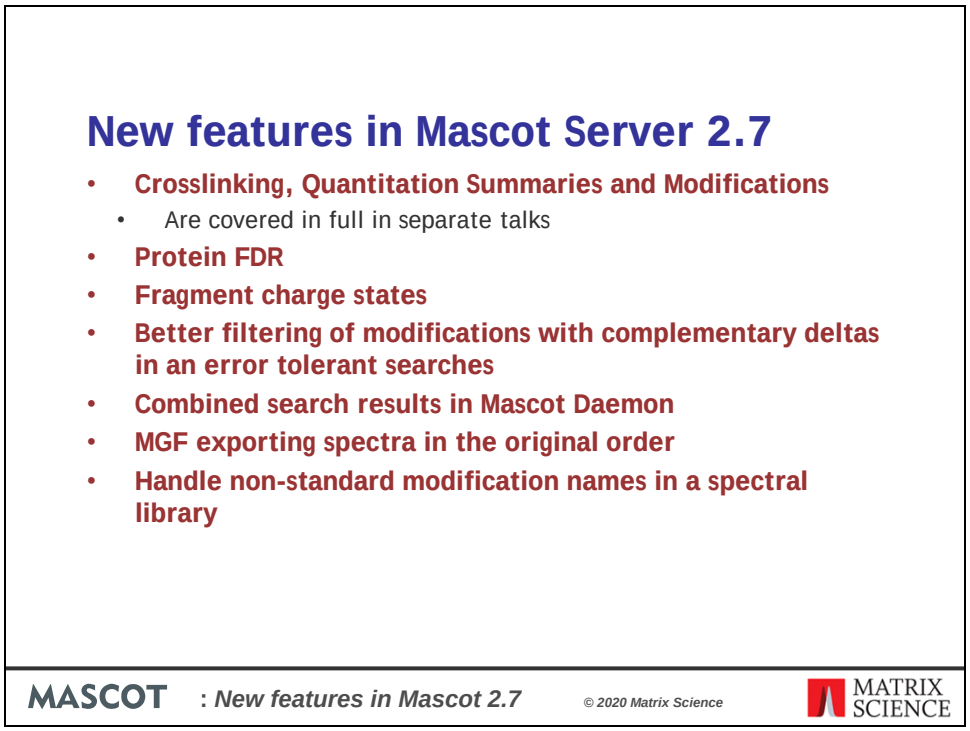
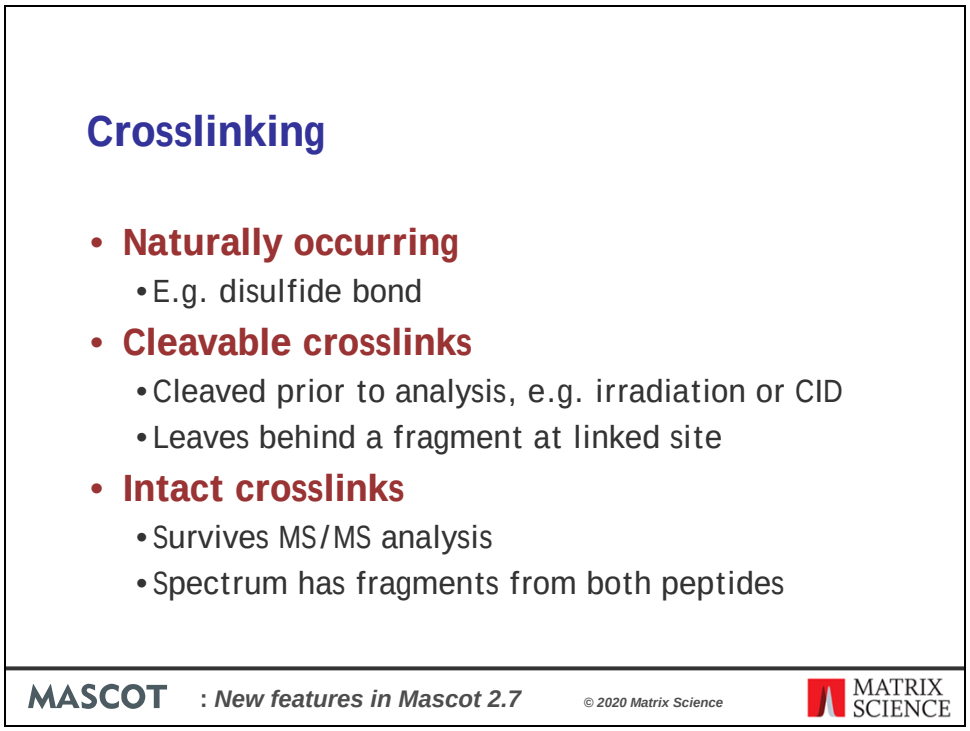
New features in Mascot Server 2.7

Richard Jacob

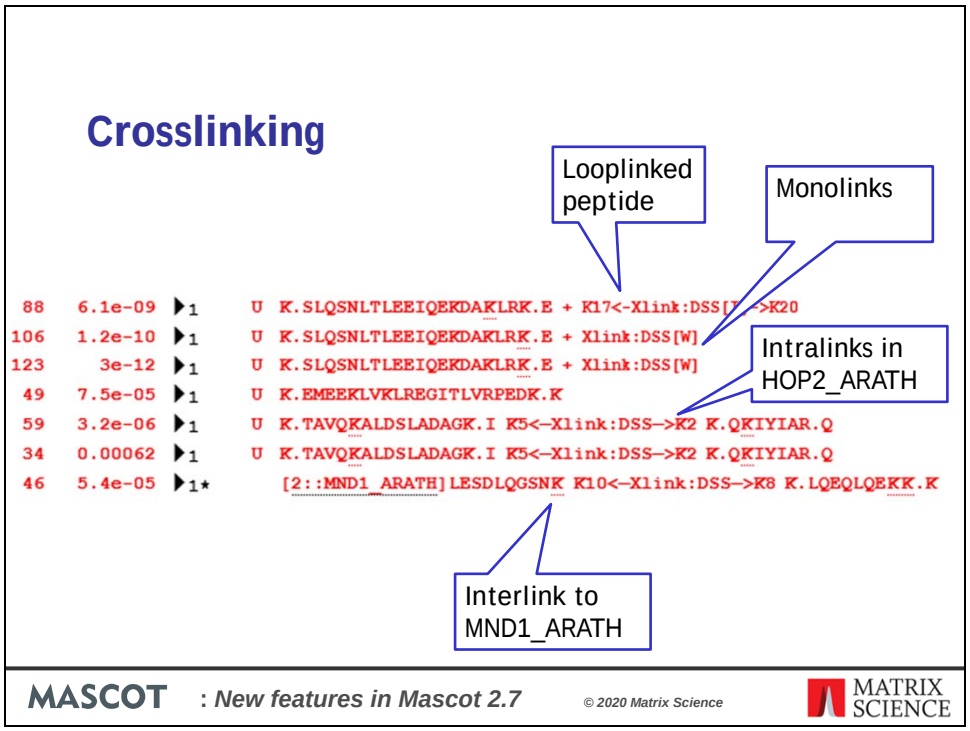


The big new feature support for crosslinking. This and the Quantitation Summaries features and improvements to how we search for Modifications are covered in separate talks.

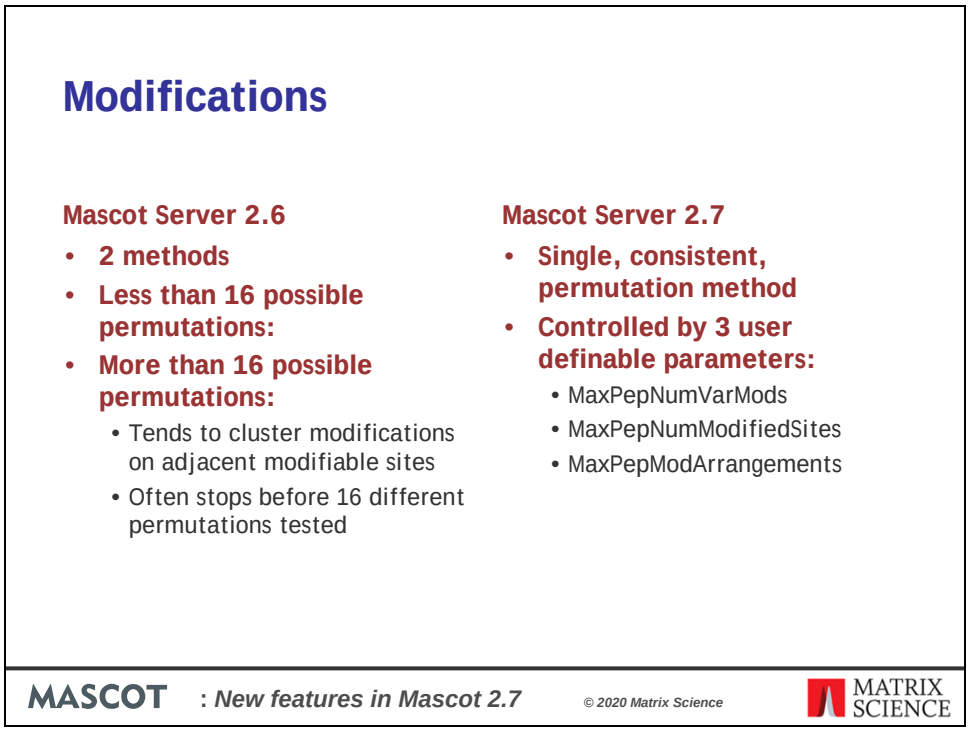
In this talk I will cover the following topics:



We've added full support for a comprehensive range of crosslinking scenarios. This includes editors for the crosslinking agents and methods, fragmentation patterns in peptide_view and exports to xiVIEW for further visualization.



Mascot can report all the different types of links from looped, Inter and Intra links and incomplete mono links.



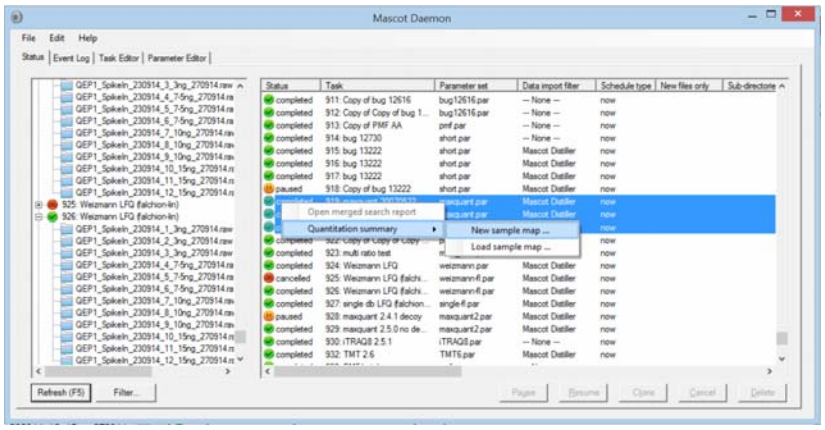
Mascot Server 2.7 has better site localisation by introducing separate caps on the number of distinct modifications in a PSM, the number of modified sites in a PSM, and the total number of arrangements to be tested.

Mascot Server 2.6 worked well but had limitations. If a peptide has less than 16 possible variable modification permutations, then there are no issues as all possibilities are tested. It's in the cases where the peptide has more than 16 possible permutations that issues can arise. The sliding window method tends to cluster modifications on adjacent modifiable sites, and it will often stop before 16 different permutations have been tested.

In Mascot 2.7 we've taken a different approach. We use a single, consistent permutation method – there's no switching between different methods. The new permutation iterator samples arrangements using a uniformly random scheme. The operation of this is controlled by 3 user definable settings.

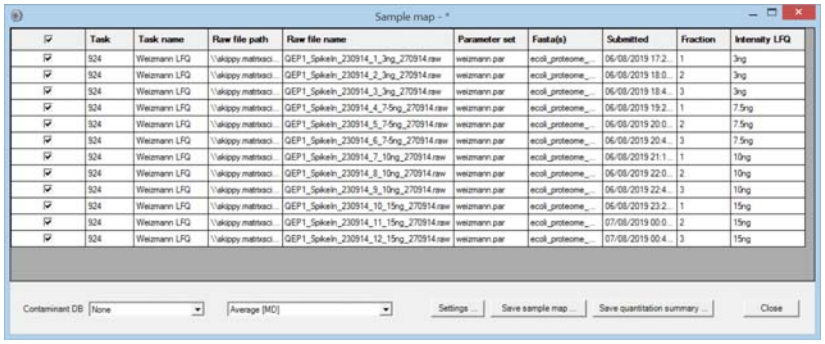
By adjusting the parameters you can choose to reduce the search depth and speed up searches of lightly modified samples or increase the search depth where site localization matters or for highly modified samples.

Quantitation Summaries

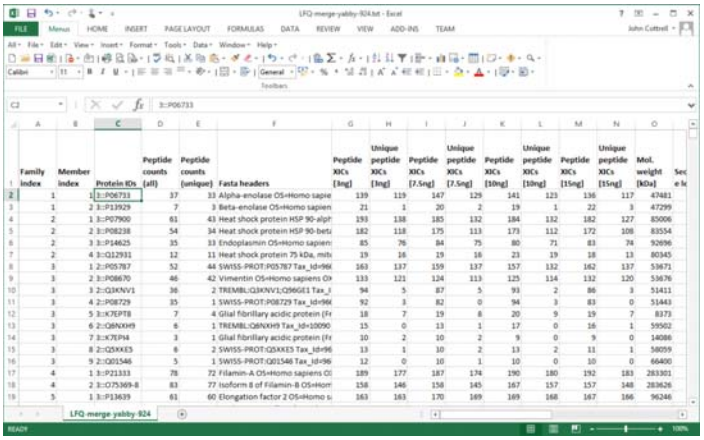


The Quantitation Summaries feature in Mascot Daemon allows you to select one or more tasks and create a sample map.

Sample map



The Sample Map is used to annotate the list of result files with recognisable sample identifiers

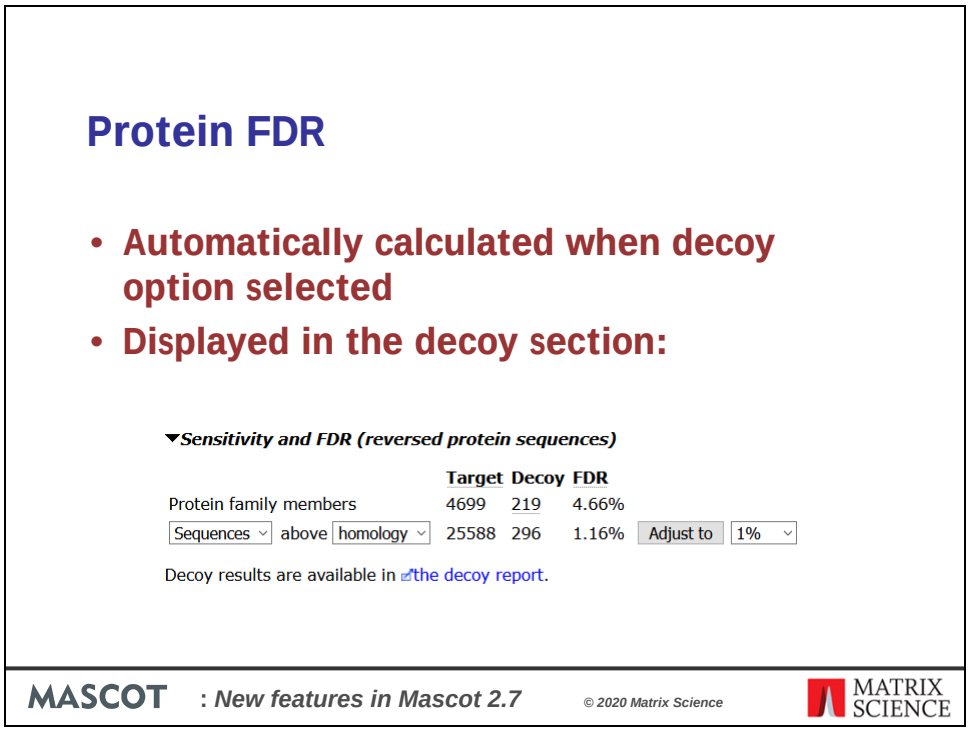


Family index	Member index	Protein IDs	Peptide counts (all)	Peptide counts (unique)	Fasta headers	Peptide XICs [1ng]	Unique peptide XICs [1ng]	Peptide XICs [7.5ng]	Unique peptide XICs [7.5ng]	Peptide XICs [10ng]	Unique peptide XICs [10ng]	Peptide XICs [15ng]	Unique peptide XICs [15ng]	Mol. weight [kDa]
1	1	1::P06733	37	33	Alpha-enolase OS=Homo sapie	139	119	147	129	141	123	136	117	47481
1	2	1::P13929	7	3	Beta-enolase OS=Homo sapien	21	1	20	2	19	1	22	3	47299
2	1	3::P07900	61	43	Heat shock protein HSP 90-alpl	193	138	185	132	184	132	182	127	85006
2	2	3::P08238	54	34	Heat shock protein HSP 90-beta	182	118	175	113	173	112	172	108	83554
2	3	3::P14625	35	33	Endoplasmin OS=Homo sapien	85	76	84	75	80	71	83	74	92696
2	4	3::Q12931	12	11	Heat shock protein 75 kDa, mit	19	16	19	16	23	19	18	13	80345
3	1	2::P05787	52	44	SWISS-PROT:P05787 Tax_Id=96	163	137	159	137	157	132	162	137	53671
3	2	3::P08670	46	42	Vimentin OS=Homo sapiens O	133	121	124	113	125	114	132	120	53676
3	3	2::Q3KNV1	36	2	TREMBL:Q3KNV1:Q96GE1 Tax_I	94	5	87	5	93	2	86	3	51411
3	4	2::P08729	35	1	SWISS-PROT:P08729 Tax_Id=96	92	3	82	0	94	3	83	0	51443
3	5	3::K7EPT8	7	4	Glial fibrillary acidic protein (Fr	18	7	19	8	20	9	19	7	8373
3	6	2::Q6NXH9	6	1	TREMBL:Q6NXH9 Tax_Id=10090	15	0	13	1	17	0	16	1	59502
3	7	3::K7EPN4	3	1	Glial fibrillary acidic protein (Fr	10	2	10	2	9	0	9	0	14088
3	8	2::Q5XKE5	8	2	SWISS-PROT:Q5XKE5 Tax_Id=96	13	1	10	2	13	2	11	1	58059
3	9	2::Q01546	5	1	SWISS-PROT:Q01546 Tax_Id=96	12	0	10	1	10	0	10	0	66400
4	1	3::P21333	78	72	Filamin-A OS=Homo sapiens O	189	177	187	174	190	180	192	183	283301
4	2	3::O75369-8	83	77	Isoform 8 of Filamin-B OS=Hom	158	146	158	145	167	157	157	148	283626
5	1	3::P13639	61	60	Elongation factor 2 OS=Homo s	163	163	170	169	169	168	167	166	96246

MASCOT : *New features in Mascot 2.7* © 2020 Matrix Science MATRIX SCIENCE

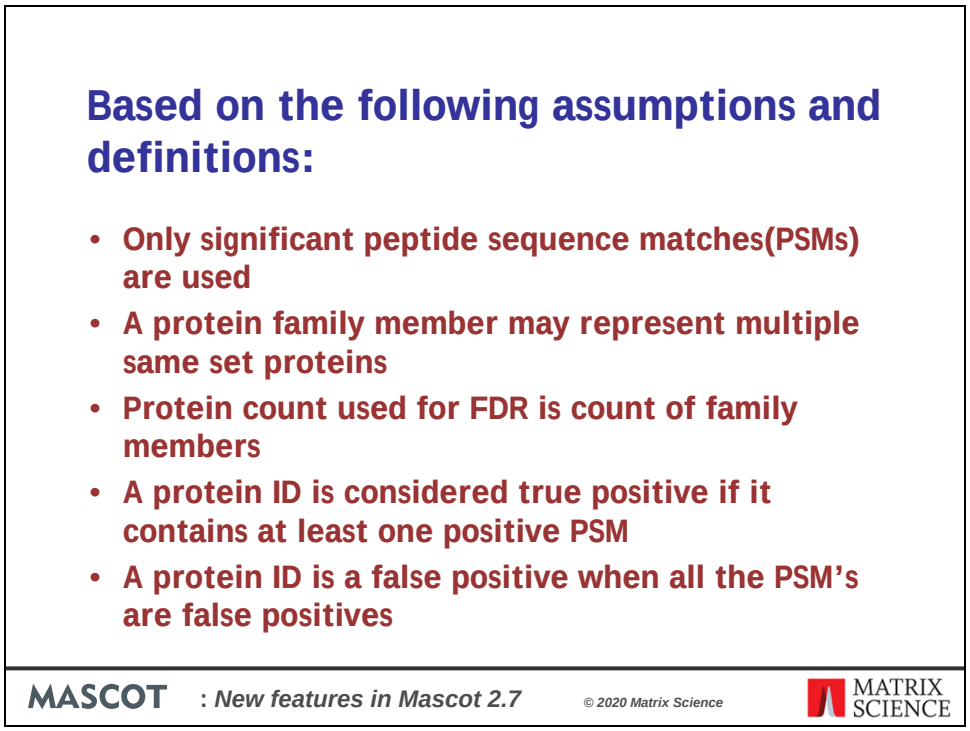
And Daemon creates and saves the Quantitation Summary, a tab delimited text file which is the input file for statistical analysis. This can then be used with anything from Excel to Perseus or a programming environment like R.

Go and see the presentation if you would like to see examples of what you can do with the output. If you would like to know more about crosslinking or the changes to modification detection check out those respective presentations as well.



The new Protein FDR value is automatically calculated when the decoy search option is selected.

It is displayed above the peptide/PSM decoy results.



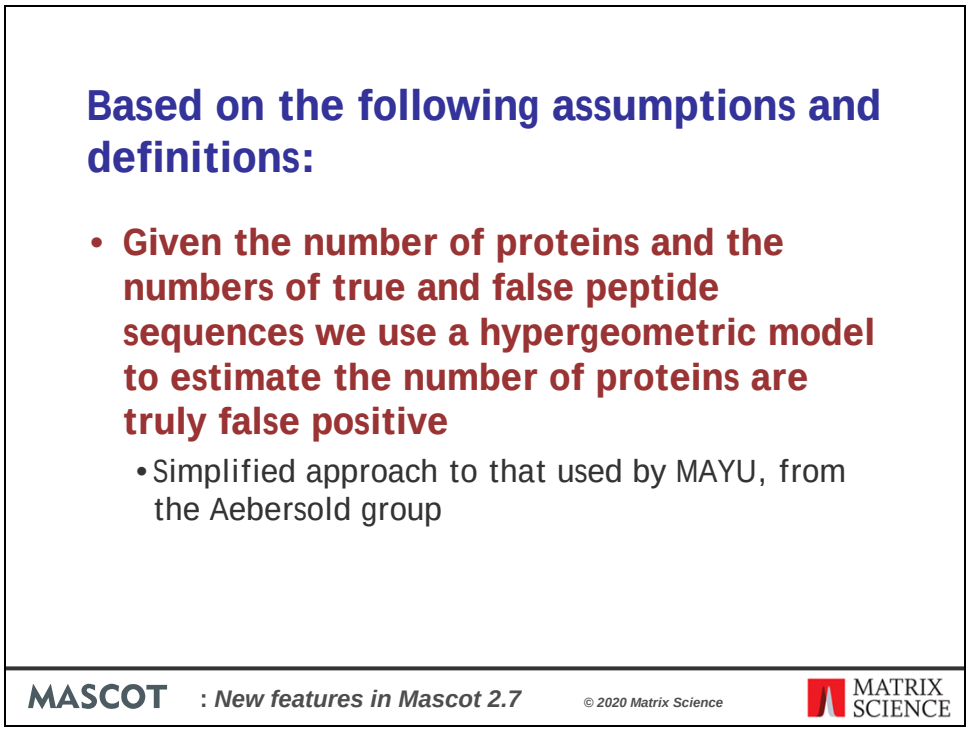
Protein FDR is based on the following assumptions and definitions:

By default the protein family report only shows peptide sequence matches (PSMs) with significant scores and are used for the protein family assignment.

A protein family member may represent multiple same set proteins. Only members of all the protein families in the report, those that contain a unique peptide, are counted.

Protein count used for FDR is count of family members. That is, if the report contains 2 families, one with 4 members and the other with a single member, this counts as a total of 5 proteins. Same-set, sub-set and intersection proteins are not counted.

While a protein ID is a false positive when all the PSM's are false positives. Just one true PMS would make the protein identification a true positive. This is very important.

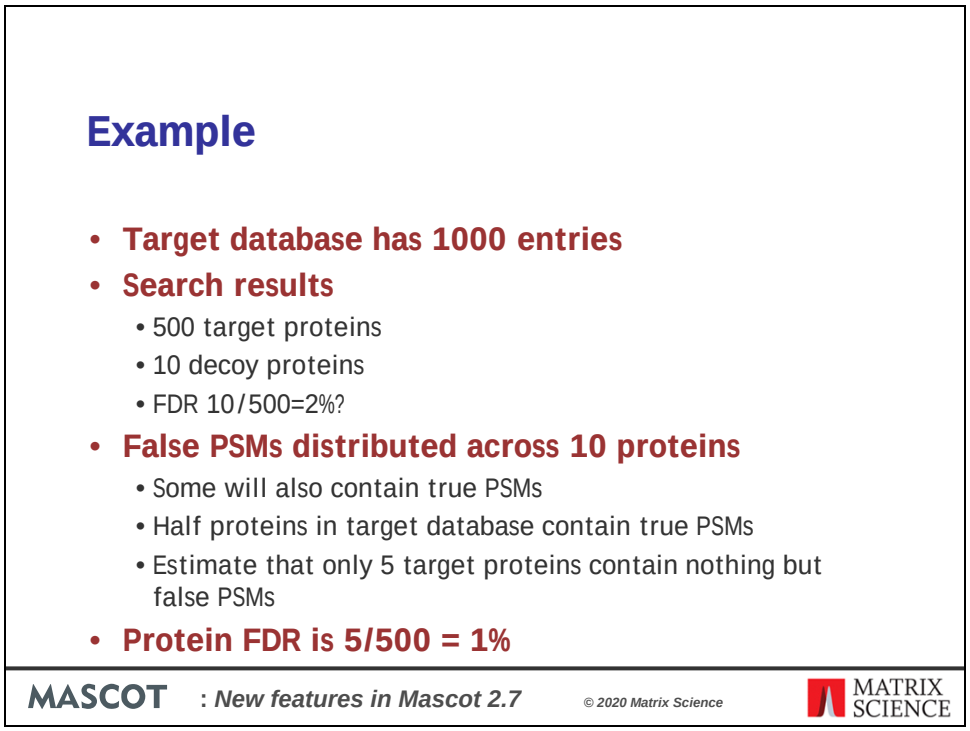


Given the number of proteins and the numbers of true and false peptide sequences we use a hypergeometric model to estimate the number of proteins that are truly false positive

The algorithm is a simplified approach to that used by MAYU, from the Aebersold group

https://www.mcponline.org/content/8/11/2405https://www.mcponline.org/content/8/11/2405

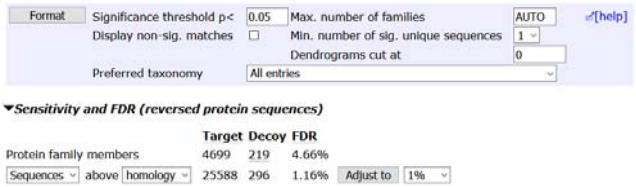
The main differences are that we do not make a separate estimate of the FDR for one-hit wonders and we do not partition the database by protein size. We use a simpler estimate for the number of false proteins in the target database, based on the assumption that the number of decoy proteins never reaches a significant proportion of the database size.



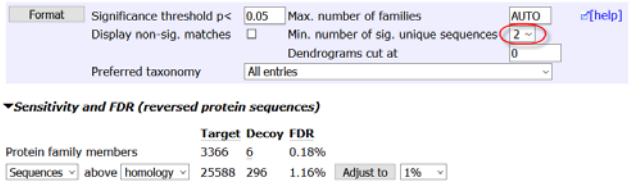
Imagine the and the search results show 500 target proteins and 10 decoy proteins. Does this mean protein FDR is 10/500 = 2%? No, it does not. We can assume the false PSMs in the target are distributed across 10 proteins, but some of these will also contain true PSMs, so should not be counted as false. Since half the proteins in the target database contain true PSMs, a reasonable estimate would be that only 5 target proteins containing nothing but false PSMs, so that the protein FDR is 5/500 = 1%.

Adjusting the Protein FDR

Default protein FDR

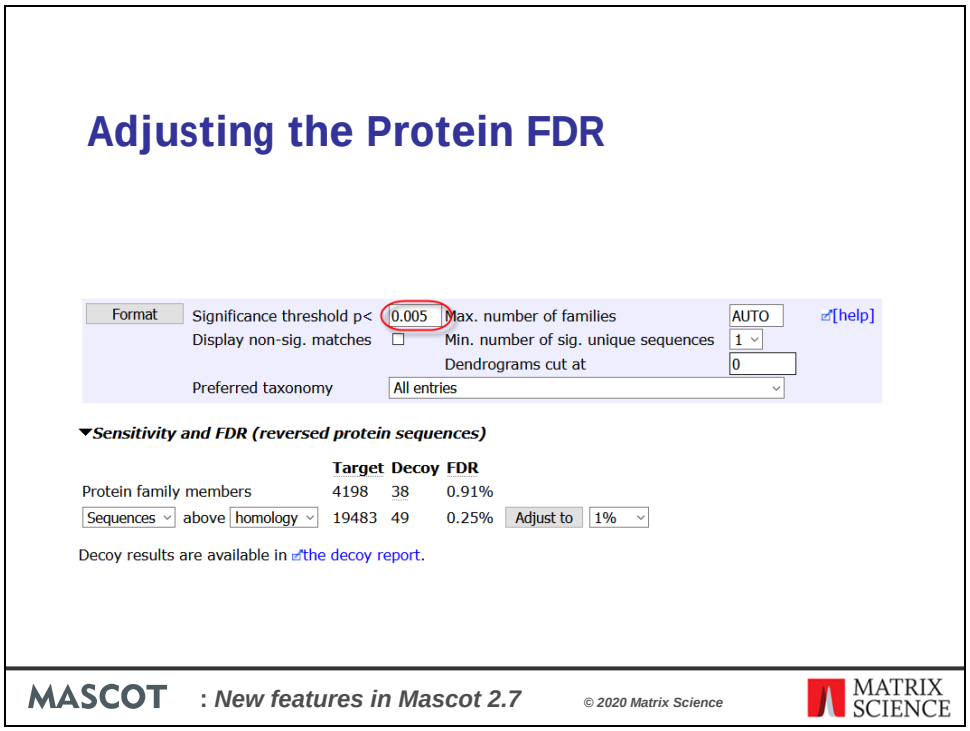


Set Min. number sig. unique peptide sequences to 2



The default significance threshold for a Mascot search is usually 0.05 and this will often give a peptide FDR in the region of 5%. In this dataset the protein FDR is ~4.5%. If we want to lower the Protein FDR we can try adjusting the peptide FDR to 1%. In this case the default peptide FDR is already close to 1% so we can try adjusting the report other ways.

We can adjust the Minimum number of significant unique sequences. This has quite a strong affect on the Protein FDR. If we change it to 2 and eliminate the “one hit wonders” the protein FDR drops to 0.18%.

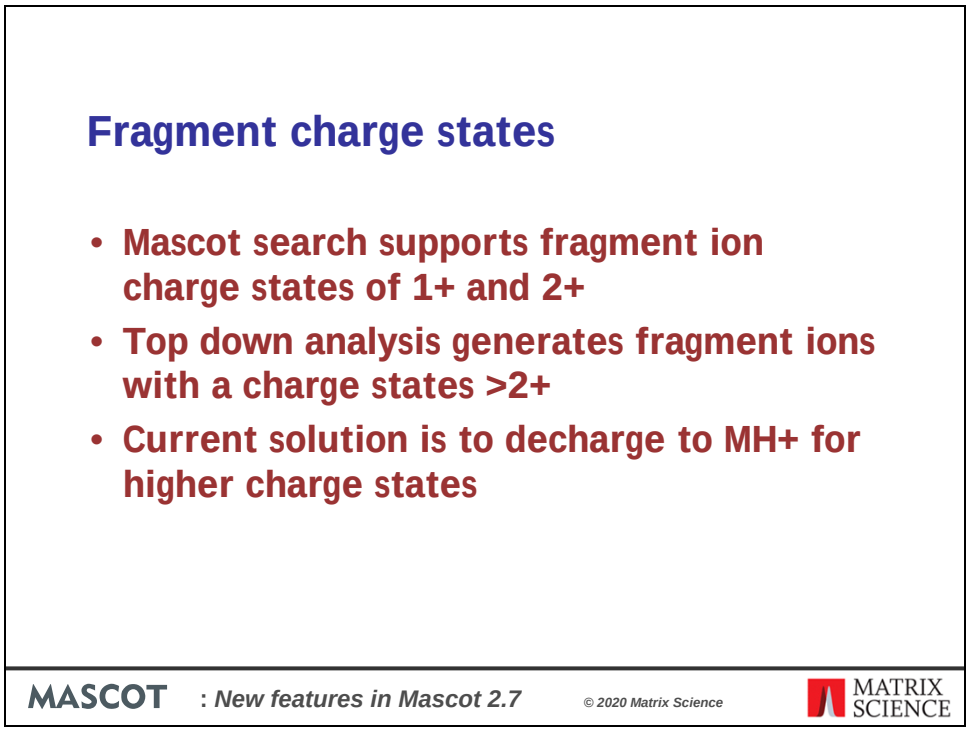


Alternatively if we can adjust the Significance threshold for the results. I took a guess and reduced it by a factor of 10 from the default values of 0.05 to 0.005 and clicked the format button. The resulting Protein FDR is approximately 1%.

The current HUPO guidelines Interpretation Guidelines for large-scale results recommend adjusting the settings to lower than 1% protein-level global FDR so after formatting this search result would meet those guidelines.

When interpreting the results a protein FDR of 1% only tells us that 1% of the proteins listed are wholly false. This doesn't mean the other 99% are "correct". In particular, where there are same-set proteins, we cannot say which one is "correct".

This is because database redundancy causes protein inference ambiguity and we can account for the PSM evidence using several sets of proteins. It is important to remember that a protein accession number in the summary report does not mean "this is the correct protein", it means "the correct protein is likely to be very similar to one of the set of proteins represented by this family member".



Mascot Server automatically searches for fragment ions as singly or doubly charged ions but not higher charge states.

Middle and top down analysis of large peptides and proteins can generate fragment ions with charge states of 3+ or higher.

In the past if you have fragment ions with higher charge states, you needed to deconvolute them to MH+.

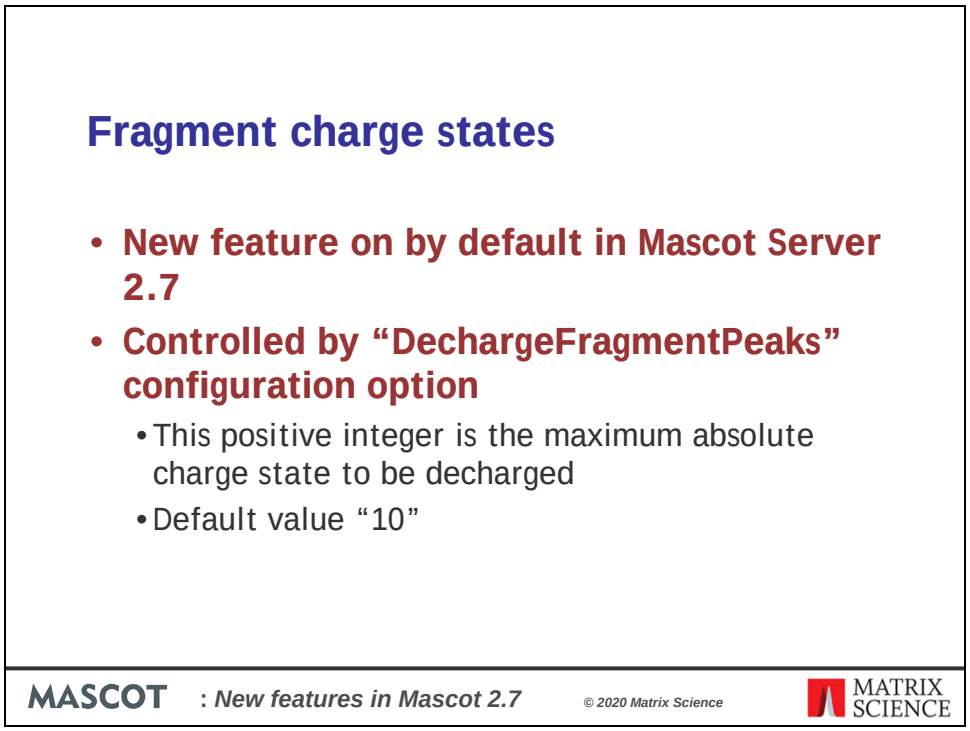
This is supported by Mascot Distiller and is still the preferred method.

Fragment charge states

- **MGF format now supports fragment ion charge states in the third column**

1052.529	265.0801	7	236.4717	7360.649	0.016802	0.710974	232
1068.542	180.6315	7	87.04799	7472.745	0.01753	0.707436	480
1072.044	126.6786	2	87.0784	2142.073	0.019016	0.728941	360
1078.925	274.3603	5	63.94548	5389.586	0.018413	0.74134	238
1082.002	620.7084	9	1.837065	9728.949	0.064662	0.511568	298
1095.742	182.8241	5	117.3447	5473.675	0.018439	0.691048	446
1104.344	446.9772	5	3610.589	5516.682	0.017342	0.954766	116
1111.563	85.08009	1	25.13187	1110.556	0.024389	0.785191	475
1132.535	539.4132	2	1581.047	2263.056	0.018836	0.978732	22
1141.594	91.67654	1	10.8601	1140.587	0.019478	0.885844	338
1141.96	364.0111	5	228.1712	5704.762	0.018316	0.941458	190
1181.524	286.7844	1	340.8262	1180.517	0.020134	0.975948	58
1264.863	256.5222	4	997.5933	5055.423	0.021817	0.950108	378

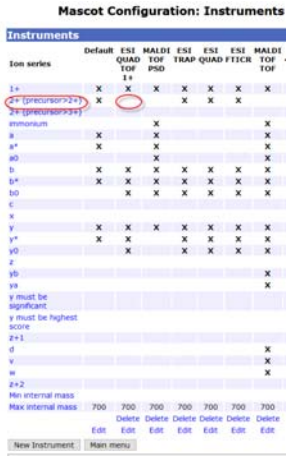
For software that does not decharge the peaks we added support for specifying the charge state in the third column of the peak list. So you have mass, intensity or area and now charge state. Any data in the later columns will be ignored.



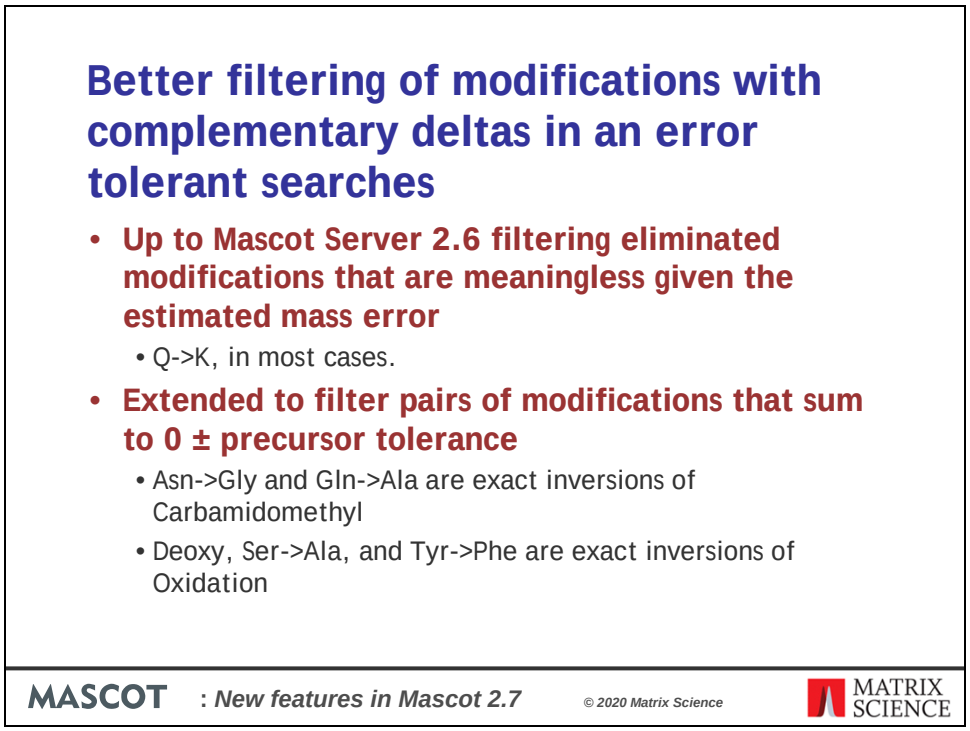
The new feature is on by default and controlled by the "DechargeFragmentPeaks" configuration option.

DechargeFragmentPeaks is the maximum absolute charge state to be decharged and is set to 10 by default.

Create a new instrument that only looks for 1+ fragments



As we expect all the fragment ions in the peak list to be MH+ there is no need to search for 2+ ions. To improve the specificity of the search we recommend creating an instrument configuration that removes support for 2+ (precursor >2+) ion series.



The next feature is “Better filtering of modifications with complementary deltas in an error tolerant searches”.

Error tolerant search results automatically remove matches where the delta of the modifications is less than smaller of the precursor or fragment ion tolerances. For example Q->K, in most cases.

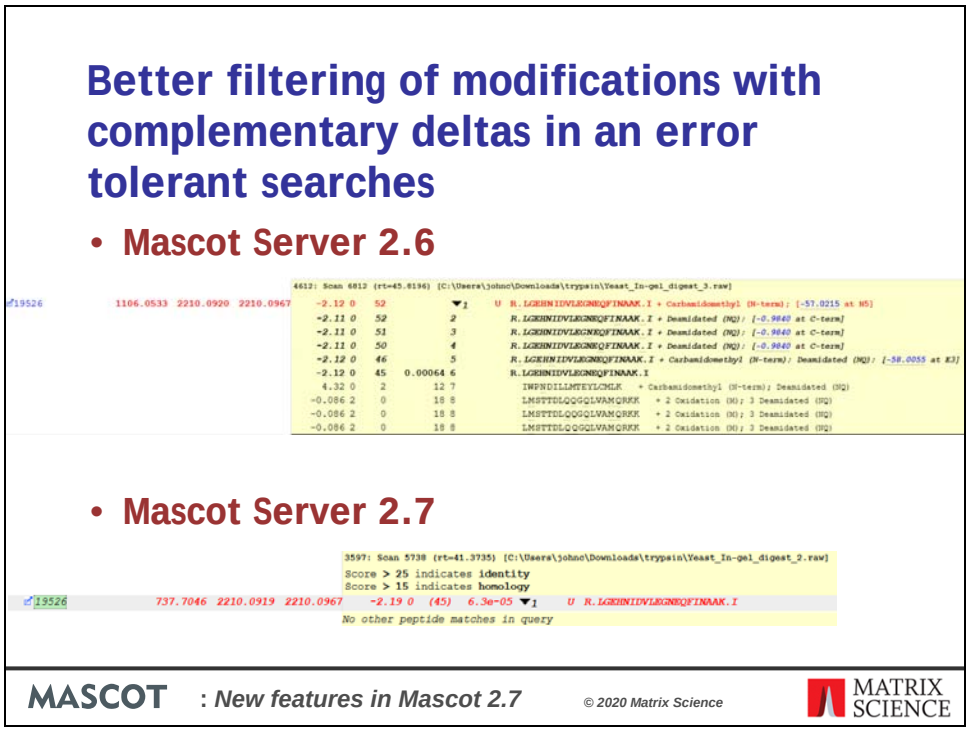
In Mascot Server 2.7 we extend that filtering to remove matches with pairs of modifications that where the deltas sum to zero within the precursor tolerance.

Such matches allow the fragment masses to be shuffled around to get a better score while leaving the parent mass unchanged. These are almost always false, with the true match, without either mod, getting a slightly lower score.

Some examples of this are issues are:

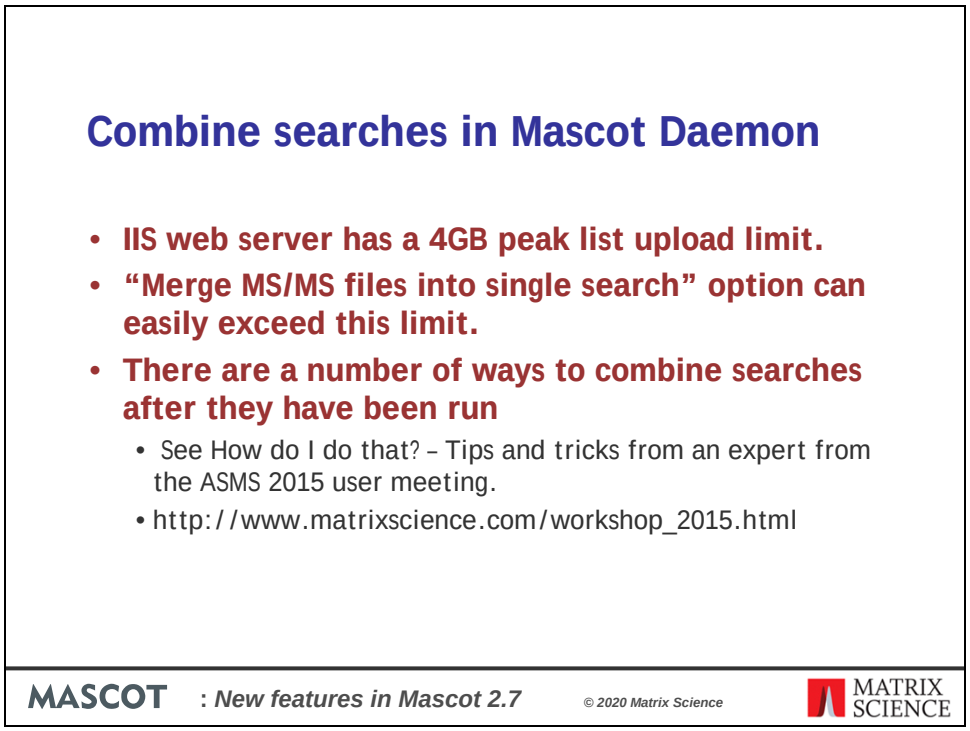
Asn->Gly and Gln->Ala are exact inversions of Carbamidomethyl

Deoxy, Ser->Ala, and Tyr->Phe are exact inversions of Oxidation



Here I show the reporting of a peptide With a Carbamidomethyl modification at the N-terminal and an equivalent loss of 57Da Asn->Gly modification at the Ans 5 position. Expanding the top 10 matches of the peptide shows the unmodified peptide with a significant score 6th on the list.

In Mascot Server 2.7 all of the top 5 matches have been filtered out, and the bottom 4 too, leaving just the unmodified sequence match.

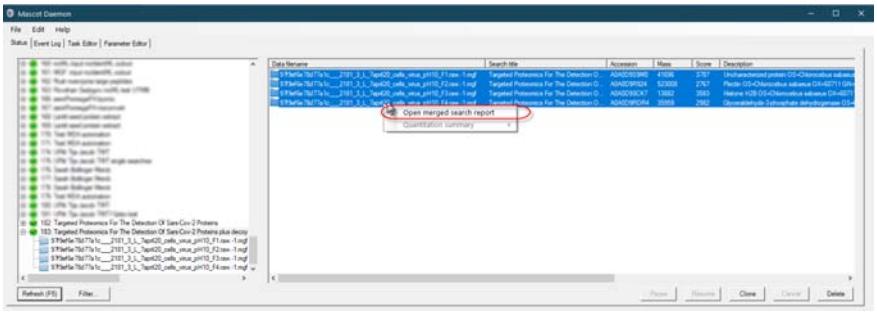


We added a new feature to Mascot Daemon which allows you to easily combine search results.

When working with fractionated samples you will want to combine the searches from multiple data files into a single result. The easy way to do this is with the "Merge MS/MS files into single search" option in Mascot Daemon. However if you are using the Windows operating system for your Mascot Server and the default IIS web server, you can hit a 4 GB upload limit.

There are a number of ways around this limit such as using Mascot Daemon on the same computer as Mascot Server, so the searches are submitted on the command line, moving from the IIS web server to Apache or using Mascot Distiller for peak picking and merging the results there. These are described in an older talk from 2015.

Combine searches in Mascot Daemon



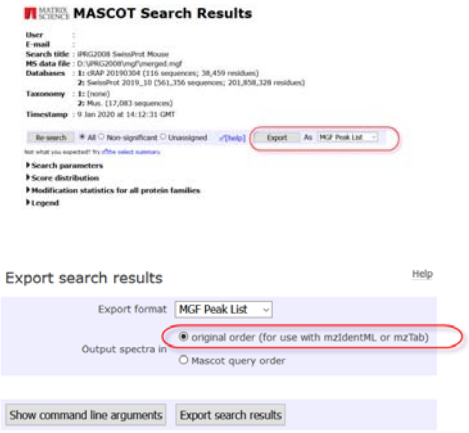
Now Mascot Daemon 2.7 gives you another way to merge searches. This is useful for when users are using msconvert or other peak picking third party tool and not using the merging features of Mascot Distiller.

Instead you can select multiple searches in a Mascot Daemon task by CTRL+click individually searches or shift+click a range then right click and choose combined report.

The combined search will open in a web page and list the results files that have been merged at the top of the report.

This will work with searches that have been processed by any peak picking software including Mascot Distiller.

MGF exporting spectra in the original order



This next feature requires some background. When a peak list is submitted to Mascot, the neutral molecular mass (Mr) is calculated for each precursor and queries are sorted lowest to highest. The precursors are given a new index number based on this sorting.

You can run into a problem related to the change in order of the precursors in an MGF file when using search results with third party analysis software. When generating a mzIdentML file from the Mascot Server search results the spectrum IDs field in the mzIdentML file refer to the original MGF file precursor order.

Some third-party analysis software, such as the PRIDE submission tool, maps the data in the MGF file to the data in the mzIdentML file. If a collaborator exports the search results from Mascot but does not have access to the original mgf file they can export it from the search results.

In Mascot Server 2.6 the newly created MGF file is exported in the sorted order of the precursors, not the original order, the mapping between the mzIdentMl file and the MGF file will fail and the collaborator is left thinking that the either Mascot Server or the third party tool is broken. To prevent this misunderstanding the MGF file can now be exported with the precursors in the original order to maintain mapping compatibility.

To put the new feature to use go to the Export menu, choose MGF Peak list, and it is the default option.

Handle non-standard modification names in a spectral library

- Proteomics Standards Initiative modification names guidelines introduced in 2008.
- Old names or acronyms did not meet the new guidelines.
- Older data sets mined for spectral libraries contained these obsolete names and the spectra were not recognized by Mascot.

In 2008 the Proteomics Standards Initiative released community standards for representation of protein modification data by use of a consensus nomenclature. Prior to that there were there were some differences in modification names between search engines and research groups. Also some of the old names or acronyms did not meet the new guidelines. Some of the modifications had a common name but it did not conform to the new standards, for example "Sodiated" = "Cation:Na". Or used a lab acronyms "CAM" = "Carbamidomethyl".

These names live on in old data sets some of which have been used to create spectral libraries that are publicly available from NIST. Mascot Server could not match these modification names to the newer ones and would report an error for those spectra, effectively excluding them from the library.

Handle non-standard modification names in a spectral library

- **mascot/config/library_mod_aliases**
  - Preconfigured with 46 common aliases and acronyms

```
# The default list provided is supplied by NIST
#
"AB_old_ICATd0" = "ICAT-D"
"AB_old_ICATd8" = "ICAT-D:2H(8)"
"Acetylation" = "Acetyl"
"CAM" = "Carbamidomethyl"
"Deamidation" = "Deamidated"
"Dehydro" = "Dehydrated"
"di-Methylation" = "Dimethyl"
"dihydroxy" = "Dioxidation"
"formylkynurenin" = "Dioxidation"
"HMVK86" = "HMVK"
"hydroxykynurenin" = "Trp->Hydroxykynurenin"
```


To fix this we have introduced an aliases file that maps the old names on to the new names.

The aliases file resides in the Mascot/config directory and is preconfigured with 46 common aliases and acronyms.

It is extendible as well.