

New Features in Mascot 2.0

ASMS 2004



New Features in Mascot 2.0

- Web page re-design
- Large searches
- Tag searches
- Enhancements to Mascot Daemon
- Iteration of B, Z, X.

ASMS 2004

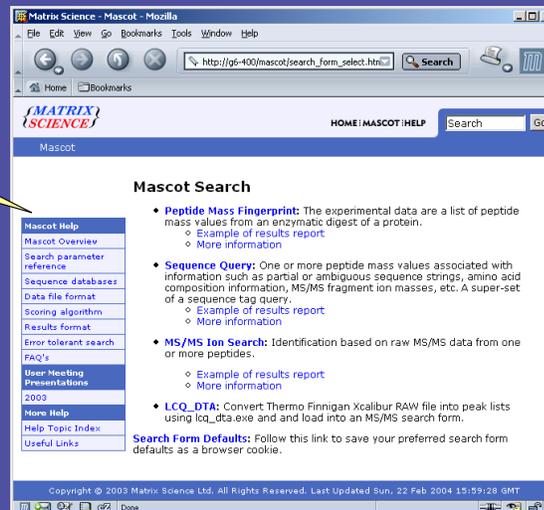
**{MATRIX}
{SCIENCE}**

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.

New Web Page Design

Context sensitive menu for easier navigation



ASMS 2004

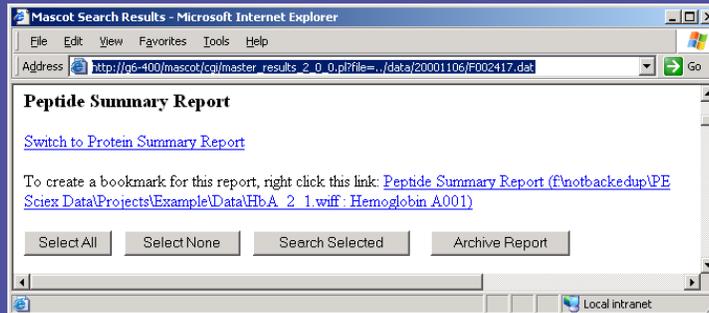


The first thing that that you will notice about Mascot 2.0 is that all the HTML pages have a new look.

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



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.

Large searches - reports

- Mascot 1.9 - had a limit of about 30,000 spectra for viewing a report on a 32 bit platform such as Windows
- All reports now use Mascot Parser compiled code
- Reports still limited to approximately 300,000 spectra
- Much faster, but not instant for huge reports.

ASMS 2004

{MATRIX}
{SCIENCE}

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

The screenshot shows a web-based form titled "Peptide Summary Report". It contains several controls for customizing the search results output:

- Format As:** A dropdown menu set to "Peptide Summary".
- Significance threshold p<:** A text input field containing "0.05".
- Max. number of hits:** A text input field containing "AUTO".
- Standard scoring:** A radio button that is selected.
- MudPIT scoring:** A radio button that is unselected.
- Ions score cut-off:** A text input field containing "0".
- Show sub-sets:** An unchecked checkbox.
- Show pop-ups:** A radio button that is selected.
- Suppress pop-ups:** A radio button that is unselected.
- Sort unassigned:** A dropdown menu set to "Decreasing Score".
- Require bold red:** An unchecked checkbox.

A "Help" link is visible in the top right corner of the form area.

- 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



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

Large searches - 'Select Summary'

Select Summary Report (./data/20020805/F018412.dat) - Microsoft Internet Explorer

Address: http://t41-dmc/mascot/cg/master_results.pl?file=.%2Fdata%2F20020805%2FF018412.dat&REPTYPE=select&sigthreshold=0.05&REPORT=

26. [gi170716](#) Mass: 13767 Score: 1001 Queries matched: 45
 histone H2B - bovine
 Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
229	451.24	900.46	900.49	-0.03	1	46	0.0058	1	LAKYKRR
385	477.27	952.53	952.60	-0.06	0	32	0.12	1	LLLPGLAK 383 386
1676	569.24	1136.46	1136.54	-0.07	0	33	0.13	1	ESYSVYVYK 1677
1986	584.77	1167.52	1167.59	-0.07	0	44	0.0084	1	QVHPDTGISSK 1985 1988 1989
3166	633.30	1264.59	1264.63	-0.04	1	58	0.0004	1	KESYSVYVYK 3159 3163 3170 3175
7405	880.37	1758.73	1758.81	-0.08	0	84	1e-006	1	AMGDMSPVPER 7407 7408 7409 7410 7411 7412 7413

Proteins matching the same set of peptides:
[gi14504257](#) Mass: 13898 Score: 1001

Multiple matches to the same peptide sequence are collapsed into one line

ASMS 2004 MATRIX SCIENCE

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.

Large searches - protein scoring

- We are confident that Mascot ions scoring is robust and reliable - tests with huge searches against random databases give close to the expected number of false positives
- Peptide scoring unchanged from Mascot 1.9
- Protein scoring is a tough problem, and is currently a hot topic
- Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.

ASMS 2004

**MATRIX
SCIENCE**

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.

Large searches - standard scoring

- Standard protein scoring is identical to Mascot 1.9 scoring
- With standard peptide summary scoring, the protein score is the sum of all the non-duplicate peptides
- Where there are duplicate peptides, the highest scoring peptide is used
- Can get proteins with a high score if there are hundreds/thousands of very weak / random ms-ms matches.

ASMS 2004

{MATRIX}
{SCIENCE}

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:

Large searches - protein scoring

Peptide Summary Report (.data/20020805/FOI8412.daa) - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/peptide/summary/summary.cgi?query=FOI8412.daa

1178. [q11193532.64](#) Score: 66 Queries matched: 10
 Check to include this hit in error tolerant search or archive report

Query	Observed	M ₀ (opt)	M ₁ (calc)	Delta	Miss	Score	Expect	Rank	Peptide
474	487.75	973.49	973.52	-0.03	1	15	8.1	4	LSNDEIKR
581	495.27	988.51	988.53	-0.04	0	(12)	13	3	DMLEQLLK
582	495.27	988.49	988.53	-0.04	0	(12)	13	3	DMLEQLLK
583	495.27	988.51	988.53	-0.01	0	14	10	3	DMLEQLLK
659	495.27	988.51	988.53	-0.01	0	(12)	14	3	DMLEQLLK
660	495.27	988.51	988.53	-0.01	0	(12)	15	2	DMLEQLLK
661	495.27	988.52	988.53	-0.01	0	(12)	15	4	DMLEQLLK
662	495.27	988.53	988.53	0.00	0	(12)	14	2	DMLEQLLK
663	495.27	988.53	988.53	0.01	0	(13)	12	2	DMLEQLLK
592	495.28	988.54	988.53	0.01	0	(13)	13	4	DMLEQLLK
1083	535.31	1068.60	1068.58	0.02	0	8	32	2	SIFQPTNALK + Deamidation (NQ)
1474	558.32	1114.62	1114.59	0.04	0	19	3	4	ILDEDLER
1475	558.33	1114.64	1114.59	0.05	0	(16)	5	4	ILDEDLER
2222	595.31	1188.60	1188.67	-0.07	1	4	94	6	ISSLNKIDYK
3510	645.33	1288.65	1288.66	-0.00	1	1	1.8e+002	3	RAQNCNILLSR + 2 Deamidation (NQ)
3724	656.91	1311.81	1311.74	-0.13	0	4	1e+002	4	IIGQIVIQNDK + Deamidation (NQ)
5295	760.33	1518.65	1518.71	-0.06	0	5	76	7	AILTRERQDQPE + Deamidation (NQ); Oxidation
7555	595.61	1788.80	1784.01	-0.21	0	3	1.6e+002	9	NITLLRVLITIVENK + Deamidation (NQ)

1179. [q11185572.73](#) Mass: 19926 Score: 66 Queries matched: 4
 (XM_093917) similar to d3475N16.3 (novel protein similar to RPL7A (60S ribosomal protein L7A)) [Hom]
 Check to include this hit in error tolerant search or archive report

Query	Observed	M ₀ (opt)	M ₁ (calc)	Delta	Miss	Score	Expect	Rank	Peptide
1080	532.78	1063.55	1063.50	0.05	1	9	29	7	NVVEQQRK + Deamidation (NQ); Oxidation

ASMS 2004



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.

Large searches - protein scoring

Peptide Summary Report (C:/data/20020805/F018412.dat) - Microsoft Internet Explorer

Address: http://r41-dmc/mascot/cgi/master_results.pl?file=...&REPTYPE=peptide&...

Peptide	Mass	Score	Queries matched						
2585	610.81	1219.59	1219.64	-0.04	1	6	63	7	IPFNNIVSKR + Oxidation (M)
4051	668.85	1335.68	1335.60	0.08	0	11	22	3	DSTVNHLLTCTK + Deamidation (N)
5968	775.86	1549.71	1549.74	-0.03	1	3	1.5e+002	4	SKDSTVNHLLTCTK + Oxidation (M)
6450	807.44	1612.86	1612.82	0.04	1	9	36	10	SGTLPQPSNRGGDISK

1259. [gi|3945306](#) Mass: 17587 Score: **63** Queries matched: 2
(NM_020300) microsomal glutathione S-transferase 1 [Homo sapiens]

Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide	
<input checked="" type="checkbox"/>	5267	1469.59	1469.64	-0.04	0	(38)	0.043	1	MHLNSTATAFYR + 3 Oxidation (M)	
<input checked="" type="checkbox"/>	5267	1469.82	1469.61	1469.64	-0.02	0	63	0.00013	1	MHLNSTATAFYR + 3 Oxidation (M)

Score: 63 Queries matched: 21

Include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input type="checkbox"/>	27	914.52	914.55	-0.04	1	24	0.79	3	ILKALNSR + Deamidation (NQ)

A protein match with just one highly significant peptide contributing to the lower score of 63

ASMS 2004



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.

Large searches - MudPIT scoring

- Duplicate peptides contribute towards score
- The score for each peptide is not its absolute score, but the amount above the threshold
- Therefore, peptides with a score below the threshold do not contribute to the score
- Finally, the average of the thresholds used is added to the score
- For each peptide, the "threshold" is the homology threshold if there is one, otherwise it is the identity threshold.

ASMS 2004

**MATRIX
SCIENCE**

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.

Select Summary Report (../data/20020805/F018412.dat) - Microsoft Internet Explorer

Address: http://41-dmc/mascot/cj/master_results.pl?file=...%2Fdata%2F20020805%2FF018412.dat&REPTYPE=select&sigthres=0.01

Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide	
10393	965.18	2892.52	2892.48	0.05	1	0	2.9e+002	8 VVQNLGSLGGQINRNLTLVTSSDEMVK	
10605	809.38	3233.50	3233.32	-0.01	0	12	18	1 SSFDWLTGSSDPLVDHTSPSSDLSLFAHK	
49. g11481515 Mass: 41786 Score: 823 Queries matched: 62									
actin beta' chain - human									
Check to include this hit in error tolerant search or archive report									
Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide	
2	400.74	799.46	799.53	-0.06	1	39	0.023	1 RGILTLK	
11	488.69	975.36	975.44	-0.08	0	67	5e-005	1 AGFAGDDAPR 463	
17	507.71	1013.41	1013.47	-0.06	0	21	1.9	1 DLTDYLHK 698	
15	566.72	1131.41	1131.52	-0.10	0	40	0.025	1 GYSFTTAAER 1626 1627 1629 1630 1631	
18	589.26	1176.51	1176.61	-0.10	0	34	0.094	1 EITALAPSTR 2104 2105	
25	599.72	1197.43	1197.51	-0.08	0	38	0.034	1 DSYVGEAQS	
2306	599.84	1197.66	1197.70	-0.04	0	39	0.027	1 AVFPSIVGRPR 2303 2305	
2358	602.26	1202.50	1202.55	-0.05	0	32	0.14	1 HQGVIVGQGR 2357 2359 2360 2362 2369 257	
4222	677.80	1353.59	1353.62	-0.02	1	33	0.15	1 DSYVGEAQS	
5659	758.34	1514.66	1514.74	-0.08	0	49	0.0034	1 IWHHTYNELR 5657 5660 5661 5662 5818 595	
5668	758.83	1515.65	1515.70	-0.04	0	73	1.2e-005	1 QEYDESGSIVHR 5666 5667 5669	
618	895.83	1789.65	1789.88	-0.23	0	95	8.9e-008	1 SYELPQGVITIGNER 7622 7623	
834	977.46	1952.90	1953.06	-0.16	0	69	3.5e-005	1 VAPEEHPVLLTEAPLNRY 8325 8338 8340 8341	
10354	1059.13	3174.38	3174.42	-0.04	0	2	2e+002	4 CPEALFQPSFLGMSGSHETTFNSIMK	
10589	800.65	3198.57	3198.60	-0.04	0	63	0.0017	1 TTGIVNDSGDGVTHVPIYEGYALPHALLR 10572 1	
10631	1101.19	3300.54	3300.57	-0.03	0	34	1	1 HTQINPT...	
50. g120538236 Mass: 97159 Score: 817 Queries matched: 51									
(XM_045170) myosin IB [Homo sapiens]									
Check to include this hit in error tolerant search or archive report									
Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide

ASMS 2004

MATRIX SCIENCE

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.

Large searches - protein grouping

Select Summary Report

Format As	Select Summary (protein hits)	Help					
Significance threshold p<	0.05	Max. number of hits	AUTO				
Standard scoring	<input type="radio"/>	MudPIT scoring	<input checked="" type="radio"/>	Ions score cut-off	15	Show sub-sets	<input type="checkbox"/>
Show pop-ups	<input checked="" type="radio"/>	Suppress pop-ups	<input type="radio"/>	Sort unassigned	Decreasing Score	Require bold red	<input type="checkbox"/>
Repeat Search							

Choose a value of, for example 15 to improve grouping of homologous proteins and speed up report loading.

ASMS 2004

MATRIX
SCIENCE

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

MudPIT searches - summary

- Memory limitations for searches on 32 bit platforms removed
- Memory / performance for large reports
- Protein scoring
- Select summary report
- Easier to choose report format.

ASMS 2004

**MATRIX
SCIENCE**

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

Sequence tag searches

- Mann, M and Wilm, M, Error-tolerant identification of peptides in sequence databases by peptide sequence tags. (Anal Chem, 66(24) 4390-9 1994).
- Enter observed 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
- 1890.2 tag(1004.1, LSADTG, 1548.5)

ASMS 2004

**MATRIX
SCIENCE**

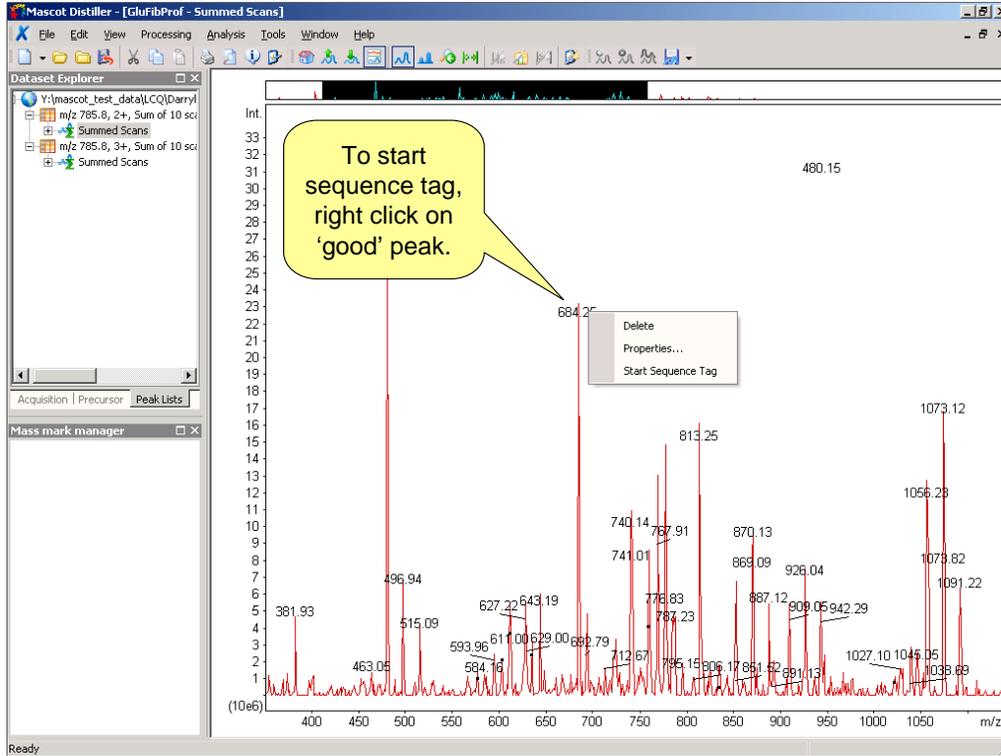
The sequence query, in which one or more peptide molecular masses are combined with sequence, composition and fragment ion data, is potentially the most powerful search of all. The usual source of the sequence information is interpretation of an MS/MS spectrum. While it is very difficult to determine 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.

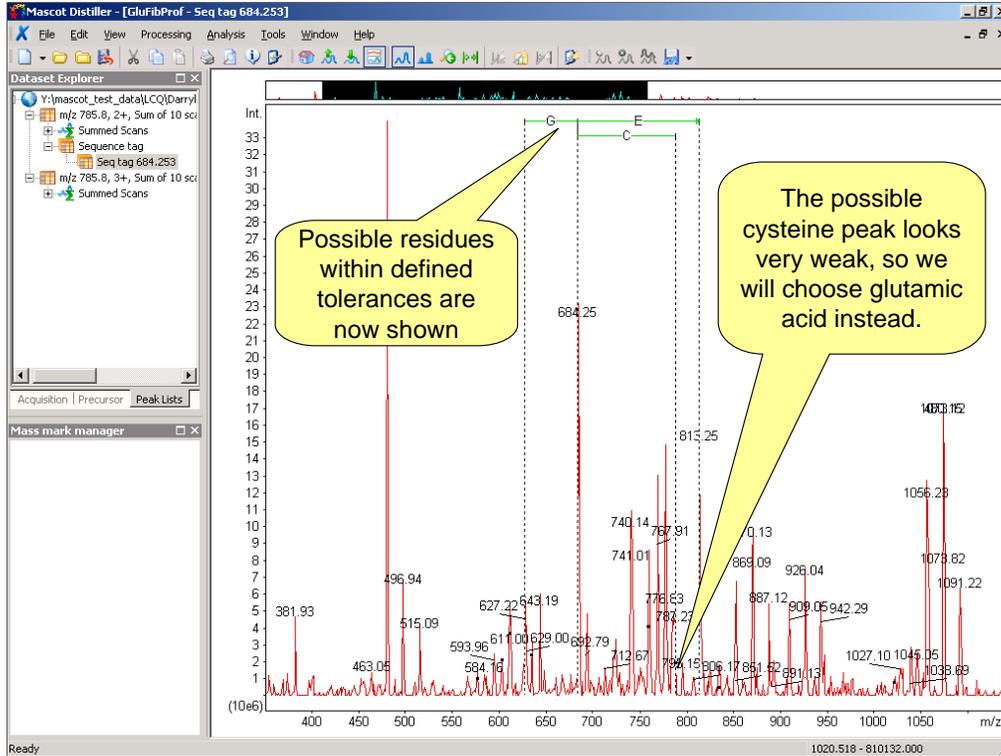
The sequence query mode of Mascot supports both standard and error tolerant sequence tags. It also allows arbitrary combinations of fragment ion mass values, amino acid sequence data and amino acid composition data to be searched.

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)



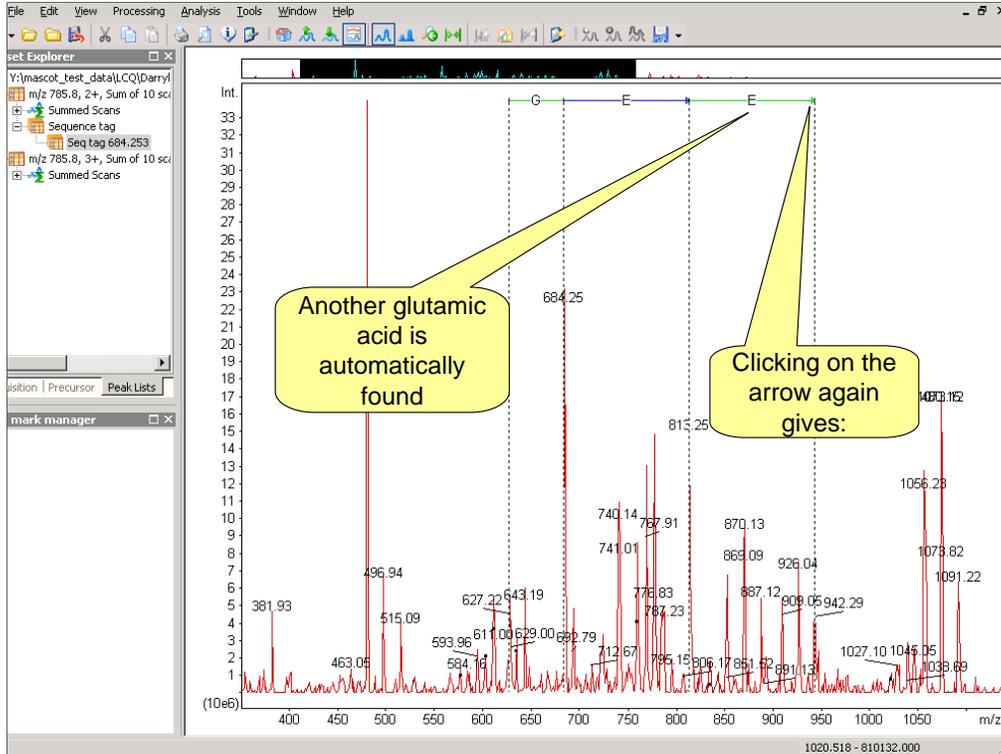
In the next version of Distiller, it is easy to generate sequence tags. Simply process the spectrum, right click on a strong peak, and select "Start Sequence Tag" from the pop-up menu



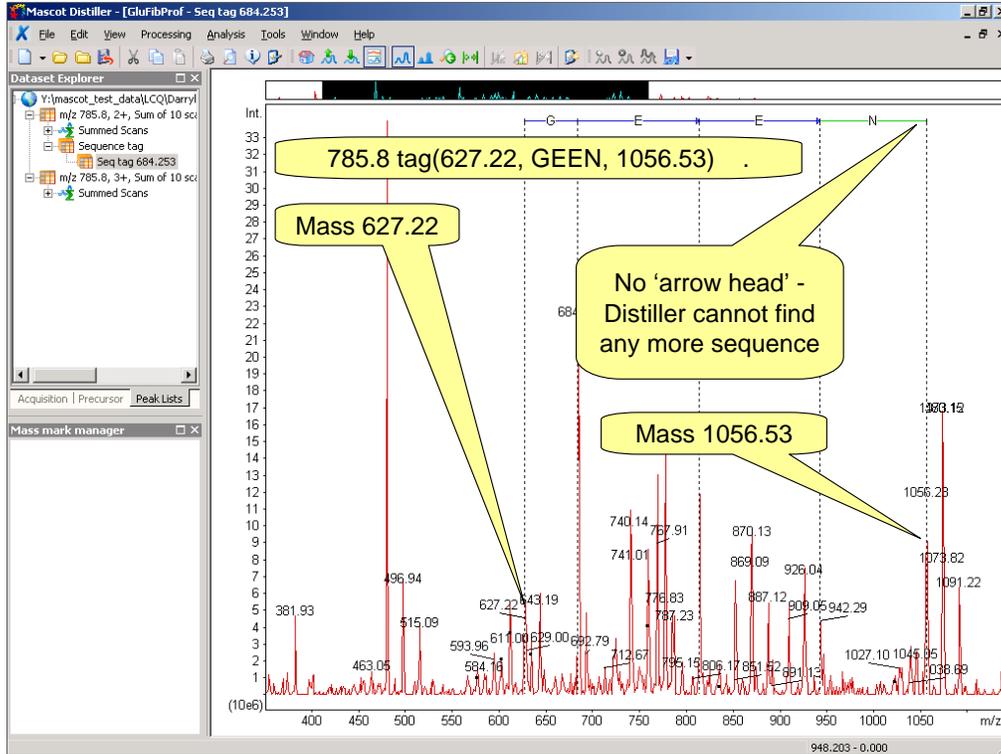
This then displays possible residue matches that fall within the specified tolerance.

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)

The image shows two screenshots of the Matrix Science Mascot web interface. The top screenshot is the 'Sequence Query' form, and the bottom screenshot is the 'Peptide Summary Report'.

Sequence Query Form:

- Protein mass: [] kDa
- Peptide tol. ±: 1.2 Da
- Peptide charge: 2+
- MS/MS tol. ±: 1 Da
- ICAT:
- Monoisotopic: Average:
- Query: 785.8 tag(627.22, GEEN, 1056.53)

Peptide Summary Report:

1. [AAA52445](#) Mass: 7050 Score: 51 Queries matched: 1
 HUNFG802 NID: - Homo sapiens
 Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/> 1	785.80	1569.59	1568.69	0.90	0	54	0.0008	1	QGVNDNEEGFFSAR

ASMS 2004

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.

Error tolerant tag

- Useful where there are unknown modifications

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

Address: http://t41-dmc/mascot/cgi/master_results.pl?file=...%2Fdata%2F20040519%2FF001319.dat&REPTYPE=peptide&sigthreshold=0.05&RE

<input checked="" type="checkbox"/>	58	537.67	1073.33	1073.63	-0.29	1	40	0.23	1	KLDCILLTR
<input checked="" type="checkbox"/>	184	537.67	1610.00	1609.88	0.12	1	25	6.1	1	KTFASLYNAIDVR
<input checked="" type="checkbox"/>	150	710.22	1418.43	1418.73	-0.30	1	21	17	1	LSCEKITHYAVR
<input checked="" type="checkbox"/>	149	472.87	1415.57	1415.74	-0.17	0	10	2e+002	1	LCTVLGVDEAALR + Carbamidomethyl (C)
<input type="checkbox"/>	201	738.27	2011.20	2010.15	0.25	1	6	4e+002	1	WVUKTCCGACGCAACGSAIPEV

Score of 21.

ASMS 2004

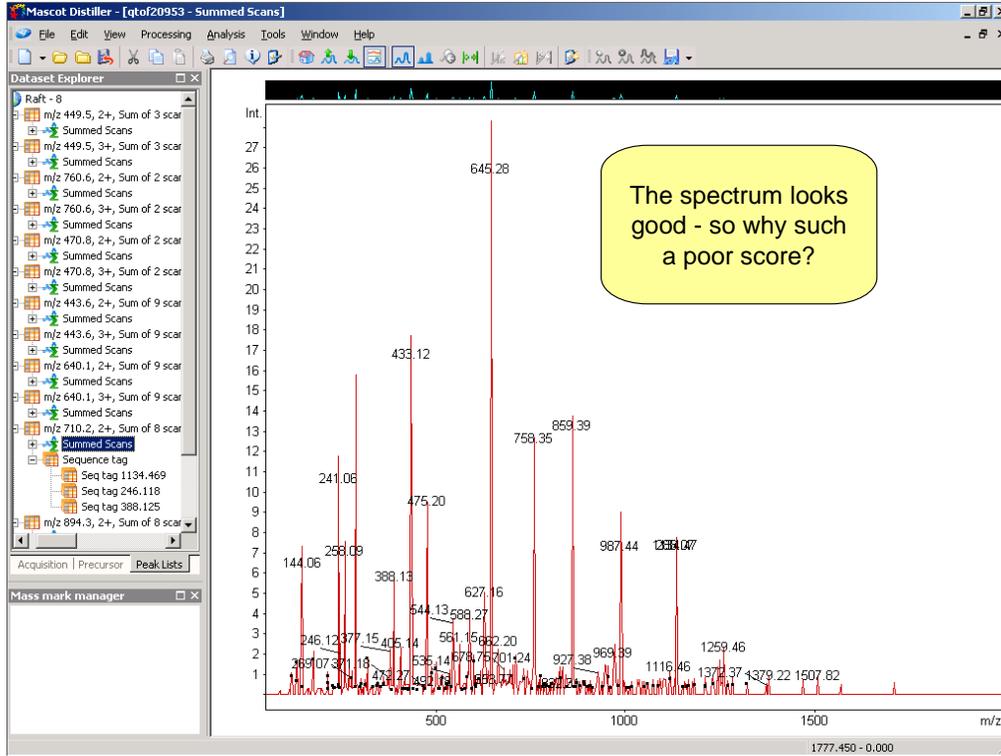
MATRIX SCIENCE

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

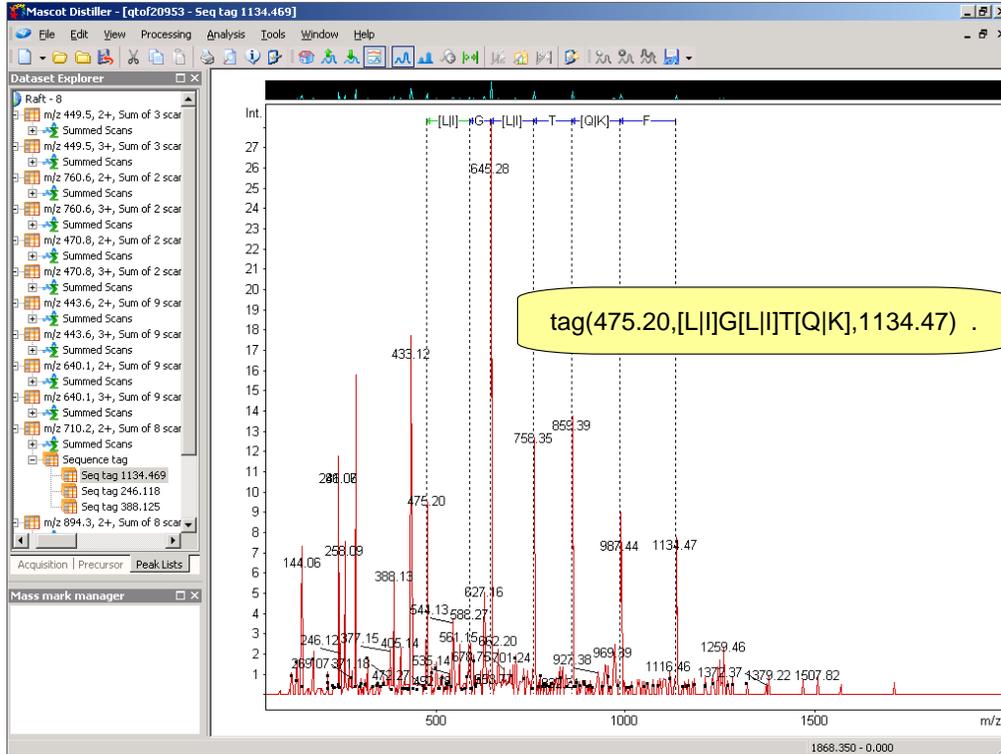
As an example, I'll take this peptide with a score of only 21 from the unassigned list.

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.

Matrix Science - Wot? No Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://t41-dmc/mascot/cgi/master_results.pl?file=.../data/20040520/F001327.dat Go Powermarks

Google Search Web Search Site PageRank Options

MATRIX SCIENCE HOME: MASCOT: HELP Search Go

Mascot > Wot? No Results

Wot? No Results

Sorry ... but this is the page you get to see when your search produces no results.

Here are some possible reasons:

- Sequence Query: None of the specified sequences occur in the database being searched. Sequence qualifiers requires an exact match for homologous sequences. If you suspect a typo, please correct it and re-submit the search.
- All searches: Match worse than expected. This is unlikely to happen. Spurious matches tend to occur because of very narrow tolerances.
- All searches: If a taxonomy filter is selected, it may be too restrictive.

No results from the tag search, so try an etag:
etag(475.20,[L|I]G[L|I]T[Q|K],1134.47)

ASMS 2004 **MATRIX SCIENCE**

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

Peptide Summary Report (.../data/20040520/F001326.dat) - Microsoft Internet Explorer

Address: http://t41-dmc/mascot/cgi/master_results.pl?file=...%2Fdata%2F20040520%2FF001326.dat&REPTYPE=peptide&sig=...

Select All Select None Search Selected Error tolerant Archive Report

1. [PAHUA](#) Mass: 57917 Score: 80 Queries matched: 1
alkaline phosphatase (EC 3.1.3.1) precursor, placental [validated] - human
 Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/> 1	710.20	1418.39	1304.68	113.70	0	80	3.6e-008	1	GNFQTIGLSAAAR

Top scoring peptide matches to query 1
etag(475.20, [L|I]G[L|I]T[Q|K]F,1134.47)
Score greater than 18 indicates identity
Status bar shows all hits for this peptide

Prot	Score	Delta	Hit	Protein	Peptide	Series matched:
PPBN	79.9	99.69	2	PAHUI	ANFQTIGLSAAAR	1
AAHO	79.9	113.70	1	PAHUA	GNFQTIGLSAAAR	1

1:PAHUA Local intranet

This shows a possible modification or substitution

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 - Microsoft Internet Explorer

Address: http://41-dmc/mascot/cgi/peptide_view.pl?file=.../data/20040520/F001326.dat&query=1&hit=2&index=PAHUA&px=1

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 $M_r(\text{calc})$: 1304.68
 Unsuspected modification: 113.70 Da, located in the region N-term to N2
 Ions Score: 80 Expect: 3.6e-008

Peptide view shows location of unsuspected modification

#	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 -

UniMod: Basic Data View - Microsoft Internet Explorer

Address: http://www.unimod.org/cgi/unimod.cgi?records_per_page=25&columns_to_view=full_name&columns_to_view=code_name&columns_to_view=mono_...

phosphorylation to amine thiol	ser_thr_DAET	87.050655	87.1866	H(9) C(4) N O(-1) S	⌵
thioacylation of primary amines (N-term and Lys)	DSP	87.998285	88.1283	H(4) C(3) O S	⌵
C13 label (Phosphotyrosine)	13C9_Phospho_Tyr	88.996524	88.9138	H C(-9) C13(9) O(3) P	⌵
N-ethyl iodoacetamide-d5	NEIAA-d5	90.084148	90.1353	H(2) H2(5) C(4) N O	⌵
Acrolein addition +94	Acrolein94	94.041865	94.1112	H(6) C(6) O	⌵
N-isopropylcarboxamidomethyl	NIPCAM	99.068414	99.1311	H(9) C(5) N O	⌵
Succinic anhydride labeling reagent light form (N-term & K)	Suc_anh_light	100.016044	100.0728	H(4) C(4) O(3)	⌵
labeling reagent light form (N-term & K)	benzoyl	104.026215	104.1061	H(4) C(7) O	⌵
Succinic anhydride labeling reagent, heavy form (+4amu, 4C13), N-term & Suc_anh+4C13 K	Suc_anh+4C13	104.029463	104.0434	H(4) C13(4) O(3)	⌵
Succinic anhydride labeling reagent, heavy form (+4amu, 4H2), N-term & K	Suc_anh+4H2	104.041151	104.0974	H2(4) C(4) O(3)	⌵
S-pyridylethylation	S-pyridylethyl	105.057849	105.1372	H(7) C(7) N	⌵
Acrolein addition +112	Acrolein112	112.052430	112.1265	H(8) C(6) O(2)	⌵
ubiquitinylation residue	GlyGly	114.042927	114.1026	H(6) C(4) N(2) O(2)	⌵
Pyridyl	Pyridyl	119.037114	119.1207	H(5) C(7) N O	⌵
N-ethylmaleimide on cysteines	NEM	125.047679	125.1253	H(7) C(6) N O(2)	⌵
Iodination	Iodination	125.896648	125.8965	H(-1) I	⌵
N-Succinimidyl-3-morpholine acetate	SMA	127.063329	127.1412	H(9) C(6) N O(2)	⌵
Quaternary amine labeling reagent light form (N-term & K)	Quat_0	127.099714	127.1842	H(13) C(7) N O	⌵
Quaternary amine labeling reagent heavy (+3amu) form, N-term & K	Quat_3	130.118544	130.2027	H(10) H2(3) C(7) N O	⌵
Hydroxyphenylglyoxal arginine	Arg1HPG	132.021129	132.1162	H(4) C(6) O(2)	⌵
Quaternary amine labeling reagent heavy form (+6amu), N-term & K	Quat_6	133.137375	133.2212	H(7) H2(6) C(7) N O	⌵
Quaternary amine labeling reagent heavy form (+9amu), N-term & K	Quat_9	136.156205	136.2397	H(4) H2(9) C(7) N O	⌵

Total Records Returned: 181 Previous 25 Next 25 Viewing Records: 76 - 100

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.

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

Address: http://t41-dmc/mascot/cgi/master_results.pl?file=.../data/20040520/F001320.dat

Show pop-ups Suppress pop-ups Sort unassigned Decreasing Score Require bold red

Select All Select None Search Selected Error tolerant Archive Report

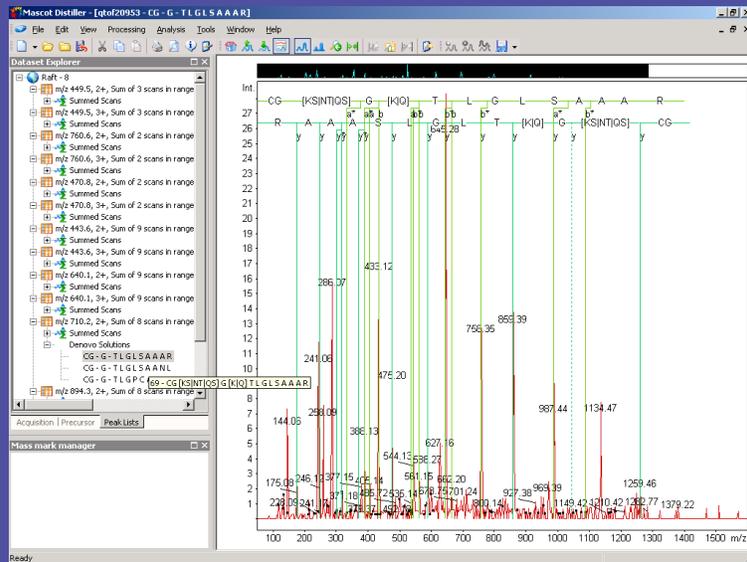
1. [AAB64400](#) Mass: 56428 Score: 1087 Queries matched: 28
 SECRETED ALKALINE PHOSPHATASE.- Cloning vector pSEAP2-Basic.
 Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/> 24	460.17	918.32	918.51	-0.19	0	49	0.0012	1	SVGVVITTR
<input checked="" type="checkbox"/> 25	462.68	923.35	923.51	-0.16	0	32	0.049	1	FPYVALSK
<input checked="" type="checkbox"/> 47	517.18	1032.34	1032.56	-0.22	0	78	1.4e-006	1	SSIFGLAPGK + Carbamidomethyl (N-term) [+57.02]
<input checked="" type="checkbox"/> 55	532.18	1062.35	1062.57	-0.22	0	63	5.9e-005	1	GSSIFGLAPGK + G->S [+30.01]
<input checked="" type="checkbox"/> 62	545.68	1089.35	1089.58	-0.23	0	50	0.0011	1	RGSSIFGLAPGK + R->G [-99.08]
<input checked="" type="checkbox"/> 78	567.66	1133.30	1133.55	-0.25	0	49	0.0014	1	GHEVISVMNR + Oxidation (M)
<input checked="" type="checkbox"/> 94	593.18	1184.35	1184.60	-0.25	0	42	0.007	1	RGFFLFVEGGR + R->G [-99.08]
<input checked="" type="checkbox"/> 121	653.21	1304.41	1304.68	-0.28	0	89	1.8e-007	1	HFQIIGLSAAR + Carbamidomethyl (N-term) [+57.02]
<input checked="" type="checkbox"/> 139	681.82	1361.63	1361.55	0.07	0	54	0.00053	1	YPDDYSQGGTR + benzoyle (N-term) [+104.03]
<input checked="" type="checkbox"/> 150	710.22	1418.43	1418.73	-0.29	0	87	2.8e-007	1	KGHFQIIGLSAAR + K->N [-14.05]
<input checked="" type="checkbox"/> 154	726.18	1450.35	1450.65	-0.30	0	69	1.9e-005	1	HWYSADAVPASAR
<input checked="" type="checkbox"/> 165	754.69	1518.35	1518.65	-0.30	0	69	1.9e-005	1	RHWYSADAVPASAR + R->G [-99.08]
<input checked="" type="checkbox"/> 175	526.15	820.73	820.73	0.00	0	32	0.049	1	ALTEIMFDDAIER + F->V [-48.00]
<input checked="" type="checkbox"/> 190	820.73	820.73	820.73	0.00	0	32	0.049	1	ALTEIMFDDAIER + Oxidation (M)
<input checked="" type="checkbox"/> 208	864.29	864.29	864.29	0.00	0	32	0.049	1	AYTVLLYGNGPGYVLK
<input checked="" type="checkbox"/> 214	586.50	586.50	586.50	0.00	0	32	0.049	1	IIPVEEENPDFWNR
<input checked="" type="checkbox"/> 215	879.24	879.24	879.24	0.00	0	32	0.049	1	IIPVEEENPDFWNR
<input checked="" type="checkbox"/> 218	593.48	593.48	593.48	0.00	0	32	0.049	1	HVPDSGATATAYLCGVK + Carbamidomethyl (C); V->N [+3
<input checked="" type="checkbox"/> 254	975.81	975.81	975.81	0.00	0	32	0.049	1	NLIIFLGDGMGVSTVTAAR + Oxidation (M)
<input checked="" type="checkbox"/> 255	651.16	651.16	651.16	0.00	0	32	0.049	1	DGARPDVTESESGSPEYR

Error tolerant search gives a significant match to the same sequence with an additional n terminus asparagine.

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

Denovo in Distiller 2.0



ASMS 2004

MATRIX
SCIENCE

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

- Process raw data files using Mascot Distiller
 - Retention times and scan numbers transferred to Mascot results
 - More robust peak detection
 - Output of peak lists not automated on some systems
- Runs as a 'service'
- Trouble shooting tips.

ASMS 2004

**MATRIX
SCIENCE**

Mascot Daemon had some major changes with version 2.0

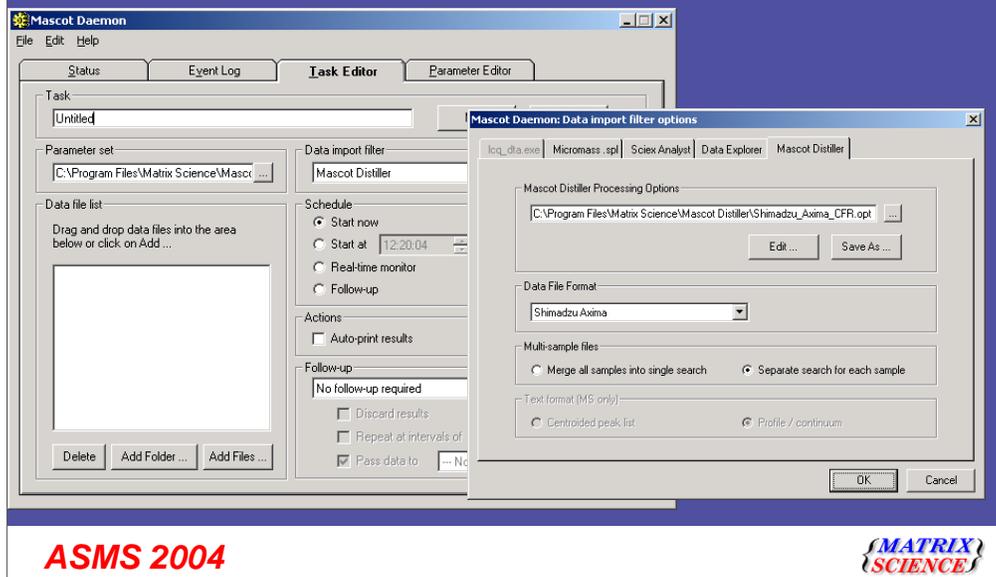
- it is now possible to process raw data files using the Distiller libraries. This means that all the common raw data formats are now supported. The advantages are that retention times and scan numbers along with other information in the raw data files are transferred in the Mascot results. This information, 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 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.

Daemon - distiller options



ASMS 2004



There is not much to say about the Distiller options. There is an additional license cost for using Distiller from Daemon, and once the additional 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.

Daemon - runs as a service

- Graphical User Interface (GUI)
 - configure Daemon, edit tasks, view search results
- Service
 - executes the tasks, in the background even if the GUI is not running
- No longer requires user to be logged in at the workstation
- If connection is lost between Daemon and the Mascot server, results are found when Daemon is re-started.

ASMS 2004



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 which 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.

Daemon - trouble shooting tips

- On-line help now available
- “File not found” error: To access files on shared network drives, you have to change the service to log in as a user with appropriate network privileges. See on-line help, In depth, Mascot Daemon Service

ASMS 2004



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

Iterating through B,Z,X

- X - iterate through all amino acids
- B - check asparagine and aspartic acid
- Z - check glutamine and glutamic acid

X => V

Rank	Score	Peptide Sequence
39	731.97	XGSIGAASMEFCFDVFKELK + Oxidation (M)
40	1097.46	XGSIGAASMEFCFDVFKELK
41	745.64	XGSIGAASMEFCFDVFKELK + Carbamidomethyl (C)
43	761.01	DILNQITKPNQVYVSLASR
44	762.01	LYAEERYPIIPEYLQCVK + Carbamidomethyl (C)

ASMS 2004

MATRIX
SCIENCE

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

Miscellaneous

- More rigorous checking of fasta files
 - Duplicate accessions reported
 - Missing taxonomy files an error instead of warning
- Use taxonomy to determine the correct genetic code
- Support for high energy side chain cleavage ions (d, v, w).

ASMS 2004

**MATRIX
SCIENCE**

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.