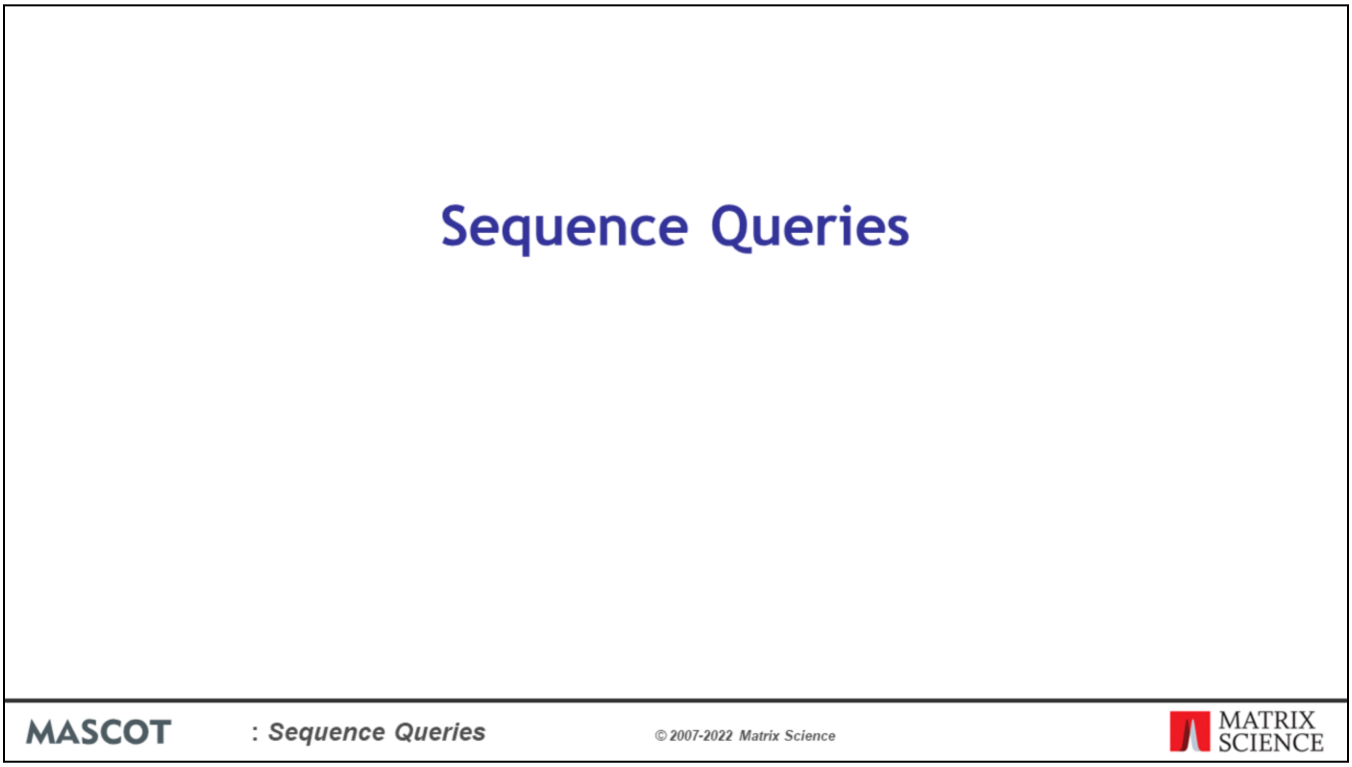
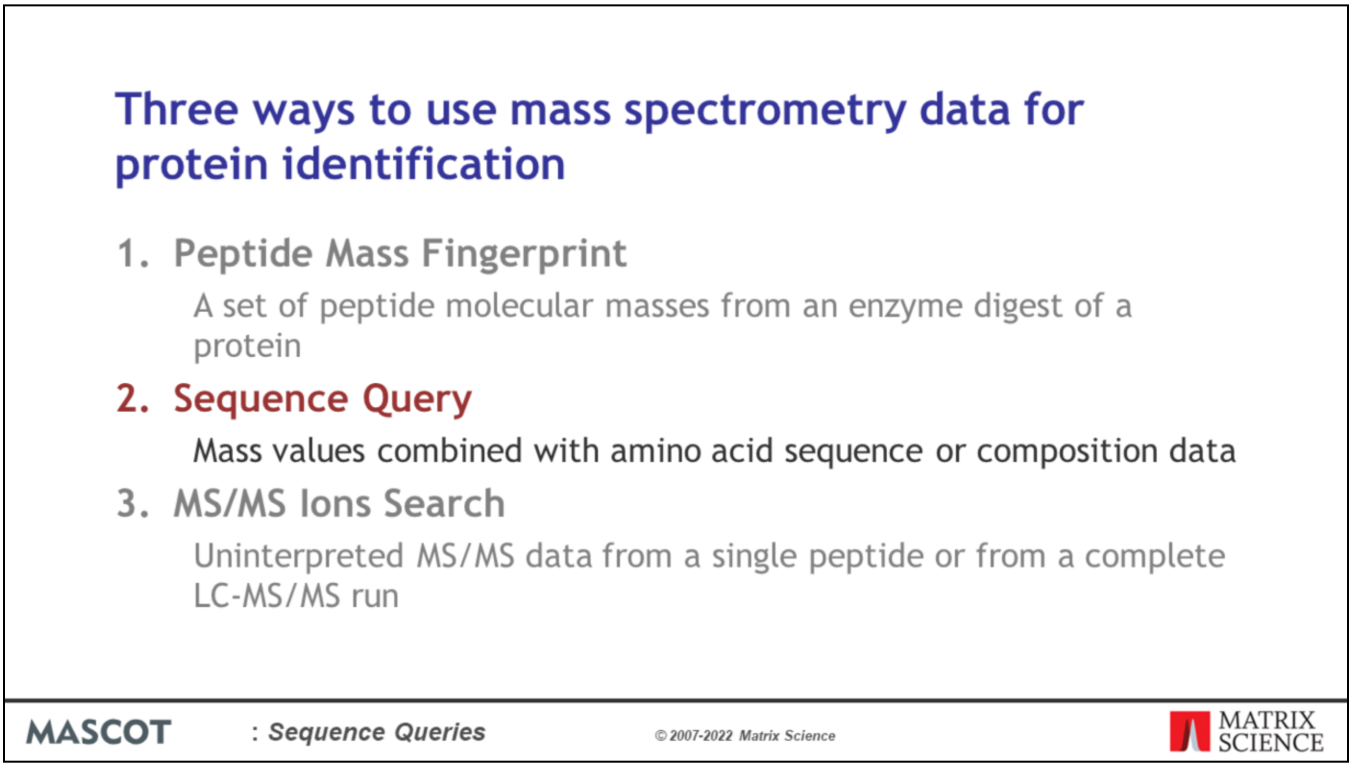


Sequence Queries



Three ways to use mass spectrometry data for protein identification

1. Peptide Mass Fingerprint

   A set of peptide molecular masses from an enzyme digest of a protein

2. Sequence Query

   Mass values combined with amino acid sequence or composition data

3. MS/MS Ions Search

   Uninterpreted MS/MS data from a single peptide or from a complete LC-MS/MS run

You will remember from the introduction, that sequence queries are searches where mass information is combined with amino acid sequence or composition information

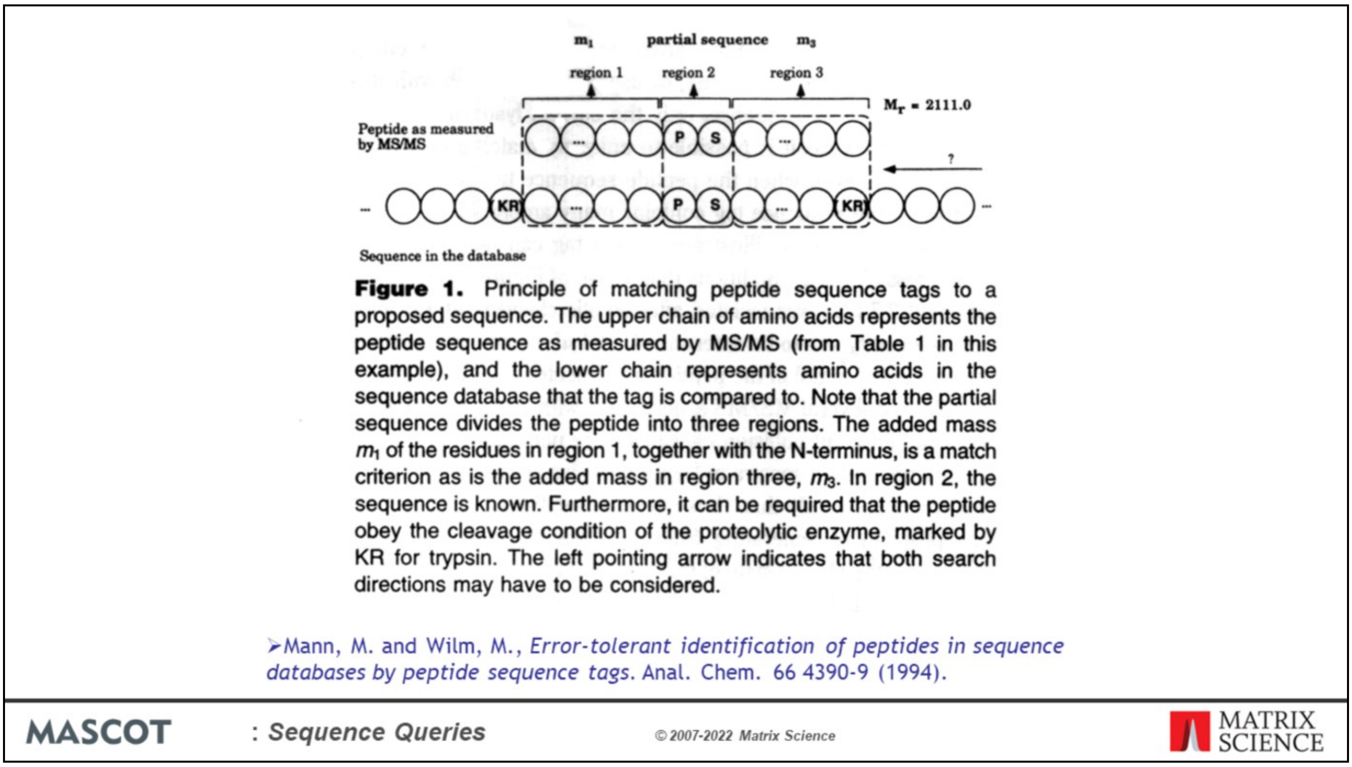


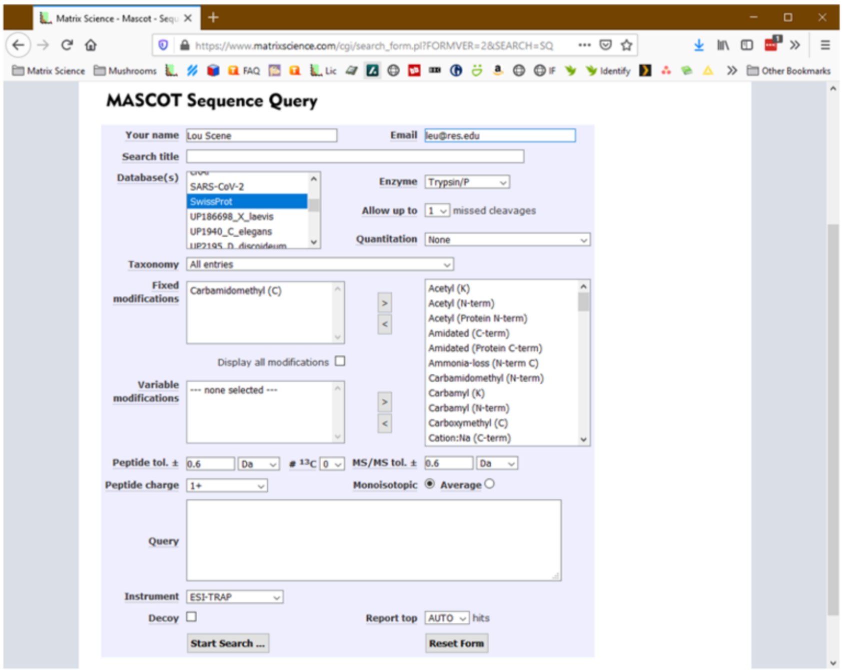
Figure 1. Principle of matching peptide sequence tags to a proposed sequence. The upper chain of amino acids represents the peptide sequence as measured by MS/MS (from Table 1 in this example), and the lower chain represents amino acids in the sequence database that the tag is compared to. Note that the partial sequence divides the peptide into three regions. The added mass m_1 of the residues in region 1, together with the N-terminus, is a match criterion as is the added mass in region three, m_3. In region 2, the sequence is known. Furthermore, it can be required that the peptide obey the cleavage condition of the proteolytic enzyme, marked by KR for trypsin. The left pointing arrow indicates that both search directions may have to be considered.

- Mann, M. and Wilm, M., *Error-tolerant identification of peptides in sequence databases by peptide sequence tags*. Anal. Chem. 66 4390-9 (1994).

The best known example is a sequence tag search, where a few residues of amino acid sequence are interpreted from the MS/MS spectrum.

Search parameters all still apply

- Enzyme
- Modifications
- Charge
- Instrument



Matrix Science - Mascot - Sequ
https://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=SQ
Matrix Science
Mushrooms
FAQ
Lic
IF
Identify
Other Bookmarks

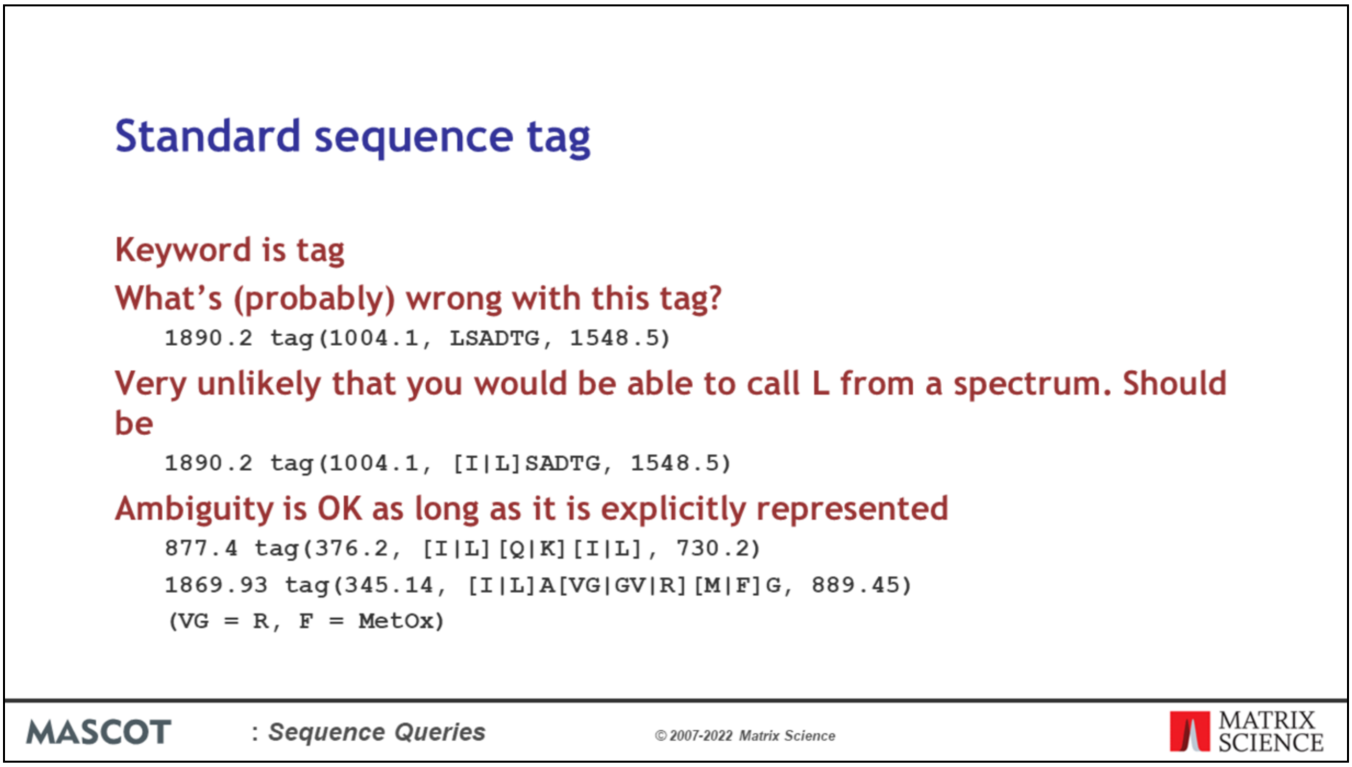
MASCOT Sequence Query

Your name	Lou Scene	Email	leu@res.edu
Search title			
Database(s)	SARS-CoV-2 SwissProt UP186698_X_laevis UP1940_C_elegans UP2195_D_discoideum	Enzyme	Trypsin/P
		Allow up to	1 missed cleavages
		Quantitation	None
Taxonomy	All entries		
Fixed modifications	Carbamidomethyl (C)		Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-term) Ammonia-loss (N-term C) Carbamidomethyl (N-term) Carbamyl (K) Carbamyl (N-term) Carboxymethyl (C) Cation:Na (C-term)
	Display all modifications □		
Variable modifications	--- none selected ---		
Peptide tol. ±	0.6 Da # ^{13}C 0	MS/MS tol. ±	0.6 Da
Peptide charge	1+	Monoisotopic ● Average ○	
Query			
Instrument	ESI-TRAP		
Decoy	□	Report top	AUTO hits

You can enter sequence tags, and other types of query, into the sequence query form.

Remember that all the search parameters, including enzyme specificity, modifications, and precursor charge, still apply to this type of search.

Mascot will look for a match between the tag and the ion series specified by the instrument type. Note that Mascot will only try to match the tag against ion series formed by a single backbone cleavage, and maybe a neutral loss, like y or b* or y++. It won’t try to match against side chain cleavage fragments, like d, v, w or internal fragments.



Standard sequence tag

Keyword is tag

What’s (probably) wrong with this tag?

```
1890.2 tag(1004.1, LSADTG, 1548.5)
```

Very unlikely that you would be able to call L from a spectrum. Should be

```
1890.2 tag(1004.1, [I|L]SADTG, 1548.5)
```

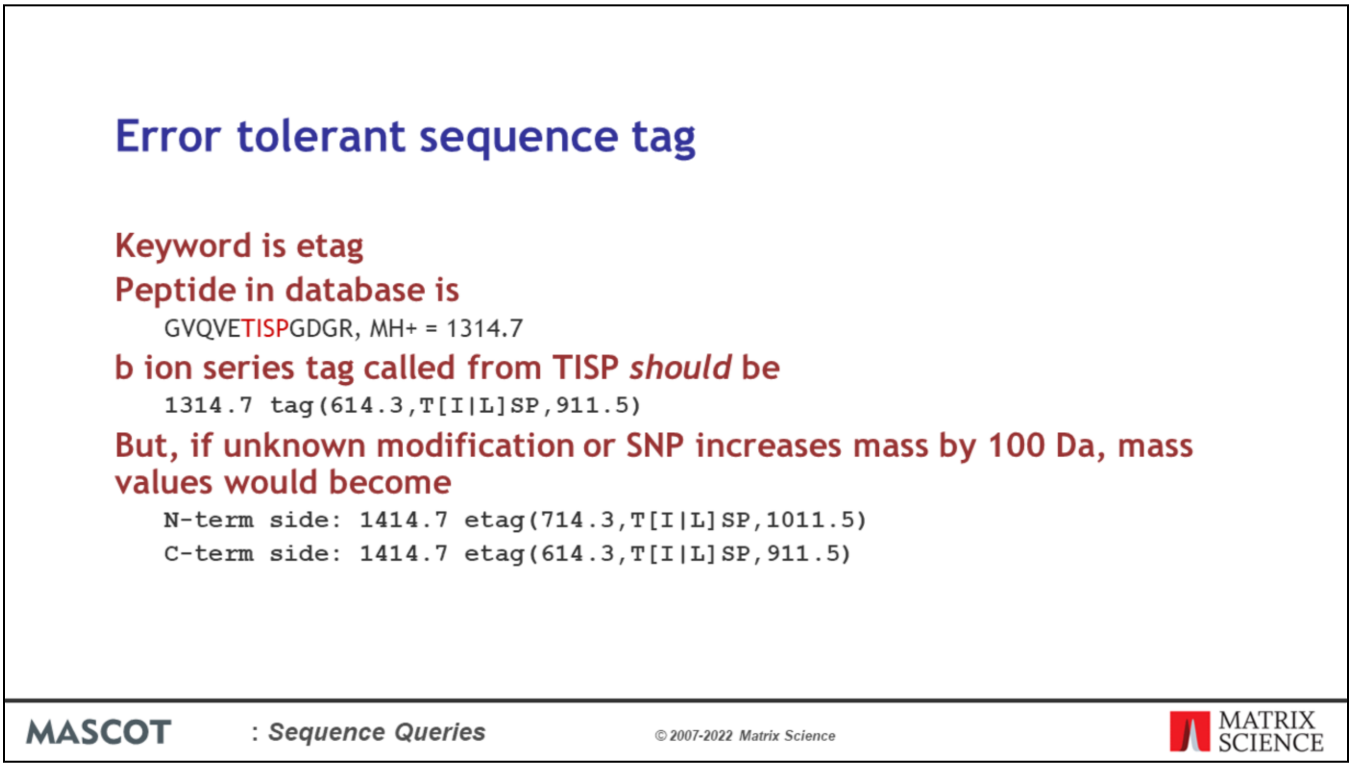
Ambiguity is OK as long as it is explicitly represented

```
877.4 tag(376.2, [I|L][Q|K][I|L], 730.2)
1869.93 tag(345.14, [I|L]A[VG|GV|R][M|F]G, 889.45)
(VG = R, F = MetOx)
```

MASCOT : *Sequence Queries* © 2007-2022 Matrix Science MATRIX SCIENCE

Unless you have high energy fragmentation, and are able to distinguish L from I by side chain cleavage fragments, then this tag is wrong. It should be I or L.

Ambiguity in a tag is fine as long as it is recognised and spelt out. Most times, you won’t know whether a residue is Q or K. F is almost identical to oxidised M. If the peaks are weak, are you sure you have a mass difference of R, or could it be VG and the intermediate peak is missing?



Error tolerant sequence tag

Keyword is etag

Peptide in database is

GVQVETISPGDGR, MH+ = 1314.7

b ion series tag called from TISP *should* be

```
1314.7 tag(614.3,T[I|L]SP,911.5)
```

But, if unknown modification or SNP increases mass by 100 Da, mass values would become

```
N-term side: 1414.7 etag(714.3,T[I|L]SP,1011.5)
C-term side: 1414.7 etag(614.3,T[I|L]SP,911.5)
```

MASCOT : *Sequence Queries* © 2007-2022 Matrix Science MATRIX SCIENCE

If the sequence is in the database, it is easier and safer to perform an MS/MS search of the peak list. In this sense, the standard sequence tag is obsolete.

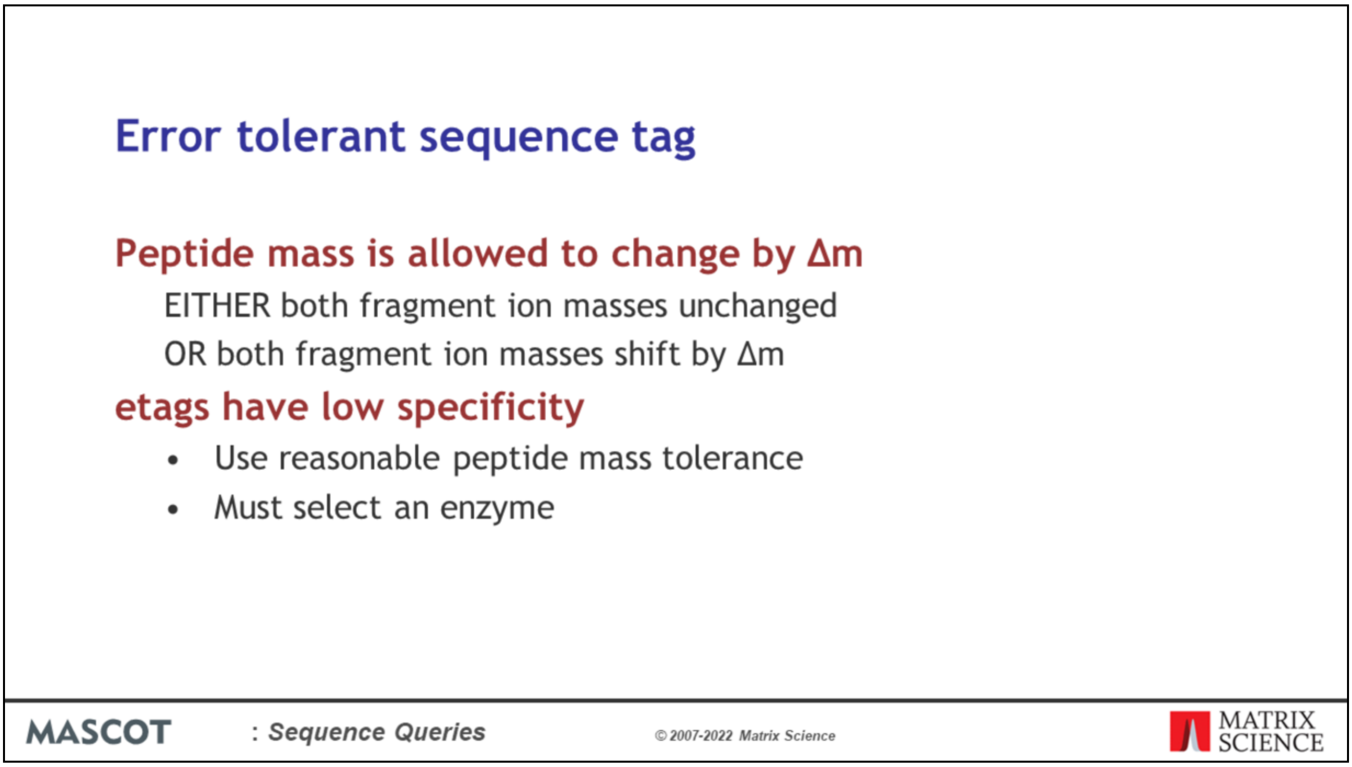
The error tolerant tag, which can find a match when there is an unsuspected modification or a small difference in the sequence, is very powerful and very useful.

Imagine we had an unmodified peptide of MH+ 1314.7 and we interpreted a tag of TISP in the b+ series between peaks at 614.3 and 911.5.

What happens if there is a modification or SNP that increases mass by 100 Da?

If the mod is on the N-term side of the tag, all the masses shift up by 100. However, if it is on the C-term side, only the peptide mass changes.

If the tag was in the y ion series, the reverse would be observed



Error tolerant sequence tag

Peptide mass is allowed to change by Δm

EITHER both fragment ion masses unchanged

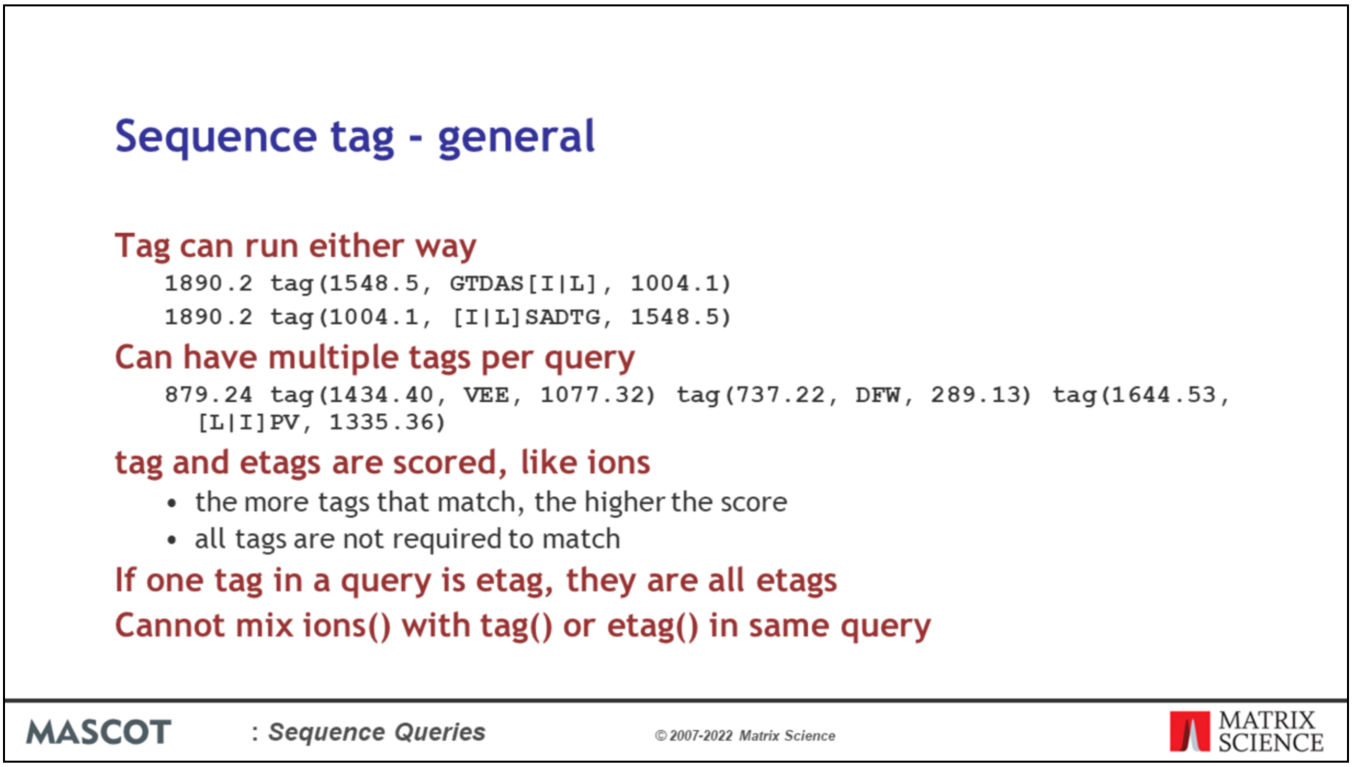
OR both fragment ion masses shift by Δm

etags have low specificity

- Use reasonable peptide mass tolerance
- Must select an enzyme

The error tolerant tag allows for this. In effect, it allows the peptide mass to vary and allows the tag to float. However, the tag must stay attached to one end or the other. Either both fragment ion masses are unchanged or both fragment ion masses shift by the same amount as the precursor.

This causes a huge loss of specificity, so we cannot allow etag searches with very wide peptide mass tolerance (> 1% or > 10 Da) or with no enzyme specificity. The enzyme specificity in an etag search is never fully specific, in any case, because one end of the peptide can just extend until it finds a cleavage point.



Sequence tag - general

Tag can run either way

```
1890.2 tag(1548.5, GTDAS[I|L], 1004.1)
1890.2 tag(1004.1, [I|L]SADTG, 1548.5)
```

Can have multiple tags per query

```
879.24 tag(1434.40, VEE, 1077.32) tag(737.22, DFW, 289.13) tag(1644.53,
  [L|I]PV, 1335.36)
```

tag and etags are scored, like ions

- the more tags that match, the higher the score
- all tags are not required to match

If one tag in a query is etag, they are all etags

Cannot mix ions() with tag() or etag() in same query

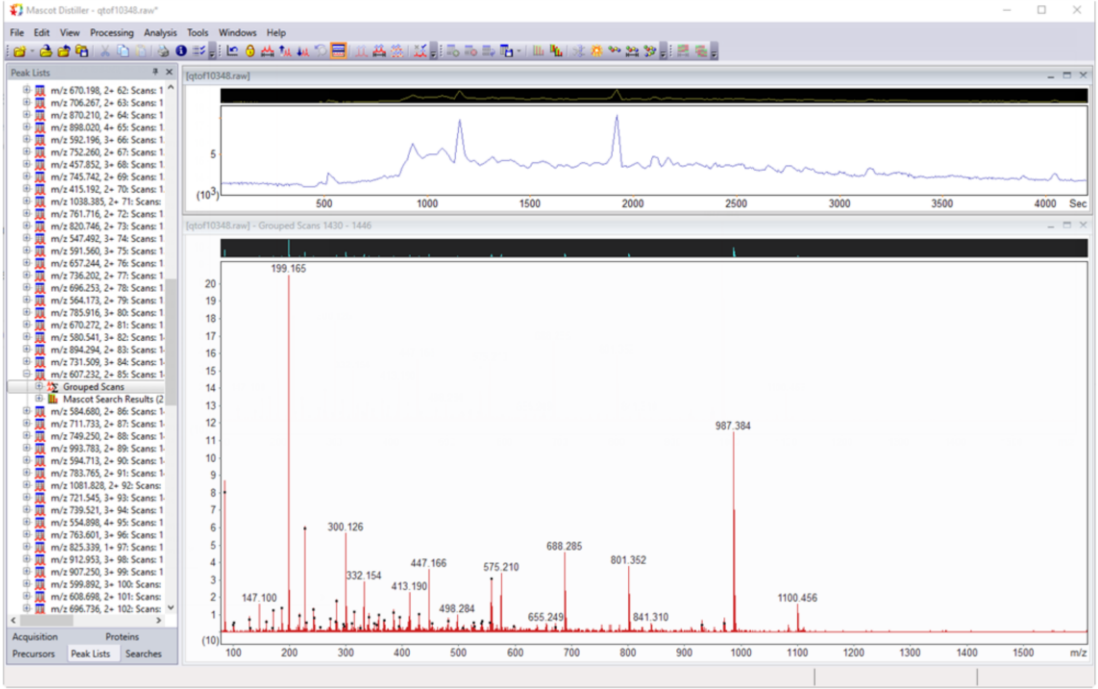
MASCOT : *Sequence Queries* © 2007-2022 *Matrix Science* MATRIX SCIENCE

Tags can be entered with the high mass fragment on the left or the right. These two tags are identical

Mascot allows multiple tags in a single query. That is, you can call multiple tags from a single MS/MS spectrum. Tags are scored probabilistically. If one tag is wrong, you can still get a good match from the tags that are correct.

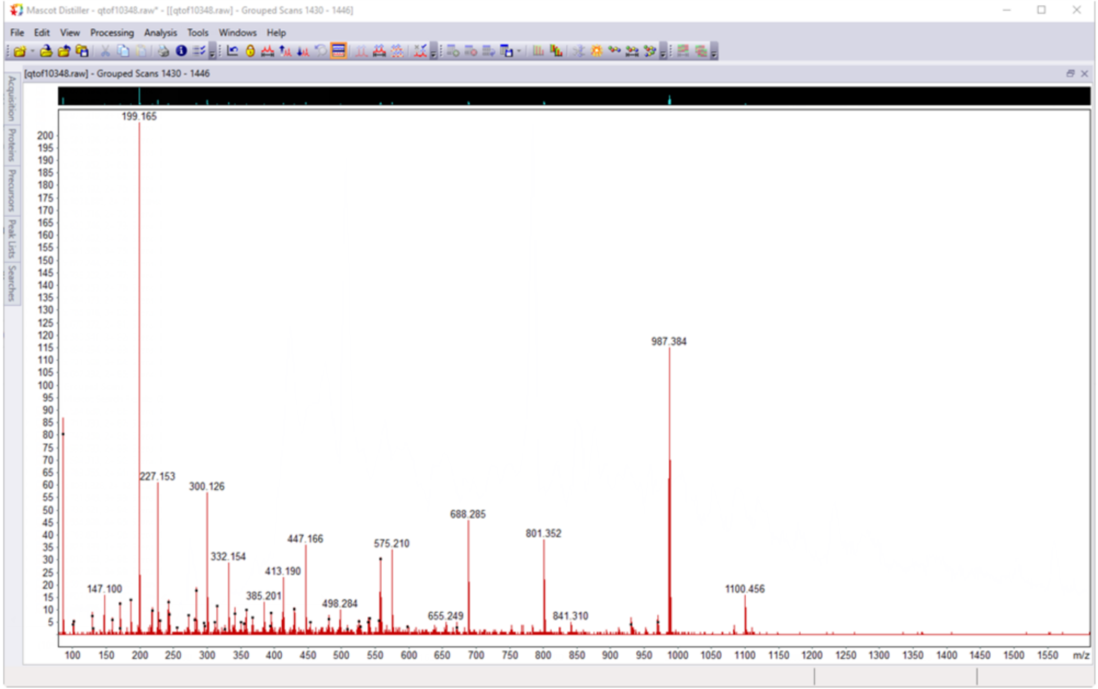
If one tag in a query is an etag than all the tags for that query are treated as etags, (not all tags in the search, just in the query)

Finally, you cannot mix ions qualifiers with tag or etag qualifiers. It would just be too complicated.



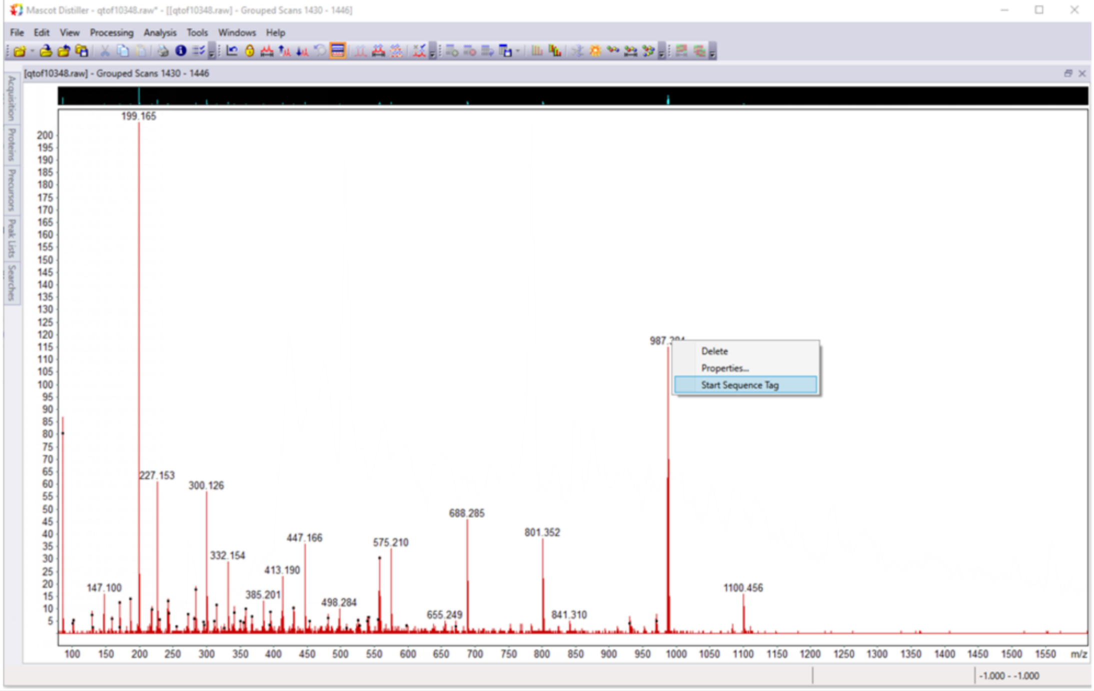
m/z 670.198, 2+ 62: Scans: 1
m/z 706.267, 2+ 63: Scans: 1
m/z 870.210, 2+ 64: Scans: 1
m/z 898.020, 4+ 65: Scans: 1
m/z 592.196, 3+ 66: Scans: 1
m/z 752.260, 2+ 67: Scans: 1
m/z 457.852, 3+ 68: Scans: 1
m/z 745.742, 2+ 69: Scans: 1
m/z 415.192, 2+ 70: Scans: 1
m/z 1038.385, 2+ 71: Scans:
m/z 761.716, 2+ 72: Scans: 1
m/z 820.746, 2+ 73: Scans: 1
m/z 547.492, 3+ 74: Scans: 1
m/z 591.560, 3+ 75: Scans: 1
m/z 657.244, 2+ 76: Scans: 1
m/z 736.202, 2+ 77: Scans: 1
m/z 696.253, 2+ 78: Scans: 1
m/z 564.173, 2+ 79: Scans: 1
m/z 785.916, 3+ 80: Scans: 1
m/z 670.272, 2+ 81: Scans: 1
m/z 580.541, 3+ 82: Scans: 1-
m/z 894.294, 2+ 83: Scans: 1-
m/z 731.509, 3+ 84: Scans: 1-
m/z 607.232, 2+ 85: Scans: 1-
Grouped Scans
Mascot Search Results (2
m/z 584.680, 2+ 86: Scans: 1-
m/z 711.733, 2+ 87: Scans: 1-
m/z 749.250, 2+ 88: Scans: 1-
m/z 993.783, 2+ 89: Scans: 1-
m/z 594.713, 2+ 90: Scans: 1-
m/z 783.765, 2+ 91: Scans: 1-
m/z 1081.828, 2+ 92: Scans:
m/z 721.545, 3+ 93: Scans: 1-
m/z 739.521, 3+ 94: Scans: 1
m/z 554.898, 4+ 95: Scans: 1
m/z 763.601, 3+ 96: Scans: 1
m/z 825.339, 1+ 97: Scans: 1
m/z 912.953, 3+ 98: Scans: 1
m/z 907.250, 3+ 99: Scans: 1
m/z 599.892, 3+ 100: Scans:
m/z 608.698, 2+ 101: Scans:
m/z 696.736, 2+ 102: Scans:

A lot of people call tags using a calculator and a table of mass values. An alternative is to use Mascot Distiller. Let’s see how this works

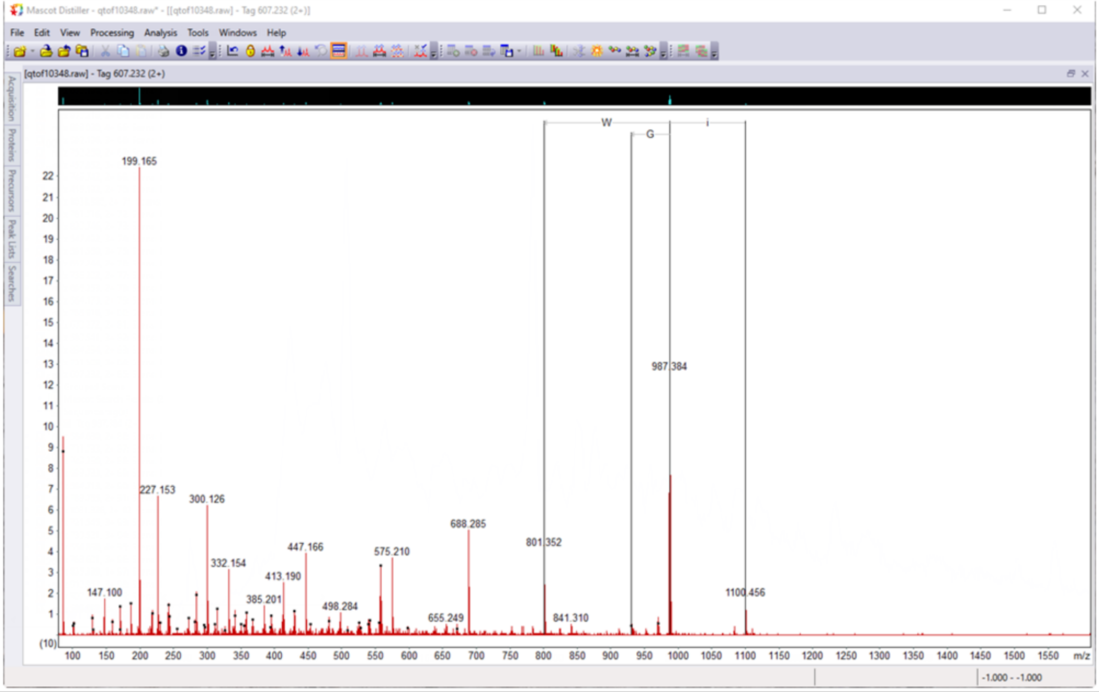

Mascot Distiller - qtof10348.raw* - [[qtof10348.raw] - Grouped Scans 1430 - 1446]
File Edit View Processing Analysis Tools Windows Help
[qtof10348.raw] - Grouped Scans 1430 - 1446
Acquisition
Proteins
Precursors
Peak Lists
Searches
199.165
987.384
227.153
300.126
688.285
447.166
575.210
801.352
332.154
413.190
147.100
385.201
498.284
1100.456
655.249
841.310
200
195
190
185
180
175
170
165
160
155
150
145
140
135
130
125
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1000 1050 1100 1150 1200 1250 1300 1350 1400 1450 1500 1550 m/z

Maximise the window

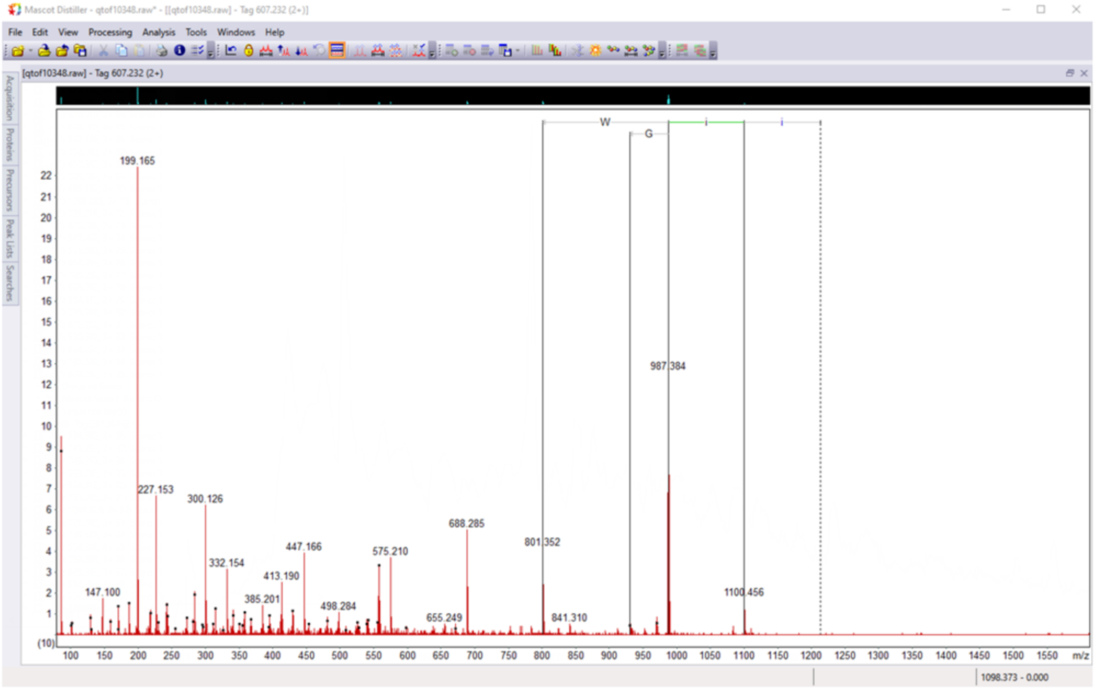
Choose a likely looking peak, such as 987.384



Right click to start a tag

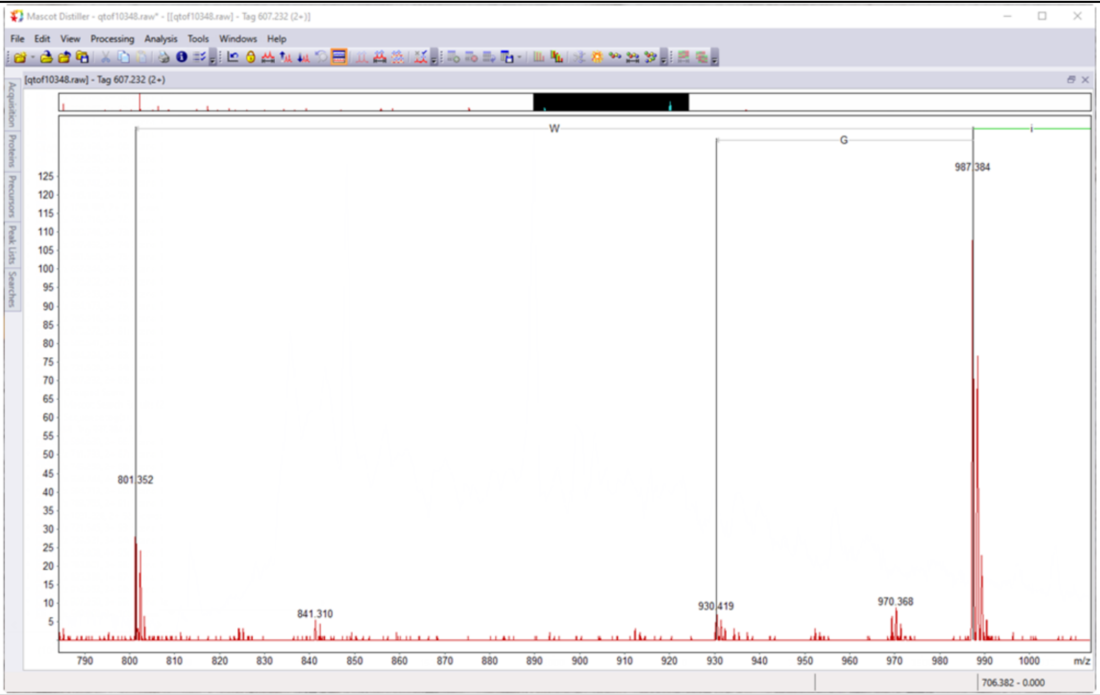


Click on any arrow to extend the tag. Where there's a choice, I'll just go for the biggest peak and stop when it starts to look tricky

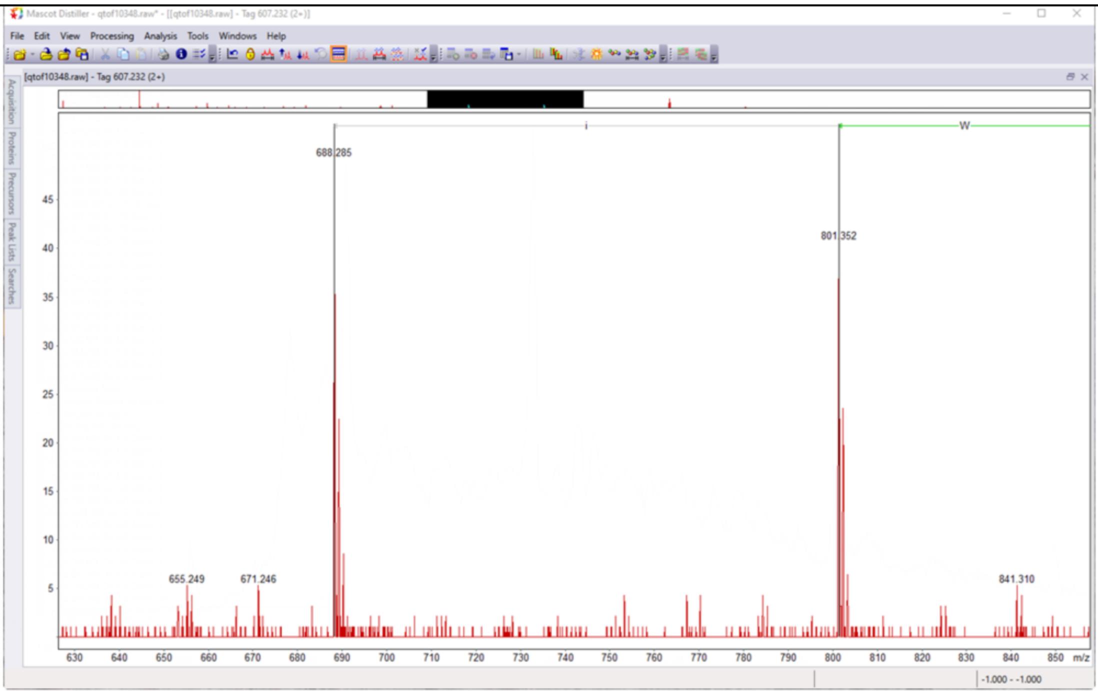


Going to the right, there's no choice and the descender from the second i (=I or L) is dotted to indicate we've reached the precursor mass, which is a bit of luck.

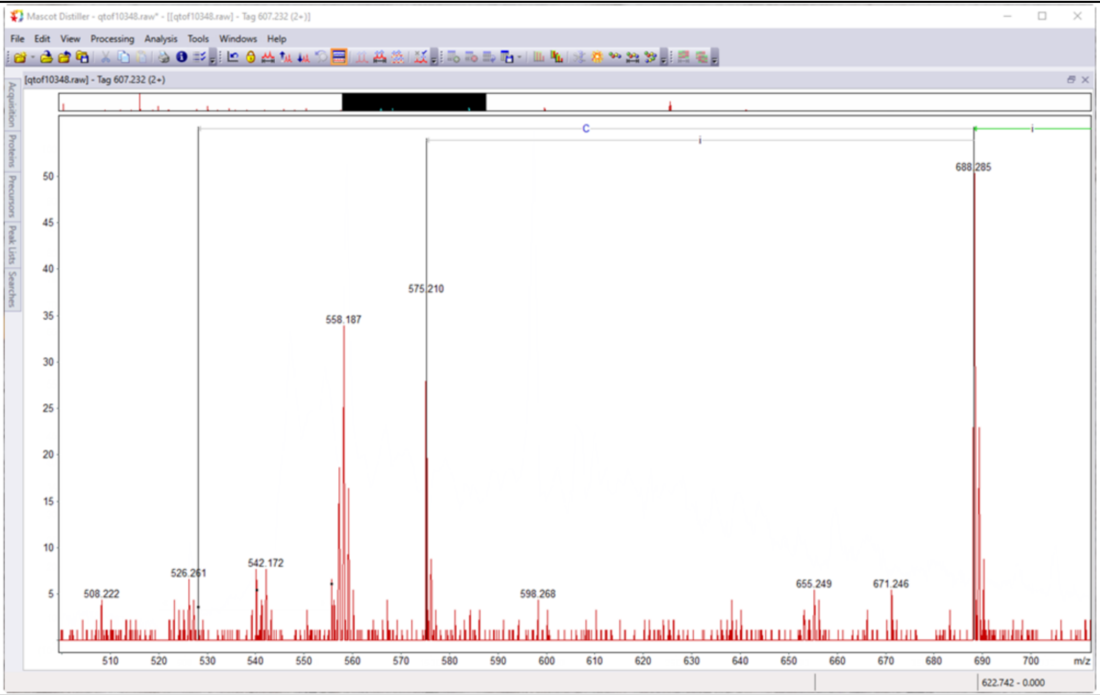
Going to the left, there is a choice of three residues. Zoom in for a closer look



No doubt that W is the most intense, so we click on the arrow head to extend it

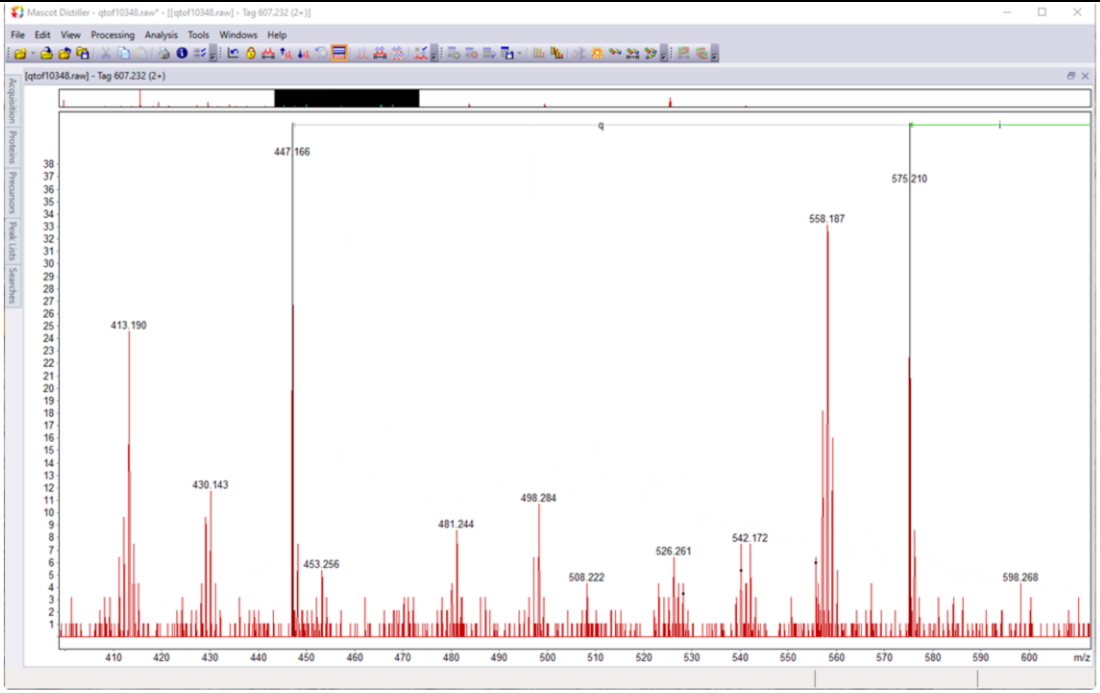


This time only one choice so we’ll go with i

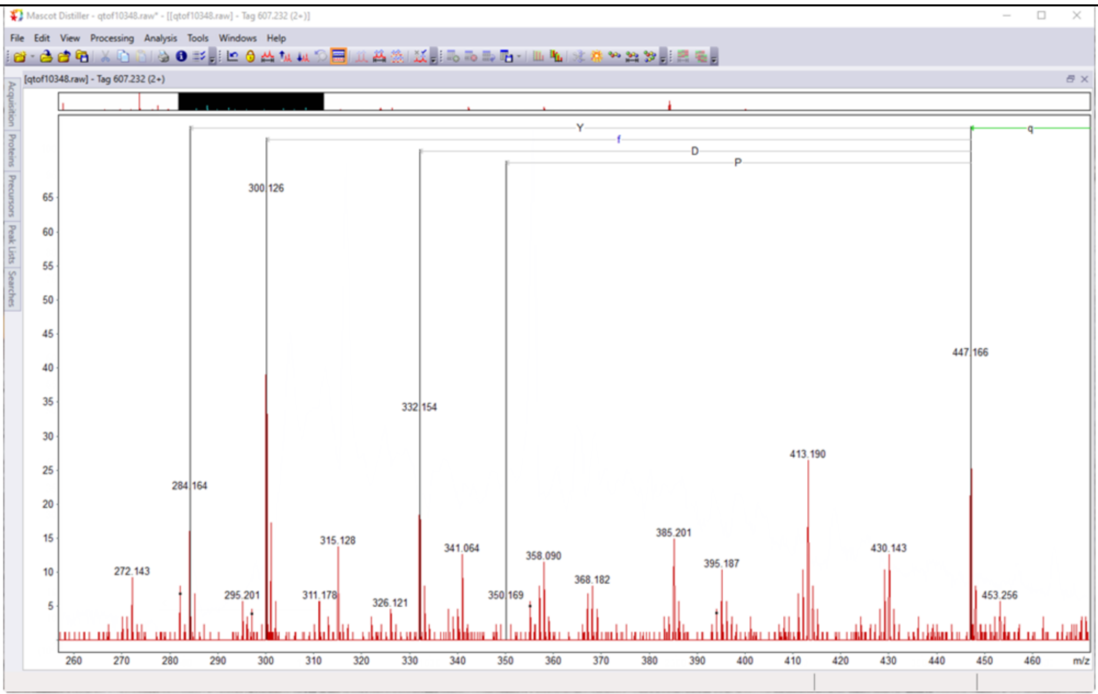


MASCOT : *Sequence Queries* © 2007-2022 Matrix Science MATRIX SCIENCE

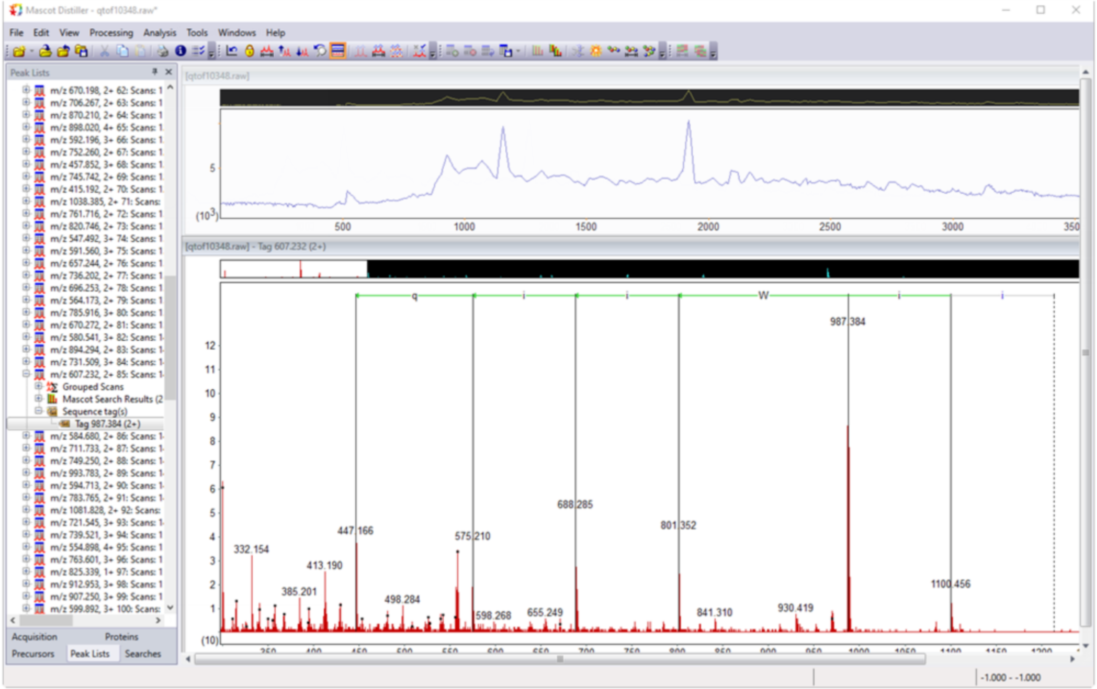
C looks a bit dodgy, so we’ll go with I again



No choice … good!



Lots of choice. Maybe time to give up



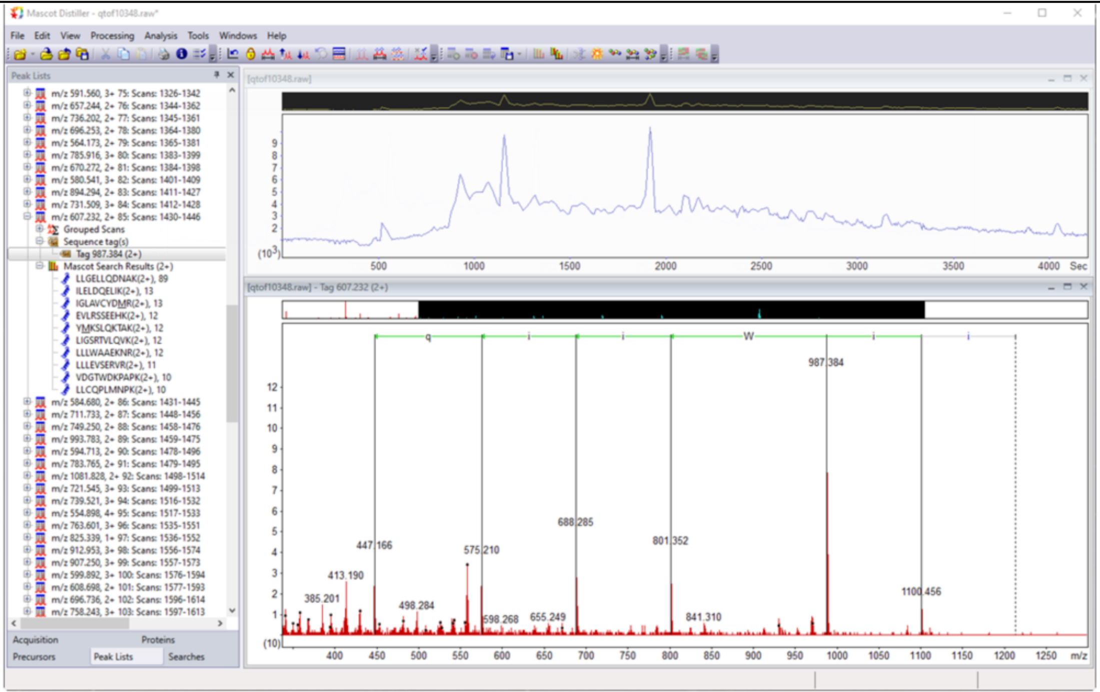
Mascot Distiller - qtof10348.raw*

Peak Lists

- m/z 670.198, 2+ 62: Scans: 1
- m/z 706.267, 2+ 63: Scans: 1
- m/z 870.210, 2+ 64: Scans: 1
- m/z 898.020, 4+ 65: Scans: 1
- m/z 592.196, 3+ 66: Scans: 1
- m/z 752.260, 2+ 67: Scans: 1
- m/z 457.852, 3+ 68: Scans: 1
- m/z 745.742, 2+ 69: Scans: 1
- m/z 415.192, 2+ 70: Scans: 1
- m/z 1038.385, 2+ 71: Scans:
- m/z 761.716, 2+ 72: Scans: 1
- m/z 820.746, 2+ 73: Scans: 1
- m/z 547.492, 3+ 74: Scans: 1
- m/z 591.560, 3+ 75: Scans: 1
- m/z 657.244, 2+ 76: Scans: 1
- m/z 736.202, 2+ 77: Scans: 1
- m/z 696.253, 2+ 78: Scans: 1
- m/z 564.173, 2+ 79: Scans: 1
- m/z 785.916, 3+ 80: Scans: 1
- m/z 670.272, 2+ 81: Scans: 1
- m/z 580.541, 3+ 82: Scans: 1-
- m/z 894.294, 2+ 83: Scans: 1-
- m/z 731.509, 3+ 84: Scans: 1-
- m/z 607.232, 2+ 85: Scans: 1-
  - Grouped Scans
  - Mascot Search Results (2
  - Sequence tag(s)
    - Tag 987.384 (2+)
- m/z 584.680, 2+ 86: Scans: 1-
- m/z 711.733, 2+ 87: Scans: 1-
- m/z 749.250, 2+ 88: Scans: 1-
- m/z 993.783, 2+ 89: Scans: 1-
- m/z 594.713, 2+ 90: Scans: 1-
- m/z 783.765, 2+ 91: Scans: 1-
- m/z 1081.828, 2+ 92: Scans:
- m/z 721.545, 3+ 93: Scans: 1-
- m/z 739.521, 3+ 94: Scans: 1
- m/z 554.898, 4+ 95: Scans: 1
- m/z 763.601, 3+ 96: Scans: 1
- m/z 825.339, 1+ 97: Scans: 1
- m/z 912.953, 3+ 98: Scans: 1
- m/z 907.250, 3+ 99: Scans: 1
- m/z 599.892, 3+ 100: Scans:

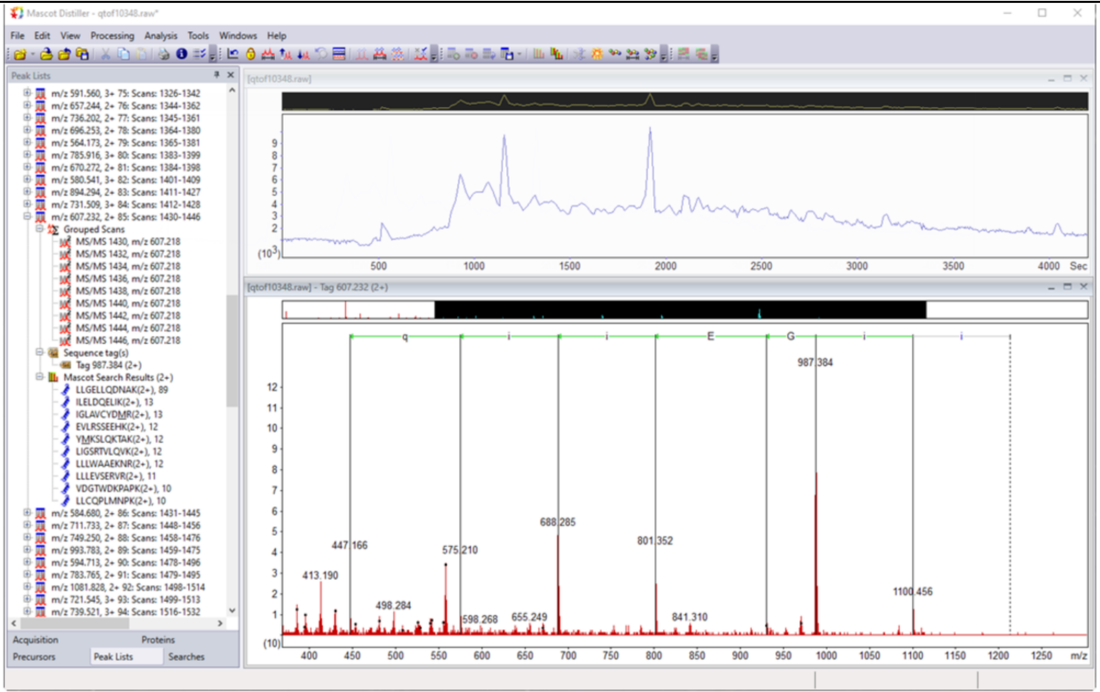
[qtof10348.raw] - Tag 607.232 (2+)

Restore the window arrangement. Here’s our tag, but is it correct? Well, we just happen to have database search results from this spectrum



m/z 591.560, 3+ 75: Scans: 1326-1342
m/z 657.244, 2+ 76: Scans: 1344-1362
m/z 736.202, 2+ 77: Scans: 1345-1361
m/z 696.253, 2+ 78: Scans: 1364-1380
m/z 564.173, 2+ 79: Scans: 1365-1381
m/z 785.916, 3+ 80: Scans: 1383-1399
m/z 670.272, 2+ 81: Scans: 1384-1398
m/z 580.541, 3+ 82: Scans: 1401-1409
m/z 894.294, 2+ 83: Scans: 1411-1427
m/z 731.509, 3+ 84: Scans: 1412-1428
m/z 607.232, 2+ 85: Scans: 1430-1446
Grouped Scans
Sequence tag(s)
Tag 987.384 (2+)
Mascot Search Results (2+)
LLGELLQDNAK(2+), 89
ILELDQELIK(2+), 13
IGLAVCYDMR(2+), 13
EVLRSSEEHK(2+), 12
YMKSLQKTAK(2+), 12
LIGSRTVLQVK(2+), 12
LLLWAAEKNR(2+), 12
LLLEVSERVR(2+), 11
VDGTWDKPAPK(2+), 10
LLCQPLMNPK(2+), 10
m/z 584.680, 2+ 86: Scans: 1431-1445
m/z 711.733, 2+ 87: Scans: 1448-1456
m/z 749.250, 2+ 88: Scans: 1458-1476
m/z 993.783, 2+ 89: Scans: 1459-1475
m/z 594.713, 2+ 90: Scans: 1478-1496
m/z 783.765, 2+ 91: Scans: 1479-1495
m/z 1081.828, 2+ 92: Scans: 1498-1514
m/z 721.545, 3+ 93: Scans: 1499-1513
m/z 739.521, 3+ 94: Scans: 1516-1532
m/z 554.898, 4+ 95: Scans: 1517-1533
m/z 763.601, 3+ 96: Scans: 1535-1551
m/z 825.339, 1+ 97: Scans: 1536-1552
m/z 912.953, 3+ 98: Scans: 1556-1574
m/z 907.250, 3+ 99: Scans: 1557-1573
m/z 599.892, 3+ 100: Scans: 1576-1594
m/z 608.698, 2+ 101: Scans: 1577-1593
m/z 696.736, 2+ 102: Scans: 1596-1614
m/z 758.243, 3+ 103: Scans: 1597-1613

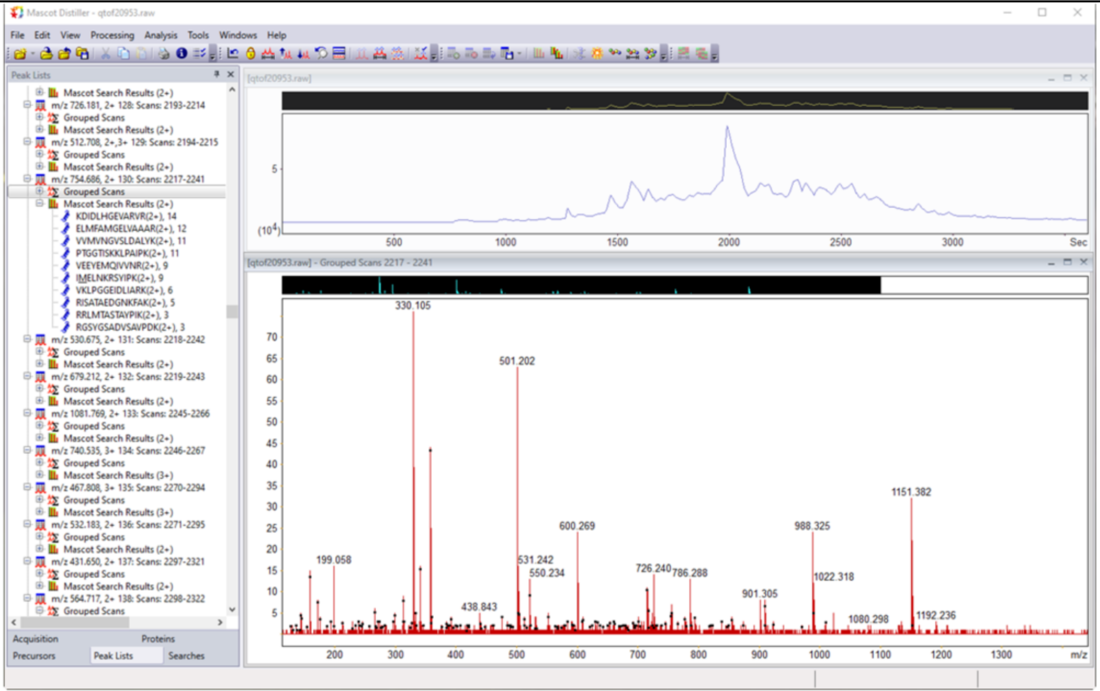
Our tag was wrong. It should have been GE in the middle, not W. Note that the sequence is running left to right, telling us this is a y ion tag and the terminus we reached was the amino terminus



```
m/z 591.560, 3+ 75: Scans: 1326-1342
m/z 657.244, 2+ 76: Scans: 1344-1362
m/z 736.202, 2+ 77: Scans: 1345-1361
m/z 696.253, 2+ 78: Scans: 1364-1380
m/z 564.173, 2+ 79: Scans: 1365-1381
m/z 785.916, 3+ 80: Scans: 1383-1399
m/z 670.272, 2+ 81: Scans: 1384-1398
m/z 580.541, 3+ 82: Scans: 1401-1409
m/z 894.294, 2+ 83: Scans: 1411-1427
m/z 731.509, 3+ 84: Scans: 1412-1428
m/z 607.232, 2+ 85: Scans: 1430-1446
    Grouped Scans
        MS/MS 1430, m/z 607.218
        MS/MS 1432, m/z 607.218
        MS/MS 1434, m/z 607.218
        MS/MS 1436, m/z 607.218
        MS/MS 1438, m/z 607.218
        MS/MS 1440, m/z 607.218
        MS/MS 1442, m/z 607.218
        MS/MS 1444, m/z 607.218
        MS/MS 1446, m/z 607.218
    Sequence tag(s)
        Tag 987.384 (2+)
    Mascot Search Results (2+)
        LLGELLQDNAK(2+), 89
        ILELDQELIK(2+), 13
        IGLAVCYDMR(2+), 13
        EVLRSSEEHK(2+), 12
        YMKSLQKTAK(2+), 12
        LIGSRTVLQVK(2+), 12
        LLLWAAEKNR(2+), 12
        LLLEVSERVR(2+), 11
        VDGTWDKPAPK(2+), 10
        LLCQPLMNPK(2+), 10
m/z 584.680, 2+ 86: Scans: 1431-1445
m/z 711.733, 2+ 87: Scans: 1448-1456
m/z 749.250, 2+ 88: Scans: 1458-1476
m/z 993.783, 2+ 89: Scans: 1459-1475
m/z 594.713, 2+ 90: Scans: 1478-1496
m/z 783.765, 2+ 91: Scans: 1479-1495
m/z 1081.828, 2+ 92: Scans: 1498-1514
m/z 721.545, 3+ 93: Scans: 1499-1513
m/z 739.521, 3+ 94: Scans: 1516-1532
```

This is what it should have looked like. I obviously need more practice!

That is a rather artificial example, of course, because we have a good match from the database search, so no-one would need to call a sequence tag

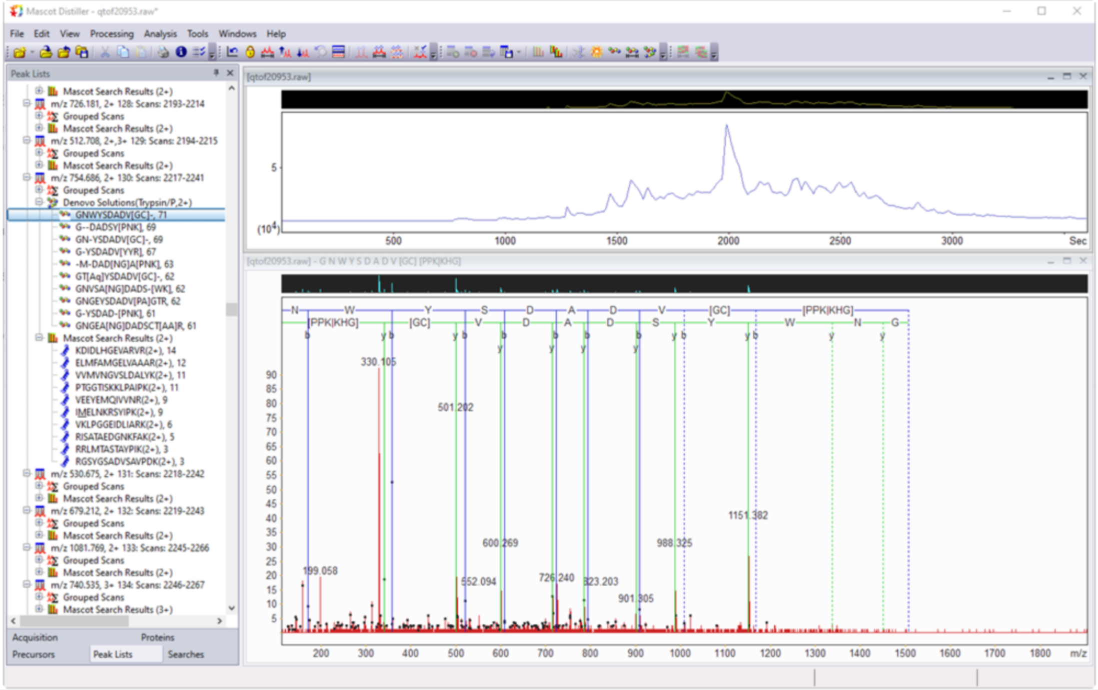


```
Mascot Search Results (2+)
m/z 726.181, 2+ 128: Scans: 2193-2214
  Grouped Scans
  Mascot Search Results (2+)
m/z 512.708, 2+,3+ 129: Scans: 2194-2215
  Grouped Scans
  Mascot Search Results (2+)
m/z 754.686, 2+ 130: Scans: 2217-2241
  Grouped Scans
  Mascot Search Results (2+)
    KDIDLHGEVARVR(2+), 14
    ELMFAMGELVAAAR(2+), 12
    VVMVNGVSLDALYK(2+), 11
    PTGGTISKKLPAIPK(2+), 11
    VEEYEMQIVVNR(2+), 9
    IMELNKRSYIPK(2+), 9
    VKLPGGEIDLIARK(2+), 6
    RISATAEDGNKFAK(2+), 5
    RRLMTASTAYPIK(2+), 3
    RGSYGSADVSAVPDK(2+), 3
m/z 530.675, 2+ 131: Scans: 2218-2242
  Grouped Scans
  Mascot Search Results (2+)
m/z 679.212, 2+ 132: Scans: 2219-2243
  Grouped Scans
  Mascot Search Results (2+)
m/z 1081.769, 2+ 133: Scans: 2245-2266
  Grouped Scans
  Mascot Search Results (2+)
m/z 740.535, 3+ 134: Scans: 2246-2267
  Grouped Scans
  Mascot Search Results (3+)
m/z 467.808, 3+ 135: Scans: 2270-2294
  Grouped Scans
  Mascot Search Results (3+)
m/z 532.183, 2+ 136: Scans: 2271-2295
  Grouped Scans
  Mascot Search Results (2+)
m/z 431.650, 2+ 137: Scans: 2297-2321
  Grouped Scans
  Mascot Search Results (2+)
m/z 564.717, 2+ 138: Scans: 2298-2322
  Grouped Scans
```

Alternatively, you can automate the process entirely by using the de novo algorithm.

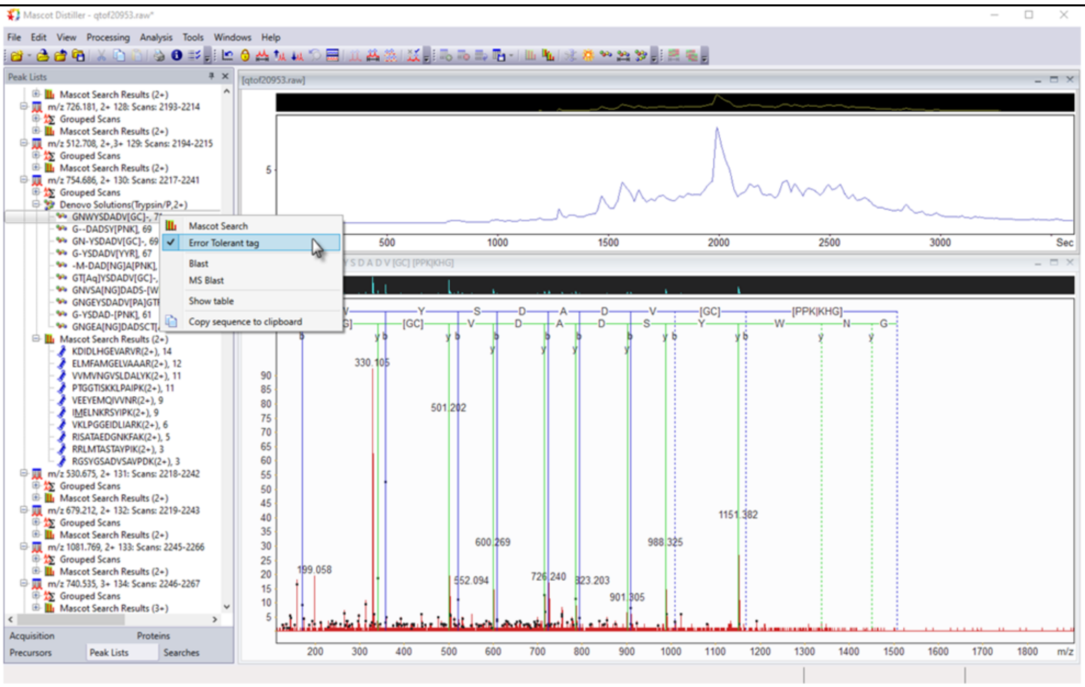
Here’s a nice spectrum in another data set where the Mascot database search has failed to find a match

If we right click the peak list and choose de novo …


Mascot Distiller - qtof20953.raw*
File Edit View Processing Analysis Tools Windows Help
Peak Lists
Acquisition
Proteins
Precursors
Peak Lists
Searches
[qtof20953.raw]
5
(10⁴)
500
1000
1500
2000
2500
3000
Sec
[qtof20953.raw] - G N W Y S D A D V [GC] [PPK|KHG]
N W Y S D A D V [GC] [PPK|KHG]
[PPK|KHG] [GC] V D A D S Y W N G
b y b b b b b y b y b y y
90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5
330.105
501.202
600.269
1151.382
988.325
199.058
552.094
726.240
823.203
901.305
200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 m/z

```
⊞ Mascot Search Results (2+)
⊟ m/z 726.181, 2+ 128: Scans: 2193-2214
  ⊞ Grouped Scans
  ⊞ Mascot Search Results (2+)
⊟ m/z 512.708, 2+,3+ 129: Scans: 2194-2215
  ⊞ Grouped Scans
  ⊞ Mascot Search Results (2+)
⊟ m/z 754.686, 2+ 130: Scans: 2217-2241
  ⊞ Grouped Scans
  ⊟ Denovo Solutions(Trypsin/P,2+)
      GNWYSDADV[GC]-, 71
      G--DADSY[PNK], 69
      GN-YSDADV[GC]-, 69
      G-YSDADV[YYR], 67
      -M-DAD[NG]A[PNK], 63
      GT[Aq]YSDADV[GC]-, 62
      GNVSA[NG]DADS-[WK], 62
      GNGEYSDADV[PA]GTR, 62
      G-YSDAD-[PNK], 61
      GNGEA[NG]DADSCT[AA]R, 61
  ⊟ Mascot Search Results (2+)
      KDIDLHGEVARVR(2+), 14
      ELMFAMGELVAAAR(2+), 12
      VVMVNGVSLDALYK(2+), 11
      PTGGTISKKLPAIPK(2+), 11
      VEEYEMQIVVNR(2+), 9
      IMELNKRSYIPK(2+), 9
      VKLPGGEIDLIARK(2+), 6
      RISATAEDGNKFAK(2+), 5
      RRLMTASTAYPIK(2+), 3
      RGSYGSADVSAVPDK(2+), 3
⊟ m/z 530.675, 2+ 131: Scans: 2218-2242
  ⊞ Grouped Scans
  ⊞ Mascot Search Results (2+)
⊟ m/z 679.212, 2+ 132: Scans: 2219-2243
  ⊞ Grouped Scans
  ⊞ Mascot Search Results (2+)
⊟ m/z 1081.769, 2+ 133: Scans: 2245-2266
  ⊞ Grouped Scans
  ⊞ Mascot Search Results (2+)
⊟ m/z 740.535, 3+ 134: Scans: 2246-2267
  ⊞ Grouped Scans
  ⊞ Mascot Search Results (3+)
```

We get a reasonably high scoring solution, but with some uncertainty

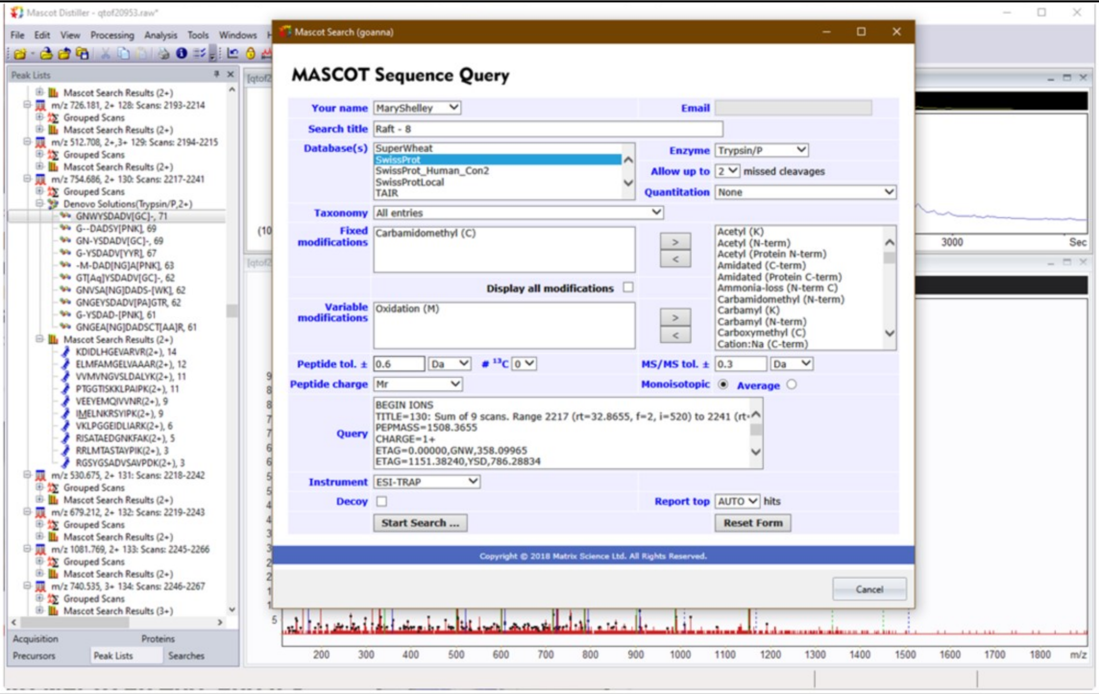


MASCOT : ***Sequence Queries***

© 2007-2022 Matrix Science

MATRIX SCIENCE

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


Mascot Distiller - qtof20953.raw*
File Edit View Processing Analysis Tools Windows
Peak Lists
Mascot Search (goanna)
Acquisition
Proteins
Precursors
Peak Lists
Searches
Sec
3000
200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 m/z

Mascot Search Results (2+)
m/z 726.181, 2+ 128: Scans: 2193-2214
Grouped Scans
Mascot Search Results (2+)
m/z 512.708, 2+,3+ 129: Scans: 2194-2215
Grouped Scans
Mascot Search Results (2+)
m/z 754.686, 2+ 130: Scans: 2217-2241
Grouped Scans
Denovo Solutions(Trypsin/P,2+)
GNWYSDADV[GC]-, 71
G--DADSY[PNK], 69
GN-YSDADV[GC]-, 69
G-YSDADV[YYR], 67
-M-DAD[NG]A[PNK], 63
GT[Aq]YSDADV[GC]-, 62
GNVSA[NG]DADS-[WK], 62
GNGEYSDADV[PA]GTR, 62
G-YSDAD-[PNK], 61
GNGEA[NG]DADSCT[AA]R, 61
Mascot Search Results (2+)
KDIDLHGEVARVR(2+), 14
ELMFAMGELVAAAR(2+), 12
VVMVNGVSLDALYK(2+), 11
PTGGTISKKLPAIPK(2+), 11
VEEYEMQIVVNR(2+), 9
IMELNKRSYIPK(2+), 9
VKLPGGEIDLIARK(2+), 6
RISATAEDGNKFAK(2+), 5
RRLMTASTAYPIK(2+), 3
RGSYGSADVSAVPDK(2+), 3
m/z 530.675, 2+ 131: Scans: 2218-2242
Grouped Scans
Mascot Search Results (2+)
m/z 679.212, 2+ 132: Scans: 2219-2243
Grouped Scans
Mascot Search Results (2+)
m/z 1081.769, 2+ 133: Scans: 2245-2266
Grouped Scans
Mascot Search Results (2+)
m/z 740.535, 3+ 134: Scans: 2246-2267
Grouped Scans
Mascot Search Results (3+)

MASCOT Sequence Query

Your name: MaryShelley
Email:
Search title: Raft - 8
Database(s): SuperWheat, SwissProt, SwissProt_Human_Con2, SwissProtLocal, TAIR
Enzyme: Trypsin/P
Allow up to 2 missed cleavages
Quantitation: None
Taxonomy: All entries
Fixed modifications: Carbamidomethyl (C)
Display all modifications □
Variable modifications: Oxidation (M)

Acetyl (K)
Acetyl (N-term)
Acetyl (Protein N-term)
Amidated (C-term)
Amidated (Protein C-term)
Ammonia-loss (N-term C)
Carbamidomethyl (N-term)
Carbamyl (K)
Carbamyl (N-term)
Carboxymethyl (C)
Cation:Na (C-term)

Peptide tol. ± 0.6 Da # ^{13}C 0
MS/MS tol. ± 0.3 Da
Peptide charge: Mr
Monoisotopic ● Average ○

Query:

```
BEGIN IONS
TITLE=130: Sum of 9 scans. Range 2217 (rt=32.8655, f=2, i=520) to 2241 (rt-
PEPMASS=1508.3655
CHARGE=1+
ETAG=0.00000,GNW,358.09965
ETAG=1151.38240,YSD,786.28834
```

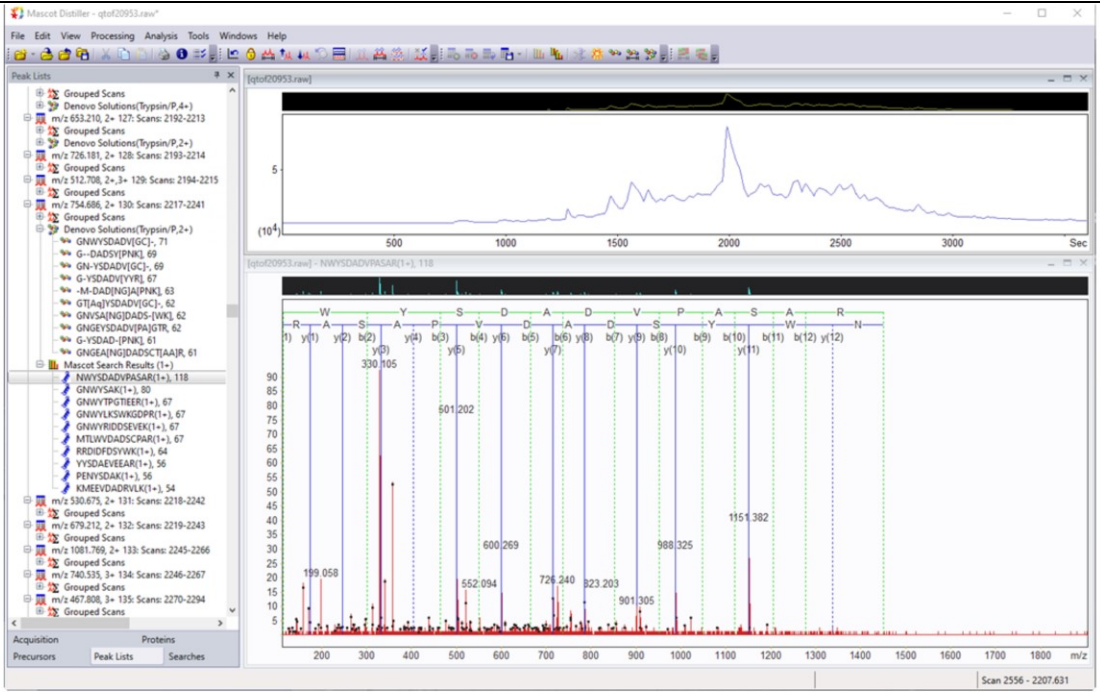
Instrument: ESI-TRAP
Decoy □
Report top AUTO hits

Start Search ... | Reset Form

Copyright © 2018 Matrix Science Ltd. All Rights Reserved.

Cancel

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

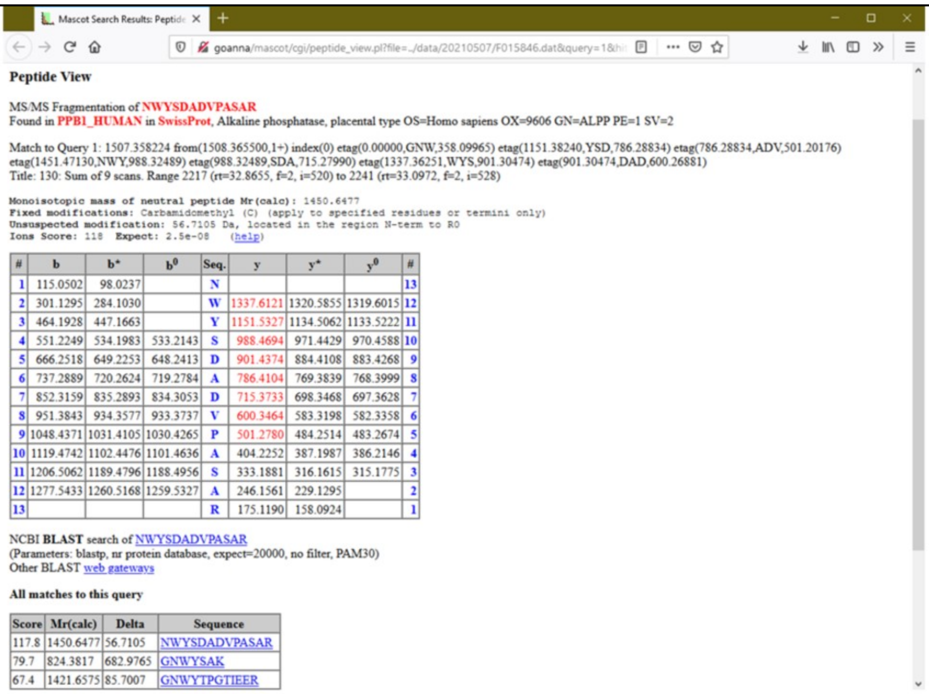


Grouped Scans
Denovo Solutions(Trypsin/P,4+)
m/z 653.210, 2+ 127: Scans: 2192-2213
Grouped Scans
Denovo Solutions(Trypsin/P,2+)
m/z 726.181, 2+ 128: Scans: 2193-2214
Grouped Scans
m/z 512.708, 2+,3+ 129: Scans: 2194-2215
Grouped Scans
m/z 754.686, 2+ 130: Scans: 2217-2241
Grouped Scans
Denovo Solutions(Trypsin/P,2+)
GNWYSDADV[GC]-, 71
G--DADSY[PNK], 69
GN-YSDADV[GC]-, 69
G-YSDADV[YYR], 67
-M-DAD[NG]A[PNK], 63
GT[Aq]YSDADV[GC]-, 62
GNVSA[NG]DADS-[WK], 62
GNGEYSDADV[PA]GTR, 62
G-YSDAD-[PNK], 61
GNGEA[NG]DADSCT[AA]R, 61
Mascot Search Results (1+)
NWYSDADVPASAR(1+), 118
GNWYSAK(1+), 80
GNWYTPGTIEER(1+), 67
GNWYLKSWKGDPR(1+), 67
GNWYRIDDSEVEK(1+), 67
MTLWVDADSCPAR(1+), 67
RRDIDFDSYWK(1+), 64
YYSDAEVEEAR(1+), 56
PENYSDAK(1+), 56
KMEEVDADRVLK(1+), 54
m/z 530.675, 2+ 131: Scans: 2218-2242
Grouped Scans
m/z 679.212, 2+ 132: Scans: 2219-2243
Grouped Scans
m/z 1081.769, 2+ 133: Scans: 2245-2266
Grouped Scans
m/z 740.535, 3+ 134: Scans: 2246-2267
Grouped Scans
m/z 467.808, 3+ 135: Scans: 2270-2294
Grouped Scans

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

This match looks promising. The sequence runs the same direction as the top de novo solution, but this is to be expected by chance if the de novo fails to reach one or other terminus.

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


Mascot Search Results: Peptide
goanna/mascot/cgi/peptide_view.pl?file=../data/20210507/F015846.dat&query=1&hi

Peptide View

MS/MS Fragmentation of NWYSDADVPASAR
Found in PPB1_HUMAN in SwissProt, Alkaline phosphatase, placental type OS=Homo sapiens OX=9606 GN=ALPP PE=1 SV=2

Match to Query 1: 1507.358224 from(1508.365500,1+) index(0) etag(0.00000,GNW,358.09965) etag(1151.38240,YSD,786.28834) etag(786.28834,ADV,501.20176) etag(1451.47130,NWY,988.32489) etag(988.32489,SDA,715.27990) etag(1337.36251,WYS,901.30474) etag(901.30474,DAD,600.26881)
Title: 130: Sum of 9 scans. Range 2217 (rt=32.8655, f=2, i=520) to 2241 (rt=33.0972, f=2, i=528)

```
Monoisotopic mass of neutral peptide Mr(calc): 1450.6477
Fixed modifications: Carbamidomethyl (C) (apply to specified residues or termini only)
Unsuspected modification: 56.7105 Da, located in the region N-term to R0
Ions Score: 118  Expect: 2.5e-08   (help)
```

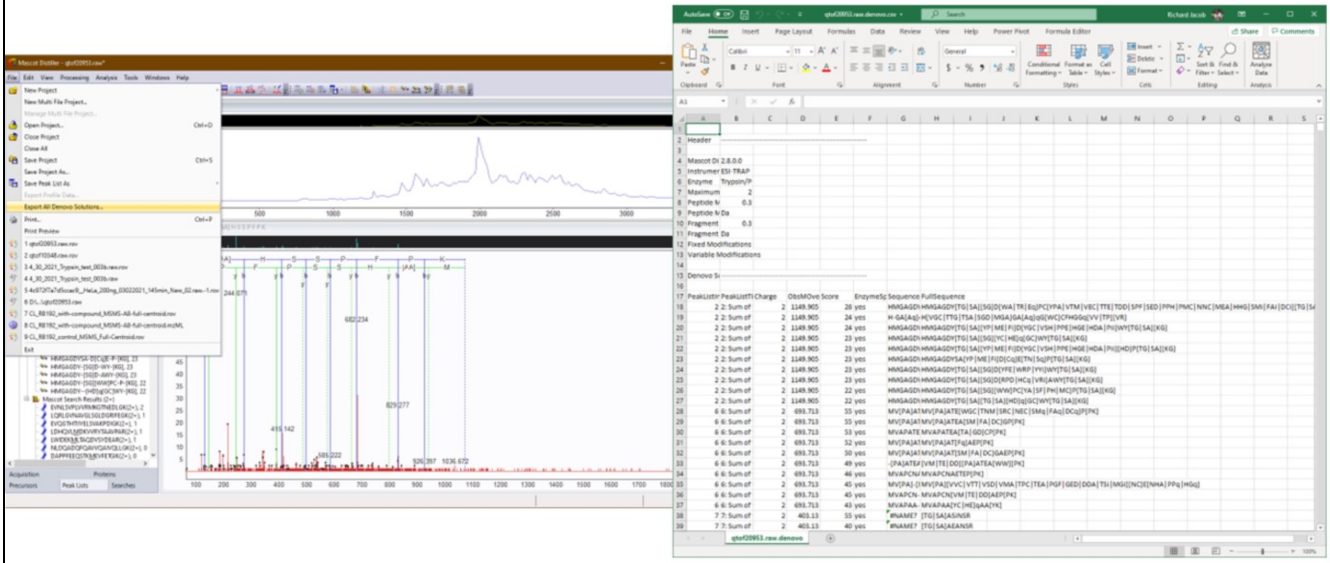
#	b	b^{*}	b^{0}	Seq.	y	y^{*}	y^{0}	#
1	115.0502	98.0237		N				13
2	301.1295	284.1030		W	1337.6121	1320.5855	1319.6015	12
3	464.1928	447.1663		Y	1151.5327	1134.5062	1133.5222	11
4	551.2249	534.1983	533.2143	S	988.4694	971.4429	970.4588	10
5	666.2518	649.2253	648.2413	D	901.4374	884.4108	883.4268	9
6	737.2889	720.2624	719.2784	A	786.4104	769.3839	768.3999	8
7	852.3159	835.2893	834.3053	D	715.3733	698.3468	697.3628	7
8	951.3843	934.3577	933.3737	V	600.3464	583.3198	582.3358	6
9	1048.4371	1031.4105	1030.4265	P	501.2780	484.2514	483.2674	5
10	1119.4742	1102.4476	1101.4636	A	404.2252	387.1987	386.2146	4
11	1206.5062	1189.4796	1188.4956	S	333.1881	316.1615	315.1775	3
12	1277.5433	1260.5168	1259.5327	A	246.1561	229.1295		2
13				R	175.1190	158.0924		1

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

All matches to this query

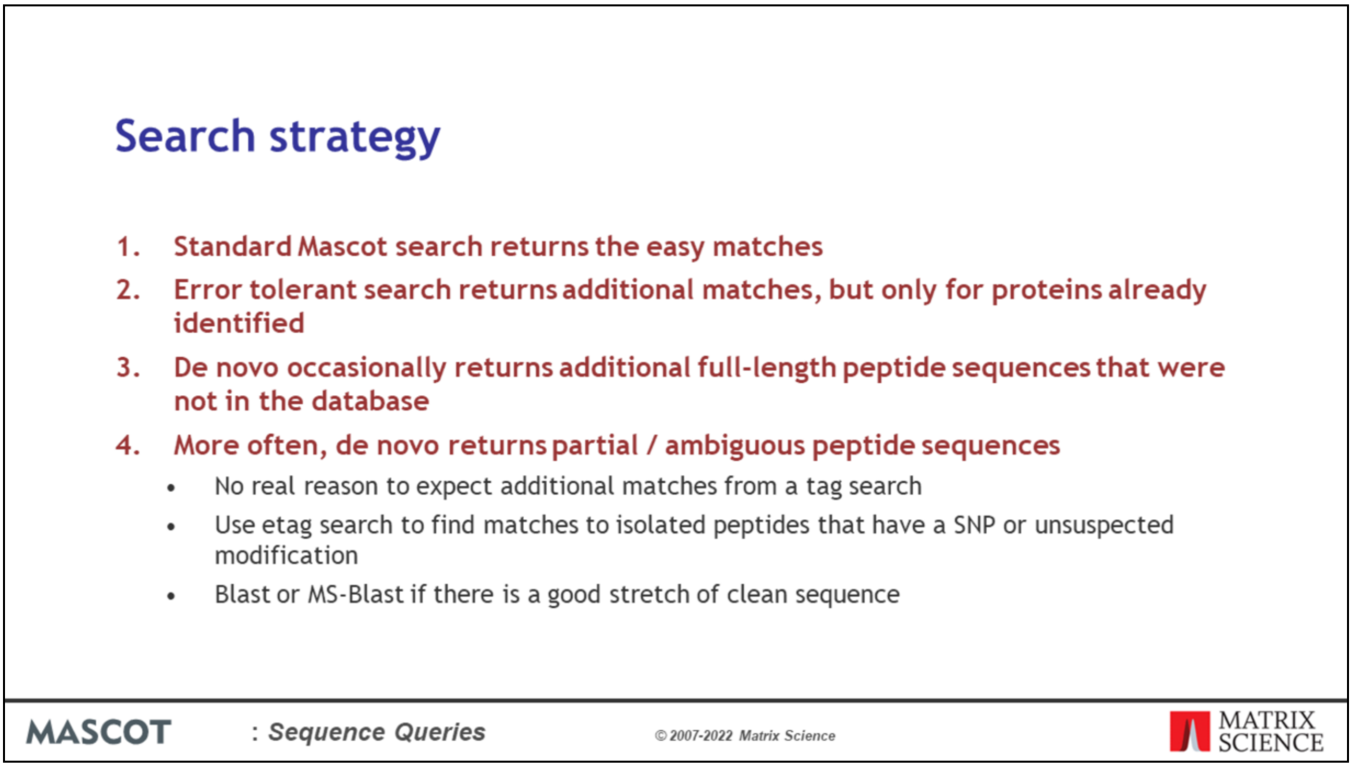
Score	Mr(calc)	Delta	Sequence
117.8	1450.6477	56.7105	NWYSDADVPASAR
79.7	824.3817	682.9765	GNWYSAK
67.4	1421.6575	85.7007	GNWYTPGTIEER

We can see from the Peptide View that the best match was obtained by placing a modification delta of +57 Da on the N-term residue. This is almost certainly carbamidomethylation, which often derivatises amino groups. This was why the original database search failed to get a match and illustrates how the error tolerant tag can get a match when the modification is unsuspected or even unknown (not in the modifications list)



Header							
Mascot Di	2.8.0.0						
Instrumer	ESI-TRAP						
Enzyme	Trypsin/P						
Maximum	2						
Peptide M	0.3						
Peptide M	Da						
Fragment	0.3						
Fragment	Da						
Fixed Modifications							
Variable Modifications							
Denovo Sc							
PeakListIr	PeakListTi	Charge	ObsMOve	Score	EnzymeSp	Sequence	FullSequence
2	2: Sum of	2	1149.905	26	yes	HMGAGD	HMGAGDY[TG|SA][SG]D[WA|TR|Eq]PC[YPA|VTM|VEC|TTE|TDD|SPF|SED|PPH|PMC|NNC|MEA|HHG|SMI|FAI|DCI][TG|S
2	2: Sum of	2	1149.905	24	yes	H-GA[Aq]-	H[VGC|TTG|TSA|SGD|MGA]GA[Aq]qG[WC]CFHGGq[VV|TP][VR]
2	2: Sum of	2	1149.905	24	yes	HMGAGD	HMGAGDY[TG|SA][YP|ME|FI]D[YGC|VSH|PPE|HGE|HDA|PII]WY[TG|SA][KG]
2	2: Sum of	2	1149.905	23	yes	HMGAGD	HMGAGDY[TG|SA][SG][YC|HE]q[GC]WY[TG|SA][KG]
2	2: Sum of	2	1149.905	23	yes	HMGAGD	HMGAGDY[TG|SA][YP|ME|FI]D[YGC|VSH|PPE|HGE|HDA|PII][HD]P[TG|SA][KG]
2	2: Sum of	2	1149.905	23	yes	HMGAGD	HMGAGDYSA[YP|ME|FI]D[Cq]E[TN|Sq]P[TG|SA][KG]
2	2: Sum of	2	1149.905	23	yes	HMGAGD	HMGAGDY[TG|SA][SG]D[YFE|WRP|YYI]WY[TG|SA][KG]
2	2: Sum of	2	1149.905	23	yes	HMGAGD	HMGAGDY[TG|SA][SG]D[RPD|HCq|VRI]AWY[TG|SA][KG]
2	2: Sum of	2	1149.905	22	yes	HMGAGD	HMGAGDY[TG|SA][SG][WW]PC[YA|SF|PH|MC]P[TG|SA][KG]
2	2: Sum of	2	1149.905	22	yes	HMGAGD	HMGAGDY[TG|SA][TG|SA][HD]q[GC]WY[TG|SA][KG]
6	6: Sum of	2	693.713	55	yes	MV[PA]AT	MV[PA]ATE[WGC|TNM|SRC|NEC|SMq|FAq|DCq]P[PK]
6	6: Sum of	2	693.713	55	yes	MV[PA]AT	MV[PA]ATEA[SM|FA|DC]GP[PK]
6	6: Sum of	2	693.713	53	yes	MVAPATE	MVAPATEA[TA|GD]CP[PK]
6	6: Sum of	2	693.713	52	yes	MV[PA]AT	MV[PA]AT[Fq]AEP[PK]
6	6: Sum of	2	693.713	50	yes	MV[PA]AT	MV[PA]AT[SM|FA|DC]GAEP[PK]
6	6: Sum of	2	693.713	49	yes	-[PA]ATE	[VM|TE|DD][PA]ATEA[WW][PK]
6	6: Sum of	2	693.713	46	yes	MVAPCN	MVAPCNAETEP[PK]
6	6: Sum of	2	693.713	45	yes	MV[PA]-[	MV[PA][VVC|VTT|VSD|VMA|TPC|TEA|PGF|GED|DDA|TSI|MGI][NC]E[NHA|PPq|HGq]
6	6: Sum of	2	693.713	45	yes	MVAPCN-	MVAPCN[VM|TE|DD]AEP[PK]
6	6: Sum of	2	693.713	43	yes	MVAPAA-	MVAPAA[YC|HE]qAA[YK]
7	7: Sum of	2	403.13	55	yes	#NAME?	[TG|SA]ASINSR
7	7: Sum of	2	403.13	40	yes	#NAME?	[TG|SA]AEANSR

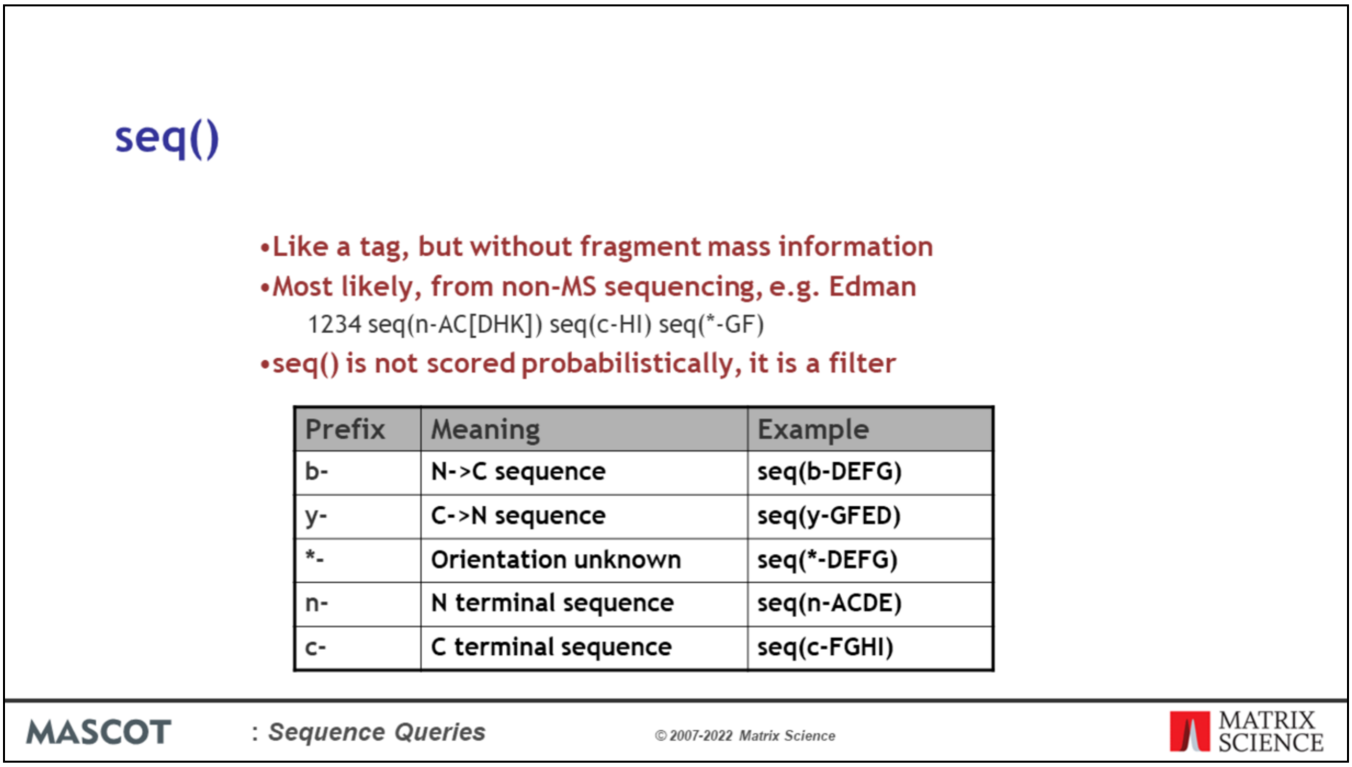
You can export the de novo solutions to a csv file that can be opened in Excel. From there it is easy to sort them by score to determine which ones are worth following up.



Search strategy

1. Standard Mascot search returns the easy matches
2. Error tolerant search returns additional matches, but only for proteins already identified
3. De novo occasionally returns additional full-length peptide sequences that were not in the database
4. More often, de novo returns partial / ambiguous peptide sequences
   - No real reason to expect additional matches from a tag search
   - Use etag search to find matches to isolated peptides that have a SNP or unsuspected modification
   - Blast or MS-Blast if there is a good stretch of clean sequence

If you want to get as many identifications as possible, as efficiently as possible, you might use a strategy similar to this.



seq()

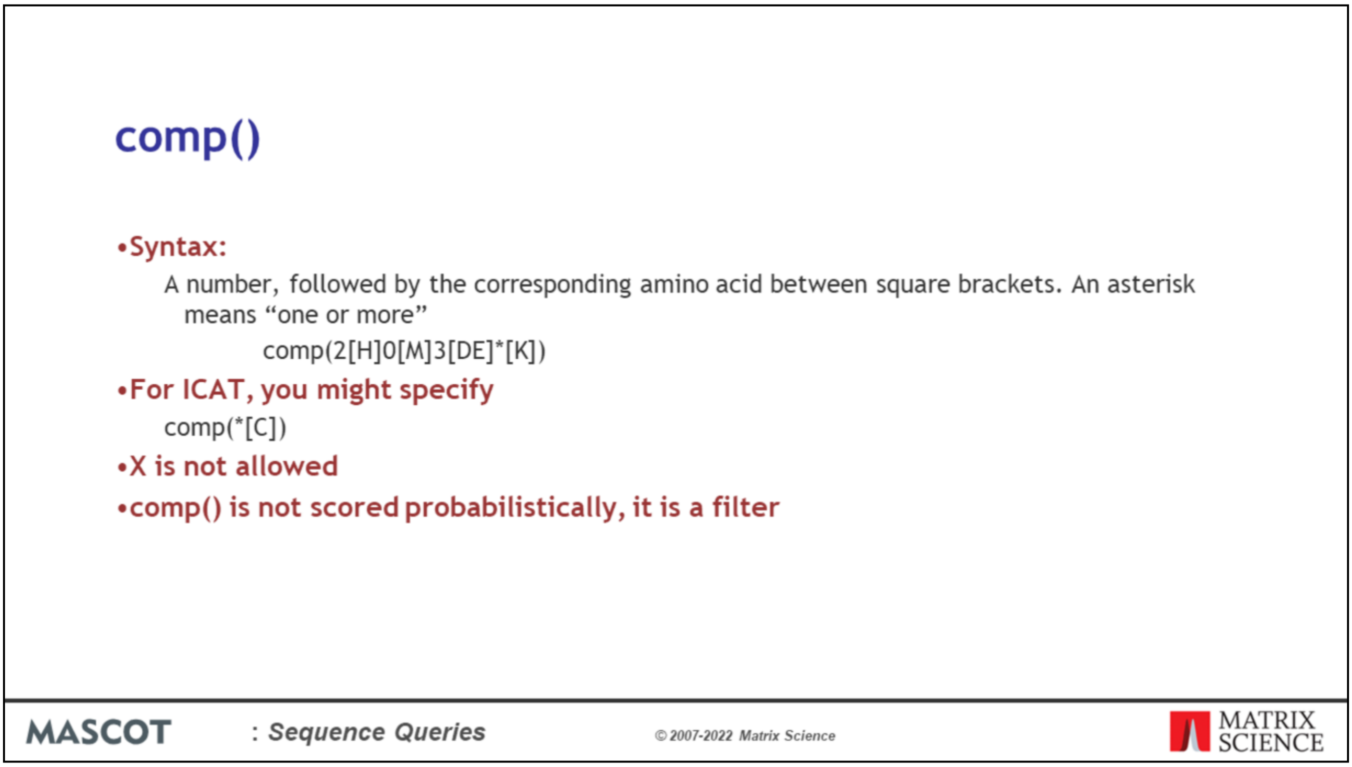
- Like a tag, but without fragment mass information
- Most likely, from non-MS sequencing, e.g. Edman

  1234 seq(n-AC[DHK]) seq(c-HI) seq(*-GF)
- seq() is not scored probabilistically, it is a filter

Prefix	Meaning	Example
b-	N->C sequence	seq(b-DEFG)
y-	C->N sequence	seq(y-GFED)
*-	Orientation unknown	seq(*-DEFG)
n-	N terminal sequence	seq(n-ACDE)
c-	C terminal sequence	seq(c-FGHI)

Besides tag and etag, Mascot supports a number of other sequence qualifiers. One of these is seq()

Note that seq() is a filter. It must be correct or there will be no match



comp()

- **Syntax:**

  A number, followed by the corresponding amino acid between square brackets. An asterisk means “one or more”

  comp(2[H]0[M]3[DE]*[K])

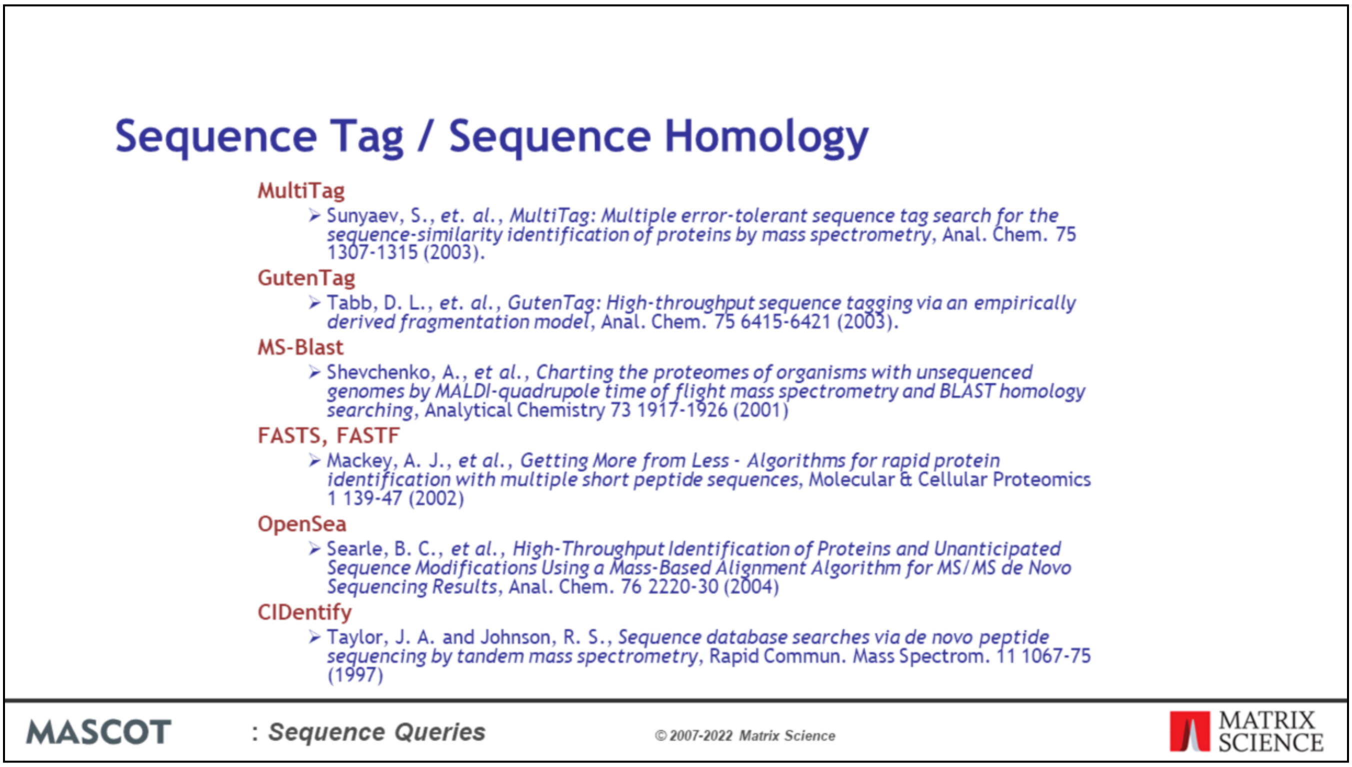
- **For ICAT, you might specify**

  comp(*[C])

- **X is not allowed**
- **comp() is not scored probabilistically, it is a filter**

The other important one is comp(). This would be useful in an ICAT search.

Note that comp() is a filter. It must be correct or there will be no match



Sequence Tag / Sequence Homology

MultiTag

- Sunyaev, S., *et. al.*, *MultiTag: Multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry*, Anal. Chem. 75 1307-1315 (2003).

GutenTag

- Tabb, D. L., *et. al.*, *GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model*, Anal. Chem. 75 6415-6421 (2003).

MS-Blast

- Shevchenko, A., *et al.*, *Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time of flight mass spectrometry and BLAST homology searching*, Analytical Chemistry 73 1917-1926 (2001)

FASTS, FASTF

- Mackey, A. J., *et al.*, *Getting More from Less - Algorithms for rapid protein identification with multiple short peptide sequences*, Molecular & Cellular Proteomics 1 139-47 (2002)

OpenSea

- Searle, B. C., *et al.*, *High-Throughput Identification of Proteins and Unanticipated Sequence Modifications Using a Mass-Based Alignment Algorithm for MS/MS de Novo Sequencing Results*, Anal. Chem. 76 2220-30 (2004)

CIDentify

- Taylor, J. A. and Johnson, R. S., *Sequence database searches via de novo peptide sequencing by tandem mass spectrometry*, Rapid Commun. Mass Spectrom. 11 1067-75 (1997)

MASCOT : *Sequence Queries* © 2007-2022 Matrix Science MATRIX SCIENCE

As always, there is more information in the Mascot help pages. These references are a good starting point if you are interested in learning more about the potential of combining mass and sequence information.