New Features in Mascot 2.0

ASMS 2004

MATRIX SCIENCE

In this session, I'll be describing the some of the enhancements that we introduced with Mascot 2.0.

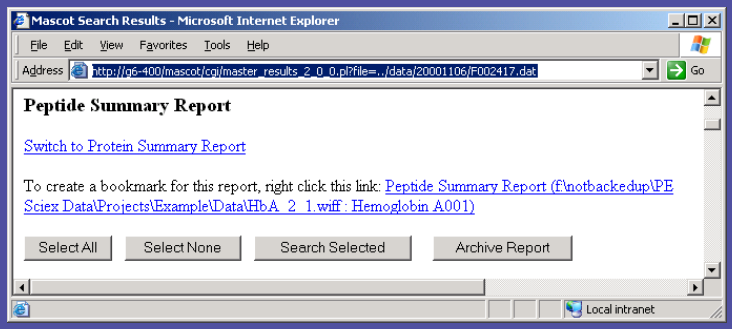
There is a 'new look' for all the web pages, support for very large MudPIT searches, sequence tag and error tolerant tag searches. There were also a number of enhancements to Mascot Daemon. Finally I'll describe a number of minor changes including iteration of B, Z and X residues.

The first thing that that you will notice about Mascot 2.0 is that all the HTML pages have a new look.

There are some practical benefits to this. For example, the context sensitive menu makes navigation around the help pages and online resources much easier

New Web Page Design

- No frames



- Removed link for bookmarking report

ASMS 2004

MATRIX SCIENCE

Another benefit is that we no longer use 'frames'. The use of separate frames for the menu and the title was quite annoying with some browsers. It meant for example that if you just chose "Add to Favourites" in your browser, you could find you'd bookmarked the home page or the menu frame.

Because of this, we added the "right click this link" in earlier versions of Mascot. Now, because we don't have frames, you will find that this link has gone.

Large searches - memory issues

- Mascot 1.01 had a limit of 300 spectra
- Limit of approx. 40,000 spectra in Mascot 1.9 on 32 bit platforms
- Mascot 2.0 - no limit (except time and compute resources).

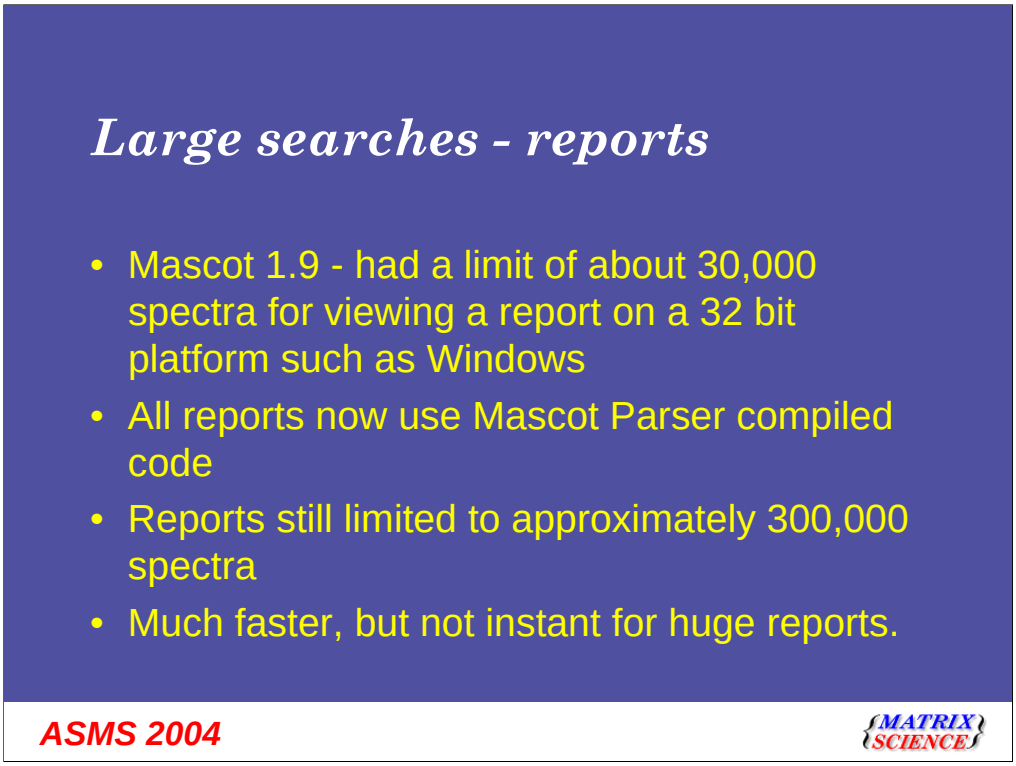
ASMS 2004

MATRIX SCIENCE

The first version of Mascot, which we released about five years ago had a limit of 300 ms-ms spectra in a single search. At the time, we couldn't conceive of anyone wanting more than this

Although Mascot 1.9 had a limit of 100,000 spectra, searches this size were only possible on 64 bit platforms. For Windows and Intel Linux systems, the practical limit was about 40,000 spectra.

Mascot 2.0 now has no such limit. The search is automatically split into chunks of 1000 spectra, and these are merged at the end of the search.



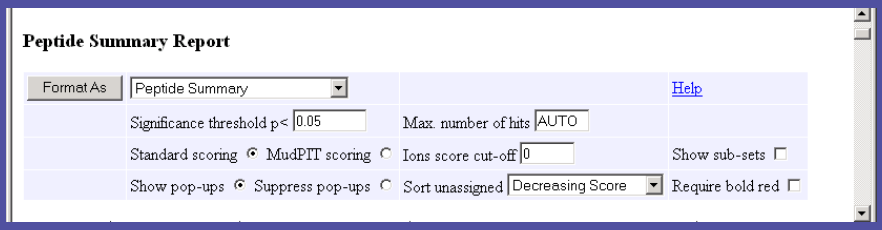
Viewing the results of huge searches in Mascot 1.9 was also an issue because of memory limitations on a 32 bit platforms

We've addressed this in Mascot 2.0 by using Mascot Parser. This makes production of the reports much less memory hungry.

There is still a limit - however, the limit is as much due to the size of a report that can be displayed in a browser as the amount of memory used by Mascot Parser.

Very large reports will still take some time to load.

Large searches - new options



- Format options introduced in Mascot 2.0.03 - upgrade patch available from web site
- New "Select Summary"
- Suppress pop-ups useful for large searches.

ASMS 2004

MATRIX SCIENCE

For the last few versions of Mascot, there have been a number of options for formatting the results output. However, the only way to invoke these options has been by adding them to the end of the URL. To make this easier, we have now added some form controls near the top of the results report

These options were only added in Mascot 2.0.03 - so if you have Mascot 2.0.0, please get the patches from the web site.

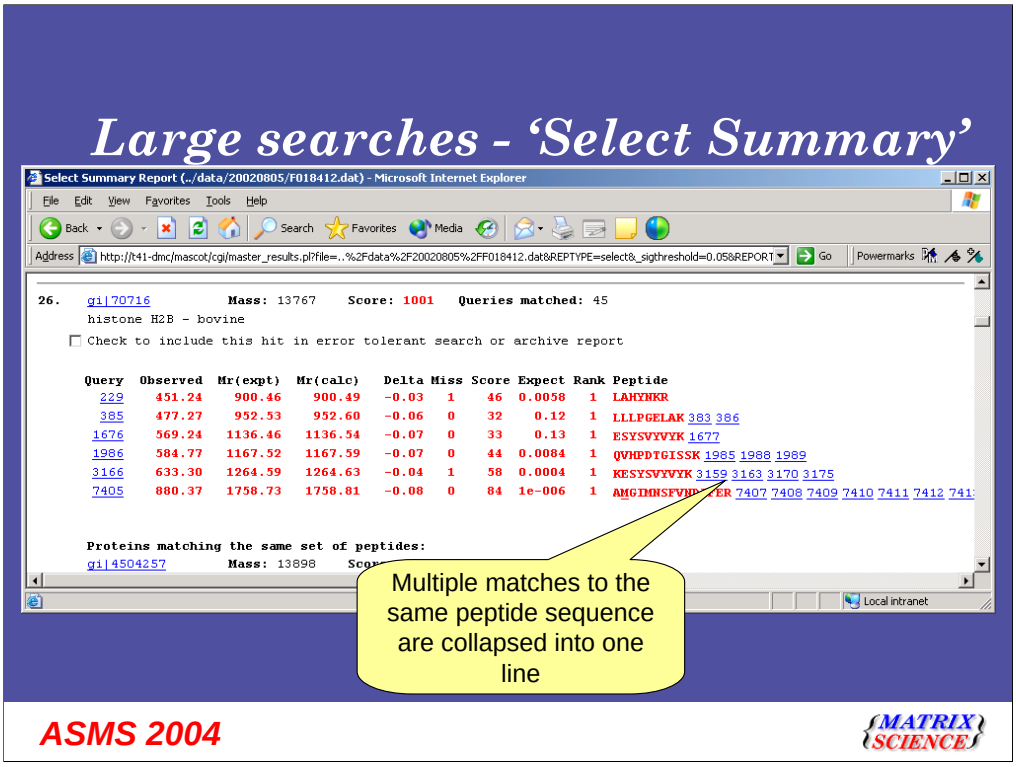
The Format As option allows choice of peptide and protein summary along with the new select summary.

The significance threshold is the protein cut off that is used if AUTO is entered as the number of hits.

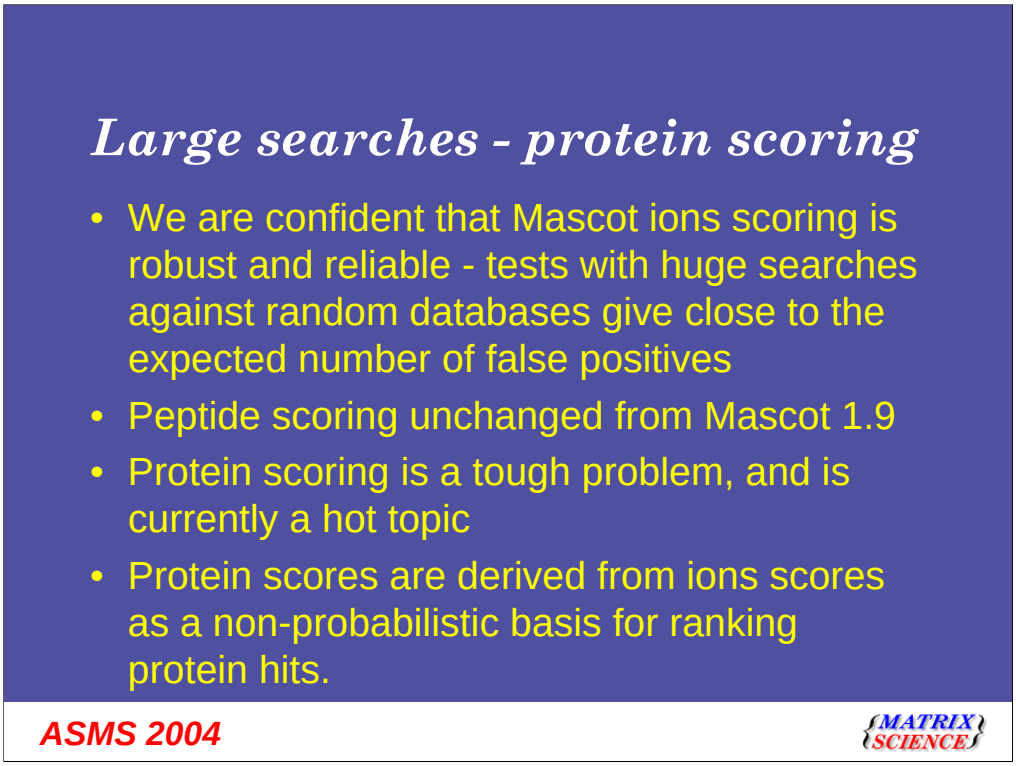
I'll describe standard versus MudPIT scoring in a minute, along with the Ions Score cut-off value.

The option to suppress the yellow pop-ups is useful with huge reports as these take up a lot of memory in Internet Explorer.

One final note, if you are using Internet Explorer with large searches, IE 5.5 is very sluggish - use 5.0 or 6.0



The Select Summary was inspired by David Tabb's DTASelect. It is very similar to the Peptide Summary, but more compact because multiple matches to the same peptide sequence are collapsed into a single line. Also, the list of peptide matches that are not assigned to a protein hit is split into a separate report.



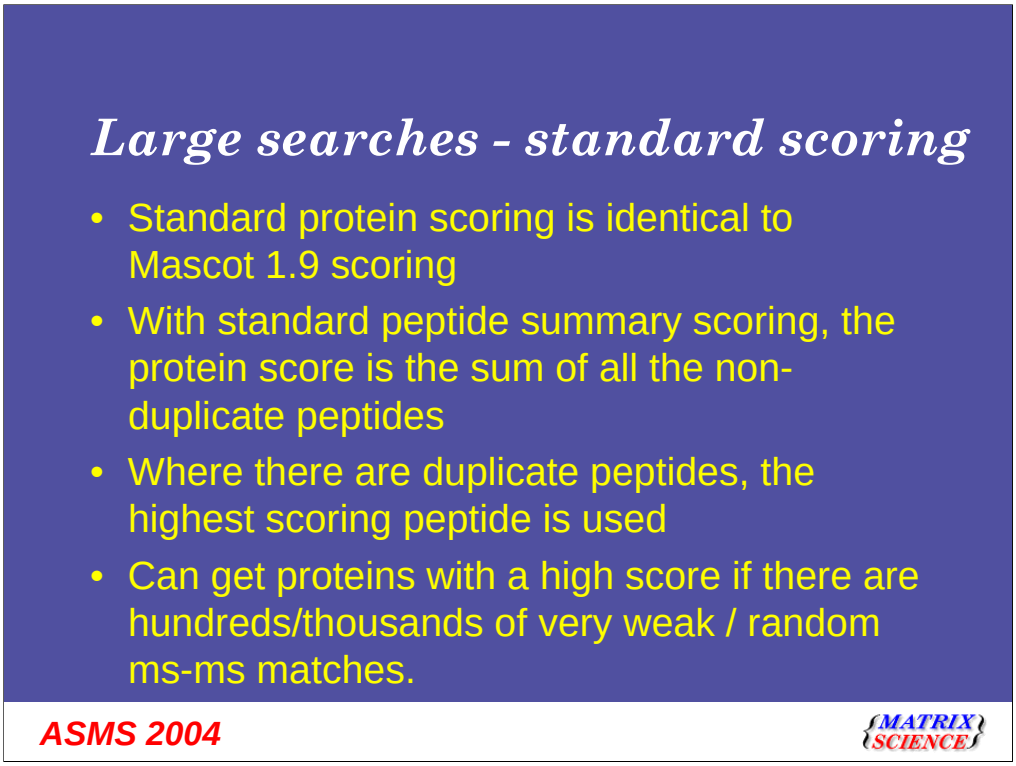
There are a number of issues with protein scoring when searches are huge, or when the number of spectra approaches the number of sequences in the database

Large MudPIT searches really test the Mascot peptide scoring - and we, along with many of our customers have run large data sets against randomised databases. We do find that we get close to the expected number of false positives, which justifies our confidence in the peptide scoring.

There are no changes to peptide scoring from Mascot 1.9

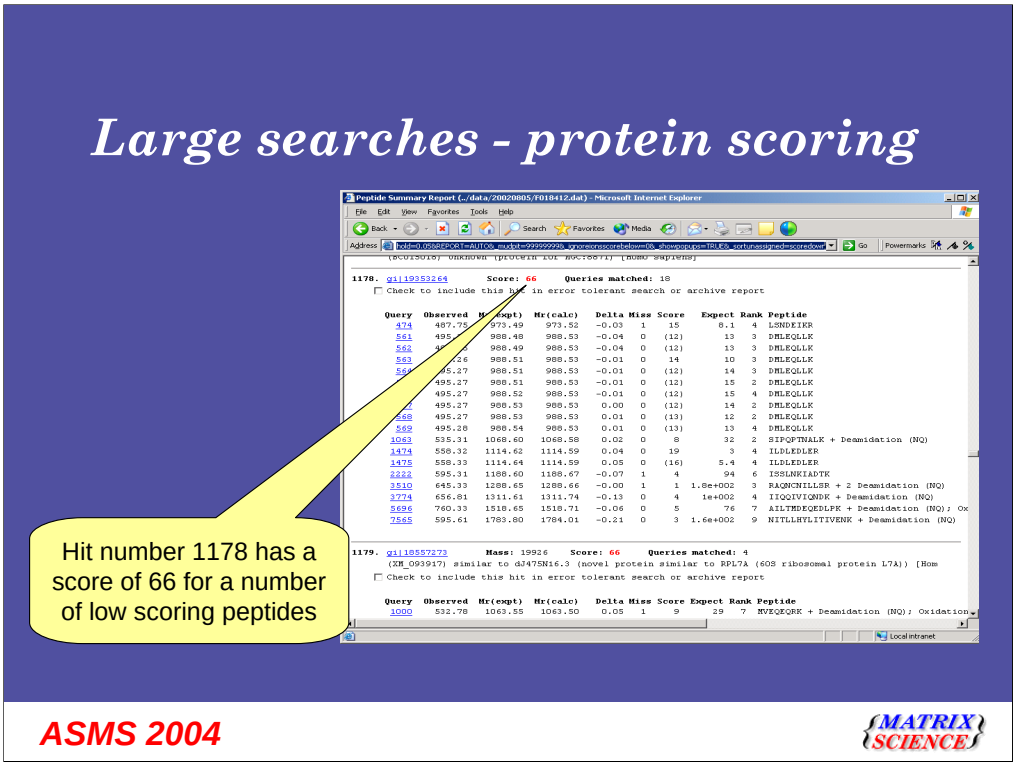
Protein scoring is a tough problem for a number of reasons, and it currently seems to be quite a hot topic.

Please note that Mascot does not assign strict probabilistic scores to proteins for MS/MS searches.

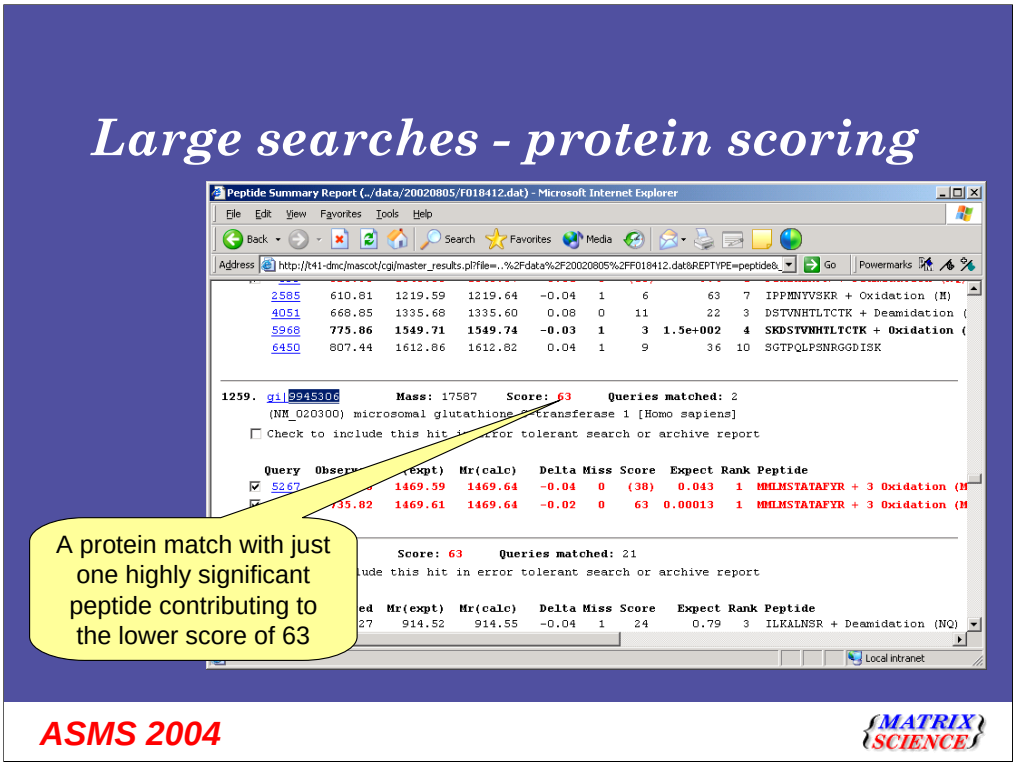


In Mascot 2.0, standard protein scoring is identical to the scoring in Mascot 1.9 - and this protein scoring works well with data sets up to several hundreds of spectra. Protein scores are derived by simply summing all of the non-duplicate peptide scores. Where there are duplicate peptides, the highest scoring peptide is used.

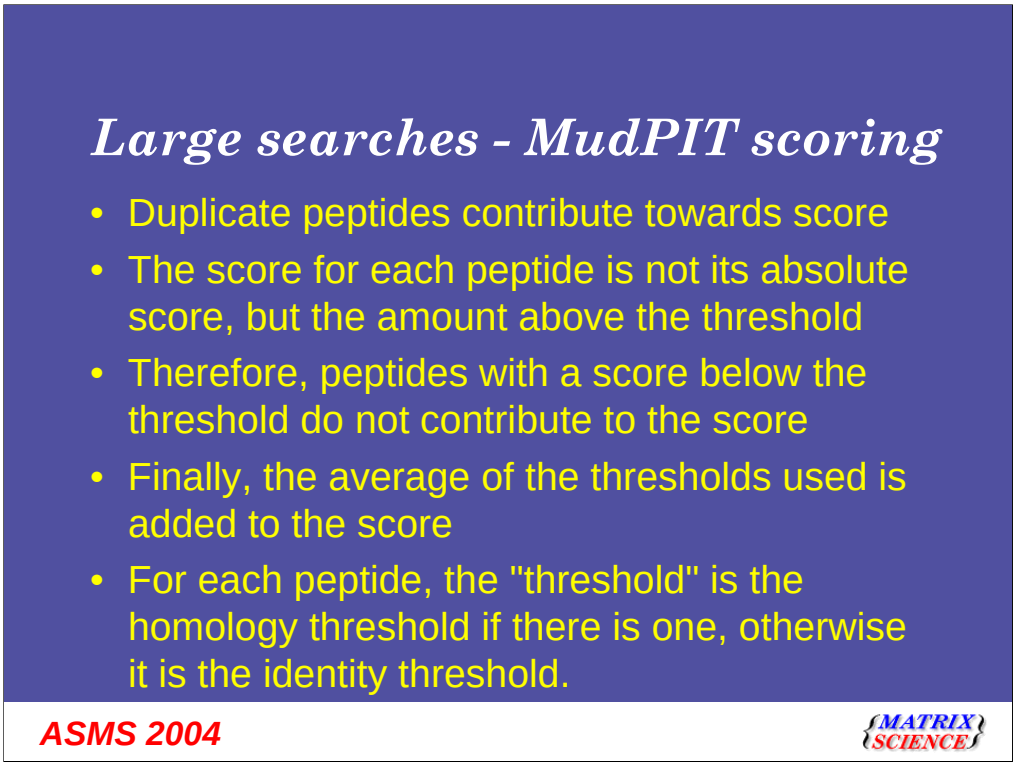
Although standard protein scoring works well with reasonable sized data sets, but it starts to show problems with huge MudPIT searches, where there are hundreds or thousands of low scoring peptide matches. I'll illustrate this problem:



This protein, at rank 1178, gets a score of 66, but you can see that the only evidence for this protein is a number of poor scoring peptides. Most people wouldn't consider this to be sufficient evidence for the existence of this protein.



Slightly further down the list, we see a protein that is much more believable. However, this protein gets a score lower than the previous one. Clearly we have a problem if we want to assign a cut-off point for proteins.



So, we have developed a new algorithm for calculating a protein score. These are the rules that we use:

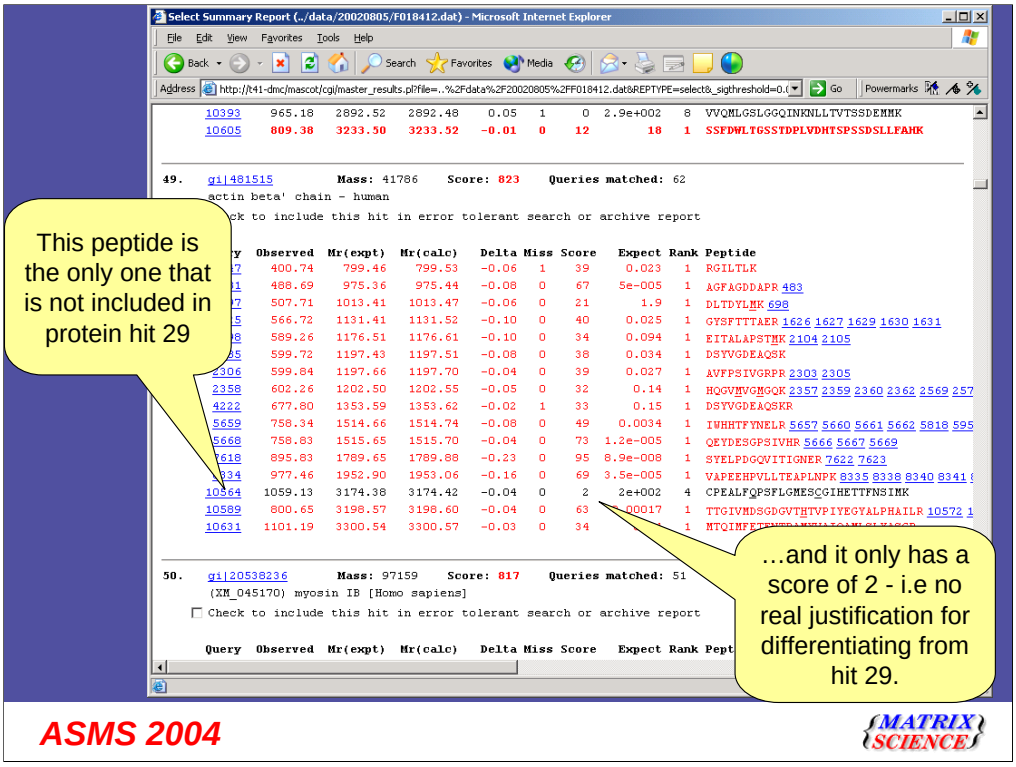
- duplicate peptides contribute towards the score. This seems more reasonable than for a gel based search because the spectra probably came from different fractions

- the score for each peptide is not its absolute score, but the difference between the score and the threshold. This means that low scoring peptides don't contribute to the score

- Finally, the average threshold is added to the score

- For each peptide, the "threshold" is the homology threshold if there is one, otherwise it is the identity threshold

Results show that this scoring does work well with large data sets - but I must emphasize that this is not true probability based scoring at the protein level.



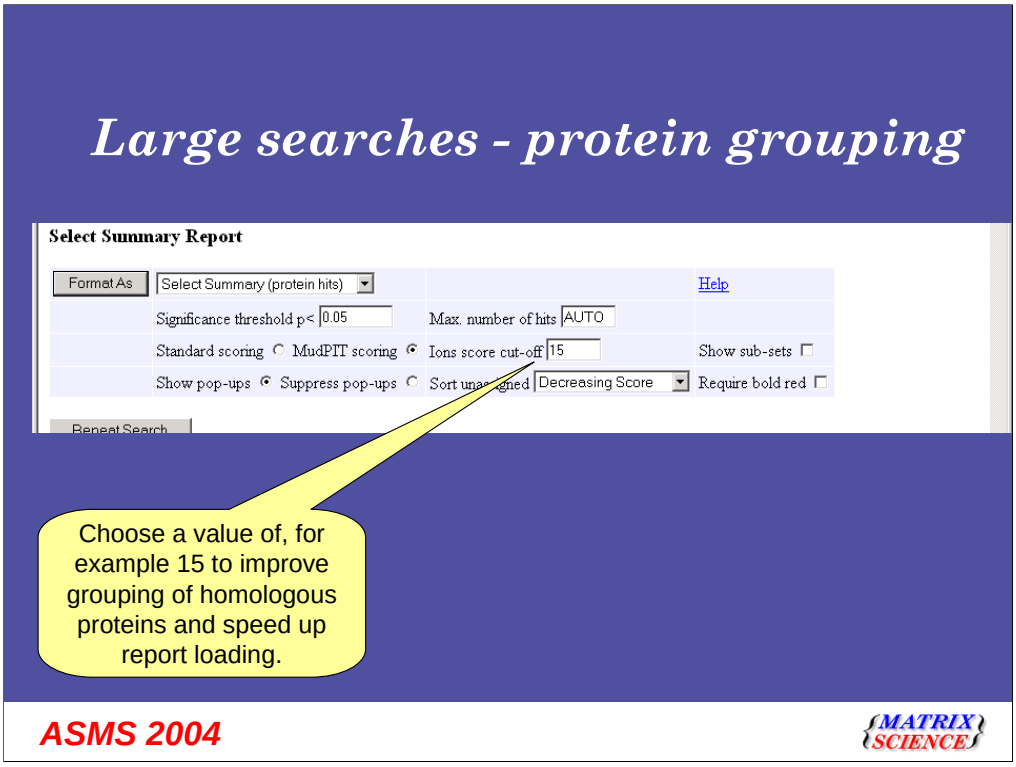
Next, I'd like to describe another problem that is particularly evident with MudPIT data, and how to prevent it.

Look at this hit - the first thing to point out is that there are no bold red or bold black peptides. Just as a reminder, the red indicates that the peptide is the highest scoring match to the MS/MS spectra. Bold typeface is used to show that this spectra has appeared in a higher scoring protein.

So, non-bold typeface indicates that all of these peptides have been seen in a protein with a higher score.

Looking in more detail, it appears that all of the peptides are identical to protein hit 29 - another actin beta chain protein.

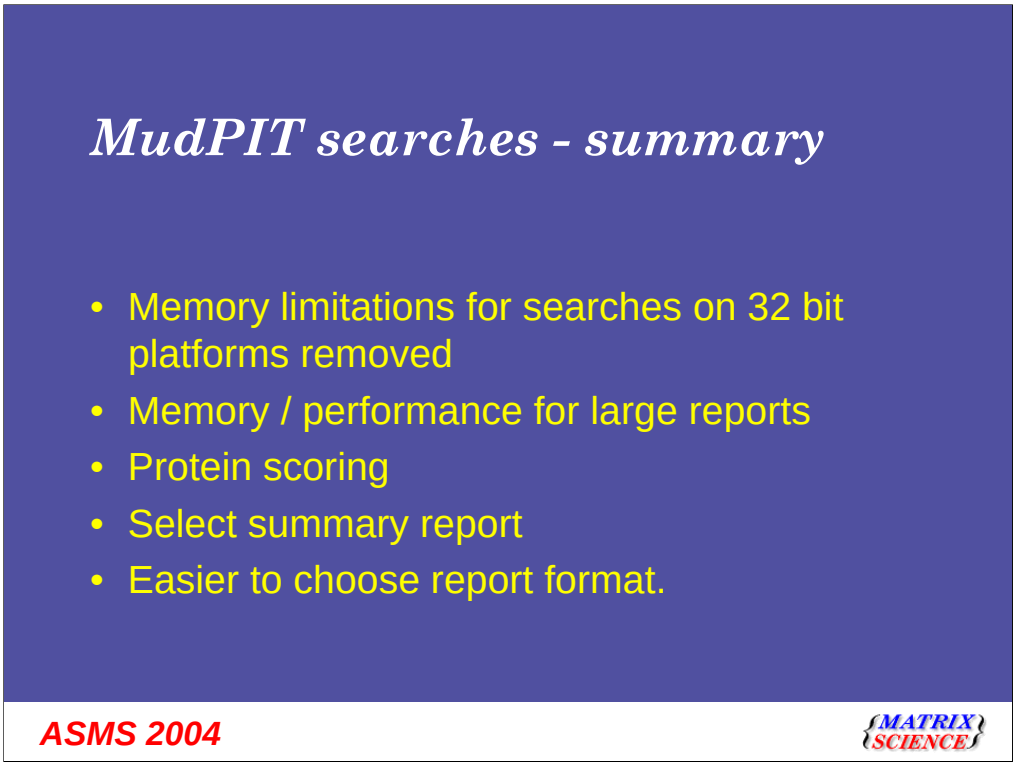
And since this peptide one only has a score of 2, there is no justification for differentiating it from protein hit 29.



So, there are two possible remedies to this - one is to choose "Require bold red", and the other is to set the ions score cut-off to a suitable value - for example 15.

Setting the ions score cut-off has the added benefit of making the reports load more quickly.

If you want to set a default value of 15, add an entry to the options section of mascot.dat - for details, click on the help



There were a number of changes to improve large MudPIT type searches in Mascot 2.0.

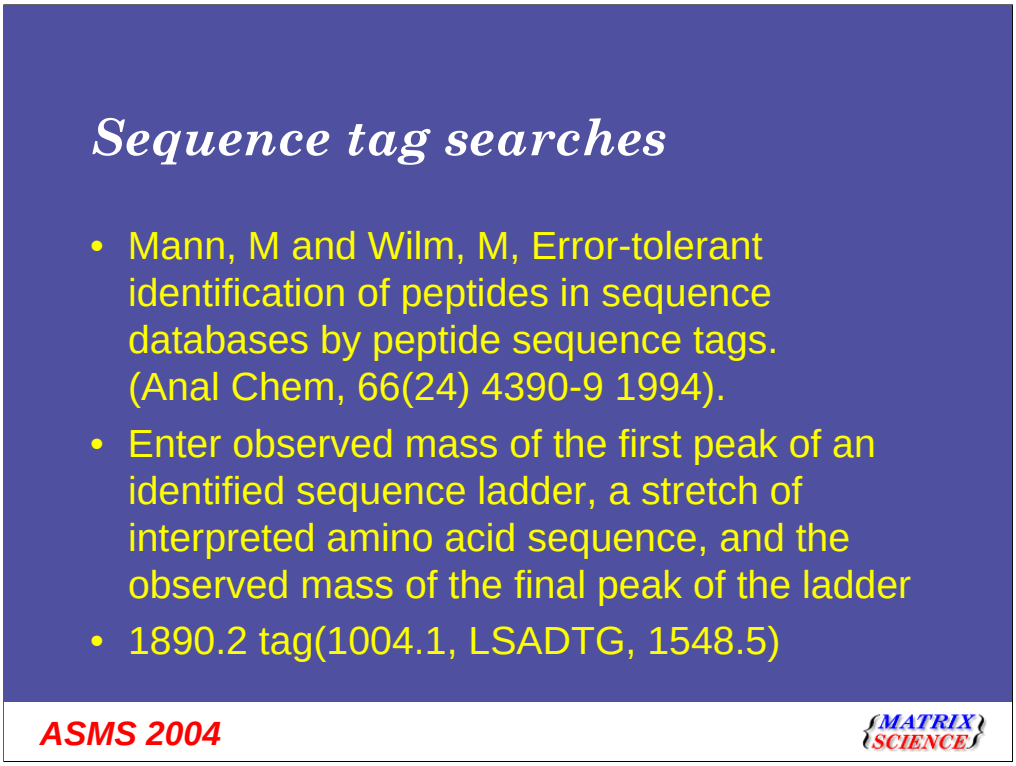
Firstly, on 32 bit platforms, some searches used to run out of memory with Mascot 1.9.

Secondly, viewing large reports was very slow and would occasionally run out of memory.

Thirdly, we have addressed some issues with protein scoring

Fourthly, we have introduced a new report

Finally, we have made it easier to switch between report formats



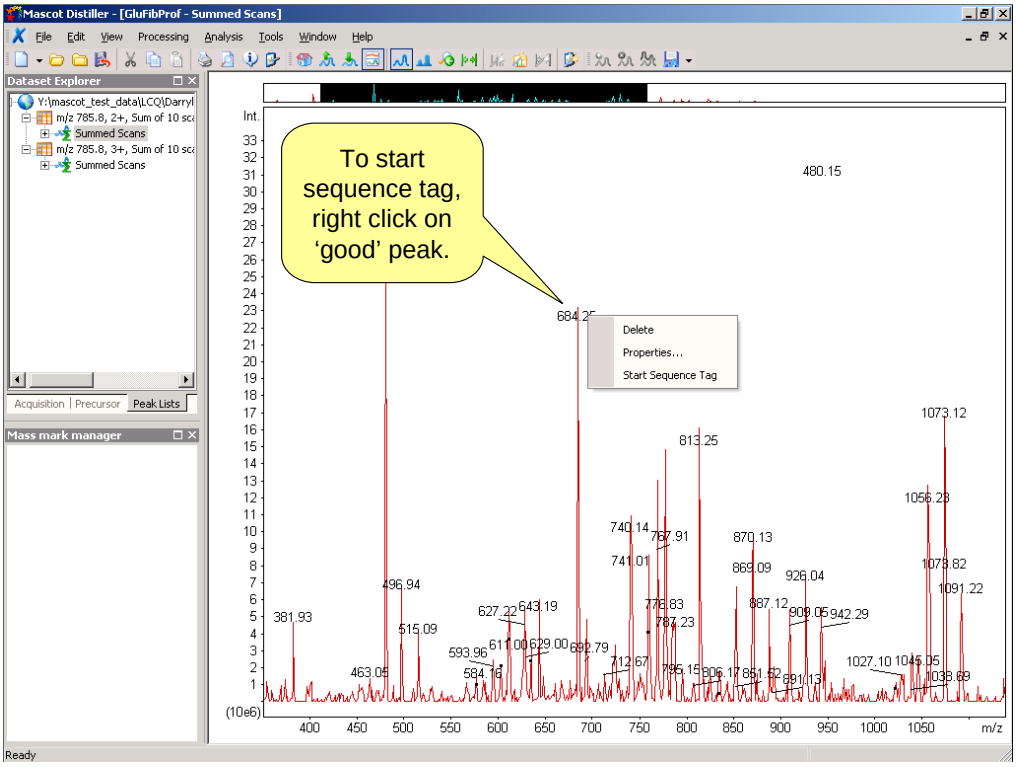
The sequence query, in which one or more peptide molecular masses are combined with sequence, composition and fragment ion data, is potentially the most powerful search of all. The usual source of the sequence information is interpretation of an MS/MS spectrum. While it is very difficult to determine a complete and unambiguous peptide sequence from an MS/MS spectrum, it is often possible to find a series of peaks providing 3 or 4 residues of sequence data.

This general approach was pioneered by Mann and co-workers at EMBL, who used the term "sequence tag" for the combination of a few residues of sequence data combined with molecular weight information. They defined a sequence tag derived from an MS/MS spectrum as the mass of the precursor peptide, the mass of the first peak of the identified sequence ladder, a stretch of interpreted sequence, and the mass of the final peak of the ladder.

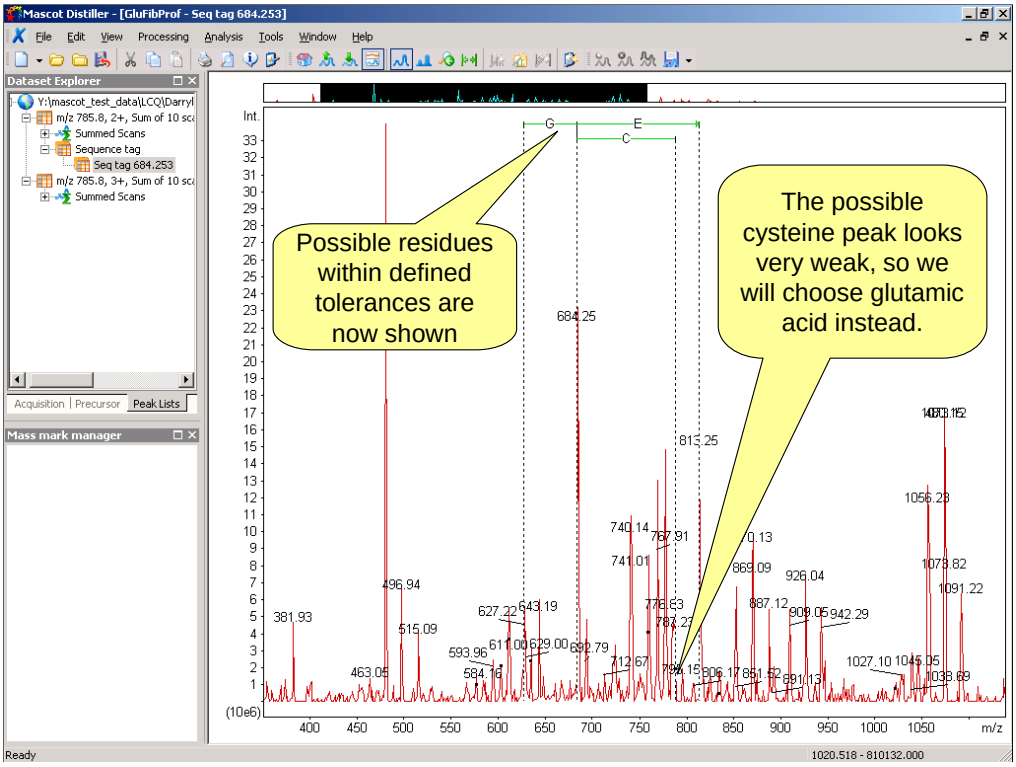
The sequence query mode of Mascot supports both standard and error tolerant sequence tags. It also allows arbitrary combinations of fragment ion mass values, amino acid sequence data and amino acid composition data to be searched.

The format is to enter the mass of the first peak of an identified sequence ladder, a stretch of interpreted amino acid sequence, and the observed mass of the final peak of the ladder - for example:

1890.2 tag(1004.1, LSADTG, 1548.5)



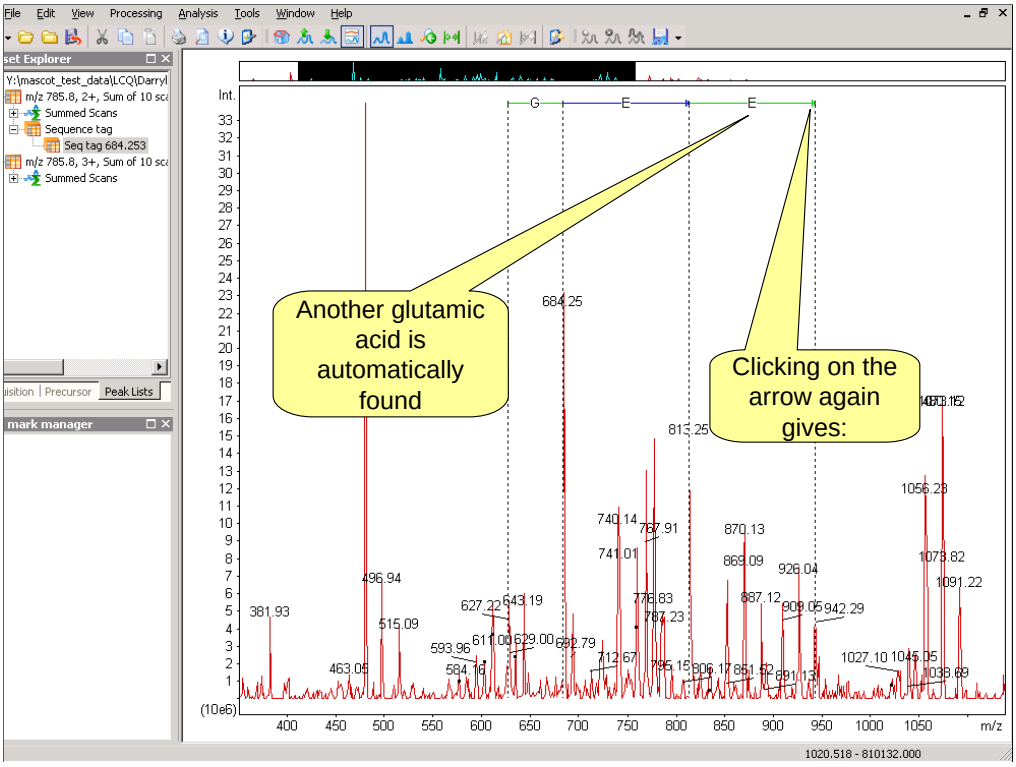
In the next version of Distiller, it is easy to generate sequence tags. Simply process the spectrum, right click on a strong peak, and select "Start Sequence Tag" from the pop-up menu



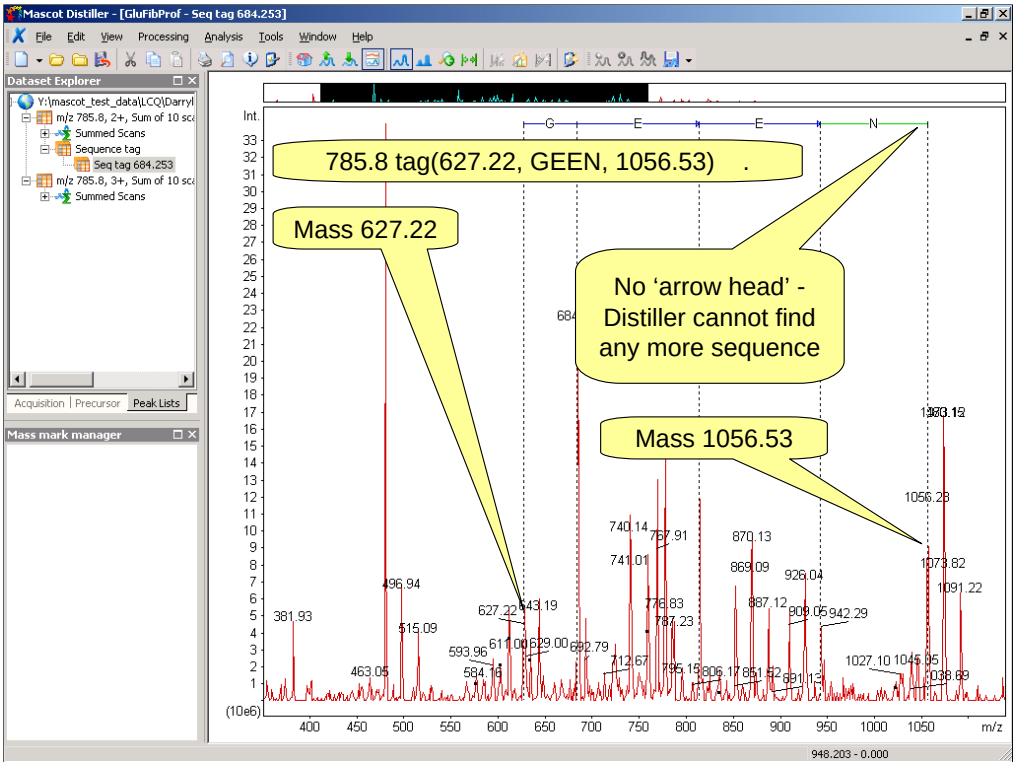
This then displays possible residue matches that fall within the specified tolerance.

The arrow head at the end of the glutamic acid indicates that Distiller can find another peptide after this one.

The lack of an arrow for the cysteine and the fact that the cysteine peak is very small means that it is best to choose the glutamic acid.



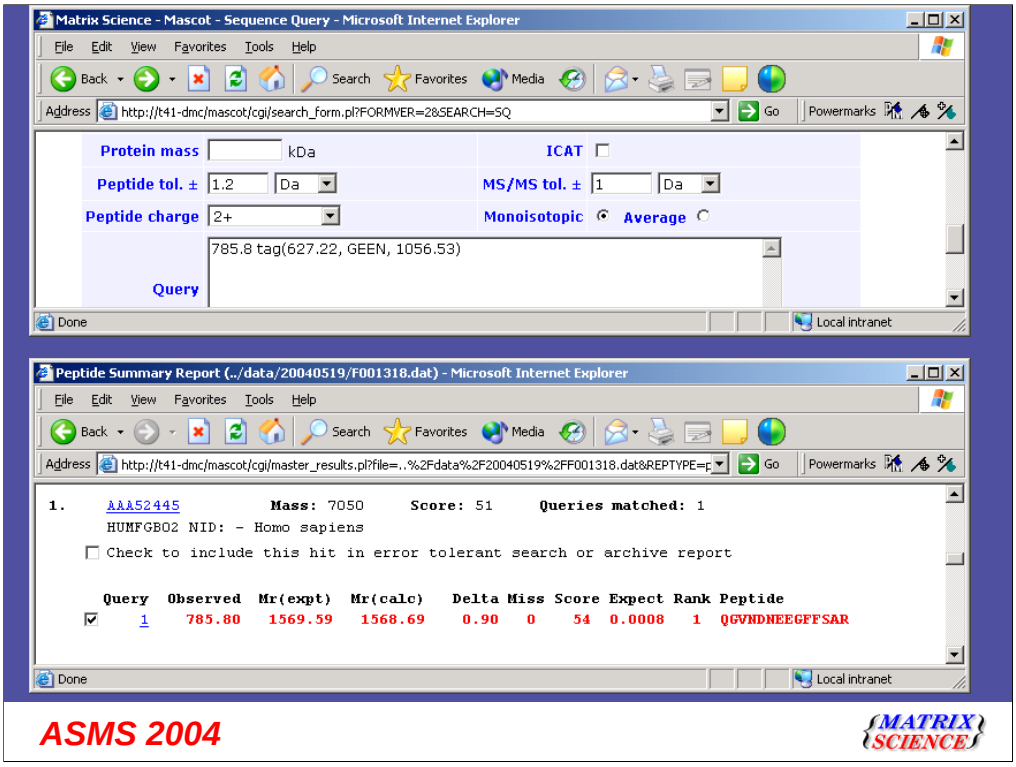
Having chosen the first glutamic acid, a second is then displayed - this time there are no other choices. Clicking on the arrow again gives:



And there are no further residues to be found in this sequence.

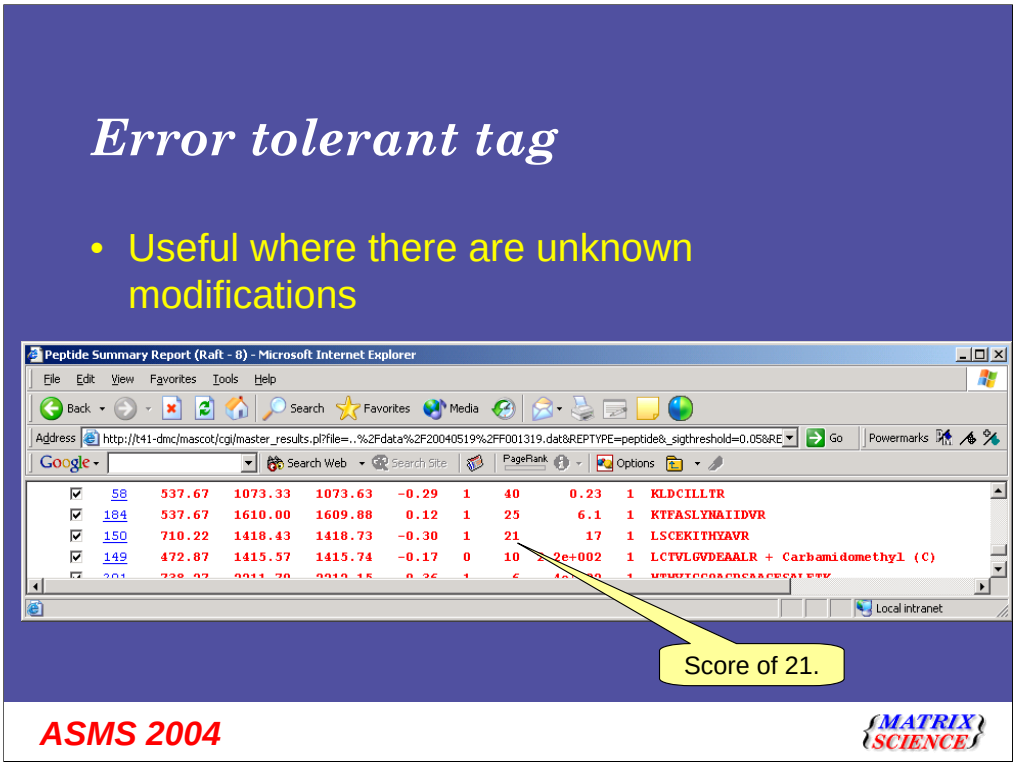
The start mass of this tag is 627.22, the end mass is 1056.53, so we can make a sequence tag like this

785.8 tag(627.22, GEEN, 1056.53)



It's possible to submit this automatically from Distiller, but it can also just put this in the query window in a sequence search form. Remember to enter the charge state of the precursor

Standard probability based Mascot scoring is used for these searches. This means that you can include several tags for each peptide and the one with the best match will get the highest score. It also means that it's possible to mix some ions and some sequence tags if desired.

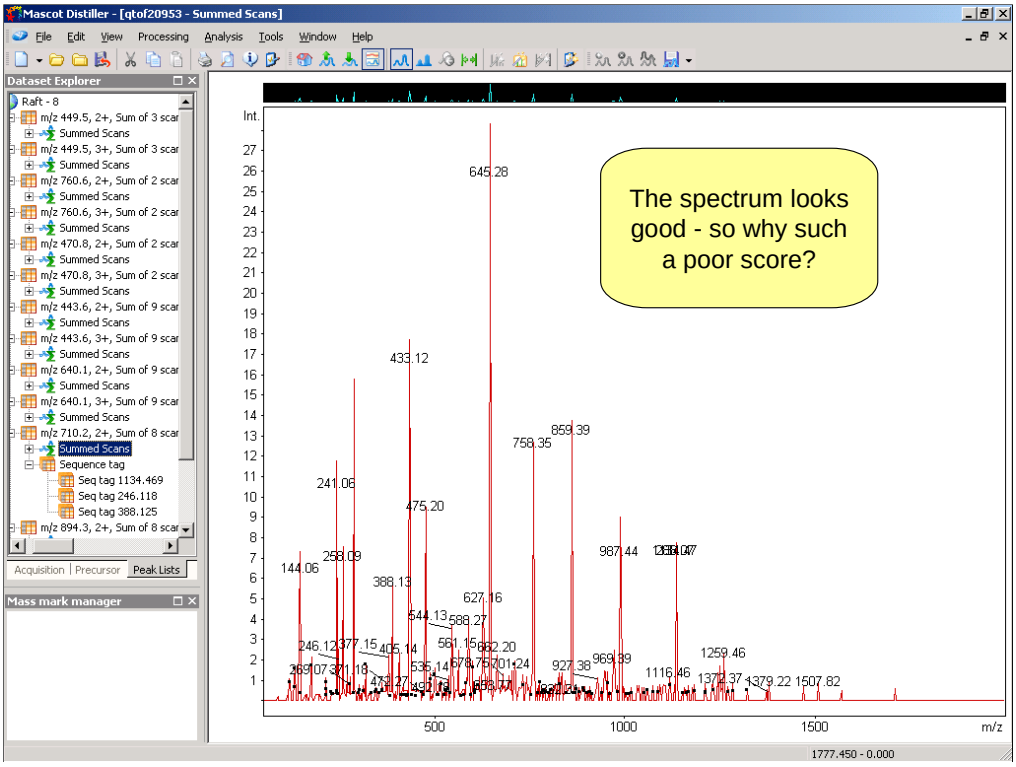


The tag approach really shows it's worth when there are unknown modifications, or when only a homologous sequence is in the database.

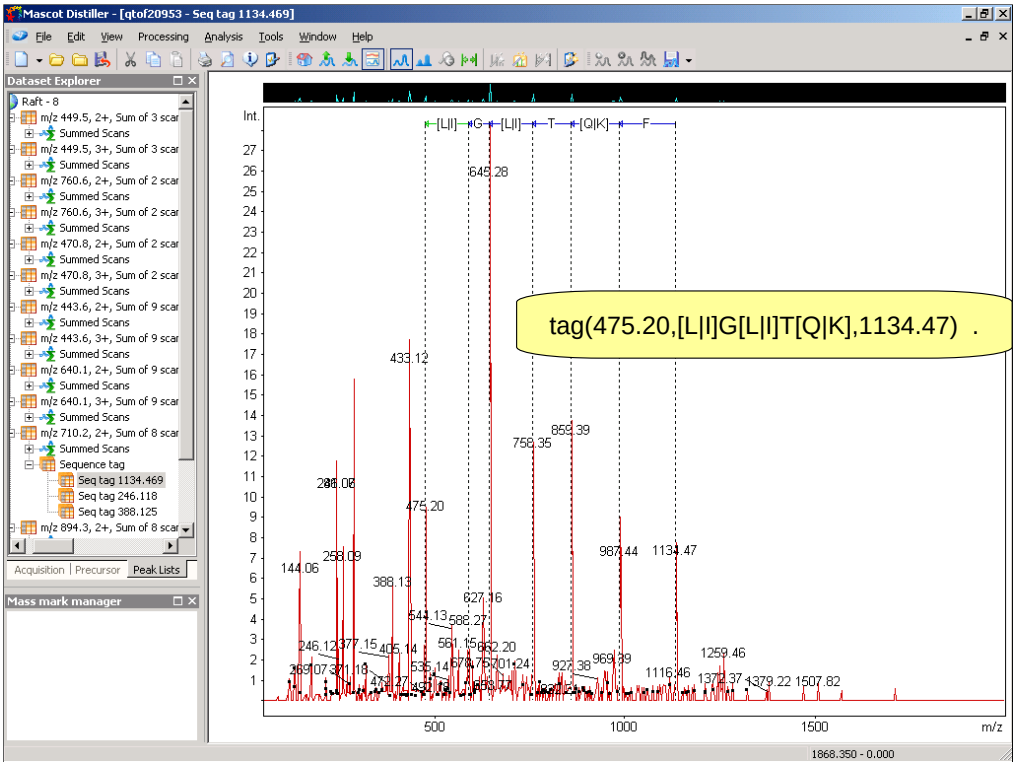
As an example, I'll take this peptide with a score of only 21 from the unassigned list.

Incidentally, one of the new report options is to sort the unassigned list by intensity. This can be quite useful for searching for the stronger spectra that didn't get database search matches.

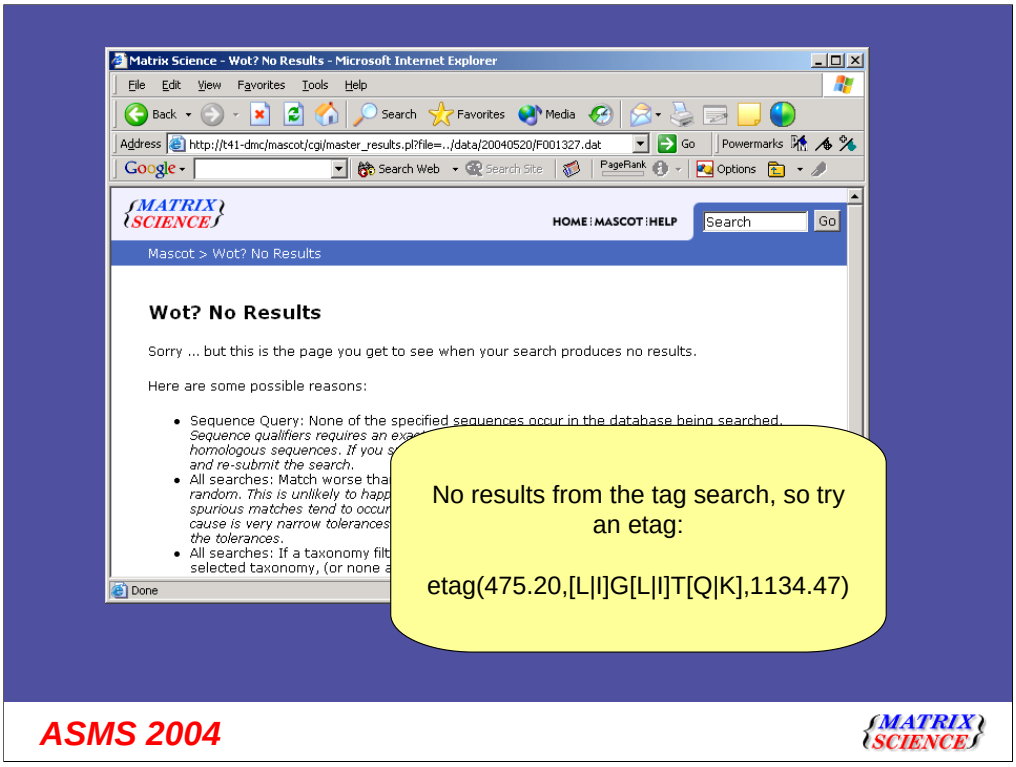
If we have a look at the spectrum



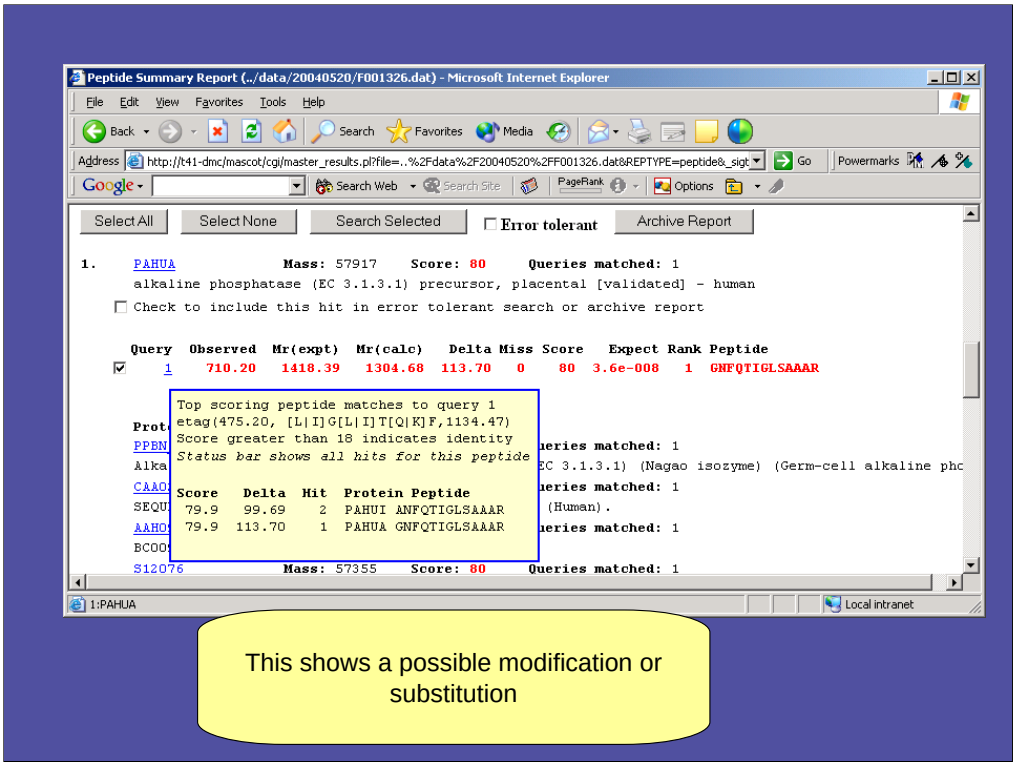
It is quite obviously of high quality, so it's not at all clear why there is no match.



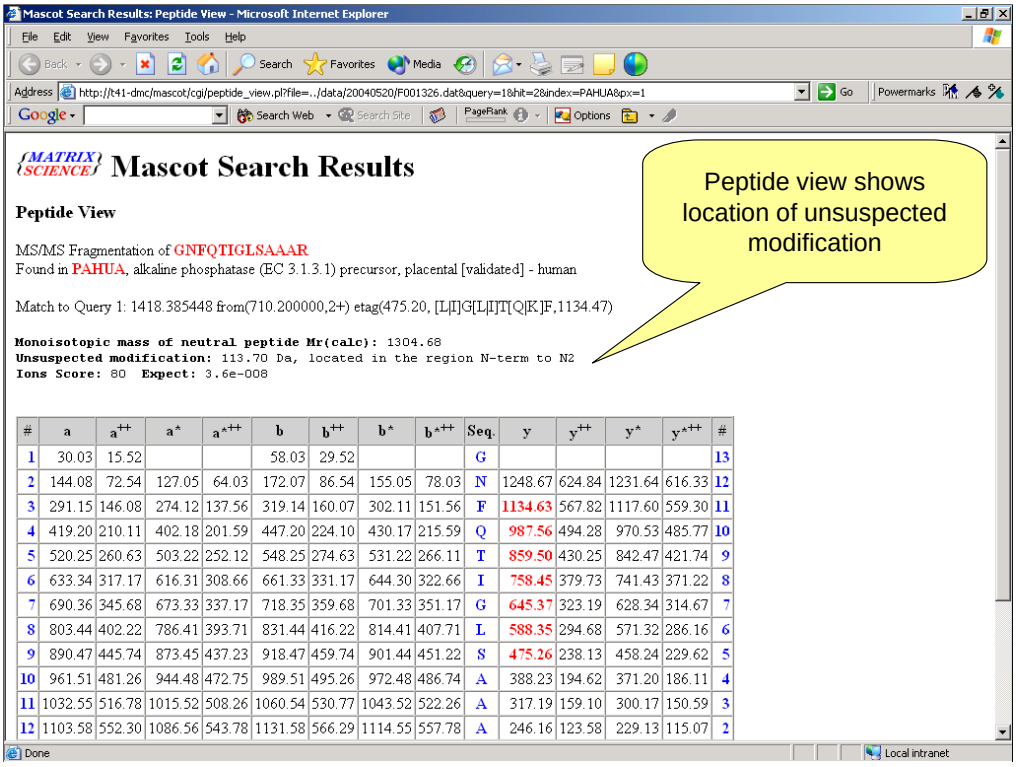
Once again, we the sequence tag is performed as before. Of course if you don't have distiller, you can always do this manually. The complication in this case is that within the tolerance specified, it's not possible to differentiate between glutamine and lysine, and of course Isoleucine and Leucine also have the same masses. So, in this case, we can specify a tag like this.



and we get absolutely no results... so we will try an error tolerant tag search:



and this time we do get a hit with quite a high score. Notice that there are two possible hits with the same score - the sequences have a different N terminus residue, and the delta in one case is 113.7 and the other is 99.69.



Mascot Search Results

Peptide View

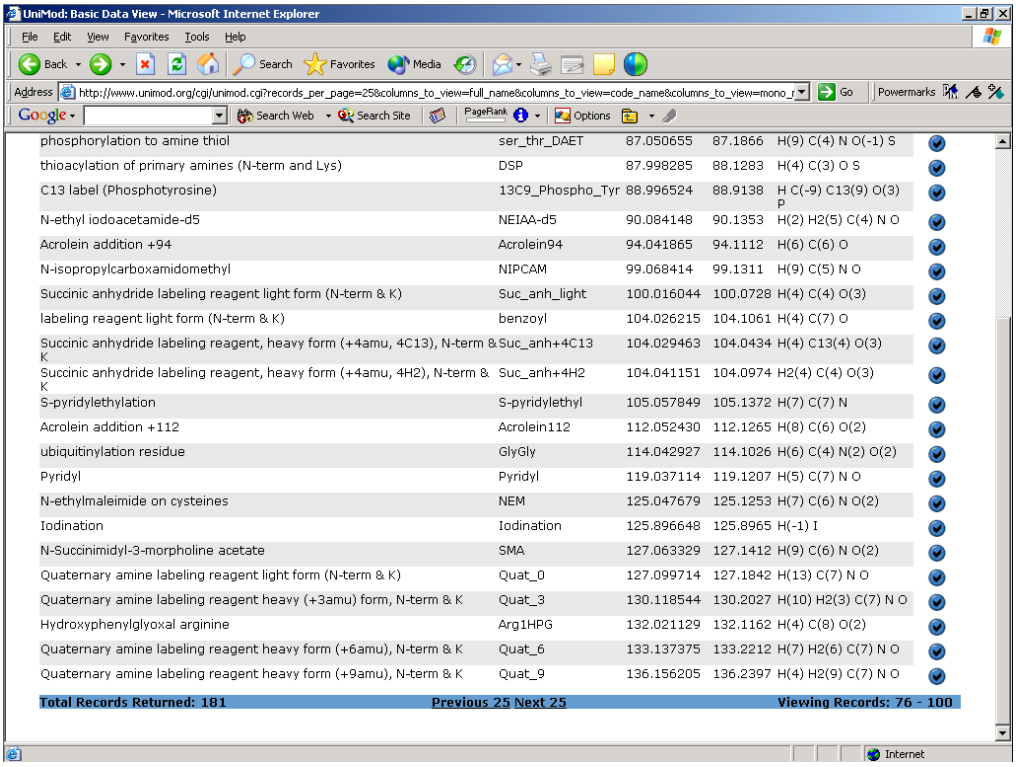
MS/MS Fragmentation of GNFQTIGLSAAAR
Found in PAHUA, alkaline phosphatase (EC 3.1.3.1) precursor, placental [validated] - human

Match to Query 1: 1418.385448 from(710.200000,2+) etag(475.20, [L|I]G[L|I]T[Q|K]F,1134.47)

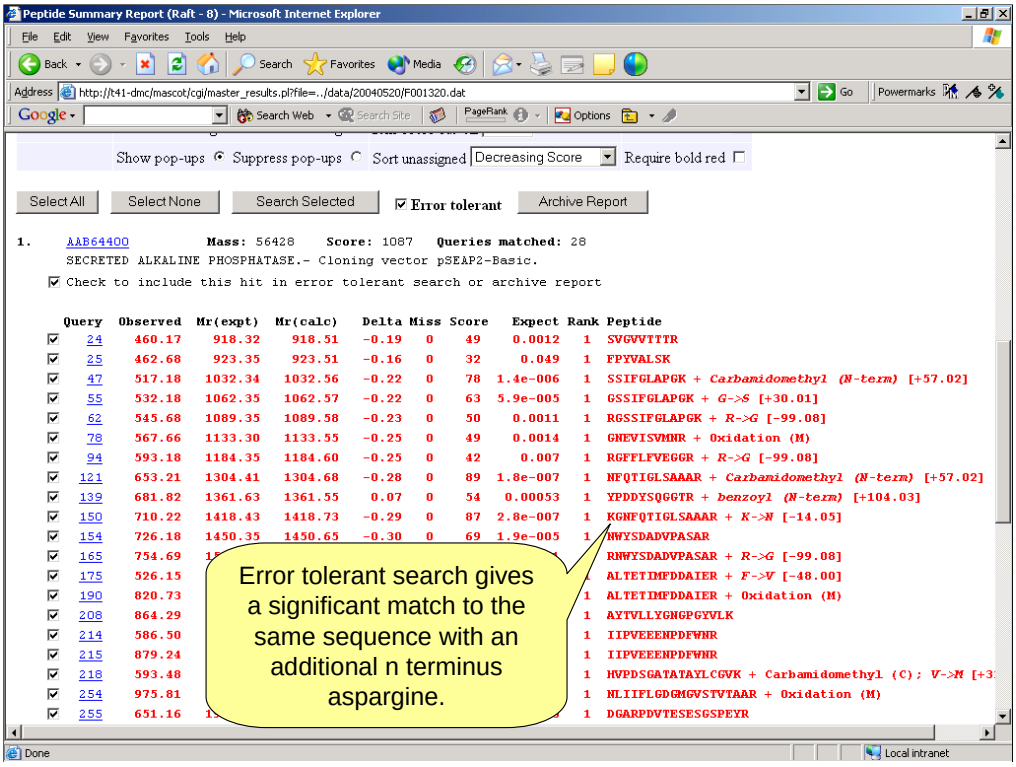
Monoisotopic mass of neutral peptide Mr(calc): 1304.68
Unsuspected modification: 113.70 Da, located in the region N-term to N2
Ions Score: 80 **Expect:** 3.6e-008

#	a	a^{++}	a*	a^{*++}	b	b^{++}	b*	b^{*++}	Seq.	y	y^{++}	y*	y^{*++}	#
1	30.03	15.52			58.03	29.52			G					13
2	144.08	72.54	127.05	64.03	172.07	86.54	155.05	78.03	N	1248.67	624.84	1231.64	616.33	12
3	291.15	146.08	274.12	137.56	319.14	160.07	302.11	151.56	F	1134.63	567.82	1117.60	559.30	11
4	419.20	210.11	402.18	201.59	447.20	224.10	430.17	215.59	Q	987.56	494.28	970.53	485.77	10
5	520.25	260.63	503.22	252.12	548.25	274.63	531.22	266.11	T	859.50	430.25	842.47	421.74	9
6	633.34	317.17	616.31	308.66	661.33	331.17	644.30	322.66	I	758.45	379.73	741.43	371.22	8
7	690.36	345.68	673.33	337.17	718.35	359.68	701.33	351.17	G	645.37	323.19	628.34	314.67	7
8	803.44	402.22	786.41	393.71	831.44	416.22	814.41	407.71	L	588.35	294.68	571.32	286.16	6
9	890.47	445.74	873.45	437.23	918.47	459.74	901.44	451.22	S	475.26	238.13	458.24	229.62	5
10	961.51	481.26	944.48	472.75	989.51	495.26	972.48	486.74	A	388.23	194.62	371.20	186.11	4
11	1032.55	516.78	1015.52	508.26	1060.54	530.77	1043.52	522.26	A	317.19	159.10	300.17	150.59	3
12	1103.58	552.30	1086.56	543.78	1131.58	566.29	1114.55	557.78	A	246.16	123.58	229.13	115.07	2

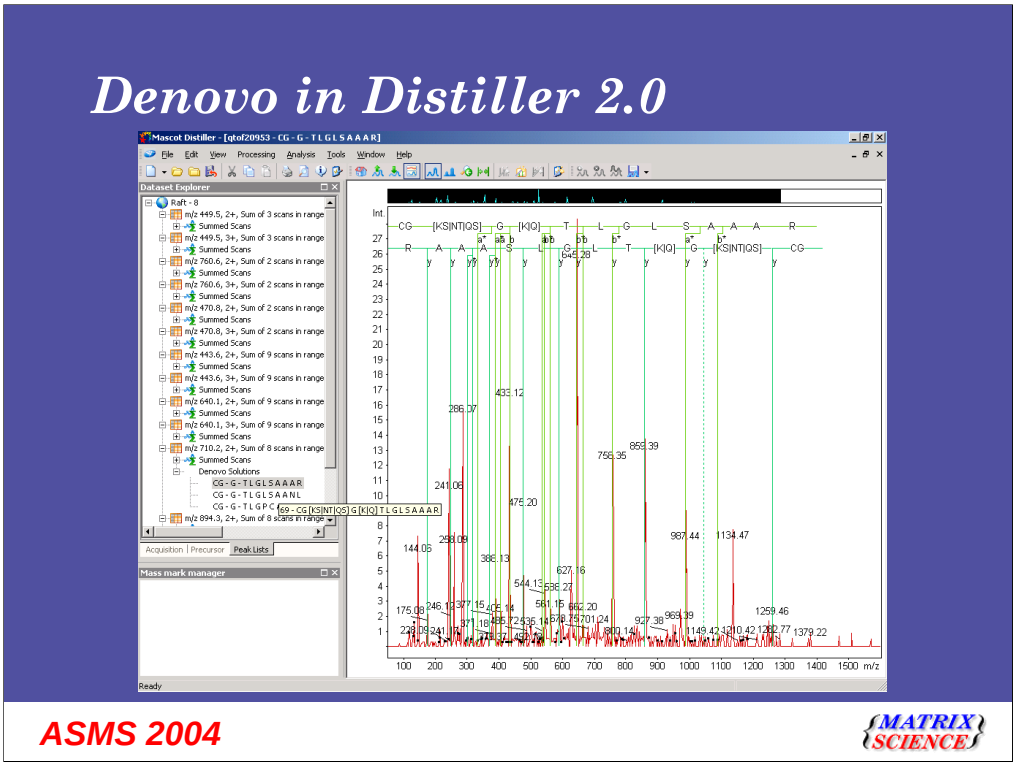
I was intrigued as to what this delta might be, and had a look on unimod.org -



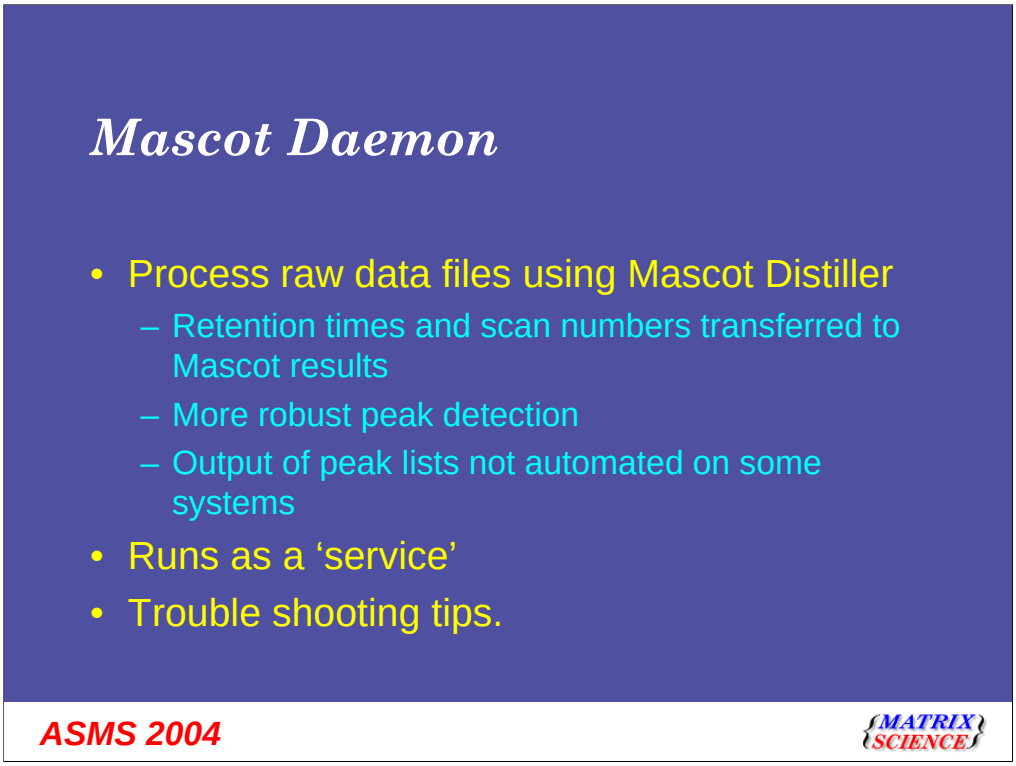
But the only thing listed here of this mass was ubiquitinylation - and this is unlikely because there's no lysine. Likewise, I found nothing suitable on Deltamass.



So, this really does seem to be unknown modification at the n terminus, and hence proves the power of the error tolerant search.



One other feature in the next version of distiller is that we have implemented a denovo algorithm - using this same spectrum, the denovo code found much of the same sequence automatically. Once again, we have used the proven Mascot scoring algorithm with Denovo. It will also be possible to submit these data as a tag search



Mascot Daemon had some major changes with version 2.0

- it is now possible to process raw data files using the Distiller libraries. This means that all the common raw data formats are now supported. The advantages are that retention times and scan numbers along with other information in the raw data files are transferred in the Mascot results. This information, which can be seen in the yellow pop-ups, makes it easy to find the original spectrum in the instrument data system or in the Distiller GUI.

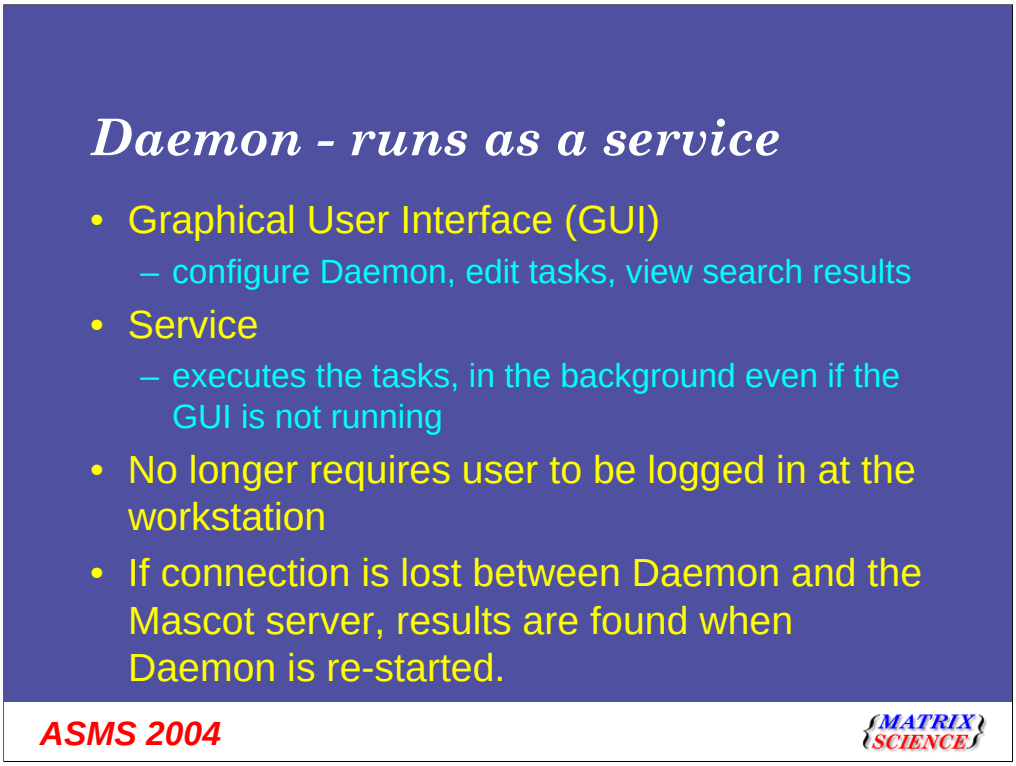
The peak detection in Distiller is proving to be among the best available. In addition, this makes it possible to automate the process of of peak detection, data reduction and searching that is not possible with some data systems

A major change was that the Daemon engine has been separated from the user interface. I'll describe this in more detail shortly.

I'll also describe a couple of trouble shooting tips.

There is not much to say about the Distiller options. There is an additional license cost for using Distiller from Daemon, and once the addition software has been installed, the "Mascot Distiller" data import filter should be available in the drop down list.

Clicking on the "Options", allows the choice and editing of processing parameters.

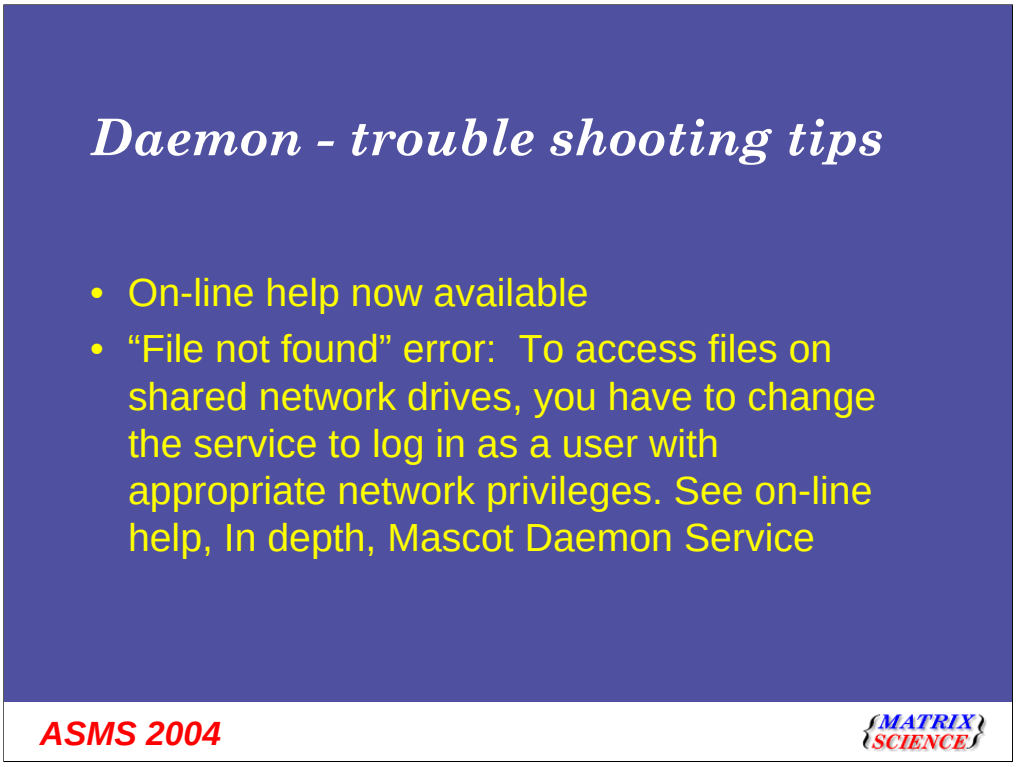


Mascot Daemon is now split into two distinct parts

- the GUI, which is used to configure Daemon, edit tasks and view search results

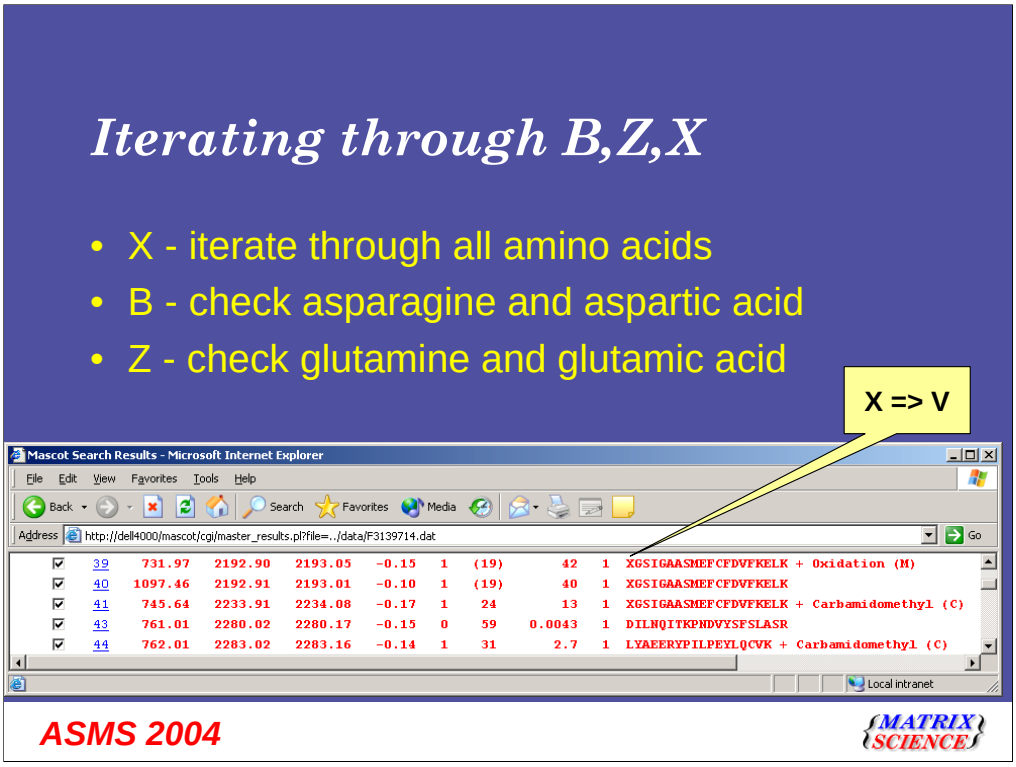
- a service with executes the tasks in the background even if the GUI is not running.

So, the advantage of this approach is that it no longer requires a user to be logged in at the workstation. Another change is that if the connection between Daemon and the Mascot server is lost, the results will be picked up by Daemon when the Daemon service is re-started.



There is added complexity in this version of Daemon, and for this reason we added comprehensive on-line help. So, if in doubt, please try reading that first.

The most common error that people experience is the "File not found" error when trying to access files on shared network drives. If you need to access remote files, you will need to change the service to log in as a user with appropriate network privileges. See on-line help, In depth, Mascot Daemon Service

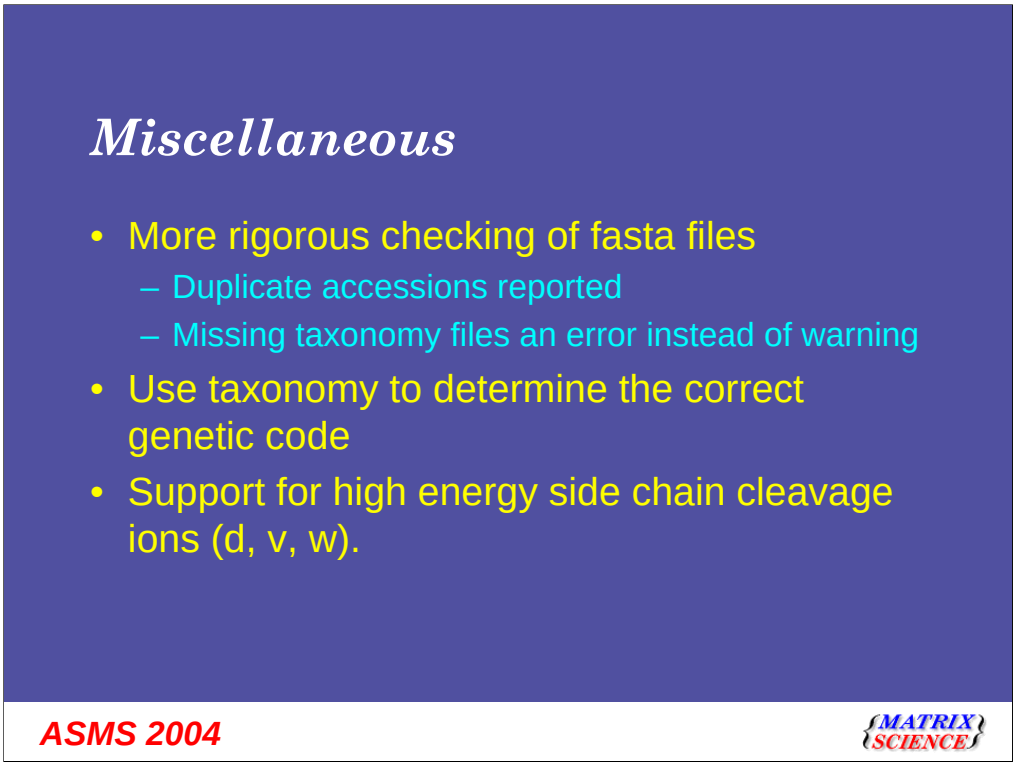


Another slightly over-due change to Mascot 2.0 is that where there is an 'X' in the sequence, the code now iterates through all possible amino acids

For B - Mascot checks asparagine and aspartic acid

For Z - Mascot checks glutamine and glutamic acid

Here's an example here of where Mascot found a peptide match by substituting an X for a valine



And finally, there are a number of small changes.

There is now more rigorous checking of fasta files. A number of our customers unfortunately had different sequences with the same accession numbers. This obviously causes confusion at best, and incorrect answers at worst, so this now causes an error, which needs to be fixed before searches can be run.

Similarly, missing taxonomy files used to just give a warning, but now it results in an error.

For EST searches and genomic data, Mascot now uses taxonomy to determine the correct genetic code translation.

Finally, we now support high energy side chain cleavage ions: d, v, and w.

Conclusions

- Huge searches now practical with Mascot 2.0
- A number of new report options
- tag and etag complementary to error tolerant searching
- Distiller support within Mascot Daemon enables highly integrated workflow
- A number of minor changes.