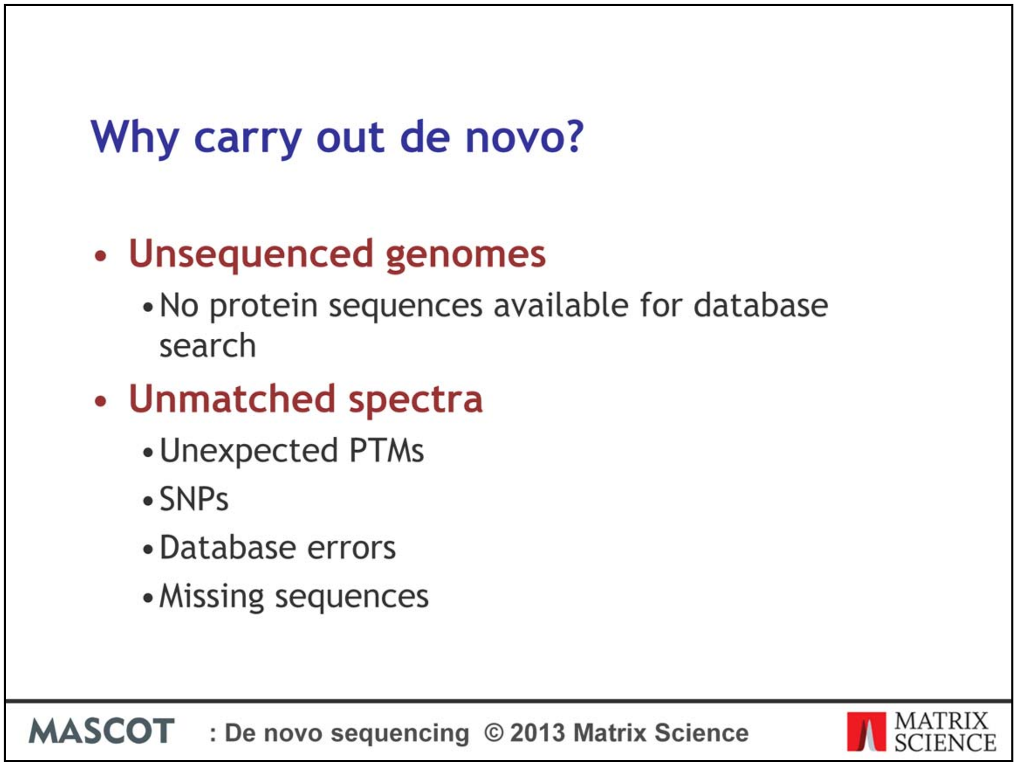
De novo sequencing in Mascot Distiller



Why carry out de novo?

- Unsequenced genomes
  - No protein sequences available for database search
- Unmatched spectra
  - Unexpected PTMs
  - SNPs
  - Database errors
  - Missing sequences

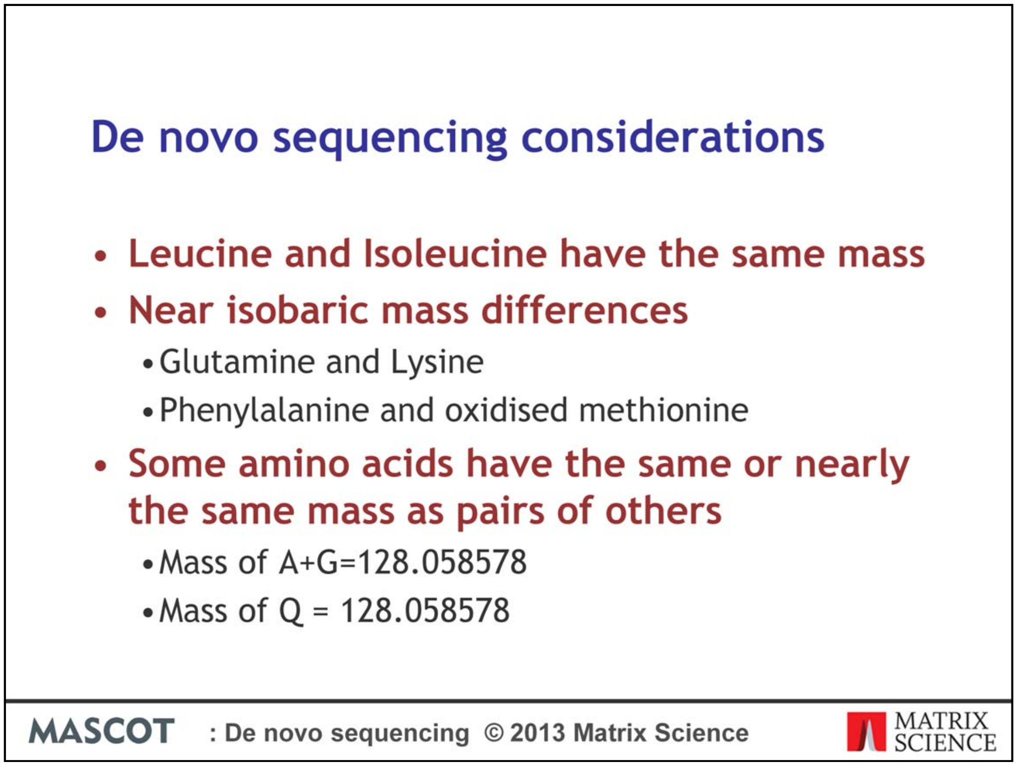
MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

A typical peptide identification protocol works by carrying out a search of the uninterpreted MS/MS spectra against a protein (or translated Nucleic acid) sequence database. This can be a very efficient and effective technique, but for this approach to work the matching peptide sequence must be available in the sequence database searched.

De novo sequencing interprets peptide sequence information directly from the MS/MS spectrum, with no sequence database required. This can be a useful approach if you are working on an organism which does not have a sequenced genome. If the sequences from a related species is available, you can then use the interpreted de novo sequence results in a database homology search, such as an MS-BLAST search.

From a standard MS/MS database search, you’ll often get a large number of unmatched spectra. Database searching of MS/MS spectra will often be caused to fail because of an unexpected post translational modification, or because of differences between the database sequence and the sequence in your dataset (often caused by sequence variations such Single Nucleotide Polymorphisms, sequencing errors in the database or sequences missing from the searched database). De novo sequencing is one approach that can help to match some of these spectra.

In general, we’d recommend carrying out a two pass error tolerant search for matching unexpected PTMs and SNPs before carrying out a de novo search as this is a much more efficient way to pick up these types of matches. If you are looking for SNPs or unexpected PTMs, only carry out a de novo search if you’re unable to get matches from a standard, first pass search – if you’re looking at endogenous peptides for example.



De novo sequencing considerations

- **Leucine and Isoleucine have the same mass**
- **Near isobaric mass differences**
  - Glutamine and Lysine
  - Phenylalanine and oxidised methionine
- **Some amino acids have the same or nearly the same mass as pairs of others**
  - Mass of A+G=128.058578
  - Mass of Q = 128.058578

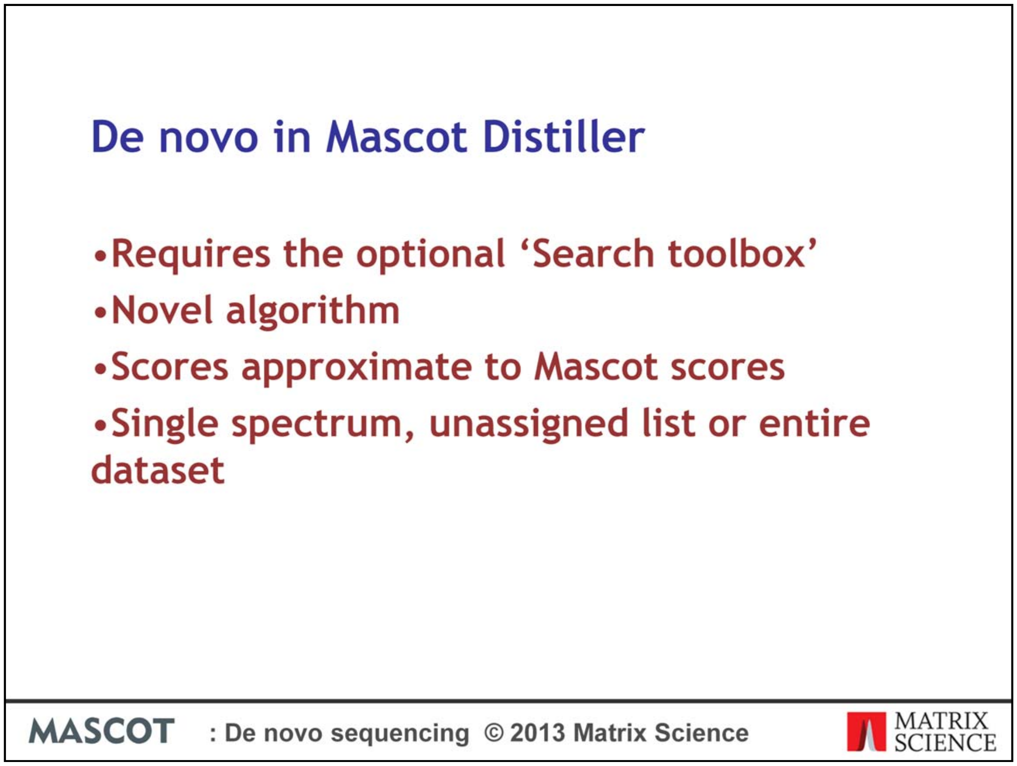
MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

When carrying out de novo sequencing there are some considerations which need to be taken which need to be taken into account and which can lead to errors and ambiguity in the generated sequence.

For example, Leucine and Isoleucine have the same mass and you cannot tell the difference between them in a standard MS/MS experiment.

Certain other amino acid pairs, such as glutamine and lysine, Phenylalanine and Oxidised methionine have near isobaric mass differences, and without a high accuracy instrument you’re not going to be able to tell the difference between them.

Some amino acids have nearly the same mass as pairs of others. For example the mass of alanine plus glycine is the same as that of Glutamine. You may have enough information in the spectrum to confidently tell the difference between the options, but often you won’t.



De novo in Mascot Distiller

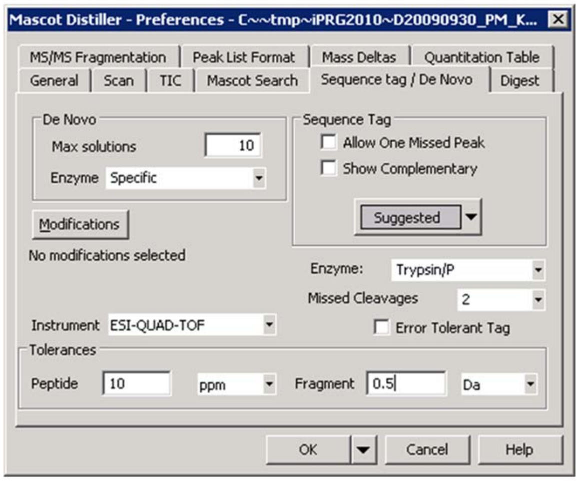
- Requires the optional ‘Search toolbox’
- Novel algorithm
- Scores approximate to Mascot scores
- Single spectrum, unassigned list or entire dataset

The Mascot Distiller search toolbox includes a powerful de novo sequencing module.

The score assigned to a de novo solution is similar to a Mascot score. In general, these scores will be higher than you expect to see in a Mascot database search, because the algorithm has selected the best matching sequence from all possible sequences, rather than the limited number of sequences found in any database. So, you should not judge the quality of the match by applying any rule of thumb or significance threshold to the score. However, if you get the same solution by de novo and by database search, using identical parameters, you should find the Mascot scores are very close.

You can de novo sequence a single spectrum, the ‘unassigned list’ from a Mascot search result or all the MS/MS spectra in the file

De novo settings



Mascot Distiller - Preferences - C~~tmp~iPRG2010~D20090930_PM_K...

De Novo

Max solutions: 10

Enzyme: Specific

No modifications selected

Instrument: ESI-QUAD-TOF

Sequence Tag

☐ Allow One Missed Peak

☐ Show Complementary

Enzyme: Trypsin/P

Missed Cleavages: 2

☐ Error Tolerant Tag

Tolerances

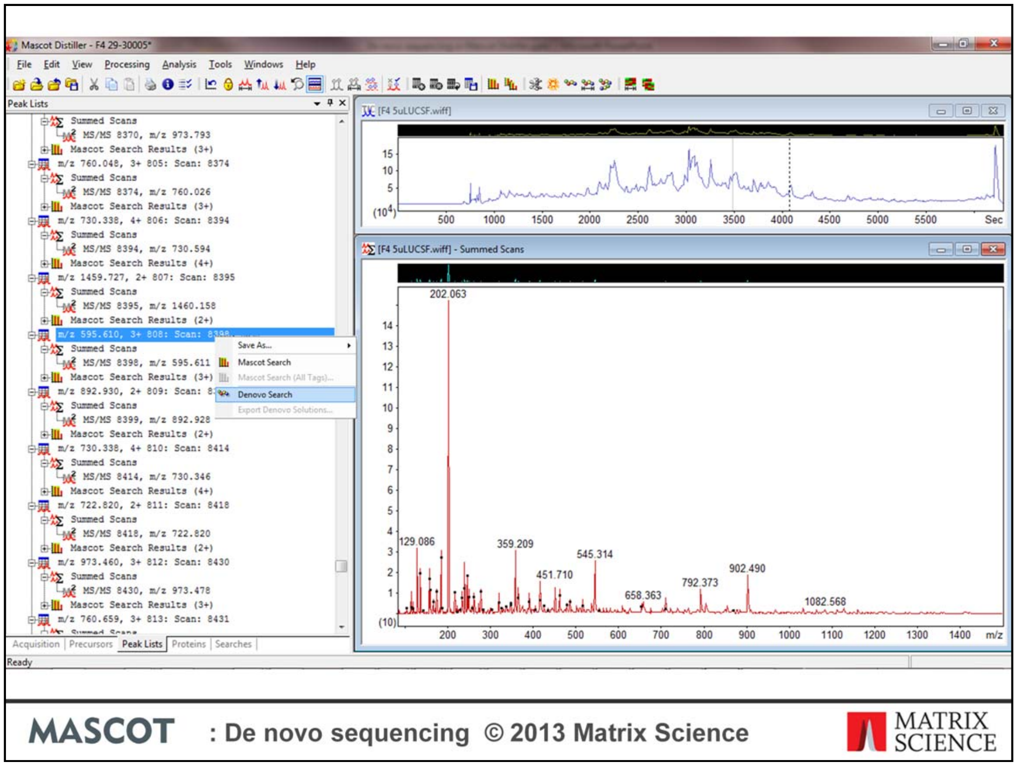
Peptide: 10 ppm

Fragment: 0.5 Da

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

As with a Mascot database search, there are various search settings which need to be set for a de novo search before carrying out the search. These can be found on the preferences dialog in Mascot Distiller, accessed from the ‘Tools’ menu.

The controls are very similar to those on the Mascot search settings form that you use when submitting a database search and it is important to set the correct de novo settings before initiating a de novo search, otherwise the default settings will be used which will probably not be appropriate for your dataset.

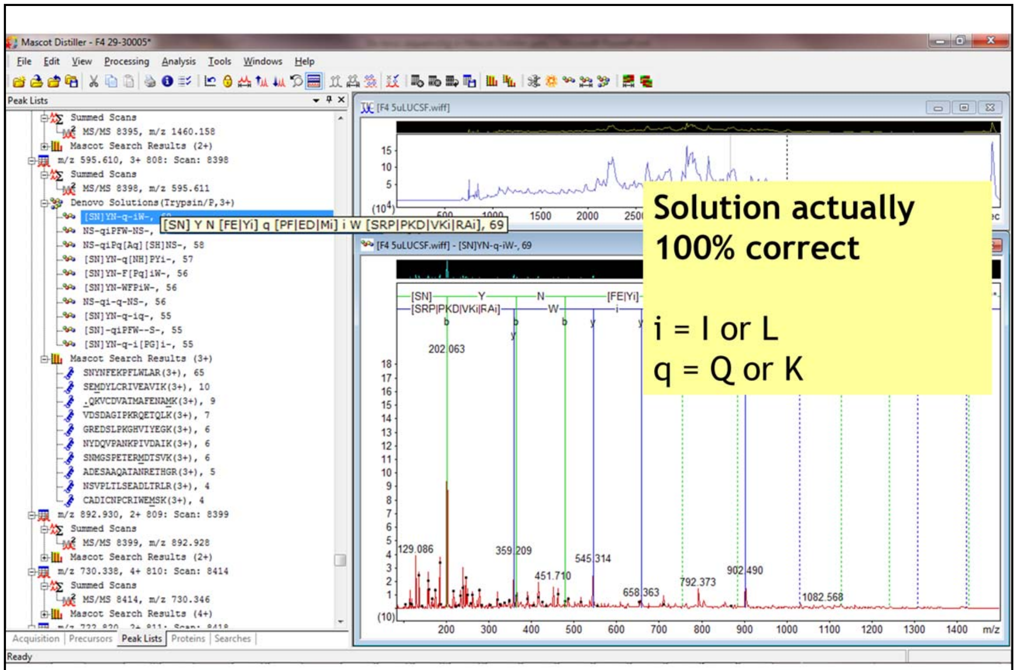


```
Summed Scans
  MS/MS 8370, m/z 973.793
Mascot Search Results (3+)
m/z 760.048, 3+ 805: Scan: 8374
  Summed Scans
    MS/MS 8374, m/z 760.026
  Mascot Search Results (3+)
m/z 730.338, 4+ 806: Scan: 8394
  Summed Scans
    MS/MS 8394, m/z 730.594
  Mascot Search Results (4+)
m/z 1459.727, 2+ 807: Scan: 8395
  Summed Scans
    MS/MS 8395, m/z 1460.158
  Mascot Search Results (2+)
m/z 595.610, 3+ 808: Scan: 8398
  Summed Scans
    MS/MS 8398, m/z 595.611
  Mascot Search Results (3+)
m/z 892.930, 2+ 809: Scan: 8
  Summed Scans
    MS/MS 8399, m/z 892.928
  Mascot Search Results (2+)
m/z 730.338, 4+ 810: Scan: 8414
  Summed Scans
    MS/MS 8414, m/z 730.346
  Mascot Search Results (4+)
m/z 722.820, 2+ 811: Scan: 8418
  Summed Scans
    MS/MS 8418, m/z 722.820
  Mascot Search Results (2+)
m/z 973.460, 3+ 812: Scan: 8430
  Summed Scans
    MS/MS 8430, m/z 973.478
  Mascot Search Results (3+)
m/z 760.659, 3+ 813: Scan: 8431
  Summed Scans
```

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

The starting point for de novo can be any MS/MS scan where Distiller has been used to create a peak list. Select the ‘Peak list’ tab in the dataset explorer in Distiller. Right click the peak list node and choose ‘de novo Search’, or choose the de novo button from the toolbar when a Summed Scans node is selected.

When the de novo search has completed, the top ten solutions are added to the ‘Peak lists’ tab.



```
Summed Scans
  MS/MS 8395, m/z 1460.158
Mascot Search Results (2+)
m/z 595.610, 3+ 808: Scan: 8398
  Summed Scans
    MS/MS 8398, m/z 595.611
  Denovo Solutions(Trypsin/P,3+)
    [SN]YN-q-iW-,
    NS-qiPFW-NS-,
    NS-qiPq[Aq][SH]NS-, 58
    [SN]YN-q[NH]PYi-, 57
    [SN]YN-F[Pq]iW-, 56
    [SN]YN-WFPiW-, 56
    NS-qi-q-NS-, 56
    [SN]YN-q-iq-, 55
    [SN]-qiPFW--S-, 55
    [SN]YN-q-i[PG]i-, 55
  Mascot Search Results (3+)
    SNYNFEKPFLWLAR(3+), 65
    SEMDYLCRIVEAVIK(3+), 10
    .QKVCDVATMAFENAMK(3+), 9
    VDSDAGIPKRQETQLK(3+), 7
    GREDSLPKGHVIYEGK(3+), 6
    NYDQVPANKPIVDAIK(3+), 6
    SNMGSPETERMDTSVK(3+), 6
    ADESAAQATANRETHGR(3+), 5
    NSVPLTLSEADLTRLR(3+), 4
    CADICNPCRIWEMSK(3+), 4
m/z 892.930, 2+ 809: Scan: 8399
  Summed Scans
    MS/MS 8399, m/z 892.928
  Mascot Search Results (2+)
m/z 730.338, 4+ 810: Scan: 8414
  Summed Scans
    MS/MS 8414, m/z 730.346
  Mascot Search Results (4+)
```

[SN] Y N [FE|Yi] q [PF|ED|Mi] i W [SRP|PKD|VKi|RAi], 69

Solution actually 100% correct

i = I or L

q = Q or K

MASCOT : De novo sequencing © 2013 Matrix Science

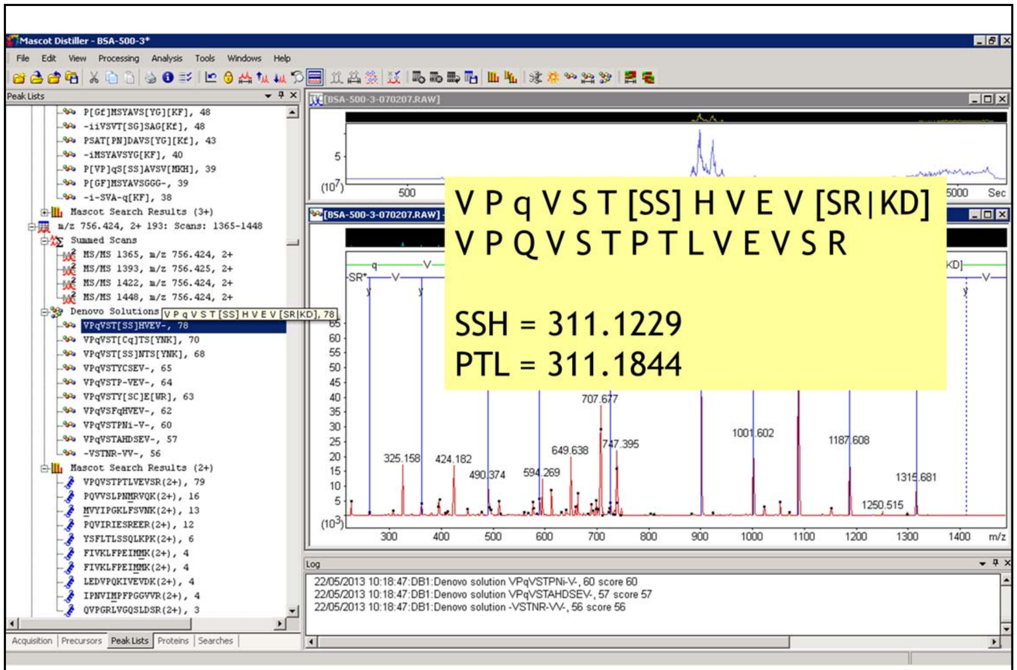
MATRIX SCIENCE

Let’s take a look at a couple of de novo results where we already have strong database matches in order to illustrate some of the differences between a de novo result and a Mascot database search result. Of course you wouldn’t normally carry out a de novo search of spectra where you have strong database matches as this would normally be a waste of time.

Good signal to noise and good mass accuracy are critical for successful de novo sequencing; much more so than in database searching. GIGO (garbage in - garbage out) is guaranteed.

In a de novo solution, i always represents I or L, and q represent Q or K when the mass tolerance does not allow these residues to be distinguished. However, K is assumed at the C terminus of a peptide when tryptic specificity applies. Also remember that phenylalanine may be oxidised methionine, and vice versa. Ambiguity is indicated by a dash in the sequence. The tooltip shows details of the ambiguity in square brackets, using pipe symbols to separate alternatives. Note that the order of any pairs and triplets is undefined, so that RAi could also be iAR.

Although at first glance, the example shown here looks very different to the Mascot database match, they are actually in perfectly agreement. Some uncertainty is unavoidable in de novo, because the search space is so very much larger.



```
P[Gf]MSYAVS[YG][KF], 48
-iiVSVT[SG]SAG[Kf], 48
PSAT[PN]DAVS[YG][Kf], 43
-iMSYAVSYG[KF], 40
P[VP]qS[SS]AVSV[MKH], 39
P[GF]MSYAVSGGG-, 39
-i-SVA-q[KF], 38
Mascot Search Results (3+)
m/z 756.424, 2+ 193: Scans: 1365-1448
Summed Scans
MS/MS 1365, m/z 756.424, 2+
MS/MS 1393, m/z 756.425, 2+
MS/MS 1422, m/z 756.424, 2+
MS/MS 1448, m/z 756.424, 2+
Denovo Solutions  V P q V S T [SS] H V E V [SR|KD], 78
VPqVST[SS]HVEV-, 78
VPqVST[Cq]TS[YNK], 70
VPqVST[SS]NTS[YNK], 68
VPqVSTYCSEV-, 65
VPqVSTP-VEV-, 64
VPqVSTY[SC]E[WR], 63
VPqVSFqHVEV-, 62
VPqVSTPNi-V-, 60
VPqVSTAHDSEV-, 57
-VSTNR-VV-, 56
Mascot Search Results (2+)
VPQVSTPTLVEVSR(2+), 79
PQVVSLPNMRVQK(2+), 16
MVYIPGKLFSVNK(2+), 13
PQVIRIESREER(2+), 12
YSFLTLSSQLKPK(2+), 6
FIVKLFPEIMMK(2+), 4
FIVKLFPEIMMK(2+), 4
LEDVPQKIVEVDK(2+), 4
IPNVIMPFPGGVVR(2+), 4
QVPGRLVGQSLDSR(2+), 3
```

```
22/05/2013 10:18:47:DB1:Denovo solution VPqVSTPNi-V-, 60 score 60
22/05/2013 10:18:47:DB1:Denovo solution VPqVSTAHDSEV-, 57 score 57
22/05/2013 10:18:47:DB1:Denovo solution -VSTNR-VV-, 56 score 56
```

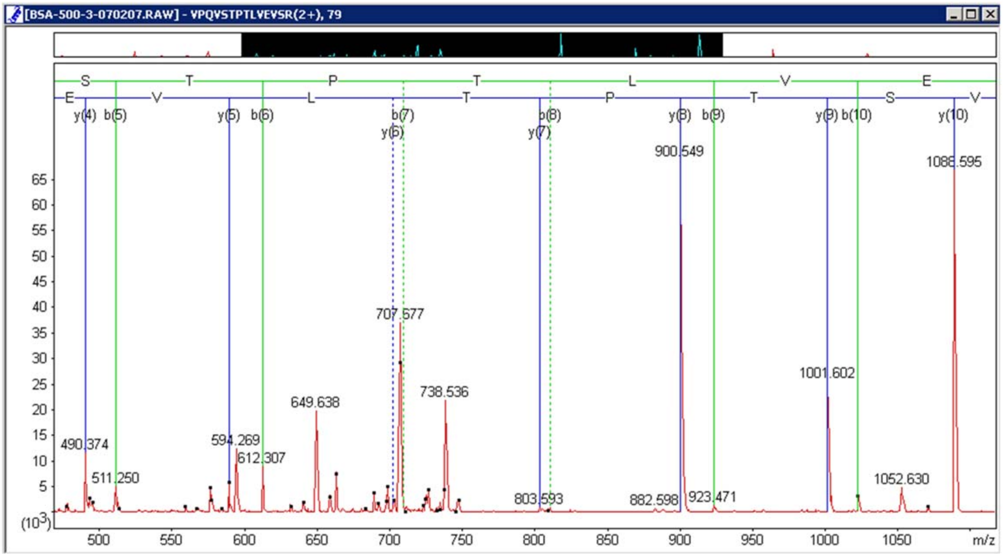
V P q V S T [SS] H V E V [SR|KD]
V P Q V S T P T L V E V S R

SSH = 311.1229
PTL = 311.1844

MASCOT : De novo sequencing © 2013 Matrix Science

MATRIX SCIENCE

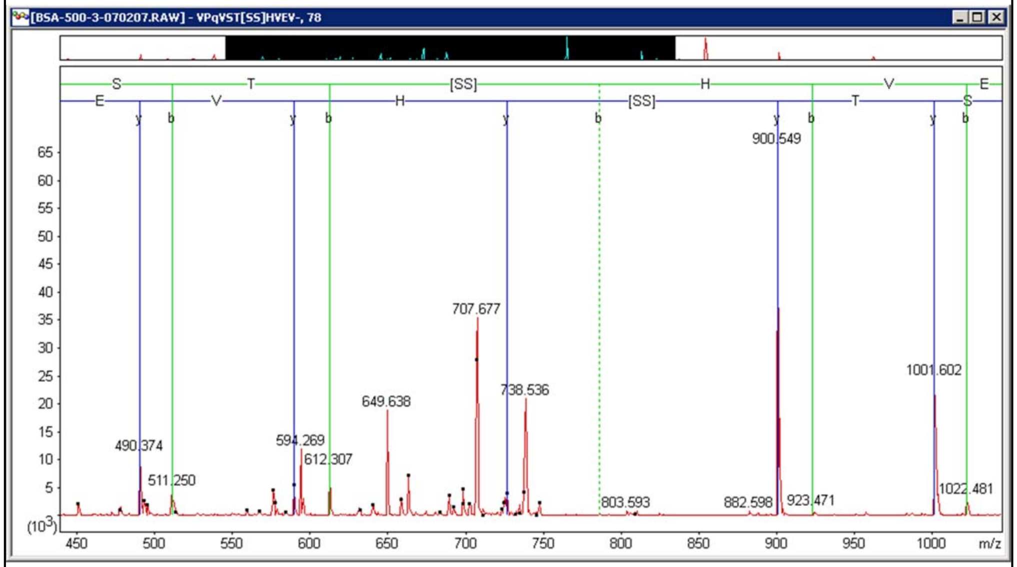
Lets take a look at another example. Here, we have a Mascot database match and a de novo match with practically the same score. The two sequences are almost identical – the difference being the de novo solution suggests SSH in the middle of the peptide where the database sequence is PTL, the masses of which are almost identical, and if you take a look at the matches in this region



[BSA-500-3-070207.RAW] - VPQVSTPTLVEVSR(2+), 79

Mascot database match

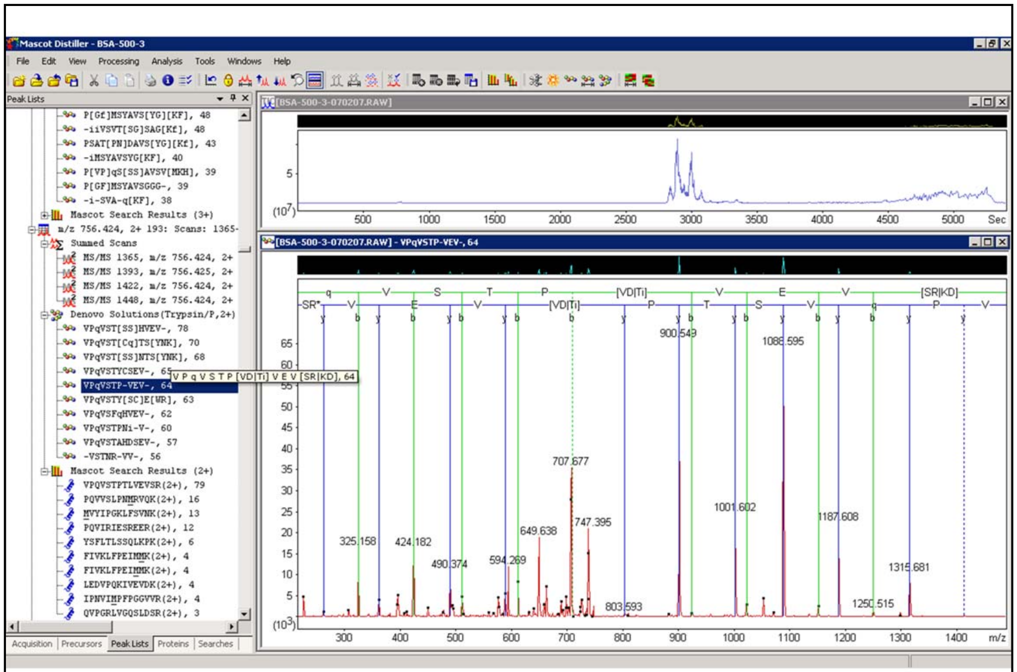
We can see that neither the Mascot database search match



De novo sequence match

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

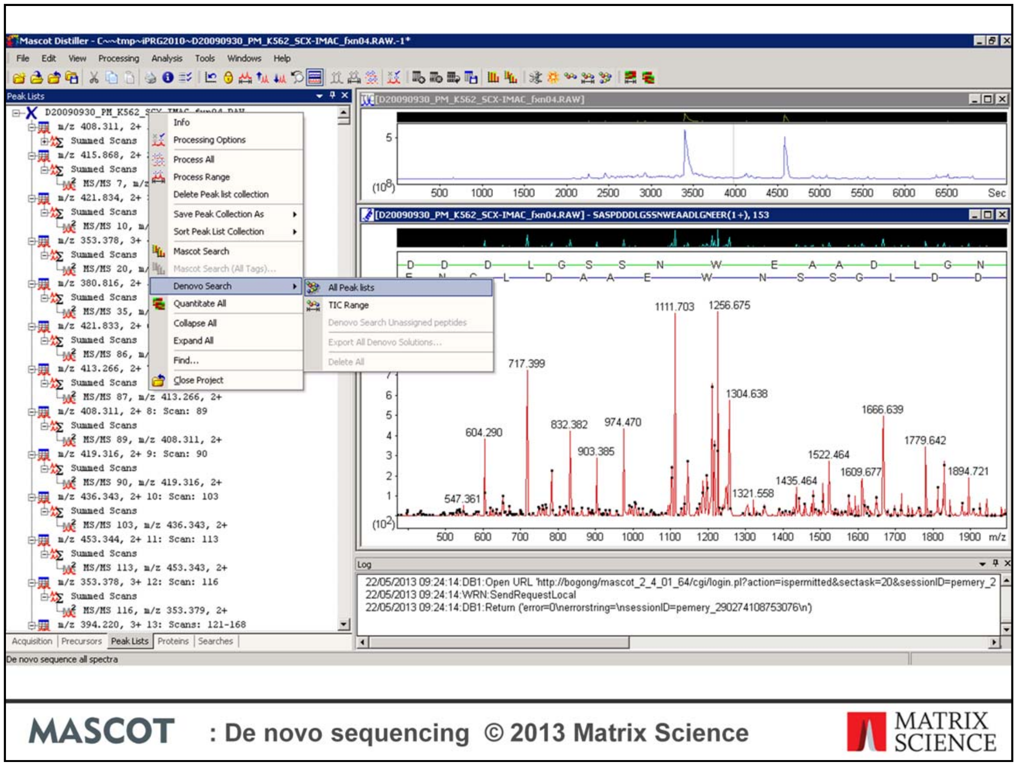
Nor the Distiller de novo search match has a complete set of fragment matches across this region.


Mascot Distiller - BSA-500-3
File Edit View Processing Analysis Tools Windows Help
Peak Lists
[BSA-500-3-070207.RAW]
(10^7)
5
500 1000 1500 2000 2500 3000 3500 4000 4500 5000 Sec
[BSA-500-3-070207.RAW] - VPqVSTP-VEV-, 64
q V S T P [VD|Ti] V E V [SR|KD]
SR* V E V [VD|Ti] P T S V q P V
900.549
1088.595
707.677
1001.602
1187.608
747.395
649.638
325.158
424.182
490.374
594.269
1315.681
1250.515
803.593
65 60 55 50 45 40 35 30 25 20 15 10 5
(10^3)
300 400 500 600 700 800 900 1000 1100 1200 1300 1400 m/z
Acquisition Precursors Peak Lists Proteins Searches

P[Gf]MSYAVS[YG][KF], 48
-iiVSVT[SG]SAG[Kf], 48
PSAT[PN]DAVS[YG][Kf], 43
-iMSYAVSYG[KF], 40
P[VP]qS[SS]AVSV[MKH], 39
P[GF]MSYAVSGGG-, 39
-i-SVA-q[KF], 38
Mascot Search Results (3+)
m/z 756.424, 2+ 193: Scans: 1365-
Summed Scans
MS/MS 1365, m/z 756.424, 2+
MS/MS 1393, m/z 756.425, 2+
MS/MS 1422, m/z 756.424, 2+
MS/MS 1448, m/z 756.424, 2+
Denovo Solutions(Trypsin/P,2+)
VPqVST[SS]HVEV-, 78
VPqVST[Cq]TS[YNK], 70
VPqVST[SS]NTS[YNK], 68
VPqVSTYCSEV-, 65
VPqVSTP-VEV-, 64
V P q V S T P [VD|Ti] V E V [SR|KD], 64
VPqVSTY[SC]E[WR], 63
VPqVSFqHVEV-, 62
VPqVSTPNi-V-, 60
VPqVSTAHDSEV-, 57
-VSTNR-VV-, 56
Mascot Search Results (2+)
VPQVSTPTLVEVSR(2+), 79
PQVVSLPNMRVQK(2+), 16
MVYIPGKLFSVNK(2+), 13
PQVIRIESREER(2+), 12
YSFLTLSSQLKPK(2+), 6
FIVKLFPEIMMK(2+), 4
FIVKLFPEIMMK(2+), 4
LEDVPQKIVEVDK(2+), 4
IPNVIMPFPGGVVR(2+), 4
QVPGRLVGQSLDSR(2+), 3

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

In fact the de novo search did find the same match as the Mascot database search, but it has a slightly lower score and has come out as the fifth ranked match. This peptide came from BSA, and this stretch of sequence is highly conserved across serum albumins from many different species, so it seems unlikely that the alternative de novo match is correct. In fact all ten reported de novo matches have reasonable scores and are, to a greater or lesser extent, a variation on the ‘correct’ (database) sequence. With such a large possible search space, this type of ambiguity is a fact of life with de novo. A sequence database search is much more constraining on the possible sequence matches.



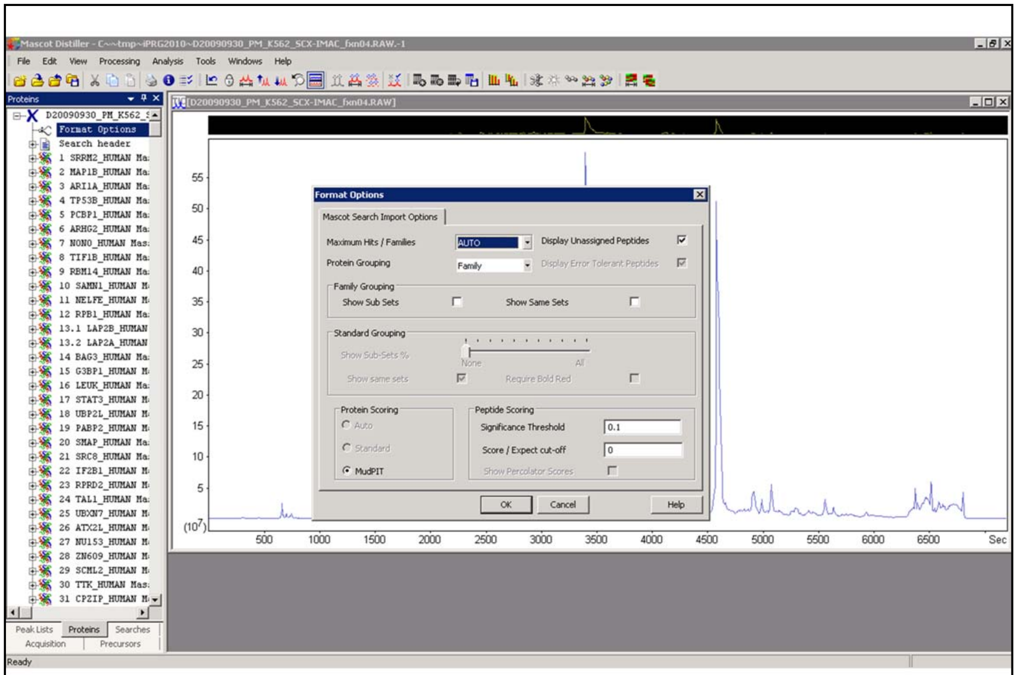
Mascot Distiller - C~~tmp~iPRG2010~D20090930_PM_K562_SCX-IMAC_fxn04.RAW.-1*
File
Edit
View
Processing
Analysis
Tools
Windows
Help
Peak Lists
D20090930_PM_K562_SCX-IMAC_fxn04.RAW
m/z 408.311, 2+
Summed Scans
m/z 415.868, 2+
MS/MS 7, m/z
m/z 421.834, 2+
MS/MS 10, m/
m/z 353.378, 3+
MS/MS 20, m/
m/z 380.816, 2+
MS/MS 35, m/
m/z 421.833, 2+
MS/MS 86, m/
m/z 413.266, 2+
MS/MS 87, m/z 413.266, 2+
m/z 408.311, 2+ 8: Scan: 89
MS/MS 89, m/z 408.311, 2+
m/z 419.316, 2+ 9: Scan: 90
MS/MS 90, m/z 419.316, 2+
m/z 436.343, 2+ 10: Scan: 103
MS/MS 103, m/z 436.343, 2+
m/z 453.344, 2+ 11: Scan: 113
MS/MS 113, m/z 453.343, 2+
m/z 353.378, 3+ 12: Scan: 116
MS/MS 116, m/z 353.379, 2+
m/z 394.220, 3+ 13: Scans: 121-168
Info
Processing Options
Process All
Process Range
Delete Peak list collection
Save Peak Collection As
Sort Peak List Collection
Mascot Search
Mascot Search (All Tags)...
Denovo Search
Quantitate All
Collapse All
Expand All
Find...
Close Project
All Peak lists
TIC Range
Denovo Search Unassigned peptides
Export All Denovo Solutions...
Delete All
Acquisition
Precursors
Peak Lists
Proteins
Searches
De novo sequence all spectra
[D20090930_PM_K562_SCX-IMAC_fxn04.RAW]
5
(10⁸)
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
Sec
[D20090930_PM_K562_SCX-IMAC_fxn04.RAW] - SASPDDDLGSSNWEAADLGNEER(1+), 153
D D D L G S S N W E A A D L G N
E N G L D A A E W N S S G L D D
1111.703
1256.675
717.399
1304.638
1666.639
604.290
832.382
974.470
903.385
1779.642
1522.464
1609.677
1894.721
1435.464
547.361
1321.558
7
6
5
4
3
2
1
(10²)
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
m/z
Log

```
22/05/2013 09:24:14:DB1:Open URL 'http://bogong/mascot_2_4_01_64/cgi/login.pl?action=ispermitted&sectask=20&sessionID=pemery_2
22/05/2013 09:24:14:WRN:SendRequestLocal
22/05/2013 09:24:14:DB1:Return ('error=0\nerrorstring=\nsessionID=pemery_290274108753076\n')
```

MASCOT : De novo sequencing © 2013 Matrix Science

MATRIX SCIENCE

To de novo sequence a complete peak list collection, or the peak lists in the currently displayed TIC range, use the context menu obtained by right-clicking the root (world) node of the peak lists tree



Format Options

Mascot Search Import Options

Maximum Hits / Families: AUTO

Display Unassigned Peptides ☑

Protein Grouping: Family

Display Error Tolerant Peptides ☑

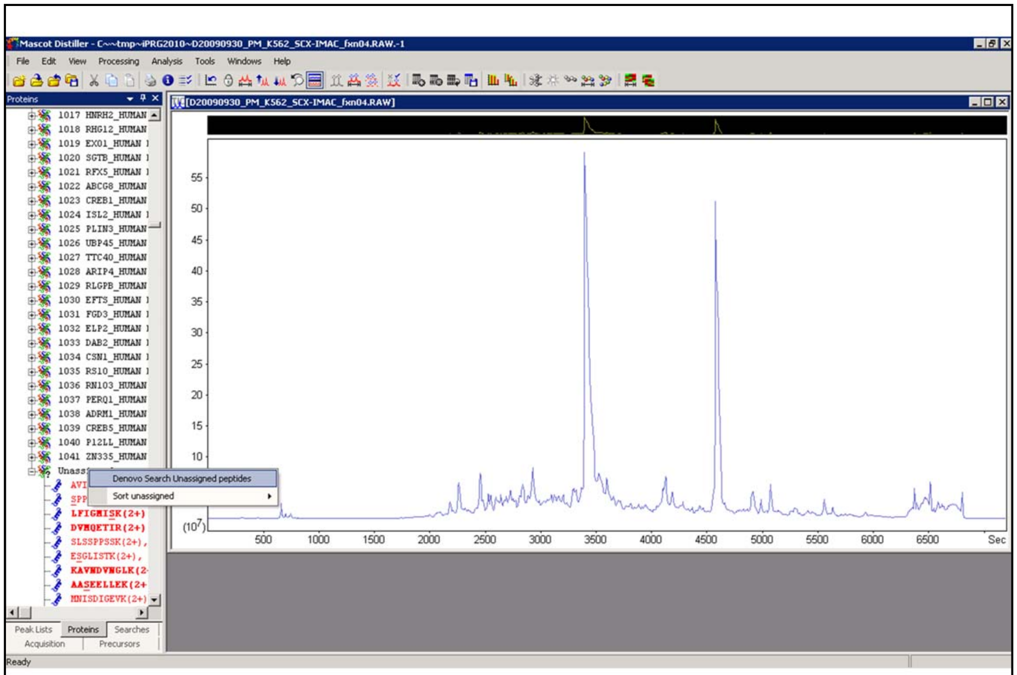
Family Grouping
Show Sub Sets □
Show Same Sets □

Standard Grouping
Show Sub-Sets %: None … All
Show same sets ☑
Require Bold Red □

Protein Scoring
○ Auto
○ Standard
● MudPIT

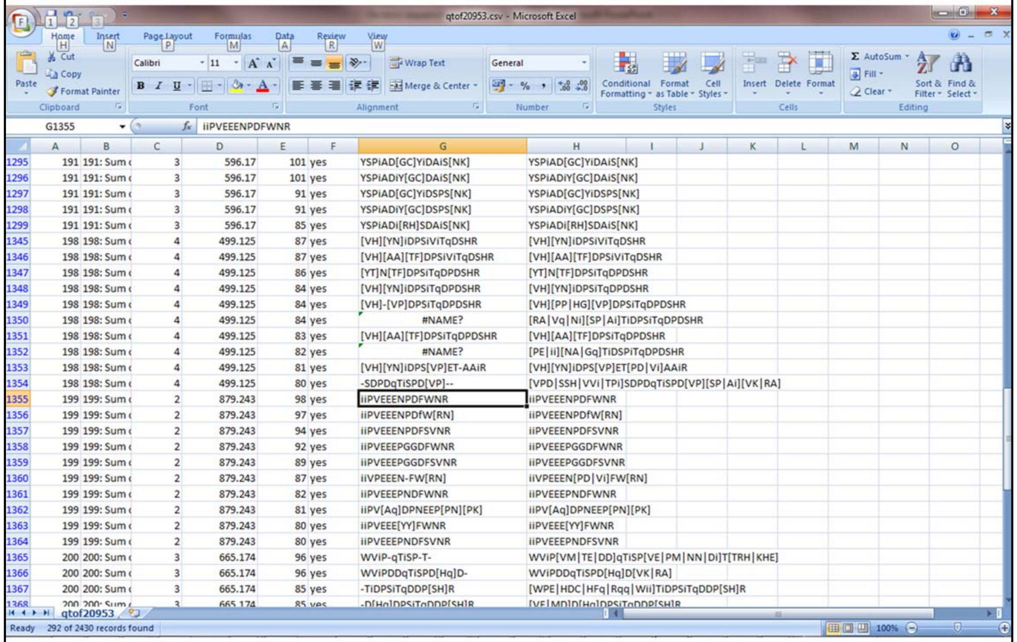
Peptide Scoring
Significance Threshold: 0.1
Score / Expect cut-off: 0
Show Percolator Scores □

One common reason for carrying out a de novo search is to try and find additional matches to spectra in the ‘Unassigned’ list in a Mascot search result – these are the spectra that failed to give decent matches. The most efficient way to de novo these spectra is to switch to the proteins tree, click on Format Options, and choose to load the unassigned queries.



MASCOT : De novo sequencing © 2013 Matrix Science

Scroll down to the bottom of the proteins tree and use the context menu obtained by right-clicking the unassigned node to de novo just the unassigned queries. Once completed, the de novo solutions are added to the ‘Peak list’ tree in the dataset explorer. From there, we can scroll down the tree and look for spectra where we have a high scoring de novo solution, with no reliable Mascot database search match.



G1355 iiPVEEENPDFWNR

	A	B	C	D	E	F	G	H
1295	191	191: Sum c	3	596.17	101	yes	YSPiAD[GC]YiDAiS[NK]	YSPiAD[GC]YiDAiS[NK]
1296	191	191: Sum c	3	596.17	101	yes	YSPiADiY[GC]DAiS[NK]	YSPiADiY[GC]DAiS[NK]
1297	191	191: Sum c	3	596.17	91	yes	YSPiAD[GC]YiDSPS[NK]	YSPiAD[GC]YiDSPS[NK]
1298	191	191: Sum c	3	596.17	91	yes	YSPiADiY[GC]DSPS[NK]	YSPiADiY[GC]DSPS[NK]
1299	191	191: Sum c	3	596.17	85	yes	YSPiADi[RH]SDAiS[NK]	YSPiADi[RH]SDAiS[NK]
1345	198	198: Sum c	4	499.125	87	yes	[VH][YN]iDPSiViTqDSHR	[VH][YN]iDPSiViTqDSHR
1346	198	198: Sum c	4	499.125	87	yes	[VH][AA][TF]DPSiViTqDSHR	[VH][AA][TF]DPSiViTqDSHR
1347	198	198: Sum c	4	499.125	86	yes	[YT]N[TF]DPSiTqDPDSHR	[YT]N[TF]DPSiTqDPDSHR
1348	198	198: Sum c	4	499.125	84	yes	[VH][YN]iDPSiTqDPDSHR	[VH][YN]iDPSiTqDPDSHR
1349	198	198: Sum c	4	499.125	84	yes	[VH]-[VP]DPSiTqDPDSHR	[VH][PP|HG][VP]DPSiTqDPDSHR
1350	198	198: Sum c	4	499.125	84	yes	#NAME?	[RA|Vq|Ni][SP|Ai]TiDPSiTqDPDSHR
1351	198	198: Sum c	4	499.125	83	yes	[VH][AA][TF]DPSiTqDPDSHR	[VH][AA][TF]DPSiTqDPDSHR
1352	198	198: Sum c	4	499.125	82	yes	#NAME?	[PE|ii][NA|Gq]TiDSPiTqDPDSHR
1353	198	198: Sum c	4	499.125	81	yes	[VH][YN]iDPS[VP]ET-AAiR	[VH][YN]iDPS[VP]ET[PD|Vi]AAiR
1354	198	198: Sum c	4	499.125	80	yes	-SDPDqTiSPD[VP]--	[VPD|SSH|VVi|TPi]SDPDqTiSPD[VP][SP|Ai][VK|RA]
1355	199	199: Sum c	2	879.243	98	yes	iiPVEEENPDFWNR	iiPVEEENPDFWNR
1356	199	199: Sum c	2	879.243	97	yes	iiPVEEENPDfW[RN]	iiPVEEENPDfW[RN]
1357	199	199: Sum c	2	879.243	94	yes	iiPVEEENPDFSVNR	iiPVEEENPDFSVNR
1358	199	199: Sum c	2	879.243	92	yes	iiPVEEEPGGDFWNR	iiPVEEEPGGDFWNR
1359	199	199: Sum c	2	879.243	89	yes	iiPVEEEPGGDFSVNR	iiPVEEEPGGDFSVNR
1360	199	199: Sum c	2	879.243	87	yes	iiVPEEEN-FW[RN]	iiVPEEEN[PD|Vi]FW[RN]
1361	199	199: Sum c	2	879.243	82	yes	iiPVEEEPNDFWNR	iiPVEEEPNDFWNR
1362	199	199: Sum c	2	879.243	81	yes	iiPV[Aq]DPNEEP[PN][PK]	iiPV[Aq]DPNEEP[PN][PK]
1363	199	199: Sum c	2	879.243	80	yes	iiPVEEE[YY]FWNR	iiPVEEE[YY]FWNR
1364	199	199: Sum c	2	879.243	80	yes	iiPVEEEPNDFSVNR	iiPVEEEPNDFSVNR
1365	200	200: Sum c	3	665.174	96	yes	WViP-qTiSP-T-	WViP[VM|TE|DD]qTiSP[VE|PM|NN|Di]T[TRH|KHE]
1366	200	200: Sum c	3	665.174	96	yes	WViPDDqTiSPD[Hq]D-	WViPDDqTiSPD[Hq]D[VK|RA]
1367	200	200: Sum c	3	665.174	85	yes	-TiDPSiTqDDP[SH]R	[WPE|HDC|HFq|Rqq|Wii]TiDPSiTqDDP[SH]R
1368	200	200: Sum c	3	665.174	85	yes	-D[Hq]DPSiTqDDP[SH]R	[VE|MD]D[Hq]DPSiTqDDP[SH]R

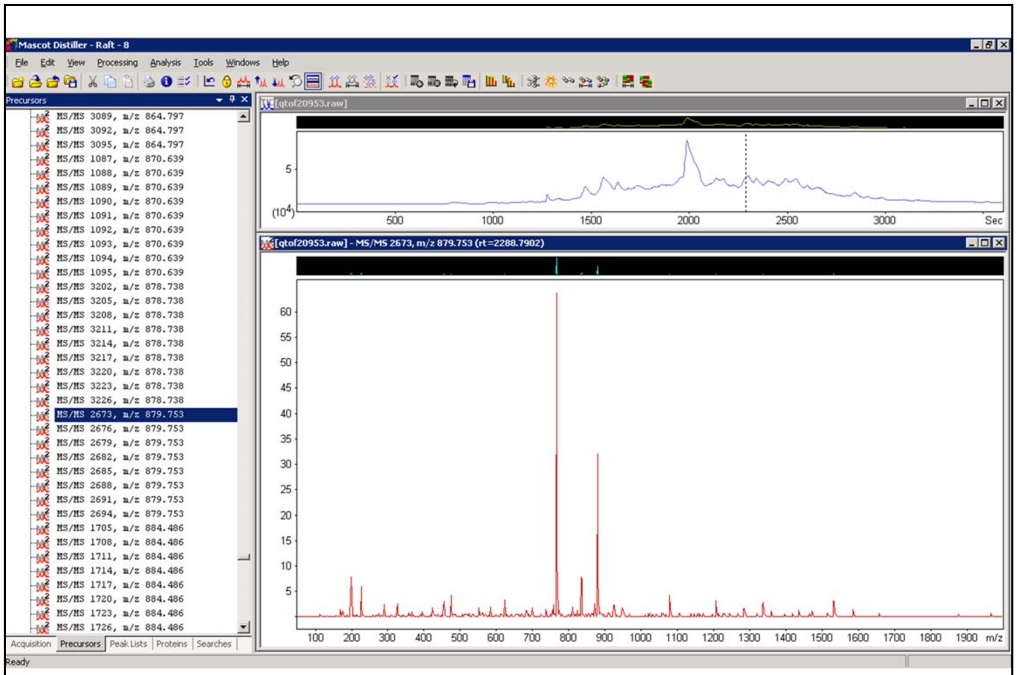
MASCOT : De novo sequencing © 2013 Matrix Science

MATRIX SCIENCE

For a large dataset, an alternative is to export the de novo solutions as a CSV file and then viewing and filtering the results in Microsoft Excel. To do this, select the ‘Export All De novo Solutions’ option from the ‘File’ menu, or from the context menu obtained by right clicking the root (world) node of the peak lists tree. Once you’ve exported the results, you can load them directly into Excel and use the filtering functions to screen the de novo matches.

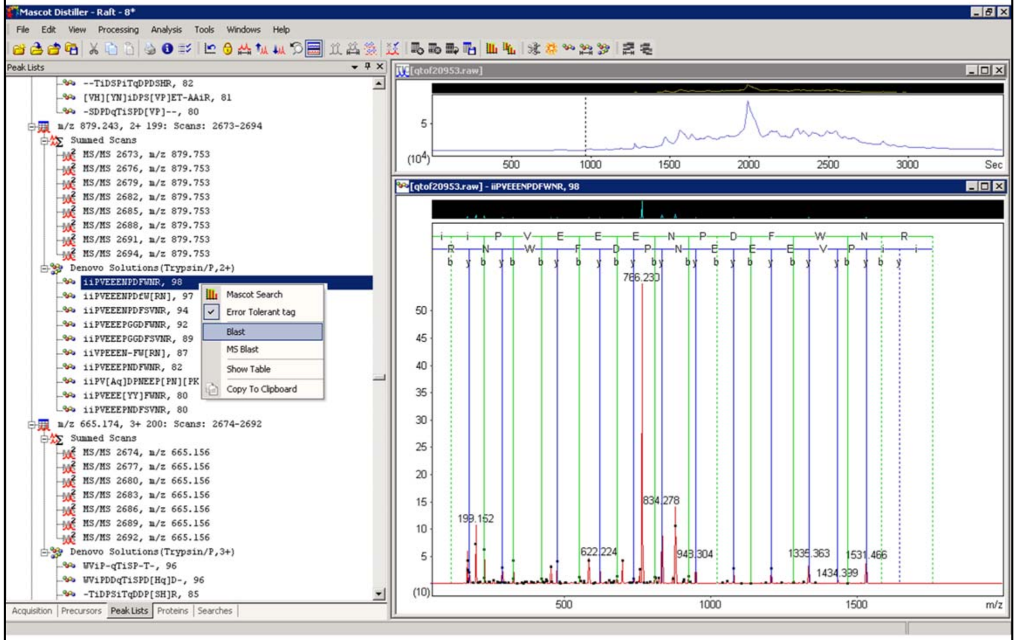
These data are from a QTof dataset. Here, I’ve filtered on the ‘Score’ column in the CSV file for values greater than or equal to 80.

With high quality data, we can get something close to a complete peptide sequence, such as the match highlighted here. To take a look at this match in Distiller, note the query m/z value of 879



```
MS/MS 3089, m/z 864.797
MS/MS 3092, m/z 864.797
MS/MS 3095, m/z 864.797
MS/MS 1087, m/z 870.639
MS/MS 1088, m/z 870.639
MS/MS 1089, m/z 870.639
MS/MS 1090, m/z 870.639
MS/MS 1091, m/z 870.639
MS/MS 1092, m/z 870.639
MS/MS 1093, m/z 870.639
MS/MS 1094, m/z 870.639
MS/MS 1095, m/z 870.639
MS/MS 3202, m/z 878.738
MS/MS 3205, m/z 878.738
MS/MS 3208, m/z 878.738
MS/MS 3211, m/z 878.738
MS/MS 3214, m/z 878.738
MS/MS 3217, m/z 878.738
MS/MS 3220, m/z 878.738
MS/MS 3223, m/z 878.738
MS/MS 3226, m/z 878.738
MS/MS 2673, m/z 879.753
MS/MS 2676, m/z 879.753
MS/MS 2679, m/z 879.753
MS/MS 2682, m/z 879.753
MS/MS 2685, m/z 879.753
MS/MS 2688, m/z 879.753
MS/MS 2691, m/z 879.753
MS/MS 2694, m/z 879.753
MS/MS 1705, m/z 884.486
MS/MS 1708, m/z 884.486
MS/MS 1711, m/z 884.486
MS/MS 1714, m/z 884.486
MS/MS 1717, m/z 884.486
MS/MS 1720, m/z 884.486
MS/MS 1723, m/z 884.486
MS/MS 1726, m/z 884.486
```

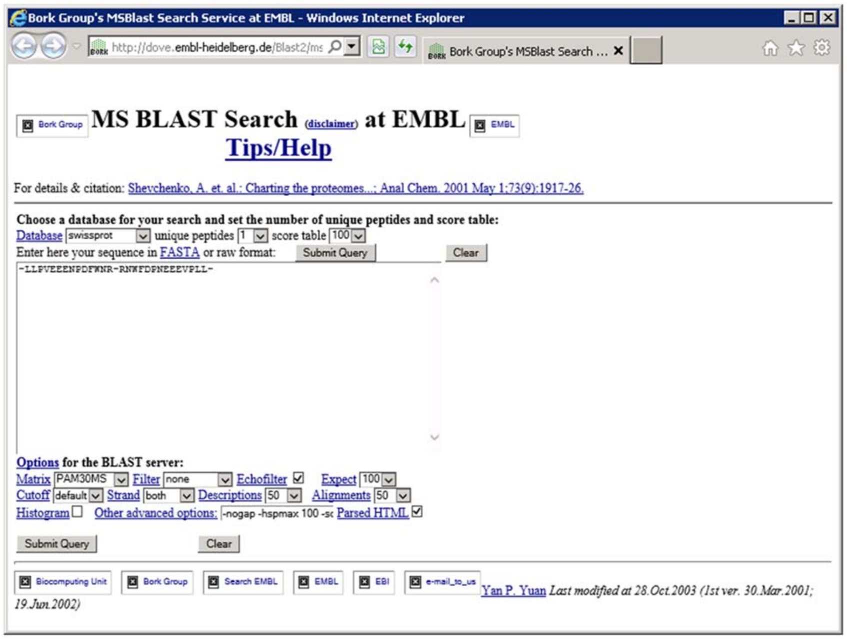
Return to Distiller, select the ‘Precursors’ tab, and scroll down or use ‘Find’ function in Distiller to find the precursor. The tabs in the Distiller dataset explorer are, as far as is possible, kept synchronised – so now select the ‘Peak Lists’ tab



```
--TiDSPiTqDPDSHR, 82
[VH][YN]iDPS[VP]ET-AAiR, 81
-SDPDqTiSPD[VP]--, 80
m/z 879.243, 2+ 199: Scans: 2673-2694
Summed Scans
MS/MS 2673, m/z 879.753
MS/MS 2676, m/z 879.753
MS/MS 2679, m/z 879.753
MS/MS 2682, m/z 879.753
MS/MS 2685, m/z 879.753
MS/MS 2688, m/z 879.753
MS/MS 2691, m/z 879.753
MS/MS 2694, m/z 879.753
Denovo Solutions(Trypsin/P,2+)
iiPVEEENPDFWNR, 98
iiPVEEENPDfW[RN], 97
iiPVEEENPDFSVNR, 94
iiPVEEEPGGDFWNR, 92
iiPVEEEPGGDFSVNR, 89
iiVPEEEN-FW[RN], 87
iiPVEEEPNDFWNR, 82
iiPV[Aq]DPNEEP[PN][PK
iiPVEEE[YY]FWNR, 80
iiPVEEEPNDFSVNR, 80
m/z 665.174, 3+ 200: Scans: 2674-2692
Summed Scans
MS/MS 2674, m/z 665.156
MS/MS 2677, m/z 665.156
MS/MS 2680, m/z 665.156
MS/MS 2683, m/z 665.156
MS/MS 2686, m/z 665.156
MS/MS 2689, m/z 665.156
MS/MS 2692, m/z 665.156
Denovo Solutions(Trypsin/P,3+)
WViP-qTiSP-T-, 96
WViPDDqTiSPD[Hq]D-, 96
-TiDPSiTqDDP[SH]R, 85
```

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

And here is our match. To see if we can get a match into a protein sequence, with a sequence this clean we could carry out a sequence homology search. To do this, right click over the de novo peptide match to bring up the context menu. This gives us various options to submit a Mascot tag search (we'll be taking a look at that in a moment), and to submit the peptide to a BLAST or MS-BLAST server. In this case we'll submit the peptide for an MS_BLAST search.


Bork Group's MSBlast Search Service at EMBL - Windows Internet Explorer
http://dove.embl-heidelberg.de/Blast2/ms
Bork Group's MSBlast Search ...
Bork Group
EMBL
Submit Query
Clear
Submit Query
Clear
Biocomputing Unit
Bork Group
Search EMBL
EMBL
EBI
e-mail_to_us

MS BLAST Search (disclaimer) at EMBL

Tips/Help

For details & citation: Shevchenko, A. et. al.: Charting the proteomes...; Anal Chem. 2001 May 1;73(9):1917-26.

Choose a database for your search and set the number of unique peptides and score table:

Database swissprot unique peptides 1 score table 100

Enter here your sequence in FASTA or raw format:

```
-LLPVEEENPDFWNR-RNWFDPNEEEVPLL-
```

Options for the BLAST server:

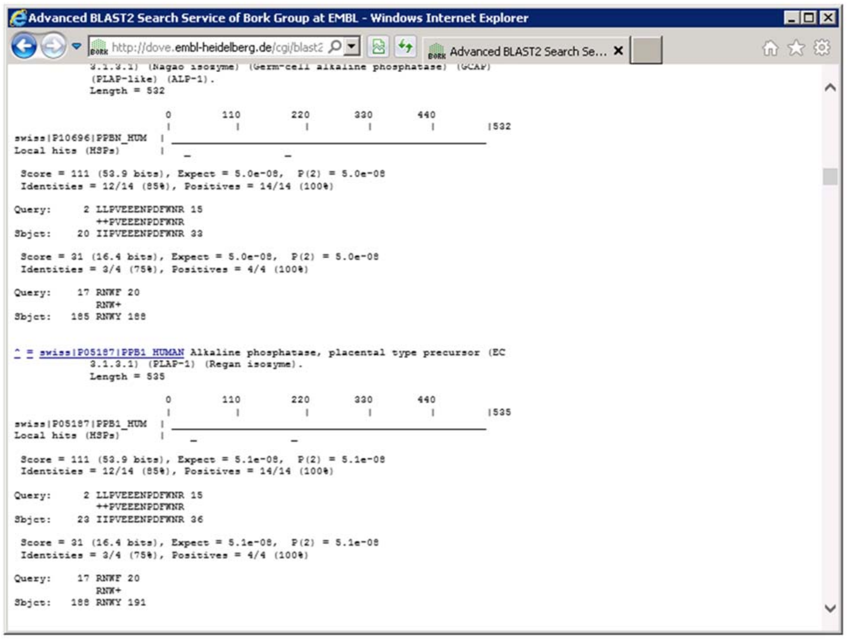
Matrix PAM30MS Filter none Echofilter ☑ Expect 100

Cutoff default Strand both Descriptions 50 Alignments 50

Histogram ☐ Other advanced options: -nogap -hspmax 100 -sc Parsed HTML ☑

Yan P. Yuan *Last modified at 28.Oct.2003 (1st ver. 30.Mar.2001; 19.Jun.2002)*

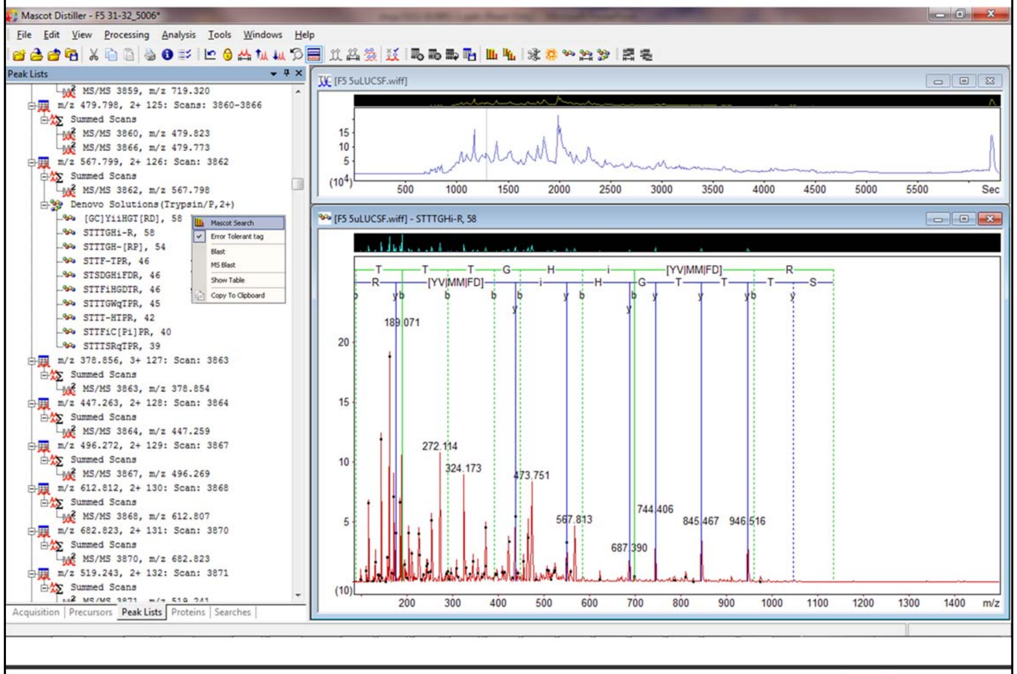
If you submit a BLAST search, Distiller will open up submission windows – one with the forward peptide sequence and one with the reverse. Also, if there is ambiguity in the sequence only a single version of the peptide will be submitted for each direction. When submitting to MS Blast however, any ambiguity in the sequence will be expanded and formatted using MS Blast ambiguity codes, and all the generated the forward and reverse peptide sequences will be submitted together.



Advanced BLAST2 Search Service of Bork Group at EMBL - Windows Internet Explorer
http://dove.embl-heidelberg.de/cgi/blast2
Advanced BLAST2 Search Se...

```
           3.1.3.1) (Nagao isozyme) (Germ-cell alkaline phosphatase) (GCAP)
           (PLAP-like) (ALP-1).
           Length = 532

                          0         110        220       330       440
                          |          |          |         |         |        |532
swiss|P10696|PPBN_HUM  |  ______________________________________________________
Local hits (HSPs)      |    _               _

 Score = 111 (53.9 bits), Expect = 5.0e-08,  P(2) = 5.0e-08
 Identities = 12/14 (85%), Positives = 14/14 (100%)

Query:     2 LLPVEEENPDFWNR 15
             ++PVEEENPDFWNR
Sbjct:    20 IIPVEEENPDFWNR 33

 Score = 31 (16.4 bits), Expect = 5.0e-08,  P(2) = 5.0e-08
 Identities = 3/4 (75%), Positives = 4/4 (100%)

Query:    17 RNWF 20
             RNW+
Sbjct:   185 RNWY 188


^ = swiss|P05187|PPB1_HUMAN Alkaline phosphatase, placental type precursor (EC
           3.1.3.1) (PLAP-1) (Regan isozyme).
           Length = 535

                          0         110        220       330       440
                          |          |          |         |         |        |535
swiss|P05187|PPB1_HUM  |  ______________________________________________________
Local hits (HSPs)      |    _               _

 Score = 111 (53.9 bits), Expect = 5.1e-08,  P(2) = 5.1e-08
 Identities = 12/14 (85%), Positives = 14/14 (100%)

Query:     2 LLPVEEENPDFWNR 15
             ++PVEEENPDFWNR
Sbjct:    23 IIPVEEENPDFWNR 36

 Score = 31 (16.4 bits), Expect = 5.1e-08,  P(2) = 5.1e-08
 Identities = 3/4 (75%), Positives = 4/4 (100%)

Query:    17 RNWF 20
             RNW+
Sbjct:   188 RNWY 191
```

When we get the results back, we find identity matches to several Alkaline phophatases, one of which PPB1_HUMAN was found in the original Mascot database search. In this case we have a nice match to a semi-tryptic peptide. An Error Tolerant Mascot database search would also have found this match.

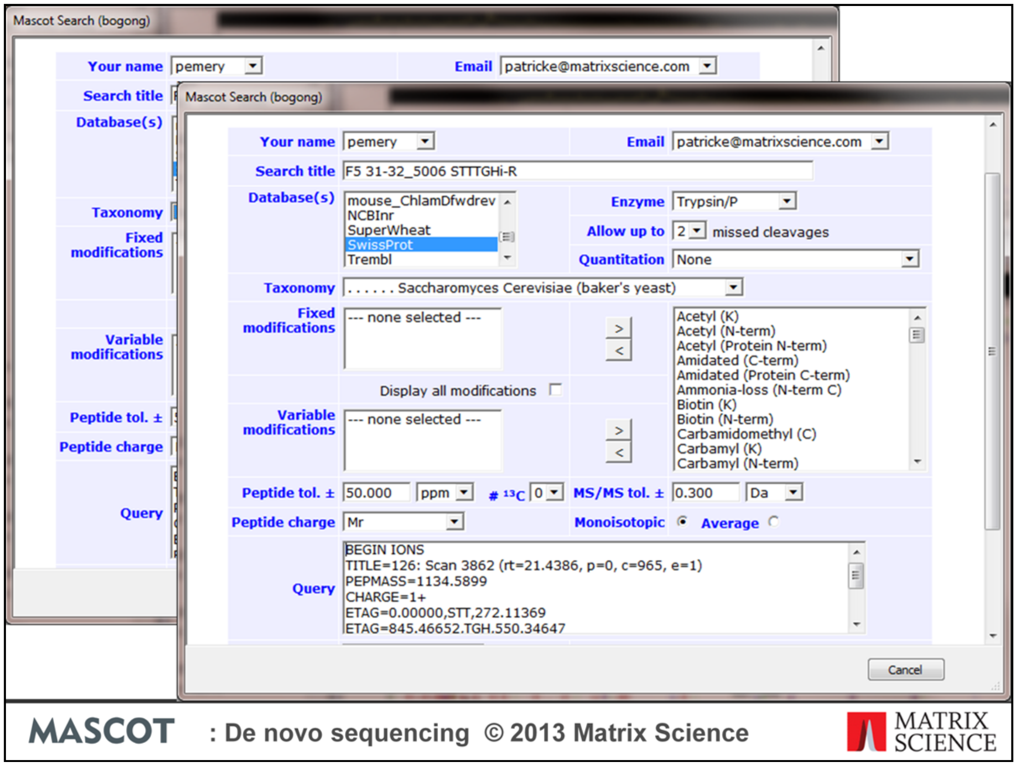


```
MS/MS 3859, m/z 719.320
m/z 479.798, 2+ 125: Scans: 3860-3866
  Summed Scans
    MS/MS 3860, m/z 479.823
    MS/MS 3866, m/z 479.773
m/z 567.799, 2+ 126: Scan: 3862
  Summed Scans
    MS/MS 3862, m/z 567.798
  Denovo Solutions(Trypsin/P,2+)
    [GC]YiiHGT[RD], 58
    STTTGHi-R, 58
    STTTGH-[RP], 54
    STTF-TPR, 46
    STSDGHiFDR, 46
    STTFiHGDTR, 46
    STTTGWqTPR, 45
    STTT-HTPR, 42
    STTFiC[Pi]PR, 40
    STTTSRqTPR, 39
m/z 378.856, 3+ 127: Scan: 3863
  Summed Scans
    MS/MS 3863, m/z 378.854
m/z 447.263, 2+ 128: Scan: 3864
  Summed Scans
    MS/MS 3864, m/z 447.259
m/z 496.272, 2+ 129: Scan: 3867
  Summed Scans
    MS/MS 3867, m/z 496.269
m/z 612.812, 2+ 130: Scan: 3868
  Summed Scans
    MS/MS 3868, m/z 612.807
m/z 682.823, 2+ 131: Scan: 3870
  Summed Scans
    MS/MS 3870, m/z 682.823
m/z 519.243, 2+ 132: Scan: 3871
  Summed Scans
    MS/MS 3871, m/z 519.241
```

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

That is a fairly unusual case, and you could have just read the peptide sequence off. With typical data you'll often find that you only get partial sequence matches with rather more ambiguity in the sequence. Here, for example, we've got a match with where the top two de novo solutions have the same score, and both have a reasonable degree of ambiguity. The next step might be to take these partial sequences and search them as error tolerant sequence tag searches. This will allow the peptide mass to vary and allows the tags to float, potentially finding matches where there are sequence variations (from SNPs etc), none specific peptides or unsuspected modifications.

We can use the de novo solution to generate the tags automatically. To do this right click over the solution to bring up the context menu. Note that we have the 'Error tolerant tag' selected.



Mascot Search (bogong)

Your name: pemery
Email: patricke@matrixscience.com
Search title
Database(s)
Taxonomy
Fixed modifications
Variable modifications
Peptide tol. ±
Peptide charge
Query

Mascot Search (bogong)

Your name	pemery	Email	patricke@matrixscience.com
Search title	F5 31-32_5006 STTTGHi-R		
Database(s)	mouse_ChlamDfwdrev NCBInr SuperWheat SwissProt Trembl	Enzyme	Trypsin/P
		Allow up to	2 missed cleavages
		Quantitation	None
Taxonomy	. . . . . . Saccharomyces Cerevisiae (baker's yeast)		
Fixed modifications	--- none selected ---	> <	Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-term) Ammonia-loss (N-term C) Biotin (K) Biotin (N-term) Carbamidomethyl (C) Carbamyl (K) Carbamyl (N-term)
	Display all modifications □		
Variable modifications	--- none selected ---	> <	
Peptide tol. ±	50.000 ppm # ^{13}C 0	MS/MS tol. ±	0.300 Da
Peptide charge	Mr	Monoisotopic ●	Average ○

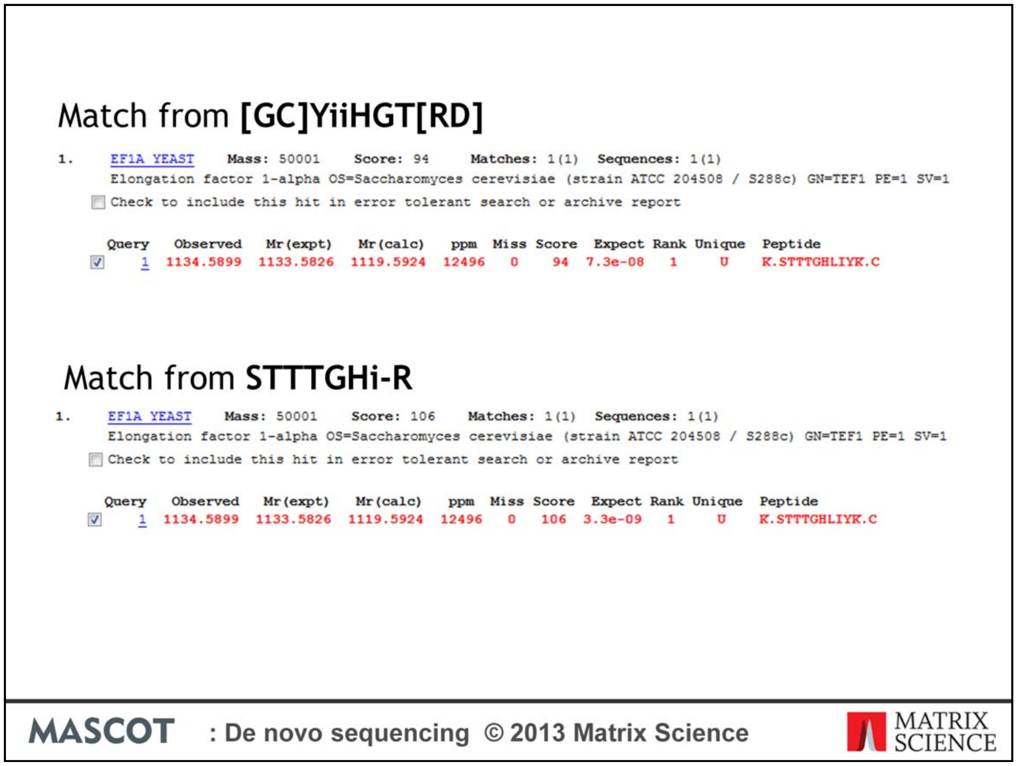
Query

```
BEGIN IONS
TITLE=126: Scan 3862 (rt=21.4386, p=0, c=965, e=1)
PEPMASS=1134.5899
CHARGE=1+
ETAG=0.00000,STT,272.11369
ETAG=845.46652,TGH,550.34647
```

MASCOT : De novo sequencing © 2013 Matrix Science

MATRIX SCIENCE

Distiller populates the query field with the tags taken from the non-ambiguous parts of the de novo solution. We do this for each of the top two de novo solutions individually. Then we submit the searches …



Match from **[GC]YiiHGT[RD]**

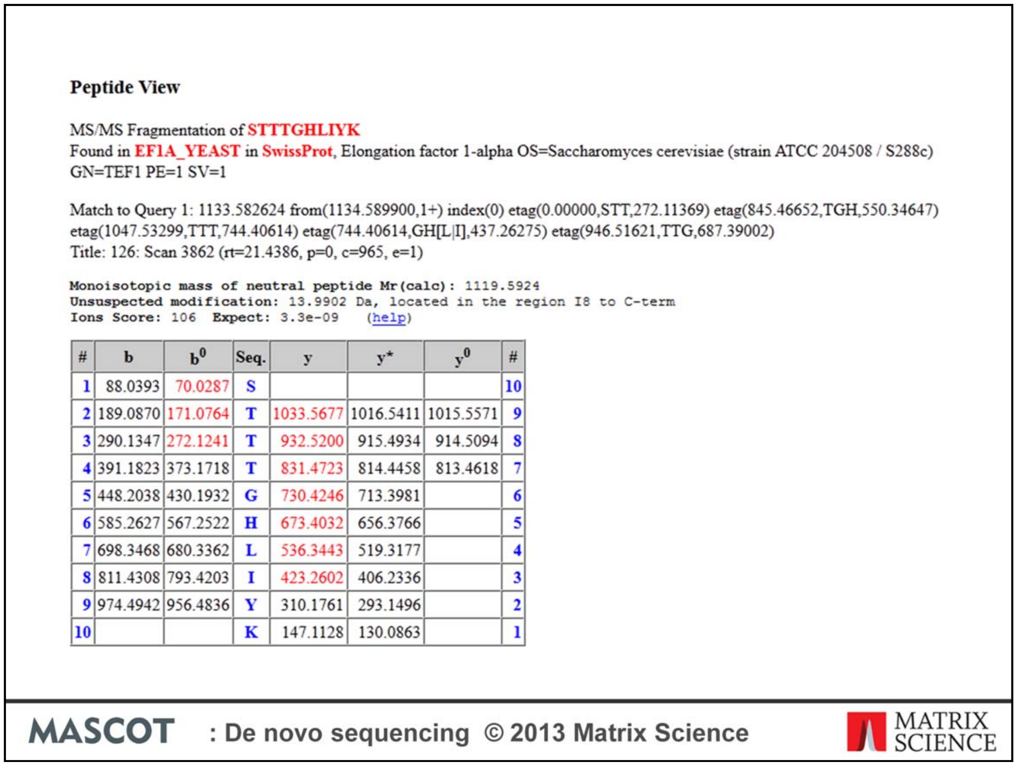
```
1.     EF1A_YEAST     Mass: 50001     Score: 94      Matches: 1(1)   Sequences: 1(1)
       Elongation factor 1-alpha OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=TEF1 PE=1 SV=1
    □ Check to include this hit in error tolerant search or archive report

      Query   Observed   Mr(expt)   Mr(calc)    ppm  Miss Score  Expect Rank Unique  Peptide
    ☑    1  1134.5899  1133.5826  1119.5924  12496    0    94  7.3e-08    1     U    K.STTTGHLIYK.C
```

Match from **STTTGHi-R**

```
1.     EF1A_YEAST     Mass: 50001     Score: 106     Matches: 1(1)   Sequences: 1(1)
       Elongation factor 1-alpha OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=TEF1 PE=1 SV=1
    □ Check to include this hit in error tolerant search or archive report

      Query   Observed   Mr(expt)   Mr(calc)    ppm  Miss Score  Expect Rank Unique  Peptide
    ☑    1  1134.5899  1133.5826  1119.5924  12496    0   106  3.3e-09    1     U    K.STTTGHLIYK.C
```

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

When the results come back, we get the same match back from the tags from the two different de novo solutions, but the match from the second de novo solution gets a better score.



Peptide View

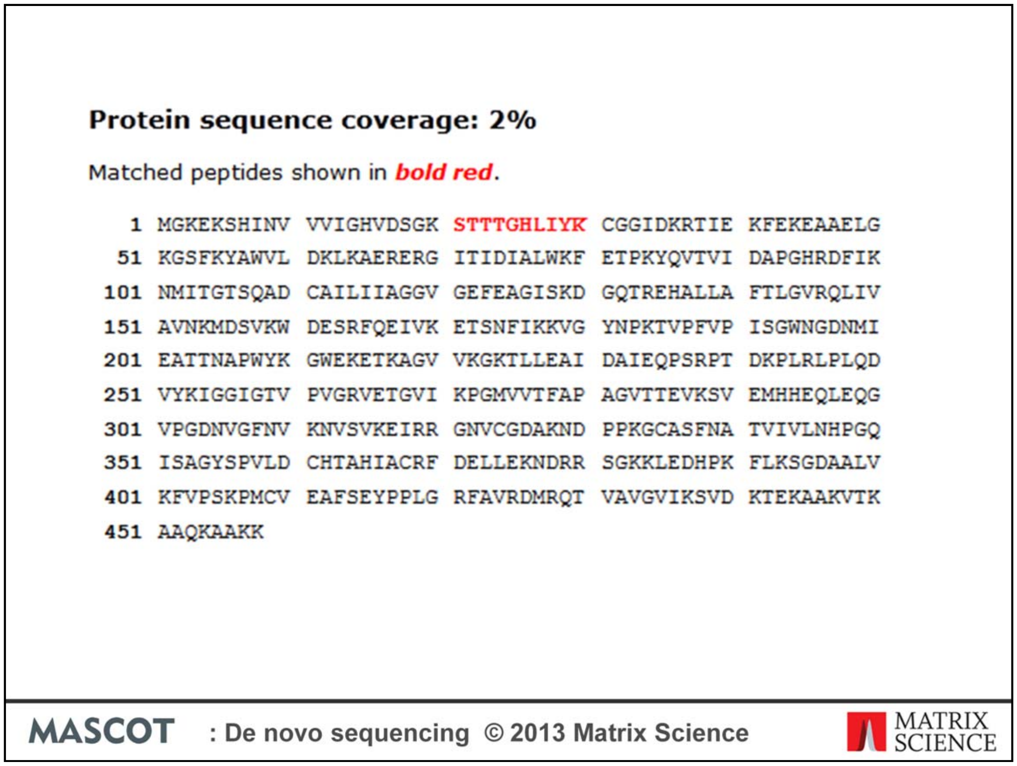
MS/MS Fragmentation of **STTTGHLIYK**
Found in **EF1A_YEAST** in **SwissProt**, Elongation factor 1-alpha OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=TEF1 PE=1 SV=1

Match to Query 1: 1133.582624 from(1134.589900,1+) index(0) etag(0.00000,STT,272.11369) etag(845.46652,TGH,550.34647) etag(1047.53299,TTT,744.40614) etag(744.40614,GH[L|I],437.26275) etag(946.51621,TTG,687.39002)
Title: 126: Scan 3862 (rt=21.4386, p=0, c=965, e=1)

```
Monoisotopic mass of neutral peptide Mr(calc): 1119.5924
Unsuspected modification: 13.9902 Da, located in the region I8 to C-term
Ions Score: 106  Expect: 3.3e-09   (help)
```

#	b	b^0	Seq.	y	y*	y^0	#
1	88.0393	70.0287	S				10
2	189.0870	171.0764	T	1033.5677	1016.5411	1015.5571	9
3	290.1347	272.1241	T	932.5200	915.4934	914.5094	8
4	391.1823	373.1718	T	831.4723	814.4458	813.4618	7
5	448.2038	430.1932	G	730.4246	713.3981		6
6	585.2627	567.2522	H	673.4032	656.3766		5
7	698.3468	680.3362	L	536.3443	519.3177		4
8	811.4308	793.4203	I	423.2602	406.2336		3
9	974.4942	956.4836	Y	310.1761	293.1496		2
10			K	147.1128	130.0863		1

If we take a look at the peptide view for this match, it is suggesting an unsuspected modification of 13.9902 Da in the region I8 to the C-terminus.



Protein sequence coverage: 2%

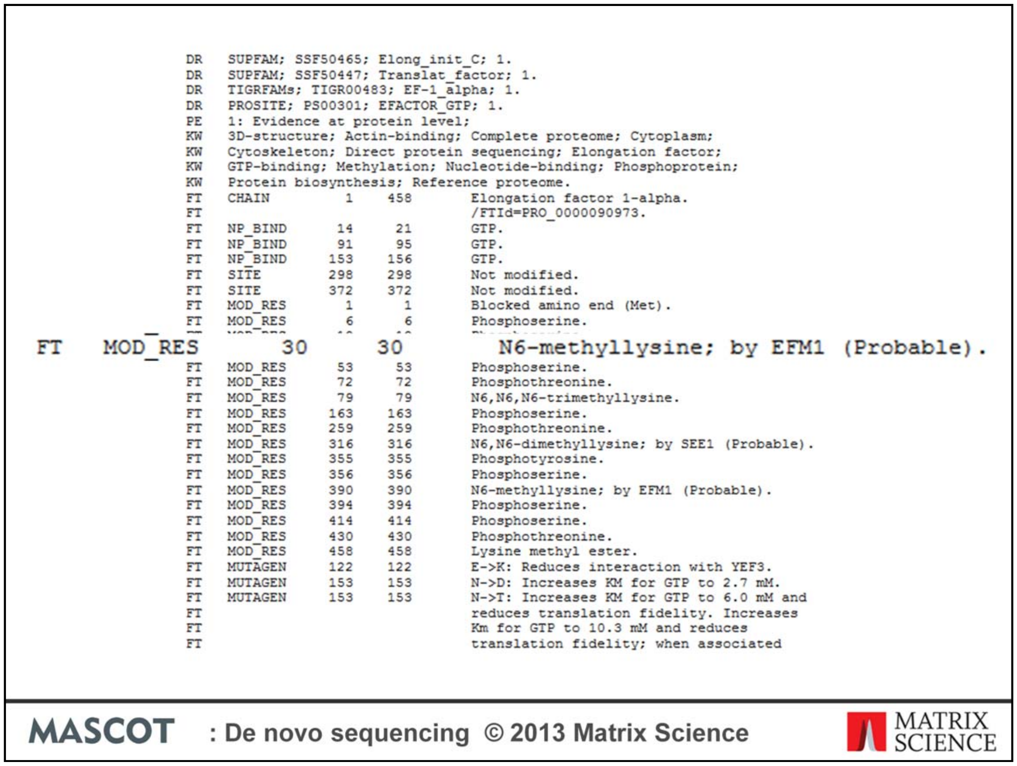
Matched peptides shown in ***bold red***.

```
  1 MGKEKSHINV VVIGHVDSGK STTTGHLIYK CGGIDKRTIE KFEKEAAELG
 51 KGSFKYAWVL DKLKAERERG ITIDIALWKF ETPKYQVTVI DAPGHRDFIK
101 NMITGTSQAD CAILIIAGGV GEFEAGISKD GQTREHALLA FTLGVRQLIV
151 AVNKMDSVKW DESRFQEIVK ETSNFIKKVG YNPKTVPFVP ISGWNGDNMI
201 EATTNAPWYK GWEKETKAGV VKGKTLLEAI DAIEQPSRPT DKPLRLPLQD
251 VYKIGGIGTV PVGRVETGVI KPGMVVTFAP AGVTTEVKSV EMHHEQLEQG
301 VPGDNVGFNV KNVSVKEIRR GNVCGDAKND PPKGCASFNA TVIVLNHPGQ
351 ISAGYSPVLD CHTAHIACRF DELLEKNDRR SGKKLEDHPK FLKSGDAALV
401 KFVPSKPMCV EAFSEYPPLG RFAVRDMRQT VAVGVIKSVD KTEKAAKVTK
451 AAQKAAKK
```

MASCOT : De novo sequencing © 2013 Matrix Science

MATRIX SCIENCE

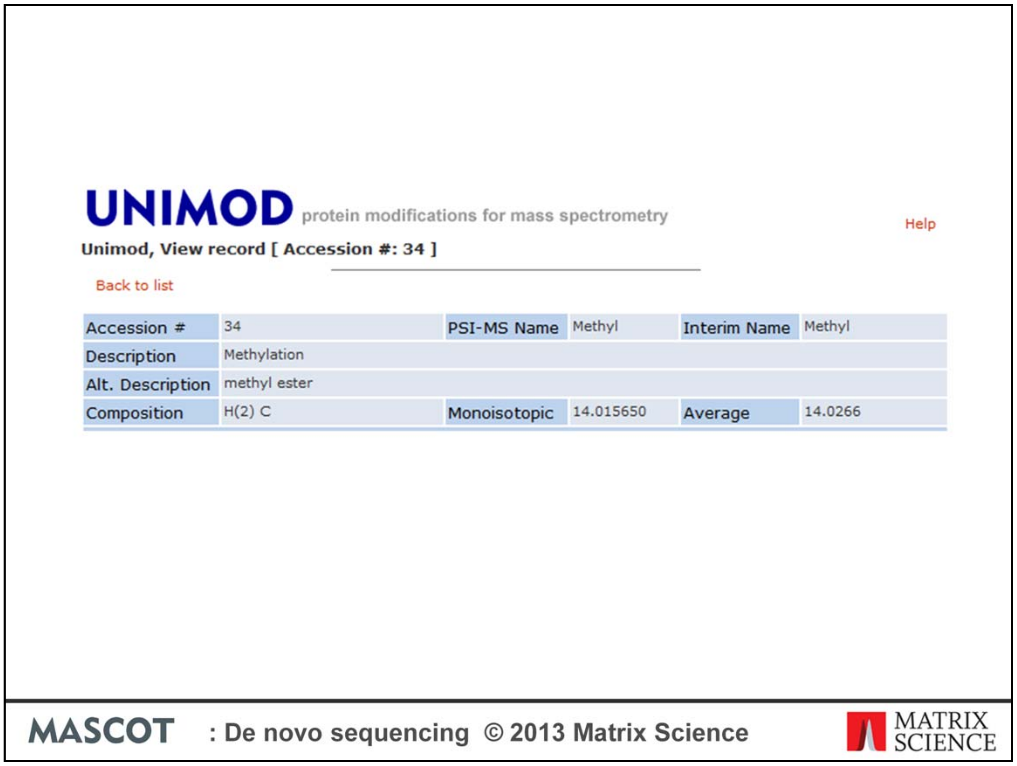
Looking at the protein view, our match runs from residues 21 to 30. Fortunately, I searched SwissProt so the protein hit has annotation data.



```
DR   SUPFAM; SSF50465; Elong_init_C; 1.
DR   SUPFAM; SSF50447; Translat_factor; 1.
DR   TIGRFAMs; TIGR00483; EF-1_alpha; 1.
DR   PROSITE; PS00301; EFACTOR_GTP; 1.
PE   1: Evidence at protein level;
KW   3D-structure; Actin-binding; Complete proteome; Cytoplasm;
KW   Cytoskeleton; Direct protein sequencing; Elongation factor;
KW   GTP-binding; Methylation; Nucleotide-binding; Phosphoprotein;
KW   Protein biosynthesis; Reference proteome.
FT   CHAIN         1    458       Elongation factor 1-alpha.
FT                                /FTId=PRO_0000090973.
FT   NP_BIND      14     21       GTP.
FT   NP_BIND      91     95       GTP.
FT   NP_BIND     153    156       GTP.
FT   SITE        298    298       Not modified.
FT   SITE        372    372       Not modified.
FT   MOD_RES       1      1       Blocked amino end (Met).
FT   MOD_RES       6      6       Phosphoserine.
FT   MOD_RES      30     30       N6-methyllysine; by EFM1 (Probable).
FT   MOD_RES      53     53       Phosphoserine.
FT   MOD_RES      72     72       Phosphothreonine.
FT   MOD_RES      79     79       N6,N6,N6-trimethyllysine.
FT   MOD_RES     163    163       Phosphoserine.
FT   MOD_RES     259    259       Phosphothreonine.
FT   MOD_RES     316    316       N6,N6-dimethyllysine; by SEE1 (Probable).
FT   MOD_RES     355    355       Phosphotyrosine.
FT   MOD_RES     356    356       Phosphoserine.
FT   MOD_RES     390    390       N6-methyllysine; by EFM1 (Probable).
FT   MOD_RES     394    394       Phosphoserine.
FT   MOD_RES     414    414       Phosphoserine.
FT   MOD_RES     430    430       Phosphothreonine.
FT   MOD_RES     458    458       Lysine methyl ester.
FT   MUTAGEN     122    122       E->K: Reduces interaction with YEF3.
FT   MUTAGEN     153    153       N->D: Increases KM for GTP to 2.7 mM.
FT   MUTAGEN     153    153       N->T: Increases KM for GTP to 6.0 mM and
FT                                reduces translation fidelity. Increases
FT                                Km for GTP to 10.3 mM and reduces
FT                                translation fidelity; when associated
```

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

Which suggests that Lysine 30 is Methylated.

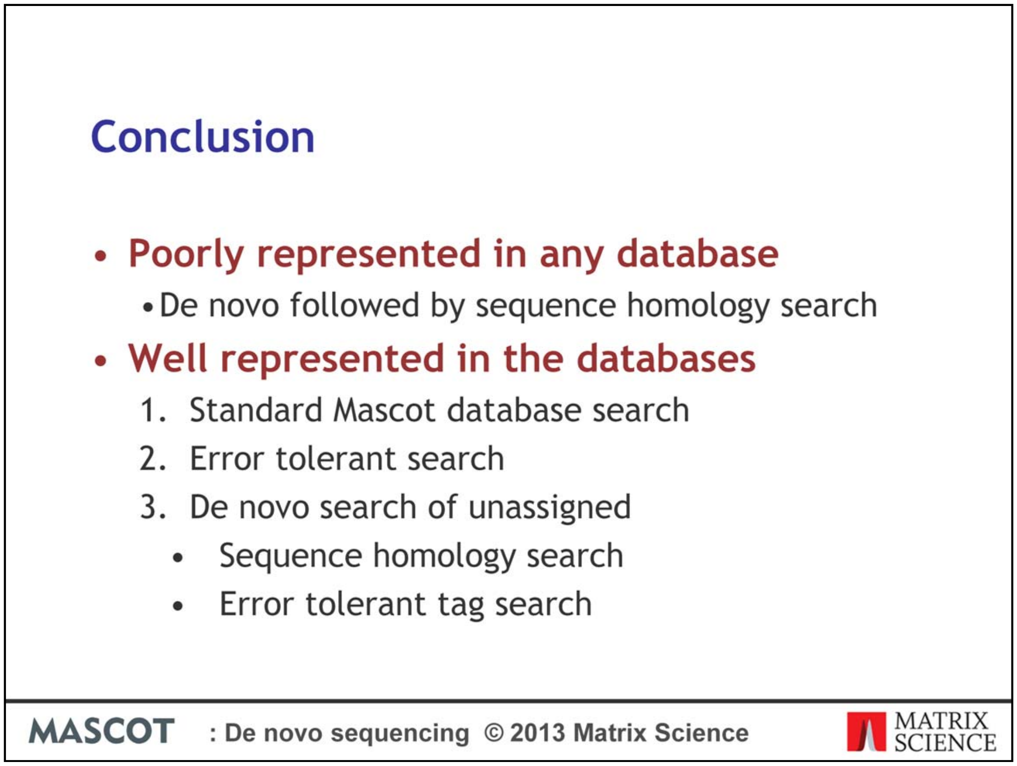


UNIMOD protein modifications for mass spectrometry

Unimod, View record [Accession #: 34]

Accession #	34	PSI-MS Name	Methyl	Interim Name	Methyl
Description	Methylation				
Alt. Description	methyl ester				
Composition	H(2) C	Monoisotopic	14.015650	Average	14.0266

A quick trip to Unimod shows that the mass shift associated with Methylation is +14.01Da, which is well within the mass tolerance for this search for the suggested mass shift of 13.9902 Da from the error tolerant tag search – so this seems like a reasonable assignment



Conclusion

- **Poorly represented in any database**
  - De novo followed by sequence homology search
- **Well represented in the databases**
  1. Standard Mascot database search
  2. Error tolerant search
  3. De novo search of unassigned
     - Sequence homology search
     - Error tolerant tag search

MASCOT : De novo sequencing © 2013 Matrix Science MATRIX SCIENCE

If you are trying to get as much coverage as possible over your data, you might come up with a search strategy like this:

If the organism you are working on is poorly represented and there isn’t a closely related species in the database, then de novo followed by a BLAST or MS-BLAST search (if you get a reasonable length region of clean sequence) is probably your best option. However, if you have protein sequences available, then you’re best off starting with a standard Mascot search, followed by an error tolerant search – this will get all of the ‘easy’ spectra. For the unassigned spectra, carry out a de novo search followed by BLAST or MS-BLAST. You can also try an etag search to find matches to isolated peptides that have a SNP or unsuspected modification.