

Very Large Searches

MASCOT



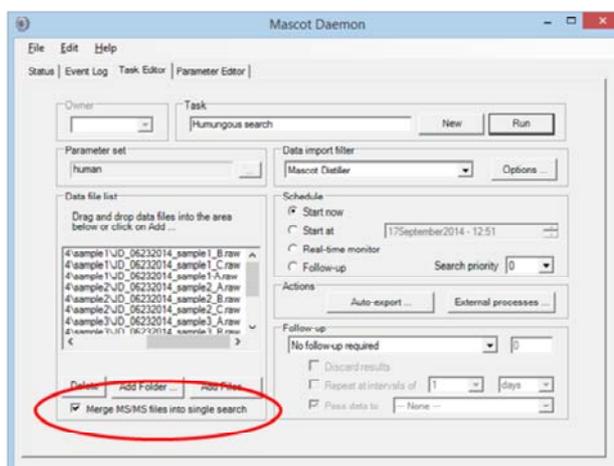
Topics

- 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

© 2007-2014 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*

© 2007-2014 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

Data files

Concatenating peak lists:

- MGF

Windows: copy



```
Command Prompt
C:\TEMP>copy *.mgf merge.out
```

Unix: cat



```
matrix@frill:~$ cat *.mgf > merge.out
```

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.

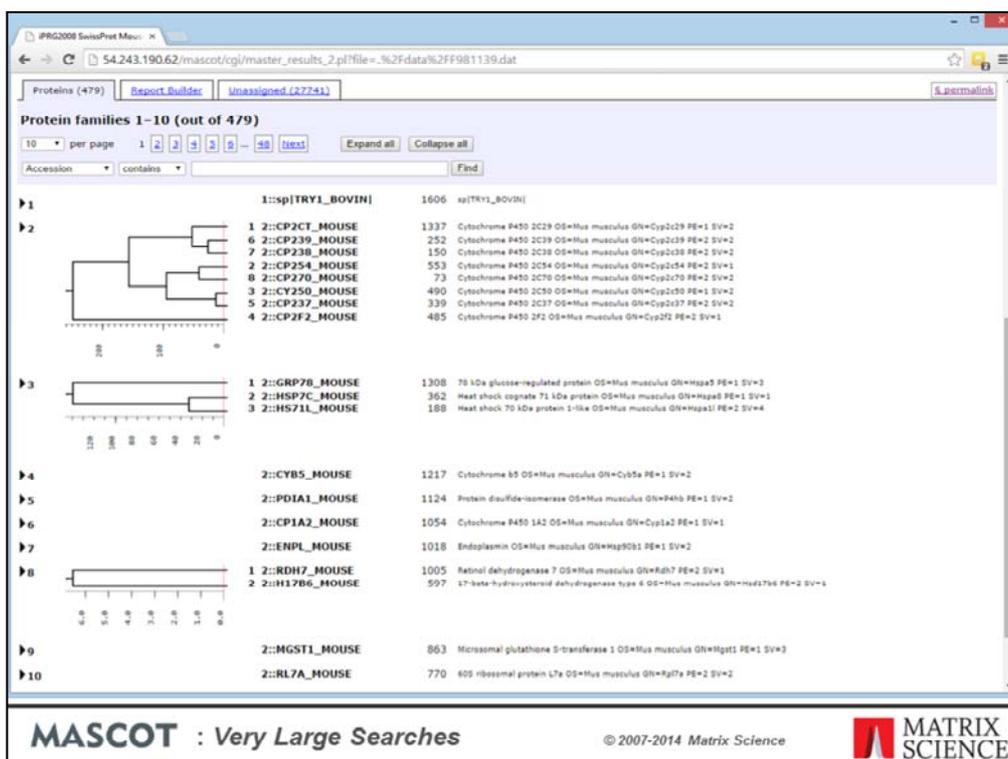
This is OK, but remember to reset these values as soon as you are able to. Otherwise, you might find you run out of memory or address space for your large searches

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

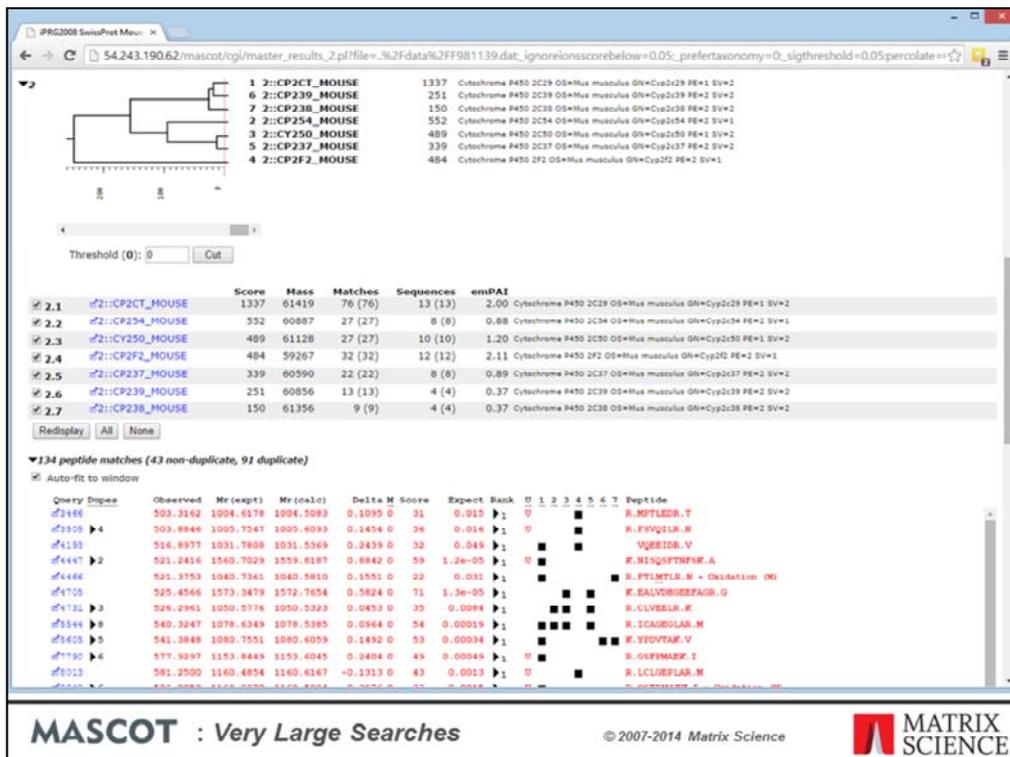
In early versions of Mascot, trying to display result reports for very large searches would often lead to problems with timeouts and running out of memory. To address this, the Protein Family Summary loads most of the information 'on demand'. This requires some index files to be created on the server,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



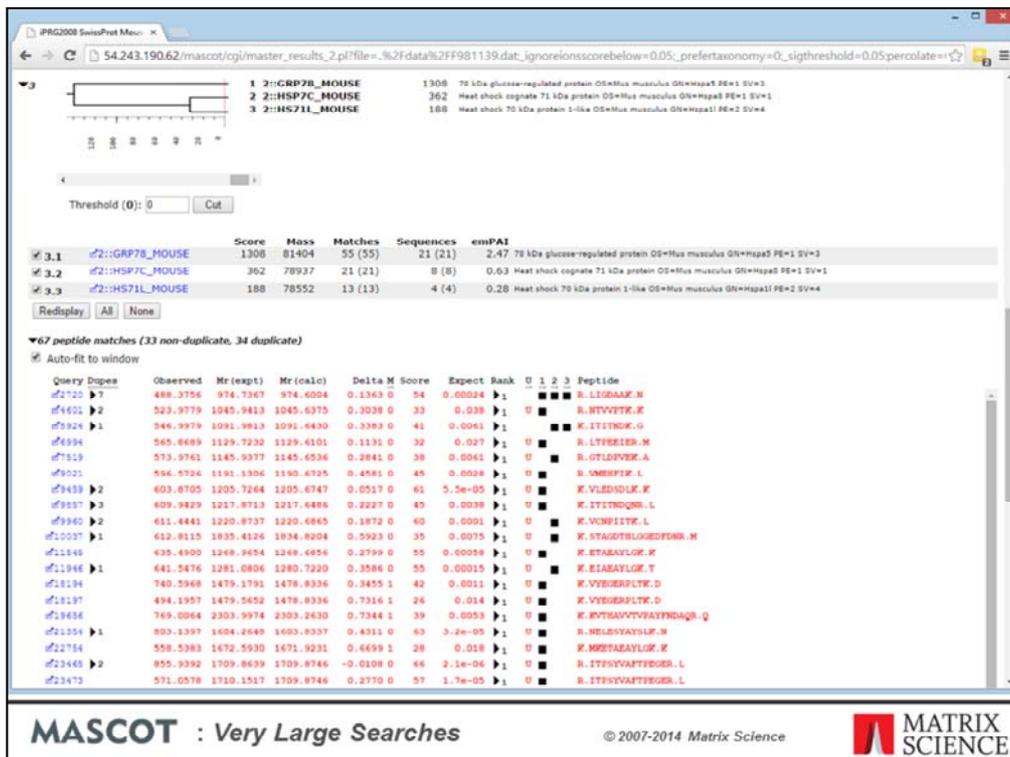
If there are 300 or more spectra, the Protein Family Summary is the default. This is the appearance of a typical family report immediately after loading. The body of the report consists of three tabs, one for protein families, one for Report Builder, and one for unassigned matches. The report is paged, 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. 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

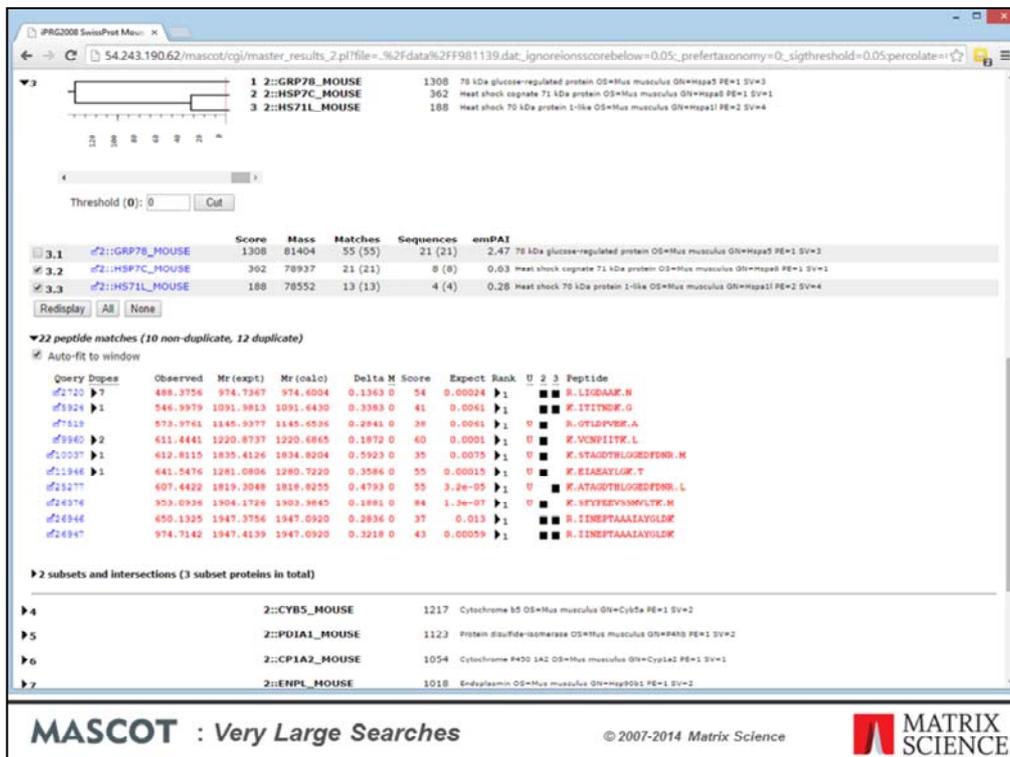


Moving down to family 3, the scale on the dendrogram is ions score, and HSP7C_MOUSE and HS71L_MOUSE join at a score of approximately 30. This represents the score of the significant matches that would have to be discarded in order to make one protein a sub-set of the other. These two proteins are much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

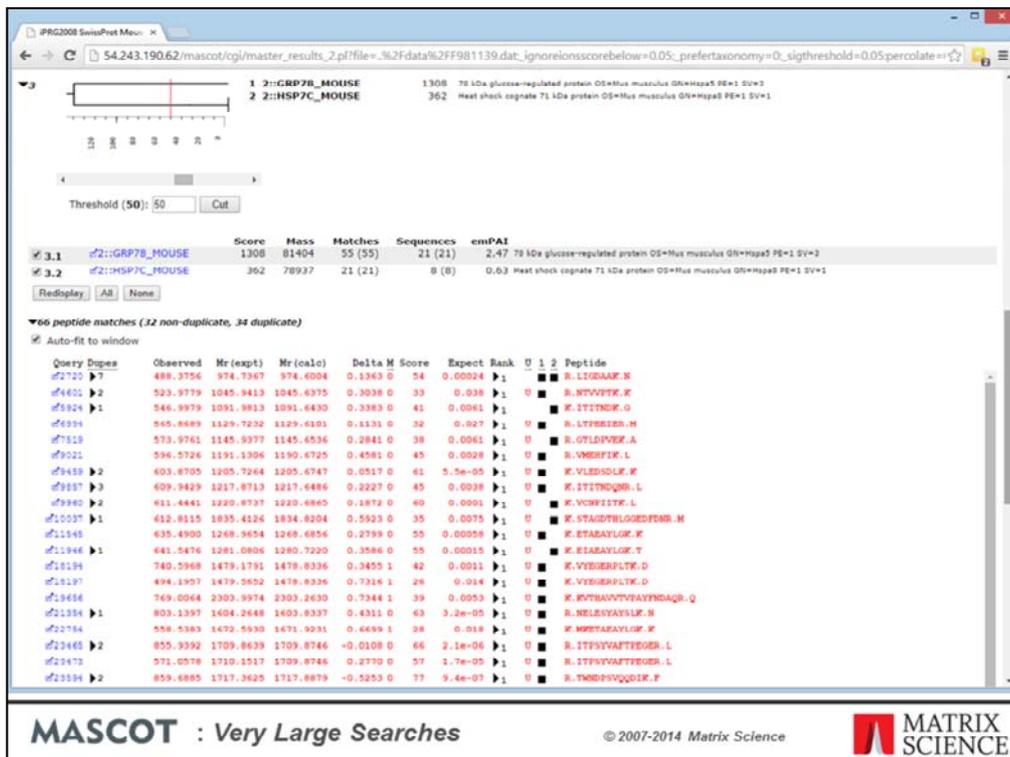
The table of peptide matches is similar to that found in the other result reports. Duplicate matches to the same sequence are collapsed into a single row. Click on the triangle to expand.

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 screenshot displays the Mascot search results interface. At the top, it shows the search parameters: "Proteins (448)", "Report Builder", and "Unassigned (30292)". The main section is titled "Protein families 41-50 (out of 448)". Below this, there are navigation controls for "per page" (set to 10) and "Find" buttons. A search bar contains the sequence "MNVLADALK".

The protein families list includes:

- 41 2::NB5R3_MOUSE 364 NADH-cytochrome b5 reductase 3 OS=Mus musculus GN=Cyb5r3 PE=1 SV=3
- 42 2::RS19_MOUSE 360 40S ribosomal protein S19 OS=Mus musculus GN=Rps19 PE=1 SV=3
- 43 2::CP2E1_MOUSE 358 Cytochrome b450 2E1 OS=Mus musculus GN=Cyp2e1 PE=3 SV=1
- 44 2::RL22_MOUSE 347 60S ribosomal protein L22 OS=Mus musculus GN=Rpl22 PE=1 SV=2
- 45 2::RS15A_MOUSE 344 40S ribosomal protein S15a OS=Mus musculus GN=Rps15a PE=1 SV=2

Below the protein list, a table shows peptide matches for the selected protein (45.1 f2::RS15A_MOUSE):

Query Dupes	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Expect	Rank	U	Peptide	
f2708	5	508.3777	1014.7407	1014.4308	0.1100	0	45	0.00053	1	K IYVNLTGK.L
f11285	5	631.9663	1261.9180	1261.7308	0.1872	0	77	2.4e-06	1	K MNVLADALK.S
f11274	1	631.8868	1261.7591	1261.7308	0.0284	0	(66)	1.8e-05	1	K MNVLADALK.S
f11276	1	631.8914	1261.7682	1261.7308	0.0375	0	(59)	3.7e-05	1	K MNVLADALK.S
f11283	1	631.9416	1261.8686	1261.7308	0.1379	0	(59)	0.00013	1	K MNVLADALK.S
f11287	1	632.0080	1262.0014	1261.7308	0.2704	0	(42)	0.0048	1	K MNVLADALK.S
f11288	1	632.0218	1262.0291	1261.7308	0.2983	0	(43)	6.4e-05	1	K MNVLADALK.S
f11604	1	636.4751	1270.9355	1270.6904	0.2452	0	38	0.03	1	K WQNLIPK.R
f11780	1	639.8954	1277.7742	1277.7257	0.0505	0	50	0.00084	1	K MNVLADALK.S + Oxidation (M)
f11780	1	639.8899	1277.9652	1277.7257	0.2398	0	(48)	0.00054	1	K MNVLADALK.S + Oxidation (M)

The interface also includes a "MASCOT : Very Large Searches" logo and "MATRIX SCIENCE" branding at the bottom.

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.

IPRG2008 SwissProt Mac...
 54.243.190.62/mascot/cgi/master_results_2.pl?file=%2Fdata%2FF981139.dat_ignoreionsscorebelow=0.05_prefertaxonomy=0_sigthreshold=0.05_percolate=1
 Proteins (448) Report Builder Unassigned (30397) [Link](#)

Protein hits (476 proteins)
 Columns: Standard (12 out of 16)
 Filters: (none)
 Export as CSV

*Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	CRAP	f1:sp TRY1_BOVIN	1606	28266	48	48	7	7	2.86	sp TRY1_BOVIN
2	1	SwissProt	f2: CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1
2	2	SwissProt	f2: CP2S4_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2
2	3	SwissProt	f2: CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1
2	4	SwissProt	f2: CP2F2_MOUSE	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=
2	5	SwissProt	f2: CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2
2	6	SwissProt	f2: CP339_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2
2	7	SwissProt	f2: CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2
3	1	SwissProt	f2: GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa1
3	2	SwissProt	f2: HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hsp
4	1	SwissProt	f2: CYB5_MOUSE	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	f2: PDI1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1
6	1	SwissProt	f2: CP1A2_MOUSE	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV
2	1	SwissProt	f2: ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmic reticulum protein OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1	SwissProt	f2: RDH7_MOUSE	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV
8	2	SwissProt	f2: H17B6_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus muscul
9	1	SwissProt	f2: MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=
10	1	SwissProt	f2: RL7A_MOUSE	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2
11	1	SwissProt	f2: RLA0_MOUSE	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein PO OS=Mus musculus GN=Rplp0 P
12	1	SwissProt	f2: CP2A2_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1
12	2	SwissProt	f2: CP2A5_MOUSE	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 PE=2 SV
13	1	SwissProt	f2: ACSL1_MOUSE	749	86078	38	38	18	18	1.00	Long-chain-fatty-acid-CoA ligase 1 OS=Mus musculus GN=Acsl
13	2	SwissProt	f2: ACSL5_MOUSE	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid-CoA ligase 5 OS=Mus musculus GN=Acsl
14	1	SwissProt	f2: RL13_MOUSE	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2
15	1	SwissProt	f2: PDI3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdi3 PE
16	1	SwissProt	f2: CP3A8_MOUSE	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1
17	1	SwissProt	f2: UDB17_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2
17	2	SwissProt	f2: UD11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a
17	3	SwissProt	f2: UD16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a
18	1	SwissProt	f2: EST3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2

MASCOT : Very Large Searches © 2007-2014 Matrix Science 

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

Protein hits (476 proteins)

Columns: Standard (12 out of 16)

Filters: (none)

Export as CSV

Family	Mass	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	Num. of matches	1606	28266	48	48	7	7	2.86	sp TRY1_BOVIN
2	Num. of significant matches	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1
2	Num. of unique sequences	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2
2	Num. of significant unique sequences	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1
2	emPAI	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1
2	Sequence coverage	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2
2	pl	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2
2	Description	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2
2	Methylation (C)	1308	81404	55	55	21	21	2.47	78 kDa glycine-regulated protein OS=Mus musculus GN=Hsp78
2	ITRAQ4plex (K)	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hsp70
4	1 SwissProt	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1 SwissProt	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1
6	1 SwissProt	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
2	1 SwissProt	1018	103744	63	63	19	19	1.53	Endoplasmic reticulum chaperone protein OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1 SwissProt	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
8	2 SwissProt	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus GN=Hsd17b6 PE=1 SV=2
9	1 SwissProt	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=UGT1A1 PE=1 SV=1
10	1 SwissProt	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=1
11	1 SwissProt	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 P
12	1 SwissProt	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1 SV=1
12	2 SwissProt	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 PE=2 SV=1
12	1 SwissProt	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid-CoA ligase 1 OS=Mus musculus GN=Acsl1
13	2 SwissProt	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid-CoA ligase 5 OS=Mus musculus GN=Acsl5
14	1 SwissProt	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2 SV=1
15	1 SwissProt	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE=1 SV=1
16	1 SwissProt	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1 SV=1
17	1 SwissProt	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2b17

MASCOT : Very Large Searches © 2007-2014 Matrix Science

MATRIX SCIENCE

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

IPRG2008 SwissProt Musc...

54.243.190.62/mascot/cgi/master_results_2.pl?file=%2Fdata%2FF981139.dat_ignoreionsscorebelow=0.05_prefertaxonomy=0_sigthreshold=0.05_percolate=1

Proteins (446) Report Builder Unassigned (30327) [Permalink](#)

Protein hits (229 proteins)

Columns: 22 out of 207

Filters: "Num. of significant sequences" >= 2

Export as CSV

Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	enPI1	Description
1	1	CRAP	f1::sp TRY1_BOVIN	1006	28266	48	48	7	7	2.86	sp TRY1_BOVIN
2	1	SwissProt	f2::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV
2	2	SwissProt	f2::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV
2	3	SwissProt	f2::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV
2	4	SwissProt	f2::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	f2::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV
2	6	SwissProt	f2::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV
2	7	SwissProt	f2::CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV
2	1	SwissProt	f2::IGRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 P
2	2	SwissProt	f2::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8
4	1	SwissProt	f2::CYB5_MOUSE	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	f2::PDI1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV
6	1	SwissProt	f2::CYP1A2_MOUSE	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
2	1	SwissProt	f2::ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmic reticulum protein OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1	SwissProt	f2::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	f2::H17B6_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus
9	1	SwissProt	f2::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mt
10	1	SwissProt	f2::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	f2::RLA0_MOUSE	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 PE=
12	1	SwissProt	f2::CP2A1_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1 SV
12	2	SwissProt	f2::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	f2::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	f2::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	f2::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	f2::PDI3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE=1
16	1	SwissProt	f2::CP3A8_MOUSE	666	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1 SV
17	1	SwissProt	f2::UGT2B1_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2b1
17	2	SwissProt	f2::UGT11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a1 f
17	3	SwissProt	f2::UGT16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a6 f
18	1	SwissProt	f2::EST3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2

MASCOT : Very Large Searches © 2007-2014 Matrix Science **MATRIX SCIENCE**

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

IPRG2008 SwissProt Mac...
 54.243.190.62/mascot/cgi/master_results_2.pl?file=%2Fdata%2FF981139.dat_ignoreionscorebelow=0.05_prefertaxonomy=0_sigthreshold=0.05_percolate=1
 Proteins (448) Report Builder Unassigned (30292) 4 permalink

Protein hits (228 proteins)
 Columns: Standard (12 out of 16)
 Filters: (NOT(Database is cRAP) AND "Num. of significant sequences" >= 2)
 Export as CSV

Family	#	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	empP1	Description
2	1	SwissProt	#2::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=
2	2	SwissProt	#2::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=
2	3	SwissProt	#2::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=
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=1
2	5	SwissProt	#2::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV=
2	6	SwissProt	#2::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=
2	7	SwissProt	#2::CP238_MOUSE	150	61350	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=
3	1	SwissProt	#2::GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE
3	2	SwissProt	#2::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	#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::PDI1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=Pdia1 PE=1 SV=
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=1
7	1	SwissProt	#2::ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmic reticulum protein OS=Mus musculus GN=Enpl PE=1 SV=2
8	1	SwissProt	#2::RDH7_MOUSE	1005	38955	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
8	2	SwissProt	#2::H17B_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus C
9	1	SwissProt	#2::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgt
10	1	SwissProt	#2::RL7_MOUSE	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=
11	1	SwissProt	#2::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	#2::CP2A2_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1 SV=
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=1
12	1	SwissProt	#2::ACSL1_MOUSE	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acsl1 P
12	2	SwissProt	#2::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	#2::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	#2::PDI3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE=1
16	1	SwissProt	#2::CP3A8_MOUSE	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1 SV=
12	1	SwissProt	#2::UGT17_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2b17
12	2	SwissProt	#2::UGT11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a1 PE
12	3	SwissProt	#2::UGT16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a6 PE
18	1	SwissProt	#2::CE3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2
19	1	SwissProt	#2::RL4_MOUSE	650	55568	34	34	11	11	1.59	60S ribosomal protein L4 OS=Mus musculus GN=Rpl4 PE=1 SV=3

MASCOT : Very Large Searches © 2007-2014 Matrix Science 

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

The screenshot shows the Mascot Report Manager interface. At the top, it displays the URL: `54.243.190.62/mascot/cgi/master_results_2.pl?file=%2Fdata%2FF981139.dat_ignoreionscorebelow=0.05_prefertaxonomy=0_sigthreshold=0.05_percolate=1`. Below the URL, there are tabs for 'Proteins (448)', 'Report Builder', and 'Unassigned (30297)'. A 'permalink' button is also visible.

The main section is titled 'Protein hits (228 proteins)'. Underneath, it says 'Columns: Standard (12 out of 16)'. There are two panes: 'Enabled' and 'Available'. The 'Enabled' pane lists columns such as Family, Member, Database, Accession, Score, Mass, Num. of matches, Num. of significant matches, Num. of sequences, Num. of significant sequences, emPAI, and Description. The 'Available' pane lists 'Protein hits', 'Num. of unique sequences', 'Num. of significant unique sequences', and 'Sequence coverage'. Buttons for '+', '-', and 'Apply' are present between the panes.

Below the panes, there are filter settings: 'Filters: (NOT(Database is cRAP) AND "Num. of significant sequences" >= 2)'. There is also an 'Export as CSV' button.

The main data table has the following columns: Family, M, DB, Accession, Score, Mass, Matches, Match(sig), Sequences, Seq(sig), emPAI, and Description. The table contains several rows of protein data, including entries for Cytochrome P450 and Heat shock coonate.

At the bottom of the interface, there is a banner for 'MASCOT : Very Large Searches' with the copyright notice '© 2007-2014 Matrix Science' and the 'MATRIX SCIENCE' logo.

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

Microsoft Excel - data_20120501_F001467_dat_r1_reportbuilder.csv

Filters: Num. of significant sequences >= 2

Family Member	Database	Accession	Score	Mass	Num. of matches	Num. of significant matches	Num. of sequences	Num. of significant sequences	emPAI	Description	
31	1	IPRG_2012	P00925	2140	46942	148	100	53	43	44.71	Enolase 2 OS=Saccharomyces cere
32	1	IPRG_2012	P00924	1059	46844	71	46	35	27	7.47	Enolase 1 OS=Saccharomyces cere
33	2	IPRG_2012	P00549	1933	54909	133	87	56	43	18.28	Pyruvate kinase 1 OS=Saccharomy
34	3	IPRG_2012	P40150	1613	66668	105	66	66	45	11.76	Heat shock protein SSB2 OS=Sacch
35	3	IPRG_2012	P11484	1590	66732	103	65	64	44	11.12	Heat shock protein SSB1 OS=Sacch
36	4	IPRG_2012	P10592	1591	69599	107	57	52	32	5.01	Heat shock protein SSA2 OS=Sacch
37	4	IPRG_2012	P10591	1161	69786	85	44	48	26	3.02	Heat shock protein SSA1 OS=Sacch
38	4	IPRG_2012	P16474	233	74479	23	8	17	6	0.32	78 kDa glucose-regulated protein hor
39	5	IPRG_2012	P00330	1453	37282	73	51	32	25	13.48	Alcohol dehydrogenase 1 OS=Sacch
40	5	IPRG_2012	P07246	101	40743	14	5	7	3	0.29	Alcohol dehydrogenase 3, mitochon
41	6	IPRG_2012	P00560	1382	44768	102	58	54	33	12.75	Phosphoglycerate kinase OS=Sacch
42	7	IPRG_2012	P00359	1361	35838	76	54	31	25	12.29	Glyceraldehyde-3-phosphate dehydr
43	7	IPRG_2012	P00358	1242	35938	69	48	29	24	9.89	Glyceraldehyde-3-phosphate dehydr
44	7	IPRG_2012	P00360	525	35842	30	20	14	12	2.47	Glyceraldehyde-3-phosphate dehydr
45	7	IPRG_2012	P04406	41	36201	4	2	4	2	0.21	Glyceraldehyde-3-phosphate dehydr
46	8	IPRG_2012	P06169	1289	61685	44	41	28	26	4.7	Pyruvate decarboxylase isozyme 1 C
47	9	IPRG_2012	P00960	1031	27692	67	44	32	26	34.97	Phosphoglycerate mutase 1 OS=Sa
48	10	IPRG_2012	P07261	1015	15881	51	38	16	13	22.71	40S ribosomal protein S19-B OS=Sa
49	10	IPRG_2012	P07200	1014	15907	51	38	16	13	22.71	40S ribosomal protein S19-A OS=Sa
50	11	contaminants	P00761	922	25078	37	27	7	6	2.89	SWISS-PROT P00761 TRYF_PIG Tr
51	12	IPRG_2012	P32324	784	93686	49	33	33	23	1.44	Elongation factor 2 OS=Saccharomy
52	13	IPRG_2012	P16521	771	116727	62	33	47	30	1.52	Elongation factor 3A OS=Saccharom
53	14	IPRG_2012	P05319	765	10739	38	29	10	9	95.65	60S acidic ribosomal protein P2- α
54	15	IPRG_2012	Q03048	721	15949	28	23	17	14	17.82	Cofin OS=Saccharomyces cerevisi
55	16	IPRG_2012	P00046	719	9797	42	29	15	12	207.43	40S ribosomal protein S21-A OS=Sa
56	16	IPRG_2012	O3E754	694	9811	41	28	15	12	148.28	40S ribosomal protein S21-B OS=Sa

MASCOT : Very Large Searches © 2007-2014 Matrix Science 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

Large search results in 2.2 and earlier

Never Peptide
Important
Simplifies

Select Summary Report

Format As	Select Summary (protein hits) Help
Significance threshold p<	0.05 Max. number of hits AUTO
Standard scoring	<input type="radio"/> MudPIT scoring <input checked="" type="radio"/> Ions score cut-off 0.5 <input type="checkbox"/> Show sub-sets
Show pop-ups	<input type="radio"/> Suppress pop-ups <input checked="" type="radio"/> Sort unassigned Decreasing Score <input checked="" type="checkbox"/> Require bold red

Reduces memory
Simplifies

http://.../master_results.pl?file=../data/20060202/F000123.dat
&REPTYPE=select &REPORT=AUTO &_showpopups=FALSE
&_ignoreionsscorebelow=0.5&_requireboldred=1

MASCOT : Very Large Searches
© 2007-2014 Matrix Science

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

Address http://141-dnc/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

MASCOT : Very Large Searches

© 2007-2014 Matrix Science

MATRIX SCIENCE

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

Reporting large search results

???

Select Summary Report

Format As	Select Summary (protein hits) ▼	Help					
Significance threshold p <	0.05	Max. number of hits	AUTO				
Standard scoring	<input type="radio"/>	MudPIT scoring	<input checked="" type="radio"/>	Ions score cut-off	0.5	Show sub-sets	<input type="checkbox"/>
Show pop-ups	<input type="radio"/>	Suppress pop-ups	<input checked="" type="radio"/>	Sort unassigned	Decreasing Score ▼	Require bold red	<input checked="" type="checkbox"/>

MASCOT : *Very Large Searches*

© 2007-2014 Matrix Science



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::IP100022283](#) Mass: 3832803 Score: 181 Matches: 51(0) Sequences: 48(0)

Tax_Id=9606 Gene_Symbol=TTN Isoform 2 of Titin

Query	Observed	Hr(expt)	Hr(calc)	ppm	Miss	Score	Expect	Rank	Unique	Peptide
28	359.7341	717.4537	717.4537	-0.09	0	7	4.2	5	U	R.LFATVR.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.EPFLTE.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.N
1029	473.2757	944.5368	944.5331	3.97	1	3	21	3	U	R.EPPSFYK.I
1058	475.7505	949.4064	949.4069	-0.47	0	4	22	5	U	R.SSVLSLGGK.P
1066	476.2790	950.5433	950.5425	0.94	0	1	23	4	U	R.PLYDLQVR.E

MASCOT : Very Large Searches

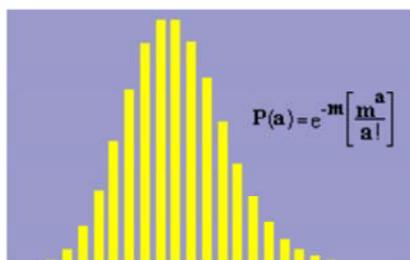
© 2007-2014 Matrix Science



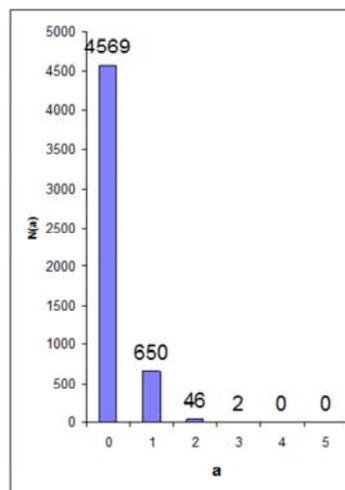
With standard peptide summary scoring, the protein score is essentially the sum of the ions scores of all the peptides assigned to the protein. Where there are duplicate matches to the same peptide, the highest scoring match is used. A correction is applied based on the number of candidate peptides that 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

Protein Inference



- Huge MudPIT data set
- Search Swiss-Prot using drosophila taxonomy filter (5268 entries)
- 75,000 matches with 1% FDR
- i.e. 750 false matches



MASCOT : Very Large Searches

© 2007-2014 Matrix Science



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 oversimplification. 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::IP100023283  Mass: 3832803  Score: 0  Matches: 51(0)  Sequences: 48(0)
Tax_Id=9606 Gene_Symbol=TTN Isoform 2 of Titin
Query  Observed  Hr(expt)  Hr(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.EFLTLR.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.VTLVDFVTR.N
840  459.2484  916.4821  916.4767  5.98  0  2  29  2  U  K.GVEFMVPR.L
844  459.7299  917.4452  917.4454  -0.24  0  4  15  6  U  K.ELEETAR.M
1029  473.2757  944.5368  944.5331  3.97  1  3  21  3  U  R.EPPSEIKK.I
1058  475.7505  949.4864  949.4869  -0.47  0  4  22  5  U  R.SSVLSWGR.P
1066  476.2790  950.5433  950.5425  0.94  0  1  23  4  U  R.PLTDLQVR.E

```

MASCOT : Very Large Searches

© 2007-2014 Matrix Science

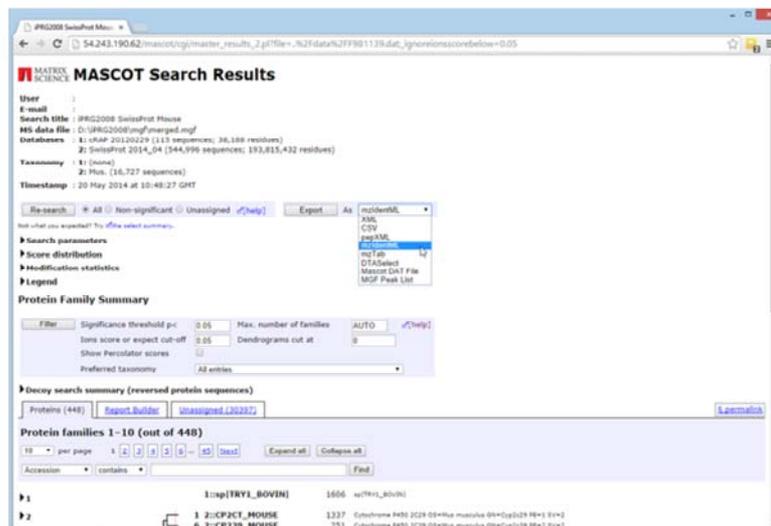


To avoid this problem, we use MudPIT protein scoring, in which the score for each peptide match is not its absolute score, but the amount that it is above the threshold. Therefore, matches with a score below the threshold do not contribute to the score. The MudPIT protein score is the sum of the score 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.

Search result export



MASCOT : Very Large Searches

© 2007-2014 Matrix Science



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

The screenshot shows the Matrix Science Mascot web interface. At the top, there is a navigation bar with the Matrix Science logo and a search box. Below the navigation bar, there are links for 'Home', 'Access Mascot Server', 'Database search help', and 'Contact'. The main content area is titled 'Export search results' and contains a form with various options for exporting search results. The 'Export format' dropdown menu is open, showing options: 'mascotXML', '2DLC', 'CSV', 'PeakList', 'nciTab', 'CITAttachment', 'Mascot DAT File', and 'LSDI Peak List'. Other options include 'Significance threshold p:', 'Target FDR', 'Max. number of hits', 'Protein scoring', 'Include same-set protein hits', 'Include sub-set protein hits', 'Group protein families', 'Show Percolator scores', and 'Preferred Taxonomy'. The 'Export format' dropdown is currently set to 'mascotXML'.

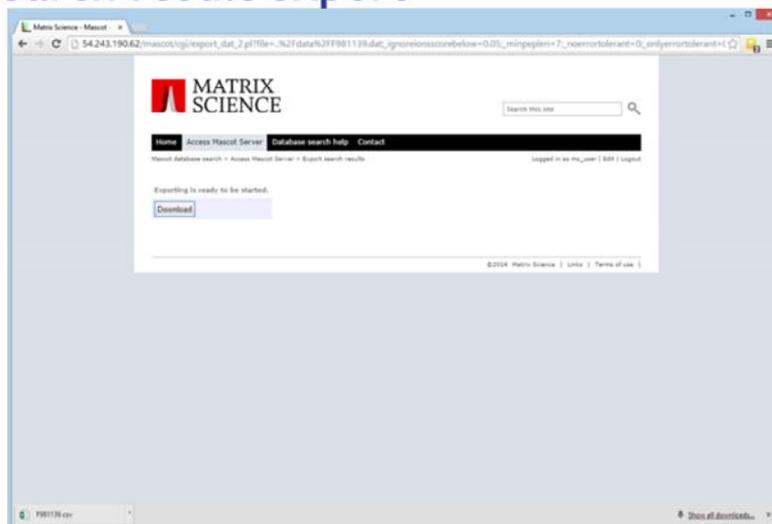
MASCOT : Very Large Searches

© 2007-2014 Matrix Science



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



MASCOT : Very Large Searches

© 2007-2014 Matrix Science



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. Once the download is finished, you can open it into Excel:

Search result export

prot_hit	prot_acc	prot_desc	prot_score	prot_mass	prot_match	pep_query	pep_exp_f	pep_exp_r	pep_exp_t	pep_calc	pep_delta	pep
1	A32600	chaperonin	1195	61016	31	11	417.1822	832.3490	2	832.3827	-0.0329	
12						12	422.7439	843.472	2	843.5065	-0.0345	
13						13	430.7328	859.451	2	859.4837	-0.0327	
15						15	451.2499	900.4853	2	900.528	-0.0427	
16						16	456.7806	911.5467	2	911.5903	-0.0337	
21						21	480.7447	959.4748	2	959.5036	-0.0288	
24						24	595.7855	1189.557	2	1189.601	-0.0447	
25						25	603.772	1205.529	2	1205.596	-0.0668	
26						26	608.7990	1214.605	2	1214.651	-0.0454	

MASCOT : Very Large Searches

© 2007-2014 Matrix Science



Much easier and safer than “screen scraping”

Search result export

pep_exp_mz	pep_exp_mv	pep_calc_mv	pep_delta	pep	pep_score	pep_expect	pep	pep	pep_seq
417.1822	832.3498	2 832.3827	-0.0329	0	45.35	0.1	1	K	APGFQDNR
451.2499	900.4953	2 900.5280	-0.0427	0	51.95	0.026	1	K	LSDGVAVLK
456.7806	911.5467	2 911.5803	-0.0337	0	59	0.0041	1	K	VGLGVAVLK
480.7447	969.4748	2 969.5036	-0.0289	0	45.33	0.11	1	R	VTDALNATR
595.7855	1189.5585	2 1189.6012	-0.0447	0	56.55	0.0068	1	K	EIGNISDAMK
603.7720	1206.6294	2 1206.6961	-0.0668	0	60.13	0.027	1	K	EIGNISDAMK
608.3099	1214.6052	2 1214.6506	-0.0454	0	73.21	0.00015	1	K	NAGVEGSLVEK
617.2657	1232.5569	2 1232.5884	-0.0315	0	80.63	2.7e-05	1	K	VGOTSQVLENEK
672.8375	1343.6605	2 1343.7085	-0.0480	0	64.38	0.001	1	R	TVIEGSGVSPK
714.8804	1427.7623	2 1427.8057	-0.0434	0	64.62	0.00006	1	R	GVMLAVDAVAELK
714.8938	1427.7730	2 1427.8057	-0.0327	0	72.61	0.00013	1	R	GVMLAVDAVAELK
722.8849	1443.7952	2 1443.8006	-0.0454	0	72.71	0.00014	1	R	GVMLAVDAVAELK
722.8934	1443.7722	2 1443.8006	-0.0284	0	70.08	0.00025	1	R	GVMLAVDAVAELK
752.8643	1503.7141	2 1503.7490	-0.0349	0	89.96	2.7e-06	1	K	TUNDELEIEGMK
760.8461	1519.6777	2 1519.7439	-0.0662	0	84.43	8.9e-06	1	K	TUNDELEIEGMK
840.3201	1917.9625	3 1918.0636	-0.1010	0	101.5	1.3e-07	1	K	ISSISGIVPALEIANHR
960.0227	1918.0509	2 1918.0636	-0.0127	0	87.34	3.2e-06	1	K	ISSISGIVPALEIANHR
1019.5106	2037.0067	2 2037.0153	-0.0086	0	52.42	0.01	1	R	ISEIEGLDVTSEYEK
1057.0637	2112.0929	2 2112.1322	-0.0393	0	115.78	4.6e-09	1	R	ALMLGGVLLADAVAVTMGPK
1065.0399	2128.0653	2 2128.1271	-0.0618	0	68.73	0.00022	1	R	ALMLGGVLLADAVAVTMGPK
1073.0477	2144.0809	2 2144.1220	-0.0411	0	69.64	0.00018	1	R	ALMLGGVLLADAVAVTMGPK
789.1062	2364.2968	3 2364.3263	-0.0296	0	55.53	0.0038	1	R	KPLVIAEDVDGEALSTLVNLR
1183.1570	2364.2994	2 2364.3263	-0.0269	0	65.46	0.00038	1	R	KPLVIAEDVDGEALSTLVNLR
789.1094	2364.3063	3 2364.3263	-0.0200	0	94.59	4.5e-07	1	R	KPLVIAEDVDGEALSTLVNLR
1078.1277	2481.3248	3 2481.3641	-0.0403	0	47.63	0.03	1	D	TALLPAAGVADLIITFAEAATEE

MASCOT : Very Large Searches

© 2007-2014 Matrix Science



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

Export search results

This utility enables MASCOT search results to be exported in a variety of "machine readable" formats. When used interactively, the file format is chosen and customized using a web browser form, displayed by choosing **Export Search Results** in the format controls of a results report and pressing **Format as...** In addition, the utility can be executed by scripts, with the options specified on the command line.

Custom XML and CSV

The information contained in these two formats is identical. XML is ideal for importing into a relational database. CSV can be opened in spreadsheets such as Microsoft Excel.

For a Peptide Mass Fingerprint, the result information is structured in a very similar way to a Concise Protein Summary report. For search results that include MS/MS data, you can choose whether to structure the protein list and associated peptide matches in a similar way to a Peptide Summary report or a Protein Family report. To create an export that contains information equivalent to a particular MASCOT HTML report, the settings of the format controls must match, also.

Type of search	HTML Report	Threshold type	Protein Scoring	Score-sets	Sub-sets	Group proteins
PMF	Concise Protein Summary	N/A	N/A	checked	1	N/A
MS/MS	Peptide Summary	Identity	As format controls	checked	As format controls	not checked
MS/MS	Protein Family Report	Homology	MS/MS	checked	1	checked

MASCOT : Very Large Searches

© 2007-2014 Matrix Science



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

Search result export



MASCOT : *Very Large Searches*

© 2007-2014 Matrix Science



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



MASCOT : *Very Large Searches*

© 2007-2014 Matrix Science



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