349	-1.72	48	3.5e-05	▶1	U R.SCAAAGTECLISGWNTK.S
#15914	628.6180	1882.8323	1882.8189	7.09	44	0.014	▶1	U R.SCAAAGTECLISGWNTK.S + [+0.9840 at N16]
#16103 ▶3	948.9313	1895.8480	1895.8506	-1.38	112	2.7e-09	▶1	U R.SCAAAGTECLISGWNTK.S + [+14.0156 at C-term K]
#16220	955.9276	1909.8407	1909.8298	5.70	79	4.4e-06	▶1	U R.SCAAAGTECLISGWNTK.S + [+27.9949 at T17]
#16229	637.6266	1909.8581	1909.8462	-4.27	54	0.0018	▶1	U R.SCAAAGTECLISGWNTK.S + [+28.0313 at N16]
#16242 ▶4	637.6288	1909.8645	1909.8462	-0.90	60	0.00041	▶1	U R.SCAAAGTECLISGWNTK.S + [+28.0313 at C-term K]
#16244 ▶9	955.9400	1909.8654	1909.8462	-0.45	117	8.5e-10	▶1	U R.SCAAAGTECLISGWNTK.S + [+28.0313 at C-term K]
#16287	956.4820	1910.9494	1910.9486	0.43	47	0.01	▶1	U E.HNIDVLEGNQFVNAAK.I
#16330 ▶1	957.9163	1913.8181	1913.8070	5.79	77	5.8e-06	▶1	U R.SCAAAGTECLISGWNTK.S + [+31.9721 at W4]
#16432 ▶1	962.9273	1923.8400	1923.8567	-8.68	85	1.2e-06	▶1	U R.SCAAAGTECLISGWNTK.S + Carbanidomethyl (N-term); [+42.0218 at C2]
#16434	642.2883	1923.8430	1923.8567	-7.13	50	0.0035	▶1	U R.SCAAAGTECLISGWNTK.S + [+42.0218 at C-term K]
#16469 ▶1	943.9348	1925.8551	1925.8611	-3.12	45	0.00011	▶1	U R.SCAAAGTECLISGWNTK.S + [+44.0262 at G15]
#16485	964.9608	1927.9669	1927.9998	-1.48	45	0.015	▶1	U K.IITDFWPNFVNTLSNDIK.L
#16568	965.9386	1937.8627	1937.8611	0.80	47	4.0085	▶1	U R.SCAAAGTECLISGWNTK.S + [+56.0262 at S12]
#16891 ▶4	970.4324	1938.8502	1938.8564	-3.21	85	4.7e-09	▶1	U R.SCAAAGTECLISGWNTK.S + Carbanidomethyl (N-term)
#16892	647.2907	1938.8503	1938.8564	-3.15	28	0.002	▶1	U R.SCAAAGTECLISGWNTK.S + Carbanidomethyl (N-term)

Possible assignments:  
Methyl (C-term) [+14.0156]  
Methyl (K) [+14.0156]

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.

▼954 peptide matches (173 non-duplicate, 781 duplicate)

Auto-fit to window

Query Dupes	Observed	Mr(expt)	Mr(calcd)	ppm	M	Score	Expect	Rank	U	Peptide
#12574 ▶2	804.4050	1606.7955	1606.8025	-4.31	0	81	3e-06	▶1	U	N.FNGLTNDLMLK.L
#12741 ▶1	812.3828	1622.7511	1622.7536	-1.51	0	39	0.04	▶1	U	R.LGHNIDVLEGNQ.F + Carbanidomethyl (N-term)
#13143 ▶4	827.3561	1652.6976	1652.6923	3.23	0	84	8.1e-07	▶1	U	R.SCAAAGTECLISGWGN.T
#13307 ▶1	830.9304	1659.8463	1659.8468	-0.25	0	68	6e-05	▶1	U	N.IDVLEGNQFINAAK.I
#13830 ▶1	855.8650	1709.7153	1709.7137	0.94	0	65	5.8e-05	▶1	U	R.SCAAAGTECLISGWGN.T + Carbanidomethyl (N-term)
#13877 ▶5	857.4082	1712.8018	1712.8006	0.70	0	58	0.0006	▶1	U	R.LGHNIDVLEGNQF.I
#14490 ▶4	883.8943	1765.7741	1765.7764	-1.29	0	48	0.005	▶1	U	R.SCAAAGTECLISGWNTK.S + 2 [-1.0078 at C2,C9]
#14608 ▶8	887.9519	1773.8892	1773.8897	-0.25	0	113	2.2e-09	▶1	U	H.NIDVLEGNQFINAAK.I
#14772	896.4172	1790.8199	1790.8258	-3.25	0	57	0.00082	▶1	U	R.SCAAAGTECLISGWNTK.S + [-33.9877 at C9]
#14785 ▶2	897.4366	1792.8586	1792.8566	1.11	0	88	6e-09	▶1	U	K.VCYVNWVQQTIAN.-
#15193 ▶5	912.4043	1822.7940	1822.7978	-2.10	0	57	0.00066	▶1	U	R.SCAAAGTECLISGWNTK.S + Carbanidomethyl (N-term); 2 [-1.0078 at C2,C9]
#15293 ▶3	916.4605	1830.9065	1830.9111	-2.56	0	93	2.6e-07	▶1	U	H.NIDVLEGNQFINAAK.I + Carbanidomethyl (N-term)
#15889 ▶9	941.9230	1881.8313	1881.8349	-1.90	0	114	8.5e-12	▶1	U	R.SCAAAGTECLISGWNTK.S
#15890	628.2845	1881.8317	1881.8349	-1.72	0	48	3.5e-05	▶1	U	R.SCAAAGTECLISGWNTK.S
#15914	628.6180	1882.8323	1882.8189	7.09	0	44	0.014	▶1	U	R.SCAAAGTECLISGWNTK.S + [+0.9840 at N16]
#16103 ▶3	948.9313	1895.8480	1895.8506	-1.38	0	112	2.7e-09	▶1	U	R.SCAAAGTECLISGWNTK.S + [+14.0156 at C-term K]
#16220	955.9276	1909.8407	1909.8298	5.70	0	79	4.4e-06	▶1	U	R.SCAAAGTECLISGWNTK.S + [+27.9949 at T17]
#16229	637.6266	1909.8581	1909.8662	-4.27	0	54	0.0018	▶1	U	R.SCAAAGTECLISGWNTK.S + [+28.0313 at N16]
#16242 ▶4	637.6288	1909.8645	1909.8662	-0.90	0	60	0.00041	▶1	U	R.SCAAAGTECLISGWNTK.S + [+28.0313 at C-term K]
#16244 ▶9	955.9400	1909.8654	1909.8662	-0.45	0	117	8.5e-10	▶1	U	R.SCAAAGTECLISGWNTK.S + [+28.0313 at C-term K]
#16287	956.4820	1910.9494	1910.9486	0.43	0	47	0.01	▶1	U	E.HNIDVLEGNQFINAAK.I
#16330 ▶1	957.9163	1913.8181	1913.8070	5.79	0	77	5.8e-06	▶1	U	R.SCAAAGTECLISGWNTK.S + [+31.9721 at N14]
#16432 ▶1	962.9273	1923.8400	1923.8567	-8.68	0	85	1.2e-06	▶1	U	R.SCAAAGTECLISGWNTK.S + Carbanidomethyl (N-term); [+42.0218 at C2] ←
#16434	642.2883	1923.8430	1923.8567	-7.13	0	50	0.0035	▶1	U	R.SCAAAGTECLISGWNTK.S + [+42.0218 at C-term K]
#16469 ▶1	963.9348	1925.8551	1925.8611	-3.12	0	65	0.00011	▶1	U	R.SCAAAGTECLISGWNTK.S + [+44.0262 at G15]
#16485	964.9608	1927.9069	1927.9098	-1.48	0	45	0.015	▶1	U	K.IITHFNGNTLNDIM.L
#16568	969.9386	1937.8627	1937.8611	0.80	0	47	0.0085	▶1	U	R.SCAAAGTECLISGWNTK.S + [+56.0262 at S12]
#16581 ▶4	970.4324	1938.8502	1938.8564	-3.21	0	85	4.7e-09	▶1	U	R.SCAAAGTECLISGWNTK.S + Carbanidomethyl (N-term) ←
#16582	647.2907	1938.8503	1938.8564	-3.15	0	28	0.002	▶1	U	R.SCAAAGTECLISGWNTK.S + Carbanidomethyl (N-term)

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.



ET larger test search training co... x

eductus/mascot/cgi/master\_results\_2.cgi?file=%2Fdata%2F20230317%2F001361.dat\_sigthreshhold=0.001

Error tolerant mod delta: -17.026532 Find Previous Next Clear

Matches found in family 5.

- 1.3 #1::P02538 1488 60293 72 (72) 24 (24) SWISS-PROT:P02538 Tax\_ID=9606 Gene\_Symbol=KRT6A Keratin, type II cytoskeletal 6A
- 1.4 #1::P04259 1328 60247 68 (68) 24 (24) SWISS-PROT:P04259 Tax\_ID=9606 Gene\_Symbol=KRT6B Keratin, type II cytoskeletal 6B
- 1.5 #1::P13647 1126 62568 55 (55) 18 (18) SWISS-PROT:P13647 Tax\_ID=9606 Gene\_Symbol=KRT5 Keratin, type II cytoskeletal 5

Redisplay All None

619 peptide matches (274 non-duplicate, 345 duplicate)

Auto-fit to window

Query Dupes	Observed	Mr (expt)	Mr (calc)	ppm	M	Score	Expect	Rank	0	1	2	3	4	5	Peptide
#3123	540.2709	1078.5272	1078.5295	-2.13	0	40	0.009	1	U	■	■	■	■	■	K.AQYEEIAQR.S + [-28.0061 at C-term R]
#3202	542.7500	1083.4855	1083.4906	-4.78	0	54	4.4e-06	1	U	■	■	■	■	■	K.DVDNAYMIR.V + Oxidation (0)
#3309	545.9705	1089.5264	1089.5302	-3.51	0	53	0.00043	1	U	■	■	■	■	■	R.TLLKQRESR.N + [+57.0215 at N-term T]
#3330	546.7531	1091.4916	1091.4956	-3.65	0	45	3.2e-05	1	U	■	■	■	■	■	R.GSOGSSGSSIQGR.G
#3347	547.2854	1092.5162	1092.5199	-3.42	0	38	0.00028	1	U	■	■	■	■	■	K.AQYEEIARR.S
#3369	547.9759	1093.5373	1093.5404	-2.80	1	65	4.4e-05	1	U	■	■	■	■	■	K.RYLDLTAAR.T + [-99
#3537	554.2736	1106.5326	1106.5356	-2.67	0	70	3.1e-07	1	U	■	■	■	■	■	K.AQYEEIAQR.S
#3587	556.2882	1110.5618	1110.5669	-4.60	0	69	2.3e-07	1	U	■	■	■	■	■	R.IISITSQGSFR.N
#3677	559.7947	1117.5748	1117.5768	-1.74	0	39	0.014	1	U	■	■	■	■	■	R.FLQQMQL.G
#3789	563.2739	1124.5332	1124.5349	-1.49	0	82	1.7e-08	1	U	■	■	■	■	■	K.AEASLYQSK.Y
#3885	566.2566	1130.4987	1130.5026	-3.48	0	56	2.4e-06	1	U	■	■	■	■	■	R.STSSPFLSRR.H
#3909	567.2815	1132.5485	1132.5546	-5.42	1	32	0.0028	1	U	■	■	■	■	■	R.KLLEQRECR
#3977	570.2805	1138.5464	1138.5506	-3.63	0	40	0.011	1	U	■	■	■	■	■	K.AEASLYQSK.Y + [+14.0156 at E2]
#3978	570.2938	1138.5730	1138.5771	-3.62	0	49	0.0014	1	U	■	■	■	■	■	Y.FEFLNLR.R
#4002	571.2608	1140.5071	1140.5121	-4.36	0	57	2.6e-06	1	U	■	■	■	■	■	R.DYQLMPTK.L
#4041	572.3143	1142.6141	1142.6183	-3.63	0	72	2e-07	1	U	■	■	■	■	■	K.LALEALQK.A
#4143	576.7781	1151.5416	1151.5458	-3.62	1	49	3.7e-05	1	U	■	■	■	■	■	NYEDELNR.K
#4146	584.8555	1151.5446	1151.5458	-1.10	1	28	0.0044	1	U	■	■	■	■	■	NYEDELNR.K
#4168	578.2687	1154.5239	1154.5278	-4.18	0	51	1.4e-05	1	U	■	■	■	■	■	R.DYQLMPTK.L + Oxidation (0)
#4194	579.2590	1156.5035	1156.5070	-3.05	0	63	6.9e-07	1	U	■	■	■	■	■	R.DYQLMPTK.L + Oxidation (0)
#4232	579.9849	1157.5053	1157.5064	-0.95	0	42	0.0064	1	U	■	■	■	■	■	K.LMDLDAQQ.A
#4314	582.7835	1163.5524	1163.5571	-3.99	0	50	0.001	1	U	■	■	■	■	■	K.AQYEEIAQR.S + [+57.0215 at N-term]
#4408	583.2958	1164.5771	1164.5775	-0.36	0	69	4.2e-07	1	U	■	■	■	■	■	K.YEELQVTAQR.H
#4437	583.7863	1165.5581	1165.5615	-2.89	1	41	0.011	2	U	■	■	■	■	■	NYEDELNR.R + [+14.0156 at E6]
#4455	584.2781	1166.5417	1166.5455	-3.22	0	47	0.0022	1	U	■	■	■	■	■	K.AEASLYQSK.Y + [+42.0106 at N-term]
#4488	584.7994	1167.5043	1167.5084	-3.53	0	47	0.0027	1	U	■	■	■	■	■	R.IISITSQGSFR.N + [+57.0215 at N-term]
#4505	585.2770	1168.5394	1168.5434	-3.41	0	33	0.0013	1	U	■	■	■	■	■	K.YEQLMPTK.L + Oxidation (0)

Possible assignments:  
 Carbamidomethyl (N-term) [+57.0215]  
 Carbamidomethyl (T) [+57.0215]  
 Gly (T) [+57.0215]

MASCOT : Modifications © 2007-2023 Matrix Science MATRIX SCIENCE

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.

Matches found in family 5.

5.1 [#2::POCS90](#) Score 2737 Mass 70585 Matches 116 (116) Sequences 34 (34) Import motor subunit, mitochondrial OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) OX=559292 GN=SSC1 PE=1 SV=1

▼116 peptide matches (70 non-duplicate, 46 duplicate)

Auto-fit to window

Query Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M Score	Expect	Rank	U Peptide
#8025	708.3600	1414.7054	1414.6940	8.09	0 42	0.011	▶1	U R.VQGGEEVNAEELK.T + [+14.0156 at E10]
#8057 ▶1	709.8384	1417.6622	1417.6685	-4.42	0 93	8.4e-10	▶1	U K.ADQLANDTENSLEK.E
#8523	729.8553	1457.6960	1457.6998	-2.61	0 44	0.006	▶1	U R.VQGGEEVNAEELK.T + [+57.0215 at N-term]
#8664	738.3498	1474.6851	1474.6899	-3.29	0 44	0.0054	▶1	U K.ADQLANDTENSLEK.E + [+57.0215 at N-term]
#9026 ▶3	754.9447	1507.8748	1507.8762	-0.95	0 84	3.6e-09	▶1	U K.LIGNFTLAGIPPAPK.G
#9046	755.4444	1508.8742	1508.8733	0.64	0 55	0.00019	▶1	U K.LIGNFTLAGIPPAPK.G + [+0.9970 at N-term L]
#9085	757.8959	1513.7772	1513.7777	-0.33	0 72	1.1e-05	▶1	U R.QAVVNPENTLEATK.R + [-17.026532 at N-term]
#9243 ▶1	766.4087	1530.8029	1530.8042	-0.87	0 87	6.2e-09	▶1	U R.QAVVNPENTLEATK.R
#9674	783.4563	1564.8980	1564.8977	0.20	0 61	7e-05	▶1	U K.LIGNFTLAGIPPAPK.G + [+57.0215
#9860	527.6145	1579.8217	1579.8279	-3.95	2 35	0.0012	▶1	U K.MRETAEAYLQKPKV.N + Oxidation (M)
#9862	790.9210	1579.8274	1579.8279	-0.31	2 61	2.6e-06	▶1	U K.MRETAEAYLQKPKV.N + Oxidation (M)
#9929	794.9176	1587.8207	1587.8257	-3.14	0 55	0.00062	▶1	U R.QAVVNPENTLEATK.R + [+57.0215 at N-term]
#10566 ▶1	549.2972	1644.8697	1644.8723	-1.54	0 72	2e-07	▶1	U R.VVNEPTAAALAYGLEK.S
#10567 ▶3	823.4424	1644.8702	1644.8723	-1.26	0 102	1.8e-10	▶1	U R.VVNEPTAAALAYGLEK.S
#10908	833.9230	1665.8314	1665.8244	4.22	0 53	0.00093	▶1	U L.STSDISEVLLVOGMSR.M + Oxidation (M)

Possible assignments:  
Gln->pyro-Glu (N-term Q) [-17.0265]

MASCOT

: Modifications

© 2007-2023 Matrix Science



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

ET target test search training

edcutus/mascot/cgi/master\_results\_2.pl?file=\_%2Fdata%2F20230317%2F%2F001361.dat\_sighreshold=0.001

Error tolerant mod delta > -48.00371 Find Previous Next Clear

Matches found in family 2.

Auto-fit to window

Query	Dipep	Observed	Mr (expt)	Mr (calc)	ppm	M Score	Expect	Rank	U	1	2	3	4	Peptide
#6374		655.8430	1309.6939	1309.6876	-2.53	48	0.002	1	U					K.ELQDIARFIMSK.L + [48.0037] at M0
#6374		657.3133	1312.6120	1312.6122	-0.15	46	0.0031	1	U					R.FEELGADLPR.S + [+14.0156 at E] Possible assignments: Dehomethyl (M) [-48.0034]
#6434		657.3726	1312.7310	1312.7305	-0.39	63	6e-05	1	U					R.STLPPVPEVLR.D + [+57.0215 at I] Dehomethyl (M) [-48.0034]
#6470		661.8514	1321.6883	1321.6877	0.42	65	4.7e-05	1	U					K.HQLRSYASYLE.H + [+57.0215 at R-term]
#6462		666.3566	1330.6387	1330.7034	-3.32	44	0.0061	1	U					R.LNNRFQPEK.R + [+57.0215 at R]
#6461		666.3572	1330.6399	1330.7034	-2.40	46	0.0038	1	U					R.LNNRFQPEK.R + [+57.0215 at R-term]
#6977		669.3299	1336.6450	1336.6510	-4.38	40	0.0003	1	U					K.ETARNLOTREV.D
#7104	2	672.3235	1342.6325	1342.6438	-8.45	69	2.6e-07	1	U					K.HERTARYLQAE.V + Oxidation (O)
#7108	2	673.3500	1344.6804	1344.6885	-2.30	87	6.1e-09	1	U					K.IYVDIVLVGGTR.L
#7140		675.8160	1349.6174	1349.6211	-0.77	44	4.8e-05	1	U					K.FGADRGDQGR.V
#7217	1	678.8463	1355.6780	1355.6755	1.90	67	2.0e-05	1	U					K.ELQDIARFIMSK.L + Oxidation (O) [-18.0106 at R-term]
#7218	1	678.8467	1355.6789	1355.6755	2.90	61	0.00011	1	U					K.ELQDIARFIMSK.L + Oxidation (O) [-18.0106 at R-term]
#7240		679.8520	1357.6895	1357.6911	-1.17	76	5.0e-08	1	U					K.ELQDIARFIMSK.L
#7236		680.3242	1358.6339	1358.6387	-3.59	70	1.2e-05	1	U					K.HERTARYLQAE.V + Oxidation (O) [+15.9949 at R]
#7216		681.8006	1361.7467	1361.7442	1.86	65	3.7e-05	1	U					S.LLSIISDIEIPEV.A
#7106	3	687.8494	1373.6843	1373.6860	-1.26	73	2.1e-07	1	U					K.ELQDIARFIMSK.L + Oxidation (O)
#7107	3	687.8500	1373.6855	1373.6860	-0.40	91	3.3e-09	1	U					K.ELQDIARFIMSK.L + Oxidation (O)
#7121		688.3390	1374.6435	1374.6627	0.45	42	0.0095	1	U					K.WTIEAGDWLRQ.D
#7142		689.8250	1377.6364	1377.6412	-3.46	52	6.7e-06	1	U					R.WNDFPQDMSK.H
#7171		691.3248	1380.6591	1380.6590	-0.02	69	2.1e-05	1	U					K.HYFDFQISNHL.G + Oxidation (O)
#7429		693.3311	1384.6476	1384.6544	-4.93	40	0.012	1	U					K.HERTARYLQAE.V + Oxidation (O) [+42.0106 at R]
#7468		694.8564	1387.6992	1387.7017	-2.94	62	0.00011	1	U					K.ELQDIARFIMSK.L + Oxidation (O) [+14.0156 at R]
#7478		695.3634	1388.7122	1388.7147	-1.80	65	4.6e-07	1	U					A.AVQAALLTODESK.T
#7482		695.8456	1389.6766	1389.6809	-3.10	64	0.0041	1	U					K.ELQDIARFIMSK.L + [31.9898 at R]
#7116		696.8339	1391.6533	1391.6568	-2.56	54	0.00051	1	U					K.SKSTETNLDN.T
#7124		697.3026	1392.5905	1392.5980	-5.33	43	5.5e-05	1	U					R.WNDFPQDMSK.H
#7174		698.8388	1395.6430	1395.6480	-3.58	40	0.016	1	U					K.ELQDIARFIMSK.L + Oxidation (O) [+21.9819 at D]
#7139		700.8366	1399.6586	1399.6653	-4.77	48	0.0019	1	U					K.HERTARYLQAE.V + Oxidation (O) [+57.0215 at R]
#7165		701.3536	1400.6627	1400.6669	-3.04	55	0.00049	1	U					K.ELQDIARFIMSK.L + [43.0058 at R]
#7194		749.2041	1404.5906	1404.5906	-0.032	34	0.023	1	U					A.GYVLEGGDFDRL.S
#7113		704.3126	1406.6106	1406.6136	-2.17	61	7.3e-07	1	U					R.WNDFPQDMSK.H
#7196	4	705.3007	1408.5869	1408.5929	-4.27	74	4.4e-08	1	U					R.WNDFPQDMSK.H + Oxidation (O)
#7173		707.3309	1412.6473	1412.6493	-1.42	37	0.026	1	U					K.HERTARYLQAE.V + Oxidation (O) [+70.0055 at R]
#6140		708.8558	1415.6971	1415.7078	-7.40	40	0.016	1	U					K.ELQDIARFIMSK.L + Oxidation (O) [+42.0218 at R-term]
#6105	1	712.3116	1422.6086	1422.6085	0.065	70	1e-07	1	U					R.WNDFPQDMSK.H + Oxidation (O)
#6179	2	712.8028	1423.5911	1423.5925	-1.04	74	2.3e-06	1	U					R.WNDFPQDMSK.H + Oxidation (O) [+40.9840 at R-term R]
#6179	2	713.8643	1425.7141	1425.7140	0.097	88	4.5e-09	1	U					K.VQQLLSEYDSE.K
#6209		719.3176	1436.6206	1436.6242	-2.51	79	1.1e-06	1	U					R.WNDFPQDMSK.H + Oxidation (O) [+14.0156 at R]

1 subset or intersection (1 subset protein in total)

MASCOT : Modifications © 2007-2023 Matrix Science

MATRIX SCIENCE

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.

The screenshot displays the Mascot search results for a query. The top match is highlighted in yellow, showing a score of 34 and a peptide sequence R.VQIISLQDQRR.I. The interface includes a search bar, a list of matches with columns for query, observed, expected, and score, and a detailed view of the top match.

Query	Observed	Mr (expt)	Mr (calc)	ppm	M Score	Expect	Rank	U	1	2	3	4	Peptide
Q5127	600.3410	1190.6677	1190.6670	0.64	109	4.3e-11	1	U					R.VQIISLQDQRR.T
Q5102	602.3382	1202.6619	1202.6659	-3.37	64	1.5e-06	1	U					R.VQIISLQDQRR.T
Q5122	605.3293	1208.6440	1208.6441	-0.096	76	6.7e-08	1	U					R.VQIISLQDQRR.T
Q5130	607.8081	1213.6016	1213.6051	-3.90	61	1.9e-06	1	U					R.VQIISLQDQRR.T
Q5100	613.3410	1224.6479	1224.6714	-2.85	47	0.002	1	U					R.VQIISLQDQRR.T
Q5101	613.3482	1224.6817	1224.6826	-0.71	74	3.6e-06	1	U					R.VQIISLQDQRR.T
Q5130	614.8150	1227.6155	1227.6207	-4.25	60	2.5e-06	1	U					R.VQIISLQDQRR.T
Q5141	615.3090	1229.6035	1229.6048	-1.03	38	0.014	1	U					R.VQIISLQDQRR.T
Q5178	617.3143	1230.6141	1230.6183	-3.41	38	0.023	1	U					R.VQIISLQDQRR.T
Q5189	620.8370	1239.6651	1239.6611	-0.82	55	0.0032	1	U					R.VQIISLQDQRR.T
Q5103	621.3471	1240.6797	1240.6886	-7.33	65	2.8e-05	1	U					R.VQIISLQDQRR.T

Score > 34 indicates identity

Score	> 34	indicates	identity
1.72	0	20	0.13
-7.35	1	27	0.15
-7.35	1	17	1.8
1.89	0	17	1.8
-4.22	1	10	8.1
7.49	0	10	8.4
1.72	0	10	8.7
0.27	1	10	8.4

Possible assignments:

- Acetyl (N-term) [+42.0106]
- Acetyl (Protein N-term) [+42.0106]
- Acetylation (M) [+34.0114 at C-term R]

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 assignments 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 known 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.

Automatic error tolerant search: x

archive-win10/mascot\_2\_7\_00/cg/master\_results\_2.pl?file=\_%2Fdata%2FF981130.dat\_ignoreion: Find Clear

1 subset or intersection (1 subset protein in total)

7 RPN2\_HUMAN 113 Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 2 OS=Homo sapie...

Score	Mass	Matches	Sequences
7.1	69355	3 (2)	3 (2)

2 same sets of RPN2\_HUMAN

3 peptide matches (3 non-duplicate, 0 duplicate)

Auto-fit to window

Query Dupes	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Expect	Rank	U	Peptide
50	539.6746	1077.3347	1077.5818	-0.2471	0	60	9.2e-05	1	R.YIANTVELR.V
139	766.7314	1531.4482	1531.7518	-0.3036	0	86	3.6e-07	1	K.TQGEVVFVAEPDNK.N
246	766.7314	2297.1723	2297.2318	-0.0594	1	68		1	R.LIHKTKQGEVVFVAEPDNK.N * [+145.1405 at R5]

1 subset or intersection (3 subset proteins in total)

8 BASI\_BOVIN 104 Basigin (Fragment) OS=Bos taurus OX=9913 GN=BSG PE=2 SV=1

9 KPVM\_HUMAN 92 Pyruvate kinase PKM OS=Homo sapiens OX=9606 GN=PKM PE=1 SV=4

10 CBPM\_HUMAN 83 Carboxypeptidase M OS=Homo sapiens OX=9606 GN=CPM PE=1 SV=2

10 per page 1 2 3 Next Expand all Collapse all

Mascot: <http://www.matrixscience.com/>

© 2007-2023 Matrix Science

MATRIX SCIENCE

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

MASCOT

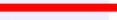
: *Modifications*

© 2007-2023 Matrix Science

 MATRIX  
SCIENCE

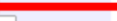
To summarise, it's best to exclude isotopic labels and very large modifications from error tolerant searching. You should specify very abundant modifications as variable modifications, so you don't miss doubly modified peptides.

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.

Format	Significance threshold p<	0.02075	Max. number of families	AUTO	<a href="#">[help]</a>
	Target FDR (overrides sig. threshold)	1%	FDR type	PSM	
	Display non-sig. matches	<input type="checkbox"/>	Min. number of sig. unique sequences	1	
	Error tolerant matches:	Reliable	Dendrograms cut at	0	
	Preferred taxonomy	All entries			

▼ **Sensitivity and FDR (reversed protein sequences)**

	Target	Decoy	FDR
Protein family members	88	20	22.73%
PSMs	4279	42	0.98%

Format	Significance threshold p<	0.02075	Max. number of families	AUTO	<a href="#">[help]</a>
	Target FDR (overrides sig. threshold)	1%	FDR type	PSM	
	Display non-sig. matches	<input type="checkbox"/>	Min. number of sig. unique sequences	2	
	Error tolerant matches:	Reliable	Dendrograms cut at	0	
	Preferred taxonomy	All entries			

▼ **Sensitivity and FDR (reversed protein sequences)**

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

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