

# Sequence Queries

**MASCOT**



## 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

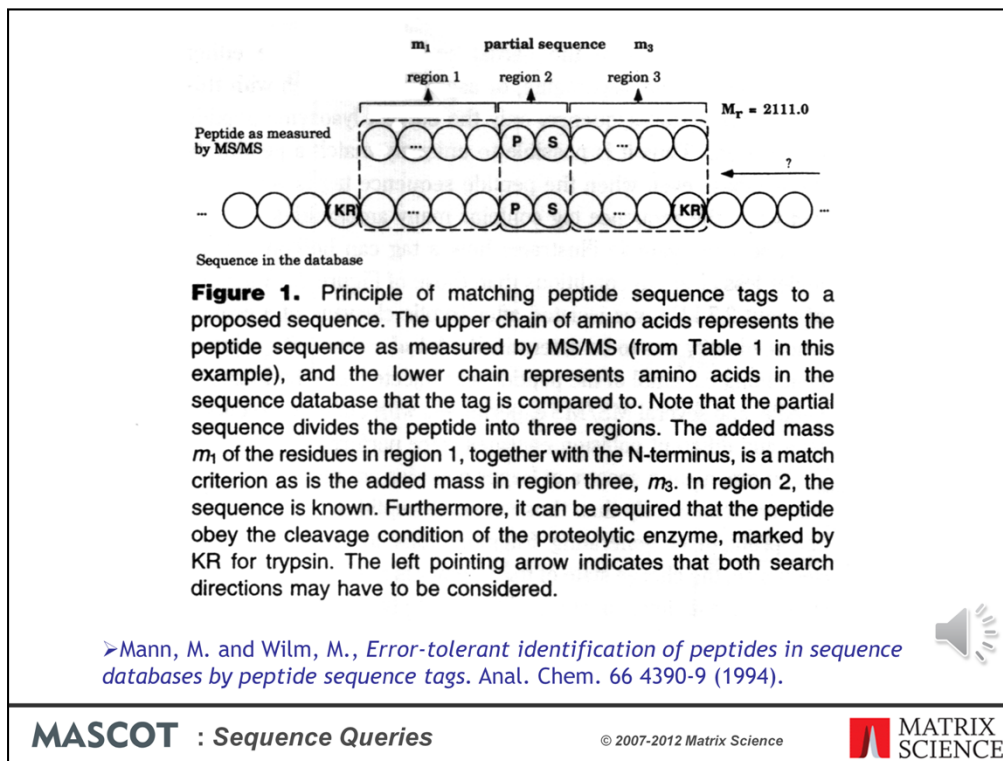
Mass values combined with amino acid sequence or composition data

### 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



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

**MASCOT** : Sequence Queries

© 2007-2012 Matrix Science



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)



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

© 2007-2012 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 $\Delta m$

EITHER both fragment ion masses unchanged

OR both fragment ion masses shift by  $\Delta m$

### 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



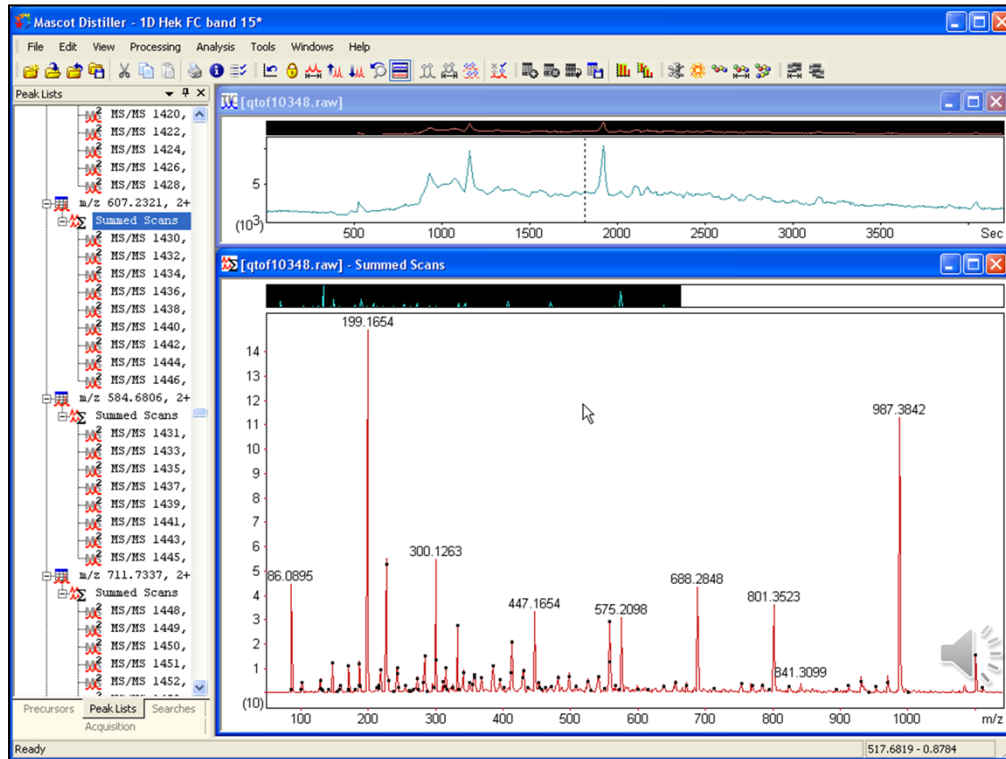
Tags can be entered with the high mass fragment on the left or the right. These two tags are identical

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.

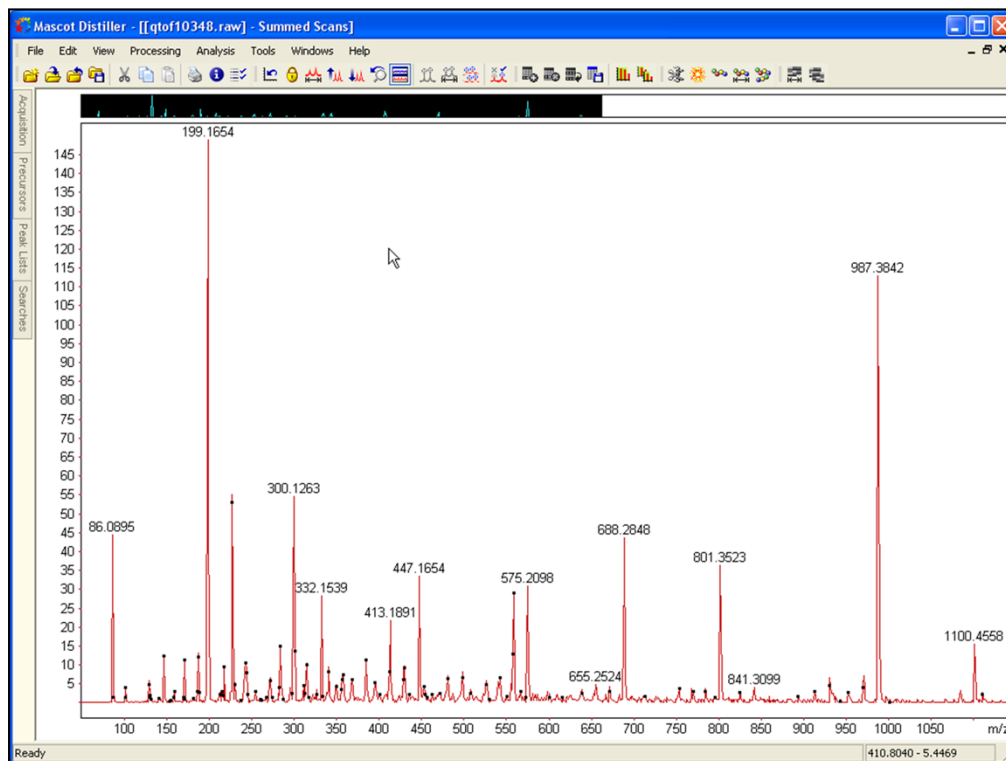
If one tag in a query is an etag then all the tags for that query are treated as etags, (not all tags in the search, just in the query)

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

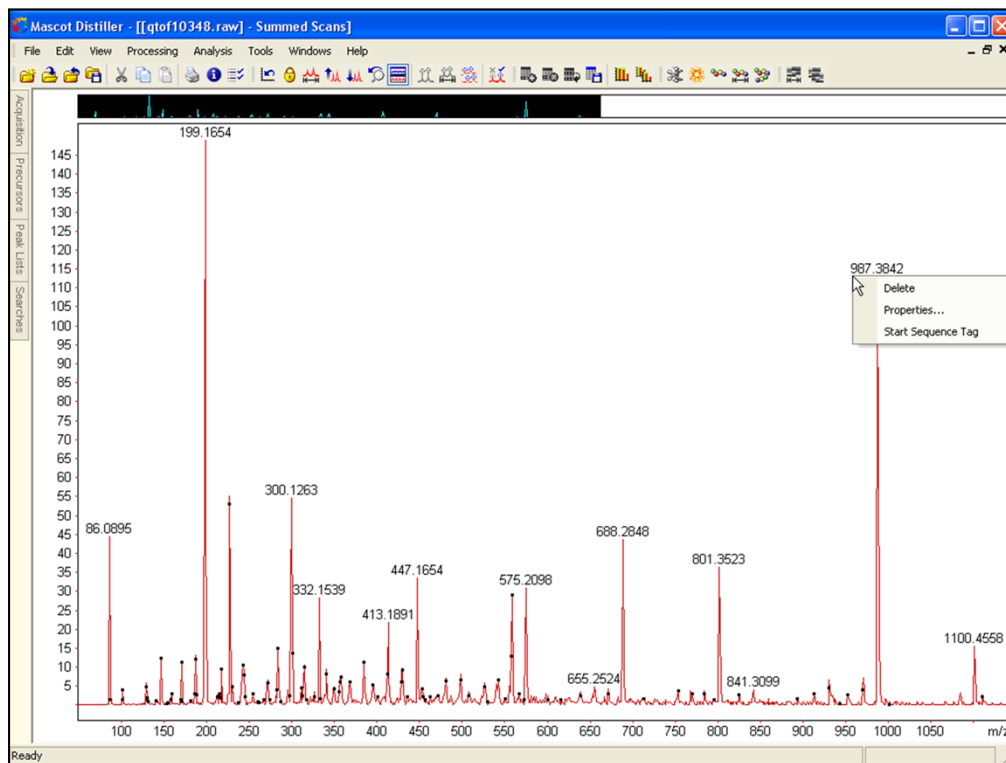




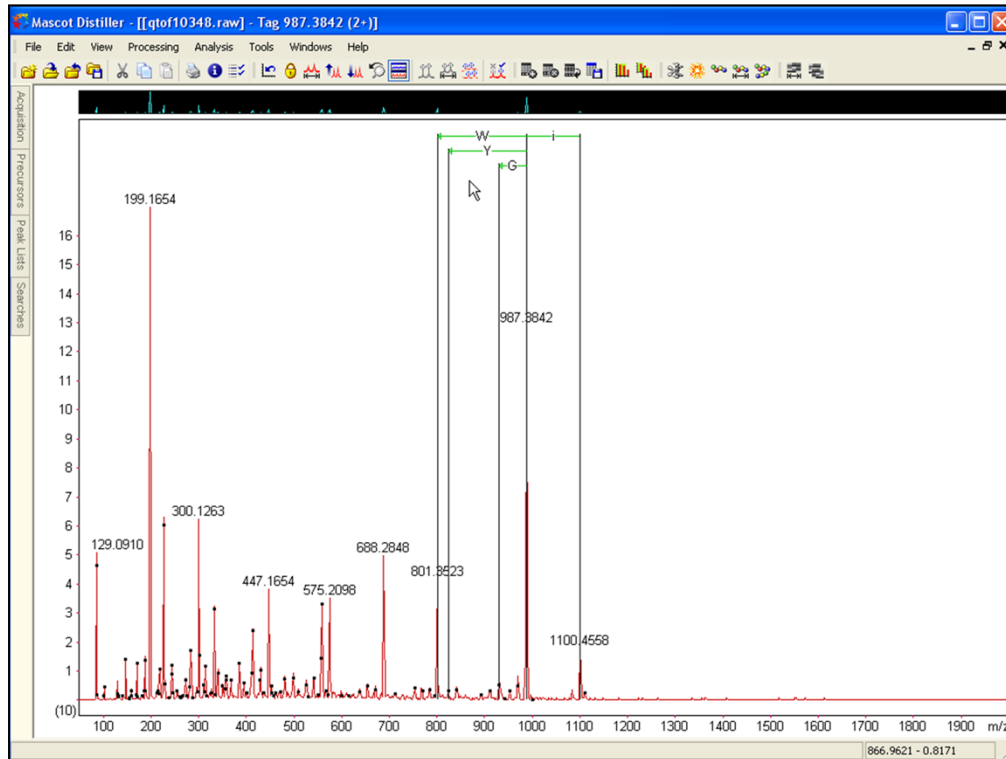
A lot of people call tags by inspection, using a calculator or a table of mass values. An alternative is to use Mascot Distiller. Here's how it works



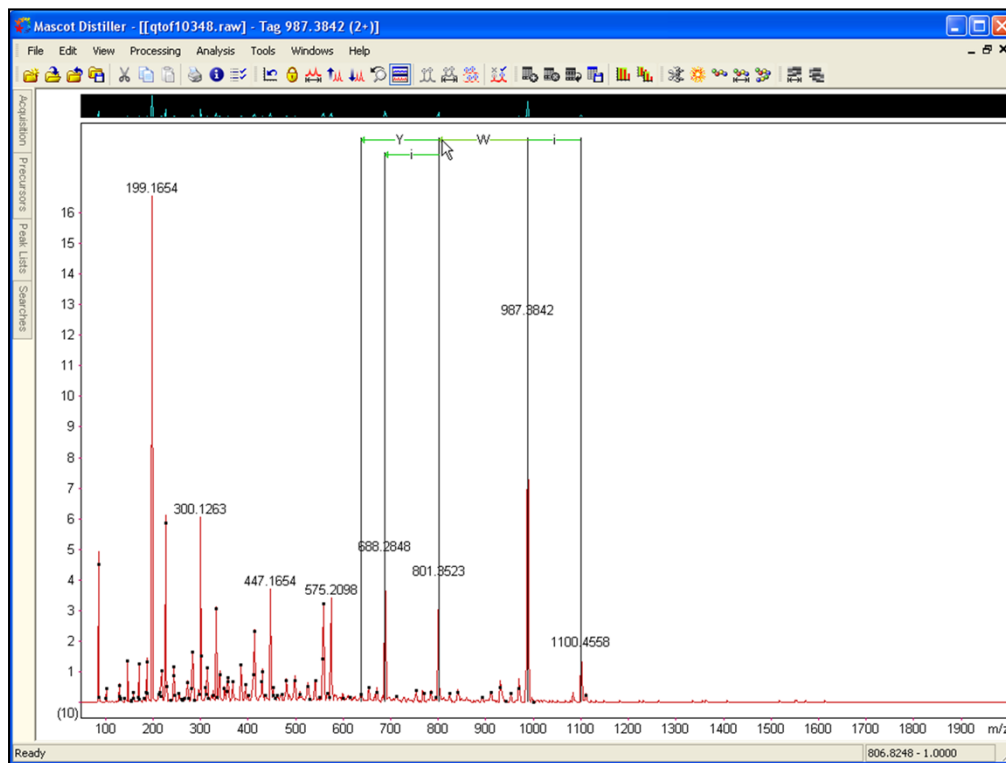
I'll maximise the window, so that we can see what's happening more clearly



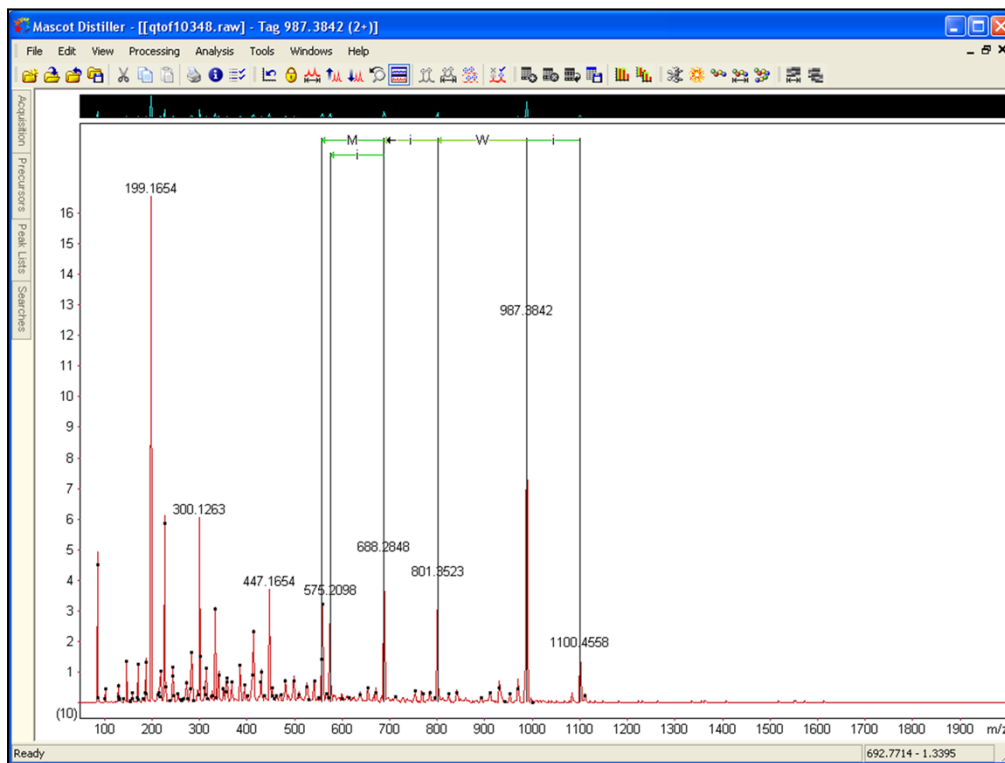
Choose a likely looking peak, such as this one at 987, right click the label, and choose 'Start sequence tag'



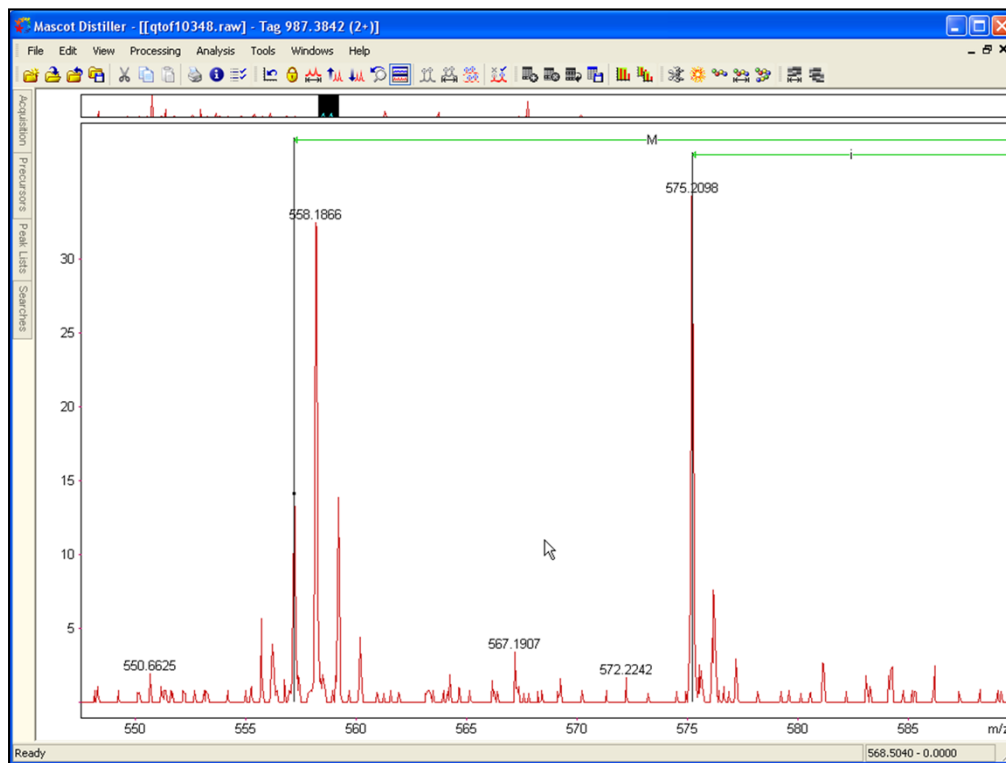
Distances to nearby peaks that correspond to amino acid residue masses are labelled. You have to choose one and click on the arrow head to select it. To keep things simple, I'm going to choose the most intense peak each time. In this case, it is the peak at 801, giving us a tryptophan



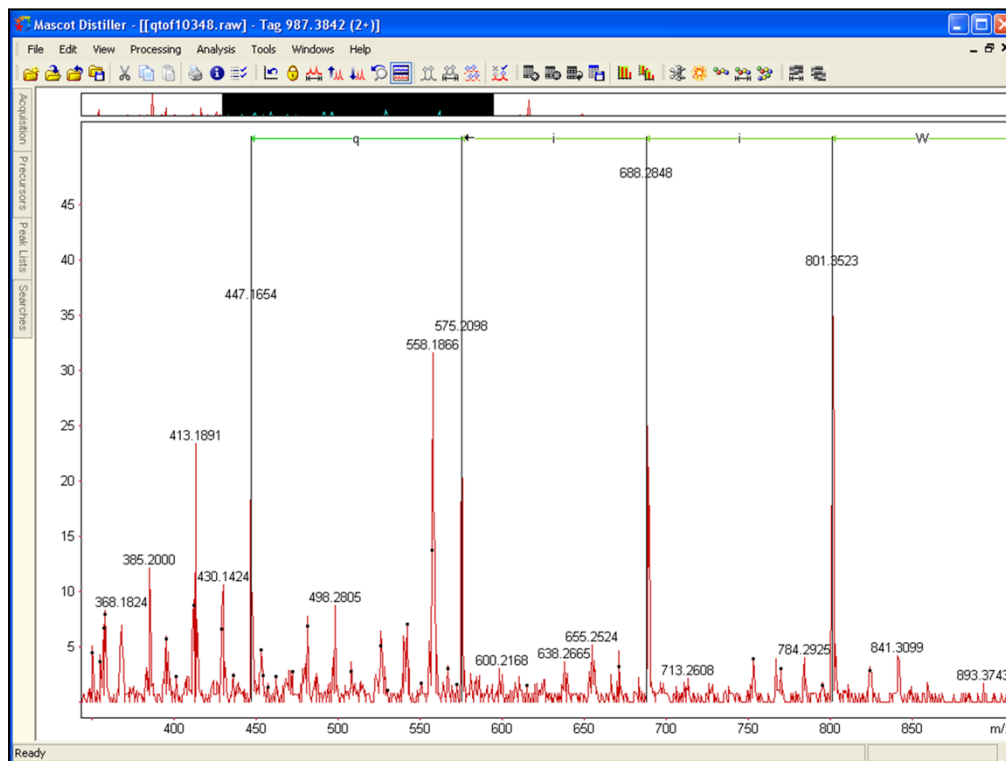
Then the lower case i for leucine or isoleucine



Zoom in to get a clearer view

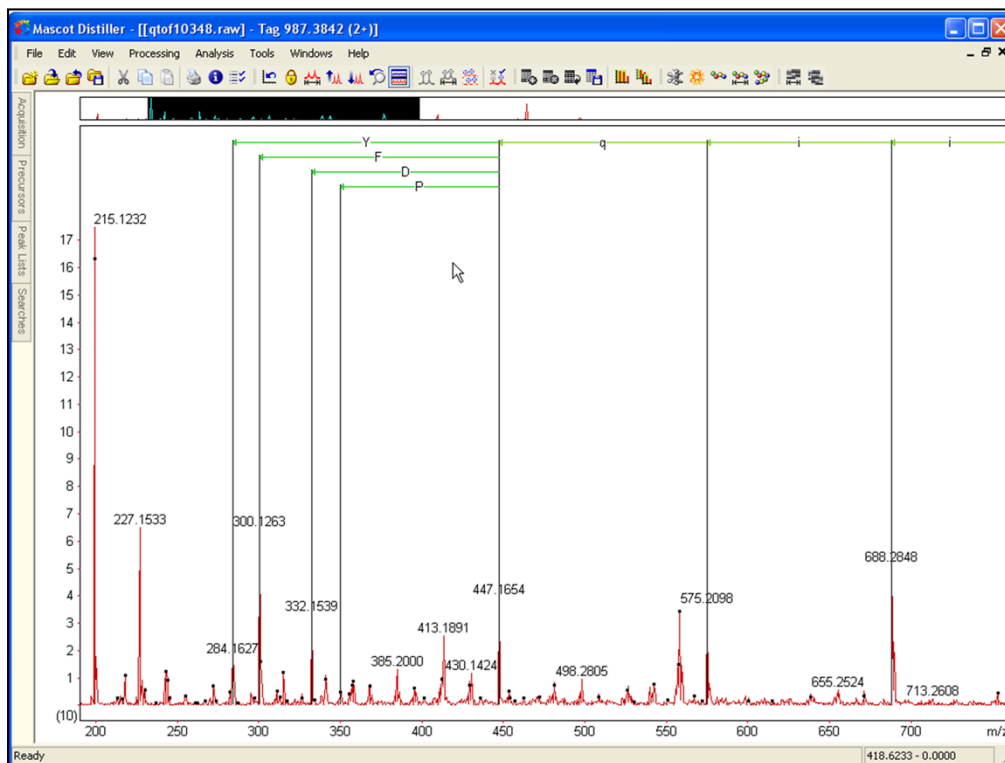


Let's try another i

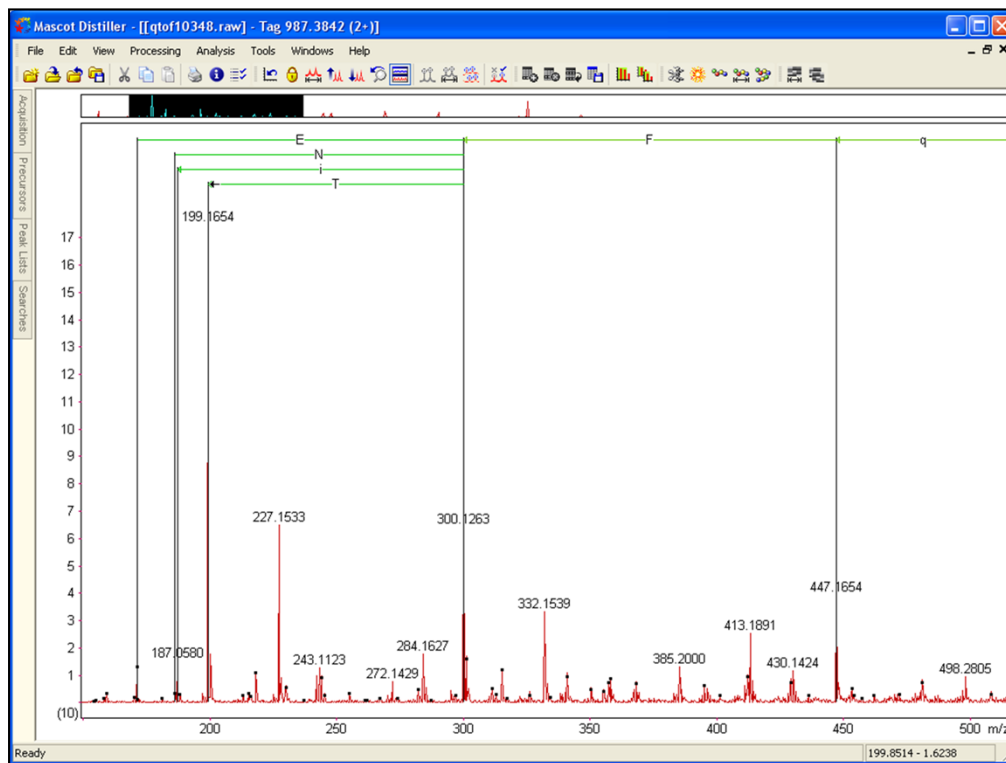


Only one option, q for glutamine of lysine

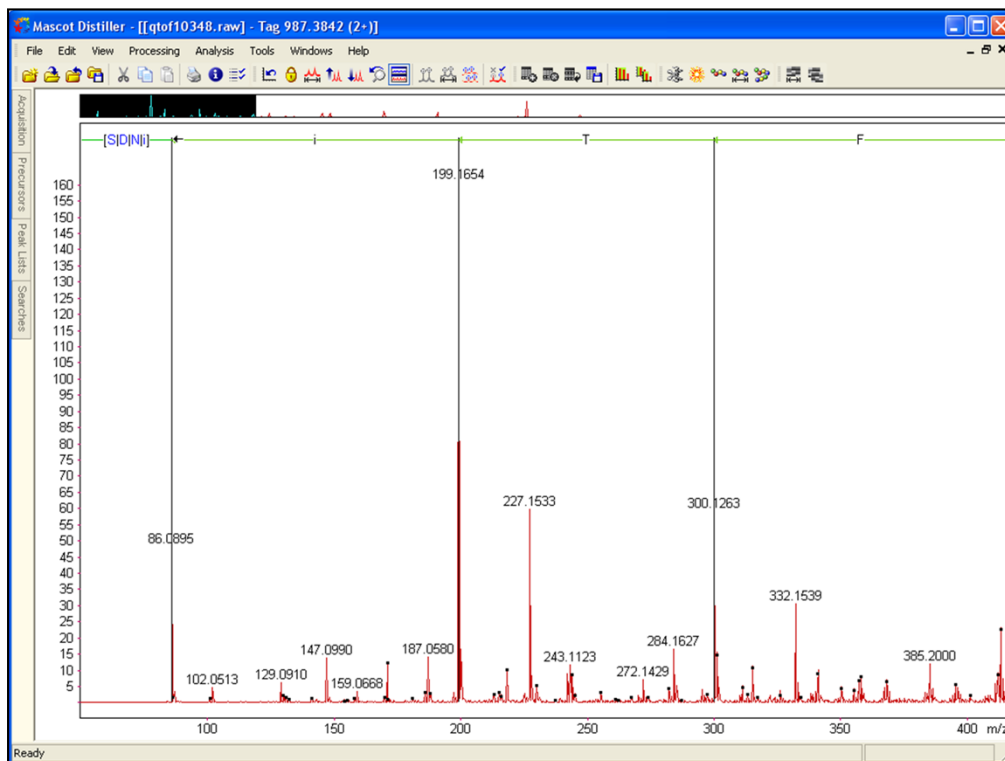




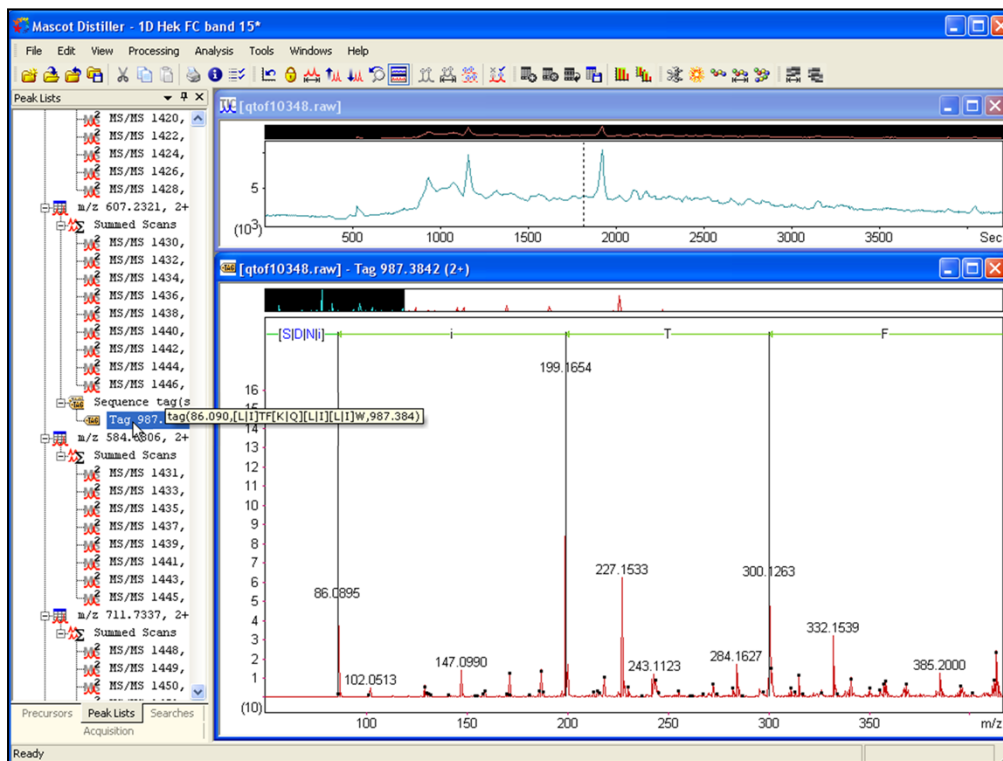
Complicated. Maybe the F for phenylalanine



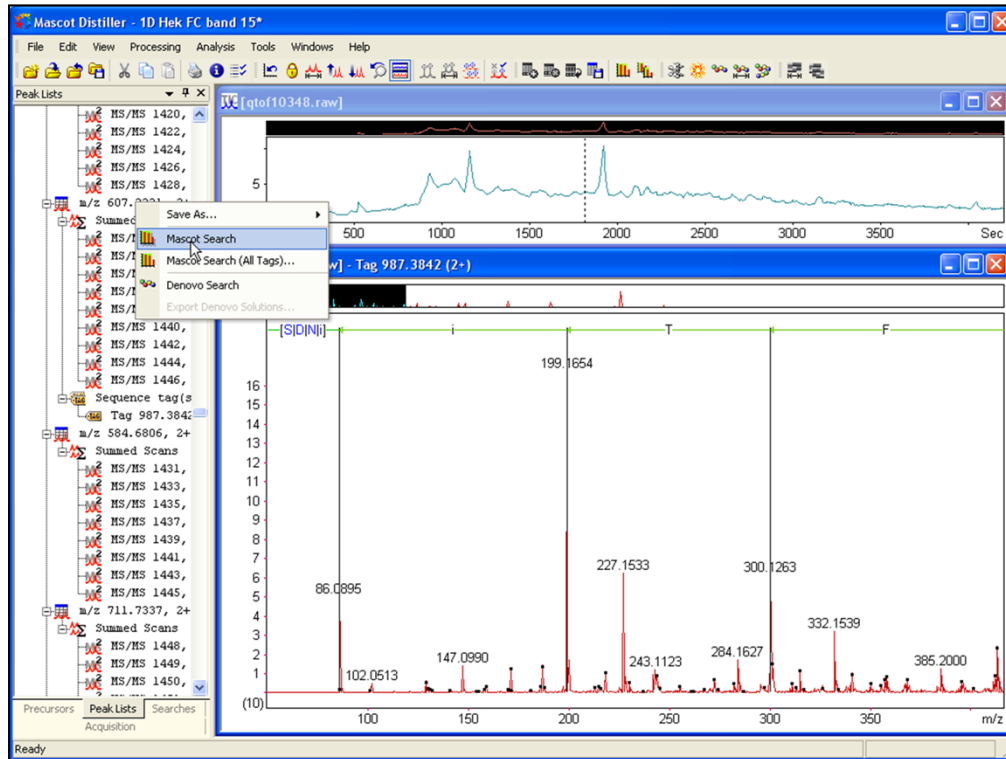
Then a threonine



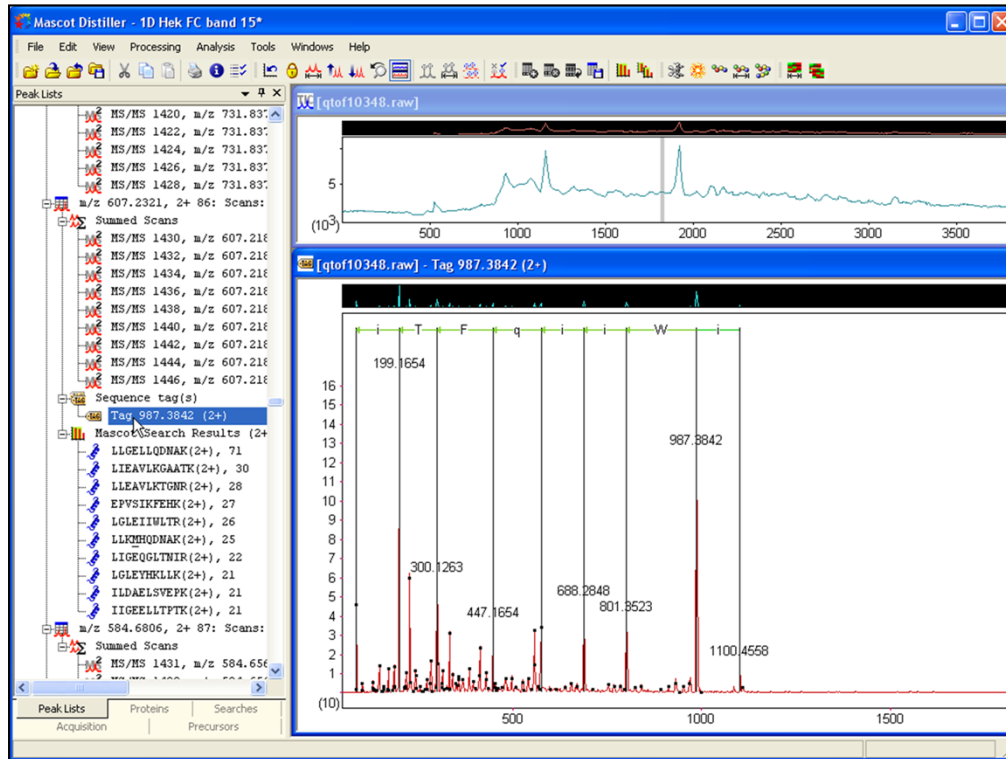
And another leucine or isoleucine. This takes us to the lowest mass peak of the spectrum. We then see a choice of residues to connect our tag to the terminus of the peptide. This is because the final mass difference depends on which ions series we happen to be looking at, which is still unknown, and there are several possibilities that fit in this case. So, we'll stop at the leucine / isoleucine



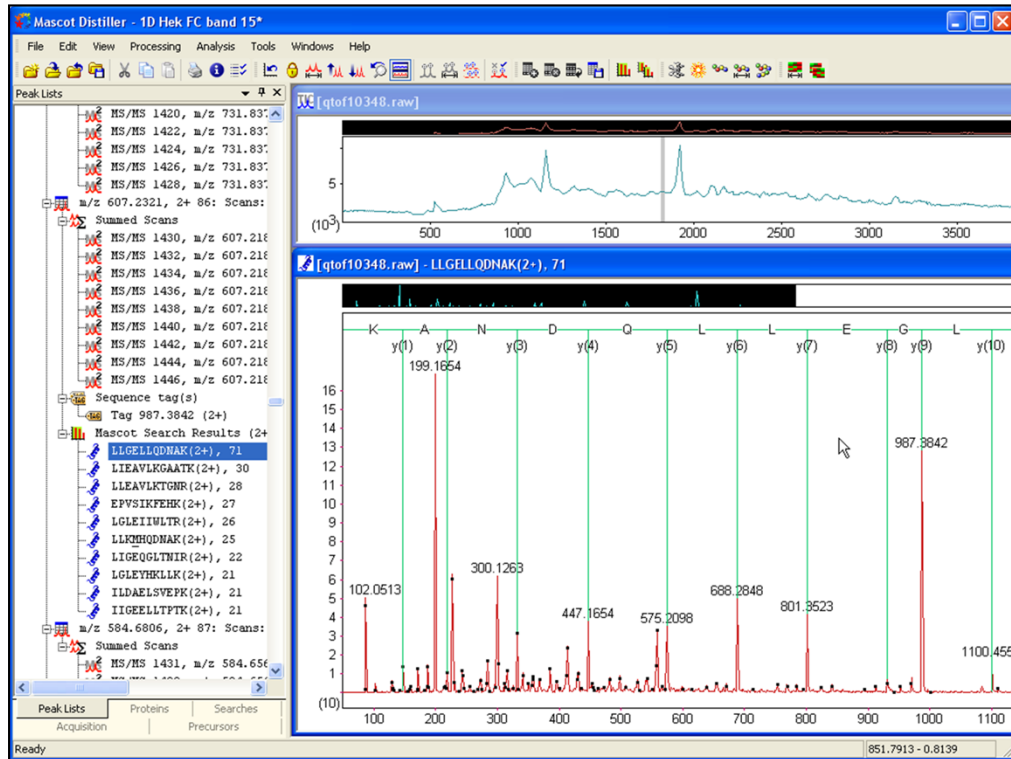
Restore the other windows and we can see our new sequence tag on the peak lists tree. But, is it correct? Lets see whether we can find a match for this spectrum in the database.



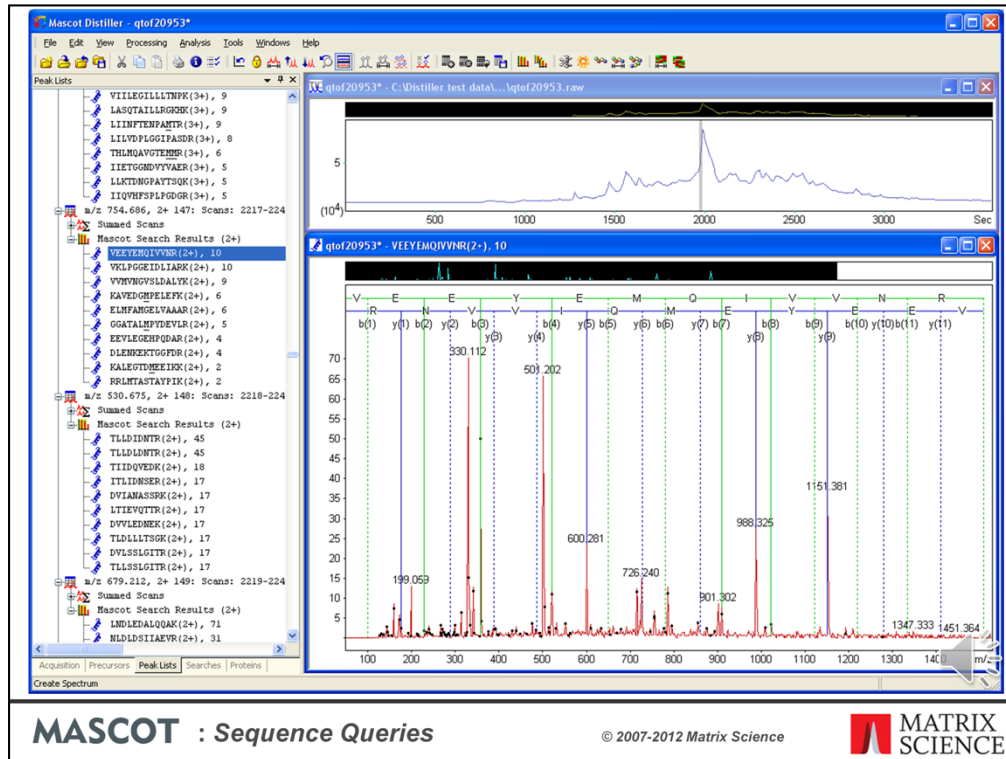
If we right click the peak list and perform a Mascot search



We get a good match with a score of 71. Compare the two sequences and we can see that we didn't do very well. We were in y ions, so the two sequences are running in opposite directions. We have LLQ correct, but not much else. At the beginning, where we chose W for the gap between 801 and 987



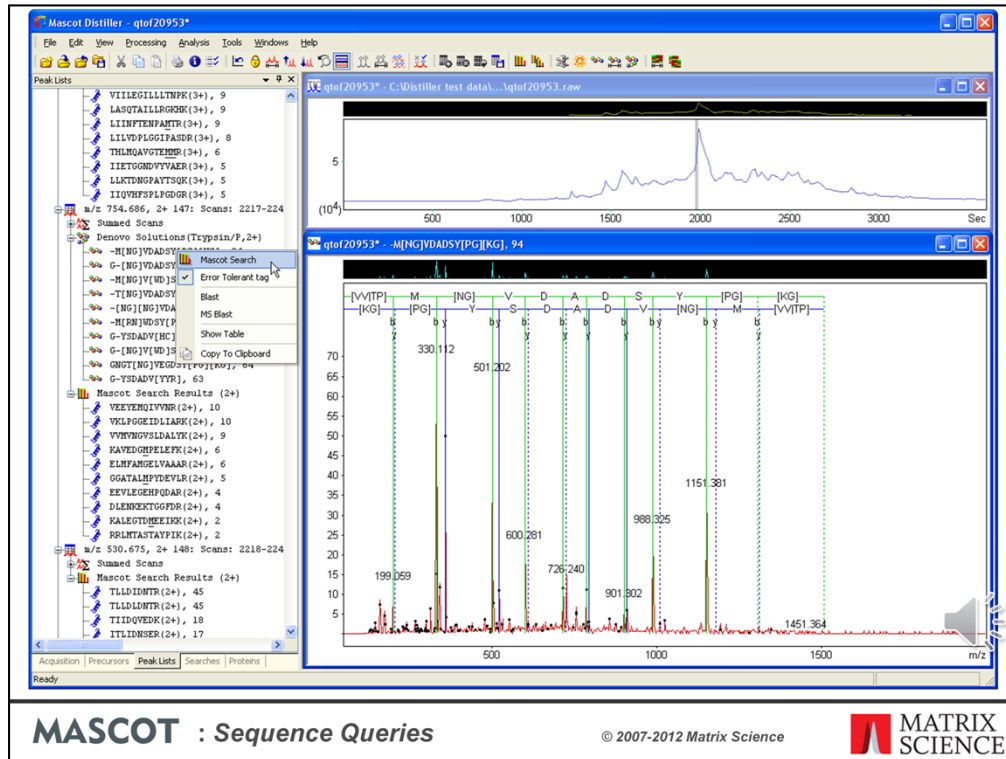
We can see from the database sequence that one of the small peaks in between was actually a y ion, making the sequence EG, not W. Calling a correct sequence tag is not nearly as easy as you might expect.



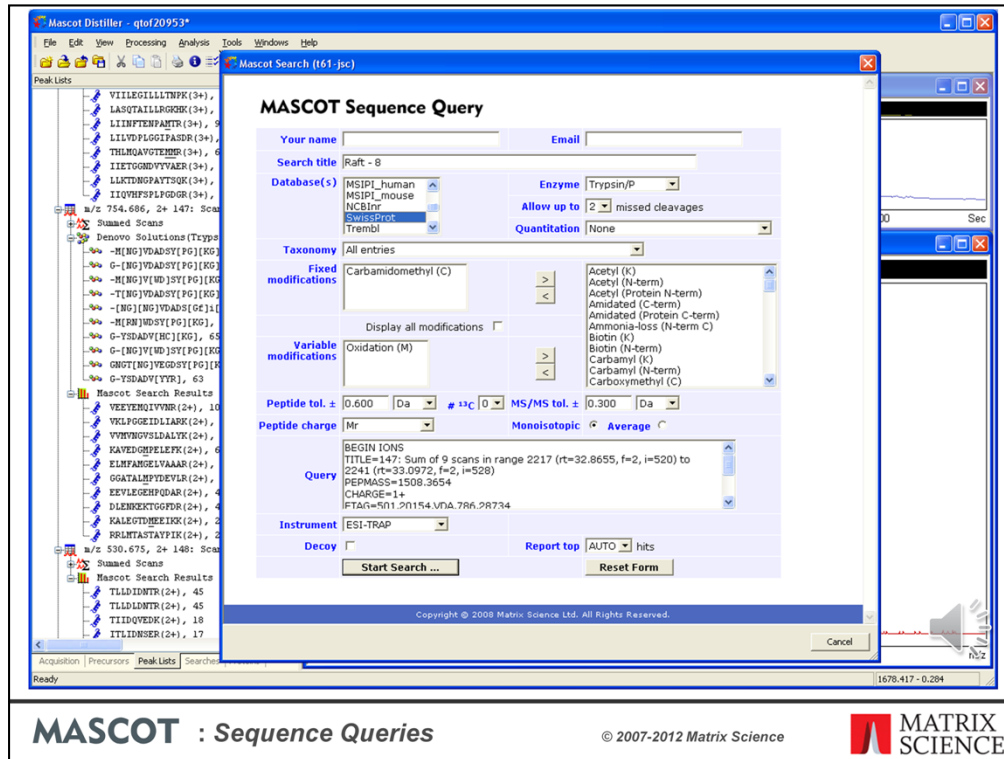
So, an alternative is to automate the process entirely by using the de novo algorithm. Here's a nice spectrum where the Mascot database search has failed to find a match. If we right click the peak list and choose de novo ...



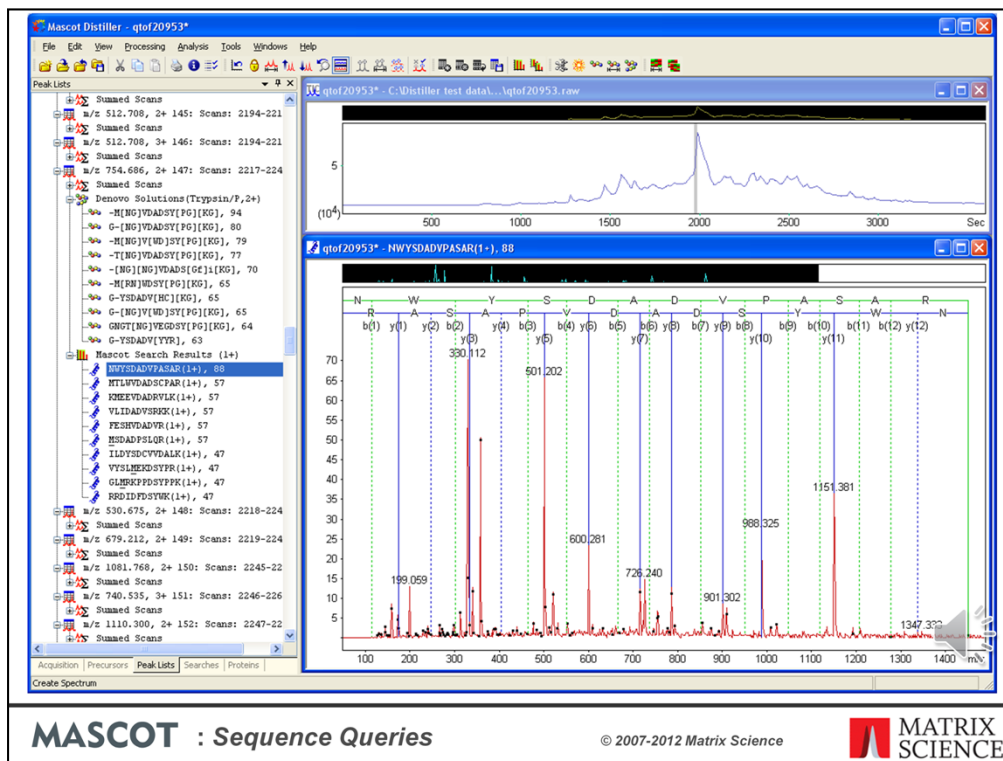




Right click the solution and choose Mascot search from the context menu. Note that we have already toggled the tag type to error tolerant



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



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



Mascot Search Results: Peptide View - Mozilla Firefox

http://r61-jsc.mascot/cgi/peptide\_view.pl?file=...&data=20100421.F01623H.dat&query=1&ht=1&index=PP61\_HUMAN&ip=1&section=5&save\_thre...

**Mascot Search Results**

**Peptide View**

MS/MS Fragmentation of **NWYSDADVPASAR**  
 Found in **PP61\_HUMAN**, Alkaline phosphatase, placental type OS=Homo sapiens GN=ALPP PE=1 SV=2

Match to Query 1: 1507.358124 from(1508.365400,1+) etag(501.20154,VDA,786.28734) etag(786.28734,DSY,1151.38063) etag(600.28103,DAD,901.30219) etag(794.22814,ADS,521.15782)  
 Title: 147: Sum of 9 scans in range 2217 (r=32.8655, f=2, i=520) to 2241 (r=33.0972, f=2, i=528)

Monoisotopic mass of neutral peptide M<sub>r</sub>(calc): 1450.6477  
 Fixed modifications: Carbamidomethyl (C) (apply to specified residues or termini only)  
 Unsuspected modification: 56.7104 Da, located in the region N-term to W2  
 Ions Score: 88 Expect: 2e-005 ([help](#))

#	b	b <sup>+</sup>	b <sup>0</sup>	Seq	y	y <sup>+</sup>	y <sup>0</sup>	#
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

Find:     ☐ Match case

Done

**MASCOT : Sequence Queries** © 2007-2012 Matrix Science **MATRIX SCIENCE**

The match was obtained by placing a modification delta of +57 Da on the N-term residue. This is almost certainly carbamidomethylation, which can derivatise amino groups if the conditions aren't right. This was why the original database search failed to get a match and this is why the error tolerant tag is so useful

## 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 come up with 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)

### CIIdentify

- 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)



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.