

A Strategy for the Efficient Identification of Modified Peptides

MASCOT

{MATRIX}
{SCIENCE}

Aim

To get matches to “additional” modified peptides

- Unsuspected chemical or post-translational modifications
- Minor sequence variants, such as SNPs
- Products of non-specific cleavage

Not intended for

- Matching peptides from a poorly represented or unsequenced genome - *de novo*
- Specific modifications, e.g. phosphorylation - *targeted experiment*

I want to discuss the most efficient way to get as many matches as possible from an LC-MS/MS run. That is, to find matches to peptides with unsuspected chemical or post-translational modifications, with minor sequence variants, such as SNPs, and peptides which are the products of non-specific cleavage.

This strategy is not applicable to identifying peptides from protein that have little similarity to those in the database. As the databases fill up, this is becoming less common. If you are in this situation, the primary tool has to be *de novo* of high quality MS/MS spectra

Neither is it applicable to investigations focused on a particular modification, such as phosphorylation. The key here would be a targeted experiment. Maybe using a neutral loss scan to identify and select the phosphorylated peptides, or an IMAC column to isolate the phosphopeptides

Search strategy

- 1. Standard Mascot search**
Returns the easy matches
- 2. Error tolerant search**
Returns additional matches, but only for proteins where we have at least one good peptide match already
Limited to a single additional SNP or modification per peptide
- 3. De novo**
If data very high quality, can return novel full-length peptide sequences
Use **Blast** to find likely parent proteins
More often, returns partial / ambiguous peptide sequences
- 4. Error tolerant tag search**
To find matches to
 1. Isolated peptides that have a SNP or unsuspected modification
 2. Peptides with multiple SNPs or unsuspected modifications(No reason to expect additional matches from a **standard tag search**)

MASCOT : *Modified Peptides*

© 2006 Matrix Science

MATRIX
SCIENCE

If you simply want to get as many identifications as possible, so as to minimise the number of unmatched spectra and maximise protein coverage, you might come up with a strategy similar to this. I'll now go through the four steps in some detail. Step 1 is, of course, a standard Mascot MS/MS ions search.

1. Standard Mascot search

Why not try to get everything in a single search?

plc dataset on dual processor 2.8 GHz P4				
CLE	peptides tested	minutes	identity matches	average threshold
trypsin	7.5E+07	10	399	41
semi-trypsin	1.2E+09	127	379	53
none	1.0E+10	1067	299	62

Above timings are for Oxidation (M) as only variable modification.
If also include Phospho (STY), Me-ester (DE), Me-ester (C-term), Acetyl (N-term), pyro-glu (N-term Q), pyro_gln (N-term E), the trypsin time goes from 10 mins to 1357 mins!

MASCOT : *Modified Peptides*

© 2006 Matrix Science

MATRIX
SCIENCE

Should we try to get as many matches as possible in the first pass search? Well, let's look at some numbers for a typical ion trap dataset when we search using loose trypsin, semi-specific trypsin, and no enzyme specificity

As you can see, the search time increases by an order of magnitude as we go from trypsin to semi-specific trypsin, and a further order of magnitude as we go to a completely non-specific search.

The reason is simple, the search space, that is the number of candidate peptides, is increasing by a factor of 10 each time. Having to wait 10 or 100 times as long for the results is bad enough. A more fundamental problem is that the significance threshold score is a function of the number of candidates, so this increase by 10 each time, and we lose marginal matches. Unless you have a high level of non-specific peptides in the sample, you lose more than you gain.

So, doing a no-enzyme search in Mascot is not a good idea unless there is a very high level of non-specific peptides. Semi-trypsin is almost always a better choice if the peptides came from a tryptic digest. Only use no enzyme if the peptides are not the products of a deliberate enzyme digest, e.g. MHC peptides or endogenous peptides.

Identical considerations apply to modifications. If we go from 1 variable mod to 7, the search time is even worse than for no enzyme. This is because of the combinatorial explosion. Having to test all the combinations and permutations of these variable mods.

So, the answer is no, do not try to get as many matches as possible in the first pass search. It just makes the search very slow and very insensitive.

Search strategy

1. **Standard Mascot search**
Returns the easy matches
2. **Error tolerant search**
Returns additional matches, but only for proteins where we have at least one good peptide match already
Limited to a single additional SNP or modification per peptide
3. **De novo**
If data very high quality, can return novel full-length peptide sequences
Use Blast to find likely parent proteins
More often, returns partial / ambiguous peptide sequences
4. **Error tolerant tag search**
To find matches to
 1. Isolated peptides that have a SNP or unsuspected modification
 2. Peptides with multiple SNPs or unsuspected modifications(No reason to expect additional matches from a standard tag search)

MASCOT : *Modified Peptides*

© 2006 Matrix Science

MATRIX
SCIENCE

If you want to get as many identifications as possible, as efficiently as possible, the first pass search must be kept simple. Usually, strict or loose trypsin. Zero or one variable modifications. Certainly not more than two unless you know for sure they really are present.

Step 2 of our strategy is an error tolerant search. This is the efficient way to find unusual modifications, as well as variations in the primary sequence and peptides from non-specific cleavage

2. Error tolerant search

First pass - simple search of entire database

- Minimal modifications
- Enzyme specificity

Second pass - exhaustive search of selected protein hits

- Wide range of modifications
- Look for SNPs
- Relax enzyme specificity

MASCOT : *Modified Peptides*

© 2006 Matrix Science

MATRIX
SCIENCE

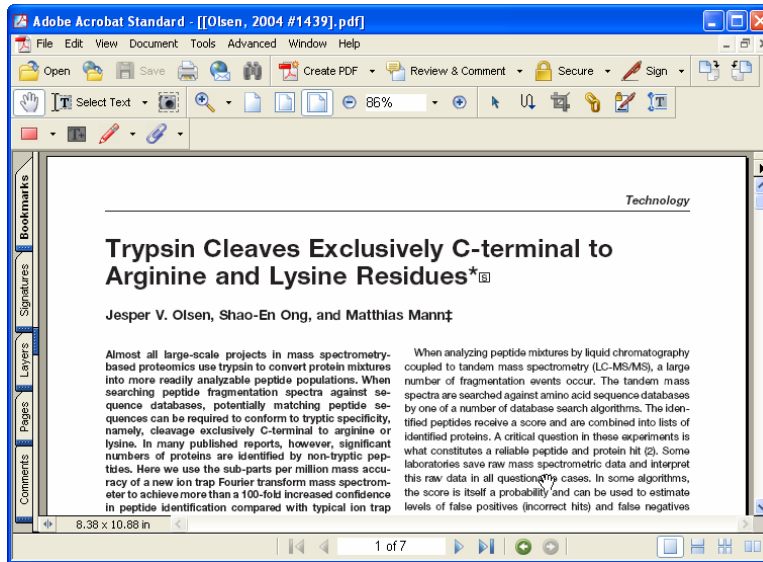
All the protein hits found in the first pass search are selected for an exhaustive second pass search.

Because only a handful of entries are being searched, search time is not an issue.

For modifications, an error tolerant search looks for one unsuspected modification per peptide in addition to those mods specified as fixed or variable. This is sufficient because it will be very, very rare to get two unsuspected mods on a single peptide.

The error tolerant search also looks for sequence variants, such as single nucleotide polymorphisms (SNPs) or sequencing errors.

You can remove enzyme specificity completely, but you have to ask yourself whether you would believe a match that was doubly non-specific.



Olsen, J. V., Ong, S.-E. and Mann, M., *Mol. and Cellular Proteomics*, 3, 608-14 (2004)

MASCOT : Modified Peptides

© 2006 Matrix Science

MATRIX
SCIENCE

I think in most cases the answer is no. Our experience is that the levels of non-specific peptides are very low, less than 3%, unless there is something seriously wrong with the trypsin or the protocol. This is also the conclusion of a very careful study by Matthias Mann's group. So, in general, I prefer to use semi-trypsin in an error tolerant search

The screenshot shows the Mascot Peptide Summary Report interface. At the top, there is a histogram of 'Probability Based House Score' with a peak around 100. Below the histogram, the 'Peptide Summary Report' section includes search parameters like 'Significance threshold p < 0.05' and 'Max. number of hits AUTO'. There are radio buttons for 'Standard scoring' and 'MudPIT scoring', and a 'Show pop-ups' checkbox. A 'Select All' button is present, and the 'Error tolerant' checkbox is checked. Below this, a table of search results is displayed with columns for 'Query', 'Observed', 'Mr (expt)', 'Mr (calc)', 'Delta Miss Score', 'Expect Rank', and 'Peptide'. The first row of the table is highlighted, and the 'Error tolerant' checkbox is checked.

1. Standard search

2. Check Error tolerant

3. Select hits

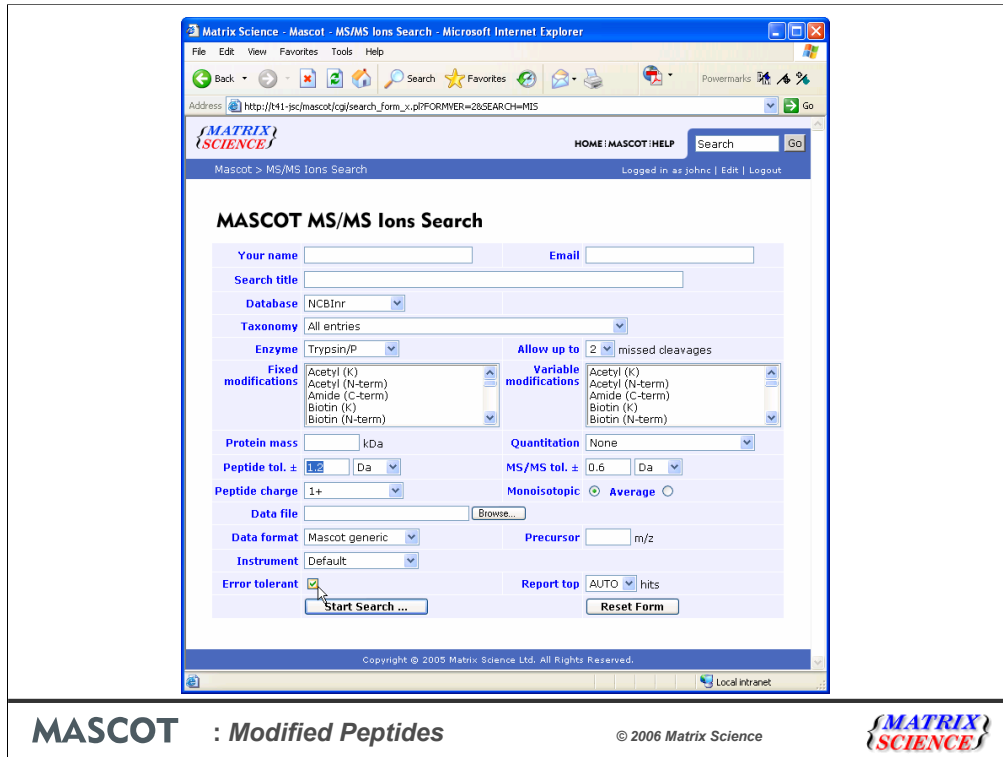
4. Search selected

MASCOT : Modified Peptides

© 2006 Matrix Science

MATRIX SCIENCE

In the current version of Mascot, an error tolerant search is literally a second pass search. You have to select the protein hits you want to search by manually checking them off in the results report



In Mascot 2.2, we will make the process integrated and automatic. You just have to check the Error tolerant box on the search form. This will perform a first pass search using the enzyme and modifications you specify in the search form. It will then automatically perform an error tolerant search on all of the proteins that contain significant peptide matches. David will say more about this in a later talk

Peptide Summary Report (Raft - 8) - Microsoft Internet Explorer

Address http://www.matrixscience.com/cgi/master_results.pl?file=

QTof data courtesy of Jyoti Choudhary, Sanger Centre

158	805.1681	1608.3216	1608.6508	-0.3	1	71	0.0051	1	G.WGNTKSSGTSYPPDLK.C + [+76.0313 at K5]
196	606.1852	1815.5339	1814.8839	0.6500	1	64	0.03	1	G.EDNINVVVEGHEQFISASK.S + [-18.0106 at N-term E]
221	987.7766	1973.5386	1973.9330	-0.3943	0	64	0.03	1	R.LGEDNINVVVEGHEQFISASK.S
240	1081.7681	2161.5216	2162.0490	-0.5274	0	157	1.6e-11	1	R.LGEDNINVVVEGHEQFISASK.S
241	721.5400	2161.5982	2162.0490	-0.4509	0	(93)	4.1e-05	1	R.LGEDNINVVVEGHEQFISASK.S
244	729.5355	2185.5846	2186.1290	-0.5444	0	113	3.3e-07	1	L.GEDNINVVVEGHEQFISASK.S + [+137.1640 at N-term G]
245	1094.8112	2187.6078	2188.0283	-0.4205	0	(102)	4.6e-06	1	R.LGEDNINVVVEGHEQFISASK.S + Acetyl (N-term): [-16.0313 at I]
247	1102.8030	2203.5914	2204.0596	-0.4682	0	(106)	1.8e-06	1	R.LGEDNINVVVEGHEQFISASK.S + [+42.0106 at N-term L]
248	735.5400	2203.5981	2203.9691	-0.3710	0	(77)	0.0015	1	L.GEDNINVVVEGHEQFISASK.S + [+155.0041 at E2]
250	740.5353	2218.5842	2219.0705	-0.4863	0	(100)	8.1e-06	1	R.LGEDNINVVVEGHEQFISASK.S + [+57.0215 at C-term K]
251	1110.2998	2218.5850	2219.0705	-0.4855	0	(114)	2.6e-07	1	R.LGEDNINVVVEGHEQFISASK.S + [+57.0215 at C-term K]
255	758.5516	2272.6331	2273.1361	-0.5031	0	69	0.009	1	K.SIVHPSPNSHTLNNDIHLIK.L + [+0.9840 at N10]

Proteins matching the same set of peptides:

[P167549](#) Mass: 24662 Score: 892 Queries matched: 22
 Trypsin (EC 3.4.21.4) precursor - bovine

[P113096615](#) Mass: 24563 Score: 892 Queries matched: 22
 Chain A, Bovine Beta-Trypsin Bound To Para-Amidino Schiff-Base Copper (Ia) Chelate

[P176615880](#) Mass: 26439 Score: 892 Queries matched: 22
 PREDICTED: similar to Cationic trypsin III precursor (Pretrypsinogen III) isoform 1 [Bos taurus]

[P15542503](#) Mass: 25388 Score: 892 Queries matched: 22
 Chain A, Trypsin Inhibitors With Rigid Tripeptidyl Aldehydes

[P1559311](#) Mass: 26093 Score: 892 Queries matched: 22
 pancreas cationic pretrypsinogen [Bos taurus]

[P11421532](#) Mass: 24659 Score: 891 Queries matched: 22
 Trypsinogen-Ca From Peg

[P161873128](#) Mass: 26453 Score: 890 Queries matched: 22
 PREDICTED: similar to Cationic trypsin III precursor (Pretrypsinogen III) isoform 1 [Bos taurus]

Possible Assignments:
 N->D [+0.9840]
 Glyc-Asn (N) [+0.9840]
 Deamidation (NQ) [+0.9840]

MASCOT : Modified Peptides © 2006 Matrix Science MATRIX SCIENCE

We need to do more work on trying to filter out the unlikely matches. One rule we plan to introduce is that you can't have an unsuspected modification in a non-specific peptide. Another is to ignore a modification if it only gives a tiny increase in score over an unmodified and specific peptide. It will still be necessary to decide between alternative assignments of observed mass differences

Fixed modifications: Carbamidomethyl (C)
 Variable modifications: Acetyl (N-term), Oxidation (M)
 Cleavage by Trypsin/P: cuts C-term side of KR
 Sequence Coverage: 41%

Matched peptides shown in **Bold Red**

1 **IIPVEEENPD** FWNREAEAL GAAKELQPAQ TAAKRLIIFL **GDGMGVSTVT**
 51 **AARILK**GQKK DELGPEIPLA MDRFPYVALS KTYNVDKHVP DSGATATAYL
 101 **CGVGNFQTI** GLSAAARFNQ CNTRRGNVEVI SVMNRKAKAG KSVGVVITTR
 151 **VQHAS**PAGTY AHYVRRHWYS DADVPASARQ EGCQDIATQL ISNMDIDVIL
 201 **GGRR**TFHF GTDPEEFDQ YSQGTSLDLC KNLQGVMLAK FGGARTVWR
 251 TELMQASLDP SVTHMLCLPE FGDNRYEIER DSTLDPSLHE NTEALLRLLS
 301 RNPFGFFLFV EGGRIDHGH ESRATRALTE **TDFDDAIER** AGQLTSEEDT
 351 **LSLV**TADSH VFSFGVPLR GSSIFGLAPG KARDRKATYV LLYGNPGTV
 401 **LKD**GARDVT ESESGSPEYR QQSAPLDEE **THAGEDVAVF** ARGPQALVH
 451 GVQECTFIAR VHAFAACLEP YTACDLAPPA GTTDAHPGR SVFPALLPLL
 501 AGTLLLETA TAP

Show predicted peptides also

Sort Peptides By Residue Number Increasing Mass Decreasing Mass

Start - End	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Sequence
1 - 14	586.4950	1756.4633	1756.8420	-0.3787	0	--IIPVEEENPDFWNR.E (Ions score 60)
1 - 14	879.2425	1756.4705	1756.8420	-0.3715	0	--IIPVEEENPDFWNR.E (Ions score 100)
35 - 53	975.8103	1949.6060	1950.0244	-0.4183	0	K.NLIFLGGDGHVSVTAAAR.I Oxidation (M) (Ions score 92)
105 - 117	653.2101	1304.4056	1304.6836	-0.2780	0	K.GNFQITGLSAAAR.F (Ions score 95)
151 - 166	427.8731	1707.4632	1707.8441	-0.3809	0	R.VQHASPAGTYAHYVNR.H (Ions score 80)
167 - 179	726.1806	1450.3465	1450.6477	-0.3011	0	R.HVYSADVPASAR.Q (Ions score 74)
180 - 204	901.5943	2701.7611	2702.3003	-0.5392	0	R.QEGCQDIATQLISNMDIDVILGGGR.K (Ions score 66)
210 - 227	1001.2025	2000.3904	2000.8058	-0.4133	0	R.HGTPDDEYPPDYSQGGTR.L Oxidation (M) (Ions score 82)
210 - 227	667.8045	2000.3918	2000.8058	-0.4139	0	R.HGTPDDEYPPDYSQGGTR.L Oxidation (M) (Ions score 80)
327 - 340	829.7282	1639.4419	1639.7763	-0.3344	0	R.ALTTDFDDAIER.A Oxidation (M) (Ions score 101)
341 - 370	809.2359	3232.9146	3233.5628	-0.6483	0	R.AGQLTSEEDTSLSVTADSHVFSFGVPLR.G (Ions score 73)
371 - 381	517.1760	1032.3374	1032.5603	-0.2229	0	R.GSSIFGLAPG.A (Ions score 76)
403 - 420	651.1554	1950.4443	1950.8555	-0.4112	1	K.DGARDVTESESGSPEYR.Q (Ions score 68)
421 - 442	790.2187	2367.6342	2368.1294	-0.4952	0	R.QQSAPLDEETHAGEDVAVFAR.G (Ions score 106)

MASCOT : Modified Peptides

© 2006 Matrix Science



The good news about the error tolerant search is that it substantially increases the number of matches. In this particular hit, from 14 to 22.

Fixed modifications: Carbamidomethyl (C)
 Variable modifications: Acetyl (N-term), Oxidation (M)
 Semi-specific cleavage, (peptide can be non-specific at one terminus only)
 Cleavage by semi-Trypsin: cuts C-term side of KR unless next residue is P
 Sequence Coverage: 42%

Matched peptides shown in **Bold Red**

1 IIPVEENPD FWNREAEAL GAARKLQPAQ TAARHLIIFL GDGHGVSTVT
51 AARILKQEK DLGPEEPLA DRFFVVALS KTYVDKHP DSGATATYL
101 CGUNGFQTI GLSAAARFQK CHITRQNEVI SVNRAKKG KSVGVITTR
151 VQHASPAGY AHTVHRWYS DADVPASARQ EGCQDIATQL ISNMDIVIL
201 GGGKYMFRM GTPDPEYDD YSGGGTRLDG KMLVQEWLAK RQGARYVNR
251 TELMQASLDP SVTHLMLGFE PDMKYEIHR DSTLDPSLME HTEAALRLLS
301 RNPFGFLV EGGRIDHGH ESRAVYALTE TDFDDAIER AGQLTSEEDT
351 LSLVADHSH VFSGGVPLR GSSIFGLAPK KARDKATTV LLYGMQFTV
401 LKDGARDDVT ESESGSPEYR QQSAVPLDEE THAGEDVAVF ARGFQHLVH
451 GVQEQTFIAH VMAFAACLEP YTACDLAPPA GTTDAHPGR SVVFPALLPLL
501 ACTLLLETA TAP

Sort Peptides By Residue Number Increasing Mass Decreasing Mass

Start - End	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Sequence
1 - 14	879.2425	1756.4705	1756.8420	-0.3715	0	-IIPVEENPDFWNR.E (Ions score 100)
35 - 53	975.8103	1949.6060	1950.0244	-0.4183	0	K.NLIIFLGDGHGVSTVTAAR.I Hydroxylation (D) [+15.99] (Ions score 92)
104 - 117	710.2234	1418.4323	1418.7266	-0.2943	1	V.KGNFQITGLSAAAR.F K->N [-14.05] (Ions score 92)
106 - 117	653.2101	1304.4056	1304.6836	-0.2780	0	G.NFQITGLSAAAR.F Carbamidomethyl (N-term) [+57.02] (Ions score 99)
132 - 166	427.8731	1787.4632	1787.8230	-0.3597	0	V.QHASDAGFYAHVNR.H S->W [+99.05] (Ions score 84)
167 - 179	726.1806	1430.3465	1430.6477	-0.3011	0	R.HWYSADVPASAR.Q (Ions score 79)
180 - 204	901.5943	2701.7611	2702.3003	-0.5392	0	R.QEGCQDIATQLISNMDIVILGGGR.K (Ions score 66)
205 - 227	681.1543	2720.5880	2720.1482	0.4398	2	R.KVMFRMGTPDPEYDDYSGGTR.L Acetyl (N-term); Oxidation (M); Y->D [-48.04] (Ions score 82)
210 - 227	1001.2025	2000.3904	2000.8058	-0.4153	0	R.MGTPDPEYDDYSGGTR.L Oxidation (M) (Ions score 82)
210 - 227	667.8045	2000.3918	2000.8058	-0.4139	0	R.MGTPDPEYDDYSGGTR.L Oxidation (M) (Ions score 80)
327 - 340	526.1536	1575.4390	1575.7814	-0.3424	0	R.ALTEYTHEDDAIER.A F->V [-48.00] (Ions score 77)
327 - 340	820.7282	1639.4419	1639.7763	-0.3344	0	R.ALTEYTHEDDAIER.A Oxidation (M) (Ions score 101)
341 - 370	809.2359	3232.9146	3232.6152	0.2994	0	R.AGQLTSEEDVLSLVADHSHVSPGGTR.G E->K [-0.95] (Ions score 73)
370 - 381	545.6818	1089.3491	1089.5818	-0.2327	1	L.RGSSIFGLAPGK.A R->G [-59.08] (Ions score 63)
371 - 381	517.1760	1032.3374	1032.5603	-0.2229	0	R.GSSIFGLAPGK.A (Ions score 76)
371 - 381	532.1836	1062.3527	1062.5709	-0.2182	0	R.GSSIFGLAPGK.A G->S [+30.01] (Ions score 65)
403 - 420	651.1554	1950.4443	1950.8355	-0.4112	0	K.DGARPDVTESESGSPEYR.Q (Ions score 68)
403 - 420	656.1752	1965.5039	1964.8711	0.6327	0	K.DGARPDVTESESGSPEYR.Q Me-ester (DE) [+14.02] (Ions score 74)
403 - 420	670.1561	2007.4464	2007.8769	-0.4306	0	K.DGARPDVTESESGSPEYR.Q Carbamidomethyl (D) [+57.02] (Ions score 84)
421 - 442	766.2128	2295.6167	2296.1083	-0.4916	0	R.QQSAVPLDEETHAGEDVAVFAR.G E->G [-72.02] (Ions score 61)
421 - 442	784.5441	2330.6104	2331.1029	-0.4925	0	R.QQSAVPLDEETHAGEDVAVFAR.G Pyro-glut (N-term Q) [-17.03] (Ions score 78)
421 - 442	790.2187	2367.6342	2368.1294	-0.4952	0	R.QQSAVPLDEETHAGEDVAVFAR.G (Ions score 105)
421 - 442	809.2209	2424.6408	2425.1509	-0.5101	0	R.QQSAVPLDEETHAGEDVAVFAR.G Carbamidomethyl (N-term) [+57.02] (Ions score 78)

MASCOT : Modified Peptides

© 2006 Matrix Science



The bad news is that the sequence coverage hardly changes, 41% to 42%

Sequence Coverage: 41%

Matched peptides shown in **Bold Red**

```

1 IIPVEEENPD FWNREAAEAL GAAKKLQPAQ TAAKNIIFI GDGMGVSTVT
51 AARILRGQKK DKLGP EIPLA MDRFPYVALS KTYNVDRKVP DSGATATAYL
101 CGVRGNFQTI GLSAAARFNQ CNTTRGNEVI SVMNRKKAG KSVGVTTR
151 VOHASTAGTY AHTVNRWFYS DADVPASARQ EGCQDIATQL ISNDDIDVIL
201 GGRKYMFRM ITPDEYRDD YSQGGTRLDG KNLVQEWLAK RQGARYVWNR
251 TELVQASLDL SVTHLMGLFE PGDMKYEIHR DSTLDPSLME MTEAALRLLS
301 RNPRGFLLFV EGGRIDHGHH ESRAYRALTE TIMFDDAIER AGQLTSEEDT
351 LSLVTADHSH VESEGGYPLR GSSIFGLAPG KARDRKAYTV LLYGNPGYV
401 LKDGARPDVT ESESGSPEYR QSSAVPLDEE THAGEDVAVF ARGPQHLVH
451 GVQEQTPIAH VMAFAACLEP YTACDLAPPA GTTDAHPGR SVVPALLPLL
501 AGTLLLLLETA TAP

```

Sequence Coverage: 42%

Matched peptides shown in **Bold Red**

```

1 IIPVEEENPD FWNREAAEAL GAAKKLQPAQ TAAKNIIFI GDGMGVSTVT
51 AARILRGQKK DKLGP EIPLA MDRFPYVALS KTYNVDRKVP DSGATATAYL
101 CGVRGNFQTI GLSAAARFNQ CNTTRGNEVI SVMNRKKAG KSVGVTTR
151 VOHASTAGTY AHTVNRWFYS DADVPASARQ EGCQDIATQL ISNDDIDVIL
201 GGRKYMFRM ITPDEYRDD YSQGGTRLDG KNLVQEWLAK RQGARYVWNR
251 TELVQASLDL SVTHLMGLFE PGDMKYEIHR DSTLDPSLME MTEAALRLLS
301 RNPRGFLLFV EGGRIDHGHH ESRAYRALTE TIMFDDAIER AGQLTSEEDT
351 LSLVTADHSH VESEGGYPLR GSSIFGLAPG KARDRKAYTV LLYGNPGYV
401 LKDGARPDVT ESESGSPEYR QSSAVPLDEE THAGEDVAVF ARGPQHLVH
451 GVQEQTPIAH VMAFAACLEP YTACDLAPPA GTTDAHPGR SVVPALLPLL
501 AGTLLLLLETA TAP

```

2720.1482	0.4398	2	68	0.01	6	R.KYHFRMGTPDEYRDDYSQGGTR.L + Acetyl (N-term); Oxidation (M): [-48.0364 at Y2]
3232.6152	0.2994	0	73	0.0031	1	R.AGQLTSEEDTSLVTDHSHRVFSGGYPLR.G + [-0.9476 at E8]

et of peptides:

1 Score: 1036 Queries matched: 23
 H A Human Phosphatase

Possible Assignments

Y->N [-49.0204]
 Y->D [-48.0364]

MASCOT : Modified Peptides © 2006 Matrix Science **MATRIX SCIENCE**

If we look more closely, we see that this is just an additional 5 residues here, KYMFR. However, when we look at the matches, there is only one match spanning this peptide and it is a weak and dubious match, requiring two mods and a SNP. In all honesty, I would not want to accept this match, so the coverage is actually unchanged. For this particular protein, the error tolerant search just gives us additional matches to the same peptides we saw in the standard search.

Peptide Summary Report (yeast_500K_Orb_50K_Orb_MS2_copy.RAW Error tolerant) - Microsoft Internet Explorer

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

Orbitrap data courtesy of Steve Gygi, Harvard Medical School

2174 905.1237
 2182 1364.1895 2726.3644 2726.3550 0.0094 1 69 3.1e-06 1 K.KSEVFSTYADNQPGLVLIQVFEGEER.A + [+14.0156 at N-term K]
 2183 909.7962 2726.3667 2726.3551 0.0116 1 (65) 0.00066 1 K.KSEVFSTYADNQPGLVLIQVFEGEER.A + [+14.0157 at V4]
 2208 924.1313 2769.3721 2769.3609 0.0112 1 (46) 0.054 3 K.KSEVFSTYADNQPGLVLIQVFEGEER.A + [+57.0215 at E3]

3. [q1|6324707](#) Mass: 93686 Score: 1120 Queries matched: 38
 Elongation factor 2 (EF-2), also encoded by EFT2; catalyzes ribosomal translocation during protein synthesis; contains diphth
 Check to include this hit in error tolerant search

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/> 42	365.7276	729.4406	729.4384	0.0022	0	57	0.0028	1	R.AGIISAAR.A
<input checked="" type="checkbox"/> 133	422.2563	842.4980	842.4974	0.0007	1	57	0.003	1	V.VIKRVDLR.A
<input checked="" type="checkbox"/> 292	486.8060	971.5975	971.5950	0.0026	1	54	0.0039	1	K.ALLKRVHR.K + [+43.0058 at K4]
<input checked="" type="checkbox"/> 441	542.7985	1083.5824	1083.5786	0.0038	0	45	0.052	1	R.LFTADHFK.K
<input checked="" type="checkbox"/> 494	555.2761	1108.5375	1108.5335	0.0041	0	76	4.5e-05	1	N.VAFIVDQHR.S + [+43.0058 at N-term V]
<input checked="" type="checkbox"/> 508	560.3054	1118.5963	1118.5931	0.0032	0	76	5e-05	1	K.STLTDQLVR.A
<input checked="" type="checkbox"/> 516	563.2738	1124.5331	1124.5284	0.0047	0	(52)	0.01	1	N.VAFIVDQHR.S + Oxidation (M); [+43.0058 at N-term V]
<input checked="" type="checkbox"/> 561	582.7733	1163.5321	1163.5280	0.0041	0	67	0.00032	1	K.EGRIFGEHR.S
<input checked="" type="checkbox"/> 564	582.8245	1163.6345	1163.6298	0.0047	1	46	0.049	1	R.VFAGVVKSGQK.V + [+43.0058 at K7]
<input checked="" type="checkbox"/> 629	603.3207	1204.6269	1204.6234	0.0036	1	57	0.0046	1	R.SLMDKTVHR.N + [+43.0058 at K5]
<input checked="" type="checkbox"/> 633	605.8005	1209.5864	1209.5818	0.0046	0	63	0.001	1	R.LWGDSEFPK.I
<input checked="" type="checkbox"/> 725	629.3507	1256.6868	1256.6836	0.0032	1	84	7.1e-06	1	R.AGIISAARAGEAR.F + [+43.0058 at R8]
<input checked="" type="checkbox"/> 733	633.8427	1265.6708	1265.6655	0.0053	0	80	1.8e-05	1	R.ATYAGELADPK.I
<input checked="" type="checkbox"/> 785	654.3312	1306.6478	1306.6432	0.0046	0	74	0.0001	1	R.NMSVIAHVDHGR.S
<input checked="" type="checkbox"/> 883	690.4420	1378.8695	1378.8660	0.0036	1	51	0.002	1	R.LKPVVVVIRKVR.A
<input checked="" type="checkbox"/> 962	711.9451	1421.8757	1421.8718	0.0039	1	(47)	0.0067	1	R.LKPVVVVIRKVR.A + [+43.00
<input checked="" type="checkbox"/> 1010	731.9082	1461.8018	1461.8051	-0.0033	2	49	0.024	1	R.VFAGVVKSGQKVR.I + [+43.0
<input checked="" type="checkbox"/> 1031	741.8754	1481.7363	1481.7303	0.0060	1	55	0.0067	1	R.LWGDSEFPKTK.K + [+43.0058 at K10]
<input checked="" type="checkbox"/> 1034	742.3951	1482.7757	1482.7693	0.0064	0	65	0.00072	1	R.AFMHFILDFIPL.L
<input checked="" type="checkbox"/> 1041	744.4293	1486.8440	1486.8394	0.0045	0	101	1.2e-07	1	K.FSVSPVQVAVEVK.N

Possible Assignments:
 Carbamyl (K) [+43.0058]

MASCOT : Modified Peptides © 2006 Matrix Science MATRIX SCIENCE

If you don't want to spend a lot of time studying the matches from an error tolerant search and deciding which you accept, you can use it as a quick way of spotting whether there are modifications which should be included in the first pass search as variable mods. Here's a nice example. Lots of matches for a modification of 43 Da, almost certainly carbamylation.

Select Summary Report [yeast_500K_Orb_50K_MS2_copy.RAW] - Microsoft Internet Explorer

Back View Favorites Tools Help

http://www.matrixscience.com/cgi/blast_results.pl?db=.data/20060524NFkoCat.dat

1. [q11312152](#) Mass: 69786 Score: 1723 Queries matched: 57
 unnamed protein product [Saccharomyces cerevisiae]
 Check to include this hit in error tolerant search

Query	Observed	M(e)pt	M(e)alc	Delta	Miss	Score	Expect	Rank	Peptide
258	888.4760	881.4887	887.4600	0.0087	0	42	0.017	1	R.SLIPPEK.V 100 191 192 193
415	534.7660	1067.5174	1067.5135	0.0040	0	29	0.13	1	K.EIAESYLGAR.V
590	592.3320	1182.6511	1182.6397	0.0114	0	31	0.071	1	K.FELSGIPPAFR.G 588
618	600.3429	1190.6712	1190.6669	0.0042	0	93	5.8e-08	1	K.DADYTAGLHWLR.I
638	607.0118	1211.6091	1211.6051	0.0040	0	69	1.4e-05	1	R.VLLIARDQGR.T
721	633.3430	1264.6714	1264.6663	0.0052	0	61	0.0001	1	K.NULESIAYS.K.M
742	637.8505	1273.6865	1273.6819	0.0046	0	50	0.002	1	R.LVNHFIQEFK.R 741
770	650.3086	1290.6026	1290.5965	0.0062	0	71	5.3e-06	1	R.FEELCADLFR.S
824	672.3317	1342.6489	1342.6430	0.0051	1	63	4.8e-05	1	K.DYFARVELDAM.V 814 815
834	704.3164	1406.6182	1406.6136	0.0046	0	75	1.1e-06	1	R.HFHDPEYQAMK.H 844
971	715.9003	1429.7860	1429.7830	0.0030	1	56	0.0003	1	R.LVNHFIQEFK.R 972
1015	736.3597	1470.7949	1470.6991	0.0059	0	60	0.5e-05	1	R.YTSPFAVFDTER.L
1111	763.8973	1525.7398	1525.7347	0.0051	1	67	1.4e-05	1	R.ADFEELCADLFR.S 1110
1151	776.8948	1531.7250	1531.7281	0.0069	0	77	2.3e-06	1	R.VTFEYRKEKDR.S 1150
1171	783.8985	1565.7893	1565.7759	0.0065	0	52	0.00069	1	K.HFDEQISVMDGK.H
1260	812.3789	1622.7432	1622.7358	0.0074	0	106	1.1e-09	1	K.WQAMHNSHTVFDAK.R 1237
1304	830.4544	1658.8942	1658.8879	0.0063	0	107	1.8e-09	1	R.LIDEPFAAAIAYGLDK.R 1303
1310	955.6474	1663.9203	1663.9144	0.0059	1	25	0.18	1	R.IASDQLFESLAYS.K.H
1320	838.3724	1674.7302	1674.7233	0.0069	0	97	5.7e-09	1	K.ATAGDTHLGGEDFDR.L 1321
1385	864.9838	1727.9530	1727.9457	0.0073	1	83	3.4e-07	1	K.LLDVDRKPIQVEFK.G 1384
1414	882.4309	1762.8472	1762.8420	0.0052	1	120	1e-10	1	K.WQAMHNSHTVFDAK.L 1411
1474	894.5014	1786.9803	1786.9828	0.0055	1	111	4.4e-10	1	R.LTFEFAKAYGLDNG.G
1492	908.4983	1814.9821	1814.9737	0.0084	1	122	4.8e-11	1	K.LDSQYDELVLGGSTR.L 1498
1503	947.9723	1893.9300	1893.9250	0.0080	0	60	7.5e-05	1	K.VHDAPVTPATYFDQGR.G 1504
1803	715.3938	2143.1597	2143.1524	0.0073	2	64	2.2e-05	1	K.LLDVDRKPIQVEFGEIK.N 1804
1821	1002.0465	2162.0784	2162.0782	0.0002	2	39	0.0009	1	R.NTFEAGDLEQADDPVTK.N 1820
1892	723.0173	2166.0300	2166.0163	0.0136	0	37	0.011	1	K.AGIDLETTESVQAFDAK.V
1895	723.0700	2166.1881	2166.1796	0.0085	2	27	0.069	1	K.TKDNMLGRFELSGIPPAFR.G 1826
2101	1286.2066	2570.3986	2570.3876	0.0110	0	33	0.014	1	K.YDRLLLDPAFLSLGLETAGGVHTK.L 2102
2105	1289.1417	2576.2688	2576.2605	0.0083	0	144	2.6e-13	1	R.SIFPDQAVATDAVQAAILTQDESSK.T
2115	1300.1424	2598.2702	2598.2601	0.0101	0	100	6.1e-09	1	K.SEIFSTADNQQVLIQVEGER.A 2116
2182	1364.1895	2726.3644	2726.3551	0.0094	1	89	7.5e-08	1	K.KSEIFSTADNQQVLIQVEGER.A 2183

2. [q116324707](#) Mass: 93686 Score: 1448 Queries matched: 68
 Elongation Factor 2 (EF-2), also encoded by EFT2; catalyzes ribosomal translocation during protein synthesis; contains diphthamide, the unique proter

Check to include this hit in error tolerant search

Query	Observed	M(e)pt	M(e)alc	Delta	Miss	Score	Expect	Rank	Peptide
-------	----------	--------	---------	-------	------	-------	--------	------	---------

MASCOT : Modified Peptides © 2006 Matrix Science MATRIX SCIENCE

Adding this in as a variable mod increases the number of peptide matches for hit 1 from 57

Select Summary Report [yeast_500K_Orb_50K_MS2_copy.RAW] - Microsoft Internet Explorer

http://www.matrixscience.com/cgi/master_results.pl?seq..._data/20060524/FacOxIDE.d

1. [211312152](#) Mass: 69786 Score: 2247 Queries matched: 92
 unnamed protein product [Saccharomyces cerevisiae]
 Check to include this hit in error tolerant search

Query	Observed	M(e)pt	M(c)alc	Delta	Miss	Score	Expect	Rank	Peptide
225	888.4760	887.4687	887.4600	0.0087	0	42	0.001	1	R.STLDPVLR.K 190 191 192 193
227	502.7679	1003.5212	1003.5186	0.0027	1	37	0.19	1	R.LSSEDIK.M
229	530.7925	1059.5704	1059.5672	0.0032	1	60	0.00082	1	K.IITTRDEGR.L
215	534.7660	1067.5174	1067.5135	0.0040	0	29	0.51	1	K.KTASTYLAK.V
202	500.8163	1175.6181	1175.6146	0.0035	1	39	0.079	1	K.SREITTRDK.G
280	592.3320	1182.6311	1182.6397	0.0114	0	31	0.24	1	K.FELSGIPPAFR.G 588
218	600.3429	1198.6732	1198.6669	0.0062	0	93	2.14e-07	1	K.DAGTLAGLWLR.I
218	607.8118	1213.6091	1213.6051	0.0040	0	69	5.7e-05	1	R.VDTIARDQGR.I
221	633.3430	1264.6714	1264.6663	0.0052	0	61	0.00038	1	K.MDLSTSLK.W
242	637.8505	1273.6865	1273.6819	0.0046	0	50	0.0068	1	R.LVNHPIQEK.R 241
270	650.3086	1298.6026	1298.5965	0.0062	0	71	2.4e-05	1	R.FEELCADLFR.S
272	650.3695	1298.7245	1298.7194	0.0051	1	52	0.0028	1	R.STLDPVLR.D 271
224	672.3317	1342.6489	1342.6438	0.0051	1	63	0.0062	1	K.HPTASTYLAK.V 814 815 817 818
209	695.3786	1388.7426	1388.7372	0.0055	2	37	0.1	1	K.SREITTRDK.G 208 209
214	704.3164	1406.6189	1406.6136	0.0046	0	75	5.9e-06	1	R.HFNEPQVQAMR.H 211
271	715.9003	1429.7860	1429.7830	0.0030	1	56	0.0011	1	R.LVNHPIQEK.R 272 1017
1015	736.2597	1470.7849	1470.6991	0.0059	0	60	0.00059	1	R.VTSPVAFPTER.L
1111	763.8713	1525.7389	1525.7347	0.0053	1	67	7.4e-05	1	R.AEELCADLFR.S 1110
1151	776.8948	1551.7750	1551.7681	0.0069	1	77	9.8e-06	1	K.LVTFYVQKPRR.S 1150 1217
1171	783.8985	1565.7823	1565.7759	0.0065	0	52	0.0028	1	K.HTFPQISNVLGR.H
1213	401.2290	1600.8071	1600.8037	0.0033	2	5	90	1	R.LVNHPIQEKR.H
1240	812.3789	1622.7432	1622.7358	0.0074	0	106	6.3e-09	1	K.HQAMHPSHTVFQAK.R 1237
1304	830.4544	1658.8942	1658.8879	0.0063	0	107	6.7e-09	1	R.IIHPFAAALATGLDK.K 1303
1316	557.6462	1669.9167	1669.9111	0.0057	1	15	9.2	1	R.QATKDAITLAGLWLR.I
1310	838.3724	1674.7302	1674.7233	0.0069	0	97	2.5e-08	1	K.ATAGDTHLGGEDFDR.I 1311
1350	854.4710	1706.9275	1706.9202	0.0072	1	49	0.0029	1	R.IASDQLLESATSLK.H 1310 1356
1385	864.9838	1727.9530	1727.9457	0.0073	1	83	1.2e-06	1	K.IIHPVQKIQVEK.G 1384
1442	882.4309	1762.8472	1762.8420	0.0052	1	120	4.5e-10	1	K.HQAMHPSHTVFQAK.R 1441 1493 1503
1474	894.5814	1786.9883	1786.9828	0.0055	1	111	1.5e-09	1	R.IIHPFAAALATGLDK.K 1520 1521
1499	908.4983	1814.9821	1814.9737	0.0084	1	122	1.7e-10	1	K.LDKSQQEIVLVGGSTR.I 1498 1548
1541	618.3169	1851.9089	1851.9222	0.0066	1	40	0.041	1	K.HTFPQISNVLGR.H 1541
1583	947.9723	1893.9300	1893.9220	0.0080	0	60	0.00035	1	K.VHDAVTVVATLVDSGR.G 1584
1775	702.9980	2105.9721	2105.9628	0.0092	1	1	1.7e+02	1	R.HFNEPQVQAMHPSHTVFQAK.R
1803	715.3938	2143.1597	2143.1524	0.0073	2	64	0.3e-05	1	K.IIDVDKQKIQVEFQKTR.H 1804 1851 1852
1822	723.0173	2166.0300	2166.0163	0.0136	0	37	0.053	1	K.AVIGLDTTSCVQAMHQR.V
1825	723.0780	2166.1881	2166.1796	0.0085	2	27	0.26	1	K.FSDNLLGRFELSGIPPAFR.G 1824 1878 1879
1832	1087.0803	2172.1460	2172.1386	0.0075	2	98	4.7e-08	1	R.DAKLDSQQEIVLVGGSTR.I 1831
1874	736.0350	2205.0931	2205.0760	0.0071	2	65	0.00012	1	K.MTISEAGDRELVQDEPTVK.R 1820 1821 1875
1921	773.7316	2318.1429	2318.1324	0.0104	1	22	2.2	1	R.IIDVAAKQAMHPSHTVFQAK.R

MASCOT : Modified Peptides © 2006 Matrix Science MATRIX SCIENCE

To 92. However, coverage only increases from 57% to 64%. Basically, just one new peptide, that was only present in carbamylated form

“One of the main conclusions from this investigation is that as a general rule, chemically modified peptides are only likely to be correctly identified if they derive from the most abundant peptides in the sample.”

The screenshot shows a research article titled "Depth of Proteome Issues" by Kenneth C. Parker et al. The article is categorized as "Research" and is a "YEAST ISOTOPE-CODED AFFINITY TAG REAGENT STUDY". The authors listed are Kenneth C. Parker, Dale Patterson, Brian Williamson, Jason Marchese, Armin Graber, Feng He, Allan Jacobson, Peter Juhasz, and Stephen Martin. The abstract discusses a test case for optimizing proteomics experiments using a yeast model system with a UPF1 gene knockout. It details the use of ICAT experiments, MALDI-MS/MS, and electrospray MS/MS to assess peptide identification reproducibility. A sidebar on the left contains navigation links for "Home", "Signatures", "Layers", "Pages", and "Comments". The bottom of the screenshot shows a browser status bar with "1 of 35" pages.

Parker, K. C. et al., *Mol. and Cellular Proteomics*, 3, 625-59 (2004)

MASCOT : Modified Peptides

© 2006 Matrix Science

MATRIX
SCIENCE

This observation is in agreement with a very careful study by Ken Parker and colleagues. As you go deeper, you tend to find modified versions of peptides that you already identified. They did detailed manual validation of matches to non-specific peptides and found approximately 2% were semi-tryptic and zero were fully non-specific.

Search strategy

1. **Standard Mascot search**
Returns the easy matches
2. **Error tolerant search**
Returns additional matches, but only for proteins where we have at least one good peptide match already
Limited to a single additional SNP or modification per peptide
3. **De novo**
If data very high quality, can return novel full-length peptide sequences
Use **Blast** to find likely parent proteins
More often, returns partial / ambiguous peptide sequences
4. **Error tolerant tag search**
To find matches to
 1. Isolated peptides that have a SNP or unsuspected modification
 2. Peptides with multiple SNPs or unsuspected modifications(No reason to expect additional matches from a standard tag search)

MASCOT : *Modified Peptides*

© 2006 Matrix Science



Where can we go from here? Well, maybe there are some peptides which can't be picked up by the error tolerant search. Maybe a peptide that spans a splice site or a peptide with a modification that is not in our list of modifications. The next step is de novo.

De Novo - Issues

Database : NCBI nr 20060518 (3647739 sequences)
Enzyme : Trypsin/P
Fixed modifications : Carbamidomethyl (C)
Variable modifications : Acetyl (N-term), Oxidation (M)
Peptide Mass Tolerance : ± 0.8 Da
Fragment Mass Tolerance : ± 0.4 Da
Max Missed Cleavages : 2

Peptide mass	Database search	De novo
~ 1000 Da	~ 6×10^5	~ 1×10^8
~ 2000 Da	~ 5×10^5	~ 4×10^{19}

MASCOT : Modified Peptides

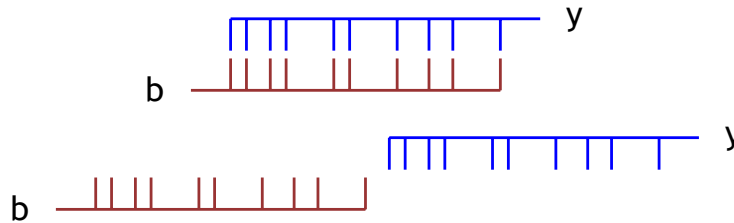
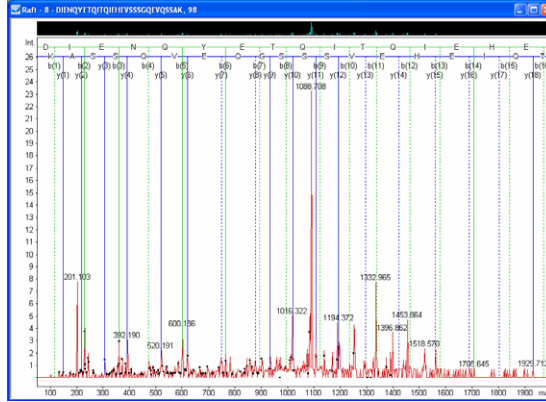
© 2006 Matrix Science

MATRIX
SCIENCE

The problem with de novo is that the search space is huge. If we assume tryptic specificity, the bigger the peptide, the fewer the candidates in a database search. With de novo, the number of candidate sequences grows geometrically with peptide length. In reality, things aren't so bad. Any practical de novo algorithm explores only a small portion of this search space. Nevertheless, you cannot expect to get de novo solutions from large peptides unless the signal to noise and mass accuracy are both very good.

De Novo - Issues

Less reliable when b and y ions don't overlap

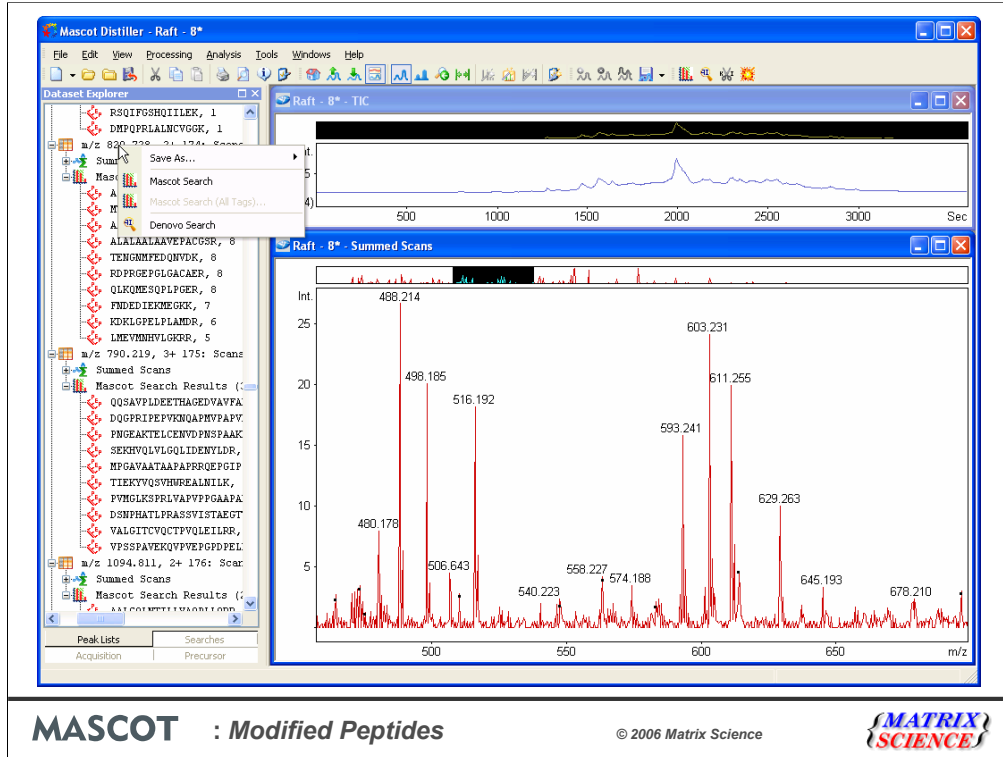


MASCOT : Modified Peptides

© 2006 Matrix Science

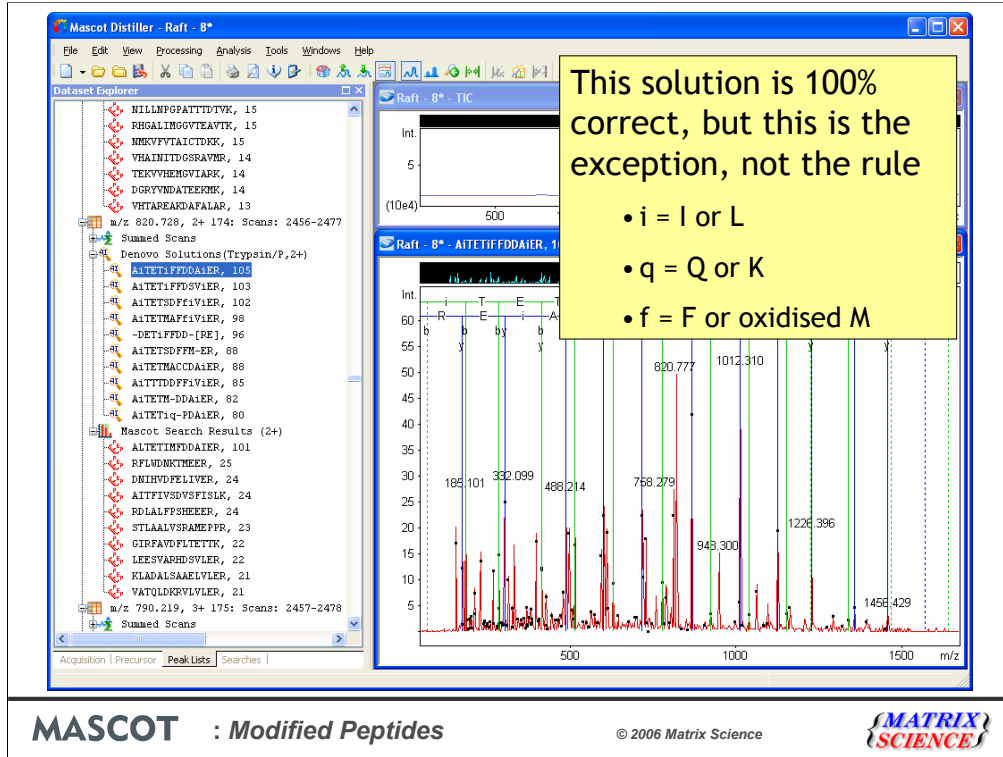
MATRIX
SCIENCE

Another important factor is coverage. It is hard to over emphasise the importance of getting both N-term and C-term matches for a stretch of de novo sequence to be reliable. This is a particular problem with large peptides, where the spectrum is often only good at (say) the low mass end. If the C-term ladder and the N-term ladder do not overlap, this is a much, much less constrained situation.



MASCOT : Modified Peptides © 2006 Matrix Science **MATRIX SCIENCE**

De novo is implemented in Mascot Distiller, because it requires very reliable peak picking. The starting point can be any MS/MS scan that has been processed to create a peak list. Right click the peak list node in Dataset Explorer and choose 'de novo Search', or choose the de novo button from the toolbar when a Summed Scans node is selected.

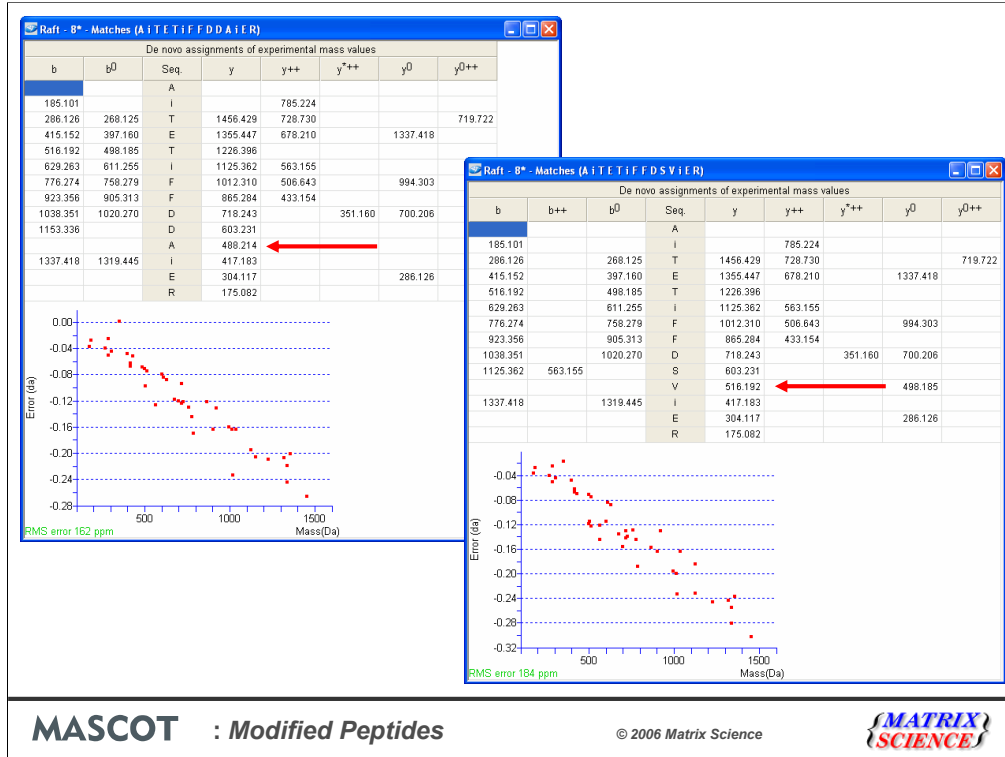


Good signal to noise and good mass accuracy are critical for successful de novo sequencing; much more so than in database searching. GIGO (garbage in - garbage out) is guaranteed.

In a de novo solution, i always represents I or L. q represent Q or K, when the mass tolerance does not allow these residues to be distinguished, although K is assumed at the C terminus of a peptide when tryptic specificity applies. f represents F or Met-Ox.

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 the pairs and triplets is undefined, so that SP could also be PS.

Although 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. For example, the score hardly changes when DA is replaced by SV

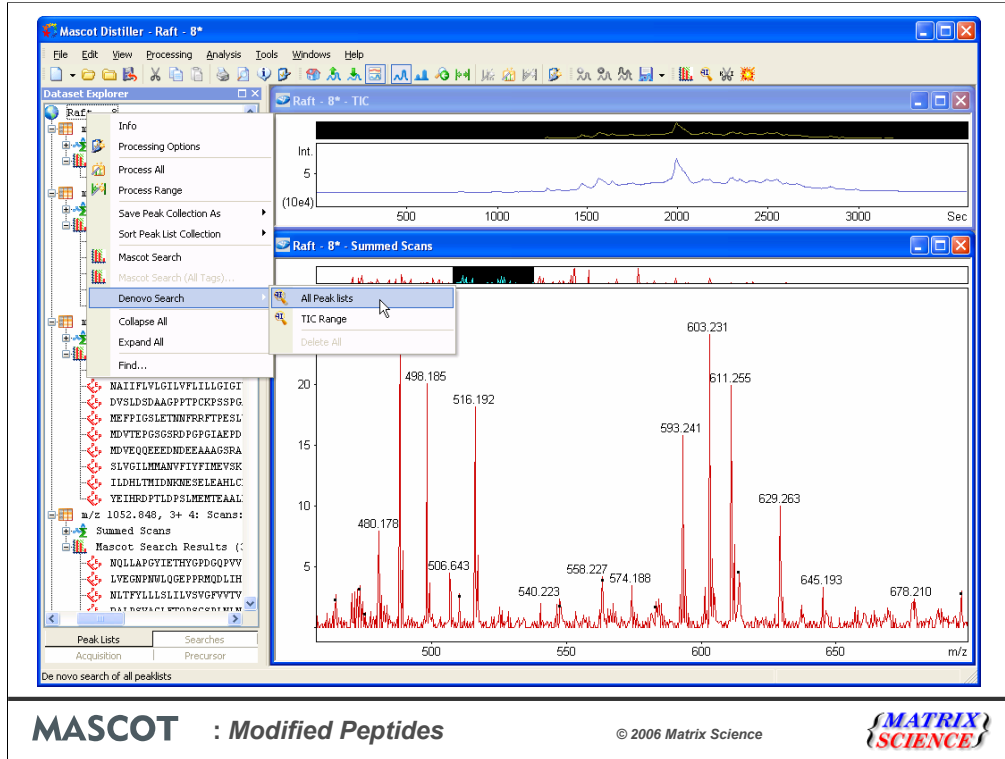


MASCOT : Modified Peptides

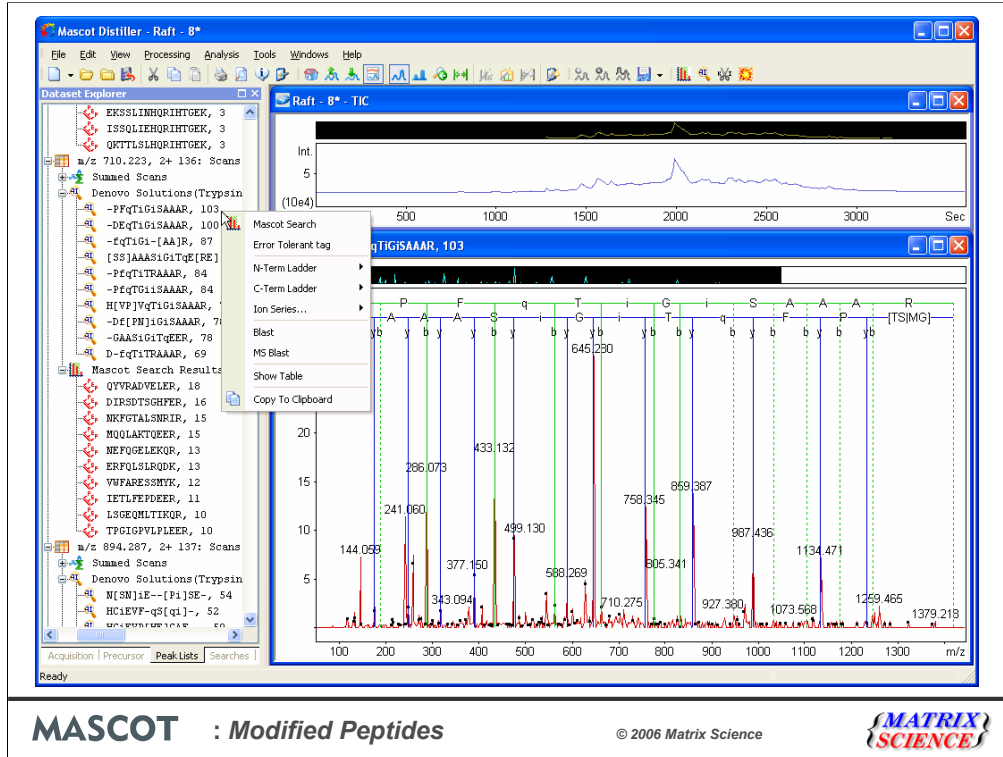
© 2006 Matrix Science



If we look at details of these two matches, we can see why. There is a y ion peak at 488.214 to support the sequence being DA, but there is also a y ion peak at 516.192 to support the alternative. This is just the nature of de novo on non-ideal data



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



You can then browse down the tree, looking for cases where the database search failed and de novo has a high score.

This looks like a promising case. The Mascot search didn't get a significant match, but de novo has a very high score.

But, is it right? And, how do we resolve the ambiguity at the N-terminus?

Search strategy

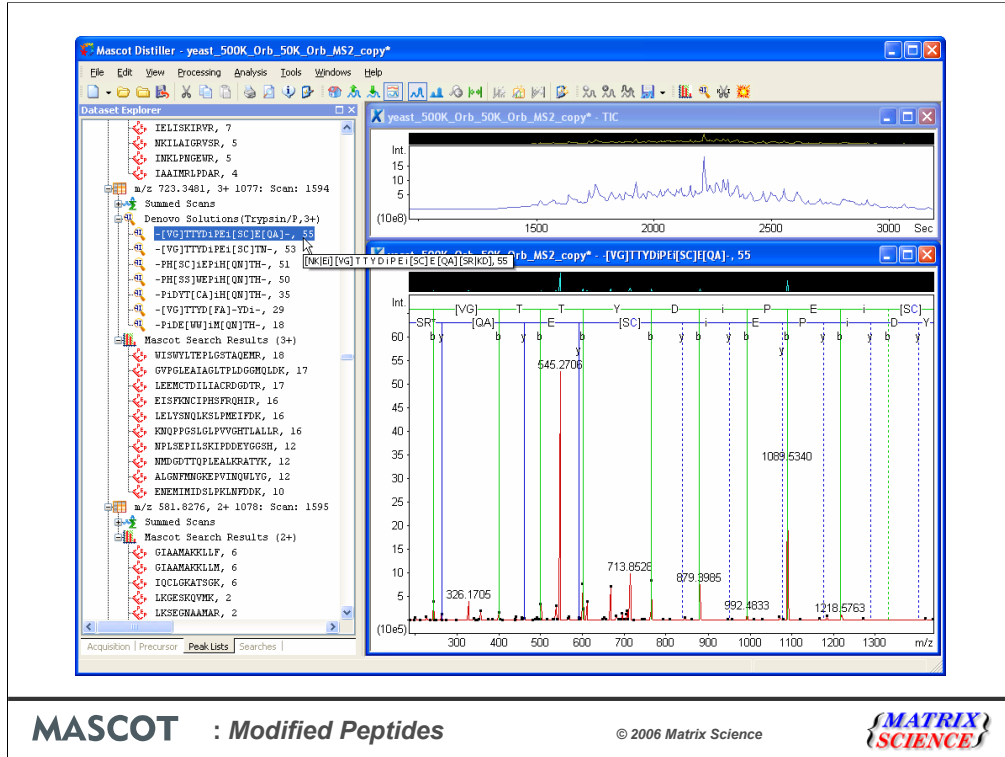
1. **Standard Mascot search**
Returns the easy matches
2. **Error tolerant search**
Returns additional matches, but only for proteins where we have at least one good peptide match already
Limited to a single additional SNP or modification per peptide
3. **De novo**
If data very high quality, can return novel full-length peptide sequences
Use Blast to find likely parent proteins
More often, returns partial / ambiguous peptide sequences
4. **Error tolerant tag search**
To find matches to
 1. Isolated peptides that have a SNP or unsuspected modification
 2. Peptides with multiple SNPs or unsuspected modifications(No reason to expect additional matches from a **standard tag search**)

MASCOT : *Modified Peptides*

© 2006 Matrix Science

MATRIX
SCIENCE

This brings us to step 4 of our strategy. As long as we are not dealing with an un-sequenced genome, the best way to test a de novo solution is an error tolerant tag search. This can often get a match even when there are multiple differences between the analyte peptide and the database sequence



Here's another example, from the Orbitrap data, where the Mascot database search has failed to find a match

The de novo solution is not a great score, and there's ambiguity at each terminus

MASCOT : Modified Peptides

© 2006 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 Sequence Query

Your name: _____ Email: _____

Search title: yeast_500K_Orb_50K_Orb_MS2_copy_RAW

Database: NCBIInr

Taxonomy: All entries

Enzyme: Trypsin/P Allow up to: 2 missed cleavages

Fixed modifications: Acetyl (N-term), Amide (C-term), Biotin (K), Biotin (N-term), Carbamidomethyl (C)

Variable modifications: N-Acetyl (Protein), N-Formyl (Protein), NIPCAM (C), O18 (C-term), Oxidation (M)

Protein mass: _____ kDa ICAT:

Peptide tol. ±: 0.020 Da MS/MS tol. ±: 0.020 Da

Peptide charge: 1+ Monoisotopic: Average

Query: ETAG=399:21020,11Y,764:37078
 ETAG=764:37078,D[L]P,1089:53398
 ETAG=500:25857,YD,879:39847
 ETAG=879:39847,[L]I]PE,1218:57634
 ETAG=601:33028,YD[L]I,992:48333
 ETAG=992:48333,PE[L]I,1331:66054

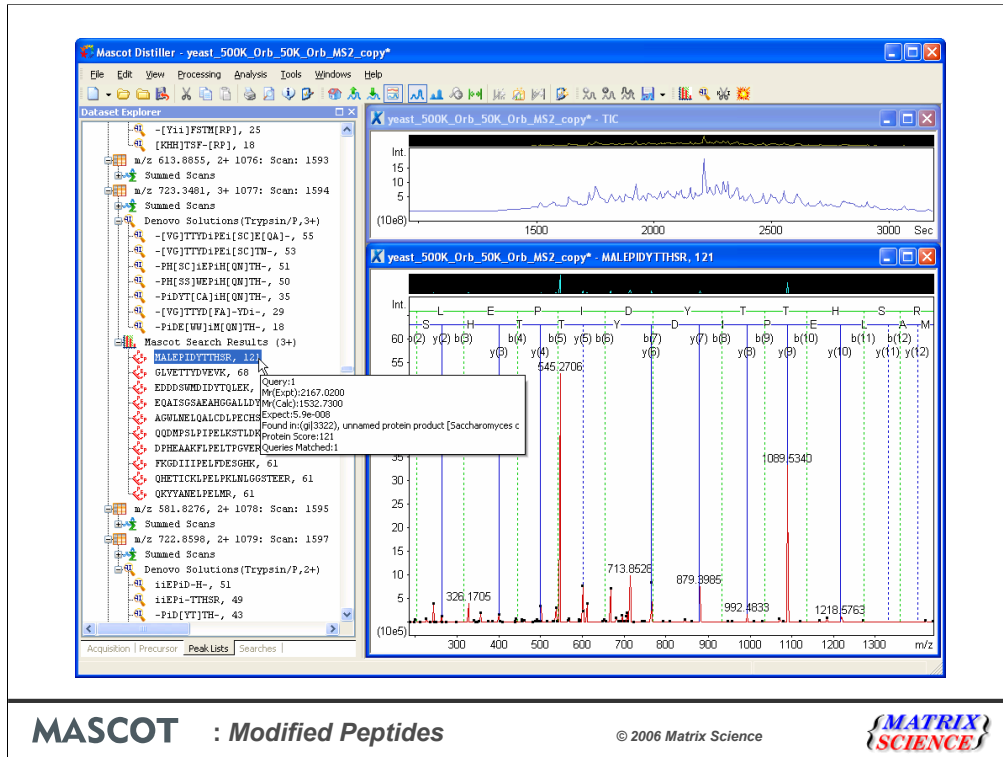
Instrument: ESI-TRAP

Overview: Report top: AUTO hits

Start Search... Reset Form

Copyright © 2005 Matrix Science Ltd. All Rights Reserved.

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



MASCOT : Modified Peptides

© 2006 Matrix Science



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 very promising. It's a high score, and it's a protein to which we already have other good matches. Notice that the de novo solution wasn't bad, but it was reversed. TTYDiPEi should have been iEPiDYTT. Unless the de novo manages to reach a terminus, there's a 50:50 chance that it will be the wrong way round

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

Peptide Summary Report (1077:ET) - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/master_results.pl?file=.../data/20060525/Faomneo0.dat&sessionID=quest_100052565474484

Select All Select None Search Selected Error tolerant

1. [g113322](#) Mass: 65552 Score: 121 Queries matched: 1
 unnamed protein product [Saccharomyces cerevisiae]
 Check to include this hit in error tolerant search

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/>	723.3481	2167.0225	1532.7293	634.2932	0	121	5.9e-08	1	-.MALEPIDYTTTHER.E

Top scoring peptide matches to query 1

Prot: 1077: Scan 1594 (rt=35.0144)

[g112](#) etag(399.21020, TTY, 764, 37078)

[g116](#) etag(500.25857, TYD, 879, 39847)

[g111](#) etag(992.48333, PE[L]I, 1331.66054)

Chai: Score greater than 61 indicates identity
 Status bar shows all hits for this peptide

2. [g118](#)
 hypo: 121.3 634.29 1 [g113322](#) -.MALEPIDYTTTHER.E
 Check: 67.6 615.34 2 [g1184994](#) K.GLVETTYDVEVK.L
 65.6 165.21 3 [g127803027](#) R.EDDSWMDIDYTOLEK.E
 Query: 65.6 -221.12 4 [g185093159](#) K.EQAISGSAEAAHGALLDYTEELK.I
 63.4 -474.22 5 [g168012661](#) R.AGVNLNELQALCDLPECHSGSKTR.A
 61.0 127.96 K.QQDMPSELPIPELSTLTK.Y
 61.0 162.00 R.DPHEAAKFLPELTPGVER.I
 61.0 223.02 K.FKQDIIPELPEESGSK.I
 3. [g112](#)
 unna: 61.0 -481.34 R.QHETICKLPELPLNLGGSTEER.G
 Check: 61.0 513.20 K.QRYIANLPELMR.T

Query Observed Mr(expt) Mr(calc) Delta Miss Score Expect Rank Peptide

1:g113322

MASCOT : Modified Peptides © 2006 Matrix Science MATRIX SCIENCE

The reason we didn't get a match from Mascot is that there is a modification, giving a delta of 634 Da. The peptide forms the protein N-terminus
 If we click on the hyperlink to see the peptide view ...

Mascot Search Results: Peptide View - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/peptide_view.pl?file=../data/20060525/FaonmeaO.dat&query=1&hit=1&index=g%7c33228px=1

MS/MS Fragmentation of **MALEPIDYTHSR**
 Found in **gi|3322**, unnamed protein product [Saccharomyces cerevisiae]

Match to Query 1: 2167.022472 from(723.348100,3+) etag(399.21020,TTY,764.37078) etag(764.37078,D[L]I[F,1089.53398) etag(500.25857,TYD,879.39847) etag(879.39847,[L]I[P,E,1218.57634) etag(601.33028,YD[L]I,992.48333) etag(992.48333,PE[L]I,1331.66054) 1077: Scan 1594 (n=35.0144)

Monoisotopic mass of neutral peptide Mr(calc): 1532.7293
Fixed modifications: Carbamidomethyl (C)
Unsuspected modification: 634.2932 Da, located in the region N-term to A2
Ions Score: 121 **Expect:** 5.9e-08

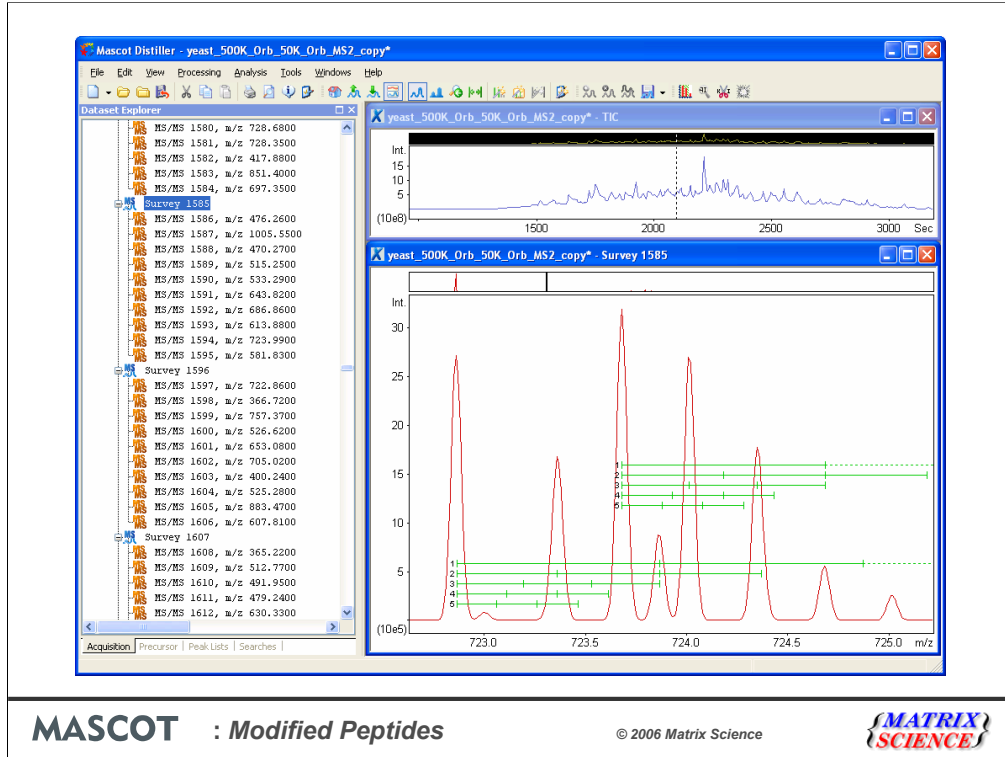
#	b	b ⁺⁺	b ⁰	b ⁰⁺⁺	Seq.	y	y ⁺⁺	y [*]	y ⁺⁺⁺	y ⁰	y ⁰⁺⁺	#
1	132.0478	66.5275			M							13
2	203.0849	102.0461			A	1402.6961	701.8517	1385.6696	693.3384	1384.6855	692.8464	12
3	316.1689	158.5881			L	1331.6590	666.3331	1314.6324	657.8199	1313.6484	657.3279	11
4	445.2115	223.1094	427.2010	214.1041	E	1218.5749	609.7911	1201.5484	601.2778	1200.5644	600.7858	10
5	542.2643	271.6358	524.2537	262.6305	P	1089.5323	545.2698	1072.5058	536.7565	1071.5218	536.2645	9
6	655.3483	328.1778	637.3378	319.1725	I	992.4796	496.7434	975.4530	488.2302	974.4690	487.7381	8
7	770.3753	385.6913	752.3647	376.6860	D	879.3955	440.2014	862.3690	431.6881	861.3850	431.1961	7
8	933.4386	467.2229	915.4280	458.2177	Y	764.3686	382.6879	747.3420	374.1747	746.3580	373.6826	6
9	1034.4863	517.7468	1016.4757	508.7415	T	601.3053	301.1563	584.2787	292.6430	583.2947	292.1510	5
10	1135.5340	568.2706	1117.5234	559.2653	T	500.2576	250.6324	483.2310	242.1191	482.2470	241.6271	4
11	1272.5929	636.8001	1254.5823	627.7948	H	399.2099	200.1086	382.1833	191.5953	381.1993	191.1033	3
12	1359.6249	680.3161	1341.6143	671.3108	S	262.1510	131.5791	245.1244	123.0659	244.1404	122.5738	2
13					R	175.1190	88.0631	158.0924	79.5498			1

MASCOT search of M A L E P I D Y T H S R

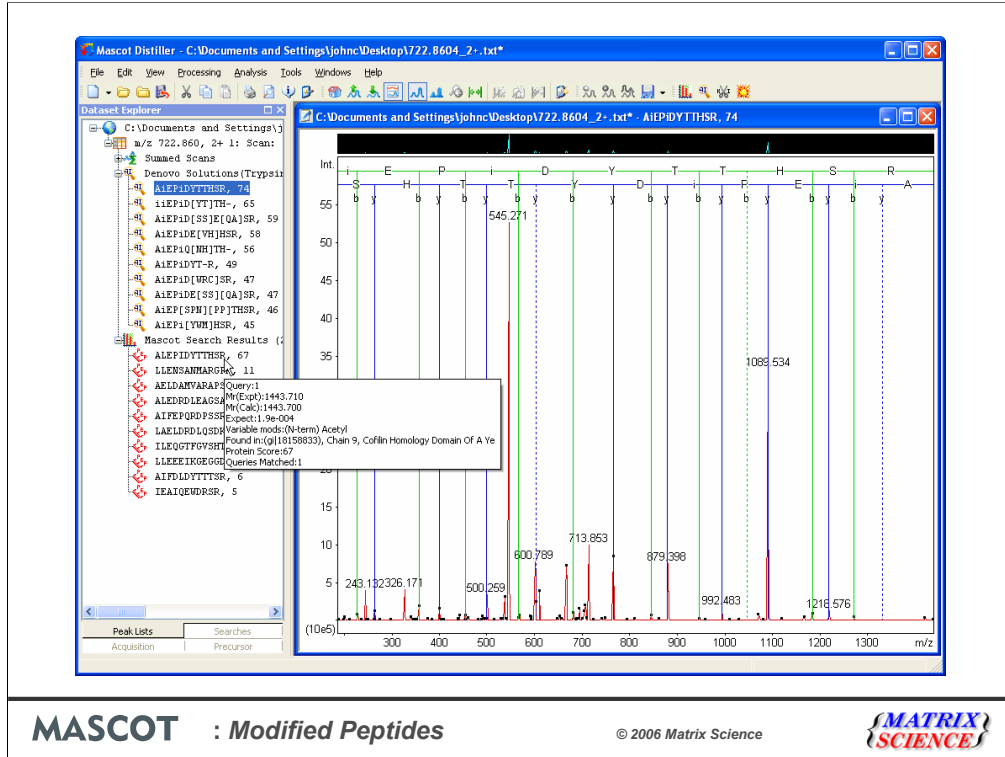
MASCOT : Modified Peptides © 2006 Matrix Science

The match was obtained by placing this modification delta on or close to the N-terminus. Remember that this peptide forms the N-terminus of the protein, so in all probability, the initiator Met is lost in the mature protein, making our actual mass delta 765.33. To further complicate the picture, the annotations for this protein report that the N-term Alanine is normally acetylated. If so, our unknown mod is actually 723.32

None of these deltas correspond to anything in Unimod, which is why this match wasn't picked up by the error tolerant search. This looked like a very solid match to me, so I started trying to figure out what the mod might be. Then the penny dropped. The delta is awfully close to the m/z value, which immediately suggests a precursor charge error.



If we go back to the survey scan, this is what we find. A 3+ peptide at 723.6803 and a 2+ peptide at 722.8604. The instrument thought it was going for 723.99, so Distiller used the 3+ peptide, which is both the closest and the more intense. If we take the correct mass and charge ...



We get the correct match from Mascot and even the de novo falls straight out. The initiator Met has indeed been removed and the Alanine acetylated. Not an unknown modification after all, but nice to get to the bottom of a small mystery. The take home message is that de novo plus sequence tag can often take you further than an error tolerant search of the uninterpreted data.

Search strategy

- 1. Standard Mascot search**
Returns the easy matches
- 2. Error tolerant search**
Returns additional matches, but only for proteins where we have at least one good peptide match already
Limited to a single additional SNP or modification per peptide
- 3. De novo**
If data very high quality, can return novel full-length peptide sequences
Use **Blast** to find likely parent proteins
More often, returns partial / ambiguous peptide sequences
- 4. Error tolerant tag search**
To find matches to
 1. Isolated peptides that have a SNP or unsuspected modification
 2. Peptides with multiple SNPs or unsuspected modifications(No reason to expect additional matches from a **standard tag search**)

MASCOT : *Modified Peptides*

© 2006 Matrix Science

MATRIX
SCIENCE

So, there we have it. Four powerful tools to help us find modified peptides. The challenge going forward is to make the workflow more integrated. It is still a bit too manual for large data sets. The other thing we need to address is the speed of an error tolerant tag. It would be great if we could find a way to speed this up so as to allow them to be fired off automatically for every de novo solution