Very Large Searches

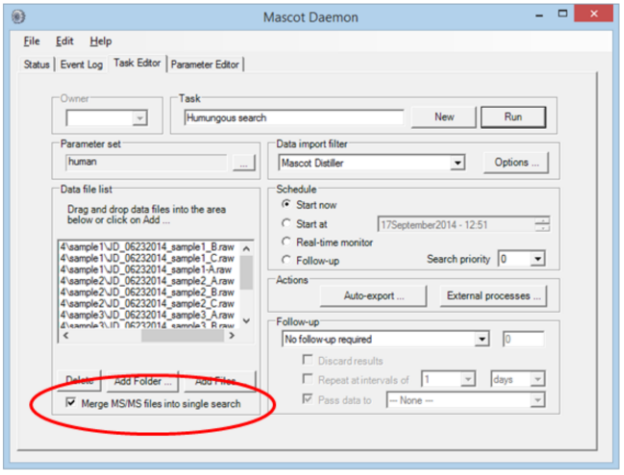
Topics

- Combining data files
- Performing large searches
- The Protein Family summary
- Protein scoring – standard vs. MudPIT
- Exporting results

Very large searches present a number of challenges. These are the topics we will cover during this presentation.

Data files

- Can use Mascot Daemon to process and merge fractions
- Use Distiller or a file specific data import filter



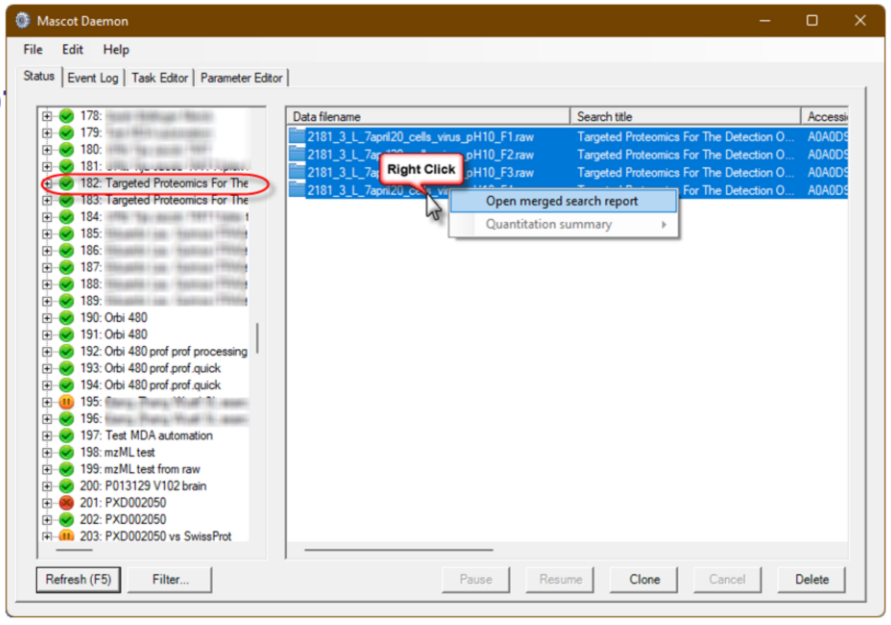
The smartest way to merge files, like fractions from a fractionated run, is using Mascot Daemon. Just tick the box at the bottom left.

The batch can be peak lists or raw files

Note that Mascot Daemon 2.1 had a file size limit of 2 GB. This was lifted in 2.2, and we have successfully merged and searched a 6 GB file on a Linux system.

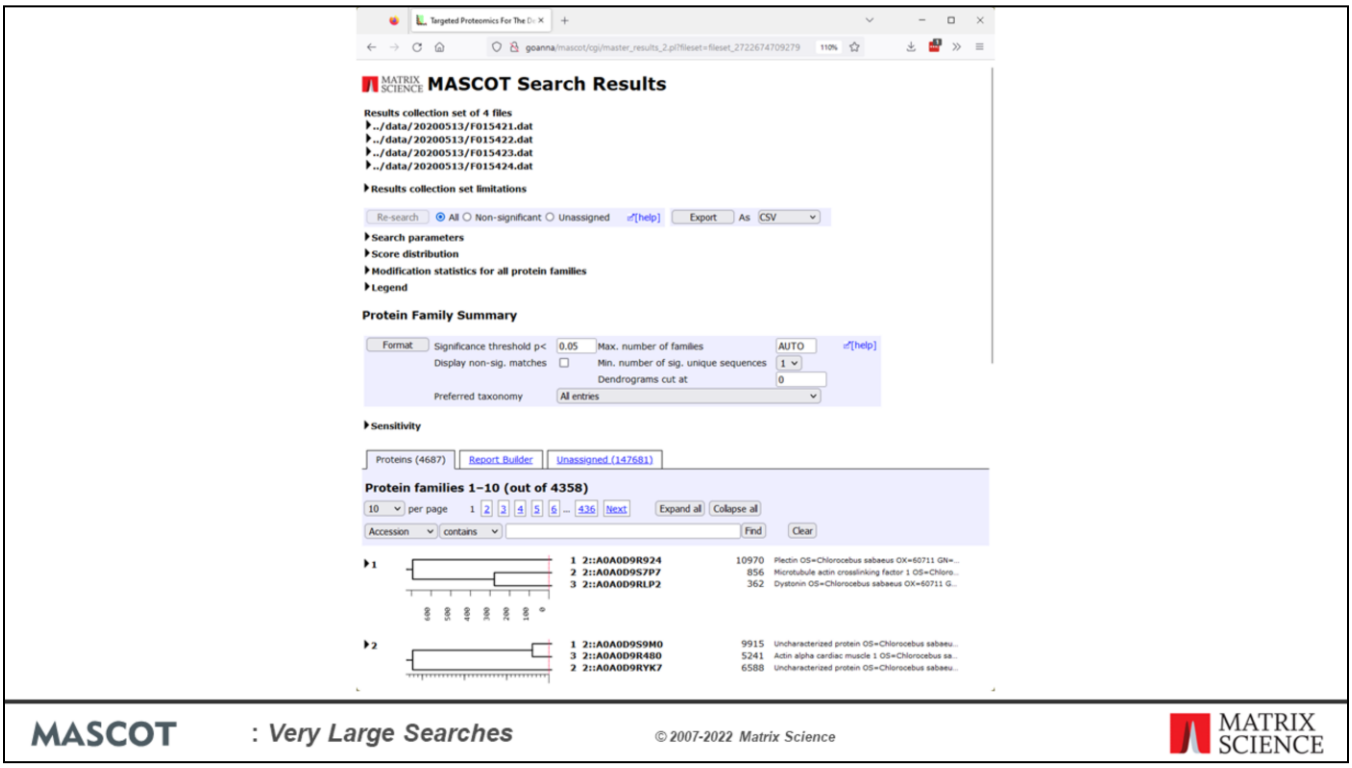
For windows web servers the upload limit is 4 GB. In Mascot Daemon 2.6 we introduced an option to run searches from the command line if Mascot Daemon and Mascot Server are installed on the same computer. This bypasses any web server file limit and search sizes are effectively unlimited.

Comb



Now Mascot Daemon 2.7 gives you another way to merge searches.

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.

Data files

Concatenating peak lists:

- DTA or PKL

  Download merge.pl from the Matrix Science Xcalibur help page
  `http://www.matrixscience.com/help/instruments_xcalibur.html`

  Retains filename as scan title

```
BEGIN IONS
TITLE=raft3031.1706.1706.2.dta
CHARGE=2+
PEPMASS=1243.577388
451.1228 5080
487.4352 3283
550.4203 5087
```


If you don't want to use Daemon, you can merge peak lists manually.

For DTA or PKL, you can download a script from our web site.

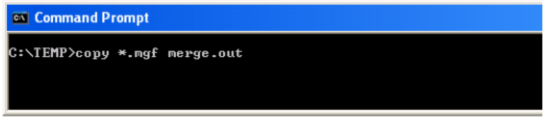
A nice feature of this script is that it puts the filename into the scan title, so you can tell which fraction a particular spectrum came from. The scan titles are displayed in the yellow pop-ups on the Mascot result report

Data files

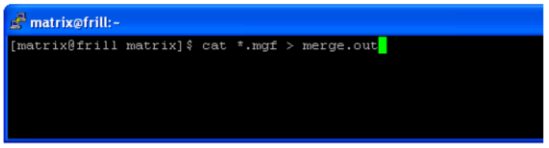
Concatenating peak lists:

- MGF

Windows: copy



Unix: cat



As long as MGF files contain only peak lists, you don't need a script. Just use copy or cat

If the MGF files have search parameters at the beginning, you'll need to remove these before merging the files. Because a number of third party utilities add commands to MGF headers, and these cause a merged search to fail, Mascot Daemon 2.3 and later strips out header lines when merging MGF files.

Data files

- Average spectrum might contain 100 real peaks
- Each peak might require ~ 20 bytes
  967.41590 [tab] 470.20193 [newline]
- 2 GB should be sufficient for ~ 1 million spectra
- If your peak list is orders of magnitude larger than 2kB / spectrum, then something is not right!

In talking to Mascot users, it is clear that peak lists files are often much bigger than they should be. In other words, the peak detection is not very good. If you do a back of the envelope calculation, you can see that 2 GB should be enough for approximately 1 million spectra.

If you intend to do a lot of large searches, its worth getting the peak detection right. Shipping unnecessarily large files around wastes both time and disk space

Performing large searches

32 bit platforms: maximum process size 2GB

Mascot divides large searches into chunks

- mascot.dat:

```
SplitNumberOfQueries 1000
SplitDataFileSize 10000000
```

Consequences:

- Search size is "unlimited" (except by disk space)
- No protein summary section in result file

32 bit platforms have a maximum process size of 2 GB on Windows or 3Gb on Linux. To get around this limit, Mascot divides large searches into smaller chunks, so as to avoid having everything in memory at the same time. The parameters to control this are SplitNumberOfQueries and SplitDataFileSize in the Options section of mascot.dat

One consequence of splitting a search is that there is no protein summary section in the result file. This is not a problem, because no-one wants a protein summary report for a large MS/MS search. However, some old client software gets confused by the missing section. The work around is to increase the values so that large searches never split. Maybe setting SplitNumberOfQueries to 1 million spectra and SplitDataFileSize to 10 billion bytes.

This is OK, but remember to reset these values as soon as you are able to. Otherwise, you might find you run out of memory or address space for your large searches

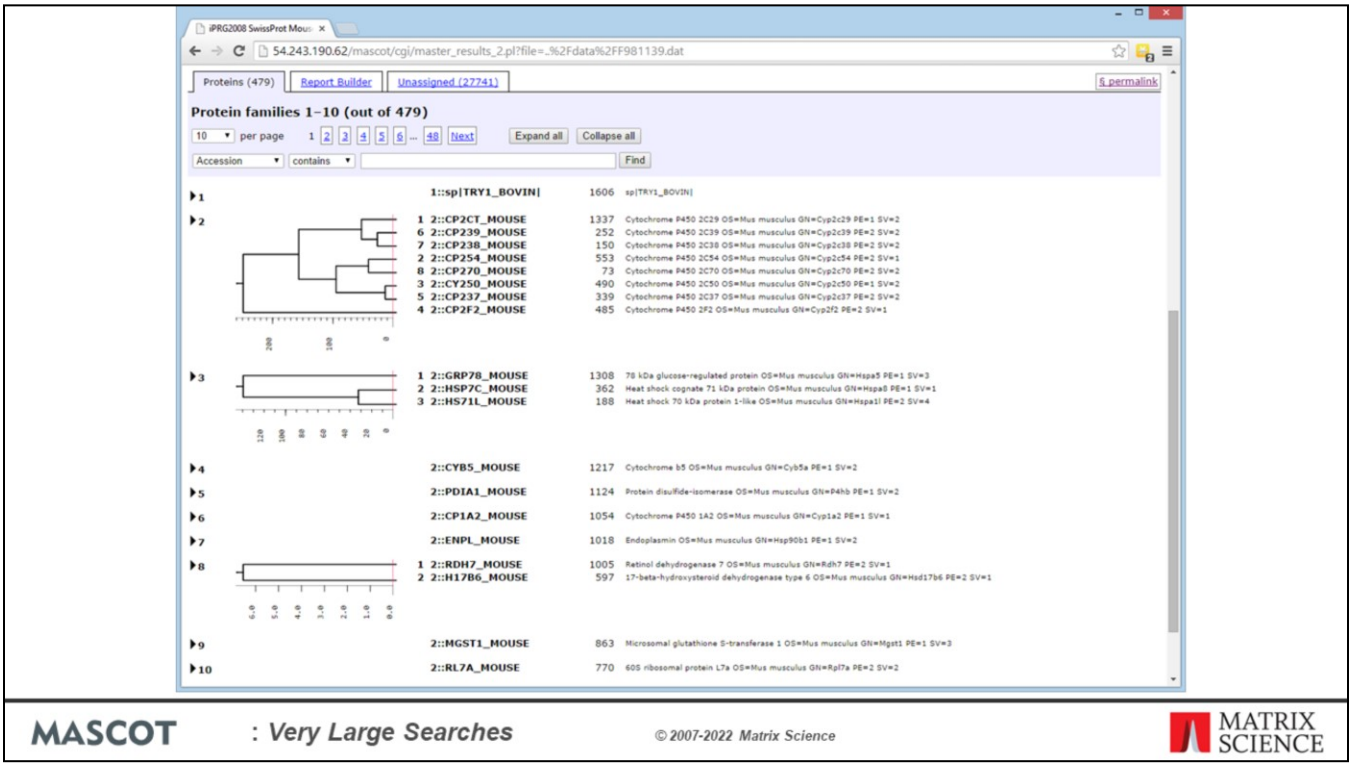
Mascot Server is now fully 64 bit. If you have enough RAM you could avoid splitting the search into chunks, however we still do by default because there is no performance penalty in doing so.

Reporting large searches

Protein Family Summary

- Paged report to conserve memory
- Detailed information is shown 'on demand'
- Index files are created and cached to speed loading in future
- Proteins grouped into families by means of shared peptide matches
- Hierarchical clustering within each protein family

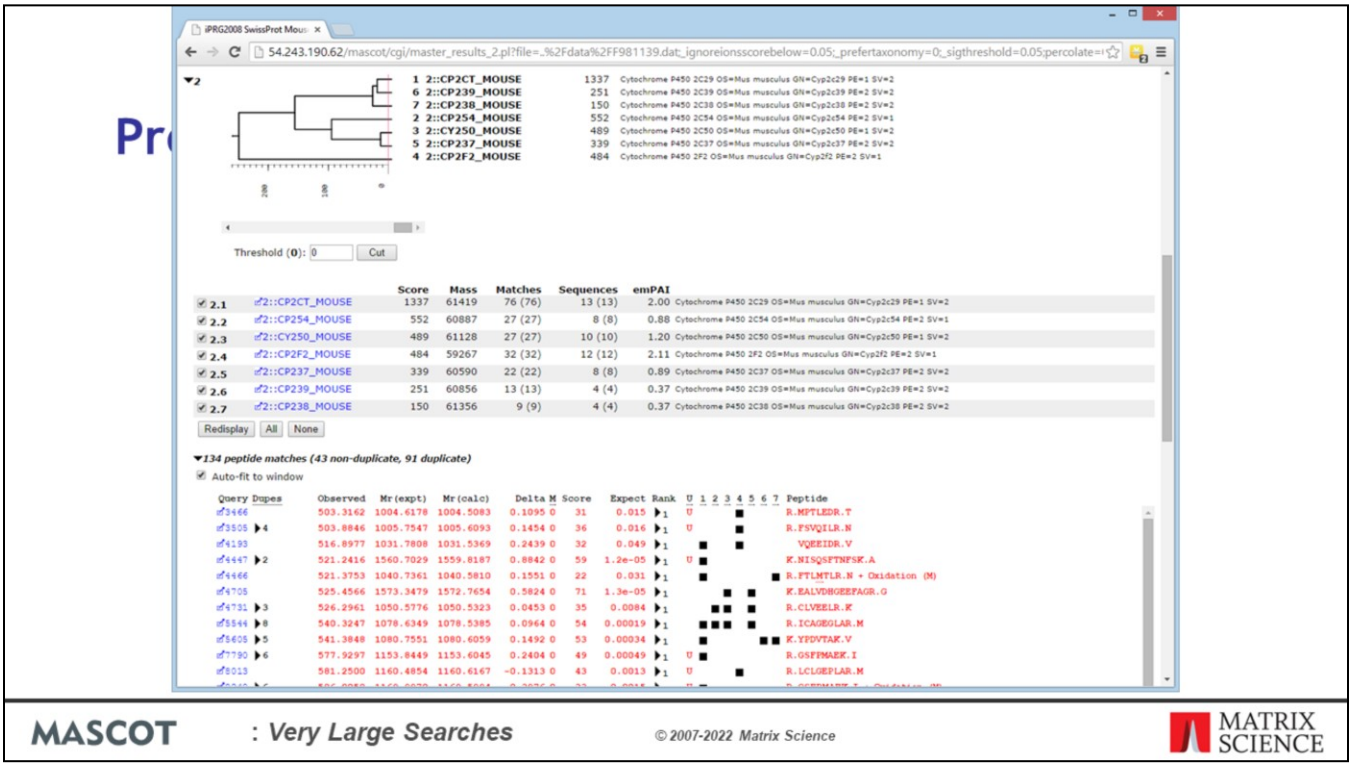
In early versions of Mascot, trying to display result reports for very large searches would often lead to problems with timeouts and running out of memory. To address this, the Protein Family Summary loads most of the information 'on demand'. This requires some index files to be created on the server, and these index files are cached, so that the report loads much faster on the second and subsequent occasions. Proteins are grouped into families by means of shared peptide matches and, within each family, hierarchical clustering is used to illustrate which proteins are closely related and which are more distant.



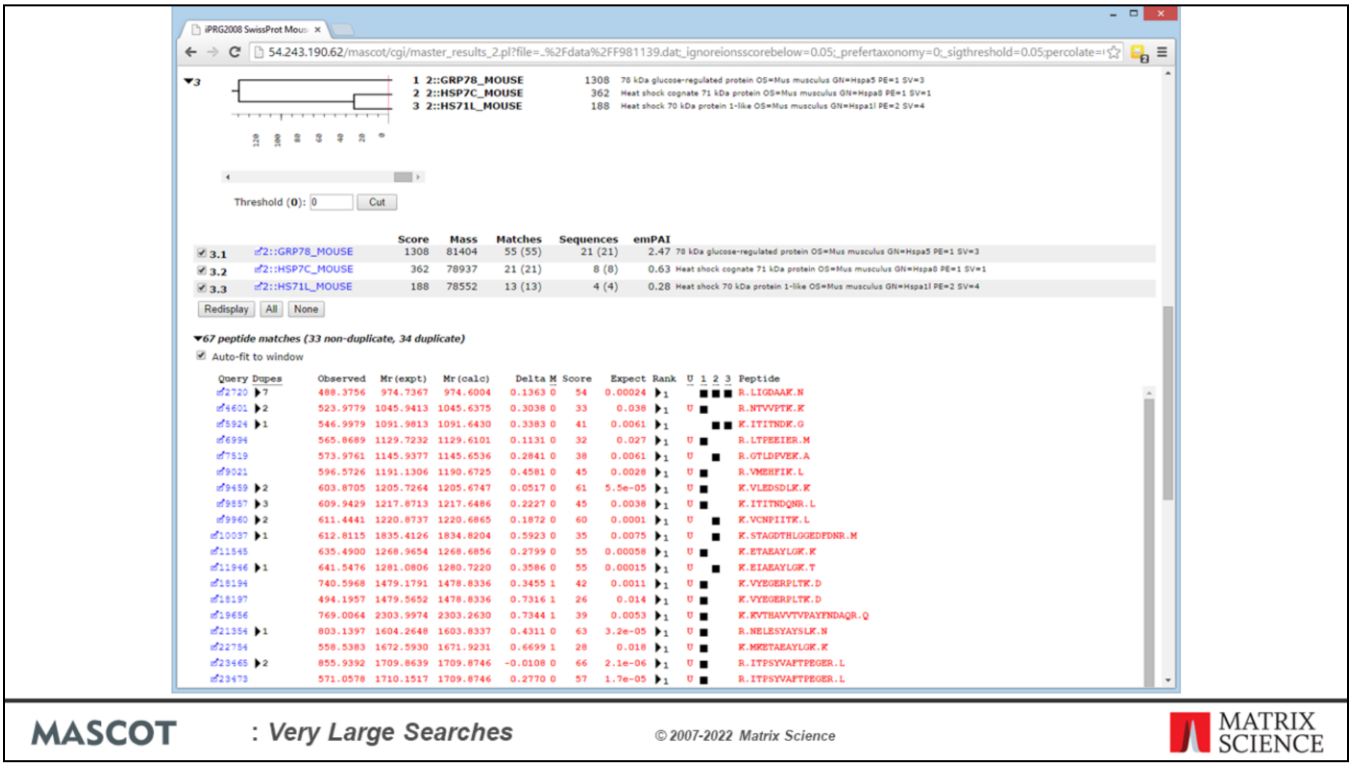
If there are 300 or more spectra, the Protein Family Summary is the default. This is the appearance of a typical family report immediately after loading. The body of the report consists of three tabs, one for protein families, one for Report Builder, and one for unassigned matches. The report is paged, with a default page size of 10 families. If you wish, you can choose to display a larger number of families on a single page.

Proteins are grouped into families using a novel hierarchical clustering algorithm. If the family contains a single member, the accession string, protein score and description are listed. If the family contains multiple members, the accessions, scores and descriptions are aligned with a dendrogram, which illustrates the degree of similarity between members.

The scores for the proteins in family 2 vary from 1337 down to 73. In the earlier Peptide Summary or Select Summary reports, these would have been at opposite ends of the report. It would have been difficult to recognise that these proteins belonged together, even though they have shared peptide matches and are all cytochrome P450.



If you are interested in family 2, then you click to expand it to show the details. Immediately under the dendrogram is a list of the proteins. The table of peptide matches is similar to that found in the other result reports. We only reports the significant peptide matches less than the p value for the search that defaults to $p<0.05$. Duplicate matches to the same sequence are collapsed into a single row. The columns headed 1, 2, 3, etc. represent the proteins and contain a black square if the peptide is found in the protein. Some matches are shared, but each protein has some unique peptide matches, otherwise it would be dropped as a sub-set.

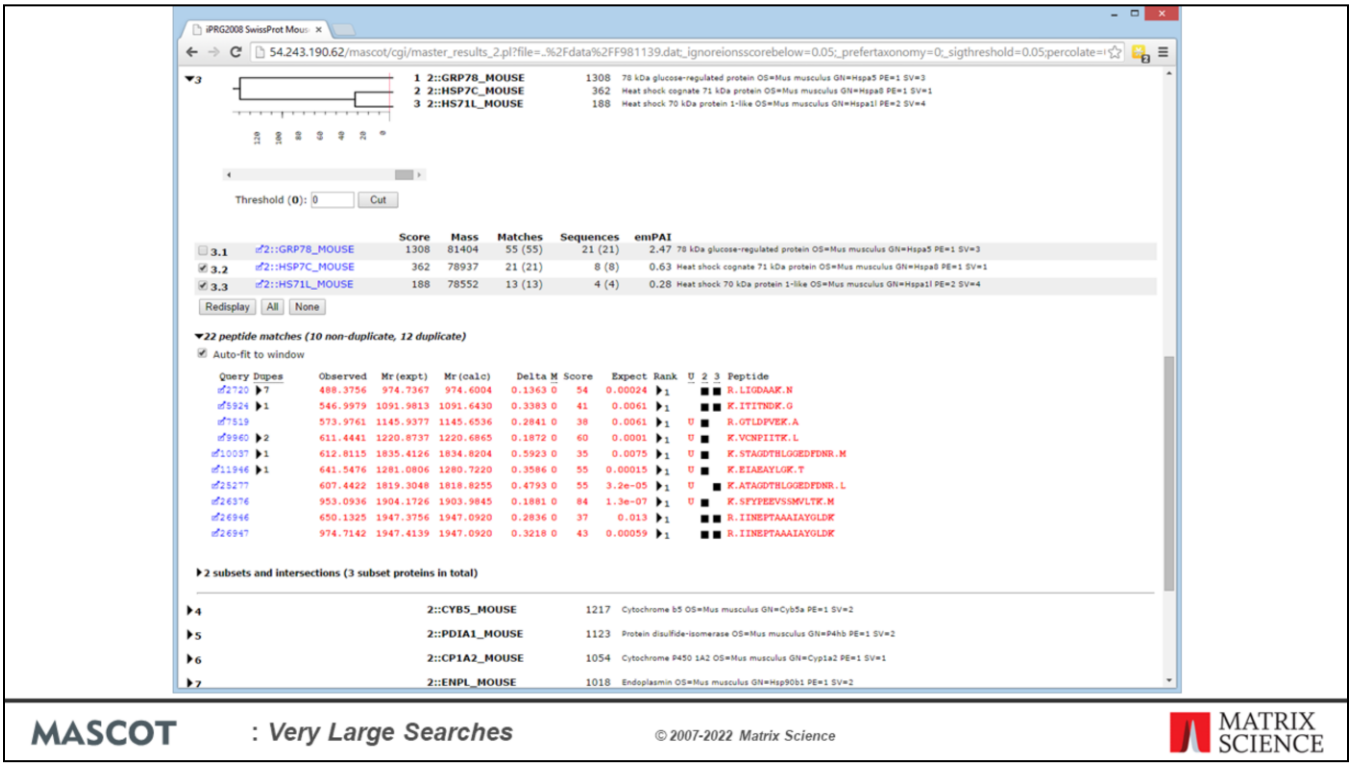


Moving down to family 3, the scale on the dendrogram is ions score, and HSP7C_MOUSE and HS71L_MOUSE join at a score of approximately 30. This represents the score of the significant matches that would have to be discarded in order to make one protein a sub-set of the other. These two proteins are much more similar to one other than to GRP78_MOUSE, which has non-shared peptide matches with a total score of approximately 145. Note that, where there are multiple matches to the same peptide sequence, (ignoring charge state and modification state), it is the highest score for each sequence that is used.

Immediately under the dendrogram is a list of the proteins. In this example, because SwissProt has low redundancy, each family member is a single protein. In other cases, a family member will represent multiple same-set proteins. One of the proteins is chosen as the anchor protein, to be listed first, and the other same-set proteins are collapsed under a same-set heading. There is nothing special about the protein picked for the anchor position. You may have a preference for one according to taxonomy or description, but all proteins in a same-set group are indistinguishable on the basis of the peptide match evidence.

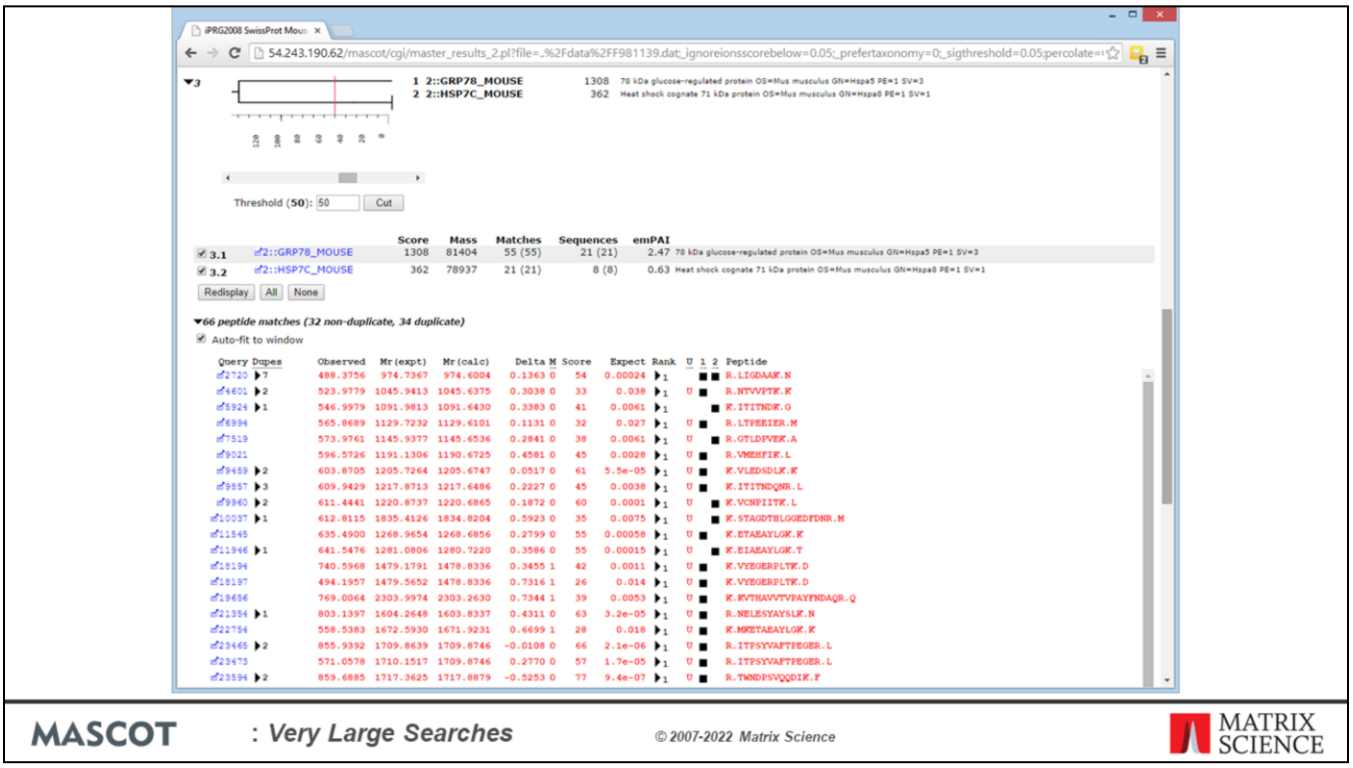
The table of peptide matches is similar to that found in the other result reports. Duplicate matches to the same sequence are collapsed into a single row. Click on the triangle to expand.

The black squares to the right show which peptides are found in which protein. To see the peptides that distinguish HSP7C_MOUSE and HS71L_MOUSE, clear the checkbox for GRP78_MOUSE and choose Redisplay.

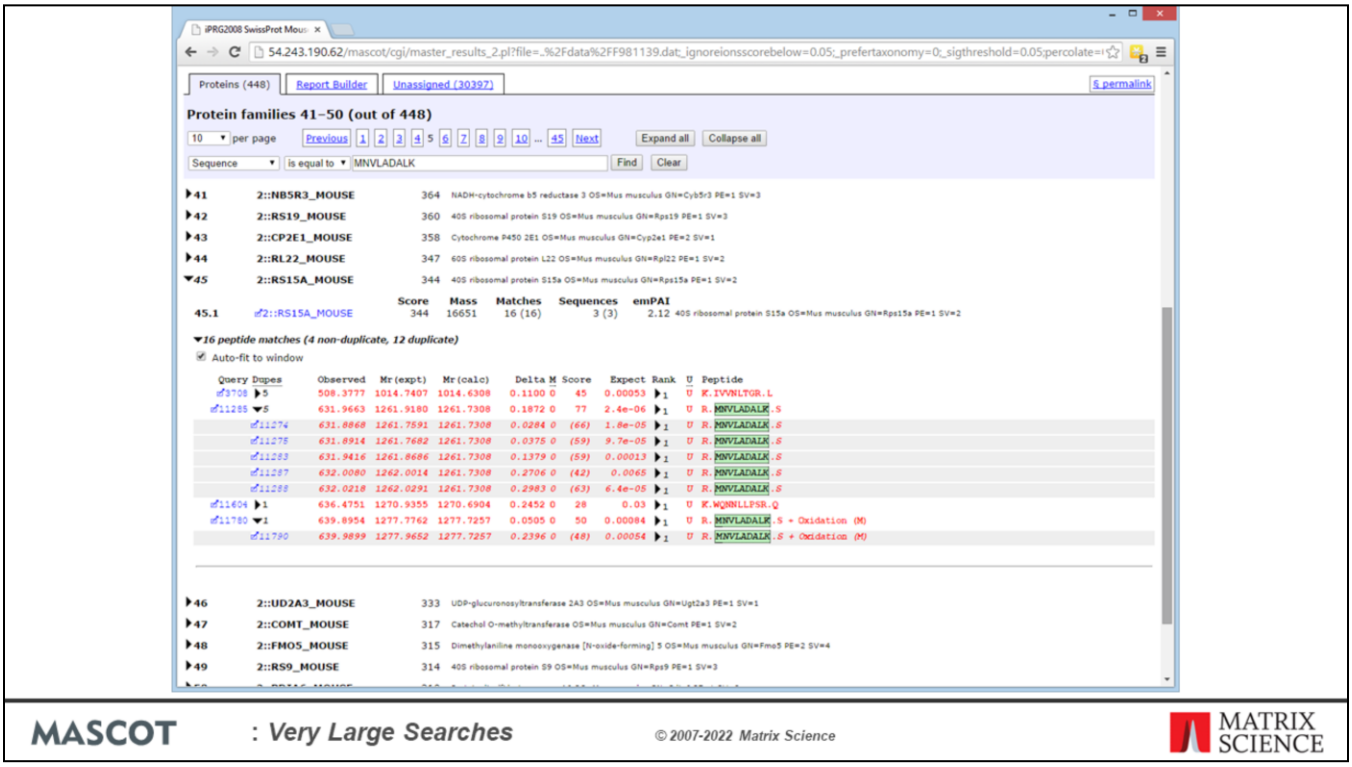


It can now be seen that HS71L_MOUSE would be a sub-set of HSP7C_MOUSE if it was not for one match, K.ATAGDTHLGGEDFDNR.L. It is the significant score for this match that separates the two proteins in the dendrogram by a distance of 32 (score of 55 - homology threshold score of 23).

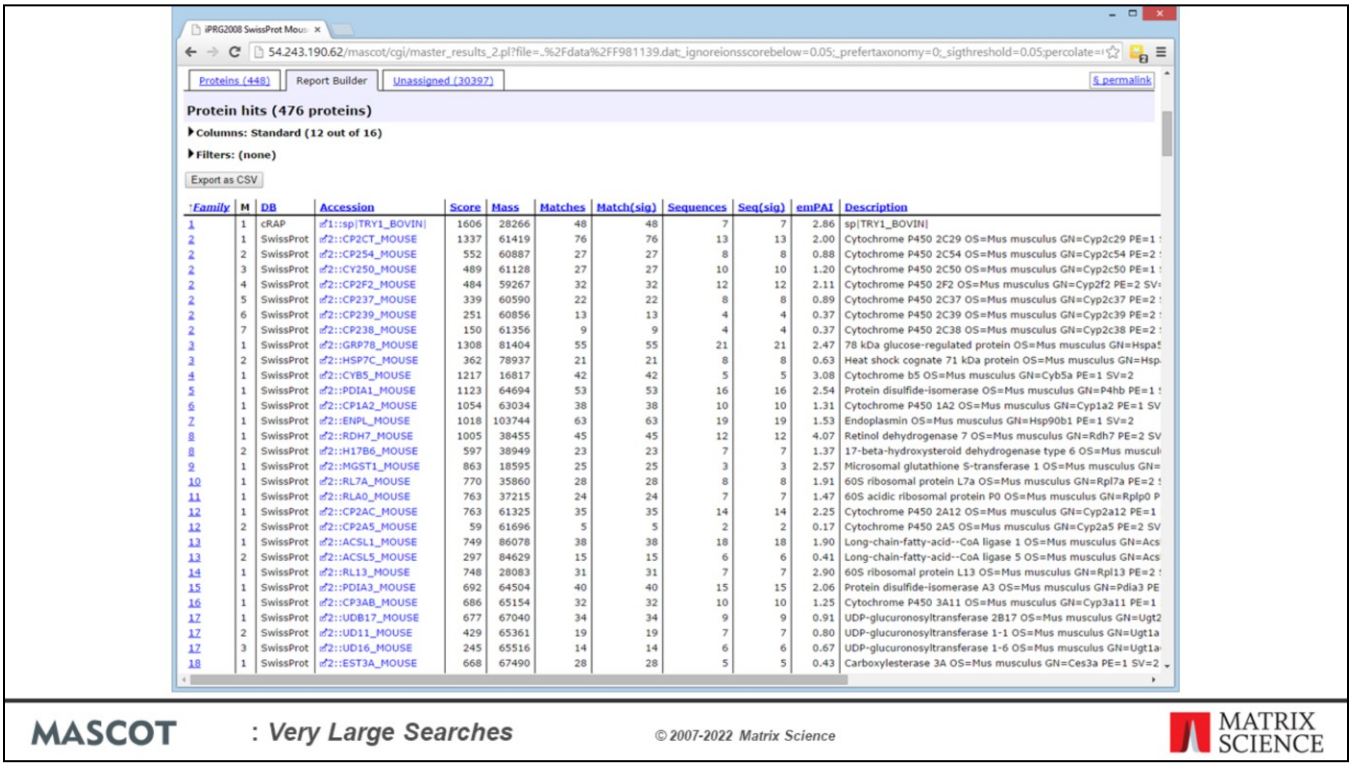
You can "cut" the dendrogram using the slider control.



If we cut the dendrogram at a score of 50, HS71L_MOUSE will be dropped because it is now a sub-set protein. If you compare the matches to HSP7C_MOUSE with those to GRP78_MOUSE, it is clear that these are very different proteins. They are part of the same family because of two shared matches, but many highly significant matches would have to be discarded for either protein to become a sub-set of the other. In summary, we can quickly deduce from the Family Summary that there is abundant evidence that both GRP78_MOUSE and HSP7C_MOUSE were present in the sample. There is little evidence for HS71L_MOUSE. It is more likely that the HSP7C_MOUSE contained a SNP or two relative to the database sequence.

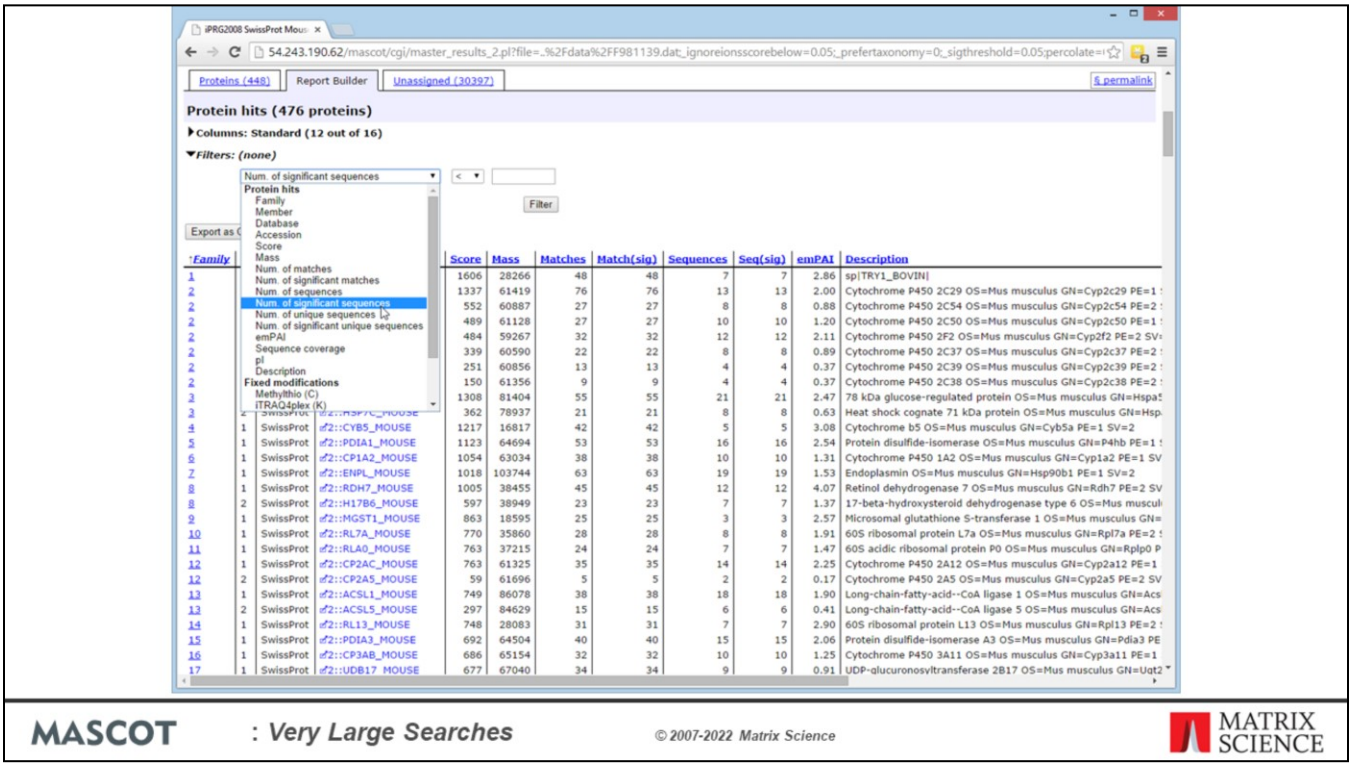


The family report also includes a text search facility, which is particularly important for a paged report. You can search by accession or description sub-string, or by query, mass or sequence. Here, for example, we searched for a peptide sequence. The display jumps to the first instance of the sequence, expands, and highlights (in green) the target peptides.

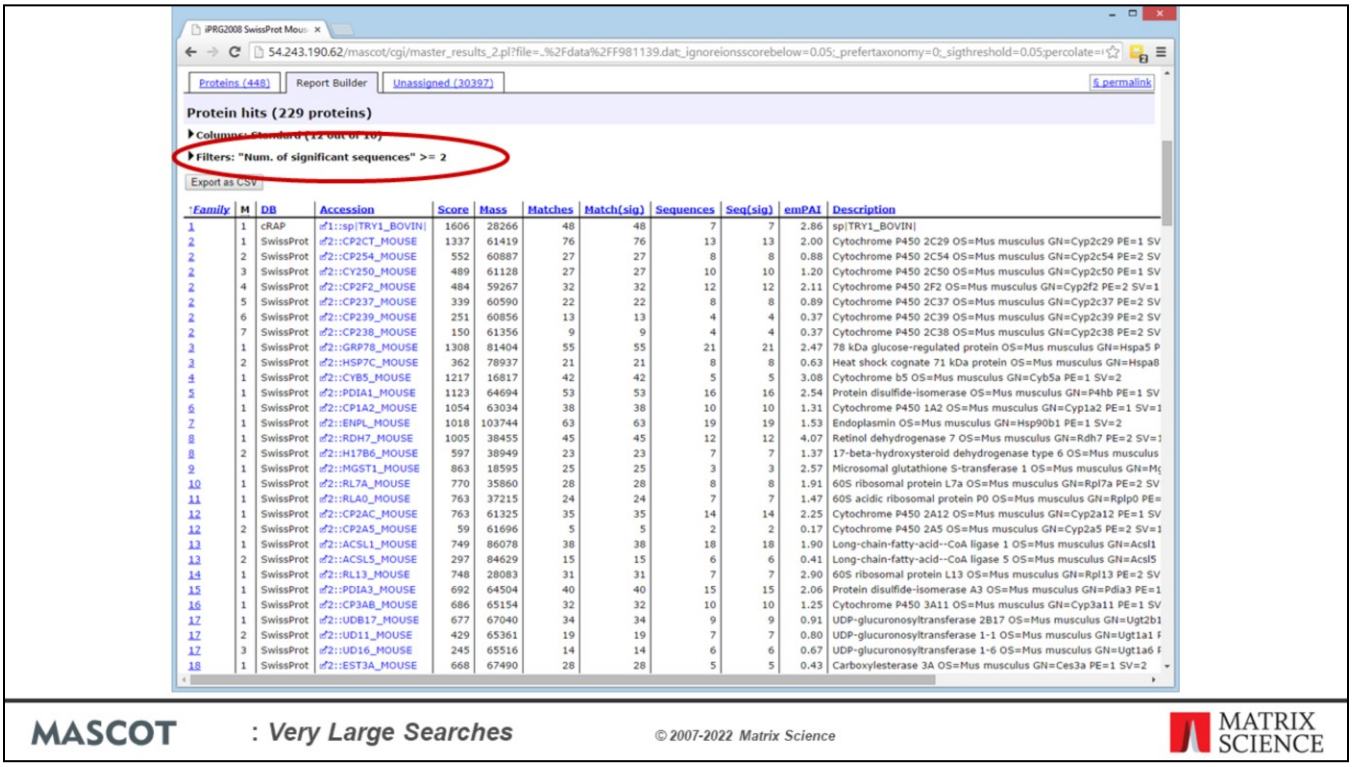


Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	cRAP	⚤1::sp|TRY1_BOVIN|	1606	28266	48	48	7	7	2.86	sp|TRY1_BOVIN|
2	1	SwissProt	⚤2::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1
2	2	SwissProt	⚤2::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2
2	3	SwissProt	⚤2::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1
2	4	SwissProt	⚤2::CP2F2_MOUSE	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=
2	5	SwissProt	⚤2::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2
2	6	SwissProt	⚤2::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2
2	7	SwissProt	⚤2::CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2
3	1	SwissProt	⚤2::GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5
3	2	SwissProt	⚤2::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hsp
4	1	SwissProt	⚤2::CYB5_MOUSE	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	⚤2::PDIA1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1
6	1	SwissProt	⚤2::CP1A2_MOUSE	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV
7	1	SwissProt	⚤2::ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmin OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1	SwissProt	⚤2::RDH7_MOUSE	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV
8	2	SwissProt	⚤2::H17B6_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus muscul
9	1	SwissProt	⚤2::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=
10	1	SwissProt	⚤2::RL7A_MOUSE	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2
11	1	SwissProt	⚤2::RLA0_MOUSE	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 P
12	1	SwissProt	⚤2::CP2AC_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1
12	2	SwissProt	⚤2::CP2A5_MOUSE	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 PE=2 SV
13	1	SwissProt	⚤2::ACSL1_MOUSE	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acs
13	2	SwissProt	⚤2::ACSL5_MOUSE	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acs
14	1	SwissProt	⚤2::RL13_MOUSE	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2
15	1	SwissProt	⚤2::PDIA3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE
16	1	SwissProt	⚤2::CP3AB_MOUSE	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1
17	1	SwissProt	⚤2::UDB17_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2
17	2	SwissProt	⚤2::UD11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a
17	3	SwissProt	⚤2::UD16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a
18	1	SwissProt	⚤2::EST3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2

The Report Builder tab is useful when you need a table of proteins suitable for publication. Lets assume we want to drop the ‘one hit wonders’ and only report proteins that have significant matches to at least 2 different peptide sequences

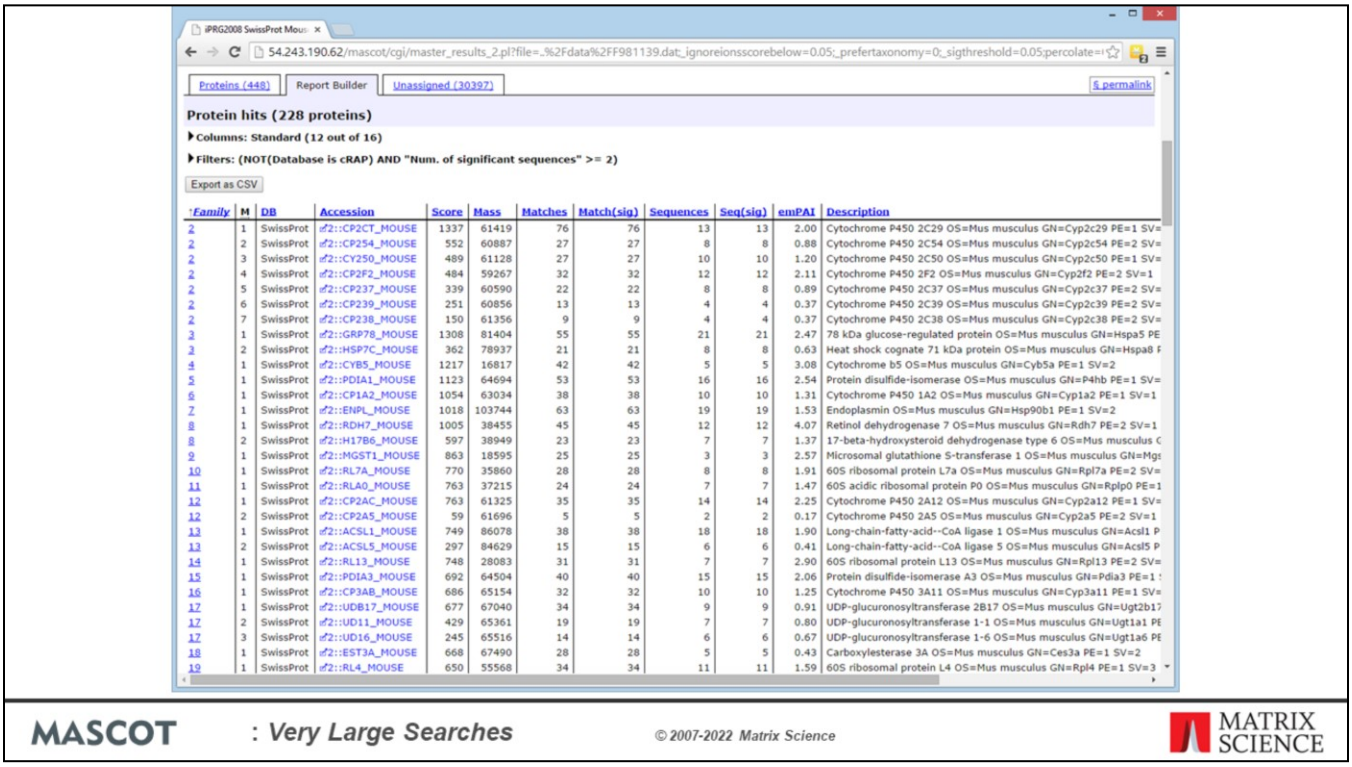


We open up the filters section and add a suitable filter.



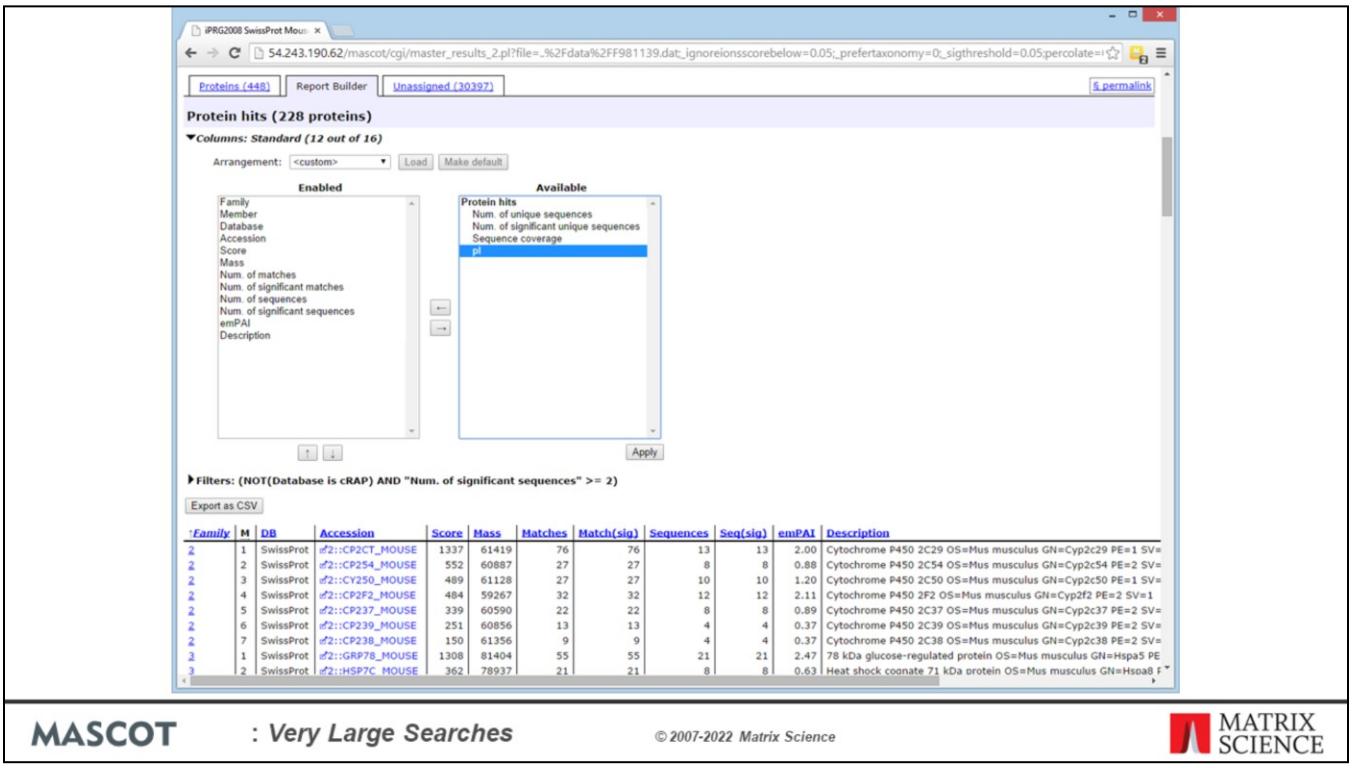
Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	cRAP	::sp|TRY1_BOVIN|	1606	28266	48	48	7	7	2.86	sp|TRY1_BOVIN|
2	1	SwissProt	::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV
2	2	SwissProt	::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV
2	3	SwissProt	::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV
2	4	SwissProt	::CP2F2_MOUSE	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1
2	5	SwissProt	::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV
2	6	SwissProt	::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV
2	7	SwissProt	::CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV
3	1	SwissProt	::GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 P
3	2	SwissProt	::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8
4	1	SwissProt	::CYB5_MOUSE	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	::PDIA1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV
6	1	SwissProt	::CP1A2_MOUSE	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
7	1	SwissProt	::ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmin OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1	SwissProt	::RDH7_MOUSE	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
8	2	SwissProt	::H17B6_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus
9	1	SwissProt	::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=M
10	1	SwissProt	::RL7A_MOUSE	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV
11	1	SwissProt	::RLA0_MOUSE	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 PE=
12	1	SwissProt	::CP2AC_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1 SV
12	2	SwissProt	::CP2A5_MOUSE	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 PE=2 SV=1
13	1	SwissProt	::ACSL1_MOUSE	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acsl1
13	2	SwissProt	::ACSL5_MOUSE	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acsl5
14	1	SwissProt	::RL13_MOUSE	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2 SV
15	1	SwissProt	::PDIA3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE=1
16	1	SwissProt	::CP3AB_MOUSE	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1 SV
17	1	SwissProt	::UDB17_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2b1
17	2	SwissProt	::UD11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a1 F
17	3	SwissProt	::UD16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a6 F
18	1	SwissProt	::EST3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2

Only proteins with significant matches to at least 2 sequences remain. The filtering is very flexible, with lots of useful terms.

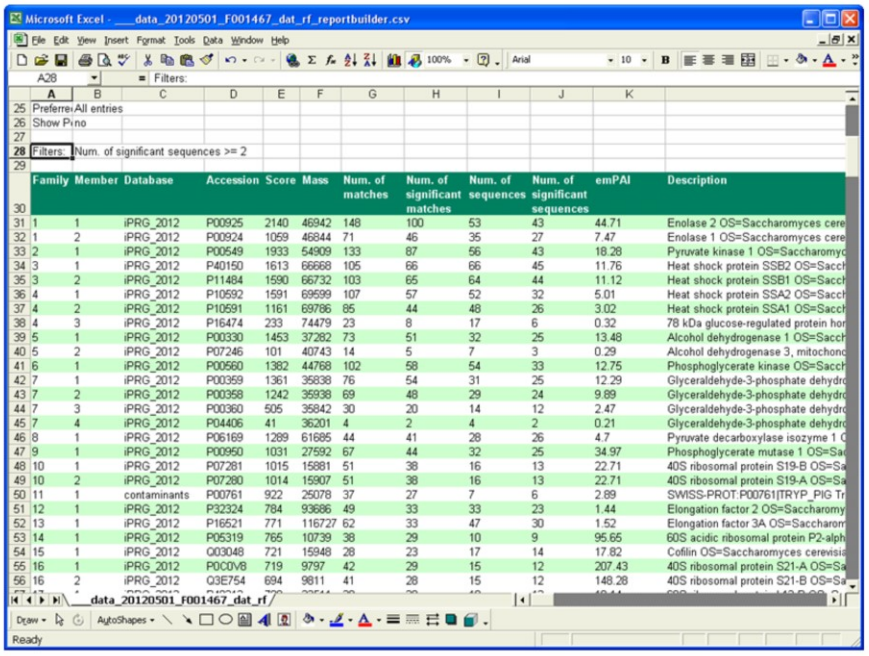


Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
2	1	SwissProt	::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=
2	2	SwissProt	::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=
2	3	SwissProt	::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=
2	4	SwissProt	::CP2F2_MOUSE	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1
2	5	SwissProt	::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV=
2	6	SwissProt	::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=
2	7	SwissProt	::CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=
3	1	SwissProt	::GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE
3	2	SwissProt	::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 F
4	1	SwissProt	::CYB5_MOUSE	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	::PDIA1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV=
6	1	SwissProt	::CP1A2_MOUSE	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
7	1	SwissProt	::ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmin OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1	SwissProt	::RDH7_MOUSE	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
8	2	SwissProt	::H17B6_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus G
9	1	SwissProt	::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgs
10	1	SwissProt	::RL7A_MOUSE	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=
11	1	SwissProt	::RLA0_MOUSE	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 PE=1
12	1	SwissProt	::CP2AC_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1 SV=
12	2	SwissProt	::CP2A5_MOUSE	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 PE=2 SV=1
13	1	SwissProt	::ACSL1_MOUSE	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acsl1 P
13	2	SwissProt	::ACSL5_MOUSE	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acsl5 P
14	1	SwissProt	::RL13_MOUSE	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2 SV=
15	1	SwissProt	::PDIA3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE=1
16	1	SwissProt	::CP3AB_MOUSE	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1 SV=
17	1	SwissProt	::UDB17_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2b17
17	2	SwissProt	::UD11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a1 PE
17	3	SwissProt	::UD16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a6 PE
18	1	SwissProt	::EST3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2
19	1	SwissProt	::RL4_MOUSE	650	55568	34	34	11	11	1.59	60S ribosomal protein L4 OS=Mus musculus GN=Rpl4 PE=1 SV=3

Another thing that you could easily do would be to exclude proteins from the contaminants database

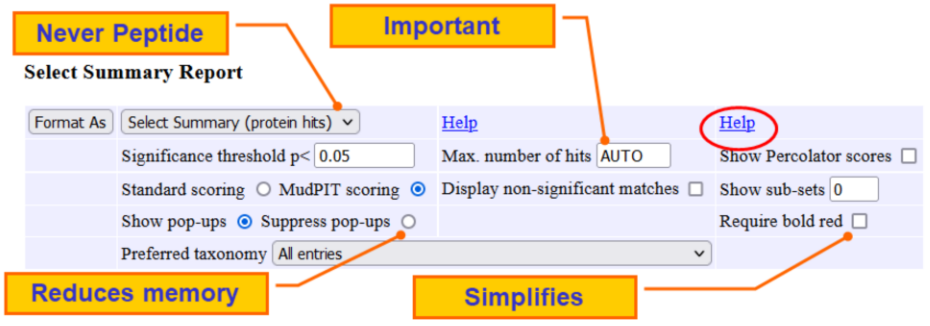


The columns section of Report Manager allows you to choose which columns to include and, if required, change their order



Once the list is filtered and the columns arranged as required, there is a button to export the table as CSV, which can be pasted into Excel and formatted to create a suitable figure for dropping into a publication

Large search results in 2.2 and earlier



http://…/master_results.pl?file=../data/20060202/F000123.dat
&REPTYPE=select &REPORT=AUTO &_showpopups=FALSE &_requireboldred=1

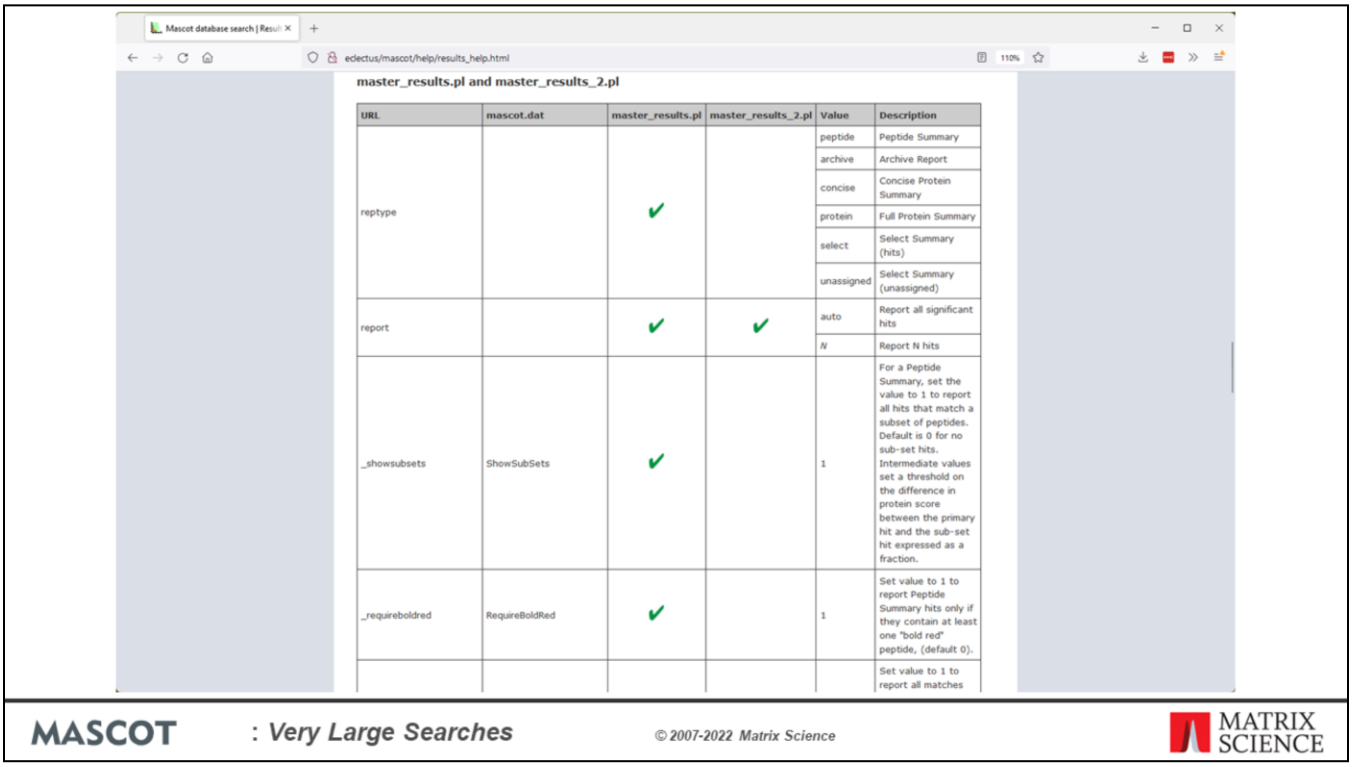
If you are still using Mascot 2.2 or if you have some application software that requires the results in the earlier format, and you are encountering problems with timeouts and running out of memory, here are some tips:

•Ensure you are using the Select report. If you are using a third party client that has specified Peptide summary or Protein summary, add this to the URL before opening the file: &REPTYPE=select

•Don't specify a huge number of hits 'just in case'. Choose AUTO to display all protein hits that contain at least one significant peptide match: &REPORT=AUTO

•Get rid of the yellow pop-ups: &_showpopups=FALSE

•Setting require bold red and leaving "Display non-significant matches" unchecked will minimize the number of hits: &_requireboldred=1



If you can't remember these URL parameters, just click on the help link

Reporting large search results

What do we mean by Standard scoring and MudPIT scoring?

Protein Scores for MS/MS Searches

Standard protein score

- the sum of the ions scores
- excluding the scores for duplicate matches, which are shown in parentheses
- correction to reduce the contribution of low-scoring random matches

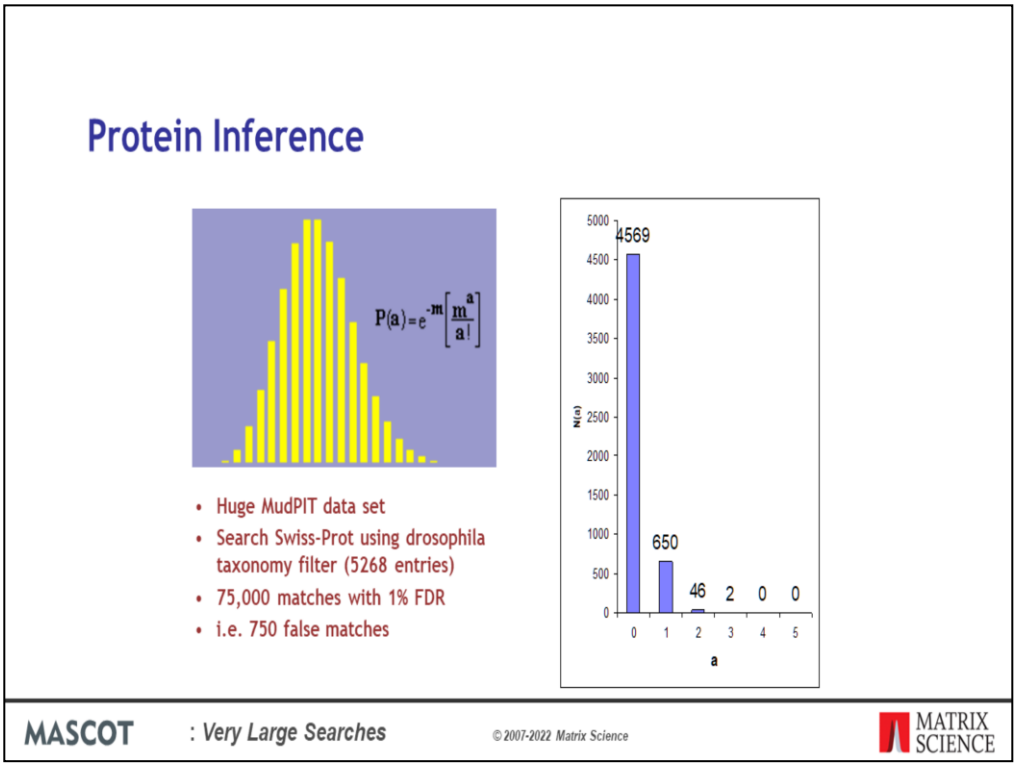
342. 2::IPI00023283 **Mass:** 3832803 **Score:** 181 **Matches:** 51(0) **Sequences:** 48(0)
Tax_Id=9606 Gene_Symbol=TTN Isoform 2 of Titin

Query	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Score	Expect	Rank	Unique	Peptide
28	359.7341	717.4537	717.4537	-0.09	0	7	4.2	5	U	R.LFAIVR.G
209	394.2371	786.4596	786.4599	-0.46	0	8	13	3	U	K.LTIADVR.A
334	411.2073	820.4000	820.3954	5.61	0	3	15	4	U	K.TDSGLVR.C
357	413.2642	824.5139	824.5135	0.48	1	12	1.1	5	U	K.RFLTLR.K
715	450.7365	899.4584	899.4588	-0.38	0	10	2.9	2	U	K.IVDVSSDR.C
740	**451.7681**	**901.5217**	**901.5233**	**-1.72**	**0**	**3**	**24**	**3**	**U**	**R.VTLVDVTR.N**
840	459.2484	916.4821	916.4767	5.98	0	2	29	2	U	K.GVEFNVPR.L
844	459.7299	917.4452	917.4454	-0.24	0	4	15	6	U	K.ELEETAAR.M
1029	**473.2757**	**944.5368**	**944.5331**	**3.97**	**1**	**3**	**21**	**3**	**U**	**R.EPPSFIKK.I**
1058	475.7505	949.4864	949.4869	-0.47	0	4	22	5	U	R.SSVSLSWGK.P
1066	**476.2790**	**950.5433**	**950.5425**	**0.94**	**0**	**1**	**23**	**4**	**U**	**R.PLTDLQVR.E**

MASCOT : *Very Large Searches* MATRIX SCIENCE

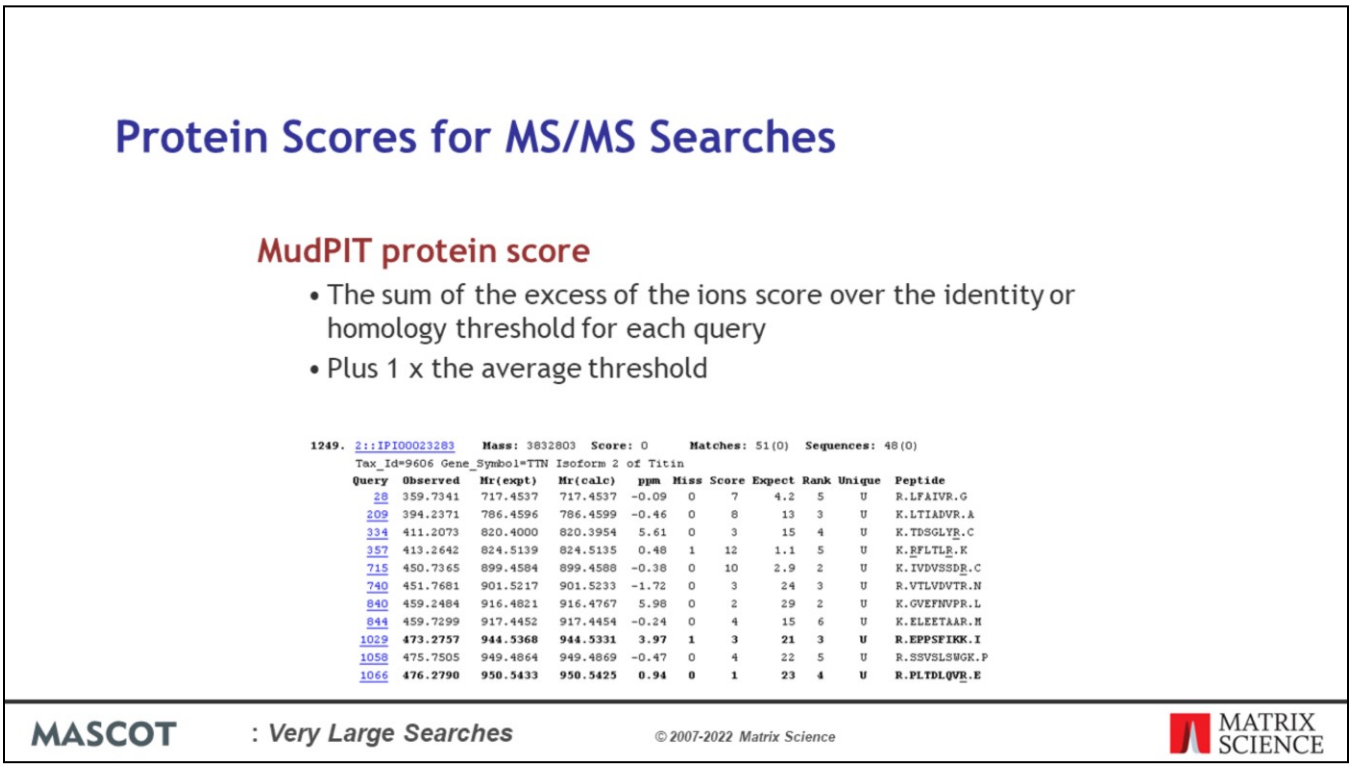
With standard peptide summary scoring, the protein score is essentially the sum of the ions scores of all the peptides assigned to the protein. Where there are duplicate matches to the same peptide, the highest scoring match is used. A correction is applied based on the number of candidate peptides that were tested. This correction is very small unless it is a very large protein, like here, or a no-enzyme search

Despite this correction, as this example shows, when we have many low scoring matches assigned to the same protein, we can still get a high protein score, even though none of the individual peptide matches are significant



A protein with matches to just a single peptide sequence is commonly referred to as a “one-hit wonder” and is often treated as suspect. This is actually a slight over-simplification. In a search with a large number of spectra and a small database, even though the peptide false discovery rate is low, a protein can pick up multiple false matches by chance. This is easily calculated using a Poisson Distribution, where m is the average number of false matches per protein. In this example, m is 750/5268, and we would expect 650 database entries to be one-hit wonders. However, 46 entries will pick up two false matches and 2 entries will pick up three, which could mean we report 48 false proteins.

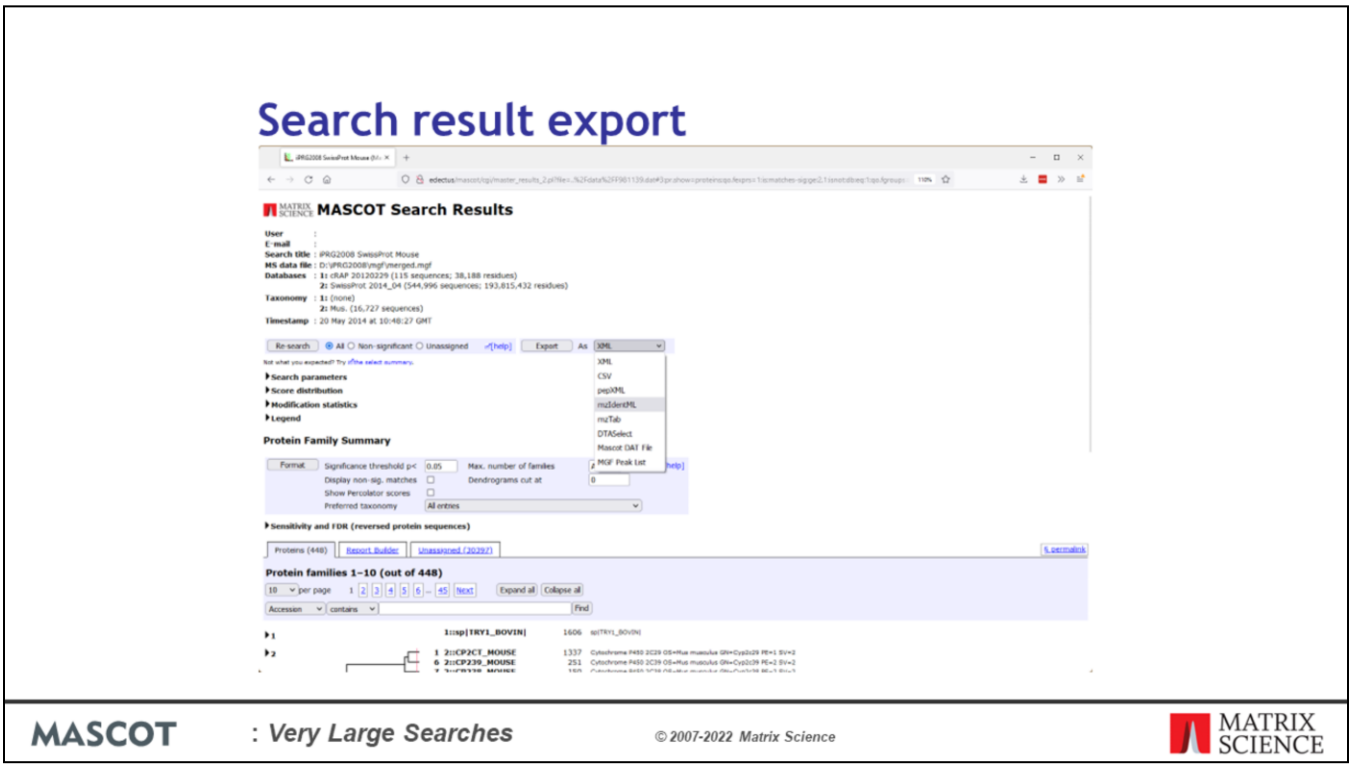
The problem isn’t limited to large searches. It is the ratio between the number of spectra and the number of entries in the database that matters. So, a small search against a small database can give similar numbers



To avoid this problem, we use MudPIT protein scoring, in which the score for each peptide match is not its absolute score, but the amount that it is above the threshold. Therefore, matches with a score below the threshold do not contribute to the score. The MudPIT protein score is the sum of the score excess over threshold for each of the matching peptides plus one times the average threshold. For each peptide, the "threshold" is the homology threshold if it exists, otherwise it is the identity threshold.

So, even though a large protein like titin may pick up several random matches, with MudPIT scoring, the protein score is zero, so you don't see it listed in the report unless you specify a huge number of protein hits, as was done here to capture this screen shot.

By default, MudPIT protein scoring is used when the ratio between the number of queries and the number of database entries, (after any taxonomy filter), exceeds 0.001. This default switching point can be moved by changing the value of MudpitSwitch in mascot.dat. You can also switch between the two scoring methods by using the format controls at the top of the report.



At some stage, it is likely that you will want to export the search results to another application or a relational database. If you want to write your own code, we provide a free library called Mascot Parser that provides a clean, object oriented programming interface to the result file. The supported languages are C++, Java, and Perl.

Mascot also includes a flexible export utility.

If you want the XML format, you probably know that this is what you want. If you've no idea what XML is, chances are you don't want it.

Choose CSV if you want to export to Excel - I'll show an example in a moment.

Choose pepXML if you want to export to Protein Prophet from ISB.

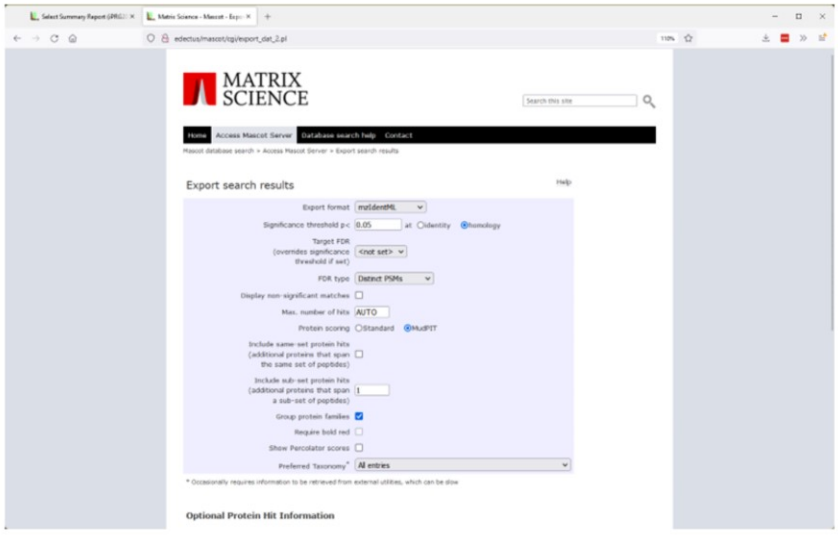
mzIdentML and mzTab are the standard formats from PSI for search result interchange. Mascot provides a very full implementation of mzIdentML and this is the one to choose if you are writing new application software that will use Mascot results

DTASelect is the tab separated format used by David Tabb's DTASelect program

The Mascot DAT file is the raw result file. If you need the result file for some reason, and don't have FTP or SCP access to your Mascot server, this is a convenient way to get the file.

MGF peak list is useful when you have the search result but can't find the peak list.

Search result export

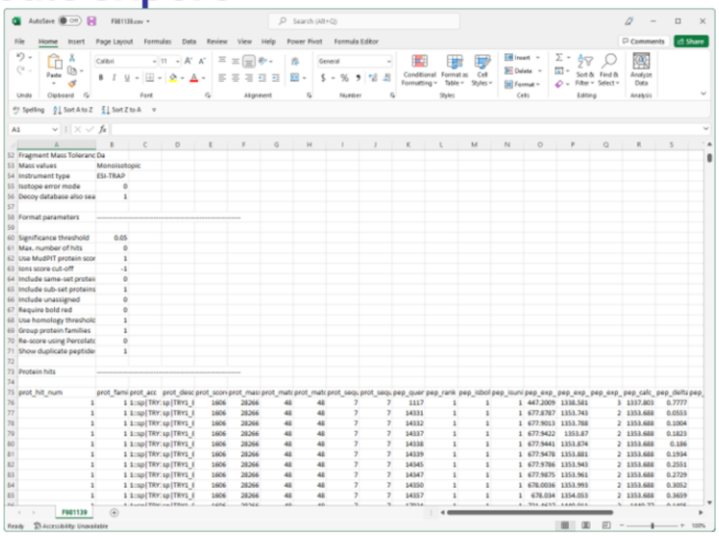


If you arrive here from one of the older reports, to begin with, you may need to select the required output format. Different formats have different options further down the page

Search result export

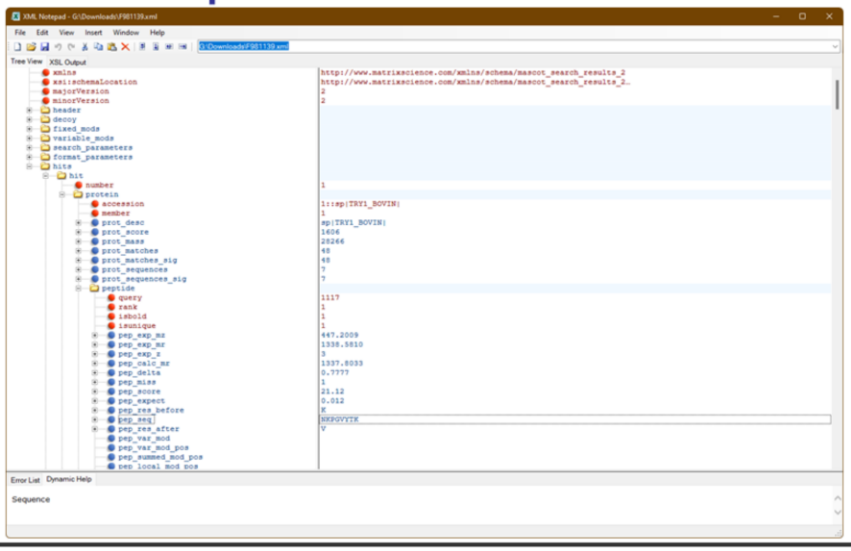
To export to Excel, simply select CSV as the format, and click on the Export Search Results button at the bottom of the page. In recent versions of Mascot, the report is prepared and then a download button is displayed. In older versions, the download would start immediately. One the download is finished, you can open it into Excel:

Search result export



Much easier and safer than “screen scraping”

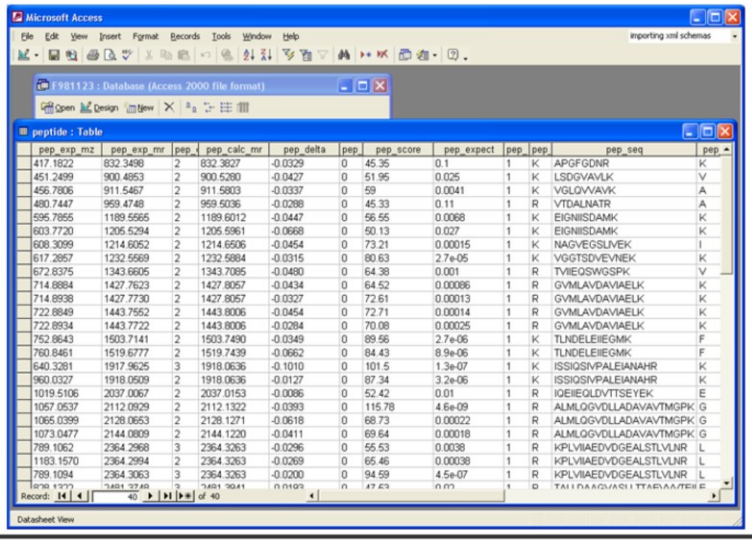
Search result export



For those of you into XML, here is a sample XML file. The schema is available from our web site or your local Mascot installation.

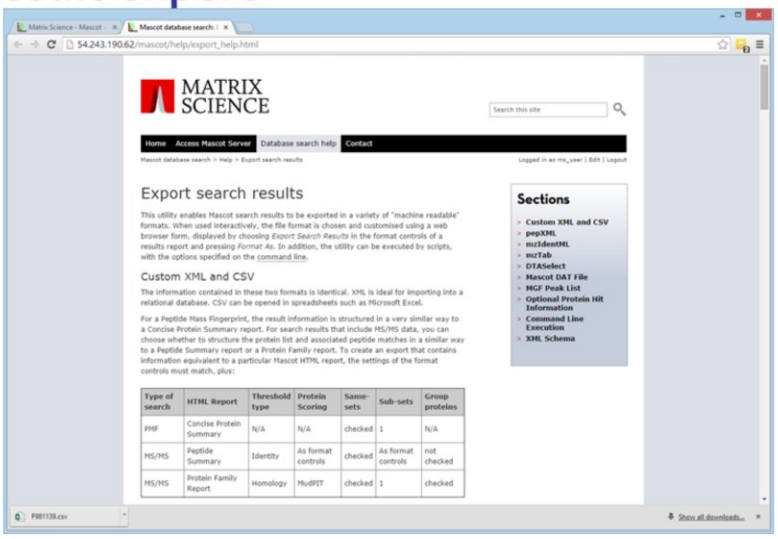
Please read the help for details.

Search result export



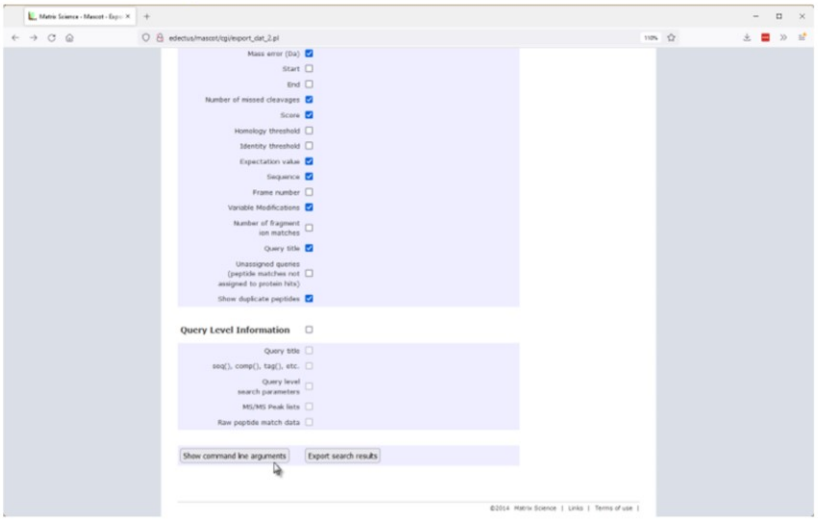
XML is ideal for transferring the results to a relational database. Even Microsoft Access can open the XML file directly into database tables

Search result export



There is a very detailed help page for all of this.

Search result export

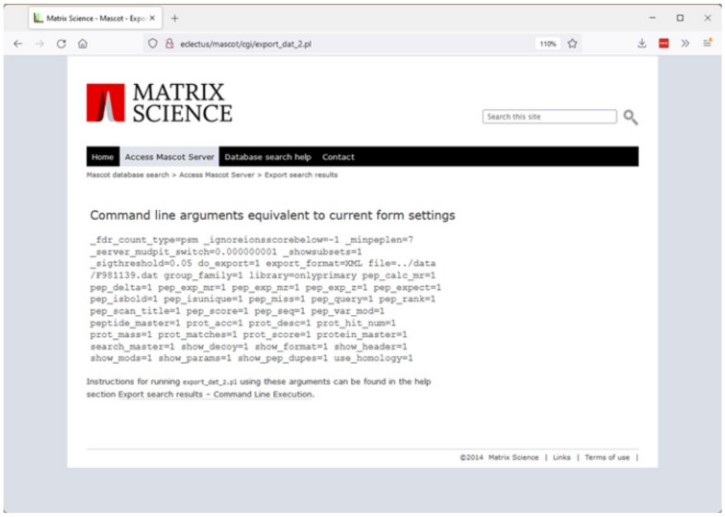


Which describes how the export script can be called from the command line or a shell prompt, as part of an automated pipeline.

I won't go into any detail here, but this means that it is possible to set up a script that will, for example, automatically convert all of your Mascot results to XML files.

Figuring out the command line arguments from the help can be tricky so, in Mascot 2.3, we added a function to display the command line corresponding to the selected options

Search result export



By the way, don't delete the original result files after exporting them or your won't be able to view the standard Mascot reports in a browser.