Very Large Searches

MASCOT

MATRIX SCIENCE

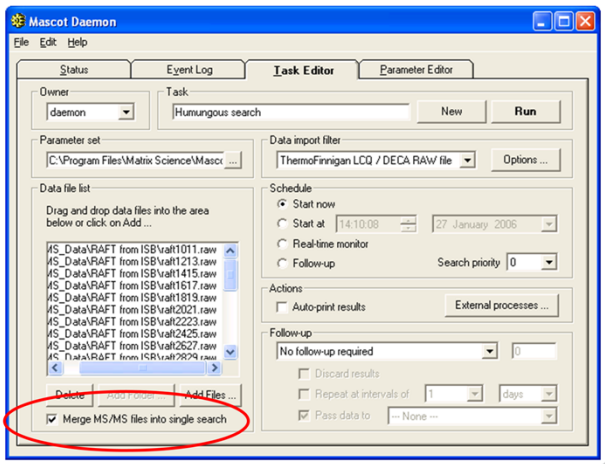
Topics

- 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 MudPIT fractions
- Use Distiller or a file specific data import filter



MASCOT : *Very Large Searches*

MATRIX SCIENCE

The smartest way to merge files, like fractions from a MudPIT 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, although note that some web servers cannot accept uploads larger than 4 GB

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
```

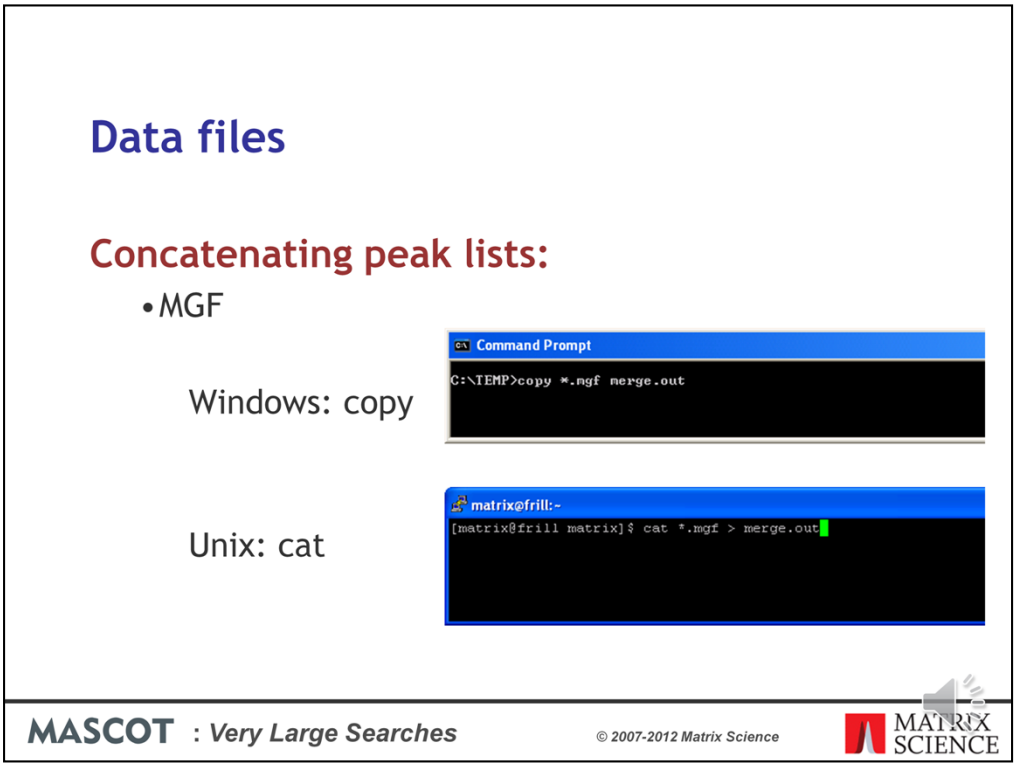
MASCOT : *Very Large Searches*

MATRIX SCIENCE

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



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!

MASCOT : *Very Large Searches* MATRIX SCIENCE

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

MASCOT : *Very Large Searches*

MATRIX SCIENCE

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

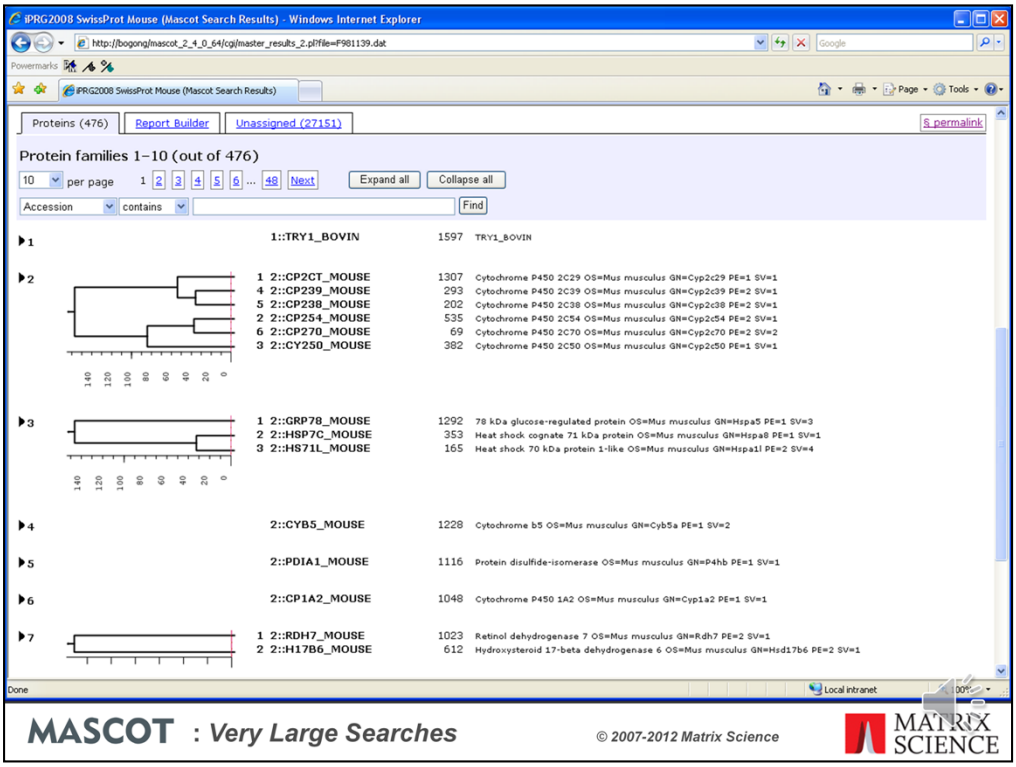
Reporting large searches

Protein Family Summary (new in Mascot 2.3)

- 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

MASCOT : *Very Large Searches* MATRIX SCIENCE

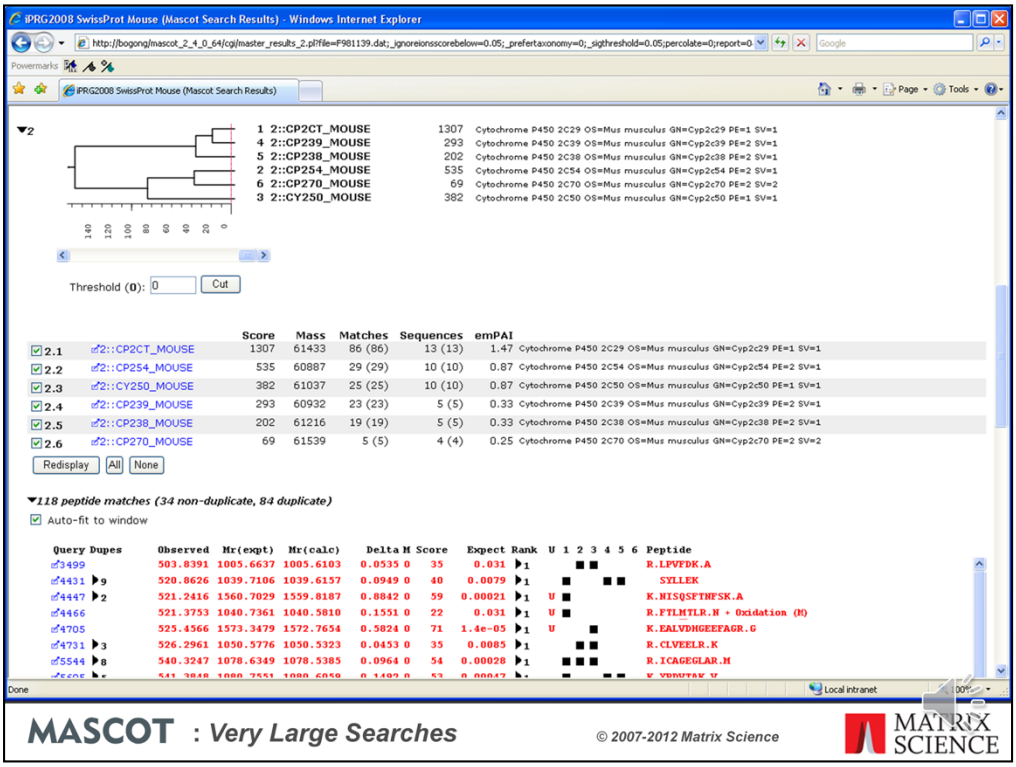
In Mascot 2.2 and earlier, trying to display result reports for very large searches would often lead to problems with timeouts and running out of memory. To address this, the Protein Family Summary loads most of the information ‘on demand’. This requires some index files to be created on the server, 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 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.

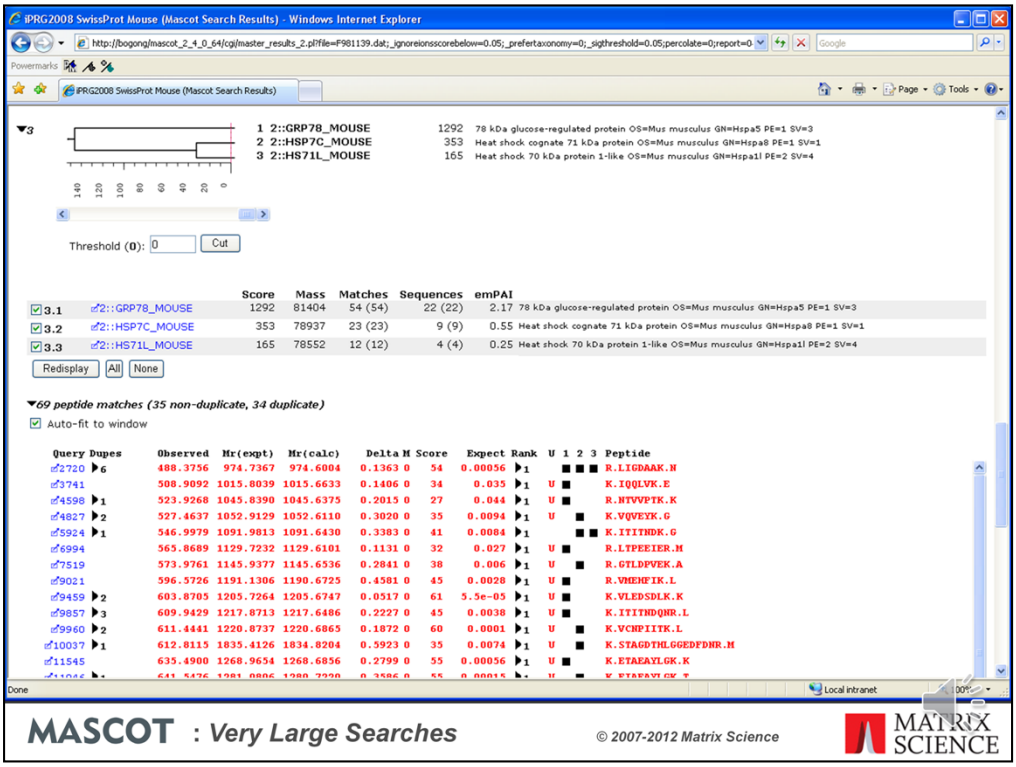
Proteins are grouped into families using a novel hierarchical clustering algorithm. If the family contains a single member, the accession string, protein score and description are listed. If the family contains multiple members, the accessions, scores and descriptions are aligned with a dendrogram, which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

The scores for the proteins in family 2 vary from 1307 down to 69. 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 2C proteins.



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. 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.

In this screen shot and the ones that follow, we've set an expect cut-off of 0.05 to simplify the picture by removing low scoring matches

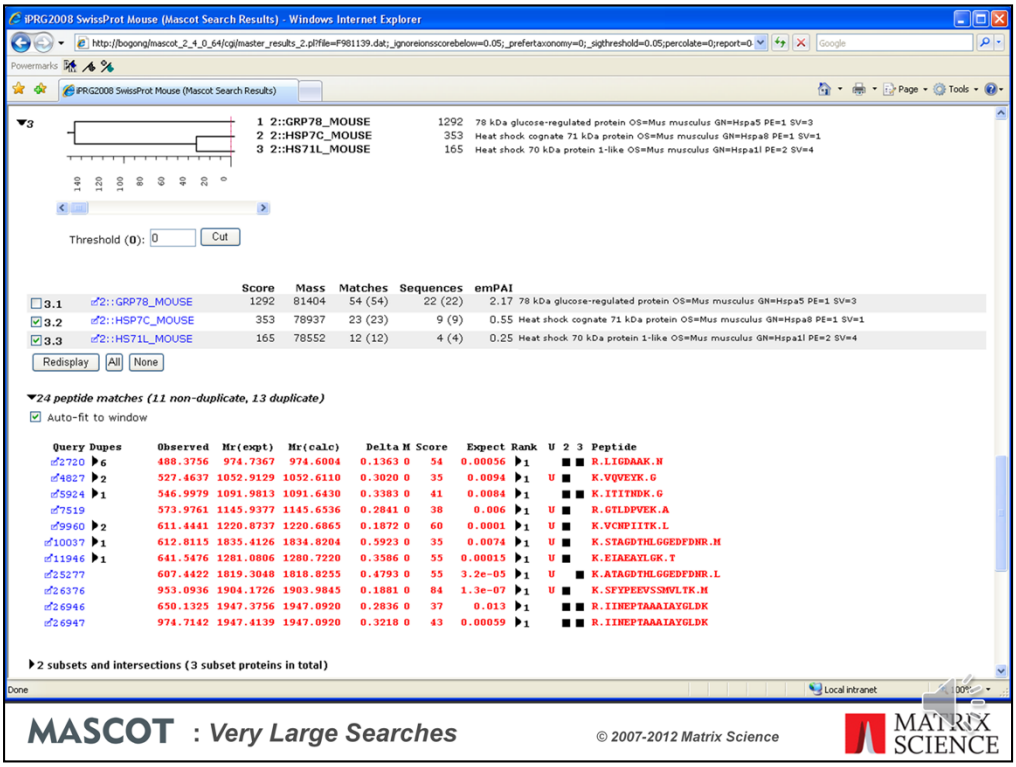


Moving down to family 3, the scale on the dendrogram is ions score, and HSP7C_MOUSE and HS71L_MOUSE join at a score of approximately 30. This represents the score of the significant matches that would have to be discarded in order to make one protein a sub-set of the other. These two proteins are much 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.

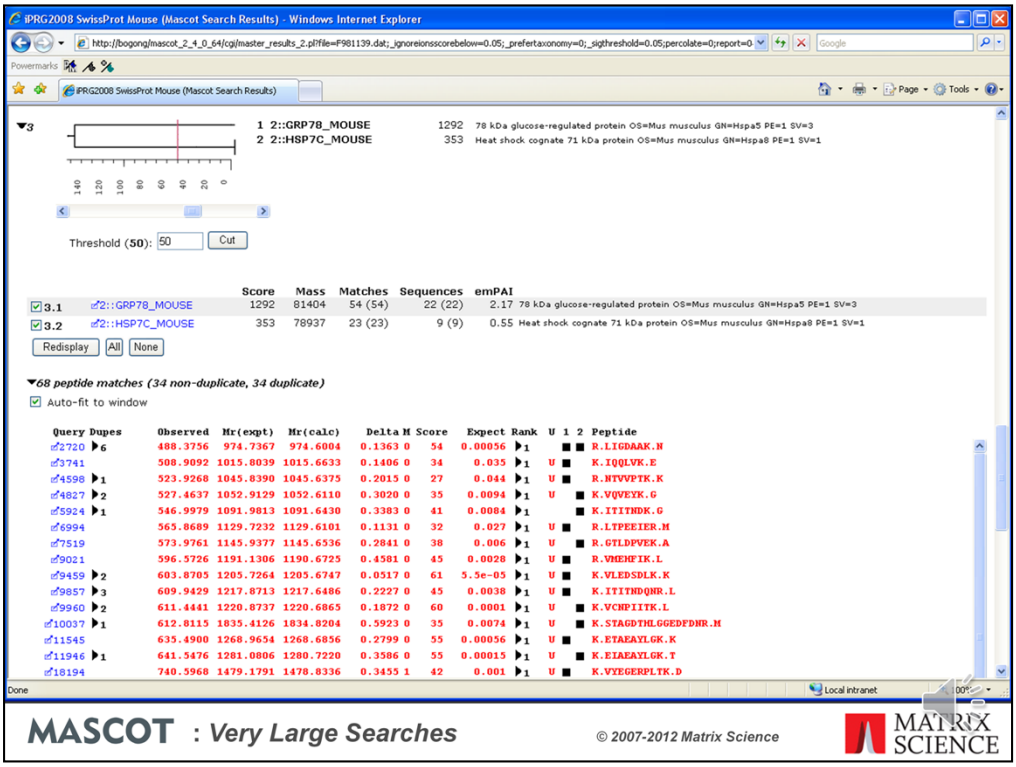
The table of peptide matches is similar to that found in the other result reports. Duplicate matches to the same sequence are collapsed into a single row. Click on the triangle to expand.

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.

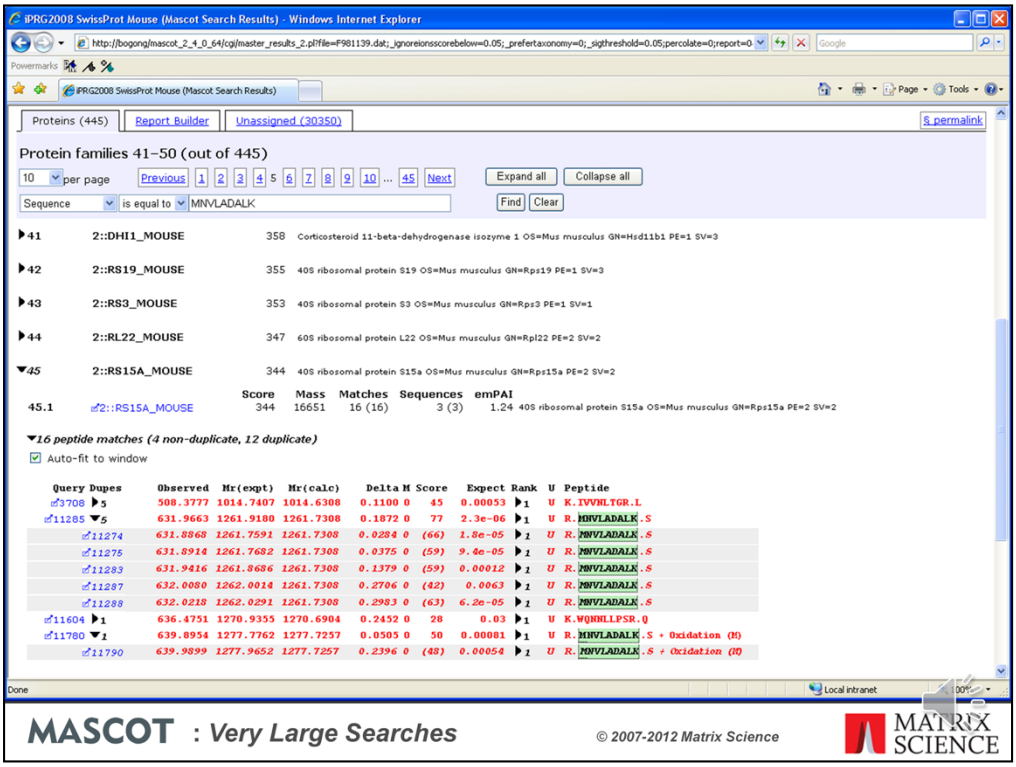


It can now be seen that HS71L_MOUSE would be a sub-set of HSP7C_MOUSE if it was not for one match, K.ATAGDTHLGGEDFDNR.L. It is the significant score for this match that separates the two proteins in the dendrogram by a distance of 32 (score of 55 - homology threshold score of 23).

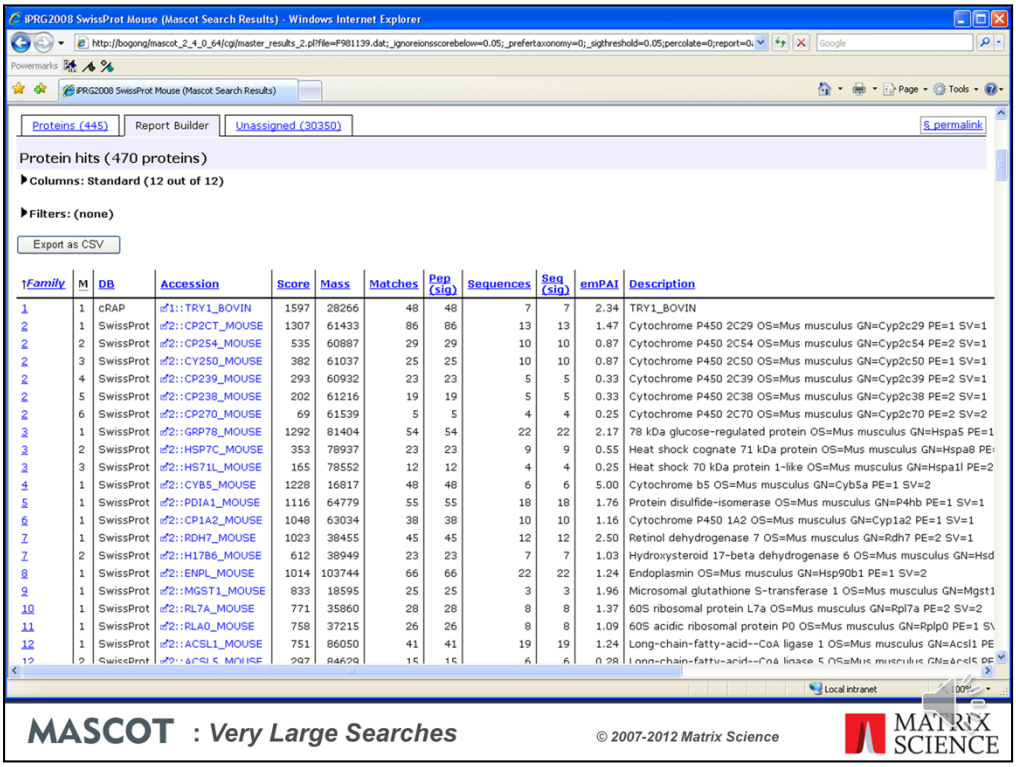
You can "cut" the dendrogram using the slider control.



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.

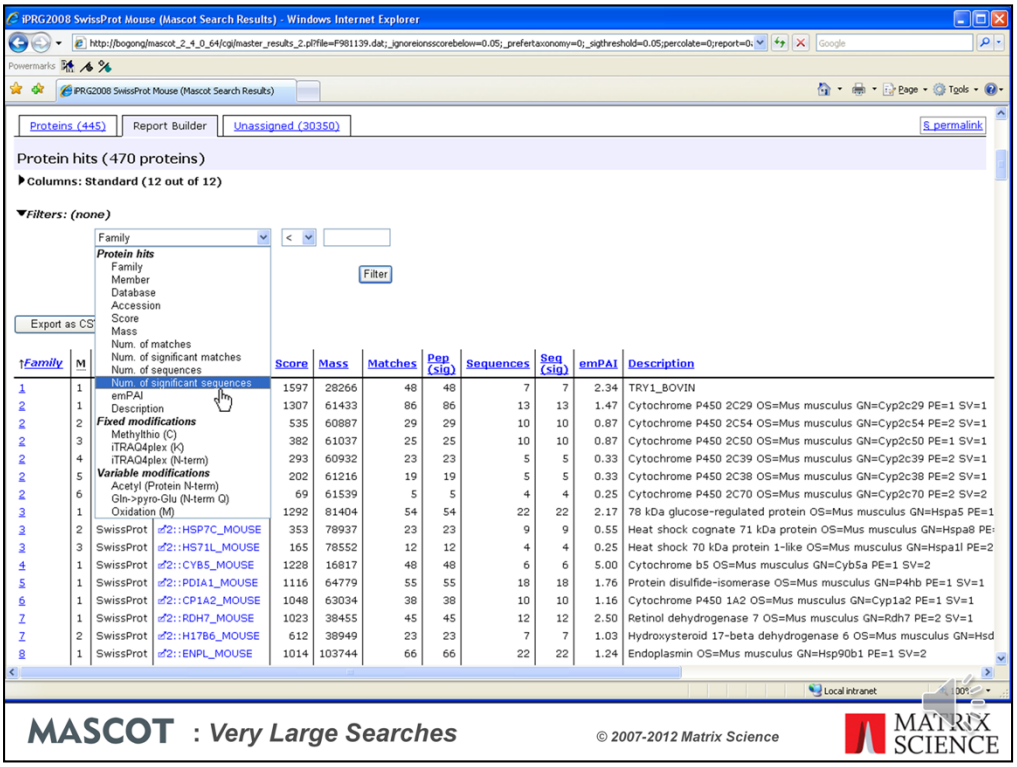


The family report also includes a text search facility, which is particularly important for a paged report. You can search by accession or description sub-string, or by query, mass or sequence. Here, for example, we searched for a peptide sequence. The display jumps to the first instance of the sequence,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

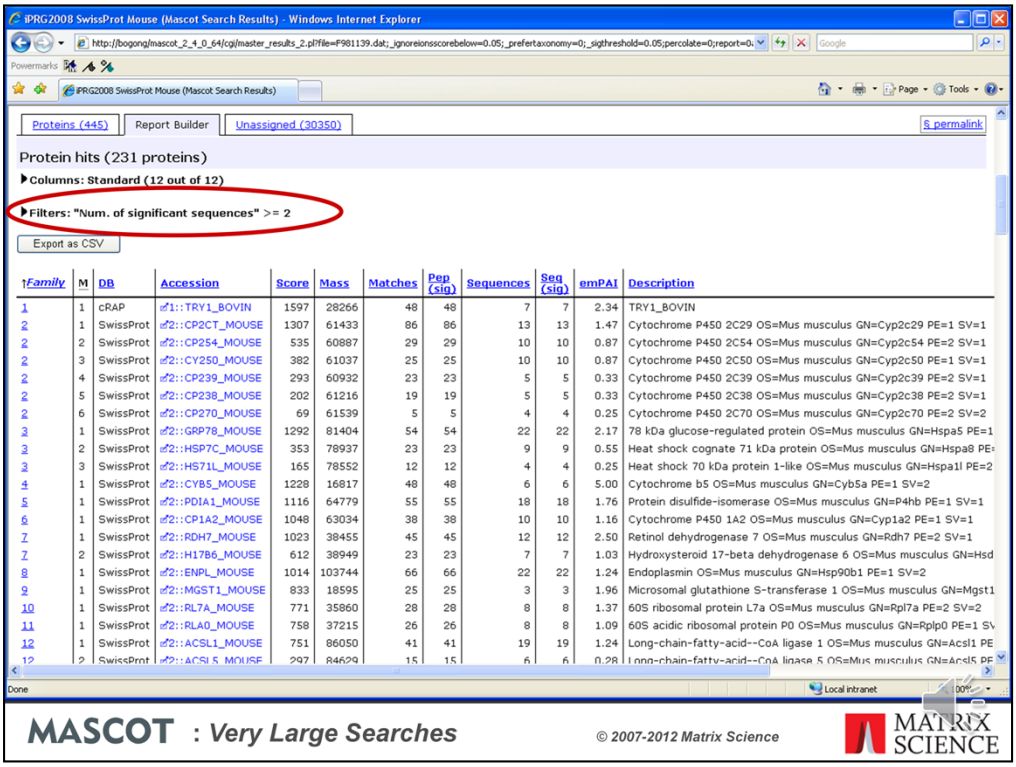


Family	M	DB	Accession	Score	Mass	Matches	Pep (sig)	Sequences	Seq (sig)	emPAI	Description
1	1	cRAP	1::TRY1_BOVIN	1597	28266	48	48	7	7	2.34	TRY1_BOVIN
2	1	SwissProt	2::CP2CT_MOUSE	1307	61433	86	86	13	13	1.47	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=1
2	2	SwissProt	2::CP254_MOUSE	535	60887	29	29	10	10	0.87	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
2	3	SwissProt	2::CY250_MOUSE	382	61037	25	25	10	10	0.87	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=1
2	4	SwissProt	2::CP239_MOUSE	293	60932	23	23	5	5	0.33	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=1
2	5	SwissProt	2::CP238_MOUSE	202	61216	19	19	5	5	0.33	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=1
2	6	SwissProt	2::CP270_MOUSE	69	61539	5	5	4	4	0.25	Cytochrome P450 2C70 OS=Mus musculus GN=Cyp2c70 PE=2 SV=2
3	1	SwissProt	2::GRP78_MOUSE	1292	81404	54	54	22	22	2.17	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1
3	2	SwissProt	2::HSP7C_MOUSE	353	78937	23	23	9	9	0.55	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE
3	3	SwissProt	2::HS71L_MOUSE	165	78552	12	12	4	4	0.25	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2
4	1	SwissProt	2::CYB5_MOUSE	1228	16817	48	48	6	6	5.00	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	2::PDIA1_MOUSE	1116	64779	55	55	18	18	1.76	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV=1
6	1	SwissProt	2::CP1A2_MOUSE	1048	63034	38	38	10	10	1.16	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
7	1	SwissProt	2::RDH7_MOUSE	1023	38455	45	45	12	12	2.50	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
7	2	SwissProt	2::H17B6_MOUSE	612	38949	23	23	7	7	1.03	Hydroxysteroid 17-beta dehydrogenase 6 OS=Mus musculus GN=Hsd
8	1	SwissProt	2::ENPL_MOUSE	1014	103744	66	66	22	22	1.24	Endoplasmin OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
9	1	SwissProt	2::MGST1_MOUSE	833	18595	25	25	3	3	1.96	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgst1
10	1	SwissProt	2::RL7A_MOUSE	771	35860	28	28	8	8	1.37	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=2
11	1	SwissProt	2::RLA0_MOUSE	758	37215	26	26	8	8	1.09	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 PE=1 SV
12	1	SwissProt	2::ACSL1_MOUSE	751	86050	41	41	19	19	1.24	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acsl1 PE
12	2	SwissProt	2::ACSL5_MOUSE	297	84629	15	15	6	6	0.28	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acsl5 PE

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

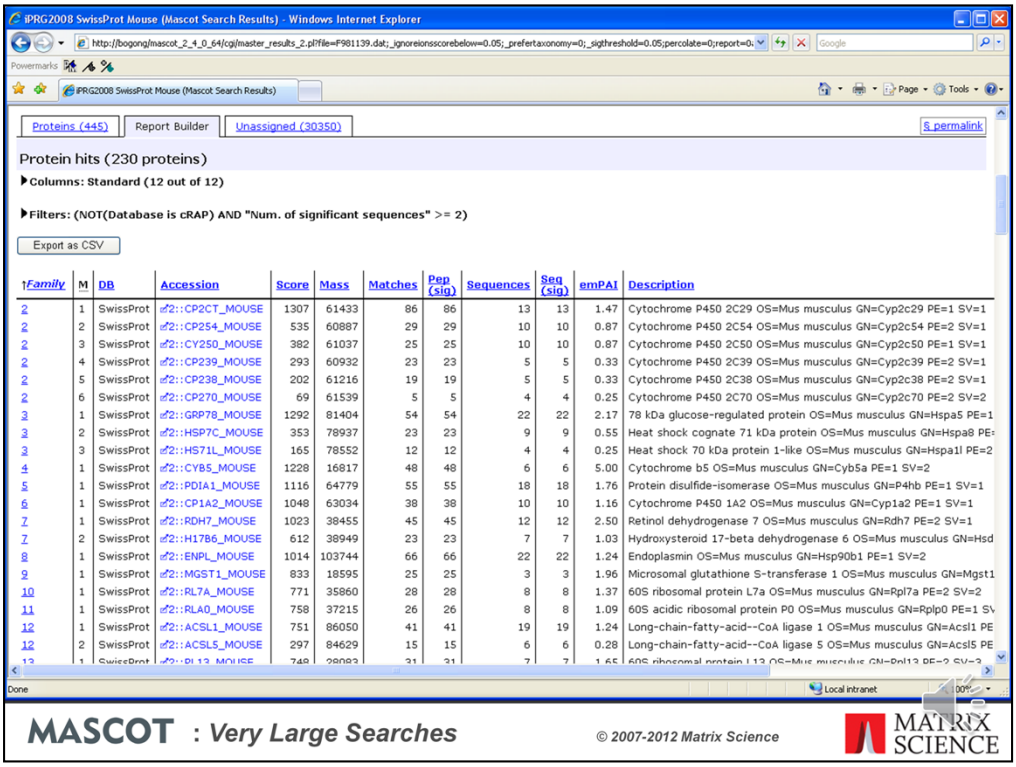


We open up the filters section and add a suitable filter.



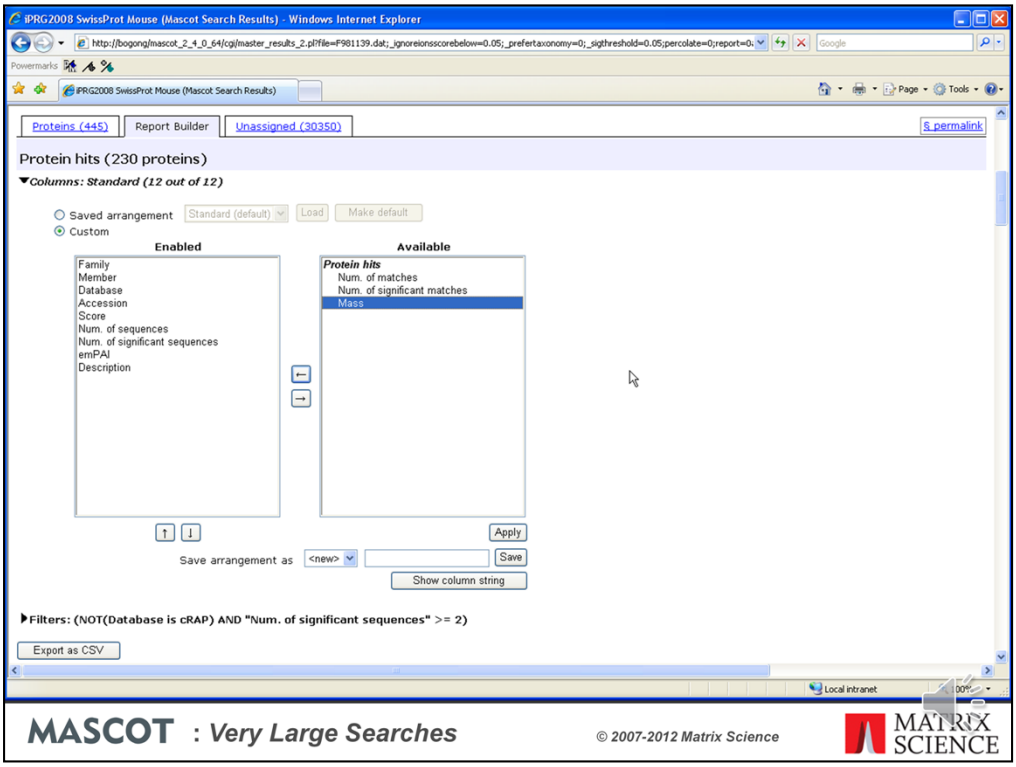
Family	M	DB	Accession	Score	Mass	Matches	Pep (sig)	Sequences	Seq (sig)	emPAI	Description
1	1	cRAP	1::TRY1_BOVIN	1597	28266	48	48	7	7	2.34	TRY1_BOVIN
2	1	SwissProt	2::CP2CT_MOUSE	1307	61433	86	86	13	13	1.47	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=1
2	2	SwissProt	2::CP254_MOUSE	535	60887	29	29	10	10	0.87	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
2	3	SwissProt	2::CY250_MOUSE	382	61037	25	25	10	10	0.87	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=1
2	4	SwissProt	2::CP239_MOUSE	293	60932	23	23	5	5	0.33	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=1
2	5	SwissProt	2::CP238_MOUSE	202	61216	19	19	5	5	0.33	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=1
2	6	SwissProt	2::CP270_MOUSE	69	61539	5	5	4	4	0.25	Cytochrome P450 2C70 OS=Mus musculus GN=Cyp2c70 PE=2 SV=2
3	1	SwissProt	2::GRP78_MOUSE	1292	81404	54	54	22	22	2.17	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1
3	2	SwissProt	2::HSP7C_MOUSE	353	78937	23	23	9	9	0.55	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=
3	3	SwissProt	2::HS71L_MOUSE	165	78552	12	12	4	4	0.25	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2
4	1	SwissProt	2::CYB5_MOUSE	1228	16817	48	48	6	6	5.00	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	2::PDIA1_MOUSE	1116	64779	55	55	18	18	1.76	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV=1
6	1	SwissProt	2::CP1A2_MOUSE	1048	63034	38	38	10	10	1.16	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
7	1	SwissProt	2::RDH7_MOUSE	1023	38455	45	45	12	12	2.50	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
7	2	SwissProt	2::H17B6_MOUSE	612	38949	23	23	7	7	1.03	Hydroxysteroid 17-beta dehydrogenase 6 OS=Mus musculus GN=Hsd
8	1	SwissProt	2::ENPL_MOUSE	1014	103744	66	66	22	22	1.24	Endoplasmin OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
9	1	SwissProt	2::MGST1_MOUSE	833	18595	25	25	3	3	1.96	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgst1
10	1	SwissProt	2::RL7A_MOUSE	771	35860	28	28	8	8	1.37	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=2
11	1	SwissProt	2::RLA0_MOUSE	758	37215	26	26	8	8	1.09	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 PE=1 SV
12	1	SwissProt	2::ACSL1_MOUSE	751	86050	41	41	19	19	1.24	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acsl1 PE
12	2	SwissProt	2::ACSL5_MOUSE	297	84629	15	15	6	6	0.28	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acsl5 PE

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

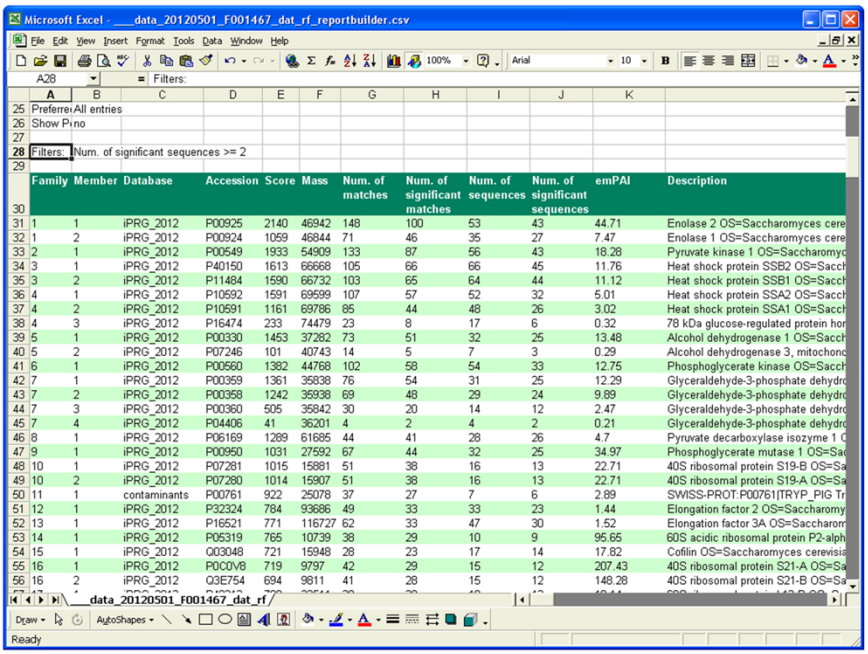


Family	M	DB	Accession	Score	Mass	Matches	Pep (sig)	Sequences	Seq (sig)	emPAI	Description
2	1	SwissProt	2::CP2CT_MOUSE	1307	61433	86	86	13	13	1.47	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=1
2	2	SwissProt	2::CP2S4_MOUSE	535	60887	29	29	10	10	0.87	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
2	3	SwissProt	2::CY2S0_MOUSE	382	61037	25	25	10	10	0.87	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=1
2	4	SwissProt	2::CP239_MOUSE	293	60932	23	23	5	5	0.33	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=1
2	5	SwissProt	2::CP238_MOUSE	202	61216	19	19	5	5	0.33	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=1
2	6	SwissProt	2::CP270_MOUSE	69	61539	5	5	4	4	0.25	Cytochrome P450 2C70 OS=Mus musculus GN=Cyp2c70 PE=2 SV=2
3	1	SwissProt	2::GRP78_MOUSE	1292	81404	54	54	22	22	2.17	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1
3	2	SwissProt	2::HSP7C_MOUSE	353	78937	23	23	9	9	0.55	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=
3	3	SwissProt	2::HS71L_MOUSE	165	78552	12	12	4	4	0.25	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2
4	1	SwissProt	2::CYB5_MOUSE	1228	16817	48	48	6	6	5.00	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	2::PDIA1_MOUSE	1116	64779	55	55	18	18	1.76	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV=1
6	1	SwissProt	2::CP1A2_MOUSE	1048	63034	38	38	10	10	1.16	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
7	1	SwissProt	2::RDH7_MOUSE	1023	38455	45	45	12	12	2.50	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
7	2	SwissProt	2::H17B6_MOUSE	612	38949	23	23	7	7	1.03	Hydroxysteroid 17-beta dehydrogenase 6 OS=Mus musculus GN=Hsd
8	1	SwissProt	2::ENPL_MOUSE	1014	103744	66	66	22	22	1.24	Endoplasmin OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
9	1	SwissProt	2::MGST1_MOUSE	833	18595	25	25	3	3	1.96	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgst1
10	1	SwissProt	2::RL7A_MOUSE	771	35860	28	28	8	8	1.37	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=2
11	1	SwissProt	2::RLA0_MOUSE	758	37215	26	26	8	8	1.09	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 PE=1 SV
12	1	SwissProt	2::ACSL1_MOUSE	751	86050	41	41	19	19	1.24	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acsl1 PE
12	2	SwissProt	2::ACSL5_MOUSE	297	84629	15	15	6	6	0.28	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acsl5 PE

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

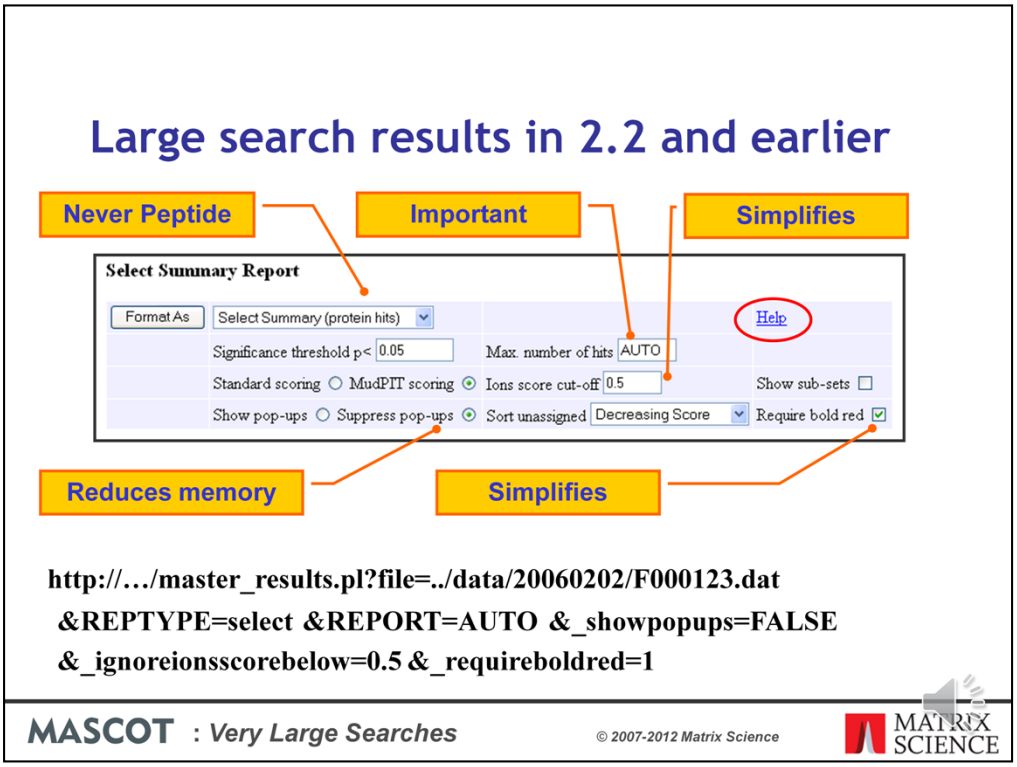


Family	Member	Database	Accession	Score	Mass	Num. of matches	Num. of significant matches	Num. of sequences	Num. of significant sequences	emPAI	Description
1	1	iPRG_2012	P00925	2140	46942	148	100	53	43	44.71	Enolase 2 OS=Saccharomyces cere
1	2	iPRG_2012	P00924	1059	46844	71	46	35	27	7.47	Enolase 1 OS=Saccharomyces cere
2	1	iPRG_2012	P00549	1933	54909	133	87	56	43	18.28	Pyruvate kinase 1 OS=Saccharomyc
3	1	iPRG_2012	P40150	1613	66668	105	66	66	45	11.76	Heat shock protein SSB2 OS=Sacch
3	2	iPRG_2012	P11484	1590	66732	103	65	64	44	11.12	Heat shock protein SSB1 OS=Sacch
4	1	iPRG_2012	P10592	1591	69599	107	57	52	32	5.01	Heat shock protein SSA2 OS=Sacch
4	2	iPRG_2012	P10591	1161	69786	85	44	48	26	3.02	Heat shock protein SSA1 OS=Sacch
4	3	iPRG_2012	P16474	233	74479	23	8	17	6	0.32	78 kDa glucose-regulated protein hor
5	1	iPRG_2012	P00330	1453	37282	73	51	32	25	13.48	Alcohol dehydrogenase 1 OS=Sacch
5	2	iPRG_2012	P07246	101	40743	14	5	7	3	0.29	Alcohol dehydrogenase 3, mitochonc
6	1	iPRG_2012	P00560	1382	44768	102	58	54	33	12.75	Phosphoglycerate kinase OS=Sacch
7	1	iPRG_2012	P00359	1361	35838	76	54	31	25	12.29	Glyceraldehyde-3-phosphate dehydro
7	2	iPRG_2012	P00358	1242	35938	69	48	29	24	9.89	Glyceraldehyde-3-phosphate dehydro
7	3	iPRG_2012	P00360	505	35842	30	20	14	12	2.47	Glyceraldehyde-3-phosphate dehydro
7	4	iPRG_2012	P04406	41	36201	4	2	4	2	0.21	Glyceraldehyde-3-phosphate dehydro
8	1	iPRG_2012	P06169	1289	61685	44	41	28	26	4.7	Pyruvate decarboxylase isozyme 1 C
9	1	iPRG_2012	P00950	1031	27592	67	44	32	25	34.97	Phosphoglycerate mutase 1 OS=Sac
10	1	iPRG_2012	P07281	1015	15881	51	38	16	13	22.71	40S ribosomal protein S19-B OS=Sa
10	2	iPRG_2012	P07280	1014	15907	51	38	16	13	22.71	40S ribosomal protein S19-A OS=Sa
11	1	contaminants	P00761	922	25078	37	27	7	6	2.89	SWISS-PROT:P00761|TRYP_PIG Tr
12	1	iPRG_2012	P32324	784	93686	49	33	33	23	1.44	Elongation factor 2 OS=Saccharomy
13	1	iPRG_2012	P16521	771	116727	62	33	47	30	1.52	Elongation factor 3A OS=Sacchario
14	1	iPRG_2012	P05319	765	10739	38	29	10	9	95.65	60S acidic ribosomal protein P2-alph
15	1	iPRG_2012	Q03048	721	15948	28	23	17	14	17.82	Cofilin OS=Saccharomyces cerevisia
16	1	iPRG_2012	P0C0V8	719	9797	42	29	15	12	207.43	40S ribosomal protein S21-A OS=Sa
16	2	iPRG_2012	Q3E754	694	9811	41	28	15	12	148.28	40S ribosomal protein S21-B OS=Sa

MASCOT : *Very Large Searches*

MATRIX SCIENCE

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



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:

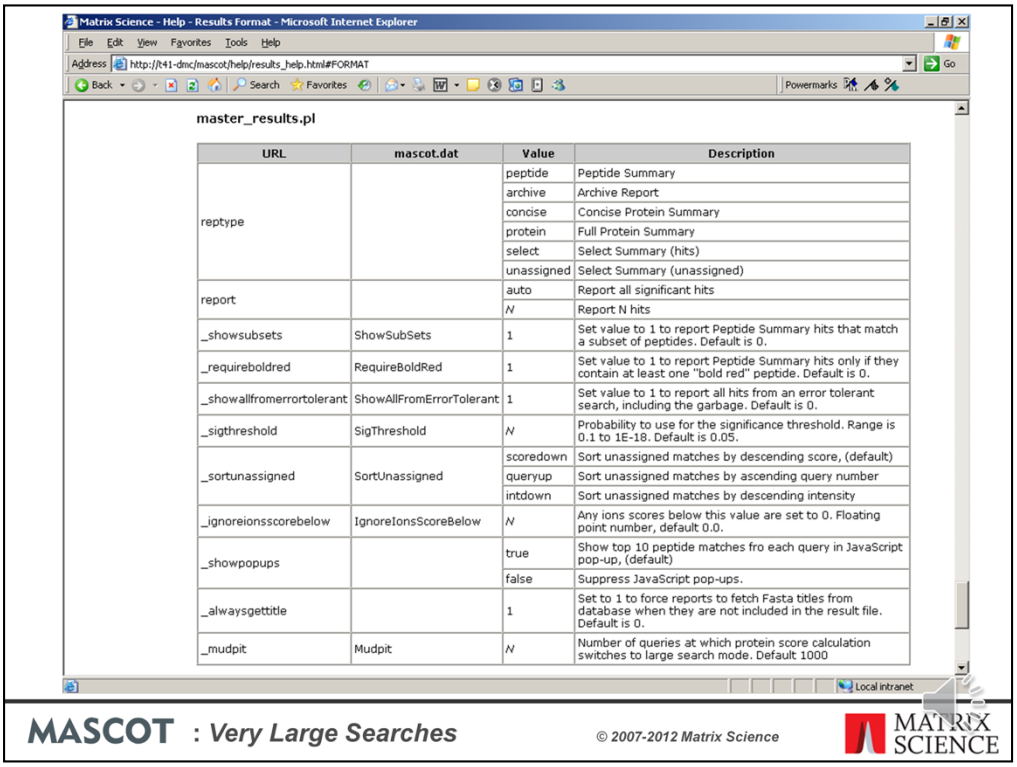
•Ensure you are using the Select report. If you are using a third party client that has specified Peptide summary or Protein summary, add this to the URL before opening the file: &REPTYPE=select

•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 an expect value cut-off will minimise the number of hits: &_ignoreionsscorebelow=0.5&_requireboldred=1

Note that the ions score cut-off is as score threshold when the value is 1 or greater. When the value is between 0 and 1, it is an expect threshold, which is often much more useful. I often set this to 0.5 to get rid of all the junk matches.



Matrix Science - Help - Results Format - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://lt41-dmc/mascot/help/results_help.html#FORMAT

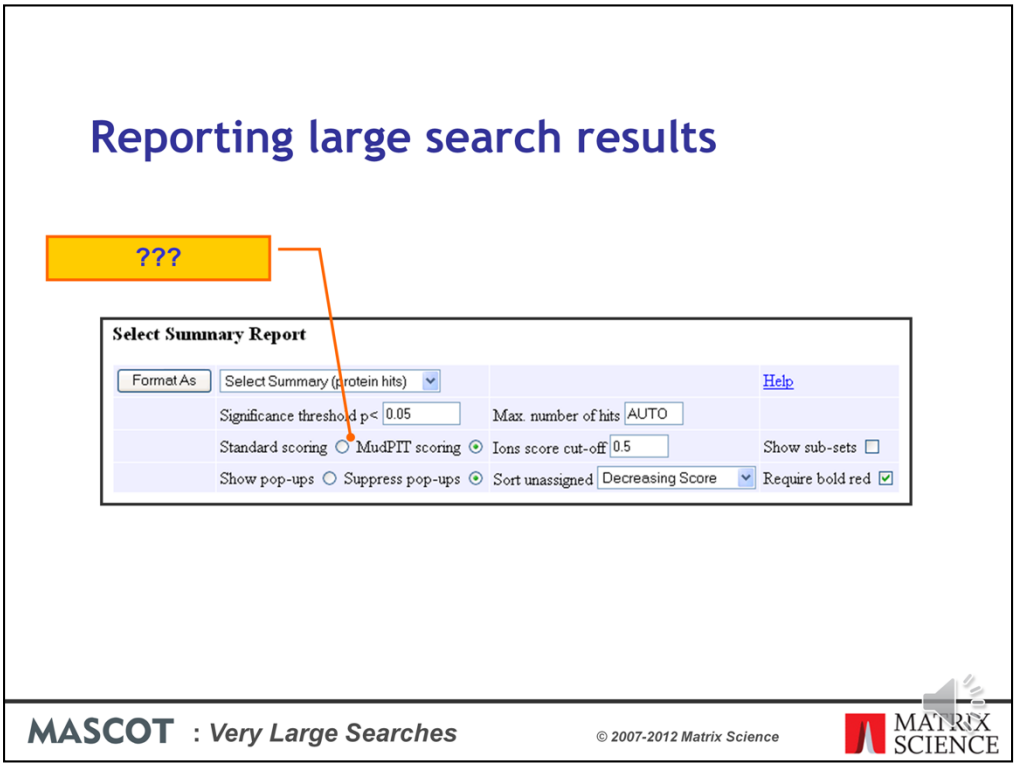
master_results.pl

URL	mascot.dat	Value	Description
reptype		peptide	Peptide Summary
		archive	Archive Report
		concise	Concise Protein Summary
		protein	Full Protein Summary
		select	Select Summary (hits)
		unassigned	Select Summary (unassigned)
report		auto	Report all significant hits
		N	Report N hits
_showsubsets	ShowSubSets	1	Set value to 1 to report Peptide Summary hits that match a subset of peptides. Default is 0.
_requireboldred	RequireBoldRed	1	Set value to 1 to report Peptide Summary hits only if they contain at least one "bold red" peptide. Default is 0.
_showallfromerrortolerant	ShowAllFromErrorTolerant	1	Set value to 1 to report all hits from an error tolerant search, including the garbage. Default is 0.
_sigthreshold	SigThreshold	*N*	Probability to use for the significance threshold. Range is 0.1 to 1E-18. Default is 0.05.
_sortunassigned	SortUnassigned	scoredown	Sort unassigned matches by descending score, (default)
		queryup	Sort unassigned matches by ascending query number
		intdown	Sort unassigned matches by descending intensity
_ignoreionsscorebelow	IgnoreIonsScoreBelow	*N*	Any ions scores below this value are set to 0. Floating point number, default 0.0.
_showpopups		true	Show top 10 peptide matches fro each query in JavaScript pop-up, (default)
		false	Suppress JavaScript pop-ups.
_alwaysgettitle		1	Set to 1 to force reports to fetch Fasta titles from database when they are not included in the result file. Default is 0.
_mudpit	Mudpit	*N*	Number of queries at which protein score calculation switches to large search mode. Default 1000

Local intranet

MASCOT : *Very Large Searches* © 2007-2012 Matrix Science MATRIX SCIENCE

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



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

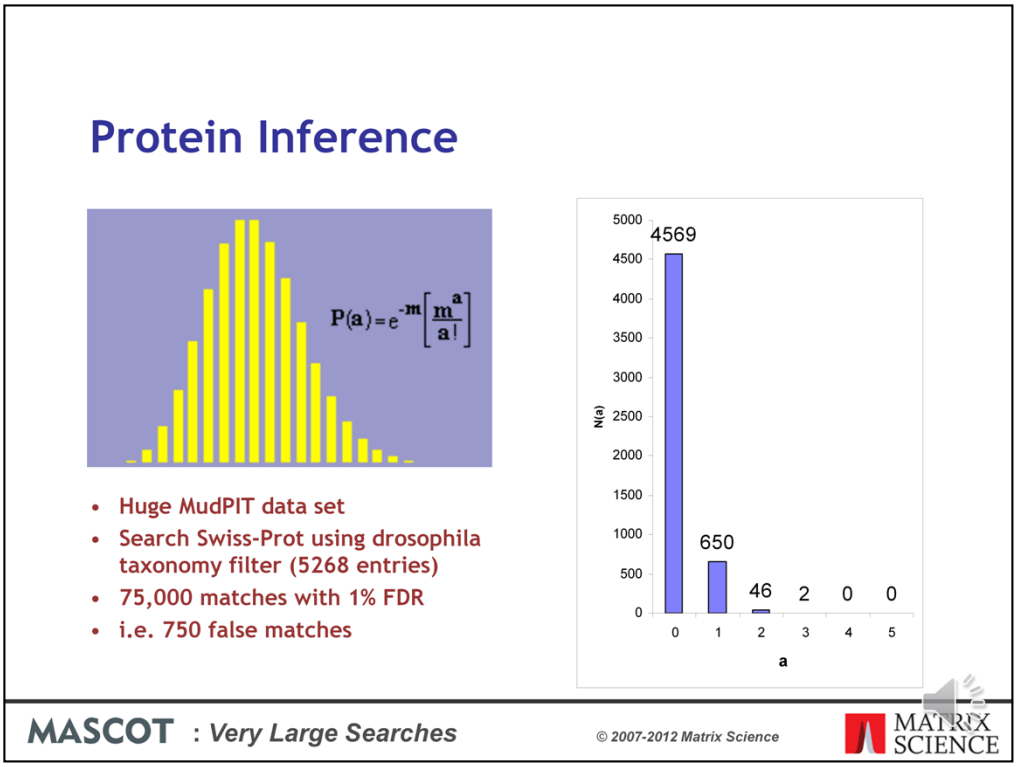
Query	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Score	Expect	Rank	Unique	Peptide
28	359.7341	717.4537	717.4537	-0.09	0	7	4.2	5	U	R.LFAIVR.G
209	394.2371	786.4596	786.4599	-0.46	0	8	13	3	U	K.LTIADVR.A
334	411.2073	820.4000	820.3954	5.61	0	3	15	4	U	K.TDSGLYR.C
357	413.2642	824.5139	824.5135	0.48	1	12	1.1	5	U	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

Protein Scores for MS/MS Searches

MudPIT protein score

- The sum of the excess of the ions score over the identity or homology threshold for each query
- Plus 1 x the average threshold

1249. 2::IPI00023283 **Mass:** 3832803 **Score:** 0 **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.TDSGLYR.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**

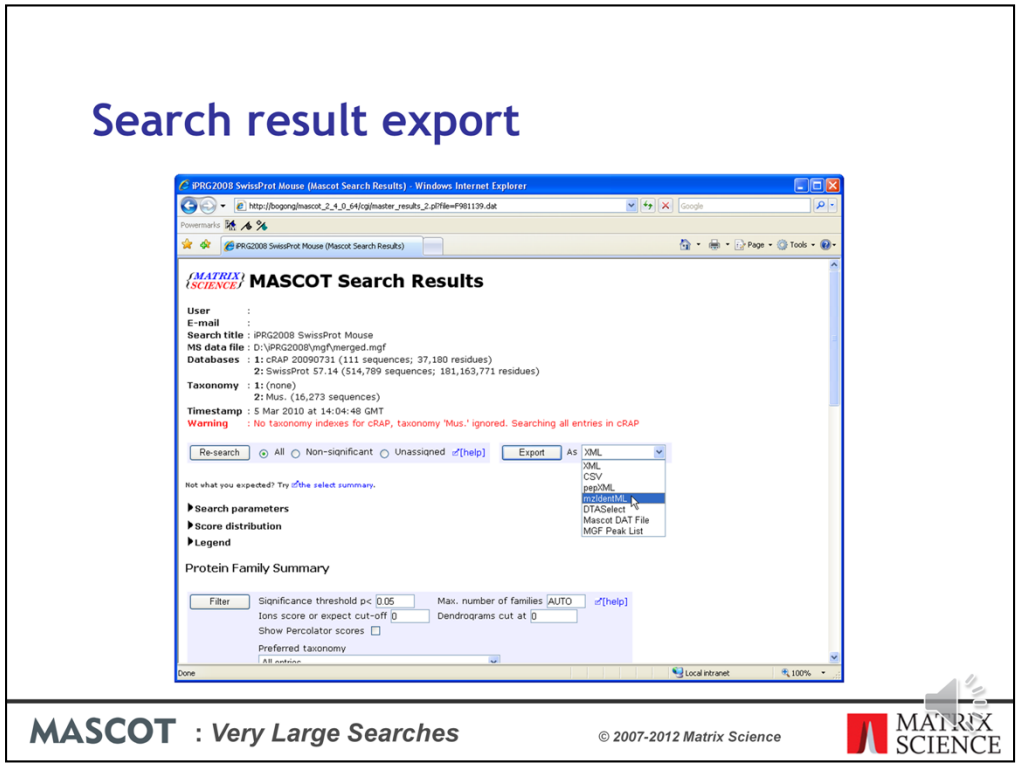
MASCOT : *Very Large Searches*

MATRIX SCIENCE

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.

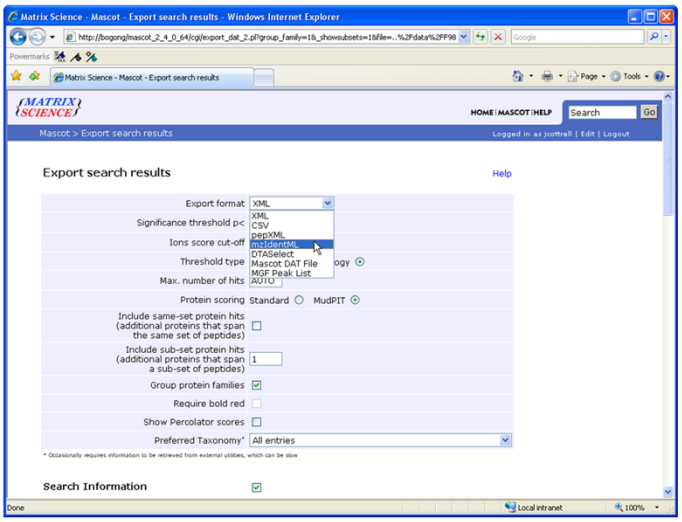
mzIdentML is the new, standard format from PSI for search result interchange. Mascot provides a very full implementation of mzIdentML and this is the one to choose if you are writing new application software that will use Mascot results

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.

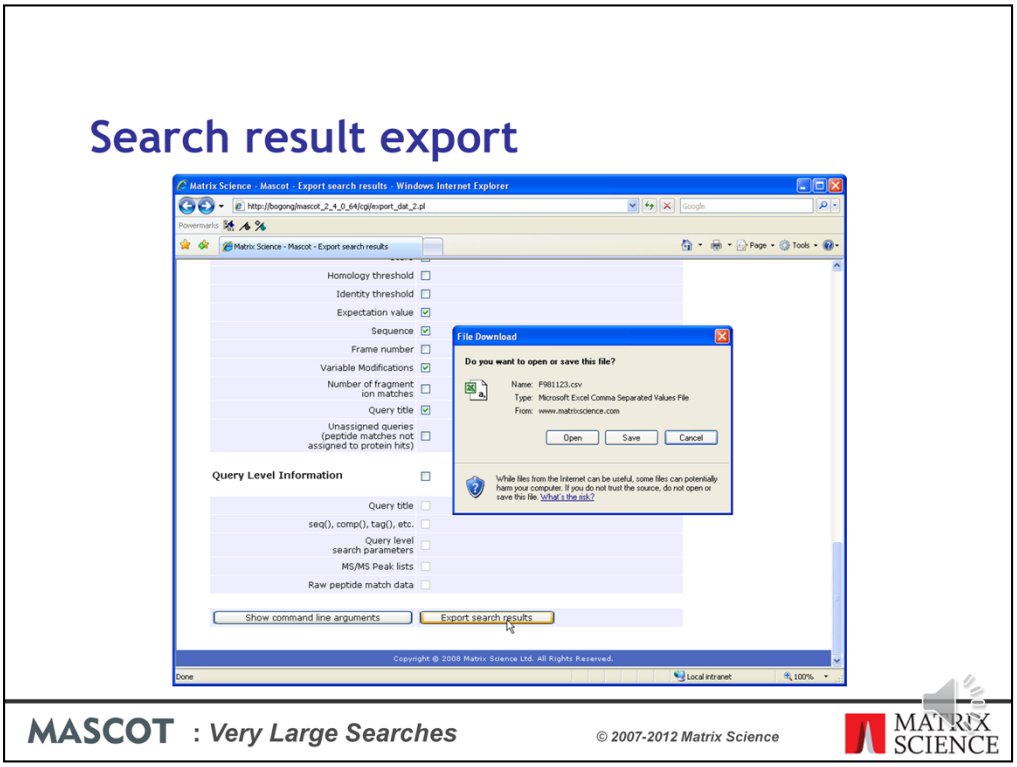
MGF peak list is useful when you have the search result but can't find the peak list.

Search result export



MASCOT : *Very Large Searches*

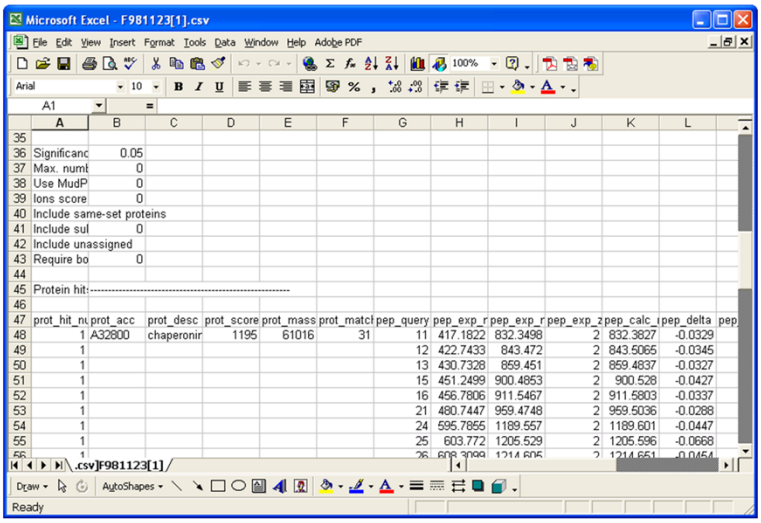
If you arrive here from one of the older reports, to begin with, you may need to select the required output format. Different formats have different options further down the page



To export to Excel, simply select CSV as the format, and click on the Export Search Results button at the bottom of the page.

You can then click on the Open button to open it into Excel:

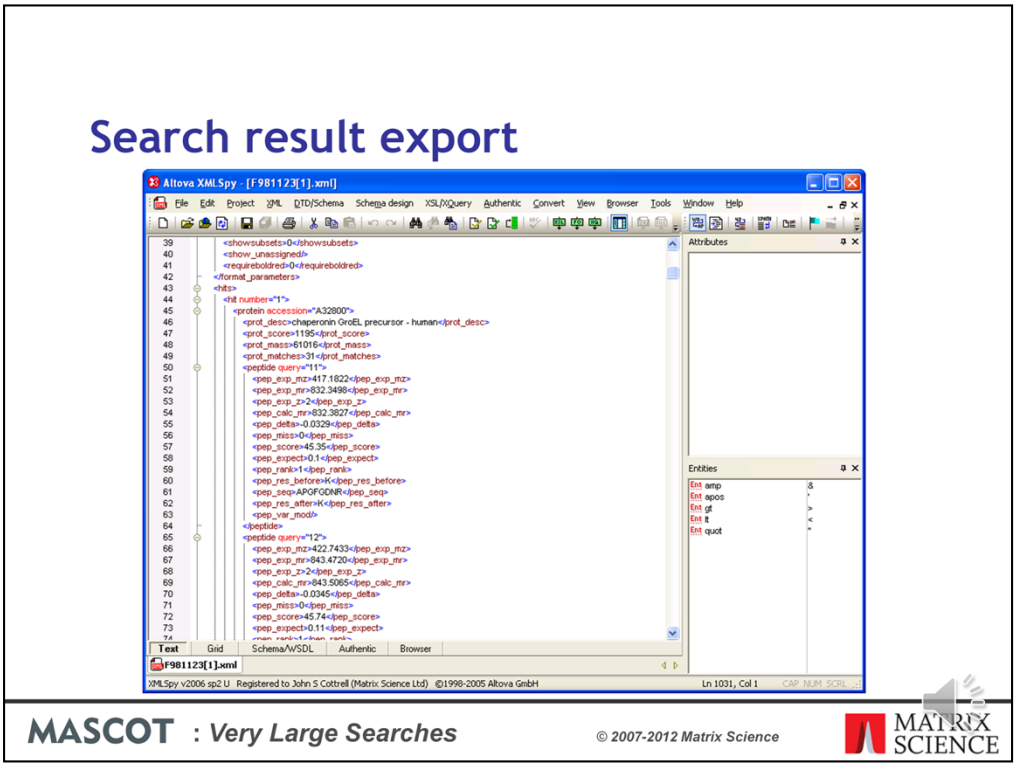
Search result export



MASCOT : *Very Large Searches*

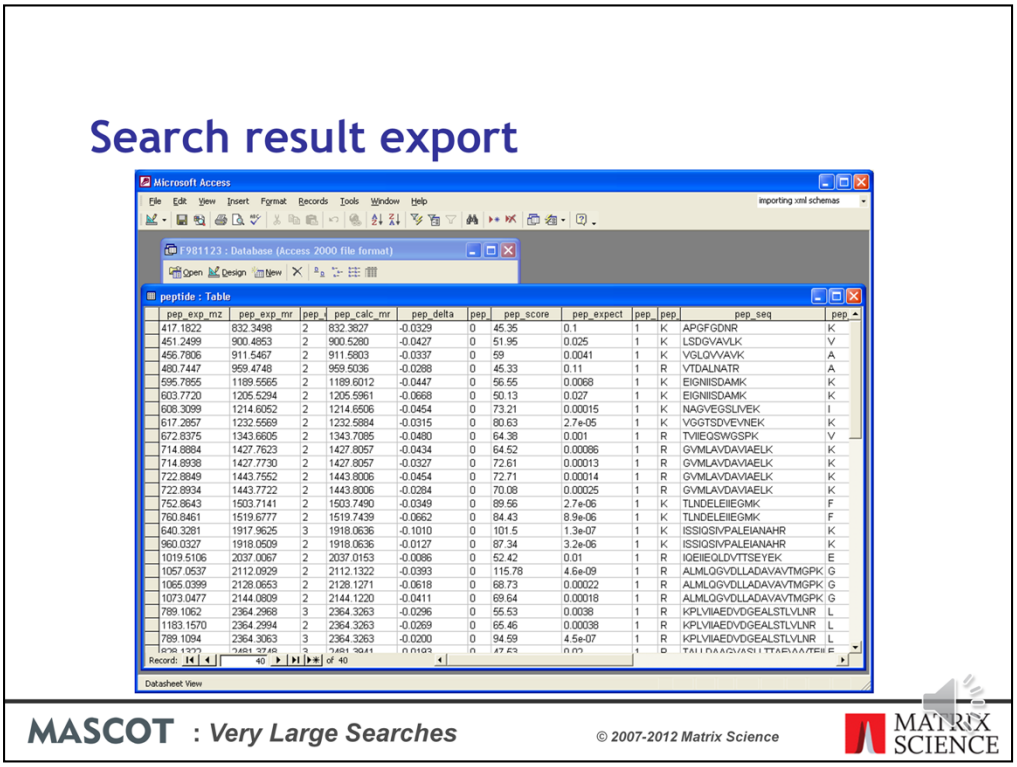
MATRIX SCIENCE

Much easier and safer than “screen scraping”



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.

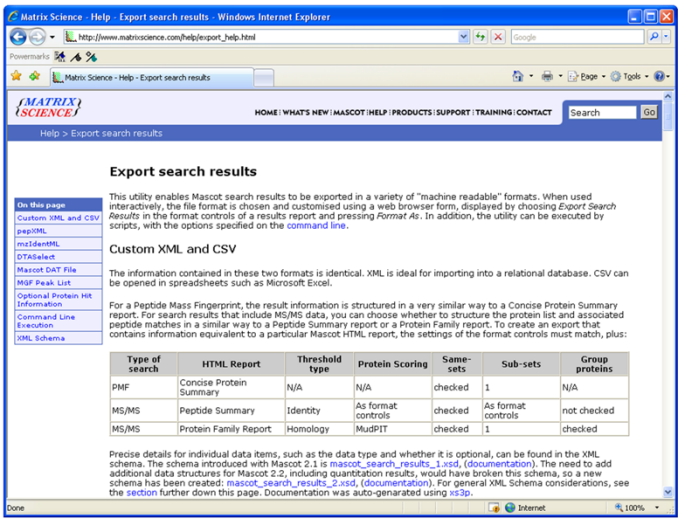
Please read the help for details.



pep_exp_mz	pep_exp_mr	pep_	pep_calc_mr	pep_delta	pep_	pep_score	pep_expect	pep_	pep_	pep_seq	pep_
417.1822	832.3498	2	832.3827	-0.0329	0	45.35	0.1	1	K	APGFGDNR	K
451.2499	900.4853	2	900.5280	-0.0427	0	51.95	0.025	1	K	LSDGVAVLK	V
456.7806	911.5467	2	911.5803	-0.0337	0	59	0.0041	1	K	VGLQVVAVK	A
480.7447	959.4748	2	959.5036	-0.0288	0	45.33	0.11	1	R	VTDALNATR	A
595.7855	1189.5565	2	1189.6012	-0.0447	0	56.55	0.0068	1	K	EIGNIISDAMK	K
603.7720	1205.5294	2	1205.5961	-0.0668	0	50.13	0.027	1	K	EIGNIISDAMK	K
608.3099	1214.6052	2	1214.6506	-0.0454	0	73.21	0.00015	1	K	NAGVEGSLIVEK	I
617.2857	1232.5569	2	1232.5884	-0.0315	0	80.63	2.7e-05	1	K	VGGTSDVEVNEK	K
672.8375	1343.6605	2	1343.7085	-0.0480	0	64.38	0.001	1	R	TVIIEQSWGSPK	V
714.8884	1427.7623	2	1427.8057	-0.0434	0	64.52	0.00086	1	R	GVMLAVDAVIAELK	K
714.8938	1427.7730	2	1427.8057	-0.0327	0	72.61	0.00013	1	R	GVMLAVDAVIAELK	K
722.8849	1443.7552	2	1443.8006	-0.0454	0	72.71	0.00014	1	R	GVMLAVDAVIAELK	K
722.8934	1443.7722	2	1443.8006	-0.0284	0	70.08	0.00025	1	R	GVMLAVDAVIAELK	K
752.8643	1503.7141	2	1503.7490	-0.0349	0	89.56	2.7e-06	1	K	TLNDELEIIEGMK	F
760.8461	1519.6777	2	1519.7439	-0.0662	0	84.43	8.9e-06	1	K	TLNDELEIIEGMK	F
640.3281	1917.9625	3	1918.0636	-0.1010	0	101.5	1.3e-07	1	K	ISSIQSIVPALEIANAHR	K
960.0327	1918.0509	2	1918.0636	-0.0127	0	87.34	3.2e-06	1	K	ISSIQSIVPALEIANAHR	K
1019.5106	2037.0067	2	2037.0153	-0.0086	0	52.42	0.01	1	R	IQEIIEQLDVTTSEYEK	E
1057.0537	2112.0929	2	2112.1322	-0.0393	0	115.78	4.6e-09	1	R	ALMLQGVDLLADAVAVTMGPK	G
1065.0399	2128.0653	2	2128.1271	-0.0618	0	68.73	0.00022	1	R	ALMLQGVDLLADAVAVTMGPK	G
1073.0477	2144.0809	2	2144.1220	-0.0411	0	69.64	0.00018	1	R	ALMLQGVDLLADAVAVTMGPK	G
789.1062	2364.2968	3	2364.3263	-0.0296	0	55.53	0.0038	1	R	KPLVIIAEDVDGEALSTLVLNR	L
1183.1570	2364.2994	2	2364.3263	-0.0269	0	65.46	0.00038	1	R	KPLVIIAEDVDGEALSTLVLNR	L
789.1094	2364.3063	3	2364.3263	-0.0200	0	94.59	4.5e-07	1	R	KPLVIIAEDVDGEALSTLVLNR	L

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

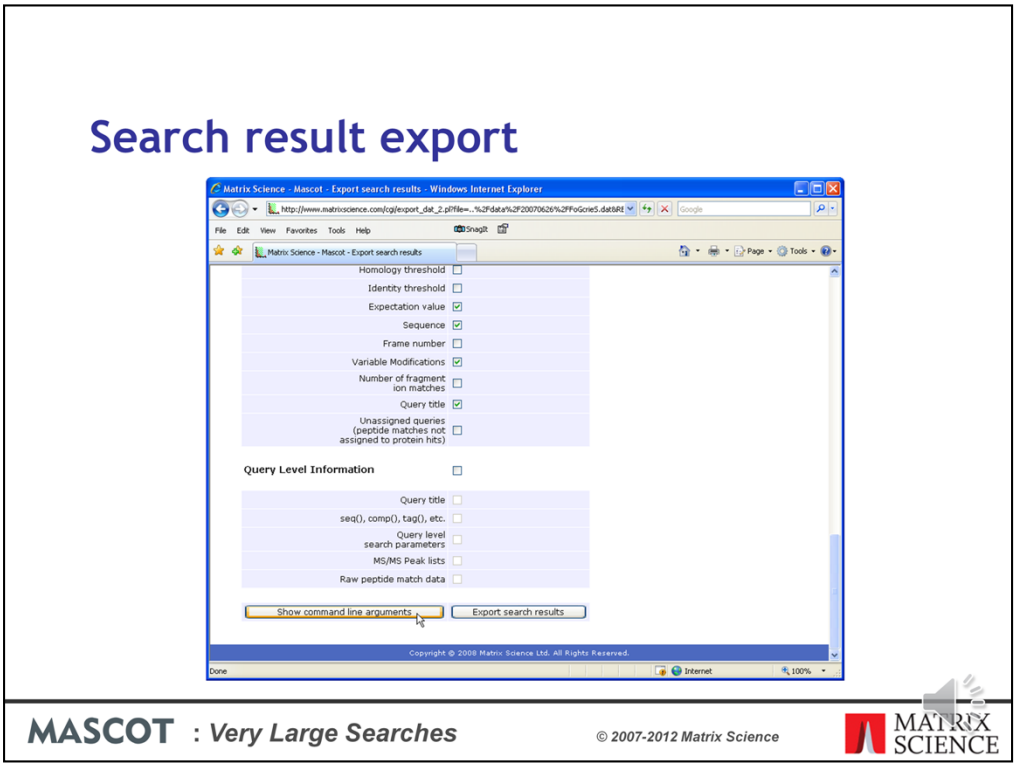
Search result export



MASCOT : *Very Large Searches*

MATRIX SCIENCE

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

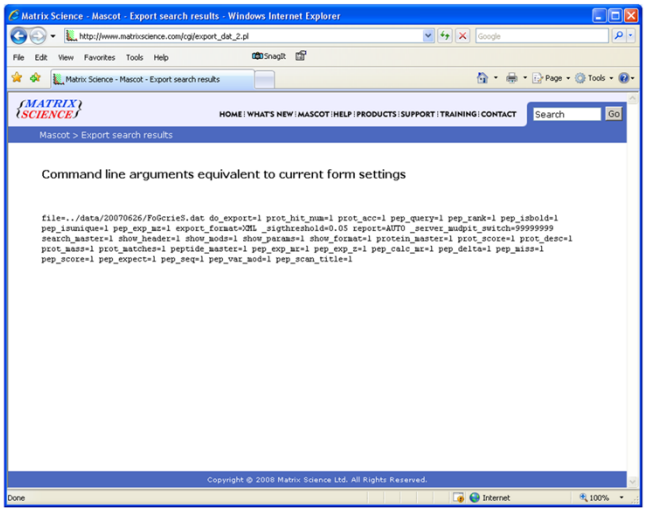


Which describes how the export script can be called from the command line or a shell prompt, as part of an automated pipeline.

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.