Top 10 Tips for Successful Searching

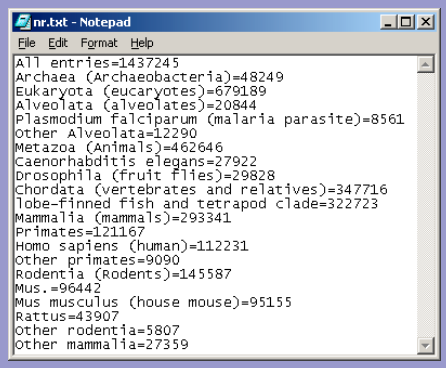
ASMS 2003

MATRIX SCIENCE

I'd like to present our top 10 tips for successful searching with Mascot.
Like any hit parade, we will, of course, count them off in reverse order

10. Don't specify a poorly represented taxonomy

- In most cases, if the correct protein is not in the database, you'd like to see the closest match … whatever the species



ASMS 2003

MATRIX SCIENCE

So, at number 10, Don't specify a poorly represented taxonomy.

Think carefully about what you are trying to achieve when specifying a taxonomy filter.

If the correct protein from the correct species is not in the database, wouldn't you want to see a good match to a protein from a different species?

This is especially important for poorly represented species. For example, look at these numbers for the NCBI nr database in June 2003: 1.4 million entries; 120,000 entries for primates, of which all but 9,000 are for human. So, even if you are studying chimps or orang-utans or yeti, you probably don't want to choose 'Other primates'.

9. Use the Peptide Summary Report for MS/MS results

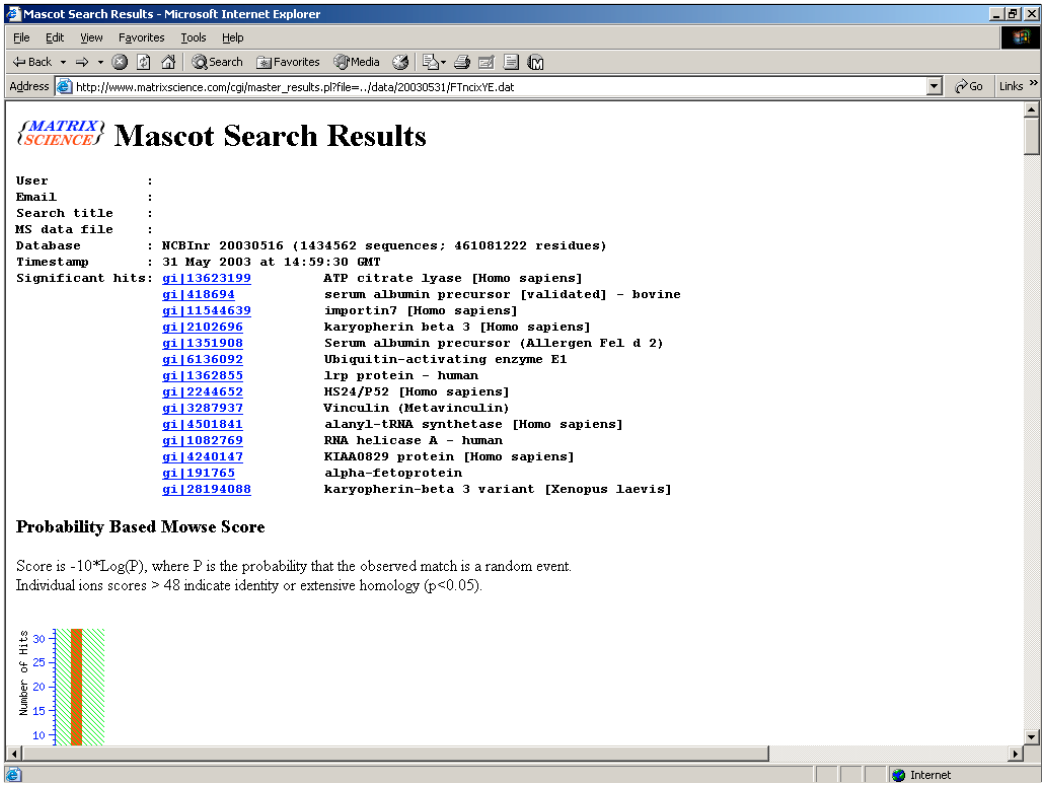
- The Protein Summary Report is intended for Peptide Mass Fingerprint results
- Worst case is a complex mixture with lots of queries
  - Protein Summary is 50 proteins max
  - Matches to sets of identical peptides are not collapsed into single protein hits, so a match may disappear off the end of the top 50
  - Weak matches may disappear into the distribution of random PMF matches

ASMS 2003

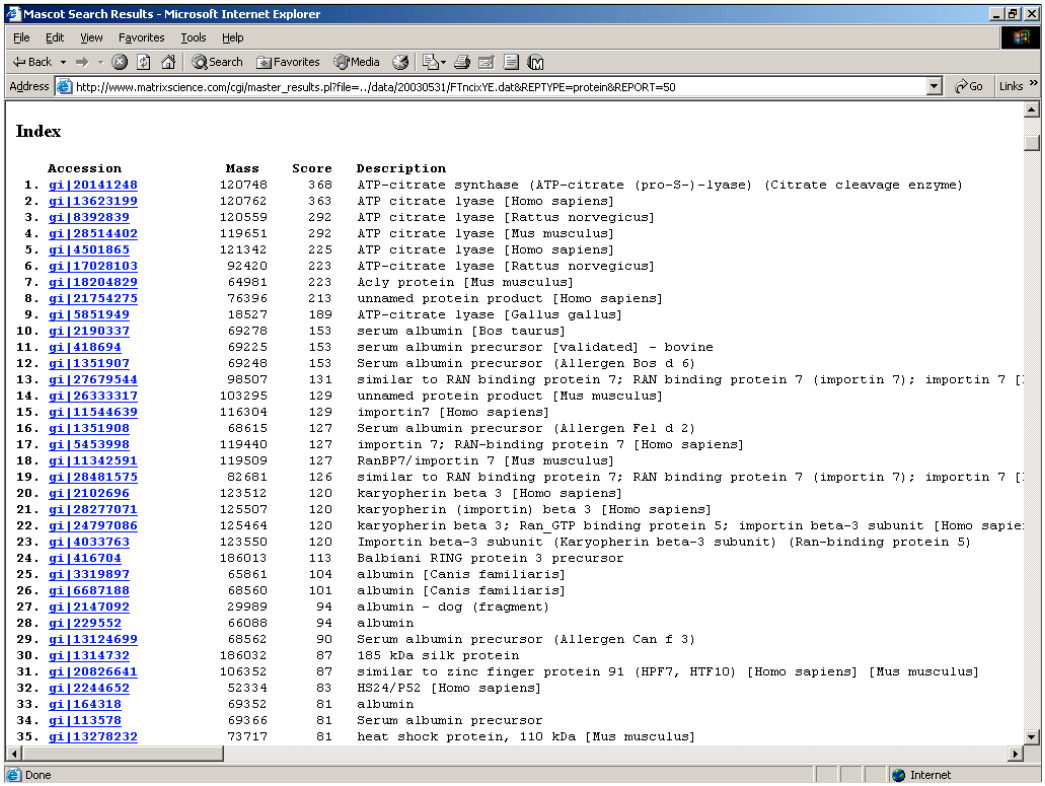
MATRIX SCIENCE

At number 9, We encourage you to Use the Peptide Summary Report for MS/MS results

There are several flavours of reports for Mascot search results. Historically, the first report was the Protein Summary, used for peptide mass fingerprint results. Because this was the first report, there are still some old clients out there that specify this report for all searches. Unfortunately, in most cases, the Protein Summary is not a good way to view MS/MS results. For example:



This result from an MS/MS search has 12 significant matches. There is a little bit of duplication, e.g. 2 serum albumins, but not much



Index

	Accession	Mass	Score	Description
1.	gi|20141248	120748	368	ATP-citrate synthase (ATP-citrate (pro-S-)-lyase) (Citrate cleavage enzyme)
2.	gi|13623199	120762	363	ATP citrate lyase [Homo sapiens]
3.	gi|8392839	120559	292	ATP citrate lyase [Rattus norvegicus]
4.	gi|28514402	119651	292	ATP citrate lyase [Mus musculus]
5.	gi|4501865	121342	225	ATP citrate lyase [Homo sapiens]
6.	gi|17028103	92420	223	ATP-citrate lyase [Rattus norvegicus]
7.	gi|18204829	64981	223	Acly protein [Mus musculus]
8.	gi|21754275	76396	213	unnamed protein product [Homo sapiens]
9.	gi|5851949	18527	189	ATP-citrate lyase [Gallus gallus]
10.	gi|2190337	69278	153	serum albumin [Bos taurus]
11.	gi|418694	69225	153	serum albumin precursor [validated] - bovine
12.	gi|1351907	69248	153	Serum albumin precursor (Allergen Bos d 6)
13.	gi|27679544	98507	131	similar to RAN binding protein 7; RAN binding protein 7 (importin 7); importin 7 [
14.	gi|26333317	103295	129	unnamed protein product [Mus musculus]
15.	gi|11544639	116304	129	importin7 [Homo sapiens]
16.	gi|1351908	68615	127	Serum albumin precursor (Allergen Fel d 2)
17.	gi|5453998	119440	127	importin 7; RAN-binding protein 7 [Homo sapiens]
18.	gi|11342591	119509	127	RanBP7/importin 7 [Mus musculus]
19.	gi|28481575	82681	126	similar to RAN binding protein 7; RAN binding protein 7 (importin 7); importin 7 [
20.	gi|2102696	123512	120	karyopherin beta 3 [Homo sapiens]
21.	gi|28277071	125507	120	karyopherin (importin) beta 3 [Homo sapiens]
22.	gi|24797086	125464	120	karyopherin beta 3; Ran_GTP binding protein 5; importin beta-3 subunit [Homo sapie
23.	gi|4033763	123550	120	Importin beta-3 subunit (Karyopherin beta-3 subunit) (Ran-binding protein 5)
24.	gi|416704	186013	113	Balbiani RING protein 3 precursor
25.	gi|3319897	65861	104	albumin [Canis familiaris]
26.	gi|6687188	68560	101	albumin [Canis familiaris]
27.	gi|2147092	29989	94	albumin - dog (fragment)
28.	gi|229552	66088	94	albumin
29.	gi|13124699	68562	90	Serum albumin precursor (Allergen Can f 3)
30.	gi|1314732	186032	87	185 kDa silk protein
31.	gi|20826641	106352	87	similar to zinc finger protein 91 (HPF7, HTF10) [Homo sapiens] [Mus musculus]
32.	gi|2244652	52334	83	HS24/P52 [Homo sapiens]
33.	gi|164318	69352	81	albumin
34.	gi|113578	69366	81	Serum albumin precursor
35.	gi|13278232	73717	81	heat shock protein, 110 kDa [Mus musculus]

If we look at the same results in protein view, there is much greater duplication, because this type of report isn't trying to collapse hits that share a common set of MS/MS matches.

Now, we have 7 or 8 representatives for the more common protein hits, which means that the lower scoring hits are pushed off the bottom of the list.

Also, you can't see the wood for the trees.

So, if you have old client software that brings up a protein summary for an MS/MS search, the first thing to do is click on the link to switch to the peptide summary

8. Submit new modifications to Unimod

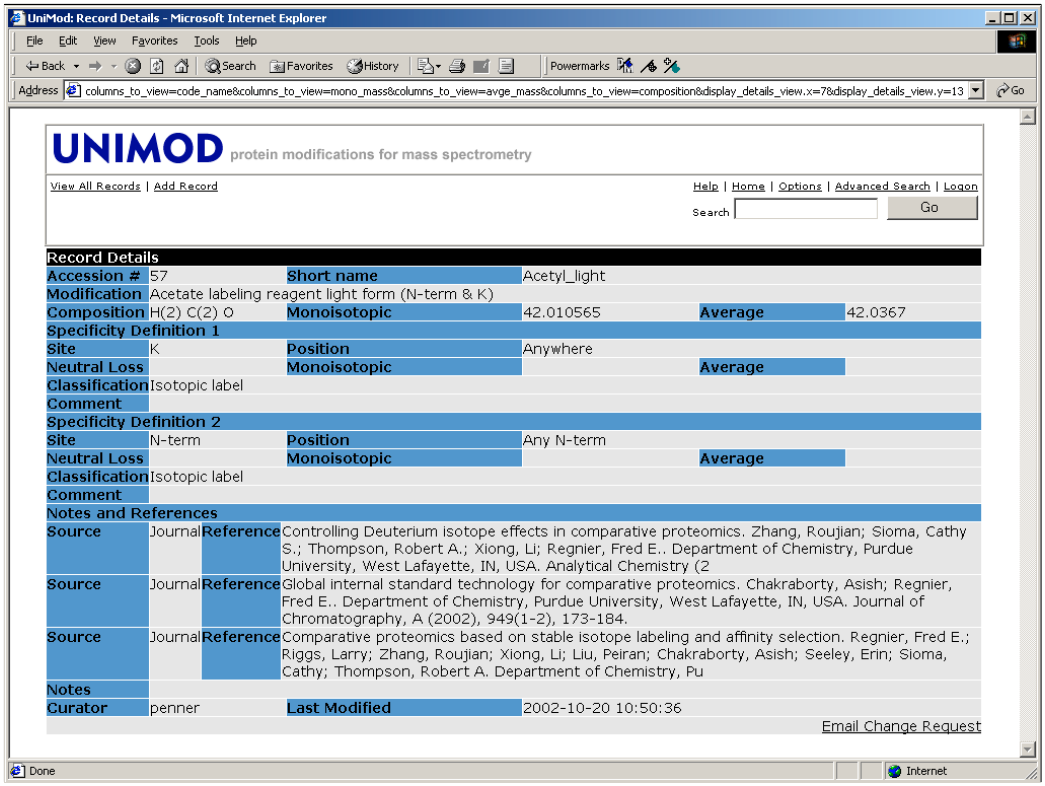
- On-line at www.unimod.org
  - Saves calculating mass values
  - Saves having to understand the syntax of the Mascot mod_file
  - Share your modification with other Mascot users
  - Provides a way to update the modifications list on the Mascot public web site

ASMS 2003

MATRIX SCIENCE

At number 8, if the modification you need isn't on the Mascot search form list, submit it to Unimod. The advantages of doing this are

<read from slide>



Unimod is a live, public domain database. If you add a modification, you become the curator of that modification. The database is used to update the Mascot mod_file every weekend. If you have an in-house Mascot server, you can download the same new mod_file

7. Be skeptical if the Mascot score is below threshold

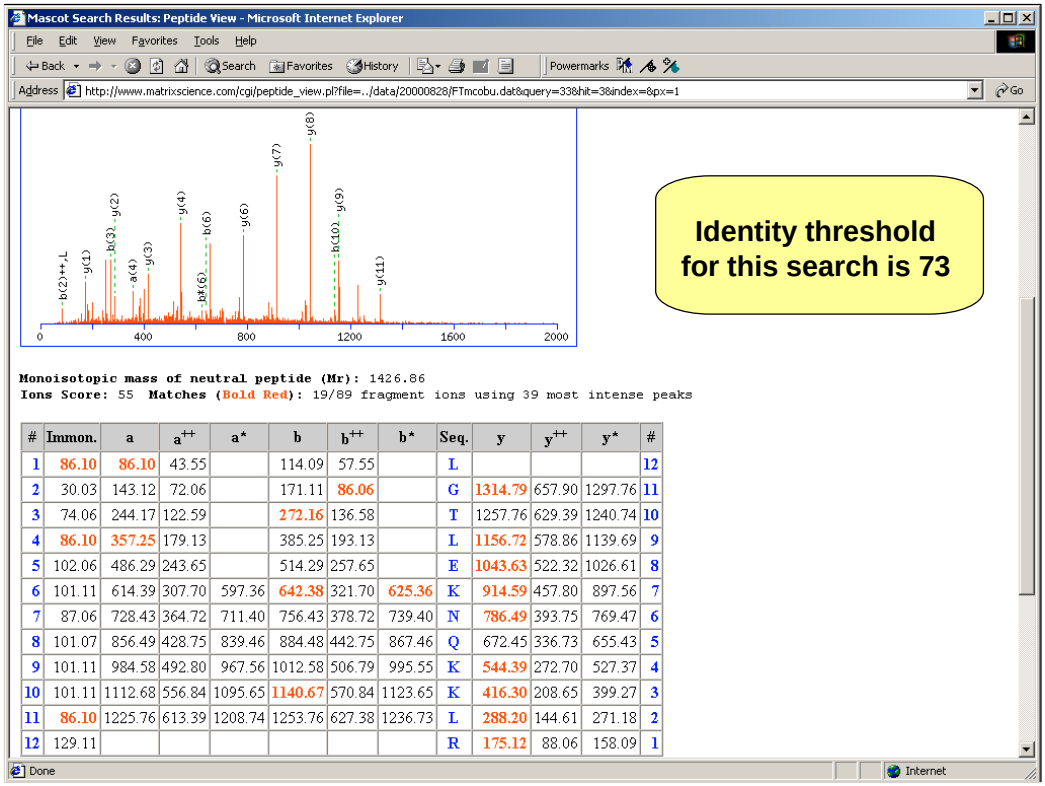
- It *may* be right …

ASMS 2003

MATRIX SCIENCE

At number 7, with a bullet, Be skeptical if you want to accept a match when the Mascot score is below threshold.

You may be right ...

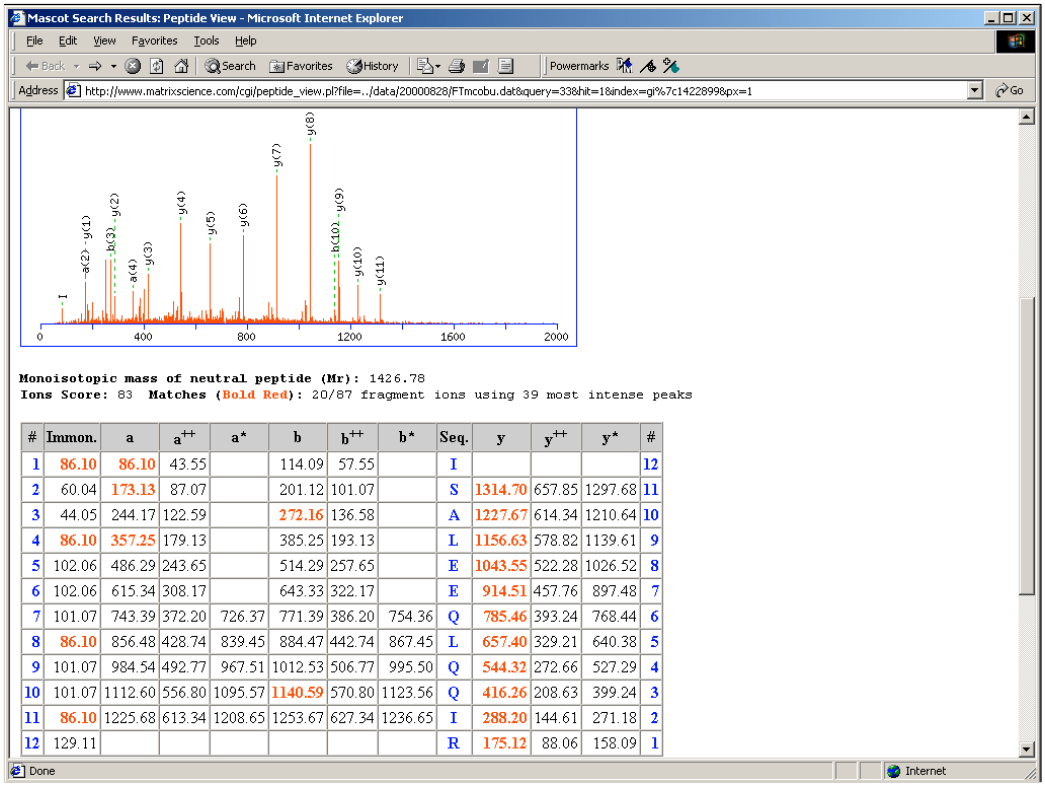


#	Immon.	a	a++	a*	b	b++	b*	Seq.	y	y++	y*	#
1	86.10	86.10	43.55		114.09	57.55		L				12
2	30.03	143.12	72.06		171.11	86.06		G	1314.79	657.90	1297.76	11
3	74.06	244.17	122.59		272.16	136.58		T	1257.76	629.39	1240.74	10
4	86.10	357.25	179.13		385.25	193.13		L	1156.72	578.86	1139.69	9
5	102.06	486.29	243.65		514.29	257.65		E	1043.63	522.32	1026.61	8
6	101.11	614.39	307.70	597.36	642.38	321.70	625.36	K	914.59	457.80	897.56	7
7	87.06	728.43	364.72	711.40	756.43	378.72	739.40	N	786.49	393.75	769.47	6
8	101.07	856.49	428.75	839.46	884.48	442.75	867.46	Q	672.45	336.73	655.43	5
9	101.11	984.58	492.80	967.56	1012.58	506.79	995.55	K	544.39	272.70	527.37	4
10	101.11	1112.68	556.84	1095.65	1140.67	570.84	1123.65	K	416.30	208.65	399.27	3
11	86.10	1225.76	613.39	1208.74	1253.76	627.38	1236.73	L	288.20	144.61	271.18	2
12	129.11							R	175.12	88.06	158.09	1

Here's a good example.

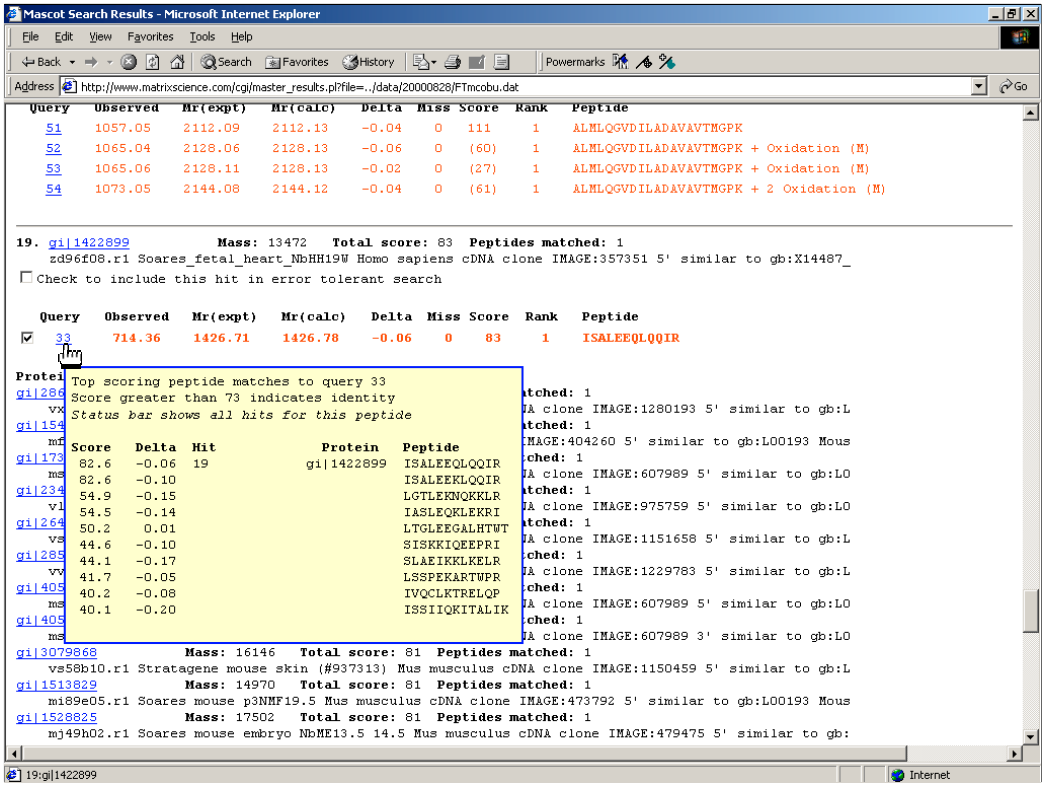
A run of 9 Y ions. Who wants to tell me that this could never happen by chance? And yet the score is below threshold!

OK, now lets take a look at a different match



#	Immon.	a	a^{++}	a*	b	b^{++}	b*	Seq.	y	y^{++}	y*	#
1	86.10	86.10	43.55		114.09	57.55		I				12
2	60.04	173.13	87.07		201.12	101.07		S	1314.70	657.85	1297.68	11
3	44.05	244.17	122.59		272.16	136.58		A	1227.67	614.34	1210.64	10
4	86.10	357.25	179.13		385.25	193.13		L	1156.63	578.82	1139.61	9
5	102.06	486.29	243.65		514.29	257.65		E	1043.55	522.28	1026.52	8
6	102.06	615.34	308.17		643.33	322.17		E	914.51	457.76	897.48	7
7	101.07	743.39	372.20	726.37	771.39	386.20	754.36	Q	785.46	393.24	768.44	6
8	86.10	856.48	428.74	839.45	884.47	442.74	867.45	L	657.40	329.21	640.38	5
9	101.07	984.54	492.77	967.51	1012.53	506.77	995.50	Q	544.32	272.66	527.29	4
10	101.07	1112.60	556.80	1095.57	1140.59	570.80	1123.56	Q	416.26	208.63	399.24	3
11	86.10	1225.68	613.34	1208.65	1253.67	627.34	1236.65	I	288.20	144.61	271.18	2
12	129.11							R	175.12	88.06	158.09	1

Now we have a run of 11 y ions and a higher score, above the significance threshold. These are not similar sequences with the same set of mass matches.



The threshold was high because this was an EST search. However, this doesn't change the fact that many people would accept the first match until shown the second.

Our subjective judgement can be misleading

I'm not suggesting that Mascot is infalible, far from it. However, if you choose to disregard the score, you should look very carefully at the match and, ideally, have some additional evidence for it being correct.

6. Peak detection, peak detection, peak detection

- Especially critical for Peptide Mass Fingerprints
- Time domain summing of LC-MS/MS data is very important
- Throw out low mass precursors in MS/MS

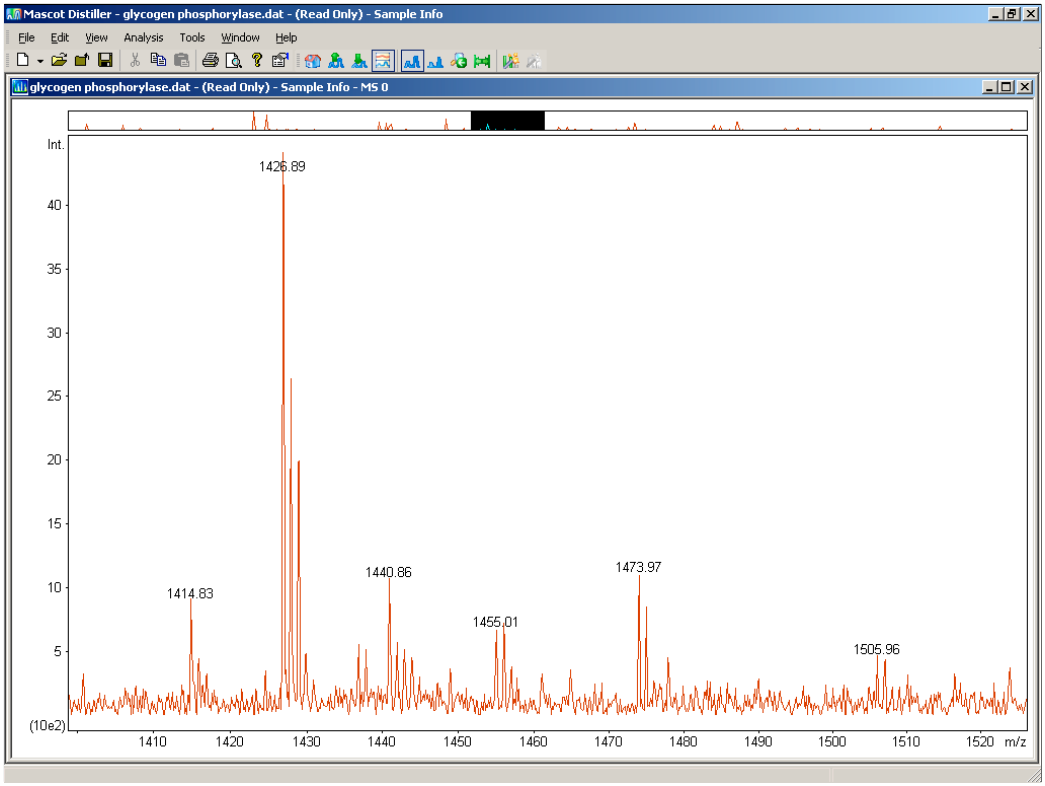
ASMS 2003

MATRIX SCIENCE

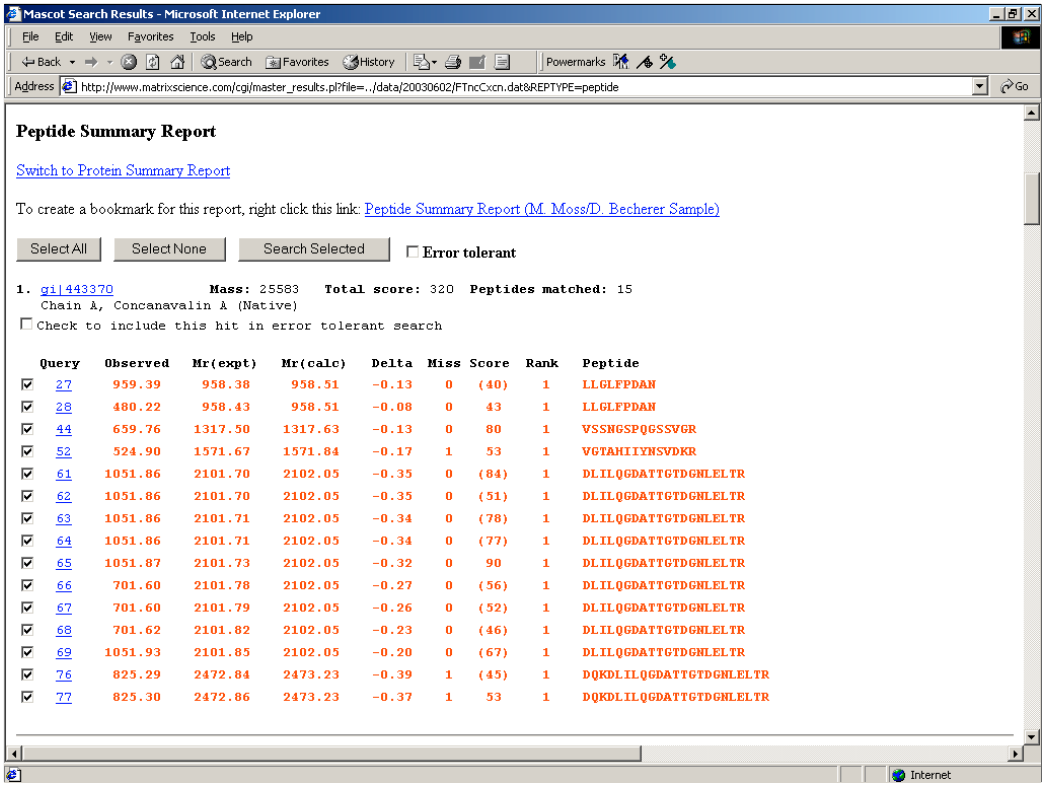
If you ask an estate agent (a realtor) in the UK what determines the price of property, they'll probably reply location, location, and location.

Well, in many ways, the quality of a Mascot result depends on peak detection, peak detection, peak detection.

<read from slide>



There is a world of difference between a good quality peak list, as you might expect from a piece of software like - random example - Mascot Distiller, and a poor quality peak list, where every spike and glitch on the baseline has been added to the peak list



In the case of MS/MS data, noise peaks aren't such a problem, because Mascot itertively determined which are signal and which are noise. However, time domain processing of LC-MS/MS data is very important.

This example shows what you don't want to see - the same peptide found over and over again. If all these spectra could be summed together, the signal to noise, and hence the Mascot score, would be greatly improved

5. Keep the taxonomy indexes up-to-date

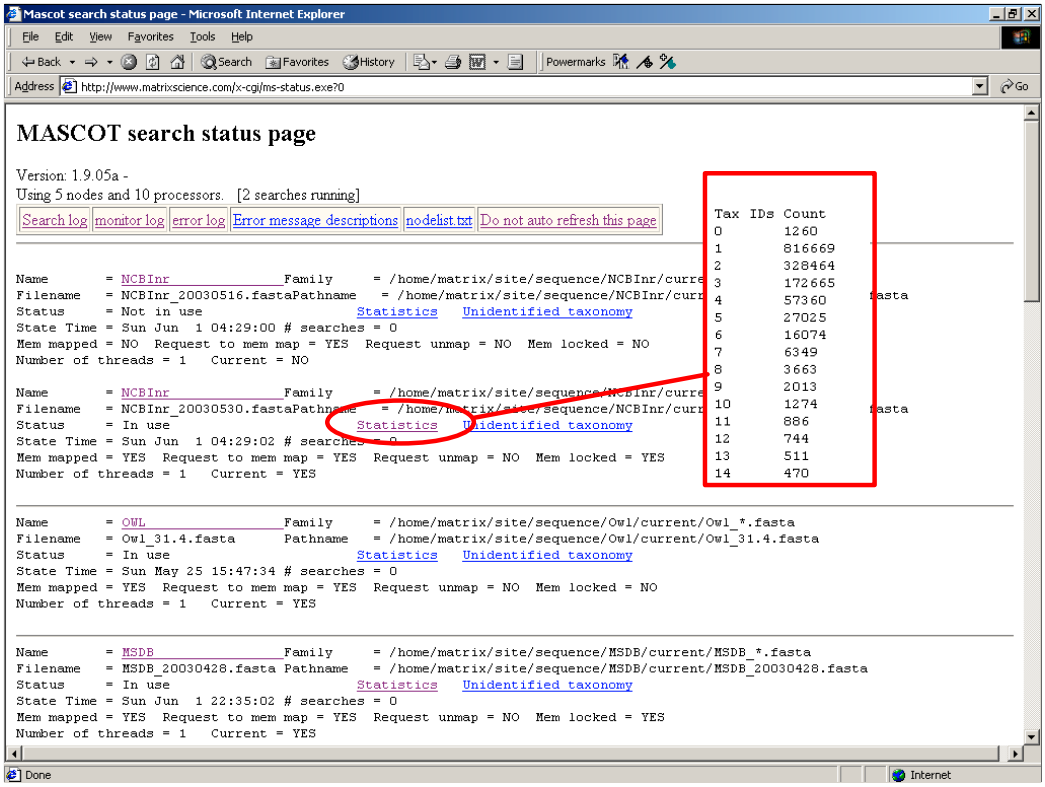
- Whenever you update a database, update the relevant taxonomy files
  - Database update script does this automatically

ASMS 2003

MATRIX SCIENCE

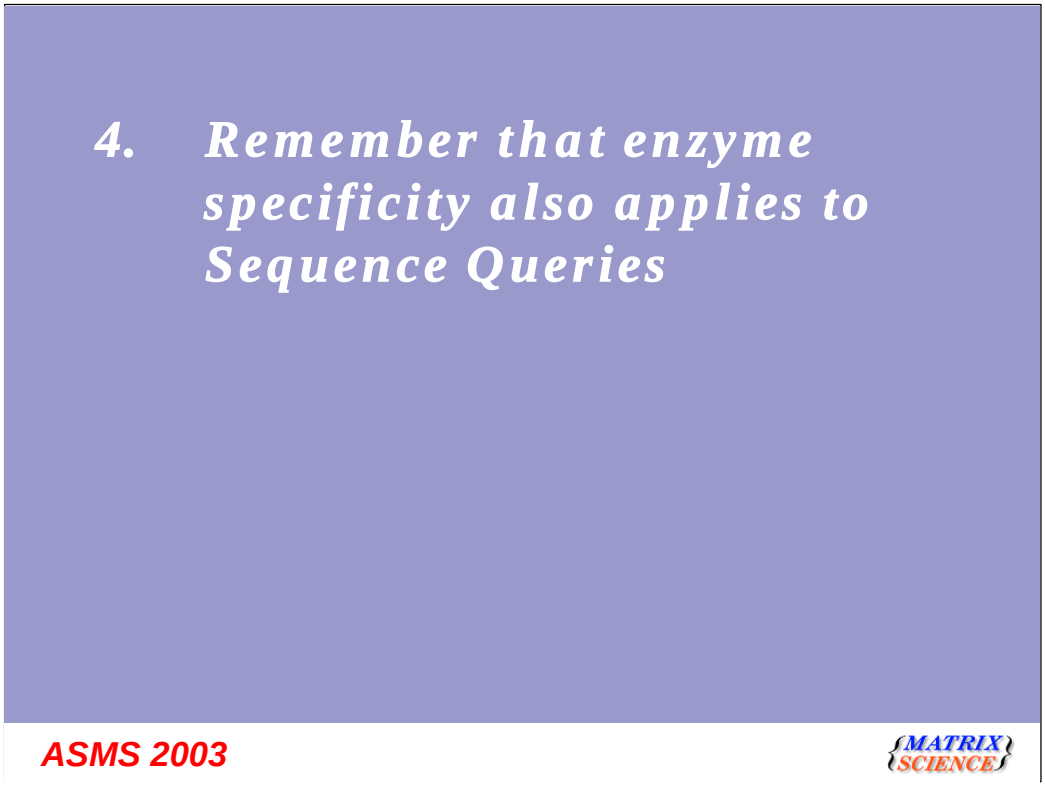
At number 5, one for the Administrators of in-house Mascot servers: Keep your taxonomy indexes up-to-date

<read from slide>



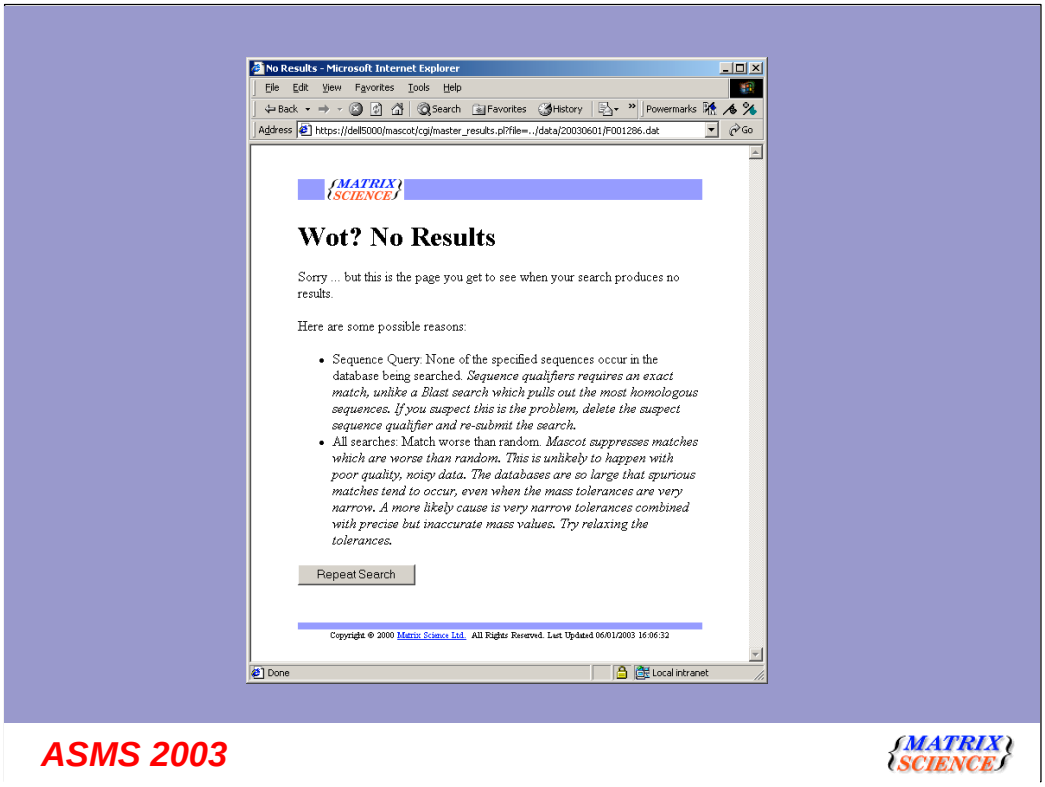
From time to time, its a good idea to check the stats file for each database. It contains lots of useful information, like whether entries contain illegal characters or whether an entry is too long.

It also tells you how good your taxonomy is. Here are the numbers for the nr database on our web site at the end of May. There are 1.4 million entries, but only 1200 have no taxonomy. In other words, better than 99.9% of the entries have a taxonomy assigned. If you look at your stats file and see that (say) 10% of the entries have no taxonomy, that's 10% of the entries that are going to be missed whenever you do a search with taxonomy specified.



Top tip number 4 is Remember that enzyme specificity also applies to Sequence Queries

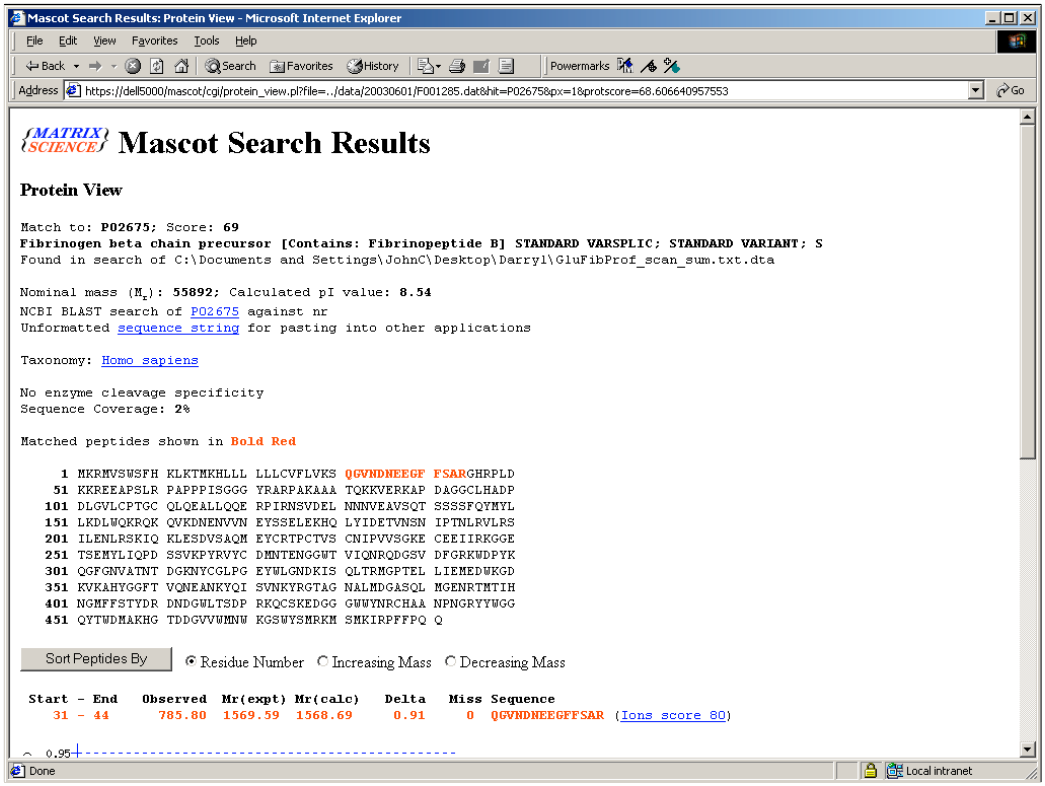
One of the most common emails we receive is "Mascot is broken. I did a search for this peptide and I know its in the database but Mascot failed to find it"



For example, here's a search for glu-fib, a very common sequencing standard. The mass is correct and the sequence is correct. But, when we do a search of Swiss-Prot -

No results!

Why



Because glu-fib in Swiss-Prot is not a tryptic peptide. The N-terminus is created by a post-translational cleavage after serine. If you now go back to the search form and select enzyme type none, bingo ... you'll get a match

3. *Don't specify a protein mass unless essential*

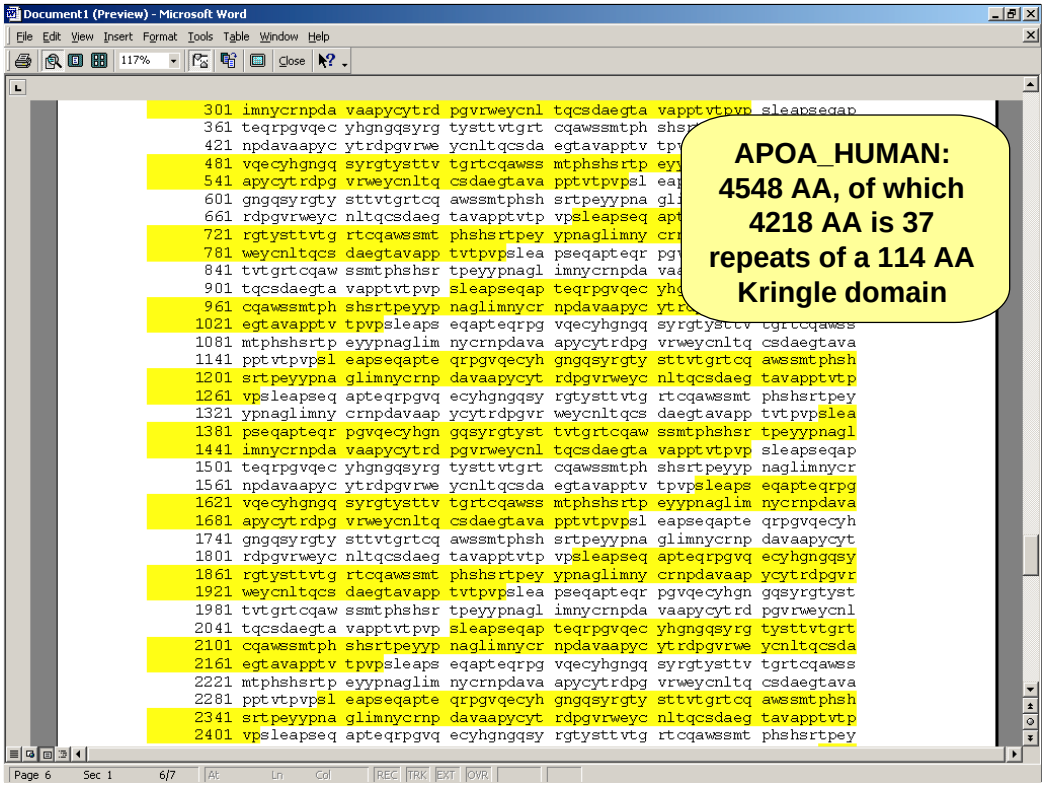
- Slows down the search
- Cannot guarantee that the mass of the database entry is close to that of the analyte
- Never useful for MS/MS search. Only useful for Peptide Mass Fingerprint when
  - Analyte is small fragment of very large entry
  - Low complexity entry

ASMS 2003

MATRIX SCIENCE

Number 3 is another very common technical support issue: Whether to specify a protein mass

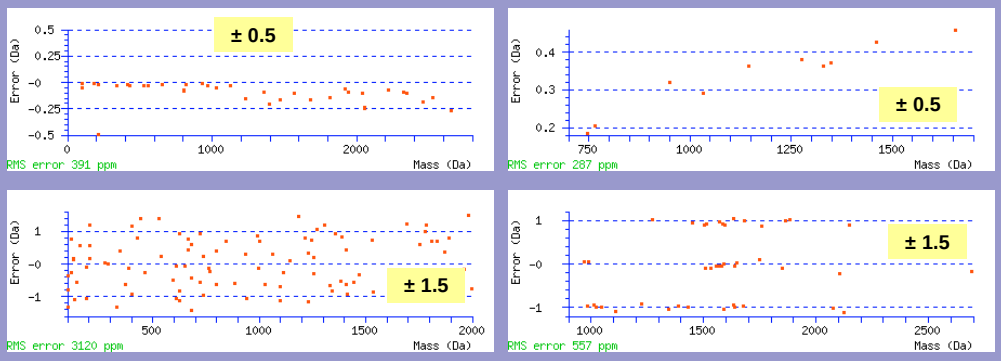
<read from slide>



Here, for example, is human Apolipoprotein a Almost all of this protein is a repeated kringle domain of just 114 residues. Statistically, this protein behaves like a much smaller protein ... for eample, it will produce many fewer unique tryptic peptides than you would expect from its size. If you had a peptide mass map of this protein, it would be very, very difficult to get a match without specifiying a small protein mass.

This, and the case where the experimental protein is a very small fragment of the database entry are the times you need to use SEG. Otherwise, much better to leave the protein mass open

2. Use the error graphs to estimate mass tolerances



ASMS 2003

MATRIX SCIENCE

Number 2 is a reminder to use the error graphs to estimate mass tolerances

1. This example is fine, the mass errors are well within the specified tolerance of +/- 0.5. You could probably increase the score slightly by going to +/- 0.3, but safer to leave it where it is

2. This is also fine! The mass values are mostly within the specified tolerance of +/- 1.5. In fact, this is the error distribution for a very good MS/MS match from an ion trap.

3. In contrast, this is not right. Although the accuracy is better than the last example, the mass scale should continue to 2500 Da. However, all the potentail matches above 1650 Da have been lost because the tolerance is too tight and is clipping the high masses. The precision suggests that some calibration is overdue

4. This is a worrying example. The accuracy is excellent, but a very wide tolerance has been specified. For a peptide mass fingerprint, this can easily create a false positive, because the distribution of mass values is is not uniform. This kind of data is playing with Mascot's mind. I don't have time to go into great detail. Suffice to say that if you see this, you should set a more appropriate tolerance, like +/- 0.5.

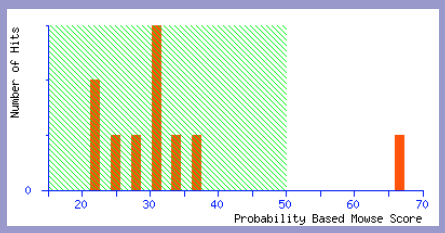
1. Be sparing with variable modifications

- Some modifications are worse than others
  - Mods that affect a terminus are less of a problem, e.g. Pyro-glu
  - Mods that apply to residue(s) with a high fractional abundance and at any position are BIG problem, e.g. Phospho (ST) = 13%
- Use an error tolerant search to pick up uncommon modifications
  - Efficient
  - Also catch non-specific peptides

ASMS 2003

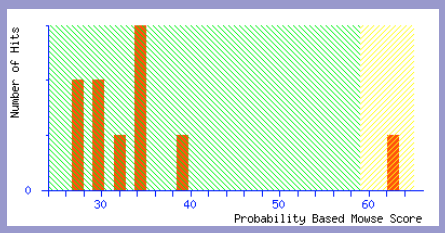
MATRIX SCIENCE

And finally, number 1, our top tip! Be sparing with variable modifications
<read from slide>



Oxidation (M)

8 sec



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

92 sec

ASMS 2003

MATRIX SCIENCE

This search of a single MS/MS spectrum, using one variable mod, gives a nice, statistically significant match.

If the search is repeated with 8 mods, the match is the same, but it is no longer so clear cut.

All of these mods have effectively increased the size of the database by a factor of 30!

Whats worse, the search takes over 10 times as long!

So, our top tip is to use variable mods sparingly. You'll get better results faster.

1. Be sparing with variable modifications

2. Use the error graphs to estimate mass tolerances

3. Don't specify a protein mass unless essential

4. Remember that enzyme specificity also applies to Sequence Queries

5. Keep the taxonomy indexes up-to-date

6. Peak detection, peak detection, peak detection

7. Be skeptical if Mascot score is below threshold

8. Submit new modifications to Unimod

9. Use the Peptide Summary Report for MS/MS results

10. Don't specify a poorly represented taxonomy

ASMS 2003

MATRIX SCIENCE

So, there we are, our top 10 tips for 2003. I hope you'll find one or two of them useful